

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

EPP 622  
Bioinformatics Applications

Bode Olukolu  
Assistant Professor  
Entomology and Plant Pathology

# Outline

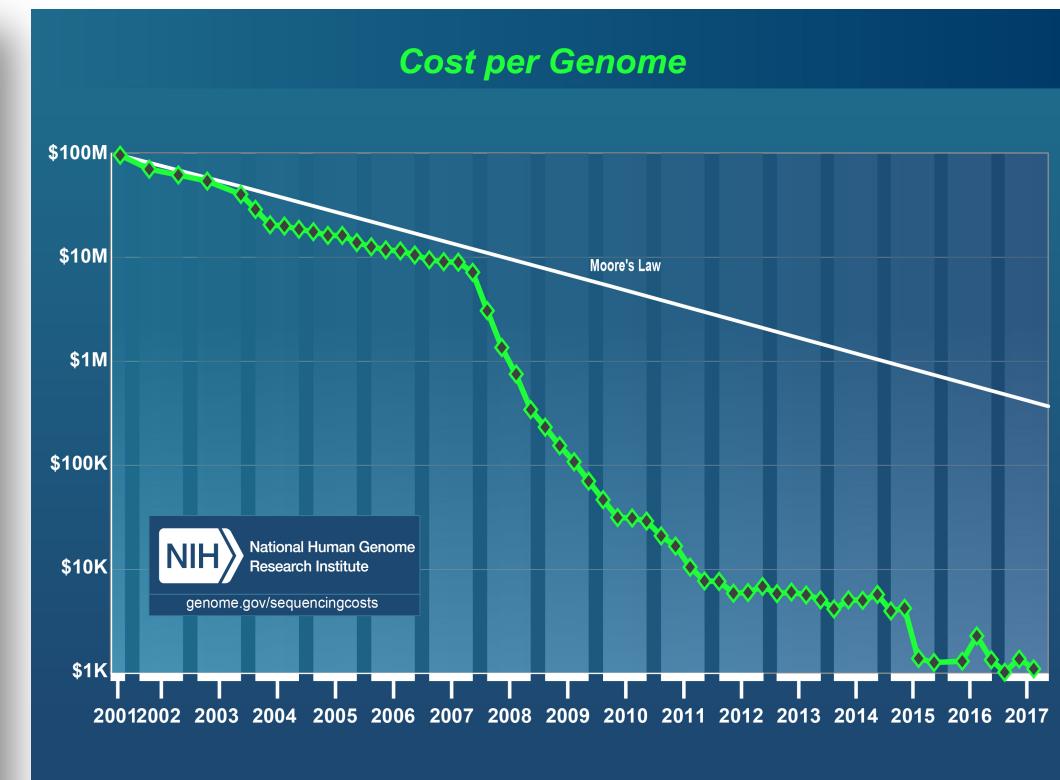
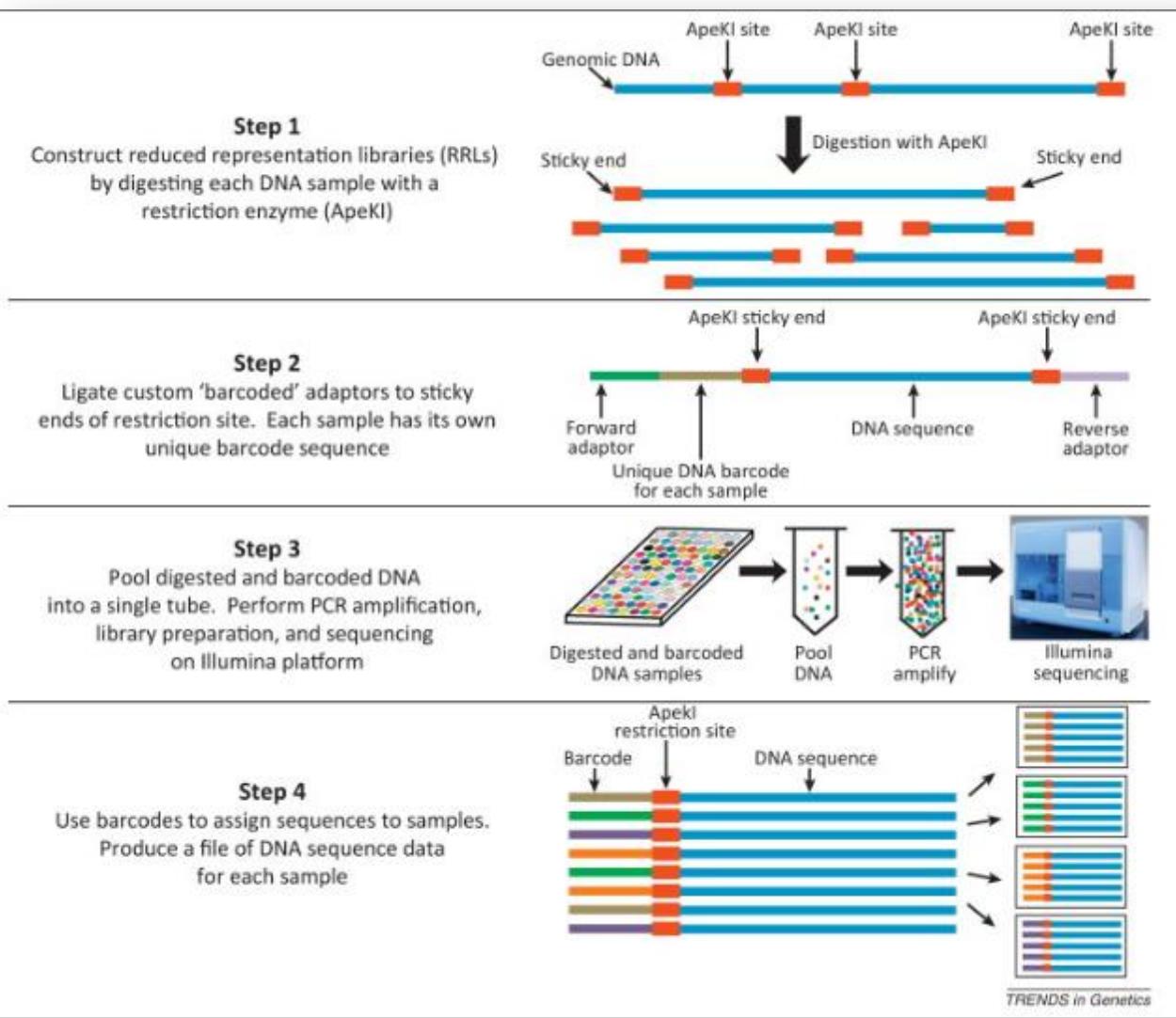
- Next-Generation Sequencing and Quality Controls (part 1)
  - Understanding multiplexed NGS libraries.
  - Base calling and quality scores.
  - Re-calibration of per-base quality scores.
- Variant calling and filtering workflows (Part 2)
  - Alignment (reference genome/assembly, assembly errors, paralogs...).
  - 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

# 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.

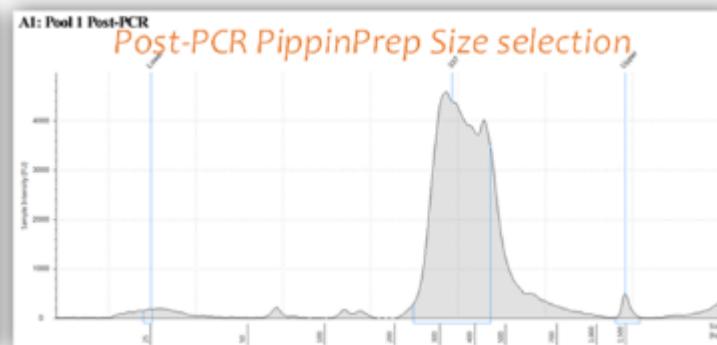
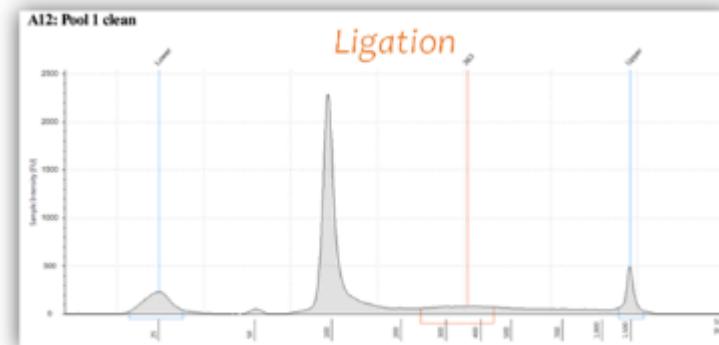
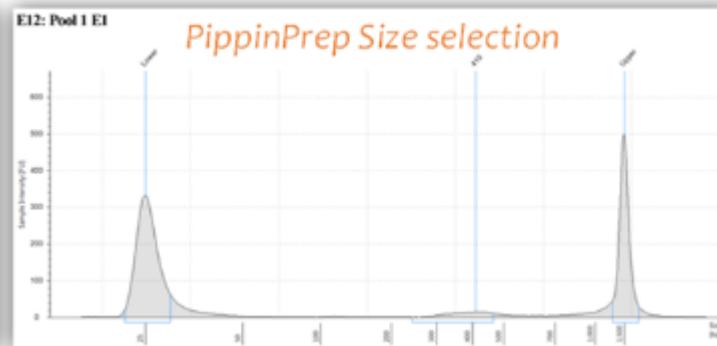
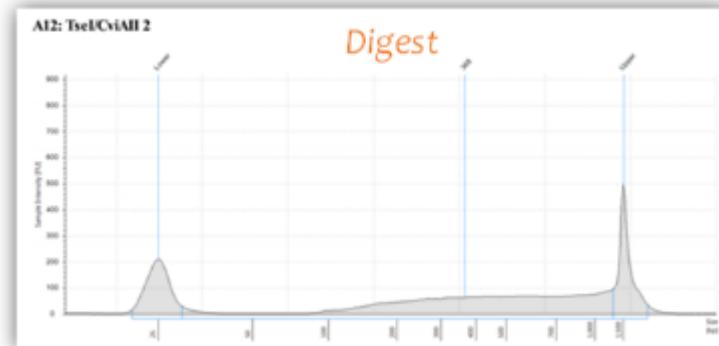
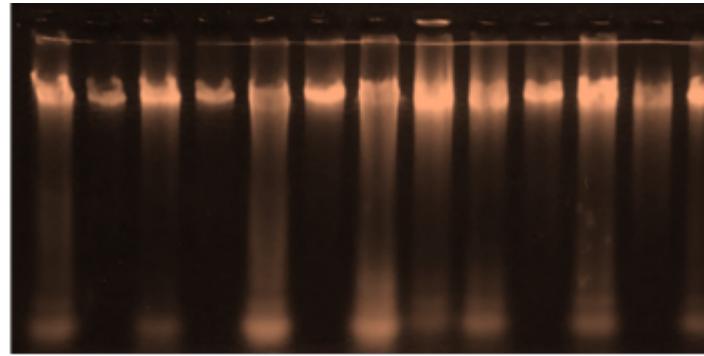
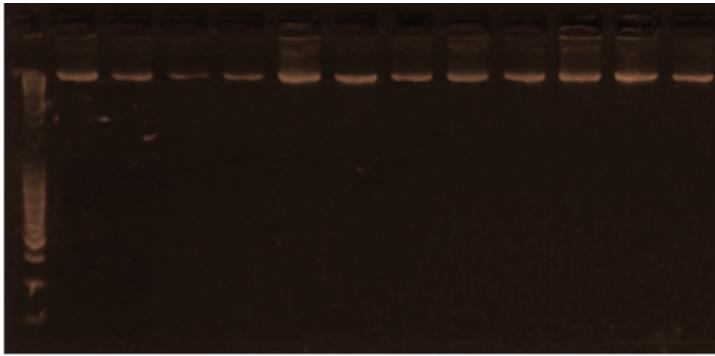
*Nonetheless, GBS assay is now commonly used and possibly the future of high-throughput genotyping.*

