Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

ANNUAL REVIEWS **Further**

Click here for quick links to Annual Reviews content online, including:

- Other articles in this volume
- Top cited articles
- Top downloaded articles
- Our comprehensive search

Next-Generation DNA Sequencing Methods

Elaine R. Mardis

Departments of Genetics and Molecular Microbiology and Genome Sequencing Center, Washington University School of Medicine, St. Louis MO 63108; email: emardis@wustl.edu

Annu. Rev. Genomics Hum. Genet. 2008. 9:387–402

First published online as a Review in Advance on June 24, 2008

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi: 10.1146/annurev.genom.9.081307.164359

Copyright © 2008 by Annual Reviews. All rights reserved

1527-8204/08/0922-0387$20.00

Key Words

massively parallel sequencing, sequencing-by-synthesis, resequencing

Abstract

Recent scientific discoveries that resulted from the application of next-generation DNA sequencing technologies highlight the striking impact of these massively parallel platforms on genetics. These new methods have expanded previously focused readouts from a variety of DNA preparation protocols to a genome-wide scale and have fine-tuned their resolution to single base precision. The sequencing of RNA also has transitioned and now includes full-length cDNA analyses, serial analysis of gene expression (SAGE)-based methods, and noncoding RNA discovery. Next-generation sequencing has also enabled novel applications such as the sequencing of ancient DNA samples, and has substantially widened the scope of metagenomic analysis of environmentally derived samples. Taken together, an astounding potential exists for these technologies to bring enormous change in genetic and biological research and to enhance our fundamental biological knowledge.

NeoGenomics
Exhibit 1104 Page 1

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

INTRODUCTION

The sequencing of the reference human genome was the capstone for many years of hard work spent developing high-throughput, high-capacity production DNA sequencing and associated sequence finishing pipelines. The approach used >20,000 large bacterial artificial chromosome (BAC) clones that each contained an approximately 100-kb fragment of the human genome, which together provided an overlapping set or tiling path through each human chromosome as determined by physical mapping (31). In BAC-based sequencing, each BAC clone is amplified in bacterial culture, isolated in large quantities, and sheared to produce size-selected pieces of approximately 2–3 kb. These pieces are subcloned into plasmid vectors, amplified in bacterial culture, and the DNA is selectively isolated prior to sequencing. By generating approximately eightfold oversampling (coverage) of each BAC clone in plasmid subclone equivalents, computer-aided assembly can largely recreate the BAC insert sequence in contigs (contiguous stretches of assembled sequence reads). Subsequent refinement, including gap closure and sequence quality improvement (finishing), produces a single contiguous stretch of high-quality sequence (typically with less than 1 error per 40,000 bases). Since the completion of the human genome project (HGP) (26, 51), substantive changes have occurred in the approach to genome sequencing that have moved away from BAC-based approaches and toward whole-genome sequencing (WGS), with changes in the accompanying assembly algorithms. In the WGS approach, the genomic DNA is sheared directly into several distinct size classes and placed into plasmid and fosmid subclones. Oversampling the ends of these subclones to generate paired-end sequencing reads provides the necessary linking information to fuel whole-genome assembly algorithms. The net result is that genomes can be sequenced more rapidly and more readily, but highly polymorphic or highly repetitive genomes remain quite fragmented after assembly.

Despite these dramatic changes in sequencing and assembly approaches, the primary data production for most genome sequencing since the HGP has relied on the same type of capillary sequencing instruments as for the HGP. However, that scenario is rapidly changing owing to the invention and commercial introduction of several revolutionary approaches to DNA sequencing, the so-called next-generation sequencing technologies. Although these instruments only began to become commercially available in 2004, they already are having a major impact on our ability to explore and answer genome-wide biological questions; more than 100 next-generation sequencing–related manuscripts have appeared to date in the peer-reviewed literature. These technologies are not only changing our genome sequencing approaches and the associated timelines and costs, but also accelerating and altering a wide variety of types of biological inquiry that have historically used a sequencing-based readout, or effecting a transition to this type of readout, as detailed in this review. Furthermore, next-generation platforms are helping to open entirely new areas of biological inquiry, including the investigation of ancient genomes, the characterization of ecological diversity, and the identification of unknown etiologic agents.

NEXT-GENERATION DNA SEQUENCING

Three platforms for massively parallel DNA sequencing read production are in reasonably widespread use at present: the Roche/454 FLX (30) (**http://www.454.com/enabling-technology/the-system.asp**), the Illumina/Solexa Genome Analyzer (7) (**http://www.illumina.com/pages.ilmn?ID=203**), and the Applied Biosystems SOLiD™ System (**http://marketing.appliedbiosystems.com/images/Product/Solid_Knowledge/flash/102207/solid.html**). Recently, another two massively parallel systems were announced: the Helicos Heliscope™ (**www.helicosbio.com**) and Pacific Biosciences SMRT (**www.pacificbiosciences.com**) instruments. The

NeoGenomics
Exhibit 1104 Page 2

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

Helicos system only recently became commercially available, and the Pacific Biosciences instrument will likely launch commercially in early 2010. Each platform embodies a complex interplay of enzymology, chemistry, high-resolution optics, hardware, and software engineering. These instruments allow highly streamlined sample preparation steps prior to DNA sequencing, which provides a significant time savings and a minimal requirement for associated equipment in comparison to the highly automated, multistep pipelines necessary for clone-based high-throughput sequencing. By different approaches outlined below, each technology seeks to amplify single strands of a fragment library and perform sequencing reactions on the amplified strands. The fragment libraries are obtained by annealing platform-specific linkers to blunt-ended fragments generated directly from a genome or DNA source of interest. Because the presence of adapter sequences means that the molecules then can be selectively amplified by PCR, no bacterial cloning step is required to amplify the genomic fragment in a bacterial intermediate as is done in traditional sequencing approaches. Importantly, both the Helicos and Pacific Biosystems instruments mentioned above are so-called "single molecule" sequencers and do not require any amplification of DNA fragments prior to sequencing.

Another contrast between these instruments and capillary platforms is the run time required to generate data. Next-generation sequencers require longer run times of between 8 h and 10 days, depending upon the platform and read type (single end or paired ends). The longer run times result mainly from the need to image sequencing reactions that are occurring in a massively parallel fashion, rather than a periodic charge-coupled device (CCD) snapshot of 96 fixed capillaries. The yield of sequence reads and total bases per instrument run is significantly higher than the 96 reads of up to 750 bp each produced by a capillary sequencer run, and can vary from several hundred thousand reads (Roche/454) to tens of millions of reads (Illumina and Applied Biosystems SOLiD). The combination of streamlined sample preparation and long run times means that a single operator can readily keep several next-generation sequencing instruments at full capacity. The following sections aim to introduce the reader to the primary features of each of the three most widely used next-generation platforms and to discuss strengths and weaknesses.

