# BIOINFORMATICS ANALYSIS OF WHOLE EXOME SEQUENCING DATA

## Authors

Peter J. Ulintz, Weisheng Wu, and Chris M. Gates

## Presented by

Namuswe
Magdalene

Semawule
Syrus

Kirabo
Gloria

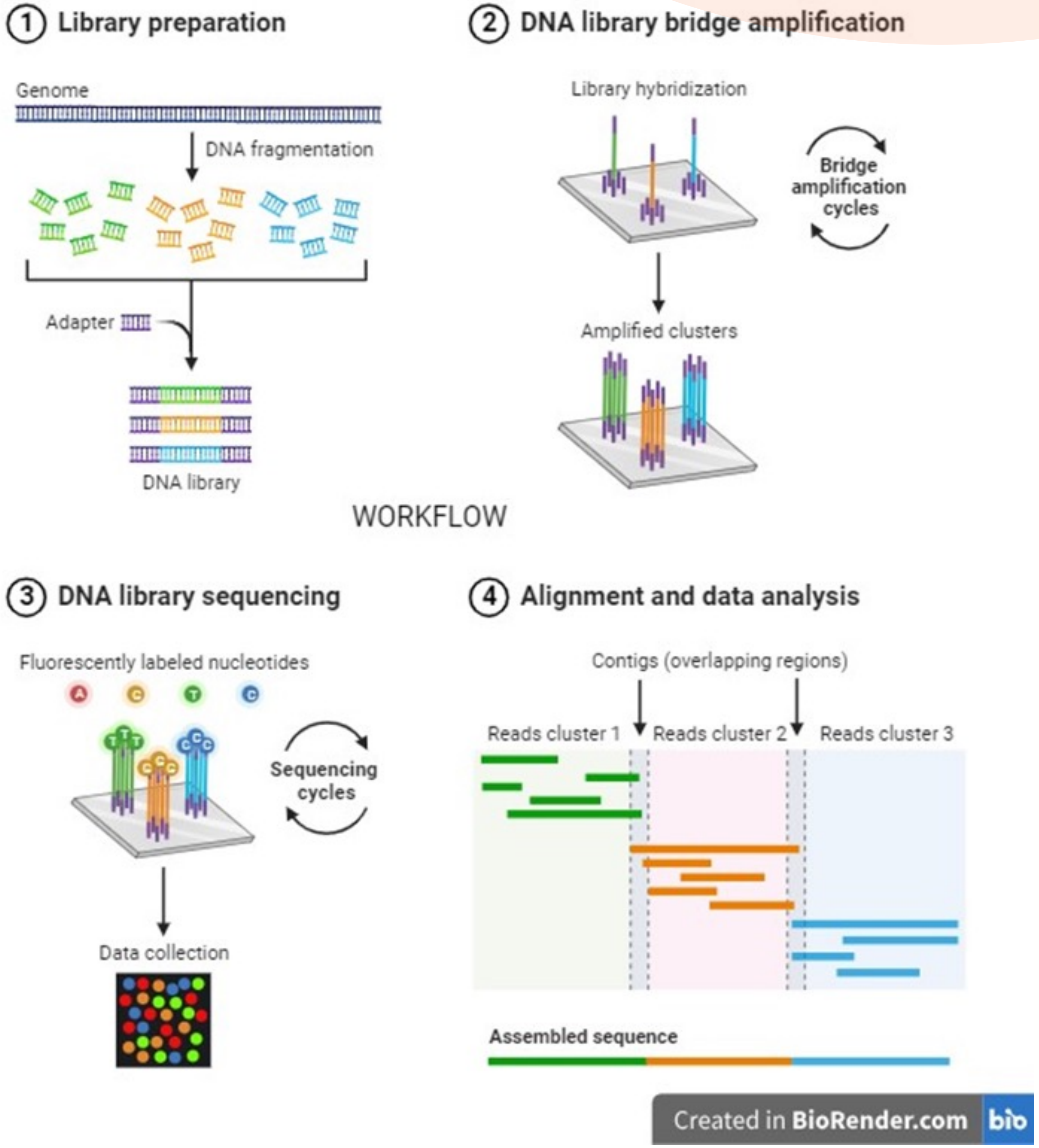
# CONTENTS

1. Introduction
2. Pre-processing
3. Methodology (what and why)
    - a). Quality analysis
    - b). Trimming
    - c). Read alignment
    - d). Alignment analysis
    - e). Variant calling and filtering
    - f). Variant annotation

