Variant Calling (Part 1 & 2): Quality Control and Pre-processing

EPP 622
Bioinformatics Applications

Bode Olukolu
Assistant Professor
Entomology and Plant Pathology

Outline

- Next-Generation Sequencing and Quality Controls (part 1)
  - Understanding multiplexed NGS libraries.
  - Base calling and quality scores.
  - Re-calibration of per-base quality scores.

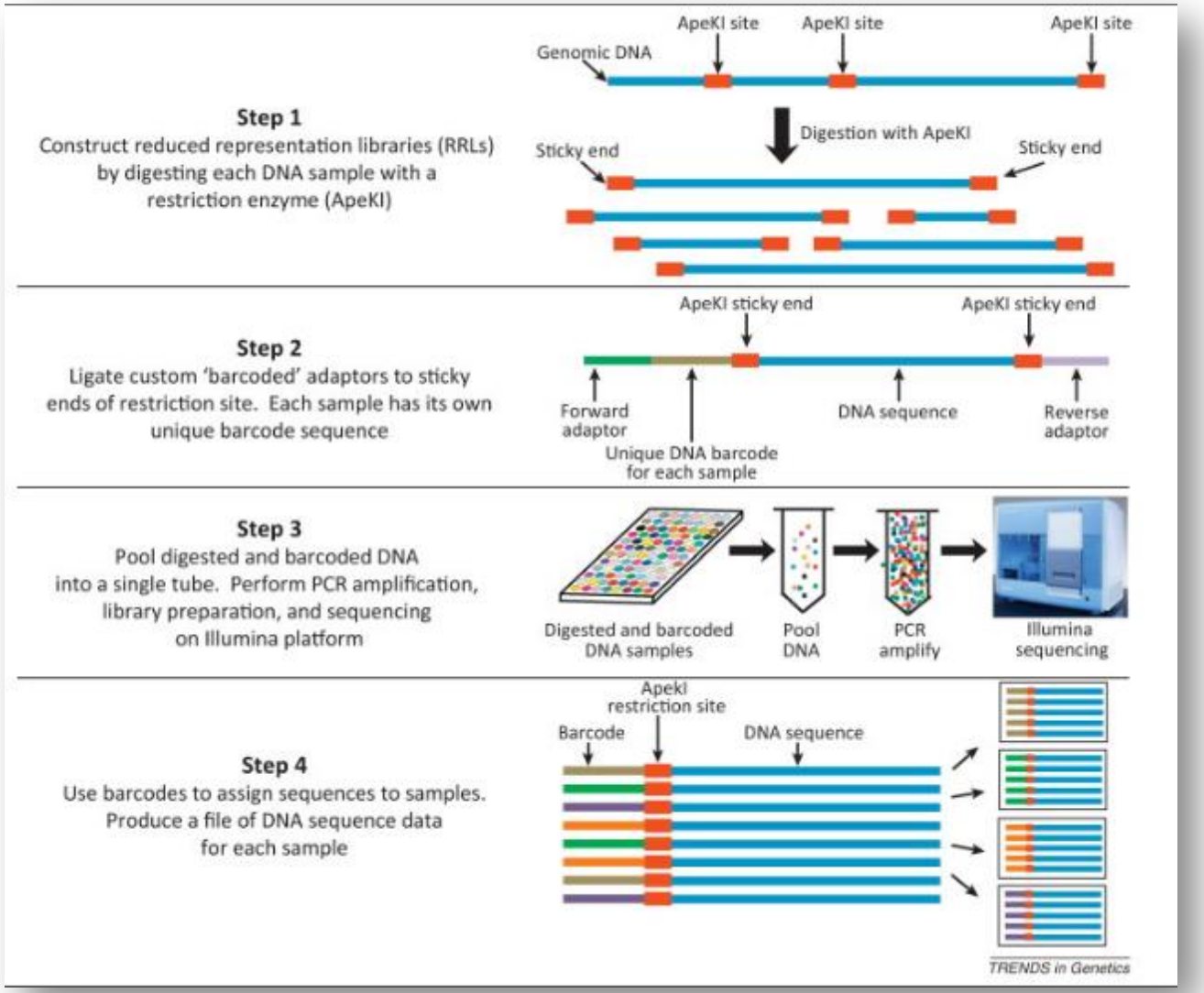
- Variant calling and filtering workflows (Part 2)
  - Alignment (reference genome/assembly, assembly errors, paralogs…).
  - SNP calling in diploids and polyploids.
  - Post-alignment quality control.
  - Post SNP calling QC
  - Filtering (read depth, quantitative genetics parameters)
  - Identifying variants types (SNP, indels, restriction site polymorphisms, epigenetic markers)

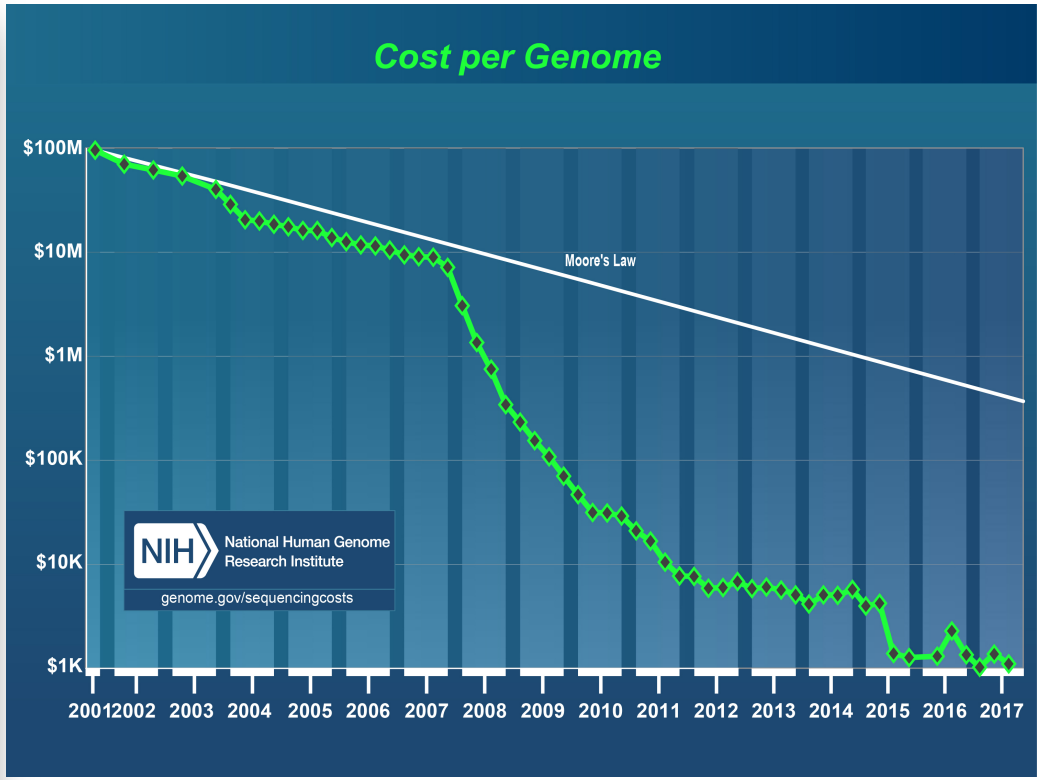
Suggested Reading:

Nielsen R, Paul JS, Albrechtsen A, Song YS. (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011 Jun;12(6):443-51.

Irina A, Boekhorst R, Orlov Y. (2017) Computational errors and biases in short read Next Generation Sequencing. J Proteomics Bioinform. 10:1

Introduction: sequencing-based genotyping





- *RAD-Seq (Restriction Site Associated DNA)*
- *GBS (Genotyping-By-Sequencing)*
- *ddRAD-Seq*
- *DArTseq*
- *GBSpoly*
- *Capture-Seq*
- *MonsterPlex*

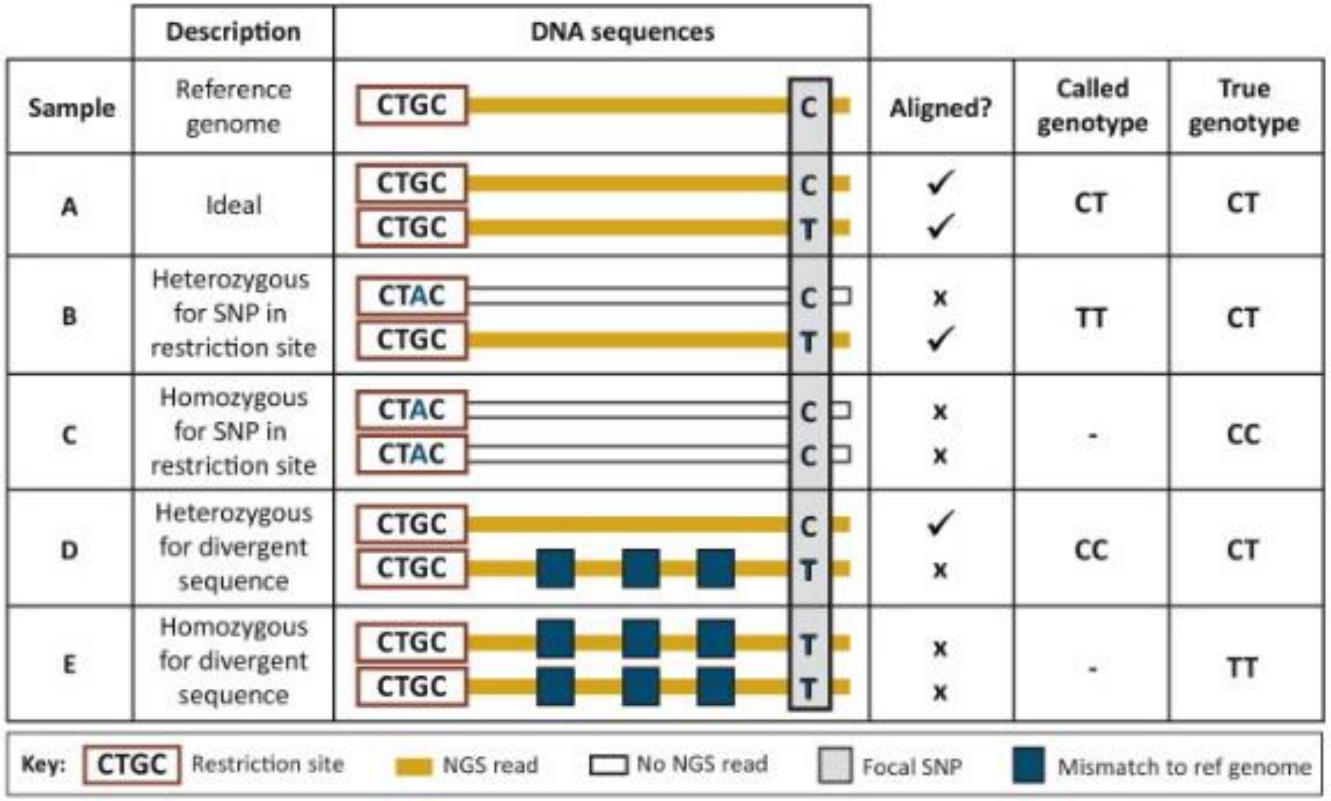
Elshire et al. (2011) PLoS ONE; *Myles 2013, Trends in Genetics*;
https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/

Sequencing-based genotyping pitfalls

- ✓ NGS produces vast amount of data with error that is difficult to distinguish from true biological variation (leads to non-existent SNPs.)
- ✓ Allele and Ascertainment bias
- ✓ Difficulty in calling heterozygotes in diploids and dosage in polyploids
- ✓ A lot of missing data.

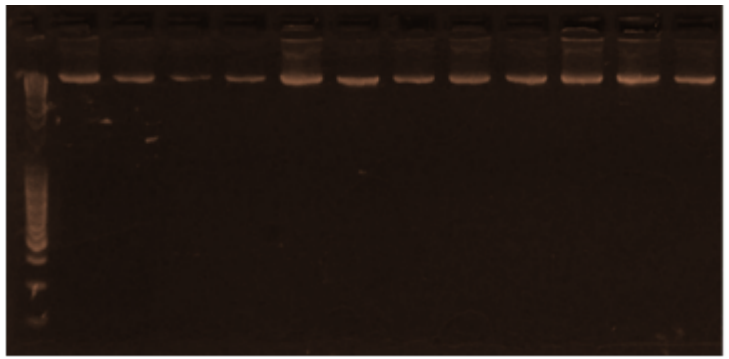
Nonetheless, GBS assay is now commonly used and possibly the future of high-throughput genotyping.

