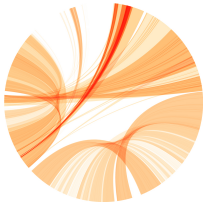
Practical Bioinformatics

Variant Calling 2

Stefan Wyder

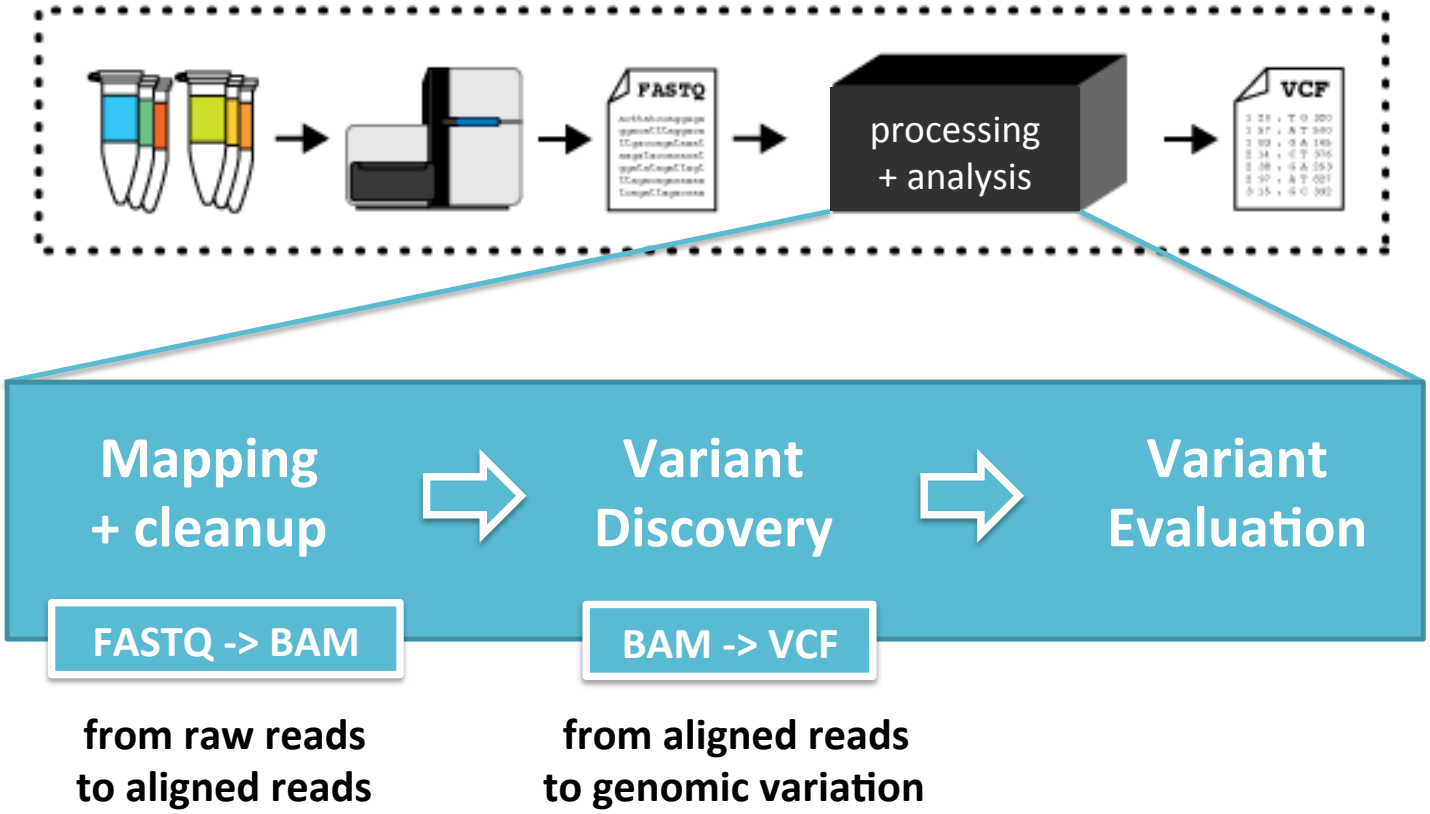
stefan.wyder@uzh.ch

Universität Zürich[UZH]



