Universitair Medisch Centrum Utrecht

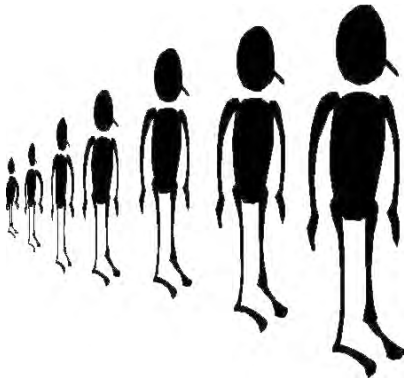
Mapping sequence reads & Calling variants

Laurent Francioli

2014-10-28

l.francioli@umcutrecht.nl

Next Generation Sequencing

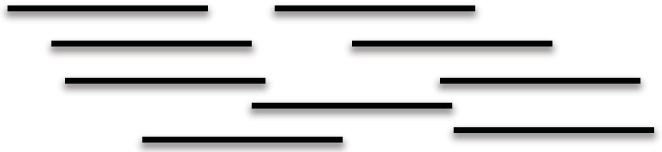


DNA

Sequencing

Mapping

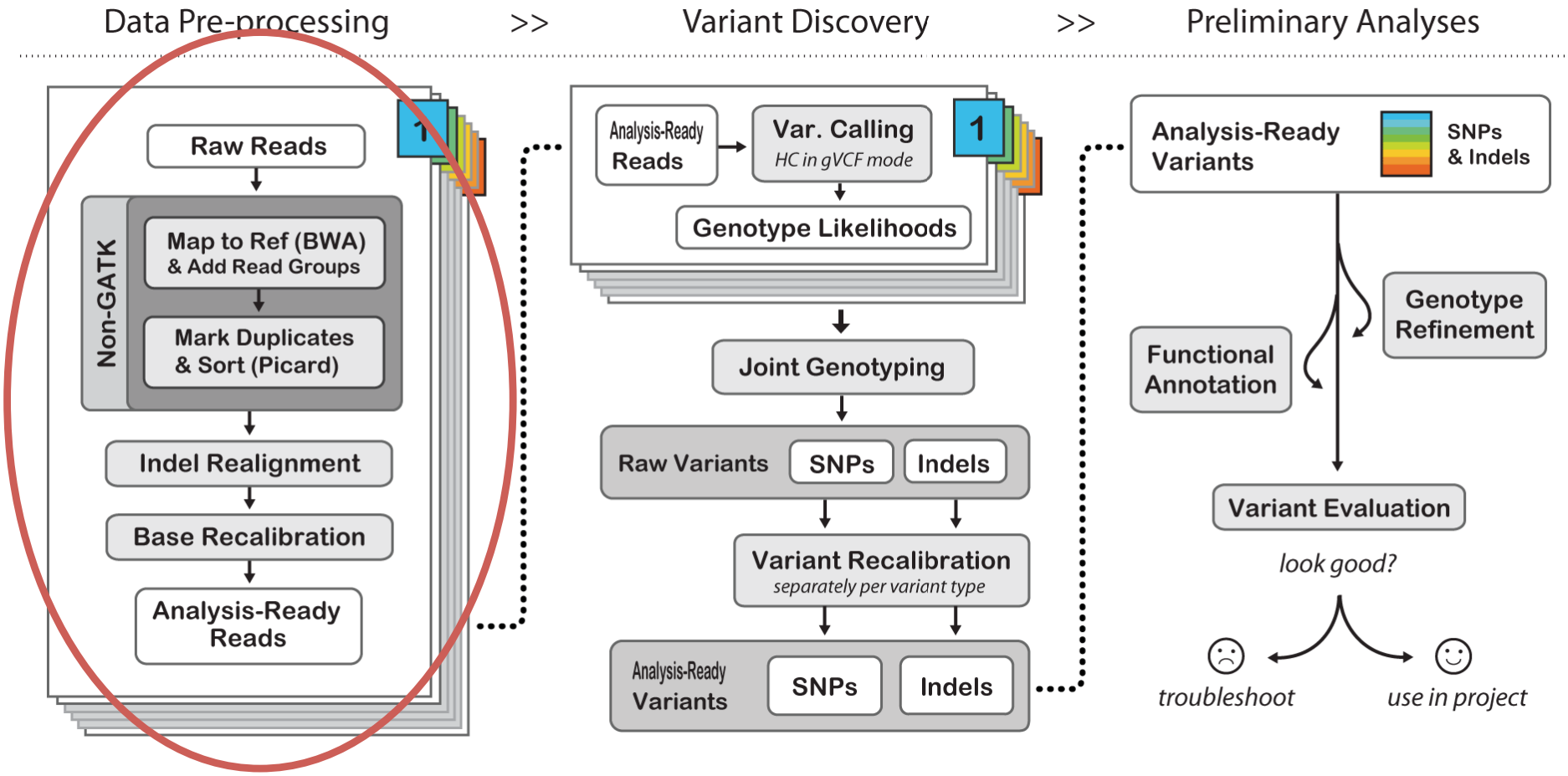
Human Reference Sequence



Variant Calling

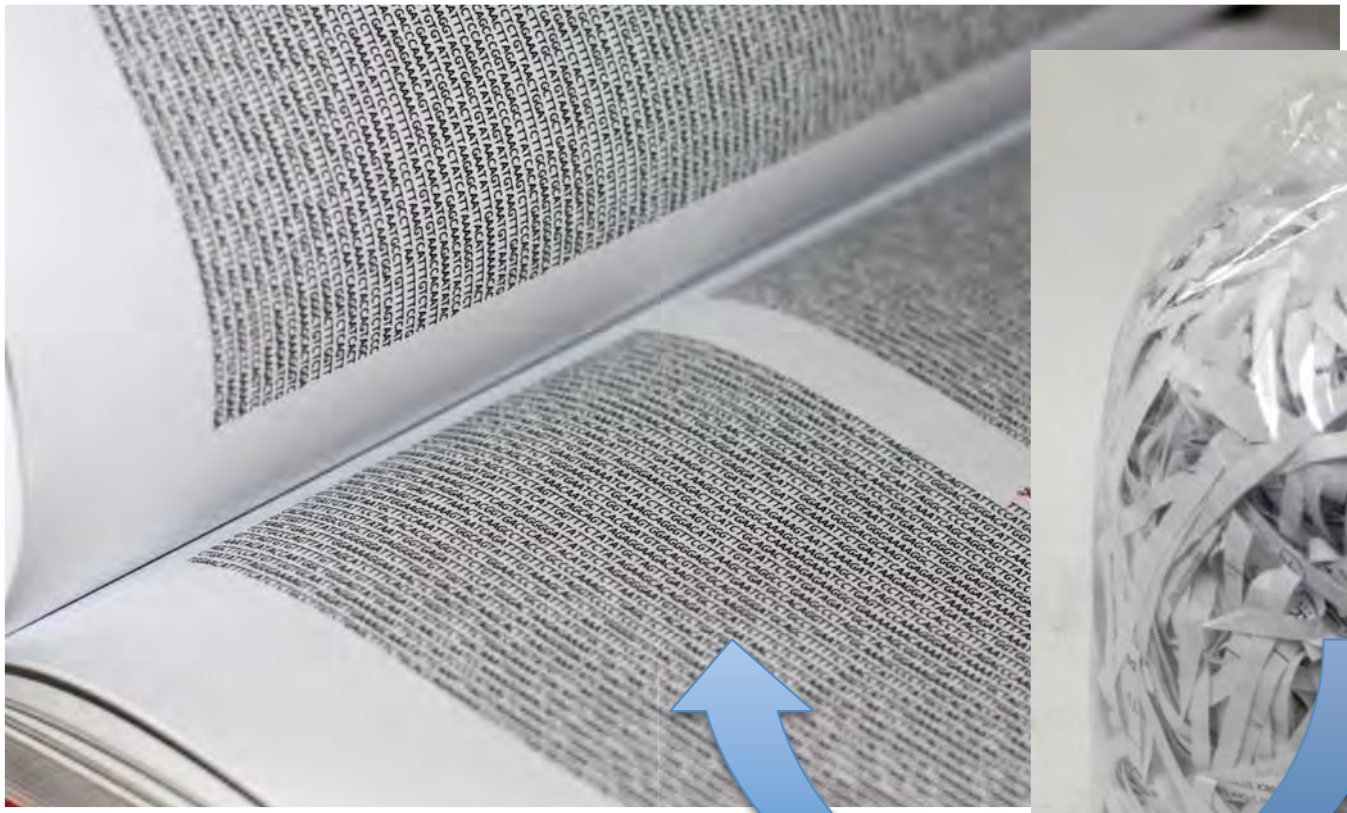
Variant	Sample 1	Sample 2
rs12625542	T/C	T/T
rs61087198	G/G	T/T

Data processing pipeline



From the Genome Analysis Toolkit: *http://www.broadinstitute.org/gatk/guide/best-practices*

