Imperial College London

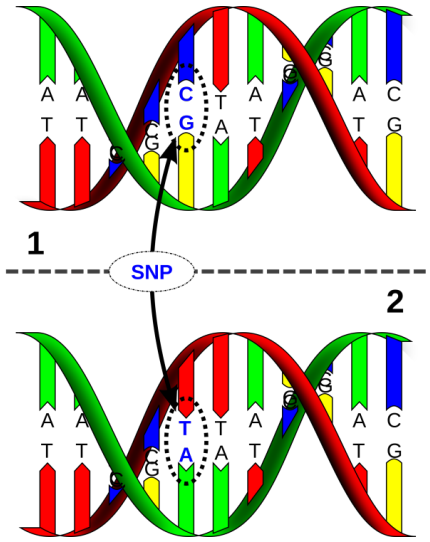
Royal Brompton & Harefield NHS
NHS Foundation Trust

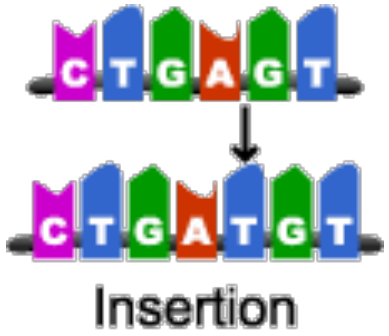
Variant calling in next-generation sequencing

