Reference Genome

8

Melanie Kappelmann-Fenzl

Contents

What You Will Learn in This Chapter

This chapter describes the relevance of the reference genome for the analysis of Next Generation Sequencing (NGS) data and how the respective reference genome can be created. You will learn which databases provide the files for the creation of a Reference Genome Index and which criteria you have to consider when choosing the database and the respective files. Depending on the chosen alignment tool to be used for further analyses, a Reference Genome Index must also be created with the same tool. The corresponding code is shown in detail using the alignment software tools *STAR* and *Bowtie2*.

M. Kappelmann-Fenzl (✉)
Deggendorf Institute of Technology, Deggendorf, Germany

Institute of Biochemistry (Emil-Fischer Center), Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany
e-mail: melanie.kappelmann-fenzl@th-deg.de

M. Kappelmann-Fenzl (ed.), *Next Generation Sequencing and Data Analysis*, Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-62490-3_8

8.1 Introduction

Depending on the samples sequenced (human, mouse, etc.) you need to generate a Genome Index of your reference genome before you are able to align your sequencing reads. Therefore, usually the comprehensive gene annotation on the primary assembly (chromosomes and scaffolds) sequence regions (PRI; .gtf file) and the nucleotide sequence (PRI, FASTA file) of the genome release of interest (e.g., GRCh38) are downloaded. Genome sequence and annotation files can be downloaded from various freely accessible databases as listed below:

- GENCODE: https://www.gencodegenes.org
- UCSC Genome Browser: https://hgdownload.soe.ucsc.edu/downloads.html
- Ensembl: https://www.ensembl.org/info/data/ftp/index.html
- NCBI RefSeq: https://www.ncbi.nlm.nih.gov/refseq/

Once the Genome sequence and annotation files have been downloaded, a Genome Index should be created. Each Genome Index has to be created by the software tool you are using for sequence alignment. In this chapter, we focus on the *STAR* and *Bowtie2* alignment tools.

8.2 Generate Genome Index via STAR

A key limitation with *STAR* [1] is its requirement for large memory space. *STAR* requires at least 30 GB to align to the human or mouse genomes. In order to generate the Genome Index with *STAR*; first, create a directory for the index (e.g., GenomeIndices/Star/GRCh38_index). Then, copy the genome FASTA and Gene Transfer Format (GTF) files into this directory.

Example:

```
STAR –runMode genomeGenerate –runThreadN 23 –genomeDir /path/to/
GenomeIndices/Star/GRCh38_index/ –genomeFastaFiles /path/to/
GenomeIndices/Star/GRCh38_index/GRCh38.primary_assembly.genome.fa –
sjdbGTFfile /path/to/GenomeIndices/Star/GRCh38_index/gencode.v28.
primary_assembly.annotation.gtf –sjdbOverhang 99
```

```
Description of parameters:
--runMode               genomeGenerate
--runThreadN            Number of Threads
--genomeDir             /path/to/genomeDir
--genomeFastaFiles      /path/to/genome/fasta1 /path/to/genome/fasta2 ...
--sjdbGTFfile           /path/to/annotations.gtf
--sjdbOverhang          Read Length -1
```

For a more detailed explanation of the different options of *STAR* to build a Genome Index read the *STAR* Manual on GitHub:

https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf, or type `STAR -h` in the terminal. After generating the Genome Index some more files with information about exonic gene sizes or chromosome sizes can be created using Bioconductor packages in R or the command line tool.

The detailed process on how to create these additional files is depicted below.

Create exonic gene sizes in R [2]

```
source("https://bioconductor.org/biocLite.R")
biocLite("GenomicFeatures")
setwd("/path/to/GenomeIndices/GRCh38")
library(GenomicFeatures)
txdb<-makeTxDbFromGFF("gencode.v24.primary_assembly.annotation.gtf",
format="gtf")
# then collect the exons per gene id
exons.list.per.gene <- exonsBy(txdb,by="gene")
# then for each gene, reduce all the exons to a set of non overlapping exons,
calculate their lengths (widths) and sum then
exonic.gene.sizes <- lapply(exons.list.per.gene,function(x)sum(width
(reduce(x))))
table <- t(exonic.gene.sizes)
write.table (t(table), file = "gencode.v24.primary_assembly.exonic.gene.
sizes.txt", sep = "  ", col.names=N
```

Extract geneID and gene symbol and gene_type from gtf annotation file

```
setwd("/path/to/GenomeIndices/GRCh38")
gtf.file = "gencode.v24.primary_assembly.annotation.gtf"
gtf.gr = rtracklayer::import(gtf.file) # creates a GRanges object
gtf.df = as.data.frame (gtf.gr)
genes = unique(gtf.df[ ,c("gene_id","gene_name","gene_type")])
library(data.table)
fwrite(genes, file="geneIDs_shortannotation.gencodeV24.txt", sep="\t")
```

Create the chromosome-size (command line):

```
samtools faidx /path/to/GenomeIndices/GRCh38/GRCh38.primary_assembly.
genome.fa cut -f1,2 /path/to/GenomeIndices/GRCh38/GRCh38.
primary_assembly.genome.fa.fai > GRCh38.chromosome.sizes
```

All the generated files should be stored in the *GenomeIndices/Star/*directory, or the name you have chosen.

8.3 Generate Genome Index via *Bowtie2*

Bowtie2 can also be used to generate Genome Index files (do not confuse *Bowtie2* indexing with *Bowtie* indexing as they are different). A more detailed description of *Bowtie* and *Bowtie2* can be found in Chap. 9. First, download FASTA files for the unmasked genome (i.e., hg38.fa.gz from http://hgdownload.cse.ucsc.edu/downloads.html) of interest if you have not already. Do NOT use masked sequences.

From the directory containing the *genome.fa* file, run the `bowtie2-build` command. The default options usually work well for most genomes. For example, for hg38:

```
bowtie2-build -threads 23 /path/to/GenomeIndices/bowtie2/GRCh38/GRCh38.
primary_assembly.genome.fa /path/to/GenomeIndices/bowtie2/GRCh38
```

This command will create 6 files with a **.bt2* file extension in your *Bowtie2* index directory. These will then be used by *Bowtie2* to map your sequencing data to the reference genome.

Take Home Message

- Generating a genome index is a time-consuming process, but you only need to do this once per reference genome.
- Organism and version of a reference genome are very important when mapping sequencing reads.
- To create an index of a reference genome you need the nucleotide sequence (FASTA) and the corresponding annotation file (GTF/GFF).
- The most common databases for reference genome download are: GENCODE, UCSC, Ensembl, and NCBI.
- Each Reference Genome Index must be created by the same software tool you want to use for alignment.

Acknowledgements We are grateful to Dr. Richa Barthi (Bioinformatician at TUM Campus Straubing, Germany) for critically reading this text. We thank for correcting our mistakes and suggesting relevant improvements to the original manuscript.

References

1. Dobin A, Gingeras TR. Optimizing RNA-Seq mapping with STAR. Methods Mol Biol. 2016;1415:245–62.
2. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118.