

From: [John West](#)
To: [Scott Kirk](#)
Cc: [Richard Chen](#); [Mark Pratt](#)
Subject: Accuracy documents to convert into project plan format
Date: Tuesday, May 29, 2012 12:58:39 PM
Attachments: [Accuracy framework JW 5Jan2012.ppt](#)
[Accuracy JW 5Mar2012.ppt](#)
[Accuracy Gantt JW 15Mar2012.ppt](#)
[Accuracy slides for mtg 13Apr2012.ppt](#)
[Accuracy slides 2May2012 JW.pptx](#)
[Methods to determine SNP error rate 3May2012 JW.pptx](#)
[Accuracy JW 16May2012.ppt](#)
[Accuracy differentiation JW 22May2012.xls](#)
[Accuracy Gantt Spreadsheet 21April2012 JW.xls](#)

Hi Scott,

I am attaching a sequence of documents related to accuracy planning which have been developed starting in early January. Since they are sequential in time, there is overlap of content, but perhaps walking through them will give you an idea of how the ideas developed and where we are now. These should complement the material you have from Mark. I would be happy to discuss these with you, Mark & Rich as you see fit, so this can become more integrated with the other project plans. Please let me know.

Thanks,

John

ACCURACY DIFFERENTIATION

Draft for discussion, JW, 22 May, 2012

ORGANIZATION OF THIS FILE

- Company differentiation around accuracy
- Better sequencing in the laboratory
- Better variant detection & reporting
- More accurate databases, for interpretation

COMPANY DIFFERENTIATION AROUND ACCURACY

- Focus on accuracy relevant to medical interpretation
- Better understanding of the issues than anyone
- Best track record of publications on the issue
- Unbiased from a platform standpoint & able to combine platforms
- More comprehensive data on accuracy than anyone
 - World's only collection of genomes sequenced on both ILMN and CG platforms, plus arrays and karyotyping
 - Largest family pedigree sequenced to high coverage
 - Only genome sequenced on Sanger, ILMN and CGI
- Databases more accurate than those publicly available
- Able to provide a detailed quantitative view of mechanisms underlying errors
- Deep understanding of accuracy issues used to create better results
 - Not just flagging errors, or filtering out those loci, but fixing the problems
 - More insightful approaches delivery accuracy affordably

BETTER SEQUENCING IN THE LABORATORY (WHEN MANAGED BY PERSONALIS)

Focus on getting the whole medically-interpretable genome, accurately, even if more expensive

- Use insight into error types & medical content to keep this affordable

Combine data from multiple different runs of a single platform

- Combine paired-end libraries made with multiple insert lengths
- Use longer read lengths (e.g. 2 x 250 bases, when available later in 2012)
 - More expensive because only MiSeq, but clearly better
- Combine with bulk shorter-read data from HiSeq
- Substantially more efficient at split-read & junction-sequence SV detection
- Key to single-base breakpoint determination

Combine data from multiple platforms

- More experience with Illumina / Complete Genomics than anyone
- May add Ion Torrent, Oxford Nanopore
- Guided by deep understanding of differential error mechanisms in each platform
 - Not tied to any one platform
 - Use whatever it takes to get the best possible combination

Combine data from outside next-gen sequencing

- Several major areas of medical genetics are not well assayed by next gen sequencing
 - Example 1 : Diseases caused by STR-expansion (e.g. Huntington's)
 - Example 2 : Robertsonian translocations
- Personalis will combine NGS data with Non-NGS technologies to create a complete assessment
 - Add karyotyping
 - Add electrophoresis where appropriate (TBD)
 - Others
- Orthogonal technologies also provide validation of NGS results
 - Integrate NGS with array (fluorescence for SV's, in addition to genotypes)
 - Sanger and/or (PCR + electrophoresis) as follow-up to SNP's / SV's of specific genomes (option TBD)

Ability to create semi-custom products focused on medically interpretable parts of the genome

- Leverages Personalis advantages in content
- Custom hybridization array
- Custom pullout set
- Other assays to fix specific error types

Question : Should there be a "Personalis exome" option ?

- More comprehensive / accurate at exome price level ?

What is proprietary about this approach ?

Personalis can focus its efforts based on the world's best content (re medical interpretation)
Personalis will develop proprietary understanding about how best to combine multiple technologies
Personalis does not face the competitive & anti-trust barriers that platform companies do
Personalis' people combine deep experience with both platforms and interpretation & can leverage the two against each other
Personalis can combine work in the lab and in bioinformatics, in a way that pure informatics companies can't

BETTER VARIANT DETECTION & REPORTING

Fewer false positive SNP's due to method of generating laboratory data

Combination of paired-end insert lengths covers more of the genome uniquely

Fewer false negative SNP's due to method of generating laboratory data

More uniform coverage by combining library prep methods & platforms

Orthogonal validation of millions of SNP genotypes by array

Integrated with next gen sequencing data, not just another separate report

Better alignment, due to better reference sequence

SNP major allele ref by ethnicity (in our first product)

InDel major allele ref by ethnicity (later)

Other advances as R&D develops them:

Iterative alignment

TBD changes in SNP alleles called with Personalis reference vs public standard

Likely more improvement in non-European ethnicities

Include (eventually) changes in InDels called as well

We provide the only support available specifically for admixed genomes

Major-allele non-ethnic reference, or even more advanced options

Focused effort to align in the presence of SNP / InDel clusters, MNP's

May leverage Hugo's BreakSeq approach

May need time in the development plan

Better SNP reporting due to better reference

We report variants when sequence is a homozygous match to the public ref but that's the minor allele

Entirely missed by systems which use the public reference

Example : Factor V Leiden

Rong had a whole paper on all the disease variants in the public ref

> 1M loci where we can be different

We should calculate the average # actual loci / genome, by ethnicity

At het loci, we report both alleles, but we report the minor allele as the variant

Not the allele which is different from the public reference

Better detection of SV's

Better lab data for SV detection:

Longer reads (better for approaches based on split-read & junction-sequences)

MiSeq 2x250 or other platform

Multiple insert lengths

Electrophoretic assay of STR-expansions

Karyotyping for Robertsonian translocations

Orthogonal technologies for validation of SV's:

Fluorescence intensity data from hybridization arrays

We combine the results from five different algorithmic approaches

We test our SV algorithms by Mendelian Inheritance in high coverage whole genome family data sets

one which was sequenced with ten different paired-end libraries spanning 200 - 40,000 bases

and validate them using fluorescent intensities from high density hybridization arrays

We don't treat all SV's as novel - we have the world's best database of known SV's and their junction sequences

Detection is better when you know exactly what you are looking for

We should have a meeting to discuss how we can (easily ?) build this

Start with 1,000 Genomes result Hugo has helped create

Large data set but low coverage may make detection less certain in low MAF SV's

Others will be able to access this eventually, potentially catching up, or claiming to

Augment this with (more confident ?) SV's from:

Full coverage (30-40x) genomes (West, Altman, 40 Koreans, others we can download)
High coverage (>60x) genomes (Snyder, CEPH1463, Venter, others ?)

Better reporting of SV's

We determine the zygosity of deletions and report it
Deletions integrated with SNP report, e.g. "A-" vs "AA" inside a het deletion
We report SV's with their allele frequencies in the ethnicity matching the sample

Flagging of potential errors

Many subtle error types not recognized by others
Error mechanisms underlying differences when the same person is sequenced twice (it's not just Poisson !)
Error loci determined from deep & multi-platform sequencing of large families
Error loci determined by extensive platform comparison, both NGS/NGS and NGS/Non-NGS
Detailed understanding of compressions, and large unpublished catalog of them

MORE ACCURATE DATABASES, FOR INTERPRETATION

Cleaner databases:

Well financed, systematic manual curation to industrial QC standard
Standardized medical language hierarchy
Extensive cross checking of databases developed independently
VariMed vs HGMD
MendelDB vs OMIM
Personalis PharmGKB vs public PGKB (need to be careful in this positioning)

Databases others will not have:

Regulome
BreakSeq (esp if augmented with private Personalis data)
Compression list (described in publications but not released)
Variant data derived from a broad collection of genomes
Multiple public data sets, some processed in proprietary ways by Personalis
Access to private data sets, sequenced by others
Access to private data sets, sequenced by Personalis

Framework for assessment & improvement of accuracy in whole genome medical interpretation

**Draft Jan 5, 2012
John West**

Research Market Products

- Comprehensive medically-focused interpretation of whole human genome sequences
 - Huge # of potential results
 - Even a low % error can swamp customer & waste time trying to track down invalid hypotheses
 - If research is in a clinical setting, particularly with return of results to patients, IRB's may want "FDA-like" quality systems in place
 - Standard quality approaches
 - Standard quality terminology & reporting
 - Accuracy / quality to "near-FDA" standards can be a sales advantage in the research market
- Maximal-accuracy analysis of a genome. Personalis :
 - Takes responsibility for the sample
 - Specifies the combination of next-gen sequencing performed
 - Augments NGS with other technologies to determine key variants
 - Conducts genotyping & assesses concordance
 - Conducts follow-up genetic & non-genetic testing to validate results

Eventual clinical product

- Also focused on comprehensive medically-focused interpretation of whole human genomes
 - FDA anticipates “Class 3” (highest) level of regulatory oversight
 - Need to undo bad reputation created by DTC genotyping companies
 - Pro-active early leadership in quality systems for whole genome may create a positive reputation for Personalis at FDA

CDC Process – Established process & terminology

(Augment with FDA processes, likely to be similar)

CDC Home



Centers for Disease Control and Prevention

CDC 24/7: Saving Lives. Protecting People. Saving Money through Prevention.

Genomic Testing

ACCE Model Process for Evaluating Genetic Tests

From 2000 – 2004, CDC's Office of Public Health Genomics (OPHG) established and supported the ACCE Model Project, which developed the first publicly-available analytical process for evaluating scientific data on emerging genetic tests. The ACCE framework has guided or been adopted by various entities in the United States and worldwide for evaluating genetic tests; the CDC-supported [EGAPP™ initiative](#) builds on the ACCE model structure and experience.

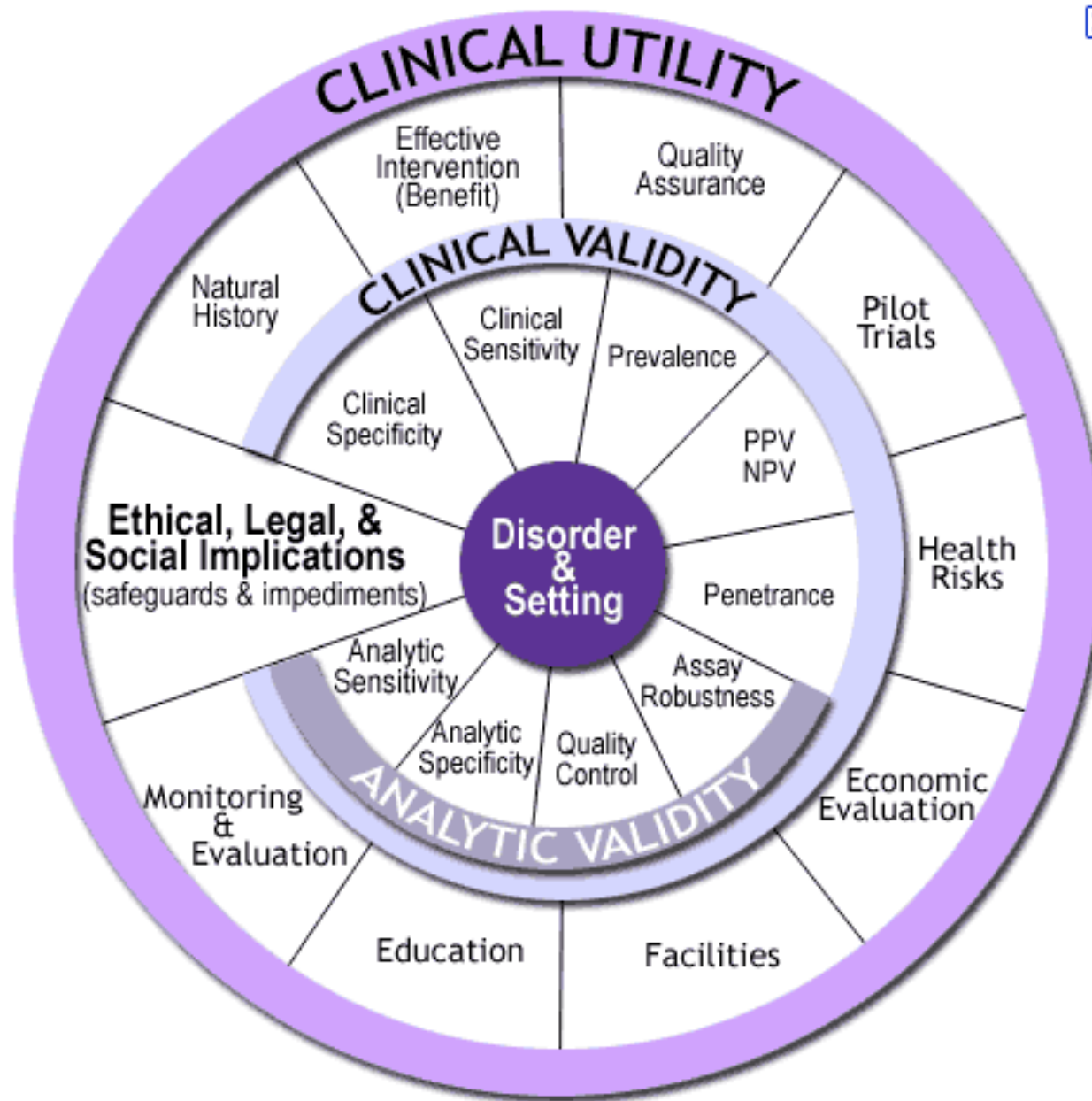
Introduction to ACCE

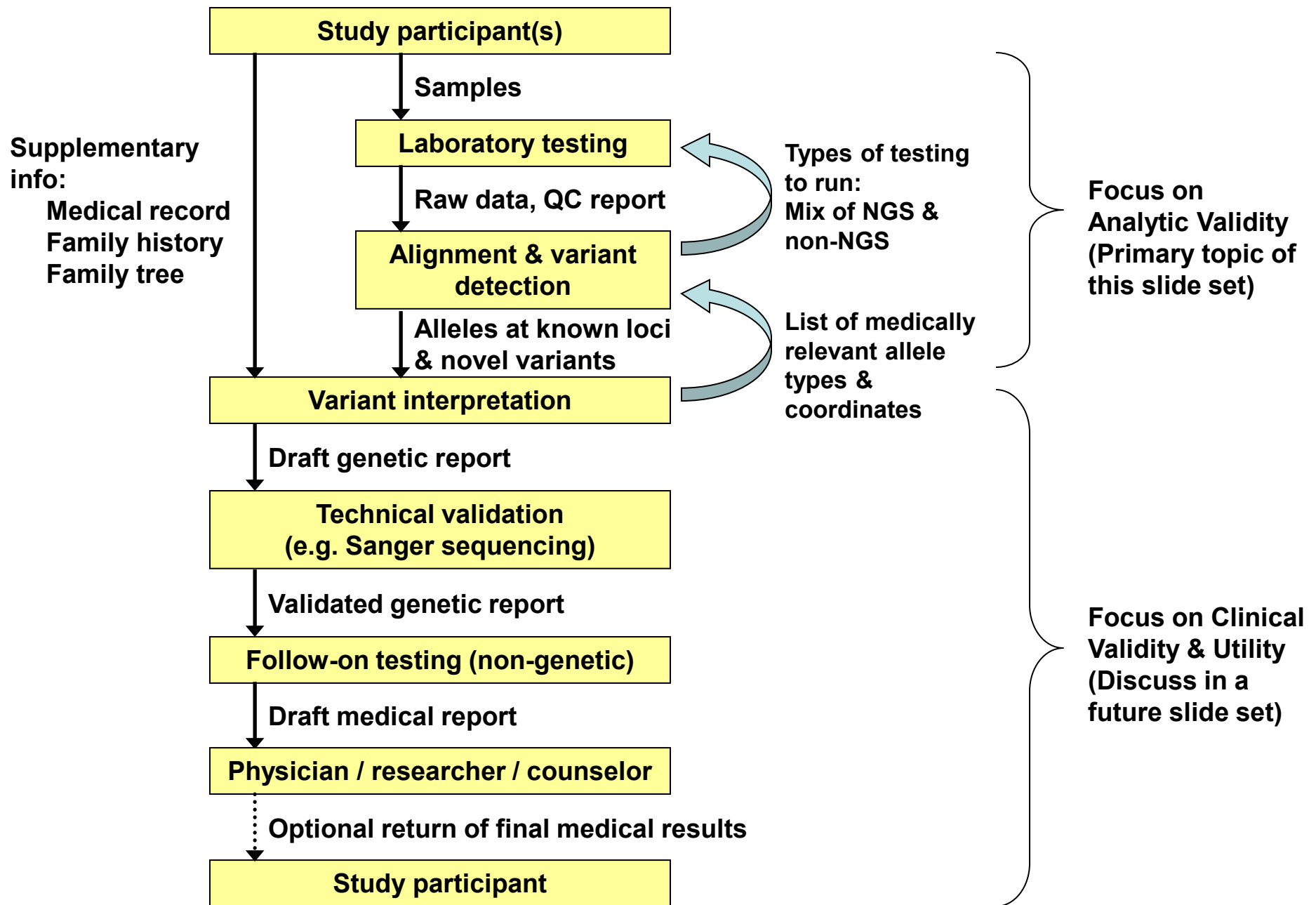
ACCE, which takes its name from the four main criteria for evaluating a genetic test — [analytic validity](#), [clinical validity](#), [clinical utility](#) and associated [ethical, legal and social implications](#) — is a model process that includes collecting, evaluating, interpreting, and reporting data about DNA (and related) testing for disorders with a genetic component in a format that allows policy makers to have access to up-to-date and reliable information for decision making. The ACCE model process is composed of a standard set of [44 targeted questions](#) (3) that address disorder, testing, and clinical scenarios, as well as analytic and clinical validity, clinical utility, and associated ethical, legal, and social issues.

An important by-product of the ACCE model process is the identification of gaps in knowledge that will help to define future research agendas. The ACCE approach builds on a methodology originally described by Wald and Cuckle (1) and on terminology introduced by the Secretary's Advisory Committee on Genetic Testing (2).

Learn more about [ACCE](#).







Elements of Analytic Validity in CDC's ACCE Model

See Appendix for additional detail

- Analytic Sensitivity : How often positive when mutation is present ?
 - One minus the false negative rate ?
- Analytic Specificity : How often negative when mutation is not present ?
 - One minus the false positive rate ?
- Assay Robustness : How often does the test fail to give a usable result ?
- Repeatability :
 - On the same sample
 - Within & between labs
 - Between sample types & other process variables
- Confirmatory testing to resolve false positives ?
 - Genetic validation tests
 - Non-genetic tests (molecular, imaging, physiological)
- Quality Control: Internal QC program defined & externally monitored ?

Crucial to develop a model which links analytic validity results to monitor-able process variables

- Allele determination performance is known to vary widely between loci
- With millions of known loci, many quite rare, it is impractical to determine the assay performance for each as an independent system.
- Novel loci, inherently, cannot be validated in advance
- Proposed approach:
 - Create a quantitative process model of allele determination
 - Incorporate known process variables (sample type, storage & preparation, coverage, read length, repeat structure, algorithm parameters, etc)
 - Augment with models of process variability & failure mechanisms
 - Instrument the model with QC observables, related to normal & failure modes
 - Compare model predictions with real data & characterize differences
 - Iterate until model can confidently predict analytic validity metrics for both known & novel alleles, including in combination
 - System performance metric should prioritize medically interpretable alleles
 - Use insights from experimentally assessed error model to :
 - Evaluate / prioritize process improvements
 - Select disease-associated alleles which can be assayed with NGS⁺

Genetic variant types to include in model

- SNP's
- InDel's
- Larger deletions
- STR / VNTR's
- Structural variants
- Copy Number Variation
- Trisomies
- Combinations:
 - Any of these can be heterozygous
 - Compound heterozygosity
 - Small variants within larger het deletions or other CNV's
 - Clusters of closely spaced variants, which might interfere with each others' detection

Examples of known error mechanisms

Many more likely to follow

- Locally low / no coverage due to:
 - GC bias of biochemistry
 - Alignment degeneracy (repeats)
 - Clusters of variant alleles exceed alignment mismatch allowance
 - Alignment poor due to problems with reference sequence
- Reads align, but incorrectly
 - InDels missed, or inconsistently placed in homopolymers
 - Misplacements lead to :
 - Phantom het SNP's
 - Apparent tri-allelic loci
 - Apparent triple haplotype structure
- Raw read errors, random or systematic
- Allelic imbalance of reads due to:
 - Random extremes of expected binomial distribution, esp where low coverage
 - Allelic biases - biochemical or bioinformatic (e.g. clusters of SNP's in LD)
- Read length too short to quantify length of VNTR, or VNTR embedded in other repeat structures

Examples of process variability

Many more likely to follow

- Sample source (blood, saliva, tissue, cell culture)
- Sample preservation, culturing & prep for sequencing
- DNA sequencing platform, lab & operator(s)
- Coverage
- Read length
- Paired-end insert length (a distribution)
- Cluster creation & Sequencing chemistry versions
- Cluster density
- Variation in data processing algorithms, input parameters, reference sequence
- DNA contamination or incomplete filtering of primer-related artifacts
- Availability of supplementary data from the same sample (e.g. genotyping)
- Inconsistent supplementary information from / about the individual sequenced

QC Readouts, to monitor the process

(Particularly given that it may be impractical to standardize all process variables completely, over multiple years)

- Paired-end insert length distribution vs expected
- Coverage non-uniformity vs % GC content; relative to a standard
- Coverage histogram in regions selected to avoid repeats & %GC variation
- Total coverage ratios of chromosomes
- Raw read Q-score %-iles vs position along a read; relative to a standard
- Allelic distribution of reads vs expected binomial, at a “diagnostic” set of het SNP loci
- Sex, ethnicity, admixture and family relationships between samples vs expected from supplementary data provided
- Raw read error rates (monitored at homozygous loci), systematic / random
- Mendelian inheritance errors (family genome data sets)
- Clusters of variants which are close / dense enough to predict higher probability of errors
- Two alleles / haplotypes in single sex chromosomes, vs a standard
- Three alleles / haplotypes in diploid chromosomes , vs a standard
- Amount and location of autozygosity , vs a standard
- Screening for non-human DNA
- Concordance with genotyping data, vs a standard

Categories of experimental data to assess error types, mechanisms, frequencies

- Repeatability :
 - Same person twice (or more ?), trying to keep everything else constant
 - Alternatives: Genomes of identical twins, or children where identical
- Controlled process experiments :
 - Same person sequenced multiple times, each time varying just one parameter of the process (many of these can be done computationally)
- Concordance of other technology platforms on the same sample
- Compare process QC metrics from many sources of raw genome data, even low coverage (e.g. 1,000 Genomes)
- Mendelian inheritance errors (MIE's) determined from family genome sets
- Absolute accuracy :
 - Genomes of individuals known to have clear Mendelian diseases
 - Sequencing of complex synthesized oligo sets (synthetic genome)
 - Targeted modifications of DNA from a naturally occurring genome

Lists of problematic regions in the genome

Capture all in one place, unify, & link to underlying error mechanisms of the model

- Mapability / alignability assessments (some on Santa Cruz browser site)
- Genes with known CNV's
- Genes with known pseudogenes
- List of problematic genes for NGS (e.g. VWF)
- Evan Eichler's list of problematic regions
- 1,000 Genomes list of regions with top 0.1% of coverage
- List of compressions
- MIE's & MIE-cluster regions from family data
- Difficult regions from phasing of a family
- Regions of high raw read error rates (monitor at homozygous loci)
- Loci apparently tri-allelic in the genome of a healthy individual, or 3 haplotypes
- Regions of allelic bias
- Regions of no or low aligned coverage, determined experimentally
- Segmental duplications & other repeat structures

List of Variants to Optimize Against

- Medically interpretable variants (not just SNP's !):
 - Mendelian “Top 200”
 - Specific loci / variant types, where known
 - Genes in which to identify novel variants
 - PharmGKB
 - Curated & novel variants
 - Alleles in top Risk-O-Grams
 - Diseases
 - Molecular or physiologic medically relevant phenotypes
 - Blood & tissue typing
- Functional elements (broad interest for discovery research)
 - RefSeq genes
 - Linc-RNA
 - Regulome elements
 - eQTL loci

Proposed Initial Priorities

- Literature review related to genome sequencing accuracy, followed by discussion with SAB
- Develop an initial list of top priority medically interpretable variants to optimize against
- Create / collect lists of problematic regions of the genome
- Screen target medical variant list against problematic regions to get 1st cut scale & composition of the accuracy problem
- Develop initial quantitative model of allele determination, including known error mechanisms. Predict & characterize accuracy & repeatability levels
- Obtain existing experimental data, reprocess with current version algorithms, and conduct 1st cut repeatability & accuracy assessment. Compare with model on a genome-wide basis and at the medically interpretable loci.
- Launch program to obtain long-lead-time experimental data, including prototype of “maximally accurate genome” product (based on NGS augmented by other technologies).

