Best practices for Variant Calling with Pacific Biosciences data

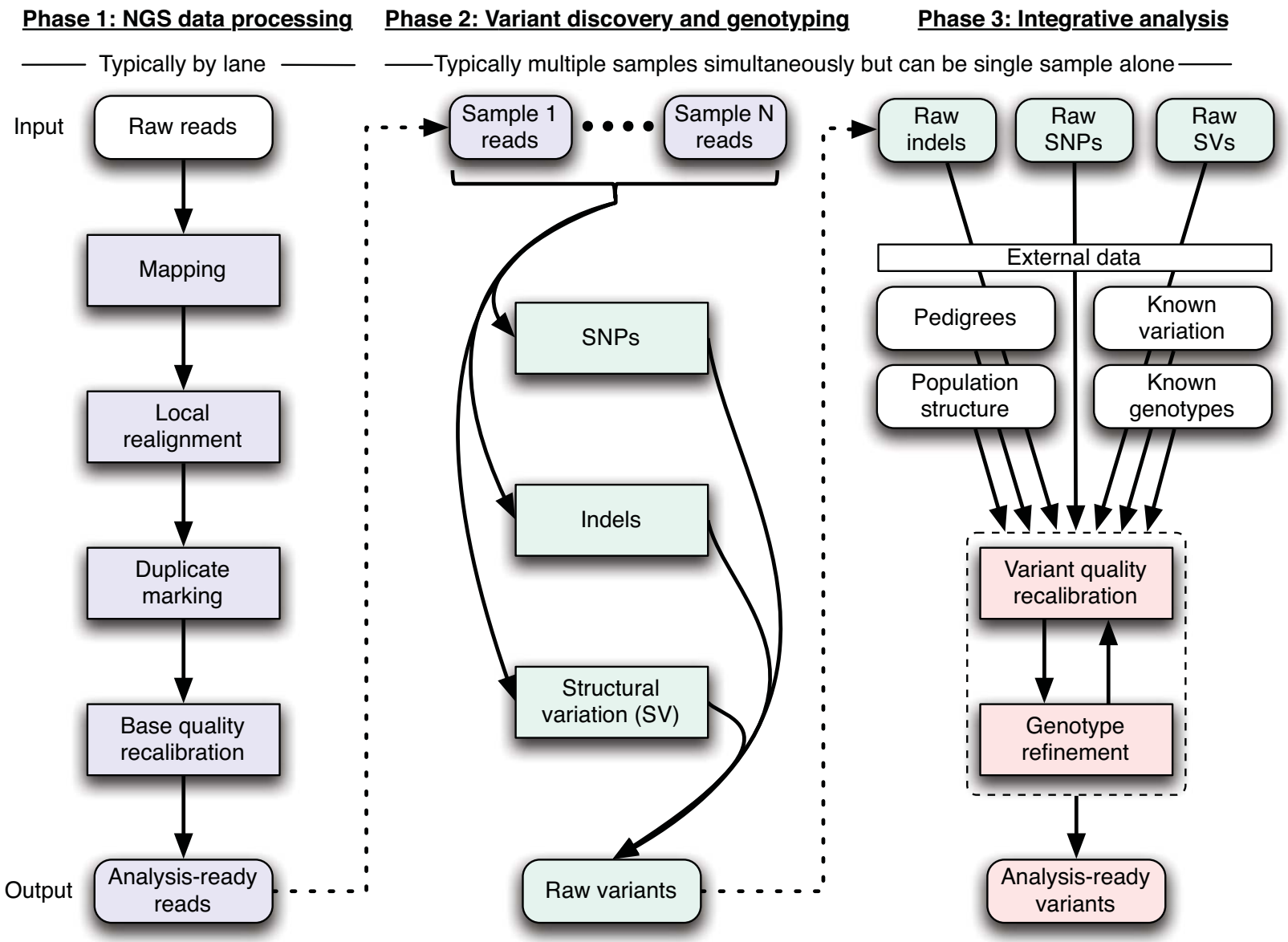
Mauricio Carneiro, Ph.D.
Mark DePristo, Ph.D.

Genome Sequence and Analysis
Medical and Population Genetics
carneiro@broadinstitute.org

General best practice data processing and variant calling using the GATK

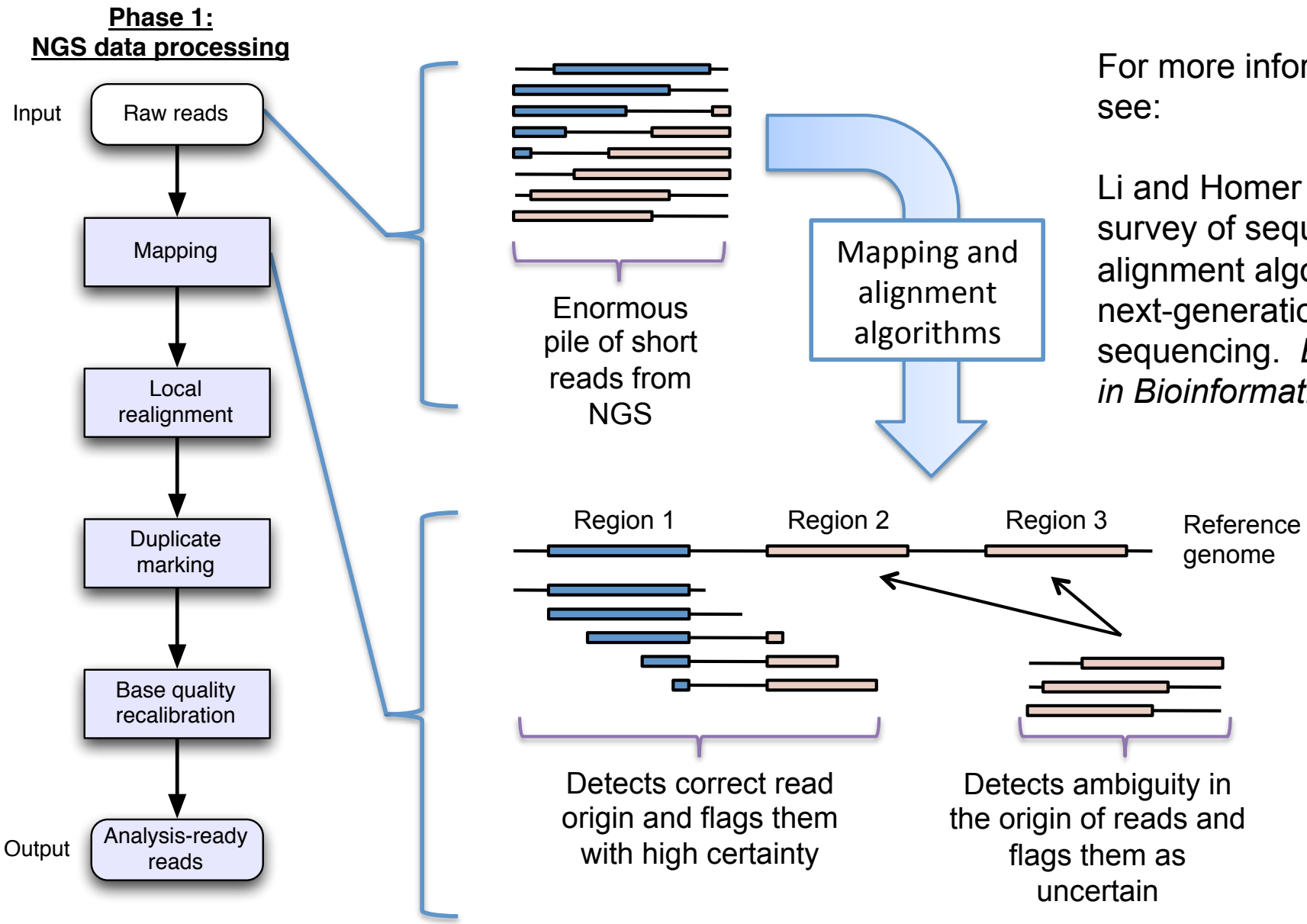
The Current Pipeline

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

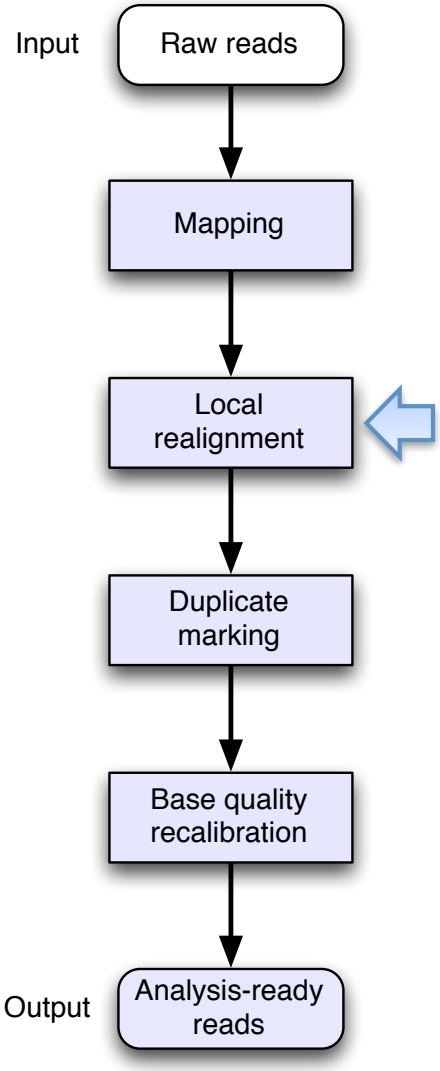


For more information see:

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

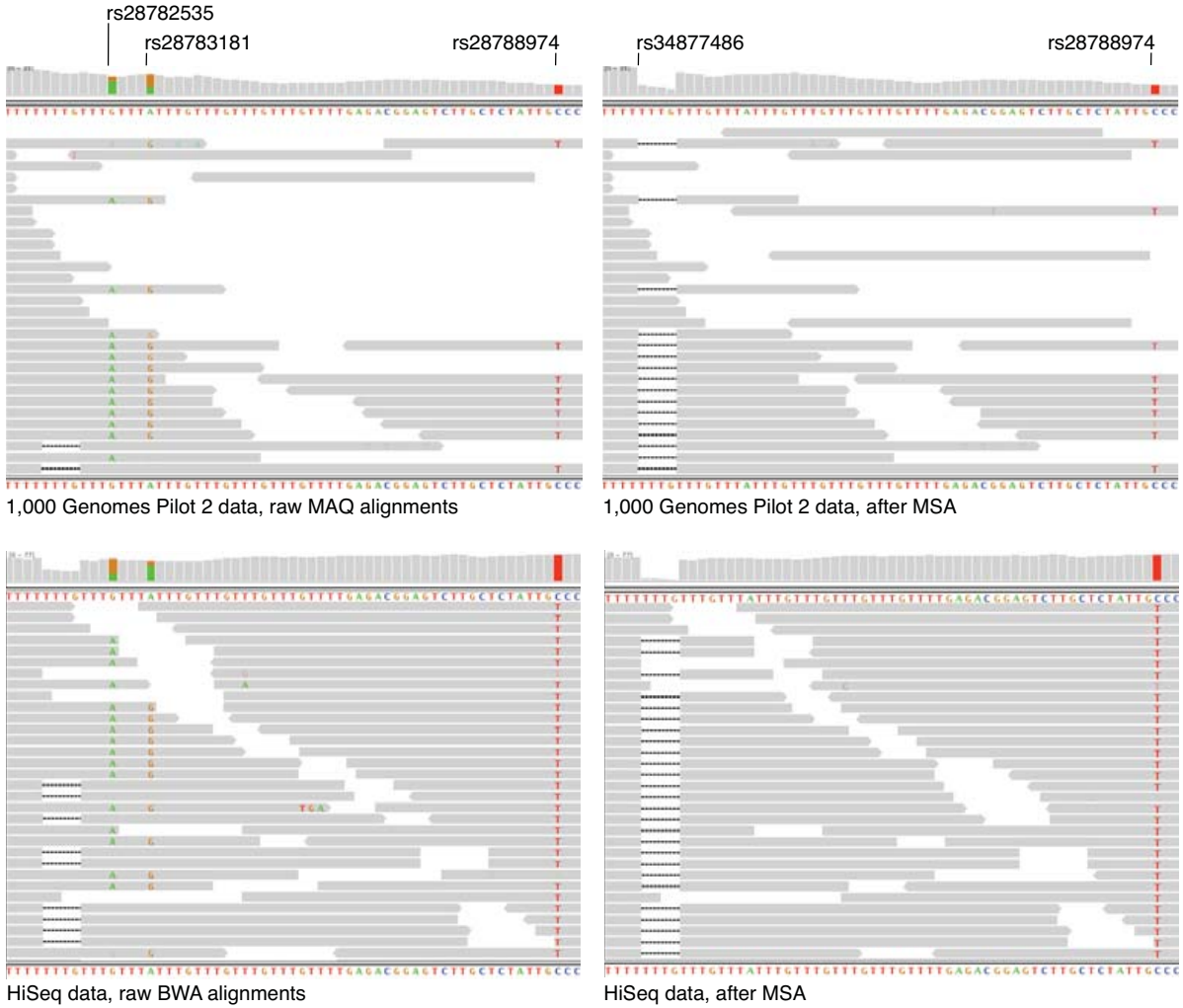
Accurate read alignment through multiple sequence local realignment

