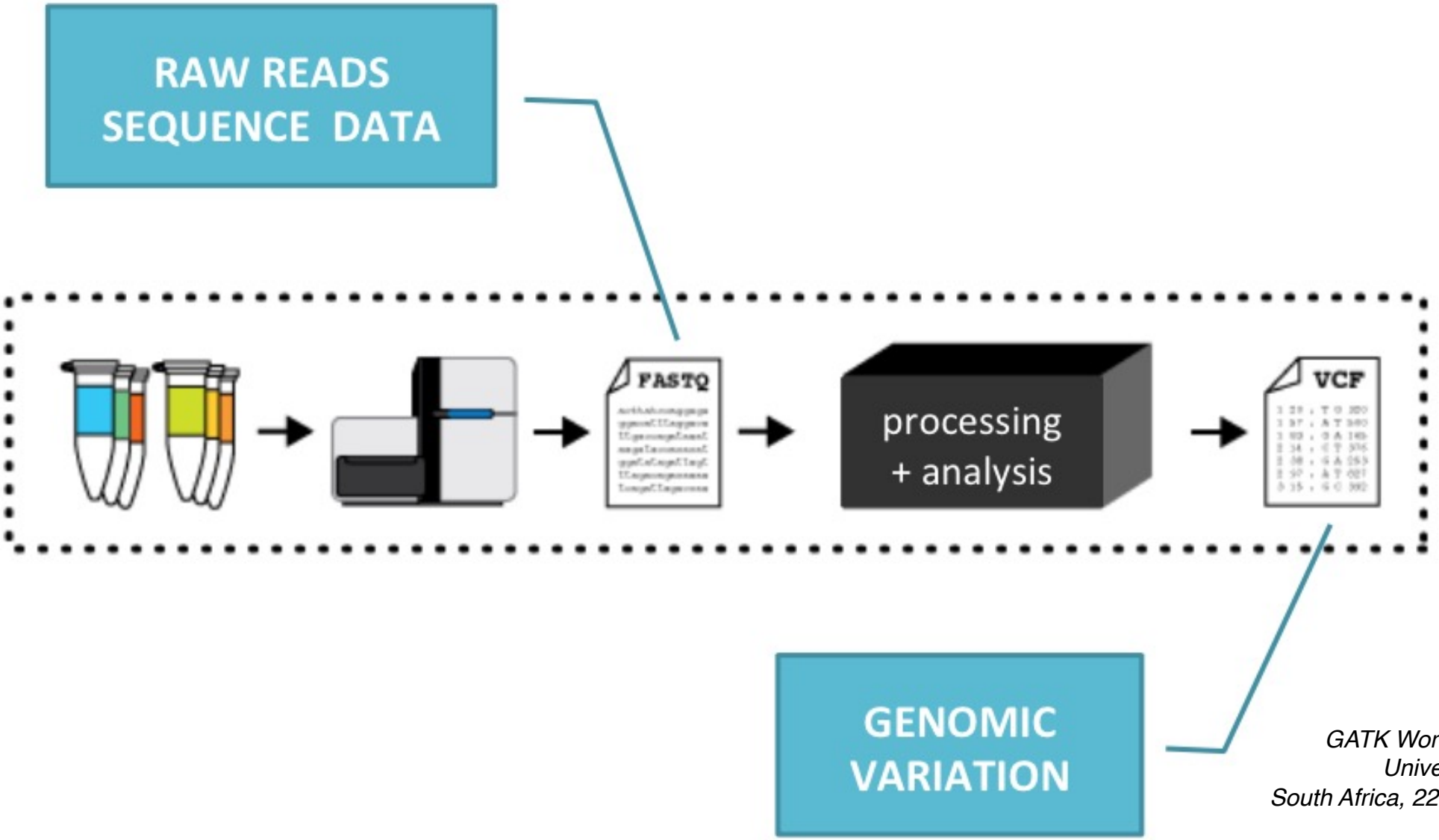
Special Topics in Biostatistics and Bioinformatics
Week XI

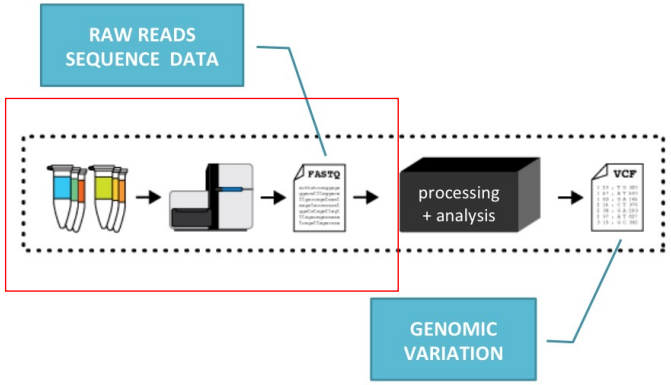
Ege Ülgen, M.D.

12 May 2022

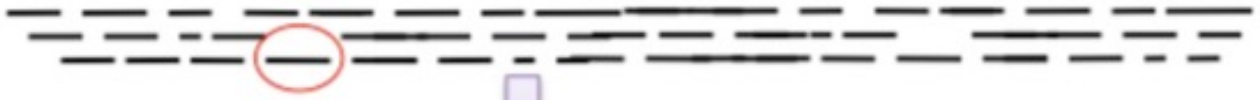
ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

Our Goal


RAW READS
SEQUENCE DATA
FASTQ
processing
+ analysis
VCF
GENOMIC
VARIATION

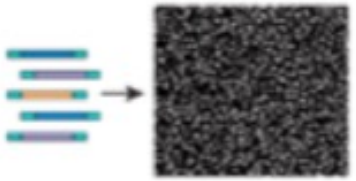


Fragmentation



Amplification

Library

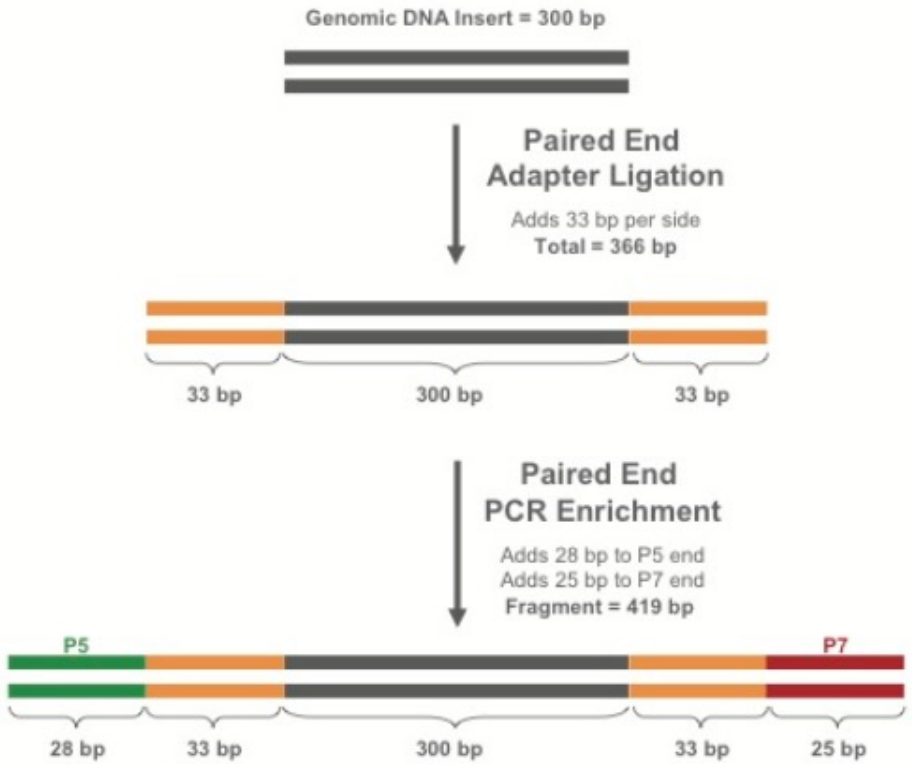


Sequencing

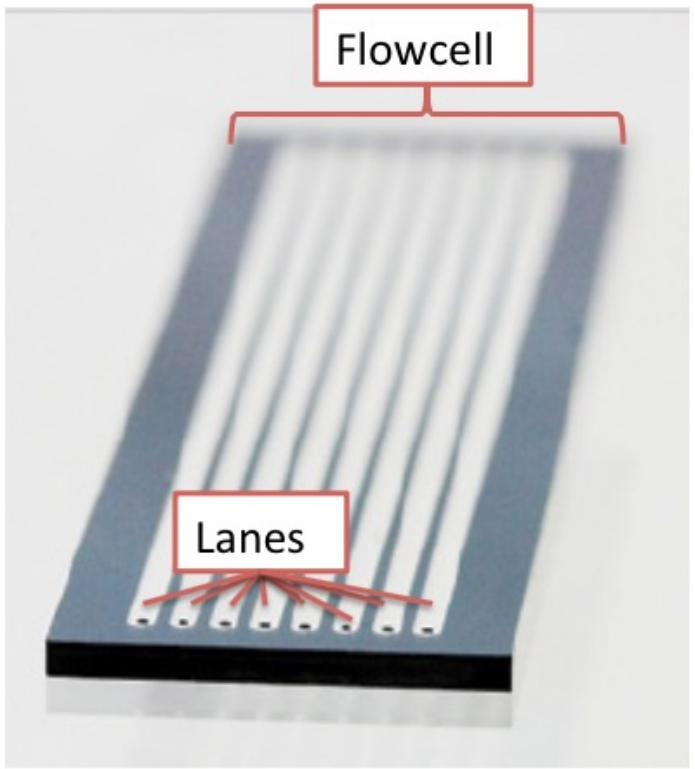
Massive parallel

GTAGGCTCGCTAT.....

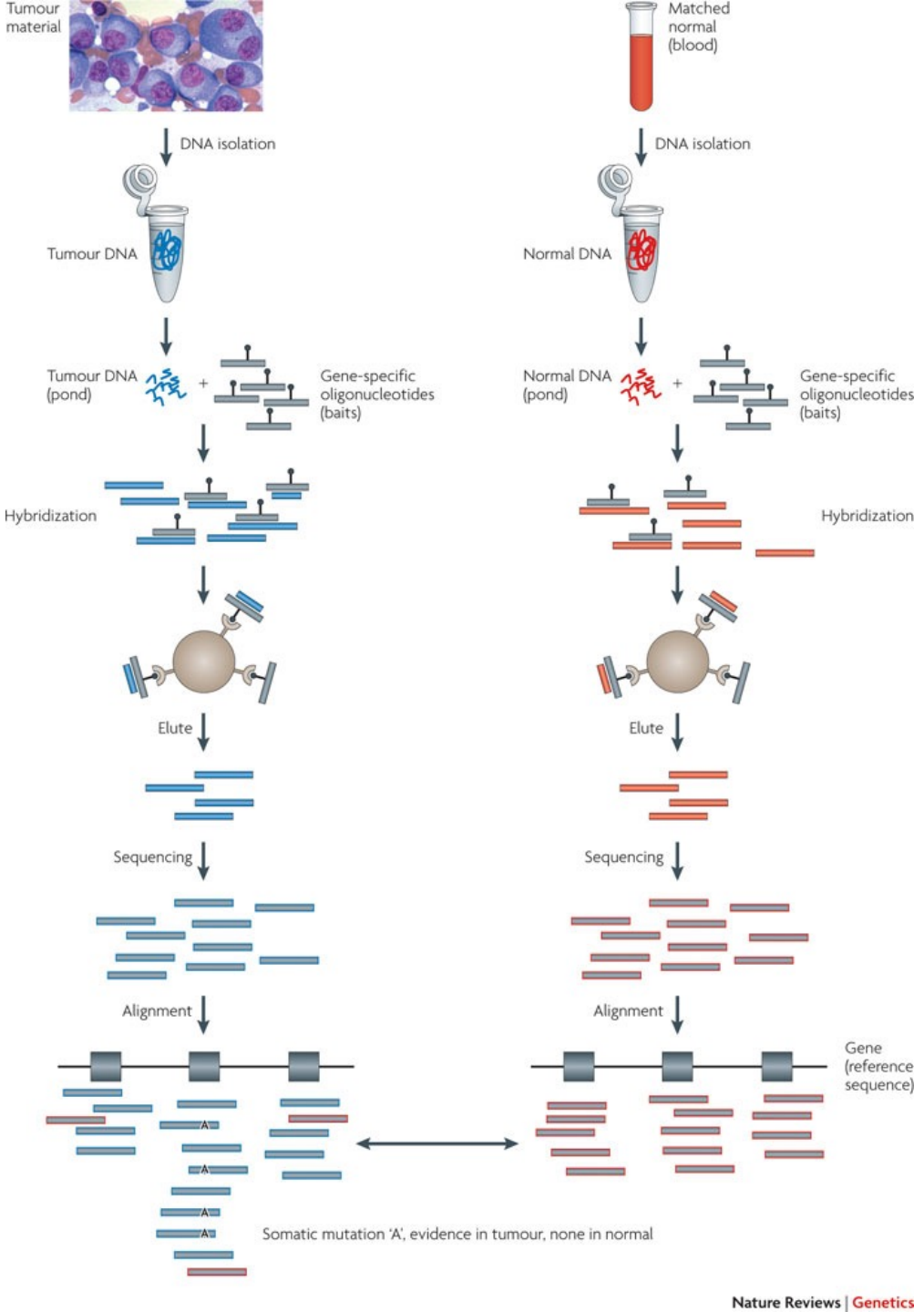
Terminology



Each reaction produces a unique **library** of DNA fragments for sequencing



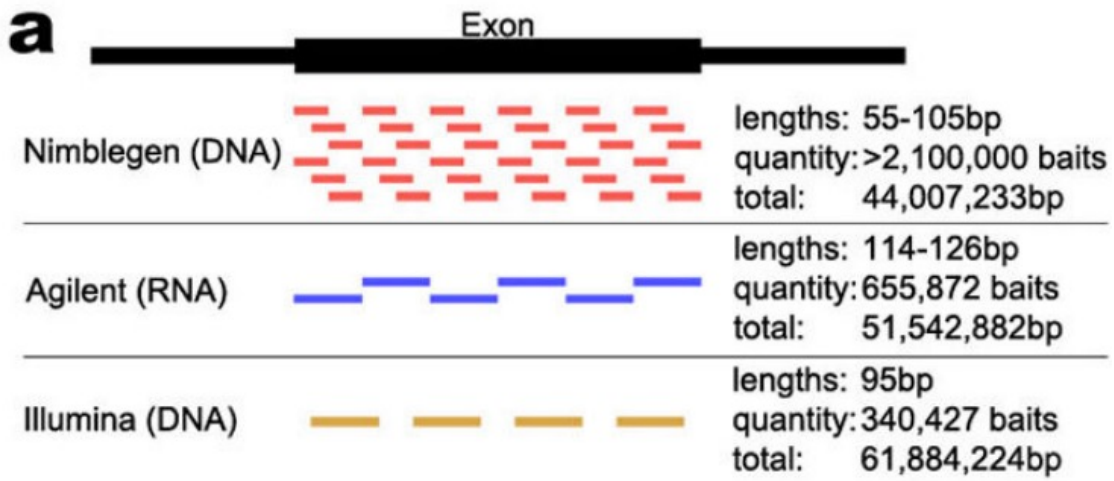
Each NGS machine processes a single **flowcell** containing several independent **lanes** during a single sequencing run

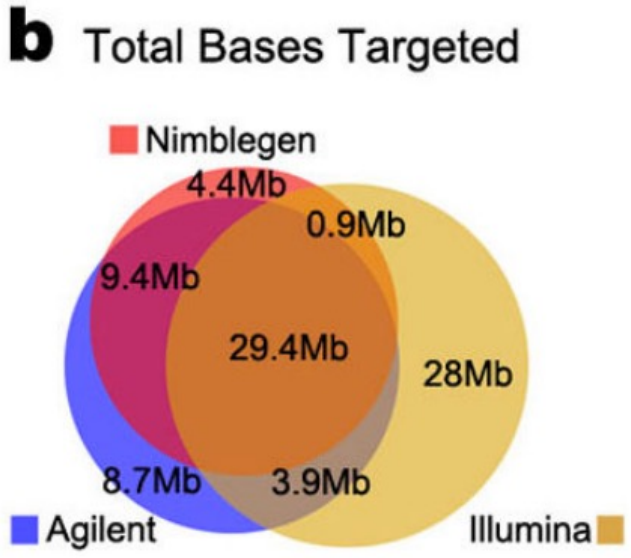


Meyerson et al., Nat Rev Genet 2010

Exome Capture

- Involves **baits** (complementary sequences) tiled to capture segments of target regions

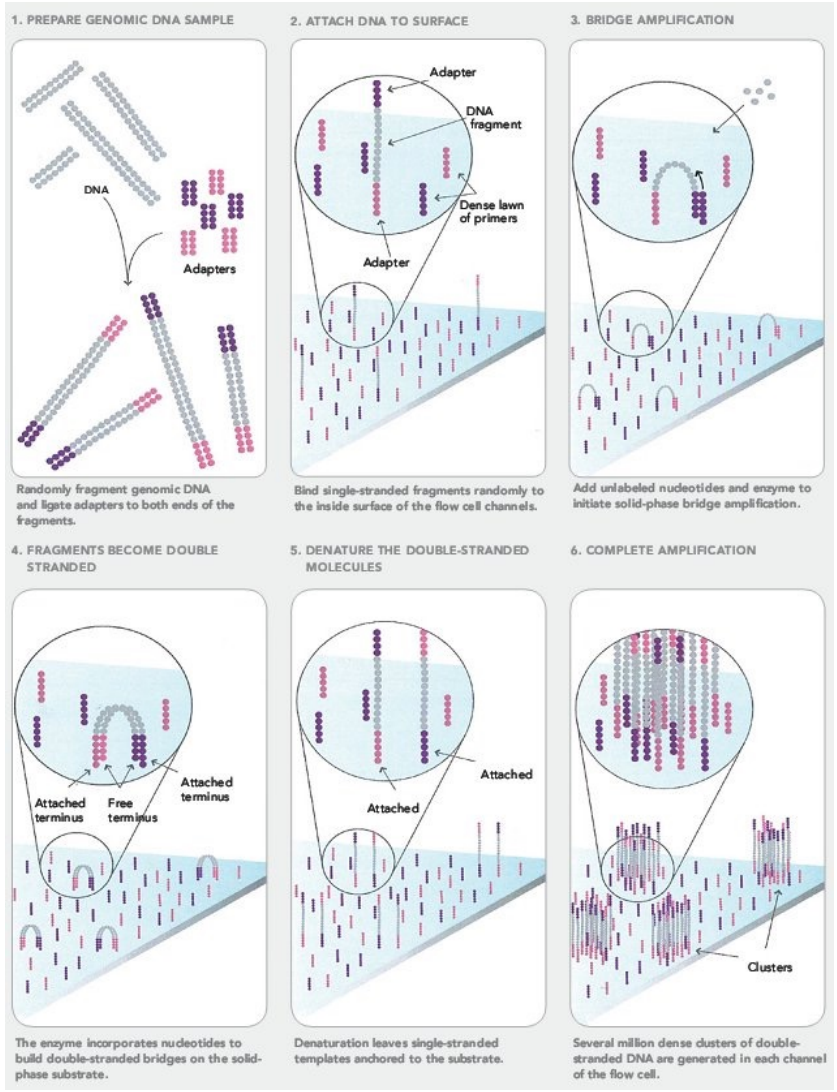


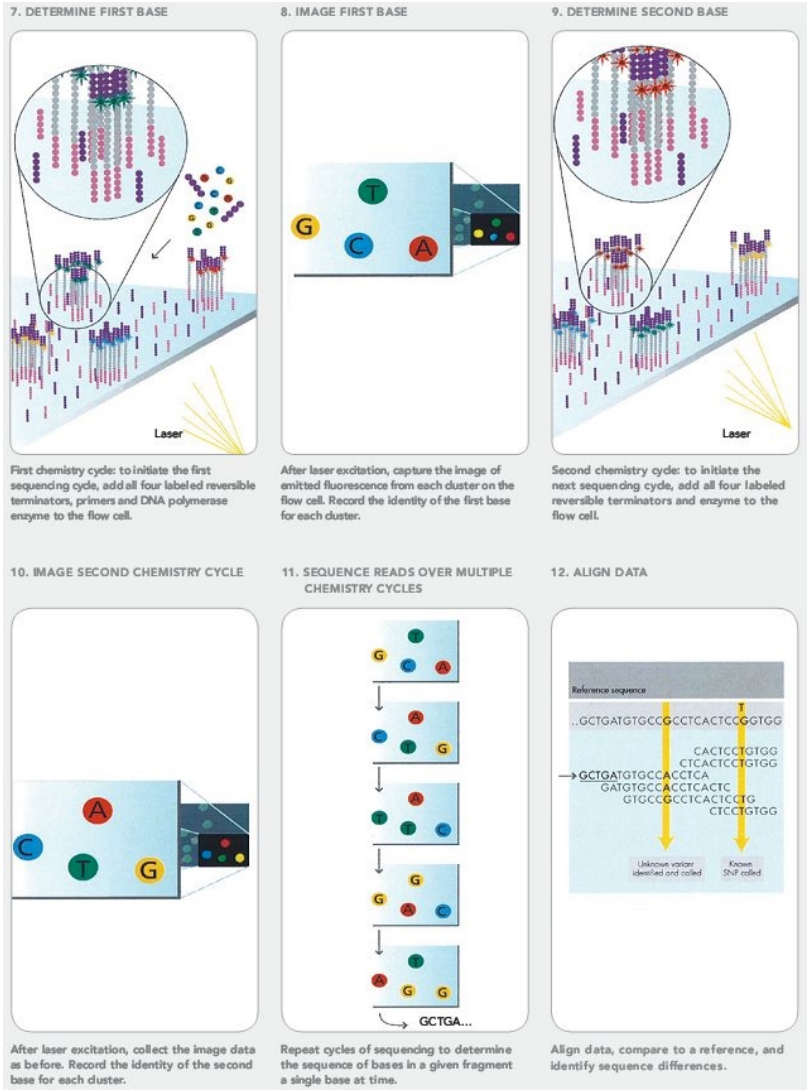


- Resulting covered intervals are **specific to capture kit manufacturer and the specific kit**

Nat Biotechnol. 2011 Sep 25;29(10):908-14. doi: 10.1038/nbt.1975.
Performance comparison of exome DNA sequencing technologies.
Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, Butte AJ, Snyder M.