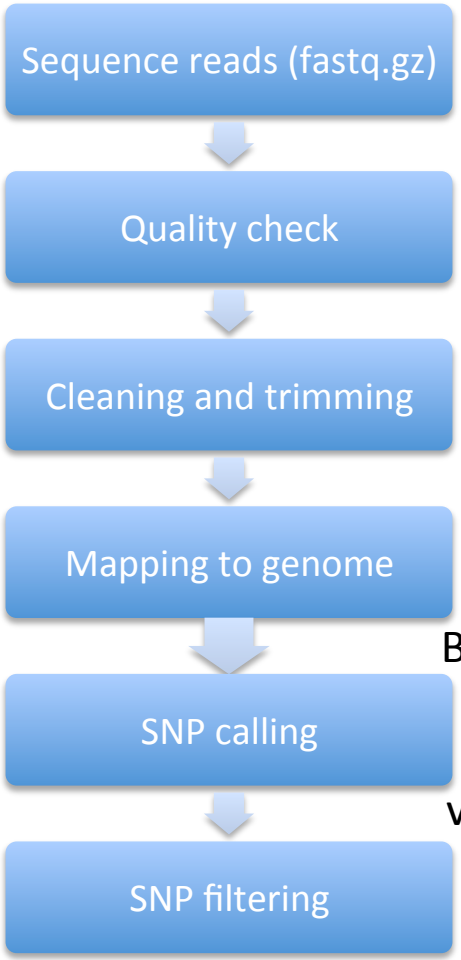
URPP Evolution in Action

From reads to variants



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

Simple workflow



fastQC

trimmomatic/
cutadapt

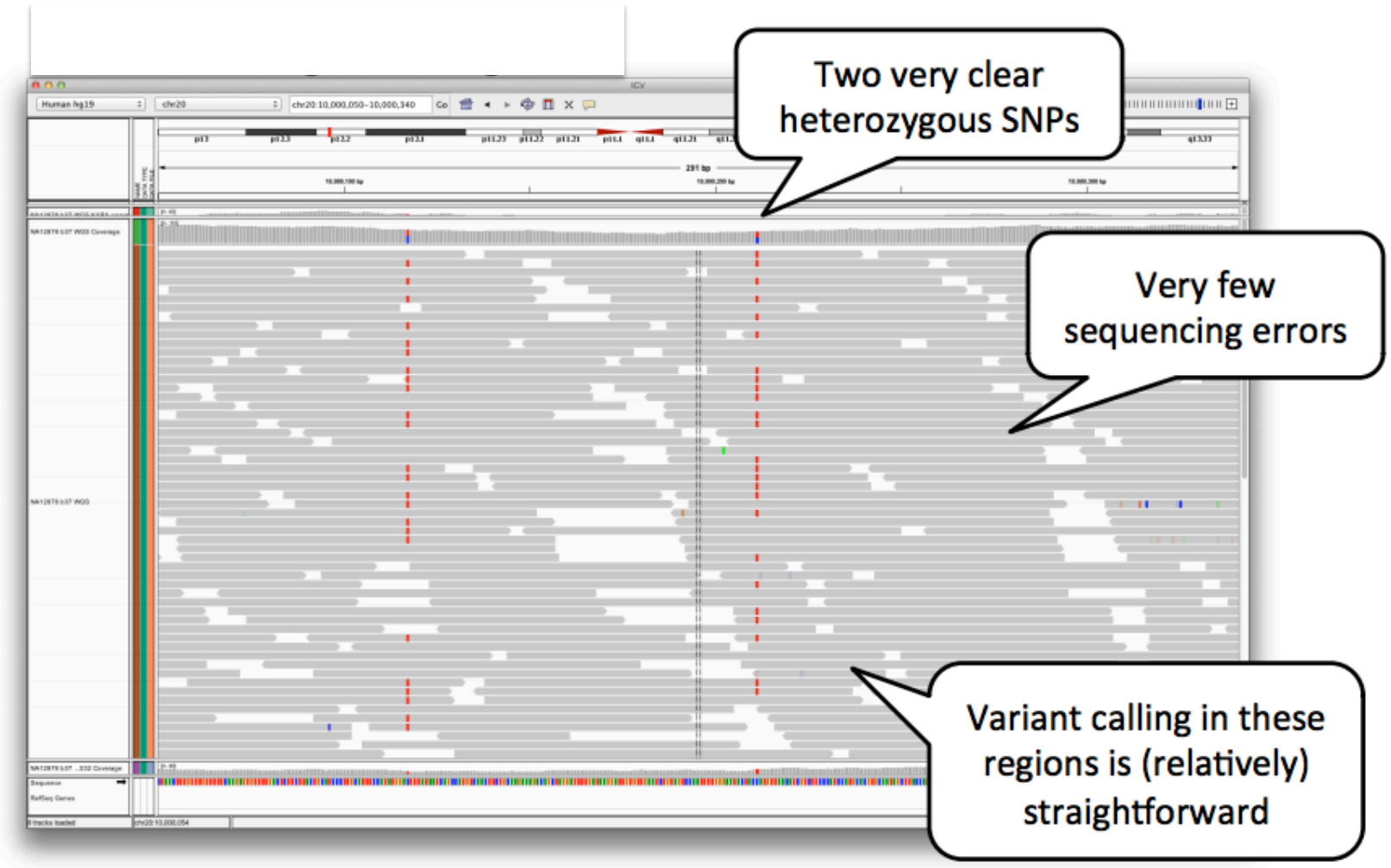
bwa mem

Freebayes /
bcftools

bcftools

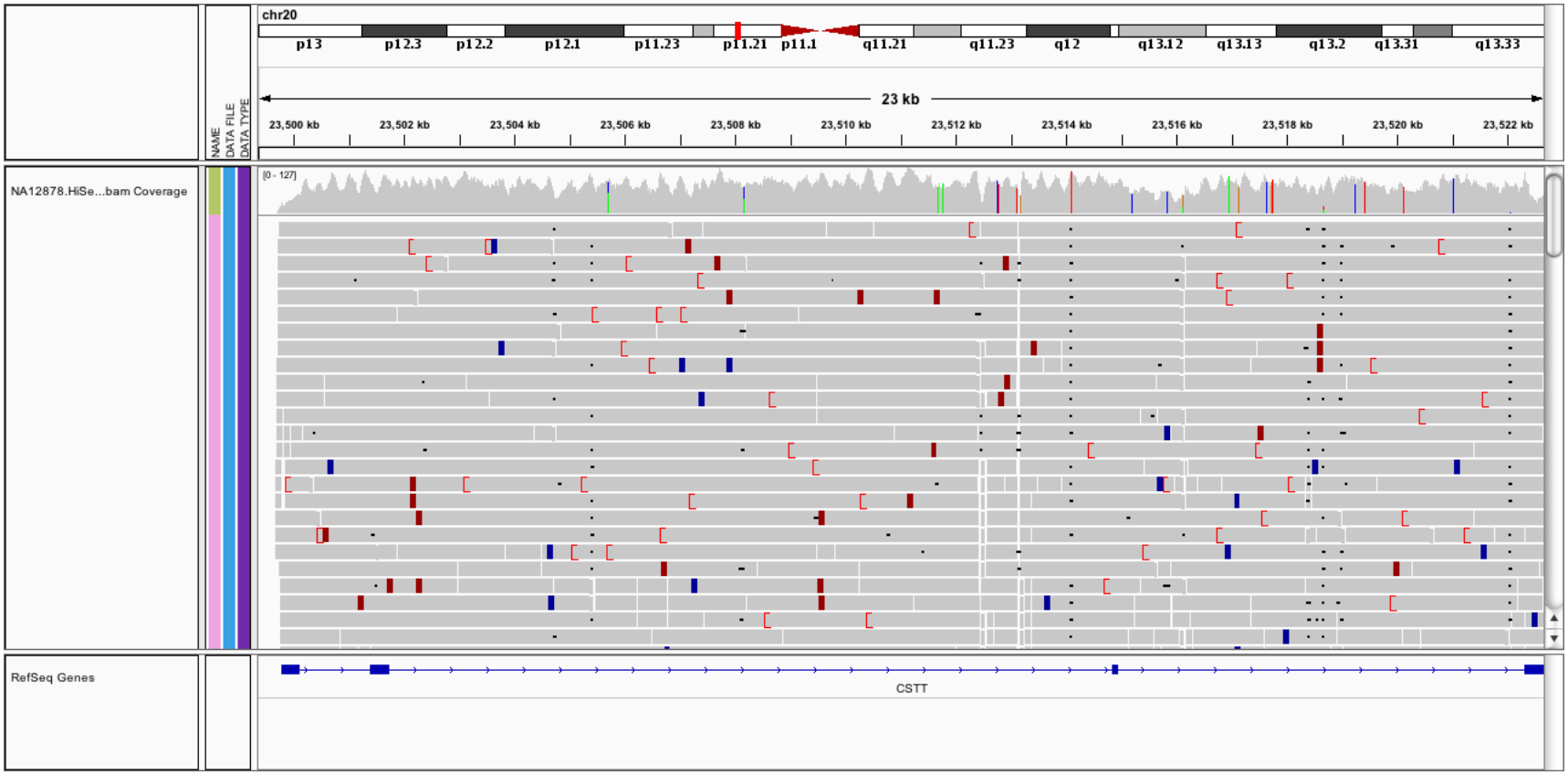
Simply counting?

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



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

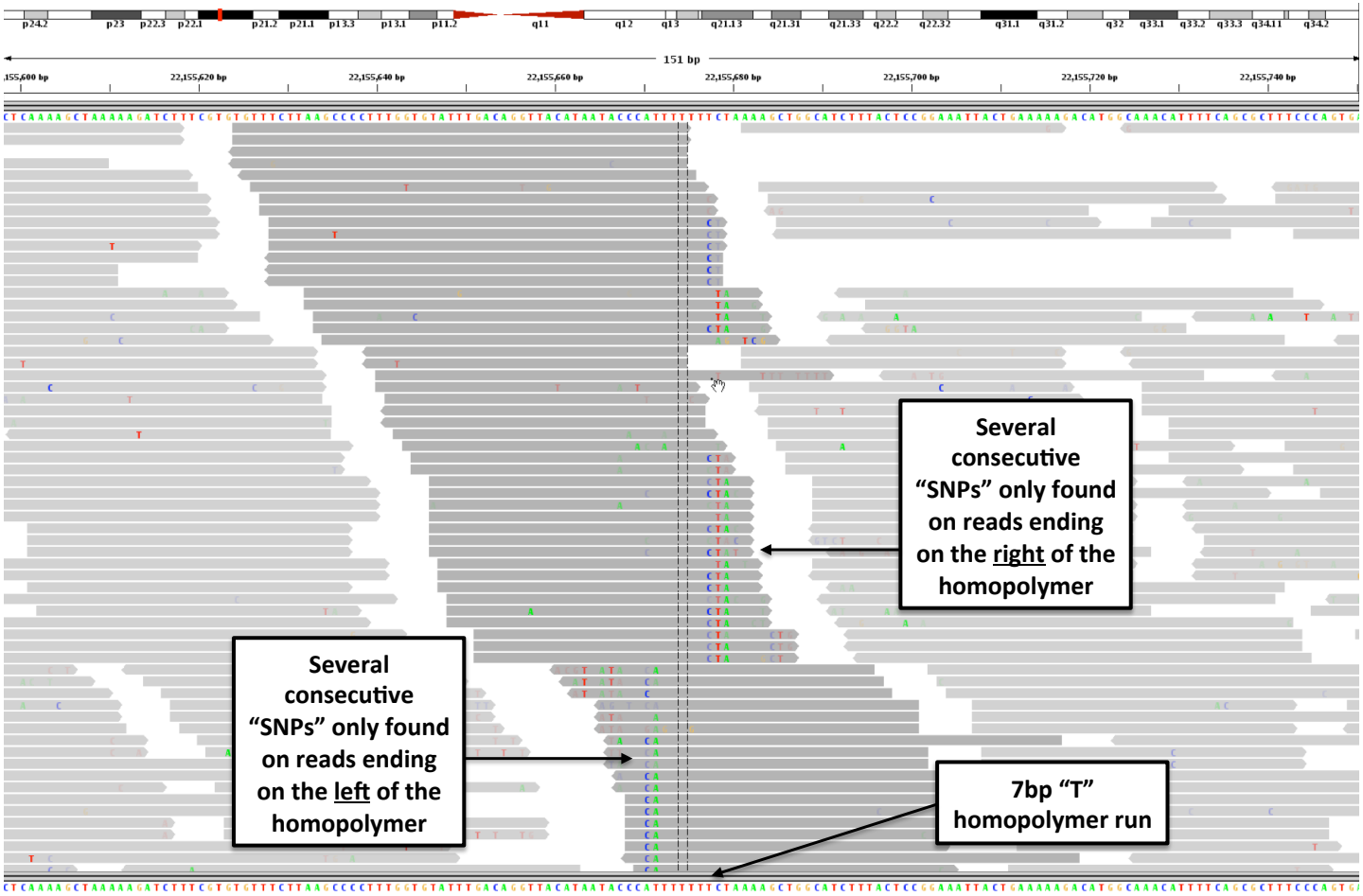
...Messy situations...



Real mutations or noise?