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

		Description	DNA sequences		Aligned?	Called genotype	True genotype
Sample	Reference genome		NGS read	No NGS read			
A	Ideal	CTGC	C		✓	CT	CT
		CTGC	T		✓		
B	Heterozygous for SNP in restriction site	CTAC	C	□	✗	TT	CT
		CTGC	T		✓		
C	Homozygous for SNP in restriction site	CTAC	C	□	✗	-	CC
		CTAC	C	□	✗		
D	Heterozygous for divergent sequence	CTGC	C		✓	CC	CT
		CTGC	T	■■■■	✗		
E	Homozygous for divergent sequence	CTGC	T		✗	-	TT
		CTGC	T	■■■■	✗		

Key: CTGC Restriction site      ■ NGS read      □ No NGS read      ■ Focal SNP      ■ Mismatch to ref genome

# 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

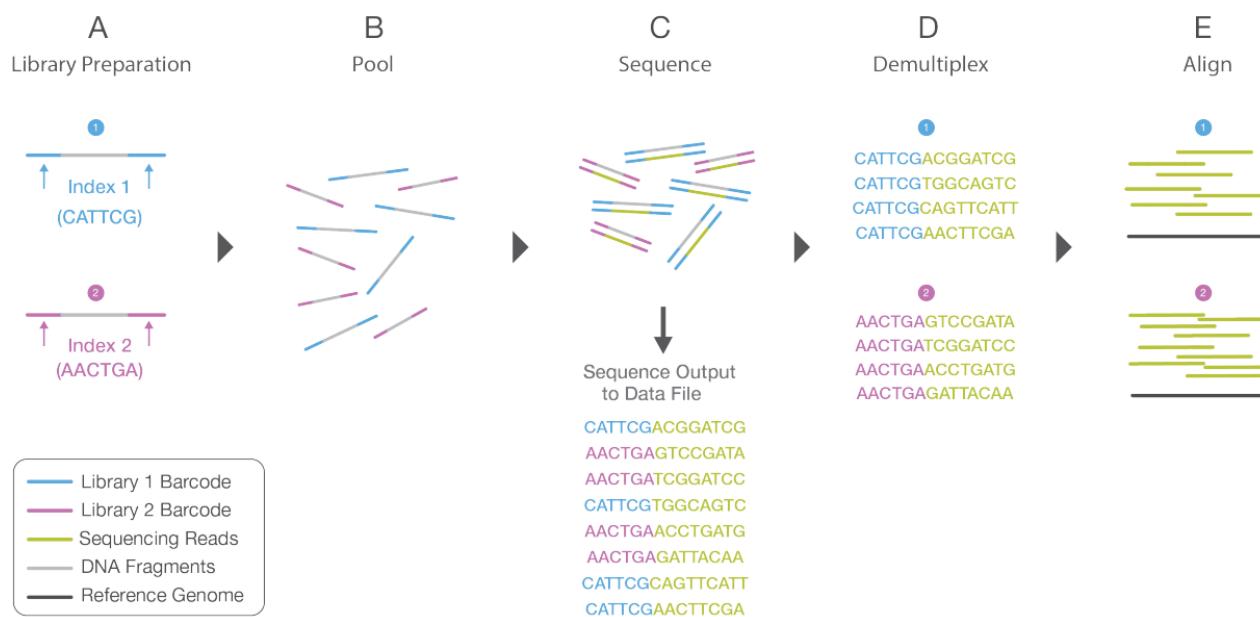
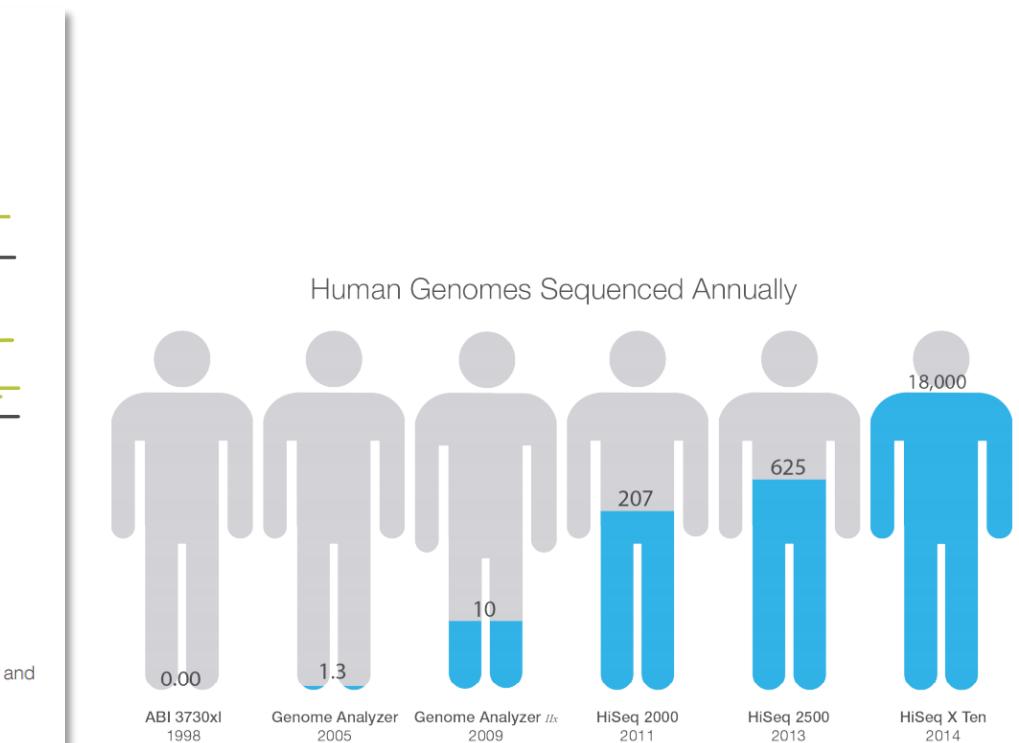
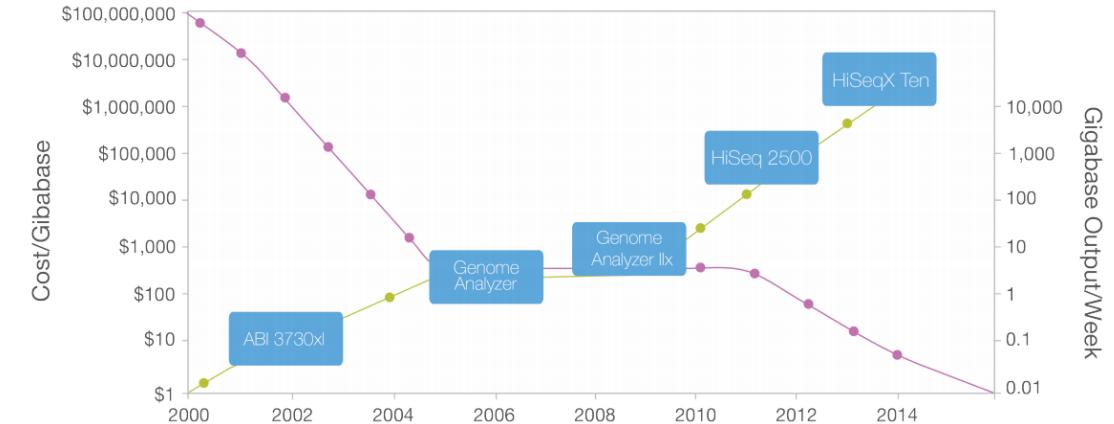
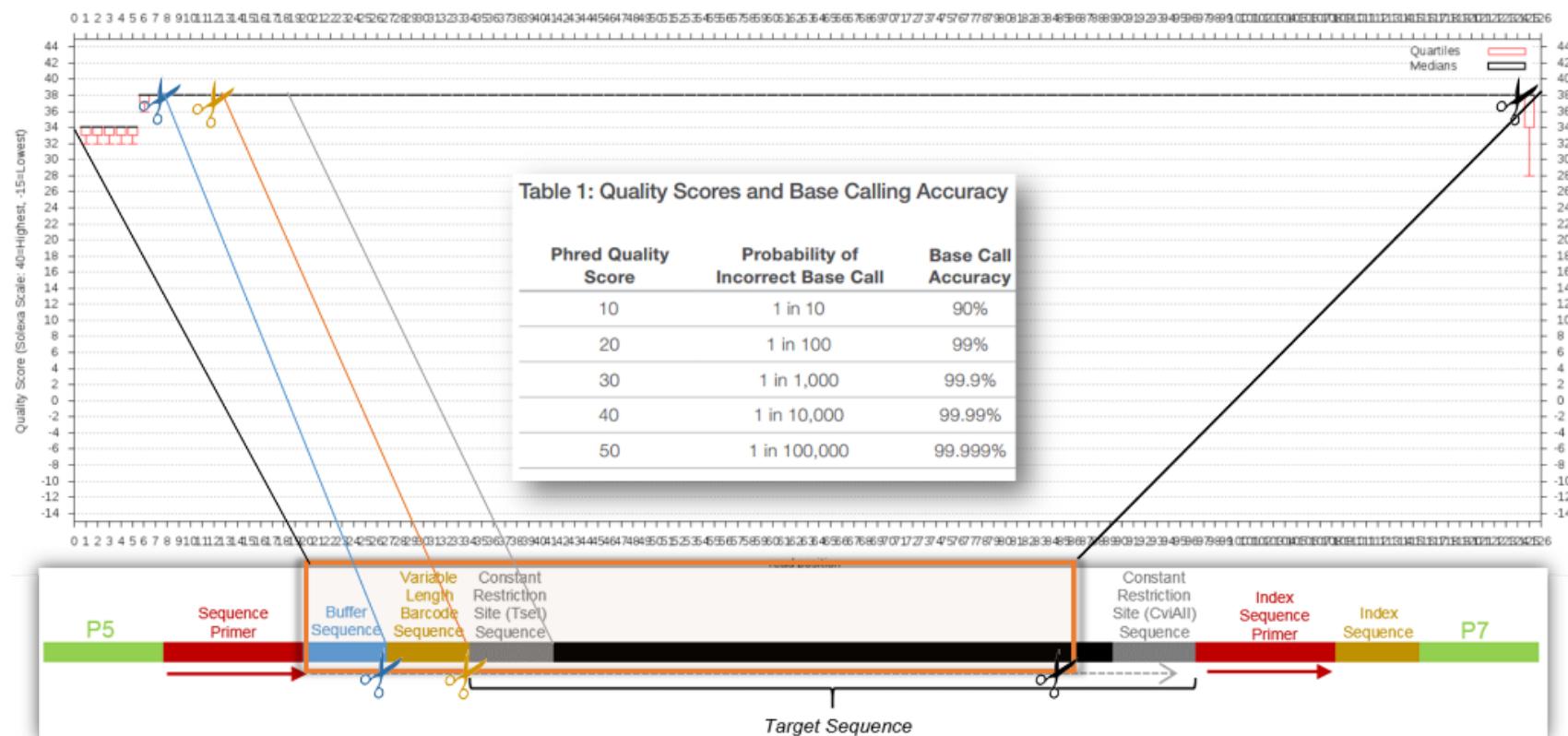
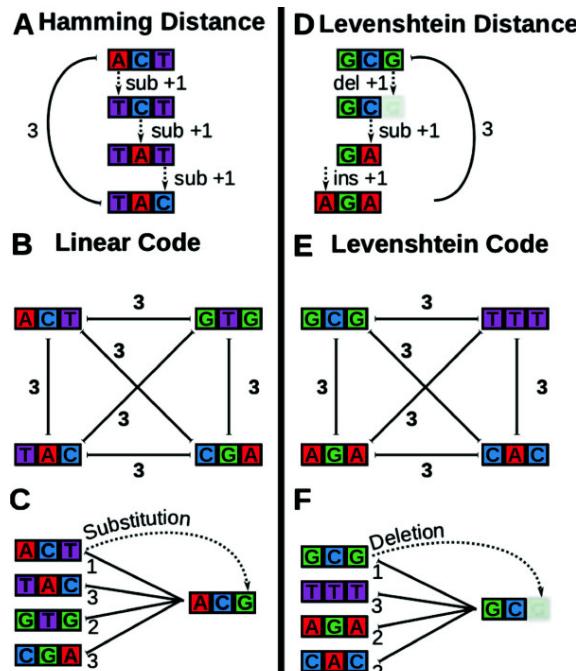