# INTRODUCTION

## Next generation sequencing.

- Next-generation sequencing (NGS) is a massively parallel sequencing technology that offers ultra-high throughput, scalability, and speed.
- The technology is used to determine the order of nucleotides in entire genomes or targeted regions of DNA or RNA.
- NGS facilitates comprehensive genetic analysis however sequencing the entire genome is cost-prohibitive.
- A more comprehensive analiysis of selected regions is done by whole exome sequencing(WES)



AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS AND DATA INTENSIVE SCIENCES

# INTRODUCTION CONT'

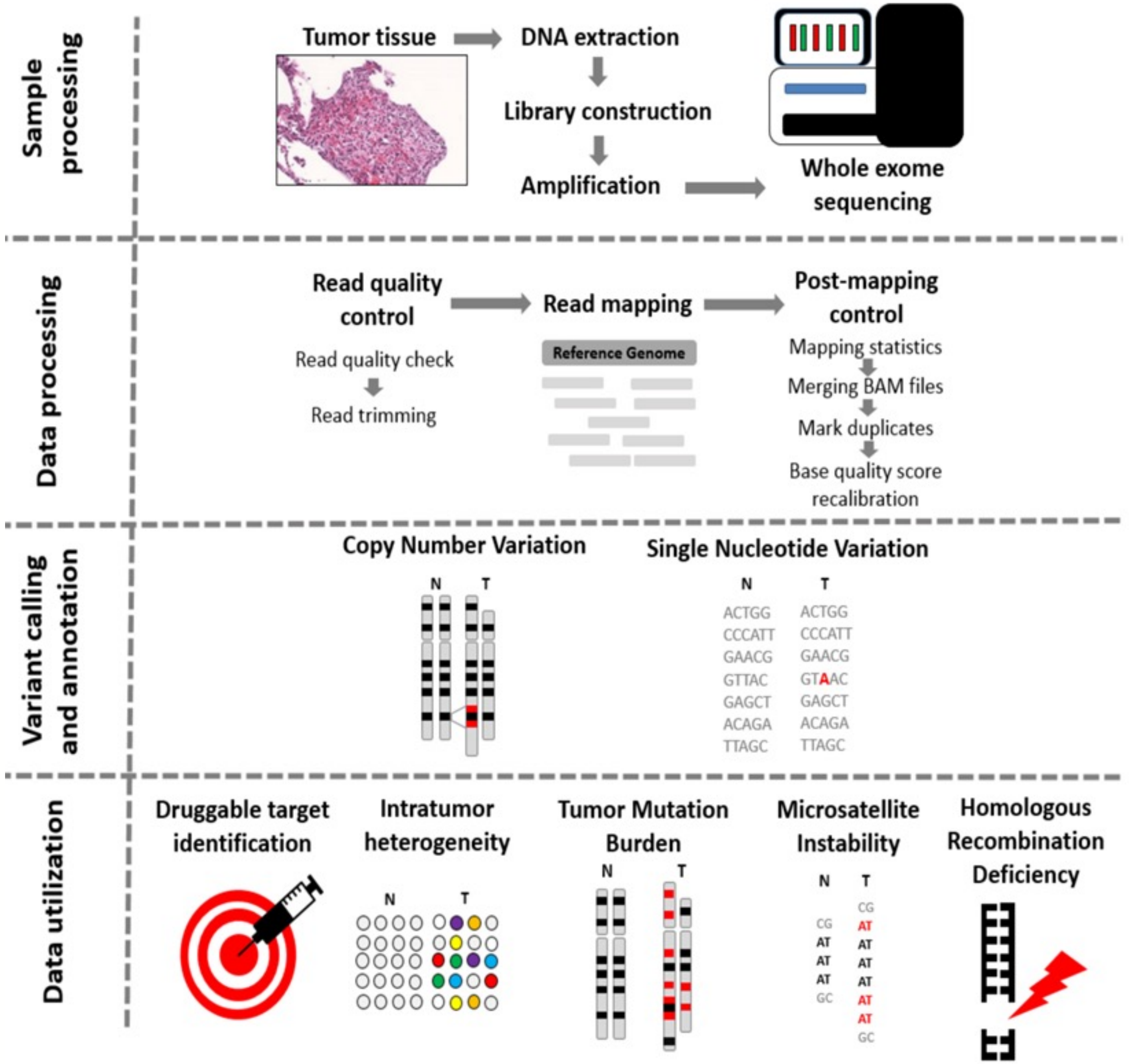
## Whole Exome Sequencing (WES)

- WES is a genomic technique for sequencing all of the protein-coding regions of genes in a genome.

- It utilizes a set of oligonucleotide hybridization probes that target known exon sequences.

