Analysis of B21000 RNA sequencing data

Stephanie Le Gras

1 Data analysed in this report

Figure 1 on the following page represents the total number of sequenced reads in each sample (51 bp reads).

Table 1 lists all samples analysed in this report, together with their associated experimental conditions.

Table 1: **Samples analysed in this report and their experimental conditions.**

Sample ID	Sample name	Condition
SRR1594531	501_shSCR1_rep1	501_shSCR1
SRR1594532	501_shSCR1_rep2	501_shSCR1
SRR1594533	501_shSCR1_rep3	501_shSCR1
SRR1594534	501_shBRG1_rep1	501_shBRG1
SRR1594535	501_shBRG1_rep2	501_shBRG1
SRR1594536	501_shBRG1_rep3	501_shBRG1
SRR1594537	501_shSCR2_rep1	501_shSCR2
SRR1594538	501_shSCR2_rep2	501_shSCR2
SRR1594539	501_shMITF_rep1	501_shMITF
SRR1594540	501_shMITF_rep2	501_shMITF
SRR1594541	H3A_shSCR_rep1	H3A_shSCR
SRR1594542	H3A_shSCR_rep2	H3A_shSCR
SRR1594543	H3A_shSCR_rep3	H3A_shSCR
SRR1594544	H3A_shMITF_rep1	H3A_shMITF
SRR1594545	H3A_shMITF_rep2	H3A_shMITF
SRR1594546	H3A_shMITF_rep3	H3A_shMITF
SRR1594547	H3A_shBRG1_rep1	H3A_shBRG1
SRR1594548	H3A_shBRG1_rep2	H3A_shBRG1
SRR1594549	H3A_shBRG1_rep3	H3A_shBRG1

2 Preprocessing

Reads were preprocessed in order to remove adapter, polyA and low-quality sequences (Phred quality score below 20). After this preprocessing, reads shorter than 40 bases were discarded for further analysis. These preprocessing steps were performed using cutadapt [1] version 1.10.

Figure 2 on page 3 provides the proportion of remaining reads after each preprocessing step.

Figure 3 on page 4 provides the number of different[1] and unique[1] reads in each sample after the preprocessing step.

3 Mapping

Reads were mapped onto the hg38 assembly of *Homo sapiens* genome using STAR [2] version 2.5.3a. Figure 4 on page 5 provides a summary of mapping results.

[1] For a given sample, the set of unique reads contains reads found only once in this sample and the set of different reads contains all distinct reads, whatever their occurrence number. For instance, for the following set of reads $\{A, B, C, C, D, E, F, F, F, G\}$, the set of unique reads is $\{A, B, D, E, G\}$ and the set of different reads is $\{A, B, C, D, E, F, G\}$.

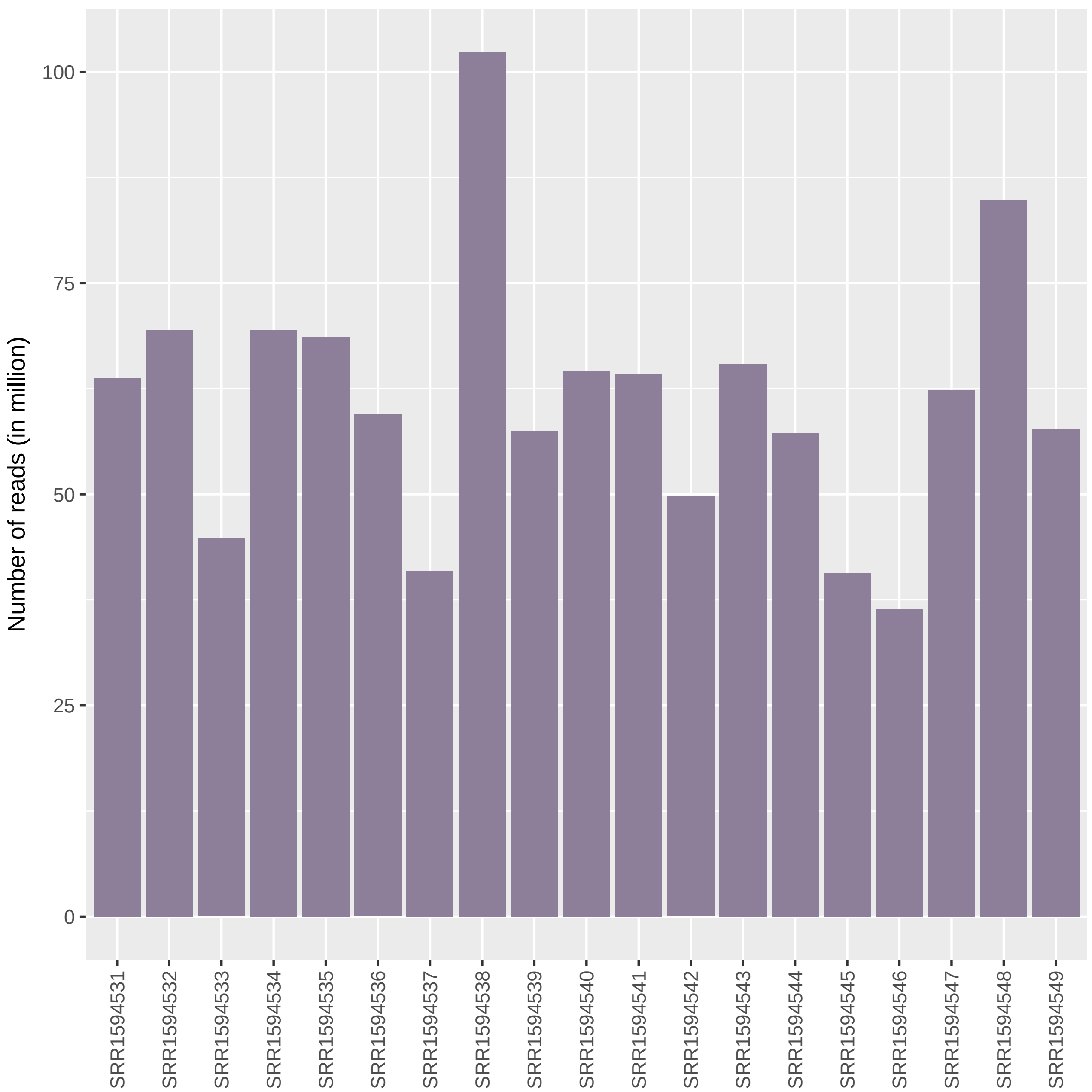

Figure 1: **Number of sequenced reads in each sample.** This barplot represents the total number of sequenced reads (in million, y-axis), in all samples (x-axis).

