

Practical Bioinformatics

Variant Calling 2

Stefan Wyder

stefan.wyder@uzh.ch

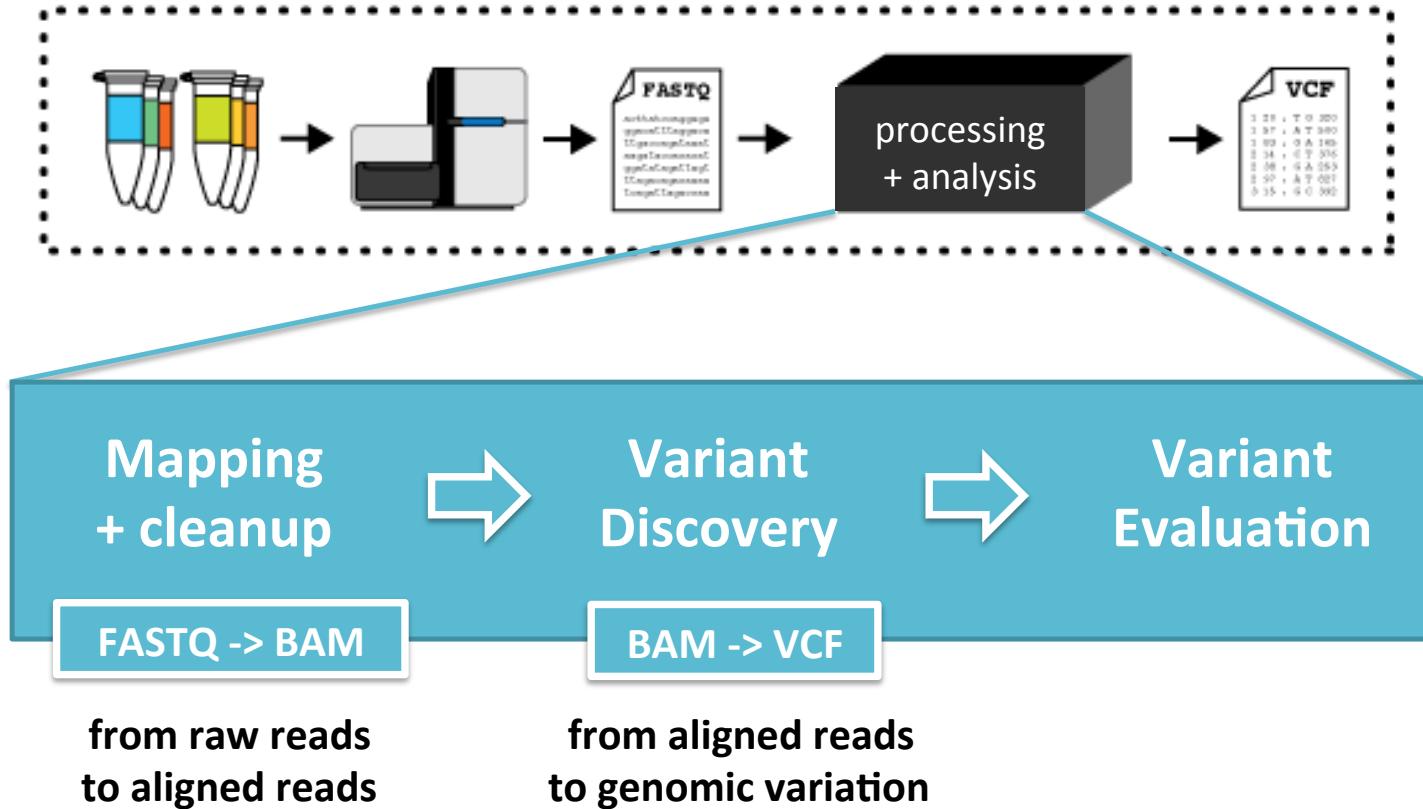


Universität
Zürich^{UZH}

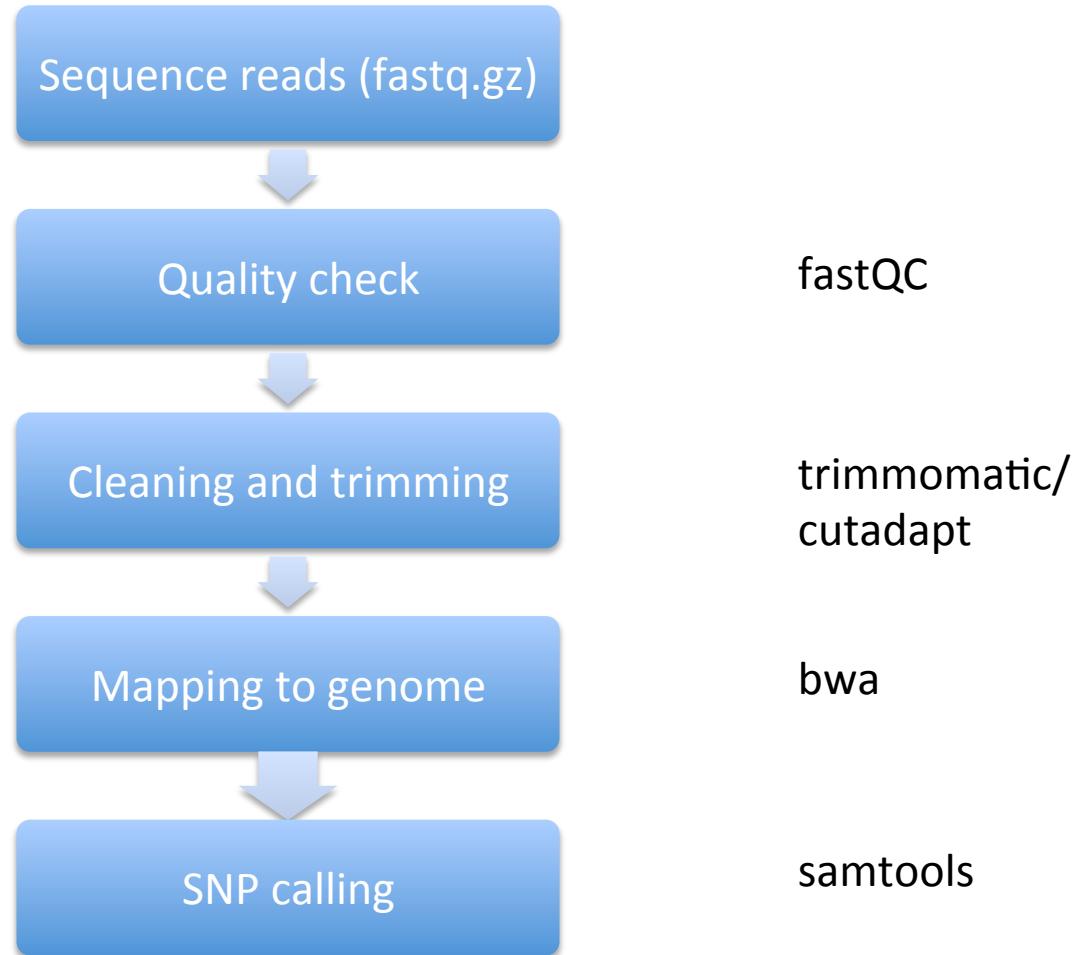


URPP
Evolution
in Action

From reads to variants



Simple workflow using samtools



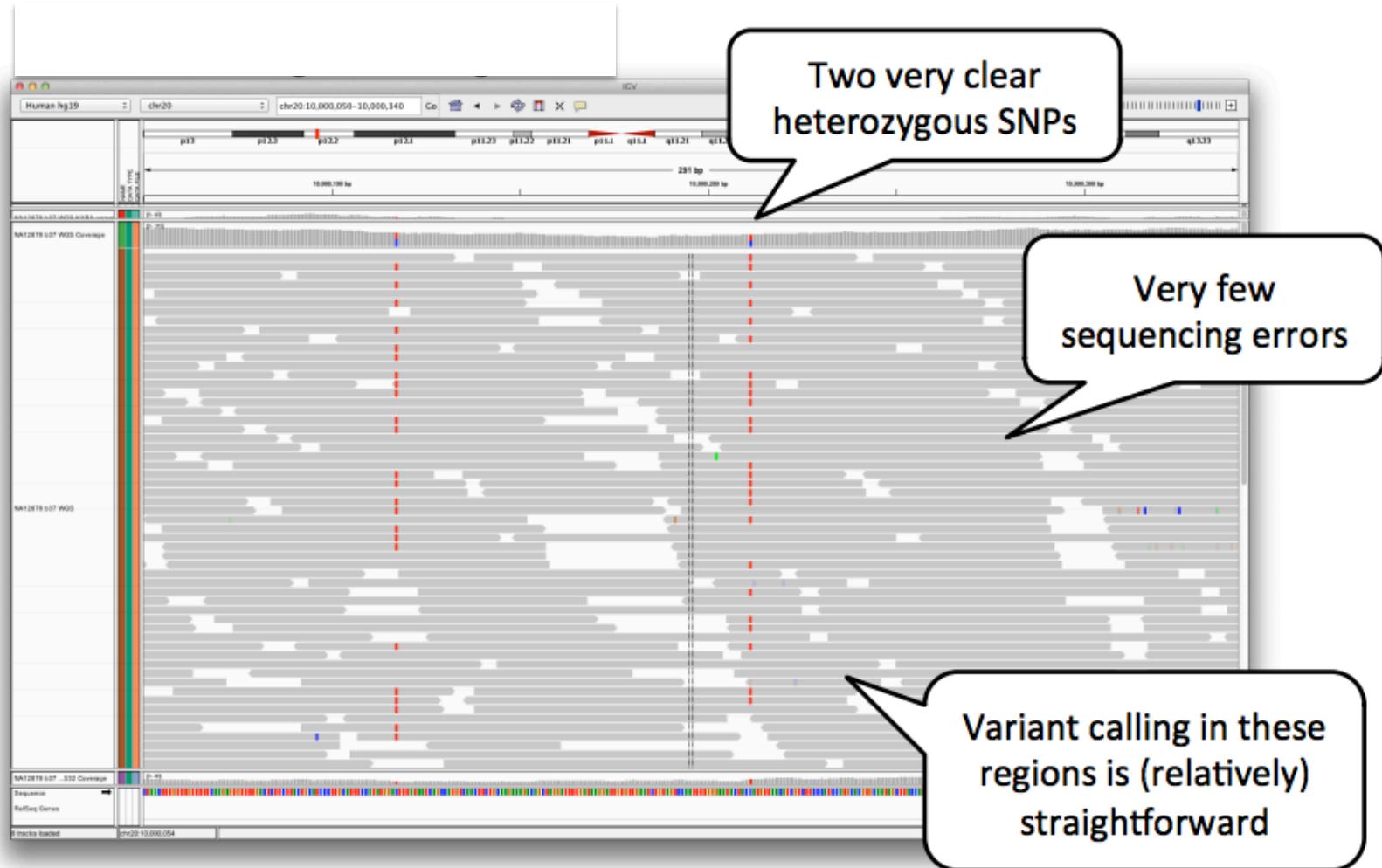
Simply counting?

GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATATACTCCACGATGTC
GTTACTGTCGTTGTAATAACTCCACGATGTC
GTTACTGTCGTTGTAATACCTCCACAATGTC
GTTACTGTCGTTGTAATgCTCCACGATGTC
GTTACTGTCGTTGTAATATACTCCACAATGTC
GTTACTGTCGTTGTAATAACTCCACGATGTC
GTTACTGTCGTGTAATATACTCCACaATGTC
GTTACTGTCGTTGTAATATACTCCACaATGTC
GTTAaTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAcTACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACaATGTC

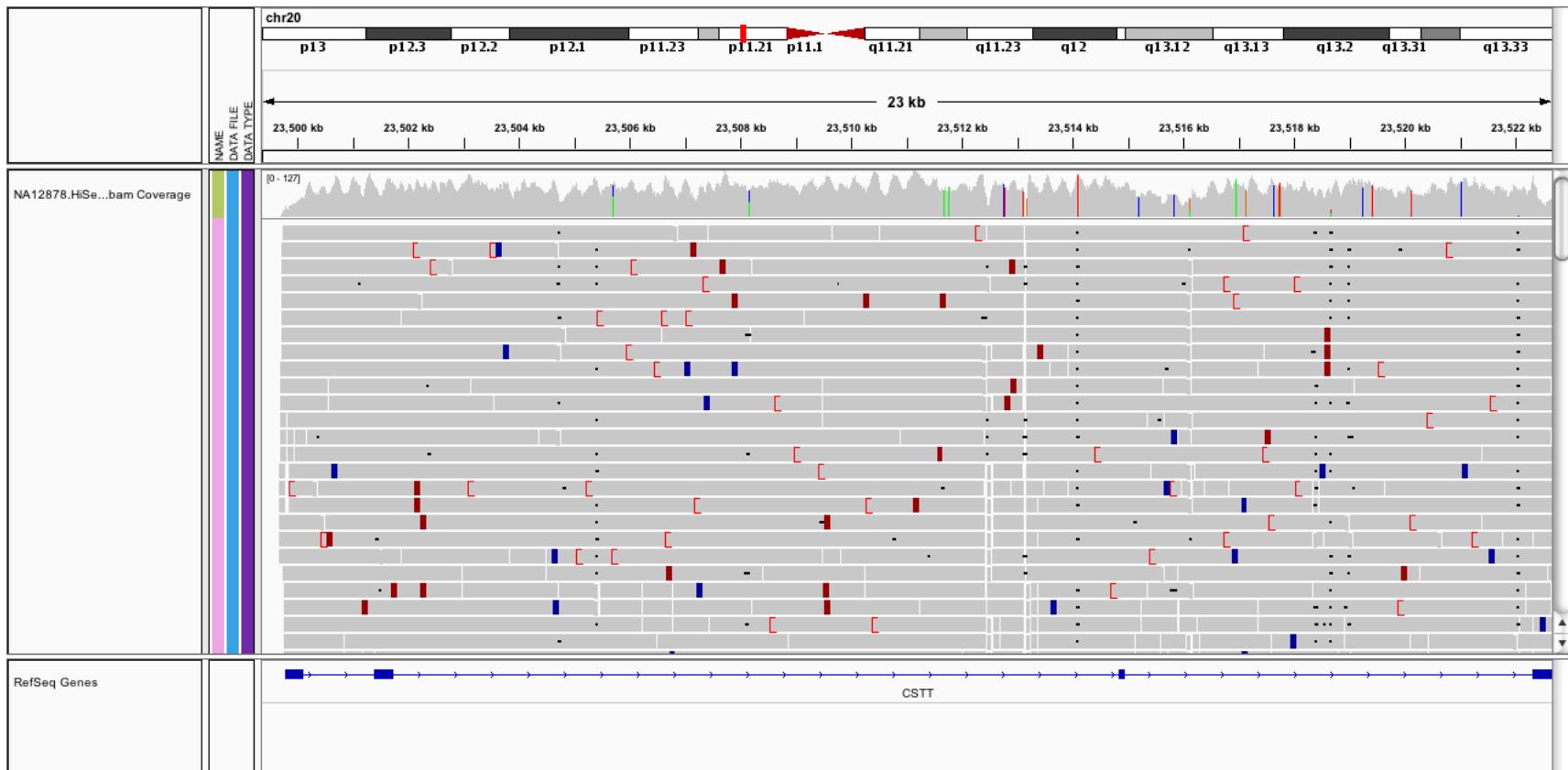
↑ ↑ ↑ ↑ ↑
sequencing errors

heterozygous
SNP

Analysis of SNPs in well-behaved regions of the genome is pretty simple