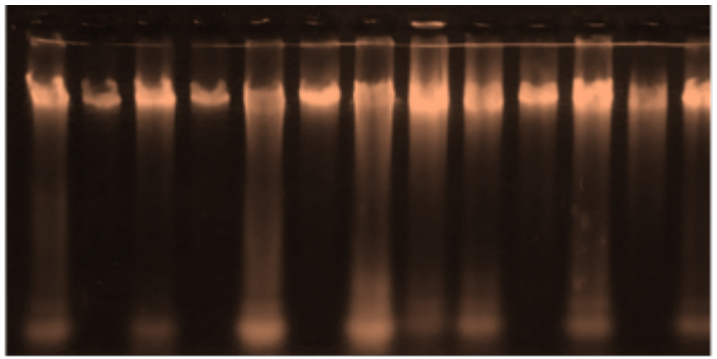
- *Inexpensive*
- *High-density markers*
- *Useful for non-model organisms*

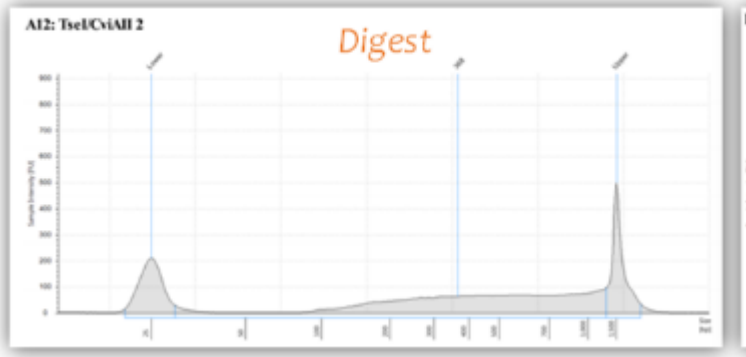


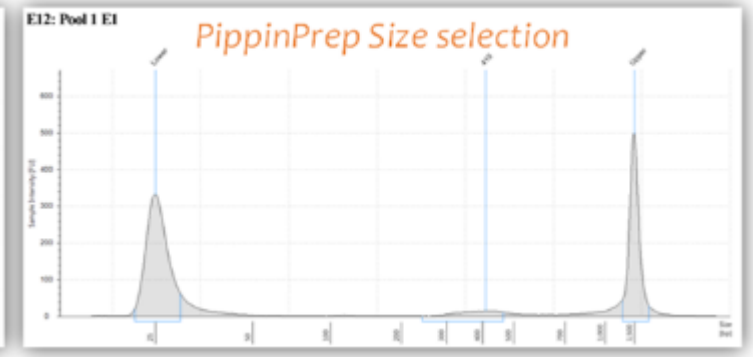
Myles 2013, Trends in Genetics

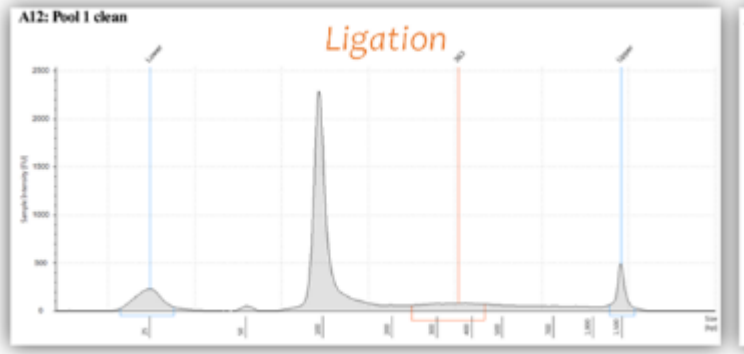
Library prep and quality control

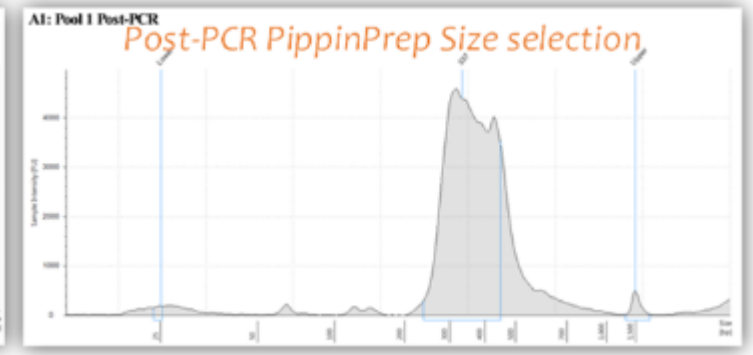












Pre-library prep:

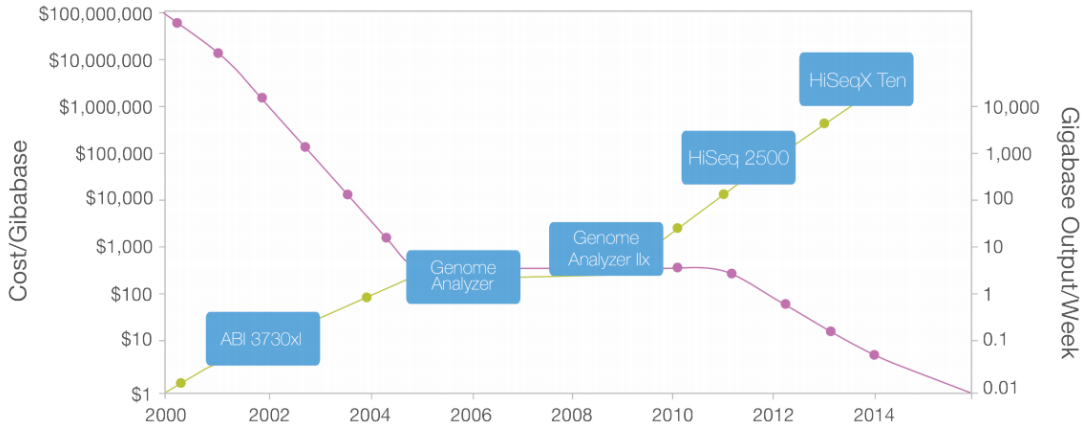
- DNA quality check
- Enzyme combination
- Barcode/adapter design

Library prep:

- Double digest
- Adapter/barcode Ligation
- Size selection
- PCR amplification
- Illumina sequencing

Why multiplex/pooled NGS libraries

- ✓ Multiplex sequencing: large number of individual samples sequenced simultaneously.
- ✓ Coverage from NGS equipment would be unnecessary for the reduced representation of a single genome (or small genome).
- ✓ Translates to inexpensive assay



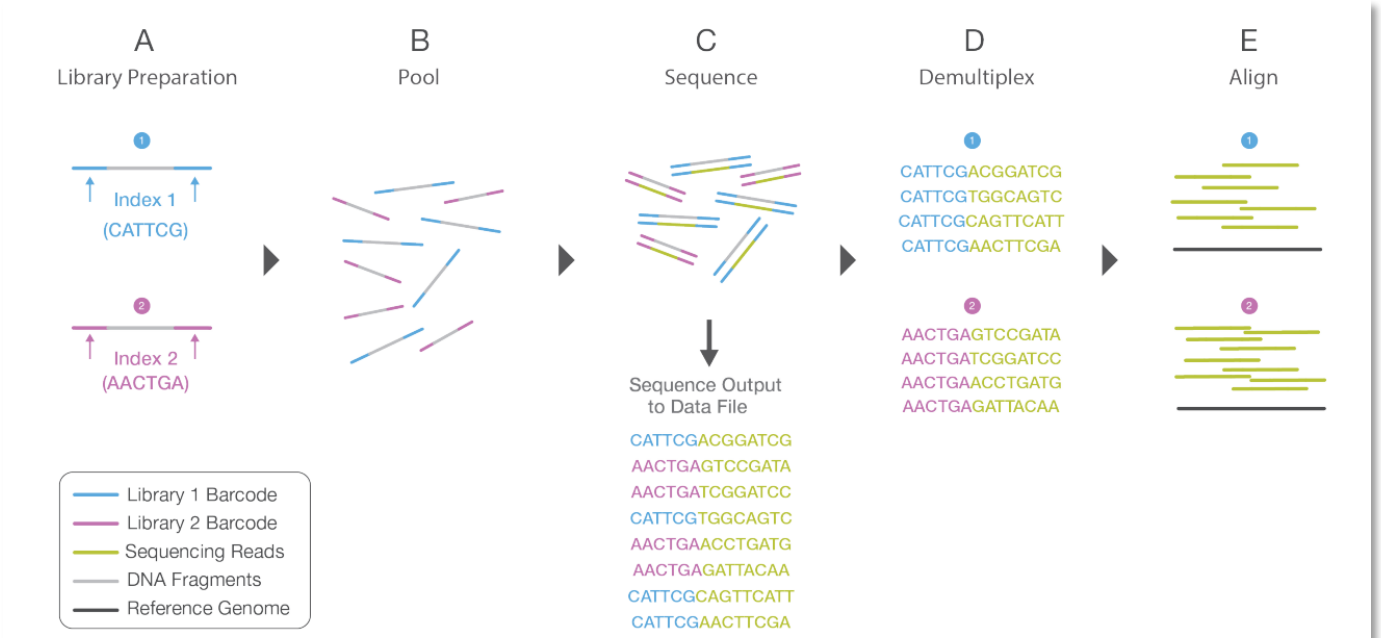
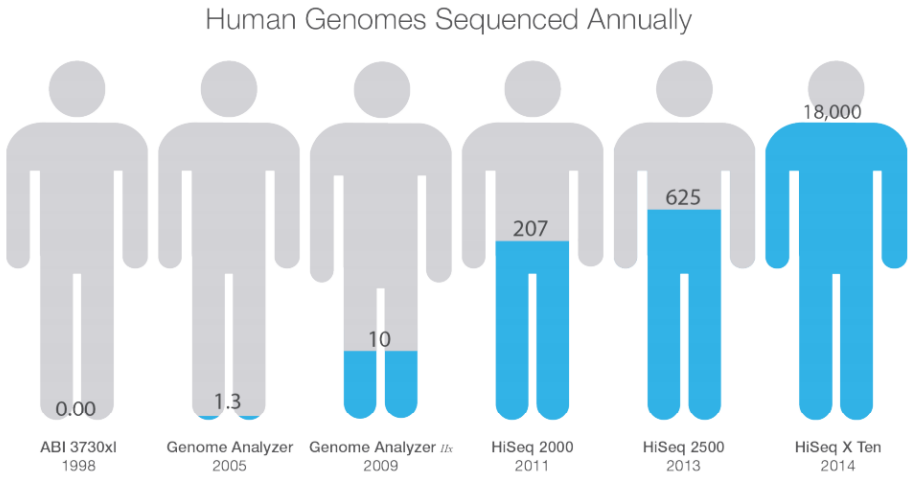


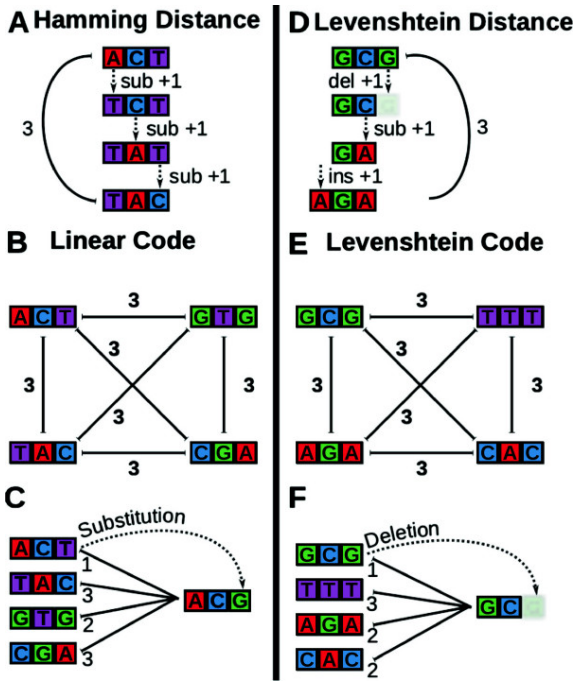
Figure 5: Library Multiplexing Overview—(A) Unique index sequences are added to two different libraries during library preparation. (B) Libraries are pooled together and loaded into the same flow cell lane. (C) Libraries are sequenced together during a single instrument run. All sequences are exported to a single output file. (D) A demultiplexing algorithm sorts the reads into different files according to their indexes. (E) Each set of reads is aligned to the appropriate reference sequence.

