

BIO634: variant calling

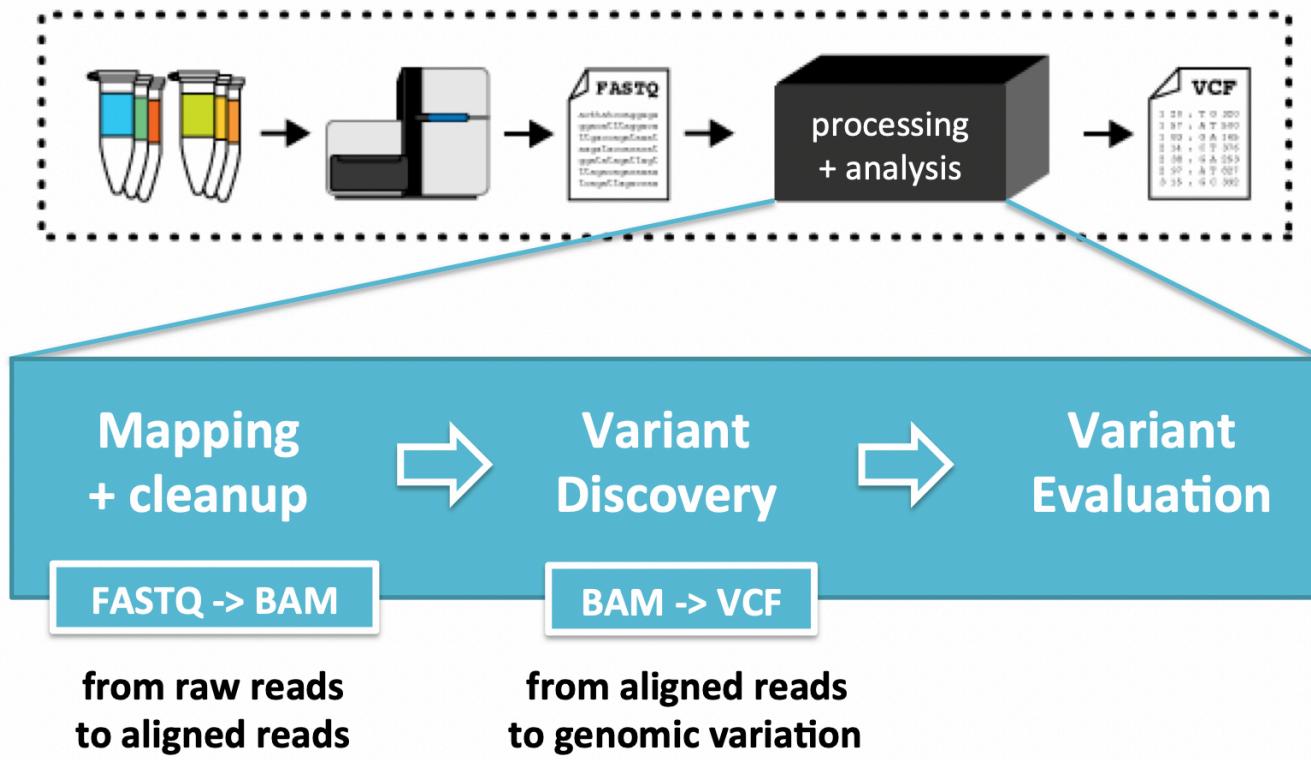


Adapted from **Stefan Wyder**
class on BIO634 (2018)



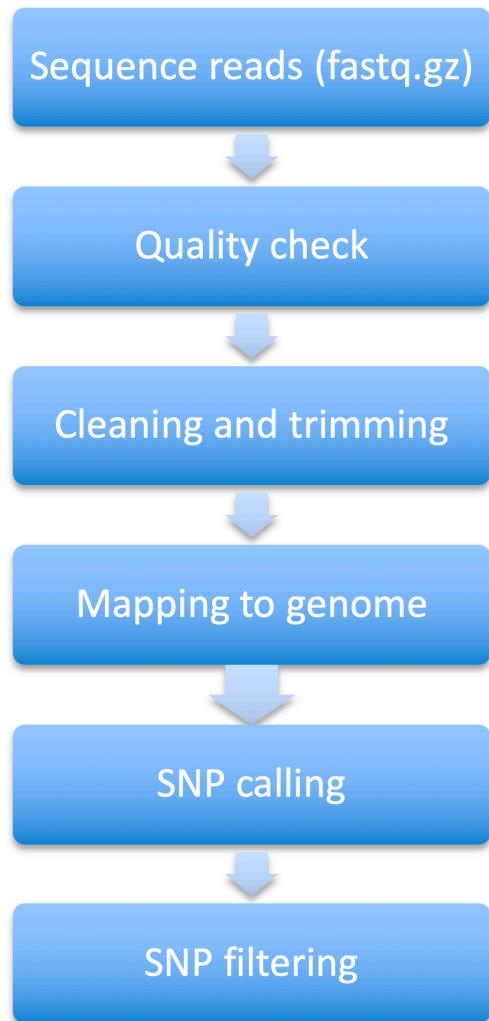
Universität
Zürich^{UZH}

From reads to variants



<https://www.broadinstitute.org/gatk/guide/best-practices>

Simple workflow



FASTQC

trimmomatic / cutadapt

bowtie2 / bwa

samtools / Freebayes / varscan / GATK

bcftools

Simply counting?