Figure 5: Library Multiplexing Overview—(A) Unique index sequences are added to two different libraries during library preparation. (B) Libraries are pooled together and loaded into the same flow cell lane. (C) Libraries are sequenced together during a single instrument run. All sequences are exported 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.



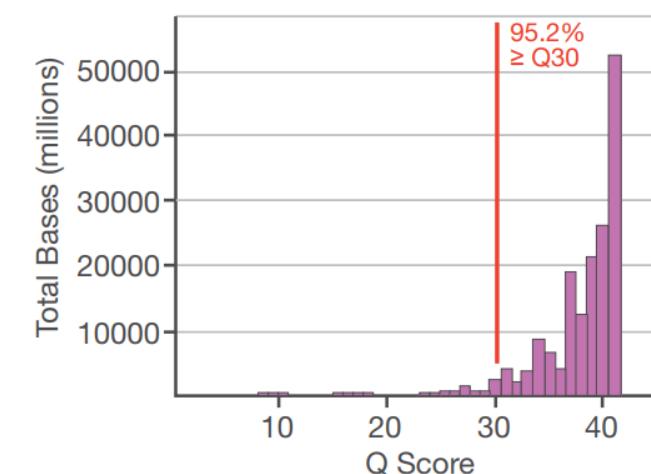
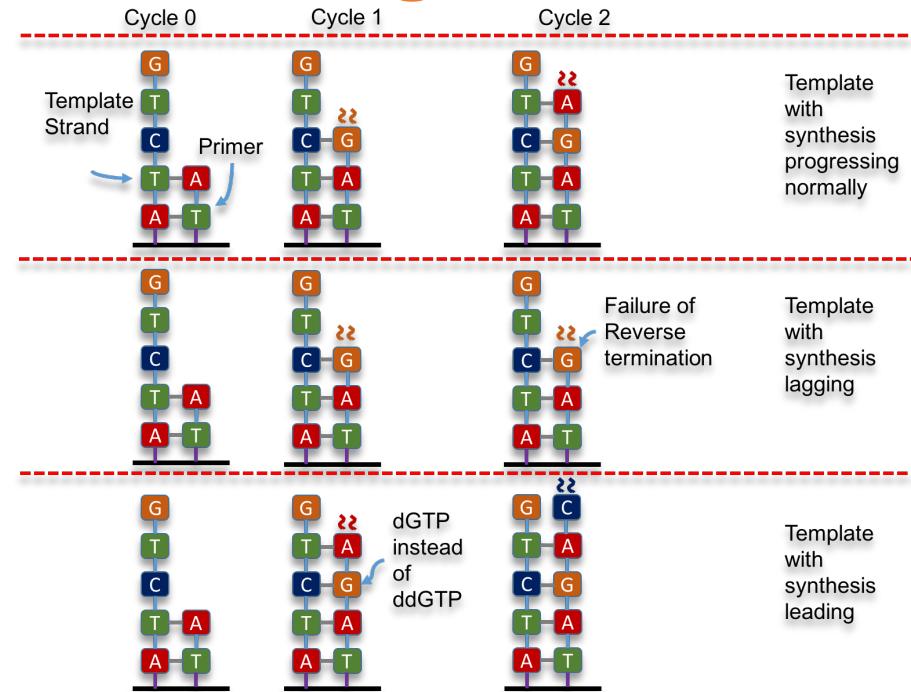
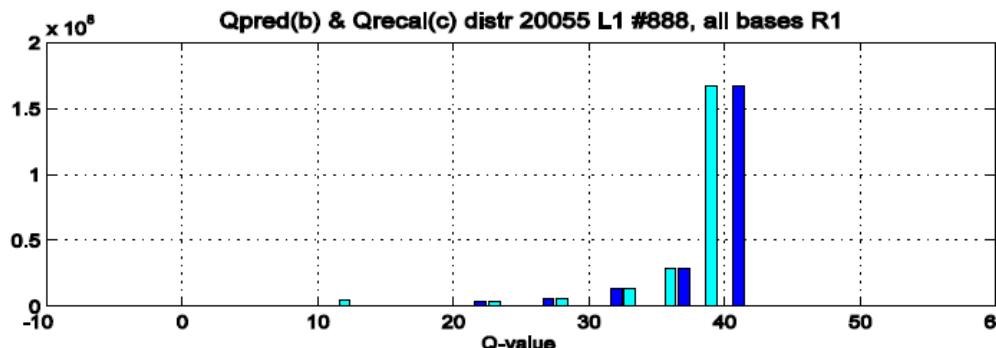
# Considerations for multiplex-sequencing

- ✓ 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



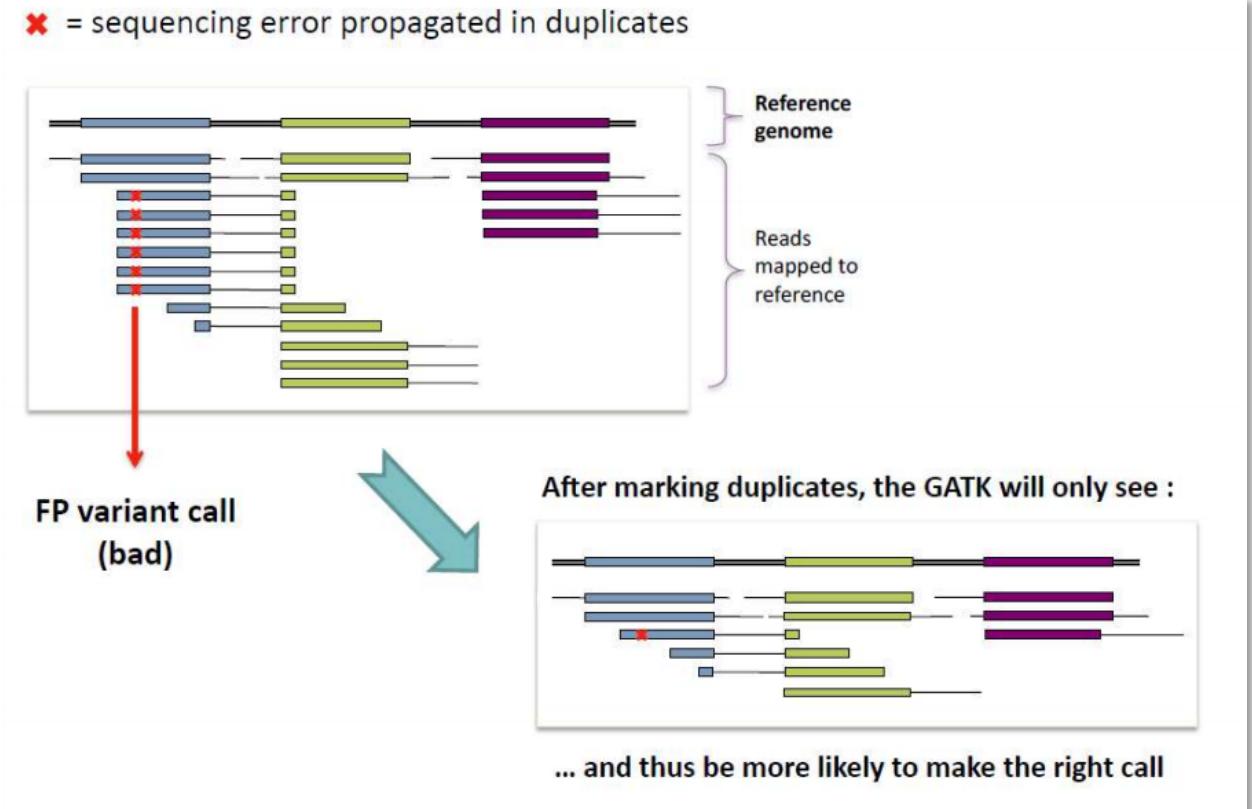
# Illumina sequencing: sources of base calling error

- ✓ 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.



