Analysis of S17305 RNA sequencing data

Celine Keime

1 Data analysed in this report

Figure 1 on the following page represents the total number of sequenced reads in each sample (50 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
BMBL57	8w_C18	Ctrl8w_NaCl
BMBL58	8w_C30	Ctrl8w_NaCl
BMBL59	8w_C53	Ctrl8w_NaCl
BMBL61	8w_C97	Ctrl8w_NaCl
BMBL62	8w_KONaCl26	KO8w_NaCl
BMBL63	8w_KONaCl31	KO8w_NaCl
BMBL64	8w_KONaCl47	KO8w_NaCl
BMBL65	8w_KONaCl71	KO8w_NaCl
BMBL66	8w_KONaCl72	KO8w_NaCl
BMBL67	8w_KONaCl76	KO8w_NaCl
BMBL68	8w_KOAAV17	KO8w_AAV
BMBL69	8w_KOAAV20	KO8w_AAV
BMBL70	8w_KOAAV49	KO8w_AAV
BMBL71	8w_KOAAV55	KO8w_AAV
BMBL72	8w_KOAAV92	KO8w_AAV
BMBL73	8w_KOAAV94	KO8w_AAV
BMBL91	3w_C3	Ctrl3w
BMBL92	3w_C5	Ctrl3w
BMBL93	3w_C15	Ctrl3w
BMBL95	3w_C193	Ctrl3w
BMBL96	3w_KO2	KO3w
BMBL97	3w_KO4	KO3w
BMBL99	3w_KO33	KO3w
BMBL100	3w_KO185	KO3w
BMBL101	3w_KO189	KO3w
BMBL113	3w_C54	Ctrl3w
BMBL114	3w_KO59	KO3w
BMBL115	8w_C93	Ctrl8w_NaCl

2 Preprocessing

Reads were preprocessed in order to remove adapter 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. Reads were mapped to rRNA sequences using bowtie [2] version 2.2.8, and reads mapping to rRNA sequences were removed for further analysis. Reads were mapped to spike sequences using bowtie [2] version 2.2.8, and reads mapping to spike sequences were removed for further analysis.

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

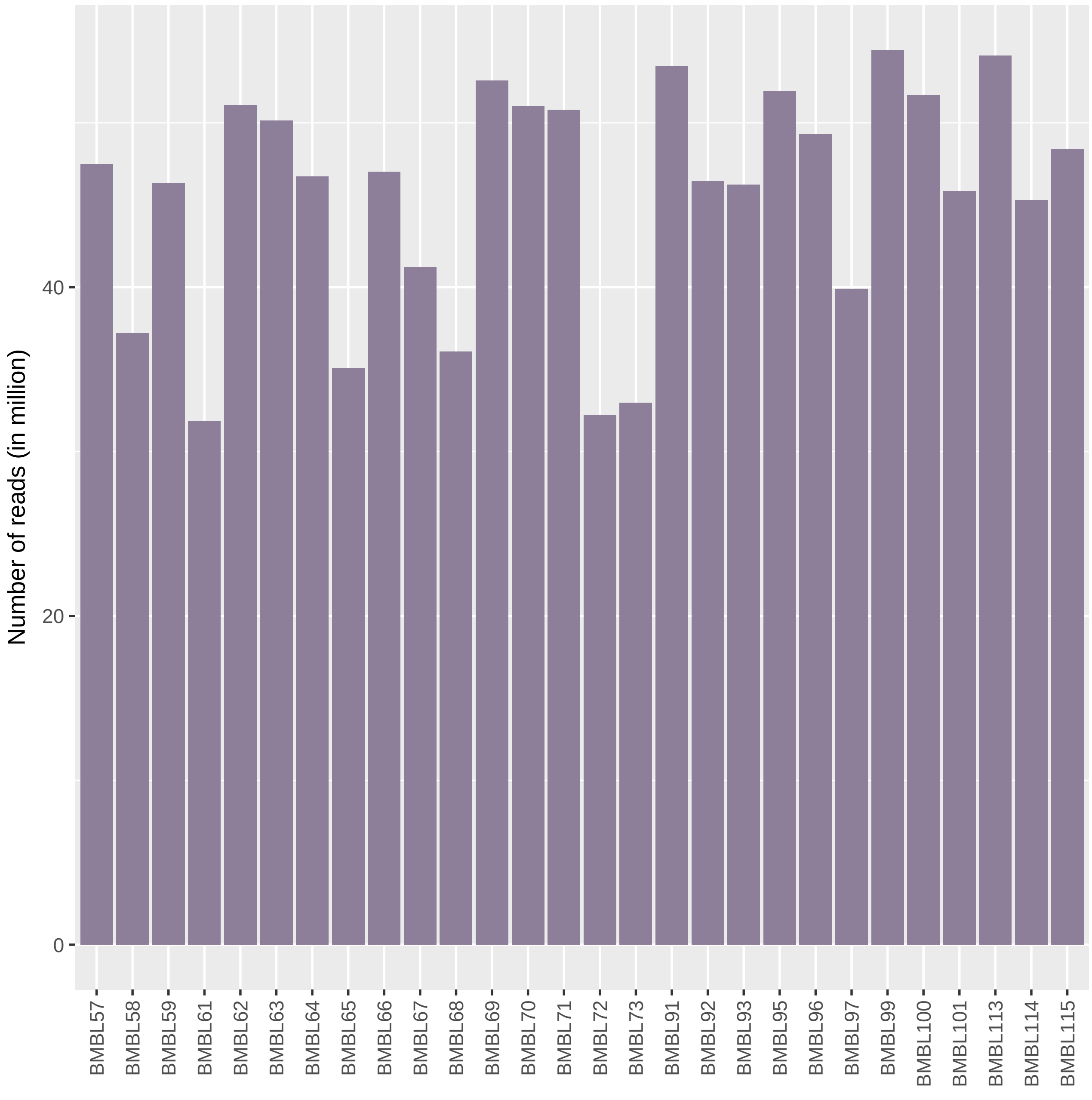

Figure 1: **Number of sequenced reads in each sample.** This barplot represents the total number of sequenced reads (in million, y-axis), in all samples (x-axis).