GTTACTGTCGTTGTAATACTCCACGATGTC

GTTACTGTCGTTGTAATACTCCACGATGTC

GTTACTGTCGTTGTAATACTCCACGATGTC

GTTACTGTCGTTGTAATACTCCACAATGTC

GTTACTGTCGTTGTAATgCTCCACGATGTC

GTTACTGTCGTTGTAATACTCCACAATGTC

GTTACTGTCGTTGTAATACTCCACGATGTC

GTTACTGTCGTGGTAATACTCCACaATGTC

GTTACTGTCGTTGTAATACTCCACaATGTC

GTtaaTGTGTCGTTGTAATACTCCACGATGTC

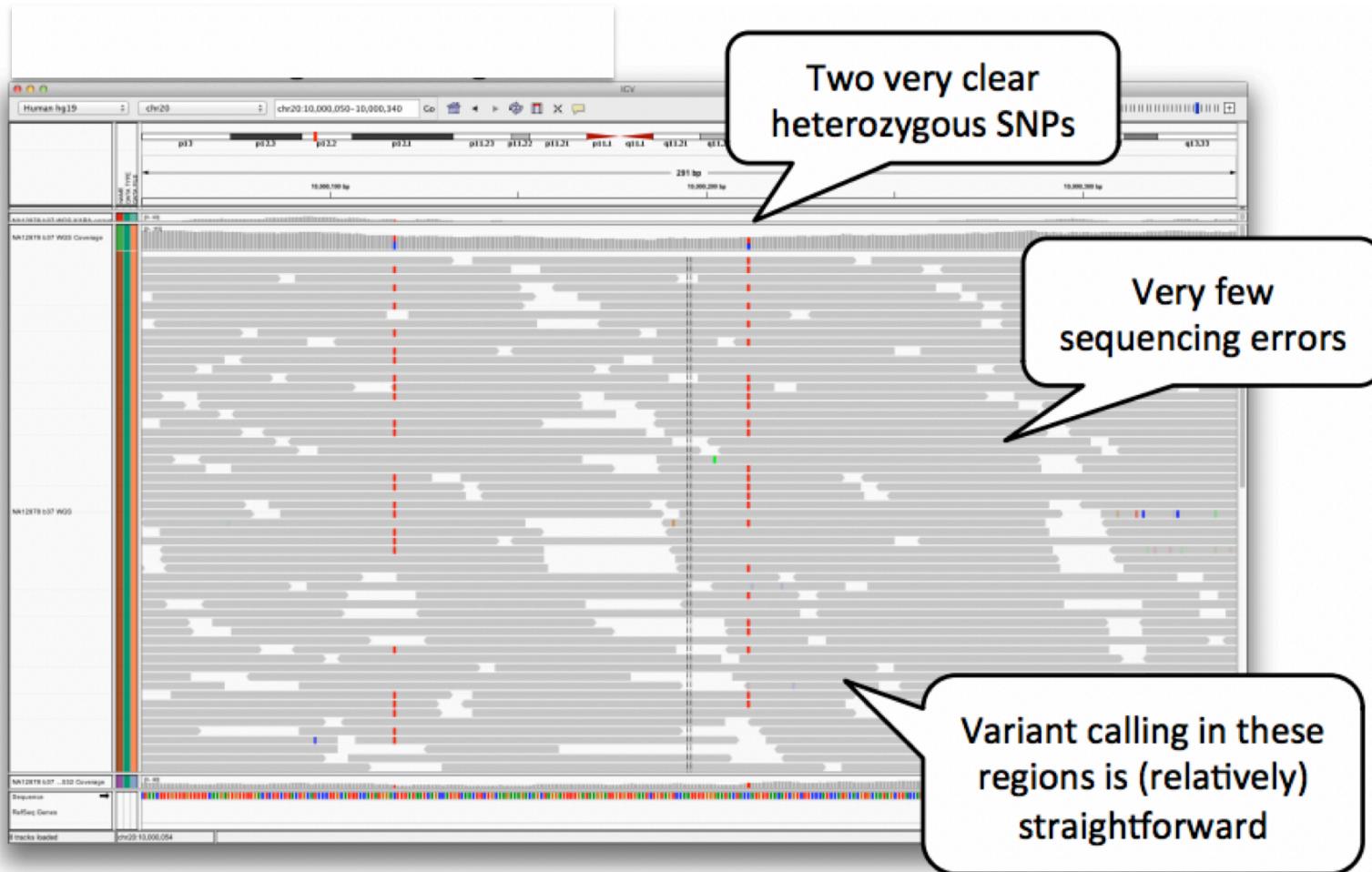
