Vilnius University

Faculty of Medicine

Report

Differential expression analyses of Human 19th cromosome gene expression in tissues HBR and UHRR

Prepared by

Laima LUKOŠEVIČIŪTĖ

Student of Systems Biology

Prepared for

Transcriptomics

Partn. Doc. Dr. G. ALZBUTAS

2020

Contents

1 Introduction

In this report I will create a workflow for bulk RNA-seq result analyses. Using reads generated by 2 sample preparation method Collibri and KAPA and two tissues HBR "normal" tissue and UHRR cancerous tissue. Each of cases has 2 biological replicates. To illustrate how workflow performs I carried out differential expression analyses of Human 19th cromosome gene expression in tissues HBR and UHRR. For each sample preparation method DE analysis comparing UHRR vs HBR was performed. Workflow was created using snakemake (Köster & Rahmann 2012) python language extention.

2 Methods

All the steps below were performed by creating rules with snakemake. Used conda (Grüning, Dale, Sjödin, Chapman, Rowe, Tomkins-Tinch, Valieris & Köster 2018) environment can be found in *envs/* directory. And make file can be found in *./Snakefile* file.

2.1 Raw data

The raw **.fastq.gz* pair-end sequencing files were taken from this location. Also reference sequence *chr19_20Mb.fa* and its dependencies *chr19_20Mb.gtf* and *chr19_20Mb.bed* were provided as well as pair of addapters, that can be found in *adapters.fa*. All the files mentioned can be found in *inputs/* folder.

2.2 Quality control and adapter trimming

Quality control of **.fastq.gz* files were carried out using 2 tools: fastqc (version 0.11.9) (Andrews, Krueger, Segonds-Pichon, Biggins, Krueger & Wingett 2012) and multiqc (version 1.8) (Ewels, Magnusson, Lundin & Käller 2016). Addapter triming have been done with bbduk tool (bbtools version 37.62) (Bushnell 2014).

2.3 Read alligment

Genome indexing and read alligment have been done with STAR program (Dobin, Davis, Schlesinger, Drenkow, Zaleski, Jha, Batut, Chaisson & Gingeras 2013). For genome indexing folowing parameters were used.

```
STAR --runThreadN 4 --runMode genomeGenerate --genomeDir outputs/indexing
--genomeFastaFiles {input.ref_genome} --sjdbGTFfile {input.gtf_file}
--sjdbOverhang 150 --genomeSAindexNbases 11
```

For reads alligment folowing parameters were used.

```
STAR --runThreadN 4 --outFileNamePrefix outputs/alligments/{wildcards.sample}/
--genomeDir {input[2]} --readFilesCommand zcat --runMode alignReads
--readFilesIn {input[0]} {input[1]} --outSAMtype BAM SortedByCoordinate
```

2.4 Count files generation

Counts files were generated using featureCounts softwear (Liao, Smyth & Shi 2014). Following parameters were used.

```
featureCounts -p -t exon -g gene_id -a {input.gtf_file}
-o {output} {input.bam_file}
```

2.5 Differential expression analyses

Differential expression analyses have been performed by creating R script (R version 3.6.3), which can be found in *bins/deseq_rscript.R*. For DE analyses DESeq2 (Love, Huber & Anders 2014) was used. To plot vulcano plots EnhancedVolcano library (Blighe 2018) was used.

3 Results

3.1 Quality control

Raw reads have reletively high adapter content (can be seen in fig. 1) and reads after adapter trimming without polyA have a lot lower adapter content (can be seen in fig. 2). Comparing the addition of polyA adapter (see fig. 3) to adapter list it can be seen that no difference is made so polyA is not found in any of the samples.

