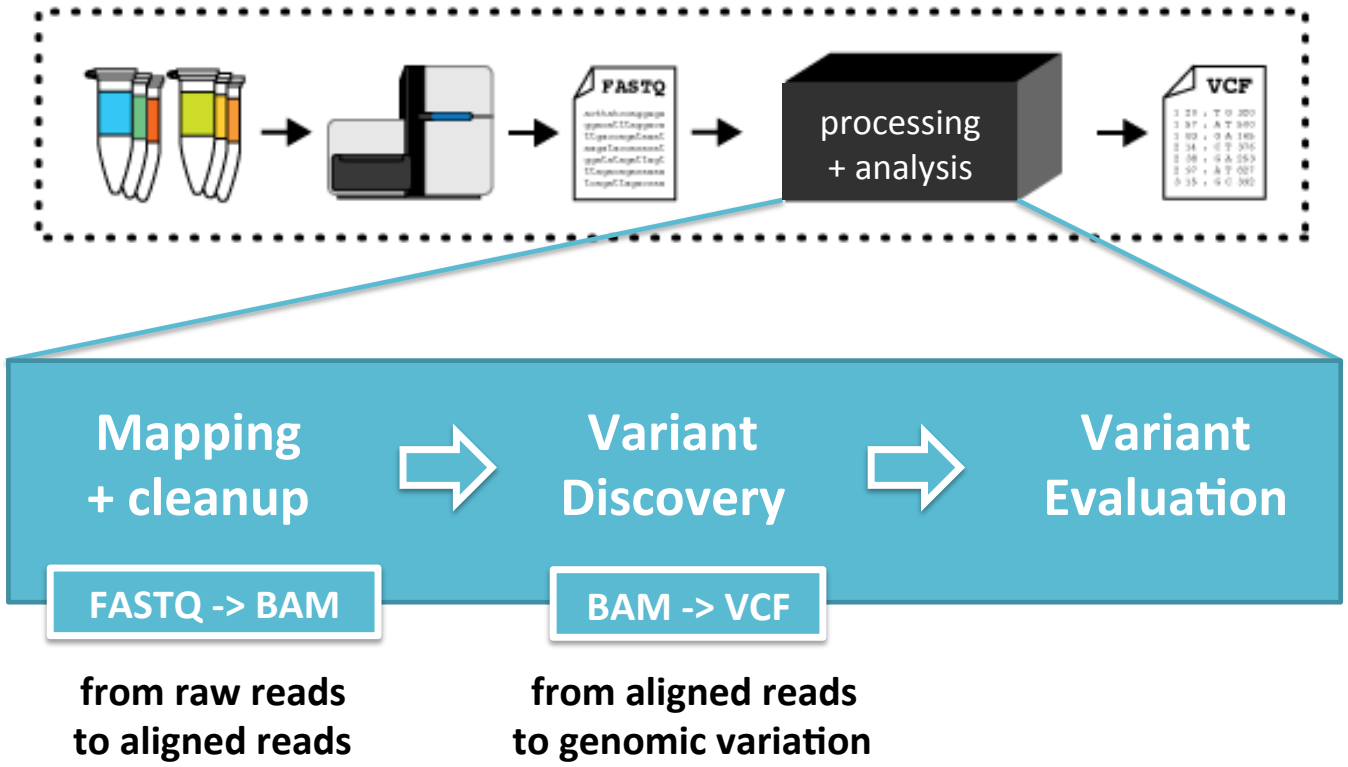
CURSO PRÁTICO DE BIOINFORMÁTICA:

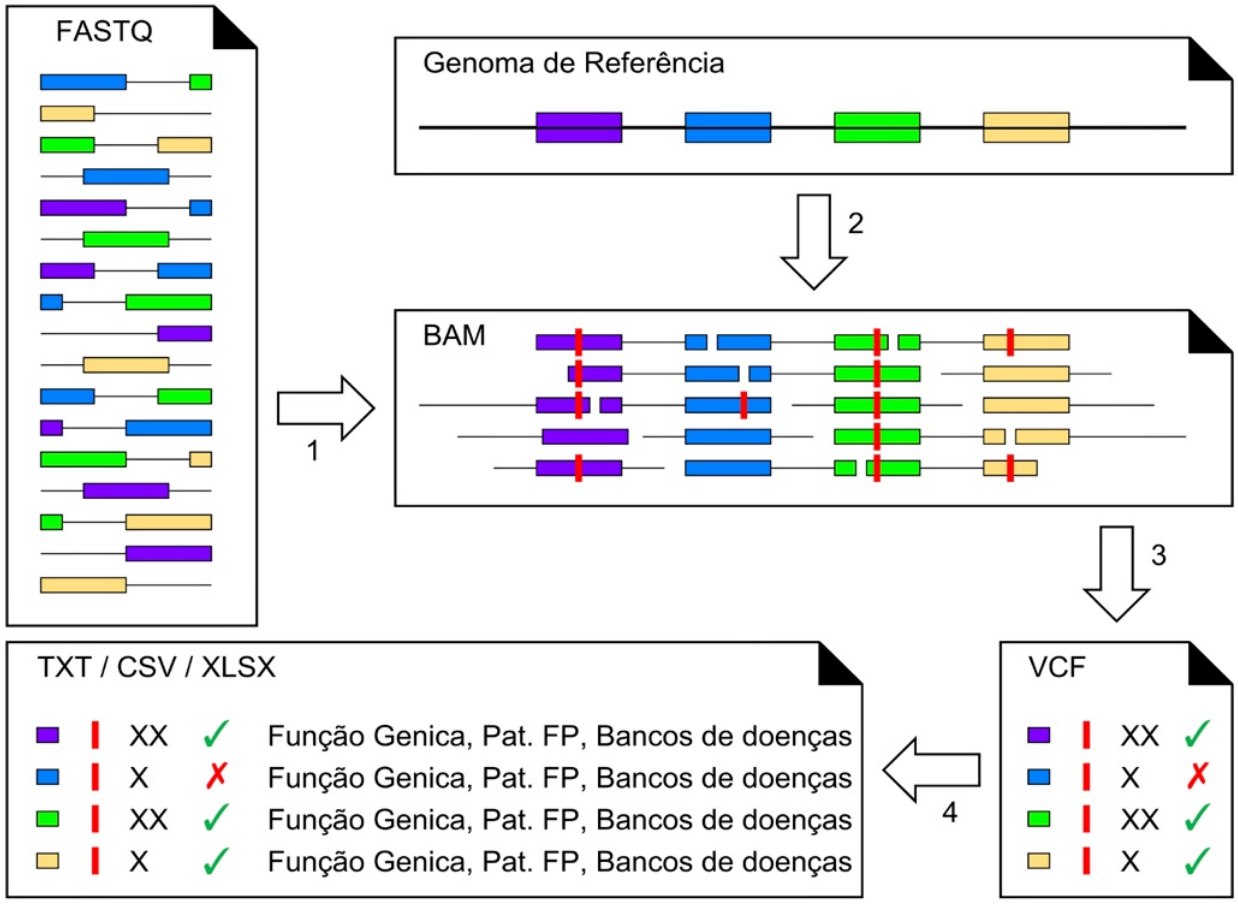
Aula 4 – Prática de Bioinformática

Renan Paulo Martin

17/07/19


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

Pipeline DNASeq



LEGENDA

- Região 1
- Região 2
- Região 3
- Região 4
- Variante
- X Heterozigose
- XX Homozigose
- Processamento
- ✓ Aprovado
- ✗ Reprovado
- Pat. Patogenicidade
- FP Frequência Populacional

Repositório do BROAD

- Genoma de Referência
- HAPMAP control
- dbSNP
- Arquivos adicionais utilizados

https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0

Origem do Arquivo FASTQ

- Projeto 1000g
- HG00096

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00096/alignment/

Geração do FASTQ

```
samtools fastq \
-1 HG00096_R1.fastq.gz \
-2 HG00096_R2.fastq.gz \
-0 /dev/null -s /dev/null –n \
HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20120522.bam
```