Alignment



Mapping to reference

- Burrow-Wheeler Algorithm (BWA)
- Reads are scored for possible alignments against reference data
- Mapping scores are computed using mismatches and gap penalties

Human Reference Sequence

G T G C C A G G A C C A G A T C G T G C C A A C G G A C A G G T G G T A A G G A A G G A G

Sequence reads

G T G C C A A G G A

C A A C G G A C A G

A G G T G G T A A G

…

Mapping to reference

- Burrow-Wheeler Algorithm (BWA)
- Reads are scored for possible alignments against reference data
- Mapping scores are computed using mismatches and gap penalties

Human Reference Sequence

G T G C C A G G A C C A G A T C G T G C C A A C G G A C A G G T G G T A A G G A A G G A G

G T G C C A A G G A

G T G C C A A G G A

Mapping to reference

- Burrow-Wheeler Algorithm (BWA)
- Reads are scored for possible alignments against reference data
- Mapping scores are computed using mismatches and gap penalties

Human Reference Sequence

G T G C C A - G G A C C A G A T C G T G C C A A C G G A C A G G T G G T A A G G A A G G A G

G T G C C A - A G G A

G T G C C A A G G A

G T G C C A A - G G A

Mapping to reference

- Burrow-Wheeler Algorithm (BWA)
- Reads are scored for possible alignments against reference data
- Mapping scores are computed using mismatches and gap penalties

Human Reference Sequence

…G T G C C A G G A C C A G A T C G T G C C A A C G G A C A G G T G G T A A G G A A G G A…

3.2Gb

Millions of sequence reads

…G T G C C A A G G A…

…C A A C G G A C A G…

…A G G T G G T A A G…

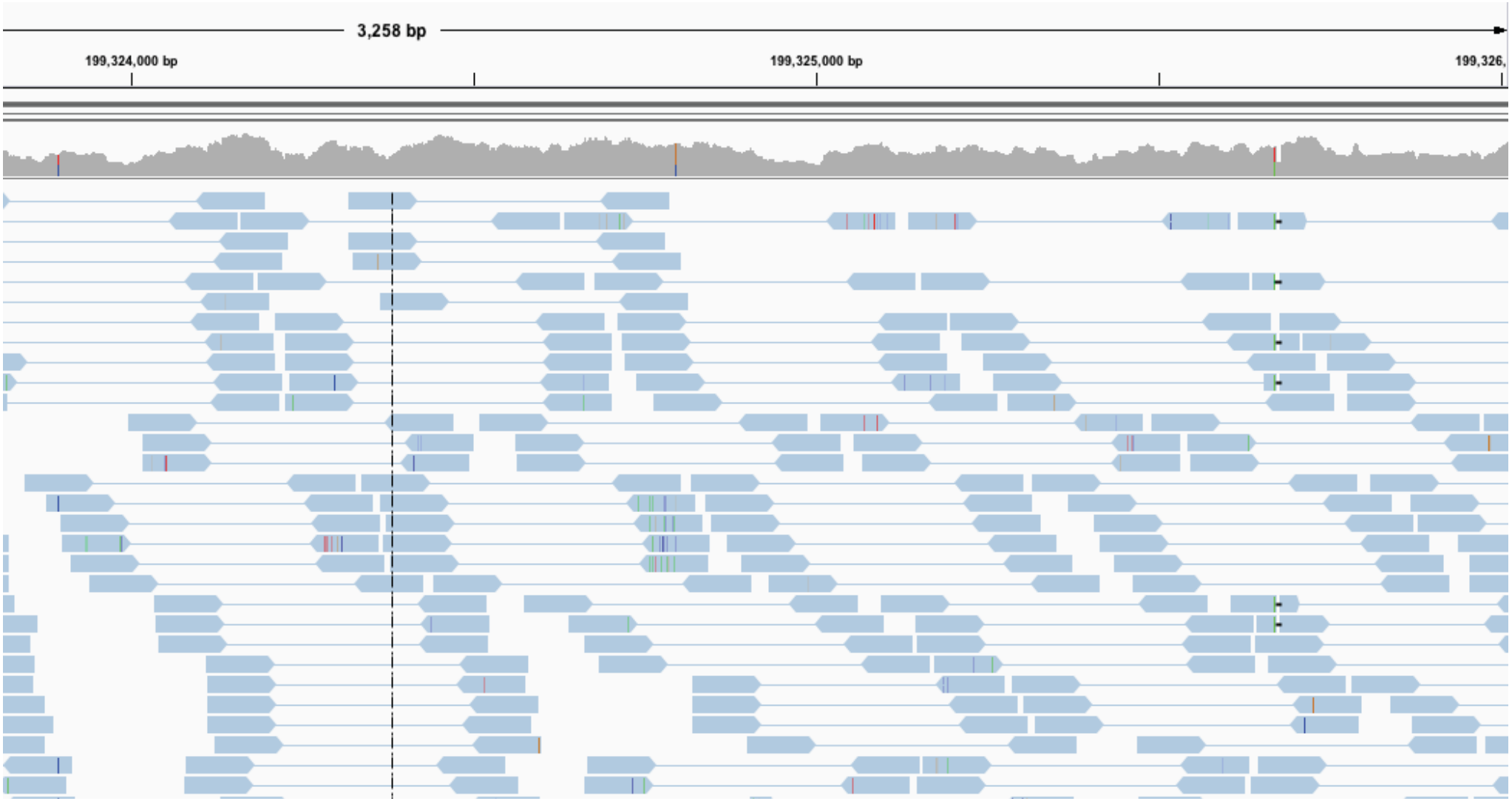
…

50 – 120bp

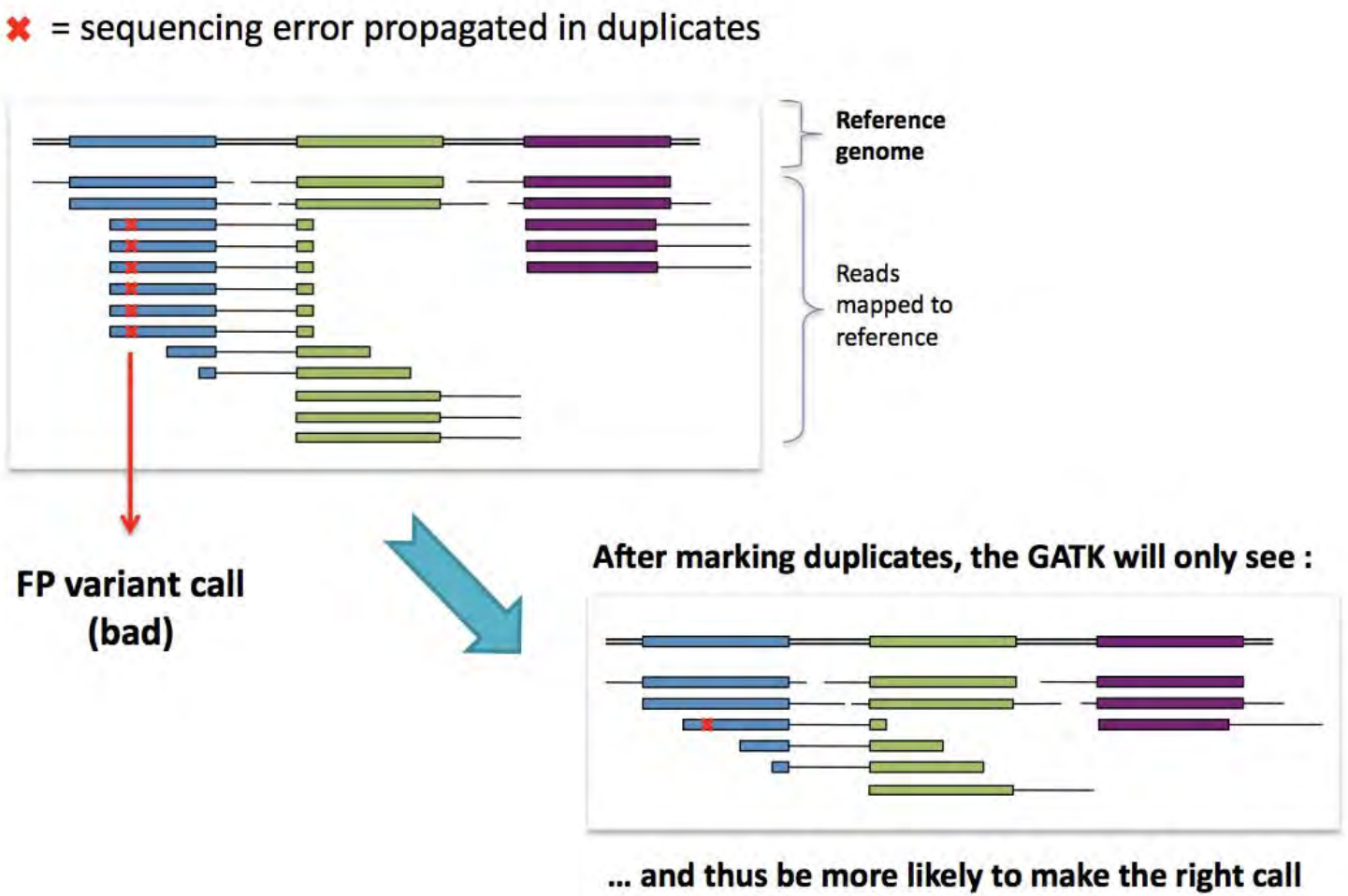
Coverage

- Number of reads covering a given base of the genome
- Overall coverage is a function of the total number of sequence reads
  - Target coverage is decided at experiment design

