Understanding and improving high-throughput sequencing data production and analysis

Der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet

Informatik

vorgelegt von

M. Sc. hon. Martin Kircher
geboren am 30. September 1983 in Erfurt

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Anton Nekrutenko (PennState University, USA)
2. Professor Dr. Peter F. Stadler (Leipzig)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 11. Juli 2011 mit dem Gesamtprädikat *summa cum laude*.

IDT Exhibit 1061.0001

Kircher, Martin

Understanding and improving high-throughput sequencing data production and analysis
Max-Planck Institute for Evolutionary Anthropology,
Leipzig University, Germany,
Dissertation 2011

226 pages, 254 references, 66 figures, 23 tables

IDT Exhibit 1061.0002

Preface

Abstract

Advances in DNA sequencing revolutionized the field of genomics over the last 5 years. New sequencing instruments make it possible to rapidly generate large amounts of sequence data at substantially lower cost. These high-throughput sequencing technologies (e.g. Roche 454 FLX, Life Technology SOLiD, Dover Polonator, Helicos HeliScope and Illumina Genome Analyzer) make whole genome sequencing and resequencing, transcript sequencing as well as quantification of gene expression, DNA-protein interactions and DNA methylation feasible at an unanticipated scale.

In the field of evolutionary genomics, high-throughput sequencing permitted studies of whole genomes from ancient specimens of different hominin groups. Further, it allowed large-scale population genetics studies of present-day humans as well as different types of sequence-based comparative genomics studies in primates. Such comparisons of humans with closely related apes and hominins are important not only to better understand human origins and the biological background of what sets humans apart from other organisms, but also for understanding the molecular basis for diseases and disorders, particularly those that affect uniquely human traits, such as speech disorders, autism or schizophrenia. However, while the cost and time required to create comparative data sets have been greatly reduced, the error profiles and limitations of the new platforms differ significantly from those of previous approaches. This requires a specific experimental design in order to circumvent these issues, or to handle them during data analysis.

During the course of my PhD, I analyzed and improved current protocols and algorithms for next generation sequencing data, taking into account the specific characteristics of these new sequencing technologies. The presented approaches and algorithms were applied in different projects and are widely used within the department of Evolutionary Genetics at the Max Planck Institute of Evolutionary Anthropology. In this thesis, I will present selected analyses from the whole genome shotgun sequencing of two ancient hominins and the quantification of gene expression from short-sequence tags in five tissues from three primates.

IDT Exhibit 1061.0003

Summary

New high-throughput sequencing technologies make it possible to apply sequence-based approaches in an unanticipated number of fields. In the field of evolutionary genetics, it is now feasible to apply sequencing-based approaches for a wide range of comparative genomic studies. For example, high-throughput sequencing can be applied to study the genomes from ancient specimens of different hominin groups, like Neandertals and Denisovans, and allow large-scale population genetics studies of present-day humans as well as measuring quantitative differences in the transcriptomes and DNA-interactome of different apes and primates.

However, while the cost and time for applying these new technologies was greatly reduced compared to traditional Sanger sequencing, the error profiles and limitations of the new instruments differ significantly from those of previous technologies. Further, the types of errors observed as well as the number and length of sequences vary considerably between these different new technologies. Therefore, data analysis requires a detailed understanding of the imperfections in the resulting sequence data and how these pose challenges and cause biases. I review current sequencing technologies and point out their conceptual limitations. In this thesis I describe very specific biases and limitations, which go back to the technical details of how DNA molecules are prepared for sequencing, sequencing templates immobilized and finally read out.

Current high-throughput technologies have an average error rate of 1/25 to 1/1,000, which is considerably higher than the 1/10,000 to 1/100,000 observed for high quality Sanger sequence read outs. The *in vitro* amplifications which are generally performed prior to sequencing introduce a higher error into the sample before it enters the actual sequencing process. In addition, currently used random-dispersal protocols for immobilization of sequencing templates using beads or other solid surfaces cause mixed signal read outs and dependence of sequencing errors from strength and distance of close-by sequencing reactions.

Most errors on the new instruments originate from signal-to-noise thresholding and signal detection issues. Further, error rate substantially increases with the position in the sequence due to reductions in reaction efficiency, molecule damage and phasing, a process in which not all molecule copies are equally extended in every sequencing step. Shorter read lengths from these new platforms limit the accurate sequence mapping and assembly of genomes. Only paired end or mate pair protocols help to overcome some of these limitations by providing information about relative location and orientation of a pair of reads.

I have analyzed the currently most frequently used high-throughput sequencing platform, the Illumina Genome Analyzer, in more detail. Based on the problems observed frequently in runs performed at the Max Planck Institute for Evolutionary Anthropology, I present simple rules, which shall enable the identification and handling of the most common problems. I describe the different sources of high variance in run quality, ranging from issues with the sequencing libraries, to incorrect instrument adjustment and handling. Particles like chemistry lumps, dust and lint can cause pseudo sequence signals which result in the analysis of low sequence complexity reads not originating from the actual sequencing library. While sequence entropy filters efficientily remove these sequence, tagging or indexing allow a superior method for filtering real library molecules and further reduce the risk of library contamination.

For sequence reads where part of the adapter sequence is included, the position in the sequence read at which the adapter sequence begins has to be identified and the read trimmed appropriately. Unfortunately, this is not part of the standard Illumina data processing and also non-trivial for short adapter fragments, especially given the increasing sequencing error

IDT Exhibit 1061.0004

at the end of reads. If reads are not filtered for known chimeras and trimmed for adapter sequences, these may interfere with mapping/alignment and thereby impact downstream analysis. For paired end reads the correct identification of the adapter set-in is eased by maximizing autocorrelation of the two reads with the outlined read merging process. In addition to the efficient identification of adapters, merging reduces error rates in the consensus called sequence part. The algorithm presented has therefore been vital for different ancient DNA studies at the Max Planck Institute. Various library preparation biases may exist and impact sequencing results. For this reason, for example PCR duplicates need to be identified and specifically handled in analysis.

Considering that differences in error profiles are one of the major differences between technologies, reduction of these errors and precise estimates for the correctness of a specific base in a sequence are very important for any type of analysis. I present a new approach to base calling, the conversion of intensity measures into bases, for Illumina sequencing instruments. The approach presented is unique and currently applies to the full range of different Illumina sequencing chemistries and platform versions, for which it reduces sequencing error by at least 10-20%.

On the Illumina platform a strong correlation of adenine and cytosine intensities and of guanine and thymine intensities as well as a dependence of the signal for a specific cycle on the signal of the cycles before and after (phasing and pre-phasing) complicate base calling. Previous approaches have either completely modeled the sequencing process or at least corrected raw intensities prior to the application of statistical learners. Therefore, all these approaches depend on a good understanding and modeling of the sequencing process. The developed base calling package, `Ibis` (`Improved base identification system`), by-passes this problem by direct training of one statistical model per sequencing cycle based on raw cluster intensities of multiple input cycles, directly incorporating the effects of phasing. Thus, `Ibis` implements the most general and flexible approach, which is of advantage when considering the vast improvements of sequencing chemistry and instrument over the last years. Further, the performance of `Ibis` on standard hardware is significantly better than for other existing alternative base callers. Increases in mappable sequences due to reduced base identification errors as well as improved and calibrated `PHRED`-like quality scores enable the direct use of the sequences in other software packages.

I present two applications of `Ibis` and other principals presented in this thesis. I analyzed one of the first applications of the Illumina sequencing platform, the *NlaIII* Digital Gene Expression (DGE) approach, which infers gene expression levels through short 17nt-tag sequencing of the 3' ends of transcripts. This protocol was used to study brain, heart, kidney, liver and testis tissues of humans, chimpanzees and rhesus macaques. The biggest analysis challenge were the short tags which are not unique to specific genomic sites or genes and for which the uniqueness of tags differs slightly between the three species. Further, annotation of tags was problematic due to very different annotation quality for the three species. Only very recent human gene annotation provided the necessary annotation of 3' untranslated regions and could be projected to the chimpanzee and rhesus macaque genomes, losing about 36% of genes annotated in human but giving similar proportions of tag counts within genes for all three species.

From comparisons to other studies of the same species and tissues, larger disagreement was observed than was expected. For example, differences in the percentage differentially expressed genes or in the symmetry of assignment of changes to evolutionary lineages were observed. It is likely that all methods have technological (experimental and analysis) biases. A comparison with the Babbitt et al. study, also using the *NlaIII* DGE protocol but for different brain samples, clearly shows that sampling variation is at least in the range of biological

IDT Exhibit 1061.0005

differences between human and chimpanzee and that analysis variation may even be as strong as differences of humans and chimpanzees when compared to rhesus macaques. Future studies will need to control sample environmental effects, sample age, and tissue sampling more thoroughly. Further, improved experimental and analysis protocols are required which allow to detect and measure subtle effects that could introduce a species bias. Currently, species specific differences may easily originate from different genome quality, genome completeness and genome annotation quality.

The second analysis presents whole genome shotgun sequencing data that was generated for two hominin genomes from ancient DNA, the Neandertal and Denisova genome. Ancient DNA sequences are generally short in length, damaged, and at low copy-number relative to co-extracted environmental DNA. For Neandertal and Denisova the challenges from sequencing ancient DNA, which include adapter sequence at the read ends, chimerical sequences and other artifacts as well as sequencing error for short molecule lengths, have been addressed using the described approaches of improved base calling, tag filtering, and short paired end read merging. In combination with experimental approaches for reducing ancient DNA damage and the consensus from PCR duplicates, the sequencing error associated with ancient DNA studies could be considerably reduced. The remaining error from sequencing and error originating from ancient DNA damage in the Denisova molecule read outs is even lower than for present-day human sequences generated with the same technology.

I show how the ancient DNA sequences can be used to study sites in the human genome which have changed since the last common ancestor of human and chimpanzee and to identify features that set fully anatomically modern humans apart from other hominin forms. The identified positions point to several regions and genes, some of which might be affected by positive selection in the recent evolutionary history of modern humans. Experimental work will be required to elucidate the physiological consequences of the identified changes. In addition, I describe an interesting subset of sites which changed on the human lineage. I identified tens of thousand of positions where Denisovans and Neandertals disagree in the ancestral state at sites where the human reference sequence carries the derived allele. These positions are inconsistent between lines that separated more than half a million years ago and at least partially, reflect variation at the point of Human-Neandertal-Denisova lineage separation that segregated differently in the three lineages (incomplete lineage sorting).

It is likely that a large proportion of these sites, which were polymorphic at the time when human, Neandertal and Denisovan lineage separated, fixed for the derived allele in present-day humans. Thus, these differently segregating sites might have been reintroduced into some present-day human populations by admixture with either Neandertals or Denisovans and can be used to test present-day human individuals whether they show more frequently the ancestral allele for the Denisova ancestral sites or the ancestral allele for Neandertal ancestral sites. When analyzing these Neandertal-Denisova discordant sites in twelve present-day populations, they turned out to be informative for detecting admixture with either of the ancient population. I could confirm that an African individual shares fewer ancestral alleles with Neandertal than do all non-African individuals, supporting the admixture signal with non-Africans described in Green et al. for the Neandertal genome. Further, I could show that Melanesians, especially the two Papuan individuals, show a signal of Denisovan admixture not shared with other sampled populations, a result in agreement with the D-statistics for population pairs presented in Reich et al. for the Denisova genome.

During analysis, I point out differences in sampling of the reads obtained for the Neandertal and the Denisovan genomes. For example, in human accelerated regions Neandertal and Denisova data both show that these regions tend to predate the human-Denisova-Neandertal split and that differences caused by biased gene conversion tend to be older in time, however

IDT Exhibit 1061.0006

we sampled much more reads covering ancestral sites just in the Denisova data. While this may point to a simple sampling effect, an excess in the number of Denisova ancestral sites was observed in the concordance analysis, which can not result from sampling and also does not originate from a human reference sequence bias (from the Neandertal admixture present in parts of the reference genome). Currently this excess might either orignate from different alignment approaches or admixture into the Denisovan individual from some archaic hominin.

Both analyses pointed out that small effects throughout the whole data generation and data analysis process may introduce sufficiently large biases to complicate drawing biological conclusions from experimental data. The information and approaches outlined in this thesis, will however help to either prevent generating such biased data sets or at least reduce the sequencing instrumentation biases.

IDT Exhibit 1061.0007

Zusammenfassung

Die Verfügbarkeit neuer Sequenzierinstrumente ermöglicht die Anwendung sequenzbasierter Methoden für eine Vielzahl biologischer Gebiete. Auf dem Gebiet der evolutionären Genetik ist es heutzutage möglich sequenzbasierte Methoden für vergleichende Genomstudien anzuwenden. Zum Beispiel erlauben diese Hochdurchsatzsequenziermethoden die Analyse der Genome archäologischer Funde der Gattung *Homo*, wie beispielsweise von Neandertalern oder Denisova-Menschen. Des Weiteren können große Populationsstudien heutiger Menschen durchgeführt werden sowie quantitative Unterschiede in den Transkriptomen und DNS-Interaktomen verschiedener Menschenaffen und Affen untersucht werden.

Während die Kosten und der zeitliche Aufwand für die Generierung von Sequenzierdaten im Vergleich zur klassischen Sanger-Sequenzierung erheblich gesunken sind, unterscheiden sich jedoch die Fehlerprofile und Limitierungen der neuen Instrumente erheblich. Die Arten beobachteter Fehler, die Anzahl bestimmbarer Sequenzen pro Instrumentenlauf sowie die Länge der Sequenzen sind sehr variabel zwischen den verschiedenen Verfahren. Die Analyse von Sequenzierdaten setzt daher ein genaues Verständnis der Unvollkommenheiten jedes Instruments voraus und wie diese die Analyse verkomplizieren oder sogar systematische Fehler verursachen können. Ich habe daher Wissen zu allen aktuell verfügbaren Sequenzierinstrumenten zusammengetragen und die konzeptionellen Schwachpunkte herausgearbeitet. In dieser Arbeit beschreibe ich spezifische systematische Fehler und Einschränkungen der verschiedenen Technologien, welche auf technische Details in der Vorbereitung von DNS für die Sequenzierung, der Art und Weise wie Moleküle für das Auslesen fixiert werden und der Methode mit welcher Sequenzen ausgelesen werden, zurückzuführen sind.

Momentan verfügbare Hochdurchsatzsequenziermethoden besitzen mittlere Lesefehlerraten von 1/25 bis 1/1'000. Diese sind beachtlich höher als die Raten von 1/10'000 bis 1/100'000 wie sie für hochqualitative Sanger-Sequenzen erreicht werden. Die *in vitro* Vervielfältigungen von DNS-Molekülen, welche normalerweise vor der Sequenzierung mit den neuen Instrumenten durchgeführt werden, verursachen eine höhere Anzahl an Sequenzfehlern noch vor dem eigentlichen Auslesen. Des Weiteren ist eine Abhängigkeit der Lesefehlerrate vom Abstand der ausgelesenen Sequenzen zu beobachten. Dieser Abstand ist durch die Verwendung von experimentellen Protokollen zur zufälligen Verteilung der Sequenzen auf festen Oberflächen (planar oder in Form von Kugeln) sehr variabel. Die meisten Lesefehler dieser Instrumente werden jedoch durch Probleme bei der Signaldetektion und der Unterscheidung von Signal und Hintergrund verursacht. Zusätzlich steigt der Lesefehler beachtlich mit der Position in der auszulesenden Sequenz. Dies wird durch einen Abfall der Reaktionseffizienz, Beschädigung von Molekülen während der Sequenzierung und dem Verlust der Synchronität ausgelesener Molekülkopien verursacht. Der Verlust von Signalsynchronität beruht darauf, dass aufgrund unvollständiger Effizienz nicht alle Molekülkopien an jedem Reaktionsschritt teilnehmen.

Die überwiegend kurzen Sequenzen, welche mit den neuen Instrumenten bestimmt werden, begrenzen die Genauigkeit mit der diese bekannten Sequenzen zugeordnet werden können oder genomische Regionen rekonstruiert werden können. Nur Protokolle welche beide Enden eines Moleküls bestimmen, entweder in dem die tatsächlichen Enden von Molekülen gelesen oder vorher experimentell weiter auseinander liegende Molekülenden rekombiniert werden, liefern die notwendige Information über relativen Abstand und Richtung, welche die fehlende Information kurzer Sequenzen ausgleichen kann.

Ich habe das derzeit am häufigsten verwendete Sequenzierinstrument, den Illumina Genome Analyzer, genauer untersucht und basierend auf den Problemen, welche bei Sequenzierläufen am Max-Planck-Institut für evolutionäre Anthropologie häufig beobachtet wurden, einfache

IDT Exhibit 1061.0008

Regeln aufgestellt die es gestatten sollen diese Fehler zu vermeiden oder in den Sequenzierdaten zu korrigieren. In der vorliegenden Arbeit beschreibe ich verschiedene Quellen für die hohe Varianz in der Qualität von Sequenzierläufen. Diese umfassen Probleme mit Sequenzierbibliotheken, mangelnde Instrumentjustierung und Fehler in der Handhabung. Partikel, wie beispielsweise Klümpchen, Staub und Fusseln in der Sequenzierchemie können Sequenzähnliche Signale hervorrufen und damit zur Analyse von niedrigkomplexen Sequenzen führen, welche nicht aus der Bibliothek stammen. Diese Sequenzen können durch Entropie-basierende Filter entfernt werden, das Filtern mittels Bibliothek-spezifischer Erkennungssequenzen ist jedoch zu bevorzugen, da hier keine falsch-positiven Sequenzen entfernt werden und zusätzlich die Gefahr von Kontamination mit anderen Sequenzierbibliotheken reduziert wird.

Für Sequenzen bei denen ein Teil des Sequenzierbibliothekadapters mitgelesen wird, muss dieser am Ende der eigentlichen Zielsequenz identifiziert werden und vor der weiteren Prozessierung entfernt werden. Dies ist leider nicht Teil der Standardprozessierung des Illumina-Instrumentes. Durch die am Sequenzende erhöhte Lesefehlerrate und häufig kurze Adapterfragmente, ist auch die Erkennung des Adapteres ein nicht-triviales Problem. Wenn Sequenzen nicht für Adapterchimären gefiltert werden oder einsetzende Adaptersequenz nicht entfernt wird, können Ergebnisse negativ beeinflusst werden. Bei der Sequenzierung von beiden Molekülenden kann eine einsetzende Adaptersequenz durch ein beschriebenes Verfahren zur Maximierung der Autokorrelation einfacher identifiziert werden. Die Berechnung einer Konsensussequenz in der Überlappung beider Endsequenzen reduziert zusätzlich die Lesefehlerrate. Daher war dieses Verfahren insbesondere für die Analyse von DNS-Molekülen aus archäologischen Proben von großer Bedeutung. Verschiedene Präparationsprotokolle für Sequenzierbibliotheken existieren und haben erheblichen Einfluss auf die Sequenzierung. Um beispielsweise systematische Fehler durch die ungleichmäßige Vervielfältigung von DNS-Molekülen zu vermeiden, müssen solche Kopien identifiziert und speziell gehandhabt werden.

Berücksichtigt man dass verschiedene Fehlerprofile eine der Hauptunterschiede zwischen den verschiedenen Technologien sind, so ist die Minderung von Lesefehlern sowie die genaue Abschätzung der Fehlerwahrscheinlichkeit gelesener Basen sehr wichtig für jegliche Art der Analyse. In dieser Arbeit stelle ich für das Illumina-Instrument einen Ansatz für die Basenbestimmung (die Konvertierung von Maschinenintensitätswerten zu Basen) vor. Die vorgestellte Methode ist einzigartig unter den bisher verwendeten Ansätzen und kann auf alle bisher verfügbaren Illumina Instrument- und Sequenzierchemieversionen angewendet werden und reduziert den Lesefehler um mindestens 10-20%.

Für diese Instrumente wird eine starke Korrelation der Adenin- und Cytosin-Intensitäten sowie der Guanin- und Thymin-Intensitäten beobachtet. Des Weiteren existiert eine Abhängigkeit der Intensitäten eines bestimmen Instrumentenzyklus von den Intensitäten der vorhergehenden und nachfolgenden Zyklen. Beide Effekte erschweren die Basenbestimmung. Bisher verwendete Methoden haben daher entweder den vollständigen Sequenzierprozess modelliert oder die Maschinenintensitäten für diese Effekte teilweise korrigiert bevor ein statistisches Verfahren zur Basenbestimmung verwendet wurde. Das von mir entwickelte Programm `Ibis` (`Improved base identification system`) umgeht das Problem der Modellierung durch das Trainieren eines statistischen Klassifikators pro Instrumentenzyklus unter Verwendung der Intensitätswerte mehrerer Zyklen. Daher implementiert `Ibis` einen sehr allgemeinen Ansatz, was im Bezug auf die Häufigkeit von Aktualisierungen der Instrument- und Sequenzierchemieversionen über die letzten Jahre von großem Vorteil ist. Zusätzlich ist `Ibis` im Vergleich zu anderen Programmen sehr effizient und besitzt keine besonderen Anforderungen an die eingesetzte Rechentechnik. Die geringere Lesefehlerrate und damit ein höherer Anteil nutzbarer Sequenzen sowie die besseren und kalibrierten Basenfehlerwahrscheinlichkeiten ermöglichen die direkte Verwendung der `Ibis`-Daten mit anderen Programmen.

IDT Exhibit 1061.0009

Ich stelle zwei Anwendungen von `Ibis` sowie der dargelegten Analyseprinzipien vor. Als erstes analysiere ich eine der ersten Anwendungen für die Illumina-Sequenzierinstrumente, das *NlaIII* Protokoll für digitale Genexpression (DGE). Dieses Protokoll bestimmt die Expression von Genen mittels kurzer 17-Nukleotid-langer Erkennungssequenzen der 3'-Enden von Transkripten. Dieses Protokoll wurde verwendet um die Genexpression in Hirn-, Herz-, Hoden-, Leber- und Nierengewebe von Menschen, Schimpansen und Rhesusaffen zu messen. Das größte Problem bei der Analyse dieser Daten waren die sehr kurzen Erkennungssequenzen, welche nicht einzigartig für bestimmte genomische Regionen oder Gene sind und für welche die Einzigartigkeit auch zwischen den untersuchten Arten leicht variiert. Zusätzlich stellte sich die Annotation der Erkennungssequenzen als schwierig heraus, da die verfügbare Genannotation für die drei Arten von sehr verschiedener Qualität ist. Nur die neusten Versionen der menschlichen Genannotation beschreiben 3'-untranslatierte Bereiche von Genen in ausreichend guter Qualität um sie für die Analyse dieser Daten verwenden zu können. Ich habe daher die menschliche Genannotation unter Verlust von 36% aller Gene auf die Genome der beiden anderen Arten projiziert. Nur so konnte ich sicherstellen, dass gleiche Anteile der bestimmten Erkennungssequenzen in allen drei Arten analysiert werden.

Im Vergleich mit anderen Studien derselben Gewebe und Arten stellte ich größere Abweichungen fest. Zum Beispiel zeigten sich Unterschiede im Anteil der unterschiedlich expremierter Gene in bestimmten Geweben sowie im Anteil der Genexpressionsveränderungen welche einer bestimmten evolutionären Linie zugeordnet werden konnten. Es ist daher wahrscheinlich, dass die verschiedenen experimentellen oder analytischen Methoden systematische Fehler aufweisen. Der Vergleich mit der Studie von Babbit et al., welche ebenfalls das *NlaIII*-DGE-Protokoll in verschiedenen Hirnproben angewendet hat, zeigte sehr klar, dass die Varianz, welche durch das Beproben von verschiedenen Individuen eingebracht wird, mindestens genauso groß ist wie die Varianz welche durch biologische Unterschiede zwischen Schimpanse und Mensch beobachtet wird. Des Weiteren zeigte sich, dass die Varianz aus der Analyse der Daten mindestens genauso groß sein kann wie die biologischen Unterschiede zwischen Mensch oder Schimpanse im Vergleich zu Rhesusaffen. Zukünftige Studien müssen daher verstärkt für verschiedene Umwelteinflüsse, Alter der beprobten Individuen und Art der Gewebebeprobung kontrollieren. Verbesserte experimentelle Protokolle sowie bessere Analysemethoden werden benötigt, welche es erlauben systematische Fehler früh zu erkennen und zu messen. Derzeit können Unterschiede in Genomqualität, Vollständigkeit der verfügbaren Genome und Qualität der verfügbaren Genannotation leicht einen Spezieseffekt verursachen.

Die zweite Analyse stellt Schrotschusssequenzierungsdaten aus archäologischen Proben zweier Hominini vor; das Genom des Neandertalers und des Denisova-Menschen. DNS wie sie aus archäologischen Funden extrahiert werden kann, ist häufig stark fragmentiert, chemisch beschädigt und besitzt einen hohen Anteil von DNS aus verschiedenen Umweltquellen. Kurze DNS-Moleküle, welche den Bibliotheksadapter am Ende der ausgelesen Sequenz aufweisen, Chimären, Artefaktsequenzen, sowie hohe Fehlerraten in Kombination mit den kurzen Sequenzen, stellen ein Problem bei der Analyse solcher Daten dar. Das verbesserte Verfahren zur Basenerkennung, Filtern für Bibliothekserkennungssequenzen und der Einsatz des Verfahrens zur Rekonstruktion des vollständigen Moleküls von Sequenzen beider Molekülenden konnte viele Probleme in diesen Daten überwinden. Kombiniert mit experimentellen Methoden zur Reduktion des chemischen Schadens und der Berechnung von Konsensussequenzen von Molekülduplikaten konnte der verbleibende Sequenzfehler soweit reduziert werden, dass er für das Genom des Denisova-Menschen kleiner ist als für Sequenzierdaten heutiger Menschen welche mit der gleichen Sequenziertechnik bestimmt wurden.

Die Sequenzierdaten des Neandertaler- und Denisova-Genomes können verwendet werden um Positionen im menschlichen Genom zu identifizieren, welche sich seit dem letzten gemeinsa-

IDT Exhibit 1061.0010

men Vorfahren von Mensch und Schimpanse geändert haben. Diese Positionen sind wichtig um Gene zu identifizieren, welche anatomisch-moderne Menschen von anderen Hominini unterscheiden. Die bisher mittels Neandertal- und Denisova-Genom identifizierten Positionen heben mehrere Gene und genomische Regionen heraus, welche in der jüngeren menschlichen Evolution von Bedeutung gewesen sein könnten. Experimentelle Tests werden nun zeigen müssen, ob die identifizierten Unterschiede auch funktionelle Bedeutung haben.

Bei der Analyse dieser Positionen beschreibe ich eine interessante Untermenge von mehreren zehntausend Positionen, welche im Neandertaler und Denisova-Genom verschiedene Zustände haben. Hier zeigt entweder das Neandertaler- oder das Densiova-Genom den in der menschlichen Referenz vorhandenen neuen Zustand, während das jeweils andere Genom den ursprünglichen Zustand des gemeinsamen Vorfahren mit Schimpansen zeigt. Diese Positionen sind damit inkonsistent zwischen evolutionären Linien welche sich vor mehr als einer halben Million Jahre getrennt haben. Sie repräsentieren damit zumindest teilweise Positionen unvollständiger Linienauftrennung welche im gemeinsamen Vorfahren von Mensch, Neandertaler und Denisova-Mensch noch variabel waren. Es ist wahrscheinlich dass ein großer Teil dieser Positionen in heute lebenden Menschen für die neue Sequenzvariante fixiert ist. Der ursprüngliche Zustand für diese unterschiedlich segregierenden Positionen könnte aber durch Vermischung mit Neandertalern und Denisova-Menschen wieder in bestimmte Menschenpopulationen eingebracht worden sein. Daher können diese Positionen, genauer der Anteil mit der ein heutiger Mensch den ursprünglichen Zustand für solche Positionen mit Neandertaler oder Denisova-Mensch teilt, genutzt werden um eine stärkere Vermischung mit einer der beiden ausgestorbenen Menschenformen zu detektieren. Durch die Analyse konnte ich die Ergebnisse von Green et al. bestätigen, dass ein untersuchter Afrikaner weniger dieser Postionen mit Neandertalern teilt als die neun getesteten Nichtafrikaner. Des Weiteren konnte ich zeigen dass Melanesier, insbesondere zwei Individuen aus Papua-Neuguinea, ein Signal für die Vermischung mit Denisova-Menschen zeigen, welches nicht in den anderen Individuen gefunden werden konnte. Dieses Ergebnis ist in Übereinstimmung mit den Ergebnissen der D-Statistik für Populationspaare wie sie in Reich et al. vorgestellt wurde.

In der Analyse weise ich auch darauf hin, dass Sequenzen des Neandertal- und Denisova-Genoms unterschiedlich beprobt werden. Beispielsweise zeigt sich bei der Analyse von Regionen die sich auf der menschlichen Linie besonders stark verändert haben, dass diese zwar in beiden Genomen häufiger den modernen Zustand zeigen als dies im Genommittel der Fall ist – ein Effekt der wahrscheinlich von Genkonversionsereignissen getrieben wird –, aber in den Denisova-Daten werden mehr Positionen mit dem ursprünglichen Zustand durch Sequenzen abgedeckt, als dies für Positionen der Fall ist, die nur in den Neandertaler-Daten abgedeckt werden. Während dies auf einen Unterschied in den erhaltenen DNS-Molekülen hinweisen könnte, kann dies nicht den Unterschied abweichender Positionen in Neandertalern oder Denisova-Menschen erklären. Auch nach Korrektur eines Effektes der humanen Referenzsequenz, welche Neandertal-Vermischung in sich trägt, verbleibt ein Überschuss von 18.4% für Positionen die den ursprünglichen Zustand in Denisova-Menschen aber den neuen Zustand im Neandertaler-Genom zeigen. Dies könnte auf Vermischung des Denisova-Menschen mit einer noch älteren Menschenform oder aber ein Alignierungsartefakt durch die verschiedenen Algorithmen hinweisen.

Beide präsentierten Analysen haben kleine Effekte in der Datengenerierung und Auswertung aufgezeigt, welche einen hinreichend großen systematischen Fehler verursachen, so dass nicht ohne weiteres biologische Schlussfolgerungen aus den experimentellen Daten gezogen werden können. Die Informationen und Methoden welche in dieser Arbeit beschrieben werden, können jedoch dabei helfen diese Effekte in zukünftigen Daten zu vermeiden oder sie in bestehenden Daten zu reduzieren.

IDT Exhibit 1061.0011

Acknowledgments

I thank my supervisors Janet Kelso, Svante Pääbo and Peter Stadler for their continuous support and helpful discussions as well as for providing funding necessary for this work. Further, I thank all current and previous members of the Evolutionary Genetics department at the Max Planck Institute for Evolutionary Anthropology for their scientific advice and support.

Additionally, I would like to say thank you to all these that mostly remain unmentioned in sciences: family and friends that indirectly contributed to this work by, for example, reading and reviewing manuscripts, discussing different topics related to my work or putting their trust in me.

IDT Exhibit 1061.0012

Note

As expressed in the German text sections below, I have produced this thesis independently and without using any other than the aids listed. In part, text, tables and figures from this thesis have been used in peer-reviewed publications of which I am one of the main authors [116, 115, 81, 186]. To date, the thesis has not been submitted to any other board of examiners in the same or a similar format. The curriculum vitae as well as lists of scientific publications and talks are available in the appendix, pages 215 and 217 respectively.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die aufgeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien und erbrachten Dienstleistungen als solche gekennzeichnet.

Erklärungen des Antragstellers

Hiermit erkenne ich die Promotionsordnung der Fakultät für Mathematik und Informatik der Universität Leipzig vom 22. Juli 2009 an. Die eingereichte Arbeit wurde in gleicher oder ähnlicher Form nicht einer anderen Prüfungsbehörde zum Zwecke einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt. Abbildungen, Tabellen und Texte dieser Arbeit wurden in Teilen bereits in Publikationen präsentiert von welchen ich einer der Hauptautoren bin (siehe Referenzen [116, 115, 81, 186]).

Ein tabellarischer Lebenslauf mit Darstellung des wissenschaftlichen Werdeganges und Lichtbild sowie das Verzeichnis der wissenschaftlichen Veröffentlichungen und Vorträge findet sich in Anhang auf Seite 215 sowie auf Seite 217.

Leipzig, den 21. Februar 2011

Martin Kircher

IDT Exhibit 1061.0013

IDT Exhibit 1061.0014

Contents

1 Introduction **19**

2 DNA sequencing technologies **21**

2.1 Sanger capillary sequencing 22
2.2 454/Roche Genome Sequencer 23
2.3 Illumina Genome Analyzer 26
2.4 Life Technologies SOLiD 29
2.5 Helicos HeliScope 31
2.6 Single Molecule Real Time (SMRT) 33
2.7 Upcoming developments 33
2.8 General considerations 36
2.9 Summary and conclusions 39

3 Computational challenges from sequencing data production **41**

3.1 Sequencing libraries, minimum insert size and adapter artifacts 44
3.2 Short-insert libraries and paired-end sequencing 48
3.3 Separation of samples from multiplex experiments 53
3.4 Sample contamination 55
3.5 Machine adjustment and run preparation 56
3.6 Image analysis 58
3.7 Low quality sequences and sequencing artifacts 59
3.8 Sequence composition and standard base calling 63
3.9 Control read spike-in and alternative base callers 63
3.10 PCR duplicates in data analysis 64
3.11 Summary and conclusions 66

4 Improving base call quality of the Genome Analyzer platform **69**

4.1 Improved base calling and quality scoring 70
4.2 Input for base calling (intensity files) 70
4.3 Illumina standard base caller 71
4.4 Direct application of a statistical learner 74
4.4.1 Simulating phasing, pre-phasing and T accumulation 75
4.4.2 Using the signal of neighboring bases 79
4.5 Comparison to other systems for base calling 82

IDT Exhibit 1061.0015

4.5.1 Performance test on data sets using v1 chemistry 83
4.5.2 Performance on a test data set from v2 chemistry 85
4.5.3 Performance test on current data sets from Genome Analyzer IIx . . . 88
4.5.4 Required computational resources 92
4.5.5 Dependence on training input data 93
4.5.6 Training without a dedicated control lane 94
4.6 Downstream quality score recalibration 95
4.7 Summary and conclusions 96

5 Quantification of gene expression from short-sequence tags 97
5.1 Samples and experimental protocol 98
5.2 Sequencing and primary data processing 99
5.3 Tag alignment 100
5.4 Annotation of expressed sequence tags 103
5.5 Gene quantification and count normalization 109
5.6 Differentially expressed genes 113
5.7 Comparison to other data sets 117
5.7.1 Comparison of brain samples with Babbitt et al. 118
5.7.2 Comparison of liver samples with Blekhman et al. 123
5.7.3 Comparison of all five tissues with Khaitovich et al. 126
5.8 Summary and conclusions 128

6 Analysis of two hominin genomes from ancient DNA 131
6.1 Neandertals and Denisovans 132
6.2 Illumina Sequencing and primary data processing 134
6.2.1 Neandertal sequence data 134
6.2.2 Denisova sequence data 136
6.2.3 Identification of endogenous molecules 136
6.2.4 Present-day human low-coverage data 138
6.3 Identification of changes on the human lineage 141
6.3.1 Identification of positions with Neandertal sequence coverage 142
6.3.2 Identification of positions with Denisova sequence coverage 142
6.3.3 Annotation of genomic features 144
6.4 Changes in protein-coding sequences analyzed from Neandertal data 144
6.4.1 Amino acid substitutions 145
6.4.2 Stop/Start codon substitutions 146
6.4.3 Indels in coding sequence 146
6.5 Changes in protein-coding sequences analyzed from Denisova data 147
6.5.1 Amino acid substitutions 147
6.5.2 Stop/Start codon substitutions 148
6.5.3 Insertions and deletions in coding sequence 148
6.6 Changes in non-protein-coding sequences 149

IDT Exhibit 1061.0016

6.6.1 5' UTR substitutions and insertion/deletions 149
6.6.2 3' UTR substitutions and insertion/deletions 150
6.6.3 microRNAs 151
6.6.4 Human Accelerated Regions 154
6.7 Neandertal-Denisova concordance 156
6.8 Allele sharing of humans at sites of Neandertal-Denisova discordance 157
6.8.1 Overrepresentation of Denisova ancestral alleles 158
6.8.2 Testing twelve present-day human individuals 161
6.8.3 Generating an African catalog 163
6.9 Summary and conclusions 166

7 Discussion and conclusions **169**
DNA sequencing and technologies 169
Computational challenges from sequencing data production 171
Improving base call quality of the Genome Analyzer platform 172
Quantification of gene expression from short-sequence tags 173
Analysis of two hominin genomes from ancient DNA 175
Outlook 176

Bibliography **178**

Appendix **197**
Tables 197
List of Figures 208
List of Tables 212
Curriculum vitae 215
List of publications and talks 217
Abbreviations 221

Index **223**

IDT Exhibit 1061.0017

IDT Exhibit 1061.0018

Chapter 1

Introduction

Over the past five years, advances in DNA sequencing have revolutionized the field of genomics. It is now possible to generate large amounts of sequence data very rapidly and at low cost. Using current high-throughput sequencing instruments, the amount of sequencing data produced by the human genome project [100] in over 13 years can be created within weeks and for a price of a few tens of thousands of dollars instead of millions of dollars.

These high-throughput sequencing technologies allow the application of sequence-based approaches in an unanticipated number of fields. Reducing sequencing costs means DNA sequencing is now available to many more researchers and projects. Common sequencing applications are wide and varied. They include whole genome sequencing, measuring population and species variation, determining transcriptome structure, quantifying gene expression, determining DNA methylation and unveiling DNA-protein interactions.

In the field of evolutionary genetics, it is now feasible to apply these different sequencing-based approaches for comparative genomic studies. In primates, and more specifically in extinct and extant ape populations. The high resolution of sequencing approaches allow studies of species and population differences that have not been accessible so far due to the small evolutionary differences studied. Further, these high-throughput sequencing technologies allow studies of ancient DNA samples with a low fraction of endogeneous DNA, studies which were too expensive when using traditional sequencing approaches. Comparative genomics studies of humans, hominins and other apes are important in order to better understand human origins and the biological background of what sets humans apart from other organisms. They can also provide insights into the basis of diseases or developmental problems that affect uniquely human traits, such as speech disorders, mental disorders such as autism and schizophrenia, or metabolic disorders such as obesity.

However, while the cost and time for applying these new technologies have been greatly reduced compared to traditional sequencing, the error profiles and limitations of the new platforms differ significantly from those of previous sequencing technologies. Further, the types of errors observed as well as the number and length of sequences vary considerably among these different new technologies. Thus, the selection of an appropriate sequencing platform for particular types of experiments is an important consideration. In addition, data analysis requires a detailed understanding of the imperfections in the resulting sequence data and how these pose challenges and cause biases in downstream analysis.

In the following chapters, I will first review the sequencing approaches implemented by different DNA sequencing instruments and discuss their inherent limitations. Then I will discuss how experimental steps in the generation of high-throughput sequencing data impact data quality and generate artifacts that may affect data analysis (chapter 3). In chapter 4,

IDT Exhibit 1061.0019

I will present an approach for improving the base calling for one of the high-throughput instruments, the Illumina Genome Analyzer, and show how it allows for a reduction of error rates and provides more informative base quality scores. In the last two chapters, I will discuss specific problems as well as selected results from the quantification of gene expression from short-sequence tags in five tissues from three primates (chapter 5) as well as the analysis of whole genome shotgun sequencing data of two ancient hominin genomes (chapter 6).

IDT Exhibit 1061.0020

Chapter 2

DNA sequencing technologies

Ambition is the last refuge of the failure. – Oscar Wilde [96](1205)

The first DNA genome, that of the 5,386-nucleotide, single-stranded bacteriophage ϕX174 was determined in 1977 [202] using one of the technologies of DNA sequencing invented at the time [203, 204, 76, 245]. Since then sequencing of whole genomes as well as of individual genomic regions and genes has become a major focus of modern biology and transformed, if not founded, the field of modern genetics.

However, in the 1970s and for almost another decade, DNA sequencing was a barely automated and therefore very tedious process which allowed determining only a few hundred nucleotides in an experiment. In the late 1980s, semi-automated sequencers with higher throughput became available [216, 223], still only able to determine a few sequences at a time. One breakthrough in the early 1990s was the development of capillary array electrophoresis and appropriate detection systems [249, 97, 105, 233, 114]. As recently as 1996, these developments converged in a commercial single capillary sequencer (ABI Prism 310). In 1998, the GE Healthcare MegaBACE 1000 and the ABI Prism 3700 DNA Analyzer became the first commercial 96 capillary sequencers, a development which then was termed high-throughput sequencing.

Only within the last five years, alternative sequencing strategies like pyrosequencing [196, 142], reversible terminator chemistry [231, 16], sequencing-by-ligation [210], virtual terminator chemistry [89] and real-time sequencing [118] were developed or converged into new instruments. These new instruments require us to completely redefine the term "high-throughput sequencing", as they outperform the older Sanger-sequencing technologies by a factor of 100 to 1,000 in daily throughput and reduce the cost of sequencing one million nucleotides (1Mb) to 4%-0.1% of that associated with Sanger sequencing. This large difference in throughput led scientists and companies to introduce new terms like "next generation sequencing" [149, 8] or "ultra-high-throughput sequencing" [71] for this group of new technologies. These terms unfortunately leave little scope for the ongoing developments, which is why I use the terms high-throughput sequencing and Sanger sequencing instead.

In the following, I will describe the concepts of currently available sequencing technologies. While high-throughput sequencing technologies have been widely reviewed [8, 149, 206, 209] and possible applications discussed [148, 172, 69, 169, 237, 44], this chapter has a more technical focus and outlines the inherent limitations that come with each of the technologies [115].

IDT Exhibit 1061.0021

2.1 Sanger capillary sequencing

Current Sanger capillary array electrophoresis (CAE) systems, like the widely used Applied Biosystems 3000 series or the GE Healthcare MegaBACE instrument, are based on the same general scheme applied in 1977 for the ϕX174 genome [202, 204]: First, millions of copies of the sequence to be determined are purified or amplified, depending on the source of the sequence. Reverse strand synthesis is performed on these copies using a known priming sequence upstream of the sequence to be determined and a mixture of deoxy-nucleotides (dNTPs, the standard building blocks of DNA) and dideoxy-nucleotides (ddNTP, modified nucleotides missing a hydroxyl group at the third carbon atom of the sugar). The dNTP/ddNTP mixture causes random, non-reversible termination of the extension reaction; creating from the different copies molecules extended to different lengths. Following denaturation and clean up of free nucleotides, primers and the enzyme, the resulting molecules are sorted by their molecular weight (corresponding to the point of termination) and the label attached to the terminating ddNTPs is read out sequentially in the order created by the sorting step. A schematic representation of this process is available in figure 2.1.

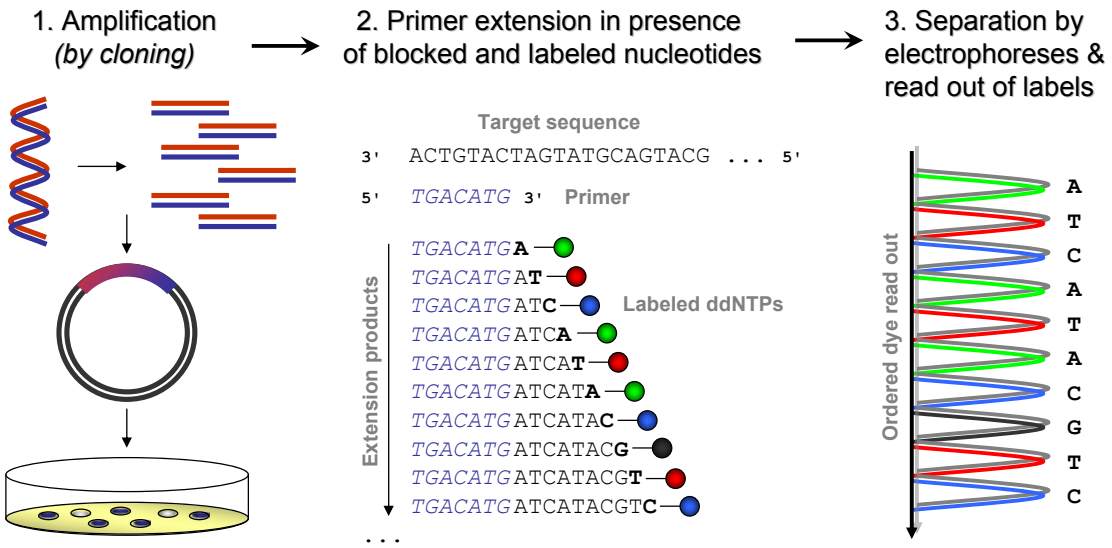

Figure 2.1: Schematic representation of the Sanger sequencing process. Input DNA is fragmented and cloned into bacterial vectors for *in vivo* amplification. Reverse strand synthesis is performed on the obtained copies starting from a known priming sequence and using a mixture of deoxy-nucleotides (dNTPs) and dideoxy-nucleotides (ddNTPs). The dNTP/ddNTP mixture randomly causes the extension to be non-reversibly terminated; creating differently extended molecules. Subsequently, after denaturation, clean up of free nucleotides, primers, and the enzyme, the resulting molecules are sorted using capillary electrophoresis by their molecular weight (corresponding to the point of termination) and the fluorescent label attached to the terminating ddNTPs is read out sequentially.

Sorting by molecular weight was originally performed using gel electrophoresis but is nowadays carried out by capillary electrophoresis [223, 74]. Originally, radioactive or optical labels were applied in four different terminator reactions (each sorted and read out separately), but today four different fluorophores, one per nucleotide (A, C, G and T) are used in a single reaction [216]. Additionally, the advent of more sensitive detection systems and several rounds of primer extensions (equivalent to a linear amplification) permit smaller amounts of starting DNA to be used for modern sequencing reactions.

IDT Exhibit 1061.0022

Unfortunately, there is still little automation for creation of the high copy input DNA with known priming sites. Typically this is done by cloning, i.e. introducing the target sequence into a known vector sequence using restriction and ligation procedures and using a bacterial strain to amplify the target sequence *in vivo* – thereby exploiting the low amplification error due to inherent proof-reading and repair mechanisms. However this process is very tedious and is sometimes hampered by difficulties with cloning specific sequences due to their base composition, length and interactions with the bacterial host system. Though not yet widely used, integrated microfluidic devices have been developed which aim to automate the DNA extraction, *in vitro* amplification and sequencing on the same chip [58, 22, 197, 144].

Using current Sanger sequencing technology, it is technically possible for up to 384 sequences [58, 213] of between 600 and 1,000 nucleotides (nt) in length [211, 92] to be sequenced in parallel. However, these 384-capillary systems are rare. The more standard 96-capillary instruments yield a maximum of approximately 6Mb of DNA sequence per day, with costs for consumables amounting to about $500 per 1Mb. The sequencing error observed for Sanger sequencing is mainly due to errors in the amplification step (a low rate when done *in vivo*), natural variance and contamination in the sample used as well as polymerase slippage at low complexity sequences like simple repeats (short variable number tandem repeats) and homopolymers (stretches of the same nucleotide). Further, lower intensities and missing termination variants tend to lead to sequencing errors accumulating towards the end of long sequences. In combination with reduced separation by the electrophoresis, base miscalls [64] and deletions increase with read length. However, the average error rate (the average over all bases of a sequence) after sequence end trimming is typically very low, with an error every 10,000-100,000 nucleotides [63].

2.2 454/Roche Genome Sequencer

The 454 Genome Sequencer (GS) platform was the first of the new high-throughput sequencing platforms on the market when released in October 2005. It is based on the pyrosequencing approach developed by Pål Nyrén and Mostafa Ronaghi at the Royal Institute of Technology Stockholm in 1996 [196]. In contrast to the Sanger technology, pyrosequencing is based on iteratively complementing single strands and simultaneously reading out the signal emitted from the nucleotide being incorporated (also called "sequencing by synthesis" or "sequencing during extension"). Electrophoresis is therefore no longer required to generate an ordered read out of the nucleotides, as the read out is done simultaneously with the sequence extension.

In the pyrosequencing process (figure 2.2 on the next page), one nucleotide at a time is washed over several copies of the sequence to be determined, causing polymerases to incorporate the nucleotide if it is complementary to the template strand. The incorporation stops if the longest possible stretch of complementary nucleotides has been synthesized by the polymerase. In the process of incorporation, one pyrophosphate per nucleotide is released and converted to adenosine triphosphate (ATP) by an ATP sulfurylase. The ATP drives the light reaction of luciferases present and the emitted light signal is measured. To prevent the deoxyadenosine triphosphate (dATP) provided in a typical sequencing reaction from being used directly in the light reaction, deoxy-adenosine-5'-(alpha-thio)-triphosphate (dATPαS), which is not a substrate of the luciferase, is used for the base incorporation reaction of adenine. Standard deoxyribose nucleotides are used for all other nucleotides. After capturing the light intensity, the remaining unincorporated nucleotides are washed away and the next nucleotide is provided.

IDT Exhibit 1061.0023

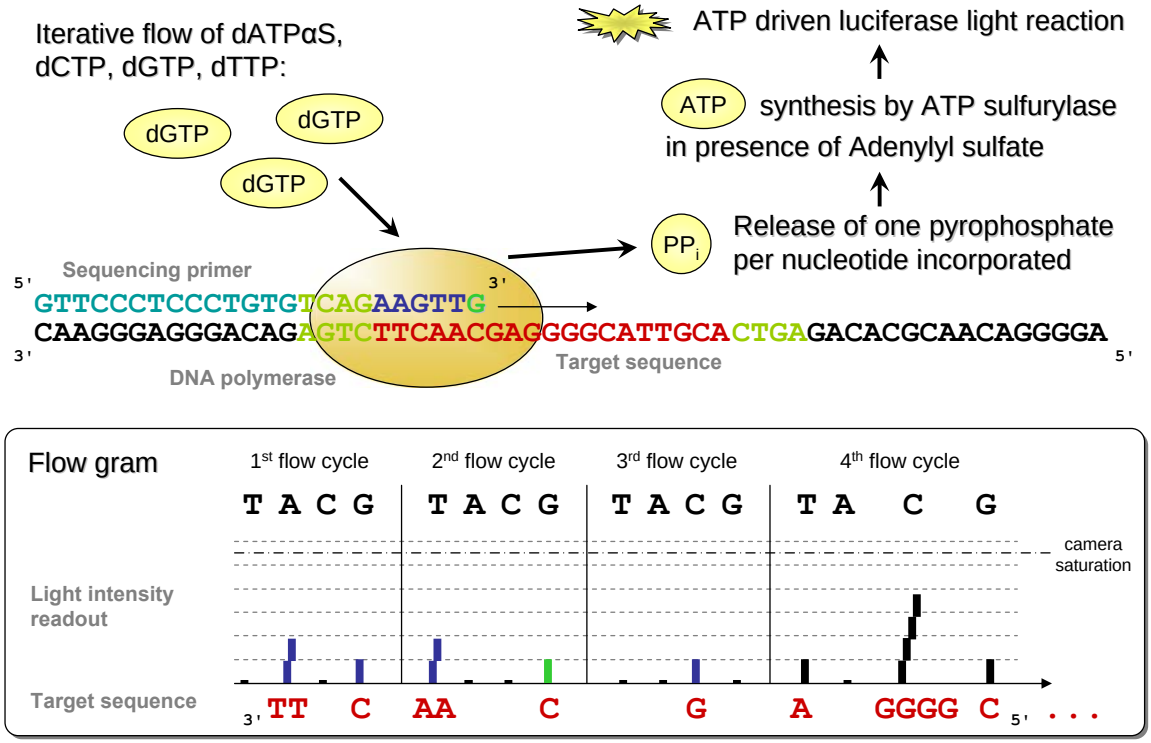

Figure 2.2: The pyrosequencing process. One of four nucleotides is washed sequentially over copies of the sequence to be determined, causing polymerases to incorporate complementary nucleotides. The incorporation stops if the longest possible stretch of the available nucleotide has been synthesized. In the process of incorporation, one pyrophosphate per nucleotide is released and converted to ATP by an ATP sulfurylase. The ATP drives the light reaction of luciferases present and a light signal proportional (within limits) to the number of nucleotide incorporations can be measured.

In 2005, pyrosequencing technology was parallelized on a picotiter plate[1] by 454 Life Sciences (later bought by Roche Diagnostics) to allow high-throughput sequencing [142]. The sequencing plate has about two million wells – each of them able to accommodate exactly one $28\mu m$ diameter bead covered with single stranded copies of the sequence to be determined. The beads are incubated with a polymerase and single-strand binding proteins and, together with smaller beads carrying the ATP sulfurylases and luciferases, gravitationally deposited in the wells. Free nucleotides are then washed over the sequencing plate and the light emitted during the incorporation is captured for all wells in parallel using a high resolution CCD camera, exploiting the light-transporting features of the plate used.

One of the main prerequisites for applying this array-based pyrosequencing approach is covering individual beads with multiple copies of the same molecule. This is done by first creating sequencing libraries in which every individual molecule gets two different adapter sequences, one at the 5' end and one at the 3' end of the molecule. In the case of the 454/Roche sequencing library preparation [142], this is done by sequential ligation of two pre-synthesized oligos. One of the adapters added is complementary to oligonucleotides on the sequencing beads and thus allows molecules to be bound to the beads by hybridization. Low molecule-to-bead ratios and amplification from the hybridized double-stranded sequence on the beads (kept separate using polymerase chain reaction in an water-in-oil emulsion, i.e emulsion PCR) makes it possible to grow beads with thousands of bound copies of a single

[1] a flat plate with millions of wells used as separate reaction chamber

IDT Exhibit 1061.0024

starting molecule. Using the second adapter, beads covered with molecules can be separated from empty beads (using capture beads with oligonucleotides complementary to the second adapter) and are then used in the sequencing reaction as described above.

The average substitution (excluding insertion/deletions) error rate is in the range of 10^{-3} to 10^{-4} [142, 184], which is higher than the rates observed for Sanger sequencing but is the lowest average substitution error rate of the new sequencing technologies discussed here. As mentioned earlier for Sanger sequencing, *in vitro* amplifications performed for the sequencing preparation cause a higher background error rate, that is, the error introduced into the sample before it enters the sequencing process. In addition, in bead preparation (i.e. emulsion PCR step) a fraction of the beads end up carrying copies of multiple different sequences. These "mixed beads" will participate in a high number of incorporations per flow cycle, resulting in sequencing reads that do not reflect real molecules. Most of these reads are automatically filtered during the software post-processing of the data. The filtering of mixed beads may however cause a depletion of real sequences with a high fraction of incorporations per flow cycle.

A large fraction of the errors observed for this instrument are small insertions or deletions (InDels), mostly arising from inaccurate calling of homopolymer length, and single base-pair deletions or insertions caused by signal-to-noise thresholding issues [184]. Most of these problems can be resolved by higher coverage. For long ($>$ 10nt) homopolymers however, there often is a consistent length miscall that is not resolvable by coverage [184, 243, 83]. Strong light signals in one well of the picotiter plate may also result in insertions in sequences in neighboring wells. If the neighboring well is empty this can generate so-called ghost wells, i.e. wells for which a signal is recorded even though they contain no sequence template, hence the intensities measured are completely caused by bleed-over signal from the neighboring wells. Computational post-processing may correct for these artifacts [82].

As for Sanger sequencing, the error rate increases with the position in the sequence. In the case of 454 sequencing, this is caused by (1) a reduction in enzyme efficiency or loss of enzymes which results in a reduction of the signal intensities, (2) some molecules on the beads no longer being elongated and (3) by an increasing, so-called, phasing effect. Phasing is observed when a population of DNA molecules amplified from the same starting molecule (ensemble) is sequenced, and describes the process whereby not all molecules in the ensemble are extended in every cycle. This causes the molecules in the ensemble to lose synchrony/phase, and results in an echo of the preceding cycles to be added to the signal as noise.

The current 454/Roche GS FLX Titanium platform makes it possible to sequence about 1.5 million such beads in a single experiment and to determine sequences of length between 300-500nt. The length of the reads is determined by the number of flow cycles, i.e. the number of times all four nucleotides have been washed over the plate, as well as by the base composition and the order of the bases in the sequence to be determined. Currently, 454/Roche limits this number to 200 flow cycles, resulting in an expected average read length of about 400nt. This is largely due to limitations imposed by the efficiency of polymerases and luciferases which drops over the sequencing run resulting in decreased base qualities. Later in 2011, 454/Roche will release new versions of their sequencing chemistry allowing for sequencing of about two times longer reads. Currently the platform allows the creation of about 750Mb of DNA sequence per day at a cost of about 20$/Mb.

IDT Exhibit 1061.0025

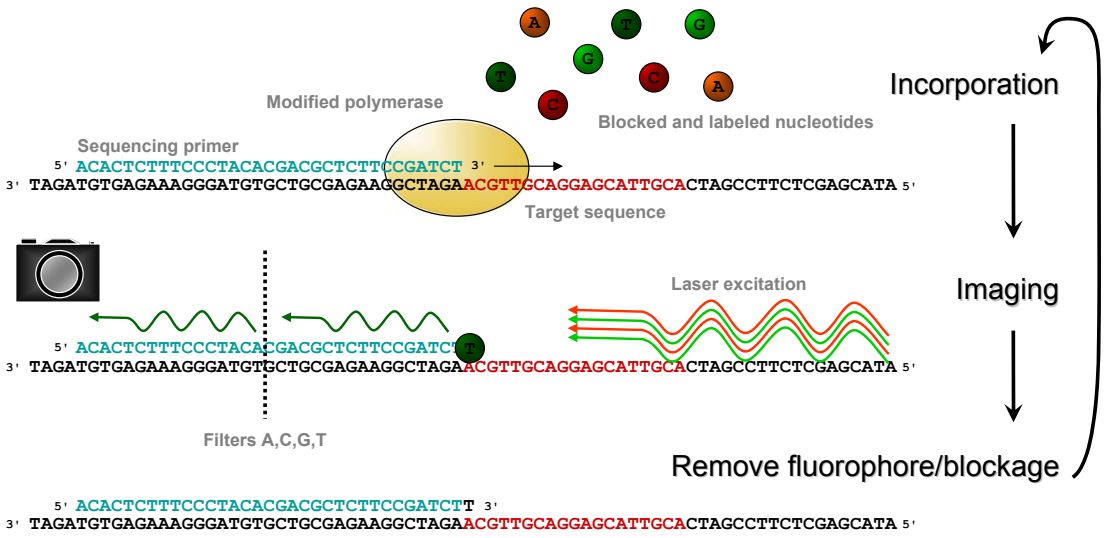

Figure 2.3: Reversible terminator chemistry applied by the Illumina Genome Analyzer. Sequencing primers are annealed to the adapters of the sequences to be determined. Polymerases are used to extend the sequencing primers by incorporation of fluorescently labeled and terminated nucleotides. The incorporation stops immediately after the first nucleotide due to the terminators. The polymerases and free nucleotides are washed away and the label of the bases incorporated for each sequence is read with four images taken through different filters (T nucleotide filter is indicated in the figure) and using two different lasers (red: A, C and green: G, T) to illuminate fluorophores. Subsequently the fluorophores and terminators are removed and the sequencing continued with the incorporation of the next base.

2.3 Illumina Genome Analyzer

The reversible terminator technology used by the Illumina Genome Analyzer employs the sequencing by synthesis concept that is most similar to that used in Sanger sequencing i.e. the incorporation reaction is stopped after each base, the label of the base incorporated is read out with fluorescent dyes and the sequencing reaction is then continued with the incorporation of the next base [16, 231] (see figure 2.3).

Like 454/Roche, the Illumina sequencing protocol requires that the sequences to be determined are converted into a sequencing library, which allows them to be amplified and immobilized for sequencing [16, 66]. For this purpose two different adapters are added to the 5' and 3' ends of all molecules using ligation of so-called forked adapters[2] . The library is then amplified using longer primer sequences which extend and further diversify the adapters, i.e. add further unique nucleotides at both adapter ends, to create the final sequence needed in subsequent steps.

This double-stranded library is melted using sodium hydroxide to obtain single stranded DNAs, which are then pumped at a very low concentration through the channels of a flow cell. This flow cell has on its surface two populations of immobilized oligonucleotides complementary to the two different single stranded adapter ends of the sequencing library. These oligonucleotides will hybridize to the single stranded library molecules. By reverse strand synthesis starting from the hybridized (double-stranded) part, the new strand being created

[2] Also called Y-Adapters. Duplex of two oligonucleotides with perfect base-pairing at one molecule end, but non-pairing/unique sequence at the other end.

IDT Exhibit 1061.0026

is covalently bound to the flow cell. If this new strand bends over and attaches to another oligonucleotide complementary to the second adapter sequence on the free end of the strand, it can be used to synthesize a second covalently bound reverse strand. This process of bending and reverse strand synthesis, called bridge amplification, is repeated several times and creates what are termed clusters, the accumulation of several thousand copies of the original sequence in very close proximity to each other on the flow cell [16, 66].

These randomly distributed clusters contain molecules that represent the forward as well as reverse strands of the original sequences. Before determining the sequence, one of the strands has to be removed to prevent it from hindering the extension reaction sterically or by complementary base pairing. Selective strand removal targets base modifications of the oligonucleotide populations on the flow cell. Following strand removal, each cluster on the flow cell consists of single stranded, identically oriented copies of the same sequence; which can be sequenced by hybridizing the sequencing primer onto the adapter sequences and starting the reversible terminator chemistry.

"Solexa sequencing", as it was introduced in early 2007, initially allowed for the simultaneous sequencing of several million very short sequences (at most 26nt) in a single experiment. In recent years there have been several technical, chemical and software updates. The product, which is now called the Illumina Genome Analyzer (GA), has increased flow cell cluster densities (more than 300 million clusters per run), a wider range of the flow cell is imaged, and sequence reads of up to 125nt can be generated.

A technical update (Paired End Module, PEM) also enabled the sequencing of the reverse strand of each molecule. This is achieved by chemical melting and washing away the synthesized sequence, repeating a few bridge amplification cycles for reverse strand synthesis and then selectively removing the starting strand (again using base modifications of the flow cell oligonucleotide populations), before blocking 3' ends and annealing another sequencing primer for the second read (see figure 2.4 on the next page). Using this "paired end sequencing" approach, approximately twice the amount of data can be generated.

The Illumina library and flow cell preparation includes several *in vitro* amplification steps which cause a high background error rate and contribute to the average error rate of about 10^{-2} to 10^{-3} [49, 116]. Further, the flow cell preparation creates a fraction of ordinary-looking clusters which are initiated from more than one individual sequence. These result in mixed signals and mostly low quality sequences for these clusters. In an effect similar to the 454 ghost wells, the Illumina image analysis software may identify reagent crystals, dust and lint particles as clusters and call sequences from these (see section 3.7 on page 59 of chapter 3).

As is the case for the other platforms, the error rate increases with increasing position in the determined sequence. This is mainly due to phasing, which increases the background noise as sequencing progresses. While the ensemble sequencing process for pyrosequencing creates unidirectional phasing from lagging, non-extended molecules, reversible terminator sequencing creates bi-directional phasing [61, 116, 106] as some incorporated nucleotides may also fail to be correctly terminated – allowing the extension of the sequence by another nucleotide in the same cycle. With increasing cycle numbers, the intensities extracted from the clusters decline [61, 198, 116]. This may be due to fewer molecules participating in the extension reaction as a result of non-reversible termination/DNA degradation, due to dimming effects of the sequencing fluorophores, or due to an increase in background noise by the accumulation of the sequencing fluorophores in the flow cell.

In early versions of the chemistry, specifically one of the fluorophores (T fluorophore) stuck to the clusters creating a biased background signal and thereby an overcall of the respective

IDT Exhibit 1061.0027

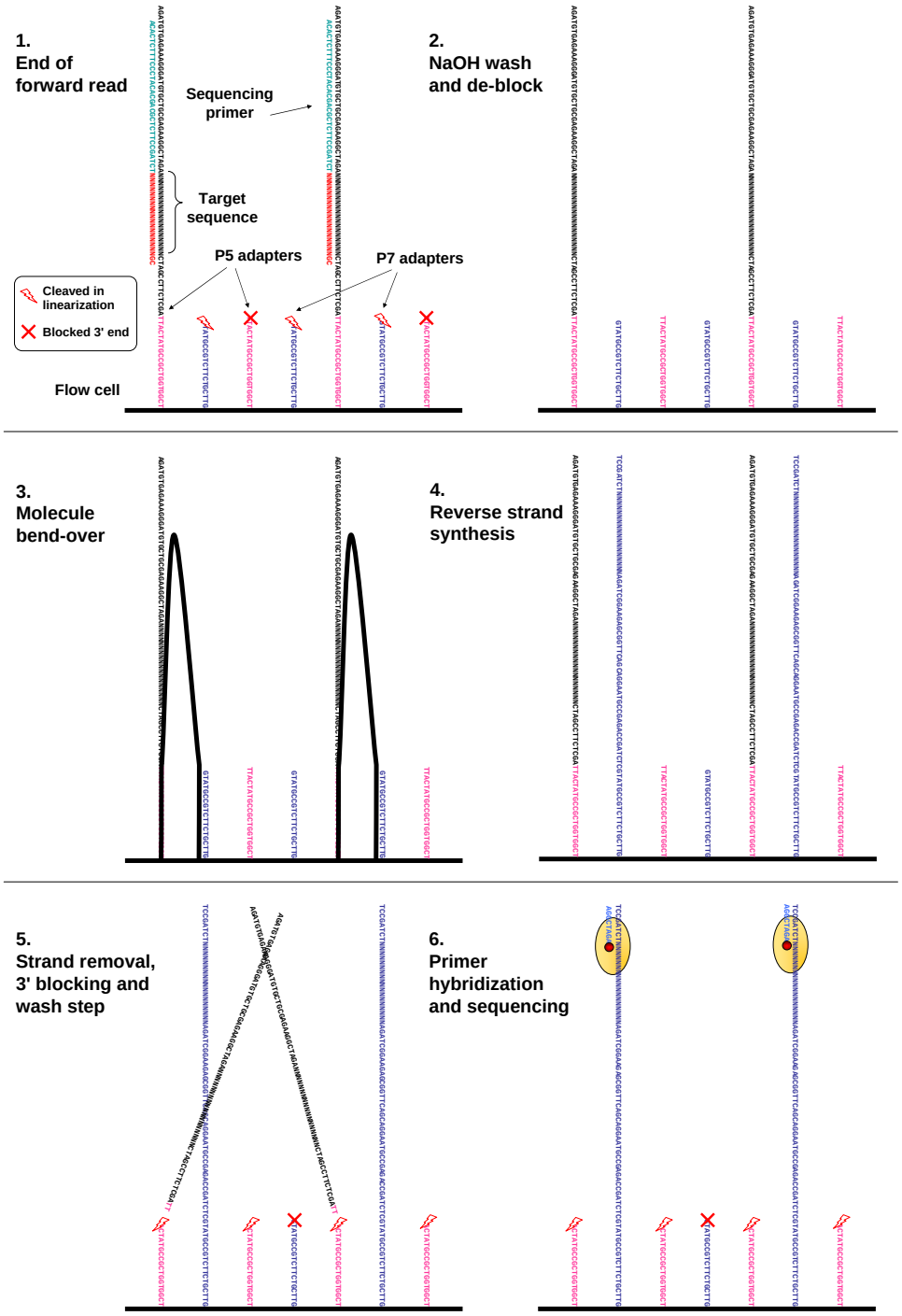

Figure 2.4: Synthesis reaction performed for Illumina paired end sequencing. After the end of the forward read **(1)**, the synthesized sequence of this read is chemically melted using sodium hydroxide, 3' blocking is removed from all sequences and the flow cell washed **(2)**, molecules bend over and a few bridge amplification cycles are performed for reverse strand synthesis **(3+4)** and then selectively removing the starting strand **(5)** by targeting base modifications of the flow cell oligonucleotide populations. Afterwards, free 3' ends are blocked and another sequencing primer annealed for sequencing of the second read **(6)**.

IDT Exhibit 1061.0028

base [116, 106]. The simultaneous identification of four different nucleotides is also a general issue. The Genome Analyzer uses four fluorescent dyes to distinguish the four nucleotides A, C, G and T. Of these, two pairs (A/C and G/T) excited using the same laser, are similar in their emission spectra and show only limited separation using optical filters. Therefore, the highest substitution errors observed are between A/C and G/T [49, 116] (see also figure 3.2 on page 43 and figure 4.3 on page 74).

Even though the Illumina Genome Analyzer reads show a higher average error rate and are considerably shorter than 454/Roche reads, this instrument determines more than 10,000Mb per day with a price of about 0.50$/Mb. This is more than ten times higher daily throughput than 454/Roche and for a considerably lower price per megabase.

2.4 Life Technologies SOLiD

The prototype of what was further developed and later sold by Applied Biosystems (ABI) and later by Life Technologies as the SOLiD sequencing platform, was developed by George Church's laboratory at Harvard Medical School and the Howard Hughes Medical Institute and published in 2005 [210]. With its commercial release in late 2007, SOLiD was the third new high-throughput system entering a highly competitive market with all three vendors selling their instruments for around half a million dollars. The Church lab at Harvard Medical School continued the development of the system and now offers a cheaper (< $200,000) open source version of the system (called Polonator) in collaboration with Dover System. In the third quarter of 2008, a biotechnology company from Mountain View, California, named Complete Genomics started offering a human genome sequencing service. Their technology is also based on the Church lab sequencing-by-ligation concept, but combines it with a new strategy of sequencing library construction and sequence immobilization using rolling circle amplification [52]. Here, I focus on the commercial SOLiD system as this is the most widespread application of this concept.

The principle behind sequencing-by-ligation is very different from the approaches discussed thus far. The sequence extension reaction is not carried out by polymerases but rather by ligases [210] (see figure 2.5 on the next page for a schematic representation of the SOLiD 2-4 platform). In the sequencing-by-ligation process, a sequencing primer is hybridized to single-stranded copies of the library molecules to be sequenced. A mixture of 8mer probes carrying four distinct fluorescent labels compete for ligation to the sequencing primer. The fluorophore encoding, which is based on the two 3' most nucleotides of the probe, is read. Three bases including the dye are cleaved from the 5' end of the probe, leaving a free 5' phosphate on the extended (by five nucleotides) primer, which is then available for further ligation. After multiple ligations (typically up to 10 cycles), the synthesized strands are melted and the ligation product is washed away before a new sequencing primer (shifted by one-nucleotide) is annealed. Starting from the new sequencing primer the ligation reaction is repeated.

The same process is followed for three other primers, facilitating the read out of the dinucleotide encoding for each start position in the sequence. Using a specific fluorescent label encoding, the dye read outs (i.e. colors) can be converted to a sequence [19]. This conversion from color space to sequence requires a known first base, which is the last base of the library adapter sequence. Given a reference sequence this encoding system allows for the detection of machine errors and the application of an error correction to reduce the average error rate. In the absence of a reference sequence, however, color conversion fails with an error in the dye read out and causes the sequence downstream of the error to be incorrect.

IDT Exhibit 1061.0029

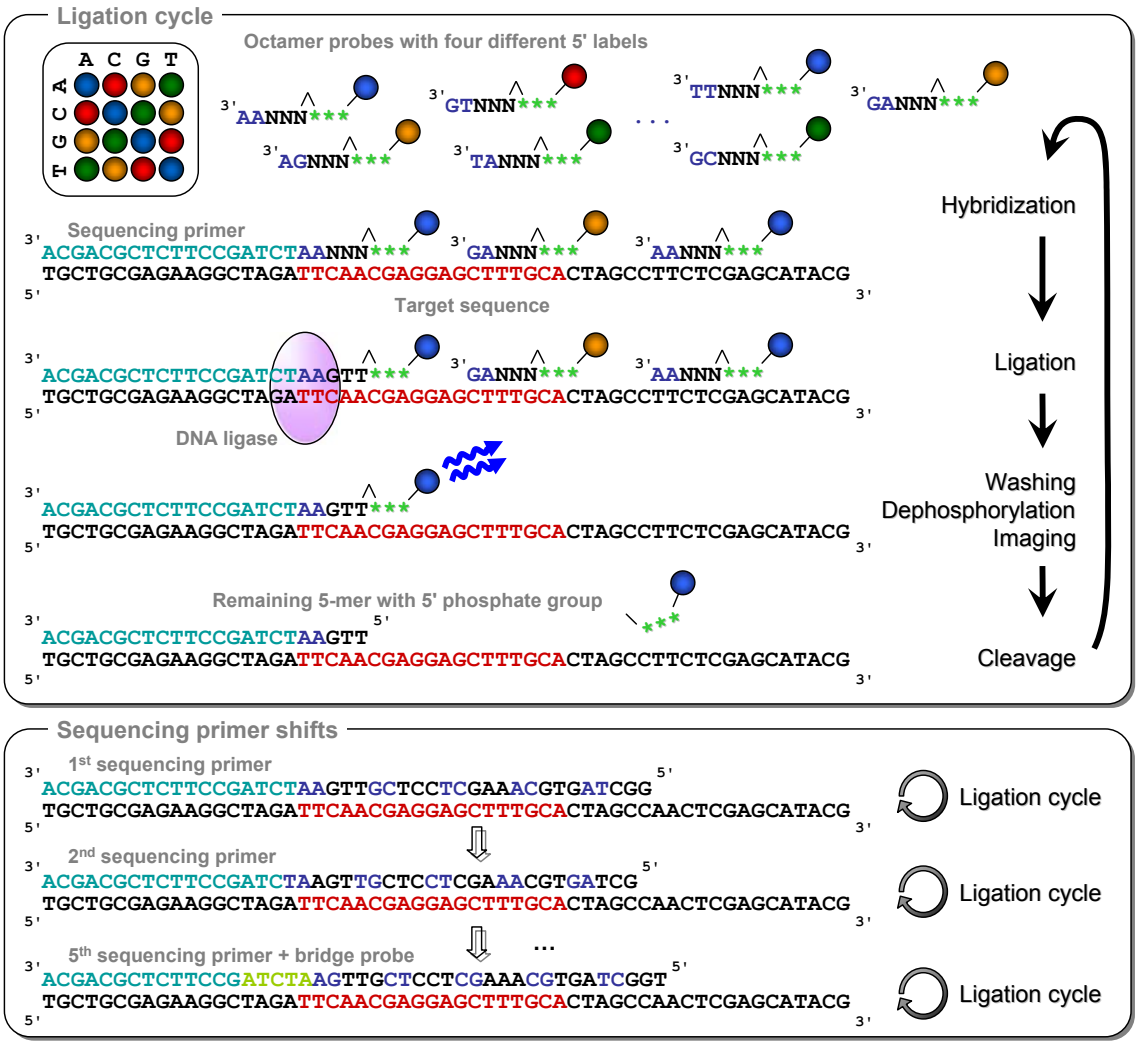

Figure 2.5: Life Technologies's SOLiD sequencing-by-ligation. A sequencing primer is annealed to single stranded copies of sequences to be determined. Octamer probes are hybridized, ligated to the sequencing primer and a fluorescent dye at the 5' end of the ligated 8-mer probes, encoding the 3' most two nucleotides of the probe, is read out. Non-extended primers are dephosphorylated. Three nucleotides of the probe including the dye are cleaved, creating a free 5' phosphate for further ligations. After multiple ligations, the synthesized strands are melted and the ligation product is washed away before a new, by-one-nucleotide-shifted sequencing primer is annealed. Starting from the new sequencing primer the ligation reaction is repeated. The same is done for three other primers, allowing the read out of the dinucleotide label for every position in the sequence.

IDT Exhibit 1061.0030

For parallelization, the sequencing process uses beads covered with multiple copies of the sequence to be determined. These beads are created in a similar fashion to that described earlier for the 454/Roche platform. In contrast to the 454/Roche technology, the SOLiD system does not use a picotiter plate for fixation of the beads in the sequencing process; instead the 3' ends of the sequences on the beads are modified in a way that allows them to be covalently bound onto a glass slide. As for the Illumina system, this creates a random dispersion of the beads in the sequencing chamber and allows for higher loading densities. However, random dispersion complicates the identification of bead positions from images, and results in the possibility that chemical crystals, dust and lint particles can be misidentified as clusters. Further, dispersal of the beads results in a wide range of inter-bead distances which then have differing susceptibility to signals from neighboring beads.

Types and causes of sequence errors are diverse: First, the *in vitro* amplification steps cause a higher background error rate than *in vivo* amplifications using the Sanger cloning approach. Secondly, beads carrying a mixture of sequences and beads in close proximity to one another create false reads and low quality bases. Further, signal decline and incomplete dye removal result in increasing error as the ligation cycles progress [48]. Phasing, as described earlier, is a minor issue on this platform as sequences not extended in the last cycle are non-reversibly terminated using phosphatases. Since hybridization is a stochastic process and probes do not necessarily hybridize adjacent to the (extended) sequencing primer, this causes a considerable reduction in the number of molecules participating in subsequent ligation reactions, and therefore substantial signal decline. Given the high efficiency of phosphatases the remaining phasing effect can be considered very low. However, incomplete cleavage of the dyes may allow cleavage in the next ligation reaction, which then allows for the extension in the next but one cycle. This causes a different phasing effect and additional noise from the previous cycle's dyes in the dye identification process.

The SOLiD system currently allows sequencing of more than 300 million beads in parallel, with a typical read length of between 25 and 75nt. At the time of writing, the ABI SOLiD system is therefore comparable to the Illumina Genome Analyzer system in terms of daily throughput and price per million nucleotides (≈10,000Mb/day, ≈0.50$/Mb). Average error rates are dependent on the availability of a reference genome for error correction (10^{-3}-10^{-4} vs. 10^{-2}-10^{-3}). In the absence of a reference genome, assembly and consensus calling may be performed based on dye read outs (so called color space sequences) to reduce the errors before conversion to the nucleotide sequence. If no reference genome is available for error correction, and no assembly and consensus calling is performed, then the average error rate is higher than for the Illumina GA.

2.5 Helicos HeliScope

Helicos was the first company to sell a sequencer able to sequence individual molecules instead of molecule ensembles created by an amplification process. Single-molecule sequencing has the advantage that it is not affected by biases or errors introduced in a library preparation or amplification step, and may facilitate sequencing of minimal amounts of input DNA. Using methods able to detect non-standard nucleotides, it could also allow for the identification of DNA modifications, commonly lost in the *in vitro* amplification process.

The HeliScope, as the Helicos sequencer is called, was first sold in March 2008, and at the end of the first quarter of 2009 only four machines had been installed world-wide. This might be surprising given the advantages of single molecule sequencing, but probably reflects both

IDT Exhibit 1061.0031

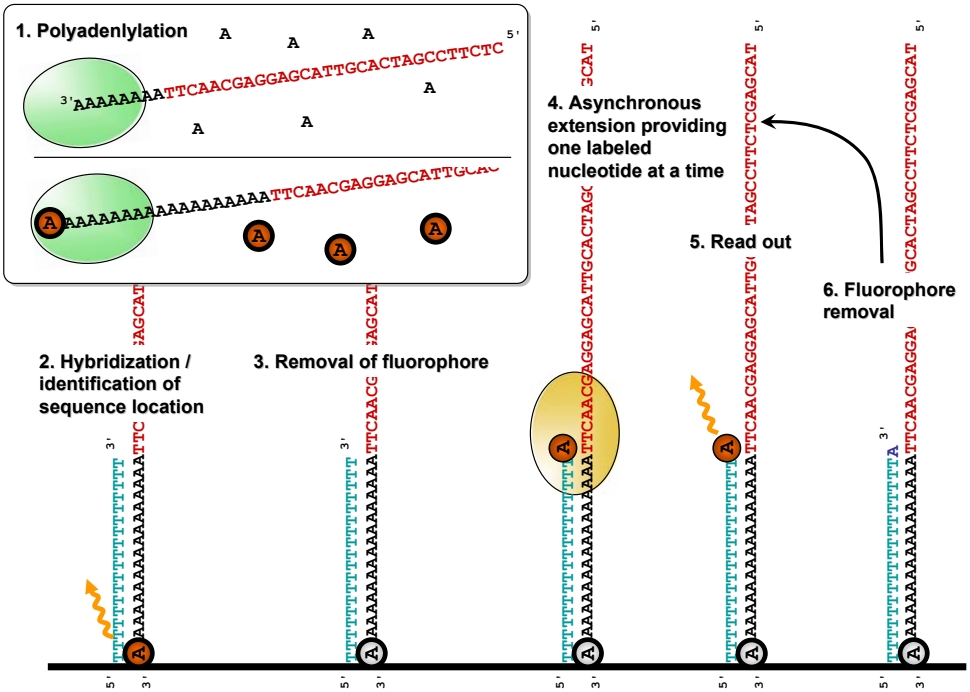

Figure 2.6: Asynchronous virtual-terminator chemistry performed by the HeliScope. Input DNA is fragmented, melted and polyadenylated. A fluorescently labeled adenine is added in the last step. This single stranded DNA is washed over a flow cell with poly-T-oligonucleotides allowing hybridization. The bound coordinates on the flow cell are determined using the fluorescently labeled adenines. Having the coordinates identified, the fluorescent label of the 3' adenines is removed. Polymerases are washed through with one type of fluorescently labeled nucleotides (A, C, G and T) at a time and the polymerases extend the reverse strand of the sequences starting from the poly-T-oligonucleotides. The nucleotide incorporation of the polymerases is slowed down by the fluorescent labeling and allows for at most one incorporation before the polymerase is washed away. The flow cell is then imaged, the fluorescent dyes removed and the reaction continued with another nucleotide.

the specific limitations of this platform, the price (about one million dollars), and a relatively small market that already has invested extensively in new sequencing technologies.