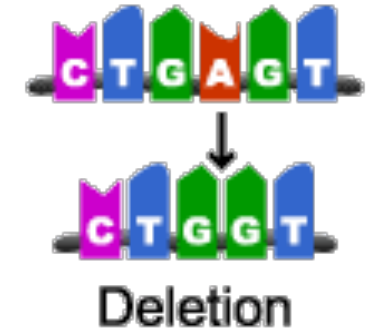
Dr Nicola Whiffin

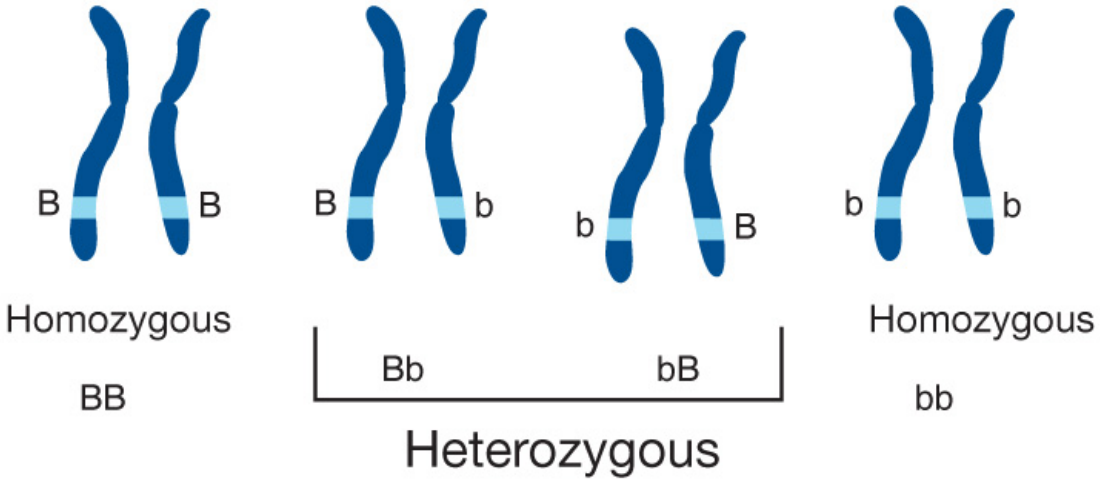
n.whiffin@imperial.ac.uk

Revision of terms

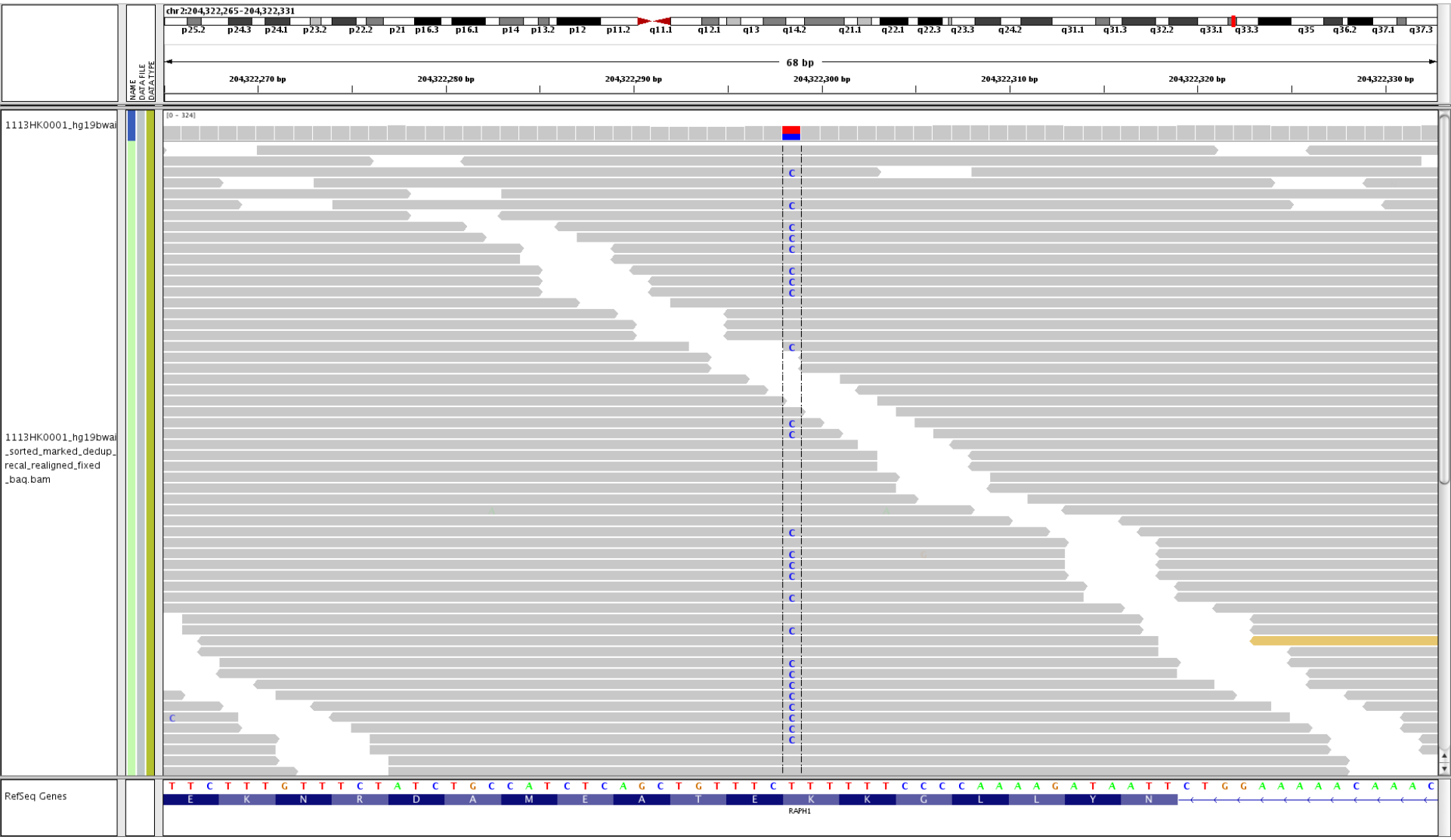




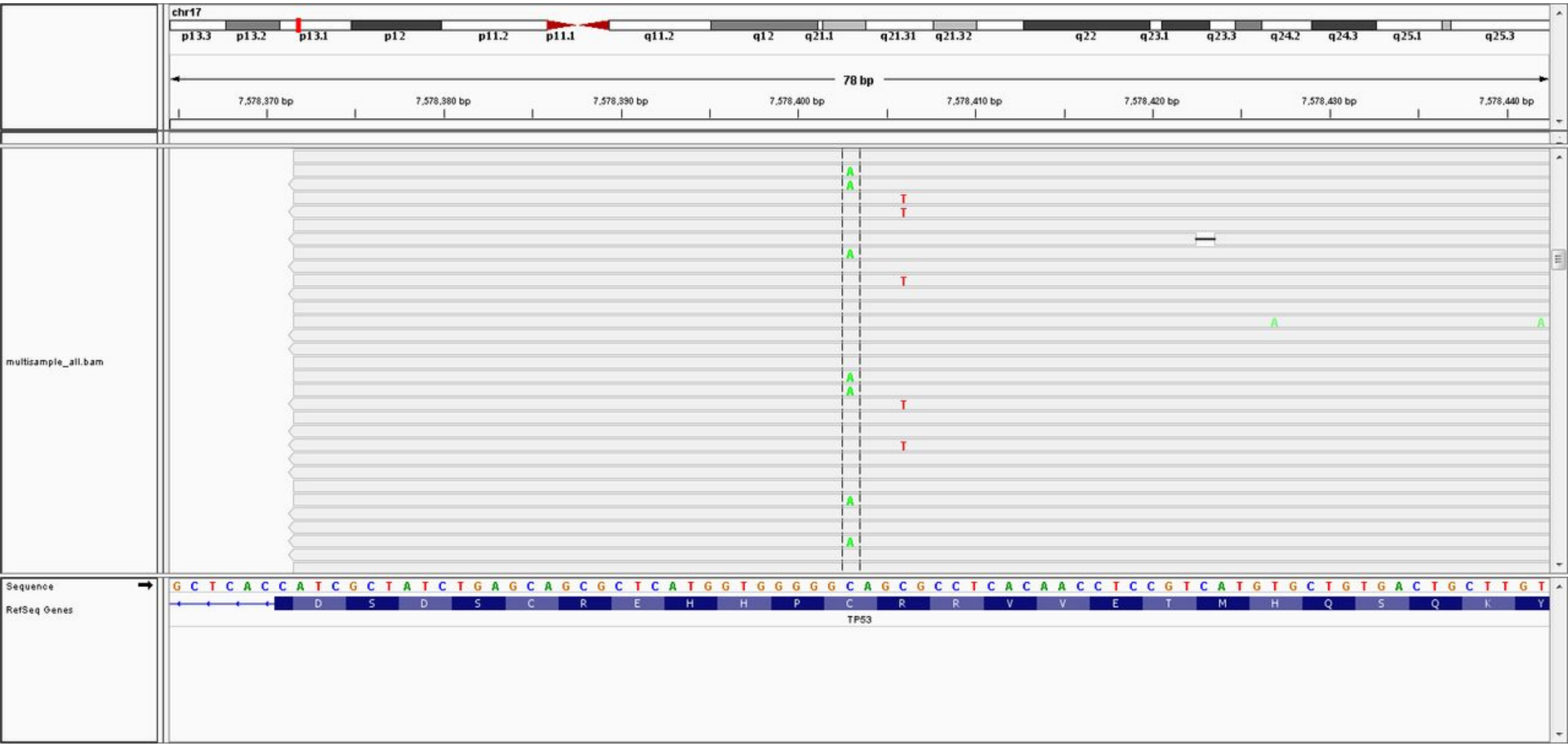




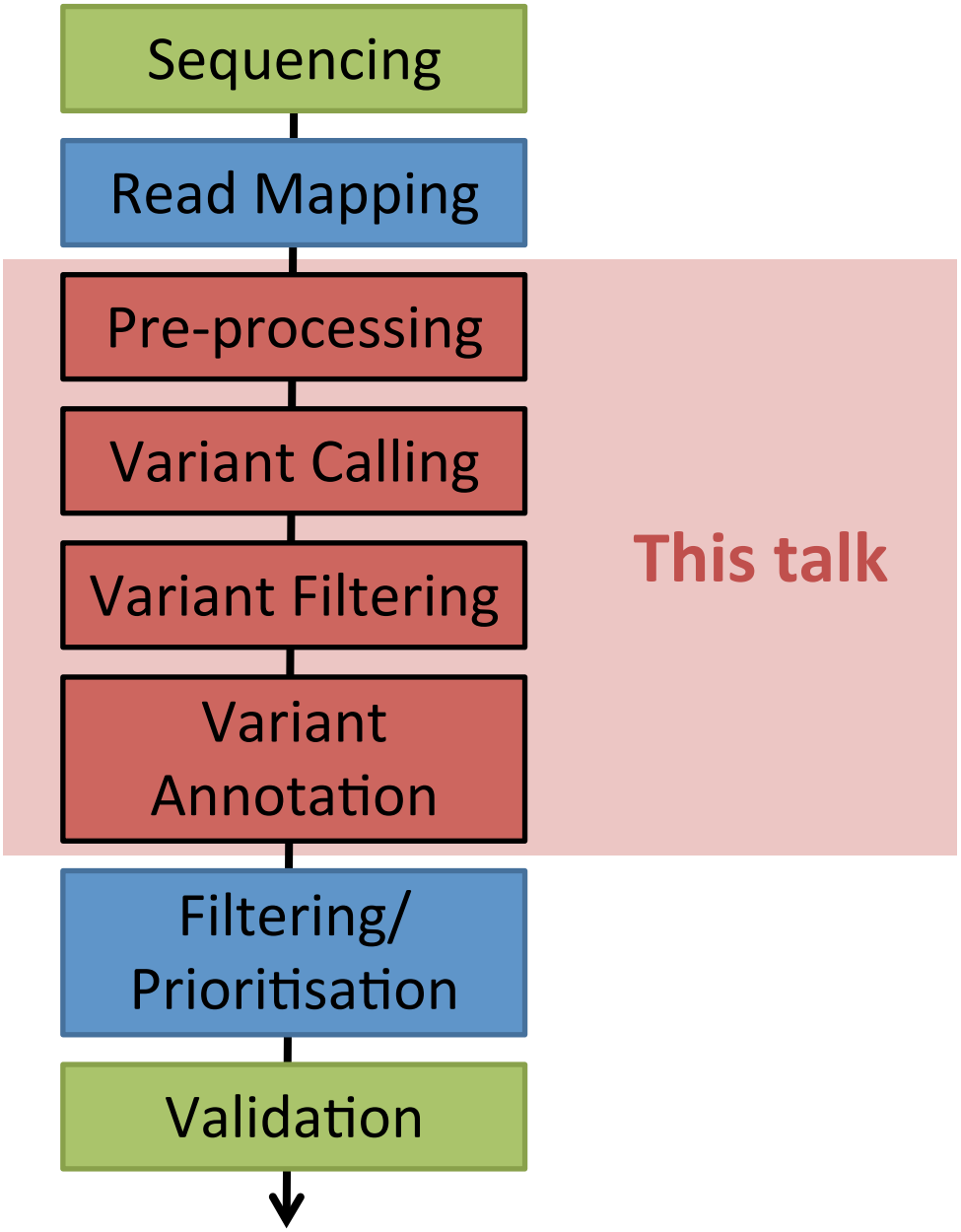
In principle it is very easy…



But the reality is somewhat different…



Workflow


Sequencing
Read Mapping
Pre-processing
Variant Calling
Variant Filtering
Variant Annotation
This talk
Filtering/ Prioritisation
Validation