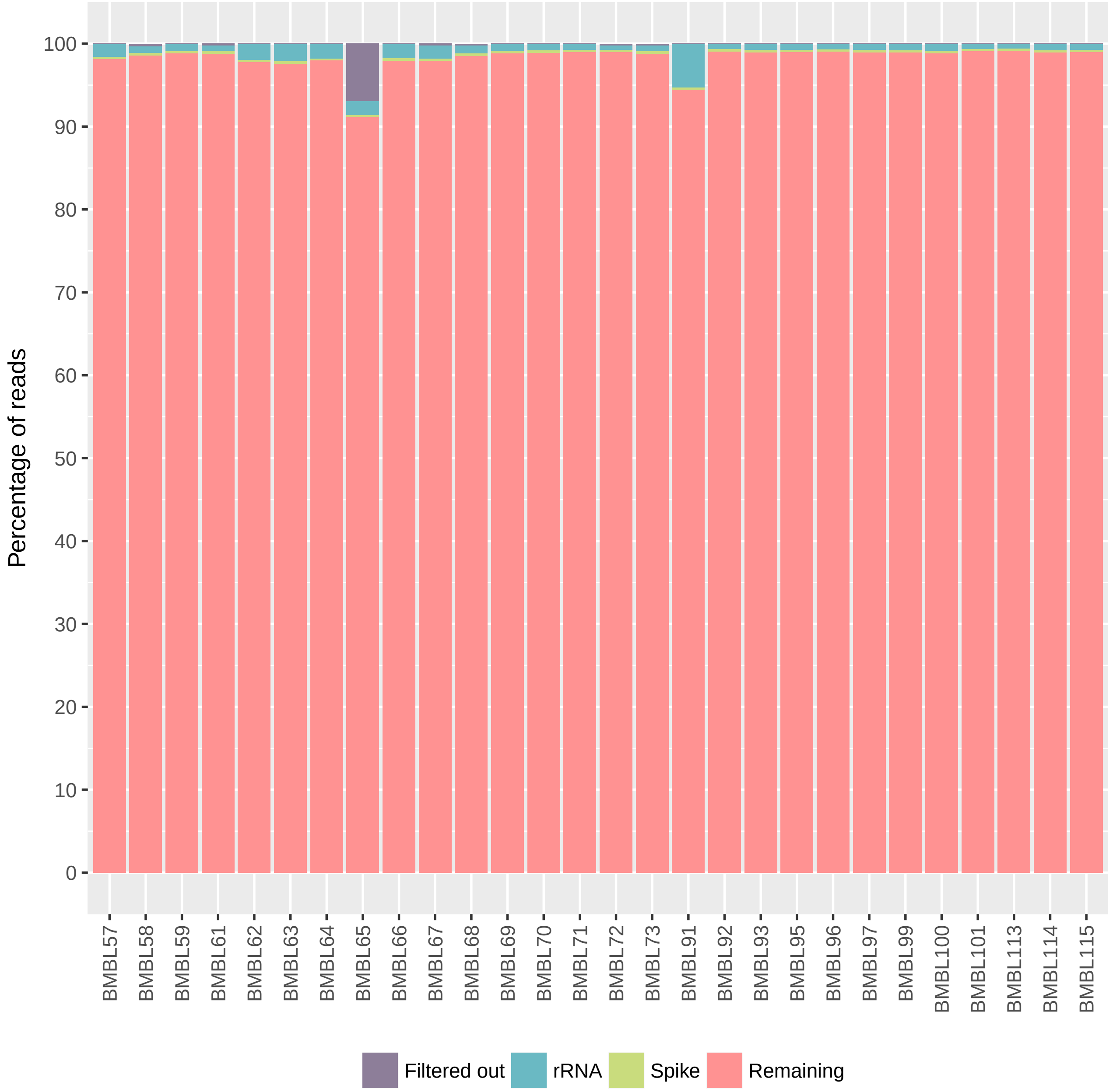

Figure 2: **Summary of preprocessing results.** "Filtered out" represents the percentage of reads shorter than 40 bases discarded after adapter and low-quality sequences (Phred quality score below 20) removal."rRNA" represents the percentage of reads mapping to rRNA sequences. "Spike" represents the percentage of reads mapping to spike-in sequences. "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.

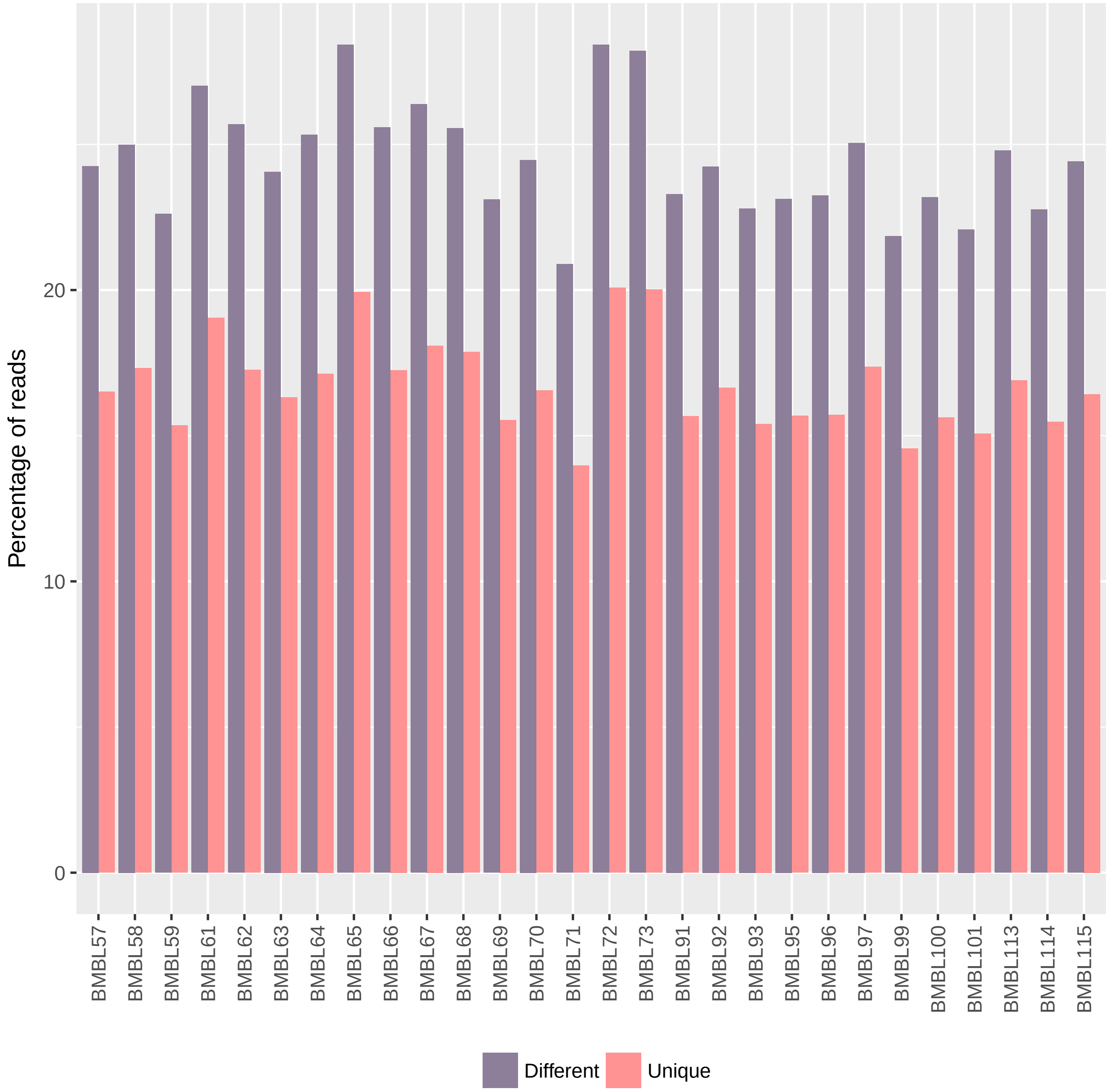

Figure 3: **Percentage of different and unique reads in each sample.** These percentages were calculated relative to the number of reads after preprocessing step.

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

3 Mapping

Reads were mapped onto the mm10 assembly of *Mus musculus* genome using STAR [3] version 2.5.3a. Figure 4 on the following page provides a summary of mapping results.

Figure 5 on page 7 represents read coverage over genes in all samples (coverage was computed for each gene percentile using geneBodyCoverage from RSeQC [4] version 2.6.4).

4 Quantification

Gene expression quantification was performed from uniquely aligned reads using htseq-count [6] version 0.6.1p1, with annotations from Ensembl version 91 and "union" mode[2]. Figure 6 on page 8 provides a summary of quantification results. Only non-ambiguously assigned reads have been retained for further analyses.

5 Data exploration

Figure 7 on page 9 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 8 on page 10 represents represents the first principal components of a Principal Component Analysis, showing the main sources of variance in the data.

6 Differential gene expression analysis

Comparisons of interest (listed in Table 2) were performed using the test for differential expression indicated in Table 2, proposed by Love et al. [8] and implemented in the Bioconductor package DESeq2 version 1.16.1.

Table 2: **Differential expression comparisons performed.**

Name of the comparison	Statistical test	Variable(s) taken into account in the model	Levels compared	Sample(s) filtered out
KO8w_NaCl vs Ctrl8w_NaCl	Wald	Condition	KO8w_NaCl vs Ctrl8w_NaCl	None
KO8w_AAV vs Ctrl8w_NaCl	Wald	Condition	KO8w_AAV vs Ctrl8w_NaCl	None
KO8w_NaCl vs KO8w_AAV	Wald	Condition	KO8w_NaCl vs KO8w_AAV	None
KO3w vs Ctrl3w	Wald	Condition	KO3w vs Ctrl3w	None
Ctrl8w_NaCl vs Ctrl3w	Wald	Condition	Ctrl8w_NaCl vs Ctrl3w	None
KO8w_NaCl vs KO3w	Wald	Condition	KO8w_NaCl vs KO3w	None

Genes with high Cook's distance were filtered out. Cook's distance is a measure of how much a single sample is influencing the fitted coefficients for a gene, and a large value of Cook's distance is intended to indicate an outlier count. These genes have no p-value in the resulting file.