GTTACTGTCGTTGTAcTACTCCACGATGTC

GTTACTGTCGTTGTAATACTCCACaATGTC

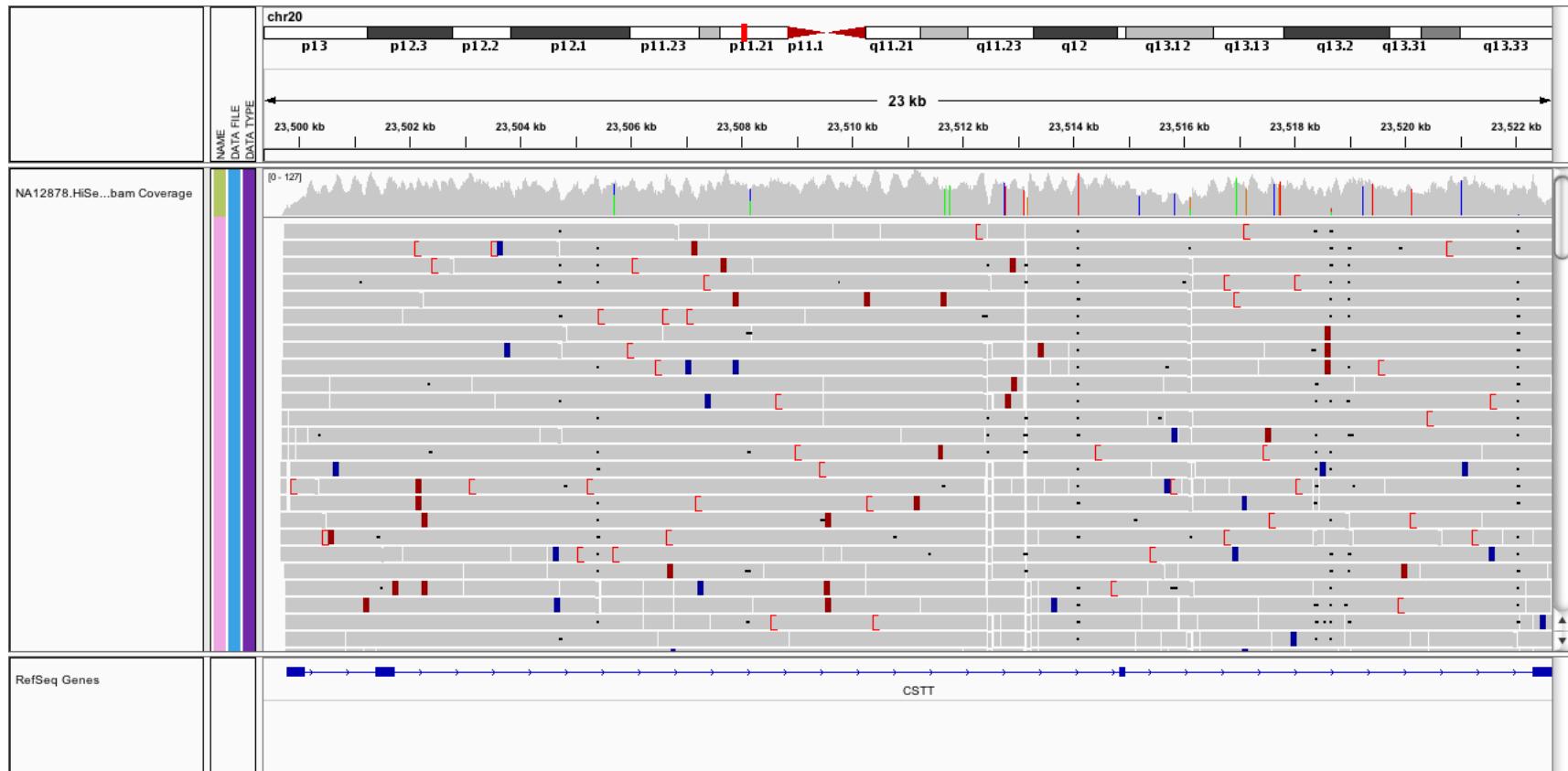
sequencing errors

heterozygous SNP

SNP analysis in well behaved regions of the genome

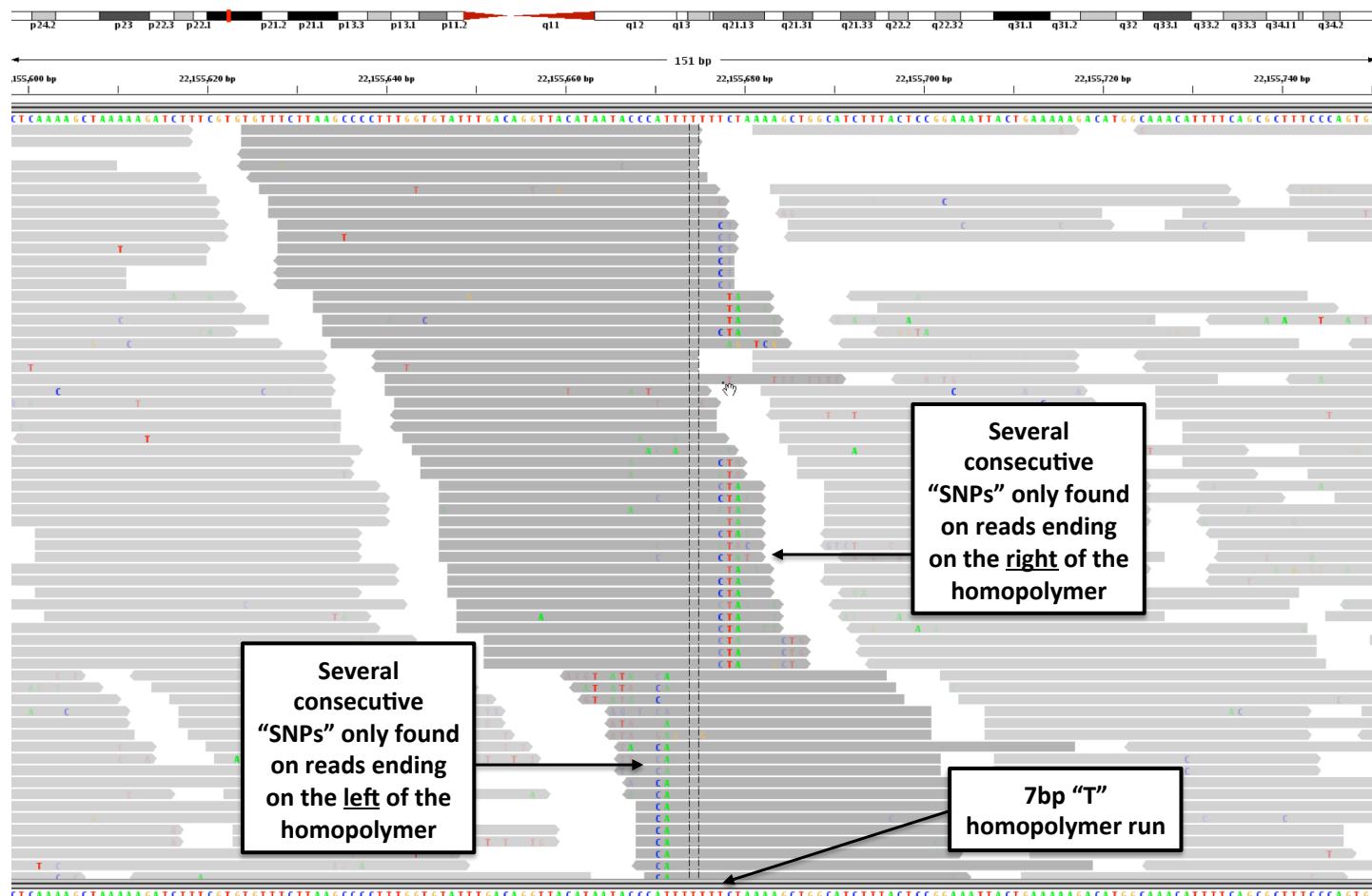


Unclear situations



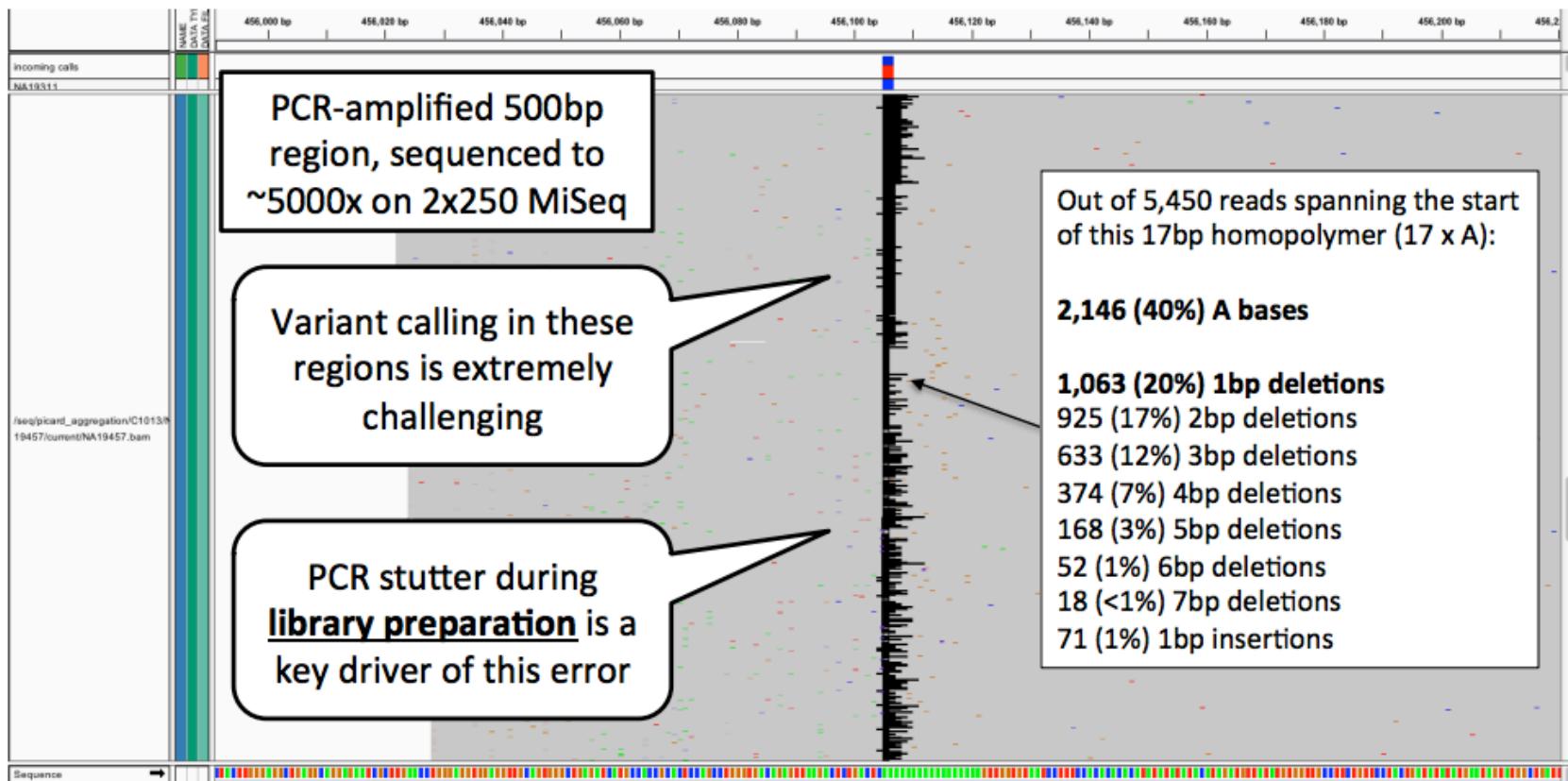
real mutations or noise?

