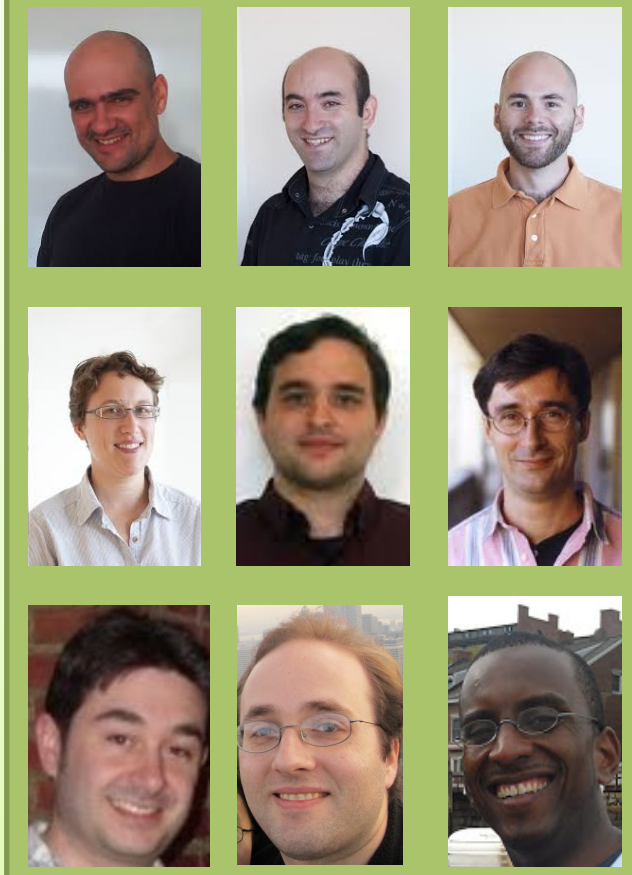
Best Practices for DNA Sequencing Data Processing and Analysis

NGS Data Analysis
October 16th, 2013

Mauricio Carneiro
Group Leader, Computational Technology Development
Broad Institute of MIT and Harvard

Acknowledgments

Genome sequencing and analysis team



Mark DePristo

Eric Banks

Stacey Gabriel

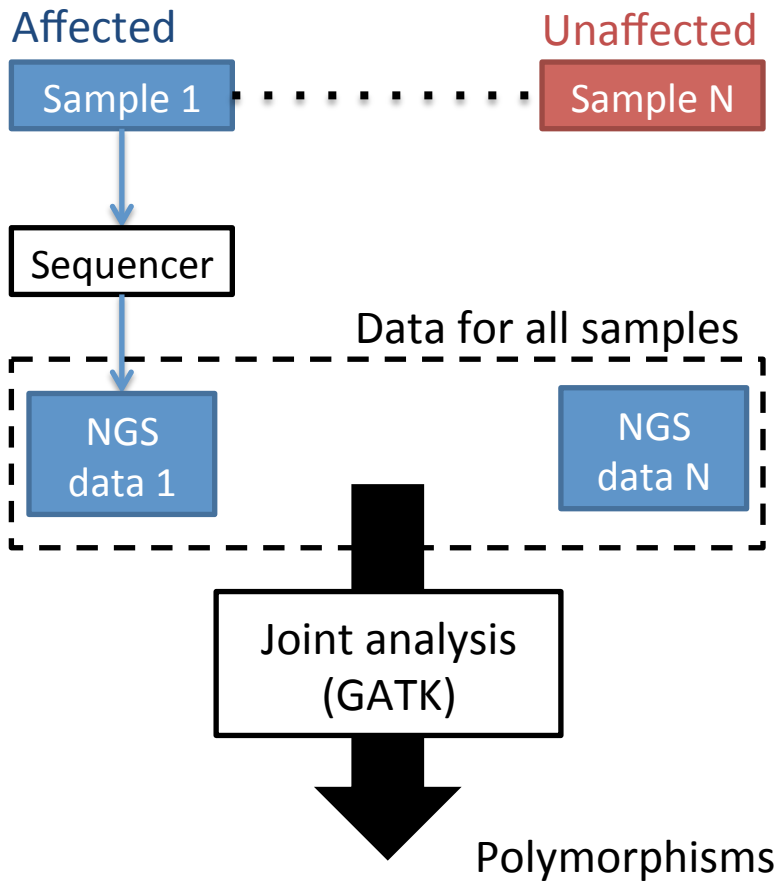
David Altshuler

There are three parts to this talk

- **Study design**
  - An overview on the different applications that make use of sequencing data and how to better design for them.
- **Analysis pipeline**
  - The full NGS data analysis pipeline, from DNA to analysis-ready variants.
- **Quality control**
  - How do I know whether my project has reliable data?

How to discover loci involved in disease in three easy steps

1. Get thousands of affected and tens of thousands of unaffected samples
2. Sequence samples with next-generation sequencer
   - Discover polymorphisms (SNPs, indels, etc) across samples
   - Determine genotype of every sample at each variant site
3. Look for systematic differences in genotypes for affected samples vs. unaffected across all sites



	SNP	Indel
Affected	10/1000	10/1000
Unaffected	25/2000	1/2000
Association strength (P)	0.72	0.0001

To fully understand **one** genome we need **tens of thousands** of genomes

Clinical genomics

Single diseased individual

vs

Compared to many control samples

Find variants consistent with disease model (dominant, recessive), often conditioning on these variants being absent / rare among the controls

Research genomics

Many affected individuals

vs

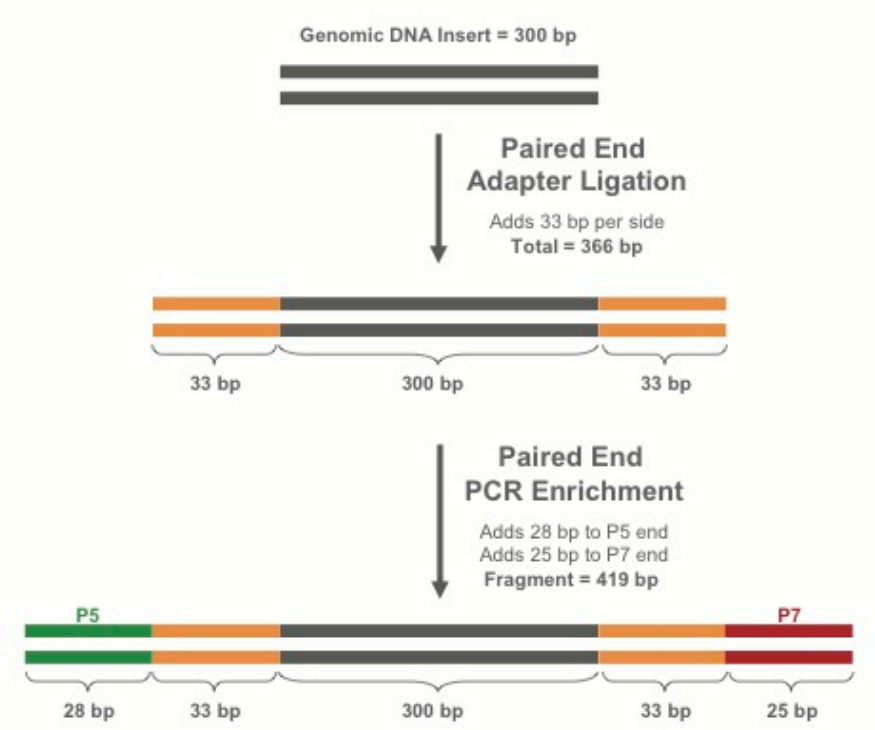
Compared to many control samples

Find variants enriched / depleted in affected individuals, relative to controls

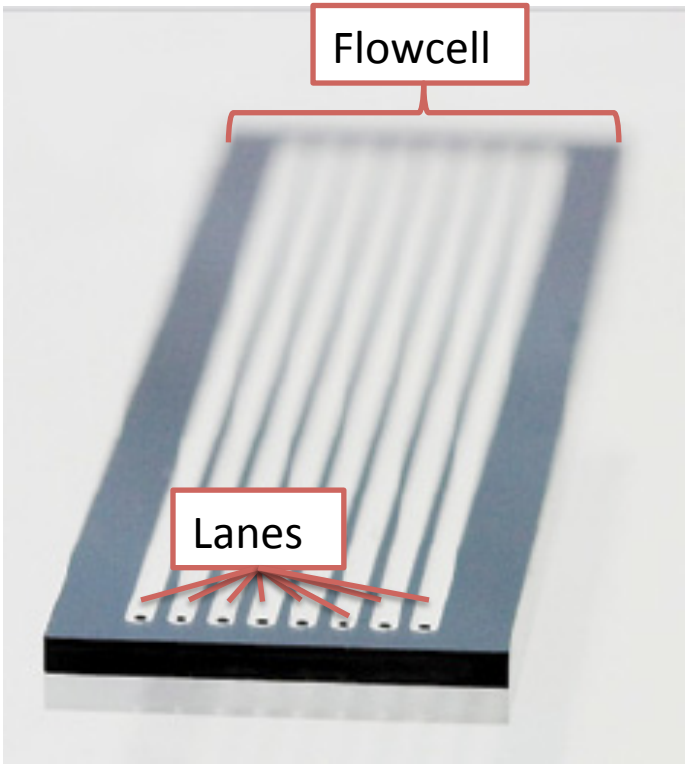
All individuals must be consistently called, so that the differences between groups are due to real genetic differences, not technical ones

