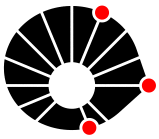

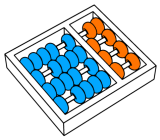

Análise transcriptômica de dados de osteossarcoma com os softwares HISAT2 e StringTie

Julia Pietro *João Meidanis*

Technical Report - IC-22-05 - Relatório Técnico
April - 2022 - Abril

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Análise transcriptômica de dados de osteossarcoma com os softwares HISAT2 e StringTie.

Julia Pietro João Meidanis*

Resumo

Análises transcriptômicas têm se mostrado necessárias no ramo da Medicina, pois faz-se viável a identificação de produtos da transcrição do DNA em diferentes moléculas. No caso de mutações, como em células cancerígenas, os transcritos e proteínas sofrem alterações, o que torna possível analisar o padrão de expressão, com base nos produtos e vias ativas. A partir disso, tratamentos podem ser desenvolvidos, visando a sobrevivência e o bem-estar de pacientes. Desse modo, com base nos dados de 35 pacientes do Centro Infantil Boldrini e com auxílio do CENAPAD, sequências de RNA de crianças com osteossarcoma foram analisadas via *softwares* e linguagens de programação. Este projeto teve como principal objetivo a obtenção de dois produtos finais: matrizes com contadores de genes e transcritos por amostras, que irão alimentar estudos sobre vias metabólicas e expressão diferencial de genes. Para isso, as amostras passaram por testes de qualidade, limpeza de sequências, mapeamento com genoma de referência e, finalmente, o processo de contagem de transcritos. Os relatórios antes da limpeza expressaram resultados com baixas estatísticas básicas. Estes percentuais serviram de auxílio para realizar os ajustes necessários nas amostras, como retirar bases de baixa qualidade. Desse modo, logo após a limpeza dos dados, o relatório de qualidade apresentou melhoras significativas na qualidade das sequências. Então, o mapeamento das leituras com o genoma de referência resultou em taxa de alinhamento com média de 95,91%. Por último, a contagem de transcritos forneceu tabelas com os níveis de transcritos e genes expressos por amostras. Dessa forma, durante o PEOp 2022 foi possível desenvolver técnicas de bioinformática e chegar ao resultado esperado.

1 Introdução

O termo Transcriptoma refere-se ao conjunto de moléculas de RNA transcritas a partir do DNA, em determinado contexto fisiológico celular. De modo que, as sequências de RNA refletem a regulação gênica e precedem um dos produtos finais da expressão gênica, as proteínas. Análises transcriptômicas proporcionam a identificação de proteínas que serão geradas e o DNA que as originou. Em caso de mutações a nível de DNA e, consequentemente, modificações proteicas e metabólicas, a análise dos transcritos permite identificar o padrão de expressão das moléculas. Atualmente, esta análise é realizada por programas de computador, pois é a melhor forma de lidar com quantidades significativas de dados.

*Inst. de Computação, UNICAMP, 13083-852 Campinas, SP.

Este processo se tornou útil para a medicina de forma geral, possibilitando conhecer as vias metabólicas das proteínas ativas em casos de doenças com origem em mutações gênicas, como o câncer, por exemplo. Assim, tratamentos específicos podem ser desenvolvidos.

Este projeto teve como objetivo principal realizar contagem de transcritos por genes em amostras do Centro Infantil Boldrini de pacientes com osteossarcoma, e disponibilizar os resultados encontrados para estudos de expressão diferencial gênica. Com o intuito de gerar os contadores, a análise transcriptômica contará com processos de testes de qualidades, limpeza, mapeamento e quantificação dos dados. Conceitos de bioinformática foram utilizados, o que possibilitou o aprendizado de linguagens de programação em *Linux*. Além disso, para cada um dos processos, softwares atuaram otimizando esta cadeia de procedimentos.

2 Metodologia

As análises realizadas durante o projeto ocorreram graças ao uso do Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD) [1]. O CENAPAD disponibiliza aos usuários ambientes computacionais para *hardwares* e *softwares*, assim como armazenamento de dados. Com isso, através do uso de filas, foi possível rodar os códigos, o que permitiu o desenvolvimento do projeto em tempo hábil para outras atividades do estágio.

