Data processing and analysis of genetic variation using next-generation DNA sequencing

Mark DePristo, Ph.D.

Genome Sequencing and Analysis Group
Medical and Population Genetics Program
Broad Institute of MIT and Harvard

IPAM Oct. 2011

Acknowledgments

Genome sequencing and analysis (GSA)
Eric Banks
Ryan Poplin
Guillermo del Angel
Kiran Garimella
Chris Hartl
Matt Hanna
Khalid Shakir
David Roazen
Mauricio Carneiro

Genome Sequencing Platform
In general but notably:
Tim Fennell
Alec Wysoker
Toby Bloom

Broad postdocs, staff, and faculty
Anthony Philippakis
Manny Rivas
Jared Maguire
David Jaffe
Bob Handsaker
Steve McCarroll
Menachem Fromer
Kristian Cibulskis
Andrey Sivachenko
Gad Getz
Aaron McKenna

IGV
Jim Robinson
Helga Thorvaldsdottir

1000 Genomes project
In general but notably:
Richard Durbin
Goncalo Abecasis
Matt Hurles
Richard Gibbs
Gabor Marth
Fuli Yu
Gil McVean
Gerton Lunter
Heng Li

MPG directorship
Stacey Gabriel
David Altshuler
Mark Daly

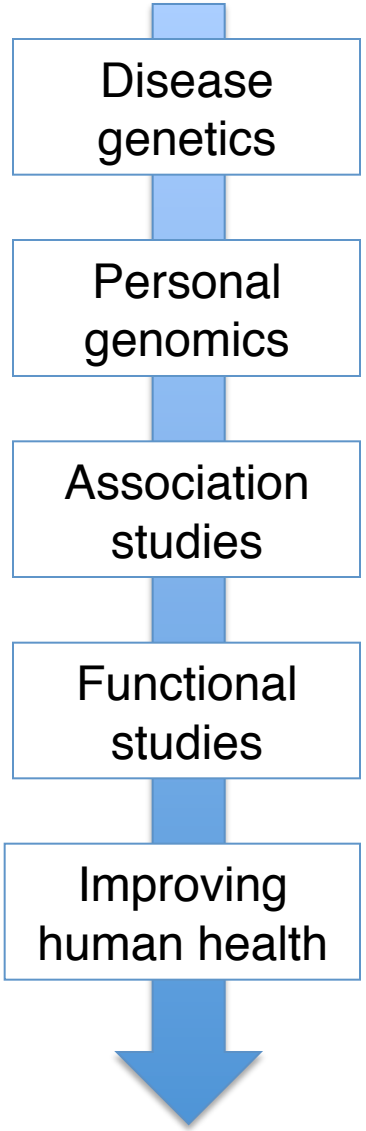
What is the BROAD INSTITUTE ?

The Broad Institute mission

1. This generation has a historic opportunity and responsibility to transform medicine by using systematic approaches in the biological sciences to dramatically accelerate the understanding and treatment of disease.

2. To fulfill this mission, we need new kinds of research institutions, with a deeply collaborative spirit across disciplines and organizations, and having the capacity to tackle ambitious challenges.

How is Medical and Population Genetics (MPG) at the Broad achieving these goals?



Many simple and complex human diseases are heritable

Affected and unaffected individuals differ systematically in their genetic composition*

These systematic differences can be identified by comparing affected and unaffected individuals

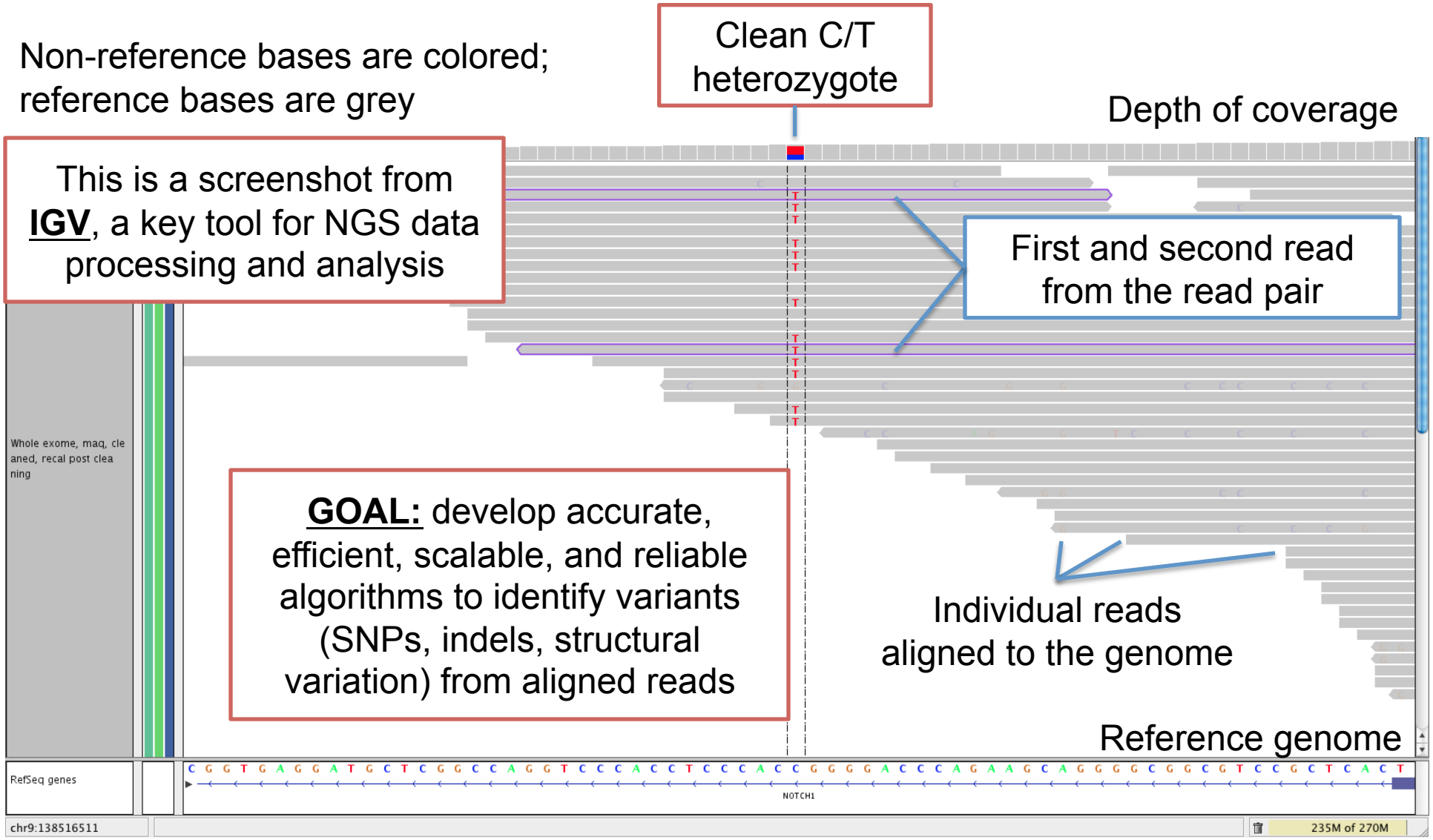
These associated variants give insight into the biological mechanisms of disease

These insights can be used to intervene in the disease process itself

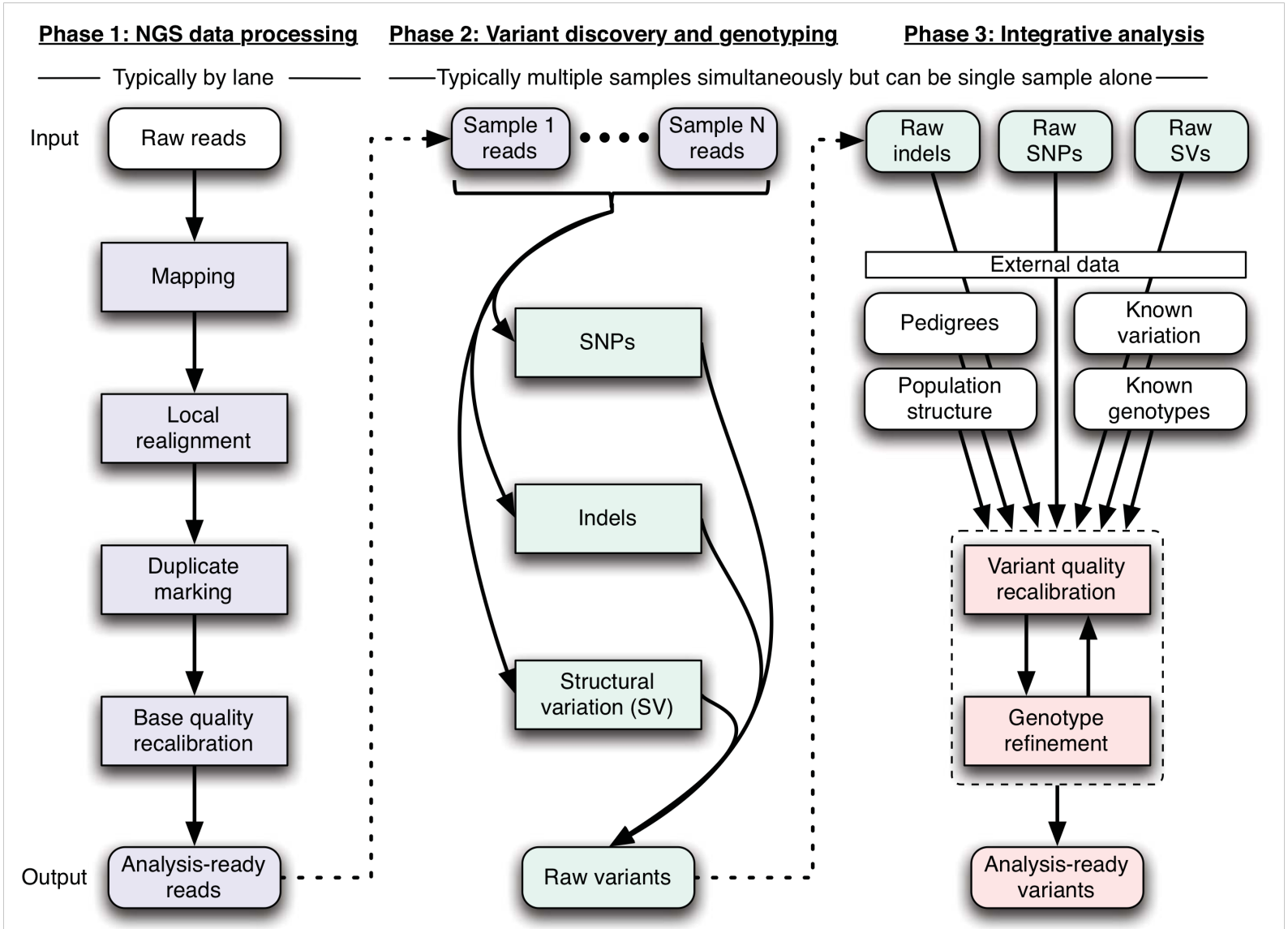
*Can be determined by next-generation sequencing

Detecting variants in next-generation sequencing data

NGS provides an unprecedented opportunity to characterize genetic variation in thousands of samples at a reasonable cost

