Week 1

Hien Le

14/01/2022

Excercise week 1: RNA-seq quantification

The exercise set includes three samples from benign prostatic hyperplasia (BPH) and three samples of primary prostate cancer (PC).

Assessing sample quality

An overall evaluation of samples quality before performing further analysis.

Tasks

Tasks 1

(*) Use the fastqc tool to generate quality reports for the reads from the six samples we are using in this exercise. Save the results in the qc folder.

```
fastqc ./data/BPH_659.chrX_R1.fq.gz -o ./qc/
fastqc ./data/BPH_665.chrX_R1.fq.gz -o ./qc/
fastqc ./data/BPH_701.chrX_R1.fq.gz -o ./qc/
fastqc ./data/PC_13943.chrX_R1.fq.gz -o ./qc/
fastqc ./data/PC_15194.chrX_R1.fq.gz -o ./qc/
fastqc ./data/PC_19403.chrX_R1.fq.gz -o ./qc/
```

Questions

Question 1: Inspect the fastqc output for the forward sequencing reads of sample PC_19403? The results: PC_19403.chrX_R1.fq.gz FastQC Report

Are failures reported for any of the analysis modules? Per base sequence quality and sequence duplication levels are reported as a failure. Per base sequence content, per sequence GC content, and overrepresented sequences are reported as warning qualities.

If so, what do these modules tell you about the sequencing reads (i.e. what is wrong with the quality of the reads precisely)? As failure: Per base sequence quality: a plot of the total number of reads in a comparison with the average quality score over full length of that read. In here, at the ending of read (54-55 bp), the quality of read decreases and a signification decreasing observation is belonged at the end of reads. Sequence duplication levels: a low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias. In here, the percentage of seqs remaining if deduplicated 27.12%. It is the quite low percentage of deduplication. Two sources of

duplicate reads can be found: - PCR duplication in which library fragments have been over-represented due to biased PCR enrichment. It is a concern because PCR duplicates misrepresent the true proportion of sequences in the input. - Truly over-represented sequences such as very abundant transcripts in an RNA-Seq library or in amplicon data (like this sample). - It is an expected case and not of concern because it does faithfully represent the input.

Can you suggest how the reads could (in theory) be pre-processed to improve their quality? We know that Fastqc lists all of the sequence which make up more than 0.1% of the toal. For each over-represented sequence fastqc will look for matches in a databased of common contaminants and will report the best hit it find. Hits must be at least around 20bp in length and have no more than 1 mismatch. Finding a hit or error is unable to conclude that the samples exist the contamination, by may point us in the right direction for further analysis. I think that it would be better if we arrange the sequences and read the sequence with the shorter reads.

Question 2 Should duplicate reads be removed from RNA-sequencing data? Why or why not? I think that NO. You are checking for differential expression of genes, number of reads mapping to a gene is a measure of its expression. By removing reads, you will bias the true expression measurements. In addition, RNA-seq often has the long-gap in the results, therefore, we can choose another solution for analyzing RNA-seq.

Quantification with an alignment based approach

Short read alignment

Hisat2 is a software for the alignment of RNA sequencing data against a reference's genome. To use Hisat2, we need two things: - Index of a reference genome - RNA sequencing data in FASTQ format Hisat2 will take the short reads and align them against the reference genome, figuring out exon-exon splice junctions along the way. The end results is map showing what parts of the genome are transcribed into RNA inside the cells and how the RNA molecules are structured.

Index generation

Command: hisat2 -build Reference genome: references/fasta/GRCh38.p13.chrX.fa

Tasks

1. Generate the index files using hisat2 -build including index aware of splicing junction, exons, haplotype, and SNPs.

Make index aware of splicing junctions, exons, haplotype, and SNPs.

```
hisat2-build \
  -f ../../references/fasta/GRCh38.p13.chrX.fa \
  --exon ../../references/gencode.v32.exons.tsv \
  --ss ../../references/gencode.v32.ss.tsv \
  --haplotype ../../references/snp144Common.haplotype \
  --snp ../../references/snp144Common.snp GRCh38_p38_chrX_index
```