NGS TECHNOLOGIES AND TERMINOLOGY

Libraries, lanes, and flowcells

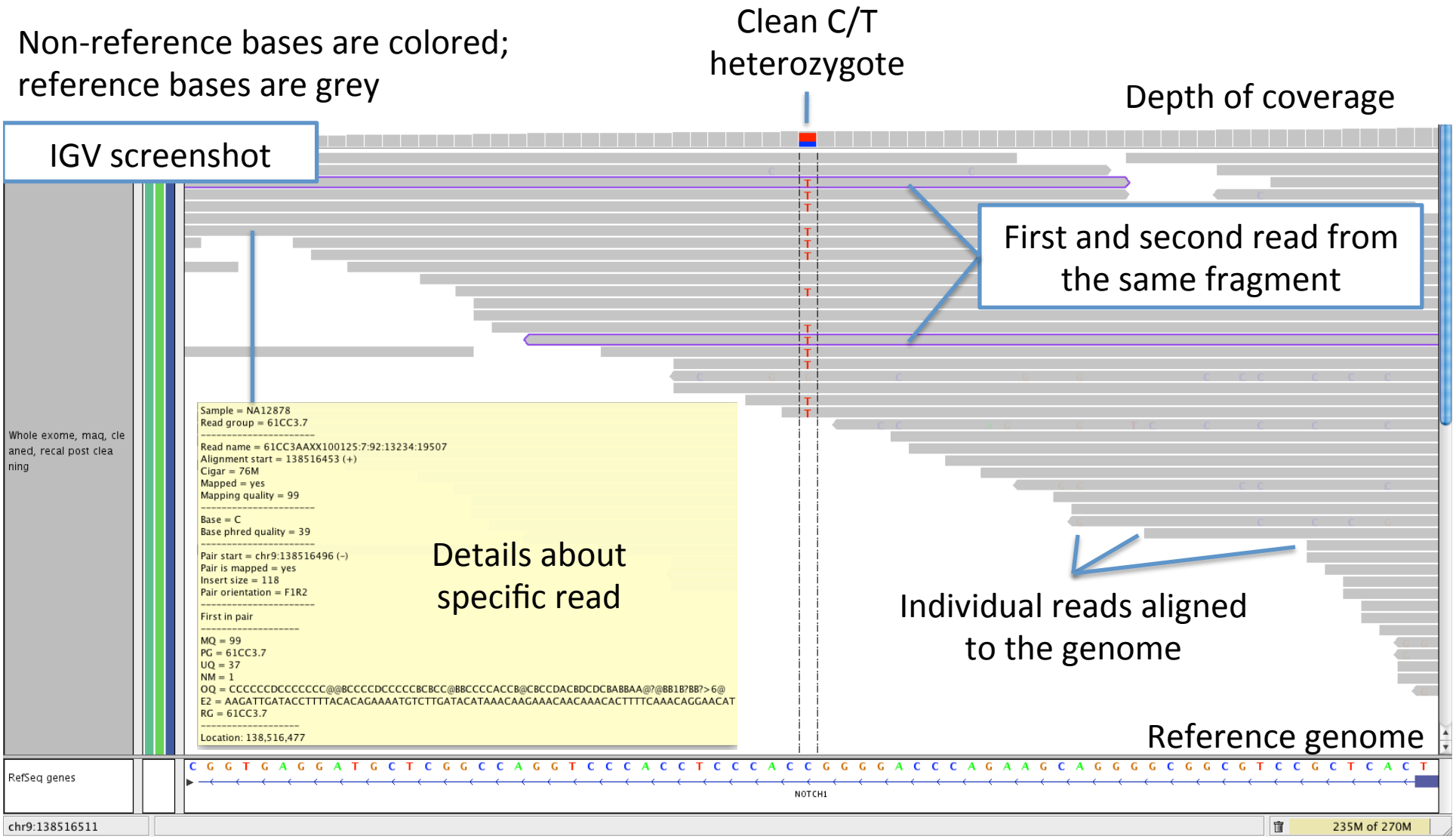


Each reaction produces a unique **library** of DNA fragments for sequencing.



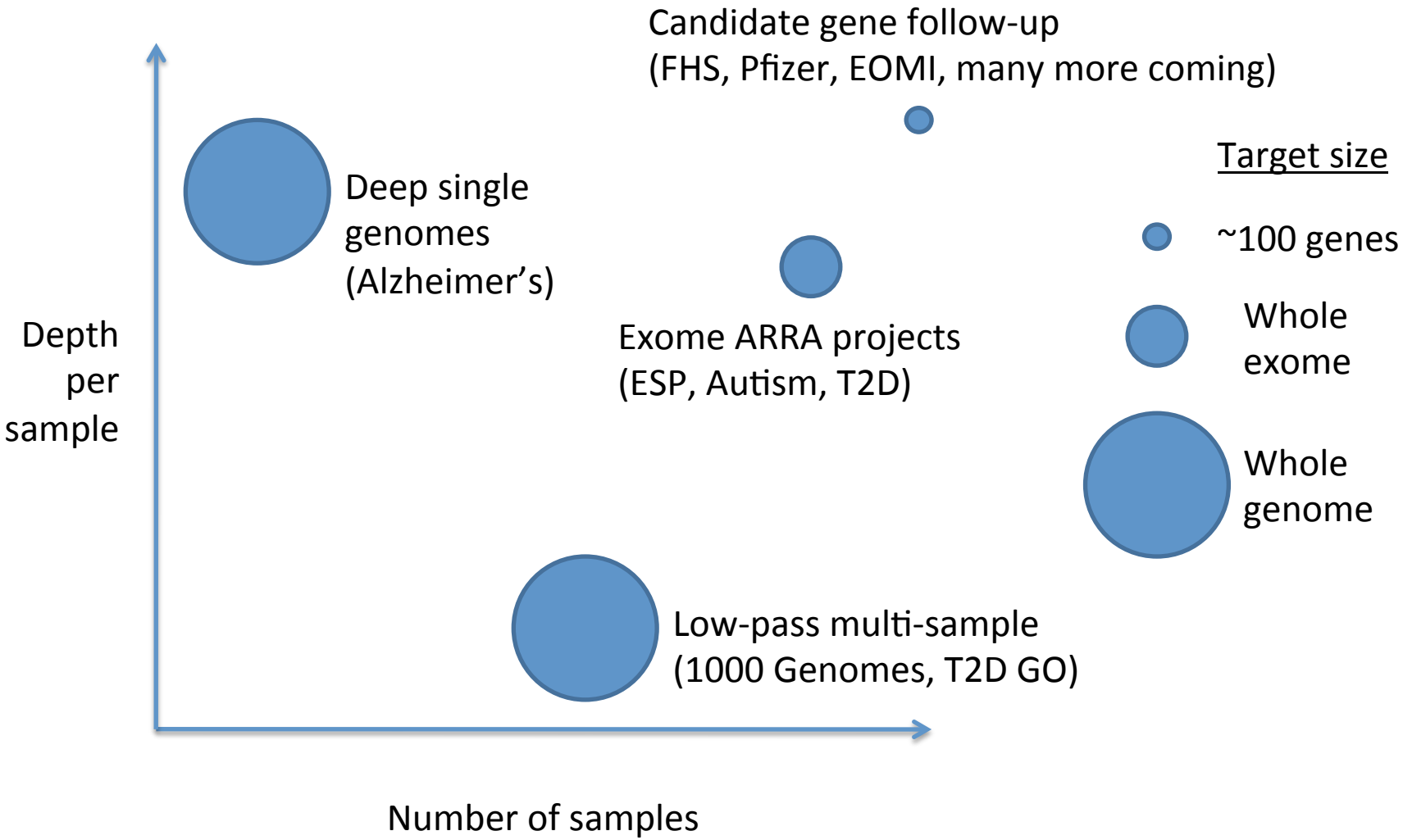
Each NGS machine processes a single **flowcell** containing several independent **lanes** during a single sequencing run

Reads and fragments

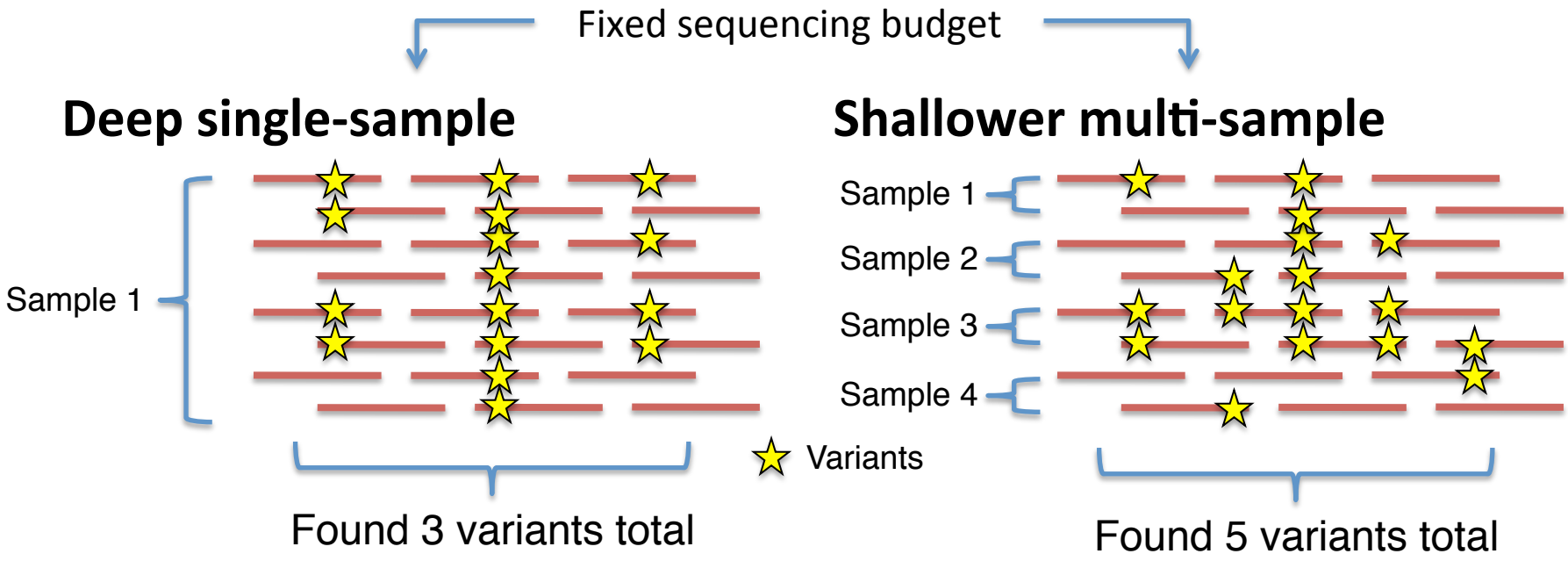


EXPERIMENTAL DESIGNS

Projects use a variety of experimental designs, mostly selected to maximize disease sample size



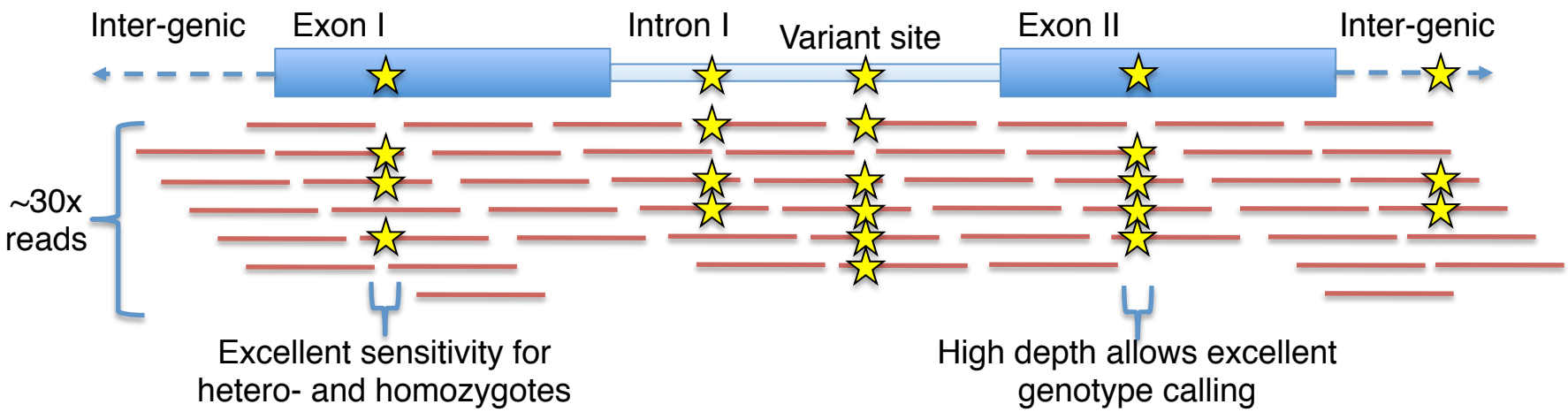
Single vs. multi-sample analysis



- Higher sensitivity for variants in the sample
- More accurate genotyping per sample
- Cost: no information about other samples

