Calling Genomic Variants from DNAseq data

Oxford Biomedical Data Science Training Programme

Genomic DNA Sequencing

- Whole genome sequencing
- Exome sequencing (2% of genome)
- Targeted panel (~10-100 genes)

- Variant types
  - Single nucleotide variants
  - Small insertions and deletion (indels)
  - Copy number variations
  - Structural variations

Germline vs Somatic Variants

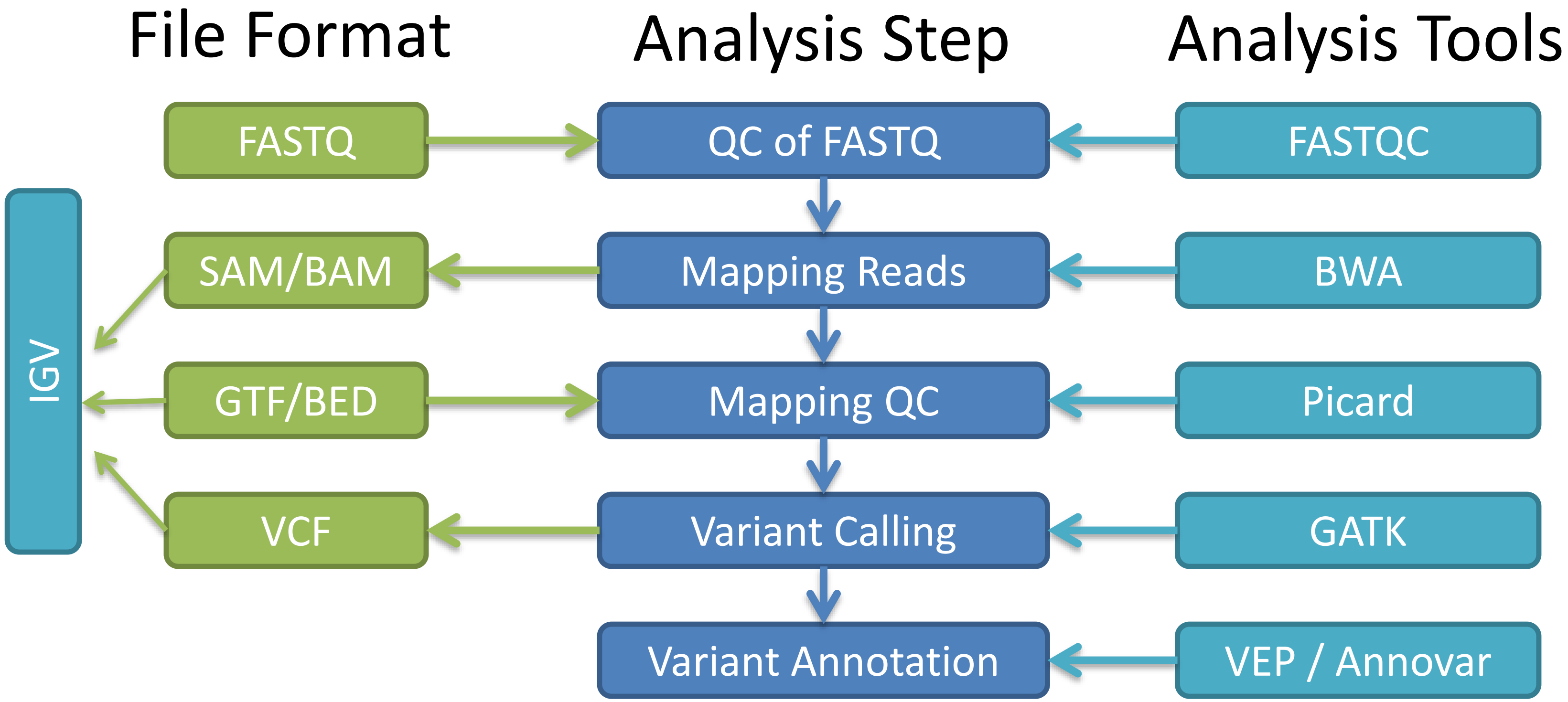
Germline

- Rare disease
- *De novo* mutations
- Trio or family
- Assume diploid

Somatic

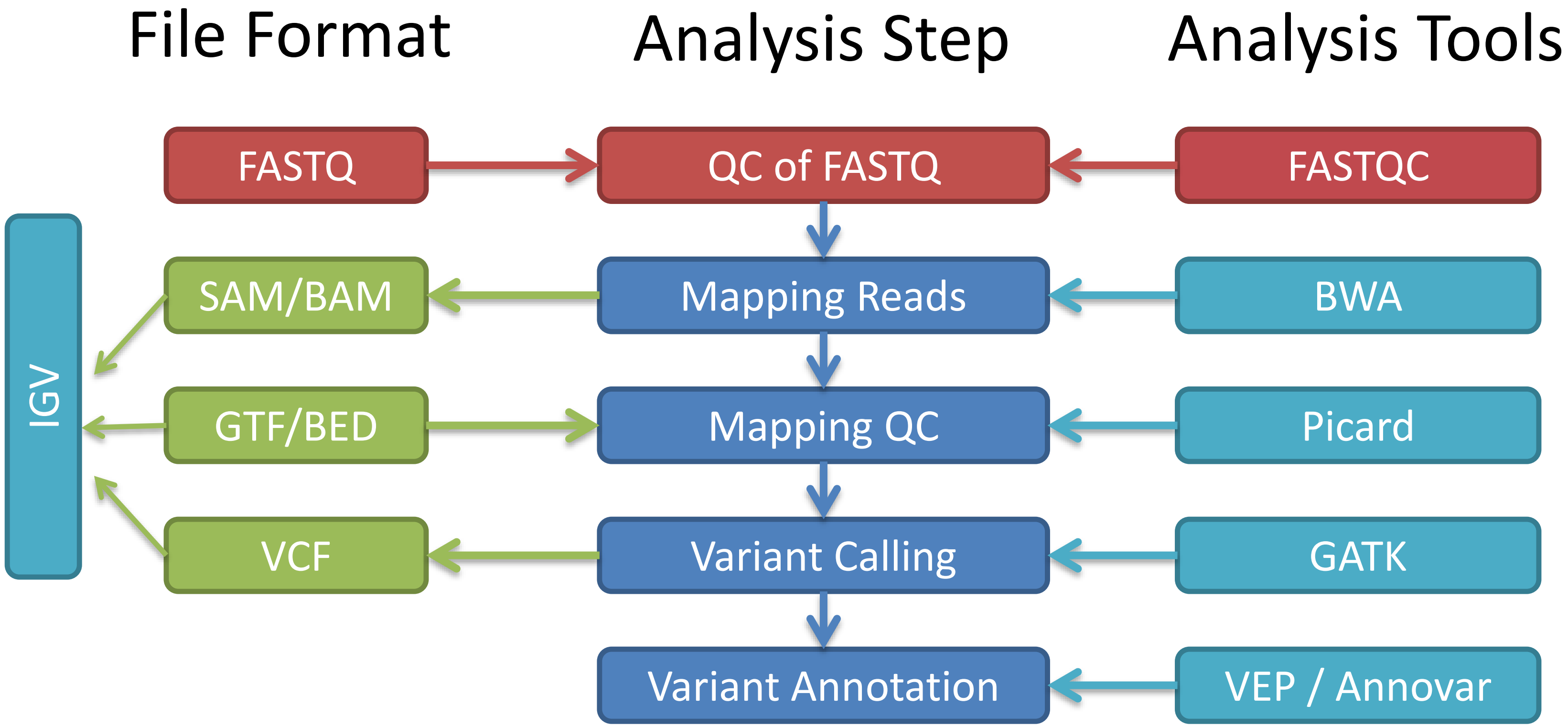
- Cancer
- Tumour vs germline
- Tumour purity
- Tumour heterogeneity
- Cannot assume diploid