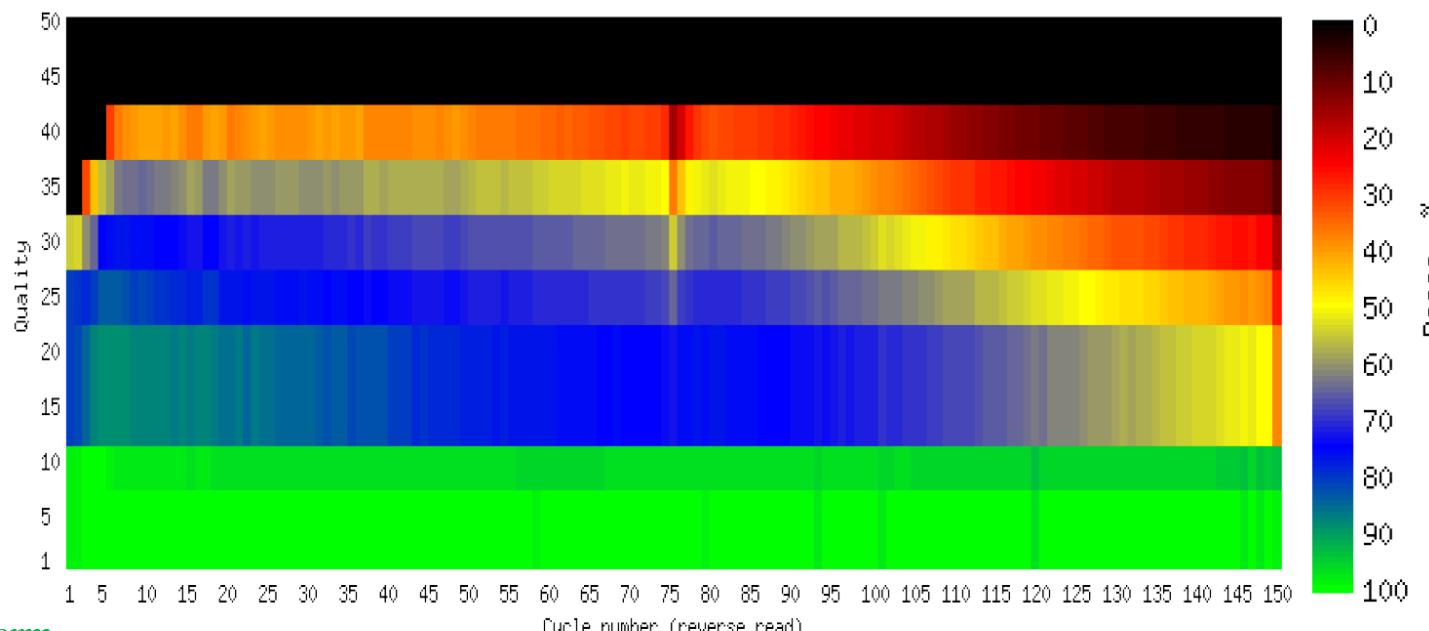
# Removing PCR/optical duplicates

- ✓ 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



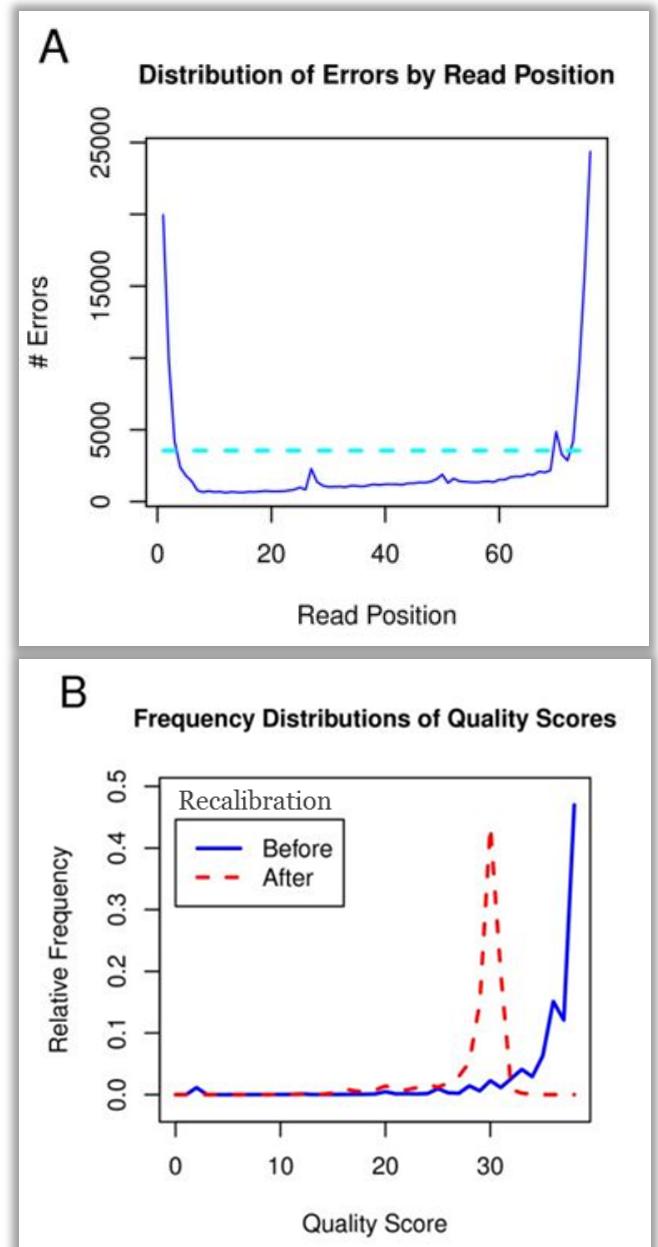
# Base calling and quality scores

- ✓ 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 minuscule, but retaining them in reads lead to non-existent SNP calls. This account for the inflated minor allele frequency observed in most GBs data.

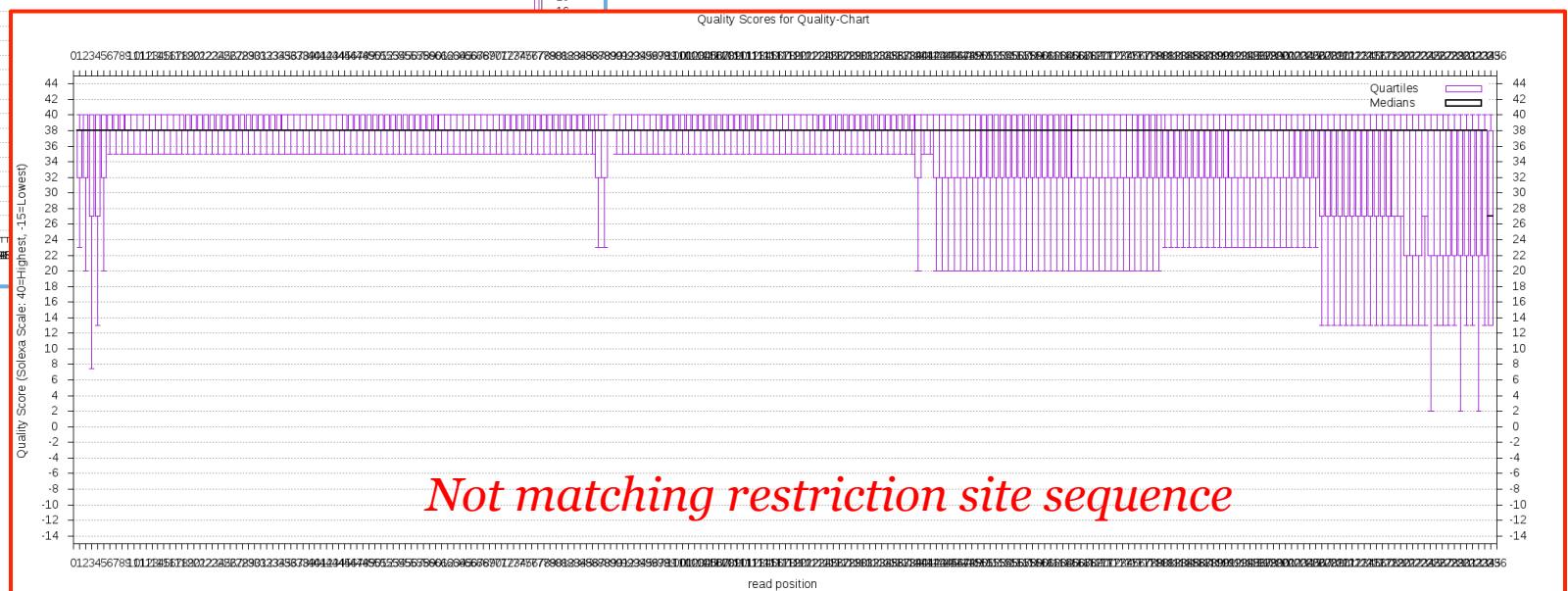
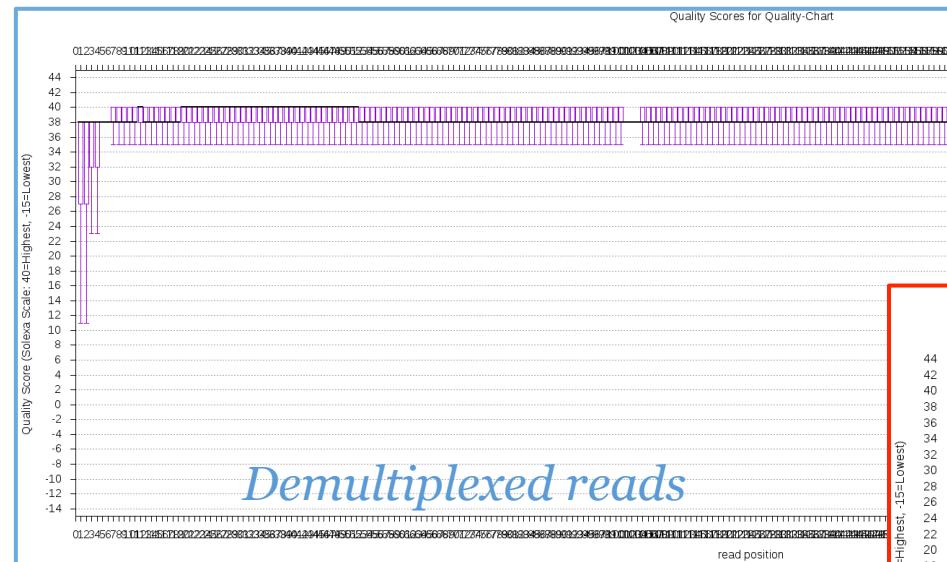