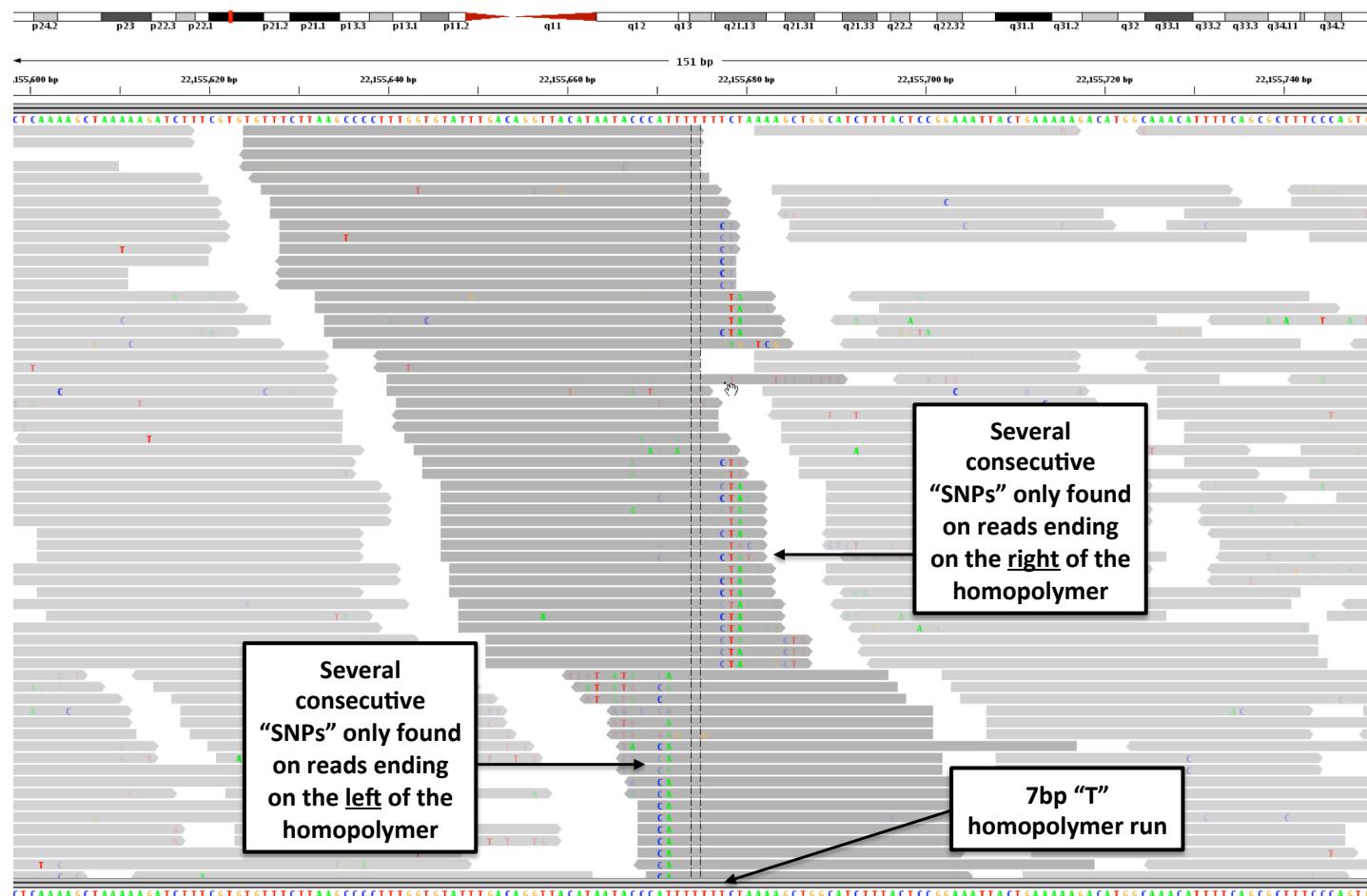


...Messy situations...



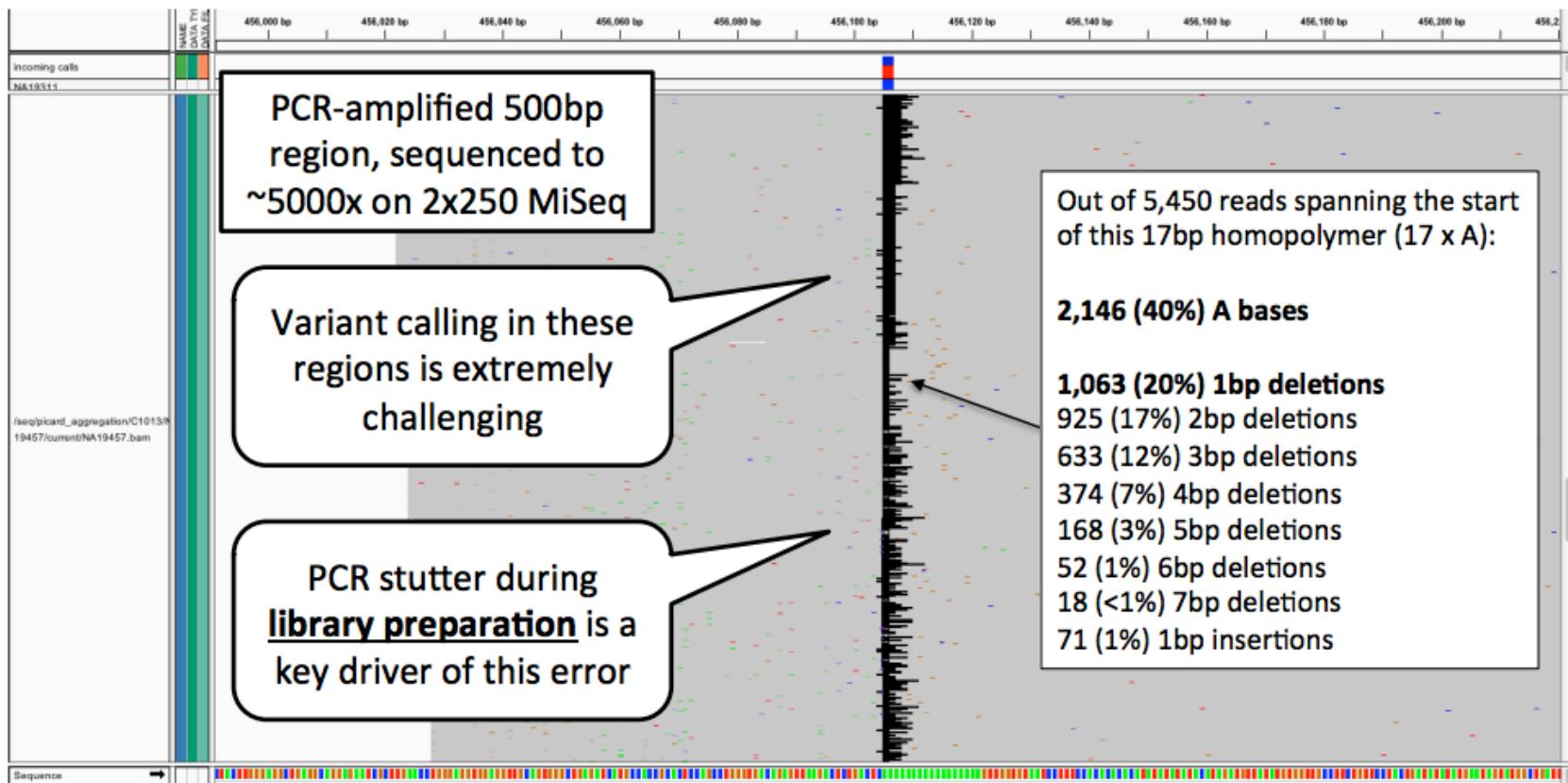
Real mutations or noise?