Charge-coupled device (CCD): a capacitor array used in optical scanners to capture images

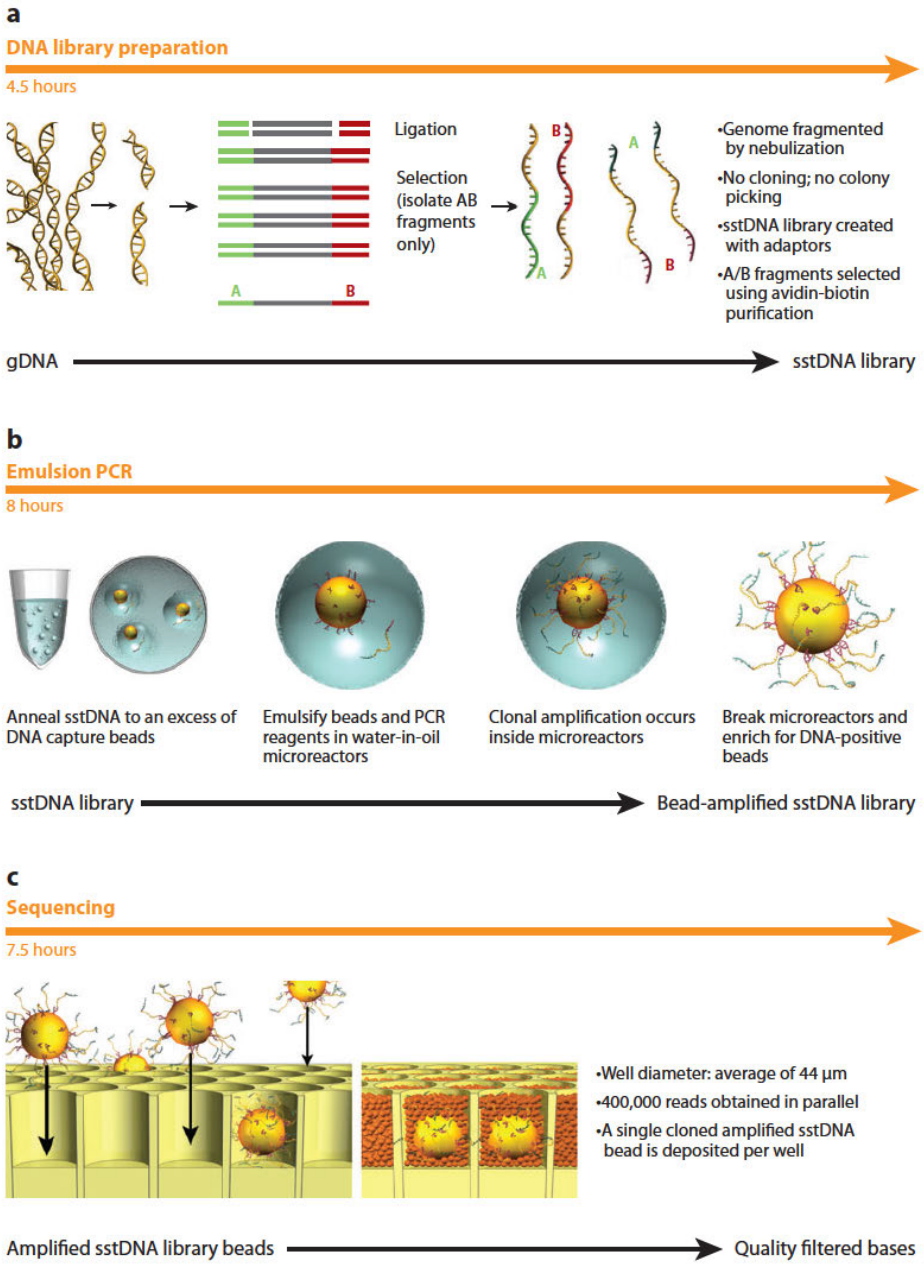
Emulsion PCR (ePCR): method for DNA amplification that uses a water in oil emulsion to isolate single DNA molecules in aqueous microreactors

Roche/454 FLX Pyrosequencer

This next-generation sequencer was the first to achieve commercial introduction (in 2004) and uses an alternative sequencing technology known as pyrosequencing. In pyrosequencing, each incorporation of a nucleotide by DNA polymerase results in the release of pyrophosphate, which initiates a series of downstream reactions that ultimately produce light by the firefly enzyme luciferase. The amount of light produced is proportional to the number of nucleotides incorporated (up to the point of detector saturation). In the Roche/454 approach (**Figure 1**), the library fragments are mixed with a population of agarose beads whose surfaces carry oligonucleotides complementary to the 454-specific adapter sequences on the fragment library, so each bead is associated with a single fragment. Each of these fragment:bead complexes is isolated into individual oil:water micelles that also contain PCR reactants, and thermal cycling (emulsion PCR) of the micelles produces approximately one million copies of each DNA fragment on the surface of each bead. These amplified single molecules are then sequenced en masse. First the beads are arrayed into a picotiter plate (PTP; a fused silica capillary structure) that holds a single bead in each of several hundred thousand single wells, which provides a fixed location at which each sequencing reaction can be monitored. Enzyme-containing beads that catalyze the downstream pyrosequencing reaction steps are then added to the PTP and the mixture is centrifuged to surround the agarose beads. On instrument, the PTP acts as a flow cell into which each pure nucleotide solution is introduced in a stepwise fashion, with an imaging step after each

NeoGenomics
Exhibit 1104 Page 3

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.



Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

nucleotide incorporation step. The PTP is seated opposite a CCD camera that records the light emitted at each bead. The first four nucleotides (TCGA) on the adapter fragment adjacent to the sequencing primer added in library construction correspond to the sequential flow of nucleotides into the flow cell. This strategy allows the 454 base-calling software to calibrate the light emitted by a single nucleotide incorporation. However, the calibrated base calling cannot properly interpret long stretches (>6) of the same nucleotide (homopolymer run), so these areas are prone to base insertion and deletion errors during base calling. By contrast, because each incorporation step is nucleotide specific, substitution errors are rarely encountered in Roche/454 sequence reads.

The FLX instrument currently provides 100 flows of each nucleotide during an 8-h run, which produces an average read length of 250 nucleotides (an average of 2.5 bases per flow are incorporated). These raw reads are processed by the 454 analysis software and then screened by various quality filters to remove poor-quality sequences, mixed sequences (more than one initial DNA fragment per bead), and sequences without the initiating TCGA sequence. The resulting reads yield 100 Mb of quality data on average. Downstream of read processing, an assembly algorithm (Newbler) can assemble FLX reads. Although shorter than reads derived from capillary sequencers, FLX reads are of sufficient length to assemble small genomes such as bacterial and viral genomes to high quality and contiguity. As mentioned, the lack of a bacterial cloning step in the Roche/454 process means that sequences not typically sampled in a WGS approach owing to cloning bias will be more likely represented in a FLX data set, which contributes to more comprehensive genome coverage.