As 35 amostras (*raw data*) provenientes de cada um dos pacientes foram gentilmente cedidas, mediante acordo de confidencialidade, pelo Centro Infantil Boldrini e correspondem a sequências de RNA de crianças entre 1 e 8 anos com osteossarcoma [2]. As informações gerais sobre os dados podem ser encontradas no NCBI [3][4](National Center for Biotechnology Information), plataforma digital que providencia acesso a informações biomédicas e genômicas. Os arquivos se encontram em formato `paired.end`, ou seja, cada transcrito foi transformado em um cDNA (DNA complementar), que tiveram suas duas fitas lidas a partir de casa ponta, ambas na direção 5' para 3'. Primeiramente, cada uma das amostras precisou passar por um controle de qualidade. Este procedimento inicial padrão possibilita identificar possíveis problemas com os *raw data*, e foi feito através de um *software* chamado `FastQC`, o qual gera um relatório de qualidade com informações sobre:

- estatística básica dos dados;
- qualidade da sequência por base;
- pontos da qualidade de sequência por base;
- conteúdo da sequência por base;
- conteúdo de sequência por GC;
- conteúdo de sequência por N;
- distribuição do tamanho da sequência;
- níveis de duplicação da sequência;

- sequência super representadas;
- conteúdo de adaptadores [5][6].

Com base nos parâmetros de baixa qualidade identificados no passo anterior, as amostras são encaminhadas para a etapa de limpeza. Para promover a retirada de sequências específicas de acordo com parâmetros explicitados nas linhas de código, foi utilizado o *software* `Trimmomatic` [7] . Todos os dados passaram pela mesma filtragem e, ao final, para cada um deles, outro relatório de controle de qualidade foi gerado, com o intuito de identificar se houve melhorias.

Com os dados devidamente limpos e com qualidade certificada, o mapeamento pode ser realizado. Conforme o protocolo sugerido pelo grupo do Dr. Salzberg da *Johns Hopkins University* para análises de sequências de RNA [8], o *software* `HISAT2` [9] mostrou-se ser um método rápido para mapear os transcritos no genoma de referência. Para tal é necessário um genoma de referência indexado, ou seja, com estruturas pré-computadas que facilitem as buscas [10]. No caso deste projeto, o genoma humano foi adquirido no Manual do `HISAT2`, que disponibiliza um banco de dados com genomas previamente indexados. Por se tratar de um estudo de pacientes com osteossarcoma, escolheu-se um genoma com marcadores genéticos (SNP) e anotações de transcritos (TRAN). Dessa forma, por meio do `HISAT2`, cada par de amostras `paired.end` foi mapeado com o genoma de referência para identificar a posição genômica dos transcritos [11] [12]. Arquivos '`.sam`' foram gerados, sendo necessário convertê-los para arquivos '`.bam`' através do software `SAMtools` [13], para assim, compactar os documentos para acesso mais rápido.

A contagem de transcritos necessita de dois processos. Um deles utiliza o *software* `StringTie` [14], que gera estimativas de transcritos por amostra em conformidade com um genoma de referência. Posteriormente, o produto gerado pelo `StringTie` é convertido em duas matrizes, uma com contadores de genes e outra análoga, mas para transcritos. A partir desta quantificação, informações sobre os transcritos de cada gene podem ser extraídas, estimulando estudos sobre vias metabólicas ativadas em casos de mutações, análises de expressão diferencial gênica e associação com tamanhos de telômeros, importantes para o avanço no tratamento oncológico.

3 Resultados e discussão

3.1 Controle de qualidade

O relatório de qualidade gerado pelo primeiro `FastQC` proporcionou o conhecimento da qualidade inicial das amostras. Assim, para o programa em questão os *inputs* (entradas) foram os 35 dados `paired.end` do Instituto, gerando 70 relatórios. Individualmente, além de gráficos e listas de dados, as estatísticas básicas são acompanhadas de avisos: *"Passed", "Warning"* e *"Fail"*, que representam, respectivamente, normalidade, medianidade e anormalidade das sequências [15]. De um modo geral, para as amostras avaliadas, os *'Fails'* dominantes são representados por qualidades das bases entre 0 e 20 Phred. O conteúdo de sequências por base também mostrou-se irregular, com porcentagens incompatíveis entre as

bases adenina, timina, citosina e guanina. Outro parâmetro que interferiu na qualidade das amostras foram as sequências super-representadas, ou seja, sequências que apareceram mais do que deveriam. Elas são comparadas com uma lista de contaminantes comuns, providenciando a origem a partir de *adapters* e *primers*. Os indicadores com *'Fail'* são esperados, pois a própria coleta pode gerar consequências indesejadas [16]. Dessa forma, o teste de qualidade informou indicadores sobre as amostras.