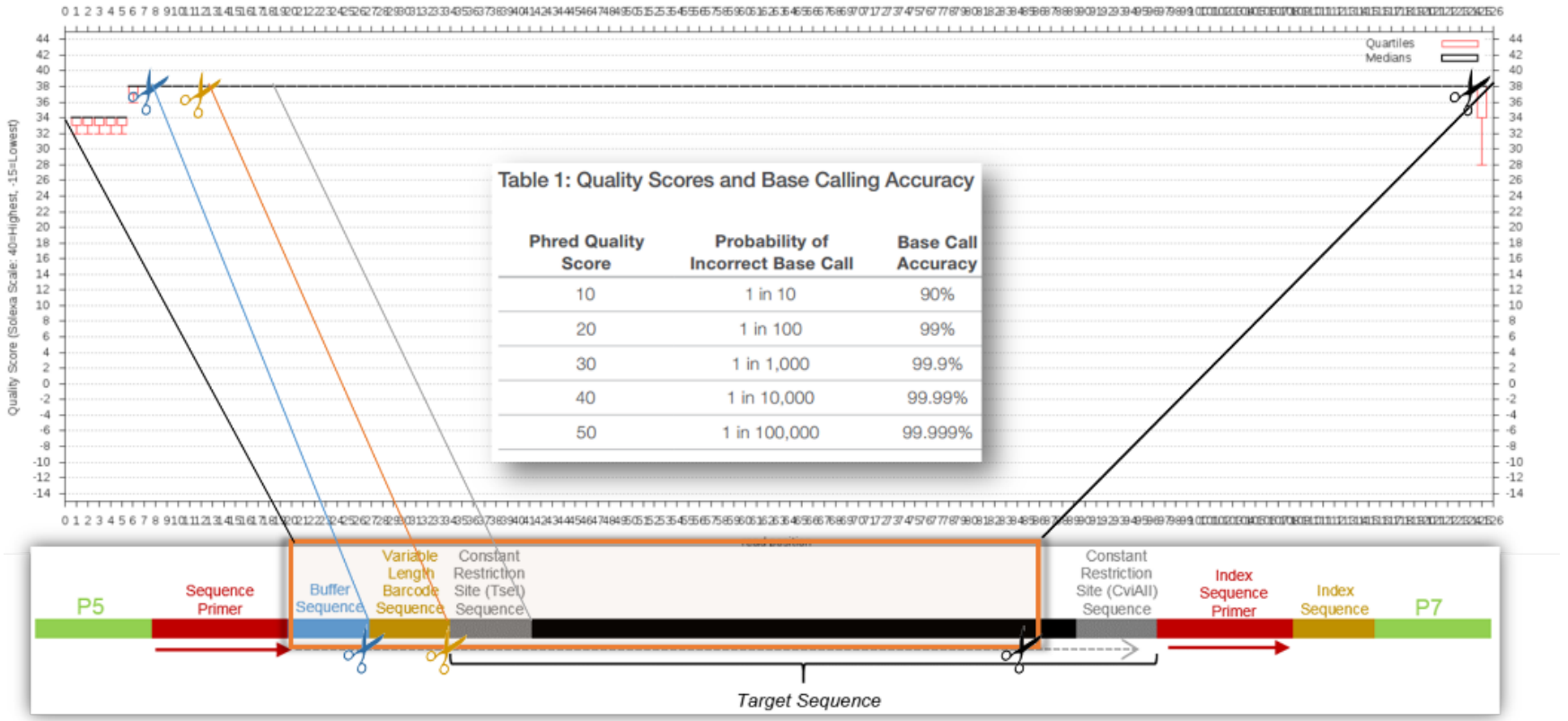


Illumina Inc.

Considerations for multiplex-sequencing

- ✓ Location of barcodes on the sequence read (problematic if located at proximal ends).
- ✓ Demultiplexing accuracy.
- ✓ “Hamming distance” vs. “edit/Levenshtein distance”.
- ✓ Diversity of barcode nucleotide composition.
- ✓ Variable length barcode to resolve phasing error.
- ✓ Chimeric reads

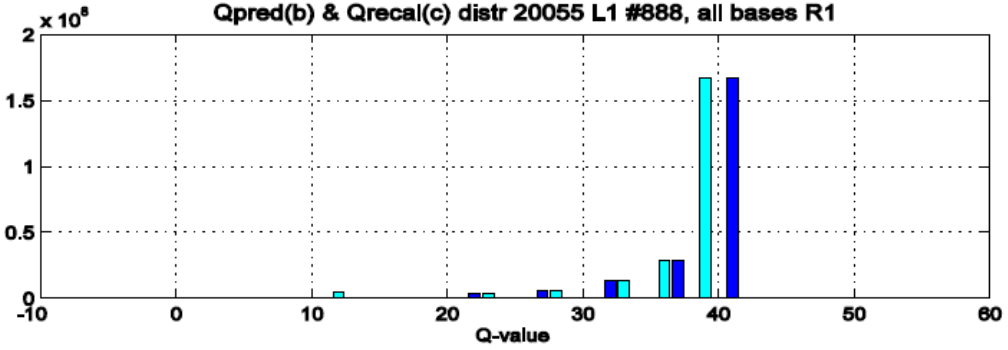


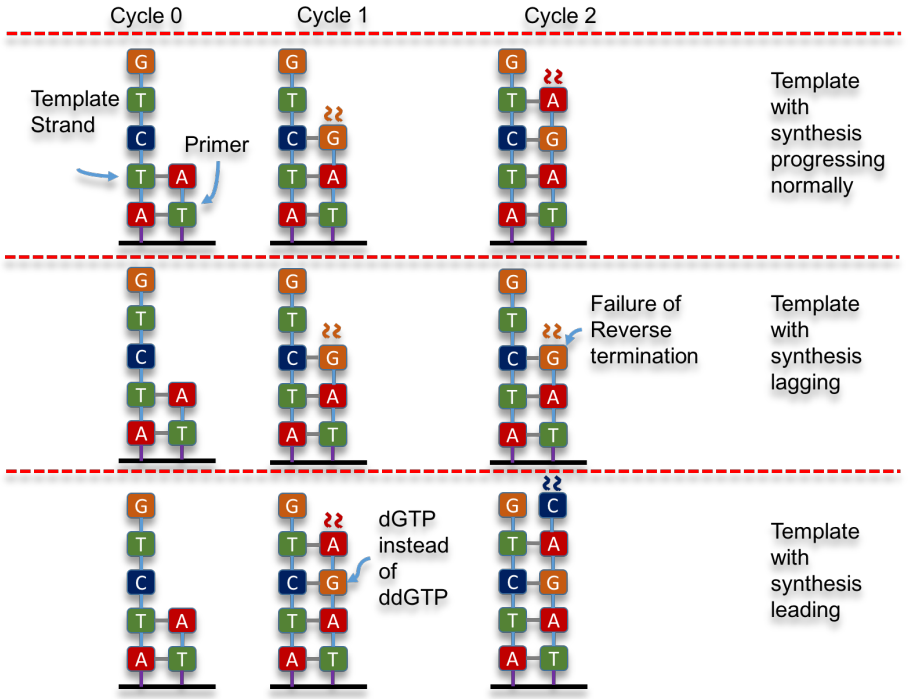


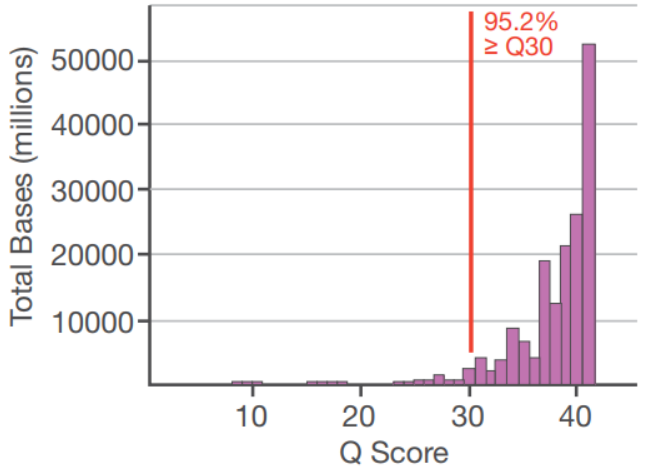
Buschmann et al. (2013) BMC Bioinformatics;

Illumina sequencing: sources of base calling error

- ✓ Phasing error
- ✓ Lack of diversity at position across samples (typical of A-tailed libraries and digested libraries retaining overhangs)
- ✓ Unpredictable random errors along entire length of read (as shown in quality scores).
- ✓ Quality scores seem to be approximations or binned values
- ✓ QC plots only represent 95% of quality scores, hence, leading to false impression about random errors.





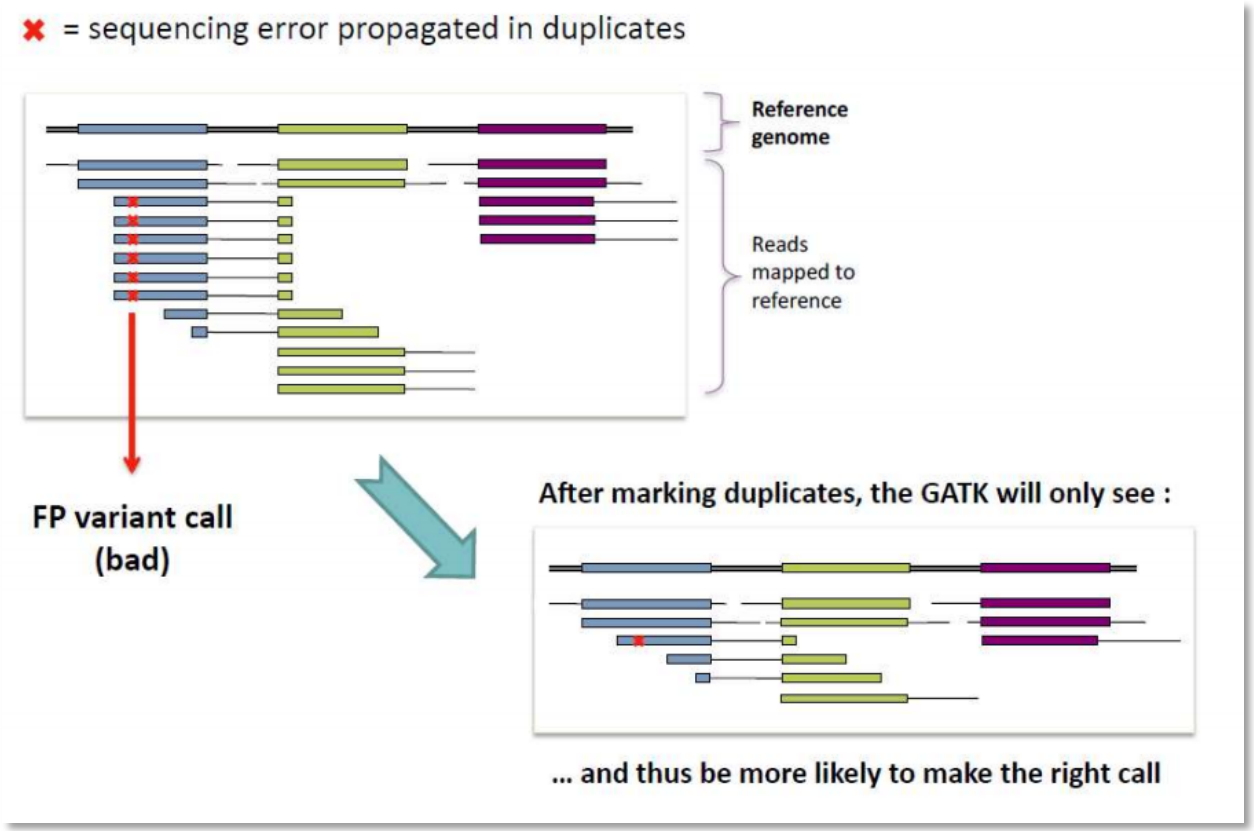


Illumina Inc; Irina et al. (2017) J. Proteomics Bioinform.; http://data-science-sequencing.github.io/Spr2016/lectures/lecture3/

Removing PCR/optical duplicates

- ✓ Marks reads that are duplicates and retains read with highest sum of base quality scores.

