Workflows Científicos de RNA-Seq em Ambientes Distribuídos de Alto Desempenho: Otimização de Desempenho e Análises de Dados de Expressão Diferencial de Genes

Lucas Cruz[1,2], **Micaella Coelho**[1], **Rafael Terra**[1], **Diego Carvalho**[2], **Luiz Gadelha**[1], **Carla Osthoff**[1], **Kary Ocaña**[1]

[1]Laboratório Nacional de Computação Científica (LNCC)
Petrópolis – RJ – Brasil

[2]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)
Rio de Janeiro – RJ – Brasil

{lucruz, micaella, rafaelst, lgadelha, osthoff, karyann}@lncc.br

d.carvalho@ieee.org

Abstract. *We present the new ParslRNA-Seq scientific workflow of high-performance computing for differential gene expression analyses, which showed improvements in the total execution time of up to 70%. ParslRNA-Seq performance was validated through a comparative analysis of differential gene expression from a real RNA-Seq experiment in cardiomyocytes. Finally, the article discusses the choice of which modifications in the workflow modeling lead to improve computational performance and scalability, based on provenance data information. ParslRNA-Seq is available at https://github.com/lucruzz/rna-seq.*

Resumo. *Apresentamos uma versão do workflow científico ParslRNA-Seq para análises de experimentos de Expressão Diferencial de Genes, acoplada a ambientes de Computação de Alto Desempenho, que mostrou melhoras no tempo total de execução de até 70%. O desempenho ParslRNA-Seq foi validado por meio de uma análise comparativa de dados da EDG em cardiomiócitos de um experimento real de RNA-Seq. Finalmente, o artigo traz discussões sobre a eleição de quais modificações na modelagem do workflow levam à melhora do desempenho e escalabilidade computacional, baseadas em dados de proveniência. ParslRNA-Seq está disponível em https://github.com/lucruzz/rna-seq.*

1. Introdução

Para a bioinformática, a modelagem e execução de experimentos de Sequenciamento RNA (RNA-Seq) representam um desafio pela complexidade na gerência e análise de grandes volumes de dados biológicos e computacionais. A técnica de RNA-Seq é utilizada para análises de Expressão Diferencial de Genes (EDG) que permite estudar o comportamento de um conjunto de transcritos de uma célula em uma dada condição fisiológica ou de desenvolvimento, tal como o câncer. A tecnologia de RNA-Seq é um grande avanço nos estudos da transcriptômica, da mesma maneira que prévias análises de microarranjos, mas ainda existem muitos desafios pela frente relacionados à complexidade, natureza e volume dos dados em experimentos de EDG [Anders and Huber 2010].

Diversos programas, algoritmos e sistemas baseados em técnicas estatísticas podem ser acoplados na análise de experimentos de RNA-Seq, mas ainda não há uma metodologia universal e cada uma dessas abordagens apresentam vantagens, limitações e um enorme desafio para a bioinformática no quesito de desempenho computacional. Por essa razão a bioinformática se alia a tecnologias como Ciência de Dados, Computação de Alto Desempenho (CAD) e Aprendizado de Máquinas visando estratégias para prover soluções de baixo custo computacional. *Workflows* científicos representam o fluxo encadeado de atividades de um experimento [Mattoso et al. 2010], o que possibilita estabelecer uma melhor modelagem, gerência de execução e análise que levam a reforçar a reprodutibilidade, confiabilidade e escalabilidade do experimento. Sistemas de Gerência de *Workflows* Científicos (SGWfC) baseados em *web* como Galaxy[1] e pacotes estatísticos do R e Bioconductor (EdgeR, DESeq2) são usados nos estudos de EDG. No quesito de automação de tarefas o uso de linguagens ou SGWfC distribuídos e paralelos como Nextflow, Tavaxy, Kepler, Pegasus, Swift e Parsl são estratégias promissoras [Ferreira da Silva et al. 2017].