Actual Sequencing - Illumina





Illumina Sequencing by Synthesis: https://youtu.be/fCd6B5HRaZ8

Terminology - Read

- A **read** is a sequence measurement
- The sequencing instrument “reads” the DNA fragment
- A read has a **size** (length) and **qualities** associated with it
- We call sequences as reads when in the context of measurements
- Some instruments produce reads of equal sizes even if the lengths of the DNA fragments vary

Terminology – Pair-end Sequencing

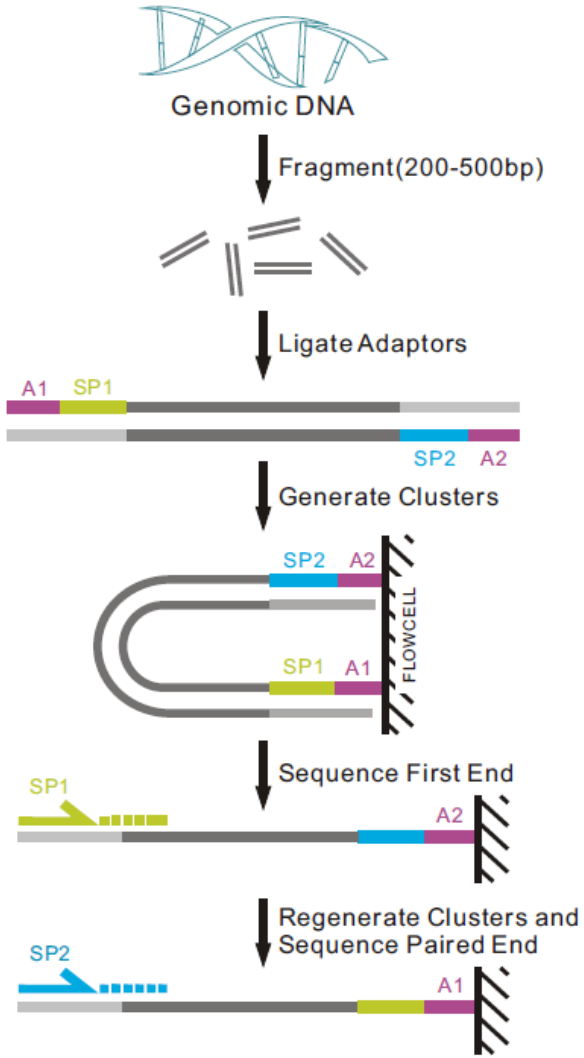


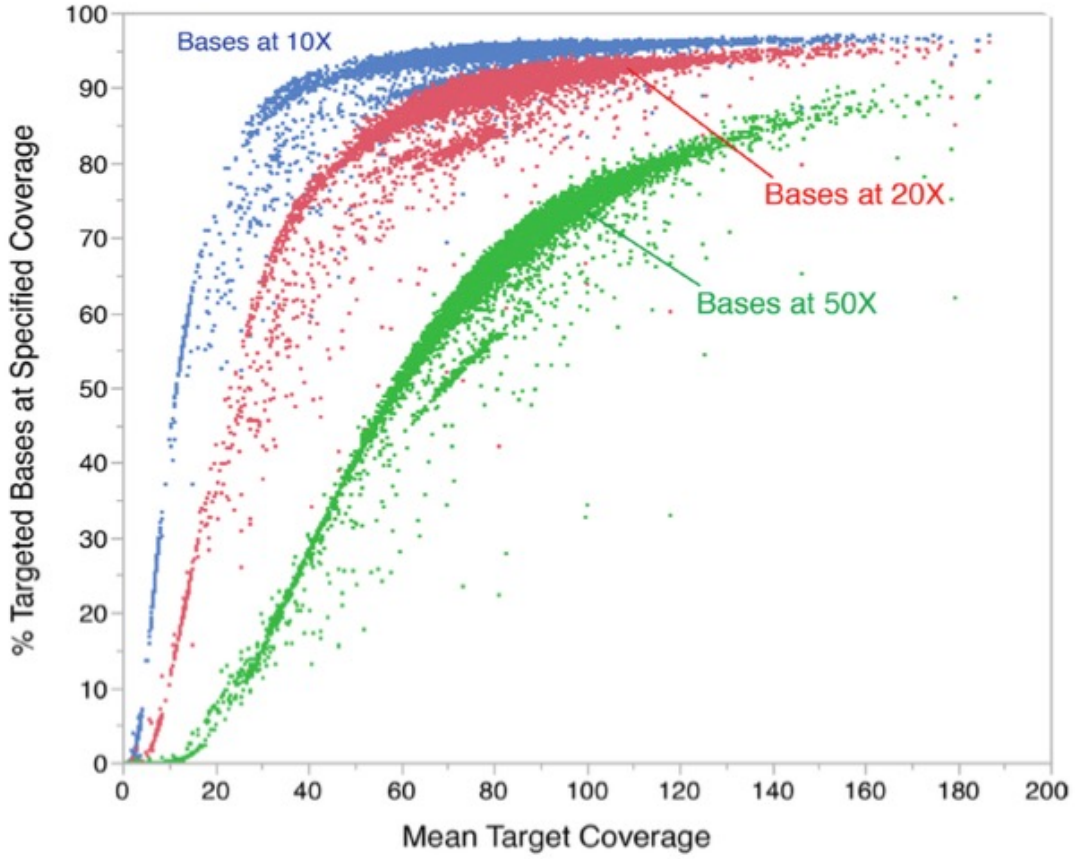
Figure 1-2-1 Pipeline of paired-end sequencing (www.illumina.com)

Pair-end Reads Advantage

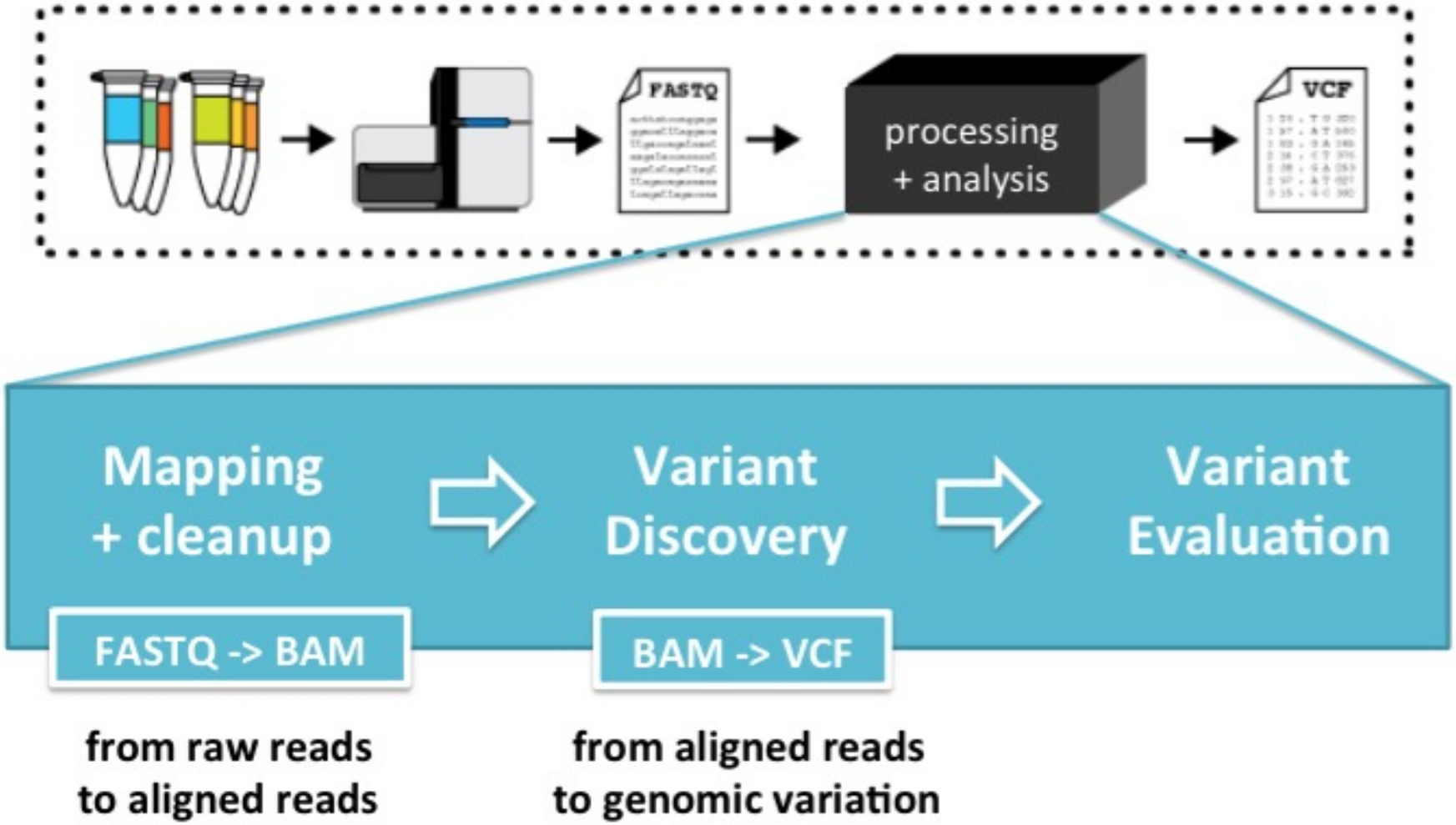
- DNA fragments are typically longer than the measured read lengths
  - For many applications, it is greatly advantageous to be able to measure (if not the entire fragment) at least both ends of it
  - Recommended for genomic variation and genome assembly analyses

Excerpt from The Biostar Handbook

Terminology – Coverage



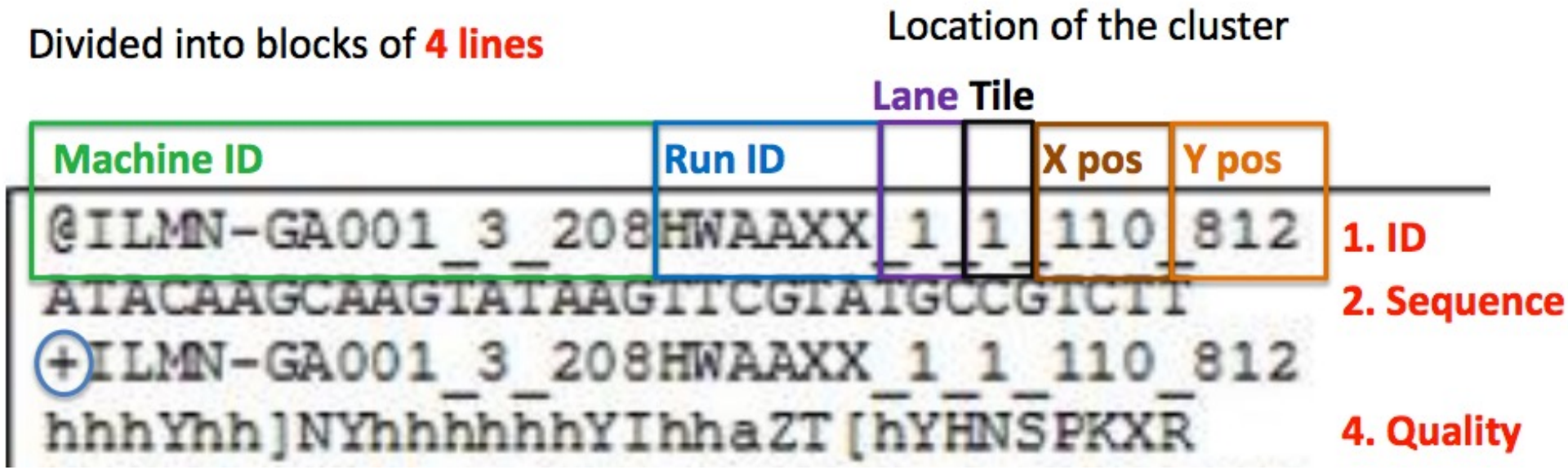
Distribution of coverage levels for targeted bases for representative samples sequenced to ~10X, ~20X, and ~50X mean target coverage


FASTQ
processing
+ analysis
VCF
Mapping
+ cleanup
Variant
Discovery
Variant
Evaluation
FASTQ -> BAM
BAM -> VCF
from raw reads
to aligned reads
from aligned reads
to genomic variation

Important File Formats

Important File Format #1: FASTQ (raw reads)

- Simple extension from traditional FASTA format.
- Each block has 4 elements (in 4 lines) :
  1. Sequence Name
  2. Sequence
  3. +(optional: Sequence name again)
  4. Associated quality score



Quality Scores

- Phred value = $-10 \times \log_{10}(\varepsilon)$

(*ε: Error Rate)

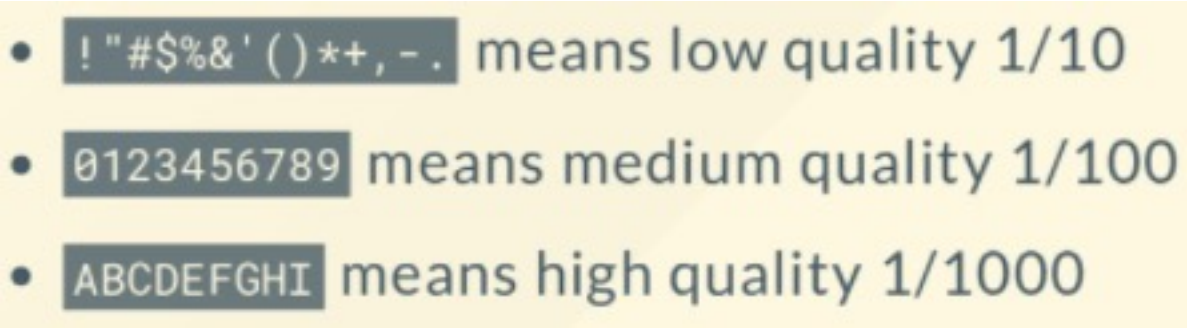
e.g.:

90% confidence >> 10% $\varepsilon = 10^{-1}$ >> Phred Q10 >> ASCII 74 = J

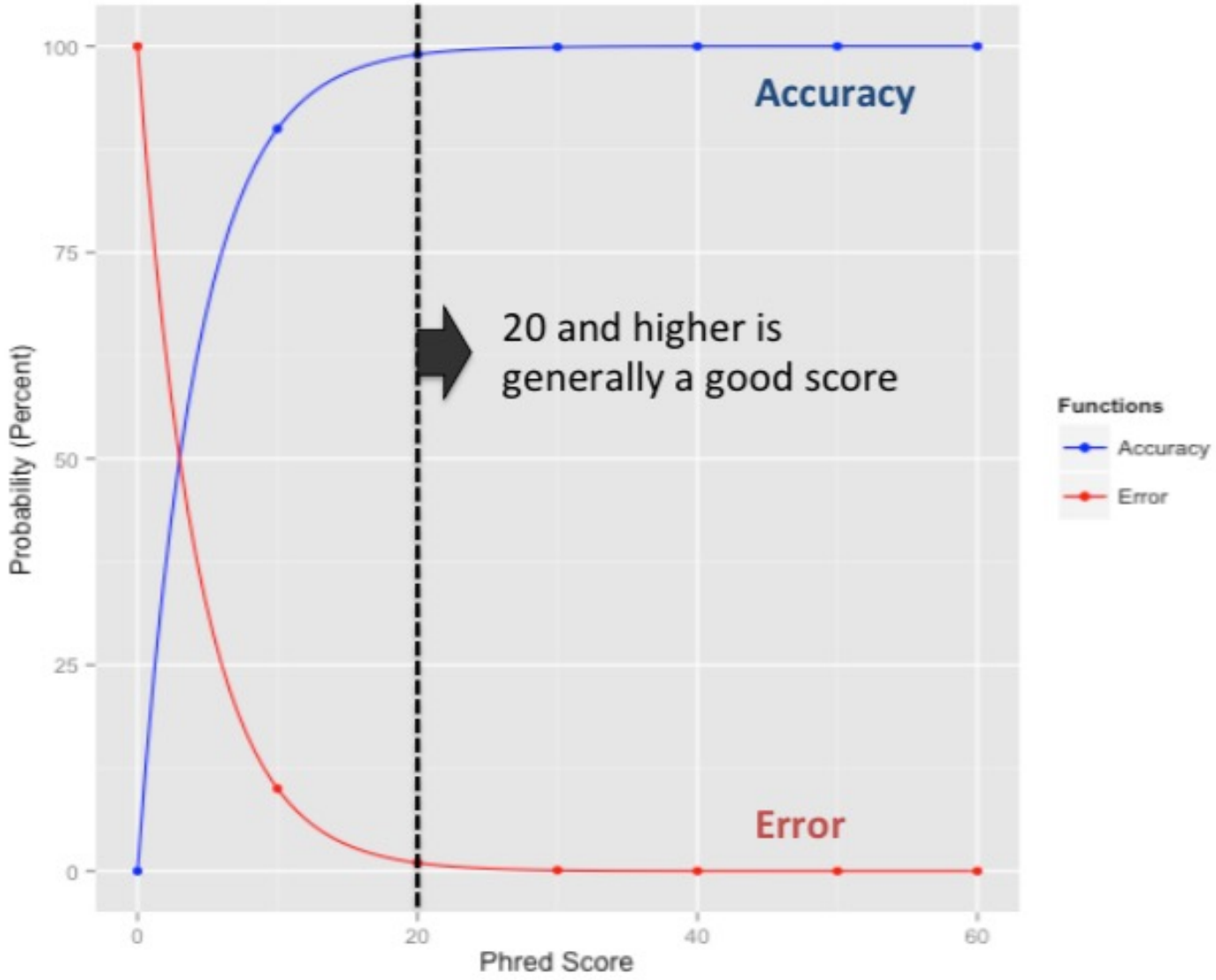
99% confidence >> 1% $\varepsilon = 10^{-2}$ >> Phred Q20 >> ASCII 84 = T

99.9% confidence >> 0.1% $\varepsilon = 10^{-3}$ >> Phred Q30 >> ASCII 94 = ^

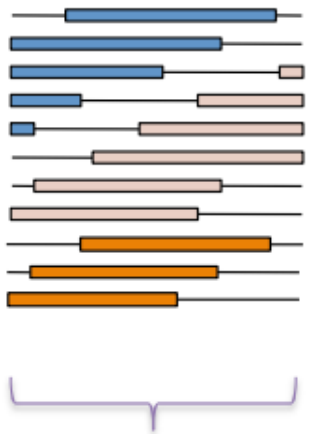
- There are different encoding schemes as well !!