Questions

Question 1

What was the meaning of each parameters you used in the indexing command? hisat2-build command was used with 5 options: -f, –exon, –ss, –snp, and –haplotype. -f : our reference genome is stored in .fa file –snp SNA file name –exon Exon file name –ss Splice site file name –haplotype haplotype file name

Question 2

How many file constitute the Hisat2 index? 8 files were created by hisat2-build command

Alignment

Tasks

1. Use the hisat2 command to perform the alignment. Save the results in hisat2/alignments/.

```
for f in $(find /student_data/BBT.BI.203_2022/data/ex1/ -type f -name '*.chrX_R1.fq.gz'); do
  echo $f;
  bn=$(basename $f);
  bam=${bn/_R1.fq.gz/_sorted.bam};
  echo $bn
  echo $bam
  hisat2 -x ../index/GRCh38_p38_chrX_index \
  -1 <(zcat $f) -2 <(zcat ${f/R1/R2}) \
  | samtools view -bS \
  | samtools sort -o $bam - \
  && samtools index $bam;
done
```

2. Convert the output to binary compressed format

```
for f in $(find /student_data/BBT.BI.203_2022/data/ex1/ -type f -name '*.chrX_R1.fq.gz'); do
  echo $f;
  bn=$(basename $f);
  bam=${bn/_R1.fq.gz/_sorted.bam};
  echo $bn
  echo $bam
  hisat2 -x ../index/GRCh38_p38_chrX_index \
  -1 <(zcat $f) -2 <(zcat ${f/R1/R2}) \
  | samtools view -bS \
  | samtools sort -o $bam - \
  && samtools index $bam \
&& gzip $bam;
done
```

3. Sotre only the final BAM files

```
rm -.bai
```

Questions

1. What was the alignment rate, number of unmapped reads, and number of reads mapped to multiple regions? Alignmant rate: This means the total mapping (alignment) ratio of RNA reads to the genome that we used. BPH_665 sample: 95.87% overeall alignment rate BPH_659 sample: 94.95% overall alignment rate BPH_701 sample: 94.86% overall alignment rate PC_13943 sample: 90.92% overall alignment rate PC_19403 sample: 92.21% overall alignment rate PC_15194 sample: 94.31% overall alignment rate

Unmapped reads: means that fail to align to file, path, or multiple regions. For 6 samples, 100% were paired

Number of reads mapped to multiple regions: Read counts and alignments mapped to multiple regions. BPH_665 sample: 1552486 reads BPH_659 sample: 1434722 reads BPH_701 sample: 1475037 reads PC_13943 sample: 1356064 reads PC_19403 sample: 1406716 reads PC_15194 sample: 1215200 reads

2. What was the meaning of each parameter you used? -x : defaul pamater before reference genome index. It indicates that index file names prefix as .x.ht2 -1 : the first file for comparing with the second file -2 : the second file for comparing with the first file zcat : for reading the compressed file withouh decompressing The output of hisat2 is SAM file. For saving the memory, we change SAM file to BAM file. In here, we use samtools view to change SAM to BAM, samtool sort to sort BAM file, and samtool index to Indexing a genome sorted BAM file allows one to quickly extract alignments overlapping particular genomic regions.

Visualization

We will take a look at the alignment results by opening the alignment files in the IGV genome browser. ## Tasks 1. Sort and index all alignment files. Save the files in hisat2/alignments and append _sorted. bam to each file.

```
... samtools sort -o $bam -  && samtools index $bam ...
```

2. Download the sorted BAM and index files. Open the IGV program and inspect the alignment results

Questions

1. How does samtools sort reads by default? Are there other options for how to sort the reads? When we align from FASTQ files with all current sequence aligners, the alignments produced are in random order with respect to their position in the reference genome. In here, samtool sort is to order the sequences occurred from input files.

Are there other options for how to sor the reads? From samtools manula pages, samtools tview could be used. From searching, I have known that sambamba-sort could do the same things as samtool sort.

2. Search the AR gene from the top dropdown menu in IGV. Zoom in to the AR locus. How are the reads distributed on the genome? What kind of genomic features do they cover? Can you explain why? As shown in IGV, AR gene expression is recored in all samples.

3. To show all transcript variants, right click on the "Gene" track at the bottom of the screen and change the view mode to "Squished". What are the names of the AR transcript variants. There have the names of 2 transcript variants: transcription variant 2 and 3.