## Applications of WES.

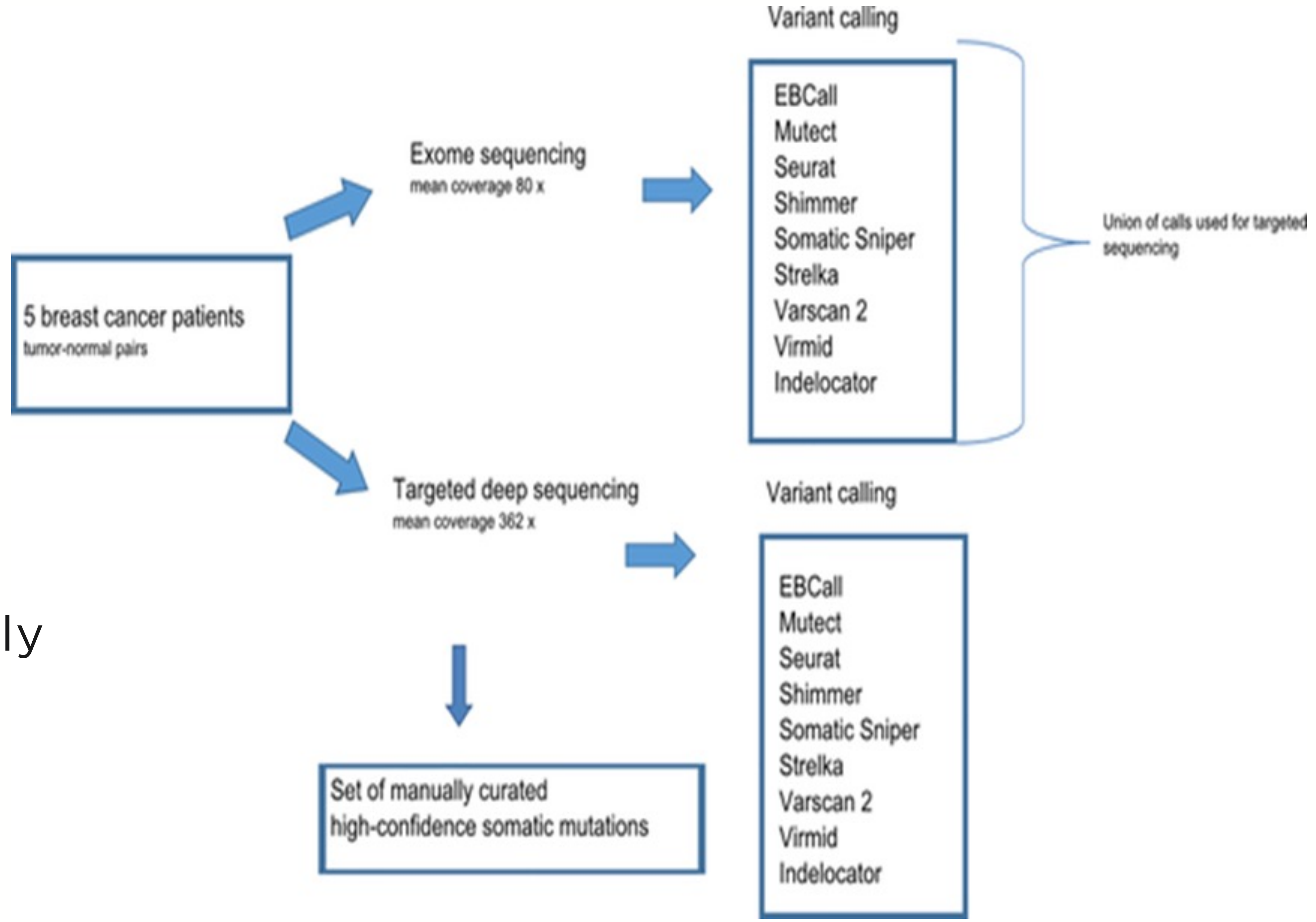
- Somatic variant detection
- Characterization of new therapeutic targets
- Profiling of copy-number variations (CNVs) and the detection of structural variations.
- Mutational analysis: the detection of single-nucleotide variants (SNVs) or small insertions and deletions (Indels).