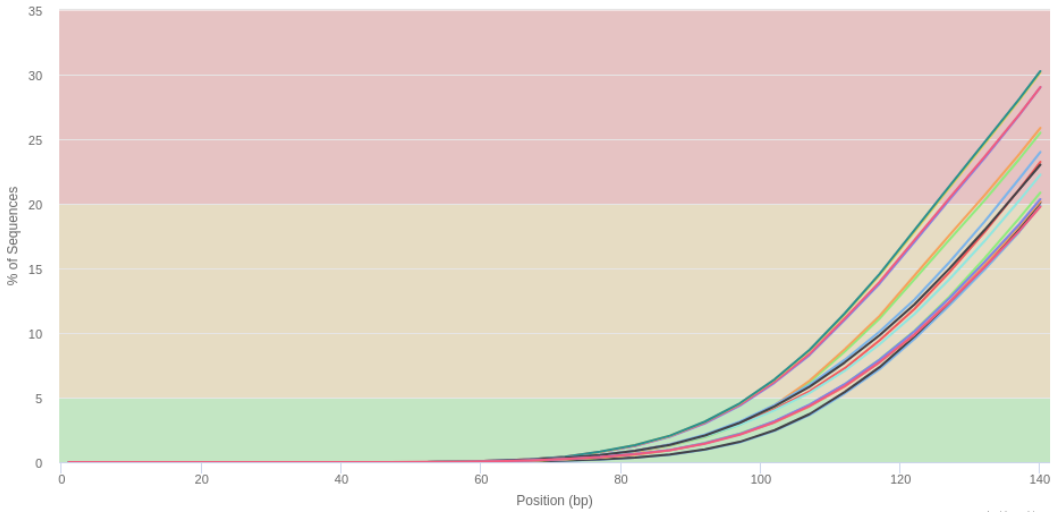

Figure 1: Adapter content in raw files.

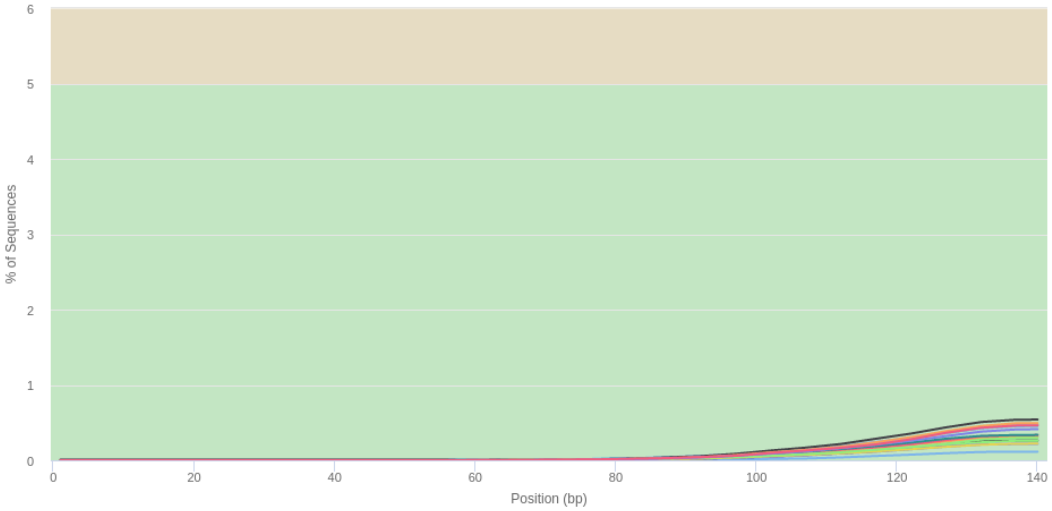

Figure 2: Adapter content in files after adapter trimming (without polyA adapter).

