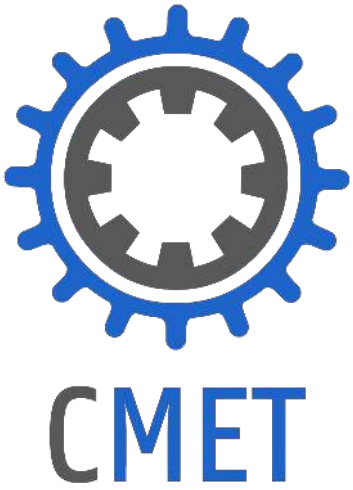

Versatile human *in vitro* triple coculture model coincubated with adhered gut microbes reproducibly mimics pro-inflammatory host-microbe interactions in the colon

S5: Processing whole-transcript gene expression assay data from triple coculture human cell models generated with GeneChip Human Gene 2.1 ST microarrays in R.

Version July 4, 2021

Annelore Beterams
Marta Calatayud Arroyo
Kim De Paepe
Laure Maes
India Jane Wise
Herlinde De Keersmaecker
Andreja Rajkovic
Debby Laukens
Tom Van de Wiele

Contents

List of Figures

List of Tables

Chapter 1

Affymetrix Human Gene 2.1 ST Array Strip

Affymetrix Hugene microarrays contain a high density of oligonucleotide probes to profile human gene expression through hybridization with labeled cDNA obtained from mRNA transcripts. Each transcript is detected by a particular group of 25mer probe sequences (median of 21 unique probes), known as a probeset/probe-group [11]. In total, the HuGene ST array consists of $>$ 1.35 million probes (Table 1.1).

Signals from multiple individual probes targeting one transcript need to be merged into a single measure of expression for a gene in a process called summarization, which is discussed in detail below [11].

1.1 Triple coculture human cell model

A reproducible, versatile triple coculture *in vitro* model consisting of immune-like, goblet and epithelial cells in indirect contact with attached gut microbes was developed to permit the mechanistic study of complex microbiota-host interactions in the human gut. The model consisted of a collagen coating with confluent THP-1, LS-174T and T84 cells in DMEM/F-12 at the basolateral side, in indirect contact with an apical bacterial biofilm adhered to a solidified agar-mucin layer covering the Transwell polycarbonate membrane filter (0.4 μm pores). Microbial biofilms were composed of gut microbes derived from a distal colon compartment of the Simulator of the

Table 1.1: Affymetrix Human Gene 2.1 ST Array Strip design specifications.

Total probes	>1.35 million
Exon-level probe sets	>418,000
Gene-level probe sets	>48,000
NM and XM – RefSeq coding transcript, well-established and provisional annotations	>33,500
NR and XR – RefSeq non-coding transcript, well-established and provisional annotations	>6,500
Total RefSeq transcripts	>40,000
RS (Entrez) gene count	>25,000
lincRNA transcripts Derived from the Broad Institute's Human Body Map lincRNAs and TUCP (transcripts of uncertain coding potential) catalog and lncRNA db	>11,000
ERCC probe sets1	92
Background probes	Antigenomic set
Poly-A controls	dap, lys, phe, thr, trpn
Hybridization controls	BioB, BioC, BioD, CreX

Human Intestinal Microbial Ecosystem (SHIME), with or without *Lactobacillus rhamnosus* GG LMG 18243 (LGG) addition. Transcriptome samples of the 16 hours coincubated triple coculture model with either SHIME or SHIME + LGG microbiota were compared to an untreated Blank condition. For more details, see Materials & Methods in the manuscript entitled 'Versatile human *in vitro* triple coculture model coincubated with adhered gut microbes reproducibly mimics pro-inflammatory host-microbe interactions in the colon' by Beterams et al. 2021.

1.2 Example data

Besides the acquired data form the triple cell coculture model, Affymetrix example data was analyzed in parallel acting as a reference to benchmark the obtained quality metrics. The example dataset consists of quadruplicate microarrays hybridized with liver, muscle, spleen and testes tissue RNA, alongside 100% of Stratagene's Universal Human Reference RNA and 100% of Ambion's Human Brain Reference RNA. The latter two are termed MAQCA and MAQCB, and are derived from the Microarray Quality Consortium (MAQC) [4].

Chapter 2

Read in .CEL files, annotation and library files

Microarray data was loaded and further processed in `R version 4.1.0 (2021-05-18)` running on a `x86\\_64-pc-linux-gnu` platform.

2.1 CEL files

The transcriptomics data consists of a series of CEL files containing raw intensities for each probe on the mircoarray (probe level intensity data on a per-chip basis) generated by the Affymetrix software. Data was imported as a GeneFeatureSet using the Oligo package (version 1.56.0). Data transformations are not applied upon data import, but were manually performed afterwards.

Metadata is provided in the phenodata field of the GeneFeatureSet. The following variables were included:

```
##  [1] "SampleName"          "Condition"
##  [3] "Replicate"           "LGG"
##  [5] "SHIME"               "HumanCells"
##  [7] "IL8"                 "RelativeTNFalpha"
##  [9] "RelativeIL1beta"     "RelativeIL10"
## [11] "FCM"                 "LGG_spike_count"
## [13] "LDH"                 "NormalizedResazurin"
## [15] "Mucus"               "Acetic_acid"
## [17] "Propionic_acid"      "Butyric_acid"
## [19] "Isobutyric_acid"     "Isovaleric_acid"
## [21] "Valeric_acid"        "Isocaproic_acid"
## [23] "Caproic_acid"        "Heptanoic_acid"
## [25] "Octanoic_acid"       "Total_SCFA"
```

2.2 Annotation files

Probe and probeset IDs (also called feature and feature set ID), as well as, probe sequence info are contained in the GeneFeatureSet created from the .CELdata thanks to the associated annotation package: 'pd.hugene.2.1.st'). This annotation package is based on the 'HuGene-2_1-st-v1.na36.hg19.probeset.csv' and 'HuGene-2_1-st-v1.na36.hg19.transcript.csv' files. Annotation (and summarization, see below) can be performed either at probeset (target = probeset; corresponding to an exon) or at transcript (target = core; corresponding to a gene) level. Traditionally, Affymetrix arrays (the so-called 3' IVT arrays) were probeset based: a certain fixed group of probes were part of a probeset which represented a certain gene or transcript (note however, that a gene can be represented by multiple probesets). The more recent 'Gene' and 'Exon' Affymetrix arrays are exon based and hence there are two levels of summarization to get to the gene level. The 'probeset' summarization leads to the exon level. The gene/transcript level is given by 'transcript clusters' [3].

Feature IDs ('fid') are assigned to individual features or probes (sequences on the microarray) and are identical between exon and gene level analysis, while the feature set IDs ('man_fsetid') differ for the exon or transcript level analysis. Feature set IDs are not unique (all the probes within a feature set have the same feature set ID). Note that feature IDs are also not unique, as the same fid can occur multiple times, in different probesets/transcript clusters. Therefore a unique identifier combining the feature and feature set IDs is constructed.

```
##             fid man_fsetid   fsetid  x y chrom type
## 16802771_6    6   16802771 16802771  5 0 chr15 main
## 16755904_7    7   16755904 16755904  6 0 chr12 main
## 17120602_9    9   17120602 17120602  8 0  chrU <NA>
## 16817074_10  10   16817074 16817074  9 0 chr16 main
## 17081850_12  12   17081850 17081850 11 0  chr8 main
## 17121732_15  15   17121732 17121732 14 0  chrU <NA>
##                uniqueid                 sequence
## 16802771_6   16802771_6 TCTAAGAGAAATCAACGGGCCGGCG
## 16755904_7   16755904_7 ACACGGTGGGAGTCTGGAATAATGT
## 17120602_9   17120602_9 ACAAAACACGTGGAACCTTGGGTAG
## 16817074_10 16817074_10 TTCAGTTCAGGGTGTCCACTGGGTC
## 17081850_12 17081850_12 TGGCTAGATGACCATCCCGACAGGG
## 17121732_15 17121732_15 CGAGGGAGATCCCTTCATGAGGGTC
##             fid man_fsetid   fsetid  x y chrom type
## 16802772_6    6   16802772 16802772  5 0 chr15 main
## 16755905_7    7   16755905 16755905  6 0 chr12 main
## 17120603_9    9   17120603 17120603  8 0  chrU <NA>
## 16817075_10  10   16817075 16817075  9 0 chr16 main
## 17081852_12  12   17081852 17081852 11 0  chr8 main
## 17121733_15  15   17121733 17121733 14 0  chrU <NA>
##                uniqueid                 sequence
## 16802772_6   16802772_6 TCTAAGAGAAATCAACGGGCCGGCG
## 16755905_7   16755905_7 ACACGGTGGGAGTCTGGAATAATGT
## 17120603_9   17120603_9 ACAAAACACGTGGAACCTTGGGTAG
## 16817075_10 16817075_10 TTCAGTTCAGGGTGTCCACTGGGTC
## 17081852_12 17081852_12 TGGCTAGATGACCATCCCGACAGGG
## 17121733_15 17121733_15 CGAGGGAGATCCCTTCATGAGGGTC
```

There are 352859 probesets and 53617 transcript clusters. The mapping of probesets to transcript cluster IDs can be viewed in the probeset csv file. In the transcript csv file the probeset ID is identical to the transcript cluster ID. The following annotation columns are present:

```
##  [1] "transcript_cluster_id" "probeset_id"
##  [3] "seqname"               "strand"
##  [5] "start"                 "stop"
##  [7] "total_probes"          "gene_assignment"
##  [9] "mrna_assignment"       "swissprot"
## [11] "unigene"               "GO_biological_process"
## [13] "GO_cellular_component" "GO_molecular_function"
## [15] "pathway"               "protein_domains"
## [17] "crosshyb_type"         "category"
##  [1] "probeset_id"
##  [2] "seqname"
##  [3] "strand"
##  [4] "start"
##  [5] "stop"
##  [6] "probe_count"
##  [7] "transcript_cluster_id"
##  [8] "exon_id"
##  [9] "psr_id"
## [10] "gene_assignment"
## [11] "mrna_assignment"
## [12] "crosshyb_type"
## [13] "number_independent_probes"
## [14] "number_cross_hyb_probes"
## [15] "number_nonoverlapping_probes"
## [16] "level"
## [17] "bounded"
## [18] "noBoundedEvidence"
## [19] "has_cds"
## [20] "fl"
## [21] "mrna"
## [22] "est"
## [23] "vegaGene"
## [24] "vegaPseudoGene"
## [25] "ensGene"
## [26] "sgpGene"
## [27] "exoniphy"
## [28] "twinscan"
## [29] "geneid"
## [30] "genscan"
## [31] "genscanSubopt"
## [32] "mouse_fl"
## [33] "mouse_mrna"
## [34] "rat_fl"
## [35] "rat_mrna"
## [36] "microRNAregistry"
## [37] "rnaGene"
## [38] "mitomap"
```

```
## [39] "probeset_type"
```

2.3 Library files

Additional information about the probes is contained in the Affymetrix library files which are also integrated in the oligo annotation package: citation ('pd.hugene.2.1.st'). The following additional information is provided in these files:

```
##  [1] "Probe.ID"              "Transcript.Cluster.ID"
##  [3] "probe.x"               "probe.y"
##  [5] "assembly"              "seqname"
##  [7] "start"                 "stop"
##  [9] "strand"                "probe.sequence"
## [11] "target.strandedness"   "category"
##  [1] "probeset_id"    "probeset_type"  "atom_id"
##  [4] "probe_id"       "probe_type"     "gc_count"
##  [7] "probe_length"   "probe_sequence" "x"
## [10] "y"
## [1] "pm:st"
## [1] "probeset_id"           "transcript_cluster_id"
## [3] "probeset_list"         "probe_count"
##         V1                V2       V3    V4 V5 V6 V7
## 1 16650001 normgene->intron 16650001       NA NA NA
## 2       NA                 1                NA NA NA
## 3       NA                     233137 pm:st  7 25 13
## 4       NA                 2                NA NA NA
## 5       NA                    1029116 pm:st  8 25 13
## 6       NA                 3                NA NA NA
## 7       NA                     395457 pm:st  7 25 13
## 8       NA                 4                NA NA NA
## 9       NA                    1075670 pm:st  9 25 13
##                          V8
## 1
## 2
## 3 TTTTTCTAAGAATCCTAGTTAACTG
## 4
## 5 GTTAACTGTACGAAGTGAAAATTGT
## 6
## 7 CTGTACGAAGTGAAAATTGTAATTT
## 8
## 9 CCTAGTTAACTGTACGAAGTGAAAA
##            V1   V2
## 1 probeset_id type
##            V1                          V2              V3 V4
## 5225 17127633 control->affx->polya_spike AFFX-TrpnX-3_st  1
##  [1] "AFFX-BioB-3_at"        "AFFX-BioB-5_at"
##  [3] "AFFX-BioB-M_at"        "AFFX-BioC-3_at"
```

```
##  [5] "AFFX-BioC-5_at"        "AFFX-BioDn-3_at"
##  [7] "AFFX-BioDn-5_at"       "AFFX-BkGr-GC12_st"
##  [9] "AFFX-CreX-3_at"        "AFFX-CreX-5_at"
## [11] "AFFX-DapX-3_st"        "AFFX-DapX-5_st"
## [13] "AFFX-DapX-M_st"        "AFFX-LysX-3_st"
## [15] "AFFX-LysX-5_st"        "AFFX-LysX-M_st"
## [17] "AFFX-PheX-3_st"        "AFFX-PheX-5_st"
## [19] "AFFX-PheX-M_st"        "AFFX-ThrX-3_st"
## [21] "AFFX-ThrX-5_st"        "AFFX-ThrX-M_st"
## [23] "AFFX-TrpnX-3_st"       "AFFX-TrpnX-5_st"
## [25] "AFFX-TrpnX-M_st"       "AFFX-r2-Bs-dap-3_st"
## [27] "AFFX-r2-Bs-dap-5_st"   "AFFX-r2-Bs-dap-M_st"
## [29] "AFFX-r2-Bs-lys-3_st"   "AFFX-r2-Bs-lys-5_st"
## [31] "AFFX-r2-Bs-lys-M_st"   "AFFX-r2-Bs-phe-3_st"
## [33] "AFFX-r2-Bs-phe-5_st"   "AFFX-r2-Bs-phe-M_st"
## [35] "AFFX-r2-Bs-thr-3_s_st" "AFFX-r2-Bs-thr-5_s_st"
## [37] "AFFX-r2-Bs-thr-M_s_st" "AFFX-r2-Ec-bioB-3_at"
## [39] "AFFX-r2-Ec-bioB-5_at"  "AFFX-r2-Ec-bioB-M_at"
## [41] "AFFX-r2-Ec-bioC-3_at"  "AFFX-r2-Ec-bioC-5_at"
## [43] "AFFX-r2-Ec-bioD-3_at"  "AFFX-r2-Ec-bioD-5_at"
## [45] "AFFX-r2-P1-cre-3_at"   "AFFX-r2-P1-cre-5_at"
```

Some more information about the data fields:

- Probex en Probey refer to the x and y coordinates for probe locations on the array.
- Assembly refers to the genome assembly version used for the array design data: in this case the UCSC hg19 build which corresponds to the NCBI GRCh37 build was used. The design-time vs. annotation-time version of the genome assembly may differ since array designs are occasionally mapped or 'lifted' to an updated version of the genome assembly for improved annotation by the NetAffxAnalysis Center.
- Seqname refers to the sequence name for the genomic location of probes (chromosome).
- Start refers to the starting coordinate of probe genomic location (1-based).
- Stop refers to the ending coordinate of probe genomic location (1-based).
- Strand refers to the sequence strand of probe genomic location (+ or -).
- Probe sequence refers to the probe sequence.
- Sense/Antisense refers to the Strandedness of the target which the probe detects.
- Category(also called probeset type) refers to the array design category of the probe: this can be either main or different types of controls (outlined below).
- An atom refers to a particular collection of probes that are interrogating the same position. For expression arrays an atom is usually a probe pair (pm and mm probe pair) for arrays of the HG-U133 series, or a single pm probe for arrays like Human Gene or Exon ST that do not contain mismatch probes.
- Group name identifies the group(s) of which the probeset is a member.

Affymetrix ATP or Array Power Tools uses the library and annotation files and creates additional CDF (Chip Description Files) containing the Chip layout information (i.e. what probes are where on the chip and how are they grouped into probesets) and CHP files containing probeset summary information (ie probeset signal, gene-level information) on a per chip basis. In R a pdInfoPackage built using the pdInfoBuilder collates the cdf, probe and annotation data together: citation ('pd.hugene.2.1.st'). This .pd information can be queried through sql commands. Alternatively probenames ('fid'), probeset names ('fset_id'), probe information (x,y,chrom,type), probe sequences and GC content can be interrogated from the .CEL data through Oligo commands. Finally, the data can be reclassified post-normalization using ChipDb objects (explained hereinafter).

Chapter 3

Data transformations

Intensities are usually log2 transformed and data from the 'perfect match' probes is extracted. Note that background probes are also contained within the 'perfect match' data matrix. The use of MisMatch (MM) probes is highly controversial and discontinued in the newer microarray designs. As a consequence we did not detect a MM feature set in our data.

While no mismatch probes were present, the 'perfect match' dataframe is still reduced (1025088) compared to the original dataset (1416100). This is due to the presence of low and high magnitude probes having no biological significance. These probes on the edge of the array are intended to provide a reference pattern of roughly alternating bright and very bright signals that enable the subsequent analysis of the scanned GeneChip to correctly allocate signals to probes.

The 'perfect match' data can be annotated at probeset level (target = probeset) or at transcript/gene level (target=core) after summarization (discussed below).

After summarization, the ChipDb package ('hugene20sttranscriptcluster.db') can be used to expand the annotation of the microarray features. The ChipDb package differs from the above described pdInfo package ('pd.hugene.2.0.st') in that it contains only annotations that could be mapped to NCBI Gene IDs, while the former contains the relatively unprocessed annotation data from Affymetrix.

Chapter 4

Quality Control

Sources of variability and anomalies within and between arrays will be identified by looking at the distribution of log2 probe intensities through diagnostic plots (array images, box plots, density plots, intensity-sequence relationships, ROC curves, MA plots, model fit residual/weight plots), probe metrics (RLE, NUSE, DABG) and Principal Component Analysis [4]. Many of the quality assessment measures are highly correlated and perform well in identifying gross outliers that should be excluded from downstream analyses. Negative and positive control probes are included in the microarray designs. They can be used in conjunction with positive spike-ins (called external/exogenous controls) and a separate microarray loaded with a Hela cell RNA control sample to trace problems to specific stages of the microarray experiment. Example datasets from Affymetrix are available as a reference and were used to benchmark the analysis pipeline.

4.1 Mircoarray Pictures of log2 probe intensities

Log2-scaled pseudo-images were used to assess the spatial distribution of the individual probe intensities on the chips (Package Oligo version 1.56.0). These microarray pictures can unravel large inconsistencies on individual arrays and between replicated arrays.

A quick glance at the microarray pictures reveals different intensities across replicates (Figure 4.1). The blank3, SHIME2 and SHIME_LGG1 replicate appear dimmer compared to the other arrays. Some markings (darker blue areas with less probe binding/expression) in the corners and one spot in the middle of all arrays stand out. Similar markings were present in the Hela cell control microarray, as well as the example microarray data provided by Affymetrix, indicating that these are biologically meaningful features (e.g. negative control probes) and not technical failures e.g. related to imaging. Finally, randomly distributed highly intense white patches are evident in some but not all of the blank (2,3) and SHIME_LGG (2) samples. The biological meaning of these observations is questionable since the patches are not replicated which is suggestive of technical artifacts. Similar large, spatially contiguous clusters of highly intense signals or so-called blob-like defects (mostly oval shaped/round/arcs) have been described to occur in 10-20% of Affymetrix ST microarrays and are likely caused by bubbles formed during array manufacturing [10, 6, 11]. Such blobs render the transcriptional information in the affected area useless and violate the distributional assumptions made by normalization methods (e.g. quantile normalization assumes that distributions of signals in all microarrays are similar). Moreover normalizing the data in this case will contaminate the signal in all of the arrays leading to subsequent spurious findings [10]. Rather than repeating assays, which is both costly and time-consuming, we implemented

outlier detection prior to differential expression analysis. A couple of dedicated specialized tools have been developed to identify and remove blob-like artifacts (Harshlight, MBR, caCORRECT) [10, 6]. While those tools were very important with old array designs where all probes contributing to a single gene expression value were arranged in contiguous regions, the newer chip layouts with randomly distributed probes do not require a blob-removal pre-processing step. The standard techniques such as RMA can adequately detect single outlier probes within a probeset through model-fitting approaches without information about spatial location. The performance of RMA with respect to artifact removal was evaluated below.

Note that even in the example data provided by Affymetrix, artifacts were apparent (although rare) and the intensities largely differed between replicate arrays (Figure 4.2-4.3).

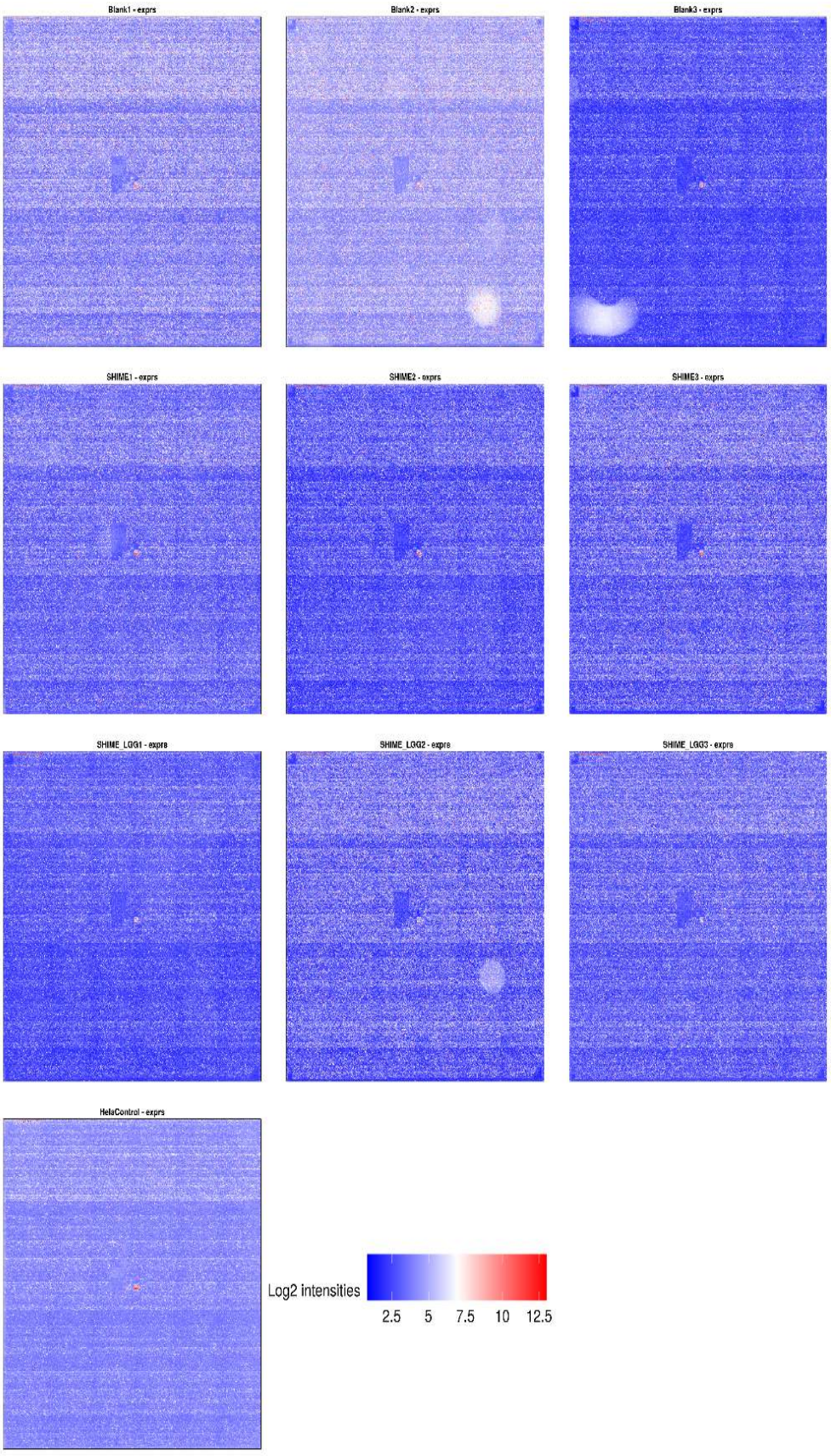

Figure 4.1: Pseudo-images of the log2 scaled probe intensities. Intensities differ across replicates and artifacts are present on a few chips. Normalization and outlier detection will be applied to deal with these issues.

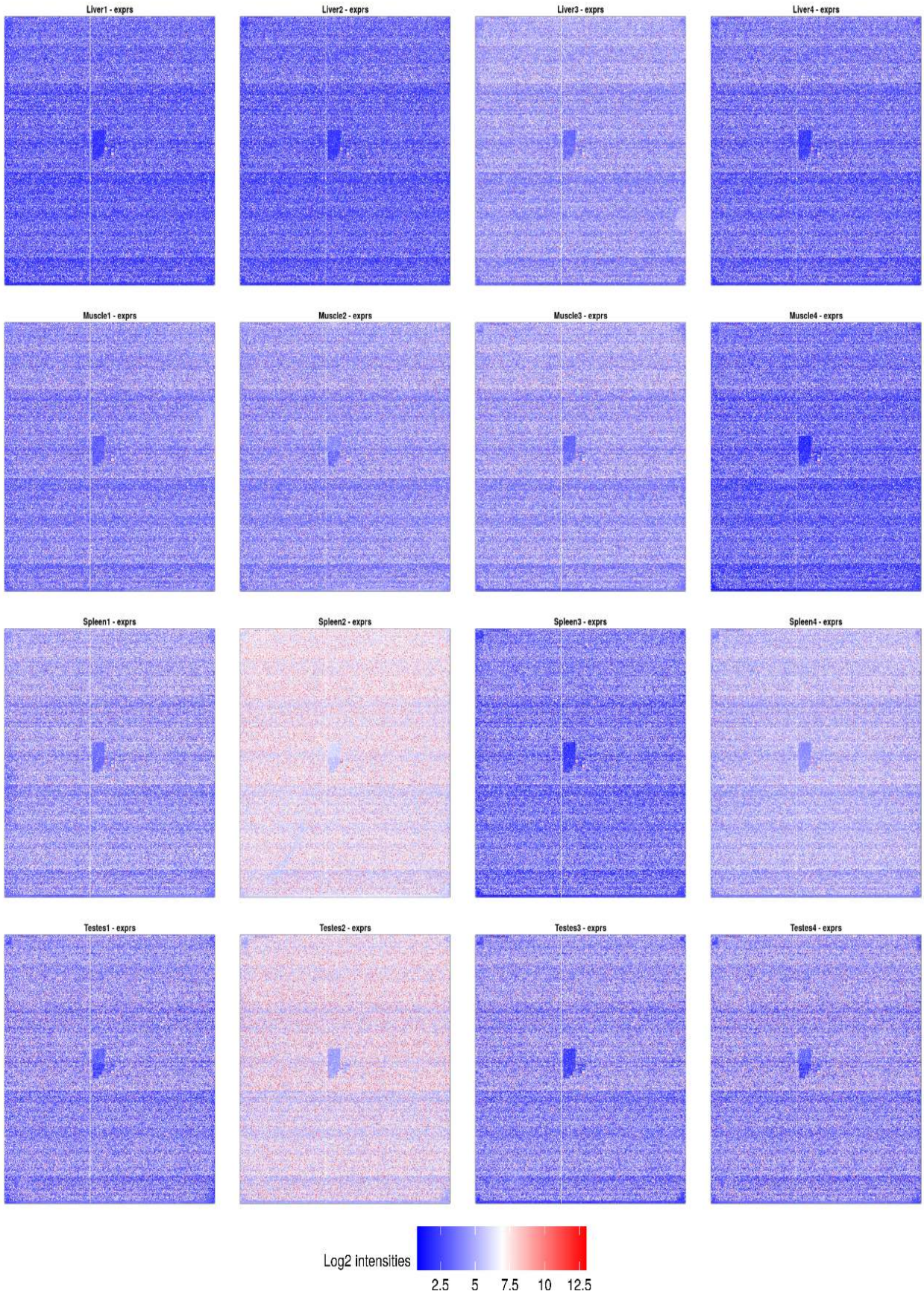

Figure 4.2: Pseudo-images of the log2 scaled probe intensities from Affymetrix tissue example data. Intensities differ across replicates and artifacts are observed on a few chips.

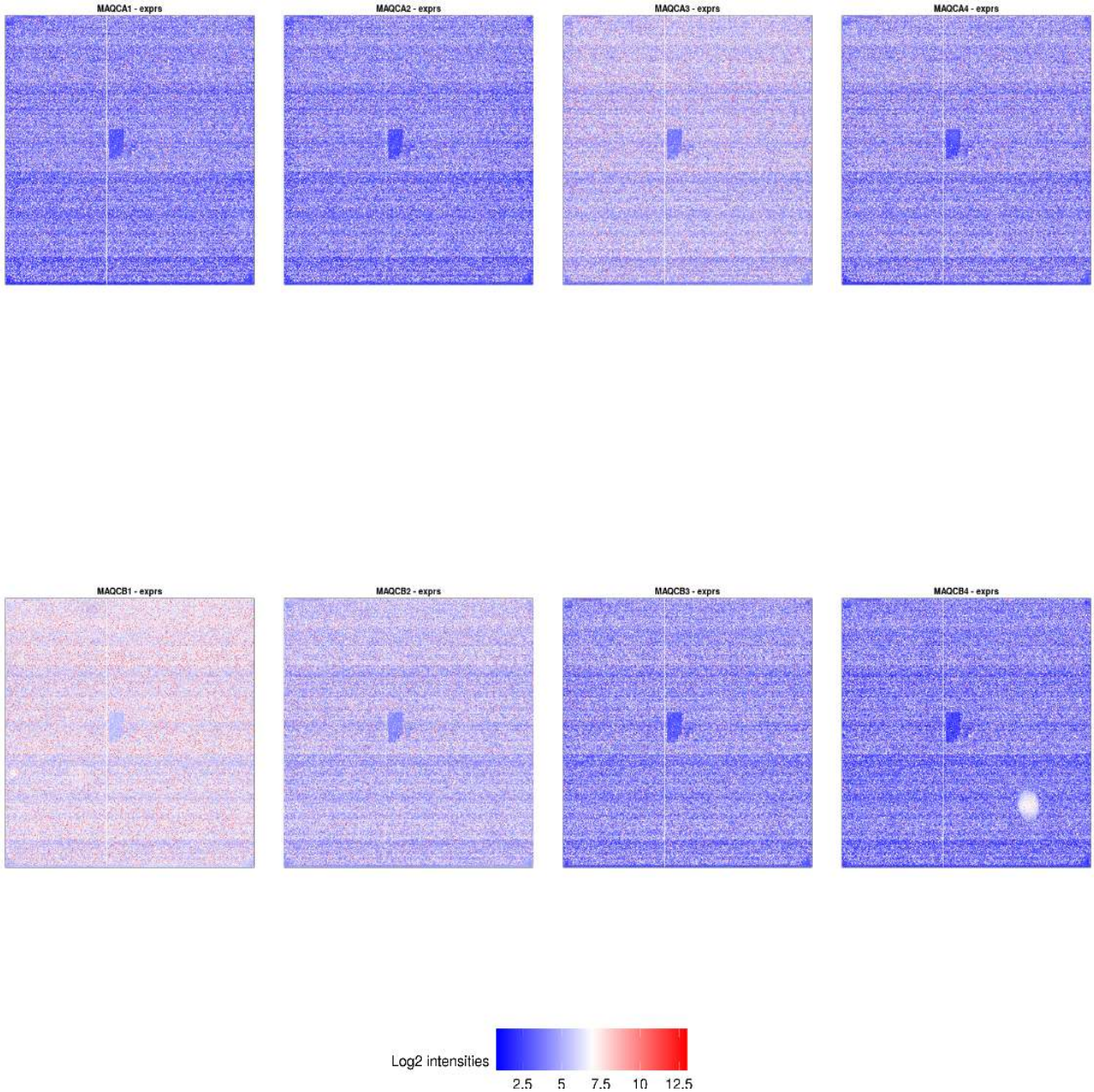

Figure 4.3: Pseudo-images of the log2 scaled probe intensities from Affymetrix MAQC example data. Intensities differ across replicates and artifacts are observed on a few chips.

4.2 Distribution of log2 probe intensities

4.2.1 Distribution of log2 'perfect match' probe intensities

The distribution of log2 intensities for all 'perfect match' probes annotated but not summarized at the transcript level (target = core) was compared between different arrays. While no outlier arrays are present, differences in shape and center of the distributions of the 'perfect match' intensities were observed (Figure 4.4). This can also be inferred from a box plot visualization, which displays differences in range and median intensities indicating that normalization of the data is required (Figure 4.5, see 5.2). The lower intensities in the Blank3 and SHIME_LGG1 control are in line with the dim appearance of those arrays (Figure 4.1,4.5). Additionally, the SHIME1 and Hela control probe log2 intensity distributions are at the lower end of the spectrum. This could point at lower cDNA concentrations loaded onto these microarrays, however total input RNA and the cDNA yield were quantified and standardized. Moreover, RNA degradation can be ruled out as RNA integrity and purity were verified based on the A260/280 ratio determined by UV Vis spectroscopy. Alternatively, a less efficient labeling, hybridization or more thorough washing could lead to variability in signal intensities. Control probes will be used to address these possibilities.

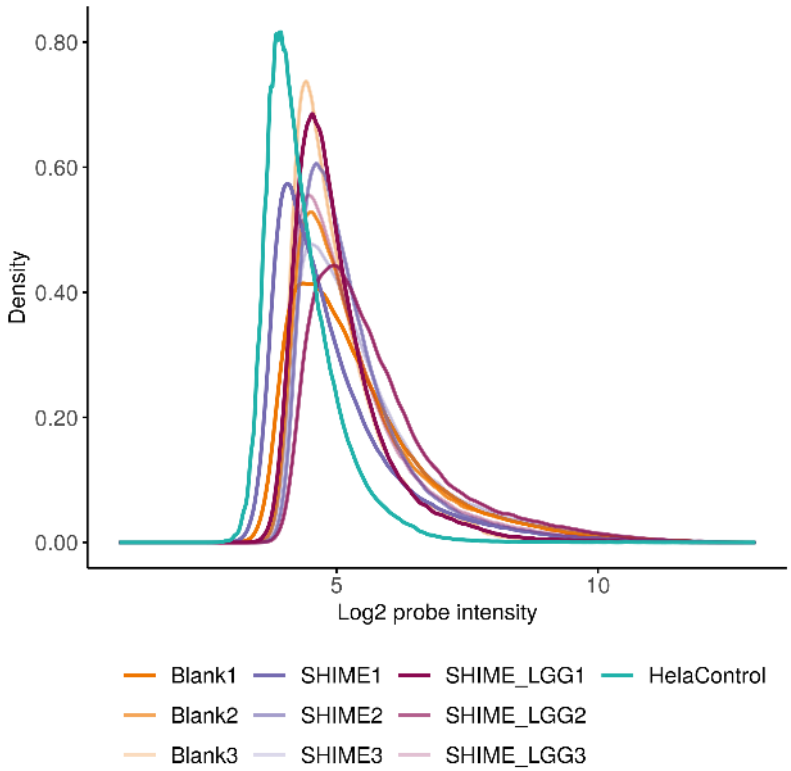

Figure 4.4: Density plot of the log2 intensities of all 'perfect match' probes.

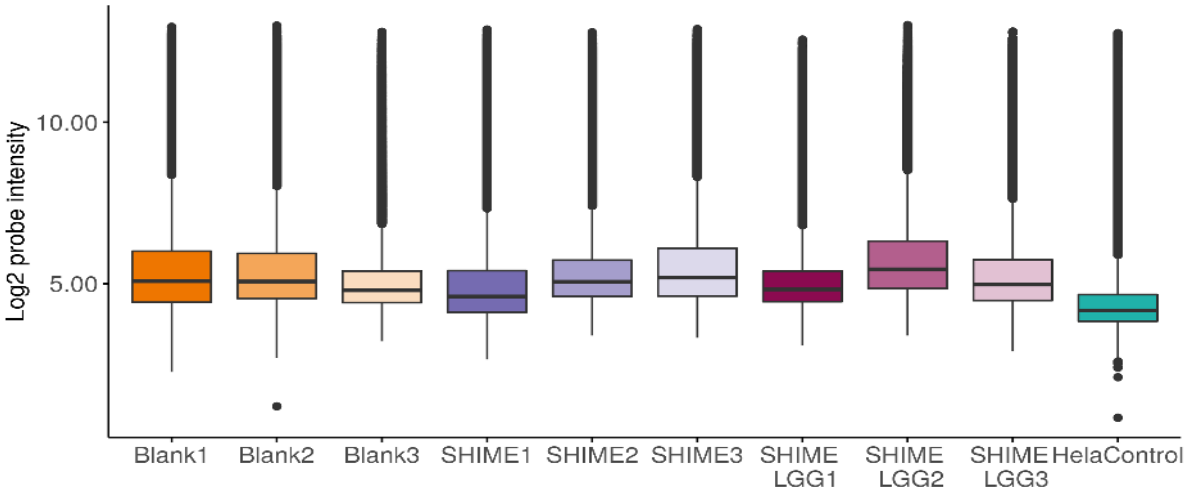

Figure 4.5: Box plots of the log2 intensities of all 'perfect match' probes.

4.2.2 Distribution of log2 background probe intensities

Histograms and boxplots can be plotted for the background probes separately (Figure 4.6,4.7). The background level is rather similar across all chips, suggesting that there were no major experimental anomalies. In line with the overall probe intensities, SHIME1 and the Hela cell control display lower background intensities. Discrepancies in background intensity may be due to differences in cDNA concentrations/quality, labeling, hybridization or washing.

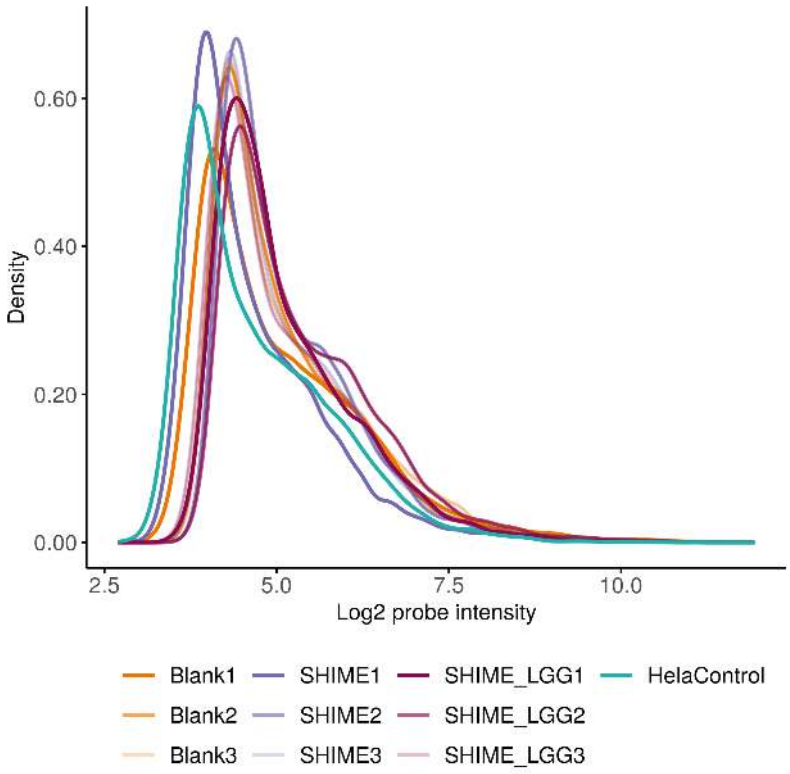

Figure 4.6: Density plot of the background probe log2 intensities

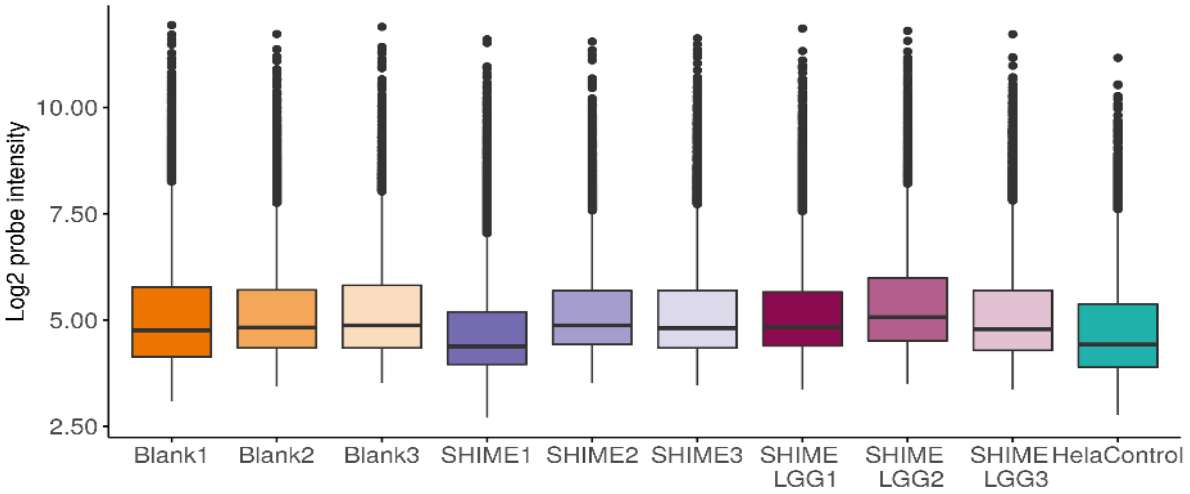

Figure 4.7: Box plots of the background probe log2 intensities

4.2.3 Distribution of log2 control probe intensities

Background probes (control→bgp→antigenomic) are only one type of control probesets included in the Affymetrix Human Gene 2.1 ST Array Strip, which additionally contains external positive controls (control→affx, control→affx→bac_spike, control→affx→polya_spike, control→affx→ercc), internal positive controls (normgene→exon), negative controls (normgene→intron) and Reporter+Rescue controls.

Table 4.1: Overview of control probesets on the Affymetrix Human Gene 2.1 ST Array Strip

Probe type	**Number of probesets**	**Number of probes**	**Probe type information**
main	343871	912340	Probesets which are part of the main design
control→affx	18	360	Standard Affymetrix spike control probeset (ie bacterial and polyA spikes)
control→affx→bac spike	18	198	Probesets which hybridize to pre-labeled Affymetrix bacterial spike controls (BioB, BioC, BioD, and Cre). This category is useful in identifying problems with the hybridization, washing, scanning and/or chip.
control→affx→polya spike	39	563	Probesets which hybridize to polyadenylated RNA spikes (Lys, Phe, Thr, and Dap). This category is useful in identifying problems with the target preparation (amplification/labeling).
normgene→exon	1626	3043	Probesets against exon regions of a set of housekeeping genes acting as positive controls. This category is useful in identifying problems with sample extraction.
normgene→intron	3575	13131	Probesets against intron regions of a set of housekeeping genes acting as negative controls.
control→bgp→antigenomic	23	16943	Antigenomic background probesets. Important to determine signal to noise ratio. Used for background correction.
control→affx→ercc	92	2322	ERCC RNA Spike-In Control Mixes defined by the External RNA Controls Consortium = pre-formulated sets of 92 polyadenylated transcripts from the ERCC plasmid reference library of NIST-certified DNA plasmids that are designed to produce a set of transcripts 250–2000 nt in length that mimic natural eukaryotic mRNAs. The transcripts are traceable through the manufacturing process to the NIST plasmid reference material.
Reporter+Rescue	3597	76188	Probesets against mRNA sequences which did not align (or poorly aligned) to the genome.

The probetypes available on the Affymetrix Human Gene 2.1 ST Array Strip, their purpose and the number of probesets of every type are shown in tables 4.1 and 4.2.

An overview of the unnormalized raw log2 intensities indicates that the antigenomic background probes (listed in the .bgp file) that are not matching against the human genome tend to have similar median log2 intensities compared to the main and substantially lower median log2 intensities compared to the positive control probesets (Figure 4.8-4.10).

Similarly, the signals resulting from probesets targeting putative intron regions of a set of putative housekeeping genes are in the same range as the main probe intensities. Intron regions act as a negative control and should have a very low signal `https://tools.thermofisher.com/content/sfs/brochures/exon_gene_arrays_qa_whitepaper.pdf`. However, some variation in this negative control may occur. While intronic regions in theory should be spliced out in the mRNA, some of these putative intronic regions may be transcribed and retained. Besides, some of the identified genes corresponding to those intronic regions may not be constitutively expressed (Figure 4.8-4.10).

ERCC RNA Spike-In positive controls defined by the External RNA Controls Consortium were not added during the sample preparation, explaining the low intensity of ERCC probesets (Figure 4.8-4.10). The ERCC RNA Spike-In Control Mixes are pre-formulated sets of 92 polyadenylated transcripts from the ERCC plasmid reference library of NIST-certified DNA plasmids that are designed to produce a set of transcripts (250–2000 nucleotides in length) that mimic natural eukaryotic mRNAs [4].

Finally, reporter and rescue probesets targetting mRNA transcripts which either did not align to the genome, or aligned poorly, behave similar to the other negative controls (Figure 4.8-4.10).

The fact that background/control probes can have a higher/similar median intensity compared to main 'perfect match' probes is not problematic because 'perfect match' probes behave differently from background probes. For example, there may be no real target for many of the main probes in the samples. Thus the median intensity of the 'perfect match' probes may indeed be very low (at or near background). In contrast, the background intensity may be skewed towards higher values due to some GC rich probes present as a control for the high GC count probes (see 4.3).

Table 4.2: Number of control probes and probesets on the Affymetrix Human Gene 2.1 ST Array Strip

	Probe type	Probe subtype	Number of probesets	Number of probes
1	control→affx→bac spike	17127689	1	11
2	control→affx→bac spike	17127693	1	11
3	control→affx→bac spike	17127697	1	11
4	control→affx→bac spike	17127701	1	11
5	control→affx→bac spike	17127705	1	11
6	control→affx→bac spike	17127709	1	11
7	control→affx→bac spike	17127713	1	11
8	control→affx→bac spike	17127717	1	11
9	control→affx→bac spike	17127721	1	11
10	control→affx→bac spike	AFFX-r2-Ec-bioB-3 at	1	11
11	control→affx→bac spike	AFFX-r2-Ec-bioB-5 at	1	11
12	control→affx→bac spike	AFFX-r2-Ec-bioB-M at	1	11
13	control→affx→bac spike	AFFX-r2-Ec-bioC-3 at	1	11
14	control→affx→bac spike	AFFX-r2-Ec-bioC-5 at	1	11
15	control→affx→bac spike	AFFX-r2-Ec-bioD-3 at	1	11
16	control→affx→bac spike	AFFX-r2-Ec-bioD-5 at	1	11
17	control→affx→bac spike	AFFX-r2-P1-cre-3 at	1	11
18	control→affx→bac spike	AFFX-r2-P1-cre-5 at	1	11
19	control→affx→polya spike	17127639	1	11
20	control→affx→polya spike	17127643	1	11
21	control→affx→polya spike	17127647	1	11
22	control→affx→polya spike	17127651	1	10
23	control→affx→polya spike	17127655	1	11
24	control→affx→polya spike	17127659	1	11
25	control→affx→polya spike	17127663	1	11
26	control→affx→polya spike	17127667	1	11
27	control→affx→polya spike	17127671	1	11
28	control→affx→polya spike	17127675	1	11
29	control→affx→polya spike	17127679	1	11
30	control→affx→polya spike	17127683	1	11
31	control→affx→polya spike	AFFX-DapX-3 st	1	20
32	control→affx→polya spike	AFFX-DapX-5 st	1	20
33	control→affx→polya spike	AFFX-DapX-M st	1	20
34	control→affx→polya spike	AFFX-LysX-3 st	1	20
35	control→affx→polya spike	AFFX-LysX-5 st	1	20
36	control→affx→polya spike	AFFX-LysX-M st	1	20
37	control→affx→polya spike	AFFX-PheX-3 st	1	20
38	control→affx→polya spike	AFFX-PheX-5 st	1	20
39	control→affx→polya spike	AFFX-PheX-M st	1	20
40	control→affx→polya spike	AFFX-r2-Bs-dap-3 st	1	11
41	control→affx→polya spike	AFFX-r2-Bs-dap-5 st	1	11
42	control→affx→polya spike	AFFX-r2-Bs-dap-M st	1	11
43	control→affx→polya spike	AFFX-r2-Bs-lys-3 st	1	11
44	control→affx→polya spike	AFFX-r2-Bs-lys-5 st	1	11
45	control→affx→polya spike	AFFX-r2-Bs-lys-M st	1	11
46	control→affx→polya spike	AFFX-r2-Bs-phe-3 st	1	11
47	control→affx→polya spike	AFFX-r2-Bs-phe-5 st	1	11
48	control→affx→polya spike	AFFX-r2-Bs-phe-M st	1	11
49	control→affx→polya spike	AFFX-r2-Bs-thr-3 s st	1	11
50	control→affx→polya spike	AFFX-r2-Bs-thr-5 s st	1	11
51	control→affx→polya spike	AFFX-r2-Bs-thr-M s st	1	11
52	control→affx→polya spike	AFFX-ThrX-3 st	1	20
53	control→affx→polya spike	AFFX-ThrX-5 st	1	20
54	control→affx→polya spike	AFFX-ThrX-M st	1	20
55	control→affx→polya spike	AFFX-TrpnX-3 st	1	20
56	control→affx→polya spike	AFFX-TrpnX-5 st	1	20
57	control→affx→polya spike	AFFX-TrpnX-M st	1	20

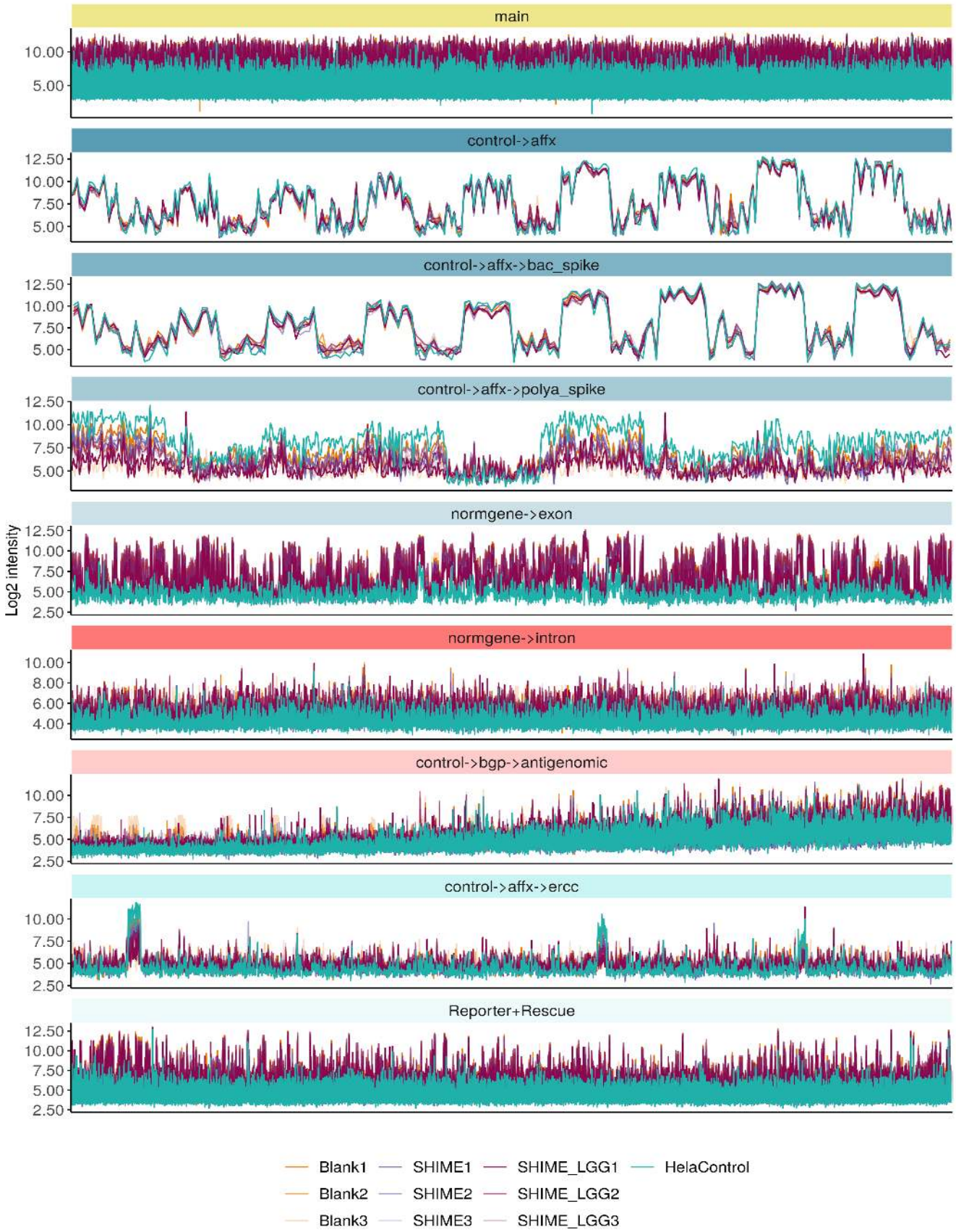

Figure 4.8: Raw log2 intensities of all probes within the probesets corresponding to the spike-in positive controls (control→affx, control→affx→bac_spike, control→affx→polya_spike) and negative controls (control→affx→ercc, control→bgp→antigenomic, normgene→intron).