Independent filtering based on the mean of normalized counts was performed in order to filter out those genes that have no or little chance of showing significance evidence of differential expression (without looking at their statistic). Indeed, genes with very low counts in all samples are not likely to be significantly

[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 $\{A, B, C, D, E, F, G\}$.

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

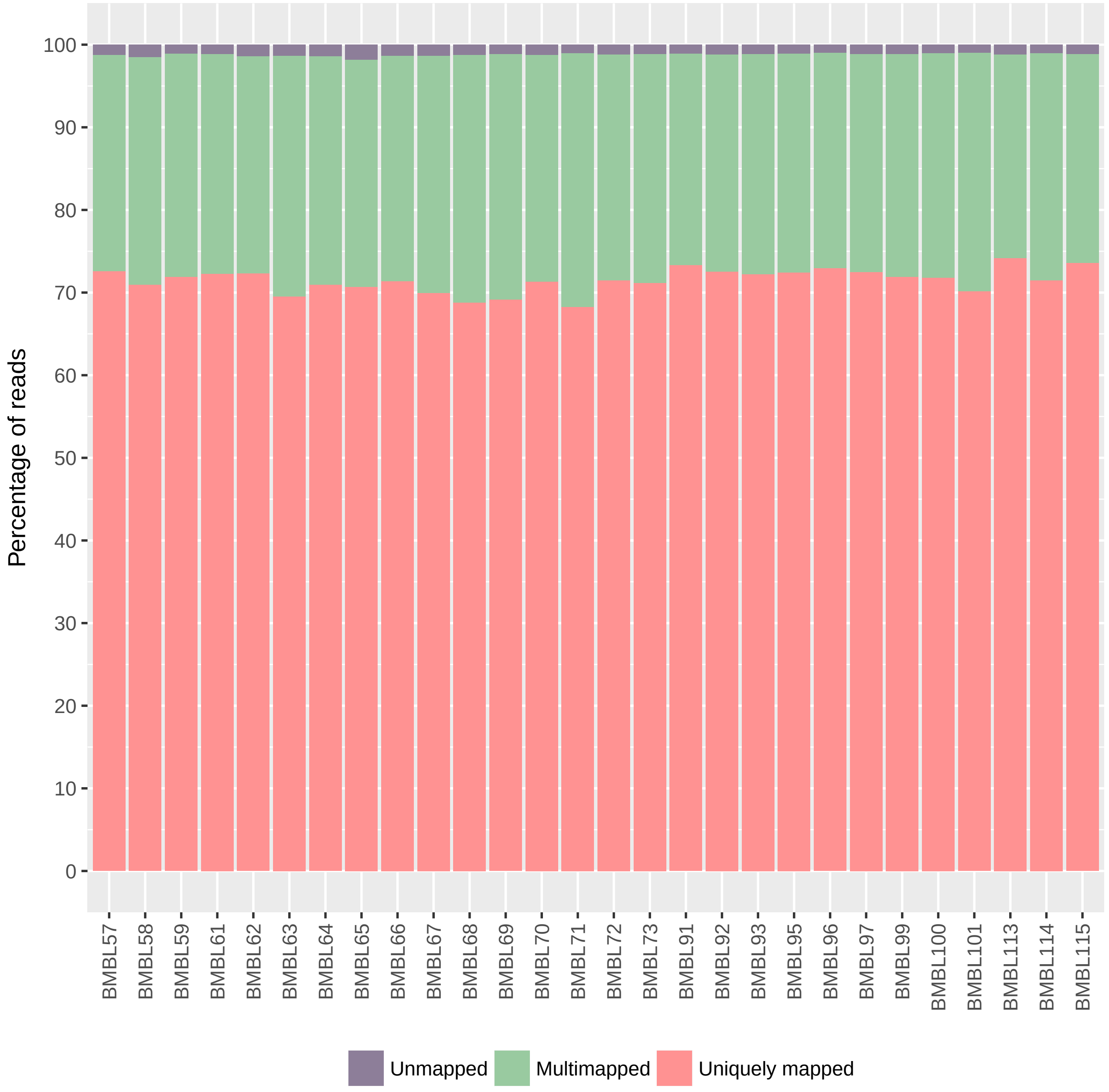

Figure 4: **Summary of mapping results.** This barplot represents the percentage of reads mapped only once on the genome (uniquely mapped), mapped at several locations on the genome (multi-mapped), or not mapped onto the genome (unmapped). These percentages were calculated relative to the number of input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

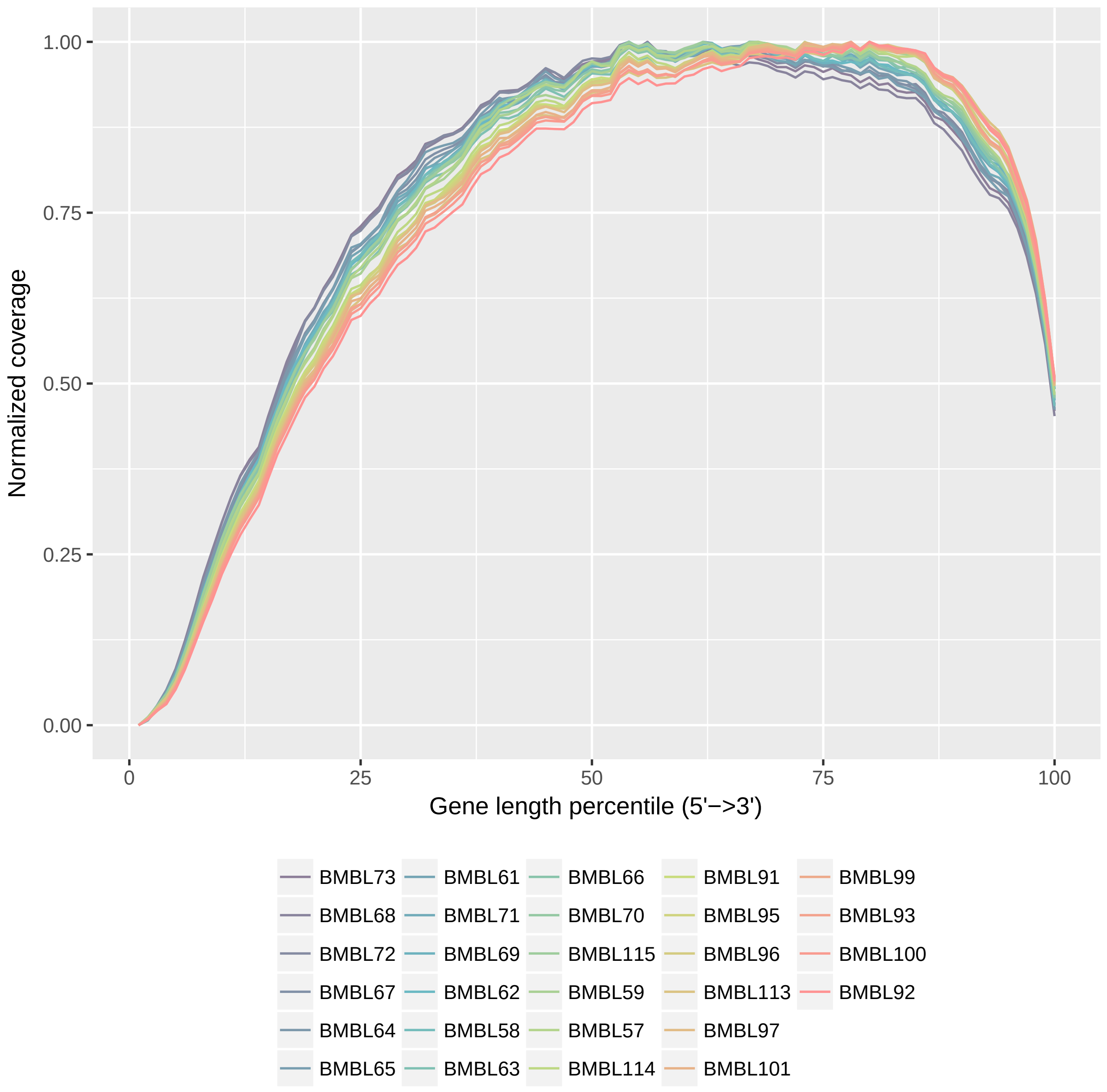

Figure 5: **Read coverage over genes in all samples.** This plot represents the normalized coverage (y-axis) at all percentiles of gene length (x-axis). Genes with mRNA length below 100bp were skipped from this analysis. In the legend, samples are ordered according to their Pearson's skewness coefficient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