Phase 1:
NGS data processing



Effect of MSA on alignments

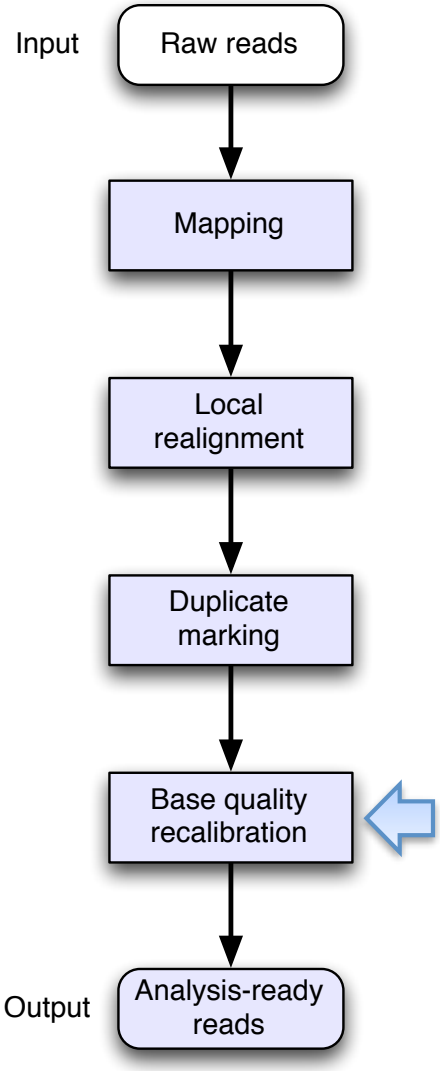
NA12878, chr1:1,510,530-1,510,589



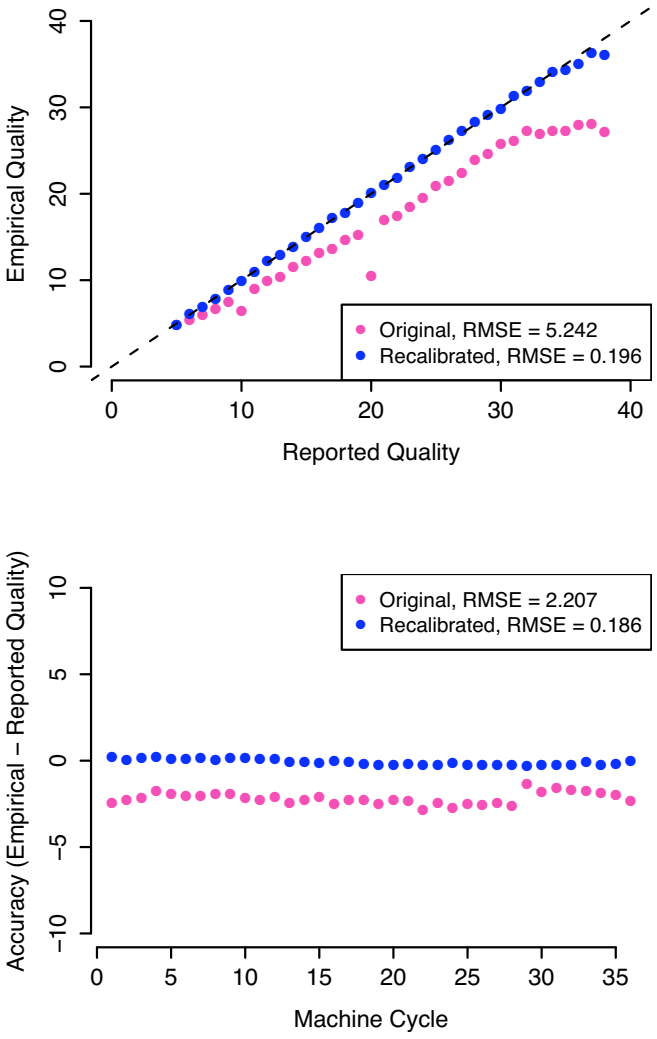
DePristo, M., Banks, E., Poplin, R. et. al, (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet.

Accurate error modeling with base quality score recalibration

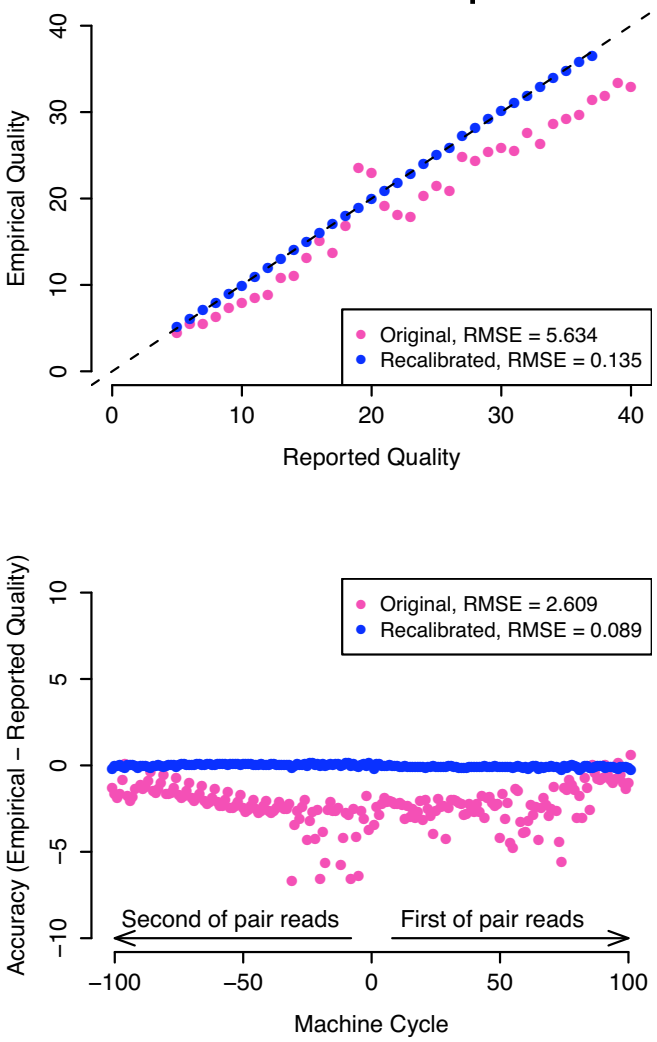
Phase 1:
NGS data processing



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.

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

Prior of the genotype

Likelihood of the genotype

Bayesian model

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

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

Diploid assumption

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

- 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

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

SNP genotype likelihoods

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

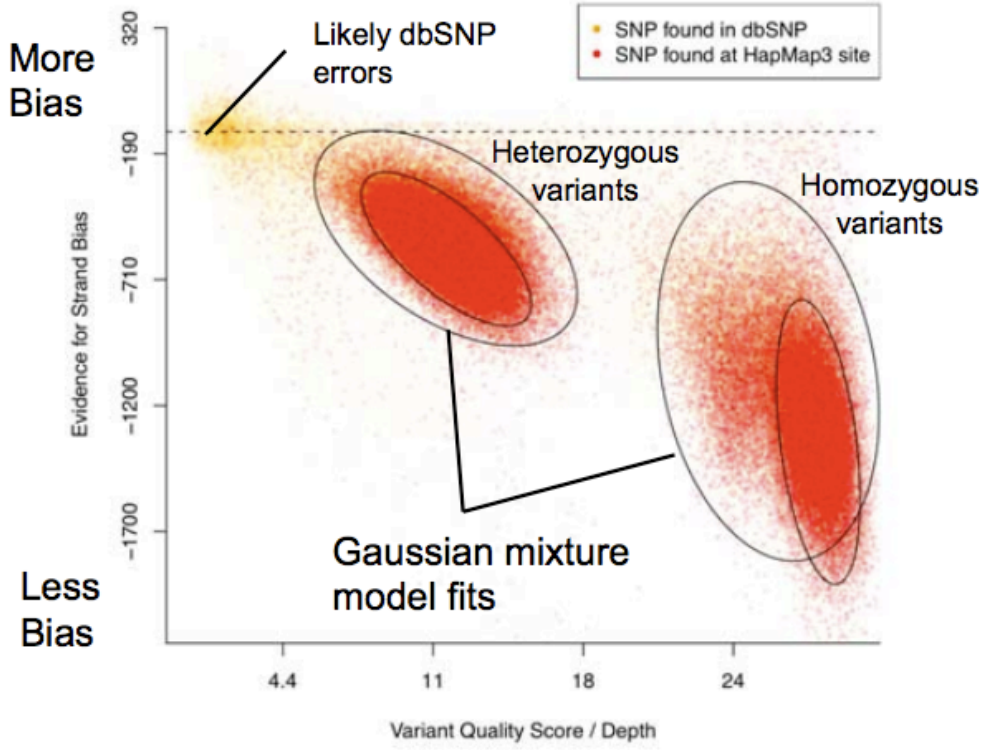
$$\Pr\{D_j|b\} = \begin{cases} 1-\epsilon_j & D_j = b, \\ \epsilon_j & \text{otherwise.} \end{cases}$$

- 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

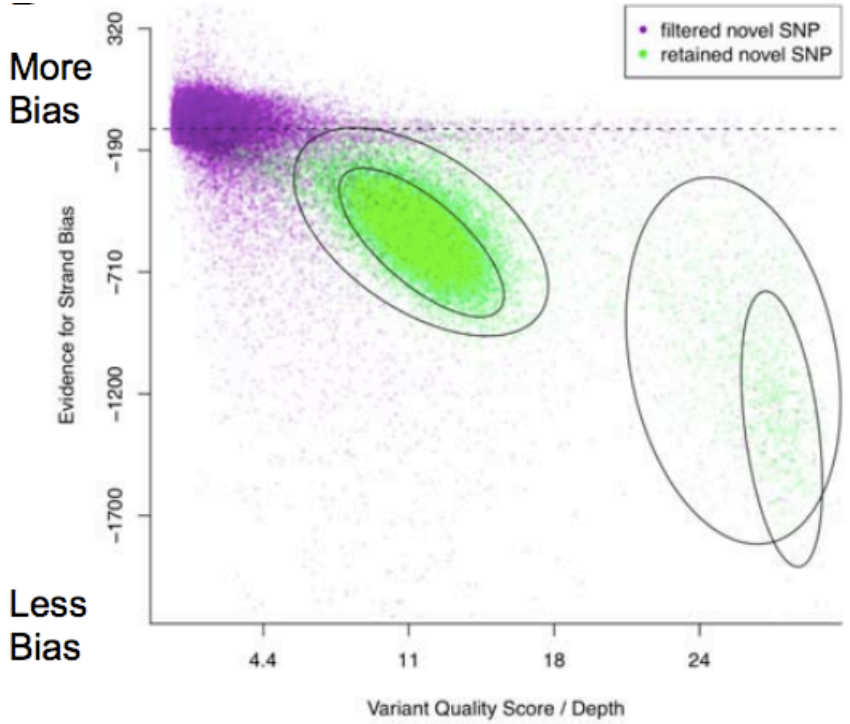
28 See http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper for more information

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.



32

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

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

Indel realignment	Base quality score recalibration
Unified Genotyper	VQSR
Variant Eval	Many other 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

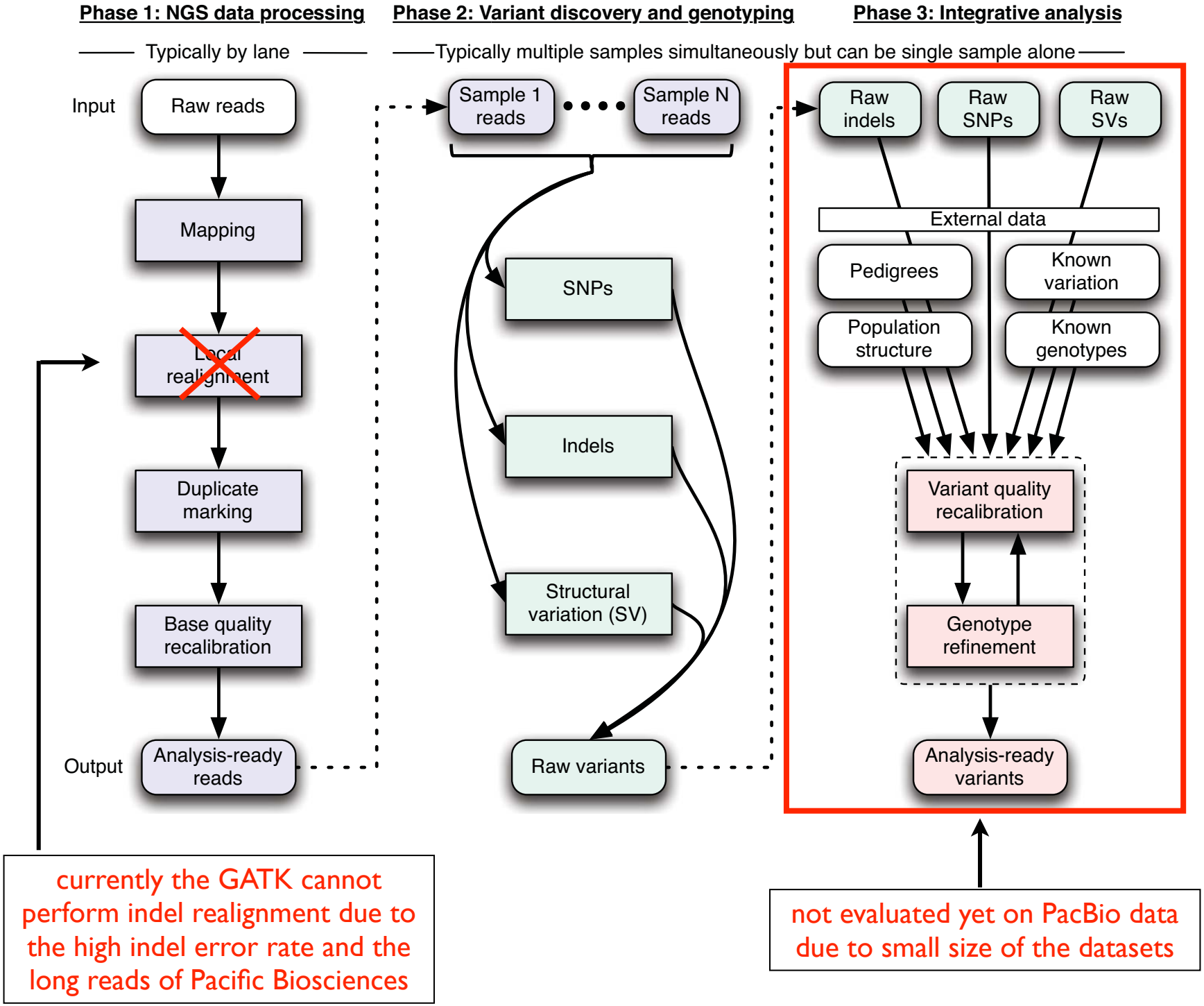
- Standard and accessible format for storing population variation and individual genotypes

http://vcftools.sourceforge.net/

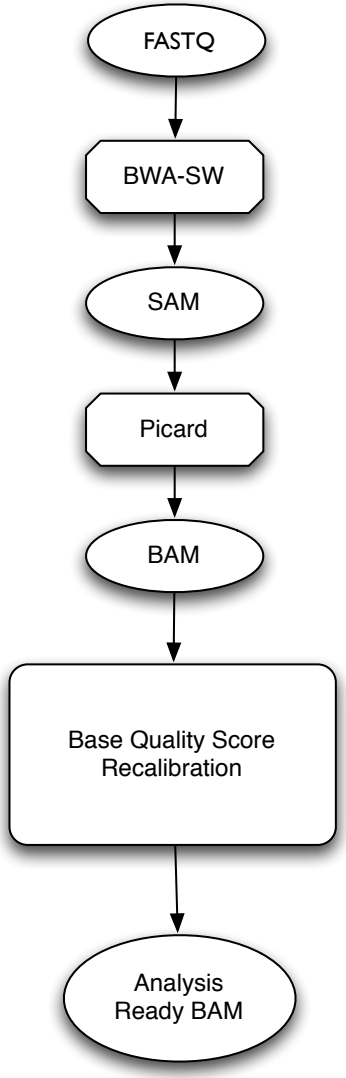
McKenna et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res.

