Algoritmos Avançados de Bioinformática

Trabalho Prático 1

Processamento de Dados NGS

Relatório de Desenvolvimento

Grupo 6

Eduardo Cunha
A71940

João Dias
PG37982

Junho 2019

Capítulo 1

Introdução

1.1 Contextualização

Neste trabalho de Algoritmos Avançados de Bioinformática, era necessário efetuar o processamento de dados NGS relativos ao cancro da mama.

Assim, estava enunciado a utilização de diversas ferramentas para este processamento e posterior análise dos datasets, como a avaliação dos respetivos datasets com o *FastQC*, o ajuste dos reads com o *Trimmomatic*, o alinhamento dos reads com o genoma ou simplesmente alguns cromossomas com o *BWA* e o cálculo do valor RPKM para cada gene com o *HTSeq*.

Por fim, pretende-se efetuar a análise da expressão diferencial com o auxílio do *R* e fazer a discussão de todos os resultados obtidos anteriormente.

1.2 Estrutura do Relatório

O relatório encontra-se dividido em três principais capítulos, sendo um deles introdutório, um capítulo referente ao desenvolvimento do projeto, com a explicação de todas as ferramentas utilizadas e discussão de todos os resultados obtidos e um capítulo conclusivo.

No capítulo de desenvolvimento está exposta a forma como abordamos este trabalho, desde a explicação de todos os comandos efetuados, juntamente com a análise dos dados obtidos de cada um, até à comparação final entre os resultados alcançados pelo nosso grupo com os resultados de artigos que desenvolveram trabalhos relacionados com este.

Por fim, no capítulo três, é apresentada uma comparação dos resultados obtidos com os resultados de artigos publicados.

Capítulo 2

Processamento e Análise de Dados

2.1 FastQC

O *FastQC* é um programa que permite o controlo de qualidade dos dados sendo criado um relatório com a análise de dados sob a forma de diferentes gráficos. Inicialmente, no âmbito deste trabalho, foi executado para todos os 7 datasets disponíveis, relativos ao cancro da mama, o seguinte comando:

```
fastqc /NGSDatasets/reads/breast/IlluminaHiSeq2000/homosapiens/*.fastqc.gz /home/group6
```

Desta forma, foram criados um ficheiro *Html* e um ficheiro *Zip* para cada dataset, com a respetiva análise dos dados.

2.1.1 Resultados

Primeiramente analisando os relatório do fastqc, pode-se denotar várias métricas que permitem avaliar as várias samples. No geral todas as reads tiveram valores aceitáveis em todas as métricas, com exceção do conteúdo de par de bases em cada sequência. No primeiro ponto podemos verificar o tamanho das reads e dos fragmentos e outras informações gerais.
Na *Per base sequence quality* é permitido ver a qualidade das sequências. Na *Per tile sequence quality* é possível ver a qualidade dos reads na sequência. Na *Per sequence quality scores* dá para ver os scores da sequência. Na *Per sequence GC contente* podemos ver o conteúdo de GC na sequência. Na *Per base N content* podemos ver o conteúdo de N na sequência. Na *Sequence Length Distribution* podemos ver a distribuição do comprimento das sequências. Na *Sequence Duplication Levels* podemos ver se há duplicação de sequências, sendo que neste caso, não existem duplicação. Na *Overrepresented sequences* podemos ver se há sequencias overexpressed, o que não existe, neste caso. Na *Adapter Content* podemos averiguar se há adaptadores nas sequências, não existindo nenhum, de acordo com os resultados obtidos.

2.2 Trimmomatic

O *Trimmomatic* é uma ferramenta que permite limpar os primers dos reads presentes nos ficheiros gerados anteriormente. Tendo em conta que o sequenciador utilizado pelo *FastqQC* foi o *IlluminaHiSeq2000*, existiam dois ficheiros *fasta* associados a este sequenciador, *TruSeq3-PE-2.fa* e *TruSeq3-SE*. Assim, foram executados os seguintes comandos para a execução do *Trimmmomatic*, para cada dataset em questão:

```
trimmomatic SE *.fastqc.gz ILLUMINACLIP:TruSeq3-PE-2.fa:30:10 LEADING:3
TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50

trimmomatic SE *.fastqc.gz ILLUMINACLIP:TruSeq3-SE.fa:30:10 LEADING:3
TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50
```

De notar que todos os parâmetros possuem os valores default de execução deste programa, com a exceção do campo *MINLEN* que foi alterado para 50, de acordo com os artigos encontrados que tratam este assunto.