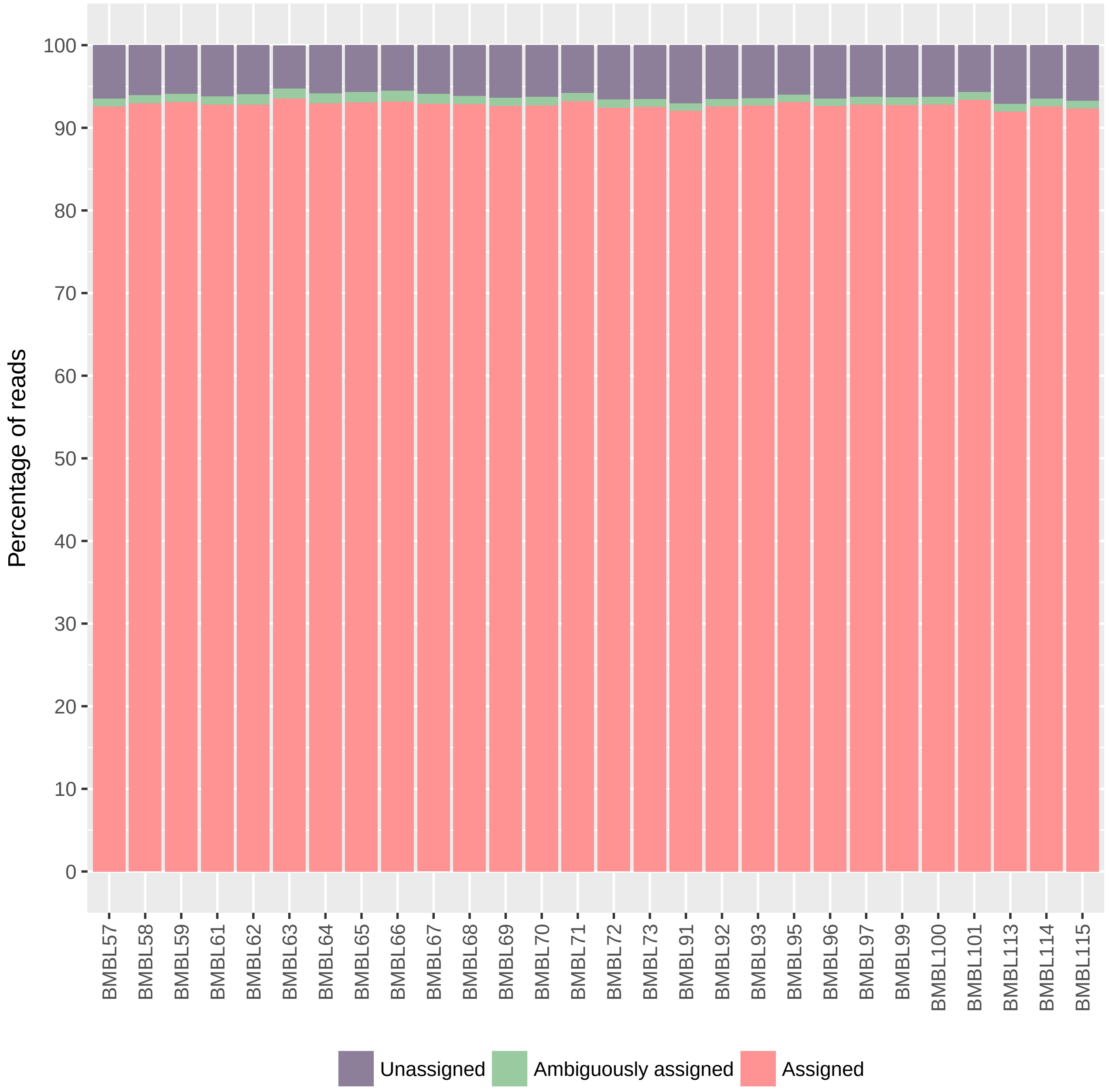

Figure 6: **Summary of quantification results.** This barplot represents the proportion of reads aligned to a transcribed region corresponding to one annotated gene (Assigned), to more than one annotated gene (Ambiguously assigned) or to no annotated gene (Unassigned), among all uniquely aligned reads.

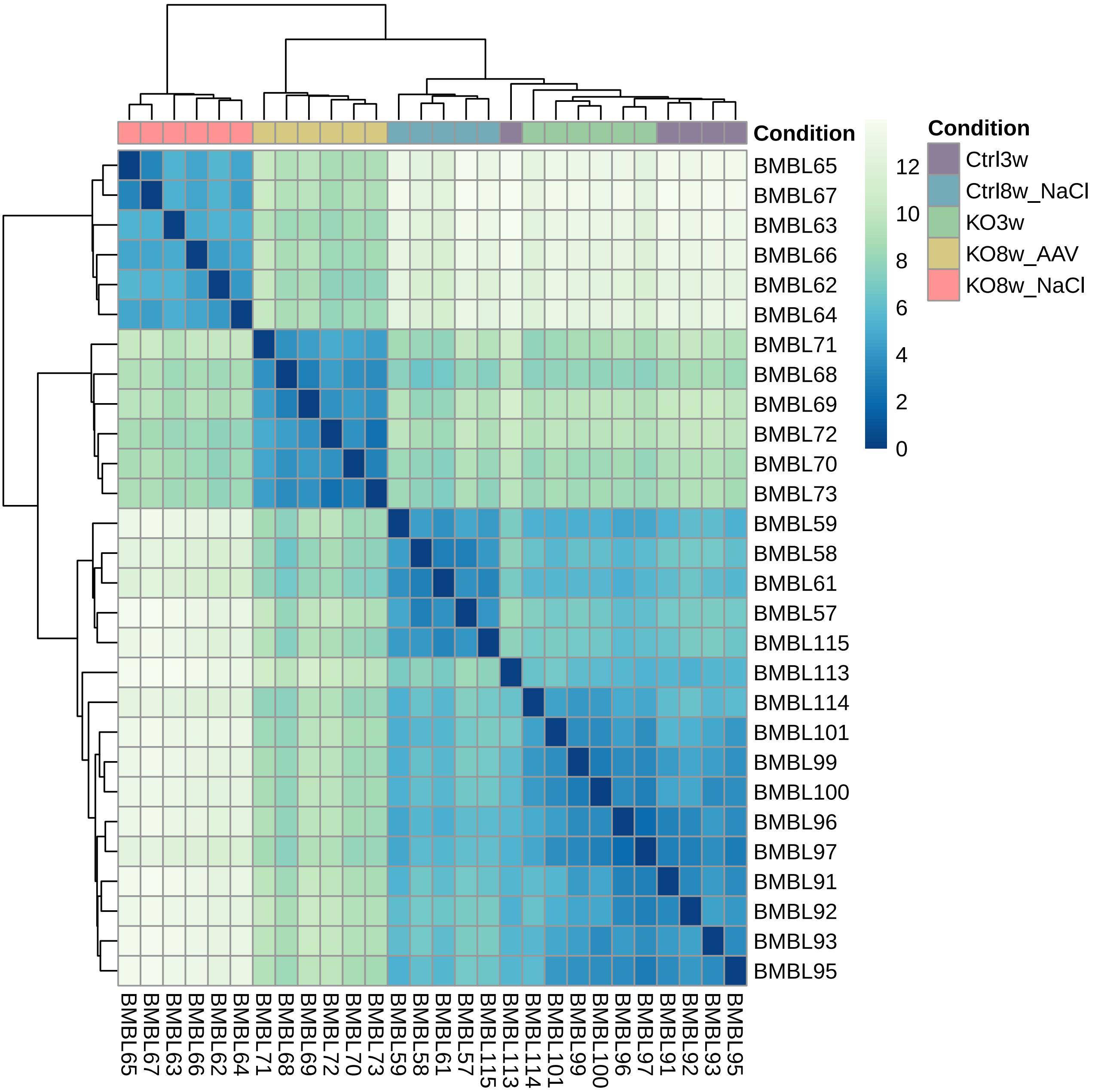

Figure 7: **Heatmap of sample-to-sample distances.** Sample-to-sample distances correspond to SERE [7] coefficient. Hierarchical clustering was performed using the Unweighted Pair Group Method with Arithmetic mean (UPGMA) algorithm.