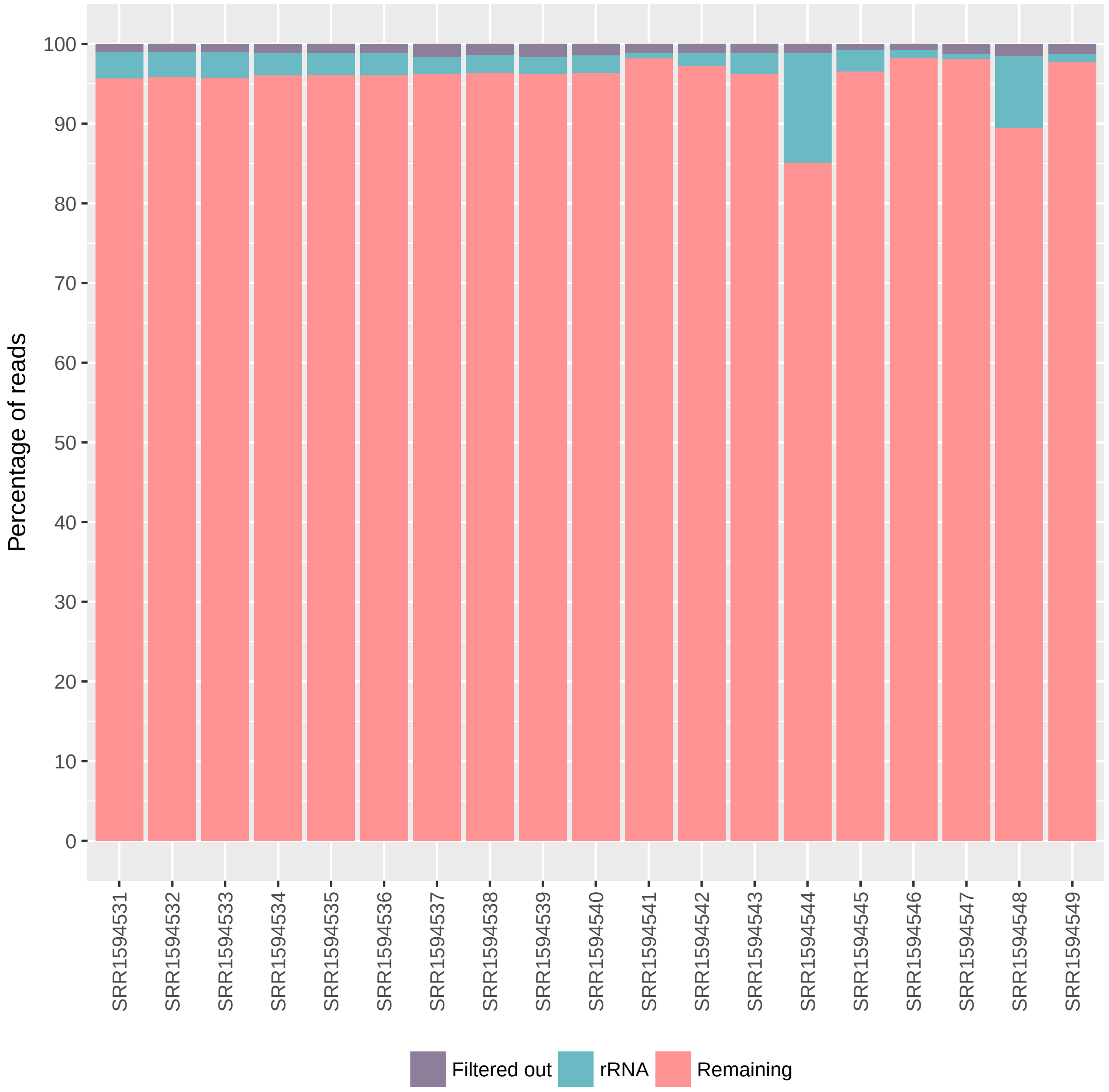

Figure 2: **Summary of preprocessing results.** “Filtered out” represents the percentage of reads shorter than 40 bases discarded after adapter and low-quality sequences (Phred quality score below 20) removal. “Remaining” indicates the percentage of reads that remain after all preprocessing steps. All percentages were calculated relative to the total number of sequenced reads in each sample.

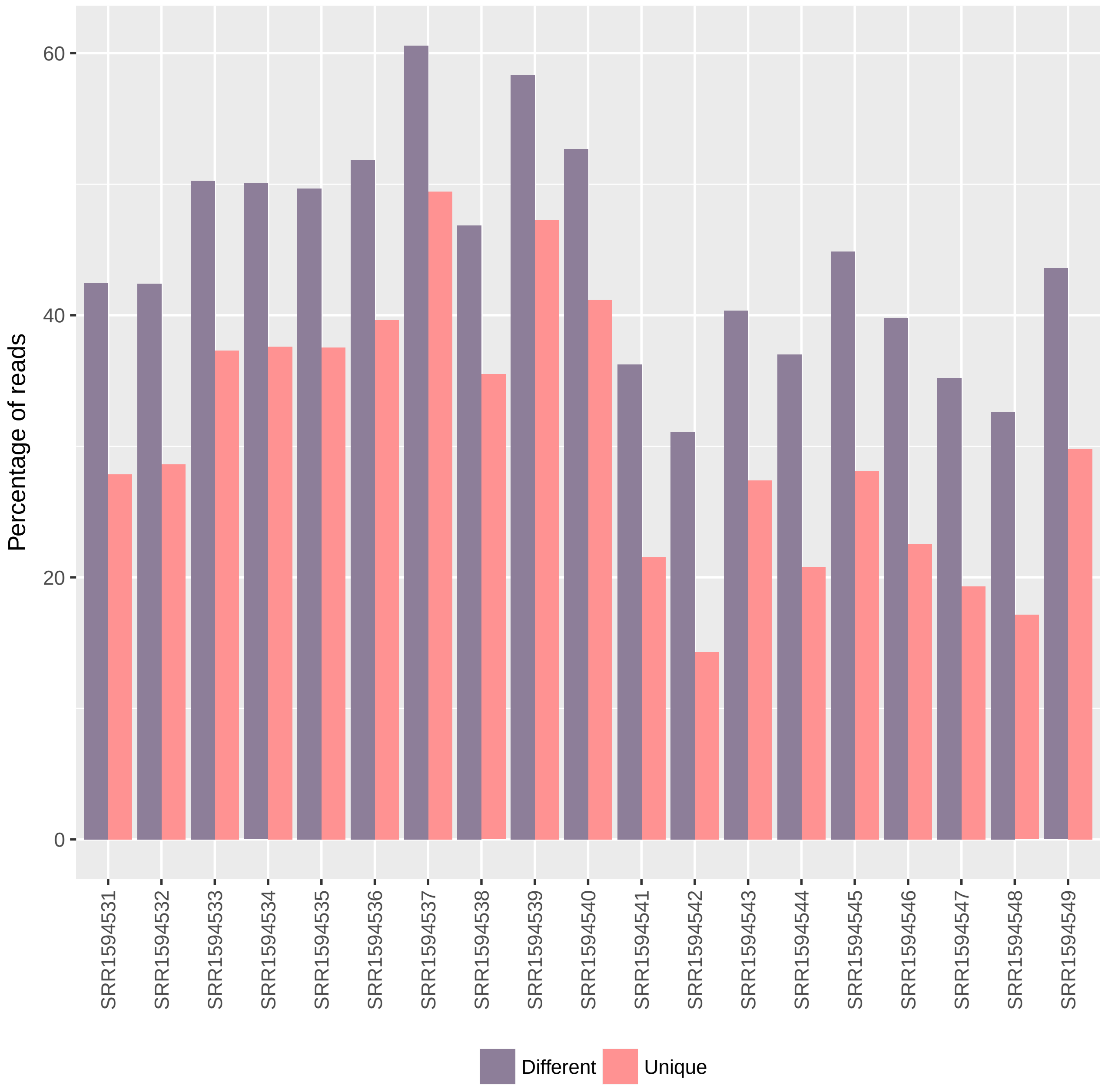

Figure 3: **Percentage of different and unique reads in each sample.** These percentages were calculated relative to the number of reads after preprocessing step.

