Annotation of a soil fungus

Redmine issue: 5123

NBIS staff: Nima Rafati

01 Mars, 2021

Redmine issue:	5123
NBIS staff:	Nima Rafati (nima.rafati@nbis.se)
Principal investigator:	Anna Rosling(anna.rosling@ebc.uu.se)
Request by:	David Manyara(david.manyara@ebc.uu.se)
Organisation:	UU
Estimated time:	100
Used time:	83

Contents

1 Project information **3**

2 Work log **3**

3 Practical information **3**

3.1 Data responsibilities . . . 3

3.2 Acknowledgements . . . 3

3.3 Closing procedures . . . 4

4 Methods **4**

4.1 Genome preparation . . . 4

4.2 RNA-seq . . . 4

4.2.1 QC (00-QC) . . . 4

4.2.2 Trimming (01-Trimmed_reads) . . . 4

4.2.3 Alignment (02-BAM) . . . 4

4.3 Transcriptome assembly (03-Expression-Transcriptome-Assembly) . . . 5

4.4 Expression analysis (03-Expression-Transcriptome-Assembly) . . . 5

5 Results **6**

5.1 QC . . . 6

5.2 Trimming . . . 6

5.3 Alignment . . . 6

5.4 Transcriptome assembly . . . 7

5.5 Expression analysis . . . 7

6 Concluding remarks **11**

7 Reproducibility **13**

8 Sessioninfo **13**

```
## Loading required package: limma
```

1 Project information

- Redmine issue: **5123**
- NBIS staff: **Nima Rafati**
- Request by: **David Manyara**
- Principal investigator: **Anna Rosling**
- Organisation: **UU**

2 Work log

The aim of this project is to improve the annotation of the genome assembly (Merce et al., 2020) by using RNA-seq data.

- **2020-08-11:** Meeting with the group to plan data analyses
- **2020-10-21:** Meeting with the group about RNAseq data
- **2020-11-10:** Meeting with David about variant quality in shared vcf files
- **2020-12-17:** Meeting with the group about status of the RNAseq data analyzed on assembly 3n, the decision was to analyze the data on assembly1
- **2021-02-11:** Last meeting with the group about status of the RNAseq data analyzed on assembly1

3 Practical information

3.1 Data responsibilities

Unfortunately, NBIS does not have resources to keep any files associated with the support request; we kindly suggest that you safely store the results delivered by us. In addition, we kindly ask that you remove the files from UPPMAX/UPPNEX. The main storage at UPPNEX is optimized for high-speed and parallel access, which makes it expensive and not the right place for long-term archiving. Please be considerate of your fellow researchers by not taking up this expensive space.

The responsibility for data archiving lies with universities and we recommend asking your local IT for support with long-term data storage. The Data Center at SciLifeLab may also be of help with discussing other options.

Please note that special considerations may apply to human-derived, sensitive personal data. This should be handled according to specific laws and regulations as outlined at the NBIS website.

3.2 Acknowledgements

If you are presenting the results in a paper, at a workshop or at a conference, we kindly remind you to acknowledge us according to the signed NBIS User Agreement:

NBIS staff should be included as co-authors if the support work leads to a publication and when this is merited in accordance to the ethical recommendations for authorship, *i.e.* the ICMJE recommendations. If applicable, please include *Nima Rafati, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University* as co-author. If the above is not applicable, please acknowledge NBIS like so: *Support by NBIS (National Bioinformatics Infrastructure Sweden) is gratefully acknowledged.*

In addition, Uppmax kindly asks you to acknowledge UPPMAX and SNIC. If applicable, please add: *The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project snic2020-16-21 (Storage) snic2020-5-245 (Computation).*

In any and all publications based on data from NGI Sweden, the authors must acknowledge SciLifeLab, NGI and Uppmax, like so: *The authors would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure, NGI, and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure.*

3.3 Closing procedures

You should soon be contacted by one of our managers, Jessica Lindvall *(jessica.lindvall@nbis.se)* or Henrik Lantz *(henrik.lantz@nbis.se)*, with a request to close down the project in our internal system and for invoicing matters. If we do not hear from you within **30 days** the project will be automatically closed and invoice sent. Again, we would like to remind you about data responsibility and acknowledgements, see the sections on data responsibilities and acknowledgments.