Appendix

**US Centers for Disease Control (CDC)
ACCE Model Process for Evaluating Genetics Tests**

**Potentially useful framework & terminology for moving towards FDA
approval of genome-based diagnostics**



<http://www.cdc.gov/genomics/gtesting/ACCE/>

Genomic Testing

ACCE Model Process for Evaluating Genetic Tests

From 2000 – 2004, CDC's Office of Public Health Genomics (OPHG) established and supported the ACCE Model Project, which developed the first publicly-available analytical process for evaluating scientific data on emerging genetic tests. The ACCE framework has guided or been adopted by various entities in the United States and worldwide for evaluating genetic tests; the CDC-supported [EGAPP™ initiative](#) builds on the ACCE model structure and experience.

Introduction to ACCE

ACCE, which takes its name from the four main criteria for evaluating a genetic test — [analytic validity](#), [clinical validity](#), [clinical utility](#) and associated [ethical, legal and social implications](#) — is a model process that includes collecting, evaluating, interpreting, and reporting data about DNA (and related) testing for disorders with a genetic component in a format that allows policy makers to have access to up-to-date and reliable information for decision making. The ACCE model process is composed of a standard set of [44 targeted questions](#) (3) that address disorder, testing, and clinical scenarios, as well as analytic and clinical validity, clinical utility, and associated ethical, legal, and social issues.

An important by-product of the ACCE model process is the identification of gaps in knowledge that will help to define future research agendas. The ACCE approach builds on a methodology originally described by Wald and Cuckle (1) and on terminology introduced by the Secretary's Advisory Committee on Genetic Testing (2).

Learn more about [ACCE](#).



Genomic Testing

ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing

Element	Component	Specific Question
Disorder/Setting		1. What is the specific clinical disorder to be studied? 2. What are the clinical findings defining this disorder? 3. What is the clinical setting in which the test is to be performed? 4. What DNA test(s) are associated with this disorder? 5. Are preliminary screening questions employed? 6. Is it a stand-alone test or is it one of a series of tests? 7. If it is part of a series of screening tests, are all tests performed in all instances (parallel) or are only some tests performed on the basis of other results (series)?

Genomic Testing

ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing

Element	Component	Specific Question
Analytic Validity		8. Is the test qualitative or quantitative?
	Sensitivity	9. How often is the test positive when a mutation is present?
	Specificity	10. How often is the test negative when a mutation is not present?
		11. Is an internal QC program defined and externally monitored? 12. Have repeated measurements been made on specimens? 13. What is the within- and between-laboratory precision? 14. If appropriate, how is confirmatory testing performed to resolve false positive results in a timely manner? 15. What range of patient specimens have been tested? 16. How often does the test fail to give a useable result? 17. How similar are results obtained in multiple laboratories using the same, or different technology?

Genomic Testing

ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing

Element	Component	Specific Question
Clinical Validity	Sensitivity	18. How often is the test positive when the disorder is present?
	Specificity	19. How often is the test negative when a disorder is not present?
		20. Are there methods to resolve clinical false positive results in a timely manner?
	Prevalence	21. What is the prevalence of the disorder in this setting?
		22. Has the test been adequately validated on all populations to which it may be offered? 23. What are the positive and negative predictive values? 24. What are the genotype/phenotype relationships? 25. What are the genetic, environmental or other modifiers?

Genomic Testing

ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing

Element	Component	Specific Question
Clinical Utility	Intervention	26. What is the natural history of the disorder?
	Intervention	27. What is the impact of a positive (or negative) test on patient care?
	Intervention	28. If applicable, are diagnostic tests available?
	Intervention	29. Is there an effective remedy, acceptable action, or other measurable benefit?
	Intervention	30. Is there general access to that remedy or action?
		31. Is the test being offered to a socially vulnerable population?
	Quality Assurance	32. What quality assurance measures are in place?
	Pilot Trials	33. What are the results of pilot trials?

Genomic Testing

ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing

Element	Component	Specific Question
	Health Risks	34. What health risks can be identified for follow-up testing and/or intervention?
		35. What are the financial costs associated with testing?
	Economic	36. What are the economic benefits associated with actions resulting from testing?
	Facilities	37. What facilities/personnel are available or easily put in place?
	Education	38. What educational materials have been developed and validated and which of these are available?
		39. Are there informed consent requirements?
	Monitoring	40. What methods exist for long term monitoring?
		41. What guidelines have been developed for evaluating program performance?

Genomic Testing

ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing

Element	Component	Specific Question
ELSI	Impediments	42. What is known about stigmatization, discrimination, privacy/confidentiality and personal/family social issues?
		43. Are there legal issues regarding consent, ownership of data and/or samples, patents, licensing, proprietary testing, obligation to disclose, or reporting requirements?
	Safeguards	44. What safeguards have been described and are these safeguards in place and effective?

Develop list of medically interpretable loci / regions for testing

Review literature

Develop Error model based on hypothesized mechanisms of action

Create QC tools

Analysis of experimental data to quantify error rates for all variant types & characterize by mechanism of action

Develop strategy

Design, implement & test technology improvements

Release version x.x

Iterate error model & refine experimental analysis

Conceptual Gantt chart for analytical accuracy development. In practice, there will be overlap of the elements.

Develop a List of Medically Interpretable Loci / Regions for Testing

- Develop a list, some of each variant-type, which we want to be able to interpret medically. An initial list can be anecdotal / representative. Eventually, it should be more comprehensive.
- PharmGKB SNP' s
- VariMed SNP' s which contribute to Risk-O-Grams
 - 1st exon, other exons, non-exonic
- Create an initial list (maybe 20 each ?) of medically interpretable:
 - Small InDels, CNV' s, VNTR' s, Deletions spanning 10 – 100k bases, Insertions, Duplications, Balanced translocations
- Create a list of genes which we want to be able to assay confidently in their entirety:
 - PhGKB VIP genes, genes for 20 major Mendelian Diseases (e.g. CFTR), 20 major cancer genes (e.g. p53, BRCA 1&2, ...)
- Variants required for blood typing
- Variants for tissue typing
- Linc-RNA regions, Regulome elements (TF binding site ranges), eQTL loci
- Loci / genes for which phasing is most likely to be necessary for interpretation

Literature Review

- Avoid re-inventing the wheel
- Literature already reviewed:
 - Marguillies (NHGRI, coverage analysis & error filtering)
 - Lam (Platform comparison)
 - Kahn (Accuracy advances, Illumina technology)
 - Complete Genomics technology whitepaper
 - Finishing of the Human Genome (2004)
 - Nature Biotech Dec 2011, Dutch group. (Filtering out errors)
 - Some literature on Huntington's disease assays
- To be reviewed:
 - Complete Genomics data pipeline paper (recent)
- To look for:
 - Papers systematically addressing the mechanisms underlying errors by type
 - Assay methods & problems for specific difficult cases
 - UCLA (Stan Nelson's group) error model (published ?)
- Mark to review algorithms and parameter settings of HugeSeq elements (Hugo may already be familiar):
 - Alignments : BWA
 - SNP & InDel detection : GATK, SAMtools
 - SV & CNV detection :
 - Breakdancer (paired-end mapping)
 - CNVnator (read-depth analysis)
 - Pindel (split-read analysis)
 - BreakSeq (junction mapping)

Develop an error model based on hypothesized mechanisms of action

- Coverage model incorporating: GC-bias, alignment degeneracy, Poisson sampling
- SNP detection error model incorporating: Binomial distribution of het alleles, Raw read error rate, proximity to InDels (detected or not), clusters of SNPs, compressions, alignment degeneracy, coverage, SNP loci within SV's (esp large deletions), aligner & variant caller parameters, allelic biases. False positive, false negative & non-call rates.
- InDel error model, including zygosity
- SV error model, including zygosity (separate models for the multiple SV & CNV algorithms used, and then their combination)
- Can the model correctly reflect the differences between Illumina & CGI data sets on the same sample ?
- Can the model reflect errors in the reference sequence & guide its improvement ?

Create QC Tools to Measure Error-Model Parameters From Experimental Data Sets

- Determinants of coverage:
 - Coverage vs % GC content
 - Aligned coverage vs reference degeneracy (multiple parameters?)
 - Coverage histogram in regions selected to avoid repeats & %GC variation
 - Should match Poisson ?
 - Look for deviations from Poisson (e.g. palindromes)
 - Total coverage ratios of chromosomes (e.g. each chromosome vs total of Chr 1-22)
- Paired-end insert length distribution vs expected
 - Tails of the distribution may impact BreakDancer
- Raw read Q-score %-iles vs position along a read; vs expected
- Raw read error rates (monitored at homozygous loci)
 - In regions devoid of known problems (i.e. baseline)
 - Profiled across the genome, to look for problematic regions
- Allelic distribution of reads vs expected binomial (alignment, biochemistry problems)
 - In regions devoid of known problems (i.e. baseline)
 - Profiled across the genome, to look for problematic regions (allelic bias in the genomic sequence reads)
 - May be best with a very high coverage genome
- Sex, ethnicity, admixture and family relationships between samples vs expected from supplementary data provided
- Clusters of variants (not just SNPs) which are close / dense enough to predict higher probability of errors
- Amount and location of autozygosity, vs expectation
- Screening for non-human DNA (esp in data sets from saliva ?)

Analysis of experimental data to quantify error rates for all variant types & characterize by mechanism of action

- **Analysis of a single genome:**
 - Coverage assessment & errors related to that, including non-calls
 - Two alleles / haplotypes in single sex chromosomes, vs a standard
 - Three alleles / haplotypes in diploid chromosomes, vs a standard
 - Comparison of variant detection between multiple algorithms used in HUGO for the same thing:
 - SNP's & InDels : GATK & SAMtools
 - SV's & CNV's : Breakdancer, CNVnator, Pindel, BreakSeq
 - Aligned with the standard reference sequence vs ethnically specific major allele
- **Comparison of genomes:**
 - Independent subsamples from a single deep genome data set
 - Same genome sequenced twice (or more) with all the same platform settings
 - Same genome vs paired-end insert length
 - Same genome vs read-length
 - Mendelian Inheritance Errors (MIE) & Inheritance State Consistency Errors (ISCE)
 - Same genome on two platforms:
 - GA IIx vs HiSeq
 - HiSeq vs Complete Genomics
 - HiSeq vs array (genotypes and/or fluorescence for SV's)
 - HiSeq vs Sanger sequencing (Venter genome ?)
- **Tools to characterize apparent errors by mechanism of action:**
 - Segmentation by degeneracy of region
 - Segment by proximity to InDel, overlap with SV, SNP cluster (interferences)
 - Correlate with coverage, zygosity (all variant types), SV length
 - Characterizations across the genome vs at the Personalis list of medically interpretable loci / regions

Develop Strategy

- Combine results from all the experiments (& characterization by underlying mechanisms) to get an overall quantitative view of errors
- From that decide:
 - What to flag (& how – many options)
 - What to filter out (e.g. if always wrong)
 - What to fix in the lab
 - What to fix bio-informatically

Potential Technology Improvements (to be prioritized)

- Advances in third party aligners & variant callers
 - Aligner for > 200 base reads
 - New version of GATK
 - Others ?
- Advances in the reference sequence (admixed, InDels, other)
- Iterative alignment
- Alignment & variant calling in :
 - Degenerate regions
 - SNP / InDel clusters
 - SNP – InDel / SV overlaps
 - VNTR regions
- Combining Illumina & CGI genome data
- Optimal combinations of sample preps:
 - Insert sizes, read lengths
 - WGS + exome
 - Custom pullout
 - Arrays (several possibilities)
 - RNA-seq

Input needed to identify a prioritized list of medically interpretable variants

- Rong:
 - List of PharmGKB SNP' s
 - List of VariMed SNP' s (with MAF' s if poss) & note which contribute to Risk-O-Grams
 - List of PharmGKB VIP genes
 - List of 20+ cancer predisposing genes (p53, BRCA, etc)
 - List of Regulome regions (e.g. TF binding site ranges)
 - List of eQTL loci (just a representative set – doesn' t have to be comprehensive)
- Sarah / Gemma:
 - A representative list of variants connected to Mendelian disease
 - Maybe 20 of each variant type : SNP, InDel, Larger deletions spanning 10 bases – 100k bases, CNV' s, VNTR (I think we have an initial version thVNTR list), insertions, duplications, balanced translocations
 - List of highest priority Mendelian genes (e.g. CFTR, ...)
- John:
 - Variants for blood typing (I have an article on this I can provide)
- Rick Dewey :
 - Variants for tissue typing (this was done for the quartet paper and has since been revised by Rick in response to feedback that there are newer methods)

Genome Data Sets Needed & How to Get Them

- Quartet data sets – Hugo has them
- Quartet daughter HiSeq - Hugo has it for comparison with GA IIx data
- SRA HapMap NA12878 – Download priority emailed separately
- Other comparison data sets (e.g. platform comparison)
 - Complete Genomics vs Illumina underlying data sets from Lam paper (accessible ?)
 - Complete Genomics data set for NA12878
 - Array genotype data for NA12878 (e.g. HapMap III) – Can Mark download this ?
 - Is fluorescence data available from the HapMap site as well ? Useful for SV comparisons
- New data sets:
 - 2 x 250 bp data (BaseSpace) – Can Mark investigate ?
 - More complete families:
 - Search for 3-generation HapMap family data sets
 - Personalis-specific optimal combination of technologies

HugeSeq Pipeline Runs Requested

- Quartet for MIE / ISCE analyses
- Quartet daughter on HiSeq, for 1st pass GA IIx vs HiSeq platform comparison
- Multiple runs related to NA12878:
 - Same person sequenced three times by Broad on HiSeq from same library (Solexa-21431) at same depth (+/- 5%) :
 - SRX028061 (46.2-fold coverage)
 - SRX028032 + SRX028015 (30.0+17.4=47.4-fold coverage)
 - SRX028060 + SRX028015 (33.7+17.4=51.1-fold coverage)
 - Same person sequenced twice by Broad on HiSeq from different libraries of same PE insert length (Solexa-21431 & Solexa-21462):
 - Combine all 5 accession #'s for Solexa-21462 library (total 108.2-fold coverage)
 - Combine the 3 biggest accession #'s for Solexa-21431 for a total of 110-fold coverage
 - Alignment of one of the above (e.g. SRX028061) against an artificial reference with just the VNTR regions (e.g. Huntington's), with settings so if either read of a read pair aligns to the region, both reads of the pair are reported out. Those reads will then be used to test VNTR-specific alignment algorithms
 - Alignment of all eight NA12878 accession numbers, which have 6.5 to 40-k-base inserts, plus representative data sets from the other (<1kbase) insert lengths vs an artificial reference. This reference is to be constructed by Mark of degenerate regions of varying length & degeneracy, and various well documented compression regions which we think might be solvable using this data. This is a pullout technique. The reads pulled out this way will then be used to test algorithms for alignment in degenerate regions.

Key Skills & Bandwidth Needed

- To be filled in

Next steps in developing Gantt / Project Plan

- To be filled in

Accuracy Gantt Spreadsheet																									
OVERVIEW																									
Develop list of medically interpretable loci / regions for testing																									
Review literature																									
Develop error model based on hypothesized mechanisms of action																									
Create QC tools																									
Analysis of experimental data to quantify error rates for all variant types & characterize by mechanism of action																									
Develop strategy																									
Design, implement & test technology improvements																									
Release version x.x																									
DETAILS																									
Develop list of medically interpretable loci / regions for testing																									
Category													Status												
PharmGKB SNP's													Mark should have the whole list												
VariMed SNP's																									
All													Mark should have the whole list												
Just those which contribute to Risk-O-Grams													Have these been identified by Rong ?												
1st exon													Has Mark been able to do this categorization ?												
Other exons													Has Mark been able to do this categorization ?												
Non-exonic													Has Mark been able to do this categorization ?												
Other medically-interpretable variant types:																									
Small InDel's													Status ? Help needed ?												
CNV's													There is a database of meically interpretable CNV's. Does Mark have it ? Does it have well defined chromosomal coordiantes ? Zygocity / copy number associated with disease ?												
VNTR's													Mark has a list of 19(?). Exact chromosomal coordinates determined for each yet ? Are there more ?												
Deletions spanning 10 bases - 100kbp													Status ? Help needed ?												
Insertions													Status ? Help needed ?												
Duplications													Status ? Help needed ?												
Balanced translocations													Status ? Help needed ?												
Genes to assay confidently in their entirety :																									
PharmGKB genes, esp VIP													Mark should have all, & list of which are VIP												
Mendelian disease diseases (e.g. CFTR)													I think Mark has a small initial list - Can become more comprehensive as our Mendelian phenotype database grows, right ? Will that leave gaps ?												
													Compare with genotyping DTC vendor lists ? Compare with assay lists from Ambry, GeneDx, etc ?												
Cancer genes (e.g. p53, BRCA 1&2,...)													I think Mark has an anecdotal list. How comprehensive (should be 3-6,000 genes, at least, eventually)												
Variants required for blood typing													JW sent Mark link to NHGRI database. Did he get a chance to use it ? SV's, not just SNP's, right ?												
Variants for tissue typing													Did Mark get input from Rick Dewey ? Help needed ?												
Linc-RNA regions													Does Mark have URL from Broad web site (did JW send it ? Check)												
Regulome elements (TF binding sites)													Status ? Help needed ?												
eQTL loci													Status ? Help needed ?												
Loci / genes for which phasing is most likely													Status ? Help needed ?												
to be necessary for interpretation																									
Review literature																									
Category													Status												
Mark to review papers on HugeSeq elements, including new / upcoming versions																									
Alignments : BWA													Status ? Help needed ?												
SNP & InDel detection : GATK, SAMtools													Status ? Help needed ?												
SV & CNV detection :													Status ? Help needed ?												
BreakDancer (paired-end mapping)													Status ? Help needed ?												
CNVnator (read depth analysis)													Status ? Help needed ?												