Illumina Genome Analyzer

The single molecule amplification step for the Illumina Genome Analyzer starts with an Illumina-specific adapter library, takes place on the oligo-derivatized surface of a flow cell, and is performed by an automated device called a Cluster Station. The flow cell is an 8-channel sealed glass microfabricated device that allows bridge amplification of fragments on its surface, and uses DNA polymerase to produce multiple DNA copies, or clusters, that each represent the single molecule that initiated the cluster amplification. A separate library can be added to each of the eight channels, or the same library can be used in all eight, or combinations thereof. Each cluster contains approximately one million copies of the original fragment, which is sufficient for reporting incorporated bases at the required signal intensity for detection during sequencing.

Bridge amplification: allows the generation of in situ copies of a specific DNA molecule on an oligo-decorated solid support

The Illumina system utilizes a sequencing-by-synthesis approach in which all four nucleotides are added simultaneously to the flow cell channels, along with DNA polymerase, for incorporation into the oligo-primed cluster fragments (see **Figure 2** for details). Specifically, the nucleotides carry a base-unique fluorescent label and the 3′-OH group is chemically blocked such that each incorporation is a unique event. An imaging step follows each base incorporation step, during which each flow cell lane is imaged in three 100-tile segments by the instrument optics at a cluster density per tile of 30,000. After each imaging step, the 3′ blocking group is chemically removed

←

Figure 1

The method used by the Roche/454 sequencer to amplify single-stranded DNA copies from a fragment library on agarose beads. A mixture of DNA fragments with agarose beads containing complementary oligonucleotides to the adapters at the fragment ends are mixed in an approximately 1:1 ratio. The mixture is encapsulated by vigorous vortexing into aqueous micelles that contain PCR reactants surrounded by oil, and pipetted into a 96-well microtiter plate for PCR amplification. The resulting beads are decorated with approximately 1 million copies of the original single-stranded fragment, which provides sufficient signal strength during the pyrosequencing reaction that follows to detect and record nucleotide incorporation events. sstDNA, single-stranded template DNA.

NeoGenomics
Exhibit 1104 Page 5

to prepare each strand for the next incorporation by DNA polymerase. This series of steps continues for a specific number of cycles, as determined by user-defined instrument settings, which permits discrete read lengths of 25–35 bases. A base-calling algorithm assigns sequences and associated quality values to each read and a quality checking pipeline evaluates the Illumina data from each run, removing poor-quality sequences.

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

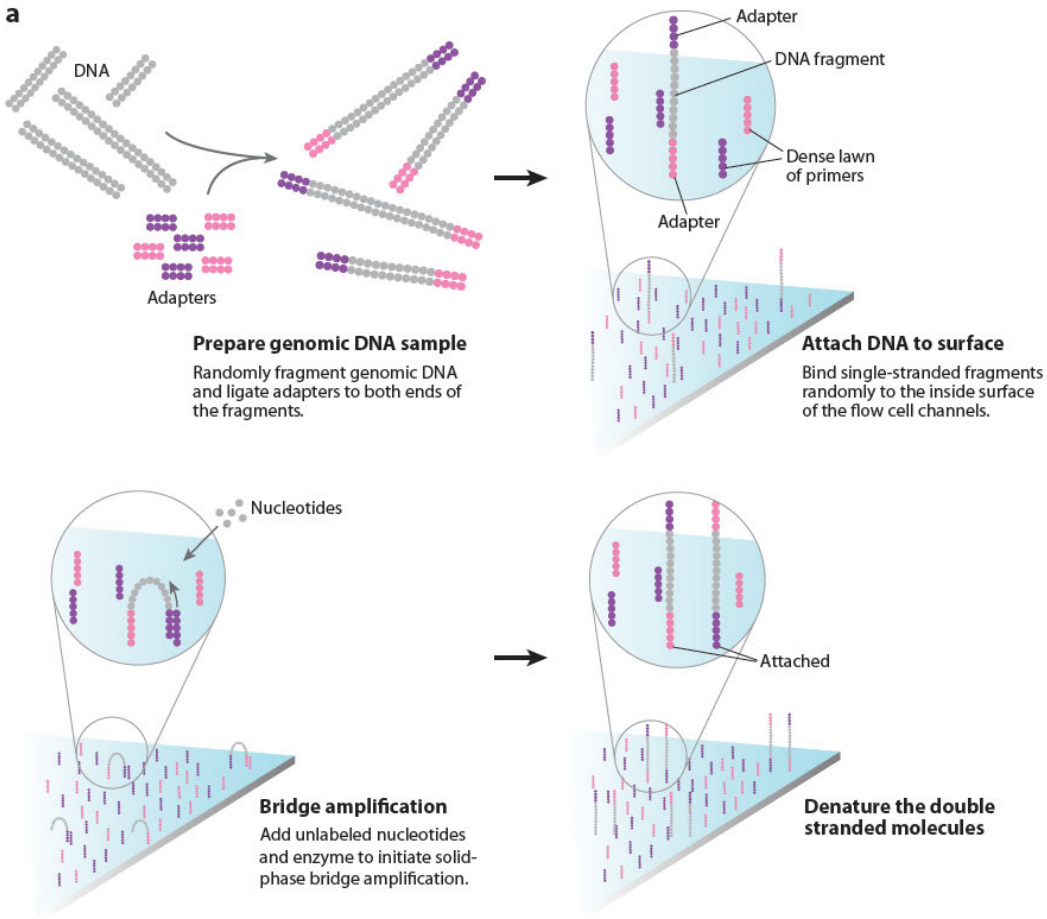


Figure 2

The Illumina sequencing-by-synthesis approach. Cluster strands created by bridge amplification are primed and all four fluorescently labeled, 3′-OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation step, the unused nucleotides and DNA polymerase molecules are washed away, a scan buffer is added to the flow cell, and the optics system scans each lane of the flow cell by imaging units called tiles. Once imaging is completed, chemicals that effect cleavage of the fluorescent labels and the 3′-OH blocking groups are added to the flow cell, which prepares the cluster strands for another round of fluorescent nucleotide incorporation.

NeoGenomics
Exhibit 1104 Page 6

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

Applied Biosystems SOLiD™ Sequencer

The SOLiD platform uses an adapter-ligated fragment library similar to those of the other next-generation platforms, and uses an emulsion PCR approach with small magnetic beads to amplify the fragments for sequencing. Unlike the other platforms, SOLiD uses DNA ligase and a unique approach to sequence the amplified fragments, as illustrated in **Figure 3*a***. Two flow cells are processed per instrument run, each of which can be divided to contain different libraries in up to four quadrants. Read lengths for SOLiD are user defined between 25–35 bp, and each sequencing run yields between 2–4 Gb of DNA sequence data. Once the reads are base called, have quality values, and low-quality sequences have been removed, the reads are aligned to a reference genome to enable a second tier of quality evaluation called two-base encoding. The principle of two-base encoding is shown in **Figure 3*b***, which illustrates how this approach works to differentiate true single base variants from base-calling errors.