# INTRODUCTION CONT'

## Somatic Variant Detection

- This is performed using algorithms and software tools specialized for the task
- Can classify a variant in a cancer sample as either germline or somatic with a second measure of likelihood
- Mutect2 somatic variant caller workflow used, largely following the Broad GATK4 Somatic SNVs + Indels Best Practices workflow
- Also a supplementary workflow based on a second popular caller: VarScan Somatic



5 breast cancer patients
tumor-normal pairs

Exome sequencing
mean coverage 80 x

Variant calling

EBCall
Mutect
Seurat
Shimmer
Somatic Sniper
Strelka
Varscan 2
Virmid
Indelocator

Union of calls used for targeted sequencing

Targeted deep sequencing
mean coverage 362 x

Variant calling

EBCall
Mutect
Seurat
Shimmer
Somatic Sniper
Strelka
Varscan 2
Virmid
Indelocator

Set of manually curated
high-confidence somatic mutations

# METHODOLOGY

## 1. Setup

- **Files**

  Fastq files: 4 lines per read

  Adapter files (TruSeq3-PE-2.fa)

  GATK resource bundles (ref, dict, VCF files)

- **Folder setup**

  Created main and sub directories

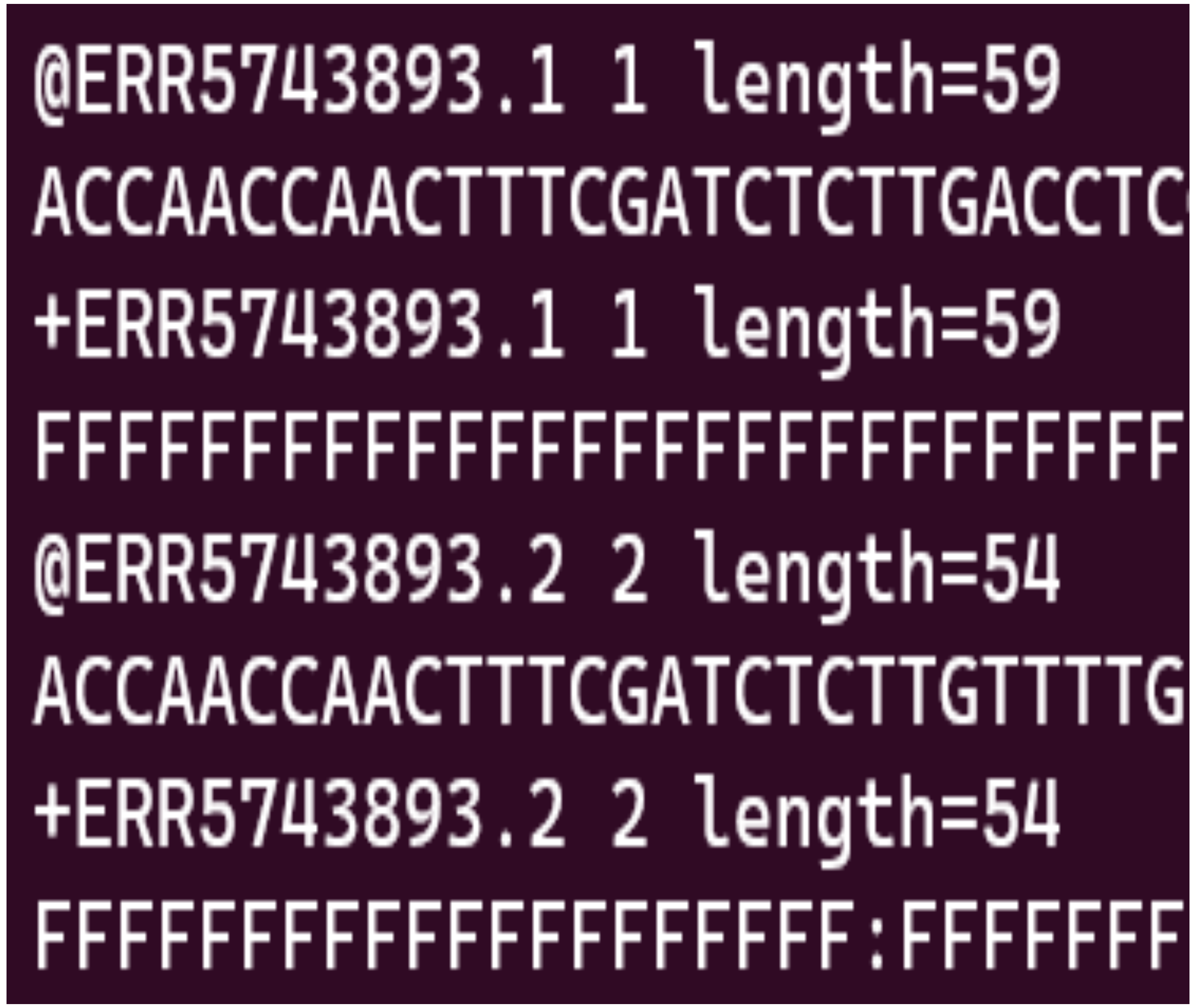
- **Software setup**

  Atleast 16gb RAM and 4 cores

  Create a new conda environment

  Configure conda channels (r, bioconda, conda-forge)

  *conda install –c bioconda "tool-name"*



```
@ERR5743893.1 1 length=59
ACCAACCAACTTTCGATCTCTTGACCTC
+ERR5743893.1 1 length=59
FFFFFFFFFFFFFFFFFFFFFFFFFFFF
@ERR5743893.2 2 length=54
ACCAACCAACTTTCGATCTCTTGTTTTG
+ERR5743893.2 2 length=54
FFFFFFFFFFFFFFFFFFFF:FFFFFFF
```

An example of a fastq file.