https://www.broadinstitute.org/gatk

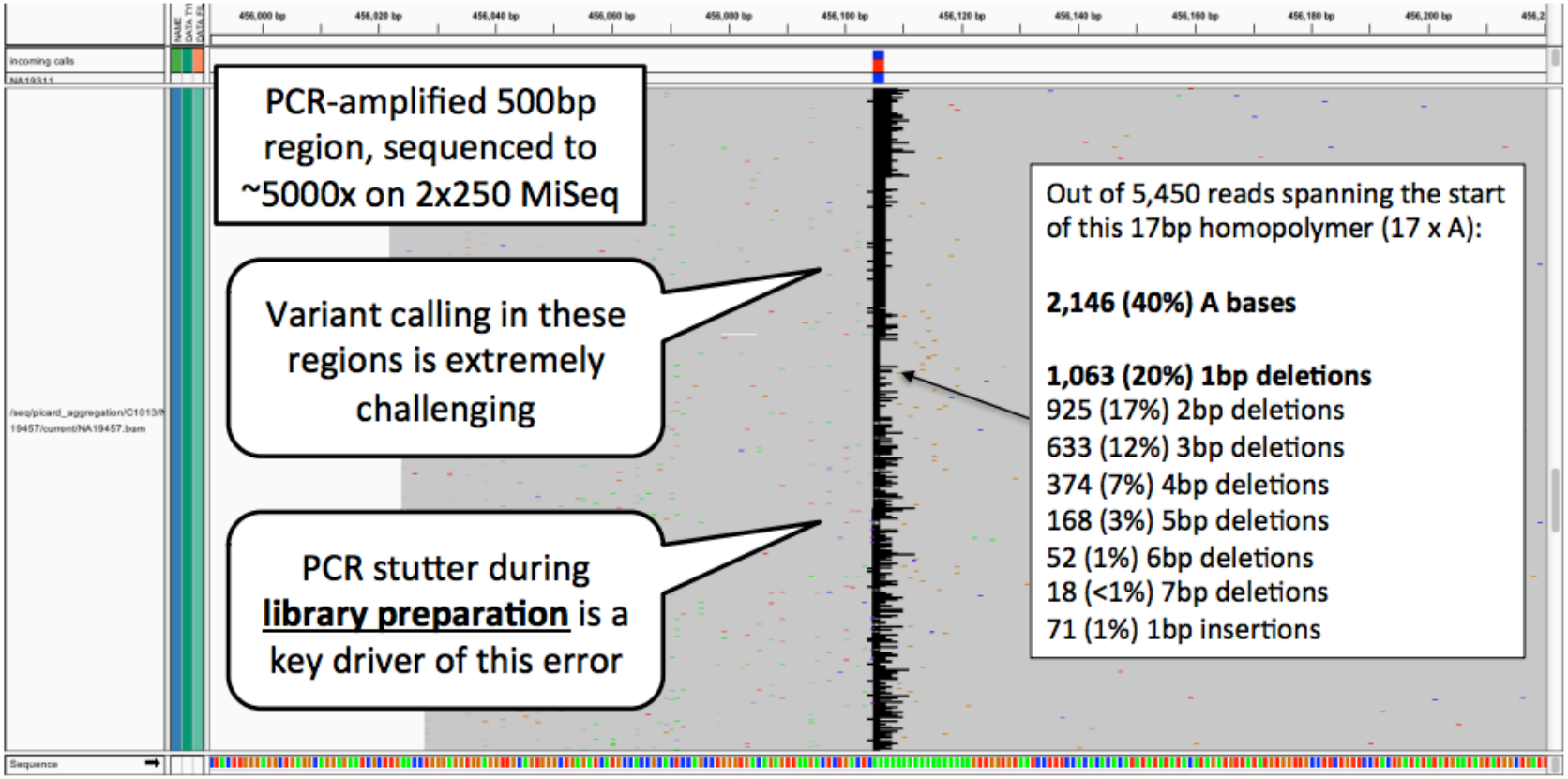
An example of a strand-discordant locus



https://www.broadinstitute.org/gatk

…it can get even worse

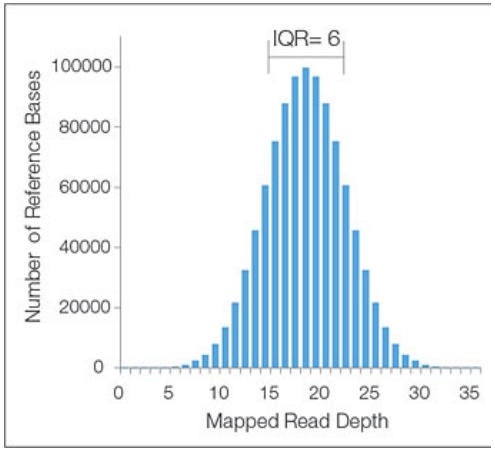
Poorly-behaved region of the genome



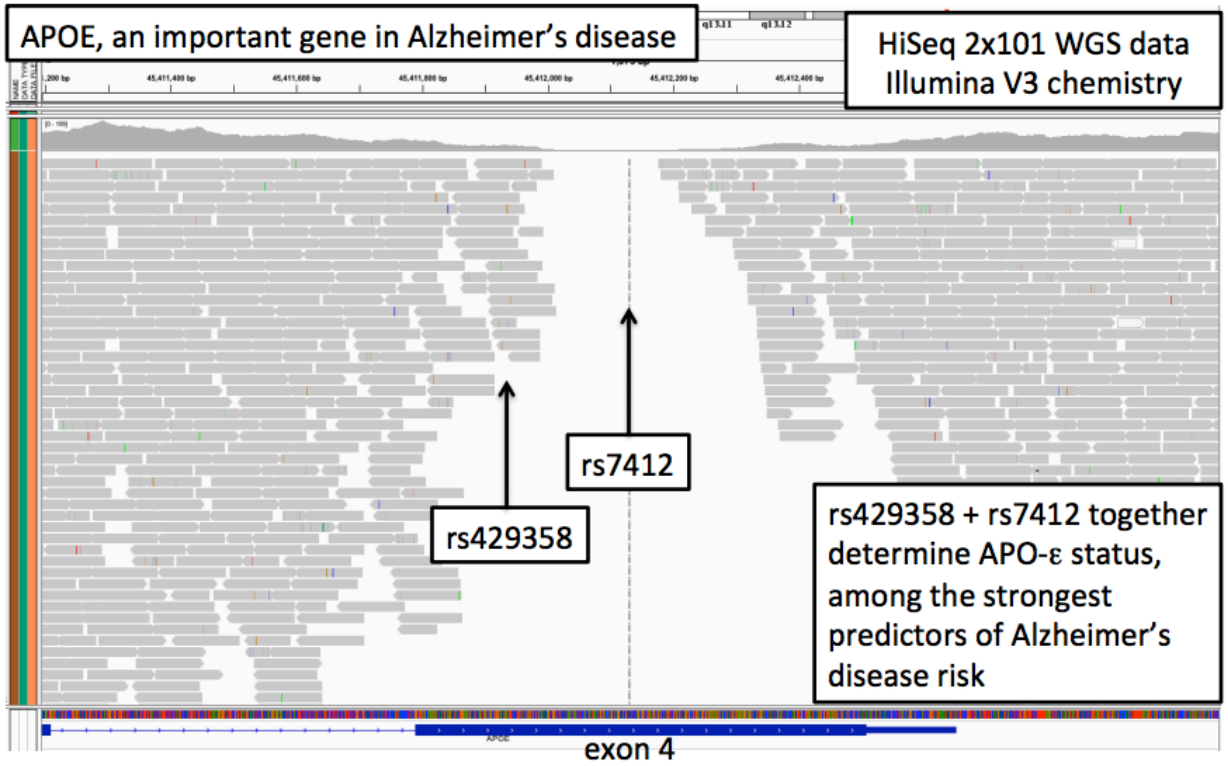
http://www.broadinstitute.org/gatk//events/2247/agbt_2013_depristo.pdf

Problem 1: Lack of coverage

Read depth histogram



Illumina

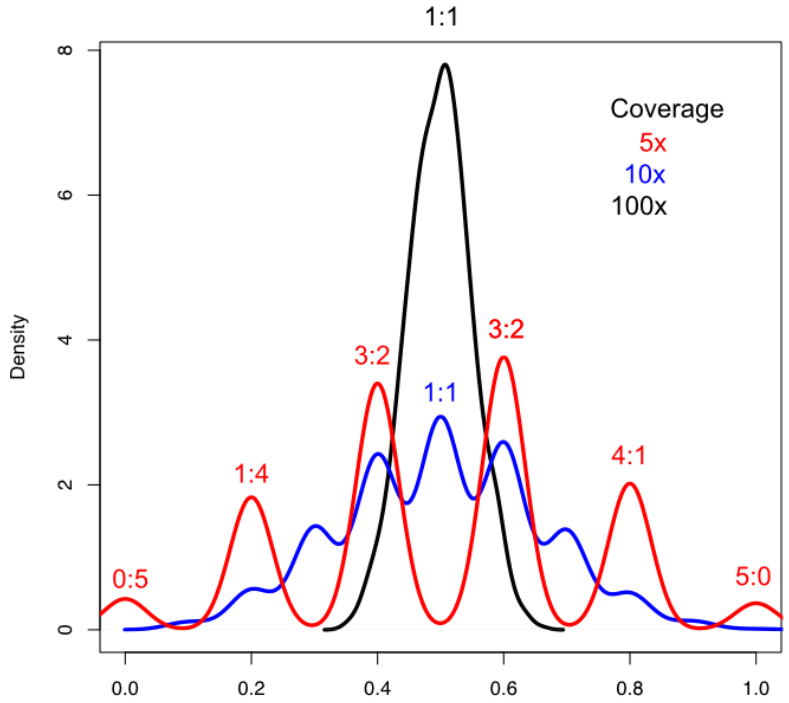


http://www.broadinstitute.org/gatk//events/2247/agbt_2013_depristo.pdf

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

Problem 2: Random 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 !