A figura 1 abaixo apresenta a qualidade das bases por sequência de uma das leituras da amostra SRR1701090. De acordo com a posição dos nucleotídeos, a qualidade da maioria das bases concentra-se entre 0 e 20 Phred. Dessa forma, a maior parte dos dados dessa amostra são representados na região vermelha e laranja do gráfico. A análise destes parâmetros permite identificar que a qualidade por sequência de bases está baixa, diferenciando-se dos maiores pontos esperados.

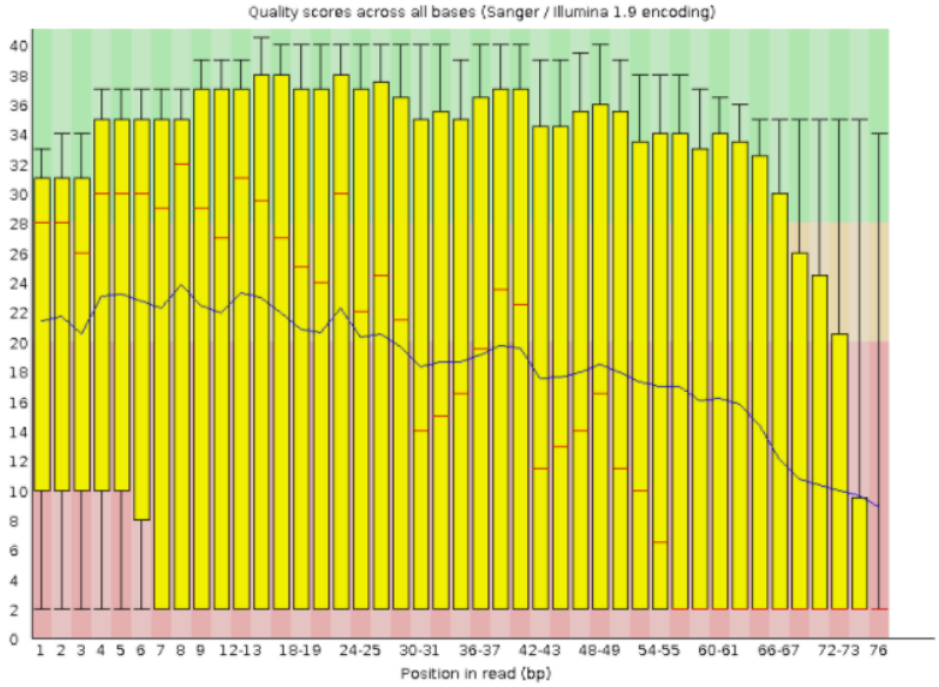

Figura 1: Qualidade por base das leituras do Relatório FastQc antes da limpeza. A amostra corresponde ao ID `SRR1701090_2.fastq`.

3.2 Limpeza de dados

Com base nas *flags* com *'Warning'* ditadas pelo relatório de qualidade, foi realizada a limpeza dos dados pelo software `Trimmomatic` [7]. Para padronizar o processo, escolheu-se seguir os mesmos parâmetros para todas as amostras. Desse modo, segundo a própria documentação do programa, bases com qualidades baixas (menor que Phred 3) foram retiradas do início e final das leituras. Além disso, removeu-se leituras com tamanhos curtos (meno-

res que 36 pares de bases) e sequências correspondentes a contaminantes comuns (*adapter TrueSeq*). Ao final desta etapa, foram gerados 4 produtos para cada 2 leituras `paired.end` das amostras: *paired forward* (_1) e *reverse* (_2), *unpaired forward* (_1) e *reverse* (_2). Os produtos *forward* correspondem às leituras feitas a partir do *primer forward*, enquanto as *reverse* são leituras feitas do *primer reverse*. Por sua vez, os formatos *paired* e *unpaired* são respectivamente, casos onde ambas as leituras de um mesmo transcrito sobreviveram à limpeza, e casos onde apenas uma das leituras sobreviveu. Dado o fato de que apenas uma insignificante minoria de transcritos ficaram na categoria *unpaired*, somente os resultados pareados foram escolhidos para dar-se continuidade.