An example of a strand-discordant locus



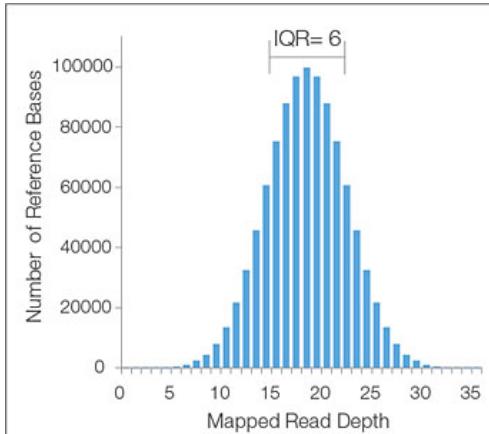
...it can get even worse

Poorly-behaved region of the genome

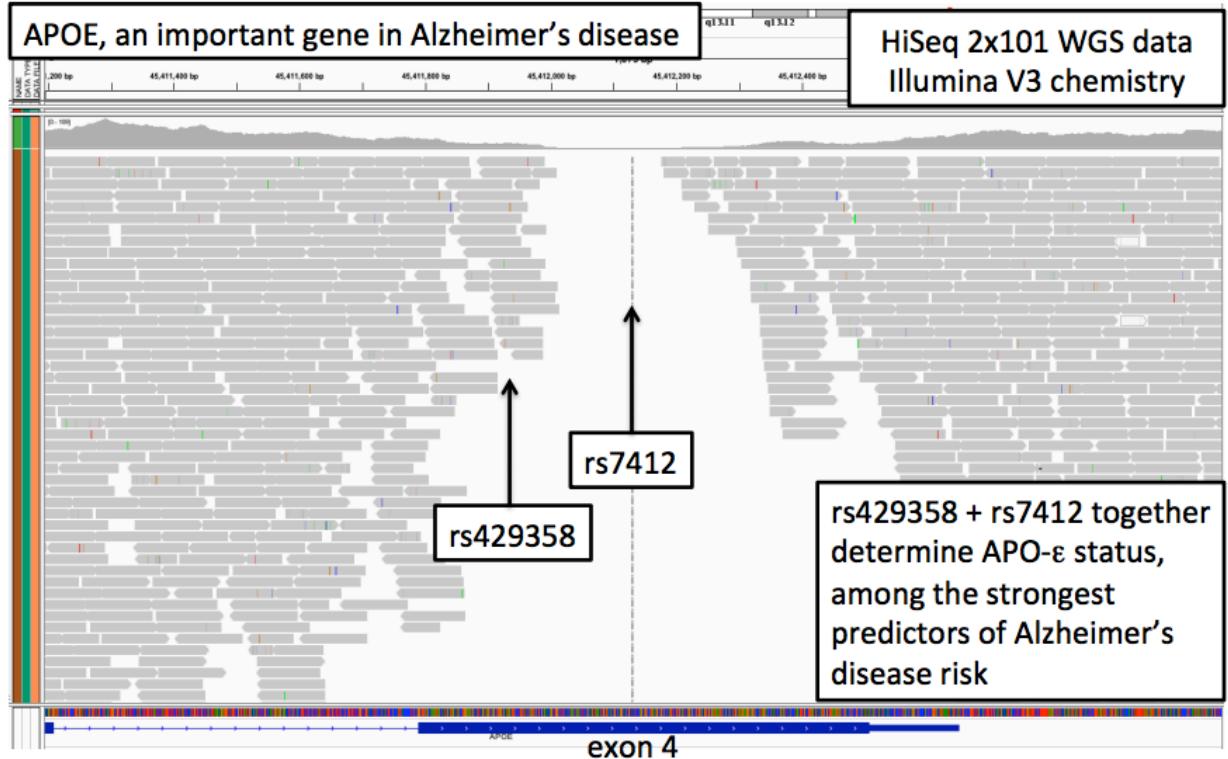


Problem 1: Lack of coverage

Read depth histogram



Illumina

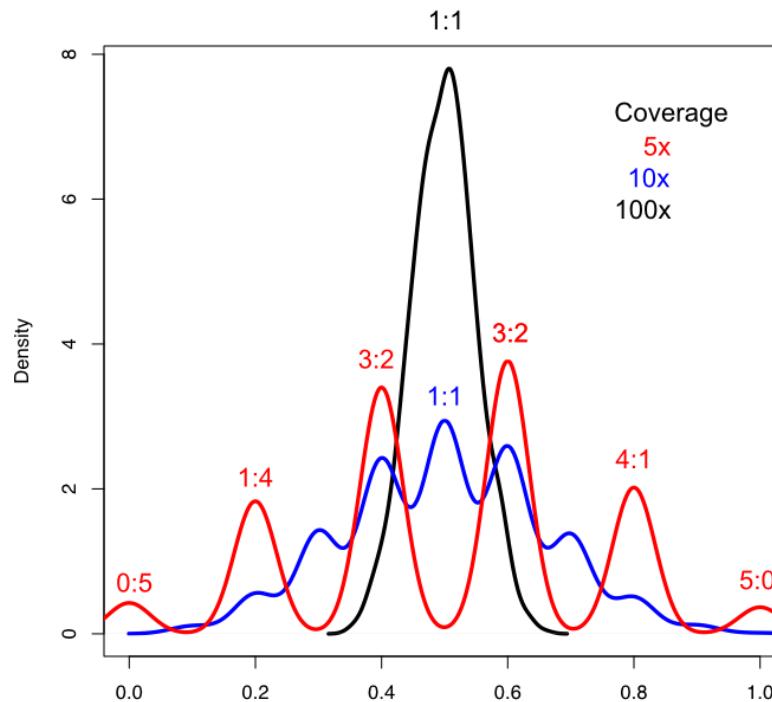


http://www.broadinstitute.org/gatk//events/2247/agbt_2013_depristo.pdf

depends on GC-content, library protocol, sampling effects, mapping problems, ...

Problem 2: Sampling

Simulation of a heterozygous Site
(Binomial Distribution, 1000 samples each)



At 5x coverage, ~10% of sites are 0:5 or 5:0 !

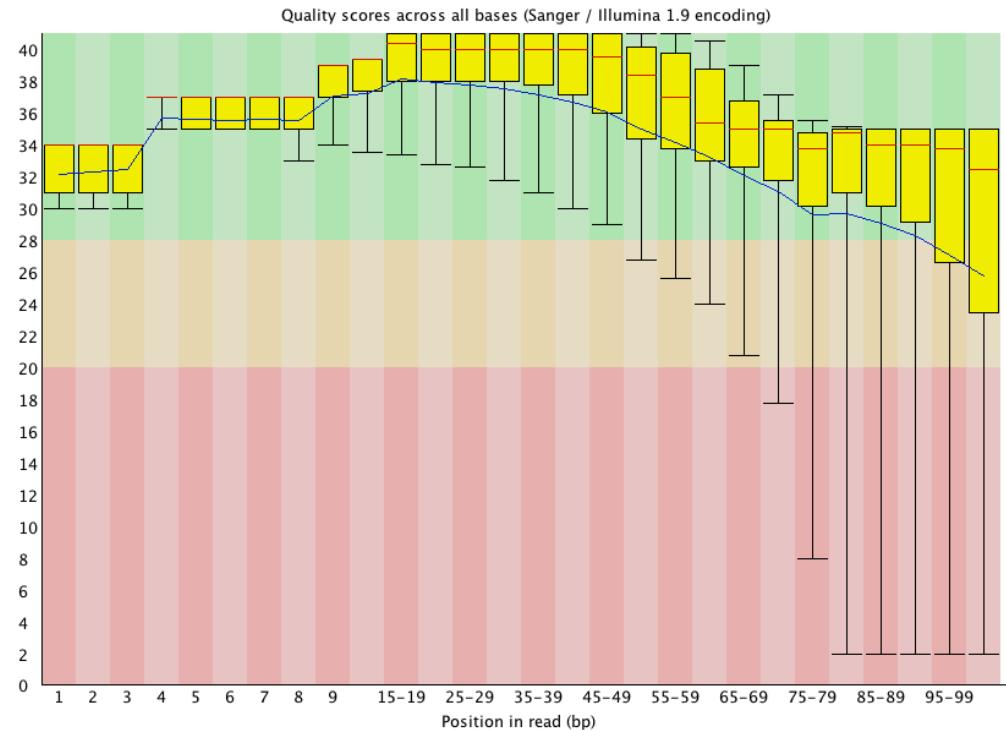
At 10x coverage, ~0.4% of sites are 0:10 or 10:0 !

Problem 3: Sequencing errors

Error rate and error profile are technology-specific

Illumina Sequencing

- Error Rate: $> 0.1\%$
(i.e. > 1 in 1000)
- mainly substitutions errors
- errors mostly at read's start or end



Problem 4: Incorrect mapping

With indels multiple alignments are possible

		Variant Region	Variant Region	
Ref	TACCGAT	CATTGGATCA	AAAAAAA-	GACCGCA
Reads	TACCGAT	CATTGGATCA	-AAAAAA-	GACCGCA
	ACCGAT	TATTGCATCG	-AAAAAA-	GACCGCA
	ACCGAT	CATTGGATCA	AAAAAA-A	GACCGCA
	ACCGAT	TATTGGATCG	-AAAAAAA	GACCGCA
	CCGAT	C-TTGGATCA	AAAAAAA-	GACCGCA
	CCGAT	CATGGGATCA	AAAAAAA A	GACCGCA