SNV / Indel Calling Workflow



File Format	Analysis Step	Analysis Tools
FASTQ	QC of FASTQ	FASTQC
SAM/BAM	Mapping Reads	BWA
GTF/BED	Mapping QC	Picard
VCF	Variant Calling	GATK
	Variant Annotation	VEP / Annovar

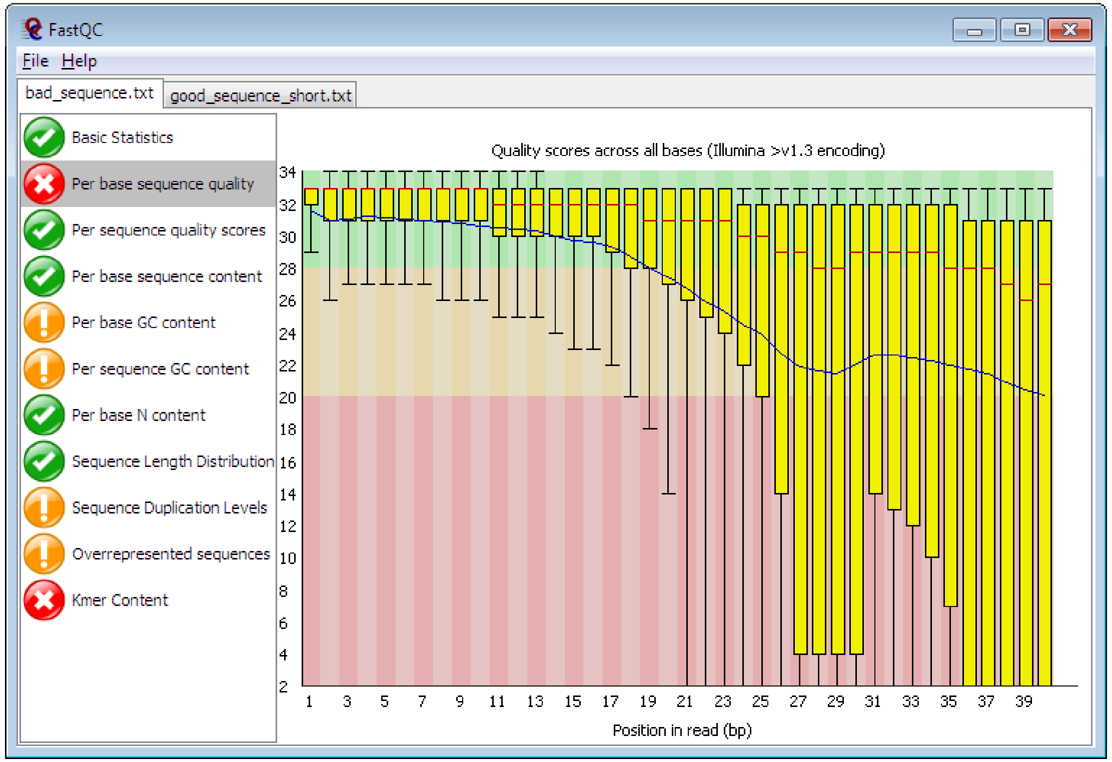
Step 1 - QC of Raw Sequencing Data



Read Quality Control

- FastQC
- Traffic light overview
- Graphical summaries
- HTML report

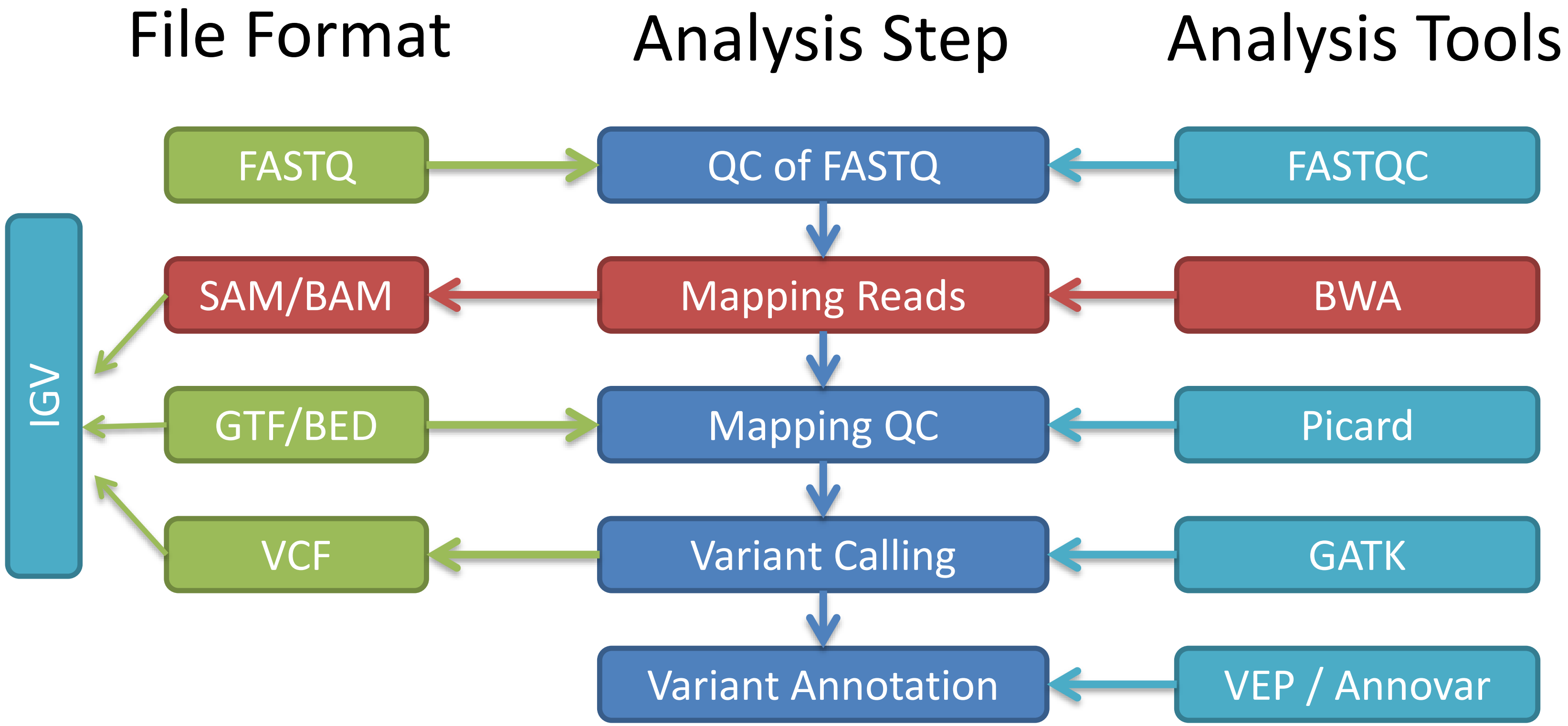
Babraham Bioinformatics



- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per base GC content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✗ Kmer Content

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Step 2 - Mapping Reads to a Reference Genome