Para concluir a análise da qualidade dos dados, eles passaram novamente pelo `FastQC`, onde gerou-se mais um novo relatório para cada uma das 70 amostras. O resultado da maioria deles mostrou-se aperfeiçoado. Os avisos *'Fail'* diminuíram consideravelmente, englobando menos de 40% das amostras. Mesmo que algumas amostras ainda apresentem *red flags*, o processo de limpeza já foi realizado, o que significa que, a partir deste momento, possíveis problemas são provenientes da amostra propriamente dita. Além disso, as porcentagens de sequências contaminadas e bases com pareamento irregular também subiram para níveis *'Warning'* e *'Passed'*. Isto mostra que o processo de limpeza teve resultados positivos.

Em relação a amostra SRR1701090, ao analisar a qualidade por sequência de bases após o software `Trimmomatic`, notou-se uma considerável diminuição da concentração dos dados na região vermelha do gráfico, como pode ser visualizado nna figura 2 . Dessa forma, os pontos passam a se localizar em regiões entre 20 e 40 Phred, deslocando o resultado encontrado no `FastQC` anterior. A amostra em questão foi selecionada pois ela permite diferenciar as estatísticas básicas antes e depois da limpeza com clareza. Assim, percebe-se que até mesmo a mediana dos *boxplots* na figura 2 gira em torno de 35 pontos, corroborando a necessidade de aumentar sua qualidade dos dados antes do mapeamento.

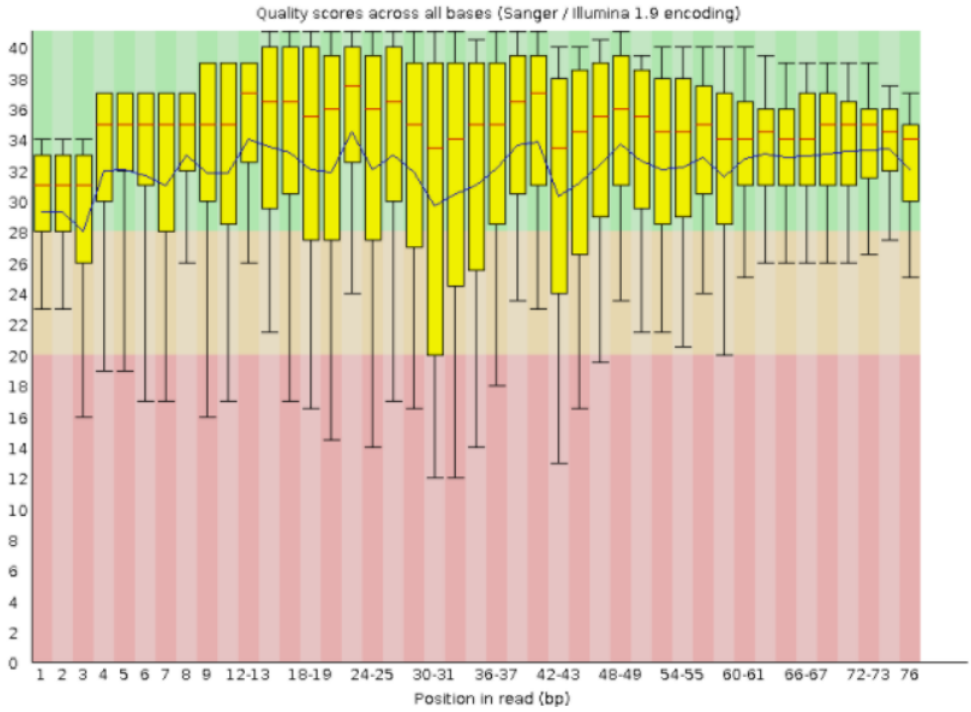

Figura 2: Qualidade por base das leituras do Relatório FastQc depois da limpeza. A amostra corresponde ao ID `SRR1701090_2.fastq`.