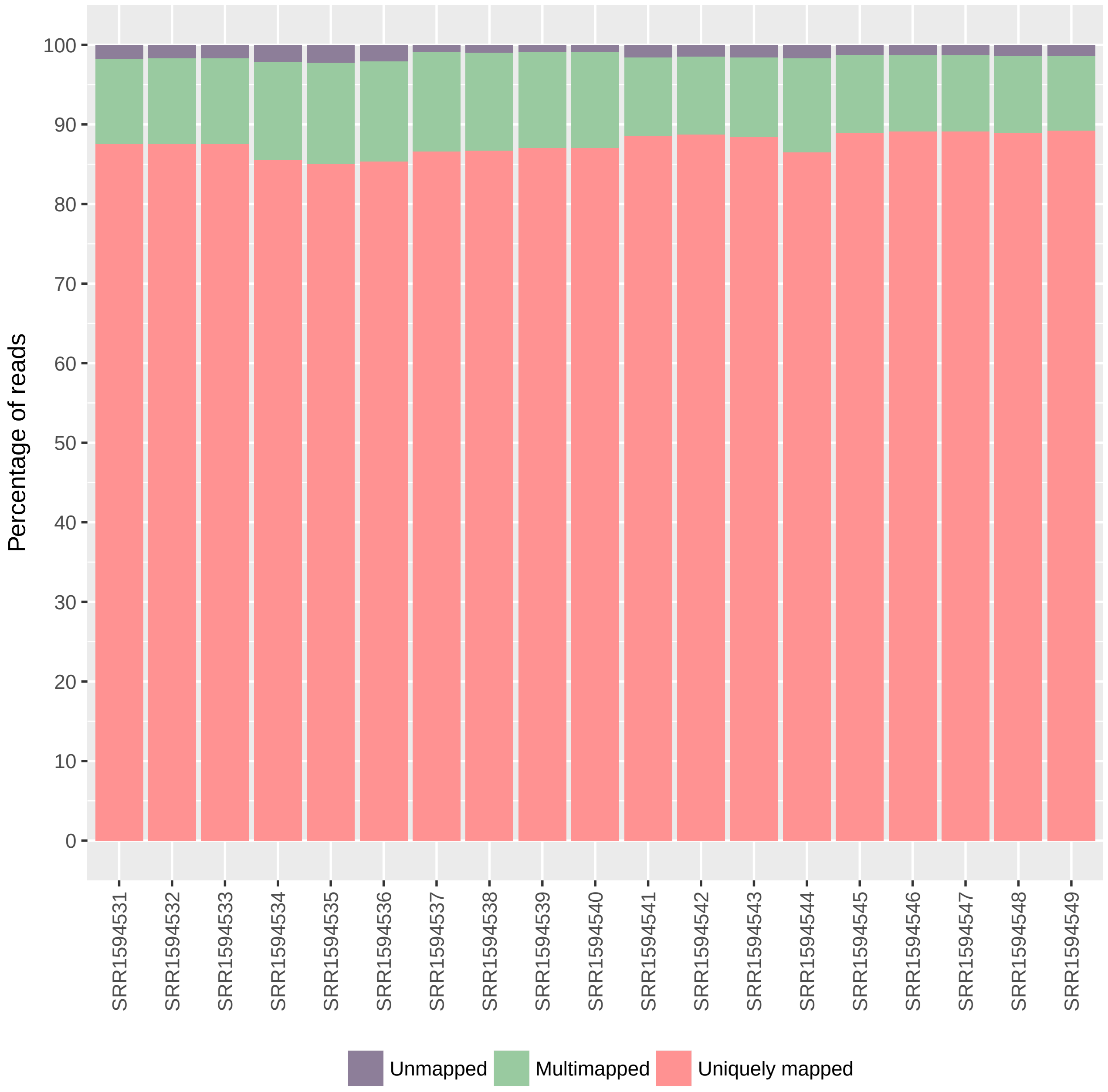

Figure 4: **Summary of mapping results.** This barplot represents the percentage of reads mapped only once on the genome (uniquely mapped), mapped at several locations on the genome (multi-mapped), or not mapped onto the genome (unmapped). These percentages were calculated relative to the number of input reads (i.e. reads kept after preprocessing).