# Re-calibration of per-base quality scores

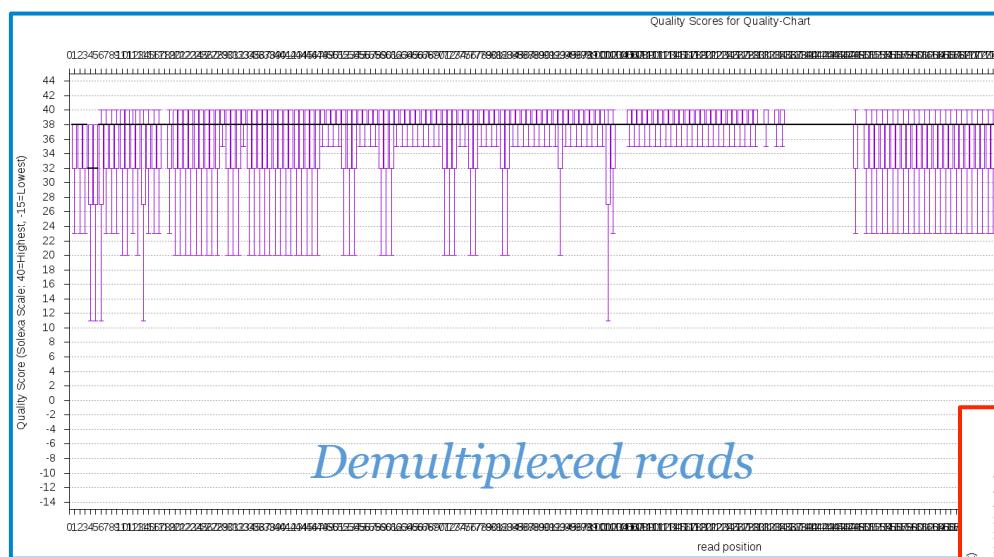
- ✓ Raw Phred-scaled quality scores produced from base-calling algorithms may not accurately reflect the true base-calling error rates.
- ✓ Hence, need for recalibration.
- ✓ **SOAPsnp:** recalibrated by comparing a sequenced genome to the reference genome at sites with no known SNPs. Algorithm estimates 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



# Error filtering: Forward Read



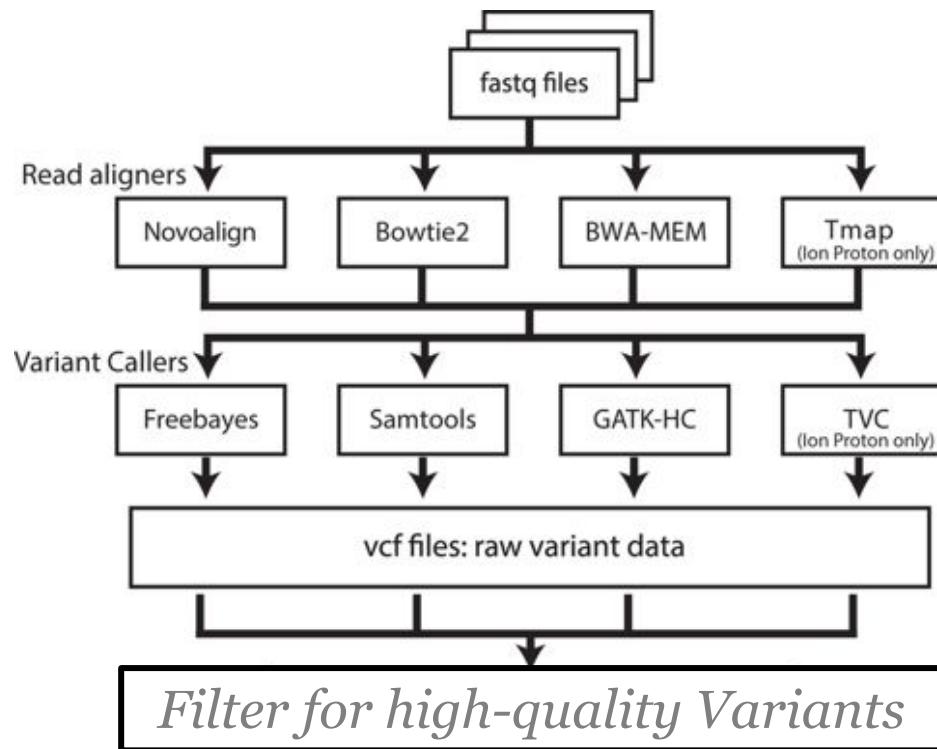
# Error filtering : Reverse Read



# 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

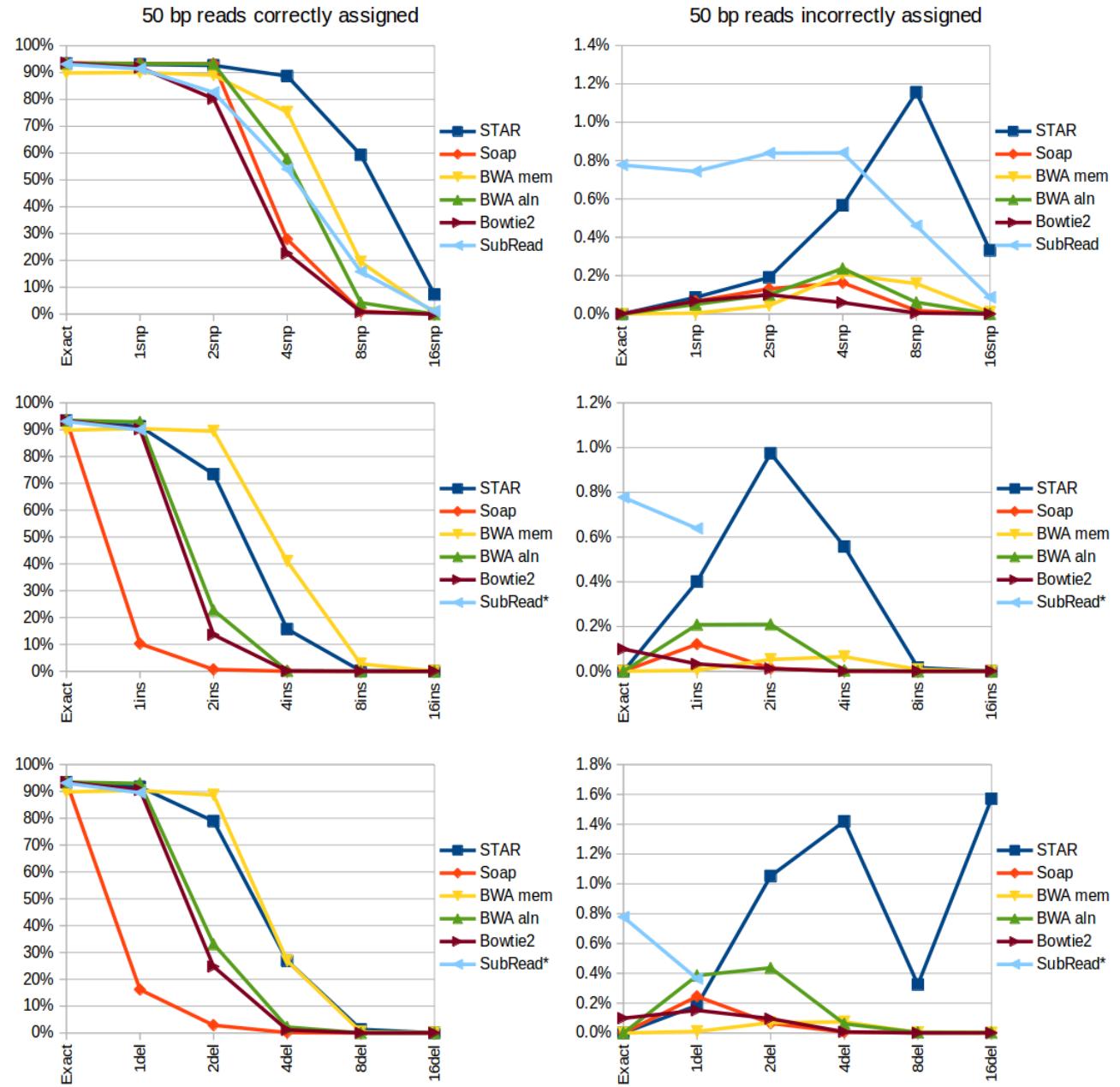


# Read alignment

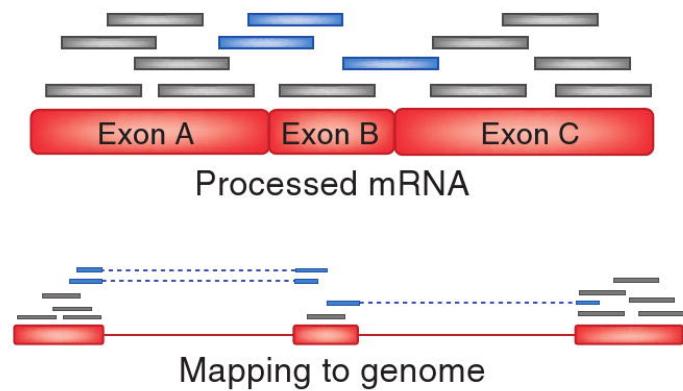
- ✓ Samples can be divergent from reference genome, even among the same species.
- ✓ Paralogs or duplicate sequences can lead to error if reads from multiple loci pile up at one locus
- ✓ In polyploids, it is important to differentiate between “subgenome-specific sequences” and “common sequences across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

	Hexaploid <i>BxT</i> <sup>1</sup>	Diploid <i>M9xM19</i> <sup>1</sup>
(%)	(%)	
After de-multiplexing pooled samples	94.6	97.3
Reads matching <i>I. trifida</i> nuclei genome <sup>2</sup>	87.5	90.8
Reads matching and unique to <i>I. triloba</i>	6.3	-