Real NGS aligned data in IGV



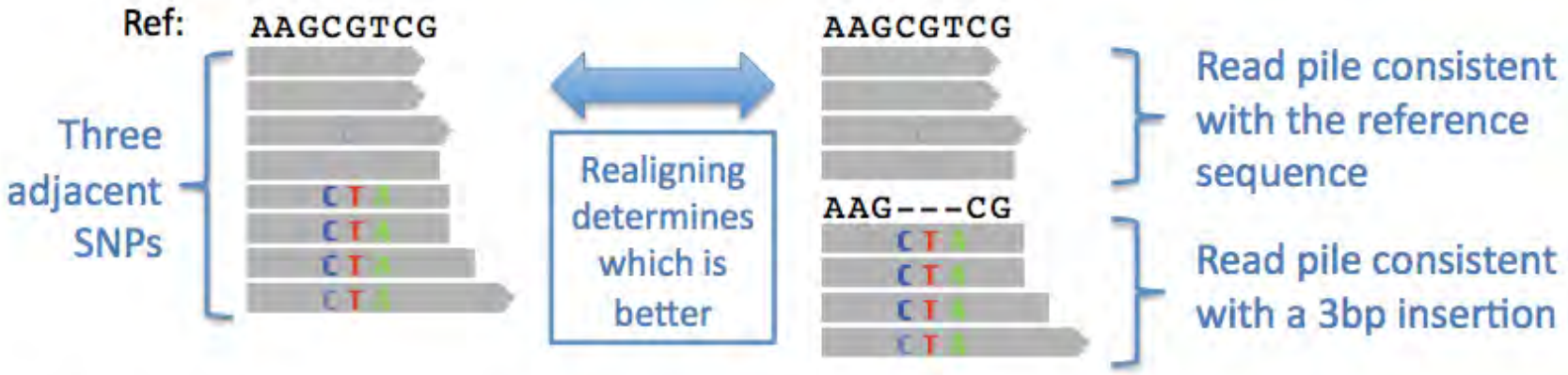
Marking Duplicates



http://www.broadinstitute.org/gatk/guide/events

Indel realignment

1. Find the best alternate consensus sequence that, together with the reference, best fits the reads in a pile (maximum of 1 indel)



2. The score for an alternate consensus is the total sum of the quality scores of mismatching bases

3. If the score of the best alternate consensus is sufficiently better than the original alignments (using a LOD score), then we accept the proposed realignment of the reads

http://www.broadinstitute.org/gatk/guide/events

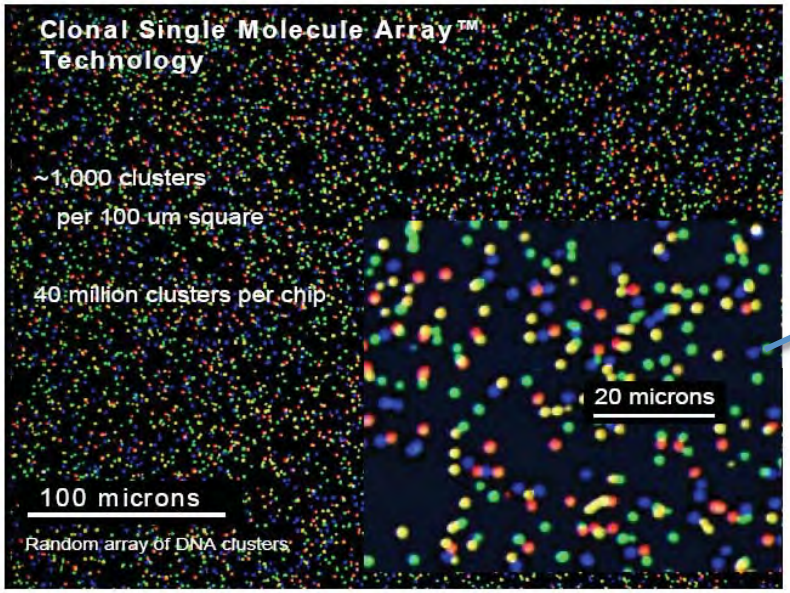
Indel realignment - example

http://www.broadinstitute.org/gatk/guide/events

Indel realignment - example

http://www.broadinstitute.org/gatk/guide/events

Base qualities



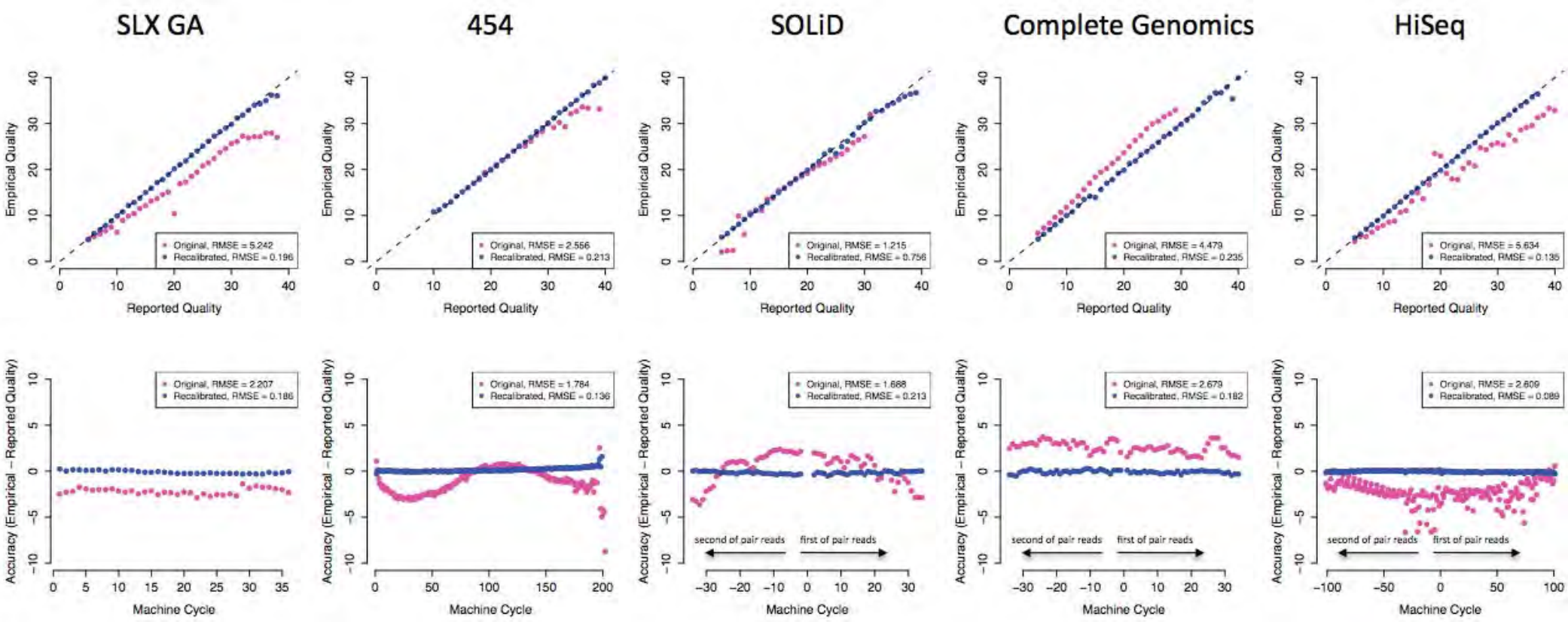
ACTGCCAGGTNTCAGTACA

Phred value = -10 Log10(E)

Quality 10 = 10% error (90% confidence)
Quality 20 = 1% error (99% confidence)
Quality 30 = 0.1% error (99.9% confidence)

Base quality recalibration

- Downstream models use base qualities as input
- Base qualities have systematic machine-specific biases