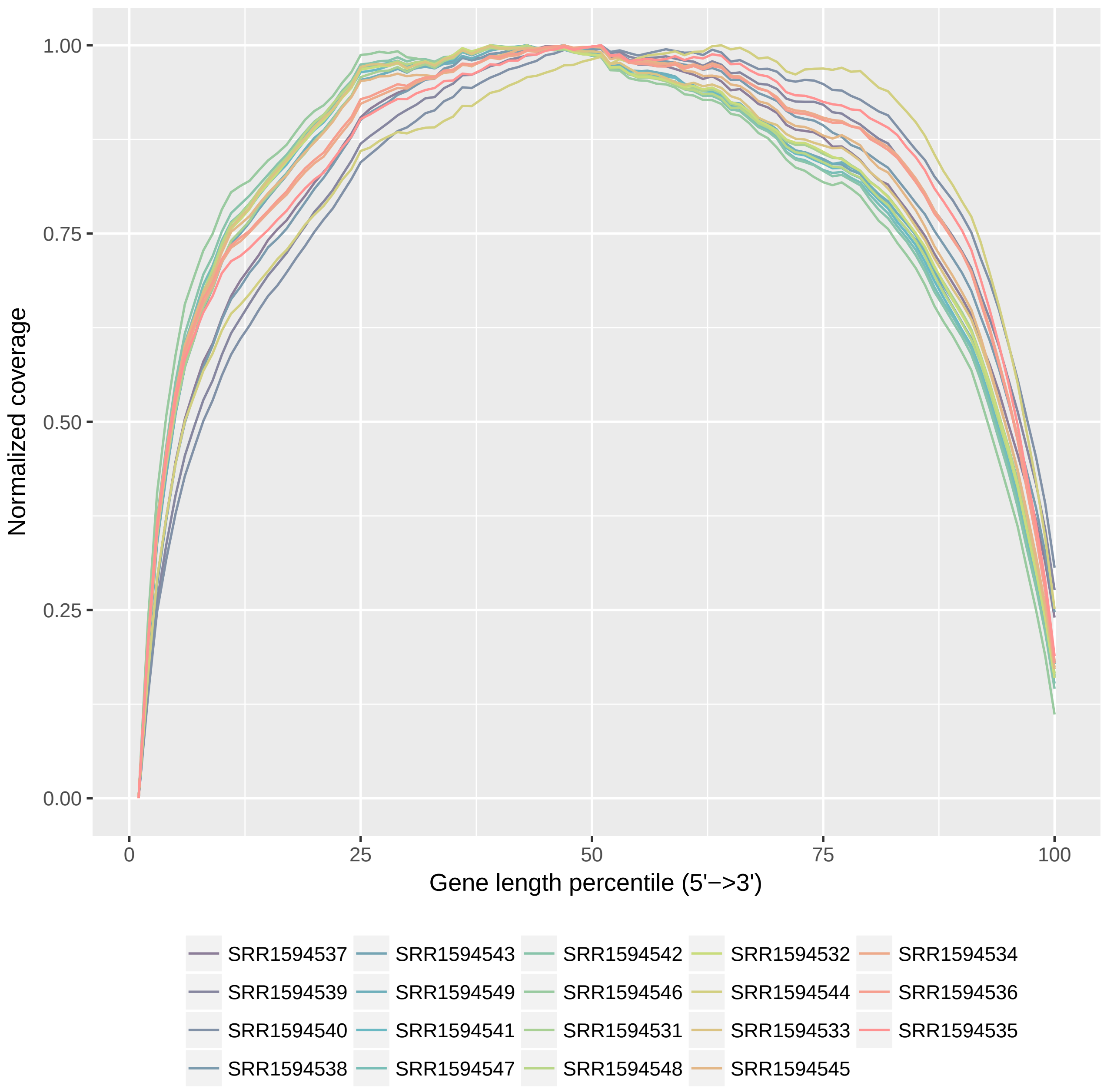

Figure 5: **Read coverage over genes in all samples.** This plot represents the normalized coverage (y-axis) at all percentiles of gene length (x-axis). Genes with mRNA length below 100bp were skipped from this analysis. In the legend, samples are ordered according to their Pearson's skewness coefficient (samples with more skewness are displayed at the begining of the legend).

Figure 5 on the previous page represents read coverage over genes in all samples (coverage was computed for each gene percentile using geneBodyCoverage from RSeQC [3] version 2.6.4).

Figure 6 on the following page provides the median of transcript integrity numbers (TIN) [4] computed across all transcripts with at least 10 mapped reads. TIN is a metric dedicated to capture the uniformity of coverage for a given transcript, and thus is a measure of RNA integrity. TIN score is the percentage of transcript that has uniform read coverage and ranges from 0 (the worst) to 100 (the best).

Figure 7 on page 9 provides the proportion of uniquely aligned reads across exonic, intronic and intergenic genomic regions (using annotations from Ensembl release 103). When genome features overlappe (e.g. the same region can be annotated as both exonic and intronic when different transcripts are overlapping), they are prioritized as: exonic > intronic > intergenic regions. For example, if a read was mapped to both exonic and intronic regions, it will be assigned to exonic region.

4 Quantification

Gene expression quantification was performed from uniquely aligned reads using htseq-count [5] version 0.6.1p1, with annotations from Ensembl version 103 and "union" mode[2]. Figure 8 on page 10 provides a summary of quantification results. Only non-ambiguously assigned reads have been retained for further analyses.

5 Normalization

Read counts have been normalized across samples with the median-of-ratios method proposed by Anders and Huber [6], to make these counts comparable between samples.

Figure 9 on page 11 represents Relative Log expression (RLE) in all samples before and after normalization.

6 Data exploration

Figure 10 on page 12 provides an heatmap of sample-to-sample distances. The Simple Error Ratio Estimate (SERE) [7] coefficient that quantifies global RNA-seq sample differences has been used. A SERE coefficient of 0 indicates data duplication, a score of 1 corresponds to faithful replication (samples differ exactly as would be expected due to Poisson variation). If RNA-Seq samples are truly different, this coefficient is greater than 1 (overdispersion), and the more the coefficient is high, the more the samples are different.

Figure 11 on page 13 represents represents the first principal components of a Principal Component Analysis, showing the main sources of variance in the data.

7 Files delivered

7.1 Alignment files

For each sample, an alignment file in BAM format and the corresponding index (BAI format) are available. The BAM files can be opened using a genome browser, for example Integrative Genomics Viewer [3].

7.2 Result file

A TSV (tab-separated values) file provides raw read counts and normalized read counts for each gene together with gene annotations and the p-value, adjusted p-value and log2 fold-change for each performed comparison. This file contains only genes with at least one read count in one sample. It can be opened with a spreadsheet software like Excel or Calc. The "," character is used as decimal separator in numeric columns. This file contains the following columns:

[2] http://htseq.readthedocs.io/en/master/count.html

[3] IGV is freely available on `http://software.broadinstitute.org/software/igv`

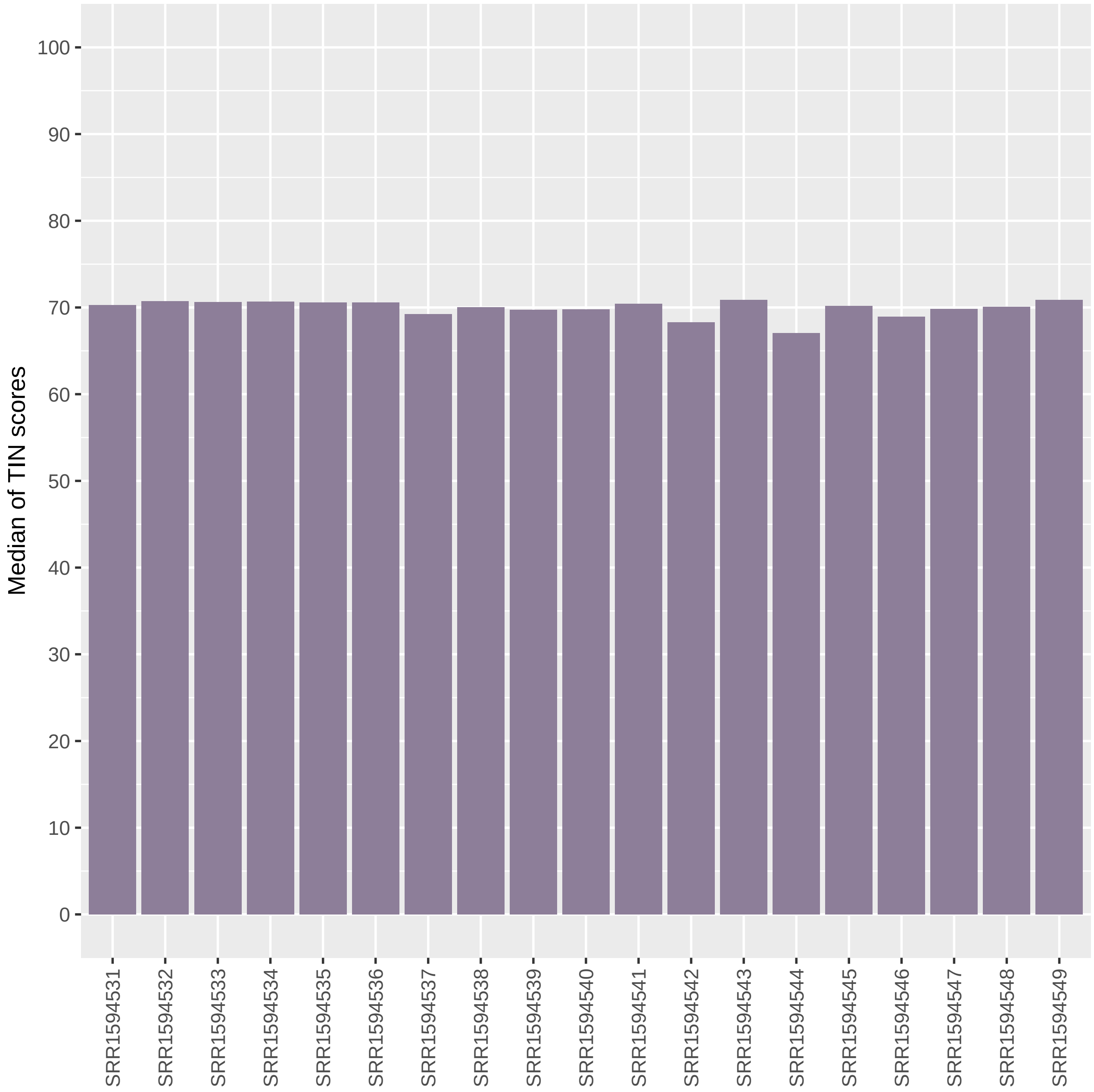

Figure 6: **Median of TIN scores for each sample.** This barplot represents the median of TIN scores (y-axis), in all samples (x-axis). TIN score is the percentage of transcript that has uniform read coverage and ranges from 0 (the worst) to 100 (the best). Transcript with a number of mapped reads below 10 were skipped from the analysis.

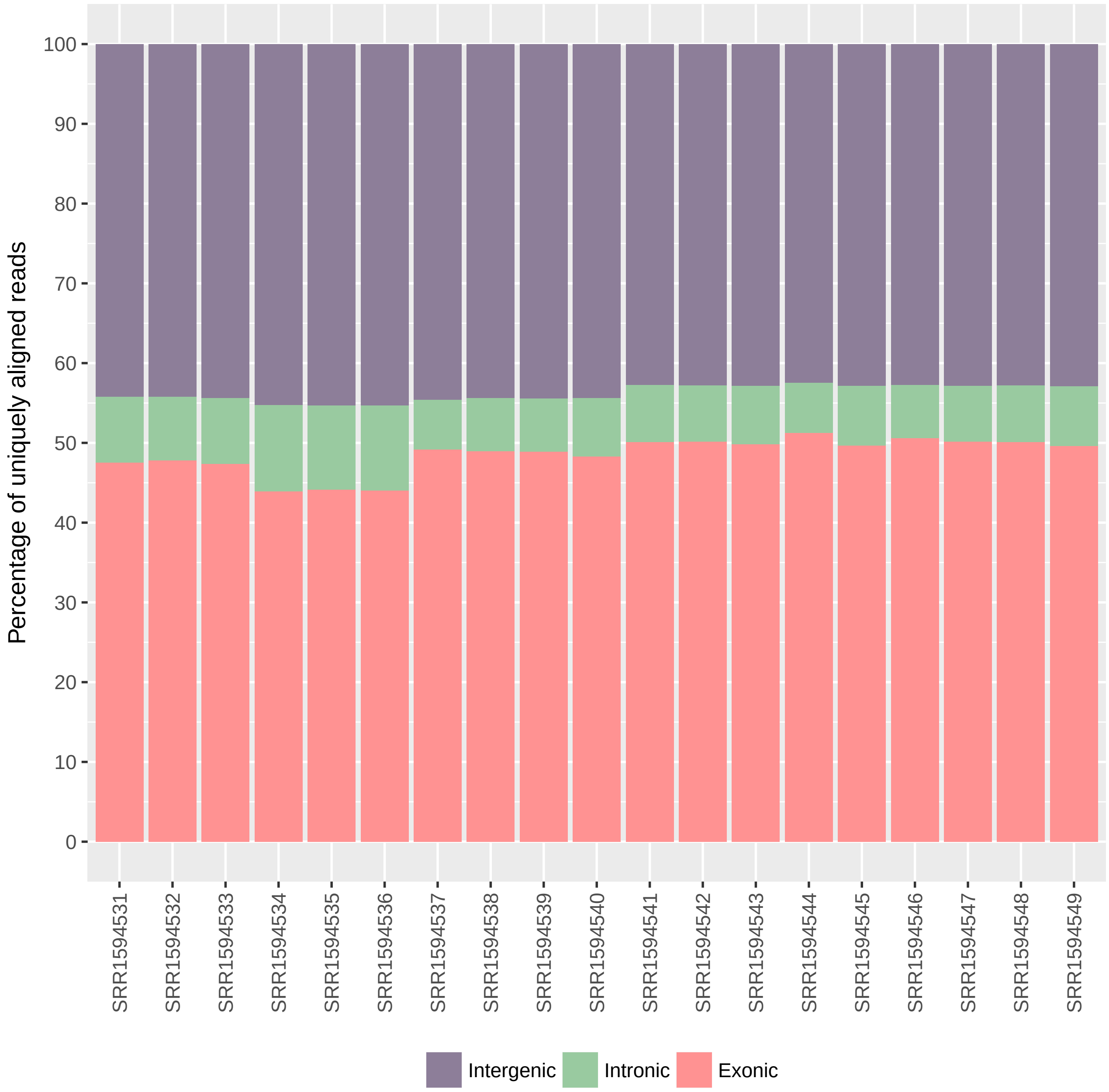

Figure 7: **Reads distribution over annotated genome features.** This barplot represents the proportion of reads aligned to exonic, intronic or intergenic regions, among all uniquely aligned reads.