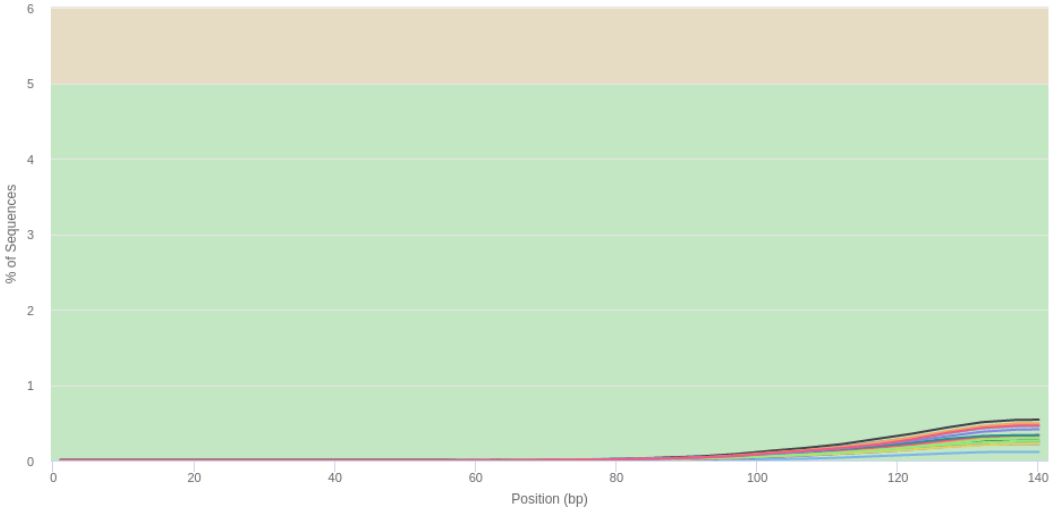

Figure 3: Adapter content in files after adapter trimming (with polyA adapter).

3.2 Differential expression analyses

The results of differential expression analyses were visualized with vulcano plots and can be seen in fig. 4 and fig. 5. The lists of differentialy expressed genes were