The technology applied by the HeliScope (figure 2.6) could be termed asynchronous virtual-terminator chemistry [89]: Input DNA is fragmented and melted before a poly-A-tail is synthesized onto each single stranded molecule using a polyadenylate polymerase. In the last step of polyadenylation, a fluorescently labeled adenine is added. The library, i.e. the polyadenylated single stranded DNA, is washed over a flow cell where the poly-A tails bind to poly-T-oligonucleotides. The bound coordinates on the flow cell are determined using a fluorescence-based read out of the flow cell. Having these coordinates identified, the fluorescent label of the 3' adenine is removed and the sequencing reaction started.

Polymerases are washed through the flow cell with one type of fluorescently labeled nucleotides (A, C, G and T) at a time and the polymerases extend the reverse strand of the sequences starting from the poly-T-oligonucleotides. The nucleotide incorporation of the polymerases is slowed down by the fluorescent labeling and allows for at most one incorporation before the polymerase is washed away together with the non-incorporated nucleotides (termed virtual termination; [253, 25]). The flow cell is then imaged again, the fluorescent dyes are removed and the reaction continued with another nucleotide. By this process not every molecule is extended in every cycle, which is why it is an asynchronous sequencing process resulting in sequences of different length (as is the case for the 454/Roche platform).

IDT Exhibit 1061.0032

Since single molecules are sequenced, the signals being measured are weak (complicating signal-to-noise thresholding), and there is no possibility that misincorporation errors can be corrected by an ensemble effect. Due to the fact that molecules are attached to the flow cell by hybridization only, there is a chance that template molecules can be lost in the wash steps. In addition, molecules may be irreversibly terminated by the incorporation of incorrectly synthesized nucleotides. Overall, reads are between 24 to 70nt long (average 32nt) [182] and thus shorter than for the other platforms. Due to the higher number of sequences determined in parallel, the total throughput per day (4150Mb/day with ≈0.33$/Mb [182]) is in a similar range as for the GA and SOLiD systems. The average error rate, which is in the range of a few percent, is slightly higher than for all other instruments and biased towards insertions and deletions rather than substitutions.

2.6 Single Molecule Real Time (SMRT)

Another technology for sequencing individual molecules is Pacific Biosciences's SMRT (Single Molecule Real Time) sequencing technology [118]. This technology performs the sequencing reaction on silicon dioxide chips with a 100nm metal film containing thousands of tens-of-nanometers diameter holes, so called zero-mode waveguides (ZMWs) [57]. Each ZMW is used as a nano visualization chamber, providing a detection volume of about 20 zeptoliters (10^{-21} liters). At this volume, a single molecule can be illuminated while excluding other labeled nucleotides in the background – saving time and sequencing chemistry by omitting wash steps.

For SMRT sequencing (figure 2.7 on the next page), a single DNA polymerase is fixed to the bottom of the surface within the detection volume of each ZMW. Nucleotides with different fluorescent dyes attached to the phosphate chain are used in concentrations allowing normal enzyme processivity. As the polymerase incorporates complementary nucleotides, the nucleotide is held within the detection volume for tens of milliseconds, orders of magnitude longer than for unspecific diffusion events. This way the fluorescent dye of the incorporated nucleotide can be identified during normal speed reverse strand synthesis [57]. Further, by attaching the fluorescent dyes to the phosphate chain of the deoxy-nucleotides the dye is released with the cleaved pyrophosphate, generating an unmodified complimentary DNA strand.

In pilot experiments, Pacific Biosciences showed that their technology allows for direct sequencing of a few thousand bases before the polymerase is denatured due to optic and thermal stress from the laser read-out of the dyes. They were also able to show that they can measure differences in polymerase kinetics to such an extend that modified nucleotides may be detected [70]. The SMRT technology was intended for release in the fourth quarter of 2010. Due to this recent release, the amount of information on the actual instrument is very limited and it is likely that further development is needed to create a robust system over the next years.

2.7 Upcoming developments

Motivated by the goal of sequencing a genome for $1,000 set by NIH/NHGRI to enable personalized medicine[3], the throughput of all systems described is constantly increasing and numbers given here are rapidly outdated. In the third quarter of 2010, Illumina started

[3] `http://www.nih.gov/news/pr/aug2007/nhgri-01.htm`

IDT Exhibit 1061.0033

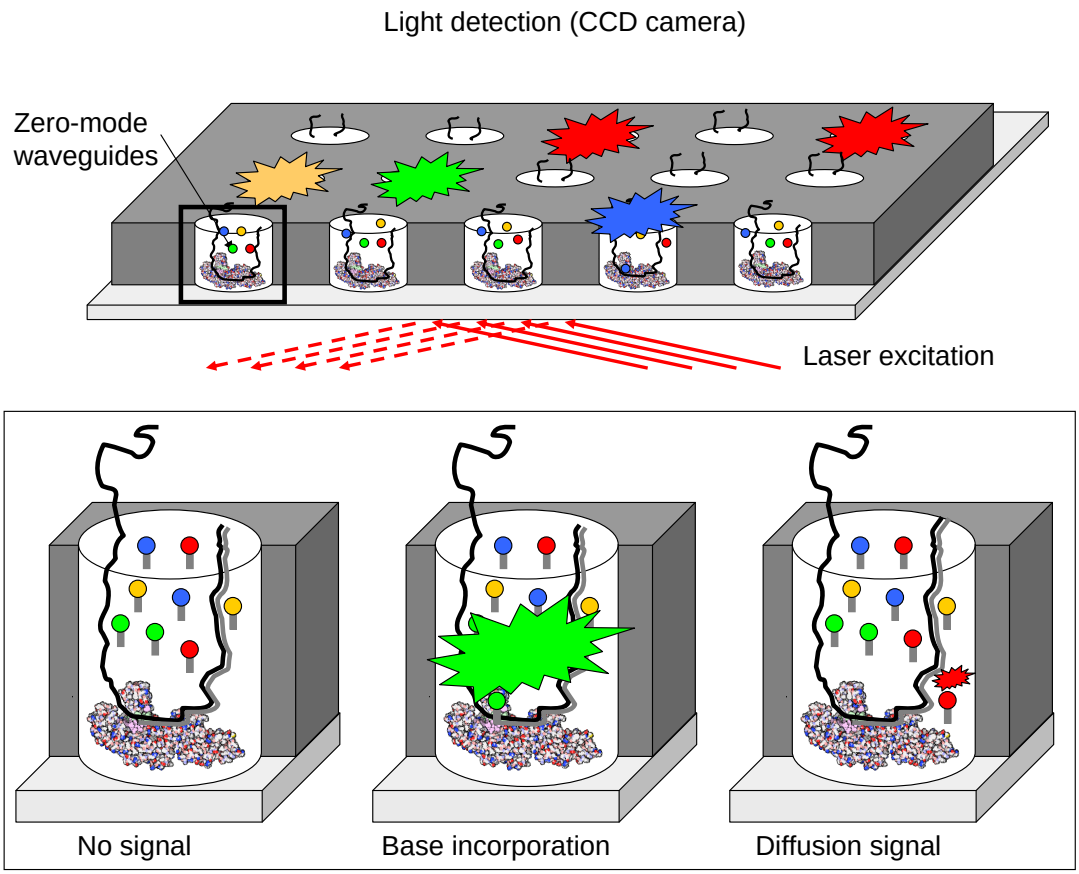

Figure 2.7: Single Molecule Real Time sequencing as implemented by Pacific Biosciences. DNA polymerases are fixed to the bottom of zero-mode waveguides (ZMW), a tens-of-nanometers diameter holes in a 100nm metal film. Each ZMW is used as a nano visualization chamber with a very small detection volume allowing a single molecule to be illuminated, while excluding other labeled nucleotides in the background. As the polymerase incorporates complementary nucleotides, the nucleotide with a fluorescent dye attached to its phosphate chain is held within the detection volume for tens of milliseconds, orders of magnitude longer than for unspecific diffusion events. The fluorescent dye attached to the phosphate chain of the deoxy-nucleotides is released into solution when the pyrophosphate is cleaved during incorporation, generating an unmodified complimentary DNA strand. [Image of DNA Polymerase taken from `http://www.rcsb.org/pdb/static.do?p=education_discussion/molecule_of_the_month/pdb3_1.html`]

IDT Exhibit 1061.0034

shipping the HiSeq2000 system, an updated version of the Genome Analyzer instrument determining sequences of clusters on bottom and top of the flow cell using confocal scanning rather than total internal reflection optics and capable of processing two flow cells in parallel. Also 2010, Life Technologies introduced SOLiD4, an update to the SOLiD platform which among other improvements now allows paired end sequencing.

While on the one end of the spectrum instrument throughput increases, some vendors also recently started to offer budget versions of their instruments (e.g. Illumina Genome Analyzer IIe, Illumina MySeq or 454/Roche GS Junior) with lower sequencing capacity. Other vendors, like the Life Technologies company Ion Torrent, also try to establish their market in "lower" high-throughput sequencing instruments with benchtop instruments. In the case of Ion Torrent, the pyrosequencing concept is used with a different detection system for the incorporation reaction. While the 454/Roche instrument uses the released pyrophosphate ions in a light reaction and measures the light intensities, the Ion Torrent instrument uses semiconductor chips to measure the pH change in the solution. Even though the instrument costs are around one quarter to one third of the price for the big instruments, a similar infrastructure is required for data processing and analysis, which keeps the financial investments considerably high. Further, the costs per base are generally higher than for the other instruments.

At the same time, a completely new generation of sequencers is already on the horizon. What started with the Helicos and Pacific Biosciences systems – the sequencing of single molecules without prior library preparation or amplification – will likely become a popular paradigm. Specifically, three other systems have captured media and scientific attention well in advance of their actual availability: Oxford Nanopore's BASE technology [37], IBM's proposal of silicon-based nanopores [189] and Life Technologies single molecule sequencing technology based on quantum dots.

Oxford Nanopore's BASE technology offers the potential to identify individual nucleotide modifications (e.g. 5-methyl-cytosine versus cytosine) during the sequencing process [37]. The idea behind this technology is the identification of individual nucleotides using a change in the membrane potential as the nucleotides pass through a modified α-hemolysin membrane pore with a cyclodextrin sensor [37, 9]. To apply this technology for sequencing, the pore has to be fused to an exonuclease which degrades single stranded DNA sequences and releases individual nucleotides into the pore. In addition, the technology needs to be parallelized in array format, before its release as a high-throughput sequencing platform. While the sensitivity for individual nucleotide modifications seems to be a major advantage, the destructive fashion of the outlined sequencing process might be considered a hindrance for applications with precious samples, and does obviously not allow a second read cycle for error reduction.

In early October 2009, IBM issued a press release [189] describing a method for controlling the speed of an individual DNA strand passing through a nanopore. For this purpose they developed a multilayer metal/dielectric nanopore device which utilizes the interaction of the DNA backbone charges with a modulated electric field to trap and slowly release an individual DNA molecule. The technology described could theoretically be combined with, for example, the Nanopore technologies developed at Harvard University [2] or the previously described BASE [37] technology where it may overcome the destructive approach followed so far.

At the 11th annual Advances in Genome Biology and Technology (AGBT 2010) meeting in Marco Island (Florida, USA), Life Technologies presented the first results on their experiments with quantum dots, light-emitting semiconductor nanocrystals of 2-10 nm diameter,

IDT Exhibit 1061.0035

attached to DNA polymerases (Joseph M. Beechem, personal communication, May 12th 2010). These dots can be laser excited with a specific wavelength and then via Fluorescence Resonance Energy Transfer (FRET) enable light emission from the fluorescently labeled nucleotides at a different wave length while the polymerase incorporates them during complementary strand synthesis. These quantum dots do not only provide/transfer the energy for the fluorescence signal, they also enhance the signal strength. While the actual sequencing process is similar to the above described SMRT technology from Pacific Biosciences, except having a free polymerase and a DNA template covalently bound to a flow cell, the sequencing run can be stalled mid-way to wash off polymerases and nucleotides and replace them with new ones, thus replacing reagent molecules affected by chemical damage. This allows a reset of the error process and thus very long reads. When also removing the so-far synthesized strand, for example by chemical denaturation with sodium hydroxide, the sequencing reaction can be reset completely – allowing long and low error sequences to be generated from multiple read outs of the same template molecule.

These new technologies on the horizon, suggest the major future directions in the field of DNA sequencing: the ability to use individual molecules without any library preparation or amplification, the identification of nucleotide modifications and the ability to generate longer sequence reads. These developments are likely to facilitate future research in many fields, make data analysis easier and further reduce per base sequencing costs.

2.8 General considerations

All current high-throughput technologies have an average error rate that is considerably higher than the typical 1/10,000 to 1/100,000 observed for high quality Sanger sequences. Further, the GS FLX Titanium, Genome Analyzer, SOLiD and HeliScope platforms each have very specific biases and limitations, making it necessary to choose a platform appropriate for a specific project or application (for a summary see figure 2.8 on the facing page). A combination of technologies [153, 188, 47, 34] and experimental protocols [254, 252, 113] may also be appropriate, and even complementary, for specific projects.

High quality Sanger sequencing is now commonly used to generate low coverage sequencing of individual positions and regions (e.g. diagnostic genotyping) or the sequencing of virus- and phage-sized whole genomes. As the Sanger sequence length is longer than most abundant short repeat classes, it allows the unambiguous assembly of most genomic regions – something which is generally not possible using the shorter read platforms. However, the technology is expensive and too slow for sequencing a large number of samples, extended genomic regions or the many molecules required for quantitative applications (e.g. gene expression quantification; ChIP-Seq and MeDip-Seq).

For quantitative applications the HeliScope provides the highest throughput in terms of sequence number and has the advantage of not requiring a multistep library preparation protocol. On the other hand, the HeliScope provides the lowest resolution in mapping accuracy for complex genomes due its short read length and error profile. The GA or SOLiD platforms may thus provide equivalent results for quantitative applications, while providing fewer but longer reads and requiring a more elaborate library preparation.

While it has not yet been fully analyzed, it is possible (and even likely) that library preparation protocols could bias the sequence representation in a library [49, 183, 139, 32], making the replacement of this step an important goal. Further, multi-step library preparation protocols require higher amounts of input material, limiting their general application. However,

IDT Exhibit 1061.0036

	Throughput	Length	Quality	Costs	Applications	Main sources of errors
Sanger	6Mb/day	800nt	10^{-4}-10^{-5}	~500$/Mb	Small sample sizes, genomes/scaffolds, InDels/SNPs, long haplotypes, low complexity regions, ...	Polymerase/amplification, low intensities/missing termination variants, contaminant sequences
454/Roche	750Mb/day	400nt	10^{-3}-10^{-4}	~20$/Mb	Complex genomes, SNPs, structural variation, indexed samples, smallRNA+, mRNAs+, ...	Amplification, mixed beads, intensity thresholding, homopolymers, phasing, neighbor interference
Illumina	10000Mb/day	125nt	10^{-2}-10^{-3}	~0.50$/Mb	Complex genomes, counting (SAGE, CNV ChIP, small RNA), mRNAs, InDels/ homopolymers, structural variation, bisulphite data, indexing, SNPs+, ...	Amplification, mixed clusters/neighbor interference, phasing, base labeling
SOLiD4	10000Mb/day	60nt	10^{-2}-10^{-3}	~0.50$/Mb	Complex small genomes, counting (SAGE, ChIP, small RNA, CNV), SNPs, mRNAs, structural variation, indexing, ...	Amplification, mixed beads, phasing, signal decline, neighbor interference
Helicos	5000Mb/day	32nt	10^{-2}	<0.50$/Mb	Non-amplifiable samples, counting (SAGE, ChIP, small RNA), ...	Polymerase, low intensities / thresholding, molecule loss/termination

+ High sequencing depth/number of runs required

Abbreviations: InDel – Insertion / Deletion, SNP – Single Nucleotide Polymorphism, SAGE – Serial Analysis of Gene Expression, mRNAs – messenger RNA/transcripts, ChIP – Chromatin immuno Precipitation, CNV – copy number variation

Figure 2.8: Comparison of high-throughput sequencing technologies available. The table summarizes throughput, length, quality and costs for the current versions of the mentioned technologies. These approximate numbers are constantly improving and based on figures available in January 2011. Costs do not include instrument acquisition and maintenance, further they may be affected by discounts and scale effects for multiple instruments. Where numbers are very similar, colors ranging from red (low performance) to green (good performance) indicate a general trend. In the last column, example applications fitting the throughput and error profiles of each of the platforms are given. Typically, this does not mean that the technology is limited to these applications, but that it is currently best suited to such applications.

protocols for library construction from limited sample amounts are available or being developed for each of the platforms, and publications demonstrate that while vendor protocols indicate the need for higher sample quantities (microgram range), many users are proceeding successfully with low input DNA amounts (nanogram to picogram range), as for example shown for ancient DNA specimens and more specifically with protocols developed in Leipzig at the Max Planck Institute for Evolutionary Anthropology [195, 26, 27, 143].

Like Sanger sequencing, the GS FLX Titanium provides a read length spanning many of the short repeat sequences – an important feature for accurate sequence mapping and assembly of genomes [241]. Despite the insertion/deletion errors, this technology has very low rates of misidentifying individual bases, making it perfectly suited for the identification of Single Nucleotide Polymorphisms (SNPs). Also geared to the identification of SNPs, at least for samples with an existing reference genome, is the SOLiD instrument with its dinucleotide encoding scheme [19]. Considerably higher coverage is needed in order to perform SNP calling with similar accuracy using the Illumina GA [88]. Neither the Illumina GA or the SOLiD sequencing systems are prone to generating high rates of small insertions or deletions, making them well suited for studying InDel variation.

As mentioned earlier, the drawback of short reads (below about 75nt) obtained from Helicos, SOLiD or Genome Analyzer instruments is in genome assembly and mapping applications, where the placement of repeated or very similar sequences cannot be resolved unambiguously. The correct placement is further complicated by high error rates introducing a requirement for a minimum sequence distance of an unambiguous placement.

Paired end or mate pair protocols (figure 2.9 on the next page) help to overcome some of these limitations of short reads [33] by providing information about relative location and orientation of a pair of reads. Currently a paired end protocol is only commonly applied on the Genome Analyzer and SOLiD4, while mate pair protocols are available for SOLiD,

IDT Exhibit 1061.0037

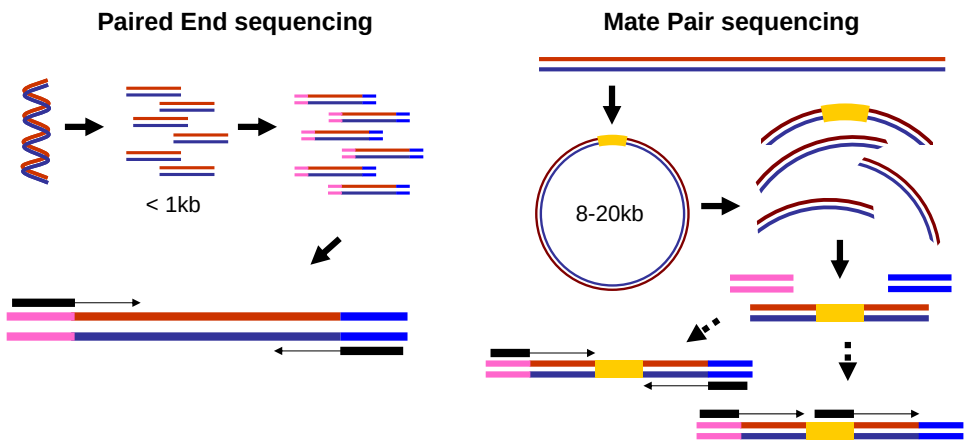

Figure 2.9: In paired end sequencing **(left)** the actual ends of rather short DNA molecules (<1kb) are determined, while for mate pair sequencing **(right)** the ends of long molecules are joined and prepared in special sequencing libraries. In these mate pair protocols, the ends of long, size-selected molecules are connected with an internal adapter sequence (i.e. linker, yellow) in a circularization reaction. The circular molecule is then processed using restriction enzymes or fragmentation. Fragments are enriched for the linker and outer library adapters are added around the two combined molecule ends. The internal adapter can then be used as a second priming site for an additional sequencing reaction in the same orientation or sequencing can be performed from the second adapter, from the reverse strand.

GS FLX Titanium and Genome Analyzer. In paired end sequencing the actual ends of rather short DNA molecules (<1kb) are determined, while mate pair sequencing requires the preparation of special libraries. In these protocols, the ends of longer, size-selected molecules (e.g. 8kb, 12kb or 20kb) are connected with an internal adapter sequence in a circularization reaction. The circular molecule is then processed using restriction enzymes or fragmentation before outer library adapters are added around the two combined molecule ends. The internal adapter can then be used as a second priming site for an additional sequencing reaction on the same immobilized molecules. Thus, mate pair sequencing provides distance information useful for assembly, but does not allow the merging of two overlapping end reads, as by design the molecule ends will not overlap in sequencing.

Due to the large amounts of sequences that can be generated on these sequencing platforms, there is interest in sequencing targeted regions (for example a genomic locus, from sequence capture experiments [77, 94, 26, 30]) in multiple individuals/samples instead of sequencing one sample to excessive depth. All technologies therefore provide a separation of their sequencing plate into defined regions or channels. However, at most sixteen or twenty-five such regions/channels are available (GS FLX Titanium and HeliScope plates), which may be not sufficient for some applications. Using different library construction protocols most platforms allow addition of sample-specific barcode (sometimes called "index") sequences to the library molecules. These molecules can then be sequenced in the same channel, and later separated (computationally) based on their barcode sequence [151, 60, 141, 150]. This facilitates highly parallel sequencing of a large number of samples beyond that possible using the physical channel separation. Currently such protocols (mostly non-vendor protocols) are available for the GS FLX Titanium, Genome Analyzer and SOLiD instruments.

Though sequencing prices per gigabase have fallen considerably in recent years, making projects like the 1,000 Human Genome Variation project [53], 1001 Arabidopsis thaliana Genomes project [239], whole genomes for 10,000 vertebrate species [91] or the International

IDT Exhibit 1061.0038

Cancer Genome Consortium [222] possible, high-throughput sequencing still has high acquisition, running and maintenance costs, which are not included in the numbers provided in figure 2.8 on page 37.

2.9 Summary and conclusions

The discussed technologies make it possible for even single research groups to generate large amounts of sequence data very rapidly and at substantially lower costs than traditional Sanger sequencing. While costs have been reduced to less than 4%-0.1% and time has been shortened by a factor of 100-1,000 based on daily throughput, the error profiles and limitations observed for the new platforms differ significantly from Sanger sequencing and between approaches. Some vendors recently started to offer budget versions of their instruments (e.g. Illumina Genome Analyzer IIe, Illumina MySeq or 454/Roche GS Junior) with lower sequencing capacity. However, financial investments remain considerably high – with costs per base generally higher than for the standard instrument, and equivalent infrastructure required. Often the choice of an appropriate sequencing platform is project-specific and sometimes even combinations can be advantageous. This may open the market further to companies and sequencing centers providing sequencing-on-demand services.

Over the last years, the whole field observed a shift from the amount of time required to prepare and run a sequencing experiment to the time required for the analysis of the generated data [176, 183, 191, 12]. It is likely that also in the future laboratories will need to invest considerable time, expertise and money in the design of experiments and the analysis of the vast quantities of data that will be generated. Smaller research groups may still find the costs of the infrastructure needed for storing, handling and analyzing several tens of gigabytes of raw sequence data and terabytes of several thousand intermediate files generated by these instruments each week too high. Even for larger groups and experienced genome centers this aspect remains an ever-increasing challenge for the ongoing use of these platforms. Thus especially financial considerations, the number of projects requiring high-throughput data and the interest of implementing own improvements to the instruments/protocols are important factors for instrument acquisition.

New technologies like SMRT sequencing by Pacific Biosciences, Quantum-dot sequencing by Life Technologies or BASE by Oxford Nanopore will allow sequencing long individual molecules without or with little preparation steps and probably even the identification of specific nucleotide modifications. Improvements to current instruments are likely to further increase throughput and reduce cost of reading out DNA molecules. Thus, the goal of a $1,000 human genome set by NIH/NHGRI for personalized medicine may soon be achieved. All these developments will hopefully facilitate future research in many fields and simplify biological data analysis.

The Max Planck Institute for Evolutionary Anthropology decided to have high-throughput technologies on site, allowing fast access and the development of application-specific protocols for these platforms. Thus very early on, two 454 FLX instruments were acquired and in late 2007 a single Illumina Genome Analyzer I instrument was installed at the institute. Shortly after this first Genome Analyzer I was updated to version II in summer 2008, four additional Genome Analyzer II instruments were bought. Having seven high-throughput and two Sanger capillary array sequencers on site provides high flexibility and fast access to these technologies, but also enabled the development of protocols and algorithms pushing instrument limits.

IDT Exhibit 1061.0039

IDT Exhibit 1061.0040

Chapter 3

Computational challenges from sequencing data production

One should never listen. To listen is a sign of indifference to one's hearers.
– Oscar Wilde [96](1204)

The described advances in DNA sequencing have revolutionized the field of molecular biology and specifically genomics, making it possible to generate large amounts of sequence data for answering biological questions very rapidly and at substantially lower costs. This brings a broader part of the scientific community into the position where a high-throughput project has to be designed or large sequence data sets have to be analyzed. The new technologies however come with some limitations and problems. For example, considerable variance in run quality, specific biases and sensitivities, pseudo-sequences, high error rates as well as adapter and chimera sequences are observed. These issues require either design of the project in a way that circumvents them, or at least considers them in data analysis. In this regard, I analyzed the most commonly[1] used high-throughput sequencing platform, the Illumina Genome Analyzer.

To recapitulate what was described in chapter 2; the Illumina Genome Analyzer is based on parallel, fluorescence-based readout of millions of immobilized sequences that are iteratively sequenced using reversible terminator chemistry [16]. A flow diagram with the steps involved from DNA sample to sequence read outs with quality score, is available in figure 3.1 on the next page.

Independent of the actual application, Illumina sequencing requires that the molecules to be determined are converted into special sequencing libraries, allowing molecules to be amplified, immobilized and primed for sequencing. Up to eight different DNA libraries can be loaded to the 8-lane flow cell. In each of the lanes, single stranded library molecules hybridize to complementary oligos which are covalently bound to the flow cell surface. Starting from the double stranded duplex, the reverse strand of each library molecule is synthesized and the now covalently bound molecule is then further amplified using bridge amplification. This generates randomly distributed sequence clusters, each containing more than 1000 copies

[1] Based on 1430 total instruments, listed world-wide by `http://pathogenomics.bham.ac.uk/hts/stats` in February 2011, 63% are Illumina instruments (638 Genome Analyzer, 265 HiSeq), 18% SOLiD, 16% 454 and 2% other.

IDT Exhibit 1061.0041

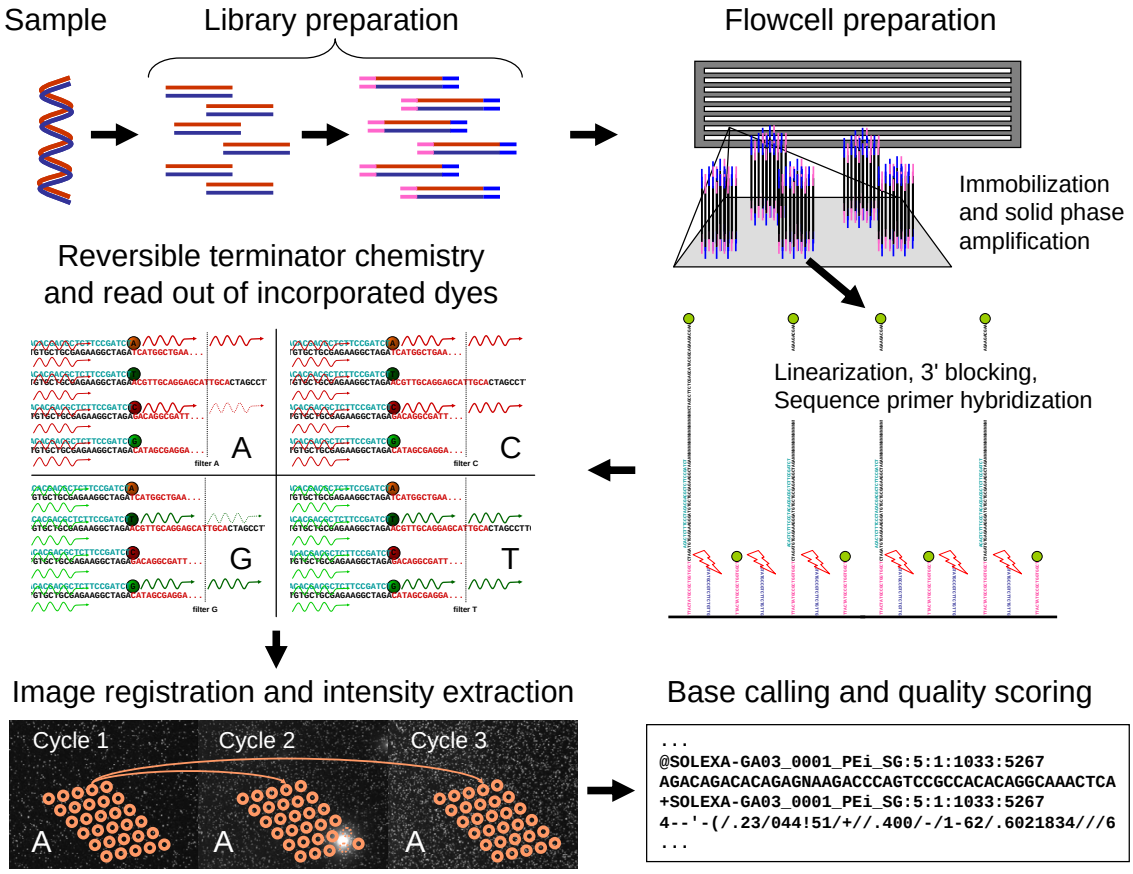

Figure 3.1: Illumina sequencing requires that a DNA sample is converted into special sequencing libraries. This is for example achieved by shearing DNA to a designated size and adding specific adapter sequences to both ends of the DNA molecules. These adapters allow molecules to be amplified and immobilized in one or more channels of the 8-channel flow cell. Immobilization and solid-phase amplification create randomly scattered clusters, consisting of a few thousand copies of the original molecule in very close proximity to each other. One of the DNA strands is removed to obtain single stranded, identically oriented copies, 3' ends of the DNA are blocked and a sequencing primer hybridized on the adapter sequences. Afterwards, the reversible terminator chemistry is performed. Here, four differently labeled nucleotides are provided and used for extension of the primers by DNA polymerases. The polymerase reaction immediately stops after the first base incorporation since nucleotides used are not only labeled, but also 3'-blocked. After washing away free nucleotides, the nucleotides incorporated are readout by piece-wise imaging of the flow cell. Then, the terminator and fluorophore is removed and another incorporation cycle started. The four images are overlaid (registered) and light intensities extracted for each cluster and cycle using a cluster position template obtained from the first instrument cycles. Resulting intensity files serve as input for base calling, the conversion of intensity values into bases and quality scores.

IDT Exhibit 1061.0042

of the starting molecule. One strand is then selectively removed, free 3' ends of the DNA are blocked and a sequencing primer is annealed onto the adapter sequences of the cluster molecules. Starting from these sequencing primers, the reversible terminator sequencing reaction is performed.

Fluorophores attached to the incorporated nucleotides are illuminated using a red and a green laser, and imaged through different filters, yielding four images per tile. The number of tiles varies; for Genome Analyzer I it is typically 300 tiles per lane, for Genome Analyzer II it is 100 tiles per lane and for Genome Analyzer IIx 120 tiles. For the new HiSeq instruments, which read the flow cell by confocal scanning rather than image tiling, the flowcell is arbitrarily divided into 32 tiles for computational purposes – 16 on the upper and 16 on the lower flowcell layer. After imaging, fluorescent labels and 3' terminators are removed and the next incorporation cycle started. Incorporation and imaging cycles are repeated up to a designated number of cycles, defining the read length for all clusters.

During progression of the sequencing run or when images for all cycles have been collected (depending on the setup and version), the four images per tile are overlaid (registered) and light intensities extracted for each cluster and cycle [16]. Briefly, clusters are identified by overlaying band-pass filtered and transformed/scaled images of the first few cycles. The resulting cluster position template is then aligned with images of all cycles and the intensities minus the surrounding background in the four different images extracted. Resulting intensity files serve as input for base calling, the conversion of intensity values into bases.

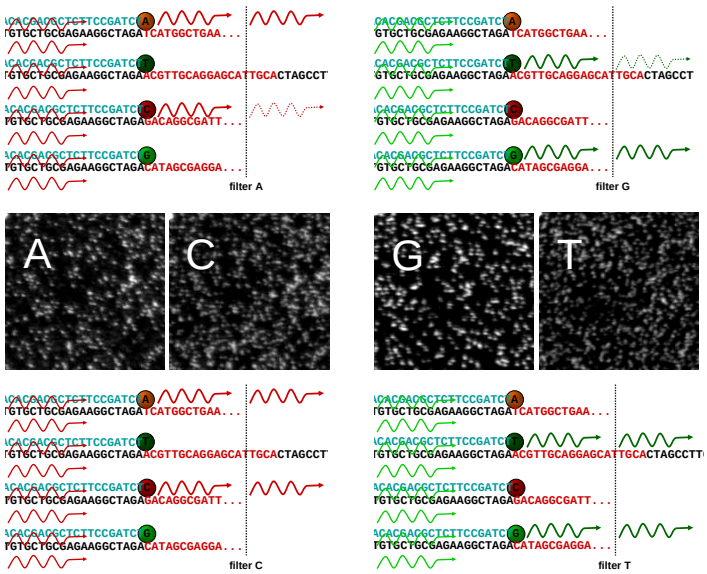

Figure 3.2: The fluorophores attached to the nucleotides are illuminated using a red and a green laser in a total internal reflection optics system. The four different fluorophores are imaged through different filters, yielding four images per tile. The A and C read outs as well as the G and T read outs are correlated (cross-talk) due to similar emission spectra of fluorophores used and their limited separation by the optical filters.

Chapter 4 discusses base calling on the Illumina platform in depth, here it is only relevant to note that base calling on this platform is complicated by at least two effects: (1) a strong correlation of the A and C intensities as well as of the G and T intensities due to similar emission spectra of fluorophores used and their limited separation by optical filters (figure 3.2), and (2) dependence of the signal for a specific cycle on the signal of the cycles before and after, known as phasing and pre-phasing respectively. As described in section 2.3 on page 26,

IDT Exhibit 1061.0043

phasing and pre-phasing describe the loss of synchrony in the readout of the sequence copies of a cluster. Phasing is caused by incomplete removal of the 3' terminators and fluorophores as well as sequences in the cluster missing an incorporation cycle. Pre-phasing is caused by the incorporation of nucleotides without effective 3'-blocking. The proportion of sequences in each cluster which are affected by phasing and pre-phasing increases with cycle number; hampering correct base identification.

In the last three years there have been several technical updates and instead of the earlier 26 sequencing cycles up to 125 sequencing cycles are currently performed. In addition, flow cell cluster densities were increased from 5-12 million clusters to about 30-50 million clusters per lane and layer. Further, a technical update made sequencing of the reverse strand of each molecule possible. Using this "paired-end sequencing" approach for determining the reverse strand, doubles the amount of sequence data generated. The technical update enabling paired-end sequencing also allows the hybridization of further sequencing primers in one strand orientation, permitting to sequence, for example, a sample index (i.e. barcode) being part of the ligated adapter or mate pair library. An index read allows for multiple samples to be sequenced in one lane (multiplexing) [141, 150], which later can be computationally separated based on their sample-specific sequence in this separate read.

From this whole process, the Illumina user typically obtains sequences and per base quality scores. The set of sequences for each lane is typically quality filtered and the user gets a summary report for judging run quality. Finally the Illumina `CASAVA` package comes with additional tools and an interface to the visualization routines in Illumina's `Genome Studio`. Different commercial as well as free programs are available that replace some parts of the processing such as image analysis (e.g. `Swift` [242]), base calling (e.g. `AltaCyclic` [61], `BayesCall` [106], `Ibis` [116], `naïveBayesCall` [107], `Rolexa` [198]), quality assessment (e.g. `TileQC` [50] or `FastQC` [7]), mapping (e.g. `bwa` [131], `bowtie` [126], `segemehl` [95], `SOAP` [136]) as well as downstream data analysis and processing (e.g. `EULER-USR` [33], `samtools` [132], `SOAPsnp` [134], `tophat` [228], `velvet` [250]). By now there is a large community of users and developers for this platform; for example the `seqanswers.com` website[2] is an excellent resource when starting to explore the variety of programs available for analyzing the data generated.

3.1 Sequencing libraries, minimum insert size and adapter artifacts

The most important requirement for a DNA library to be sequenced on the Illumina platform is the presence of specific outer adapter sequences complementary to the oligonucleotides on the flow cell used for cluster generation, the so-called "grafting sequences". As different sequencing primers can be used, the rest of the library design is very flexible and various library preparation protocols with partially distinct adapter sequences are used for specific applications. Library adapters can be added by single strand ligation (e.g. Illumina small RNA protocol), double strand blunt-end ligation (e.g. for a multiplex protocol [150]), double strand overhang ligation (e.g. A-overhang for Illumina genomic library protocols, and restriction enzyme overhangs in the Illumina *NlaIII* DGE protocol), or by extension from overhanging primers (e.g. multiplex PCR or molecular inversion probes [177, 232]). Each of these approaches has a different susceptibility to the creation of library adapter dimers, chimeric sequences and other library artifacts. Each therefore requires a different approach to

[2] `http://seqanswers.com/wiki/SEQanswers`

IDT Exhibit 1061.0044

enrich for only those molecules with correctly added adapters, and to remove short/no insert molecules and molecules which are too long (>800nt) from the library before sequencing.

Failure to perform this enrichment during library preparation has two potential effects: (1) these artifact sequences may have a negative impact on the image analysis and base-calling which are both challenged by an over-representation of one sequence population (see below) and (2) sequencing of large numbers of such artifacts is uneconomical and lowers the potential number of informative sequences that can be generated per run. Libraries prepared from small amounts of input material tend to suffer from a higher fraction of library artifacts due to the relative abundance of adapter oligonucleotides compared to insert molecules. Computational post-processing of sequencing data where enrichment is/can not be performed is possible.

Figure 3.3 on the following page exemplifies for the Illumina *NlaIII* DGE protocol (a protocol for digital gene expression tag profiling, see chapter 5 for details) that adapter chimeras might be created which are of comparable length as the targeted library molecules and thus may not be removed by selecting a specific library insert-size (e.g. by gel length selection, silica column purification or Solid Phase Reversible Immobilization (SPRI) purification [46]). In this case, a program like `TagDust` [128] can be used with the original adapter and primer oligonucleotide sequences to identify such artifacts in a library (Figure 3.3B). This program can be either used to directly remove these sequences or, for a representative lane, its results can be clustered and the most frequent ones used with other software tools. Figure 3.4 on page 47 shows the results of clustering the most frequent sequences identified by `TagDust` for an Illumina Multiplex library. In this case, the sequencing library has been enriched for sequence similarity to mitochondrial genomes by a hybridization approach prior to sequencing. Again, different sequence reads resulting from adapter chimeras are observed. However, probably due to the enrichment and the resulting over-representation of mitochondrial k-mers, one also sees a large proportion of false positive sequences. This clearly shows that the blind application of computational filtering approaches has its limits.

Inappropriate size selection during library preparation may complicate analysis due to partial sequencing of the adaptor at the sequence ends. Thus, when selecting for insert-size, it should be considered that current experimental methods generally do not provide precise length cutoffs. The lower cutoff selected should therefore be well-above the desired sequencing length. For sequence reads where part of the adapter sequence is included, the position in the sequence read at which the adapter sequence begins has to be identified and the read trimmed appropriately. Unfortunately, this is not part of the standard Illumina data processing and also non-trivial for short adapter fragments, especially given the increasing sequencing error at the end of reads. If reads are not filtered for known chimeras and trimmed for adapter sequences, these may interfere with mapping/alignment.

In order to test how Illumina's `ELAND` mapper [16] as well as the widely used mapping program `BWA` [131] are impacted by adapter sequence at the read end, 101-cycle reads of an Illumina paired end genomic library with 10,000 reads were simulated as follows: ten thousand 350nt long sequences not containing N characters were extracted from all chromosomes and contigs of at least 1Mb in the human hg19/GRCh37 assembly. These sequences were then trimmed for the different molecule lengths and paired end reads created. For sequences below the read length of 101, the forward and the reverse read adapter sequences were added (forward: `AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG`, reverse: `AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG`) and if required extended by further A bases at the end. On the resulting 3.5 million reads, the error profile extracted from the control reads of a 2x101 cycle Illumina version 4 sequencing chemistry run was applied by randomly mutating bases at the observed rate for each position. Considering

IDT Exhibit 1061.0045

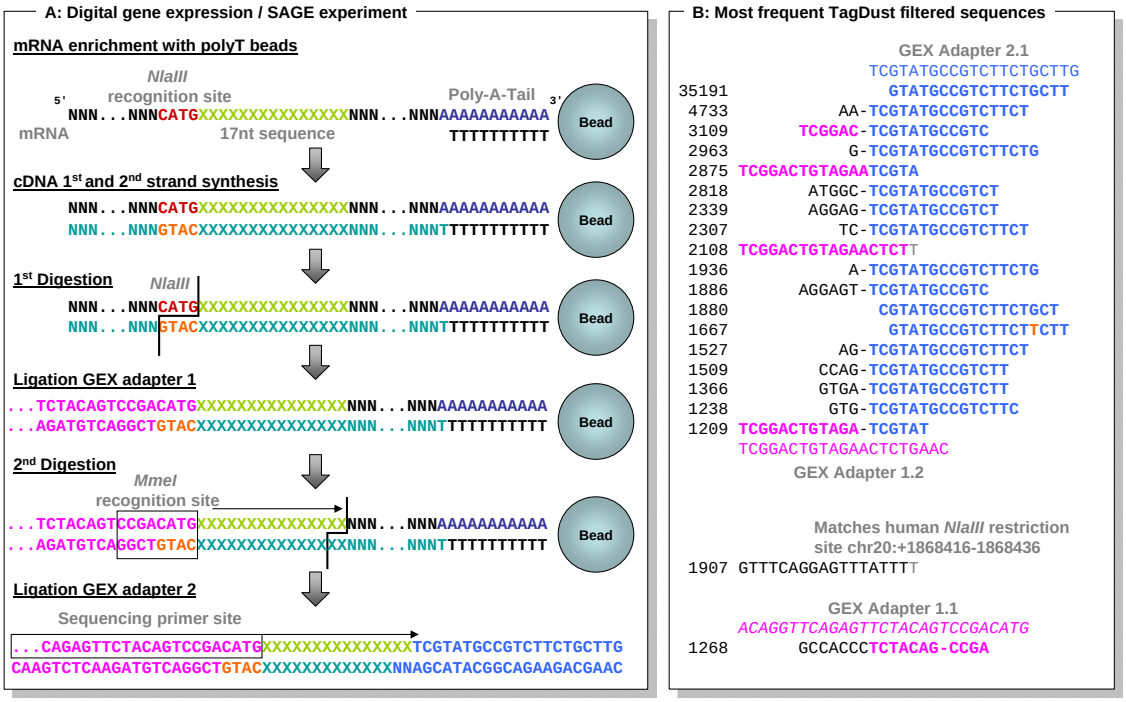

Figure 3.3: The Illumina *NlaIII* DGE tag protocol illustrated here, is a protocol for digital gene expression tag profiling based on Serial Analysis of Gene Expression (SAGE) [234]. The protocol uses short adapters which allow only single read sequencing and are added by overhang ligation **(A)**. For this protocol the majority of adapter dimers are removed by a gel excision step after library preparation. However, the protocol may also create adapter chimeras with a length comparable to the targeted library molecules. The resulting chimera sequences also show the sequences required for cluster generation as well as the necessary priming site, causing them to be sequenced together with the real tags. The program `TagDust` was used with the original adapter and primer oligonucleotide sequences to identify such artifacts **(B)**. Shown are the twenty most frequent identified artifacts from one lane with human tags as well as the oligosequences they might be based on. One of the 20 sequences seems to be a real tag that was incorrectly identified as artifact.

IDT Exhibit 1061.0046

Figure 3.4: Adapter dimer variants as well as a population of false positive sequences (italic) identified by `TagDust` for an Illumina Multiplex library **(top panel)**. Prior to sequencing, this Illumina Multiplex library was enriched for the sequence similarity to mitochondrial genomes by a hybridization approach. `TagDust` was used with either only the first 30nt **(left)** or 50nt **(right)** of the forward read. The figure shows the 30 most frequently identified artifact sequences. It is likely that high number of false positive sequences is due to biased k-mer frequencies in the read data after sequence enrichment.

IDT Exhibit 1061.0047

that both programs implement very different approaches (seed alignment versus semi-global alignment of the whole read respectively), the performance of Illumina's ELAND mapper is expected to be different from BWA. Since ELAND requires only a fixed seed in the beginning of the read (typically of 32nt length) adapters starting after this seed region should not affect ELAND's mapping.

Indeed, ELAND maps 98% of all simulated reads of at least 30nt insert size (2nt of adapter sequence being compensated by two mismatches being allowed in the seed), while BWA only reports 98% successful mappings for reads with an insert size of at least 97nt and not a single alignment for 30nt insert size (figure 3.5 on the facing page). More relevant for many analyses, however, is the number of mappings reported to be uniquely placed and whether they are mapped at the correct position in the genome. ELAND reports a uniquely placed 20nt-insert-size read, but it is placed incorrectly, as are all uniquely placed reads reported up to an insert size of 67nt. BWA reports the first three uniquely placed fragments (mapping quality above 20) for an insert size of 83nt (two of them are correctly placed). If one requires that 98% of the reads are correct placed, ELAND achieves this for insert sizes of 83nt and above (14nt of adapter), while BWA can only compensate with mismatches for 4nt of adapter sequence (97nt insert size). However, BWA provides a lower total number of false positive placements due to the inclusion of adapter sequence (8490 vs. 6308). Moreover, for an insert size of at the least read length, BWA reports 99.999% of uniquely placed reads (94.2% of all reported alignments) at the correct genomic positions, while ELAND only reports 98.757% of the uniquely placed reads (83.8% of all reported alignments) at the correct genome coordinates. BWA therefore provides a more accurate mapping of these reads for downstream analysis.

While length selection and dimer removal are important for the cost-effective sequencing of a library and downstream data analysis, experimental methods to achieve these generally consume sample material and may bias molecule representation. It is therefore often only practical to apply a minimum of these purification steps in order to maintain library quantity and complexity. In such cases, downstream sequence processing prior to data analysis becomes extremely important.

3.2 Short-insert libraries and paired-end sequencing

While adapter dimers and chimeras from library preparation should be directly removed, short insert-size molecules (e.g. ancient DNA) result in sequence reads with adapter sequence at the read end. Here, the adapter start has to be identified and the read trimmed back to the actual insert length. Unfortunately, this is non-trivial for short adapter pieces and increasing sequencing error at the end of reads. However, if reads are not trimmed for adapter sequences, these may interfere with mapping or alignment (as shown before) and reads are either excluded or placed incorrectly. In both cases downstream data analysis will be negatively affected.

When libraries containing inserts shorter than the sum of forward and reverse read cycles are created, these can be sequenced from both ends to obtain higher quality sequence information for the overlapping sequence part. For such paired-end reads the correct identification of the adapter is eased by maximizing autocorrelation of the two reads as well as requiring identical adapter start positions for both reads. Figure 3.6 on page 50 outlines the approach implemented for ancient DNA libraries, which has already been applied in different studies [81, 186, 121, 30, 27, 119]. Briefly, the two reads from a cluster are merged providing the expected adapter sequences and requiring more than 10nt overlap between the two reads. The overlap is determined by sliding the reverse complement sequence of the

IDT Exhibit 1061.0048

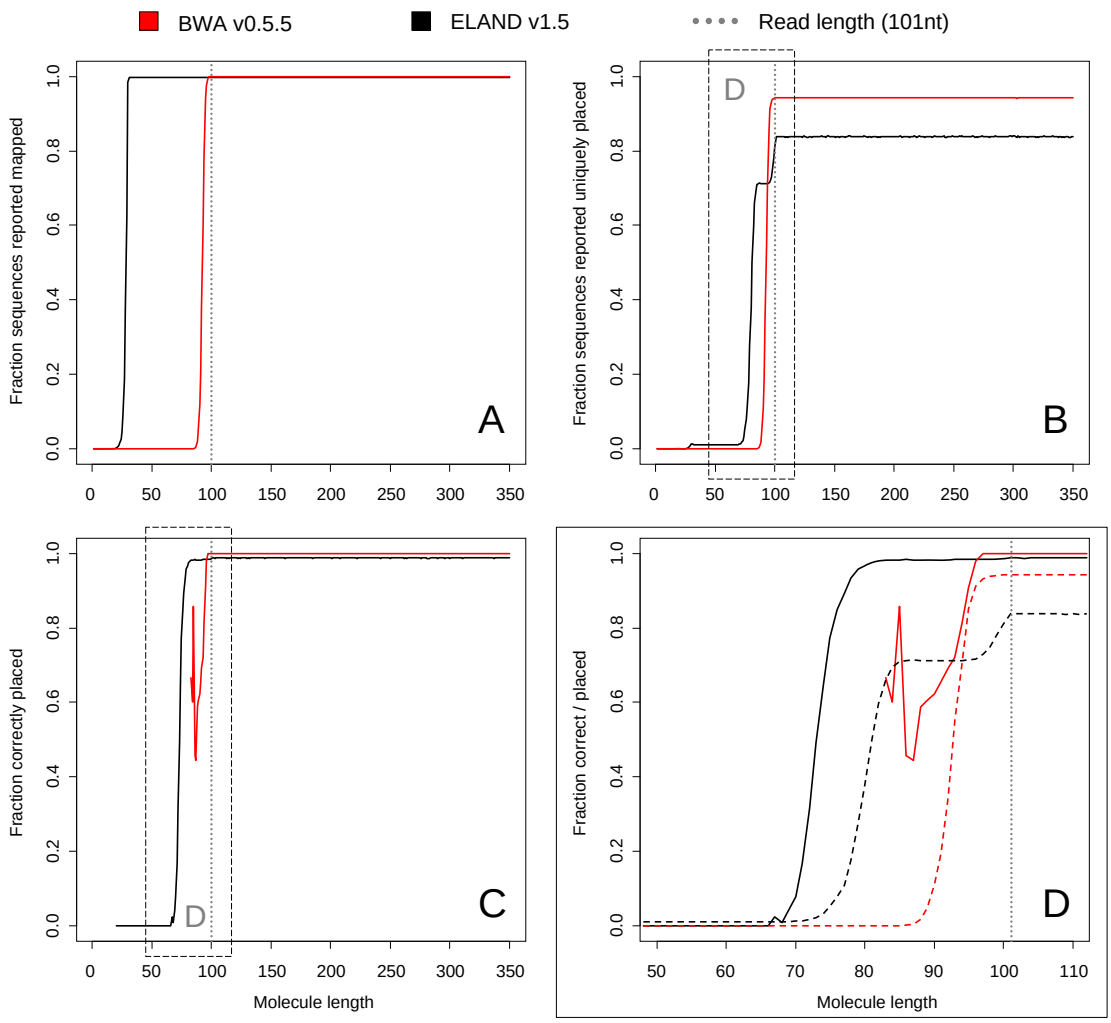

Figure 3.5: Untrimmed adapter sequence at the read ends can interfere with alignment or mapping. On a simulated data set it was tested how `ELAND` and `BWA` are affected by inclusion of adapter sequence in read mapping: **(A)** `ELAND` requires only a fixed seed (here 32nt) in the beginning of the read. Adapters beginning after this seed region may therefore have no effect on the output. `ELAND` reports 98% successful mappings for all simulated reads of at least 30nt insert size (2nt of adapter sequence being compensated by two mismatches allowed in the seed), `BWA` only reports 98% successful mappings for reads with an insert size of at least 97nt. **(B)** Considering only uniquely placed molecules, `ELAND` reports the first uniquely placed fragment for 20nt insert size. `BWA` reports the first three uniquely placed fragments (mapping quality above 20) for an insert size of 83nt. **(C)** All uniquely placed reads reported by `ELAND` up to an insert length of 67nt are placed incorrectly, as is one of the three reported by `BWA` for an insert size of 83nt. When requiring 98% correct placements, `ELAND` handles up to 14nt of adapter, while `BWA` can only compensate with mismatches for 4nt of adapter sequence. **(D)** For analysis purposes, `BWA` shows the better performance due to the lower number of false positive placements. Moreover, for an insert size of at least the read length (i.e. no adapters interfering with the alignment), `BWA` reports 99.999% of uniquely placed reads (94.2% of all reported alignments) at the designated genomic positions, while `ELAND` only reports 98.757% of the uniquely placed reads (83.8% of all reported alignments) at the correct position.

IDT Exhibit 1061.0049

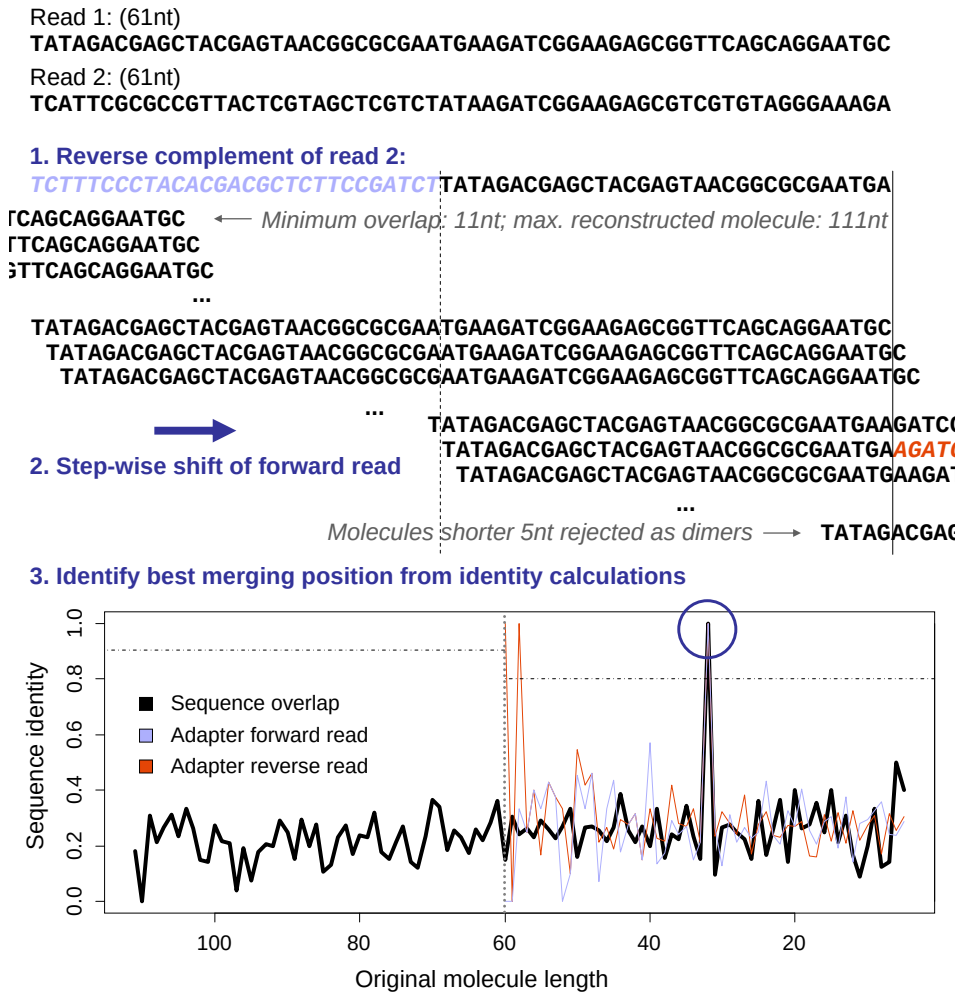

Figure 3.6: The identification of the adapter start is eased by searching for overlapping sequence of the two paired end reads and expecting identical adapter start points for both reads. The figure illustrates how the forward read is shifted along the reverse complement of the second read for identifying the original molecule length and thereby the start of the adapters. In every step, the sequence identity in the overlapping sequence part as well as the identity of the remaining adapter sequences is calculated. This is the approach applied in several ancient DNA studies [81, 186, 121, 30, 27, 119], except for two modifications: (1) Instead of the normal sequence identity, a sequence identity corrected for the observed quality scores was calculated (equation 3.1 and 3.2). (2) A heuristic was implemented by first searching the variants of decreasing length with adapter sequence present, before checking the longer variants with no adapter sequence present by increasing length. The search is aborted when a corrected sequence identity of 0.95 is observed, while otherwise the maximum sequence identity is searched as described above and read merging performed if a maximum value of least 0.9 corrected sequence identity, when no adapter is observed, or 0.8 corrected sequence identity, when at least one of the adapters was observed with 0.9 corrected sequence identity, was obtained. The actual implementation also required length of more than 10nt for the overlap and rejects inserts shorter than 5nt as adapter dimers.

IDT Exhibit 1061.0050

reverse read along the forward read and determining the quality score adjusted sequence identity (equation 3.1 and 3.2) of forward and reverse read for the different adapter start positions. Merging is performed if the highest observed sequence identity in the read overlap was at least 90%. In the overlapping sequence, quality scores are combined (assuming equal likelihood for non-observed bases, equation 3.3) and the base with the highest base quality score called. This strategy is more powerful than alignment(-like) approaches used for identifying adapter starts in single reads, which frequently remove sequence from the read ends that match the adapter by chance, or which do not identify real adapter sequence due to the higher sequencing error at the end of reads. Thus, for short insert libraries, paired end sequencing is preferable.

$$ID_{QS}(seq1, seq2) = \sum_i \begin{cases} 1 & | \; seq1_i = seq2_i \\ 1 - min\left(p(QS_{seq1,i}), p(QS_{seq2,i})\right) & | \; else \end{cases} \tag{3.1}$$

$$p_{base}\left(QS\right) = 1 - 10^{\frac{QS}{-10}} \tag{3.2}$$

$$p_{cons}(base) = \frac{p_{1,base} \cdot p_{2,base}}{\sum\limits_{n}^{\{A,C,G,T\}} \left(p_{1,n} \cdot p_{2,1}\right)} \tag{3.3}$$

Read merging performed for short-insert libraries considerably decreases the number of errors and creates sequences reflecting the original outer molecule length (e.g. of interest for authenticity of ancient DNA samples [119]). Applying the outlined merging approach to the simulated data set described above, but this time using both paired-end reads, a factor of 5 reduction in the error rate of all merged sequences is observed (average error of 0.24% reduces to 0.05%; figure 3.7 on the following page). For sequences shorter or equal to the read length a reduction by a factor of about 21 (0.146% to 0.007%) is observed. One should however caution the application of merging for long insert libraries, as false positive merging of simple repeat sequences may cause a wrong reconstruction of such regions. In this simulated data set, on average 0.29% of longer sequences (192-350nt) were incorrectly reported as merged reads.

If no paired-end data is available and adapter sequence has to be identified from single reads from a library with insert-size ranging shorter than read length, additional measures may be required. For example, it might be necessary to require at least 5nt of adapter sequence to be identified (likelihood from false adapter identification in 5nt random sequence <0.1%) and to exclude the longer sequences for all downstream analyses, guaranteeing that only a minimal fraction of erroneous adapter sequences is propagated into alignment and data analysis. Alternatively, it could be advisable to combine read trimming with the actual alignment procedure, i.e. offering that the alignment of the 3' end can be either extended to the genome or to the adapter. This way, the reference genome provides additional information for the identification of the adapter set in point. For short adapter pieces the adapter and the reference may easily result in equally good alignments. In these cases, the alignment should be terminated and a trimmed alignment reported. Using such an alignment approach and the information from a reference sequence may introduce a reference bias, an effect that will be more dominant if alignments scoring equally well for reference and adapter are assigned to the reference.

IDT Exhibit 1061.0051

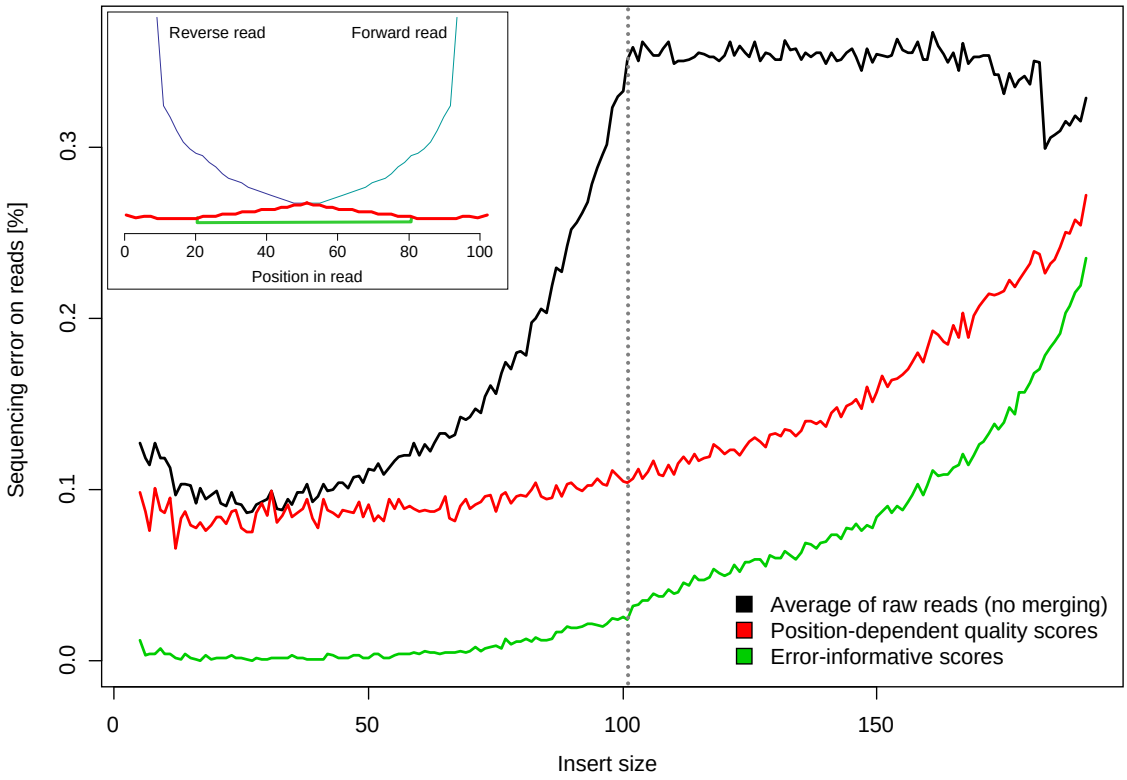

Figure 3.7: Merging of paired end reads efficiently removes adapter sequence for short insert libraries and increases read accuracy. Shown is the average sequencing error of the two simulated raw reads (black) in comparison to the sequencing error remaining after read merging for different adapter start points. The development is shown for two different types of simulated quality scores (red and green). In red, the quality score is the average error observed for the specific base-type in this cycle (i.e. all Adenines at this position in the read have the same quality score), while in green an error-informative quality score was simulated. For this type of quality score a random number between 0 and 10 (uniform sampling) was added to the average quality score of this base when the correct base was simulated and a random number between 0 and 10 (uniform sampling) was subtracted if a wrong base was simulated. The average reduction of error (starting from 0.244%) is 1.93x (0.126%) for the position-dependent quality scores and 4.98x (0.049%) for the error-informative quality scores. For sequences shorter or equal to read length (5-101nt) a reduction of error (0.146%) by a factor of 1.62x (0.090%) and 20.88x (0.007%) is observed, respectively. Sequences are required to have more than 10nt overlap for merging and merged sequences below 5nt are discarded as adapter dimers by the program.

IDT Exhibit 1061.0052

3.3 Separation of samples from multiplex experiments

With increasing sequencing capacity, multiplex experiments and sample pooling are getting more and more important for efficient use of the increasing throughput. Especially if individual loci or only small genomes are studied, sample barcoding is required as the sequencing regions of current instruments are typically too large for cost efficient sequencing of one sample per region. Sequencing results from multiplex libraries have to be computationally separated based on typically 6nt to 8nt index nucleotides that are either part of the actual sequence reads (index adjacent to insert) or performed as separate technical reads (index embedded in the adapter sequence).

On the Illumina sequencing platform, multiplex protocols [141, 151] have been established which include the index sequence embedded into the adapter sequence – separated from the actual insert (see figure 3.8A). Thus for such a typical Illumina Multiplex library, the index is determined after the forward read of the insert, in a so-called index read, for which a new sequencing primer is annealed. This decoupling of actual insert sequence read and index read, makes it possible to easily leave out the index if it is not required for the experiment, but it also results in a lower sequencing error as one of the main factors of sequencing error on the Illumina platform, namely phasing, is reset with the annealing of a new sequencing primer. In addition, with this approach image analysis and the estimation of base calling parameters are not affected by the frequently unbalanced base composition of a barcode.

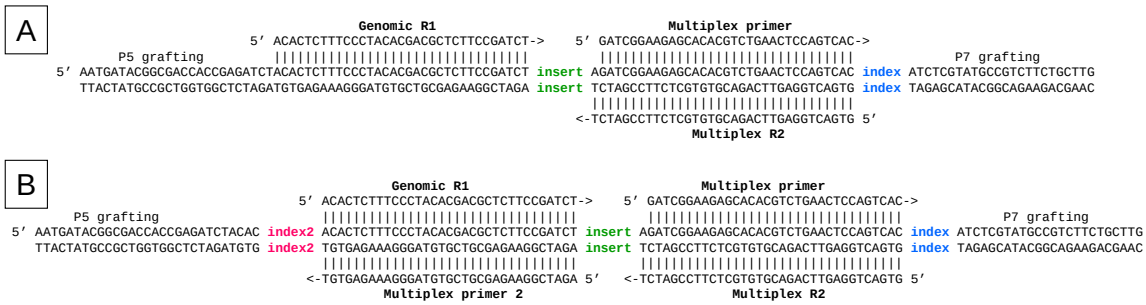

Figure 3.8: **(A)** Illumina multiplex library with grafting sequences (P5 and P7) for immobilization and amplification as well as different priming sites for forward (Genomic R1), reverse (Multiplex R2) and index read (Multiplex primer). The regular index read is performed directly after the forward read. **(B)** Modified Illumina multiplex library with an additional second index as well as another sequencing primer (Multiplex primer 2) for determining this index read after the sequencing of the template starting from Multiplex R2.

Current demultiplex approaches mostly differ in whether only exact matches from a list of used/available index sequences are identified and whether quality scores in this part of the sequencing run are evaluated. Frequently, quality scores are ignored and only perfectly matching indexes considered. Depending on the application more advanced approaches might be in place. For long barcodes the sequencing error of current instruments will cause an increasing fraction of sequences to be excluded. Theoretical considerations assuming a uniform substitution error rate of 0.5% and a 6nt index predict ≈3% erroneous sequence read outs and ≈5% erroneous read outs for 10nt. Imbalanced usage of barcodes [150] and the non-uniform distribution of errors across clusters [115, 49] will however cause higher fractions of erroneous index reads with some clusters showing close to random sequence. Ideally, complete alignments (considering quality scores in alignment scoring) to the set of indexes used should be performed. However, due to the large number of sequences that need to be processed and the computational complexity of an alignment approach, an intermediate

IDT Exhibit 1061.0053

solution is the application of a quality filter and matching to index variants with very few substitutions (stored in efficient look-up data structures and hash tables). Applying a filter to the quality scores in the index read out seems to be of special importance.

The barcodes used for multiplexing are designed to be highly distinguishable [151, 221, 141, 151], i.e. to require several sequence edits before on index read would be converted into another valid barcode sequence. With typically at least three edits between different index sequences, they are designed for accurate molecule-to-sample assignments even for high sequencing error or other errors in the index read (e.g. library amplification or synthesis errors). Assuming an independent and unbiased average error of at most 0.5% (higher than the advertised sequencing error on current versions of the Illumina instrument), three edits would correspond to at most 1.26E-07 misassignments (5 per 40 million sequences ≈5 per lane) for a 7nt index read. Such designated low misassignment rates may be required for some applications where conclusions are drawn from individual molecules (e.g. in ancient DNA or diagnostics from low quality samples), especially if positive controls or samples with high coverage are sequenced in the same pool.

However, other processes than sequencing, synthesis or amplification errors may also cause wrong assignments: (1) contamination of barcodes during synthesis or contamination of barcodes during handling, (2) overwriting of barcodes due to contamination or "jumping" in amplification, and (3) incorrect sequence read outs due to mixed intensity read outs (e.g. in the case of mixed clusters, clusters of close proximity, or any intermediate).

The contamination of barcodes during synthesis can be due to the sequential clean up of the synthesized indexing primers/adapters on the same HPLC (high purity liquid chromatography) column. Even though columns are washed between oligonucleotides, low levels of carry over are difficult to prevent. Handling contamination and overwriting in amplification steps results largely from common errors in liquid handling.

Overwriting by "jumping" however does not require any physical contamination by other indexing primers; here the overwriting sequences originate from incompletely extended amplification products or broken template molecules [166, 152, 163, 124]. Jumping was first described as a problem in the analysis of heterogeneous genetic material such as RNA viruses, multi-gene families, or repetitive sequences [152] as well as for damaged ancient DNA [152] when using polymerase chain reaction. This indicates that low amounts of DNA and the presence of similar sequences facilitate chimera formation.

The pooled amplification of libraries with different barcodes, especially after an enrichment process, is similar in several respects. Due to the library preparation protocol all molecules show common sequence parts. Further, the enrichment causes the inserts of many molecules to be similar. In addition, from an enrichment process or even from some sample sources used for library preparation, only little material is obtained (picograms to nanograms of DNA). Thus the amplification of such libraries can be considered more critical then the amplification of high quantity and complex libraries. In addition, the Illumina library design might be especially prone to the effects of jumping. While approaches having the index adjacent to the insert can only form mislabeled chimera sequences if a sufficiently similar molecule is encountered during amplification; when the index is separated from the insert by a common priming site, an incomplete extension within the priming site (figure 3.8 on the previous page, Multiplex priming site) creates an index primer template that can overwrite the index of any other library molecules without any mismatches at its 3' end.

Index misidentification due to incorrect sequence read outs from mixed intensity clusters has not previously been considered as a source of error. However loading densities on current instruments are very high (600,000-800,000 cluster per mm^2) and clusters are distributed

IDT Exhibit 1061.0054

randomly on the Illumina flow cells. Thus a considerable proportion of mixed clusters, i.e. one physical cluster entity originating from two different starting sequences, and very densely packed clusters exist on every flow cell. Considering the typical results of Illumina's default quality filter for signal purity (Pass Filter flag), an upper estimate of such impure clusters is in the range of 10-15%. For these clusters, the bridge amplification process during second read synthesis of a paired end sequencing run may also change the ratio of sequence populations and sequence mismatches at the read priming sites may change the effective ratio of the sequences during sequencing.

In order to disentangle the effect of all factors that influence the accuracy of assigning sequence reads to the samples, we[3] recently developed a new multiplex protocol with a second index read at the opposite molecule end (figure 3.8 on page 53) and analyzed the index pairs obtained under three different experimental conditions. We identified that mixed clusters contribute most of the false pairs at a rate of about 0.4-0.5% relative to sequences with error free index read outs. Applying the widely used Illumina Pass Filter flag, removed 11-30% of these false index pairs, while a filter directly applied on the quality scores of the index reads removed 92%-94% of the attributed false index pairs (while maintaining similar numbers of quality filtered sequences). The second strongest effect, PCR jumping was only observed in one of the experiments, where we pooled a subset of libraries prior to target enrichment, and hence captured, amplified and sequenced in a multiplex setup. In this specific setup, we estimated that PCR jumping caused about 0.4% chimerical molecules per correctly paired molecules after 24 amplification cycles. With rates lower than these two effects, we observed different levels of index cross-contamination from synthesis and handling (probably partially also including PCR jumping). We assigned rates of 0.04% false pairs in two experiments. For the third experiment, where libraries were enriched individually for mitochondrial genomes – the experiment involving the most sample handling –, we determined a rate of 0.10%. These results indicate that it is essential to quality filter raw clusters specifically on the index read(s), to reduce the number of false sample assignments.

3.4 Sample contamination

Another problem is sample contamination during library preparation from other DNA/RNA sources. Such contamination may be introduced by the experimenter or stem from lab chemicals. Hence, contamination from food sources, humans, parasites and bacterial contamination is frequently observed. Library preparations starting from low amounts of endogenous DNA and protocols using single strand ligation procedures can be considered most prone for contamination.

Even though there is no good way of handling contamination except for its avoidance, it has been suggested before (e.g. [42]) that reads can be filtered by the alignment to the putative contaminant sequence before data analysis. However, such a filtering may introduce biases in the data, especially if the evolutionary distance between contaminant and sample is low. This is a frequent problem in ancient DNA studies of early modern humans, Neandertals or closely-related primate species. Here the fraction of contamination is deduced from informative sites (i.e. know sites of fixed differences between species/populations) and a fraction of contaminant molecules is determined [119, 26, 81, 186]. Once the frequency of contamination in a sample is known, this information can be incorporated within statistical models during data analysis. If no informative sites are known, estimates of contamination may be obtained

[3]together with Matthias Meyer, Susanna Rankin/Sawyer, Svante Pääbo and Janet Kelso

IDT Exhibit 1061.0055

from biallelic or triallelic sites in haploid/diploid sequences. Further, for example Y chromosomal sequences in a female sample as well as heterozygous sites on the X chromosome can be used to estimate the presence of contaminant DNA in the sample [83, 84, 81, 186].

Cross-contamination of libraries before sequencing may also be a source of contamination, however this type of cross-contamination can be largely identified and filtered from the final sequencing data, when sample specific barcodes are used and determined during sequencing (see section 3.3 on page 53).

3.5 Machine adjustment and run preparation

The correct adjustment of the Genome Analyzer instrument is an important prerequisite for producing high quality sequencing data. The individual instruments as well as sequencing kits and flow cells used come with some variance. Therefore, the correct instrument adjustment should be frequently checked and preparation of a sequencing run done with much attention to detail. While liquids and optics of the instrument typically get higher attention, the correct function of other components such as thermal elements and cooling devices is equally important for high quality runs.

When loading the chemistry and flow cell, all connectors should be checked for leaks, the correct priming of the reagents validated and long waiting times evaded. This is to prevent air bubbles in the pump, tubing and finally flow cell, which could cover parts of the images (figure 3.9A) or reduce chemistry efficiency, due to smaller effective volumes and incomplete coverage of the inner flow cell surface. In case several sequencing kits are required (total read length above 36 cycles), all sequencing kits required for all sequencing reads should be thoroughly mixed before splitting them by tube volume to not introduce later problems in base calling due to different fluorophore intensities. Further the incorporation mix should be filtered and centrifuged before adding the polymerase and filling it in the final tubing. Thereby chemistry crystal and lint artifacts on the images are reduced. Otherwise, during illumination such artifacts my result in strong light signals (figure 3.9B) shining over wide parts of the image and thereby overlaying actual clusters or cause cluster like structures being mis-identified as clusters in the image analysis process (section 3.7 on page 59).

After performing the first base incorporation, the correct adjustment of the flow cell stage (flatness), flow cell tilt, the complete illumination of tiles (footprint) and oil application, the adjustment of the focus laser, the maximum focus range and stage tilt should be checked with the first cycle report and also manually before final first cycle imaging. Some of the necessary adjustments can be done directly by a skilled lab technician, if a problem is correctly identified. It is advisable to temporarily store run images (for a few weeks); in case problems with a run are observed, images still provide the most information for troubleshooting.

Problems with the stage flatness can be identified if on multiple tiles image distortions are seen, i.e. only part of the image is sharp while the rest is blurry (figure 3.9C). If this effect is limited to tiles at the flow cell edges, oil covering the flow cell surface could be a more reasonable explanation. While in the first case the adjustment by an Illumina technician is required, in the later case, the flow cell has to be removed, cleaned and reinserted into the instrument (otherwise the oil will be spread by the thermal element over the course of the run). Commonly, a band of brighter clusters is observed on the right side in the second track of lane 8 (figure 3.9D). It is probably caused by a reflection of laser light on the right flow cell edge and seems not problematic for data analysis.

Flow cell tilt is measured automatically with current software versions; if a too high value is determined, a wrong alignment of the flow cell in the instrument is a likely source. However,

IDT Exhibit 1061.0056

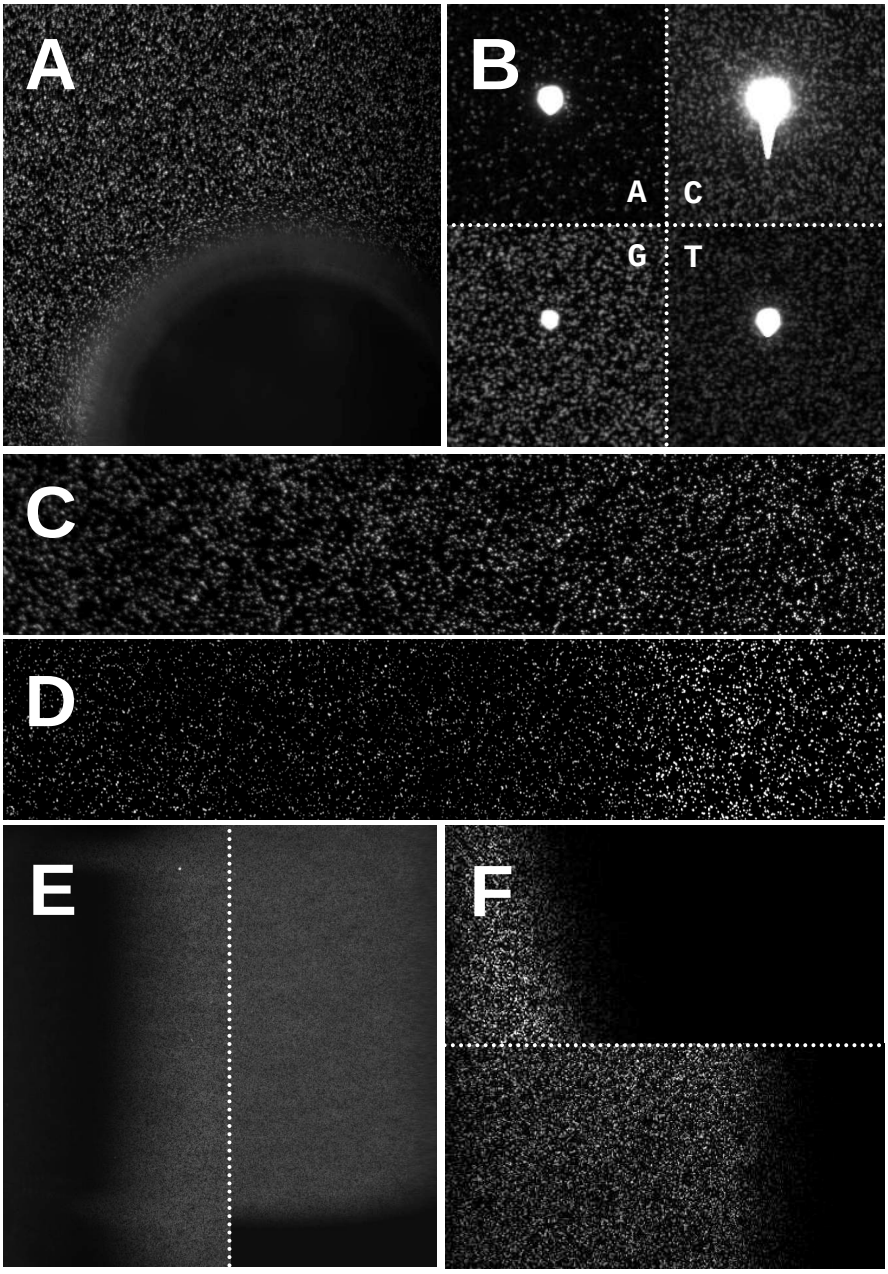

Figure 3.9: Correct instrument adjustment is an important prerequisite for producing high quality sequencing data. Preparation and start of a sequencing run has to be done with careful attention to avoid or identify the following instrumentation artifacts: **(A)** Air bubbles, caused by leaks, insufficient priming and long waiting times. Bubbles can obscure parts of the images or reduce chemistry efficiency. **(B)** Particles in the sequencing chemistry (e.g. crystals from an unfiltered incorporation mix) frequently result in image artifacts. **(C)** Incorrect adjustment of stage flatness and stage tilt can cause distortions, i.e. parts of the image are sharp while the rest is out of focus. A similar effect limited to tiles at the flow cell edges, can originate from oil covering the flow cell surface. **(D)** Reflections in the instrument can cause variation in cluster brightness, like the commonly observed band of bright clusters in column 2 of lane 8. **(E)** If the position of laser excitement is not in sync with imaging (footprint), a black straight band can be observed at the edges of multiple tiles (partially with comb like slots). **(F)** If this effect is limited to tiles at the flow cell edges, oil coverage is insufficient.