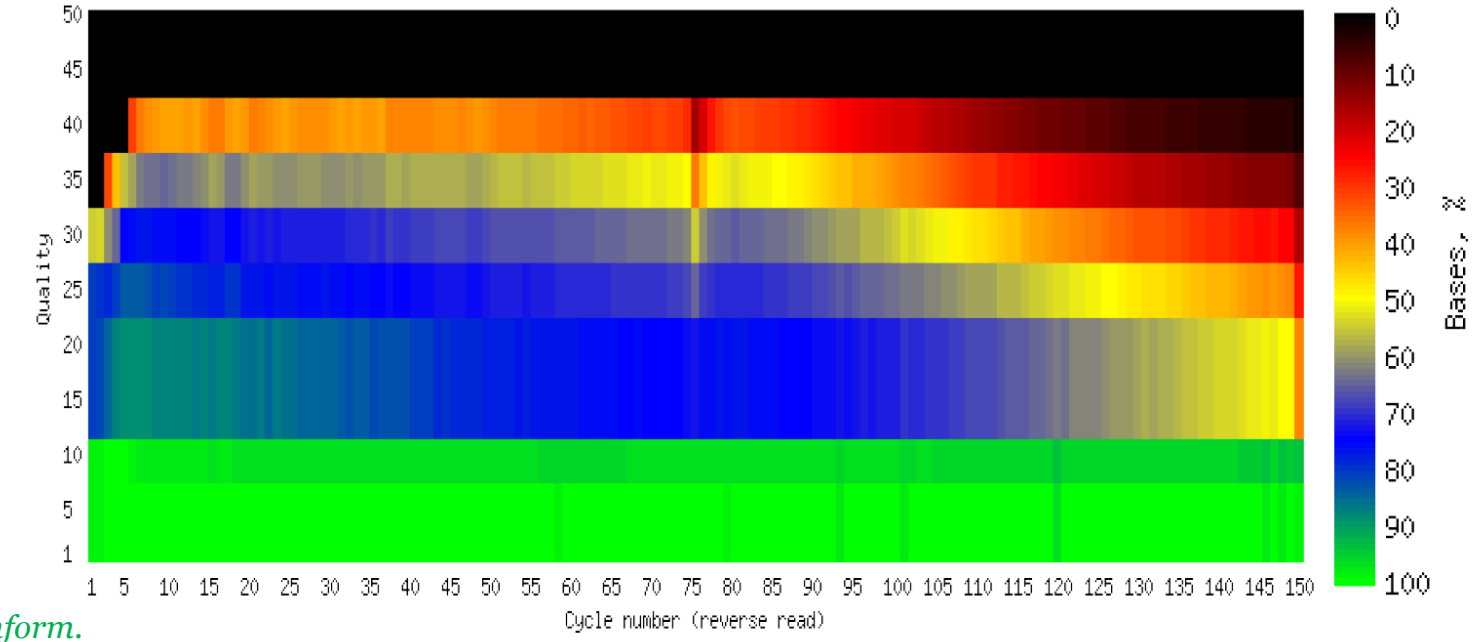
- ✓ Not perfect:
  - Does not account for sequencing errors
  - Does not account for natural duplicates (paralogs)
  - Does not account for duplicate reads with different mapping locations



Metzker ML (2010) Nat Rev Genet. 11(1):31-46

Base calling and quality scores

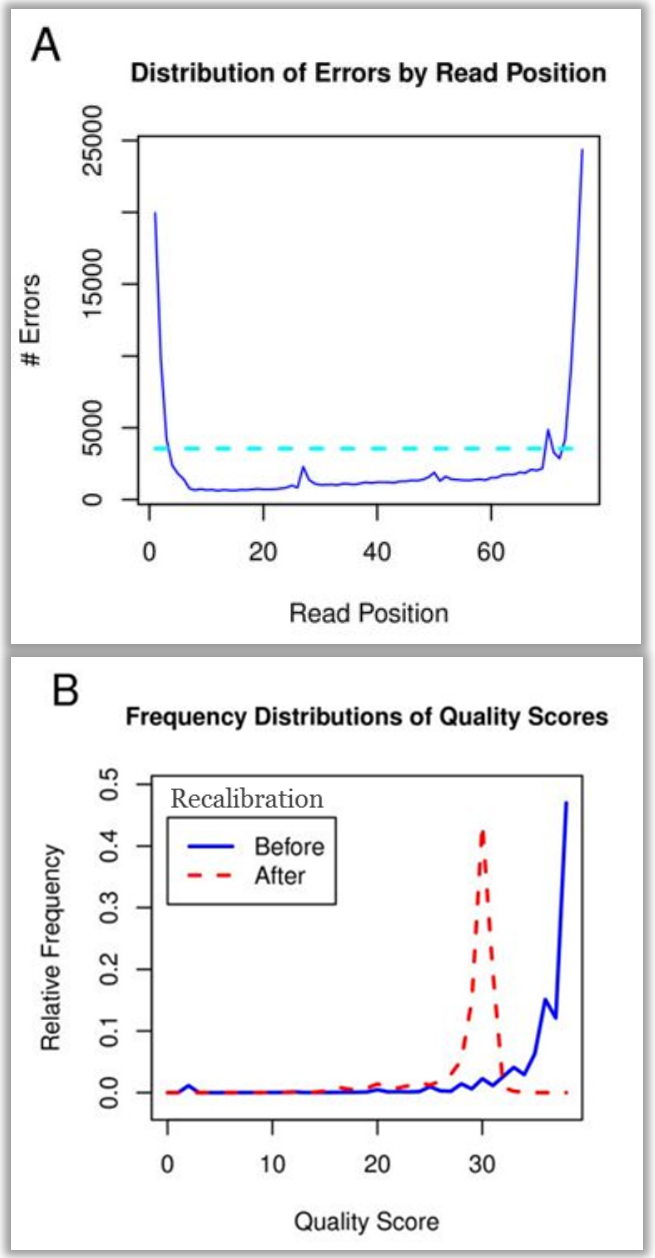
- ✓ How to estimate miscall error rate?
- ✓ High error rate at proximal ends of read inflated and crucial for demultiplexing pooled libraries.
- ✓ A 1% error rate might seem miniscule, but retaining them in reads lead to non-existent SNP calls. This account for the inflated minor allele frequency observed in most GBs data.



Irina et al. (2017) J. Proteomics Bioinform.

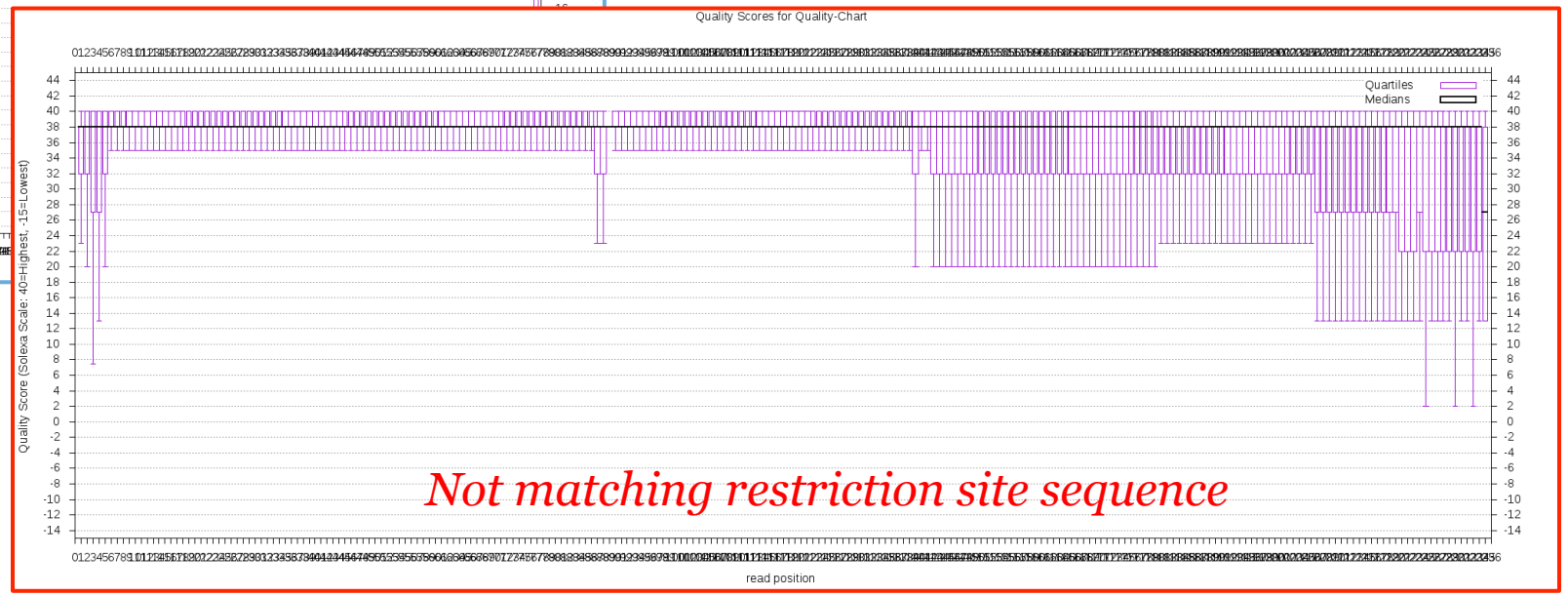
Re-calibration of per-base quality scores

- ✓ Raw Phred-scaled quality scores produced from base-calling algorithms may not accurately reflect the true base-calling error rates.
- ✓ Hence, need for recalibration.
- ✓ **SOAPsnp:** recalibrated by comparing a sequenced genome to the reference genome at sites with no known SNPs. Algorithm estimates the empirical quality score by using the number of mismatches with respect to the reference genome.
- ✓ **GATK:** A related alignment-based recalibration algorithm has been implemented in the GATK software. It also takes into account several covariates such as machine cycle and dinucleotide context