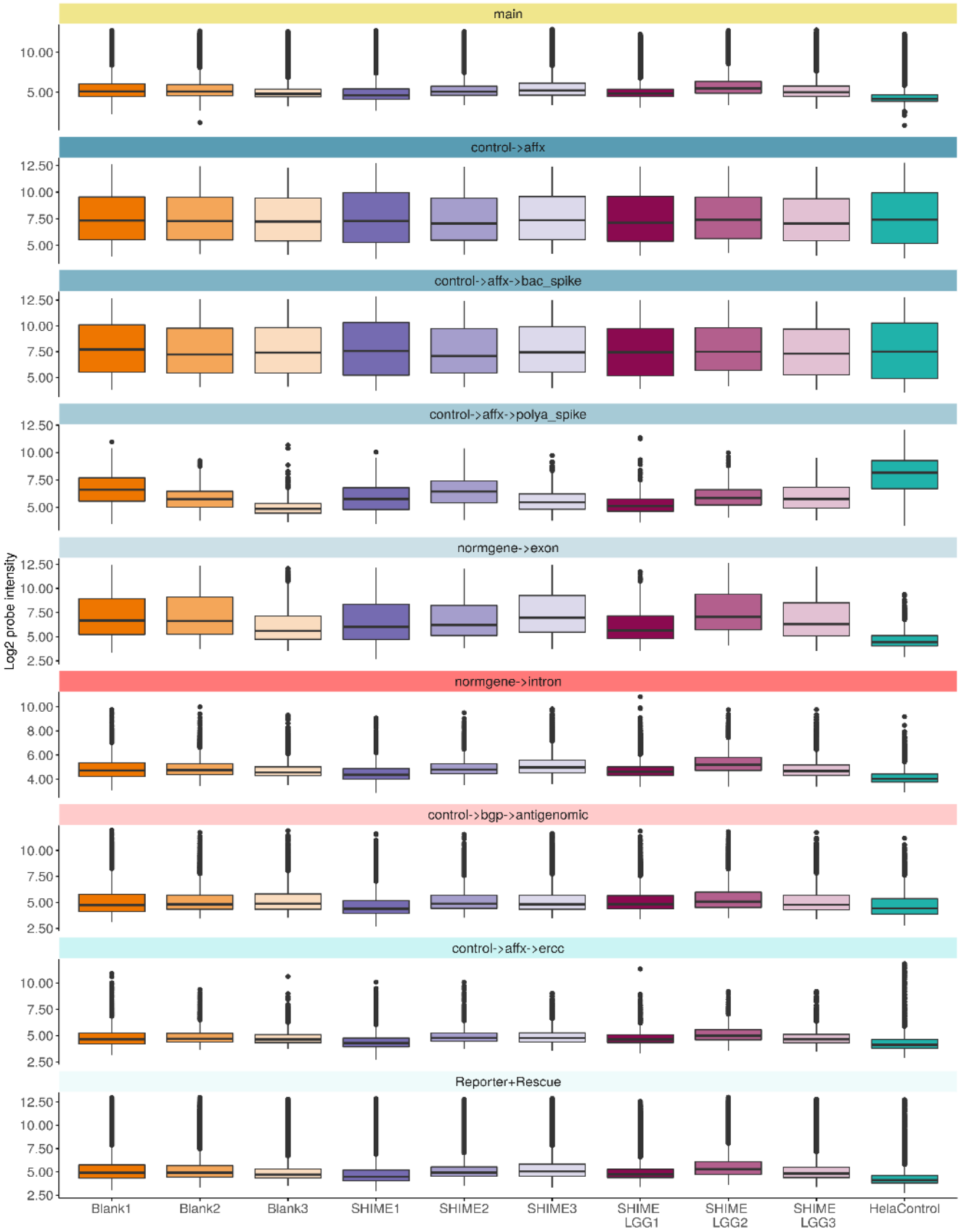

Figure 4.9: Box plots of the raw log2 intensities of all control probes within the probesets corresponding to the spike-in positive controls (control→affx, control→affx→bac_spike, control→affx→polya_spike) and negative controls (control→affx→ercc, control→bgp→antigenomic,normgene→intron) faceted by the controls.

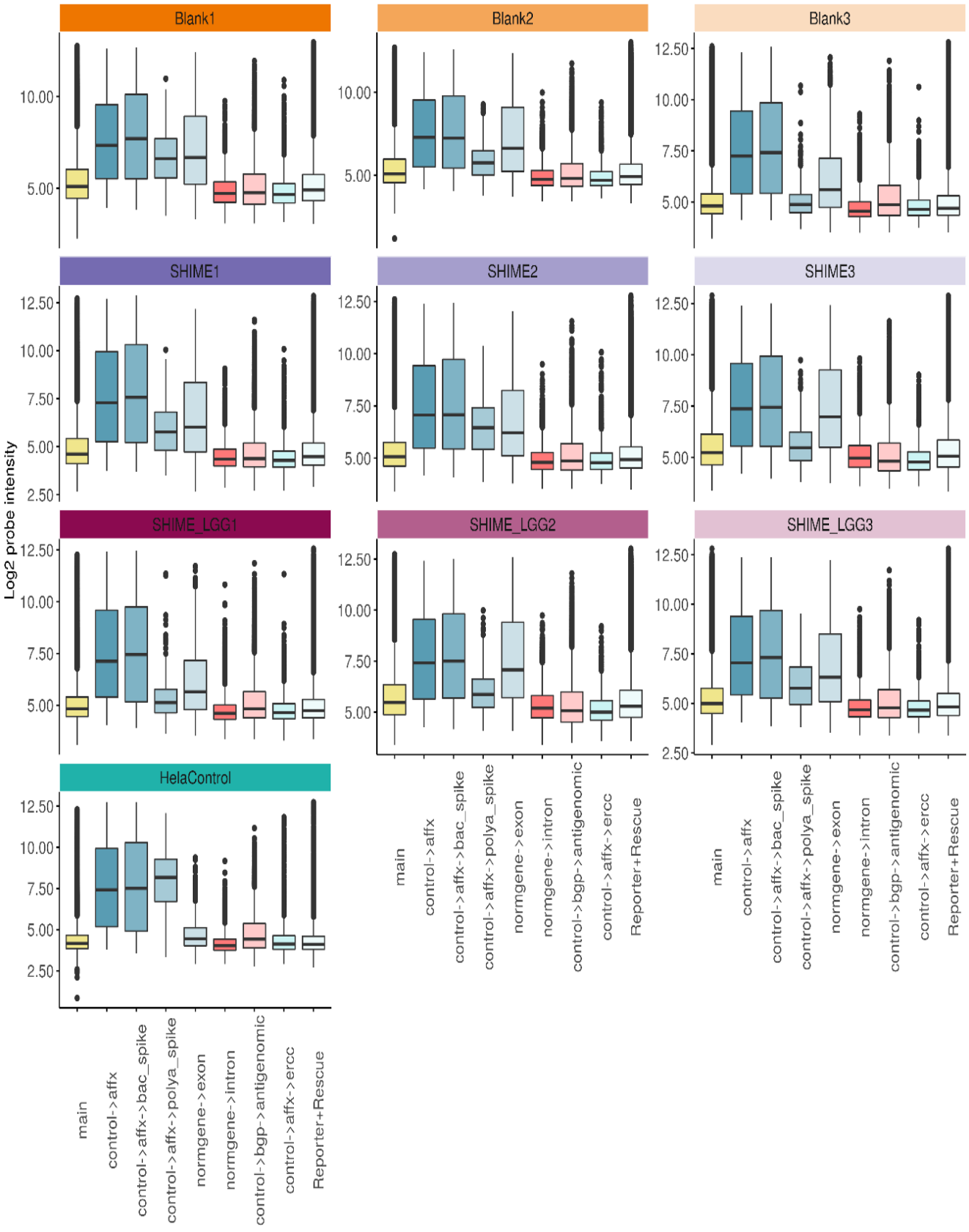

Figure 4.10: Box plots of the raw log2 intensities of all control probes within the probesets corresponding to the spike-in positive controls (control→affx, control→affx→bac_spike, control→affx→polya_spike) and negative controls (control→affx→ercc, control→bgp→antigenomic, normgene→intron) faceted by the samples.

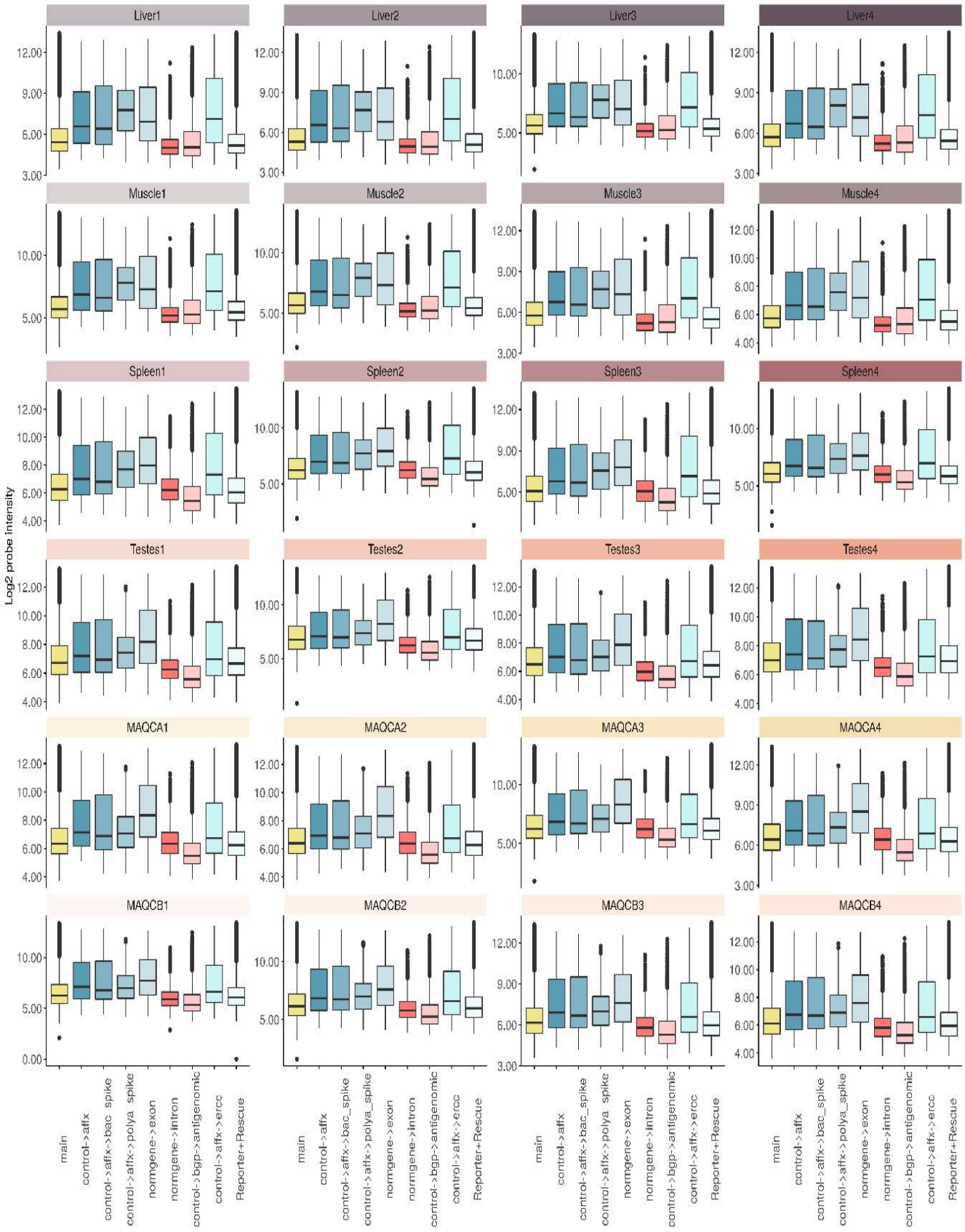

Figure 4.11: Box plots of the raw log2 intensities of all control probes within the probesets corresponding to the spike-in positive controls (control→affx, control→affx→bac_spike, control→affx→polya_spike) and negative controls (control→affx→ercc, control→bgp→antigenomic, normgene→intron) faceted by the controls in the example data.

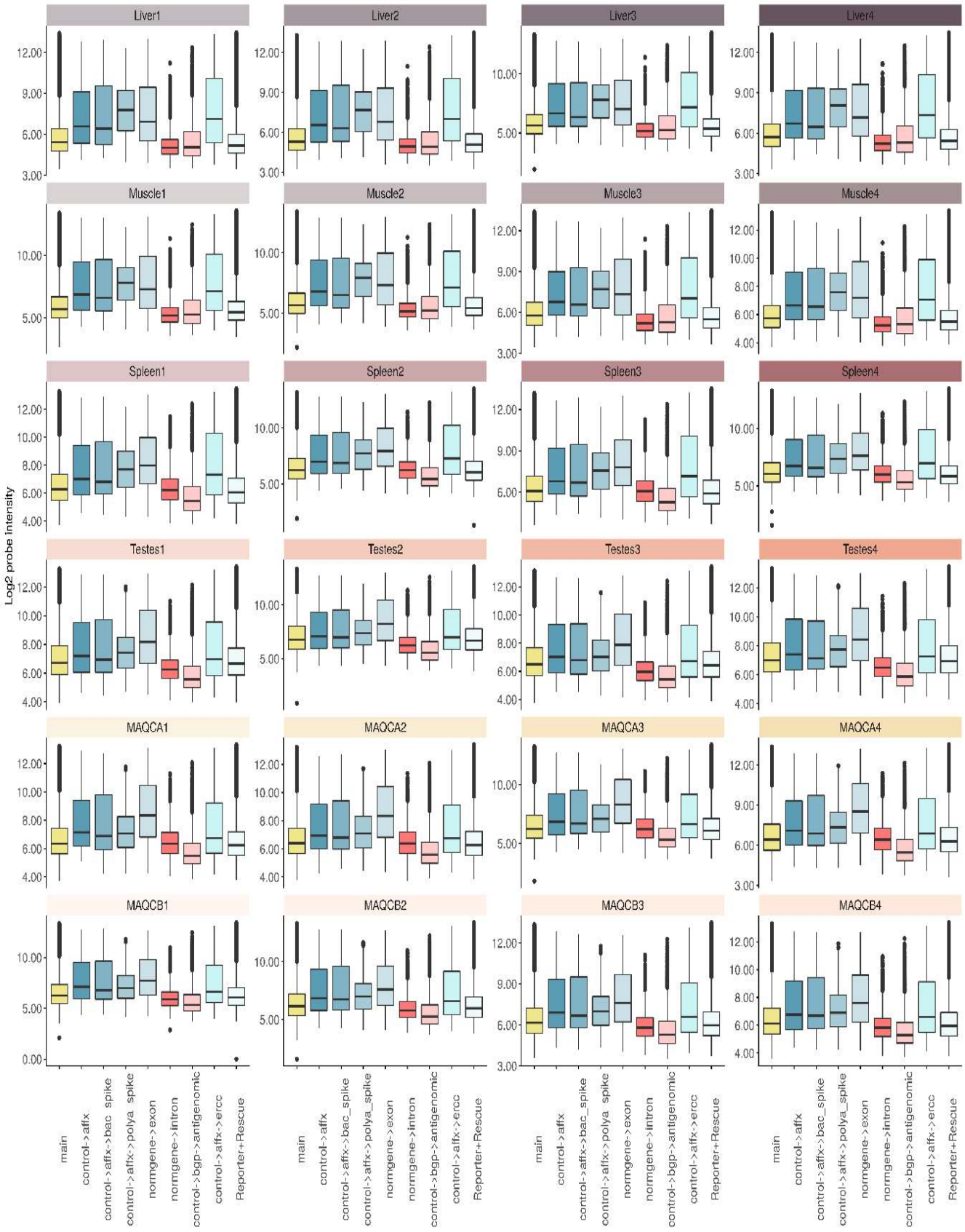

Figure 4.12: Box plots of the raw log2 intensities of all control probes within the probesets corresponding to the spike-in positive controls (control→affx, control→affx→bac_spike, control→affx→polya_spike) and negative controls (control→affx→ercc, control→bgp→antigenomic, normgene→intron) faceted by the example samples.

4.2.3.1 Distribution of log2 bac and polyA spike probe intensities

Each eukaryotic GeneChip array also contains probesets for several *E. coli* (BioB, BioC, BioD), bacteriophage P1(Cre) and modified *B. subtilis* genes (trpn, lys, phe, thr, and dap) that are absent in eukaryotic samples. These bac_spike (18 probesets) and polya_spike (39 probesets) probesets act as external positive controls through the addition of complementary targets: the affymetrix bacterial and polyA spike-in controls. The pre-labeled bacterial/phage spike-in controls (BioB, BioC, BioD, and Cre) are added during the hybridization step in increasing concentrations of 1.5, 5, 25, and 100 pM respectively. They act as an internal hybridization control to confirm the low-end assay sensitivity and evaluate the dose-response across the dilution range. They can be used to identify potential problems with the hybridization, washing, scanning or the chip itself. Issues arising during the target preparation phase can be detected by means of the *in vitro* synthesized polyadenylated RNA spikes (Lys, Phe, Thr, and Dap) that are added to the total RNA prior to the *in vitro* first-Strand cDNA Synthesis step. The polyadenylated transcripts for the *B. subtilis* genes are premixed at staggered concentrations (to obtain final copy number ratio of 1:100000,1:50000,1:25000,1:6667). Note that although Trpn probesets are present, the corresponding polyA transcript was not added and hence this acts as a negative control.

The replicate arrays show high reproducibility for the bacterial spike-in controls indicating that the hybridization, fluidics methods as well as the chips and scanning and griding procedure were fine. This is confirmed by the bacterial spikes displaying the expected rank order (BioB<BioC<BioD<Cre) and an approximate linear relationship between the log2 intensities and bac spike concentrations (Figure 4.13,4.14). A larger between-array variability was observed in the polyA spike. This is in line with the expectations since the hybridization controls are added in the last experimental stage, whereas the polyA spikes are subjected to additional sources of variation in the sample amplification and labeling steps [4]. Remarkably, the spike intensities peaked in the Hela cell control array, which exhibited smaller median 'perfect match' and background probe intensities compared to the samples. This could be due to differences in the quality and ratio of total RNA vs polyA spikes. Interestingly, the blank3 and SHIME_LGG1 sample generated lower polyA spike signals, which is in correspondence with the lower overall intensities, suggesting a lower efficiency during the target preparation (Figure 4.15). Despite this variability, the polyA spikes display the expected rank order (trpn<lys<phe<thr<dap) indicating no major aberrancies during the preparation of the sample RNA and spikes (Figure 4.15,4.16). The poor fit of a linear regression model (min r^2 value of 0.15) suggests that background correction is required to improve the linear signal intensity-RNA concentration relationship. While our data behaves similar to the example data with respect to the bacterial spike-in controls, the example data contains abnormal high lys signals (Figures 4.17-4.20). This might relate to changes in the experimental workflow since the acquisition of the example data. Indeed earlier resources report a different spike-in order (lys>phe>dap>thr) [4].

As can be observed from the large interquartile ranges (Figure 4.10), metrics in these categories of exogenous positive controls have more variability than other categories due to the limited number of spikes and probesets. Thus they should only be used to troubleshoot specific problems whereas the Hela cell control and replicates of the other arrays were used to assess overall quality.

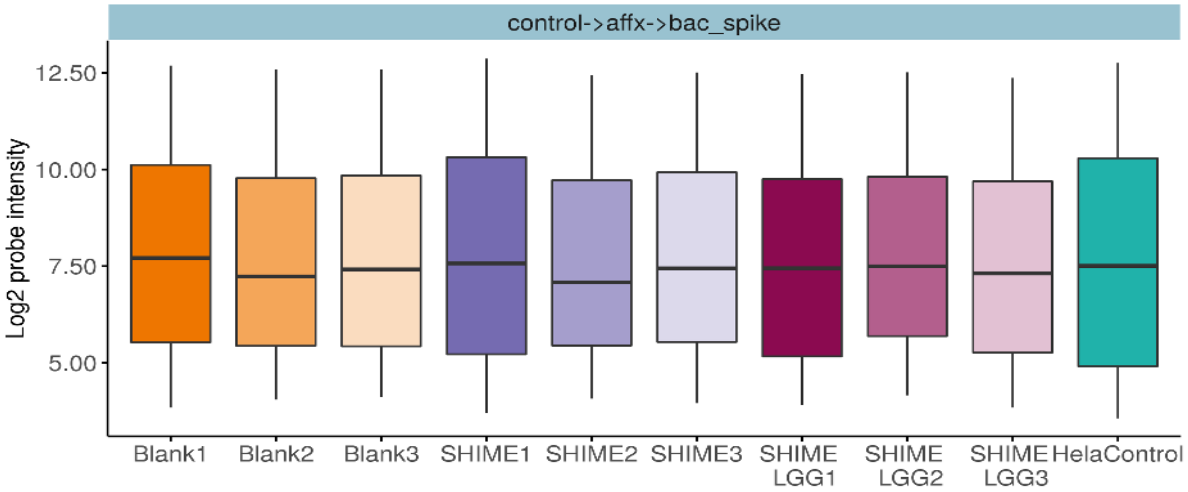

Figure 4.13: Raw log2 intensities of all probes within the probesets corresponding to the bacterial Affymetrix spike-in faceted by the samples.

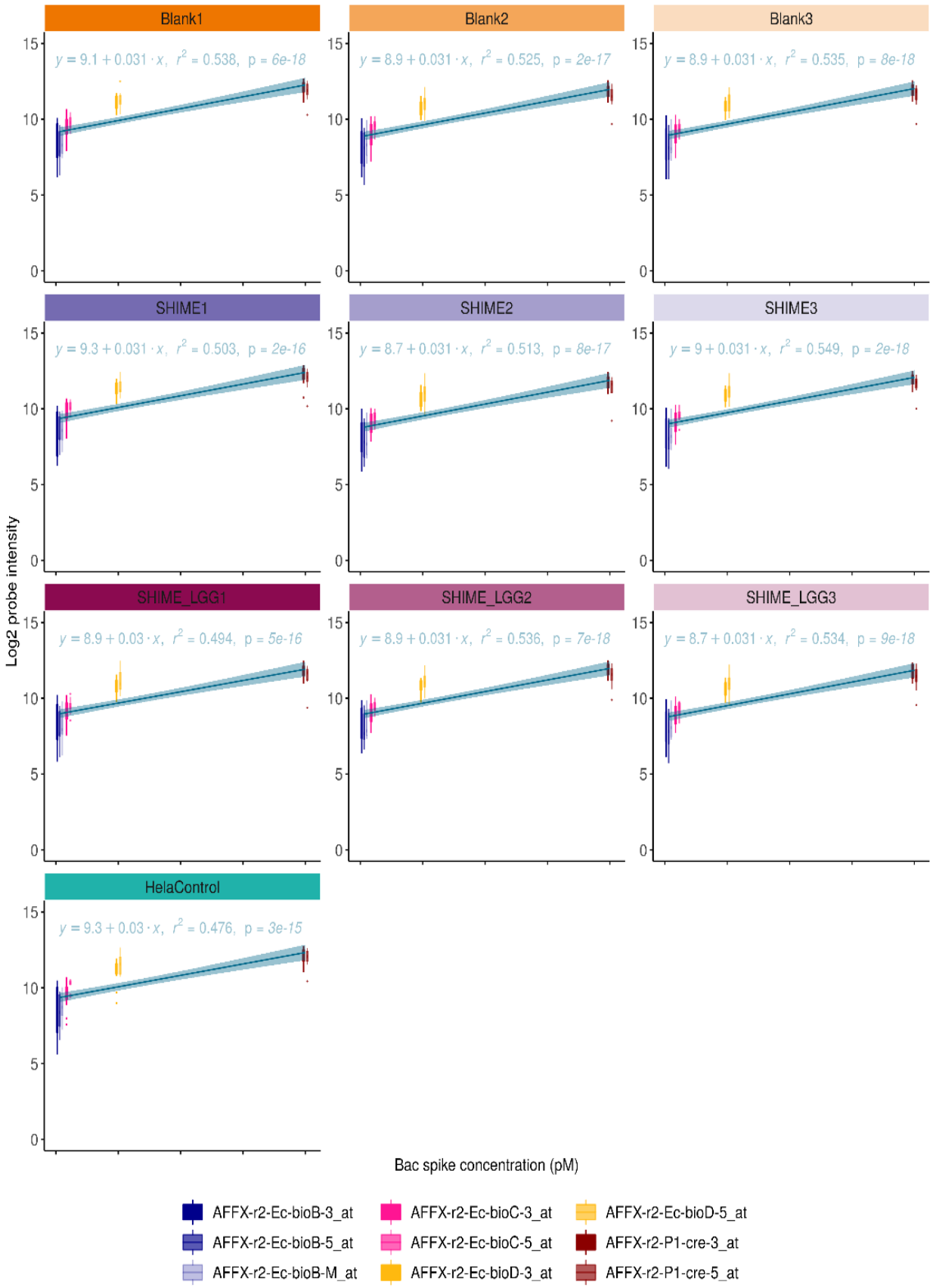

Figure 4.14: Raw log2 intensities of the BioB, BioC, BioD and Cre bacterial Affymetrix spike-in probes faceted by the samples.

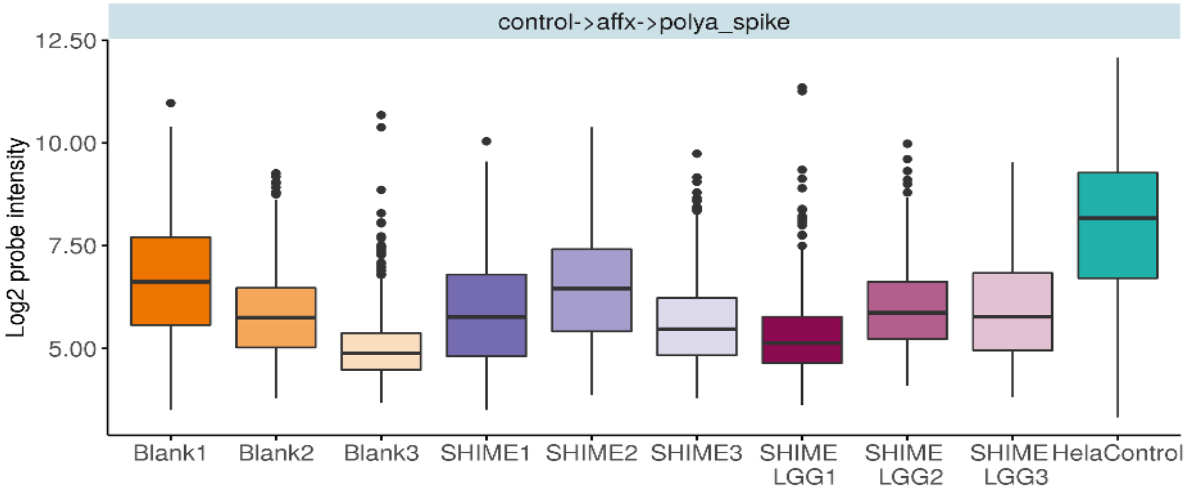

Figure 4.15: Raw log2 intensities of all probes within the probesets corresponding to the polyA Affymetrix spike-in faceted by the samples.

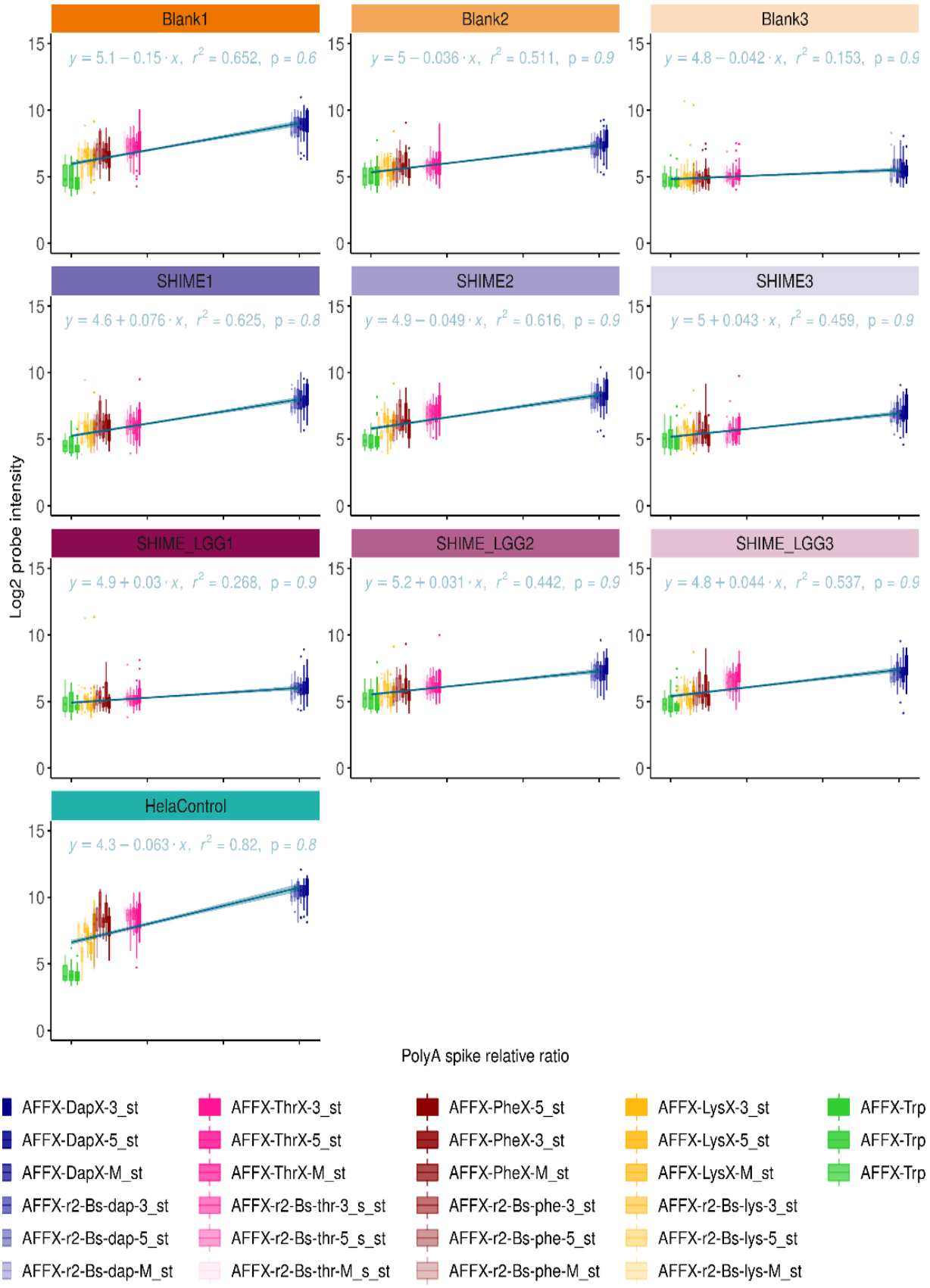

Figure 4.16: Raw log2 intensities of all probes within the probesets corresponding to the polyA Affymetrix spike-in faceted by the samples.

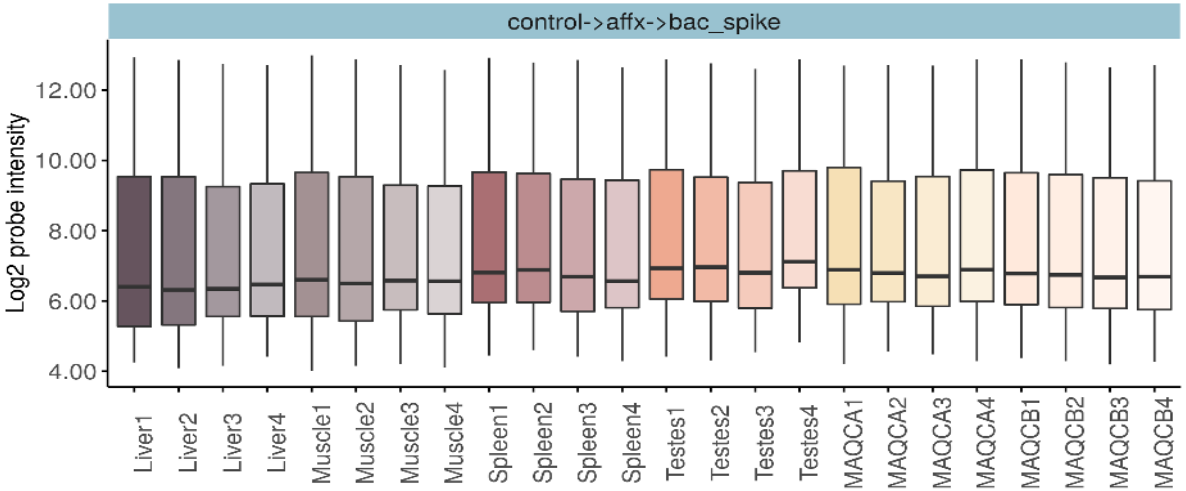

Figure 4.17: Raw log2 intensities of all probes within the probesets corresponding to the bacterial Affymetrix spike-in faceted by the samples in the example data.

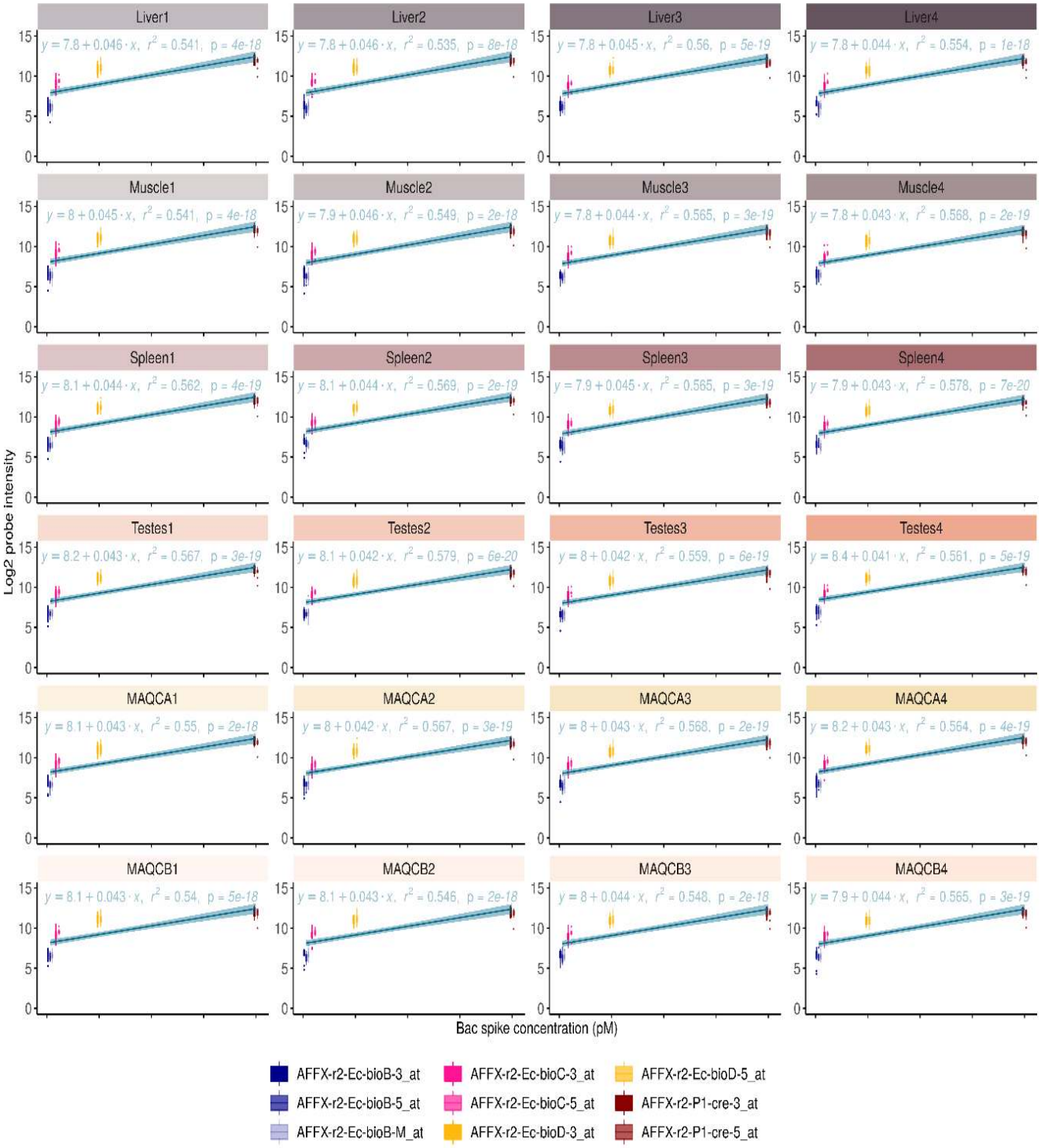

Figure 4.18: Raw log2 intensities of the BioB, BioC, BioD and Cre bacterial Affymetrix spike-in probes faceted by the samples in the example data.

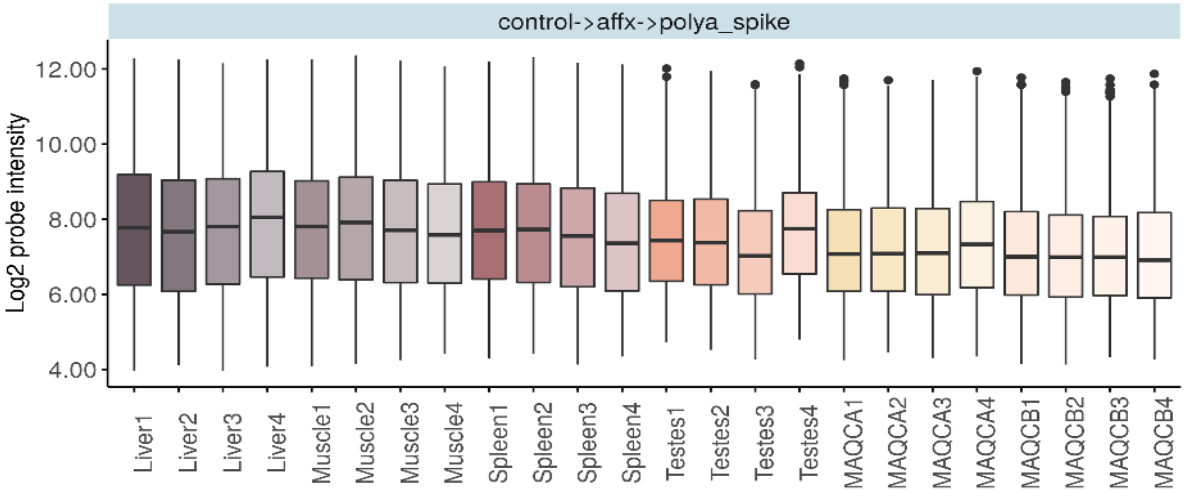

Figure 4.19: Raw log2 intensities of all probes within the probesets corresponding to the polyA Affymetrix spike-in faceted by the samples in the example data.

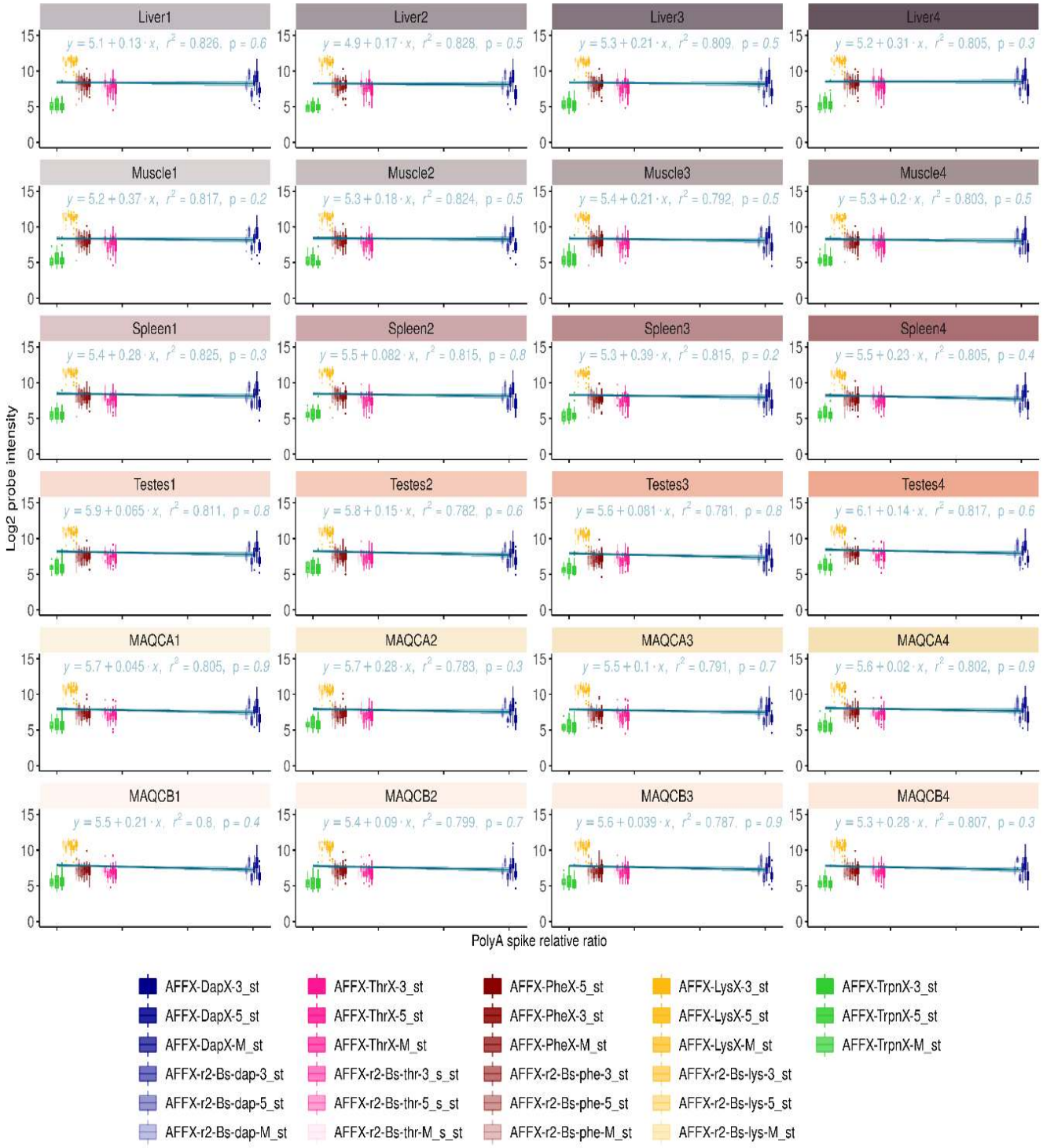

Figure 4.20: Raw log2 intensities of all probes within the probesets corresponding to the polyA Affymetrix spike-in faceted by the samples in the example data.

To conclude, a between-array comparison of all probe types reveals a similar pattern for all microarrays in this study. Moreover, the example microarrays portray similar trends with respect to the controls and main probes, indicating that our data is of good quality (Figure 4.11 and 4.12). Moreover, the positive and negative controls also display similar signal patterns across all arrays, with exception of the polyA spikes, which follow the expected dose-response trajectory in our samples and not the example data (Figure 4.9). This is likely due to recently implemented modifications to the experimental workflow.

4.2.3.2 Distribution of log2 housekeeping probe intensities

Besides the exogenous control probesets, it can be interesting to look at some housekeeping genes with expected constant expression levels. Some endogenous controls (exon regions of constitutively expressed genes) are included in the chip design. These 'normgene exons' behave similarly for all arrays and confirm the lower intensity values for the blank3,SHIME_LGG1 and Hela cell control (Figure 4.10). Additionally, the raw log2 intensities of probesets corresponding to some specific housekeeping genes (ACTB, GAPHD and HMBS) were explored. Median intensities slightly differed between arrays for all three genes and the blank3, SHIMELGG1 and Hela control array produced the smallest signals (in line with the overall trend). The variability in log2 intensities was similar for ACTB and GAPHD, but for HMBS the interquartile range was notably larger for the samples compared to the Hela cell control (Figure 4.21). Despite the differences in signal magnitudes, all arrays hybridizing with the triple coculture cell samples displayed very similar patterns of individual probe intensities for all three genes across the different conditions (blank, SHIME, SHIME + LGG). Deviations were most pronounced for HMBS and particularly the Hela cell control exhibited slightly deviating patterns. While the between array differences were negligible, large differences were noted between probes in the same probeset targeting one and the same gene, which is also reflected in the considerable interquartile ranges (Figure 4.22). Large interquartile ranges and between-array differences were also observed in the example datasets (Figure 4.23-4.24).

To conclude, the spike-in controls suggest that hybridization, scanning and target preparation were adequate. The fact that the sample arrays produce a similar (or even more intense) overall signal compared to the Hela cell positive control and example data demonstrates that the quality and nature of the original RNA samples was satisfactory. Differences in intensity between arrays and probes will be corrected through a normalization and summarization approach (see 5).

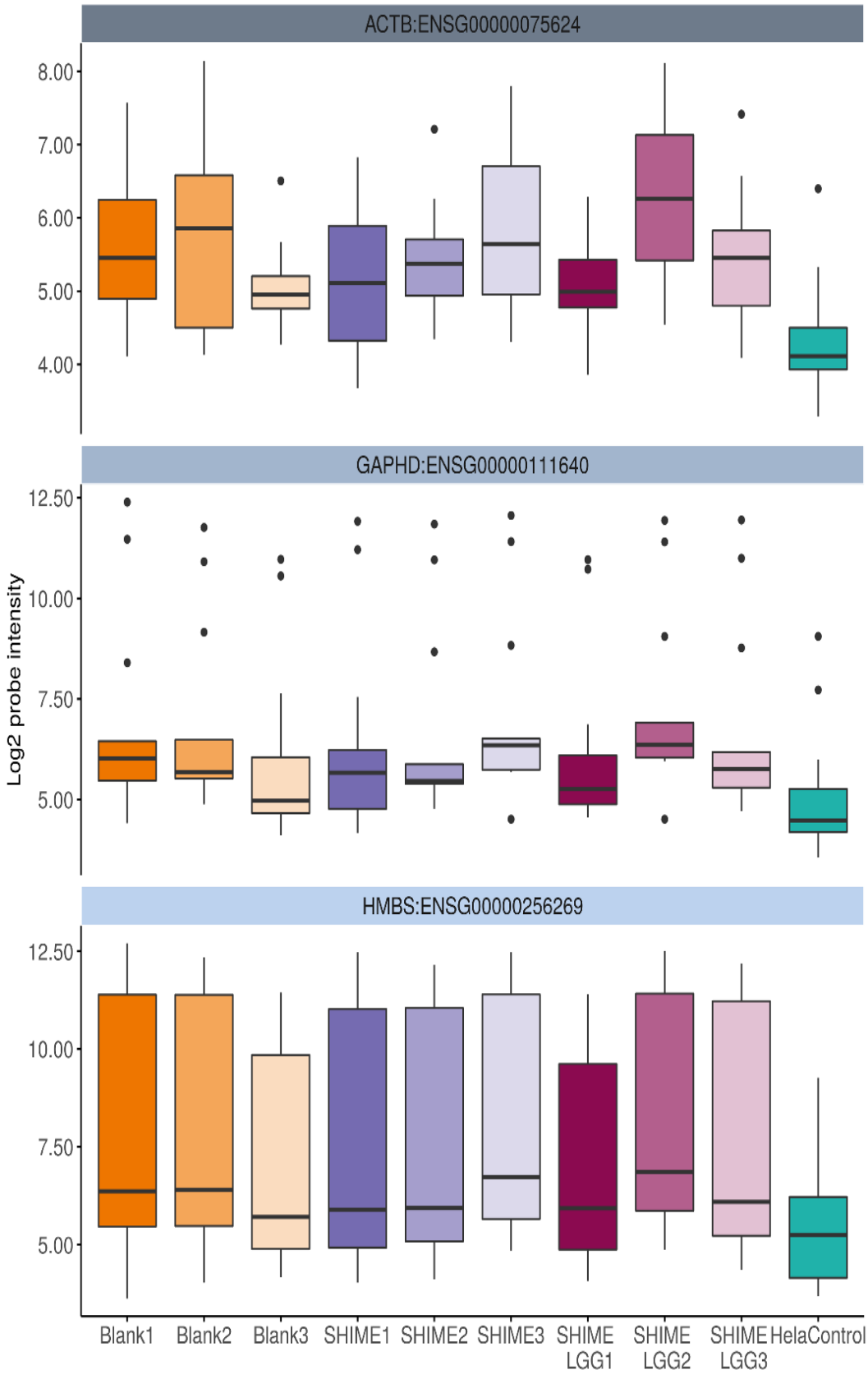

Figure 4.21: Box plots of the raw log2 intensities of all probes within the probesets corresponding to the housekeeping genes ACTB, GAPHD and HMBS.

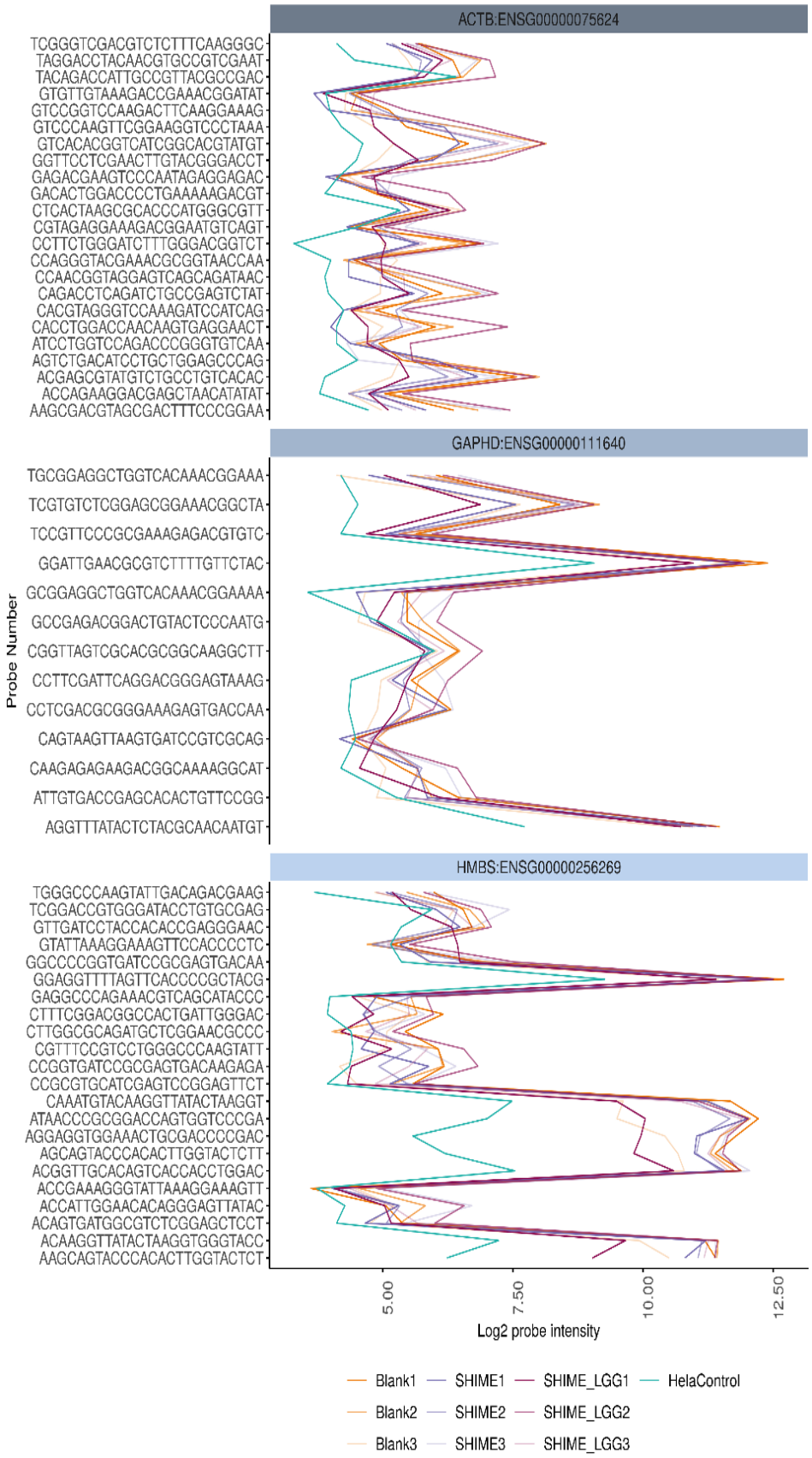

Figure 4.22: Raw log2 intensities of all probes within the probesets corresponding to the housekeeping genes ACTB, GAPHD and HMBS.

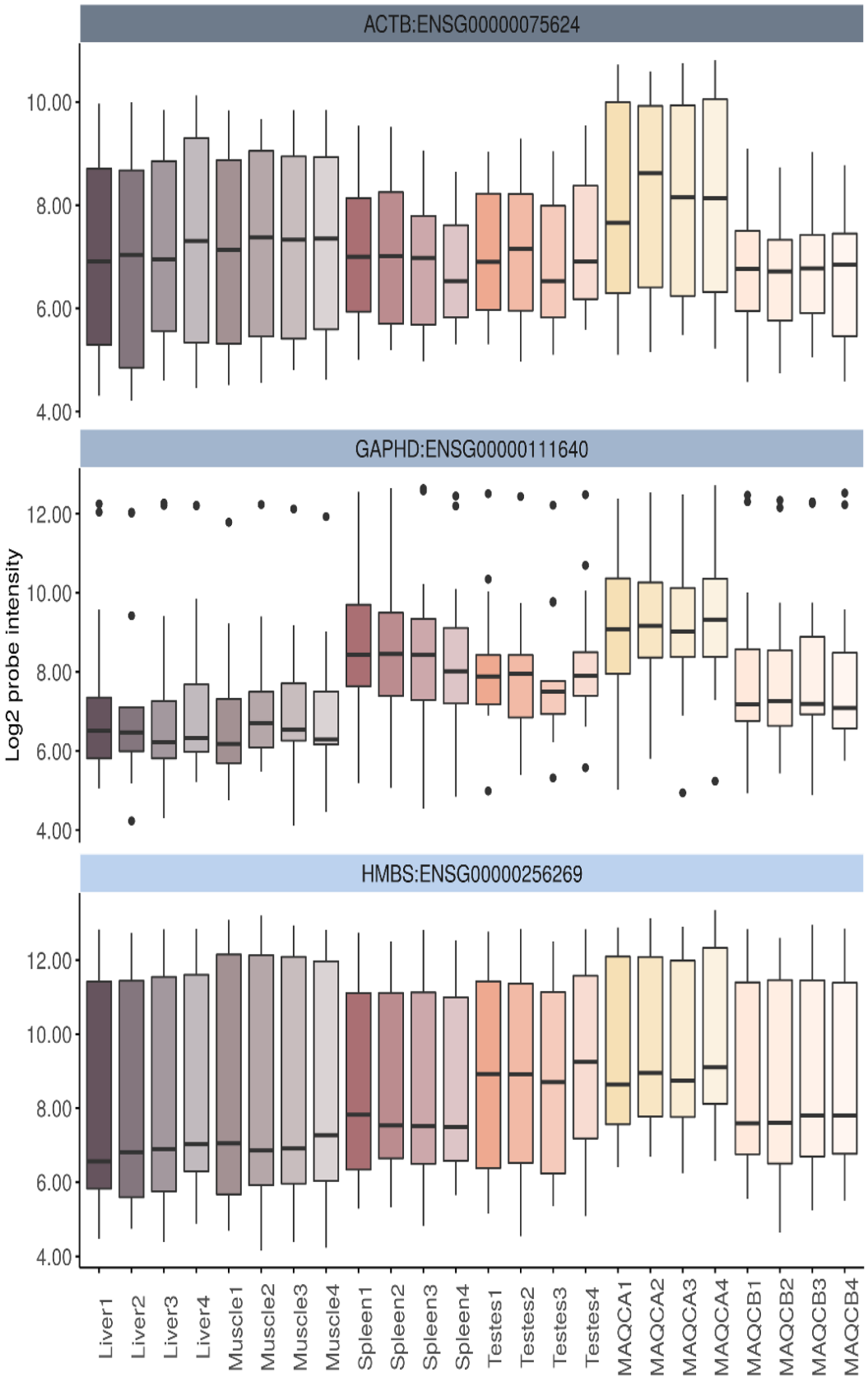

Figure 4.23: Box plots of the raw log2 intensities of all probes within the probesets corresponding to the housekeeping genes ACTB, GAPHD and HMBS in the example data.