# METHODOLOGY

## 2. Preprocessing A

- **Quality Checks**

  *fastqc**

  Checks the quality of our reads.

- **Trimming**

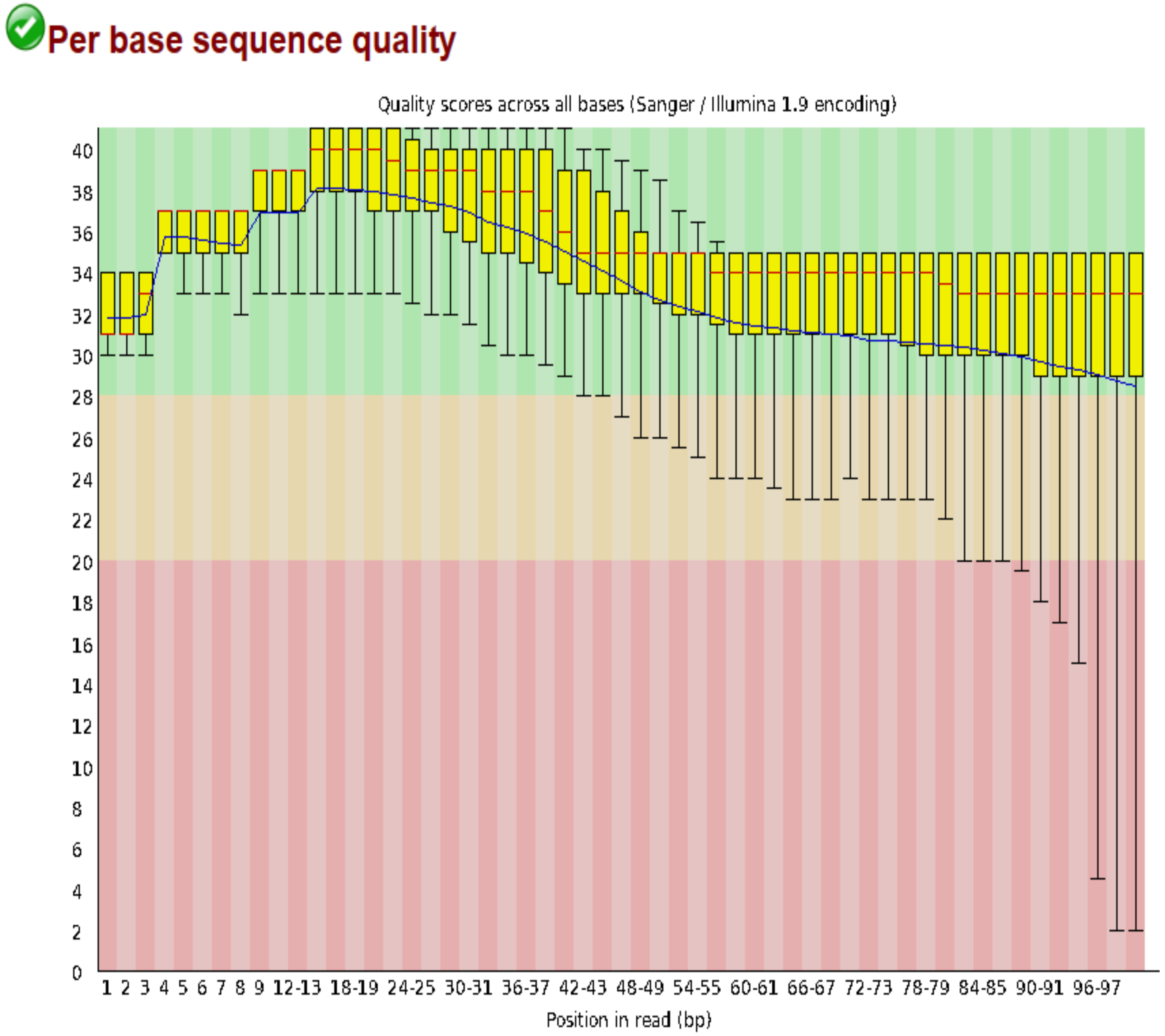
  *Trimmomatic*,* cutadapt, trim-galore, fastp

  Chop 5’ & 3’ ends, removes adapters, poor quality, and short reads.