- Sensitivity dependent on frequency of variation
- Worse genotyping
- More total variants discovered

High-pass sequencing design



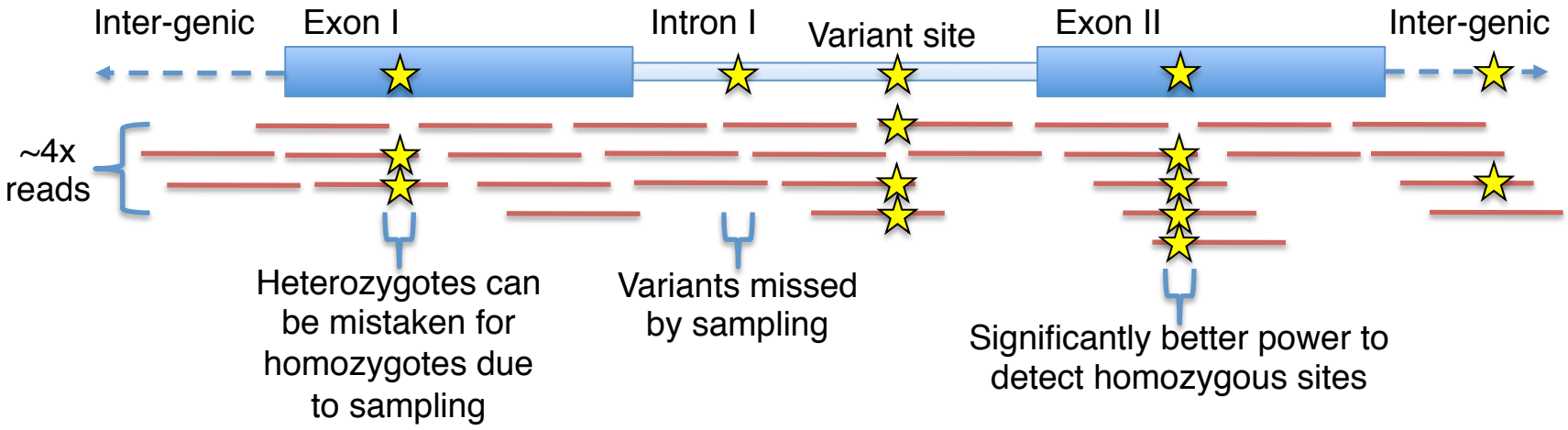
Data requirements per sample

Targeted bases	~3 Gb
Coverage	Avg. 30x
# sequenced bases	100 Gb
# lanes of HiSeq	~8 lanes

Variant detection among multiple samples

Variants found per sample	~3-5M
Percent of variation in genome	>99%
Pr{singleton discovery}	>99%
Pr{common allele discovery}	>99%

Low-pass sequencing design



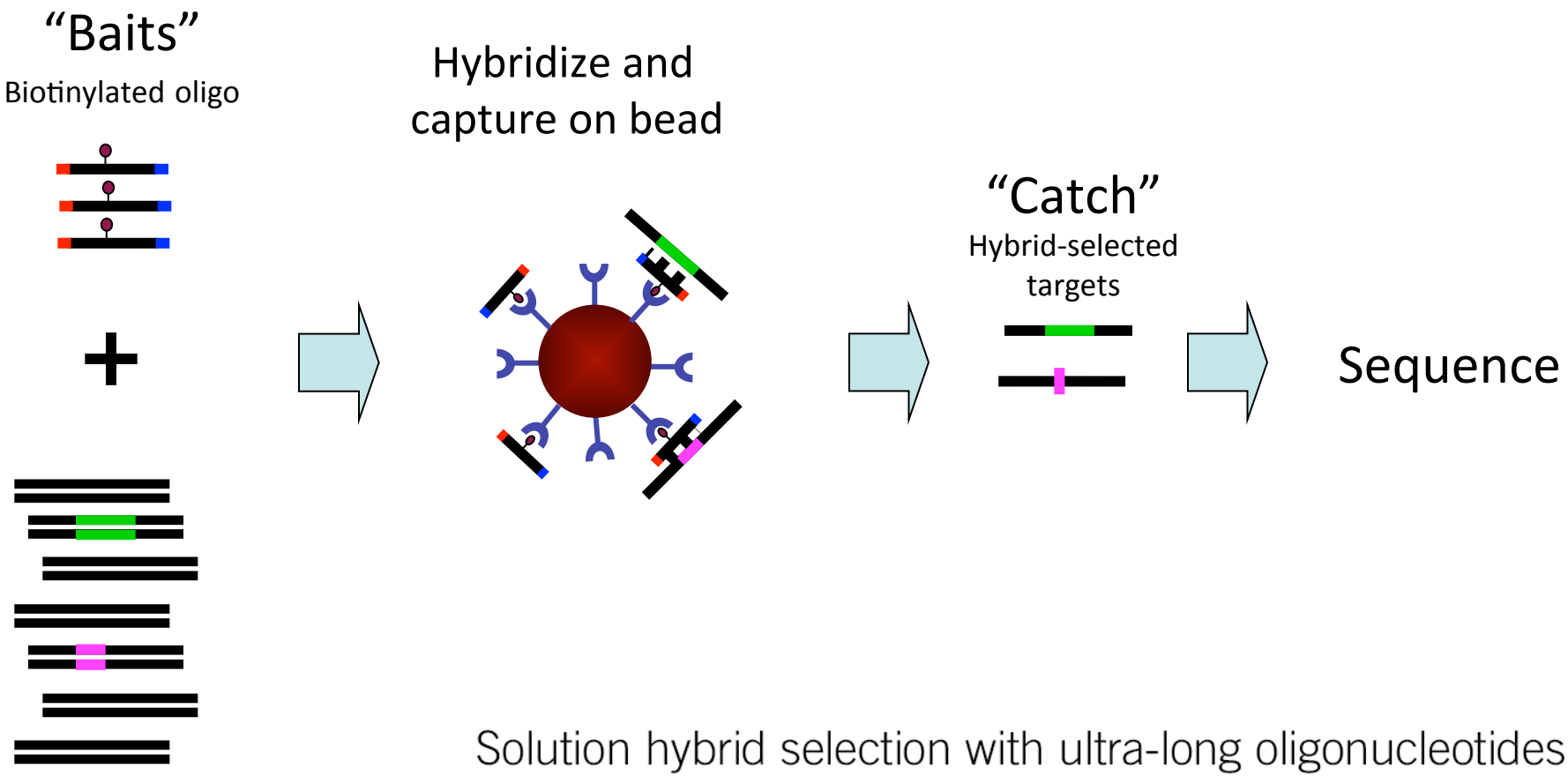
Data requirements per sample

Targeted bases	~3 Gb
Coverage	Avg. 4x
# sequenced bases	20 Gb
# lanes of HiSeq	~1.25

Variant detection among multiple samples

Variants found per sample	~3M
Percent of variation in genome	~90%
Pr{singleton discovery}	<50%
Pr{common allele discovery}	~99%

Hybrid Selection: an Approach to "Fish" for Exons



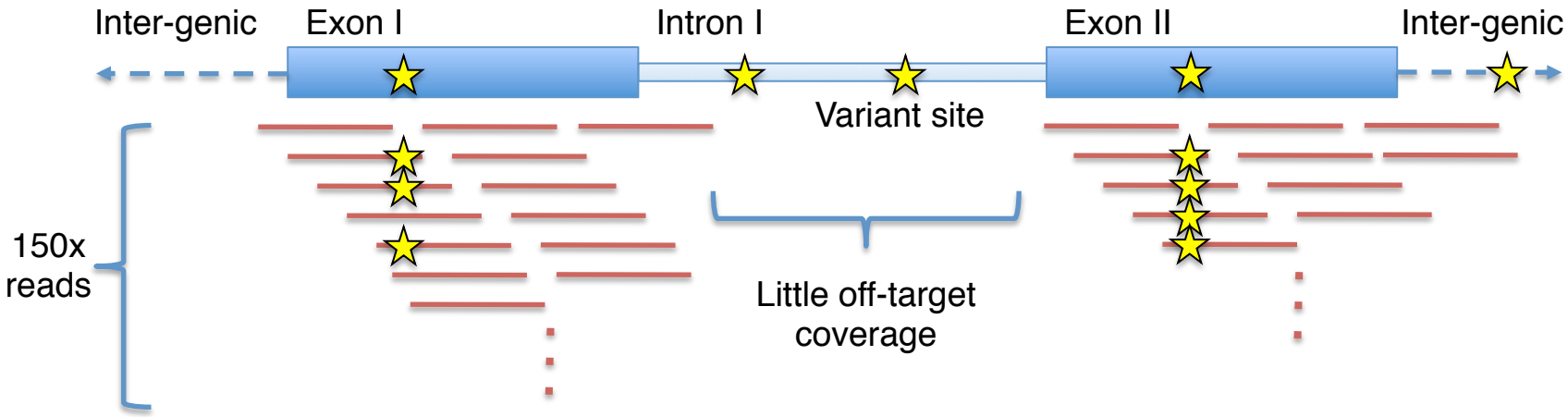
Stacey Gabriel

Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing

Andreas Gnirke[1], Alexandre Melnikov[1], Jared Maguire[1], Peter Rogov[1], Emily M LeProust[2], William Brockman[1,5], Timothy Fennell[1], Georgia Giannoukos[1], Sheila Fisher[1], Carsten Russ[1], Stacey Gabriel[1], David B Jaffe[1], Eric S Lander[1,]