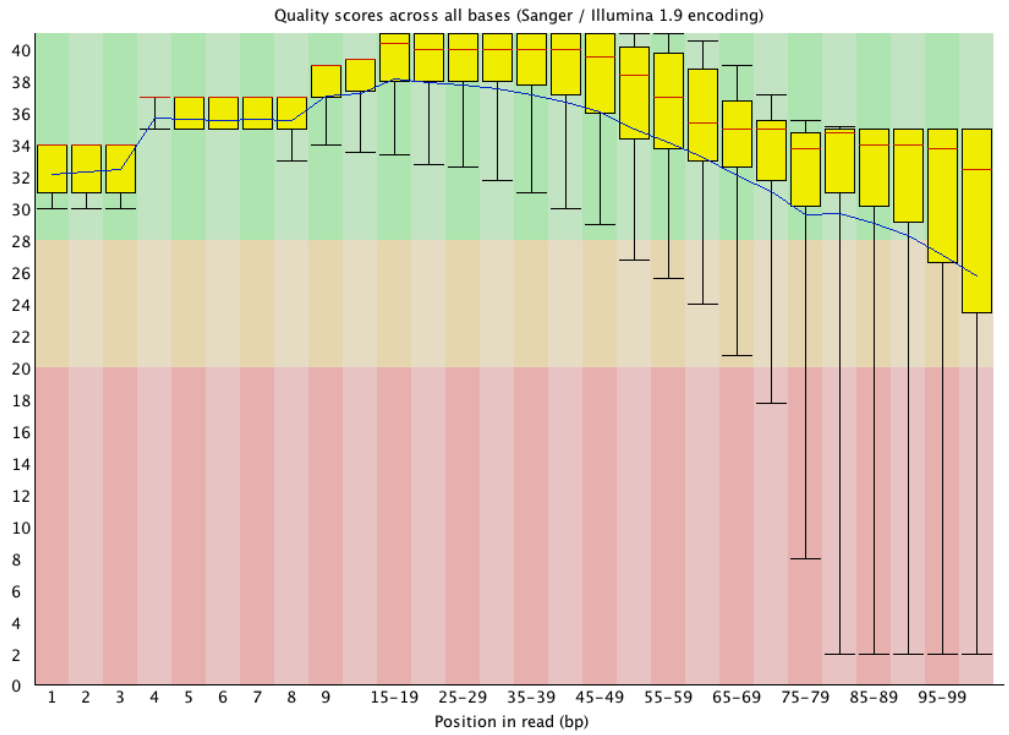
Problem 3: Sequencing errors

Error rate and error profile are technology-specific

Illumina Sequencing

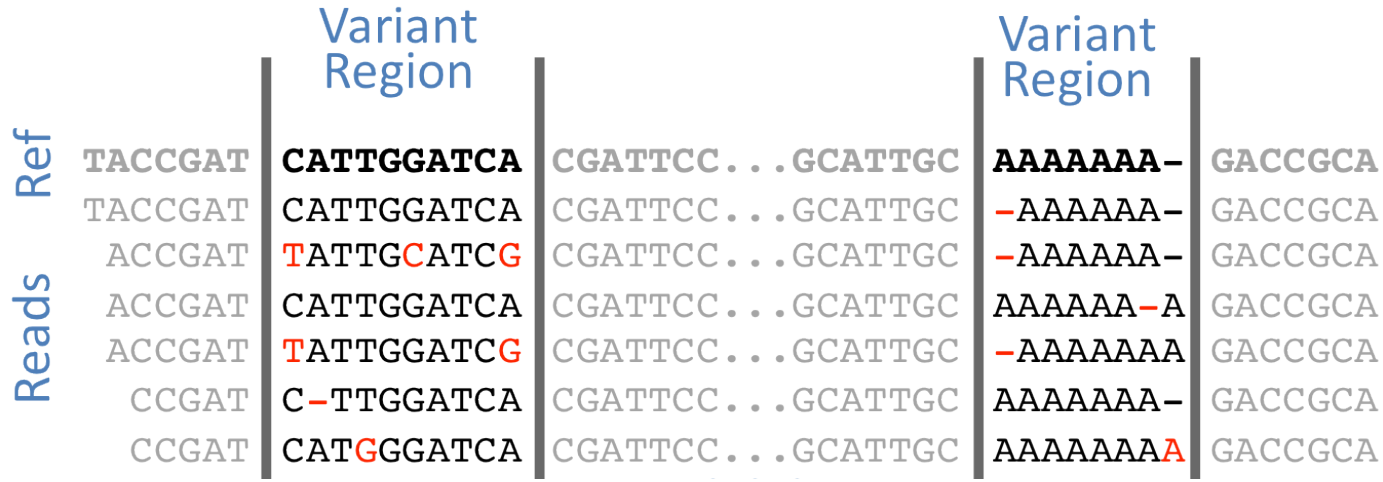
- Error Rate: > 0.1% (i.e. > 1 in 1000)
- mainly substitutions errors
- errors mostly at read's start and end

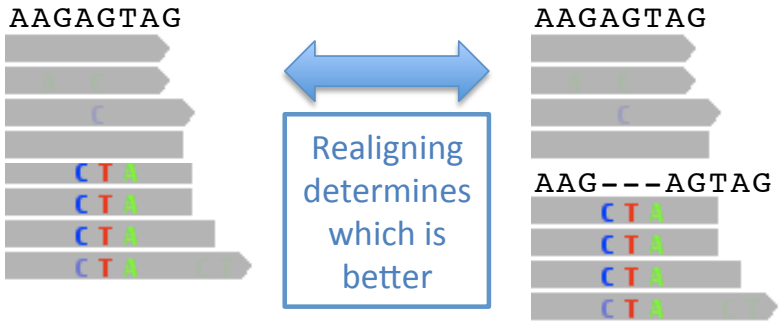
random?



Problem 4: Incorrect mapping

With indels multiple alignments are possible





-> Indel Realignment

Problem 4: Incorrect mapping cont.

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

- incomplete/missassembled reference genome

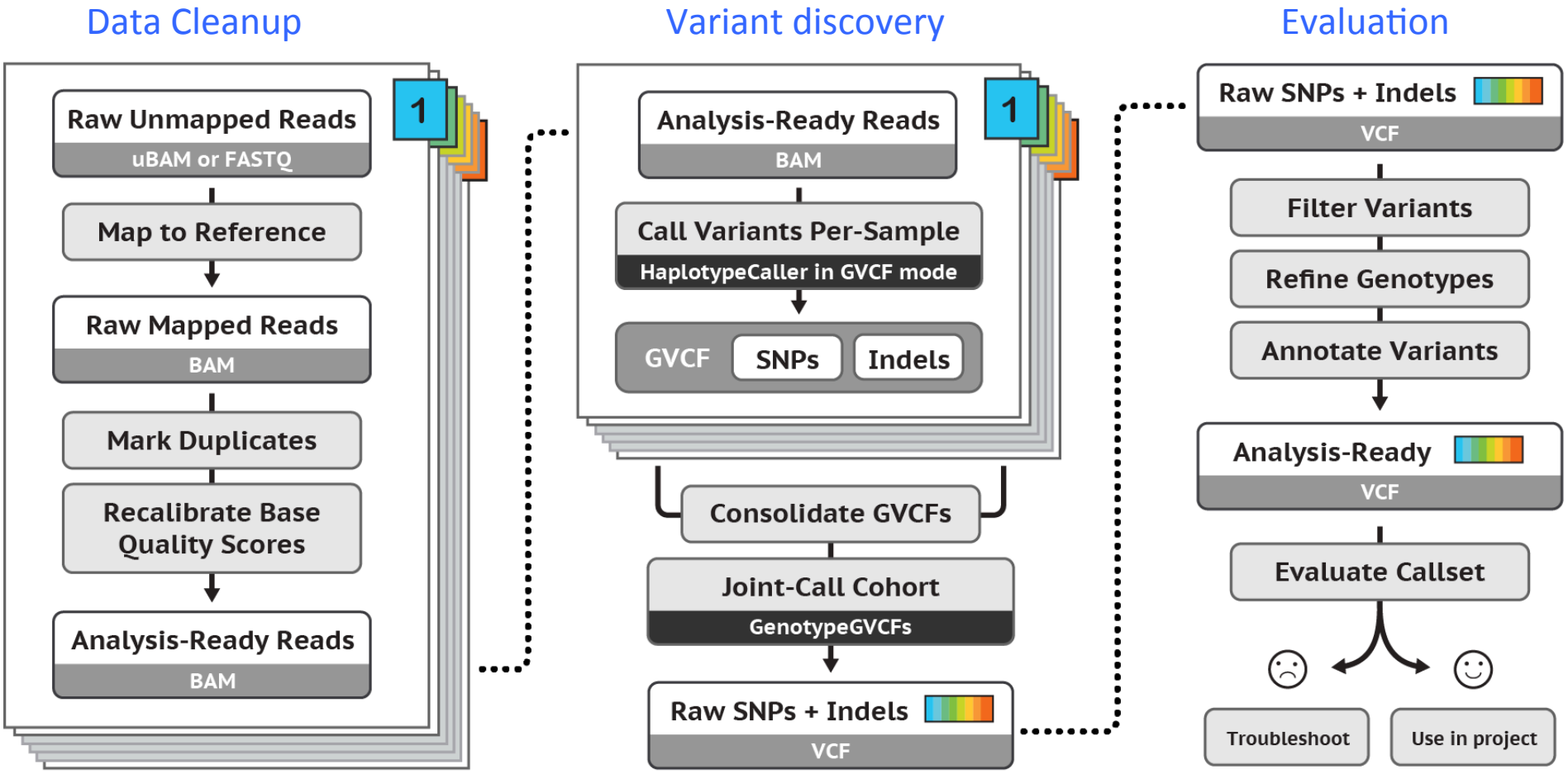
GATK

- 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 (germline)

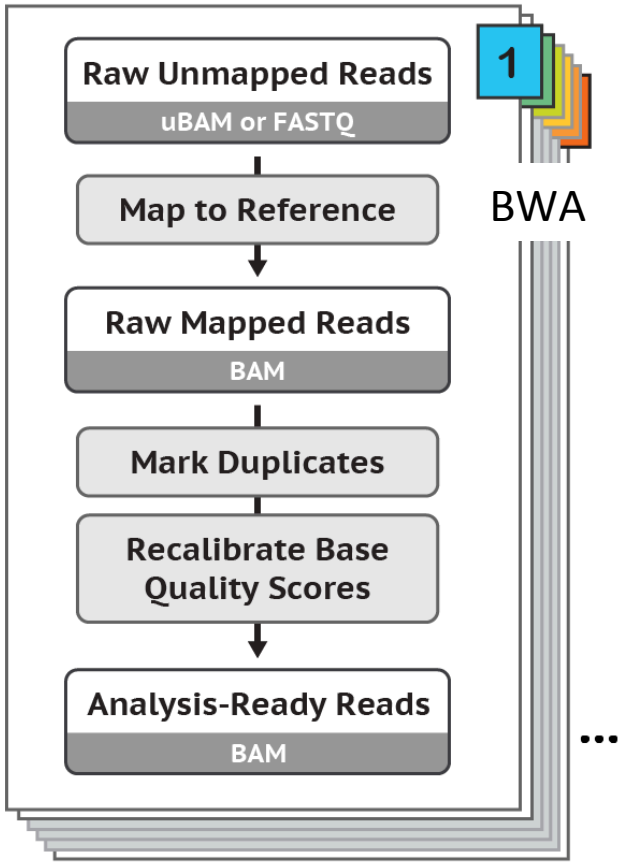


per lane (read group)

