Introduction to transcriptome sequencing analysis methods

Anamaria Necsulea

March 20, 2024

1 Introduction

In this practical session, we will analyze transcriptome sequencing data from a publication that studied the functions of a mouse long non-coding RNA [1]. The studied long non-coding RNA is named **Hotair**. It was previously proposed that this long non-coding RNA is involved in the regulation of *HOX* genes. These genes, which in vertebrates are organized in 4 genomic clusters (*HOXA*, *HOXB*, *HOXC* and *HOXD*, each containing around 10 protein-coding genes) are developmental transcription factors, important for the establishment of the body plan in bilaterian organisms. *Hotair* is located within the *HOXC* cluster, in antisense with respect to protein-coding genes.

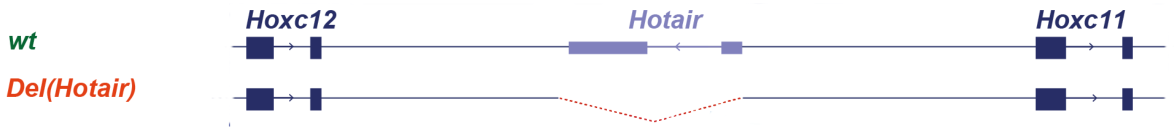

Figure 1: Position of *Hotair* in the mouse genome, on chromosome 15. *Hotair* is transcribed on the reverse strand and is found between *Hoxc11* and *Hoxc12*, which are transcribed on the forward strand. The first line represents the wild type genome configuration. The second line represents the region that was deleted in the genetic manipulation. Image from [1].

To study Hotair functions, a group of researchers generated a knock-out mouse, which includes a deletion of this locus [2] (figure 1). By comparing the transcriptomes of wild-type and mutant mice, the researchers observed an increase in the expression level of *HOXD* genes, and deduced that Hotair is a negative regulator of *HOXD* genes. This analysis was done *in vitro*, on fibroblast cell cultures. To verify whether these conclusions are also valid *in vivo*, another group of researchers analyzed the transcriptomes of wildtye and knock-out mice in embryonic tissues [1]. Several tissues were studied: the forelimbs (FL) and the hindlimbs (HL), the genital tubercle (GT), and three regions of the trunk (T1, T2 and T3). For each tissue, two wild-type and two mutant individuals were sampled. The data was generated with the Illumina TruSeq protocol, single-end, with polyA selection, strand-specific. The data are publicly accessible in the SRA database with the identifier données NCBI SRA avec le numéro d'accession SRP071333 (`https://www.ncbi.nlm.nih.gov/sra?term=SRP071333`). In this practical session, we will analyze forelimb and posterior trunk samples (FL and T3).

Identifier	Tissue	Genotype
FL_wt_1	forelimbs	wild-type
FL_wt_2	forelimbs	wild-type
FL_ko_1	forelimbs	knock-out
FL_ko_2	forelimbs	knock-out
T3_wt_1	posterior trunk	wild-type
T3_wt_2	posterior trunk	wild-type
T3_ko_1	posterior trunk	knock-out
T3_ko_2	posterior trunk	knock-out

Table 1: RNA-seq data used in the practical.

In addition to RNA-seq data, you also provided with the fasta sequence of the genome, the genomic annotations and the cDNA sequences. You can download the data from the following URL:
`http://pbil.univ-lyon1.fr/members/necsulea/MADGT_2024/`

2 RNA-seq data processing and visualisation

1. Using the `FastQC` tool, analyze the quality of the 8 available RNA-seq samples.
2. Using `HISAT2`, for which the manual is available online (`https://daehwankimlab.github.io/hisat2/manual/`), construct a genomic index for the genome sequence that you were given. To construct the index, in addition to the genomic sequence, you can use the known exon and intron coordinates.
3. Using `HISAT2`, align the RNA-seq data on the genome. Note that the RNA-seq library is stranded, and that the reads are coming from the antisense strand of the transcripts. Create an output file in SAM format, with an explicit name.
4. Using `samtools sort`, sort the resulting output file by genomic coordinate.
5. View the data in the IGV navigator. Do you confirm the deletion of the *Hotair* locus in the knockout individuals?
6. How are the splice junctions shown in the IGV navigator? Can you visually identify alternative splicing events around the *Hotair* locus?

3 Gene expression level estimation

1. In R, use the `featureCounts` function in the `Rsubread` package to count the number of reads that align on the exons, for each gene.
2. Analyze the results obtained with `featureCounts`. What are the most highly expressed genes? Do you confirm the deletion of the *Hotair* gene?
3. Compute the TPM expression levels starting from the read counts given by `featureCounts`. The exonic length is provided by feature counts.

4 Gene and transcript annotation

1. Download the binary package of the `StringTie` tool here: `http://ccb.jhu.edu/software/stringtie/dl/stringtie-2.2.1.Linux_x86_64.tar.gz`. Unpack it and add the resulting directory to your `PATH` variable.
2. Using StringTie, annotate genes and transcripts starting from the alignments obtained with HISAT2. The alignments should be sorted by position and in BAM format. Remember that the RNA-seq libraries are strand-specific and that we are sequencing the antisense strand.
3. Merge the resulting annotations with the `stringtie - - merge` command, asking for a minimal exonic length of 200 nucleotides *per* transcript.
4. Import the annotation in IGV. Can you visually detect any differences with the reference annotation?
5. You can also use the GFF utilities here (`https://ccb.jhu.edu/software/stringtie/gff.shtml#gffcompare`) to compare the annotations.
6. Re-compute the expression levels with `featureCounts`. Are the expression levels of the *Hoxc11* and *Hoxc12* genes unchanged between the reference annotation and the new annotation?

References

[1] A. R. Amândio, A. Necsulea, E. Joye, B. Mascrez, and D. Duboule. Hotair Is Dispensible for Mouse Development. *PLoS Genet.*, 12(12):e1006232, December 2016.

[2] L. Li, B Liu, O. L. Wapinski, M Tsai, K. Qu, J. Zhang, M. Lin, F. Fang, R. A. Gupta, J. A. Helms, and H. Y. Chang. Targeted disruption of Hotair leads to homeotic transformation and gene derepression. *Cell Rep*, 5(1):3–12, October 2013.