Meaning of Phred Scores



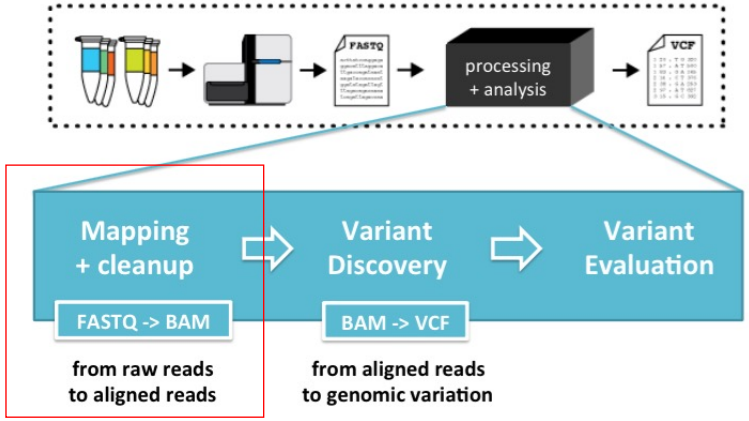
Alignment

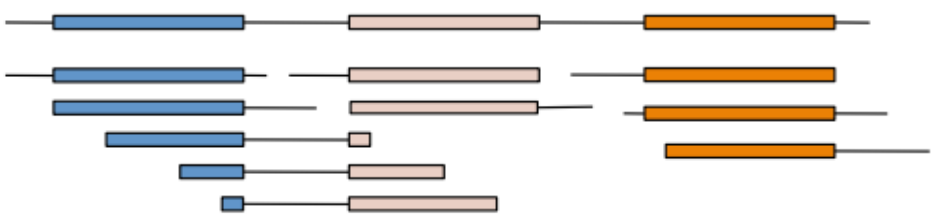


Enormous pile of short reads from HTS

Mapping and alignment algorithms

bwa
bowtie2
soap2 …



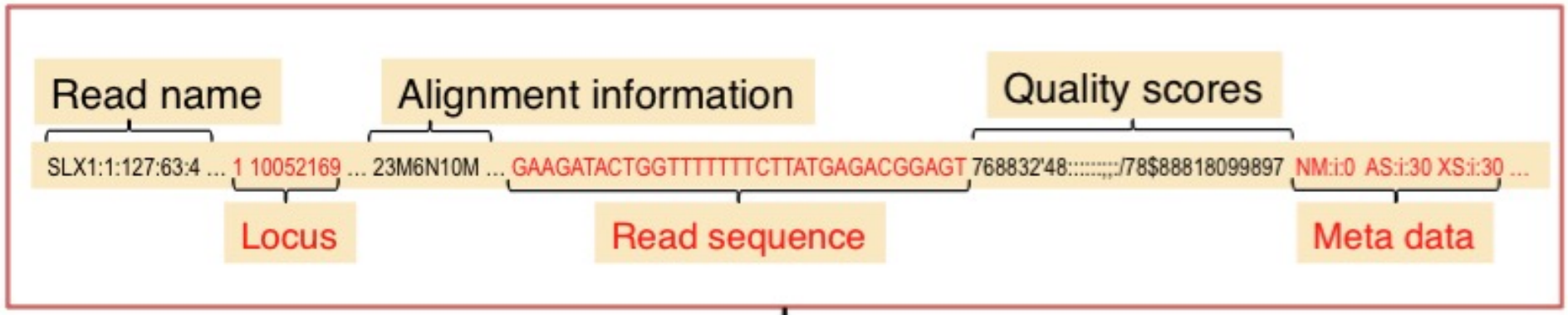


Reference genome

Reads mapped to reference

Important File Format #2: SAM/BAM(aligned reads)

Sequence **A**lignment **M**ap **/ B**inary **A**lignment **M**ap



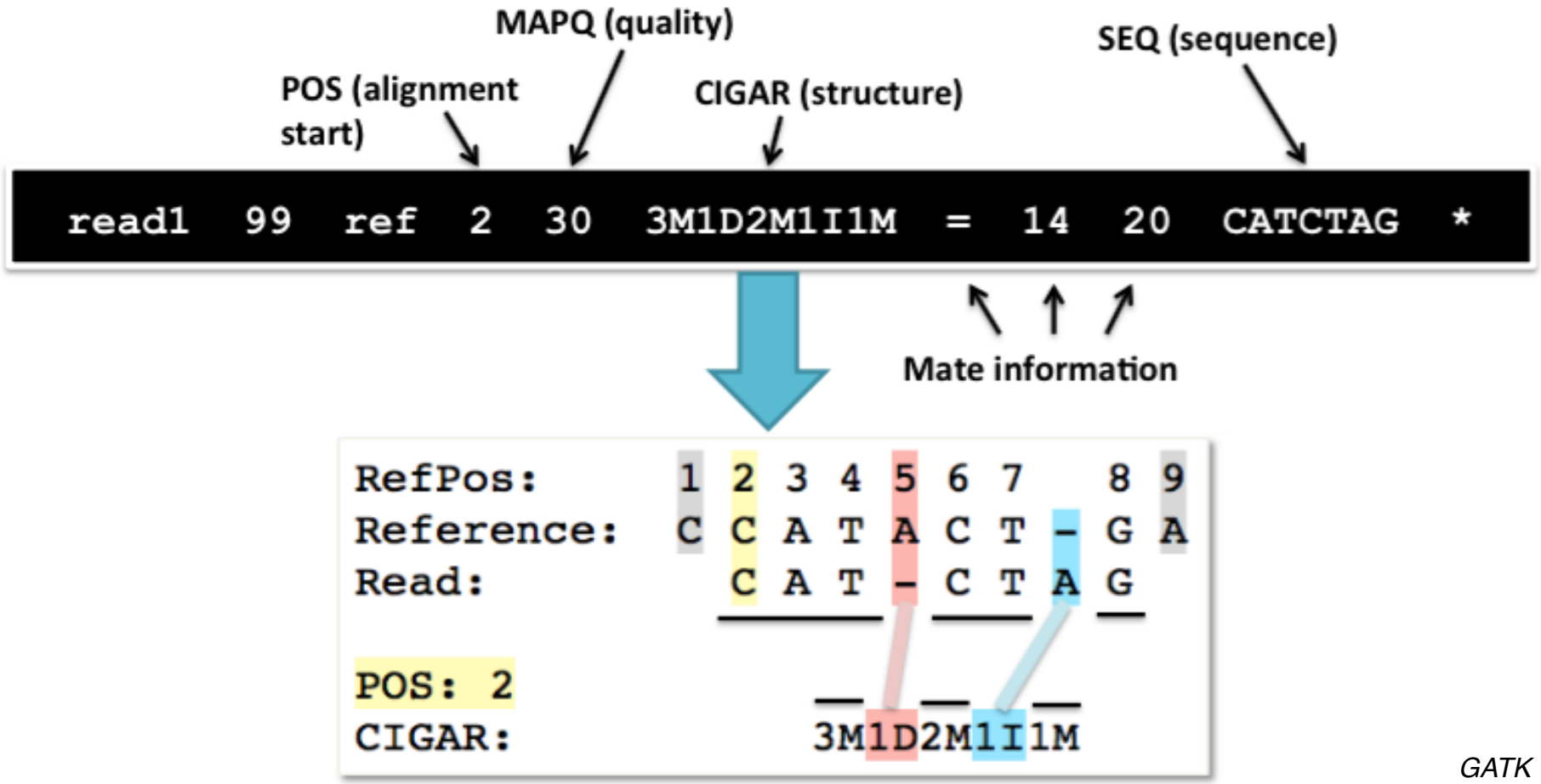
BAM file allows us to represent the data of any sequencer. Analyses can then be conducted largely agnostic to the particular sequencer used.

-> technology-independent

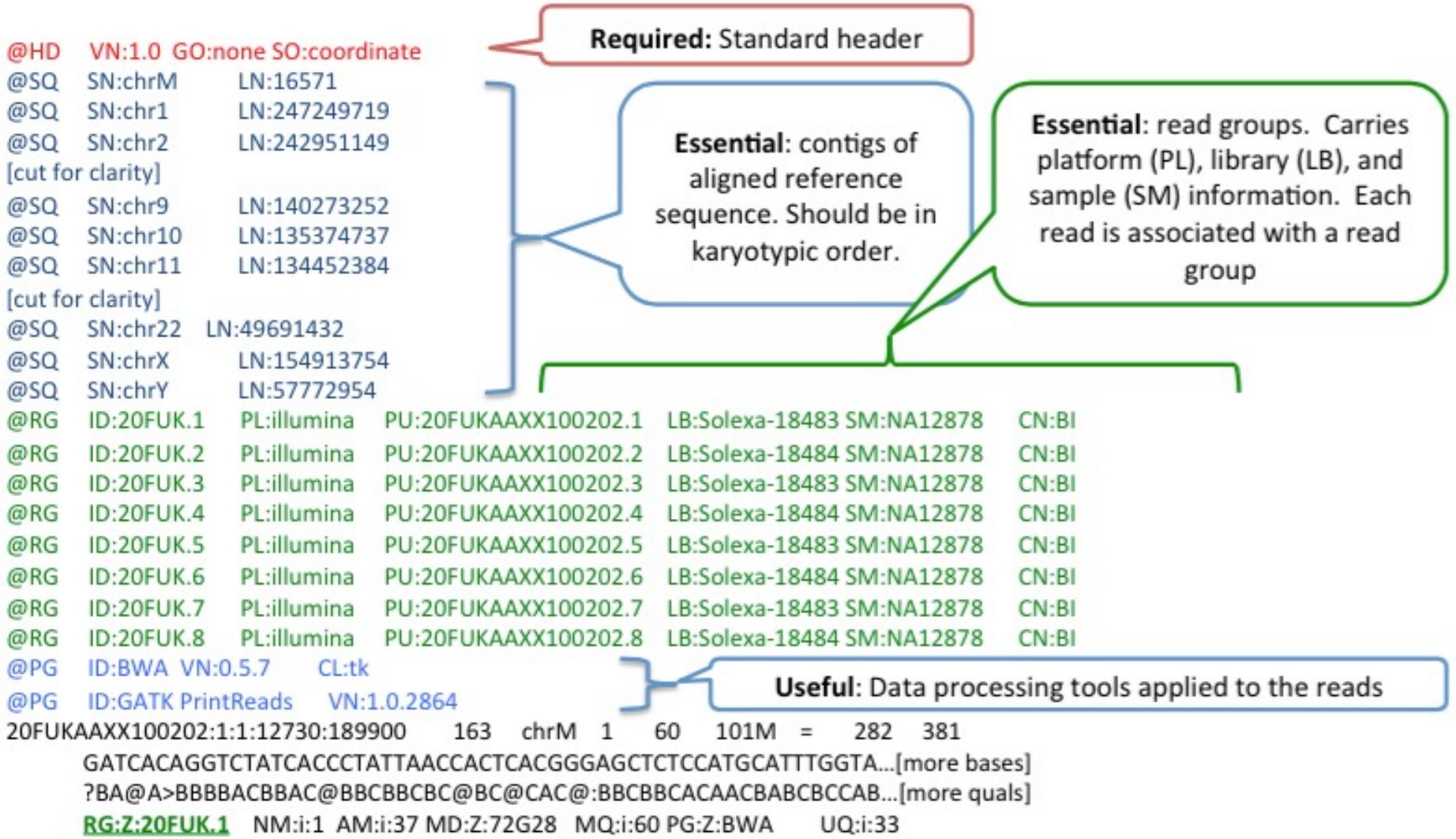
Data processing and analysis

A BAM file can contain data from a single or from several samples

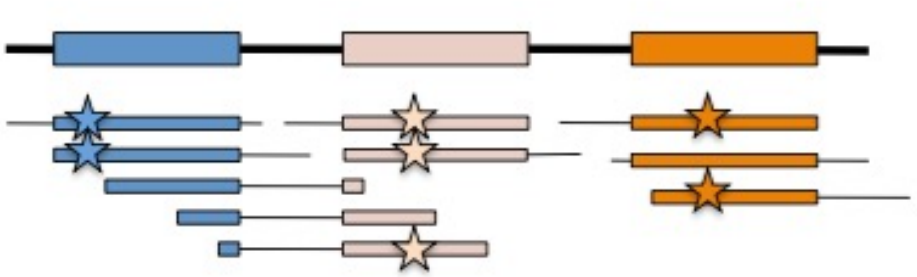
CIGAR – Explanation

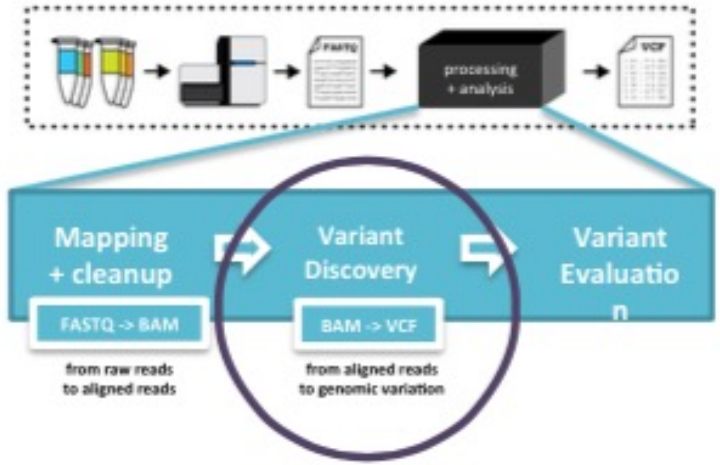


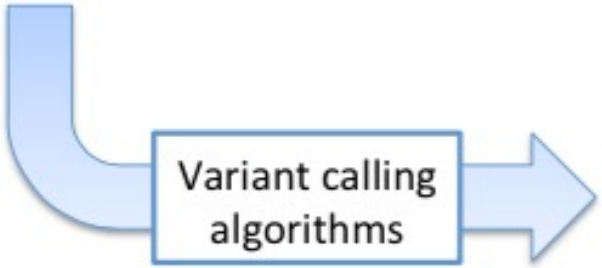
SAM Headers


@HD VN:1.0 GO:none SO:coordinate
@SQ SN:chrM LN:16571
@SQ SN:chr1 LN:247249719
@SQ SN:chr2 LN:242951149
[cut for clarity]
@SQ SN:chr9 LN:140273252
@SQ SN:chr10 LN:135374737
@SQ SN:chr11 LN:134452384
[cut for clarity]
@SQ SN:chr22 LN:49691432
@SQ SN:chrX LN:154913754
@SQ SN:chrY LN:57772954
@RG ID:20FUK.1 PL:illumina PU:20FUKAAXX100202.1 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.2 PL:illumina PU:20FUKAAXX100202.2 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.3 PL:illumina PU:20FUKAAXX100202.3 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.4 PL:illumina PU:20FUKAAXX100202.4 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.5 PL:illumina PU:20FUKAAXX100202.5 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.6 PL:illumina PU:20FUKAAXX100202.6 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.7 PL:illumina PU:20FUKAAXX100202.7 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.8 PL:illumina PU:20FUKAAXX100202.8 LB:Solexa-18484 SM:NA12878 CN:BI
@PG ID:BWA VN:0.5.7 CL:tk
@PG ID:GATK PrintReads VN:1.0.2864
20FUKAAXX100202:1:1:12730:189900 163 chrM 1 60 101M = 282 381
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTA...[more bases]
?BA@A>BBBBACBBAC@BBCBBCBC@BC@CAC@:BBCBBCACAACBABCBCCAB...[more quals]
RG:Z:20FUK.1 NM:i:1 AM:i:37 MD:Z:72G28 MQ:i:60 PG:Z:BWA UQ:i:33
Required: Standard header
Essential: contigs of aligned reference sequence. Should be in karyotypic order.
Essential: read groups. Carries platform (PL), library (LB), and sample (SM) information. Each read is associated with a read group
Useful: Data processing tools applied to the reads