	Pindel (split-read analysis)	Status ? Help needed ?
	BreakSeq (junction mapping)	Status ? Help needed ?
	Review other tools available for next-gen sequencing genome QC	Status ? Help needed ?
	Papers addressing new alignment methods (attend bioinformatics conference this summer)	Status ? Help needed ?
	Papers on hairpin strength / locations	
	Develop error model based on hypothesized mechanisms of action	
	Category	Status
	Coverage model incorporating :	
	GC bias	Exploration with smaller data sets, until whole genome 1-base-resolution coverage results can be routinely generated : GC coverage bias determined from old alignment of quartet data, found to differ more person to person than expected Assessment only done at 250-base chunk size. Noticed apparent coverage bias related to hairpins / palindromes, including those with surprisingly long loops SAM files generated for Chr1 of quartet - supports generation of some coverage data post-HugeSeq Coverage graphed at approx 250-base resolution based on SAM data - Showed surprising high coverage in degenerate Amylase region Coverage being determined at 1-base resolution over a limited range (1M-base) from SAM file using spreadsheet Tool for early study of GC bias mechanism: GC bias at differing spatial scales Possible palindrome / hairpin mechanism (also requires method to identify all palidromes) Comprehensive modelling using whole genome 1-base-resolution coverage results routinely from multiple genomes: Is there systematic coverage variation (seen across multiple genomes) not accounted for by mechanisms in the model ? If exploratory analysis shows palindrome / hairpin coverage mechanism, need to generate genome-wide palindrome map
	Alignment degeneracy	Exploration with smaller data sets, until whole genome 1-base-resolution coverage results can be routinely generated : Post BWA coverage assessed in one quartet genome in one degenerate region (Amylase) showed: Coverage substantially above that of the surroundings in part of the region, about 1/2 normal in the rest Coverage levels appear erratic in degenerate region, though not examined very closely yet These results were a little surprising & so will need some exploration to understand & develop a model for Exploration planned to understand this: Examine coverage in whole quartet, other genomes as well (NA12878 & others) Identify apparently het loci in these regions: Look at allelic fractions & compare with binomial expectation, for insight on structure Look for potential to link these apparently het loci by paired-ends of various insert sizes Use data from NA12878, sequenced with very long insert paired-end reads, to reduce degeneracy & link apparently het SNP loci Look for Mendelian inheritance of coverage in the quartet, esp at apparently het loci, in this degenerate region Examine other degenerate regions, perhaps simpler tests cases (e.g. clean 2x degenerate region) Can Mark identify some simple / clean test regions to study ? (Easiest if on Chr 1 for now) Use of third party tools to generate coverage data or visualize it, from BAM files (IGV 2.0, SAMtools, others Christian or Nan Leng may know) Comprehensive modelling using whole genome 1-base-resolution coverage results routinely from multiple genomes: Comprehensive assessment of BWA-aligned coverage in degenerate regions
	SNP / InDel clusters	Mark's hypothesis : reads may not be aligned if # SNP loci / read length exceeds aligner mismatch allowance and occur together Mark has suggested this might be fixed by BreakSeq-like approach Mark has dbSNP loci list from Rong, but no phasing data Michael Clark thought a quite recent 1,000 Genomes data release included phasing - Mark to follow up with Michael. Possible quick anecdotal test using phased SNP's observed in the quartet ? Anticipate that this effect, if it exists, would lead to allelic bias at het SNP's, biased in favor of reference sequence alleles (check for this along with coverage) Start by looking at clusters of just SNPs, then expand to combinations with InDels ?
	Poisson sampling	Not studied yet

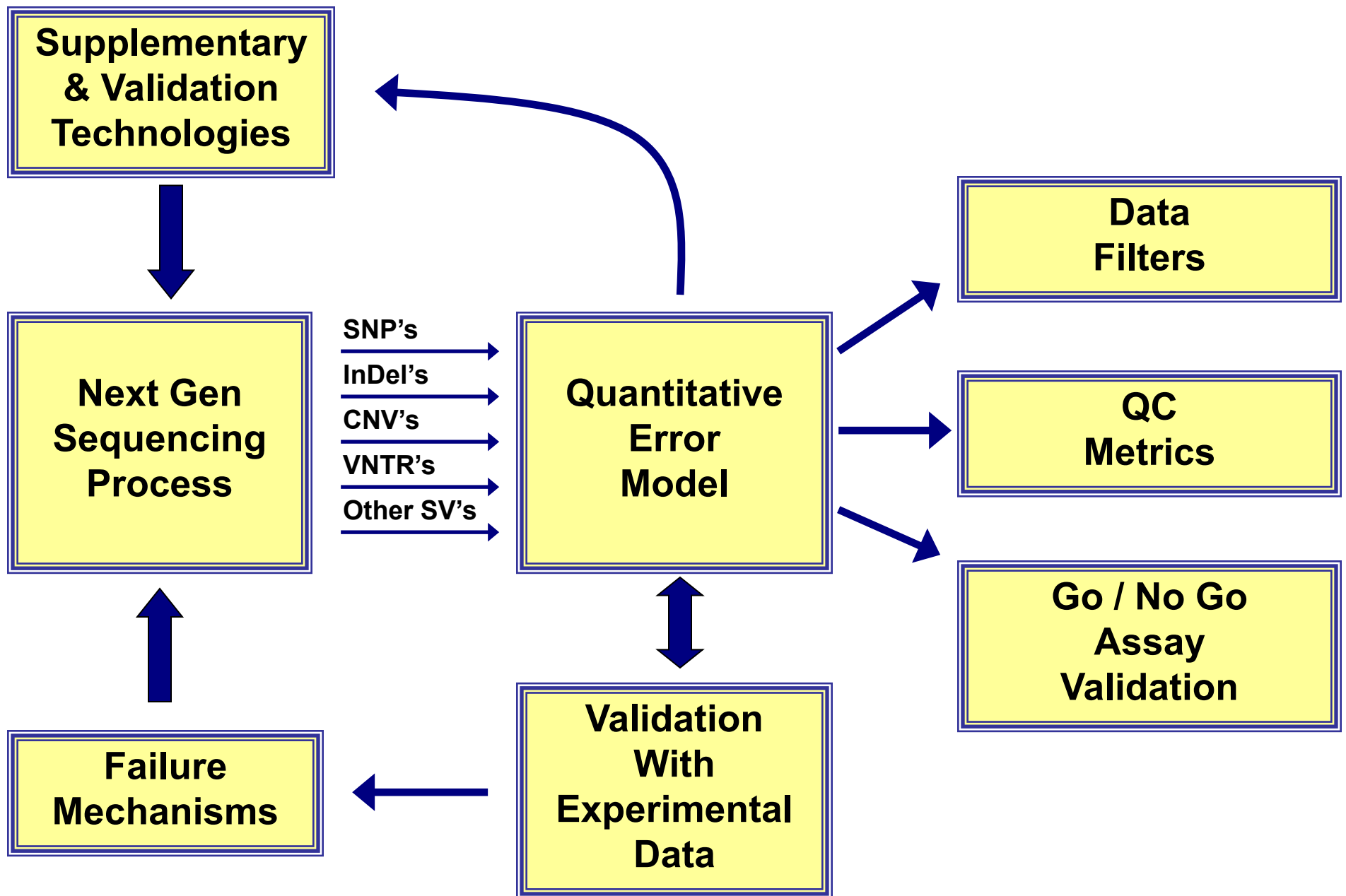
	Coverage	A very basic model of SNP detection, based on overlapping binomial probability distributions, was developed. It showed a huge increase in error probability at low coverage While a much more sophisticated SNP detection error model could be developed, I think this conclusion would likely remain This can be tested once we have SNP MIE data and can correlate it with coverage (requires 1-base resolution coverage across the genome) If a problematic sequence (e.g. a homopolymer) creates a systematic depression in raw Q-scores in a region, it may be thought of as lowering the "effective coverage". The studies described above, on Q-score depression around problematic sequence motifs, should illuminate this.
	SNP loci within SV's (esp large deletions)	The mechanism leading to SNP allele calling errors in SV's (at least large deletions) is clear. To fix this, Hugo is adding: Determination of SV zygosity to HugeSeq SV overlap with SNP alleles reported (e.g. "A-", instead of an incorrect "AA") The real challenge here is getting accurate SV detection (see below) and making sure our medical interpretation can use calls like "A-" Modelling : We need a simple model of MIE generation due to unrecognized deletions (and more generally SV's) Can Mark overlap MIE's with SV's (both from the quartet data) to test this mechanism ?
	Aligner & variant caller parameters	Need to complete this part of the literature review before we can model this & test it experimentally
	False positive, negative & non-call rates	False positive & negative rates could be determined from the binomial model but given the systematic depression of raw Q-scores around certain sequence motifs, would be too low Systematic raw Q-score depression around these motifs likely needs to be understood, and applied to the whole genome, before we can make projections of the rates of the various error types Non-call rates would be determined by the coverage or consensus allele Q score cutoff, whichever is used (discuss with Hugo - this is a HugeSeq setting)
	InDel error model, including zygosity	The binomial SNP error model has been extended to detection of InDels by the split-read method. Conclusions: False negative (i.e. missed) het InDel's should decline significantly with read length (e.g. 100 to 250 bases) False negative (i.e. missed) het InDel's are more coverage-sensitive than SNP's A BreakSeq approach to InDels could help, particularly when we have MAF data for the most common ones New approaches, e.g. local re-alignment around any potential InDel detected in a small number of reads, need to be understood in detail This is part of the literature review yet to be completed (GATK & SAMtools advances in InDel detection)
	SV error model (including zygosity) :	Generally needs data from SV MIE analysis, focused on mechanisms underlying these errors, before we can do very much useful on modelling
	Split reads	Split-read SV detection error modeling should be similar to InDel detection by split reads (discussed above)
	Read-pairs	A graph of the moving average (& moving median) read-pair spacing for experimental data derived from HugeSeq (a SAM file) was generated and compared to HugeSeq SV results, to look for examples and identify underlying error mechanisms Mechanisms found this way could be used as the basis for an error model This will require some focused time with Hugo A hypothesized mechanism, based on electrophoretic GC bias of fragment length, was not borne out by the experimental data described above
	Read depth	The read-depth algorithm seems anecdotally to be the source of a lot of SV MIE This may be based on systematic coverage variation due to GC bias and difficult sequence motifs (discussed above) Modelling of this SV detector's error mechanism should probably wait until after we make progress on the coverage model We had discussed a concept of building up a database of coverage values over the genome, to use as a reference The variation in GC bias across the four quartet genomes (which we might have expected would be very similar) is a warning that coverage while it clear has some systematic variation, may be more variable person to person than previously appreciated. This concept may need to be on hold until we make more progress on the coverage model
	Junction mapping	Junction mapping SV detection error modeling should be similar to InDel detection by split reads (discussed above) Experimental data on mechanisms underlying these errors, and their frequencies, would be helpful before putting much effort into modeling
	Combination of the above	We need progress understanding (and reducing) the error mechanisms for the individual SV detection algorithms before we can work on their combination
	Can the model reflect :	
	Illumina vs CGI differences ?	Eventually, but we need to make progress on just one platform, which is fairly open and for which we have good access to raw data, before moving on to the other & comparison
	Guide how/whether to combine them ?	Ditto
	Errors in the reference sequence ?	Potential areas include:
	Guide how to improve it ?	SNP clusters (see work on these, above) High MAF InDels (need MAF data, e.g. from 1,000 Genomes, but also need better algorithms to apply to that data, and maybe more sequencing depth) Compressions (see work on these, above)

[illegible]

Framework for assessment & improvement of accuracy in whole genome medical interpretation

March 5, 2012

John West



Template for assessment of genome sequencing accuracy

Variant Type	# Expected Per Genome	MIE & ISCE		Same Genome Twice	
		# Errors	% Errors	# Differences	% Differences
SNP	2,800,000				
InDel	400,000				
Large Del	800				
VNTR	5,000				
CNV	500				
Balanced Translocation	50				
Other SV	200				

Other Error Metrics:

Loci Tri-Allelic in a Genome :

bases insufficient aligned coverage (<15-fold) :

Others

List of Variants to Optimize Against

- Medically interpretable variants (not just SNP's !):
 - Mendelian “Top 200”
 - Specific loci / variant types, where known
 - Genes in which to identify novel variants
 - PharmGKB
 - Curated & novel variants
 - Alleles in top Risk-O-Grams
 - Diseases
 - Molecular or physiologic medically relevant phenotypes
 - Blood & tissue typing
- Functional elements (broad interest for discovery research)
 - RefSeq genes
 - Linc-RNA
 - Regulome elements
 - eQTL loci

Proposed Initial Priorities (Jan 5th)

- Literature review related to genome sequencing accuracy, followed by discussion with SAB
- Develop an initial list of top priority medically interpretable variants to optimize against
- Create / collect lists of problematic regions of the genome
- Screen target medical variant list against problematic regions to get 1st cut scale & composition of the accuracy problem
- Develop initial quantitative model of allele determination, including known error mechanisms. Predict & characterize accuracy & repeatability levels
- Obtain existing experimental data, reprocess with current version algorithms, and conduct 1st cut repeatability & accuracy assessment. Compare with model on a genome-wide basis and at the medically interpretable loci.
- Launch program to obtain long-lead-time experimental data, including prototype of “maximally accurate genome” product (based on NGS augmented by other technologies).

Near-Term Experiments to Assess Accuracy Levels Currently Achieved

- Mendelian Inheritance Errors & Inheritance-State Consistency errors from a family of four: SNP's, InDels, Large Deletions, VNTR's, CNV, other SVs
 - Ethics committee supports technical analysis of quartet data as long as no medical interpretation by Personalis
 - Hugo has all the files to begin processing
- Repeatability (sequence the same person twice)
 - Quartet daughter was sequenced twice (GA IIx and later HiSeq)
 - Quartet son & daughter (both GA IIx) same over $\frac{1}{4}$ of genome
 - HapMap NA12878 sequenced many times by Broad on HiSeq: download
 - Mike Snyder sequenced several times (same platform & different)
- Problems determinable from individual genomes
 - No aligned coverage (or not enough to call reliably)
 - Loci tri-allelic in an autosome, or bi-allelic in a haploid X or Y
 - Many QC readouts (see QC list)
- Key is to quantitatively characterize these and understand the underlying mechanisms

Near-Term Testing of Potential Solutions

- Iterative alignment
 - Use the consensus sequence from one pass of alignment as the reference sequence against which to align on a 2nd pass
- Pull out reads of degenerate region:
 - Use BWA and an artificial “reference” to pull out reads which align, even if not uniquely, to a degenerate region (e.g. Amylase)
- Pull out reads of VNTR regions:
 - Use a similar approach to pull out reads in VNTR regions (e.g. Huntington’s) to see if there are reads in the poly-Q region which we might custom-assemble / align.
- “Ground truth” for Amylase region:
 - Solve the Amylase region using NA12878 data from 10 different PE insert lengths – test case

Elements of Analytic Validity in CDC's ACCE Model

Key is Quantitative Assessment

- Analytic Sensitivity : How often positive when mutation is present ?
 - One minus the false negative rate ?
- Analytic Specificity : How often negative when mutation is not present ?
 - One minus the false positive rate ?
- Assay Robustness : How often does the test fail to give a usable result ?
- Repeatability :
 - On the same sample
 - Within & between labs
 - Between sample types & other process variables
- Confirmatory testing to resolve false positives ?
 - Genetic validation tests
 - Non-genetic tests (molecular, imaging, physiological)
- Quality Control: Internal QC program defined & externally monitored ?

Crucial to develop a model which links analytic validity results to monitor-able process variables

- Allele determination performance is known to vary widely between loci
- With millions of known loci, many quite rare, it is impractical to determine the assay performance for each as an independent system.
- Novel loci, inherently, cannot be validated in advance
- Proposed approach:
 - Create a quantitative process model of allele determination
 - Incorporate known process variables (sample type, storage & preparation, coverage, read length, repeat structure, algorithm parameters, etc)
 - Augment with models of process variability & failure mechanisms
 - Instrument the model with QC observables, related to normal & failure modes
 - Compare model predictions with real data & characterize differences
 - Iterate until model can confidently predict analytic validity metrics for both known & novel alleles, including in combination
 - System performance metric should prioritize medically interpretable alleles
 - Use insights from experimentally assessed error model to :
 - Evaluate / prioritize process improvements
 - Select disease-associated alleles which can be assayed with NGS⁺

Genetic variant types to include in model

- SNP's
- InDel's
- Larger deletions
- STR / VNTR's
- Structural variants
- Copy Number Variation
- Trisomies
- Combinations:
 - Any of these can be heterozygous
 - Compound heterozygosity
 - Small variants within larger het deletions or other CNV's
 - Clusters of closely spaced variants, which might interfere with each others' detection

QC Readouts, to monitor the process

(Particularly given that it may be impractical to standardize all process variables completely, over multiple years)

- Paired-end insert length distribution vs expected
- Coverage non-uniformity vs % GC content; relative to a standard
- Coverage histogram in regions selected to avoid repeats & %GC variation
- Total coverage ratios of chromosomes
- Raw read Q-score %-iles vs position along a read; relative to a standard
- Allelic distribution of reads vs expected binomial, at a “diagnostic” set of het SNP loci
- Sex, ethnicity, admixture and family relationships between samples vs expected from supplementary data provided
- Raw read error rates (monitored at homozygous loci), systematic / random
- Mendelian inheritance errors (family genome data sets)
- Clusters of variants which are close / dense enough to predict higher probability of errors
- Two alleles / haplotypes in single sex chromosomes, vs a standard
- Three alleles / haplotypes in diploid chromosomes , vs a standard
- Amount and location of autozygosity , vs a standard
- Screening for non-human DNA
- Concordance with genotyping data, vs a standard

Lists of problematic regions in the genome

Capture all in one place, unify, & link to underlying error mechanisms of the model

- Mapability / alignability assessments (some on Santa Cruz browser site)
- Genes with known CNV's
- Genes with known pseudogenes
- List of problematic genes for NGS (e.g. VWF)
- Evan Eichler's list of problematic regions
- 1,000 Genomes list of regions with top 0.1% of coverage
- List of compressions
- MIE's & MIE-cluster regions from family data
- Difficult regions from phasing of a family
- Regions of high raw read error rates (monitor at homozygous loci)
- Loci apparently tri-allelic in the genome of a healthy individual, or 3 haplotypes
- Regions of allelic bias
- Regions of no or low aligned coverage, determined experimentally
- Segmental duplications & other repeat structures

Categories of experimental data to assess error types, mechanisms, frequencies

- Repeatability :
 - Same person twice (or more ?), trying to keep everything else constant
 - Alternatives: Genomes of identical twins, or children where identical
- Controlled process experiments :
 - Same person sequenced multiple times, each time varying just one parameter of the process (many of these can be done computationally)
- Concordance of other technology platforms on the same sample
- Compare process QC metrics from many sources of raw genome data, even low coverage (e.g. 1,000 Genomes)
- Mendelian inheritance errors (MIE's) determined from family genome sets
- Absolute accuracy :
 - Genomes of individuals known to have clear Mendelian diseases
 - Sequencing of complex synthesized oligo sets (synthetic genome)
 - Targeted modifications of DNA from a naturally occurring genome

Examples of known error mechanisms

Many more likely to follow

- Locally low / no coverage due to:
 - GC bias of biochemistry
 - Alignment degeneracy (repeats)
 - Clusters of variant alleles exceed alignment mismatch allowance
 - Alignment poor due to problems with reference sequence
- Reads align, but incorrectly
 - InDels missed, or inconsistently placed in homopolymers
 - Misplacements lead to :
 - Phantom het SNP's
 - Apparent tri-allelic loci
 - Apparent triple haplotype structure
- Raw read errors, random or systematic
- Allelic imbalance of reads due to:
 - Random extremes of expected binomial distribution, esp where low coverage
 - Allelic biases - biochemical or bioinformatic (e.g. clusters of SNP's in LD)
- Read length too short to quantify length of VNTR, or VNTR embedded in other repeat structures

Examples of process variability

Many more likely to follow

- Sample source (blood, saliva, tissue, cell culture)
- Sample preservation, culturing & prep for sequencing
- DNA sequencing platform, lab & operator(s)
- Coverage
- Read length
- Paired-end insert length (a distribution)
- Cluster creation & Sequencing chemistry versions
- Cluster density
- Variation in data processing algorithms, input parameters, reference sequence
- DNA contamination or incomplete filtering of primer-related artifacts
- Availability of supplementary data from the same sample (e.g. genotyping)
- Inconsistent supplementary information from / about the individual sequenced

Proposal for near-term accuracy plan

For coordination with rest of team

John West
May 16, 2012

Overall Goal & Proposed Timing

- Identify Specific Mechanisms Leading to Errors
 - Ideally leading to a calibrated quantitative model which can predict the rate of each error type vs process parameters we can control
 - Identify all regions subject to each error type of error, for potential flagging (even if error not actually seen there)
 - Quantify the absolute rate of errors by mechanism
 - Tool to prioritize our approaches to solving errors
 - Develop approaches to reduce error rate in a deterministic fashion
 - Lab method choices, e.g. level of coverage, combination of insert lengths...
 - Improvement in algorithms
 - Optimization of reference sequence
 - etc
- 
- Complete 1st pass by end of June
- Complete 1st pass by end of August

5 Categories of Methods to Identify Errors

Each Provides a Different View

Each is Discussed on a Subsequent Slide

- Inspection of individual aligned genomes
 - Zero coverage, tri-allelic/triple haplotype loci in a single person, het's on a single X or Y, clusters of depressed raw Q-scores, development of a coverage model
- Same person sequenced twice, using normal methods
 - Errors due to: random variation (Poisson), systematic process variables (coverage, GC-bias, degeneracy vs PE-length, platform-specific error modes)
- Comparison vs a “Gold Standard”
 - Minimize known error sources (Poisson by huge coverage, degeneracy by multiple insert lengths, etc)
 - Provides visibility of the next level of error indicators (e.g. reads not 50/50 at het loci)
- Mendelian Inheritance Errors
 - Identify systematic false positive & false negative errors by use of specific inheritance states
- Population statistics / SNP databases
 - Leverage work already done by others and orthogonal methods
 - dbSNP problematic SNP loci, violation of Hardy-Weinberg in GWAS, loci with > 2 alleles in a population (esp 2 different rsID's e.g. A/C & A/AAA)

Inspection of Individual Aligned Genomes

- Work with genomes sequenced using new/recent technology
 - CEPH1463 families sequenced by Illumina, in SRA
 - Recent Stanford faculty genomes ?
- Zero (or very low) coverage (Hugo/pipeline or BAM crawler)
- Non-call regions (Hugo)
- Tri-allelic in a single individual (Chr 1-22 & XX) (BAM crawler)
- Het in single X & Y chromosomes (Pipeline ?)
- Clusters of depressed raw Q-scores (BAM crawler)
 - Work to identify underlying sequence motifs & calibrate strength of Q-score depression vs motif metrics (e.g. length of homopolymer)
 - Identify these motifs in the reference (potentially including hairpins)
- Development of a coverage model
 - Explore using smaller regions & then test at genome scale
- Develop a set of standard QC reports for each genome
 - Use publicly available where possible

Same Person Sequenced Twice

NA12878 data from Broad via SRA

- Same sequencing library at same coverage (MP already organized)
- Same sequencing library, progressively increasing coverage
- Different sequencing library, same coverage (MP already organized)
- Different paired-end insert lengths, same coverage
- Platform comparisons:
 - ILMN / CGI on NA12878
 - Genotyping array data ?
 - Fluorescent array data (Mike Snyder, Hugo)

Gold Standard Genome(s)

- Develop NA12878 Gold Standard Genome in 2 Stages:
 - 500x coverage, all short PE inserts
 - Best possible results with algorithms most sensitive to coverage (e.g. split reads and BreakSeq)
 - Cleanest possible detection of alignment degeneracy via allelic balance of # reads at het SNP's (not 50/50)
 - Cover regions which normally get little (but not zero) coverage
 - Combining longer inserts (> 1 k-base)
 - Solving degenerate & compressed regions
- Comparisons vs Gold Standard
 - Does this identify different errors vs “Same Person Sequenced Twice” ?
 - Helpful to understand ILMN / CGI differences ?

Mendelian Inheritance Errors

- West quartet
 - Original SNP MIE's obtained from Rick Dewey
 - Overlapped with VariMed
 - Loci being investigated for underlying mechanism
 - Repeating SNP analysis with modern HugeSeq and adding InDel analysis requires "Union" of VCF's
 - SV's started manually, suggests SV algorithms need optimization
 - Gold Standard more useful for that ?
- CEPH1463 / CGI data set
 - Already downloaded, practical in Excel
- Link MIE rates to overall error rates – disconnect ?

Population Statistics / Public Databases

Based on Discussion with Rong

- SNP loci labeled in dbSNP as problematic
 - About 200k loci
 - Various labels
 - Rong has the data
- Loci with > 2 alleles in the population
 - Can be real, but likely enriched for errors
 - Rong suspicious of loci with multiple rsID's (e.g. A/C & A/AAA)
 - Rong can easily create this list
- SNP's which do not meet Hardy Weinberg in GWAS studies based on genotyping arrays
 - Rong has contacted Carlos

Progression of Genome Data Sets for Accuracy Studies

Founder Genomes



Downloading



Custom Sequenced

West quartet (2009)

MIE analysis:

SNP (Dewey)

SV's (Manual, WG)

InDel (TBD)

Anecdotal error studies on
portions of the genome:

Phenomena

Error rates

Underlying mechanisms

GA IIx vs HiSeq (early 2011)

HapMap NA12878

Broad data set (late 2010):

Hi-Seq 2x100

10 PE insert lengths

Total coverage > 500

Same person "twice":

Same library

Different library

High coverage

SV validation (Snyder)

Fluor. Array data

HapMap genotypes

Some relatives' data available

CEPH1463

3 generations, 17 people

Mother is HapMap NA12878

To be "gold std"

All 17 sequenced by CGI

Test-bed for Personalis max
accuracy combination of
technologies

Use to assess other tech's:

Ion Proton

Oxford Nanopore

MiSeq 2x250

Anonymous samples (ethics)

No Personalis relationship

Snyder son/mother

Data behind publications:

Illumina vs CGI

Snyderome paper

Some inheritance validation

Korean PGP

40 HiSeq 2x100 genomes

Venter Genome

Only individual whole genome
on Sanger technology

Overview of Personalis Accuracy Program

Characterize errors

Analysis

Rates by variant type & Q-score
Underlying mechanisms
Medical interpretation impact
Regulatory impact

Methods

Comparison with literature
Studies on limited regions
Whole genome analysis:

- Single genome metrics
- Same person twice
- Family inheritance (MIE)
- Platform comparison
 - Arrays
 - ILMN / CGI / Ion T...
 - Sanger capillary

Prioritize solutions

Bio-informatic approaches
Laboratory approaches
Content approaches

Informatic Solutions

Advances in 3rd party alg's
Aligner for > 200bp reads
New versions, current alg's
Other alg's (e.g. lobSTR)

Advances in our alg's:
SV's : Zygosity, accuracy
Combine SV's & SNP's
Broader use of Breakseq

New Personalis alg's
Alignment degeneracy
Genome-aware alignment
VNTR-specific

Advances in reference
Ethnic major-allele ref's
Iterative
Admixed
InDel's / SV's

Flag or Filter Out
For problems not yet solved

Lab, Content Solutions

Combine technologies
WGS + exome
Longer reads and/or inserts
Multi-vendor
Arrays (genotypes & SV's)
Electrophoretic sizing
Sanger validation

Content-based semi-custom
Array (genotypes & SV's)
Pullout

Custom sample preps
Problematic motifs
Degenerate regions
Phasing

Content
Gold standard genomes
Known healthy genomes

Determining SNP Error Rate

Quartet data

CEPH1463
CGI data

NA12878 SRA data

CEPH1463
family new
data

Process with
modern
HugeSeq (done)

MIE Analysis

Link MIE rate to
error rate & type

VCF's aligned
across four
family members,
including non-
calls, coverage,
allele Q-scores

MIE Analysis

Link MIE rate to
error rate, type,
mechanism

Comparisons

Same library, two
runs

Same protocol,
two libraries

Broad HiSeq vs
CGI (vs array ?)