You are naturally more than welcome to come back to us with further data analysis request at any time via the support form. Thank you for using NBIS, we wish you the best of luck with your future research!

4 Methods

4.1 Genome preparation

We analyzed the data on genome assembly 3n and assembly1. Results in this report is based on alignment on assembly1 and the annotation of this genome was updated by using Funannotate (Palmer 2017) by the group.

4.2 RNA-seq

4.2.1 QC (00-QC)

We checked the quality of the reads by using FastQC (Andrews, n.d.) and merged the results by MultiQC(Ewels et al. 2016).

4.2.2 Trimming (01-Trimmed_reads)

We trimmed the reads by trimmomatic(Bolger, Lohse, and Usadel 2014) to trim the adapters and filter low quality reads.

4.2.3 Alignment (02-BAM)

We aligned trimmed reads on Sorghum and AM_fungi genome by STAR(Dobin et al. 2012) and GSNAP(Wu and Watanabe 2005). These two genomes were provided by David Manyara. We first evaluated the aligners and the results showed that GSNAP had a better statistics. Thus, all the results provided here is based on GSNAP alignment. To select species specific reads, we used Disambiguate (Ahdesmäki et al. 2017) and XenofilteR(Kluin et al. 2018) tools. These tools assign reads to corresponding species based on edit distance. We evaluated performance of these tools and XenofilteR could rescue more accurate alignments. We tested different edit distances to assign the reads. After mapping we used QoRTs (Hartley and Mullikin 2015) to check the quality of the alignments.

4.3 Transcriptome assembly (03-Expression-Transcriptome-Assembly)

After assigning the reads to AM_fungi genome, we assembled transcriptome by StringTie (Pertea et al. 2015) using available annotation generated by Funannotate. All the generated gtf files by StringTie were then merged. We compared the merged gtf file with the available annotation by BEDTools (Quinlan and Hall 2010). Also, this gtf was used to extract gene expression level in all the samples.

4.4 Expression analysis (03-Expression-Transcriptome-Assembly)

We extracted fragment counts of all genes by using featurecounts(Liao, Smyth, and Shi 2014). We used reads with mapping quality +20 and pairs that are properly mapped on the same contig. For downstream analysis we used edgeR(Robinson, McCarthy, and Smyth 2009).

All the downstream analysis and visualisation was done in R 3.6.0.

Table 1: Summary of sequencing data; Number of trimmed reads.

Sample	Trimmed_paired_reads
31B	30405701
32A	21132972
32B	36903891
33A	28592911
33B	27192272
34A	33302494
34B	27758826
35A	29492958
35B	28209350
36A	31472821
37B	35355301
38B	27676494
39B	27188497
40B	34563906
41B	33230341
42B	36158716
43B	36548792
44B	36143450
45B	31164207
46B	38477147
48B	42409786

5 Results

There were 24 samples (Sorghum) incubated with AM_Fungi listed below:

5.1 QC

The QC results is available in */crex/proj/uppstore2017083/private/SMS_5123_20_AM_fungi_annotation/results/00-QC/* The duplication rate is a bit high in the raw reads (Please check the multiQC report).

5.2 Trimming

By using Trimmomatic, we kept reads that both pairs survived the trimming and reads with +36 bases length. Trimmed reads are located in */crex/proj/uppstore2017083/private/SMS_5123_20_AM_fungi_annotation/data/Trimmed_reads/*

5.3 Alignment

We first aligned the reads on AM_Fungi (assembly1) and Sorghum separetly. Then, we used XenofilteR with different MM_THreshould implemented in the tool to be used for filtering reads (4-28). Figure 1 shows number of mismatches in the reads assigned to corresponding genomes. Based on distribution in this figure we selected a cut-off value of 16 to filter the reads because values above 16 do not show significant improvement in terms of reliable matches and higher values will only add noise to the data.