O presente trabalho apresenta uma nova versão do *workflow* científico ParslRNA-Seq [Cruz et al. 2020], com o desempenho validado por análises comparativas computacionais e de inferência em análises de EDG. A versão atual é composta por seis atividades principais, onde as modificações foram realizadas para utilização da nova atualização do programa HTSeq, a qual permite o particionamento dos dados de entrada para distribuição e execuções paralelas em múltiplos *cores*. Embora tenha sido necessário incluir mais três novas atividades para paralelização do HTSeq, a nova versão do ParslRNA-Seq pode levar as execuções a alcançarem um ganho em tempo computacional de até 70%. A grande vantagem da versão atual se dá pela execução paralela e distribuída em múltiplos nós, pois na versão anterior não haviam ganhos de desempenho significativos nas execuções multinós. O ambiente computacional usado para os testes é o supercomputador Santos Dumont[2] (SDumont). O artigo está organizado da seguinte forma: a Seção 2 traz os trabalhos relacionados; a Seção 3 apresenta conceitos sobre experimentos de RNA-Seq; a Seção 4 descreve o *workflow* científico ParslRNA-Seq; na Seção 5 são apresentados os resultados e análises experimentais; e, por fim, a Seção 6 traz as considerações finais e os trabalhos futuros.

2. Trabalhos Relacionados

[Cruz et al. 2020] traz a modelagem e análises de desempenho do ParslRNA-Seq_*alfa* executado no ambiente do SDumont. As análises envolvem a gerência do Parsl para o uso eficiente dos recursos computacionais e a melhor exploração do parâmetro *multithread* do Bowtie2, que em conjunto levam a uma melhora de desempenho significativa do *workflow*. Outros trabalhos para análises de EDG incluem *pipeline* seveaseq gerenciado por *scripts* e segundo as nossas pesquisas executado de forma serial. Pacote RSEM para a identificação e quantificação de transcritos em análises de RNA-Seq, não usa genomas de referência, usa Bowtie e um algoritmo do RSEM que calcula a abundância foi otimizado, mas sem estudos em ambientes de CAD. MyExperiment hospeda uma biblioteca de diferentes tipos de *workflows* de bioinformática e RNA-Seq. Galaxy é uma plataforma fortemente Web que apresenta diversos *workflows* de RNA-Seq, com opções de paralelização quando integradas com o Taverna (Tavaxy). É o mais similar ao nosso ParslRNA-Seq,

[1]https://galaxyproject.org/
[2]https://sdumont.lncc.br/

que foi explorado intensivamente em ambientes de supercomputação. Dentre as outras ferramentas temos Tximeta, Salmon, Sailfish e featureCounts [Liao et al. 2014]

3. Background sobre Análises de Expressão Diferencial de Genes

A tecnologia de Sequenciamento de Nova Geração (SNG) revoluciona o campo das análises genômicas e transcritômicas devido ao sequenciamento de forma massiva em grande escala. A técnica conhecida como RNA-Seq se baseia na análise da EDG de genes usando ferramentas de modelização estatísticas dos dados relacionando com a quantidade de transcritos. Estudos de RNA-Seq tem facilitado o estudo de *splicing* alternativo, Polimorfismos de Nucleotídeo Simples (PNS), modificações postranscripcionais e mudanças na expressão gênica através do tempo ou entre grupos de tratamentos ou progressão de uma doença. Análises de EDG permitem elucidar o nível de expressão entre condições experimentais diferentes e estabelecer se existe diferença significativa entre elas. Realizar estudos de EDG indica a formalização de um fluxo de atividades ou processos encadeados que podem ser representados pelo uso de diferentes *software* ou algoritmos, sendo essencial estabelecer uma correlação biológica para os resultados estatísticos resultantes.

