LBB-Mendelics-2021 Dia 2

**Questão 1: Quais variantes deverão ser desconsideradas no seu VCF? - Qualquer métrica do software de escolha poderá ser utilizada. Discorra sobre a métrica utilizada.**

Resposta: As variantes desconsideradas vão ser aquelas que ocorrem em posições que não estejam nos intervalos das regiões de captura. Também serão desconsideradas as variantes que não passaram pelo critério de qualidade do GATK (campo FILTER do vcf). Por fim, serão desconsideradas variantes com baixo valor de cobertura ( < 20).

# Descompactar o arquivo bed

$ bgzip -d coverage.bed.gz

# Selecionar apenas as variantes presente nos intervalos da região alvo

# Motivo do uso: Se o sequenciamento da amostra for um exoma é de interesse apenas reportar variantes presentes na região de captura.

$ intersectBed -a amostra-lbb\_normalized.vcf.gz -b coverage.bed -header > amostra-lbb\_intersectedBed.vcf

# Selecionar apenas as variantes com a flag PASS no campo FILTER

# Motivo do uso: O campo FILTER contém a informação de que se a variante passou ou não passou pelos critérios de qualidade da chamada de variante.

$ vcftools --vcf amostra-lbb\_intersectedBed.vcf --remove-filtered-all --recode --recode-INFO-all --out amostra-lbb.vcftools\_filter\_step1

# Desconsiderar variantes com cobertura menor do que 20

# Motivo do uso: Para uma maior confiança na chamada de variante deve se considerar uma profundidade mínima de 20.

$ vcftools --vcf amostra-lbb.vcftools\_filter\_step1.recode.vcf --minDP 20 --max-missing 1 --recode --recode-INFO-all --out amostra-lbb.vcftools\_filter\_step2

**Questão 2: Discorra sobre as regiões com baixa cobertura e quais foram seus critérios. Figuras são bem-vindas.**

Resposta: Regiões com baixa cobertura podem ter algumas origens:

1. Problema na probe de captura
2. Sequenciamento de baixa qualidade

Critério para a escolha do valor de corte: Gerar um gráfico com o percentual de bases on-target cobertas para os valores os seguintes valores de cobertura: 5, 10, 20, 50, 100, 150 e 250. O valor escolhido vai ser aquele valor em que houver o primeiro declínio brusco da reta.

# Indexar o arquivo bam

$ samtools index amostra-lbb.processed.bam

# Calcular o percentual de bases cobertas on-target para cada limiar de cobertura.

$ java -jar gatk-package-4.2.0.0-local.jar DepthOfCoverage -I amostra-lbb.processed.bam -L coverage.bed -R grch38.chr22.fasta -O amostra-lbb\_depth --summary-coverage-threshold 5 --summary-coverage-threshold 10 --summary-coverage-threshold 20 --summary-coverage-threshold 50 --summary-coverage-threshold 100 --summary-coverage-threshold 150 --summary-coverage-threshold 250

# Rodar o script para gerar gráfico

python gerar\_plot.py

Chart, line chart

Description automatically generated# Para gerar o bed das regiões com baixa cobertura segue os comandos:

$ bgzip -c amostra-lbb.vcftools\_filter\_step1.recode.vcf > amostra-lbb.vcftools\_filter\_step1.recode.vcf.gz

$ bgzip -c amostra-lbb.vcftools\_filter\_step2.recode.vcf > amostra-lbb.vcftools\_filter\_step2.recode.vcf.gz

$ tabix -p vcf amostra-lbb.vcftools\_filter\_step1.recode.vcf.gz

$ tabix -p vcf amostra-lbb.vcftools\_filter\_step2.recode.vcf.gz

$ bcftools isec -p isec\_out amostra-lbb.vcftools\_filter\_step1.recode.vcf.gz amostra-lbb.vcftools\_filter\_step2.recode.vcf.gz

$ cd isec\_out

$ grep -v "#" 0000.vcf | awk '{print $1"\t"$2"\t"$2}' > low\_coverage.bed

**Questão 3: Obter informações sobre seu alinhamento. Quantos reads? Qual a porcentagem deles que foi mapeada corretamente? Muitos alinharam em mais de um local do genoma com a mesma qualidade?**

Resposta:

# Quantos reads?

$ samtools view amostra-lbb.processed.bam -c

1731238

# Qual a porcentagem mapeada corretamente?

$ samtools view amostra-lbb.processed.bam -f 2 -c

1729046

(1729046/1731238) \* 100 = 99.87%

# Muitos alinharam em mais de um local do genoma com a mesma qualidade?

$samtools view amostra-lbb.processed.bam -f 1024 -c

258976

$ samtools view amostra-lbb.processed.bam -f 1024 -bS > duplicates.bam

# Mostra quantos reads duplicados alinharam na mesma posição com a mesma qualidade

$ samtools view duplicates.bam | cut -f 4,5 | sort | uniq -c

# Contar quantos reads com MAPQ igual a 0

$ samtools view -h amostra-lbb.processed.bam | cut -f 5 | grep -w "0" | uniq -c

# A mesma resposta pode ser obtida usando o seguinte comando

$ samtools flagstat amostra-lbb.processed.bam

1731238 + 0 in total (QC-passed reads + QC-failed reads)

0 + 0 secondary

0 + 0 supplementary

258976 + 0 duplicates

1731238 + 0 mapped (100.00% : N/A)

1731238 + 0 paired in sequencing

865619 + 0 read1

865619 + 0 read2

1729046 + 0 properly paired (99.87% : N/A)

1731238 + 0 with itself and mate mapped

0 + 0 singletons (0.00% : N/A)

0 + 0 with mate mapped to a different chr

0 + 0 with mate mapped to a different chr (mapQ>=5)