Strand-discordant locus

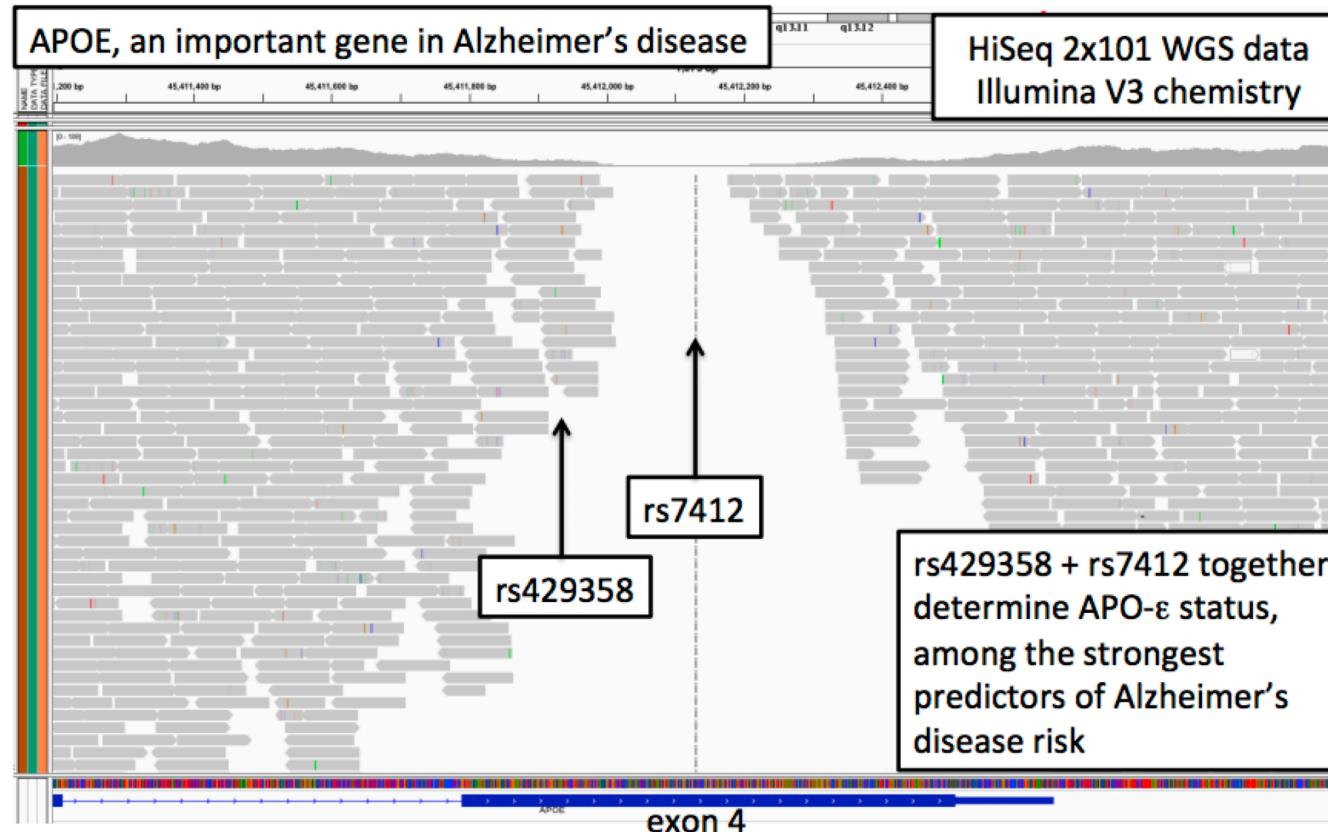


Becomes challenging

Poorly-behaved region of the genome



Lack of coverage



Sequencing errors

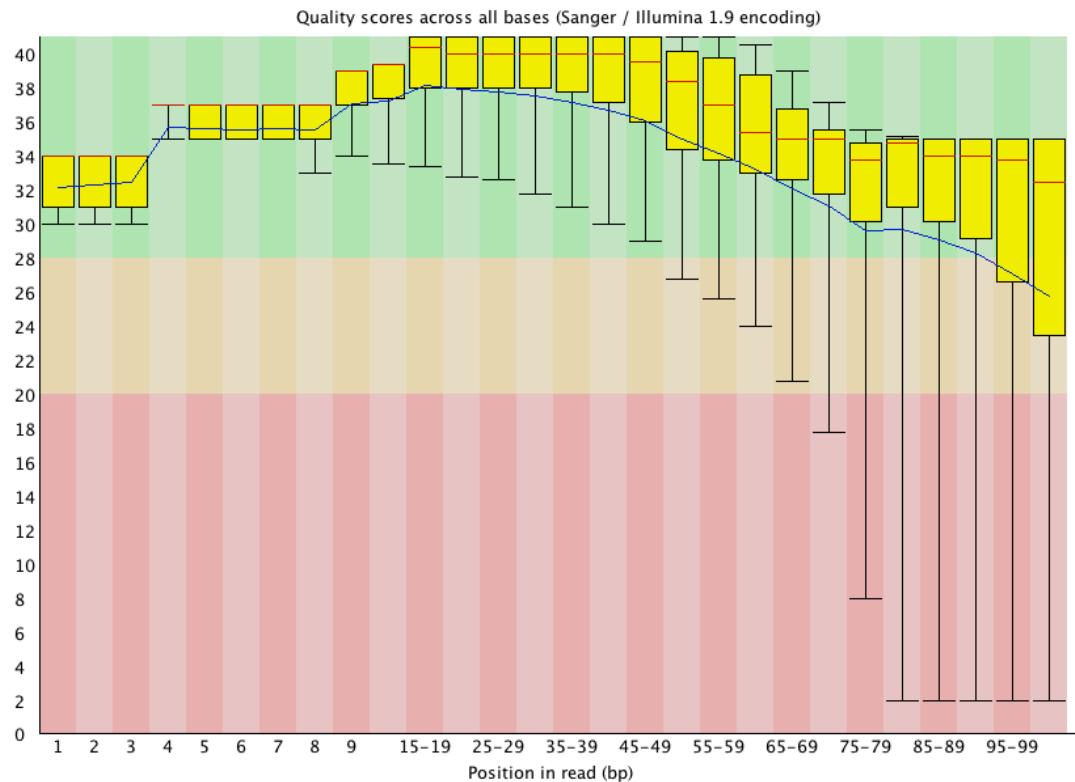
Illumina Sequencing

Error Rate: > 0.1%

(i.e. > 1 in 1000)

mainly substitutions errors

errors mostly at 5' and 3' end



Ambiguous mapping, indels

		Variant Region		Variant Region
Reads	Ref			
TACCGAT		CATTGGATCA	CGATTCC...GCATTGC	AAAAAAA-
TACCGAT		CATTGGATCA	CGATTCC...GCATTGC	-AAAAAA-
ACCGAT		TATTGCATCG	CGATTCC...GCATTGC	-AAAAAA-
ACCGAT		CATTGGATCA	CGATTCC...GCATTGC	AAAAAA-A
ACCGAT		TATTGGATG	CGATTCC...GCATTGC	-AAAAAAA
CCGAT		C-TTGGATCA	CGATTCC...GCATTGC	AAAAAAA-
CCGAT		CATGGGATCA	CGATTCC...GCATTGC	AAAAAAAA

Misaligned reads

segmental duplication
pseudogenes
close paralogs
repetitive sequences
small but complex indels
allelic bias towards reference