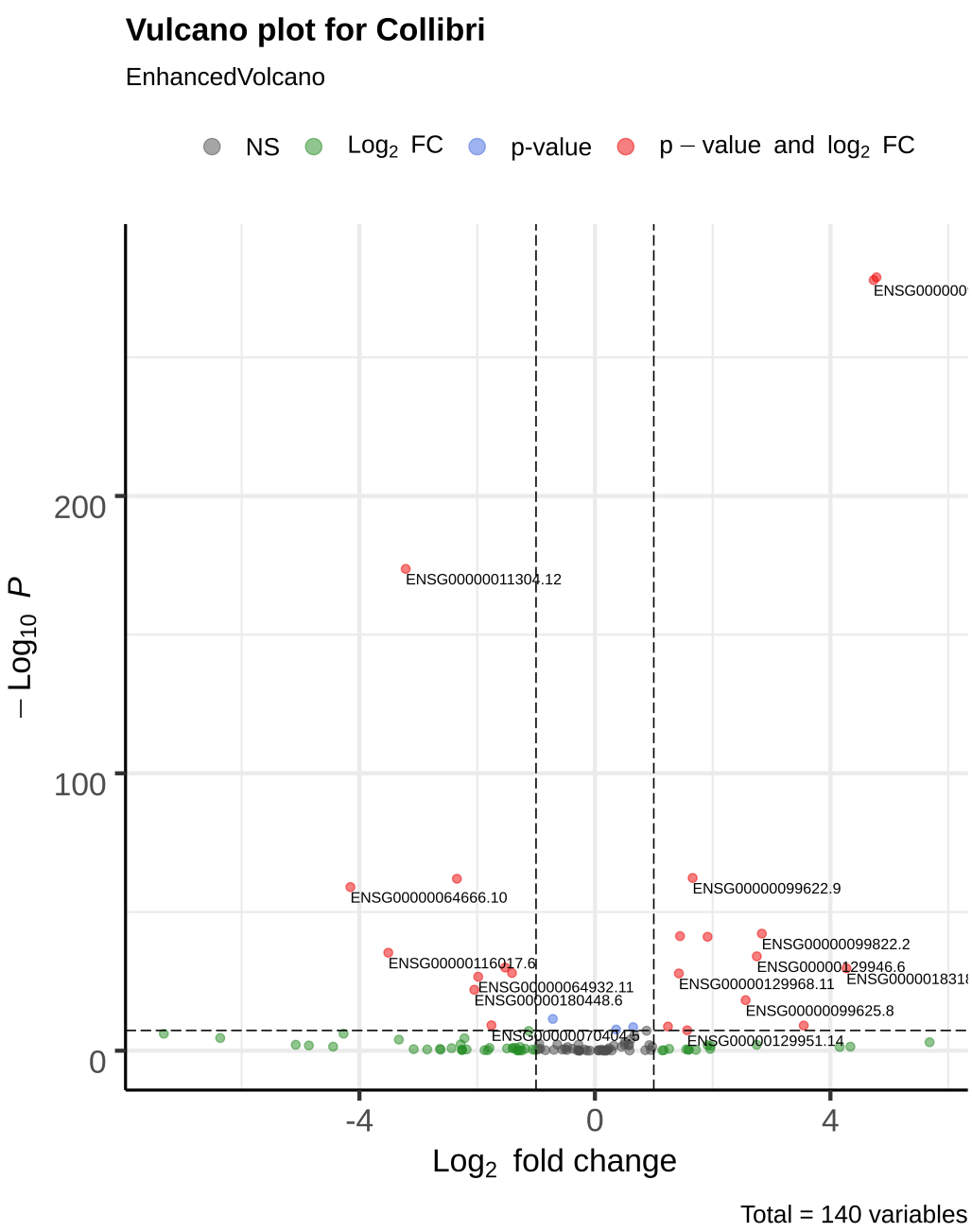

Figure 4: Volcano plot for samples prepered by Collibri method. logFC normal/cancerous.