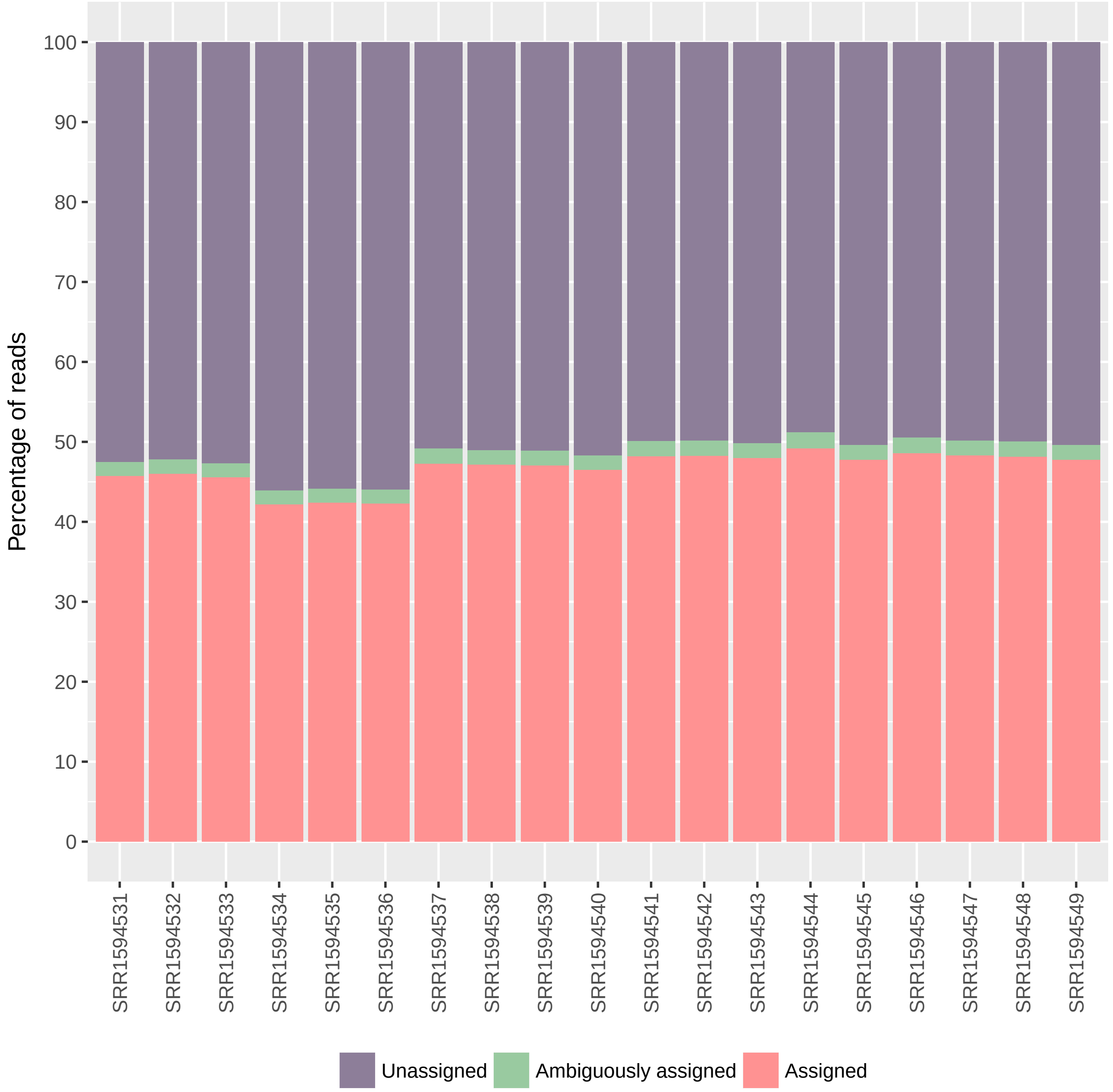

Figure 8: **Summary of quantification results.** This barplot represents the proportion of reads aligned to a genomic region corresponding to one annotated gene (Assigned), to more than one annotated gene (Ambiguously assigned) or to no annotated gene (Unassigned), among all uniquely aligned reads.

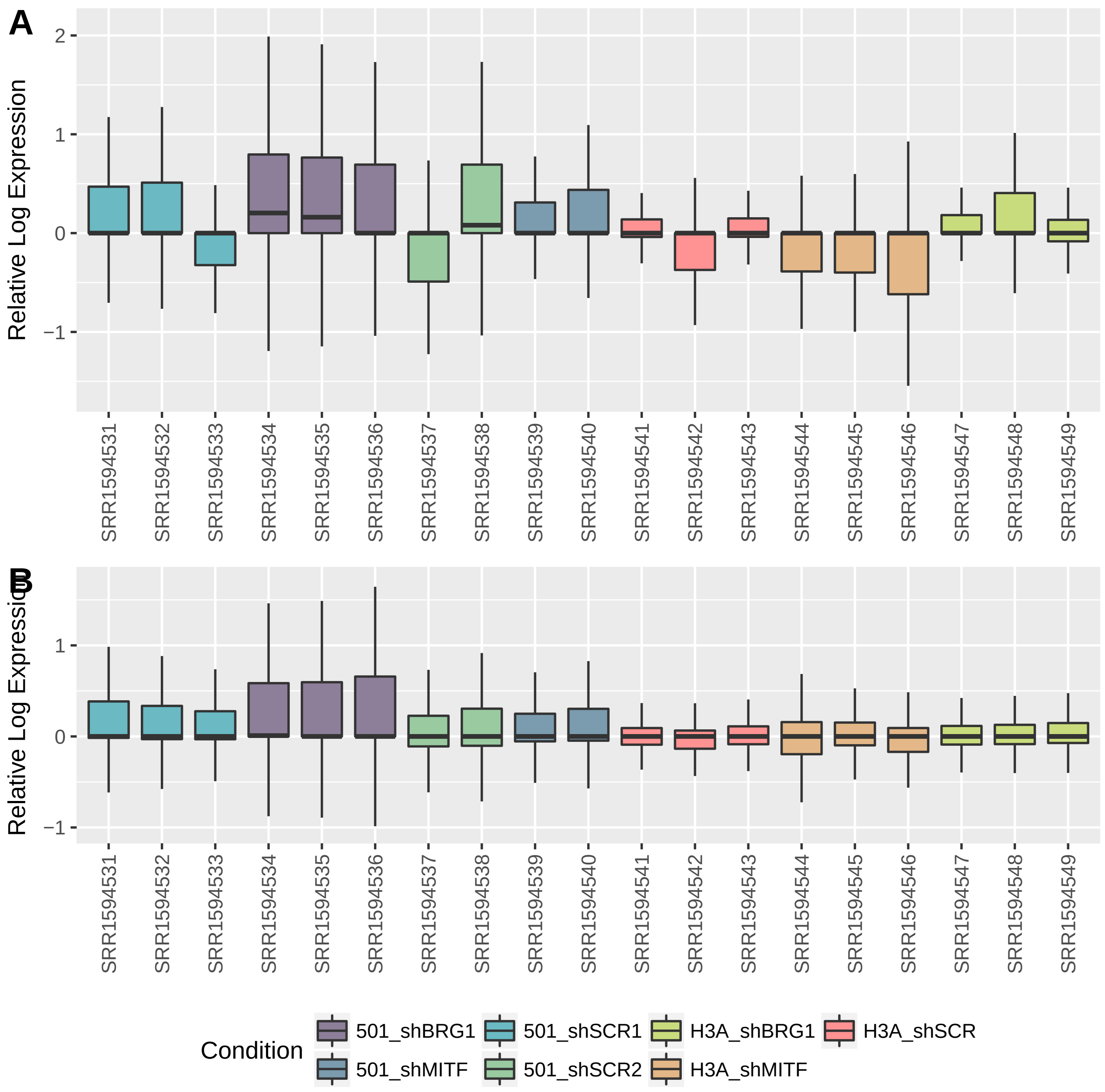

Figure 9: **Relative Log Expression plots in all samples, before (A) and after (B) normalization.** An RLE plot is constructed as follows: for each gene, its median expression (after log transformation) across all samples is calculated, then the deviations from this median is computed ; for each sample, a boxplot of all the deviations for that sample is generated.