IDT Exhibit 1061.0057

an uneven outer flow cell edge may also cause this problem; in this case, the flow cell tilt has to be manually set by averaging out the uneven edge. If a black straight band is observed on one of the four edges of multiple tiles (partially comb like slots can be observed, figure 3.9E) the position of laser excitement (i.e. footprint) may not be in synchronization with imaging. In this case the laser spot can be corrected by two adjustable screws. If this effect is limited to tiles at the flow cell edges, oil coverage below the flow cell is not sufficient and additional oil has to be applied (figure 3.9F).

Focus calibration reports should be checked for continuous X and Y values. Jumps in these values are caused by confusion of the focus laser spot with its reflection and, like error messages of the spot being close to the image edge, these are the result of an incorrect focus laser adjustment. If the focus laser is not readjusted by an Illumina technician, unfocused tiles will be obtained. High values for the maximum focus range in the first cycle report (Z-axis change above 12,000-18,000) might hint at incorrect flow cell alignment causing higher bending of the flow cell. In this case, Z-values considerably decrease from the middle of the flow cell towards both ends. However if Z-values decrease or increase monotonically from top to bottom of the flow cell the stage's tilt can be adjusted by precision mechanics screws. If not readjusted, the maximum focus range will be exceeded during sequencing and unfocused tiles will be the result.

When low intensities are observed in the first cycle report (e.g. due to long handling times caused by one of the problems described before), the primer hybridization should be repeated. This step can be performed with a sodium hydroxide wash and hybridization protocol either in a Cluster Station or cBot (devices used for preparing the Illumina flow cell with the immobilized clusters; from now on only referred to as Cluster Station), or directly in the Genome Analyzer instrument, if a PE module is available and identical sequencing primers are used for all lanes.

3.6 Image analysis

The images for the four fluorophores, for the more than 100 tiles per lane and for each cycle performed, have to be overlaid (registered) and light intensities extracted for each cluster and cycle [16]. When all sequences, or a vast majority of the sequences start identically, the image analysis will consider a higher fraction of the clusters as being grown into each other and remove them – reducing the overall yield from a sequencing run by 10% to 30% depending on the loading density and software version (figure 3.10 on the next page). Such effects are, for example, observed if libraries are made from restriction digested molecules or if tag/barcode sequences are added on the outer molecule edges and these are determined in the first read cycles. Starting image analysis (`Firecrest` module) with a user-defined `nr` parameter, thereby setting the number of cycles used for cluster identification, can be used as a work-a-round. The default value of this parameter depends on the analysis pipeline version (below v1.3: 1 and not configurable; v1.3 to v1.5: 2, v1.6: 4). However, the design of project specific primers should be preferred over changing image analysis parameters, as the data transfer and offline analysis of images causes additional investments of money and time. Changing this parameter may also increase the fraction of artifacts being identified as clusters (see sequencing artifacts section). Further, when a majority of sequences are identical in the first cycles this may cause problems in base calling (see section 3.8 on page 63 and chapter 4 section 4.3 on page 71).

The optimal number of clusters per tile varies depending on the Genome Analyzer version and the library being sequenced. With the current software version (`RTA/OLB` v1.8), a complex

IDT Exhibit 1061.0058

Lane	Sample	Conc. [pM]	1st Cycle	4th Cycle	Ratio
1	SL7	1.0	6,941,328	7,238,465	96%
2	SL7	2.0	8,871,763	12,646,584	70%
3	SL6	1.0	7,675,074	8,123,010	94%
5	SL6	1.5	9,061,321	11,200,209	81%
6	SL6	2.0	9,190,638	14,167,561	65%
7	SL8	1.0	6,773,787	6,834,065	99%
8	SL8	2.0	9,003,716	11,369,083	79%
4	PhiX	2.0	11,279,178	11,478,043	98%

G channel, 1st cycle

G channel, 4th cycle

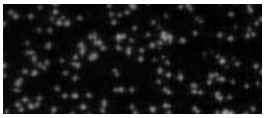

Figure 3.10: If all or the vast majority sequences start identical in the first read of a sequencing run, image analysis will consider a higher fraction of the clusters as being grown into each other and remove them. This effect is for example observed if libraries are made from restriction digested molecules or if tag/barcode sequences are added on the outer molecule edges and read in the first read. Changing parameters for an image offline analysis (`Firecrest` module) can be used as a work-a-round. The figure table shows cluster counts as well as a section of the image of the same tile in cycle 1 and 4 for a run from the Neandertal Genome project [81] (080902_BIOLAB29_Run_PE51_1) in which the tag 'GAC' was read in the beginning of the first read. Cluster counts were obtained from `IPAR` v1.01 image analysis (done only based on the first cycle of the run) and the results for a version of the `Firecrest` v1.9.5 algorithm, in which cluster identification was done in cycle 4.

library, i.e. with sufficiently many different molecules that no base composition biases is observed when averaged across them, may be loaded with 330,000 to 400,000 clusters per tile (600,000-800'000 per mm^2), while a low complexity library should be loaded in the range of 250,000 to 330,000 clusters per tile (450,000-600'000 per mm^2). Differences between low and high complexity libraries, are caused by an increased background signal (lower base qualities) and cluster tracking issues (N bases) if a majority of reads shows the same base in a cycle and therefore are imaged together. If cluster densities are reduced, the background signal from close-by clusters is reduced and thus purer intensity values obtained.

To precisely load the correct amount of library DNA for obtaining these designated cluster counts, a precise quantification by quantitative PCR (qPCR) or high-resolution chip-based capillary electrophoresis is required [183]. Hence, a stable quantification procedure is also one of the main prerequisites for performing high-quality sequencing runs.

3.7 Low quality sequences and sequencing artifacts

The random dispersion cluster generation process currently performed for the Genome Analyzer platform allows for high loading densities but also complicates the identification of cluster positions from images. Image analysis and cluster identification algorithms used for this purpose can incorrectly identify sequencing chemistry crystals, dust and lint particles as well as other flow cell features as sequence clusters (figure 3.11 on the following page).

The fraction of such artifact clusters is increased for low cluster densities (as the number of these artifacts does not necessarily increase with cluster density) and for low intensity runs. Low cluster intensities can have multiple sources: (1) reduced cluster growth during bridge amplification, (2) wide spread clusters (e.g. due to large library insert sizes), (3) inefficient sequencing primer hybridization, (4) degraded/bleached fluorophores or bad performance of

IDT Exhibit 1061.0059

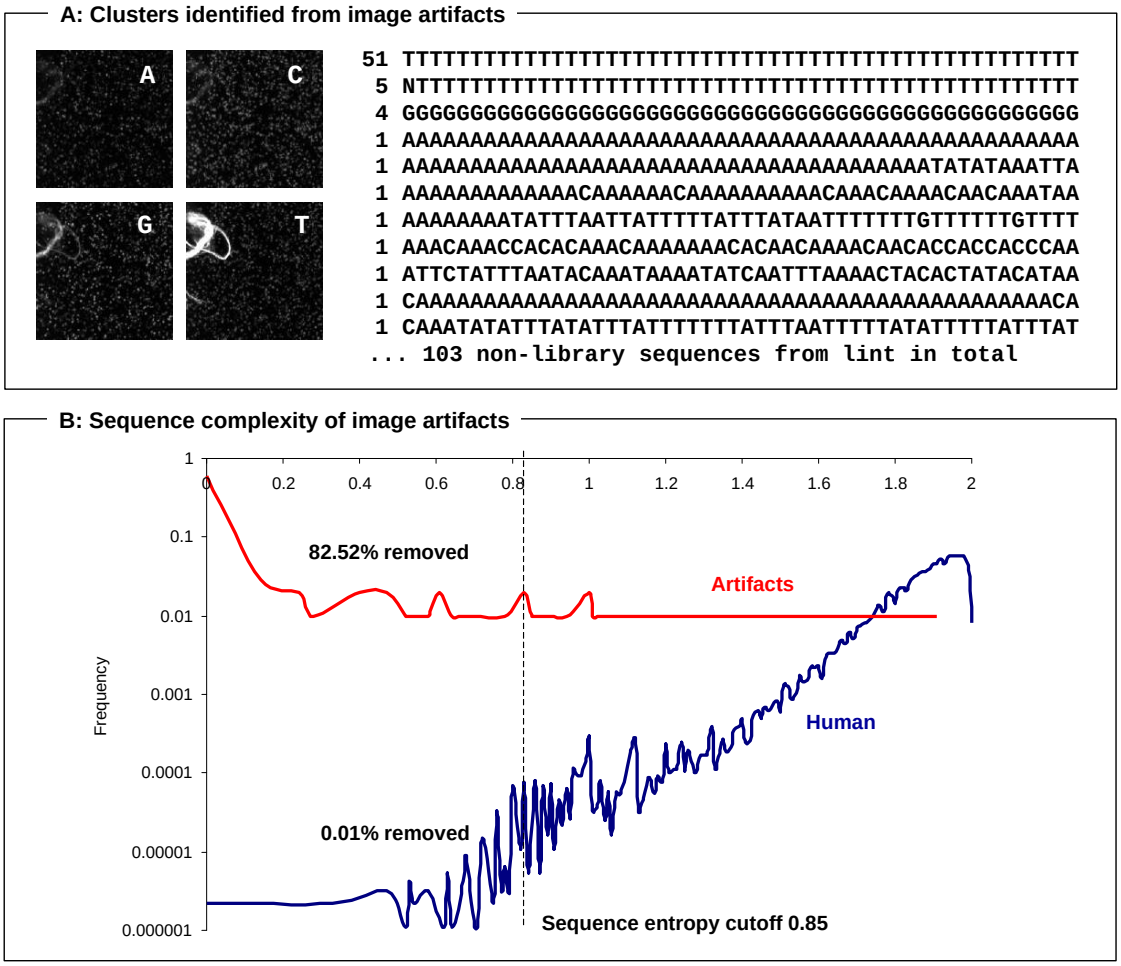

Figure 3.11: Cluster identification can identify crystals, dust and lint particles as well as other flow cell features as sequence clusters **(A)**. Indicated are 103 non-library sequences originating from a lint particle that has been observed in a library that was sequenced with a three base pair tag ('GAC') in the beginning of each read. In this case, non-library sequences could therefore be distinguished based on these first three bases. The fraction of such artifact clusters is increased for low loading density and low intensity runs. A sequence entropy filter (equation 3.4) is efficient for removing the majority of these sequences (82.52% for a cutoff of 0.85), but also removes non-artifact sequences **(B)** – as indicated in the figure, 0.01% of the human reference genome (GRCh37/hg19).

IDT Exhibit 1061.0060

Figure 3.12: Sequences resulting from crystals, dust and lint particles as well as other flow cell features are typically of low complexity (figure 3.11 on the preceding page) but only partially of low quality. Plotted is the quality score frequency distribution (`PHRED`-scale [63], `Ibis` base caller [116]) for all reads matching the 'GAC' library tag in the beginning of the read (black, n = 557,466,159 bases from 10,930,709 reads) as well as all sequences not matching the tag sequence and its one base pair substitutions (red, n = 3,481,668 bases from 68,268 reads). The data was obtained from lane 5 of the 080902_BIOLAB29_Run_PE51_1 run from the Neandertal Genome project [81].

polymerase enzymes due to production, storage or handling issues, or (5) increased background signal. If low intensities are observed for the first time, primer hybridization and first base incorporation should be repeated to exclude the most frequent sources.

If cluster identification picks up sequencing chemistry crystals, dust and lint particles, the resulting sequences are typically of low sequence complexity (i.e. consisting of long stretches of identical bases; figure 3.11 on the preceding page) and only partially also of low quality (figure 3.12). Thus they are not completely removed by signal purity/quality filters. Further, freely movable versions of these features may also appear in later cycles and overlay the signal of regular clusters. Depending on their size, these may even cover a larger fraction of a tile and thereby prohibit the correct read out of many clusters in one or several cycles (figure 3.9 on page 57). In combination with air bubbles (figure 3.9A) caused by leaks or a chemistry running low, these impermanent features are a frequent source of missing base calls (Ns) and sequencing error. The number of these fixed and movable artifacts can be reduced by a clean sequencing set-up and the above described steps. In extreme cases, the exclusion of complete tiles from analysis should be considered.

Due to the low sequence complexity obtained from the read out of most fixed image features identified as clusters, these can be removed by a sequence entropy filter (equation 3.4) or another base composition/base frequency filter. For indexed sequencing libraries [141, 150], such clusters are efficiently removed by an index sequence filter step. The same applies for libraries with tag sequences. Filtering for index and tag sequences should be considered superior, as other filters may also remove non-artifact sequences of low complexity (figure 3.11B).

$$H(x) = -\sum_i p(x_i) \cdot \log_2 p(x_i) \tag{3.4}$$

IDT Exhibit 1061.0061

Further, random cluster dispersal results in a wide range of inter-cluster distances, causing different susceptibility of clusters to neighboring signals. In the most extreme two clusters can be completely grown into each other, resulting in the read out of a mixture of signals from the different sequences. Depending on the ratio of the two sequences and the sequence similarity, the resulting sequence can be close to random with overall low base quality scores or be close to one of the original molecules and show low base quality scores/higher error rate for some positions along the read.

This effect of cluster distance on signal purity causes sequencing errors to be non-randomly distributed, i.e. the fraction of reads with two errors is not equal to the squared fraction of reads with one error - but considerably higher (e.g. figures 4.9 on page 84, 4.12 on page 89, 4.13 on page 90, 4.14 on page 91, 4.15 on page 92 as well as figure 3.13).

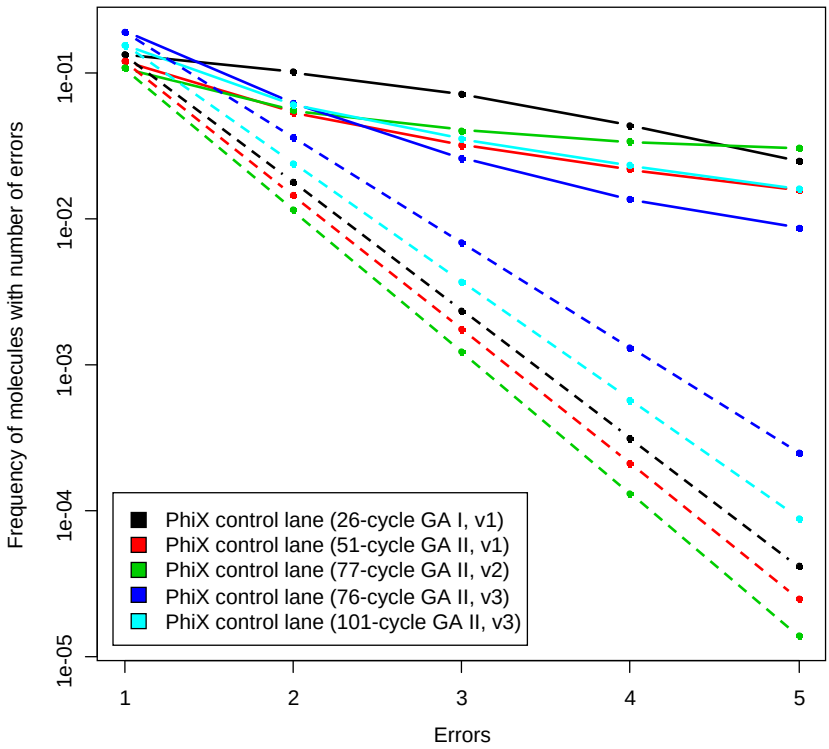

Figure 3.13: Random cluster generation results in a wide range of inter-cluster distances, causing sequencing error to be non-randomly distributed across clusters. The fraction of reads with two errors is not equal to the squared fraction of reads with one error. Shown are the observed rates for reads with 1 to 5 errors for different Illumina Genome Analyzer data sets **(solid lines)** presented as test data sets for the `Ibis` base caller (chapter 4, [116]) and the expected rates when extrapolating from the fraction of molecules with one error **(dashed line)**.

Hence, there are clusters accumulating error due to their close proximity to another sequence cluster. These can be identified by a high frequency of low quality bases. Filtering is commonly done by the default Illumina signal purity filter called "chastity", or also referred to as Pass Filter flag. This filter requires that for the first twelve cycles (in later versions of the analysis pipeline the first 25 cycles and allowing one outlier) corrected intensities for the bases called are 1.5 times higher than the next highest intensity. However, preferentially a simple quality-score-based filter should be applied over all reads and not only the first up to 25 bases of the sequencing run. A quality score filter is by design highly correlated with signal purity, but also incorporates signal strength.

IDT Exhibit 1061.0062

3.8 Sequence composition and standard base calling

The Illumina base caller uses a model-based approach for the conversion of intensity values into bases (for more details see also chapter 4 section 4.3 on page 71). The run-specific parameters of this model (so-called cross-talk matrix and phasing/pre-phasing values) are determined form the first few cycles of each read. The cross-talk matrix is typically estimated from cycle 2, phasing and pre-phasing values from the first 20 cycles. The estimate is easily confused by a library having an unbalanced base composition in this part of the read. For an extreme example of a false estimation due to an unbalanced base composition see also figure 4.1 on page 72. Such problems, are for example, caused by a restriction site or some tag sequence in either the forward or the reverse read (in case of paired-end sequencing). The only read type for which this parameter estimation is not used/not done is the index read. For the index, parameters of the preceding read are applied. Therefore for at least one lane in each run the base composition has to be balanced and the average constant for each read over the thousands of clusters per tile, or a separate control lane has to be sequenced for estimating these base calling parameters.

This control lane library is not limited to only the commonly used ϕX174 variant; however the choice should be a higher complexity shot-gun library from an organism with close to 50% GC content, to account for assumptions in the parameter estimation process [16, 116, 106]. A genomic shot-gun library from most species can be used for this purpose. While mRNA libraries are sufficiently complex to produce a balanced base composition, prepared from the standard Illumina-protocol, they show a biased base composition in the first twelve bases of the reads originating from second strand cDNA synthesis with "not-so-random" hexamers [87] and cannot be used for base calling parameter estimation.

3.9 Control read spike-in and alternative base callers

In addition to the estimation of base calling parameters, a control lane also provides useful statistics for sequencing quality. Therefore, spiking-in ϕX174 control sequences in every lane is recommended, even though a high complexity shot-gun library lane is available for parameter estimation and no control lane necessary. Given a known high quality control library, the obtained per lane statistics for this library can be easily compared, even between different runs. The choice of ϕX174 as a quality control is arbitrary; however a spiked-in sample should have the following features: (1) small genome with no similarity to the actual library sequenced for later identification/filtering by alignments, (2) a completely known sequence of the genome of the exact sample for determining error development, (3) high complexity and balanced base composition to study error patterns. If condition (1) is fulfilled, the spike-in can be performed without the need of multiplex sequencing for later separation of the control reads from other library molecules.

A fraction of 1-2% control reads is sufficient for creating quality statistics over all lanes and for using a reference-based base calling approach like `AltaCyclic` [61] and `Ibis` [116] for increasing the base calling accuracy (see chapter 4). When turning off automatic parameter estimation for the Illumina base caller and using default values from runs of comparable sequencing chemistry, this low fraction of control reads in combination with an alternative base caller allows the omission of the dedicated control lane even for whole runs of libraries with unbalanced base composition.

Frequently projects use sequence data generated on different sequencing platforms, with varying versions of the sequencing chemistry and instrument software, or data produced in

IDT Exhibit 1061.0063

different facilities. This creates a need for assuring data quality and consistency. Quality score recalibration based on alignments to a reference genome has been identified as one solution to this problem [53, 147]. Spike-in control reads provide an unbiased source for the required alignments (see chapter 4 section 4.6 on page 95).

3.10 PCR duplicates in data analysis

Some samples contain very little DNA material (e.g. DNA extracted from ancient specimens), PCR amplification of such sequencing libraries is therefore often unavoidable when using these samples in sequencing experiments. This may lead to problems in downstream applications, as non-random amplification from only few starting template molecules may limit the capacity to identify polymorphisms (or damage [28, 27]) or alter their frequency in the sample. Figure 3.14 on the facing page demonstrates the uneven representation of PCR duplicates for a deeply sequenced ancient DNA library. Quial et al. [183] described that such biases may also originate from gel purified modern DNA samples and Mamanova et al. [141] described them for PCR amplification of modern DNA samples. In the worst case, the uneven representation can lead to incorrect consensus sequence calls or false estimates for variant frequencies at polymorphic sites. The best way to deal with this problem is to identify and remove duplicate sequences that are the result of PCR amplification. However, in quantitative applications this is frequently not an option as the dynamic range of the measurement exceeds the number of possible different molecules and technical (PCR) duplicates can therefore not be separated from biological replicas.

Independent molecules are frequently identified based on their outer alignment coordinates (e.g. [132, 147, 81, 186]); however sequence-based approaches such as clustering [56, 137, 13], may also be used. When PCR duplicates of the same molecule have been identified (a PCR cluster), either a representative sequence is chosen or a consensus determined. A representative sequence should be selected based on the lowest sequencing error probability, i.e. the highest sum of quality scores. Both the `samtools` [132] and `Genome Analysis Tool Kit` [147] (`GATK`) packages implement routines for this type of filtering. A consensus sequence, which may reduce sequencing error considerably (e.g. [27, 81, 186]), should only be calculated if the identified duplicates are very unlikely to originate from different molecules.

If non-identical sequences originating from different DNA molecules are clustered together, a consensus approach will average these. This may result in incorrect haplotype calls and low quality scores for sites where variation is present. For ancient DNA samples with a few million endogenous molecules, large megabase-sized genomes and random fragment ends, the assumption of independence is probably valid. Large amounts of endogenous DNA, small genomes, or protocols that generate non-random fragment ends (such as the use of restriction enzymes) may, however, conflict with this assumption.

Identifying PCR duplicates in the whole library or the alignable fraction of the data may also be of interest for determining the complexity of a library, i.e. the total number of different molecules. The number of unique molecules u and the number of sampled molecules s can be used to estimate the total number of different sequences M. Assuming sampling with replacement and thus assuming that the above described biases have a small effect, one can fit sub-sampled values to equation 3.5 for this purpose.

$$u \approx 0 + M \cdot \left(1 - e^{-\frac{s}{M}}\right) \tag{3.5}$$

IDT Exhibit 1061.0064

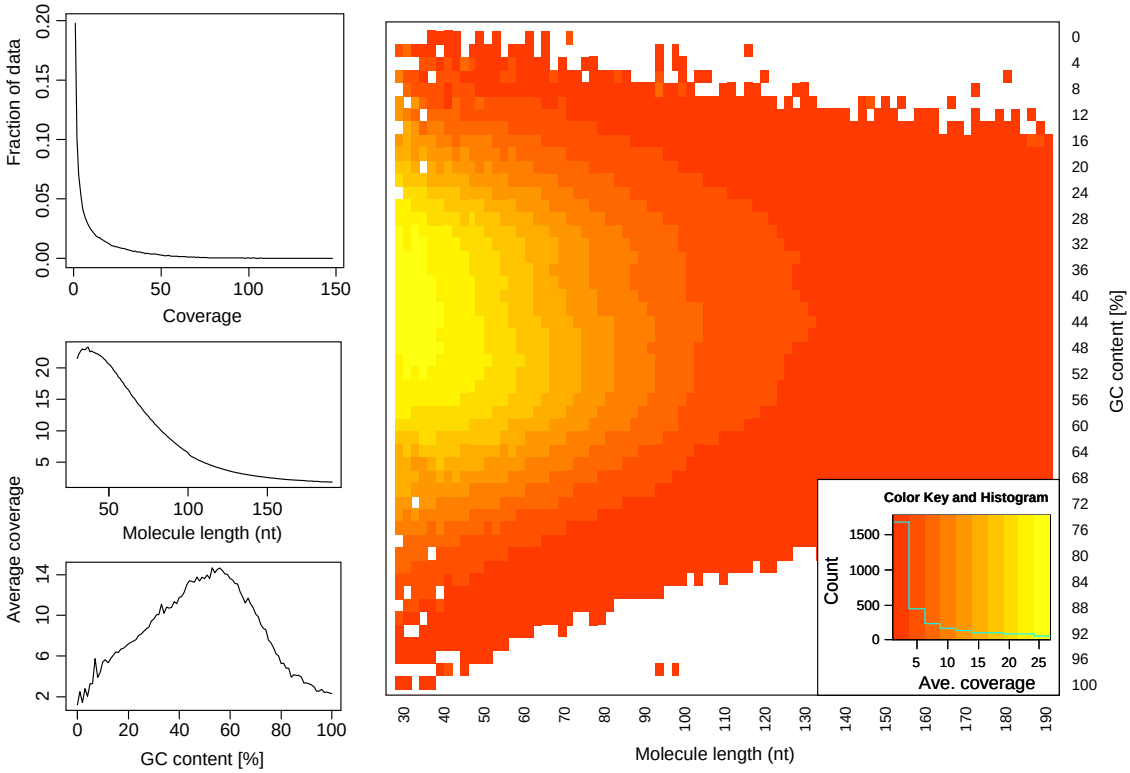

Figure 3.14: From in total 44 Illumina Genome Analyer IIx lanes of the Denisova SL3004 library ([186], 2×101+7 cycles, v4 sequencing chemistry), reads aligned with `BWA` to the human reference genome (NCBI 36/UCSC hg18) were analyzed for PCR duplicates. The SL3004 library was amplified with Phusion High-Fidelity DNA Polymerase (Finnzymes) during library preparation. Analyzing molecules aligning with the same outer coordinates, a mapping quality of at least 30 and a length of at least 30nt, resulted in an average coverage of 12.9 per PCR duplicate and an empirical coverage distribution similar to an exponential/power law distribution **(left upper panel)**. This indicates that many molecules are only observed for deeper sequencing while other molecules are available at higher frequencies. Analyzing length **(left middle panel)** and GC content **(left lower panel)** patterns as well as the combination **(right panel)** shows higher PCR duplicate counts for a GC content between 30% to 70% as well as for shorter molecules compared to longer molecules. This effect may be due to an amplification bias from the polymerase or the cluster generation process necessary for Illumina sequencing.

IDT Exhibit 1061.0065

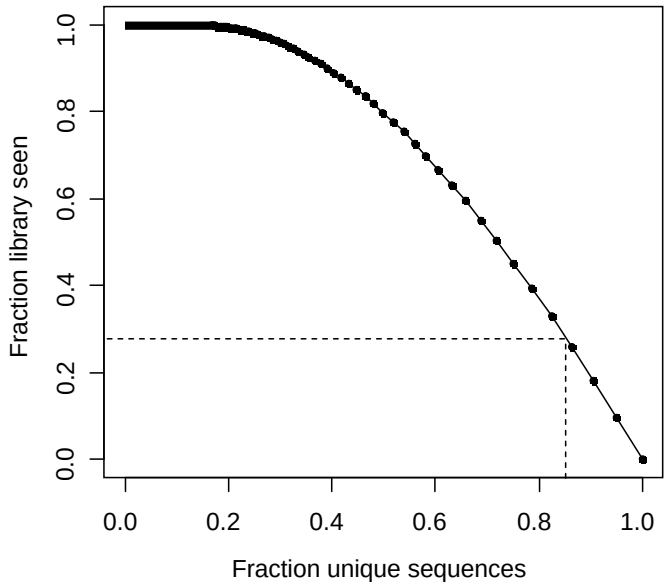

Figure 3.15: Interaction of the number of unique molecules (u) obtained from PCR duplicate analysis and the fraction of the library seen (u/M) when assuming sampling with replacement. Dots show equal amounts of sampled data (s) being added. The number of unique molecules (u) relates to the total number of different sequences (M) and the number of sampled sequences as given in equation 3.5. If one observes about 85% unique molecules, this indicates that about 28% of the total library (dashed lines) was sampled.

Such complexity estimates are of interest when determining the fraction of a library that has been sequenced. The results of such an analysis may indicate that libraries have not been sequenced deeply and that further sequencing is still efficient, e.g. when 85% of sequence reads are identified to be unique, only about 28% of the different molecules in a library have been sequenced (see figure 3.15).

3.11 Summary and conclusions

Library design for the Illumina Genome Analyzer and HiSeq platforms is very flexible due to custom priming sites and even allows applications-specific library preparation protocols. Creating libraries without artifacts like adapter dimers, chimeras and contamination from external DNA/RNA sources as well as a sufficient insert size is, however, a general aim for every library preparation protocol.

What is not widely known is that, the Illumina software does not handle artifact sequences, nor does it filter or trim adapters. Thus, some fraction of insert-adapter-chimeras or pure adapter dimers often ends up in the final data analysis or may introduce a bias when reduced insert-size reads are excluded during mapping. Explicit identification of starting adapter sequence and adapter chimeras is hampered by reads showing only a few bases of the adapter and by higher error rates at the end of reads. For paired end reads the correct identification of the adapter start is eased by maximizing autocorrelation of the two reads with the outlined read merging process, which has already been applied in different ancient DNA studies [27, 119, 121, 81, 186]. In addition to the efficient identification of adapter start points, merging performed for these short insert libraries allows for improved error rate in the called consensus

IDT Exhibit 1061.0066

sequence part (figure 3.7 on page 52, [27]). Thus, for short insert libraries, paired-end sequencing is to be preferred over single-read sequencing.

In addition to creating a high-quality sequencing library and quantifying it, the correct adjustment of the machine, handling, as well as particles in the sequencing chemistry have a considerable impact on run quality. Reflections, uneven application of oil, air bubbles and an imperfectly-adjusted machine cause varying data quality. Particles like chemistry lumps, dust and lint can cause pseudo sequence signals which then result in the analysis of artifactual reads not originating from DNA molecules in the library sequenced. Tagging or indexing allows filtering for real library molecules and should be preferred over sequence complexity-based methods. Sequence complexity-based methods provide high removal rates, but they may introduce a bias due to the removal of real low complexity sequences.

When an index/tag is placed in the beginning of the read, it may increase effective sequencing costs due to problems introduced in image data analysis and base calling. Such tags may reduce the number of clusters identified in image analysis and negatively impact base calling parameter estimation, thereby reducing the total amount of usable sequence. Correctly performed as separate reads [141, 150], the error profile of the actual reads is not altered and multiplexing allows for the optimal usage of the increasing sequencing throughput, especially if subsets of a large genome or several small genomes are studied. When using an indexed spike-in control library in all lanes of a run, the standard Illumina analysis pipeline provides useful measures to assess run quality between individual lanes and whole runs. Further, these reads allow for quality score calibration and the application of alternative base callers. Improved base callers should be considered to obtain sequences of increased quality (chapter 4).

`PHRED`-like base quality scores [63] should be used for quality-based filtering based on the complete read and specifically also on the index read(s) of multiplex experiments. Quality score based filters are equally suited for filtering clusters accumulating error due to their close proximity to another sequence cluster as the Illumina Pass Filter flag based on the first run cycles, but may also remove reads affected by freely movable artifacts in later sequencing cycles.

The most important principles presented, can be summarized as follows: (1) Filter reagents for undesirable particles and carefully start a run with checking for leaks, oil coverage and instrument adjustment, (2) Regularly check quality statistics and images for artifacts as well as the correct adjustment of the instrument, (3) Filter sequence data for library artifacts such as adapters and chimeras (4) Remove artificial clusters by filtering for sequence complexity, or if possible, filtering for specific tags/indexes used, (5) Filter low quality reads based on quality scores of all reads performed, and (6) use alternative base callers to obtain the maximum yield of high quality sequences from a run. These simple rules laid out here, shall enable the identification and handling of the most common problems in sequencing runs encountered on Illumina sequencing instruments.

IDT Exhibit 1061.0067

IDT Exhibit 1061.0068

Chapter 4

Improving base call quality of the Genome Analyzer platform

To define is to limit – Oscar Wilde [96](147-148)

As described in chapters 2 and 3, the Illumina Genome Analyzer is based on parallel, fluorescence-based readout of millions of immobilized sequences that are iteratively sequenced in a base-wise manner. After sequencing or while the sequencing run proceeds, images are analyzed and intensities extracted for each sequence cluster (see chapter 3 section 3.6 on page 58). For this purpose, the four images per tile, showing the cluster fluorophores after illuminating them with different wavelengths and filtering the emitted light with different optical filters, are scaled and overlaid (registered). Then the light intensities minus the surrounding background are extracted for each cluster and cycle. Resulting intensity files serve as input for base calling, the conversion of intensity values into bases. Base calling on the Illumina platform is complicated by at least two effects:

- A strong correlation of the A and C intensities as well as of the G and T intensities due to similar emission spectra of fluorophores and a limited separation by optical filters.
- Dependence of the signal for a specific cycle on the signal of the cycles before and after, known as phasing and pre-phasing respectively (chapter 2 section 2.2 on page 23 and section 2.3 on page 26). Phasing and pre-phasing describe the loss of synchrony in the readout of the sequence copies of a cluster. Phasing is caused by incomplete removal of the 3' terminators and fluorophores as well as sequences in the cluster missing an incorporation cycle. Pre-phasing is caused by the incorporation of nucleotides without effective 3'-blocking. The fraction of molecules in cluster affected by phasing and pre-phasing increases with the number of cycles, hampering correct base identification as the run proceeds [61, 198, 116, 106].

Technical improvements in the optical filters and camera of the Genome Analyzer II/IIx, have helped with distinguishing the A and C as well as G and T fluorophores. Phasing and pre-phasing were addressed by continuous improvement of the sequencing chemistry kits that became available over the last three years (currently the fifth chemistry version is used). Both improvements reduced the overall error rate and allow more sequencing cycles.

IDT Exhibit 1061.0069

4.1 Improved base calling and quality scoring

During the first year after the release of the Genome Analyzer I platform, two publications [61, 198] addressed the base calling of the platform; both using statistical learners trained on sequences called by the standard base caller, `Bustard`. Erlich et al. [61] published `AltaCyclic` – the first machine-learning based approach to base calling for the Genome Analyzer. Their approach applies Support Vector Machines (SVM) trained for each individual cycle. `Rolexa` [198], a base caller for the statistical software package R [224], applies Gaussian mixture models, similar to the approach used by Cokus et al. [38] for the analysis of bisulphite sequencing data. The two base callers differ further in that `Rolexa` generates IUPAC ambiguity codes[1] for ambiguous base calls, while `AltaCyclic` produces unambiguous bases with quality scores.

End of 2008, we[2] started developing `Ibis` (`Improved base identification system`), an accurate, fast and easy-to-use base caller for the Illumina sequencing system. We aimed to significantly reduce the error rate, to provide better quality scores with each base, as well as to provide a more generalized and computationally lighter approach than the ones presented before. Our developments converged in the publication and release of `Ibis` in Genome Biology end of summer 2009 [116]. Its improvements allow increased output of usable reads due to reduced error rates and facilitate better quality filtering of the data, sequence read mapping, *de novo* assembly and downstream data analysis like SNP calling due to calibrated `PHRED`-like [63] quality scores.

With the publication of `Ibis` in Genome Biology [116], another base caller called `BayesCall` was published in Genome Research [106]. Unlike `Ibis`, `AltaCyclic` and `Rolexa`, this base caller does not use statistical learners, instead a more complex model for the base calling process is described and model parameters fitted from the sequencing data. The full model of `BayesCall` is computationally too expensive and only the, in 2010 published, simplified model seems useful for a wider application of this approach [107, 129]. As outlined in a recent base calling review [129], `naïveBayesCall` and `Ibis` are the only two actively maintained alternative base callers for the Illumina sequencing platform and both outperform the vendor base caller about equally well. The latter result was also confirmed by an internal comparison done right after publication of the two base callers (Wei-Chun Kao personal communication, September 9th 2009). `Ibis` is constantly adapted to changes of the Illumina pipeline and updated versions are available on the project website `http://bioinf.eva.mpg.de/ibis`.

4.2 Input for base calling (intensity files)

As briefly described above and with more detail in chapter 3 section 3.6 on page 58, the Genome Analyzer instrument takes four images per tile[3] and cycle during a sequencing run. Depending on the exact instrument and software version, the `Firecrest` program of the Genome Analyzer `Analysis Pipeline`, the `IPAR` (`Integrated Primary Analysis and Reporting`) machine, the `OLB` (`Off-Line Basecaller`) or the `RTA` (`Real Time Analysis`)

[1] The Nomenclature Committee of the International Union of Pure and Applied Chemistry (IUPAC) has assigned one letter codes to express the possibility of different combinations of bases for a specific DNA sequence position. A table of this assignment is for example available from `http://en.wikipedia.org/wiki/Nucleotide`

[2] together with Udo Stenzel and Janet Kelso

[3] Actual numbers vary for various technical and practical reasons. Typical are 330 tiles on Genome Analyzer I, 100 tiles on Genome Analyzer II, 120 on tiles Genome Analyzer IIx and 32 on tiles HiSeq for each lane of the flowcell.

IDT Exhibit 1061.0070

software registers the four images, which are slightly scaled and shifted due to the different optical filters used, and identifies the clusters in the images of each tile. The images of one tile are then further registered between cycles and the intensity values extracted from the four images for each of the clusters identified. In addition to the intensity value (measured as the maximum brightness of the central cluster pixels), also the mean intensity value of the surrounding pixels (called noise) is determined. The difference of these two values is stored as the actual intensity value. This results in four floating point numbers per clusters and cycle. A cluster is identified by the quadruple of lane number, tile number and x-y coordinates of the cluster in the superimposed reference image.

Depending on the image analysis software (`Firecrest`, `IPAR`, `RTA` or `OLB`) the created output files vary, but otherwise provide the same input for the base calling process. The original `Firecrest` format (nowadays refered to as legacy format by Illumina) stores this information as one `GZip`-compressed tab-separated text file per tile with the first four columns being lane number, tile number and x-y coordinates and then having one column per cycle. Within each cycle column the four intensity values are separated by space characters.

For `IPAR/OLB` output, the cluster coordinates are separately stored in one text file and the intensity values are stored in another `GZip`-compressed text file per tile. The latter text file contains one line per cluster and cycle, i.e. intensity values for one cluster and cycle are printed as space-separated text in one line and cycles are separated by a comment line (starting with a # character). `RTA` also uses the separate file with cluster coordinates, but stores intensities in a binary format, called `Cluster Intensity Format` (`CIF`), with one file per cycle and tile. This binary format contains a 13 byte header: the characters "CIF" as the first three bytes, followed by one byte for a version number, one byte for the block size, two bytes for the cycle number, another two bytes for the number of cycles stored in the file (currently 1), and 4 bytes for the number of clusters stored in the file. The file header is followed by four blocks (A,C,G,T), of size *block size* bytes times *number of clusters*, containing the actual intensity float values for one of the four bases.

4.3 Illumina standard base caller

As shown previously by Erlich et al. [61] and Rougemont et al. [198], the intensities of each two channels of the Illumina instrument are highly correlated due to the similar fluorophores used for A, C and G, T. In order to separate these channels and normalize their individual intensities, Illumina's `Bustard` base caller transforms the raw intensity values using a so-called cross-talk matrix estimated from the second imaging cycle (first cycle in earlier software versions). This estimate is based on the assumption that the four nucleotides are almost equally frequent at this sequence position in the library being sequenced. If this assumption is violated, the inaccurate estimates can lead to incorrect base calling.

Figure 4.1 on the next page shows an extreme case example of a wrong estimate. Here the vast majority of clusters incorporated a cytosine in the cycle where the estimate was done. This caused the software to scale up intensity of all other bases (diagonal values of the matrix) and to see a big contribution of the A/C signal to the G/C signal, which should be close to zero otherwise. To prevent such an incorrect estimation due an unbalanced base composition of the sequencing library, the cross-talk matrix is commonly estimated using a control lane in which a variant of ϕX174 (GC content of 44.7%) is sequenced (see also section 3.9 on page 63).

`Bustard` also estimates phasing (the fraction of cluster molecules falling one sequence position back in one base incorporation step) and pre-phasing (the fraction of cluster molecules

IDT Exhibit 1061.0071

Estimate for sequences showing mainly Cs (~99%)

	A	C	G	T
A	1.31	0.76	7.26	23.49
C	1.00	0.92	5.31	19.85
G	0.02	0.02	1.68	1.08
T	0.03	0.02	0.77	2.09

0.50%

Estimate PhiX control lane

	A	C	G	T
A	1.16	0.52	0.02	0.03
C	0.92	1.09	0.03	0.03
G	0.03	0.06	1.59	0.06
T	-0.02	0.01	0.60	0.84

84.05%

of raw sequences mapped to by PhiX genome

Figure 4.1: Cross-talk matrix estimates of the Illumina `Bustard` base caller, here done for cycle 1 (left) and cycle 4 (right) of the same lane and then applied for base calling. The lane contained ϕX control reads with a tag sequence (CAG) in the first three cycles of the run. Thus in the first cycle, most clusters incorporated a cytosine. When applying the obtained cross-talk matrix for base calling, only 0.5% of sequences aligned to the ϕX174 reference genome. The matrix on the right was obtained from the fourth cycle, one cycle after the tag ended, considering actual ϕX bases. When this matrix is applied, 84.05% of sequences align. The "correct" matrix also indicates that the separation within the A/C laser channel is worse than within the G/T laser channel: C signal (0.52) has to be removed from the A signal, while almost no T signal (0.06) has to be removed from the G signal.

running one cycle ahead in one base incorporation step) as two channel-independent parameters from the increasing correlation of intensities in the first few cycles of the sequencing run. Using the cross-talk matrix and the two phasing parameters, `Bustard` first creates corrected intensity values and then calls the base with the highest corrected intensity for each cluster and cycle. In the case of equal intensity values or small intensity differences an 'N' is called. Further, a trust value, i.e. a base quality score, is assigned to each base call.

The `Bustard` base calling process described here is based on two additional, implicit assumptions: first, that the cross-talk matrix can be considered constant over the run, and second, that phasing affects all nucleotides in the same way. Erlich et al. [61] have previously shown that this first assumption is violated. Another argument for this is the commonly observed decrease in intensities over the course of the run (see figure 4.2 on the next page). This is likely to be a result of degradation of the fluorophores, the effect of a decreasing number of sequences being elongated in each cluster when nucleotides for which the termination cannot be removed are incorporated (as also suggested by Erlich et al. [61]) or cluster molecules being lost due to degradation events. Further, phasing does not affect all nucleotides equally. With the chemistries FC-104-100x or FC-204-20xx (sequencing chemistry versions 1 and 2), the fluorophores used for thymine show a lower removal rate after treatment with TCEP (tris-(2-carboxyethyl)-phosphine) [16] and accumulate over the sequencing run (T accumulation, see figure 4.2 on the facing page and section 4.4.1 on page 75).

The effects of cross-talk, declining intensities, pre-phasing and phasing, as well as T accumulation complicate the identification of the correct base, especially in later sequencing cycles. When mapping raw reads of ϕX174 RF1 sequenced with 51 cycles, 79.4% map to the corresponding reference genome allowing up to five mismatches. Only 39.8% map without any mismatches. Analyzing the different types of mismatches, a non-random distribution (figure 4.3A on page 74) is observed. Starting around cycle 25, guanine is increasingly confused with thymine (illuminated using the same laser); in later cycles adenine and cytosine show also a high rate of erroneous thymine calls due to increasing T accumulation. The error rate of the first base is especially high due to the higher handling time when starting

IDT Exhibit 1061.0072

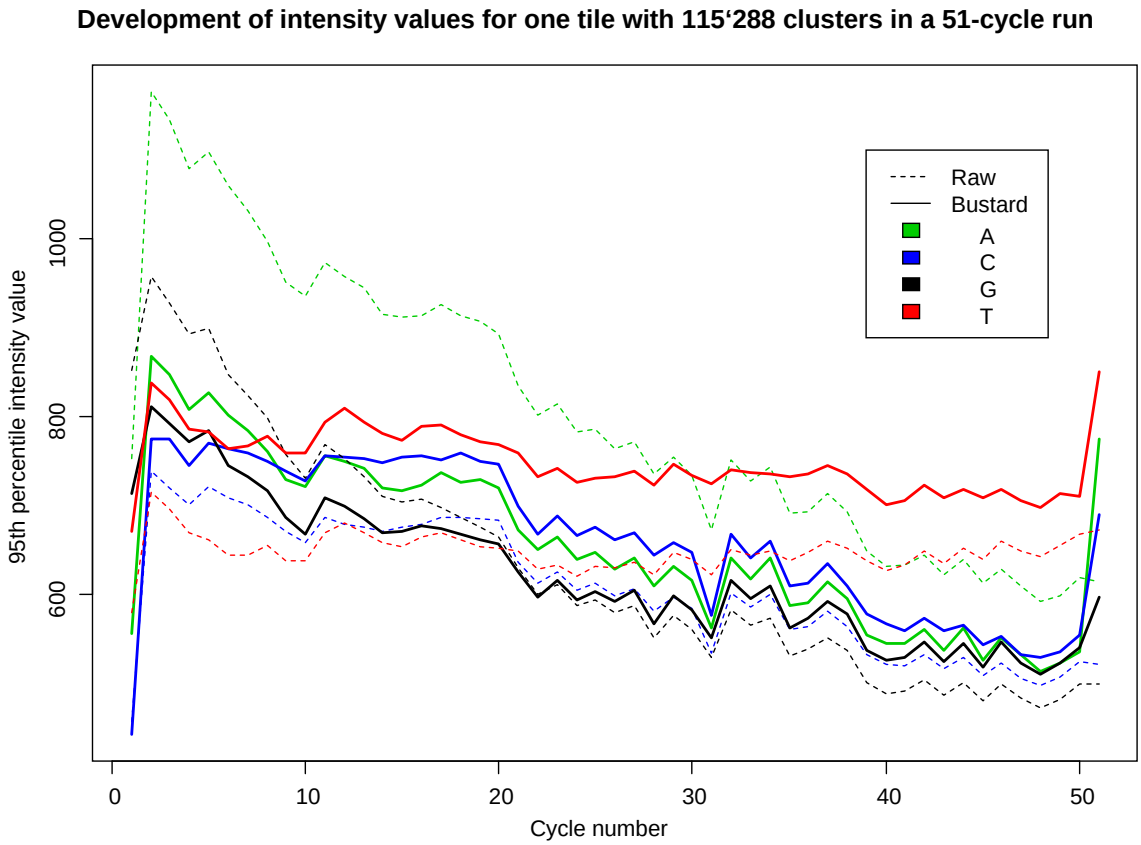

Figure 4.2: Intensity values for one tile of a ϕX174 RF1 lane from a 51-cycle Genome Analyzer II run before and after correction by `Bustard`. On this tile 115'288 clusters were identified by the image analysis software `Firecrest`. Shown are the 95th percentile for the signal intensities in each channel and cycle. The raw intensities are shown with dashed lines, the intensities after transformation by `Bustard` are shown with solid lines. Intensities for A, C, and G decline over the run while the intensities for T stay nearly constant. Both effects can be explained by degradation of the fluorophores or non-reversible termination/degradation of sequences over the run as well as the accumulation of T fluorophores on the synthesized strand. Intensities for the first cycle are lower than in other cycles due to dimming and bleaching caused by longer handling times before imaging of the first cycle. Corrected intensities for the last and first cycle do not follow the normal trend, since full phasing correction cannot be applied.

IDT Exhibit 1061.0073

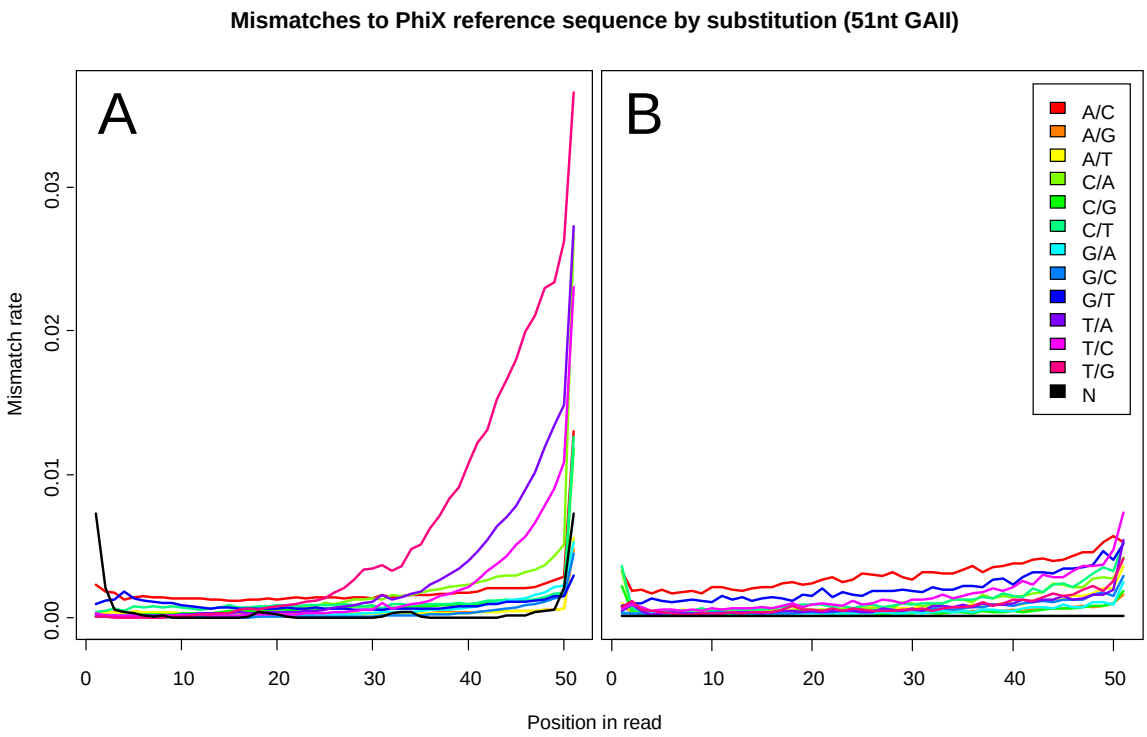

Figure 4.3: Analysis of mismatches seen for `Bustard` raw reads **(A)** and `Ibis` raw reads **(B)** of a lane with 11,478,043 ϕX174 RF1 raw reads sequenced with 51 cycles and mapped with `SOAP` [135] to the corresponding reference genome allowing up to five mismatches (including 'N' characters). For `Bustard` 9,110,666 (79.4%) of raw reads can be mapped, for `Ibis` 9,695,354 (84.5%) of raw reads. The sequencing error, measured as the mismatch rate, increases with cycle number. For `Bustard`, starting around cycle 25, guanine is mistaken as thymine. In later cycles adenine and cytosine are also mistaken as thymine, due to increasing T accumulation. The error rate of the last base is especially high due to incomplete phasing correction. The patterns of T accumulation are not observed when `Ibis` is used for base calling this lane **(B)**.

the sequencing run (e.g. focusing and first cycle report, see also chapter 3 section 3.5 on page 56); the last base is especially high due to the inability to correct phasing without a successive cycle read out.

4.4 Direct application of a statistical learner

When designing a base caller which can cope with the cycle-dependent problems discussed above, we considered constructing a more complex model of the sequencing chemistry than available in `Bustard` – including T accumulation, declining intensities and the specific characteristics of the first and last cycle. All previously available base callers and the `BayesCall` program developed in parallel by Kao et al. [106] followed this general approach, although the complexity of the model and the modeled parameters differs.

However, increasing model complexity has two major disadvantages. First, building a correct model for the Illumina sequencing platform requires a deep understanding of the causes for sequencing error and thus a model is likely to be incomplete. Secondly, a sufficiently complex model will depend on the chemistry or platform version used and has to be adjusted when either one changes. We instead chose to estimate the sequencing chemistry model as a parameter directly from the data using statistical learners and a training data set derived

IDT Exhibit 1061.0074

from the `Bustard` output. Even though our approach may not provide interpretable model parameters values, which measure for example the efficiency of specific sequencing chemistry steps and may thereby inspire further experimental improvements or serve as sequencing run quality measures, it is the most general and flexible approach, which is of advantage when considering the vast improvements of sequencing chemistry and instrument over the last years.

Statistical learners [90], also called machine learning approaches, describe a wide range of mathematical models and algorithms used to extract patterns and rules from huge data sets. In general, statistical learning can facilitate a better understanding of the basics underlying data or can be applied for predicting both qualitative (i.e. discrete labels) and quantitative descriptors (i.e. values out of a continuous range) from data. In the context of statistical learning, base calling can be seen as predicting discrete labels, finding the correct nucleotide label given the intensity values observed for a specific cycle (i.e. a four-class classification problem).

Previous approaches [61, 198] corrected raw intensities prior to the application of the statistical learner and used only the intensities of one cycle as input. This causes these approaches to be highly dependent on a correct modeling, or at least very good modeling, of the sequencing process to obtain the corrected intensities. This problem is bypassed by directly basing training on the raw cluster intensities. In this case, the statistical learner has to be provided with the intensities of multiple cycles to incorporate the effects of phasing and pre-phasing.

4.4.1 Simulating phasing, pre-phasing and T accumulation

To identify the correct number of cycles as input for the statistical learner, clusters of a thousand identical sequences and the fluorophore attachment over several sequencing cycles were simulated for 10,000 clusters. The 10,000 sequences for these different clusters were created randomly using the GC content of ϕX (44.7%) as a guide. The sequencing model, with pre-phasing and phasing as described above, can be expressed with the following recurrence:

$$Basecase:$$

$$f_0(0) = 100\%, f_{p\neq 0}(0) = 0\%, f_0(c \neq 0) = 0\%$$

$$Recurrence:$$

$$\begin{aligned} f_{p\in\{1,\ldots,l\}}(c) &= f_{p-1}(c-1) \\ &\quad + (f_{p-2}(c-1) - f_{p-1}(c-1)) \cdot p_{pre} \\ &\quad + (f_p(c-1) - f_{p-1}(c-1)) \cdot p \end{aligned}$$

For phasing and pre-phasing parameters in this model, the values reported by `Bustard` (available in the params `XML` files of the `Bustard` subfolder of a run) for the last ten runs done on the MPI Genome Analyzer II sequencers at the time were checked (FC-204-20xx chemistry, see table 4.1 on the next page). While some runs showed balanced phasing and pre-phasing values, others showed higher phasing values. I picked the mean of the pre-phasing values and simulated symmetrical phasing/pre-phasing with a rate of 0.4% per cycle.

It was also of interest whether fluorophore accumulation impacts this model. Thus, I estimated fluorophore accumulation from the same 51 cycle run also presented in figures 4.2 and 4.3 above. From one tile with 115,288 clusters, the 5th percentile of raw intensities in

IDT Exhibit 1061.0075

Table 4.1: Phasing and pre-phasing values determined for ten different Genome Analyzer II runs (FC-204-20xx chemistry) done at the MPI for Evolutionary Anthropology in late 2008. Values have been extracted from the `params.xml` files of the `Bustard` subfolder of each run.

Run	Phasing	Pre-phasing
1	0.0072	0.0049
2	0.0053	0.0053
3	0.0039	0.0038
4	0.0039	0.0038
5	0.0044	0.0035
6	0.0043	0.0042
7	0.0062	0.0043
8	0.0070	0.0033
9	0.0048	0.0030
10	0.0052	0.0036
Mean	0.0052	0.0040
Std deviation	0.0012	0.0007
Median	0.0050	0.0038

each of the 51 cycles was extracted. Due to the effects of bleaching and dimming in the first cycle (as a result of longer handling times), the first cycle was excluded from the analysis. Afterwards, values for each channel (A, C, G and T) were normalized separately, so that the 5th percentile raw intensity value of the second cycle is 100 in each of the four channels (figure 4.4 on the facing page).

As phasing and pre-phasing increases, the background noise in each cycle increases. Thus an exponential increase of the background noise measured with the 5th percentile is expected. Fitting a logarithmic function ($noise = a \cdot \log(cycle) + n$; inverse of the exponential function) to the values extracted, a very good fit for A and C was obtained, but G and T seemed to miss another linear factor: the described T accumulation. I therefore fitted $noise = a \cdot \log(cycle) + b \cdot cycle + n$, resulting in the fits shown in figure 4.4 on the next page. With the linear factor the function fits G and T better, for A and C there is no significant increase with respect to the Akaike information criterion (AIC, [90]). Due to cross-talk the linear effect observed for G may be completely caused by the accumulation of T fluorophores. Together with the observation of the base substitution patterns seen for `Bustard` (figure 4.3A on page 74), I therefore only model fluorophore accumulation of Ts, with a rate estimated from the fit as 3.8% per cycle.

For each cycle in the simulation, the number of fluorophores attached to all sequences of the cluster was determined and the fraction of fluorophores representing the current cycle, the previous cycle and the next cycle, as well as representing cycles more than one ahead and more than one behind calculated. Furthermore, the number fluorophores attached due to T accumulation was calculated. A visual representation of the simulation results is available in figure 4.5 on page 78, a textual representation is available in table 1 on page 197.

From the simulation results, it turns out that for a read length of 50 cycles 59.5% of the fluorophores reflect the current cycle, 17.4% are exactly one cycle behind and the same fraction is one cycle ahead. In a sequencing chemistry with T accumulation, on average 339 T fluorophores stuck to the cluster by cycle 50. In cycle 100, 52.4% of the clusters are in sync and 19.5% are exactly one cycle behind or one cycle ahead. At this point, on

IDT Exhibit 1061.0076

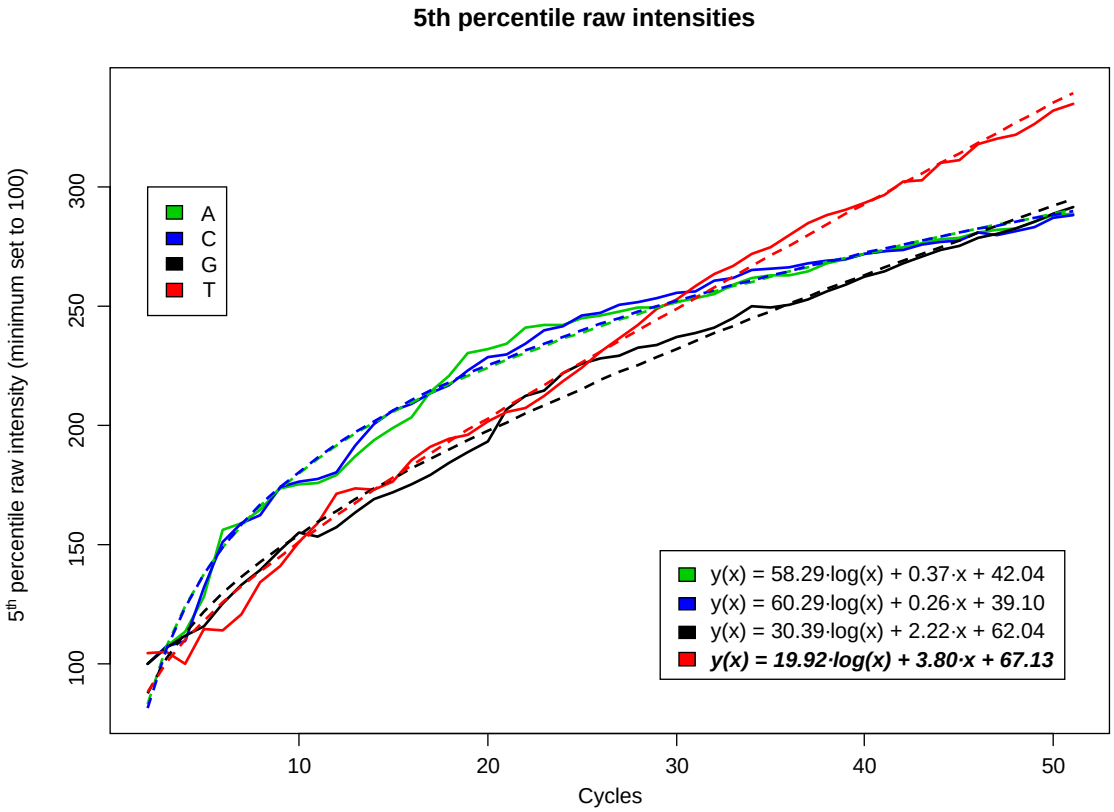

Figure 4.4: Development of the 5th percentile of the raw intensities over 50 cycles. Due to the bleaching/dimming effect of the first cycle, the first of 51 cycles has been excluded. The remaining values were normalized within each channel (A, C, G and T) separately, so that the 5th percentile raw intensity of the second cycle is 100. A function consisting of a linear (describing the accumulation of fluorophores) and a logarithmic part (describing noise increase due to phasing) was fitted. For A and C the linear part does not significantly increase the quality of the fit (AIC, [90]). The linear effect is strongest in the T channel and accounts for an increase of 3.8% per cycle; the effect in the G channel is probably also caused by the T channel - due to the read out with the same laser (cross-talk).

IDT Exhibit 1061.0077

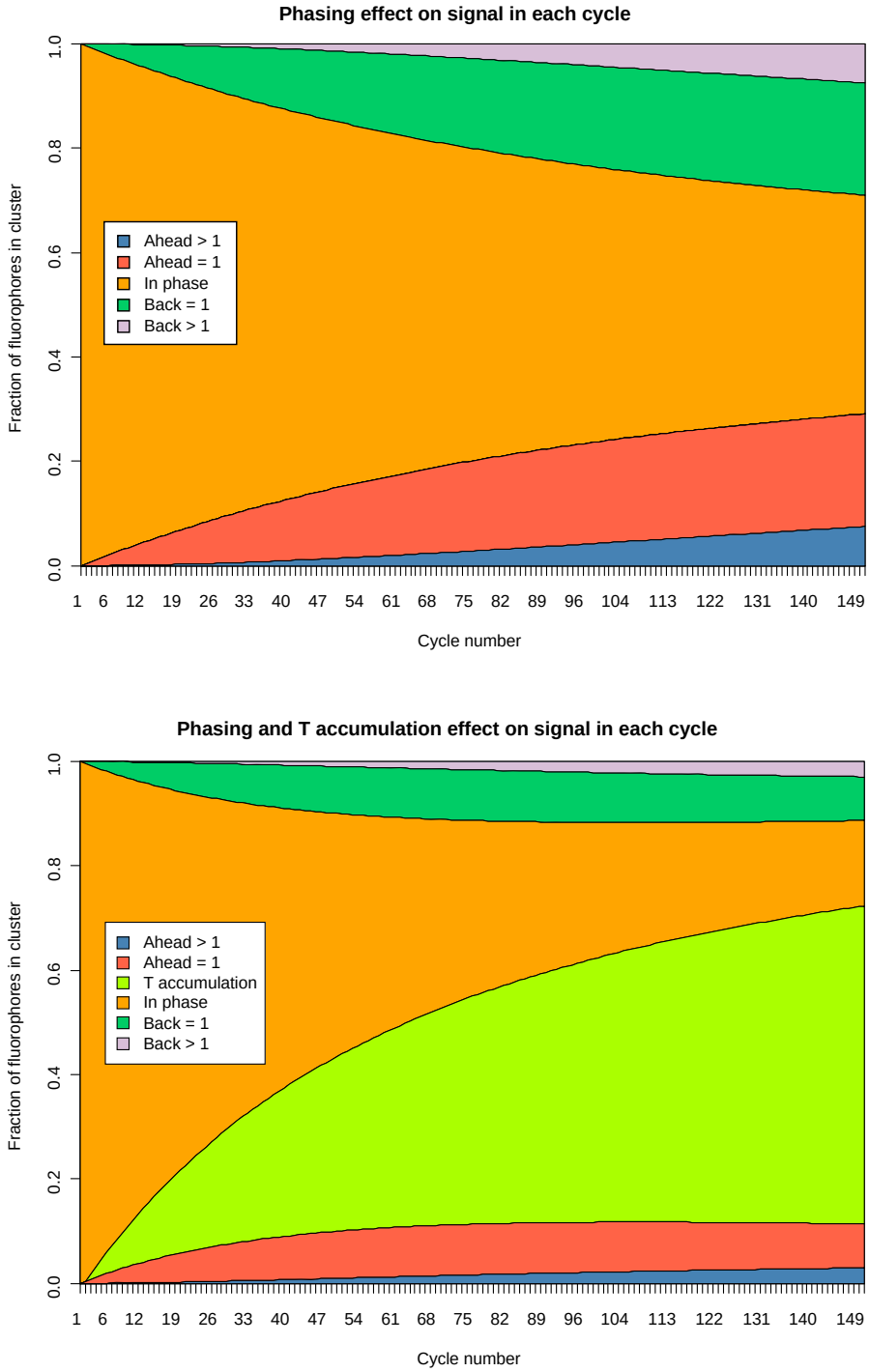

Figure 4.5: Development of the fraction of fluorophores representing the current cycle, the previous and the next cycles, cycles more than one ahead and more than one behind **(top)**, as well as fluorophores attached due to T accumulation **(bottom)**. The numbers are based on simulations of 10,000 clusters with a thousand identical sequences each. Sequences were created randomly, only using the GC content of ϕX174 (44.7%) as reference. A model with pre-phasing (0.4% per cycle), phasing (0.4% per cycle) and T accumulation (3.8% per cycle) was used and the fluorophore read outs for 150 sequencing cycles simulated.

IDT Exhibit 1061.0078

average 1,039 T fluorophores are bound to the 1,000 sequences in each cluster, exceeding the real incoporation signal. However, even after 150 cycles 85.1% of the fluorophores account for the previous, the current or the next base to be sequenced (in a sequencing chemistry without T accumulation). The conclusion from this simulation was that most of the signal to be captured by a statistical learner is contained in the raw intensities of the previous, the current and the next cycle.

4.4.2 Using the signal of neighboring bases

I implemented a base caller with Support Vector Machine (SVM) classifiers for each cycle which have the intensity values of the current cycle and its two neighbors as input. The exceptions are the first and last cycle, where only one of the neighbors can be included. For the SVM classifiers of each cycle, a computationally fast implementation of multi-class SVMs [40] with linear kernels, called $\texttt{SVM}^{multiclass}$ [230] and available from Thorsten Joachims[4] was used.

Support vector machines [39] are a generalization of the very basic principle of separating two classes by putting the thickest possible board between them, figuratively speaking. In a mathematical sense, they are a generalization of linear decision boundaries of separating hyperplanes that are placed in a way to obtain a maximum margin between classes. By maximizing the margin, the generalization error of the classifier, i.e. the error from labeling new examples based on their position relative to the decision boundary, is minimized. Frequently, real data is sufficiently complex or contains labeling errors, not allowing the separation using linear decision boundaries/hyperplanes. Therefore, support vector machines exploit that the input features (i.e. the input variables) may be combined to new features using so-called kernel functions, which then project the data in a higher dimensional space in which they may be separated by a hyperplane. Further, SVMs extend the maximum margin principle by allowing a hyperplane in the final feature space which mislabels some samples of the training data set, but still maximizes the distance to the nearest correctly separated examples.

By increasing the feature space, complex decision boundaries can be obtained from the combination of simple kernel functions and simple separating hyperplanes. Figure 4.6 on the following page provides a visual example for applying some kernel to a two dimensional data set and obtaining a hyperplane separating the data, which, if projected back to the two dimensional feature space, corresponds to a higher order decision boundary.

The implementations of support vector machines for two classes do unfortunately not directly translate to SVMs for multiple class labels. Typically, when requiring a n-class classifier based on SVMs, n classifiers are trained, each separating the one class from all other classes ("One-against-all"). Another approach of converting the multi-class problem to a binary problem is the training of pair-wise classifiers ("one-against-one"), each sub classifier separating two classes of the class set from each other. This pair-wise approach vastly increases in computational complexity for higher numbers of class labels, even though also a reduced subset of the data is used with each of the classifiers. Most commonly used are "one-against-all" SVM classifiers, which are then frequently referred to as multi-class SVMs.

However, actual multi-class SVM implementations, even though with constraints on the kernel functions, exist as so called structural SVMs [41], which also extend SVMs from independent and categorical labels to structured objects with relationships defined between them. To distinguish them from these other approaches, they are sometimes referred to as "True multi-class". The $\texttt{SVM}^{multiclass}$ package by Thorsten Joachims implements this latter

[4] `http://svmlight.joachims.org/svm_multiclass.html`

IDT Exhibit 1061.0079

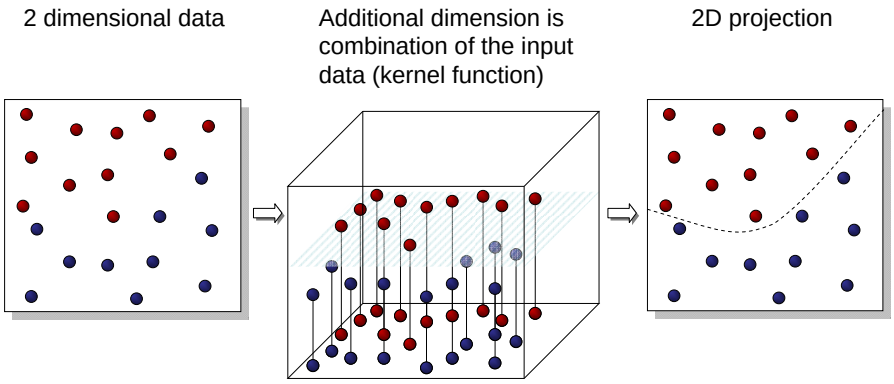

Figure 4.6: The n input features of a classification problem can be combined using kernel functions to add additional features to the input space. This extends the R^N features space by m dimensions (R^{n+m}). The idea is to put the data in a new extended feature space, in which class labels may be separated by a hyperplane. This way, complex decision boundaries are the result of the combination of simple kernel functions and simple separating hyperplanes.

type of multi-class SVMs [230]. Figure 4.7 on the next page provides a visual example, summarizing the three types of multi-class support vector machines.

For training a classifier, a set of examples is required, i.e. multiple examples of up to twelve intensity values (the input features) and the assigned base (the class label) need to exist for each cycle of the sequencing process. A putative training data set is created by aligning the `Bustard` raw reads with mismatches for a fraction of the tiles to an appropriate reference sequence (e.g. ϕX174 RF1) using `SOAP` v1.11 [135]. Half of this data set is kept as a test data set and the other half used for training the classifiers (`svm_multiclass_learn`) separating all four nucleotide classes (A, C, G, and T) in each cycle.

The result of the training is verified by using the test data set with the trained models (`svm_multiclass_classify`) and comparing the predicted labels with the ones obtained from the reference sequence. Evaluating this information, one can also estimate parameters for calculating a quality score for each called base given the class assignment and the distances to the classification/decision boundary reported by `SVM`^{multiclass}. Based on this measure, I use the density distributions for the four distances to the decision boundary seen for each correct class label (16 in total, in an example data set each was observed to follow a normal distribution based on Shapiro Wilk Normality test [208, 199]). Given the four distances d_Z ($Z \in \{A, C, G, T\}$) and the parameters determined for the normal distributions estimated from the actual test data set, the likelihood of the called base being wrong was defined as:

$$p(\neg base) = \frac{\sum_{Z \neq base} p(Z \mid base)}{\sum_{Z \in \{A,C,G,T\}} p(Z \mid base)} \Bigg| \, p(Z \mid base) = p(Z \wedge base) \cdot cdf(d_Z, \mu_Z, \sigma_Z) \quad (4.1)$$

It was necessary to extend the `SVM`^{multiclass} C/C++ package by routines that are able to handle several classifiers in parallel for the individual cycles, parse `Firecrest`, `IPAR`, `RTA` and `OLB` output files, calculate quality scores and create `FastQ` output files on a per tile basis. `FastQ` is currently the most common sequence file format used both in data exchange and as input for post-sequencing software. This text format is an extension of the `FASTA` sequence format, where each sequence in the file is associated with an identifying tag/identifier, and also with an additional line for quality scores (figure 4.8 on page 82). Unfortunately, there is

IDT Exhibit 1061.0080

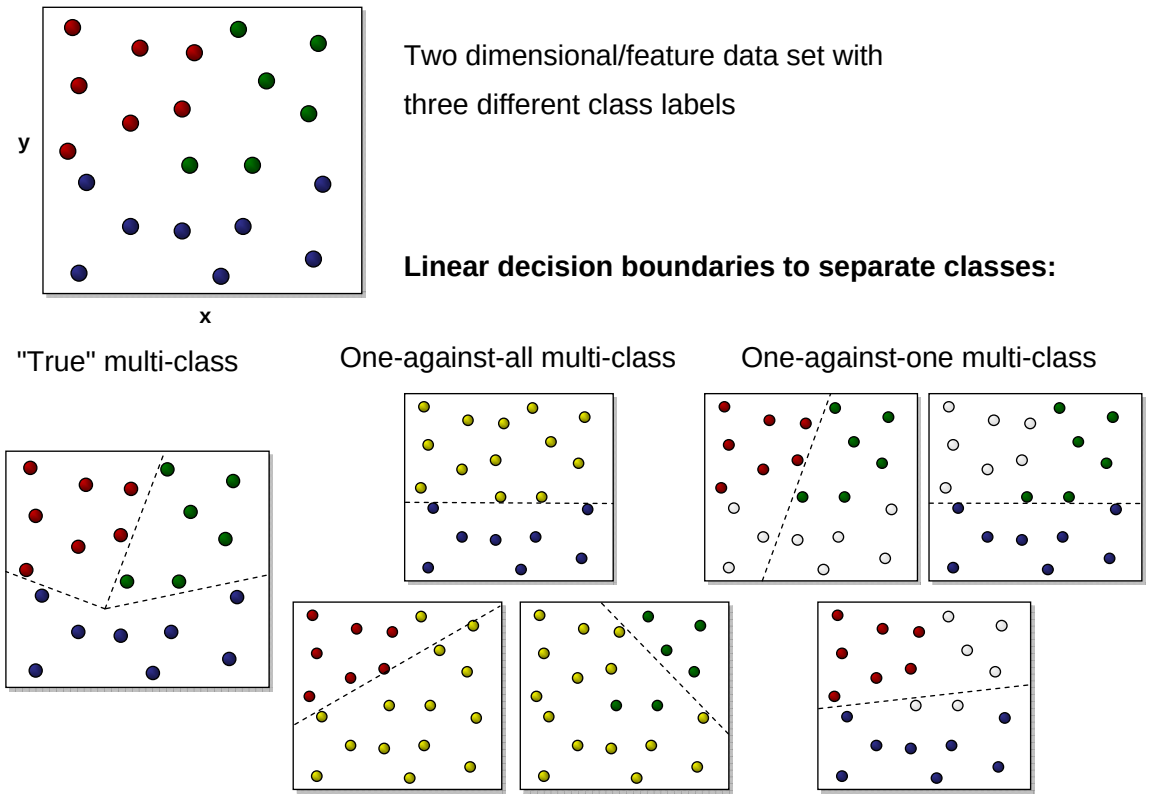

Figure 4.7: Most implementations of multi-class support vector machines are not actually separating multiple classes in one classifier. Typically, when requiring a n-class classifier based on SVMs, n binary classifiers are trained, each separating the one class from all other classes ("one-against-all"). Another approach of converting the multi-class problem to a binary problem is the training of pair-wise classifiers ("one-against-one"), each sub classifier separating two classes of the class set from each other. This pair-wise approach vastly increases in computational complexity for higher numbers of class labels, even though also a reduced subset of data is used with each of the sub classifiers. Actual multi-class SVM implementations are rare and constraint in the kernel functions used.

IDT Exhibit 1061.0081

```
...
@SOLEXA-GA03_0001_PEi_SG:5:1:1033:5267
AGACAGACACAGAGNAAGACCCAGTCCGCCACACAGGCAAACTCACGCAGTACGCGCCG
+SOLEXA-GA03_0001_PEi_SG:5:1:1033:5267
4--'-(/.23/044!51/+//.400/-/1-62/.6021834///62126313/2230+2
...
```

Figure 4.8: A FastQ file starts with an '@' character followed by a unique sequence identifier (platform specific, shown is an Illumina Genome Analyzer read ID providing run ID, lane, tile and X-Y-coordinates), the next line contains the sequence, then a line starting with a '+' character, either followed again by the complete read identifier or just containing the '+', indicates that the next line contains the quality scores. This example encodes quality scores in the Sanger standard, which uses ASCII characters from 33-127 to encode base qualities in `PHRED`-scale between 0 and 94 (e.g. '4' (ASCII 52) corresponds to `PHRED` quality score 19 thus an error likelihood of 1.26%, while '-' (ASCII 45) corresponds to `PHRED` 12 and an error probability of 6.31% (equation 4.2).

no universally accepted rule regarding how quality scores are encoded in this format. `Ibis` follows the Sanger standard of encoding `PHRED`-like quality scores [63] with one character per quality score and with an offset of 33 ('!'), as this is a widely accepted encoding:

$$char(round(-10 \cdot \log_{10}(p(\neg base))) + 33) \tag{4.2}$$

In its first versions, the Illumina software did not use such `PHRED`-like probabilities (equation 4.3). Instead, they used a quality score model with negative values (equation 4.4), and therefore set the zero quality to 64 ('@'). In current software versions, `PHRED`-like probabilities are used, but still printed with the offset of 64. Other encodings include the actual quality score numbers separated by space in the quality line (e.g. `AltaCyclic` base call files). This encoding is considered to be inefficient and is used rarely today.

$$Q_{PHRED} = -10 \cdot \log_{10} p(\neg base) \tag{4.3}$$

$$Q_{OldIllumina} = 10 \cdot \log_{10}\left(\frac{p(base)}{1 - p(base)}\right) \tag{4.4}$$

Applying the outlined approach, i.e. `Ibis`, to the lane shown in figure 4.3A on page 74 increases the number of perfectly mapped sequences from 39.8% to 60.2% (from 4,564,039 to 6,908,856) and shows an error profile of all mapped sequences (9,695,354 out of 11,478,043) as depicted in figure 4.3B on page 74 without a noticeable T accumulation effect.

4.5 Comparison to other systems for base calling

Even though the application of statistical learning for the base calling of Illumina sequences is not novel, `Ibis` differs significantly in its concept and its performance from earlier proposed approaches.

`AltaCyclic` [61] uses a model of phasing/pre-phasing, fluorescent decay and cycle-dependent cross-talk to correct raw intensities before classification. It then uses SVM classifiers trained

IDT Exhibit 1061.0082

individually for every cycle on those corrected raw intensities. The `AltaCyclic` model used for this correction step does not include base specific phasing parameters and therefore cannot correct raw intensities for the observed T accumulation effect.

Similarly, the `Rolexa` package [198] corrects the raw intensities prior to the application of Gaussian mixture models as classifiers. Deviating from the models of sequencing chemistry implemented in `AltaCyclic` and `Bustard`, `Rolexa` models only cross-talk and single-parameter phasing (pre-phasing is not modeled). In contrast to `AltaCyclic`, `Bustard` and `Ibis`, `Rolexa` applies a transformation to the intensities within each tile to correct for local differences in the illumination of clusters. Further `Rolexa` uses IUPAC ambiguity codes to encode uncertainty in base calling, while `AltaCyclic`, `Bustard` and `Ibis` try to call one correct base and reflect the associated uncertainty in the quality scores.

Even though IUPAC amiguity codes provide information on alternative base calls, the base quality approach is superior when the sequences are mapped and analyzed with software that is unable to handle these codes (like most available fast mappers or SNP calling software) and it provides a higher resolution for the trust in unambiguous calls (the majority of base calls). Unlike `AltaCyclic` and `Bustard`, `Ibis` does not call an 'N' character for low quality bases, as the most likely base can still be informative and as its uncertainty is already captured in the quality score. For `Ibis`, 'N' characters are only returned if all intensity values obtained from the intensity files are exactly zero for a specific cycle. In this case the image analysis software failed in recovering the cluster after image registration with the template of cluster positions.