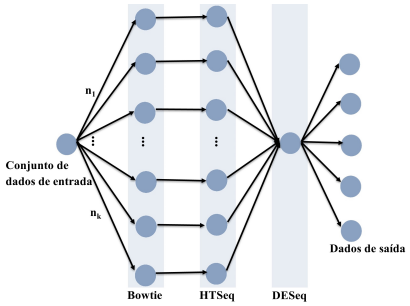
Uma tarefa básica na análise de EDG é a detecção e contagem de dados de RNA-seq. Os dados de contagem são apresentados em forma de tabela com cada amostra relacionada ao número de fragmentos de sequência atribuídos a cada gene. Uma questão de análise importante é a quantificação e inferência estatística das mudanças sistemáticas entre as condições, em comparação com a variabilidade dentro das condições. O pacote DESeq2 fornece métodos para testar a expressão diferencial pelo uso de modelos lineares generalizados binomiais negativos; as estimativas de dispersão e alterações logarítmicas incorporam distribuições anteriores baseadas em dados.

4. ParslRNA-Seq: *Workflow* para Análises de Expressão Diferencial de Genes

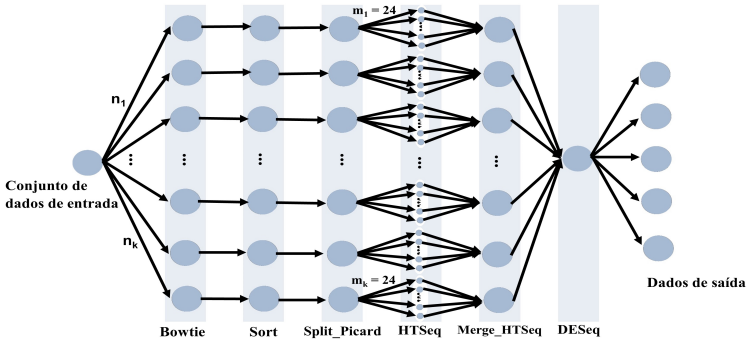
Neste trabalho, a modelagem conceitual do ParslRNA-Seq é inspirada na versão ParslRNA-Seq_*alfa* de [Cruz et al. 2020] com três atividades (Bowtie, HTSeq, DESEq) apresentada na Figura 1(a). ParslRNA-Seq da Figura 1(b) é composto de seis atividades, incluídas Sort, Split_Picard e Merge_HTSeq que visam a melhora de desempenho sobre HTSeq. ParslRNA-Seq recebe como entrada o genoma de referência do *Mus musculus*, o arquivo de formato GTF (Gene Transfer Format) com metadados genômicos e os arquivos de sequenciamento em formato FASTQ. Um arquivo de formato CSV foi criado para relacionar os FASTQ e as condições experimentais: três FASTQ de controle e três FASTQ de condição Wnt (*Wingless pathway*, via metabólica de sinalização de transcrição Wnt).

A atividade 1 executa o programa Bowtie2[3] que mapeia e compara as leituras dos genomas caractere por caractere. A atividade 2 executa o programa Samtools[4] versão 1.10 que realiza uma ordenação nas leituras. A atividade 3 executa o programa Picard[5] versão 2.25.0 usado para a manipulação e divisão dos arquivos de leituras. As atividades 2 e 3 são complementares para o ganho de desempenho, de forma que com as leituras ordenadas, o tempo computacional de execução para o particionamento das leituras é reduzido. A atividade 4 executa o programa htseqcount do HTSeq[6] versão 0.13.5 para

[3]http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
[4]http://www.htslib.org/doc/samtools.html
[5]http://broadinstitute.github.io/picard/
[6]https://htseq.readthedocs.io/



(a) ParslRNA-Seq_*alfa* (3 atividades)



(b) ParslRNA-Seq (6 atividades)

Figura 1. Modelagem Conceitual do *Workflow* Científico ParslRNA-Seq.

a contagem do número de leituras mapeadas por cada gene. Com n arquivos de leituras mapeadas, o HTSeq envia cada um para n *cores*, gerando um único arquivo de saída com $n + 1$ colunas, onde a primeira coluna representa o gene e as demais colunas representam contagens realizadas em cada arquivo. A atividade 5 (HTSeq-Merge) é um *script* em Python que faz a junção dos dados gerados pela execução *multicore* do HTSeq, unindo em uma única coluna todas as contagens realizadas. A atividade 6 executa o pacote DESeq2[7] que aplica estatísticas de EDG sobre as condições experimentais.

