

From: [John West](#)
To: [Mark Pratt](#)
Subject: My slides for meeting with Mike Snyder yesterday
Date: Tuesday, May 22, 2012 6:38:43 AM
Attachments: [Overview for Mike S 21May2012 JW.ppt](#)

Hi Mark,

I am attaching my slides from yesterday's meeting with Mike Snyder - some maybe useful for bringing Jason up to speed ?

I didn't realize Gabor had initial MIE data from the quartet genomes and am sorry we ran out of time for that, though it seemed so new that you all are somewhat still sifting through it.

When you have, could we walk through it ?

Thanks,

John

Personalis, Inc.

Accuracy Overview

John West

May 21, 2012

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

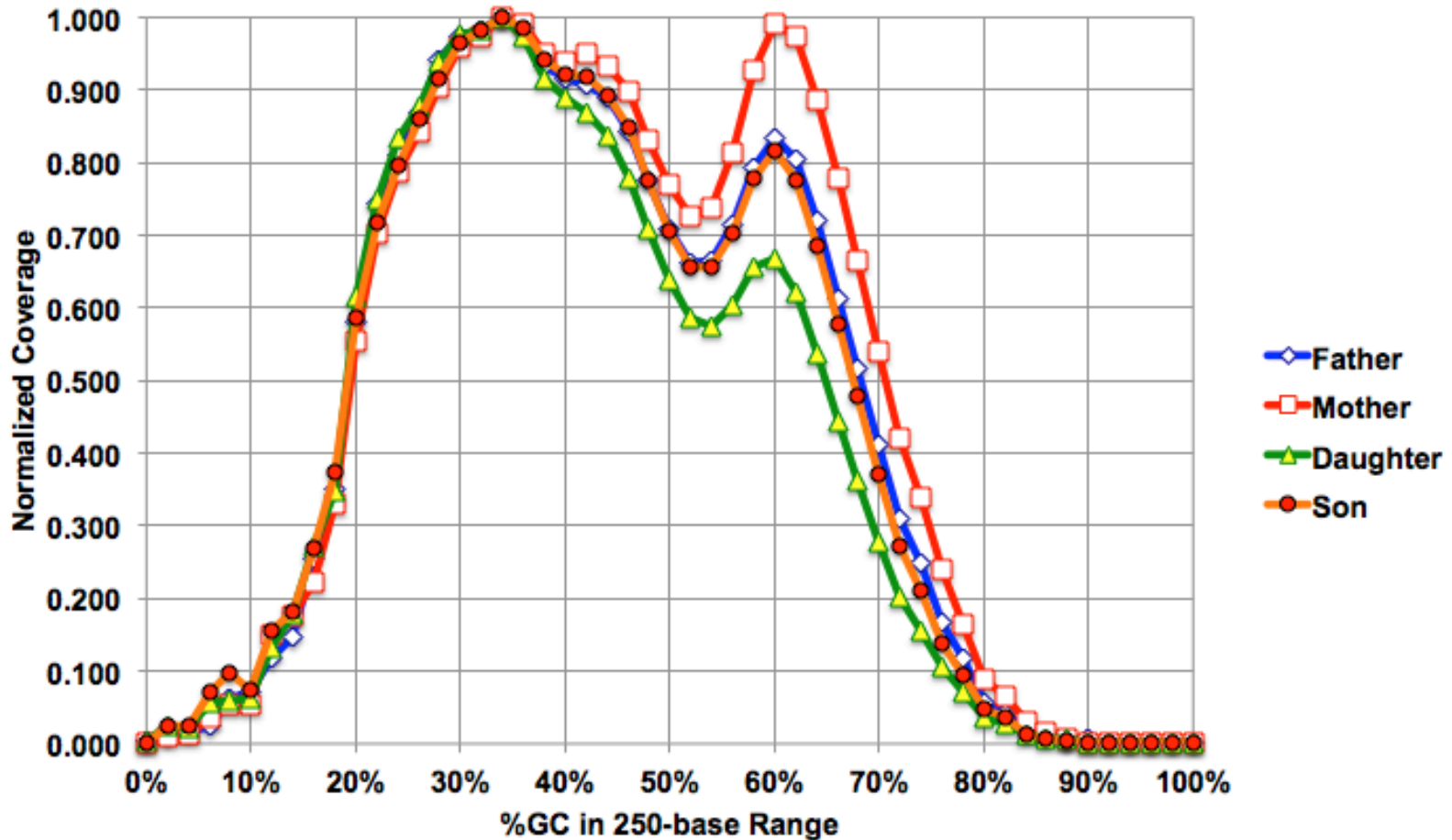
Gold standard genomes

Known healthy genomes

Overview

- (Mis)Alignment in degenerate regions
- Raw read errors (systematic)
- Problems in the reference
- Tools for identifying error sources & rates

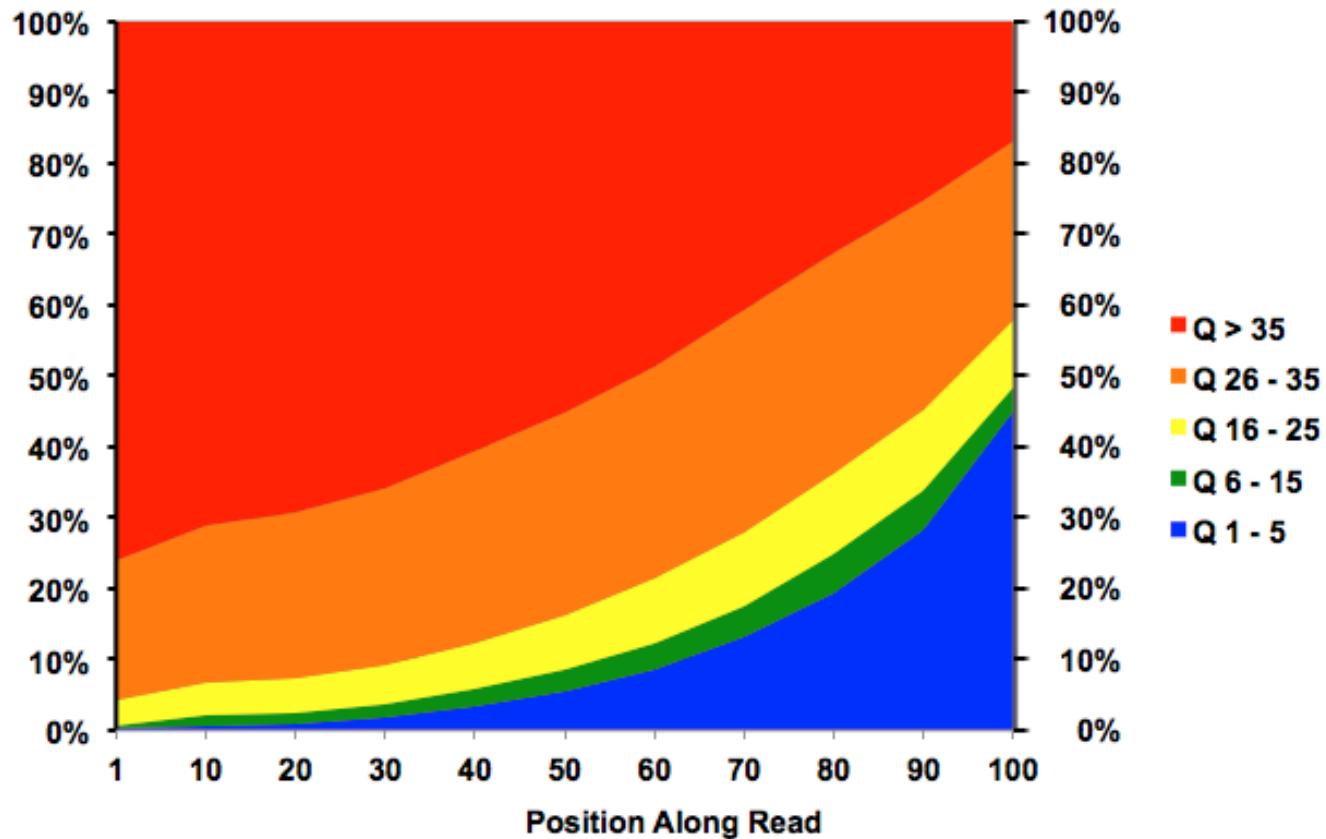
Above 50% GC, Quartet Genomes Have Substantially Different Coverage