how we apply our pipeline to Pacific Biosciences data
a step-by-step tutorial

Pacbio Processing Pipeline


Phase 1: NGS data processing
Phase 2: Variant discovery and genotyping
Phase 3: Integrative analysis
Typically by lane
Typically multiple samples simultaneously but can be single sample alone
Input
Raw reads
Mapping
Local realignment
Duplicate marking
Base quality recalibration
Output
Analysis-ready reads
Sample 1 reads
Sample N reads
SNPs
Indels
Structural variation (SV)
Raw variants
Raw indels
Raw SNPs
Raw SVs
External data
Pedigrees
Known variation
Population structure
Known genotypes
Variant quality recalibration
Genotype refinement
Analysis-ready variants
currently the GATK cannot perform indel realignment due to the high indel error rate and the long reads of Pacific Biosciences
not evaluated yet on PacBio data due to small size of the datasets

Pacbio Processing Pipeline

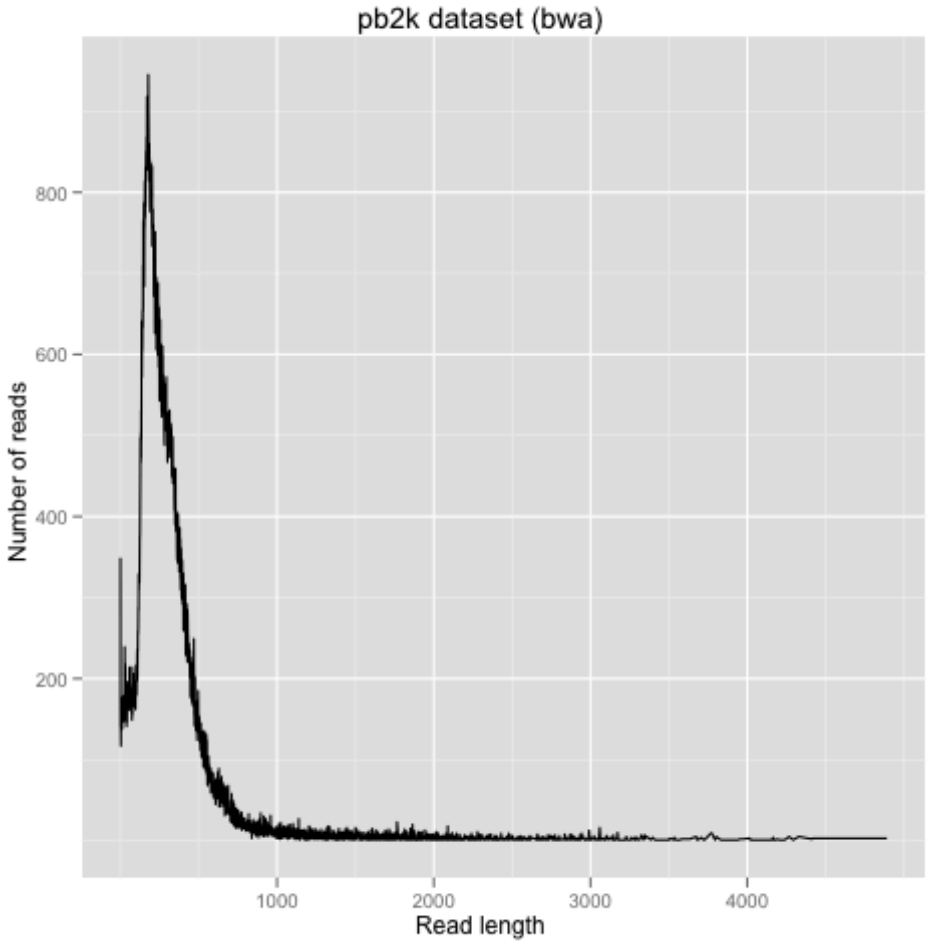


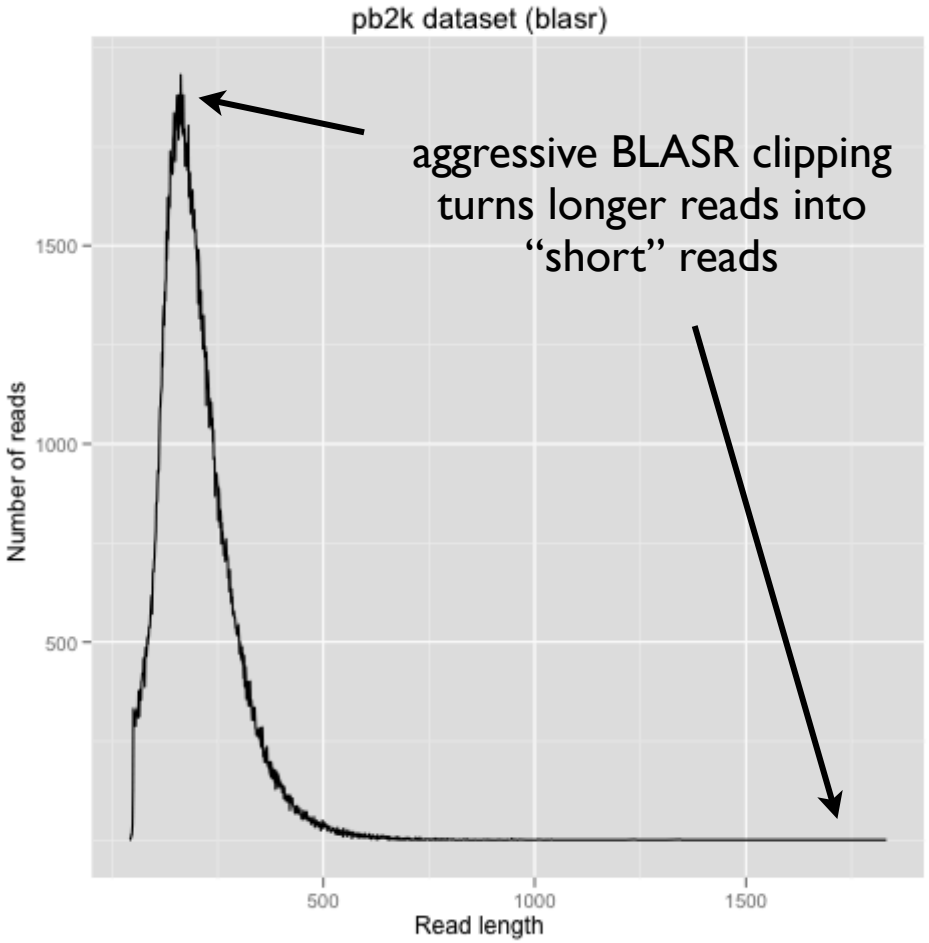
1. We start our processing pipeline with the filtered_subreads.fasta file produced by PacBio software and turn into a fastQ using SMRT Pipeline scripts provided by PacBio.
2. Mapping and Alignment are done using BWA with a heuristic smith waterman algorithm (bwa-sw)
3. We sort the bam file, add read group and sample information using Picard Tools: `SortSam` **and** `AddOrReplaceReadGroups`.
4. We recalibrate base qualities using the GATK's Base Quality Score recalibration framework.

Why do we align with BWA and not BLASR?

- BWA is the standard aligner in the Broad's sequencing platform.
- BLASR is still responsible for generating the filtered sub-reads.
- With recent updates, BLASR generated BAM files are a reasonable alternative for this step of the pipeline
  - optional pipeline starts with a BLASR generated BAM (skipping BWA and Picard steps).
  - Read Group information and BQSR are still required steps.
  - Works well, but generally smaller yield.
  - We anticipate further development in BLASR generated BAMs could improve this alternate pipeline in the future.

Strict BLASR filtering reduces yield and eliminates the longer reads





total mapped coverage: 74,735,274 bp

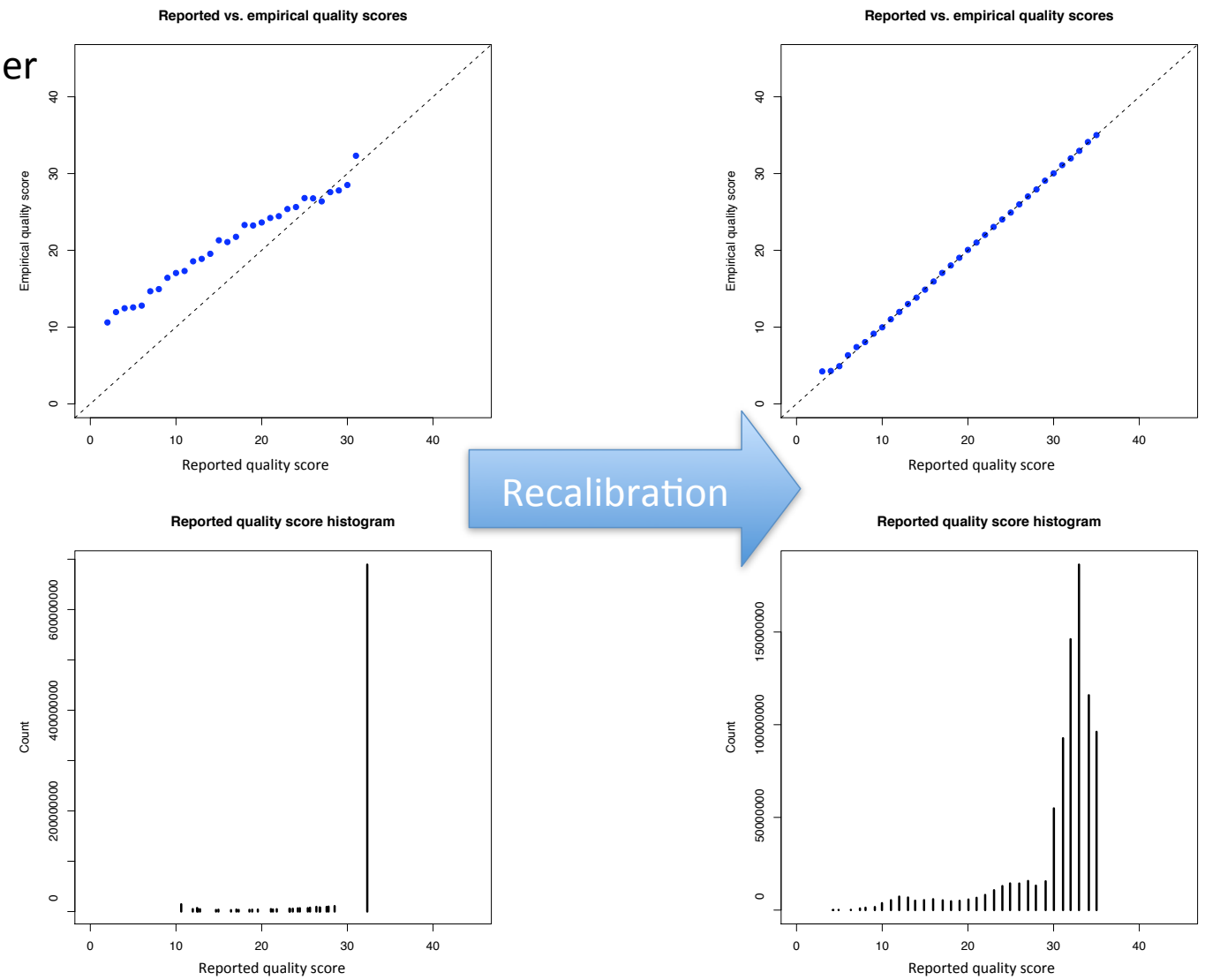
total mapped coverage: 19,562,290 bp

Introduction to Base Quality Score Recalibration

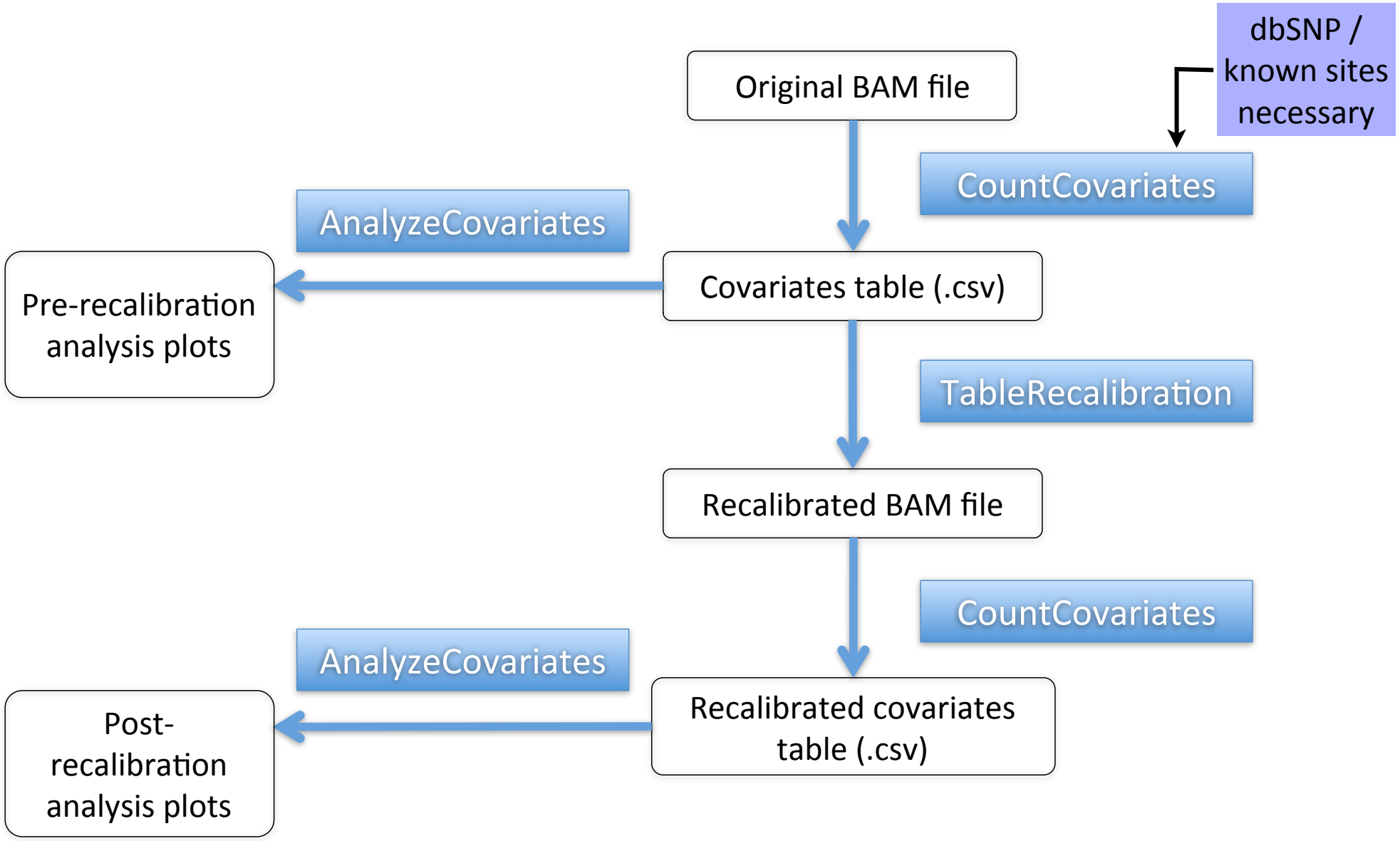
Sequencers provide estimates of error rate per nucleotide