4.5.1 Performance test on data sets using v1 chemistry

The difference in introducing IUPAC ambiguity codes complicates the direct comparison of `AltaCyclic` v0.1.1, `Bustard` v1.9.5, `Ibis` 1.0.0 and `Rolexa` v1.1.6 (with `R` v2.8.0). Therefore, `Rolexa` was configured to call sequences without using ambiguity codes:

```
Rolexa.env$HThresholds <- c(2.0,2.0,2.0);
Rolexa.env$IThresholds <- (log2(41:nrcycles/6));
Rolexa.env$iupac <- c("A", "C", "G", "T", "N", "N",
    "N", "N", "N", "N", "N", "N", "N", "N")
```

Further, 'N' characters are specifically considered for a direct comparison. I tested the performance of the four different base callers first on three data sets using the first sequencing chemistry version (FC-104-100x):

(A) ϕX control lane of a 26 cycle Genome Analyzer I run

(B) Human shotgun lane from the same Genome Analyzer I run

(C) ϕX control lane of a 51 cycle Genome Analyzer II run

For these data sets I mapped all control lane sequences to the ϕX reference sequence allowing up to five mismatches but no gaps using `SOAP` v1.11 [135]. For the lane with human shotgun sequences, the sequences were mapped to the human reference genome (hg18/NCBI Build 36.1) allowing five mismatches without any gaps, but then restricting the further analysis to sequences mapping with at most two mismatches to reduce the number of false positive

IDT Exhibit 1061.0083

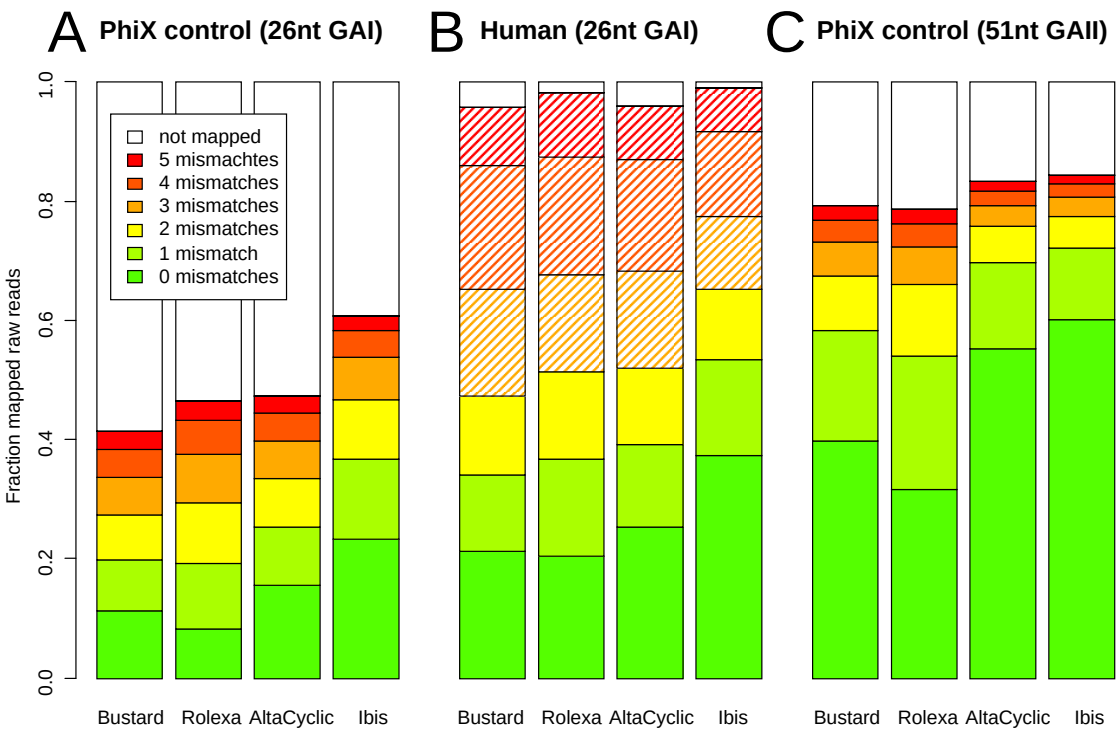

Figure 4.9: Fraction of mapped reads and corresponding number of mismatches for the three tested data sets. On the left the result for the ϕX control lane of 26 cycle Genome Analyzer I run **(A)**, in the middle a lane of human shot gun sequence analyzed on the same 26 cycle Genome Analyzer I run **(B)** and on the right the ϕX control lane of a 51 cycle Genome Analyzer II run **(C)**. The raw sequences of all three data sets were mapped to the corresponding reference genome (hg18/NCBI Build 36.1 and ϕX174 RF1) with up to five mismatches but no gaps using `SOAP` v1.11. For **(B)**, mappings with more than two mismatches are shown with dashed lines since a high number of false positive placements is expected when mapping short reads to a large genome sequence.

IDT Exhibit 1061.0084

placements expected when using a genome with almost three billion bases and short reads. The fraction of mapped raw reads and corresponding number of mismatches for the three lanes is shown in figure 4.9 on the preceding page.

The number of correct reads when using `Ibis` compared to `Bustard` increased about 2.1 fold in (A) (11.3% to 23.4%), 1.8 fold in (B) (21.2% to 37.4%), and 1.5 fold in C (39.8% to 60.2%). When comparing the error profiles of (C) (again figure 4.3 on page 74), one clearly sees that `Ibis` was able to correct for the T accumulation pattern present in the v1 chemistry version. Assuming that all reads belong to the corresponding reference, a (lower) estimate of the error rate in the run (assuming the remaining reads would be matched when allowing one more mismatch) can be obtained. For (A) these are 15.2%, 16.4%, 12.3% and 16.0% for `AltaCyclic`, `Bustard`, `Ibis` and `Rolexa` respectively. For (B) (assuming to match the rest with 3 mismatches) these are 7.1%, 7.6%, 5.5%, and 7.4%. In the third data set, the 51 cycle ϕX control lane the error rate is much lower (due to the better quality of the run and the technical improvements of the GAII instrument); the rates for `AltaCyclic`, `Bustard`, `Ibis` and `Rolexa` are 3.0%, 4.0%, 2.8% and 4.3% respectively.

The development of the mismatch rates per cycle observed in the mapping for each of the three data lanes is available in figure 4.11 on page 87. For data set (C), I also compared the quality scores reported by `Bustard`, `AltaCyclic` and `Ibis`. While `Ibis` provides `PHRED`-like quality scores, `Bustard` and `AltaCyclic` use the Illumina-specific encoding of quality scores with a different offset (64 instead of 33) and a different formula (Illumina Analysis Pipeline 1.0 and earlier versions). Therefore, quality scores from `AltaCyclic` and `Bustard` (equation 4.4 on page 82) were converted to `PHRED`-like quality scores (equation 4.3 on page 82) and compared in `PHRED` scale, equation 4.5.

$$Q_{PHRED} = -10 \cdot \log \left(\frac{1}{1 + 10^{\frac{Q_{OldIllumina}}{10}}} \right) \tag{4.5}$$

The results are available in figure 4.10 on the next page. When measuring the deviation from the optimal line, `Bustard` has a root mean square deviation (RMSD) of 84.9, `AltaCyclic` of 19.3 and `Ibis` of 0.9. Hence, `Ibis` provides useful quality scores for further analyses.

4.5.2 Performance on a test data set from v2 chemistry

Here, a comparison of three base callers (`AltaCyclic`, `Bustard` and `Ibis`) is shown on the ϕX control lane of a 77 cycle run, which was created using the FC-204-20xx chemistry (v2). This comparison does not include the `Rolexa` base caller as `Rolexa` could not be used succesfully on these longer reads. Even though this chemistry shows improved phasing values which allow for a longer read length (this chemistry was officially supported for up to 51 cycles), it still shows the T accumulation effect of the previously shown v1 sequencing chemistry.

For this 77 cycle run, the base miscalls by substitution are available for each of the base callers in figure 4.12 on page 89. This figure also includes the performance on a training data set directly obtained from the `Bustard` raw sequences without using a reference sequence. In this case about 3.7 million `Bustard` reads with at most three 'N's were used as training data. This procedure can be considered as the 'last resort' for using `Ibis` in cases where no reference sequence is available for creating a better training data set. A reduction of the error rate by about 20% compared to `Bustard` is seen, further about 10% more reads can be mapped with the same number of mismatches allowed. As expected due to large T accumulation effect on the last bases of those reads, this is inferior to the results obtained

IDT Exhibit 1061.0085

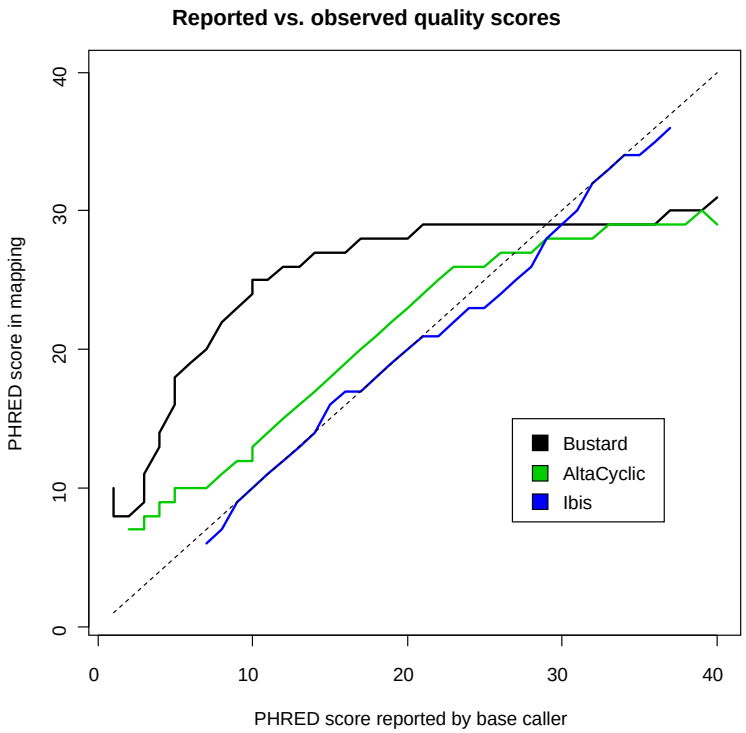

Figure 4.10: Comparison of quality scores for the 51 cycle ϕX control lane data set (C). Quality scores reported by `Bustard`, `AltaCyclic` and `Ibis` are compared in `PHRED` scale. For all three base callers only quality scores reported with 100,000 and more observations were considered. Calculating the deviation from the optimal line `Bustard` has a RMSD of 84.9, `AltaCyclic` of 19.3 and `Ibis` of 0.9.

IDT Exhibit 1061.0086

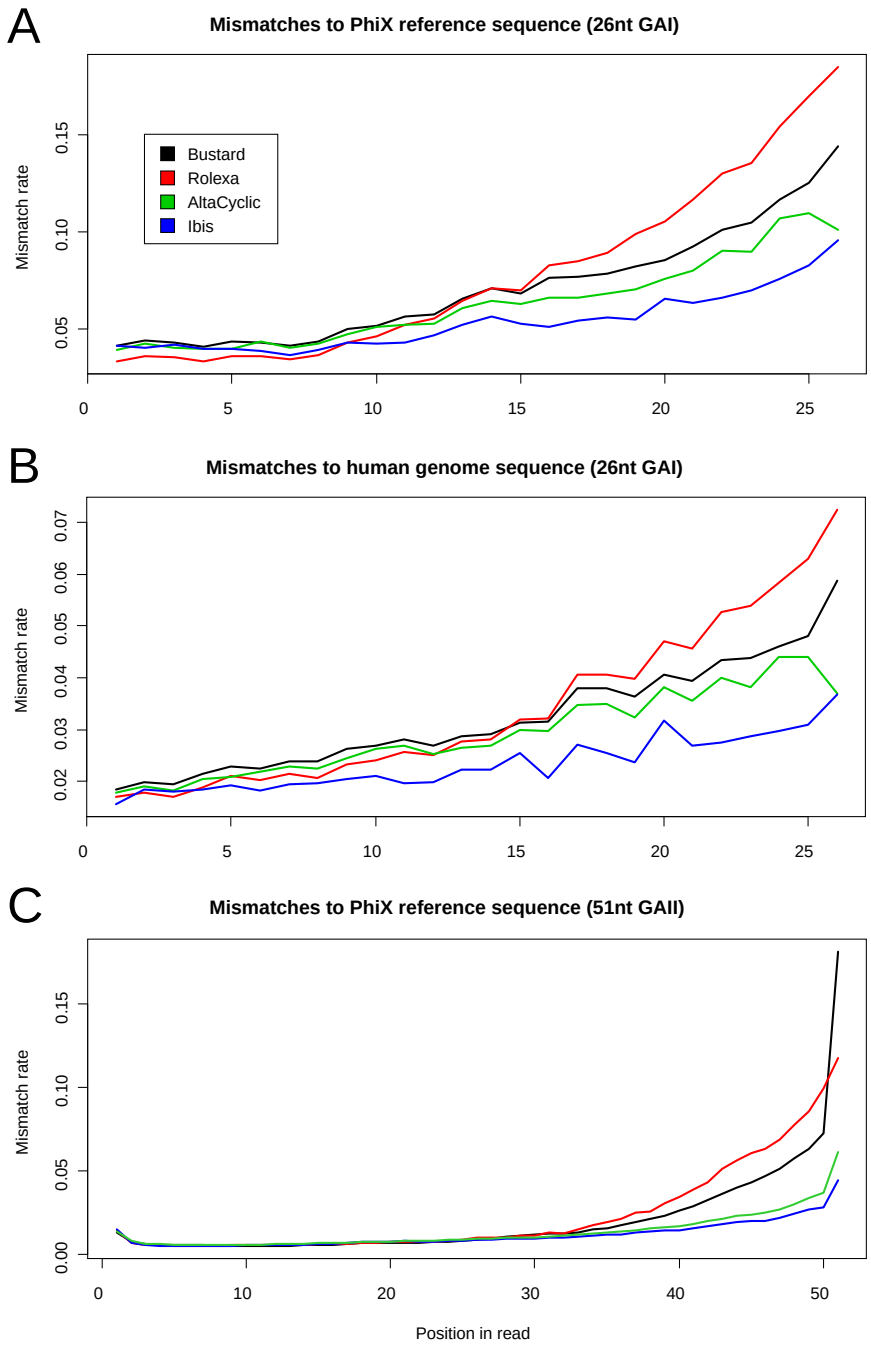

Figure 4.11: The development of error rates was determined by assuming no sample divergence from the reference (i.e. error rate equals mismatch rate) from alignments of reads called using the four different base callers (`Bustard`, `Rolexa`, `AltaCyclic` and `Ibis`) to the reference genomes. **(A)** Shows a 26 cycle Genome Analyzer I run (FC-104-100x, v1) of which the ϕX control lane was analyzed as well as one lane with human shot gun sequences **(B)**. For a second, a 51 cycle Genome Analyzer II run (FC-104-100x, v1), only the ϕX control lane was analyzed **(C)**.

IDT Exhibit 1061.0087

using a reference (in which error rate reduction is more than 70% and 52% more reads are mapped). However, even without a reference, base calling using `Ibis` provides a considerable improvement compared to the `Bustard` base caller.

4.5.3 Performance test on current data sets from Genome Analyzer IIx

Even though `Ibis` was originally developed to handle the T accumulation in sequencing chemistry (FC-103-300x, v1) and (FC-204-20xx, v2) which have been replaced by new versions during the last two years, its application is not limited to the reprocessing of data created with the older chemistries.

In spring 2009, Illumina released a new chemistry version (FC-103-300x, v3) in which the T accumulation effect could be greatly reduced. This was achieved by replacing the cleavage reagent and wash buffers used in the previous generation chemistry. In late summer 2009, Illumina started field tests of a new polymerase used for base incorporation and then released a new chemistry (FC-104-40xx, v4) including this polymerase in winter 2009. This new polymerase reduces the pre-phasing effect and therefore allows for longer reads. In autumn 2010, Illumina released sequencing chemistry version 5 as the first step of the TruSeq marketing regime.

To show that the application of `Ibis` extends to new chemistries, a 76nt cycle run created with the FC-103-300x chemistry is shown in figure 4.13 on page 90 and a 101 cycle run using the new polymerase provided by Illumina within the early access program in summer 2009 is shown in figure 4.14 on page 91. In these figures, the base miscalls by substitution as well as the fraction of reads mapped for each base caller is presented. These numbers were obtained by mapping the raw control lane sequences to the ϕX reference sequence allowing up to five mismatches but no gaps using `SOAP` v1.11. Again 'N's in the sequence were considered as mismatches to the reference and the five mismatch cutoff reinforced before further analysis. Error rate estimates are based on reads mapped with `Bustard` and each of the base callers.

Comparing the results obtained for the Illumina base caller (`Bustard`), `AltaCyclic` and `Ibis` with the results for training `Ibis` without a reference, the no-reference approach performed surprisingly well. It is considerably better than the `Bustard` results, even though SVM training being completely based on the `Bustard` error profile: 6%/16% more reads mapped and a reduction of the error rate by 28%/24% respectively.

In addition, for these two runs I have shown that it is possible to use mitochondrial reads from shotgun experiments as an alternative way to create a training data set. In the first example (76 cycle run), about 50,000 sequences extracted from only one lane and in the second example (101 cycle run) 1.8 million human mitochondrial sequences extracted from several lanes of the run were used. In both cases, the results for using the ϕX control versus using the mitochondrial reads as input for training are similar and outperform the no-reference approach: 27% vs. 24.7% reduction of the error rate and 9.8% vs. 9.1% more reads mapped for the 76 cycle run and 45.6% vs. 45% reduction of the error rate and 49% vs. 44% more reads mapped for the 101 cycle run.

Figure 4.15 on page 92 shows the error development of one run using the final v4 sequencing chemistry (instead of v3 chemistry with the early access polymerase shown in figure 4.14 on page 91, v3RDP). The only known difference between v3RDP and v4 chemistry is the exchange of Tag polymerase by Phusion polymerase in the the first extension reaction of cluster generation. Again the error rate can be considerably reduced (43.48% reduction) and the number of mapped reads increased (5.8%) by applying `Ibis` instead of `Bustard` v1.6.0 (OLB/RTA v1.6.0).

IDT Exhibit 1061.0088

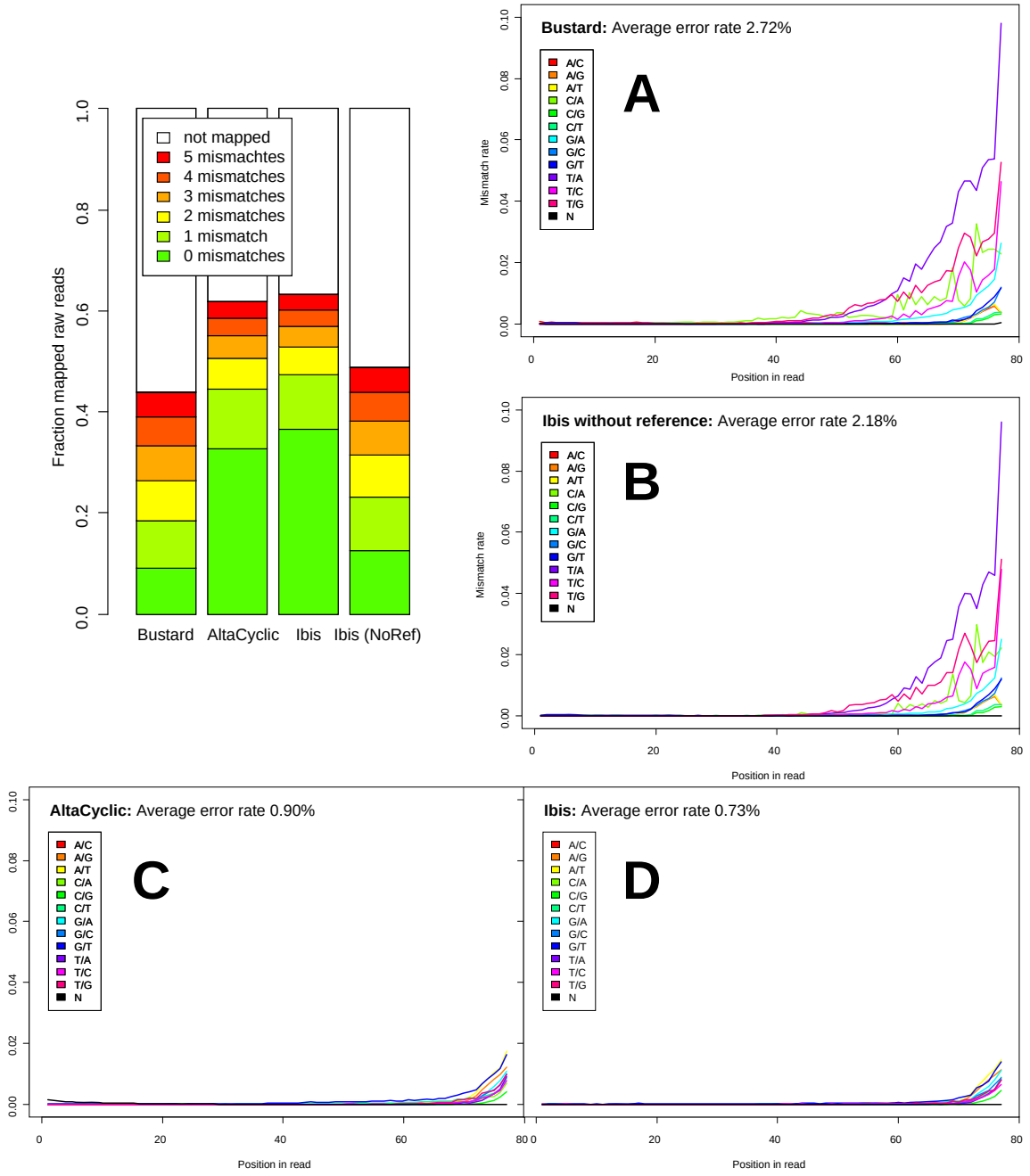

Figure 4.12: Mismatches to the reference sequence observed with different base calling strategies for the ϕX control lane of a 77 cycle Genome Analyzer II run with T accumulation (v2, FC-204-20xx chemistry). Plot **A** shows the results for the standard Illumina base caller (`Bustard`), plots **C** (`AltaCyclic`) and **D** (`Ibis`) show strategies using a reference sequence (ϕX174); in plot **B** `Bustard` reads have been used directly as input for the `Ibis` training process, without using a reference sequence. The reduction of the error rate in **B** is about 20% compared to `Bustard` and about 10% more reads can be mapped with the same number of mismatches allowed. This approach is, as expected, inferior to the results obtained using a reference (**D**, in which error rate reduction is more than 70% and 52% more reads are mapped).

IDT Exhibit 1061.0089

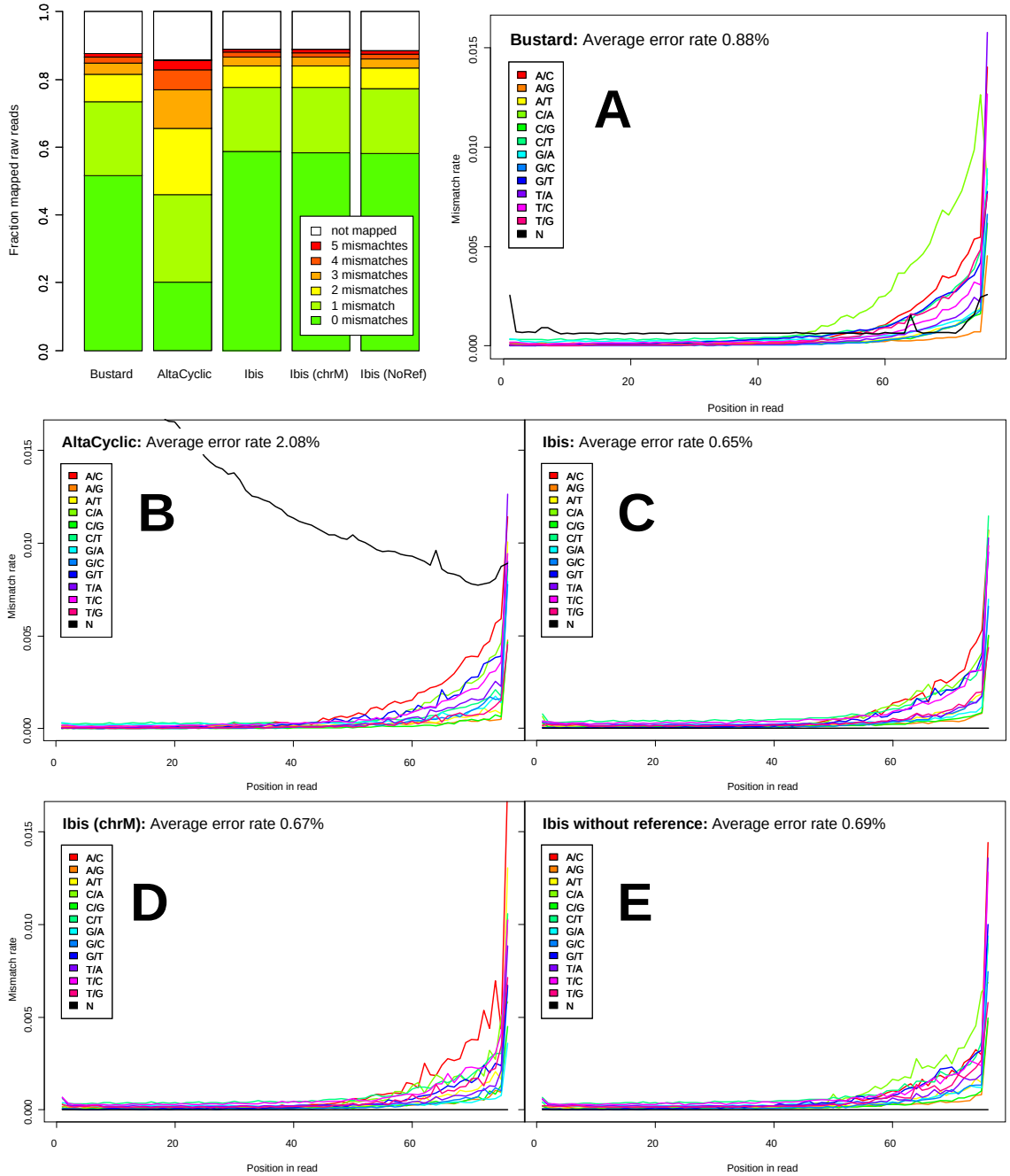

Figure 4.13: Mismatches to the reference sequence observed with different base calling strategies for the ϕX control lane of a 76 cycle Genome Analyzer II run using the current chemistry (FC-103-300x). Plot **A** shows the results for the standard Illumina base caller (`Bustard`). Plots **B**, **C** and **D** show strategies using a reference sequence (**B** and **C** ϕX174 and **D** the human mitochondrial reference sequence). In plot **E**, `Bustard` reads have been used directly as input for the `Ibis` training process, without using a reference sequence.

IDT Exhibit 1061.0090

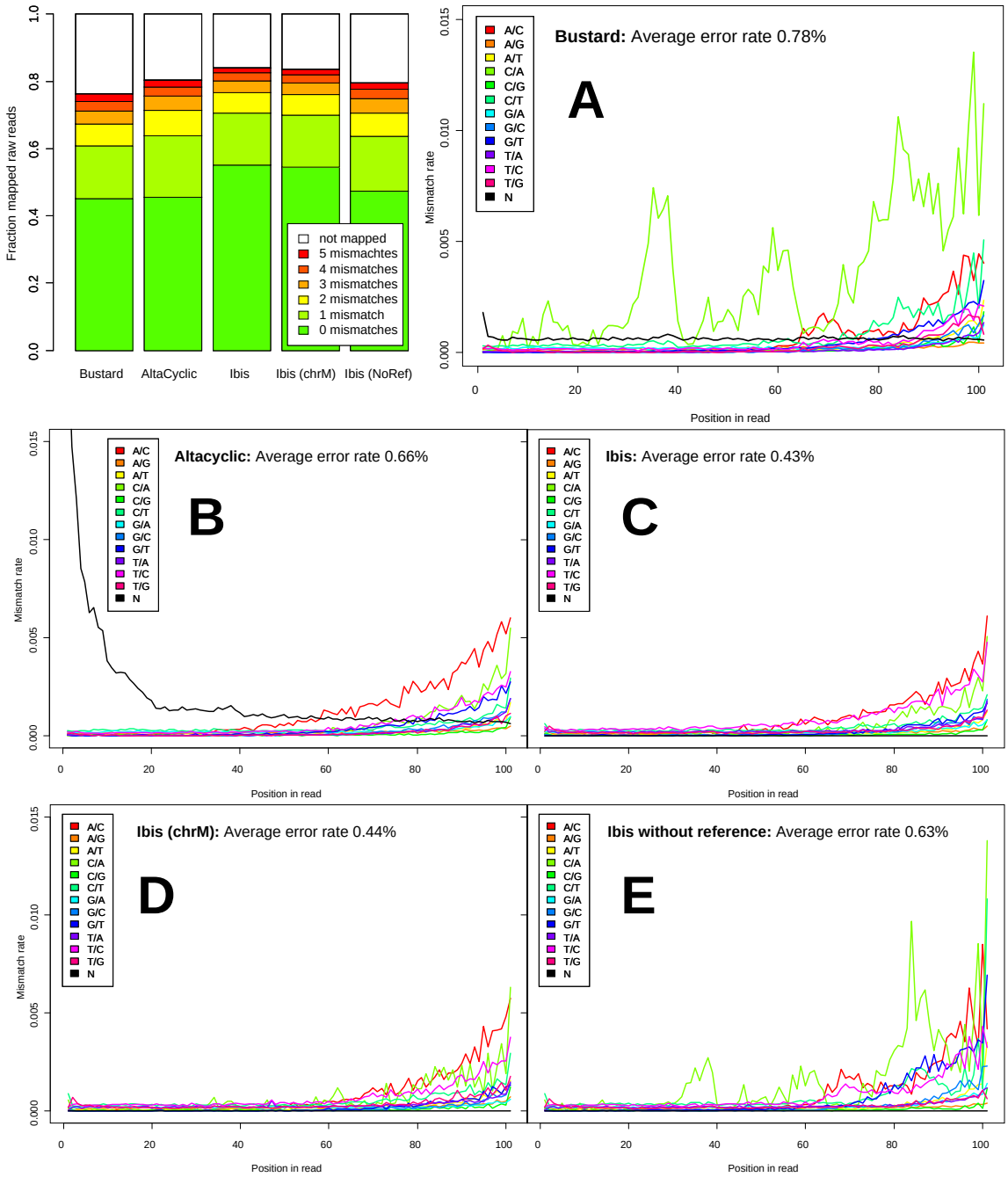

Figure 4.14: Mismatches to the reference sequence observed with different base calling strategies for the ϕX control lane of a 101 cycle Genome Analyzer II run using v3 chemistry (FC-103-300x) and new polymerase (which was later released in v4 sequencing chemistry kits, FC-104-40xx). Plot **A** shows the results for the standard Illumina base caller (`Bustard`). Plots **B**, **C** and **D** show strategies using a reference sequence (**B** and **C** ϕX174 and **D** the human mitochondrial reference sequence). In plot **E**, `Bustard` reads have been used directly as input for the `Ibis` training process, without using a reference sequence.

IDT Exhibit 1061.0091

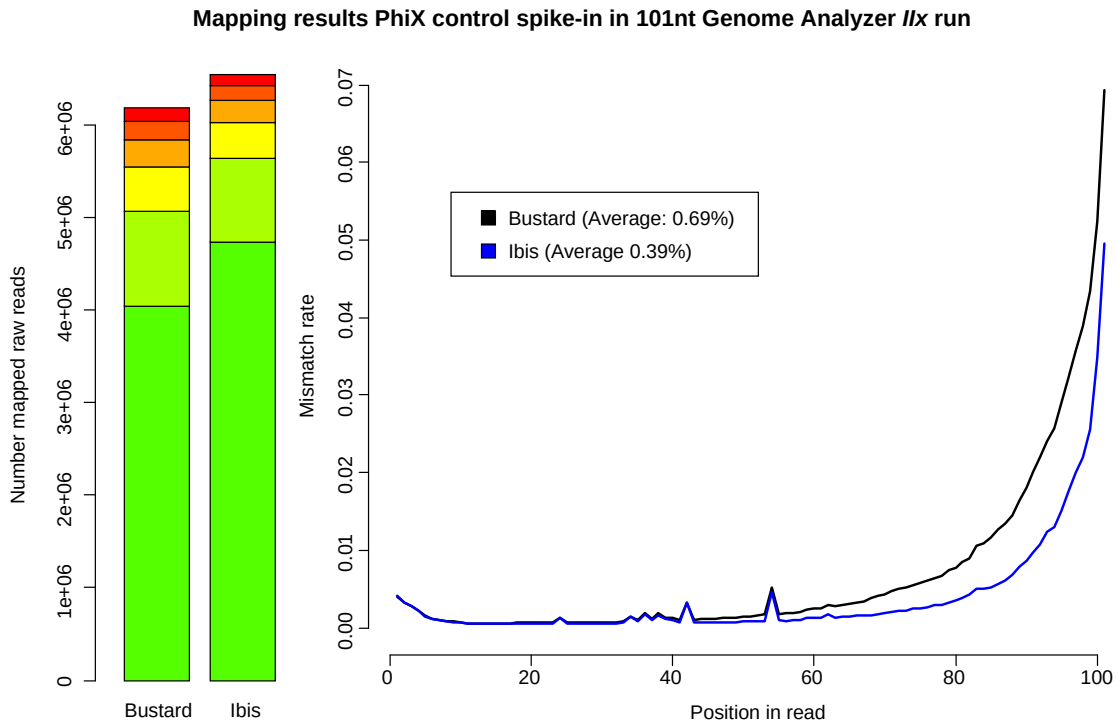

Figure 4.15: Mismatches to the ϕX reference sequence observed for `Bustard` and `Ibis` base calls of the ϕX spike-in control reads of a 101 cycle Genome Analyzer IIx run using v4 chemistry (FC-104-40xx).

4.5.4 Required computational resources

To compare the computational resources required for base calling, I measured the time for training and predicting the 51 cycle ϕX control lane created with v1 chemistry with each of the base callers. Base calling this lane using `Bustard` on an eight core system took 50 minutes (including estimation of cross-talk and phasing parameters) and created the input needed for all three other base callers. `AltaCyclic` needs a cluster system to run. Using about 80 cores of an institute cluster (machines with 4 x 2.6GHz cores, 16Gb RAM) at the MPG Rechenzentrum in Garchingen, `AltaCyclic` took about 5.5 hours for the parameter estimation and 40 minutes for base calling. On an eight core system (8 x 3.0GHz, 32Gb RAM) these would correspond to at most 61 hours in total (Amdahl's law [4]). Running `Rolexa` on an eight core machine took 17.5 hours. `Ibis` took 89 minutes for parameter estimation and 12 minutes for prediction, in total about 1.7 hours. In other words, using `Ibis` one has to invest three times more time for base calling, for `Rolexa` 21 times more time and for `AltaCyclic` 73 times more time compared to `Bustard`.

Memory consumption is difficult to evaluate for the different architectures and approaches, but for all programs at least 4Gb RAM should be available. This requirement is easily met by current computer systems used for high-throughput data analysis, since genome index structures required for alignment often exceed this limit. In terms of disk space, recalling bases will of course require additional space for storing the new base calls. The disk space required for storing the model parameters is in the range of mega bytes and can thus be neglected. However, the actual training and base callling process might require transformations of the intensity files or `Bustard` base call data and thus produce intermediate files of significant size (up to several 100Gb). Therefore, a fast and large temporary file system is advantageous.

IDT Exhibit 1061.0092

4.5.5 Dependence on training input data

As is the case for `Bustard`, `AltaCyclic` and `Rolexa`, the results shown before support the assumption that training on the ϕX extends well to the prediction of other lanes using the same estimated models. To further verify this, I did a specific test for overtraining (e.g. learning base composition of a specific reference) and undertraining on a 51 cycle run using v2 chemistry.

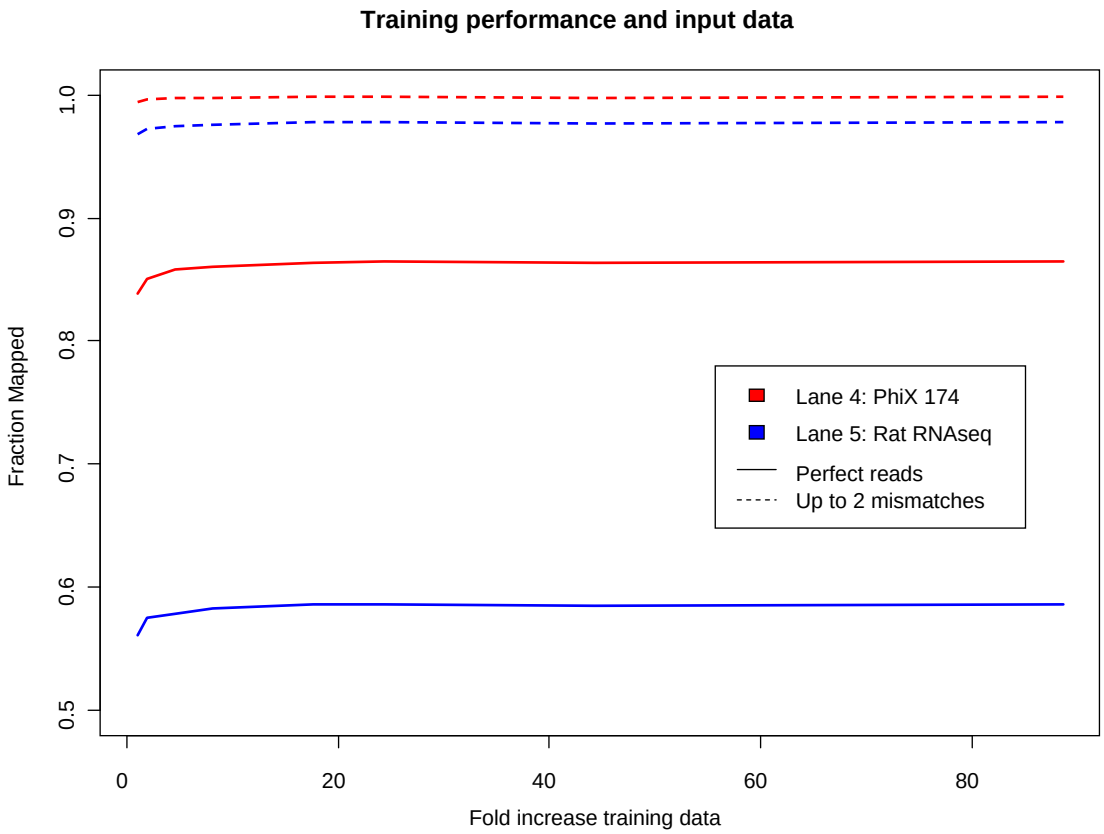

Figure 4.16: The plot shows the fraction of raw reads mapped (with up to two mismatches and without any mismatches) for a 51 cycle run for which several `Ibis` runs were performed with different amounts of training data from the ϕX control (lane 4). The base calling models obtained for the ϕX control were applied to recall sequences of the ϕX lane as well as of a neighboring lane (lane 5) with rat RNAseq reads. Reads were aligned to the ϕX174 reference and Rat Baylor 3.4/rn4 chromosome 1 using `SOAP` v1.11. For rat the fraction mapped was corrected for the length of chromosome 1 relative to the complete genome assembly (267,910,886 out of 2,571,531,505nt).

I trained several models from the ϕX in lane 4 using different numbers of tiles (1, 2, 5, 9, 19, 27, 50, 100 of 100 available tiles) for training and predicted with the resulting models the ϕX lane as well as a neighboring lane (5) which contained rat (*Rattus norvegicus*) RNAseq reads. I then examined the number of sequences mapped to the two different reference genomes and the number of mismatches observed (see figure 4.16). From this data set, there was no evidence for overtraining; however undertraining was observed, affecting the prediction of both lanes. In this test, undertraining resulted in 3-5% fewer perfect reads and only up to 1% less mappable raw reads than obtained when using at least 1,000,000 sequences for training (typically about 5-15 tiles of dedicated control lane, depending on the loading density/instrument version).

IDT Exhibit 1061.0093

4.5.6 Training without a dedicated control lane

As shown in the sections before, `Ibis` is not dependent on the inclusion of a control lane. In the case of resequencing projects or projects where some subset of the sequences generated come from a previously characterized genome, it is possible to use this sequence data for obtaining a training dataset for `Ibis`. I have shown (figures 4.13 on page 90 and 4.14 on page 91) that it is possible to use the mitochondrial sequences generated as part of a shotgun sequencing experiment for creating such an alternative training set. In the example presented, the `Ibis` results are very comparable between the models obtained from the ϕX reads and the models from the mitochondrial reads. Further, I have shown that the raw `Bustard` output can be used as training data in cases where there is no reference sequence available (figures 4.12 on page 89, 4.13 on page 90, and 4.14 on page 91). Although the reduction in error rate is less than obtained when using a reference (especially when the `Bustard` base calls are frequently erroneous due to a specific bias like T accumulation), improved base calls were also obtained in this setup.

In general a control lane, loaded with a DNA sequencing library created from the well-characterized ϕX174 strain, is only required by Illumina for runs containing samples with unbalanced base composition (e.g. smallRNA, RNAseq, ChIP-Seq, sodium-bisulfite treated DNA). However, if this control lane is available, it does not only allow for correct parameter estimates of the `Bustard` base caller, it also allows for between-run quality control and comparisons. It is thus also of interest when requesting replacement of sequencing run chemicals and materials from Illumina due to material defects. For `Ibis`, the control reads are used for calibrating the quality scores of a run. This all creates an actual benefit in using the same control library on all runs.

Though a flow cell has eight lanes, only seven lanes can be used for libraries of interest when having a dedicated control. This increases effective sequencing costs and lowers run throughput. Further, it is recommended that the control lane be in the middle of the flow cell, which typically has the highest quality and is rarely affected by run problems (see chapter 3). Thus the control uses the most favorable region and the quality statistics obtained for the control and their generalization to all lanes is put into question.

When using a multiplex library preparation protocol [150, 141] for creating an indexed control DNA library, this library can be spiked at equal concentration into each of the eight lanes and the control reads identified based on their unique index/barcode sequence. When spiking 1-2% of such a control library into all lanes (section 3.9 on page 63 in chapter 3), the resulting about 250,000 to 1 million control reads in each lane allow for lane-specific quality measurement and provide, with a total of up to 8 million reads, more than sufficient training data for `Ibis` (see figure 4.15 on page 92 for a run with spike-in control reads and no dedicated control lane). This increases usable (i.e non-control) sequence output compared to a dedicated control lane and provides lane-specific but library-independent quality measurements. Using the same high quality control library for all sequencing runs, per lane statistics obtained can be easily compared, even between different runs. When applying `Ibis`, the quality scores are then also adjusted for each run based on the same control library. Thus `Ibis`'s `PHRED`-like quality scores are comparable between sequencing runs and libraries and generally do not require further normalization (see also section 4.6 on the next page).

Even for whole runs of libraries with unbalanced base composition a dedicated control lane can be omitted. In such a case, `Bustard` model parameters from a run with the same sequencing chemistry version can be used for obtaining a training data set on which `Ibis` base calling can be performed. Even though the application of standard parameters or parameters of a different sequencing run for `Bustard` base calling will not result in the best

IDT Exhibit 1061.0094

possible base calling results, it will be sufficient to obtain a large enough training data set from spike-in controls. Thus a dedicated control lane can be completely omitted independent of the base composition of the actual libraries being sequenced, making spike-in ϕX control reads valuable for run QC and base calling with quality score calibration while enabling maximum instrument throughput.

4.6 Downstream quality score recalibration

Having different sequencing platforms, different versions of sequencing chemistry and different data production facilities creates a need for assuring data quality consistency. Recently, quality score recalibration based on alignments has been identified as one solution to this problem. Currently the most widely applied algorithm is part of the `Genome Analysis Tool Kit`[5] (`GATK`) [147] and used for studies of the 1000 Genomes project [53]. However, the quality score recalibration based on alignments coincides with steps taken for the quality score calculation in `Ibis` and generally has to be applied with caution:

1. Quality scores as provided by todays base callers are specific to the platform. On 454 the quality score concept breaks with homopolymers which are only one machine signal but several bases when converted to sequence space. On SOLiD quality scores have to incorporate a signal deduced from dinucleotides (i.e two read outs with the chance of one of them being wrong). Illumina base call qualities are actually inferred for a specific base, however they are changing with each software update that provides new quality score tables (determined from runs in Illumina R&D and probably not completely transferable to other instruments and sites due to between instrument, production and handling variance). Even though this argues for the rescoring, it also illustrates that the concept of a quality score for an individual base in a sequence alignment is not a genuine and general concept.

2. Calibration based on alignments with some divergence to the actual sample is problematic. Specifically if the divergence of multiple samples to the reference varies and thus may impact the downstream processing. This might for example cause a biased correction of quality scores and at the end fewer SNP calls for the samples with higher reference divergence.

3. If quality scores are used in an inter-species comparison and genome quality or within species variation varies, then qualities obtained from such a calibration will have a species bias (i.e will be lower for the species with lower reference genome quality or higher diversity).

This argues that quality score recalibration based on alignments can only be done if sequence divergence and population diversity is estimated at the same time. Thus, a recalibration of quality scores should only be done on a common sequence population with no divergence (e.g. spiked control reads of all lanes or control reads in a dedicated lane).

The quality scores that `Ibis` produces are a function of the decision boundary distances obtained when the base intensities are applied to the cycle-specific classifiers (see equation 4.1 in section 4.4.2 on page 79). I model the base quality using a normal approximation of the distances seen in a dedicated test data set for this cycle (i.e these are reads that have not been used for model training, but are then used with the models for testing the prediction

[5] `https://www.broad.harvard.edu/gsa/wiki/index.php/Base_quality_score_recalibration`

IDT Exhibit 1061.0095

performance). Thus, `Ibis` quality scores are normalized in a per run fashion using control reads. They should not require any further between run normalization. However, quality scores might not scale perfectly over the whole `PHRED` score range and a normalization between different sequencing platforms might still be required.

4.7 Summary and conclusions

Even though `Ibis` was originally developed to handle the T accumulation in a sequencing chemistry which has been replaced by several subsequent versions, its application is not limited to the reprocessing of data created with the older chemistries. I have shown that `Ibis` improves the output of sequencing runs from the Genome Analyzer I, which due to their short read length are barely affected by T accumulation but by a generally lower image and sequencing quality. I have also shown that it improves base calling accuracy for runs using recent sequencing chemistries without T accumulation and increased sequencing length. The reason is the sequencing model independent training process of `Ibis`, which only relies on the assumption that the vast majority of the signal needed for base calling is captured by the intensity values of the previous, the current and the next cycle.

The presented approach is unique in that the causes of sequencing error are not modeled separately, but captured by incorporating neighboring signals in the statistical learning procedure. Due to this design, `Ibis` works on a wide range of different sequencing chemistries and platform versions. The performance of `Ibis` on standard hardware is significantly better than for other existing base callers [129], enabling it to be run by research laboratories without access to large computational clusters. The increase in mappable sequences, without ambiguity codes as well as improved and calibrated quality scores enables direct use of the sequences in other software packages. Thus, there is a considerable benefit in investing the computational time in `Ibis` re-base-calling for sequencing runs of all so far available chemistry versions and Illumina sequencing instruments.

IDT Exhibit 1061.0096

Chapter 5

Quantification of gene expression from short-sequence tags

The well-bred contradict other people. The wise contradict themselves.
– Oscar Wilde [96](1205)

The study of gene expression differences at a genome-wide level using microarrays [205] allowed the characterization of human-chimpanzee phenotypic differences at a scale not possible before. The analysis of five tissue transcriptomes showed that differences between humans and chimpanzees relative to variation within species are comparably constant in brain, liver, kidney, heart, but not in testis, in which an excess of differences between species relative to within species was found [111]. In most of the investigated tissues the extent of gene expression divergence between humans and chimpanzees was observed to be proportional to the extent of expression diversity within species. The patterns of transcriptome differences between humans and chimpanzees in testis and brain were, however, identified as not being compatible with a neutral model of evolution. In both tissues it was suggested that positive selective forces have shaped transcriptomic differences. In testis this is supported by an excess of between species differences relative to the diversity within species. In brain fewer differentially expressed genes were observed, and these expression changes seem to have occurred to a larger extent on the human evolutionary lineage than on the chimpanzee lineage [111, 59, 110].

The technological advances in sequencing described in chapter 2 offer the opportunity to complement and extend this already existing body of knowledge with new insights that come from the direct sequencing of transcriptomes. The advantage of such sequencing approaches is that microarrays are static in their design and limited to the analysis of genomic features known at the time of design, whereas sequence-based gene expression profiling represents a shotgun approach to identifying the expressed molecules present in a tissue or sample. Further, the short probes used for hybridization-based technologies are rather sensitive to polymorphisms in the transcribed region that is probed [15, 45] – a problem that is amplified in comparative studies in which expression patterns of species with some evolutionary sequence differences are inferred and compared. In contrast, sequence-based transcriptome profiling technologies generate sequences from all available molecules and in a separate step determine the identity of molecules from their sequence and thus facilitate the measurement and comparison of gene expression from species whose genomes show different levels of genetic divergence.

IDT Exhibit 1061.0097

Gene expression data from humans, chimpanzees, and a genetic outgroup, such as rhesus macaques, allows the triangulation of transcriptomic differences between the outgroup, human and chimpanzee and thus enables differential gene expression to be assigned to the human lineage or to the chimpanzee lineage. Thus, in order to complement previously generated tissue transcriptome profiles from primates [59, 111, 112, 218, 10, 23, 240], we[1] used serial analysis of gene expression (SAGE) in conjunction with Illumina Genome Analyzer sequencing (chapter 2 section 2.3 on page 26) in order to infer gene expression levels. Using the *NlaIII* Digital Gene Expression (DGE) approach, we generated tag sequences of 17bp to study brain, heart, kidney, liver and testis tissues of humans, chimpanzees and rhesus macaques.

The Illumina *NlaIII* DGE protocol was one of the first applications of the Illumina sequencing platform. The first experiments with this protocol were performed at the Max Planck Institute for Evolutionary Genetics in late 2007. At this point 18 cycle/nucleotide reads were the common read length of the instrument. The protocol was discontinued by Illumina in spring 2009, after sufficiently long reads and transcriptome/messenger RNA shotgun sequencing protocols, i.e. RNAseq, had been established for the platform.

5.1 Samples and experimental protocol

Expression data was generated from brain (pre-frontal cortex), heart, kidney, liver, and testis tissues for each five male humans, five male chimpanzees and five male rhesus macaques. For the human samples informed consent for use of the tissues for research was obtained in writing from all donors or their next of kin. All non-human samples were obtained through opportunistic sampling. These individuals suffered sudden deaths for reasons other than their participation in this study. For all samples, the cause of death was unrelated to the tissues used. We note that in human tissues from 25 different individuals were used, while for the two other primates, tissues largely originate from the same donors (each six chimpanzee and rhesus macaque individuals were used across the 5 tissues). We may assume largely unrelated human samples, but chimpanzees and rhesus macaque samples show relatedness to the half- and full-sibling level. Due to the limited access to samples, we could not restrict the analysis to individuals of similar age. Human individuals vary between 5 and 88 years of age, chimpanzees between 6 years and 35 years of age and rhesus macaques between 3 and 9 years of age. Samples are named by a lower case letter h (human), c (chimpanzee) or r (rhesus macaque), followed by a upper case letter B (brain), H (heart), K (kidney), L (liver), and T (testis), followed by numbers 1-5 in human and numbers 1-6 for chimpanzee and rhesus macaque (where numbers are associated with a specific individual).

For the Illumina *NlaIII* Digital Gene Expression protocol, the original SAGE protocol [234], where the most 3' restriction tags from transcripts are ligated, cloned and the concatenated sequence tags read out using Sanger sequencing, was modified and optimized for sequencing on the Illumina Genome Analyzer I platform. This DGE protocol relies on the generation of a library of short (17bp) cDNA tags each corresponding to a sequence located immediately 3' of the 3'-most *NlaIII* restriction site of every transcript in a cell or tissue sample. Figure 3.3 on page 46 in chapter 3 illustrates the sequencing library preparation protocol. Briefly, messenger RNA (mRNA) is isolated from total RNA (1-2μg) by binding to magnetic oligo(dT) beads. The attached mRNA is used as a template for reverse transcription from the oligo(dT) primers, creating a bead-bound mRNA/cDNA hybrid. Next, in the second

[1]The project was designed by Esther Lizano González and Thomas Giger. Esther Lizano González also prepared all sequencing libraries for this project.

IDT Exhibit 1061.0098

strand cDNA synthesis, the mRNA strand is removed and a replacement strand synthesized generating a double-stranded cDNA bound to the oligo(dT) bead. Double stranded cDNA is digested with the restriction enzyme *NlaIII* and all fragments other than the 3' fragment attached to the oligo(dT) beads are washed away. Subsequently an adapter (GEX Adapter 1.2 in figure 3.3 on page 46) is ligated at the site of *NlaIII* cleavage. This adapter ligation creates the recognition sequence for the restriction enzyme *MmeI* at the adapter-cDNA junction. By restriction with the *MmeI* enzyme, which cuts 21bp downstream from its binding site, the resulting construct is no longer attached to the oligo(dT) bead and is free in solution. A second adapter (GEX Adapter 2.1 in figure 3.3 on page 46) is ligated at the site of *MmeI* cleavage. Finally, the adapter-ligated cDNA construct is enriched using PCR and the amplified cDNA construct is gel purified (see also section 3.1 on page 44) prior to sequencing on an Illumina instrument.

5.2 Sequencing and primary data processing

Libraries for all samples were generated in tissue batches, randomizing species in library preparation and sequencing. In early 2008, brain and liver samples were sequenced on a Genome Analyzer I instrument according to the manufacturer's instructions using in total seven different flow cells. Sequencing data was analyzed starting from Illumina 'chastity' filtered `FastQ` files (see also chapter 3 section 3.7 on page 59). For these files, quality scores were converted to `PHRED`-scale [63] and the offset for ASCII encoding changed to Sanger-style (see chapter 4, specifically equations 4.5 on page 85 and 4.2 on page 82). Heart, kidney and testes samples were sequenced on Genome Analyzer II instruments according to the manufacturer's instructions on ten different flow cells from January 2009 throughout August 2009. Here, the sequencing data was analyzed starting from `IPAR 1.01`, `IPAR 1.3` as well as `SCS 2.4/RTA 1.4` sequence files and intensity files. The raw reads of a dedicated ϕX174 control lane on each run were aligned to the corresponding ϕX174 reference sequence to obtain a training data set for the base caller `Ibis` (chapter 4 on page 69, [116]), which was then used to recall bases and quality scores for these Genome Analyzer II runs. `Ibis` could not be applied to the earlier Genome Analyzer I runs as intensity files were not archived for these old runs.

The so-obtained `PHRED`-scaled `FastQ` files were adapter and chimera filtered. Even though, the majority of adapter dimers is removed by the gel excision step, the *NlaIII* DGE protocol may also create adapter chimeras with a length comparable to the targeted library molecules as well as the necessary grafting and priming sites, causing them to be sequenced together with the real SAGE tags. If not removed, these artifacts can, due to their short read length, easily be aligned to the reference genome and cause false positive counts. Thus, reads were filtered for one of the following dimer/chimera sequences that can be created during library preparation and which are the most frequently identified from a representative lane using `TagDust` [128] (see figure 3.3 on page 46): `TCGTATGCCGTCTTCTGCTTG`, `TCGGACTGTAGAACTCTGAAC`, `CAGGTTCAGAGTTCTACAGTCCGACATG`, `GTATGCCGTCTTCTGCTT`, `AATCGTATGCCGTCTTCT`, `TCGGACTCGTATGCCGTC`, `GTCGTATGCCGTCTTCTG`, `TCGGACTGTAGAATCGTA`, `ATGGCTCGTATGCCGTCT`, `AGGAGTCGTATGCCGTCT`, `TCTCGTATGCCGTCTTCT`, `TCGGACTGTAGAACTCTT`, `ATCGTATGCCGTCTTCTG`, `AGGAGTTCGTATGCCGTC`, `CGTATGCCGTCTTCTGCT`, `GTATGCCGTCTTCTTCTT`, `AGTCGTATGCCGTCTTCT`, `CCAGTCGTATGCCGTCTT`, `GTGATCGTATGCCGTCTT`, `GTGTCGTATGCCGTCTTC`, `TCGGACTGTAGATCGTAT`, `GCCACCCTCTACAGCCGA`.

Reads were also filtered for the presence of at least 9nt of the adapter sequence (`TCGTATGCCGTCTTCTGCTTG`) at the read end. Further, the last read base (the first adapter base) of

IDT Exhibit 1061.0099

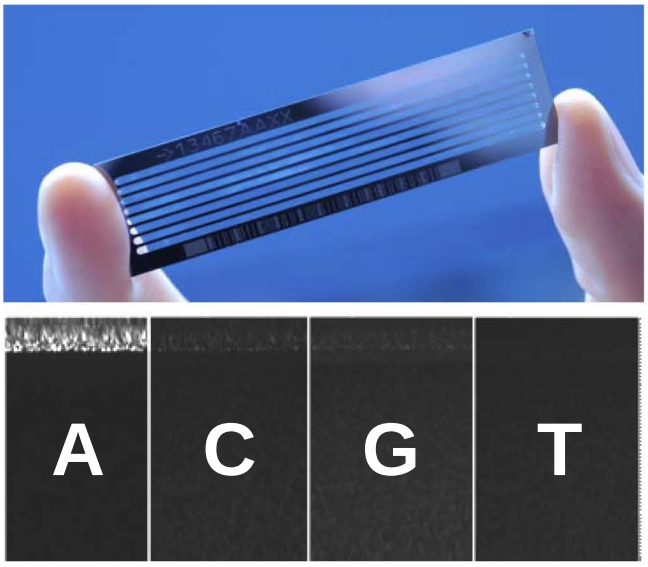

Figure 5.1: A reflection layer surrounding the actual lanes of version 1 flow cells (mostly used with Genome Analyzer I instruments), caused the image analysis algorithm to pick up artificial clusters at the tile edges. Sequences obtained from these artifacts are frequently adenine homopolymers. Upper picture of flow cell taken by Frank Vinken, MPG.

the remaining 18nt reads was clipped off and reads with more than 1 base below a quality score of 10 discarded. Subsequently, reads with sequence entropy below 0.85 were removed, as these might also originate from flow cell artifacts (see chapter 3 section 3.7 on page 59). Specifically, flow cells used with Genome Analyzer I had a reflection layer surrounding the actual lanes (figure 5.1), which caused the image analysis to identify a larger proportion of such artifacts at the outer tile edges. Later flow cell versions are transparent to reduce this effect.

Figure 5.2 on the next page provides the fraction raw sequences remaining after adapter chimera and dimer removal for the different tissue batches as well as the fraction of adapter filtered sequences remaining after quality and complexity filtering. For the kidney and heart batches some samples showed a large number (up to 40%) of adapter dimers/chimeras. Data quality also shows considerable variation between batches, supporting the need for the applied quality filter.

5.3 Tag alignment

The ultra-short sequences obtained from this *NlaIII* DGE protocol complicate data analysis. Assuming that all sequences remaining after read processing originate from the library preparation protocol, 21nt long tag sequences (`'CATG'` + 17nt) have to be aligned to at least three gigabase-sized mammalian genomes. Due to sequencing error in these tag sequences and genomes as well as due to biological variation in the restriction sites, exact matches are not necessarily expected. However, the number of putative *NlaIII* restriction sites and the uniqueness of the resulting tag sequences in the human, chimpanzee and rhesus macaque genomes also varies (table 5.1 on the facing page). Variation in genome quality and genome completeness are likely to contribute to these differences.

Different processing versions of the data have been considered and partially also tested.

IDT Exhibit 1061.0100

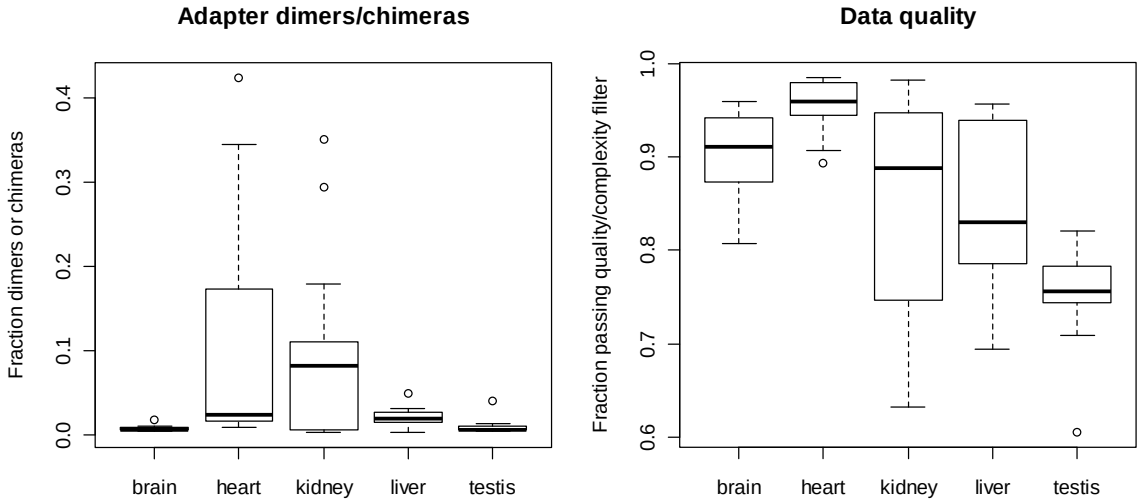

Figure 5.2: Summary statistics from the raw data processing of the Illumina DGE data by tissue. Shown are boxplots for the fraction adapter dimers and chimeras removed from the raw reads **(left)** as well as the fraction of adapter filtered reads passing the described quality and complexity filters **(right)**. Libraries for all samples were generated in tissue batches, randomizing species in library preparation and sequencing. In early 2008, brain and liver samples were sequenced on a Genome Analyzer I instrument, heart, kidney and testes samples were sequenced on Genome Analyzer II instruments from January 2009 throughout August 2009.

Table 5.1: The number of *NlaIII* restriction sites varies for the three primate genomes (human, chimpanzee, rhesus macaque), as does the fraction of unambiguous tag sequences obtained from these sites for the three different genomes. It is likely that genome quality and genome completeness contribute to these differences. Judging from these numbers some proportion of sites with identical tag sequences occur with high frequency throughout the genome.

	`'CATG'` sites	Tag sequences	Unambiguous	Ratio
Human	27,561,798	20,418,202	19,348,954	70.20%
Chimp	27,861,673	20,414,200	19,123,594	68.64%
Rhesus	24,847,282	19,485,936	18,753,873	75.48%

IDT Exhibit 1061.0101

It was, for example, discussed whether sequences should be aligned to a virtual and pre-annotated tag library or whether sequences should rather be aligned to the complete genome of each species. For generating a virtual tag library, each reference genome is scanned once and, for all *NlaIII* restriction sites, the 21nt sequence corresponding to the *MmeI* digestion product extracted. The tag library can be annotated with gene annotation. Sequencing tags can be aligned with this static virtual library using, for example, a tool that is fast for near-perfect matches (e.g. `PatMan` [180]). Illumina followed such an approach in their analysis pipeline for the DGE protocol, for which they provided pre-compiled virtual tag libraries for the mouse and human genome and aligned using `ELAND`. For this five tissues project, I used a similar approach for the first analyses. However, such a virtual tag library assumes that there is no sequence variation in the *NlaIII* restriction sites and that further these sites were identified without error from the corresponding reference sequence. Considering that a mammalian genome contains about 30 million *NlaIII* restriction sites (table 5.1 on the previous page) and considering typical diversity across human populations (which is lower than across different chimpanzee populations, [68]) about 0.14% (42,000) of these sites will differ between any two individuals.

In another approach, the recognition motif of *NlaIII* (`'CATG'`) can be attached to beginning of all sequences and the full SAGE tags aligned to the corresponding reference genome. While this may overcome the variation in the restriction sites between individuals, at least for tag sequences which are unique and sufficiently informative for one genomic position when allowing mismatches, it does not overcome the general problem with these short tags which will for two genomic sites frequently only differ in one substitution. Considering that sequencing error as low as 1% will generate about one difference in every sixth tag, this effect is considerably larger than variation in *NlaIII* sites. There is only one way to counteract this short-tag ambiguity problem: including an assumption that limits the putative sources/regions for the origin of these tags.

Based on the experimental protocol enriching for 3' most restriction sites of transcripts, one such assumption would be that tags may only originate from the last 1,000nt of known transcripts (assuming rather equal spacing of the restriction sites, complete transcripts and complete knowledge of all transcripts in the cell). Considering about 79,000 mRNAs annotated in `Ensembl v59`, this assumption would reduce the search space to less than 3% of the complete genome. Alternatively, the alignment could also be performed to sets of known full transcripts, which limits the search space to also about 3% of all genome bases. Further, aligning to actual transcript sequence has the advantage that tags spanning exons could be handled correctly and tags extending into a polyadenylation site predicted. However, with this strategy the vast majority of the genome is not considered during alignment and tags actually originating from outside annotated transcripts will likely be aligned with few sequence differences to known transcripts. This is possibly a concern when considering that projects like the Encyclopedia of DNA Elements (ENCODE) have shown that essentially the whole genome is transcribed, even though largely varying in the number of transcribed molecules [226].

In the final analysis, I extended sequences by including the `'CATG'` restriction motif and aligned them to the complete genomes (hg19 / GRCh37 excluding additional haplotypes, pantro2 / CGSC 2.1 and rhemac2 / MGSC Merged 1.0 with the addition of a mitochondrial genome sequence from `Ensembl MMUL_1.54` build) using `bowtie 0.12.4` [126]. `Bowtie` is a fast short-read aligner which allows substitutions but no insertion/deletions to the reference genome. This aligner was configured to report up to 100 equally best alignments (with up to two mismatches) and to discard reads that produce more than 100 alignments (parameters `-a -m 100 --best --strata`). Therefore, high frequency tags are excluded from analysis,

IDT Exhibit 1061.0102

while tags appearing at most 100 times in the genome were reported for each of the genomic positions. Alignments outside of gene annotation are excluded from quantification.

5.4 Annotation of expressed sequence tags

Typically gene annotation can be obtained from one of the genome browsers or bioinformatics databases maintained by the National Center for Biotechnology Information (NCBI), the University of California, Santa Cruz (UCSC) or the European Bioinformatics Institute (EBI)/European Molecular Biology Laboratory (EMBL). These databases differ in which gene prediction routines are used and how experimental transcript evidence is incorporated. We used the `Ensembl`[2] database at EBI/EMBL which provides gene annotation for all three species as well as an assignment of orthologous genes across species [43]. When the first experiments of this project were analyzed the `Ensembl` database was in version 50 (released July 2008). When analyzing one of the human brain samples, the observed tag site distribution along genes and neighboring sequence shown in figure 5.3 on the following page was obtained.

The figure does not consider actual tag frequency, but simply plots the position of tag sites observed in this sample and is therefore independent of the actual gene expression values. As expected for a protocol targeting the most 3' restriction site in each transcript, the vast majority of tag sites is observed right before the gene end. However, for some genes, tags fall right outside the 3' gene end, indicating incomplete annotation of 3' untranslated regions (3' UTRs) in this `Ensembl` database version. In addition, a higher frequency of tag sites is observed at the gene start position, which also extends into sequence upstream of the actual gene. The presence of these 5' tags might indicate incomplete mRNA extension. The extension of this effect into upstream sequence is likely caused by incomplete annotation of 5' untranslated regions (5' UTRs). Further, a mirroring of tag positions within genes is observed on the antisense strand of genes. Even though antisense transcription has been described [247, 108, 127], the exact clustering of their 3' ends with the 3' ends of the sense transcripts is unexpected.

When exploring the position of tags on the antisense strand, I find that most of them even mirror counts for the same *NlaIII* restriction site on the sense-strand, indicating that they are likely to originate from a carry-over effect rather than actual antisense transcripts. Carry-over of the upstream *NlaIII* digestion fragments on the tube walls and beads will allow adapter ligation and propagation in the protocol; creating tags that seem to originate from the opposite strand. More stringent wash protocols after *NlaIII* digestion might reduce this effect. Figure 5.4 on page 105 shows for *NlaIII* restriction sites covered with tag counts on both strands the Spearman correlation across all samples and compares it to the correlation when randomizing the assignment of sense to antisense tags.

Carry-over of upstream *NlaIII* digestion fragments may also cause non-3'-most tags to be observed from both strands of the cDNA. Further, the incomplete digestion or enzyme hindrance at close restriction sites[3] may also cause the distribution of the actual count signal over multiple restriction sites. Figure 5.4 on page 105 shows a clear correlation of counts for neighboring sites. When analyzing very close *NlaIII* restriction sites where a second or even third 'CATG' is within the 21 bases obtained from *MmeI* digestion, the spread of tag counts across the different sites due to hindrance and incomplete digestion can be observed (figure 5.5 on page 106). The effects of incomplete digestion and steric hindrance can be

[2] `http://www.ensembl.org/`

[3] In the human genome 5.5% of all tag sites contain a second *NlaIII* restriction site within their sequence.

IDT Exhibit 1061.0103

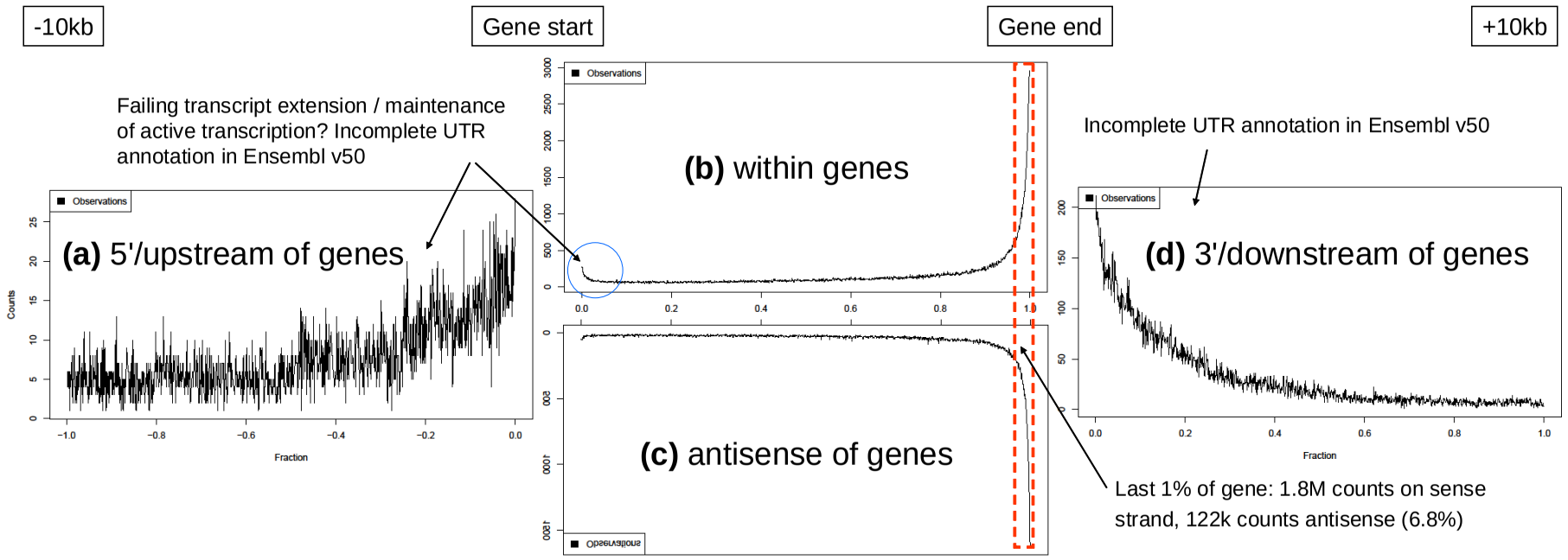

Figure 5.3: Observed tag site distribution along genes **(b+c)** and neighboring sequence (10kb upstream of the gene start **(a)** and 10kb downstream of the gene end **(d)**). This figure does not consider actual tag frequency, it shows the distribution of observed tag sites. Due to the different gene lengths, tag positions are plotted as fractions of the total length. As expected for protocols targeting the most 3' restriction site in each transcript, the vast majority of tags sites is observed right before the gene end **(b)**. However, for some genes tags fall right outside the gene end **(d)**, indicating potentially incomplete annotation of 3' untranslated regions. Further, a mirroring of expressed tag positions within genes can be seen on the antisense strand of genes **(c)**. In addition, a higher frequency of tags is observed around the transcription start site **(b)** and extends to the upstream region of genes **(a)**, indicating incomplete annotation of 5' untranslated regions.

IDT Exhibit 1061.0104

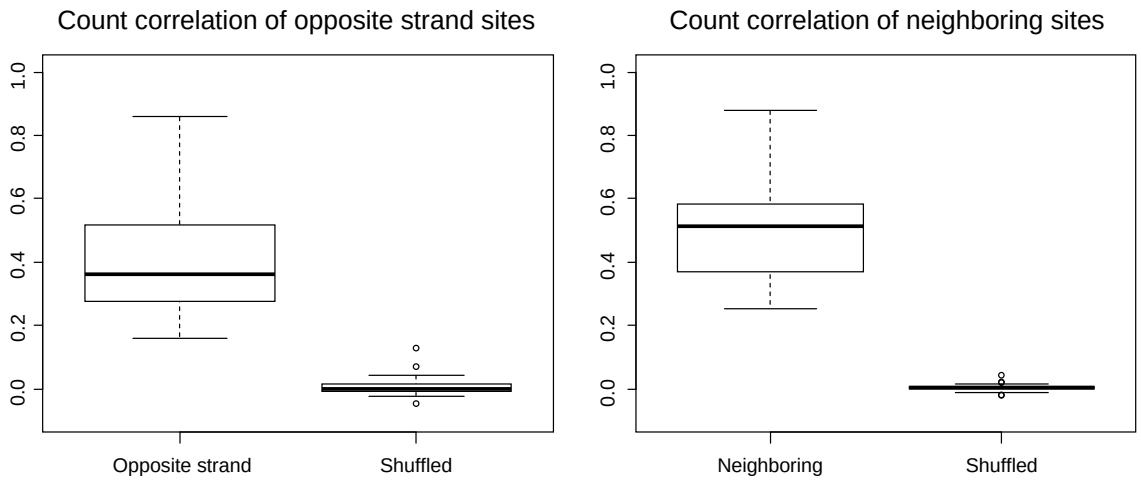

Figure 5.4: Boxplot of correlation values across all samples for tag counts orignating from the two strands of the same *NlaIII* restriction site **(left)** or neighboring *NlaIII* sites on the same strand **(right)**. For unique *NlaIII* restriction sites covered with tags on both strands the Spearman correlation of tag counts for the two strands is compared to the Spearman correlation when randomizing the assignment of sense and antisense tags. For neighboring unique *NlaIII* sites, observed together in the same sample, the Spearman correlation of these counts is compared to the Spearman correlation when randomly reassigning neighboring *NlaIII* sites.

compensated for by summing over all tags in a gene. However, when summing all tags in a gene, the carry over effects for upstream *NlaIII* digestion fragments and incompletely extended transcripts may cause a gene length bias in gene expression quantification as well as a noisy gene expression read out due to variance in the extend of carry over between experiments. Such a length bias is problematic when gene expression values are compared across genes, but not when the same genes are compared between samples.

Due to the observation of incomplete annotation of 3' UTRs, we considered extending genes by a fixed number of bases after their annotated end. However, while further data was generated and different analysis approaches tested, the `Ensembl` gene annotation was updated several times. The updates from `Ensembl v50` to `Ensembl v59` resulted in about 1.5kb increase in human median gene length (see table 5.2 on the following page). The annotation of chimpanzee and rhesus macaque did not change at all between these versions.

The number of tags mapped within 1kb downstream of 13,387 human protein-coding genes decreased by up to 65.7% between `Ensembl v50` and `Ensembl v59`. The exact number differed depending on the tissue with a reduction by 36.1% in brain, 47.1% in heart, 65.7% in kidney, 51.6% in liver and 40.2% in testis observed for the newer annotation. Since human gene annotation improved, the extension of genes for missing 3' UTR annotation was no longer necessary. So I tried applying the species-specific annotation available from `Ensembl v59`, limiting to genes with orthologous gene identifiers uniquely assigned between all three species. Figure 5.6 on page 107 shows the results when averaged over all brain samples for the three species. Only the recent human genome annotation builds provide sufficient annotation of UTR sequences. The median gene length of protein-coding genes in chimpanzee and rhesus macaque is considerably shorter (table 5.2 on the following page). Thus, gene annotation including UTRs has to be transferred from the human genome to both other species to prevent an annotation bias.

Gene annotation can be transfered by either aligning the individual gene or transcript sequences to the other genomes or by transfering gene coordinates using pair-wise whole

IDT Exhibit 1061.0105

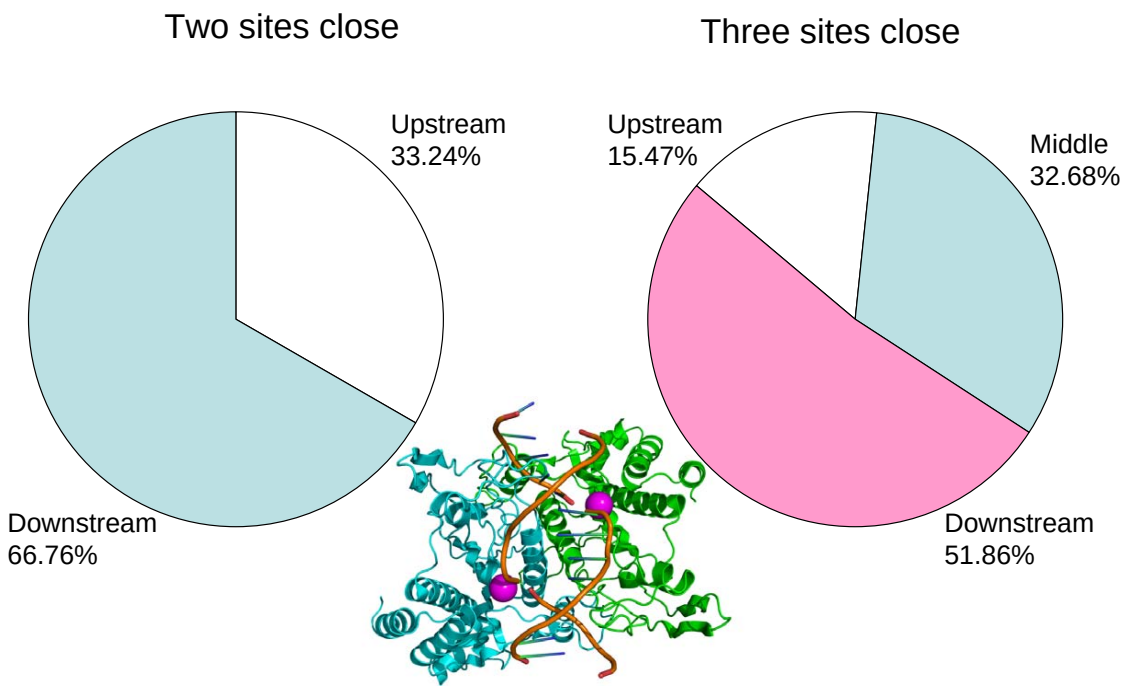

Figure 5.5: For very close *NlaIII* restriction sites, where a second or third 'CATG' is within the 21 bases obtained from *MmeI* digestion, the distribution of transcript counts across multiple tags due to hindrance and incomplete digestion can be observed. *NlaIII* is a Type-2 restriction enzyme with a four base pair recognition motif of which no crystal structure is currently available. However, *EcoRI*, also a Type-2 restriction enzyme with a four base recognition motif, covers a full turn of the DNA double helix (about 10nt). The enzyme image for *EcoRI* was downloaded from `Wikipedia` (`http://en.wikipedia.org/`) as `1QPS.png`, where it was released into public domain by the author.

Table 5.2: Gene length of protein-coding genes in `Ensembl v50` and `Ensembl v59`. Human median gene length increased by about 1.5kb in the newer version and at least one suspiciously long gene was removed. The annotation of chimpanzee and rhesus macaque did not change between these versions.

`Ensembl v50`							
	Genes	Min.	25th	Median	Mean	75th	Max.
Human	21,528	62	6,416	20,040	58,220	55,380	50,940,000
Chimp	19,829	59	6,024	19,970	54,850	55,240	2,341,000
Rhesus	21,905	62	3,014	14,450	45,530	46,090	2,034,000
`Ensembl v59`							
Human	21,727	8	6,999	21,500	59,150	60,390	2,305,000
Chimp	19,829	59	6,024	19,970	54,850	55,240	2,341,000
Rhesus	21,905	62	3,014	14,450	45,530	46,090	2,034,000

IDT Exhibit 1061.0106

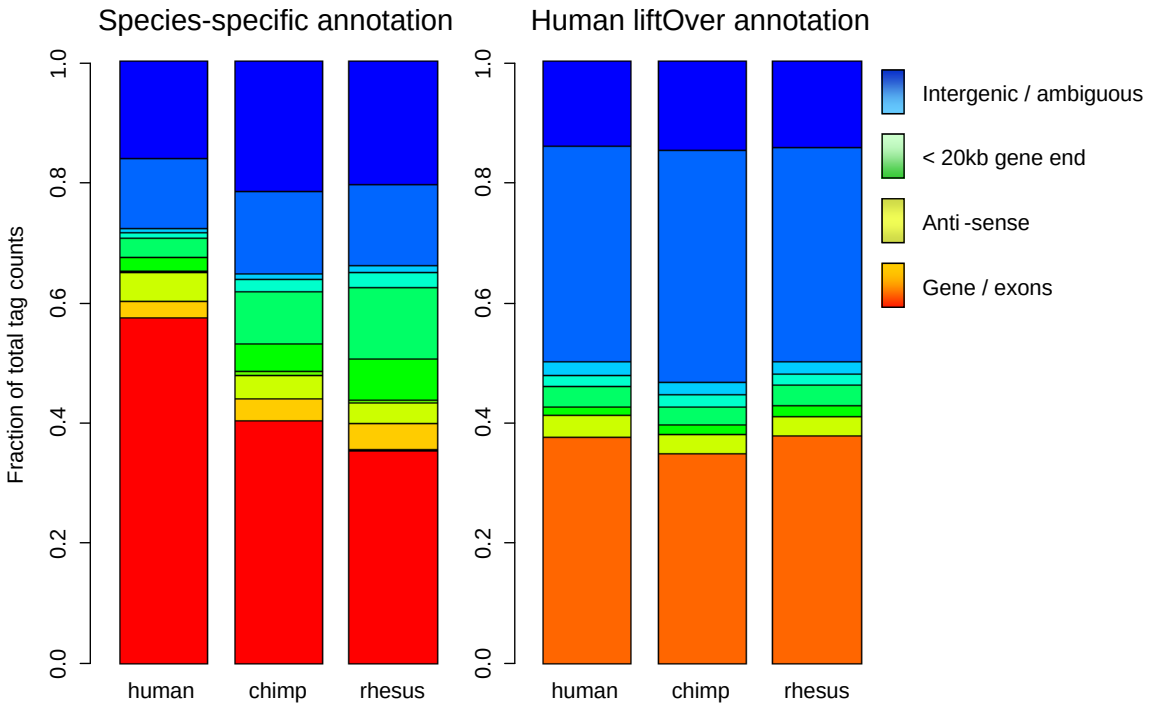

Figure 5.6: Annotation of expressed sequence tags using species-specific `Ensembl v59` annotation **(left)** and human `Ensembl v59` annotation projected to the other two species **(right)**. Chimpanzee and rhesus macaque annotation largely miss annotation of 3' UTRs and thus show a larger proportion of tags falling downstream of actual gene annotation. Thus, a smaller proportion of counts is considered in genes for those two species. When gene annotation is projected from human to chimpanzee and rhesus macaque, very similar proportions of tags are assigned to genes in each species.