VOLUME 27 NUMBER 2 FEBRUARY 2009 NATURE BIOTECHNOLOGY

Exome capture sequencing design



Data requirements per sample

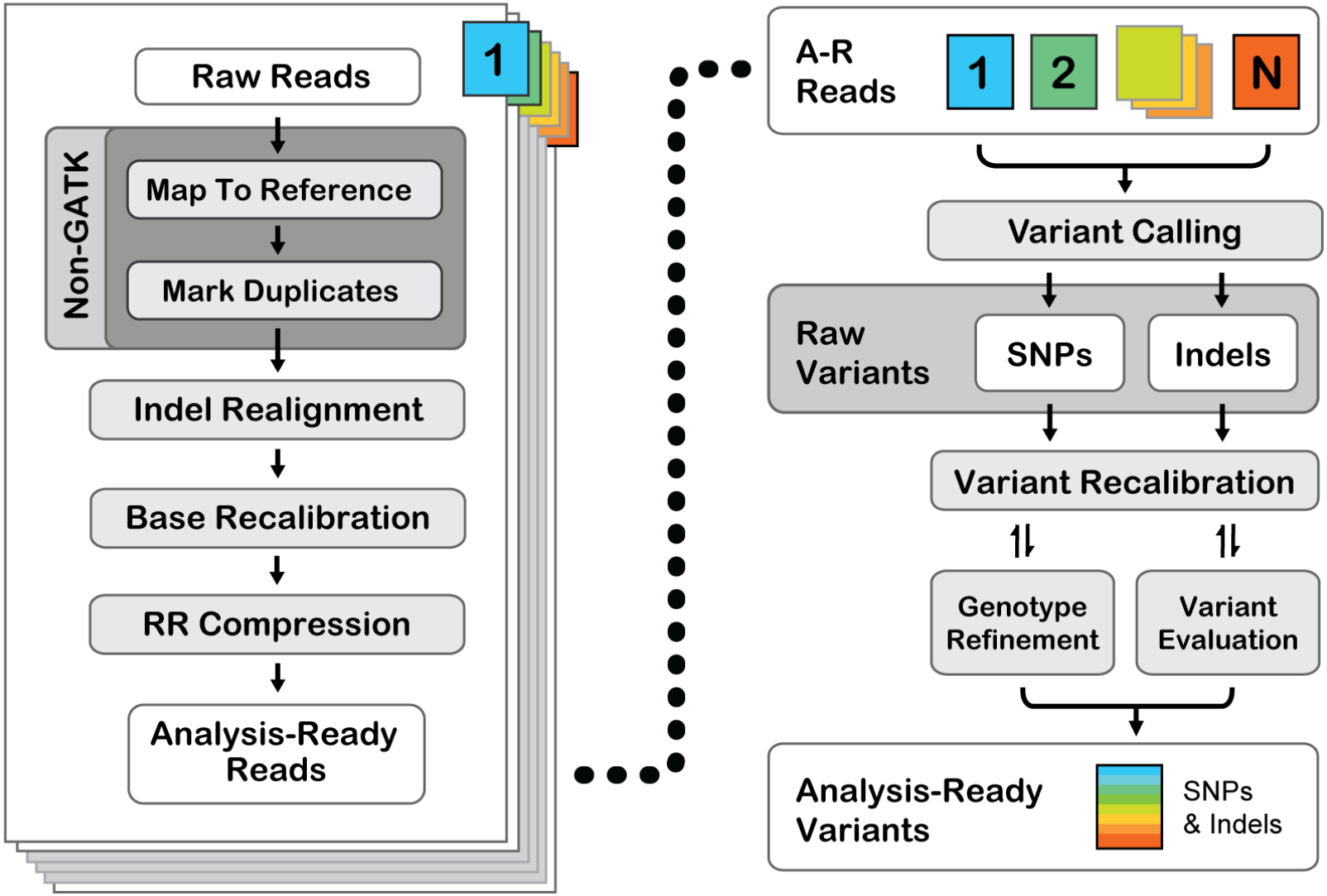
Targeted bases	~32Mb
Coverage	>80% 20x
# sequenced bases	5 Gb
# lanes of HiSeq	~0.33

Variant detection among multiple samples

Variants found per sample	~20K
Percent of variation in genome	0.5%
Pr{singleton discovery}	~95%
Pr{common allele discovery}	~95%

Detecting variants in next-generation sequencing data

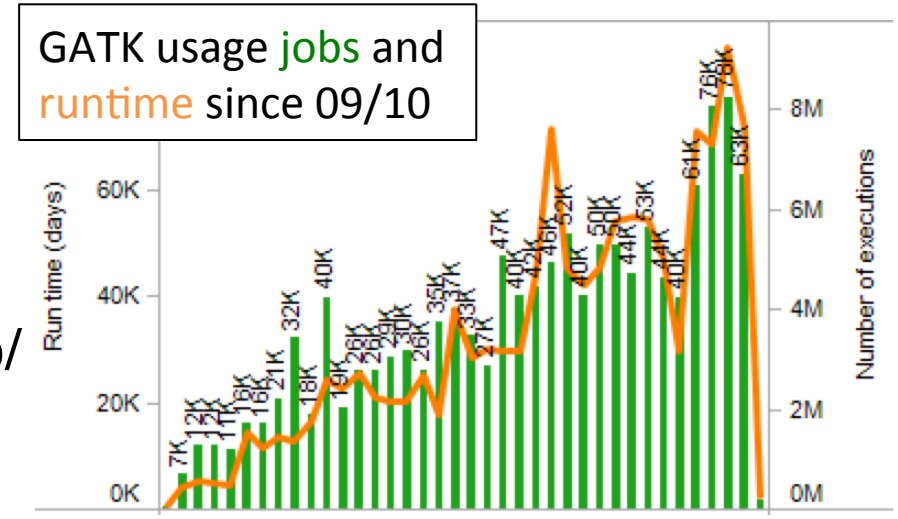
The framework for variation discovery



GATK (Genome Analysis Toolkit)

The GATK is:

- A suite of powerful NGS analysis tools distributed by the Broad Institute
- Built on the GATK open-source map/reduce programming framework



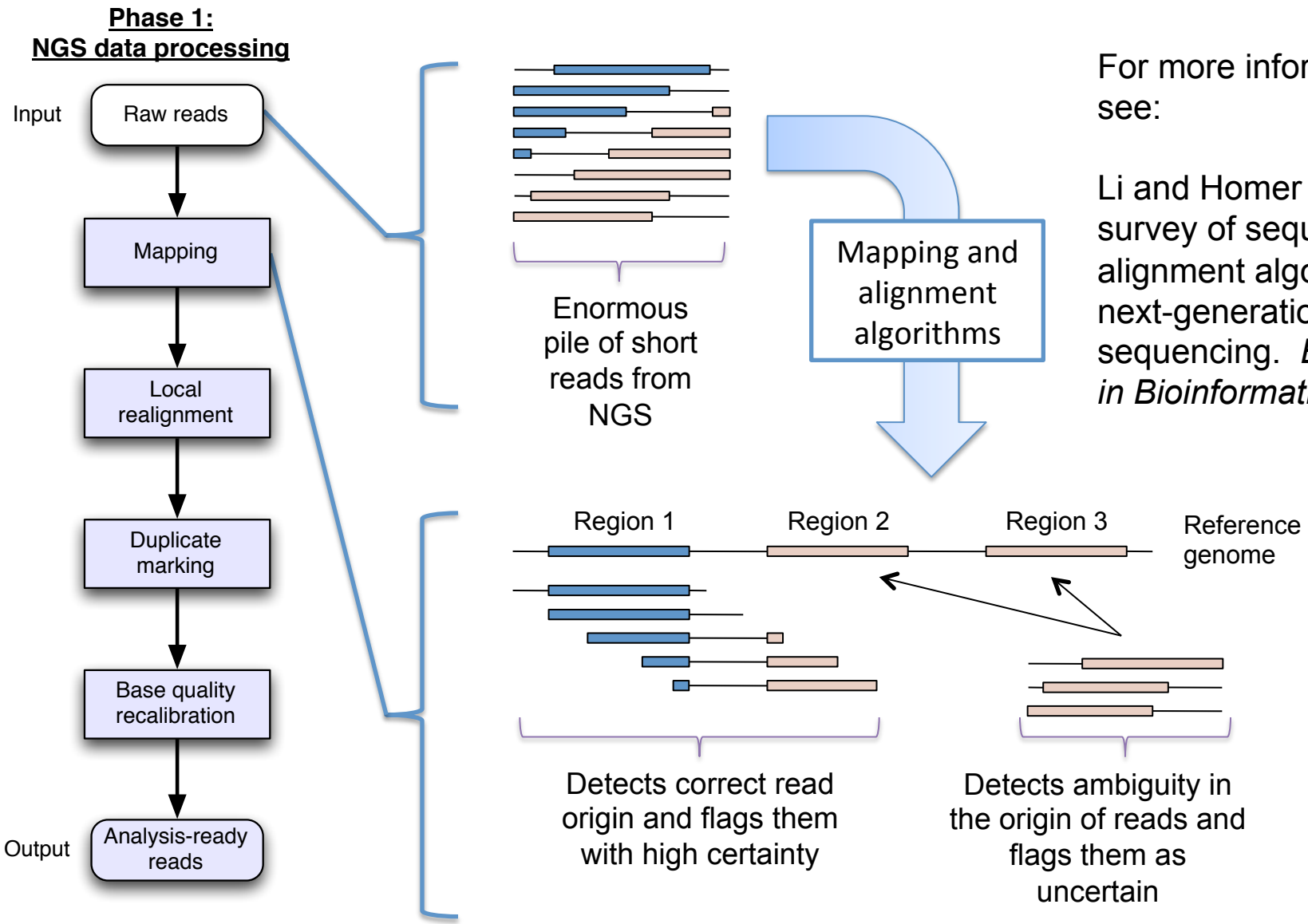
Impact

- >800 of GATK tools, some developed outside of the Broad
- Broad's suite of GATK-based tools is industry standard for NGS analysis
  - >10000 users since 2009, ~2500 active in last month
  - 10M jobs run, using >1M CPU days
- Core technology for 1000 Genomes, TCGA, ESP, and most large NGS projects
- Basis for Archon X Prize, Genomes in a Bottle and other reference standards

Publications

- DePristo et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. (2011) *Nat. Genet*.
- McKenna et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*.