Cabanski et al. (2012) BMC Bioinformatics

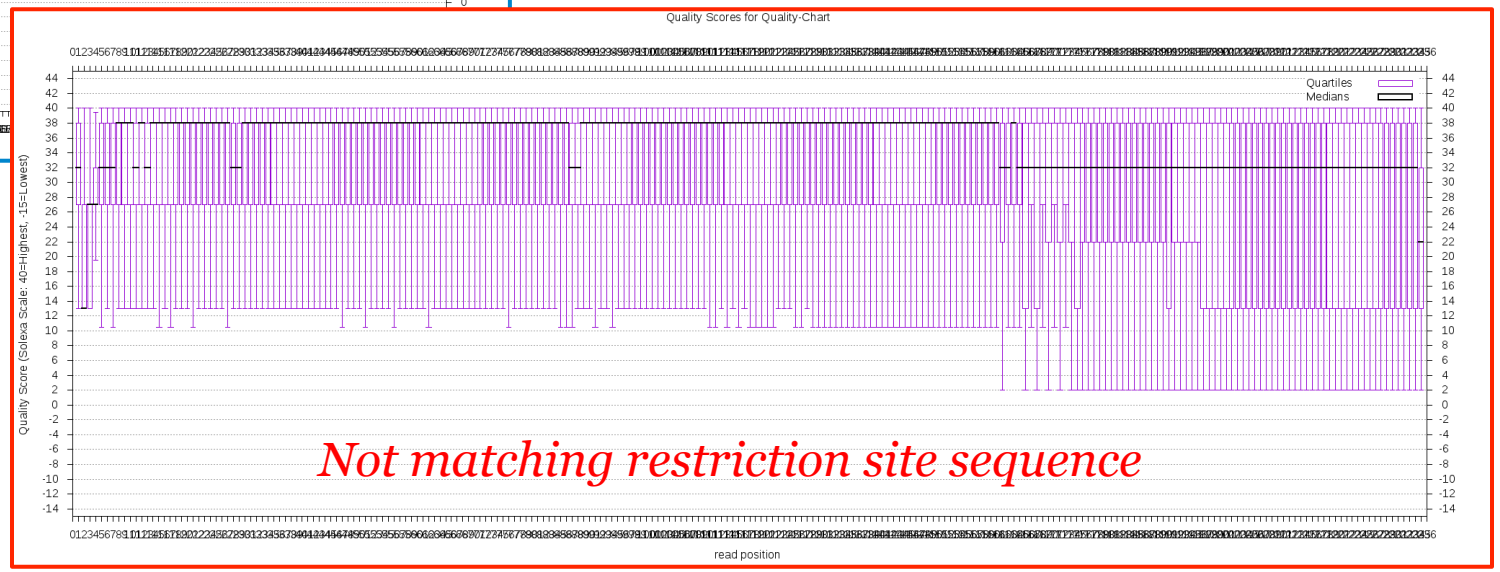
Error filtering:
Forward Read



Error filtering : **Reverse Read**

Matching restriction site sequence

Demultiplexed reads



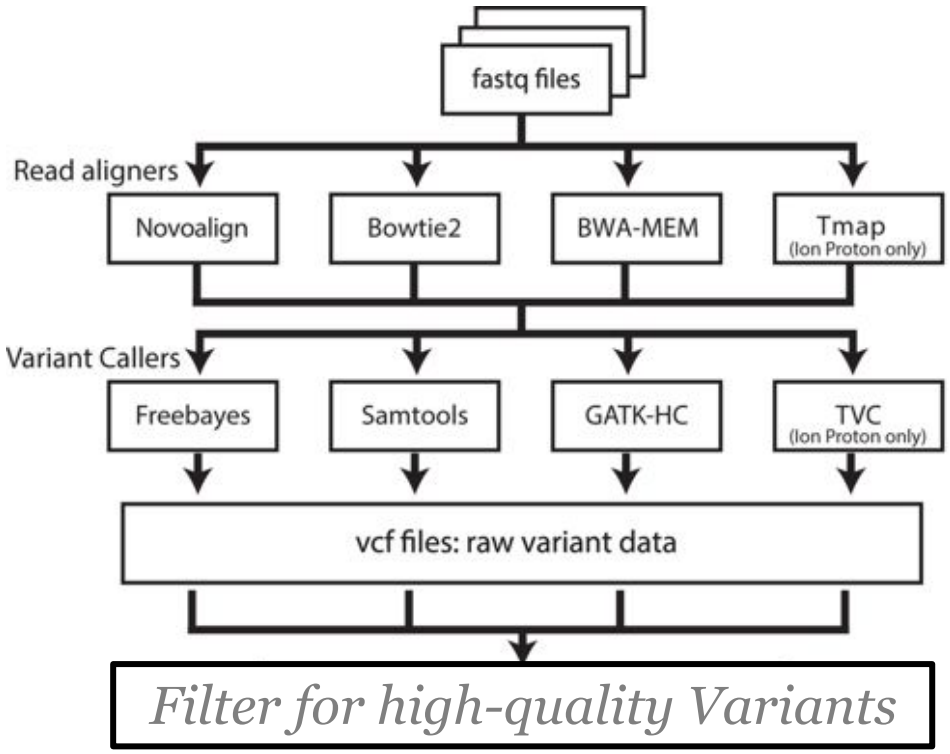
Not matching restriction site sequence

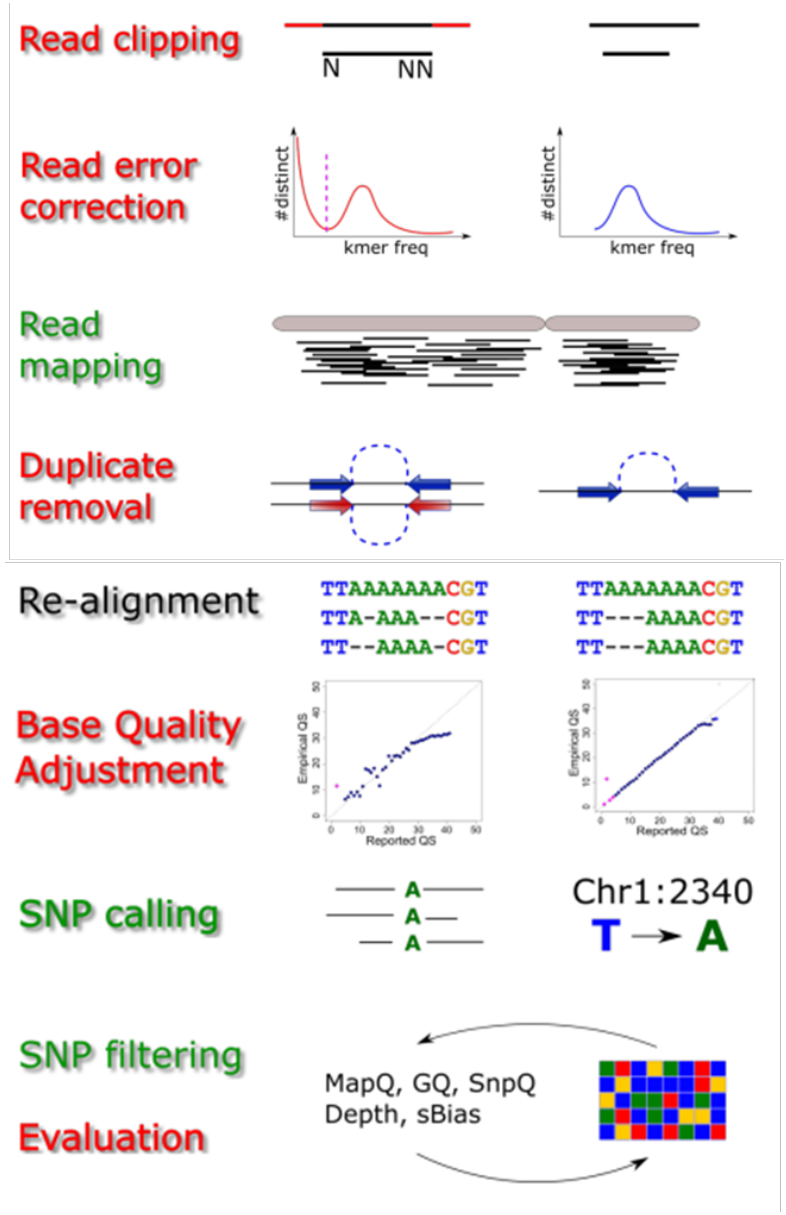
Outline

- Next-Generation Sequencing and Quality Controls (part 1)
  - Understanding multiplexed NGS libraries.
  - Base calling and quality scores.
  - Re-calibration of per-base quality scores.

- Variant calling and filtering workflows (Part 2)
  - Alignment (reference genome/assembly, assembly errors, paralogs…).
  - SNP calling in diploids and polyploids.
  - Post-alignment quality control.
  - Post SNP calling QC
  - Filtering (read depth, quantitative genetics parameters)
  - Identifying variants types (SNP, indels, restriction site polymorphisms, epigenetic markers)

SNP calling workflows





Types of variants:

- SNPs - indels
- CNVs (copy number variants)
- PAVs (Presense/Absence Variants)
- Epigenetic markers

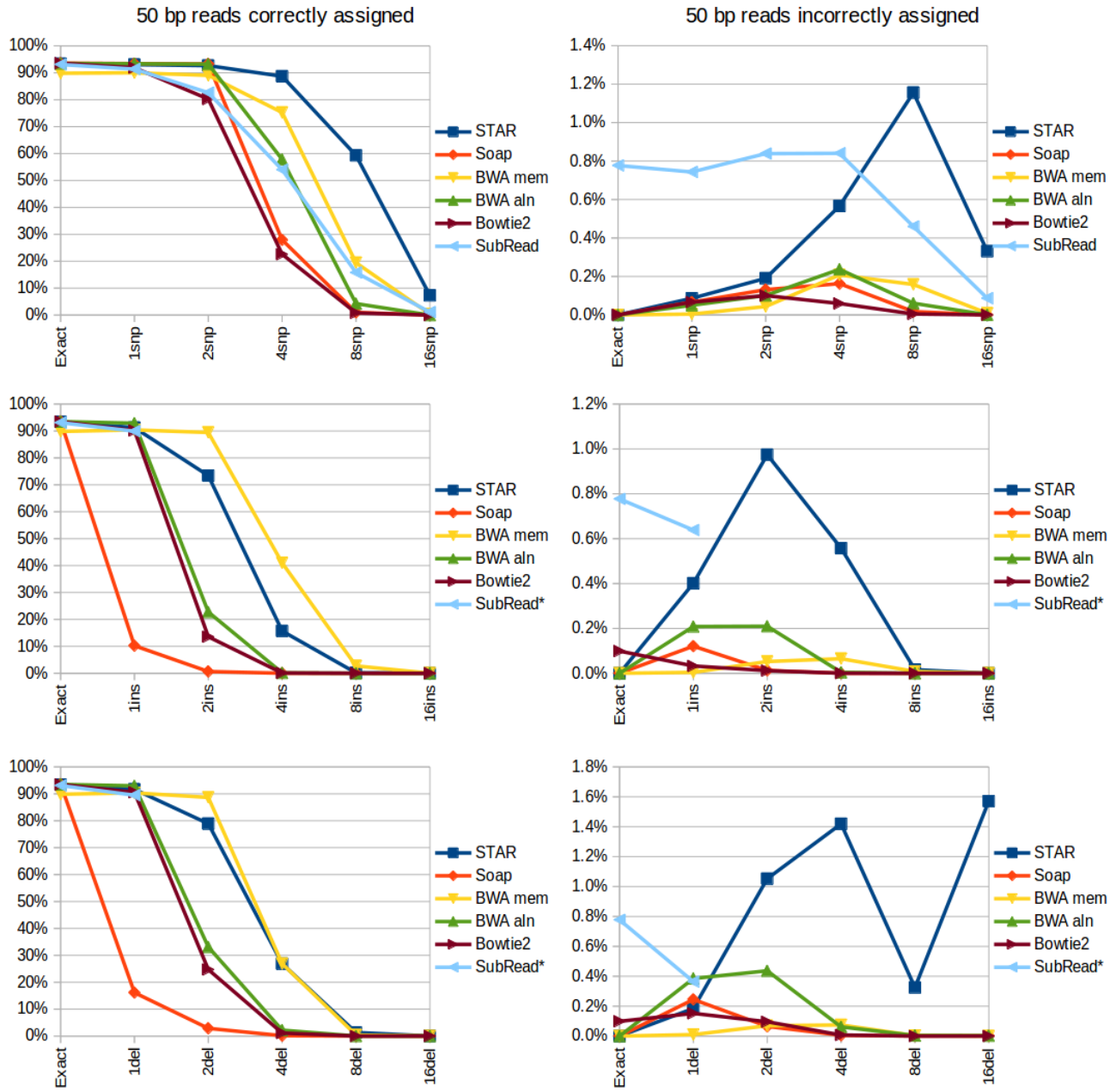
Hwang et al. (2015) Scientific Reports; Nielsen et al. (2011) Nat Rev Genet; http://www.transplantdb.eu/sites/transplantdb.eu/files/SNPcallingTutorial-Haberer_010715.pdf

Read alignment

- ✓ Samples can be divergent from reference genome, even among the same species.
- ✓ Paralogs or duplicate sequences can lead to error if reads from multiple loci pile up at one locus
- ✓ In polyploids, it is important to differentiate between "subgenome-specific sequences" and "common sequences across subgenomes.
- ✓ SNP calling approaches in species lacking a reference genome assembly are typically suboptimal.