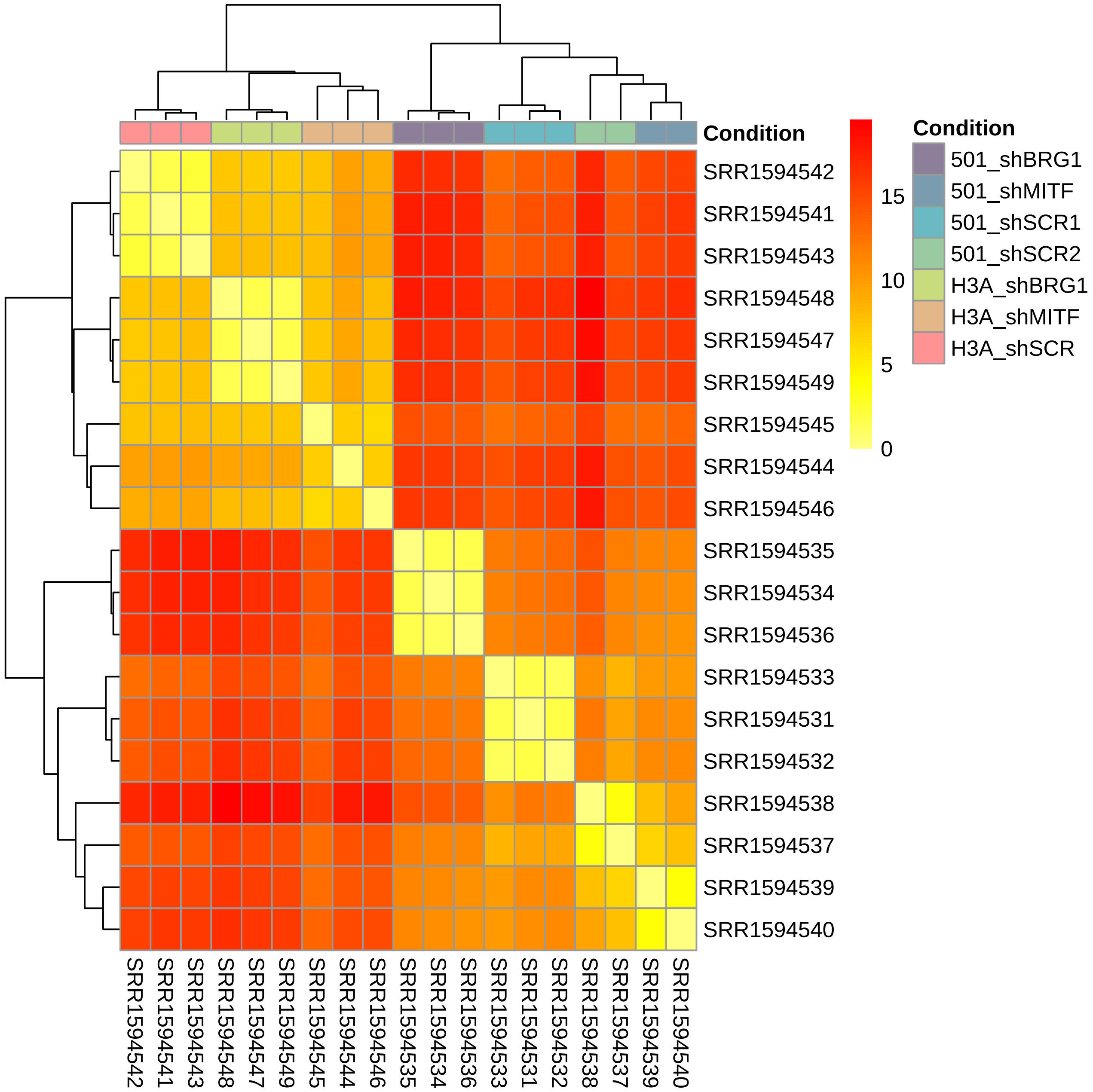

Figure 10: **Heatmap of sample-to-sample distances.** Sample-to-sample distances correspond to SERE [7] coefficient. Hierarchical clustering was performed using the Unweighted Pair Group Method with Arithmetic mean (UPGMA) algorithm.

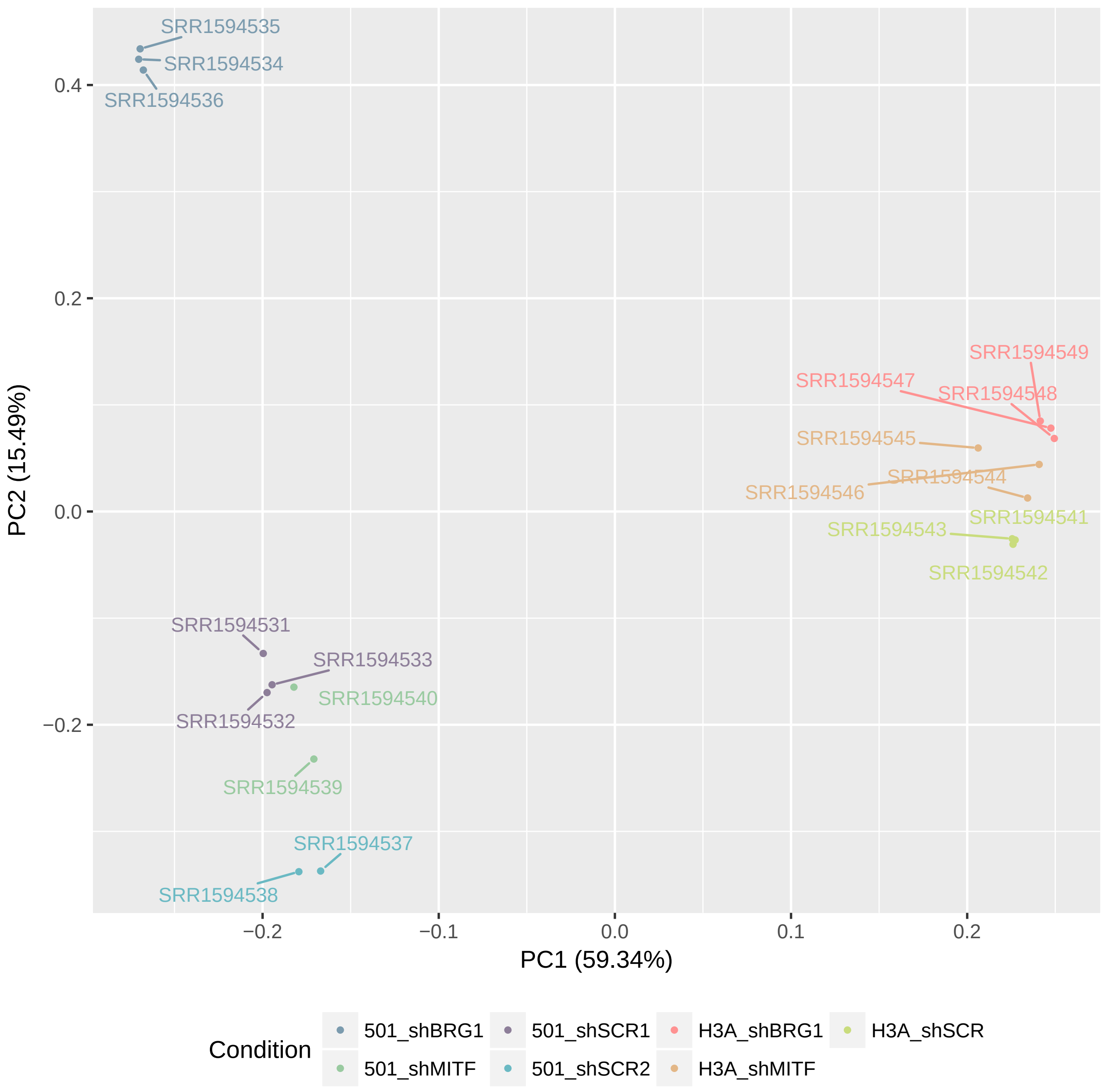

Figure 11: **Principal component analysis.** PCi axis represents the principal component i and the number into brackets indicates the percentage of explained variance associated with this axis. Principal Component Analysis was computed on regularized logarithm transformed data calculated with the method proposed in [8].