3.3 Mapeamento

Desse modo, dois arquivos (`paired_1` e `paired_2`) são os inputs para o mapeamento de cada amostra. Mediante o estudo da documentação do `HISAT2`, descobrimos que, além das sequências de RNA, também é necessário um genoma de referência indexado. O próprio programa disponibiliza-o pré indexado. Como o estudo trabalha com amostras humanas, nas quais esperamos encontrar e buscar possíveis diferenças no nível de transcrição, informações como: marcadores genéticos (SNP) e anotações de transcritos, são importantes para complementar os resultados. Assim, o genoma de referência escolhido foi '`Grch38_snp_tran`'. O mapeamento alinhou as amostras de RNA com o genoma, o que resulta na posição dos transcritos em relação à referência. Como *output* (saída/resultado), um arquivo será gerado para cada par de amostras `paired.end`, de modo a resultar em 35 arquivos. Os 35 dados gerados no mapeamento são encontrados em formato `.sam`, porém, para facilitar os próximos passos para o computador, recomenda-se a conversão dos mesmos para o formato `.sam`, ou seja, binário. Ademais, para cada processo de mapeamento, é dada uma taxa de alinhamento, que corresponde à quantidade de leituras que foram mapeadas com o genoma de referência. Neste caso, a taxa tende a ser maior que 90%, pois se tratam de amostras da mesma espécie.

Como pode ser visto na figura 3, somente uma das amostras se encontra com porcentagem menor que 90%, ou seja, em torno de 87%. Porém, o restante dos dados concentra-se

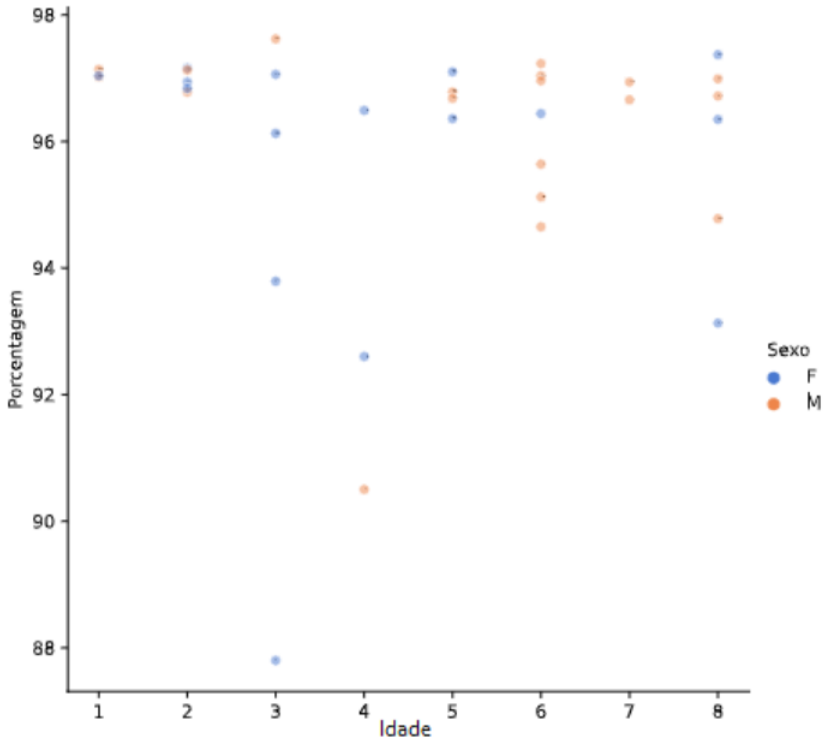

Figura 3: Taxa de mapeamento das amostras em relação ao sexo e idade.

em regiões com taxas de alinhamento maiores que 96%. Por meio de análises estatísticas, foi possível verificar que a média equivale a 95,91%. Sobre a divisão de sexos, 16 amostras são do sexo feminino, enquanto 19 são do sexo masculino. Em relação à idade, os dados concentram-se nas idades 6 e 8, de forma que a idade com menos representantes seriam de pacientes com 1, 4 e 7 anos.