… but they aren't very accurate

… and they aren't very informative



Recalibration workflow



Running CountCovariates

```
java -Xmx4g -jar GenomeAnalysisTK.jar
  -R reference.fasta
  -D dbsnp.vcf
  -I original.bam
  -T CountCovariates
  -cov ReadGroupCovariate
  -cov QualityScoreCovariate
  -cov DinucCovariate
  -cov CycleCovariate
  -recalFile table.recal_data.csv
```

List of known polymorphic sites is necessary so these sites do not count against bases mismatch rate

List of covariates to be used in the recalibration calculation

CSV file containing covariate counts

Table recalibration file (table.recal_data.csv)

```
# Counted Bases    143745620
ReadGroup,QualityScore,Dinuc,Cycle,nObservations,nMismatches,Qempirical
SRR001802,2,AA,-8,165,17,10
SRR001802,2,AA,-2,91,10,10
SRR001802,2,AA,3,5,4,1
SRR001802,2,AA,4,9,4,4
SRR001802,2,AA,7,12,4,5
```

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

Running TableRecalibration

```
java -Xmx4g -jar GenomeAnalysisTK.jar
  -R Homo_sapiens_assembly18.fasta
  -I original.bam
  -T TableRecalibration
  -recalFile table.recal_data.csv
  -outputBam recal.bam
```

Table recalibration file from CountCovariates step

The full recalibrated bam file

A recalibrated copy of the original BAM file

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

Running AnalyzeCovariates

A separate .jar file distributed with the GATK

```
java -Xmx4g -jar AnalyzeCovariates.jar
  -outputDir /path/to/output_dir/
  -resources resources/
  -recalFile table.recal_data.csv
```

The directory in which to place the output analysis plots

Points to the GATK installation's directory of R scripts which are used for plotting the data

Table recalibration file from either the before or after CountCovariates step

Many plots of base quality versus each covariate

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

The Pacbio Processing Pipeline is available for educational purposes (but not supported)

```
java
    -Xmx4g
    -jar Queue.jar
    -S PacbioProcessingPipeline.scala
    -i filtered_subreads.fastq
    -D dbSNP.vcf
    -R reference.fasta
    -run
```

or blasr.bam with extra -blasr option

Queue is part of the GATK and is a pipeline manager used internally at the Broad in most analysis projects

(see http://www.broadinstitute.org/gsa/wiki/index.php/Queue)

Calling snps and indels using pacbio data with the Unified Genotyper

```
java
    -Xmx4g
    -jar GenomeAnalysisTK.jar
    -T UnifiedGenotyper
    -I input.recal.bam
    -R reference.fasta
    -D dbsnp.vcf
    -deletions 0.5
    -o myCalls.vcf
    -mbq 10
```

allows sites with 50% deletions to be analyzed

minimum base quality 10 calibrates the UG for PacBio data (avg base qual is 20)

The ideal *deletions* and *minimum base quality* parameters for this specific dataset were determined systematically by measuring sensitivity/specificity to known variant calls in *NA12878*.

more information available at the poster session of AGBT

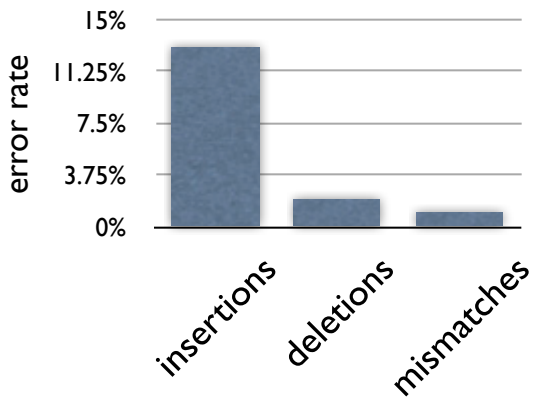
(presentation thursday 1:10 - 2:40pm)

Analyzing PacBio data

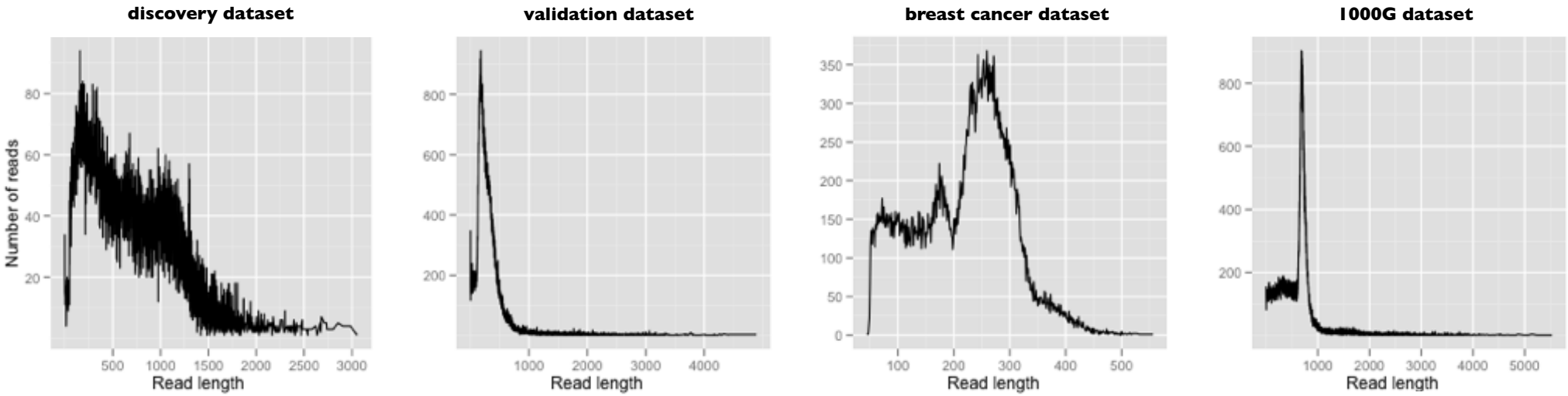
A quick look at Pacific Biosciences data

Notice the SNP

indels are the primary error mode (all purple markers)



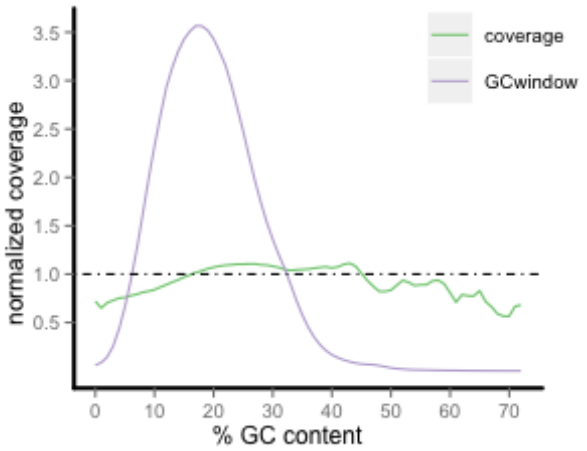
Long reads and deep coverage on all PacBio datasets



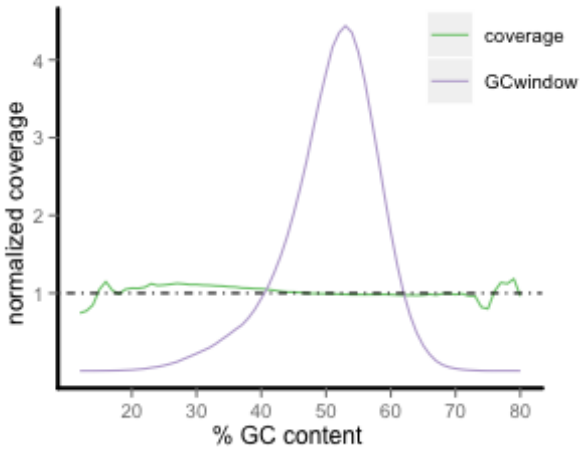
	discovery	validation	cancer	1000G
average coverage	120x	104x	~120x per sample	~500x per sample
number of reads	36,918	305,581	89,934 per sample	256,989 per sample

Sequencing bias is a known problem with NGS technologies that PacBio does not share

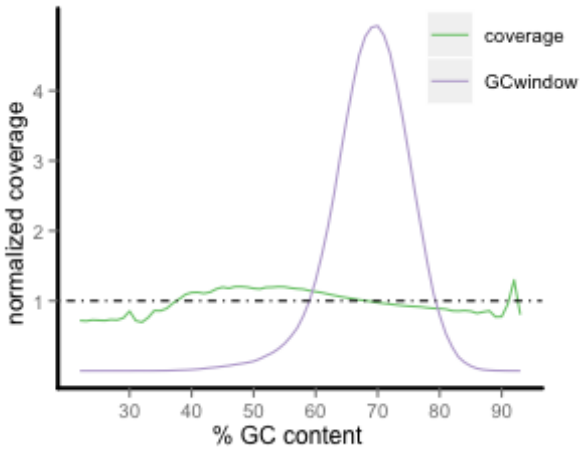
P. falciparum



E. coli



R. sphaeroids



normalized coverage by GC content contrasted with GC content of the genome

come to Michael Ross' talk on tuesday @ 7pm for a more thorough exploration on bias in the different sequencing technologies today

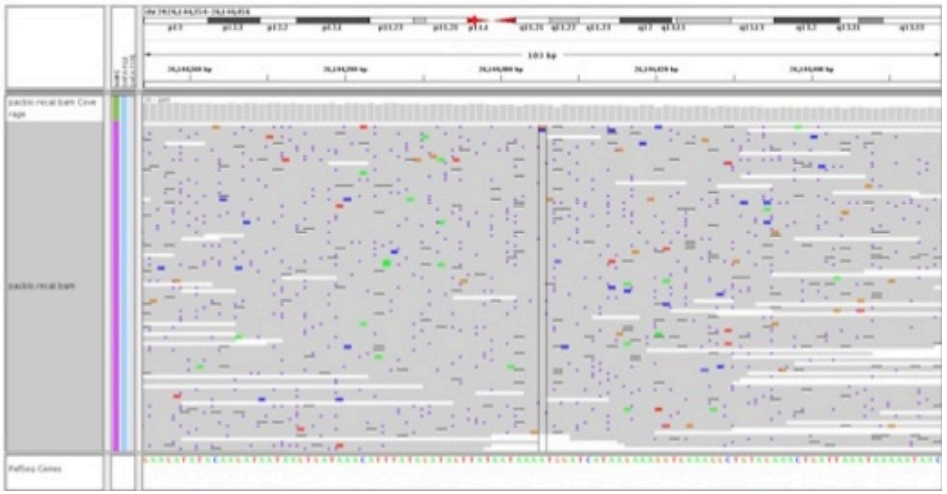
Random error profile of PacBio is much preferred by the GATK bayesian model to systematic errors