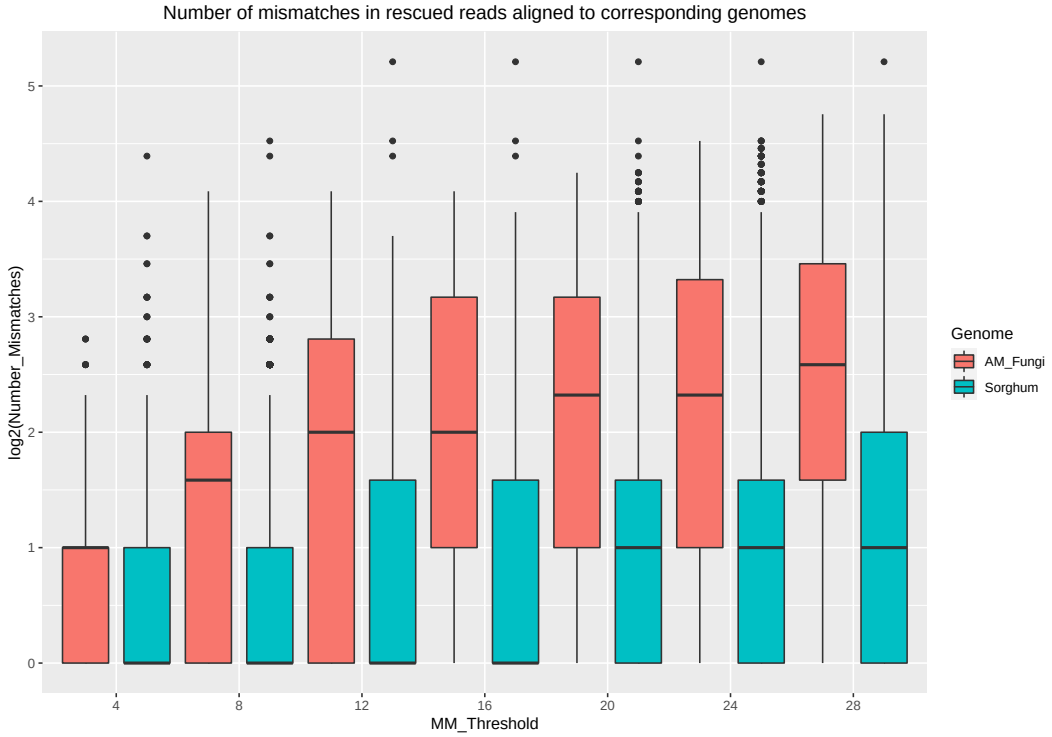

Figure 1: Distribution of nM (number of mismatches) in aligned reads after assigning to corresponding genomes.

All the bam files are in */crex/proj/uppstore2017083/private/SMS_5123_20_AM_fungi_annotation/results/02-BAM/*

5.4 Transcriptome assembly

We assembled the transcripts by using StringTie in each sample and then merged to compare with funannotate annotation. The comparison shows that there are 28 novel transcripts/genes. Merged and individual transcriptome assembly files (gtf) are located in */crex/proj/uppstore2017083/private/SMS_5123_20_AM_fungi_annotation/results/03-Expression-Transcriptome-Assembly/StringTie/*

merged_XenofilteR.gtf consists of all assembled transcripts (novel and previously annotated) and merged_XenofilteR_unique.gtf consists of novel transcripts/genes.

5.5 Expression analysis

We extracted expression values of all transcripts by using featurecounts. We used edgeR to noramlize the data and generate TMM values (trimmed mean of M-values), we checked overall expression pattern among all samples. Figure 2 shows clustering of the samples where samples 35A and 40B are not clustered with the rest of samples.