Counting reads for genes and transcripts

We use featureCounts for generating counts of reads for genes. In general, the classical use case for featureCounts is to quantify the expression of a gene as a sum of the reads mapping to the exons of that gene. Thus, it is not the best option for figuring out the expression of each transcript quantification. A single gene has several transcripts. ## Tasks 1. Take a loot at the GFF annoatation file. GFF is an BED annotation file. It includes 9 columns, and emptying column in GFF file is denoting by dot (.) character, plus optional track definition lines. Meta-feature information in GFF file is in 3rd columns typed as name, e.g. Gene, Variation, Similarity.

2. Count read pairs for annotated genes based on exons

```
featureCounts -g gene_id -t exon -p --countReadPairs \
-a /student_data/BBT.BI.203_2022/references/ex1/gff/gencode.v32.annotation.chrX.gff3 \
-o quantification_genes.bed \
/student_data/BBT.BI.203_2022/students/leh/exercise_1/histat2/alignments/*.bam
```

3. Cound read pairs for annotated transcripts based on exons

```
featureCounts -g transcript_id -t exon -p --countReadPairs \
-a /student_data/BBT.BI.203_2022/references/ex1/gff/gencode.v32.annotation.chrX.gff3 \
-o quantification_transcripts.bed \
/student_data/BBT.BI.203_2022/students/leh/exercise_1/histat2/alignments/*.bam
```

4. Count read pairs for annotated transcripts based on exons with multi-overlapping reads.

```
featureCounts -g transcript_id -t exon -p -O --countReadPairs \
-a /student_data/BBT.BI.203_2022/references/ex1/gff/gencode.v32.annotation.chrX.gff3 \
-o quantification_transcripts_multioverlap.bed \
/student_data/BBT.BI.203_2022/students/leh/exercise_1/histat2/alignments/*.bam
```

Questions

1. Take a look at the output file hisat2/quantification_genes.bed from Task 2. Briefly describe what information the columns of the file contain. How are the features and meta-features shown in the output? This file contains 12 columns with the sequences as: gene_id, Chr, start, end, strand, length, and 6 other columns for 6 samples. Gene_id: shown genes id with signs ENS. Chr: shown which located chromosome Start: shown the starting location End: shown the ending location Strand: shown strand of genes Length: shown the nominal fragment length 6 other colums: shown the number of reads regarind the gene_id in each sample (or we can understand that it is quantification)

2. What do you notice about the percentages of sucessfully assigned alignments in Task 2 and Task3 ? The percentages of succesfully assigned alignments in task 2 are always higher than in taks 3 in every samples.

3. Take a look at the output file hisat2/quantification_transcripts.bed from Task 3.

a. Gene: PLCXD1
b. As shown in IGV, PLCXD1 in BPH_659 sample has the expression, but not too much.
c. Can you suggest why featureCounts has counted zero reads for ENST00000399012.6? Because we do not put the option of overlapping in the running commands, therefore, as the default, the multiple-overlapping will be ingored and set up as 0.

4. Multi-over lapping read Multi-over lapping read is a read that overlaps more than one meta-feature when counting reads at meta-feature level or overlaps more than on feature when counting reads at feature level.

Genes can have multiple transcripts. So gene expression could mean the overal expression of all transcripts of a gene. Transcript expression is the expression of a specific transcript. In RNA-seq, specific transcript expression can be identified. In addition, in RNA-seq, several freagments overlapping will be recorded. Any single fragment must originate from only one of the target genes but the identity of the true target gene cannot be confidently determined.

Will counting multi-overlapping reads result in accurate transcript expression estimates? Why or why not? Yes, we should count mul-overlapping read results because we can use these resutls in a comparision with other methods after adjusting and signing multi-overlappaing. We could asign multi-overlapping points, and count them as 1. If we do not notice the multi-overlapping, the final expression level is not exact. The simplest ways of handling multireads are to ignore them all together, count them once for each alignment, randomly assign them to one of the best alignments or split them equally between each alignment .

What is the read count reported for ENST00000399012.6 for sample BPH_659 when multi-overlapping reads are counted in Task 4? 5287