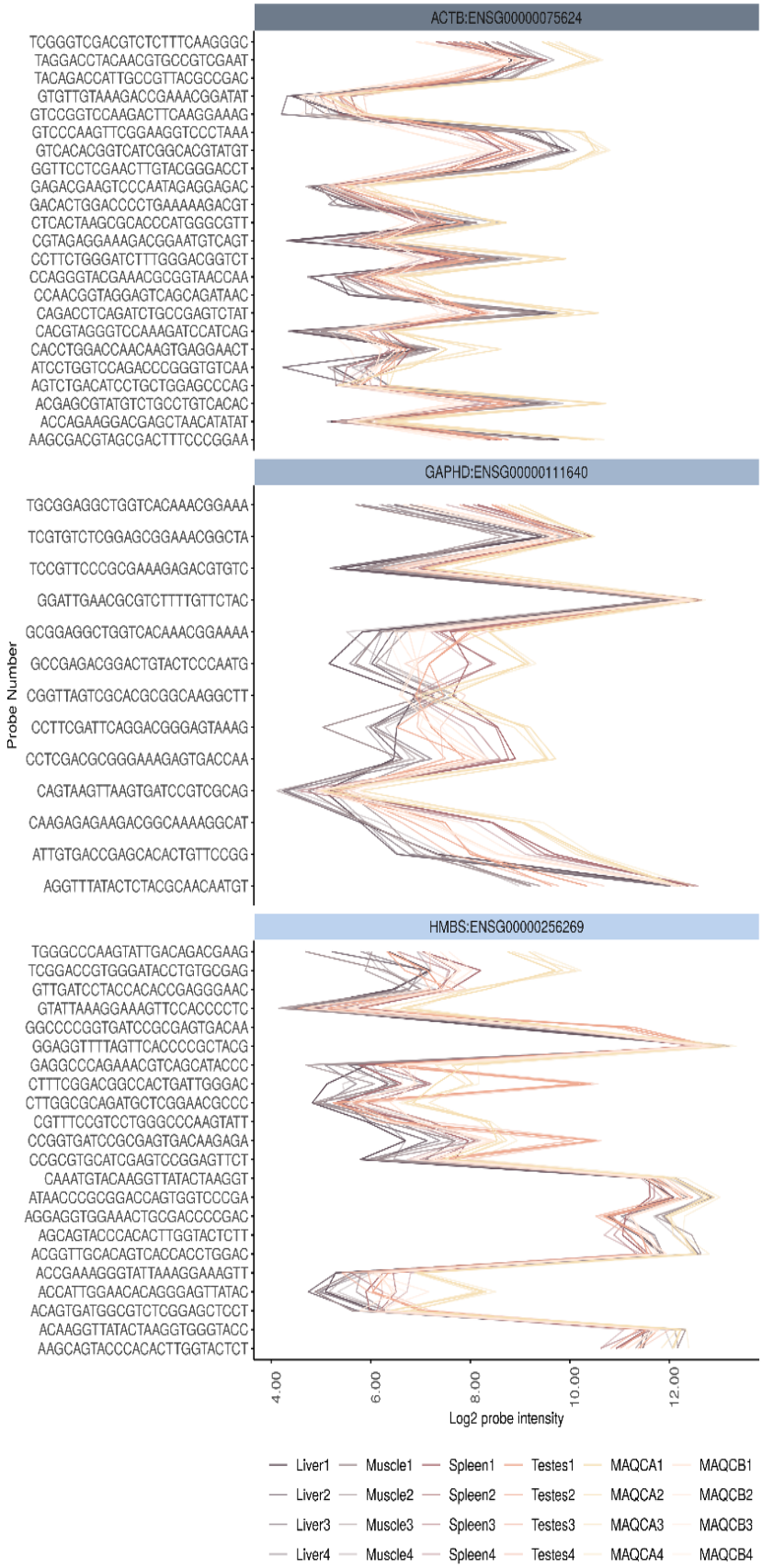

Figure 4.24: Raw log2 intensities of all probes within the probesets corresponding to the housekeeping genes ACTB, GAPHD and HMBS in the example data.

4.3 Intensity-Sequence association

The dependency of the intensity on the probe sequence is well established (Figure 4.25). Estimated affinity splines coefficients can be used to estimate the base-position effects on the log2 intensities. A clear sequence effect on the log2 intensities is revealed. Gs and Cs are associated with positive splines coefficients and thus higher log2 intensities (Figure 4.25). There is moreover a positional effect, with Gs at higher base positions showing larger effects. A boxplot stratified by GC content confirmed the strong dependency of log2-intensities on the number of G or C bases observed in the probe sequence (Figure 4.26). Our results are in line with the findings that probes containing runs of guanine show abnormal binding affinities and are typically outliers with respect to the rest of the probeset to which they are assigned [11].

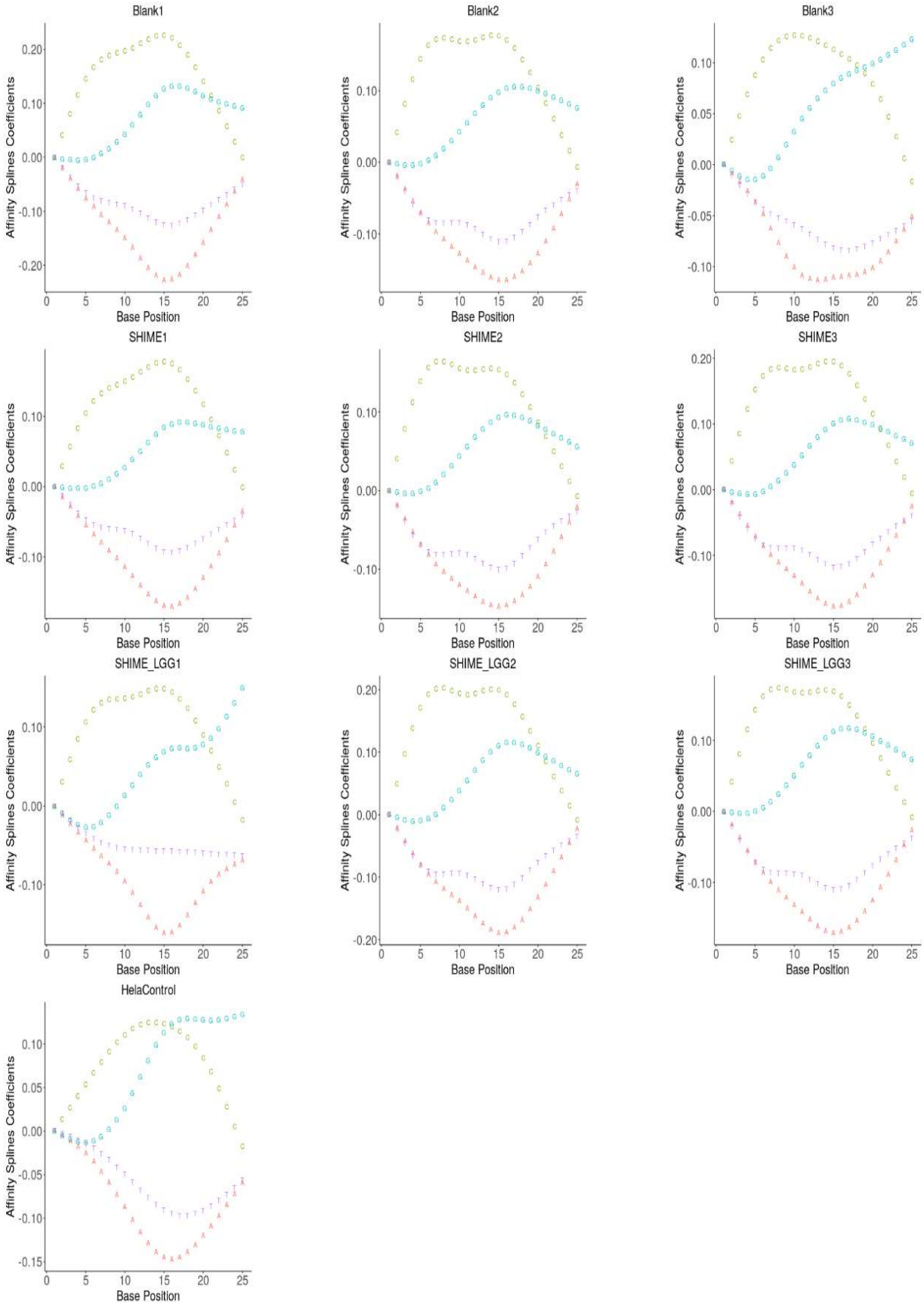

Figure 4.25: Affinity splines coefficients indicating base position effects on the log2 intensities.

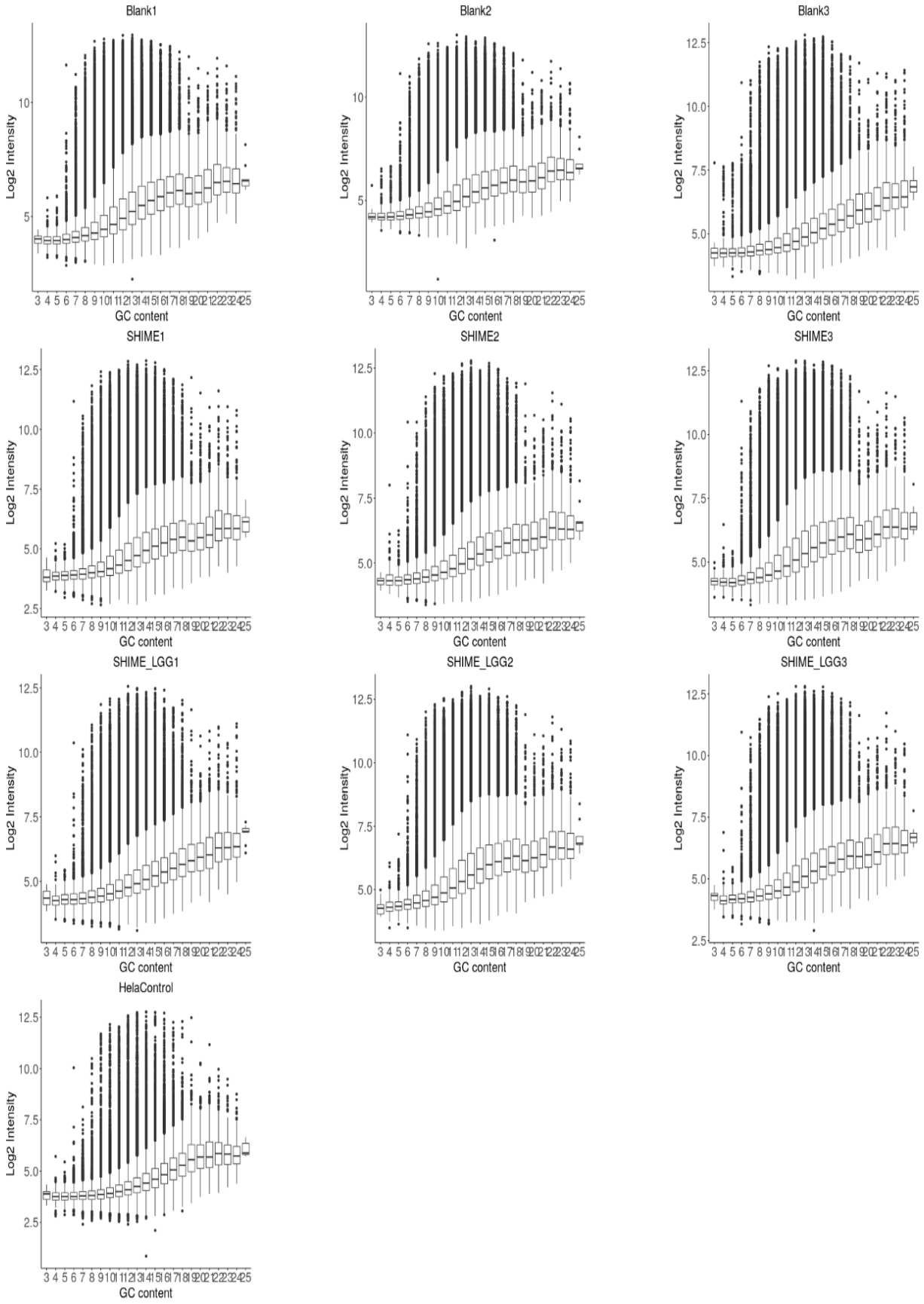

Figure 4.26: Box plots of the log2 intensities in function of the probe GC content.

4.4 ROC curves

As previously mentioned probesets targeting intron regions (normgene→intron) are incorporated in the HuGene chip design and annotated as negative controls in the library files. While 'normgene→exon' was not registered as a type of probesets in our dataset, positive controls were included in the Affymetrix library files within the main category probeset types. These correspond to the 'normgene→exon' loci, an exonic region of normalization control genes (putative housekeeping genes) which are shown to have constitutive expression over a large number of samples. While in any given sample some (or many) of these putative exonic regions may not be transcribed or may be spliced out and some (or many) of the genes may not be constitutive within certain data sets, this collection of probesets in general has moderate to high signal values.

A comparison of signal values between the positive and negative controls reflects the quality of the whole experiment (RNA, target preparation, chip defects, hybridization, scanning) and is interesting given the high number of probes in the positive and negative control groups (in contrast with the spike-in and antigenomic background control probes).

Positive and negative controls are somewhat separated based on raw log2 intensities (Figure 4.27). Hence the log2 unnormalized intensities have some diagnostic ability as binary classifier. This predictive value of log2 intensities for classification into positive vs negative controls is confirmed by Receiver Operator Curves (ROC) (Figure 4.28).

ROC curves plot the true positive fraction (TPF, proportion of observations correctly predicted to be positive out of all positive observations) on the y-axis in function of the false positive fraction (FPF, proportion of observations wrongfully predicted to be positive out of all negative observations) on the x-axis at all classification thresholds. The true positive fraction is also called the sensitivity and $1 - false\,positive\,fraction$ is known as the specificity (= true negative fraction). The sensitivity and specificity can be used to assess the performance and strength of classification models at different classification thresholds. Those thresholds correspond to values of the log2 intensities and are not fixed in advance (range of observed intensities). It is implicitly assumed that the subpopulation of positive controls tends to have larger log2 intensities compared to the subpopulation of negative controls. Hence, a positive control is predicted by a log2 intensity exceeding or equaling some threshold, while a negative control probe is defined by a log2 intensity smaller than that threshold. The classification accuracy is evaluated considering the confusion matrix (Table 4.3). Based on the raw log2 intensity values, probabilities can be assigned that observations (probes) belong to the positive or negative control probesets 4.1. Assuming that the negative controls are a measure of false positives and the positive controls are a measure of true positives, ROC curves are constructed to evaluate how well the probe set signals separate the positive controls from the negative controls.

Table 4.3: The confusion matrix cross-classifies the predicted outcome Log2 intensity $\geq$ threshold versus the true class of probes. This matrix is used to construct ROC curves by calculting sensitivity and specificity according to equation 4.1.

		Predicted class	
Actual class	Negative control	True Negative	False Positive
	Positive control	False Negative	True Positive

$$
\begin{aligned}
True\,positive\,rate &= \frac{True\ Positive}{True\ Positive + False\ Negative} \\
&= P(Log2\ intensity \geq threshold \mid Probe\ is\ a\ positive\ control) \\
True\,negative\,rate &= \frac{False\ Positive}{True\ Negative + False\ Positive} \\
&= P(Log2\ intensity \geq threshold \mid Probe\ is\ a\ negative\ control)
\end{aligned}
\tag{4.1}
$$

The ROC curves for the raw log2 intensities are not on the diagonal, implying some predictive value (Figure 4.28). The diagonal in the ROC curve is a random classifier that does not have any ability to distinguish between the two classes i.e. the predicted probabilities of the two classes overlap and therefore, TPF = FPF at any threshold. A common way to summarize ROC curves in a single value is to calculate the area under the ROC curve (AUC). This presents an aggregate measure of performance of a model across all possible classification thresholds. While the AUC facilitates comparisons between arrays/classifiers, a curve with a lower AUC and lower predictive value across the range of thresholds could still yield a higher predictive value at a specific threshold value. The greater the AUC, the more informative the classifier. A model whose predictions are 100% correct has an AUC of 1.0; one whose predictions are 100% wrong has an AUC of 0.0. A random models has an AUC = 0.5. In our case, the areas under the curve for a predictive model with positive vs negative controls is around 0.8 for all microarrays with smaller values (0.78, 0.77 and 0.68 for SHIME_LGG1, blank3 and the Hela control) (Figure 4.28). Values for microarray data typically range between 0.80 and 0.90, which is in line with the observed values in our experiment and the example dataset (Figure 4.30).

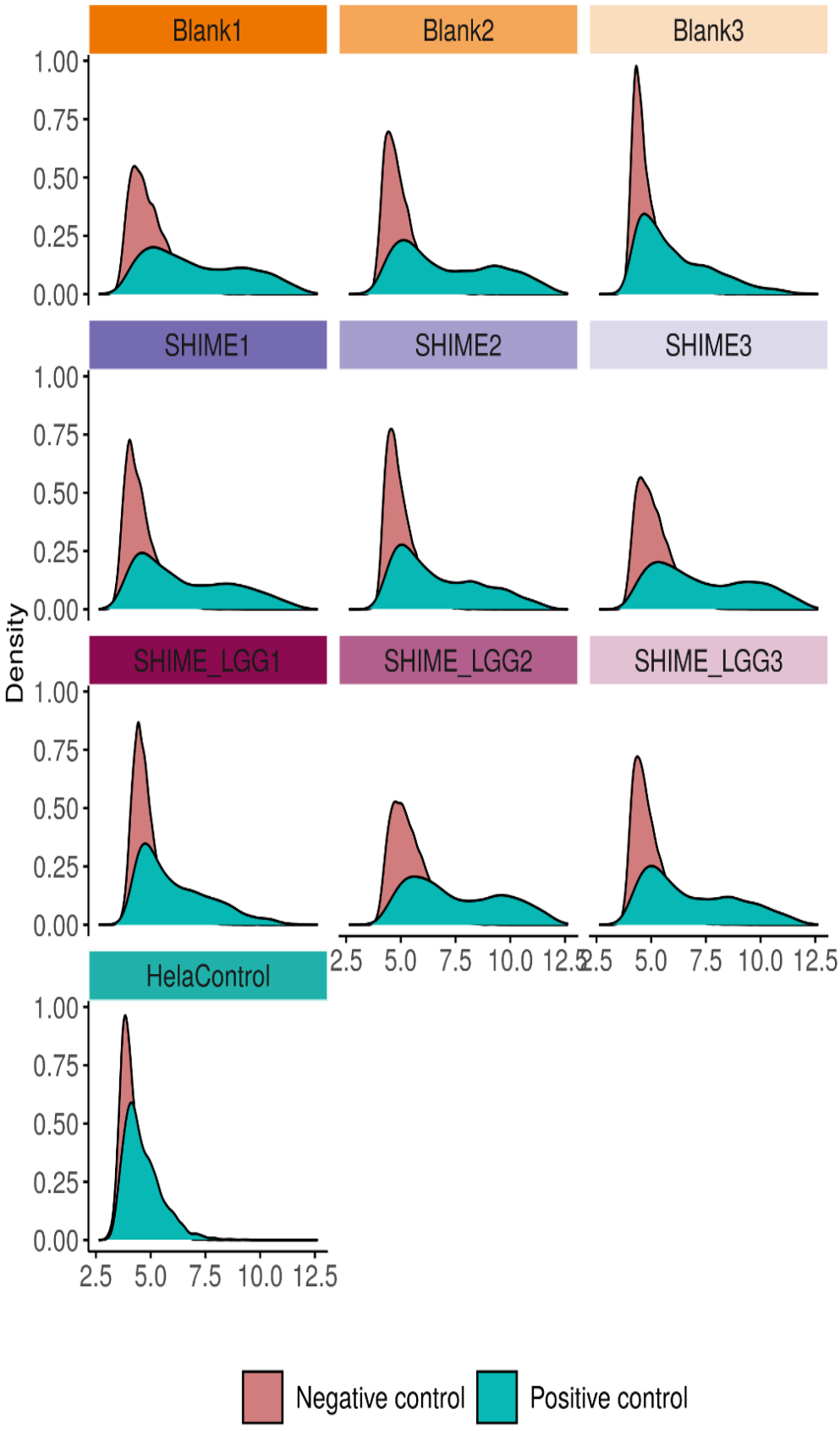

Figure 4.27: Density plot of Log2 intensities for positive and negative controls.

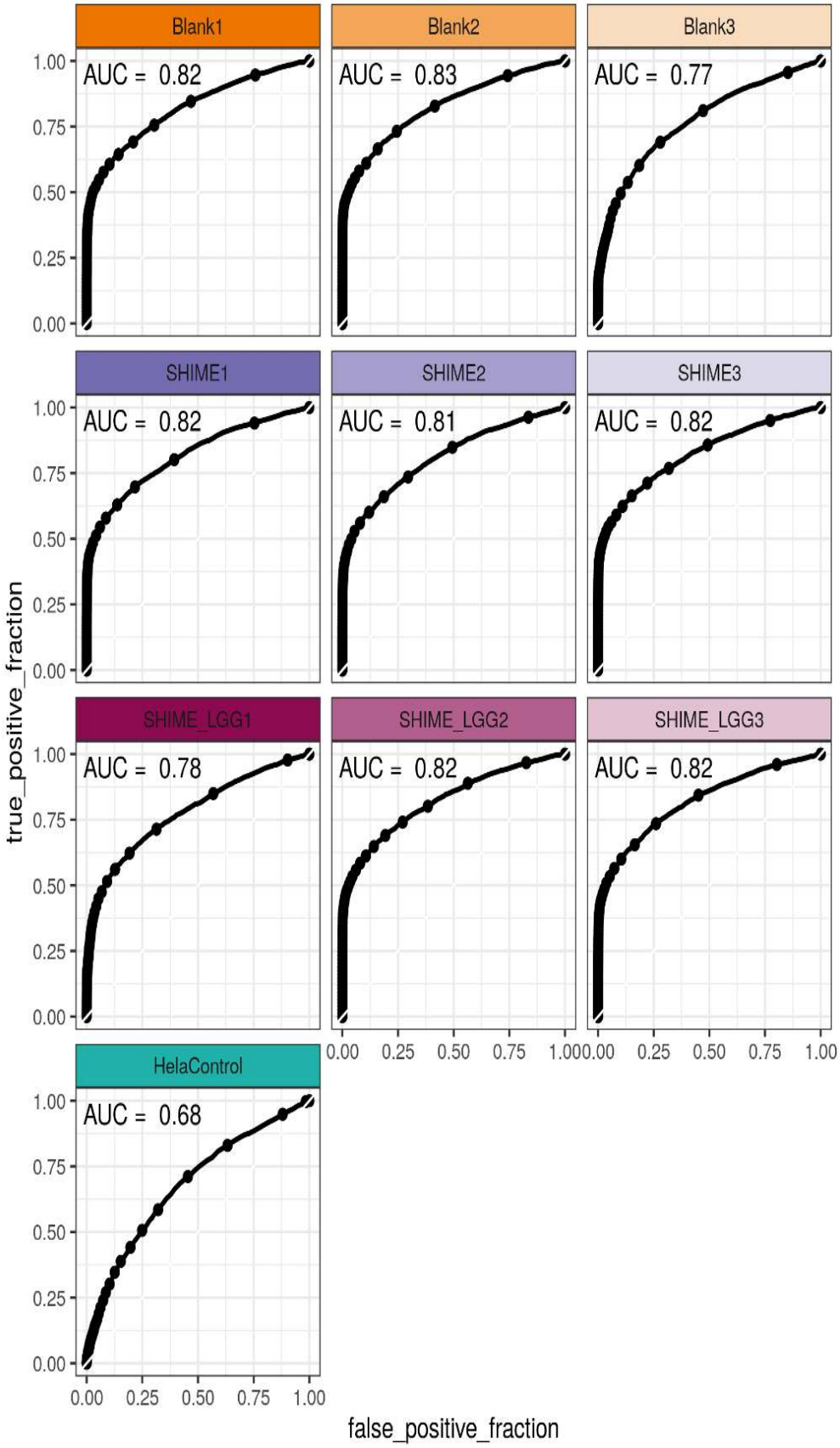

Figure 4.28: ROC curves.

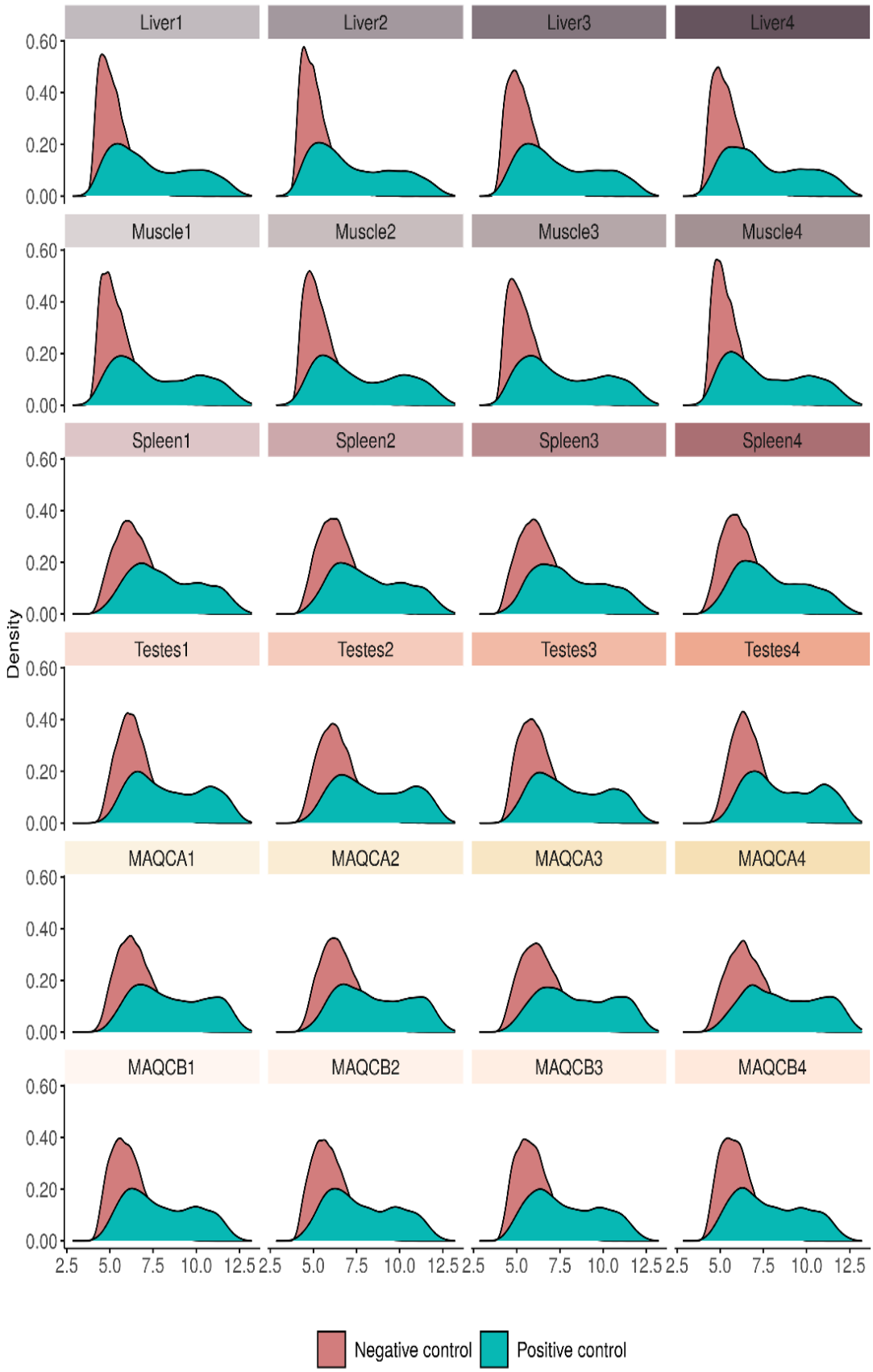

Figure 4.29: Density plot of Log2 intensities for positive and negative controls in the example data.

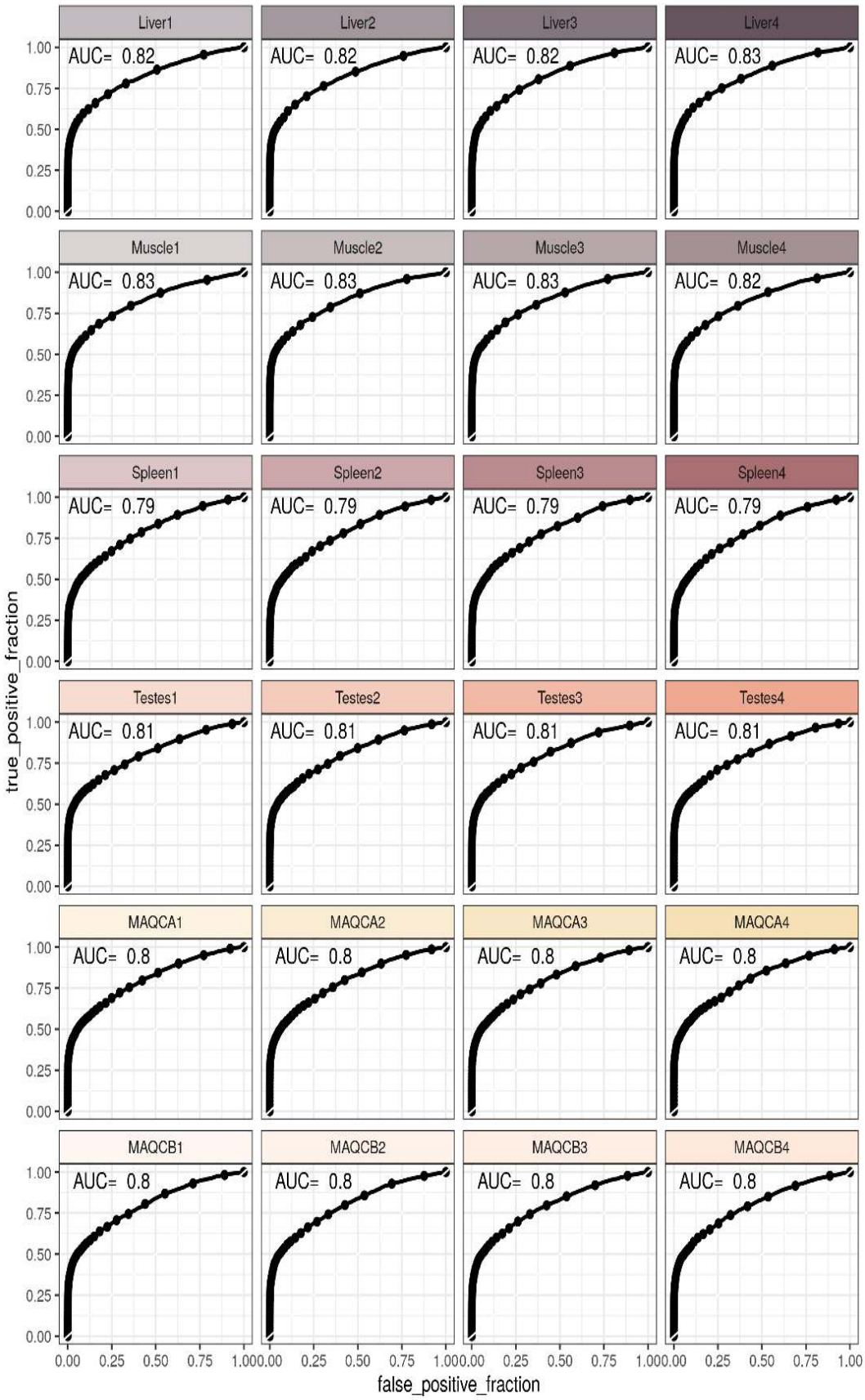

Figure 4.30: ROC curves from the example data.

4.5 MA plots

MA plots were constructed to determine the variability in expression across different expression levels. The MA plot shows to what extent the variability in expression depends on the expression level (e.g. if there is more variation on high expression values). In an MA-plot, A (the average of the intensity of a probe on that array and the median intensity of that probe over all arrays) is plotted versus M (the difference between the intensity of a probe on that array and the median intensity of that probe over all arrays) 4.2.

$$A = \frac{\log_2(PMInt\_array) + \log_2(PMInt\_medianarray)}{2} \tag{4.2}$$

$$M = \log_2(PMInt\_array) - \log_2(PMInt\_medianarray) \tag{4.3}$$

Ideally, the cloud of data points should be centered around M=0 (blue line) since we assume that the majority of the genes is not differentially expressed and that the number of up-regulated genes is similar to the number of down-regulated genes. Additionally, the variability of the M values should be similar for different A values. We observe that the spread of the cloud increases with the average intensity: the loess curve (red line) moves further and further away from M=0 with increasing A 4.31, which is in line with the example data (Figures 4.32 and 4.33). To remove (some of) this dependency, we will normalize the data (see 5.2).

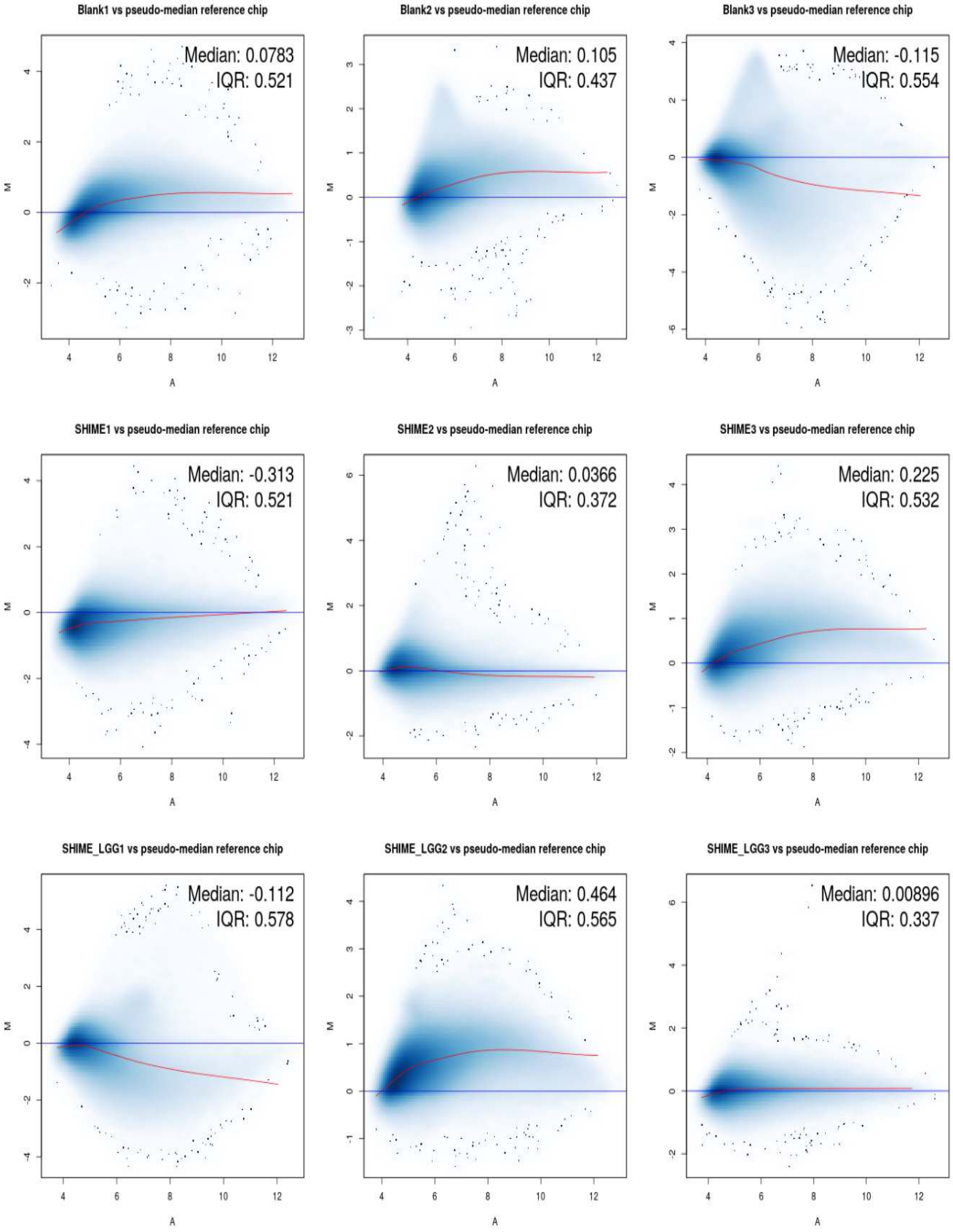

Figure 4.31: MA plots.

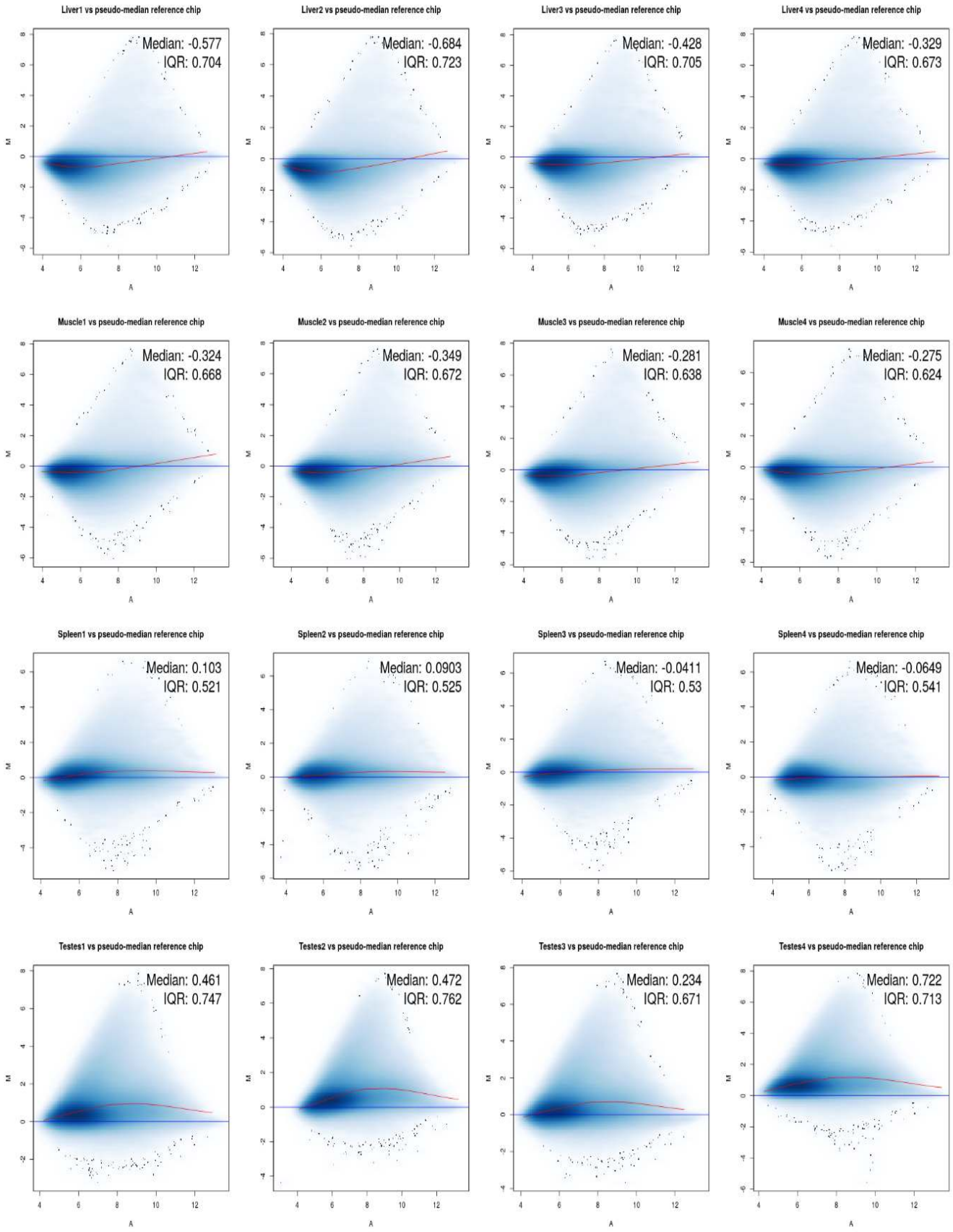

Figure 4.32: MA plots of the tissue example data.

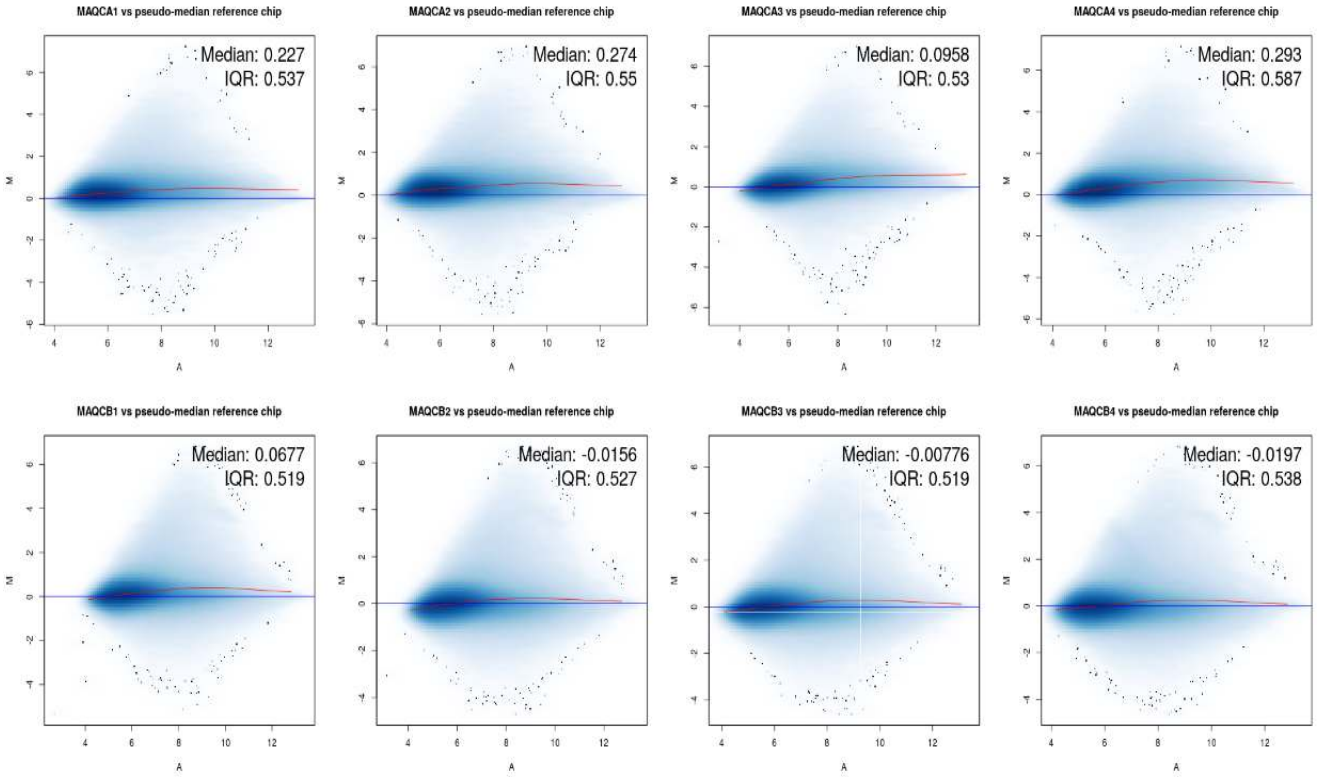

Figure 4.33: MA plots of the MAQC example data.

4.6 Presence/Absence calls - at probe and summarized probeset level

The less clear separation between positive and negative controls for the blank3,SHIME_LGG1 and Hela cell control as observed in ROC curve analysis is in agreement with findings from a presence/absence analysis. Presence/absence calls can be made by applying statistical hypothesis testing to assess whether or not each of the PM intensities are compatible with observations generated by background probes 4.4.

$$H_0 : PM\ intensity = BG\ intensity \tag{4.4}$$
$$H_1 : PM\ intensity > BG\ intensity \tag{4.5}$$

The p-value of this hypothesis test expresses the chance of observing a "present looking" probeset when the complementary RNA is in fact absent. Small p-values imply presence while large ones imply absence of transcripts from a gene/depending on whether the analysis is conducted at probeset (PSDABG) or individual probe level (DABG). Probeset level analysis implies summarization (see 5.3).

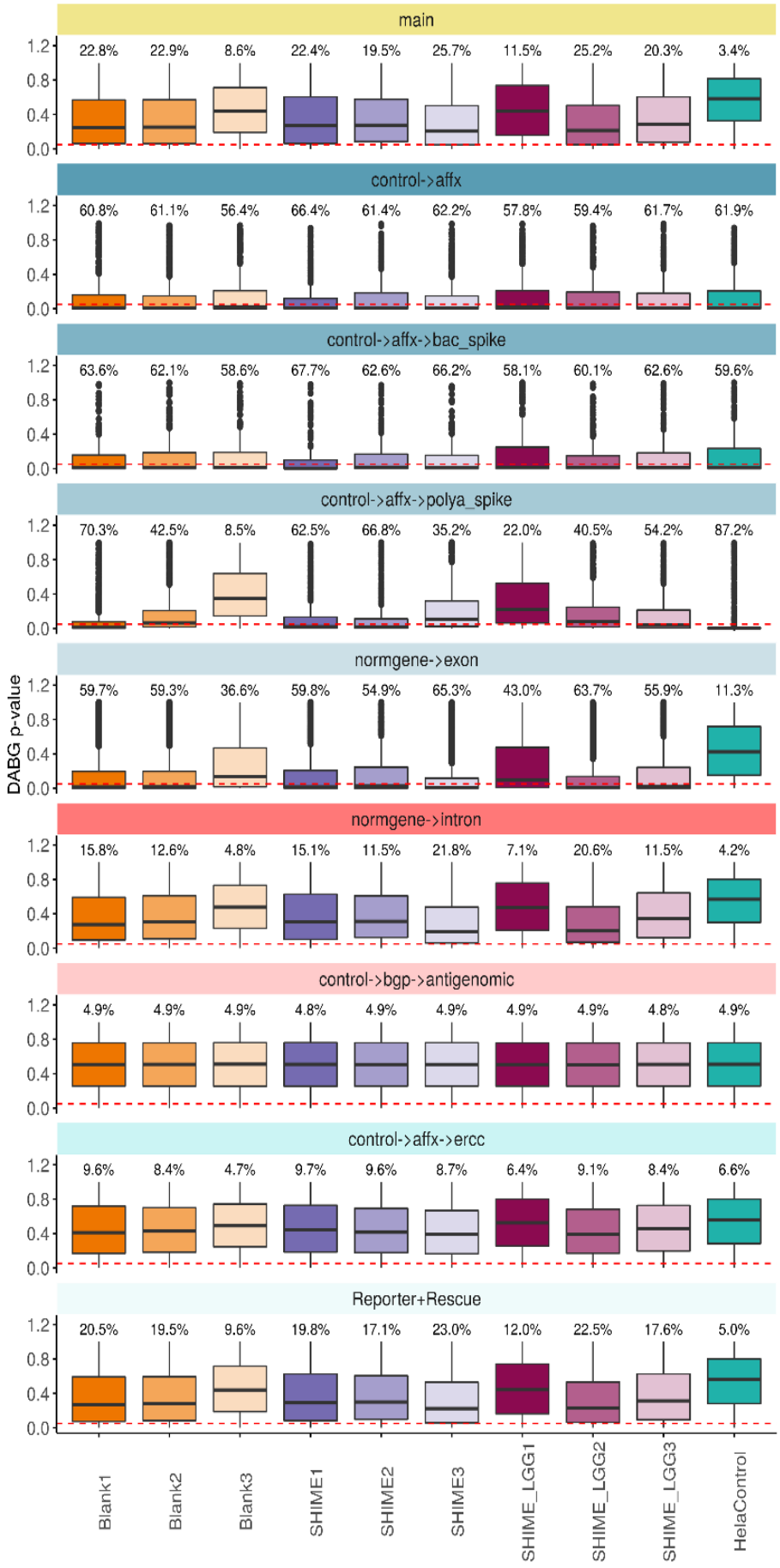

Figure 4.34: Percent present calls detected above background at probe level.

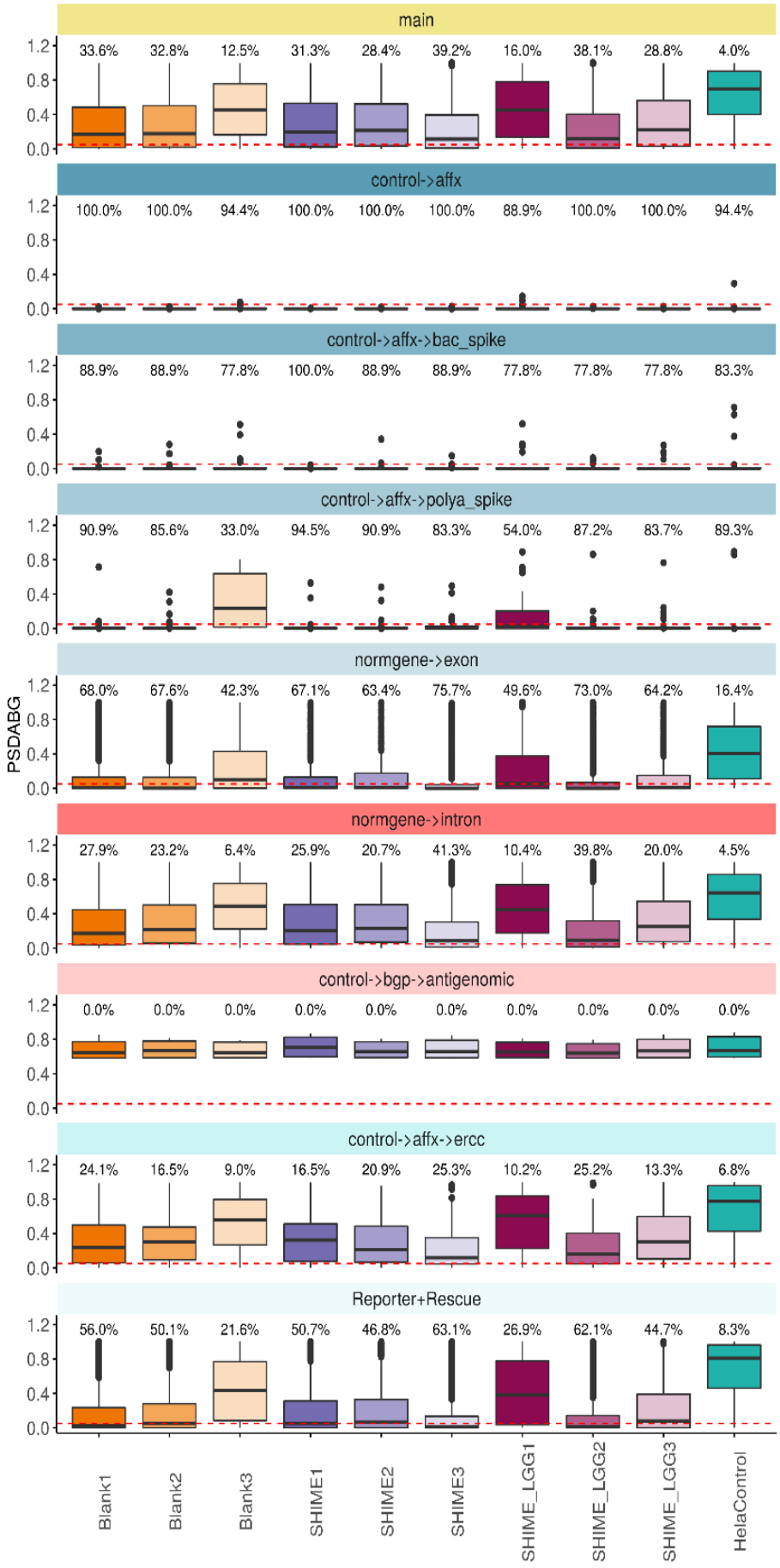

Figure 4.35: Percent present calls detected above background at probeset level.

Table 4.4: Percentage of 'present calls' at probe level.

	Blank1	Blank2	Blank3	SHIME1	SHIME2	SHIME3	SHIME_LGG1	SHIME_LGG2	SHIME_LGG3	HelaControl
main	22.80	22.91	8.61	22.44	19.49	25.75	11.48	25.18	20.26	3.36
control→affx	60.83	61.11	56.39	66.39	61.39	62.22	57.78	59.44	61.67	61.94
control→affx→bac spike	63.64	62.12	58.59	67.68	62.63	66.16	58.08	60.10	62.63	59.60
control→affx→polya spike	70.34	42.45	8.53	62.52	66.79	35.17	22.02	40.50	54.17	87.21
normgene→exon	59.71	59.25	36.61	59.81	54.91	65.30	42.95	63.69	55.90	11.30
normgene→intron	15.81	12.57	4.81	15.12	11.51	21.77	7.11	20.65	11.46	4.15
control→bgp→antigenomic	4.88	4.88	4.92	4.81	4.89	4.88	4.89	4.90	4.84	4.86
control→affx→ercc	9.60	8.35	4.74	9.69	9.65	8.66	6.37	9.09	8.44	6.63
Reporter+Rescue	20.54	19.50	9.56	19.78	17.10	23.04	12.04	22.51	17.59	5.01

Table 4.5: Percentage of 'present calls' at probeset level.

	Blank1	Blank2	Blank3	SHIME1	SHIME2	SHIME3	SHIME_LGG1	SHIME_LGG2	SHIME_LGG3	HelaControl
main	33.57	32.81	12.53	31.30	28.41	39.24	16.04	38.12	28.83	3.97
control→affx	100.00	100.00	94.44	100.00	100.00	100.00	88.89	100.00	100.00	94.44
control→affx→bac spike	88.89	88.89	77.78	100.00	88.89	88.89	77.78	77.78	77.78	83.33
control→affx→polya spike	90.94	85.61	33.04	94.49	90.94	83.30	54.00	87.21	83.66	89.34
normgene→exon	67.99	67.56	42.29	67.10	63.39	75.75	49.59	72.99	64.21	16.40
normgene→intron	27.94	23.17	6.40	25.87	20.67	41.30	10.43	39.77	20.02	4.52
control→bgp→antigenomic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
control→affx→ercc	24.07	16.54	8.96	16.54	20.89	25.28	10.16	25.15	13.31	6.80
Reporter+Rescue	55.97	50.13	21.58	50.66	46.78	63.13	26.91	62.10	44.68	8.30

Table 4.6: Percentage of 'present calls' at probe level in the spike-in controls.