File Format

Analysis Step

Analysis Tools

FASTQ

QC of FASTQ

FASTQC

SAM/BAM

Mapping Reads

BWA

GTF/BED

Mapping QC

Picard

VCF

Variant Calling

GATK

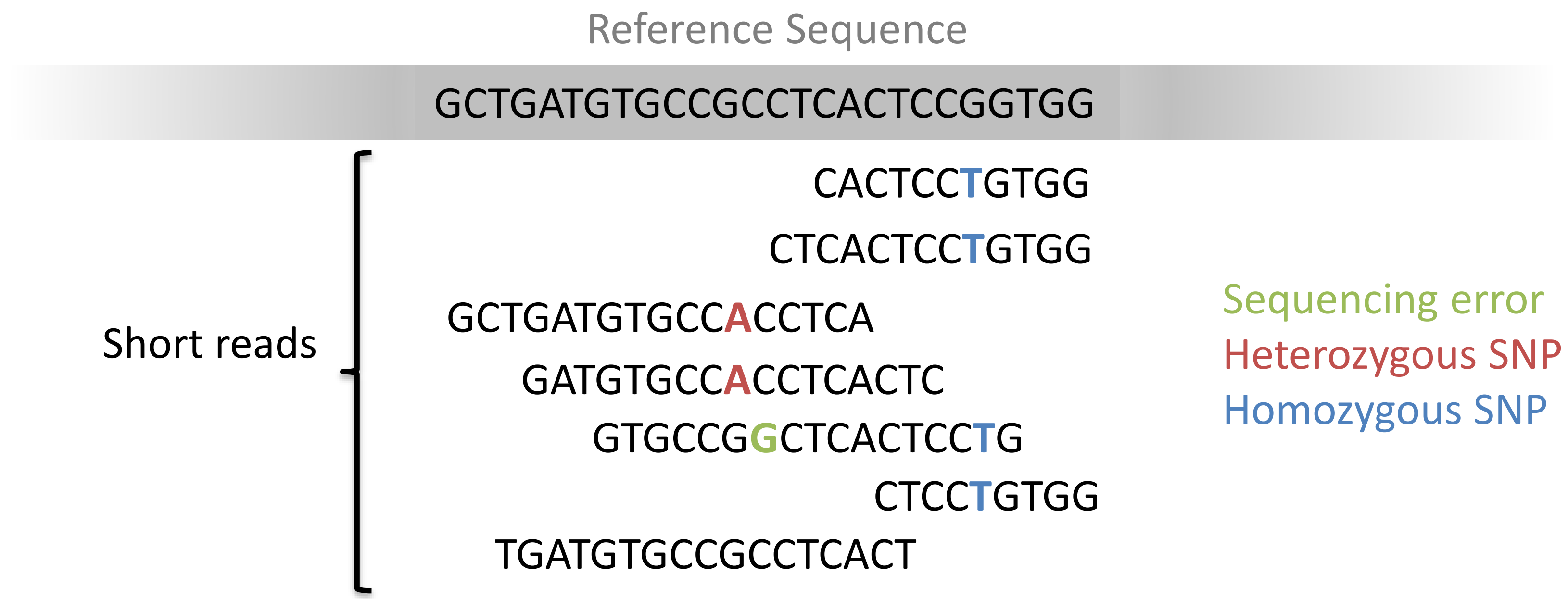
Variant Annotation

VEP / Annovar

IGV

Mapping Reads

- Find the position(s) in the reference genome where each short read sequence aligns with the fewest mismatches
- Must be fast (millions of short reads)
- Must allow small differences (sequencing errors or polymorphisms)
- String matching problem



Reference Sequence

GCTGATGTGCCGCCTCACTCCGGTGG

Short reads

CACTCC**T**GTGG

CTCACTCC**T**GTGG

GCTGATGTGCC**A**CCTCA

GATGTGCC**A**CCTCACTC

GTGCCG**G**CTCACTCC**T**G

CTCC**T**GTGG

TGATGTGCCGCCTCACT

Sequencing error

Heterozygous SNP

Homozygous SNP

Short Read Mapping Tools

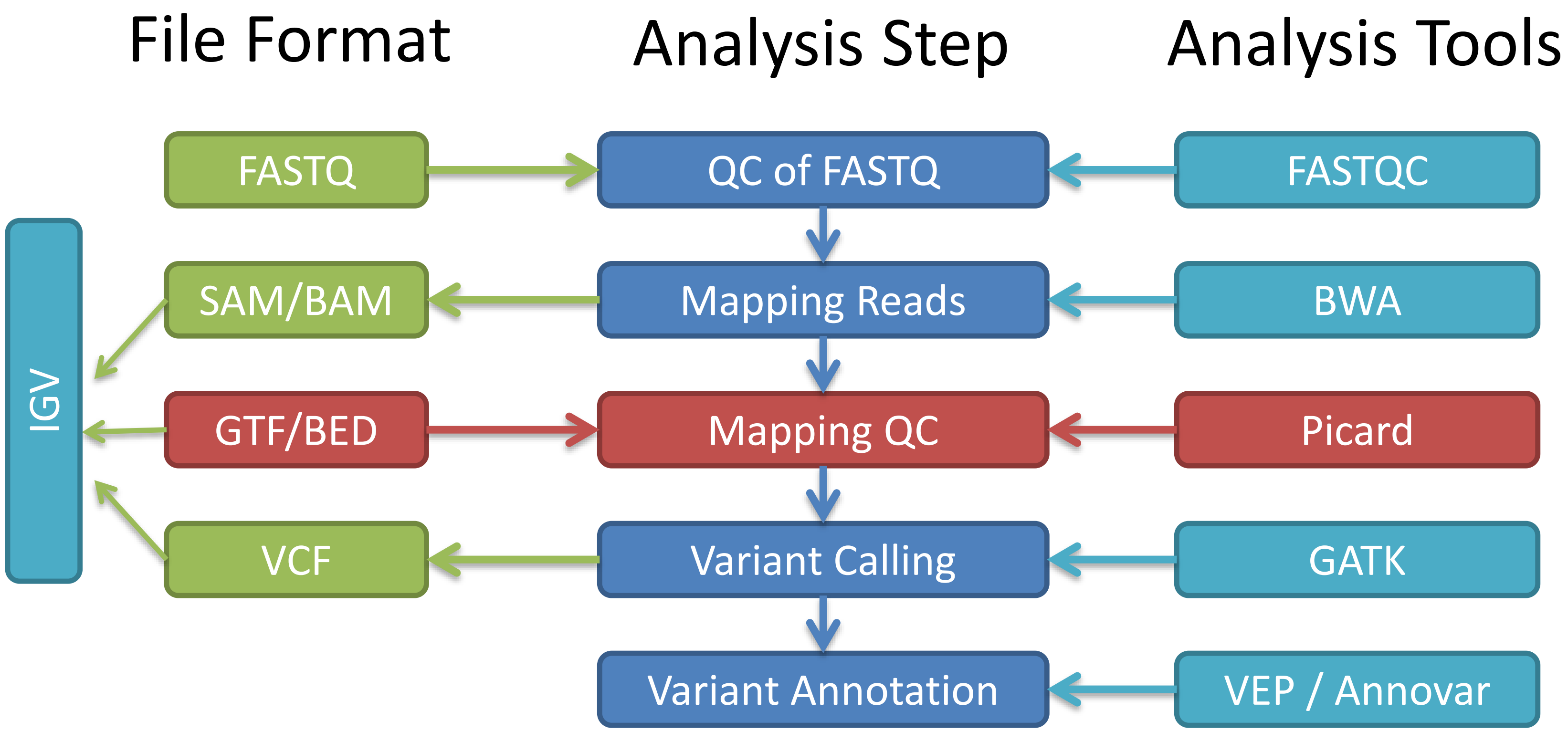
- General purpose alignment tools: BLAST, BLAT
- First short read specific tools:
  - Eland, MAQ – use hash tables
- Second generation tools:
  - Bowtie, BWA – Burrows-Wheeler Transform
- Third generation tools:
  - SOAP3 – uses GPU processing
- Trade off between sensitivity, specificity and processing time
- For DNAseq we want accurate SNP/indel detection so specificity is key