- **Alignment**

  *bwa mem*,* bowtie

  Map R1 & R2 reads to reference genome to generate sequence alignment map (SAM)



A glimpse of a qc report file.

# METHODOLOGY

## 2. Preprocessing A

- **Compress, sort, and index the alignment file.**

 - *gatk-launch sortsam**

 - Save space, arrange and tag the reads

- **Mark duplicates**

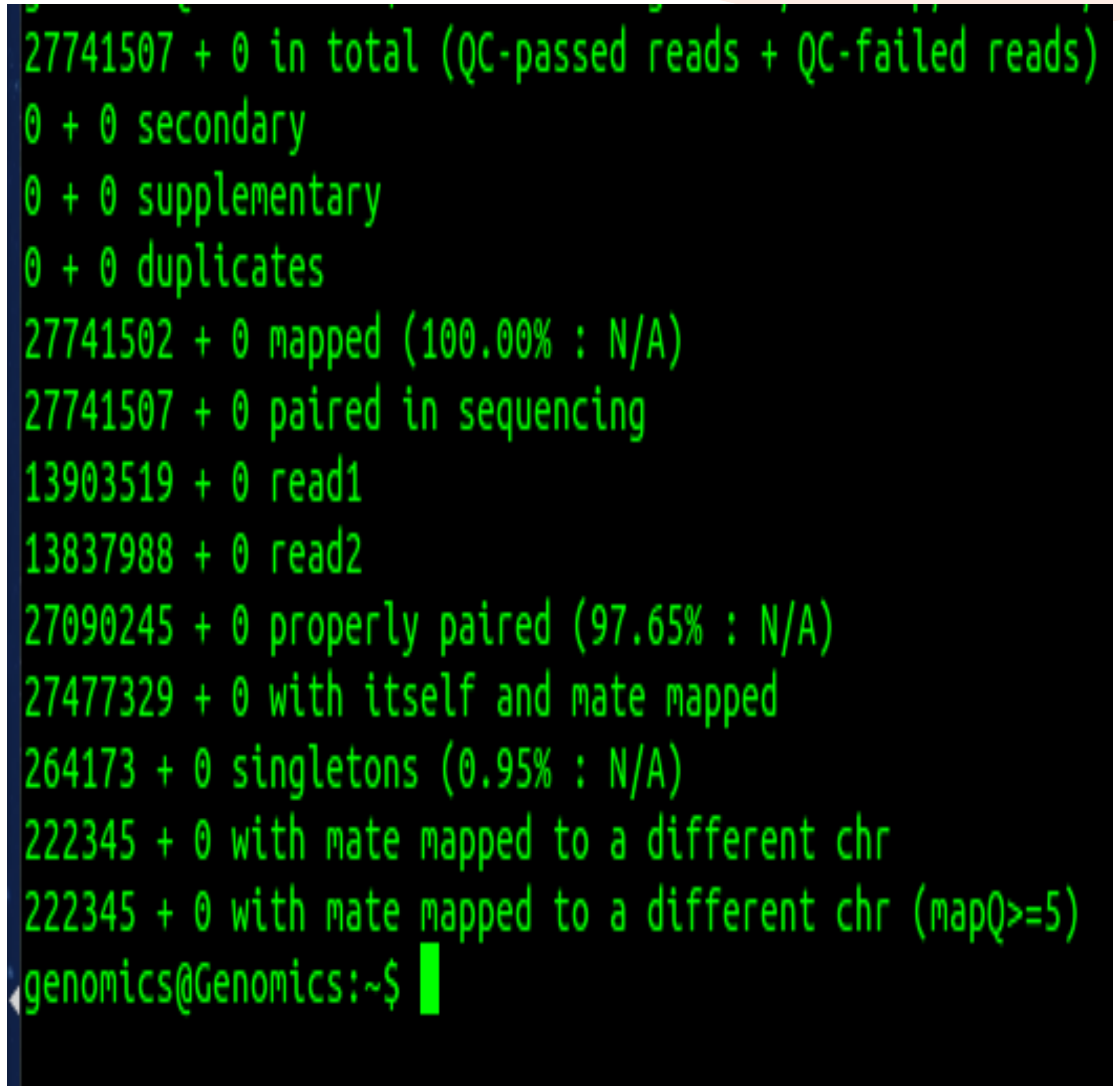
 - *gatk-launch MarkDuplicates**

 - Mark reads with the same coordinates as duplicates and retain only the highest scoring read.