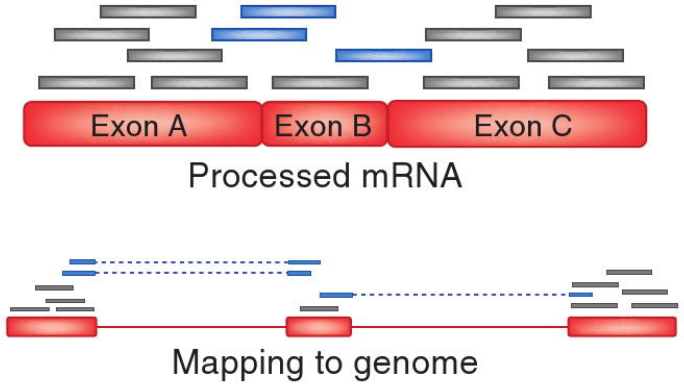
	Hexaploid BxT[1] *(%)*	*Diploid M9xM19*[1] *(%)*
After de-multiplexing pooled samples	94.6	97.3
Reads matching I. trifida nuclei genome[2]	87.5	90.8
Reads matching and unique to I. triloba	6.3	-

→ ~5% PhiX

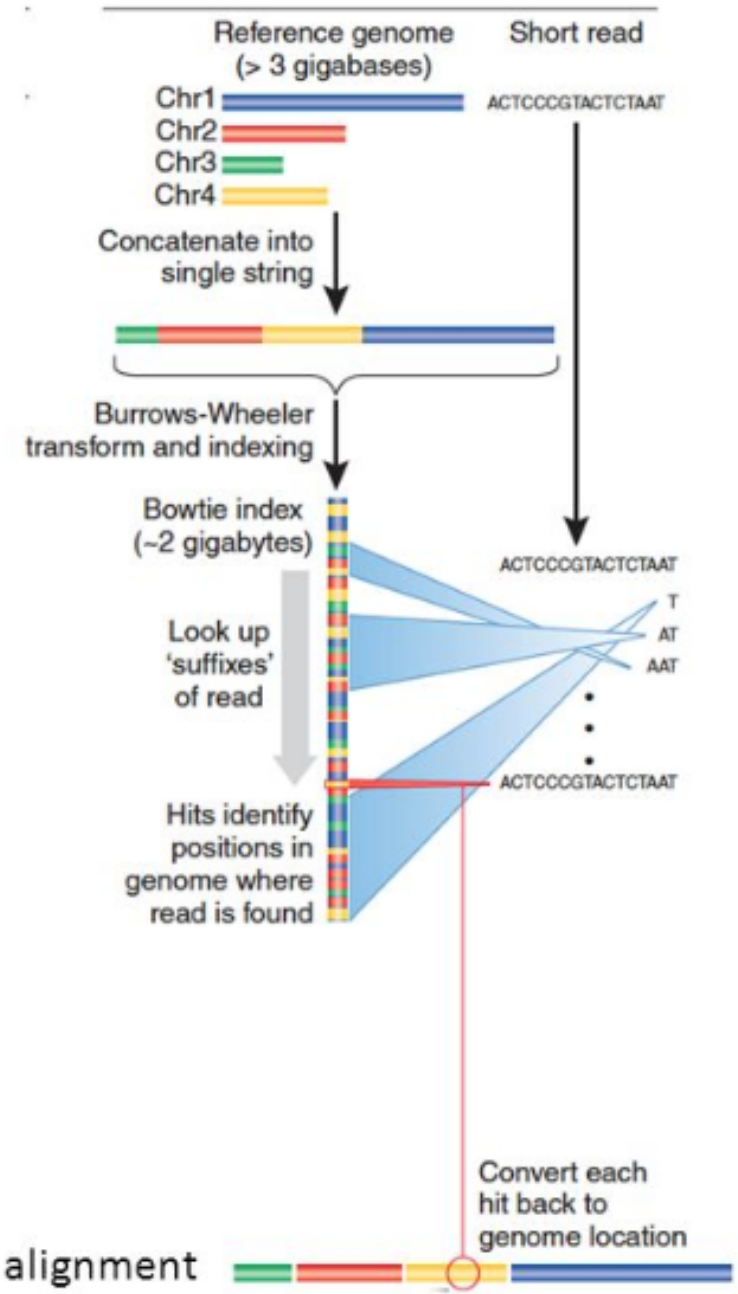


http://genomespot.blogspot.com/2014/11/dna-aligner-accuracy-bwa-bowtie-soap.html

Why Genome Indexing?

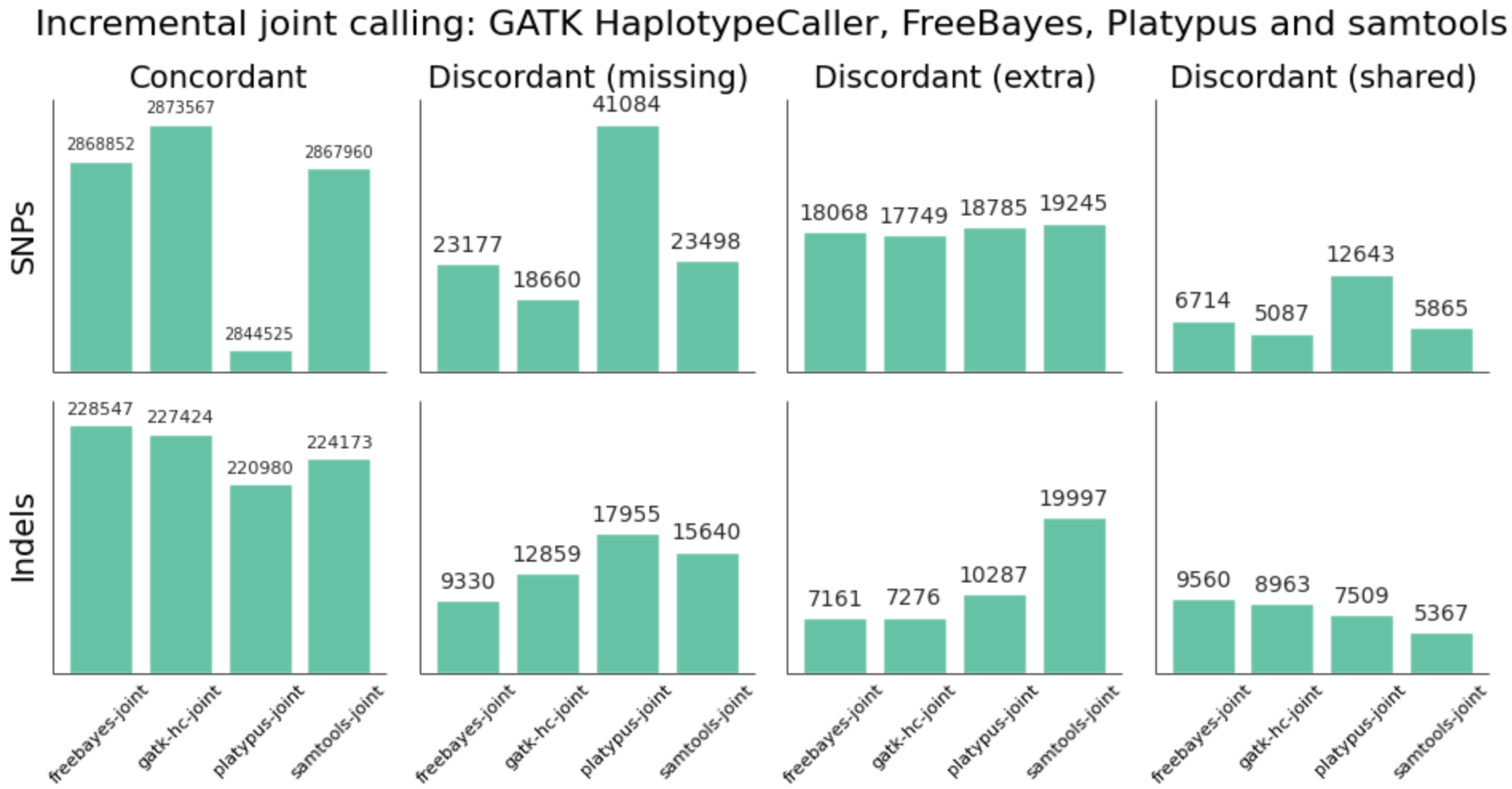


- ✓ Simplifies genome sequence
- ✓ Results in indexed genome i.e. hits/matches identify positions in genome where reads are found.
- ✓ Very fast and efficient.



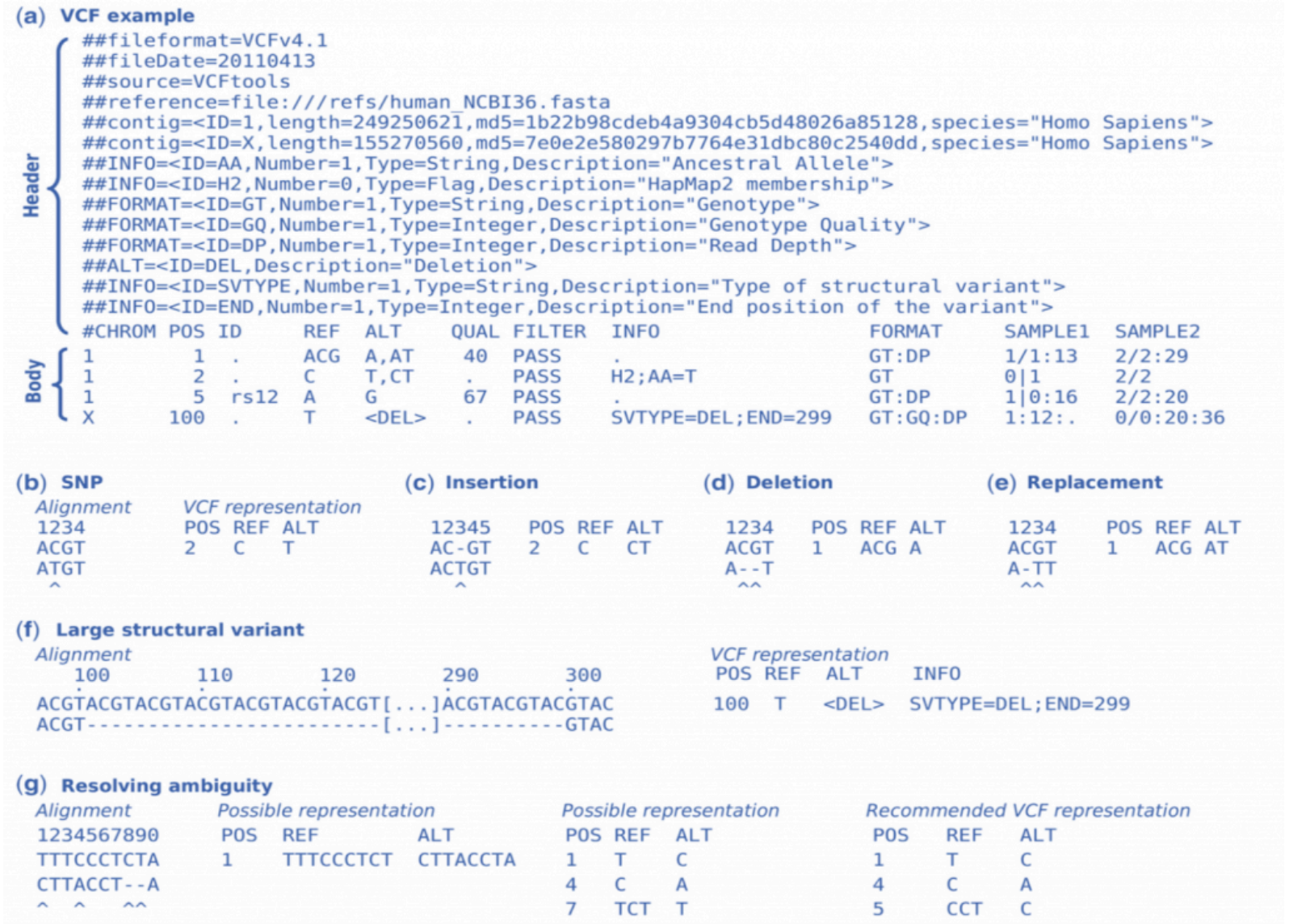
Trapnell and Salzberg (2009) Nat Biotechnology 27:455