Burrows-Wheeler Aligner

- Recommended by the Broad Institute best practice guidelines
- Aligns short sequences (< 400nt) against long reference genome
- Fast (if not too many errors)
- Gapped alignment (enables short indel calling)
- Soft clipping (ends of reads do not have to align)
- Takes FASTQ as input
  - Requires Sanger quality score format
- Produces SAM as output
- Default parameters optimised for mammalian DNA sequencing
- New algorithm BWA-MEM now recommended for read length > 70bp

http://bio-bwa.sourceforge.net/bwa.shtml

Step 3 - QC of Mapped Reads



Picard Tools

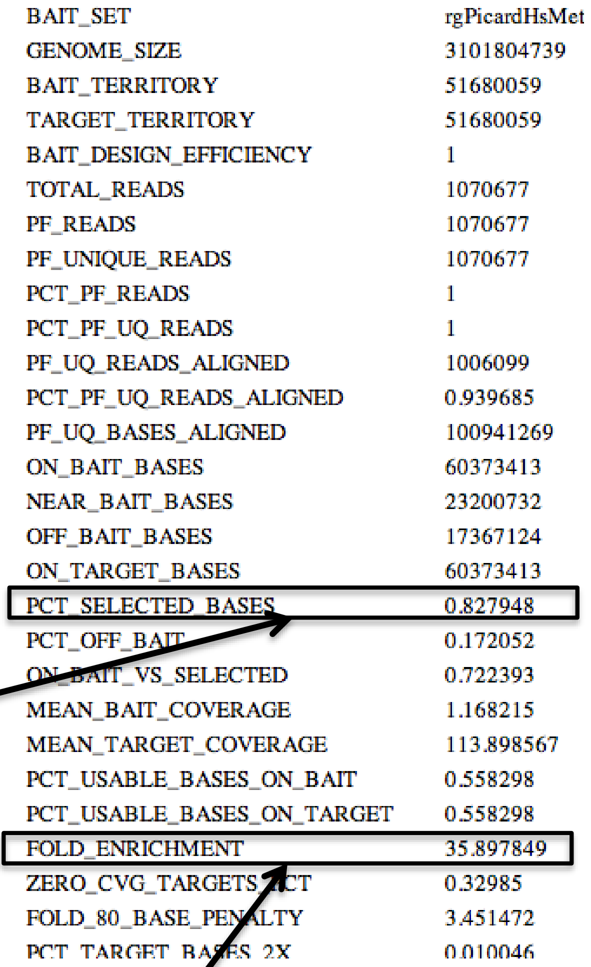
- SAM/BAM/CRAM & VCF processing
- Overlapping functionality with Samtools/Pysam
- Written in Java
- Broad Institute
- Many BAM Quality control tools
  - CollectAlignmentSummaryMetrics
  - CollectBaseDistributionByCycle
  - CollectGcBiasMetrics
  - CollectHsMetrics

CollectMultipleMetrics

https://broadinstitute.github.io/picard/command-line-overview.html

Target Coverage Metrics

- ON_BAIT_BASES: The number of PF aligned bases that mapped to a baited region of the genome.
- NEAR_BAIT_BASES: The number of PF aligned bases that mapped to within a fixed interval of a baited region, but not on a baited region.
- OFF_BAIT_BASES: The number of PF aligned bases that mapped to neither on or near a bait.
- ON_TARGET_BASES: The number of PF aligned bases that mapped to a targeted region of the genome.
- MEAN_BAIT_COVERAGE: The mean coverage of all baits in the experiment.
- MEAN_TARGET_COVERAGE: The mean coverage of targets that received at least coverage depth = 2 at one base.
- FOLD_ENRICHMENT: The fold by which the baited region has been amplified above genomic background.
- ZERO_CVG_TARGETS_PCT: The number of targets that did not reach coverage=2 over any base.
- PCT_TARGET_BASES_20X: The percentage of ALL target bases achieving 20X or greater coverage.



BAIT_SET	rgPicardHsMet
GENOME_SIZE	3101804739
BAIT_TERRITORY	51680059
TARGET_TERRITORY	51680059
BAIT_DESIGN_EFFICIENCY	1
TOTAL_READS	1070677
PF_READS	1070677
PF_UNIQUE_READS	1070677
PCT_PF_READS	1
PCT_PF_UQ_READS	1
PF_UQ_READS_ALIGNED	1006099
PCT_PF_UQ_READS_ALIGNED	0.939685
PF_UQ_BASES_ALIGNED	100941269
ON_BAIT_BASES	60373413
NEAR_BAIT_BASES	23200732
OFF_BAIT_BASES	17367124
ON_TARGET_BASES	60373413
PCT_SELECTED_BASES	0.827948
PCT_OFF_BAIT	0.172052
ON_BAIT_VS_SELECTED	0.722393
MEAN_BAIT_COVERAGE	1.168215
MEAN_TARGET_COVERAGE	113.898567
PCT_USABLE_BASES_ON_BAIT	0.558298
PCT_USABLE_BASES_ON_TARGET	0.558298
FOLD_ENRICHMENT	35.897849
ZERO_CVG_TARGETS_PCT	0.32985
FOLD_80_BASE_PENALTY	3.451472
PCT_TARGET_BASES_2X	0.010046

Expect > 60% selected bases

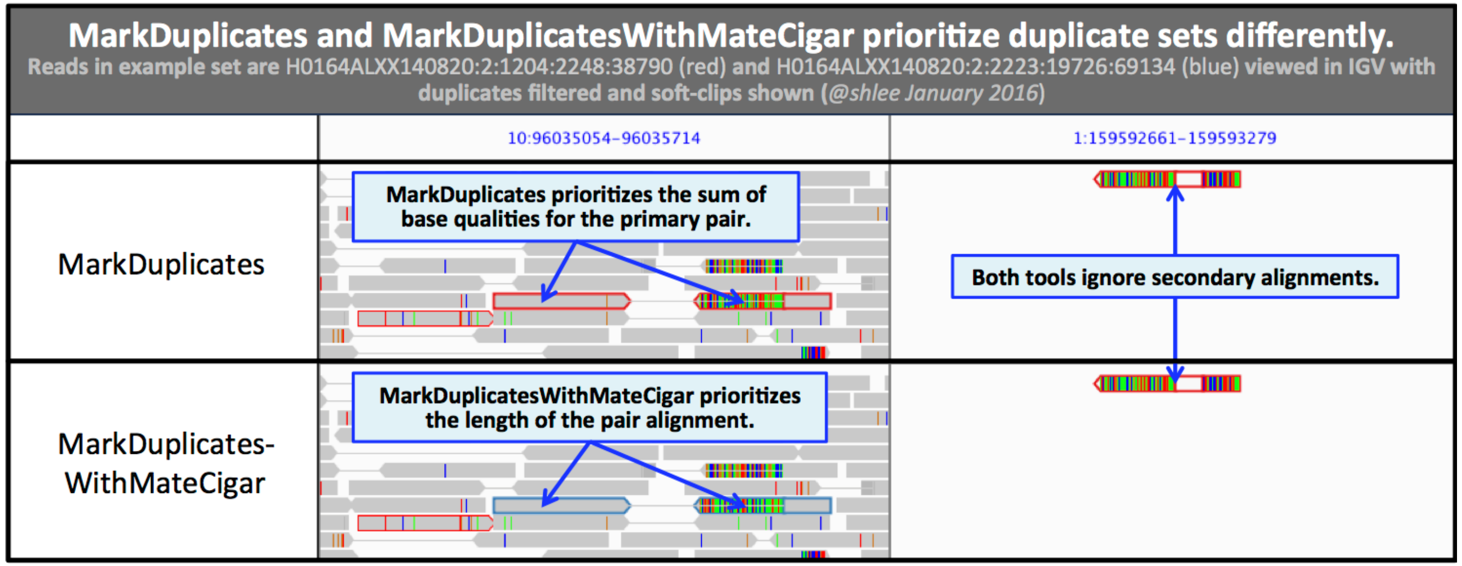
Expect > 80% bases covered at 20x

Expect enrichment > 30x

picard.sourceforge.net/picard-metric-definitions.shtml

Removing Duplicate Reads

- Remove read pairs with identical mapping coordinates
  - Assumed to be PCR duplicates
  - Unlikely to happen by chance in WGS
- Sets duplicate flag in BAM file
- Can also remove duplicates from file



MarkDuplicates and MarkDuplicatesWithMateCigar prioritize duplicate sets differently.