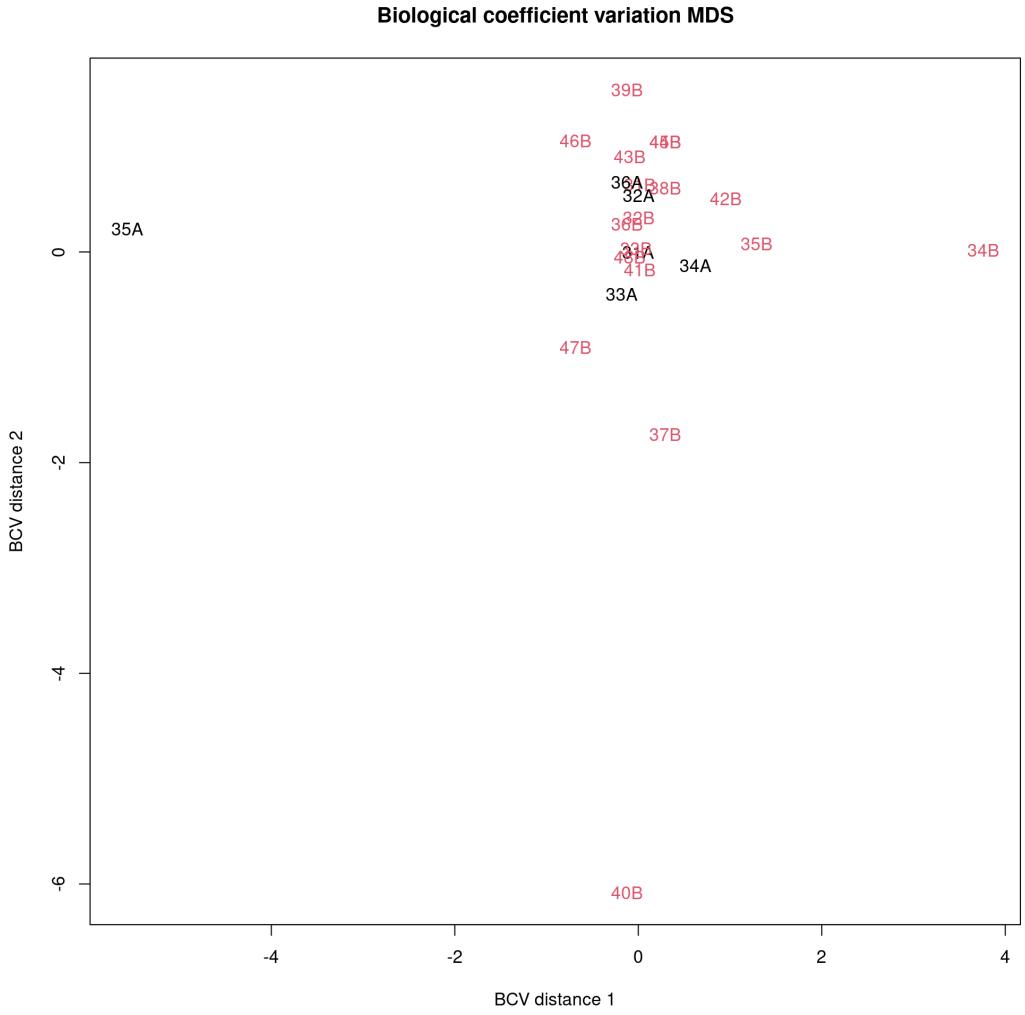

Figure 2: Multidimensional scaling (MDS) plot showing expression profile and biological variation among samples.

We also checked the expression distribution of genes in all samples (Figure 3). The variation between samples is fairly high; for instance samples 31A, 31B and 37B behave differently. In figure 4 you can also see that these samples show different pattern compared to rest of the samples.

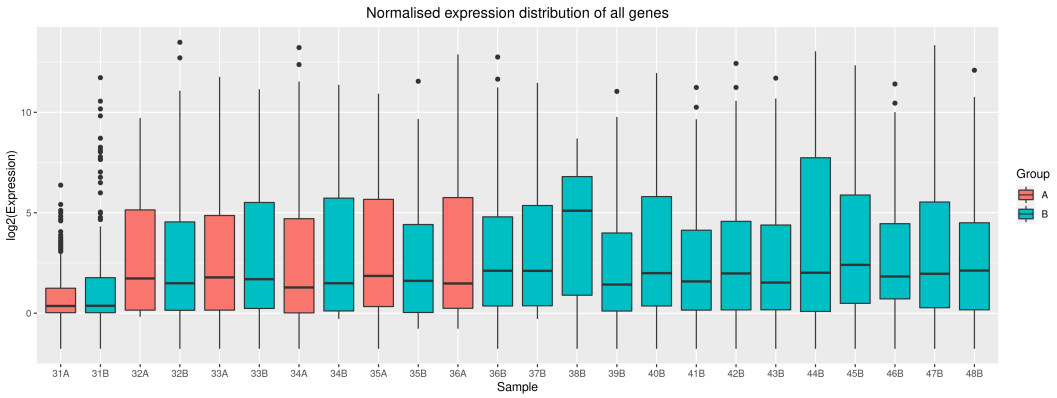

Figure 3: Expression distribution of genes across all samples. Samples are colored based on categories A and B provided in samples name.