		Blank1	Blank2	Blank3	SHIME1	SHIME2	SHIME3	SHIME_LGG1	SHIME_LGG2	SHIME_LGG3	HelaControl
control→affx	**AFFX-BioB-3 at**	90.00	90.00	90.00	95.00	90.00	95.00	95.00	90.00	90.00	95.00
control→affx	**AFFX-BioB-5 at**	80.00	85.00	65.00	95.00	80.00	85.00	75.00	75.00	80.00	90.00
control→affx	**AFFX-BioB-M at**	95.00	95.00	100.00	100.00	95.00	95.00	100.00	95.00	95.00	100.00
control→affx	**AFFX-BioC-3 at**	90.00	90.00	90.00	95.00	95.00	95.00	90.00	90.00	90.00	95.00
control→affx	**AFFX-BioC-5 at**	95.00	90.00	95.00	100.00	85.00	95.00	85.00	90.00	90.00	95.00
control→affx	**AFFX-BioDn-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	**AFFX-BioDn-5 at**	100.00	95.00	95.00	100.00	100.00	95.00	95.00	95.00	100.00	100.00
control→affx	**AFFX-CreX-3 at**	100.00	100.00	95.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	**AFFX-CreX-5 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioB-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	90.91	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioB-5 at**	81.82	90.91	81.82	100.00	90.91	90.91	90.91	90.91	90.91	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioB-M at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioC-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioC-5 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioD-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioD-5 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-P1-cre-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-P1-cre-5 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-DapX-3 st**	100.00	95.00	10.00	100.00	100.00	90.00	50.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-DapX-5 st**	100.00	90.00	20.00	95.00	95.00	95.00	60.00	90.00	90.00	100.00
control→affx→polya spike	**AFFX-DapX-M st**	100.00	95.00	20.00	100.00	100.00	95.00	45.00	80.00	90.00	100.00
control→affx→polya spike	**AFFX-LysX-3 st**	95.00	75.00	15.00	100.00	90.00	55.00	15.00	50.00	80.00	100.00
control→affx→polya spike	**AFFX-LysX-5 st**	55.00	15.00	5.00	50.00	25.00	5.00	5.00	5.00	15.00	90.00
control→affx→polya spike	**AFFX-LysX-M st**	80.00	40.00	0.00	55.00	60.00	20.00	0.00	25.00	40.00	100.00
control→affx→polya spike	**AFFX-PheX-3 st**	80.00	60.00	10.00	80.00	90.00	55.00	25.00	40.00	65.00	100.00
control→affx→polya spike	**AFFX-PheX-5 st**	85.00	15.00	0.00	60.00	50.00	15.00	0.00	10.00	25.00	100.00
control→affx→polya spike	**AFFX-PheX-M st**	90.00	40.00	5.00	75.00	70.00	25.00	5.00	35.00	30.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-dap-3 st**	100.00	81.82	18.18	100.00	100.00	81.82	72.73	81.82	90.91	100.00
control→affx→polya spike	**AFFX-r2-Bs-dap-5 st**	100.00	90.91	36.36	100.00	100.00	90.91	72.73	90.91	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-dap-M st**	100.00	100.00	27.27	100.00	100.00	100.00	45.45	90.91	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-lys-3 st**	72.73	63.64	9.09	81.82	90.91	54.55	45.45	54.55	45.45	100.00
control→affx→polya spike	**AFFX-r2-Bs-lys-5 st**	54.55	27.27	9.09	36.36	27.27	27.27	9.09	18.18	27.27	63.64
control→affx→polya spike	**AFFX-r2-Bs-lys-M st**	81.82	36.36	0.00	81.82	90.91	36.36	0.00	36.36	45.45	100.00
control→affx→polya spike	**AFFX-r2-Bs-phe-3 st**	72.73	54.55	0.00	72.73	72.73	63.64	27.27	72.73	63.64	100.00
control→affx→polya spike	**AFFX-r2-Bs-phe-5 st**	90.91	45.45	9.09	81.82	81.82	36.36	45.45	54.55	45.45	100.00
control→affx→polya spike	**AFFX-r2-Bs-phe-M st**	81.82	45.45	0.00	72.73	63.64	18.18	9.09	36.36	54.55	90.91
control→affx→polya spike	**AFFX-r2-Bs-thr-3 s st**	81.82	18.18	0.00	81.82	81.82	18.18	0.00	27.27	72.73	100.00
control→affx→polya spike	**AFFX-r2-Bs-thr-5 s st**	90.91	27.27	9.09	63.64	72.73	9.09	27.27	36.36	72.73	100.00
control→affx→polya spike	**AFFX-r2-Bs-thr-M s st**	90.91	36.36	0.00	90.91	90.91	0.00	0.00	36.36	81.82	100.00
control→affx→polya spike	**AFFX-ThrX-3 st**	85.00	40.00	10.00	80.00	85.00	30.00	10.00	45.00	80.00	100.00
control→affx→polya spike	**AFFX-ThrX-5 st**	80.00	50.00	10.00	65.00	80.00	10.00	10.00	40.00	65.00	85.00
control→affx→polya spike	**AFFX-ThrX-M st**	95.00	55.00	0.00	85.00	85.00	25.00	0.00	60.00	90.00	100.00
control→affx→polya spike	**AFFX-TrpnX-3 st**	0.00	15.00	0.00	0.00	15.00	10.00	0.00	15.00	10.00	0.00
control→affx→polya spike	**AFFX-TrpnX-5 st**	15.00	0.00	0.00	10.00	0.00	15.00	10.00	10.00	10.00	5.00
control→affx→polya spike	**AFFX-TrpnX-M st**	15.00	5.00	5.00	15.00	10.00	10.00	15.00	5.00	5.00	5.00

Table 4.7: Percentage of 'present calls' at probeset level in the spike-in controls.

		Blank1	Blank2	Blank3	SHIME1	SHIME2	SHIME3	SHIME_LGG1	SHIME_LGG2	SHIME_LGG3	HelaControl
control→affx	**AFFX-BioB-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	**AFFX-BioB-5 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	**AFFX-BioB-M at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	**AFFX-BioC-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	**AFFX-BioC-5 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	**AFFX-BioDn-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	**AFFX-BioDn-5 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	**AFFX-CreX-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	**AFFX-CreX-5 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioB-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioB-5 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioB-M at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioC-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioC-5 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioD-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-Ec-bioD-5 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-P1-cre-3 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	**AFFX-r2-P1-cre-5 at**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-DapX-3 st**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-DapX-5 st**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-DapX-M st**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-LysX-3 st**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-LysX-5 st**	100.00	100.00	0.00	100.00	100.00	0.00	0.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-LysX-M st**	100.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-PheX-3 st**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-PheX-5 st**	100.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-PheX-M st**	100.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-dap-3 st**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-dap-5 st**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-dap-M st**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-lys-3 st**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-lys-5 st**	100.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-lys-M st**	100.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-phe-3 st**	100.00	100.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-phe-5 st**	100.00	100.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-phe-M st**	100.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-thr-3 s st**	100.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-thr-5 s st**	100.00	100.00	0.00	100.00	100.00	0.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-r2-Bs-thr-M s st**	100.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-ThrX-3 st**	100.00	100.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-ThrX-5 st**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-ThrX-M st**	100.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00
control→affx→polya spike	**AFFX-TrpnX-3 st**	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
control→affx→polya spike	**AFFX-TrpnX-5 st**	0.00	0.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	0.00
control→affx→polya spike	**AFFX-TrpnX-M st**	100.00	0.00	0.00	100.00	0.00	100.00	0.00	0.00	0.00	0.00

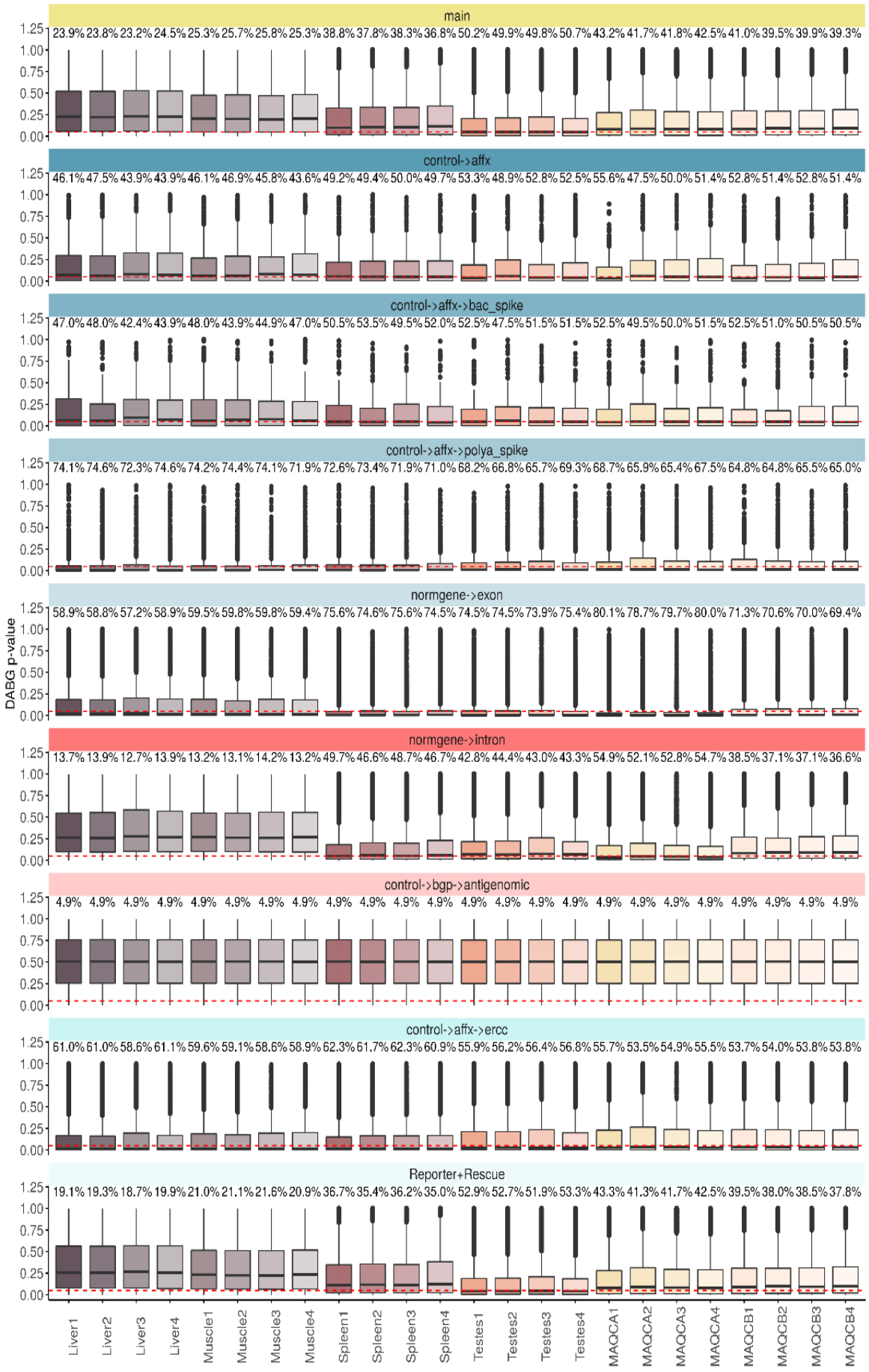

Figure 4.36: Percent present calls detected above background at probe level in the example data.

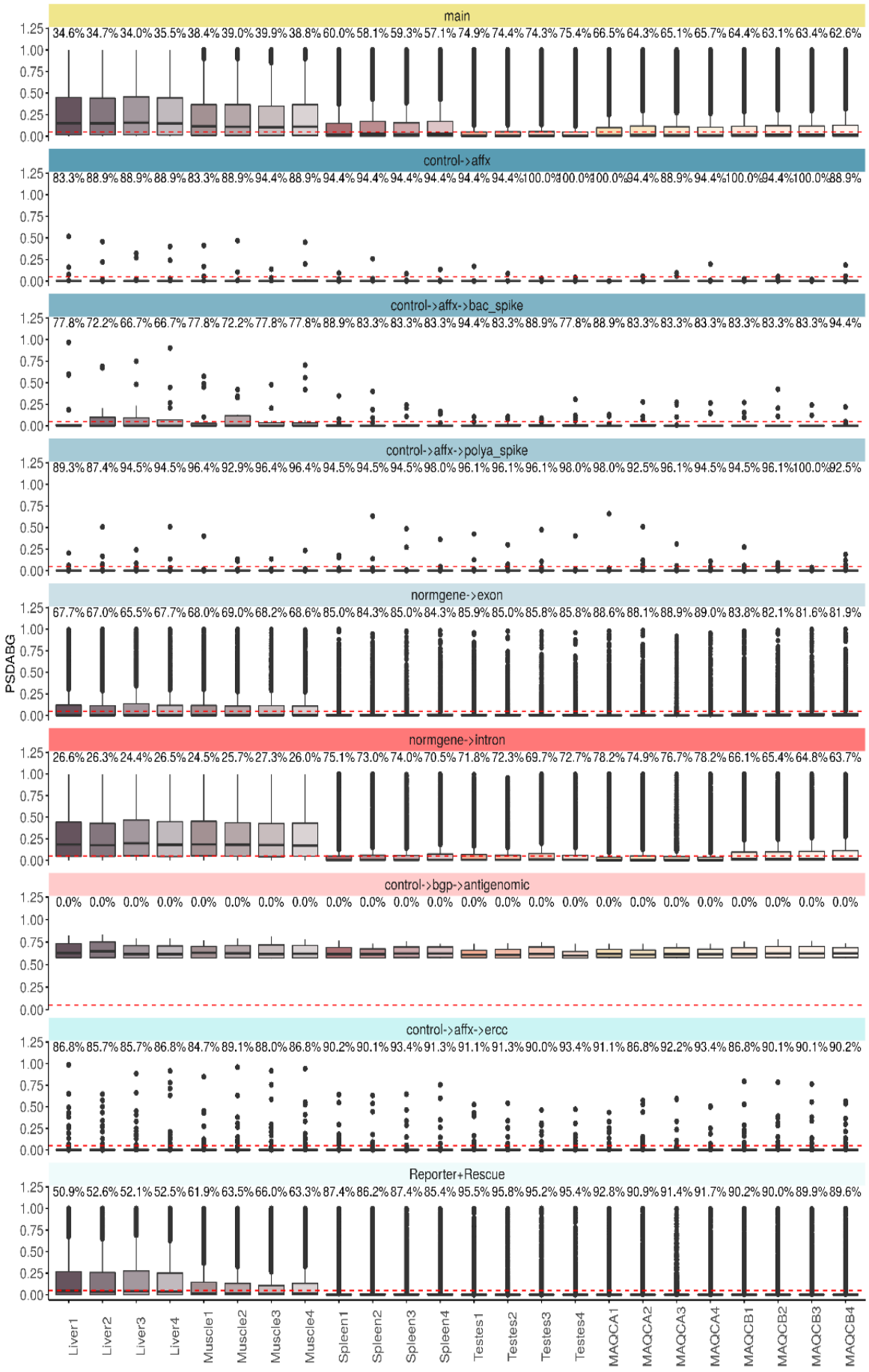

Figure 4.37: Percent present calls detected above background at probeset level in the example data.

Table 4.8: Percentage of 'present calls' at probe level.

	Liver1	Liver2	Liver3	Liver4	Muscle1	Muscle2	Muscle3	Muscle4	Spleen1	Spleen2	Spleen3	Spleen4	Testes1	Testes2	Testes3	Testes4	MAQCA1	MAQCA2	MAQCA3	MAQCA4	MAQCB1	MAQCB2	MAQCB
main	23.87	23.79	23.20	24.51	25.32	25.67	25.76	25.29	38.75	37.80	38.26	36.78	50.22	49.90	49.81	50.72	43.18	41.75	41.84	42.50	41.00	39.52	39.87
control→affx	46.11	47.50	43.89	43.89	46.11	46.94	45.83	43.61	49.17	49.44	50.00	49.72	53.33	48.89	52.78	52.50	55.56	47.50	50.00	51.39	52.78	51.39	52.78
control→affx→bac spike	46.97	47.98	42.42	43.94	47.98	43.94	44.95	46.97	50.51	53.54	49.49	52.02	52.53	47.47	51.52	51.52	52.53	49.49	50.00	51.52	52.53	51.01	50.51
control→affx→polya spike	74.07	74.60	72.29	74.60	74.25	74.42	74.07	71.94	72.65	73.36	71.94	71.05	68.21	66.79	65.72	69.27	68.74	65.90	65.36	67.50	64.83	64.83	65.54
normgene→exon	58.86	58.76	57.25	58.92	59.48	59.84	59.81	59.38	75.58	74.56	75.65	74.47	74.50	74.50	73.94	75.42	80.09	78.67	79.72	79.99	71.34	70.56	69.96
normgene→intron	13.65	13.85	12.71	13.85	13.22	13.14	14.20	13.17	49.69	46.65	48.71	46.66	42.78	44.40	43.02	43.29	54.94	52.13	52.76	54.66	38.48	37.10	37.11
control→bgp→antigenomic	4.91	4.89	4.91	4.90	4.89	4.89	4.88	4.91	4.91	4.89	4.90	4.87	4.92	4.92	4.90	4.92	4.90	4.90	4.90	4.91	4.90	4.90	4.89
control→affx→ercc	60.98	60.98	58.61	61.11	59.60	59.09	58.57	58.87	62.32	61.71	62.32	60.94	55.90	56.24	56.37	56.85	55.73	53.53	54.87	55.51	53.66	53.96	53.79
Reporter+Rescue	19.11	19.29	18.74	19.94	21.03	21.10	21.63	20.93	36.69	35.43	36.18	34.96	52.89	52.72	51.92	53.35	43.29	41.31	41.72	42.46	39.47	37.97	38.53

Table 4.9: Percentage of 'present calls' at probeset level.

	Liver1	Liver2	Liver3	Liver4	Muscle1	Muscle2	Muscle3	Muscle4	Spleen1	Spleen2	Spleen3	Spleen4	Testes1	Testes2	Testes3	Testes4	MAQCA1	MAQCA2	MAQCA3	MAQCA4	MAQCB1	MAQCB2	MAQCB
main	34.62	34.73	33.98	35.46	38.42	38.98	39.92	38.79	60.00	58.11	59.32	57.10	74.93	74.44	74.27	75.43	66.47	64.27	65.15	65.72	64.41	63.10	63.45
control→affx	83.33	88.89	88.89	88.89	83.33	88.89	94.44	88.89	94.44	94.44	94.44	94.44	94.44	94.44	100.00	100.00	100.00	94.44	88.89	94.44	100.00	94.44	100.00
control→affx→bac spike	77.78	72.22	66.67	66.67	77.78	72.22	77.78	77.78	88.89	83.33	83.33	83.33	94.44	83.33	88.89	77.78	88.89	83.33	83.33	83.33	83.33	83.33	83.33
control→affx→polya spike	89.34	87.39	94.49	94.49	96.45	92.90	96.45	96.45	94.49	94.49	94.49	98.05	96.09	96.09	96.09	98.05	98.05	92.54	96.09	94.49	94.49	96.09	100.00
normgene→exon	67.70	67.01	65.53	67.66	67.99	69.04	68.19	68.58	85.05	84.32	84.98	84.29	85.90	85.05	85.80	85.77	88.60	88.07	88.93	88.99	83.80	82.12	81.56
normgene→intron	26.62	26.30	24.38	26.53	24.52	25.67	27.32	25.99	75.05	72.98	74.01	70.50	71.85	72.31	69.71	72.66	78.23	74.92	76.68	78.21	66.09	65.39	64.77
control→bgp→antigenomic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
control→affx→ercc	86.82	85.75	85.75	86.82	84.67	89.10	88.03	86.82	90.18	90.05	93.41	91.26	91.13	91.26	90.01	93.41	91.13	86.78	92.20	93.41	86.82	90.05	90.05
Reporter+Rescue	50.93	52.64	52.14	52.53	61.85	63.47	65.96	63.26	87.38	86.22	87.42	85.44	95.47	95.81	95.20	95.39	92.75	90.89	91.36	91.72	90.22	89.98	89.86

Table 4.10: Percentage of 'present calls' at probe level in the spike-in controls.

		Liver1	Liver2	Liver3	Liver4	Muscle1	Muscle2	Muscle3	Muscle4	Spleen1	Spleen2	Spleen3	Spleen4	Testes1	Testes2	Testes3	Testes4	MAQCA1	MAQCA2	MAQCA3	MAQCA4	MAQCB1	MAQCB2	MAQCB3	MAQCB4
control→affx	AFFX-BioB-3 at	45.00	45.00	35.00	40.00	40.00	40.00	45.00	40.00	40.00	50.00	55.00	45.00	40.00	45.00	40.00	40.00	70.00	45.00	45.00	40.00	55.00	55.00	50.00	45.00
control→affx	AFFX-BioB-5 at	5.00	25.00	0.00	15.00	10.00	15.00	10.00	5.00	20.00	15.00	15.00	35.00	30.00	25.00	35.00	35.00	45.00	20.00	20.00	20.00	15.00	30.00	20.00	35.00
control→affx	AFFX-BioB-M at	40.00	40.00	30.00	40.00	40.00	45.00	40.00	35.00	50.00	45.00	50.00	60.00	50.00	40.00	55.00	45.00	60.00	45.00	45.00	50.00	65.00	55.00	50.00	45.00
control→affx	AFFX-BioC-3 at	85.00	90.00	85.00	85.00	90.00	90.00	80.00	80.00	85.00	85.00	85.00	85.00	90.00	80.00	90.00	85.00	90.00	85.00	85.00	85.00	90.00	90.00	90.00	85.00
control→affx	AFFX-BioC-5 at	85.00	85.00	85.00	80.00	85.00	85.00	80.00	80.00	85.00	85.00	85.00	80.00	85.00	80.00	85.00	85.00	80.00	80.00	85.00	85.00	85.00	90.00	85.00	85.00
control→affx	AFFX-BioDn-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	AFFX-BioDn-5 at	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00
control→affx	AFFX-CreX-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	AFFX-CreX-5 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-Ec-bioB-3 at	54.55	54.55	36.36	45.45	63.64	54.55	36.36	54.55	63.64	72.73	54.55	63.64	63.64	36.36	63.64	54.55	63.64	63.64	54.55	63.64	54.55	63.64	54.55	63.64
control→affx→bac spike	AFFX-r2-Ec-bioB-5 at	18.18	9.09	0.00	0.00	27.27	0.00	0.00	27.27	18.18	45.45	36.36	45.45	18.18	0.00	36.36	36.36	36.36	9.09	18.18	18.18	27.27	36.36	18.18	27.27
control→affx→bac spike	AFFX-r2-Ec-bioB-M at	27.27	36.36	18.18	18.18	27.27	9.09	18.18	18.18	27.27	36.36	27.27	27.27	45.45	36.36	36.36	36.36	45.45	36.36	45.45	45.45	36.36	54.55	45.45	45.45
control→affx→bac spike	AFFX-r2-Ec-bioC-3 at	100.00	100.00	100.00	90.91	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-Ec-bioC-5 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-Ec-bioD-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-Ec-bioD-5 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-P1-cre-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-P1-cre-5 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-DapX-3 at	95.00	80.00	75.00	95.00	90.00	90.00	80.00	80.00	70.00	75.00	75.00	70.00	75.00	70.00	70.00	75.00	65.00	60.00	65.00	65.00	55.00	75.00	70.00	65.00
control→affx→polya spike	AFFX-DapX-5 at	95.00	100.00	95.00	100.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00
control→affx→polya spike	AFFX-DapX-M st	95.00	95.00	95.00	100.00	100.00	95.00	100.00	100.00	95.00	100.00	95.00	100.00	85.00	85.00	80.00	85.00	95.00	90.00	90.00	90.00	85.00	95.00	90.00	85.00
control→affx→polya spike	AFFX-LysX-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-LysX-5 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-LysX-M st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-PheX-3 at	95.00	90.00	90.00	90.00	80.00	85.00	80.00	85.00	85.00	90.00	85.00	90.00	85.00	80.00	80.00	80.00	85.00	80.00	80.00	80.00	85.00	80.00	75.00	70.00
control→affx→polya spike	AFFX-PheX-5 at	100.00	95.00	100.00	100.00	100.00	100.00	100.00	95.00	100.00	95.00	100.00	95.00	85.00	80.00	90.00	100.00	90.00	90.00	90.00	90.00	80.00	85.00	85.00	90.00
control→affx→polya spike	AFFX-PheX-M st	95.00	100.00	95.00	95.00	95.00	95.00	95.00	95.00	100.00	95.00	100.00	100.00	90.00	90.00	90.00	100.00	95.00	90.00	90.00	90.00	90.00	100.00	80.00	95.00
control→affx→polya spike	AFFX-r2-Bs-dap-3 st	72.73	90.91	72.73	81.82	72.73	81.82	72.73	72.73	81.82	81.82	81.82	63.64	72.73	63.64	54.55	54.55	72.73	63.64	54.55	72.73	54.55	72.73	54.55	63.64
control→affx→polya spike	AFFX-r2-Bs-dap-5 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-dap-M st	100.00	100.00	100.00	100.00	90.91	100.00	90.91	90.91	100.00	100.00	100.00	100.00	90.91	90.91	81.82	90.91	90.91	90.91	81.82	81.82	81.82	72.73	81.82	90.91
control→affx→polya spike	AFFX-r2-Bs-lys-3 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-lys-5 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-lys-M st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-phe-3 st	90.91	81.82	81.82	90.91	90.91	90.91	81.82	90.91	81.82	90.91	81.82	81.82	72.73	72.73	63.64	81.82	72.73	72.73	63.64	72.73	63.64	63.64	63.64	72.73
control→affx→polya spike	AFFX-r2-Bs-phe-5 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	90.91	100.00	100.00	100.00	90.91	81.82	90.91	90.91	90.91	81.82	81.82	100.00	90.91	72.73	90.91	81.82
control→affx→polya spike	AFFX-r2-Bs-phe-M st	90.91	90.91	90.91	90.91	90.91	90.91	90.91	90.91	90.91	90.91	90.91	90.91	90.91	90.91	90.91	81.82	81.82	81.82	81.82	90.91	100.00	81.82	72.73	81.82
control→affx→polya spike	AFFX-r2-Bs-thr-3 s st	81.82	90.91	63.64	72.73	54.55	72.73	72.73	63.64	81.82	72.73	81.82	72.73	72.73	63.64	63.64	63.64	63.64	54.55	54.55	54.55	45.45	45.45	63.64	54.55
control→affx→polya spike	AFFX-r2-Bs-thr-5 s st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	90.91	100.00	100.00	100.00	90.91	90.91	90.91	90.91	90.91	90.91	90.91	81.82	90.91	90.91	90.91	90.91	90.91
control→affx→polya spike	AFFX-r2-Bs-thr-M s st	100.00	100.00	90.91	90.91	90.91	90.91	90.91	90.91	100.00	100.00	90.91	90.91	90.91	81.82	72.73	81.82	90.91	81.82	81.82	81.82	81.82	72.73	81.82	81.82
control→affx→polya spike	AFFX-ThrX-3 at	85.00	80.00	75.00	85.00	60.00	75.00	80.00	55.00	75.00	75.00	70.00	65.00	60.00	60.00	50.00	55.00	50.00	35.00	55.00	40.00	40.00	35.00	40.00	40.00
control→affx→polya spike	AFFX-ThrX-5 at	85.00	90.00	90.00	85.00	85.00	85.00	85.00	85.00	85.00	85.00	85.00	85.00	90.00	80.00	85.00	85.00	85.00	80.00	80.00	85.00	80.00	75.00	80.00	75.00
control→affx→polya spike	AFFX-ThrX-M st	100.00	100.00	100.00	95.00	90.00	95.00	100.00	90.00	90.00	90.00	90.00	95.00	75.00	75.00	70.00	85.00	85.00	85.00	65.00	75.00	75.00	80.00	75.00	75.00
control→affx→polya spike	AFFX-TrpnX-3 st	5.00	10.00	10.00	15.00	15.00	10.00	20.00	10.00	10.00	15.00	15.00	20.00	20.00	20.00	15.00	20.00	20.00	20.00	15.00	15.00	20.00	15.00	15.00	10.00
control→affx→polya spike	AFFX-TrpnX-5 st	5.00	0.00	5.00	10.00	15.00	15.00	15.00	15.00	25.00	15.00	25.00	20.00	25.00	20.00	20.00	25.00	25.00	20.00	15.00	20.00	25.00	20.00	20.00	15.00
control→affx→polya spike	AFFX-TrpnX-M st	5.00	5.00	10.00	5.00	10.00	15.00	15.00	5.00	15.00	10.00	5.00	5.00	5.00	15.00	10.00	10.00	20.00	15.00	10.00	5.00	5.00	5.00	10.00	15.00

Table 4.11: Percentage of 'present calls' at probe level in the spike-in controls.

		Liver1	Liver2	Liver3	Liver4	Muscle1	Muscle2	Muscle3	Muscle4	Spleen1	Spleen2	Spleen3	Spleen4	Testes1	Testes2	Testes3	Testes4	MAQCA1	MAQCA2	MAQCA3	MAQCA4	MAQCB1	MAQCB2	MAQCB3	MAQCB4
control→affx	AFFX-BioB-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	AFFX-BioB-5 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	AFFX-BioB-M at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	AFFX-BioC-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	AFFX-BioC-5 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	AFFX-BioDn-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	AFFX-BioDn-5 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	AFFX-CreX-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx	AFFX-CreX-5 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-Ec-bioB-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-Ec-bioB-5 at	100.00	100.00	100.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-Ec-bioB-M at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-Ec-bioC-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-Ec-bioC-5 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-Ec-bioD-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-Ec-bioD-5 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-P1-cre-3 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→bac spike	AFFX-r2-P1-cre-5 at	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-DapX-3 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-DapX-5 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-DapX-M st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-LysX-3 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-LysX-5 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-LysX-M st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-PheX-3 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-PheX-5 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-PheX-M st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-dap-3 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-dap-5 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-dap-M st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-lys-3 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-lys-5 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-lys-M st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-phe-3 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-phe-5 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-phe-M st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-thr-3 s st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-thr-5 s st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-r2-Bs-thr-M s st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-ThrX-3 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-ThrX-5 st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-ThrX-M st	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-TrpnX-3 st	0.00	0.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-TrpnX-5 st	0.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
control→affx→polya spike	AFFX-TrpnX-M st	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0.00	100.00	100.00	0.00	100.00	100.00	0.00

As expected the antigenomic background probes have high p-values and the percent present calls (in the absence of transcripts) is fixed at 5% (Figures 4.34 and 4.35). The negative controls and main probes are within the same range (4-20%), whereas the percentage of present calls in the positive controls is around 60%. Probeset level analysis decreases the observed variability and at transcript level, the antigenomic background controls all display 0% present calls. Positive controls range between 70 to 100 %. Negative controls and main probe present calls fall within the same range. Prior to normalization blank3, SHIME_LGG1 and the Hela cell control show deviation patterns and a consistently lower percent present calls in every probe type group, except for the bac spike and background controls. Whereas this variability in the percentage of present calls might represent actual biological variability in the Hela cell control versus the coculture cell samples, the number of probesets called present relative to the total number of probesets in the replicate coculture samples should be similar. The higher polyA signal reported earlier in the Hela control translates in an increased percentage of presence calls, setting the Hela control apart from the blank3 andSHIME_LGG1 sample. As mentioned before this suggests an increased polyA/total intact RNA ratio in the Hela control and is indicative of a less efficient target preparation in the deviation blank and SHIME_LGG1 replicates. Hence, normalization will be required (see 5.2). This is standard in microarray analysis as can be seen from the example dataset displaying similar presence profiles (Figures 4.36 and 4.37).

4.7 Probe-level model fitting: Chip pseudo-images, RLE and NUSE - summarized data

In order to better visualize discrepancies between replicates on the individual arrays, pseudo-images are generated by fitting robust Probe Level linear Models (PLM) to all probesets in the data (e.g. to the summarized data see 5.3) [1]. Probe-level models consist of probe-level (assuming that each probe should behave the same on all arrays) and chip/array/sample-level (taking into account that a gene can have different expression levels in different samples) parameters. Probe level models assume that all probes of a probe group behave the same in the different samples (Package Oligo version 1.56.0). Probes that bind well to their target should do so on all arrays, probes that bind with low affinity should do so on all arrays. PLM implements iteratively re-weighted least squares M-estimation regression and the following model was considered:

$$y_{ijk} = \beta_{jk} + \alpha ik + \eta ijk \tag{4.6}$$
$$with\ i\ index\ for\ probes,\ j\ index\ for\ arrays\ and\ k\ index\ for\ probesets \tag{4.7}$$
$$with\ y_{ijk}:\ a\ (pre-processed\ if\ normalization\ and\ background\ correction\ applied) \tag{4.8}$$
$$probe\ intensity\ on\ log2\ scale \tag{4.9}$$
$$with\ \alpha ik:\ the\ probe\ effect\ parameter\ for\ probe\ i \tag{4.10}$$
$$with\ \beta_{jk}:\ the\ array\ (chip)\ effect \tag{4.11}$$
$$with\ \eta ijk:\ an\ error\ term \tag{4.12}$$

Two types of pseudo-images were created based on the residuals (Figure 4.38) respectively the weights (Figure 4.39) resulting from a comparison of the model (the ideal data, without any noise) to the actual data. These weights or residuals are graphically displayed using the image() function in Bioconductor. Weights represent how much the original data contribute to the model: outliers are strongly downweighted because they are largely different from the ideal data. Weights have values between 0 and 1. The smaller the weight of a probe, the more the probe is not showing the typical behavior that it shows on the other arrays and the more its intensity can be considered an outlier. Residuals represent the difference between the original data and the ideal data according to the model. So the more a residual deviates from 0, the larger the difference between the data and the model. Residuals can be positive indicating that the intensity of the probe is larger than the ideal value according to the model or negative indicating that the intensity of the probe is smaller than the ideal value according to the model. Residuals can be further formatted in different ways:

- type="resids": higher red intensities corresponding with higher positive residuals, white corresponding with residuals close to 0 and more intense blues corresponding with high negative residuals.
- type="pos.resids": Extreme positive residuals plotted in red, negative and near 0 residuals in orange
- type="neg.resids": Extreme negative residuals plotted in blue, positive, negative and near 0 residuals plotted in white
- type="sign.resids": all negative residuals regardless of magnitude are indicated by blue and all positive residuals in red

The artifacts observed in the microarray pictures of the raw log 2 probe intensities (Figure 4.1), are also visible in the residuals and weights plots. When a large number of outliers (affected probes in a large surface area) is present, ambiguity may arise when the regressive model cannot determine whether the artifact data are too high, or the non-artifact data are too low. Since the patches do not show up in the other replicate areas, the model likely correctly identified the artifacts (Figure 4.38- 4.39).

Other important QC metrics derived from PLM fitting are the Relative Log Expression (RLE) and the Normalized Unscaled Standard Errors (NUSE), displayed in box plots (Figure 4.40-4.41). The RLE is computed for each sample on every probeset by comparing the expression level of one probeset against the median expression of that probeset across samples. Assuming that most genes are not changing in expression across arrays, ideally most of these RLE values will be near 0. Typically arrays with poorer quality show up with boxes that are not centered around 0 and/or are more spread out. Boxes with larger whiskers indicate an unusually high deviation from the median in a lot of transcripts, suggesting that these arrays are different from most of the others in some way. Boxes that are shifted indicate a systematically higher or lower expression of the majority of transcripts in comparison to most of the other arrays. This could be caused by quality issues or batch effects [3]. Therefore, if shape and median of a given box varies too much from the bulk, they should be inspected and potentially removed.This is the case for blank3, SHIME_LGG1 and the Hela cell control (Figure 4.40).

To determine NUSE, the standard error estimates obtained for each probeset on each array from fitPLM are standardized across arrays so that the median standard error for that probe group is 1 across all arrays. This process accounts for differences in variability between genes. Arrays whose NUSE values are significantly higher than other samples are often lower quality chips. The interquartile ranges are slightly larger in the blank3,SHIME_LGG1 and Hela cell control arrays without normalization (Figure 4.41). Focusing on the control probes, the RLE values are near 0 and the NUSE values are rather small, with the largest variation again observed in the Hela cell control (Figure 4.38 and 4.39).

The RLE and NUSE values in the example data are similar, strengthening the confidence in the obtained data (Figures 4.48-4.51).

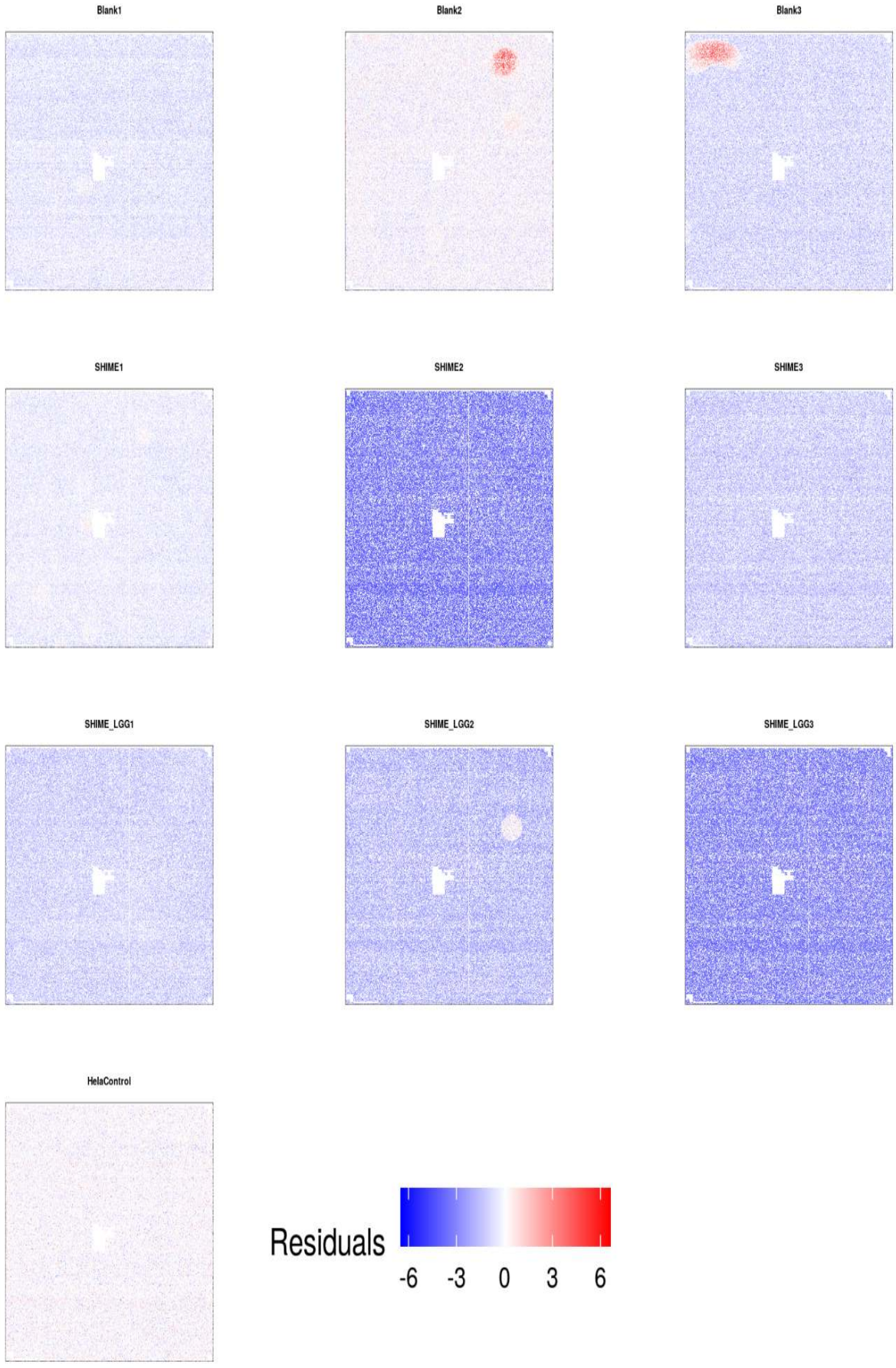

Figure 4.38: Pseudo-images of the estimated residuals from a probe-level model fitting

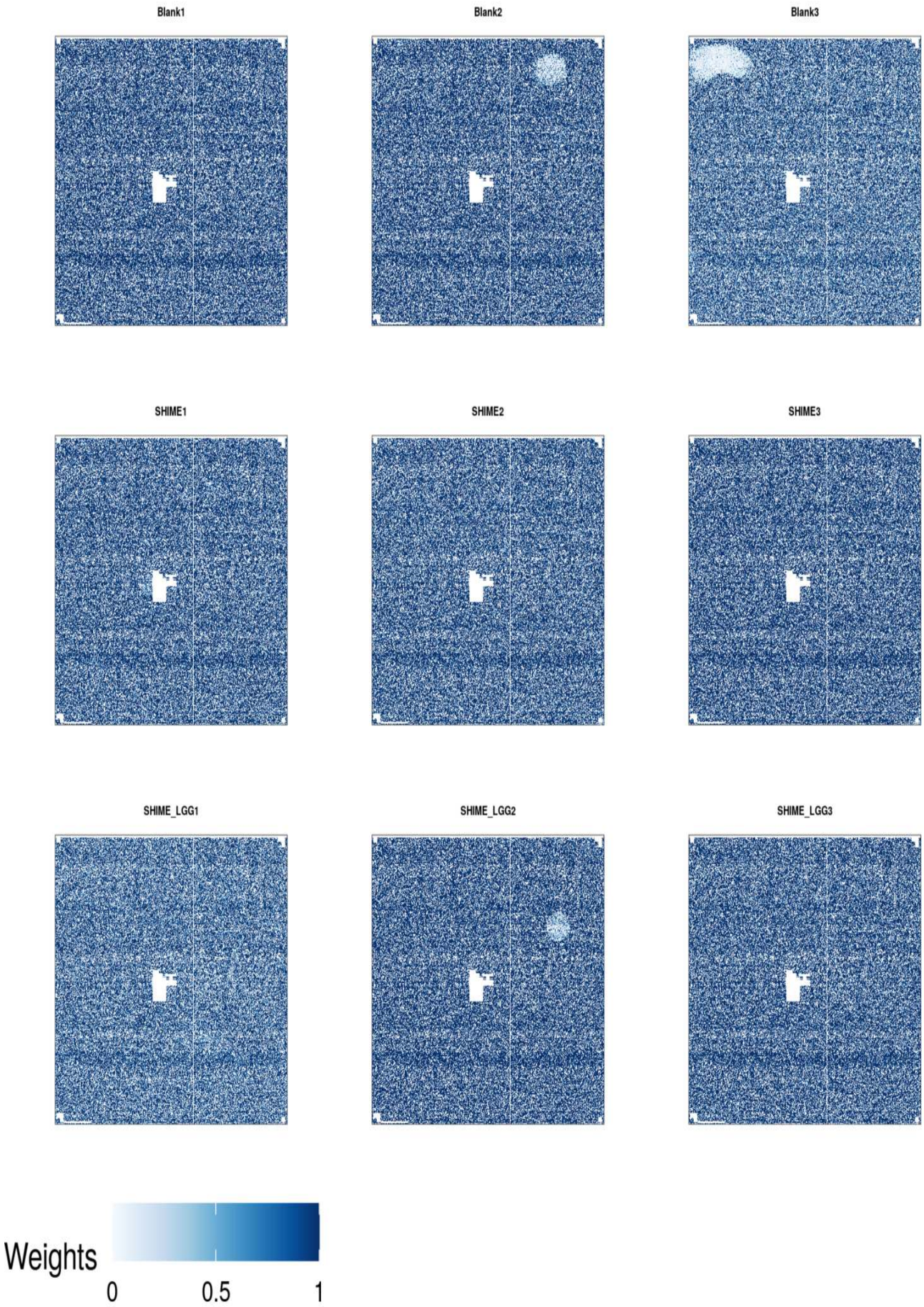

Figure 4.39: Pseudo-images of the estimated weights from a probe-level model fitting

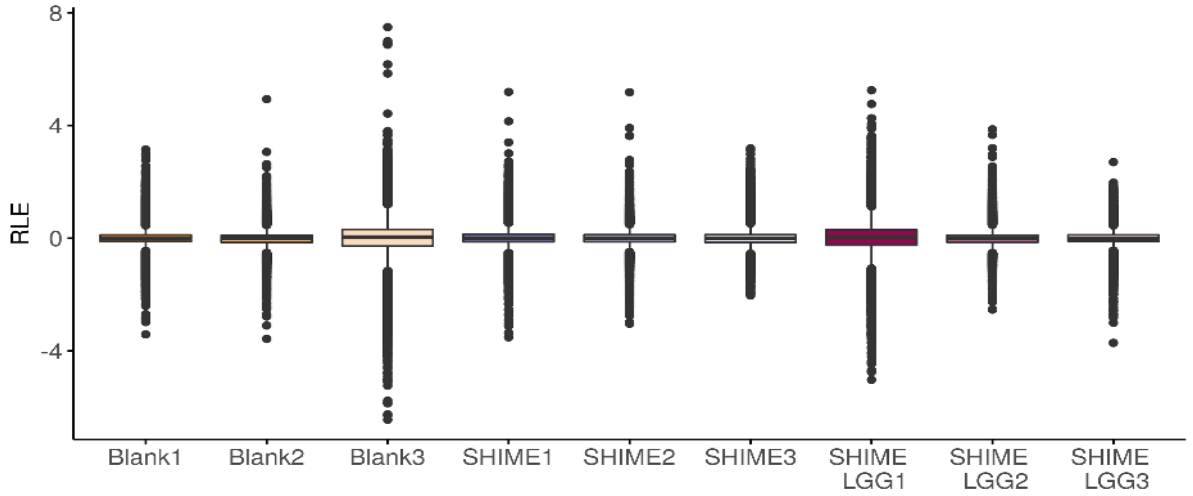

Figure 4.40: Box plot of the Relative Log Expression (RLE) values based on a probe-level model fitting

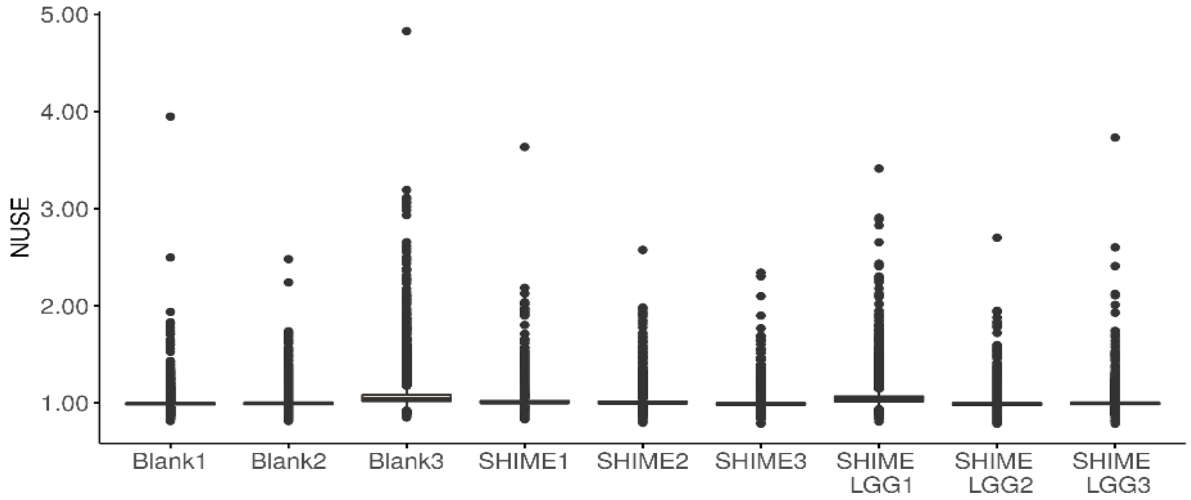

Figure 4.41: Box plot of the Normalized Unscaled Standard Errors (NUSE) values based on a probe-level model fitting

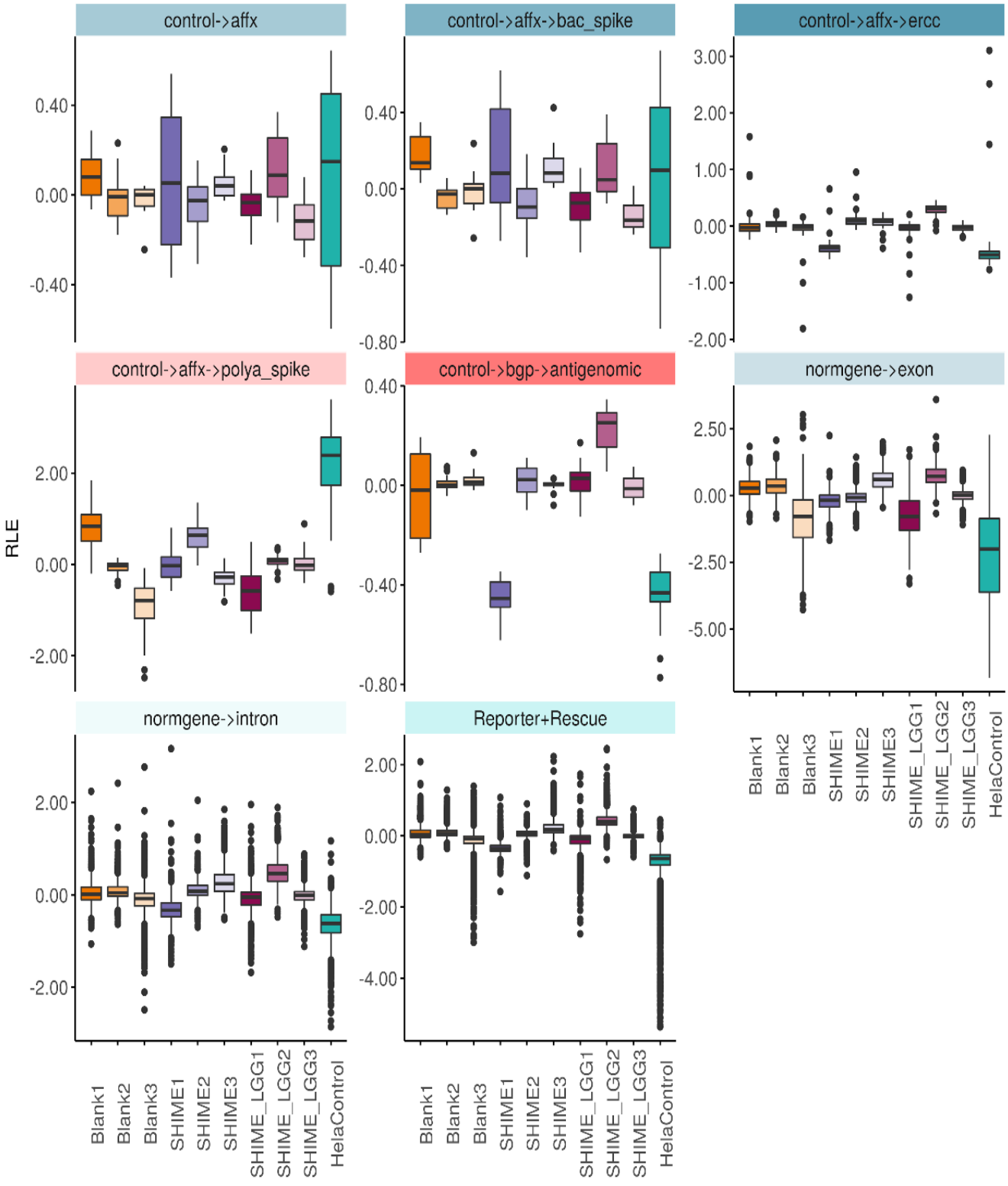

Figure 4.42: Box plot of the Relative Log Expression (RLE) values of control probes based on a probe-level model fitting

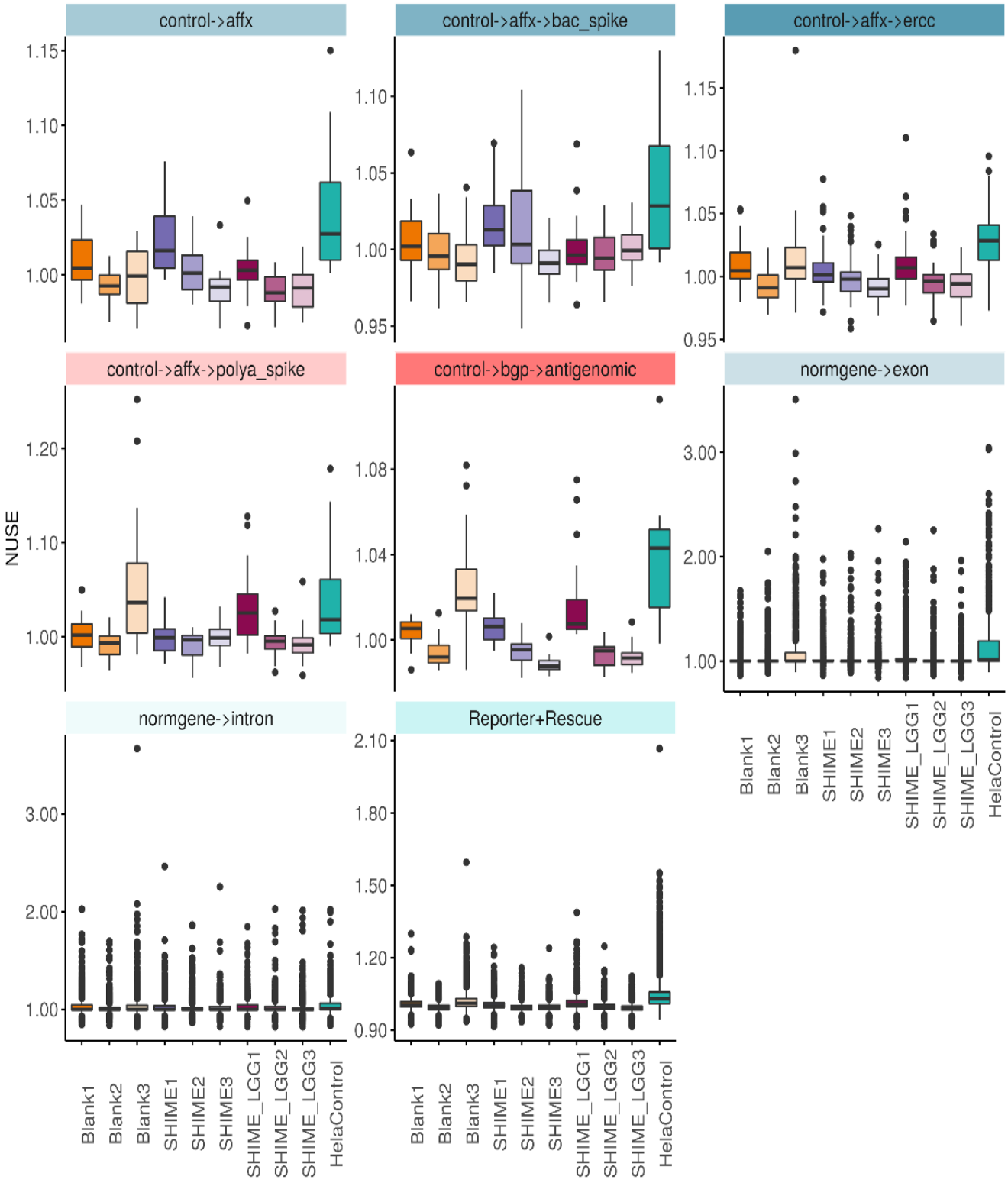

Figure 4.43: Box plot of the Normalized Unscaled Standard Errors (NUSE) values of control probes based on a probe-level model fitting

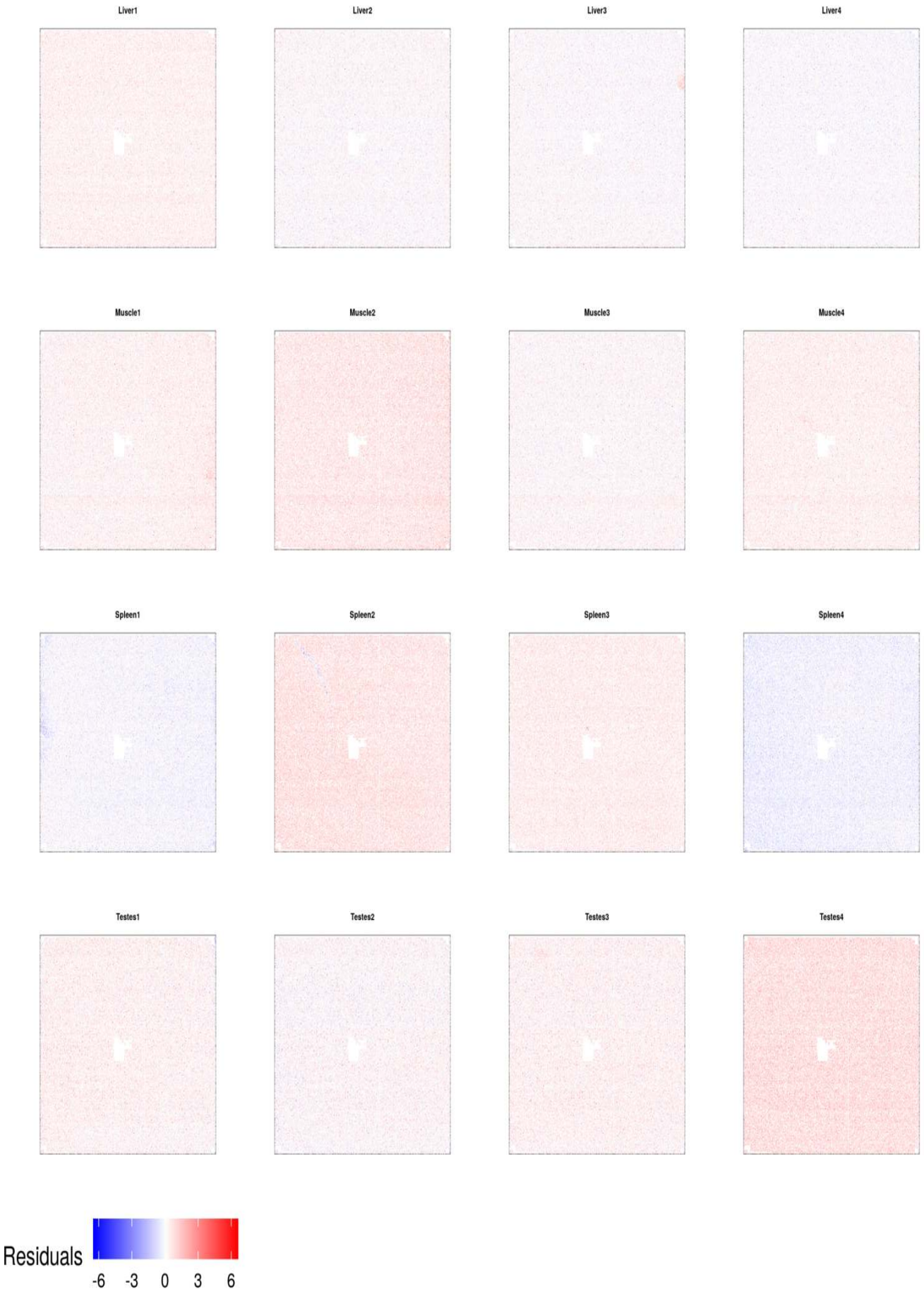

Figure 4.44: Pseudo-images of the estimated residuals from a probe-level model fitting of the tissue example data.