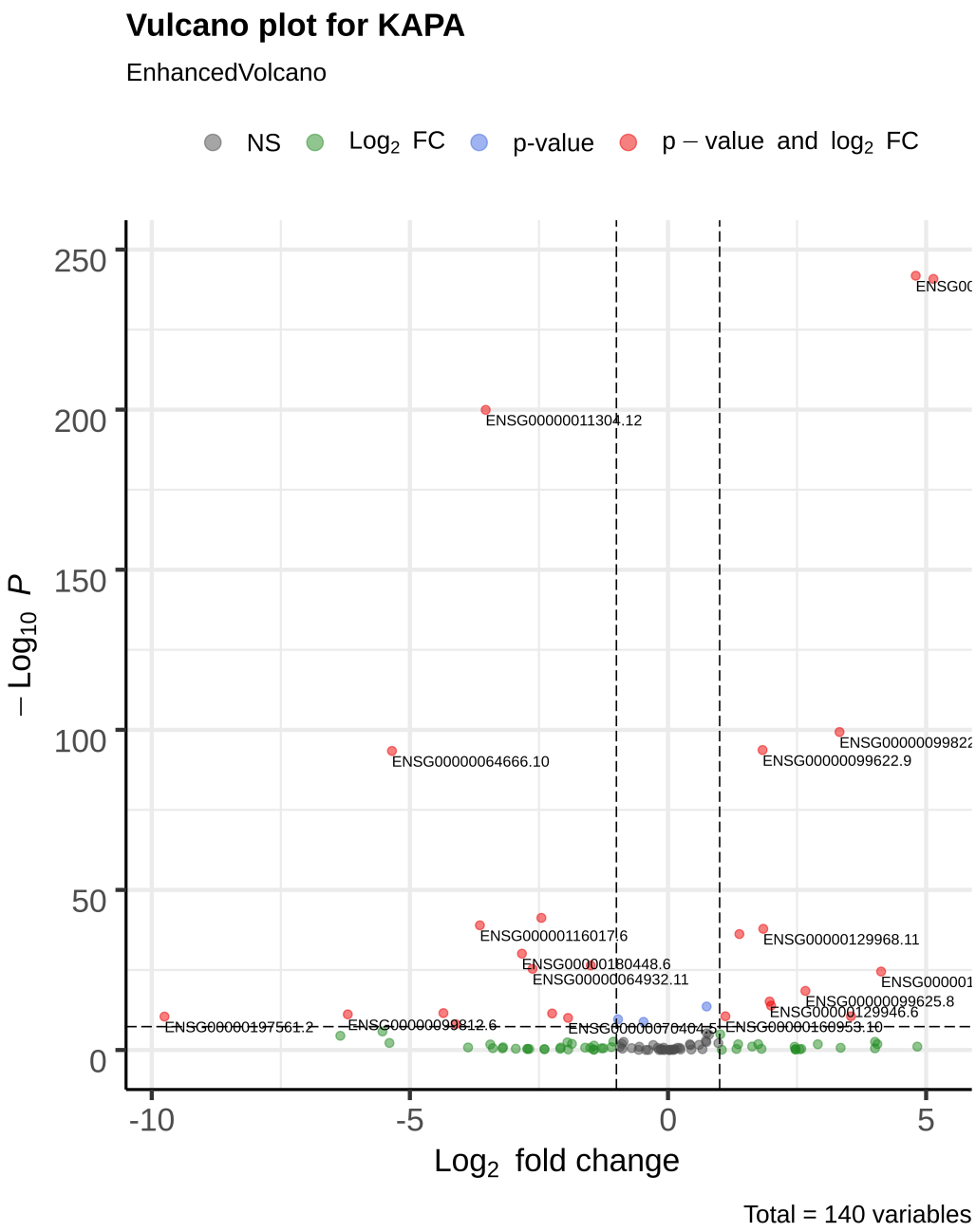

Figure 5: Volcano plot for samples prepered by KAPA method. logFC normal/cancerous.

3.3 Comparing sample preparation methods

Samples were compared by creating Venn diagram which can be seen fig. 6. From fig. 6 we can see that the results do not vary a lot but still some variance are present.

3.4 GO annotations

With my selected p value ($p = 5 \cdot 10^{-8}$) I have got 25 significantly up/down-regulated genes in samples prepared by Collibri method and 28 up/down-regulated genes in samples prepared by KAPA method. With GO annotation enritchment search no statisticaly significant results of enritchment were found in neither of comparissons. The

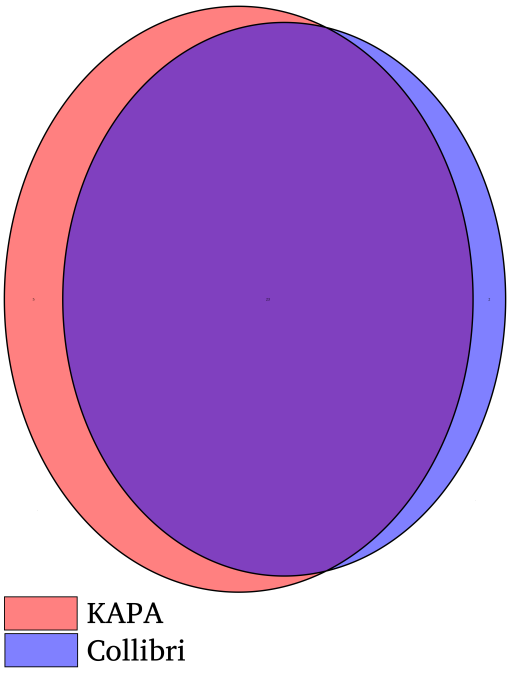

Figure 6: Venn diagram comparing results between different sample preparation methods.

lists of statisticaly significantly up/down-regulated genes can be found in *outputs/DE_genes/Collibri_pval.csv* and *outputs/DE_genes/KAPA_pval.csv* . Probably no significant results were found due to small sample size.

4 Conclusion

Reads quality were inproved after adapter trimming. After DE analyses detected genes not really depend on the sample preparation method. With GO annotation enritchment search no statisticaly significant results of enritchment were found in neither of comparissons. Most likely no significant results were found due to small comparison sample size.

References

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C. & Wingett (2012), 'Fastqc', Babraham Institute.

Blighe, K. (2018), 'Enhancedvolcano: Publication-ready volcano plots with enhanced colouring and labeling'.

Bushnell, B. (2014), 'Bbtools software package', *URL http://sourceforge. net/projects/bbmap* .

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. (2013), 'Star: ultrafast universal rna-seq aligner', *Bioinformatics* **29**(1), 15–21.

Ewels, P., Magnusson, M., Lundin, S. & Käller, M. (2016), 'Multiqc: summarize analysis results for multiple tools and samples in a single report', *Bioinformatics* **32**(19), 3047–3048.

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R. & Köster, J. (2018), 'Bioconda: sustainable and comprehensive software distribution for the life sciences', *Nature methods* **15**(7), 475–476.

Köster, J. & Rahmann, S. (2012), 'Snakemake—a scalable bioinformatics workflow engine', *Bioinformatics* **28**(19), 2520–2522.

Liao, Y., Smyth, G. K. & Shi, W. (2014), 'featurecounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics* **30**(7), 923–930.

Love, M. I., Huber, W. & Anders, S. (2014), 'Moderated estimation of fold change and dispersion for rna-seq data with deseq2', *Genome biology* **15**(12), 550.