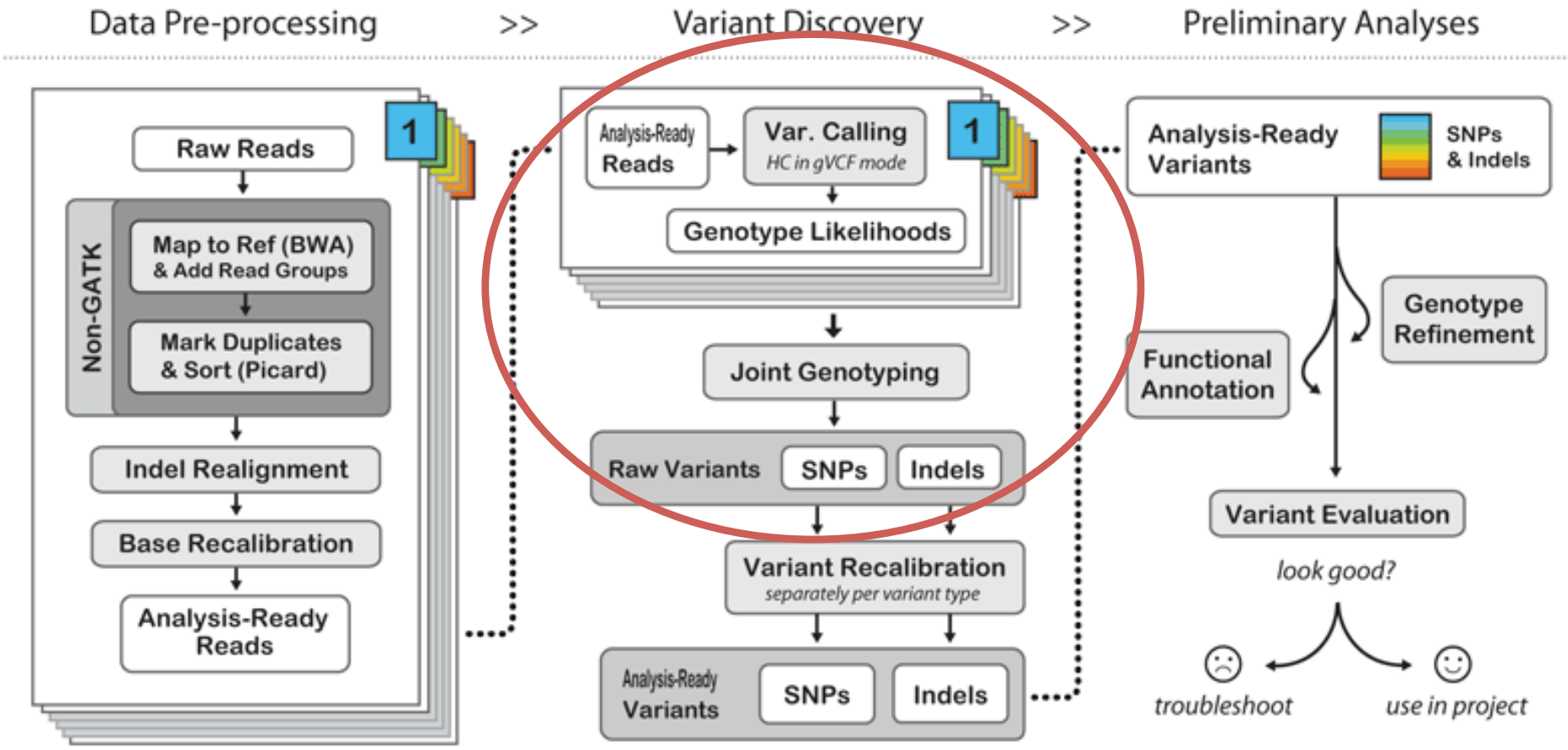


http://www.broadinstitute.org/gatk/guide/best-practices

Analysis-ready reads

- Reads are correctly placed in the genome
  - Mapping qualities model the uncertainty of mapping
- Reads are independent
  - Clonal (duplicate) reads are marked
- Base qualities are calibrated
  - Base qualities model the uncertainty of the base call

Data processing pipeline



SNPs and indels

- Single Nucleotide Variants (SNVs)

  ATATTATTGCCCAGACCAATGTCTTGGAGTTATTCCCCTGT

  ATATTATTACCCAGACCAATGTCTCGGAGTTATTCCCCTGT

- Insertions

  ATATTATTGCCCAGACCAATGTCTTGGAGTTATTCCCCTGT

  ATATTATTGCCCAGACCAGTTATGTCTTGGAGTTATTCCCCTGT

- Deletions

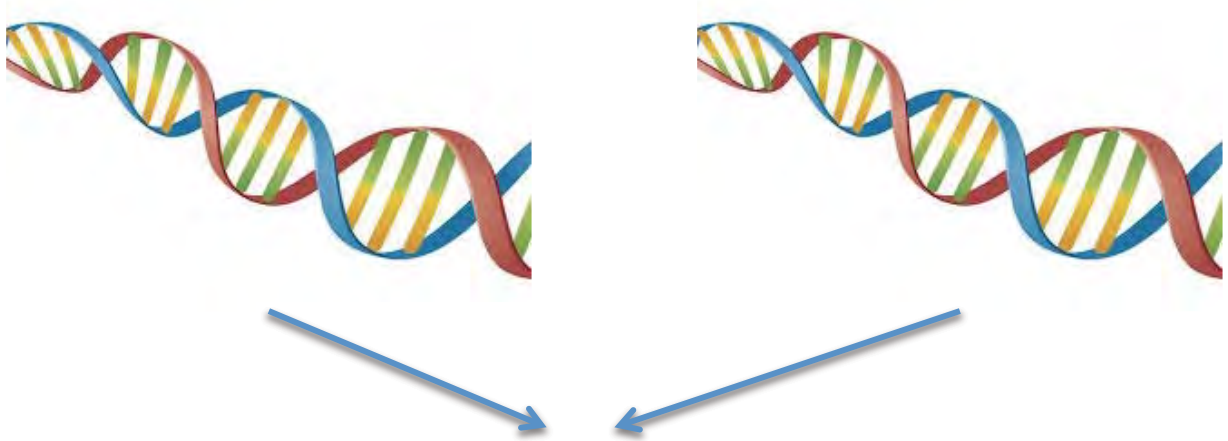
  ATATTATTGCCCAGACCAATGTCTTGGAGTTATTCCCCTGT

  ATATTATTGCCCAGACCAATGGAGTTATTCCCCTGT

NGS data is noisy

http://www.broadinstitute.org/gatk//events/2038/GATKwh0-BP-5-Variant_calling.pdf

Diploid is difficult



Chromosome pair reads are mixed in sequence data

- Homozygous
  - All reads should show the same allele
- Heterozygous
  - ~50% of the reads should show each allele

SNP Calling and Genotyping

- For each base in the genome, find the sample genotype given the observed reads.
- NGS reads contain errors (typically ~1% - 0.1% bases)

Homozygous reference (GG)

- 9 correct reads
- 1 error

Heterozygous (GT)

- 9 Correct reads (allele G)
- 1 Correct read (allele T)

Homozygous non-reference (TT)

- 1 correct reads
- 9 errors

SNP Calling and Genotyping

- For each base in the genome, find the sample genotype given the observed reads.
- NGS reads contain errors (typically ~1% - 0.1% bases)

Homozygous reference (GG)

- 6 correct reads
- 4 errors

Heterozygous (GT)

- 6 correct reads (allele G)
- 4 correct reads (allele T)

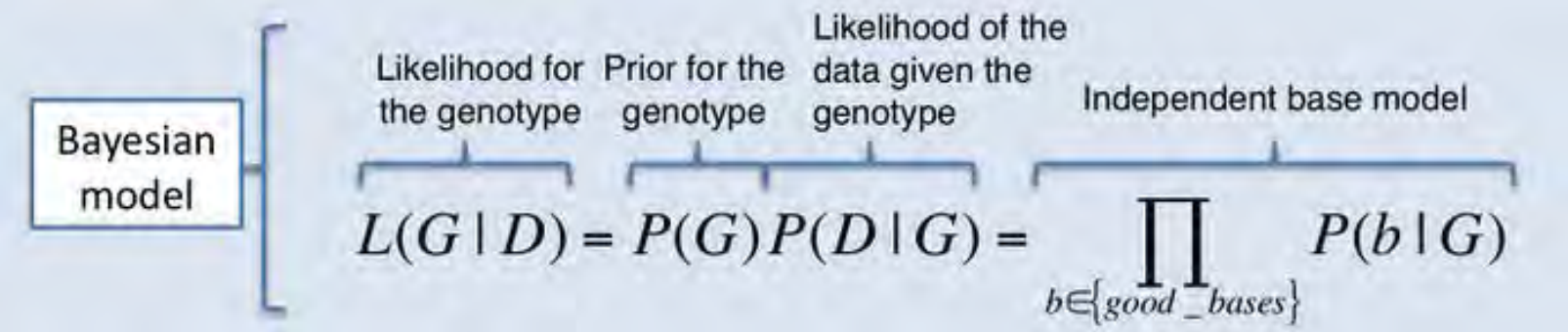
Homozygous non-reference (TT)

- 4 correct
- 6 errors

Genotyping Formalism (GATK)