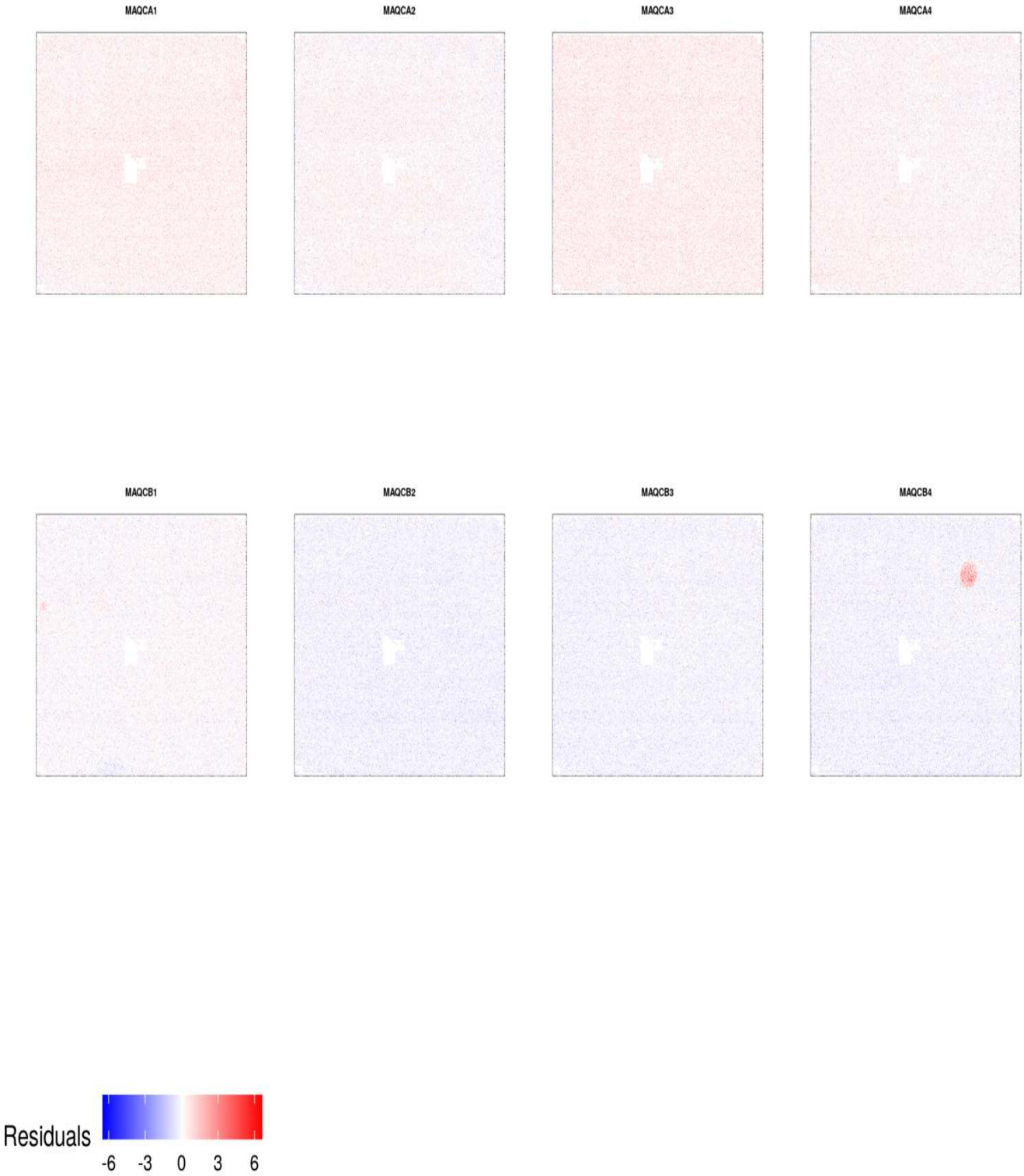

Figure 4.45: Pseudo-images of the estimated residuals from a probe-level model fitting of the MAQC example data.

Figure 4.46: Pseudo-images of the estimated weights from a probe-level model fitting of the tissue example data.

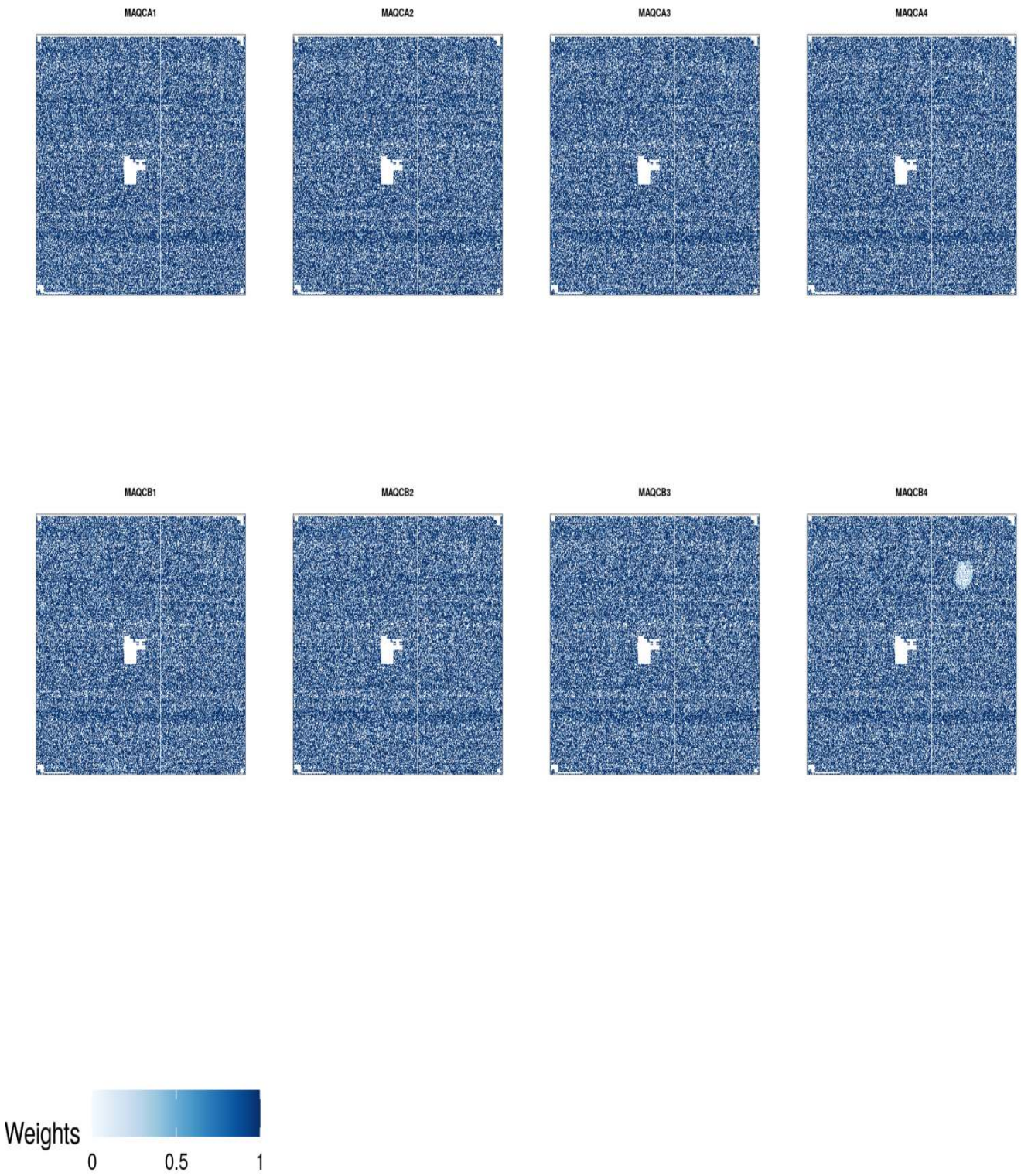

Figure 4.47: Pseudo-images of the estimated weights from a probe-level model fitting of the MAQC example data.

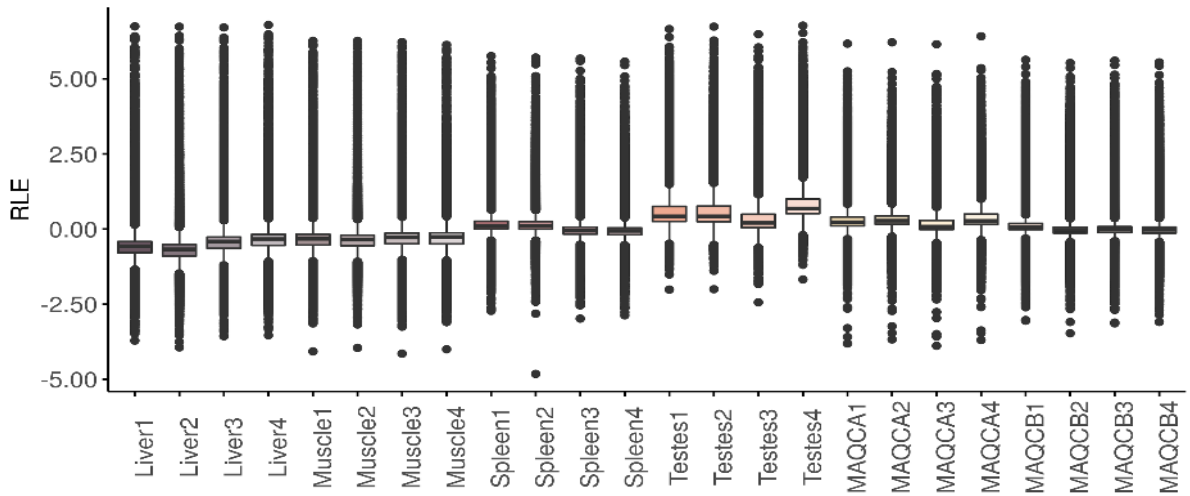

Figure 4.48: Box plot of the Relative Log Expression (RLE) values based on a probe-level model fitting of the example data.

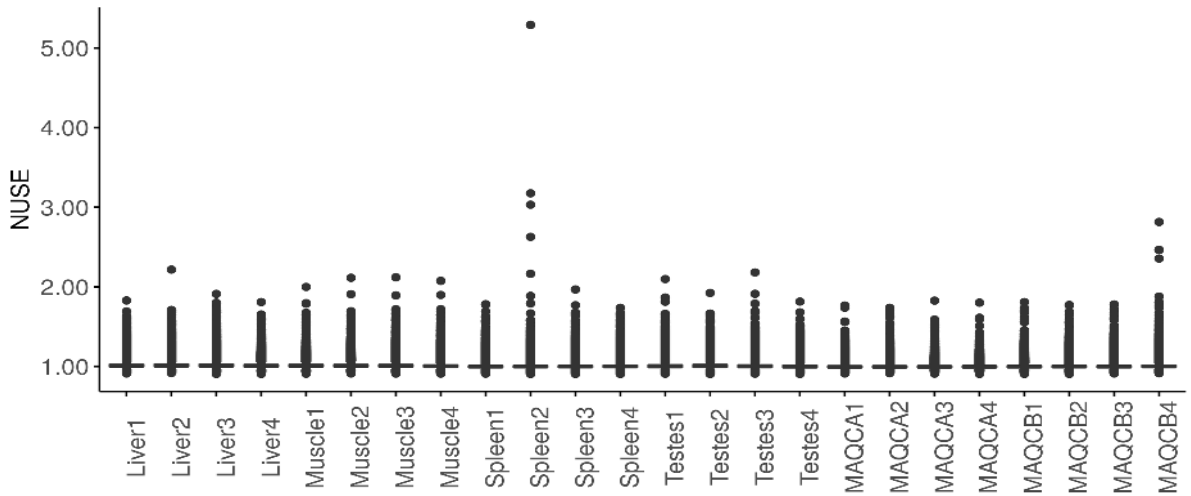

Figure 4.49: Box plot of the Normalized Unscaled Standard Errors (NUSE) values based on a probe-level model fitting of the example data.

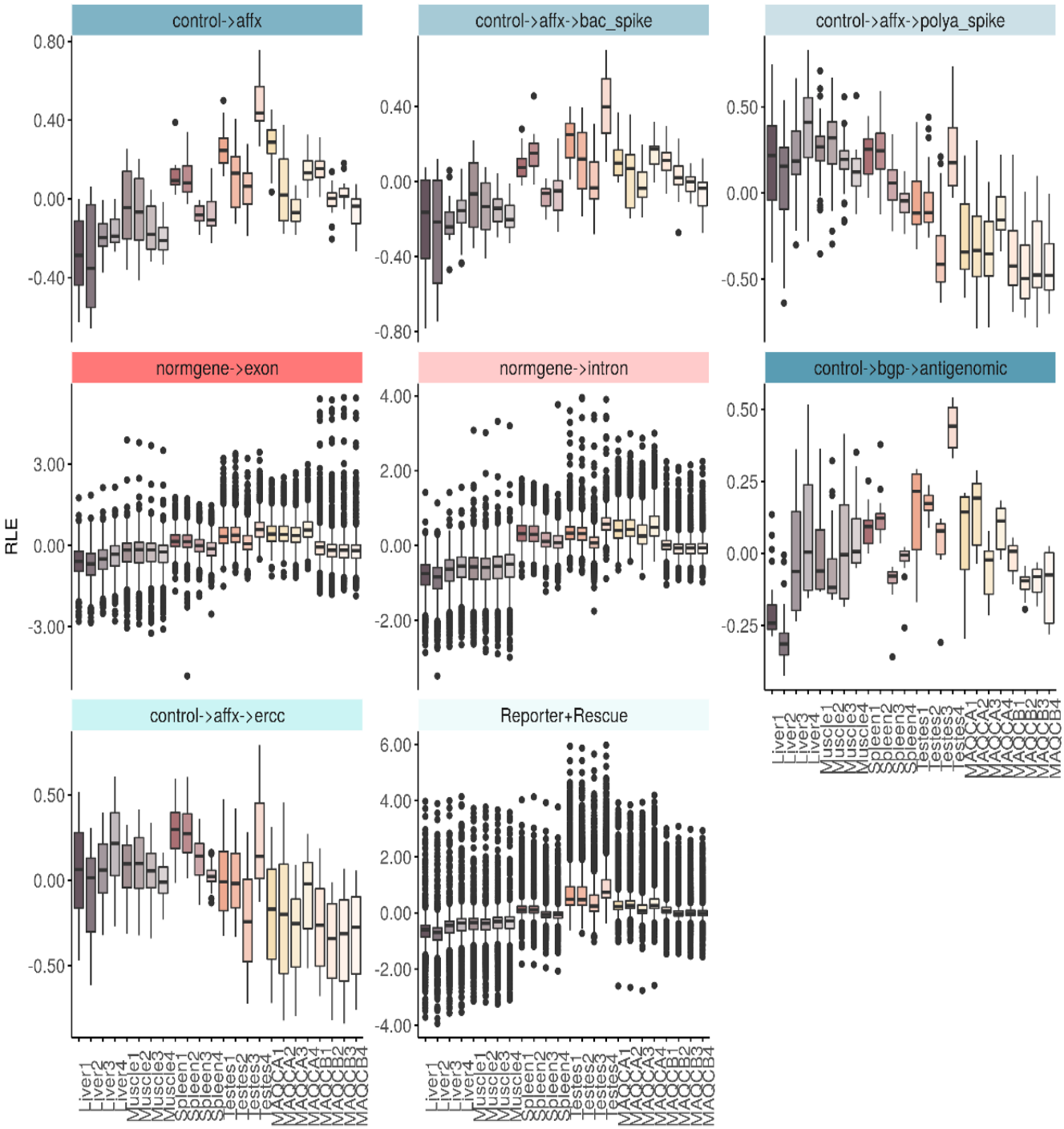

Figure 4.50: Box plot of the Relative Log Expression (RLE) values of control probes based on a probe-level model fitting of the example data.

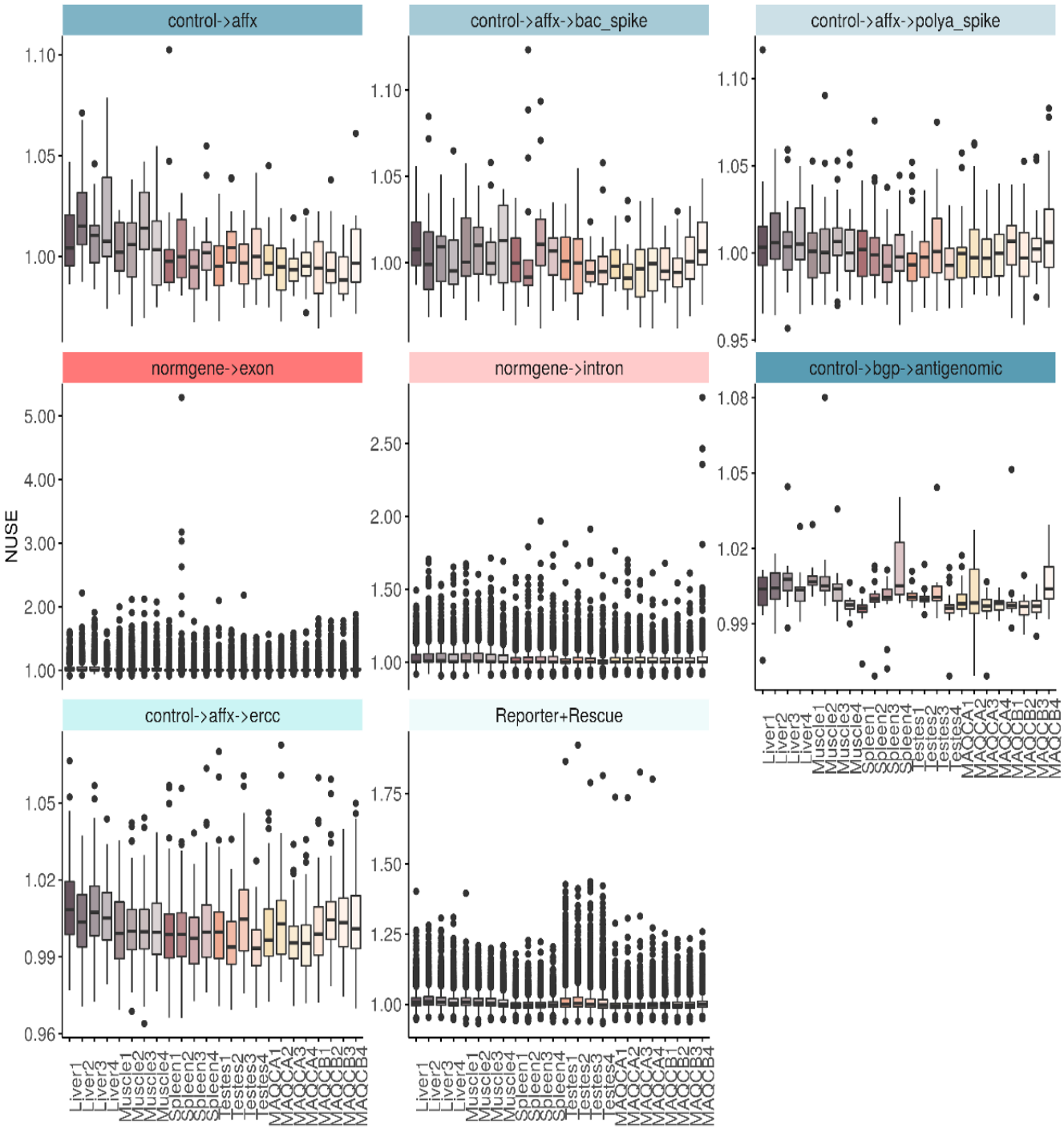

Figure 4.51: Box plot of the Normalized Unscaled Standard Errors (NUSE) values of control probes based on a probe-level model fitting of the example data.

4.8 PCA

Finally, after assessing the distributional assumptions, Principle Component Analysis (PCA) was performed and confidence ellipses were constructed to identify outliers based on the Hotelling's T2 test statistic in the unnormalized log2 transformed standardized data. Besides a PCA model based on all probes, subsets of the control probes were explored in a series of PCAs to identify specific steps in the experimental process that may have resulted in aberrant expression patterns. While no outlier samples were detected based on the Hotelling's T2 criterium, the Hela cell control is clearly separated from the coculture model samples in all PCA models and will be omitted during downstream pre-processing. Blank3 and SHIME_LGG1 also seem to cluster separately in all, but the bac PCA. This indicates that hybridization went well, but anomalies may have occurred during sample preparation. This is not the case for sample SHIME1 displaying a diverging pattern in all PCAs except for polyA, suggesting that a deviation in starting RNA yield/quality may underlie the observed deviating intensity signals of the SHIME1 microarray. These moderately deviating microarrays (Blank3, SHIME1 and SHIME_LGG1) will be re-evaluated after normalization 4.52 (see 6). Outliers were also observed in the example dataset 4.53.

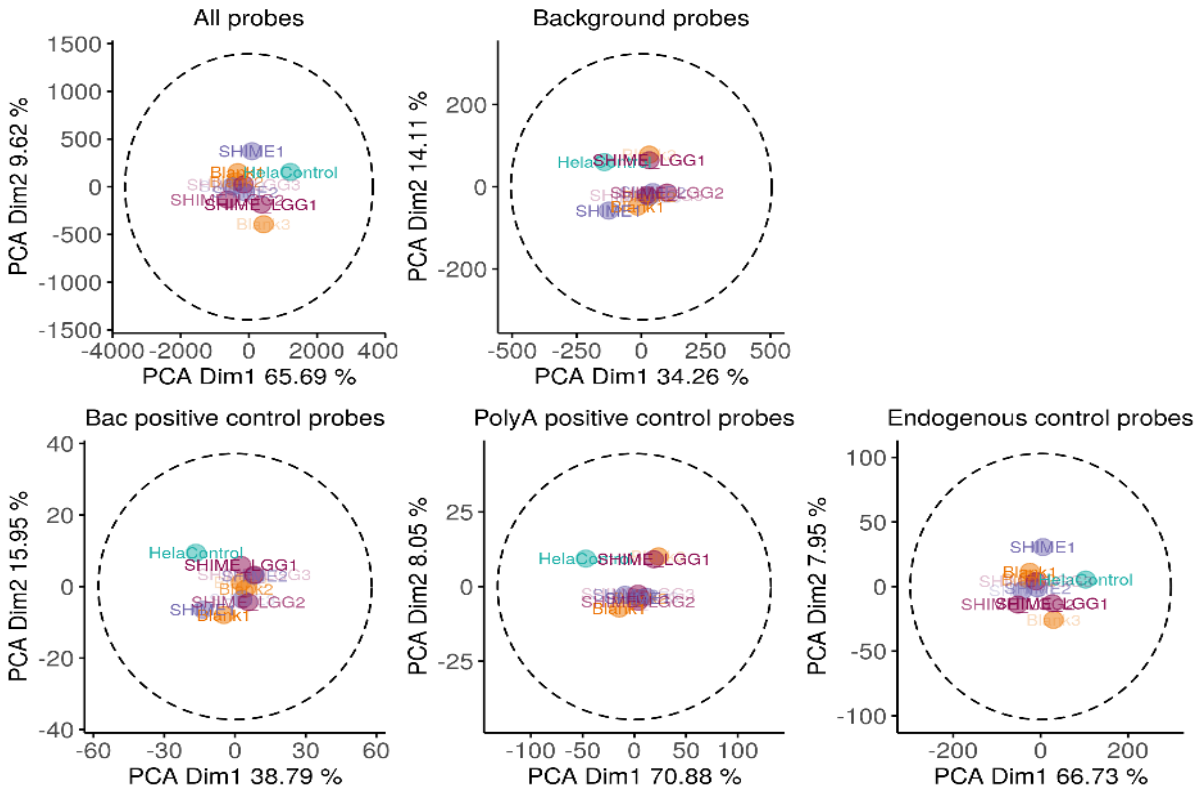

Figure 4.52: PCA scores plot of the log2 transformed standardized data.

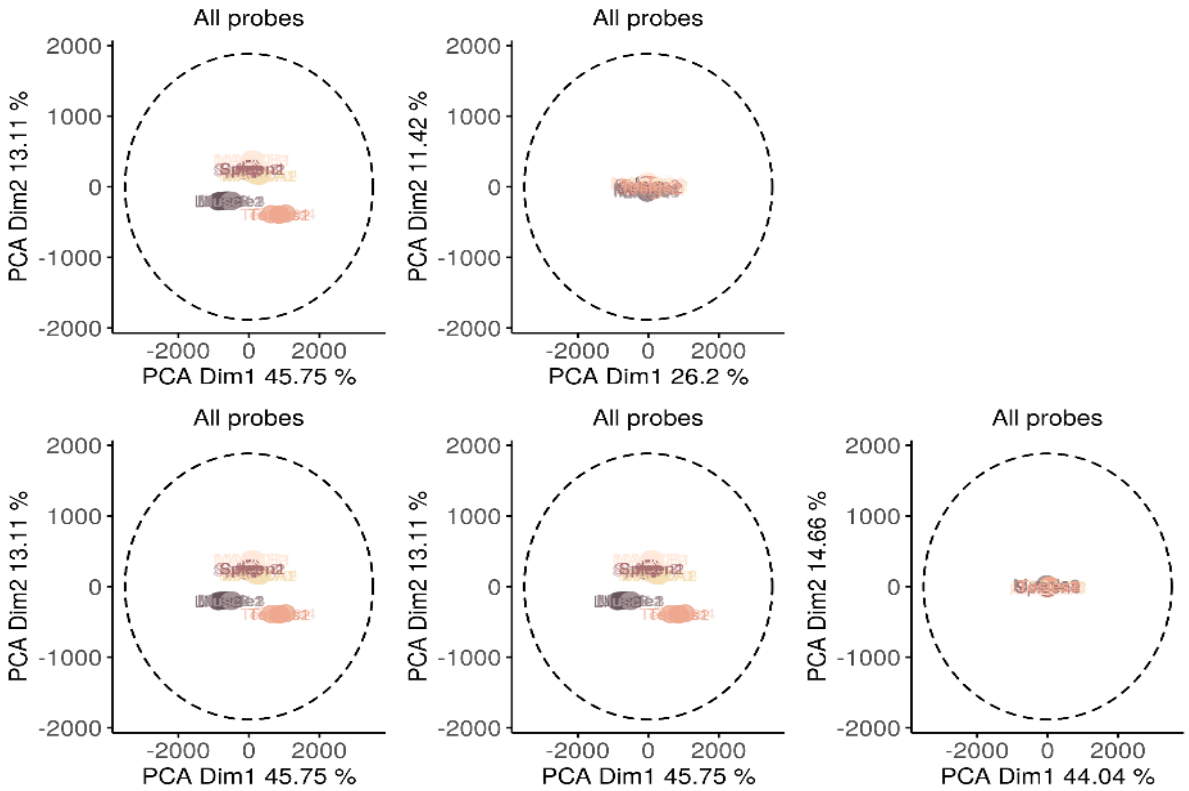

Figure 4.53: PCA scores plot of the log2 transformed standardized example data.

4.9 arrayQualityMetrics tool results

The arrayQualityMatrics tool confirmed our conclusions and additionally indicated the Hela cell control array as an outlier based on 2 different criteria (exceptionally large sum of the distances to all other arrays and a Kolmogorov-Smirnov test statistic pointing at a significantly different array distribution compared to the distribution of the pooled data).

Chapter 5

Pre-processing: Background correction, Normalization, Summarization and RMA

There are several sources of variability (noise) in microarray experiments requiring background correction and normalization and data from individual probes need to be summarized at gene level. Each of the 3 pre-processing steps can be performed separately and sequentially in a modular approach. While each step may be optimal, the combination is not guaranteed to be optimal and errors are not propagated through the sequence of steps. Alternatively, the 3 pre-processing steps can be performed simultaneously to obtain an optimal final result.

Popular data pre-processing implementations are MAS5 and RMA. MAS are Microarray Suite is software that comes with the Affymetrix Genechips. RMA or Robust Multi-array Average consists of convolution background correction, quantile normalization and summarization based on a multi-array model fit using robust median-polish for estimation.

5.1 Background correction

Within microarrays background correction is required to deal with:

- Optical noise
- Cross-hybridization (= non-specific hybridization): labeled target binds to non-complementary probe resulting in measurable intensities
- Free biotin may attach to the array, even after washing, resulting in measurable intensities
- Differences in probe affinities: G-C binding occurs through 3 H-bonds and is stronger compared to A-T pairing with 2 H-bonds. Probes with a large GC content may hence display a larger intensity. The relation between intensities and mRNA concentrations is confounded by the probe-affinity

Background correction aims to improve the linear relationship between the log intensities and the concentration of mRNA (cDNA). In general a background corrected intensity of probe i is given by:

$$X_i = Y_i - Bi \tag{5.1}$$
$$with\ X_i\ the\ background\ corrected,\ true\ intensity\ of\ probe\ i \tag{5.2}$$
$$with\ Y_i\ the\ measured\ intensity\ for\ probe\ i \tag{5.3}$$
$$with\ B_i\ the\ estimated\ background\ intensity\ for\ probe\ i \tag{5.4}$$

The background correction method currently available in the oligo package is the one used in RMA, which treats the PM intensities as a convolution of noise and true signal. While the noise and true signal are not known at probe-level, we can make distribution assumptions about X and B. X follows an exponential distribution and B a normal distribution. Parameters of these distributions are estimated ad-hoc by fitting distributions to the observed intensities (Figures 4.4 and 5.1).

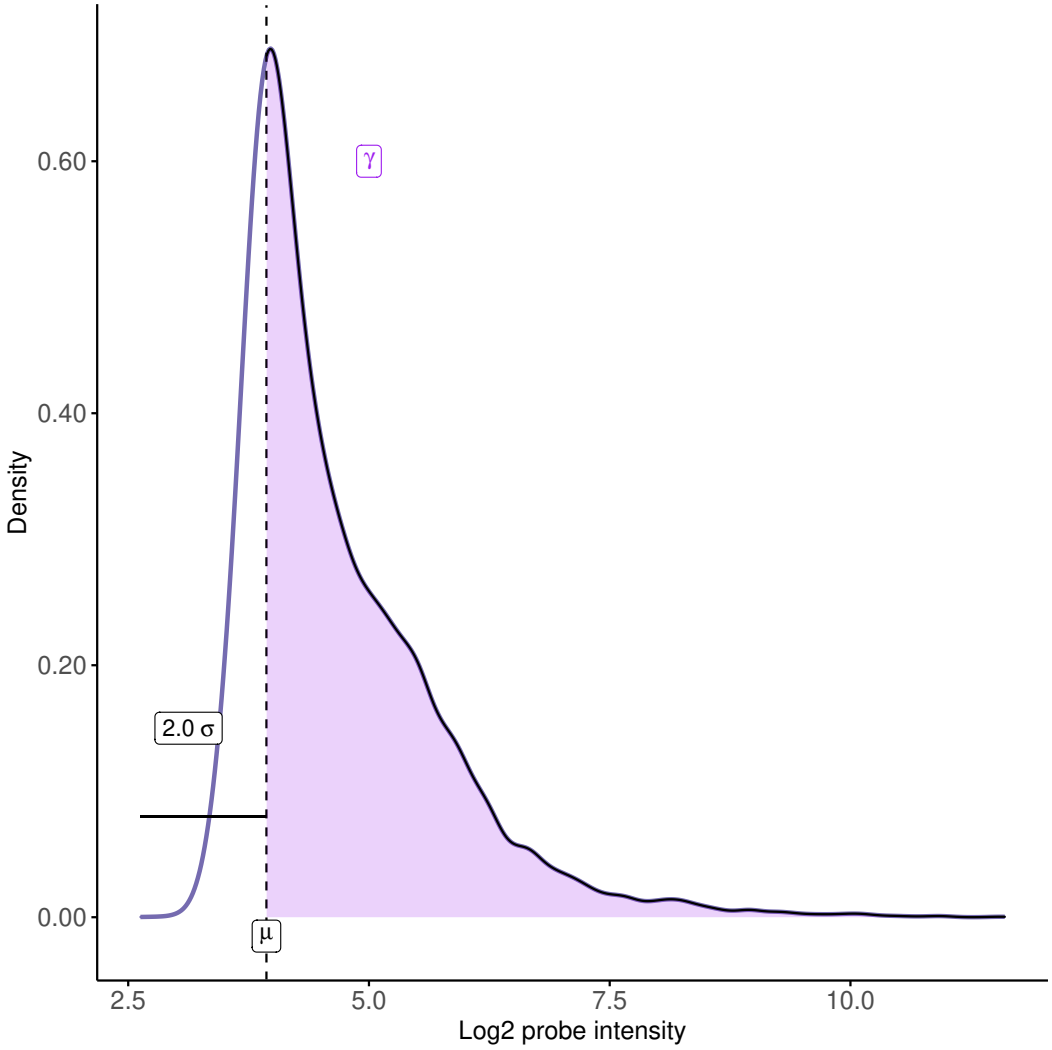

Figure 5.1: The parameters of the normal and exponential distribution used to deconvolute the probe intensity signals are estimated ad-hoc from the observed probe intensity distributions. μ is estimated as the mode, σ^2 is estimated as twice the variance in the lower tail of the distribution and γ is estimated by fitting an expenonential distribution to the right tail of the observed density.

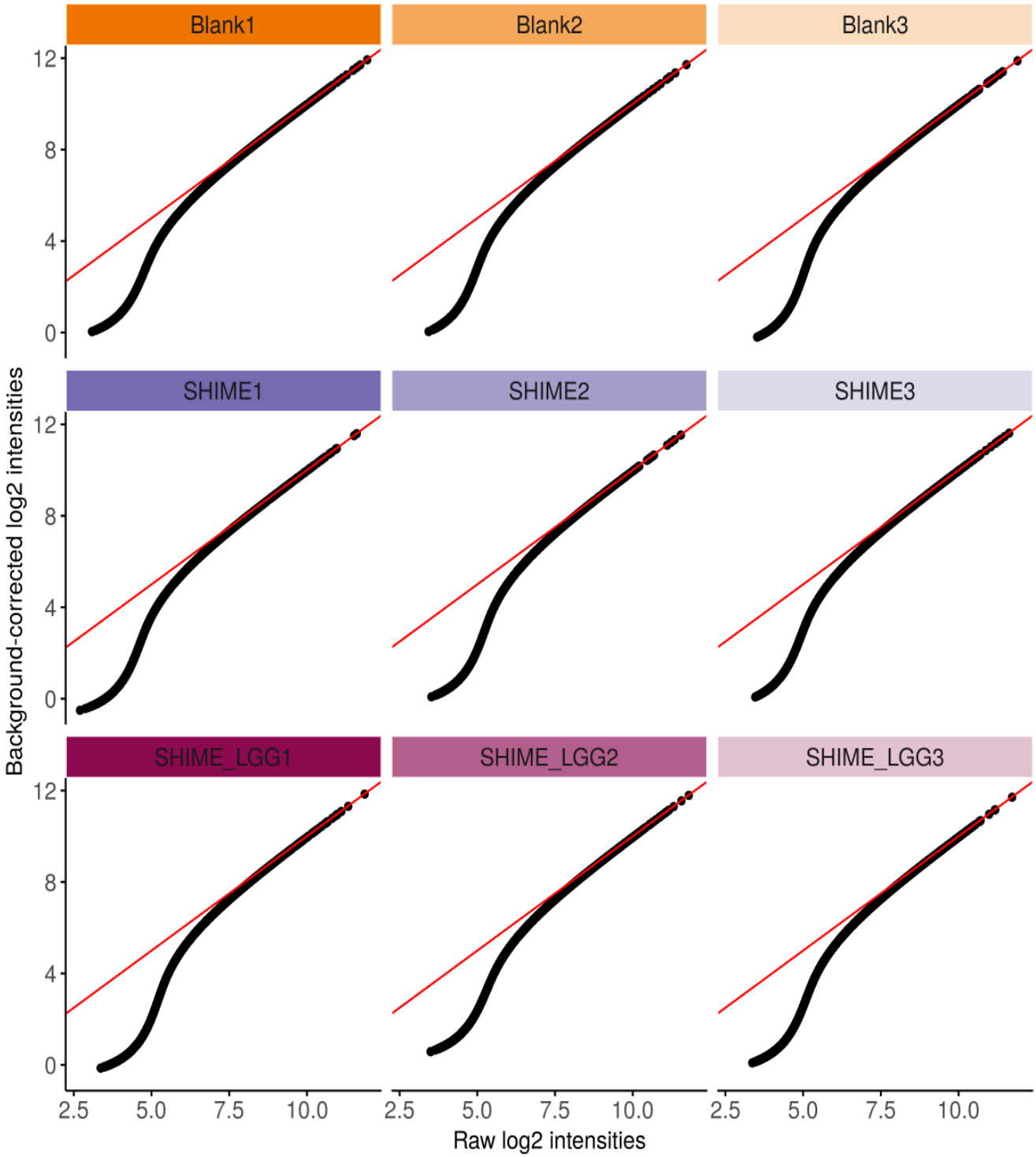

Figure 5.2: Biplot of the raw and background corrected intensities.

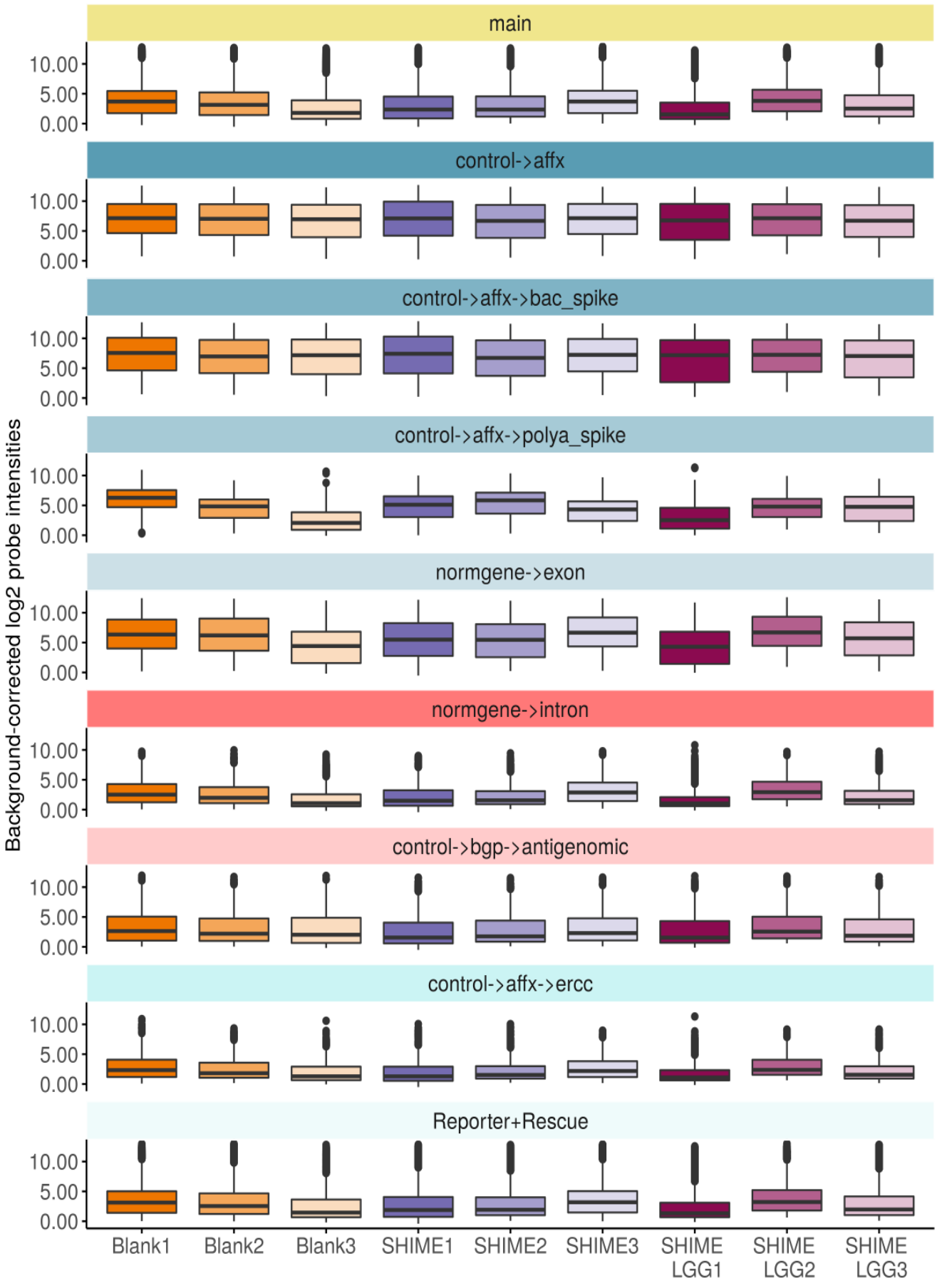

Figure 5.3: Boxplot of the background-corrected log2 intensities

If there would be no difference between raw and background corrected data, the data points should end up on the diagonal. This is of course not the case (Figure 5.2). Low intensities are strongly affected by the background subtraction since the background intensity is a small value and subtracting a small value from a small value has a much bigger impact than subtracting a small value from a high value.

5.2 Normalization

Normalization across arrays is required to deal with:

- Technical variation:
  - Imperfections on the array surface
  - Imperfect synthesis of the probes (probe fixation) during the array production process
  - Differences in hybridization conditions
  - Different scanning conditions
- Biological variation:
  - Different amounts of mRNA used for labeling and hybridization
  - Different efficiencies of reverse transcription, labeling during sample preparation or hybridization reactions
  - Reagent batch differences
  - Different laboratory conditions

Normalization removes systematic differences between the samples that are due to noise rather than true biological variability in order to make biologically meaningful conclusions about the data.

Several normalization methods are available. Quantile normalization is a commonly used method assuming the same probe level intensity distribution across chips. Basically, data points are projected onto the diagonal in a QQ-plot of the intensities of the arrays being compared (Figures 5.4). All arrays can be compared to one baseline array (2D), which is selected by the algorithm to represent a typical array or all arrays can be compared simultaneously (n dimensions). Note that normalization is generally performed at probe level prior to summarization.

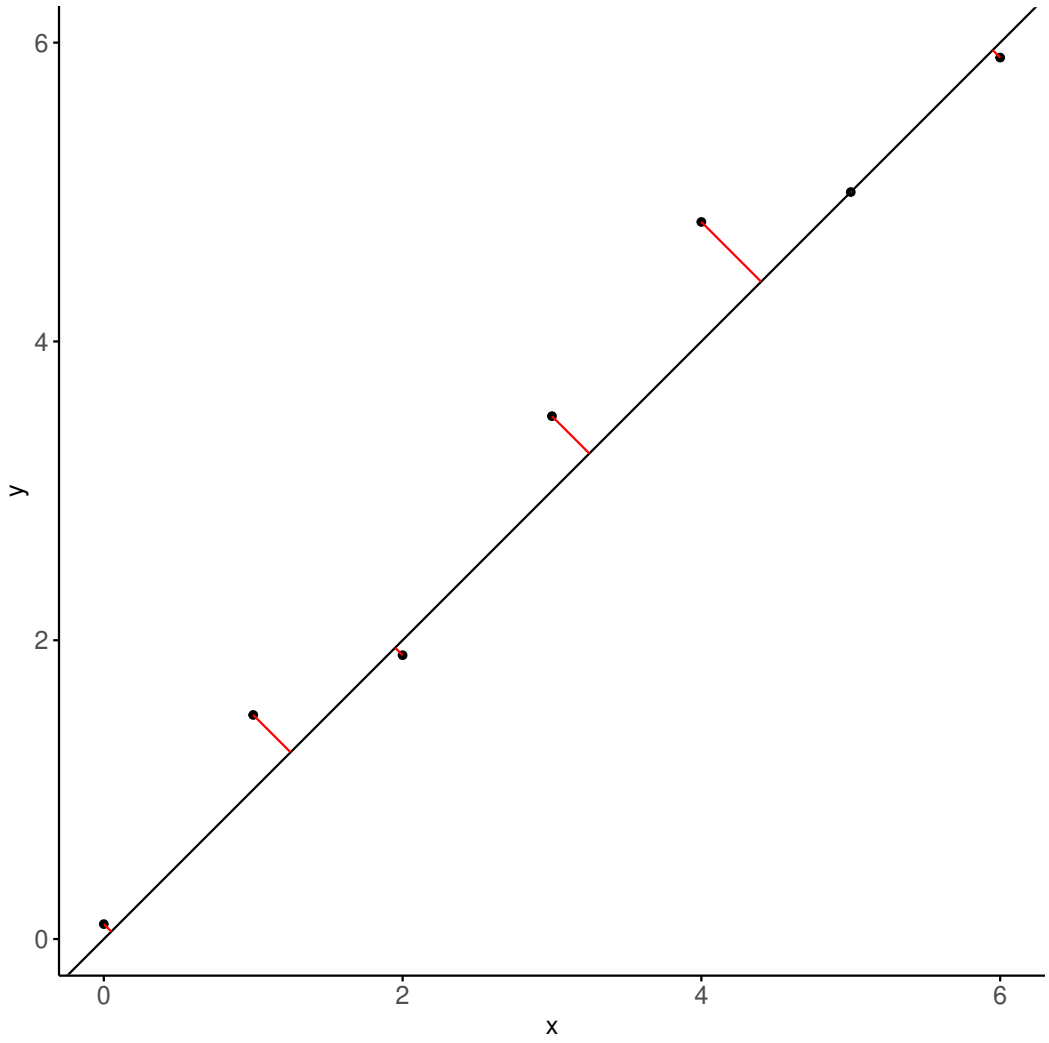

Figure 5.4: The principle of Quantile Normalization in 2D.

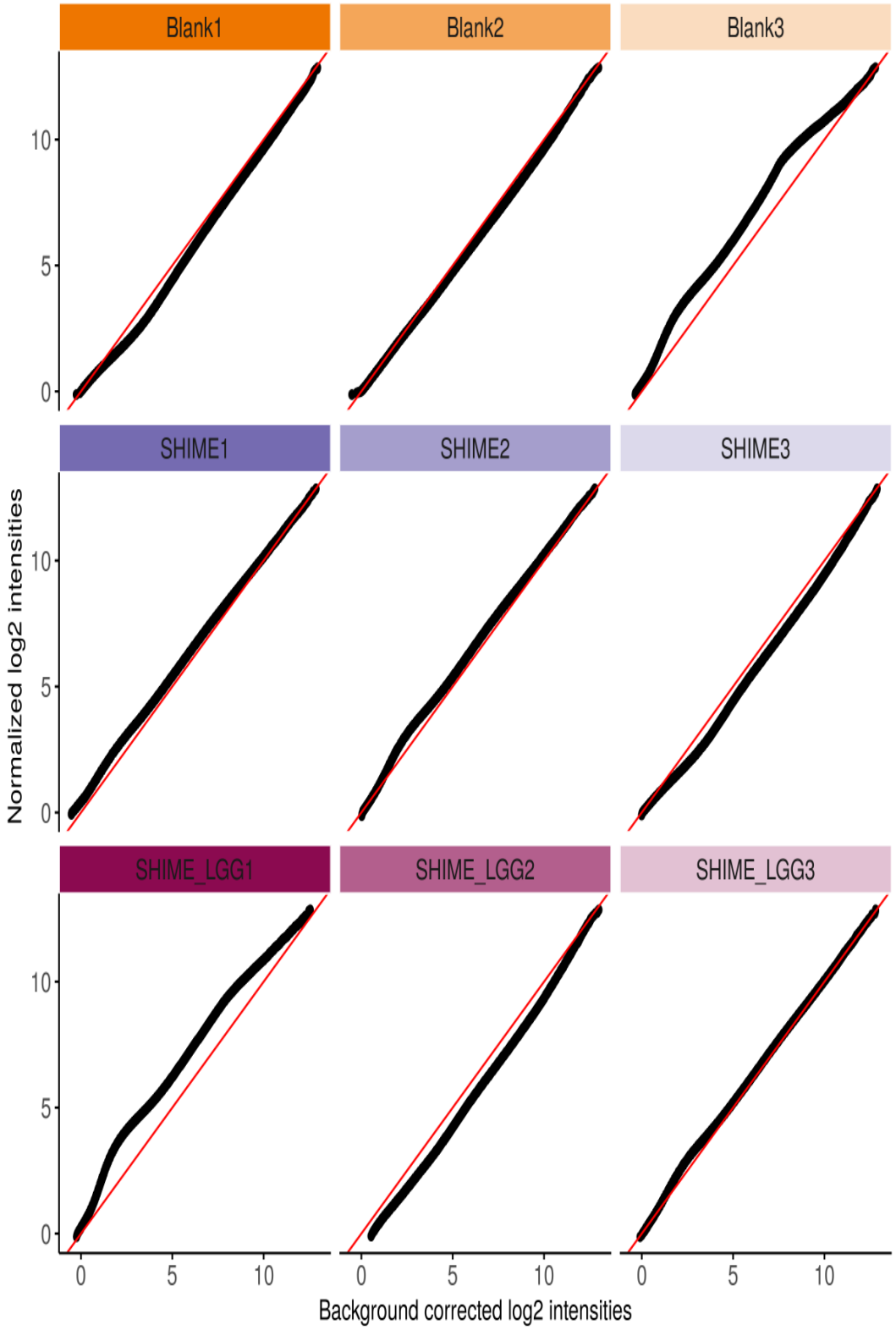

Figure 5.5: Biplot of the background-corrected and normalized intensities.