5. Resultados Experimentais

Nesta seção são apresentados análises de desempenho e escalabilidade do *workflow* no SDumont e análises biológicas e estatísticas da EDG das amostras frente à condição Wnt.

5.1. Configuração do Ambiente Computacional

O SDumont está entre as 500 máquinas mais poderosas do mundo. Ele possui uma capacidade de processamento de 5.1 Petaflop/s, com 34.688 CPU *multicores* distribuídas em 1.132 nós computacionais que são interligados por uma rede de interconexão *Infiniband* FDR/HDR. Os nós computacionais possuem duas CPUs Ivy Bridge Intel Xeon E5-2695v2 (12c @2.4GHz) e 64Gb de memória RAM e uma GPU Nvidia K40. As execuções foram realizadas em nós computacional de duas CPUs Intel Xeon E5-2695v2 Ivy Bridge, 24 núcleos (12 por CPU) e 64 GB de memória RAM. Todos os *software*, algoritmos, dependências de bioinformática (Bowtie, Samtools, Picard, HTSeq e DESeq2) e os componentes do Parsl foram alocados e instalados no ambiente do SDumont.

[7]https://bioconductor.org/packages/release/bioc/html/DESeq2.html

5.2. Configuração do Experimento

Os dados pertencem a um experimento real de RNA-Seq[8], extraídos do repositório público *Gene Expression Omnibus*[9] (GEO) e divididos em: (1) grupo de controle: Control_1 (SRR5445794); Control_2 (SRR5445795); Control_3 (SRR5445796) e (2) grupo de condições das vias Wnt: Wntup_1 (SRR5445797); Wntup_2 (SRR5445798); Wntup_3 (SRR5445799). O organismo é *Mus musculus* e o GEO.ID é GSE97763 (Plataforma Illumina HiSeq 2000 - Mus musculus). As leituras de sequência foram alinhadas ao genoma de referência do rato (UCSC versão mm9) com Bowtie2. Para cada gene, o número de leituras mapeadas foi contado com htseqcount. DESeq2 analisou a EDG a partir das matrizes das contagens do alinhamento e do mapeamento das sequências frente ao genoma de referência. Essas matrizes (arquivo GTF) contém o número de leituras que foram alinhadas de forma única (colunas) com os exones de cada gene nas amostras (colunas).

5.3. Análises de Desempenho e Escalabilidade

A Figura 2 apresenta a análise de desempenho e escalabilidade em minutos de um experimento até 24 *threads* do ParslRNA-Seq_*alfa* proposto previamente por [Cruz et al. 2020].

Desempenho do ParslRNA-Seq_*alfa*: Atividade Bowtie. Apresenta a melhora no Tempo Total de Execução (TTE) da atividade Bowtie do ParslRNA-Seq_*alfa* sobre três cenários: (a) Serialização do Parsl (Bowtie com opção *multithread*); (b) Serialização do Bowtie (Parsl com opção *multithread*) e (c) n *threads* Parsl x n *threads* Bowtie (Parsl&Bowtie com opção *multithread*). Esse último cenário (c) explora a parametrização *multithread* do Bowtie atrelado ao uso do paralelismo de tarefas do Parsl, que levou a uma dupla paralelização combinando Bowtie&Parsl, apresentando o melhor resultado, pelo que foi o eleito e acoplado no ParslRNA-Seq proposto, a ser usado no subseção seguinte.