Useful tools

SAMtools

- Utilities for manipulating SAM/BAM files

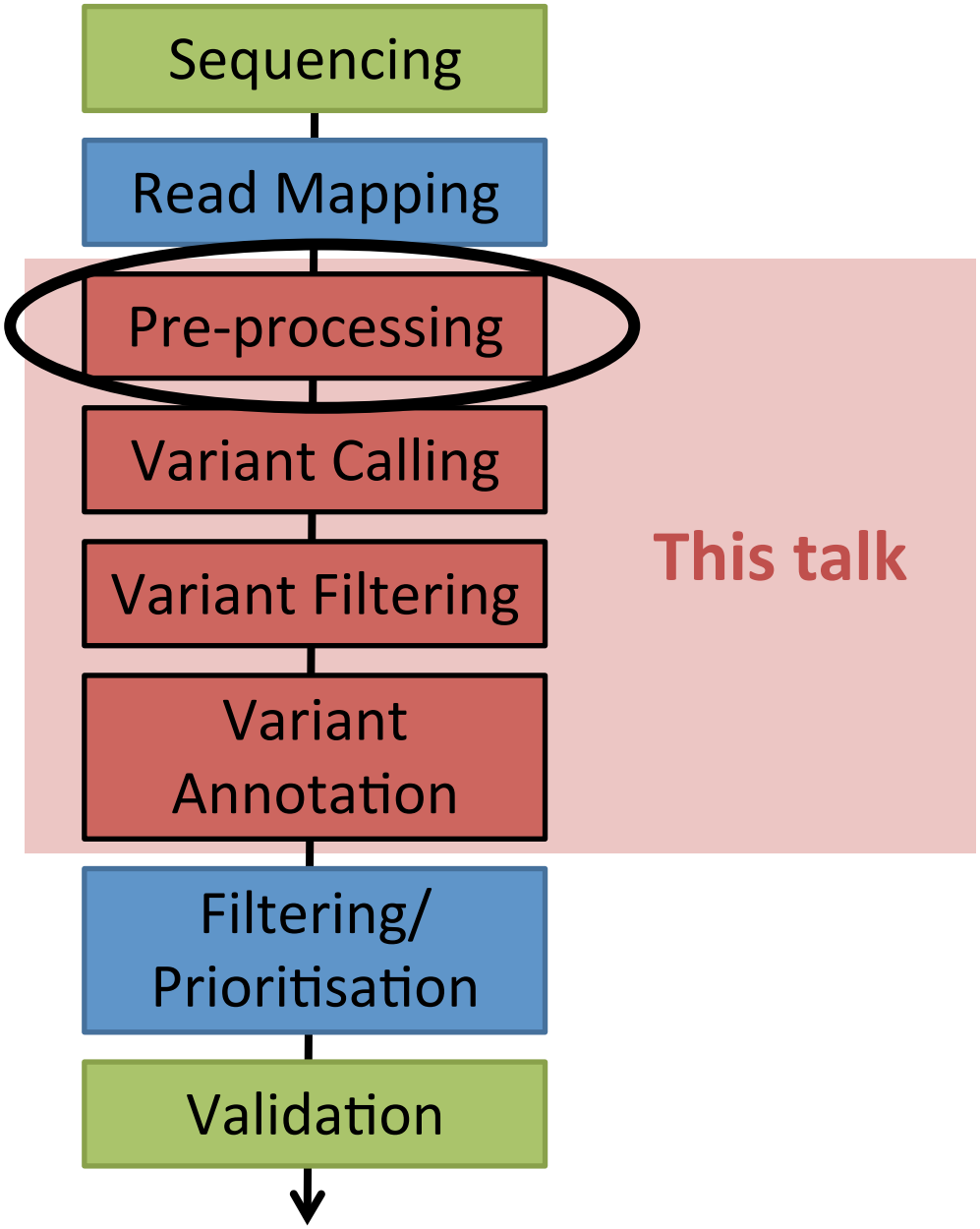
GATK

- Genome analysis toolkit – variety of tools for variant discovery, genotyping and quality

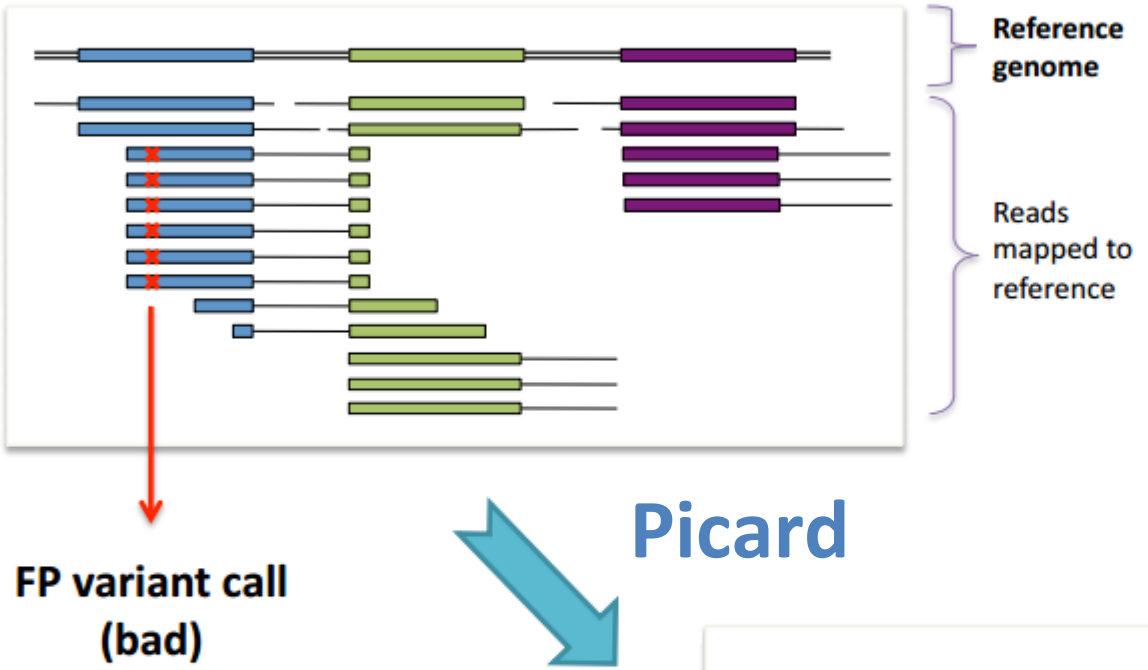
Picard

- Utilities for manipulating SAM/BAM files

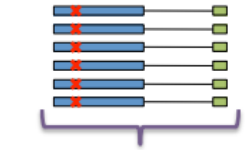
Workflow


Sequencing
Read Mapping
Pre-processing
Variant Calling
Variant Filtering
Variant Annotation
Filtering/ Prioritisation
Validation
This talk