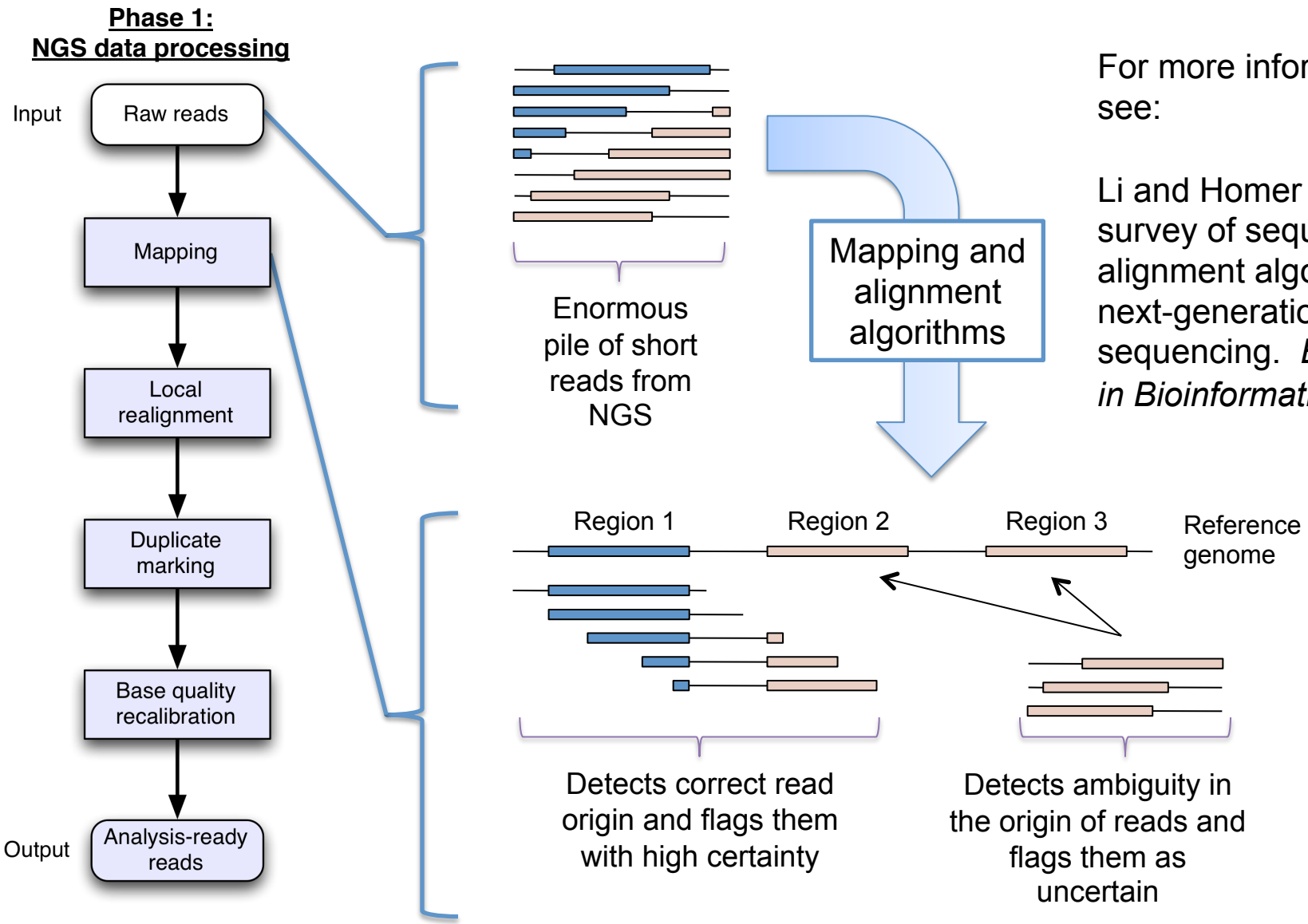


Our framework for variation discovery



DePristo, M., Banks, E., Poplin, R. et. al, (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet.

Finding the true origin of each read is a computationally demanding first step

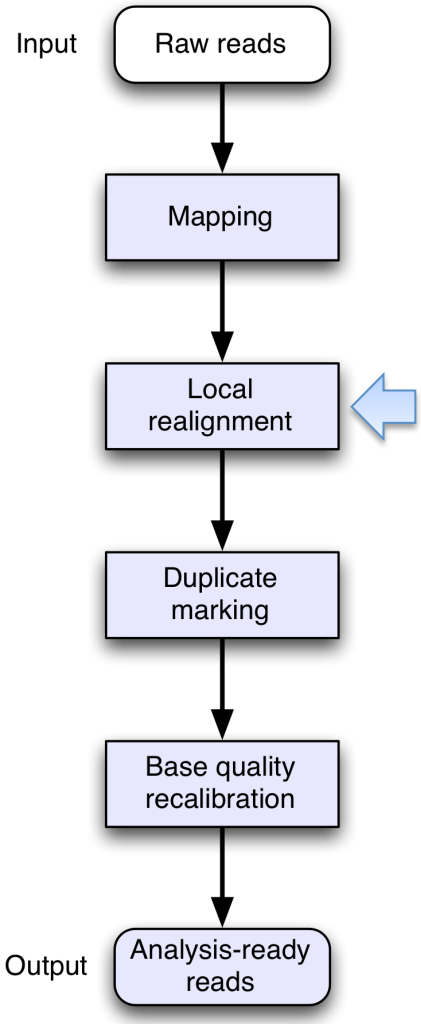


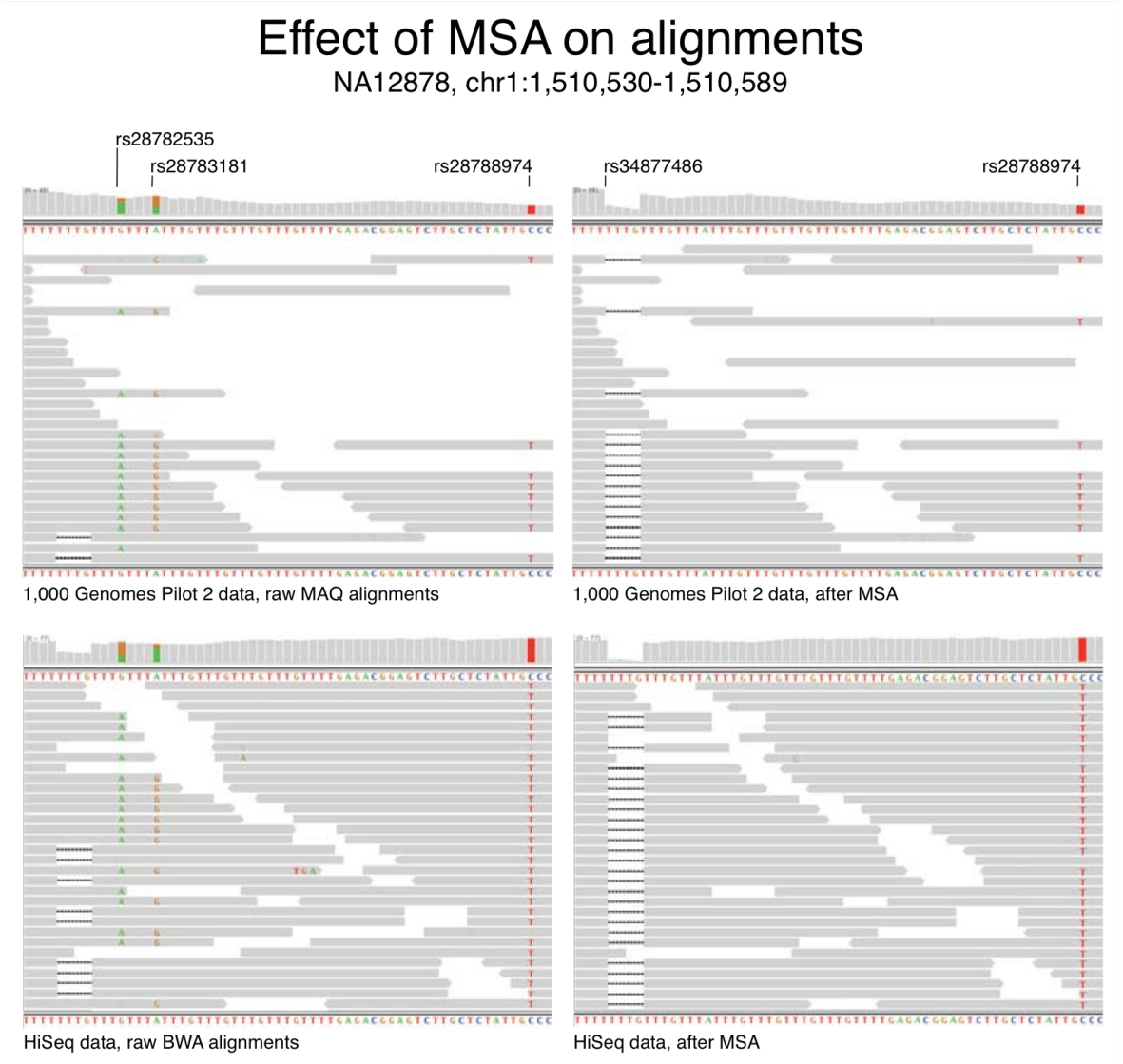
For more information see:

Li and Homer (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*.