Important File Format #3: VCF Genomic Variation







site 1 description
+ sample genotypes

site 2 description
+ sample genotypes

site 3 description
+ sample genotypes

VCF Format

```
##fileformat=VCFv4.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership">
##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">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO
    FORMAT      NA00001        NA00002        NA00003

20     14370   rs6054257 G      A       29   PASS   DP=14;AF=0.5;DB
    GT:GQ:DP 0/0:48:1 1/0:48:8 1/1:43:5
20     1110696 rs6040355 A      G,T     67   PASS   DP=10;AF=0.333,0.667;DB
    GT:GQ:DP 1/2:21:6 2/1:2:0   2/2:35:4
20     1230237 .         T      .       47   PASS   DP=13
    GT:GQ:DP 0/0:54:7 0/0:48:4 0/0:61:2
20     1234567 microsat1 GTCT   G,GTACT 50   PASS   DP=9
    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Header

Variant records

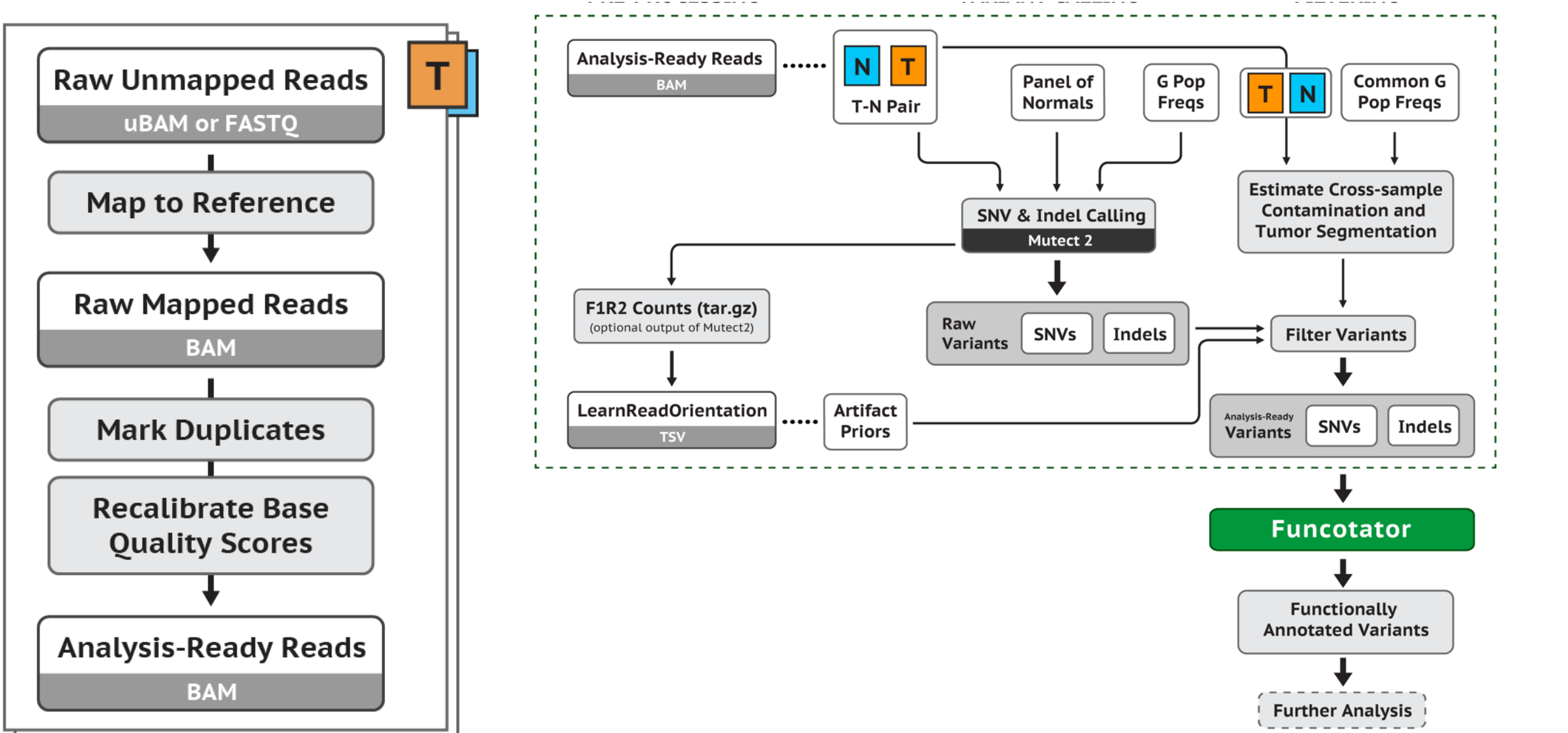
Our Goal

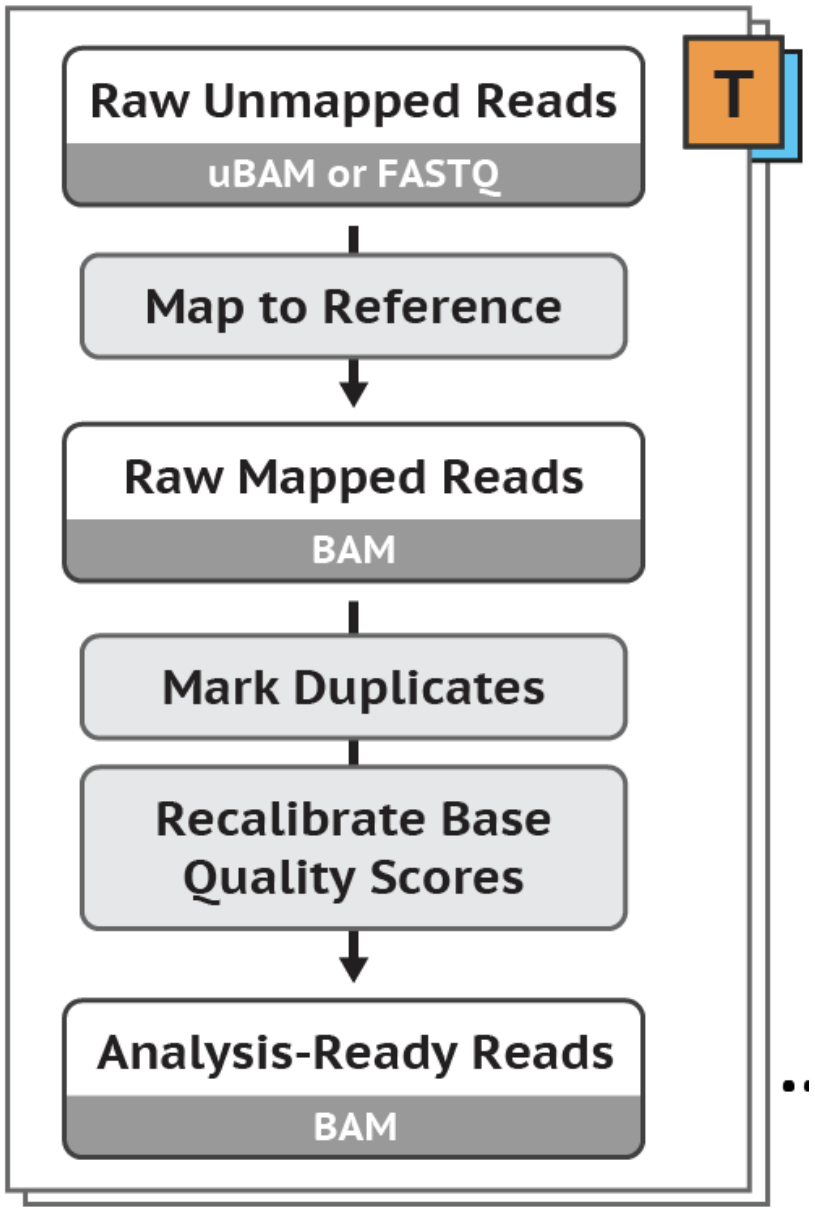
(or some other workflow)

Genome Analysis Toolkit

- Developed at Broad Institute
- Offers a wide variety of tools with a primary focus on variant discovery and genotyping
- Excellent material (Seminar slides, discussion forum, documentations, tutorials, blog…) and helpful staff
- Offers “Best Practices”:
  - pipelines optimized for accuracy and performance

Current Best Practices for Somatic Variations





Mapping and Preprocessing

Before all – Quality Control

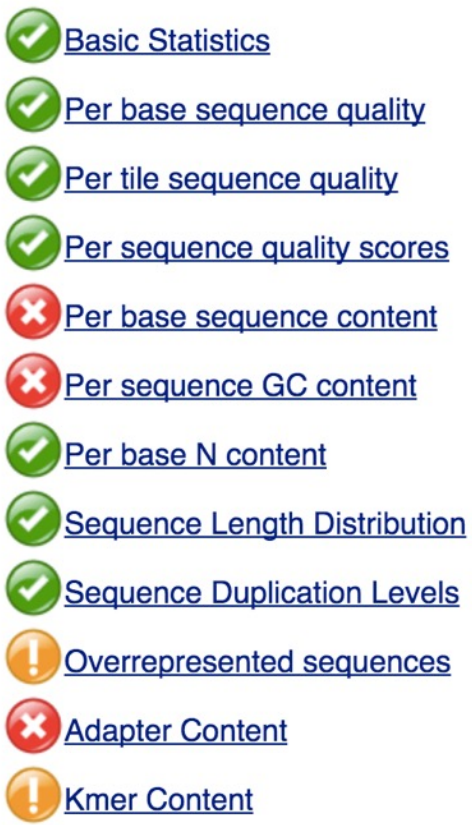
- The first step performed after data acquisition
- The process of evaluating and improving data by removing identifiable errors from it
- **QC cannot turn bad data into good data**, and we can never salvage what appears to be a total loss

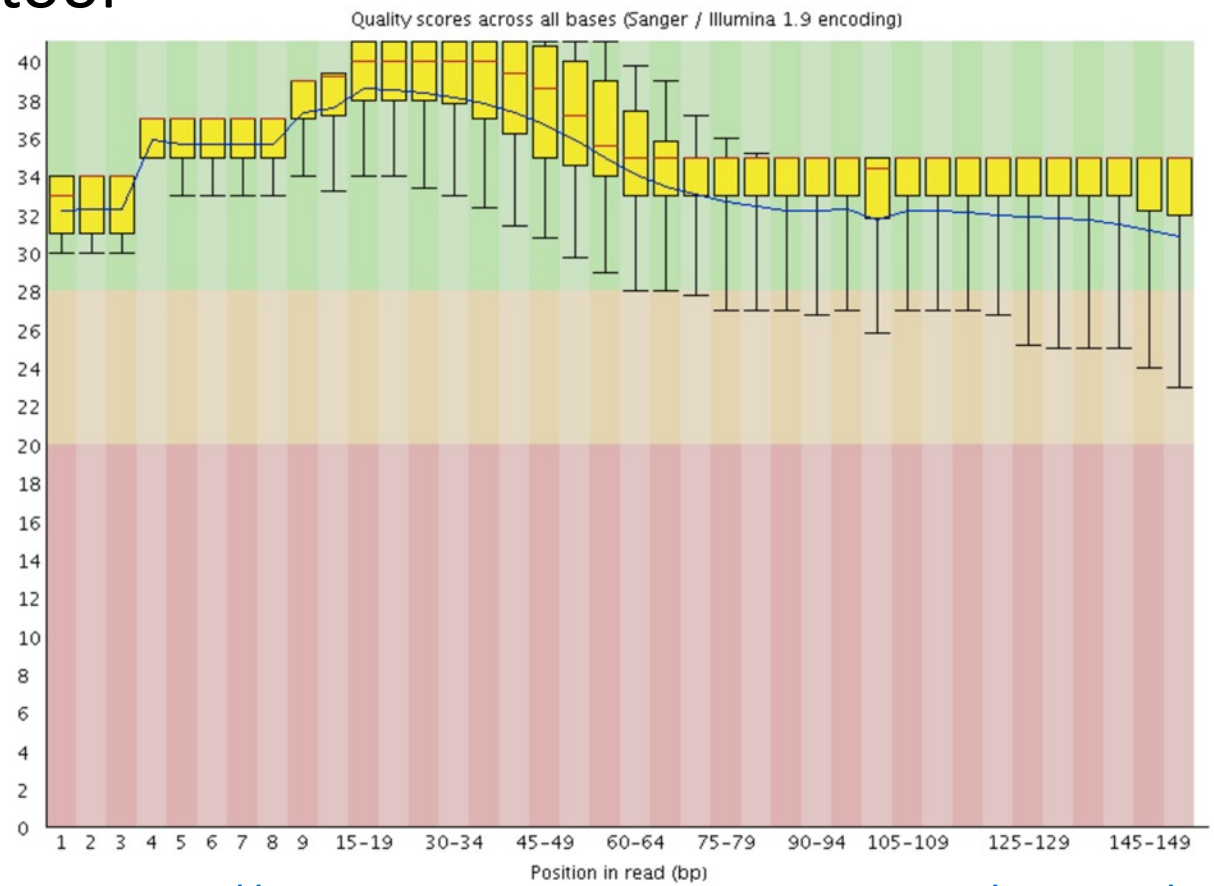
Quality Control

- QC typically follows this process:
  - Evaluate (visualize) data quality
  - Stop QC if the quality appears to be satisfactory
  - If not, execute one or more data altering steps then go to step 1

QC – Evaluation of Data Quality

- Developed by Babraham Institute, **FastQC** is the most commonly used quality control visualization tool

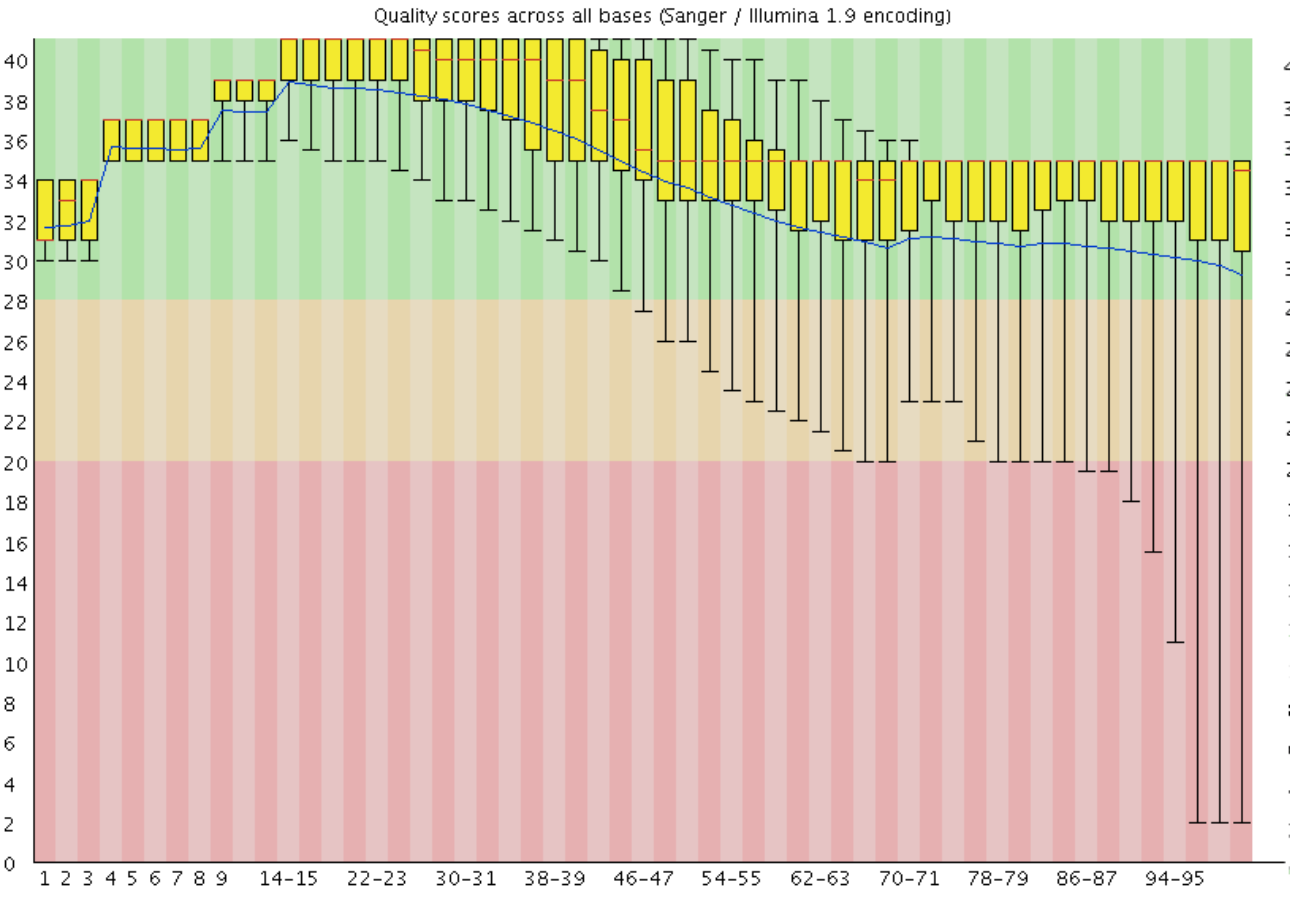


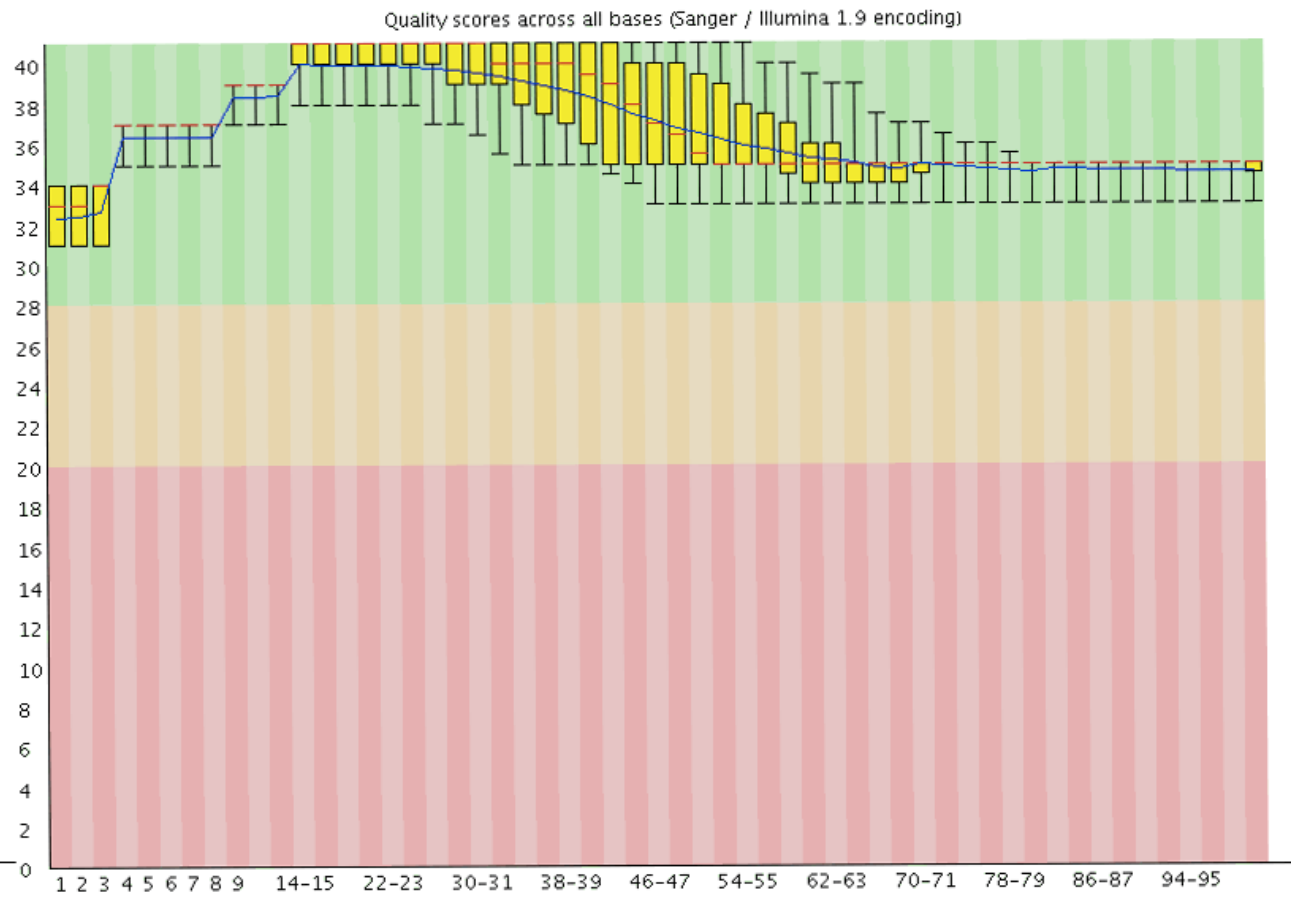


http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

QC – Quality Trimming

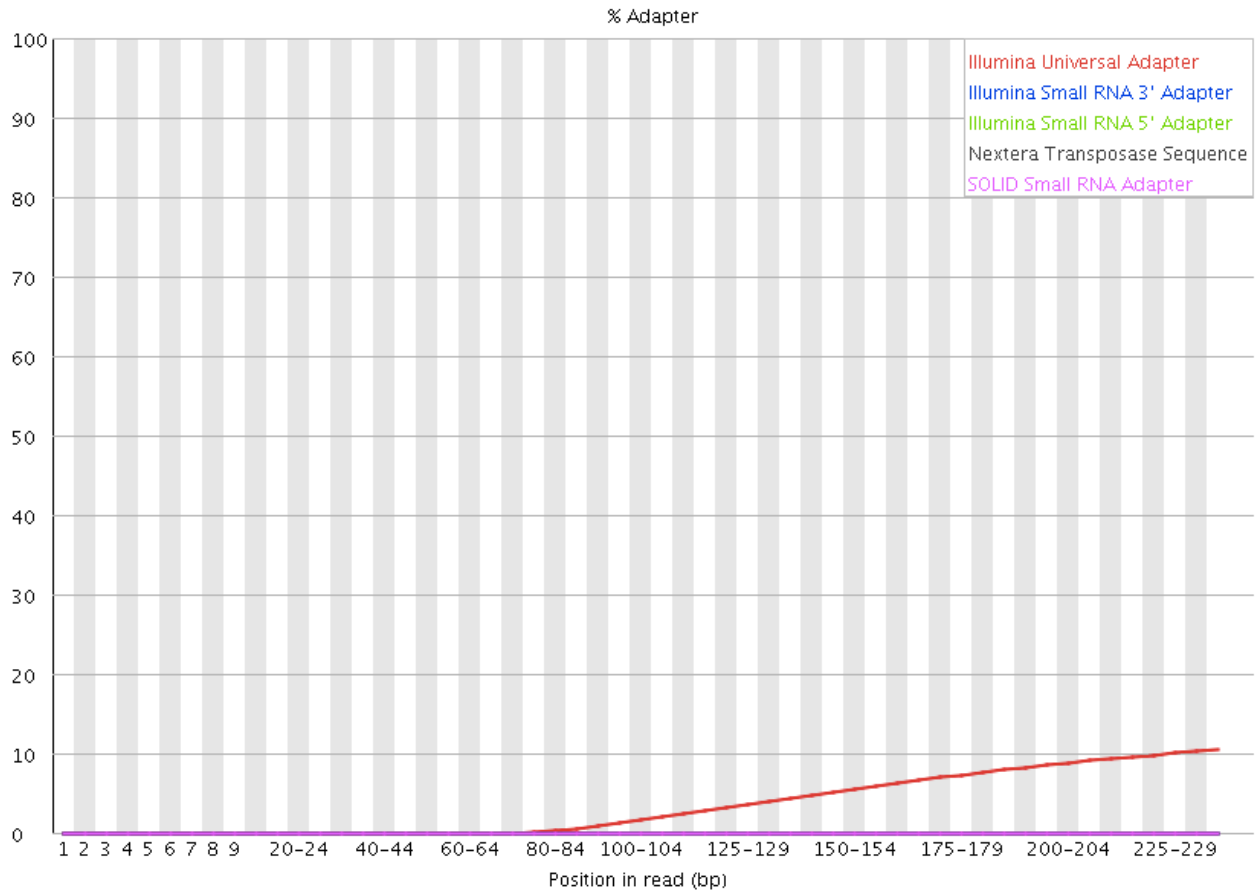
- Can be done with trimmomatic, BBDuk, Fastx Toolkit …

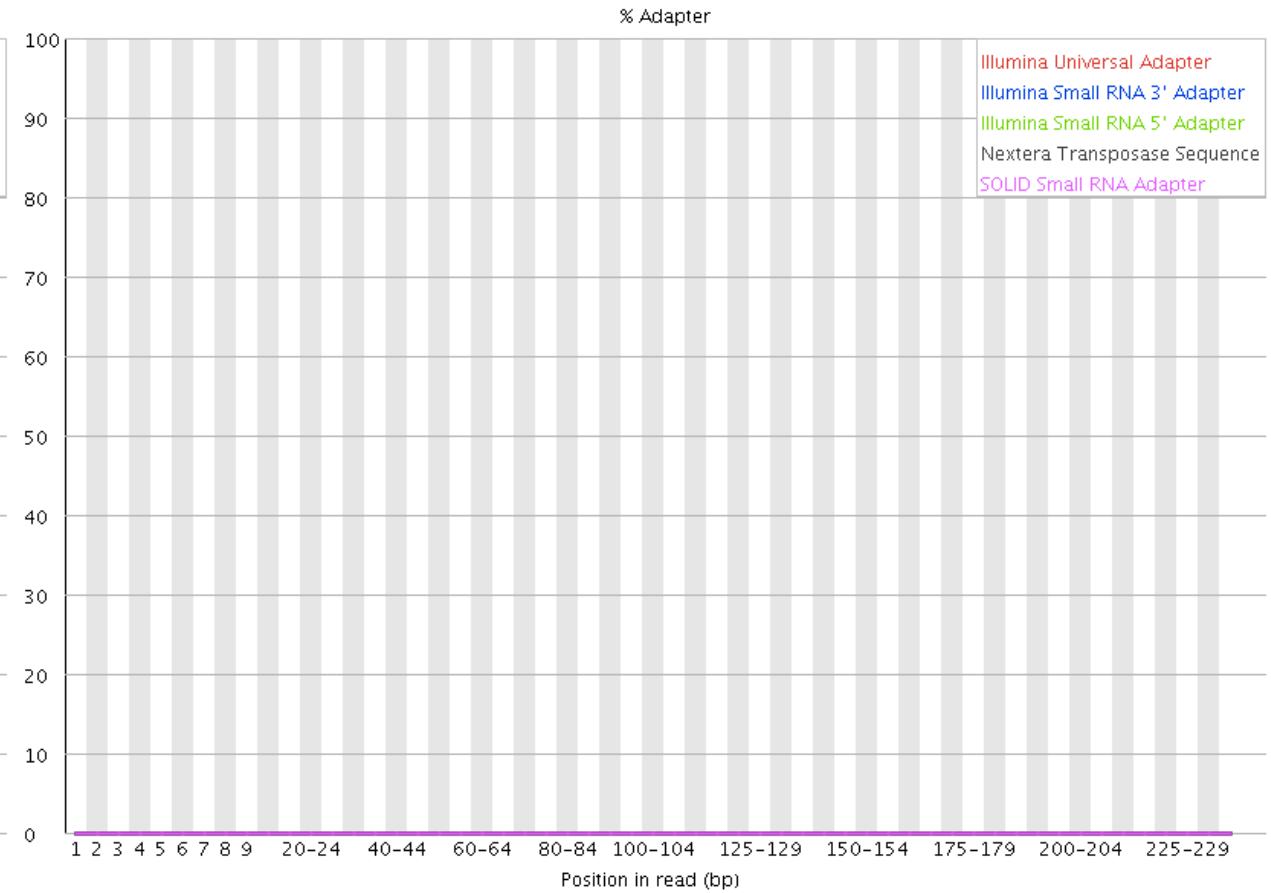


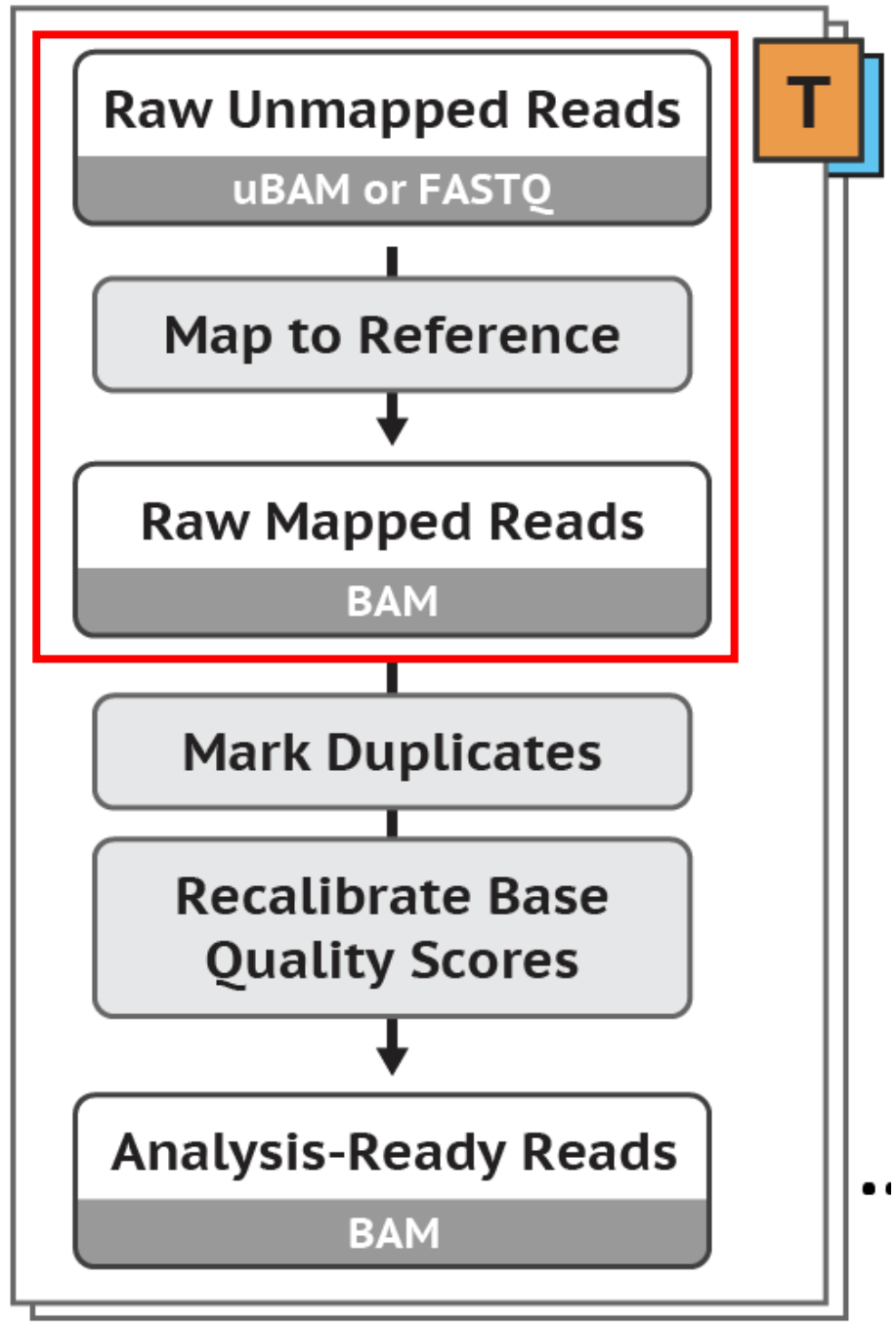


QC – Adapter Removal

- Adapter trimming can be done with trimmomatic

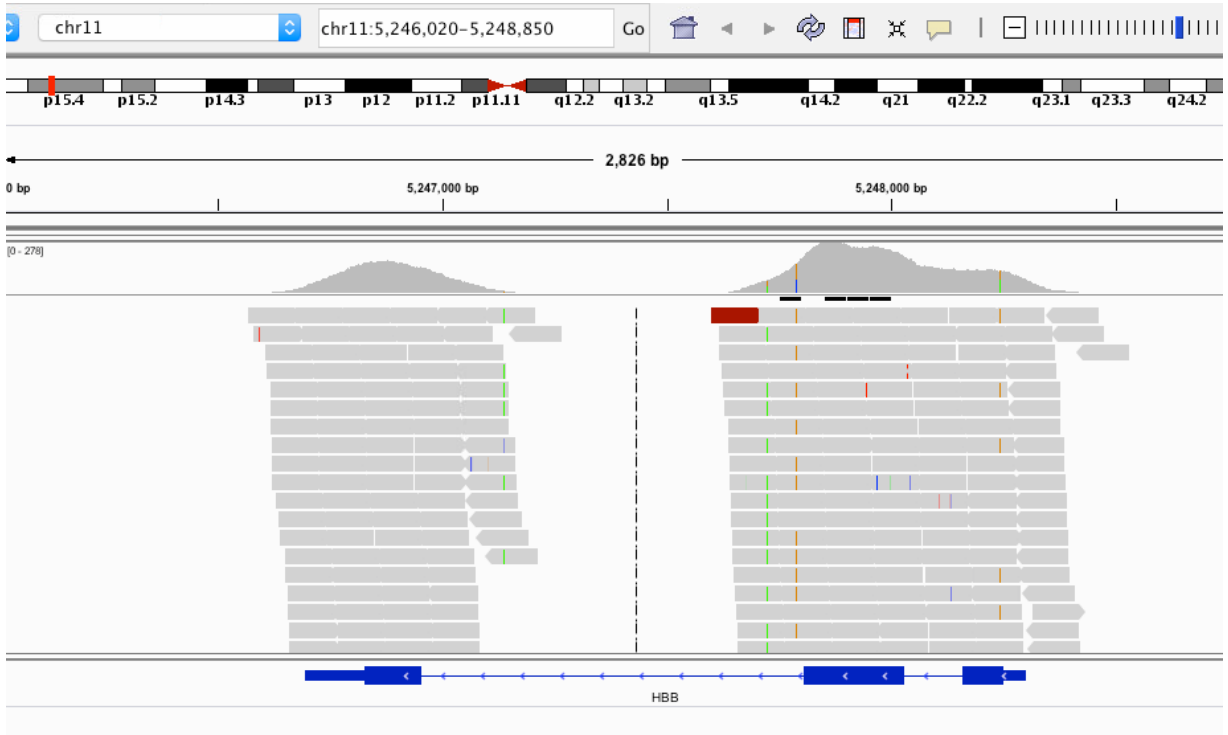





Raw Unmapped Reads
uBAM or FASTQ