2.3 BWA

A terceira tarefa desta trabalho consiste na execução do programa *BWA*, que permite alinhar os reads resultantes, com um genoma de referência.

No âmbito deste trabalho, foi efetuado o alinhamento com 3 cromossomas diferentes, nomeadamente os cromossomas 2, 11 e 17. A escolha destes cromossomas baseou-se no estudo da literatura encontrada sobre a análise desta linha celular. Por falta de tempo, não nos foi possível efetuar o alinhamento com todo o genoma humano.

Inicialmente, para cada amostra, foi efetuado o alinhamento com os cromossomas selecionados, sendo guardado num ficheiro com a extensão *sai*, através dos seguintes comandos:

```
for fq in ls *.fastq.gz; do echo $fq;
bwa aln Homo_sapiens.GRCh38.dna.chromosome.2.fa.gz $fq> "2_"$fq".sai"; done

for fq in ls *.fastq.gz; do echo $fq;
bwa aln Homo_sapiens.GRCh38.dna.chromosome.11.fa.gz $fq> "11_"$fq".sai"; done

for fq in ls *.fastq.gz; do echo $fq;
bwa aln Homo_sapiens.GRCh38.dna.chromosome.17.fa.gz $fq> "17_"$fq".sai"; done
```

Seguidamente, para cada cromossoma, e cada amostra, foi efetuado o alinhamento *single ended*, sendo guardado diretamente no formato *bam*. Abaixo está demonstrado o comando efetuado, como exemplo de um alinhamento, sendo que os restantes se fizeram da mesma forma.

```
bwa samse Homo_sapiens.GRCh38.dna.chromosome.17.fa.gz 17_SRR6638905.fastq.gz.sai
SRR6638905.fastq.gz | samtools view -S - -b > SRR6638905_17.bam
```

2.3.1 Cromossomas analisados

A escolha dos genes foi feita através da consulta de vários artigos onde foram analisados diferentes genes, proteínas e as suas expressões. A escolha do cromossoma 2 deve-se à leitura do artigo " LncRNA BLAT1 is Upregulated in Basal-like Breast Cancer through Epigenetic Modifications" , onde os autores analisam a presença da sobre expressão do gene BLAT1, que se localiza no cromossoma 2.

A escolha do cromossoma 11 deve-se à leitura do resumo do nosso dataset onde os autores fazem referência ao gene TENM4, que localiza no cromossoma 11, e que este se encontra sobre expresso na linha celular de triple negative breast cancer.

Por fim, a escolha do cromossoma 17 deveu-se ao artigo "Epigenetic and transcriptional profiling of triple negative breast cancer", onde os autores fazem referência à Pol II e aos seus resultados.

2.4 HTSeq

O *HTSeq* é um package do Python cujo objetivo é o processamento de dados de sequenciamento. Desta forma, no âmbito deste projeto, o HTSeq foi utilizado com o objetivo de alinhar o ficheiro *bam* resultante do passo anterior, com o ficheiro *gtf* do ser humano, sendo guardado no respetivo ficheiro.

```
htseq-count -f bam -s no -i gene_id SRR6638905.bam Homo_sapiens.GRCh38.96.chr.gtf.gz > C2_R1.readcounts
```

Seguidamente, foi criado um ficheiro *tab* que tem como primeira linha identificação de cada coluna e as restantes linhas consistem na junção de todos os ficheiros *readcounts* criados no passo anterior. Foram utilizados os seguintes comandos:

```
echo "geneid C2_R1 C2_R2 C2_R3 C2_R4 C2_R5 C2_R6 C11_R1 C11_R2 C11_R3 C11_R4 C11_R5 C11_R6 C17_R1
C17_R2 C17_R3 C17_R4 C17_R5 C17_R6" > C2C11C17.tab
```

```
paste *.readcounts | cut -f1,2,4,6,8,10,12,14,16,18,20,22,24,26,28,30,32,34,36 >> C2C11C17.tab
```

Depois das várias analises do dataset, o ficheiro resultante foi o "C2C11C17.tab". Neste ficheiro estão contidos os resultados dos *HTSeq*, ou seja, as várias reads para cada amostra (SRR6638905_11.bam, SRR6638906_11.bam, SRR6638907_11.bam, etc) e para cada cromossoma avaliado (2,11,17).