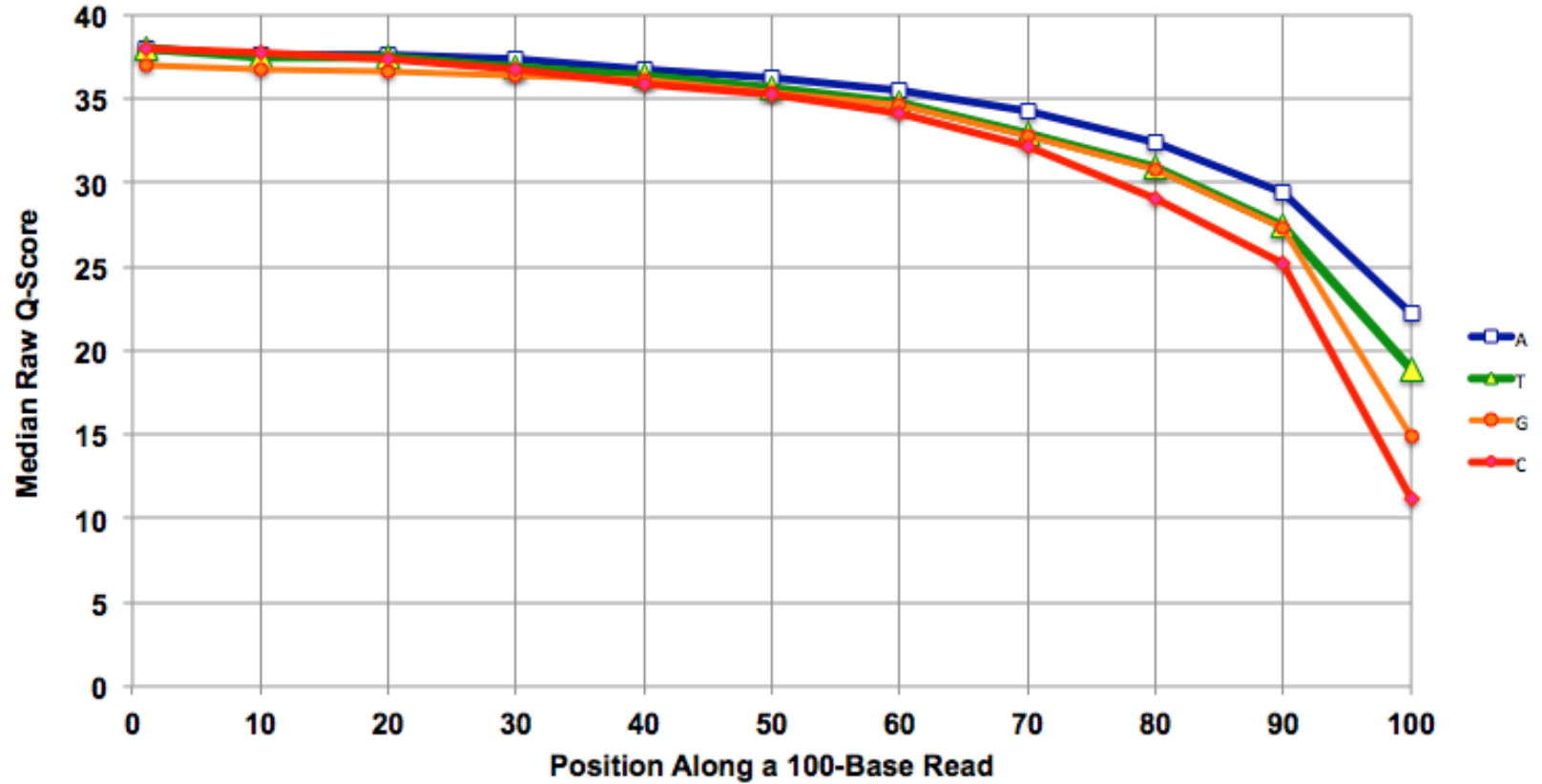
Example of process variation in genome sequencing

Four individuals had blood drawn together at the same hospital, sent to the same CLIA / CAP accredited lab, and sequenced on the same set of instruments.

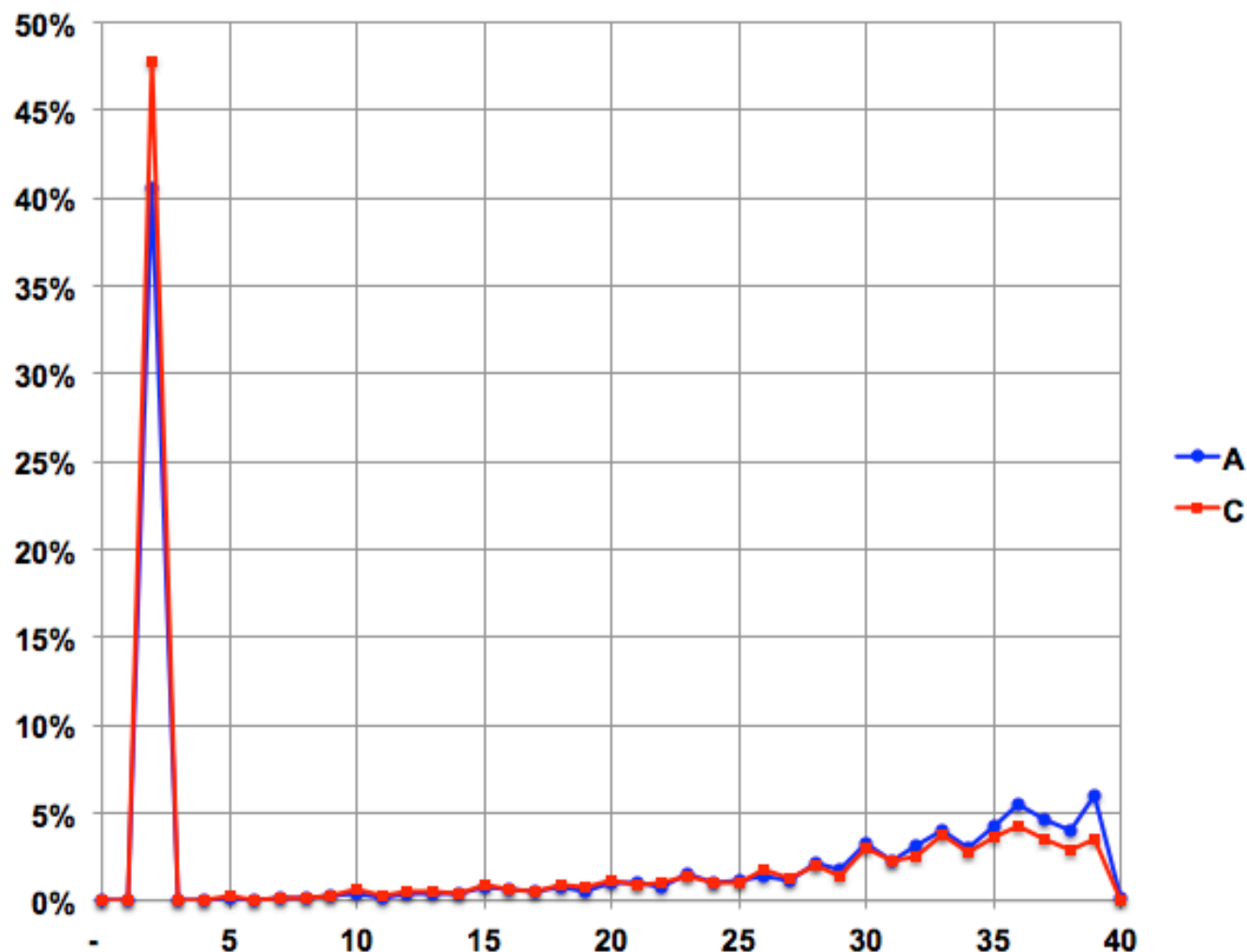
Distribution of Raw Q-Scores Along 100-base Reads



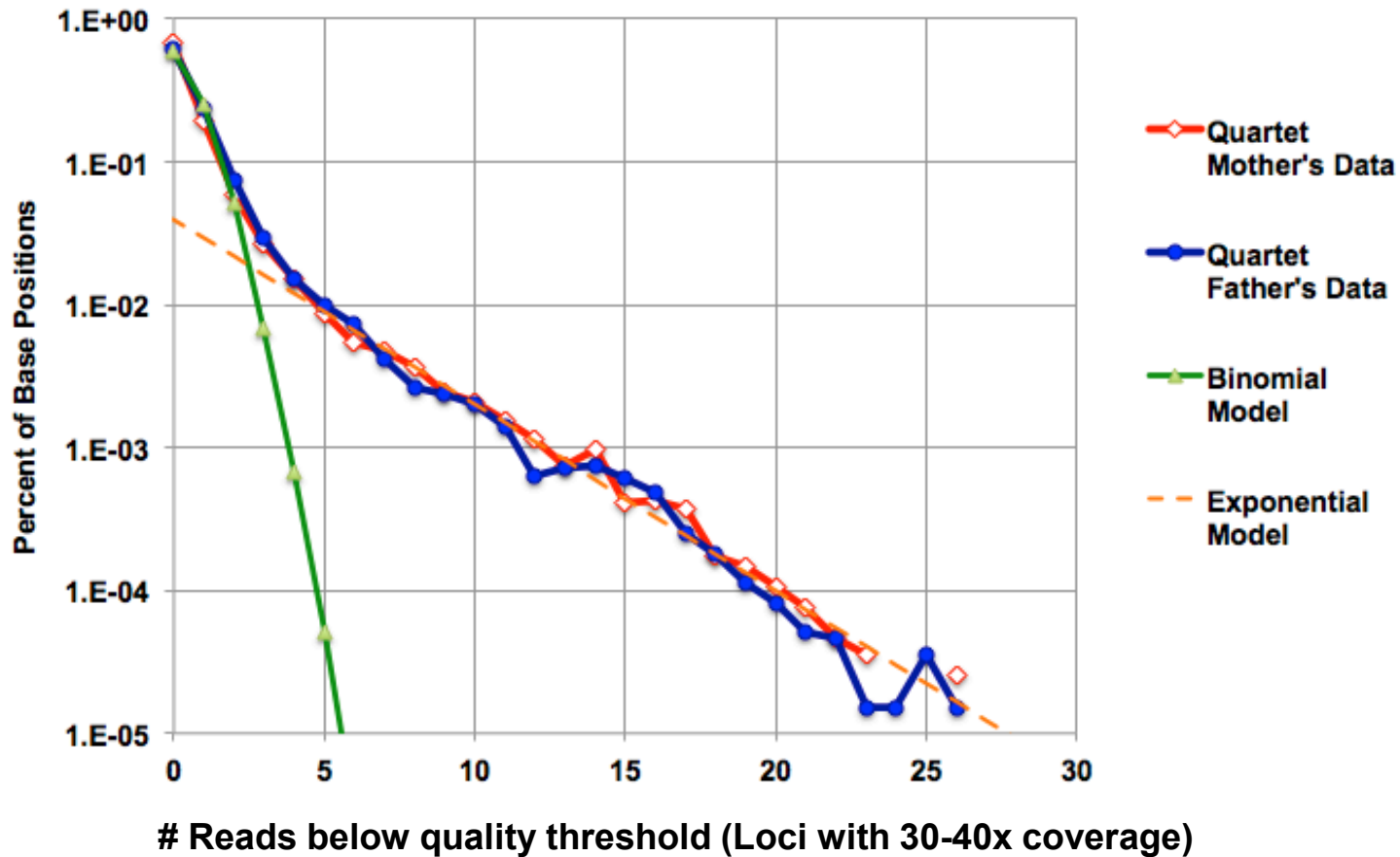
Median Raw Q-Score Along a Read, By Base