FreeBayes

-> Indel Realignment

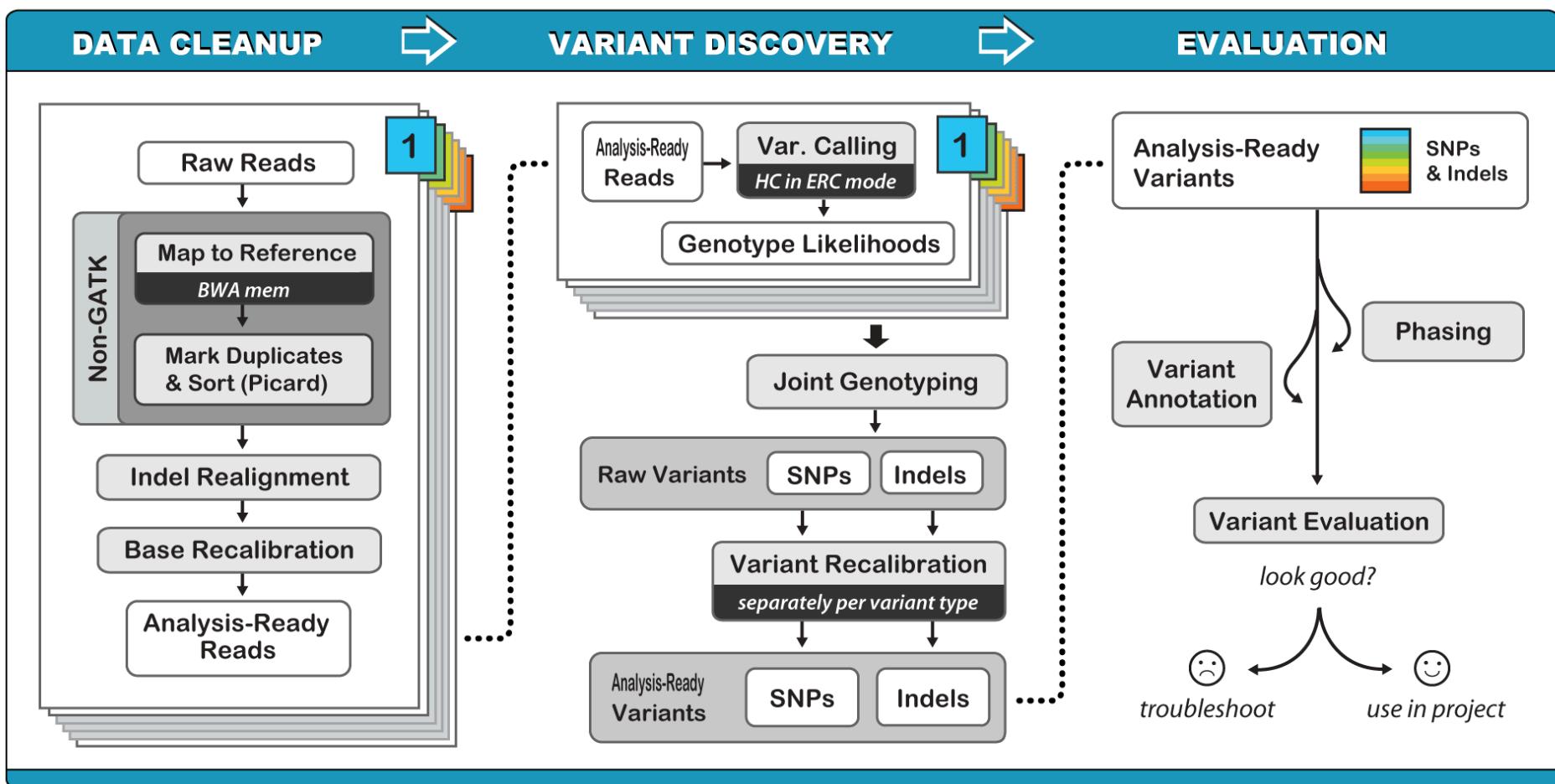
Problem 4: Incorrect mapping cont.

- mismapped reads / errors in the alignment
 - segmental duplication
 - processed pseudogenes
 - close paralogues
 - repetitive sequences
 - small but complex indels
 - allelic bias towards reference
- incomplete/missassembled reference genome

GATK

- Genome Analysis Toolkit (GATK)
- Toolkit focused on variant discovery in DNA and RNA
- initially developed for human 1000 genomes project
- handles any organism with any ploidy (<-> samtools/bcftools)
- Java-based command line tool
- Multi-sample SNP calling to increase power
- Automatic filtering ("Variant recalibration") for human data

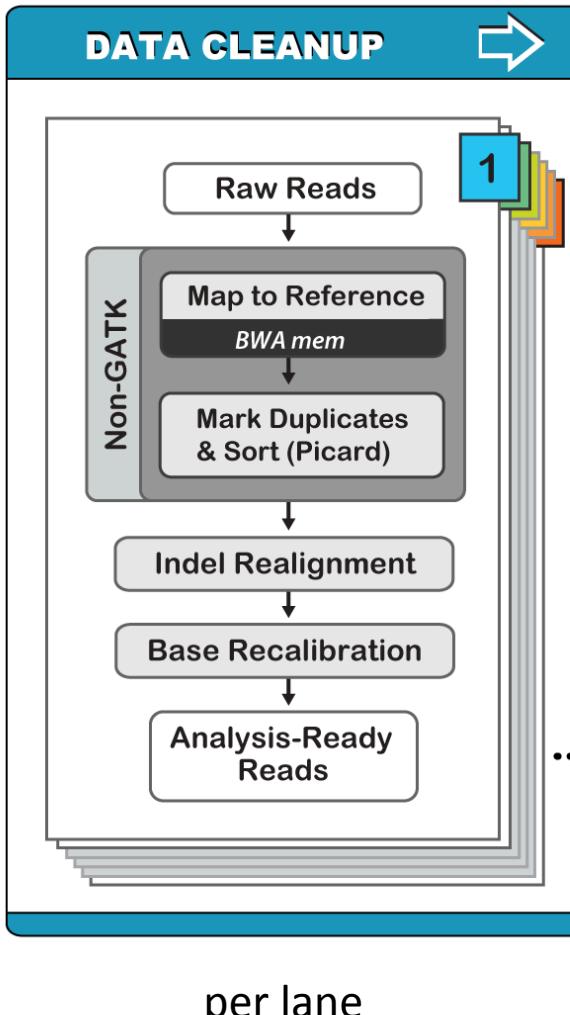
GATK Workflow for DNA



per lane

per sample

Data Cleanup

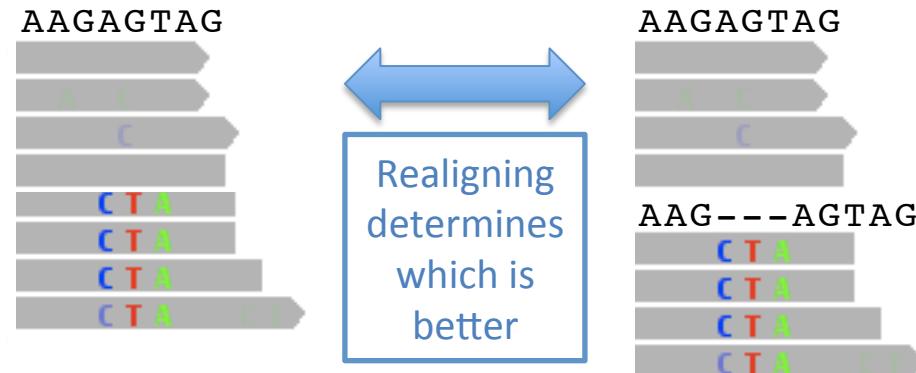


Mark PCR Duplicates

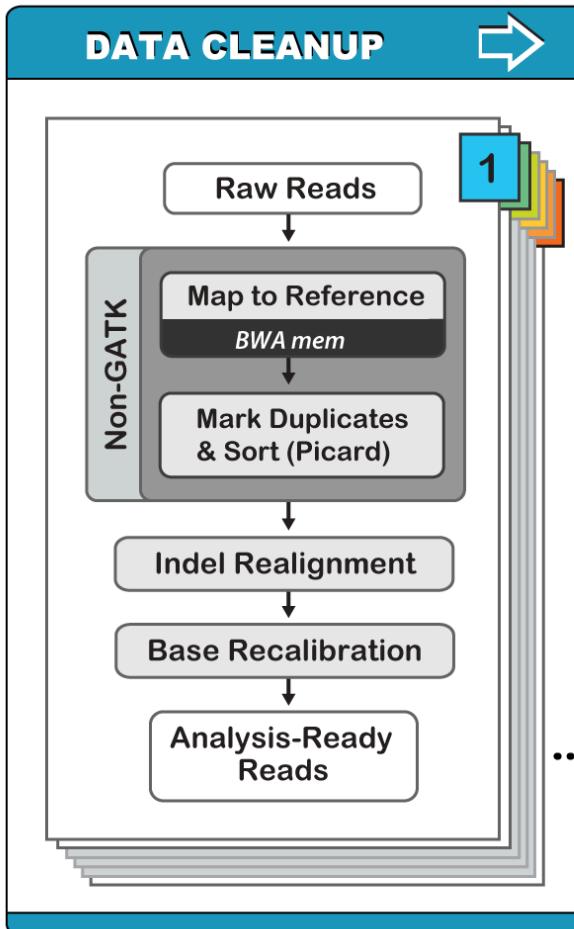
- non-independent measurements
- violate statistical assumptions
- come from same input DNA template - have same start position on reference

Indel Realignment

- Indels in reads (especially near the ends) can trick the mappers into mis-aligning with mismatches



Data Cleanup cont.