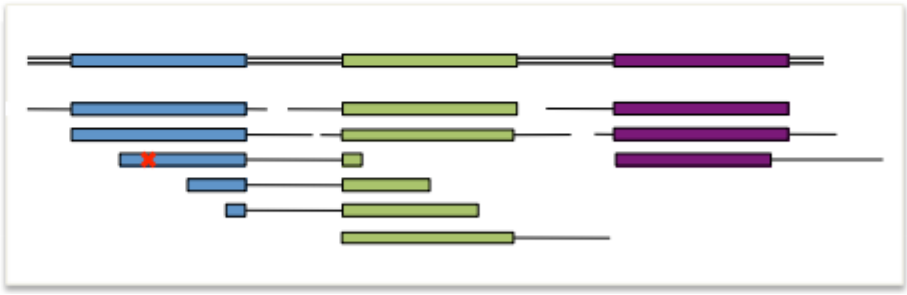
Pre-processing – mark duplicates



Duplicates have the same start position and CIGAR string – easy to tag



Generated as a consequence of PCR



https://www.broadinstitute.org/gatk/events/2038/GATKwh0-BP-1-Map_and_Dedup.pdf

Pre-processing – score recalibration

- Sequencing machines give quality scores to each base in isolation based on noise in base calling images - systematic bias
- Variant callers use quality scores to assign confidence to a call – need to be accurate
- Corrected using machine learning approach to model errors and adjust scores
- Takes into account position in read (more errors at ends) and surrounding base calls

GATK

Phred quality scores

Characterise the quality of DNA sequences

$$q = -10\log_{10}(p)$$

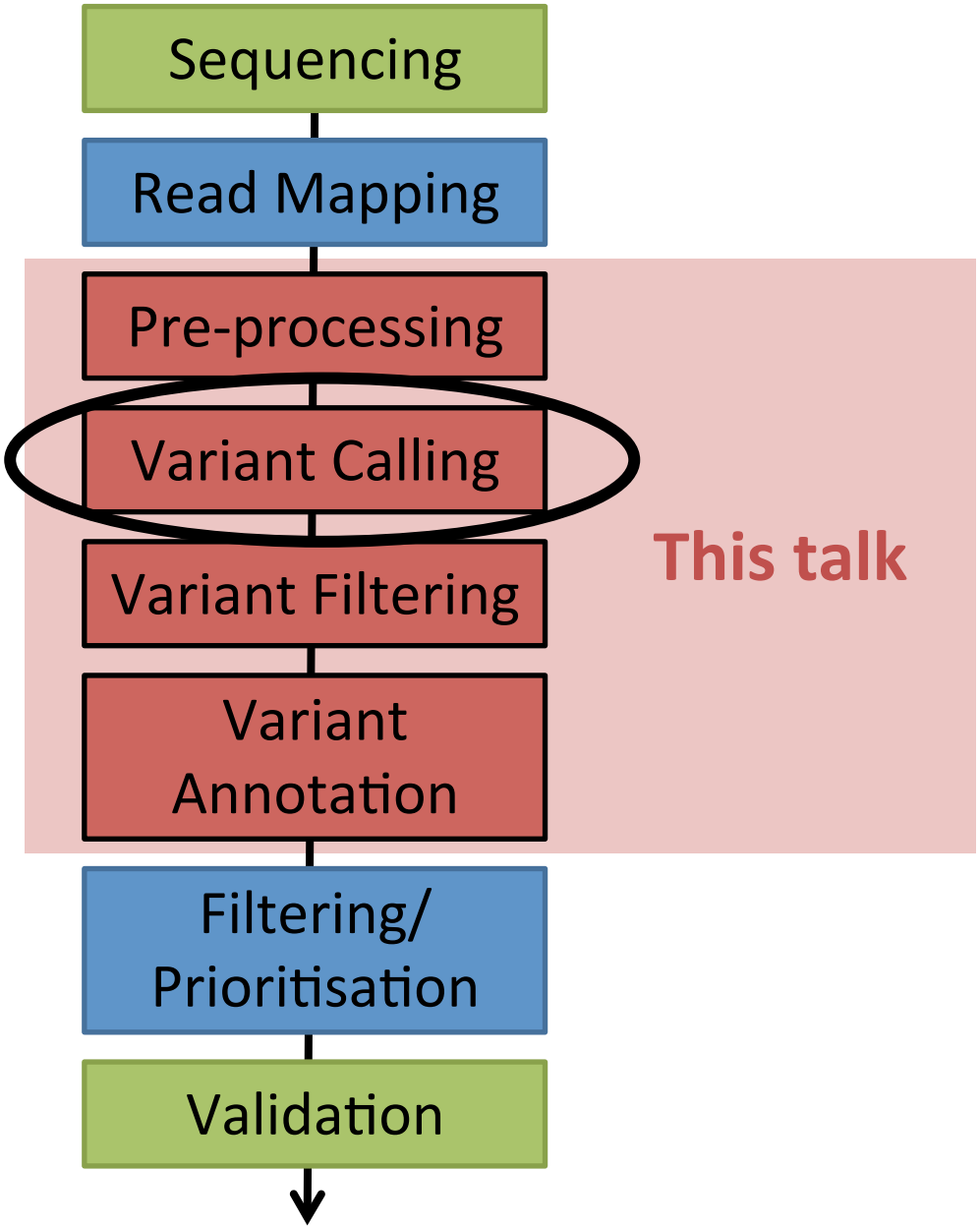
p = error probability for the base

Phred quality score	Probability	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Pre-processing - others

- Add or replace read groups (which read belongs to which sample) – **Picard**
- Sort reads by start position - **Picard**
- Local realignment around indels - mismatched reads at indel boundaries might look like evidence for SNPs [originally each read is individually mapped to the reference to cut computational cost] – **GATK**

Workflow


Sequencing
Read Mapping
Pre-processing
Variant Calling
Variant Filtering
Variant Annotation
Filtering/ Prioritisation
Validation
This talk

Too many tools!

- GATK unified genotyper
- GATK haplotype caller
- SAMtools
- SOAPsnp
- SOAPindel
- Pindel
- Dindel
- Scalpel
- Platypus
- varScan
- varDict
- Glftools
- Atlas2
- Mpileup
- MuTect
- BayesCall
- BreakDancer
- And many, many more...

Reasons for a mismatch

- True SNP

OR

- Error in library prep
- Base calling error – sequencing
  - Affected by coverage
- Mapping error (misalignment)
  - Hard in repetitive regions
  - Try local realignment
- Error in reference genome

General principles

- Identify sites that differ from the reference
- Estimate likelihood this is variant or sequencing error taking into account:
  - Base quality score
  - Proximity to indel
  - Repetitive regions (e.g. Homopolymers)
  - Mapping qualities of supporting reads
  - Read length
  - Position in read
  - Paired reads
  - Coverage
  - Strand bias
- Local *de novo* realignment in 'active' regions
- Output quality score for variant call