Accurate read alignment through multiple sequence local realignment

Phase 1:
NGS data processing

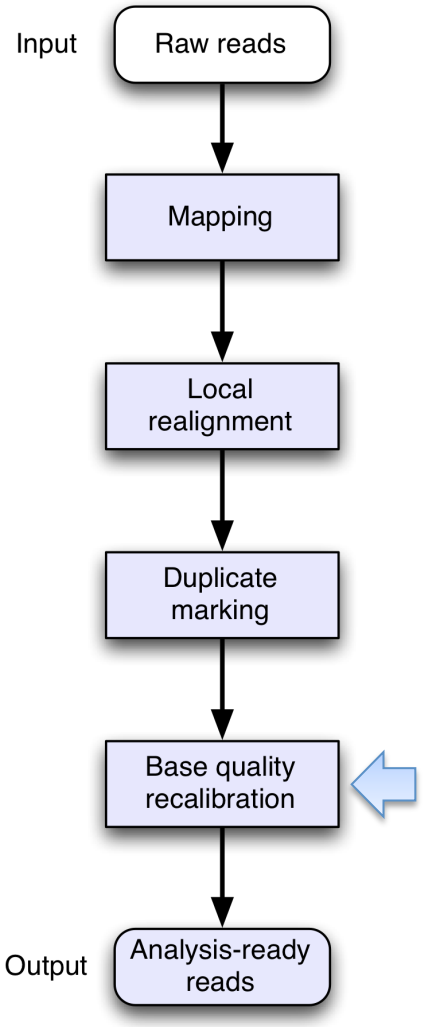


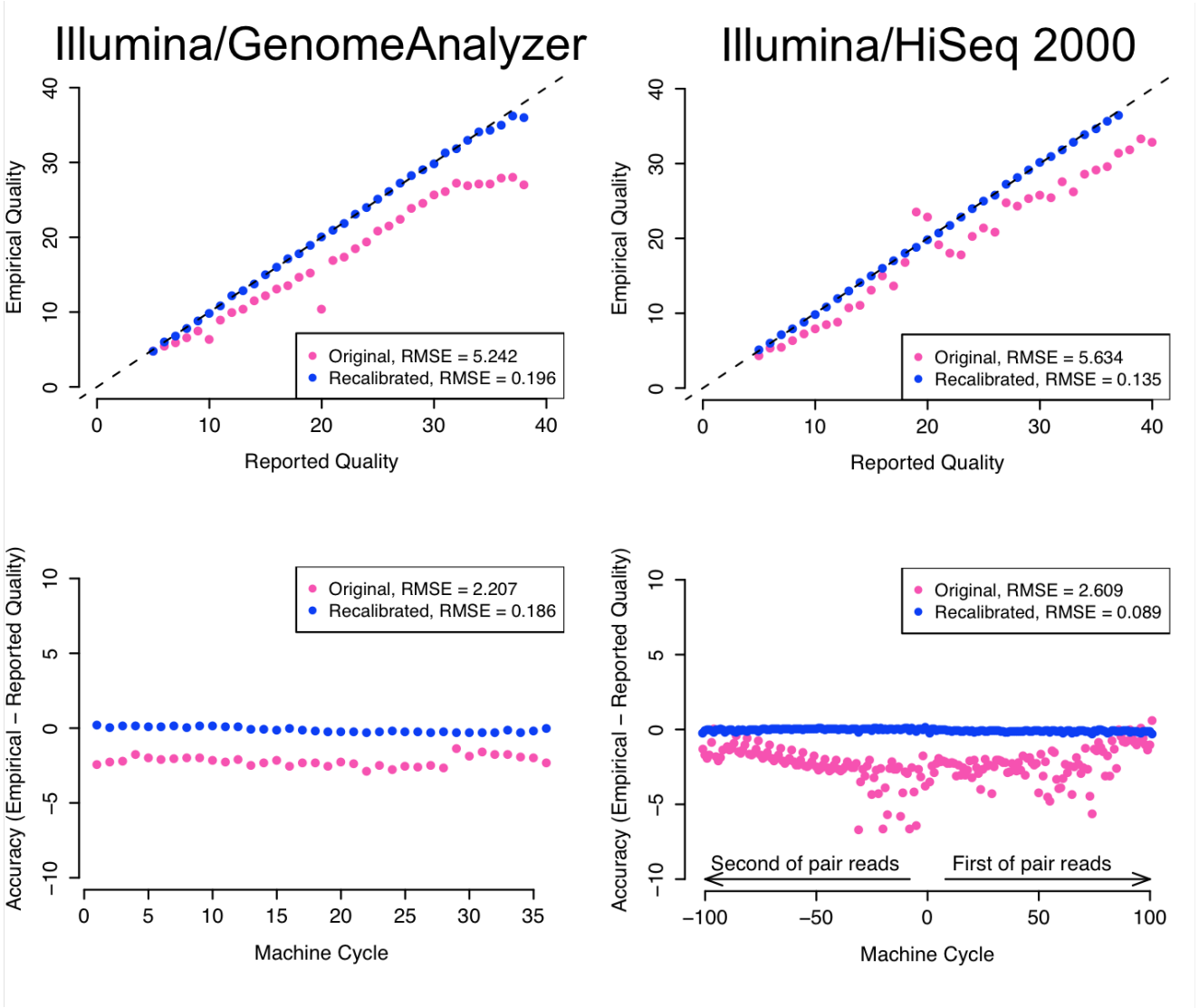


DePristo, M., Banks, E., Poplin, R. et. al, (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet.

Accurate error modeling with base quality score recalibration

Phase 1:
NGS data processing





Ryan Poplin

DePristo, M., Banks, E., Poplin, R. et. al, (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet.

SNP and Indel calling is a large-scale Bayesian modeling problem

Prior of the genotype

Likelihood of the genotype

Diploid assumption

Bayesian model

$$\Pr\{G|D\} = \frac{\Pr\{G\}\Pr\{D|G\}}{\Sigma_i \Pr\{G_i\}\Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$

$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2}\right) \text{ where } G = H_1H_2$$

$$\Pr\{D|H\} \text{ is the haploid likelihood function}$$

- Inference: what is the genotype G of each sample given read data D for each sample?
- Calculate via Bayes' rule the probability of each possible G
- Product expansion assumes reads are independent
- Relies on a likelihood function to estimate probability of sample data given proposed haplotype

See http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper for more information

SNP genotype likelihoods

$$\Pr\{D_j|H\} = \Pr\{D_j|b\}, \text{ [single base pileup]}$$

$$\Pr\{D_j|b\} = \begin{cases} 1-\epsilon_j & D_j = b, \\ \epsilon_j & \text{otherwise.} \end{cases}$$

- All diploid genotypes (AA, AC, …, GT, TT) considered at each base
- Likelihood of genotype computed using only pileup of bases and associated quality scores at given locus
- Only “good bases” are included: those satisfying minimum base quality, mapping read quality, pair mapping quality, NQS

See http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper for more information

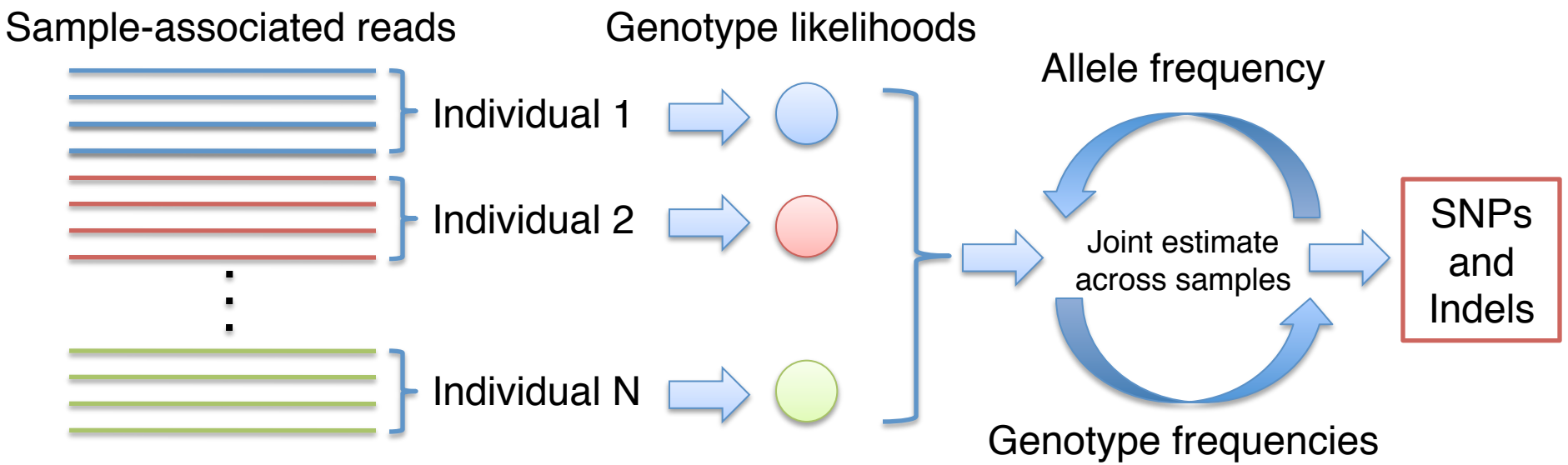
Indel genotype likelihoods

$$\Pr\{D_j|H\} = \sum_{\substack{\text{alignments } \pi \\ \text{of } D_j \text{ to H}}} Pr\{D_j, \pi\}$$

- Haplotypes H_i are discovered from indels in the reads
- Diploid genotypes G for all haplotype H_iH_j combinations
- For each haplotype H_i, calculate likelihood of each read D_j marginalizing over all possible alignments π
- Sum computed by a standard HMM with context-dependent affine gap penalties using haplotype and read bases and quality scores

See http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper for more information

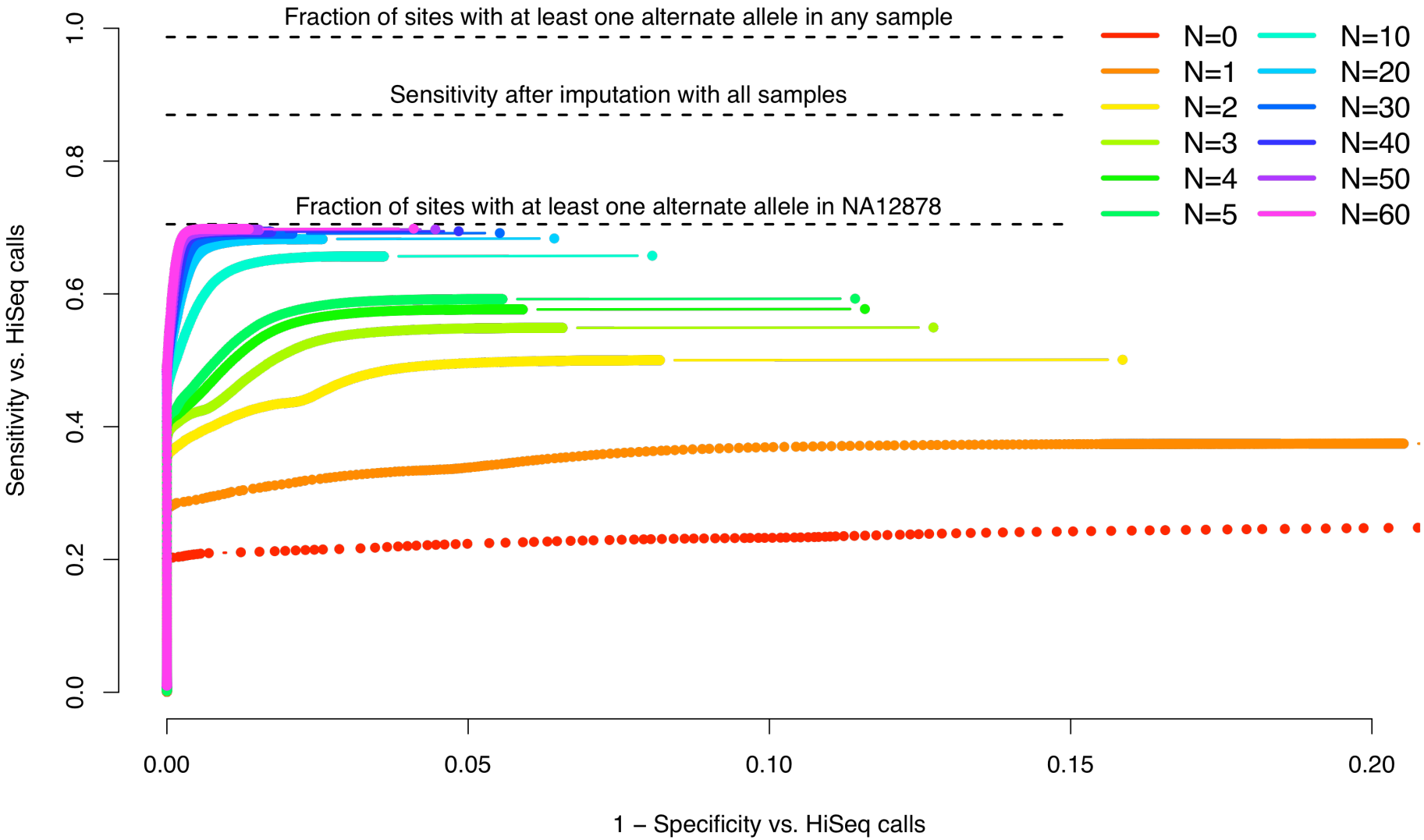
Multi-sample calling integrates per sample likelihoods to jointly estimate allele frequency of variation