2.4.1 Cálculo do RPKM

Relativamento ao cálculo do RPKM de cada gene, não nos foi possível finalizar esta fase por falta de tempo. No entanto, a ideia consistiria em, através do ficheiro *gtf* já existente, calcular o tamanho do gene através da subtração entre a quinta e a quarta coluna desse ficheiro. Posteriormente, através do ficheiro *tab* resultante do processamento efetuado, somar todos os valores, correspondentes a cada gene, sendo multiplicado por 10^6 (um milhão). Por fim, esse valor seria dividido pelo tamanho do gene em questão.

2.5 R

O *R* é uma linguagem de programação que permite a manipulação, análise e visualização gráfica de dados em diferentes formatos. Para este trabalho, foi utilizado com a finalidade de análise de expressão diferencial dos resultados alcançados nas fases anteriores deste trabalho. Após a criação do ficheiro *tab*, foi feita a análise dos dados no R, de onde resultou um pheatmap onde é possível ver os primeiros 20 genes (por exemplo ENSG00000214391) e a sua expressão em relação às condições (cromossomas c2, c11, c17).

Todo o código criado e análise efetuada estão no ficheiro RMarkdown, juntamente com a discussão dos resultados obtidos nesta fase.

2.6 Resultados

Foram selecionados os primeiros 5 genes (de acordo com a sua média), para uma melhor análise dos dados existentes. Comparando com a base de dados presente no *ensemble*, pode-se chegar aos nomes dos genes correspondentes a cada id.

Primeiramente, temos TUBAP2 (ENSG00000214391) presente no cromossoma 11. Este gene é um pseudogene da tubulina alpha. Apesar de não haver grande informação para este pseudogene, a família das tubulinas costuma ter alta expressão em vários tipos de cancro, o que vai em concordância com os nossos resultados.

O gene ACTG1 (ENSG00000184009) está presente no cromossoma 17. Este gene é responsável por uma actina citoplasmática. As proteínas resultantes deste gene, estão altamente expressos em casos de breast cancer, o que vai em concordância com os nossos resultados.

O gene EIF4G2 (ENSG00000110321) está presente no cromossoma 11. Este gene é responsável por fatores iniciais de transcrição e que normalmente serve de repressor da transcrição. As proteínas resultantes deste gene estão altamente expressas em casos de breast cancer, o que também vai em concordância com os nossos resultados.

O gene DDX5 (ENSG00000108654) está presente no cromossoma 17. Este gene é responsável por várias funções celulares como, por exemplo, na iniciação da tradução, splicing mitocondrial, entre outras. As proteínas desta resultante deste gene estão altamente expressas em casos de breast cancer.

Por fim, temos o AHNAK (ENSG00000124942) está presente no cromossoma 11. Este gene é responsável por várias funções como formação da "blood-brain barrier" ou estrutura da célula. As proteínas resultantes deste gene estão altamente expressas em casos de breast cancer, o que vai em concordância com os nossos resultados.

Capítulo 3

Discussão dos Resultados

Por fim, é feita a comparação dos resultados dos artigos com os nossos resultados.

Em primeiro lugar, no artigo para cromossoma 2, os autores fazem referência para o gene BLAT1 (ENSG00000258910) e para a sobre expressão deste. Comparando com os nossos resultados, podemos averiguar que o gene não aparece na expressão dos resultados do R. Isto pode-se dever ao facto da falta de reads, mudanças de parâmetros entre outras causas. Assim, seria necessário mais análises e mais amostras para poder tirar mais conclusões sobre a expressão do gene neste tipo de breast cancer.

No cromossoma 11, o artigo faz referência ao gene TENM4 (ENSG00000149256). Comparando com os nossos resultados, pode-se concluir que o gene se encontra expresso, mas em poucas quantidades, contrariando a análise feita pelos autores. Esta discrepância pode surgir devido ao facto da análise feita pelos alunos e pelos autores do trabalho ser diferente em termos de parâmetros, entre outros aspetos. Assim, pode haver variância na comparação do nível de expressão. Para além disso foram utilizados parâmetros a escolha dos alunos nos vários passos de processamento dos dados o que pode fazer com os resultados variem. No entanto o gene encontra-se expresso, mas não aparece como sobre expresso.

Por fim no cromossoma 17, o artigo faz referência ao POL II (ENSG00000099817). No nosso caso a expressão deste gene não foi denotada nos resultados. Isto pode surgir devido a diferenças de parâmetros, diferenças de processamento, entre outras. Seriam necessárias mais amostras, réplicas e métodos específicos para comparar resultados e poder tirar novas conclusões.