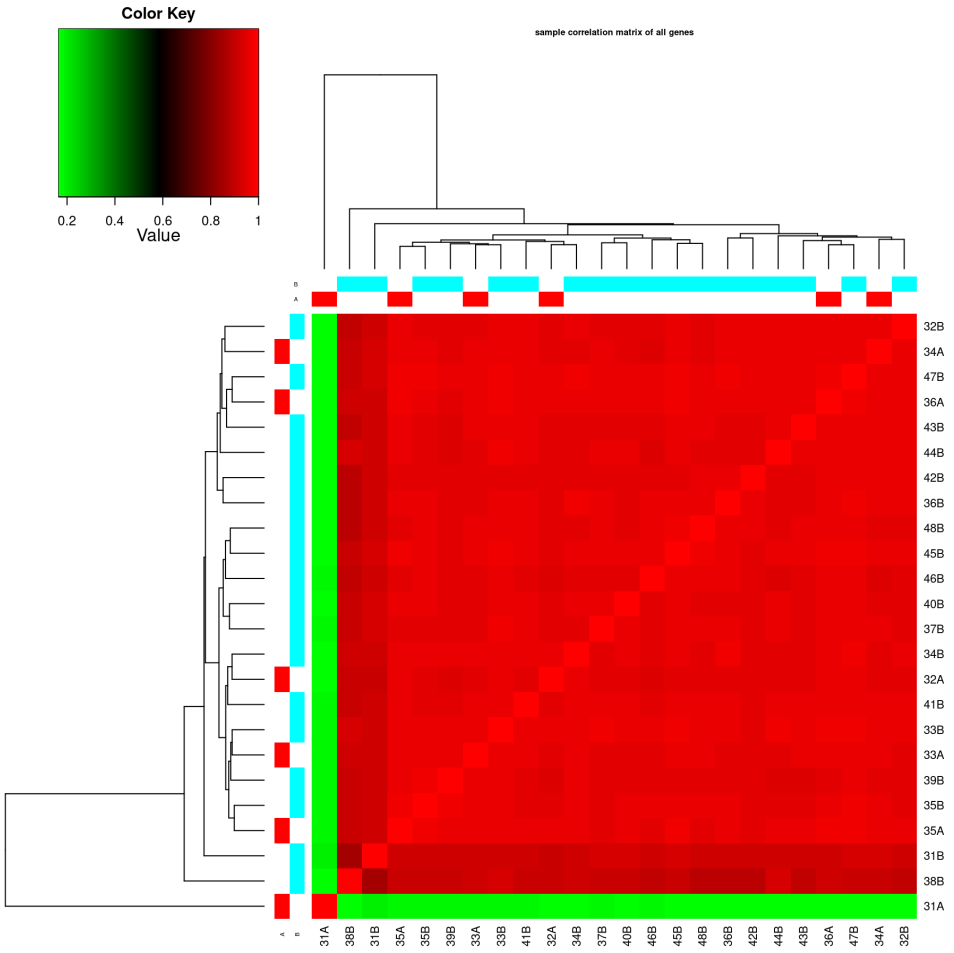

Figure 4: The correlation between samples based on overall expression pattern across the genome.

The expression profile seems to be quite variable among the samples. There are only 89 transcripts/genes where all the samples have expression values above 10 (TMM >= 10). Figure 5 shows an example suggesting two novel genes identified on contig008316 which was not annotated by funannotate.