Raw Q-Score Histogram by Base at Position 100 in Each Read

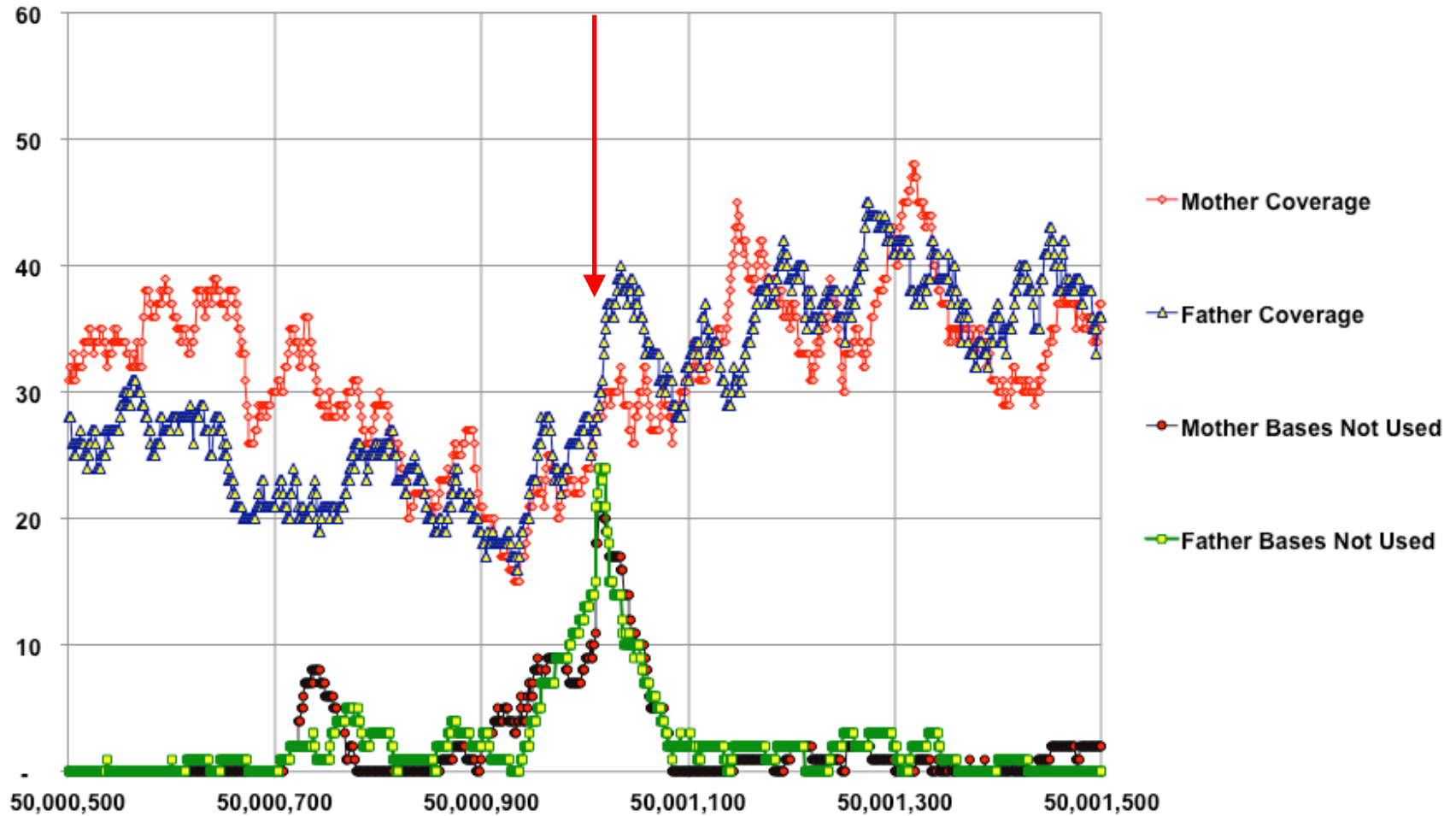


Occurrence of bases with very low raw quality scores deviates markedly from a random (binomial) model



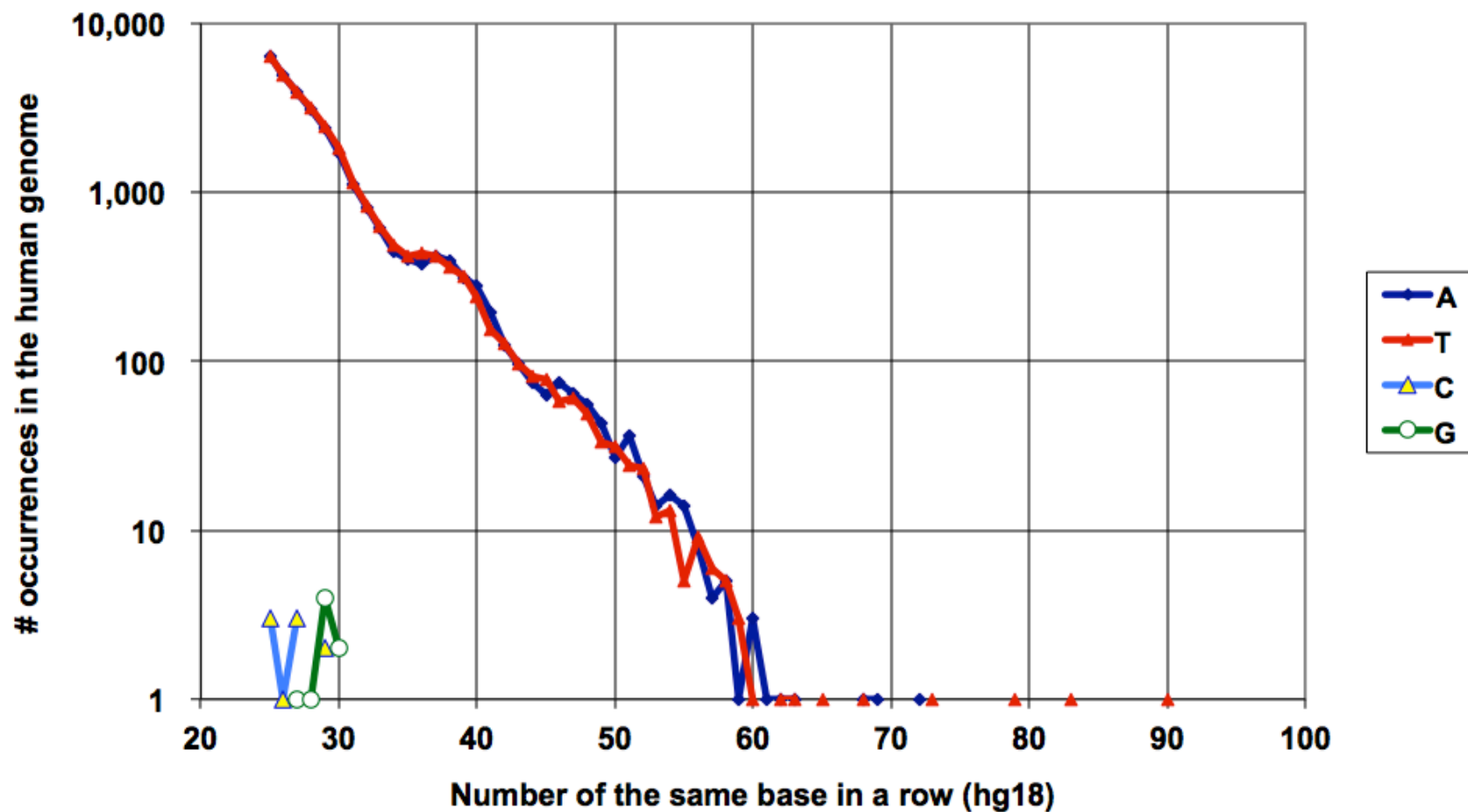
Profile of systematic errors is remarkably similar between these two genomes