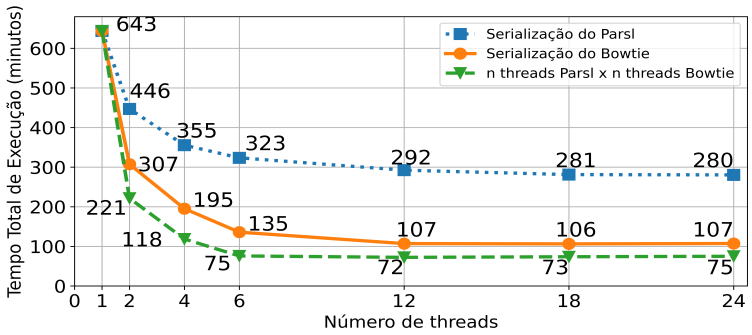


Figura 2. Escalabilidade em minutos do ParslRNA-Seq_*alfa*.

Desempenho do ParslRNA-Seq Proposto: Atividade HTSeq. O *workflow* chama a versão mais atualizada da atividade HTSeq que permite uma paralelização *multicore* das entradas. Diferentemente da versão anterior, a qual cada entrada da atividade HTSeq era executada em um único *core*, não havendo paralelização, na versão atual cada entrada foi particionada em 24 subentradas, de modo que cada subentrada fosse alocada e executada em um único núcleo de CPU do SDumont. Essa estratégia resultou em uma diminuição do tempo computacional dessa atividade de 305,3283 a 30,4161 minutos, o

[8]https://sfb1002.med.uni-goettingen.de/production/literature/publications/201
[9]https://www.ncbi.nlm.nih.gov/geo/

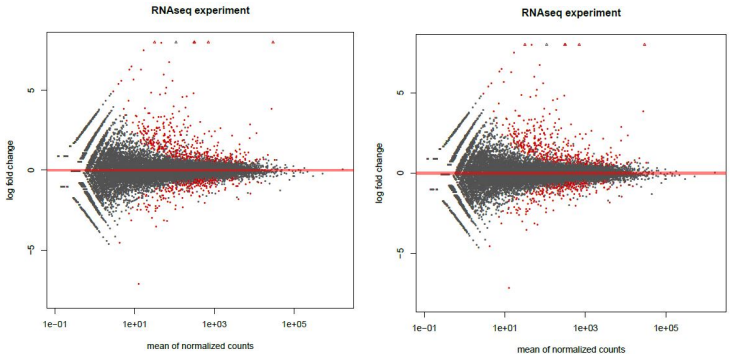
que representa aproximadamente 90% de melhora (Tabela 1). As demais atividades não apresentam gargalos de execução: Bowtie e Samtools usam parâmetros *multithreads* e Picard, HTSeq-Merge e DESeq são de baixo tempo e custo computacional. A Tabela 1 apresenta os resultados da execução serial do *workflow*. O TTE do ParslRNA-Seq_*alfa* foi de 326,0738 minutos frente ao TTE de 95,6427 minutos do ParslRNA-Seq, o que representa uma diminuição de 70,67%. Esse resultado demonstra que o ganho é três vezes maior, mesmo com a inclusão das três atividades (Sort, Picard e HTSeq-Merge). Em um cenário de execução paralela e distribuída do *workflow* a nova versão executa em cerca de 24 minutos, enquanto a antiga versão executa em 72 minutos.

Tabela 1. Tempo Total de Execução em minutos do ParslRNA-Seq proposto.

Workflow	**Bowtie**	**Sort**	**Picard**	**HTSeq**	**HTSeq-Merge**	**DESeq**	**TTE**
versão 1	19,2698	-	-	305,3283	-	1,4757	326,0738
versão 2		5,8768	38,56	30,4161	0,0443		95,6427

5.4. Análises de EDG baseada em Contagem para Dados de RNA-Seq

Os resultados da análise para a seleção de genes diferencialmente expressos sob DESeq2 são apresentados como gráficos MA-Plot para a média das leituras normalizadas de cada gene em relação ao log_2 da alteração dobrada (Figura 3). Os pontos correspondentes aos genes identificados como expressos diferencialmente (valor p ajustado menor que 0,05) são destacados em vermelho. Os pontos que ficam fora da janela são plotados como triângulos abertos apontando para cima ou para baixo, dependendo se o valor de $logFC$ é maior que 2 ou menor que -2, respectivamente.