Incomplete / misassembled reference genome

GATK

Genome Analysis Toolkit

focused on variant discovery in DNA and RNA

initially developed for the human 1000x genomes project

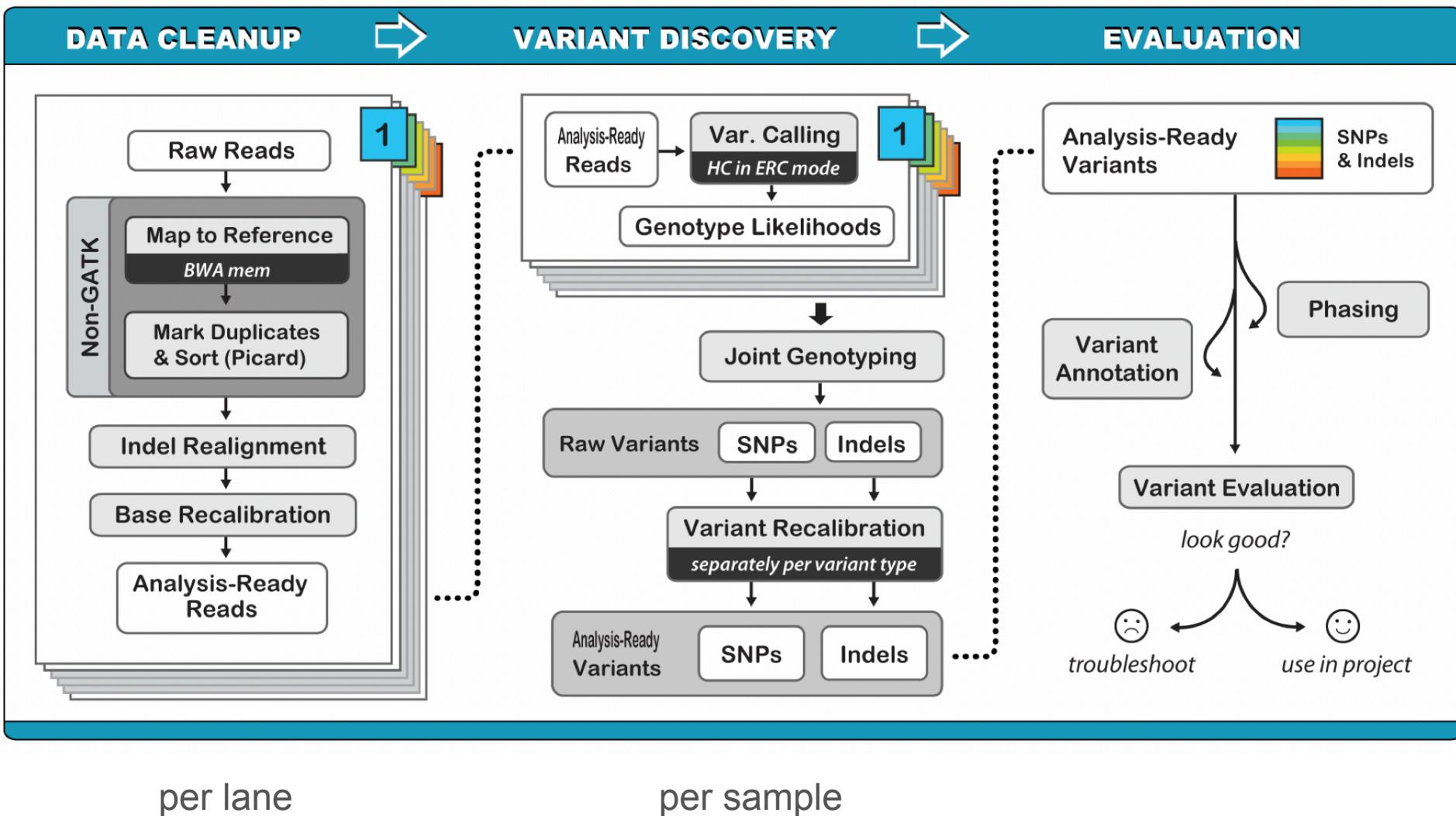
handles any organism with any ploidy

java based command line tool

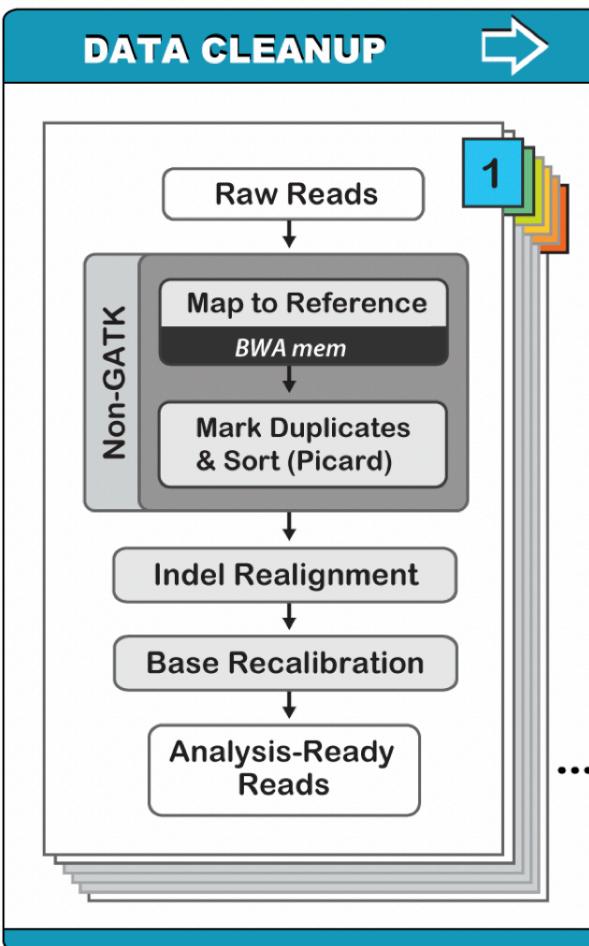
multi sample SNP calling to increase power

automatic filtering / "variant recalibration" for human data

GATK workflow



Data cleanup



per lane

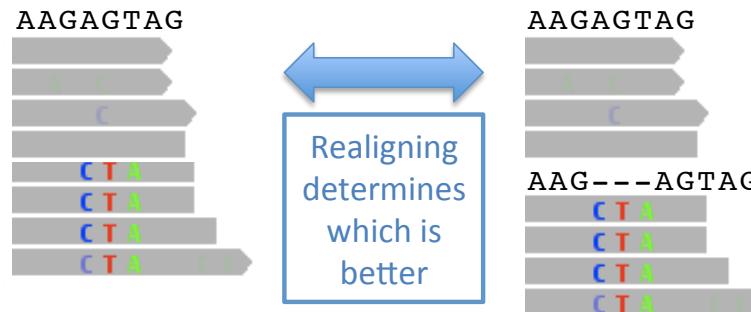
Mark PCR duplicates

come from same input DNA temp. have same start position on reference

non-independent measurements violate statistical assumptions
not applicable in amplicon seq

Indel realignment

indels (near 5'/3') can guide mappers into misaligning with mismatches



Base recalibration

base quality scores are per-base estimates of error emitted by the sequencing machines