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

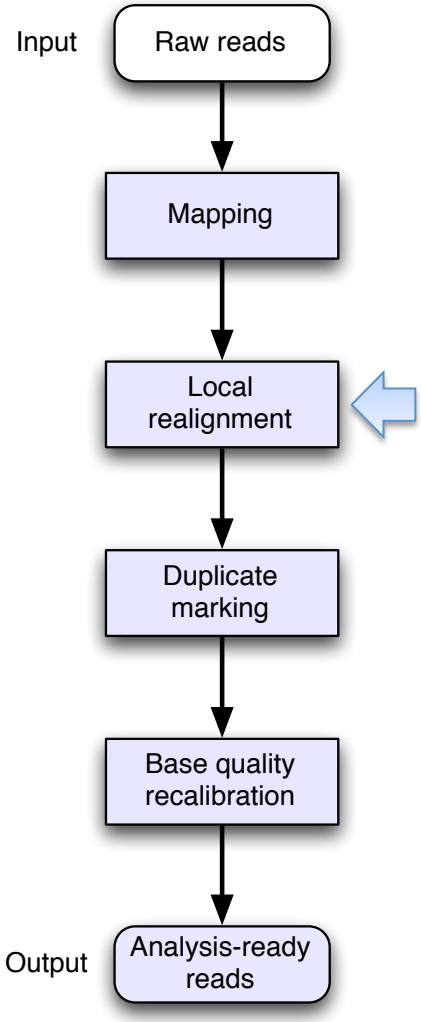


For more information see:

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

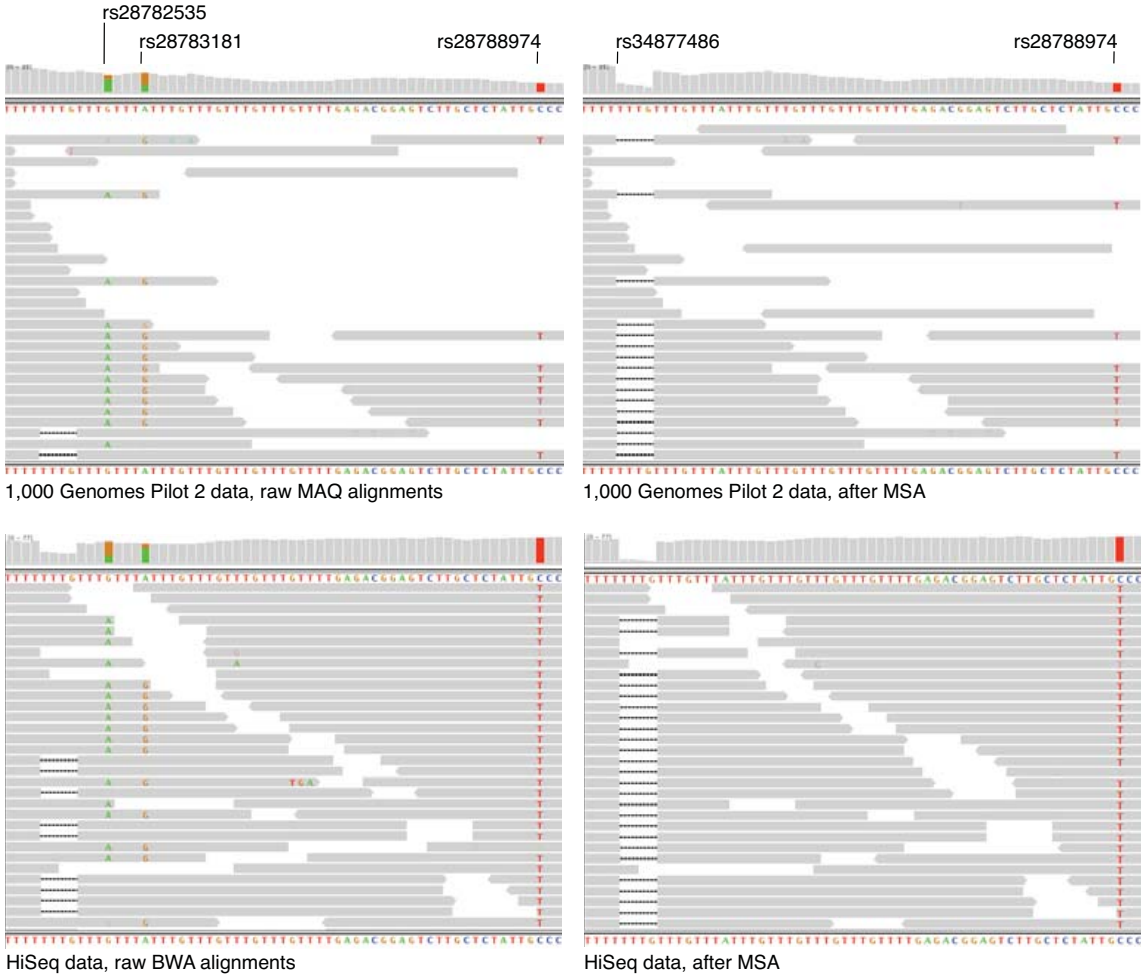
Accurate read alignment through multiple sequence local realignment

Phase 1:
NGS data processing



Effect of MSA on alignments

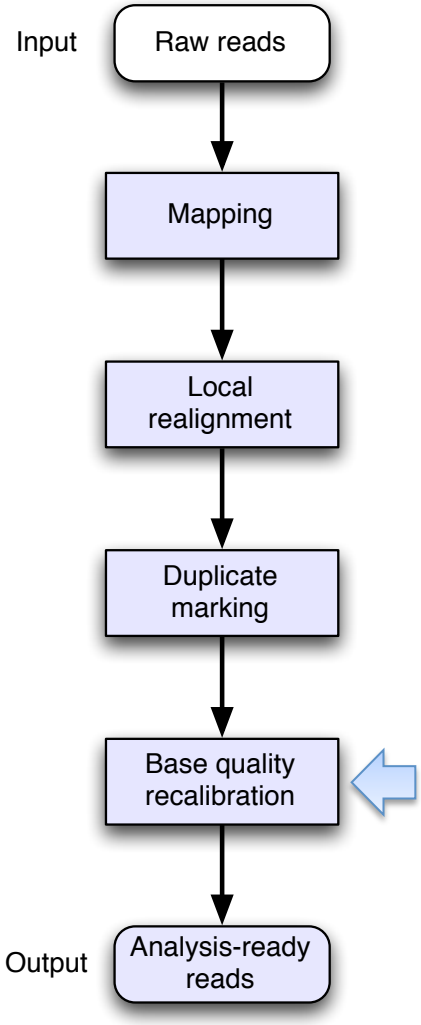
NA12878, chr1:1,510,530-1,510,589



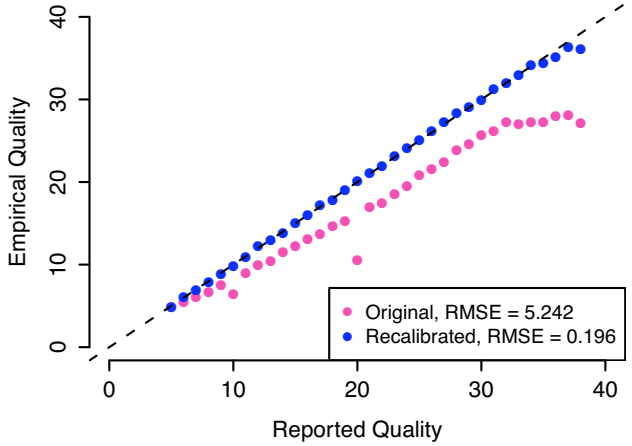
DePristo, M., Banks, E., Poplin, R. et. al, (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet.

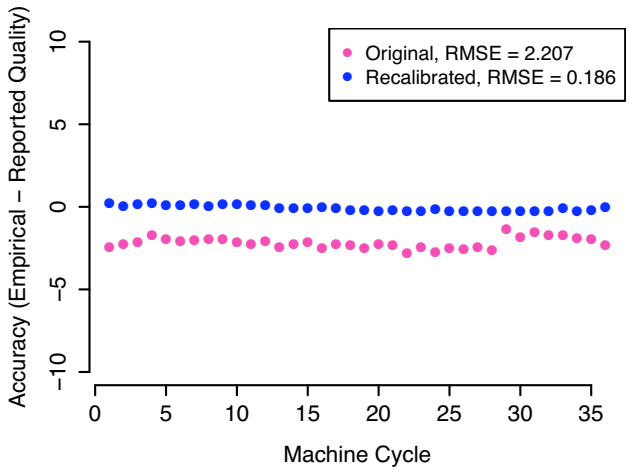
Accurate error modeling with base quality score recalibration