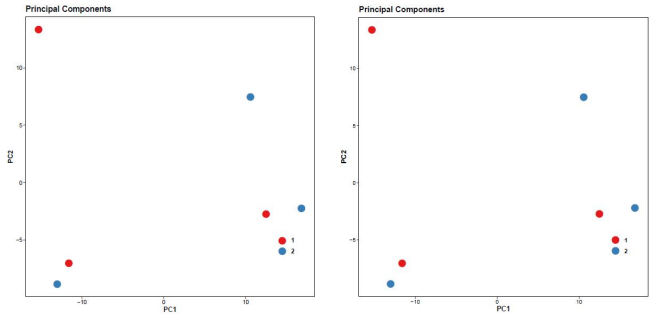
(a) ParslRNA-Seq_*alfa* (b) ParslRNA-Seq

Figura 3. Média de contagens normalizadas gene/mudança $log_2 foldChange$.

Análise de Expressão Diferencial. DESeq2 usa o modelo probabilístico Binomial Negativo para a normalização para executar análises de EDG. DESeq2 normaliza dados estimando tamanho e dispersão da amostra, ajusta dados a um Modelo Linear Generalizado (MLG) binomial negativo e verifica a EDG usando o teste paramétrico de Wald. DESeq2 calcula as funções $baseMean$ (média das leituras normalizadas); $log_2 foldChange$ (proporção de leituras em função de log_2); $lfcSE$ (erro padrão); $stat$ (Wald); $pvalue$ e $padj$ (valores p e p ajustados das transcrições DE). Na Figura 3, os genes diferencialmente expressos (DE) aparecem em vermelho e os demais em preto. Algumas considerações

são: (1) Haverá genes DE (em vermelho), acima e abaixo da linha que delimita os valores $log_2 foldChange$. Os genes acima obtiveram mais contagens na condição Controle que na condição Wnt, e os pontos abaixo o oposto. (2) Quanto mais altas as contagens médias (mais à direita do gráfico), os genes DE estarão mais próximos da linha limite, influenciado pelo $log_2 foldChange$, que quanto maiores as médias, embora sejam diferentes, o logaritmo será menos diferente e, portanto, o limite para determinar que um gene é DE será menor. (3) Há uma tendência de não haver genes DE à esquerda do gráfico. Quanto mais à esquerda do gráfico estivermos, menores serão as contagens observadas para os genes e, quando quase não há contagens, quase nenhuma diferença pode ser mostrada.

Análise de Escala Multidimensional (Gráfico MDS). É uma técnica multivariada que permite analisar visualmente a proximidade entre as amostras de um mesmo estudo, colocando-as em determinadas dimensões. Em um gráfico MDS, a primeira dimensão representa a magnitude da alteração inicial que melhor separa as amostras e, portanto, explica a maior proporção de variação nos dados. O gráfico MDS da Figura 4 apresenta a relação entre as amostras para detectar a EDG. O que mais chama a atenção no gráfico é a separação entre os dois grupos. As amostras Wnt (em vermelho), estão no geral com maiores valores positivos no eixo X, do que as amostras do grupo de controle (em azul). A aproximação entre alguns grupos pode se dever a efeitos como o gênero do rato (macho/fêmea), mas sem afetar a condição geraldo experimento.