Two key differences that speak to the utility of next-generation sequence reads are (*a*) the length of a sequence read from all current next-generation platforms is much shorter than that from a capillary sequencer and (*b*) each next-generation read type has a unique error model different from that already established for

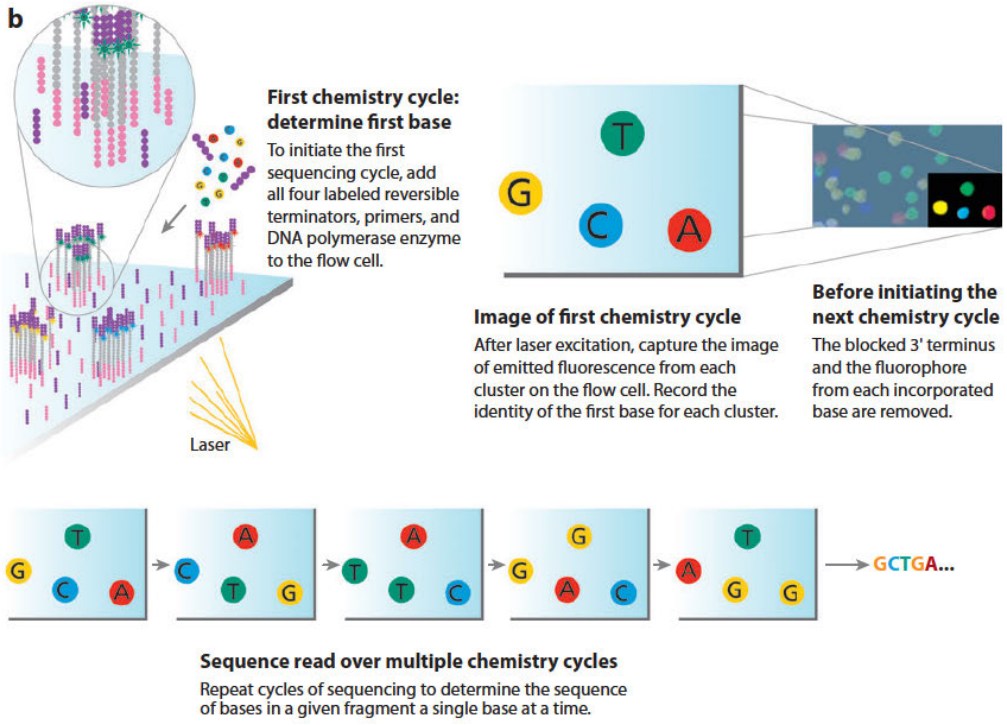
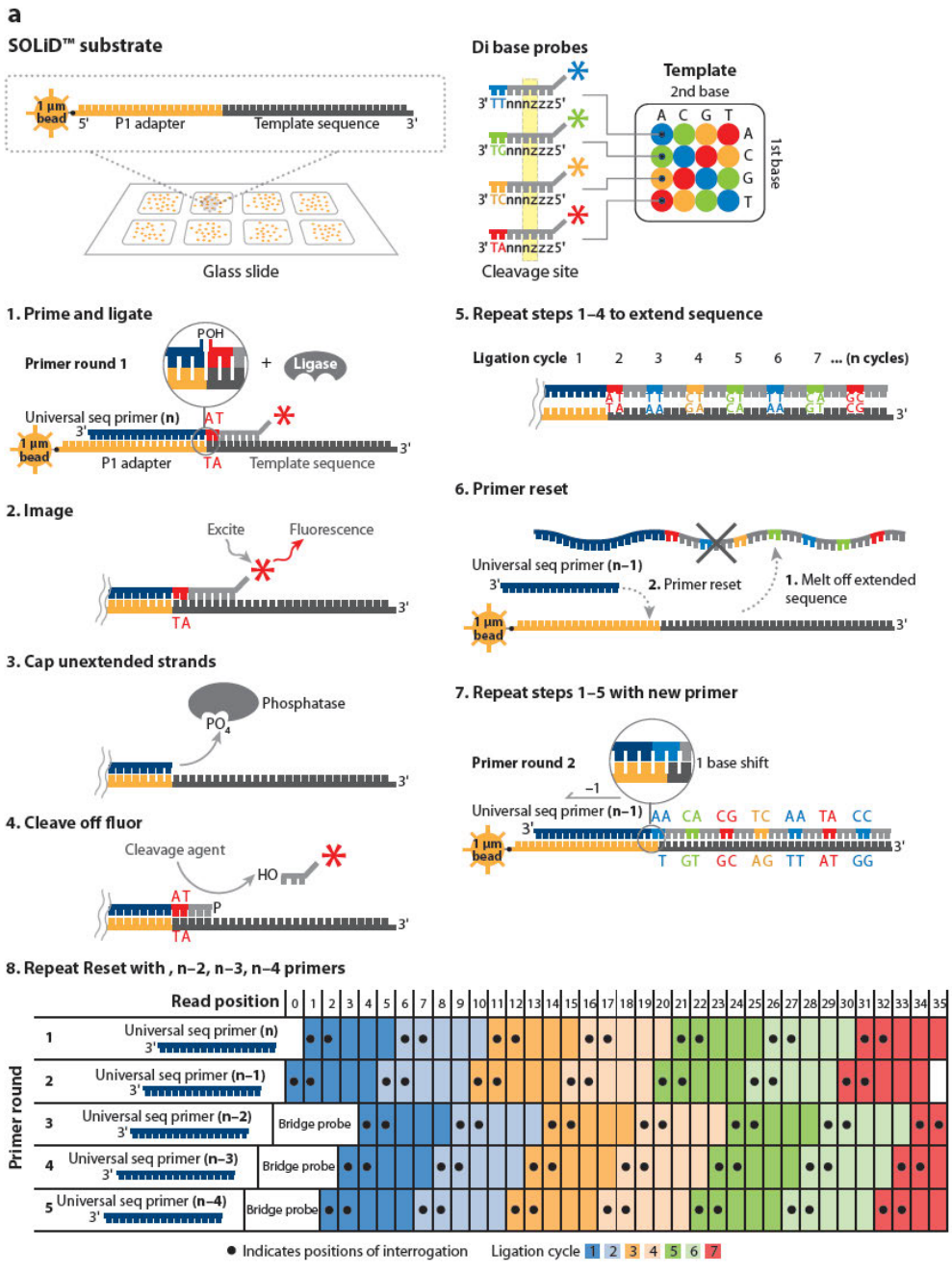


Figure 2

(*Continued*)

NeoGenomics
Exhibit 1104 Page 7

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.