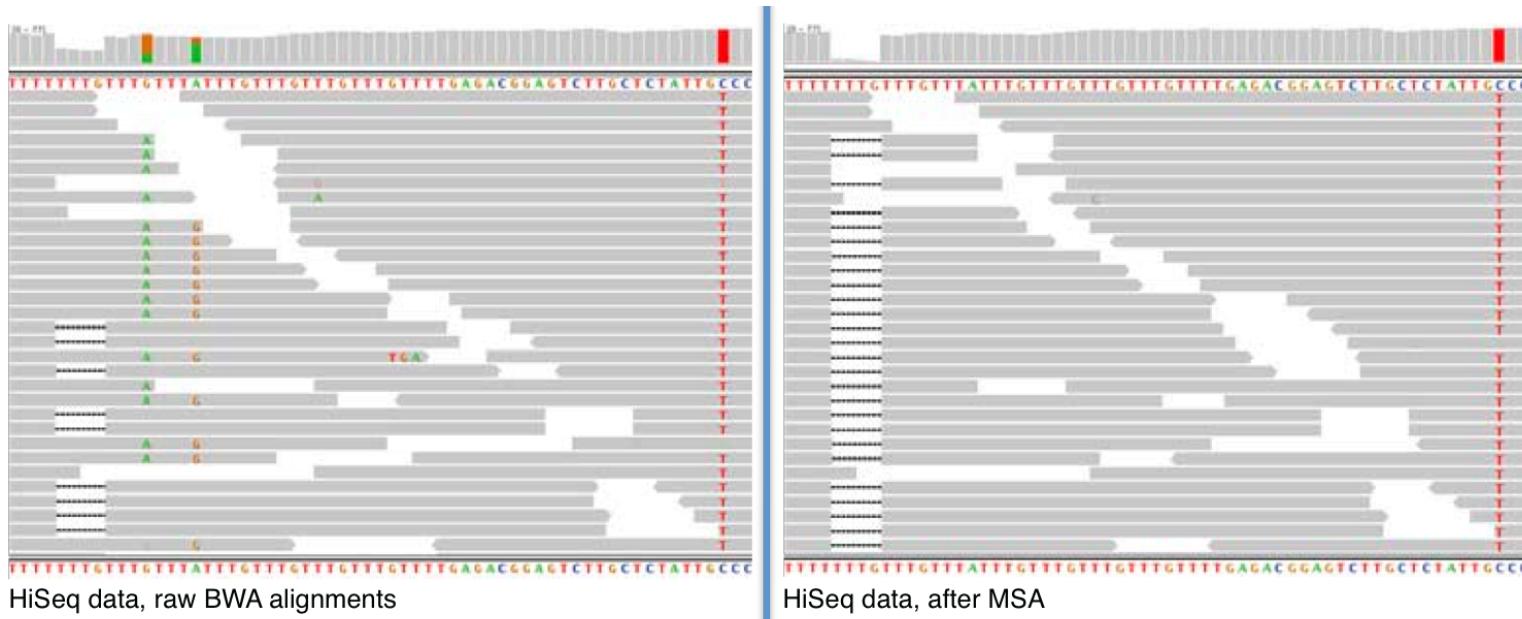


per lane

Base Recalibration

- Base quality scores are per-base estimates of error emitted by the sequencing machines
 - > various sources of systematic error
 - > over- or under-estimated base qualities
- Base quality score recalibration apply machine learning to model errors empirically and adjust the quality scores

Before and after indel realignment



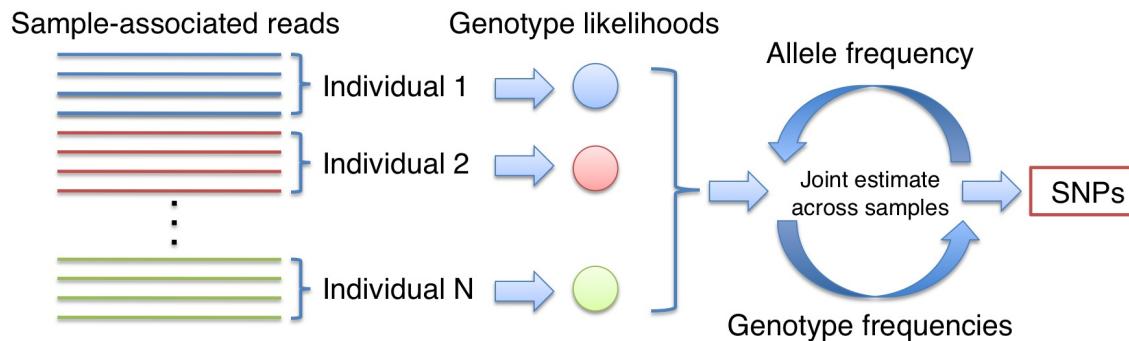
Variant Calling

- modelling various error types
- expected distribution of calls
(homozygous AA, homozygous variant BB, heterozygous AB)
- GATK comprises 2 variant calling tools:

UnifiedGenotyper	HaplotypeCaller (newer)
calls SNPs and indels separately	calls SNPs and indels simultaneously
each variant locus independently	performs a local de-novo assembly
any ploidy	any ploidy
Pooled calling	more accurate (especially for indels)
High sample numbers	computationally more intense

Multi-sample analysis

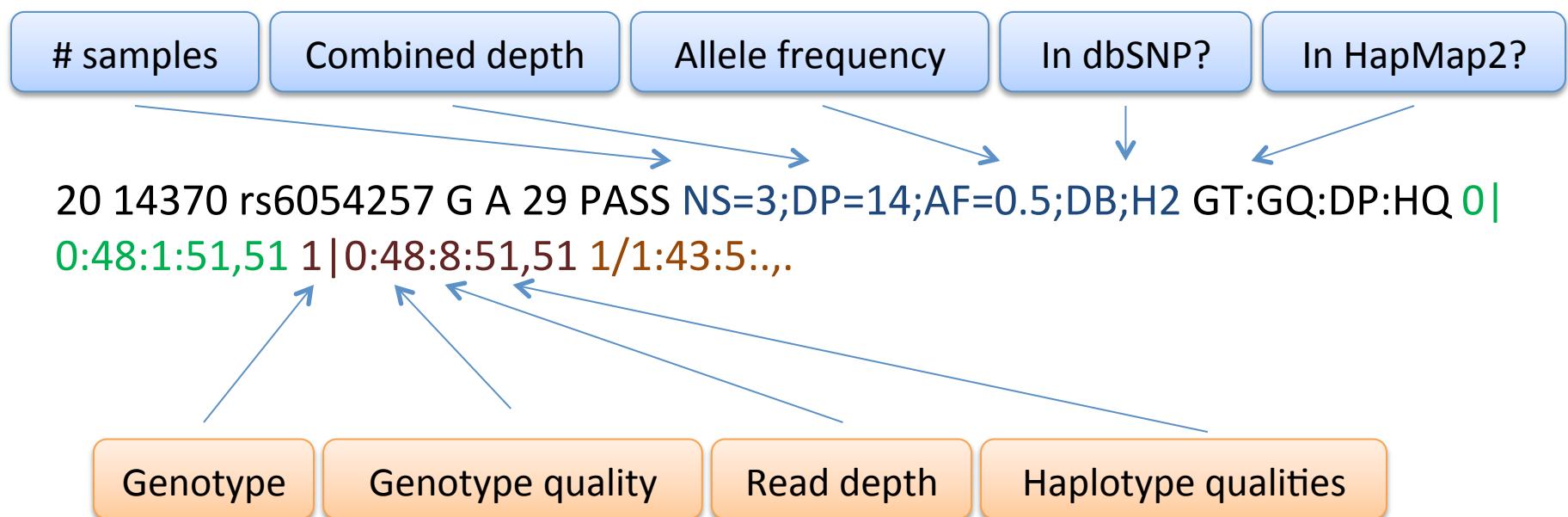
- To gain sensitivity some SNP callers allow **multi-sample** variant calling (multiple individuals/samples from the same or closely related species)



- ~ Hardy-Weinberg Equilibrium
- Genotypings like this: AB, AB, AB, AB, AB, AB have much lower probability than AA, AA, AB, BB, AA, AB, AA
- in reality: multiple alleles,...