3.4 Contagem de transcritos

Para a contagem de transcritos, o *software* `StringTie` [17] [18] [19] foi instalado e utilizado. Primeiramente, foi preciso realizar uma etapa de Montagem, ou seja, as sequências de RNA mapeadas tiveram suas definições estruturais montadas em relação a um genoma de referência (`Homo_sapiens.GRCh38.84.gtf`). Por fim, 35 arquivos são gerados neste momento. À vista disso, o próximo passo, realizado também pelo `StringTie`, recebe uma lista com todos os produtos gerados na **Montagem**. Esta fase é chamada de **União**, pois irá unir todos arquivos em uma lista com conjunto não redundante de transcritos. Assim, as especificações das sequências de cada amostra são geradas e, posteriormente, reunidas em uma única lista. Juntamente com o genoma de referência, as amostras mapeadas e a lista de transcritos, pode-se seguir para o próximo passo. Da mesma forma, o `StringTie` esteve presente, no entanto, de acordo com os parâmetros expressos na linha de comando, conseguiu-se estimar a expressão de acordo com o genoma de referência para cada uma das

35 amostras, o que acarretou nomear este processo de **Estimativa**.

Gene ID	SRR				
	1701090	170110	1701117	1701137	1701153
MSTRG.11|MIB2	84	1166	321	3210	857
MSTRG.113	13	22	0	383	182
MSTRG.112|RP11-345P4.9	26	501	58	1164	222
MSTRG.114|MMP23B	3	20	59	77	17
MSTRG.114	203	1454	164	1599	3422
MSTRG.115|CDK11B	224	810	266	1869	5195
MSTRG.115	0	0	0	4805	32
MSTRG.117	38	1875	159	6467	4972
MSTRG.117|SLC35E2B	357	3	526	978	1496
MSTRG.120	0	0	0	0	1588

Tabela 1: Contadores dos genes por amostra. Foi selecionado apenas 10 genes de 5 amostras.

O próximo passo seria uma etapa intermediária para a utilização do `DESeq2`, *software* que testa a expressão diferencial com base em um modelo usando a distribuição binomial negativa. Durante este trabalho, devido a um curto espaço de tempo (apenas 6 semanas), não foi possível analisar as sequências de RNA via `DESeq2`. No entanto, a matriz de leituras mapeadas com um genoma de referência - *input* do `DESeq2` - foi gerada. Para isso, este passo, conhecido como **prepDE** (preparação para o `DESeq2`) [19], utilizou 35 arquivos 'GTF' do processo anterior, produzindo duas matrizes: uma com os contadores de genes expressos por amostras e outra com os contadores de transcritos expressos por amostras. As tabelas foram reduzidas a apenas 5 dados, 10 genes e 10 transcritos, isso pois os dados completos apresentam tamanho elevado, impossibilitando a visualização das mesmas no presente documento. As Tabelas 1 e 3.4 apresentam o conjunto de dados selecionado, porém, ambas completas podem ser vistas no *link* de suas respectivas legendas.

Por conseguinte, todas as etapas discutidas acima constituem a pipeline do projeto desenvolvido. Isso pois, é formada por uma cadeia de processamentos, com funções e processos variados, onde a saída de um procedimento é a entrada para outro.

Transcripts ID	SRR				
	1701090	170110	1701117	1701137	1701153
ENST00000416931	2465	22137	1569	36378	27857
ENST00000457540	9066	186458	13360	244346	200996
ENST00000617238	1	2	1	8	8
ENST00000414273	17586	175755	26918	442951	259423
ENST00000427426	1575	14942	8547	40810	18116
MSTRG.25.1	57	0	553	17787	0
MSTRG.25.2	1539	26205	110	1461	10005
ENST00000621981	1	1	1	1	1
ENST00000514057	14904	111068	22053	277945	134169
ENST00000416718	4684	40966	5563	110108	46819

Tabela 2: Contadores dos transcritos por amostra. Foi selecionado apenas 10 transcritos de 5 amostras.