per sample

https://software.broadinstitute.org/gatk/best-practices/

Data Cleanup / Pre-processing



per sample

Correct for technical biases

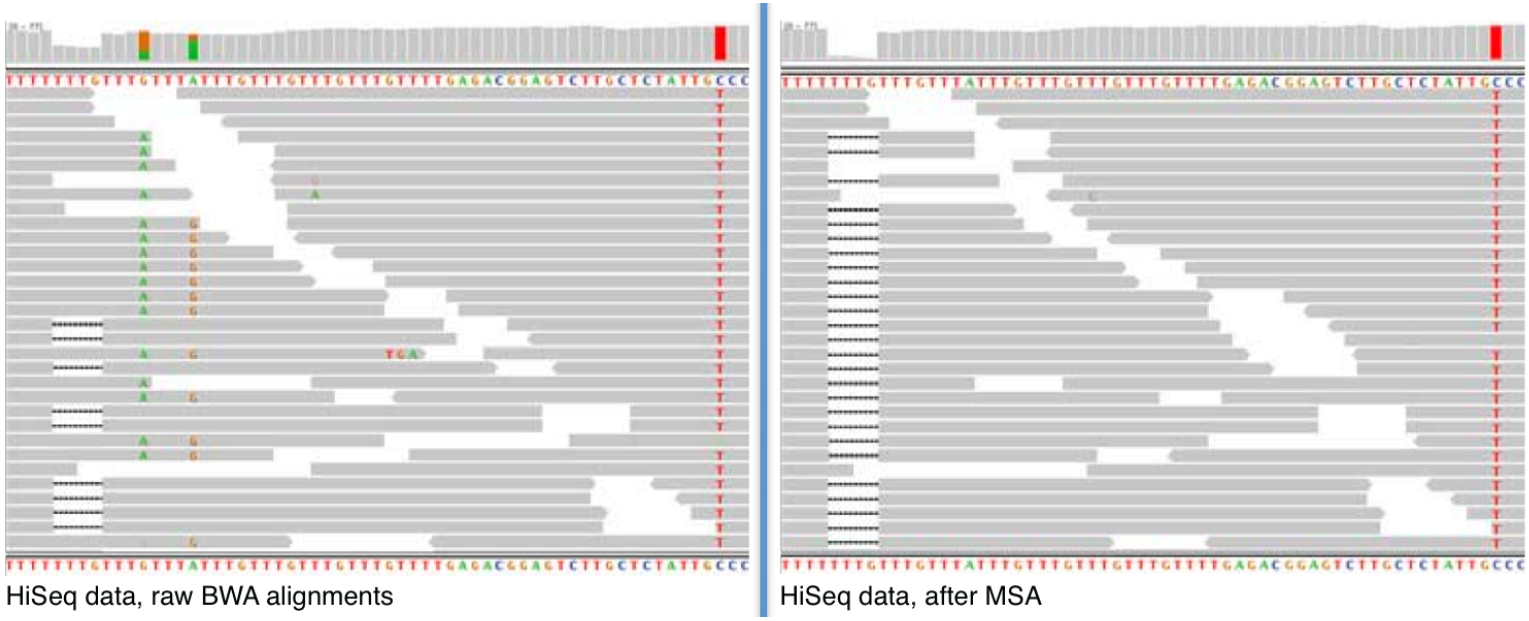
Mark PCR Duplicates

- come from same input DNA template- have same start position on reference
- non-independent measurements violate statistical assumptions
- not applicable in amplicon seq

Indel Realignment (may disappear)

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

Before and after indel realignment

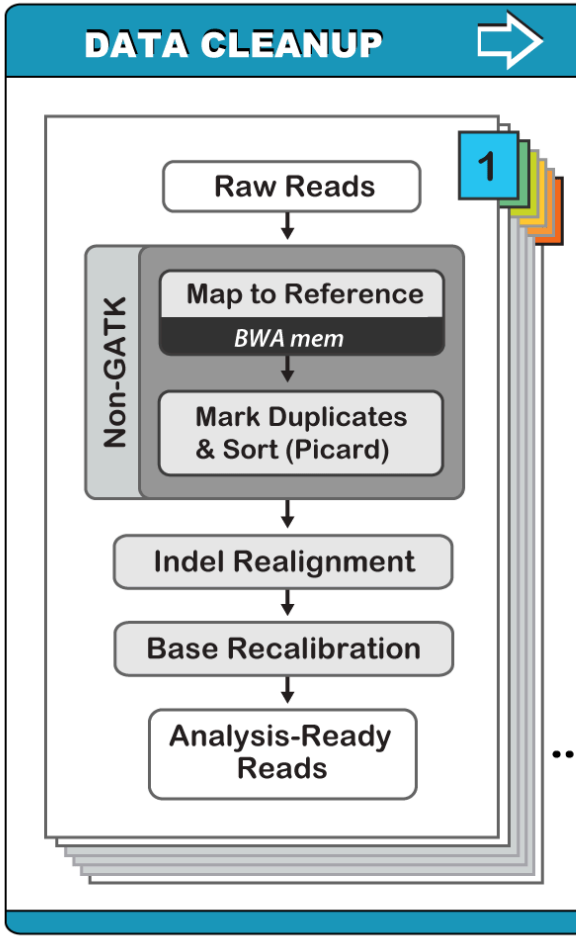


HiSeq data, raw BWA alignments

HiSeq data, after MSA

https://www.broadinstitute.org/gatk

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

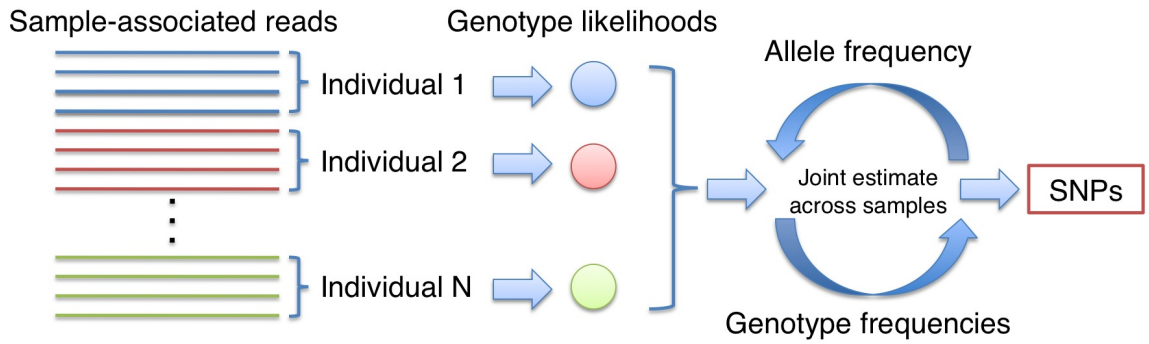
Variant Calling

- modelling various error types
- expected distribution of calls
  (homozygous AA, homozygous variant BB, heterozygous AB)
- from GATK v3.3 HaplotypeCaller is recommended for all cases

HaplotypeCaller
calls SNPs and indels simultaneously
performs a local de-novo assembly
any ploidy
more accurate (especially for indels)
up to 100s samples (-> GVCF mode)

Multi-sample analysis

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

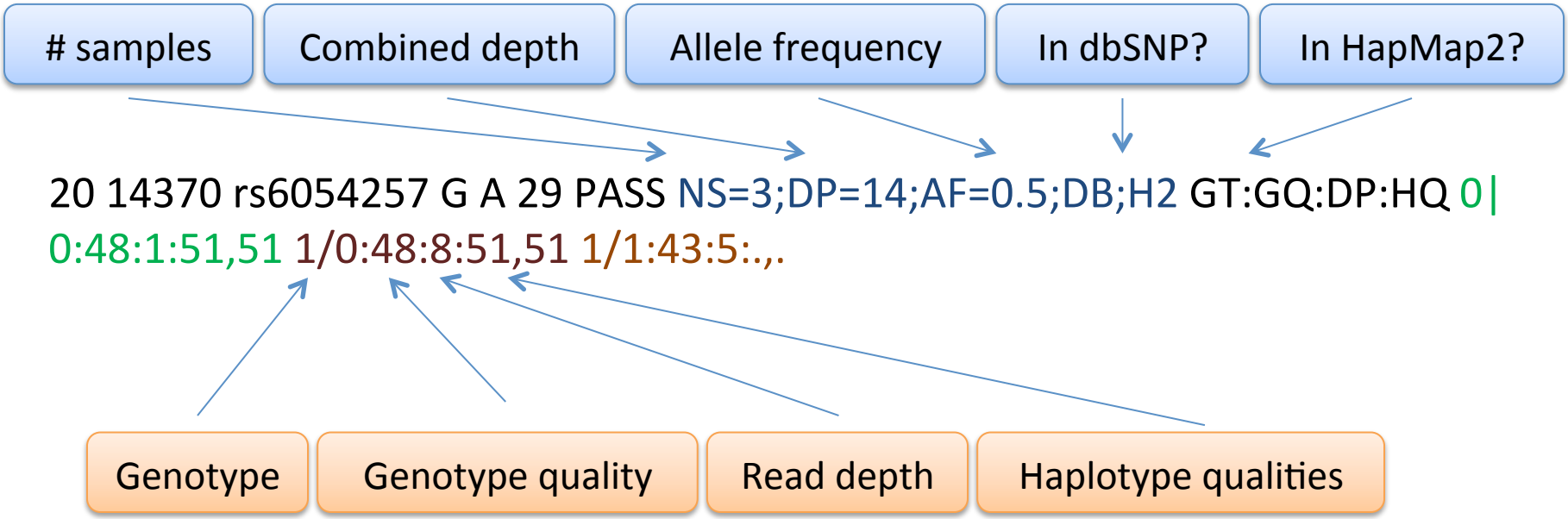


- ~ Hardy-Weinberg Equilibrium
- Genotypings like this: AB, AB, AB, AB, AB, AB have much lower probability than AA, AA, AB, BB, AA, AB, AA
- in reality: multiple alleles,...