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

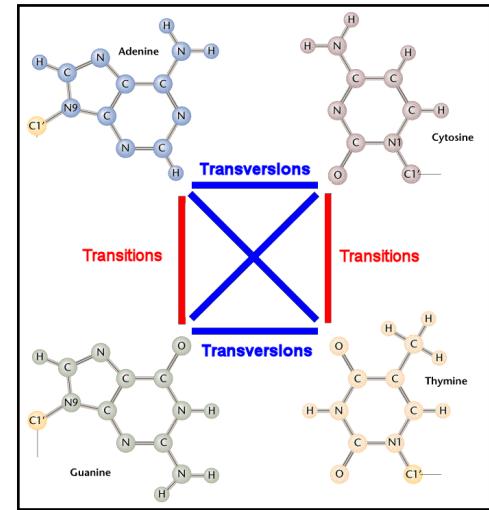
22 49582364	.	A	0	G	1	198.96	0
AB=0.67;	AC=3;	AF=0.50;		AC	No. chromosomes carrying alt allele	AB	Allele balance of ref/alt in hets
AN=6;	DP=87;	Dels=0.00;		AN	Total no. of chromosomes	Hrun	Length of longest contiguous homopolymer
HRun=1;	MQ=71.31;	MQ0=22;		AF	Allele frequency	MQ	RMS MAPQ of all reads
QD=2.29;	SB=-31.76	GT:DP:GQ		DP	Depth of coverage	MQ0	No. of MAPQ 0 reads at locus
		0/1:12:99.00		QD	QUAL score over depth	SB	Estimated SB score

Variant Filtering

- The optimal threshold for filtering has to be determined empirically
- trade-off sensitivity <-> specificity
- which metric of variant call confidence?

Intrinsic

- Transitions:transversions ratio (T_i/T_v)
(e.g. nuclear genes in humans close to 2)



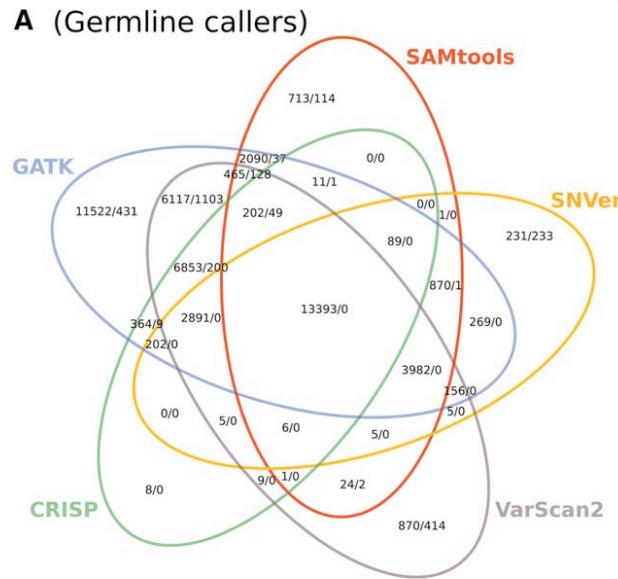
Experimental Validation

- Small-scale validation (Sanger seq, qPCR, pyrosequencing, ...)
- Orthogonal data (e.g. microarrays, different seq platform)
- Concordance among Trios (2 parents + 1 child)

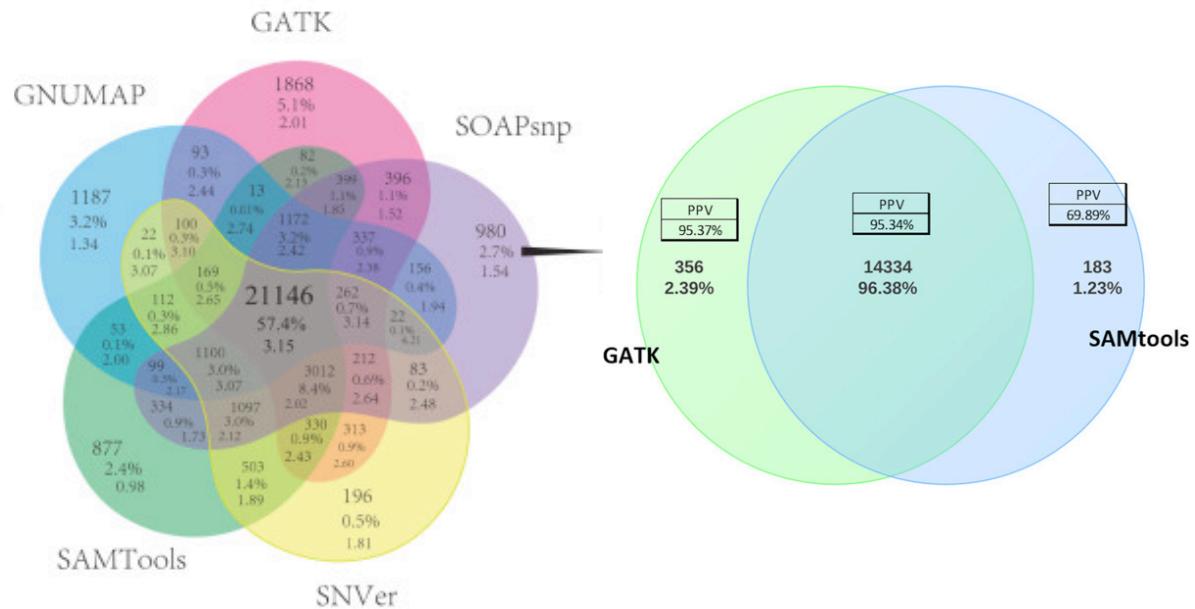
Comparison

- BAM preprocessing steps (e.g., 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?



PMID:23341494



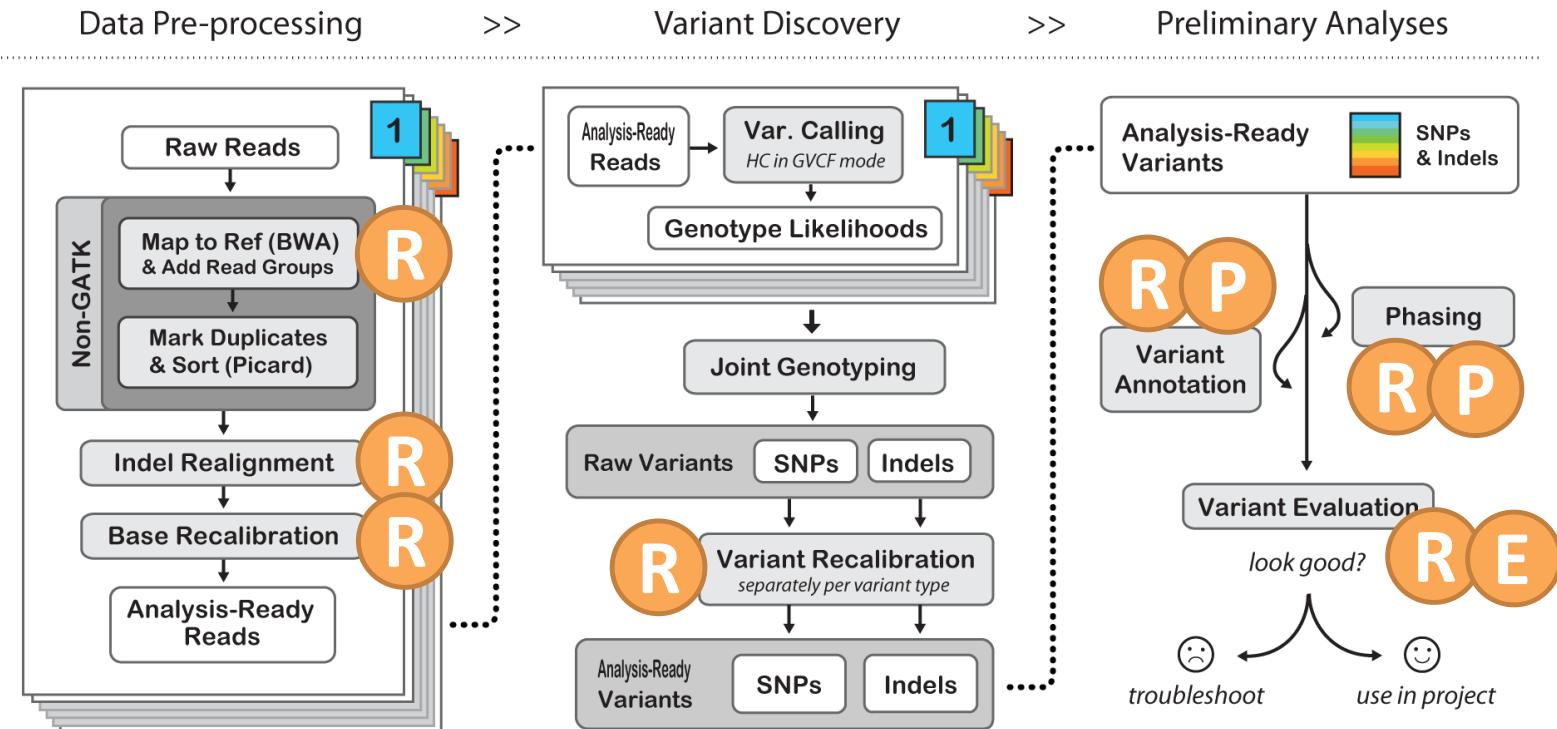
PMID:23537139

PMID:25078893

- depends on who you ask
 - GATK is 'gold-standard' acc to many but often only slightly better
 - overlap between multiple callers? time-consuming!