- Simultaneous estimation of:
  - Allele frequency (AF) spectrum Pr{AF = i | D}
  - The probability that a variant exists Pr{AF > 0 | D}
  - Assignment of genotypes to each sample

Excellent mathematical treatment at http://lh3lh3.users.sourceforge.net/download/samtools.pdf

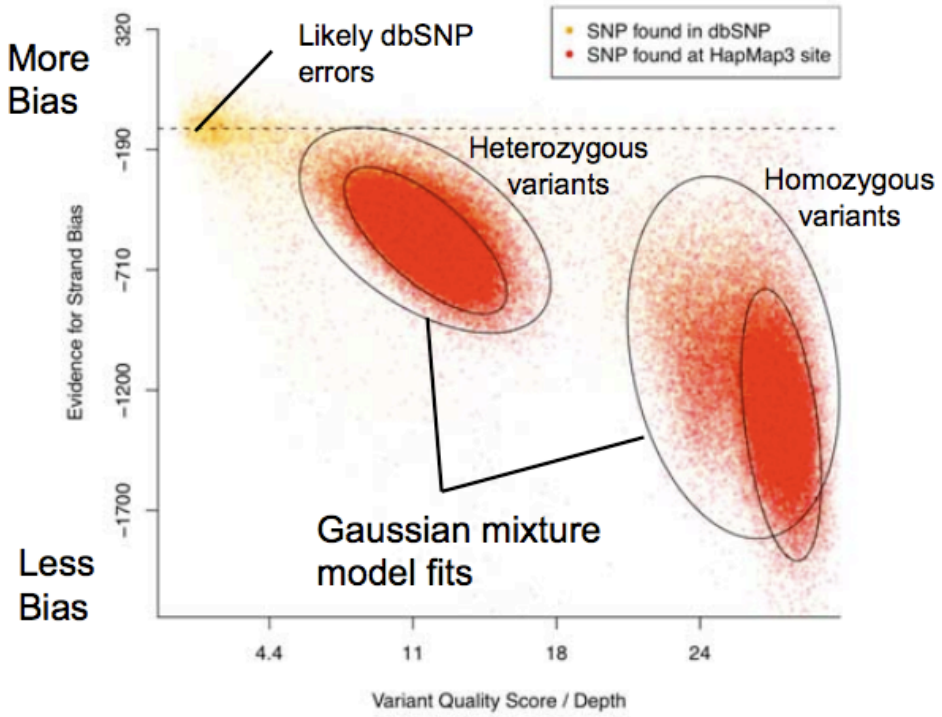
Discovery of HiSeq sites for NA12878 + N other samples



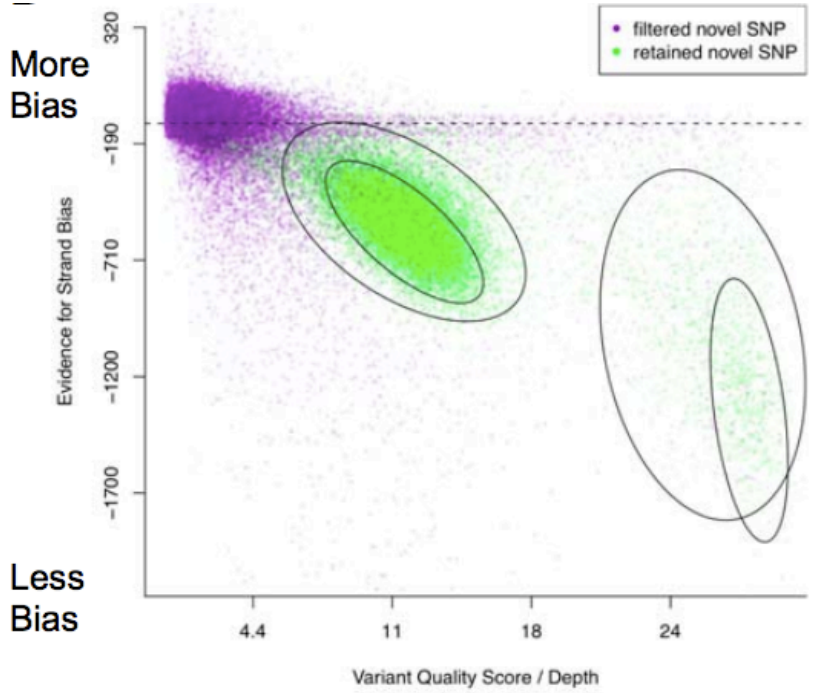
DePristo, M., Banks, E., Poplin, R. et. al, (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet.

Variant Quality Score Recalibration: modeling error properties of real polymorphism to determine the probability that novel sites are real

The HapMap3 sites from NA12878 HiSeq calls are used to train the GMM. Shown here is the 2D plot of strand bias vs. the variant quality / depth for those sites.



Variants are scored based on their fit to the Gaussians. The variants (here just the novels) clearly separate into good and bad clusters.



DePristo, M., Banks, E., Poplin, R. et. al, (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet.

Standard error modeling annotations

Annotation	Error mode it detects
QD	Little evidence per sample for variant triggering call
HaplotypeScore	Local read misalignment
ReadPosRankSum	The variant occurs only at specific cycles in the reads
FS	The variant occurs only on one read orientation (strand)
InbreedingCoeff	The sample genotypes are unlikely under Hardy-Weinberg
MQ	The absolute mapping quality of the data is low
MQRankSum	Variant reads are worse mapped than reference reads

- Train positive model with HapMap3, Omni, and best tranche of 1000 Genomes calls
- Train negative model with worst 1% of calls and worst tranche of 1000 Genomes calls

See http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v3 for more information

Methods and data are rapidly improving for all technologies

		Known		Novel		
Call set	**Date**	**# calls**	**Ti/Tv**	**# calls**	**Ti/Tv**	**Changes**
1000 Genomes 629 samples (Illumina, SOLiD, 454 data) on chr20	August 2010	183K	2.40	364K	2.30	+ Contrastive VQSR
	January 2011	184K	2.39	473K	2.28	+ EXACT allele frequency calculation
	August 2011	188K	2.34	506K	2.39	+ 1000G phase I samples and error covariates
HiSeq NA12878 at 64x coverage	May 2010	3.19M	2.10	351K	1.97	Hand-crafted hard filters
	August 2010	3.22M	2.15	362K	2.05	+ Initial version of VQSR
	December 2010	3.21M	2.16	335K	2.13	+ Contrastive VQSR
	August 2011	3.25M	2.13	283K	2.10	+ 1000G error covariates and decoy reference

Similar improvements for all experimental types
(low-pass and high-pass, targeted and whole-genome)

How well does this all work?

- SNPs
  - > 98.5% confirmation rate for variation discovery in 1100 4x samples in 1000G*
    - At least for “easy” sites in the genome
  - 98% of singletons in 2500 deep exomes**
  - 78/79 *de novo* SNPs confirmed in Autism trios***
- Indels
  - 1000G validation underway, but likely ~ 5%
    - At least for “easy” sites in the genome
    - Only for biallelic indels, not multi-allelic indels
  - Significant false negative rates for large events, especially large insertions
    - E.g., ~50% false negative rate for large (>15 bp) indels
  - Indel calling is the future challenge
    - More later

*BWA+GATK consensus calling and VQSR filtering from multiple input methods; **BWA + GATK data processing + UMich variant calling in ESP; *** Standard Broad pipeline in Autism ARRA project

These methods are available in the Genome Analysis Toolkit (GATK)

Genome Analysis Toolkit (GATK)

- Open-source map/reduce programming framework for developing analysis tools for next-gen sequencing data
- Easy-to-use, CPU and memory efficient, automatically parallelizing Java engine

1000 genomes GATK tools

Local realignment	Base error modeling	Q-score recalibration
HLA typer	Genotyping and SNP discovery	Indel genotyper
SNP filtering	De novo mutation finder	Many smaller analysis tools

- Most Broad Institute tools for the 1000 Genomes have been developed in the GATK

http://www.broadinstitute.org/gsa/wiki/

SAM/BAM format

- Technology agnostic, binary, indexed, portable and extensible file format for NGS reads
- Also used in the Broad production pipeline

http://samtools.sourceforge.net/

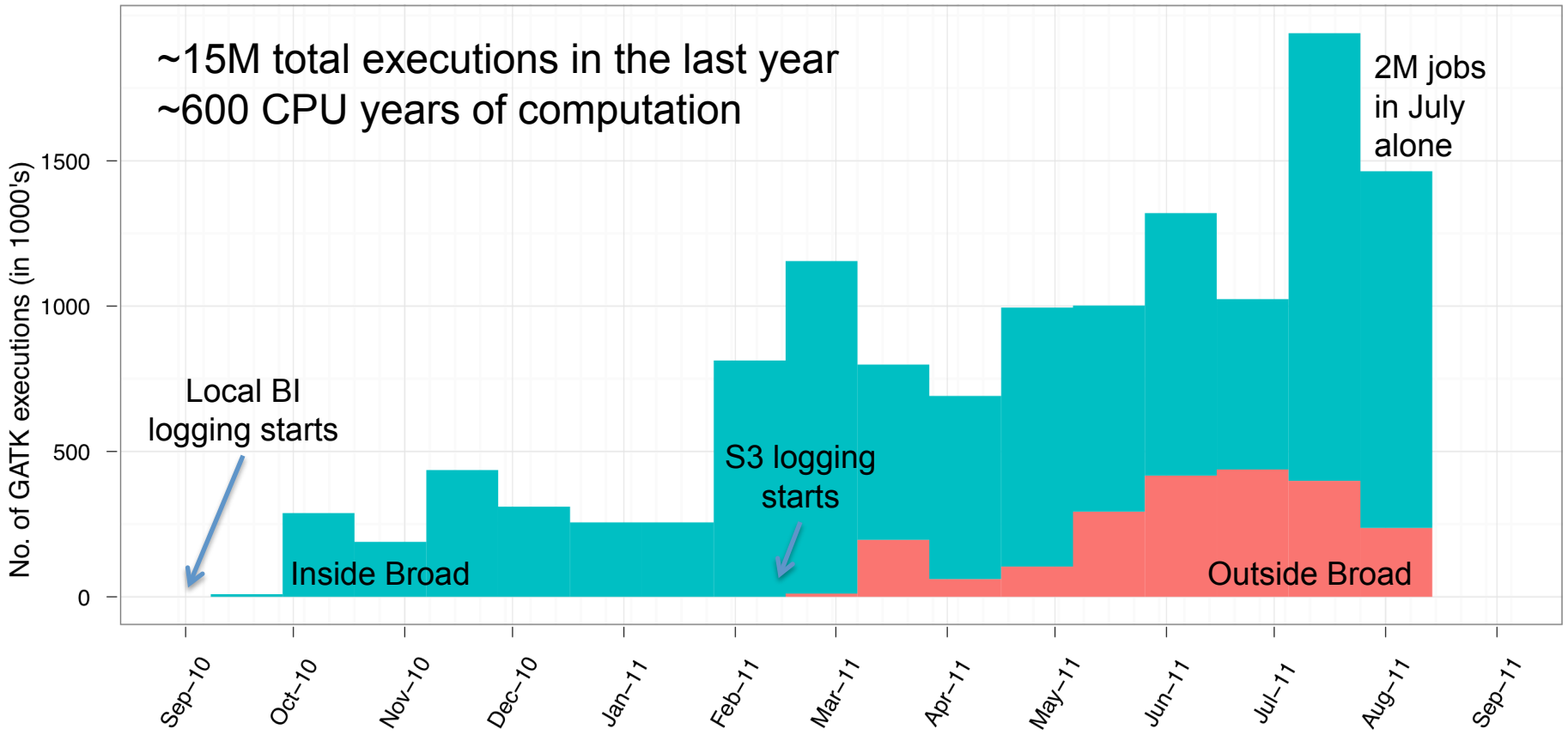
VCF format

- Standard and accessible format for storing population variation and individual genotypes

http://vcftools.sourceforge.net/

McKenna et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res.