4 Conclusão

O relatório de qualidade gerado com as amostras sem limpeza proporciona o conhecimento dos possíveis problemas das mesmas. Entre eles, pode-se notar baixa qualidade por sequência de bases, baixos pontos de sequências por base, sequências super representadas e conteúdos de adaptadores. Como explicitado, este resultado é esperado em quaisquer amostras coletadas. Por isso, a necessidade de gerar relatórios e descobrir a qualidade dos dados. Haja vista os problemas identificados, estes são excluídos das amostras. Assim, escolheu-se retirar bases de baixa qualidade no início e fim das leituras, tal como cortar leituras correspondentes a contaminantes e leituras curtas, com menos de 36 pares de bases. Como comprovação do processo de higienização dos dados, outro relatório de qualidade foi produzido. Percebe-se a nítida melhora na qualidade por sequência de bases, pois há uma mudança de avisos *'Fail'* para *'Warning'* ou *'Passed'*, como é o caso da amostra selecionada como exemplo. A porcentagem de sequências super-representadas diminui consideravelmente e o conteúdo das bases se mostrou similar ao regular. Visto isso, pode-se considerar que o processo de limpeza das amostras cumpriu com o esperado, trazendo melhorias para a qualidade das mesmas.

O processo de mapeamento, com o intuito de localizar a posição dos transcritos, rendeu uma taxa de porcentagem maior que 90% para a maioria das amostras. A média foi de 95,91%, o que prova compatibilidade das amostras com o genoma de referência. Resultado esperado, dado que ambos são genomas humanos.

O produto do mapeamento é contabilizado pelos processos de **Montagem**, **União** e **Estimativa**. Ao final de cada um deles, gera-se arquivos 'gtf', lista de transcritos e outros

arquivos 'gtf' com estimativas de transcritos, respectivamente. Assim, calculou-se o nível de transcritos para as 35 amostras. A contagem de transcritos tem seu fim com o passo chamado de **prepDE** (preparação para o `DESeq2`), que irá resultar em matrizes com os contadores para genes e transcritos. Estes contadores são específicos para cada amostra de genes e para cada um dos dados do Hospital.

Por conseguinte, com o apoio do Centro Infantil Boldrini, que forneceu os dados sobre os casos de osteossarcoma, juntamente com o auxílio do CENAPAD, foi possível analisar as sequências de RNA. *Softwares* de bioinformática foram utilizados, como `FastQC`, `Trimmomatic`, `HISAT2`, `SAMtools` e `StringTie`, que promoveram a geração de relatórios de qualidade, limpeza de dados, mapeamento, conversão e, finalmente, contagem de transcritos. Para que tudo isso fosse realizado, conhecimentos em programação foram necessários, o que promoveu o aprendizado de diversas linguagens e documentos a nível introdutório em computação. De um modo geral, o programa de estágio no Boldrini (PEOp 2022) permitiu conhecer a rotina de um Centro de Pesquisa, bem como dos pesquisadores. Ademais, garantiu o desenvolvimento de técnicas em biologia, programação e bioinformática. Dessa forma, criou-se vínculos com o conteúdo estudado, o que abriu portas para possíveis trabalhos futuros [2].

Agradecimentos

Esse trabalho usou recursos do "Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD-SP)." Agradecemos ao Centro Infantil de Investigações Hematológicas "Dr Domingos A Boldrini" e à sua equipe pela disponibilização de dados e infraestrutura. Agradecemos especialmente à Dra Mariana Maschietto por valiosas discussões e sugestões.

Referências

[1] *Centro Nacional de Processamento de Alto Desempenho em São Paulo*. URL: `https://www.cenapad.unicamp.br` (acesso em 24/04/2022).

[2] Jennifer A. Perry et al. "Complementary genomic approaches highlight the PI3K/mTOR pathway as a common vulnerability in osteosarcoma". en. Em: *Proc Natl Acad Sci USA* 111.51 (dez. de 2014), E5564–E5573. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.1419260111`. URL: `http://www.pnas.org/lookup/doi/10.1073/pnas.1419260111` (acesso em 08/02/2022).

[3] Eric W Sayers et al. "Database resources of the national center for biotechnology information". en. Em: *Nucleic Acids Research* 50.D1 (jan. de 2022), pp. D20–D26. ISSN: 0305-1048, 1362-4962. DOI: `10.1093/nar/gkab1112`. URL: `https://academic.oup.com/nar/article/50/D1/D20/6447242` (acesso em 24/04/2022).

[4] NCBI Sequence Read Archive. *Osteosarcoma Genomics*. 2014. URL: `https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP050453&o=assay_type_s%3Aa%253Bacc_s%3Bacc_s%3Aa.` (acesso em 22/10/2021).

[5] Babraham Bioinformatics. *FastQC*. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (acesso em 10/01/2022).