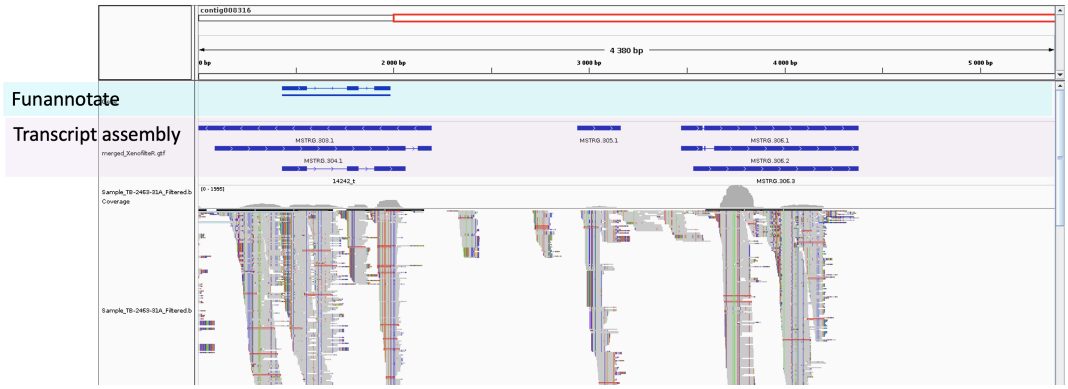

Figure 5: A screenshot of IGV showing read alignments of sample 31A and annotated gene by funannotate and assembled trasncripts by StringTie

6 Concluding remarks

The pipeline implemented in this project seem to improve the assignment of reads to two different genomes (AM_Fungi and Sorghum). However, QC of the results suggest that there is a very minor improvement in annotation using RNA-seq data. One of the concern is very small fraction of reads generated from AM_Fungi in the samples. Figure 6 shows fraction of reads that are assigned to different features on the genome. "Assigned" shows fraction of reads that are uniquely assigned to features (genes). While there are many reads that are mapped on different places which cannot be used for downstream expression analysis or annotation.

All the scripts, this report, and results are available on Uppmax:
/crex/proj/uppstore2017083/private/SMS_5123_20_AM_fungi_annotation/

Also you can find scripts and this report and results (except bam files) on github:
_https://github.com/NBISweden/SMS_5123_20_AM_fungi_annotation_

Scripts are under *doc*:
generate_commands_GSNAP_Sorghum.sh for alignment of reads on Sorghum genome.
generate_commands_GSNAP_assembly1.sh for alignment of reads on assembly1 genome.

As a side note, during analysis I tried to use variant calling data generated in the group in order to identify genes with non-synonymous mutations. Some suggestions were provided to the group which can be found in email communications.