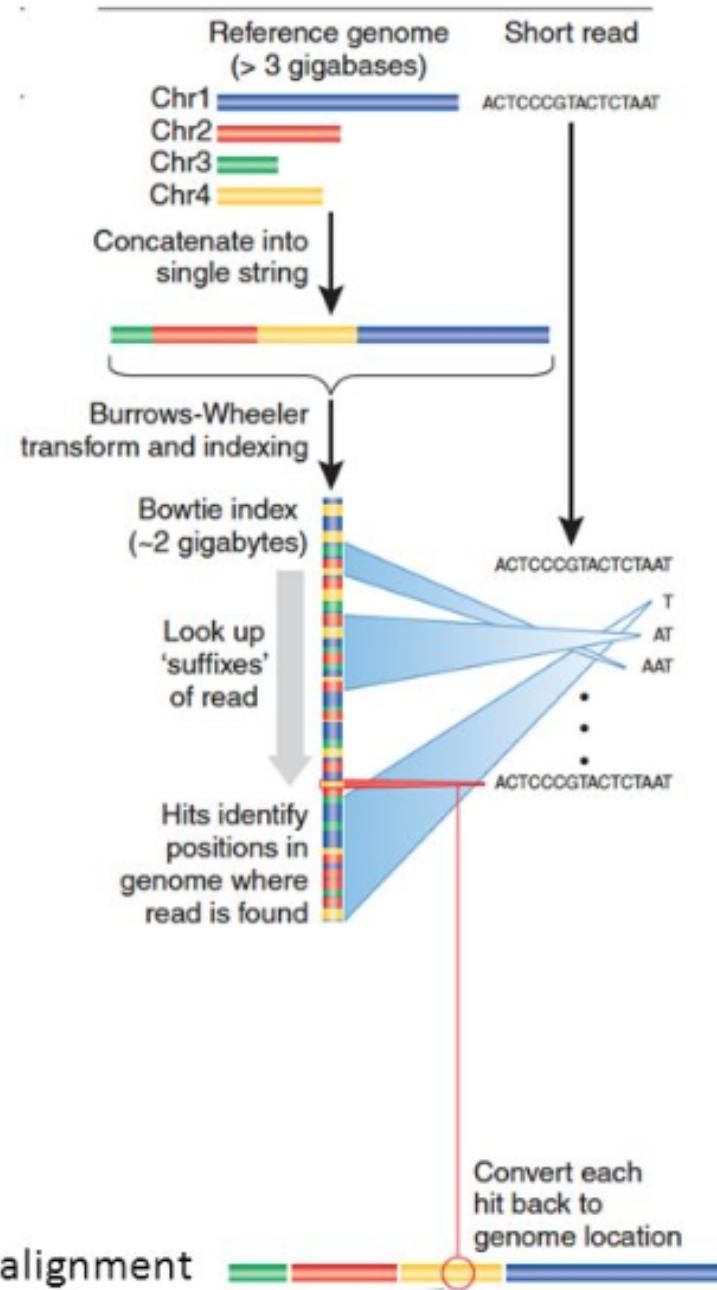
→ ~5% *PhiX*



# Why Genome Indexing?

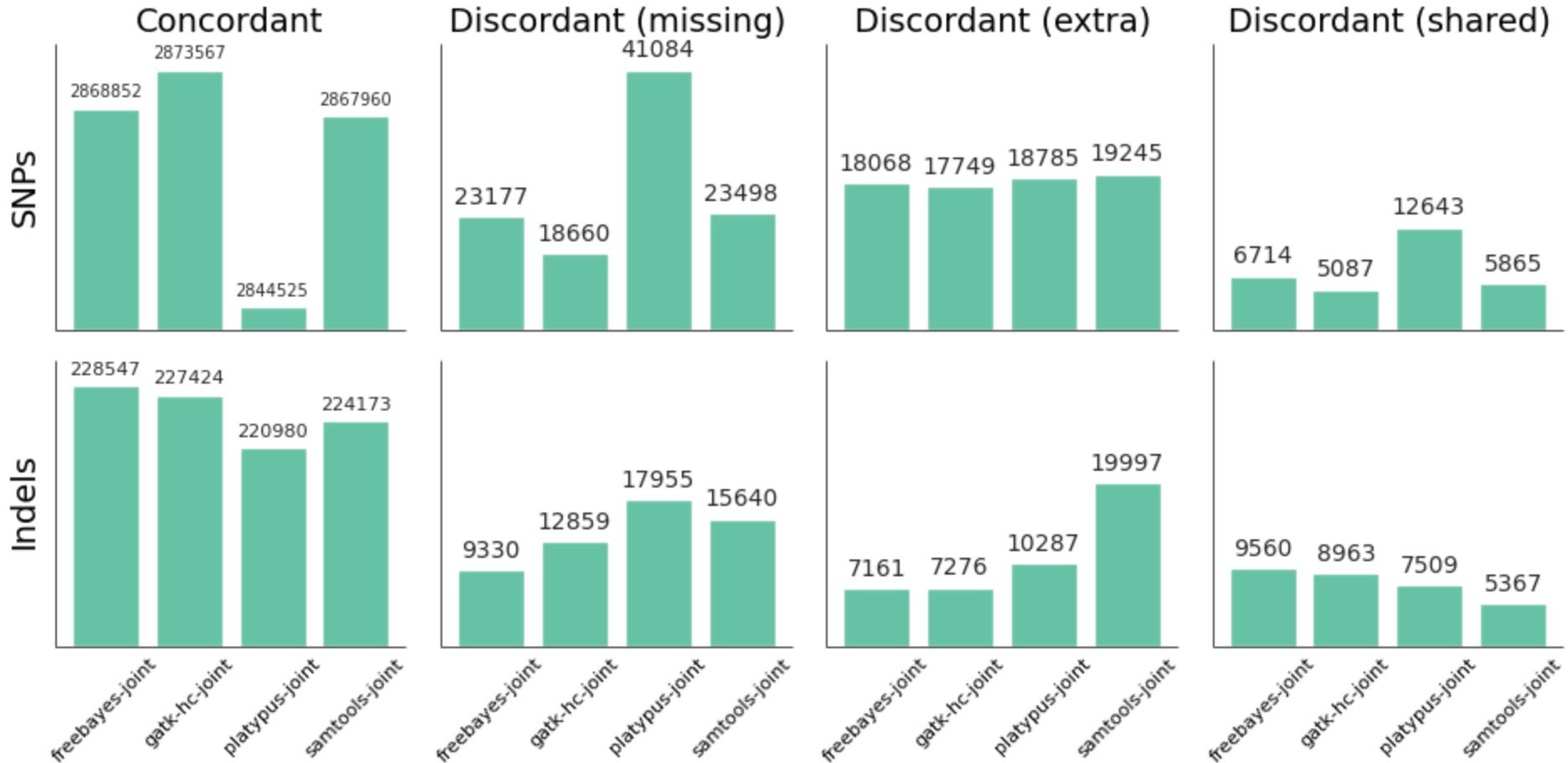


- ✓ Simplifies genome sequence
- ✓ Results in indexed genome i.e. hits/matches identify positions in genome where reads are found.
- ✓ Very fast and efficient.



# SNP calling workflows

Incremental joint calling: GATK HaplotypeCaller, FreeBayes, Platypus and samtools



# VCF file format (Variant Call Format)

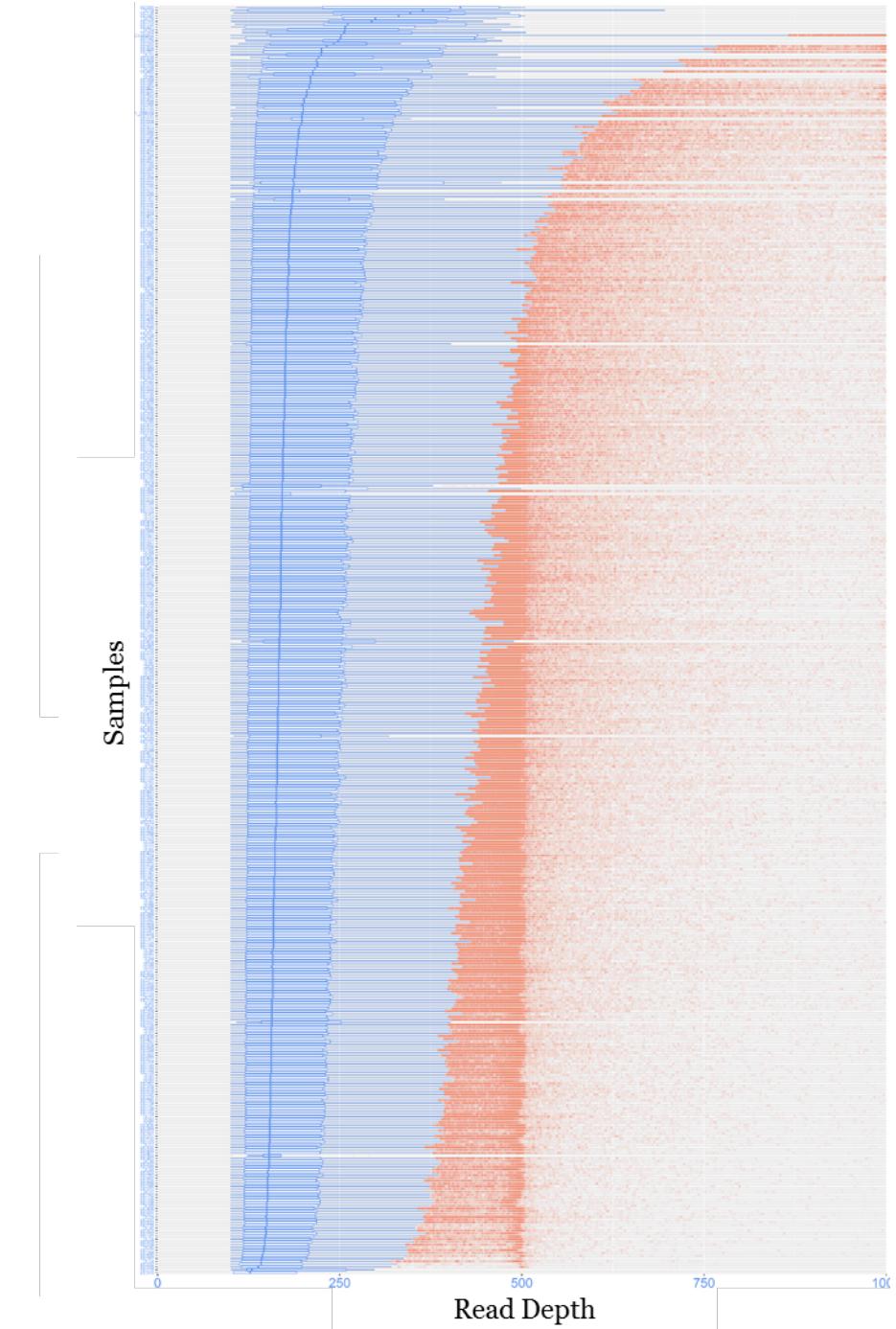
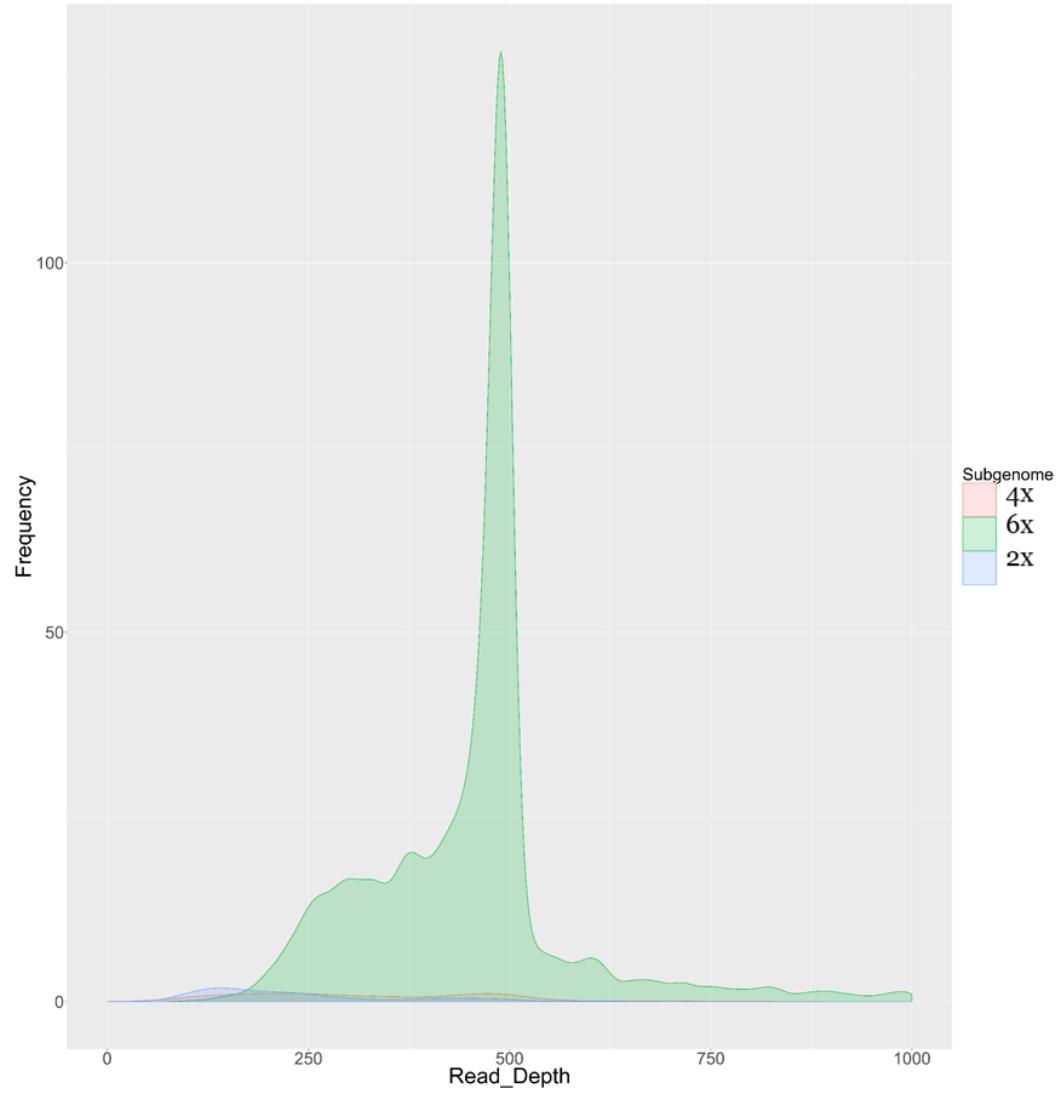
(a) VCF example												
Header												
##fileformat=VCFv4.1 ##fileDate=20110413 ##source=VCFtools ##reference=file:///refs/human_NCBI36.fasta ##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens"> ##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo 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">												
#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			(c) Insertion			(d) Deletion			(e) Replacement			
Alignment		VCF representation			12345	POS	REF	ALT	1234	POS	REF	ALT
1234		POS	REF	ALT	12345	POS	REF	ALT	1234	POS	REF	ALT
ACGT		2	C	T	AC-GT	2	C	CT	ACGT	1	ACG	A
ATGT					ACTGT				A-T			
^					^				^^			
(f) Large structural variant						VCF representation						
Alignment			VCF representation			POS	REF	ALT	INFO			
100	110	120	290	300		100	T	<DEL>	SVTYPE=DEL;END=299			
ACGTACGTACGTACGTACGTACGT[...]												
ACGT-----	[...]	-----GTAC										
(g) Resolving ambiguity												
Alignment		Possible representation			Possible representation			Recommended VCF representation				
1234567890		POS	REF	ALT	POS	REF	ALT	POS	REF	ALT		
TTTCCCTCTA	1	TTTCCCTCT	CTTACCTA		1	T	C	1	T	C		
CTTACCT--A					4	C	A	4	C	A		
^ ^ ^					7	TCT	T	5	CCT	C		