Bi-Allelic Model

- *A*: Reference allele
- *B*: Alternate allele



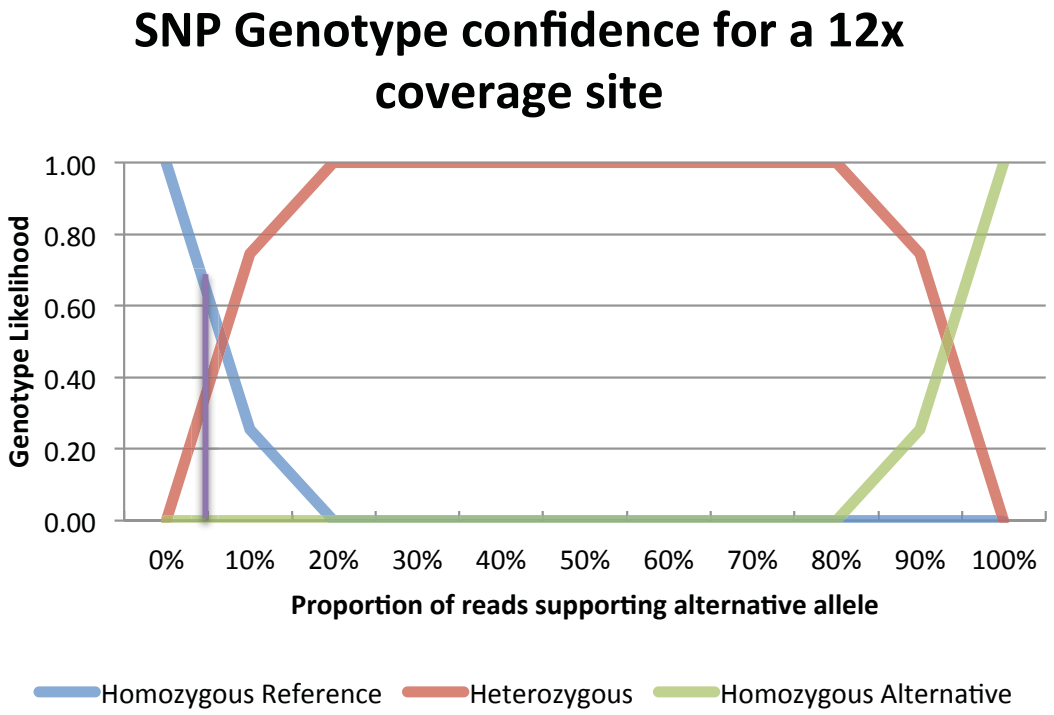
Homozygous

$$P(b \mid AA) = \begin{cases} 1 - e_b, \text{ if } b = A \\ e_b / 3, \text{ if } b \neq A \end{cases}$$

Heterozygous

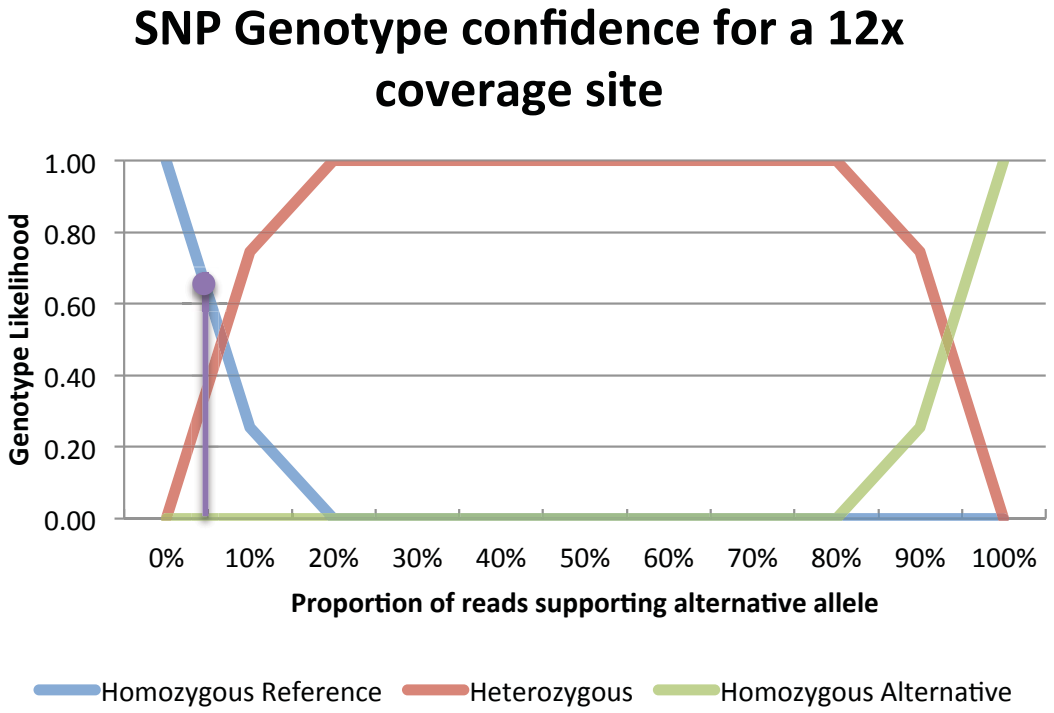
$$P(b \mid AB) = \begin{cases} (1 - e_b)/2 + e_b/6, \text{ if } b \in \{A, B\} \\ e_b / 3, \text{ if } b \notin \{A, B\} \end{cases}$$

Interpreting NGS genotypes



Genotyping a site given N supporting reads

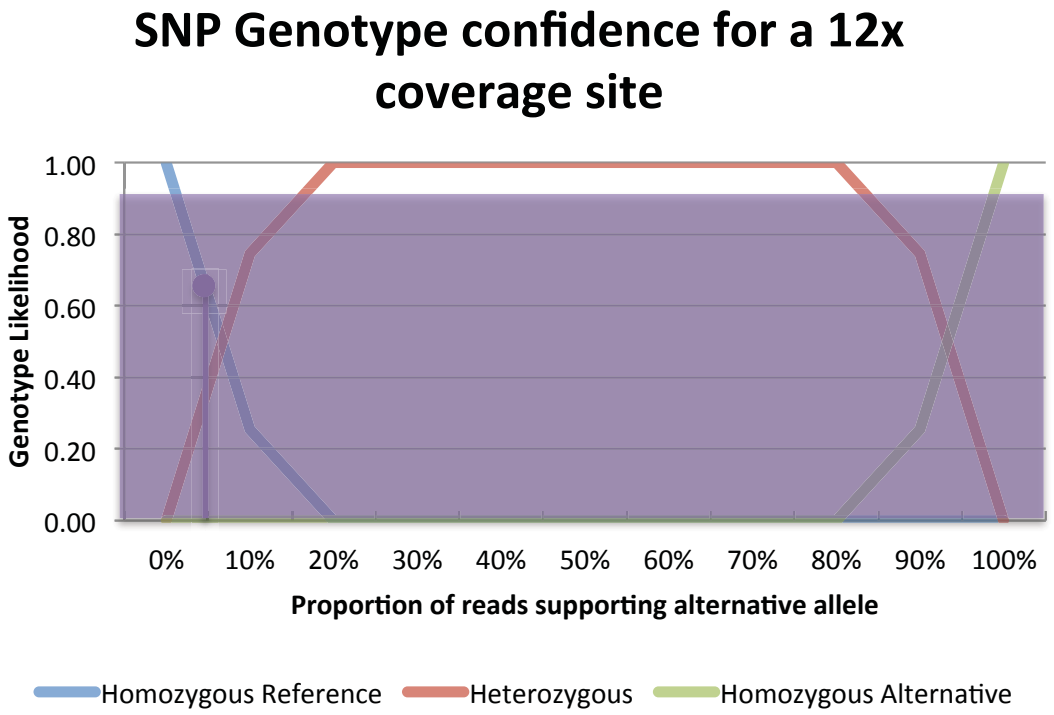
Interpreting NGS genotypes



Genotyping a site given N supporting reads

- **Most likely genotype**

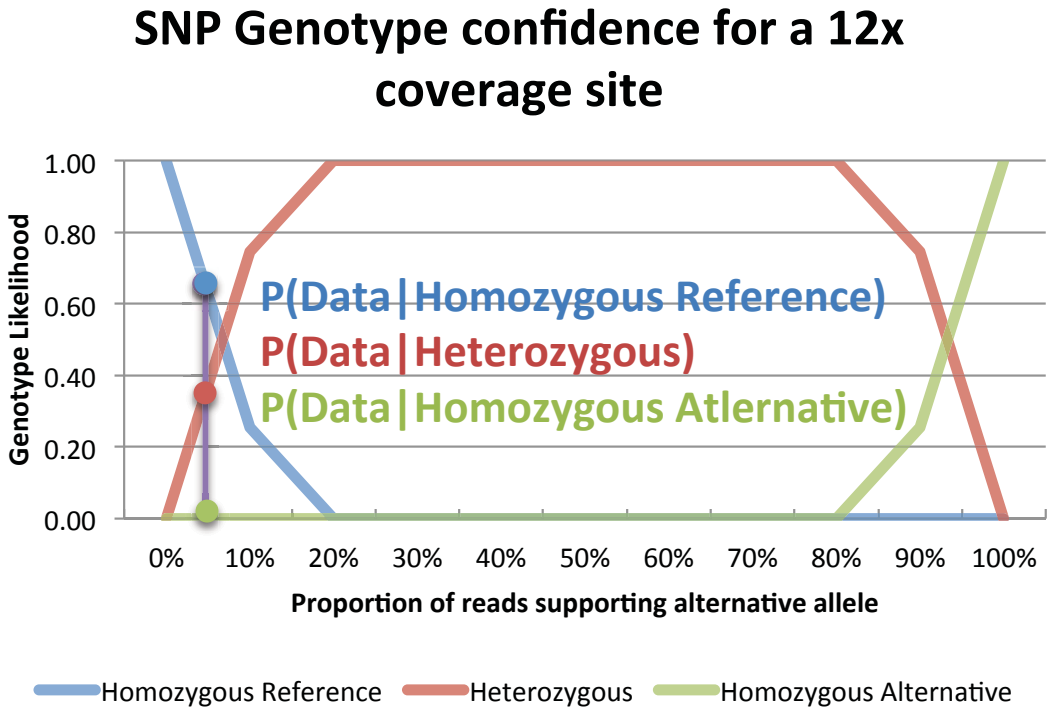
Interpreting NGS genotypes



Genotyping a site given N supporting reads

- Most likely genotype
- **Most likely genotype, only if within set confidence threshold**