IDT Exhibit 1061.0107

Table 5.3: Results from projecting human `Ensembl v59` gene annotation to chimpanzee and rhesus macaque. Counts for transfer RNAs only include the ones encoded in the mitochondrial genome, cytoplasmic transfer RNAs encoded in the nuclear genome do not have an explicit biotype in `Ensembl`. While nuclear genes were projected using UCSC's `liftOver` tool, all 37 mitochondrially encoded genes were projected using a `ClustalW` mitochondrial genome alignment of the three species.

`Ensembl` biotype	Human	Chimp	Fraction	Chimp & Rhesus	Fraction
protein-coding genes	22,099	17,989	81%	14,682	66%
pseudo genes	12,599	10,179	81%	6,858	54%
ribosomal RNAs	537	469	87%	365	68%
transfer RNAs	22	22	100%	22	100%
sn/sno/miRNAs	5,283	4,759	90%	3,804	72%
lincRNAs	1,451	1,257	87%	891	61%
other	9,724	8,345	86%	6,637	68%
total	51,715	43,020	83%	33,259	64%

genome alignments between the reference genomes. Such whole genome alignments are for example provided by UCSC[4] [190], generated using their `blastZ` and `autoMultiZ` pipelines[5], or available from `Ensembl`, using the `Enredo-Pecan-Ortheus` (`EPO`) [171, 170] pipeline. I have used the UCSC whole genome alignments, more precisely the `liftOver chain` files. These pre-generated files are available for selected assemblies and can be used with genome coordinates in `BED`[6] format using the UCSC command line tool `liftOver`. These `chain` files are based on whole genome alignments which differentiate between a reference and a target genome in the generation process, requiring a position of the reference genome to appear at most once while sequences from the target genome can be used multiple times. Therefore, `liftOver` output is non-symmetrical between species.

I used `liftOver` to project gene start and end coordinates from human to the other two species and projected coordinates back from chimpanzee and rhesus macaque to validate that they match the original coordinates of the human genome. Gene start and end coordinates were projected separately, requiring the projected coordinates to be on the same chromosome and strand. Further, the new gene length was not permitted to change by more than a factor of 10, or at most 100kb. The rhesus genome does not include a mitochondrial genome in the assembly version provided by UCSC. Hence, also the whole genome alignments do not include it and therefore `liftOver` fails for mitochondrial genes.

I therefore supplemented the whole genome alignment by aligned mitochondrial genomes of the three species generated using `ClustalW` [227], manually "synchronizing" mitochondrial genome start positions between human, chimpanzee and rhesus macaque and then transferred coordinates of 37 mitochondrially encoded genes using this alignment. The `liftOver` gene annotation projection step shows different efficiency for different gene classes, especially when projecting over the evolutionary larger distance to rhesus macaque (table 5.3). When projecting to chimpanzee 83% of genes are recovered, when projecting to chimpanzee and rhesus macaque only 64% of genes remain. Applying annotation for the genes which can be successfully projected, similar proportions of tags are assigned to genes in each species (figure 5.6 on the preceding page).

[4] `http://hgdownload.cse.ucsc.edu/downloads.html`
[5] part of `Kent-tools`, `http://hgdownload.cse.ucsc.edu/admin/exe/`
[6] `http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#BED`

IDT Exhibit 1061.0108

5.5 Gene quantification and count normalization

For quantifying gene expression, the `bowtie` output files from each lane were combined by individual (as samples hK4, rK4, rK5, rK6, cL1, cL2, hL1, rL2, rL3 and rL4 were pooled from two lanes) and mapped tags sorted by coordinates. Sites observed fewer than three times were removed to reduce the effect of adding up erroneous alignments in long transcripts, and per-site-count files created. Some individuals with similar number of reads in the same tissue ended up with fewer observed sites. This could indicate that part of the transcriptional complexity of those samples was lost during library preparation.

I therefore required that no sample had fewer observed sites than 50% of the median value of observed sites for all samples of the same tissue. This excluded samples hH5, hK2, hK3, cH5, cK5, and rK5 which did not cluster with samples of the same tissue in a principal component analysis (PCA) of raw gene counts (figure 5.7). Another sample which did not cluster with samples from the same tissue is rT2, a sample which did not show any abnormalities in experimental logs and sequencing statistics. It is the testis sample of an 4.85 years old rhesus macaque, even though this is only the second youngest rhesus macaque sample the differences in gene expression may be explained by a late onset of sexual maturity, which might not be uncommon in rhesus monkeys. Rhesus macaques typically reach puberty after 3-3.5 years of age, but show a wide age range for the birth of their first offspring (3.8-18.8a), with only 26% of males reproducing before reaching the age of 6 years [17]. Hence, this sample was excluded from further analyses.

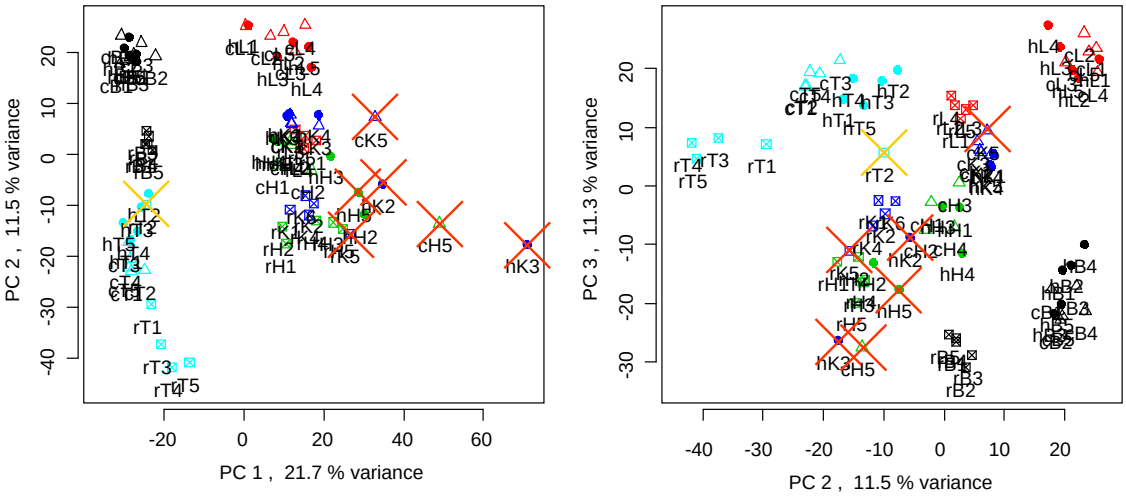

Figure 5.7: First three principal components obtained from the analysis of normalized gene expression counts for all 75 individuals across five species and five tissues (prior to variance stabilization using `DESeq`). Six outliers excluded due to a low number of covered genomic sites are marked with red crosses (hH5, hK2, hK3, cH5, cK5, and rK5). rT2 is marked with a yellow cross and was also excluded. This is a rhesus macaque testis sample which does not cluster with the other testis samples. Even though it did not show any abnormalities in experimental logs and sequencing statistics, it is possible that sexual maturity was not reached by this individual.

The per-site-count files for the remaining individuals were overlaid with the reciprocally projected `Ensembl v59` human gene annotation. Within each gene I summed all counts on the correct strand. For genes overlapping on the same strand, I divided counts between the two genes after all non-overlapping counts had been considered. The overlapping counts were split between genes based on either the ratio of the non-overlapping counts (if the overlapping counts are lower) or evenly split between genes otherwise. Again, genes receiving fewer than

IDT Exhibit 1061.0109

three counts were discarded.

The outlined approach considers tag sequences appearing in multiple genes, but at most 100 times in the complete genome, for the counts of multiple genes. The proportion of genes that are quantified by unambiguous counts only, depends on the exact tissue and species. As little as 12.7% (rhesus macaque kidney) and up to 41.2% (human liver) of genes are quantified from tags unique to one gene. When requiring at least 50% of the counts to originate from unambiguous tag sequences, at least 60.4% and at most 69.1% of genes remain, with the extremes being observed in the previously-mentioned tissues. Table 5.4 on the facing page provides a summary for different cutoff values, tissues and all three species. It is worth noting that these numbers include the removal of about 34% of human genes due to the annotation projection to rhesus macaque. Species differences observed for the different cutoffs, indicate that any kind of filters on uniqueness of tag sequences might introduce (further) species biases. This exemplifies why short-tags and the associated alignment ambiguity are so problematic for data analysis.

Instead of requiring some proportion of counts to be unambiguous for a gene, one might also quantify genes only from unambiguous counts – ignoring the proportion of tags excluded and accepting that occasionally tags quantifying the dominant transcript version are lost. Both (counting tags multiple times or excluding them from quantification) generate an inaccurate quantification result. While counting sites multiple times will overestimate gene expression, removing them will underestimate gene expression (figure 5.8). This alone would not be problematic for an inter-species study as identifying gene expression changes between species and tissues does not require a correct ranking of gene expression, however table 5.4 on the facing page indicates that different tag frequencies for the three genomes (see also table 5.1 on page 101) impacts quantification in a species-specific manner. Since already one cutoff, even though not stringent, has been applied for tag sequences occurring more than 100 times in the genome, additional filters may further put off expression quantification between species.

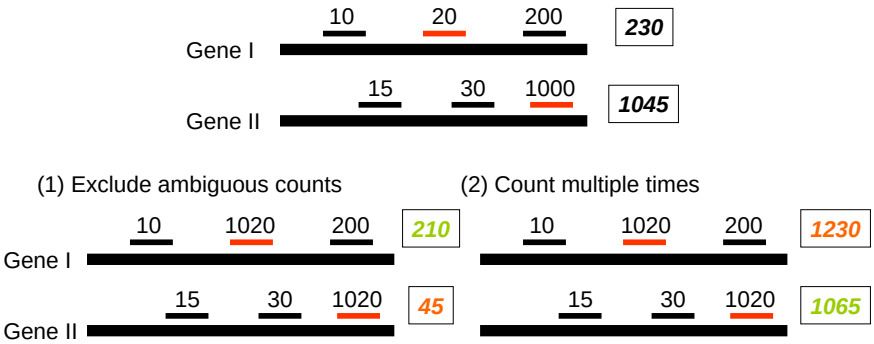

Figure 5.8: Both excluding ambiguous tags, or counting tags multiple times, results in gene quantification that is either too low or too high. The presence of ambiguously mapped tag sequences within transcribed regions means that no correct ranking of gene expression values can be obtained from this data set.

Table 5.5 and figure 5.9 look at the differences between the two extremes in gene quantification in presence of ambiguous tags: excluding all unambiguous tags or counting them multiple times (with the restriction that tag sequences more frequent than 100 times in the genomes were excluded in mapping). Spearman correlation between both estimates is high, showing that gene ranks are mostly maintained (figure 5.9 on page 112). Results for differential expression between human and chimpanzee show some variation (except for testis) between the two types of analysis (table 5.5 on page 112). However, the relative ordering of tissues in the number of expressed genes as well as the fraction of differentially expressed

IDT Exhibit 1061.0110

Table 5.4: Fractions of expressed genes remaining for each tissue when requiring different proportions of the gene expression counts to originate from unambiguous tag sequences (column 'Purity'). Species and tissue differences can be observed for different cutoff values, e.g. in kidney and liver. The differences observed indicate that filters on uniqueness of tags might introduce a species and tissue bias.

Tissue	Purity	human	chimp	rhesus
brain	100%	32.9%	34.6%	30.4%
	95%	43.9%	44.5%	41.6%
	90%	49.7%	50.1%	47.7%
	85%	53.7%	53.8%	51.6%
	75%	59.3%	59.2%	57.2%
	50%	67.9%	68.1%	66.6%
	25%	74.2%	74.5%	73.6%
heart	100%	25.7%	22.7%	25.4%
	95%	38.2%	37.0%	37.2%
	90%	43.2%	42.6%	42.5%
	85%	46.9%	46.5%	46.3%
	75%	52.5%	52.6%	52.4%
	50%	63.4%	64.1%	64.1%
	25%	73.0%	74.5%	74.8%
kidney	100%	19.1%	16.0%	12.7%
	95%	33.9%	31.6%	28.8%
	90%	40.4%	38.5%	35.8%
	85%	45.0%	43.2%	40.4%
	75%	51.8%	50.2%	47.2%
	50%	63.6%	62.8%	60.4%
	25%	74.2%	74.5%	73.9%
liver	100%	41.2%	36.1%	35.6%
	95%	48.3%	43.4%	43.1%
	90%	52.6%	47.9%	47.6%
	85%	55.6%	51.2%	50.7%
	75%	60.0%	56.1%	55.6%
	50%	67.6%	64.2%	64.0%
	25%	73.0%	70.7%	70.9%
testis	100%	20.1%	17.6%	20.9%
	95%	35.7%	35.0%	37.1%
	90%	43.3%	42.9%	44.2%
	85%	48.5%	48.2%	48.9%
	75%	56.1%	55.6%	55.7%
	50%	69.1%	68.7%	68.2%
	25%	79.3%	79.7%	78.5%

IDT Exhibit 1061.0111

genes is identical.

Table 5.5: The treatment of ambiguous tags affects the number of expressed genes and the fraction of differentially expressed genes (p-value cutoff of 0.01 after Benjamini-Hochberg FDR correction, see section 5.6 on the next page for details). The relative ordering by the number of expressed genes (≈ transcriptional complexity) or by the fraction differentially expressed (≈ gene expression divergence) is however stable with respect to this difference in data processing.

Counting multiple times				Excluding ambiguous counts			
Genes	Diff.Exp.	Tissue	Fraction	Genes	Diff.Exp.	Tissue	Fraction
10646	1387	brain	13.0%	8765	870	brain	9.9%
8361	224	heart	2.7%	6053	82	heart	1.4%
13401	2040	kidney	15.2%	10661	1256	kidney	11.8%
8285	742	liver	9.0%	6480	349	liver	5.4%
14056	4109	testis	29.2%	11531	3353	testis	29.1%

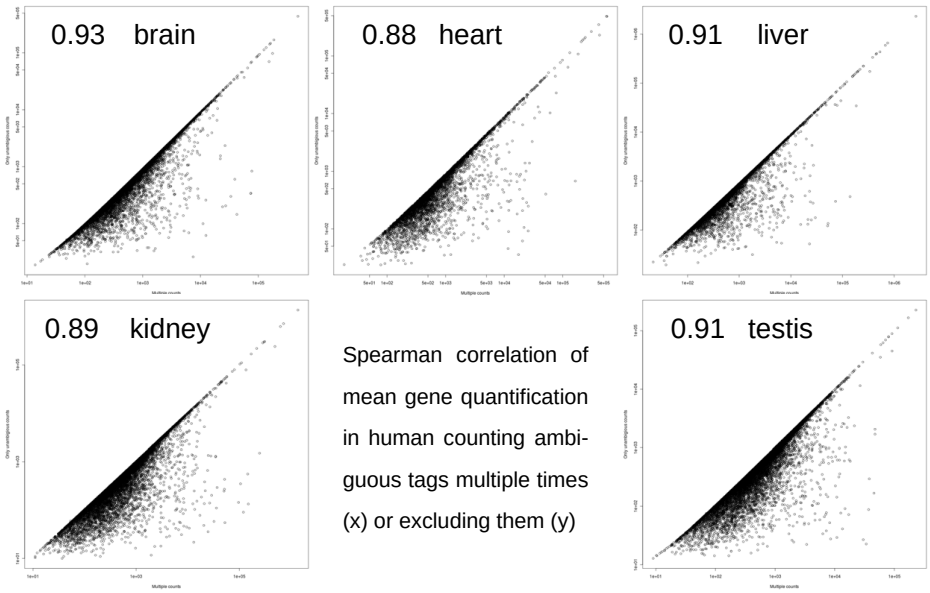

Figure 5.9: Spearman correlation of gene expression quantifications for human data when ambiguous tags are excluded (y axis) or counted multiple times (x axis). When excluding ambiguous counts gene expression values are underestimated, when counted multiple times gene expression values are overestimated (see also figure 5.8 on page 110). High Spearman rank correlations show that gene ranks are mostly stable with respect to this difference in data processing.

From the above results, specifically table 5.4 on the preceding page, it seems likely that excluding ambiguous counts would cause species-specific effects and thus negatively impact an inter-species study. I therefore restrict further analyses to the "multiple counting" approach to quantification. From these counts, the 25th and 75th percentiles of gene counts for all samples belonging to the same tissue (independent of species labels) were adjusted (scaled using a linear function to the same values) to correct for the different number of reads for each sample. Between tissues the median of medians, i.e. the median of each of the 5 tissue groups, was adjusted (i.e. scaled) to the maximum value observed across the five tissue groups. The estimates for the normalization factors and offset were based on protein-coding/non-mitochondrial genes only, but then applied for all genes. Protein-coding genes are the main target of the poly-A capture used in the DGE protocol, and mitochondrial genes are expected to show higher variance in their expression due to varying mitochondrial

IDT Exhibit 1061.0112

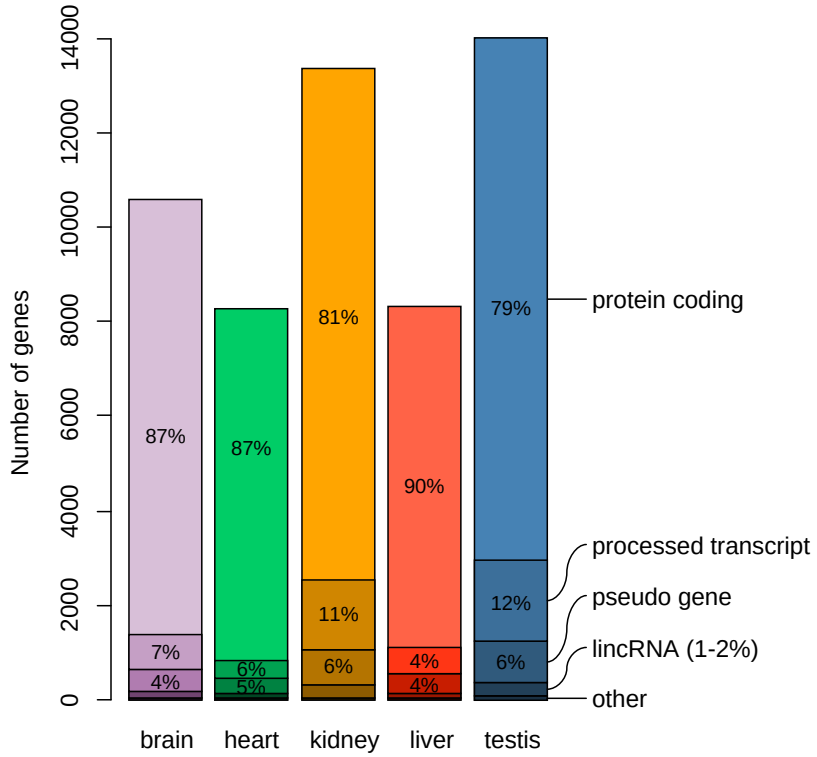

Figure 5.10: Number of expressed genes by tissue and `Ensembl` biotype. Within each tissue, genes not expressed in all individuals across the three species were excluded. Further, after mapping tag sequences observed fewer than three times have been excluded. Therefore, expressed genes in one tissue are supported by reads in each individual. Tissues with the lowest number of expressed genes are liver and heart; brain, kidney and testis show more expressed genes, with kidney and testis also having higher proportions of non-coding genes.

genome copies per diploid nuclear genome.

As required for packages used in downstream analysis (section 5.6), gene expression counts were converted back to integer values (with the minimum observed count being set to 1). Further, within each tissue group, genes with missing observations, i.e. genes not expressed in all individuals, were excluded. Excluding genes not observed in all individuals across species, assures that quantification of the gene is not hampered by sequence-specific effects (e.g. the restriction site falling outside of the gene due to a polymorphism) and indirectly serves as a lower expression value cutoff.

After this last step, the number of expressed genes is defined for each tissue. Figure 5.10 summarizes these numbers and provides the most frequent `Ensembl` gene biotypes in the different tissues. Tissues with the lowest number of expressed genes are liver and heart with 8,285 and 8,361 expressed genes, respectively. Brain (10,646), kidney (13,401) and testis (14,056) have more expressed genes, with kidney and testis also having higher proportions of non-coding genes. The exact numbers may vary due to the counting of ambiguous tags multiple times which may inflate numbers.

5.6 Differentially expressed genes

After applying the described processing steps, we obtain count data for each gene and individual. For testing whether counts observed in one species are different from counts ob-

IDT Exhibit 1061.0113

served in the other species, one needs to estimate the variation in gene quantification in each species and ask whether the obtained distributions are largely, i.e. some p-value cutoff, non-overlapping. We have counts, which relate to tags sampled from a much larger tag population. From random sampling the tag counts should follow a multinomial distribution, which can be approximated by a Poisson distribution. Previously tests based on the Poisson distribution (e.g. [145, 238]) have been used for testing differential expression. However, the single parameter of a Poisson distribution, its mean which equals its variance, seems to generate distributions with a variance too small for the variation seen in real data [193, 157]. To address this so-called overdispersion problem, sequencing count data can be modeled with negative binomial distributions [194, 5, 192].

Negative binomial distributions have two parameters (mean and variance), which need to be estimated for each individual gene from the data; data that frequently has too few replicates for a reliable per-gene estimate of these parameters. Therefore, Robinson and Smyth proposed [194], and also implemented in the `bioconductor R` [73, 224] package `edgeR` [192], the assumption that mean and variance should be related by $\sigma^2 = \mu + \alpha \cdot \mu^2$, with α being constant for the whole experiment. Therefore, effectively only one parameter needs to be estimated for each gene, allowing application to experiments with small numbers of replicates. `DESeq` [5], a more recently released `bioconductor R` package, extends this model to a variance parameter that is related to the mean expression value of the gene by a smooth function. Using this assumption, observations for genes with similar expression values can be pooled for parameter estimation and experimental effects impacting the variance of highly expressed genes differently than the variance of lowly expressed genes can be modeled.

I used the processed primate gene expression count data with the `DESeq`[7] [5] package to obtain variance stabilized expression values as well as calling differentially expressed genes between species pairs for each tissue. Due to the normalization procedure described before, the internal count normalization of the `DESeq` package was not used. In order to account for multiple hypothesis testing when identifying differentially expressed genes, p-values were Benjamini Hochberg [14] corrected and a false discovery rate adjusted p-value cutoff of 1% applied. Further, as suggested by the `DESeq` manual, differentially expressed genes with residuals from the estimated variance function greater or equal to 15 were excluded. Table 5.6 on the next page provides a summary with the fraction of differentially expressed genes identified in all pair-wise species tests for each tissue. Based on these results, gene expression between human and chimpanzee changed most in testis, followed by kidney, brain and then liver. Fewest changes in gene expression are observed in heart. When comparing human and chimpanzee to rhesus macaque, the order changes slightly, with most changes observed in testis. More changes are observed in brain than in liver and in kidney, and the fewest changes are observed in heart.

To assign human-chimpanzee differences in gene expression to a specific lineage, I considered genes as changed on the human lineage when for this gene a significant expression change was observed between humans and chimpanzees and between humans and rhesus macaques, but not between chimpanzees and rhesus macaques. Correspondingly, I defined a gene to be changed on the chimpanzee lineage, when it showed a significant difference in the human-chimpanzee comparison and the chimpanzee-rhesus macaque comparison but not in the human-rhesus macaque comparison. I assign a gene expression change as outside of the human-chimpanzee lineage, when the change was observed in the human-rhesus macaque and the chimp-rhesus macaque comparison but not between humans and chimpanzees. Further, I filtered the differentially expressed genes to show the same direction of change in

[7] `http://www-huber.embl.de/users/anders/DESeq/`

IDT Exhibit 1061.0114

Table 5.6: Percentage genes differentially expressed in each tissue, from pair-wise comparisons between species using the `DESeq` package and an adjusted p-value cutoff of 1% FDR. This table partially reprints numbers from table 5.5 on page 112.

Tissue	Genes	Human vs. Chimp	Human vs. Rhesus	Chimp vs. Rhesus
brain	10,646	13.0%	34.9%	37.4%
heart	8,361	2.7%	12.1%	13.1%
kidney	13,401	15.2%	30.0%	28.0%
liver	8,285	9.0%	31.6%	28.7%
testis	14,056	29.2%	44.6%	45.3%

these comparisons. Figure 5.11 on the following page illustrates this assignment procedure and presents the results for the different tissues.

Changes were assigned mostly symmetrical between lineages, with liver showing the largest skew with about 8% more changes being assigned to the human lineage. In kidney about 5% more changes are assigned to the human lineage, in testis about 4% and in heart 3%. In brain about 2% more changes are assigned to the chimpanzee lineage. For all tissues and for all changes assigned to either human and chimpanzee, we observe more down then up-regulation. This is consistent with the idea that random mutations may rather have negative impact on gene expression (e.g. by weakening a transcription factor binding sites) than enhancing gene expression. Surprisingly, for genes assigned to the long lineage from the human-chimpanzee common ancestor to rhesus macaque more up-regulation is seen in brain and kidney. It is unclear whether this might be a result of positive selection on gene expression in these tissues or whether it is an artifact of the long lineage to rhesus macaque. The lineage from the human-chimpanzee common ancestor to rhesus macaque spans more than 50 million years of divergent evolution [220] and may have allowed multiple expression changes for the same gene. Such multiple changes will cause an underestimate of gene expression for larger divergences (saturation effect).

It is of interest that assigning human-chimp differential expression differences was not equally efficient for the five tissues, i.e. not the same proportion of differentially expressed genes could be assigned to a lineage by triangulation. The lowest efficiency is observed in testis, where only 53.2% of changes could be assigned. Kidney and heart showed 59.6% and 64.3% assignment rates. The highest rates were observed for brain and liver, with 66.3% and 69.1%, respectively. It is striking that the lowest assignment rate is observed for testis, the tissue with most differentially genes, suggesting some saturation effect when using the rhesus macaque outgroup.

Since gene expression seems to evolve at different rates between tissues, with testis and kidney showing the highest rates, one can assess saturation by comparing the rates of differential expression observed between human and chimpanzee to the rates observed between either of them and rhesus macaque[8]. If gene expression differences behave in a clock-like manner with constant rates, one should obtain a ratio of $6.5Ma/(60Ma - 6.5Ma)$ [220] for all tissues. Saturation would cause higher ratios for tissues with higher rates as changes on the long branch will be underestimated. Table 5.7 on the next page provides the results.

The ratios follow the rates of differentially expressed genes for each tissue, indicating that a saturation effect is masking changes on the longer lineage. Only in heart, where the fewest

[8] Assumes neutral evolution with little or no selection (purifying or positive/functional) in all tissues as well as constant rates.

IDT Exhibit 1061.0115

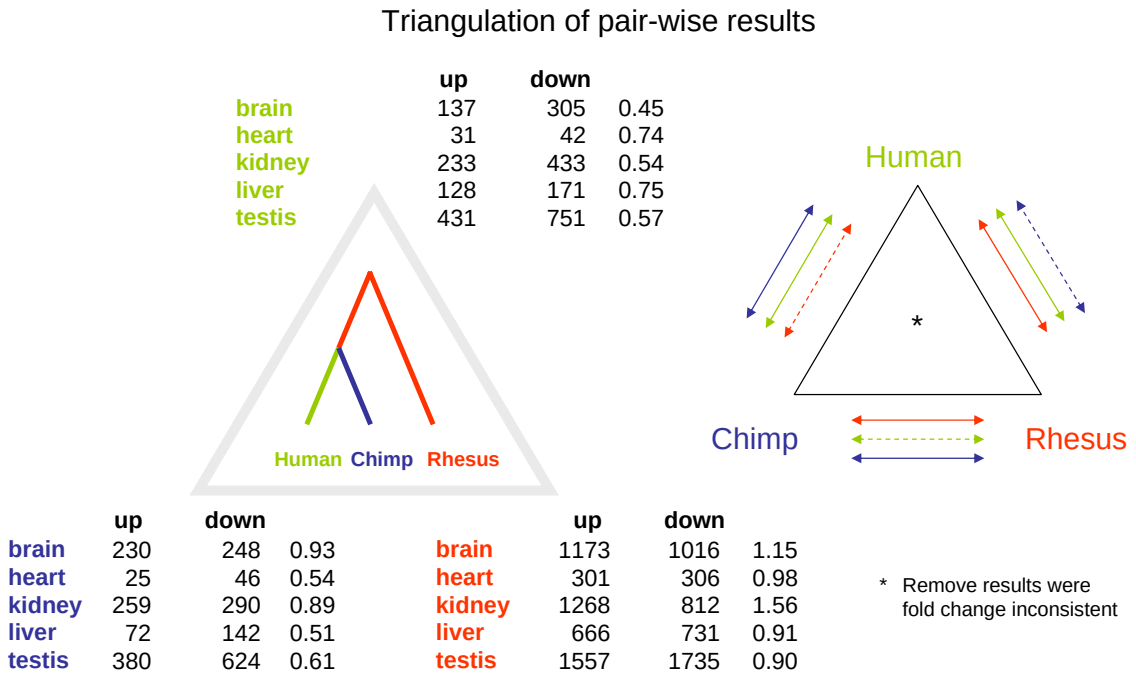

Figure 5.11: Triangulation of pair-wise expression differences. A genes is considered changed on the human lineage **(green)** when for this gene a significant expression change is observed between humans and chimpanzees and between human and rhesus macaques, but not between chimpanzees and rhesus macaques. Correspondingly, a gene changed on the chimpanzee lineage **(blue)** is defined by a significant difference in the human-chimpanzee comparison and the chimpanzee-rhesus macaque comparison, but no significant difference in the human-rhesus macaque comparison. Gene expression changes were assigned as outside of the human-chimpanzee lineage **(red)**, when the change was observed in the human-rhesus macaque and the chimp-rhesus macaque comparison, but not between humans and chimpanzees. Further, differentially expressed genes had to show the same direction of change, i.e. up- or down-regulation, in these comparisons. Decimal numbers next to the number of up- and down-regulated genes, give the ratio of number of up-regulated by number down-regulated genes.

Table 5.7: Checking saturation/clock-like behavior in differentially expressed genes by comparing the rates of differential expression observed on the human lineage (hsa_sp) and chimpanzee lineage (ptr_sp) to the rates observed on the lineage from the common ancestor of human and chimpanzee to rhesus macaque (shared). Rates of differentially expressed genes in each tissue indicate a saturation effect which masks changes on the longer lineage to rhesus macaque. Based on these results, only heart evolves clock-like while kidney and testis suffer from a stronger saturation effect caused by their higher rates of expression change accumulation.

Tissue	Genes	hsa_sp	ptr_sp	shared	$\frac{hsa_sp}{shared}$	$\frac{ptr_sp}{shared}$	Expectation
brain	10,646	4.2%	4.5%	20.6%	20%	22%	
heart	8,361	0.9%	0.8%	7.3%	12%	12%	12%
kidney	13,401	5.0%	4.1%	15.5%	32%	26%	=
liver	8,285	3.6%	2.6%	16.9%	21%	15%	$\frac{6.5Ma}{60Ma-6.5Ma}$
testis	14,056	8.4%	7.1%	23.4%	36%	30%	

IDT Exhibit 1061.0116

Table 5.8: Checking saturation/clock-like behavior based on expression distance measured as the average euclidean distance per expressed gene. Ratios of human-chimp expression distance and either human-rhesus macaque distance or chimpanzee-rhesus macaque distance are compared to species sequence divergence [35, 75] ratios or the earlier used species divergence time ratios [220], which are both similar in value. Average expression distance as defined by Khaitovich et al. [111] includes the non-significant differences, the actual amount of expression change and the number of genes expressed in calculation. On this data set the number of expressed genes seems to a have strong effect on this expression distance, as tissues end up with lower ratio values the more expressed genes are observed.

tissue	HC	HR	CR	HC/HR	HC/CR	Expectation
brain	2.198 ±0.525	3.250 ±0.645	3.502 ±0.424	67.6%	62.8%	21.7% =
heart	2.341 ±0.820	3.298 ±0.899	3.024 ±0.648	71.0%	77.4%	$\frac{2 \cdot 6.5Ma}{60Ma}$
kidney	2.228 ±0.553	3.899 ±0.793	3.556 ±0.471	57.1%	62.7%	
liver	2.441 ±0.524	3.379 ±0.407	3.220 ±0.456	72.2%	75.8%	19.0% =
testis	1.811 ±0.224	4.520 ±1.076	4.127 ±0.957	40.1%	43.9%	$\frac{1.23}{6.46}$

expression changes are observed, do the calculated ratios match the ratio expected from species divergence time. The same analysis can be done using expression distance measured as the average euclidean distance per expressed gene [111] instead of the fraction differentially expressed genes. Table 5.8 provides the results.

Using expression distance between species, one should be able to compare the ratio of human/chimpanzee distances and human/rhesus macaque or chimp/rhesus macaque to the ratio of species sequence divergences [35, 75] or the earlier used species divergence times. However, the expression distance behaves largely different from the fraction of differentially expressed genes, as it includes the non-significant differences, the actual amount of expression change and the number of genes expressed in calculation. This measure is therefore very difficult to interpret in this context. Values are not constant between tissues, indicating some problems in linearly measuring expression distance to rhesus macaque.

5.7 Comparison to other data sets

In the last sections, I have discussed how different factors impact the quantification of gene expression in three species from short tags. Even though several imperfections of the approach and the data set have been outlined, results also indicated that two important features like ranking transcriptional complexity of tissues and the fraction of differentially expressed genes are stable for distinct data analysis approaches. Even though this provides some trust in the internal consistency of the data set, it does not rule out problems from sampling and experimental strategy. The only way to assess the data set in this respect is its comparison to already published data sets including different samples and technologies. In this section, I will compare our DGE data set of five tissues and three species to:

1. A brain data set of three male human, chimpanzee and rhesus macaque individuals that was generated using the same *NlaIII* DGE protocol and published by Babbitt et al. in Genome Biology and Evolution in 2010 [10].

2. A liver data set of three male and three female human, chimpanzee and rhesus macaque individuals that uses the non-stranded[9] Illumina RNAseq protocol and was published

[9]Non-stranded RNAseq protocols do not provide information on the DNA strand a transcript originates from, thus they may identify chimeric genes from transcripts overlapping on different strands

IDT Exhibit 1061.0117

by Blekhman et al. in Genome Research 2009 [23].

3. An Affymetrix HG-U133+ 2.0 array[10] data set of all five tissues each studied in six human and five chimpanzee individuals and published by Khaitovich et al. in Science 2005 [111].

With these three data sets, we have one data set also using the DGE protocol in different individuals (1), allowing us to test for the effects of sample selection and and sample processing. One data set (2) using the next generation of gene expression quantification technologies – RNAseq, which is a "RNA shot-gun sequencing approach" and therefore less affected by sequence biases (see section 5.7.2). Unfortunately, different individuals were used in this study, causing any discrepancies to either originate from technology or sampling. Further, as the third data set we have the classical approach of measuring gene expression in a genome-wide manner using hybridization arrays. Since this last data set was also generated at the MPI for Evolutionary Anthropology, samples used partially overlap with the samples used for the DGE study. Therefore, fewer samples differences are to be expected in the comparison with this dataset.

5.7.1 Comparison of brain samples with Babbitt et al.

In 2010, Babbitt et al. [10] published a brain data set of three human, chimpanzee and rhesus macaque individuals in Genome Biology and Evolution. Their data set was generated using the same Illumina *NlaIII* DGE protocol and they also used frontal cortex tissue of male individuals. Only three individuals have been studied in each species; individuals different from the ones analyzed in our five tissues DGE study. The authors of this study have analyzed data slightly differently than described here. Briefly, Babbitt et al. aligned reads using `Maq` [133], allowing at most four equally good alignments. Multiple sites within human NCBI `RefSeq` annotation (requiring correct strand information) and non-coding RNA annotation (UCSC's `Genome Browser RNAGene` track) were summed. Human `RefSeq` annotation was projected to the other two species using `blat` [109] alignments. From this processing of the data, raw counts assigned to `RefSeq` identifiers were obtained. To compare them with the five tissues data set, I converted `RefSeq` identifiers to `EnsemblGeneIDs` and used their raw counts with the `DESeq` package as described before (section 5.5 on page 109).

To identify differences caused from processing, raw reads of their study (Courtney Babbitt, personal communication, Oct 24 2010) were reanalyzed with the above described analysis pipeline for our DGE data (section 5.2 on page 99). Table 5.9 on the next page provides the number of genes and the number of differentially expressed genes (including the assignment to lineages) for both analyses and compares them to the results for the five tissues DGE data set.

Reprocessing from raw reads using our analysis pipeline identified more expressed genes and resulted in a larger proportion of differentially expressed genes. However, reprocessing reduced the excess of changes assigned to the human lineage in the Babbitt et al. study from 9.7% to 5.0%. This fraction is however still in conflict with the result from the five tissues DGE data, where about 2% more changes are assigned to chimpanzee in brain. In table 5.9 on the facing page it appears that reprocessing caused the data to look more dissimilar between studies. This is misleading as correlation between the two data sets increased considerably with the new processing (figure 5.12 on the next page).

[10] 3'-amplification-based proprietary *in vitro* transcription and labeling system; single dye detection

IDT Exhibit 1061.0118

Table 5.9: Comparison of the two cortex data sets, the data set comprised of three individuals per species published by Babbitt et al. and the, each five, individuals from the five tissues DGE study. From the processing presented by Babbitt et al., raw gene counts were used with the `DESeq` package as described before. In addition, raw reads of this study were reanalyzed with the outlined analysis pipeline (section 5.2 on page 99). Differentially expressed genes were assigned to lineages as outlined in figure 5.11 on page 116.

	Genes	HC	%Diff	HR	CR	hsa_sp	ptr_sp	shared	H/sh	C/sh
Paper	9,961	1,323	13.3%	3,185	3,209	474	320	1,925	24.6%	16.6%
Reanalysis	11,469	1,939	16.9%	3,470	3,553	629	515	1,907	33.0%	27.0%
This study	10,646	1,387	13.0%	3,716	3,978	442	478	2,189	20.2%	21.8%

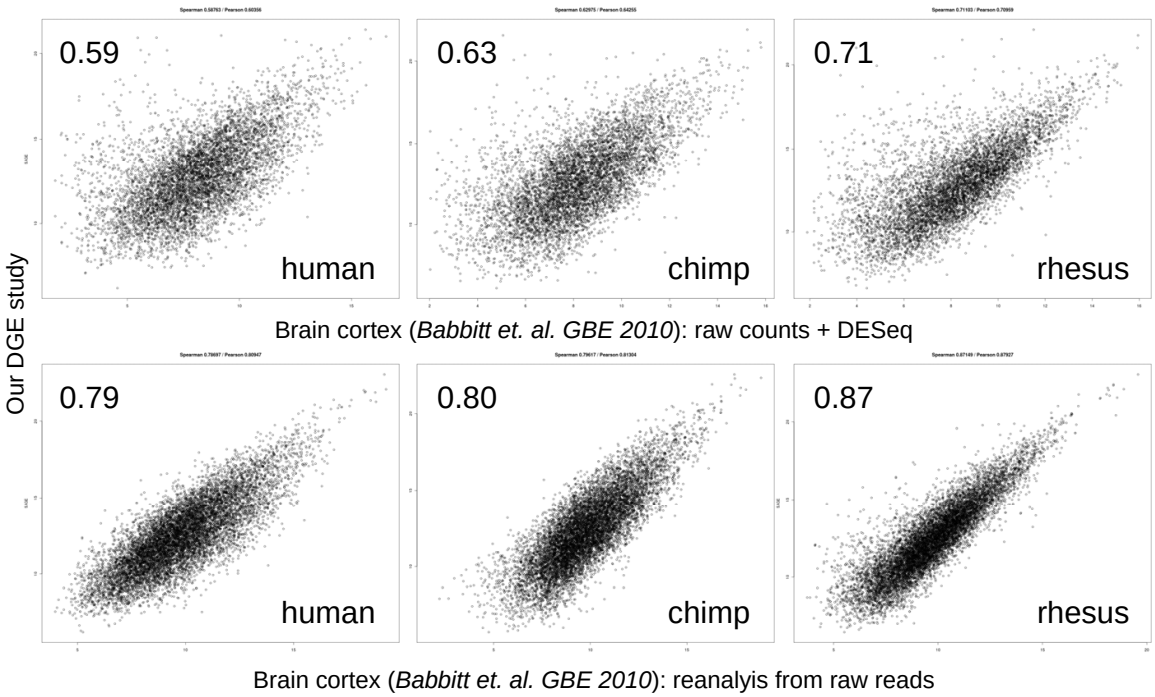

Figure 5.12: Scatter plots and Spearman correlations of brain cortex expression values from Babbitt et al. (raw counts presented in the paper **(upper panel)** and counts after reanalysis from raw reads **(lower panel)**) with the brain samples from the five tissues DGE study. Without reanalysis a larger proportion of genes with higher counts in our data is observed, this is probably due to the different treatment of tags mapping to multiple genomic sites.

IDT Exhibit 1061.0119

Reanalysis of the Babbitt et al. data set increased Spearman correlation between data sets by about 0.2, indicating a large impact of data processing on gene quantification. Further, without reanalysis a large proportion of genes with higher counts in our data is observed. This is probably due to the different treatment of tags mapping to multiple genomic sites and likely the source of the lower correlation. After reanalysis correlation ranges from 0.79/0.80 in human and chimpanzee to almost 0.87 in rhesus macaque. From sequencing and library preparation replicates in our data set (not shown), a Spearman correlation of 0.86 was observed when a library from a second tissue sample of the same individual was prepared and sequenced. A Spearman correlation of 0.89 and 0.96 was observed when sequencing the same two sequencing libraries on a different sequencing instrument version. Hence, the correlation for the rhesus samples is in the range of technical variation observed in our data. The difference in correlation between human/chimpanzee and rhesus, is likely to originate from differences in sampling individuals and tissues.

Since the third data set from Khaitovich et al. also includes brain samples for human and chimpanzee, one can calculate all pair-wise Spearman correlations for human and chimpanzee mean gene expression between the three data sets. Table 5.10 gives the results. The reanalysis does not improve correlation with the Khaitovich et al. array data set, which are mostly lower as correlations between the data sets generated using the same technology.

Table 5.10: Pair-wise Spearman correlation of three data sets, the brain samples from the five tissues DGE study, the Babbitt et al. DGE brain samples and the brain samples from the reanalyzed Khaitovich et al. array data set. For Babbitt et al. brain samples results starting from paper raw counts and after reanalysis from raw reads are presented.

	Human			Chimpanzee		
	Paper	Reanalysis	Khaitovich	Paper	Reanalysis	Khaitovich
this study	0.59	0.79	0.49	0.63	0.79	0.47
Khaitovich	0.60	0.57	-	0.49	0.49	-

Using Principal Component Analysis (PCA) one can analyze whether the brain samples collected by Babbitt et al. cluster with the brain samples in the five tissues DGE data set. Figure 5.13 on the next page shows the first five principal components when using the raw counts provided with the paper, figure 5.14 on page 122 provides the equivalent plots when using the reanalyzed samples.

When the raw gene expression counts available from the Babbitt et al. publication are normalized and variance stabilized with the five tissues DGE data set prior to PCA, the third principle component separates the Babbitt et al. samples from all other brain samples and only the fourth component separates human/chimpanzee individuals from rhesus macaque individuals. When combining the reanalyzed data with the five tissues DGE data, the third principle component separates human/chimpanzee individuals from rhesus macaque individuals and only the fourth and fifth component separate the Babbitt et al. samples from the five tissues DGE samples. The difference of rhesus macaque samples from human and chimpanzee samples is supposed to be the strongest biological signal after the tissue differences. This emphasizes how strongly data analysis impacts the results. Figure 5.15 on page 123 shows the result of the Principal component analysis when only brain samples are considered.

Reducing PCA to only the brain samples shows that even after reanalyzing the data from raw reads, the two sources of the data can be clearly identified for human and chimpanzee samples. While two of the three rhesus macaque samples can also be separated after reanalysis, one of the samples can not be separated using the first two principle components. Since

IDT Exhibit 1061.0120

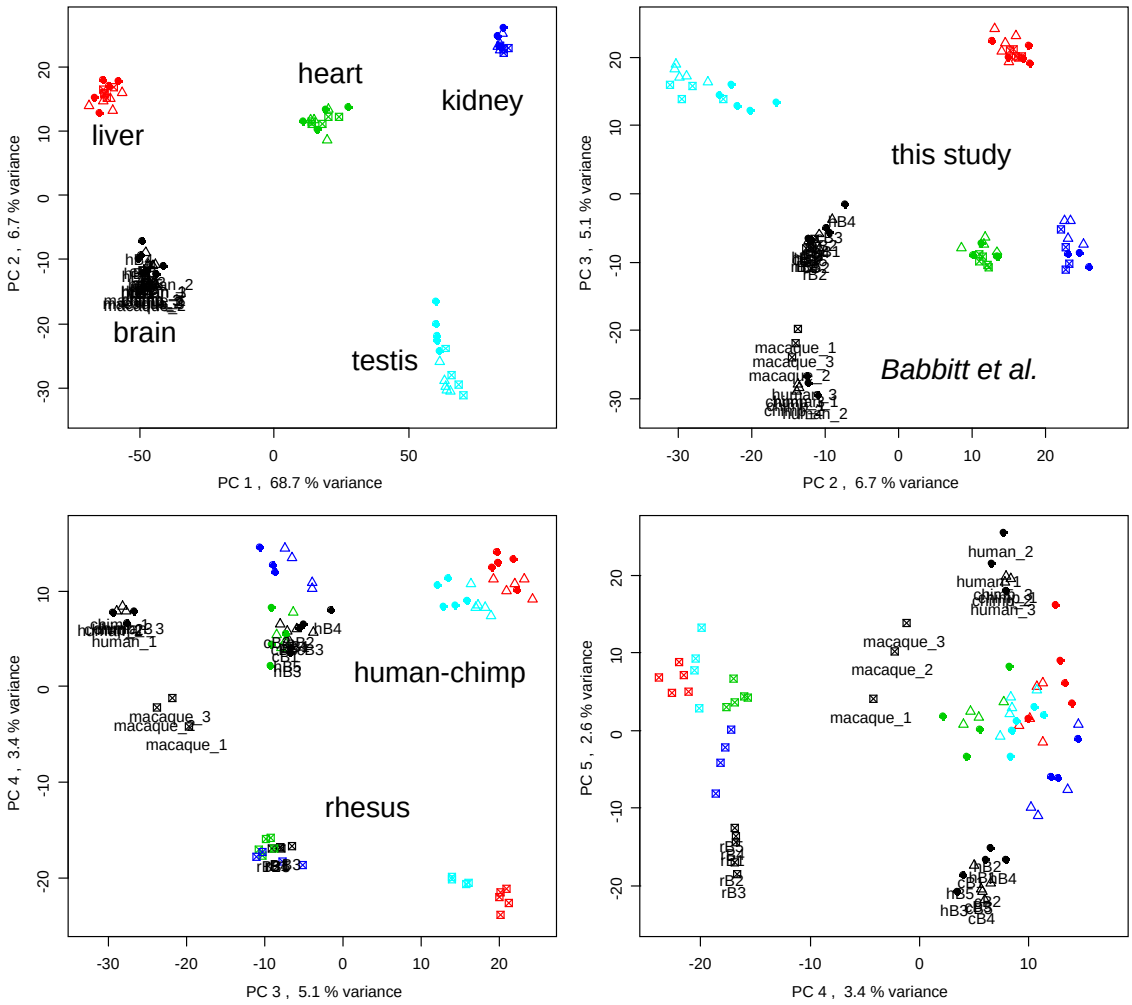

Figure 5.13: Principal Component analysis of the five tissues DGE data set when including the Babbitt et al. data starting from the raw gene counts presented in the paper. Raw gene expression counts have been normalized together with the five tissues data and variance stabilized gene expression values, obtained from `DESeq`, used as input for PCA. Sample labels are only plotted for brain samples, colors indicate tissue and symbols species (humans – dots, chimps – triangles, rhesus – crossed squares). The first two principle components clearly separate tissues, with the second component separating brain and testis from all the other tissues. The third principle component separates the Babbitt et al. samples from all other brain samples and only the fourth component separates human/chimpanzee individuals from rhesus macaque individuals.

IDT Exhibit 1061.0121

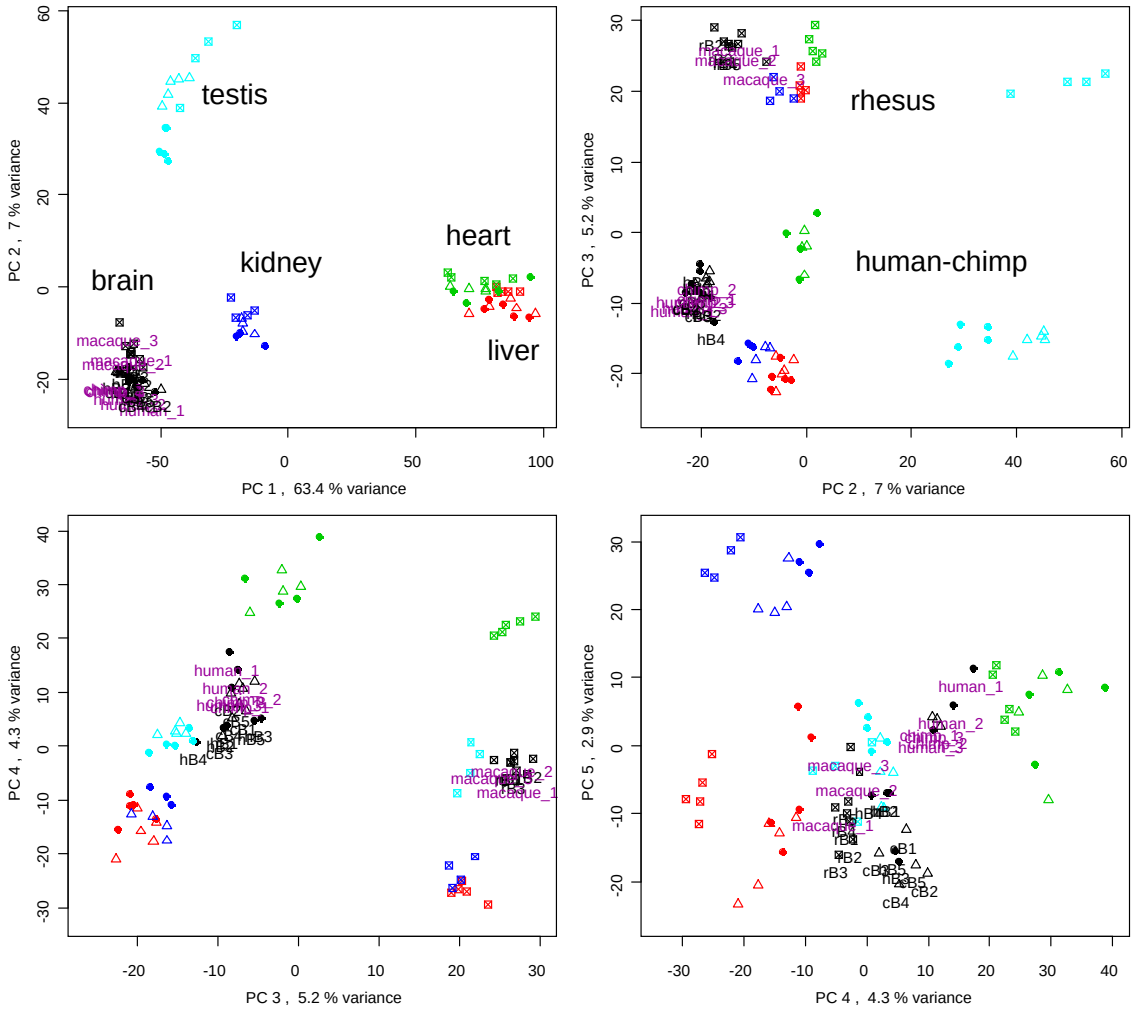

Figure 5.14: Principal Component analysis of the five tissues DGE data set when including the completely reanalyzed Babbitt et al. data. Variance stabilized gene expression values, obtained from `DESeq`, are used as input for PCA. Sample labels are only plotted for brain samples, colors indicate tissue and symbols the species (humans – dots, chimps – triangles, rhesus – crossed squares). To ease identification, labels of the Babbitt et al. samples were colored in purple. The first two principle components clearly separate tissues, with the second component separating testis from all the other tissues. The third principle component separates human/chimpanzee individuals from rhesus macaque individuals. The Babbitt et al. samples are separated from the five tissues DGE samples in the fourth and fifth component.

IDT Exhibit 1061.0122

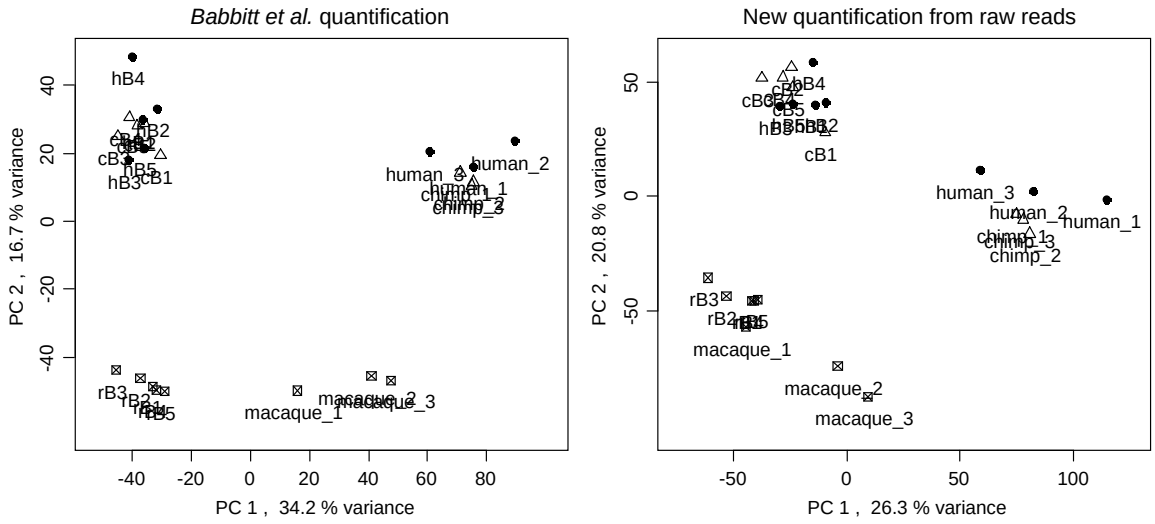

Figure 5.15: PCA of five tissues DGE and Babbitt et al. brain data using gene raw counts presented in the paper **(left)** or counts after reanalysis from raw reads **(right)** as input for normalization/variance stabilization. Labels are provided for all samples. Symbols represent species with dots for humans, triangles for chimpanzees and crossed squares for rhesus macaques. Without reanalysis samples from the two studies can be clearly separated using principal component one, after reanalysis the separation of data sets is not as complete, but still contributing more variance in the data set than human/chimpanzee differences in brain.

we can rule out processing differences, the remaining signal of sample source must originate from sampling individuals and tissue as well as subtle differences in experimental processing. Even though self-evident from the PCA, it is worth noting that the variation introduced from the two data sets after reanalysis is still considerably larger than the variation from human-chimpanzee differences in brain.

5.7.2 Comparison of liver samples with Blekhman et al.

Blekhman et al. published the first comparative primate RNAseq data set with non-pooled individuals in Genome Research 2009 [23]. They generated a liver data set of three male and three female human, chimpanzee and rhesus macaque individuals which uses the non-stranded Illumina RNAseq protocol. RNAseq can be described as the equivalent of a genome shot-gun sequencing approach for the transcriptome. As RNA fragments from the full length transcript are sequenced, this approach is affected less by sample and library preparation (e.g. molecule GC biases from PCR amplification and gel excision, see section 3.10 on page 64 of chapter 3 as well as [32, 140]) and analysis biases from ambiguous alignments in specific parts of the transcript, as the quantification can use the full length of the transcript and correct for regions of increased or reduced read coverage. Recent versions of the RNAseq quantification tool `cufflinks`[11] [229] implement such correction procedures.

Restricting this data set to the male individuals, the data set comprises of 18 Illumina Genome Analyzer II lanes, two for each individual with on average 6.7 million 36nt reads per lane (minimum 4.4 million, maximum 8.0 million). Considering the number of reads and the short length, this data set was probably generated in summer/autumn 2008 shortly after the release of the Genome Analyzer II update and is therefore of limited sequencing data quality. As for the Babbitt et al. dataset the authors have provided a table with counts per

[11] http://cufflinks.cbcb.umd.edu/

IDT Exhibit 1061.0123

gene, but for this study raw reads were also deposited in the `Short Read Archive` of NCBI. I have analyzed data starting from both sources.

For the raw gene counts provided, numbers from the two different lanes per individual were summed, genes with zero counts in one of the individuals were excluded and counts used with the `DESeq` package for obtaining variance stabilized data as well as differentially expressed genes. For the reanalysis of raw reads, reads were mapped with `tophat 1.0.13` [228] to the three reference genomes. Genes were quantified using `cufflinks 0.9.1` [229], providing the projected annotation (as described in section 5.4 on page 103) for exons of the longest transcript of each gene. Successful projection (same chromosome, strand and maximum intron size of 300kb) of 75% of exons across the three species was required for a gene to be included in the annotation. `Cufflinks` gene quantification was converted to count data (keeping one decimal of the Fragments Per Kilo base of exon and Million mapped reads (FPKM) precision), genes with zero counts in one of the individuals were excluded and the remaining counts used with `DESeq` for obtaining variance stabilized data as well as differentially expressed genes. Table 5.11 provides number of genes and number of differentially expressed genes (including the assignment to lineages) for both analyses and compares them to the results for the liver samples from the five tissues DGE data set.

Table 5.11: Comparison of the two liver data sets, the data set comprised of three male individuals per species published by Blekhman et al. and the liver subset of the five tissues DGE study. From the processing presented by Blekhman et al., raw gene counts for the same individual were summed and used with the `DESeq` package as described before. In addition, raw reads of this study were reanalyzed with `tophat` and `cufflinks`, gene quantification converted to count data and again used with `DESeq`. Differentially expressed genes were assigned to lineages as outlined in figure 5.11 on page 116.

	Genes	HC	%Diff	HR	CR	hsa_sp	ptr_sp	shared	H/sh	C/sh
Reanalysis	12,163	934	7.7%	1,625	1,527	231	249	753	30.7%	33.1%
Paper	8,283	522	6.3%	1,094	834	142	139	426	33.3%	32.6%
This study	8,285	742	9.0%	2616	2,380	299	214	1397	21.4%	15.3%

Reanalysis identified more genes and slightly increased the number of differentially expressed genes. Otherwise results of the reanalysis are very similar to the results from using the counts provided with the publication. In comparison with the DGE expression data, fewer differentially expressed genes are observed and the lineage to rhesus macaque is shortened in the RNAseq data. Figure 5.16 on the next page provides scatter plots and Spearman correlations for both analyses.

Spearman correlation increased by more than 0.1 with the reanalysis of the Blekhman et al. samples using `tophat` and `cufflinks`. This points to some problems in the original analysis of the data set. Since the third data set from Khaitovich et al. also includes liver samples, one can again calculate pair-wise correlations for human and chimpanzee mean gene expression in the three data sets. Table 5.12 on the facing page has the results. The reanalysis increased correlation with the array and the DGE data set, supporting a problem in the original analysis presented by Blekhman et al. [23]. The correlation with the reanalyzed array data set is always higher than for the DGE data set. Taking into account that RNAseq sequencing could be considered the current gold standard in gene expression quantification, the quality of quantifications obtained from DGE could be questioned. However, since different individuals are studied, an effect from sample selection can not be excluded.

IDT Exhibit 1061.0124

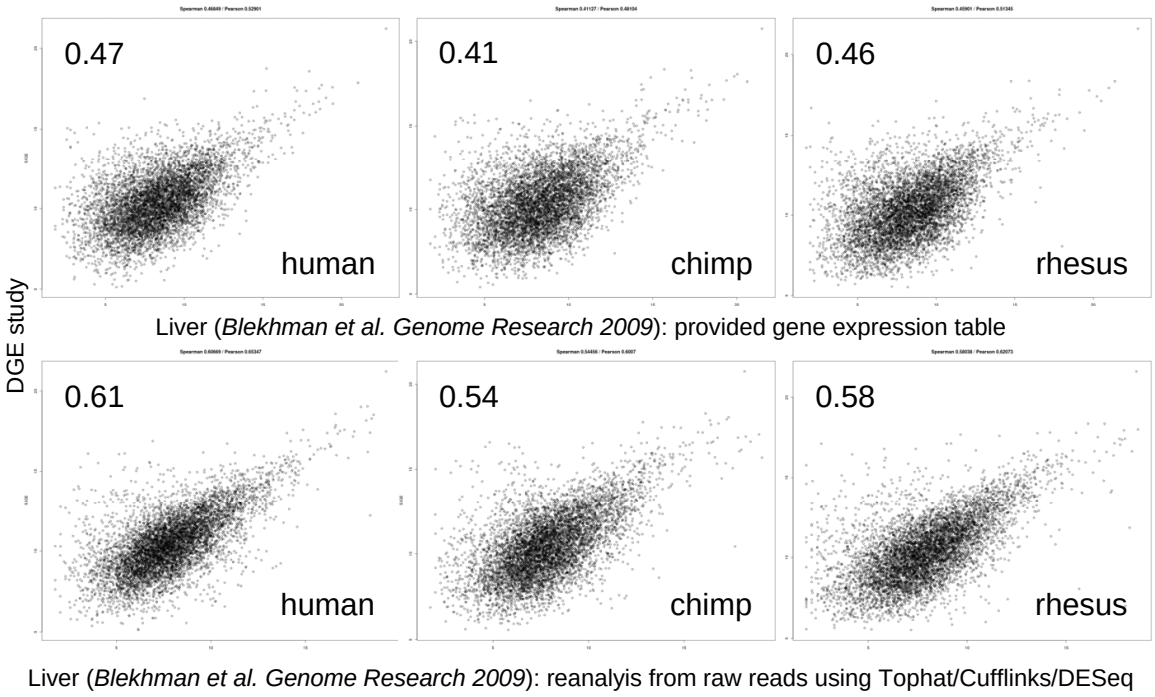

Figure 5.16: Scatter plots and Spearman correlation of DGE and Blekhman et al. data using gene quantifications presented in the paper **(upper panel)** or quantification results after reanalysis from raw reads **(lower panel)**.

Table 5.12: Spearman correlation of three different liver data sets, the RNAseq data set from Blekhman et al. (analyzed from provided gene counts and reanalyzed from raw reads), the liver samples from the reanalyzed Khaitovich et al. array data set and the liver samples from the DGE five tissues study.

	human			chimpanzee		
	Paper	Reanalysis	Khaitovich	Paper	Reanalysis	Khaitovich
this study	0.47	0.61	0.45	0.41	0.54	0.41
Khaitovich	0.56	0.70	-	0.57	0.71	-

IDT Exhibit 1061.0125

5.7.3 Comparison of all five tissues with Khaitovich et al.

Khaitovich et al. [111] used Affymetrix HG-U133+ 2.0 arrays to measure gene expression in the same five tissues for six humans (30 different human individuals across tissues) and five chimpanzees (12 different individuals across tissues). Gene expression arrays are a well-established technology (first used in 1995 [205]) and are therefore considered to be reasonably well understood in their GC hybridization and labeling biases as well as non-specific cross-hybridization to the probes [246]. However, as already mentioned in the introduction of this chapter, hybridization-based technologies are sensitive to polymorphisms in the probed transcript region [15] – a problem that is amplified in comparative studies in which expression patterns of species with much higher levels of sequence differences are inferred and compared on the same array. In such inter-species setups, probes differing in sequence and frequency in the genomes between species have to be excluded from analysis [45]. This exclusion is typically done by comparing the designed probes to the reference sequences in each species, assuming no within-species variation as well as complete and error-free genomes. In combination with serious array design issues due to incomplete and incorrect transcripts in databases at the time of design [11], arrays are now considered less suitable for inter-species comparisons, especially since the advent of the new high-throughput sequencing approaches.

The Affymetrix HG-U133+ 2.0 array was designed in 2003. When using recent genome and annotation builds more than 40% of the >54,000 probe sets can no longer be annotated to genes, while 6% of the remaining 25-mer probe sets also measure more than one transcript [11]. When this data set was analyzed for its publication in 2005, probes were aligned to the human genome NCBI build 35[12] and the first chimpanzee genome build, removing about 32% of probes (kept 412,301 out of 604,258) from the probe sets that did not match human or chimpanzee equally good [111]. Between arrays intensities have been normalized to an equal average intensity measured on the filtered probes. Probe sets were quantified using the `affy bioconductor` package and quantile normalization [24]. Gene annotation was kept as assigned by Affymetrix.

Reanalyzing the orginal data set, I kept 310,073 of the 604,258 probes from all probe sets after mapping them with `bowtie` [126] to human genome GRCh[13] build 37 (UCSC hg19, without additional haplotypes) and chimpanzee genome CGSC[14] build 2.1 (UCSC pantro2), where the remaining probes can be aligned without any mismatches and exactly once in both genomes. The so-identified probes were used with the `R` probe masking package `mask` [45] and the bioconductor `EMA` package [207]. Probe sets were normalized and quantified tissue-wise using the `GC-RMA` [246] background adjustment procedure. After normalization the default expression cutoff was applied, not removing any of the quantified probe sets. For genes quantified by multiple probe sets, the median value was considered for the gene. Genes were tested for differential expression between human and chimpanzee using Welch t-statistics allowing possibly unequal variances as well as Student's t-test assuming equal variances in the two species. Benjamini Hochberg [14] FDR adjusted p-values were used to correct for multiple testing. Table 5.13 on the next page compares the original analysis results with the results obtained from the reanalysis using current genome builds and annotation information for the Affymetrix HG-U133+ 2.0 array.

Comparing the old and the new analysis, ordering of tissues by the percentage of differentially expressed genes is slightly altered: in the new analysis brain ends up with more changes than kidney. The order in expression divergences, calculated as the average euclidean expression

[12] this array was designed from experimentally identified transcripts and human NCBI genome build 30
[13] Genome Reference Consortium for human
[14] Chimpanzee Sequencing and Analysis Consortium

IDT Exhibit 1061.0126

Table 5.13: Comparisons of published results with reanalysis results for the Khaitovich et al. Affymetrix HG-U133+ 2.0 array data set. The original probe intensity files were masked using current genome builds and updated annotation information for the Affymetrix HG-U133+ 2.0 array was applied. Further, an improved background normalization method (`GC-RMA` [246]) was used. Genes were tested for differential expression between human and chimpanzee using Welch t-statistics (columns Diff.exp (W) and %Diff.exp (W)) for possibly unequal variances as well as the Student's t-test assuming equal variances in the two species (columns Diff.exp and %Diff.exp). Benjamini Hochberg [14] (BH) FDR adjusted p-values were calculated to correct for multiple testing and an adjusted p-value cutoff of 0.01 applied. Expression divergence is calculated as average euclidean expression distance [111].

Khaitovich et al. 2005, Supporting Online Material Table S2 [111]				
Tissue	Genes	Diff.exp	%Diff.exp	Divergence
brain	16775	1306	7.8%	0.245
heart	14988	1436	9.6%	0.450
kidney	17865	1605	9.0%	0.405
liver	15046	1060	7.0%	0.586
testis	21731	7036	32.4%	0.538

Reanalysis (masking, `GC-RMA`, t-test BH FDR 0.01)						
Tissue	Genes	Diff.exp (W)	Diff.exp	%Diff.exp (W)	%Diff.exp	Divergence
brain	14138	255	457	1.80%	3.2%	0.085
heart	14138	379	701	2.68%	5.0%	0.162
kidney	14138	218	428	1.54%	3.0%	0.116
liver	14138	67	219	0.47%	1.5%	0.162
testis	14138	3665	4648	25.92%	32.9%	0.186

distance [111], is almost identical with exception that liver and heart come out with equal values in the reanalysis but have different values in the original analysis (with non-overlapping confidence levels, see Supporting Online Material Table S2 of Khaitovich et al. [111]). From the reanalysis, I could not determine differences in transcription complexity, the previously reported numbers of expressed genes are however in agreement with the five tissues DGE expression data set. The fraction of differentially expressed genes (table 5.6 on page 115) as well as the divergence estimates (table 5.8 on page 117) are not in agreement with the DGE data set.

While both data sets see most differential expression in testis (with about 30% of expressed genes between human and chimpanzee), the DGE data set orders the tissues from highest fraction of differentially expressed genes to lowest as: testis, kidney, brain, liver, heart while the original array study orders tissues as testis, heart, kidney, brain, liver (placing heart differently). Since the reanalysis also reorders kidney and brain[15] and since the fraction of differentially expressed genes are very low after FDR correction (except for testis), it might be argued that the array data set does not have enough power to detect differential expression in tissues other than testis.

To compare more general patterns between the two data sets, figure 5.17 on the following page compares gene quantification results from the reanalysis of the array data set with the quantification in the DGE data set. Spearman rank correlation varies for the five tissues between 0.47 in heart to 0.51 in testis. In three tissues, namely brain, liver and testis, it seems that lowly expressed genes do not follow a linear relation between the two data sets. For those genes, the DGE data set spans a wider value range than the array data. Taking

[15]This reordering is also seen in a reanalysis of the Khaitovich et al. data set by Nowick et al. [161].

IDT Exhibit 1061.0127

into account that this effect is not observed in figure 5.16 on page 125 where the DGE data is compared to RNAseq data in liver, the source of this effect is likely to be found in the array data set. It is possible that it for example originates from cross-hybridization signal overlaying the actual signal of lowly expressed genes.

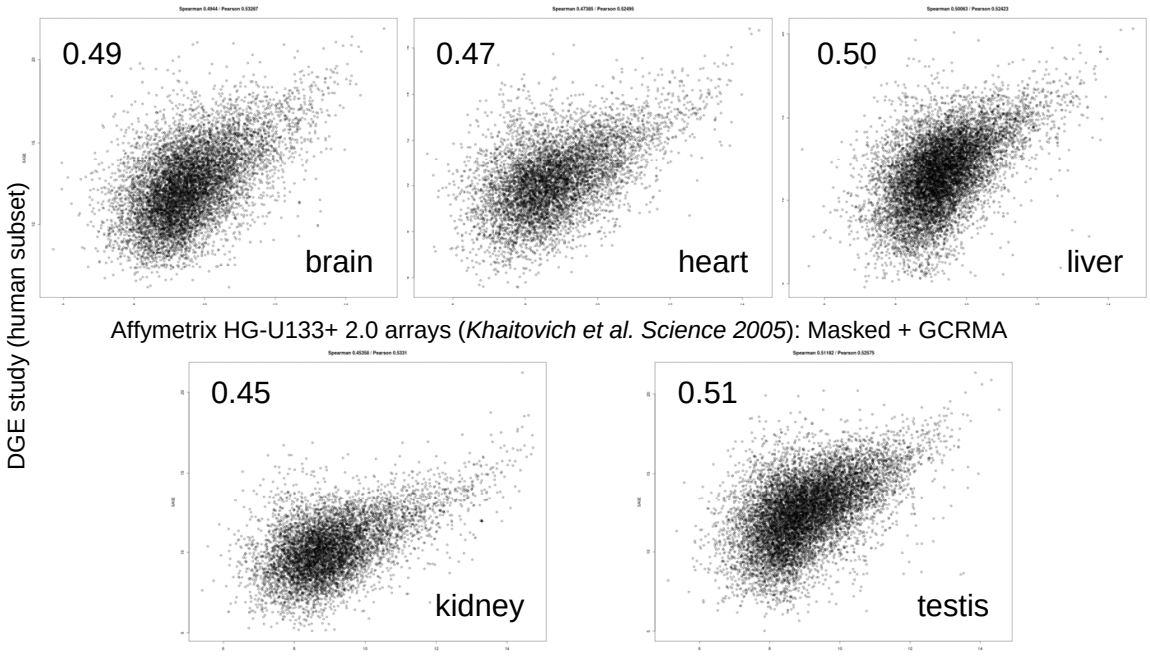

Figure 5.17: Scatter plots and correlations of the human DGE data with Khaitovich et al. array data reanalyzed from probe intensities. Spearman rank correlation vary between 0.47 and 0.51 for the different tissues. In brain, liver and testis, lowly expressed genes do not seem to follow a linear relation in the two data sets. For those genes, the DGE data set spans a wider expression value range than the array data. This effect is not observed in figure 5.16 on page 125 where the DGE data is compared to RNAseq data in liver, thus the source of this effect is likely to be found in the array data set.

5.8 Summary and conclusions

The technological advances in sequencing make it attractive to use a sequencing-based approach for identifying and quantifying the transcriptome of a cell rather than using arrays which are limited to the analysis of specific transcripts known at the time of design. Further, since hybridization-based technologies are rather sensitive to polymorphisms in the probed region of transcripts, inter-species studies were always challenging with arrays as special analysis procedures have to be implemented to prevent species biases, e.g. masking probes with differences in the species under consideration. Even though the described Illumina Digital Gene Expression protocol overcomes limitations of a static design as well as the hybridization and GC biases observed for array technologies, specific features of this type of data also complicate analysis.

I could show that incomplete digestion and enzyme hindrance cause dispersion of sequenced tags across multiple restriction sites and that non-specific carry-over of upstream *NlaIII* digestion fragments between experimental steps causes a false signal of antisense transcription and may also cause additional non-3'-most tag counts on the sense strand. Further, even though there was no evidence from the analysis of this data set, it is likely that the DGE protocol has amplification and gel excision biases dependent on the GC content of the short

IDT Exhibit 1061.0128

tags (see section 3.10 on page 64 in chapter 3 as well as two recent studies on the Illumina smallRNA protocol by Linsen et al. [140] and Caiment et al. [32], a protocol which handles molecules of comparable length).

What was considered a disadvantage of arrays – their static features with an associated annotation – turned out to remain a challenge in the DGE protocol. While array probes are by-design annotated with some gene (even though sometimes incorrectly), SAGE tags need to be mapped and annotated across multiple species. This annotation is problematic due to very different annotation quality for the different species and annotation being inferred for some species from the annotation of a close reference species (like human), thereby not overcoming the species-annotation bias seen with arrays. Even human annotations as available in summer 2008 were not appropriate for analysis, since many genes missed the annotation of 3' UTR sequences. Recent human gene annotation is considerably more complete and could be projected to the chimpanzee and rhesus, losing about 36% of genes annotated in human but giving similar proportions of tag counts within genes for all three species.

The biggest challenge, however, is that the short tag lengths means many tags are not unique to specific genomic sites or genes. In addition, the uniqueness of tags differs slightly between the three species. This causes the tags of two and more different genes to be collapsed into one measurement. As this 3' tag protocol generates mostly one tag site per transcript, the original gene expressions values cannot be reconstructed from neighboring counts. Hence, approaches of counting tags multiple times or removing ambiguous tags both result in incorrect gene quantifications, i.e. a false ranking of gene expression values. Considering the Spearman rank correlations between the two types of processing are very high (0.88-0.93), overall effects may be stable to the problem of ambiguous tags. Further, the impact on the inter-species analysis is expected to be low, as this ranking bias will be largely the same in all three species. Variation between species is to be expected from evolution and loss of *NlaIII* sites in each genome, which are expected at rates close to species sequence divergence (1.23% human-chimpanzee nucleotide divergence [35] and 6.46% human-rhesus macaque nucleotide divergence [75]).

From comparison to other studies of the same species and tissues, larger disagreement is observed than was expected. It is likely that all methods have technological (experimental and analysis) biases and that the obtained variance estimates are too low. Sampling the individuals representing each species (e.g. age, environmental and population differences), tissue sampling including for example sampling time after death, storage and tissue regions selection as well as experimental protocols varying between different studies might have a larger effect than previously expected. The comparison of the Babbitt et al. [10] study with our results clearly shows that sampling variation is in the same range as biological differences between human and chimpanzee and that analysis variation may even be as strong as differences between human/chimpanzee and rhesus macaque. Reanalyzing the Blekhman et al. [23] also showed issues with the analysis presented in the original publication, since reanalysis increased Spearman correlation with the DGE five tissues data and the earlier five tissues array study published by Khaitovich et al. [111] by at least 0.13.

Analyzing agreement between technologies showed disagreement in different measures like the symmetry of assignment of changes to lineages or the percentage differentially expressed genes. The extend of disagreement can not easily be explained from false discovery rates of statistical tests. This can have at least two sources: limits of a multiple testing correction based on false discovery rate and underestimates of experimental and biological variance. False discovery rates only measure type I errors, i.e. only false positive results and not false negative results, thus differences could be due to false negatives. Low variance estimates might be a problem of the statistics/algorithms used, but could also result from different

IDT Exhibit 1061.0129

sampling of individuals or variation in experimental protocols between experimenters and labs.

Comparing the two five tissue studies, the strongest and most consistent pattern in the data sets are the approximately 30% differentially expressed genes between human and chimpanzee testis. One biological reason for this difference could be the more promiscuous mating behavior of chimpanzees, which is associated with a high level of sperm competition, which again might spur an increased sperm production [6, 158, 156]. While the cellular composition of human testis tubules (the places of meiosis and creation of gametes) seems similar to that of other primates, the number of interstitial cells may vary between human and chimpanzee. The study by Mulugeta Achame et al. [156] suggests that per unit of tissue the human testis may contain around 20% less germ cells than the chimpanzee testis. If this is correct, the differential expression observed may be driven by changes in tissue composition between these species. That expression differences originates from variation in cell-type composition [78] is an interesting finding in itself and may also be of relevance for other tissues. However, this does not provide the hoped-for insight into how DNA sequence evolution impacts gene expression. Further, differences in tissue composition could be studied directly using other approaches like cell staining and counting. To which extend cell-type differences and differences in tissue cell-type composition contribute to the phenotypic and functional differences between species can be addressed by future studies using tissue micro-dissections.

For future studies it will be of interest to minimize all mentioned sources of variation. The presented result of human-chimpanzee expression differences being smaller than the variation from sampling and experimental protocols for at least one tissue, challenges current findings from such inter-species studies. Therefore, experimental and biological variation need to be considerably reduced in future studies. To achieve this it will be of interest to stringently control sample environmental effects and age, increase the number of samples, study specific cell-types rather than tissues and to use improved experimental and analysis protocols. The presented analyses have shown that inter-species studies are also very sensitive to small differences in data processing. Such differences may easily originate from different genome quality, genome completeness and genome annotation quality. Measures have to be established to check for such effects in the analysis.

IDT Exhibit 1061.0130

Chapter 6

Analysis of two hominin genomes from ancient DNA

Humanity takes itself too seriously. It is the world's original sin. If the cavemen had known how to laugh, History would have been different. – Oscar Wilde [96](44)

Ancient DNA sequences are generally short in length, damaged [28, 93], and at low copy-number relative to co-extracted environmental DNA. The high-throughput approaches discussed in chapter 2 therefore offer a tremendous advantage over traditional sequencing approaches in that they enable a complete characterization of an ancient DNA extract.

Shotgun sequencing of ancient DNA extracts [154, 81, 186, 185] and sequencing of ancient DNA libraries that have been enriched for specific loci [120, 119, 121, 26, 30], provide a new window into preserved genetic material. For example, the first high coverage mitochondrial genomes [173, 83] made it possible to characterize DNA preservation, contamination and damage [28, 27, 93] to an extent that had not been achieved previously. As the cost of sequencing continues to decrease, it has become feasible to analyze entire genomes of ancient samples [154, 81, 186, 185], including those for which the endogenous DNA makes up only a very small percentage of the total extracted DNA, e.g. the Neandertal genome [81].