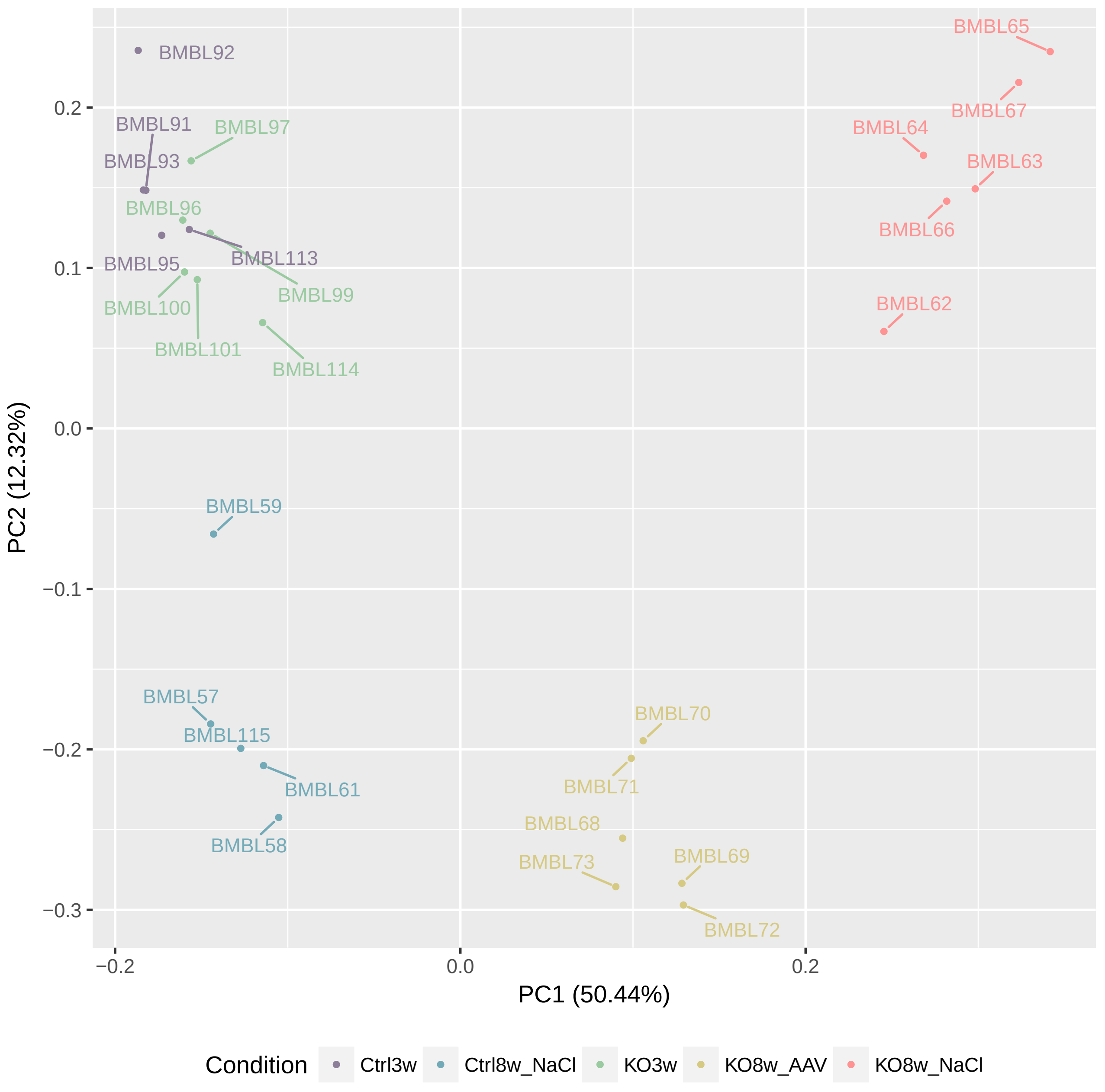

Figure 8: **Principal component analysis.** PCi axis represents the principal component i and the number into brackets indicates the percentage of explained variance associated with this axis. Principal Component Analysis was computed on regularized logarithm transformed data calculated with the method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

differentially expressed. This independent filtering results in increased detection power. P-values were adjusted for multiple testing using the Benjamini and Hochberg method [9]. Genes filtered out in the independent filtering step have no adjusted p-value in the resulting file.

Table 3: **Number of significantly differentially expressed genes.** These genes were selected using the following thresholds: adjusted p-value lower than 0.05 and absolute value of log2 Fold-Change greater than 1.

Name of the comparison (A vs B)	Number of over-expressed genes (A>B)	Number of under-expressed genes (A<B)	Number of significantly differentially expressed genes
KO8w_NaCl vs Ctrl8w_NaCl	1652	1527	3179
KO8w_AAV vs Ctrl8w_NaCl	1044	420	1464
KO8w_NaCl vs KO8w_AAV	651	814	1465
KO3w vs Ctrl3w	59	10	69
Ctrl8w_NaCl vs Ctrl3w	493	296	789
KO8w_NaCl vs KO3w	1805	1508	3313

Tables 3 provides the number of significantly differentially expressed genes in all comparisons. Figures 9 on the following page to 14 on page 17 represent the results of these comparisons.

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:

Ensembl Gene ID Ensembl identifier of the gene, corresponding to Ensembl release 91.

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.