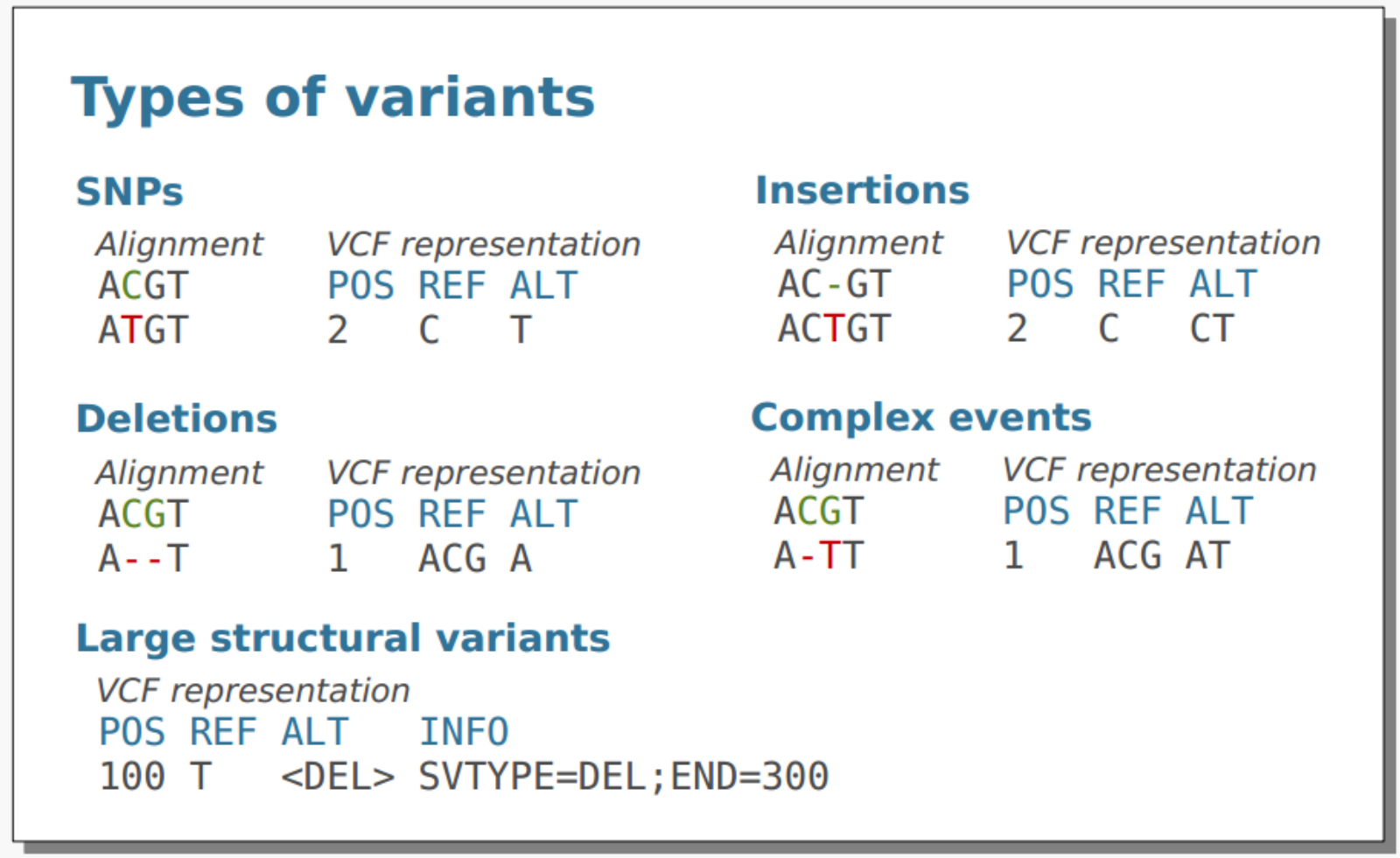
General principles

- Early methods
  - Simply count numbers of reference and alternate reads – simple cut off to identify variants
- Bayesian methods
  - Take into account counts as well as base and mapping qualities
  - Posterior probability of each possible genotype is output – used to calculate quality score

VCF file format

```
##fileformat=VCFv4.1
##FILTER=<ID=FSFilter,Description="FS > 60.0">
##FILTER=<ID=InDel,Description="Overlaps a user-input mask">
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=MQFilter,Description="MQ < 40.0">
##FILTER=<ID=MQRankSumFilter,Description="MQRankSum < -12.5">
##FILTER=<ID=QDFilter,Description="QD < 2.0">
##FILTER=<ID=ReadPosFilter,Description="ReadPosRankSum < -8.0">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##INFO=<ID=HRun,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of Variant Allele In Either Direction">
##INFO=<ID=HW,Number=1,Type=Float,Description="Phred-scaled p-value for Hardy-Weinberg violation">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), fo
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), f
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=OND,Number=1,Type=Float,Description="Overall non-diploid ratio (alleles/(alleles+non-alleles))">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##reference=file:///data/Store/reference/human/UCSC_hg19/allchrom.Chr1ToChrM.validated.fa
##source=SelectVariants
#CHROM→POS→ID→REF→ALT→QUAL→FILTER→INFO→FORMAT→14G000197
chr1→2985885→rs7413494→C→G→48.77→PASS→MQ=60.00;MQ0=0;MQRankSum=0.307;QD=2.57;ReadPosRankSum=-0.307→GT:AD:GQ:PL→0/1:13,6:77:77,0,308
chr1→3301721→rs2282198→C→T→2885.77→PASS→MQ=60.00;MQ0=0;MQRankSum=-1.100;QD=9.85;ReadPosRankSum=2.245→GT:AD:GQ:PL→0/1:148,145:99:2914,0,2
chr1→3303446→rs2245703→T→C→4454.77→PASS→MQ=60.00;MQ0=0;MQRankSum=0.774;QD=11.88;ReadPosRankSum=0.009→GT:AD:GQ:PL→0/1:179,196:99:4483,0,3
chr1→3328358→rs870124→T→C→7516.77→PASS→MQ=60.00;MQ0=0;QD=29.13→GT:AD:GQ:PL→1/1:0,258:99:7545,772,0
chr1→3334598→rs188132529→C→T→766.77→PASS→MQ=60.00;MQ0=0;MQRankSum=-0.998;QD=12.17;ReadPosRankSum=0.378→GT:AD:GQ:PL→0/1:30,33:99:795,0,645
chr1→3341540→rs2483236→C→T→1168.77→PASS→MQ=60.00;MQ0=0;MQRankSum=0.798;QD=12.30;ReadPosRankSum=-0.940→GT:AD:GQ:PL→0/1:42,53:99:1197,0,896
chr1→3341639→.→CTTTTTTTT→C,CTTTTTTTTT→0.01→LowQual;QDFilter→MQ=60.00;MQ0=0;MQRankSum=0.067;QD=0.00;ReadPosRankSum=0.762→GT:AD:GQ:PL→0/1
```