However, the qualities specific to ancient DNA present limitations that require careful consideration in data analysis. For example, sequence data of ancient DNA libraries may include chimeric sequences, larger proportions of library adapter sequence at the read ends, sequencing error and artifacts, damage, and alignment ambiguities due to the short read lengths. Partially, these topics have already been discussed in chapter 3. For ancient DNA, the short molecule length, DNA damage patterns [27, 28], the low fraction of endogenous DNA as well as the divergence to the closest modern reference sequence are the dominant sources of analysis problems [181].

I will discuss specific problems as well as selected results from the analysis of whole genome shotgun sequencing data of two ancient hominin genomes, the Neandertal [81] and Denisova genomes [186].

IDT Exhibit 1061.0131

6.1 Neandertals and Denisovans

Neandertals are the closest evolutionary relatives of present-day humans. They lived in large parts of Europe and western Asia before they disappeared around 30,000 years ago. Morphological features typical of Neandertals are described for European fossils dated as old as 400,000 years. Subsequently, the fossil record also contains more distinctive Neandertal forms until Neandertals completely disappear. During their later history, Neandertals could have had contact with anatomically modern humans in the Middle East from at least 80,000 years ago and subsequently in Europe and Asia. [81]

For eastern Asia there is no consensus on which groups were present some 20,000 years ago and before [186]. *Homo floresiensis*, a short-statured hominin which likely represents an early divergence from the lineage leading to present-day humans, lived on the island Flores in Indonesia until at least 17,000 years ago. In China, it was pointed out that morphological affinities between Neandertals and the specimen of Maba, or between *Homo heidelbergensis* and the Dali skull exist. However, these very same specimens were also classified as early *Homo sapiens* by others. [186] DNA evidence indicates that hominins carrying mtDNAs typical of Neandertals were present as far east as the Altai Mountains in southern Siberia [122].

The distal manual phalanx of a juvenile hominin was excavated at Denisova Cave in the Altai Mountains of southern Siberia in 2008 [186]. Systematic excavations at this site over the last 25 years suggest that human occupation of this cave started up to 280,000 years ago. The phalanx was found in layer 11, which is dated to 50,000 to 30,000 years ago. In early 2010, a DNA capture approach in combination with high-throughput sequencing was used to determine the complete mitochondrial DNA (mtDNA) from this Denisova phalanx [121]. Phylogenetic analysis of its sequence, the sequence of multiple Neandertal mtDNA genomes, as well as present-day human mtDNA genomes, determined that the Denisova mtDNA genome diverged from the common lineage leading to present-day human and Neandertal mtDNAs about one million years ago, i.e. about twice as far back in time as the deepest divergence between present-day human mtDNAs and all currently known Neandertals [121]. Since mtDNA is inherited as a single unit and only maternally, gene flow, positive selection and chance events of genetic drift, can have large effects on these measurements. Interpretations of this single locus may therefore not hold for the whole nuclear genome.

Sequencing the nuclear genome of this phalanx [186], showed that Neandertals and Denisovans both belong to the same, currently unnamed, sister group to all present-day humans (figure 6.1A). Comparisons of the genomes of Neandertals, Denisovans and apes to the human genome therefore allow the identification of genomic features specific to fully anatomically modern humans. These ancient genome sequences can provide a catalog of genomic changes which became fixed, or rose to high frequency in present-day humans after their separation from these other hominin forms. Identified changes might also point to regions that were positively selected since modern humans diverged from their last common ancestor with Neandertals and Denisovans.

Studying the origin of anatomically modern humans, their relationship with other human forms is a debated question [236, 186]. One model, the replacement model, postulates that modern humans evolved in a single location in Africa and, from there, spread and replaced all other existing hominins. The competing model, the multiregional model, claims that modern humans evolved at different places and from various archaic human groups. In addition, there is a range of intermediate models that predict an African origin of modern humans but with some contribution from other hominin groups. A primary question is therefore whether there was any admixture between anatomically modern humans and archaic hominins, and to what extent.

IDT Exhibit 1061.0132

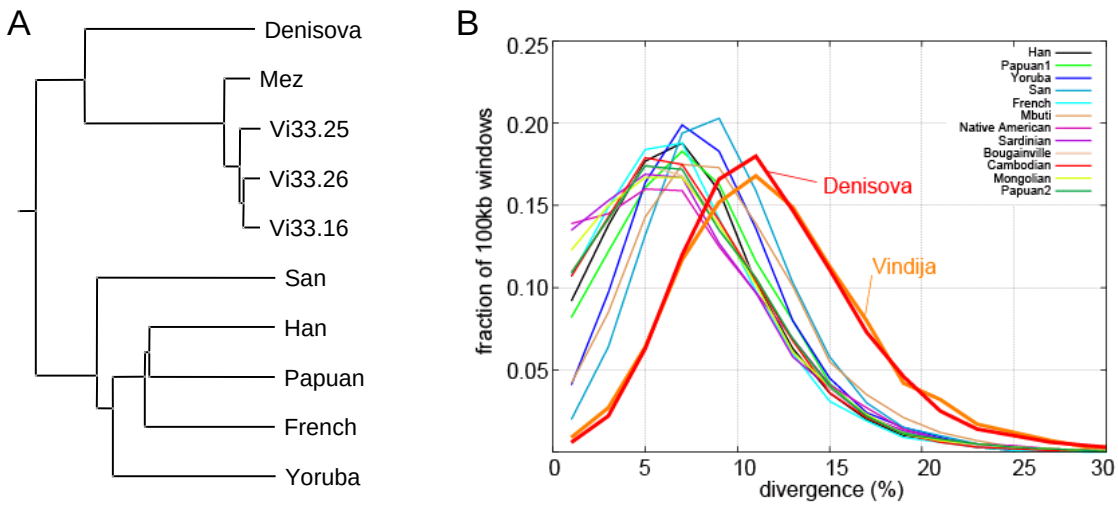

Figure 6.1: **(A)** Neighbor joining tree based on pairwise autosomal DNA sequence divergences for five ancient and five present-day hominins. The tree clearly shows that two human populations (e.g. San and French) are about as closely related as Neandertals and the Denisova individual, suggesting Neandertals and Denisovans are populations of one extinct *Homo* subspecies, rather than two separate subspecies. This, further to be defined *Homo* subspecies, is a sister clade to all present-day humans. **(B)** Variation in genetic divergence over 100kb windows for Denisova, Vindija, and a diverse set of present-day humans as a fraction of human-chimpanzee divergence. A challenge in detecting signals of gene flow between Neandertals/Denisovans and present-day human ancestors is that they share common ancestors within the last 500,000 years, which is no deeper than the nuclear DNA sequence variation within present-day humans. Figure adapted from Figure 1 and Supplemental Figure 2.4 in Reich et al. [186].

IDT Exhibit 1061.0133

The mtDNA genome is the only part of the genome which was examined earlier from multiple Neandertals and the first Denisova individual. The mtDNA of the sampled Neandertal and Denisova individuals falls outside of the variation found in present-day humans and therefore provides no evidence for admixture [81, 121]. This observation does however not rule out that Neandertals or Denisovans contributed other parts of their genomes to present-day humans by admixture. The nuclear genome is composed of several ten thousands recombining and independently evolving DNA segments, which allow for a higher resolution in testing the genetic relationship between these ancient hominin forms and present-day humans.

Detecting such signals of gene flow between Neandertals/Denisovans and present-day human ancestors is, however, complicated by the fact that the two groups share common ancestors within the last 500,000 years [186]. This divergence is not deeper than the nuclear DNA sequence variation found within present-day humans (figure 6.1B). Thus, even if no gene flow occurred, these ancient hominins are expected to be more closely related to some present-day humans than they are to each other in a number of genomic segments [165]. However, if Neandertals and Denisovans are across multiple genomic regions more closely related to humans living in some region of the world than compared to others, this indicates that the ancestors of these present-day humans exchanged parts of their genome with archaic hominins [81, 186].

6.2 Illumina Sequencing and primary data processing

Even though the Neandertal Genome project started in 2005 on the 454 sequencing platform [82], in 2008 the project changed over to the Illumina sequencing platform in order to take advantage of the higher throughput in the number of sequences obtained from this platform. While a single 454 sequencing run yields about one million molecules, the Illumina Genome Analyzer II instrument as it became available in summer 2008 already allowed the parallel sequencing of about 100 million sequences. Due to the short molecule length of ancient DNA, most of the DNA molecules in the Neandertal libraries were accessible with the shorter reads of this technology due to possibility of sequencing molecules from both ends. Thus, DNA sequencing of the final data sets used in the Neandertal and Denisova projects was performed only on the Illumina Genome Analyzer platform [81, 186]. In the following subsections, I will describe how sequencing data was generated for the two projects and sequence alignments to the human [235, 125] and chimpanzee reference genomes [35] obtained.

6.2.1 Neandertal sequence data

To find extracts suitable for sequencing larger parts of the Neandertal genome, 89 different Neandertal bones from 19 sites were analyzed and in total 201 DNA extracts were made from between 5 and 560mg of bone. Extracts were converted to 454 sequencing libraries with a project-specific key (tag) in the beginning of the 454 sequencing read (default '`TCAG`' changed to '`TGAC`'). This project-specific key was introduced to discriminate ancient DNA libraries prepared in a clean room from contamination with other 454 sequencing libraries (see chapter 3 section 3.4 on page 55). In initial screening, the proportion of human contamination in each extract was tested using a PCR-based approach, later also a targeted direct sequencing approach using Primer Extension Capture ([26], PEC) was used [119]. As an additional estimate of the proportion of human contamination and for estimating the percentage of endogenous Neandertal DNA for each Neandertal extract direct sequencing of amplified 454 libraries on a single 16th lane of the Roche 454 FLX platform was performed.

IDT Exhibit 1061.0134

Extracts that were estimated to contain more than 1.5% endogenous Neandertal DNA with less than 5% human contamination in this endogenous fraction were considered for high-throughput shotgun sequencing. To increase the fraction of endogenous Neandertal DNA in these sequencing libraries, restriction enzymes were used to deplete libraries of microbial DNA. A 454 adapter-ligated library molecule cannot be amplified or sequenced if it is cut by a restriction enzyme because the two product molecules will carry only a single 454 adapter each. Two different restriction mixes were used for enrichment. All restriction enzymes used had at least one CG dinucleotide in their recognition sequence – exploiting the underrepresentation of CG dinucleotides due to cytosine methylation in mammalian genomes compared to the environmental background [81].

To sequence the 454 Neandertal sequencing libraries on the Illumina GAII platform, they were converted to Illumina libraries. For this purpose, a PCR primer pair was constructed that is complementary to the 454 A and B primers on the 3'-ends and has tails carrying the Illumina P5 and P7 grafting sequences [119]. In total, 33 Illumina-converted 454-sequencing-libraries carrying the project-specific adapters as well as a converted ϕX control library carrying the 454 standard adapters were sequenced. Instead of the standard Paired End sequencing primers, project-specific primers were used for the forward and the reverse sequencing read, which anneal to the 454 adapter sequences and allow for sequencing the project-specific key at the beginning of each read. For the ϕX control library sequenced in one dedicated lane per run, a sequencing primer covering the 454 standard key was used for the forward read (allowing for Bustard base calling parameter estimates, chapter 4 section 4.3 on page 71).

In total 33 flow cells were sequenced, 24 with 2×51 cycles (FC-204-20xx sequencing chemistry, v2) and nine with 2×76 cycles (FC-103-300x sequencing chemistry, v3). Sequencing was performed at the MPI in Leipzig as well as at Broad institute of Harvard University and Massachusetts Institute of Technology (MIT) (2×76 cycle runs) in Boston from September 2008 through March 2009.

I analyzed these sequencing runs from raw images using the Illumina `Genome Analyzer Pipeline 1.0` and `1.3.2`. To overcome analytical challenges introduced by identical key sequences at the beginning of the first read (see chapter 3 section 3.6 on page 58), I used the first five (instead of two) sequencing cycles for cluster identification with the `Genome Analyzer Pipeline 1.3.2`. For runs analyzed with the earlier `Genome Analyzer Pipeline 1.0`, I modified the `Firecrest` algorithm [16] to perform cluster identification in cycle 4 and then extract intensities from the clusters identified starting with cycle 1.

After standard base calling using `Bustard` with parameter estimation done on the ϕX control sequenced in lane 4 of each flow cell, I aligned the ϕX174 reads to the corresponding reference sequence for creating a training data set for `Ibis` (chapter 4). `Ibis` removed the T accumulation effect observed for v2 and v3 sequencing chemistry (section 4.3 on page 71 and following), generated calibrated quality scores and reduced the overall error rate. I filtered all raw sequences from the new base calling for the three bases of the project-specific key ('`GAC`') and used the algorithm outlined in section 3.2 on page 48 for read merging. Only successfully merged sequences were considered in downstream analysis. By using project-specific libraries and a project-specific key, library contamination (section 3.4 on page 55) is excluded by incompatible sequencing priming sites as well as filtering for the project key in the sequence reads. The latter filter also efficiently reduces the library and sequencing artifacts discussed in section 3.1 on page 44 and section 3.7 on page 59. Merging removes adapter sequence starting at the read ends and further reduces sequencing error.

IDT Exhibit 1061.0135

6.2.2 Denisova sequence data

The Denisova sequencing data presented in Reich et al. [186] originates from two libraries (SL3003 and SL3004) sequenced using 2×101 + 7 cycles on two flow cells[1] sequenced in January 2010. For generating these libraries a total of 40mg of bone was removed from beneath the surface of the Denisova phalanx, DNA was extracted as described by Rohland et al. [195] and treated with the enzymes *Uracil-DNA Glycosylase* (*UDG*) and *Escherichia coli endonuclease VIII* (*EndoVIII*). The *UDG/EndoVIII* treatment leads only to a moderate reduction in the average lengths of the molecules in the library but a several-fold reduction in nucleotide misincorporation, i.e. ancient DNA damage, due to the removal of uracil residues from the library [27]. From the extracts, two independent libraries were created with a modified Illumina multiplex protocol ([150], section 3.3 on page 53). A 7nt-index ('`GTCGACT`') not available outside of the clean room, as well as outer adapter sequences required for sequencing were then added by a PCR reaction.

Libraries were sequenced according to the manufacturer's instructions for multiplex experiments on the Genome Analyzer IIx platform (FC-104-400x v4 sequencing chemistry and PE-203-4001 cluster generation kit v4). The standard protocol was followed except that an indexed control ϕX174 library (index '`TTGCCGC`') was spiked into each lane (see section 3.9 on page 63), yielding 2-3% control reads in each lane.

I analyzed the Denisova sequencing data starting from `QSeq` sequence files and `CIF` intensity files generated by the Illumina Genome Analyzer `RTA 1.6` software. The `Ibis` base caller (chapter 4 on page 69) was trained from the ϕX174 control reads and then used to call bases and quality scores. Raw sequences called by `Ibis 1.1.1` for the two paired end reads were subjected to an index read filtering step where the index read was required to match the index with at most one error [150]. The two reads in each cluster were then merged, including removal of adapter sequences and dimers as well as requiring more than 10nt overlap between the two reads (section 3.2 on page 48). Only successfully merged sequences were considered in downstream analysis.

6.2.3 Identification of endogenous molecules

One of the largest challenges for ancient DNA studies is to distinguish endogenous DNA sequences from those of environmental contaminants. Here, it is useful to know whether the potentially contaminating sequences have features distinct from the endogenous DNA. GC content and length-based filters are typically insufficient to distinguish endogenous from exogenous, i.e. microbial, fungal and other, DNA. K-mer frequencies, in particular the presence of longer k-mers, may be more distinctive. Most commonly used are alignment approaches, in which the sequence reads are aligned to the closest available reference sequence, while allowing sufficient substitutions and gaps to compensate for evolutionary divergence and ancient DNA damage.

More generally and independent of the exact approach applied, there are two different ways in which filters can be applied: (1) negative/subtractive filtering where reads identified by some criterion are removed from the read set versus (2) positive filtering where only reads identified by some criterion are kept in the read set. Negative filtering may leave many false sequences in and remove highly conserved sequences, while positive filtering may miss insertions and divergent regions as well as include highly similar (e.g. conserved) false sequences.

[1] 12 lanes SL3003, 4 lanes SL3004

IDT Exhibit 1061.0136

Alignment approaches may fail when the reference genome and the actual sample have regions of high sequence divergence or differ in larger insertions/deletions events. Eventually these missing alignments may also bias genome-wide estimates like species divergence [181]. For cases of higher divergence sequences or differences in large insertions/deletions events, overlap extension from the alignments is required for identification of endogenous molecules. Such an overlap extension step (e.g. [167]) tries to extend the aligned reads in incompletely covered regions using non-aligned reads. Ideally, alignment and extension steps must be performed in an iterative manner, however, there is no guarantee that such a process converges to the correct sequence.

The same applies to k-mer filters when used instead of alignment approaches. Additional k-mers identified in all reads passing an initial k-mer filter must be used for an iterative identification step to reduce problems with regions of high sequence divergence. Alternatively, de Bruijn graph approaches (e.g. [250, 21, 33, 102]) can be used to first build contigs from overlapping k-mers. These contigs can then be filtered for specific characteristics of endogenous DNA. Alignments may outperform k-mer/de Bruijn graph approaches in cases where samples are experimentally enriched for their similarity to specific sequences, as the enrichment approaches will also enrich for the same k-mers in the environmental/background DNA molecules.

For the Denisova and Neandertal projects [81, 186], a positive filtering by alignment to the human and chimpanzee reference genomes was performed. Due to the high fragmentation and the low coverage obtained in both projects, the direct alignment of sequences (in contrast to a *de novo* assembly) is the only feasible approach for the analysis of the ancient DNA molecules.

For Neandertal, Illumina reads were mapped to the human genome (NCBI 36/hg18) and chimpanzee genome (CGSC 2.1/pantro2) using a custom mapper called `ANFO`. This custom alignment program, written by Udo Stenzel, was developed to take the characteristics of ancient DNA into account and is available from `http://bioinf.eva.mpg.de/anfo`. `ANFO` builds an index of short words of the target genome, in a fashion similar to the method described by Morgulis et al. [155]. The query sequences are broken up into their constituent words for index lookup. Adjacent or near adjacent words are combined into longer matches, and any match that is considered "long enough" serves as the seed for a semi-global alignment. Every fourth word of length 12nt was indexed and at least 16nt long seeds required, which gives sensitivity better than `megablast` [251] and a useful compromise between sensitivity and required computational effort. When building alignments `ANFO` extends in both directions from all seeds using a best-first search. This takes advantage of the fact that only the two best alignments (not every alignment) of each query is needed for calculating map quality scores (MAPQ). The search terminates when two alignments have been found, or when all remaining alignments are guaranteed to be less good than the alignment score cutoff. The score of an alignment is defined as its negative log-likelihood (i.e. better alignments have lower scores). `ANFO` mapping quality was defined as the difference in score between the two best alignments. Following the observation and implementation by Briggs et al.[2] [28], `ANFO` uses different substitution matrices for DNA thought to be double stranded versus single stranded and switches between them if doing so affords a better alignment score. The expected length distribution of single stranded stretches is modeled as geometrically

[2]The model presented by Briggs et al. describes a considerably higher cytosine deamination rate in single stranded DNA overhangs as compared to the inner double stranded molecule parts. For single stranded 5' overhangs of molecules the double strand is repaired by complementary strand synthesis during sequencing library preparation. The 3' overhangs of molecules are removed during library preparation. Thus, part of the determined sequence may either originate from the DNA that was double or single stranded in the original ancient molecule.

IDT Exhibit 1061.0137

distributed. To separate hominin sequences from random similarities, the distribution of alignment scores depending on read length were analyzed. The scores are clearly a mixture of two distinct distributions, with the distinction becoming much less pronounced for shorter reads. We[3] therefore required a minimum read length of 30nt and a score no worse than $7.5 \cdot (length - 20)$ to distinguish spurious alignments of bacterial, fungal and non-mammalian reads from actual hominin sequences.

A total of 2,460,140,110 raw clusters was obtained from 214 lanes with libraries from six different Neandertal bones (El Sidron 1253, Feldhofer 1, Mezmaiskaya 1, Vindija 33.16, Vindija 33.25 and Vindija 33.26). The three Vindija bones contributed 85% of the reads, in about equal shares. From the 2.46 billion raw clusters, 1.35 billion (55%) merged reads were obtained. This low fraction is not unexpected due the large fraction of low quality reads observed for these earlier Illumina Genome Analyzer versions. After alignment and duplicate consensus (all libraries were sequenced with some redundancy of individual molecules and PCR duplicates identified based on their outer genomic coordinates were consensus-called incorporating the sequence quality scores, see section 3.10 on page 64, performed by Udo Stenzel), 86,810,371 (6.4%) unique molecules aligning to hg18 and 87,326,955 (6.5%) unique molecules aligning to pantro2 were reported. These 87 million molecules correspond to about 4.2Gb (roughly 1.5x) of Neandertal sequence.

For Denisova, 447,964,927 index filtered merged reads (originating from 562,650,846 raw reads, 80%) were aligned with `BWA` [131] to the human genome (NCBI 36/hg18) and chimpanzee genome (CGSC 2.1/pantro2) using default parameters. `ANFO` alignments got considerably more time-consuming, probably due to the higher length of the Denisova molecules. Further, ancient DNA damage was greatly reduced by enzymatic treatment in these reads, permitting the use of `BWA`. The `BWA` alignments were converted to `SAM/BAM` format [132] with `BWA`'s samse command and subsequently analyzed for PCR duplicates. Both libraries were sequenced with low redundancy of individual molecules and the few PCR duplicates were consensus-called (see also section 3.10 on page 64, performed by Udo Stenzel). For the two libraries, this resulted in a total of 111,466,516 (24.9%) unique sequences that were mapped to the human genome, altogether resulting in 6.6Gb of sequence. After we restricted to the 82,227,320 (18.4%) sequences with a `PHRED`-scaled mapping quality of at least 30, this resulted in a total of 5.2Gb (roughly 1.9x) of filtered sequence data. The number of sequences unambiguously mapped to the chimpanzee genome with a mapping quality of at least 30 is 72,304,848, which is 87.9% of that for the human genome.

6.2.4 Present-day human low-coverage data

To put divergence measures of the Neandertal and Denisova genomes into perspective with regard to present-day humans as well as to test whether Neandertals and Denisovans are on average across many independent regions of the genome more closely related to present-day humans in certain parts of the world, we[4] sequenced individual genomes from multiple human populations for the Neandertal and Denisova genome projects.

For the Neandertal Genome project [81], we sequenced one San from Southern Africa, one Yoruba from West Africa, one Papua New Guinean, one Han Chinese, and one French from Western Europe. For the Denisova Genome project [186], we sequenced another seven present-day humans, a Mbuti genome from Africa, a Sardinian genome from Europe, a Mongolian genome from Central Asia, a Cambodian genome from South-East Asia, an additional

[3] The Neandertal Genome Analysis Consortium, specifically Udo Stenzel
[4] Neandertal and Denisova Genome Consortia

IDT Exhibit 1061.0138

Papuan genome from Melanesia, a Bougainville islander genome from Melanesia, and a Karitiana genome from South America. The geographical distribution of all modern and ancient DNA samples studied in these two projects is shown in figure 6.2.

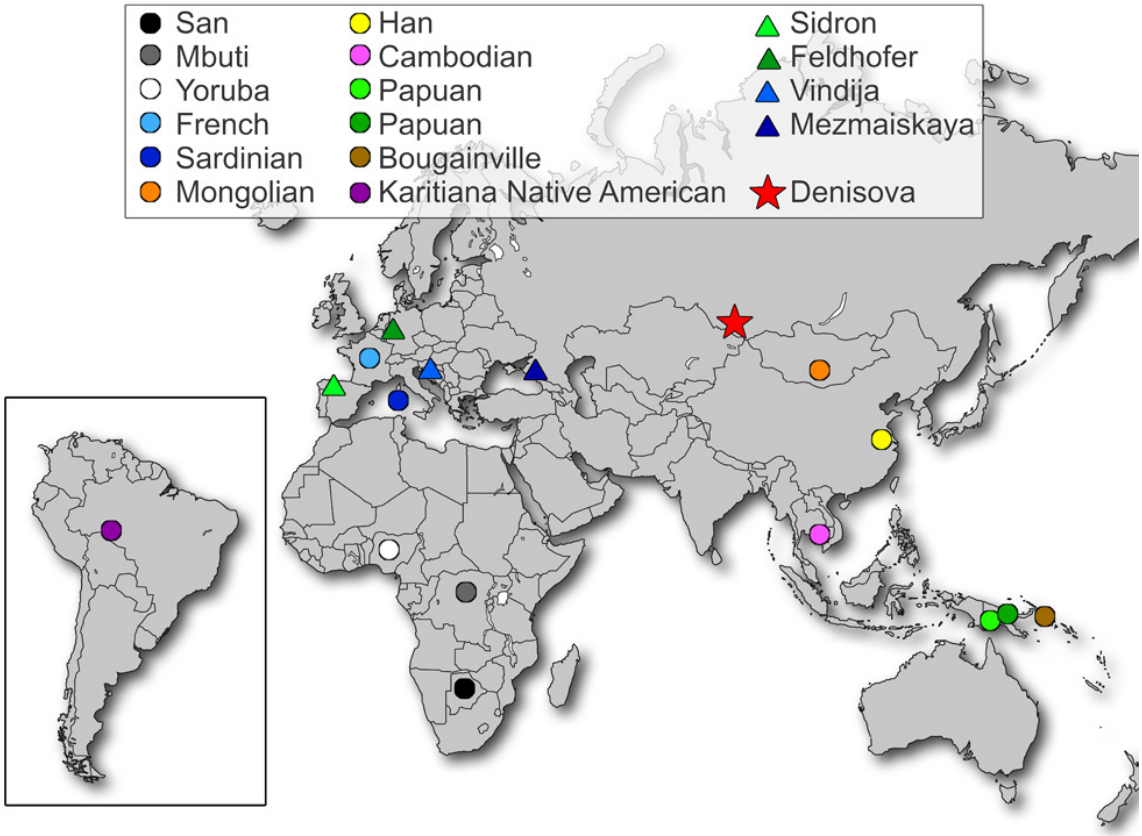

Figure 6.2: Geographic distribution of samples studied in the Denisova [186] and Neandertal Genome [81] projects. Individuals from different human populations were obtained from the Human Genome Diversity Panel (HGDP) by the Centre d'Etude du Polymorphisme Humain (CEPH): HGDP-00491 (Bougainville Melanesian), HGDP00711 (Cambodian), HGDP00521 (French), HGDP00778 (Han Chinese), HGDP00998 (Karitiana Native American), HGDP00456 (Mbuti), HGDP01224 (Mongolian), HDGP01029 (San), HGDP00665 (Sardinian), HGDP00542 (Papuan), HGDP00551 (Papuan) and HGDP00927 (Yoruba). Figure kindly provided by Knut Finstermeier.

Low-coverage data of 5 humans presented in Green et al.

In Green et al. [81], we presented whole genome shot-gun data from five Human Genome Diversity Panel (HGDP) individuals, for whom cultured lymphoblastoid cell lines are provided by the Centre d'Etude du Polymorphisme Humain (CEPH). One microgram of DNA was obtained for each of the individuals. DNA was sheared, Multiplex Illumina sequencing libraries constructed [150], and length excision performed (300bp using 2%-agarose gel). Illumina sequencing of these five libraries was performed on the Genome Analyzer II platform with 2×76 cycles on each one flow cell (FC-103-300x v3 sequencing chemistry) from May to June 2009. Due to low cluster densities for the run of the French individual (HGDP00521), another four lanes were sequenced from this library on an additional sequencing run (same chemistry and protocols).

I analyzed the runs starting from intensity files (`CIF` format) using the Illumina Genome Analyzer `Pipeline 1.4.0`. Base calling parameter estimation for the Illumina base caller

IDT Exhibit 1061.0139

Bustard was done on the ϕX174 control sample in sequenced in lane 4 of each run. The ϕX174 reads were used to obtain a training data set for Ibis (chapter 4). Raw sequences called from Ibis for the two paired end reads of each sequencing cluster were aligned separately with BWA [131] to the human (NCBI 36/hg18) and chimpanzee (CGSC 2.1/pantro2) genome with default parameters. Using BWA's sampe command the alignments for two reads were combined and converted to SAM/BAM format. In this step, missing paired alignments were searched within a window of 800nt around one aligned read (BWA sampe parameter -a 800). Subsequently, the BAM output files of all lanes from the same library were merged and the resulting files filtered by removing read pairs for which either the forward or reverse read failed one of the following criteria:

- Missing the "properly paired" bit in the BAM file.
- Mapping quality of at least 30.
- "Duplicated molecules", i.e. read pairs for which another, higher or equal quality, read pair had boundaries that map to the same outer coordinates (samtools rmdup command, see also PCR duplicate section 3.10 on page 64).
- Sequence entropy of at least 1.0 (see also equation 3.4 of sequencing artifact section 3.7 on page 59).

We obtained 101-161 million raw sequence clusters for the different samples, of which 87.6%-90.6% of reads could be aligned to the human reference genome and 81.6%-86.0% to the chimpanzee genome. When applying the described filters, these numbers reduced to 73.9%-76.2% and 65.3%-68.2%, respectively. This corresponds to on average 16Gb and 13Gb of sequence data aligned to the human and chimpanzee reference genomes.

Low-coverage data of 7 humans presented in Reich et al.

In Reich et al. [186], we presented another seven individuals which we again selected from the CEPH-HGDP panel. Also for these samples, Illumina multiplex sequencing libraries were prepared from the sheared DNA (200-400bp) according to the protocol described in Meyer and Kircher [150]. Further, for each sample (except HGDP00998 which was used without size selection due to the low amount of DNA obtained after library preparation) a narrow band around 300bp was excised from a 2% agarose gel after adapter ligation to obtain inserts of optimal size for sequencing.

Each of the Illumina multiplex libraries was sequenced in one lane using $2\times101 + 7$ cycles on one flow cell according to the manufacturer's instructions for multiplex sequencing on the Genome Analyzer IIx platform (FC-104-400x v4 sequencing chemistry and PE-203-4001 cluster generation kit v4) in March 2010. Instead of having a dedicated control lane, as done for the earlier samples, an indexed control ϕX174 library was spiked into each lane, yielding 2-3% control reads. I analyzed the sequencing run starting from QSeq sequence files and CIF intensity files from the Illumina Genome Analyzer RTA 1.6 software. Ibis was trained on the Bustard control reads aligned to the ϕX174 reference sequence and used to call bases and quality scores from intensity files. The raw paired-end reads were merged (including adapter removal, section 3.2 on page 48). The index reads used for the sequencing runs were not further evaluated for downstream analysis, they were used only to validate the correct assignment of samples to lanes.

IDT Exhibit 1061.0140

The paired-end read merging resulted in two sets of reads for each sample: regular paired-end reads and merged reads. The paired-end reads were aligned using `BWA` [131] to the human (NCBI 36/hg18) and chimpanzee (CGSC 2.1/pantro2) genomes using default parameters. Using `BWA`'s sampe command, the alignments of the first and second read were combined and converted to `SAM/BAM` format. In this step, missing paired alignments were searched within a window of 500nt around one aligned read. Merged reads were aligned separately to these genomes, again using `BWA` with default parameters. Using `BWA`'s samse command, these alignments were also converted to `SAM/BAM` format and combined with the `BAM` output files for paired-end using `samtools` [132]. Alignments to hg18 and pantro2 were filtered using the criteria for the five higher coverage humans, with the following modifications:

- Instead of removing reads missing the "properly paired" flag, non-mapped merged reads and paired-end reads missing one of the alignments were removed.
- "Duplicated" reads with the same outer coordinates (chapter 3 and section 3.10 on page 64) lower or equal sum of quality scores were removed using a custom script handling both paired end and merged reads.

We obtained 30.6-39.7 million raw sequences for these seven samples, of which 76.83%-82.12% of reads aligned to the human reference genome and 68.25%-73.52% also passed the described filters. For the chimpanzee, 68.74%-73.99% of reads aligned and 59.26%-63.26% remained after filters. This corresponds to on average 4.4Gb and 3.8Gb of sequence data aligned to the human and chimpanzee reference genomes.

6.3 Identification of changes on the human lineage

Studying sites in the human genome which changed since the last common ancestor of human, chimpanzee and bonobo allow to define the genetic background of what sets humans apart from other primates. Further, the comparison of the human genome to the genomes of Neandertals and the Denisova individual allows to identify the subset of genomic sites which set fully anatomically modern humans apart from other hominin forms. These might point to uniquely human traits and physiological changes that allowed humans to become the dominant species on this planet.

Positions that have changed on the hominin lineage since separation from apes and more distantly related primates can be identified from multi-species whole genome alignments. Such multi-species whole genome alignments can be created from pair-wise alignments available from the UCSC Genome browser[5] using `autoMultiZ` from the UCSC `Kent-tools`[6]. Whole genome alignments differentiate between a reference and a target genome. While each position of the reference genome appears at most once in the alignment, sequences from the target genome can be used multiple times. This causes whole genome alignments to be non-symmetrical. I therefore generated and used two multi-species whole genome alignments, one based on hg18 as reference and the other based on pantro2 as reference.

These alignments I screened for differences between the human and chimpanzee, and assigned the lineage on which the change occurred based on two out-groups (the orangutan and rhesus macaque). I extracted 15,216,383 single nucleotide changes (SNCs) and 1,364,433 insertion or deletion differences from the human-based alignment that were inferred to happen on

[5] `http://hgdownload.cse.ucsc.edu/downloads.html`
[6] `http://hgdownload.cse.ucsc.edu/admin/exe/`

IDT Exhibit 1061.0141

the human lineage, and 15,523,445 SNCs and 1,507,910 InDels from the chimpanzee-based alignment that were inferred to happen on the human lineage. Only positions identified in both human-based and chimpanzee-based alignments, where no gaps were present within a 5nt-window of the event, and where both out-groups agree with the chimpanzee base, were retained. In the case of InDels, I required that the InDel length does not vary between species and that the InDel sequence is not marked as a repeat (based on UCSC lower case genome soft masking). This generates a set of 10,535,445 SNCs and 479,863 InDels inferred to have occurred on the human lineage.

6.3.1 Identification of positions with Neandertal sequence coverage

As described above, Neandertal merged reads from Vindija 33.16, Vindija 33.25 and Vindija 33.26 were aligned to the human (hg18) and chimpanzee genomes (pantro2) using `ANFO`. To further reduce the effects of sequencing error, we used the alignments of Neandertal reads to the human and chimpanzee reference genomes to construct human-based and chimpanzee-based consensus "minicontigs". To generate these consensus sequences, we selected uniquely placed, overlapping alignments (`ANFO` MAPQ $\geq$ 90) and merged these into a single multi-sequence alignment using the common aligned genome sequence. At each position in the resulting alignment, for each observed base, and for each possible original base, we calculated the likelihood of the observation, estimated the likely length of single stranded overhangs, and considered the potential for ancient DNA damage using the Briggs-Johnson model [28]. If most observations in a given position showed a gap, the consensus became a gap, otherwise the base with the highest quality score (calculated by dividing each likelihood by the total likelihood, equation 3.3 on page 51) was used as the consensus. At the current coverage of the Neandertal sequence ($\approx$1.5x), heterozygous sites will appear as low quality bases with the second base having a similar likelihood to the consensus base.

We extracted the Neandertal sequence for the identified human-lineage-specific changes from minicontig alignments to both the human and the chimpanzee genomes. To ensure high quality and consistency between the human chimpanzee alignments, I filtered the data:

1. The Neandertal sequence at the same position in both human and chimpanzee-based alignments was required to be identical and to have a `PHRED` base quality score > 30.
2. All positions that fall within 5 nucleotides of the ends of minicontigs were excluded to minimize alignment errors and substitutions due to the nucleotide misincorporations, which are frequent close to the ends of ancient DNA molecules.
3. Positions that fall within 5 nucleotides of insertions or deletions (i.e. gaps) in the minicontig alignments were excluded.

Using this filtered dataset, Neandertal sequence data covers 3,202,190 of the 10,535,445 substitutions and 69,029 of 479,863 InDels inferred to have occurred on the human lineage, respectively. Figure 6.3 on the next page summarizes these numbers for different parts of genes.

6.3.2 Identification of positions with Denisova sequence coverage

Like for the Vindija Neandertal reads, we used the alignments of the Denisova phalanx reads to the human and chimpanzee reference genomes to construct human-based and chimpanzee-based consensus sequences from multiple reads of the same Denisova molecule, and joined

IDT Exhibit 1061.0142

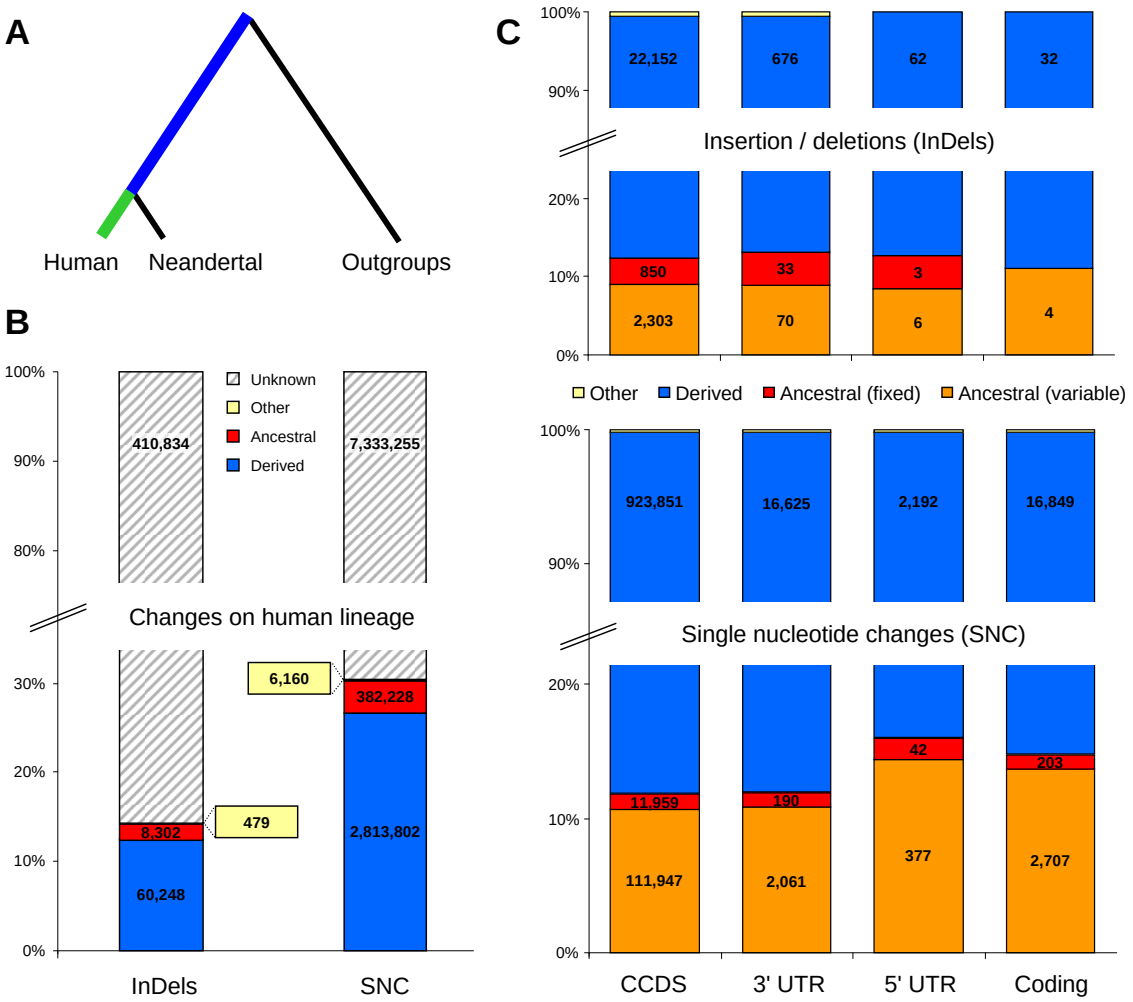

Figure 6.3: Single nucleotide changes (SNC) and insertion/deletion changes (InDels) on the human lineage inferred from whole genome alignments and their state obtained from the minicontigs created from all Vi33.16, Vi33.25 and Vi33.26 reads.

IDT Exhibit 1061.0143

overlapping fragments to construct "minicontigs". In this process, overlapping alignments were merged along the common reference to create a single multi-sequence alignment. For each column of the alignment, the number of gaps was counted, and if half the reads or more showed a gap, a gap (resulting in a deletion or no insertion, as appropriate) was called. If fewer than half the reads showed a gap, the most likely diallele (IUPAC ambiguity codes) per column was calculated as outlined in Reich et al. Supplementary Information 4 [186].

We used the resulting minicontigs to extract the Denisova sequence homologous to the human-lineage-specific changes from both the human and the chimpanzee minicontig alignments and the filtered the data as outlined for the Neandertal before. Using this filtered dataset, we have Denisova sequence coverage for 4,267,431 of the 10,535,445 substitutions and 105,372 of the 479,863 InDels inferred to have occurred on the human lineage.

6.3.3 Annotation of genomic features

We annotated all SNC and InDel events using the `Ensembl v54` annotation for hg18 and `Ensembl v55` for pantro2 (in cases where no human annotation was available). Further, a set of 16,762 human CCDS genes (Consensus Coding Sequence project of EBI, NCBI, WTSI, and UCSC), each representing the longest annotated coding sequence for the respective gene, was used for downstream analyses of protein-coding genes. If not otherwise indicated by specific citations, functional information for genes was obtained from `GeneCards` ([200], `http://www.genecards.org/`).

Due to the low coverage of the Neandertal and Denisova genomes, mostly one allele is sampled from the ancient genomes. We therefore may miss polymorphic derived alleles and consider the Neandertal and Denisova genome ancestral for these positions, thereby incorrectly assigning the arrival of the derived alleles to the human lineage after its separation from Neandertal and Denisova.

6.4 Changes in protein-coding sequences analyzed from Neandertal data

From the Neandertal data, we identified 19,780 SNCs in the coding regions of the human CCDS set. Six of these occur in two overlapping transcripts, while one occurs in three overlapping transcripts. These positions result in 11,337 synonymous substitutions and 8,451 non-synonymous substitutions. Non-synonymous amino acid substitutions that rose to nearly 100% frequency (are fixed) in present-day humans since the separation of humans from Neandertals and Denisovans might be of special interest as they may represent targets of recent selection in humans.

Recently, other group members of the department of Evolutionary Genetics identified in a separate study, a number of amino acid substitutions in signal peptides of secreted proteins where most or all present-day humans differ from the El Sidron 1253 Neandertal individual. Signal peptides direct the import of proteins into the endoplasmic reticulum, an extensive membrane network important for the transport of synthesized proteins as well as facilitation of protein folding. In the absence of other trafficking signals, membrane proteins carrying signal peptides are transported to the cell surface, while the signal peptide is cleaved off after the successful insertion of the polypeptide into the membrane of the endoplasmic reticulum. Gralle et al. [79] analyzed both the ancestral and the derived forms of these protein domains with respect to their efficiency in mediating transport of proteins to the cell surface. This

IDT Exhibit 1061.0144

study could not identify any significant functional differences, suggesting that also most non-synonymous changes in other protein domains may be functionally neutral.

Even though the very positions studied by Gralle et al. [79] are included in the set of identified human SNCs, we did not obtain reliable Neandertal state information from the genome shotgun data of the three Vindija individuals. In the 19,780 SNCs falling in coding sequences in the present genome analysis, we find 175 changes in signal peptides, 106 are non-synonymous. For 91 of these positions Neandertal shows the derived state, and for 15 sites Neandertal shows the ancestral state. All the latter changes are known to be polymorphic in modern humans (`dbSNP 130`, [212]) and are therefore possibly functionally equivalent and may not have been relevant in modern human evolution.

6.4.1 Amino acid substitutions

We excluded all non-synonymous substitutions where current humans are known to vary, and identified 78 fixed, i.e. not known to be variable in present-day humans with respect to `dbSNP 130`, non-synonymous amino substitutions from a total of 2,910 positions where the Neandertal carries the ancestral (chimpanzee-like) allele (table 2 on page 201). We identify five genes affected by two substitutions that either change amino acids or introduce a stop codon, and that have become fixed among humans since the divergence from Neandertals:

- *DCHS1* (CCDS7771), which encodes *fibroblast cadherin-1*, a calcium-dependent cell-cell adhesion molecule that may be involved in wound healing.
- *RPTN* (CCDS41397), which encodes *repetin*, an epidermal matrix protein that is expressed in the epidermis and particularly strongly in eccrine sweat glands, the inner sheaths of hair roots and the filiform papilli of the tongue.
- *SPAG17* (CCDS899) *sperm-associated antigen-17* that is thought to be important for the structural integrity of the central apparatus of the sperm axoneme, which is important for flagellar movement.
- *TTF1* (CCDS6948), a terminator of ribosomal gene transcription and regulator of RNA polymerase I transcription.
- *SOLH* (CCDS10410), which encodes a protein of unknown function.

It is striking that two of these genes are expressed primarily in the skin. This may suggest that modern humans and Neandertals differed with respect to skin morphology and physiology. Besides the number of changes in each gene, the potential physico-chemical impact of exchanging an amino acid in a protein is relevant for prioritizing these 78 positions. We have categorized the amino acid replacements into classes of chemical similarity (table 2 on page 201) using Grantham scores [80]. Based on the classification proposed by Li [138] scores from 0-50 are considered conservative, 51-100 are moderately conservative, 101-150 moderately radical and >151 are considered radical.

On this basis, only one of the substitutions in the five genes with multiple SNCs is considered radical, resulting in the change of codon 431 in *sperm associated antigen 17* from the ancestral aspartic acid to the derived tyrosine. A further four of the complete list of 78 nucleotide substitutions result in radical amino acid changes, 7 in moderately radical changes, 33 in moderately conservative, 32 in conservative changes and a single one affects a stop-codon (table 2 on page 201). The genes showing radical amino acid substitutions are involved in

IDT Exhibit 1061.0145

reproduction, hormone response, olfaction, and immunity – groups which have been shown in human-chimpanzee genome comparisons to have undergone positive selection [36]:

- *GREB1* (CCDS42655, *gene regulated in breast cancer 1*), an estrogen-responsive gene which is an early response gene in the estrogen receptor-regulated pathway. The amino acid substitution in *GREB1* occurs in a serine-rich region of the protein.

- *OR1K1* (CCDS35132, *olfactory receptor, family 1, subfamily K, member 1*), an olfactory receptor, has an exchange of arginine to cysteine in one of the extracellular domains of the protein.

- *NLRX1* (CCDS8416, *nucleotide-binding oligomerization domain, leucine rich repeat containing X1*) acts as a modulator of the innate immune response elicited from the mitochondria in response to viral challenge. Expression of *NLRX1* results in inhibition of the *RLH* and *MAVS*-mediated *interferon-beta* promoter activity and in the disruption of virus-induced *RLH-MAVS* interactions. The amino acid substitution is in the *NACHT* domain of the protein which is thought to interact with *MAV* in order to bring about the innate viral response.

- *NSUN3* (CCDS2927, *NOL1/NOP2/Sun domain family 3*) is a protein of unknown function which seems to have S-adenosyl-L-methionine-dependent methyl-transferase activity.

6.4.2 Stop/Start codon substitutions

We identified only one gene (*RPTN*, CCDS41397) in which a fixed, non-synonymous substitution introduces a stop codon in the human protein which is not seen in Neandertal. We examined the Neandertal minicontigs in the region surrounding the position of the human stop codon and identified only one stop codon – 108 amino acids downstream of the position at which the human stop codon is observed. The earlier mentioned human *RPTN* protein is thus shortened by 108 amino acids – from 892 amino acids in Neandertal to 784 amino acid residues in humans. A second gene (*KIAA1751*, CCDS3097) carries a stop codon showing a non-synonymous change which is known to be polymorphic in modern humans.

We identified no fixed, non-synonymous changes in start codons where Neandertal shows the ancestral allele. Just one non-synonymous change which is not fixed in modern humans was identified in *melastatin-1* (*TRPM1*). Functional variants of the *TRPM1* (CCDS10024) that use alternative start positions have been described in human tissues and may be able to compensate for the loss of the specific transcript variant [162]. *TRPM1* encodes an ion channel with the function of maintaining normal melanocyte pigmentation.

6.4.3 Indels in coding sequence

We identified 36 insertion/deletion events within coding sequences. In four cases the Neandertal is ancestral, and in all of these cases are modern humans known to be polymorphic for the position.

IDT Exhibit 1061.0146

6.5 Changes in protein-coding sequences analyzed from Denisova data

From the Denisova data, 35,523 SNCs, were identified in the coding regions of the human CCDS set. Thus, about 1.8 times more SNCs were covered in the Denisova data than in the Neandertal data. Of these 21,354 were synonymous substitutions and 14,169 non-synonymous substitutions. In the Neandertal data a slightly higher proportion (43% vs. 40%) of non-synonymous substitutions was obtained, indicating non-random sampling in at least one of the data sets.

6.5.1 Amino acid substitutions

Again excluding all non-synonymous substitutions where current humans are known to vary (`dbSNP v131`), we identified 129 fixed, non-synonymous amino substitutions from a total of 2,176 positions in 119 genes where the Denisova individual carries the ancestral (chimpanzee-like) allele (table 3 on page 204). We identify 10 genes affected by two amino acid substitutions that are consistent with being fixed in present-day humans since divergence from the common ancestors of Denisovans:

- *AN30A* (CCDS7193), *Ankyrin repeat domain-containing protein 30A*
- *HPS5* (CCDS7836), *Hermansky-Pudlak syndrome 5 protein*
- *ITB4* (CCDS11727), *Integrin beta-4 precursor*
- *PIGZ* (CCDS3324), *GPI mannosyltransferase 4*
- *RGS14* (CCDS43405), *Regulator of G-protein signaling 14*
- *RP1L1* (CCDS43708), *Retinitis pigmentosa 1-like 1 protein*
- *SPTA1* (CCDS41423), *Spectrin alpha chain, erythrocyte*
- *SSH2* (CCDS11253), *Protein phosphatase Slingshot homolog 2*
- *TTF1* (CCDS6948), *Transcription termination factor 1*
- *ZN333* (CCDS12316), *Zinc finger protein 333*

Interestingly, two of these genes are associated with skin diseases (*HPS5* and *ITB4*), which is similar to the high representation of genes associated with skin morphology and physiology in the Neandertal-based catalog presented in the section above (page 145). Using Grantham scores to categorize the 129 amino acid replacements into classes of chemical similarity [80, 138], we classified 54 sites as conservative (scores of 0-50), 65 as moderately conservative (scores of 51-100), 8 as moderately radical (scores of 101-150), and 1 as radical (score of >151) (table 3 on page 204). The only gene with an amino acid substitution that is classified as radical is *OR1K1* (*olfactory receptor, family 1, subfamily K, member 1*, CCDS35132), an olfactory receptor with a replacement of arginine by cysteine in one of the extracellular domains. This change was also identified in the Neandertal data described above.

IDT Exhibit 1061.0147

6.5.2 Stop/Start codon substitutions

We identified one fixed non-synonymous change in a stop codon. In *OLM2B* (*Olfactomedin-like protein 2B precursor*, CCDS1236) the loss of a stop-codon at amino acid 470, a change that is observed in all present-day humans, is required for the protein to contain the olfactomedin-like domain (amino acids 493-750). In Denisova, the ancestral stop-codon is present and the protein therefore does not include this domain.

We did not identify fixed, non-synonymous changes in start codons where the Denisova individual carries the ancestral allele. However, at one gene, *Riboflavin kinase* (*RIFK*, CCDS35044), Denisova carries an ancestral start codon (`dbSNP 131` rs2490582) that is lost in about 98% of present-day humans. In addition, there are two genes where some (but not all) present-day humans have gained a start codon relative to Denisova. This includes the *melastatin-1* gene (*TRPM1*, CCDS10024; `dbSNP 131` rs4779816 derived allele frequency 88%) and *zinc finger protein 211* (*ZNF211*, CCDS12957; `dbSNP 131` rs9749449 derived allele frequency 77%). The difference in *TRPM1* was also observed for Neandertal (see section 6.4.2 on page 146). *ZNF211*, is an as-yet uncharacterized zinc finger protein probably involved in transcriptional regulation.

6.5.3 Insertions and deletions in coding sequence

We identified 69 insertion/deletion events within coding sequences. In 15 cases the Denisova state is ancestral, and for 14 of these, present-day humans are not known to vary in `dbSNP 131` (table 6.1 on the facing page). Twelve of these 14 InDels are 3 bases long. Of these, 6 delete exactly one amino acid and the other 6 affect two amino acids while maintaining the reading frame.

In *HADHB/ECHB* (CCDS1722, *hydroxyacyl-CoA dehydrogenase / 3-ketoacyl-CoA thiolase / enoyl-CoA hydratase, β subunit*), the β subunit of an enzyme essential for the metabolism of long-chain fatty acids, the first amino acid, which is part of the mitochondrial transit peptide region of the protein, is removed. Since the mitochondrial transit peptide is responsible for the transport of the protein from the cytoplasm to the mitochondrion, it is possible that this change affects the cellular localization or the efficiency of the localization of this protein. Mutations in this gene are associated with hypoglycaemia, hypotonia and lethargy [164].

An entire codon is deleted from *RTTN* (CCDS42443, *rotatin*), a protein required for the early developmental processes of left-right specification and axial rotation and which may play a role in notochord development [65]. Examples of other three-base deletions are in *AHNK* (CCDS31584, *Desmoyokin*), a protein involved in neuroblast differentiation, in *EME1* (CCDS11565, *essential meiotic endonuclease 1 homolog 1*), involved in DNA replication and repair, *SNG1* (CCDS13989, *synaptogyrin 1*) involved in short and long-term regulation of neuronal synaptic plasticity, and the spermatogenesis-associated protein *SPT21* (CCDS172, *spermatogenesis associated 21*). Interestingly, several genes in which present-day humans appear to have undergone deletions while Denisova carries the ancestral state are involved in neuronal development and function, spermatogenesis and metabolism.

Particularly striking are single base deletion within coding sequence, as these destroy reading frame. One such InDel that we detected is in one of the final codons of the membrane protein *ADAM8* (CCDS31319, *a disintegrin and metalloproteinase domain 8*). This InDel is predicted to lead to a change of frame in the cytoplasmic portion of the protein, 6 amino acids from the derived C-terminus. Disintegrin and metalloprotease proteins are involved

IDT Exhibit 1061.0148

Table 6.1: Table of 15 insertion/deletion changes in coding sequences where Denisova has the ancestral, chimpanzee-like, state. For type 'del' (deletion) the sequence reported in column 'Seq.' was lost on the human lineage. For type 'ins' (insertion), the reported sequence was gained on the human lineage.

		Human (hg18)			Database identifier		
Type	Seq.	Chr	Start	End	CCDS ID	SwissProt	Exon
del	CTT	1	16599892	16599892	172	SPT21	9
del	ACT	2	26330629	26330629	1722	ECHB	1
del	GAG	6	151715809	151715809	5229	AKA12	3
del	GAC	8	101275635	101275635	34930	SPAG1	9
del	C	10	134926669	134926669	31319	ADAM8	23
del	CTC	11	62060131	62060131	31584	AHNK	1
del	AGC	17	45807977	45807977	11565	EME1	1
del	CTC	18	66014830	66014830	42443	RTTN	7
del	ATC	19	14913983	14913983	12320	OR7C2	1
del	CAG	19	55573634	55573634	42593	NR1H2	4
del	ACT	19	58146287	58146287	33096	Z816A	3
del	CAA	22	38107768	38107768	13989	SNG1	4
ins	AGC	2	79990299	79990302	42703	CTNA2	6
ins	GCG	2	95210767	95210770	42712	ZNF2	4
ins	G	17	21087327	21087328	42286	GTL3B	3

in a variety of biological processes involving cell-cell and cell-matrix interactions, including fertilization, muscle development, and neurogenesis. *ADAM8* has also been linked to inflammation and remodelling of the extracellular matrix (including cancers and respiratory diseases) [117]. A single base pair insertion in *chromosome 17 open reading frame 103* (also known as *gene trap locus F3b – GTL3B*, CCDS42286), a protein of unknown function, also results in a change in reading frame.

6.6 Changes in non-protein-coding sequences

In the following, I will discuss changes in the non-coding parts, the so-called untranslated regions (UTRs), of protein-coding human transcripts, microRNAs which are one class of short non-coding RNAs and associated with post-transcriptional regulation [219, 104], as well as Human Accelerated Regions (HARs) which are defined as regions of the genome conserved throughout vertebrate evolution, but which have changed radically since humans and chimpanzees split from their common ancestor [20, 31, 174, 175, 178]. There are different other transcribed and non-transcribed regions of the genome which are of functional relevance, the three features presented are just a small and probably even a non-representative subset.

6.6.1 5' UTR substitutions and insertion/deletions

Neandertal data

We have reliable Neandertal sequence data for 2,616 of the 12,045 substitutions changes in 5' untranslated regions of human transcripts. Of these, 42 affect positions where the ancestral allele is observed in Neandertals, and humans are fixed derived (`dbSNP 130`).

IDT Exhibit 1061.0149

Two genes show multiple changes. *TMEM105* (CCDS11781, *transmembrane protein 105*), a transmembrane protein with no known function, has three changes in the 5' UTR, and *SLC25A2* (CCDS4258, *solute carrier family 25 member 2*), which is thought to have a role in metabolism as a mitochondrial transport protein, has two such changes. Neandertal state information was also obtained for 71 of 810 InDels in 5' UTRs; three of them show the ancestral state retained in Neandertals (*ARHGEF11 – Rho guanine nucleotide exchange factor (GEF) 11*/CCDS1163, *ZNF564 – zinc finger protein 564*/CCDS42505, *RIBC2 – RIB43A domain with coiled-coils 2*/CCDS14066) while present-day humans are fixed derived.

Denisova data

We have Denisova sequence data for 5,654 of the 12,045 substitutions in 5' untranslated regions occurring on the human lineage. Of these, there are 66 positions in 64 genes where the ancestral allele is observed, and present-day humans are consistent with being fixed for the derived allele.

Two genes each carry two changes in the 5' UTR: *ETS2* (CCDS13659, *human erythroblastosis virus oncogene homolog 29*), a transcription factor that is involved in stem cell development, apoptosis and tumorigenesis, and *FNBP4* (CCDS41644, *formin binding protein 4*) a gene with roles in a cell adhesion and G-protein coupled receptor signaling. Denisova state information was also obtained for 198 of 810 InDels in 5' UTRs. For 24 of these (each in a different gene) the Denisova individual retains the ancestral state while present-day humans are fixed for the derived allele (`dbSNP 131`).

6.6.2 3' UTR substitutions and insertion/deletions

Neandertal data

We have reliable Neandertal sequence data for 18,909 of 55,883 substitutions in 3' UTRs. Among these, there are 190 positions where Neandertal shows the ancestral state and modern humans are fixed derived (`dbSNP 130`).

Twelve genes show multiple substitutions, with one gene having four substitutions (*CCDC117 – coiled-coil domain containing 117*, CCDS13846) and three genes having three substitutions each (*ATP9A – ATPase, class II, type 9A*/CCDS33489, *LMNB2 – lamin B2*/CCDS12090, *RCOR1 – REST corepressor 1*/CCDS9974). We also identify 784 of 5,972 InDels in 3' UTRs, 33 of which show the ancestral state in Neandertals while modern humans are fixed derived. Each of the 33 InDels is found in a different gene.

Denisova data

We have Denisova data for 26,113 of 55,883 SNCs in 3' UTRs. Among these, there are 283 positions (in 234 genes) where the Denisova individual shows the ancestral state and present-day humans are consistent with being fixed for the derived allele (`dbSNP 131`).

We also find 37 genes with multiple substitutions, with one gene having 4 substitutions (*PRDM10 – PR domain containing 10*, CCDS8485), 10 genes with 3 substitutions, and 26 genes with 2 substitutions. The protein encoded by *PRDM10* is a transcription factor that is implicated in somite and craniofacial formation during embryonic development [168] and that may be involved in the development of the central nervous system as well as in the pathogenesis of gangliosidosis (GM2, neuronal storage disease) [215].

IDT Exhibit 1061.0150

We also have Denisova data for 1,271 of 5,972 InDels in 3' UTRs, 109 of which show the ancestral state in Denisova while present-day humans are fixed for the derived allele. These InDels are located in 108 different genes. Two InDels are present in the 3' UTR of *MPP5* (*MAGUK p55 subfamily member 5*, CCDS9779), a protein that may play a role in tight junction biogenesis and in the establishment of cell polarity in epithelial cells.

6.6.3 microRNAs

MicroRNAs (miRNA) are small non-coding RNAs that regulate gene expression by mRNA cleavage or repression of mRNA translation. MiRNAs have been shown to have important role in mammalian brain and embryonic development [219, 104]. In 1,685 miRNAs annotated in `Ensembl 54` (including 670 `miRBase`-derived miRNAs [85]), I identified 357 single nucleotide changes and 17 insertion/deletion events from the whole genome alignments which occurred on the human lineage.

Neandertal data

We have Neandertal sequence data for 103 of the 357 single nucleotide changes. In 88 cases the Neandertal carries the derived allele. The remaining 15 alleles are ancestral. In only one case, ENSG00000221170 (*hsa-mir-1304*), is the Neandertal state ancestral and the human state fixed derived (`dbSNP 130`). For *hsa-mir-1304*, the substitution occurs in the seed region of the mature miRNA (see figure 6.4 on the next page), suggesting that it is likely to have altered target specificity in present-day humans relative to Neandertals and the outgroups [130]. Folding is unlikely to be changed since base pairing is unaffected by the substitution.

Reliable Neandertal sequence data is also available for two of the 17 insertion/deletion events in miRNAs. In ENSG00000211530 (*AL354933.8*) the Neandertal has the derived allele, while in ENSG00000212045 (*AC109351.3*) the allele is ancestral while modern humans are fixed derived. *AC109351.3* has a large loop that is one base shorter in Neandertal than in human (figure 6.4 on the following page). This change is not expected to alter folding near the mature miRNA or to change target specificity of the putative miRNA.

Denisova data

We have Denisova sequence for 143 of the 357 single nucleotide differences, and Denisova shows the derived state of the human reference sequence (hg18) at 125 of these sites. Out of the remaining 18 sites, 17 are polymorphic in present-day humans, while one change in miRNA *hsa-mir-564* is fixed in present-day humans for the derived allele (`dbSNP 131`). The substitution, however, is unlikely to affect microRNA function as it is located in a small bulge outside of the mature sequence. This substitution does however slightly change the estimated minimum free energy of *hsa-mir-564*, indicating that the derived version is slightly more stable (figure 6.5 on page 153).

Denisova sequence is also available for 5 of the 17 insertion/deletion events in miRNAs that occurred on the human lineage. In one case, *hsa-mir-1260*, Denisova carries the ancestral allele while present-day humans are apparently fixed for an insertion of adenine. This insertion is outside of the mature sequence in an inferred loop structure and is thus not likely to affect function (figure 6.6 on page 153).

IDT Exhibit 1061.0151

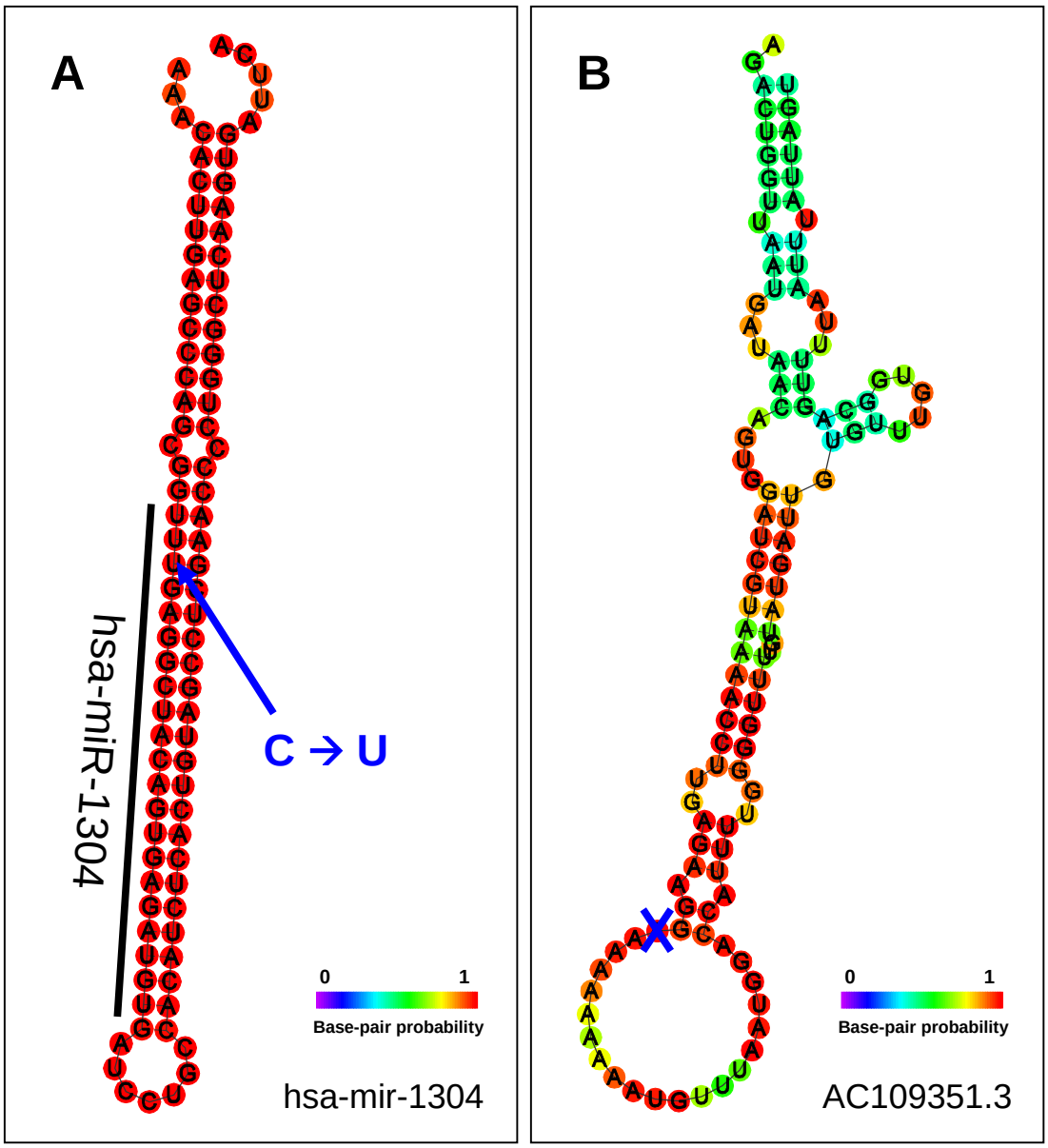

Figure 6.4: `RNAfold` (`http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi`, [86]) output for the two microRNAs showing the ancestral state in Neandertal while being fixed derived in modern humans with respect to `dbSNP 130`. For *hsa-mir-1304* **(A)** the ancestral cytosine is observed in Neandertal while modern humans are fixed derived for thymine (uracil in the microRNA transcript). The change is located in the seed of the mature microRNA, thus it likely alters target specificity of the derived version present in modern humans. In *AC109351.3* **(B)**, Neandertal shows an ancestral adenine as an additional base in the big bulge. This change is not likely to effect folding and function of this putative microRNA.

IDT Exhibit 1061.0152

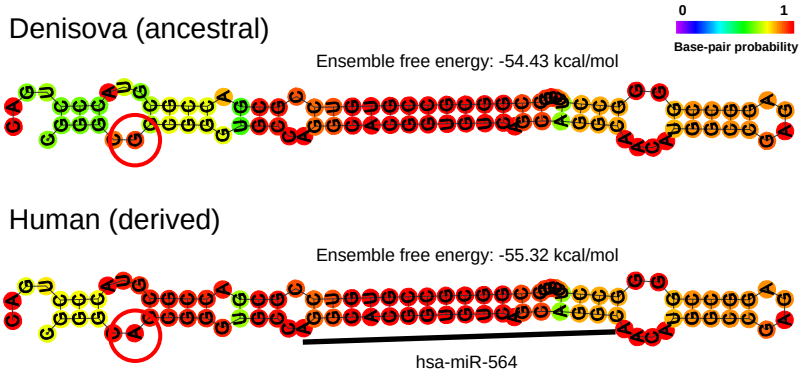

Figure 6.5: `RNAfold` (`http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi`, [86]) output for the human microRNA *hsa-mir-564* showing the ancestral state in Denisova while being fixed derived in modern humans with respect to `dbSNP 131`. The free energy of the thermodynamic ensemble is −54.43kcal/mol for the ancestral version, and slightly better for the derived version with −55.32kcal/mol. The minimum free energy structure (MFE) has a frequency of 22.05% and 31.33% in the ensemble for Denisova and human, respectively.

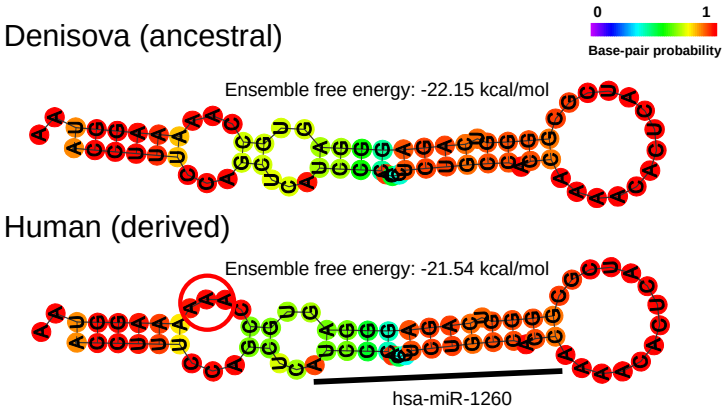

Figure 6.6: `RNAfold` (`http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi`, [86]) output for the human microRNA *hsa-mir-1260* showing the ancestral state in Denisova (i.e. missing a human derived adenine insertion) while being fixed in modern humans with respect to `dbSNP 131`. The free energy of the thermodynamic ensemble is −22.15kcal/mol for the ancestral version, and slightly worse for the derived version with −21.54kcal/mol.

IDT Exhibit 1061.0153

6.6.4 Human Accelerated Regions

Human Accelerated Regions (HARs) are defined as regions of the genome conserved throughout vertebrate evolution, but which have changed radically since humans and chimpanzees split from their common ancestor. Whether the acceleration is functionally relevant and thus driven by positive selection or whether it is a byproduct of biased gene conversion is a matter of intense debate [72, 179, 55, 54, 160]. Biased gene conversion may happen during recombination of two chromosomes/DNA molecules when two new double stranded DNA molecules with partially false-paired nucleobases (heteroduplexes) are created [3]. These false-paired bases are identified by cellular repair mechanisms, which preferentially repair C-A mispairings to C-G and G-T misparings to G-C [18]. Even though molecular and biochemical processes are not yet completely understood [217, 101], this "biased" conversion/repair is argued to be a result of the deamination of methylated cytosines in vertebrate genomes [29, 1], which if deaminated, a C^{5me}-G pairing results in a T-G mispairing. Should human accelerated regions largely be the reminiscence of recombination hotspots, i.e. frequent sites of homologous recombination of parental chromosomes, then biased gene conversion may have driven a larger number of changes in these regions.

In order to determine whether the acceleration took place before or after the human-Denisova-Neandertal split, we[7] examined a total set of 2,613 Human Accelerated Regions identified in five different studies [20, 31, 174, 175, 178].

Neandertal data

We are restricted in our coverage of the HAR regions due to the filtering of the whole-genome alignments (see section 6.3.1 on page 142) and by the Neandertal coverage within each HAR region. We identified a total of 8,949 single nucleotide changes and 213 InDels on the human lineage in these HARs. Neandertal sequence was available for 3,259 sites (3,226 substitutions and 33 InDels).

If we calculate the percentage of positions showing the derived state (2,977; 91.35% [90.32%, 92.28%] – 95% Wilson two-sided confidence interval for a proportion including continuity correction [244, 159]), we observe that this is significantly higher than for the genome-wide set (87.86% [87.82%, 87.90%]) of all derived substitutions (2,813,802) and all derived deletions (60,248). When we count the percentage of positions in which Neandertal shows the derived state only at the positions which may be the sites of biased gene conversion (A/T in chimpanzee to G/C in human, 62% of the positions under consideration), we find that this effect is even more extreme. In these substitutions, we find 1,971 derived out of 2,038 (96.84% [97.49%, 96.03%]). However, such sites show also genome wide a higher frequency of derived alleles (1,714,366 out of 1,796,587 = 95.42% [95.39%,95.45%]). The incidence of changes from ancestral G/C to a derived A/T in human is 519 derived out of 683 (76.0% derived [72.57%, 79.11%]) compared to the genome-wide average of 658,717 out of 895,453 (73.56% [73.47%,73.65%]).

The fact that the vast majority of A/T to G/C changes in human accelerated regions are shared between Neandertal and human suggests that positions affected by biased gene conversion probably predate the human/Neandertal split considerably. This could be interpreted as evidence for a relocation of recombination hotspots to previously highly conserved sites, prior to the human-Neandertal split. This relocation of recombination hotspots might date back to the speciation of all human forms from their last common ancestor with chimpanzees

[7] work in collaboration with Hernán Burbano and Janet Kelso

IDT Exhibit 1061.0154

and bonobos, or the fusion of two ape chromosomes to form human chromosome 2 [248, 99]. The fusion of two chromosomes is likely to affect sites of recombination [146] and there is evidence that the fusion occurred in the common ancestor of humans and Neandertals as recently as 740,000 years ago and no more than about 3 million years ago [51]. The number of sites are to small to test for a larger effect on chromosome 2, however it is encouraging that chromosome 2 has the third highest number of annotated HARs per megabase and the highest number of total HARs.

Even though the majority of sites in HARs are observed to be derived in the Neandertal individual, there is still a considerable number of ancestral states observed (269 SNCs and 9 InDels). There are 51 positions (in 45 HARs) where Neandertal is ancestral and humans are derived and not known to vary (`dbSNP 130`). These represent recent changes that likely occurred since the human-Neandertal split.

Denisova data

The above results from the Neandertal genome analysis indicate that the acceleration of HARs predates the Neandertal-human split. Denisova sequence is available for 3,494 changes (3,445 substitutions and 49 InDels). Of these, 3,128 are derived in Denisova (89.52% [88.45%, 90.51%]), which is significantly higher than for the complete set (86.64% [86.61%, 86.67%]) of all derived substitutions (3,696,534) and all derived InDels (91,985). Thus, we continue to find that Denisova carries the derived allele more often in HARs than elsewhere in the genome. We note that confidence levels for the HARs overlap with the earlier reported Neandertal values, while being non-overlapping and smaller in Denisova for genome-wide sites. This indicates a generally larger fraction of ancestral states in Denisova compared to Neandertal, while showing equally frequent derived states in HARs.

Testing for effects of biased gene conversion, we again restricted our analysis to SNCs with A/T in chimpanzee and G/C in human (50% of covered SNCs in HARs). We find that 1,554 out of 1,719 (90.4% [88.89%,91.73%]) changes in HARs have the derived state in Denisova, which is also significantly higher than for the 1,532,287 out of 1,750,152 (87.55% [87.5%,87.6%] sites genome-wide that have the derived state in Denisova. Checking the opposite pattern (derived A/T and ancestral G/C), we obtain 1,044 out of 1,188 (87.88% [85.86%,89.65%]) for HARs, a fraction in agreement with the genome-wide average of all SNCs, and 1,620,005 out of 1,875,789 (86.36% [86.31%,86.41%]) a number lower than the genome-wide average of all sites.

For Denisova, we identify 104 positions (98 SNPs and 6 InDels) where the Denisova individual is ancestral while present-day humans are consistent with being fixed for the derived allele (`dbSNP 131`). These represent recent changes that probably occurred since the Denisova-modern human split and merit further study.

Taken together, these results support that changes in HARs tend to predate the human-Denisova-Neandertal split and that differences caused by biased gene conversion tend to be older in time. Nevertheless, the results for Denisova are different from the results in the Vindija Neandertals in that more ancestral sites are covered in the Denisova genome. From these numbers, it is unclear whether the high frequency of derived states observed in the Vindija is an artifact of sample processing (lab protocols and computational analysis) or whether they are due to real differences of the two ancient genomes. In total there are 3,722 HAR SNC sites with information on ancestral/derived states for both Neandertals and Denisovan. In 96% of the cases the observed states agree, of 153 discordant single nucleotide changes, Denisova is derived in 91 (59%) cases. In the sites covered in only one data set,

IDT Exhibit 1061.0155

Denisova shows 9.6 times more derived than ancestral sites and the Vindija Neandertals 12.5 times more derived sites. Therefore, it is likely an ancient molecule sampling bias causes the difference between data sets.

6.7 Neandertal-Denisova concordance

Of the 10,535,445 SNCs inferred to have occurred on the lineage leading to the human reference genome (hg18), 4,267,431 (40.51%) positions are covered in the Denisova data while 3,202,190 (30.39%) are covered in Neandertal. The expected overlap from random sampling is 12.31% (40.51% · 30.39%), and thus the actual overlap of 15.61% is higher than expected. We hypothesize this may be due to higher coverage of GC-rich sequences in both data sets. The overlap of InDels of 6.05% is also higher than expected from random sampling (3.16%).

The Neandertal and the Denisova specimens carry the same assigned state at SNCs in 87.91% of the ancestral positions (Neandertal = Ancestral (A) | Denisova = A) and 97.69% of the derived positions (Neandertal = Derived (D) | Denisova = D). Similarly for InDels, $p\,(Neandertal = A \,|\, Denisova = A) = 87.64\%$ and $p\,(Neandertal = D \,|\, Denisova = D) = 98.60\%$. Table 6.2 provides individual counts.

Table 6.2: Concordance between Denisova and Neandertal for single nucleotide changes (SNCs) and insertion/deletion changes (InDels) assigned to the human lineage. A = ancestral, D = derived, M = missing, N = neither chimpanzee nor human state, P = polymorphic in Denisova. Disagreements are highlighted.

Single nucleotide changes			Insertion/deletion changes		
Count	Denisova	Neandertal	Count	Denisova	Neandertal
190836	A	A	2532	A	A
32785	A	D	**365**	A	D
339171	A	M	9937	A	M
227	A	N	12	A	N
26245	D	A	**357**	D	A
1389396	D	D	25642	D	D
2279365	D	M	65957	D	M
1528	D	N	29	D	N
164555	M	A	5409	M	A
1389996	M	D	34218	M	D
3204	M	N	382	M	N
534	N	A	4	N	A
818	N	D	23	N	D
1517	N	M	458	N	M
12	N	N	56	N	N
58	P	A			
807	P	D			
2943	P	M			
1189	P	N			

Positions where the Neandertal and Denisova data disagree on the ancestral state may be of special interest (32,785 + 365 Denisova = A & Neandertal = D; 26,245 + 357 Denisova = D & Neandertal = A). These sites show a derived state in the human reference sequence, as well as the derived state in either Denisova or Neandertal but not in both. Considering

IDT Exhibit 1061.0156

that Neandertals and Denisovans form a common clade, these positions could for example be due to variation present at the time when the lineages of modern humans, Neandertals and Denisovans separated.

On positions where Neandertals show the derived allele, Neandertals are more closely related to modern humans and for positions which show the derived allele in Denisova, Denisovans are more closely related to modern humans. Since Denisovans and Neandertals are genome-wide more closely related with each other than they are to modern humans, i.e. they form a common clade, this effect is described as incomplete lineage sorting. Incomplete lineage sorting is not unexpected, especially as it also exists between for example humans, chimpanzees and gorillas [201], which are not as closely related. For Neandertals, Denisovans and humans these sites are interesting in the respect that the ancestral allele observed in Neandertals or Denisova might have been reintroduced into some present-day human populations by admixture with individuals from either Neandertal or Denisovan populations. By comparing the ratio of frequencies with which a present-day human shares the ancestral state for these sites with either Neandertal or Denisova one may therefore detect whether this individual carries a larger admixture signal for either of the two ancient hominins (section 6.8).

Of the 59,030 single nucleotide differences where Neandertal and Denisova disagree, 61 overlap with the coding regions of 63 `Ensembl` annotated genes (49 of which belong to described the CCDS longest transcript set) and result in a non-synonymous change in the amino acid sequence. Three genes have two such sites:

1. *RPTN* (*repetin*, CCDS41397), an matrix protein that is expressed in the epidermis and particularly strongly in eccrine sweat glands, the inner sheaths of hair roots and the filiform papilli of the tongue [98]. *Repetin* was described as one of five genes with two amino acid altering substitutions that have become fixed among humans since the divergence from Neandertals. The same positions are present in the derived state, however, in the Denisova specimen.

2. *RGS14* (*regulator of G-protein signaling 14*, CCDS43405), an integrator of *G protein* and *MAPKinase* (*Ras*/*Raf*) signaling [214], carries two non-synonymous substitutions that are fixed in present-day humans, ancestral in the Denisova individual, and derived in the Neandertals.

3. *ZN333* (*zinc finger protein 333*, CCDS12316) carries two non-synonymous substitutions that are fixed in present-day humans, ancestral in Denisova, and derived in the Neandertals. *ZN333* is the only known gene containing two *KRAB* domains, which function in transcriptional repression [103]. In addition to the two coding positions there are several other positions located in the introns, which are also ancestral in the Denisova individual and derived in the Neandertals.