Reads in example set are H0164ALXX140820:2:1204:2248:38790 (red) and H0164ALXX140820:2:2223:19726:69134 (blue) viewed in IGV with duplicates filtered and soft-clips shown (*@shlee January 2016*)

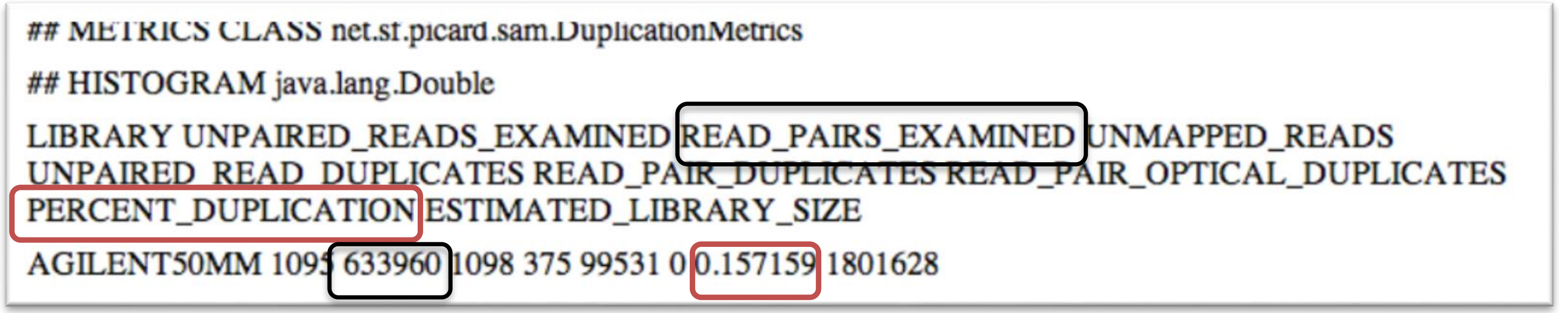
MarkDuplicates prioritizes the sum of base qualities for the primary pair.

MarkDuplicatesWithMateCigar prioritizes the length of the pair alignment.

Both tools ignore secondary alignments.

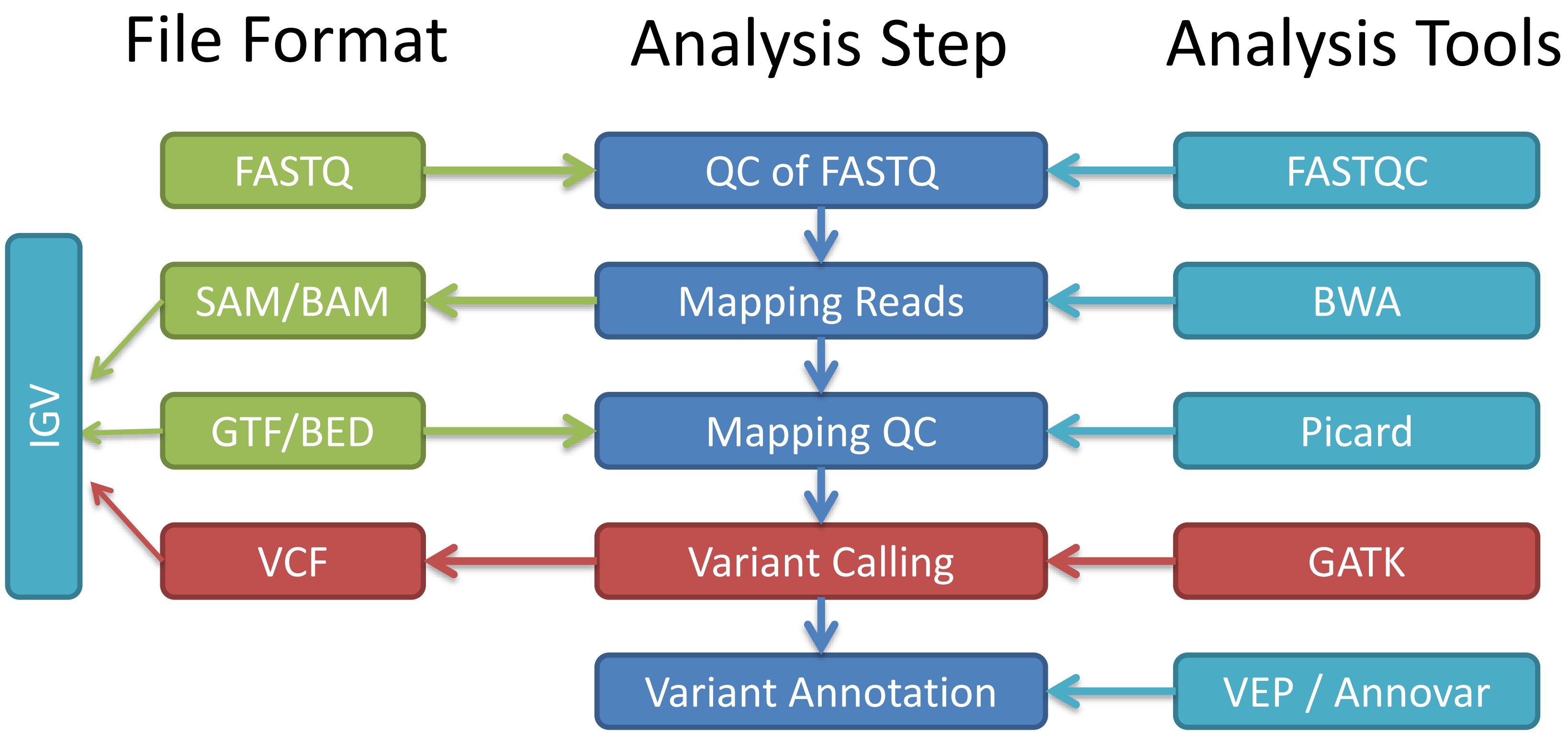
Output from Picard MarkDuplicates

- READ_PAIR_DUPLICATES:
  - The number of read pairs that were marked as duplicates.
- PERCENT_DUPLICATION:
  - The percentage of mapped sequence that is marked as duplicate.
- ESTIMATED_LIBRARY_SIZE:
  - The estimated number of unique DNA molecules in the library based on paired end duplication.



```
## METRICS CLASS net.sf.picard.sam.DuplicationMetrics
## HISTOGRAM java.lang.Double
LIBRARY UNPAIRED_READS_EXAMINED READ_PAIRS_EXAMINED UNMAPPED_READS
UNPAIRED_READ_DUPLICATES READ_PAIR_DUPLICATES READ_PAIR_OPTICAL_DUPLICATES
PERCENT_DUPLICATION ESTIMATED_LIBRARY_SIZE
AGILENT50MM 1095 633960 1098 375 99531 0 0.157159 1801628
```