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
T

Mapping

- Tumor/Normal reads are aligned to the reference human genome sequence separately with **BWA-MEM algorithm** (Heng Li & Richard Durbin)
  - Specifying read group information
- Produces SAM file

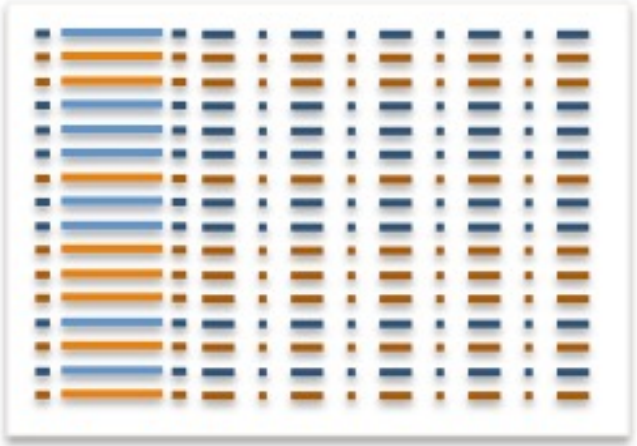


IGV screenshot of a germline sample

Preprocessing (Optional but recommended)

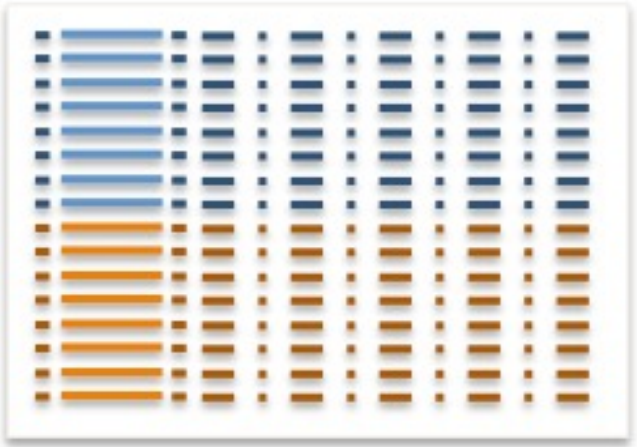
The information for this:

… is actually stored as a text file with one line per read which from far away looks like this:



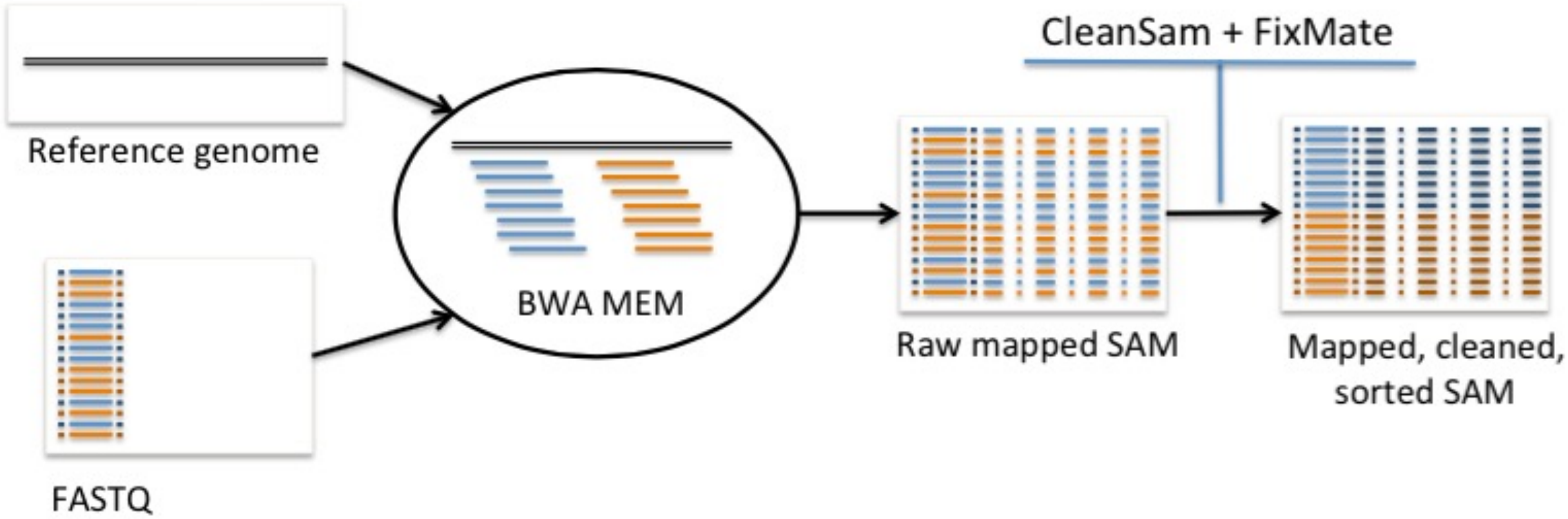
The reads are in no particular order…

… but the GATK wants reads to be sorted by starting position like this:



So we need to explicitly sort the SAM file…

Preprocessing (Optional but recommended)

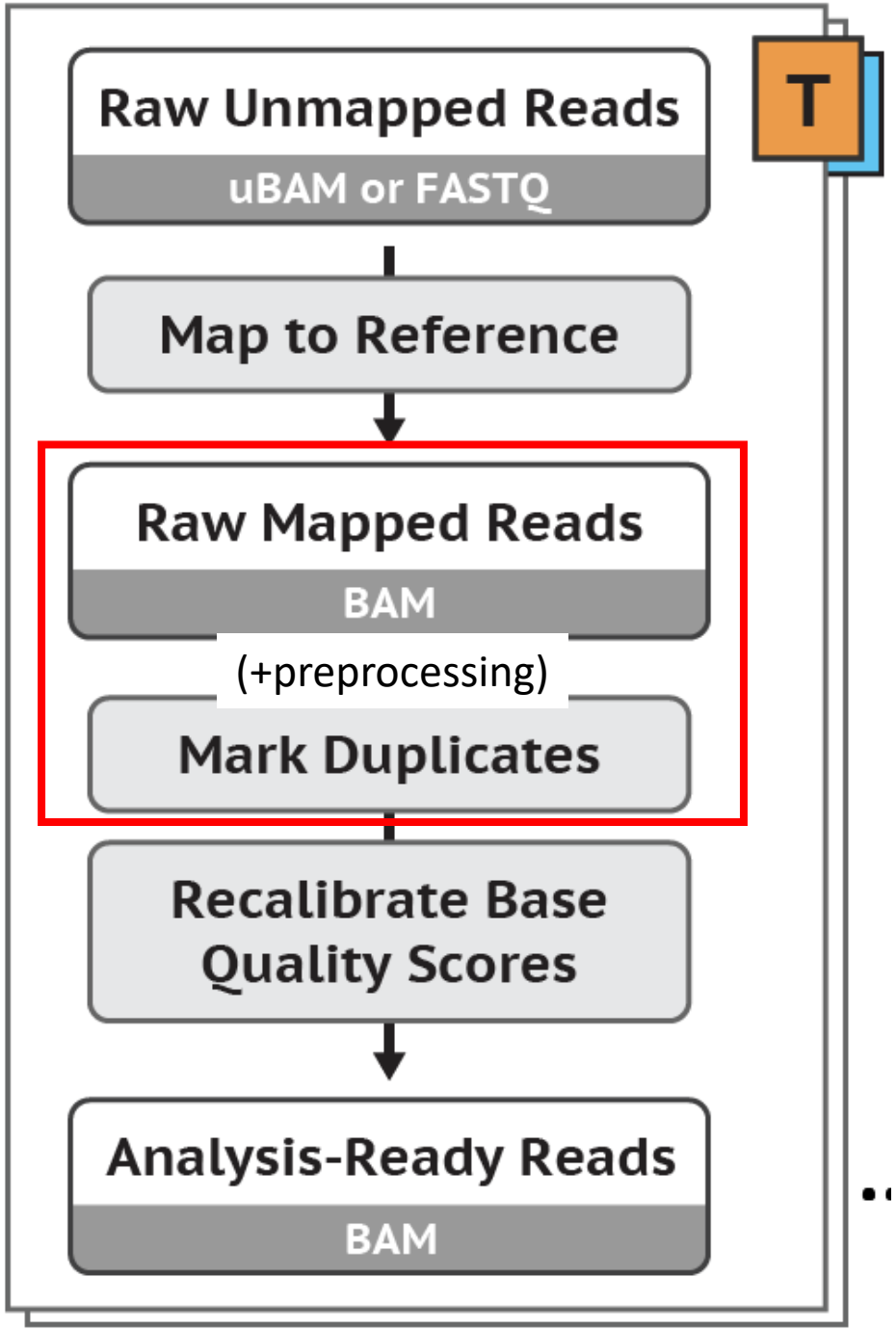


Preprocessing (Optional but recommended)

- For each aligned reads from T/N samples:
  1. SAM is cleaned using **Picard - CleanSam**

     Cleans the provided SAM, soft-clipping beyond-end-of-reference alignments and setting MAPQ to 0 for unmapped reads
  2. SAM is converted to BAM using **Picard - Sort and Index**

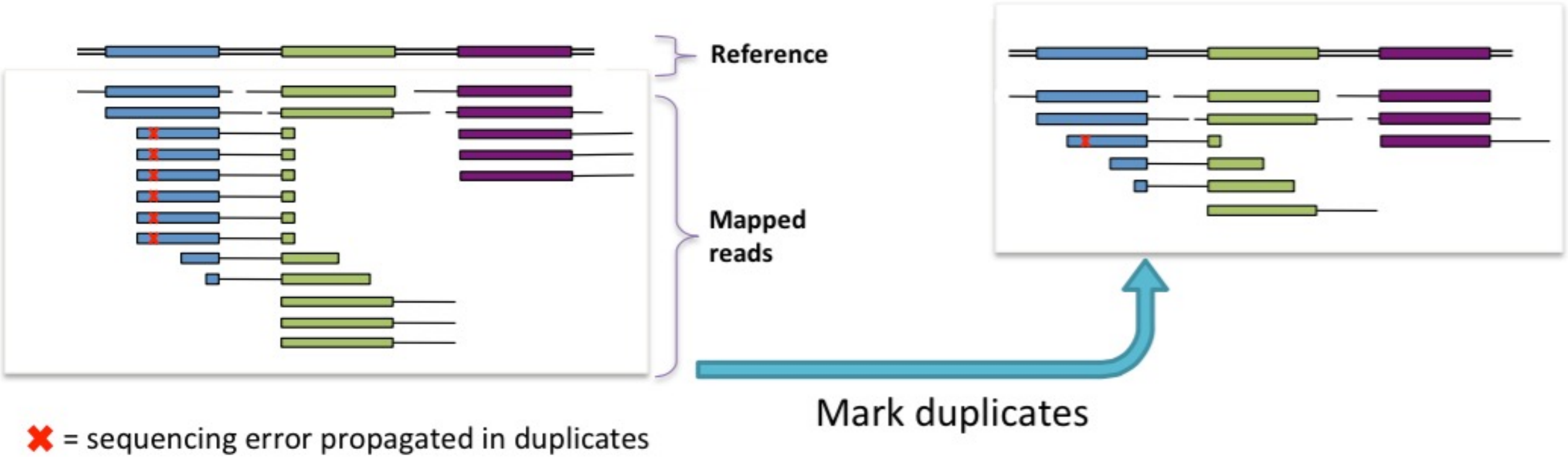
     Sorted by coordinate
  3. Mate information is fixed using **Picard – FixMateInformation**

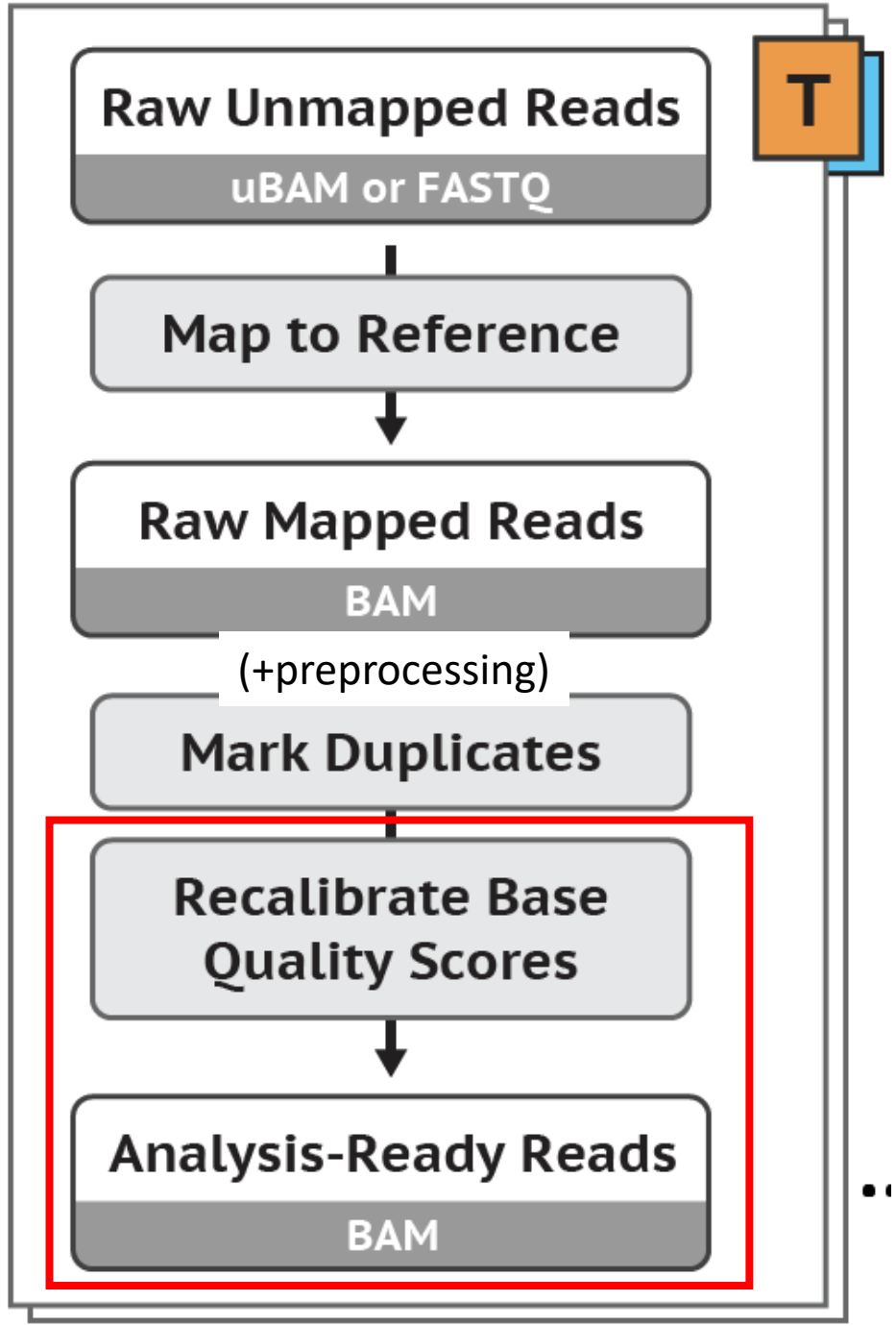

Raw Unmapped Reads
uBAM or FASTQ
T
Map to Reference
Raw Mapped Reads
BAM
(+preprocessing)
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM

Mark Duplicates

- Duplicates are sets of reads pairs that have the **same alignment start and end**
- They're suspected to be non-independent measurements of a sequence
  - Sampled from the **exact same template of DNA**
  - **Violates assumptions of variant calling**
- What's more, **errors in sample/library prep will get propagated to all the duplicates**
  - Just pick the "best" copy – mitigates the effects of errors

Mark Duplicates

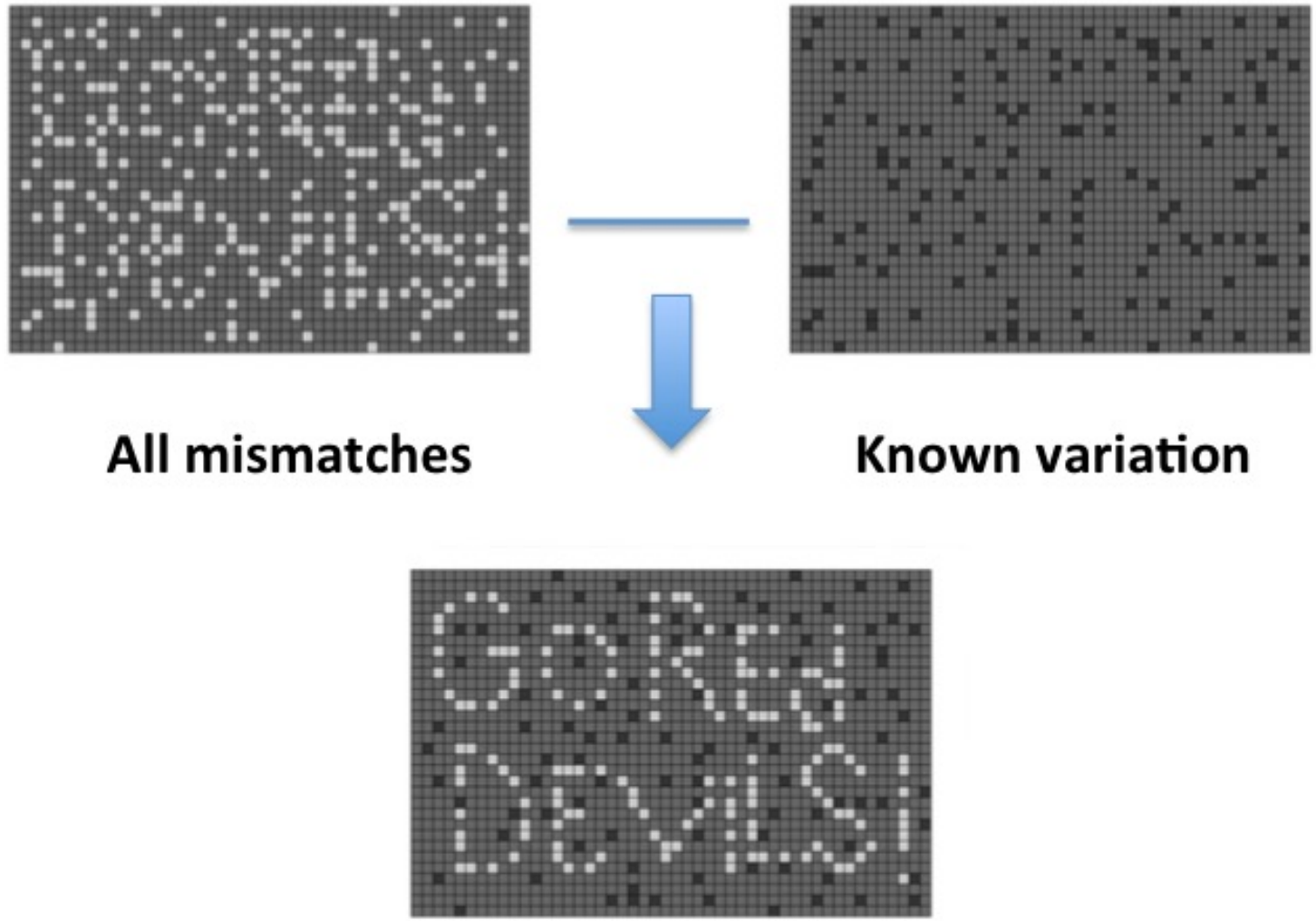


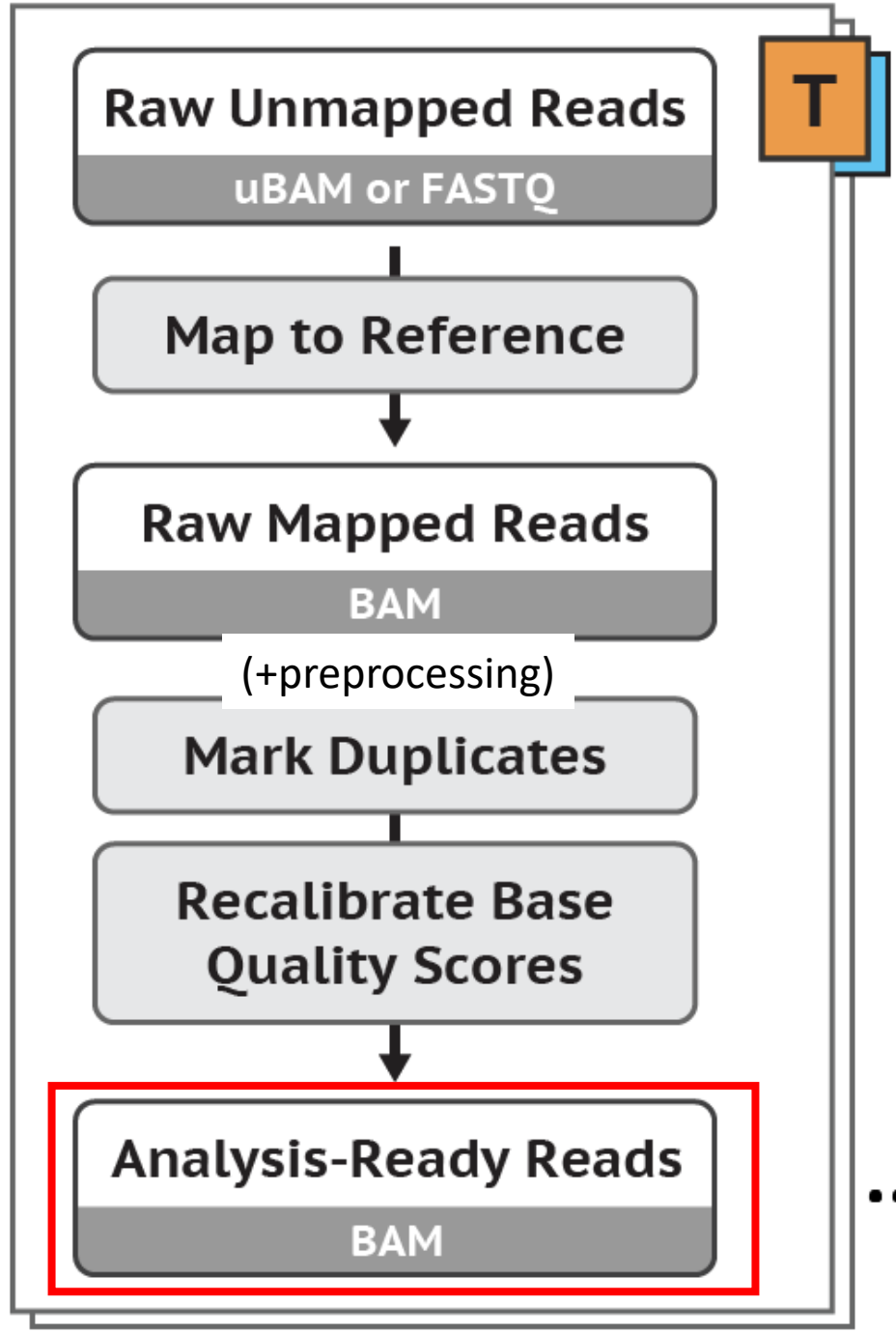

Raw Unmapped Reads
uBAM or FASTQ
T
Map to Reference
Raw Mapped Reads
BAM
(+preprocessing)
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM

Base Quality Score Recalibration

- Quality scores are **critical for all downstream analyses**
- Quality scores emitted by sequencing machines are **biased** and **inaccurate**
- Systematic biases are a major contributor to bad calls
- Base Quality Score Recalibration provides a **calibrated error model from which to make mutation calls**

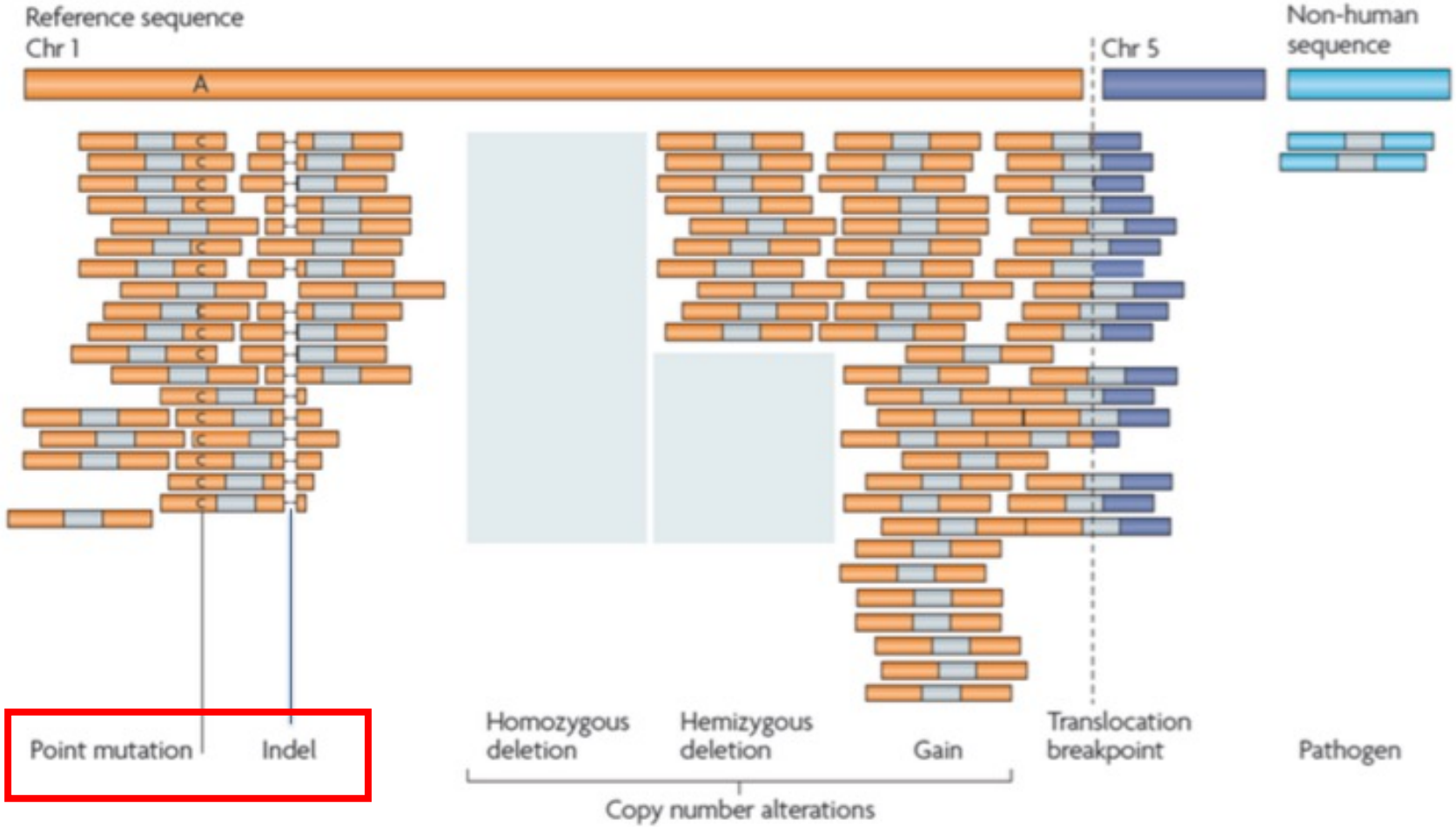
- Goal is to identify the signal within the noise



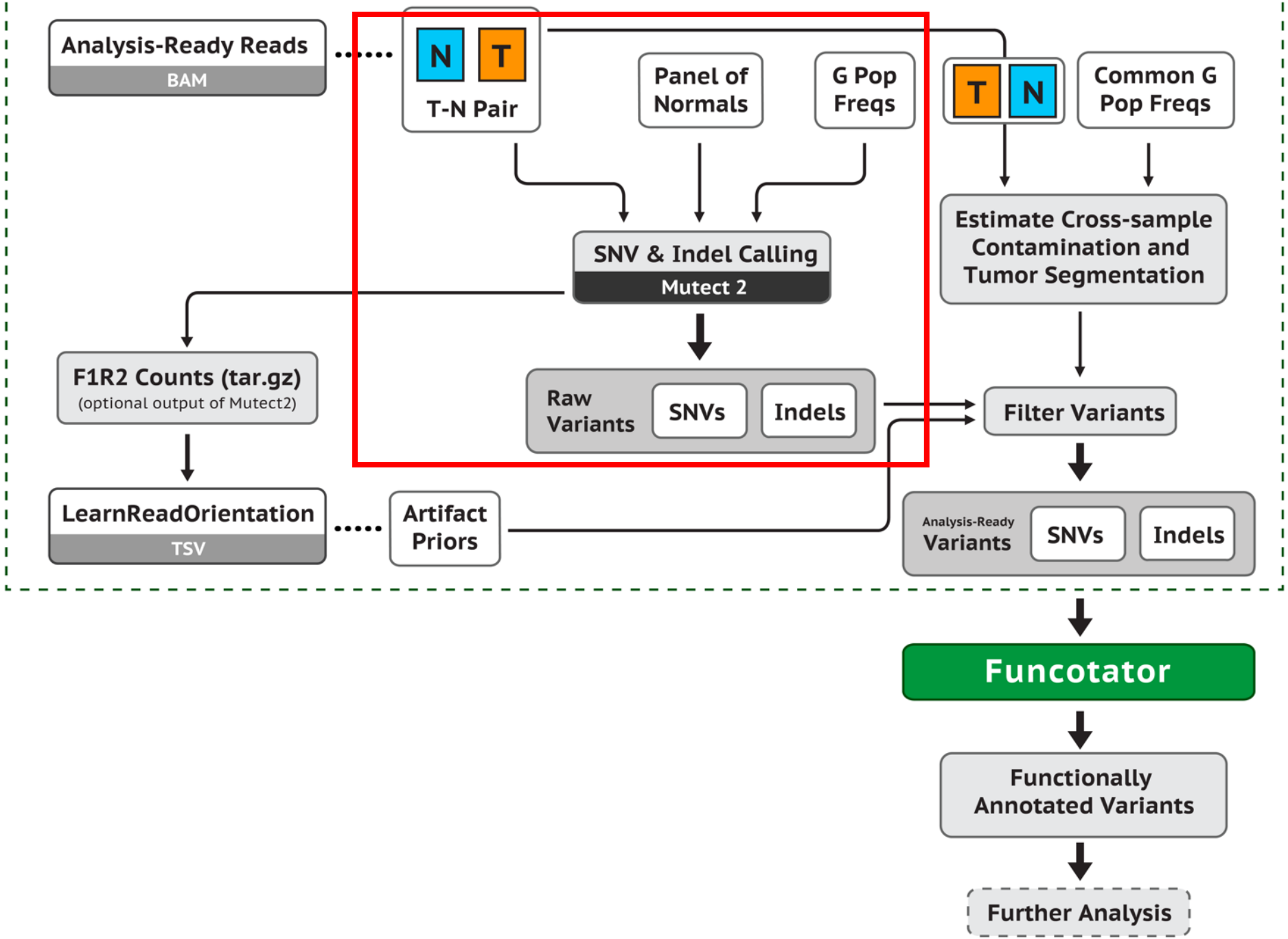

Raw Unmapped Reads