Interpreting NGS genotypes



Genotyping a site given N supporting reads

- Most likely genotype
- Most likely genotype, only if within set boundaries
- **Genotype likelihoods for all possible genotypes**

Two approaches for calling SNPs and indels

Unified Genotyper

- Calls SNPs and INDELS **separately**
- Evaluates each base **independently**

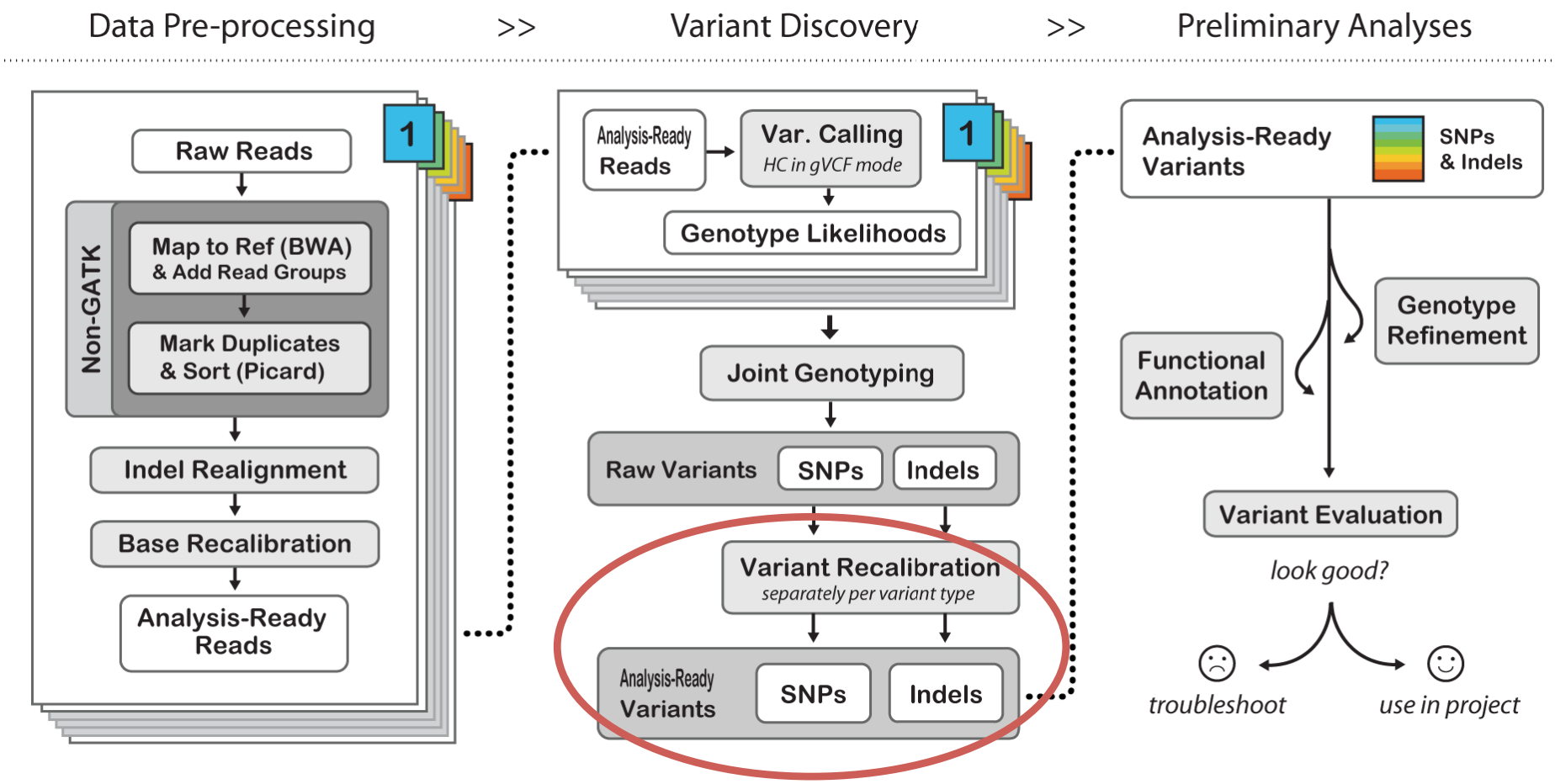
  – Fast!

Haplotype Caller

- Calls SNPs and INDELs **simultaneously**
- Uses all bases connected through local *de novo* assembly

  – No need for indel realignment
  – More accurate, especially for indels
  – Supports “N+1” genotyping

Data processing pipeline



From the Genome Analysis Toolkit: *http://www.broadinstitute.org/gatk/guide/best-practices*

Variant Annotations

- Reads
  - Strand bias
  - Quality / Depth
  - Mapping quality ranksum
  - …
- Biology / Population
  - Hardy-Weinberg Equilibrium
  - Haplotype Score

See GATK guidelines for a full list of recommended annotations to use!