6.8 Allele sharing of humans at sites of Neandertal-Denisova discordance

In the previous section, tens of thousand of positions where Denisovans and Neandertals disagree at sites where the human reference sequence (hg18) carries the derived allele were identified. These positions are inconsistent between lines that separated more than half a million years ago (Reich et al. [186] Supplemental Information Table S6.2). Therefore, these sites might, at least partially, reflect variation at the point of Human-Neandertal-Denisova lineage separation that segregated/drifted differently in the three lineages.

IDT Exhibit 1061.0157

For these sites it is clear that at least some human (the reference sequence in which they were identified) shows the derived allele, while Neandertal and Denisova show either the human (derived) or the chimpanzee (ancestral) state. Requiring that the ancient DNA sequences agree with the human or the chimpanzee reference reduces the chance that we are looking at a sequencing error at these sites. This is especially relevant for sites of ancient DNA damage, which tend to generate derived alleles rather than ancestral, as substitution changes are frequently due to transitions (A→G, G→A, C→T, T→C) of which G→A and C→T are also caused by ancient DNA damage.

Further, it is likely that a large proportion of sites which were polymorphic at the time when human, Neandertal and Denisovan lineages separated, fixed the derived allele in present-day in humans[8] due to drift and population bottlenecks. Thus, these differently segregating sites might have been reintroduced into some present-day human populations by admixture with either Neandertals or Denisovans. Hence, they can be used to test present-day human individuals whether they more frequently show the ancestral allele for the Denisova ancestral sites or the ancestral allele for Neandertal ancestral sites. A difference would indicate admixture contributing this excess of ancestral alleles to these populations.

Instead of using such ancestral positions, it might be more intuitive to identify positions which show a derived state in either Denisova or Neandertals, which is not observed in the human reference genome, the chimpanzee reference or the other ancient genome, and to test present-day human individuals for sharing these derived alleles. However, this would enrich for positions of sequencing error which are much more frequent in the Neandertal data (see table 6.4 on page 160).

6.8.1 Overrepresentation of Denisova ancestral alleles

A striking observation is that for these sites, Denisova is ancestral more often than Neandertal. I will first discuss the fact that there are more cases where Denisova has the ancestral and Neandertal the derived allele at sites where the human reference sequence shows a human-lineage derived allele than there are cases where Neandertal has the ancestral and Denisova the derived allele (32,785 vs. 26,245 SNCs and 365 vs. 357 InDels; table 6.2 on page 156). The difference in counts could have the following explanations:

1. A higher sequencing or alignment error for the Denisova sequence. Since most genomic sites are derived, i.e. match the reference sequence, errors will cause alignments of higher divergence and more ancestral sites.

2. An earlier split time of Denisova from the other lineages (that is, Denisova is an outgroup to both present-day humans and Neandertals) and the higher number of ancestral sites goes back to its deeper divergence from the human lineage.

3. Admixture into the Denisova individual from some archaic hominin (as supported by the mitochondrial data [121]) contributing these ancestral sites.

4. The human reference sequence being closer to Neandertals than to Denisovans because of a Neandertal-related contribution of genetic material to non-Africans, as described in Green et al. [81]. This could restore ancestral sites in the human reference and thereby reduce the number of Neandertal-specific ancestral sites.

[8] Assuming neutral evolution and constant population size, i.e. no bottlenecks and no population expansion, the average number of generations for losing an ancestral allele at $p = \frac{1}{3}$ frequency is $\frac{-4 \cdot N_e \cdot p \cdot \ln p}{(1-p)} = 2.197 \cdot N_e$ [62]. Assuming N_e between 3,100-7,500 [225] and considering a generation time of 28.6a [67] for ancient humans, an ancestral allele is lost after on average 195,000-471,000 years by drift.

IDT Exhibit 1061.0158

5. The Neandertal genome being closer to the human reference at some positions because of Eurasians contributing genetic material into Neandertals, i.e. introducing modern human derived sites in Neandertal for which Denisova is ancestral.

Explanation (1), higher sequencing or alignment error in Denisova, is unlikely to explain the data, since such processes would alter substitution frequencies (as sequencing and alignment errors cause substitutions which do not follow substitution rates of biological processes), and yet the signal is consistent across all classes for the two genomes as shown in table 6.3.

Table 6.3: Substitution frequencies for single nucleotide changes on the human lineage where Neandertal and Denisova disagree on their state. D_A columns show the substitution frequencies for sites where Denisova carries the ancestral allele and Neandertal the derived. N_A columns show the substitution frequencies for sites where Neandertal carries the ancestral allele and Denisova the derived allele.

D_A	Derived	Ancestral	Fraction	N_A	Derived	Ancestral	Fraction
588	A	C	1.80%	461	C	A	1.80%
2119	A	G	6.50%	1795	G	A	6.80%
411	A	T	1.30%	342	T	A	1.30%
1800	C	A	5.50%	1521	A	C	5.80%
2050	C	G	6.30%	1572	G	C	6.00%
8082	C	T	24.70%	6472	T	C	24.70%
7916	G	A	24.10%	6230	A	G	23.70%
2110	G	C	6.40%	1672	C	G	6.40%
1813	G	T	5.50%	1426	T	G	5.40%
737	T	A	2.20%	584	A	T	2.20%
4060	T	C	12.40%	3354	C	T	12.80%
1099	T	G	3.40%	816	G	T	3.10%
32785	All sites			26245	All sites		
10608	Transversions only			8394	Transversions only		

Further, the processing of the Denisova sample (section 6.2.2 on page 136) includes several steps that reduce its error compared to Neandertal (e.g. *UDG/EndoVIII* treatment, v4 sequencing chemistry and longer read overlaps in the read merging process). In fact, the sequencing error rate estimates obtained in Reich et al. [186], indicate that the error rate in the Denisova data set is 37 times lower than in the Neandertal data set. The alignment error is difficult to assess, especially as two different aligners have been used for the data sets.

Considering the very low sequence error rate of the Denisova sample and its estimated divergence of 12% of the lineage from present-day humans to the human-chimpanzee common ancestor, divergence adds 0.074% (1.23% human-chimpanzee SNP differences [35] $\cdot$ $0.5 \cdot 12\%$) differences to the sequencing error. This leaves Denisova with fewer substitution differences in the alignment than the present-day human samples (table 6.4 on the following page) due to their higher sequencing error rate. However, as all these numbers are actually obtained from our alignments, estimates are conditioned on the alignments and might also be biased.

InDel error rates are much lower on the Illumina sequencing platform (chapter 2 section 2.3 on page 26) and also happen at a lower evolutionary rate. If there was a higher sequence or alignment error in the Denisova sample, one might expect the count difference on InDels to be informative for validating the difference in numbers of ancestral sites. While there is an excess of 25% of ancestral Denisova SNC sites, only 2% more InDels are observed for the Denisova individual. However, gap penalties are different between `ANFO` and `BWA`, with

IDT Exhibit 1061.0159

Table 6.4: Sequencing error rates inferred for present-day humans and ancient DNA data sets obtained from the autosomes, at positions of one-to-one human and chimpanzee orthology whose common ancestor sequence is supported by at least one outgroup sequence (Reich et al. [186] Supplementary Information 2, Table 2.4).

Data set	All sites	Transversions
Denisova	0.041%	0.013%
Neandertal (Vindija)	1.547%	0.094%
∅ 5 humans	0.287%	0.176%
∅ 7 humans	0.189%	0.122%

`BWA` having higher penalties and thus putatively excluding gapped alignments from analysis. On the other hand the Denisova reads are longer and may thus allow for more gapped reads. In total 52.6% more insertion/deletion sites are covered for Denisova data compared to the Neandertal data, while 33.3% more SNCs are covered in Denisova compared to the Neandertal data. Putting numbers into perspective, the Denisova data comprises about 24% more aligned bases than the Neandertal data set. A minor alignment effect can therefore not be completely rejected. From these numbers, it is also likely that InDels might be more strongly affected by an alignment effect due to the different scoring. Considering the low number of InDel events observed, it is therefore reasonable to exclude them from analysis.

Explanation (2), that Denisovans are an outgroup to both present-day humans and Neandertals, is also unlikely to explain the data, since, as shown by other analyses in Reich et al. and also clear in figure 6.1 on page 133, Neandertals and Denisovans are sister groups on a genome-wide scale.

Explanations (3), (4) and (5), which involve ancient admixture events, could easily contribute to the higher number of Denisova ancestral sites. Archaic admixture into the Denisova individual (3) would be an explanation for the high divergence time of its mitochondrial genome (see section 6.1 on page 132 and [121]) and would also contribute to the number of sites where the human reference is derived but Denisova ancestral. Due to the Neandertal contribution of genetic material to non-Africans (4) as described in Green et al. [81], the non-African proportion of the composite human reference sequence will be closer to Neandertals than to Denisovans. This effect will decrease the number of sites derived in the reference but ancestral in Neandertal, while not affecting the number of derived sites in the reference that are ancestral in Denisova (thus also causing a relative increase of sites ancestral in Denisova). The same is true for modern human admixture into the Neandertal sequence (5), which introduces modern human derived sites into Neandertal and thereby increases the number of sites where only Denisova is ancestral.

While the archaic admixture (3), which might be assumed to predate a putative admixture event with modern humans, alters the Denisova genome and therefore adds Denisova ancestral sites, it does not impact the frequency with which a present-day individual may match this Denisova ancestral state due to archaic admixture on the Denisovan lineage. If archaic admixture was introduced to the Denisova genome after a putative admixture event with modern humans, the additional ancestral Denisova sites from this late archaic admixture cannot be found in present-day humans and the ancestral allele sharing with Denisova would always be lower for Denisova than the ancestral allele sharing with Neandertals. This could be tested if a human population without any admixture from Neandertals and Denisovans was available.

IDT Exhibit 1061.0160

The Neandertal admixture into the human reference sequence (4) will affect the frequency of matching the Neandertal ancestral state, as specifically those sites of variation at the point of Neandertal-Denisova lineage separation contributed by Neandertals to present-day humans will be depleted in the above identified positions. Denisova ancestral sites observed due to modern human derived sites in the Neandertal sequence (5) may only influence the frequency with which a present-day individual matches the Denisova ancestral state if these derived sites of the human reference are private to some human populations. Due to (4) and (5), I will specifically analyze the impact of the human reference sequence.

6.8.2 Testing twelve present-day human individuals

To learn more about the relationships of a diverse group of present-day humans to these two ancient hominins, one can extract the genotype of an individual P of known ancestry for the identified Denisova or Neandertal ancestral positions, i.e. determine whether the individual P from some population matches the ancestral or the derived state. Determining the genotype is equivalent to generating a virtual 6-way alignment P-hg18-Denisova-Neandertal-Chimpanzee-Outgroups. Since our catalog ascertained sites where hg18 has a derived allele (D) specific to the human lineage and Denisova and Neandertal are discordant for the derived (D) and ancestral (A) allele, we only analyze sites with the base patterns $?DADAA$ and $?DDAAA$, using the terminology shown in figure 6.7 on the following page.

To quantify the rate at which Neandertal or Denisova share ancestral alleles with a present-day human P at these sites, I compute two rates: the fraction of P carrying the ancestral allele at sites where the Neandertal genome is ancestral, $N_A = ADDAAA/(DDDAAA + ADDAAA)$ and the same fraction for sites where the Denisova genome is ancestral, $D_A = ADADAA/(DDADAA + ADADAA)$. These statistics allow testing whether individual P matches ancestral alleles in Neandertal or Denisova more often, controlling for the fact that ancestral alleles occur with different numbers for these two ancient hominins. Further, the ratio N_A/D_A corrects for different population histories of each individual, as well as any processing effects such as differences in error rate which may cause the individual to share the derived state with the human reference sequence more or less frequently. Therefore, the ratio provides a value comparable between different present-day humans.

I obtained each P from the low-coverage sequencing data of 5 present-day humans published with Green et al. [81] (Han, French, Papuan, San and Yoruba; section 6.2.4 on page 139) and 7 additional individuals sequenced for Reich et al. [186] (Cambodian, Native Karitiana, Mbuti, Bougainville Melanesian, Mongolian, Papuan and Sardinian; section 6.2.4 on page 140). From the reads aligned for each of these individuals to the human (hg18) and chimpanzee (pantro2) reference sequences, I extracted the genotype at the identified Neandertal-Denisova discordant positions from both alignments. At sites with multiple read coverage, I used the base from the read with the highest sum of quality scores, and required the base obtained from the human and chimpanzee alignments to agree. Taking the read of the highest sum of quality scores is motivated by the sampling from possibly heterozygous regions. Calculating a consensus or considering the alignment with the higher alignment score would bias against variable sites, or introduce a bias towards the reference sequence (see also earlier considerations on handling PCR duplicates in chapter 3 section 3.10 on page 64).

Variance estimates are obtained by a Block Jackknife approach (100 blocks) [123, 187]. Briefly, I sorted all sites by physical position, i.e. genome coordinates, divided them into 100 blocks by their coordinate (each with an equal number of sites and a block size larger than the extent of linkage disequilibrium of sites), and calculated the statistics for all 100

IDT Exhibit 1061.0161

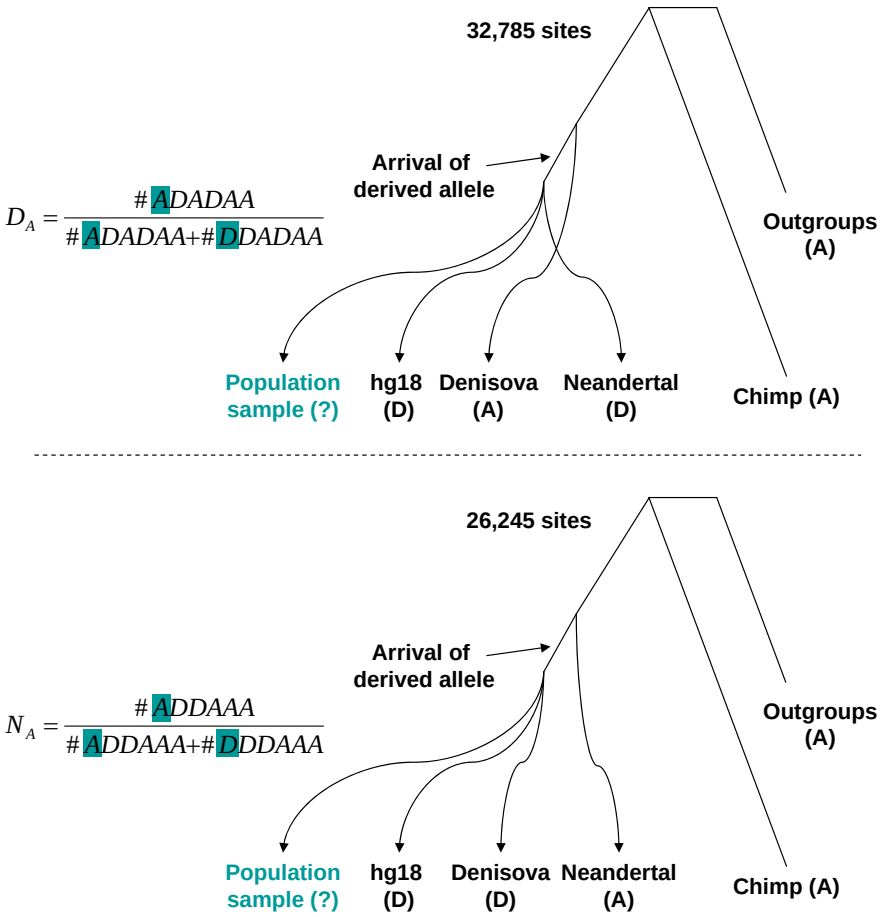

Figure 6.7: Genealogies consistent with Neandertal-Denisova discordant site classes and the calculation of ancestral allele sharing of an individual P with the Denisova (D_A) or Neandertal genomes (N_A). These two statistics allow testing whether P matches ancestral alleles in Neandertal or Denisova more often, controlling for the different number of ancestral alleles identified from the two ancient hominins. The ratio N_A/D_A provides a measure on how much closer P is to Neandertals as compared to the Denisova individual at these sites. By taking the ratio of N_A and D_A, population demographic and sample specific effects cancel and a value comparable between different individuals P is obtained.

IDT Exhibit 1061.0162

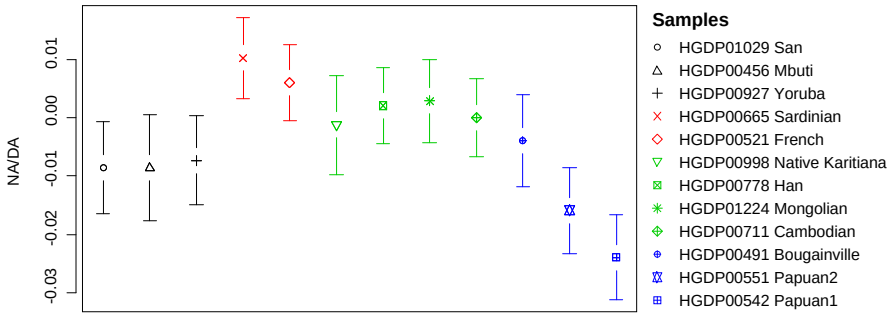

Figure 6.8: The ratio of N_A/D_A (error bars give ± 1 standard deviation) for sites obtained from the catalog of changes shows a distinct relationship of present-day Africans and non-Africans to archaic hominins. Without a reference bias, N_A/D_A should be consistent with zero when equal or no admixture is detected in an individual.

different selections of 99 blocks (i.e. leaving one out). I then took the variance observed over the 100 samples, multiplied it by 100, and determined the square root (for obtaining a standard deviation).

Figure 6.8 shows N_A/D_A for each sample with ± 1 standard deviation. Ignoring exact values, all European and most Eurasians samples show a stronger allele sharing with Neandertals, while the Papuan individuals show a stronger allele sharing with the Denisova individual. The three African individuals are intermediate. Considering actual values, Africans have a negative point estimate for N_A/D_A, suggesting more sharing of Denisova ancestral sites. Based on the results of the analyses of the Neandertal genome in Green et al. [81], one might have expected the African samples to be equally closely related to Neandertals and Denisovans (that is, the N_A/D_A ratio should be consistent with zero). The difference from zero could be due to an African admixture involving relatives of Denisovans, or the analysis being affected by a bias of the human reference which is comprised of non-African segments, which are closer to Neandertal and result in a reduction of introgressed N_A sites.

6.8.3 Generating an African catalog

To minimize any effect of the Eurasian ancestry in hg18 on this analysis, which might cause the shift towards negative N_A/D_A values in figure 6.8 (see arguments (4) and (5) in section 6.8.1 on page 158), I generated a new catalog of positions that changed on the human lineage in which I replaced the human reference base of the whole genome alignments by bases that I required to agree in the Mbuti and San individuals and which I extracted from human and chimpanzee short read alignments as described above. Thus, I sampled the read with the highest sum of base quality scores and considered only positions where the base obtained in the chimpanzee alignment disagrees with the base obtained from the human alignments.

This results in a catalog of sites where the two African individuals consistently show the derived allele. I obtained 5,711,830 single nucleotide changes on the human lineage (compared to 10,535,445 for the hg18-based catalog). There are 15,173 D_A sites where Denisova is ancestral and Neandertal derived and 12,819 N_A sites where Denisova is derived and Neandertal ancestral. The difference in counts is less than that obtained using hg18 (24.9% excess of D_A sites versus 18.4% excess in the African-genome-based catalog), suggesting that some of the earlier observed asymmetry may be due to the inclusion of non-African segments of the human reference sequence, which from Green et al. [81], we know are more

IDT Exhibit 1061.0163

Table 6.5: N_A and D_A counts from sites ascertained using the San and Mbuti individuals. These two must not show any ancestral states, as derived positions were identified for those two individuals.

Sample	Region	$N_{A,A}$	$N_{A,D}$	$D_{A,A}$	$D_{A,D}$
HGDP01029 San	Africa	0	12819	0	15173
HGDP00456 Mbuti	Africa	0	12819	0	15173
HGDP00927 Yoruba	Africa	1494	3981	1744	4659
HGDP00665 Sardinian	Europe	1222	2763	1303	3407
HGDP00521 French	Europe	1596	3797	1682	4613
HGDP00998 NatAmerican	Asia	894	2092	920	2528
HGDP00778 Han	Asia	1555	3871	1666	4674
HGDP01224 Mongolian	Asia	1210	2857	1302	3420
HGDP00711 Cambodian	Asia	1291	3074	1382	3779
HGDP00491 Bougainville	Melanesia	1294	2951	1404	3544
HGDP00542 Papuan1	Melanesia	1518	3843	1770	4509
HGDP00551 Papuan2	Melanesia	1264	2838	1481	3364

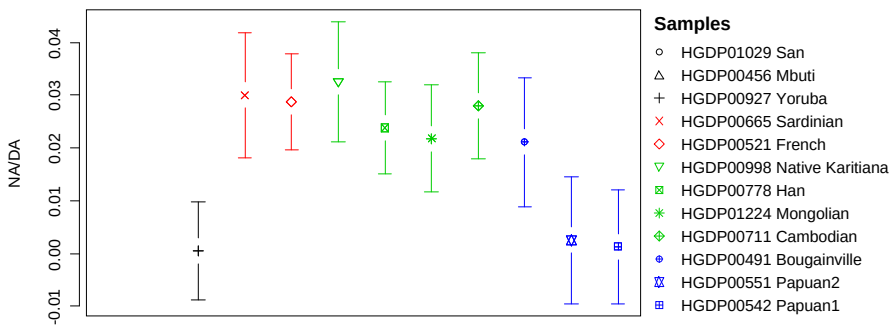

Figure 6.9: The ratio of N_A/D_A (error bars give ± 1 standard deviation) for sites obtained from the African catalog of changes again shows a distinct relationship of present day Africans and non-Africans to archaic hominins.

closely related to Neandertal. Table 6.5 provides counts observed for these sites in the twelve present-day human samples. By design San and Mbuti can not show any ancestral states for these positions and define the number of Denisova and Neandertal ancestral sites.

Figure 6.9 shows the result for the N_A/D_A analysis using this "African" catalog of positions. With sites ascertained from the two deepest-divergence human populations, the value for the remaining African individual, the Yoruba HGDP00927, is centered around zero. This indicates that the earlier observed negative shift was a result of the Neandertal introgressed sites in the human reference sequence or non-African admixture in the Neandertal genome rather than a result of an affinity to the Denisova genome. Further, this indicates that no significant proportion of archaic admixture *after* a putative admixture event with modern humans is present in the Denisova genome (which would cause a positive shift in N_A/D_A, see argument (3) in section 6.8.1 on page 158).

The majority of non-Africans being more closely related to Neandertals than to Denisovans ($N_A/D_A > 0$) is distinctly observed (Sardinian HGDP00665 Z=2.5 standard deviations, French HGDP00521 Z=3.2, Native Karitiana HGDP00998 Z=2.9, Han HGDP00778 Z=2.7, Mongolian HGDP01224 Z=2.2, Cambodian HGDP00711 Z=2.8). However, the two individuals from Papua New Guinea (Papuan1 HGDP00542 Z=0.1 and Papuan2 HGDP00551 Z=0.2) are, like the one remaining African sample (Yoruba HGDP00927 Z=0.1), consistent

IDT Exhibit 1061.0164

Table 6.6: N_A and D_A statistics from sites ascertained using the San and Mbuti individuals. Since San and Mbuti represent the deepest divergences in human populations, all populations are about equally distant from this artificial reference and the actual N_A and D_A values are similar between individuals. N_A and D_A values are however still affected by sample specific effects like sequencing error.

Sample	Region	N_A	*sd*	D_A	*sd*	N_A/D_A	*sd*
HGDP01029 San	Africa	-	-	-	-	-	-
HGDP00456 Mbuti	Africa	-	-	-	-	-	-
HGDP00927 Yoruba	Africa	27.29%	1.09%	27.24%	0.92%	0.05%	0.90%
HGDP00665 Sardinian	Europe	*30.66%*	1.13%	27.66%	1.03%	3.00%	1.14%
HGDP00521 French	Europe	*29.59%*	1.23%	26.72%	0.99%	2.87%	0.91%
HGDP00998 NatAmerican	Asia	*29.94%*	1.29%	26.68%	1.13%	3.26%	1.18%
HGDP00778 Han	Asia	*28.66%*	1.04%	26.28%	0.90%	2.38%	0.85%
HGDP01224 Mongolian	Asia	*29.75%*	1.11%	27.57%	0.92%	2.18%	1.02%
HGDP00711 Cambodian	Asia	*29.58%*	1.07%	26.78%	1.07%	2.80%	1.13%
HGDP00491 Bougainville	Melanesia	*30.48%*	1.14%	*28.38%*	1.07%	2.11%	1.21%
HGDP00542 Papuan1	Melanesia	*28.32%*	1.22%	*28.19%*	0.94%	0.13%	1.12%
HGDP00551 Papuan2	Melanesia	*30.81%*	1.22%	*30.57%*	1.08%	0.25%	1.13%

with either compensating admixture proportions with both the Neandertal and Denisova lineages, or no admixture at all. The individual from Bougainville Island is also not different from $N_A/D_A = 0$ when considering a typical Z-score cutoff of two (HGDP00491 Z = 1.7), but intermediate between the two individuals from Papua New Guinea and other EurAsians.

Taking into account that Neandertal admixture was observed for Papuan1 (HGDP00542) and described by Green et al. [81], N_A/D_A being not significantly different from zero for all three Melanesians is likely to reflect a second gene flow event from relatives of Denisovans into the ancestors of Melanesians, which balances out the effect of gene flow into the ancestors of all non-Africans from the perspective of this analysis.

Since I have used two of the deepest divergence human populations to identify the positions analyzed in the N_A/D_A ratio, all populations are about equally distant from this artificial reference and the N_A and D_A values are similar between individuals (table 6.6). Considering actual N_A and D_A values, which are still affected by sample specific effects like sequencing error, it is clear that the Yoruban individual (HGDP00927) has the lowest N_A value and a very similar D_A value. Thus, he shows neither Neandertal nor Denisova admixture, while all non-Africans show increased N_A values, and Melanesian individuals in addition also increased D_A values.

One may quantify the fraction of N_A and D_A sites that carry the Neandertal admixture signal for non-Africans and the Denisovan admixture signal for Melanesians, by fixing one of the parameters, i.e. N_A or D_A, and thereby scaling the other to be comparable between different individuals. Enforcing identical N_A values for all non-African individuals, the D_A values of Papuan1 and Papuan2 are 9.7% (±3.7%) and 9.3% (±3.9%) increased compared to the non-African samples. Enforcing identical D_A values for all individuals excluding Melanesians, the Neandertal admixture signal on N_A not present in the Yoruban individual can be quantified with 9.1% (±3.6%). This indicates similar fractions of N_A and D_A sites which could serve as admixture markers. Considering the ascertainment scheme of these sites from human lineage derived positions, they are however unlikely to represent a genome-wide average and thus these fractions do not measure the actual genome-wide admixture proportions.

Although error bars are fairly large using this method, the African individual shares fewer

IDT Exhibit 1061.0165

ancestral alleles with Neandertal than do all non-African populations, supporting the signal of Green et al. [81]. Further, Melanesians, especially the two Papuan individuals, show a signal of Denisovan admixture not shared with other sampled populations, a result in agreement with the D-statistics for population pairs presented in Reich et al. [186]. This supports the idea that a larger proportion of the Neandertal-Denisovan discordant sites go back to sites that reflect variation at the point of Human-Neandertal-Denisova lineage separation, which then segregated differently in these lineages.

6.9 Summary and conclusions

Due to the short nature of ancient DNA molecules as well as their low copy-number relative to co-extracted environmental DNA, high-throughput approaches provide a tremendous advantage over traditional sequencing approaches in that they enable a complete characterization of an ancient DNA extract. As the cost of sequencing decreased, it became feasible to analyze entire genomes from ancient samples, like the genomes of Neandertals [81] and the Denisova phalanx [186] discussed in this chapter. The high-throughout data generated from ancient DNA libraries includes adapter sequence at the read ends, chimeric sequences, sequencing errors and artifacts, and alignment ambiguities due to the short read lengths and damage. For the Neandertal and Denisova genomes these challenges have been addressed largely using the approaches outlined in chapter 3. In combination with experimental approaches, the sequencing error associated with ancient DNA studies could be considerably reduced during the course of this project. For example, sequencing and ancient DNA damage errors remaining in the Denisova genome sequences are 4.6 times lower than for present-day human sequences analyzed with the same instrument and sequencing chemistry version.

The comparison of the human genome to the genomes of Neandertals and the Denisova individual allows the identification of features that set fully anatomically modern humans apart from other hominin forms. In particular, these ancient genome sequences provide a catalog of changes that have become fixed, or have risen to high frequency, in present-day humans and point to regions and genes affected by positive selection in the recent evolutionary history of modern humans. Further, these genomes allow the investigation of whether Neandertals or Denisovans contributed parts of their genomes to present-day humans by admixture.

Even though it is likely that most of the identified positions are functionally neutral, we believe that each of the rather small number of changes in functional regions of the human genome that have become fixed in humans since the divergence from the common ancestor with the Denisova and Neandertal individuals are of sufficient interest to warrant functional investigations. Once the Neandertal and Denisova genomes are sequenced to higher coverage, the number of these candidate positions will approximately double, e.g. the non-synonymous fixed derived amino acid changes that happened after the human lineage separated from Neandertals and Denisovans will increase to around 200 positions. Continuing studies of human genome diversity, for example the 1,000 Genome project [53], will further narrow down the number of putatively relevant sites and experimental work will be required to elucidate the physiological consequences of the remaining changes.

Analyzing the concordance of Neandertals and the Denisova individual with respect to genomic sites that have changed on the human lineage identified a considerable number of sites which show a derived state in the human reference sequence, as well as the derived state in either Denisova or Neandertal but not in both. These may, at least partially, reflect standing

IDT Exhibit 1061.0166

variation at the time of the separation of the modern human and the Neandertal/Denisova ancestors which segregated differently in these lineages.

It is of interest that there are more sites where Denisova has the ancestral allele than where Neandertal does at sites where they are discordant. This observation does not appear to be an artifact of differential error between the Denisova and Neandertal samples, since the patterns are consistent across substitution classes, however I could show a bias from the ancestry of the composite human reference sequence. Even when eliminating the reference bias by ascertainment of positions from two high divergence African genomes, an excess of 18.4% remained. This excess might either originate from alignment and sampling artifacts, or actually originate from admixture into the Denisova individual from some archaic hominin. Even though supported by the high divergence time of the Denisovan mitochondrial genome from present-day humans [121], this mitochondrial sequence may also be explained by genetic drift and a sufficiently large ancient population size [186]. Thus results are inconclusive at this point.

When analyzing the Neandertal-Denisova discordant sites in present-day populations, they were informative for detecting admixture with either of the ancient populations. I could confirm that an African individual (Yoruba HGDP00927) shares fewer ancestral alleles with Neandertal than do all non-African populations, supporting the signal described in Green et al. [81]. Further, I could show that Melanesians, especially the two Papuan individuals HGDP00542 and HGDP00551, show a signal of Denisovan admixture not shared with other sampled populations, a result in agreement with the D-statistics for population pairs presented in Reich et al. [186].

IDT Exhibit 1061.0167

IDT Exhibit 1061.0168

Chapter 7

Discussion and conclusions

In this thesis, I reviewed the sequencing concepts implemented by different high-throughput sequencing instruments and discussed their inherent limitations. I described how specific experimental steps in the generation of high-throughput sequencing data impact data quality and generate artifacts that may challenge data analysis. I introduced filters and algorithms which I have implemented at the Max Planck Institute for Evolutionary Anthropology to handle different sequencing artifacts and I have outlined approaches that can be used to reduce them in data generation. Further, I presented a machine learning approach for improving the base calling of the Illumina sequencing platform which I have implemented in the program `Ibis`. I explained how `Ibis` allows for a reduction of error rates and more informative base quality scores, independent of the actual Illumina sequencing instrument version and chemistry.

In the last two chapters, I discussed specific problems and selected results from the quantification of gene expression from short-sequence tags in five tissues from human, chimpanzee and rhesus macaque as well as the analysis of whole genome shotgun sequencing data of two ancient hominin genomes, the Neandertal and the Denisovan genome.

DNA sequencing and technologies

Only within the last five years, the development of alternative sequencing strategies like pyrosequencing, reversible terminator chemistry, sequencing-by-ligation, virtual terminator chemistry and single molecule real-time sequencing converged into new sequencing instruments. These new high-throughput sequencing instruments have a daily throughput which is a factor of 100 to 1,000 higher, and reduced the cost of sequencing one million nucleotides to 4%-0.1% of that of Sanger sequencing. In chapter 2, I have described the concepts of all currently available sequencing instruments. I have shown that each of these platforms has very specific biases and limitations, which go back to the technical details of how DNA molecules are prepared for sequencing, sequencing templates immobilized and finally how these are read out.

Current high-throughput technologies, and specifically the three most widely used: Roche 454 FLX, Life Technologies SOLiD and Illumina Genome Analyzer instruments, have an average error rate that is considerably higher than the typical 1/10,000 to 1/100,000 observed for high quality Sanger sequence reads. The sequencing error observed for Sanger sequencing is mainly due to errors in the amplification step, natural variance and contamination in the sample used as well as polymerase slippage at low complexity sequences like simple repeats and homopolymers.

IDT Exhibit 1061.0169

A large fraction of the errors observed for the Roche 454 technology are small insertions or deletions, mostly arising from inaccurate calling of homopolymer length, and single base-pair deletions or insertions caused by signal-to-noise thresholding issues. Error rate increases with the position in the sequence. This is due to reduction in reaction efficiency, molecule damage and phasing, a process whereby not all molecule copies are extended in every sequencing step. Phasing causes the molecules in the ensemble to lose synchrony/phase, and results in an echo of the preceding cycles as signal noise.

While the ensemble sequencing process for 454 pyrosequencing creates unidirectional phasing, reversible terminator sequencing as applied for the Illumina instruments creates bidirectional phasing, as incorporated nucleotides may fail effective termination – allowing immediate extension by another nucleotide. Further, the Genome Analyzer uses four fluorescent dyes to distinguish the four nucleotides of which two pairs (A/C and G/T) are more difficult to separate. Phasing, as described before, is mostly not an issue for the SOLiD platform, as sequences not extended are non-reversibly terminated. Since hybridization is a stochastic process and probes do not necessarily hybridize adjacent to the primer, this termination causes a considerable reduction in the number of molecules participating in subsequent ligation reactions, and therefore a substantial signal decline and error increase. Incomplete cleavage of dyes may allow cleavage after the next ligation, which then allows for the extension in the next but one cycle – a different kind of phasing.

Generally, the *in vitro* amplifications performed prior to sequencing for these three technologies, cause a higher error to be introduced into the sample before it enters the actual sequencing process. In addition, currently used random-dispersal protocols for immobilization of sequencing templates using beads or other solid surfaces cause mixed signal read outs and dependence of sequencing errors on the strength and distance from close-by sequencing reactions.

High quality Sanger sequencing is still commonly used to generate low coverage sequencing of individual positions, regions or very small genomes. Since Sanger sequence length (700-1,000bp) is longer than most abundant short repeat classes, it allows the unambiguous assembly of most genomic regions. However, Sanger technology is too expensive and too slow for sequencing a large number of samples, extended genomic regions or the many molecules required for quantitative applications. The 454 GS FLX Titanium provides a read length (400-500bp) still spanning many short repeat sequences – an important feature for accurate sequence mapping and assembly of genomes. Paired end or mate pair protocols help to overcome some of the limitations of short reads by providing information about relative location and orientation of a pair of reads.

Despite the insertion/deletion errors, the 454 technology has very low rates of misidentifying individual bases (1/1,000 - 1/10,000), making it suited for the identification of single nucleotide polymorphisms. Also geared to the identification of SNPs, at least for samples with an existing reference genome, is the SOLiD platform. Higher coverage is needed in order to perform SNP calling with similar accuracy using the Illumina Genome Analyzer. Neither the Illumina Genome Analyzer nor the SOLiD sequencing systems are prone to generating high rates of insertions or deletions, making them well suited for studying InDel variation.

Over the last years, the field of genetics experienced a shift from the amount of time required to prepare and run a sequencing experiment to the time required for the analysis of the generated data. It is likely that also in the future, laboratories will need to invest considerably more time, expertise and money in the design of experiments and the analysis of the vast quantities of data than in generating sequences. Smaller research groups may still not be able to afford the infrastructure needed for storing, handling and analyzing several tens of

IDT Exhibit 1061.0170

gigabytes of pure sequence data from these sequencing platforms. Even for larger groups and experienced genome centers this aspect remains an ever-increasing challenge.

New technologies like SMRT sequencing by Pacific Biosciences, Quantum-dot sequencing by Life Technologies or BASE by Oxford Nanopore will allow sequencing long individual molecules without or with little preparation steps and probably even the identification of specific nucleotide modifications. Improvements to current instruments are likely to further increase throughput and reduce cost of determining DNA sequences. Thus, the goal of a $1,000 human genome set by NIH/NHGRI for personalized medicine may soon be achieved. All these developments will hopefully facilitate future research in many fields and hopefully also simplify biological data analysis, to put these technologies into more hands.

Computational challenges from sequencing data production

The described advances in DNA sequencing already now bring a broader part of the scientific community in situations where a high-throughput project has to be designed or large sequence data sets have to be analyzed. The new technologies however come with some limitations and problems like for example considerable variance in run quality, specific biases and sensitivities, pseudo-sequences, high error rates as well as adapter and chimera sequences. These issues require to either design the project in a way to circumvent them or to consider them in data analysis before answering the actual biological question. In this regard, I analyzed the currently most frequently used high-throughput sequencing platform, the Illumina Genome Analyzer, in chapter 3.

The different Illumina sequencing instruments provide a high flexibility for creating functional sequencing libraries, as the only requirement, specific outer grafting sequences, can be added using various experimental approaches. This makes library design very flexible and allows application-specific protocols. Most preparation protocols require different DNA/RNA ligation, PCR amplification, purification and length selection steps. These can cause non-random sampling of the original molecules, which may limit the capacity to identify sequence variants or alter molecule frequency and quantification. Having different library preparation protocols requires that all of them are optimized for even sampling, a minimum of artifacts like adapter dimers, chimeras and contamination from other DNA/RNA sources as well as a sufficient insert size.

Illumina software does not handle artifact sequences nor does it filter or trim adapters. This is the reason why insert-adapter-chimeras or pure adapter dimers often end up in final data analysis. When adapter sequence starting at the read end is not identified for short-insert size libraries, this introduces a bias as reads are either excluded during mapping or reported with wrong alignment coordinates. When explicitly done for a sequencing run, the identification of starting adapter sequence and adapter chimeras is hampered by reads showing only a few bases of the adapter and by higher error rates at the end of a read. For paired end reads the correct identification of the adapter start position is eased by maximizing autocorrelation of the two reads with the outlined read merging process. Hence, for short insert libraries, paired end sequencing is to be preferred over single read sequencing. In addition to the efficient identification of adapters, merging reduces error rates in the consensus called sequence part. The presented algorithm has therefore been vital for ancient DNA studies at the Max Planck Institute.

In addition to creating a high-quality sequencing library and quantifying it, the correct adjustment of the machine, handling, as well as particles in the sequencing chemistry do have an impact on run quality. Reflections, uneven application of oil, air bubbles and a not

IDT Exhibit 1061.0171

perfectly-adjusted machine cause varying data quality and increased error rates. Particles like chemistry lumps, dust and lint can cause pseudo sequence signals which then result in the analysis of low sequence complexity reads which do not originate from the library sequenced. Tagging or indexing allows to filter for real library molecules and should be preferred over sequence-complexity-based methods. Even though filter based on sequence entropy provides high removal rates, it may introduce a bias due to the removal of real low complexity sequences. Further, filtering sequence tags or indexes is advantageous as it may also exclude contamination of sequencing lanes with molecules originating from other sequencing libraries and samples.

When an index/tag is placed in the beginning of the read, it can increase sequencing costs due to problems introduced in image data analysis and base calling. Correctly performed as separate reads, the error profile of the actual reads is not altered and multiplexing allows for the optimal usage of the increasing sequencing throughput, especially if subsets of a large genome or several small genomes are studied. When using an indexed spike-in control library in all lanes of a run, measures of run quality can be obtained and compared between individual lanes and whole runs. Further, these reads allow for quality score calibration and the application of alternative base callers. `PHRED`-like base quality scores should be used for quality-based filtering based on the complete read and specifically also on the index read outs of multiplex experiment. Quality-score-based filters remove clusters accumulating error due to their close proximity to another sequence cluster, but also remove reads affected by freely movable artifacts in specific sequencing cycles.

Improving base call quality of the Genome Analyzer platform

The Illumina Genome Analyzer is based on parallel, fluorescence-based readout of millions of immobilized sequences that are iteratively sequenced in a base-wise manner. After sequencing or while the sequencing run proceeds, images are analyzed and intensities extracted for each sequence cluster and all cycles of the run. The measured light intensities minus the surrounding background are extracted and resulting intensity values serve as input for base calling, the conversion of intensity values into bases. Base calling on the Illumina platform is complicated by at least two effects. First, a strong correlation of the adenine and cytosine intensities as well as of the guanine and thymine intensities due to similar emission spectra of fluorophores and a limited separation by optical filters. Second, a dependence of the signal for a specific cycle on the signal of the cycles before and after, known as phasing and pre-phasing, respectively. The fraction of molecules in a cluster affected by this loss of synchrony in the readout of the sequence copies increases with the number of cycles, hampering correct base identification.

Already during the first year after the release of the Illumina platform, statistical learners trained on sequences called by the standard base caller were suggested for improving the base calling of the platform. In chapter 4, I presented `Ibis` (`Improved base identification system`), a more accurate, fast and easy-to-use base caller for all Illumina sequencing instruments.

Previous approaches corrected raw intensities prior to the application of statistical learners and used only the intensities of one cycle as input. This causes these approaches, and other model-based approaches, to depend on a good understanding and modeling of the sequencing process for obtaining corrected intensities. `Ibis` by-passes this problem by direct training of one statistical model per sequencing cycle based on raw cluster intensities of multiple input cycles, directly incorporating the effects of phasing and pre-phasing. Even though

IDT Exhibit 1061.0172

this approach may not provide interpretable model parameters, `Ibis` implements the most general and flexible approach. This is of advantage when considering the vast improvements of sequencing chemistry and instrument over the last years.

`Ibis` was originally developed to handle the T accumulation, in a sequencing chemistry which has been replaced by several subsequent versions, and still its application is not limited to the reprocessing of data created with the older chemistries. I have shown that `Ibis` improves the output of sequencing runs from the Genome Analyzer I, which due to their short read length are barely affected by T accumulation, but by a generally lower image and sequencing quality. Further, I have shown that it improves base calling accuracy for runs using recent sequencing chemistries without T accumulation and increased sequencing length. The reason is the sequencing model independent training process of `Ibis`, which only relies on the assumption that the vast majority of the signal needed for base calling is captured by the intensity values of the previous, the current and the next cycle.

The presented approach is unique and currently applies to the full range of different sequencing chemistries and platform versions, where it reduces sequencing error by at least 10-20%. The performance of `Ibis` on standard hardware is significantly better than for other existing alternative base callers, enabling it to be run by research laboratories without access to large computational clusters. The increase in mappable sequences as well as improved and calibrated `PHRED`-like quality scores enable the direct use of the sequences in other software packages. Thus, there is a considerable benefit in investing the computational time in `Ibis` re-base-calling for sequencing runs.

Quantification of gene expression from short-sequence tags

Comparative genomics studies of humans, hominins and other apes are important in order to better understand human origins and the biological background of what sets humans apart from other primates. They may also provide insights for the basis of diseases or developmental problems that affect uniquely human traits, such as speech disorders, mental disorders like autism and schizophrenia or metabolic disorders like obesity. The first studies of gene expression differences using microarrays allowed the characterization of human-chimpanzee gene expression at a genome-wide scale, which was not possible before. The analysis of five tissue transcriptomes showed that while gene expression may mostly evolve neutrally in the two species, positive selective forces may shape brain and testis transcriptomes.

The technological advances in sequencing allowed to complement these original studies with new insights from direct sequencing of transcriptomes. While microarrays are static in their design and limited to the analysis of genomic features known at the time of design, sequence-based gene expression profiling represents a more direct approach for identifying the expressed molecules. Further, hybridization-based technologies are rather sensitive to polymorphisms in the transcribed region that is probed – a problem that is amplified in comparative studies in which expression patterns of species with some sequence differences are inferred and compared. The Illumina *NlaIII* Digital Gene Expression (DGE) approach was used to study brain, heart, kidney, liver and testis tissues of humans, chimpanzees and rhesus macaques. This *NlaIII* DGE protocol was one of the first applications of the Illumina sequencing platform and is based on the Serial Analysis of Gene Expression (SAGE) which infers gene expression levels through short (here 17nt) tag sequencing from the 3' end of transcripts.

The tag data was generated in tissue batches over a period of one and a half years. Sequencing reads were processed following the principles outlined in chapter 3. Reads were

IDT Exhibit 1061.0173

filtered for dimer/chimera sequences and for the presence of adapter sequence at the read end. Further, reads were filtered based on `PHRED` quality scores and sequence entropy. Low complexity sequences were of increased frequency for older Genome Analyzer runs, due to a reflection layer surrounding the actual lanes of version 1 flow cells. For the kidney and heart batches, samples showed up to 40% of adapter dimers/chimeras. Data quality also showed considerable variation between batches, supporting the need of the applied quality filters.

Even though the Illumina Digital Gene Expression protocol could overcome limitations of a static microarray design, I could show that specific features of this type of data also complicate analysis. Incomplete digestion and enzyme hindrance cause the dispersion of transcript counts across multiple restriction sites. Unspecific carry-over of upstream *NlaIII* digestion fragments between experimental steps causes a false signal of antisense transcription and may also cause additional, i.e. non-3'-most, tag counts on the sense strand. Further, the free annotation of genomic features, which was considered an advantage of a sequencing approach, turned out to remain a challenge. Annotation of tags was problematic due to ambiguities from their short length and due to very different annotation quality for the different species. Only very recent human gene annotation provided the necessary 3' UTR annotation and could be projected to the chimpanzee and rhesus, losing about 36% of genes annotated in human but giving similar proportions of tag counts within genes for all three species.

The biggest challenge, however, was that many tags are not unique to specific genomic sites or genes, and that the uniqueness of tags differs slightly between the three species. This causes the tags of two and more different genes to be collapsed into one measurement. Different approaches of counting tags multiple times or removing ambiguous tags do all result in wrong gene quantifications, i.e. a false ranking of gene expression values. The effect seems not very strong as rank correlations around 0.9 are observed between two extreme approaches. Further, the impact on the inter-species analysis is expected to be low, as this ranking problem will be largely the same in all three species. Variation between species is only expected from evolution and loss of *NlaIII* sites in each genome, with rates expected to be close to average species sequence divergence.

From comparisons to other studies of the same species and tissues, larger disagreement was observed than expected. Differences in the symmetry of assignment of changes to lineages or the percentage differentially expressed genes were observed. The extend of disagreement could not easily be explained with the false discovery rates of the applied statistical tests. It is likely that all methods have technological (experimental and analysis) biases, and that the obtained variance estimates are too low. Specifically, the comparison with the Babbitt et al. study (which uses the same *NlaIII* DGE protocol for different brain samples), clearly showed that sampling variation is larger than variance from biological differences between human and chimpanzee and that analysis variation may even be as strong as differences between human/chimpanzee and rhesus macaque.

Comparing the previous microarray and the new DGE five tissue study, the strongest and most consistent pattern in the data sets are the about 30% differentially expressed genes between human and chimpanzee testis. Other studies have however suggested that human testes may contain around 20% less germ cells per volume than chimpanzee testes. If this proves to be correct, the differential expression effect may be largely driven from changes in tissue composition. Even though it is completely valid that expression differences originate from variation in cell-type composition, expression differences from changes in cell-type composition conflict with the goal of understanding how DNA sequence evolution impacts gene expression in fully developed tissue.

IDT Exhibit 1061.0174

The presented result of human-chimpanzee expression differences being smaller than the variation from sampling and experimental protocols for at least one tissue, challenges current findings from these inter-species studies. For future studies it will be of interest to minimize all sources of variation. To achieve this, it will be necessary to stringently control sample environmental effects and sample age, increase the number of samples, study specific cell-types rather than tissues and to use improved experimental and analysis protocols. The analyses presented clearly showed that inter-species studies are very sensitive to small differences in data processing. Such differences may easily originate from different genome quality, genome completeness and genome annotation quality. Measures have to be established to check for such effects in the analysis.

Analysis of two hominin genomes from ancient DNA

Since ancient DNA sequences are generally short in length, damaged, and at low copy-number relative to co-extracted environmental DNA, high-throughput sequencing approaches offer a tremendous advantage over traditional Sanger sequencing in that they enable a complete characterization of an ancient DNA extract. As the cost of sequencing continues to decrease, it has become feasible to analyze entire genomes of ancient samples, including those for which the endogenous DNA makes up only a very small percentage of the total extracted DNA. The Max Planck Institute for Evolutionary Anthropology has used high-throughput sequencing for whole genome shotgun sequencing of two extinct hominin genomes, the Neandertal and Denisova genome. Chapter 6 described the processing of the ancient DNA as well as present-day human Illumina sequencing data which I have performed for both projects.

The high-throughout data obtained from ancient DNA libraries includes reads with adapter sequence at the ends, chimerical sequences and other artifacts, sequencing error as well as alignment ambiguities due to the short read lengths and DNA damage. For Neandertal and Denisova these different challenges have been addressed using the approaches outlined in chapter 3 and 4 (e.g. improved base calling, tag filtering, and read merging). In combination with experimental approaches, the sequencing error associated with ancient DNA studies could be considerably reduced. For example, the remaining error from sequencing and ancient DNA damage in the Denisova molecule read outs are 4.6x lower than for present-day human sequences analyzed with the same instrument and sequencing chemistry version.

I exemplified how the sequence data can be used to study sites in the human genome which have changed since the last common ancestor of human, chimpanzee and bonobo. The comparison of the human genome to the genomes of Neandertals and the Denisova individual allows to identify features that set fully anatomically modern humans apart from other hominin forms. In particular, I generated a catalog of changes that have become fixed or have risen to high frequency in present-day humans since the divergence from these other human forms. The identified positions point to several regions and genes, some of which might be affected by positive selection in the recent evolutionary history of modern humans. Once the Neandertal and Denisova genomes are sequenced to higher coverage, the number of these candidate positions will approximately double as currently some regions are not sufficiently covered to determine the exact state of these two ancient hominins for all positions which changed on the human lineage. Continuing studies of human genome diversity like for example the 1,000 Genome project, will help to reduce the number of putatively relevant sites. However, most importantly, experimental work will be required to elucidate the physiological consequences of the identified changes.

In addition, I described an interesting subset of sites which changed on the human lineage.

IDT Exhibit 1061.0175

These sites were identified from analyzing the concordance of Neandertals and the Denisova individual. This subset is seen in the ancestral state in Neandertals or the Denisova individual but not in both. Considering that Neandertals and Denisovans form a common clade, these positions at least partially represent variation present at the time when the lineages of modern humans, Neandertals and Denisovans separated. Since positions where differently fixed in these three lineages, allele sharing patterns were generated that do not follow the genome-wide phylogeny of the three species. Hence, these sites can be described as incomplete lineage sorting. Thus, the ancestral states might have been reintroduced into some present-day human populations by admixture with either Neandertals or Denisovans. Therefore, these sites can be used to test present-day human individuals whether they show more frequently the ancestral allele for the Denisova ancestral sites or the ancestral allele for Neandertal ancestral sites.

I pointed out that an excess in the number of Denisova ancestral sites exists in the catalog of changes, even when removing a bias from the human reference genome due to Neandertal admixture in its non-African parts. This excess might originate from admixture into the Denisova individual from some other archaic hominin, however biases from alignment can not be ruled out and results are inconclusive at this point. When analyzing the Neandertal-Denisova discordant sites in twelve present-day populations, they turned out to be informative for detecting admixture with either of the ancient population. I could confirm that an African individual shares fewer ancestral alleles with Neandertal than do all non-African individuals, supporting the admixture signal with non-Africans described in Green et al. for the Neandertal genome. Further, I could show that Melanesians, especially the two Papuan individuals, show a signal of Denisovan admixture not shared with other sampled populations, a result in agreement with the D-statistics for population pairs presented in Reich et al. for the Denisova genome.

Outlook

The field of high-throughput sequencing data analysis emerged in 2005, when the 454 Genome Sequencer became available. It is obvious that such a young field has a lot of potential for further improvements and even major changes. We will most likely see several improvements to instruments and library preparation protocols over the next few years. These will hopefully reduce amplification, ligation and purification biases, or at least lower the DNA/RNA input requirements or the number of experimental steps, to reduce their effects. In this respect more transparency from vendors will be required as thorough studies of biases should not need to be done independently by all groups applying these protocols.

Further, more experimental controls will be developed. What is already standard at the Max Planck Institute for Evolutionary Anthropology, the sequencing of a spike-in control library with every sample, will likely be more widely applied. Additional controls for library preparation and specific steps of sequencing preparation will become available and make library quality a more transparent measure. The goal will be to better understand biases and to actually measure and detect them. Without any measure of the actual biases, one might be lucky to detect them (as the case for the presented studies) but may not be able to correct the data appropriately.

These developments will hopefully contribute to a better planning of experiments, including the controls for small effects and the comparability of conditions. Once data is generated, it is frequently very difficult to identify these problems and too expensive to repeat part of the experiment. Specifically in the context of comparative studies, downstream analysis

IDT Exhibit 1061.0176

needs to be outlined before data generation and sources of species bias tested. As long as genome quality, completeness and annotation vary between species, there will always be this danger. If these problems are resolved, a concern remains about different levels of species diversity, and how one reference sequence can represent this diversity. In the analysis of Neandertal and Denisova, the population history of some genomic regions of the human reference influenced the results. Either a population history-free genome, i.e. the "average human genome", has to be generated or all types of analyses have to be checked that they do not depend on such small signals. Another approach might be the idea of genome graphs incorporating all known sequence variation and appropriate algorithms to work on these.

The outlined adapter and chimera handling needs to be integrated into sequencing analysis pipelines and possibly also the vendor pipelines. Alternative base callers and quality score calibration from control reads should be a standard step in sequencing analysis pipelines. Further, improved aligners and analysis software is required which fully uses quality scores and provides sensitive measures on mapping accuracy and mapping likelihood. More generally, likelihood and error propagation in data analysis needs be to strengthened, otherwise reported significance measures will not be correct.

For `Ibis`, future developments may include the reimplementation using free SVM packages, instead of a package which is only free for non-commercial use. This should then allow the integration into the standard Illumina analysis pipeline. Using one-against-all multi-class approaches may also allow more direct measures of base quality scores. It might be of interest to report up to four base quality scores instead of one base. Currently, however, effectively no downstream software exists that would handle such data.

Quantification of gene expression by tag profiling or microarrays are likely to be replaced by RNAseq approaches. RNAseq is the equivalent of a genome shot-gun sequencing approach for the transcriptome and determines RNA fragments from the full length transcripts. It is therefore less affected by sample and library preparation biases (e.g. molecule GC biases from PCR amplification and gel excision) or analysis biases due to ambiguous alignments. Quantification can use the full length of the transcript and correct for regions of increased or reduced read coverage, as is already implemented in recent RNAseq quantification tools. In addition, RNAseq provides both quantification and information on different splice forms as well as variation in transcription start and polyadenylation sites. This information may enable *de novo* transcriptome annotation, e.g. for poorly annotated primate genomes. Differences in genome quality and completeness between different primate genomes will however still complicate comparative analyses. While some studies currently aim for low-coverage sequencing of many different individuals and species, improvements to the genome quality of non-reference and model species will require bigger investments of money and time.

Improvements in sequencing technologies will most likely include considerable increases in read length. While this development is of advantage for many application using recent DNA and RNA sources, the field of ancient DNA will probably not profit much from these improvements due to the generally short ancient molecule lengths. Here approaches with high parallelization and sensitivity to base modifications will be of interest. If DNA damage could be directly identified during read out, the original base can be inferred during data analysis and experimental steps prior to sequencing reduced. While currently the treatment of ancient DNA for deaminated cytosines provides great reduction in sequence error, it also reduces molecule length and causes some molecules to be lost in the additional experimental steps. More efficient extraction and preparation methods in combination with highly parallelized sequencing could permit access to the small proportion of longer molecules and hopefully allow for high coverage and high quality genomes from ancient DNA samples.

IDT Exhibit 1061.0177

IDT Exhibit 1061.0178

Bibliography

[1] M. Abu and T.R. Waters. The main role of human thymine-DNA glycosylase is removal of thymine produced by deamination of 5-methylcytosine and not removal of ethenocytosine. *Journal of Biological Chemistry*, 278(10):8739, 2003.

[2] F. Albertorio, M. E. Hughes, J. A. Golovchenko, and D. Branton. Base dependent DNA-carbon nanotube interactions: activation enthalpies and assembly-disassembly control. *Nanotechnology*, 20(39):395101, 2009.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell. 4th edition.* Garland Science, New York, USA, 2002.

[4] Gene Amdahl. Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities. *AFIPS Conference Proceedings*, 30:483–485, 1967.

[5] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[6] M.J. Anderson, S.J. Chapman, E.N. Videan, E. Evans, J. Fritz, T.S. Stoinski, A.F. Dixson, and P. Gagneux. Functional evidence for differences in sperm competition in humans and chimpanzees. *American journal of physical anthropology*, 134(2):274–280, 2007.

[7] S. Andrews. FastQC: A quality control tool for high throughput sequence data. `http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/`, 2010.

[8] W. J. Ansorge. Next-generation DNA sequencing techniques. *Nature Biotechnology*, 25(4):195–203, 2009.

[9] Y. Astier, O. Braha, and H. Bayley. Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5'-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. *J Am Chem Soc*, 128(5):1705–10, 2006.

[10] C.C. Babbitt, O. Fedrigo, A.D. Pfefferle, A.P. Boyle, J.E. Horvath, T.S. Furey, and G.A. Wray. Both noncoding and protein-coding RNAs contribute to gene expression evolution in the primate brain. *Genome Biology and Evolution*, 2(0):67, 2010.

[11] B. Ballester, N. Johnson, G. Proctor, and P. Flicek. Consistent annotation of gene expression arrays. *BMC genomics*, 11(1):294, 2010.

[12] J. Batley and D. Edwards. Genome sequence data: management, storage, and visualization. *Biotechniques*, 46(5):333–4, 336, 2009.

[13] N. Beifang, F. Limin, S. Shulei, and L. Weizhong. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics*, 11.

IDT Exhibit 1061.0179

[14] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[15] D. Benovoy, T. Kwan, and J. Majewski. Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments. *Nucleic Acids Research*, 36(13):4417, 2008.

[16] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9, 2008.

[17] F.B. Bercovitch, A. Widdig, A. Trefilov, M.J. Kessler, J.D. Berard, J. Schmidtke, P. Nürnberg, and M. Krawczak. A longitudinal study of age-specific reproductive output and body condition among male rhesus macaques, Macaca mulatta. *Naturwissenschaften*, 90(7):309–312, 2003.

[18] C.A. Bill, W.A. Duran, N.R. Miselis, and J.A. Nickoloff. Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese hamster ovary cells: competition between long-patch and GT glycosylase-mediated repair of GT mismatches. *Genetics*, 149(4):1935, 1998.

[19] Applied Biosystems. A Theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction, 2008.

[20] C.P. Bird, B.E. Stranger, M. Liu, D.J. Thomas, C.E. Ingle, C. Beazley, W. Miller, M.E. Hurles, and E.T. Dermitzakis. Fast-evolving noncoding sequences in the human genome. *Genome Biology*, 8(6):R118, 2007.

[21] I. Birol, S.D. Jackman, C.B. Nielsen, J.Q. Qian, R. Varhol, G. Stazyk, R.D. Morin, Y. Zhao, M. Hirst, J.E. Schein, et al. De novo transcriptome assembly with ABySS. *Bioinformatics*, 25(21):2872, 2009.

[22] R. G. Blazej, P. Kumaresan, and R. A. Mathies. Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proceedings of the National Academy of Sciences USA*, 103(19):7240–5, 2006.

[23] R. Blekhman, J.C. Marioni, P. Zumbo, M. Stephens, and Y. Gilad. Sex-specific and lineage-specific alternative splicing in primates. *Genome Research*, 20(2):180, 2010.

[24] B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185, 2003.

[25] J. Bowers, J. Mitchell, E. Beer, P. R. Buzby, M. Causey, J. W. Efcavitch, M. Jarosz, E. Krzymanska-Olejnik, L. Kung, D. Lipson, G. M. Lowman, S. Marappan, P. McInerney, A. Platt, A. Roy, S. M. Siddiqi, K. Steinmann, and J. F. Thompson. Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods*, 6(8):593–5, 2009.

[26] A. W. Briggs, J. M. Good, R. E. Green, J. Krause, T. Maricic, U. Stenzel, C. Lalueza-Fox, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, R. Schmitz, V. B. Doronichev, L. V. Golovanova, M. de la Rasilla, J. Fortea, A. Rosas, and S. Pääbo. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*, 325(5938):318–21, 2009.

IDT Exhibit 1061.0180

[27] A. W. Briggs, U. Stenzel, M. Meyer, J. Krause, M. Kircher, and S. Pääbo. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Research*, 2009.

[28] A.W. Briggs, U. Stenzel, P.L.F. Johnson, R.E. Green, J. Kelso, K. Prüfer, M. Meyer, J. Krause, M.T. Ronan, M. Lachmann, et al. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences USA*, 104(37):14616, 2007.

[29] T.C. Brown and J. Jiricny. A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell*, 50(6):945–950, 1987.

[30] H. A. Burbano, E. Hodges, R. E. Green, A. W. Briggs, J. Krause, M. Meyer, J. M. Good, T. Maricic, P. L. Johnson, Z. Xuan, M. Rooks, A. Bhattacharjee, L. Brizuela, F. W. Albert, M. de la Rasilla, J. Fortea, A. Rosas, M. Lachmann, G. J. Hannon, and S. Pääbo. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science*, 328(5979):723–5, 2010.

[31] E.C. Bush and B.T. Lahn. A genome-wide screen for noncoding elements important in primate evolution. *BMC Evolutionary Biology*, 8(1):17, 2008.

[32] F. Caiment, C. Charlier, T. Hadfield, N. Cockett, M. Georges, and D. Baurain. Assessing the effect of the CLPG mutation on the microRNA catalog of skeletal muscle using high-throughput sequencing. *Genome Research*, 20(12):1651, 2010.

[33] M.J. Chaisson, D. Brinza, and P.A. Pevzner. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Research*, 19(2):336, 2009.

[34] W. Chen, R. Ullmann, C. Langnick, C. Menzel, Z. Wotschofsky, H. Hu, A. Doring, Y. Hu, H. Kang, A. Tzschach, M. Hoeltzenbein, H. Neitzel, S. Markus, E. Wiedersberg, G. Kistner, C. M. van Ravenswaaij-Arts, T. Kleefstra, V. M. Kalscheuer, and H. H. Ropers. Breakpoint analysis of balanced chromosome rearrangements by next-generation paired-end sequencing. *European Journal of Human Genetics*, 2009.

[35] Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69, 2005.

[36] A.G. Clark, S. Glanowski, R. Nielsen, P.D. Thomas, A. Kejariwal, M.A. Todd, D.M. Tanenbaum, D. Civello, F. Lu, B. Murphy, et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, 302(5652):1960, 2003.

[37] J. Clarke, H. C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnolgy*, 4(4):265–70, 2009.

[38] S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452(7184):215–9, 2008.

[39] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[40] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2(2):265–292, 2002.

IDT Exhibit 1061.0181

[41] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, 2002.

[42] C. J. Creighton, J. G. Reid, and P. H. Gunaratne. Expression profiling of micrornas by deep sequencing. *Briefings in Bioinformatics*, 10(5):490–7, 2009.

[43] V. Curwen, E. Eyras, T.D. Andrews, L. Clarke, E. Mongin, S.M.J. Searle, and M. Clamp. The Ensembl automatic gene annotation system. *Genome Research*, 14(5):942, 2004.

[44] A. V. Dalca and M. Brudno. Genome variation discovery with high-throughput sequencing data. *Briefings in Bioinformatics*, 11(1):3–14, 2010.

[45] M. Dannemann, A. Lorenc, I. Hellmann, P. Khaitovich, and M. Lachmann. The effects of probe binding affinity differences on gene expression measurements and how to deal with them. *Bioinformatics*, 25(21):2772, 2009.

[46] M. M. DeAngelis, D. G. Wang, and T. L. Hawkins. Solid-phase reversible immobilization for the isolation of pcr products. *Nucleic Acids Research*, 23(22):4742–3, 1995.

[47] S. Diguistini, N. Y. Liao, D. Platt, G. Robertson, M. Seidel, S. K. Chan, T. R. Docking, I. Birol, R. A. Holt, M. Hirst, E. Mardis, M. A. Marra, R. C. Hamelin, J. Bohlmann, C. Breuil, and S. J. Jones. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology*, 10(9):R94, 2009.

[48] Eileen T Dimalanta, Lei Zhang, Cynthia L Hendrickson, Tanya D Sokolsky, Adam E Sannicandro, Jonathan M Manning, Stephen F McLaughlin, Haoning Fu, Clarence C Lee, Alan P Blanchard, Gina L Costa, and Kevin J McKernan. Increased Read Length on the SOLiD Sequencing Platform, 2009.

[49] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105, 2008.

[50] P.C. Dolan and D.R. Denver. TileQC: a system for tile-based quality control of Solexa data. *BMC Bioinformatics*, 9(1):250, 2008.

[51] T.R. Dreszer, G.D. Wall, D. Haussler, and K.S. Pollard. Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome research*, 17(10):1420, 2007.

[52] R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, et al. *Science*, 327(5961):78–81, 2010.

[53] R.M. Durbin, D.L. Altshuler, G.R. Abecasis, D.R. Bentley, A. Chakravarti, A.G. Clark, F.S. Collins, F.M. De La Vega, P. Donnelly, M. Egholm, et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[54] L. Duret and N. Galtier. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, 10:285–311, 2009.

[55] L. Duret and N. Galtier. Comment on Human-specific gain of function in a developmental enhancer. *Science*, 323(5915):714c, 2009.

IDT Exhibit 1061.0182

[56] R.C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460, 2010.

[57] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–8, 2009.

[58] C. A. Emrich, H. Tian, I. L. Medintz, and R. A. Mathies. Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. *Analytical Chemistry*, 74(19):5076–83, 2002.

[59] W. Enard, P. Khaitovich, J. Klose, S. Zollner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, et al. Intra-and interspecific variation in primate gene expression patterns. *Science*, 296(5566):340, 2002.

[60] Y. Erlich, K. Chang, A. Gordon, R. Ronen, O. Navon, M. Rooks, and G. J. Hannon. DNA Sudoku–harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Research*, 19(7):1243–53, 2009.

[61] Y. Erlich, P. P. Mitra, M. delaBastide, W. R. McCombie, and G. J. Hannon. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nature Methods*, 5(8):679–82, 2008.

[62] W.J. Ewens. *Mathematical population genetics: theoretical introduction.* Springer Verlag, 2004.

[63] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8(3):186–94, 1998.

[64] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, 8(3):175–85, 1998.

[65] A.M. Faisst, G. Alvarez-Bolado, D. Treichel, and P. Gruss. Rotatin is a novel gene required for axial rotation and left-right specification in mouse embryos. *Mechanisms of development*, 113(1):15–28, 2002.

[66] M. Fedurco, A. Romieu, S. Williams, I. Lawrence, and G. Turcatti. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research*, 34(3):e22, 2006.

[67] J.N. Fenner. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American journal of physical anthropology*, 128(2):415–423, 2005.

[68] A. Fischer, J. Pollack, O. Thalmann, B. Nickel, and S. Pääbo. Demographic history and genetic differentiation in apes. *Current Biology*, 16(11):1133–1138, 2006.

[69] P. Flicek and E. Birney. Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, 6(11 Suppl):S6–S12, 2009.

[70] B.A. Flusberg, D.R. Webster, J.H. Lee, K.J. Travers, E.C. Olivares, T.A. Clark, J. Korlach, and S.W. Turner. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods*, 7(6):461–465, 2010.

IDT Exhibit 1061.0183

[71] S. Fox, S. Filichkin, and T.C. Mockler. Applications of ultra-high-throughput sequencing. *Methods in Molecular Biology*, 553:79–108, 2009.

[72] N. Galtier and L. Duret. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics*, 23(6):273–277, 2007.

[73] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

[74] K. S. George, X. Zhao, D. Gallahan, A. Shirkey, A. Zareh, and B. Esmaeli-Azad. Capillary electrophoresis methodology for identification of cancer related gene expression patterns of fluorescent differential display polymerase chain reaction. *Journal of Chromatography B: Biomedical Sciences and Applications*, 695(1):93–102, 1997.

[75] R.A. Gibbs, J. Rogers, M.G. Katze, R. Bumgarner, G.M. Weinstock, E.R. Mardis, K.A. Remington, R.L. Strausberg, J.C. Venter, R.K. Wilson, et al. Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, 316(5822), 2007.

[76] W. Gilbert and A. Maxam. The nucleotide sequence of the lac operator. *Proceedings of the National Academy of Sciences USA*, 70(12):3581–4, 1973.

[77] A. Gnirke, A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander, and C. Nusbaum. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, 27(2):182–9, 2009.

[78] JM Good, T. Giger, MD Dean, MW Nachman, and G.S. Barsh. Widespread Over-Expression of the X Chromosome in Sterile F1 Hybrid Mice. *PLoS Genetics*, 6(9), 2010.

[79] M. Gralle and S. Pääbo. A Comprehensive Functional Analysis of Ancestral Human Signal Peptides. *Molecular Biology and Evolution*, 2010.

[80] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185:862–864, 1974.

[81] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Fritz, N. F. Hansen, E. Y. Durand, A. S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, et al. A draft sequence of the Neandertal genome. *Science*, 328(5979):710–22, 2010.

[82] R. E. Green, J. Krause, S. E. Ptak, A. W. Briggs, M. T. Ronan, J. F. Simons, L. Du, M. Egholm, J. M. Rothberg, M. Paunovic, and S. Pääbo. Analysis of one million base pairs of Neanderthal DNA. *Nature*, 444(7117):330–6, 2006.

[83] R. E. Green, A. S. Malaspinas, J. Krause, A. W. Briggs, P. L. Johnson, C. Uhler, M. Meyer, J. M. Good, T. Maricic, U. Stenzel, K. Prüfer, M. Siebauer, H. A. Burbano, M. Ronan, J. M. Rothberg, M. Egholm, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, M. Wikstrom, L. Laakkonen, J. Kelso, M. Slatkin, and S. Pääbo. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 134(3):416–26, 2008.

[84] R.E. Green, A.W. Briggs, J. Krause, K. Prüfer, H.A. Burbano, M. Siebauer, M. Lachmann, and S. Pääbo. The Neandertal genome and ancient DNA authenticity. *The EMBO Journal*, 28(17):2494–2502, 2009.

IDT Exhibit 1061.0184

[85] S. Griffiths-Jones, R.J. Grocock, S. Van Dongen, A. Bateman, and A.J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(suppl 1):D140, 2006.

[86] A.R. Gruber, R. Lorenz, S.H. Bernhart, R. Neubock, and I.L. Hofacker. The Vienna RNA websuite. *Nucleic Acids Research*, 36(Web Server issue):W70, 2008.

[87] K.D. Hansen, S.E. Brenner, and S. Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12):e131, 2010.

[88] O. Harismendy, P. C. Ng, R. L. Strausberg, X. Wang, T. B. Stockwell, K. Y. Beeson, N. J. Schork, S. S. Murray, E. J. Topol, S. Levy, and K. A. Frazer. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, 10(3):R32, 2009.

[89] T. D. Harris, P. R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. Dimeo, J. W. Efcavitch, E. Giladi, J. Gill, J. Healy, M. Jarosz, D. Lapen, K. Moulton, S. R. Quake, K. Steinmann, E. Thayer, A. Tyurina, R. Ward, H. Weiss, and Z. Xie. Single-molecule DNA sequencing of a viral genome. *Science*, 320(5872):106–9, 2008.

[90] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[91] D. Haussler, S.J. O'Brien, O.A. Ryder, F.K. Barker, M. Clamp, A.J. Crawford, R. Hanner, O. Hanotte, W.E. Johnson, J.A. McGuire, et al. Genome 10K: A proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, 100(6):659–674, 2009.

[92] D. G. Hert, C. P. Fredlake, and A. E. Barron. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis*, 29(23):4618–26, 2008.

[93] P. Heyn, U. Stenzel, A.W. Briggs, M. Kircher, M. Hofreiter, and M. Meyer. Road blocks on paleogenomes–polymerase extension profiling reveals the frequency of blocking lesions in ancient DNA. *Nucleic Acids Research*, 2010.

[94] E. Hodges, M. Rooks, Z. Xuan, A. Bhattacharjee, D. Benjamin Gordon, L. Brizuela, W. Richard McCombie, and G. J. Hannon. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protocols*, 4(6):960–74, 2009.

[95] S. Hoffmann, C. Otto, S. Kurtz, C.M. Sharma, P. Khaitovich, J. Vogel, P.F. Stadler, and J. Hackermuller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Computational Biology*, 5(9):e1000502, 2009.

[96] Vyvayn Holland. *The Complete Works of Oscar Wilde*. Collins, London, UK, 1992.

[97] X. C. Huang, M. A. Quesada, and R. A. Mathies. DNA sequencing using capillary array electrophoresis. *Analytical Chemistry*, 64(18):2149–54, 1992.

[98] M. Huber, G. Siegenthaler, N. Mirancea, I. Marenholz, D. Nizetic, D. Breitkreutz, D. Mischke, and D. Hohl. Isolation and characterization of human repetin, a member of the fused gene family of the epidermal differentiation complex. *Journal of Investigative Dermatology*, 124(5):998–1007, 2005.

IDT Exhibit 1061.0185

[99] JW Ijdo, A. Baldini, DC Ward, ST Reeders, and RA Wells. Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proceedings of the National Academy of Sciences USA*, 88(20):9051, 1991.

[100] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[101] R.R. Iyer, A. Pluciennik, V. Burdett, and P.L. Modrich. DNA mismatch repair: functions and mechanisms. *Chem. Rev*, 106(2):302–323, 2006.

[102] W.R. Jeck, J.A. Reinhardt, D.A. Baltrus, M.T. Hickenbotham, V. Magrini, E.R. Mardis, J.L. Dangl, and C.D. Jones. Extending assembly of short DNA sequences to handle error. *Bioinformatics*, 23(21):2942, 2007.

[103] Z. Jing, Y. Liu, M. Dong, S. Hu, S. Huang, et al. Identification of the DNA binding element of the human ZNF333 protein. *Journal of biochemistry and molecular biology*, 37(6):663–670, 2004.

[104] B. John, A.J. Enright, A. Aravin, T. Tuschl, C. Sander, D.S. Marks, et al. Human microRNA targets. *PLoS Biology*, 2(11):e363, 2004.

[105] H. Kambara and S. Takahashi. Multiple-sheathflow capillary array DNA analyser. *Nature*, 361(6412):565–6, 1993.

[106] W. C. Kao, K. Stevens, and Y. S. Song. BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. *Genome Research*, 2009.

[107] W.C. Kao and Y. Song. naiveBayesCall: An efficient model-based base-calling algorithm for high-throughput sequencing. In *Research in Computational Molecular Biology*, pages 233–247. Springer, 2010.

[108] S. Katayama, Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, M. Nakamura, H. Nishida, CC Yap, M. Suzuki, J. Kawai, et al. Antisense transcription in the mammalian transcriptome. *Science*, 309(5740):1564–1566, 2005.

[109] W.J. Kent. BLAT – the BLAST-like alignment tool. *Genome research*, 12(4):656, 2002.

[110] P. Khaitovich, W. Enard, M. Lachmann, and S. Pääbo. Evolution of primate gene expression. *Nature Reviews Genetics*, 7(9):693–702, 2006.

[111] P. Khaitovich, I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Pääbo. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, 309(5742):1850, 2005.

[112] P. Khaitovich, K. Tang, H. Franz, J. Kelso, I. Hellmann, W. Enard, M. Lachmann, and S. Pääbo. Positive selection on gene expression in the human brain. *Current Biology*, 16(10):356–358, 2006.

[113] J. I. Kim, Y. S. Ju, H. Park, S. Kim, S. Lee, J. H. Yi, J. Mudge, N. A. Miller, D. Hong, C. J. Bell, H. S. Kim, I. S. Chung, W. C. Lee, J. S. Lee, S. H. Seo, J. Y. Yun, H. N. Woo, et al. A highly annotated whole-genome sequence of a Korean individual. *Nature*, 460(7258):1011–5, 2009.

[114] S. Kim, H. J. Yoo, and J. H. Hahn. Postelectrophoresis capillary scanning method for DNA sequencing. *Analytical Chemistry*, 68(5):936–9, 1996.

IDT Exhibit 1061.0186

[115] M. Kircher and J. Kelso. High-throughput DNA sequencing-concepts and limitations. *BioEssays*, 32(6):524–536, 2010.

[116] M. Kircher, U. Stenzel, and J. Kelso. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biology*, 10(8):R83, 2009.

[117] G. Koller, U. Schlomann, P. Golfi, T. Ferdous, S. Naus, and J.W. Bartsch. ADAM8 / MS2 / CD156, an emerging drug target in the treatment of inflammatory and invasive pathologies. *Current pharmaceutical design*, 15(20):2272–2281, 2009.

[118] J. Korlach, P. J. Marks, R. L. Cicero, J. J. Gray, D. L. Murphy, D. B. Roitman, T. T. Pham, G. A. Otto, M. Foquet, and S. W. Turner. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proceedings of the National Academy of Sciences USA*, 105(4):1176–81, 2008.

[119] J. Krause, A. W. Briggs, M. Kircher, T. Maricic, N. Zwyns, A. Derevianko, and S. Pääbo. A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia. *Current Biology*, 2009.

[120] J. Krause, P.H. Dear, J.L. Pollack, M. Slatkin, H. Spriggs, I. Barnes, A.M. Lister, I. Ebersberger, S. Pääbo, and M. Hofreiter. Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature*, 439(7077):724–727, 2005.

[121] J. Krause, Q. Fu, J.M. Good, B. Viola, M.V. Shunkov, A.P. Derevianko, and S. Pääbo. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, 464(7290):894–897, 2010.

[122] J. Krause, L. Orlando, D. Serre, B. Viola, K. Prüfer, M.P. Richards, J.J. Hublin, C. Hänni, A.P. Derevianko, and S. Pääbo. Neanderthals in central Asia and Siberia. *Nature*, 449(7164):902–904, 2007.

[123] H.R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

[124] D.J.G. Lahr and L.A. Katz. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques*, 47(4):857, 2009.

[125] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[126] B. Langmead, C. Trapnell, M. Pop, and S.L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[127] M. Lapidot and Y. Pilpel. Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO reports*, 7(12):1216–1222, 2006.

[128] T. Lassmann, Y. Hayashizaki, and C. O. Daub. Tagdust–a program to eliminate artifacts from next generation sequencing data. *Bioinformatics*, 25(21):2839–40, 2009.

IDT Exhibit 1061.0187

[129] C. Ledergerber and C. Dessimoz. Base-calling for next-generation sequencing platforms. *Briefings in Bioinformatics*, 2011.

[130] B.P. Lewis, C.B. Burge, and D.P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, 2005.

[131] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

[132] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, 2009.

[133] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851, 2008.

[134] R. Li, Y. Li, X. Fang, H. Yang, J. Wang, K. Kristiansen, and J. Wang. SNP detection for massively parallel whole-genome resequencing. *Genome Research*, 19(6):1124, 2009.

[135] R. Li, Y. Li, K. Kristiansen, and J. Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–4, 2008.

[136] R. Li, C. Yu, Y. Li, T.W. Lam, S.M. Yiu, K. Kristiansen, and J. Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966, 2009.

[137] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658, 2006.

[138] W.H. Li, C.I. Wu, and C.C. Luo. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*, 2(2):150, 1985.

[139] S. E. Linsen, E. de Wit, G. Janssens, S. Heater, L. Chapman, R. K. Parkin, B. Fritz, S. K. Wyman, E. de Bruijn, E. E. Voest, S. Kuersten, M. Tewari, and E. Cuppen. Limitations and possibilities of small RNA digital gene expression profiling. *Nature Methods*, 6(7):474–6, 2009.

[140] S.E. Linsen, E. de Wit, G. Janssens, S. Heater, L. Chapman, R.K. Parkin, B. Fritz, S.K. Wyman, E. de Bruijn, E.E. Voest, et al. Limitations and possibilities of small RNA digital gene expression profiling. *Nature methods*, 6(7):474, 2009.

[141] L. Mamanova, A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure, and D. J. Turner. Target-enrichment strategies for next-generation sequencing. *Nature Methods*, 7(2):111–8, 2010.

[142] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80, 2005.