(a) ParslRNA-Seq_*alfa* (b) ParslRNA-Seq

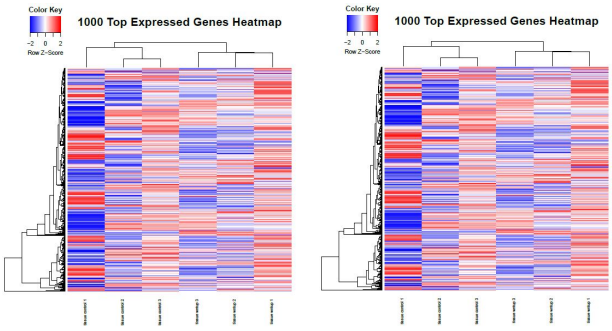
Figura 4. Escala multidimensional de distâncias da relação entre as amostras.

Heatmaps (Mapas de calor). O dendrograma da Figura 5 permite visualizar o agrupamento das amostras com base em um grupo hierárquico junto com os níveis de expressão dos genes individuais. Neste tipo de gráfico, a variância em cada uma das linhas da matriz $log_2 - CPM$ foi previamente calculada e o número de genes a serem exibidos foi estabelecido. A seleção dos 1000 genes mais variáveis agrupou as amostras de acordo com o grupo experimental. Os genes superexpressos são representados em vermelho, os subregulados em azul e a cor branca indica ausência de mudança de expressão. Cada linha da grade representa um gene e cada coluna uma amostra.

6. Conclusão

Com o presente trabalho foi apresentada a nova versão do *workflow* científico ParslRNA-Seq para análises de GDE em experimentos de RNA-Seq. Essa modelagem visou a otimização e a exploração dos recursos computacionais, principalmente, na atuação em ambientes de CAD. Além disso, a fim de manter também a integridade dos dados gerados pela aplicação foram levantadas análises comparativas da otimização realizada frente

a modelagem da versão anterior e mostra que os resultados experimentais não se modificam. A nova versão do *workflow* apresenta uma melhora de 70,67% no tempo computacional, em relação a versão anterior para uma execução serial. Futuramente serão feitas análises de desempenho para execuções paralelas e distribuídas do *workflow*, após esse estudo o ParslRNA-Seq será disponibilizado para a comunidade científica através do Bioinfo-Portal[10], um portal hospedado no LNCC, voltado ao fortalecimento das pesquisas envolvendo o uso da bioinformática. Também serão realizados estudos para explorar a execução de múltiplos experimentos de RNA-Seq sendo executados em paralelo.



(a) ParslRNA-Seq_*alfa* (b) ParslRNA-Seq

Figura 5. Mapas de calor dos 1000 genes mais variáveis.

7. Agradecimentos

Ao LNCC (MCTI, Brasil) por prover os recursos do supercomputador SDumont. As agências de fomento brasileiras CNPq (Projeto Universal) e FAPERJ (Projeto JCNE). Aos pesquisadores Laura Zelarayan-Behrend e Harald Kusch (Universidade de Göttingen, Alemanha) pela colaboração científica nas análises de RNA-Seq.

Referências

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1.

Cruz, L., Coelho, M., Gadelha, L., Ocaña, K., and Osthoff, C. (2020). Avaliação de desempenho de um workflow científico para experimentos de rna-seq no supercomputador santos dumont. In *Anais Estendidos do XXI Simpósio em Sistemas Computacionais de Alto Desempenho*, pages 86–93, Porto Alegre, RS, Brasil. SBC.

Ferreira da Silva, R., Filgueira, R., Pietri, I., Jiang, M., Sakellariou, R., and Deelman, E. (2017). A characterization of workflow management systems for extreme-scale applications. *Future Generation Computer Systems*, 75:228–238.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Mattoso, M., Werner, C., Travassos, G., Braganholo, V., Ogasawara, E., de Oliveira, D., Cruz, S., Martinho, W., and Murta, L. (2010). Towards supporting the life cycle of large-scale scientific experiments. *International Journal of Business Process Integration and Management*, 5:79–92.

[10]https://bioinfo.lncc.br/