VCF

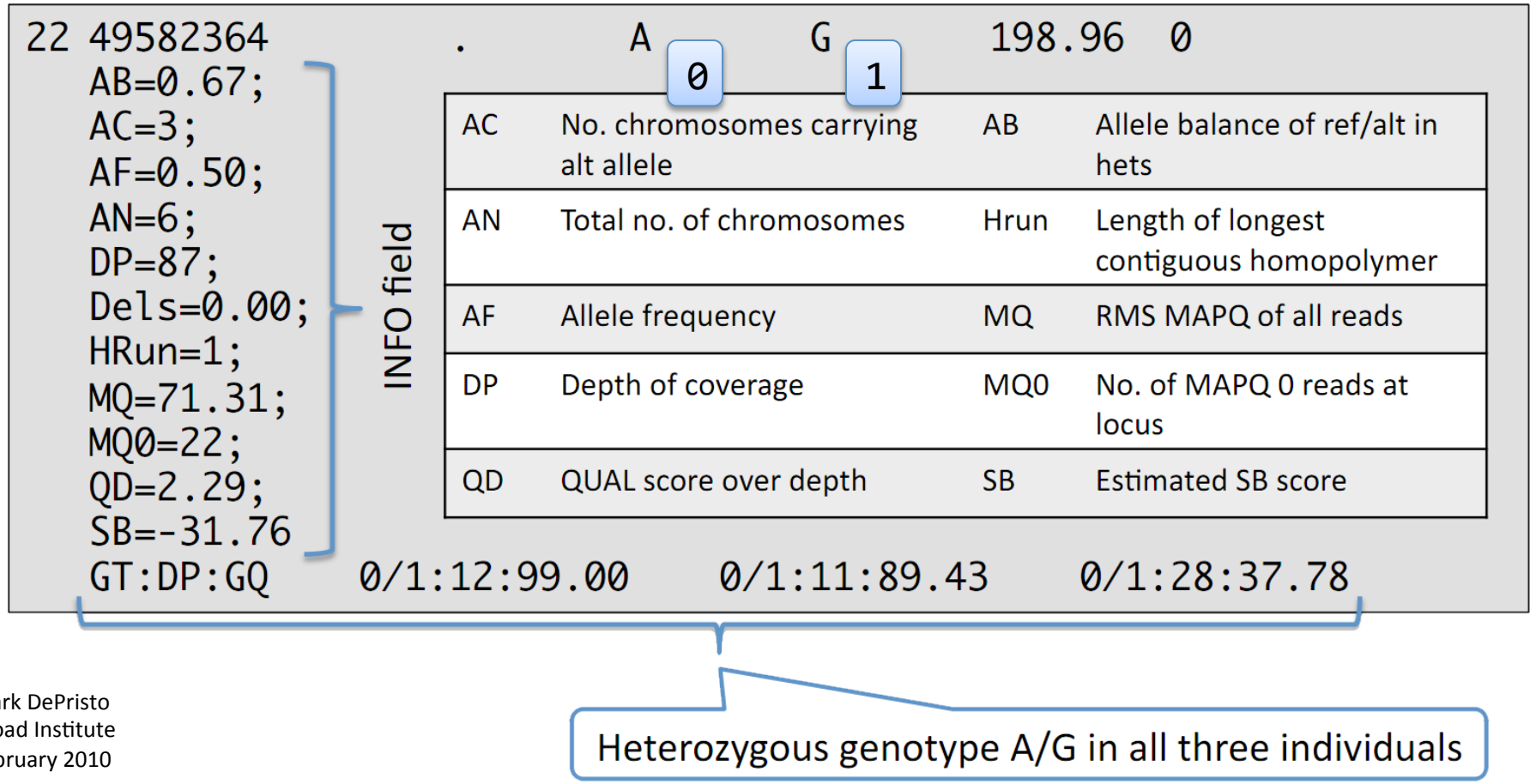
##fileformat=VCFv4.0

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



VCF info field

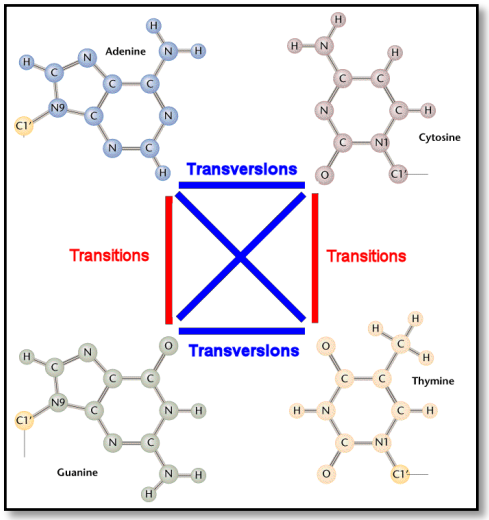
VCF record for an A/G SNP at 22:49582364



Mark DePristo
Broad Institute
February 2010

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 (Ti/Tv)
  (e.g. nuclear genes in humans close to 2)

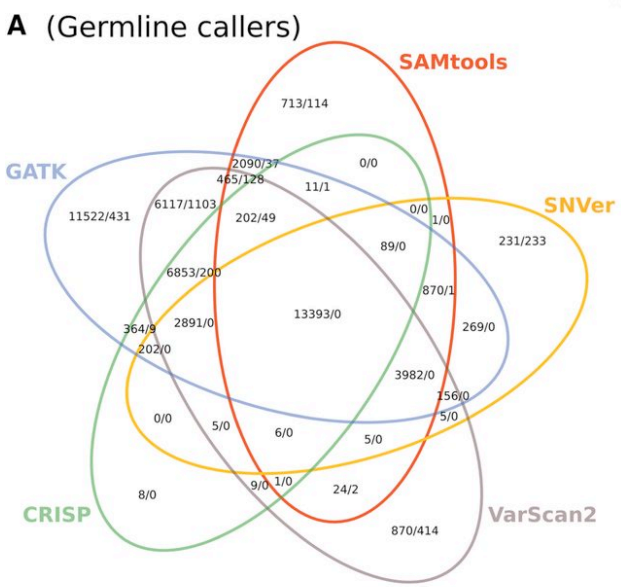
Experimental Validation

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

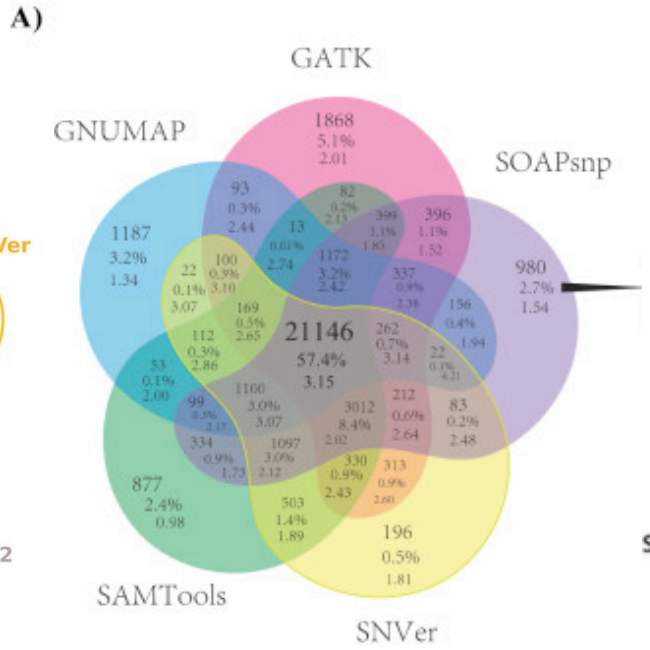
Comparison

- 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)

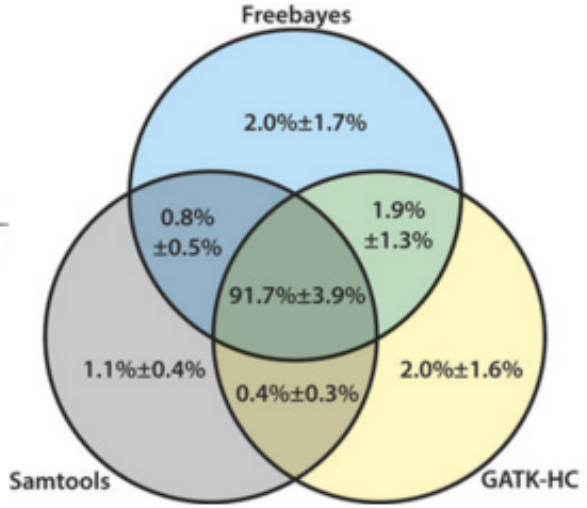
Who performs best?