[6] Karl W. Kroll et al. "Quality Control for RNA-Seq (QuaCRS): An Integrated Quality Control Pipeline". en. Em: *Cancer Inform* 13s3 (jan. de 2014), CIN.S14022. ISSN: 1176-9351, 1176-9351. DOI: 10.4137/CIN.S14022. URL: http://journals.sagepub.com/doi/10.4137/CIN.S14022 (acesso em 24/04/2022).

[7] Anthony M. Bolger, Marc Lohse e Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". en. Em: *Bioinformatics* 30.15 (ago. de 2014), pp. 2114–2120. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btu170. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu170 (acesso em 24/04/2022).

[8] Mihaela Pertea et al. "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown". en. Em: *Nat Protoc* 11.9 (set. de 2016), pp. 1650–1667. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/nprot.2016.095. URL: http://www.nature.com/articles/nprot.2016.095 (acesso em 24/04/2022).

[9] Yun Zhang et al. "Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N". Em: *Genome Research* 31.7 (2021), pp. 1290–1295. URL: https://daehwankimlab.github.io/hisat2/.

[10] Dan Brown e Burkhard Morgenstern, ed. *Algorithms in Bioinformatics: 14th International Workshop, WABI 2014, Wroclaw, Poland, September 8-10, 2014. Proceedings*. en. Vol. 8701. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. ISBN: 978-3-662-44753-6. DOI: 10.1007/978-3-662-44753-6. URL: http://link.springer.com/10.1007/978-3-662-44753-6 (acesso em 24/04/2022).

[11] Ali Mortazavi et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". en. Em: *Nat Methods* 5.7 (jul. de 2008), pp. 621–628. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1226. URL: http://www.nature.com/articles/nmeth.1226 (acesso em 24/04/2022).

[12] Guangzheng Wen. "A Simple Process of RNA-Sequence Analyses by Hisat2, Htseq and DESeq2". en. Em: *Proceedings of the 2017 International Conference on Biomedical Engineering and Bioinformatics - ICBEB 2017*. Bangkok, Thailand: ACM Press, 2017, pp. 11–15. ISBN: 978-1-4503-5297-0. DOI: 10.1145/3143344.3143354. URL: http://dl.acm.org/citation.cfm?doid=3143344.3143354 (acesso em 24/04/2022).

[13] H. Li et al. "The Sequence Alignment/Map format and SAMtools". en. Em: *Bioinformatics* 25.16 (ago. de 2009), pp. 2078–2079. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btp352. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352 (acesso em 24/04/2022).

[14] Mihaela Pertea et al. "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads". en. Em: *Nat Biotechnol* 33.3 (mar. de 2015), pp. 290–295. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3122. URL: http://www.nature.com/articles/nbt.3122 (acesso em 24/04/2022).

[15] Research Technology Support Facility. *FastQC Tutorial & FAQ*. Michigan State University. URL: https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/ (acesso em 20/01/2022).

[16] Rafael S. Gonçalves e Mark A. Musen. "The variable quality of metadata about biological samples used in biomedical experiments". en. Em: *Sci Data* 6.1 (mar. de 2019), p. 190021. ISSN: 2052-4463. DOI: 10.1038/sdata.2019.21. URL: http://www.nature.com/articles/sdata201921 (acesso em 24/04/2022).

[17] Francesco Crea et al. "Integrated analysis of the prostate cancer small-nucleolar transcriptome reveals SNORA55 as a driver of prostate cancer progression". en. Em: *Molecular Oncology* 10.5 (mai. de 2016), pp. 693–703. ISSN: 15747891. DOI: 10.1016/j.molonc.2015.12.010. URL: http://doi.wiley.com/10.1016/j.molonc.2015.12.010 (acesso em 24/04/2022).

[18] Sam Kovaka et al. "Transcriptome assembly from long-read RNA-seq alignments with StringTie2". en. Em: *Genome Biol* 20.1 (dez. de 2019), p. 278. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1910-1. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1910-1 (acesso em 24/04/2022).

[19] Johns Hopkins University Center for Computational Biology. *StringTie: Transcript assembly and quantification for RNA-Seq*. URL: http://ccb.jhu.edu/software/stringtie/ (acesso em 28/01/2022).