Preparação Genoma de Referência

- /usr/local/bin/bwa-0.7.17/bwa index hg19_g1k_v37_decoy/human_g1k_v37_decoy.fasta

Comando alinhamento

```
/usr/local/bin/bwa-0.7.17/bwa mem \
-t 4 -B 4 -O 6 -E 1 -M -a \
-R
"@RG\tID:HG00096\tPL:Illumina\tPU:TOTO2\tLB:TOTO\t
SM:TOTO3" \
/data/hg38_baseFiles/hg38/hg38.fa\
/data/bioinfo2019/HG00096/SRR081241_1.filt.fastq.gz\
/data/bioinfo2019/HG00096/SRR081241_2.filt.fastq.gz \
| samtools view -Shb -o HG00096.bam
```

samtools view

samtools –Shb –o HG00096.bam

-S auto detectar o formato de arquivo

-h incluir o cabeçalho

-b saída no formato BAM

samtools index

samtools index -b HG00096.bam

-b index no formato bai

samtools sort

samtools sort -O BAM -o HG00096_sorted.bam HG00096.bam

-O formato de saída

-o nome do arquivo de saída

Extrair CHR 20

samtools view –Sb HG00096_sorted.bam chr20 > HG00096_CHR20_sorted.bam

Picard MarkDuplicates

java -jar /data/bioinfo2019/picard.jar MarkDuplicates I=/data/bioinfo2019/HG00096/HG00096_CHR20_sorted.bam O=/data/bioinfo2019/HG00096/HG00096_CHR20_sorted_MKD.bam METRICS_FILE=/data/bioinfo2019/HG00096/HG00096_mark_dups_metrics.txt ASSUME_SORT_ORDER=queryname QUIET=true

GATK BaseRecalibrator

java -jar /data/bioinfo2019/GenomeAnalysisTK.jar -T BaseRecalibrator -I /data/bioinfo2019/HG00096/HG00096_CHR20_sorted_MKD.bam -R /data/hg38_baseFiles/hg38/hg38.fa -o /data/bioinfo2019/HG00096/HG00096_CHR20.recalfile -nct 8 --downsample_to_fraction .1 -L chr20 -knownSites /data/hg38_baseFiles/gold_standard_indels/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz -knownSites /data/hg38_baseFiles/known_indels/Homo_sapiens_assembly38.known_indels.vcf.gz

GATK PrintReads

- java -jar /data/bioinfo2019/GenomeAnalysisTK.jar -T PrintReads -I /data/bioinfo2019/HG00096/HG00096_CHR20_sorted_MKD.bam -R /data/hg38_baseFiles/hg38/hg38.fa -BQSR /data/bioinfo2019/HG00096/HG00096_CHR20.recalfile -o /data/bioinfo2019/HG00096/HG00096_CHR20_sorted_MKD_GATKrecal.bam --globalQScorePrior -1.0 --useOriginalQualities --preserve_qscores_less_than 6 --static_quantized_quals 10 --static_quantized_quals 20 --static_quantized_quals 30 --disable_indel_quals --filter_bases_not_stored

HaplotypeCaller

java -jar /data/bioinfo2019/GenomeAnalysisTK.jar -T HaplotypeCaller -R /data/hg38_baseFiles/hg38/hg38.fa -I /data/bioinfo2019/HG00096/HG00096_CHR20_sorted_MKD_GATKrecal.bam --emitRefConfidence GVCF --dbsnp /data/hg38_baseFiles/dbsnp151/Homo_sapiens_assembly38.dbsnp151.vcf.gz -L chr20 -o HG00096_CHR20.raw.snps.indels.g.vcf

IGV

- Integrative Genomic Viewer
- Visualização dos reads e das variantes

https://software.broadinstitute.org/software/igv/download