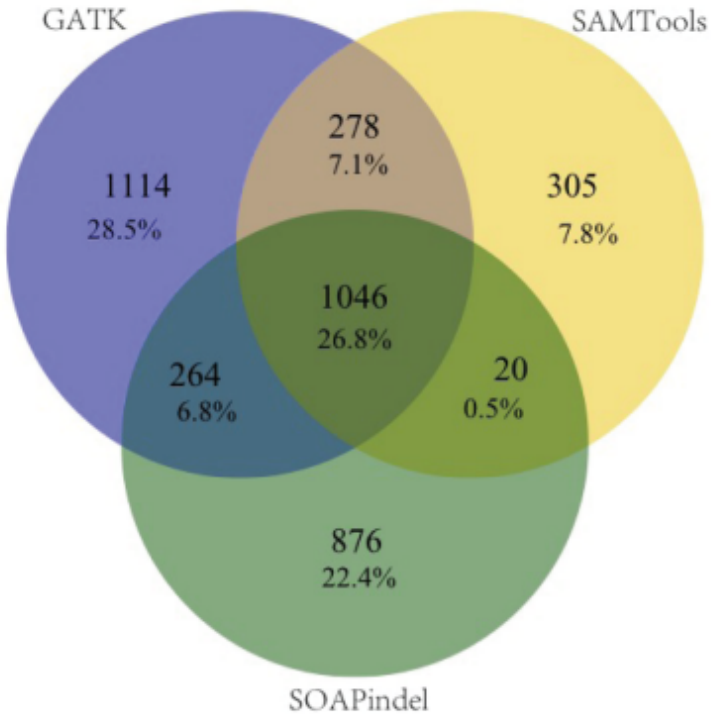
Types of variants



http://vcftools.sourceforge.net/VCF-poster.pdf

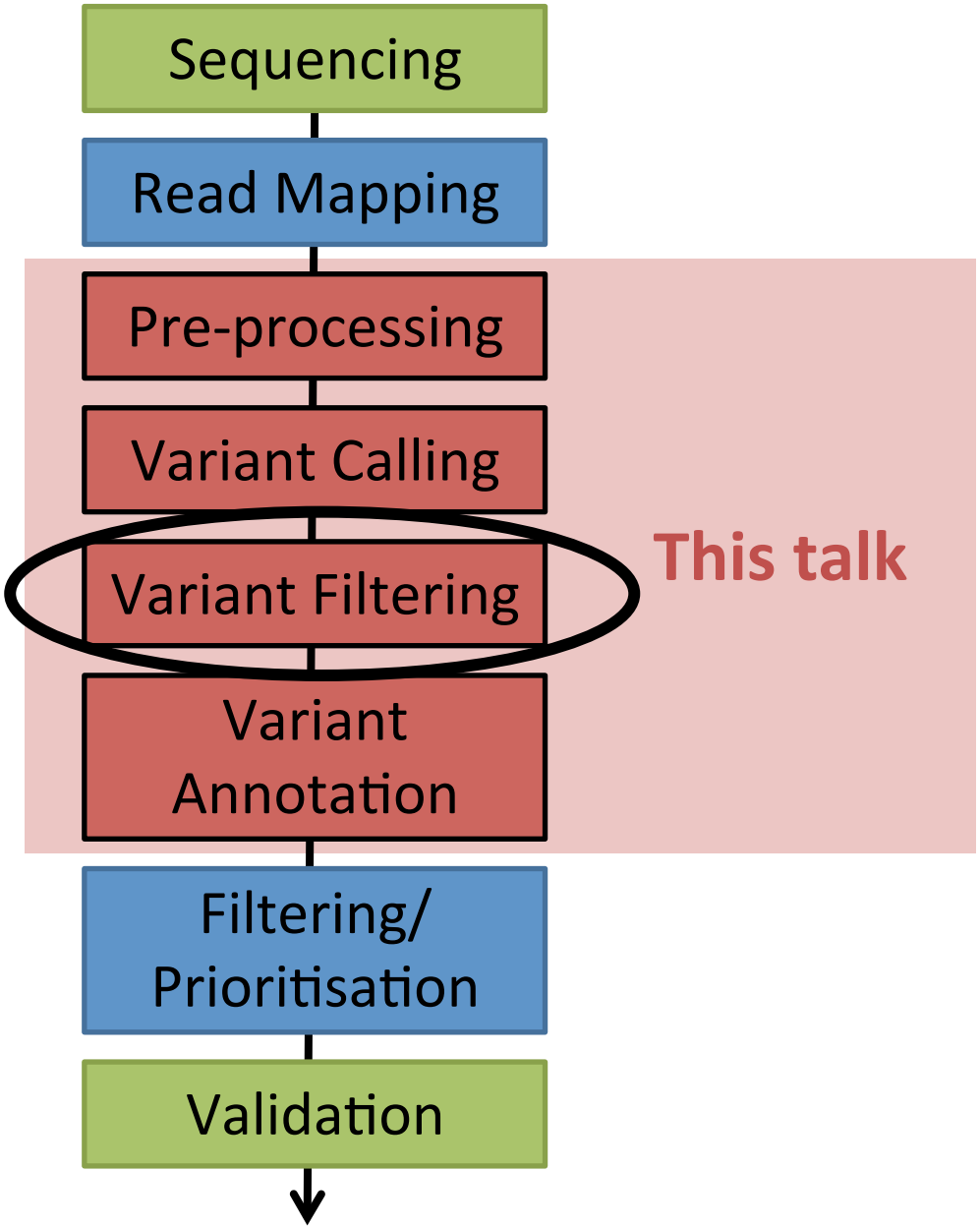
Indels are a different problem

- Lots of FP SNPs near true indels
  - Mismatches penalised less than gaps
  - Local *de novo* realignment
- Little concordance between popular callers
  - Lag behind SNV callers



O'Rawe, Jiang *et al*. Genome Medicine. 2013

Workflow


Sequencing
Read Mapping
Pre-processing
Variant Calling
Variant Filtering
Variant Annotation
This talk
Filtering/ Prioritisation
Validation

Need for good filters

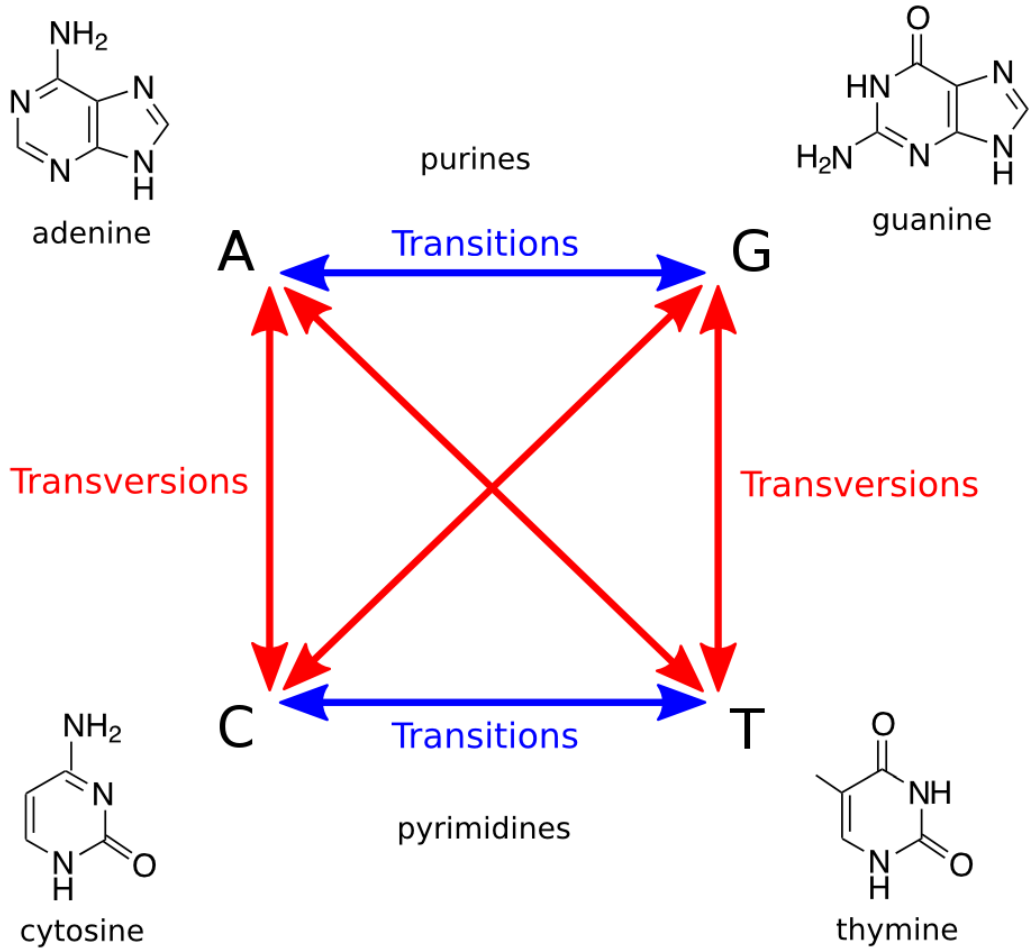
- NGS technologies and variant callers are far from perfect!
  - False discovery rate (FDR) ~0.2-0.6% (7,000-17,000 errors per genome (Complete Genomics))
- Errors occur as rare/novel variants
  - Expect disease causing variants to be rare/novel too
  - Removing common variants increases proportion of errors
- Callers often designed with high sensitivity
- Try to remove sequencing errors but retain large proportion of true variants