SNP calling workflows



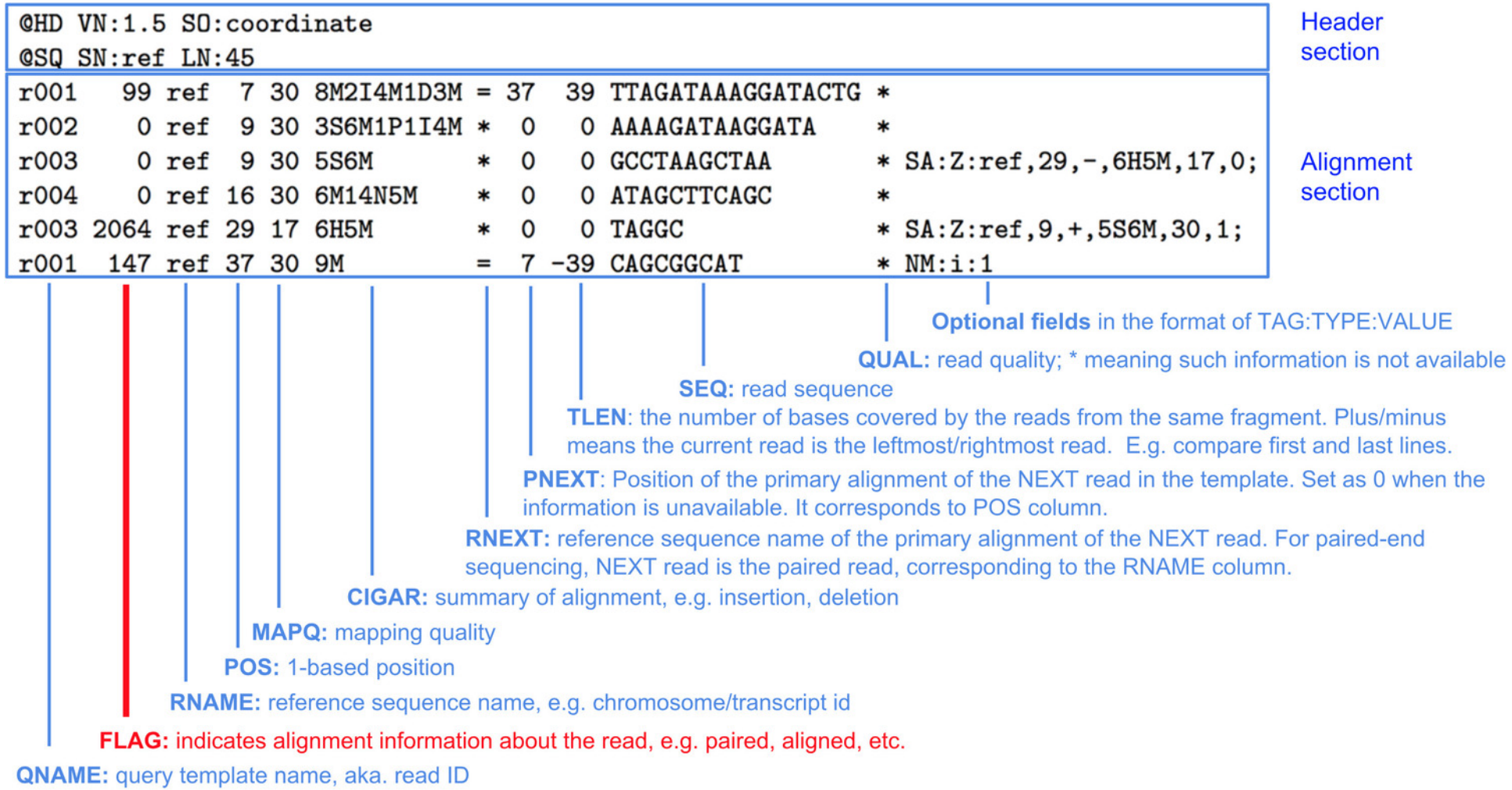
https://bcbio.wordpress.com/2014/10/07/joint-calling/

VCF file format (Variant Call Format)



https://learn.gencore.bio.nyu.edu/ngs-file-formats/vcf-format/

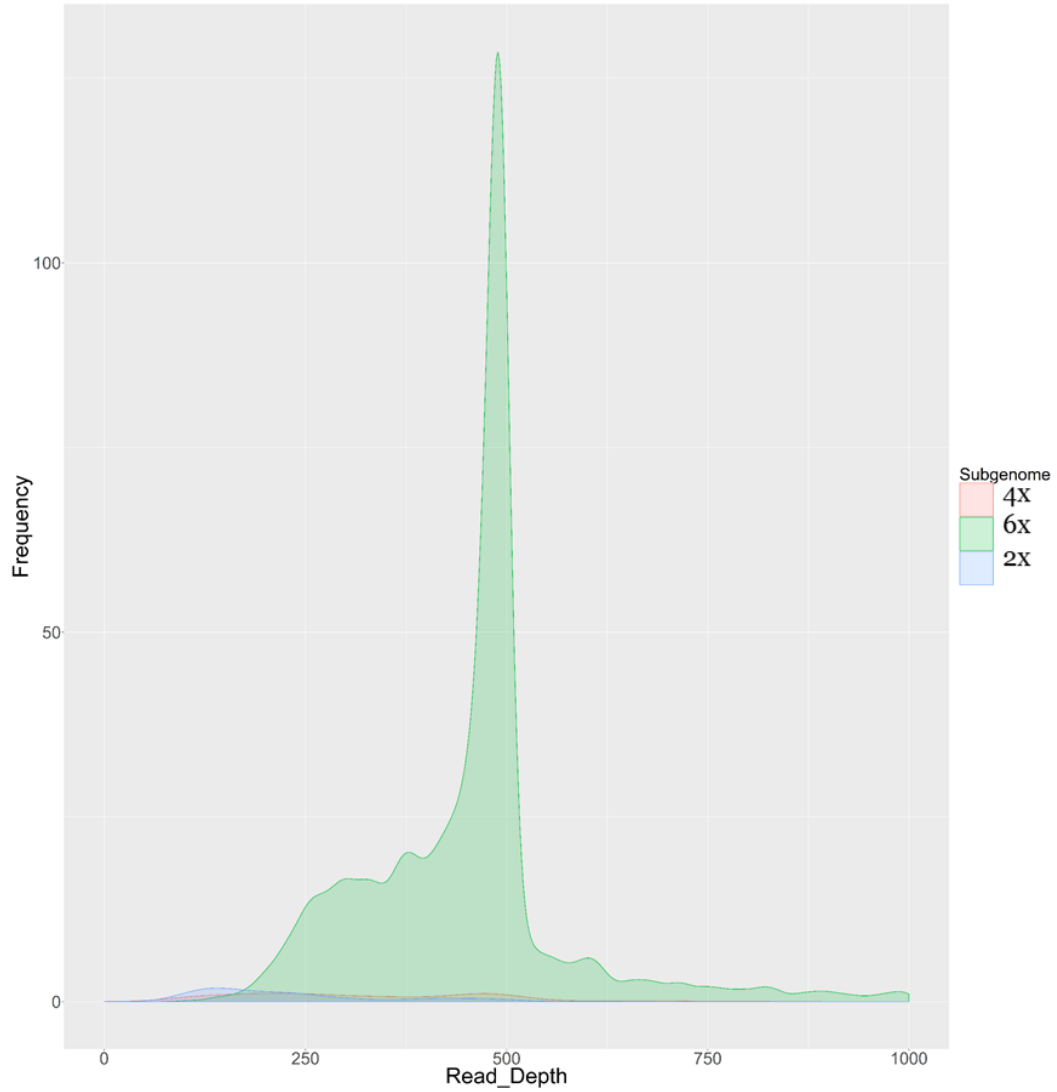
SAM file format (Sequence Alignment Map)