Max quality
alignment by
combining 10
paired-end insert
lengths & very
high coverage;
Then compare
individual runs to
this standard

Single genome

Where is
coverage
insufficient ?

Where are
allele Q-
scores too
low ?

Tri-allelic /
triple
haplotype
sites

Allelic bias
(requires
very high
coverage)

Comparison
ILMN / CGI /
Genotyping

MIE path like
quartet

ARTICLE PREVIEW

[view full access options](#) ▶

NATURE BIOTECHNOLOGY | RESEARCH | ANALYSIS

Performance comparison of whole-genome sequencing platforms

Hugo Y K Lam, Michael J Clark, Rui Chen, Rong Chen, Georges Natsoulis, Maeve O'Huallachain, Frederick E Dewey, Lukas Habegger, Euan A Ashley, Mark B Gerstein, Atul J Butte, Hanlee P Ji & Michael Snyder

[Affiliations](#) | [Contributions](#) | [Corresponding author](#)

Nature Biotechnology 30, 78–82 (2012) | doi:10.1038/nbt.2065

Received 01 September 2011 | Accepted 15 November 2011

| Published online 18 December 2011



FACULTY^{of}1000
POST-PUBLICATION PEER REVIEW

Isaac Kohane, Harvard Medical School, USA. **F1000 Physiology**

11 Jan 2012 | Controversial, Confirmation

8 MUST READ



This comparison of two whole-genome sequencing (WGS) efforts on two different platforms is a reminder of the hazard of whole-genome clinical sequencing today. The (lack of) overlap between the results of the two sequencing efforts (on the same genome) at the single nucleotide level but also in copy number variants serves as a cautionary note regarding this growing incidentalome {1}.

com/13440958

2

Performance comparison of whole-genome sequencing platforms. - F1000

In comparing the variants found on sequencing on an Illumina platform and by Complete Genomics, over 10% of the SNPs were not concordant (i.e. hundreds of thousands of SNPs called by one platform and not the other). For copy number variants, the story is even more dramatic, more than 70% of the 800,000 CNVs were NOT concordant.

Technology-specific error signatures in the 1000 Genomes Project data

Michael Nothnagel · Alexander Herrmann · Andreas Wolf · Stefan Schreiber ·
Matthias Platzer · Reiner Siebert · Michael Krawczak · Jochen Hampe

Hum Genet
DOI 10.1007/s00439-011-0971-3

Received: 4 October 2010 / Accepted: 10 February 2011

Table 5 Estimated proportion, α , of false SNV detections, in percentages
95% confidence intervals are given in parentheses.
1000 Genomes Project Pilot 2 data

	NA12878
	All SNVs
454 FLX TM	6.3 (6.1–6.5)
GA IIx TM	8.4 (8.0–8.7)
SOLiD TM	17.1 (16.9–17.4)

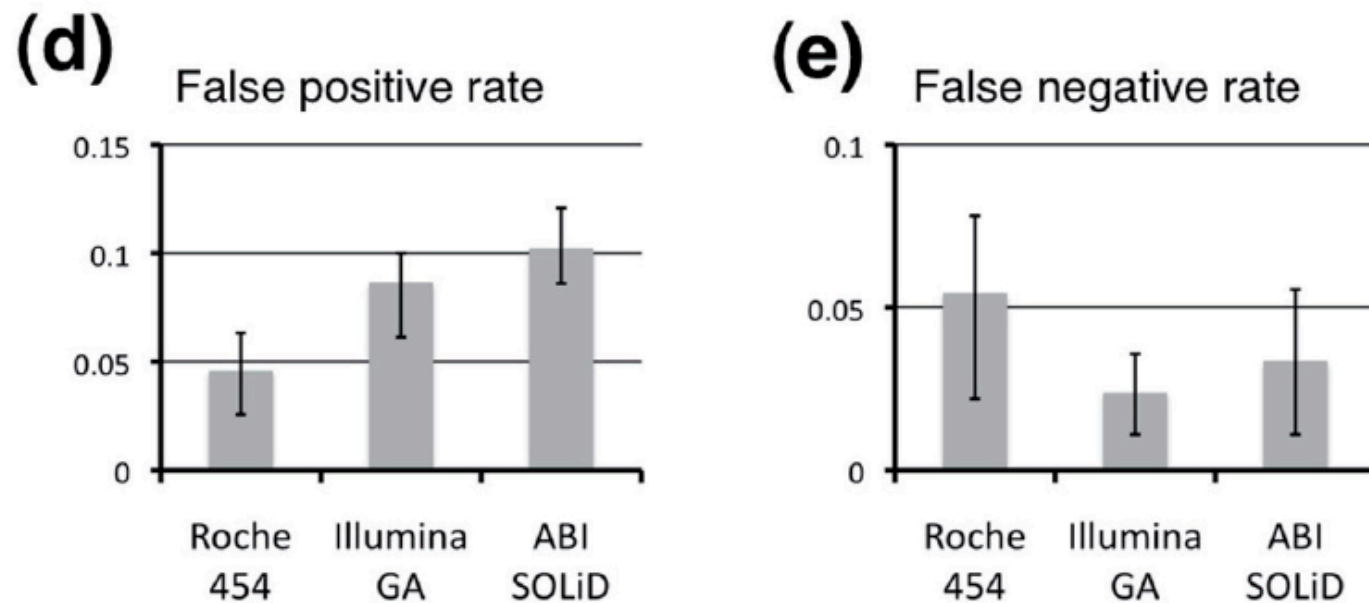
Estimated 8.4% of SNV's on
Illumina Genome Analyzer IIx are
false positive or false negative.
Quality Scores for these SNV's
were reported as 126.7 +/- 50.9

Evaluation of next generation sequencing platforms for population targeted sequencing studies

Olivier Harismendy^{✉*}, Pauline C Ng^{✉†}, Robert L Strausberg[†], Xiaoyun Wang^{*}, Timothy B Stockwell[†], Karen Y Beeson[†], Nicholas J Schork^{*}, Sarah S Murray^{*}, Eric J Topol^{*}, Samuel Levy[†] and Kelly A Frazer^{*}

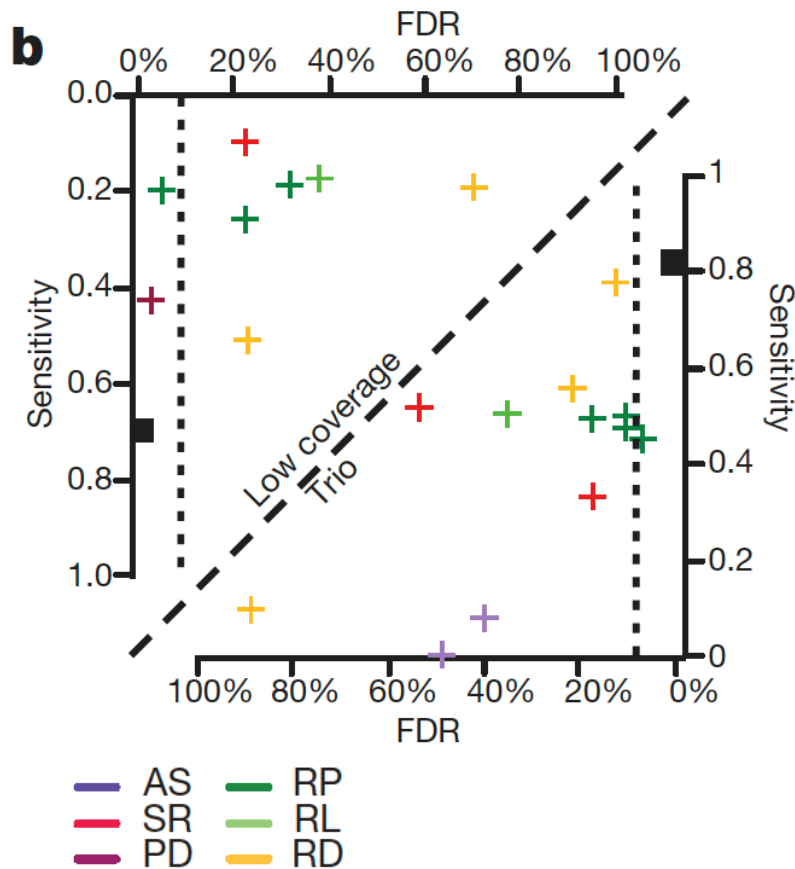
Genome Biology 2009, **10**:R32 (doi:10.1186/gb-2009-10-3-r32)

Published: 27 March 2009



False positive & negative rates for the Illumina Genome Analyzer system were 8% & 2% (Data published in 2009)

False positive & false negative rates for various approaches to structural variant detection.



AS = Assembly
RP = Read pair
SR = Split Read
RL = Read pair with long (>1kbp) inserts
RD = Read depth
PD = RP + RD

Trio = High coverage

Figure 2 | Comparative assessment of deletion discovery methods.

Mapping copy number variation by population-scale genome sequencing

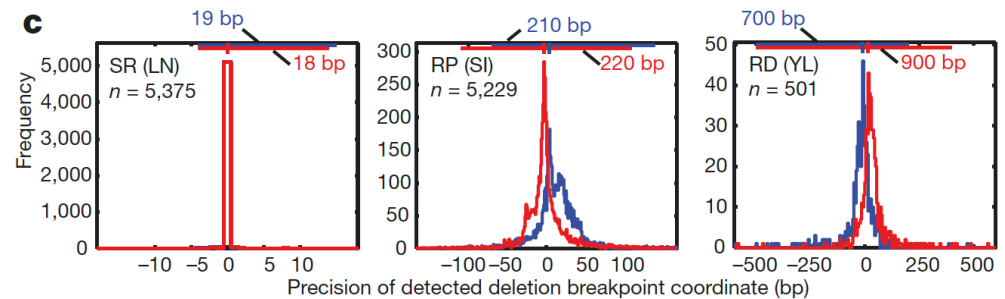
Ryan E. Mills^{1*}, Klaudia Walter^{2*}, Chip Stewart^{3*}, Robert E. Handsaker^{4*}, Ken Chen^{5*}, Can Alkan^{6,7*}, Alexej Abyzov^{8*}, Seungtae Chris Yoon^{9*}, Kai Ye^{10*}, R. Keira Cheatham¹¹, Asif Chinwalla⁵, Donald F. Conrad², Yutao Fu¹², Fabian Grubert¹³, Iman Hajirasouliha¹⁴, Fereydoun Hormozdizadeh¹⁴, Lilia M. Iakoucheva¹⁵, Zamin Iqbal¹⁶, Shuli Kang¹⁵, Jeffrey M. Kidd⁶, Miriam K. Konkel¹⁷, Joshua Korn⁴, Ekta Khurana^{8,18}, Deniz Kural³, Hugo Y. K. Lam¹³, Jing Leng⁸, Ruiqiang Li¹⁹, Yingrui Li¹⁹, Chang-Yun Lin²⁰, Ruibang Luo¹⁹, Xinmeng Jasmine Mu⁸, James Nemesh⁴, Heather E. Peckham¹², Tobias Rausch²¹, Aylwyn Scally², Xinghua Shi¹, Michael P. Stromberg³, Adrian M. Stütz²¹, Alexander Eckehart Urban^{13,27}, Jerilyn A. Walker¹⁷, Jiantao Wu³, Yujun Zhang², Zhengdong D. Zhang⁸, Mark A. Batzer¹⁷, Li Ding^{5,22}, Gabor T. Marth³, Gil McVean²³, Jonathan Sebat¹⁵, Michael Snyder¹³, Jun Wang^{19,24}, Kenny Ye²⁰, Evan E. Eichler^{6,7}, Mark B. Gerstein^{8,18,25}, Matthew E. Hurles², Charles Lee¹, Steven A. McCarroll^{4,26}, Jan O. Korbel²¹ & 1000 Genomes Project†

Authors include:

Mike Snyder, Hugo Lam, Mark Gerstein

1000 Genomes Project

3 FEBRUARY 2011
VOL 470 | NATURE | 59

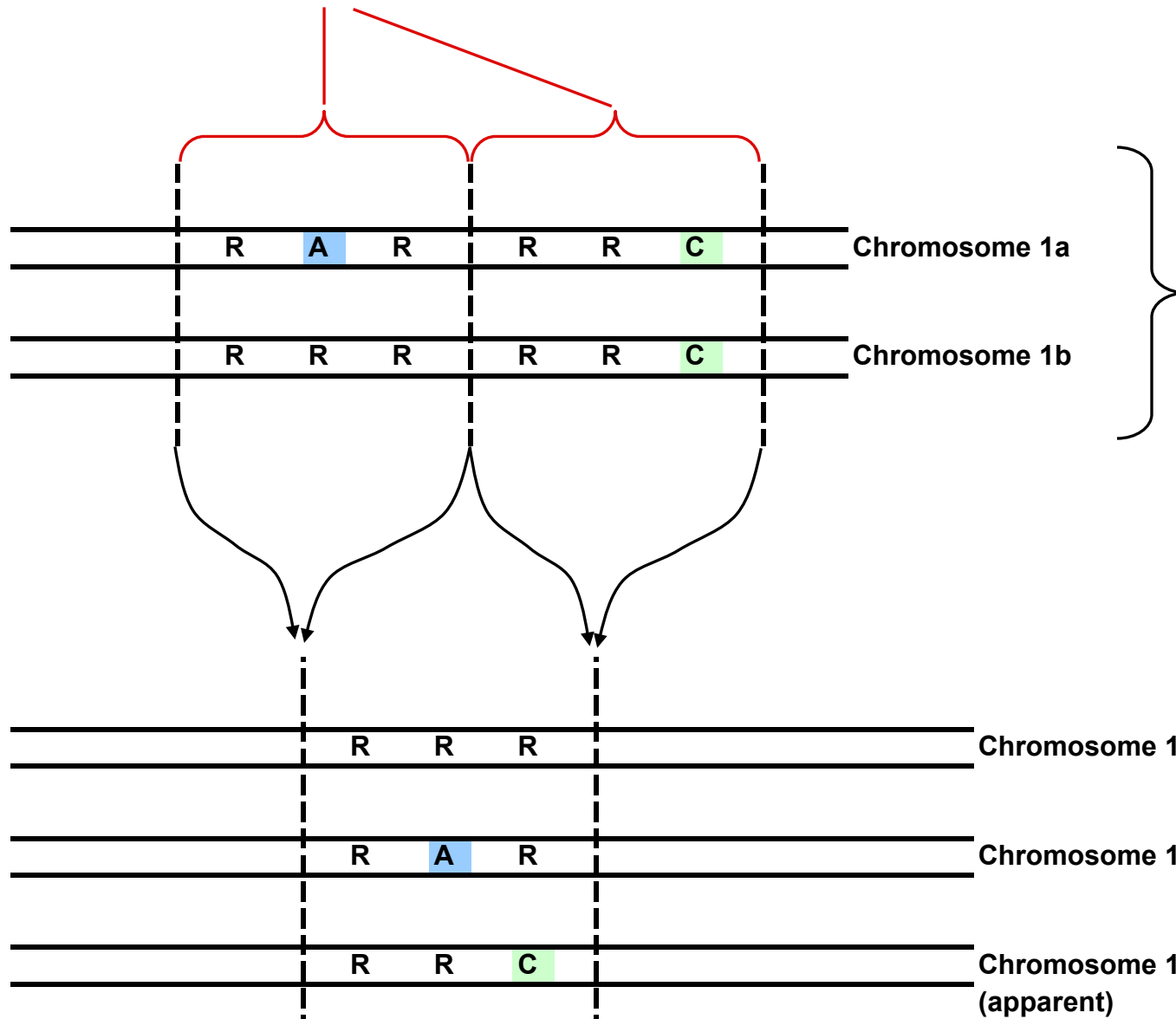


Structural variants:

False positive & negative rates are in the 10's of %
Position / length precision, except for breakpoint methods, is in the 100's of bases

Concept of a 2x Compression

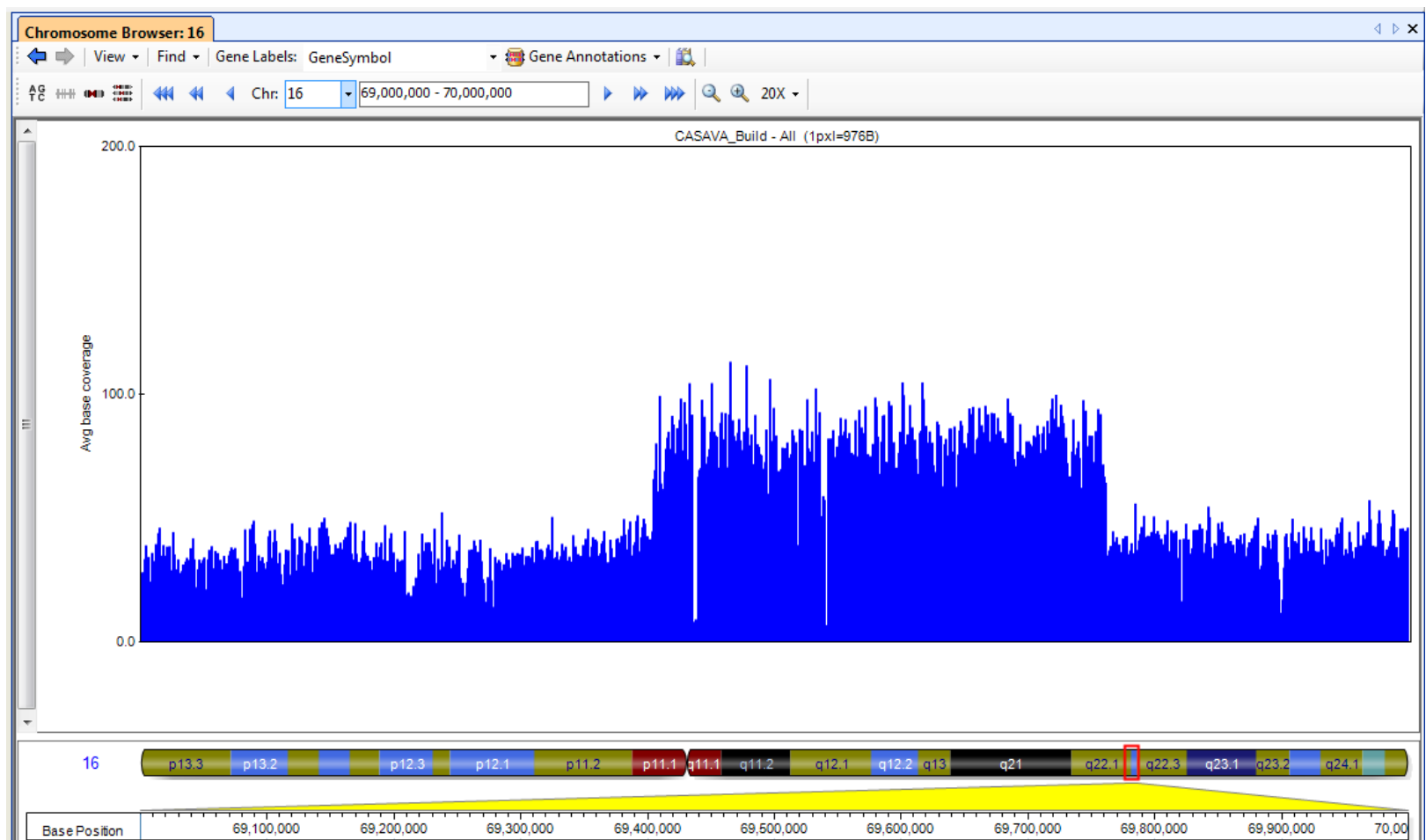
2 copies of region, not
recognized in the reference sequence



Actual diploid DNA has
two copies of region,
with differences at SNP
loci. If the duplication is
old, many SNP loci will
become fixed
(homozygous)

Sequenced
Sample,
Mapped to
Reference,
Appears
Tri-ploid

358 k-base 2x
Compression
& genes in the
region.