Signature Around 24 A's in a Row

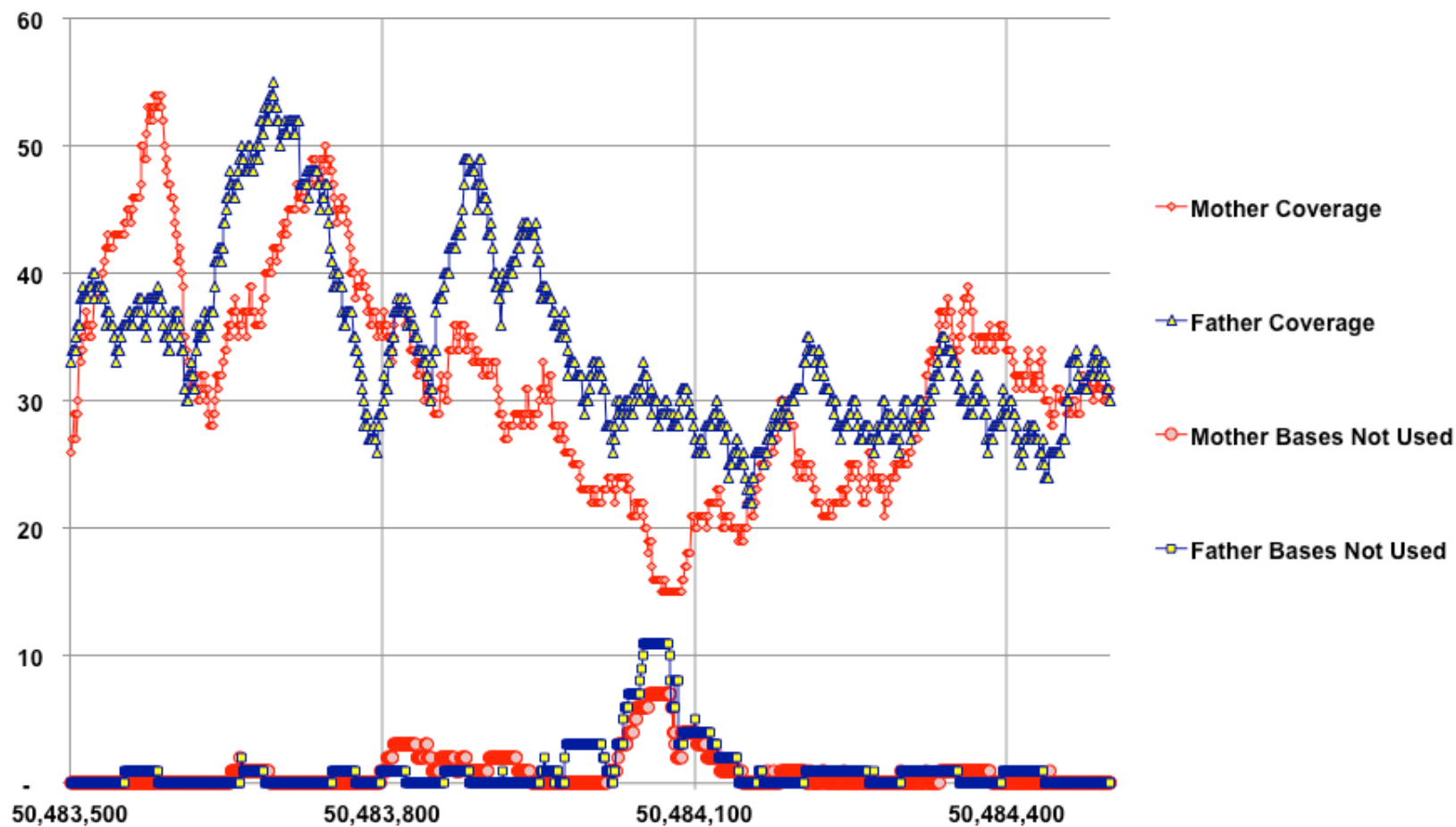


There are over 50,000 homopolymers with length > 24 in the human genome
Homopolymers of G's or C's are expected to be more severe

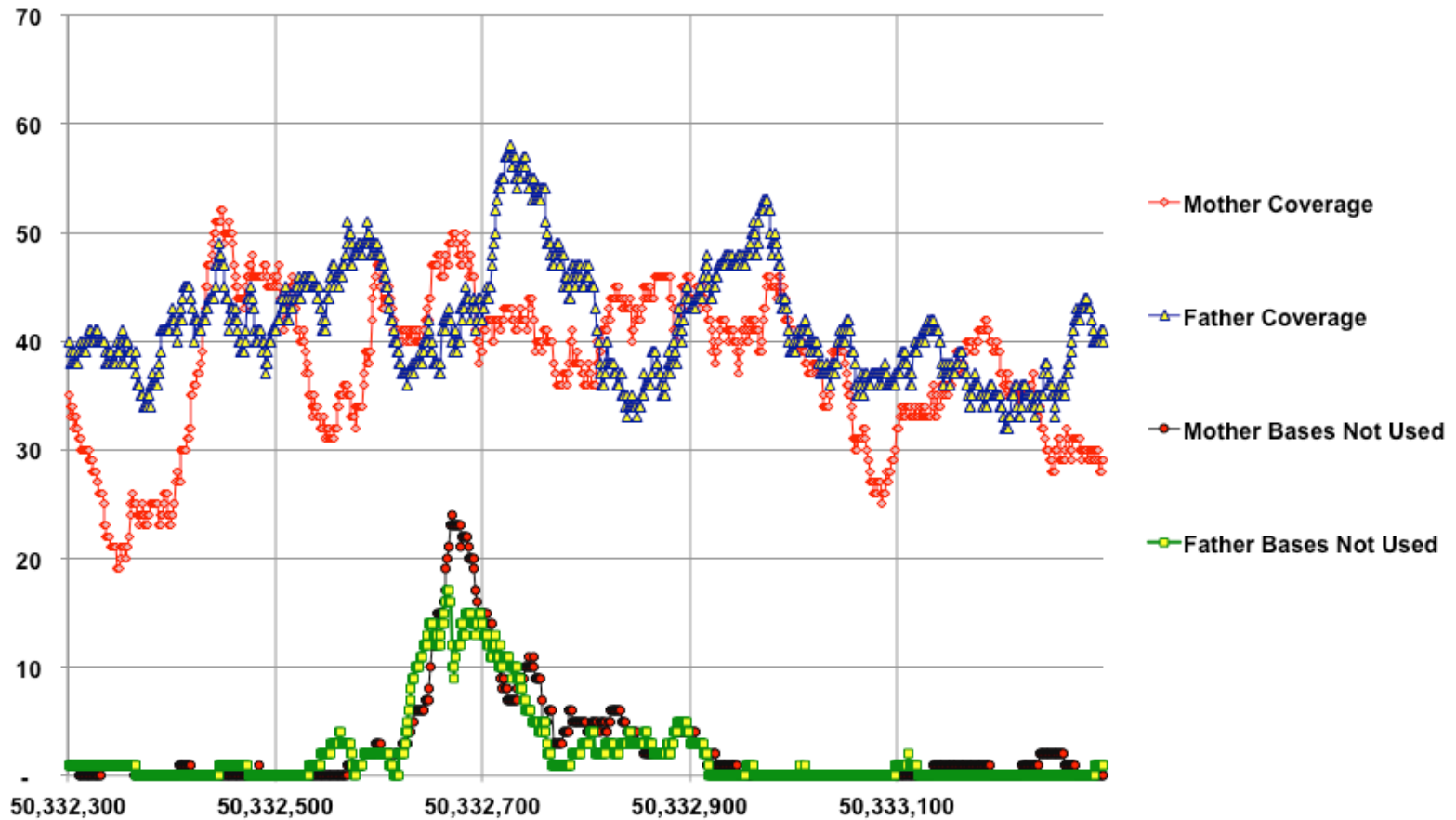
Histogram of Homopolymer Sequence Runs by Base
(Minimum length for histogram = 25; Total occurrences = 57,272)