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

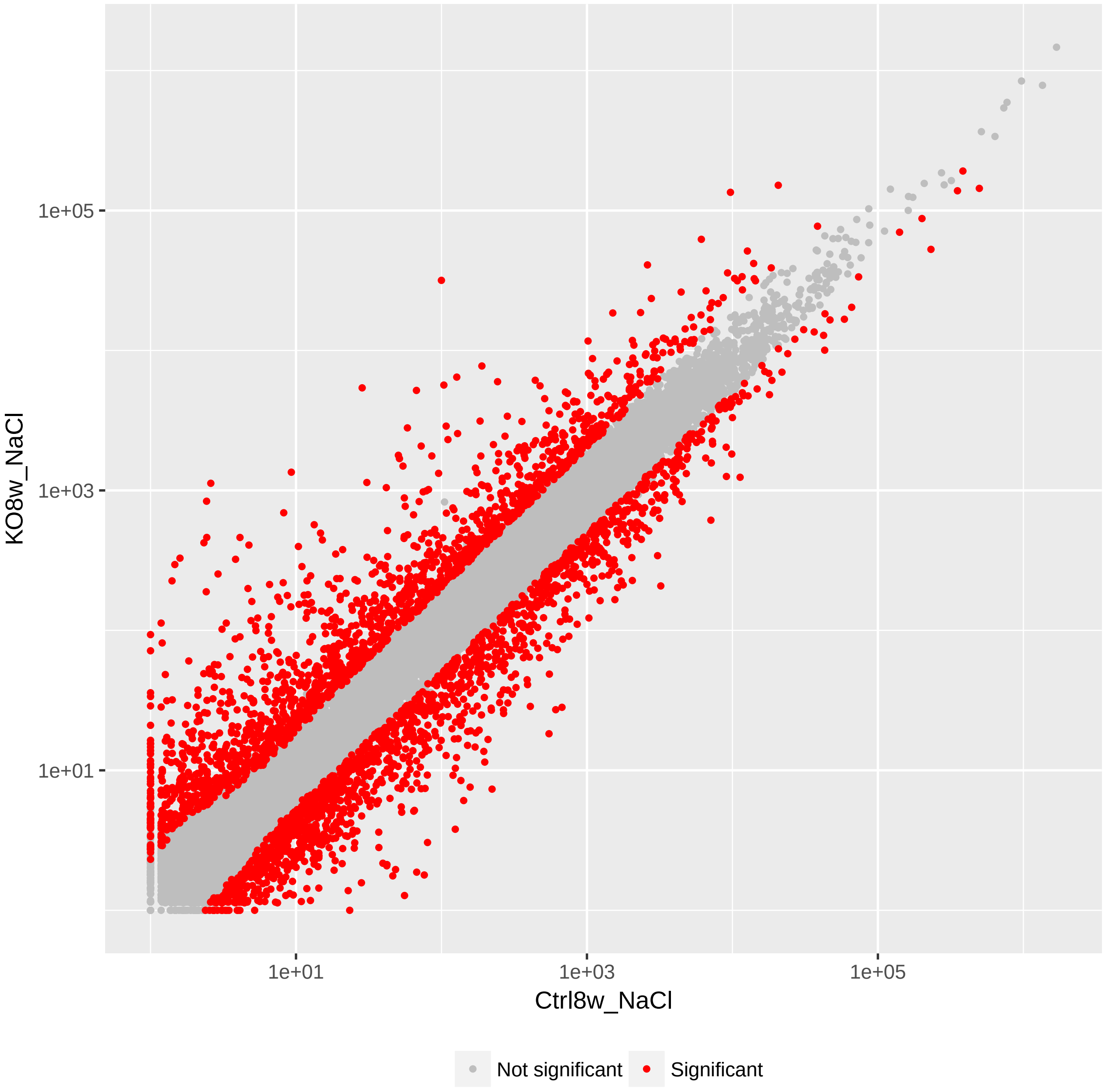

Figure 9: **KO8w_NaCl vs Ctrl8w_NaCl comparison.** Scatter-plot comparing the mean normalized counts for each condition. A pseudocount of 1 was added to all values in order to represent genes that are not expressed in one condition. Significant genes were selected using the following thresholds: adjusted p-value lower than 0.05 and absolute value of log2 Fold-Change greater than 1.

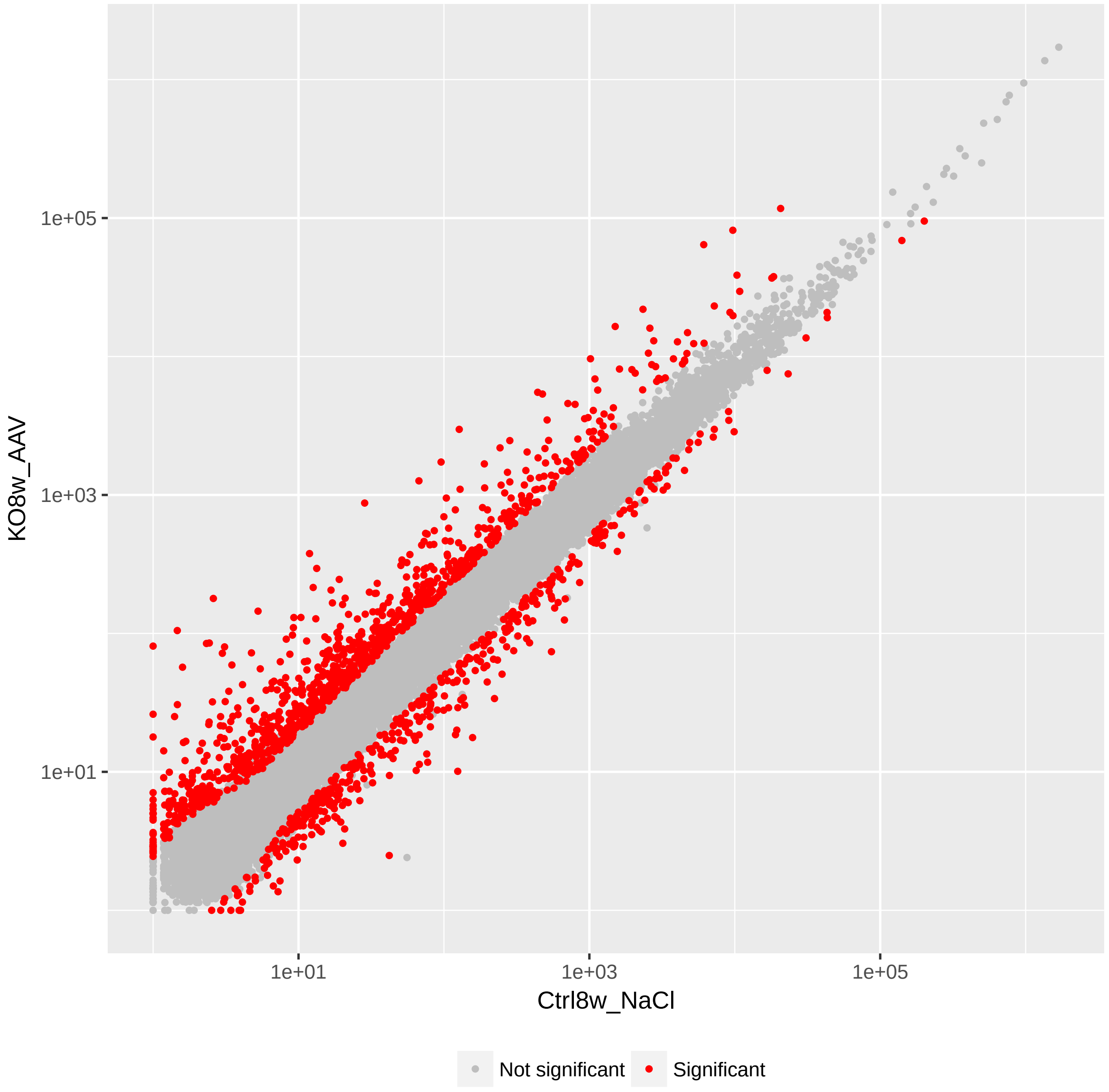

Figure 10: **KO8w_AAV vs Ctrl8w_NaCl comparison.** Scatter-plot comparing the mean normalized counts for each condition. A pseudocount of 1 was added to all values in order to represent genes that are not expressed in one condition. Significant genes were selected using the following thresholds: adjusted p-value lower than 0.05 and absolute value of log2 Fold-Change greater than 1.