The GATK is widely used in the community



Most popular GATK tools

MuTect	UnifiedGenotyper	SomaticMutation	CombineVariants	VariantFiltration
Indel Realigner	DepthOfCoverage	IndelGenotyperV2	Base quality score recalibration	

The 1000 Genomes Project and medical genetics projects at Broad

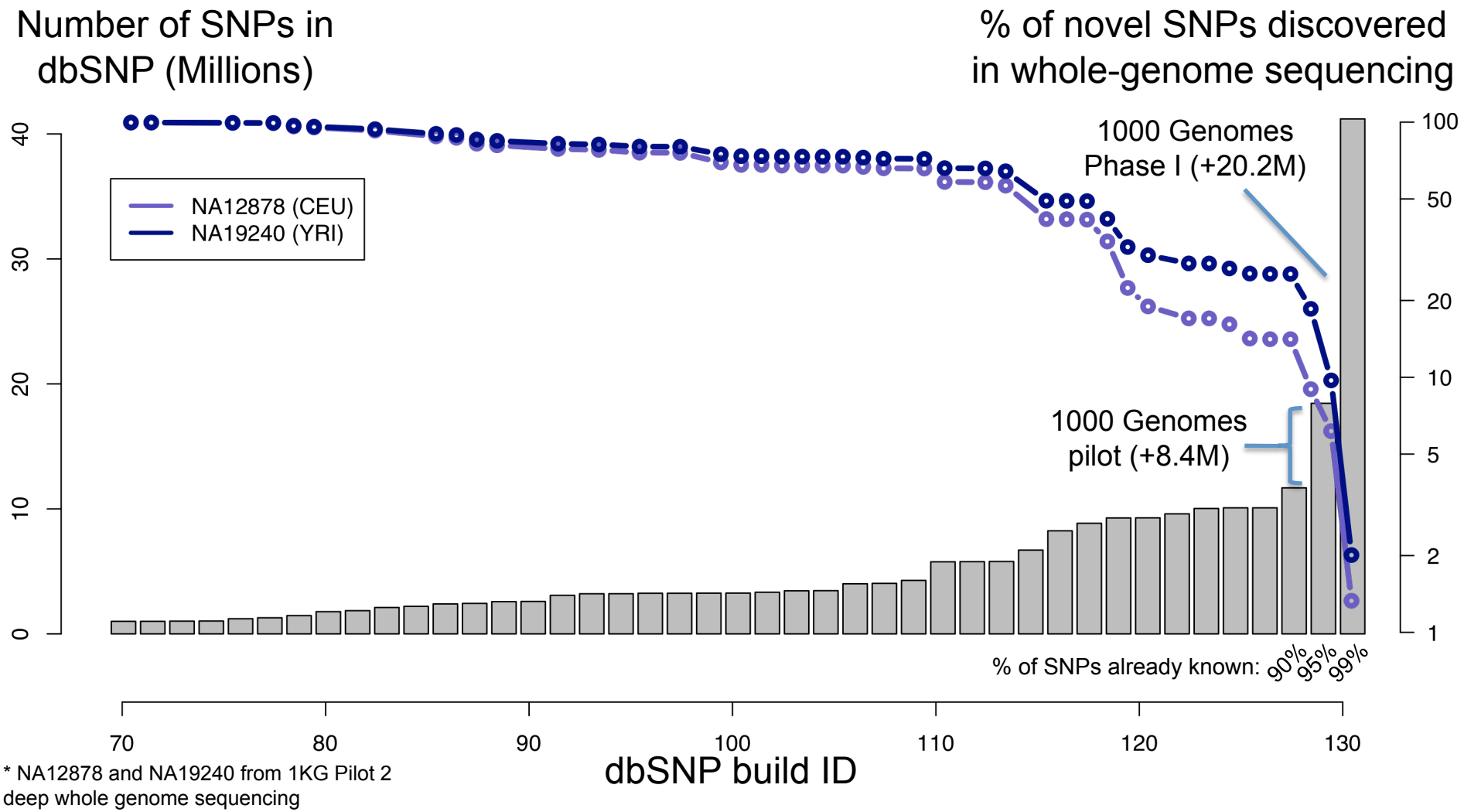
Large-scale applications of NGS sequencing

The 1000 Genomes Project is cataloging all human genetic variation with >1% MAF

- Goal: A public database of essentially all SNPs, indels, and CNVs with allele frequency >1% in each of multiple human populations
- Pioneer and evaluate methods for:
  - NGS data generation and exchange
  - Discovering and genotyping of SNPs, indels, and CNVs
  - Imputation with and from NGS data

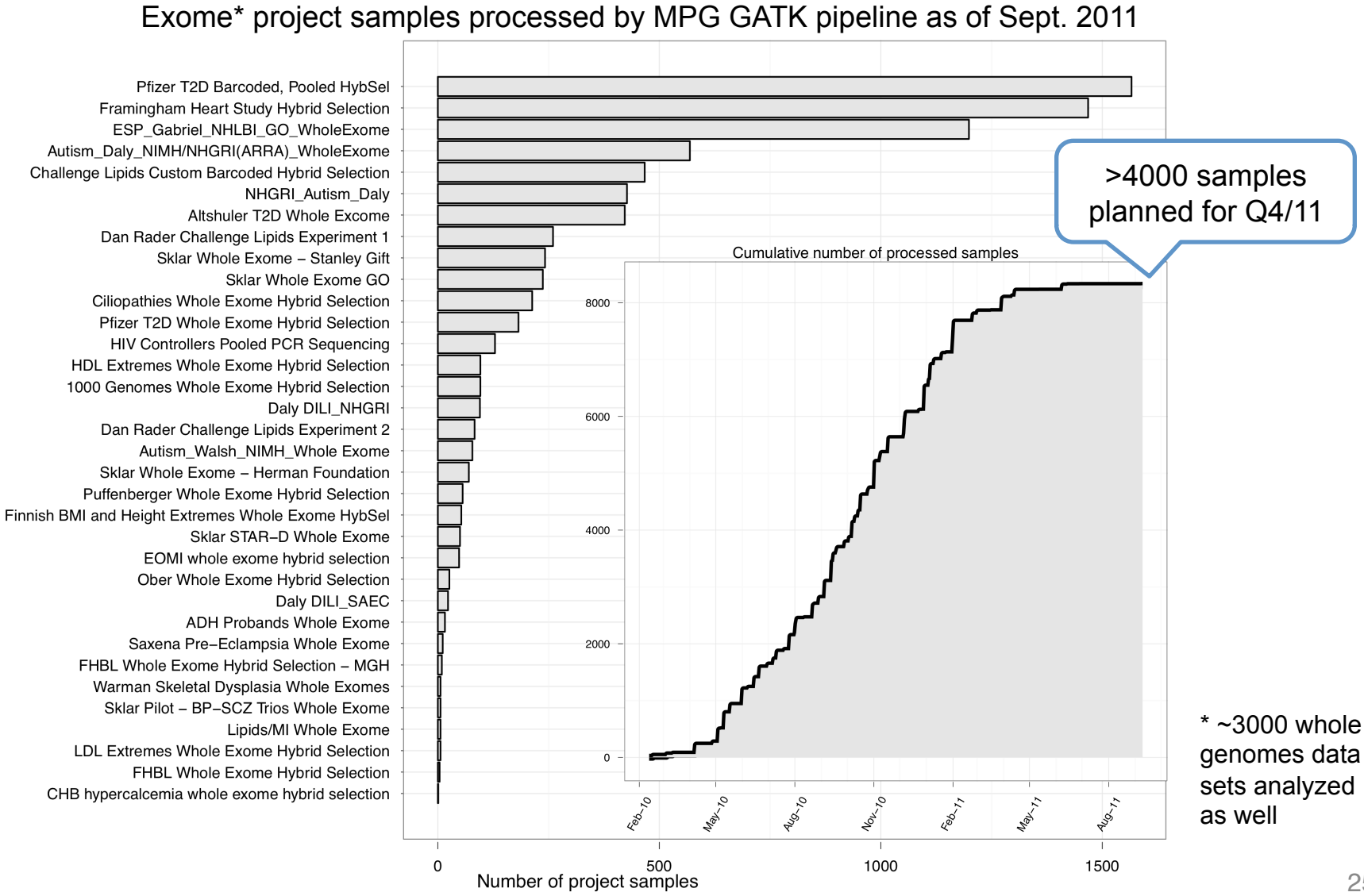
	Pilot (Nov. 2010)	Phase I (Nov. 2011)
Samples	~180	~1100
Data types	4x WGS	4x WGS, 150x WEx, 2.5M genotype chips
SNPs	15M	38M
Indels	1.5M	4M
SVs	22K	14K (genotype-able)
Completeness of catalog (% variants per sample cataloged)	95%	99%
Imputation	Imputation separate per data type (SNPs, Indels)	Integrated imputation into “best genome” per sample

1000 Genomes discovered 29M new SNPs; now ~99% of variation in each person is already known



* NA12878 and NA19240 from 1KG Pilot 2 deep whole genome sequencing

At the Broad, we analyzed over 10K medical samples in 2011, with another >20K planned for 2012



* ~3000 whole genomes data sets analyzed as well

All of this NGS data processing requires a lot of CPU power and high-performance storage

Per sample BAM data processing

The read data for each sample can be aligned, locally realigned, and recalibrated independently, for only **a few CPU days per exome**.

Multi-sample SNP + indel calling for one batch of exomes (~100 samples)

Just the multi-sample SNP and indel calling requires anywhere from **100-500 CPU days** to process.

$$\text{storage} \approx 2\frac{\text{bytes}}{\text{bp}} \times \text{targeted area}$$

Example Storage requires

Data type	Target	Storage
Per sample		
Single WEx	32 Mb	25 Gb
Deep WGS	2.85 Gb	250 Gb
Complete Project		
700 EOMI exomes	32 Mb	20 Tb
Deep Trio WGS	2.85 Gb	750 Tb

Resources required for 10,000 samples in T2D-GENES (starting Oct. 2011)

Resource	Processing	Storage
Requirement	~25,000 CPU days	250 Tb

Please note that this material is unpublished, under active development. We are sharing the details here in good faith.