GATK for non-human organisms

Potential problems



R: Lack of known resources

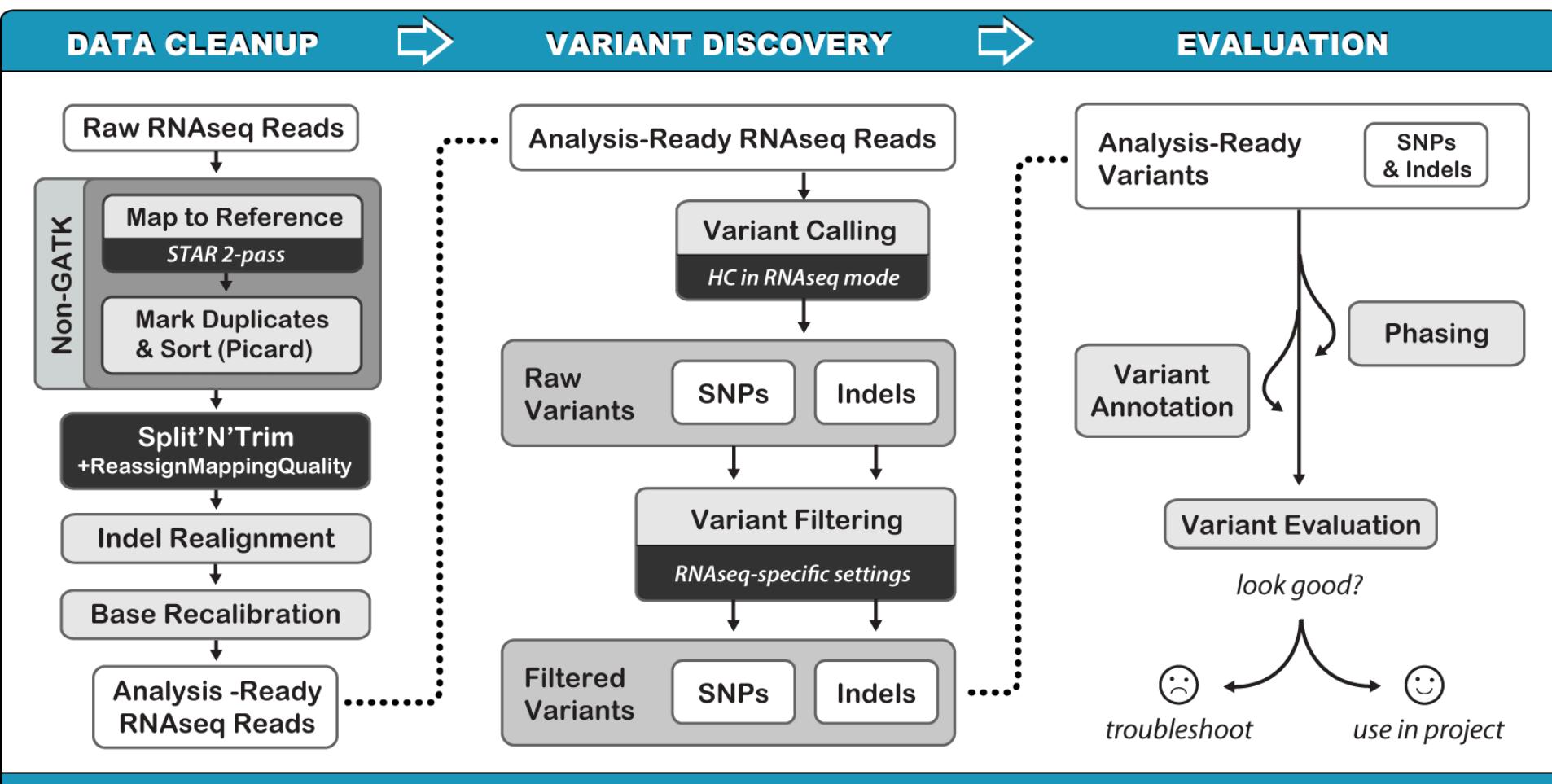
P: Ploidy assumptions in calculations

E: Lack of clear expectations

Non-human organisms

- GATK needs a reference genome
- very slow with many contigs (make supercontigs, remove/mask transposons/repeats)
- Indel Realignment by default uses indels identified in reads (known inDels not required)
- Base Quality Score Recalibration: Bootstrap until convergence:
 1. Call variants on realigned, unrecalibrated data
 2. Filter resulting variants with stringent filters
 3. 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



Summary

- Variants tend to be enriched with artifacts 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 an art): Hard-filtering for non-human organisms

Sources & Links

GATK

- Presentations <https://www.broadinstitute.org/gatk/guide/presentations>
- Documentation <https://www.broadinstitute.org/gatk/guide/>
- Ask the GATK team <http://gatkforums.broadinstitute.org/categories/ask-the-team>

Article Collections

- Review Articles from Nature Reviews Genetics
- PLoS Computational Biology: Education

Material

- SEQanswers NGS forum <http://seqanswers.com/>
- Biostar <http://biostars.org/>
- List of Applications <http://seqanswers.com/wiki/Special:BrowseData/>

FASTQ format & base qualities

@read1

TTGTGTTCAAAATATATAATTATTTATAAGCTATAATCTTATGNNNNNNNCTCCTTAGCTT

+

@C@DDDDDFHHHHJJJDHIIII@HHGGIDGEBDEIEIIIIJJII#####008BGGGHIIGGH>



@ = ASCII code 64

BQ = ASCII code – 33 = **31**

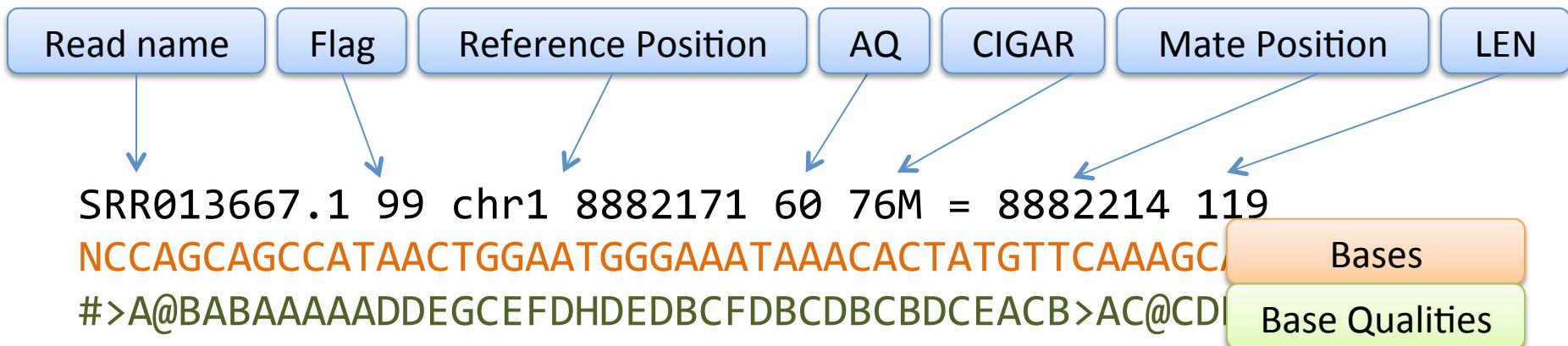
Base Quality: Phred Score Q_{phred}

$$Q_{\text{phred}} = -10 * \log_{10} (P_{\text{error}})$$

Base Quality	P_{error}
3	50 %
5	32 %
10	10 %
20	1 %
30	0.1 %
40	0.01 %

Output Formats: SAM & BAM

- SAM <http://samtools.sourceforge.net/SAMv1.pdf>



- BAM

- binary version of SAM