various sources of systematic error; over/under estimated base quals

apply machine learning to model errors empirically and adjust the quality scores

Before / after indel realignment



Variant Calling

Modelling various error types

Expected distribution of calls

homozygous AA

homozygous BB

heterozygous AB

GATK v3.3 HaplotypeCaller is recommended for all cases

call SNPs and indels simultaneously

performs a local de-novo assembly

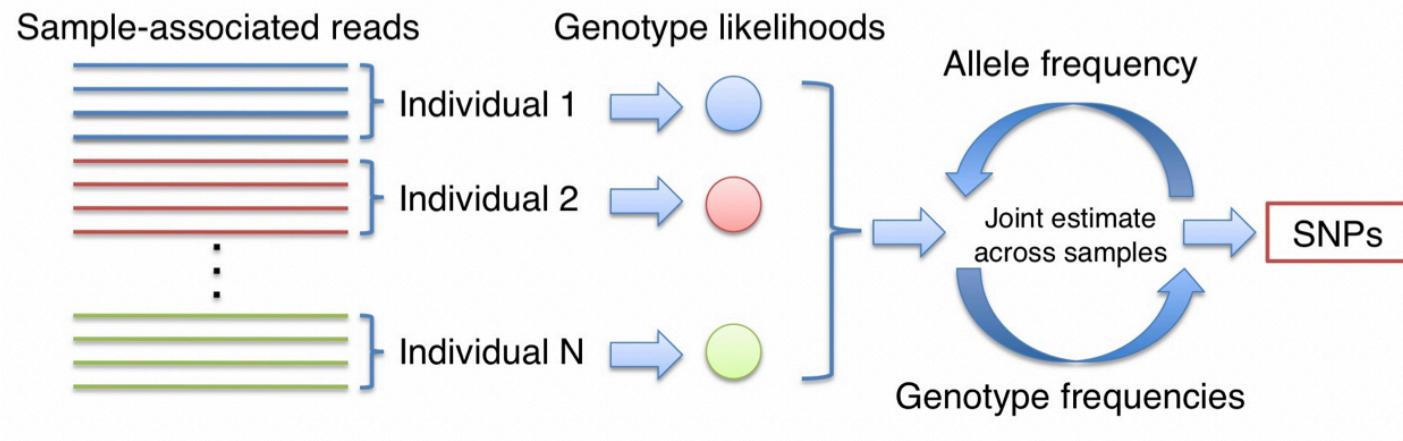
any ploidy

more accurate, especially for indels

up to 100 samples, GVCF mode

Multi-sample analysis

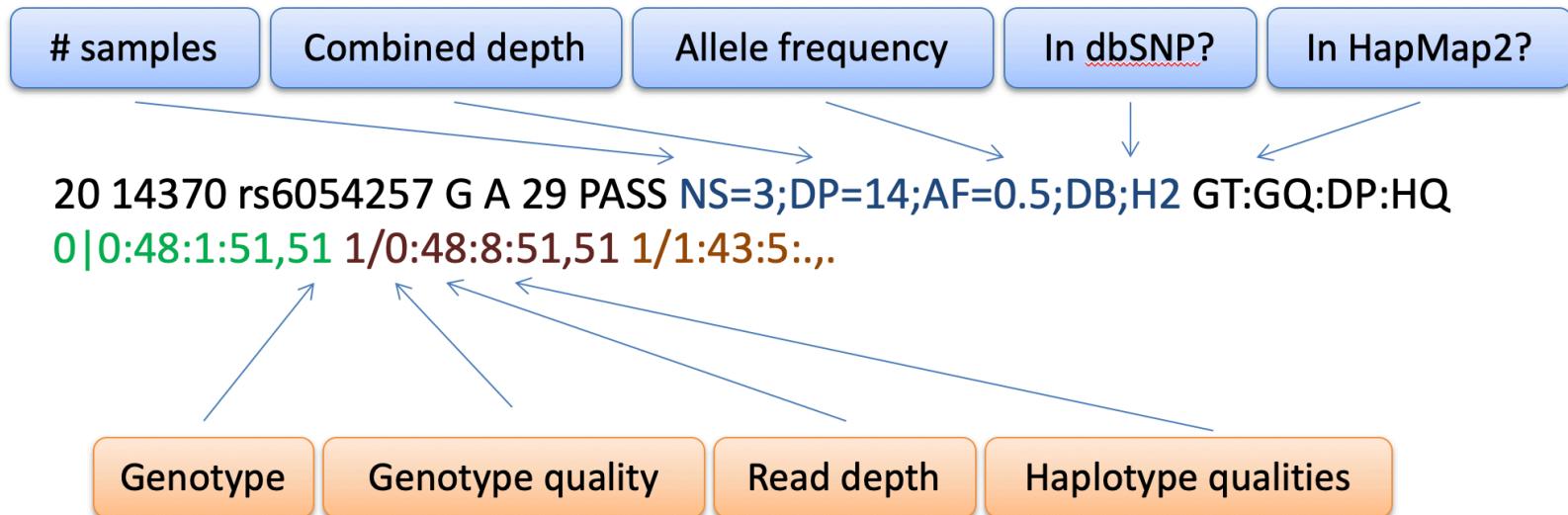
To increase sensitivity some SNP callers allow multi-sample variant calling (multiple individuals / samples from the same or closely related species)



VCF

##fileformat=VCFv4.0

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002
NA00003



VCF info field

VCF record for an A/G SNP at 22:49582364

22 49582364	.	A	0	G	1	198.96	0
AB=0.67;							
AC=3;							
AF=0.50;							
AN=6;							
DP=87;							
Dels=0.00;							
HRun=1;							
MQ=71.31;							
MQ0=22;							
QD=2.29;							
SB=-31.76							
GT:DP:GQ	0/1:12:99.00	0/1:11:89.43	0/1:28:37.78				

INFO field

AC	No. chromosomes carrying alt allele	AB	Allele balance of ref/alt in hets
AN	Total no. of chromosomes	Hrun	Length of longest contiguous homopolymer
AF	Allele frequency	MQ	RMS MAPQ of all reads
DP	Depth of coverage	MQ0	No. of MAPQ 0 reads at locus
QD	QUAL score over depth	SB	Estimated SB score

Heterozygous genotype A/G in all three individuals

Variant Filtering

The optimal threshold for filtering has to be determined empirically

Trade-off between sensitivity and specificity

Which metric of variant call confidence

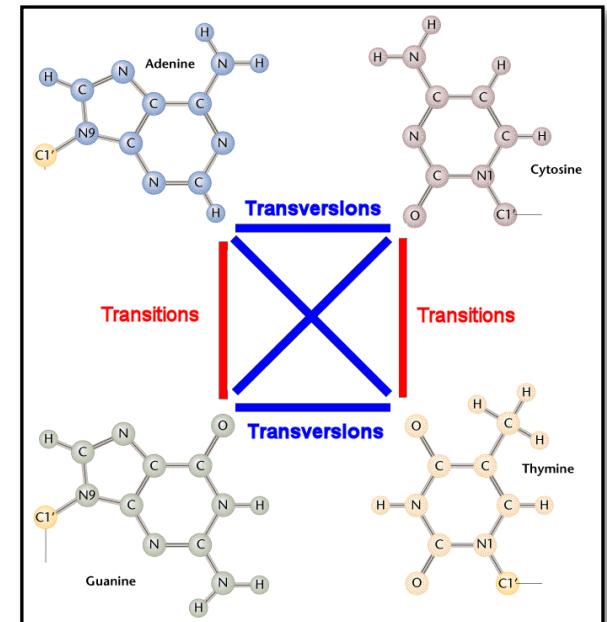
Intrinsic: transitions:transversions ratio (Ti/Tv) (e.g. nuclear genes in humans close to 2)

Experimental validation

Small-scale validation (Sanger seq, qPCR)

Orthogonal data (e.g. microarrays, diff. seq. platform)

Concordance among Trios (2 parents + 1 child)

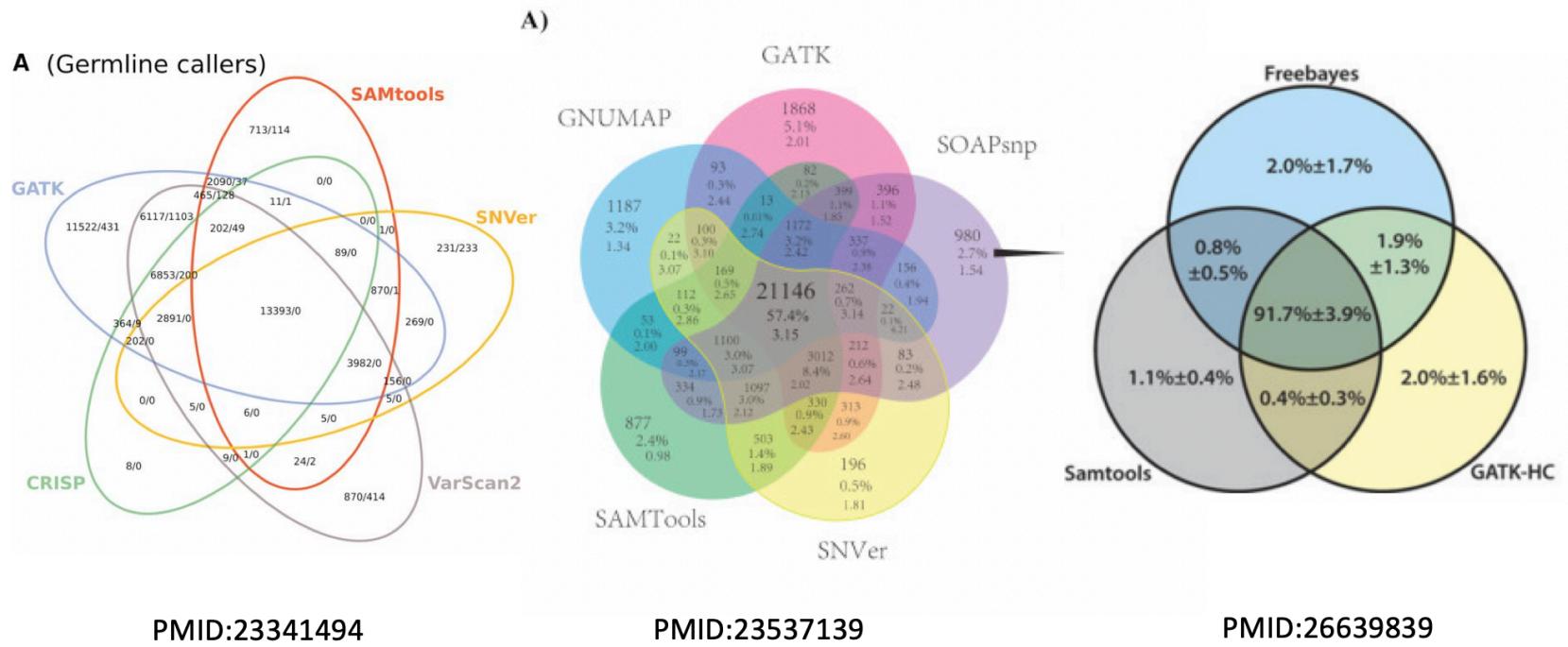


Comparison

BAM pre-processing steps (indel realignment and quality score base recalibration using GATK) had only a modest impact on the variant calls (PMID:25289185)