Phase 1:
NGS data processing

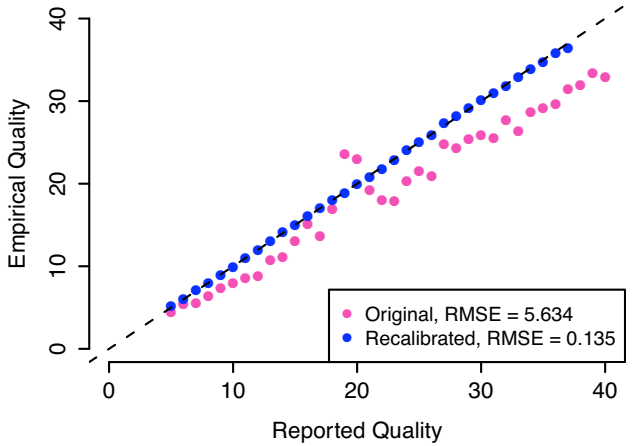


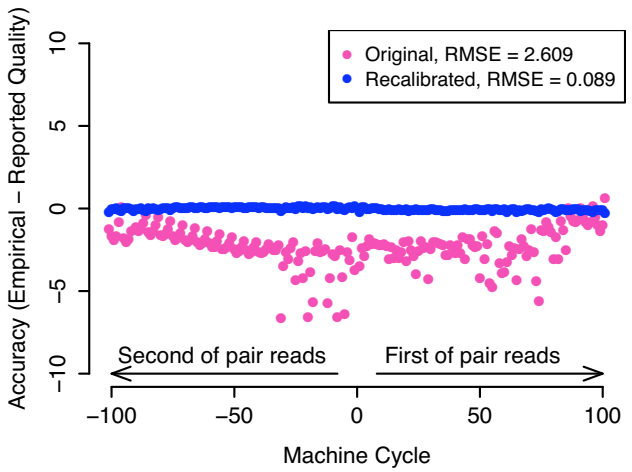
Illumina/GenomeAnalyzer





Illumina/HiSeq 2000



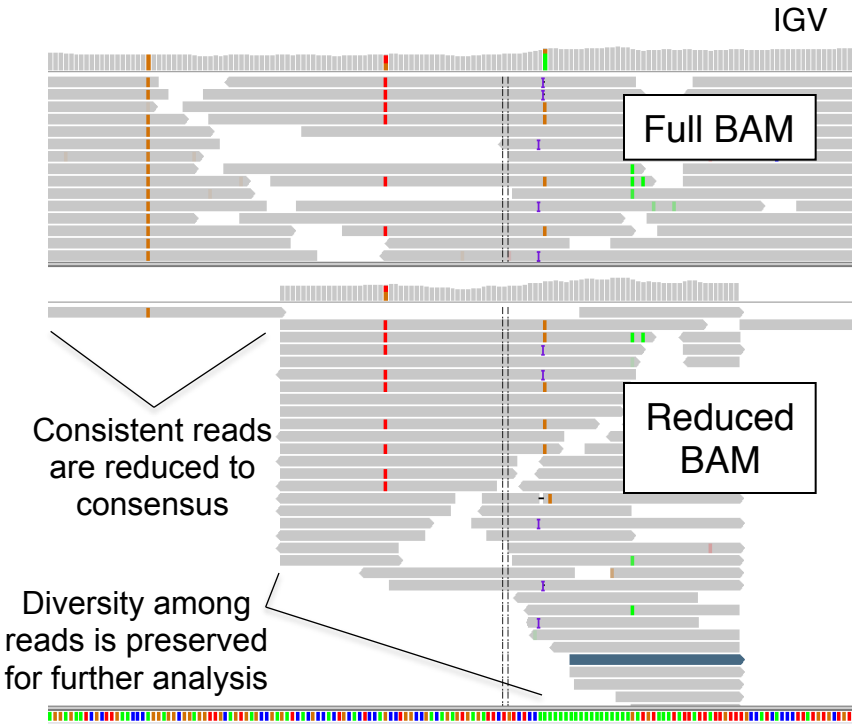


Ryan Poplin

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

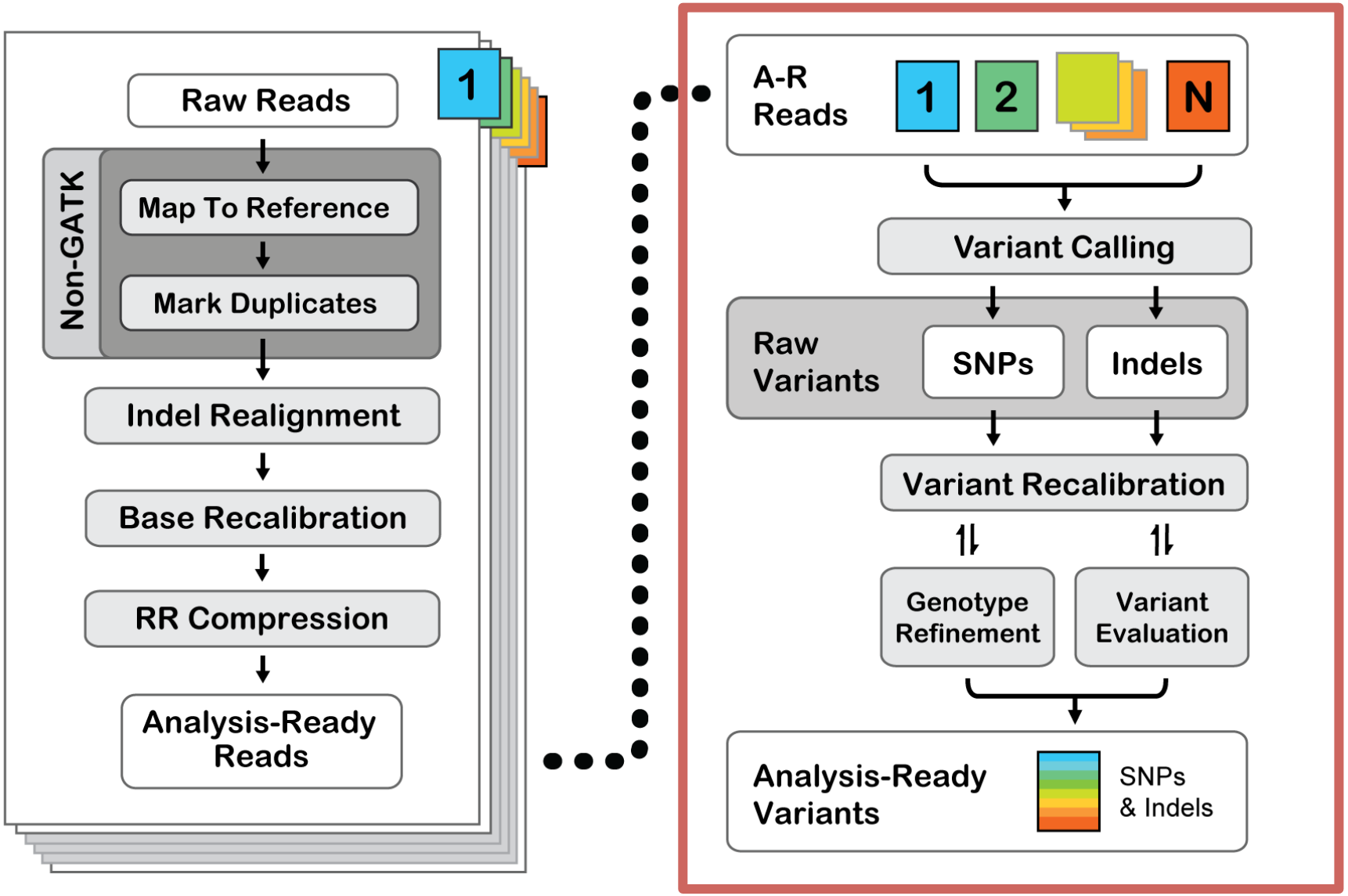
Reduced representation BAM is a novel, highly compressed intermediate format

- BAMs are problematically large
  - Expensive, e.g., 10K exomes requires 250 TB of storage ($250K / year)
  - Hard to analyze (lots of CPUs)
  - Hard to transfer (takes months)
- Reduced read compression addresses these problems
- Potentially large cost savings
  - Archival BAM on cheap storage
  - Reduced bam on high-performance storage
  - Potential cost savings of 10x
    - Maybe more depending on how well we can compress archival BAM free of need for frequent random access for analytics



Full BAM size	~ 10 GB (exome)
Reduced BAM size	~ 100 MB
Compression ratio	~ 50x
SNP calling	As accurate
Indel calling	As accurate
Calling runtime	>50x faster

From reads to variants



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

Prior of the genotype

Likelihood of the genotype

Diploid assumption

Bayesian model

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

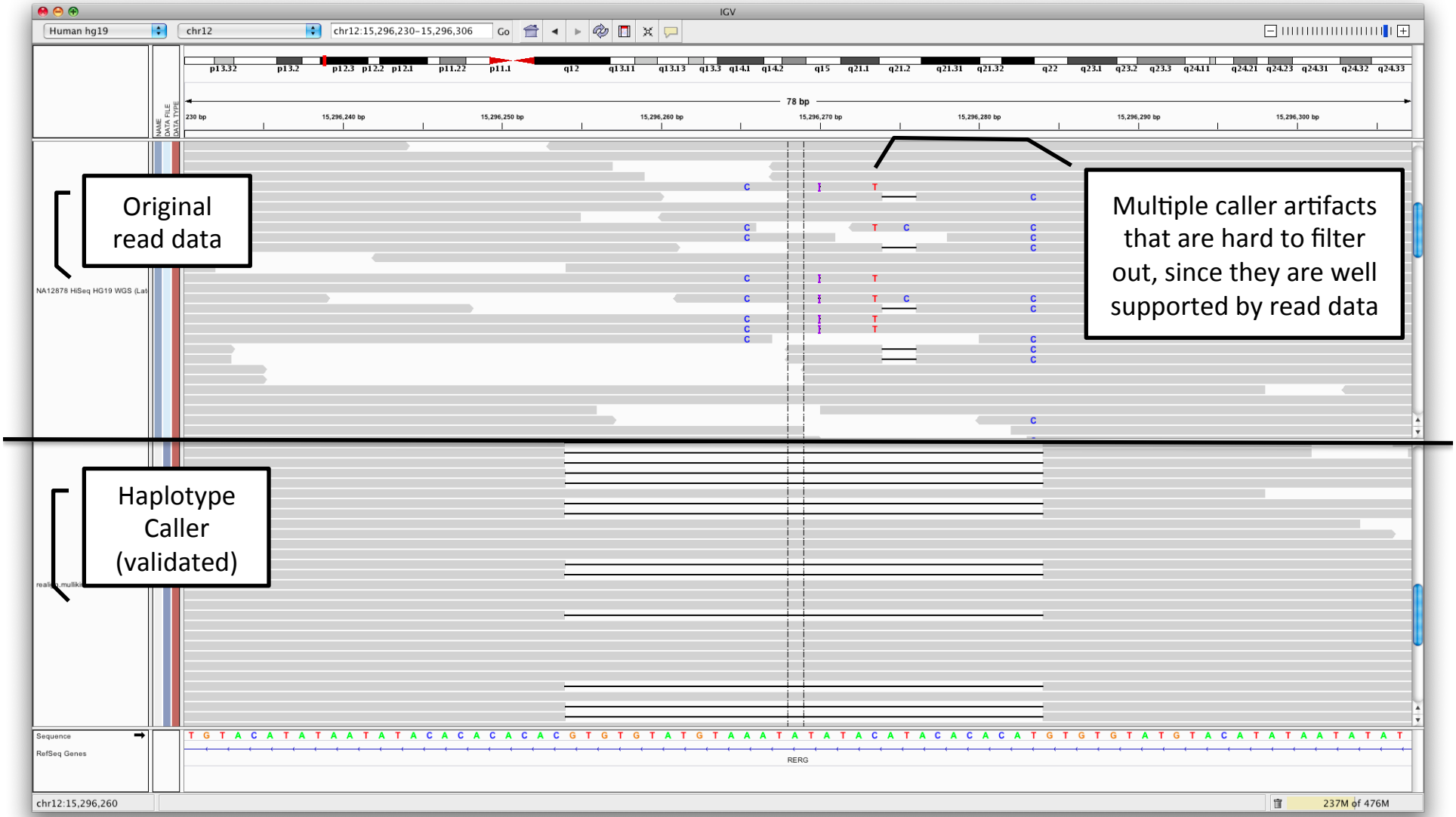
$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2} \right) \text{ where } G = H_1 H_2$$

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

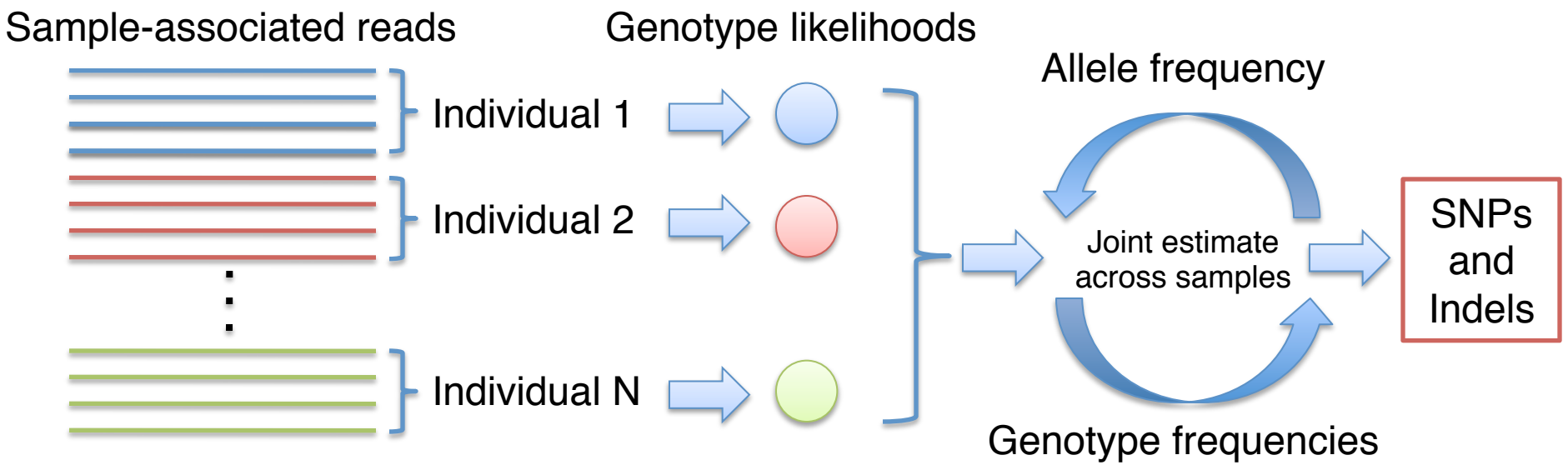
- Inference: what is the genotype G of each sample given read data D for each sample?
- Calculate via Bayes' rule the probability of each possible G
- Product expansion assumes reads are independent
- Relies on a likelihood function to estimate probability of sample data given proposed haplotype