uBAM or FASTQ
T
Map to Reference
Raw Mapped Reads
BAM
(+preprocessing)
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM

Variant Discovery

Types of Somatic Variants

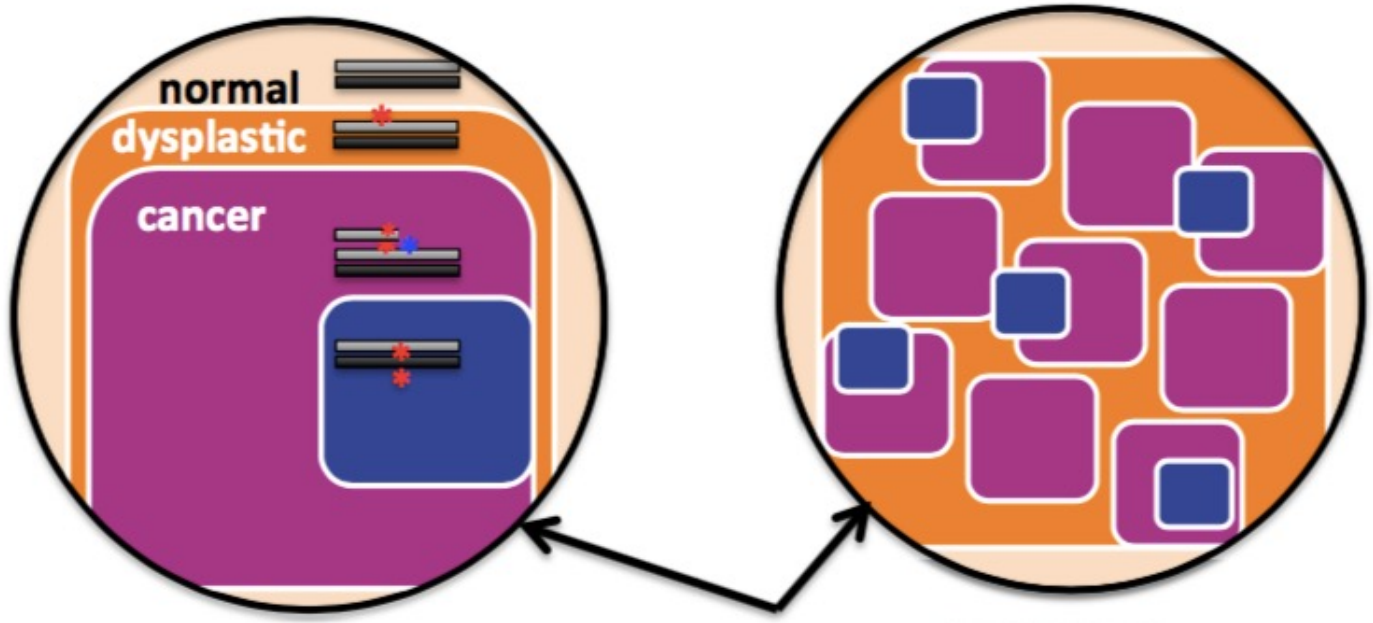


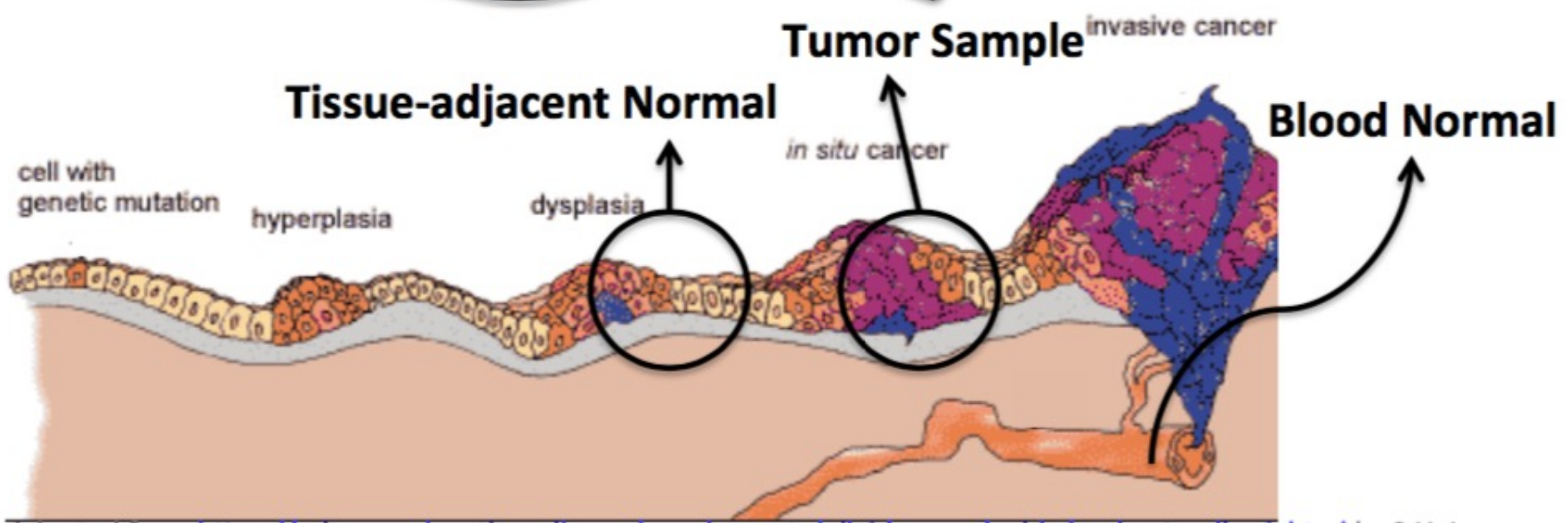
Meyerson, Gabriel and Getz. Nat. Rev. Genet. (2010)


Analysis-Ready Reads
BAM
N
T
T-N Pair
Panel of Normals
G Pop Freqs
T
N
Common G Pop Freqs
Estimate Cross-sample Contamination and Tumor Segmentation
SNV & Indel Calling
Mutect 2
F1R2 Counts (tar.gz)
(optional output of Mutect2)
Raw Variants
SNVs
Indels
Filter Variants
LearnReadOrientation
TSV
Artifact Priors
Analysis-Ready Variants
SNVs
Indels
Funcotator
Functionally Annotated Variants
Further Analysis

Tumor **purity** = $\frac{\text{(tumor cells)}}{\text{(normal + tumor cells)}}$

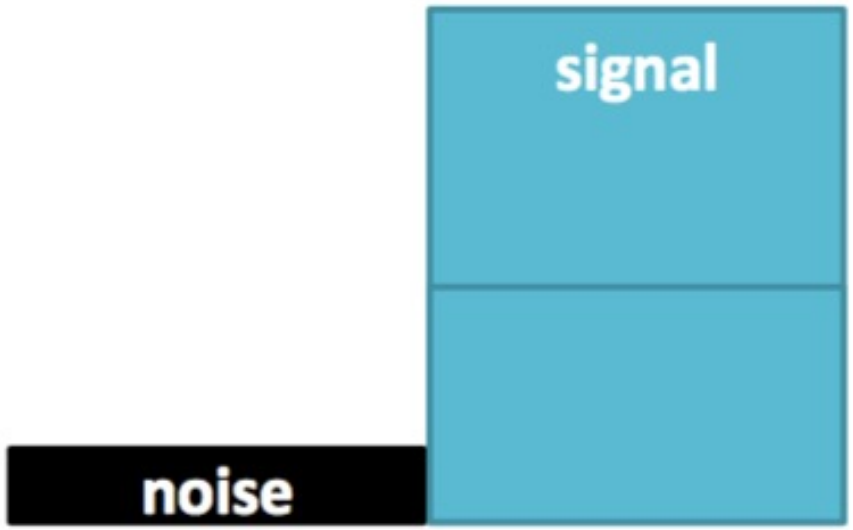
Tumor **heterogeneity** is based on polygenomic populations, segregated or intermixed, due to ongoing subclonal evolution.



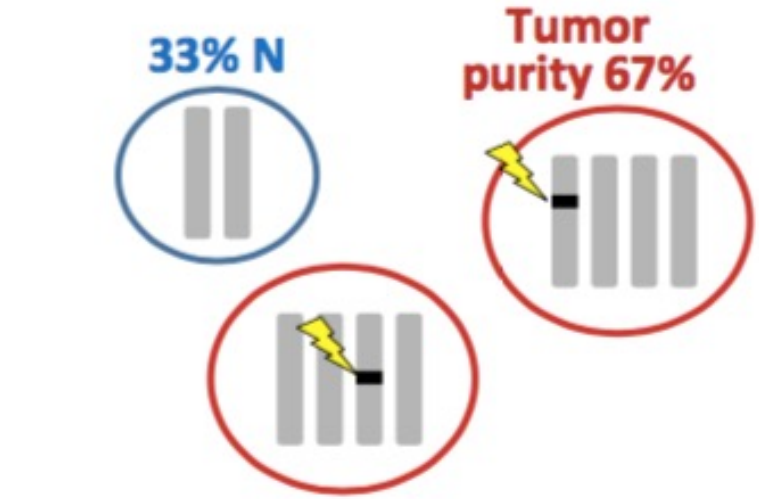


Adopted from https://science.education.nih.gov/supplements/nih1/cancer/guide/understanding1.*html* by S.H. Lee

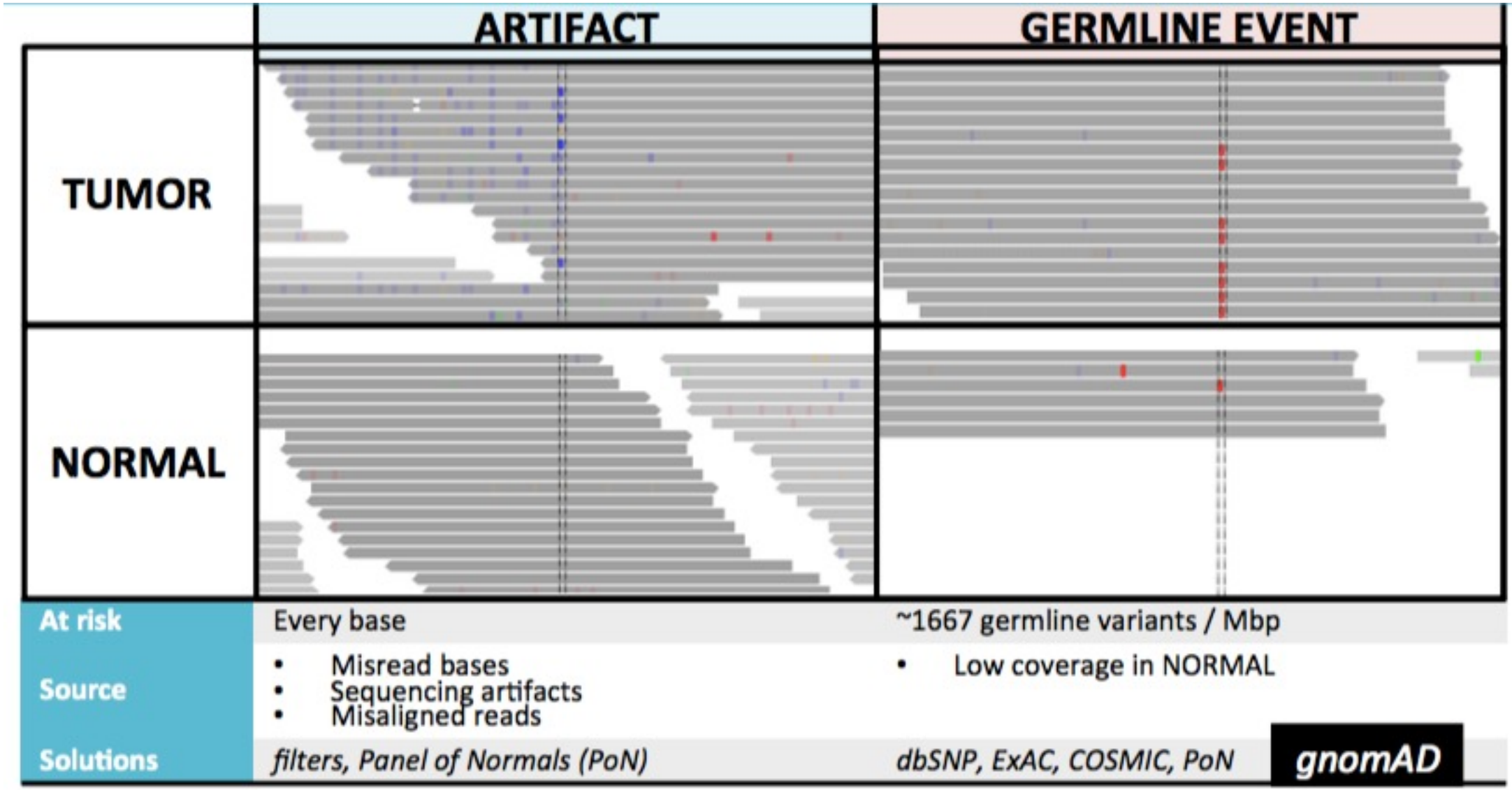
Expectation for germline variants



Expectation for somatic variants




Germline SNP
Somatic SNV?
NORMAL
TUMOR
L G R G E P S I P S N R I K T D G A
CNTN1


ARTIFACT
GERMLINE EVENT
TUMOR
NORMAL
At risk
Every base
~1667 germline variants / Mbp
Source
• Misread bases
• Sequencing artifacts
• Misaligned reads
• Low coverage in NORMAL
Solutions
filters, Panel of Normals (PoN)
dbSNP, ExAC, COSMIC, PoN
gnomAD

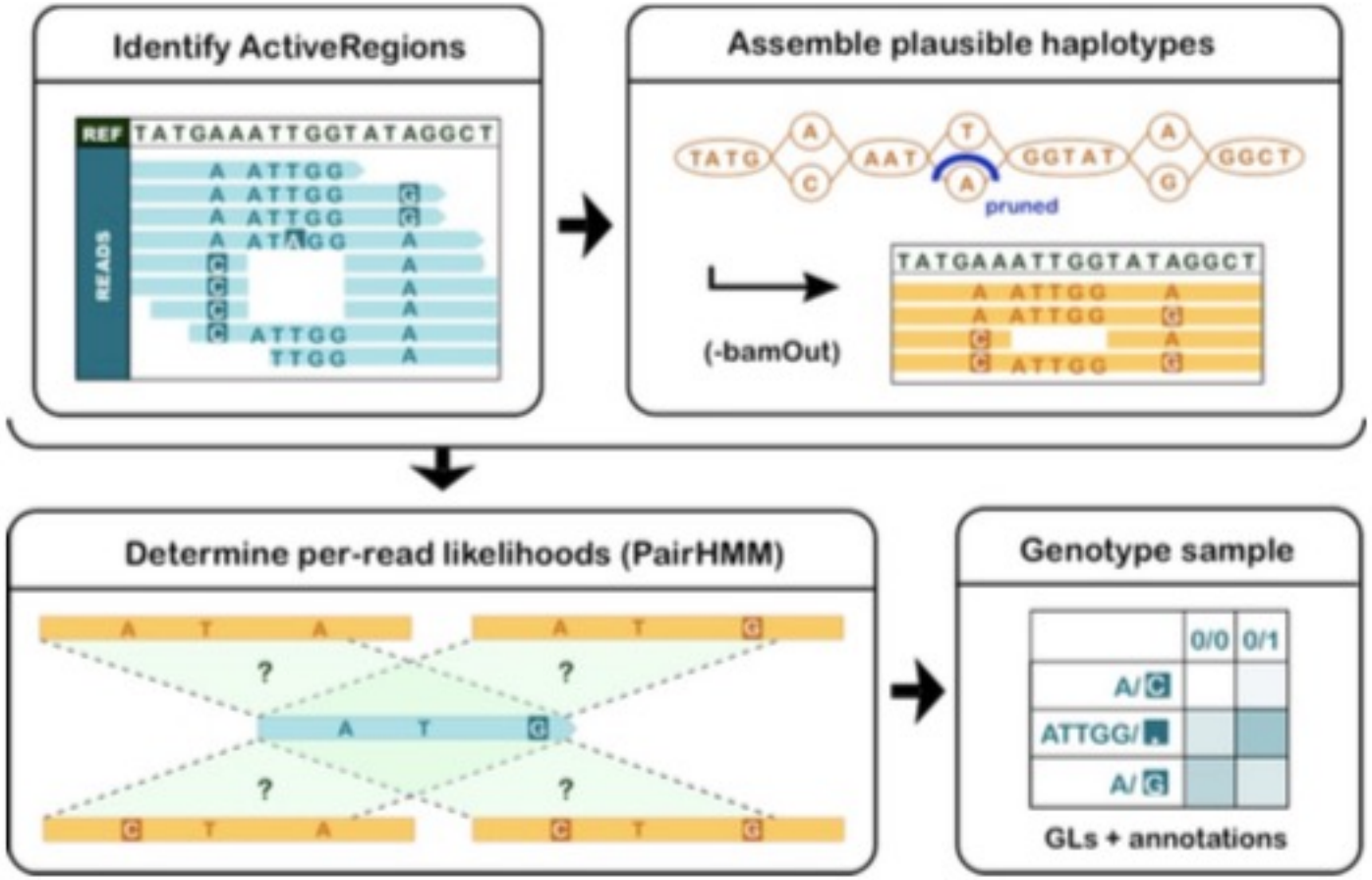
Mutect2 is based on HaplotypeCaller

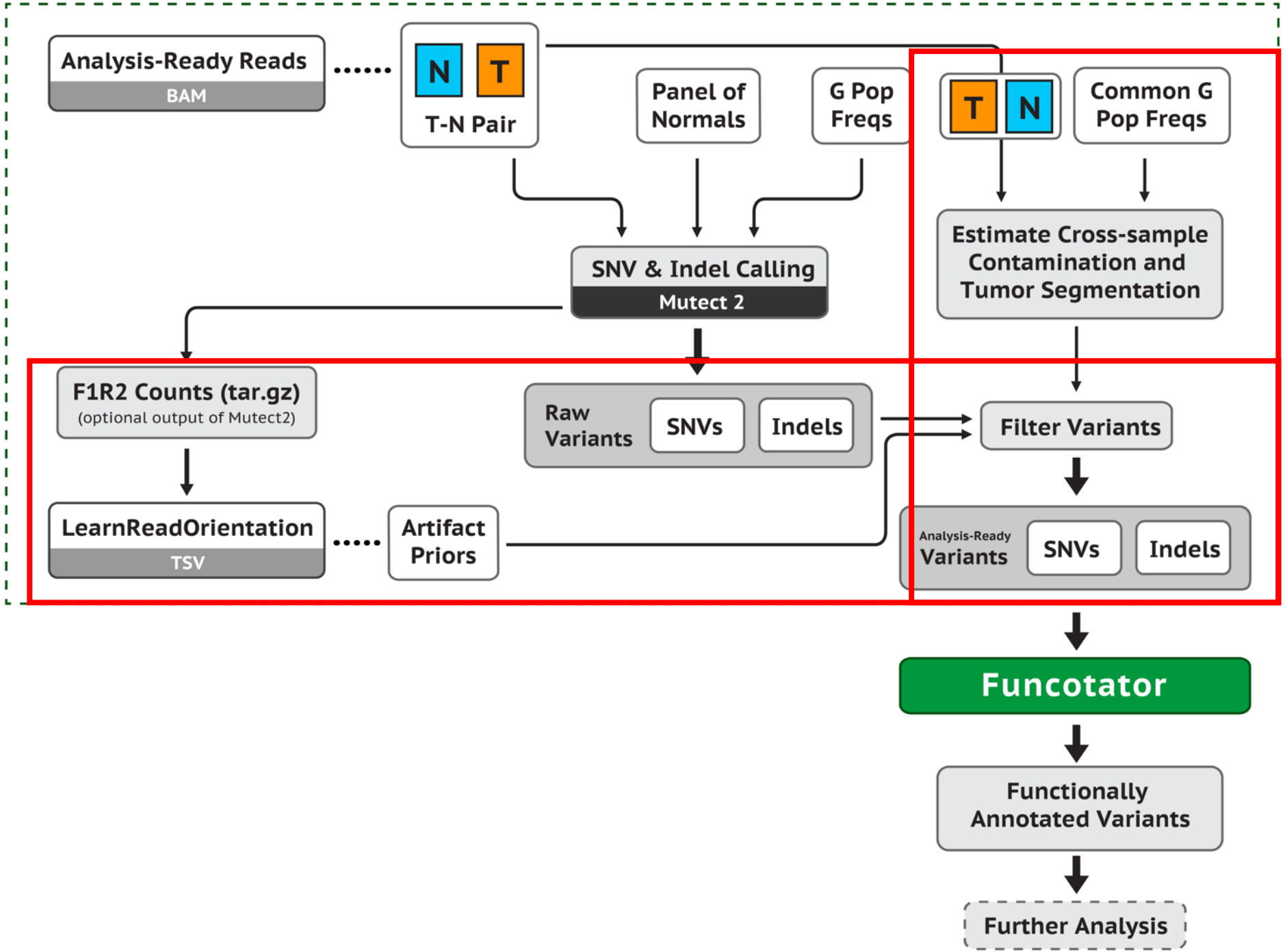
Skip :

- Sites in *PoN*
- Sites with high fraction alt alleles in normal

Allele-specific calling:

- Distinguishes alleles in *the germline population frequency* resource and uses AF in calculating probability variant exists in normal *and* tumor




Analysis-Ready Reads
BAM
N
T
T-N Pair
Panel of Normals
G Pop Freqs
T
N
Common G Pop Freqs
Estimate Cross-sample Contamination and Tumor Segmentation
SNV & Indel Calling
Mutect 2
F1R2 Counts (tar.gz)
(optional output of Mutect2)
Raw Variants
SNVs
Indels
Filter Variants
LearnReadOrientation
TSV
Artifact Priors
Analysis-Ready Variants
SNVs
Indels
Funcotator
Functionally Annotated Variants
Further Analysis

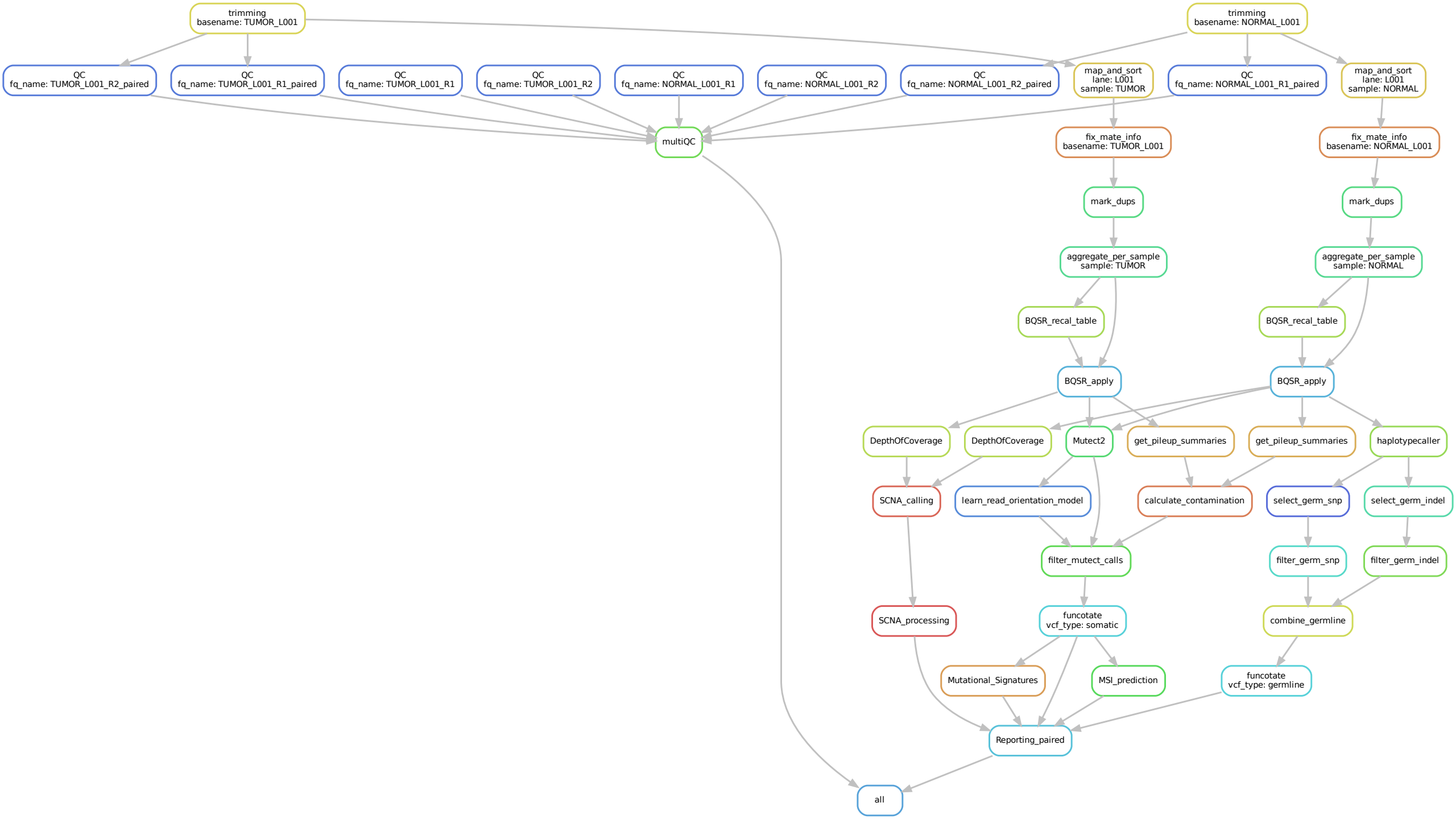
FilterMutectCalls

FILTER	Description
artifact_in_normal	artifact_in_normal
base_quality	alt median base quality
clustered_events	Clustered events observed in the tumor
contamination	contamination
duplicate_evidence	evidence for alt allele is overrepresented by apparent duplicates