# SAM file format (Sequence Alignment Map)

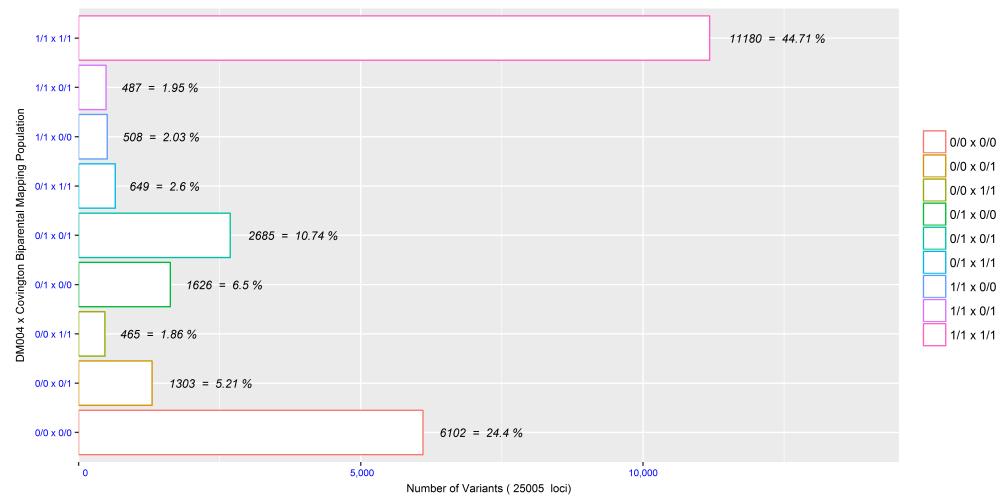
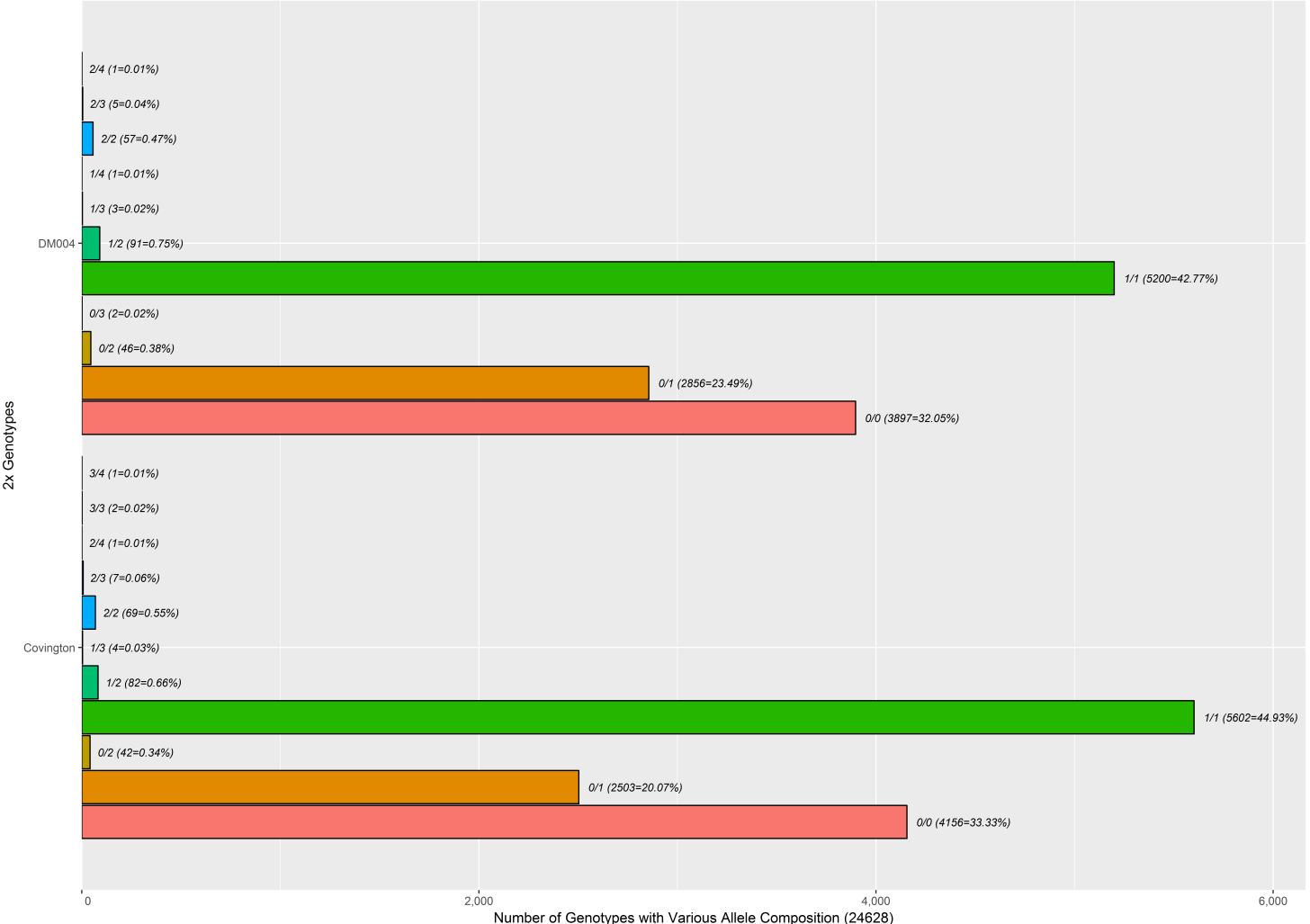
@HD VN:1.5 SO:coordinate												Header section
@SQ SN:ref LN:45												Alignment section
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAAGGATACTG *												
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *												
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;												
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *												
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;												
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1												
												Optional fields in the format of TAG:TYPE:VALUE
												QUAL: read quality; * meaning such information is not available
												SEQ: read sequence
												TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.
												PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.
												RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.
												CIGAR: summary of alignment, e.g. insertion, deletion
												MAPQ: mapping quality
												POS: 1-based position
												RNAME: reference sequence name, e.g. chromosome/transcript id
												FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.
												QNAME: query template name, aka. read ID