Example filters

- Repetitive regions – sequencing and mapping difficult
- Strand bias – low frequency of reads with errors
- Coverage
- Quality scores
- Proximity to SNV/Indel

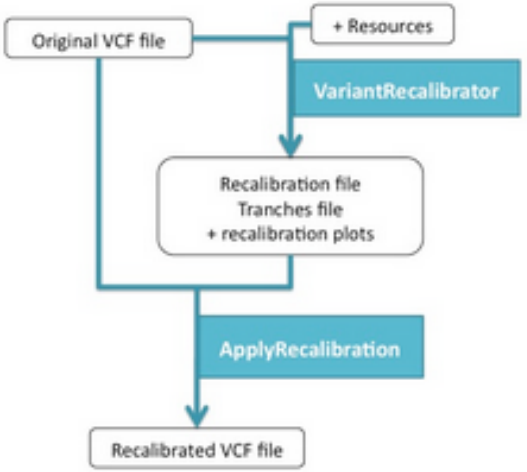
Can be applied through calling algorithm (filters may vary) or subsequent VCF file filtering

Ti/Tv ratio



- Transition is more frequent than transversion
  - Ti/Tv ~1.5 for whole genome
  - Ti/Tv ~2.0 for exome
  - Ti/Tv = 2/4 = 0.5 for random, uniform, sequencing error

GATK soft filtering - VQSR



- Variant quality score recalibration
  - Machine learning to assign well calibrated probabilities to each variant using a high quality set of know variants as training and truth resources – *VariantRecalibrator*
  - Filtering based on this new quality score - *ApplyRecalibration*
- Requires large, high-quality set of variants from organism of interest
- Needs a large number of samples run at one time to learn profiles of good and bad variants

GATK hard filtering best practice

Quality/Depth

FisherStrand – Phred scaled *P*-value to detect strand bias

Root mean square of mapping qualities of all reads

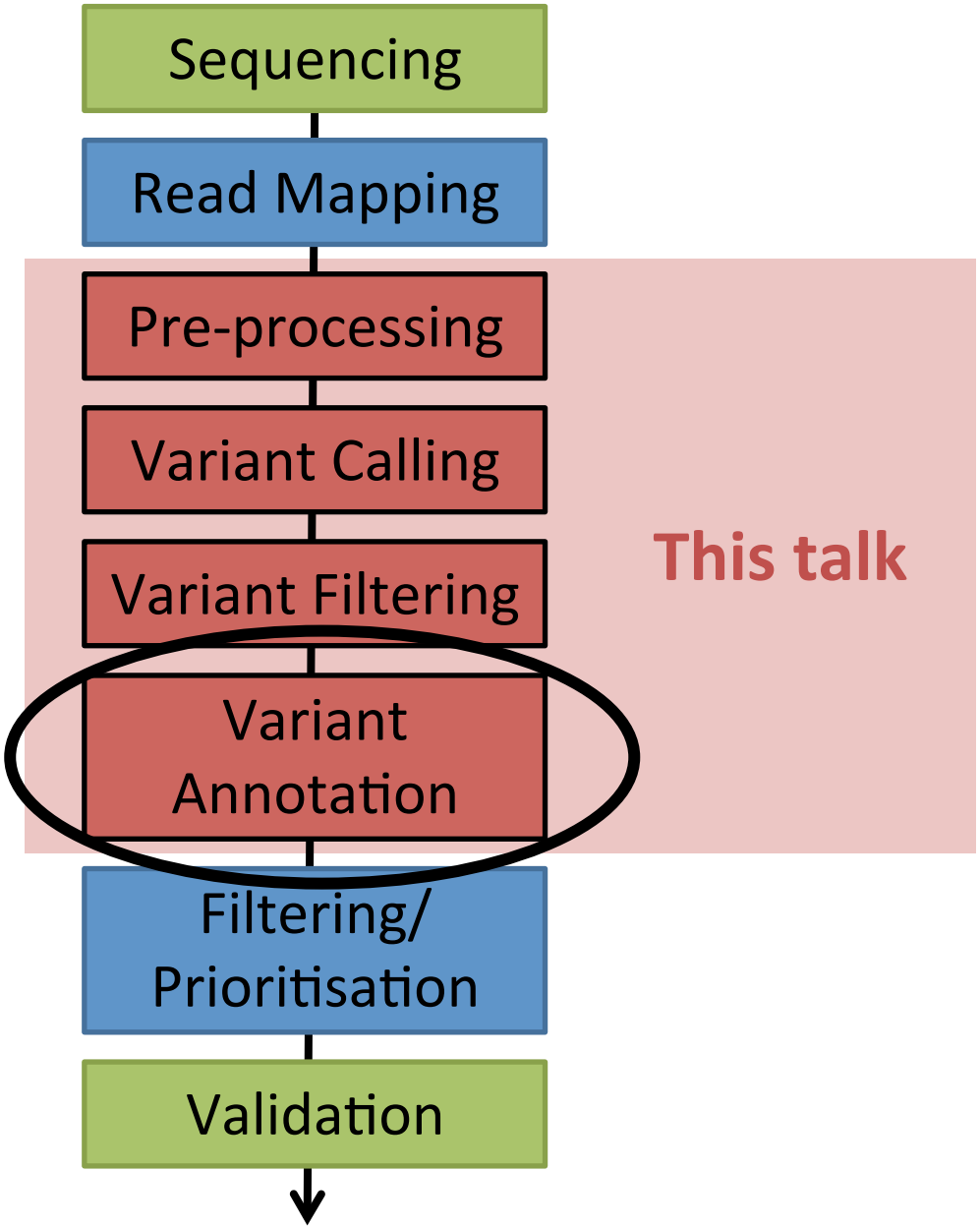
```
java -jar GenomeAnalysisTK.jar \
    -T VariantFiltration \
    -R reference.fa \
    -V raw_snps.vcf \
    --filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || HaplotypeScore > 13.0 || MappingQualityRankSum < -12.5
|| ReadPosRankSum < -8.0" \
    --filterName "my_snp_filter" \
    -o filtered_snps.vcf
```