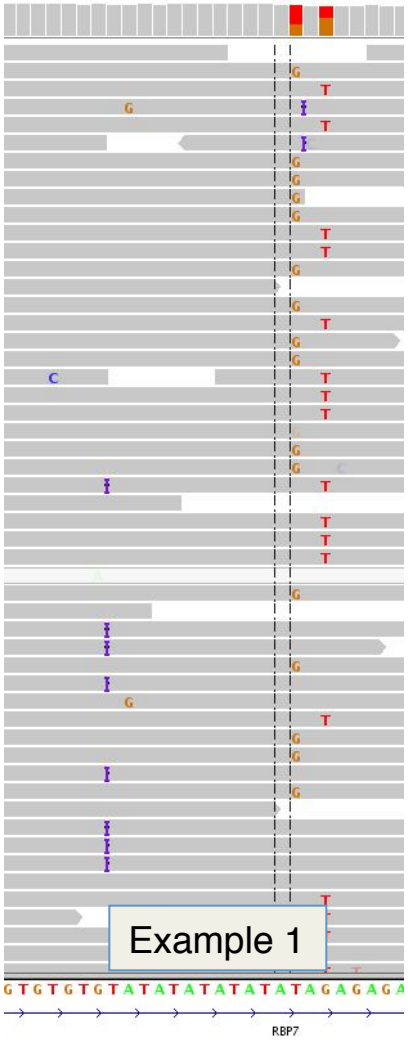
What the GSA team is working on right now

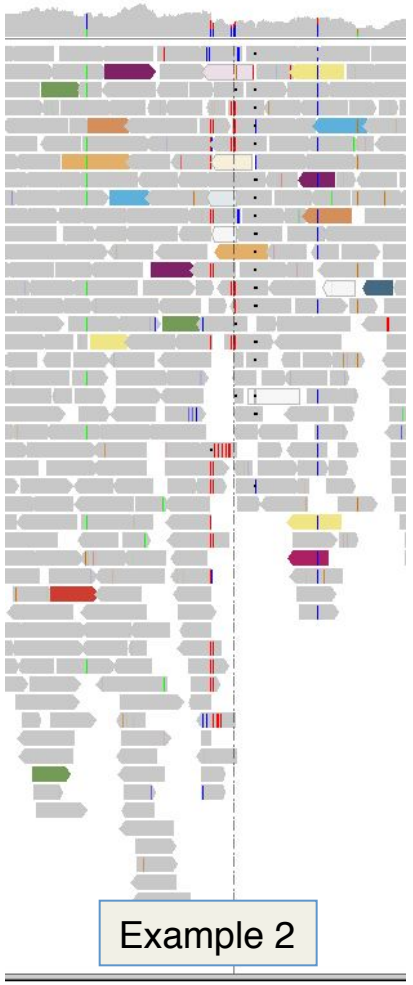
Challenges in primary data analysis

From reads to alleles: the first frontier

- Can't calculate a likelihood for a hypothesis you don't consider
- How do I know what genetic variant I'm looking at, given the read data alone?
  - A SNP, an INDEL, an SV, or something else?
- General problem, but acute for medium-sized events and insertions

Too systematic to be machine errors, but the haplotype for Pr{D|H} is unclear



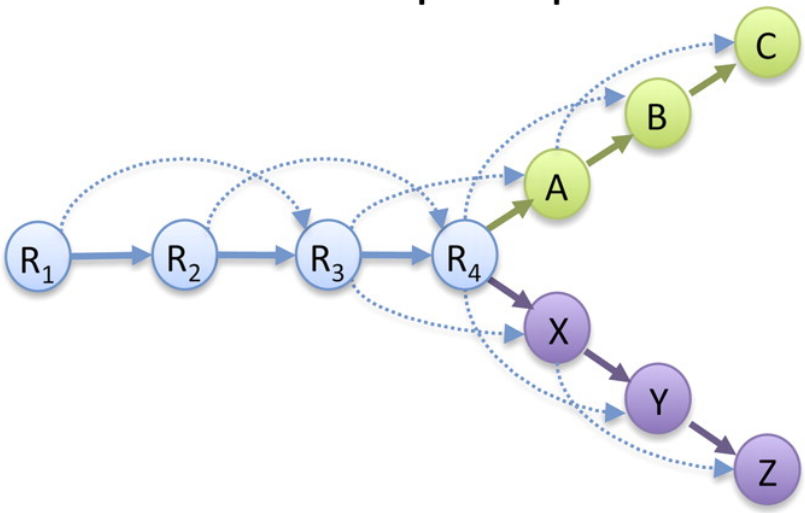


Using local de novo haplotype assembly via DeBruijn graphs

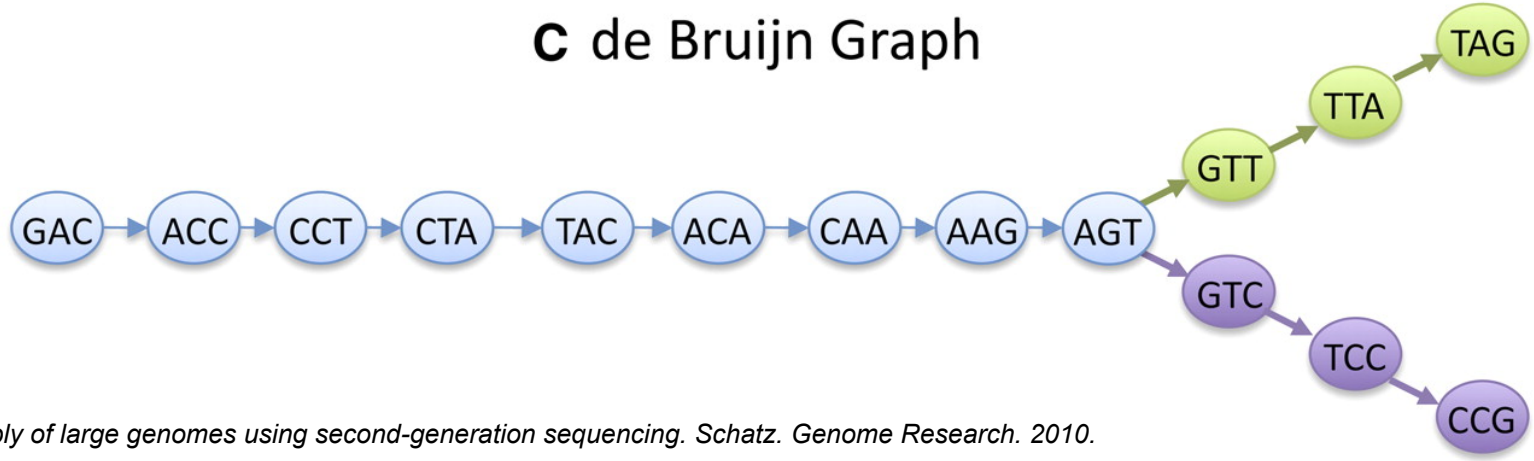
A Read Layout

```
R1: GACCTACA
R2:   ACCTACAA
R3:    CCTACAAG
R4:     CTACAAGT
A:       TACAAGTT
B:        ACAAGTTA
C:         CAAGTTAG
X:       TACAAGTC
Y:        ACAAGTCC
Z:         CAAGTCCG
```

B Overlap Graph



C de Bruijn Graph



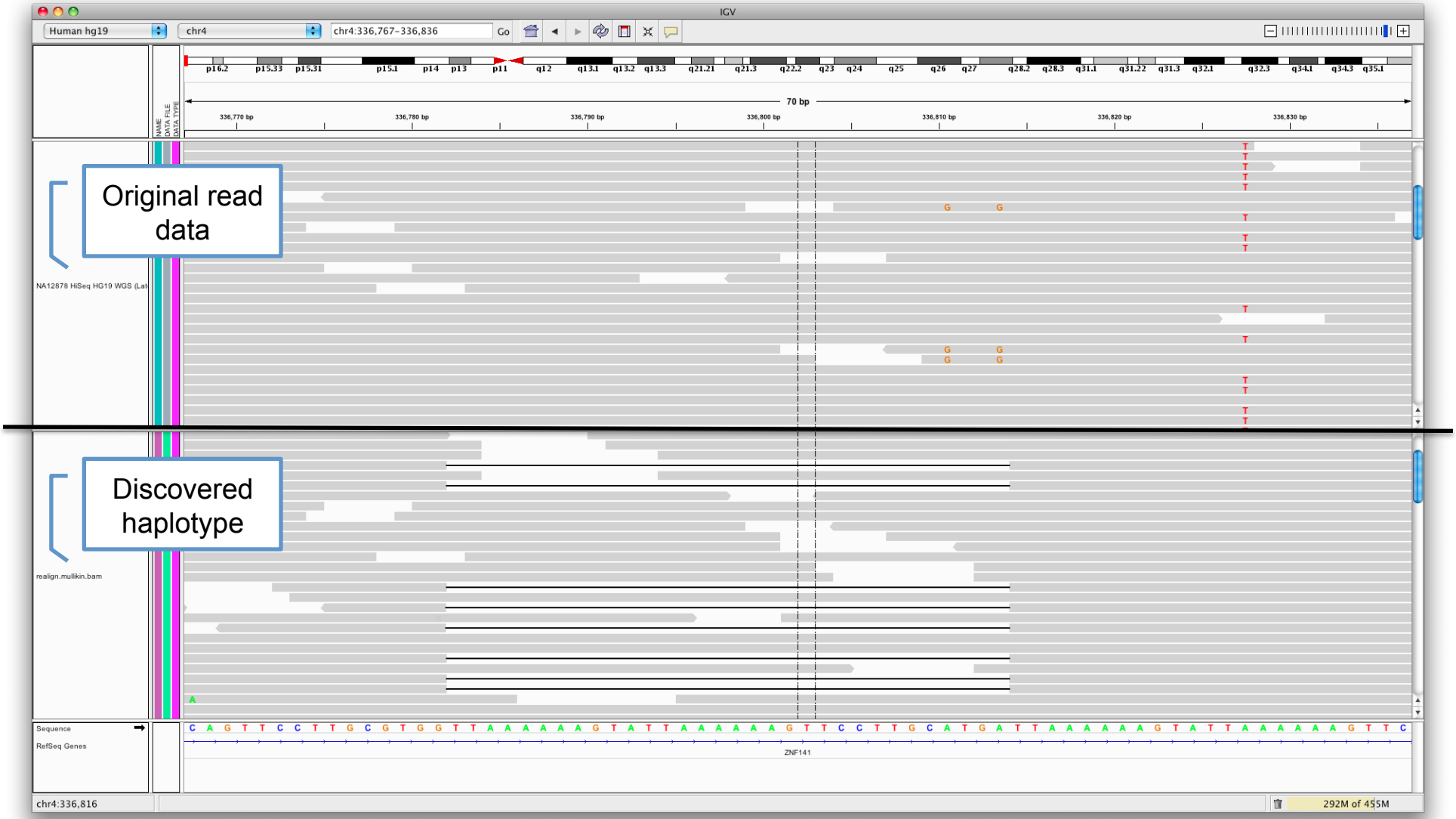
Assembly of large genomes using second-generation sequencing. Schatz. Genome Research. 2010.

