Descrição da Primeira Fase do Workflow CellHeap para Análise Single-Cell RNA-seq

Gabriel P. T. Silva[1], Maria Clicia S. de Castro[1], Vanessa S. Silva[2], Fabricio A. B. Silva[2]

[1]Instituto de Matemática e Estatística – Universidade do Estado do Rio de Janeiro (UERJ) Rio de Janeiro – RJ – Brazil

[2]Fundação Oswaldo Cruz (Fiocruz) Rio de Janeiro – RJ – Brazil

gabrielplaza@hotmail.com.br, mariaclicia@gmail.com, vsantos223@gmail.com, fabricio.silva@fiocruz.br

Abstract. *This paper describes the initial step of the CellHeap workflow used in the data analysis of the single-cell sequencing (scRNA-seq). Phase 1 of this workflow consists of the download and validation of the accession data to be used in subsequent phases. Finally, some planned improvements to this step are described.*

Resumo. *Este artigo descreve a etapa inicial do workflow CellHeap utilizado na análise dos dados de sequenciamento de células únicas (scRNA-seq). A Fase 1 consiste no download e validação dos dados de amostras a serem utilizadas nas fases subsequentes. Ao fim, são descritas algumas melhorias planejadas para esta etapa.*

1. Introdução

Atualmente, diversas centenas de terabytes de dados de sequenciamento de células únicas (scRNA-seq) estão disponíveis em repositórios públicos. Estes dados são relacionados a tecidos, comorbidades e condições.

Os protocolos de sequenciamento de células únicas (scRNA-seq) desenvolvidos nos últimos anos permitem um melhor conhecimento dos sistemas biológicos. Eles podem requerer infraestrutura de software e de sistemas computacionais de larga escala, dependendo da quantidade de dados utilizada.

O *workflow* CellHeap (Silva et al, 2021) foi concebido para realizar análises single-cell RNA-seq (scRNA-seq) voltadas ao estudo da COVID-19. A relevância deste tipo de análise reside na compreensão mais precisa do transcriptoma em células individuais. As tecnologias scRNA-seq geram conjuntos de dados que descrevem o estado de células individuais com resolução sem precedentes (Nicolás e Costa, 2021).

Neste contexto, temos um grande volume de dados por amostra que deve ser armazenado e tratado, o que acarreta desafios de natureza computacional a medida que a disponibilidade de dados públicos continua aumentando e os estudos passam a lidar rotineiramente com conjuntos de dados cada vez maiores. Portanto, é necessária uma infraestrutura computacional poderosa.

O *workflow* CellHeap foi projetado e desenvolvido para execução no supercomputador Santos Dumont, um dos mais importantes da América Latina, situado no Laboratório Nacional de Computação Científica (LNCC).

Na Seção 2 apresentamos o *workflow* CellHeap e na Seção 3 concluímos apresentando os próximos passos de implementação do *workflow*.

2. Fases do *Workflow*

O *workflow* CellHeap é dividido em 5 fases como mostra a Figura 1. Ele segue um esquema de *pipeline* onde as saídas de uma fase são as entradas das fases subsequentes e podem ser usadas diferentes ferramentas em cada fase. Cada uma das fases está descrita a seguir.

A Fase 1 é responsável pela curadoria, *download* e validação dos dados brutos. A Fase 2 gera informações sobre os dados da amostra, que podem ou não ser agregados. Por exemplo, um dos arquivos de saída da Fase 2 é uma matriz de contagem de genes. A Fase 3 é responsável pelo controle de qualidade na geração das células de interesse. É nesta fase que devem ser removidos da amostra artefatos como códigos com células duplicadas. A Fase 4 reduz a dimensionalidade e forma *cluster* com os genes selecionados. A Fase 5 contém uma grande quantidade de análises diferentes em nível celular ou genético.

Observamos que as Fases 1 e 2 são as que consomem maior quantidade de tempo de execução e armazenamento de memória se comparados às Fases 3, 4 e 5. As Fases 1 e 2 são as que manipulam os dados brutos das amostras.

