Laboratory of Genetics and Molecular Cardiology

Desafio Técnico – Bioinformata

O controle de qualidade é uma etapa fundamental em todo processo de sequenciamento. O(a) candidato(a) deverá implementar um *pipeline* automatizado que realiza o controle de qualidade de dados de *Whole Exome Sequencing* (WES). Para tal, deve utilizar um *dataset* público de exoma completo, que será fornecido. O pipeline deve processar os dados em ambiente Linux, que deve ser disponibilizado em um repositório bem documentado no GitHub.

O *pipeline* deverá realizar:

1. **Cálculo de cobertura genômica**;
2. **Inferência do sexo genético a partir dos dados de sequenciamento**;
3. **Estimativa de contaminação por DNA exógeno ou de outros indivíduos**.

Dataset

Utilize o seguinte arquivo público de WES, proveniente do projeto 1000 Genomes:

- NA06994.alt_bwamem_GRCh38DH.20150826.CEU.exome.cram
- NA06994.alt_bwamem_GRCh38DH.20150826.CEU.exome.cram.crai

Disponível para download em:

http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data/CEU/NA06994/exome_alignment/

Utilize o arquivo BED das regiões exônicas para o GRCh38 disponível em:

https://www.twistbioscience.com/resources/data-files/twist-exome-20-bed-files

Hashes md5 para checagem de integridade dos *downloads*:

- arquivo cram: 3d8d8dc27d85ceaf0daefa493b8bd660
- arquivo crai: 15a6576f46f51c37299fc004ed47fcd9
- arquivo bed: c3a7cea67f992e0412db4b596730d276

Tarefas Obrigatórias

O pipeline deve ser executado em ambiente Linux, com ferramentas de linha de comando, e pode utilizar linguagens como Bash, Python ou R. O *pipeline* automatizado deve ser capaz de realizar:

1. **Conversão do arquivo CRAM** para um formato apropriado (BAM ou FASTQ), se necessário. O CRAM pode ser usado diretamente, sem conversão, de acordo com suas escolhas de projeto;
2. **Cálculo de cobertura** nas regiões exônicas (do arquivo BED), incluindo:
    2.1. Profundidade média;
    2.2. Percentual do exoma coberto a pelo menos 10x e 30x;
3. **Inferência do sexo genético**, com base em cobertura dos cromossomos X e Y. Utilize ferramentas preexistentes ou implemente sua própria solução;
4. **Estimativa de contaminação**, selecione as ferramentas apropriadas e justifique sua escolha no README;
5. **Geração de relatórios** textuais e/ou gráficos com os resultados;
6. **Automação** do pipeline via *scripts* Bash, Makefile, Snakemake ou Nextflow (à sua escolha).

Entrega

O projeto deverá ser hospedado em um **repositório público do GitHub** contendo:

1. Scripts e arquivos de configuração;
2. Um README.md com:
    2.1. Sua identificação pessoal;
    2.2. Descrição do pipeline;
    2.3. Instruções detalhadas de uso;
    2.4. Dependências e ferramentas utilizadas;
    2.5. Comandos de exemplo;
    2.6. Explicações sobre os *outputs* e resultados esperados;
    2.7. Resultados obtidos com o processamento da amostra;
    2.8. Não é necessário adicionar ao repositório os arquivos de entrada (CRAM, genoma de referência e BED), mas indique no README a origem dos mesmos;
3. *Logs* e exemplos de arquivos de saída (preferencialmente em diretórios separados);
4. Boa organização da estrutura de diretórios e nomeação clara dos arquivos.

Outras orientações:

- Adicionar o usuário LGCM-dev no projeto
- **Prazo: até às 23h59 do dia 17/05/2025. Envie um e-mail para rogerio.rosa@hc.fm.usp.br confirmando a entrega e fornecendo a URL do projeto. Não realize *commits* depois desse horário.**

Critérios de Avaliação

Critério	Peso
Correção e robustez do pipeline	30%
Clareza da documentação e reprodutibilidade	25%
Organização e modularidade do código	20%
Uso adequado de ferramentas e boas práticas	15%
Criatividade e completude dos relatórios	10%

Dicas

- Automatize o máximo possível e evite etapas manuais;
- Evite *hardcoding* de caminhos e parâmetros fixos;
- Teste seu *pipeline* em múltiplas execuções para garantir consistência.