Tests if alternative allele only seen at ends of reads

Consistency of site with only 2 haplotypes – assumes diploid

Mapping qualities of read with alternate vs reference alleles

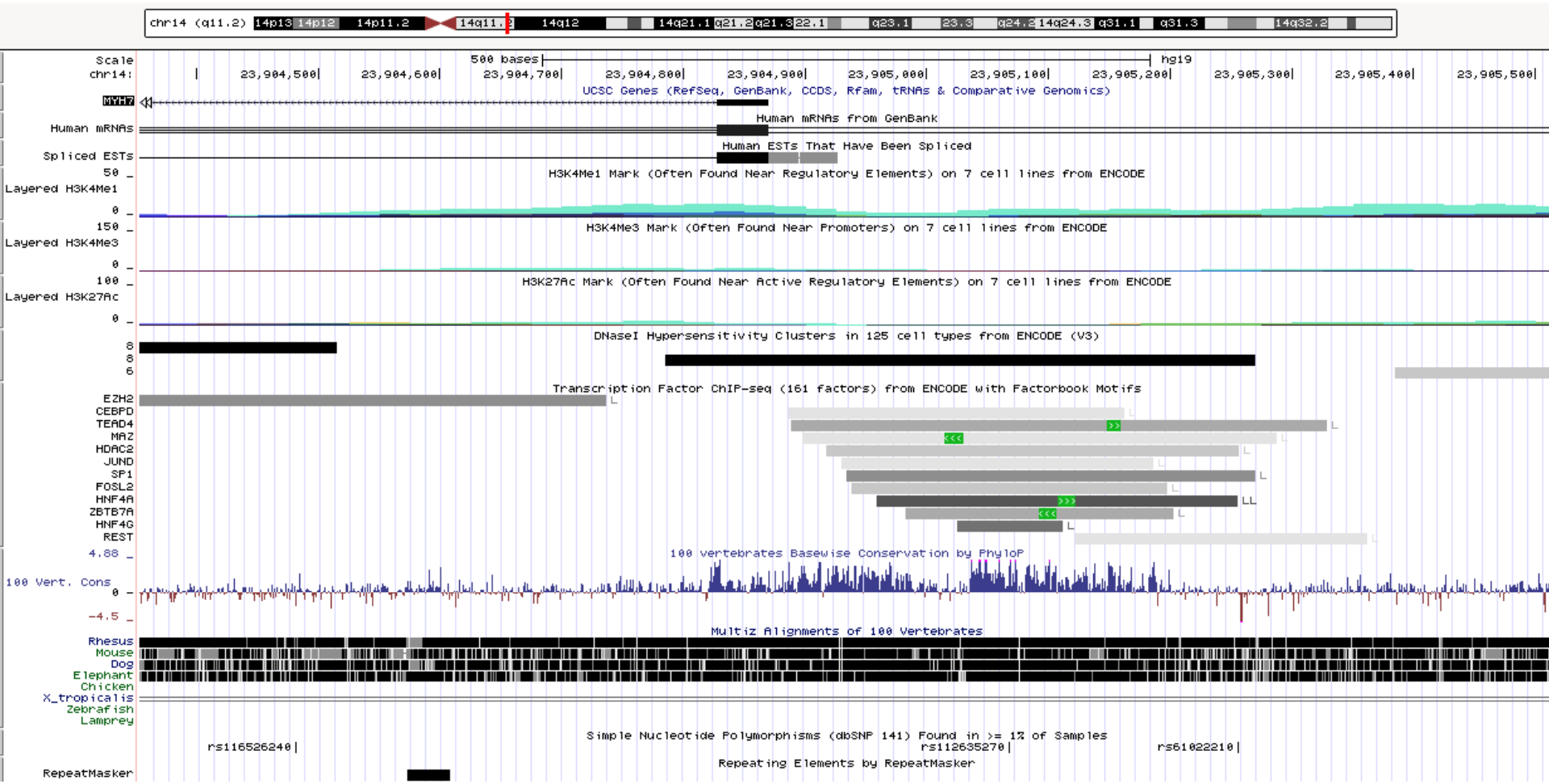
Workflow


Sequencing
Read Mapping
Pre-processing
Variant Calling
Variant Filtering
Variant Annotation
Filtering/ Prioritisation
Validation
This talk

Functional annotation

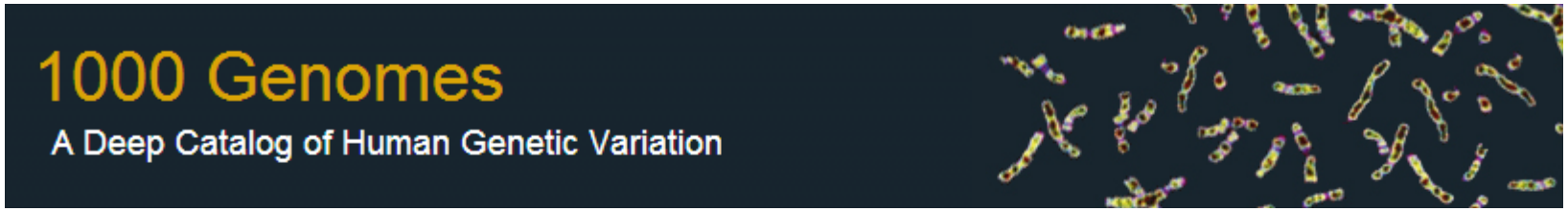
- What/where in the genome does the variant map?
  - Protein coding?
  - Synonymous or non-synonymous?
  - Frame-shift or frame-preserving?
  - Other functional regions (e.g. Splice sites, promoter/ enhancer regions, ncRNAs)
  - Annovar/VEP
- Is the position conserved?
  - PhastCons
  - GERP
  - PhyloP

UCSC genome browser



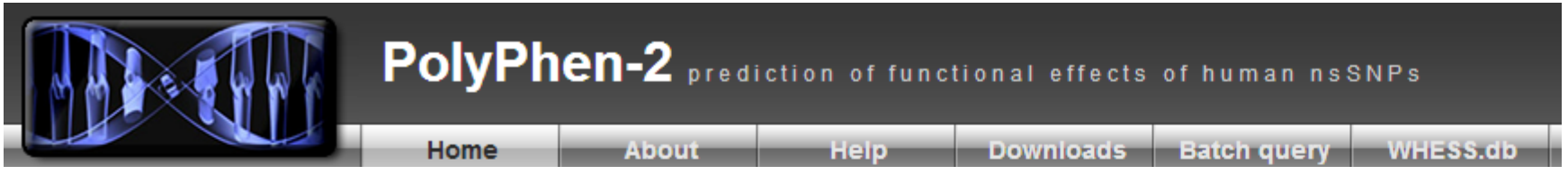
Annotation - frequency

- Frequency in public variant resources
  - EVS (exome variant server)
  - dbSNP
  - 1000 genomes
- Can be inaccurate and incomplete



Annotation - deleteriousness

- Various tools attempt to assign scores of deleteriousness to variants
  - SIFT
  - Polyphen2
  - CADD
  - CONDEL
  - MutationTaster
  - MutationAssessor
  - Grantham
  - SuSPect
- Mainly only for protein coding regions
- Look for a consensus between tools



Summary

- Pre-processing of BAM files is necessary before variant calling
- Variant callers estimate the probability a difference is a variant rather than a sequencing error – lots of tools
- Indels are more difficult to call than SNVs
- Algorithms have high sensitivity so filtering is needed to remove any remaining errors
- Annotation allows us to prioritise variants that may have a role in a trait/disease

Any Questions?