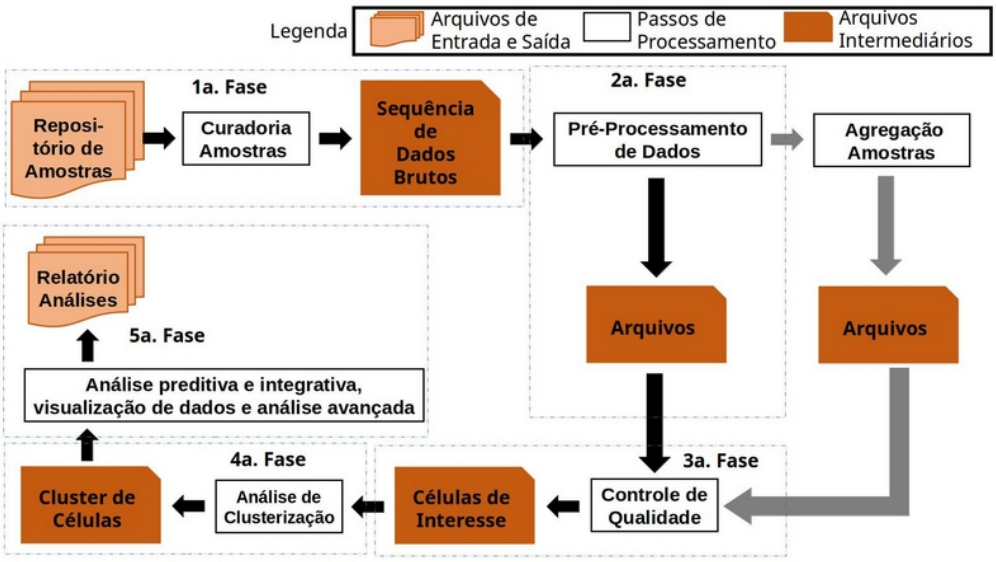

Figura 1. Diagrama conceitual do workflow Cellheap. As caixas pontilhadas representam diferentes fases.

2.1. Fase 1

A primeira fase consiste no *download* dos dados brutos referentes às amostras selecionadas. Para esta fase foram criados dois *scripts*: um com informação sobre a

amostra, *download* e validação, e outro com a geração de arquivos de arquivos divididos e nomeados de acordo com a entrada esperada na Fase 2.

Para obter maior visibilidade no processo, incialmente, o comando *vdb-dump --info* é invocado com a identificação da amostra considerada. O seu resultado é escrito na saída padrão e contém diversas informações como tamanho total da amostra, *link* de onde ela foi recebida, sua data de entrada no banco de dados entre outras informações. O *download* é realizado com o comando *prefetch* do pacote de software *sratoolkit* do *National Center for Biotechnology Information* (NCBI). Este pacote é disponibilizado como módulo no computador Santos Dumont e inclui todos os comandos utilizados nesta fase. Para garantir a ausência de falhas no *download*, o comando *vdb-validate* é executado em seguida. Como não existe dependência entre diferentes amostras, é possível usar este *script* de forma paralela usando o paradigma *bag-of-tasks*.

Obtidos os arquivos .sra e validados, é preciso convertê-los para o formato mais apropriado à leitura da ferramenta *Cellranger count* na Fase 2. Para isto, o segundo *script* da Fase 1 emprega o comando *fasterq-dump*. Este comando substitui o comando *fastq-dump* cuja versão é sequencial. O *fasterq-dump* é *multithreaded* e permite realizar a conversão com processamento paralelo. Dessa forma, a ferramenta explora os recursos computacionais do supercomputador de forma mais eficiente. Por fim, os arquivos .fastq são compactados com a ferramenta *gzip* e renomeados para o formato esperado pelo *Cellranger count*.

3. Próximos Passos do *Workflow*

Nos procedimentos iniciais, cada invocação do primeiro *script* da Fase 1 aloca um nó computacional do computador Santos Dumont e realiza o *download* e validação de uma única amostra. Para melhorar a eficiência na utilização de recursos e melhorar o tempo de obtenção das amostras está sendo avaliada a possibilidade de especificar múltiplos identificações de amostra como argumento à este *script* e distribuir o *download* das amostras entre vários nós diferentes. Mesmo com a alta capacidade das interfaces de rede do supercomputador, o *download* é seguramente a etapa mais longa desta primeira fase, visto que uma única amostra pode representar a transferência de mais de 100GB de dados. Otimizar esta etapa representa um ganho de agilidade significativo na execução do *workflow*.

Referências

Silva, Vanessa S. et al. (2021) “CellHeap: a Workflow for Optimizing COVID-19 Single-Cell RNA-seq Data Processing in the Santos Dumont Supercomputer”, In: Proceedings of the Brazilian Symposium on Bioinformatcs, Lecture Notes in Bioinformatics serires, Springer, 2021.

Nicolás, Marisa F. e Costa, Maiana O. C. (2021) “Single-cell RNA sequencing (scRNA-seq)”, Notas de Aula “Seminário Inova Covid19”, Brazil.