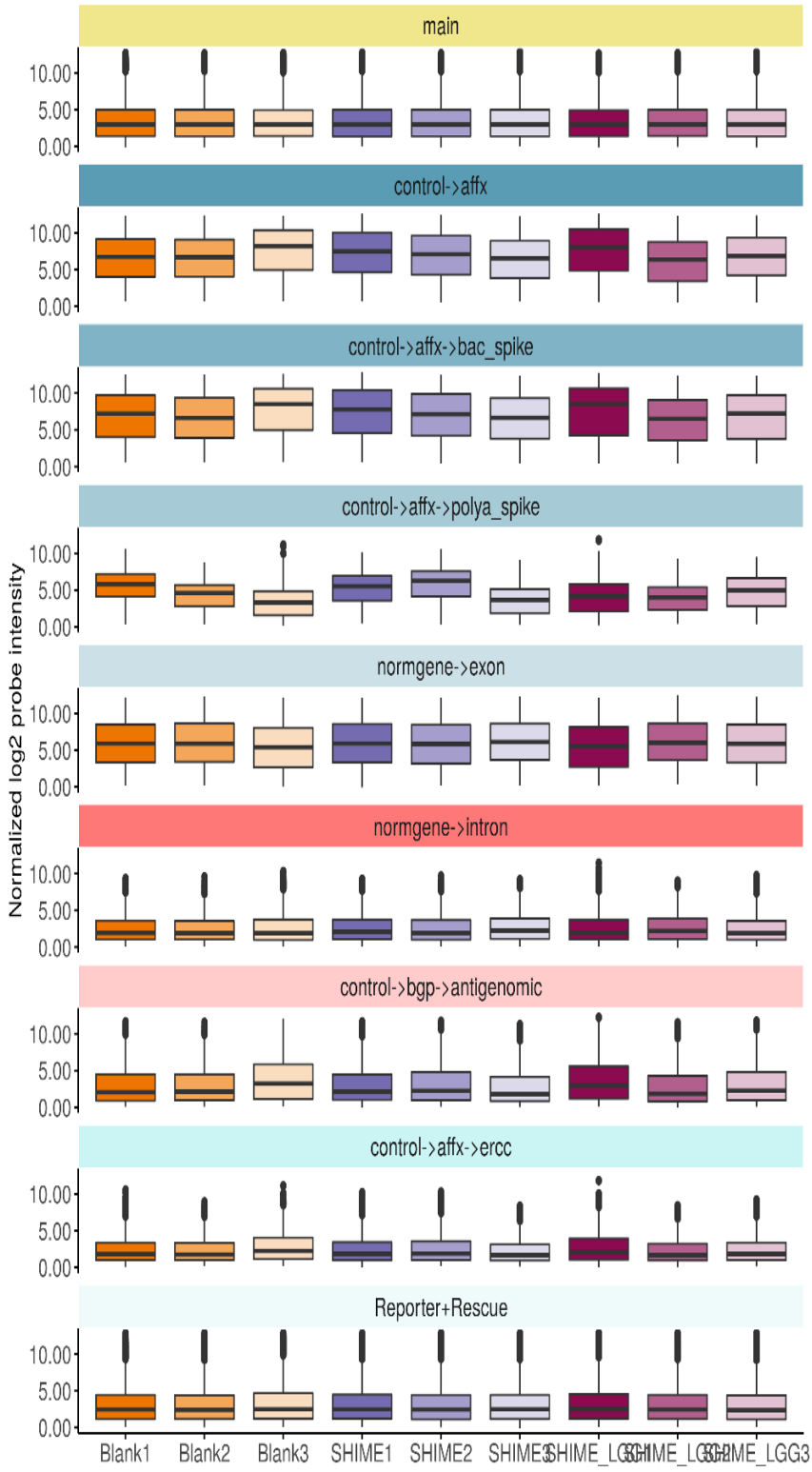

Figure 5.6: Boxplot of the normalized log2 intensities

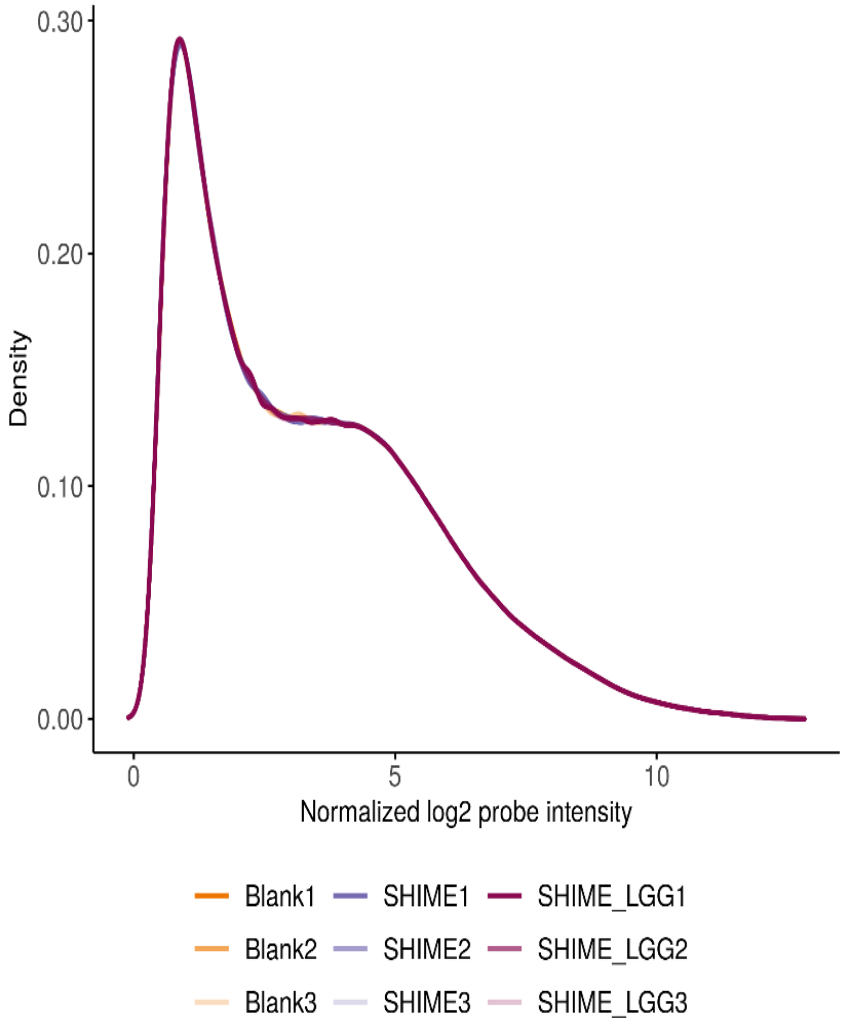

Figure 5.7: Histogram of the normalized log2 intensities

After normalization, none of the samples should stand out from the rest. The different arrays should have the same (or at least a very comparable) median expression level. Also the scale of the boxes should be almost the same indicating that also the spread of the intensity values on the different arrays is comparable (Figures 5.5-5.7).

5.3 Summarization

In the pre-processing step, data is summarized at the gene/transcript level (option target = core) or probeset level. This is necessary for downstream analysis because there are multiple probes per probeset. The summary measure is a proxy for the transcript concentration or the expression value of a gene.

Since the Human Gene 2.1 ST microarrays were created as an affordable version of the 'Exon' arrays by reducing the number of probes keeping only the 'good' probes from the Exon arrays, probesets or exons are represented by only a few probes and summarization on this level is not recommended [3].

Several single- and multi-chip methods exist, including the robust average (MAS), robust es-

timation, median polish (RMA) and plm. The robust average method calculates the average log2 intensity for all probes within a probeset excluding outlier probes. However, the parallel behavior in probe response across arrays and the relationship between concentration and expression level on each array (Figure 4.22) motivates the use of multi-chip models with probe and chip response parameters. RMA makes use of such advanced additive regression models. Since there are often probes on individual arrays that behave discordantly due to non-biological causes, it is advantageous to fit the model robustly and use median polish instead of common linear regression. This is done by using the median polish algorithm to fit the more general model displayed in equation 5.5 [1]. See `https://mgimond.github.io/ES218/Week11a.html` for more information on the median polish algorithm. Alternatively, PLM described above (4.6) can be used to summarize the data.

$$y_{ijk} = \mu_k + jk + \alpha ik + \eta ijk \tag{5.5}$$
$$with\ i\ index\ for\ probes,\ j\ index\ for\ arrays\ and\ k\ index\ for\ probesets \tag{5.6}$$
$$with\ y_{ijk}:\ a\ (pre-processed\ if\ normalization\ and\ background\ correction \tag{5.7}$$
$$applied\ according\ to\ RMA)\ probe\ intensity\ on\ log2\ scale \tag{5.8}$$
$$with\ \mu_k:\ the\ overall\ expression\ value \tag{5.9}$$
$$with\ jk:\ the\ array\ effect\ for\ array\ j\ (column\ effect\ in\ median\ polish) \tag{5.10}$$
$$with\ \alpha ik:\ the\ probe\ effect\ parameter\ for\ probe\ i\ (row\ effect\ in\ median\ polish) \tag{5.11}$$
$$with\ \eta ijk:\ an\ error\ or\ residual\ term \tag{5.12}$$
$$with\ constraints:\ median(\alpha ik)\ =\ median(jk)\ =\ median(\eta ijk) \tag{5.13}$$
$$where\ \beta_{jk}\ = \mu_k + jk \tag{5.14}$$
$$with\ \beta_{jk}:\ the\ log2\ expressions \tag{5.15}$$
$$\tag{5.16}$$

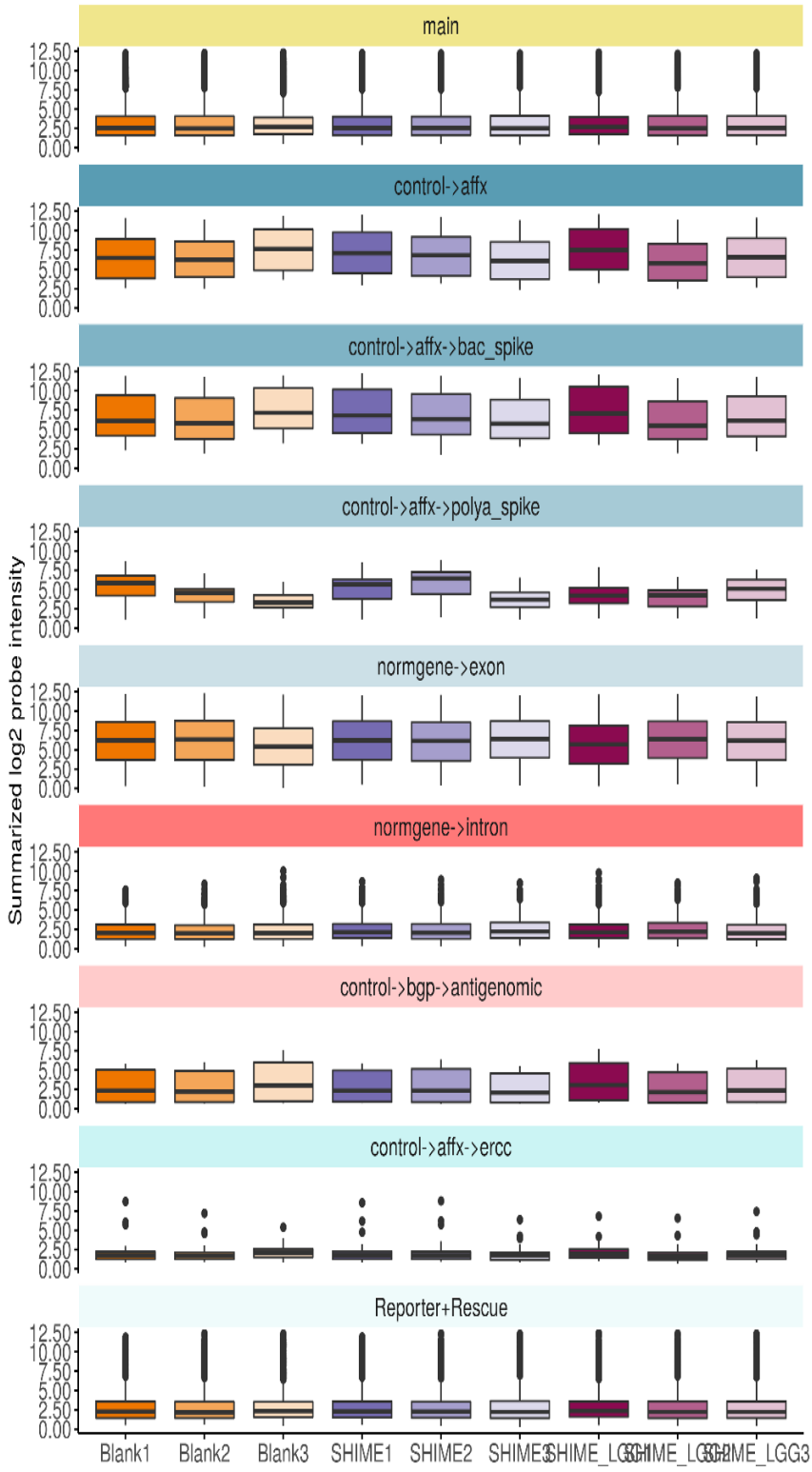

Figure 5.8: Boxplot of the summarized log2 intensities.

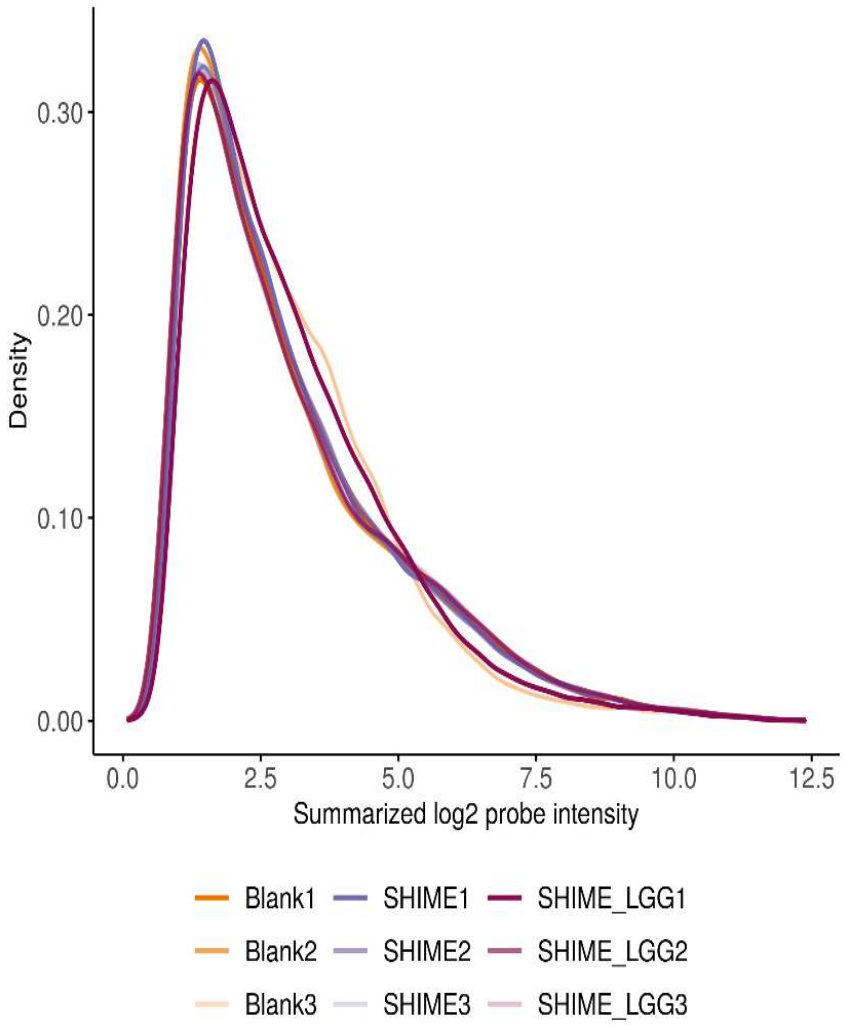

Figure 5.9: Density plot of the summarized log2 intensities.

5.4 RMA

As mentioned before background subtraction, quantile normalization and summarization via median polish can be combined through RMA. RMA was performed at the gene level (target = core) and after omission of the Hela cell control sample.

The RMA algorithm performs

- Background correction to correct for spatial variation within individual arrays: a background-corrected intensity is calculated for each PM probe in such a way that all background corrected intensities are positive
- Log transformation to improve the distribution of the data: the base-2 logarithm of the background corrected intensity is calculated for each probe. The log transformation will make the data less skewed and more normally distributed and provide an equal spread of up- and down-regulated expression ratios
- Quantile normalization to correct for variation between the arrays: equalizes the data distributions of the arrays and make the samples completely comparable
- Probe normalization to correct for variation within probesets: equalizes the behavior of the probes between the arrays and combines normalized data values of probes from a probeset into a single value for the whole probeset

Chapter 6

Quality Control of the RMA pre-processed data

Quality was inspected after RMA processing with the same QC tools as before (array images, box plots, density plots, model fit residual/weight plots, MA plots, RLE, NUSE, DABG, PCA).

6.1 Distribution of the RMA pre-processed intensities

After normalization, the distribution of the samples should be homogeneous. The different arrays should have a very comparable median expression level and the scale of the boxes should be very comparable indicating that the spread of the intensity values on the different arrays is equalized. Samples blank3 and SHIME_LGG1 still stand out from the other samples (Figures 6.1-6.2), displaying a slightly higher median expression and smaller standard deviations. Most of the background probes shifted to lower expression values after RMA pre-processing, but some probes retained higher signals resulting in a bimodal distribution of the background probes post-RMA processing. The blank3 and SHIME_LGG1 samples have higher medians and spreads, and therefore do not align well with the other samples (Figures 6.3-6.4). The reduced background signal also affected the negative control samples and is reflected in decreased median signal intensities (Figures 6.5 and 6.6). While the endogenous and bac spike positive controls have comparable distributions with the exception of samples blank3 and SHIME_LGG1, the polyA spike still shows a high level of between-array variability (also within replicate arrays). Different ratios of the spike-in compared to the input RNA amount may explain the observed differences. The R^2 of the linear regression model of the RMA pre-processed signal intensities in function of the spike-in concentration improved for all microarrays (Figures 6.7-6.10).

6.1.1 Distribution of the RMA pre-processed all-transcript intensities

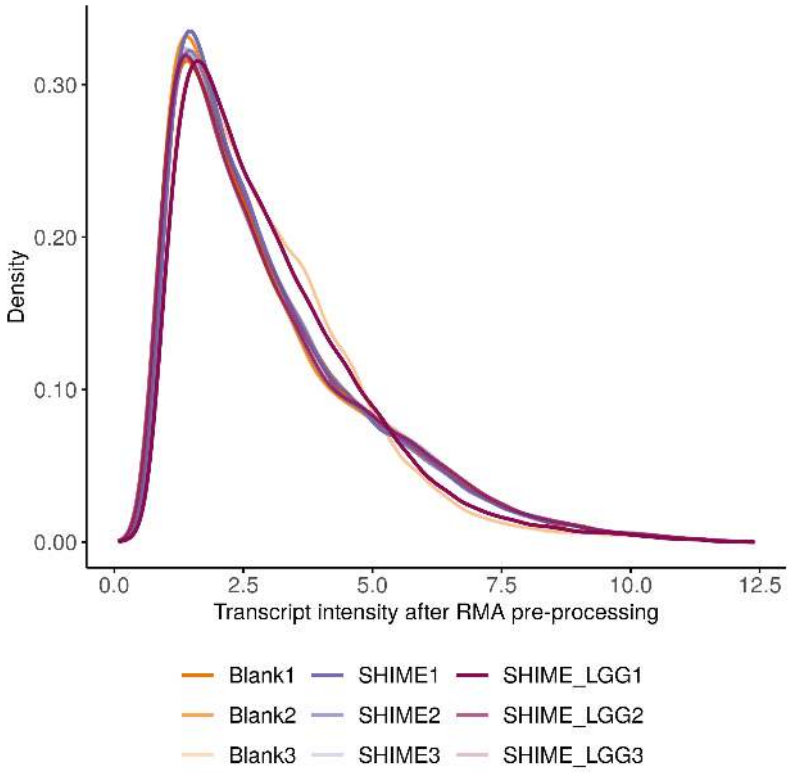

Figure 6.1: Density plot of the transcript intensities after RMA pre-processing.

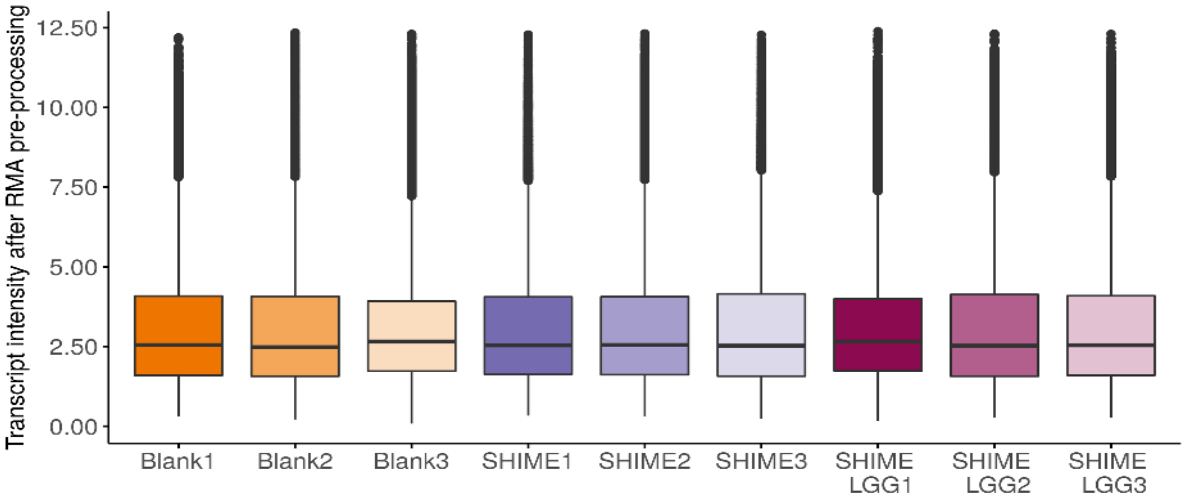

Figure 6.2: Box plots of the transcript intensities after RMA pre-processing.

6.1.2 Distribution of the RMA pre-processed background transcript intensities

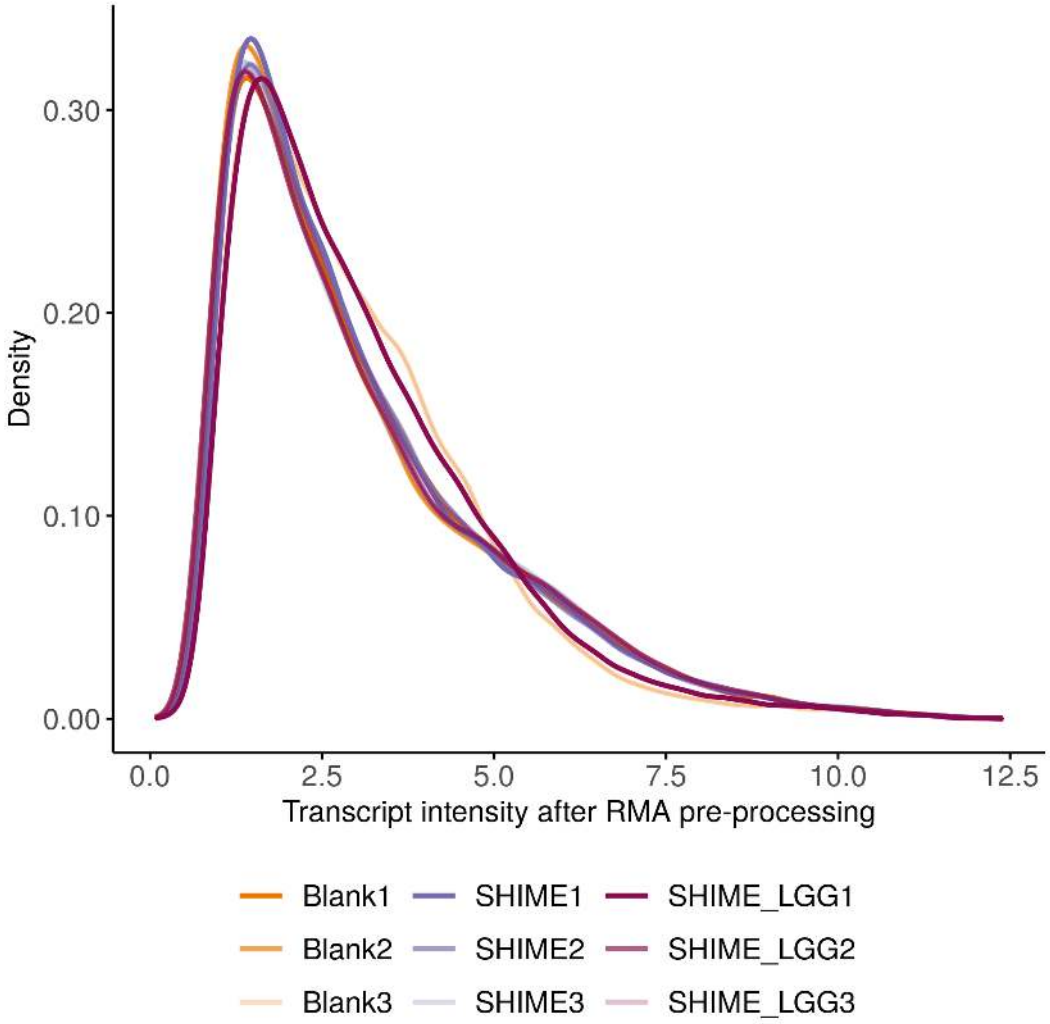

Figure 6.3: Density plot of the background probe intensities after RMA pre-processing.

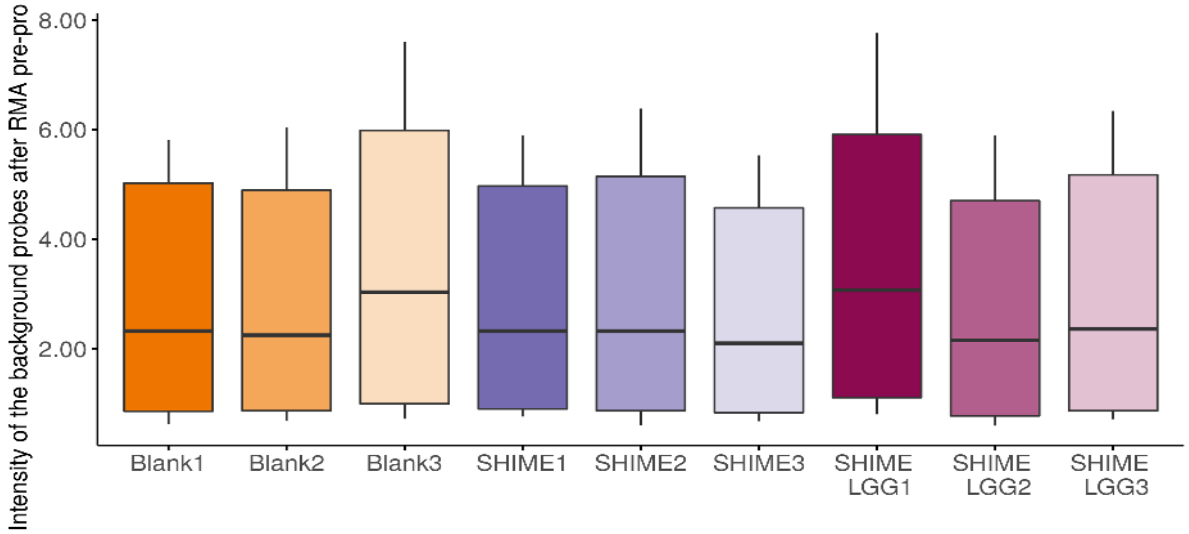

Figure 6.4: Box plots of the background probe log2 intensities after RMA pre-processing.

6.1.3 Distribution of the RMA pre-processed control transcript intensities

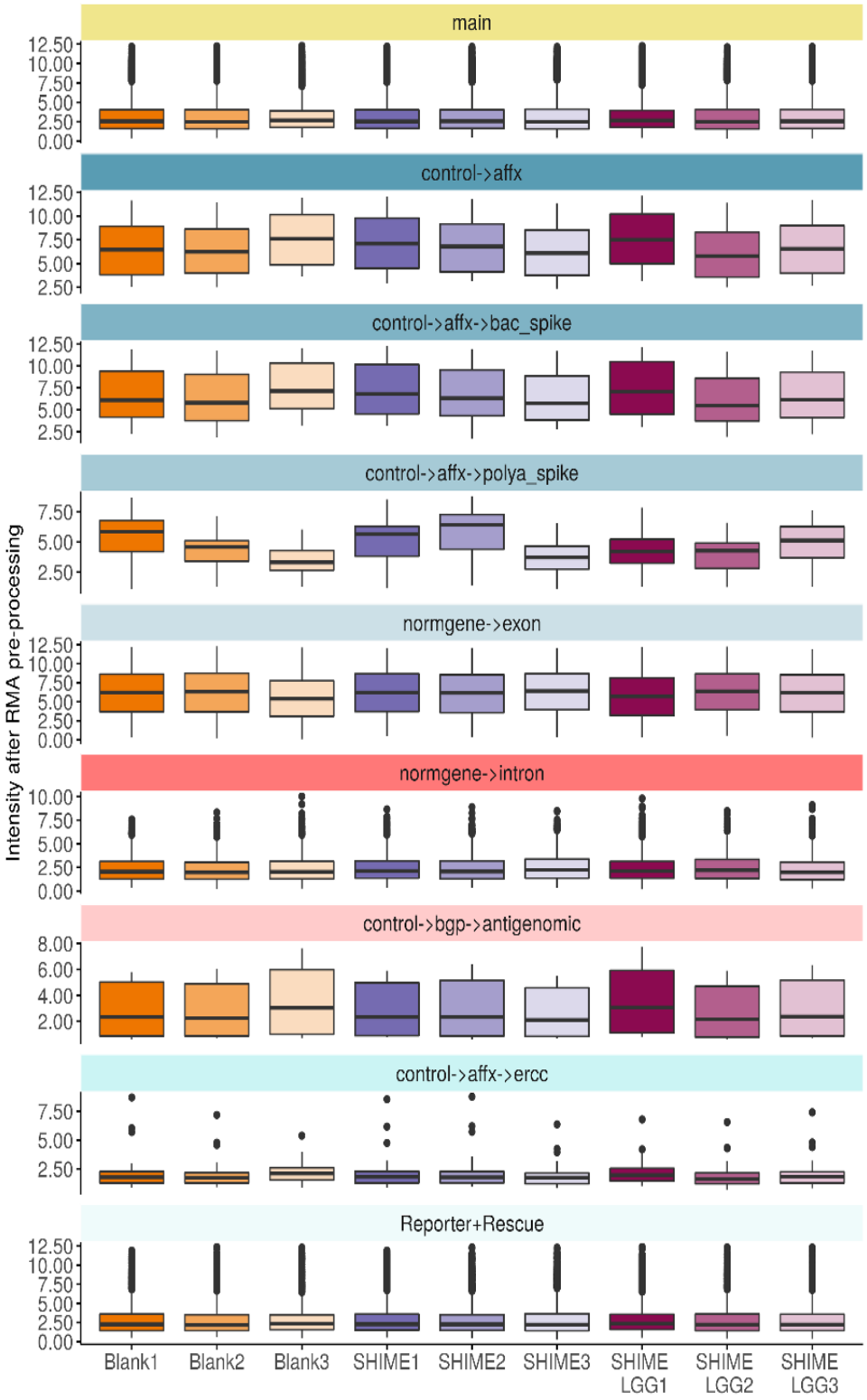

Figure 6.5: Box plots of the intensities of all control probes after RMA pre-processing within the probesets corresponding to the spike-in positive controls (control→affx, control→affx→bac_spike, control→affx→polya_spike) and negative controls (control→affx→ercc, control→bgp→antigenomic,normgene→intron) faceted by the controls.

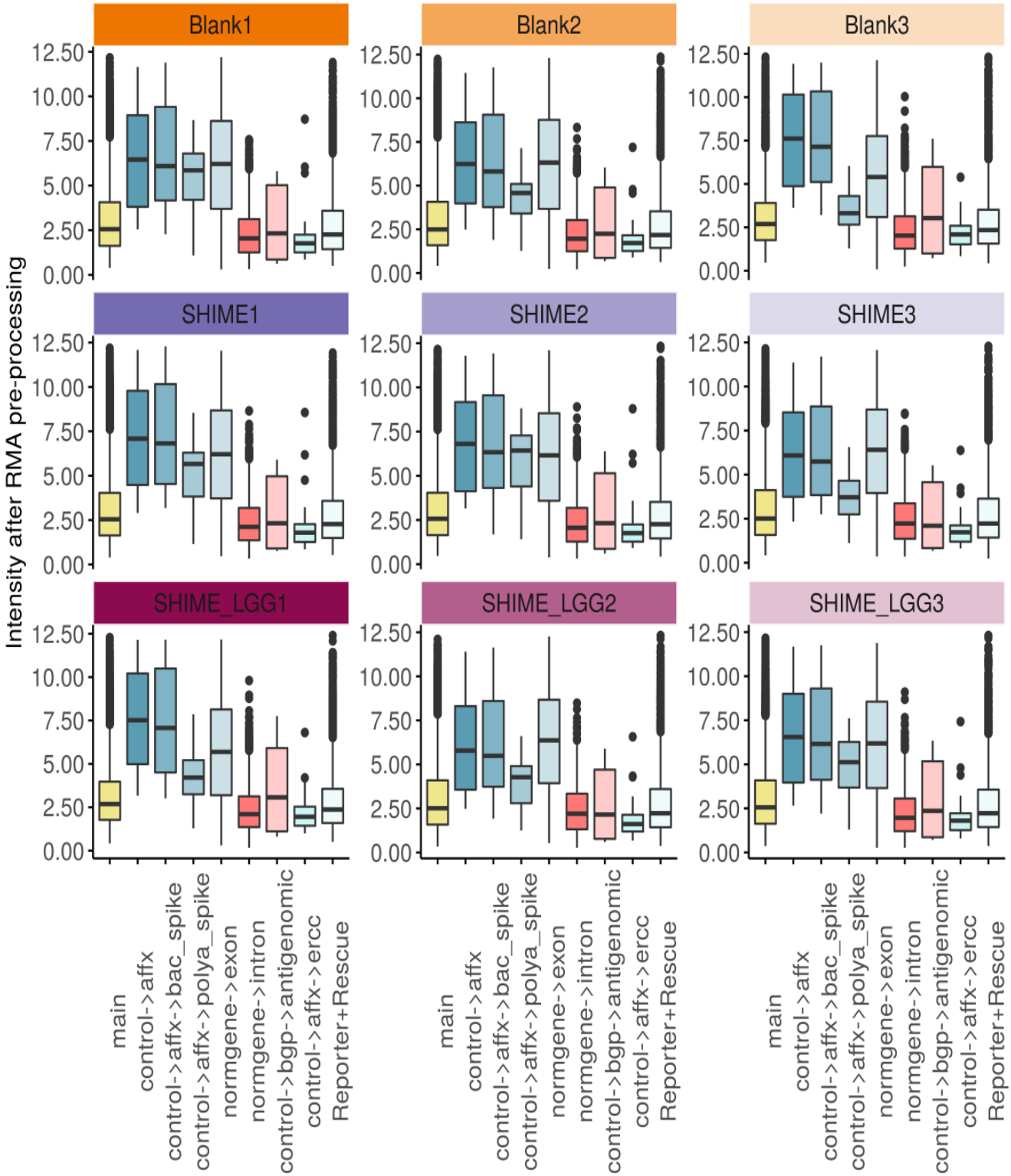

Figure 6.6: Box plots of the intensities of all control probes after RMA pre-processing within the probesets corresponding to the spike-in positive controls (control→affx, control→affx→bac_spike, control→affx→polya_spike) and negative controls (control→affx→ercc, control→bgp→antigenomic, normgene→intron) faceted by the samples.

6.1.3.1 Distribution of the RMA pre-processed bac and polyA transcript intensities

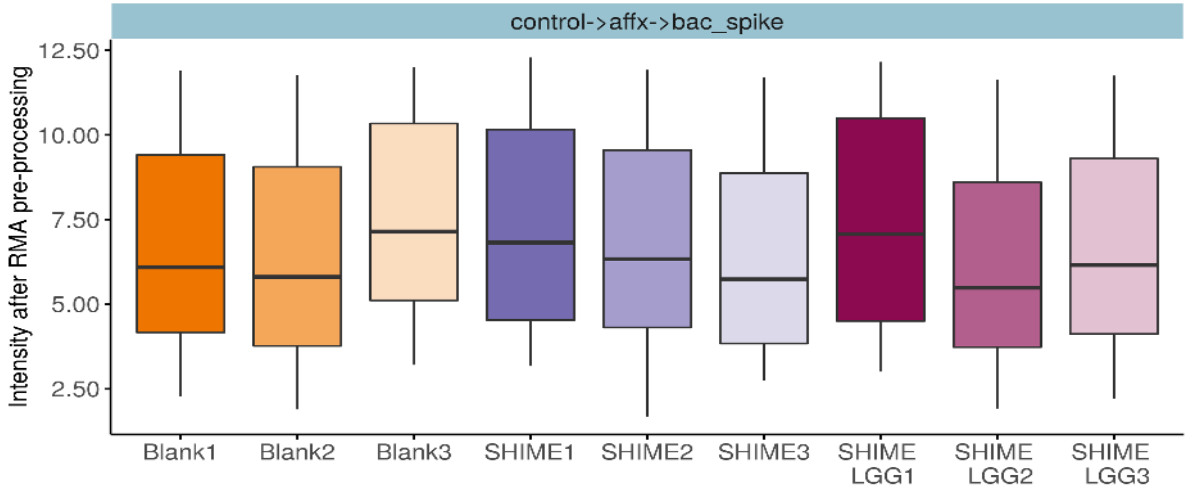

Figure 6.7: Intensities of all probes after RMA pre-processing within the probesets corresponding to the bacterial Affymetrix spike-in faceted by the samples.

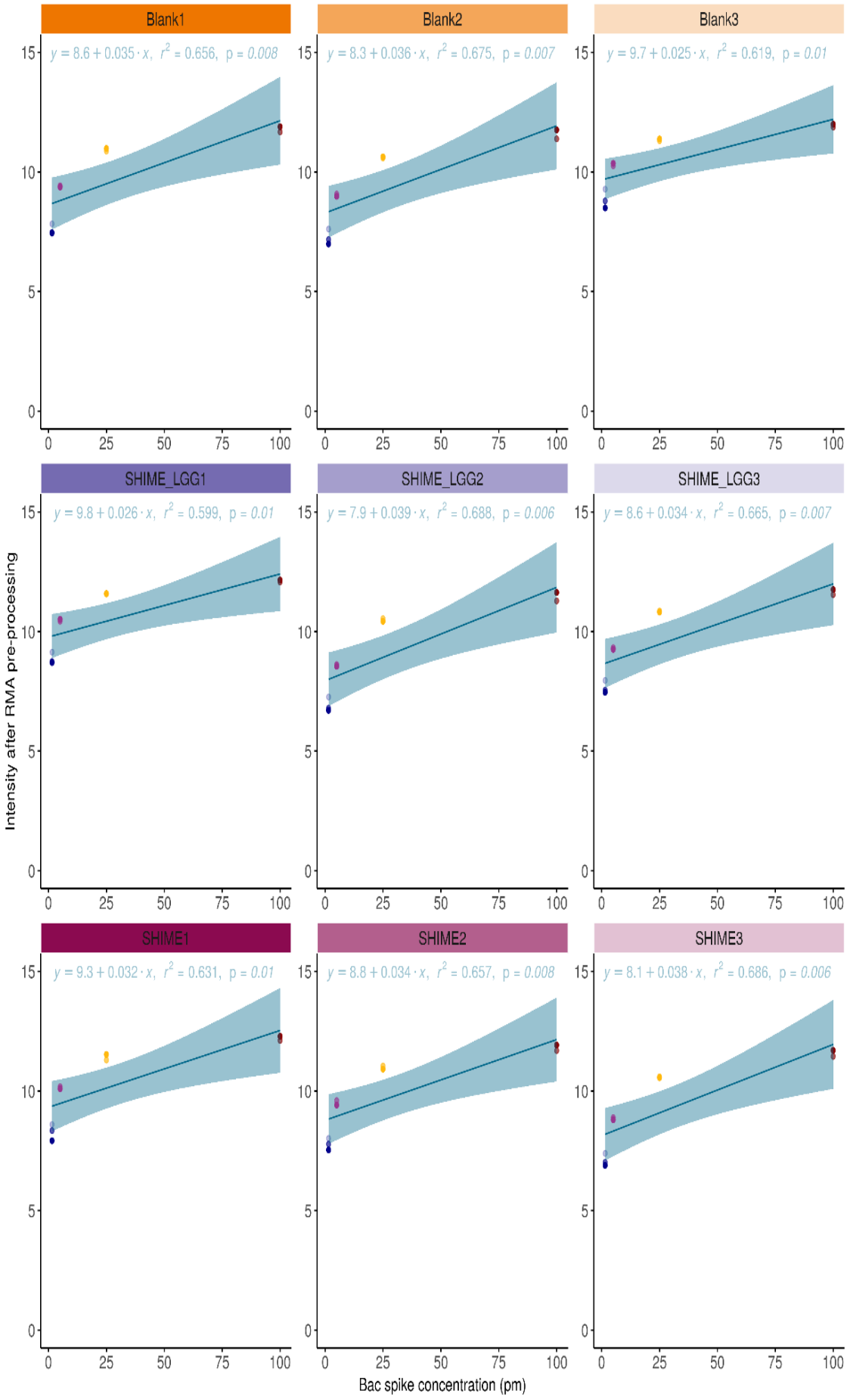

Figure 6.8: Intensities of the BioB, BioC, BioD and Cre bacterial Affymetrix spike-in probes after RMA pre-processing faceted by the samples.

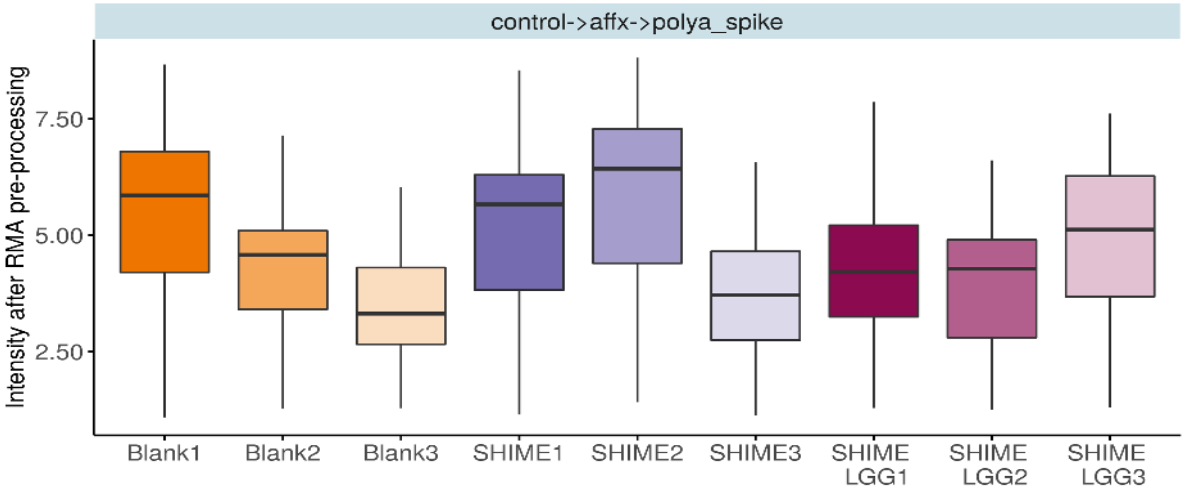

Figure 6.9: Intensities of all probes after RMA pre-processing within the probesets corresponding to the polyA Affymetrix spike-in faceted by the samples.

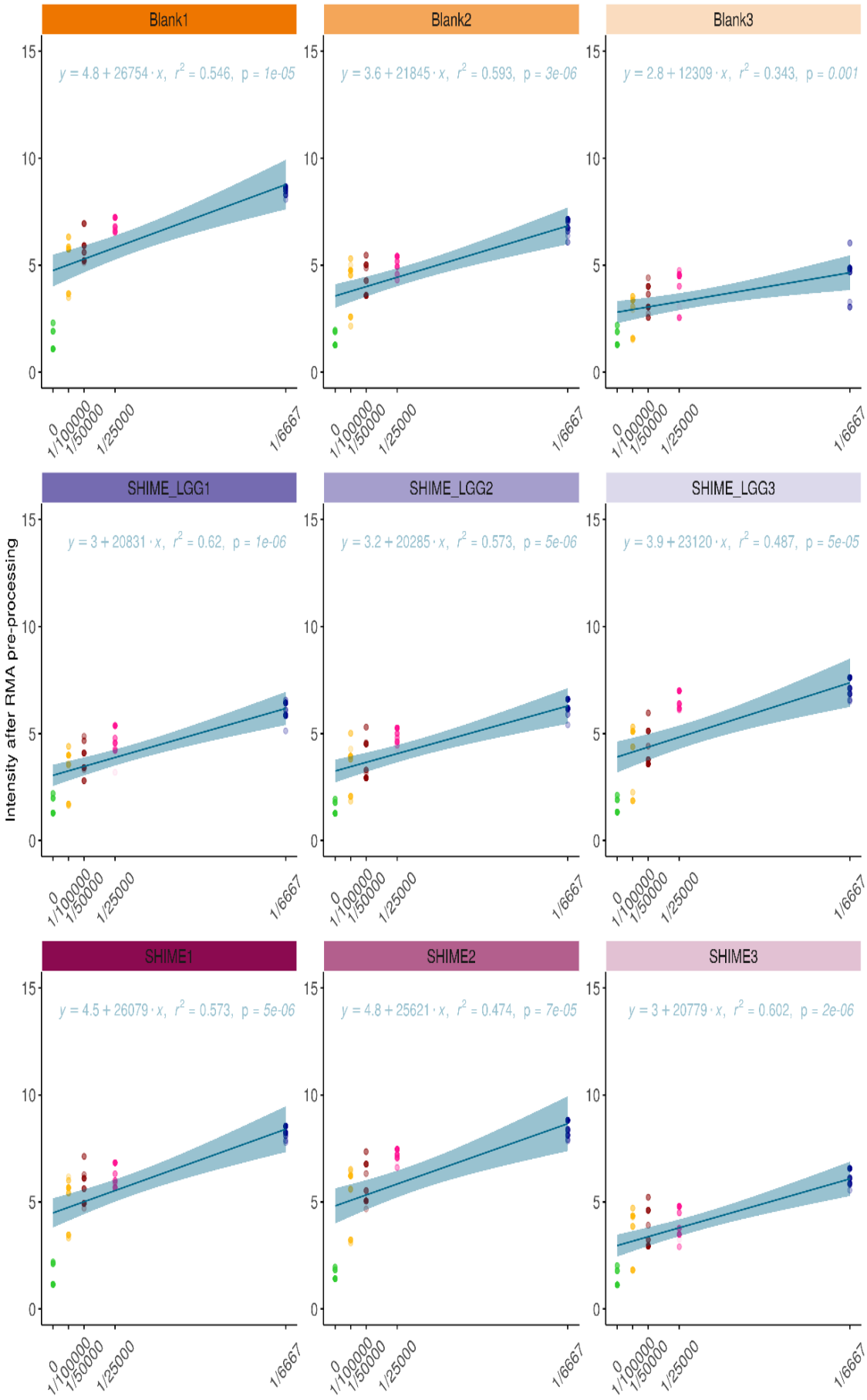

Figure 6.10: Intensities of all probes after RMA pre-processing within the probesets corresponding to the polyA Affymetrix spike-in faceted by the samples.

6.1.3.2 Distribution of the RMA pre-processed housekeeping gene transcript intensities

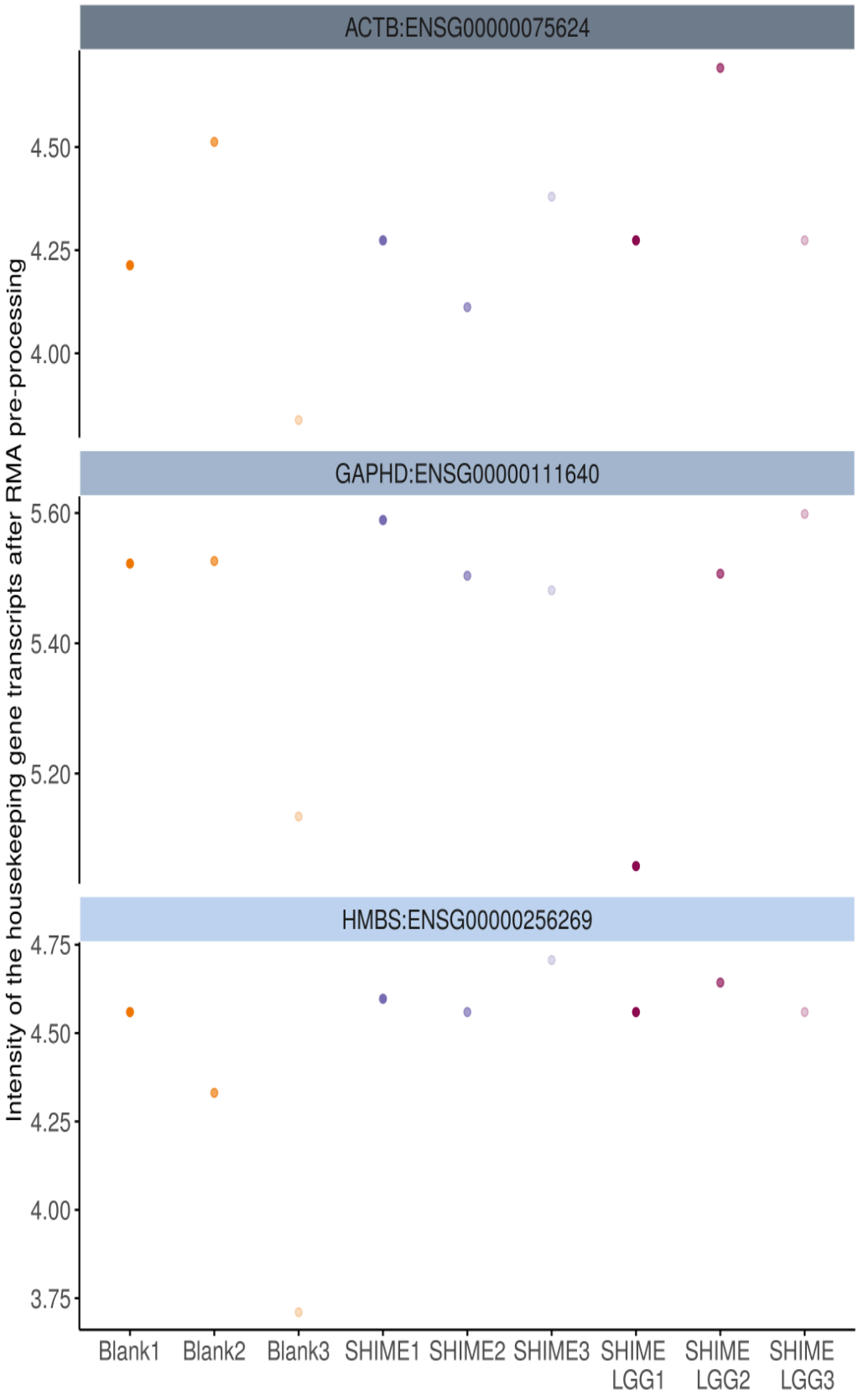

Figure 6.11: Intensities of all probes after RMA pre-processing within the probesets corresponding to the housekeeping genes ACTB, GAPHD and HMBS.

6.2 ROC curves

RMA pre-processing increased the AUC values of ROC curves assessing the predictive value of the log2 probe intensities for classification of the negative and positive control samples. The AUC value increased from around 0.8 to 0.86/0.87 for most microarrays, except for sample Blank3 and SHIME_LGG1, which increased from 0.77/0.78 to 0.82 (Figures 6.12 and 6.13).

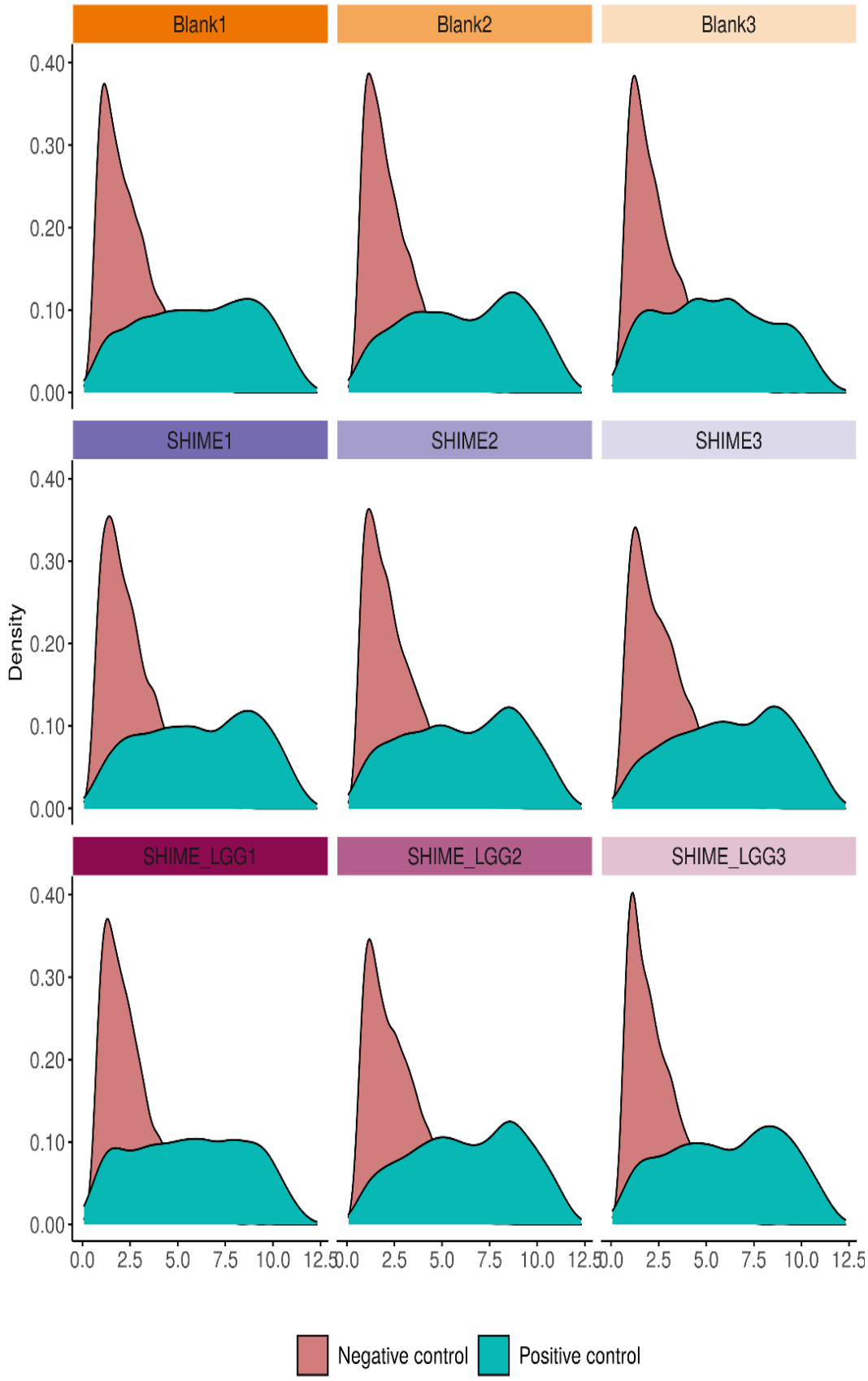

Figure 6.12: Density plot of the RMA pre-processed intensities for positive and negative controls.