Signature Around a 28 x GT Di-Nucleotide Repeat



Signature in Hairpin Area: 16-base stem & 56-base loop

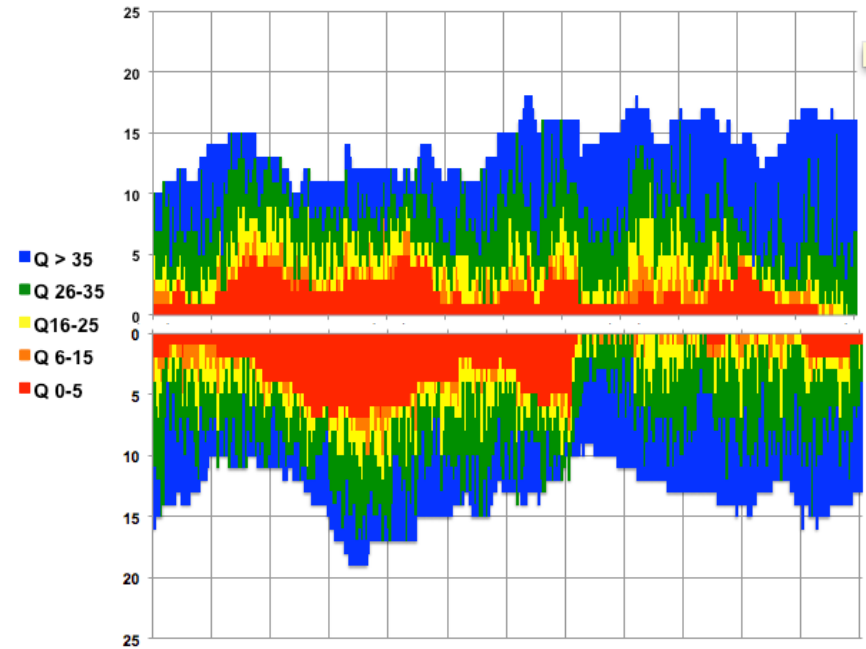


Q-Score Distribution of Reads Across a Region



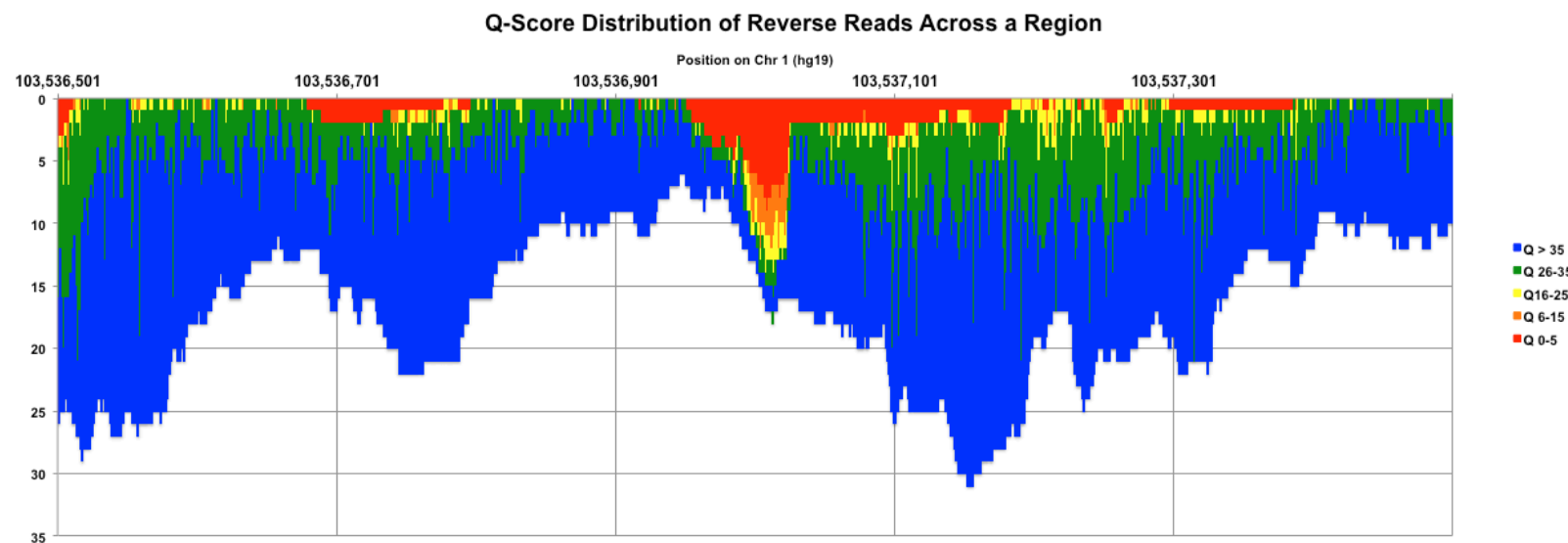
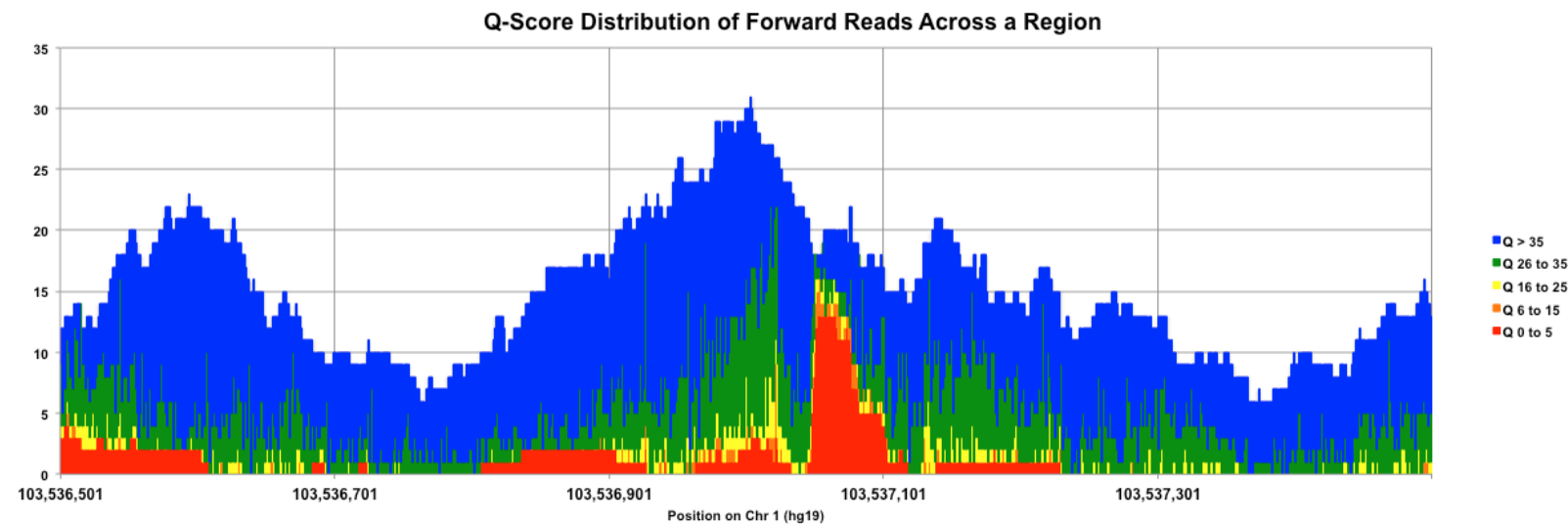
Region with “normal”
sequence