Local assembly with HMM haplotype likelihoods underlies the successor to the GATK UnifiedGenotyper

	Mullikin		Mills	
Caller	**Variant Sensitivity (strict)**	**Genotype Concordance (strict)**	**Variant Sensitivity (strict)**	**Genotype Concordance (strict)**
Unified Genotyper	51.9% (40 / 77)	51.9% (40 / 77)	49.0% (97 / 198)	49.0% (97 / 198)
Haplotype Caller (Sept. 2011)	90.9% (70 / 77)	89.6% (69 / 77)	80.8% (160 / 198)	80.8% (160 / 198)

- Input data is NA12878 b37 WGS HiSeq high coverage
  - Allele discovery with assembly, likelihoods with GATK indel HMM
- Sites chosen to be very difficult (het) but high confidence in being real (require family transmission)
- Evaluation sets
  - Mullikin Fosmids and Mills et al, GR, 2011 (2x hit, double center)
  - Large events (> 15 bp), largest is 106bp (which we fail to call)

Example Mullikin het deletion we now call chr4:336781 TTAAAAAAGTATTAAAAAAGTTCCTTGCATGA/-

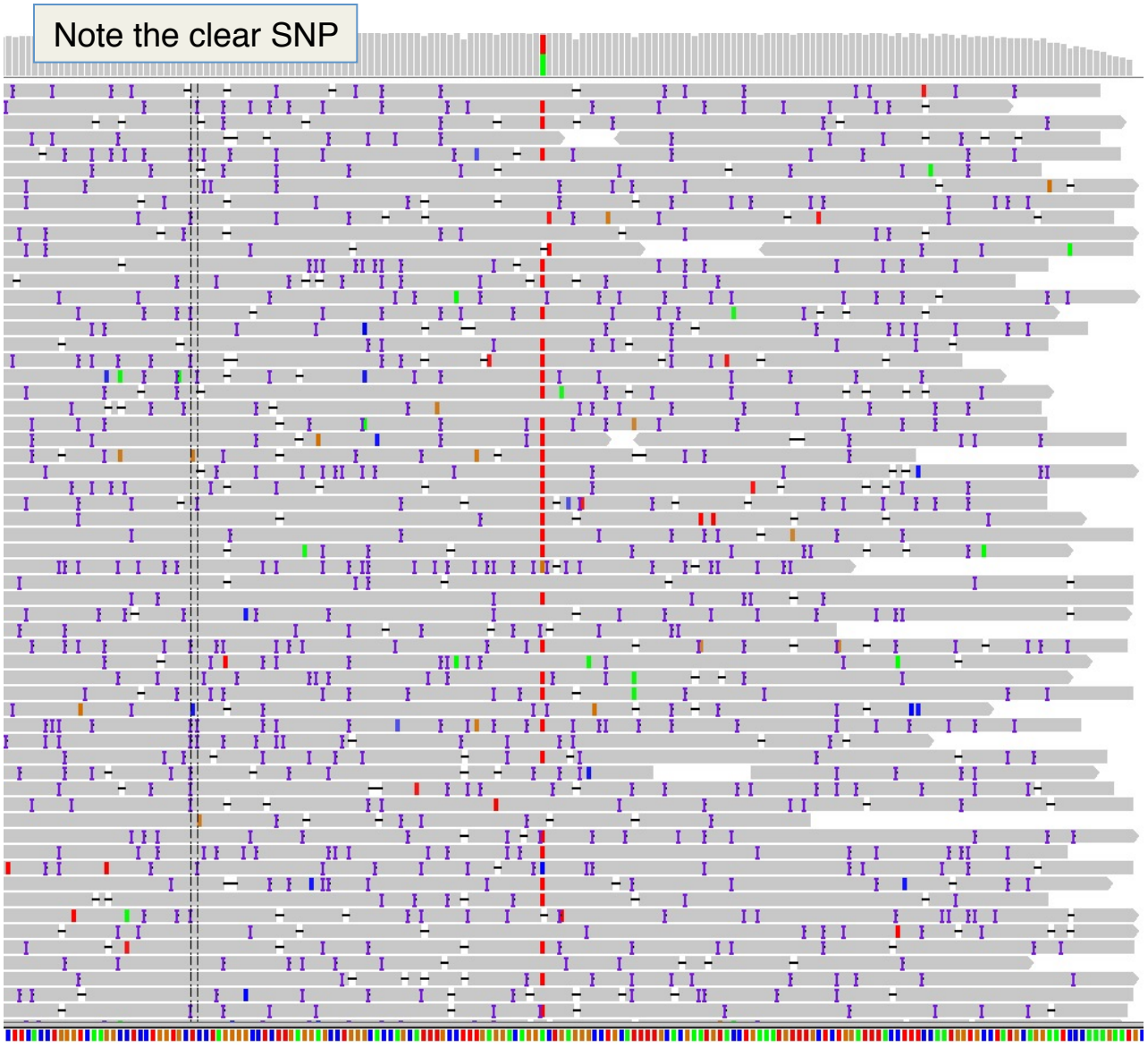


Please note that this material is unpublished, under active development. We are sharing the details here in good faith.

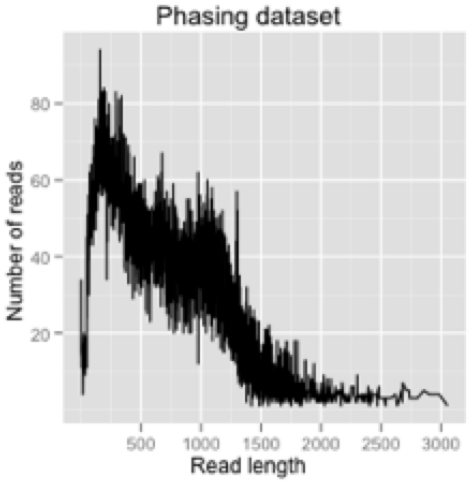
Very long, very indel rich reads

PACIFIC BIOSCIENCES

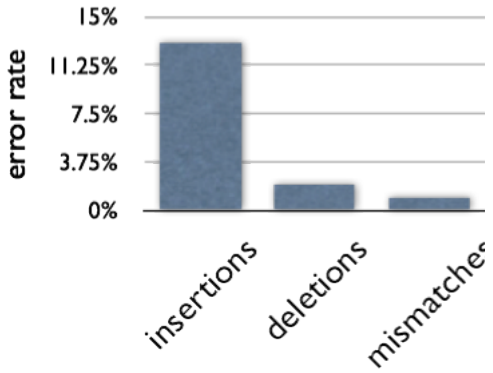
Pacific Bioscience RS produces long single molecule reads full of insertion errors



Median read lengths of 3000 bp today

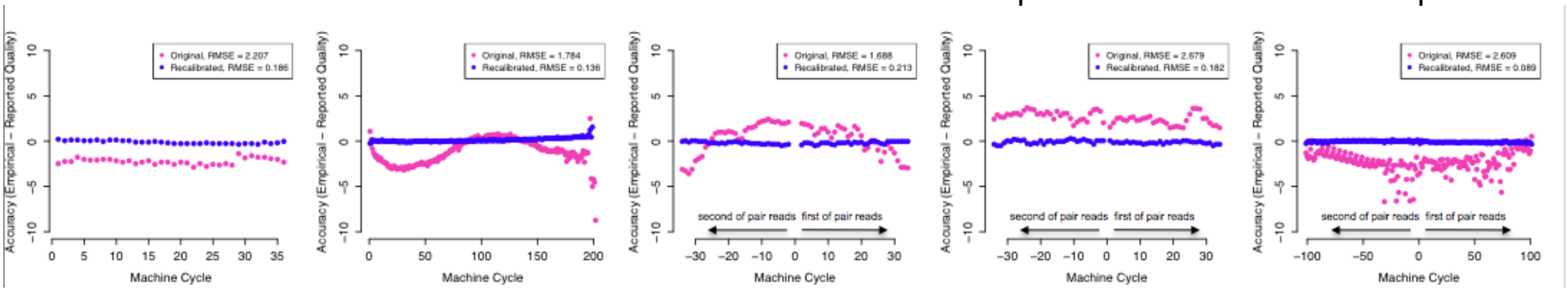


Indels are the main error mode (see purple markers)



PacBio substitution errors are independent of machine cycle

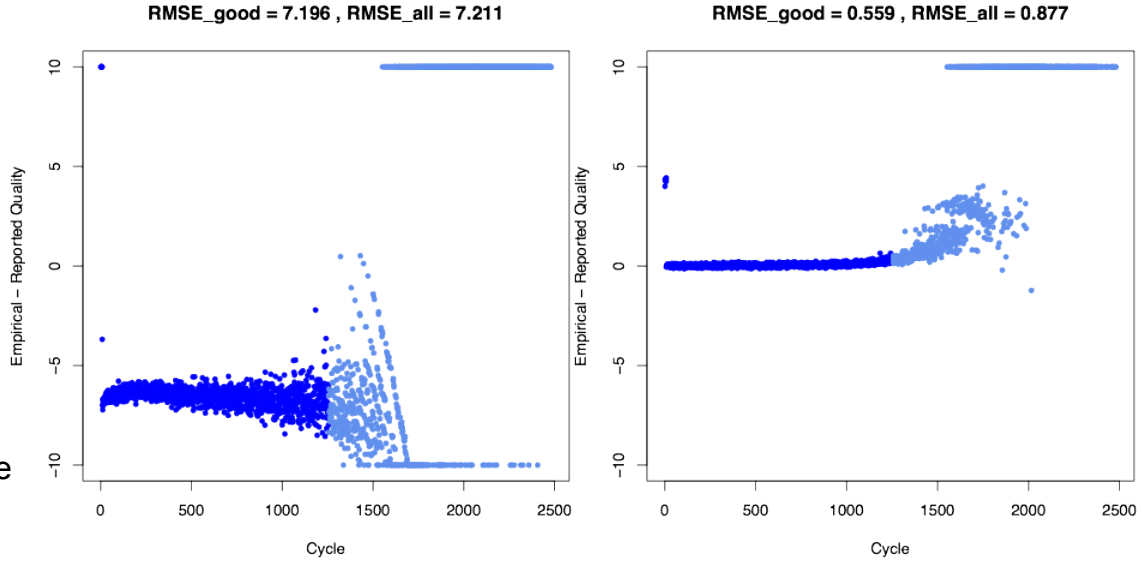
SLX GA | 454 | SOLiD | Complete Genomics | HiSeq



PacBio

before recalibration:

- even before recalibration PacBio reads do not seem to be affected by the length of the read like other technologies.
- The steady straight line breaks after 1250bp because we have very few reads that go that long (hence the light blue colored dots)



after recalibration:

- recalibration helps make the straight line more dense and clear.
- the lack of data points still breaks the recalibrated line after 1250bp.

phasing dataset

Pacific Biosciences support sensitive and specific SNP calling with the GATK

SNP call comparison at region surround TP and FP de novo mutation calls from Conrad et al.*

	Confirmed *de novo* SNP	Confirmed artifact
Pacbio ALT	48	5
Pacbio REF	0	67
HiSeq ALT	48	35
HiSeq REF	0	37

	PacBio	HiSeq
Sensitivity	100%	100%
Specificity	93%	51%
PPV	91%	58%
NPV	100%	100%

SNP calls in 117 Kbp regions surrounding high-quality SNP sites from in DePristo et al.**

	Unfiltered SNP calls	HapMap calls	TiTv known	TiTv novel
Pacbio	531	38	3.22	1.96
HiSeq	547	40	3.00	1.93

** Somewhat biased against HiSeq, as these false positive calls were made on Illumina GA instruments in 1000G pilot*

***Missed called due to alignment reference bias; can be corrected with haplotype-based caller*

Please note that this material is unpublished, under active development. We are sharing the details here in good faith.

A intermediate format for efficient data processing

REDUCED REPRESENTATION READS

Why do the file (BAM) sizes matter?