Example: strand bias

Positive (or forward) strand

Negative (or reverse) strand

Example: strand bias

Positive (or forward) strand

Negative (or reverse) strand

Allele	Positive strand	Negative strand
A	30	5
C	35	30

Fisher's test p-value = 0.002

Two ways of filtering

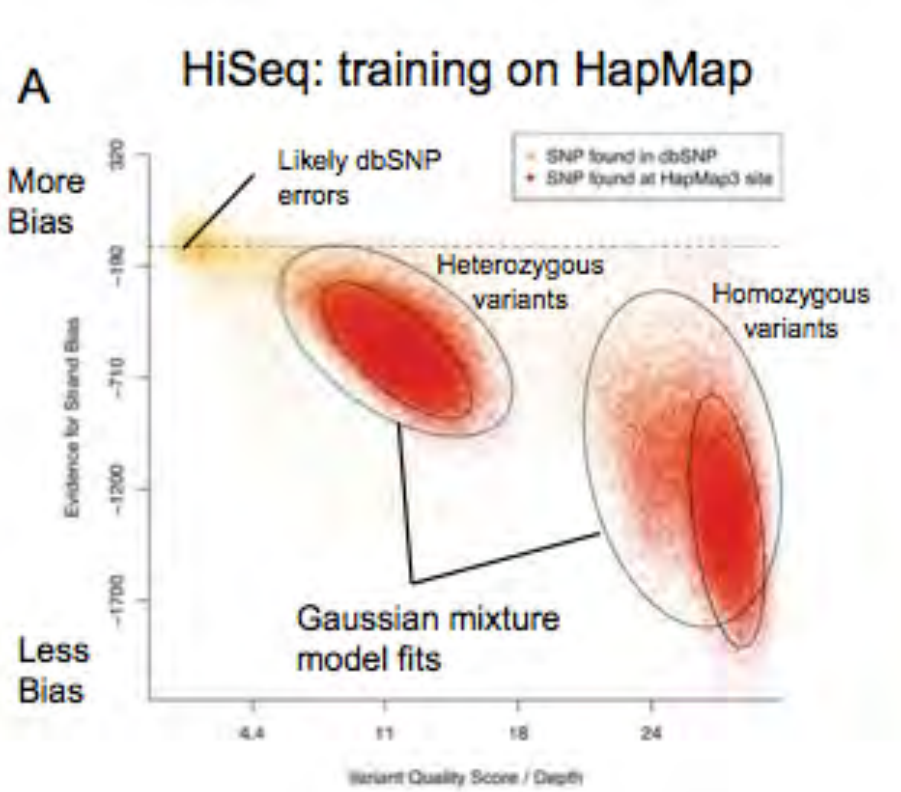
Thresholds

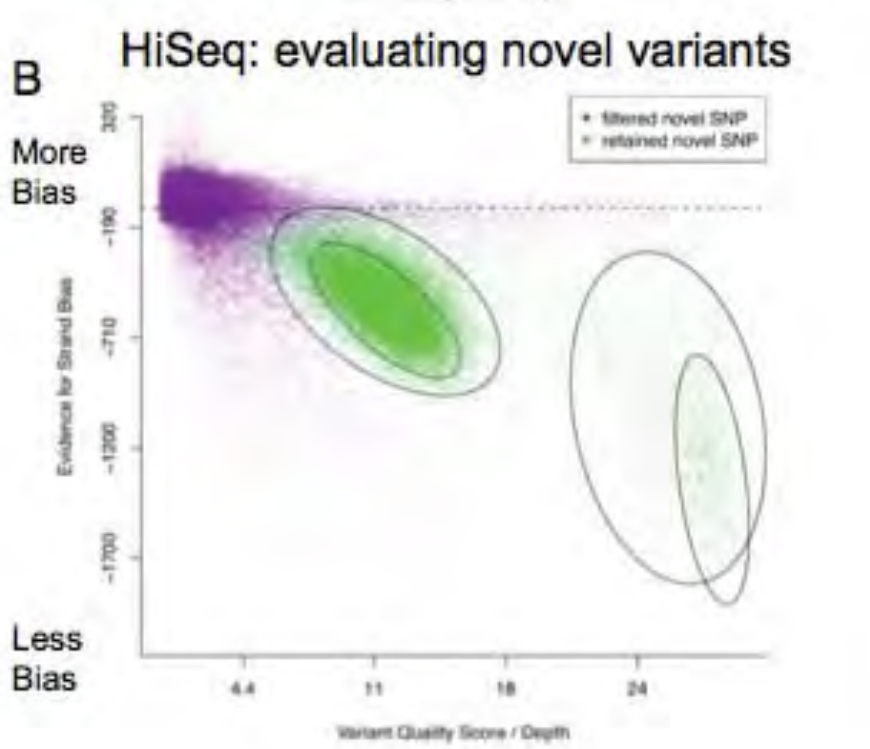
- Pros
  - Work for small datasets
  - No need for external resources
- Cons
  - Does not adapt to particular data
  - Manual tuning
  - Human bias

Machine-learning

- Pros
  - Multi-dimensional and complex model can fit the data better
  - Ranking of variants using a single metric
- Cons
  - Need training data
  - Need enough data

Machine learning filtering

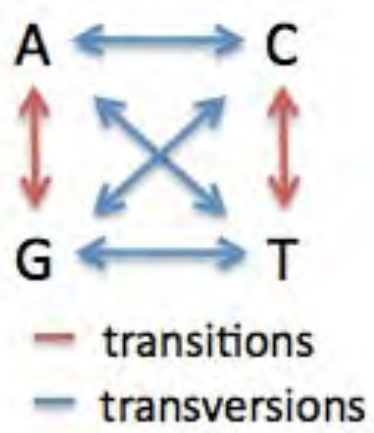




DePristo *et al.*, Nature Genetics 2011

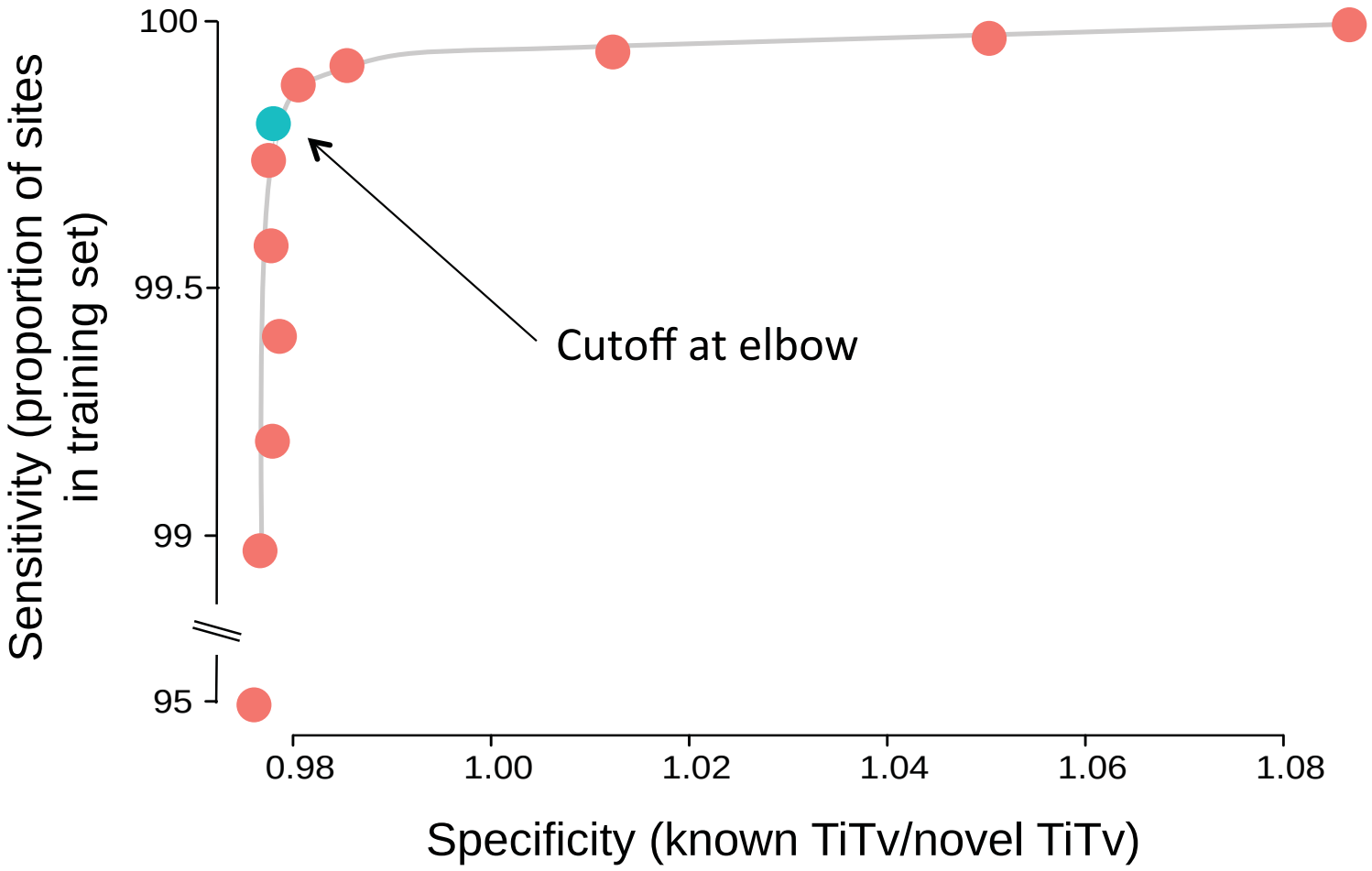
Filtering Evaluation

- Transition / Transversion (Ti/Tv)
- Proportion of novel variants
- Heterozygosity
- Allele frequency spectrum of variants
- Genotype concordance against chip (if available)