Step 4 - Variant Calling



File Format

Analysis Step

Analysis Tools

FASTQ

QC of FASTQ

FASTQC

SAM/BAM

Mapping Reads

BWA

GTF/BED

Mapping QC

Picard

VCF

Variant Calling

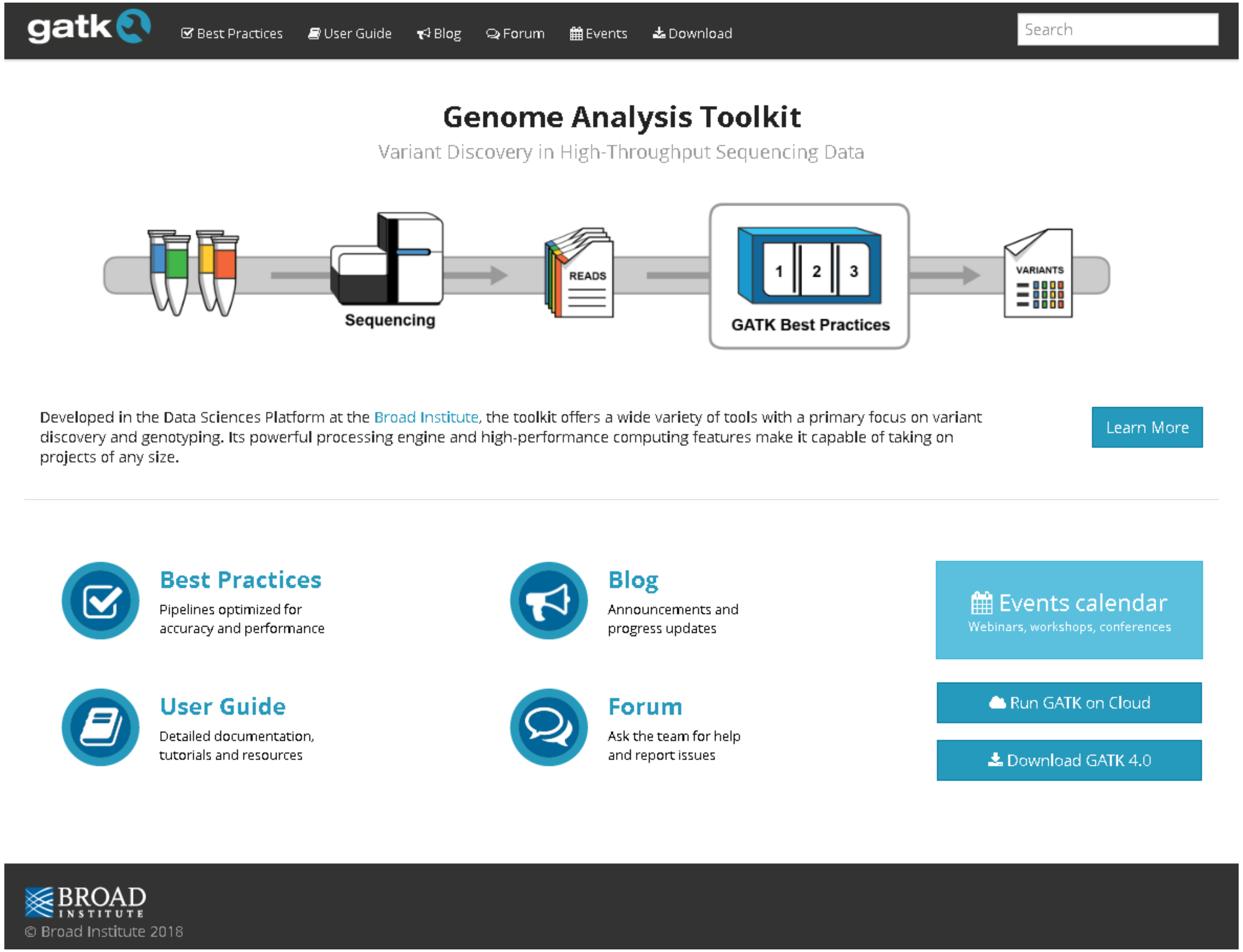
GATK

Variant Annotation

VEP / Annovar

IGV

Genome Analysis Toolkit



Genome Analysis Toolkit

Variant Discovery in High-Throughput Sequencing Data

Developed in the Data Sciences Platform at the Broad Institute, the toolkit offers a wide variety of tools with a primary focus on variant discovery and genotyping. Its powerful processing engine and high-performance computing features make it capable of taking on projects of any size.

