Pilot-case #1:

Replication of reference genome index files for NGS data analysis

Integrating ELIXIR reference datasets within the European Grid Infrastructure

Contents

[1. Short Description 2](#_Toc421612612)

[2. Data 2](#_Toc421612613)

[Option 1: Pre-build files 2](#_Toc421612614)

[Option 2: Build from source 2](#_Toc421612615)

[*Step 1: Download the reference genome* 3](#_Toc421612616)

[*Step 2: Get gene models annotation* 3](#_Toc421612617)

[*Step 3: Build the reference index* 3](#_Toc421612618)

[3. Applications 3](#_Toc421612619)

[4. Workflow 4](#_Toc421612620)

[*Step 1: Data Annotation per Sample* 4](#_Toc421612621)

[*Step 2: Assembly of the transcripts* 4](#_Toc421612622)

[*Step 3: Final transcriptome assembly* 4](#_Toc421612623)

[*Step 4: Differential Expression results* 4](#_Toc421612624)

[*Step 5: Statistical analysis and visualization* 5](#_Toc421612625)

[References 5](#_Toc421612626)

# 1. Short Description

Next-generation sequencing (NGS) technologies are revolutionizing genome research, and in particular, their application to transcriptomics (RNA-seq) is increasingly being used for gene expression profiling as a replacement for microarrays [1]. The established workflow utilizes reference genomic data in order to annotate the input NGS sequences. The annotated data is consequently analyzed by the Cufflinks tool which is a spliced aligner. It can be used to identify different splice variants from RNA-Seq data, and determine their relative frequencies. Ultimately, Cufflinks allows us to determine differential gene expression between tissues, individuals or disease states.

# 2. Data

There are two options for the obtaining reference index data:

## Option 1: Pre-build files

Use the pre-build indexes provided by Illumina and available through several sites such as [TopΗat](https://ccb.jhu.edu/software/tophat/igenomes.shtml) (*see also Section* ). There are datasets for each of the major species, including Homo Sapiens, as well as based on reference data from the three major data sources: Ensembl, NCBI and UCSC. These files are updated every time a new major version of the reference data is released. Average compressed size of the pre-build index files ranges between 15GB and 21GB, depending on version and source.

## Option 2: Build from source

Reference genome index for bowtie2/tophat2, can be build from the source FASTA files through the following process. This procedure needs to be run only once for each reference genome used. As mentioned, pre-built indices for many commonly used genomes are available as pre-build files (Option #1).

The user needs to download the reference genome sequence for the organism under study in (compressed) FASTA format. This can be done from Ensembl and UCSC databases among many others.

* Ensembl FTP server: http://www.ensembl.org/info/data/ftp/index.html
* UCSC FTP server: ftp://hgdownload.cse.ucsc.edu/goldenPath/currentGenomes/

It is important to note that for Ensembl, choose the "FASTA (DNA)" link instead of "FASTA (cDNA)", since alignments to the genome, not the transcriptome, are desired. The following steps assume that the target species is Homo Sapiens, and the selected source is **Ensembl** (release 80, assembly version 38).

### *Step 1: Download the reference genome*

wget ftp://ftp.ensembl.org/pub/release-80/fasta/homo\_sapiens/dna/Homo\_sapiens.GRCh38.dna.toplevel.fa.gz gunzip Homo\_sapiens.GRCh38.dna.toplevel.fa.gz

### *Step 2: Get gene models annotation*

It is important to note that the FASTA reference sequence and the GTF annotation data must come from the same resource provider.

wget ftp://ftp.ensembl.org/pub/release-80/gtf/homo\_sapiens/Homo\_sapiens.GRCh38.80.gtf.gz gunzip Homo\_sapiens.GRCh38.80.gtf.gz

### *Step 3: Build the reference index*

Before reads can be aligned, the reference FASTA files need to be preprocessed into an index that allows the aligner easy access. To build a bowtie2-specific index from the FASTA file use the command:

bowtie2-build -f Homo\_sapiens.GRCh38.dna.toplevel.fa. Homo\_sapiens\_GRCh38

# 3. Applications

The following is a list of applications required for the whole process:

1. **BowTie2** http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequence. For the human genome, its memory footprint is typically around 3.2 GB.

1. **TopHat2** https://ccb.jhu.edu/software/tophat/index.shtml (also contains Bowtie)

fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

1. **Cufflinks** http://cole-trapnell-lab.github.io/cufflinks/

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.

1. **R** and **BioConductor** http://www.r-project.org/ and http://www.bioconductor.org/

R is a free software environment for statistical computing and graphics. Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. It uses the R statistical programming language, and is open source and open development.

# 4. Workflow

The following procedure employs the reference index files in order to perform a differential analysis of two NGS samples (Sample1 and Sample 2), using TopHat for the data annotation and Cufflinks for the analysis, and Homo\_sapiens\_GRCh38 as the reference data. The overall execution time for the following workflow is approximately 37 hours, for Sample data of ~60 million reads each and utilizing 8 cores per machine (parameter -p 8 where applicable). The **reference data files** are marked with **bold** in each command.

### *Step 1: Data Annotation per Sample*

tophat -p 8 -G **Homo\_sapiens\_GRCh38.gtf** -o OutputSample1 **Homo\_sapiens\_GRCh38** Sample1\_R1.fastq Sample1\_R2.fastq

tophat -p 8 -G **Homo\_sapiens\_GRCh38.gtf** -o OutputSample2 **Homo\_sapiens\_GRCh38** Sample2\_R1.fastq Sample2\_R2.fastq

### *Step 2: Assembly of the transcripts*

cufflinks -p 8 -o OutputSample1Assembled OutputSample1/accepted\_hits.bam

cufflinks -p 8 -o OutputSample2Assembled OutputSample2/accepted\_hits.bam

### *Step 3: Final transcriptome assembly*

Create the file [assemblies.txt] containing the following two lines:

./OutputSample1Assembled/transcripts.gtf

./OutputSample2Assembled/transcripts.gtf

cuffmerge -g **Homo\_sapiens\_GRCh38.gtf** -s **Homo\_sapiens\_GRCh38.fa** -p 8 assemblies.txt

### *Step 4: Differential Expression results*

cuffdiff -o diff\_out -b **Homo\_sapiens\_GRCh38.fa** -p 8 -L S1,S2 -u merged\_asm/merged.gtf ./OutputSample1/accepted\_hits.bam ./OutputSample2/accepted\_hits.bam

### *Step 5: Statistical analysis and visualization*

[list of R scripts and commands]

# References

1. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A., "**Differential expression in RNA-seq: a matter of depth**", *Genome Res. 2011* Dec;21(12):2213-23. doi: 10.1101/gr.124321.111. Epub 2011 Sep 8.