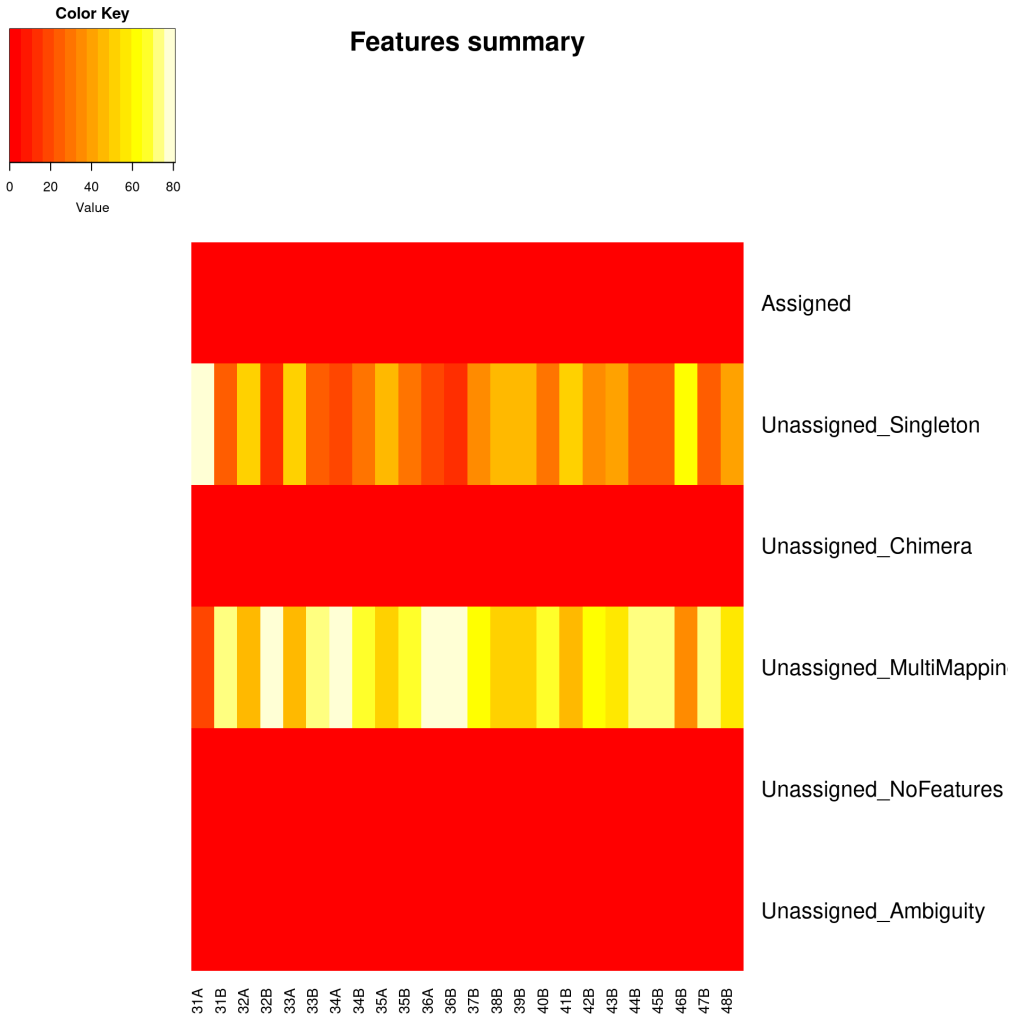

Figure 6: Summary of read assignments reported by featurecounts

7 Reproducibility

List of tools:
- FastQC 0.11.9
- MultiQC 1.9
- Trimmomatic 0.36
- STAR 2.7.2b
- GSNAP gmap-gsnap/2017-09-11
- samtools 1.10
- QoRTs 1.3.6
- StringTie 2.1.4
- featureCounts 2.0.0
- Disambiguate 1.0
- XenofilteR 0.0.99
- R 3.6.0

8 Sessioninfo

sessionInfo()

Ahdesmäki, Miika J., Simon R. Gray, Justin H. Johnson, and Zhongwu Lai. 2017. "Disambiguate: An open-source application for disambiguating two species in next generation sequencing data from grafted samples." *F1000Research* 5 (January): 2741. https://doi.org/10.12688/f1000research.10082.2.

Andrews, S. n.d. "FastQC A Quality Control tool for High Throughput Sequence Data." citeulike-article-id:11583827%20http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2012. "STAR: ultrafast universal RNA-seq aligner." *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: summarize analysis results for multiple tools and samples in a single report." *Bioinformatics* 32 (19): 3047–48. https://doi.org/10.1093/bioinformatics/btw354.

———. 2016. "MultiQC: summarize analysis results for multiple tools and samples in a single report." *Bioinformatics* 32 (19): 3047–48. https://doi.org/10.1093/bioinformatics/btw354.

Hartley, Stephen W, and James C Mullikin. 2015. "QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments." *BMC Bioinformatics* 16 (1): 224. https://doi.org/10.1186/s12859-015-0670-5.

Kluin, Roelof J C, Kristel Kemper, Thomas Kuilman, Julian R de Ruiter, Vivek Iyer, Josep V Forment, Paulien Cornelissen-Steijger, et al. 2018. "XenofilteR: computational deconvolution of mouse and human reads in tumor xenograft sequence data." *BMC Bioinformatics* 19 (1): 366. https://doi.org/10.1186/s12859-018-2353-5.

Liao, Yang, Gordon K Smyth, and Wei Shi. 2014. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features." *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/bioinformatics/btt656.

Palmer, J. 2017. "Funannotate: Fungal genome annotation scripts." https://doi.org/10.5281/zenodo.2604804.

Pertea, Mihaela, Geo M. Pertea, Corina M. Antonescu, Tsung Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. 2015. "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads." *Nature Biotechnology* 33 (3): 290–95. https://doi.org/10.1038/nbt.3122.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: a flexible suite of utilities for comparing genomic features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2009. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Wu, Thomas D., and Colin K. Watanabe. 2005. "GMAP: a genomic mapping and alignment program for mRNA and EST sequences." *Bioinformatics* 21 (9): 1859–75. https://doi.org/10.1093/bioinformatics/bti310.