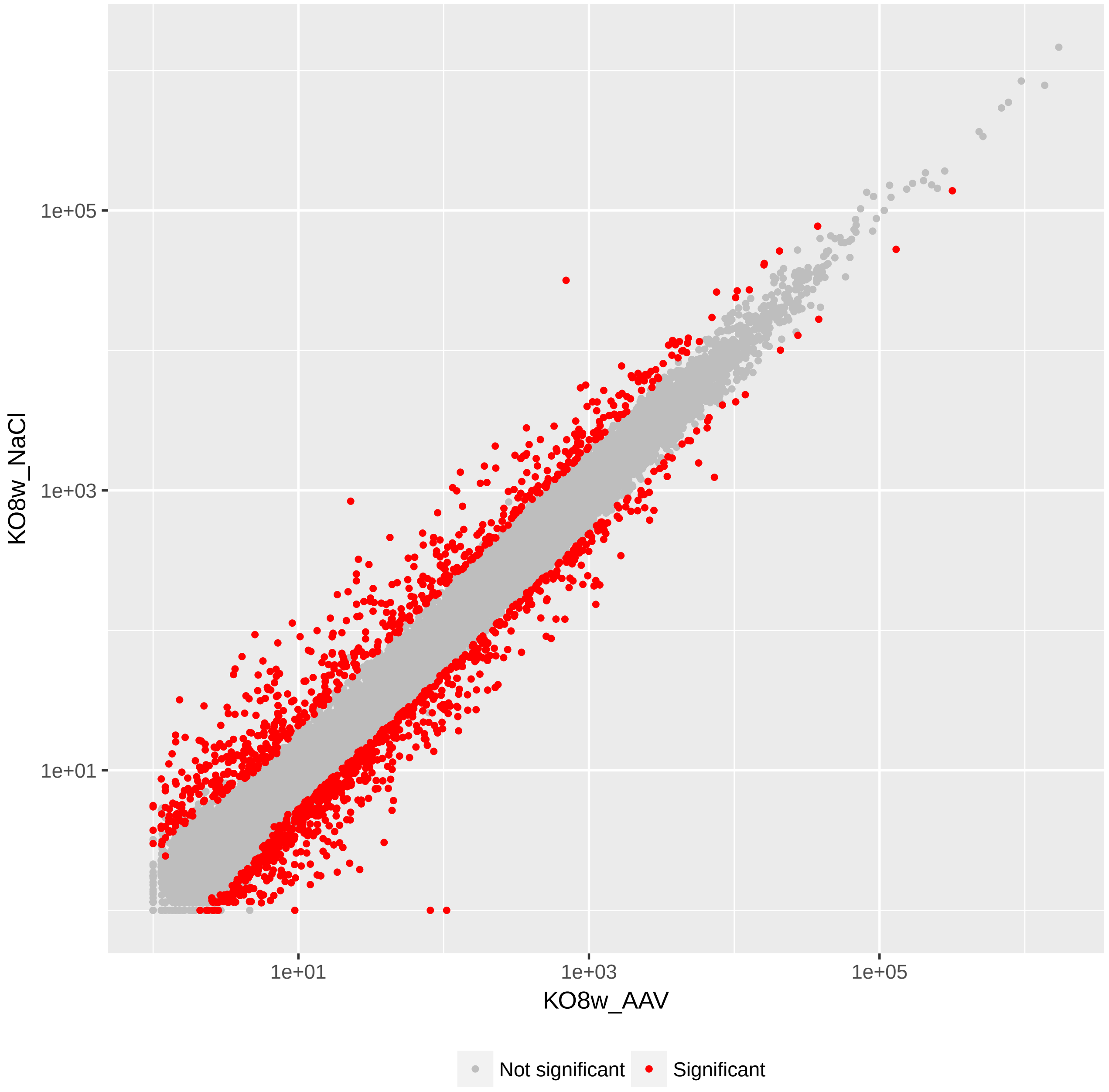

Figure 11: **KO8w_NaCl vs KO8w_AAV comparison.** Scatter-plot comparing the mean normalized counts for each condition. A pseudocount of 1 was added to all values in order to represent genes that are not expressed in one condition. Significant genes were selected using the following thresholds: adjusted p-value lower than 0.05 and absolute value of log2 Fold-Change greater than 1.

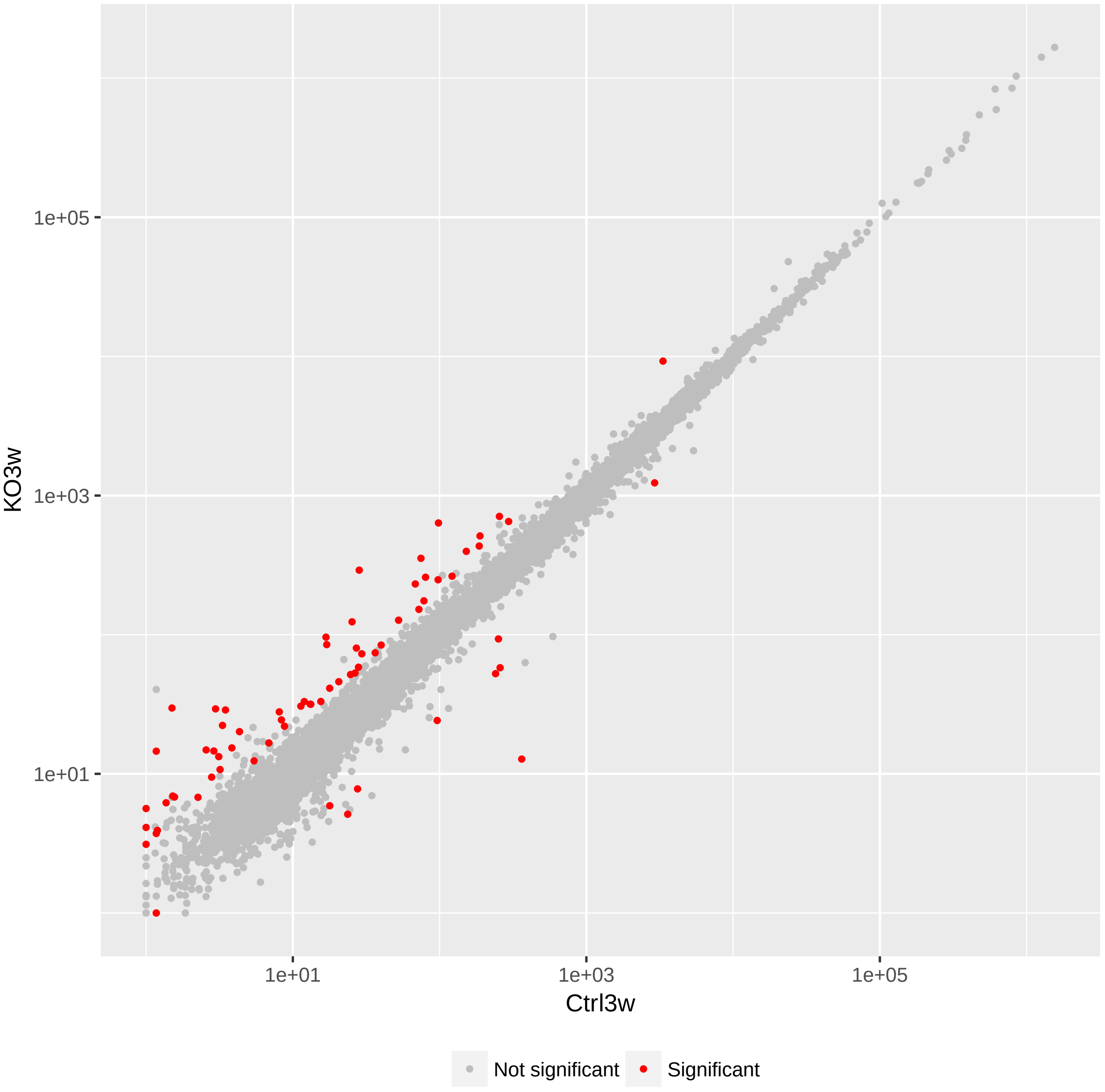

Figure 12: **KO3w vs Ctrl3w comparison.** Scatter-plot comparing the mean normalized counts for each condition. A pseudocount of 1 was added to all values in order to represent genes that are not expressed in one condition. Significant genes were selected using the following thresholds: adjusted p-value lower than 0.05 and absolute value of log2 Fold-Change greater than 1.