Best Practices
Pipelines optimized for accuracy and performance

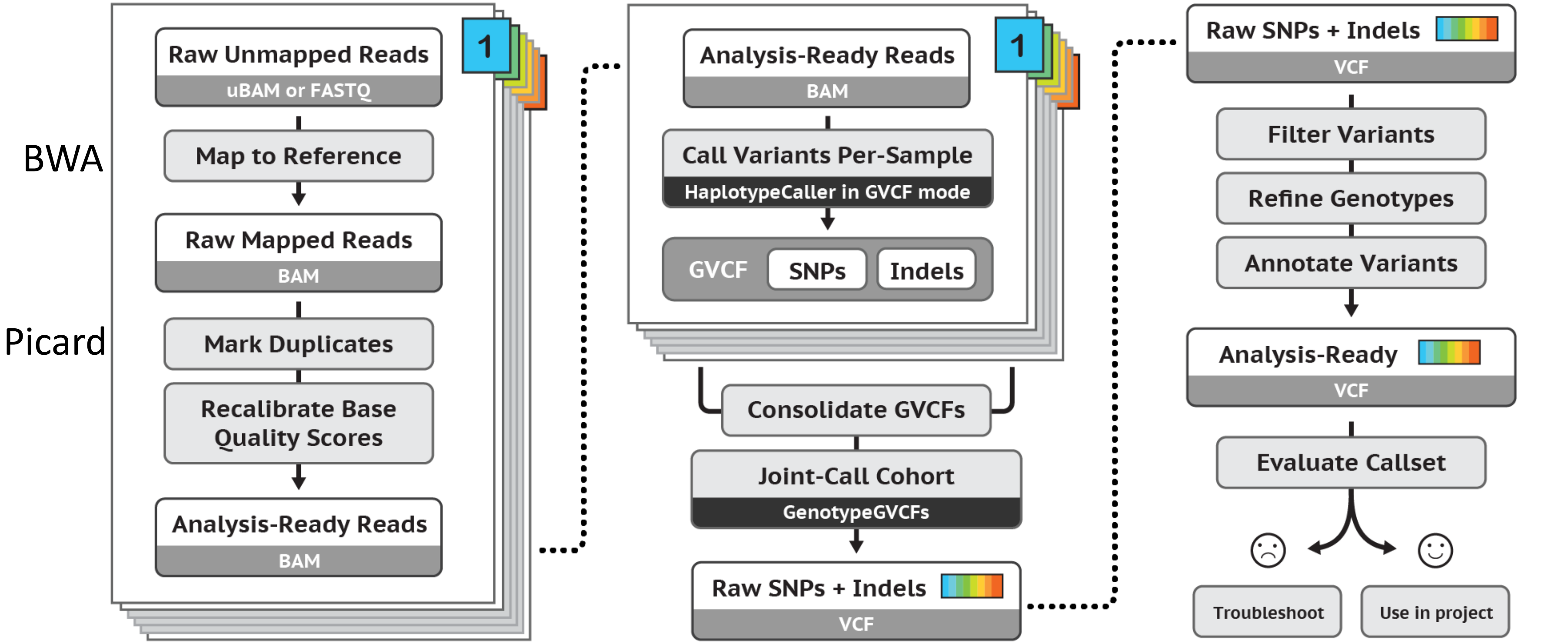
Blog
Announcements and progress updates

User Guide
Detailed documentation, tutorials and resources

Forum
Ask the team for help and report issues

GATK Best Practices Workflow

Germline Variants



1. Pre-processing

2. Variant Calling

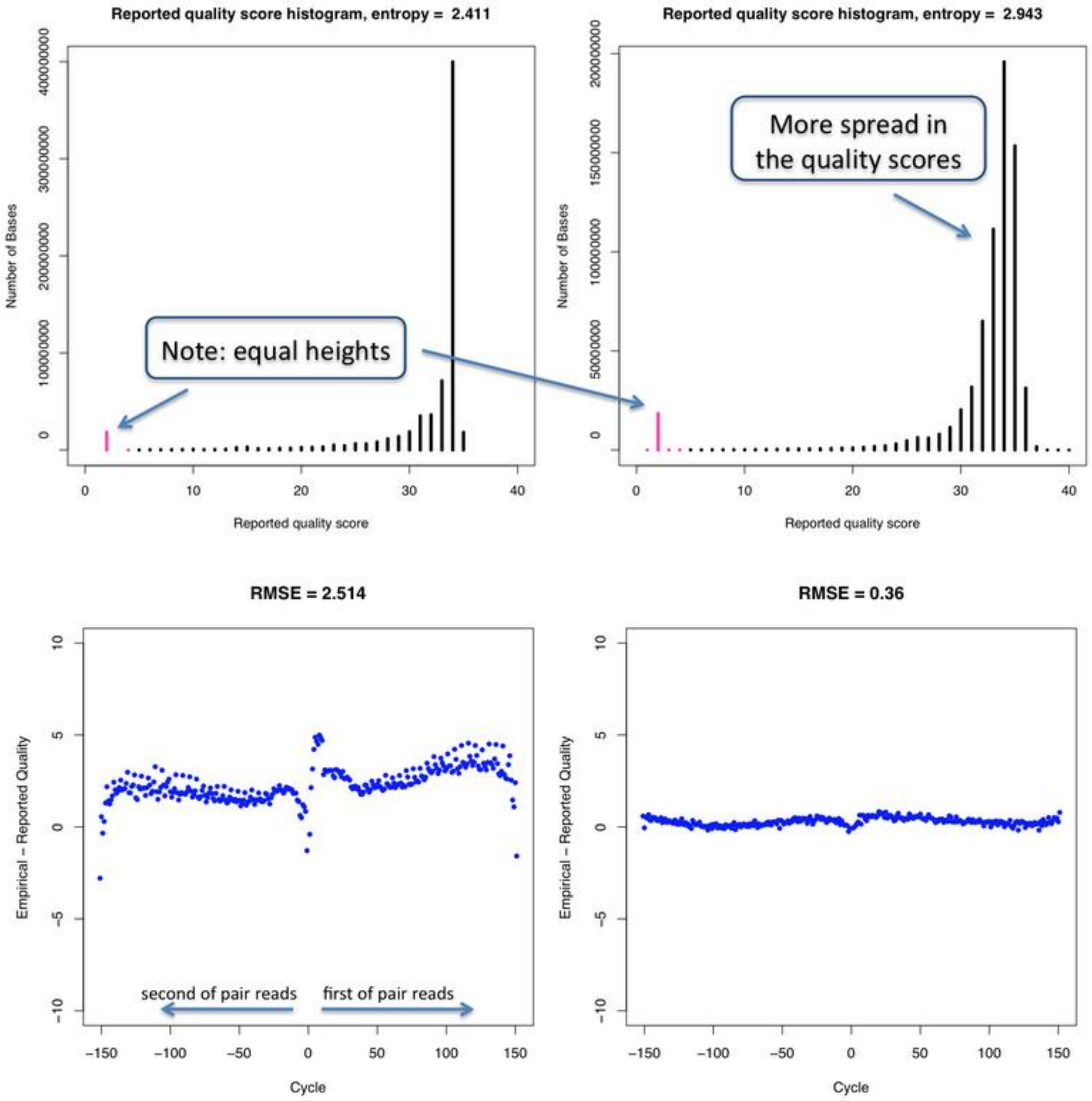
3. Annotation & Filtering

Base Quality Score Recalibration

- **Improves accuracy of base quality scores to reduce false positive variant calls**
- Analyse the variation among several features of a base:
  - Reported quality score
  - The position within the read
  - The preceding and current nucleotide (sequencing chemistry effect)
- These covariates are used to recalibrate the quality scores of all reads in a BAM file
- For example, a pre-calibration file could contain only reported Q25 bases
- These bases actually mismatch the reference at a 1 in 100 rate, so are actually Q20 empirically
- Base mismatches with the reference occur at the end of the reads more frequently than at the beginning
- Mismatches are strongly associated with sequencing context, in that the dinucleotide AC is often much lower quality than TG
- The recalibration tool corrects average Q score (shifting from Q25 to Q20)
- Also reduces quality of end of read AC bases compared to TG bases at the start of the read

Before

After



Note: equal heights

More spread in the quality scores

Reduced error per machine cycle

GATK HaplotypeCaller

- Call SNVs & indels by performing a local de-novo assembly
- Offers improved indel detection & phases haplotypes (vs UnifiedGenoyper)
- De-novo assembly method
  - Determine if a region has the potential to be variable
  - Construct a de Bruijn graph assembly of the region
  - The paths in the graph are potential haplotypes to be evaluated
  - Calculate haplotype likelihoods given the data
  - Determine if there are any variants on the most likely haplotypes
  - Compute the allele frequency distribution to determine most likely allele count
- Performed on each sample
- Produces a Genome VCF (.gvcf) file
- GVCFs can be combined using CombineGVCFs

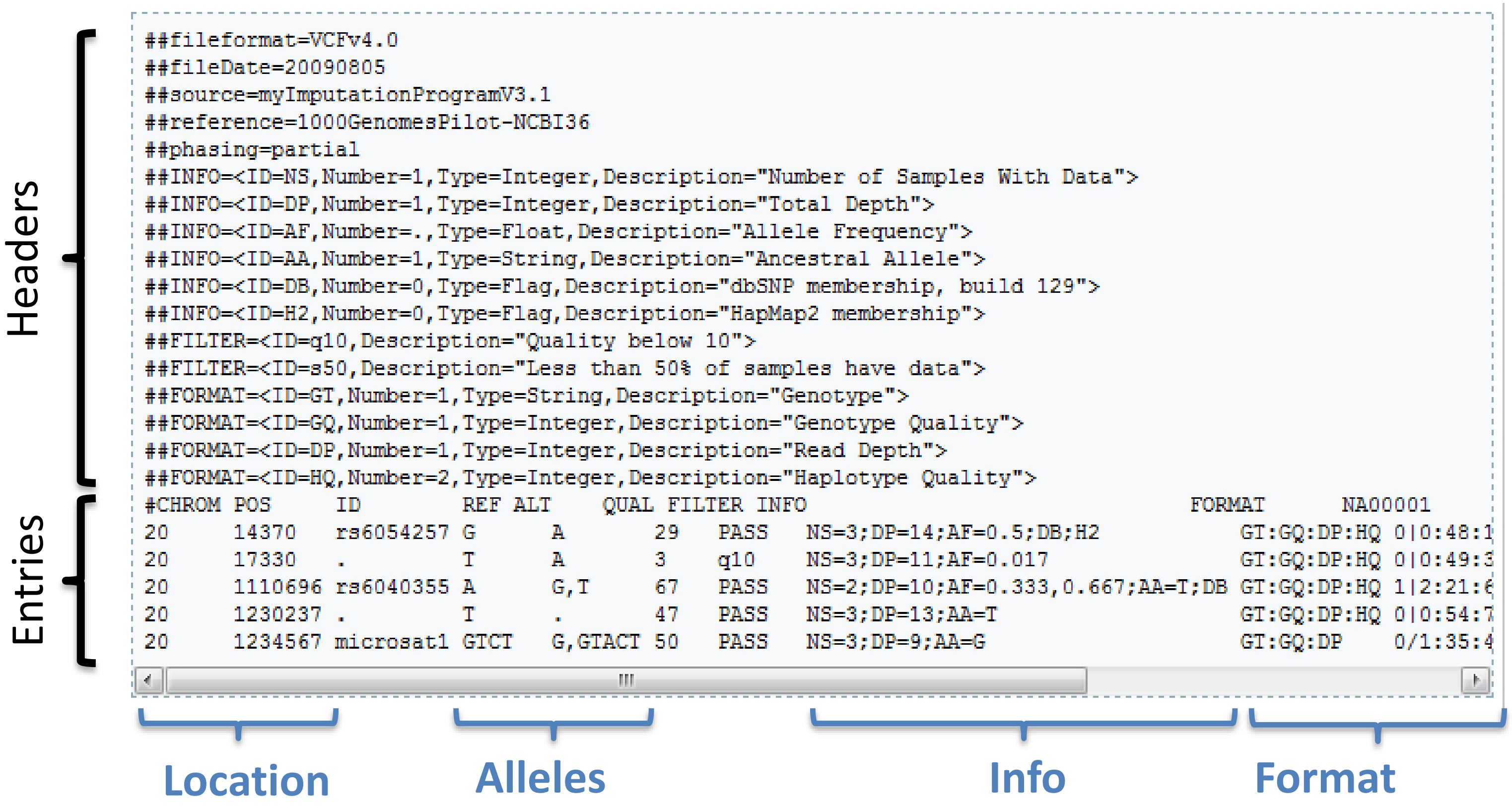
Joint Genotyping

- Gain power by calling variants on multiple samples
- Computationally inexpensive
- Works on combined GVCFs
- GenotypeGVCFs