Q-Score Distribution of Reads Across a Region

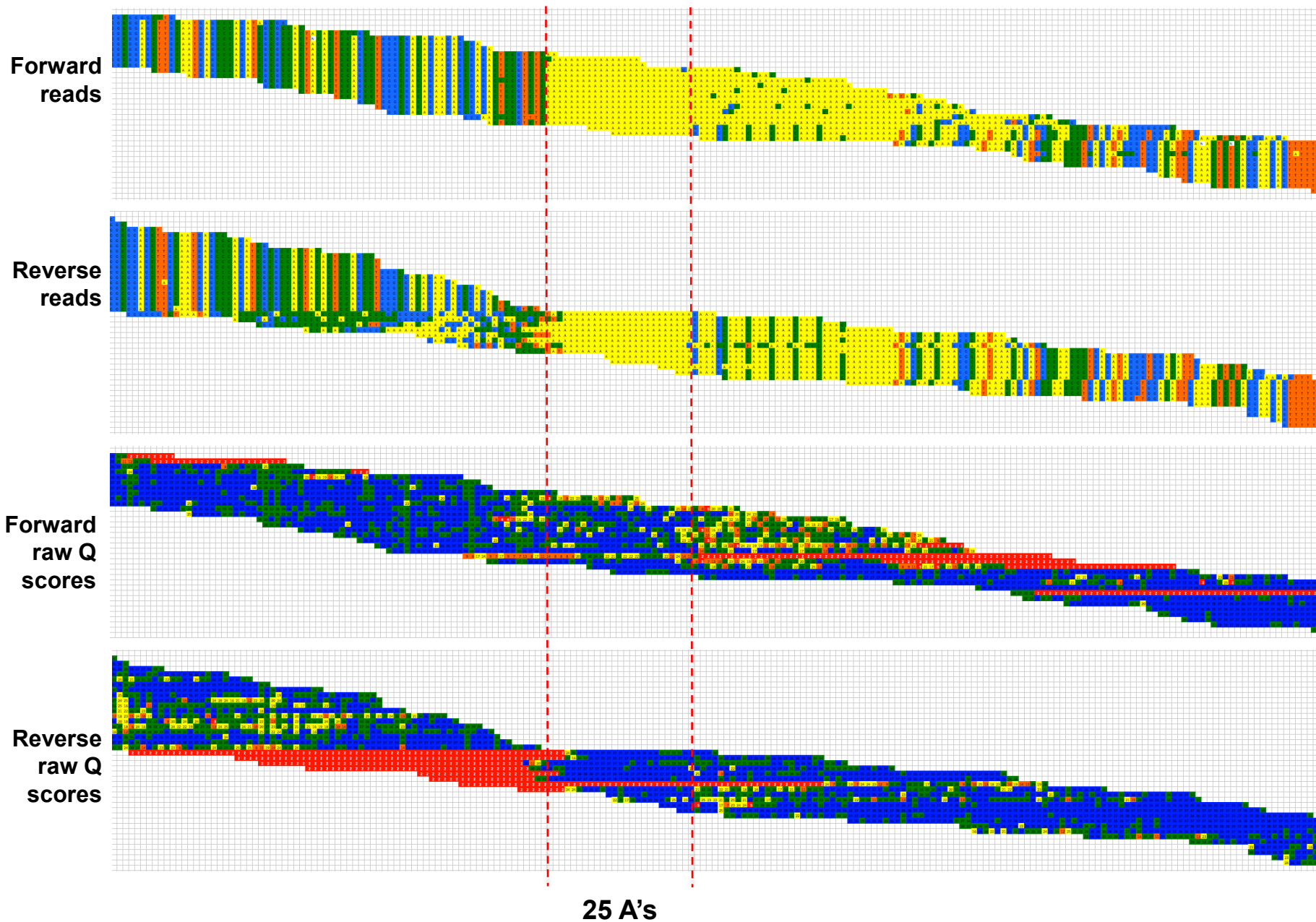


Region with “problematic”
Sequence motif

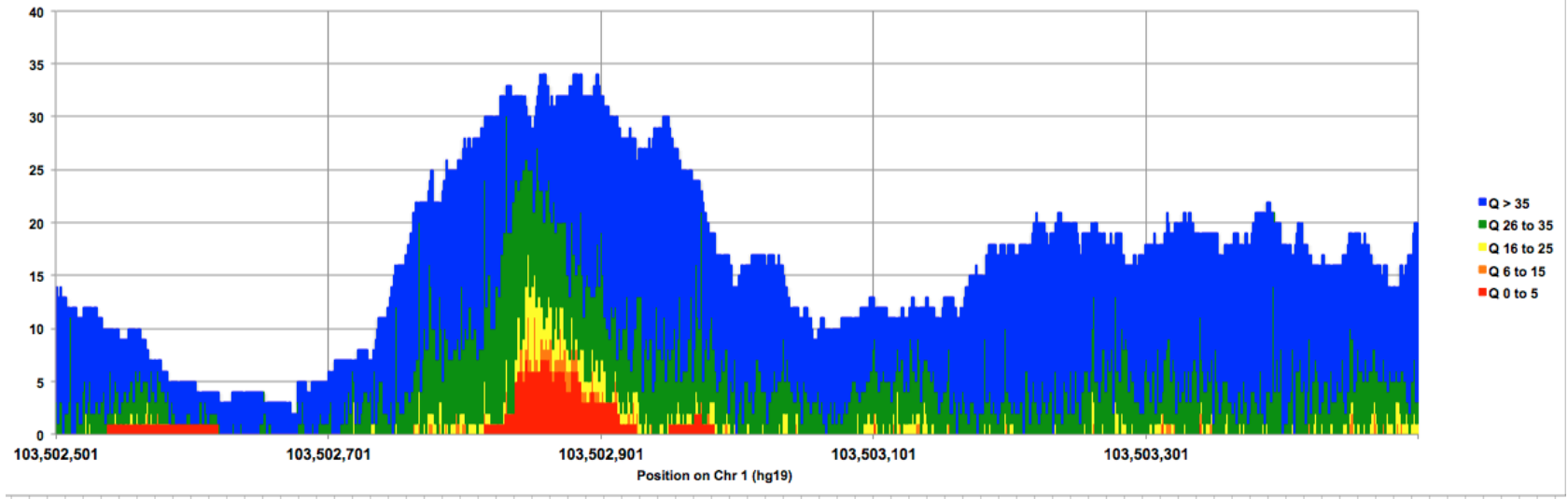
Tools to search for systematic errors across the genome, with their underlying mechanisms of action, will be followed by bio-informatic flags and (more importantly) bio-informatic and laboratory solutions.



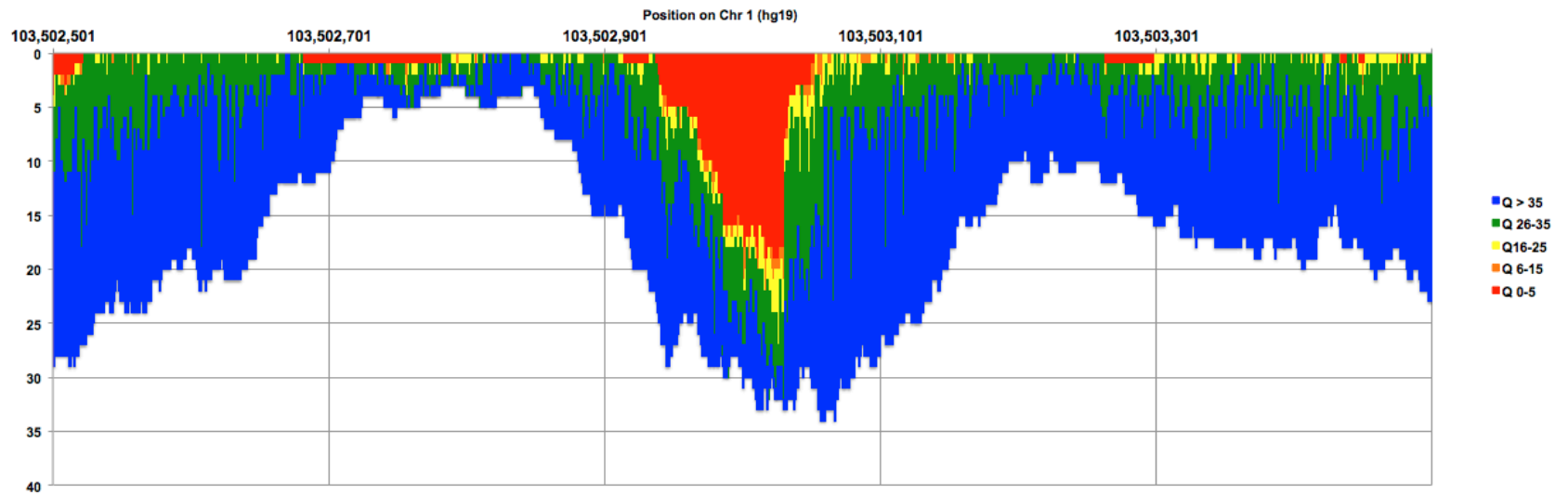
Directional raw Q-score distribution around a 25 A homopolymer
The impact is downstream, not in the feature



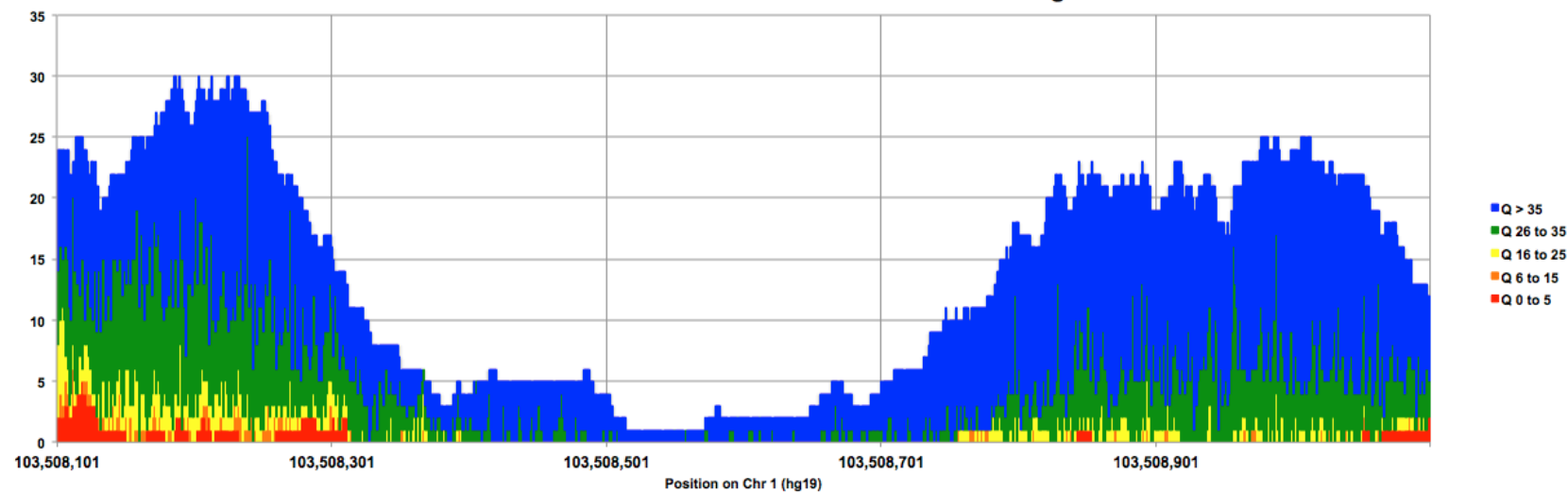
Q-Score Distribution of Forward Reads Across a Region