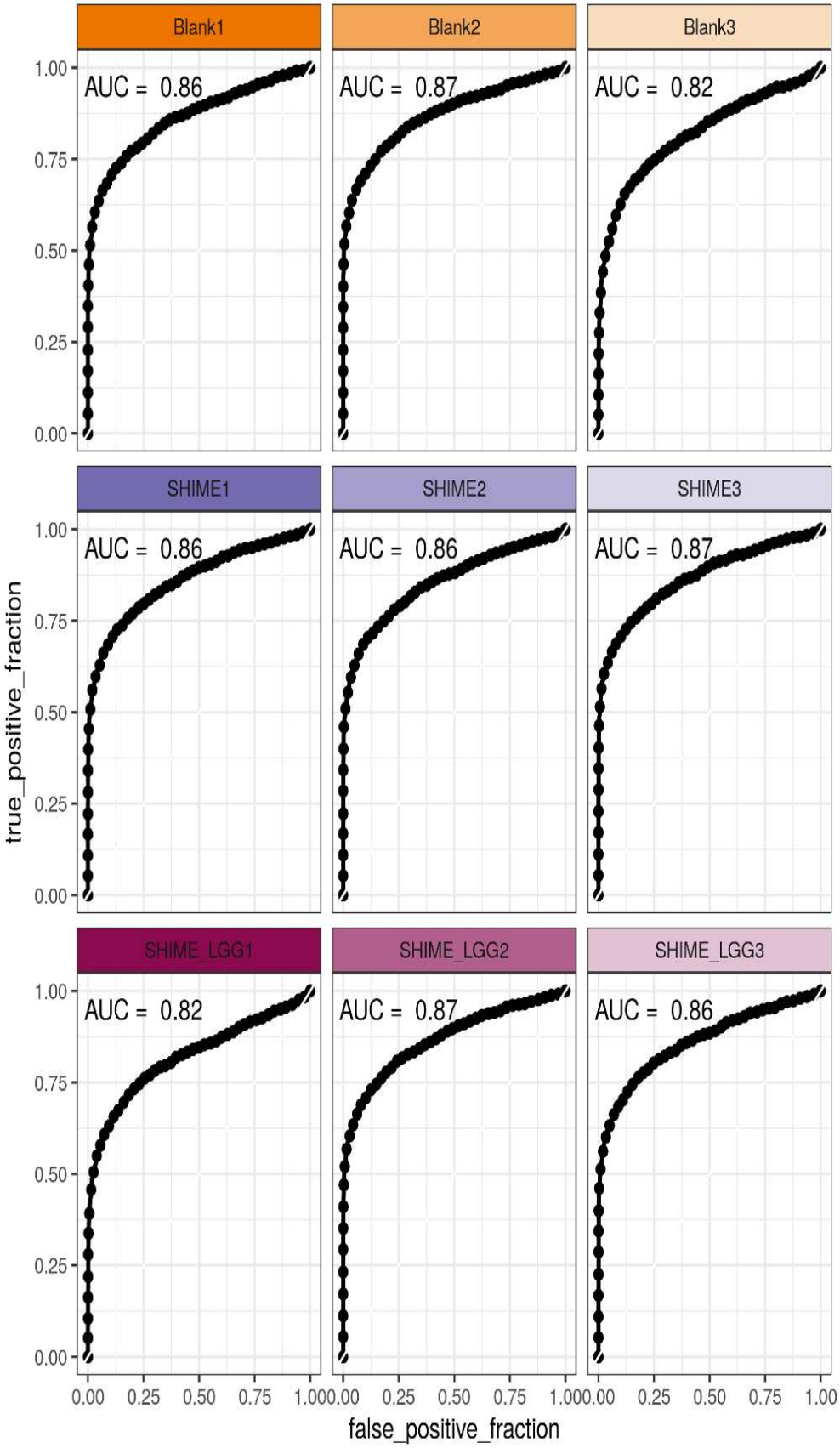

Figure 6.13: ROC curves of the RMA pre-processed data.

6.3 MA plots

Compared to the raw intensities, a more symmetric and even spread of the data is observed indicating that the dependence of the variability on the average expression level is less strong than before normalization. The loess fit coincides with the M=0 line for most samples, except Blank3 and SHIME_LGG1 (Figure 6.14).

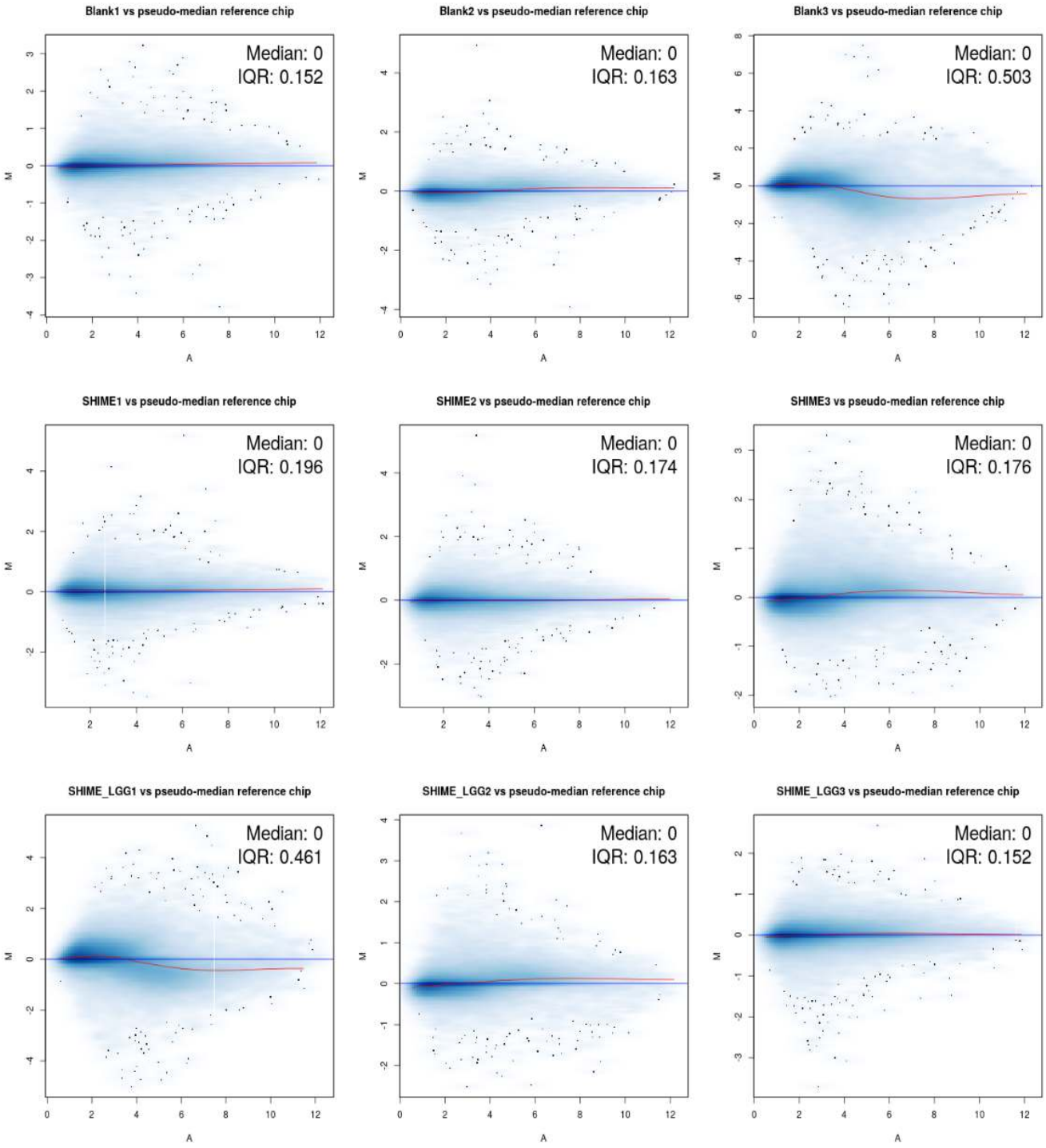

Figure 6.14: MAplots.

6.4 PA calls

To the best of our knowledge, functions to perform presence/absence calling for RMA preprocessed data at transcript level are not available (except in the xps package). We implemented Wilcoxon signed rank-based gene expression presence/absence detection to give an indication about which genes are not expressed in the entire dataset. The following hypotheses were considered:

$$H_0 : PM\ intensity <= BG\ intensity \tag{6.1}$$
$$H_1 : PM\ intensity > BG\ intensity \tag{6.2}$$

Note that use of a global universal background estimate for all probes is not acceptable and results in a large distribution of errors in a PLM type of model. Instead, background estimates of probes with similar GC content as the probe to be analyzed (GC-bins) performed equally good as the mismatch probe methods in earlier versions of Affymetrix arrays. Therefore, the expression of all probes corresponding to a gene/transcript were compared to background probes with a similar GC count.

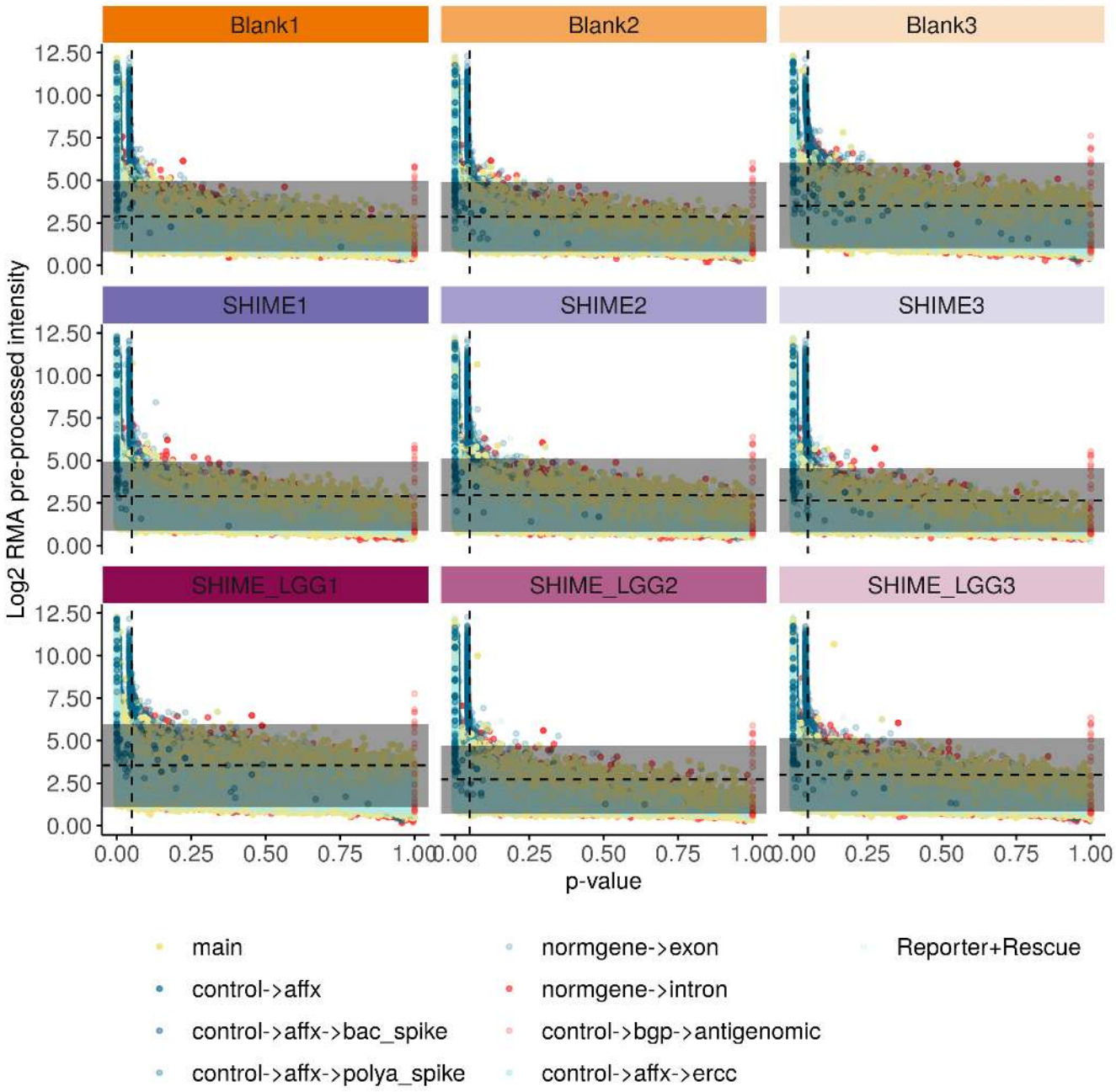

Figure 6.15: Detected Above Background plots, with probe intensities per microaray. The $\alpha = 0.05$ significance level is indicated by a vertical dashed line. The per-sample average background probe intensity and standard deviation are marked by a horizontal dashed line with shaded interval.

Only 28.72 percent of the transcripts significantly exceeds the background probe signal. The p-value increases with decreasing probe intensity and as expected the negative control probes are enriched at p>0.05 (Figure 6.15).

6.5 Probe-level model fitting: Chip pseudo-images, RLE and NUSE - summarized data

Although the median polish algorithm (used in RMA) is not available in the fit PLM model, the pseudochip images generated with the PLM method suggest that the blobs are properly disambiguated since the identified overly intense artefacts are not showing up in the unaffected high-quality arrays as under-expressed areas (Figures 6.16-6.17). To ensure that blobs were properly dealt with, intensities of the probes in the blob region were manually inspected prior to and after RMA normalization. Blob regions were gated with the gglocator function and probes with coordinates within the demarcated polygon were identified. Boxplots comparing the intensities in the blob area before and after RMA pre-processing confirmed that the artefact was succesfully removed, without affecting the other arrays (Figures 6.23 and 6.24) . The output from RMA suggests that RMA does a fair job of disambiguating the scratch from the surrounding data, but some areas of blue still exist near the scratch, suggesting that the gene expression values for these probe sets have been overestimated. All NUSE values are around 1, with slight deviations for sample blank3 and SHIME_LGG1, indicating that the skewdness in raw intensities in not properly corrected by normalization in these chips (Figure 6.19). The bigger spread for blank3 and SHIME_LGG1 is also apparent in the RLE values (Figure 6.18), which are still nicely centered around zero.

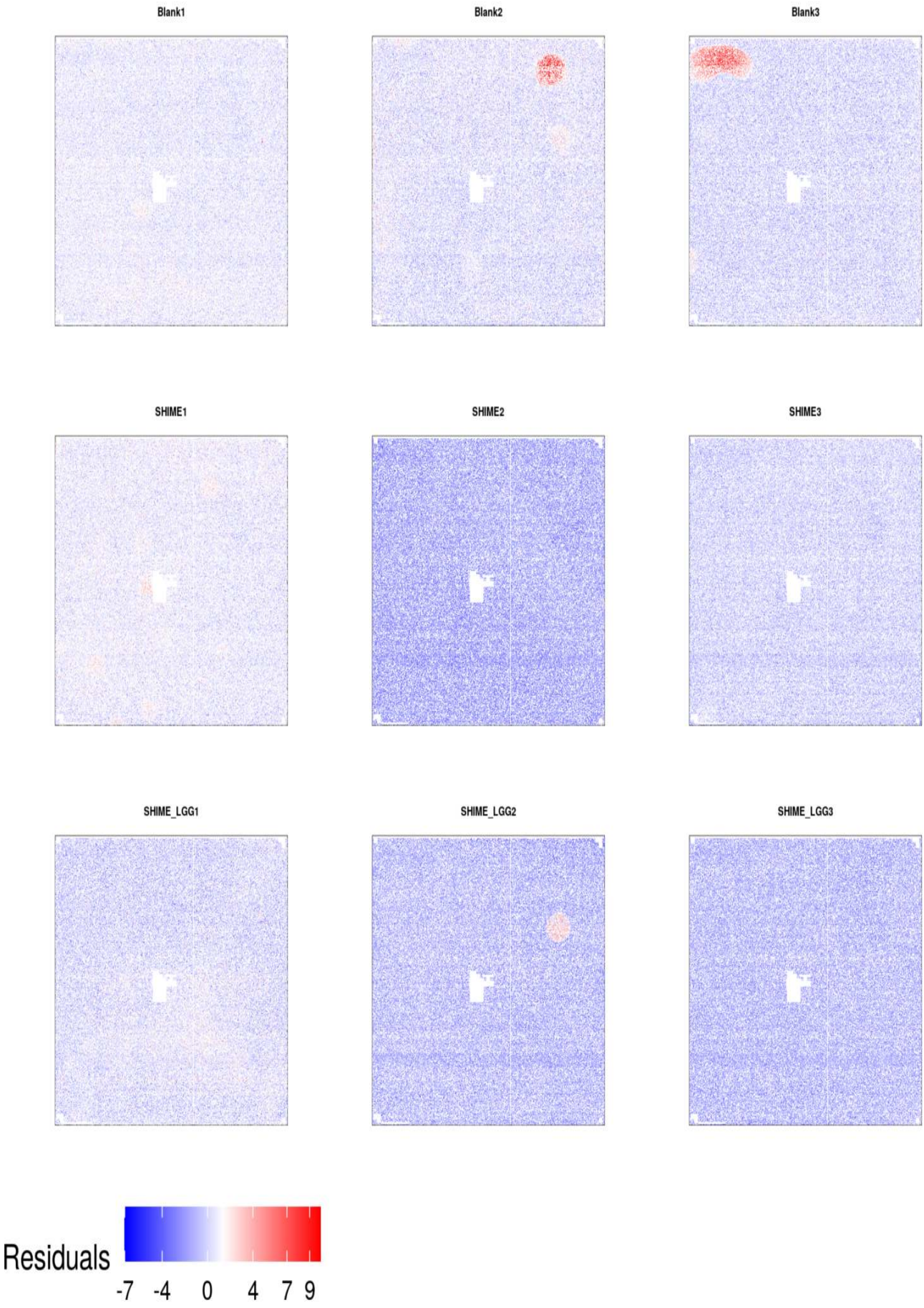

Figure 6.16: Pseudo-images of the estimated residuals from a probe-level model fitting after background correction and normalization

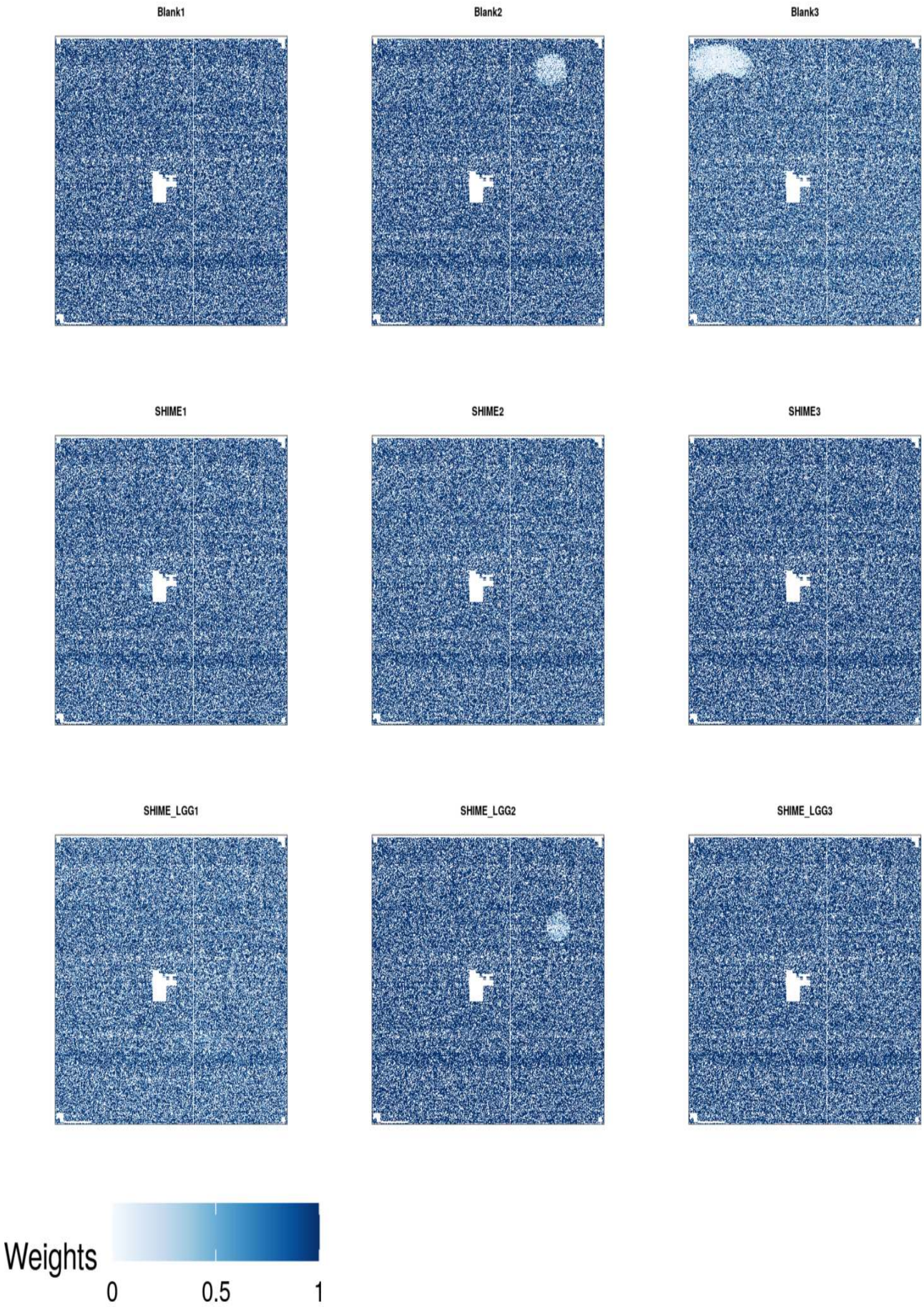

Figure 6.17: Pseudo-images of the estimated weights from a probe-level model fitting after background correction and normalization

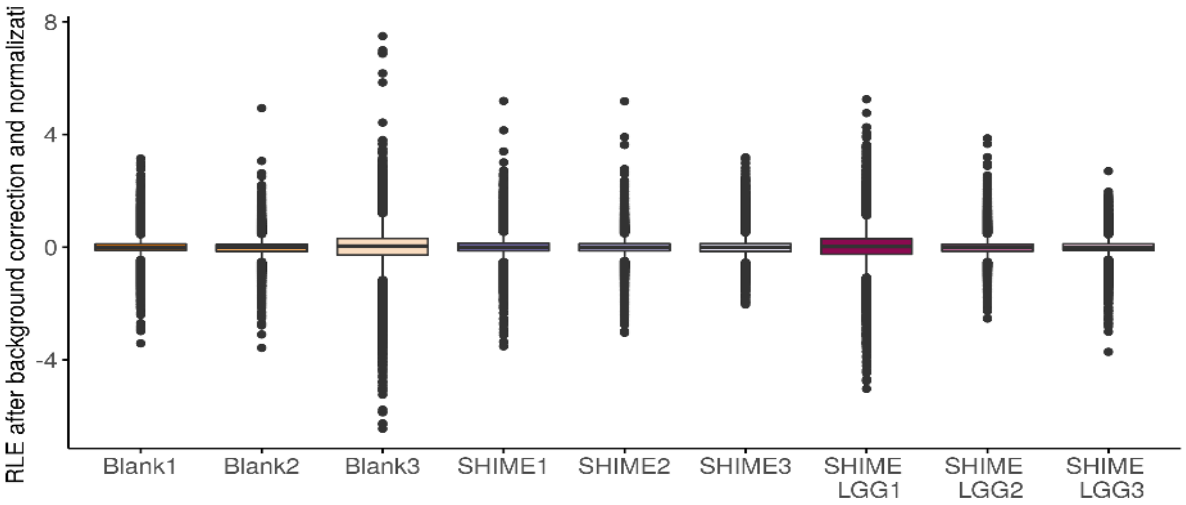

Figure 6.18: Box plot of the Relative Log Expression (RLE) values based on a probe-level model fitting after background correction and normalization

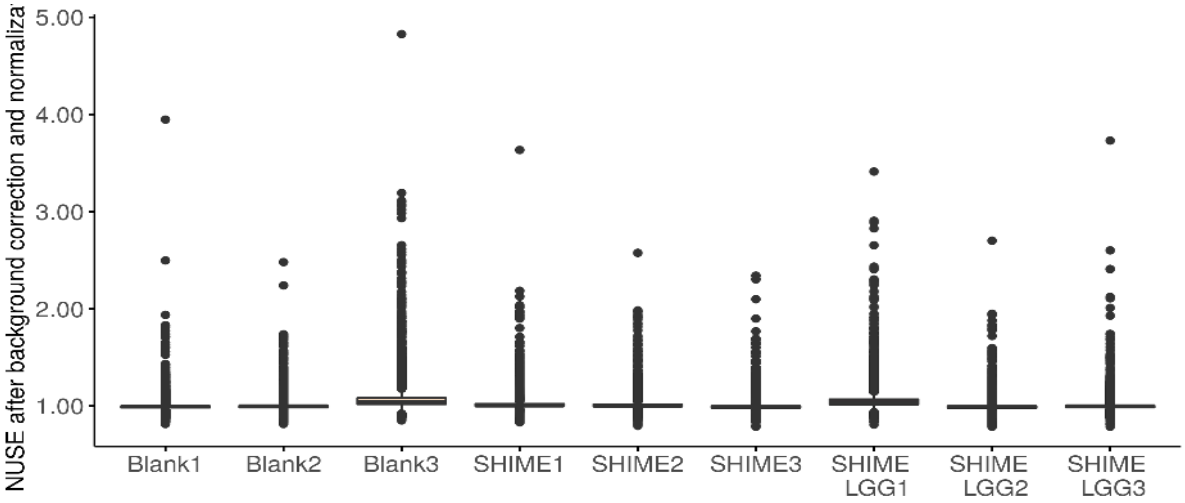

Figure 6.19: Box plot of the Normalized Unscaled Standard Errors (NUSE) values based on a probe-level model fitting after background correction and normalization

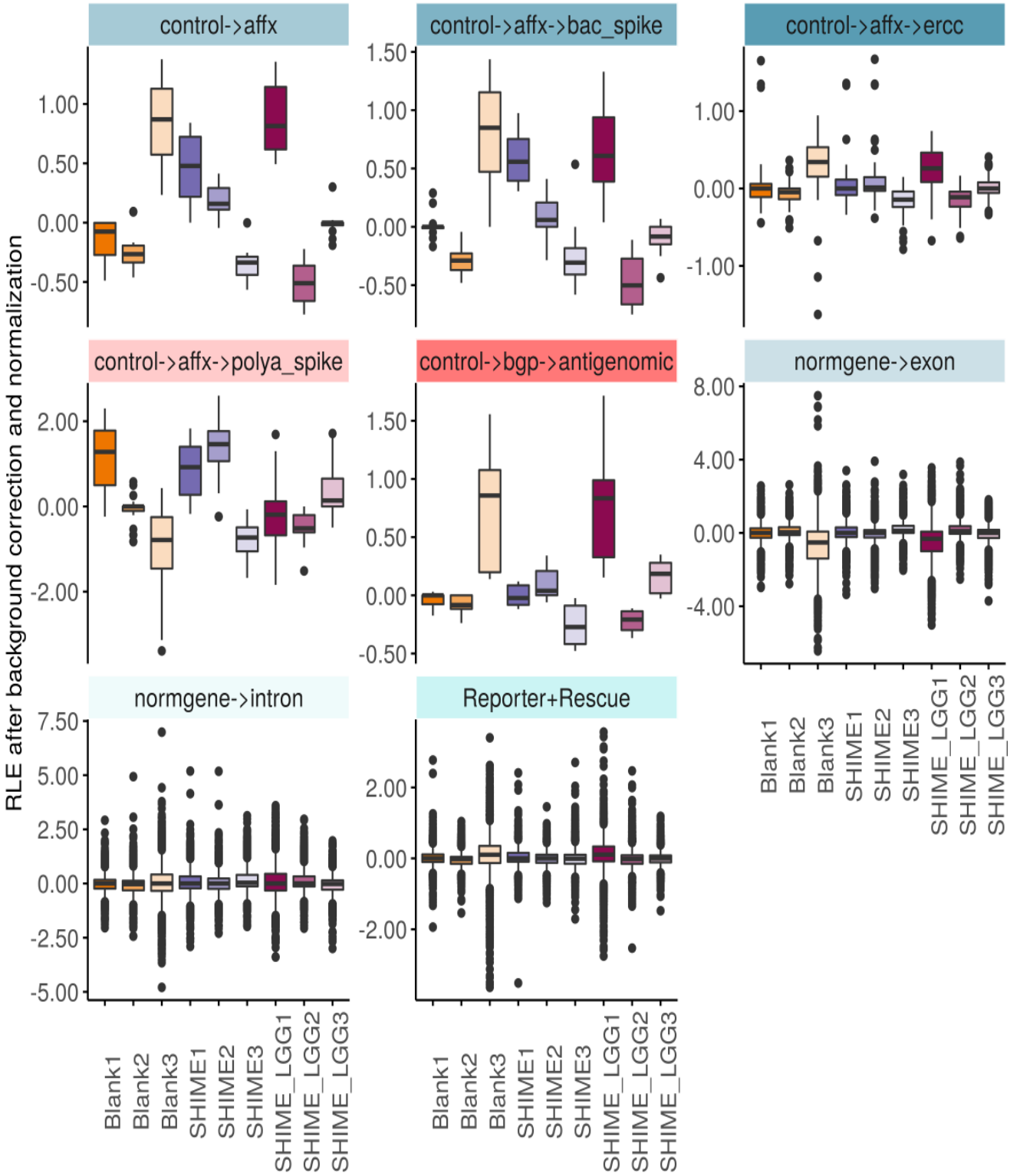

Figure 6.20: Box plot of the Relative Log Expression (RLE) values of control probes based on a probe-level model fitting after background correction and normalization

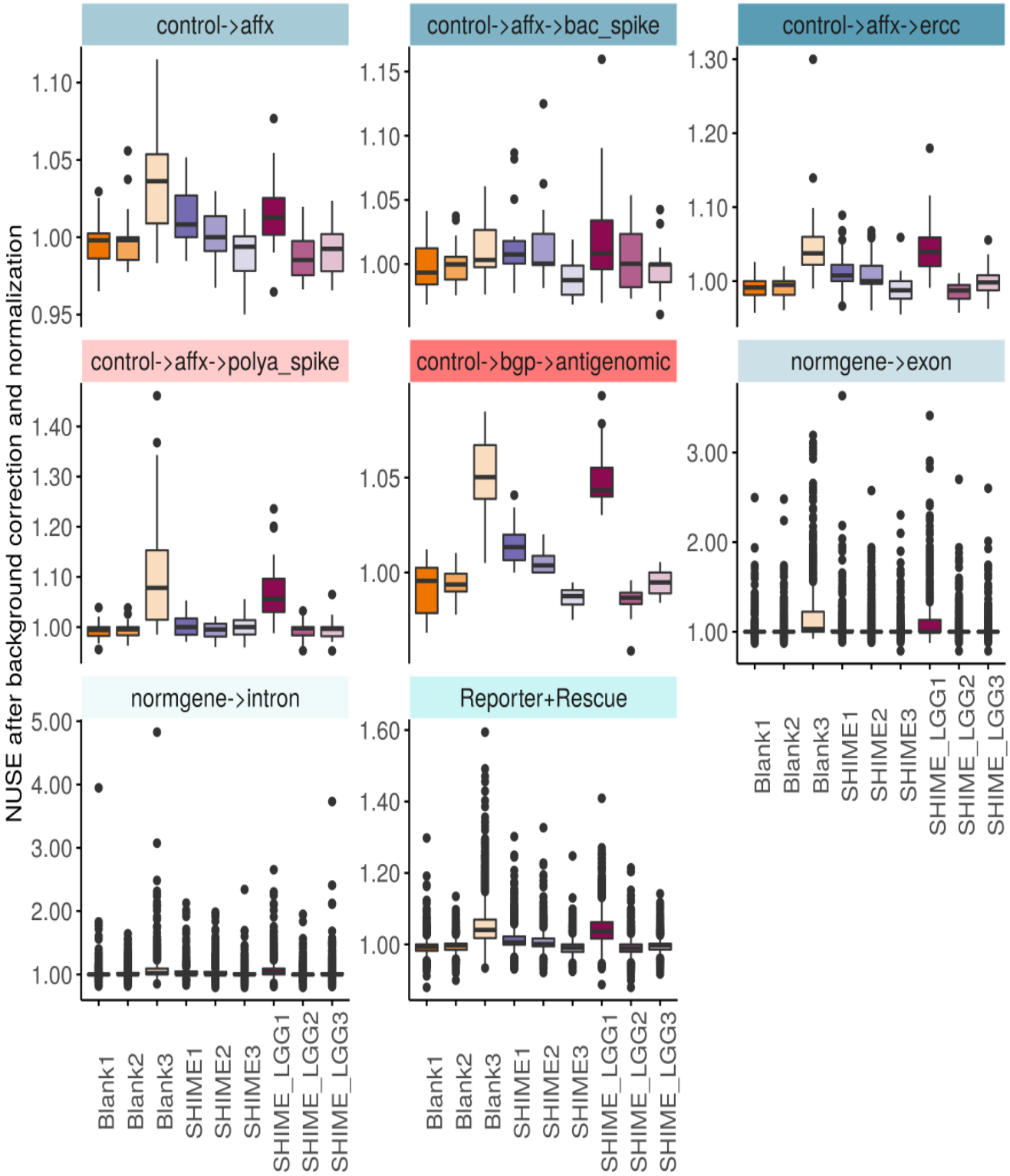

Figure 6.21: Box plot of the Normalized Unscaled Standard Errors (NUSE) values of control probes based on a probe-level model fitting after background correction and normalization

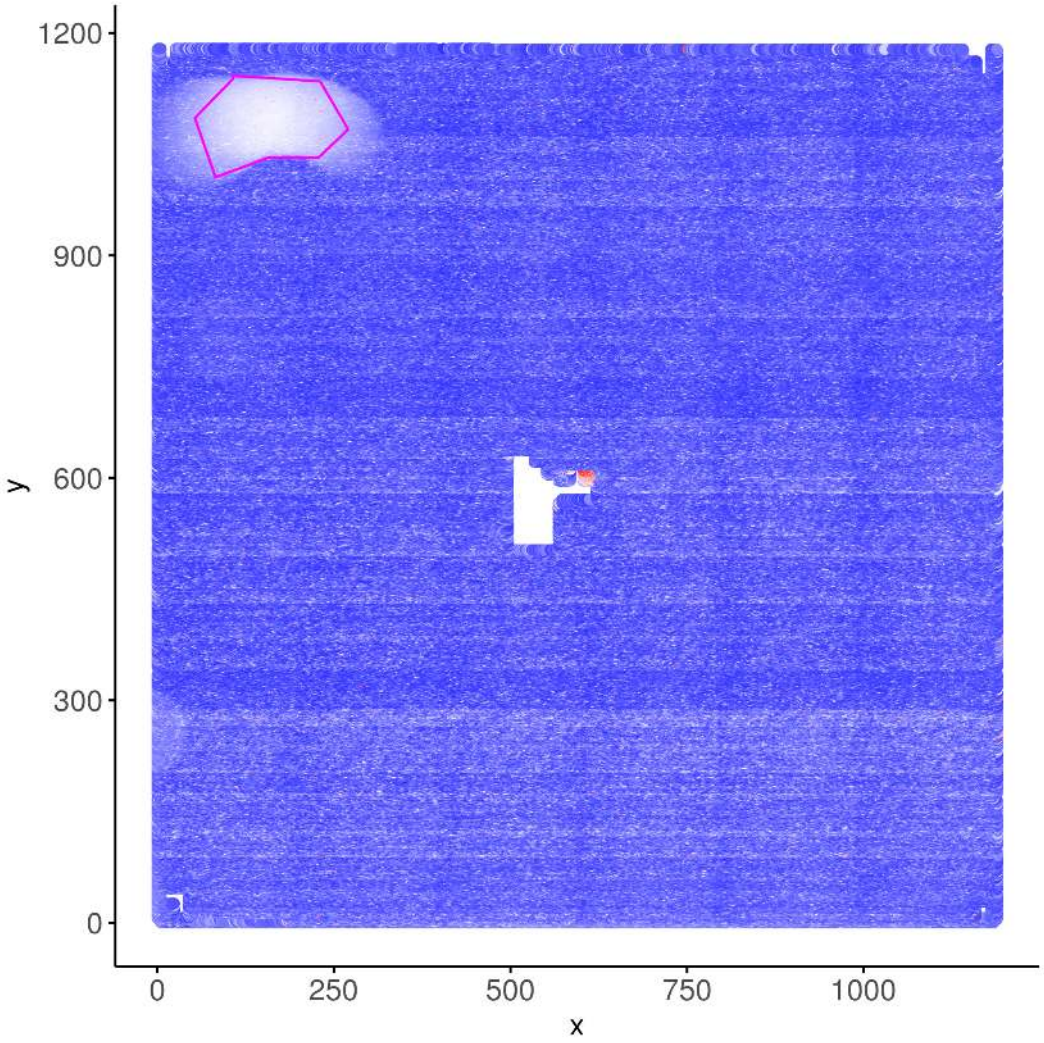

Figure 6.22: Gating the blob region to extract data of the affected probes.

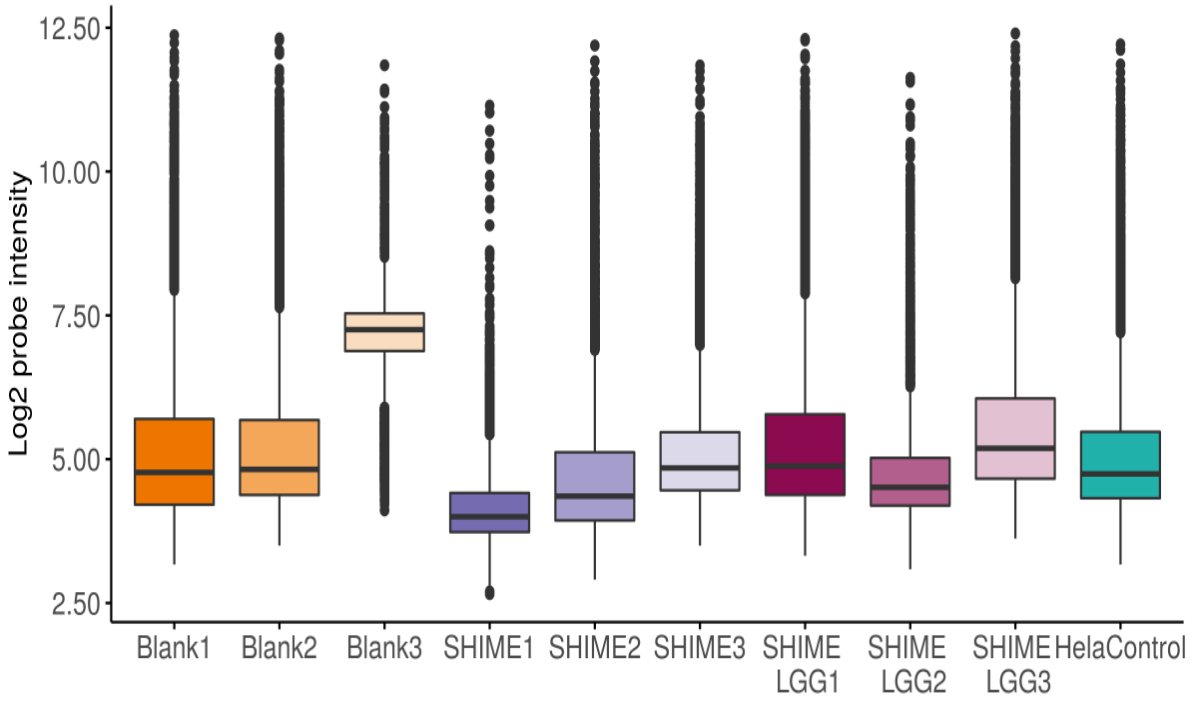

Figure 6.23: Box plot of the unnormalized raw intensities of the blob probes

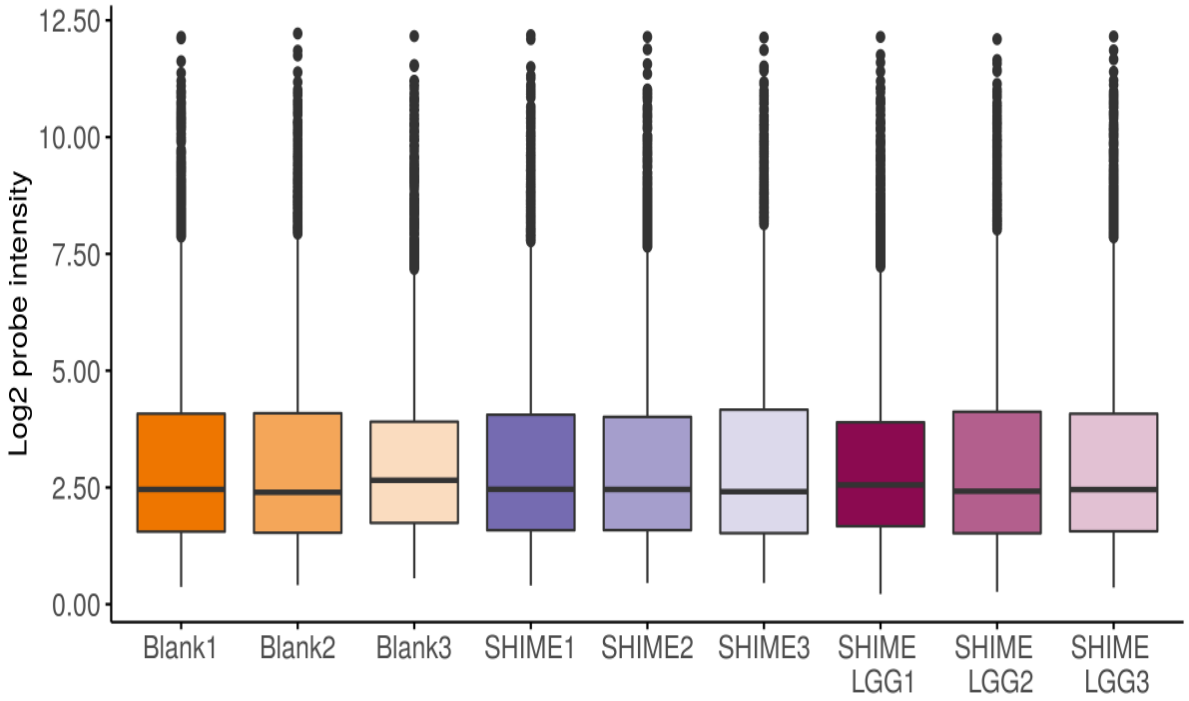

Figure 6.24: Box plot of the RMA pre-processed intensities of the blob probes

6.6 PCA

While no outlier samples were identified according to the Hotelling's T2 criterion, sample Blank3 and SHIME_LGG1 are separated from the other samples in most probesets, including the bac spike-in controls suggesting that differences in hybridization efficiency underlie the observed discrepancies (Figure 6.25).

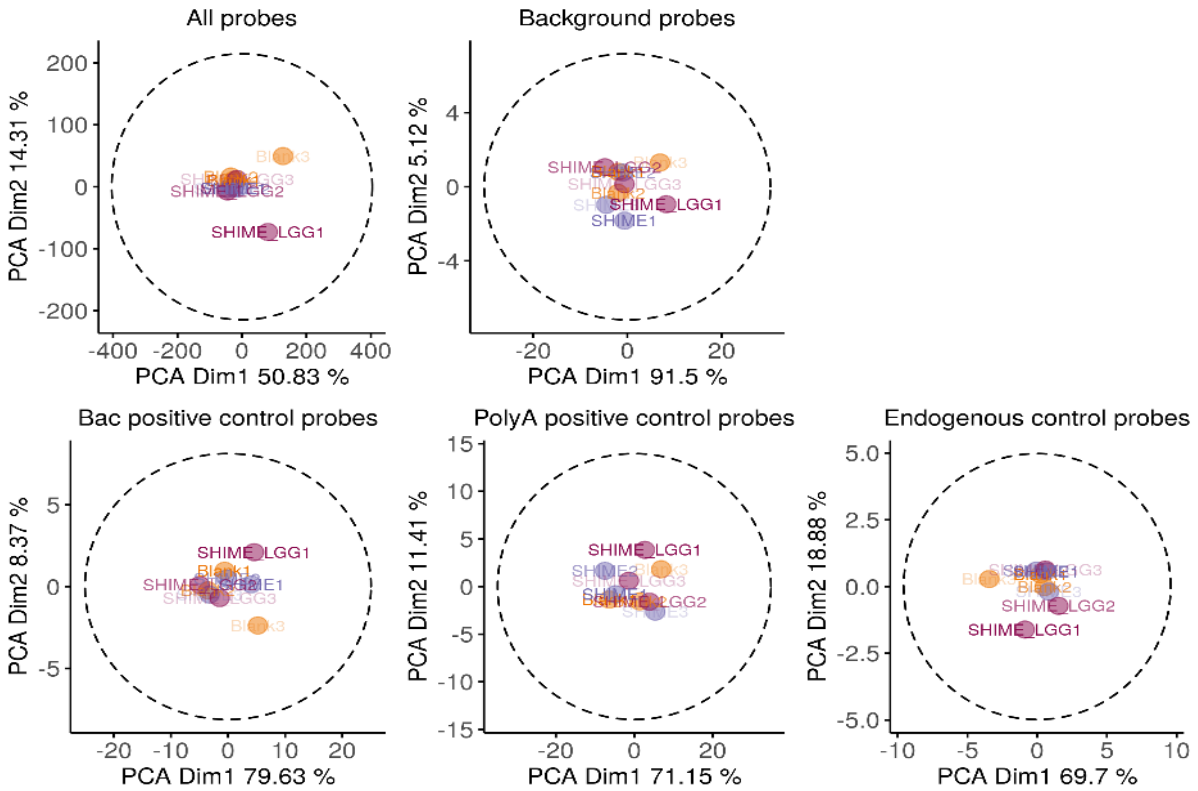

Figure 6.25: PCA scores plot of the RMA pre-processed standardized data.

Chapter 7

Triple coculture cell model validation

To validate the triple coculture cell model, expression above background was evaluated for several genes of interest related to epithelial barrier integrity (TJP1 = ZO-1), mucus expression (MUC2), cell surface receptors (MS4A12), immune signalling (TLR4) and transport (SLC16A1 = MCT1), as well as, marker genes for the presence of macrophages (CD68, ITGAM = CD11b). Results from the presence/absence calling for RMA pre-processed data at transcript level were used (see 6.4).

Table 7.1: Average above background expression for several genes of interest related to epithelial barrier integrity (TJP1 = ZO-1), mucus expression (MUC2), cell surface receptors (MS4A12), immune signalling (TLR4) and transport (SLC16A1 = MCT1), as well as, marker genes for the presence of macrophages (CD68, ITGAM = CD11b).

Ensembl gene ID	**Entrez gene ID**	**Gene name**	**Average expression across all conditions**
ENSG00000071203	54860	MS4A12	2.325954323244329380316
ENSG00000104067	7082	TJP1	6.913905037860886615420
ENSG00000104067	7082	TJP1	1.518063995563541768163
ENSG00000129226	968	CD68	8.333876454735577254951
ENSG00000136869	7099	TLR4	3.734387203672488997341
ENSG00000169896	3684	ITGAM	2.368340009667618595302
ENSG00000198788	4583	MUC2	6.351978112886125238390

Chapter 8

Differential expression analysis with Limma limma)

Differential expression was assessed by means of the Limma package 3.48.0) [8, 7]. The central principle in Limma is to fit a linear model to the expression data of each gene, which can be either log-ratios or log-intensities. A moderated t-test or ANOVA with F-test statistic is then used to identify differential expression. An empirical Bayes method is used to moderate/smooth the standard errors of the estimated log-fold changes (e.g. standard errors are squeezed towards a common value by borrowing information across genes), making the analyses stable while at the same time improving the power, even for experiments with a small number of arrays.

As we have 3 groups of samples (Blank/SHIME/SHIME_LGG), a One-way ANOVA will be performed, followed by pairwise comparisons of contrasts.

8.1 Running limma on the raw data

The ANOVA moderated F-statistic (F) combines the t-statistics for all the contrasts (3 groups) into an overall test of significance for that gene. The F-statistic tests whether any of the contrasts are non-zero for that gene, i.e., whether that gene is differentially expressed on any contrast. The ANOVA F-test p-value is significant (< 0.05) for 1057, 3 genes. To establish which of the 3 groups differ, the ANOVA will be followed by a series of pairwise comparisons to determine the t-statistics and corresponding p-values. The output of the moderated t-tests together with the Log2 fold Changes (logFC) stored in the coefficients data slot are used to construct volcano plots. Volcano plots arrange genes according to biological and statistical significance. The X-axis shows the logFC between the two groups (a positive logFC indicates up-regulated expression in the FC numerator), which represents biological impact of the change. The Y-axis displays the p-value of a t-test comparing samples (on a negative log scale so smaller p-values appear higher up) and hence indicates the statistical evidence of the change. A large number of significant genes was observed in all individual pairwise comparisons (contrasts), but only a fraction of genes displayed absolute log2FC >2 (Figure 8.1).

The moderated t-test statistic is computed for each contrast and for each probe, resulting in a large number of hypotheses to be tested and the need for multiple testing correction. The most popular form of multiple testing adjustment is Benjamini and Hochberg's method to control the false discovery rate. The adjusted values are often called q-values if the intention is to control or

estimate the false discovery rate. If all genes with q-value below a threshold, say 0.05, are selected as differentially expressed, then the expected proportion of false discoveries in the selected group is controlled to be less than the threshold value, in this case 5%.

Implementing FDR control with Benjamini and Hochberg's method resulted in adjusted p-values equal to 1. This is an indication that there is no evidence of differential expression in the data after adjusting for multiple testing. This can occur even though many of the raw p-values may seem highly significant when taken as individual values. This situation typically occurs when none of the raw p-values are less than $\frac{1}{G}$, where G is the number of probes included in the fit.

Since the microarray design includes many genes 53617, it is common practice to apply a filtering prior to differential expression analysis to reduce the number of pairwise comparisons to be tested. Genes that are not expressed in any of the assessed conditions are by definition not differentially expressed, and such genes are unlikely to be of biological interest in a study. Removing such genes from further consideration increases the power of differential expression analysis [3]. Filtering can be done based on the average log expression values computed by the fit or through a more rigorous Detected Above Background (DABG) approach comparing the main probe intensities to the background probe intensities. Additionally a LFC cut-off can be defined to select genes with a strong biological meaning.

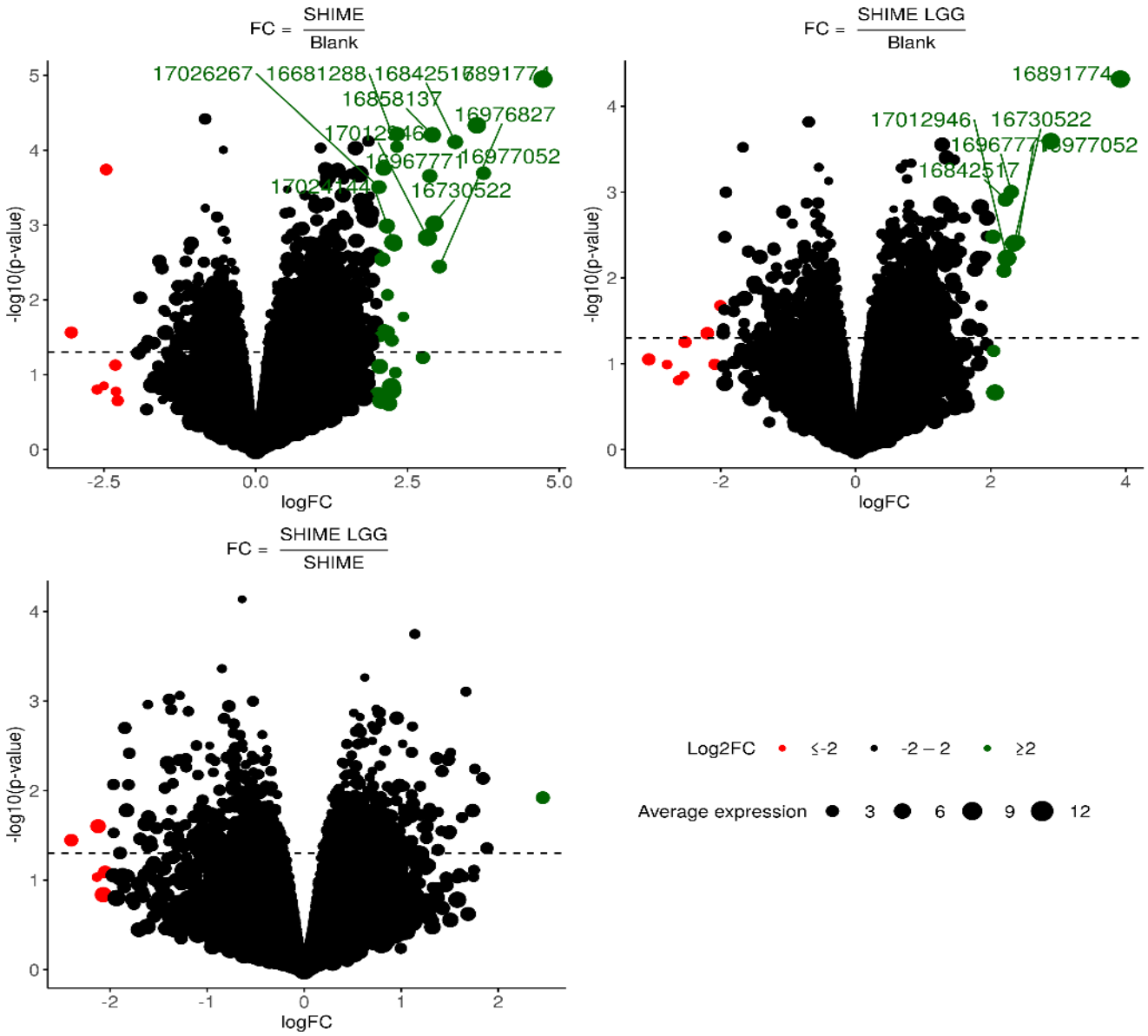

Figure 8.1: Volcano plots.

8.2 Filtering criteria

In order to assess the need for filtering, a histogram of the gene-wise median expression was constructed [3]. A large fraction of genes with a median expression below the average background signal was observed (Figure 8.2).