Realignment of mapped reads and recalibration of base quality scores before SNV calling proved to be crucial to accurate variant calling (PMID: 25078893)

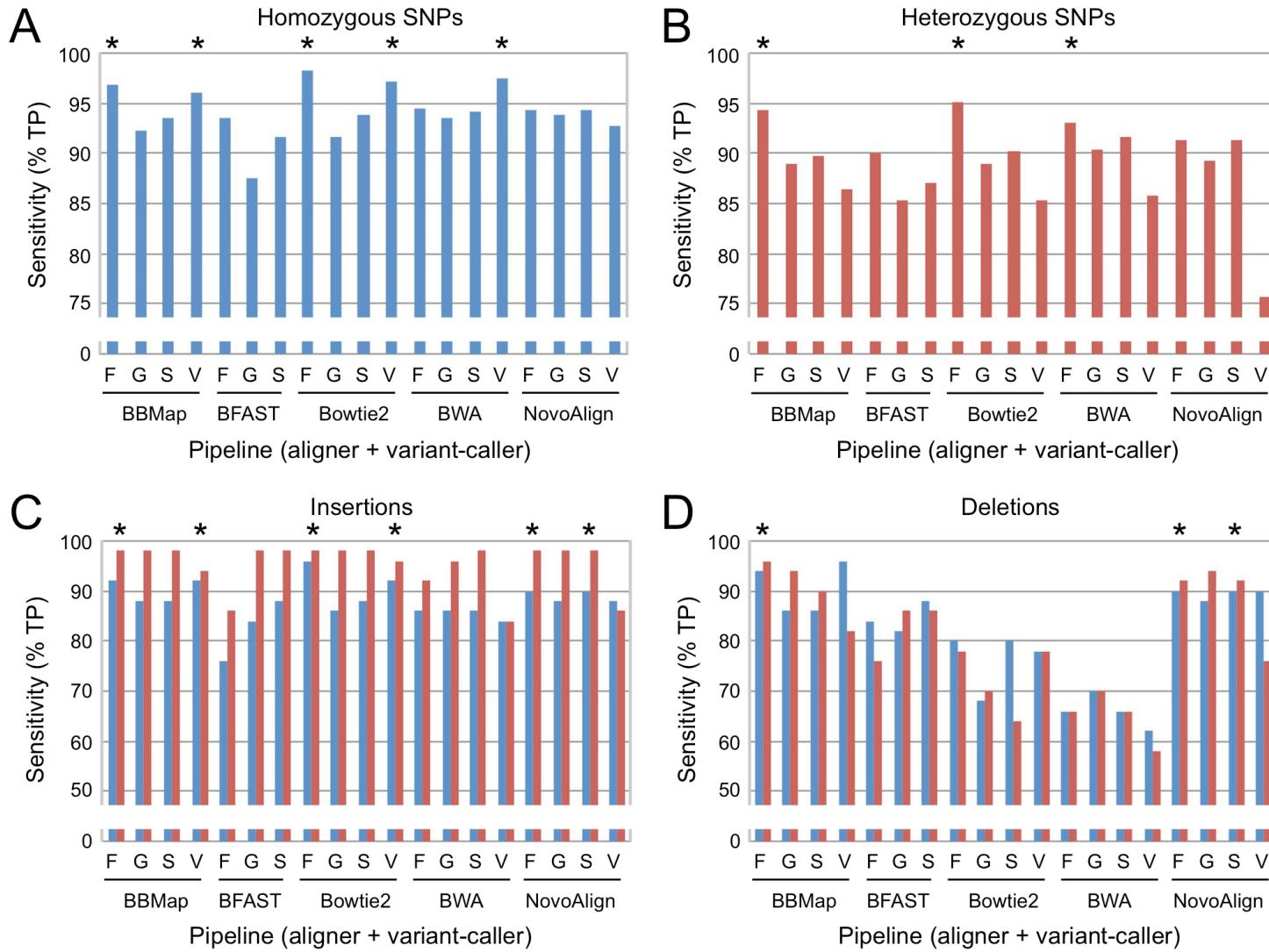
Who performs best?



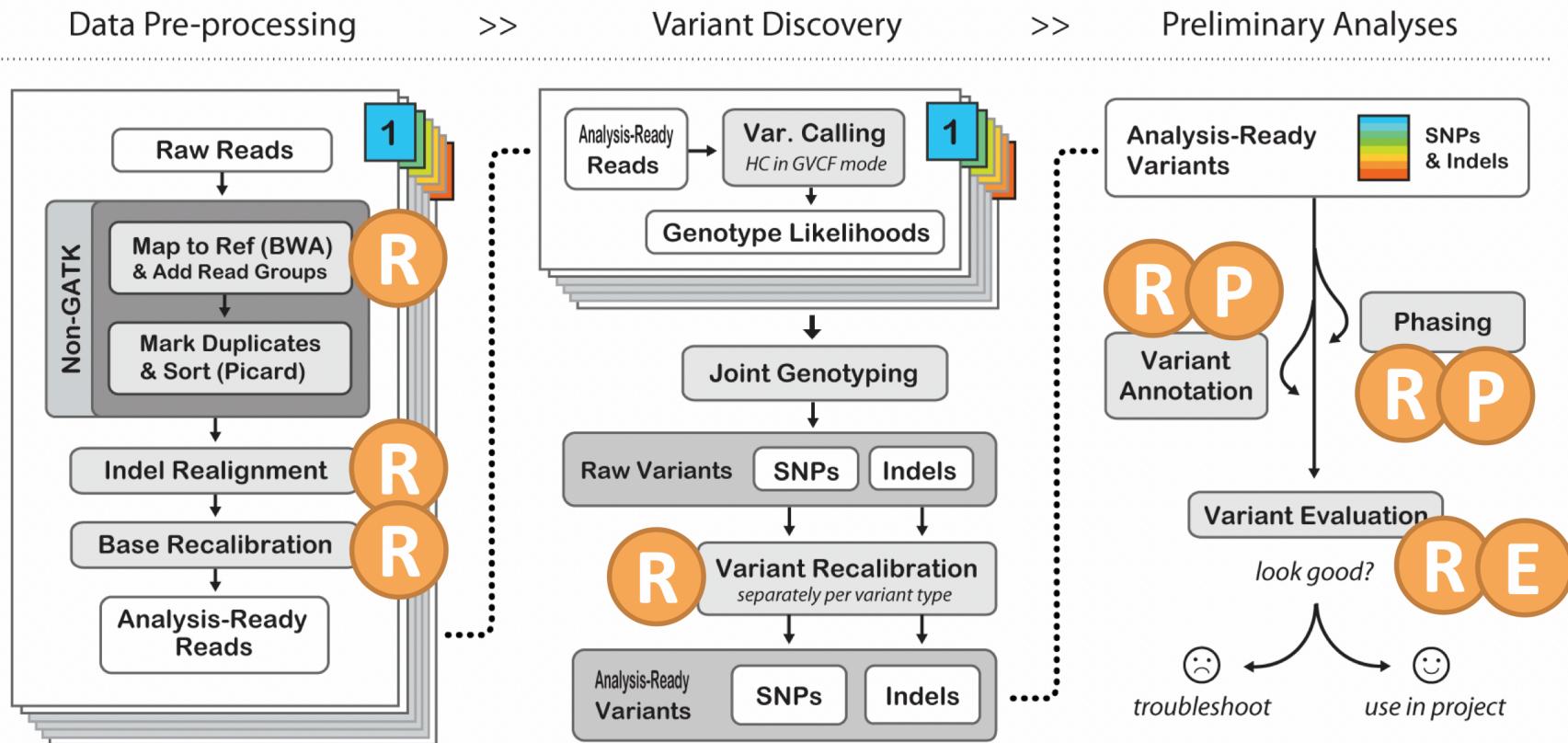
GATK is 'gold-standard' according to many but often only slightly better