PMID:23341494



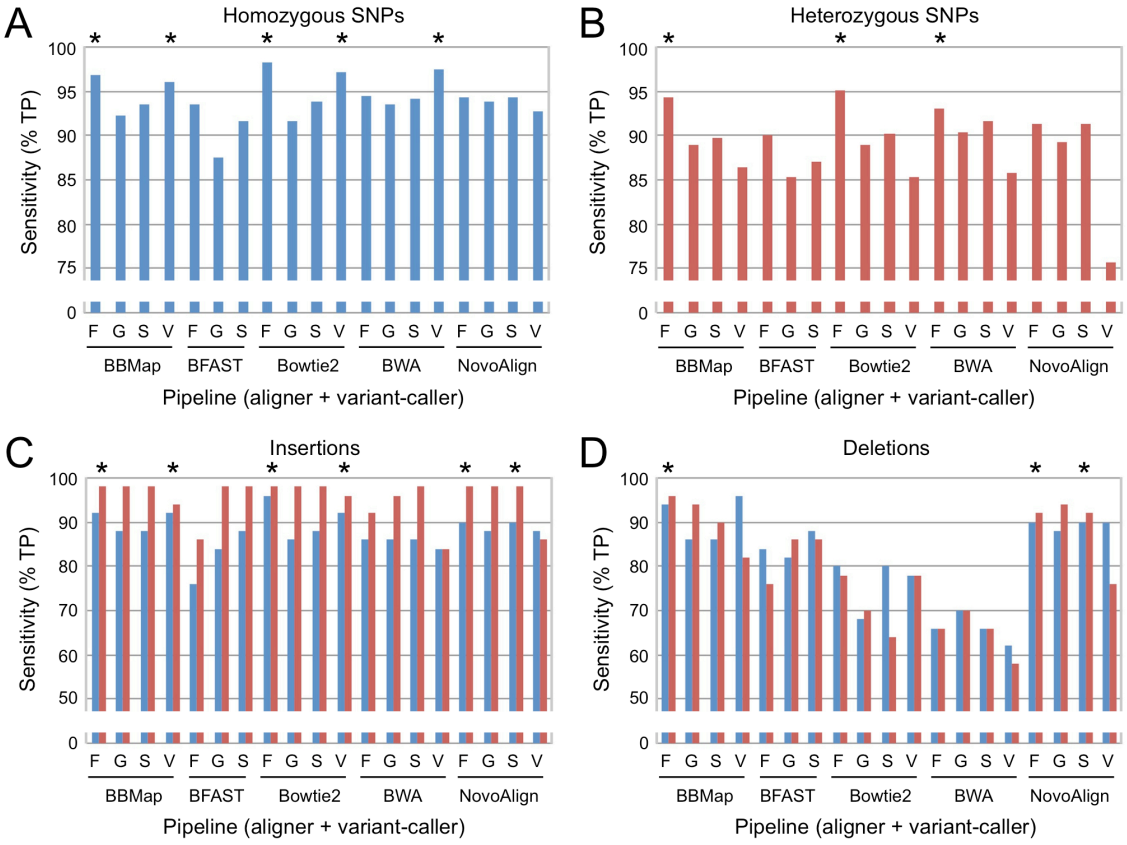
PMID:23537139



PMID:26639839

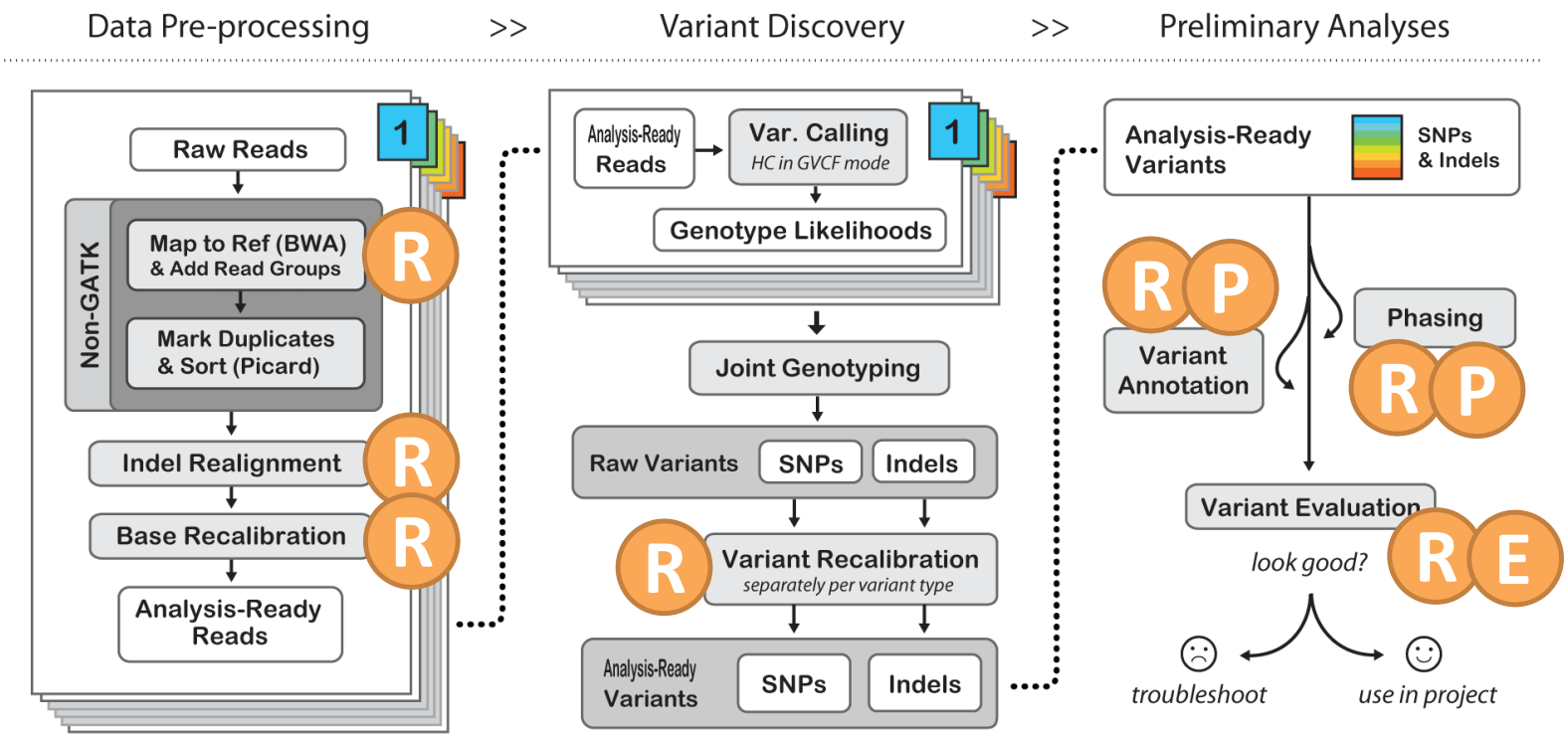
- depends on who you ask
- GATK is 'gold-standard' acc to many but often only sligthly better
- overlap between multiple callers? time-consuming!

Combinations of mappers-callers



PMID:28333980

GATK for non-human organisms
Potential problems

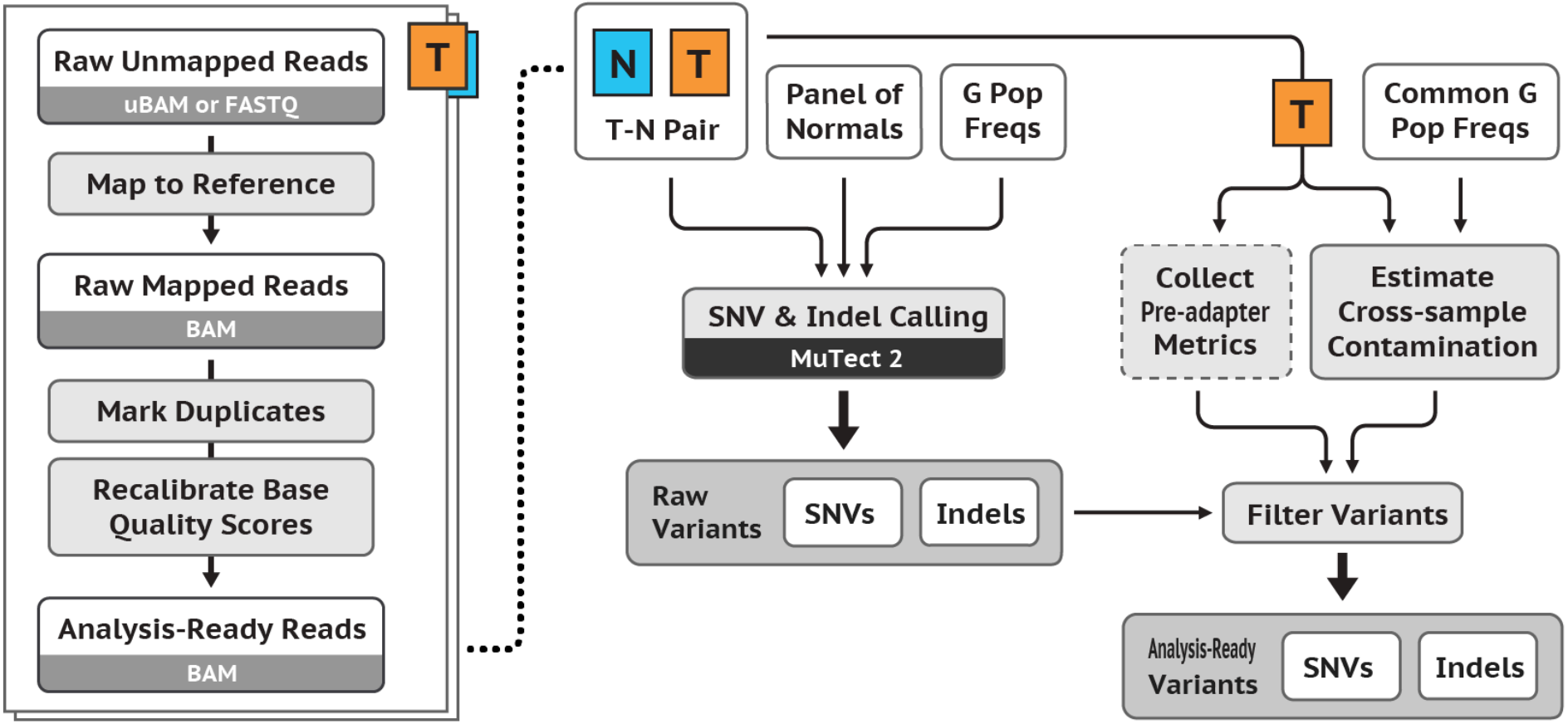


R: Lack of known resources
P: Ploidy assumptions in calculations
E: Lack of clear expectations

