Universidade Federal do Pará
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Genética e Biologia Molecular

Disciplina: Bioinformática Aplicada à Ciências Ômicas
Professor: Gilderlanio Santana de Araújo

Exercício I: Análise de Dados do 1000 Genomas

1. Escolha um cromossomo e faça o download do arquivo.vcf.

Acesse o site do FTP do 1000 Genomas e escolha um cromossomo específico para suas análises:

- https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
- Ex.: ALL.chr22.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz

2. Utilize a ferramenta PLINK ou VCFTools para calcular a quantidade de genótipos faltantes em cada SNP no arquivo .vcf que você analisará. Remova os SNPs com mais de 10% de dados ausentes com o PLINK.

Característica	Sumário	Filtro
Missing genotype por indivíduo	--missing	--mind *N*
Missing genotype por SNP	--missing	--geno *N*

4. Calcular a frequência dos SNPs bialélicos (VCFTools ou PLINK v1.9) (--freq).

5. Classificar os SNPs em raros e comuns. Baseado nas frequências dos alelos, classifique os SNPs como raros (frequência < 0.01) ou comuns (frequência >= 0.01).

6. Verificar estruturação genética por Análises de Componentes Principais (PCA), após feito o controle de qualidade dos SNPs na etapa anterior. Utilize uma biblioteca em R ou Python para gerar um gráfico de dispersão e verificar os grupos por ancestralidade genômica.

Obs. Todas as amostras do .vcf e suas respectivas regiões geográficas estão no arquivo 1kgp_genomes_metadata.txt