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

The Haplotype Caller uses local assembly to recover true large indel events and avoid their associated artifacts



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

The product is a squared-off matrix of variants x samples with likelihoods

All case and control samples

~3M variants

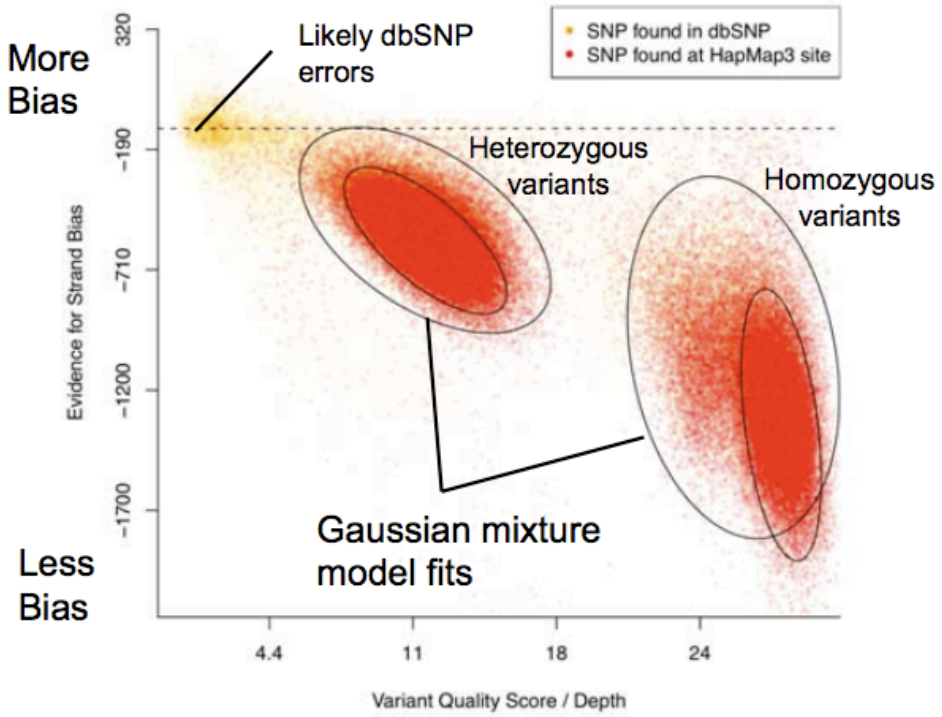
	Site	Variant	Sample 1	Sample 2	…	Sample N
SNP	1:1000	A/C	0/0 0,10,100	0/1 20,0,200	…	0/0 0,100,255
Indel	1:1050	T/TC	0/0 0,10,100	0/0 0,20,200	…	1/0 255,0,255
SNP	1:1100	T/G	0/0 0,10,100	0/1 20,0,200	…	0/0 0,100,255
	…	…	…	…	…	…
SNP	X:1234	G/T	0/1 10,0,100	0/1 20,0,200	…	1/1 255,100,0

Genotypes:
0/0 ref
0/1 het
1/1 hom-alt

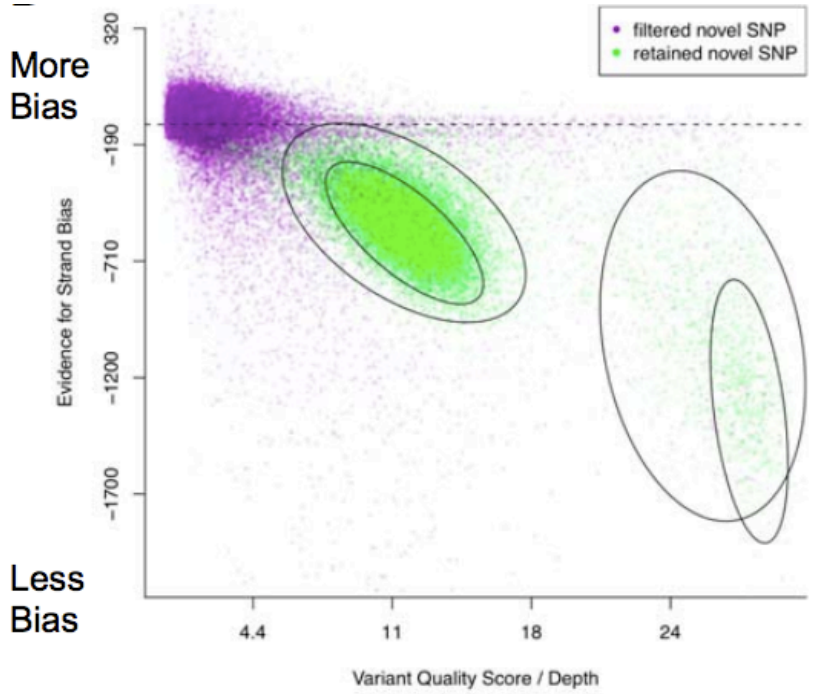
Likelihoods:
A/B/C phred-scaled probability of hom (A), het (B), hom-alt (C) genotypes given NGS data

Variant Quality Score Recalibration (VQSR): 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.

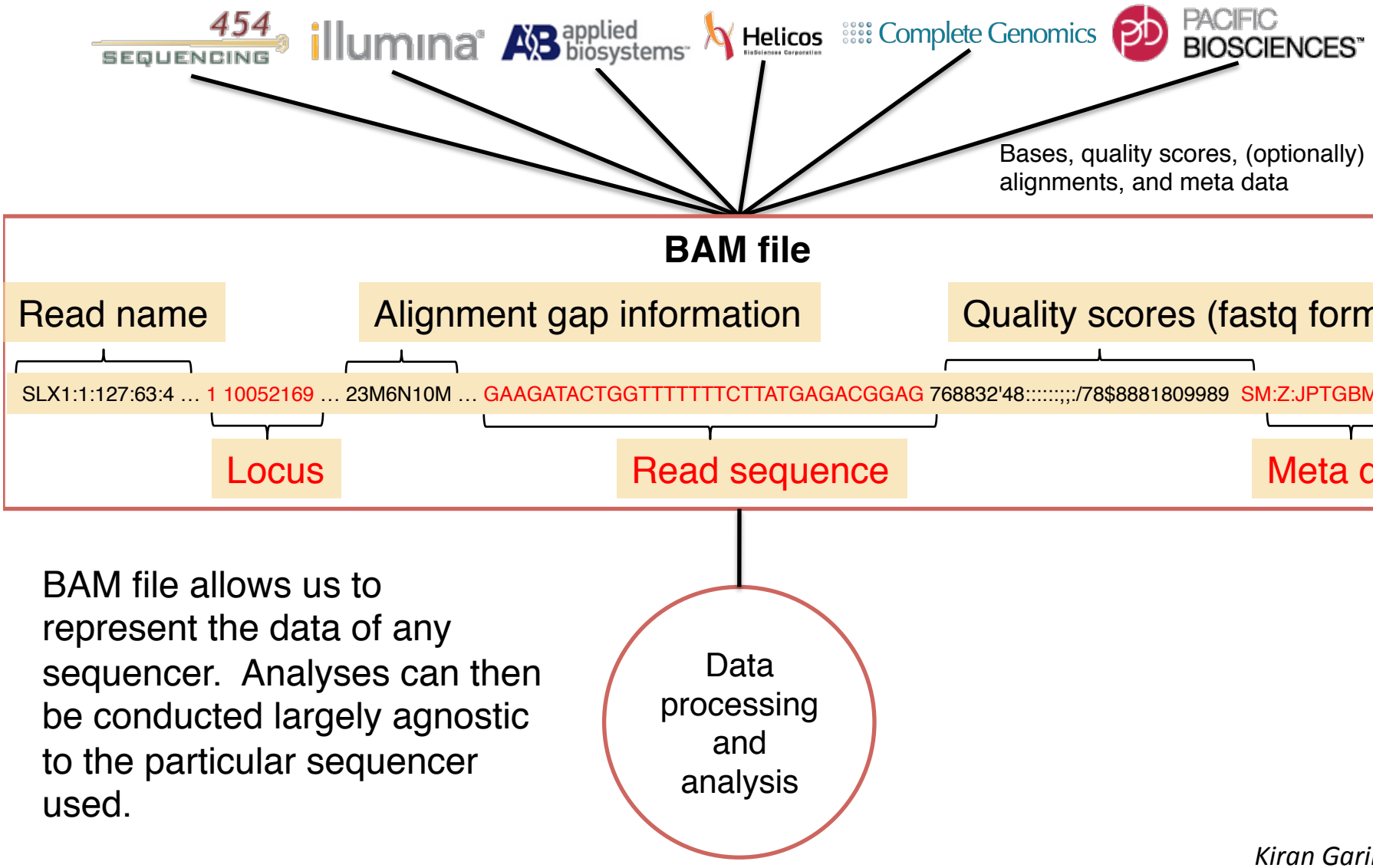


DePristo, M., Banks, E., Poplin, R. et. al, (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet.

BAM: NGS reads, VCF: variant calls and genotypes

KEY FILE FORMATS

The BAM format facilitates sequencer-agnostic analyses



BAM file allows us to represent the data of any sequencer. Analyses can then be conducted largely agnostic to the particular sequencer used.