NeoGenomics
Exhibit 1104 Page 8

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

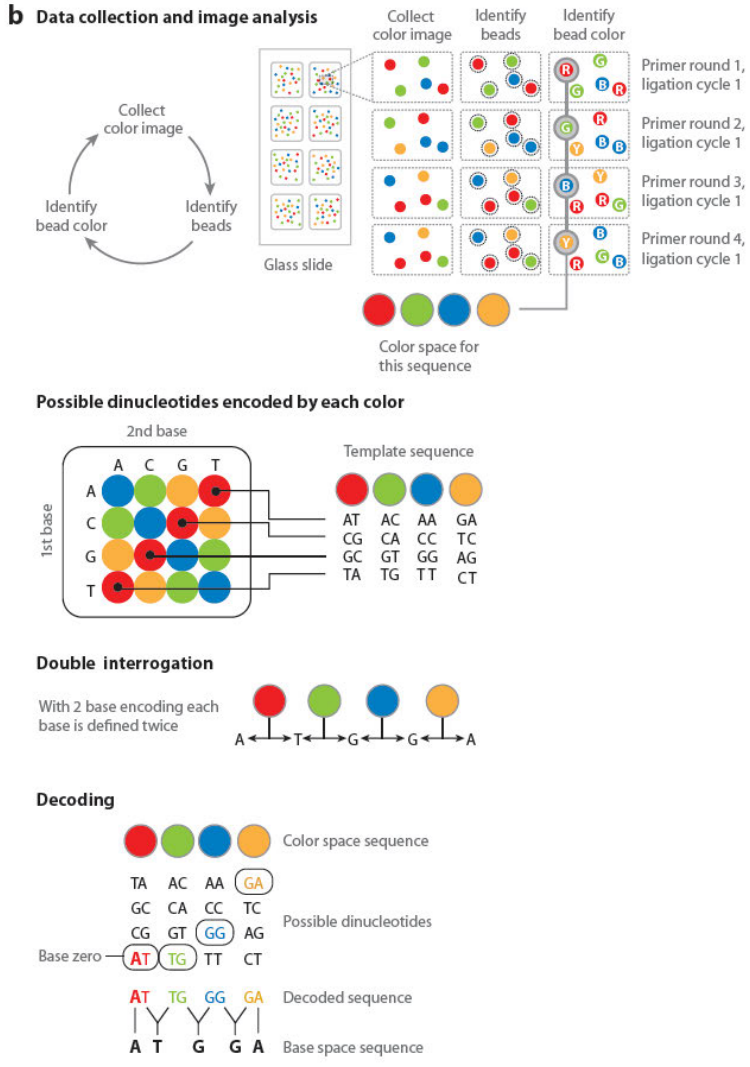


Figure 3

(*a*) The ligase-mediated sequencing approach of the Applied Biosystems SOLiD sequencer. In a manner similar to Roche/454 emulsion PCR amplification, DNA fragments for SOLiD sequencing are amplified on the surfaces of 1-μm magnetic beads to provide sufficient signal during the sequencing reactions, and are then deposited onto a flow cell slide. Ligase-mediated sequencing begins by annealing a primer to the shared adapter sequences on each amplified fragment, and then DNA ligase is provided along with specific fluorescent-labeled 8mers, whose 4th and 5th bases are encoded by the attached fluorescent group. Each ligation step is followed by fluorescence detection, after which a regeneration step removes bases from the ligated 8mer (including the fluorescent group) and concomitantly prepares the extended primer for another round of ligation. (*b*) Principles of two-base encoding. Because each fluorescent group on a ligated 8mer identifies a two-base combination, the resulting sequence reads can be screened for base-calling errors versus true polymorphisms versus single base deletions by aligning the individual reads to a known high-quality reference sequence.

NeoGenomics
Exhibit 1104 Page 9

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

Chromatin immunoprecipitation (ChIP): chemical crosslinking of DNA and proteins, and immunoprecipitation using a specific antibody to determine DNA:protein associations in vivo

Quantitative polymerase chain reaction (qPCR): rapidly measures the quantity of DNA, cDNA, or RNA present in a sample through cycle-by-cycle measurement of incorporated fluorescent dyes

capillary sequence reads. Both differences affect how the reads are utilized in bioinformatic analyses, depending upon the application. For example, in strain-to-reference comparisons (resequencing), the typical definition of repeat content must be revised in the context of the shorter read length. In addition, a much higher read coverage or sampling depth is required for comprehensive resequencing with short reads to adequately cover the reference sequence at the depth and low gap size needed.

Some applications are more suitable for certain platforms than others, as detailed below. Furthermore, read length and error profile issues entail platform- and application-specific bioinformatics-based considerations. Moreover, it is important to recognize the significant impacts that implementation of these platforms in a production sequencing environment has on informatics and bioinformatics infrastructures. The massively parallel scale of sequencing implies a similarly massive scale of computational analyses that include image analysis, signal processing, background subtraction, base calling, and quality assessment to produce the final sequence reads for each run. In every case, these analyses place significant demands on the information technology (IT), computational, data storage, and laboratory information management system (LIMS) infrastructures extant in a sequencing center, thereby adding to the overhead required for high-throughput data production. This aspect of next-generation sequencing is at present complicated by the dearth of current sequence analysis tools suited to shorter sequence read data; existing data analysis pipelines and algorithms must be modified to accommodate these shorter reads. In many cases, and certainly for new applications of next-generation sequencing, entirely new algorithms and data visualization interfaces are being devised and tested to meet this new demand. Therefore, the next-generation platforms are effecting a complete paradigm shift, not only in the organization of large-scale data production, but also in the downstream bioinformatics, IT, and LIMS support required for high data utility and correct interpretation. This paradigm shift promises to radically alter the path of biological inquiry, as the following review of recent endeavors to implement next-generation sequencing platforms and accompanying bioinformatics-based analyses serves to substantiate.

ELUCIDATING DNA-PROTEIN INTERACTIONS THROUGH CHROMATIN IMMUNOPRECIPITATION SEQUENCING

The association between DNA and proteins is a fundamental biological interaction that plays a key part in regulating gene expression and controlling the availability of DNA for transcription, replication, and other processes. These interactions can be studied in a focused manner using a technique called chromatin immunoprecipitation (ChIP) (43). ChIP entails a series of steps: (*a*) DNA and associated proteins are chemically cross-linked; (*b*) nuclei are isolated, lysed, and the DNA is fragmented; (*c*) an antibody specific for the DNA binding protein (transcription factor, histone, etc.) of interest is used to selectively immunoprecipitate the associated protein:DNA complexes; and (*d*) the chemical crosslinks between DNA and protein are reversed and the DNA is claimed for downstream analysis. In early applications, typical analyses examined the specific gene of interest by qPCR (quantitative PCR) or Southern blotting to determine if corresponding sequences were contained in the captured fragment population. Recently, genome-wide ChIP-based studies of DNA-protein interactions became possible in sequenced genomes by using genomic DNA microarrays to assay the released fragments. This so-called ChIP-chip approach was first reported by Ren and coworkers (38). Although utilized for a number of important studies, it has several drawbacks, including a low signal-to-noise ratio and a need for replicates to build statistical power to support putative binding sites.