HiSeq

SYSTEMATIC ERROR

same genome region on both datasets

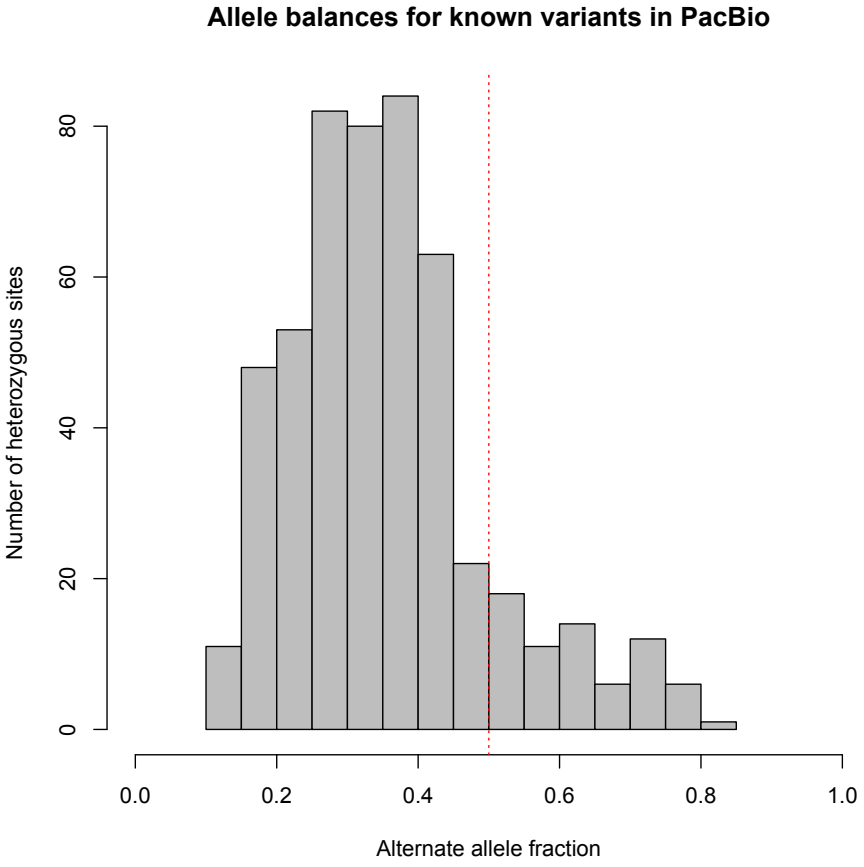
PacBio

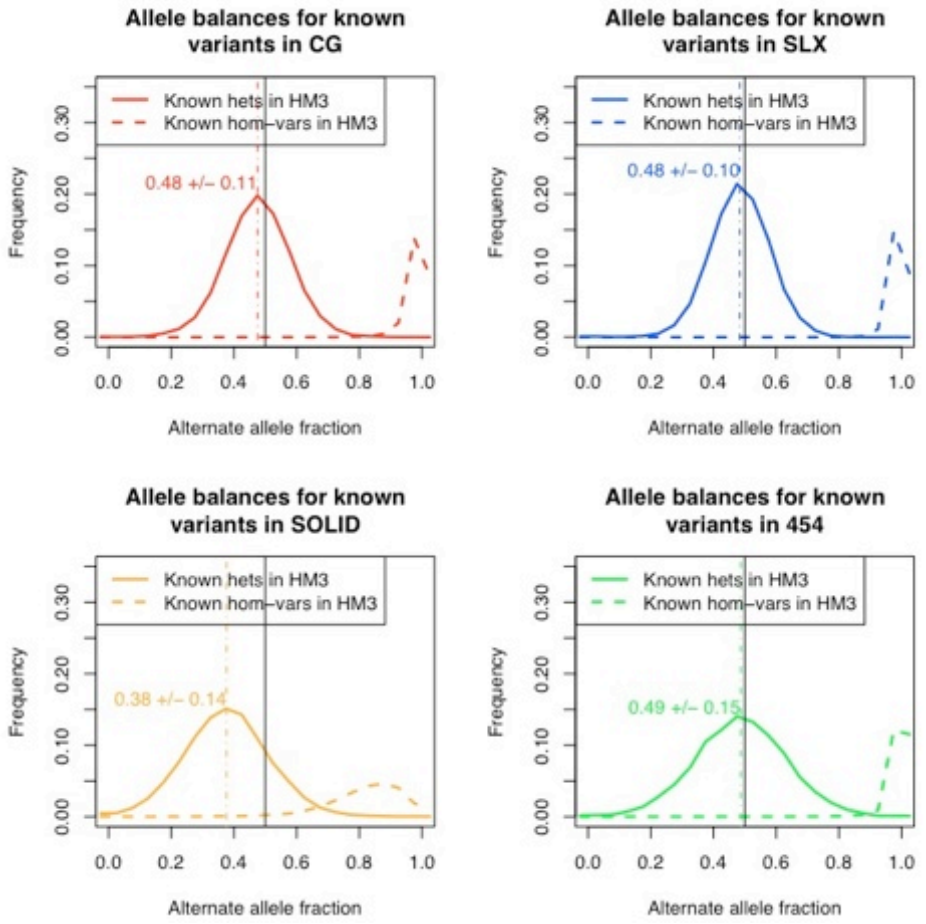


RANDOM ERROR

phasing dataset

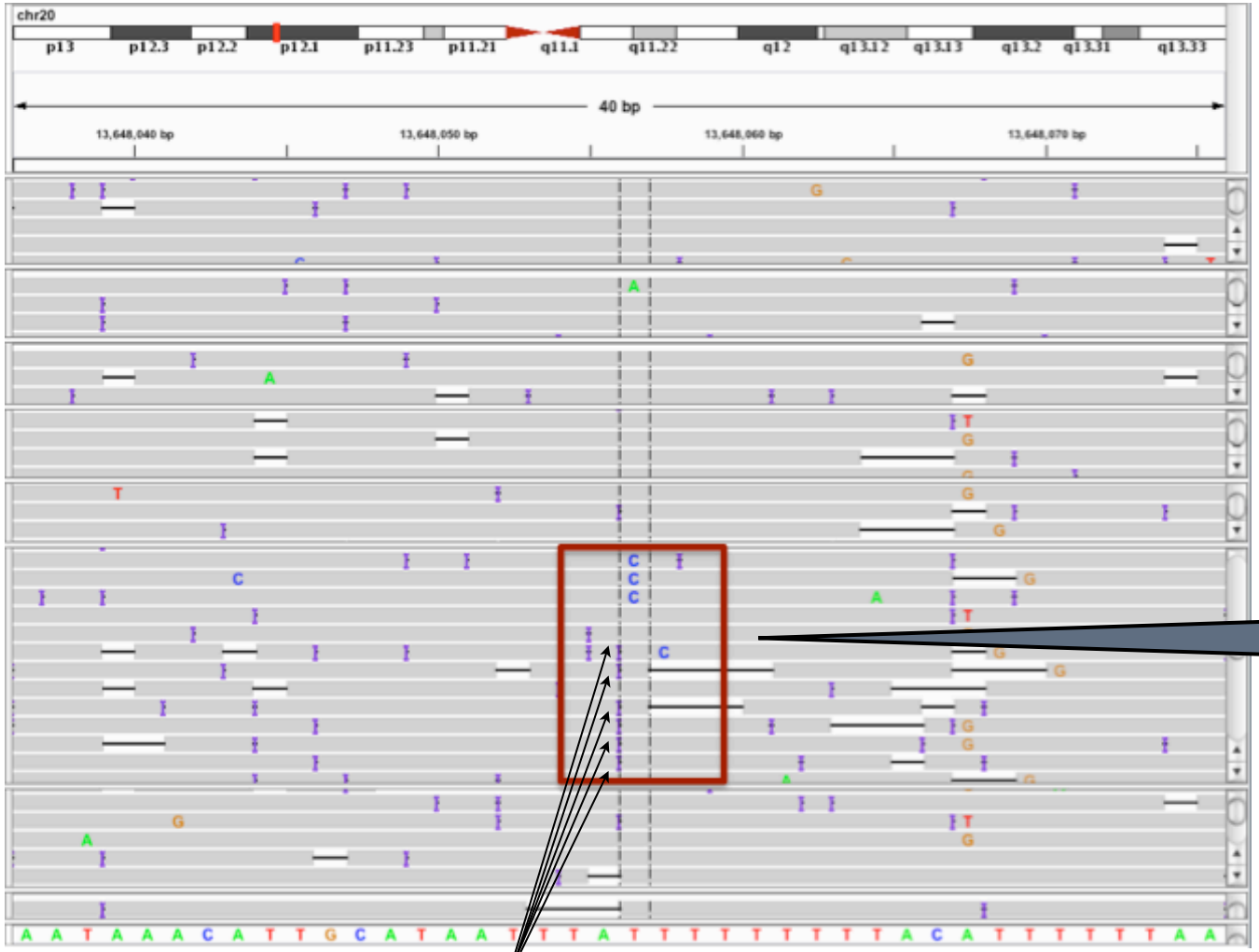
Long reads with a high indel error rate have a side effect: reference bias





Current tools are not capable of locally realigning PacBio data, but we anticipate that newer tools will improve this issue.

True variation missed by Pacbio due to reference bias



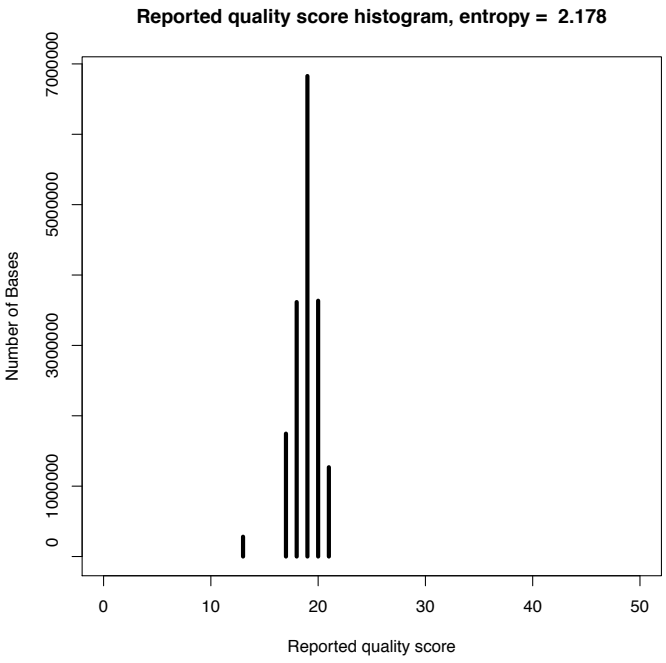
the alternate allele is hiding inside the insertions due to the low gap open penalty of the aligner.

"C" INSERTIONS

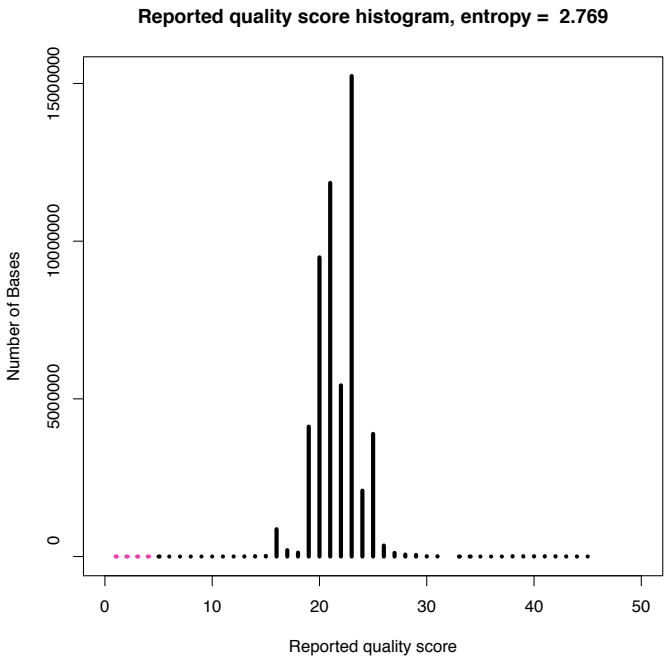
1000G dataset

PacBio produces Q20 bases on average across datasets

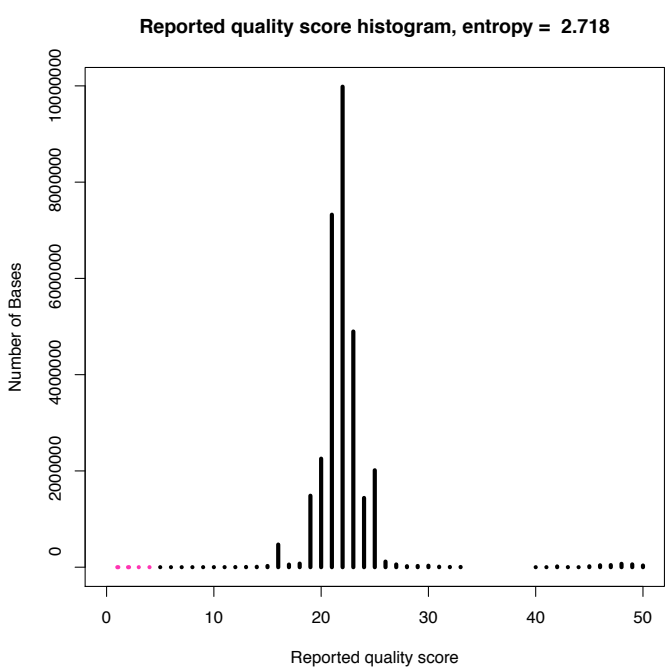
discovery dataset



1000G dataset



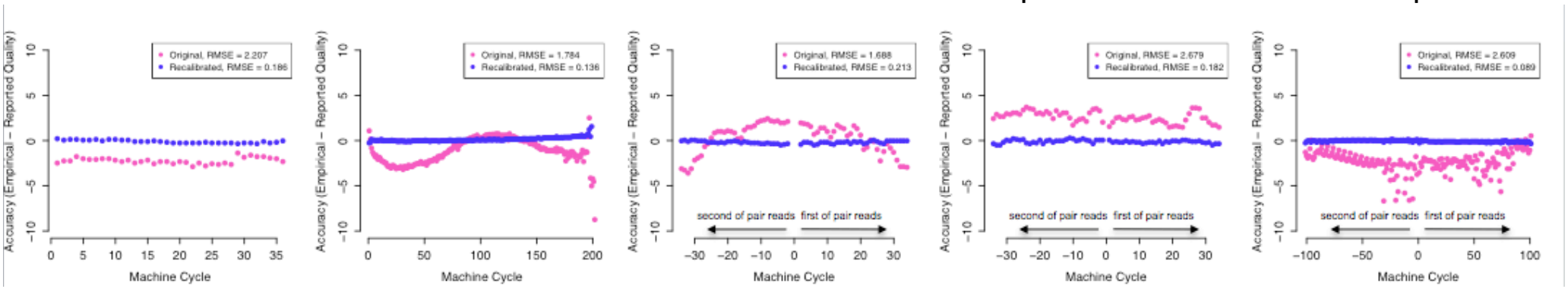
validation dataset



The Base Quality Score framework does not account for indel errors

PacBio base qualities are not affected by the length of the read

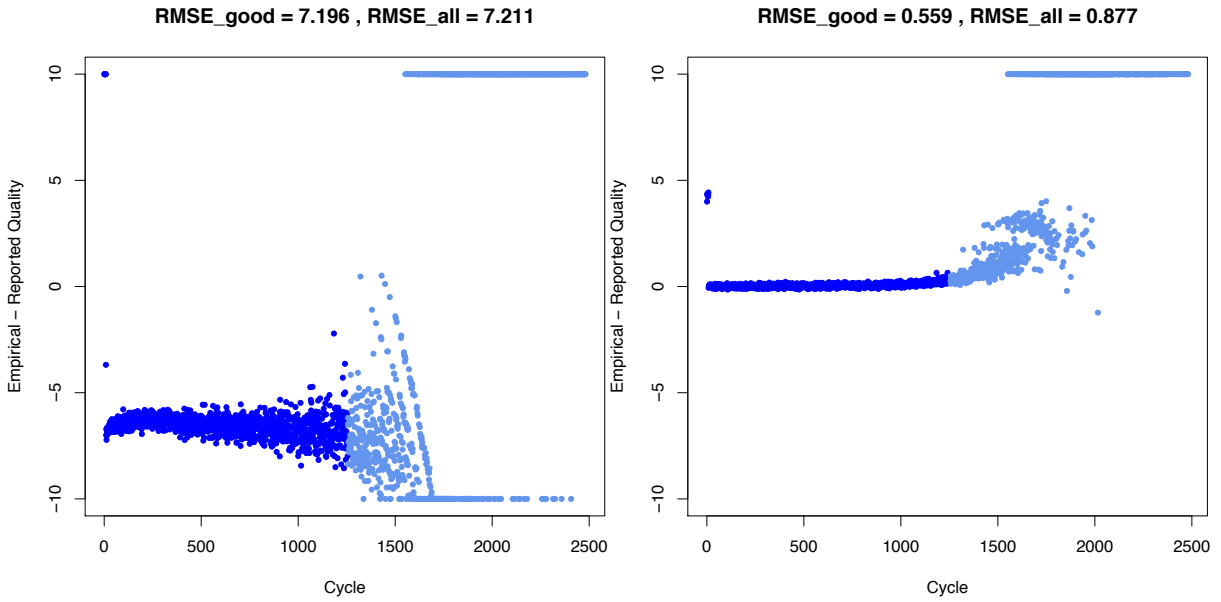
SLX GA | 454 | SOLiD | Complete Genomics | HiSeq



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.

discovery dataset

validating hard-to-call-sites and a look at variation discovery using PacBio

PacBio variation discovery and validation

How can we use PacBio data for human analysis?

- Is PacBio a good platform for follow-up **validation** today?
- Can we do SNP **discovery** with PacBio data?
- How does PacBio compare to **other technologies**?

Data and Definitions

- We have performed a number of experiments at the Broad using PacBio for human data analysis.
  - **discovery dataset** (12/23/2010)
    61 amplicons covering 177 kb from regions across chromosome 20 of NA12878 (1000G sample).
  - **validation dataset** (1/20/2011)
    a set of hard to call NA12878 snps targeted with 2Kbp amplicons
  - **breast cancer dataset** (6/17/2011)
    24 samples for tumor/normal validation analysis of 15 events against HiSeq, 454 and Sequenom.
  - **1000G dataset** (8/25/2011)
    8 samples resequenced at 250 sites for follow up validation against Illumina, Sanger and Sequenom.