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

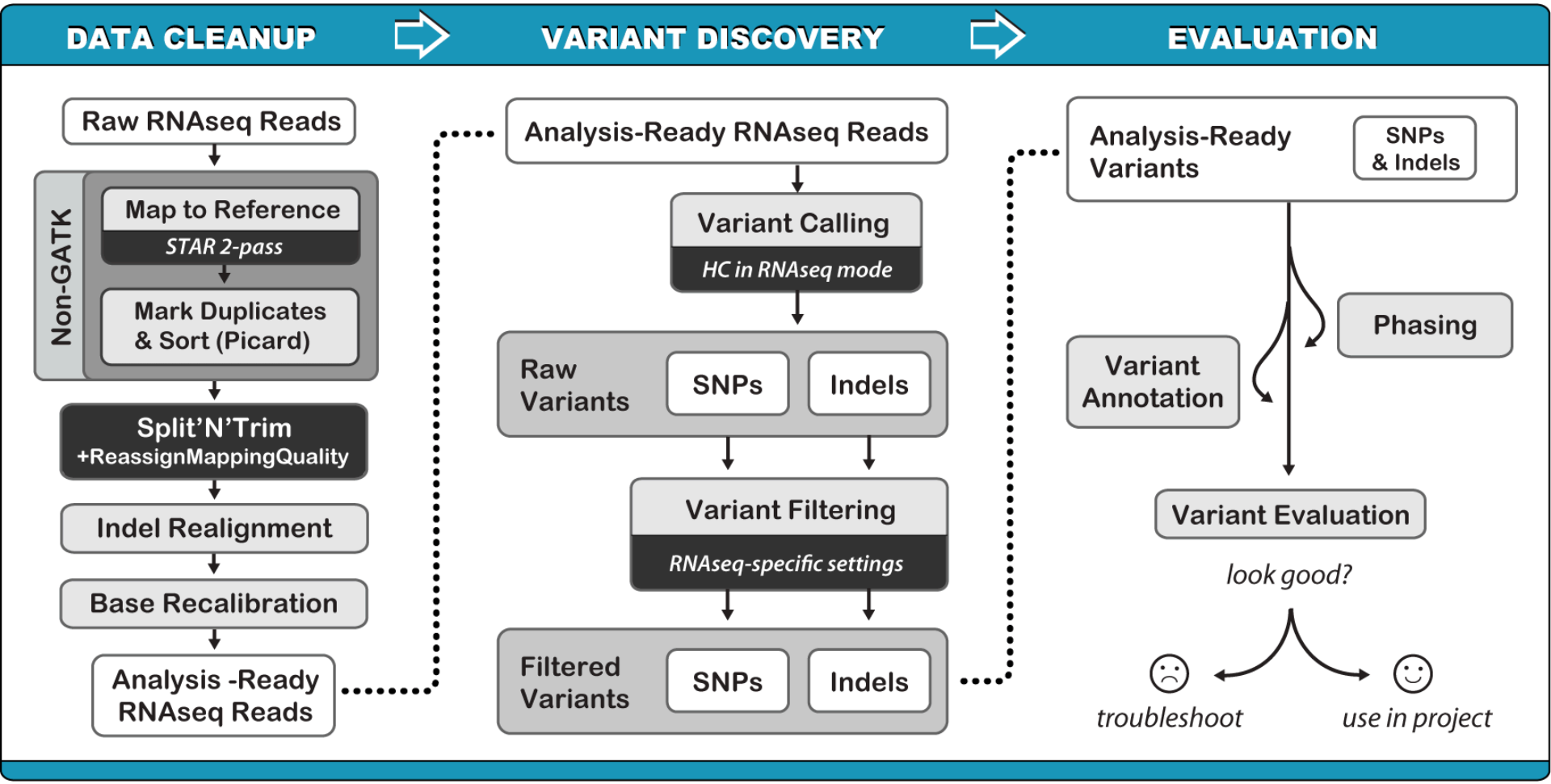
somatic GATK variant discovery



for tumors
requires a panel of normals

https://software.broadinstitute.org/gatk/best-practices/

GATK variant discovery for RNA-seq

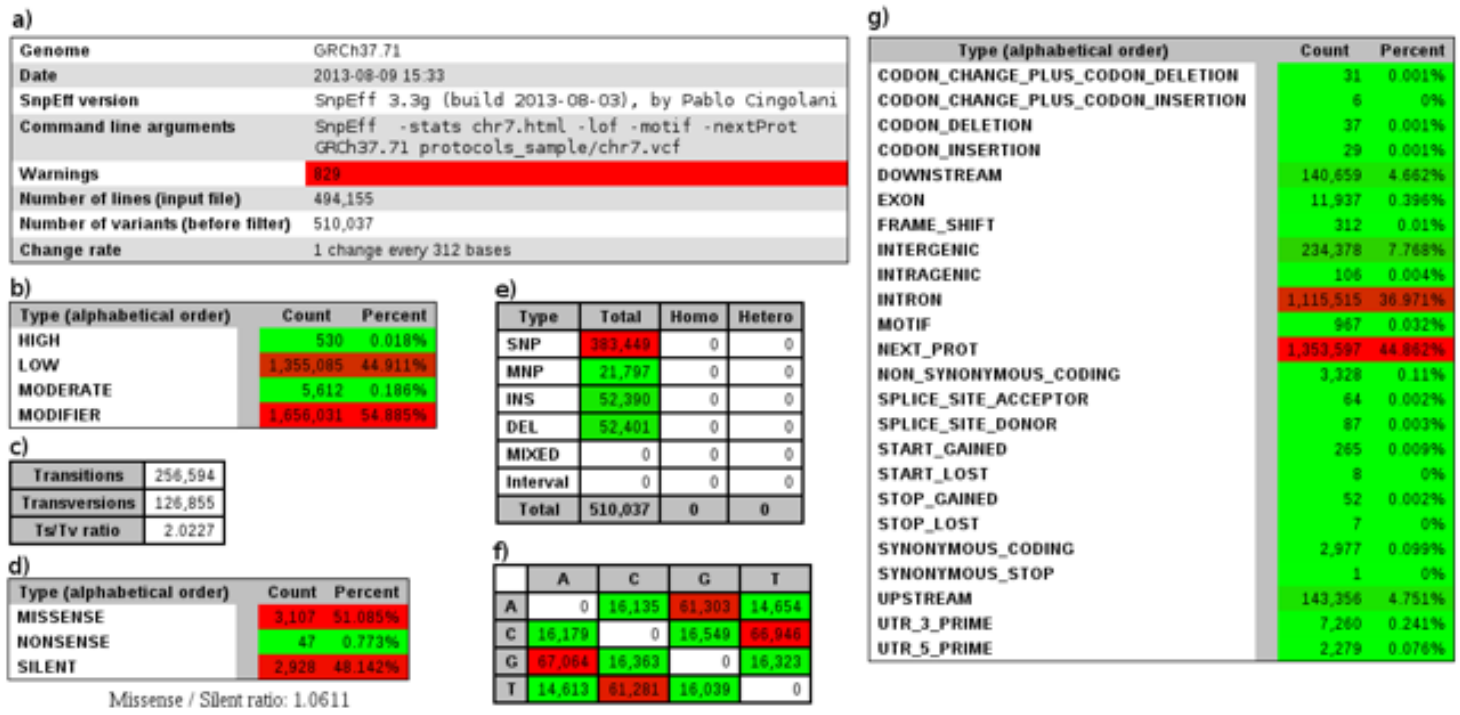


in development for GATK 4

https://software.broadinstitute.org/gatk/best-practices/

Variant Effect Prediction

- software tools that annotate and predict the effects of variants on genes (e.g. SnpEff & SnpSift, ensembl VEP)



a)

Genome	GRCh37.71
Date	2013-08-09 15:33
SnpEff version	SnpEff 3.3g (build 2013-08-03), by Pablo Cingolani
Command line arguments	SnpEff -stats chr7.html -lof -motif -nextProt GRCh37.71 protocols_sample/chr7.vcf
Warnings	829
Number of lines (input file)	494,155
Number of variants (before filter)	510,037
Change rate	1 change every 312 bases

b)

Type (alphabetical order)	Count	Percent
HIGH	530	0.018%
LOW	1,355,085	44.911%
MODERATE	5,612	0.186%
MODIFIER	1,656,031	54.885%

c)

Transitions	256,594
Transversions	126,855
Ts/Tv ratio	2.0227

d)

Type (alphabetical order)	Count	Percent
MISSENSE	3,107	51.085%
NONSENSE	47	0.773%
SILENT	2,928	48.142%

Missense / Silent ratio: 1.0611

e)

Type	Total	Homo	Hetero
SNP	383,449	0	0
MNP	21,797	0	0
INS	52,390	0	0
DEL	52,401	0	0
MIXED	0	0	0
Interval	0	0	0
Total	510,037	0	0

f)

	A	C	G	T
A	0	16,135	61,303	14,654
C	16,179	0	16,549	66,946
G	67,064	16,363	0	16,323
T	14,613	61,281	16,039	0

g)

Type (alphabetical order)	Count	Percent
CODON_CHANGE_PLUS_CODON_DELETION	31	0.001%
CODON_CHANGE_PLUS_CODON_INSERTION	6	0%
CODON_DELETION	37	0.001%
CODON_INSERTION	29	0.001%
DOWNSTREAM	140,659	4.662%
EXON	11,937	0.396%
FRAME_SHIFT	312	0.01%
INTERGENIC	234,378	7.768%
INTRAGENIC	106	0.004%
INTRON	1,115,515	36.971%
MOTIF	967	0.032%
NEXT_PROT	1,353,597	44.862%
NON_SYNONYMOUS_CODING	3,328	0.11%
SPLICE_SITE_ACCEPTOR	64	0.002%
SPLICE_SITE_DONOR	87	0.003%
START_GAINED	265	0.009%
START_LOST	8	0%
STOP_GAINED	52	0.002%
STOP_LOST	7	0%
SYNONYMOUS_CODING	2,977	0.099%
SYNONYMOUS_STOP	1	0%
UPSTREAM	143,356	4.751%
UTR_3_PRIME	7,260	0.241%
UTR_5_PRIME	2,279	0.076%

SnpEff

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 difficult): Hard-filtering for non-human organisms
- Required precision depends on application
  Population Genetics < Mutagenesis << Diagnosis

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
TTGTGTTCAAAATATATAATTTATTTATAAGCTATAATCTTATGNNNNNNNCTCCTTCTTAGCTT
+
@C@DDDDDFHHHHJJDHIIIJI@HHGGIDGEBDEIEIIIIJJII#######008BGGGHIIGGH>
```

@ = ASCII code 64
BQ = ASCII code – 33 = **31**

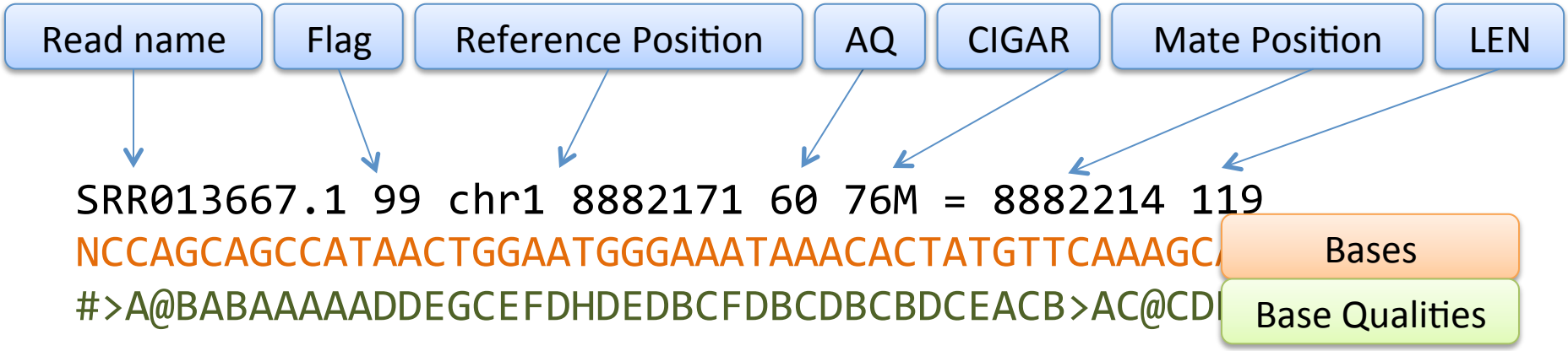
Base Quality: Phred Score Q_{phred}

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

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