Resources required for T2D GENES

Resource	Requirement
Processing	100 * 100 batches = Minimum 10,000 CPU days
Storage	25 Gb * 10,000 = 250 Tb

BAM files are archival and very large

- Transferring 250 Tb of data is a time consuming and errorful process
- Processing the data is resource intensive for multi-sample calling: large I/O burden moving the data and calling requires much memory
  - Because of this, we are limited to batches of ~100 samples even when samples are homogenous (e.g. same cohort)
  - There are technical incompatibilities among the disparate batches (e.g. filters) that are not trivial to address

Reduced Reads is a new GATK capability being pioneered now

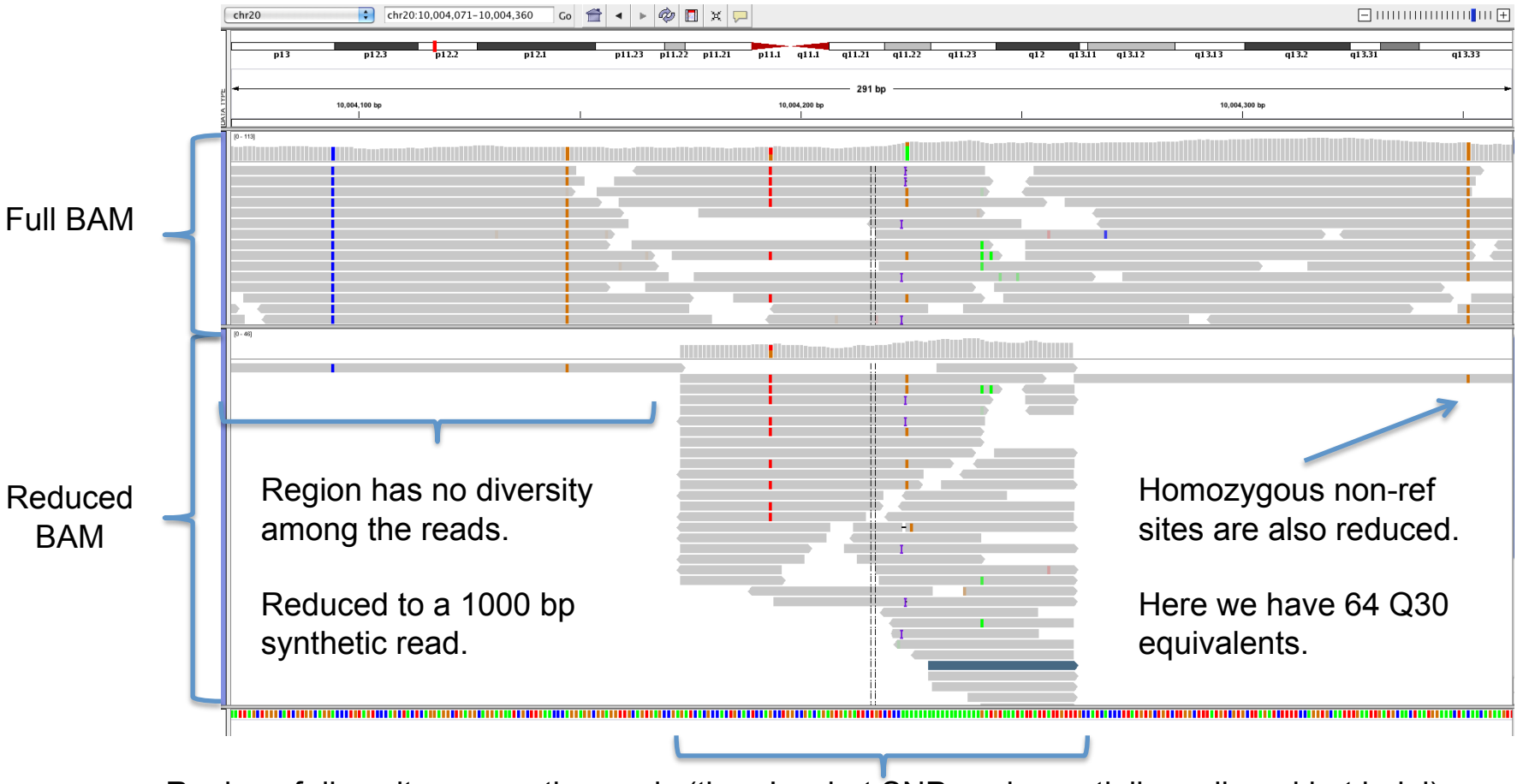
- Produces an analysis-compatible “ReducedBAM” file that
  - Is 100x smaller than a complete exome BAM
  - Does not impact SNP/indel calling sensitivity or specificity
  - Can be used to assess coverage and “callability”

- Many benefits to ReducedBAMs
  - SNP/indel calling time reduced by many orders of magntitude
  - Enables the simultaneous calling of >10,000 samples
  - Vastly easier data transfer for project files

- Reduced BAMs are not suitable for primary data archiving
  - NCBI and EBI are working on compressed archival BAMs

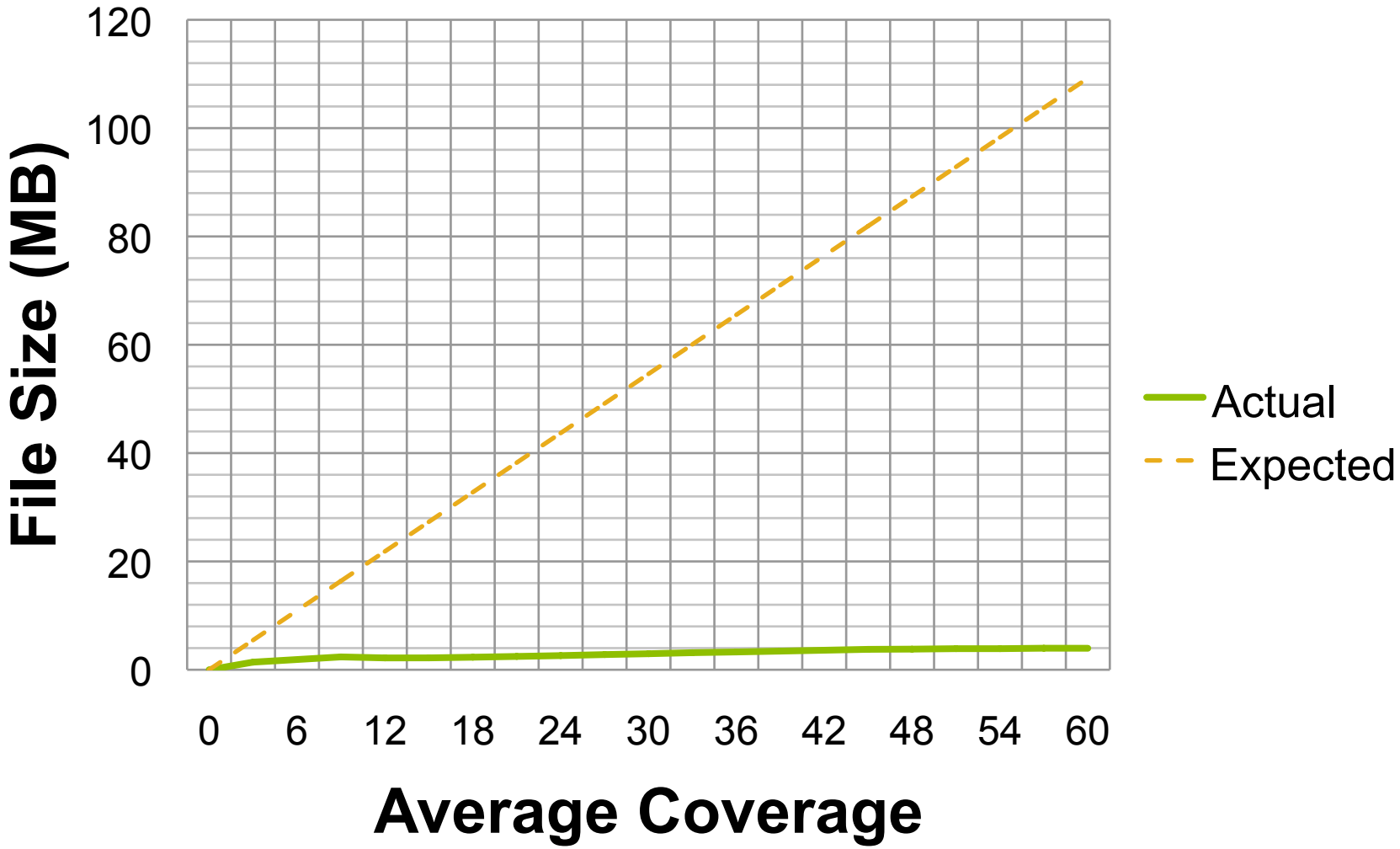
Example WGS: zoomed out

Detail of reduced reads around a variable region



Region of diversity among the reads (there's a het SNP, and a partially realigned het indel) results in a variable region. Reads spanning the region (sites of diversity +/- 40 bp) are clipped to the region, and emitted. Reads are downsampled to 50x avg. coverage across the region.

Only a marginal increase in file size with coverage



Preliminary results: runtime and file sizes

Simultaneous SNP and indel calling

	Single exome	63 CEU exomes	822 1000G exomes
File size			
Full BAM(s)	11 Gb	1.2 Tb	11.2 Tb
Reduced BAM(s)	114 Mb	12 Gb	133 Gb
Relative improvement (x)	100x	100x	85x
Calling runtime			
Full BAM(s)	36 min	48 hours	140 days
Reduced BAM(s)	9 min	6 hours	8 days
Relative improvement (x)	5x	8x	18x

Conclusions

- Data processing
  - Developed experimental and analytical methods enabling today's medical genetics projects
  - Methodological advances continue at rapid pace
  - All available in open-source GATK project

- The 1000 Genomes Project:
  - The pilot project completed with excellent results
  - More than doubled the number of cataloged SNPs, indels, and CNVs
  - Phase I release scheduled for May 2011
    - On track to find ~37M SNPs, up from ~15M in pilot

Challenges in NGS and medical genetics I

What would I work on if I were just starting out in this field?

- Most errors in NGS data processing are now due to mismapping and misalignment.
  - What experimental approaches and algorithms will fix this?
- Collectively we've sequenced more than 25K samples this last year, and likely 100K in 2012.
  - How we can make the most of the available information?
  - How can I do joint analyze of 100K samples?
- What does it mean to have a complete genome?
  - Can we "account for all reads"?
- How much are our resequencing results influenced by the quality of the reference genome?
  - Are a lot of our point mutations actually larger events?

Challenges in NGS and medical genetics II

- Suppose I have a 50 putative disease associations today. How can I replicate and extend this results in 100x more samples quickly and with reasonable cost?

- Sample size is king (at least in complex genetics)
  - How can I combine NGS variation with 100x larger GWAS data sets, particular for low-frequency variation?
  - What experimental designs are most empowered to identify disease-associated variation?

- Soon all novel mutations will be private to families
  - How can we analyze variation we only ever see once?
  - What are appropriate experimental designs given this?

- Biology-free statistically association is very robust and can identify regions of interest
  - Figuring out causal variation is not at all obvious, though. Can we do this at scale?

Help develop and apply methods in NGS to medical genetics projects

- The Genome Sequencing and Analysis group in Medical and Population Genetics at the Broad Institute is hiring

Computational Biologist	Ph.D.-level research scientist focused on project analysis and methods development in medical genetics
Senior Software Engineer	B.A./M.A./Ph.D in computer science with 5+ years of experience to lead MPG software development projects
Test and Documentation Engineer	B.A. in computer science or equiv. to lead MPG software QC and manage user interactions

Talk to me for more information or email depristo@broadinstitute.org