Pacbio as a validation tool

- Follow up validation is a major unmet need at the Broad and other centers.
- We carried out a follow-up validation assay using the *de novo* mutations previously validated by the 1000G project.
  - Some are real *de novo* mutations
  - Most are machine artifacts already identified by follow up validation in 1000G.
  - These are hard-to-call sites that are prone to errors and really challenge sequence technology accuracy.

PacBio demonstrates great performance on hard-to-call sites

PacBio	known *true* variant site	known *false* variant site	predictive value
called alt	48	5	91%
called ref	0	67	100%

HiSeq	known *true* variant site	known *false* variant site	predictive value
called alt	48	35	58%
called ref	0	37	100%

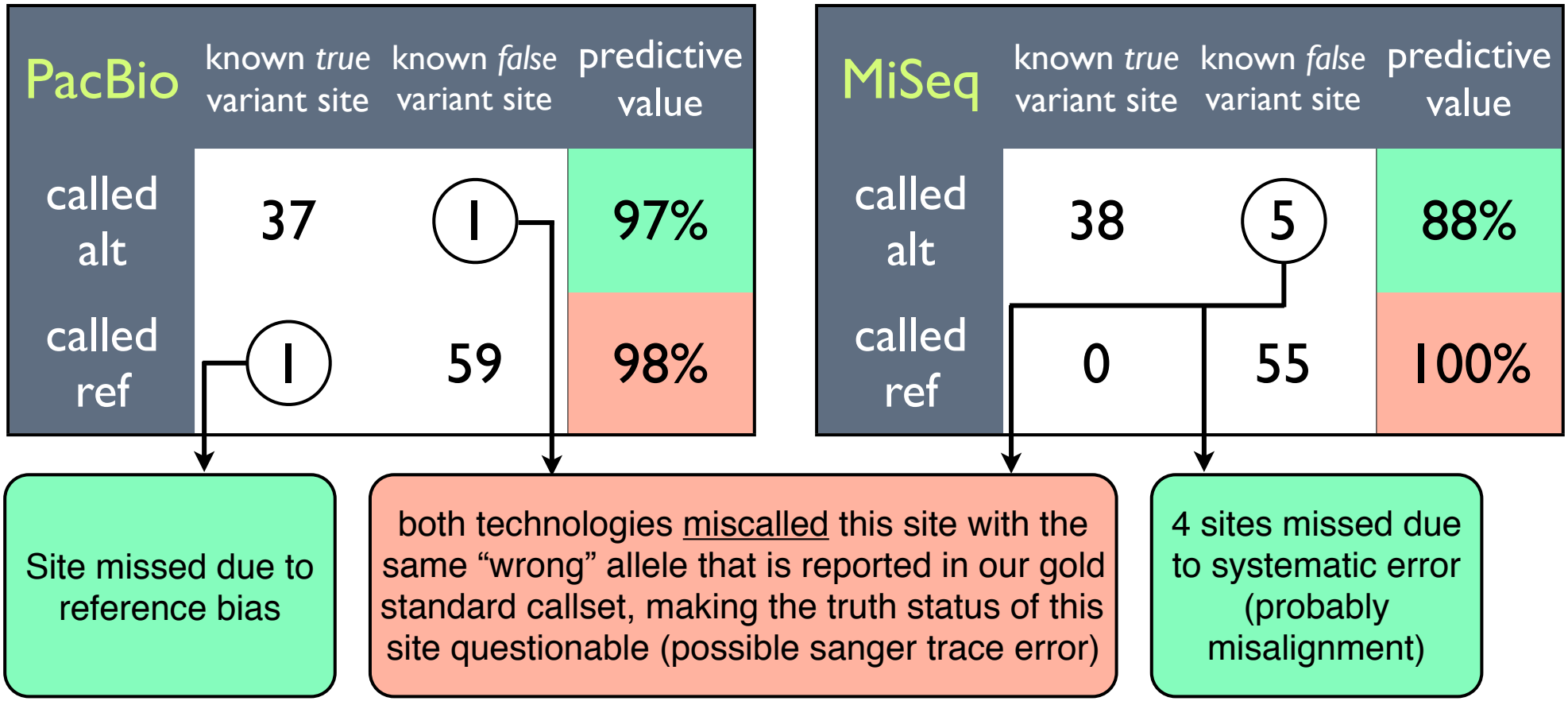
positive predictive value, or **precision rate** is the proportion of subjects with positive test results who are correctly diagnosed

negative predictive value (NPV) is the proportion of subjects with a negative test result who are correctly diagnosed.

same sites on both tests

validation dataset

Pacbio performs well in "apples to apples" comparison with MiSeq data



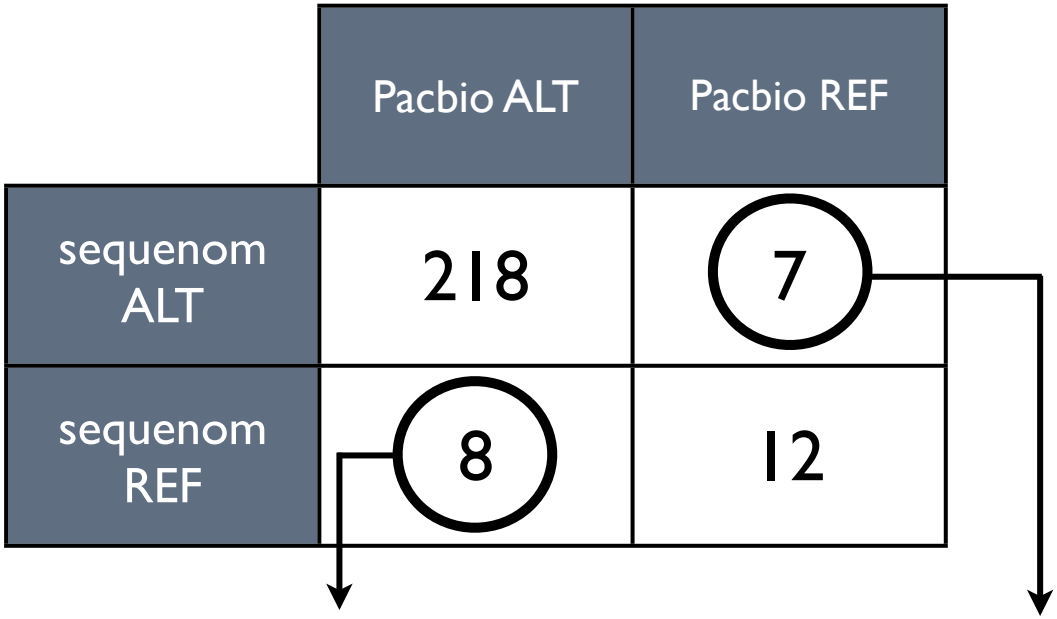
same sites on both tests

validation dataset

1000G project validation experiment

- First we used Sequenom to validate 300 well-behaving SNP sites chosen to be polymorphic in at least 1 out of 8 specific samples from Illumina low pass data.
  - Sequenom is the current standard validation tool at the Broad.
- Sequenom only had data for 250 sites.
- We used PacBio to validate all 300 sites and looked at the agreement between Sequenom and Pacbio.

Pacbio adds valuable information to Sequenom validation



Visual classification	Result from Pacbio
6 look incredibly good	5 ALTs, 1 Reference Bias
1 bad mapping quality	ALT
1 has nearby deletion (unclear)	Reads actually didn't belong at location
50 sites not called by sequenom	Many sites were ALT, others mismapped

Result Pacbio	No. occurrences	what went wrong
good sequencing	1	Sequenom was wrong
Alt allele placed on insertion	4	Pacbio Reference Bias
No coverage	1	Reads actually didn't belong at location
Wrong ALT allele called	1	UG triallelic issue

1000G dataset

Pacific Biosciences, Ion Torrent and MiSeq have good potential for validation experiments

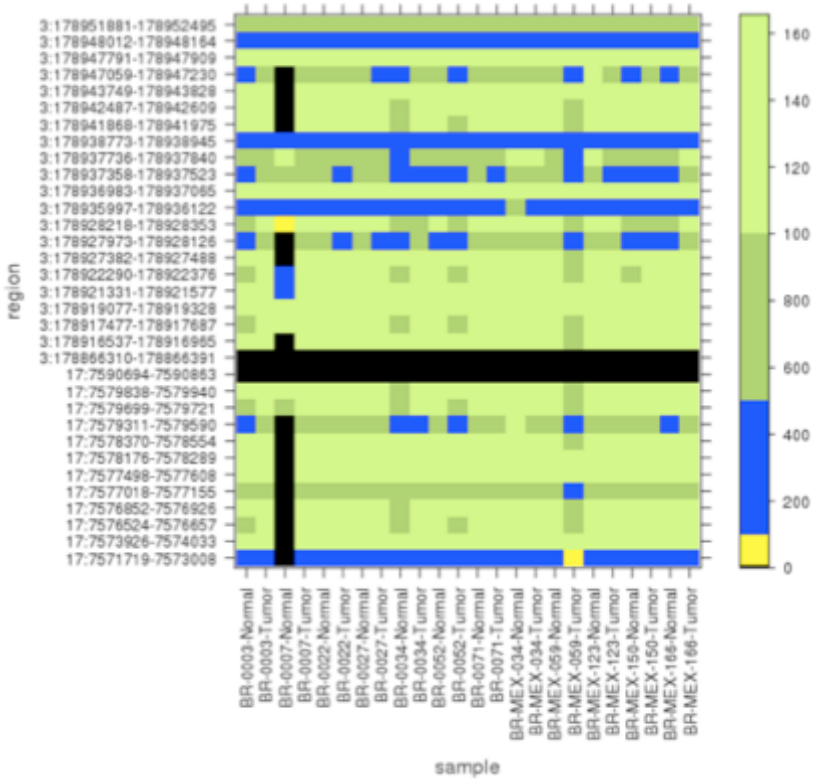
Ion Torrent has a low NPV but is good in most other metrics. NPV

	sensitivity	specificity	PPV	NPV
Ion (bwa-sw)	96.2%	100%	100%	54.5%
Ion (tmap)	96.2%	100%	100%	54.5%
MiSeq	98.1%	92.3%	99.6%	70.5%
PacBio	98.1%	100%	100%	68.7%

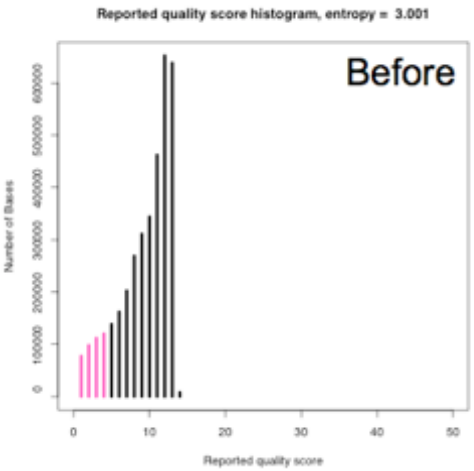
Low specificity indicates artifactual calls outside the scope of the validation

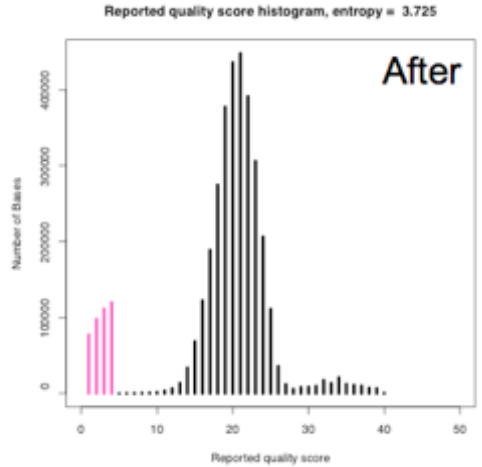
breast cancer validation experiment

high coverage and high specificity to targets



base qualities are severely under calculated





	Illumina	Sequenom	Pacbio	454
somatic	15	6	12	8
wildtype	0	6	1	0
unknown	0	3	2	7

Pacbio correctly identified a false positive in the original dataset
(unknown in sequenom and 454)

cancer dataset

GATK performs very well for SNP discovery with PacBio data

	discovery		
	MiSeq	HiSeq	PacBio
Gold Standard SNP calls	222	225	197
calls on HapMap	43	43	38
Sensitivity	99.1%	100.0%	87.6%

- Reference bias (17) and lack of coverage (11) were the reasons for missed sites in Pacbio
- MiSeq missing data are due to mismapping/artifact (2) or low coverage (1).

Broad's somatic mutation caller (muTect) successfully calls pacbio data

- One tumor/normal pair called:
  - 6,459 sites examined
  - 4,837 sites covered (14x/8x)
  - 1 true somatic mutation called
    (previously validated)
  - 0 False Positives called

muTect is a GATK based caller developed by the cancer group at the Broad Institute
(https://confluence.broadinstitute.org/display/CGATools/MuTect)

PacBio data performs well with the GATK because...

- The error rate is random (despite being high).
  - Such non-systematic error mode is well handled by the GATK SNP calling mathematics.
- very long reads make mapping very clear.
  - less mismappings of paralogous sequences.
  - structural variants are less prone to appear as SNPs.