Variant Quality Score Recalibration

- Variant calling is very sensitive
- Machine learning to identify variants that are likely to be real
  - Trains a model based on annotations of known variants
  - Applies the model to the entire dataset
  - Needs large, high quality set of known variants

VCF File Format



```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7
20     1234567 microsat1 GTCT   G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

http://www.1000genomes.org/node/101

Alterative Variant Callers

Joint Genotyping

- Samtools
- SomaticSniper
- FaSD-somatic
- JointSNVMix2
- Virmid
- SNVSniffer
- CaVEMan

Heuristic

- VarScan2
- SOAPsnv
- VarDict
- qSNP
- RADIA
- Shimmer

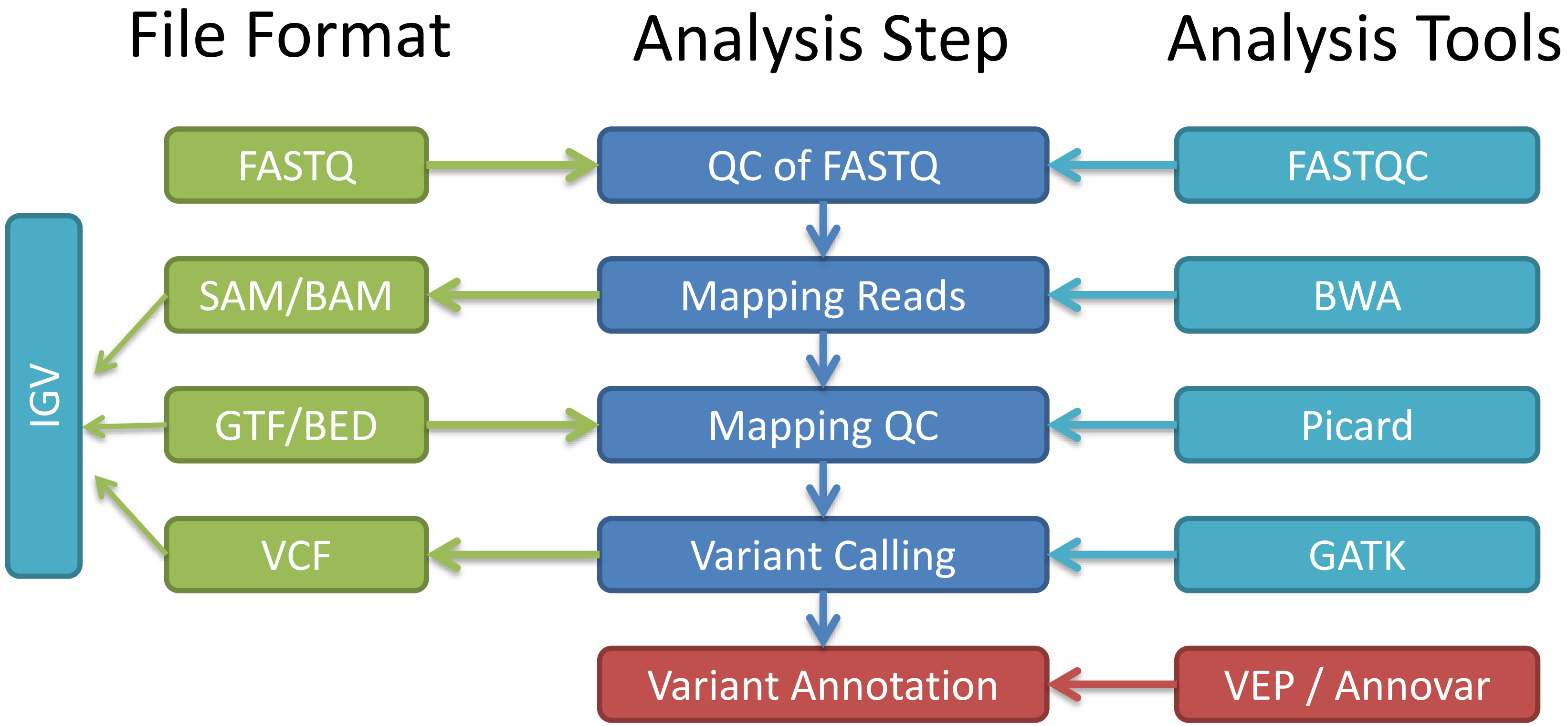
Joint Allele Freq

- Mutect2
- Strelka2
- LoFreq
- EBCall
- deepSNV
- LoLoPicker
- MuSE

Machine learning / ensemble methods: MutationSeq, SomaticSeq, SNooPer, BAYSIC

https://www.sciencedirect.com/science/article/pii/S2001037017300946

Step 5: Variant Annotation & Filtering



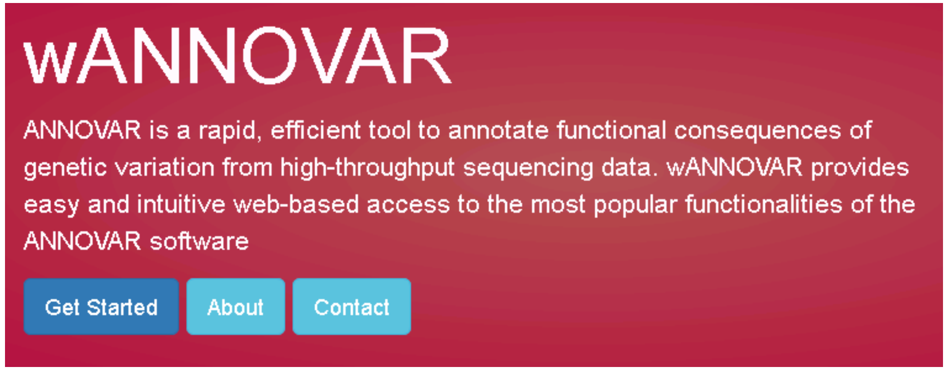
Variant Annotation

- Determining the potential biological action of SNVs and small indels
- Known variants
  - dbSNP identifier (rs number)
  - 1000 genomes allele frequency (rare / common)
- Variant location
  - Intron, UTR, CDS, splice site, promoter etc
- Variant effect
  - Non-synonymous coding / missense
  - Predicted effect (SIFT, Polyphen2, Provean)
  - Nonsense (premature stop codon)
  - Frameshift (indels)
- Variant evolutionary conservation
  - GERP, PhastCons
- Regulatory effects
  - Transcription factor binding sites
  - microRNAs
  - CpG islands
- Phenotype and disease association

Variant Annotation Tools

Ve!P

http://www.ensembl.org/info/docs/tools/vep/index.html



wANNOVAR

ANNOVAR is a rapid, efficient tool to annotate functional consequences of genetic variation from high-throughput sequencing data. wANNOVAR provides easy and intuitive web-based access to the most popular functionalities of the ANNOVAR software

http://wannovar.usc.edu/index.php

Mutation Taster

http://www.mutationtaster.org/StartQueryEngine.html

AVIA

Annotation, Visualization, and Impact Analysis

http://avia.abcc.ncifcrf.gov/apps/site/sub_analysis/?id=3

SNPnexus

http://www.snp-nexus.org/

VCF Filtering Tools

- Filtering of VCF files can be done using several tools
    1. SnpSift filter
    2. VCFfilter (part of the VCFlib toolkit)
    3. GATK Select Variants
- Can filter on any logical combination of fields
- Syntax varies between tools

VCF Filtering Schemes for Different Genetic Models

- De novo variants
  - Homozygous reference in unaffected parents and heterozygous in proband
- Recessive variants
  - Heterozygous in unaffected parents and homozygous non-reference in affected proband
  - Can also be compound heterozygous in proband (two different mutations in the same gene)
- Dominant variants
  - Heterozygous in affected individuals but homozygous reference in unaffected

Related workflows

- RNAseq variant calling
- UMI-based variant calling
- Structural variants
- Copy number variants