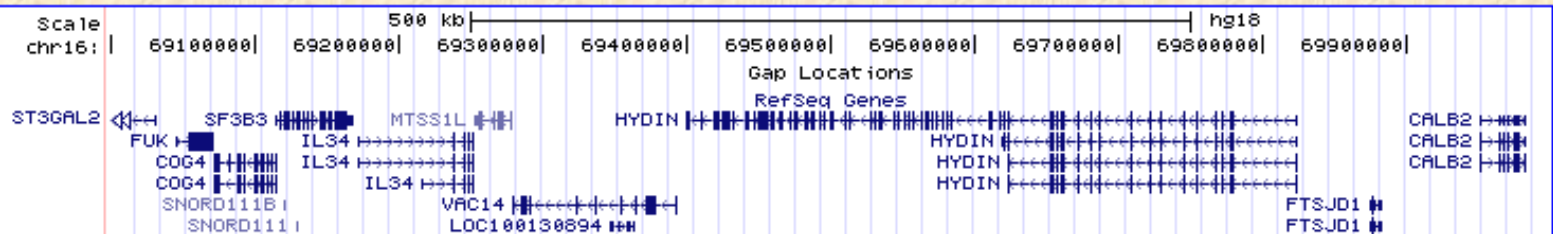


UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

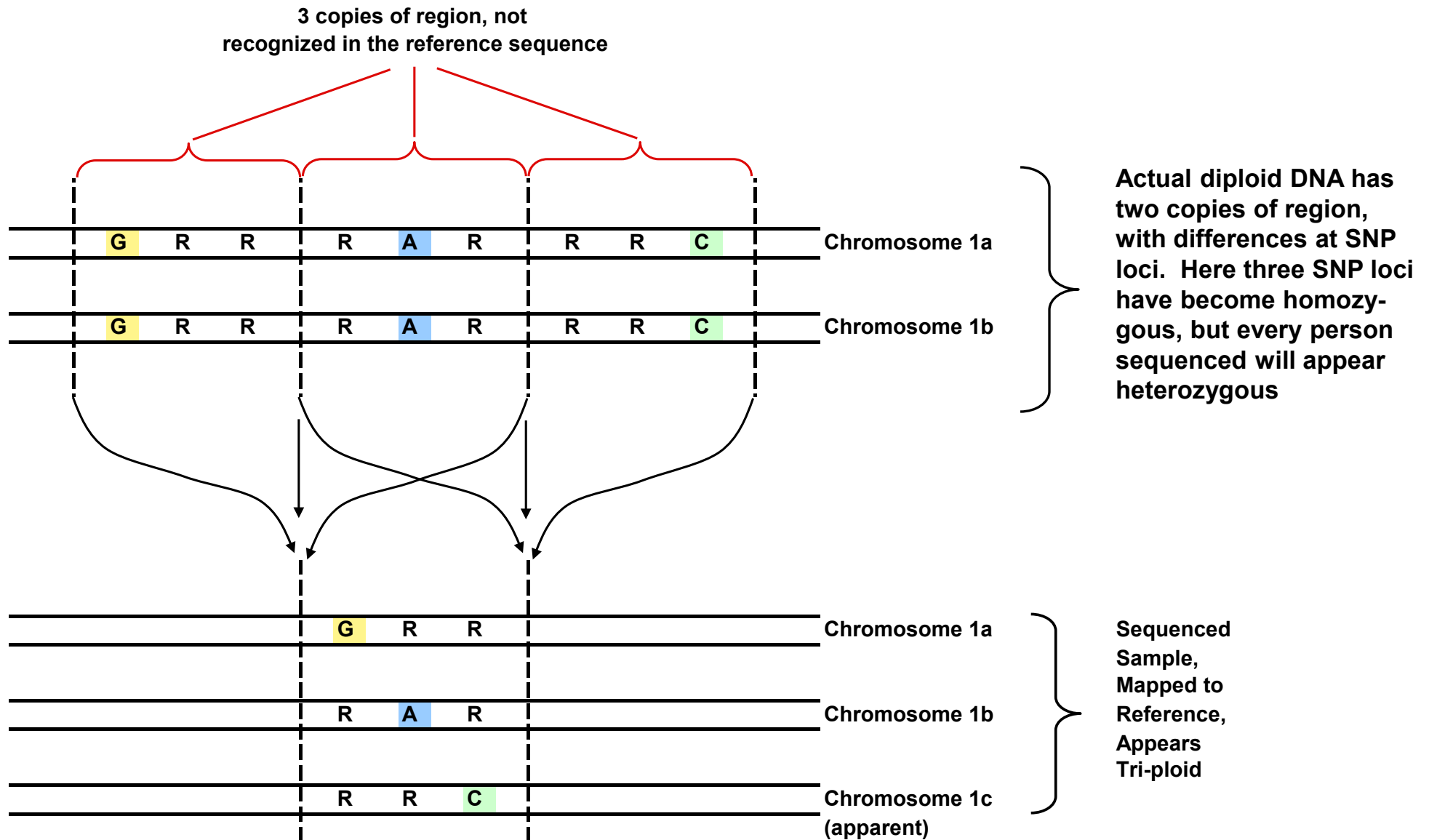
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr16:69,000,000-70,000,000 gene jump clear size 1,000,001 bp. configure

chr16 (q22.1-q22.3) p13.3 p13.2 p12.3 p12.1 p11.2 p11.1 q11.1 q11.2 q12.1 q12.2 q13 q21 q22.1 q22.3 q23.1 q23.2 q24.1

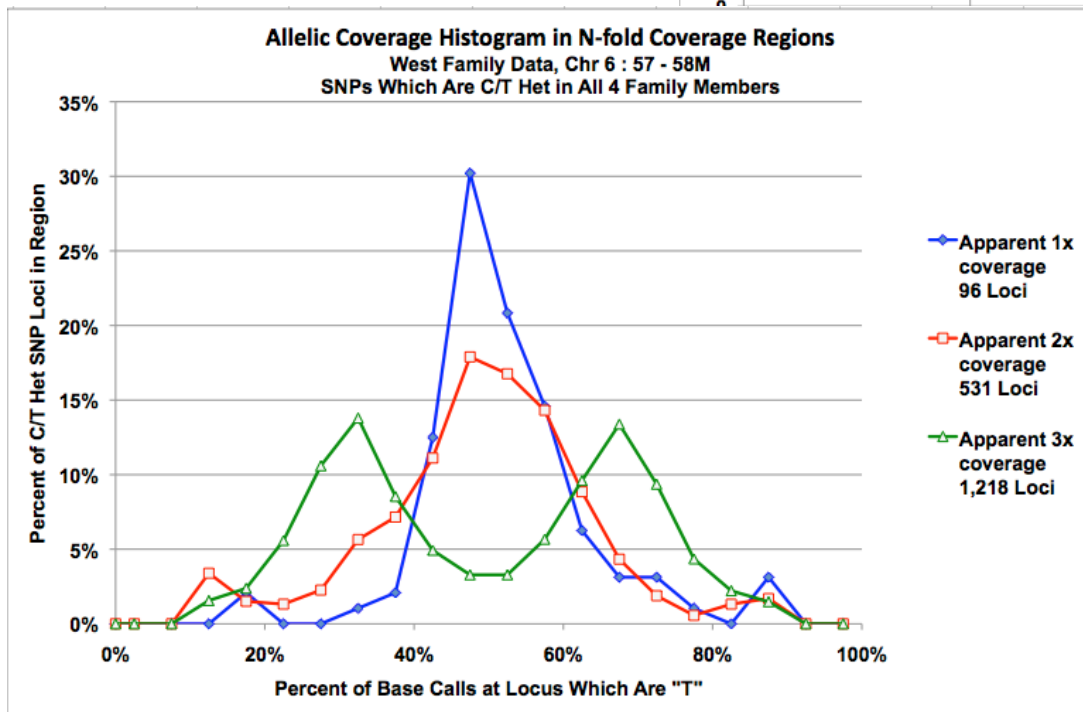
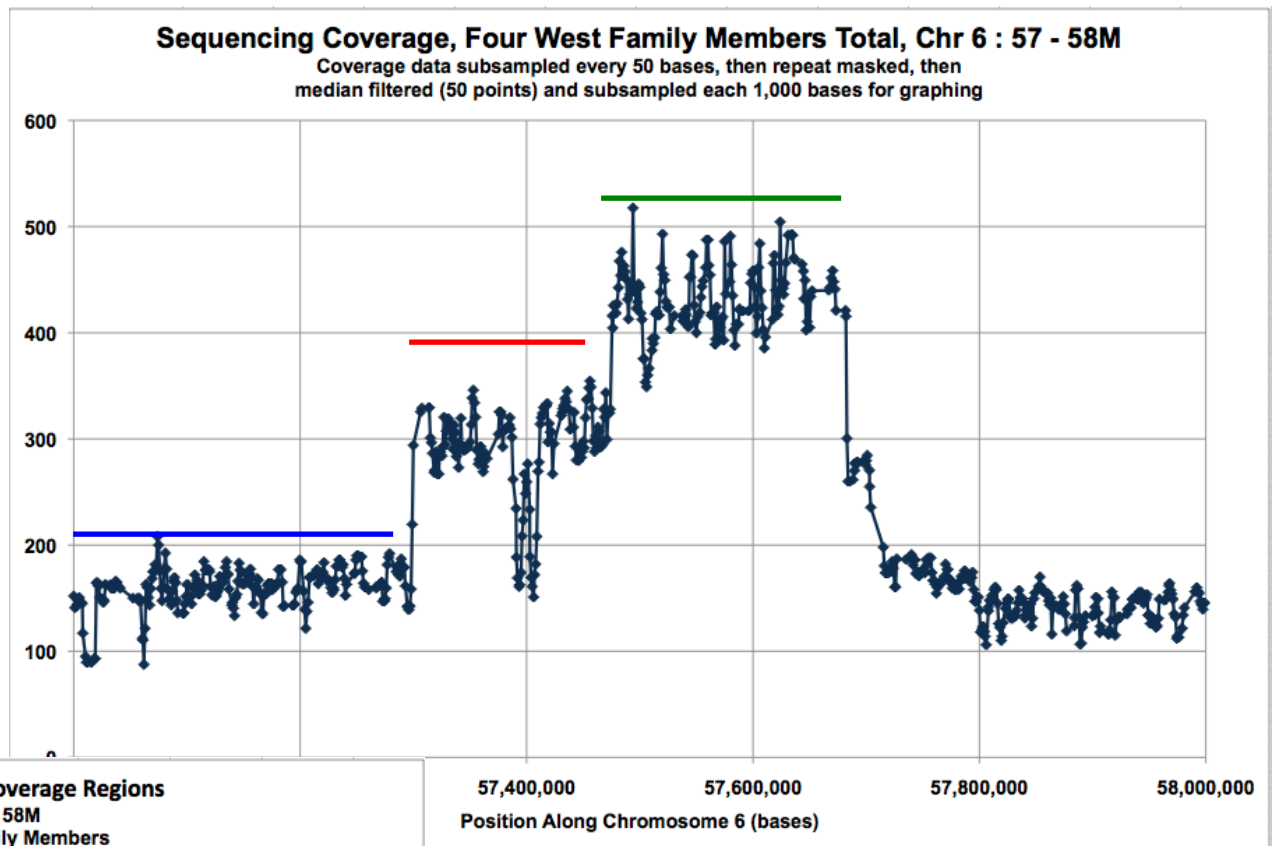


Concept of a 3x Compression



Compound 2x / 3x Compression

Almost 400,000 bases long
Encompasses PRIM2 gene.



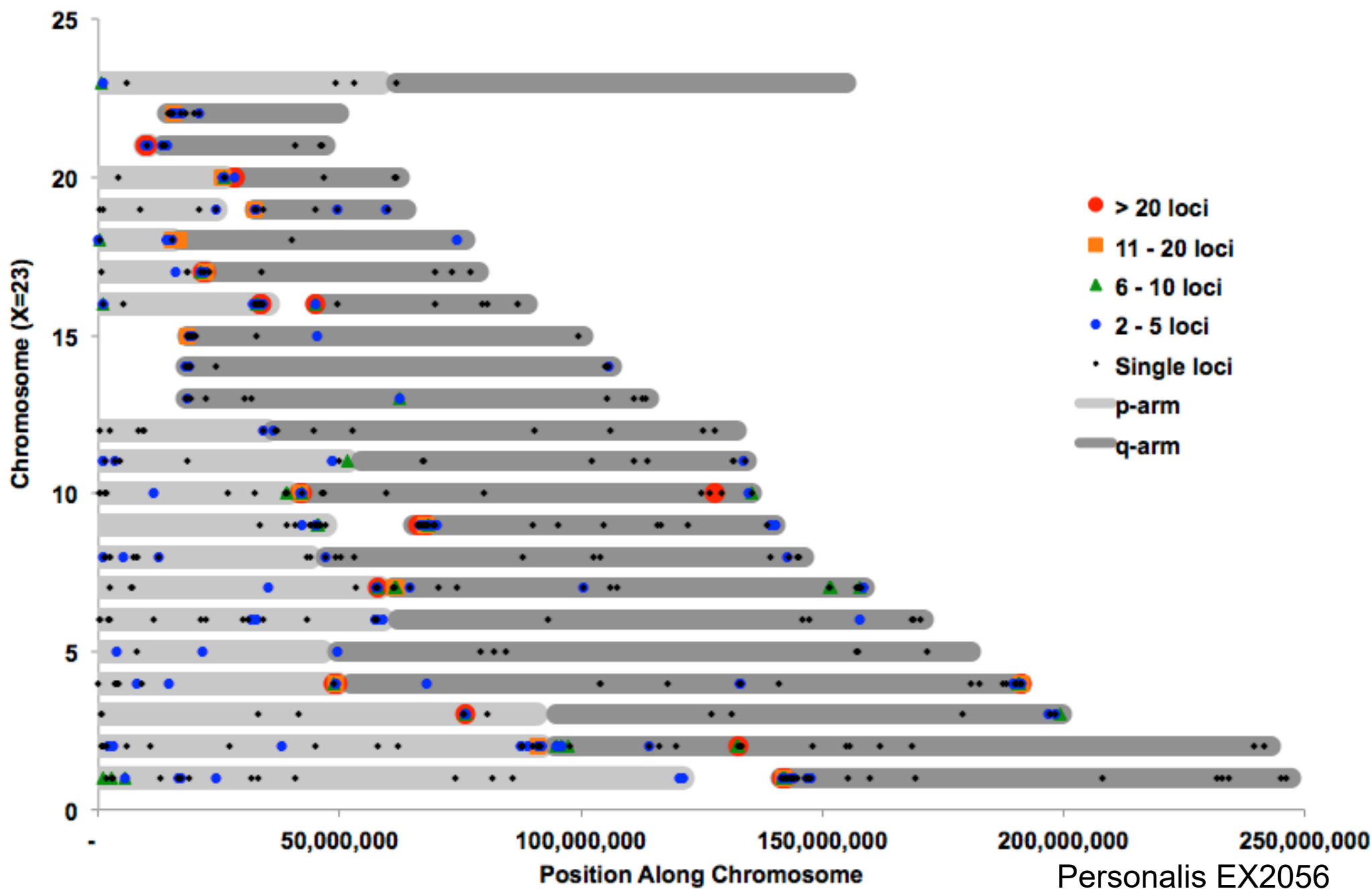
Distribution of base calls in 1x & 2x
vs 3x regions shows “triploid”
characteristic in 3x region.



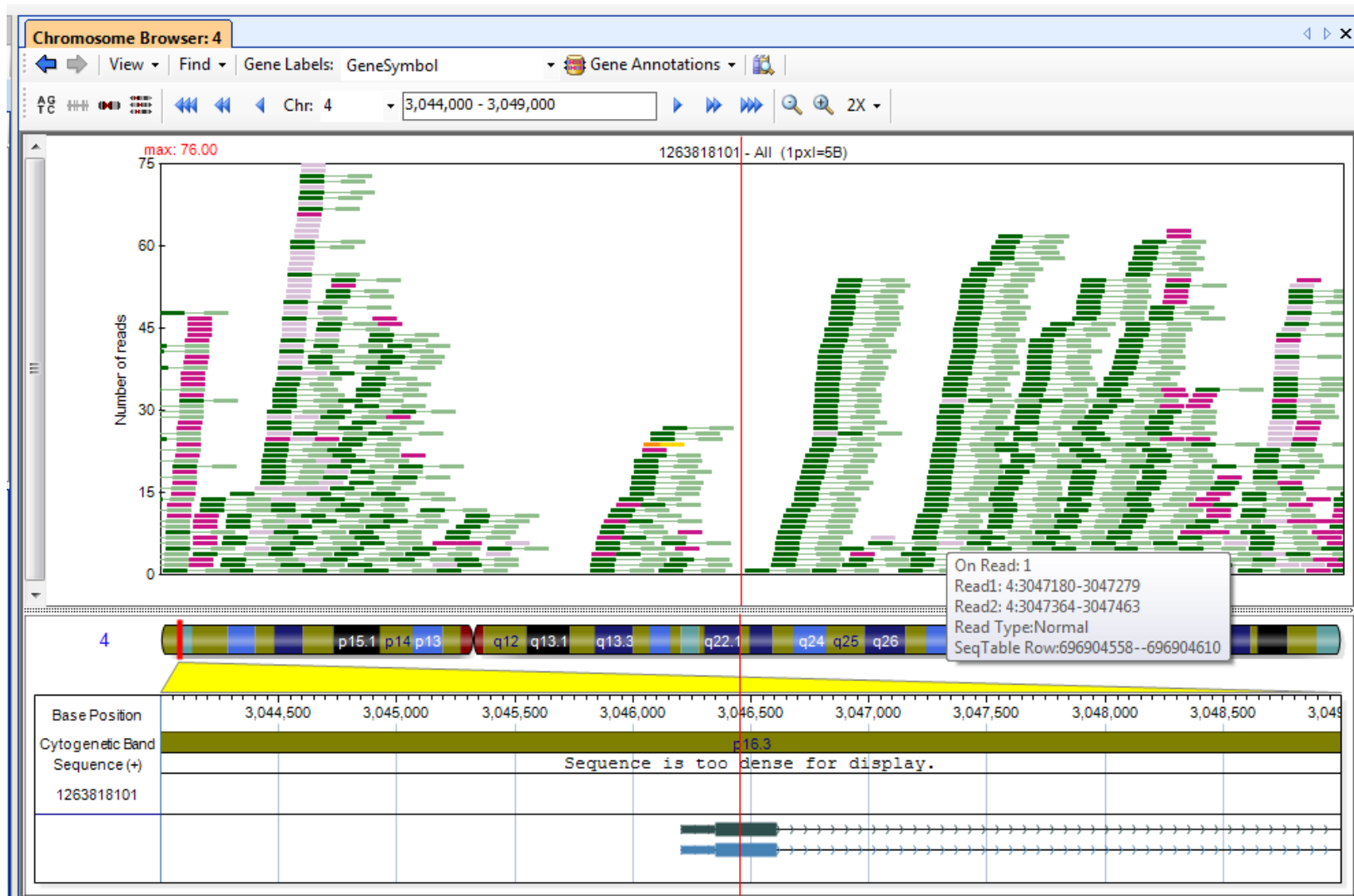
Apparent triple haplotype in a compression

This set of reads has good coverage and seems to be cleanly aligned, but the four SNP loci cluster into three haplotypes in a single individual : GGGT, GACC and AACC. The individual is not triploid. Evidence suggests this represents an ancient 3-fold duplication not recognized in the reference sequence. Interpreting genes in these regions requires special care to avoid high confidence false positive conclusions.

Clusters of Loci Tri-Allelic in the Genome Sequence of a Single Individual



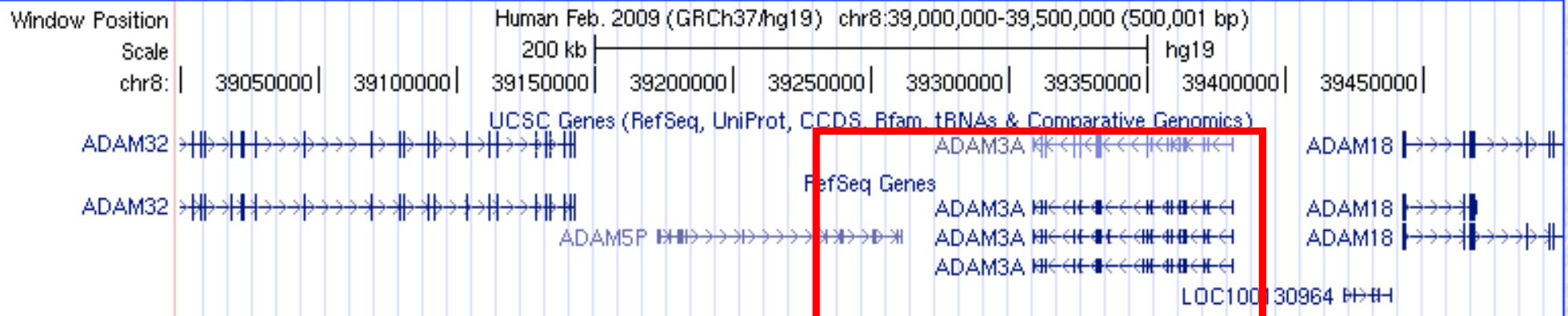
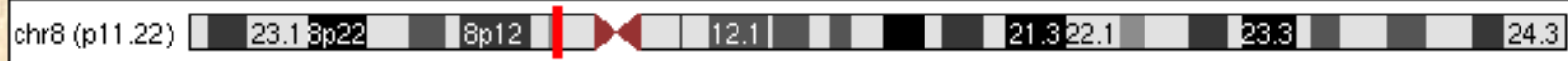
Zero DNA Sequencing Coverage at the Huntington's Disease locus



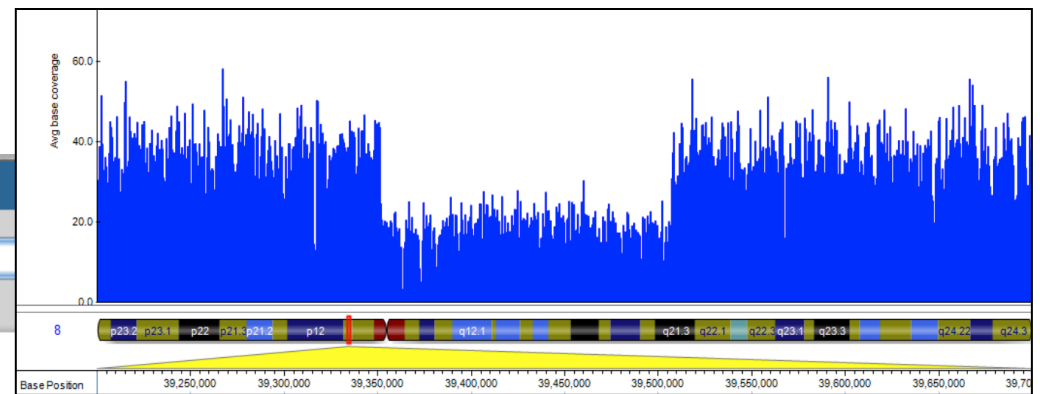
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr8:39,000,000-39,500,000 [gene](#) jump clear size 500,001 bp. configure



Deletion spans 155k bases but SV algorithms don't report zygosity



NCBI Resources How To

PubMed.gov

US National Library of Medicine
National Institutes of Health

PubMed

Limits Advanced

Display Settings: Abstract

Neuro Oncol. 2011 Feb;13(2):212-22. Epub 2010 Dec 7.

Homozygous loss of ADAM3A revealed by genome-wide analysis of pediatric high-grade glioma and diffuse intrinsic pontine gliomas.

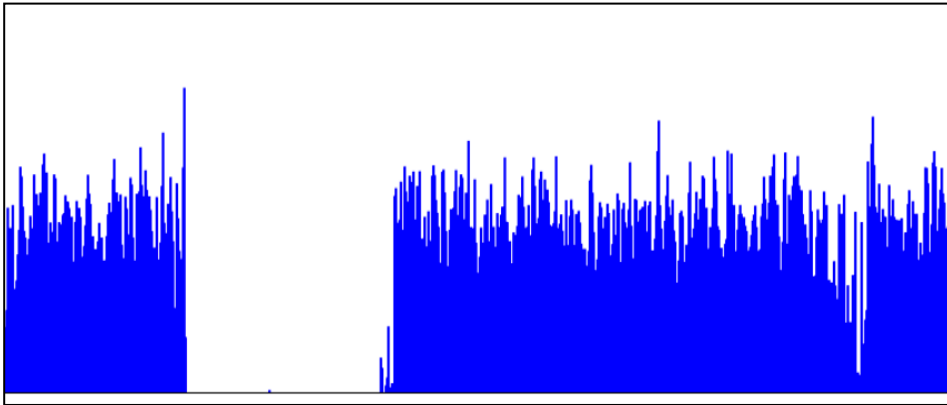
Barrow J, Adamowicz-Brice M, Cartmill M, MacArthur D, Lowe J, Robson K, Brundler MA, Walker DA, Coyle B, Grundy R.