[143] T. Maricic and S. Pääbo. Optimization of 454 sequencing library preparation from small amounts of DNA permits sequence determination of both DNA strands. *Biotechniques*, 46(1):51–2, 54–7, 2009.

IDT Exhibit 1061.0188

[144] Jr. Mariella, R. Sample preparation: the weak link in microfluidics-based biodetection. *Biomedical Microdevices*, 10(6):777–84, 2008.

[145] J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509, 2008.

[146] E. Martinez-Perez and M.P. Colaiácovo. Distribution of meiotic recombination events: talking to your neighbors. *Current opinion in genetics & development*, 19(2):105–112, 2009.

[147] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297, 2010.

[148] P. Medvedev, M. Stanciu, and M. Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6(11 Suppl):S13–20, 2009.

[149] M. L. Metzker. Sequencing technologies — the next generation. *Nature Review Genetics*, 11(1):31–46, 2010.

[150] M. Meyer and M. Kircher. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols*, 2010(6), 2010.

[151] M. Meyer, U. Stenzel, and M. Hofreiter. Parallel tagged sequencing on the 454 platform. *Nature Protocols*, 3(2):267–78, 2008.

[152] A. Meyerhans, J.P. Vartanian, and S. Wain-Hobson. DNA recombination during PCR. *Nucleic Acids Research*, 18(7):1687, 1990.

[153] J. R. Miller, A. L. Delcher, S. Koren, E. Venter, B. P. Walenz, A. Brownley, J. Johnson, K. Li, C. Mobarry, and G. Sutton. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, 24(24):2818–24, 2008.

[154] W. Miller, D.I. Drautz, A. Ratan, B. Pusey, J. Qi, A.M. Lesk, L.P. Tomsho, M.D. Packard, F. Zhao, A. Sher, et al. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, 456(7220):387–390, 2008.

[155] A. Morgulis, G. Coulouris, Y. Raytselis, T.L. Madden, R. Agarwala, and A.A. Schäffer. Database indexing for production MegaBLAST searches. *Bioinformatics*, 24(16):1757, 2008.

[156] E. Mulugeta Achame, W.M. Baarends, J. Gribnau, J.A. Grootegoed, and J.G. Umen. Evaluating the Relationship between Spermatogenic Silencing of the X Chromosome and Evolution of the Y Chromosome in Chimpanzee and Human. *PloS one*, 5(12):424–434, 2010.

[157] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344, 2008.

[158] J.M. Nascimento, L.Z. Shi, S. Meyers, P. Gagneux, N.M. Loskutoff, E.L. Botvinick, and M.W. Berns. The use of optical tweezers to study sperm competition and motility in primates. *Journal of The Royal Society Interface*, 5(20):297, 2008.

IDT Exhibit 1061.0189

[159] R.G. Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine*, 17(8):857–872, 1998.

[160] J.P. Noonan. Regulatory DNAs and the evolution of human development. *Current opinion in genetics & development*, 19(6):557–564, 2009.

[161] K. Nowick, T. Gernat, E. Almaas, and L. Stubbs. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proceedings of the National Academy of Sciences*, 106(52):22358, 2009.

[162] E. Oancea, J. Vriens, S. Brauchi, J. Jun, I. Splawski, and D.E. Clapham. TRPM1 forms ion channels associated with melanin content in melanocytes. *Science Signaling*, 2(70), 2009.

[163] S.J. Odelberg, R.B. Weiss, A. Hata, and R. White. Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. *Nucleic Acids Research*, 23(11):2049, 1995.

[164] K.E. Orii, T. Aoyama, K. Wakui, Y. Fukushima, H. Miyajima, S. Yamaguchi, T. Orii, N. Kondo, and T. Hashimoto. Genomic and mutational analysis of the mitochondrial trifunctional protein β-subunit (HADHB) gene in patients with trifunctional protein deficiency. *Human molecular genetics*, 6(8):1215, 1997.

[165] S. Pääbo. Human evolution. *Trends in Cell Biology*, 9(12):M13–M16, 1999.

[166] S. Pääbo, D.M. Irwin, and A.C. Wilson. DNA damage promotes jumping between templates during enzymatic amplification. *Journal of Biological Chemistry*, 265(8):4718, 1990.

[167] L.E. Palmer, M. Dejori, R. Bolanos, and D. Fasulo. Improving de novo sequence assembly using machine learning and comparative genomics for overlap correction. *BMC Bioinformatics*, 11(1):33, 2010.

[168] J.A. Park and K.C. Kim. Expression patterns of PRDM10 during mouse embryonic development. *Development*, 43(1):29–33, 2010.

[169] P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature Review Genetics*, 10(10):669–80, 2009.

[170] B. Paten, J. Herrero, K. Beal, and E. Birney. Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics*, 25(3):295–301, 2009.

[171] B. Paten, J. Herrero, K. Beal, S. Fitzgerald, and E. Birney. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*, 18(11):1814–28, 2008.

[172] S. Pepke, B. Wold, and A. Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, 6(11 Suppl):S22–32, 2009.

[173] H.N. Poinar, C. Schwarz, J. Qi, B. Shapiro, R.D.E. MacPhee, B. Buigues, A. Tikhonov, D.H. Huson, L.P. Tomsho, A. Auch, et al. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, 311(5759):392, 2006.

IDT Exhibit 1061.0190

[174] K.S. Pollard, S.R. Salama, B. King, A.D. Kern, T. Dreszer, S. Katzman, A. Siepel, J.S. Pedersen, G. Bejerano, R. Baertsch, et al. Forces shaping the fastest evolving regions in the human genome. *PloS Genetics*, 2(10):e168, 2006.

[175] K.S. Pollard, S.R. Salama, N. Lambert, M.A. Lambot, S. Coppens, J.S. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel, et al. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, 443(7108):167–172, 2006.

[176] M. Pop and S. L. Salzberg. Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, 24(3):142–9, 2008.

[177] G. J. Porreca, K. Zhang, J. B. Li, B. Xie, D. Austin, S. L. Vassallo, E. M. LeProust, B. J. Peck, C. J. Emig, F. Dahl, Y. Gao, G. M. Church, and J. Shendure. Multiplex amplification of large sets of human exons. *Nature Methods*, 4(11):931–6, 2007.

[178] S. Prabhakar, J.P. Noonan, S. Pääbo, and E.M. Rubin. Accelerated evolution of conserved noncoding sequences in humans. *Science*, 314(5800):786, 2006.

[179] S. Prabhakar, A. Visel, J.A. Akiyama, M. Shoukry, K.D. Lewis, A. Holt, I. Plajzer-Frick, H. Morrison, D.R. FitzPatrick, V. Afzal, et al. Human-specific gain of function in a developmental enhancer. *Science*, 321(5894):1346, 2008.

[180] K. Prüfer, U. Stenzel, M. Dannemann, R.E. Green, M. Lachmann, and J. Kelso. PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, 24(13):1530, 2008.

[181] K. Prüfer, U. Stenzel, M. Hofreiter, S. Pääbo, J. Kelso, and R.E. Green. Computational challenges in the analysis of ancient DNA. *Genome Biology*, 11(5):R47, 2010.

[182] D. Pushkarev, N. F. Neff, and S. R. Quake. Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, 27(9):847–52, 2009.

[183] M. A. Quail, I. Kozarewa, F. Smith, A. Scally, P. J. Stephens, R. Durbin, H. Swerdlow, and D. J. Turner. A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, 5(12):1005–10, 2008.

[184] A. R. Quinlan, D. A. Stewart, M. P. Stromberg, and G. T. Marth. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature Methods*, 5(2):179–81, 2008.

[185] M. Rasmussen, Y. Li, S. Lindgreen, J.S. Pedersen, A. Albrechtsen, I. Moltke, M. Metspalu, E. Metspalu, T. Kivisild, R. Gupta, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, 463(7282):757–762, 2010.

[186] D. Reich, R.E. Green, M. Kircher, J. Krause, N. Patterson, E.Y. Durand, B. Viola, A.W. Briggs, U. Stenzel, P.L.F. Johnson, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–1060, 2010.

[187] D. Reich, K. Thangaraj, N. Patterson, A.L. Price, and L. Singh. Reconstructing Indian population history. *Nature*, 461(7263):489–494, 2009.

[188] J. A. Reinhardt, D. A. Baltrus, M. T. Nishimura, W. R. Jeck, C. D. Jones, and J. L. Dangl. De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae. *Genome Research*, 19(2):294–305, 2009.

IDT Exhibit 1061.0191

[189] IBM Research. IBM Research Aims to Build Nanoscale DNA Sequencer to Help Drive Down Cost of Personalized Genetic Analysis, 2009.

[190] B. Rhead, D. Karolchik, R.M. Kuhn, A.S. Hinrichs, A.S. Zweig, P.A. Fujita, M. Diekhans, K.E. Smith, K.R. Rosenbloom, B.J. Raney, et al. The UCSC genome browser database: update 2010. *Nucleic Acids Research*, 2009.

[191] B. G. Richter and D. P. Sexton. Managing and analyzing next-generation sequence data. *PLoS Computational Biology*, 5(6):e1000369, 2009.

[192] M.D. Robinson, D.J. McCarthy, and G.K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139, 2010.

[193] M.D. Robinson and G.K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881, 2007.

[194] M.D. Robinson and G.K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321, 2008.

[195] N. Rohland and M. Hofreiter. Comparison and optimization of ancient DNA extraction. *Biotechniques*, 42(3):343–52, 2007.

[196] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlen, and P. Nyren. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, 242(1):84–9, 1996.

[197] M. G. Roper, C. J. Easley, L. A. Legendre, J. A. Humphrey, and J. P. Landers. Infrared temperature control system for a completely noncontact polymerase chain reaction in microfluidic chips. *Analytical Chemistry*, 79(4):1294–300, 2007.

[198] J. Rougemont, A. Amzallag, C. Iseli, L. Farinelli, I. Xenarios, and F. Naef. Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics*, 9:431, 2008.

[199] JP Royston. An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31(2):115–124, 1982.

[200] M. Safran, I. Dalah, J. Alexander, N. Rosen, T. Iny Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug, et al. GeneCards Version 3: the human gene integrator. *Database*, 2010(0), 2010.

[201] A.H. Salem, D.A. Ray, J. Xing, P.A. Callinan, J.S. Myers, D.J. Hedges, R.K. Garber, D.J. Witherspoon, L.B. Jorde, and M.A. Batzer. Alu elements and hominid phylogenetics. *Proceedings of the National Academy of Sciences of the United States of America*, 100(22):12787, 2003.

[202] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–95, 1977.

[203] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–8, 1975.

[204] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA*, 74(12):5463–7, 1977.

IDT Exhibit 1061.0192

[205] M. Schena, D. Shalon, RW Davis, and PO Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467, 1995.

[206] S. C. Schuster. Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1):16–8, 2008.

[207] N. Servant, E. Gravier, P. Gestraud, C. Laurent, C. Paccard, A. Biton, I. Brito, J. Mandel, B. Asselain, E. Barillot, et al. EMA- A R package for Easy Microarray data analysis. *BMC Research Notes*, 3(1):277, 2010.

[208] S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591, 1965.

[209] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–45, 2008.

[210] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–32, 2005.

[211] J. A. Shendure, G. J. Porreca, and G. M. Church. Overview of DNA sequencing strategies. *Current Protocols in Molecular Biology*, Chapter 7:Unit 7 1, 2008.

[212] ST Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, EM Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308, 2001.

[213] K. Shibata, M. Itoh, K. Aizawa, S. Nagaoka, N. Sasaki, P. Carninci, H. Konno, J. Akiyama, K. Nishi, T. Kitsunai, H. Tashiro, N. Sumi, Y. Ishii, S. Nakamura, M. Hazama, T. Nishine, A. Harada, R. Yamamoto, H. Matsumoto, S. Sakaguchi, T. Ikegami, K. Kashiwagi, S. Fujiwake, K. Inoue, and Y. Togawa. RIKEN integrated sequence analysis (RISA) system–384-format sequencing pipeline with 384 multicapillary sequencer. *Genome Research*, 10(11):1757–71, 2000.

[214] F. Shu, S. Ramineni, and J.R. Hepler. RGS14 is a multifunctional scaffold that integrates G protein and Ras/Raf MAPkinase signalling pathways. *Cellular signalling*, 22(3):366–376, 2010.

[215] D.A. Siegel, M.K. Huang, and S.F. Becker. Ectopic dendrite initiation: CNS pathogenesis as a model of CNS development* 1. *International Journal of Developmental Neuroscience*, 20(3-5):373–389, 2002.

[216] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–9, 1986.

[217] B.A. Sokhansanj, G.R. Rodrigue, J.P. Fitch, and M.W. David III. A quantitative model of human DNA base excision repair. I. Mechanistic insights. *Nucleic acids research*, 30(8):1817, 2002.

[218] M. Somel, H. Franz, Z. Yan, A. Lorenc, S. Guo, T. Giger, J. Kelso, B. Nickel, M. Dannemann, S. Bahn, et al. Transcriptional neoteny in the human brain. *Proceedings of the National Academy of Sciences USA*, 106(14):5743, 2009.

IDT Exhibit 1061.0193

[219] A. Stark, J. Brennecke, N. Bushati, R.B. Russell, and S.M. Cohen. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6):1133–1146, 2005.

[220] M.E. Steiper and N.M. Young. Primate molecular divergence dates. *Molecular phylogenetics and evolution*, 41(2):384–394, 2006.

[221] M. Stiller, M. Knapp, U. Stenzel, M. Hofreiter, and M. Meyer. Direct multiplex sequencing (DMPS) – a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. *Genome Research*, 19(10):1843, 2009.

[222] M.R. Stratton, P.J. Campbell, and P.A. Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.

[223] H. Swerdlow and R. Gesteland. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research*, 18(6):1415–9, 1990.

[224] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[225] A. Tenesa, P. Navarro, B.J. Hayes, D.L. Duffy, G.M. Clarke, M.E. Goddard, and P.M. Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, 17(4):520, 2007.

[226] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007.

[227] J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673, 1994.

[228] C. Trapnell, L. Pachter, and S.L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105, 2009.

[229] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. Van Baren, S.L. Salzberg, B.J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[230] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(2):1453, 2006.

[231] G. Turcatti, A. Romieu, M. Fedurco, and A. P. Tairi. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research*, 36(4):e25, 2008.

[232] E. H. Turner, C. Lee, S. B. Ng, D. A. Nickerson, and J. Shendure. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature Methods*, 6(5):315–6, 2009.

[233] K. Ueno and E. S. Yeung. Simultaneous monitoring of DNA fragments separated by electrophoresis in a multiplexed array of 100 capillaries. *Analytical Chemistry*, 66(9):1424–1431, 1994.

IDT Exhibit 1061.0194

[234] V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484, 1995.

[235] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304, 2001.

[236] J.D. Wall and M.F. Hammer. Archaic admixture in the human genome. *Current opinion in genetics & development*, 16(6):606–610, 2006.

[237] P. K. Wall, J. Leebens-Mack, A. S. Chanderbali, A. Barakat, E. Wolcott, H. Liang, L. Landherr, L. P. Tomsho, Y. Hu, J. E. Carlson, H. Ma, S. C. Schuster, D. E. Soltis, P. S. Soltis, N. Altman, and C. W. dePamphilis. Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, 10:347, 2009.

[238] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136, 2010.

[239] D. Weigel and R. Mott. The 1001 genomes project for Arabidopsis thaliana. *Genome Biology*, 10(5):107, 2009.

[240] A. Wetterbom, A. Ameur, L. Feuk, U. Gyllensten, and L. Cavelier. Identification of novel exons and transcribed regions by chimpanzee transcriptome sequencing. *Genome Biology*, 11(7):R78, 2010.

[241] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X. Z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs, and J. M. Rothberg. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–6, 2008.

[242] N. Whiteford, T. Skelly, C. Curtis, M.E. Ritchie, A. Löhr, A.W. Zaranek, I. Abnizova, and C. Brown. Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics*, 25(17):2194, 2009.

[243] T. Wicker, E. Schlagenhauf, A. Graner, T. J. Close, B. Keller, and N. Stein. 454 sequencing put to the test using the complex genome of barley. *BMC Genomics*, 7:275, 2006.

[244] E.B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[245] R. Wu and A. D. Kaiser. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *Journal of Molecular Biology*, 35(3):523–37, 1968.

[246] Z. Wu, R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909–917, 2004.

[247] R. Yelin, D. Dahary, R. Sorek, E.Y. Levanon, O. Goldstein, A. Shoshan, A. Diber, S. Biton, Y. Tamir, R. Khosravi, et al. Widespread occurrence of antisense transcription in the human genome. *Nature Biotechnology*, 21(4):379–386, 2003.

IDT Exhibit 1061.0195

[248] J.J. Yunis and OM Prakash. The origin of man: a chromosomal pictorial legacy. *Science*, 215(4539):1525, 1982.

[249] R. J. Zagursky and R. M. McCormick. DNA sequencing separations in capillary gels on a modified commercial DNA sequencing instrument. *Biotechniques*, 9(1):74–9, 1990.

[250] D.R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821, 2008.

[251] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, 7(1-2):203–214, 2000.

[252] X. Zhou, Z. Su, R. D. Sammons, Y. Peng, P. J. Tranel, Jr. Stewart, C. N., and J. S. Yuan. Novel software package for cross-platform transcriptome analysis (CPTRA). *BMC Bioinformatics*, 10 Suppl 11:S16, 2009.

[253] Z. Zhu and A. S. Waggoner. Molecular mechanism controlling the incorporation of fluorescent nucleotides into DNA by PCR. *Cytometry*, 28(3):206–11, 1997.

[254] A. V. Zimin, A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biology*, 10(4):R42, 2009.

IDT Exhibit 1061.0196

Appendix

Tables

Table 1: Results of the simulation of 10,000 clusters, each with a thousand identical sequences. Random sequences were created using the GC content of ϕX 174 (44.7%) as a reference. I used a model with pre-phasing (0.4% per cycle), phasing (0.4% per cycle) and T accumulation (3.8% per cycle) and simulated the fluorophores attached for 150 sequencing cycles. For each cycle, the number of fluorophores attached to the sequences of the cluster was determined and the fraction of fluorophores representing the current cycle, the previous and the next cycles as well as representing cycles more than one ahead and more than one behind calculated. Further, the fraction of fluorophores attached due to T accumulation was determined. The results were averaged over all 1,000 sequences. T accumulation is given as a fraction of the starting fluorophores in a cluster.

Cycle	Back>1	Back=1	InPhase	Ahead=1	Ahead>1	T accumulation
1	0.000	0.004	0.992	0.004	0.000	0.000
2	0.000	0.008	0.984	0.008	0.000	0.010
3	0.000	0.012	0.976	0.012	0.000	0.021
4	0.000	0.016	0.969	0.016	0.000	0.032
5	0.000	0.019	0.961	0.019	0.000	0.042
6	0.000	0.023	0.954	0.023	0.000	0.053
7	0.000	0.027	0.946	0.027	0.000	0.063
8	0.000	0.030	0.939	0.030	0.000	0.073
9	0.001	0.034	0.931	0.034	0.001	0.083
10	0.001	0.037	0.924	0.037	0.001	0.094
11	0.001	0.041	0.917	0.040	0.001	0.105
12	0.001	0.044	0.910	0.044	0.001	0.115
13	0.001	0.047	0.903	0.047	0.001	0.126
14	0.001	0.050	0.897	0.050	0.001	0.136
15	0.002	0.054	0.890	0.053	0.002	0.147
16	0.002	0.057	0.883	0.057	0.002	0.157
17	0.002	0.060	0.877	0.060	0.002	0.168
18	0.002	0.063	0.870	0.063	0.002	0.178
19	0.002	0.066	0.864	0.066	0.002	0.188
20	0.003	0.069	0.857	0.069	0.003	0.199
21	0.003	0.072	0.851	0.071	0.003	0.210
22	0.003	0.075	0.845	0.074	0.003	0.220
23	0.004	0.077	0.839	0.077	0.003	0.231
24	0.004	0.080	0.833	0.080	0.004	0.241
25	0.004	0.083	0.827	0.082	0.004	0.252

IDT Exhibit 1061.0197

Cycle	Back>1	Back=1	InPhase	Ahead=1	Ahead>1	T accumulation
26	0.004	0.085	0.821	0.085	0.004	0.262
27	0.005	0.088	0.815	0.088	0.005	0.273
28	0.005	0.091	0.809	0.090	0.005	0.283
29	0.005	0.093	0.803	0.093	0.005	0.294
30	0.006	0.096	0.798	0.095	0.006	0.304
31	0.006	0.098	0.792	0.098	0.006	0.315
32	0.007	0.100	0.786	0.100	0.006	0.325
33	0.007	0.103	0.781	0.102	0.007	0.336
34	0.007	0.105	0.776	0.105	0.007	0.346
35	0.008	0.107	0.770	0.107	0.008	0.357
36	0.008	0.110	0.765	0.109	0.008	0.367
37	0.008	0.112	0.760	0.111	0.008	0.378
38	0.009	0.114	0.755	0.114	0.009	0.388
39	0.009	0.116	0.750	0.116	0.009	0.399
40	0.010	0.118	0.744	0.118	0.010	0.409
41	0.010	0.120	0.739	0.120	0.010	0.420
42	0.011	0.122	0.734	0.122	0.011	0.430
43	0.011	0.124	0.730	0.124	0.011	0.441
44	0.012	0.126	0.725	0.126	0.011	0.451
45	0.012	0.128	0.720	0.128	0.012	0.462
46	0.012	0.130	0.715	0.130	0.012	0.472
47	0.013	0.132	0.711	0.132	0.013	0.482
48	0.013	0.134	0.706	0.133	0.013	0.493
49	0.014	0.136	0.701	0.135	0.014	0.503
50	0.014	0.137	0.697	0.137	0.014	0.514
51	0.015	0.139	0.692	0.139	0.015	0.524
52	0.015	0.141	0.688	0.140	0.015	0.535
53	0.016	0.143	0.684	0.142	0.016	0.545
54	0.016	0.144	0.679	0.144	0.016	0.556
55	0.017	0.146	0.675	0.145	0.017	0.566
56	0.017	0.147	0.671	0.147	0.017	0.577
57	0.018	0.149	0.667	0.149	0.018	0.588
58	0.019	0.151	0.662	0.150	0.018	0.598
59	0.019	0.152	0.658	0.152	0.019	0.609
60	0.020	0.154	0.654	0.153	0.019	0.619
61	0.020	0.155	0.650	0.155	0.020	0.629
62	0.021	0.157	0.646	0.156	0.020	0.640
63	0.021	0.158	0.642	0.157	0.021	0.650
64	0.022	0.159	0.639	0.159	0.022	0.661
65	0.022	0.161	0.635	0.160	0.022	0.671
66	0.023	0.162	0.631	0.162	0.023	0.682
67	0.023	0.163	0.627	0.163	0.023	0.692
68	0.024	0.165	0.624	0.164	0.024	0.702
69	0.025	0.166	0.620	0.165	0.024	0.713
70	0.025	0.167	0.616	0.167	0.025	0.723
71	0.026	0.168	0.613	0.168	0.026	0.734
72	0.026	0.169	0.609	0.169	0.026	0.745
73	0.027	0.171	0.606	0.170	0.027	0.755
74	0.027	0.172	0.602	0.171	0.027	0.766

IDT Exhibit 1061.0198

Cycle	Back>1	Back=1	InPhase	Ahead=1	Ahead>1	T accumulation
75	0.028	0.173	0.599	0.172	0.028	0.776
76	0.029	0.174	0.595	0.174	0.028	0.787
77	0.029	0.175	0.592	0.175	0.029	0.797
78	0.030	0.176	0.589	0.176	0.030	0.807
79	0.030	0.177	0.585	0.177	0.030	0.818
80	0.031	0.178	0.582	0.178	0.031	0.828
81	0.032	0.179	0.579	0.179	0.031	0.839
82	0.032	0.180	0.576	0.180	0.032	0.849
83	0.033	0.181	0.572	0.181	0.033	0.860
84	0.033	0.182	0.569	0.182	0.033	0.870
85	0.034	0.183	0.566	0.183	0.034	0.881
86	0.035	0.184	0.563	0.183	0.034	0.891
87	0.035	0.185	0.560	0.184	0.035	0.902
88	0.036	0.186	0.557	0.185	0.036	0.912
89	0.037	0.187	0.554	0.186	0.036	0.923
90	0.037	0.188	0.551	0.187	0.037	0.934
91	0.038	0.188	0.548	0.188	0.038	0.944
92	0.038	0.189	0.545	0.189	0.038	0.955
93	0.039	0.190	0.543	0.189	0.039	0.965
94	0.040	0.191	0.540	0.190	0.039	0.976
95	0.040	0.192	0.537	0.191	0.040	0.986
96	0.041	0.192	0.534	0.192	0.041	0.997
97	0.042	0.193	0.532	0.192	0.041	1.008
98	0.042	0.194	0.529	0.193	0.042	1.018
99	0.043	0.195	0.526	0.194	0.042	1.029
100	0.043	0.195	0.524	0.195	0.043	1.039
101	0.044	0.196	0.521	0.195	0.044	1.050
102	0.045	0.197	0.518	0.196	0.044	1.061
103	0.045	0.197	0.516	0.197	0.045	1.071
104	0.046	0.198	0.513	0.197	0.046	1.082
105	0.047	0.199	0.511	0.198	0.046	1.092
106	0.047	0.199	0.508	0.199	0.047	1.103
107	0.048	0.200	0.506	0.199	0.047	1.113
108	0.048	0.201	0.503	0.200	0.048	1.123
109	0.049	0.201	0.501	0.200	0.049	1.134
110	0.050	0.202	0.498	0.201	0.049	1.145
111	0.050	0.202	0.496	0.201	0.050	1.155
112	0.051	0.203	0.494	0.202	0.051	1.166
113	0.052	0.203	0.491	0.202	0.051	1.176
114	0.052	0.204	0.489	0.203	0.052	1.186
115	0.053	0.204	0.487	0.203	0.053	1.197
116	0.054	0.205	0.485	0.204	0.053	1.207
117	0.054	0.205	0.482	0.204	0.054	1.217
118	0.055	0.206	0.480	0.205	0.054	1.228
119	0.055	0.206	0.478	0.205	0.055	1.238
120	0.056	0.207	0.476	0.206	0.056	1.249
121	0.057	0.207	0.474	0.206	0.056	1.259
122	0.057	0.207	0.472	0.207	0.057	1.270
123	0.058	0.208	0.469	0.207	0.057	1.281

IDT Exhibit 1061.0199

Cycle	Back>1	Back=1	InPhase	Ahead=1	Ahead>1	T accumulation
124	0.059	0.208	0.467	0.208	0.058	1.291
125	0.059	0.209	0.465	0.208	0.059	1.302
126	0.060	0.209	0.463	0.208	0.059	1.312
127	0.060	0.210	0.461	0.209	0.060	1.322
128	0.061	0.210	0.459	0.209	0.061	1.333
129	0.062	0.210	0.457	0.209	0.061	1.343
130	0.062	0.211	0.455	0.210	0.062	1.354
131	0.063	0.211	0.453	0.210	0.063	1.365
132	0.064	0.211	0.451	0.211	0.063	1.375
133	0.064	0.212	0.449	0.211	0.064	1.385
134	0.065	0.212	0.447	0.211	0.064	1.396
135	0.066	0.212	0.446	0.212	0.065	1.406
136	0.066	0.213	0.444	0.212	0.066	1.417
137	0.067	0.213	0.442	0.212	0.066	1.428
138	0.067	0.213	0.440	0.212	0.067	1.438
139	0.068	0.214	0.438	0.213	0.068	1.448
140	0.069	0.214	0.436	0.213	0.068	1.459
141	0.069	0.214	0.435	0.213	0.069	1.469
142	0.070	0.214	0.433	0.213	0.069	1.480
143	0.071	0.214	0.431	0.214	0.070	1.491
144	0.071	0.215	0.429	0.214	0.071	1.501
145	0.072	0.215	0.428	0.214	0.071	1.512
146	0.072	0.215	0.426	0.214	0.072	1.522
147	0.073	0.215	0.424	0.215	0.072	1.533
148	0.074	0.216	0.423	0.215	0.073	1.543
149	0.074	0.216	0.421	0.215	0.074	1.553
150	0.075	0.216	0.419	0.215	0.074	1.560

IDT Exhibit 1061.0200

Table 2: Table of 78 single nucleotide differences in coding sequences of CCDS genes for which Neandertal shows the ancestral (Chimpanzee-like state) while modern humans are fixed for the derived state. The table is sorted by Grantham scores (GS). Based on the classification proposed by Li [138], 5 amino acid substitutions are considered radical (> 150), 7 moderately radical (101-150), 33 moderately conservative (51-100), 32 conservative (1-50) and one change in stop-codon falls out of this scoring scheme. Genes showing multiple substitution changes are marked with bold database identifiers.

Human (derived)				Chimpanzee (ancestral)				Database identifier			Amino acid information			
Base	Chr	+/-	Pos	Base	Chr	+/-	Pos	EnsemblTranscript	CCDS ID	SwissProt	Pos	Base	AA	GS
A	1	+	150393846	G	1	+	131234541	**ENST00000316073**	**41397**	**RPTN**	785	1	*/R	-
C	2	+	11675942	T	2a	+	11846259	ENST00000381486	42655	GREB1	1164	1	R/C	180
C	9	+	124603021	T	9	+	122451445	ENST00000277309	35132	OR1K1	267	1	R/C	180
A	1	+	118435820	C	1	-	119478811	**ENST00000336338**	**899**	**SPG17**	431	1	Y/D	160
T	11	+	118550510	G	11	+	118054004	ENST00000292199	8416	NLRX1	330	1	Y/D	160
C	3	+	95285751	T	3	+	97827124	ENST00000314622	2927	NSUN3	78	2	S/F	155
T	1	+	180836069	G	1	+	162274446	ENST00000367558	1348	RGS16	197	2	D/A	126
C	4	+	13206636	G	4	+	13452091	ENST00000040738	3411	BOD1L	2684	1	G/R	125
T	6	+	121644484	A	6	+	123350109	ENST00000368464	43501	CF170	505	1	S/C	112
G	7	+	89631971	C	7	+	89789078	ENST00000297205	5614	STEA1	336	2	C/S	112
C	11	+	6195544	A	11	+	6063773	ENST00000265978	7760	F16A2	630	3	R/S	110
G	15	+	39585013	T	15	+	38515020	ENST00000263800	10077	LTK	569	1	R/S	110
A	X	+	18131667	C	X	+	18244953	ENST00000380033	14184	BEND2	261	2	V/G	109
C	11	+	6177181	T	11	+	6045584	ENST00000311352	31407	O52W1	51	2	P/L	98
T	16	+	538119	C	16_r.	+	5709532	**ENST00000219611**	**10410**	**CAN15**	427	2	L/P	98
A	3	+	47444153	G	3	+	48489458	ENST00000265565	2755	SCAP	140	2	I/T	89
A	9	+	134265413	G	9	+	132451336	**ENST00000334270**	**6948**	**TTF1**	474	2	I/T	89
C	3	+	99555910	G	3	+	102255730	ENST00000354924	33802	OR5K4	175	1	H/D	81
C	X	+	17678236	T	X	+	17782496	ENST00000380041	35210	SCML1	202	2	T/M	81
A	1	+	1110351	C	1	+	1106337	ENST00000400931	8	TTL10	394	2	K/T	78
A	2	+	99577084	G	2a	+	100577578	ENST00000356421	33258	AFF3	516	1	S/P	74
T	20	+	45078304	C	20	+	44410094	ENST00000360649	42887	EYA2	131	1	S/P	74
G	4	+	2919072	C	4	+	3060926	ENST00000314262	33945	NOP14	493	2	T/R	71
T	11	+	129277503	G	11	+	128968019	ENST00000360871	8484	PRDM10	1129	2	N/T	65
T	3	+	113671299	G	3	+	116666154	ENST00000334529	33819	BTLA	197	2	N/T	65
A	11	+	74477736	G	11	+	73459069	ENST00000305159	31639	O2AT4	224	2	V/A	64

IDT Exhibit 1061.0201

Base	Chr	+/-	Pos	Base	Chr	+/-	Pos	EnsemblTranscript	CCDS ID	SwissProt	Pos	Base	AA	GS
T	16	+	537906	C	16_r.	+	5709319	**ENST00000219611**	**10410**	**CAN15**	356	2	V/A	64
T	2	+	220087788	C	2b	+	225458007	ENST00000358078	33384	ACCN4	160	2	V/A	64
T	22	+	39090924	C	22	+	39366824	ENST00000216194	14001	PUR8	429	2	V/A	64
G	6	+	100475589	A	6	+	101531965	ENST00000281806	5044	MCHR2	324	2	A/V	64
T	7	+	17341917	C	7	+	17496236	ENST00000242057	5366	AHR	381	2	V/A	64
C	1	+	46650472	G	1	+	47202157	ENST00000243167	535	FAAH1	476	2	A/G	60
T	1	+	118360155	C	1	-	119555018	**ENST00000336338**	**899**	**SPG17**	1415	1	T/A	58
C	15	+	40529604	T	15	+	39548722	ENST00000263805	32208	ZF106	697	1	A/T	58
T	16	+	65504565	C	16	+	66617267	ENST00000299752	10823	CAD16	342	1	T/A	58
T	17	+	37020864	C	17	-	15934889	ENST00000301653	11401	K1C16	306	1	T/A	58
T	2	+	128113346	C	2b	+	128489293	ENST00000324938	2147	LIMS2	360	1	T/A	58
A	3	+	44737863	G	3	+	45716980	ENST00000296091	2719	ZN502	184	1	T/A	58
G	4	+	88986215	A	4	+	90763412	ENST00000361056	3625	MEPE	391	1	A/T	58
T	5	+	132562844	C	5	+	134834351	ENST00000265342	34238	FSTL4	791	1	T/A	58
C	8	+	51628204	G	8	+	48568603	ENST00000338349	6147	SNTG1	241	2	T/S	58
T	1	+	150393996	C	1	+	131234775	**ENST00000316073**	**41397**	**RPTN**	735	1	K/E	56
T	11	+	118278035	C	11	+	117781441	ENST00000334801	8403	BCL9L	543	1	S/G	56
T	17	+	24983160	C	17	-	27682886	ENST00000269033	11253	SSH2	1033	1	S/G	56
T	19	+	62017061	C	19	+	62668946	ENST00000326441	12948	PEG3	1521	1	S/G	56
T	21	+	33782703	G	21	+	33236795	ENST00000381947	13626	DJC28	290	1	K/Q	53
T	13	+	48179102	G	13	+	48595018	ENST00000282018	9412	CLTR2	50	1	F/V	50
A	3	+	44831503	G	3	+	45821006	ENST00000326047	33744	KIF15	827	2	N/S	46
T	1	+	32052458	C	1	+	32221622	ENST00000360482	347	SPOC1	355	2	Q/R	43
C	9	+	134267344	T	9	+	132453251	**ENST00000334270**	**6948**	**TTF1**	229	2	R/Q	43
T	9	+	139259702	G	9	+	137500097	ENST00000344774	35186	F166A	134	1	T/P	38
C	12	+	62873951	G	12	-	25278113	ENST00000398055	41803	CL066	426	1	V/L	32
C	11	+	6611387	G	11	+	6490781	**ENST00000299441**	**7771**	**PCD16**	763	1	E/Q	29
T	11	+	2383887	C	11	+	2449712	ENST00000155858	31340	TRPM5	1088	1	I/V	29
T	11	+	92535568	C	11	+	91690970	ENST00000326402	8291	S36A4	330	2	H/R	29
C	14	+	104588537	G	14	+	105590623	ENST00000392585	9997	GP132	328	1	E/Q	29
T	14	+	67344044	C	14	+	67361749	ENST00000347230	9788	ZFY26	237	2	H/R	29
A	7	+	134293531	G	7	+	135457097	ENST00000361675	5835	CALD1	671	1	I/V	29
A	8	+	25416950	G	8	+	21955896	ENST00000330560	6049	CDCA2	606	1	I/V	29
G	8	+	145211313	C	8	+	144064958	ENST00000355091	43776	GPAA1	275	1	E/Q	29
A	X	+	3012475	G	X_r.	+	1786421	ENST00000381127	14123	ARSF	200	1	I/V	29

IDT Exhibit 1061.0202

Base	Chr	+/-	Pos	Base	Chr	+/-	Pos	EnsemblTranscript	CCDS ID	SwissProt	Pos	Base	AA	GS
G	11	+	59039869	A	11	+	57731876	ENST00000329328	31564	OR4D9	303	2	R/K	26
G	18	+	2880589	A	18	-	13843773	ENST00000254528	11828	EMIL2	155	2	R/K	26
T	9	+	124622444	C	9	+	122471604	ENST00000259467	6845	PHLP	216	2	K/R	26
G	X	+	153196802	A	X	+	153627492	ENST00000369915	35448	TKTL1	317	2	R/K	26
C	1	+	12012533	G	1	+	12228263	ENST00000235332	143	MIIP	280	3	H/Q	24
T	1	+	156914834	C	1	+	137934885	ENST00000368148	41423	SPTA1	265	1	N/D	23
C	11	+	6611345	T	11	+	6490739	**ENST00000299441**	**7771**	**PCD16**	777	1	D/N	23
C	19	+	3498315	G	19	+	3591211	ENST00000398558	42464	CS028	326	3	L/F	22
G	3	+	198158892	A	3	+	202594008	ENST00000238138	3324	PIGZ	425	1	L/F	22
G	1	+	221244597	A	1	+	203711670	ENST00000284476	1536	DISP1	1079	1	V/M	21
A	14	+	20581121	G	14	+	19941316	ENST00000298690	41914	RNAS7	44	1	M/V	21
C	21	+	30576509	T	21	+	30082864	ENST00000340345	42915	KR241	205	1	V/M	21
A	20	+	31275867	G	20	+	30217610	ENST00000375454	13216	SPLC3	108	3	I/M	10
A	20	+	32801190	C	20	+	31822769	ENST00000374796	13241	NCOA6	823	3	I/M	10
G	4	+	184423847	T	4	+	187919923	ENST00000281445	34109	WWC2	479	3	M/I	10
T	10	+	73557900	A	10	+	71258470	ENST00000394919	31219	ASCC1	301	3	E/D	0
C	2	+	95309419	G	2a	+	96196093	ENST00000317668	2012	PROM2	458	3	D/E	0

IDT Exhibit 1061.0203

Table 3: Table of 129 single nucleotide differences in coding sequences of CCDS genes for which Denisova shows the ancestral (Chimpanzee-like state) while modern humans are fixed for the derived state. The table is sorted by Grantham scores (GS). Based on the classification proposed by Li [138], 1 amino acid substitution is considered radical (> 150), 8 moderately radical (101-150), 65 moderately conservative (51-100), 54 conservative (1-50) and one change in stop-codon falls out of this scoring scheme. Genes showing multiple substitution changes are marked with bold database identifiers. Genomic coordinates are zero-based.

Human (derived)				Chimpanzee (ancestral)				Database identifier			Amino acid information			
Base	Chr	+/-	Pos	Base	Chr	+/-	Pos	EnsemblTranscript	CCDS ID	SwissProt	Pos	Base	AA	GS
C	1	+	160234303	T	1	+	141210333	ENST00000294794	1236	OLM2B	470	2	W/*	-
C	9	+	124603020	T	9	+	122451444	ENST00000277309	35132	OR1K1	267	1	R/C	180
T	16	+	55853364	A	16	+	56701776	ENST00000219207	10777	PLLP	85	2	N/I	149
A	6	+	79634102	G	6	+	79859308	ENST00000369940	34488	IKBP1	31	1	R/G	125
A	6	+	28033607	T	6	+	28474934	ENST00000244623	4642	OR2B6	204	2	E/V	121
T	19	+	56195777	A	19	+	56666060	ENST00000391806	42600	KLK8	27	1	S/C	112
G	5	+	118513136	C	5	+	120549991	ENST00000311085	4125	DMXL1	1239	2	C/S	112
G	15	+	39585012	T	15	+	38515019	ENST00000263800	10077	LTK	569	1	R/S	110
T	1	+	89500647	G	1	+	90748162	ENST00000370459	722	GBP5	497	2	E/A	107
A	17	+	71516433	G	17	+	75622154	ENST00000301607	11737	EVPL	1483	1	W/R	101
A	13	+	83352655	C	13	+	84340069	ENST00000377084	9464	SLIK1	330	1	S/A	99
A	1	+	1221067	G	1	+	1209836	ENST00000354980	19	ACAP3	497	2	L/P	98
A	1	+	89370660	G	1	+	90615166	ENST00000294671	720	GBP7	559	2	L/P	98
C	10	+	37548307	T	10	+	38070616	**ENST00000361713**	**7193**	**AN30A**	1165	2	P/L	98
T	16	+	538118	C	16_r.	+	5709531	ENST00000219611	10410	CAN15	427	2	L/P	98
C	17	+	23943903	T	17	-	28754402	ENST00000321765	32594	SPAG5	162	2	G/E	98
A	19	+	59495378	G	19	+	60019602	ENST00000391745	12887	LIRA3	103	2	L/P	98
C	22	+	49002257	T	22	+	49544877	ENST00000248846	14087	GCP6	886	2	G/E	98
A	5	+	86600232	G	5	-	28406904	ENST00000274376	34200	RASA1	70	2	E/G	98
T	9	+	2719704	C	9	+	2767007	ENST00000382082	6447	KCNV2	539	2	L/P	98
T	10	+	118311044	A	10	+	117285929	ENST00000369221	7594	LIPP	414	2	M/K	95
C	17	+	59644188	T	17	+	63501641	ENST00000258991	11658	TEX2	374	2	G/D	94
T	2	+	241112138	G	2b	+	246936269	ENST00000391987	2536	ANKY1	467	3	K/N	94
C	4	+	4250211	T	4	+	4284217	ENST00000296358	3372	OTOP1	417	2	G/D	94
C	1	+	35351279	T	1	+	35619202	ENST00000359858	41302	ZMYM1	421	2	T/I	89
C	1	+	40499233	T	1	+	40903941	ENST00000372759	449	FACE1	87	2	T/I	89

IDT Exhibit 1061.0204

Base	Chr	+/-	Pos	Base	Chr	+/-	Pos	EnsemblTranscript	CCDS ID	SwissProt	Pos	Base	AA	GS
G	2	+	40510859	A	2a	+	41372434	ENST00000378715	1806	NAC1	22	2	T/I	89
A	21	+	29226747	G	21	+	28747960	ENST00000361371	33527	RN160	1662	2	I/T	89
A	3	+	47444152	G	3	+	48489457	ENST00000265565	2755	SCAP	140	2	I/T	89
A	9	+	134265412	G	9	+	132451335	**ENST00000334270**	**6948**	**TTF1**	474	2	I/T	89
C	17	+	3066433	T	17	+	3240106	ENST00000304094	11022	OR1A1	257	2	T/M	81
C	3	+	99555909	G	3	+	102255729	**ENST00000354924**	**33802**	**OR5K4**	175	1	H/D	81
G	3	+	198159340	A	3	+	202594456	**ENST00000238138**	**3324**	**PIGZ**	275	2	T/M	81
C	9	+	126152975	G	9	+	124035499	ENST00000320246	6854	NEK6	291	1	H/D	81
C	X	+	17678235	T	X	+	17782495	ENST00000380041	35210	SCML1	202	2	T/M	81
C	X	+	22928705	T	X	+	23118523	ENST00000327968	35214	DDX53	204	2	T/M	81
C	5	+	75627399	A	5	-	39561669	ENST00000322285	43331	SV2C	460	2	P/H	77
A	1	+	63831784	C	1	+	64730850	ENST00000371084	625	PGM1	13	2	Q/P	76
A	14	+	95842916	G	14	+	96604596	ENST00000359933	9944	ATG2B	1465	1	S/P	74
T	3	+	121952209	C	3	+	125367499	ENST00000283875	3002	T2EA	41	1	S/P	74
G	18	+	64715493	C	18	+	65628674	ENST00000360242	11996	C102B	371	2	R/T	71
G	4	+	2919071	C	4	+	3060925	ENST00000314262	33945	NOP14	493	2	T/R	71
A	1	+	159117069	G	1	+	140167930	ENST00000326245	1211	ITLN1	206	2	V/A	64
A	17	+	32988030	G	17	-	19834375	ENST00000346661	11321	SYNG	636	2	V/A	64
A	21	+	41788292	G	21	+	41197009	ENST00000332149	33564	TMPS2	33	2	V/A	64
T	22	+	39090923	C	22	+	39366823	ENST00000216194	14001	PUR8	429	2	V/A	64
G	6	+	100475588	A	6	+	101531964	ENST00000281806	5044	MCHR2	324	2	A/V	64
T	7	+	17341916	C	7	+	17496235	ENST00000242057	5366	AHR	381	2	V/A	64
A	8	+	10507836	G	8_r.	+	5845849	**ENST00000382483**	**43708**	**RP1L1**	394	2	V/A	64
G	X	+	3249673	A	X	+	3261751	ENST00000217939	14124	MXRA5	1351	2	A/V	64
G	X	+	50394175	A	X	+	50693925	ENST00000376020	35277	SHRM4	546	2	A/V	64
C	1	+	46821521	G	1	+	47378043	ENST00000371946	538	MKNK1	34	2	G/A	60
C	1	+	26564052	T	1	+	26589352	ENST00000329206	279	ZN683	176	1	A/T	58
A	1	+	43584841	G	1_r.	+	8286746	ENST00000372470	483	TPOR	374	1	T/A	58
T	1	+	118360154	C	1	-	119555017	ENST00000336338	899	SPG17	1415	1	T/A	58
A	11	+	7463757	G	11	+	7318919	ENST00000329293	7779	OLFL1	26	1	T/A	58
C	11	+	18295977	T	11	+	18263443	**ENST00000352460**	**7836**	**HPS5**	2	1	A/T	58
G	12	+	6754050	A	12	+	7000753	ENST00000203629	8561	LAG3	181	1	A/T	58
A	14	+	75319511	G	14	+	75532002	ENST00000298832	32124	TTLL5	958	1	T/A	58
C	15	+	40529603	T	15	+	39548721	ENST00000263805	32208	ZF106	697	1	A/T	58
G	15	+	78960362	A	15	+	78847888	ENST00000356249	10315	K1199	150	1	A/T	58

IDT Exhibit 1061.0205

Base	Chr	+/-	Pos	Base	Chr	+/-	Pos	EnsemblTranscript	CCDS ID	SwissProt	Pos	Base	AA	GS
A	16	+	19637267	G	16	+	19834441	ENST00000320394	10580	IQCK	47	1	T/A	58
T	16	+	65504564	C	16	+	66617266	ENST00000299752	10823	CAD16	342	1	T/A	58
G	17	+	71264629	A	17	+	75364635	**ENST00000200181**	**11727**	**ITB4**	1689	1	A/T	58
G	19	+	54363018	A	19	+	54888107	ENST00000252826	33073	TRPM4	101	1	A/T	58
G	4	+	89627245	A	4	+	91410717	ENST00000264350	3630	HERC5	619	1	A/T	58
G	10	+	102666423	A	10	+	101264246	ENST00000238961	7500	F178A	98	1	E/K	56
T	17	+	24983159	C	17	-	27682885	**ENST00000269033**	**11253**	**SSH2**	1033	1	S/G	56
G	19	+	14671033	A	19	+	15117465	**ENST00000292530**	**12316**	**ZN333**	83	1	E/K	56
C	4	+	5693149	T	4	+	5786383	ENST00000344408	3382	LBN	488	1	G/S	56
A	5	+	176731622	G	5	+	179753998	**ENST00000398128**	**43405**	**RGS14**	549	1	K/E	56
G	8	+	19266005	A	8	+	15599024	ENST00000265807	6009	SH24A	284	1	E/K	56
T	1	+	94337038	G	1	+	95593362	ENST00000370225	747	ABCA4	223	1	K/Q	53
G	21	+	42770559	T	21	+	42171255	ENST00000291536	13688	RSPH1	213	1	Q/K	53
G	7	+	88261633	T	7	+	88409220	ENST00000297203	34678	CG062	187	1	Q/K	53
T	16	+	82720768	C	16	+	84354684	ENST00000219439	10942	HSDL1	260	2	N/S	46
G	19	+	40449692	A	19	+	40767034	ENST00000361790	12450	LSR	424	2	S/N	46
A	19	+	63256998	G	19	+	63932677	ENST00000282326	12969	ZSCA1	332	2	N/S	46
G	20	+	47001360	A	20	+	46380523	ENST00000371917	13411	BIG2	124	2	S/N	46
T	22	+	45019740	C	22	+	45430536	ENST00000314567	33670	CV040	95	2	N/S	46
A	10	+	37548646	G	10	+	38070955	**ENST00000361713**	**7193**	**AN30A**	1278	2	Q/R	43
C	19	+	60400460	T	19	+	60916698	ENST00000376350	33110	PTPRH	609	2	R/Q	43
T	4	+	46431919	C	4	-	85955456	ENST00000396533	3472	CX7B2	16	2	Q/R	43
C	8	+	10506552	T	8_r.	+	5844565	**ENST00000382483**	**43708**	**RP1L1**	822	2	R/Q	43
C	9	+	134267343	T	9	+	132453250	**ENST00000334270**	**6948**	**TTF1**	229	2	R/Q	43
T	9	+	139259701	G	9	+	137500096	ENST00000344774	35186	F166A	134	1	T/P	38
G	1	+	55125322	C	1	+	55868939	ENST00000371269	600	DHC24	20	1	L/V	32
C	11	+	74024946	G	11	+	73005395	ENST00000263681	8233	DPOD3	393	1	L/V	32
C	19	+	11352605	G	19	+	11682327	ENST00000222139	12260	EPOR	261	1	V/L	32
C	22	+	41158219	G	22	+	41499623	ENST00000329021	14034	NFAM1	30	1	V/L	32
A	13	+	49103140	G	13	+	49528651	ENST00000282026	9419	ARL11	186	2	H/R	29
C	14	+	25987939	T	14	+	25373578	ENST00000267422	32061	NOVA1	197	1	V/I	29
C	14	+	104588536	G	14	+	105590622	ENST00000392585	9997	GP132	328	1	E/Q	29
G	15	+	38700151	A	15	+	37608632	ENST00000346991	42023	CASC5	159	2	R/H	29
G	17	+	71264899	A	17	+	75364905	**ENST00000200181**	**11727**	**ITB4**	1748	2	R/H	29
G	19	+	14667458	A	19	+	15113915	**ENST00000292530**	**12316**	**ZN333**	70	2	R/H	29

IDT Exhibit 1061.0206

Base	Chr	+/-	Pos	Base	Chr	+/-	Pos	EnsemblTranscript	CCDS ID	SwissProt	Pos	Base	AA	GS
T	22	+	45183663	C	22	+	45600477	ENST00000262738	14076	CELR1	1707	1	I/V	29
C	3	+	47137662	T	3	+	48170152	ENST00000330022	2749	SETD2	653	2	R/H	29
G	3	+	99466160	A	3	+	102166233	**ENST00000359776**	**33800**	**OR5H6**	115	1	V/I	29
G	5	+	176731601	C	5	+	179753977	**ENST00000398128**	**43405**	**RGS14**	542	1	E/Q	29
A	7	+	134293530	G	7	+	135457096	ENST00000361675	5835	CALD1	671	1	I/V	29
G	7	+	146456810	A	7	+	147715341	ENST00000361727	5889	CNTP2	345	1	V/I	29
T	8	+	19360349	C	8	+	15695152	ENST00000332246	6010	CGAT1	240	1	I/V	29
A	8	+	22076124	G	8	+	18495549	ENST00000318561	43722	PSPC	46	1	I/V	29
G	8	+	145211312	C	8	+	144064957	ENST00000355091	43776	GPAA1	275	1	E/Q	29
C	1	+	156879241	G	1	+	137899626	**ENST00000368148**	**41423**	**SPTA1**	1531	1	A/P	27
C	6	+	2841353	G	6	+	2911117	ENST00000380698	4478	SPB9	80	1	A/P	27
C	17	+	24983383	T	17	-	27682661	**ENST00000269033**	**11253**	**SSH2**	958	2	R/K	26
G	8	+	39683508	A	8	+	36409058	ENST00000265707	6113	ADA18	649	2	R/K	26
G	X	+	153196801	A	X	+	153627491	ENST00000369915	35448	TKTL1	317	2	R/K	26
T	1	+	156914833	C	1	+	137934884	**ENST00000368148**	**41423**	**SPTA1**	265	1	N/D	23
C	11	+	6611344	T	11	+	6490738	ENST00000299441	7771	PCD16	777	1	D/N	23
A	14	+	57932515	G	14	+	57740382	ENST00000360945	9734	TO20L	30	1	N/D	23
C	2	+	231682274	T	2b	+	237323413	ENST00000258400	2483	5HT2B	216	1	D/N	23
A	6	+	160425195	G	6	+	163027524	ENST00000356956	5273	MPRI	2020	1	N/D	23
A	X	+	150843587	G	X	+	151476159	ENST00000393921	14702	MAGA4	266	1	N/D	23
A	11	+	18265762	T	11	+	18233187	**ENST00000352460**	**7836**	**HPS5**	871	2	F/Y	22
C	19	+	3498314	G	19	+	3591210	ENST00000398558	42464	CS028	326	3	L/F	22
G	3	+	198158891	A	3	+	202594007	**ENST00000238138**	**3324**	**PIGZ**	425	1	L/F	22
T	1	+	6622636	C	1	+	6699799	ENST00000377577	87	DJC11	389	1	M/V	21
T	12	+	93975518	C	12	+	96042577	ENST00000393102	9051	NR2C1	242	1	M/V	21
C	16	+	87474655	T	16	+	89288970	ENST00000268679	10972	MTG16	482	1	V/M	21
T	12	+	44607998	C	12	-	43858515	ENST00000369367	8748	SFRIP	584	3	I/M	10
A	20	+	31275866	G	20	+	30217609	ENST00000375454	13216	SPLC3	108	3	I/M	10
A	20	+	32801189	C	20	+	31822768	ENST00000374796	13241	NCOA6	823	3	I/M	10
G	4	+	184423846	T	4	+	187919922	ENST00000281445	34109	WWC2	479	3	M/I	10
C	5	+	54620969	T	5	-	60628686	ENST00000251636	34158	DHX29	317	3	M/I	10
A	11	+	128345808	T	11	+	128028901	ENST00000392657	31718	RICS	1140	3	D/E	0
T	4	+	57471854	A	4	-	73606113	ENST00000309042	3509	REST	98	3	D/E	0

IDT Exhibit 1061.0207

IDT Exhibit 1061.0208

List of Figures

2.1 Schematic representation of the Sanger sequencing process 22

2.2 Pyrosequencing as applied for the 454/Roche sequencer 24

2.3 Reversible terminator chemistry applied by the Illumina Genome Analyzer 26

2.4 Synthesis reaction performed for Illumina paired end sequencing 28

2.5 Life Technologies's SOLiD sequencing-by-ligation 30

2.6 Asynchronous virtual-terminator chemistry performed by the HeliScope 32

2.7 Single Molecule Real Time implemented by Pacific Biosciences 34

2.8 Comparison of high-throughput sequencing technologies available 37

2.9 Differences of paired end sequencing and mate pair sequencing 38

3.1 Flow diagram of steps from DNA sample to sequence read outs with quality score for the Illumina sequencing instruments 42

3.2 Fluorophore read out and fluorophore cross-talk on the Illumina sequencing platform . 43

3.3 Adapter chimeras observed for the Illumina *NlaIII* DGE tag protocol 46

3.4 Adapter chimeras and false positives identified by `TagDust` for an Illumina Multiplex library 47

3.5 Untrimmed adapter sequence at read ends interferes with alignment 49

3.6 Paired-end reads of short-insert libraries ease the correct identification of the adapter start by maximizing autocorrelation of the two reads as well as requiring identical adapter start positions for both reads 50

3.7 Read merging efficiently removes adapter sequence for short insert libraries and increases read accuracy 52

3.8 Illustration of Illumina Multiplex libraries and Double Index libraries 53

3.9 Examples of instrumentation artifacts of the Illumina Genome Analyzer 57

3.10 Reduction of the number of clusters identified in image analysis due to identical sequences 59

3.11 Identification of crystals, dust and lint particles as sequencing clusters 60

3.12 Quality score distributions of artifact reads largely overlaps with quality score distribution of regular reads 61

3.13 Non-random distribution of sequencing error across clusters 62

3.14 Biases in PCR duplicate representation 65

3.15 Estimating library complexity from PCR duplicates 66

4.1 Cross-talk matrix estimate of the Illumina `Bustard` base caller 72

IDT Exhibit 1061.0209

4.2 Intensity values for one tile of a ϕX174 RF1 lane from a 51-cycle Genome Analyzer II run before and after correction by `Bustard` . . . 73

4.3 Analysis of mismatches seen for `Bustard` raw reads and `Ibis` raw reads of a lane with ϕX174 RF1 reads . . . 74

4.4 Determining T accumulation effect from raw intensities . . . 77

4.5 Simulation results for a chemistry with phasing, pre-phasing and T accumulation . . . 78

4.6 Feature space extension for obtaining simple decision boundaries . . . 80

4.7 Multi-class classification using support vector machines . . . 81

4.8 FastQ sequence file format . . . 82

4.9 Comparsion of base caller performance on two Genome Analyzer I lanes (26nt) and one Genome Analyzer II lane (51nt) . . . 84

4.10 Comparison of quality scores for the 51 cycle ϕX control lane data set . . . 86

4.11 Comparsion of error rates on two Genome Analyzer I lanes (26nt) and one Genome Analyzer II lane (51nt) . . . 87

4.12 Base calling performance on a test data set from v2 chemistry . . . 89

4.13 Base calling performance on 76 cycle run using v3 chemistry . . . 90

4.14 Base calling performance on 101 cycle run using v3 chemistry with an improved polymerase . . . 91

4.15 Base calling performance on ϕX spike-in control reads of a 101 cycle run using v4 chemistry . . . 92

4.16 Dependence of training performance on amount of training input data . . . 93

5.1 Sequence artifacts from the flow cell reflection layer . . . 100

5.2 Raw processing summary statistics for DGE expression data set . . . 101

5.3 Observed tag site distribution along genes and neighboring sequence . . . 104

5.4 Correlation of tag counts originating from the two strands of one *NlaIII* restriction site or neighboring *NlaIII* sites . . . 105

5.5 Count distribution across very close *NlaIII* restriction sites . . . 106

5.6 Annotation of expressed sequence tags using species-specific `Ensembl` annotation and human `Ensembl` annotation projected to the other species . . . 107

5.7 Principal component analysis of gene expression counts . . . 109

5.8 False quantification results due to ambiguous tags . . . 110

5.9 Spearman correlation for gene expression quantification results including and excluding ambiguous tag sequences . . . 112

5.10 Number of expressed genes by tissue and `Ensembl` biotype . . . 113

5.11 Triangulation of pair-wise expression differences . . . 116

5.12 Scatter plots and correlations of our DGE five tissues data and Babbitt et al. data using gene raw counts from the paper or counts after reanalysis . . . 119

5.13 PCA of five tissues DGE data set including the Babbitt et al. data using gene raw counts presented in the paper . . . 121

5.14 PCA of five tissues DGE data set including the Babbitt et al. data using gene counts after reanalysis from raw reads . . . 122

5.15 PCA of five tissues DGE data set including the Babbitt et al. data limited to brain samples . . . 123

IDT Exhibit 1061.0210

5.16 Scatter plots and correlation of DGE liver data and Blekhman et al. liver RNAseq data 125

5.17 Scatter plots and correlations of the DGE data set with the reanalyzed Khaitovich et al. array data for human across all five tissues 128

6.1 DNA sequence divergence for ancient and present-day hominins 133

6.2 Geographic distribution of samples studied in the Denisova and Neandertal Genome projects 139

6.3 Single nucleotide changes and insertion/deletion changes on the human lineage and their state in Neandertal 143

6.4 `RNAfold` predictions for two microRNAs showing the ancestral state in Neandertal . . 152

6.5 `RNAfold` predictions for microRNA *hsa-mir-564* showing the ancestral state in Denisova 153

6.6 `RNAfold` predictions for microRNA *hsa-mir-1260* showing the ancestral state in Denisova 153

6.7 Genealogies consistent with Neandertal-Denisova discordant site classes 162

6.8 N_A/D_A for sites obtained from the catalog of changes shows a distinct relationship of present-day Africans and non-Africans to archaic hominins 163

6.9 The N_A/D_A measure for sites obtained from an African catalog of changes confirms the distinct relationship of present day Africans and non-Africans to archaic hominins 164

IDT Exhibit 1061.0211

IDT Exhibit 1061.0212

List of Tables

4.1 Phasing and pre-phasing values determined for ten different Genome Analyzer II runs 76

5.1 Number of *NlaIII* restriction sites in three primate genomes 101

5.2 Gene length of protein-coding genes in `Ensembl v50` and `Ensembl v59` 106

5.3 Results from projecting human `Ensembl v59` gene annotation to chimpanzee and rhesus macaque 108

5.4 Proportion of expression counts originating from unambiguous tag counts 111

5.5 Effect of ambiguous tags on number of expressed genes and the fraction of differentially expressed genes in each tissue 112

5.6 Percentage of genes differentially expressed between species 115

5.7 Checking saturation/clock-like behavior based on the fraction of differentially expressed genes 116

5.8 Checking saturation/clock-like behavior based on average euclidean expression distance 117

5.9 Comparison of brain samples from the five tissues DGE study with the three species cortex data set from Babbitt et al. 119

5.10 Spearman correlation of three different brain data sets 120

5.11 Comparison of the Blekhman et al. liver data set with the five tissues DGE liver data 124

5.12 Spearman correlation of three different liver data sets 125

5.13 Reanalysis of the Khaitovich et al. Affymetrix HG-U133+ 2.0 array data set 127

6.1 InDel changes in coding sequences where Denisova has the ancestral state 149

6.2 Concordance analysis for Neandertal and Denisova at positions changed on the human lineage 156

6.3 Substitution patterns of single nucleotide changes where Neandertal and Denisova disagree on their state 159

6.4 Sequencing error rates inferred for present-day human and ancient DNA data sets . . 160

6.5 N_A and D_A counts from sites ascertained using the San and Mbuti individuals 164

6.6 N_A/D_A statistics from sites ascertained using the San and Mbuti individuals 165

1 Results of phasing and T accumulation simulation of 10,000 clusters, each with a thousand identical sequences 197

2 Single nucleotide changes in coding sequences of CCDS genes for which Neandertal shows the ancestral state 201

3 Single nucleotide changes in coding sequences of CCDS genes for which Denisova shows the ancestral state 204

IDT Exhibit 1061.0213

IDT Exhibit 1061.0214

Curriculum vitae

Name	Martin Kircher
Year / place of birth	1983 / Erfurt, Germany
Nationality	German
School-leaving certificate	General qualifying for university entrance (Grade 1.1[1]) by Spezialschulteil mathematisch-naturwissenschaftlicher Richtung[2] am Albert-Schweitzer-Gymnasium, Erfurt
Studies	04/2006 – 07/2007 *Master of Science (Honor's Degree)* Saarland University, Germany Computational Molecular Biology (Grade 1.3[3])/A-) 10/2003 – 03/2006 *Bachelor of Science* at Saarland University, Germany Computational Molecular Biology (Grade 1.7[3])/B+)
Affiliations	since 09/2007 Janet Kelso, PhD Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany *PhD Student (Bioinformatics, Evolutionary Genetics)* 05/2007 – 07/2007 Professor Dr. Thomas Lengauer, PhD Max Planck Institute for Informatics, Saarbrücken, Germany *Student assistant* 09/2006 – 10/2006 Professor Dr. Andreas Zeller Software Engineering Chair, Saarland University, Germany *Tutor of practical training in software design*

[1] diploma grade from German secondary/high school: 1 (very good/A) to 6 (insufficient/F)
[2] part of secondary/high school with focus on sciences; entrance examination needed
[3] diploma grade from German university: 1 (very good/A) to 5 (insufficient/F)

IDT Exhibit 1061.0215

Affiliations (continued)	08/2005 – 01/2006 Professor Dr. Hans-Peter Lenhof Center for Bioinformatics, Saarland University, Germany *Student assistant*
Visits	09/2010 – 11/2010 Dr. Philipp Khaitovich CAS-MPG Partner institute for Computational Biology, Shanghai, China *Analysis of Illumina Digital Gene Expression data*
	11/2009 Dr. Philipp Khaitovich CAS-MPG Partner institute for Computational Biology, Shanghai, China *Consultant for high-throughput sequencing*
	03/2008 Dr. Dirk Schübeler Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland *Measuring DNA methylation using MeDIP*
	08/2005 – 09/2005 Professor Dr. scient. Inge Jonassen Computational Biology Unit of Bergen Center for Computational Science, Bergen, Norway *Web application for analyzing expression data*
Honors and distinctions	Doctoral Scholarship (2007) International Max Planck Research School of Human Origins, Leipzig, Germany
	Scholarship from Sparkasse[1] Erfurt (2003) objective: promotion in natural and computer sciences
	Software Award of the Thuringian Minister of Education (2002)
	Albert Schweitzer School Award in the field of mathematics and computer sciences (2002)
	Second Award in the national competition (Thuringia) of youth research ("Jugend forscht") in the field of geography and space sciences and a special award in software systems technology (2002)
	Second Award of the 5th federal physics competition (1999)
Language skills	German (native language) English (very good speaking and writing skills) French (basic skills)

[1] German bank and credit institute

IDT Exhibit 1061.0216

Lists of publications and talks

List of accepted journal articles

2010 D. Reich, R.E. Green, **M. Kircher**[1], J. Krause, N. Patterson, E.Y. Durand, B. Viola, A.W. Briggs, U. Stenzel, P.L.F. Johnson, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–1060, 2010.

M. Kircher and J. Kelso. High-throughput DNA sequencing-concepts and limitations. *BioEssays*, 32(6):524–536, 2010.

P. Heyn, U. Stenzel, A.W. Briggs, **M. Kircher**, M. Hofreiter, and M. Meyer. Road blocks on paleogenomes – polymerase extension profiling reveals the frequency of blocking lesions in ancient DNA. *Nucleic Acids Research*, 2010.

M. Meyer and **M. Kircher**. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols*, 2010(6), 2010.

R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, **M. Kircher**[2], N. Patterson, et al. A draft sequence of the Neandertal genome. *Science*, 328(5979):710–22, 2010.

2009 A. W. Briggs, U. Stenzel, M. Meyer, J. Krause, **M. Kircher**, and S. Pääbo. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Research*, 2009.

J. Krause, A. W. Briggs, **M. Kircher**, T. Maricic, N. Zwyns, A. Derevianko, and S. Pääbo. A complete mtDNA genome of an early modern human from Kostenki, Russia. *Current Biology*, 2009.

M. Kircher, U. Stenzel, and J. Kelso. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biology*, 10(8):R83, 2009.

2008 **M. Kircher**, C. Bock, and M. Paulsen. Structural conservation versus functional divergence of maternally expressed microRNAs in the Dlk1/Gtl2 imprinting region. *BMC Genomics*. 2008 Jul 23;9:346.

[1]Shared first author
[2]Shared second author

IDT Exhibit 1061.0217

Publications in preparation

M. Kircher[1], S. Sawyer, and M. Meyer. *Double indexing improves the scope and accuracy of multiplex sequencing on the Illumina platform.* Journal article.

M. Kircher, P. Heyn, and J. Kelso. *Computational challenges in the production and analysis of Illumina Genome Analyzer data.* Journal article.

M. Kircher. *Analysis of high-throughput ancient DNA sequencing data.* Book chapter in "Methods in Molecular Biology: Ancient DNA" edited by Beth Shapiro and Michael Hofreiter

List of posters

2010 **M. Kircher**, E. Lizano, T. Giger, S. Pääbo, and J. Kelso. Challenges in the comparative analysis of gene expression in apes using Illumina Digital Gene Expression. *Genome Informatics.* Hinxton, Cambridge, UK

J. Kelso and **M. Kircher**. High-throughput DNA sequencing – concepts and limitations. *18th Annual International Conference on Intelligent Systems for Molecular Biology.* Boston, USA

M. Kircher, U. Stenzel, and J. Kelso. Increasing the Genome Analyzer's output using IBIS. *18th Annual International Conference on Intelligent Systems for Molecular Biology.* Boston, USA

U. Stenzel, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, T. Maričić, M. Meyer, J. Kelso, and **M. Kircher**. Increasing the Genome Analyzer's output using IBIS – with or without a dedicated control lane. *2010 Europe User Symposium.* Sitges, Barcelona, Spain

M. Kircher and J. Kelso. Illumina Genome Analyzer: Artifacts and good analysis practice. *The Biology of Genomes.* Cold Spring Harbor, New York, USA

2009 **M. Kircher**, U. Stenzel, and J. Kelso. IBIS — Improved base calling for the Illumina Genome Analyzer. *German Conference on Bioinformatics.* Halle (Saale), Germany

M. Kircher, U. Stenzel, and J. Kelso. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *17th Annual International Conference on Intelligent Systems for Molecular Biology & 8th European Conference on Computational Biology.* Stockholm, Sweden

M. Kircher, U. Stenzel, and J. Kelso. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *The Biology of Genomes.* Cold Spring Harbor, New York, USA

[1]Shared first author

IDT Exhibit 1061.0218

List of scientific talks

2011 High throughput sequencing – concepts and limitations.... *5th International Symposium on the Biology and Immunology of Cutaneous Lymphoma 2011.* January 14 2011, Berlin, Germany

2010 Applying high-throughput sequencing to ancient and modern DNA samples for studying the genetic history of humankind. *Illumina Inc.* September 14 2010, Little Chesterford, Cambridge, UK

Studying Modern Human Origins from Neandertal DNA. *Illumina Next Generation Seminar.* September 9 2010, Berlin, Germany

Applying high-throughput sequencing to ancient and modern DNA samples. *Leipzig University, Faculty of Biosciences, Pharmacy and Psychology, Prof. Dr. Mario Mörl.* June 11 2010, Leipzig, Germany

MPI EVA: High-throughput sequencing of ancient and modern DNA samples. *The First Galaxy Developer Conference.* May 16 2010, Cold Spring Harbor, New York, USA

2009 One and a half year of Illumina processing pipelines. *Lab seminar, Department of Evolutionary Genetics, Max Planck for Evolutionary Anthropology.* December 3 2009, Leipzig, Germany

Improving Illumina Genome Analyzer data quality by alternative base calling. *Leipzig University, Interdisciplinary Centre for Bioinformatics, PD Dr. habil. Hans Binder.* December 1 2009, Leipzig, Germany

Improving data quality of the Illumina Genome Analyzer platform. *Bioinformatics Autumn Seminar, Chair for Bioinformatics Leipzig University*, October 24 2009, Vysoká Lípa, Děčín, Czech Republic

Applying high-throughput sequencing to ancient and modern DNA samples for studying the genetic history of humankind. *Genomics with Illumina: Arrays and Next Generation Sequencing.* September 22 2009, Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung, Gatersleben, Germany

High throughput sequencing. *Leipzig University, Faculty of Biosciences, Pharmacy and Psychology, Prof. Dr. Mario Mörl.* June 5 2009, Leipzig, Germany

High throughput sequencing – concepts and limitations. *Deutsches Rheuma-Forschungszentrum Berlin.* May 26 2009, Berlin, Germany

Improving Illumina base calling using Statistical Learners *Lab seminar, Department of Evolutionary Genetics, Max Planck for Evolutionary Anthropology.* January 29 2009, Leipzig, Germany

2008 Illumina sequencing – A technology growing up.... *Saarland University, Faculty Natural Sciences and Technology III, Genetics / Epigenetics, Prof. Dr. Jörn Walter.* December 12 2008, Saarbrücken, Germany

Solexa sequencing – a technology growing up.... *Bioinformatics Autumn Seminar, Chair for Bioinformatics Leipzig University.* November 3 2008 Česká Kamenice, Děčín, Czech Republic

IDT Exhibit 1061.0219

2007 Coin flipping for selection - finding regions of positive selection in Human using the Neanderthal genome. *Lab seminar, Department of Evolutionary Genetics, Max Planck for Evolutionary Anthropology.* December 20 2007, Leipzig, Germany

Attendence of scientific workshops

- Workshop "Building Next Generation Sequencing platforms and pipeline solutions", November 18-20 2009, Rome, Italy
- Workshop on Molecular Evolution (Europe), January 12-23 2009, Český Krumlov, Czech Republic
- Autumn School on Epigenetics, 2007, Humboldt University & Charité, Berlin, Germany

IDT Exhibit 1061.0220

Abbreviations

A/C/G/T	Deoxyadenosine, deoxycytosine, deoxyguanosine, deoxythymidine
AIC	Akaike information criterion
ATP	Adenosine triphosphate
dATP	Deoxyadenosine triphosphate
dATPαS	Deoxy-adenosine-5'-(alpha-thio)-triphosphate
bp	base pair(s)
CAE	Capillary Array Electrophoresis
CAS	Chinese Academy of Sciences
cDNA	complementary DNA synthesized from a mRNA template
CEPH	Centre d'Etude du Polymorphisme Humain
CCD	Charge-Coupled Device (semi-conductor device used in digital cameras)
ChIP-Seq	Chromatin Immuno-Precipitation sequencing
CNV	Copy Number Variation
DGE	Digital Gene Expression
DNA	DeoxyriboNucleic acid
dNTPs/NTPs	deoxy-nucleotides
ddNTPs	dideoxy-nucleotides (modified nucleotides missing a hydroxyl group at the third carbon atom of the sugar)
EndoVIII	Escherichia coli endonuclease VIII
FDR	False Discovery Rate
FRET	Fluorescence Resonance Energy Transfer
GA	Short for Illumina Genome Analyzer
GC	Frequency of Guanine and Cytosine nucleobases
HAR	Human Accelerated Region
HGDP	Human Genome Diversity Panel
InDel	Insertion/Deletion
IPAR	Integrated Primary Analysis and Reporting (Illumina software)
IUPAC	International Union of Pure and Applied Chemistry
kb/Mb/Gb	kilo base (10^3 nt) / mega base (10^6 nt) / giga base (10^9 nt)
MeDIP-Seq	Methylation-Dependent Immuno-Precipitation sequencing
miRNA	microRNA
MPG	Max Planck Society (*German: Max-Planck-Gesellschaft*)
MPI	Max Planck Institute
mRNA	messenger RNA/transcripts
mtDNA	mitochondrial DNA
nt	nucleotide(s)
OLB	Off-Line Basecaller (Illumina software)
PCA	Principal Component Analysis
PEC	Primer Extension Capture
PCR	Polymerase Chain Reaction

IDT Exhibit 1061.0221

qPCR	quantitative Polymerase Chain Reaction
QC	Quality control
RMSD	Root Mean Square Deviation
RNA	RiboNucleic acid
RNAseq	Sequencing of mRNAs/transcripts
RTA	Real Time Analysis (Illumina software)
SAGE	Serial Analysis of Gene Expression
SCS	Sequencing Control Software (Illumina software)
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
SNC	Single Nucleotide Change (on the human lineage)
SVM	Support Vector Machine
TCEP	tris-(2-carboxyethyl)-phosphine)
UDG	Uracil-DNA Glycosylase
UTR	Untranslated region of mRNA

IDT Exhibit 1061.0222

Index

ϕX174, 21, 63, 71, 94, 135, 136
3' UTRs, *see* Untranslated region
454 Life Sciences, 23
454 sequencing, 23, 35, 36, 134, 169
5' UTRs, *see* Untranslated region

ABI, *see* Applied Biosystems
ABI Prism, 21
Admixture, 157, 166
Amplification bias, 64
ancient DNA, 131
Applied Biosystems, 21, 29
Artifact, 25, 44, 61
Asynchronous sequencing, 32

Background error rate, *see* Error rate
Barcode, 38, 44, 53
BASE, 35
Base calling, 43, 70
Beads, 24, 31
    Mixed, 25
Block Jackknife, 163
Bridge amplification, 27, 41
Burrows-Wheeler transformed alignment, *see* BWA
BWA, 45, 138, 141

Capillary Array Electrophoresis (CAE), *see* Sanger sequencing
Catalog of changes, 141
cBot, *see* Cluster station
CCDS, 144
Centre d'Etude du Polymorphisme Humain, 139
CEPH, *see* Centre d'Etude du Polymorphisme Humain
Chastity filter, 62
CIF, *see* Cluster Intensity Format
Cluster, *see* Sequence cluster
Cluster Intensity Format, 71, 136
Cluster station, 58
Color-space encoding, 29
Complete Genomics, 29
Cross-talk matrix, 71

ddNTP, 22
Denisovans, 132, 136, 156, 166
DGE, *see* Digital Gene Expression
Digital Gene Expression, 45, 98, 173
DNA sequencing, *see* Sequencing
dNTP, 22

ELAND, 45, 102
Electrophoresis, 22
Emulsion PCR, 24
Endoplasmic reticulum, 144
Ensemble, 25, 33
Error rate
    Background, 25
    Sequencing, 23, 25, 27, 31, 33, 51, 62

Firecrest, 58, 70
Flatness, 56
Flow cell, 26, 32, 35, 42, 44, 56
Flow cell tilt, 56
Flow cycle, 25
Fluorescence Resonance Energy Transfer, 36
Footprint, 56, 58
Forked adapters, 26
FRET, *see* Fluorescence Resonance Energy Transfer

GATK, *see* Genome Analysis Tool Kit
GE MegaBACE, 21
General Electrics (GE) Healthcare, 21
Genome Analysis Tool Kit, 64, 95
Genome Analyzer, 26, 36, 41, 69, 98, 135, 169, 172, 173
Ghost well, 25
GS FLX, *see* 454 sequencing
GS Junior, 35

HAR, *see* Human accelerated region
Helicos, 31
HeliScope, 31, 36
HGDP, *see* Human Genome Diversity Panel
HiSeq2000, 35

IDT Exhibit 1061.0223

Homopolymer, 23
Human accelerated region, 149, 154, 155
Human Genome Diversity Panel, 139

Ibis, 70, 172
IBM, 35
Illumina, *see* Genome Analyzer
Image analysis, 43, 70
In vitro amplification, 23, 27
In vivo amplification, 23
InDels, 25, 142
Index, *see* Barcode
Integrated Primary Analysis and Reporting, 59, 70
Intensities, 43
Intensity files, 71
Ion Torrent, 35
IPAR, *see* Integrated Primary Analysis and Reporting

Life Technologies, 29, 35

Machine learning, 75
Mate pair sequencing, 37, 170
Messenger RNA, 98, 151
microRNA, 151
miRNA, *see* microRNA
Mitochondria, 131, 148
mitochondrial DNA, 132, 167
Mixed beads, 25
mRNA, *see* Messenger RNA
mtDNA, *see* mitochondrial DNA
MySeq, 35

Nanopores
    BASE, 35
    Silicon-based, 35
Neandertals, 132, 134, 156, 166
Non-reversible termination, 22
Nucleotide modification, 35

Off-Line Basecaller, 58, 70
OLB, *see* Off-Line Basecaller
Oxford Nanopore, 35

Pacific Biosciences, 33, 36
Paired End Module, 27
Paired end sequencing, 27, 35, 37, 170
Pass Filter flag, 62
PCA, *see* Principal component analysis
PCR duplicates, 64
PEC, *see* Primer Extension Capture
Personal Genome Maschine, 35
PF, *see* Pass Filter
PGM, *see* Personal Genome Machine
Phasing, 25, 27, 31, 44, 53, 71, 75, 172
    bi-directional, 27
    unidirectional, 27
phiX 174, *see* ϕX174
Picotiter plate, 24
Polonator, 29
Pre-phasing, *see* Phasing
Primer Extension Capture, 134
Principal component analysis, 109
Pyrosequencing, 23

qPCR, 59

Real Time Analysis, 58, 70, 136
Restriction enzyme enrichment, 135
Reversible terminator chemistry, 26
RNAseq, 98, 123, 177
Roche Diagnostics, 24
Roche sequencing, *see* 454 sequencing
RTA, *see* Real Time Analysis

SAGE, *see* Serial Analysis of Gene Expression
Sanger sequencing, 22, 36, 169
Sequence cluster, 27, 41, 58, 61
Sequencing, 21, 41, 99, 134
Sequencing by synthesis, 23
Sequencing error rate, *see* Error rate
Sequencing library, 24, 26, 42, 44, 134
Sequencing-by-ligation, 29
Serial Analysis of Gene Expression, 98, 173
Signal peptide, 144
Simple repeat, 23
Single Molecule Real Time, 33, 36
Single nucleotide polymorphism, 37, 170
Single-molecule sequencing, 31, 35, 36
SMRT, *see* Single Molecule Real Time
SMS, *see* Single-molecule sequencing
SNC, 141
SNP, *see* Single nucleotide polymorphism
Solexa sequencing, *see* Genome Analyzer
SOLiD sequencing, 29, 35, 36, 169
Stage tilt, 58
Statistical learner, 75

T accumulation, 72, 75, 135
Tilt, 56, 58
Titanium, *see* 454 sequencing

IDT Exhibit 1061.0224

Untranslated region, 103, 105, 149, 150
UTR, *see* Untranslated region

Virtual-terminator chemistry, 32

Well (picotiter plate), 25

Zero-mode waveguides, 33
ZMW, *see* Zero-mode waveguides

IDT Exhibit 1061.0225