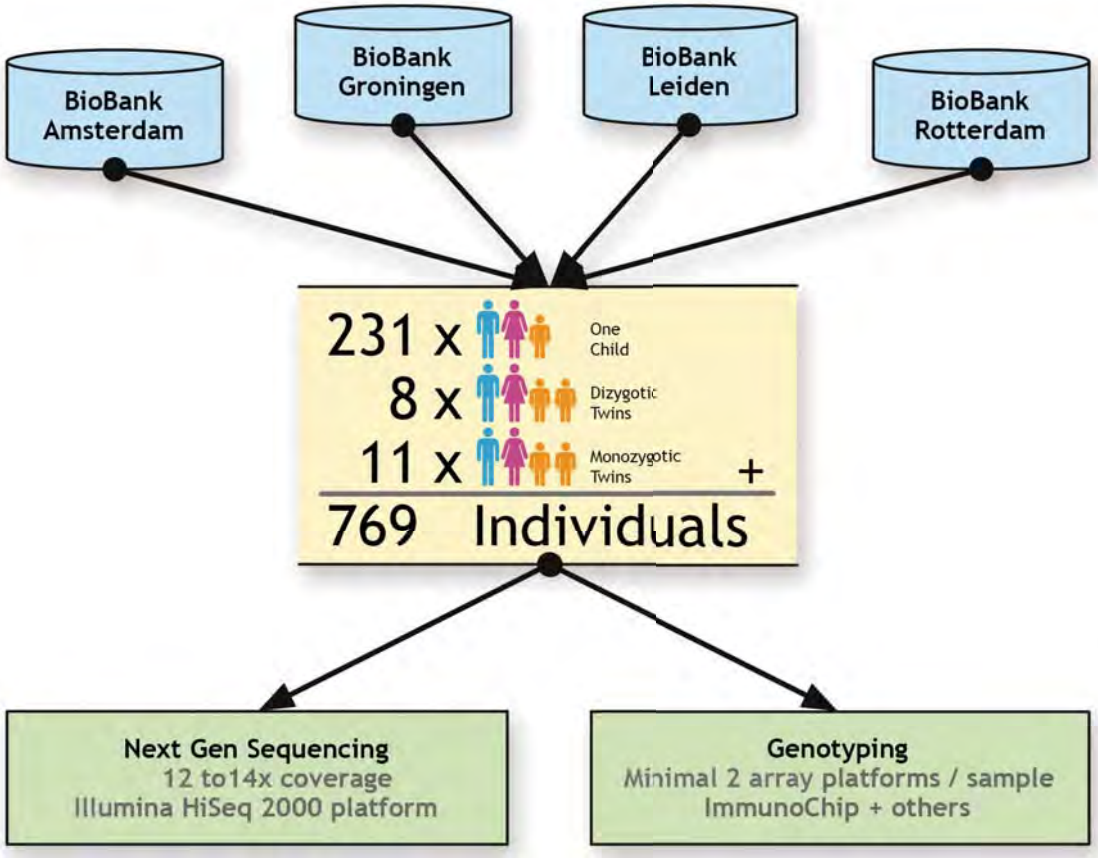


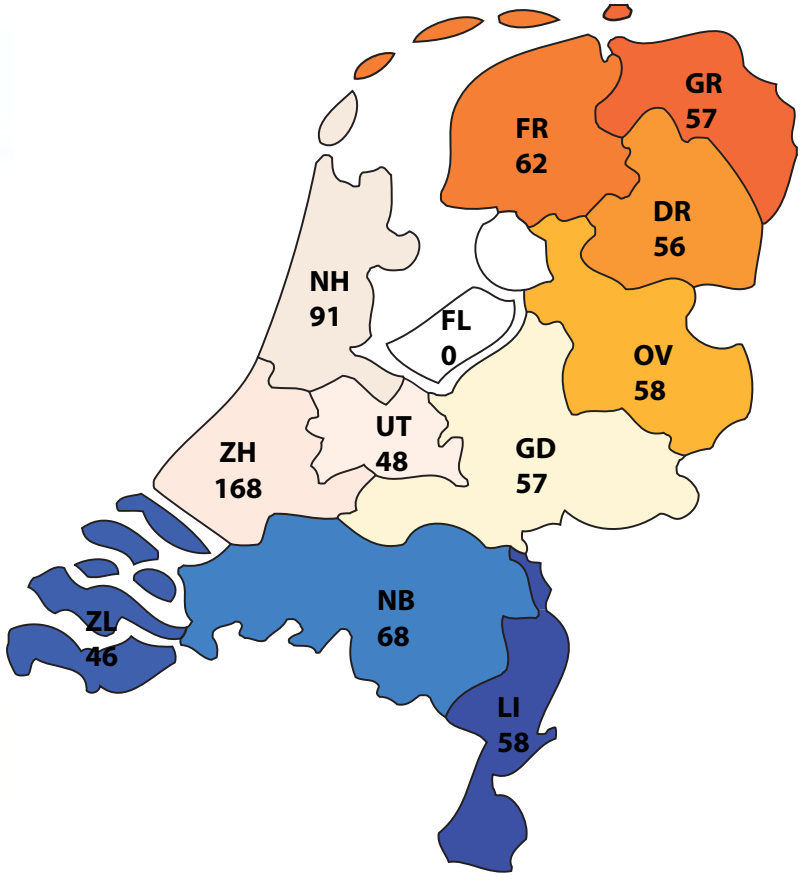
- Expected human Ti/Tv ratio
  - Whole-genome: 2.1
  - Whole-exome: 3.0
- FP SNPs should have a Ti/Tv of 0.5

Ti/Tv-based filtering threshold



Genome of the Netherlands (GoNL)



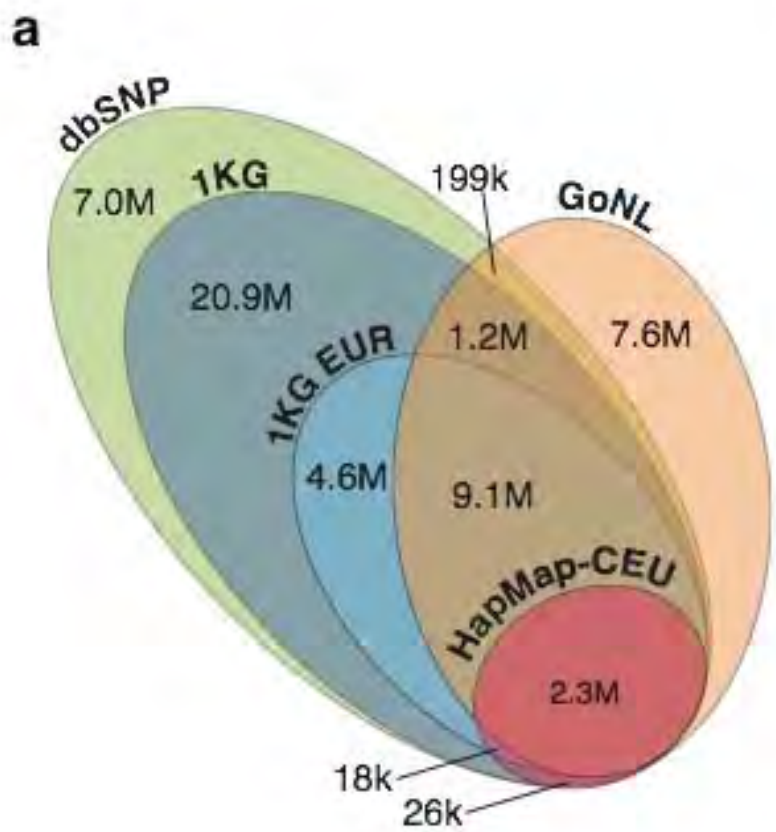


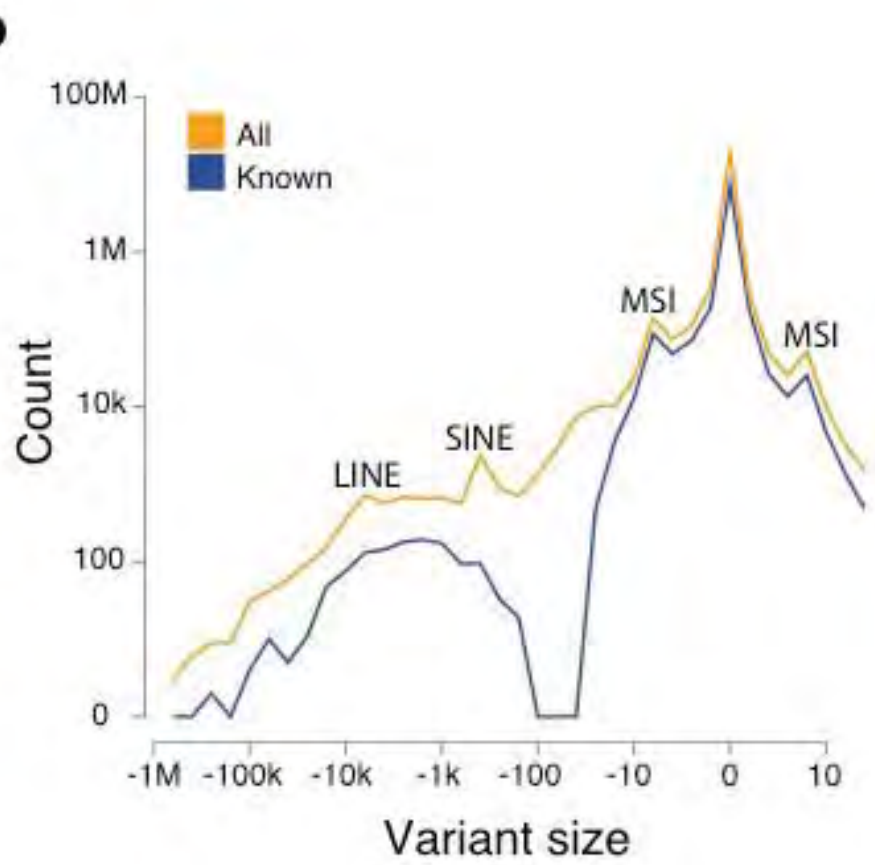
GoNL SNPs

250 Dutch trios sequenced at 12x coverage on Illumina HiSeq

Step	SNPs (mln)	% in dbSNP 137	Ti/Tv Known	Ti/Tv Novel	#novel SNPs per sample
Initial SNP Calling *Unified Genotyper*	22.10	44.71	2.18	1.85	154k
Variant Filtering *Variant Quality Score Recal.*	**19.76**	**60.01**	**2.25**	**2.14**	**26k**

Variation captured by GoNL





Francioli *et al.*, Nature Genetics (2014)

Resources

- GATK homepage
  - http://www.broadinstitute.org/gatk/index.php
- GATK video workshop
  - http://www.broadinstitute.org/gatk/guide/events?id=2038#materials
- GoNL homepage
  - http://www.nlgenome.nl