## BAM: Binary format/compression version of SAM

# Uniformity of coverage across loci, samples



# Distribution of multi-allelic and bi-allelic markers

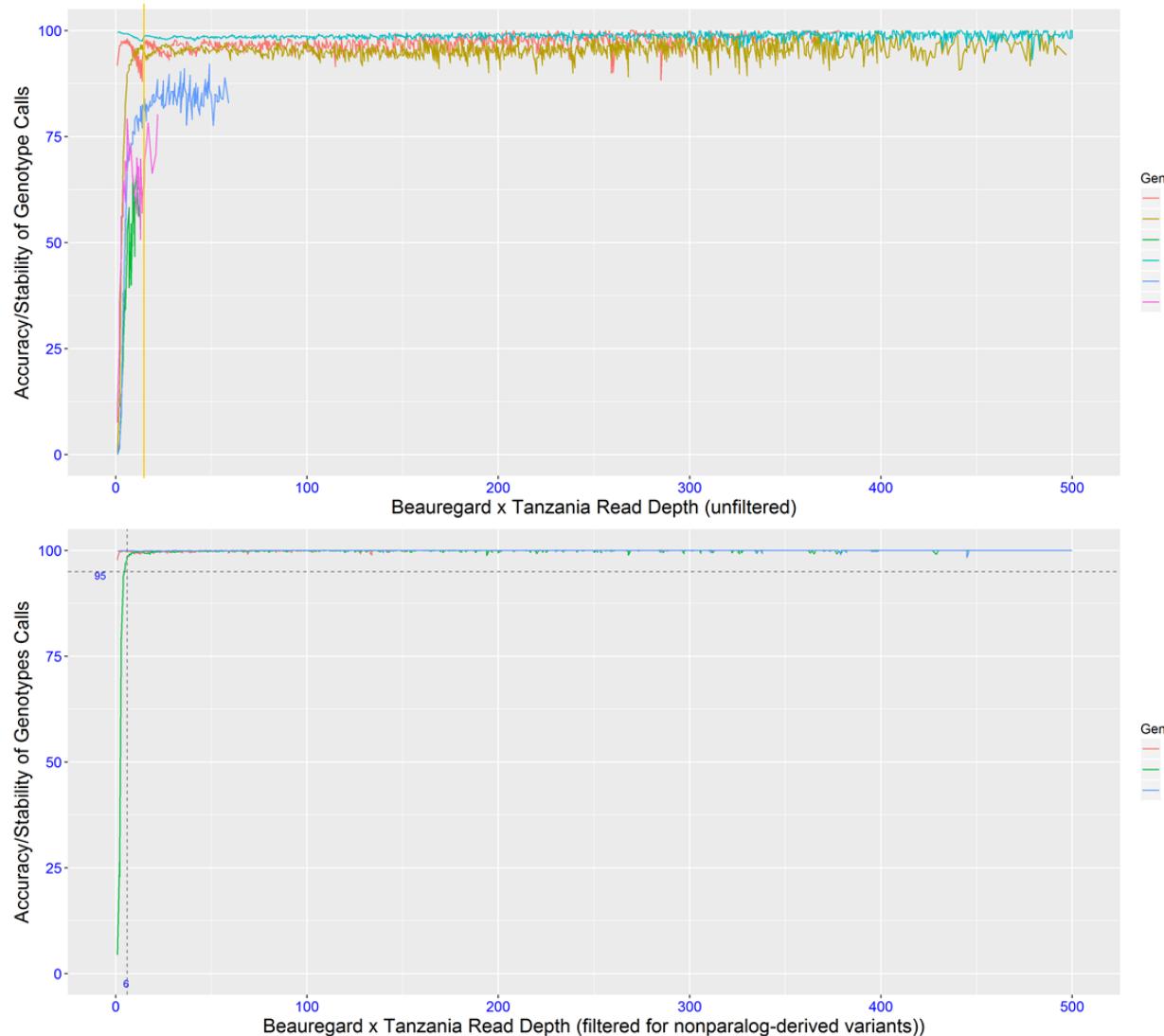
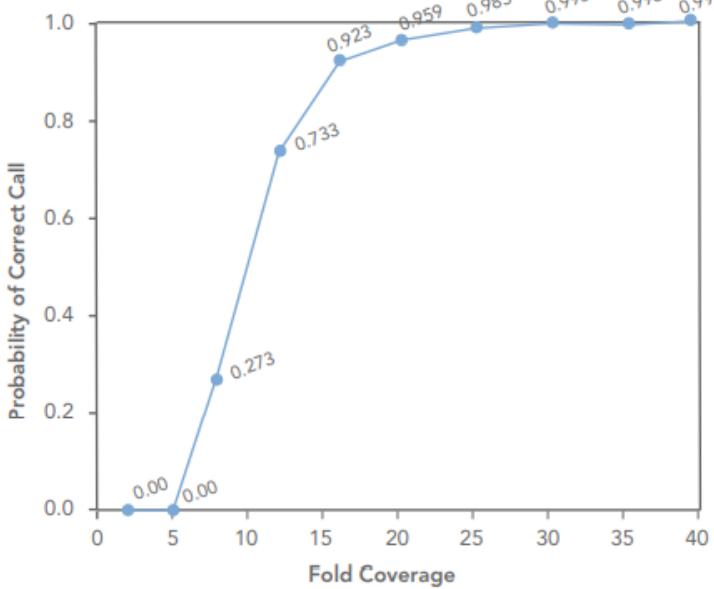


Genotype



# Read depth and probability of correct SNP calls (diploid)

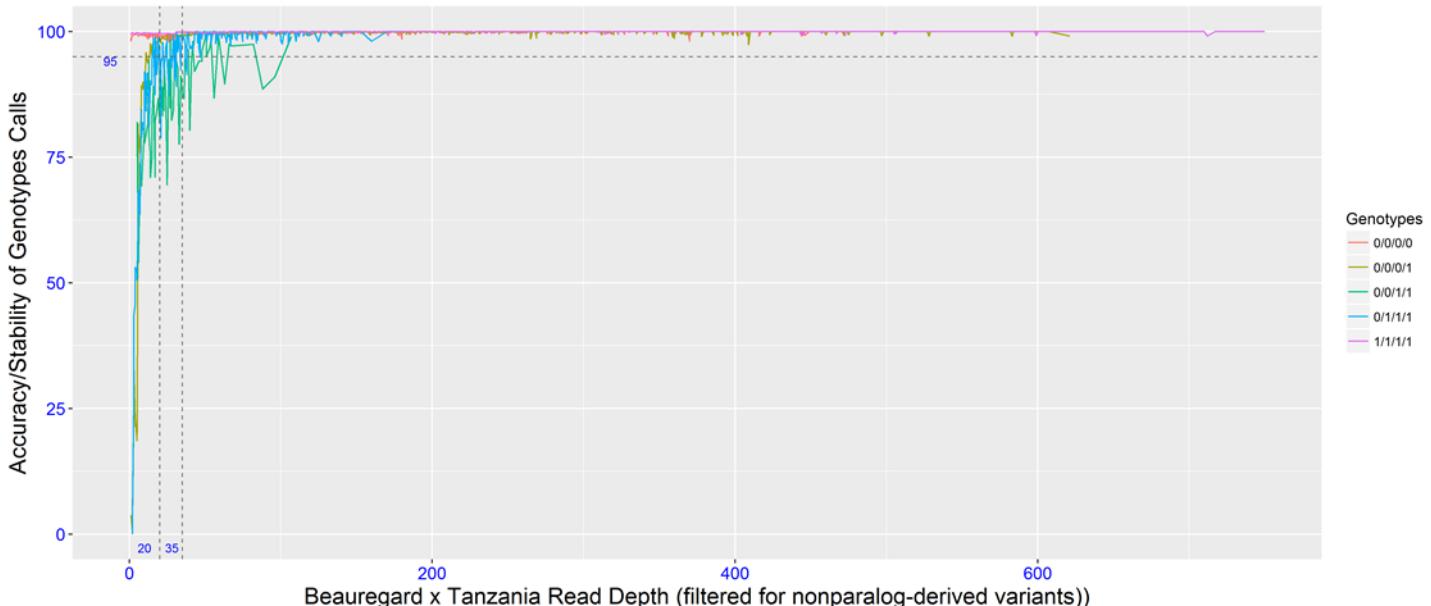
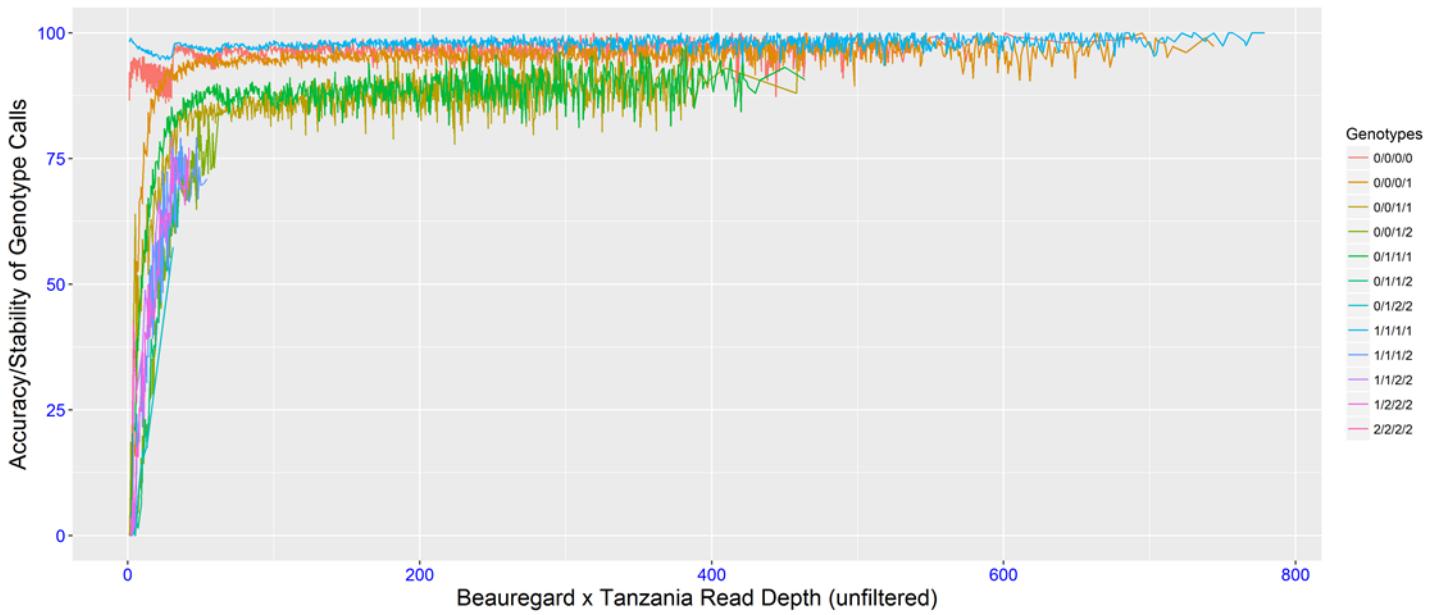
Figure 3: Probability of Correct SNP Call  
Calculation of the probability of a correct SNP call at different coverage levels for a theoretical heterozygote position. The quality of the base calls was assumed at Q30.



*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*