E
NeoGenomics
Exhibit 1104 Page 10

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

Many of these drawbacks were addressed by shifting the readout of ChIP-derived DNA sequences onto next-generation sequencing platforms. The precedent-setting paper for this paradigm was published by Johnson and colleagues (21), who used the model organism *Caenorhabditis elegans* and the Roche platform to elucidate nucleosome positioning on genomic DNA. This study established that sequencing the micrococcal nuclease–derived digestion products of genomic DNA carefully isolated from mixed stage hermaphrodite populations of *C. elegans* was sufficient to generate a genome-wide, highly precise positional profile of chromatin. This capability enables studies of specific physiological conditions and their genome-wide impact on nucleosome positioning, among other applications. Subsequent studies have utilized a ChIP-based approach and the Illumina platform to provide insights into transcription factor binding sites in the human genome such as neuron-restrictive silencer factor (NRSF) (20) and signal transducer and activator of transcription 1 (STAT1) (39). In a landmark study, Mikkelsen and coworkers (32) explored the connection between chromatin packaging of DNA and differential gene expression using mouse embryonic stem cells and lineage-committed mouse cells (neural progenitor cells and embryonic fibroblasts), providing a next-generation sequencing-based framework for using genome-wide chromatin profiling to characterize cell populations. This group demonstrated that trimethylation of lysine 4 and lysine 27 determines genes that are either expressed, poised for expression, or stably repressed, effectively reflecting cell state and lineage potential. Also, lysine 4 and lysine 9 trimethylation mark imprinting control regions, whereas lysine 9 and lysine 20 trimethylation identifies satellite, telomeric, and active long-terminal repeats. These early studies demonstrated that the ability to map genome-wide changes in transcription factor binding or chromatin packaging under different environmental conditions offers a profound opportunity to couple evidence of altered DNA:protein interactions to expression level changes of specific genes in the context of specific environmental stimuli, thereby enhancing our understanding of gene expression–based cellular responses.

GENE EXPRESSION: SEQUENCING THE TRANSCRIPTOME

Historically, mRNA expression has been gauged by microarray or qPCR-based approaches; the latter is most efficient and cost-effective for a genome-wide survey of gene expression levels. Even the exquisite sensitivity of qPCR, however, is not absolute, nor is it straightforward or reliable to evaluate novel alternative splicing isoforms using either technology. In the past, serial analysis of gene expression (SAGE) (50) and variants have provided a digital readout of gene expression levels using DNA sequencing. These approaches are powerful in their ability to report the expression of genes at levels below the sensitivity of microarrays, but have been limited in their application by the cost of DNA sequencing.

Serial analysis of gene expression (SAGE): measures the quantitative expression of genes in an mRNA population sample by generating SAGE tags that are then sequenced

By contrast, the rapid and inexpensive sequencing capacity offered by next-generation sequencing instruments meshes perfectly with SAGE tagging or conventional cDNA sequencing approaches, as evidenced by several studies that used Roche/454 technology (6, 11, 46, 53). Undoubtedly, the shorter read lengths offered by the Illumina and Applied Biosystems instruments will be utilized with these approaches in the future, offering the advantage of sequencing individual SAGE tags rather than requiring concatenation of the tags prior to sequencing. Indeed, one might imagine combining the data obtained from isolating and sequencing ChIP-derived DNA bound by a transcription factor of interest to the corresponding coisolated and sequenced mRNA population from the same cells. Such experiments will be entirely feasible with next-generation technologies, especially given the low input amount of each type of biomolecule required for a suitable library and the high sensitivity afforded by the sequencing method.

E

NeoGenomics
Exhibit 1104 Page 11

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

Noncoding RNA (ncRNA): any native RNA that is transcribed but not translated into a protein

microRNA (miRNA): 21–23-base RNA molecules that participate in RNA-induced silencing of gene expression

DISCOVERING NONCODING RNAs

One of the most exciting areas of biological research in recent years has been the discovery and functional analysis of noncoding RNA (ncRNA) systems in different organisms. First described in plants, ncRNAs are providing new insights into gene regulation in animal systems as well, as recognized by the awarding of the Nobel Prize in Medicine and Physiology to Andrew Fire and Craig Mello in 2006. Perhaps the most profound impact of next-generation sequencing technology has been on the discovery of novel ncRNAs belonging to distinct classes in an extraordinarily diverse set of species (3, 8, 9, 18, 22, 29, 41, 55). In fact, this approach has been responsible for the discovery of ncRNA classes in organisms not previously known to possess them (41). These discoveries are being coupled with an ever-expanding comprehension of the functions embodied by these unique RNA species, including gene regulation by a variety of mechanisms. In this regard, studying the roles of specific microRNAs (miRNAs) in cancer is helping to uncover certain aspects of the disease (10, 44).

Noncoding RNA discovery is best accomplished by sequencing because the evolutionary diversity of ncRNA gene sequences makes it difficult to predict their presence in a genome with high certainty by computational methods alone. The unique structures of the processed ncRNAs pose difficulties for converting them into next-generation sequencing libraries (29), but remarkable progress has already been made in characterizing these molecules. With these barriers dissolving, the high capacity and low cost of next-generation platforms ensure that discovery of ncRNAs will continue at a rapid pace and that sequence variants with important functional impacts will also be determined. Because the readout from next-generation sequencers is quantitative, ncRNA characterization will include detecting expression level changes that correlate with changes in environmental factors, with disease onset and progression, and perhaps with complex disease onset or severity, for example. Importantly, the discovery and characterization of ncRNAs will enhance the annotation of sequenced genomes such that, especially in model organisms and humans, the impact of mutations will become more broadly interpretable across the genome.

ANCIENT GENOMES RESURRECTED

Attempts to characterize fossil-derived DNAs have been limited by the degraded state of the sample, which in the past permitted only mitochondrial DNA sequencing and typically involved PCR amplification of specific mitochondrial genome regions (1, 15, 23, 24, 35, 40). The advent of next-generation sequencing has for the first time made it possible to directly sample the nuclear genomes of ancient remains from the cave bear (34), mammoth (37), and the Neanderthal (17, 33). Several non-trivial technical complications arise in these inquiries, most notably the need to identify contaminating DNA from modern humans in the case of Neanderthal remains. Although even next-generation sequencing of these sample remains is quite inefficient, owing largely to bacterial DNA that is coisolated with the genomic preparation and the degraded nature of the ancient genome, important characterizations are being made. So far, one million bases of the Neanderthal genome have been sequenced, starting from DNA obtained from a single fossil bone (17).