- **Generate metrics and coverage data.**

 - *samtools flagstat & gatk-launch CollectHsMetrics*

 Statistics on read counts, mapped, duplicates, and txt files with means and median targe coverages, % of off-bait reads and % of targets that achieve particular coverage depths. (20X, 50X, 100X)



```
27741507 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
27741502 + 0 mapped (100.00% : N/A)
27741507 + 0 paired in sequencing
13903519 + 0 read1
13837988 + 0 read2
27090245 + 0 properly paired (97.65% : N/A)
27477329 + 0 with itself and mate mapped
264173 + 0 singletons (0.95% : N/A)
222345 + 0 with mate mapped to a different chr
222345 + 0 with mate mapped to a different chr (mapQ>=5)
genomics@Genomics:~$ 
```

Results of samtools flagstat

# METHODOLOGY

## 2. Preprocessing B

- **Base Quality Score Recalibration (BQSR).**

- *gatk-launch BaseRecalibrator*

- Correct systematic base scoring errors by first making a recalibration model and applying it to the bam file.

- **Re-build a recalibration model on the recal_bam**

- *gatk-launch BaseRecalibrator**

- For comparison purposes.

- **Compare the pre- and post BQSR tables.**

  - *gatk-launch AnalyzeCovariates*

  - For comparison purposes.

Weighing balance

# METHODOLOGY

## 3. Variant calling step 1

- **Somatic algorithms.**

- *gatk-launch Mutect2**

- It’s able to call variants as it compares the tumor and normal samples to the reference.

- It accommodates data from germline variant resources and an unmatched Panel of Normal datasets (PoN).

- **Create a Panel of Normals**

- *gatk-launch Mutect2**

  *gatk-launch CreateSomaticPanelOfNormals**

- Used to detect systematic experimental errors.

- Normal unrelated samples run on the same instrument **NOT the normal tissue samples.**

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS AND DATA INTENSIVE SCIENCES

# METHODOLOGY

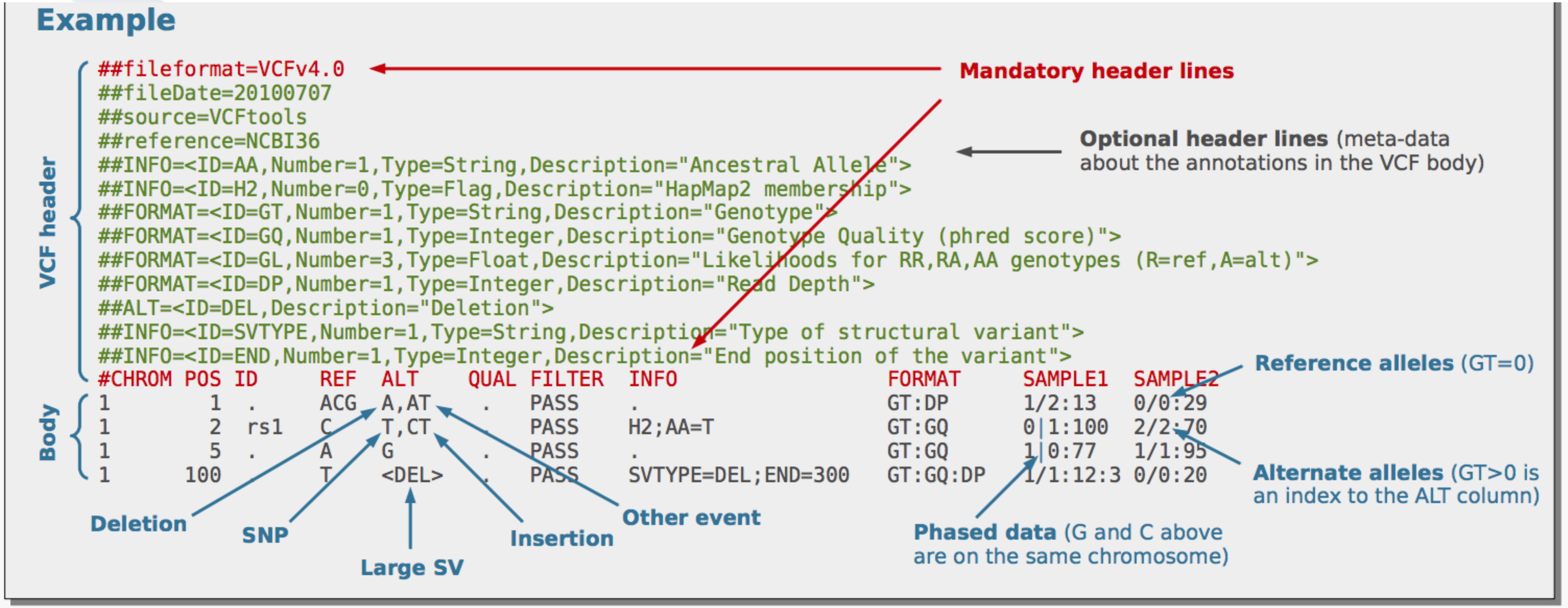
## 3. Variant calling step 2

- **Perform variant calling.**

- *gatk-launch Mutect2*

- It’s able to call variants for the Tumor/Normal pair.