Children's Brain Tumour Research Centre, School of Clinical Sciences, Queen's Medical Centre, University of Nottingham, Nottingham NG7 2UH.

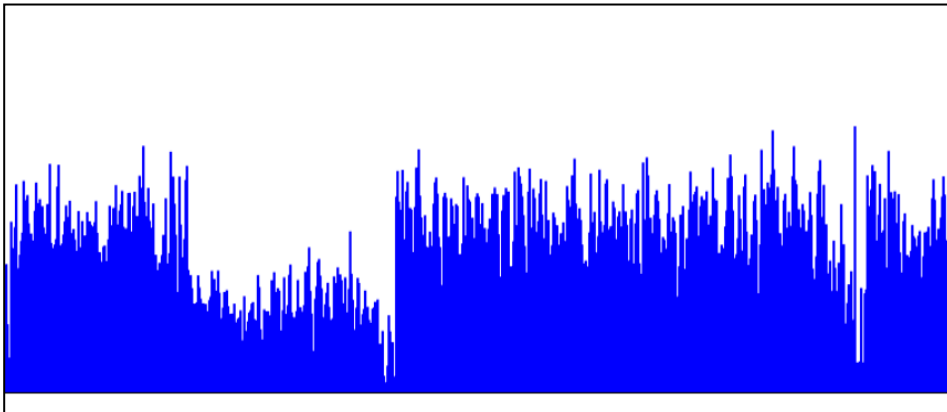
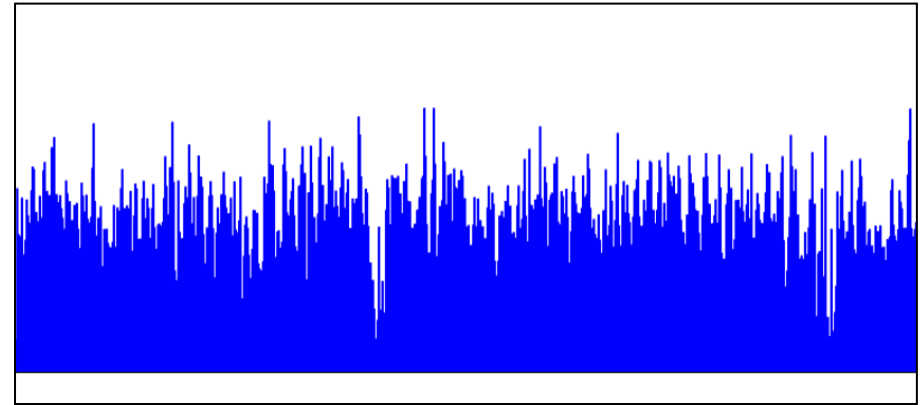
Family data confirms HugeSeq SV detection

Deletion detected on Chr 1 : Length 32,239 both homozygous states & heterozygosity

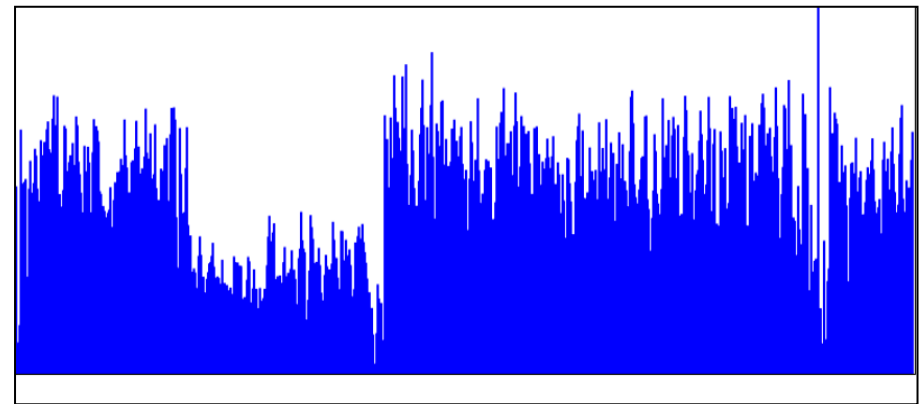
Father



Mother

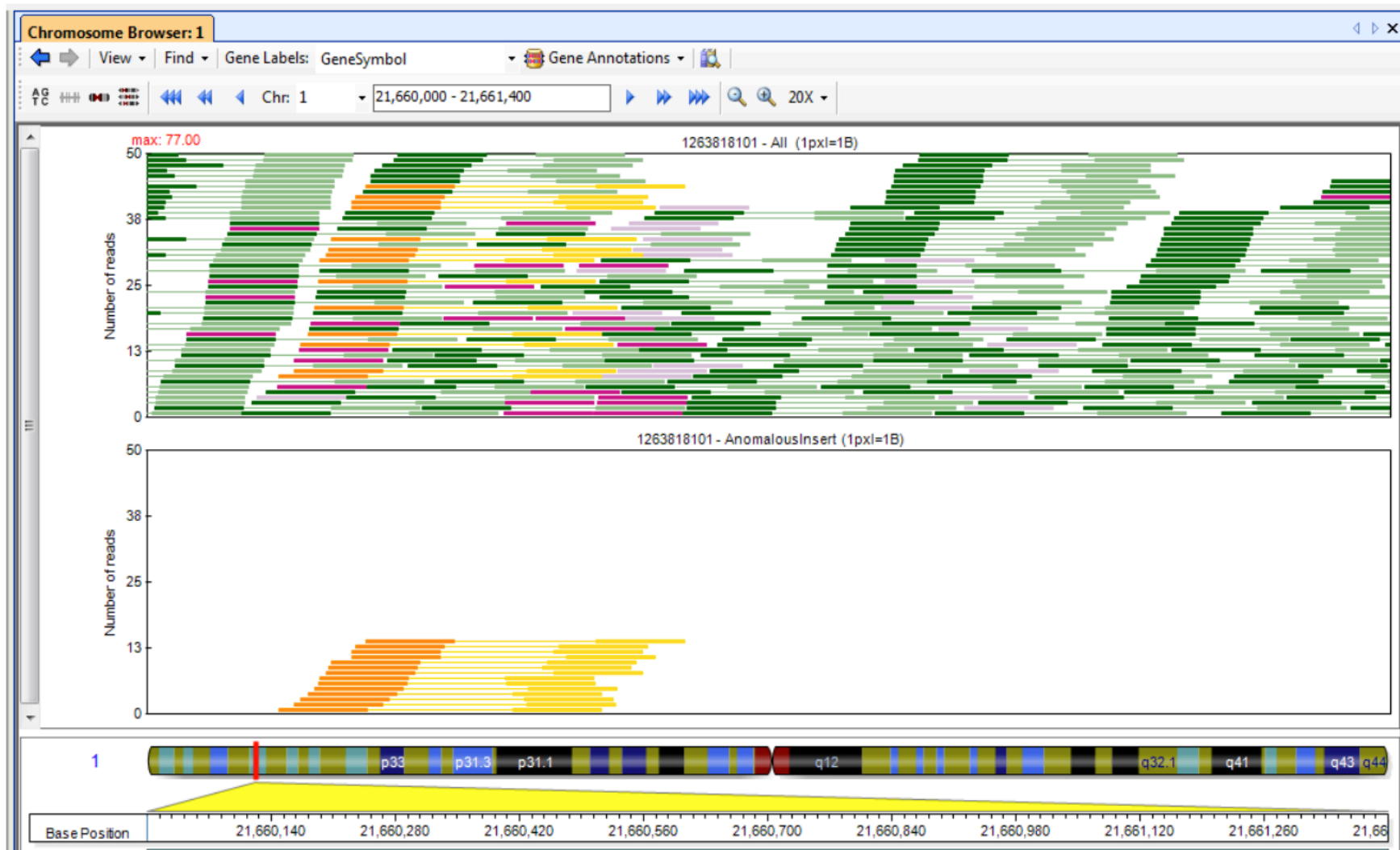


Daughter

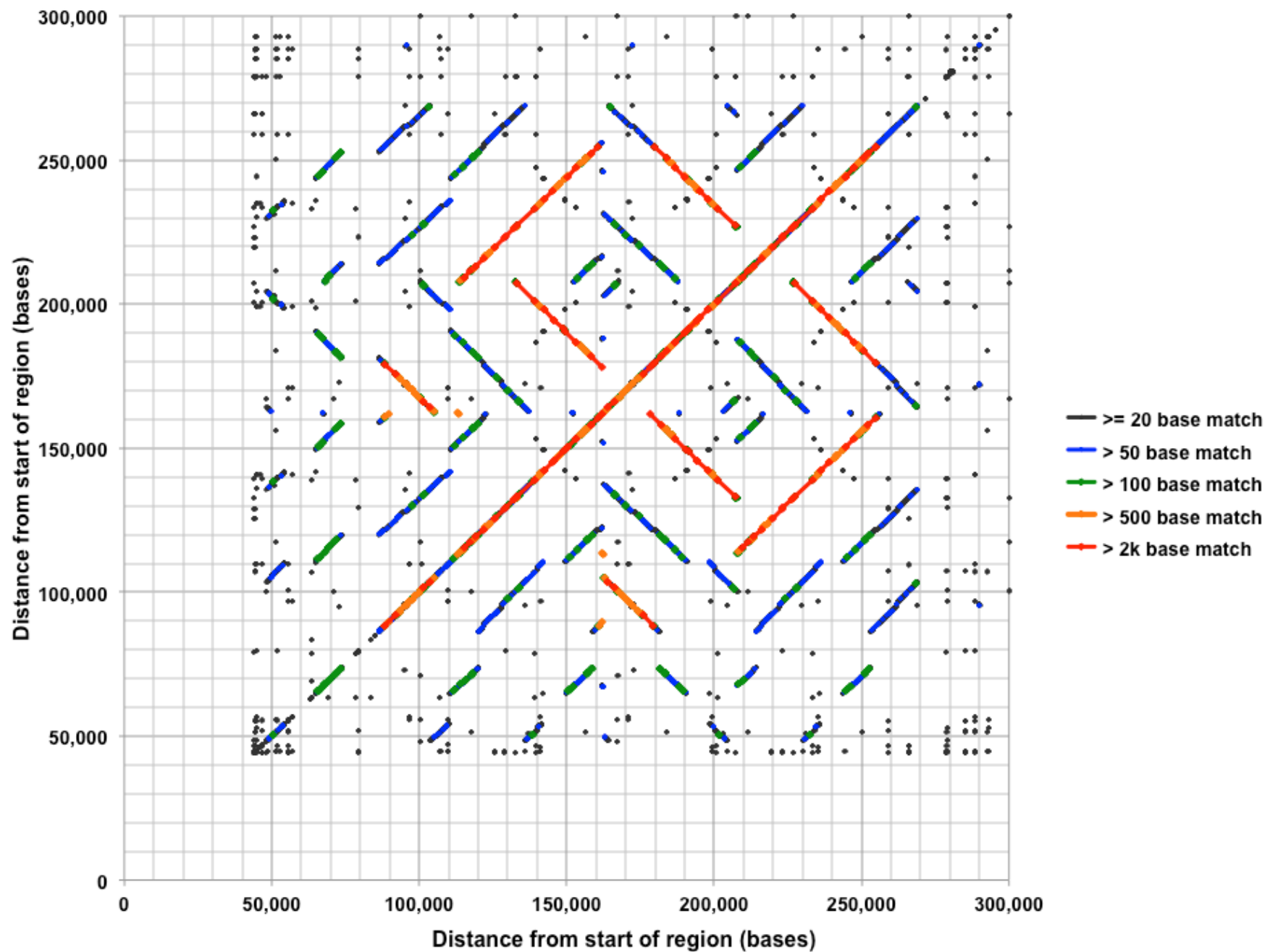


Son

HugeSeq family validation included heterozygous deletions down to 63 bases long...



Forward & Reverse repeat structure of the Amylase Gene Region



Duplications underlie many failure mechanisms

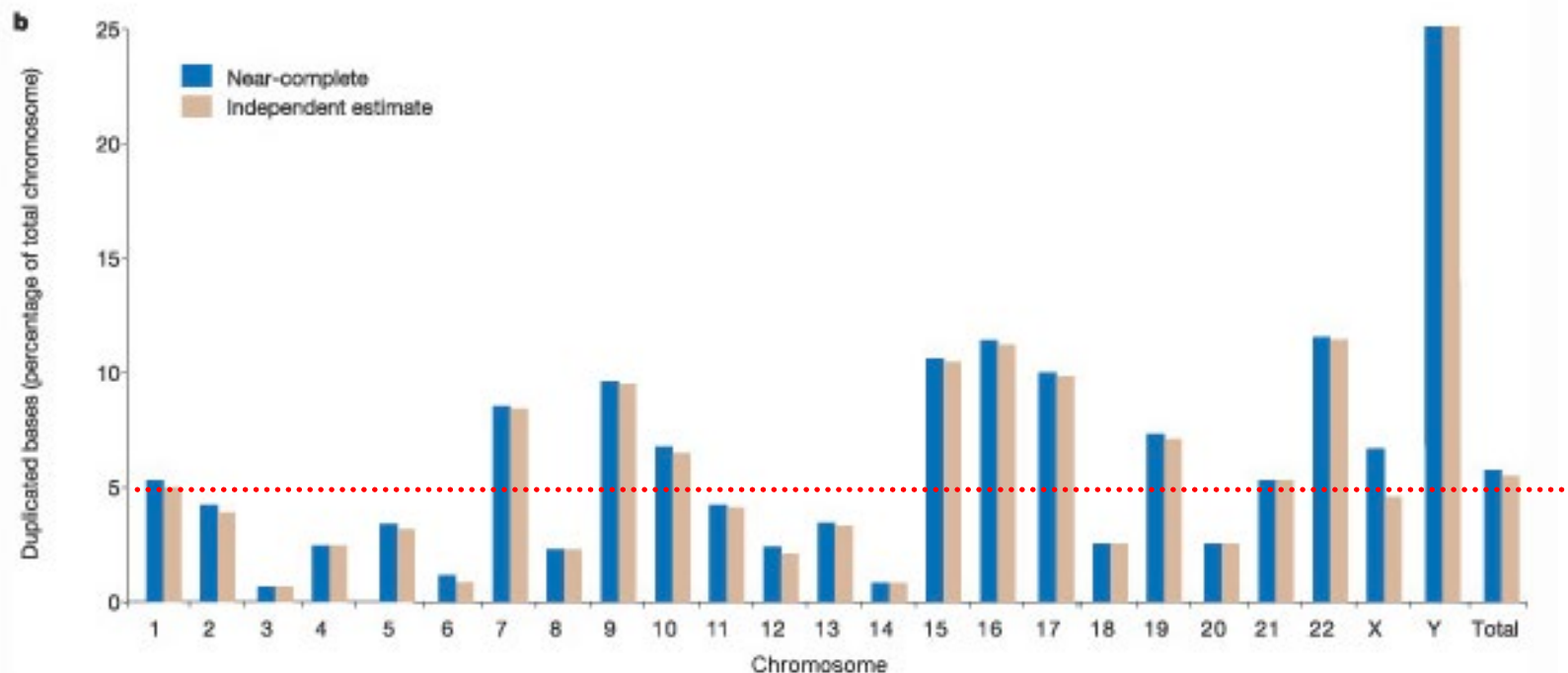


FIGURE 4. Segmental duplications across the genome.

Segmental duplications (length ≥ 10 kb and sequence identity $\geq 95\%$)

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium

Nature **431**, 931-945 (21 October 2004) | doi:10.1038/nature03001; Received 29 July 2004; Accepted 7 September 2004

Gene birth in the human lineage

The birth of new genes is of interest because it provides raw material for adaptive evolution, with extra copies of genes able to undergo functional divergence in response to positive selection. The quality and completeness of the current sequence make it possible to study this question; such analysis would have been unreliable with the earlier draft sequence, because the extensive artefactual local duplication would have given rise to many false positives.

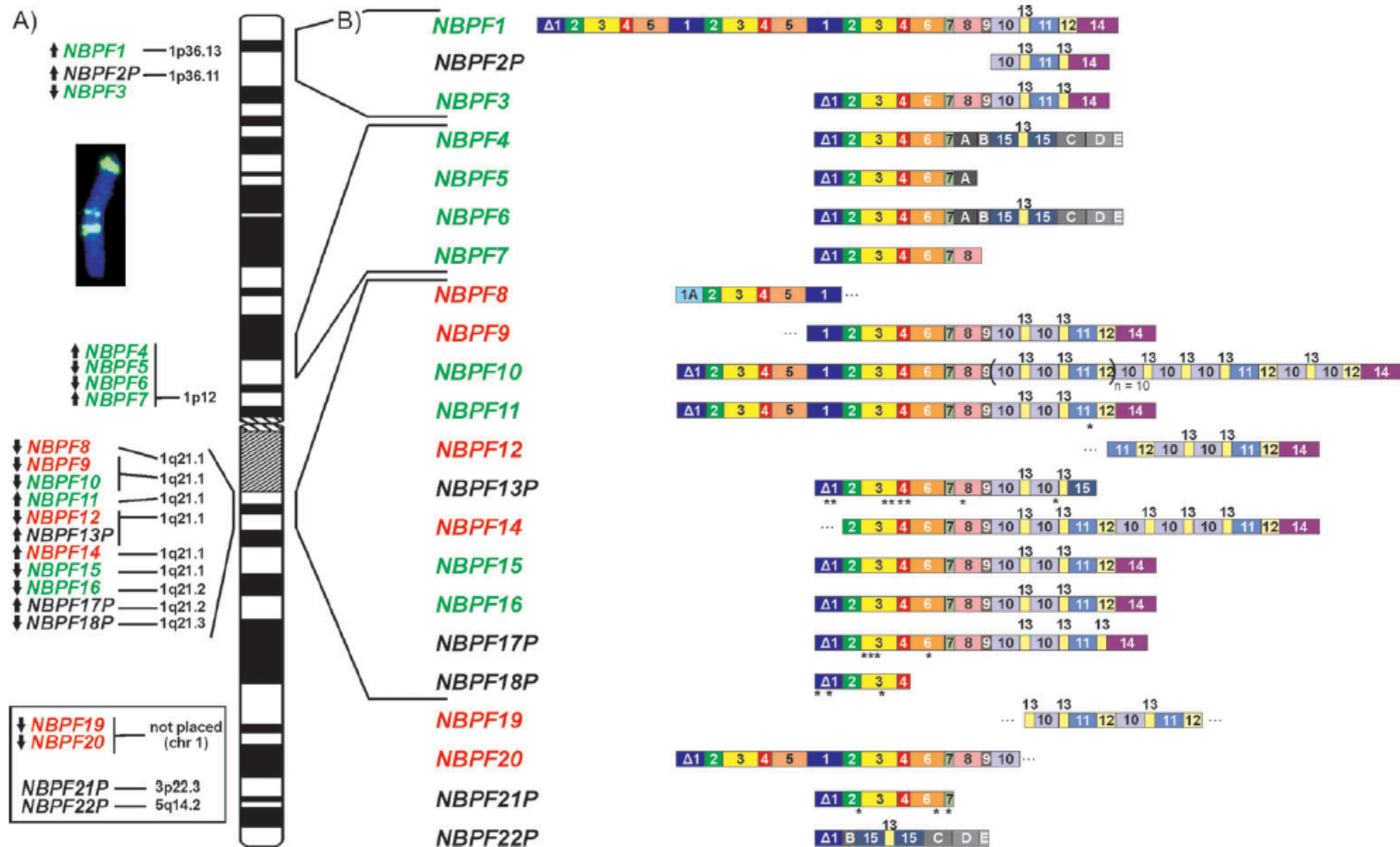
We searched for clusters of nearby homologous genes, indicative of local gene duplication. The divergence between such genes was assessed at sites likely to be selectively neutral, by measuring the estimated substitution rate per synonymous site (K_S). We looked for nearby human gene pairs differing from one another by $K_S < 0.30$, implying that each differs from the common ancestral source gene by an average $K_S < 0.15$. This threshold corresponds roughly to duplications arising after divergence from the rodent lineage, either by recent gene duplication or perhaps recent gene conversion of older duplications (see Methods in [Supplementary Information](#)). A total of 1,183 genes exhibit such divergence from a neighbouring gene (see Methods in [Supplementary Information](#)). These genes often fall within larger clusters of paralogous genes including genes with greater divergence and reflecting older duplications. These clusters contain ~3,300 genes, and those having at least five genes involved in recent duplication events are shown in [Table 4](#). Analysis of phylogenetic trees containing the related human and mouse genes confirms that the genes are more closely related within each species than between the two species in nearly all cases (97%), as would be expected for genes arising by duplication after the divergence of the human and rodent lineages.

~3,300 genes,

About 15%...

22 Member Neuroblastoma Breakpoint Family (NBPF) of Genes

Numerous regions with good DNA sequencing coverage, but tri-allelic SNP's in a single individual's genome, an indication of alignment errors



A Novel Gene Family NBPF: Intricate Structure Generated by Gene Duplications During Primate Evolution

Karl Vandepoel,* Nadine Van Roy,† Katrien Staes,* Frank Speleman,† and Frans van Roy*

Mol. Biol. Evol. 22(11):2265–2274. 2005

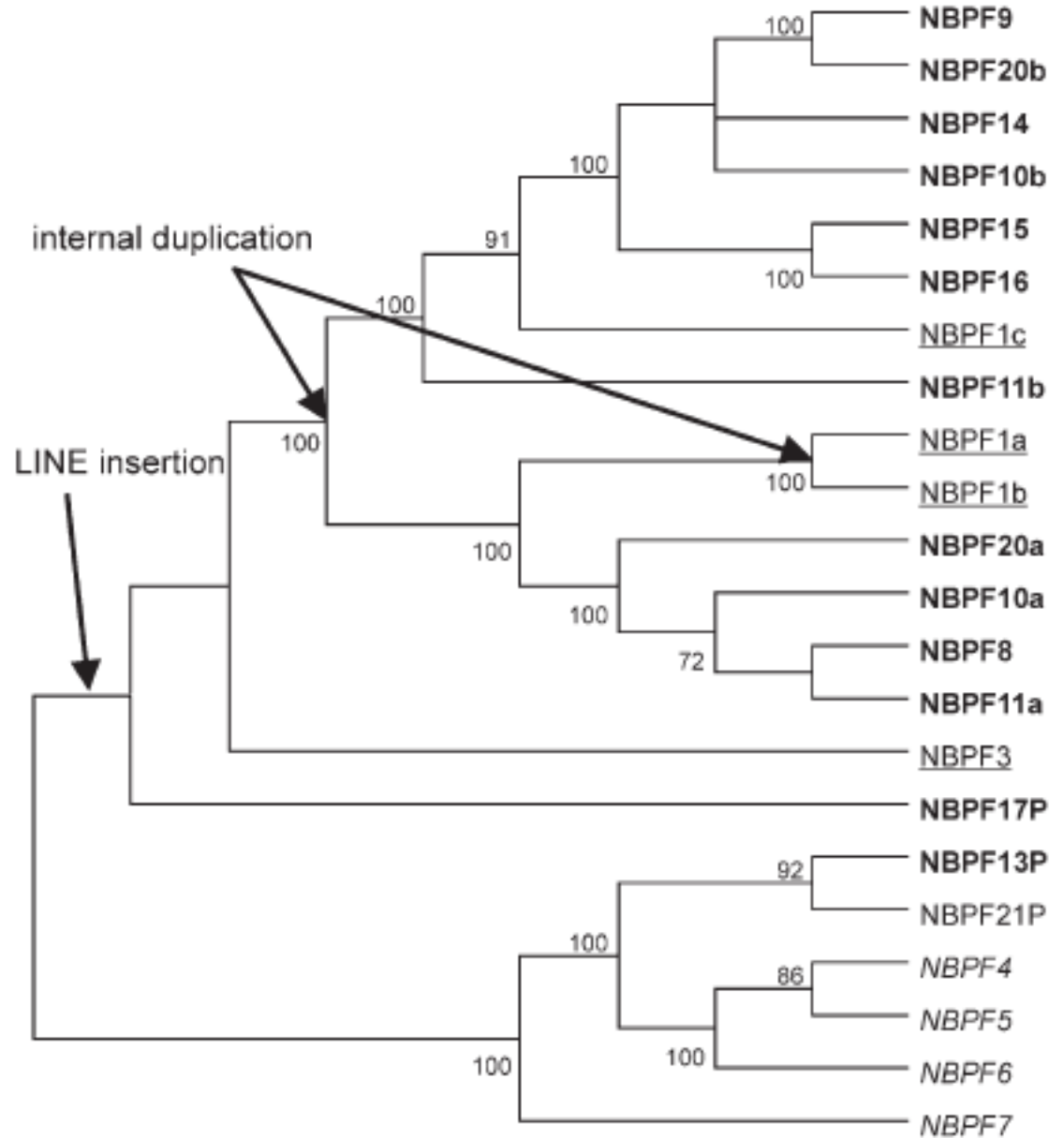
doi:10.1093/molbev/msi222

Advance Access publication August 3, 2005

**Evolutionary view
of how this gene
family developed
by duplication &
divergent mutation**

Duplications can be internal to a gene, in addition to duplications of the whole gene

Transposon insertions (e.g. LINE) can further complicate the repeat structure of a gene family



Amylase measurement, serum

What is this test?