Overlap between multiple callers? Time-consuming

Combination of mapper + caller



GATK for non-human organisms: potential problems



R: lack of known resources

P: ploidy assumptions in calculations

E: lack of clear expectations

Non-human organisms

GATK needs a reference genome

very slow with many contigs, make supercontigs, remove or mask transposons and repeats

Indel realignment by default uses indels identified in reads (known indels not required)

Base quality score recalibration: bootstrap until convergence

call variants on realigned, unrecalibrated data

filter resulting variants with stringent filters

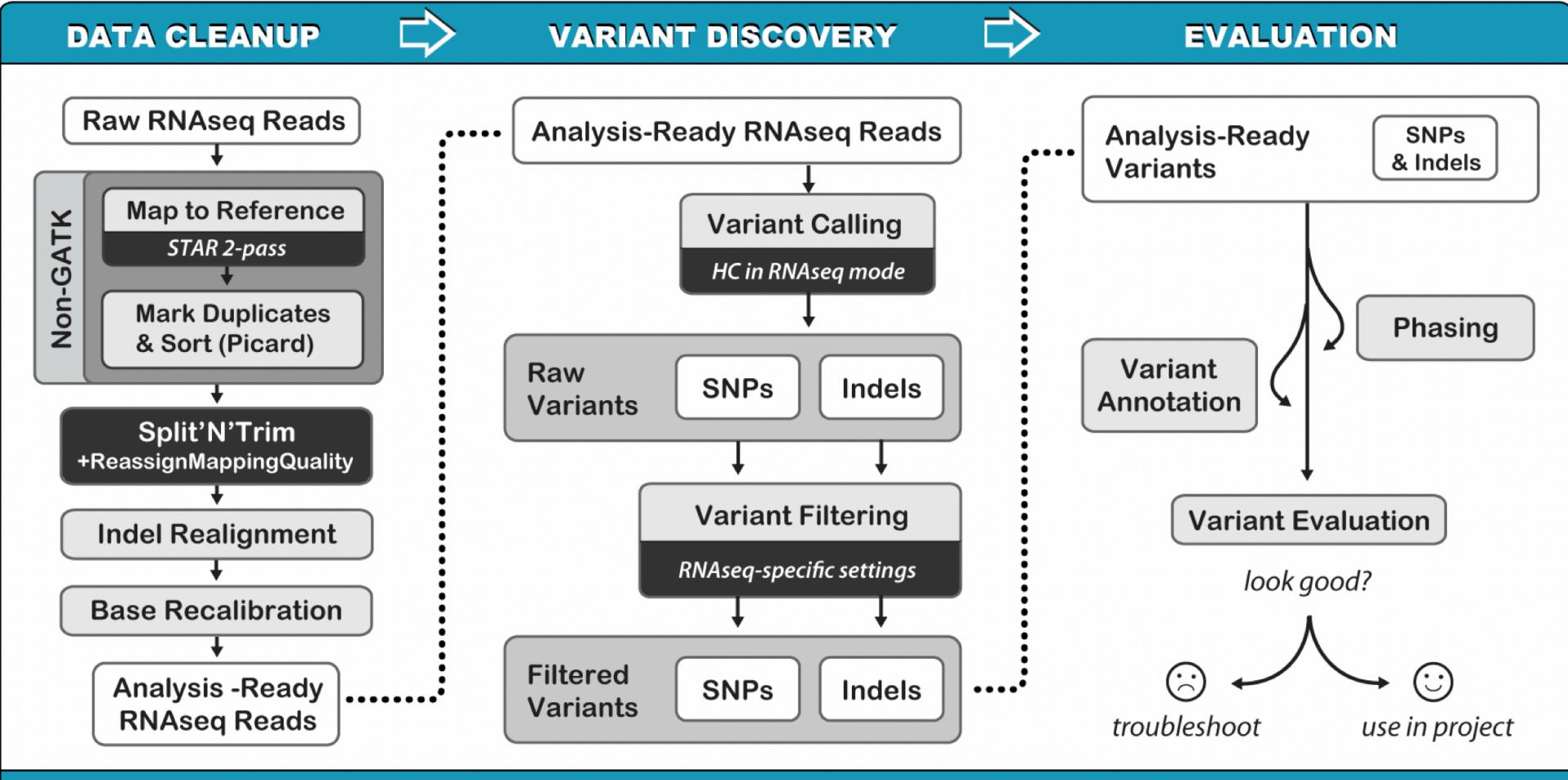
use variants that pass filters as known for BQSR

Ploidy: HaplotypeCaller has --ploidy argument since v3.2

Use hard filtering (No Variant quality score recalibration)

Variant Annotation/Phasing only work for diploid organisms

GATK variant discovery for RNA-seq



Variant Effect Prediction

Software tools that annotate and predict the effects of variants on genes

SnpEff, SnpSift

Ensembl VEP

Type	Region	
Type (alphabetical order)	Count	Percent
CODON_CHANGE_PLUS_CODON_DELETION	632	0%
CODON_CHANGE_PLUS_CODON_INSERTION	162	0%
CODON_DELETION	1,093	0.001%
CODON_INSERTION	282	0%
DOWNSTREAM	11,666,711	9.115%
EXON_DELETED	130,393	0.102%
FRAME_SHIFT	2,643	0.002%
INTERGENIC	15,542,512	12.143%
INTRAGENIC	165,204	0.129%
INTRON	60,208,947	47.041%
NON_SYNONYMOUS_CODING	852,356	0.666%
NON_SYNONYMOUS_START	138	0%
SPlice_SITE_ACCEPTOR	13,375	0.01%
SPlice_SITE_DONOR	15,019	0.012%
START_GAINED	38,052	0.03%
START_LOST	1,621	0.001%
STOP_GAINED	14,815	0.012%
STOP_LOST	814	0.001%
SYNONYMOUS_CODING	624,703	0.488%
SYNONYMOUS_START	1	0%
SYNONYMOUS_STOP	529	0%
TRANSCRIPT	26,447,595	20.663%
UPSTREAM	11,213,085	8.761%
UTR_3_DELETED	2,542	0.002%
UTR_3_PRIME	840,979	0.657%
UTR_5_DELETED	3,684	0.003%
UTR_5_PRIME	204,733	0.16%

Summary

Variants tend to be enriched with artefacts because

short reads are noisy / alignments are noisy / sampling effects

BUT when careful, we still get mostly correct SNP calls

BAM preprocessing is recommended, but the effect is disputed in some publications

Calling indels is error-prone, calling structural variants from short-reads even more (we miss many)

Filtering variants is key (and difficult): Hard-filtering for non-human organisms

Required precision depends on application: Population Genetics < Mutagenesis < Diagnosis