- It can also call variants on tumor samples in absence of a matching normal



Example

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID      REF  ALT    QUAL FILTER INFO               FORMAT    SAMPLE1  SAMPLE2
1        1 .       ACG  A,AT    .   PASS   .                  GT:DP     1/2:13   0/0:29
1        2 rs1     C    T,CT    .   PASS   H2;AA=T            GT:GQ     0|1:100  2/2:70
1        5 .       A    G       .   PASS   .                  GT:GQ     1|0:77   1/1:95
1      100 .       T    <DEL>   .   PASS   SVTYPE=DEL;END=300 GT:GQ:DP  1/1:12:3 0/0:20
```

VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

# METHODOLOGY

## 3. Filtering variants

- **Generating a contamination file.**

- *gatk-launch GetPileupSummaries**

- To generate pile information for samples at sites of known mutations for both T & N.

- **Estimate contamination.**

- *gatk-launch CalculateContamination**

- To estimate the proportion of reads originating from other samples.

- **Apply the main set of filters.**

- *gatk-launch FilterMutectCalls*

- Passing variants will be labeled with PASS and those that fail shall be retained but with the ‘FILTER’ field populated with a list of filters of which the variant failed.

- **Apply second pass filter to mark sequencing artifacts**

- *gatk-launch FilterByOrientationBias*

- To remain with variants that passed all filters.

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS AND DATA INTENSIVE SCIENCES

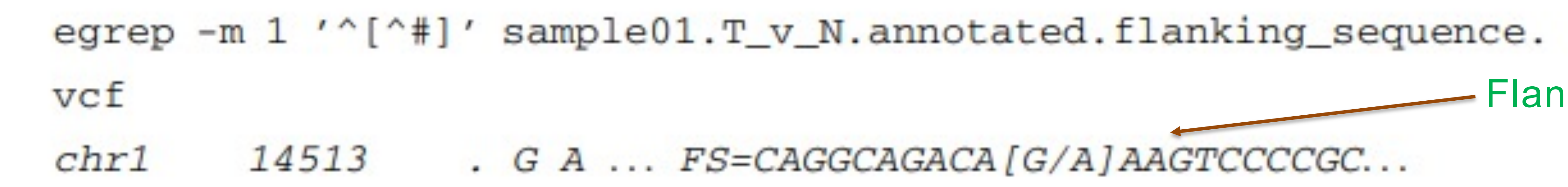
# METHODOLOGY

## 4. Variant Annotation

- **Add flanking sequence information.**

- *fill-fs**

- Use VCFtools to add the flanking genomic sequence around the variant locus which are useful for orthogonal confirmation of variants, and sometimes for custom analysis.



```
egrep -m 1 '^[^#]' sample01.T_v_N.annotated.flanking_sequence.
vcf
chr1     14513     . G A ... FS=CAGGCAGACA[G/A]AAGTCCCCGC...
```

Flanking sequences

- **Add basic annotations and impact predictions.**

- *snpEff*, Variant Effect Predictor (VEP), ANNOVAR, SVS/VarSeq***

- To predict the impact of a variant on the transcription or translation of a gene.

- SnpEff adds the following fields to the INFO field of each variant.

- ANN: Effect annotation always present

- LOF: Present only if the variant is predicted to cause loss of function.

- NMD: Present if the variant would result in Nonsense Mediated Decay

# CRITIQUE

## Strength

- The paper had strong relevance.
- The paper had a robust methodology.
- The paper had a clear research objective.
- The authors showed elements of innovation and originality.
- The authors demonstrated understanding of existing research in this field.

## Weaknesses

- Absence of example data for the trial of the given methodology.
- Some of the provided links for accessing the code are not functional.

# RECOMMENDATIONS

- The working group can achieve similar objectives with our own innovations and originality.
- We should look into having clear objectives and robust methodologies to answer research questions.
- We should share our resources openly with the cancer community to allow more contributions in the field.

Kiitos!
Grazie!
Obrigado!
Dzięki!
Tak!
Thank you!
Bedankt!
Gracias!
Merci!
Díky!
Danke!
Tack!
Novus