This test measures the amount of amylase in blood. It is used to evaluate disorders of the pancreas, and gastrointestinal (digestive) tract[1][2][3][4][5][6]. It may also be used to evaluate autoimmune and viral illnesses[7], shock, trauma, and hypersensitivity reactions[8].

Why do I need this test?

Laboratory tests may be done for many reasons. Tests are performed for routine health screenings or if a disease or toxicity is suspected. Lab tests may be used to determine if a medical condition is improving or worsening. Lab tests may also be used to measure the success or failure of a medication or treatment plan. Lab tests may be ordered for professional or legal reasons. The following are possible reasons why this test may be done:

- Acute vascular insufficiency of intestine
- Gallbladder inflammation
- HIV infection
- HUS - Hemolytic uremic syndrome
- Inflammatory bowel disease
- Injury of pancreas ←
- Mumps
- Pancreatitis
- Peptic ulcer
- Perforation of esophagus
- Toxic epidermal necrolysis

Amylase is a key enzyme in human digestion of starch.

40-45% of all salivary protein.

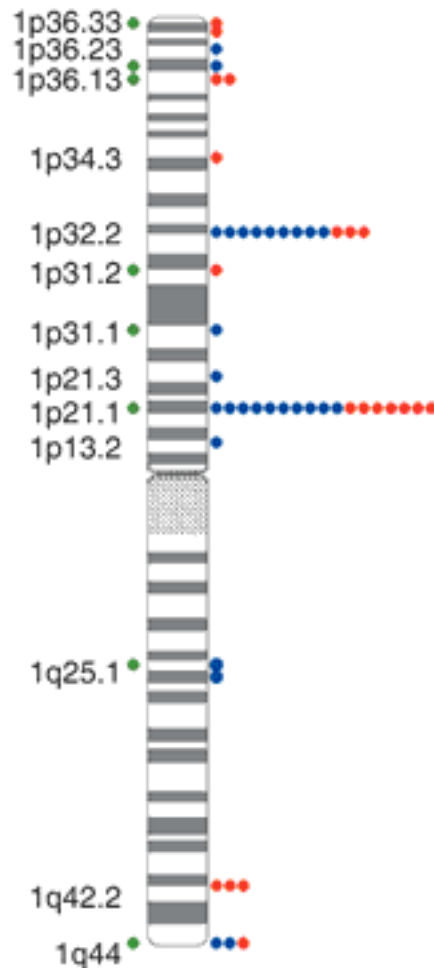
Used as an indicator of numerous diseases, but not part of a standard blood chemistry panel.

Figure 1

Nature Genetics **36**, 949 - 951 (2004)
Published online: 1 August 2004; | doi:10.1038/ng1416

Detection of large-scale variation in the human genome

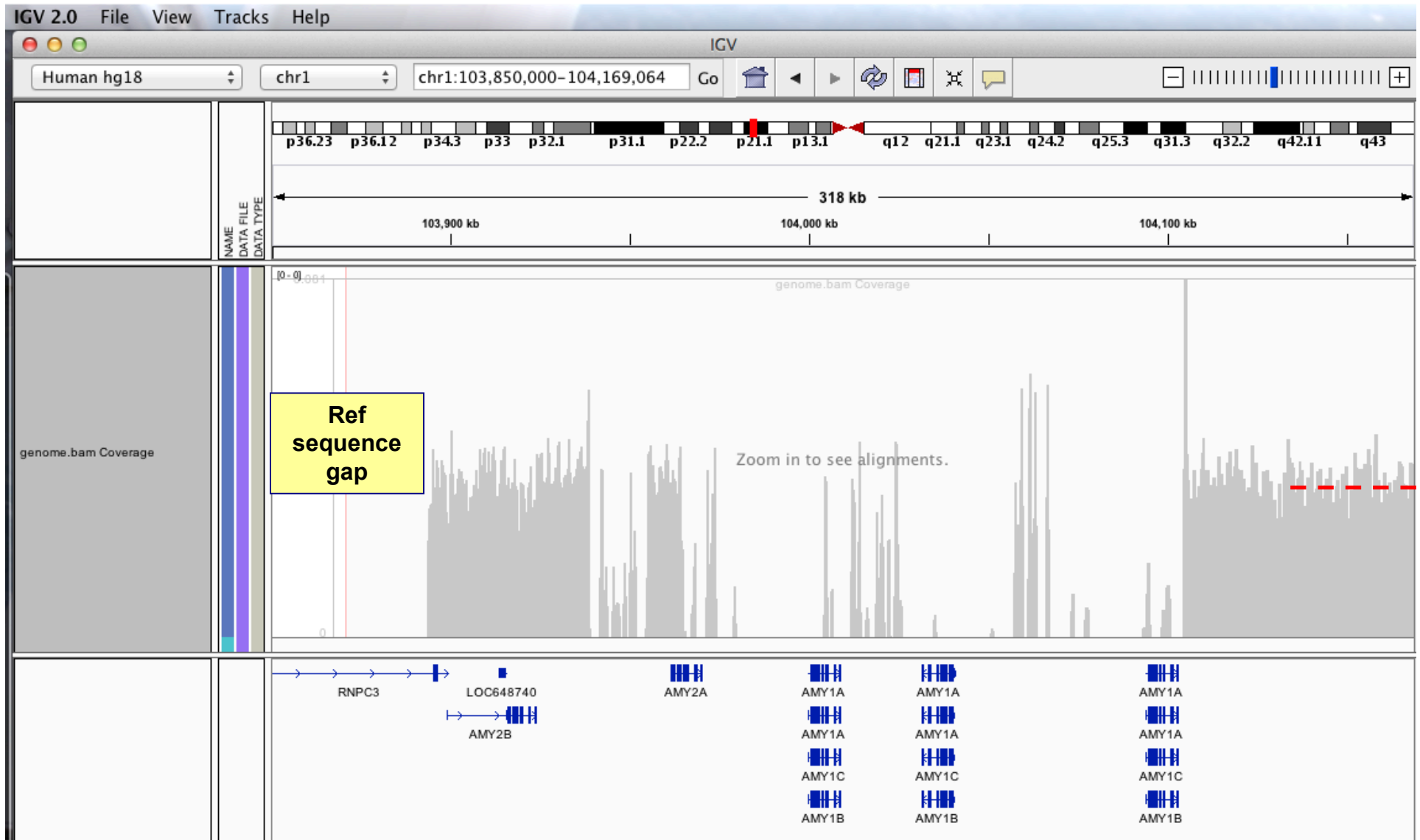
A John Iafrate, Lars Feuk, Miguel N Rivera, Marc L Listewnik, Patricia K Donahoe, Ying Qi, Stephen W Scherer & Charles Lee



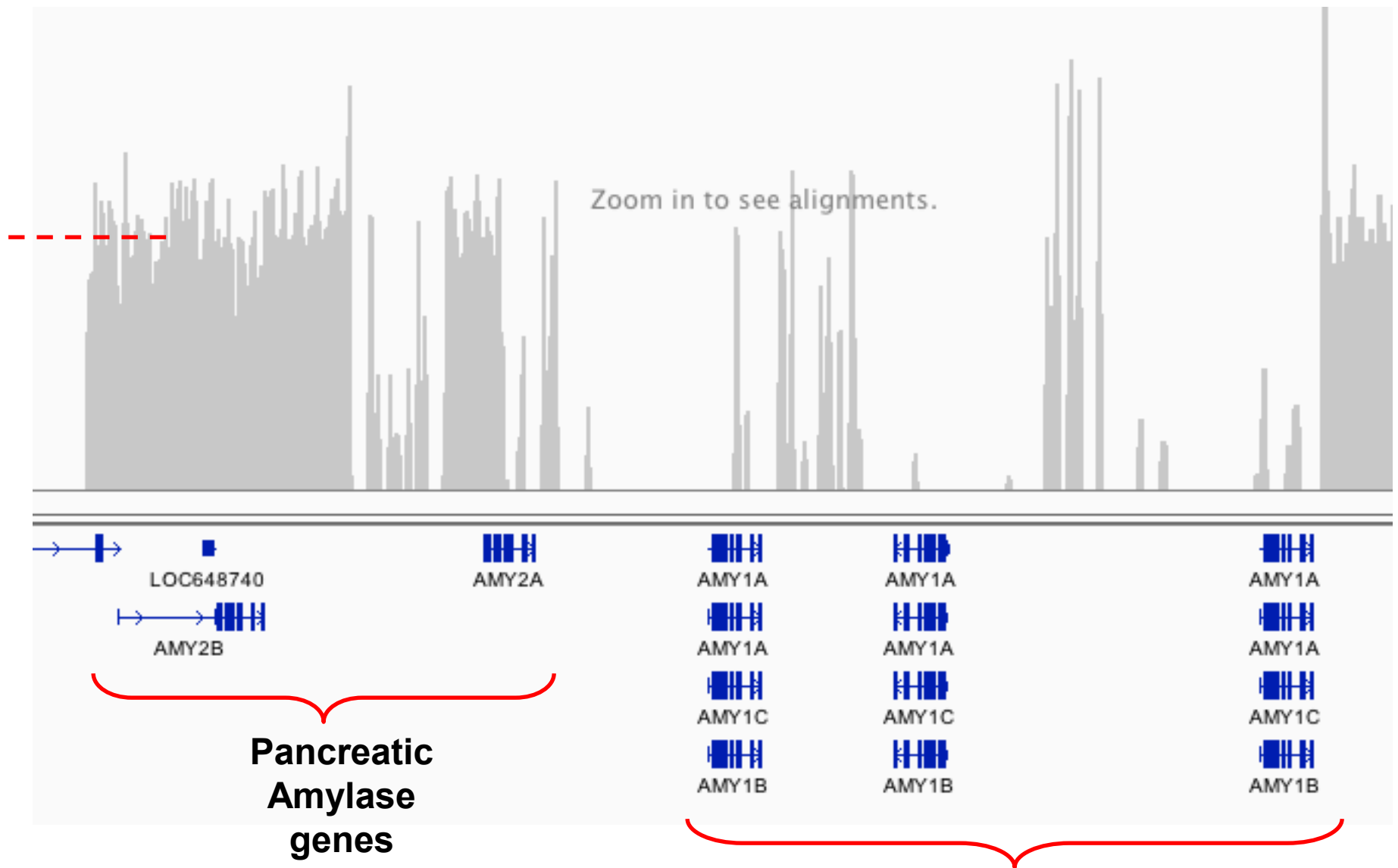
Amylase gene cluster, showing both copy number gains (blue) and losses (red).

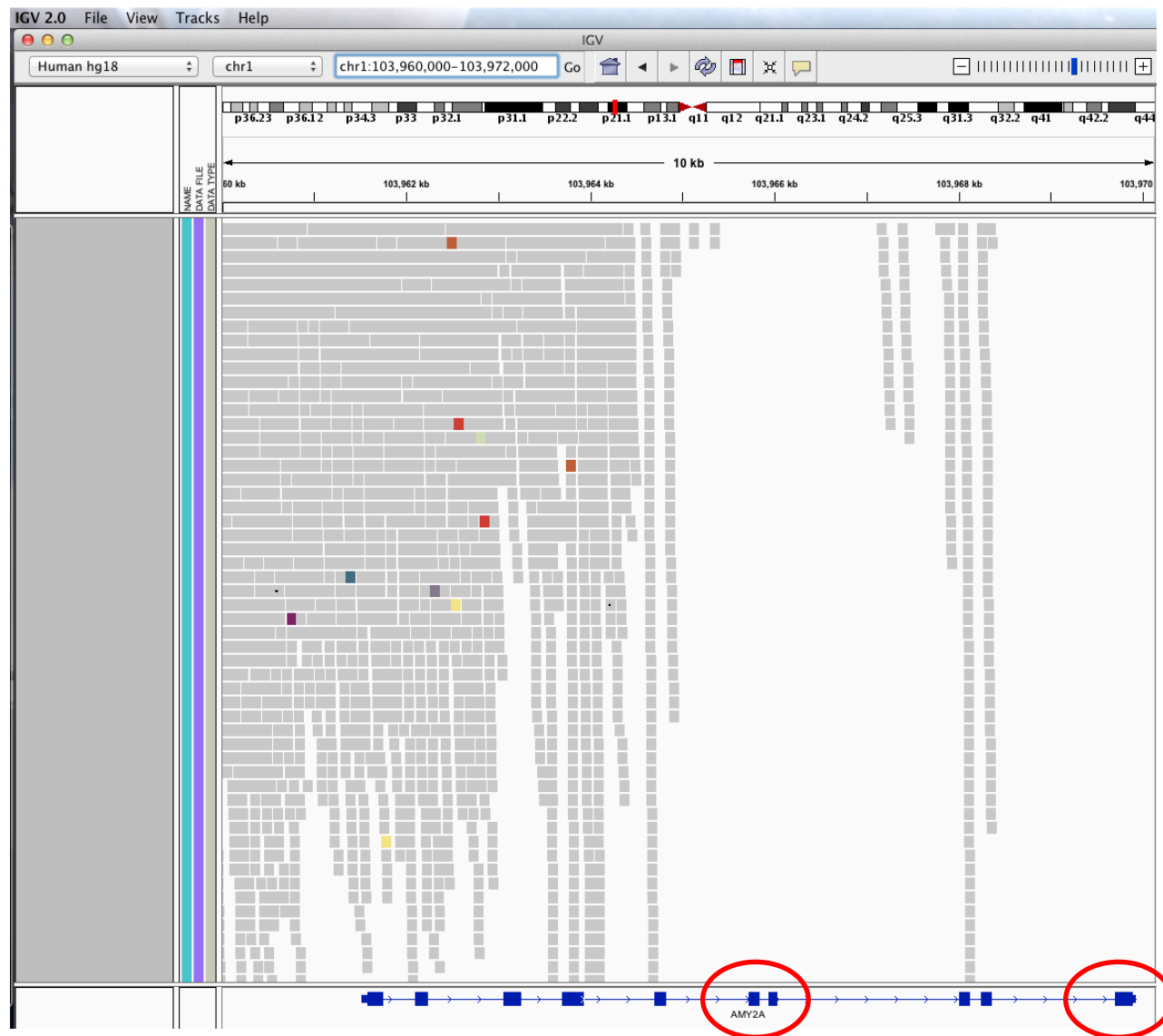
CNV in 49% of samples.

CNV is the dynamic outcome of the gene duplication mechanism which also leads to many alignment problems.

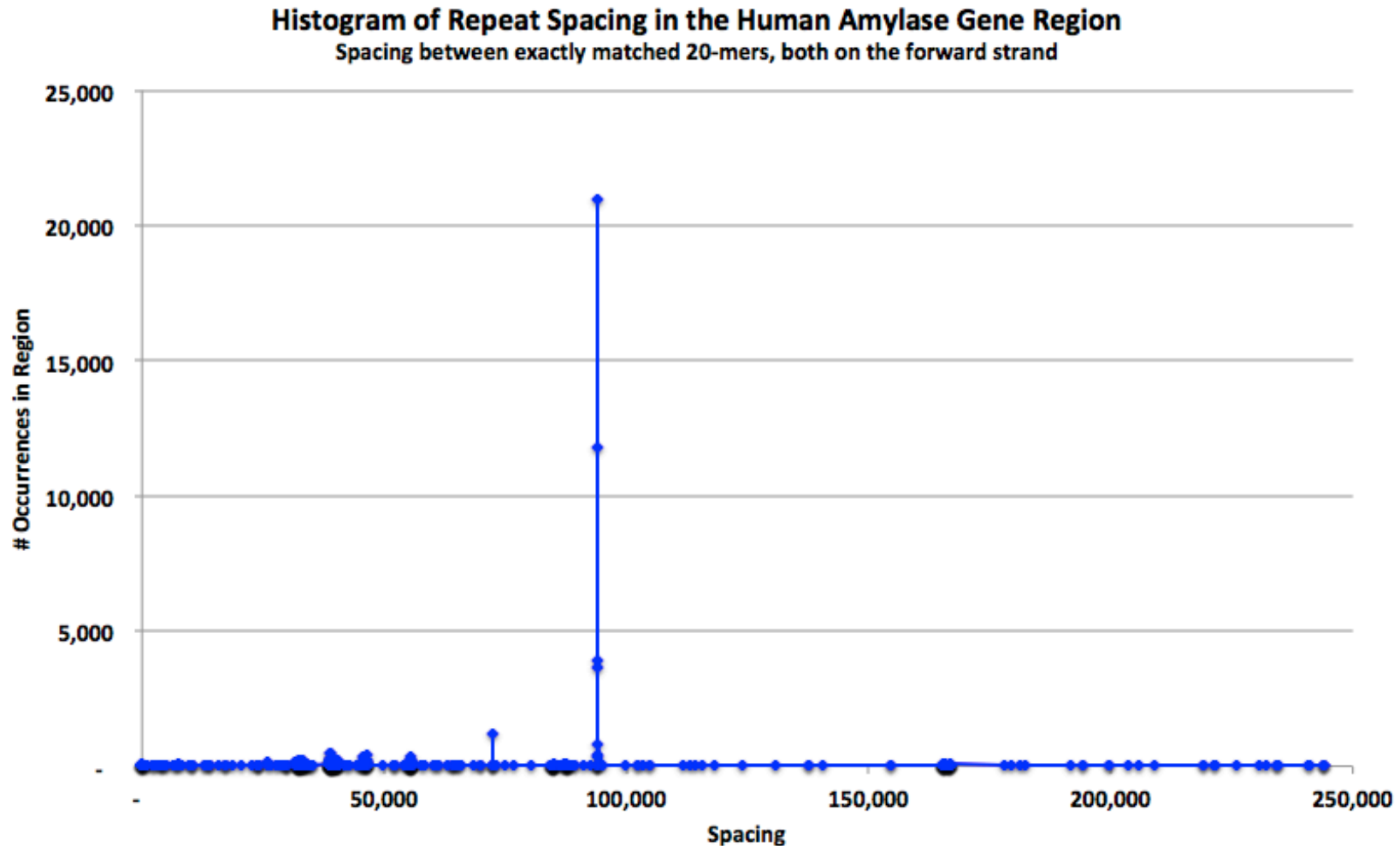


DNA sequencing coverage in a Yoruban genome sequenced by Illumina with 2 x 100 base paired-end reads, in the region of the Amylase gene cluster. Sequence reads likely exist for this region, but cannot be uniquely aligned due to sequence homology.