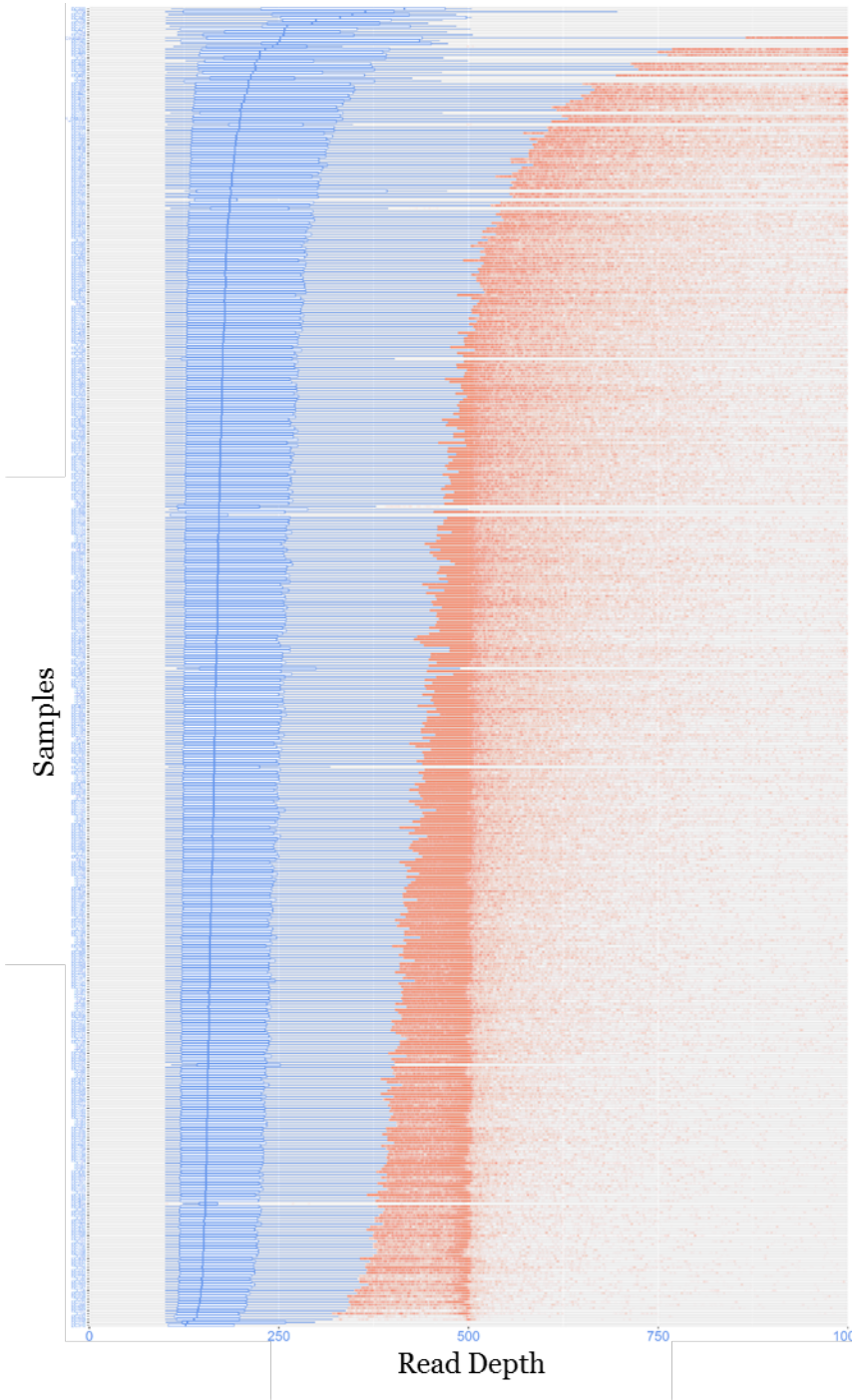


BAM: Binary format/compression version of SAM

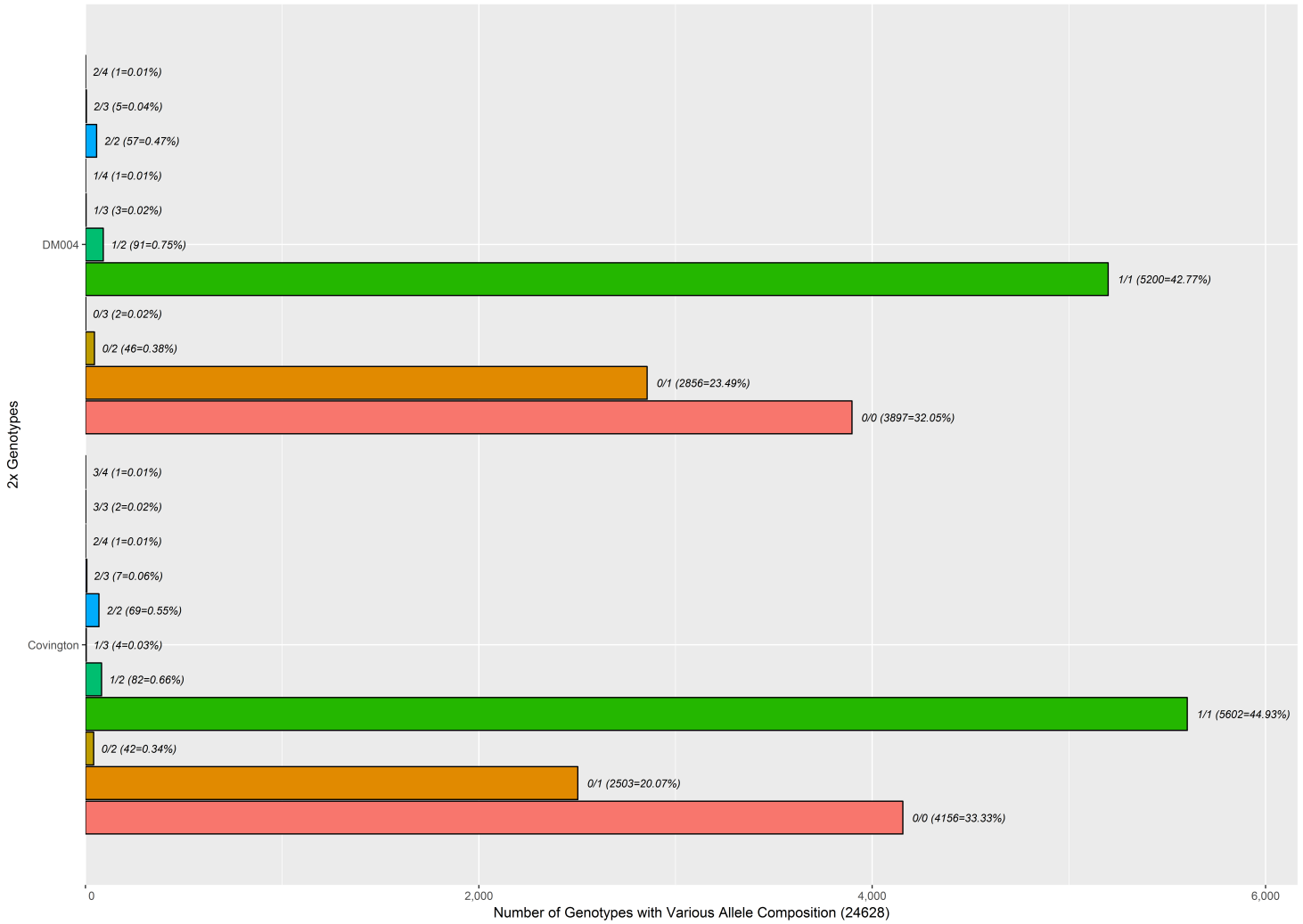
https://learn.gencore.bio.nyu.edu/ngs-file-formats/vcf-format/

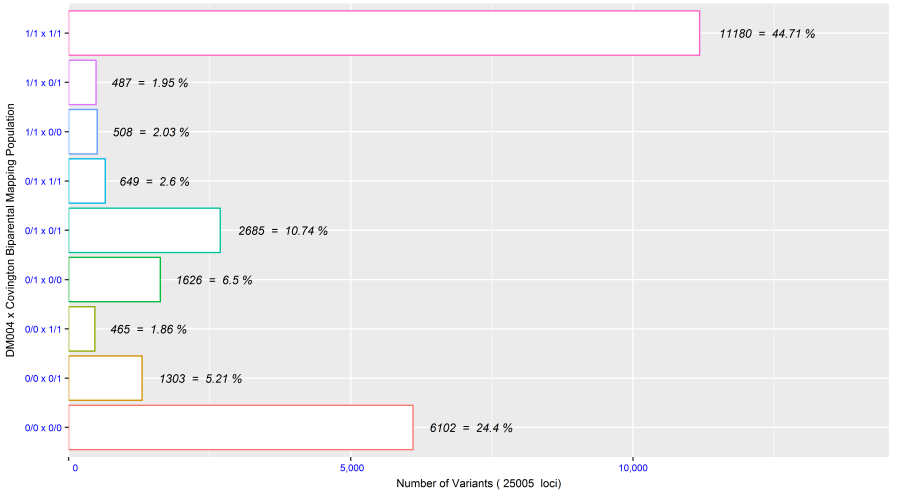
Uniformity of coverage across loci, samples

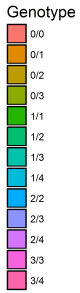




Distribution of multi-allelic and bi-allelic markers

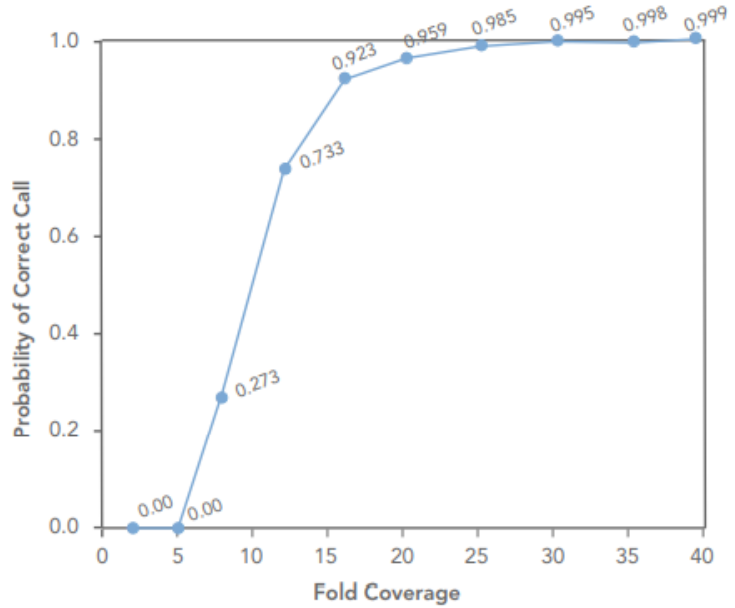


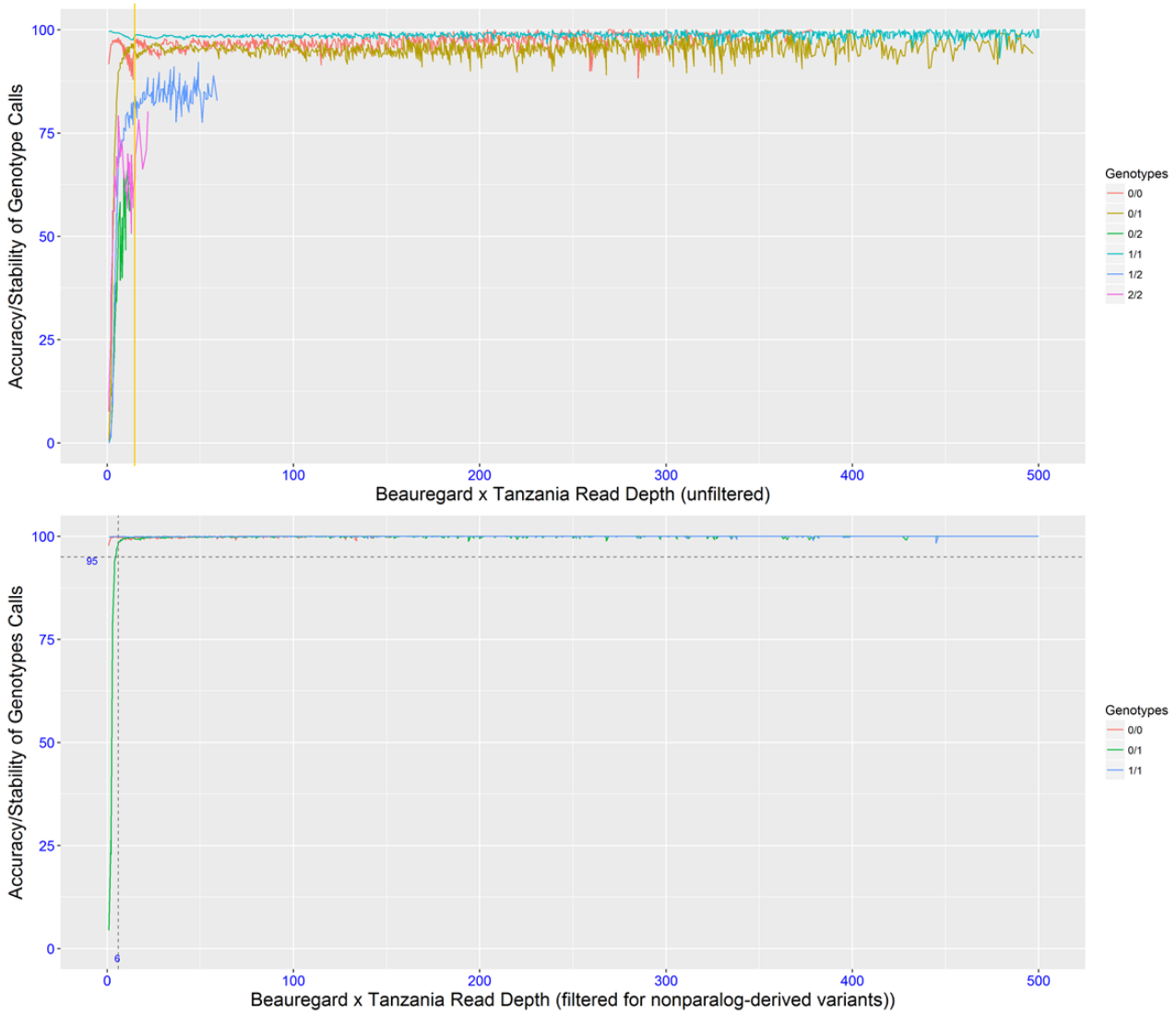




Read depth and probability of correct SNP calls (diploid)

Figure 3: Probability of Correct SNP Call
Calculation of the probability of a correct SNP call at different coverage levels for a theoretical heterozygote position. The quality of the base calls was assumed at Q30.

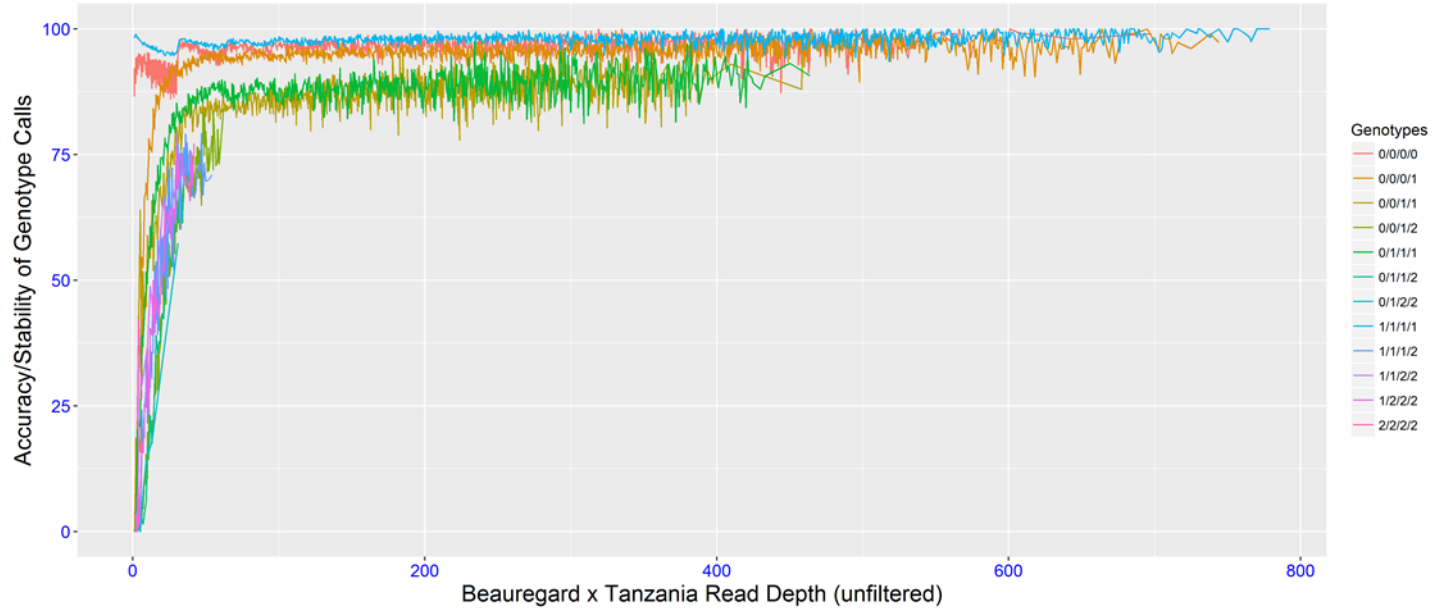




Removing SNP derived from paralogs

Read depth and probability of correct SNP calls (tetraploid)

Removing SNP derived from paralogs



Read depth and probability of correct SNP calls (hexaploid)

Removing SNP derived from paralogs