METAGENOMICS EMERGES

Characterizing the biodiversity found on Earth is of particular interest as climate changes reshape our planet. DNA- or RNA-based approaches for this purpose are becoming increasingly powerful as the growing number of sequenced genomes enables us to interpret partial sequences obtained by direct sampling of specific environmental niches. Such investigations are referred to as metagenomics, and are typically aimed at answering the question: who's there? Conventionally, this question is

E

NeoGenomics
Exhibit 1104 Page 12

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

addressed by isolating DNA from an environmental sample, amplifying the collective of 16S ribosomal RNA (rRNA) genes with degenerate PCR primer sets, subcloning the PCR products that result, and classifying the taxa present according to a database of assigned 16S rRNA sequences. As an alternative, DNA (or RNA) is isolated, subcloned, and then sequenced to produce a fragment pool representative of the existing population. These sequences can then be translated in silico into protein fragments and compared with the existing database of annotated genome sequences to identify community members. In both approaches, deep sequencing of the population of subclones is necessary to obtain the full spectrum of taxa present, and is limited by potential cloning bias that can result from the use of bacterial cloning. By sampling RNA sequences from a metagenomic isolate, one can attempt to reconstruct metabolic pathways that are active in a given environment (45, 47). Several early metagenomic studies utilized DNA sequence sampling by capillary sequencing to investigate an acidophilic biofilm (49), an acid mine site (12) and the Sargasso Sea (52). Although these studies defined metagenomics as a scientific pursuit, they were limited in the breadth of diversity that could be sampled owing to the expense of the conventional sequencing process. By contrast, the rapid, inexpensive, and massive data production enabled by next-generation platforms has caused a recent explosion in metagenomic studies. These studies include previously sampled environments such as the ocean (2, 19, 42) and an acid mine site (13), but soil (14, 27) and coral reefs (54) also were studied by Roche/454 pyrosequencing.

Another metagenomic environment that is being characterized by next-generation sequencing is the human microbiome; the human body contains several highly specific environments that are inhabited by various microbial, fungal, viral, and eukaryotic symbiont communities, the inhabitants of which may vary according to the health status of the individual. These environments include the skin, the oral and nasal cavities, the gastrointestinal tract, and the vagina, among others. Particularly well-studied is the lower intestine of humans, first characterized by 16S rRNA classification (4, 5, 28) and more recently by 454 pyrosequencing in adult humans (16, 36, 48) and in infants (25, 36). Supporting these characterization efforts will be a large-scale project to sequence hundreds of isolated microbial genomes that are known symbionts of humans as references (**http://www.genome.gov/25521743**). The early successes in human microbiome characterization and the apparent interplay between the human host and its microbial census have resulted in the inclusion of a Human Microbiome Initative in the NIH Roadmap (**http://nihroadmap.nih.gov/hmp/**). The associated funding opportunities, with the advantages offered by next-generation sequencing instrumentation, should initiate a revolution in our understanding of how the human microbiome influences our health status.

Metagenomics: the genomics-based study of genetic material recovered directly from environmentally derived samples without laboratory culture

Epigenomics: seeks to define the influence of changes to gene expression that are independent of gene sequence

FUTURE POSSIBILITIES

As this review has described, the advent and widespread availability of next-generation sequencing instruments has ushered in an era in which DNA sequencing will become a more universal readout for an increasingly wide variety of front-end assays. However, more applications of next-generation sequencing, beyond those covered here, are yet to come. For example, genome resequencing will likely be used to characterize strains or isolates relative to high-quality reference genomes such as *C. elegans*, *Drosophila*, and human. Studies of this type will identify and catalog genomic variation on a wide scale, from single nucleotide polymorphisms (SNPs) to copy number variations in large sequence blocks (>1000 bases). Ultimately, resequencing studies will help to better characterize, for example, the range of normal variation in complex genomes such as the human genome, and aid in our ability to comprehensively view the range of genome variation in clinical isolates of pathogenic microbes, viruses, etc.

Epigenomic variation, as an extension of genome resequencing applications, also will be

E

NeoGenomics
Exhibit 1104 Page 13

investigated using next-generation sequencing approaches that enable the ascertainment of genome-wide patterns of methylation and how these patterns change through the course of an organism's development, in the context of disease, and under various other influences.

Perhaps the most exciting possibility engendered by the ability to use DNA sequencing to rapidly read out experimental results is the enhanced potential to combine the results of different experiments—correlative analyses of genome-wide methylation, histone binding patterns, and gene expression, for example—owing to the similar data type produced. The power in these correlative analyses is the power to begin unlocking the secrets of the cell.

DISCLOSURE STATEMENT

The author serves as a Director of the Applera Corporation.

LITERATURE CITED

1. Adcock GJ, Dennis ES, Easteal S, Huttley GA, Jermiin LS, et al. 2001. Mitochondrial DNA sequences in ancient Australians: implications for modern human origins. *Proc. Natl. Acad. Sci. USA* 98:537–42
2. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. 2006. The marine viromes of four oceanic regions. *PLoS Biol.* 4:e368
3. Axtell MJ, Snyder JA, Bartel DP. 2007. Common functions for diverse small RNAs of land plants. *Plant Cell* 19:1750–69
4. Bäckhed F, Ding H, Wang T, Hooper LV, Koh GY, et al. 2004. The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl. Acad. Sci. USA* 101:15718–23
5. Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. 2005. Host-bacterial mutualism in the human intestine. *Science* 307:1915–20
6. Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, et al. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* 7:246
7. Bentley DR. 2006. Whole-genome resequencing. *Curr. Opin. Genet. Dev.* 16:545–52
8. Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, et al. 2006. Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* 38:1375–77
9. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, et al. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128:1089–103
10. Calin GA, Liu CG, Ferracin M, Hyslop T, Spizzo R, et al. 2007. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* 12:215–29
11. Cheung F, Haas BJ, Goldberg SM, May GD, Xiao Y, Town CD. 2006. Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics* 7:272
12. Edwards KJ, Bond PL, Gihring TM, Banfield JF. 2000. An archaeal iron-oxidizing extreme acidophile important in acid mine drainage. *Science* 287:1796–99
13. Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, et al. 2006. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7:57
14. Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, et al. 2007. Metagenomic and small-subunit rRNA analyses of the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl. Environ. Microbiol.* 73:7059–66
15. Gilbert MT, Tomsho LP, Rendulic S, Packard M, Drautz DI, et al. 2007. Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science* 317:1927–30
16. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. 2006. Metagenomic analysis of the human distal gut microbiome. *Science* 312:1355–59
17. Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, et al. 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* 444:330–36
18. Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, et al. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. *Cell* 129:69–82