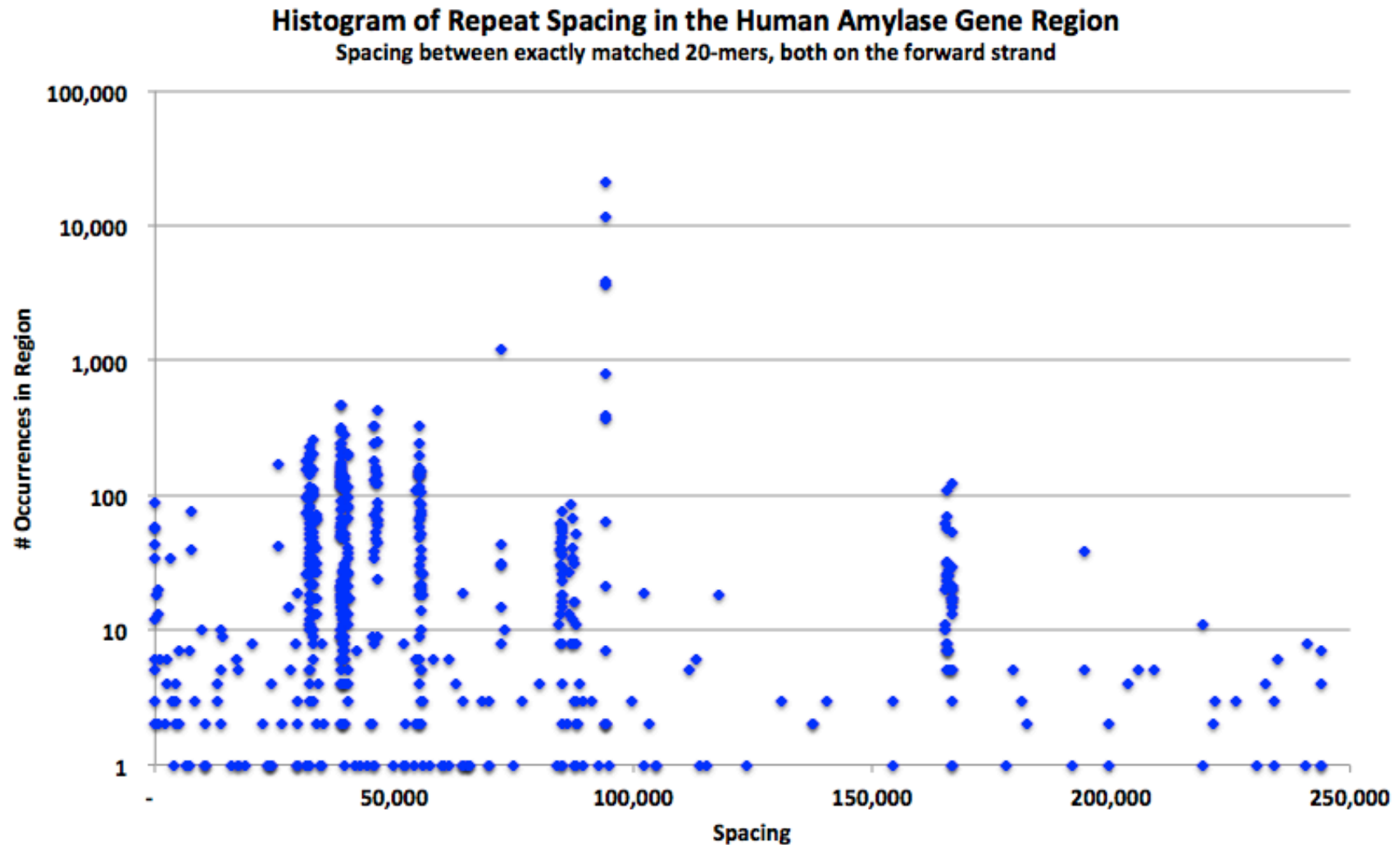


**Coverage in the AMY2A gene is zero in exons 6, 7 & 10.
The coverage gap which includes exons 6 & 7 is almost 2kbases long.**



Looking at 300,000 20-mers in the region, it seems there is one main duplication at a spacing of about 94k bases

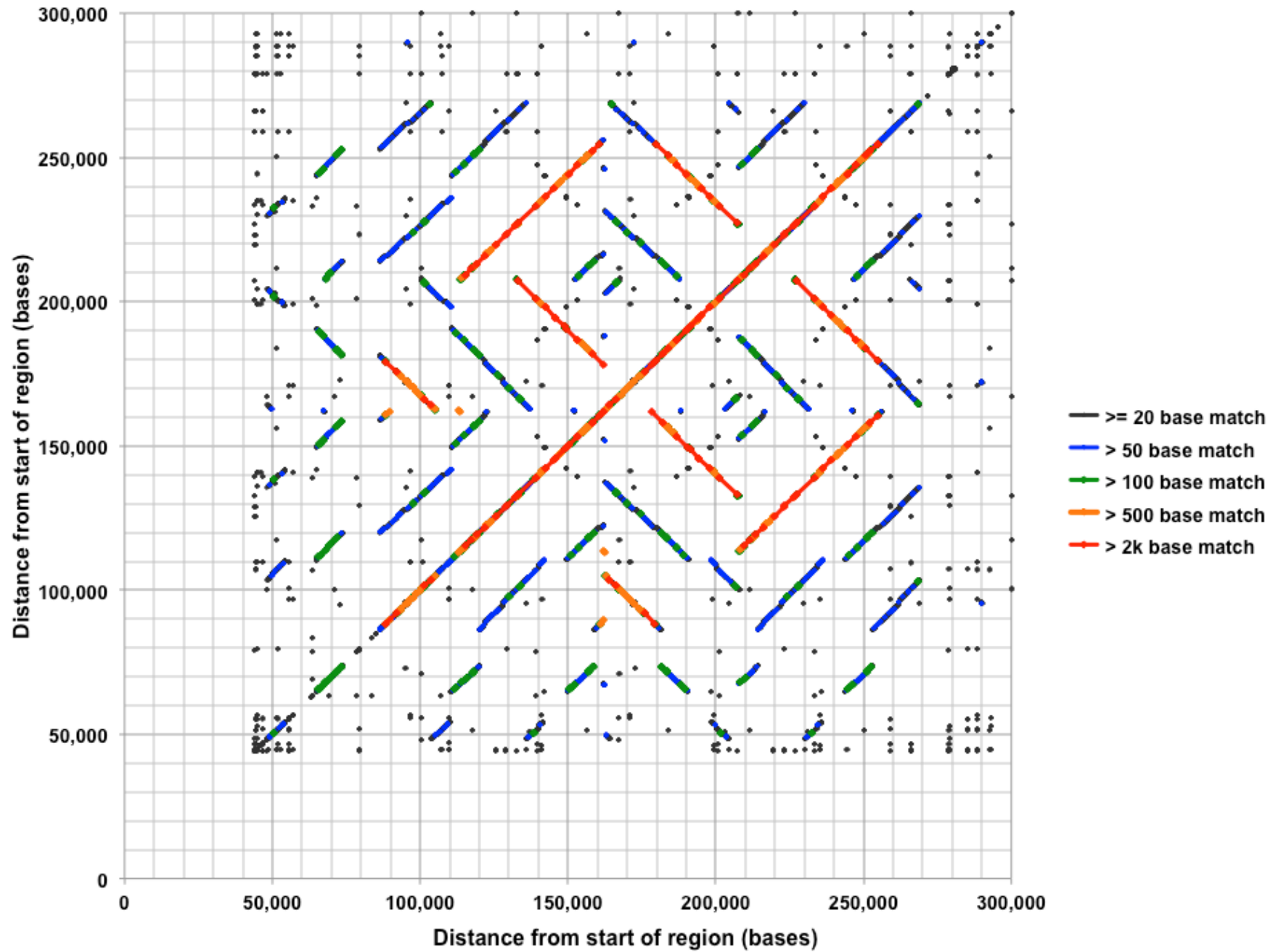
20,986 20-mers were found spaced by 94,138 bases

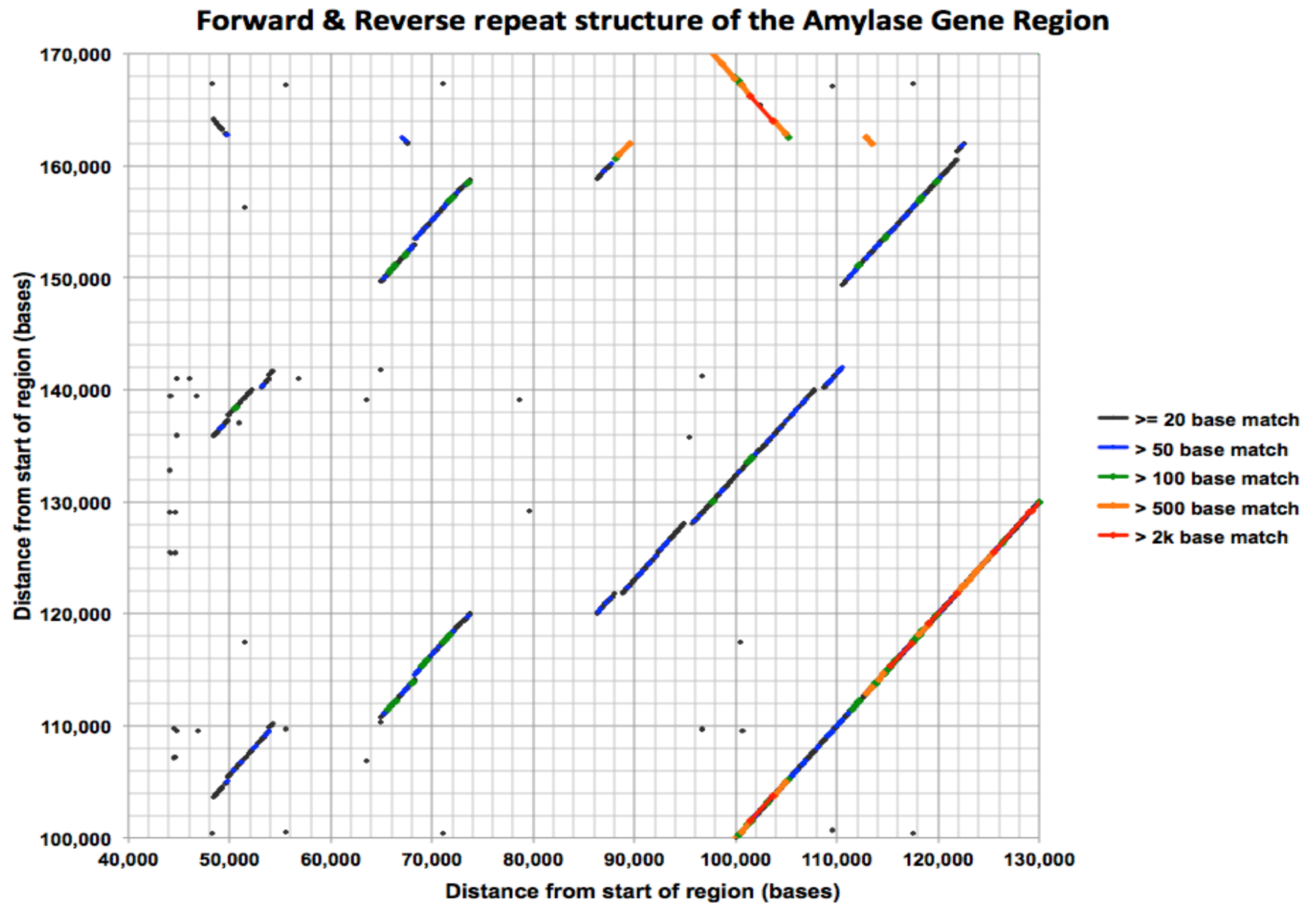


The same histogram, on a log scale.

Additional clusters suggests more complex repeat structure, fragmented by InDels.

Forward & Reverse repeat structure of the Amylase Gene Region





Zooming in, broken diagonals show InDel's within the repeat structure

Relative Spacing Between Pairs of Exactly Matched 20-mers
Object Offsets = 1,000 / object

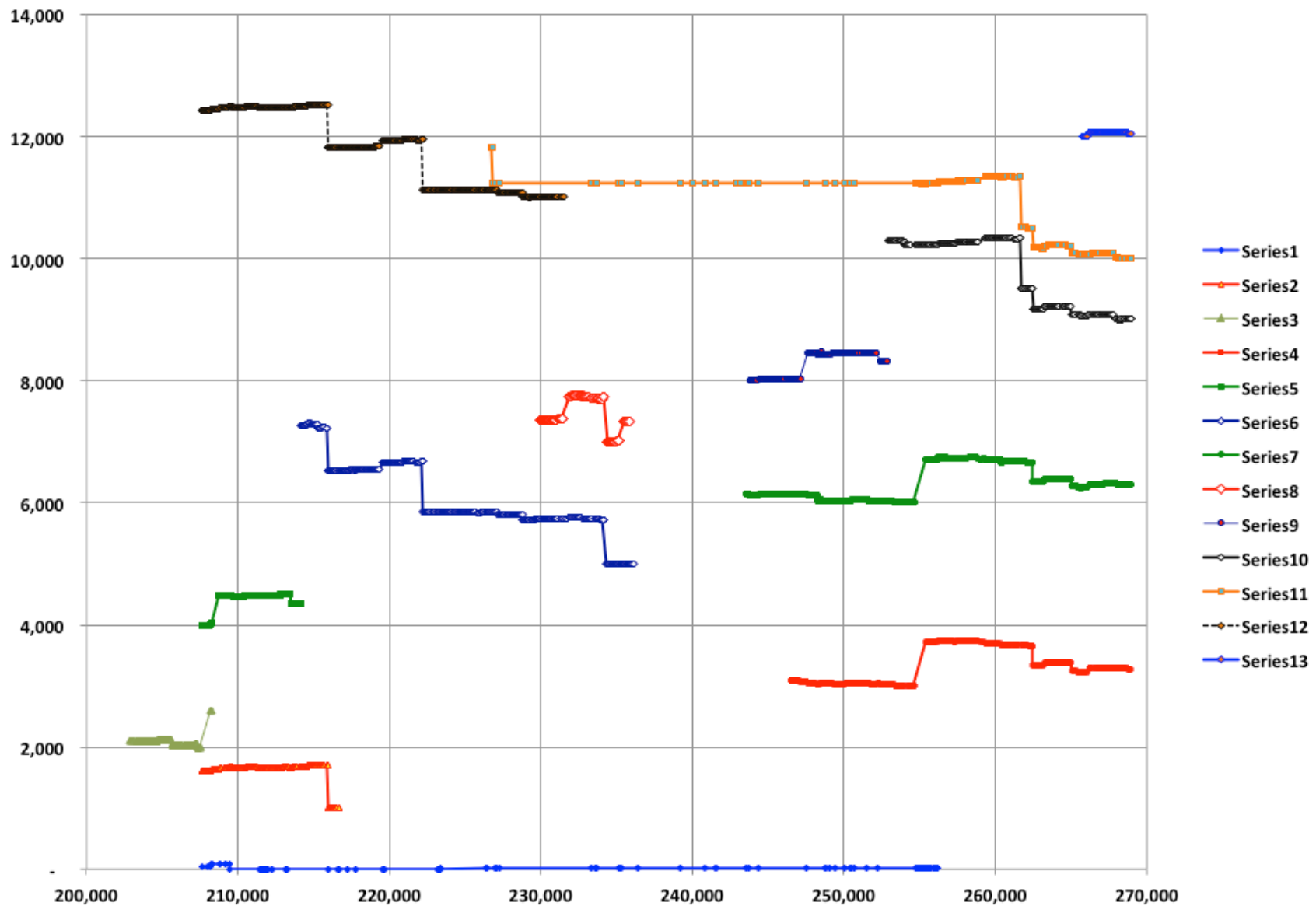
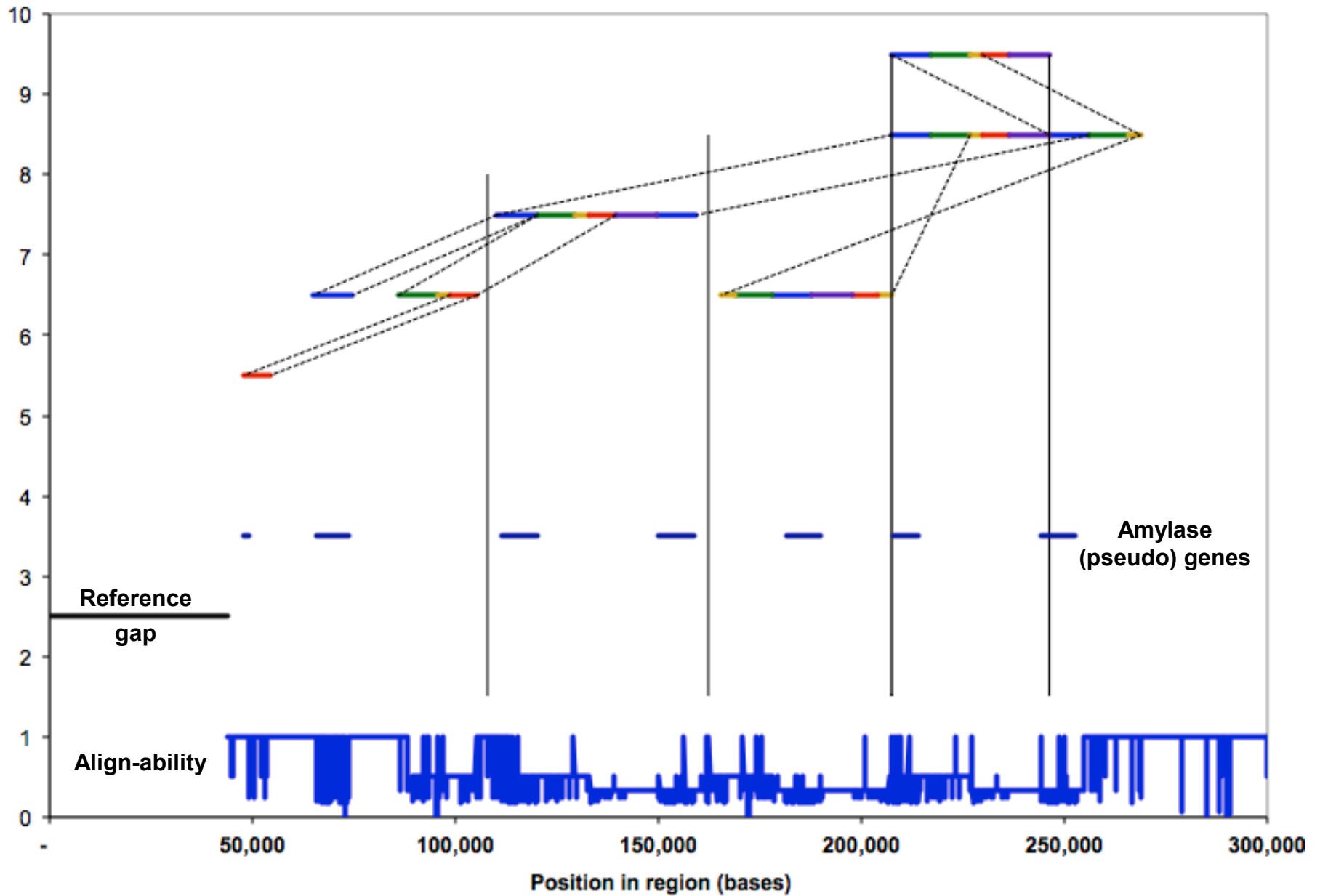
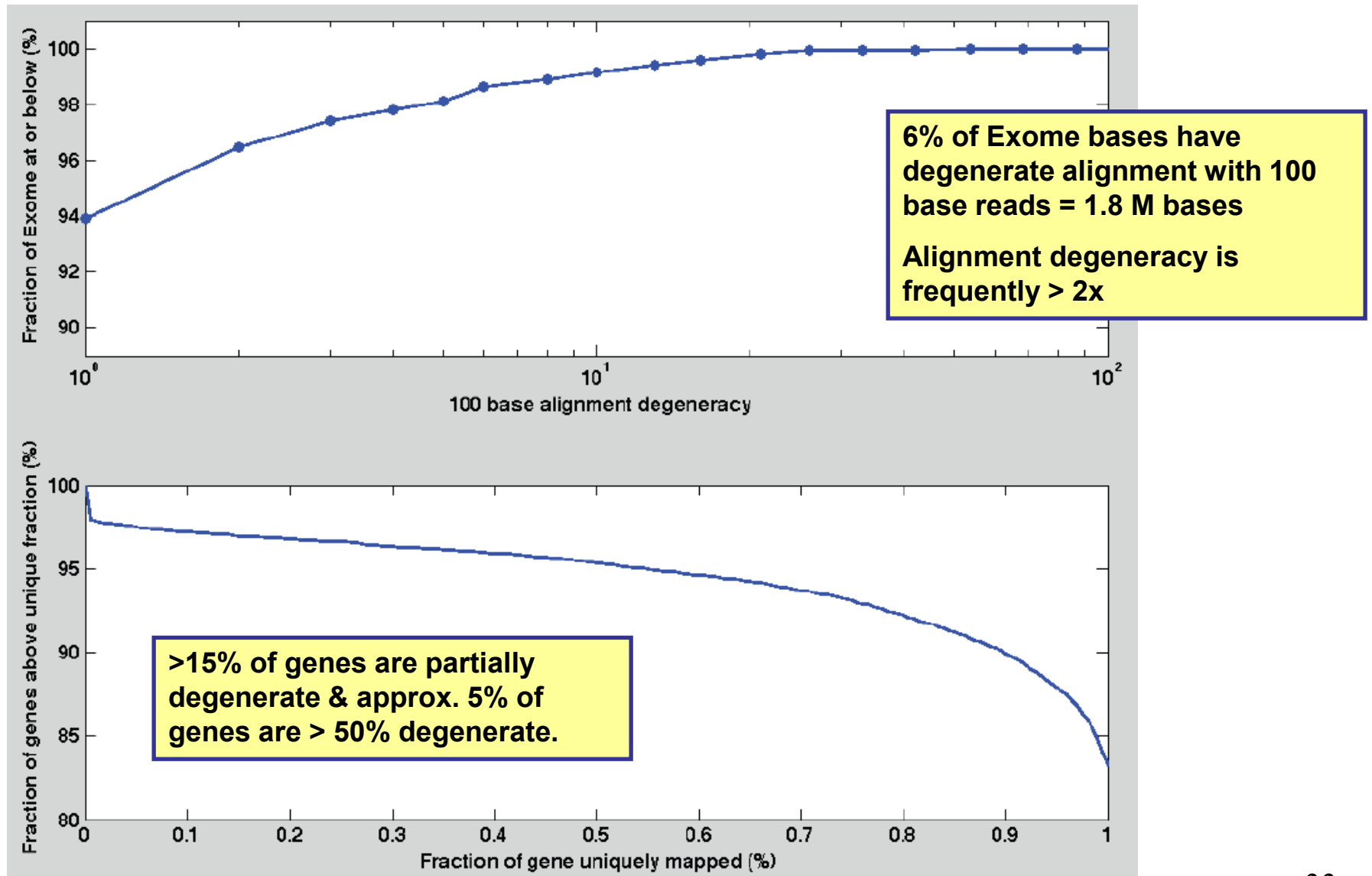


Diagram of Repeat Structure in Amylase Gene Region



Quantifying Alignment Degeneracy Genome-Wide



Near-Term Product Development to Address Alignment Degeneracy in Gene Families

- **Categorize genes with degenerate alignment:**
 - Have sequencing coverage but variants may be suspect
 - Systematically low sequencing coverage, so calls may be poor or non-existent
- **Compare degeneracy vs other lists of problematic regions:**
 - Mendelian Inheritance Errors
 - Compressions,
 - Tri-allelic loci in a single individual
- **Algorithms to align successfully in degenerate regions**
- **Laboratory methods to solve degeneracy**

Summary of Problems

- Errors (False pos/neg):
 - SNP's
 - InDel's
 - SV's (existence, exact position & length)
 - Triple haplotypes & loci tri-allelic in a single individual
 - Reference sequence
 - Suboptimal alignment
 - Minor & rare alleles in reference (including InDels & SV's)
 - Compressions
 - Millions of bases in chrN_random
 - Sanger sequencing errors
- Information missing:
 - Zygosity of SV's
 - Simple (?) for deletions
 - More complex for CNV
 - VNTR's & other problematic sequences
 - Coverage in degenerate regions (incl. 3,300 genes)
 - Coverage in high %GC regions
 - Phasing (in individual genomes)
 - Allele-specific expression
 - Many arrays don't cover difficult regions

What Personalis plans to do about it

- Comprehensively understand the problem – Current studies include:
 - MIE of all variant types in a family quartet
 - Comparison of the same person sequenced multiple times
 - Same person sequenced on multiple platforms (e.g. ILMN vs CG)
- Fix the reference
 - Ethnically specific major allele references (patent pending)
 - Expand on this theme (InDel's, SV's, admixed, etc)
- Combine multiple genetic analysis technologies to capture the genome:
 - ILMN + CG if that helps
 - Multiple insert sizes per sample
 - Sequencing + array data + other technologies (e.g. electrophoretic)
 - Considering a custom-content array (e.g. VariMed, PharmGKB & JM library)
- Focus on what matters for interpretation
 - PharmGKB, VariMed & Mendelian loci
- Leverage our expertise in SV's

Determining SNP Error Rate

Quartet data

CEPH1463
CGI data

NA12878 SRA data

CEPH1463
family new
data

Process with
modern
HugeSeq (done)

MIE Analysis

Link MIE rate to
error rate & type

VCF's aligned
across four
family members,
including non-
calls, coverage,
allele Q-scores

MIE Analysis

Link MIE rate to
error rate, type,
mechanism

Comparisons

Same library, two
runs

Same protocol,
two libraries

Broad HiSeq vs
CGI (vs array ?)

Max quality
alignment by
combining 10
paired-end insert
lengths & very
high coverage;
Then compare
individual runs to
this standard

Single genome

Where is
coverage
insufficient ?

Where are
allele Q-
scores too
low ?

Tri-allelic /
triple
haplotype
sites

Allelic bias
(requires
very high
coverage)

Comparison
ILMN / CGI /
Genotyping

MIE path like
quartet