fragment_length	abs(ref - alt) median fragment length
germline_risk	Evidence indicates this site is germline, not somatic
mapping_quality	ref - alt median mapping quality
multiallelic	Site filtered because too many alt alleles pass tumor LOD
orientation_bias	Orientation bias (in one of the specified artifact mode(s) or complement) seen in one or more samples.
panel_of_normals	Blacklisted site in panel of normals
read_position	median distance of alt variants from end of reads
str_contraction	Site filtered due to contraction of short tandem repeat region
strand_artifact	Evidence for alt allele comes from one read direction only
t_lod	Tumor does not meet likelihood threshold

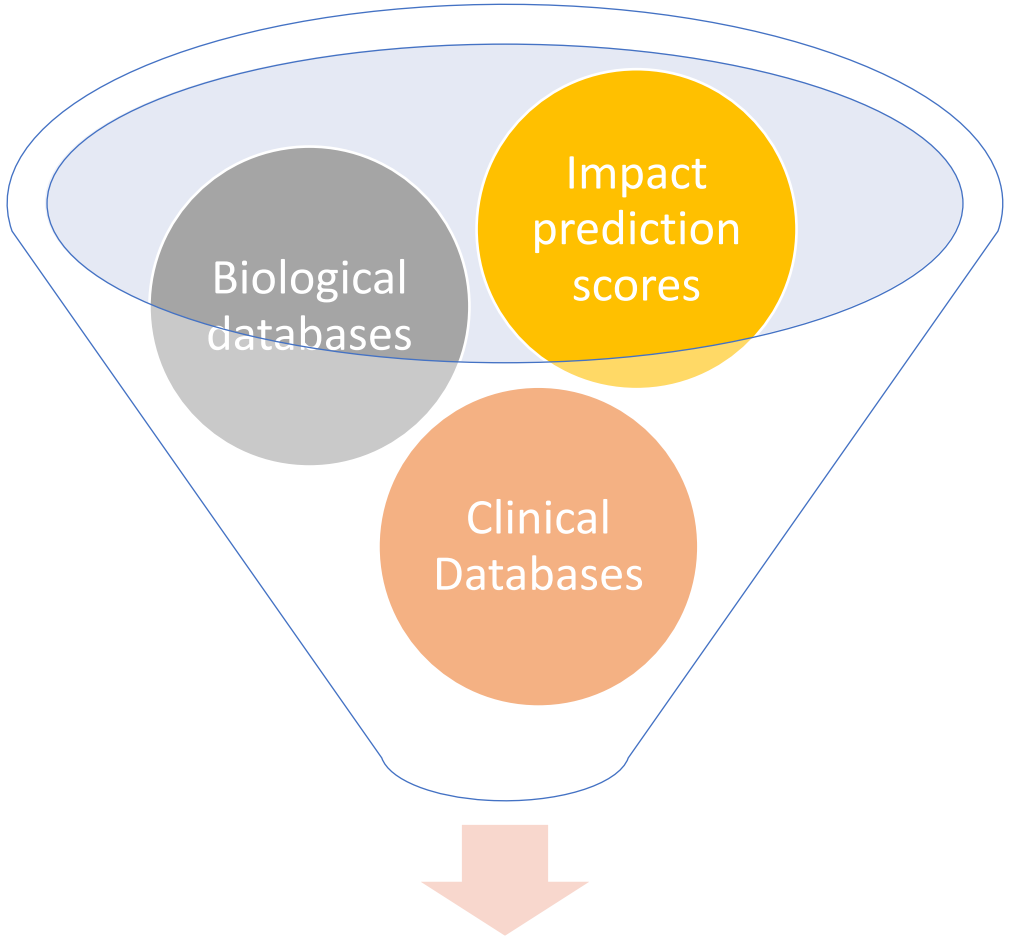
(How to) Call somatic mutations using GATK4 Mutect2

Regularly updated:

https://gatk.broadinstitute.org/hc/en-us/articles/360035531132

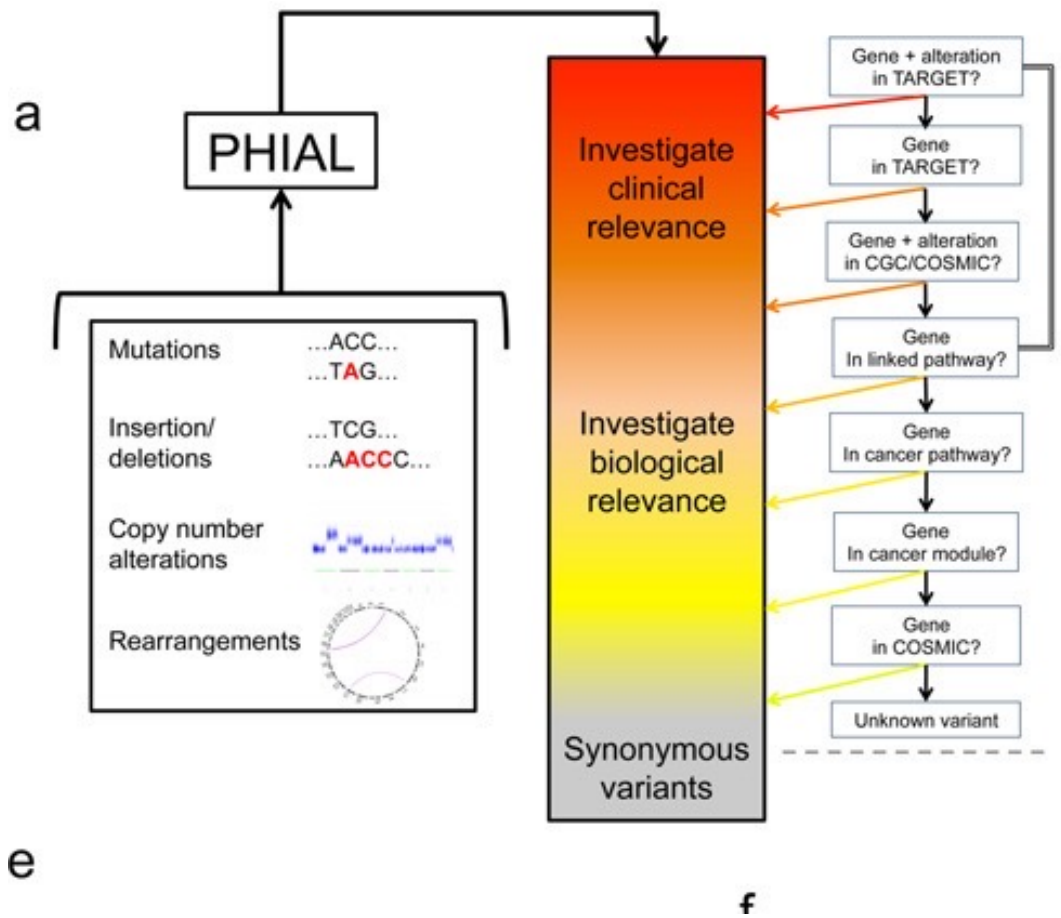

trimming
basename: TUMOR_L001
trimming
basename: NORMAL_L001
QC
fq_name: TUMOR_L001_R2_paired
QC
fq_name: TUMOR_L001_R1_paired
QC
fq_name: TUMOR_L001_R1
QC
fq_name: TUMOR_L001_R2
QC
fq_name: NORMAL_L001_R1
QC
fq_name: NORMAL_L001_R2
QC
fq_name: NORMAL_L001_R2_paired
map_and_sort
lane: L001
sample: TUMOR
QC
fq_name: NORMAL_L001_R1_paired
map_and_sort
lane: L001
sample: NORMAL
multiQC
fix_mate_info
basename: TUMOR_L001
fix_mate_info
basename: NORMAL_L001
mark_dups
mark_dups
aggregate_per_sample
sample: TUMOR
aggregate_per_sample
sample: NORMAL
BQSR_recal_table
BQSR_recal_table
BQSR_apply
BQSR_apply
DepthOfCoverage
DepthOfCoverage
Mutect2
get_pileup_summaries
get_pileup_summaries
haplotypecaller
SCNA_calling
learn_read_orientation_model
calculate_contamination
select_germ_snp
select_germ_indel
filter_mutect_calls
filter_germ_snp
filter_germ_indel
SCNA_processing
funcotate
vcf_type: somatic
combine_germline
Mutational_Signatures
MSI_prediction
funcotate
vcf_type: germline
Reporting_paired
all

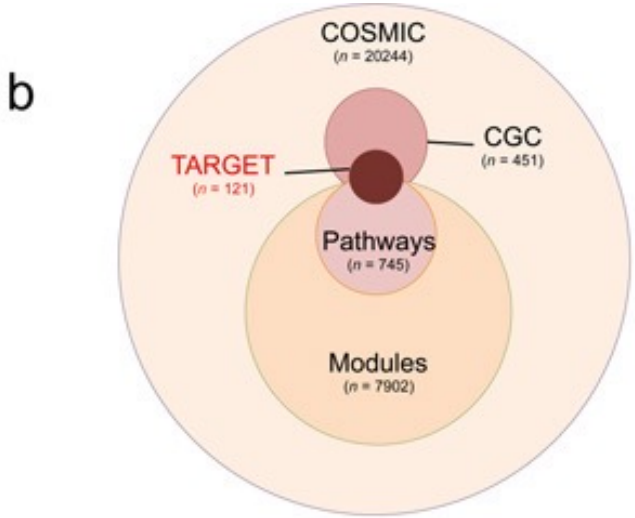
Finding driver variants is not easy

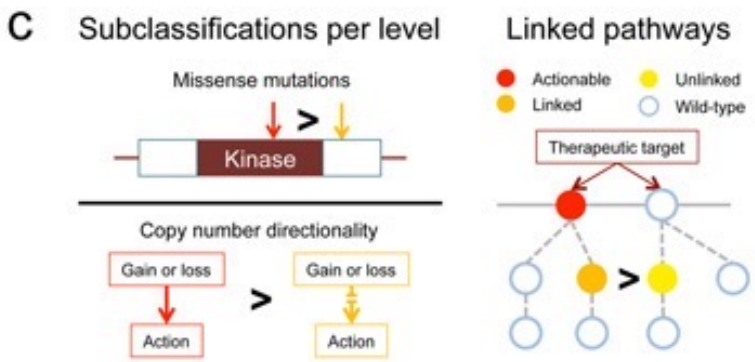


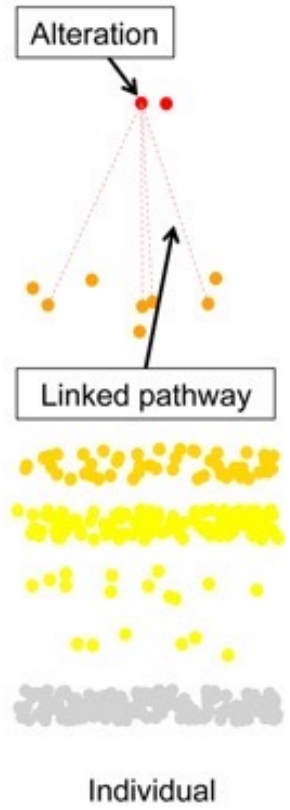
Prioritized variants

PHIAL



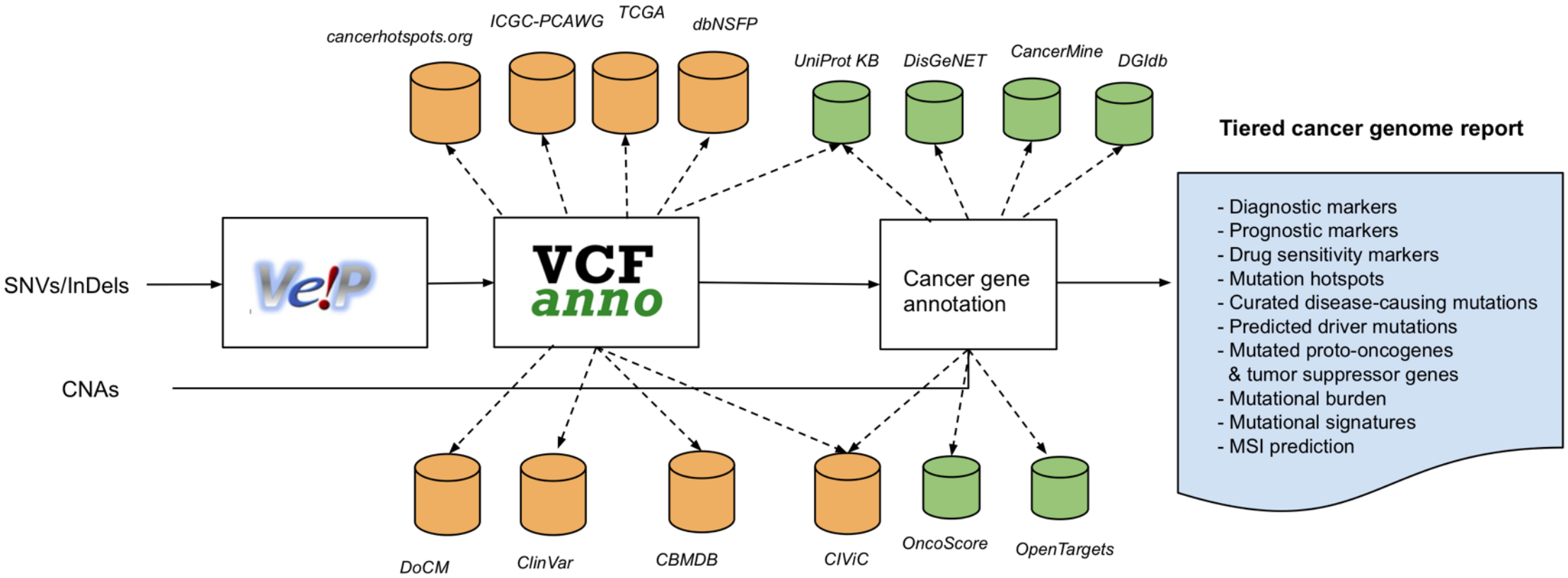






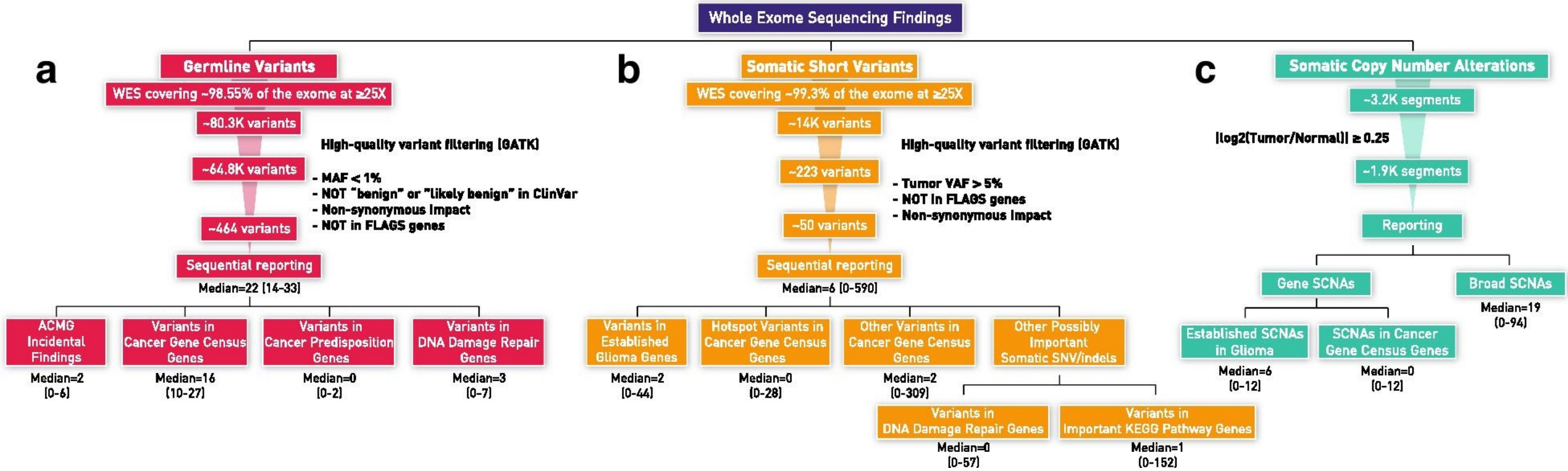
Van allen EM, Wagle N, Stojanov P, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med. 2014;20(6):682-8.

PCGR



Nakken S, Fournous G, Vodák D, Aasheim LB, Myklebost O, Hovig E. Personal Cancer Genome Reporter: variant interpretation report for precision oncology. Bioinformatics. 2018;34(10):1778-1780.

NOTATES