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

E

NeoGenomics
Exhibit 1104 Page 14

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

19. Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, et al. 2007. Microbial population structures in the deep marine biosphere. *Science* 318:97–100
20. Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316:1497–502
21. Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ. 2006. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res.* 16:1505–16
22. Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, et al. 2007. Genome-wide profiling and analysis of *Arabidopsis* siRNAs. *PLoS Biol.* 5:e57
23. Krause J, Dear PH, Pollack JL, Slatkin M, Spriggs H, et al. 2006. Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature* 439:724–27
24. Krings M, Geisert H, Schmitz RW, Krainitzki H, Paabo S. 1999. DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen. *Proc. Natl. Acad. Sci. USA* 96:5581–85
25. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, et al. 2007. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* 14:169–81
26. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
27. Leininger S, Urich T, Schloter M, Schwark L, Qi J, et al. 2006. Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* 442:806–9
28. Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. 2005. Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. USA* 102:11070–75
29. Lu C, Meyers BC, Green PJ. 2007. Construction of small RNA cDNA libraries for deep sequencing. *Methods* 43:110–17
30. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–80
31. McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, et al. 2001. A physical map of the human genome. *Nature* 409:934–41
32. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448:553–60
33. Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, et al. 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science* 314:1113–18
34. Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, et al. 2005. Genomic sequencing of Pleistocene cave bears. *Science* 309:597–99
35. Ovchinnikov IV, Gotherstrom A, Romanova GP, Kharitonov VM, Liden K, Goodwin W. 2000. Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* 404:490–93
36. Palmer C, Bik EM, Digiulio DB, Relman DA, Brown PO. 2007. Development of the human infant intestinal microbiota. *PLoS Biol.* 5:e177
37. Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, et al. 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* 311:392–94
38. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* 290:2306–9
39. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* 4:651–57
40. Rogaev EI, Moliaka YK, Malyarchuk BA, Kondrashov FA, Derenko MV, et al. 2006. Complete mitochondrial genome and phylogeny of Pleistocene mammoth *Mammuthus primigenius*. *PLoS Biol.* 4:e73
41. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, et al. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* 127:1193–207
42. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, et al. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. USA* 103:12115–20
43. Solomon MJ, Larsen PL, Varshavsky A. 1988. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* 53:937–47
44. Stahlhut Espinosa CE, Slack FJ. 2006. The role of microRNAs in cancer. *Yale J. Biol. Med.* 79:131–40

E

NeoGenomics
Exhibit 1104 Page 15

45. Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, et al. 2006. Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* 440:790–94
46. Torres TT, Metta M, Ottenwälder B, Schlötterer C. 2007. Gene expression profiling by massively parallel sequencing. *Genome Res.* 18:172–77
47. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. 2005. Comparative metagenomics of microbial communities. *Science* 308:554–57
48. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444:1027–31
49. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43
50. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. *Science* 270:484–87
51. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
52. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74
53. Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. 2007. Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiol.* 144:32–42
54. Wegley L, Edwards R, Rodriguez-Brito B, Liu H, Rohwer F. 2007. Metagenomic analysis of the microbial community associated with the coral *Porites astreoides*. *Environ. Microbiol.* 9:2707–19
55. Zhao T, Li G, Mi S, Li S, Hannon GJ, et al. 2007. A complex system of small RNAs in the unicellular green alga *Chlamydomonas reinhardtii*. *Genes Dev.* 21:1190–203

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

E

NeoGenomics
Exhibit 1104 Page 16

Annual Review of Genomics and Human Genetics

Volume 9, 2008

Contents

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

Human Telomere Structure and Biology
Harold Riethman 1

Infectious Disease in the Genomic Era
Xiaonan Yang, Hongliang Yang, Gangqiao Zhou, and Guo-Ping Zhao 21

ENU Mutagenesis, a Way Forward to Understand Gene Function
Abraham Acevedo-Arozena, Sara Wells, Paul Potter, Michelle Kelly, Roger D. Cox, and Steve D.M. Brown 49

Clinical Utility of Contemporary Molecular Cytogenetics
Bassem A. Bejjani and Lisa G. Shaffer 71

The Role of Aminoacyl-tRNA Synthetases in Genetic Diseases
Anthony Antonellis and Eric D. Green 87

A Bird's-Eye View of Sex Chromosome Dosage Compensation
Arthur P. Arnold, Yuichiro Itoh, and Esther Melamed 109

Linkage Disequilibrium and Association Mapping
B. S. Weir 129

Positive Selection in the Human Genome: From Genome Scans to Biological Significance
Joanna L. Kelley and Willie J. Swanson 143

The Current Landscape for Direct-to-Consumer Genetic Testing: Legal, Ethical, and Policy Issues
Stuart Hogarth, Gail Javitt, and David Melzer 161

Transcriptional Control of Skeletogenesis
Gerard Karsenty 183

A Mechanistic View of Genomic Imprinting
Ky Sha 197

Phylogenetic Inference Using Whole Genomes
Bruce Rannala and Ziheng Yang 217

N
Ext

NeoGenomics
Exhibit 1104 Page 17

Transgenerational Epigenetic Effects
Neil A. Youngson and Emma Whitelaw 233

Evolution of Dim-Light and Color Vision Pigments
Shozo Yokoyama 259

Genetic Basis of Thoracic Aortic Aneurysms and Dissections: Focus on Smooth Muscle Cell Contractile Dysfunction
Dianna M. Milewicz, Dong-Chuan Guo, Van Tran-Fadulu, Andrea L. Lafont, Christina L. Papke, Sakiko Inamoto, and Hariyadarshi Pannu 283

Cohesin and Human Disease
Jinglan Liu and Ian D. Krantz 303

Genetic Predisposition to Breast Cancer: Past, Present, and Future
Clare Turnbull and Nazneen Rahman 321

From Linkage Maps to Quantitative Trait Loci: The History and Science of the Utah Genetic Reference Project
Stephen M. Prescott, Jean Marc Lalouel, and Mark Leppert 347

Disorders of Lysosome-Related Organelle Biogenesis: Clinical and Molecular Genetics
Marjan Huizing, Amanda Helip-Wooley, Wendy Westbroek, Meral Gunay-Aygun, and William A. Gahl 359

Next-Generation DNA Sequencing Methods
Elaine R. Mardis 387

African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping
Michael C. Campbell and Sarah A. Tishkoff 403

Indexes

Cumulative Index of Contributing Authors, Volumes 1–9 435

Cumulative Index of Chapter Titles, Volumes 1–9 438

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* articles may be found at http://genom.annualreviews.org/

Annu. Rev. Genom. Hum. Genet. 2008.9:387-402. Downloaded from www.annualreviews.org
Access provided by 74.202.51.11 on 03/15/23. For personal use only.

N
Exh
NeoGenomics
Exhibit 1104 Page 18