Kiran Garimella

BAM headers: an optional (but essential) part of a BAM file

```
@HD    VN:1.0  GO:none SO:coordinate
@SQ    SN:chrM        LN:16571
@SQ    SN:chr1        LN:247249719
@SQ    SN:chr2        LN:242951149
[cut for clarity]
@SQ    SN:chr9        LN:140273252
@SQ    SN:chr10       LN:135374737
@SQ    SN:chr11       LN:134452384
[cut for clarity]
@SQ    SN:chr22    LN:49691432
@SQ    SN:chrX        LN:154913754
@SQ    SN:chrY        LN:57772954
@RG    ID:20FUK.1    PL:illumina    PU:20FUKAAXX100202.1    LB:Solexa-18483 SM:NA12878    CN:BI
@RG    ID:20FUK.2    PL:illumina    PU:20FUKAAXX100202.2    LB:Solexa-18484 SM:NA12878    CN:BI
@RG    ID:20FUK.3    PL:illumina    PU:20FUKAAXX100202.3    LB:Solexa-18483 SM:NA12878    CN:BI
@RG    ID:20FUK.4    PL:illumina    PU:20FUKAAXX100202.4    LB:Solexa-18484 SM:NA12878    CN:BI
@RG    ID:20FUK.5    PL:illumina    PU:20FUKAAXX100202.5    LB:Solexa-18483 SM:NA12878    CN:BI
@RG    ID:20FUK.6    PL:illumina    PU:20FUKAAXX100202.6    LB:Solexa-18484 SM:NA12878    CN:BI
@RG    ID:20FUK.7    PL:illumina    PU:20FUKAAXX100202.7    LB:Solexa-18483 SM:NA12878    CN:BI
@RG    ID:20FUK.8    PL:illumina    PU:20FUKAAXX100202.8    LB:Solexa-18484 SM:NA12878    CN:BI
@PG    ID:BWA  VN:0.5.7       CL:tk
@PG    ID:GATK TableRecalibration      VN:1.0.2864
20FUKAAXX100202:1:1:12730:189900      163    chrM    1      60     101M    =      282    381
        GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTA…[more bases]
        ?BA@A>BBBBACBBAC@BBCBBCBC@BC@CAC@:BBCBBCACAACBABCBCCAB…[more quals]
        RG:Z:20FUK.1    NM:i:1  SM:i:37 AM:i:37 MD:Z:72G28      MQ:i:60 PG:Z:BWA       UQ:i:33
```

Required: Standard header

Essential: contigs of aligned reference sequence. Should be in karotypic order.

Essential: read groups. Carries platform (PL), library (LB), and sample (SM) information. Each read is associated with a read group

Useful: Data processing tools applied to the reads

VCF (Variant Call Format) is a standard format to represent SNPs, Indels, SV calls from NGS data

(a) VCF example

Header

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	40	PASS	.	GT:DP	1/1:13	2/2:29
1	2	.	C	T,CT	.	PASS	H2;AA=T	GT	0|1	2/2
1	5	rs12	A	G	67	PASS	.	GT:DP	1|0:16	2/2:20
X	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=299	GT:GQ:DP	1:12:.	0/0:20:36

(b) SNP

Alignment

```
1234
ACGT
ATGT
 ^
```

VCF representation

POS	REF	ALT
2	C	T

(c) Insertion

```
12345
AC-GT
ACTGT
  ^
```

POS	REF	ALT
2	C	CT

(d) Deletion

```
1234
ACGT
A--T
 ^^
```

POS	REF	ALT
1	ACG	A

(e) Replacement

```
1234
ACGT
A-TT
 ^^
```

POS	REF	ALT
1	ACG	AT

QUALITY CONTROL

Evaluating SNP call quality

Expected number of calls?

- The number of SNP calls should be close to the average human heterozygosity of 1 variant per 1000 bases
- Only detects gross under/over calling

Concordance with genotype chip calls?

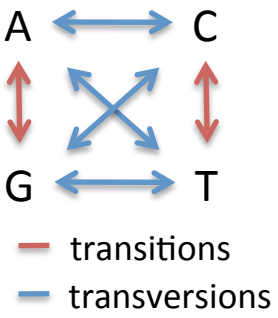
- Often we have genotype chip data that indicates the hom-ref, het, hom-var status at millions of sites
- Good SNP calls should be >99.5% consistent these chip results, and >99% of the variable sites should be found
- The chip sites are in the better parts of the genome, and so are not representative of the difficulties at novel sites

What fraction of my calls are already known?

- dbSNP catalogs most common variation, so most of the true variants found will be in dbSNP
- For single sample calls, ~90 of variants should be in dbSNP
- Need to adjust expectation when considering calls across samples

Transition to transversion ratio (Ti/Tv)?

- Transitions are twice as frequent as transversions (see *Ebersberger, 2002*)
  - Validated human SNP data suggests that the Ti/Tv should be ~2.1 genome-wide and ~2.8 in exons
- FP SNPs should has Ti/Tv around 0.5
- Ti/Tv is a good metric for assessing SNP call quality

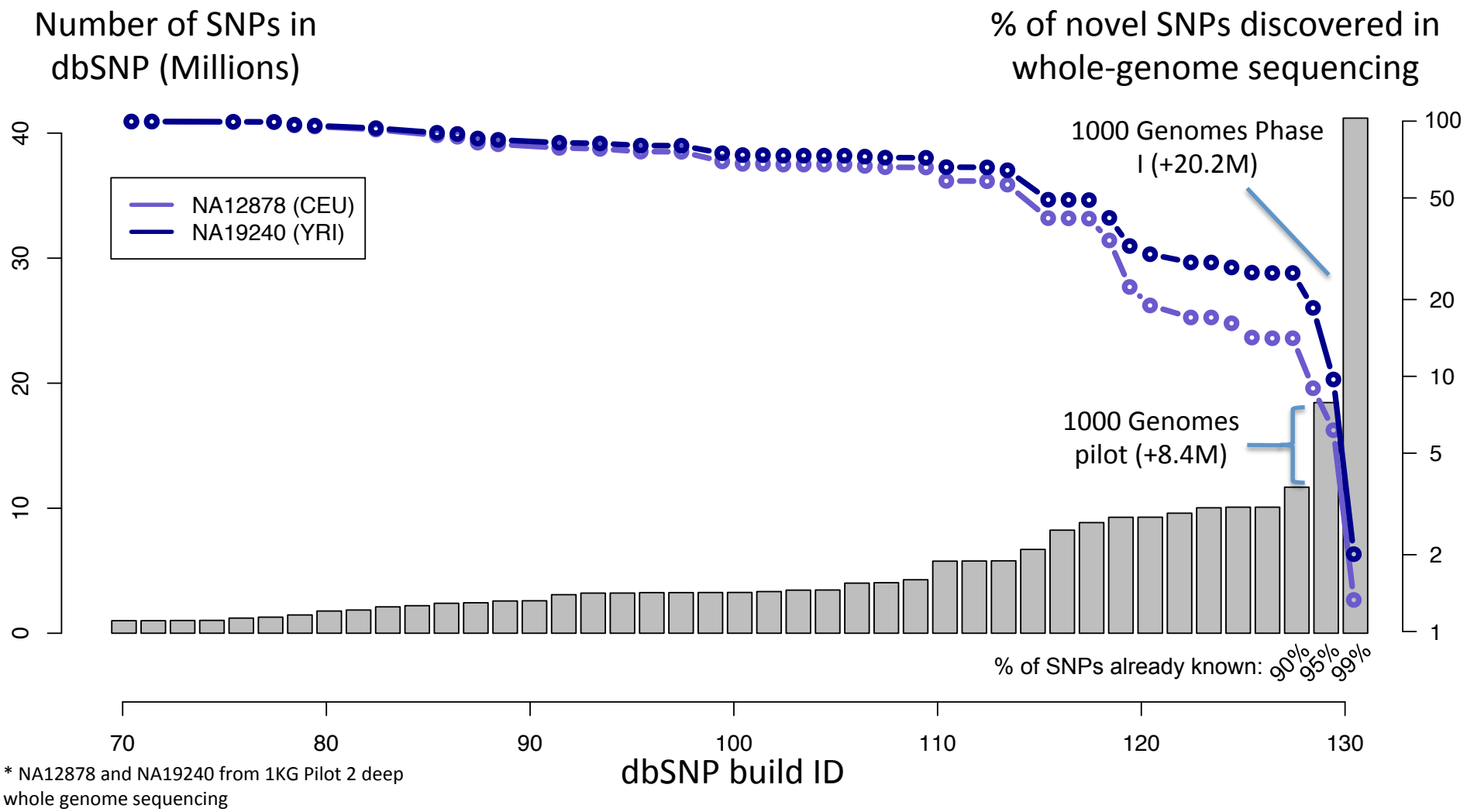


How many calls should I expect?

$$\text{Number of polymorphic sites} \approx L \cdot \theta \sum_{i=1}^{2N} 1/i$$

Sample ethnicity	Target (L)	Heterozygosity (θ)	Expected number of variants
Single sample => 2N = 2			
European	Whole genome	0.78×10^{-3}	~3.3M
European	32Mb exome	0.42×10^{-3}	~20K
African	Whole genome	1.00×10^{-3}	~4.3M
African	32Mb exome	0.48×10^{-3}	~23K
100 samples => 2N = 200			
European	Whole genome	0.78×10^{-3}	~13M

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

Genotype sensitivity and concordance

- Genotype chips (Illumina 1M, e.g.) reliably identify genotypes at many sites
  - >99% accuracy at known polymorphic sites
- Sensitivity measure: what fraction of variant sites in sample on chip did I found?
- Genotype quality: how concordant are the genotype calls from NGS with the chip?
- Caveats: not comprehensive, only at easy sites, does not apply to novel calls

Transition/transversion ratio (Ti/Tv) can be used to estimate false-positive rate for SNPs

Data set	Sequencing tech(s)	Year	SNP caller(s)	WGS			Exome		
				N sites	Known Ti/Tv	Novel Ti/Tv	N sites	Known Ti/Tv	Novel Ti/Tv
Complete Genomics NA19240[1]	CGI	2009	CGI	4.1M	2.14	2.09	20.2K	3.42	2.98
1000 Genomes CEU trio	Solexa, SOLiD and 454	2010	glfTrio, GATK	3.6M	2.08	2.02	17.6K	3.54	2.74
1000 Genomes YRI trio	Solexa, SOLiD and 454	2010	glfTrio, GATK	4.5M	2.09	2.07	25K	3.51	3.18
Weighted average					**2.10**	**2.07**	**-**	**3.49**	**2.98**

Because random errors will have Ti/Tv ratio of ~0.5, we can estimate the accuracy of novel SNP calls by their difference from expectation

Indel QC: active area of research

- How do you know your indel call set is good?
- 200-500 indels per exome
  - 200 (Yu), 300 (Eichler, CGI, 1000G) but up to 500 in ESP
  - Depends on coding definition => more conserved genes have many fewer indels
- Indels in exome should be enriched for in-frame events
  - The majority of events should preserve frame (~50-60%)
- Up to ~10% of indels are multi-allelic across samples
- True indels should look approximately similar across samples
  - E.g., number of indels, length distribution, ratio to SNPs

Summary

- Next-generation sequencing is a maturing technology
  - Multiple robust NGS technologies
  - Several canonical experimental designs
- Sophisticated data processing algorithms are available with reliable statistical approaches for SNP and indel discovery and genotyping
- Suite of quality-control metrics to assess the reliability of your experimental results