Pacbio's reference bias is currently the major limiting factor

What is the GSA team working on right now

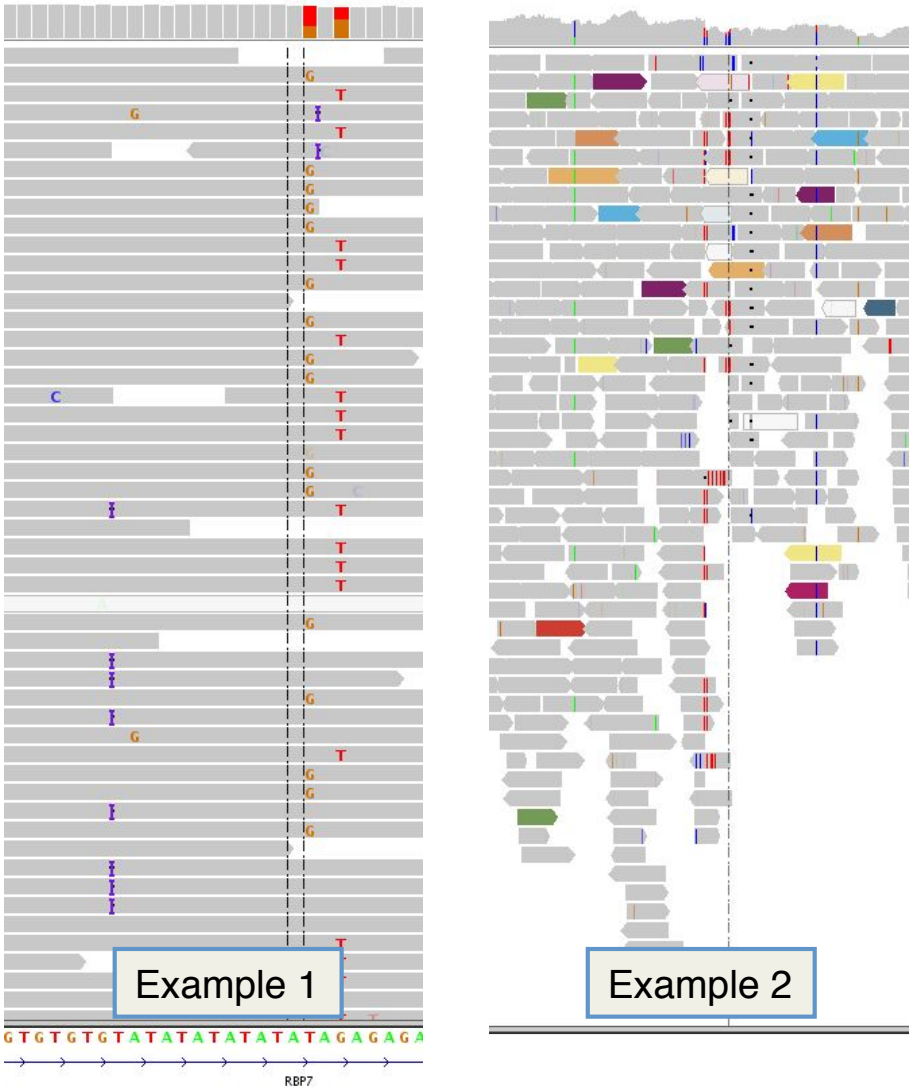
(that will impact PacBio data analysis)

Future of the GATK

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

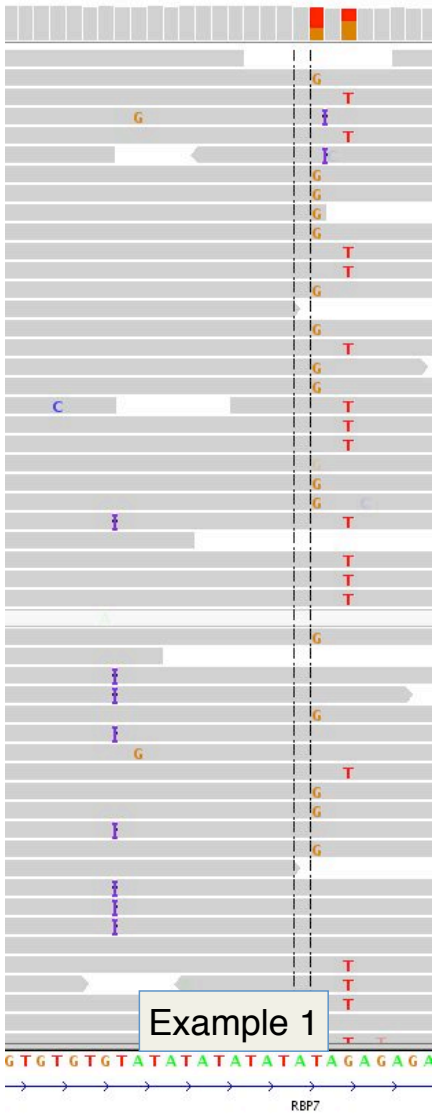
Too systematic to be machine errors, but the haplotype for Pr{D|H} is unclear

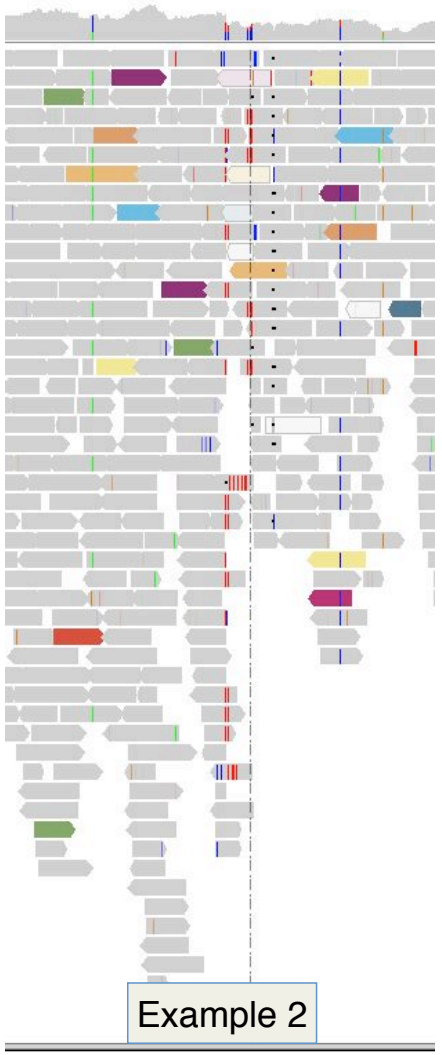


From reads to alleles: the next frontier

- Can't calculate a likelihood for a hypothesis you don't consider
- How do I know what genetic variant I'm looking at, given each locus independently?
  - A SNP, an INDEL, an SV, or something else?
- General problem, but acute for medium-sized events as we not only miss the true event but also generate many smaller false events
- Reference bias can be addressed from a haplotype approach

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



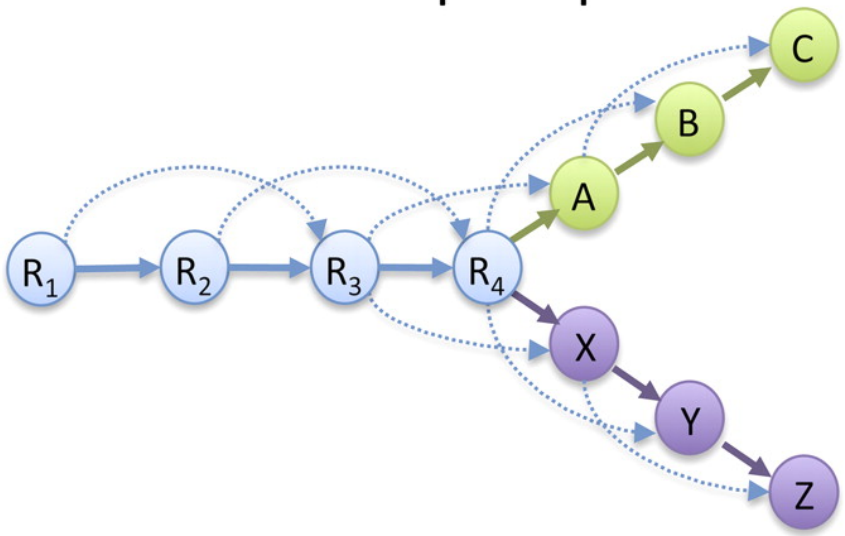


Using local de novo haplotype assembly via DeBruijn graphs

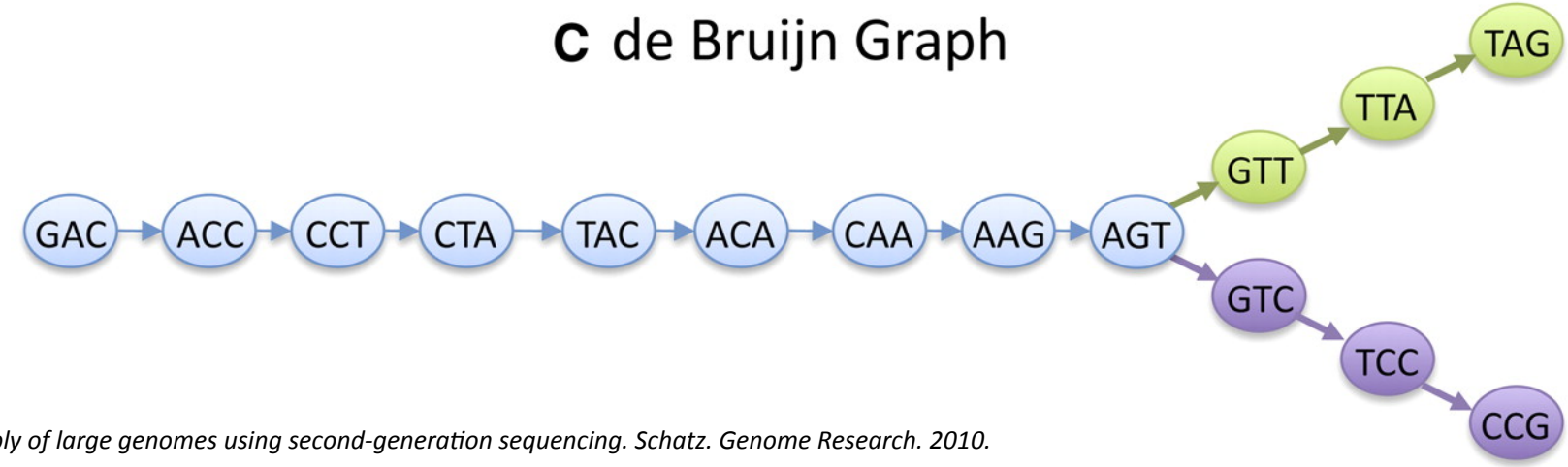
A Read Layout

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

B Overlap Graph

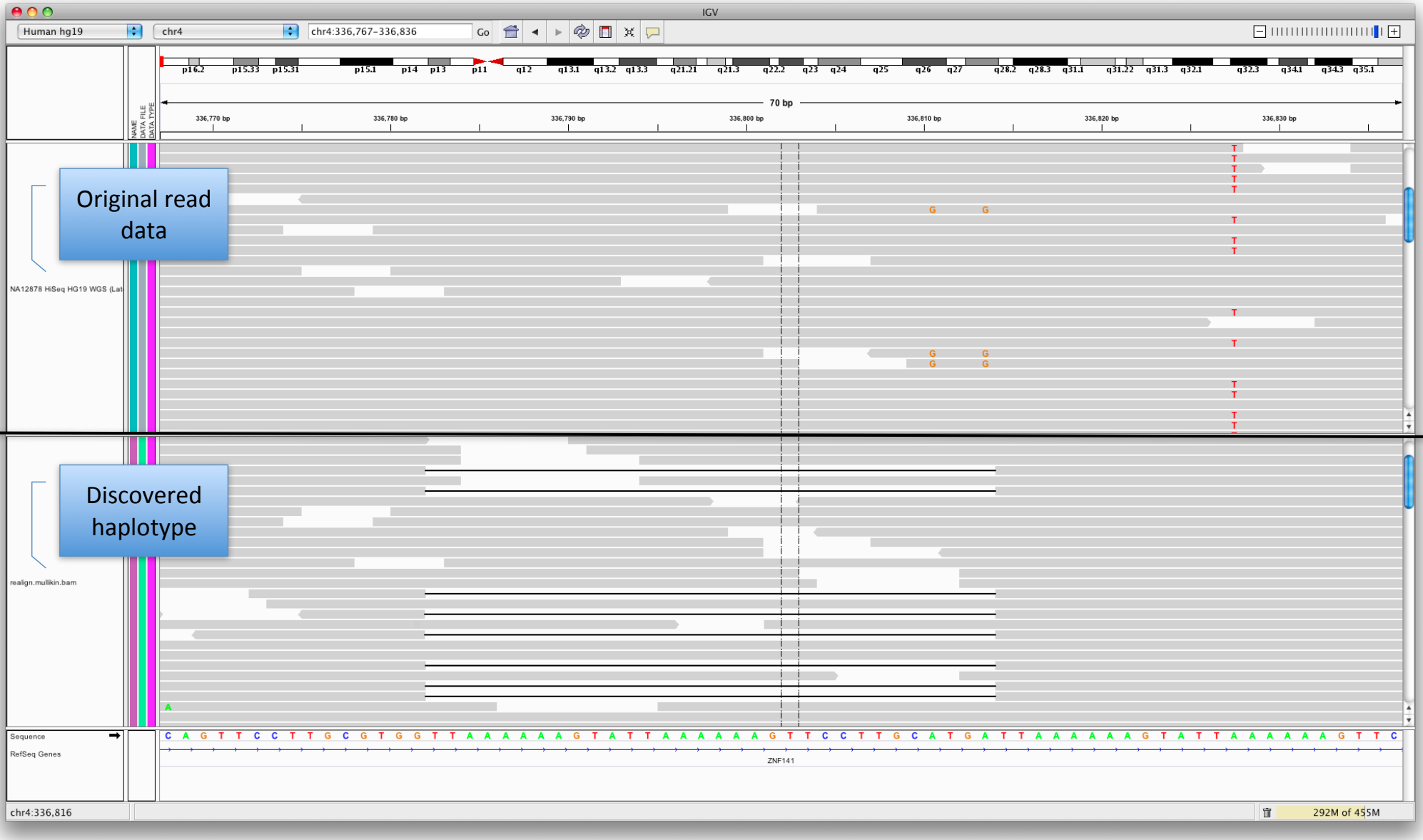


C de Bruijn Graph

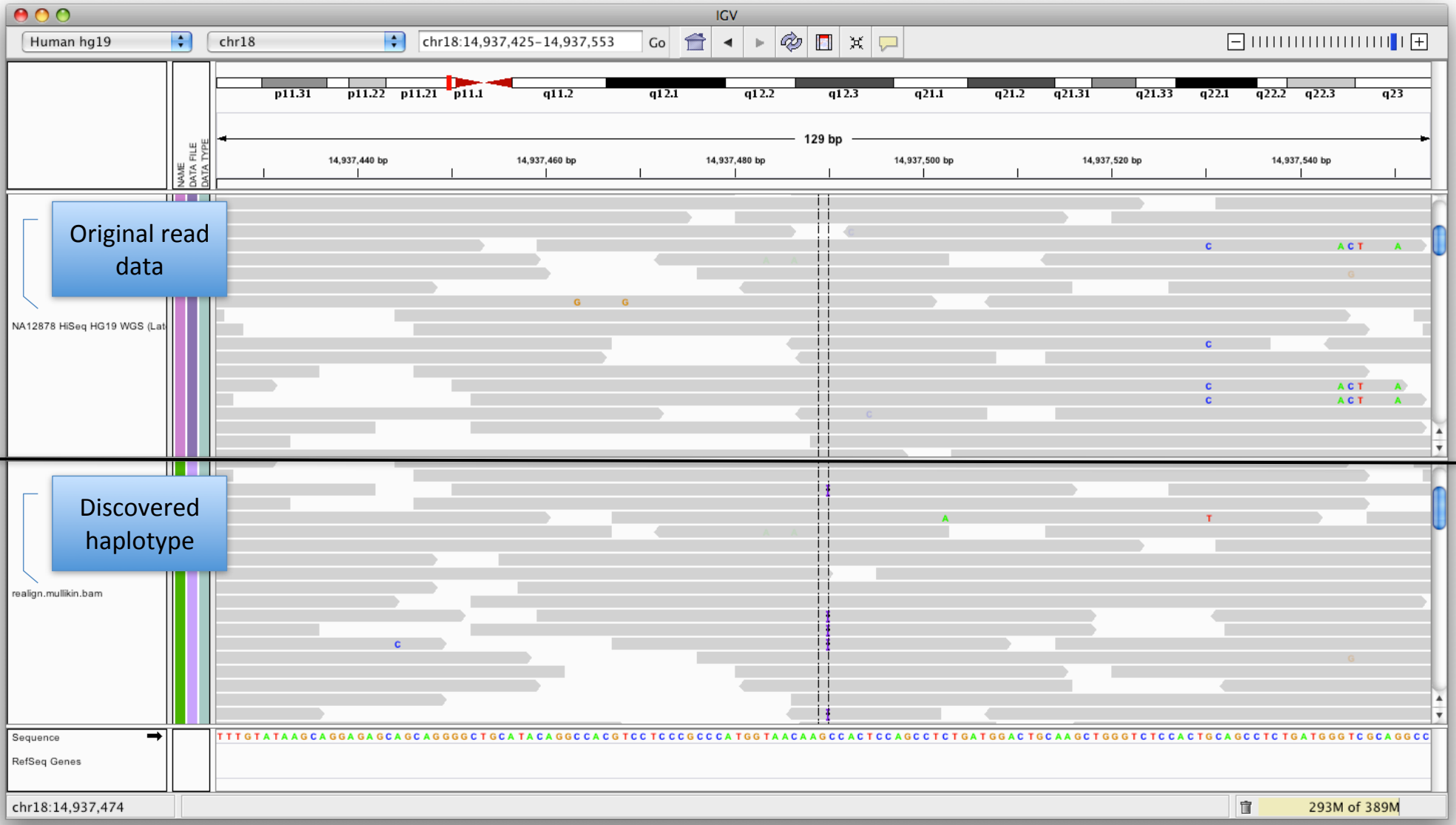


29 *Assembly of large genomes using second-generation sequencing. Schatz. Genome Research. 2010.*