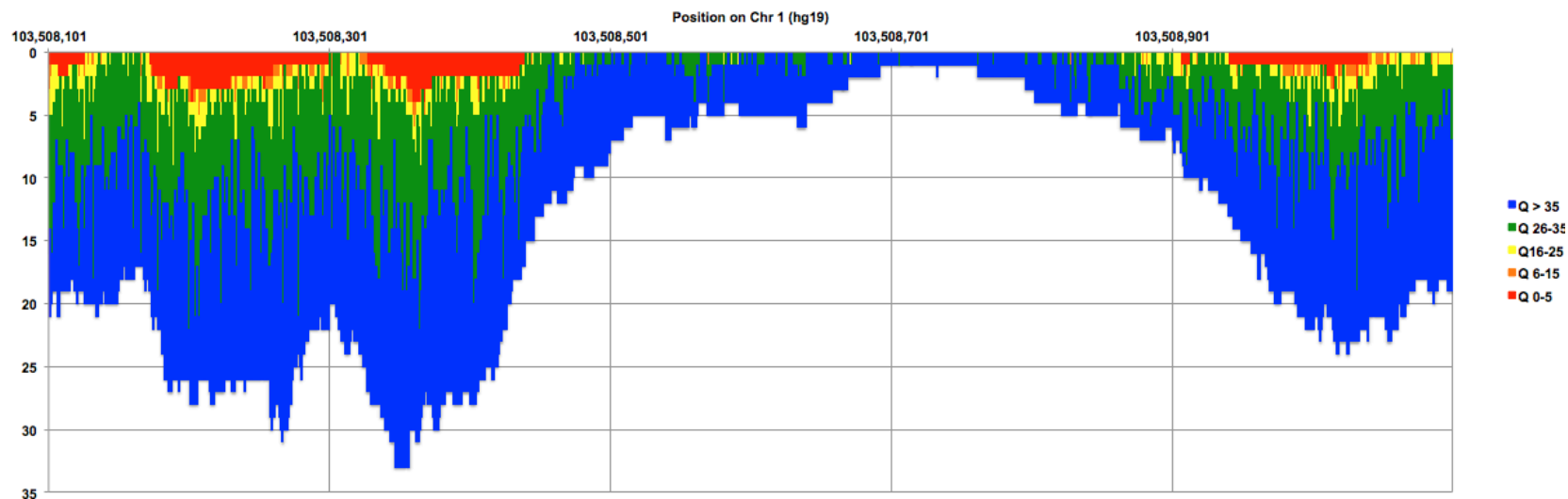
Q-Score Distribution of Reverse Reads Across a Region



Q-Score Distribution of Forward Reads Across a Region



Q-Score Distribution of Reverse Reads Across a Region



Chr 1 : 103,508,600.

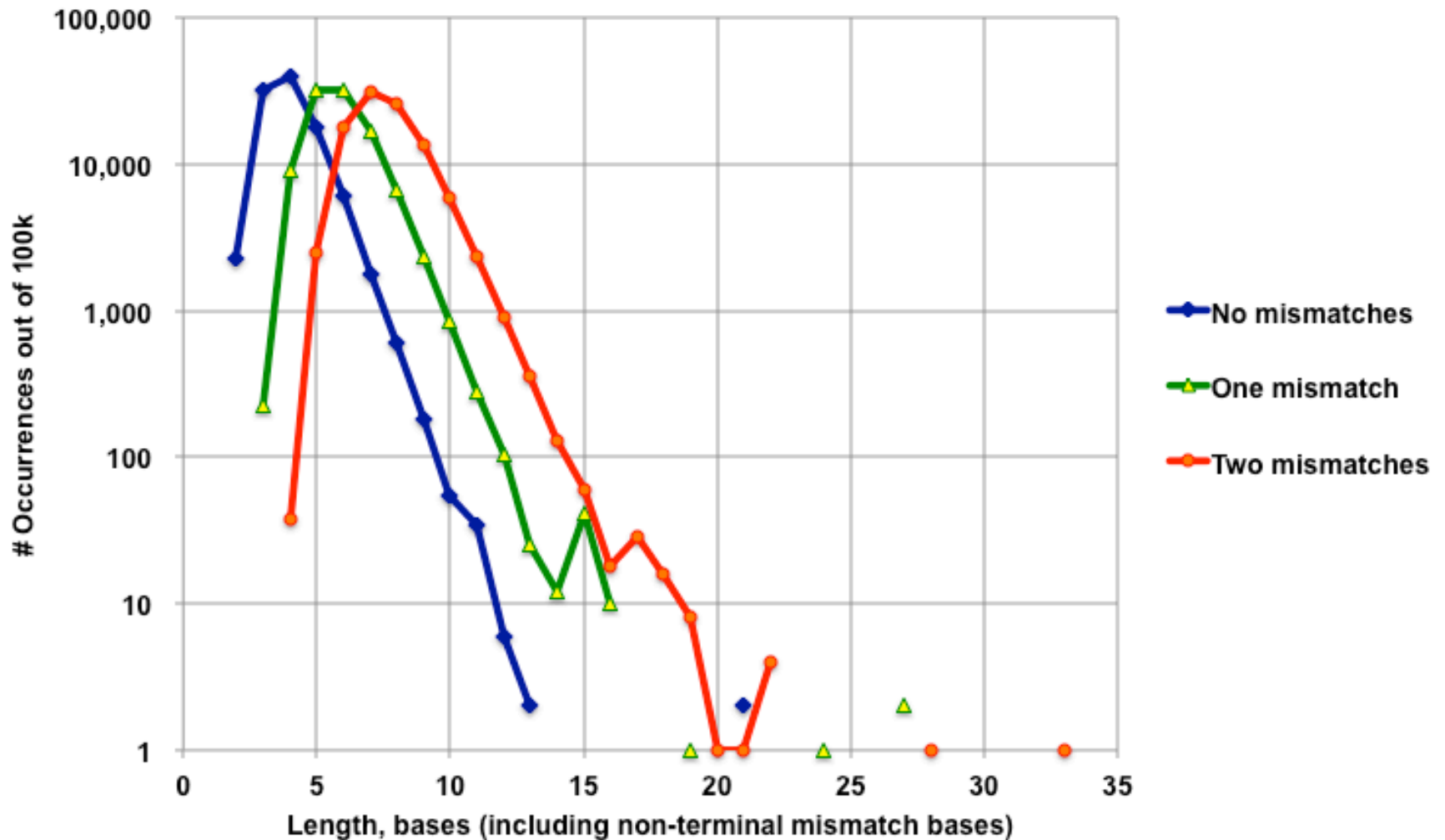
Coverage almost goes to zero, in each direction separately and somewhat offset, at an 18 x 18-mer repeat.

PG20. Major cluster of palindromes (74 of them), most with 16-19 base matching lengths.

Personalis EX2055

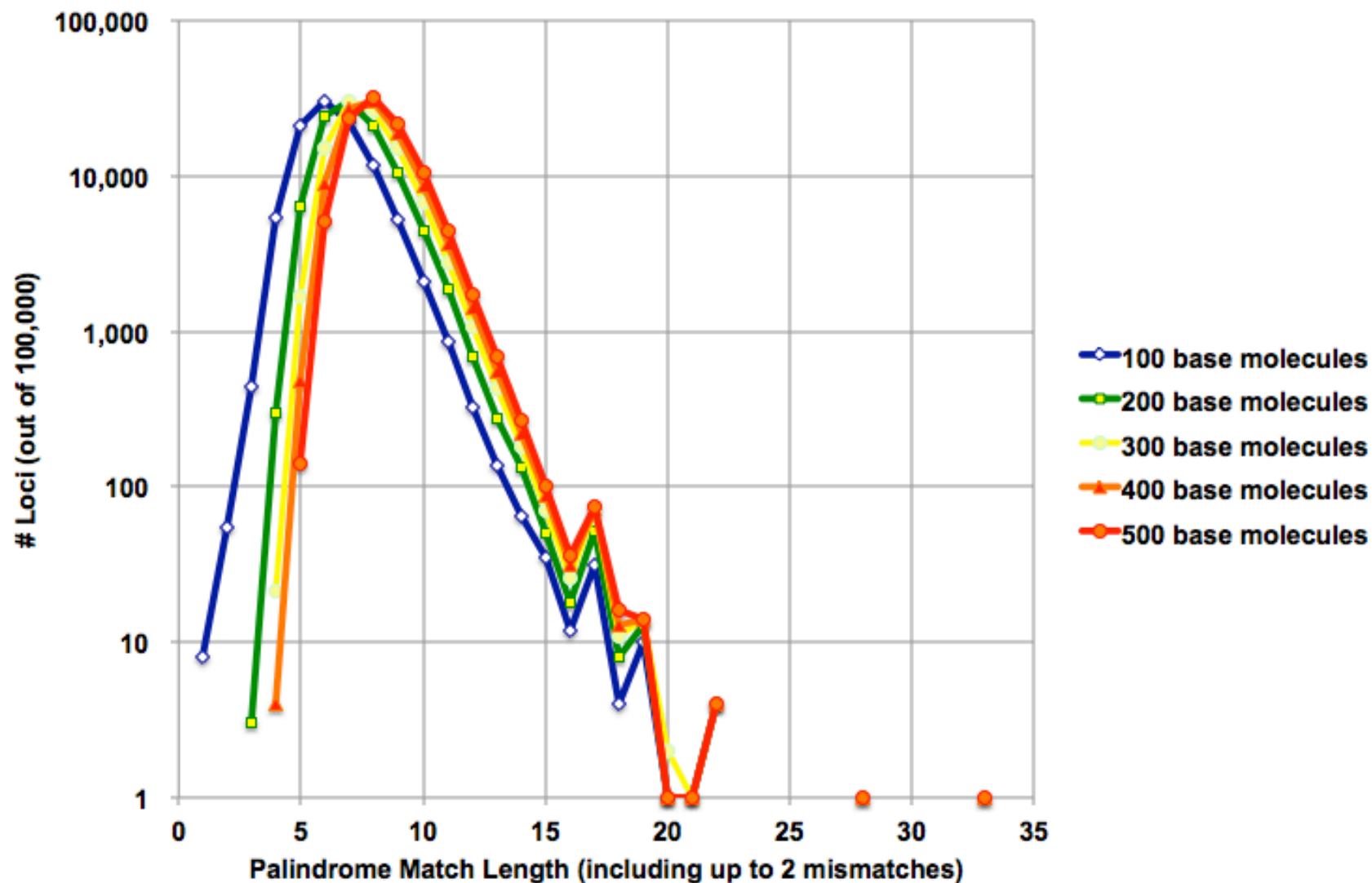
Palindrome Length Histogram - 500bp molecules