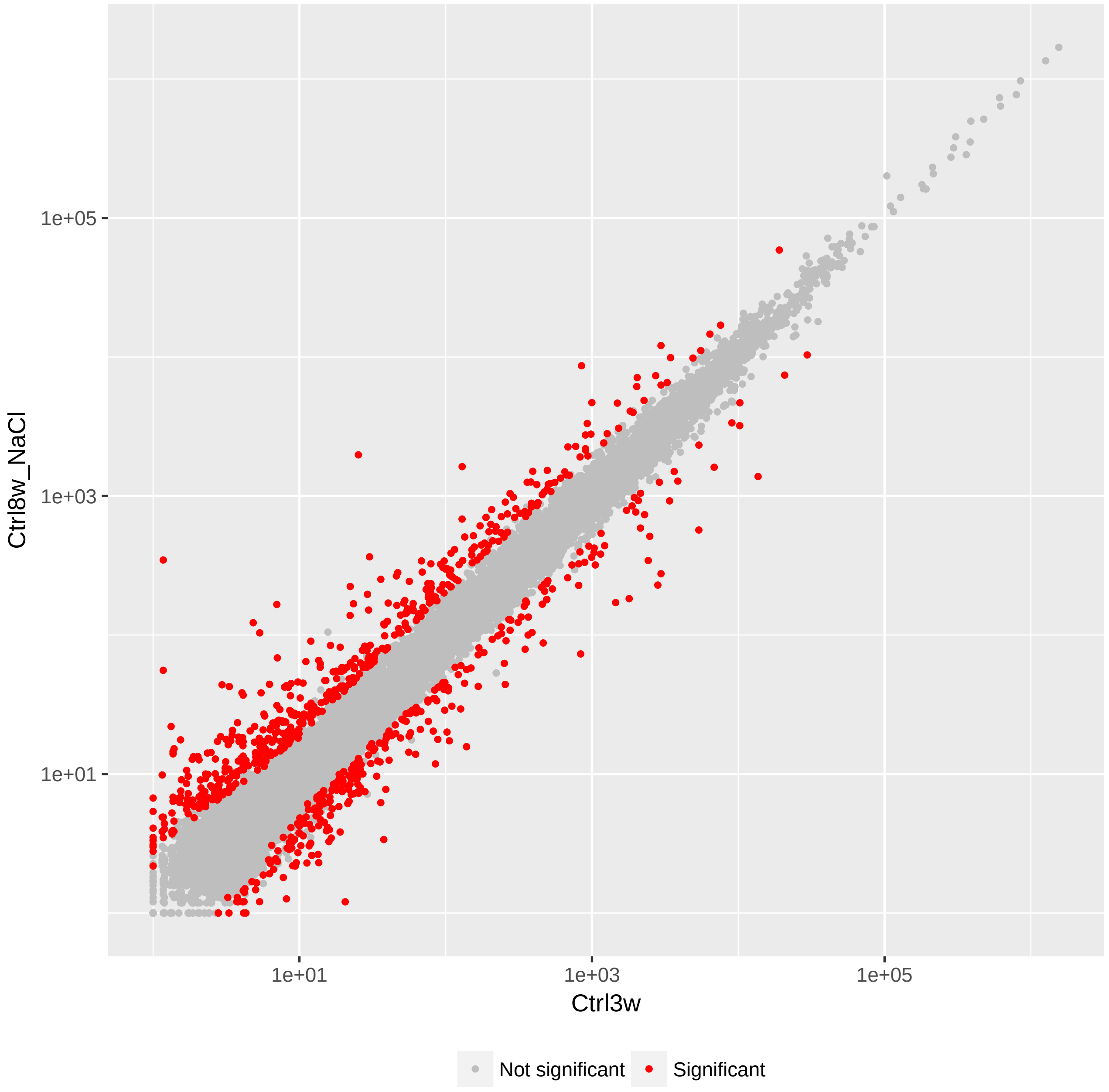

Figure 13: **Ctrl8w_NaCl vs Ctrl3w comparison.** Scatter-plot comparing the mean normalized counts for each condition. A pseudocount of 1 was added to all values in order to represent genes that are not expressed in one condition. Significant genes were selected using the following thresholds: adjusted p-value lower than 0.05 and absolute value of log2 Fold-Change greater than 1.

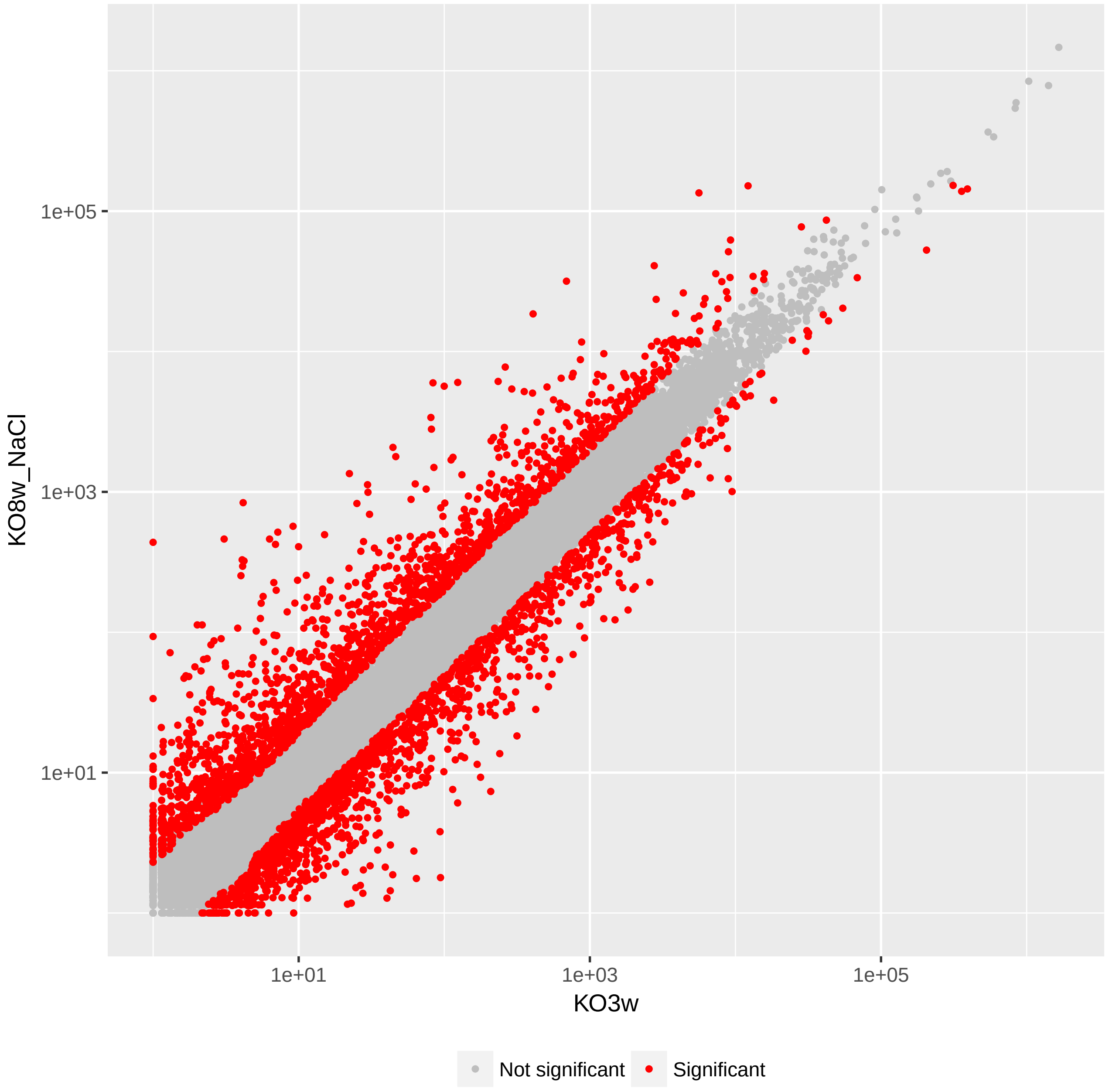

Figure 14: **KO8w_NaCl vs KO3w comparison.** Scatter-plot comparing the mean normalized counts for each condition. A pseudocount of 1 was added to all values in order to represent genes that are not expressed in one condition. Significant genes were selected using the following thresholds: adjusted p-value lower than 0.05 and absolute value of log2 Fold-Change greater than 1.

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.

log2 FC Log2 of the expression fold change estimated, reflecting differential expression between the two compared conditions.

p-value P-value of the statistical test.

Adjusted p-value P-value of the statistical test, adjusted for multiple testing.

8 Version information

8.1 Version of used tools

Table 4 provides the tools used in GenomEast RNA-seq pipeline version 1.1.2 (used to perform the analyses described in the report) and their corresponding version.

Table 4: **Tools used for the analyses presented in this report.**

Tool	Release	Description
bowtie2	2.2.8	To align reads onto a set of reference sequences.
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.

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
```

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

```
##  [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
## [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] DESeq2_1.16.1              SummarizedExperiment_1.4.0
##  [3] Biobase_2.34.0             GenomicRanges_1.26.4
##  [5] GenomeInfoDb_1.10.3        IRanges_2.8.2
##  [7] S4Vectors_0.12.2           BiocGenerics_0.22.0
##  [9] ggrepel_0.6.5              ggfortify_0.4.1
## [11] pheatmap_1.0.8             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] locfit_1.5-9.1       Rcpp_0.12.13         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] annotate_1.50.0      data.table_1.10.4    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] B. Langmead and S. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nat Methods*, vol. 9, no. 4, pp. 357–359, 2012.

[3] 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.

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

[5] 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.

[6] 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.

[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.

[9] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *J R Statist Soc B*, vol. 57, no. 1, pp. 289–300, 1995.