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



Example Mullikin het insertion we now call chr18:14937489 -/CCACTCCAGCCTCTGATGGACTGCAAGCTGGGTCT

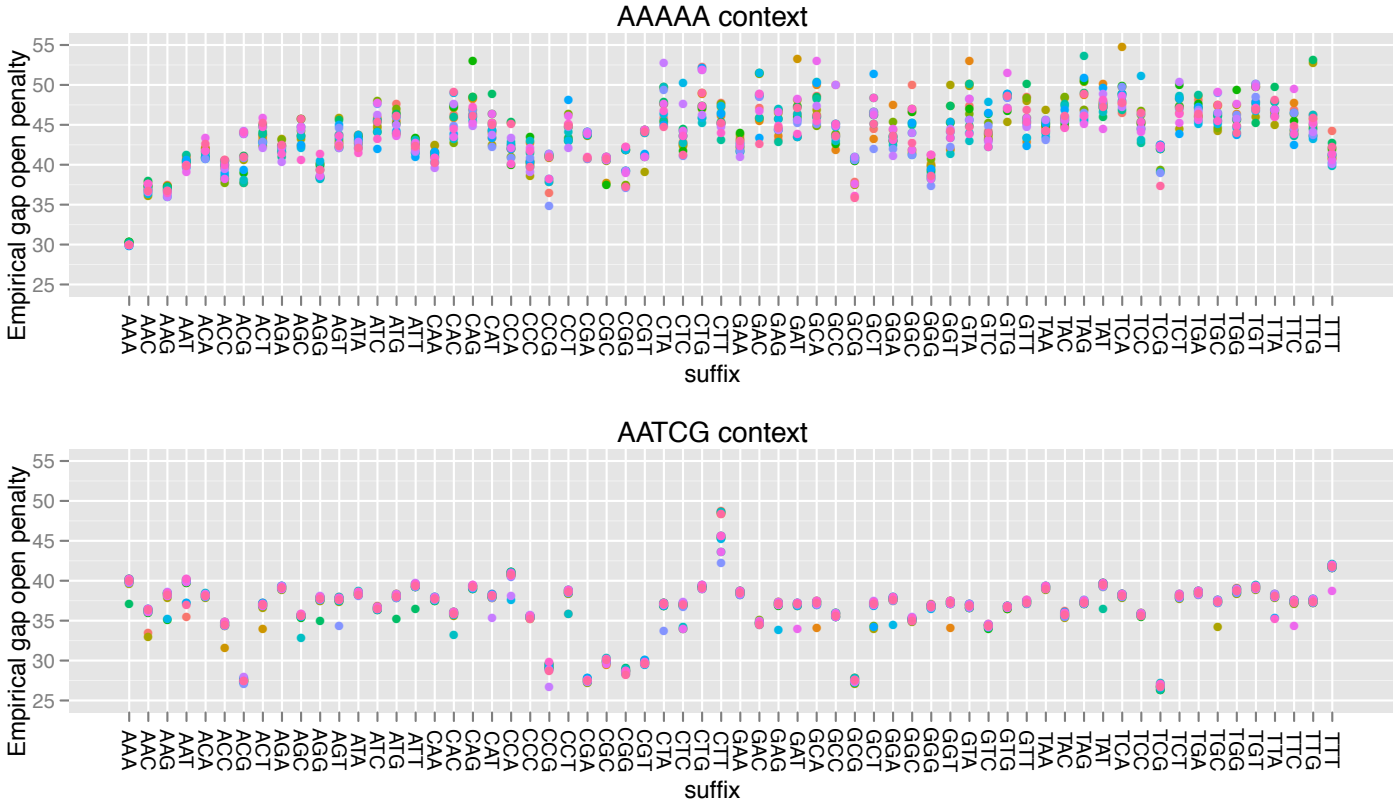


Haplotype Caller greatly increases sensitivity to larger indel events over the Unified Genotyper

	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	90.9% (70 / 77)	89.6% (69 / 77)	81.8% (162 / 198)	81.8% (162 / 198)

- Input data is NA12878 b37+decoy WGS HiSeq high coverage
- 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 don't yet call)

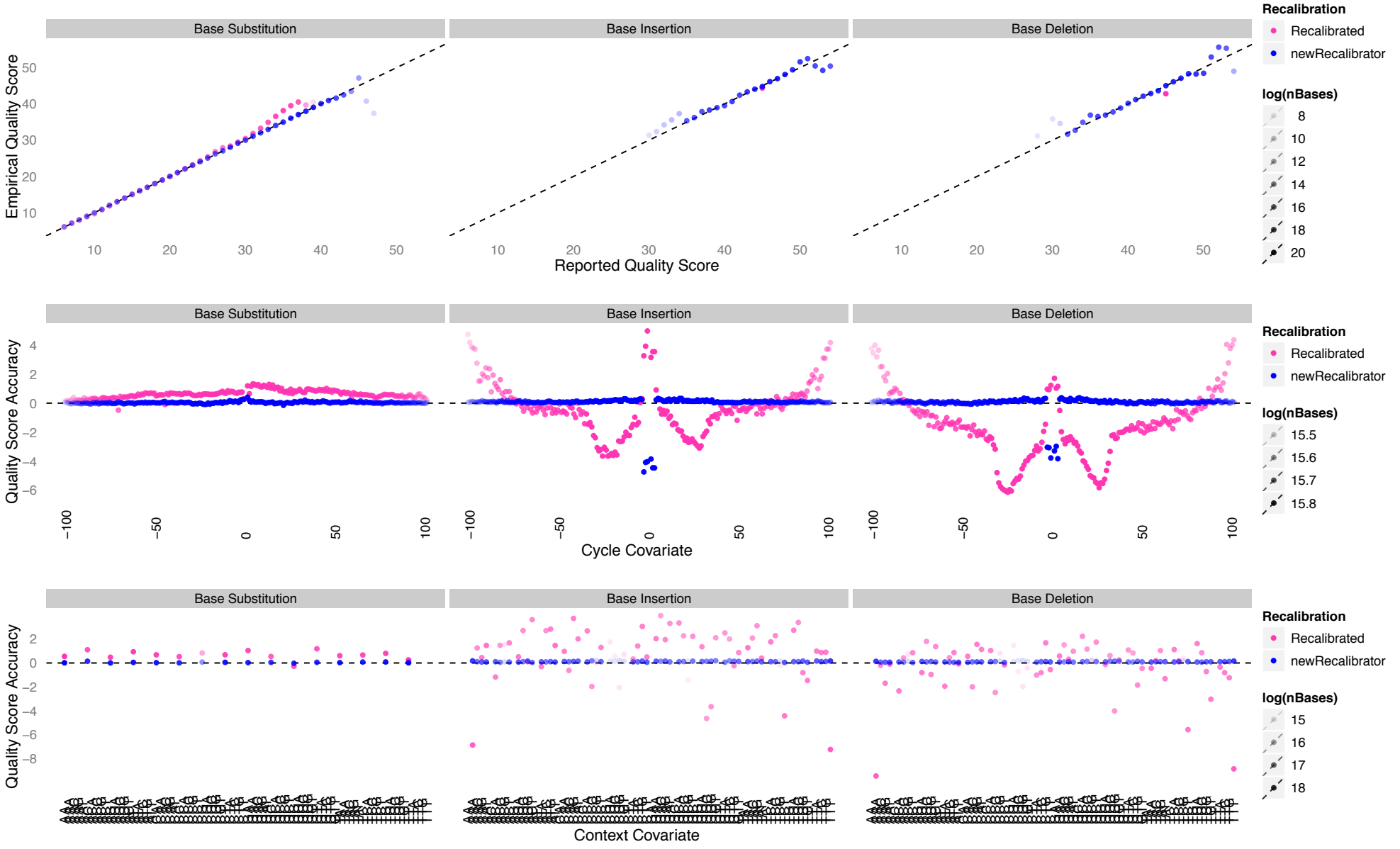
A new BQSR that also recalibrates "indel qualities" qualities

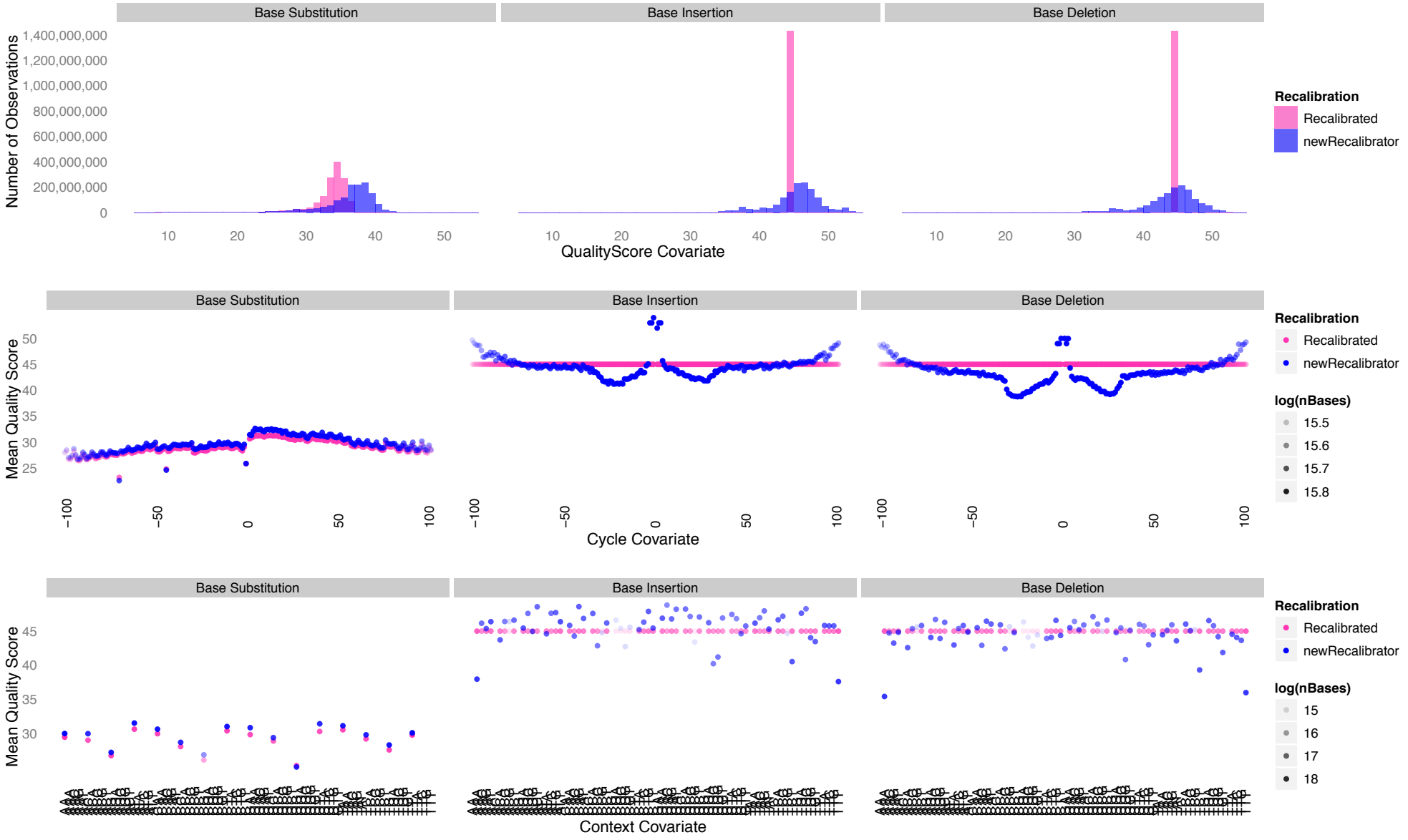


other improvements

- "auto-recalibration" mode for organisms without known callsets
- improved covariate models
- simpler command line pipeline with a single tool instead of three.

there is significant difference in the empirical probability of starting an insertion or deletion due to context


Base Substitution
Base Insertion
Base Deletion
Empirical Quality Score
Reported Quality Score
Quality Score Accuracy
Cycle Covariate
Context Covariate
Recalibration
Recalibrated
newRecalibrator
log(nBases)


Base Substitution
Base Insertion
Base Deletion
Number of Observations
1,400,000,000
1,200,000,000
1,000,000,000
800,000,000
600,000,000
400,000,000
200,000,000
0
10
20
30
40
50
QualityScore Covariate
Recalibration
Recalibrated
newRecalibrator
Mean Quality Score
50
45
40
35
30
25
−100
−50
0
50
100
Cycle Covariate
log(nBases)
15.5
15.6
15.7
15.8
15
16
17
18
Context Covariate

Thank you!

Stay up to date with the GSA team through our wiki

- the **latest releases** of our tools and version changelogs
- **tutorials** on our best practices for data processing and analysis
- further information on how to use the GATK engine for your own research or to **collaborate** with us

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