NCBI build 37, Sample region - Chr 1 : 103,500,300 - 103,600,300 (hg19)



Palindrome Match-Length Histogram vs Molecule Length

Test region - Chr 1 : 103,500,300 - 103,600,300 (hg19)



Sequence Motifs Which Underlie Some Errors

Discussion

- Studies Mike is aware of which identify these
- Bioinformatic methods to read through these
- Laboratory methods to read through these
- Contribution of this class of error mechanism to the overall next-gen sequencing error rate

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)

Snyder son/mother

Data behind publications:

Illumina vs CGI

Snyderome paper

Some inheritance validation

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

Korean PGP

40 HiSeq 2x100 genomes

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

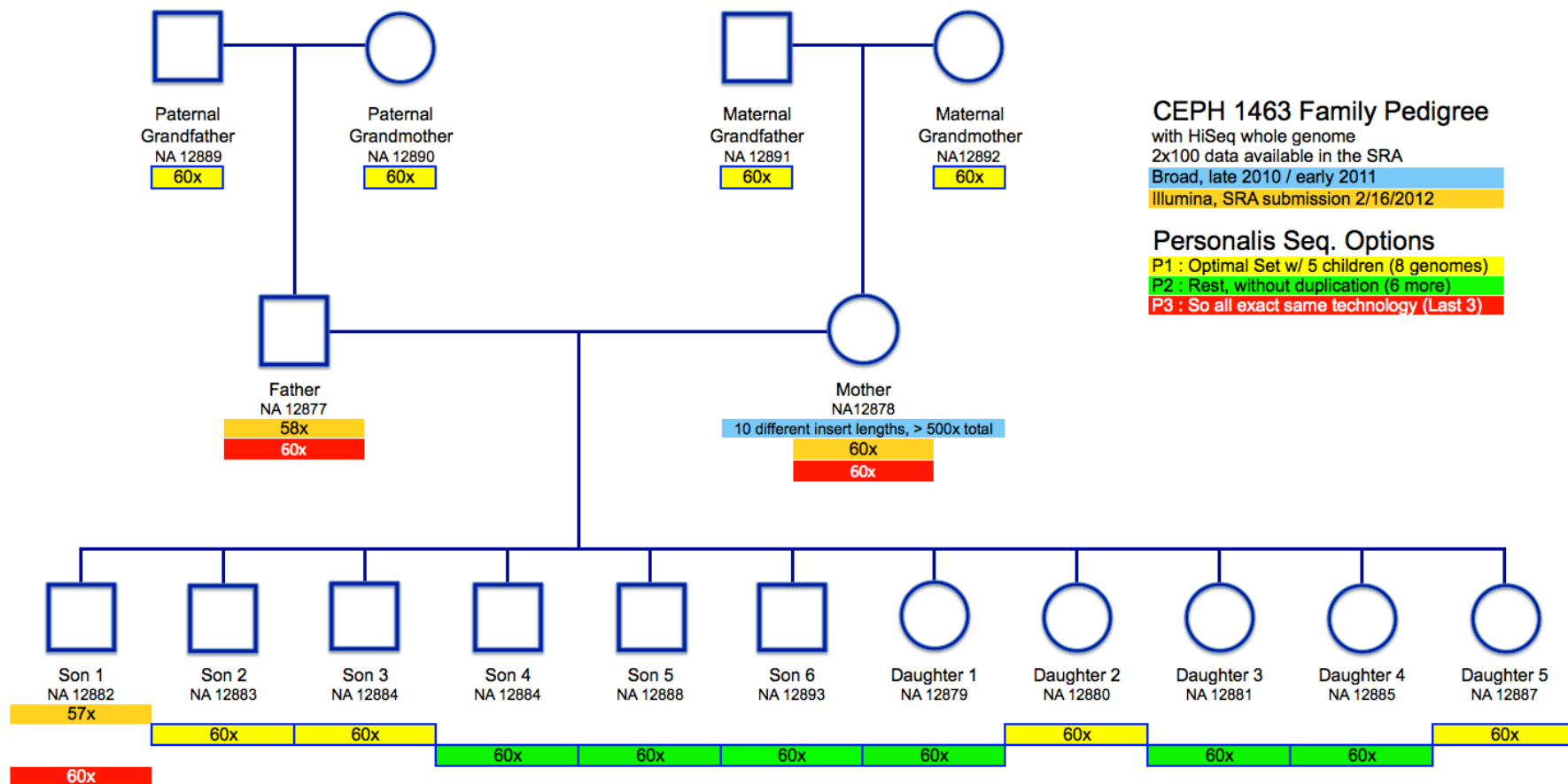
MiSeq 2x250

Anonymous samples
(ethics)

No Personalis relationship

Venter Genome

Only individual whole
genome on Sanger
technology



Segmentation of the 17-member pedigree by priority for sequencing

CEPH1463 Pedigree Analysis

Percent of the Genome Over Which We Can See, Across the Data From All of the Children Included:						
	All Inheritance states		All Parental Alleles		All Genotypes possible from parents	
	Once	Twice	Once	Twice	Once	Twice
All 11 Children	99.9%	99.9%	100.0%	96.8%	84.0%	32.0%
The combination of 5 children which maximizes % genome where all Inheritance States are represented at least once (Children # 1,3,7,8,11)	91.0%	36.8%	94.7%	35.1%	22.0%	0.0%
As above, but at least twice (Children # 2,8,9,10,11)	87.4%	46.0%	89.1%	35.9%	20.6%	0.0%
All I.S. at least once, but meeting the criteria from X chromosome analysis (Children # 2,3,5,6,11)	89.5%	36.8%	92.9%	38.6%	26.4%	0.0%
As above, but using data from Son # 1 (recently sequenced by Illumina & submitted to the SRA) (Children 1,2,3,8,11)	88.3%	35.5%	92.6%	38.0%	24.9%	0.0%

All 11 children

Several optimal subsets with 5 of the 11 children

Mendelian Inheritance Error analysis can be confounded by genomic variation induced by the cell culture process. Being able to see each parental allele in two separate children should let us separate most of these errors from sequencing artifacts, letting us better identify them.

"GOLD STANDARD GENOME" PROJECT BUDGET

Amount is within budget but requires Board approval

Additional semi-custom work based on Personalis content may be proposed later in the year.

Samples de-identified before shipment to vendors

Whole genome sequencing

By Illumina Genome Network at Illumina	\$ 122,000
20 samples at 60x coverage	
17 CEPH1463 pedigree	
2 replicates	
Venter	
Complete Genomics Sequencing	\$ 10,000
2 Samples at 80x coverage	
Venter	
CEPH1463 Mother (replicate)	
Array analysis	\$ 5,000
Genotypes	
SV validation	
Karyotyping, exomes & other assays	\$ 20,000
TOTAL	\$ 157,000

Personalis Laboratory Strategy

- Laboratory goals:
 - Direct relationship with the customer
 - Quality control, sample to results (important for research and later clinical)
 - Position as “Value-Added Genome”, not “a bio-informatics company”
 - Offer a superior genome, by combining technologies from multiple companies
 - Develop superior sample preps, esp based on Personalis content
- Outsourcing:
 - Whole genome sequencing service pricing now < reagents to do it yourself
 - Avoid infrastructure costs (instruments, initial computation, staffing)
 - Turn-around time is a disadvantage (quotes at 12-16 weeks)
 - Other assays we can outsource:
 - Arrays (standard & semi-custom)
 - Karyotyping
 - Exomes, RNAseq
- Personalis lab
 - Receive, de-identify samples
 - QC, sample tracking
 - Ship aliquots to multiple vendors
 - Validations following primary analysis
 - Develop custom sample preps