Ensembl Gene ID Ensembl identifier of the gene, corresponding to Ensembl release 103.

Raw read counts Number of reads that have been assigned to the gene.

Normalized read counts Number of reads that have been assigned to the gene, normalized to make these counts comparable between samples.

Normalized read counts divided by median of transcripts length in kb Number of reads that have been assigned to the gene, normalized between samples and divided by transcript length in kb (calculated as the median of the length of all transcripts corresponding to this gene). These expression estimates can be compared across genes and samples.

Median of transcripts length Median of the length of all transcripts corresponding to this gene (in bp).

Gene name Common gene name.

Description Description of the gene.

Chromosome name Name of the chromosome where the gene is located.

Start gene position Start coordinate of the gene.

End gene position End coordinate of the gene.

Gene biotype Biotype of the gene as defined in Ensembl[4].

GO:biological process Biological process Gene Ontology terms associated with this gene. A biological process term describes a series of events accomplished by one or more organized assemblies of molecular functions.

GO:molecular function Molecular function Gene Ontology terms associated with this gene. A molecular function term describes activities that occur at the molecular level.

GO:cellular component Cellular component Gene Ontology terms associated with this gene. A cellular component term describes a location, relative to cellular compartments and structures, occupied by a macromolecular machine when it carries out a molecular function.

8 Version information

8.1 Version of used tools

Table 2 on the following page provides the tools used in GenomEast RNA-seq pipeline version 1.2.2 (used to perform the analyses described in the report) and their corresponding version.

8.2 Version of used R packages

```
## R version 3.3.2 (2016-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Scientific Linux release 6.7 (Carbon)
##
## locale:
##  [1] LC_CTYPE=fr_FR.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=fr_FR.UTF-8        LC_COLLATE=fr_FR.UTF-8
##  [5] LC_MONETARY=fr_FR.UTF-8    LC_MESSAGES=fr_FR.UTF-8
##  [7] LC_PAPER=fr_FR.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
```

[4] https://www.ensembl.org/Help/Faq?id=468

Table 2: **Tools used for the analyses presented in this report.**

Tool	Release	Description
cutadapt	1.10	To trim low quality bases and adapter sequences from the reads and to remove too-short reads after trimming.
FastQC	0.11.5	To perform quality controls on the reads.
HTSeq	0.6.1p1	To compute the number of reads in annotated transcribed regions.
R	3.3.2	To perform statistical analysis, graphics and to generate this report.
RSeqQC	2.6.4	To perform quality controls on the alignments.
samtools	1.3.1	To manipulate SAM/BAM files.
STAR	2.5.3a	To perform spliced alignment of reads onto a reference genome.

```
## [11] LC_MEASUREMENT=fr_FR.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
##  [1] parallel  stats4    methods   grid      stats     graphics  grDevices
##  [8] utils     datasets  base
##
## other attached packages:
##  [1] ggrepel_0.6.5              ggfortify_0.4.1
##  [3] pheatmap_1.0.8             DESeq2_1.16.1
##  [5] SummarizedExperiment_1.4.0 Biobase_2.34.0
##  [7] GenomicRanges_1.26.4       GenomeInfoDb_1.10.3
##  [9] IRanges_2.8.2              S4Vectors_0.12.2
## [11] BiocGenerics_0.22.0        reshape2_1.4.2
## [13] VennDiagram_1.6.18         futile.logger_1.4.1
## [15] knitr_1.12.3               xtable_1.8-2
## [17] cowplot_0.8.0              ggplot2_2.2.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.13        locfit_1.5-9.1      lattice_0.20-34
##  [4] tidyr_0.6.1         assertthat_0.1      digest_0.6.12
##  [7] R6_2.2.0            plyr_1.8.4          futile.options_1.0.0
## [10] backports_1.0.5     acepack_1.4.1       RSQLite_1.1-2
## [13] evaluate_0.10       zlibbioc_1.20.0     lazyeval_0.2.0
## [16] data.table_1.10.4   annotate_1.50.0     rpart_4.1-10
## [19] Matrix_1.2-7.1      checkmate_1.8.2     labeling_0.3
## [22] splines_3.3.2       BiocParallel_1.6.6  geneplotter_1.50.0
## [25] stringr_1.2.0       foreign_0.8-67      htmlwidgets_0.5
## [28] RCurl_1.95-4.8      munsell_0.4.3       base64enc_0.1-3
## [31] htmltools_0.3       nnet_7.3-12         tibble_1.2
## [34] gridExtra_2.2.1     htmlTable_1.9       Hmisc_4.0-3
## [37] XML_3.98-1.6        dplyr_0.5.0         bitops_1.0-6
## [40] gtable_0.2.0        DBI_0.6-1           magrittr_1.5
## [43] formatR_1.4         scales_0.4.1        stringi_1.1.2
## [46] XVector_0.14.1      genefilter_1.58.1   latticeExtra_0.6-28
## [49] Formula_1.2-1       lambda.r_1.1.7      RColorBrewer_1.1-2
## [52] tools_3.3.2         survival_2.40-1     AnnotationDbi_1.38.0
## [55] colorspace_1.3-2    cluster_2.0.6       memoise_1.1.0
```

References

[1] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal*, vol. 17, no. 1, pp. 10–12, 2011.

[2] A. Dobin, C. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. Gingeras, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[3] L. Wang, S. Wang, and W. Li, “RSeQC: quality control of RNA-seq experiments,” *Bioinformatics*, vol. 28, no. 16, pp. 2184–5, 2012.

[4] L. Wang, J. Nie, H. Sicotte, Y. Li, J. Eckel-Passow, S. Dasari, P. Vedell, P. Barman, L. Wang, R. Weinshiboum, J. Jen, H. Huang, M. Kohli, and J. Kocher, “Measure transcript integrity using RNA-seq data,” *BMC Bioinformatics*, vol. 17, no. 58, 2016.

[5] S. Anders, P. Pyl, and W. Huber, “HTSeq-a python framework to work with high-throughput sequencing data,” *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.

[6] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biol*, vol. 11, no. 10, p. R106, 2010.

[7] S. Schulze, R. Kanwar, M. Glzenleuchter, T. Therneau, and A. Beutler, “SERE: Single-parameter quality control and sample comparison for RNA-Seq,” *BMC Genomics*, vol. 13, p. 524, 2012.

[8] M. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biol*, vol. 15, no. 12, p. 550, 2014.