Ülgen E, Can Ö, Bilguvar K, Akyerli Boylu C, Kılıçturgay Yüksel Ş, Erşen Danyeli A, et al. Sequential filtering for clinically relevant variants as a method for clinical interpretation of whole exome sequencing findings in glioma. BMC Med Genomics. 2021 Feb 23;14(1):54.

NOTATES

Whole Exome Sequencing Summary Report

Report date: 29 December, 2020
Patient ID: NOT-XXXX

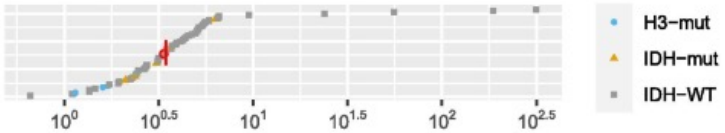
Indication for testing: Diffuse glioma clinical WES analysis
Treatment status: Primary (untreated) tumor

Tumor Sample: Formalin-Fixed Paraffin-Embedded (FFPE) tissue specimen
Normal Sample: Peripheral venous blood
DNA extraction method: QIAGEN DNeasy Blood & Tissue kit

Whole exome sequencing of this individual's tumor and normal samples were performed and covered 99.7% (tumor) and 99.6% (normal) of all exonic positions at 25X or more. Selected somatic findings are presented below.

I. Tumor Mutational Burden

TMB is defined as the number of somatic mutations per megabase. TMB is a predictive biomarker being studied to evaluate its association with response to immunotherapy.



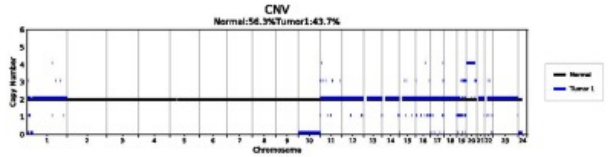
TMB for the current tumor sample is: **3.36 mutations/Mb** (Normal (<= 20 mutations/Mb)). The predicted MSI status of this tumor was **MSS (microsatellite stable)**.

II. Somatic Short Variants

Gene	Classification	Protein Change	Genome Change	VAF	Category
PTEN	Missense_Mutation	p.D24H	g.chr10:87864539G>C	0.716	Established gene
NF1	Missense_Mutation	p.A2623P	g.chr17:31357329G>C	0.500	Established gene
PTPN11	Missense_Mutation	p.T507K	g.chr12:112489096C>A	0.367	Hotspot in CGC
TGFBR2	Missense_Mutation	p.I170V	g.chr3:30671691A>G	0.412	CGC gene
PLCG1	Missense_Mutation	p.I459S	g.chr20:41165091T>G	0.234	CGC gene

(CGC: Cancer Gene Census, DDR: DNA-damage repair)

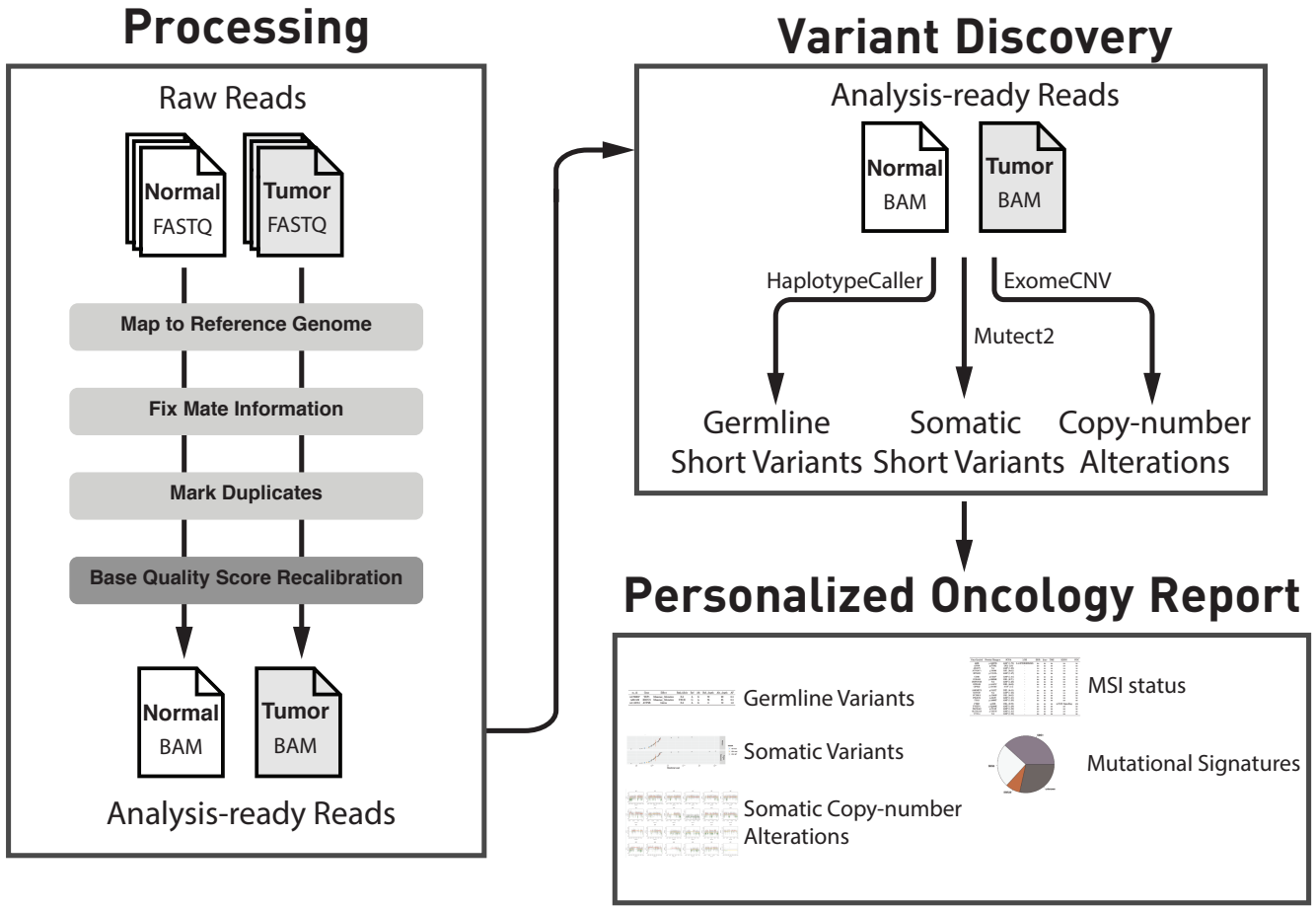
III. Somatic Copy Number Alterations



Gene	Segment	ratio	CN	av_cov	Category
CDK6	chr7:44138921-130733861	1.4097116	3	205.0379	Established SCNA
CDKN2A	chr9:20620654-22451911	0.1458543	0	249.2528	Established SCNA
CDKN2B	chr9:20620654-22451911	0.1458543	0	249.2528	Established SCNA
EGFR	chr7:44138921-130733861	1.4097116	3	205.0379	Established SCNA
EZH2	chr7:131142913-151087072	1.4323658	3	209.1155	Established SCNA
MET	chr7:44138921-130733861	1.4097116	3	205.0379	Established SCNA
BRAF	chr7:131142913-151087072	1.4323658	3	209.1155	Established SCNA

Ülgen E, Can Ö, Bilguvar K, Akyerli Boylu C, Kılıçturgay Yüksel Ş, Erşen Danyeli A, et al. Sequential filtering for clinically relevant variants as a method for clinical interpretation of whole exome sequencing findings in glioma. BMC Med Genomics. 2021 Feb 23;14(1):54.

EpiCANCER



- Raw reads to personalized report workflow
- Results are clinically prioritized