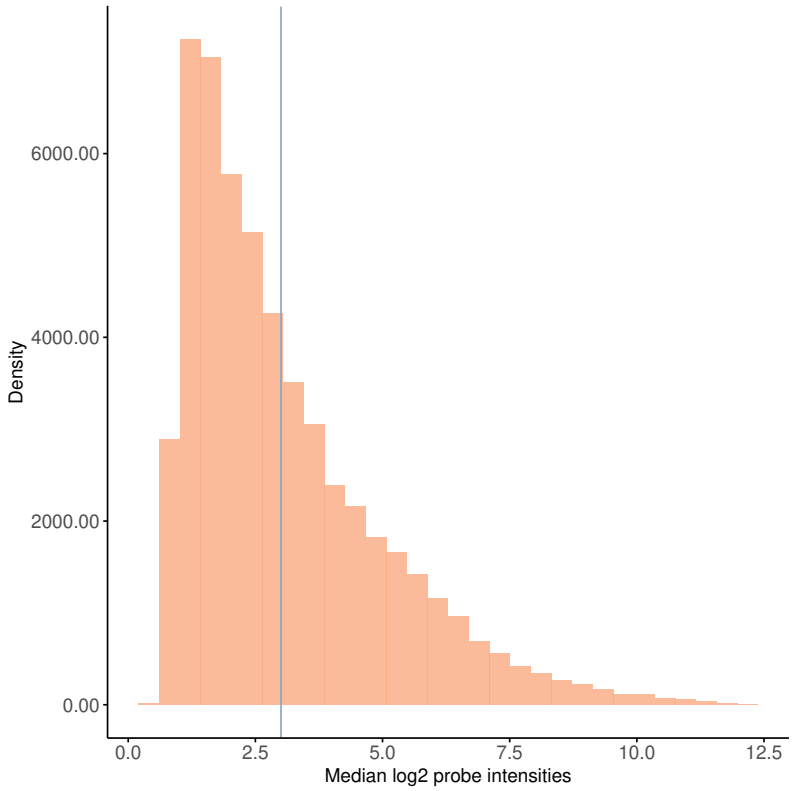

Figure 8.2: A large fraction of genes with a median expression below the average background signal (threshold line of 3) was observed.

To avoid having to set a random cut-off for the selection of genes to proceed with hypothesis testing, results from the DABG approach presented above were be used to only keep genes that are significantly (p¡0.05) expressed above background on at least 3 arrays (3 replicates per condition).

Additionally, probes that don't correspond to any known gene, i.e. that do not have a gene symbol assigned or probes with an ambiguous annotation that map to multiple gene symbols were excluded [3].

5403 genes were retained for differential expression analysis using Limma as described before (Figure 8.3).

8.3 Running limma on the DABG filtered data

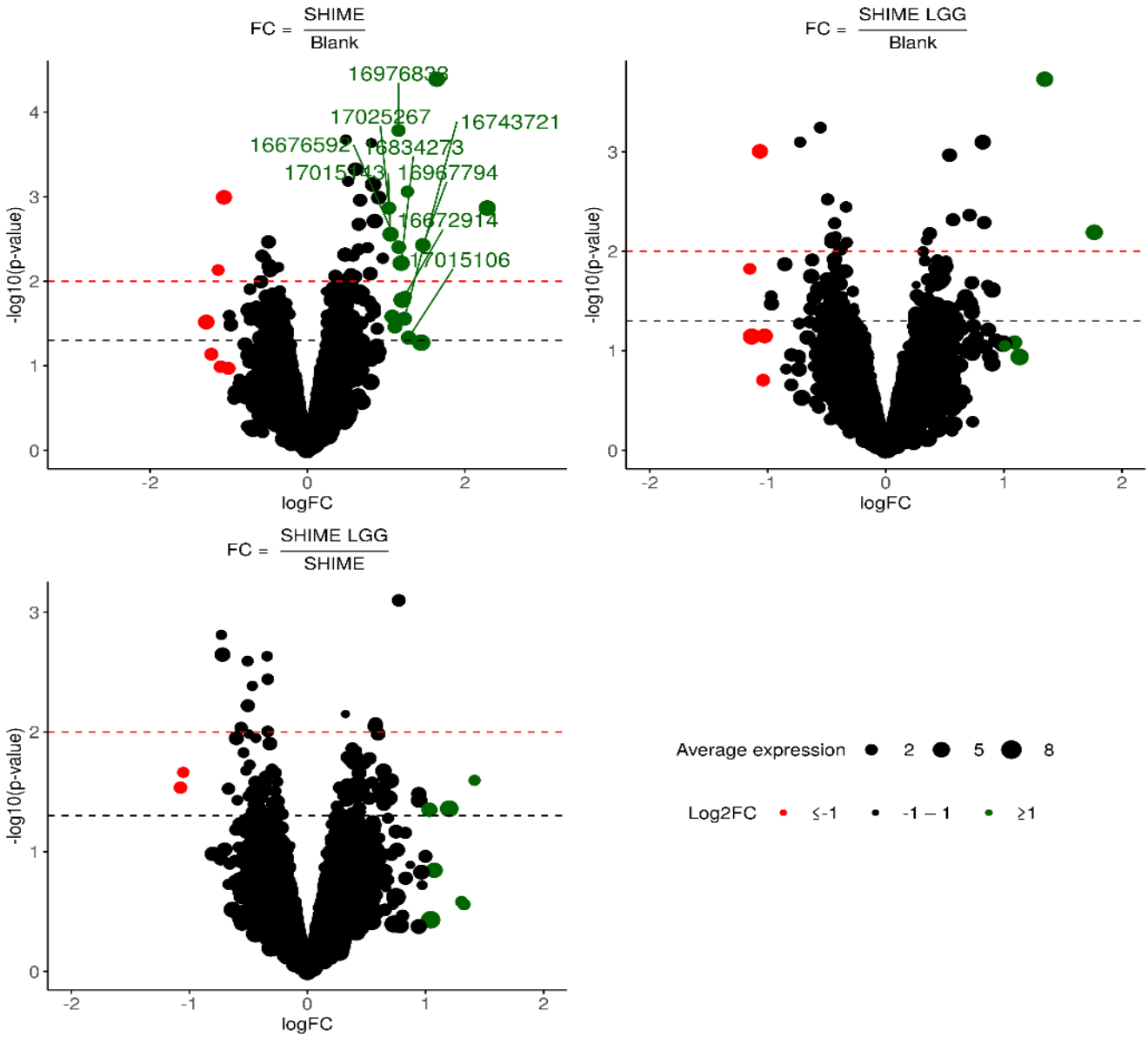

Figure 8.3: Volcano plots of the pre-filtered gene expression data.

A shortlist of top genes considered 'most' differentially expressed (with the lowest adjusted p-values and the most extreme log fold changes) was compiled for further visualization and analysis.

A reduced number of significant genes ($p<0.05$), with a LogFC exceeding 1 was visualized in heatmaps (8.4).

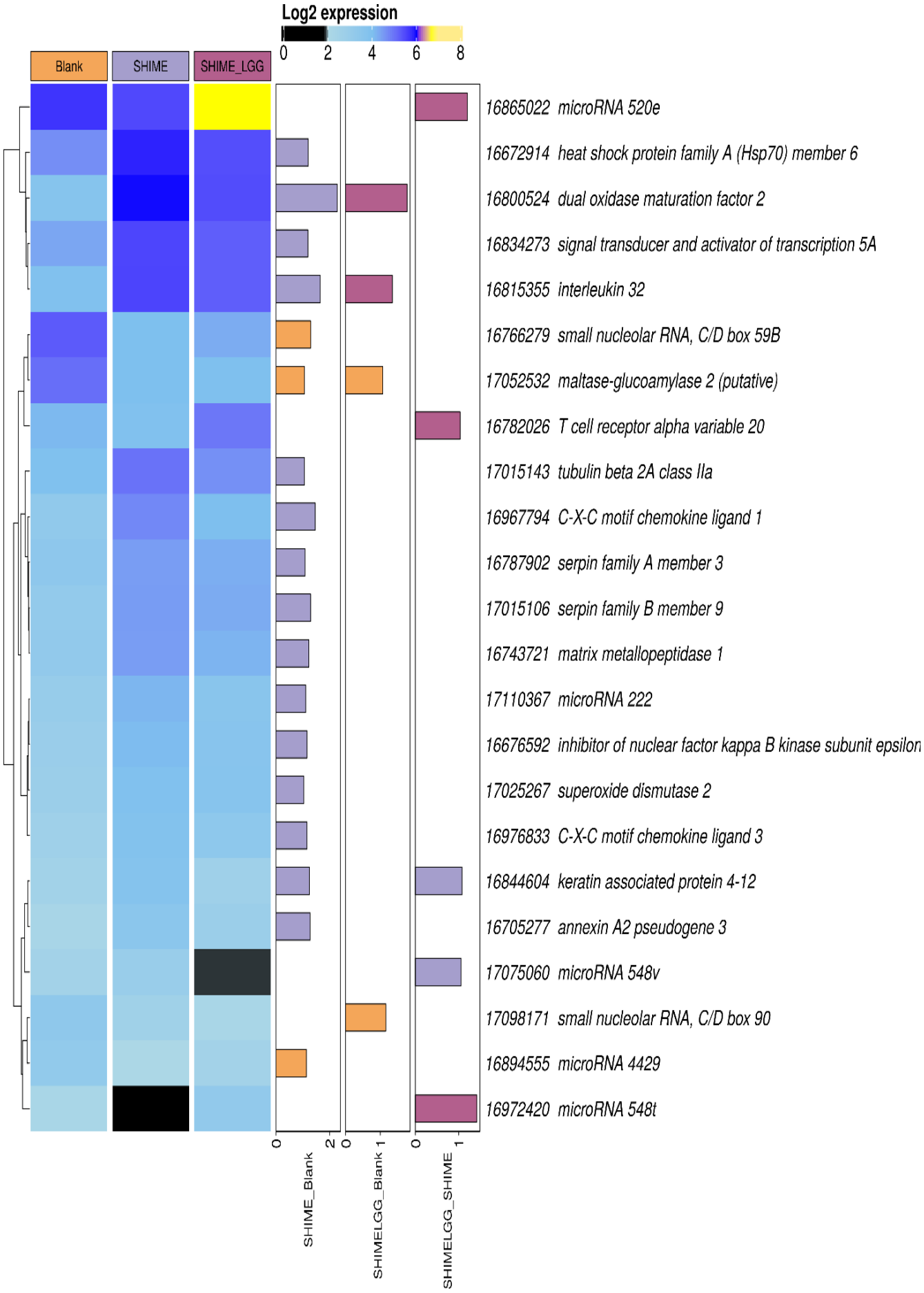

Figure 8.4: Heatmaps displaying the gene expression of significant genes with abs(LogFC) exceeding 1.

Chapter 9

Gene ontology enrichment analysis

The identified significantly differentially expressed genes in response to SHIME or SHIME+LGG exposure are not confined entities and several gene products may positively or negatively interact and can even be involved in a signalling cascade or contribute to the same biological process. For instance, a rapid visual inspection of the differential expression analysis results reveals a number of gene names that relate to a pro-inflammatory response. To automate the detection of functional relations between the retrieved significantly differentially expressed genes, a gene ontology (GO) enrichment analysis was performed with the topGO package version (2.44.0). Essentially the gene ontology (`http://www.geneontology.org/`) is a hierarchically organized collection of gene annotations organized in three domains: Molecular Function, Cellular Component and Biological Process [3].

Molecular function GO terms describe activities performed by individual (or sometimes complexes composed of multiple) gene products (i.e. proteins or RNA) occurring at the molecular level without specifying the location and context in which these activities take place. Cellular Component GO terms refer to the locations relative to cellular structures in which a gene product performs a function, either cellular compartments (e.g., mitochondrion), or stable macromolecular complexes of which they are parts (e.g., the ribosome). Biological Process GO terms, finally, delineate the larger processes, or 'biological programs' accomplished by multiple molecular activities. This is not equivalent to pathways since the dynamics and dependencies between genes involved in the same biological process are not specified by gene ontology.

Gene ontology terms (BP,CC,MF) were build based on the GOTERM environment from package GO.db version 3.13.0 and extracted from the gene-to-GO mapping for the Affymetrix Human Gene 2.1 ST Array Strip (hugene20sttranscriptcluster.db) using annFUN.db.

To find gene ontology terms that are enriched in the significantly differentially expressed genes, a background set of non-differentially expressed genes with similar average expression strength is defined using the genefinder function from the genefilter package version 1.74.0. The background was shown to have roughly the same distribution as the significant genes of interest (Figure 9.1). Due to the prior filtering step, the distribution was even similar for all genes retained after DABG filtering used as input for Limma based differential expression hypothesis testing (See section 8.3). This in contrast to the unfiltered gene population which is clearly enriched in unexpressed genes (expressions $\leq$ background probe intensities).

Enrichment of gene ontology terms of interest (BP, MF, CC) in the significantly differentially expressed genes (Limma p-value ¡0.01, see **??**) compared to the background genes was determined

with Fisher's exact tests for every GO category independently and Kolmogorov-Smirnov tests in combination with a more complex elim algorithm which take the hierarchical GO dependencies into account. The hierarchical gene ontology structure was pruned to remove terms with less than 10 annotated genes.

Over-represented GO terms were exported in a supplementary excel file and visualized in a GO graph displaying the most significant GO terms (rectangles) and their distribution in the GO hierarchy (Figure 9.2- Figure 9.3).

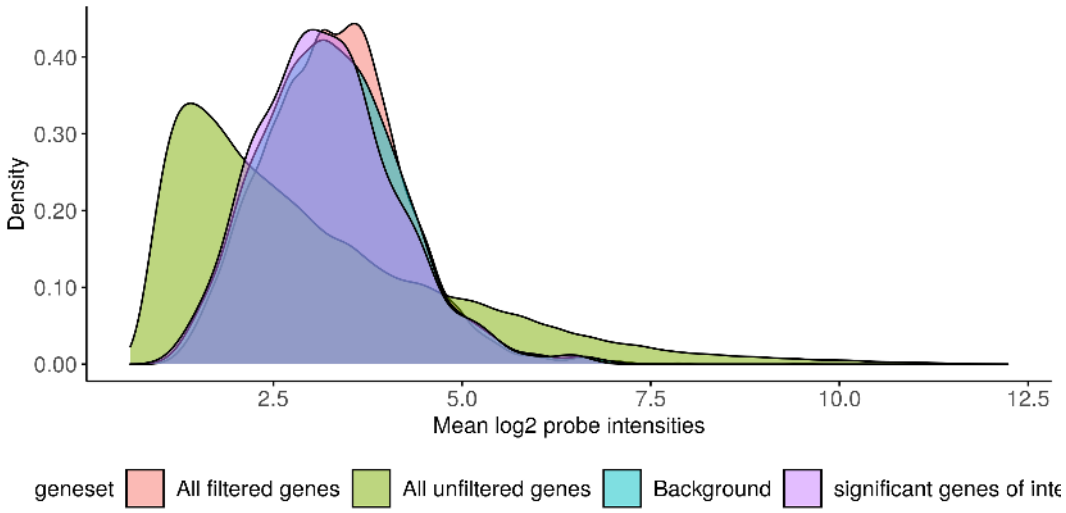

Figure 9.1: Multidensity plot of the average expression strength of the significantly differentially expressed genes of interest, a background set of non-differentially expressed genes with similar average expression strength is defined using the genefinder function from the genefilter package version .

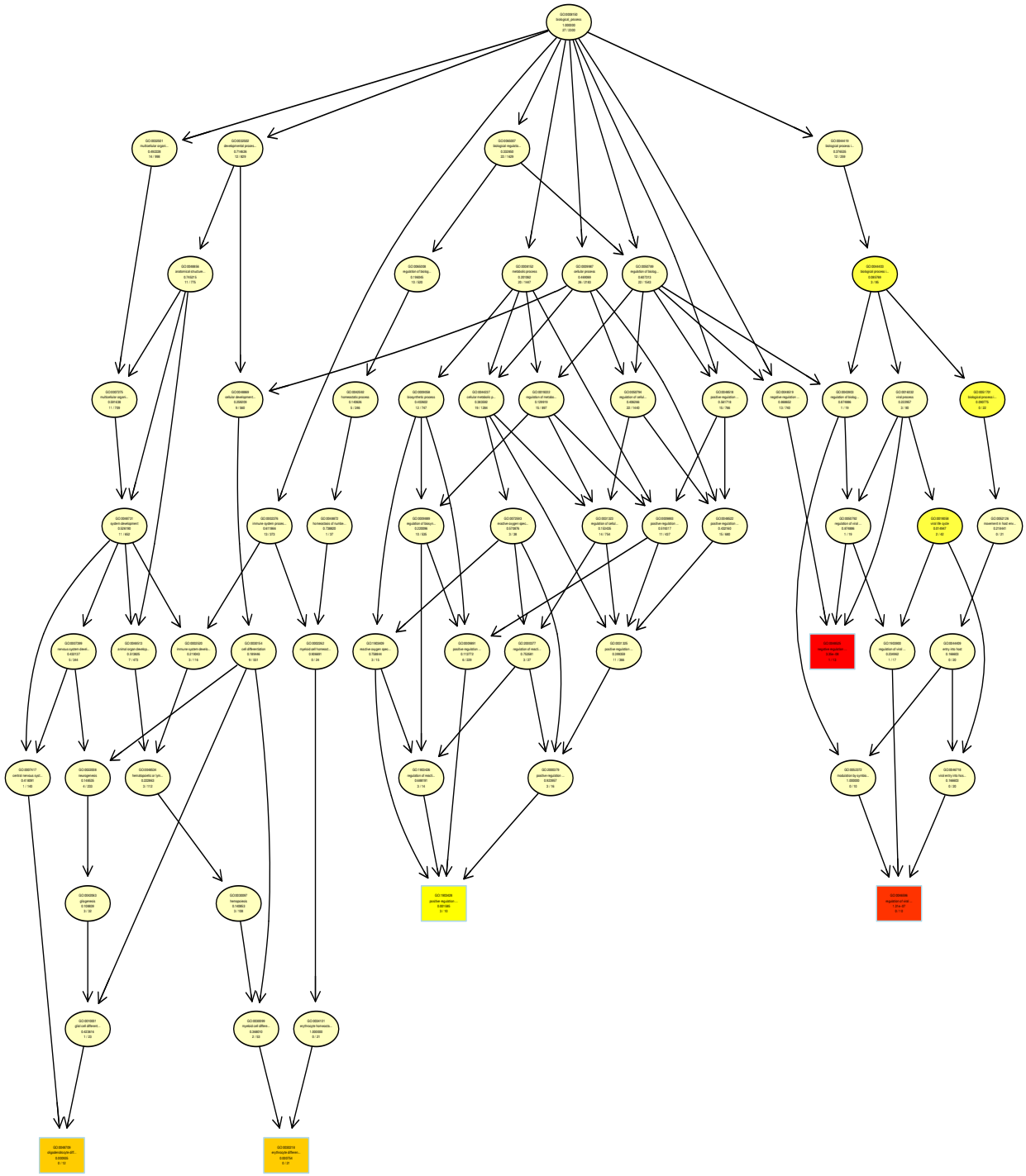

Figure 9.2: GOgraph of the 5 most significant GO BP terms, displayed as square boxes. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). Black arrows indicate is-a relationships and red arrows part-of relationships. The first two lines show the GO BP identifier and a trimmed GO BP name. In the third line the raw p-value is shown. The forth line is showing the number of significant genes and the total number of genes annotated to the respective GO BP term.

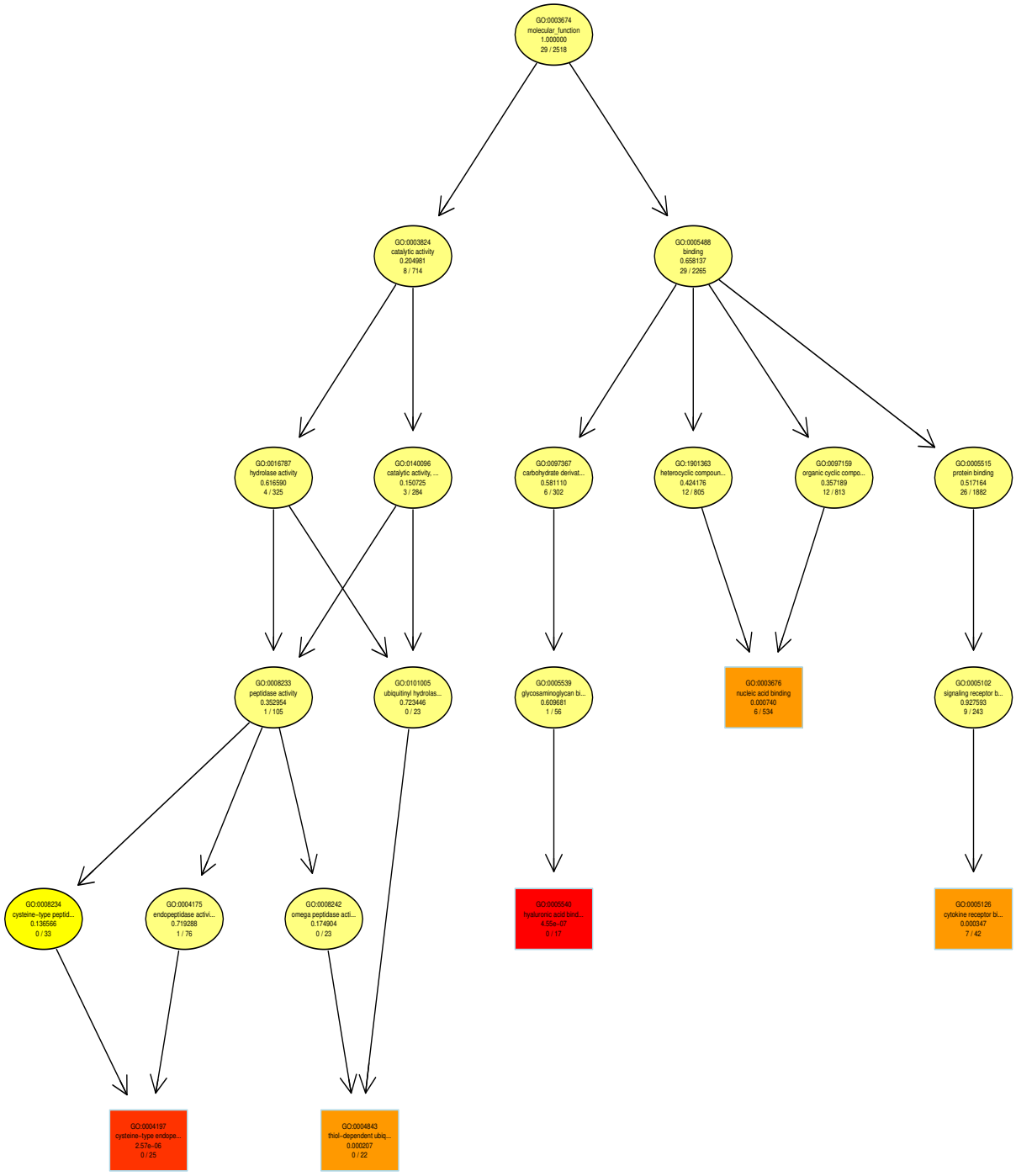

Figure 9.3: GOgraph of the 5 most significant GO MF terms, displayed as square boxes. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). Black arrows indicate is-a relationships and red arrows part-of relationships. The first two lines show the GO MF identifier and a trimmed GO MF name. In the third line the raw p-value is shown. The forth line is showing the number of significant genes and the total number of genes annotated to the respective GO MF term.

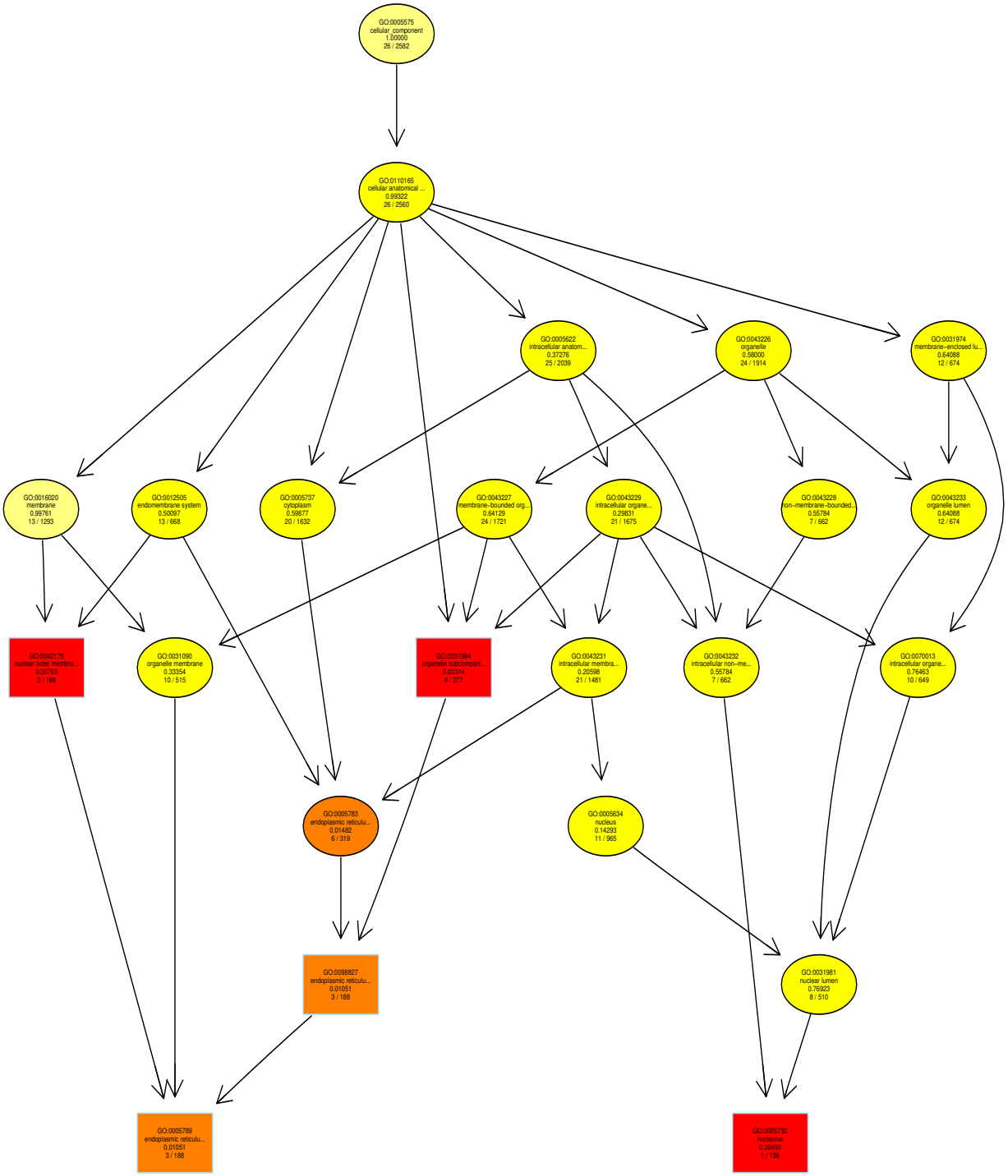

Figure 9.4: GOgraph of the 5 most significant GO CC terms, displayed as square boxes. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). Black arrows indicate is-a relationships and red arrows part-of relationships. The first two lines show the GO CC identifier and a trimmed GO CC name. In the third line the raw p-value is shown. The forth line is showing the number of significant genes and the total number of genes annotated to the respective GO CC term.

Chapter 10

KEGG pathway and gene ontolgy based gene-set analysis

Just like GO enrichment analysis, gene-set analysis (GSA) takes into account functional relationships between genes. But unlike GO enrichment analysis which starts from the results of a prior per-gene differential expression analysis, GSA focuses on sets of related genes and aims to identify significantly up- or downregulated gene-sets. Such coordinated differential expression analysis over gene set offers a number of advantages compared to classical individual gene approaches, including greater robustness, sensitivity and biological relevance [5]. This is especially true considering the functional redundancy of individual genes and the level of noise in high-throughput microarray data. Moreover, consistent perturbations over such gene sets frequently suggest mechanistic changes.

Gene-sets are pre-defined groups of genes, which are functionally related. Functional relationships can be derived from KEGG pathways, Gene Ontology terms, gene groups that share some other functional annotations, including common transcriptional regulators (like transcription factors, small interfering RNA's or siRNA, genomic locations etc.

GSA was performed with up-to-date (Sun Jul 4 14:43:35 2021) KEGG pathway and gene ontology based gene-sets in combination with a parametric gene randomization procedure implemented in the GAGE (generally applicable gene set enrichment for pathway analysis) package, version 2.42.0.

Prior to GSA in GAGE, Affymetrix gene identifiers were mapped to the matching Entrez identifiers used in the KEGG and GO databases. One-directional (either up- or down-regulation) expression changes were assessed with parametric unpaired two-sample t-tests on gene sets using LogFC as per-gene statistics, followed by robust p-value summarization using Stouffer's method and Benjamini & Hochberg FDR correction (q-value). The average of the individual LogFC statistics from multiple single array based gene-set tests was displayed in bar graphs for the significant KEGG pathways (Figure) and gene ontology terms (Figure) (p¡0.01). FDR adjusted p-values were reported and bars were color-coded according to the pathway/gene ontology categories.

There are frequently multiple significant gene-sets that share multiple member genes or represent the same regulatory mechanism. Therefore, we identified non-redundant gene-sets, not displaying significant overlap with other gene-sets in the core genes contributing most to the gene-set overall significance. The core genes were identified in the volcanoplots obtained with Limma and the LogFC in individual genes confirmed the general trends at higher levels (pathways, gene

ontology). The non-redundant KEGG-based gene-sets were also explored with pathview version 1.32.0. The perturbed expression patterns were mapped onto the KEGG pathway view [5].

10.1 KEGG pathway analysis

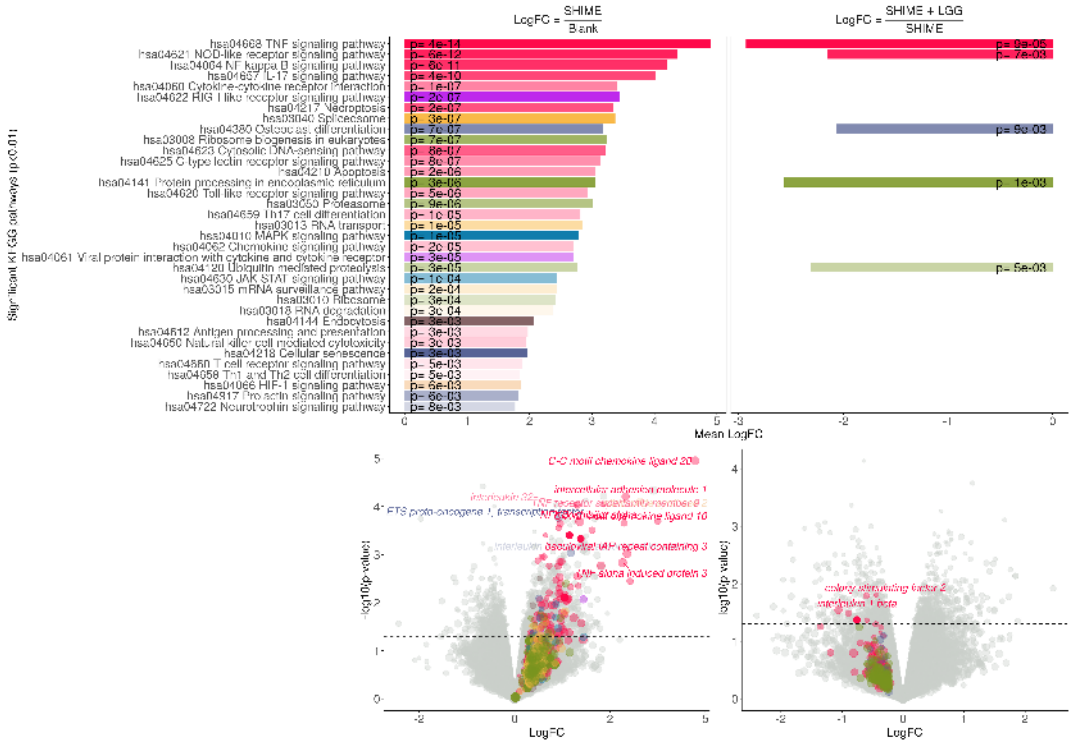

Figure 10.1: Significantly (p<0.01) up- (positive LogFC) and downregulated (negative LogFC) KEGG pathways as assessed by gene-set analysis (GSA) in GAGE. P-values are adjusted for multiple testing with Benjamini & Hochberg FDR correction. Bars are colored by the pathway category: pro-inflammatory response (pink), viral interaction (purple), RNA (green) and protein (yellow) processing, signaling patways (teal), cellular transport (brown) and cellular differentiation (blue).

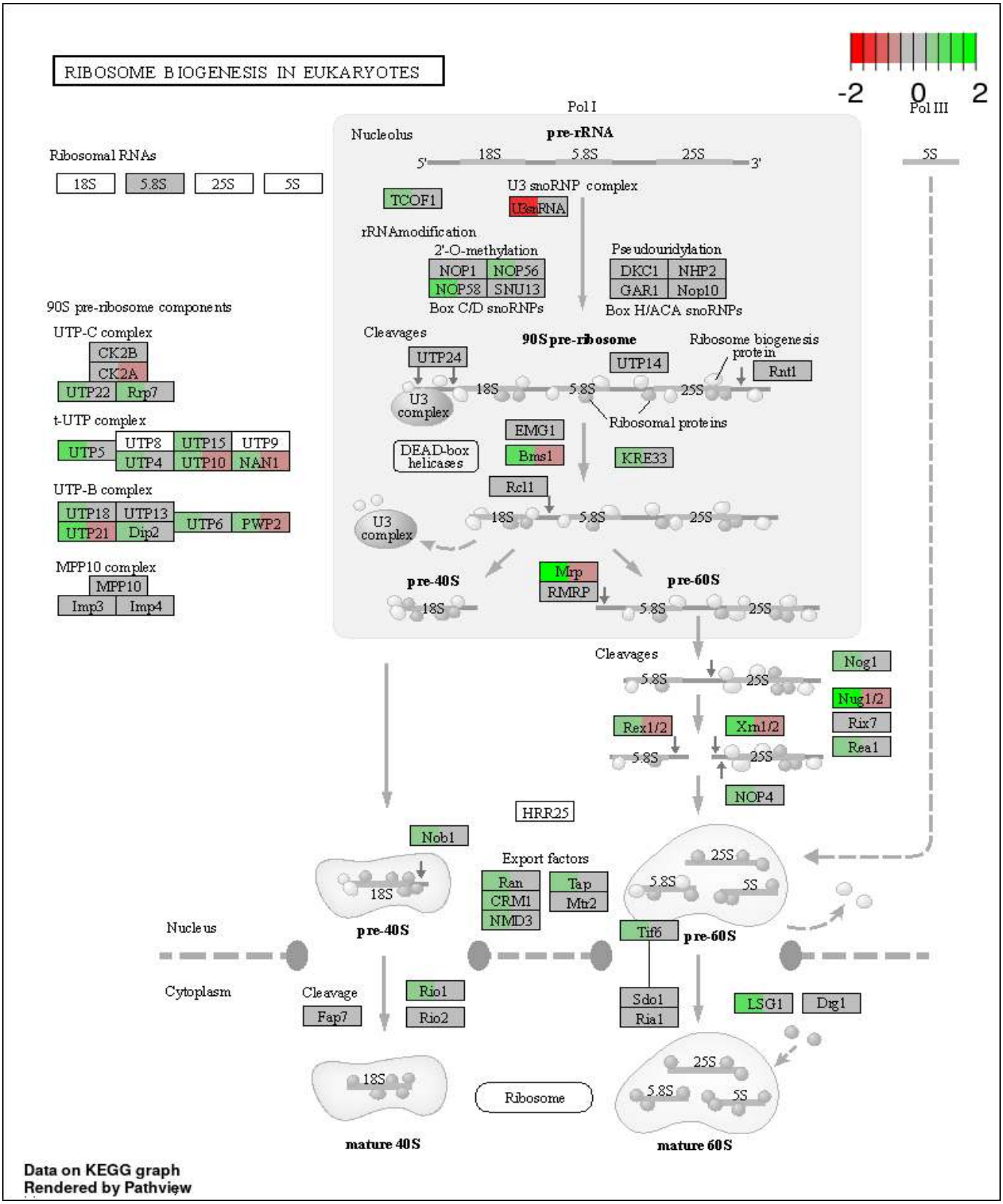

Figure 10.2: KEGG pathview of the ribosome biogenesis pathway in eukayrotes, which was significantly (Adjusted p-value = 2e-06) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

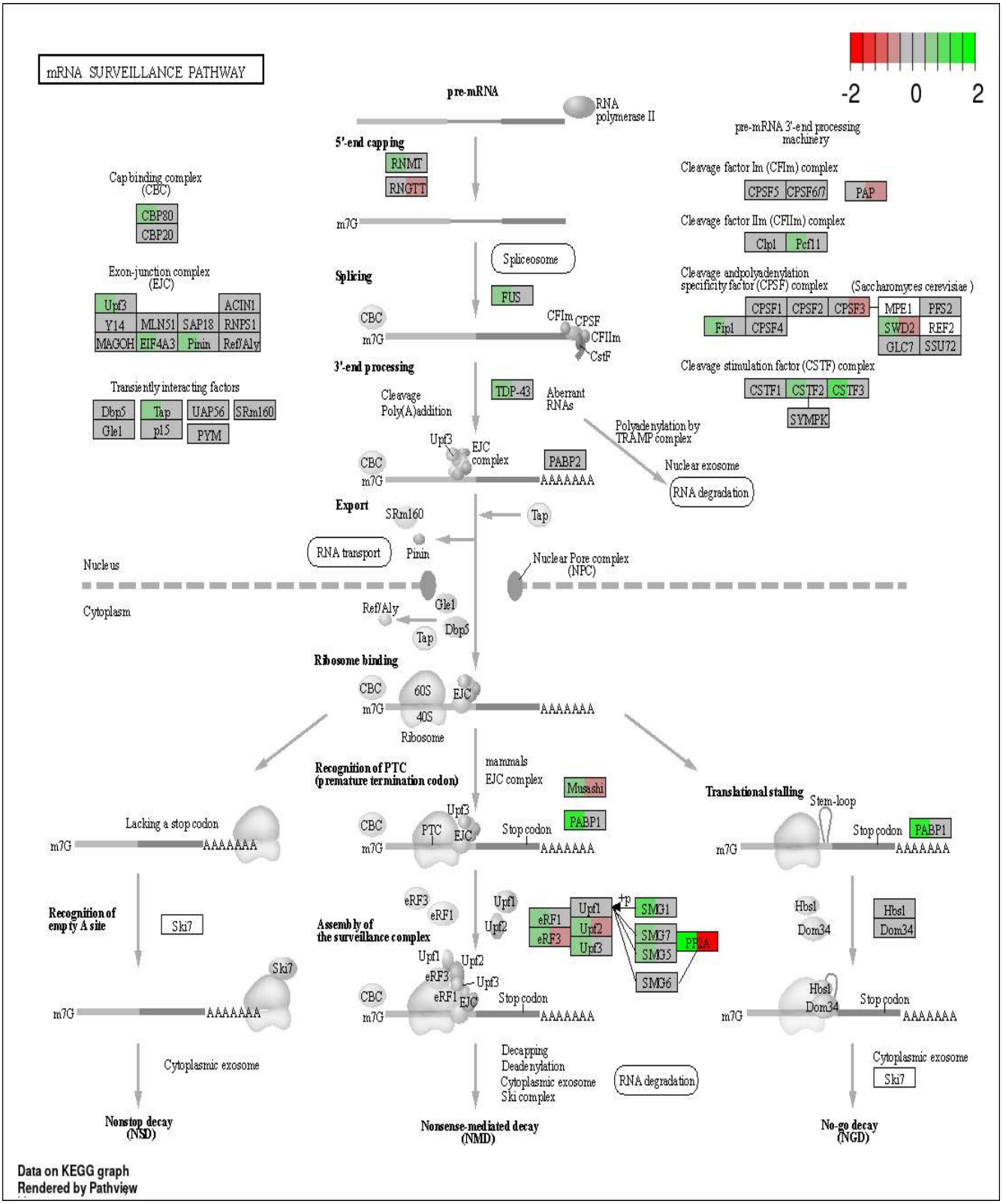

Figure 10.3: KEGG pathview of the mRNA surveillance pathway, which was significantly (Adjusted p-value = 0.00026) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

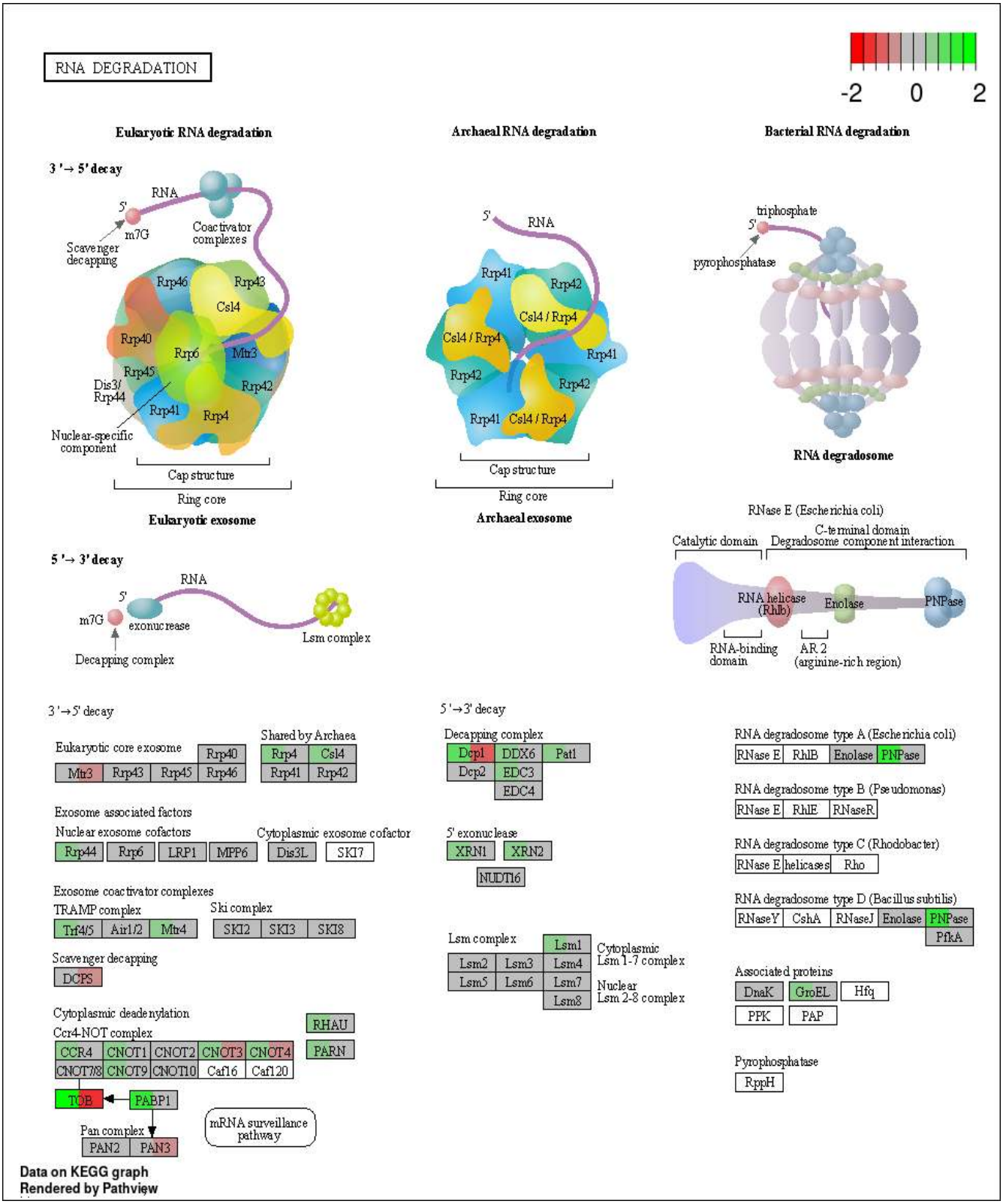

Figure 10.4: KEGG pathview of the RNA degradation pathway, which was significantly (Adjusted p-value = 0.00031) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

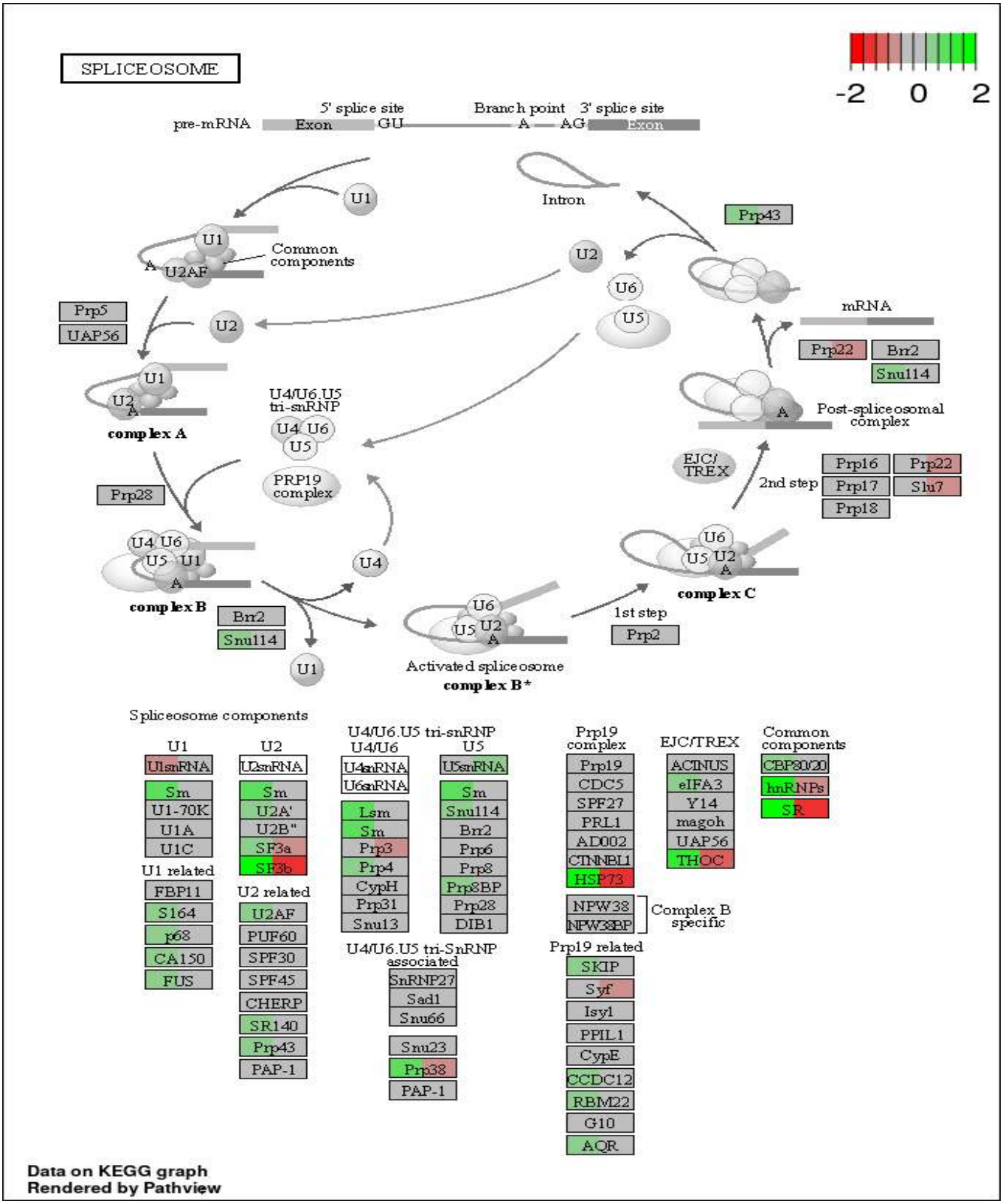

Figure 10.5: KEGG pathview of the spliceosome pathway, which was significantly (Adjusted p-value = 4.1e-07) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

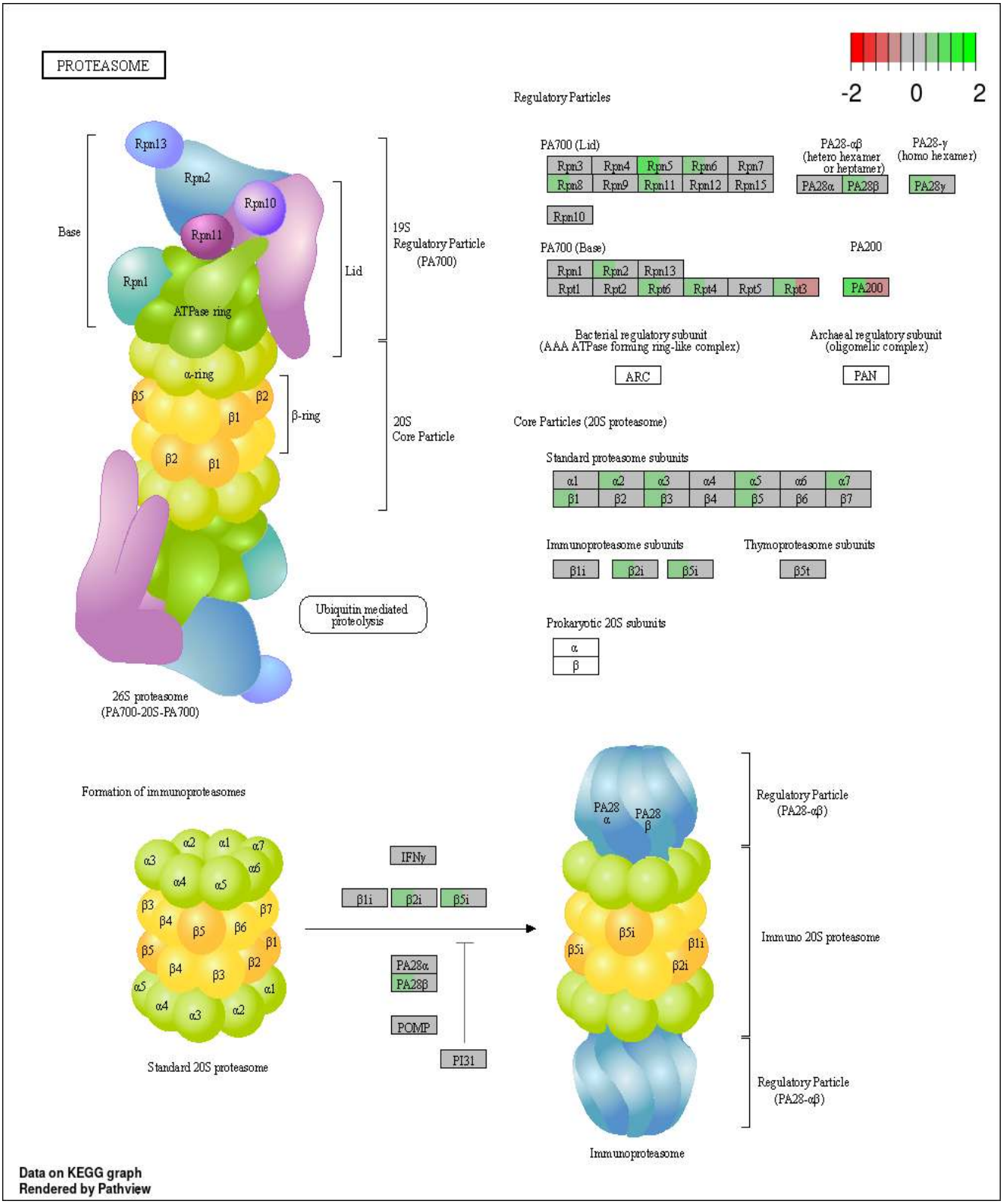

Figure 10.6: KEGG pathview of the proteasome pathway, which was significantly (Adjusted p-value = 8.3e-06) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

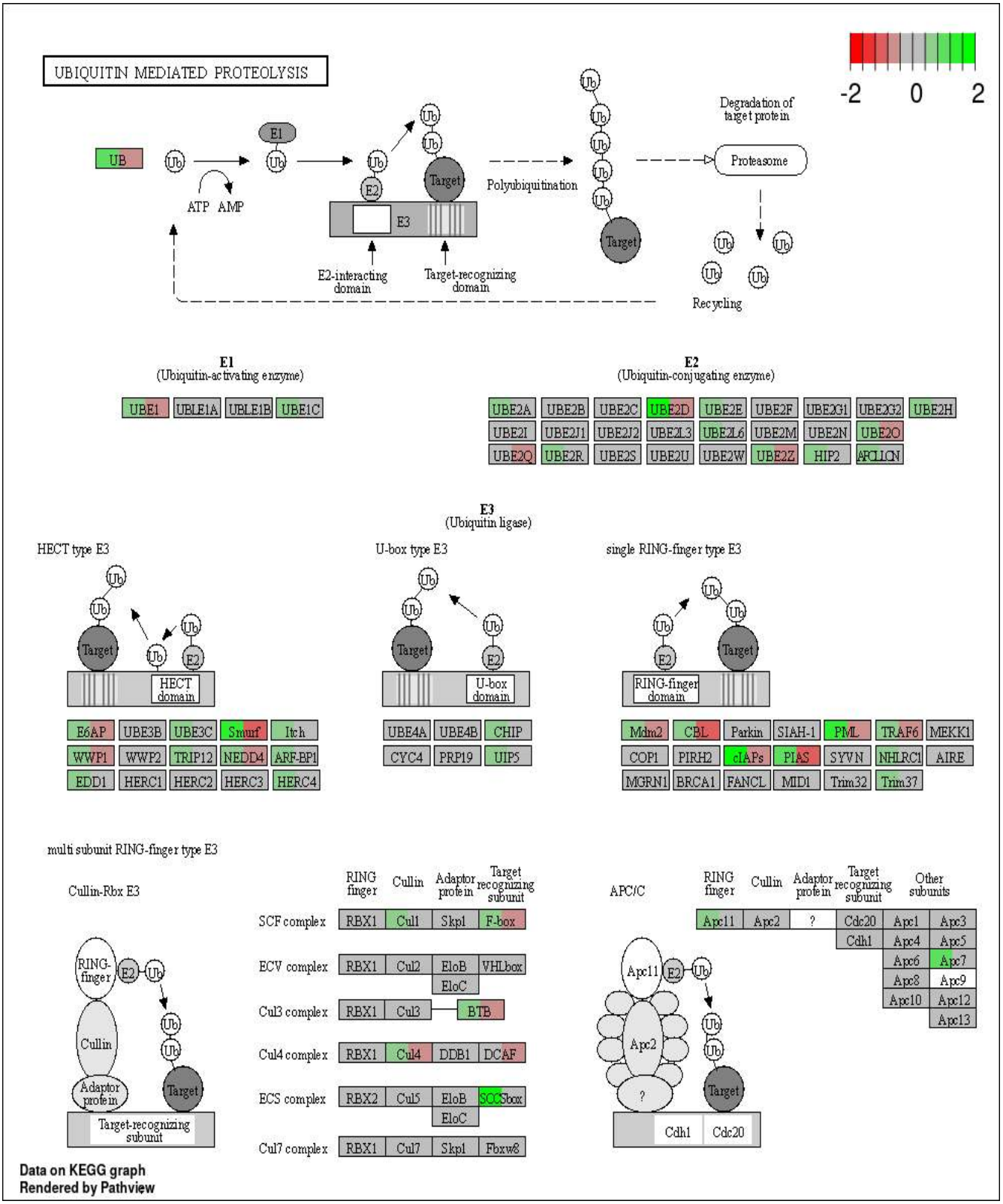

Figure 10.7: KEGG pathview of the ubiquitin mediated proteolysis pathway, which was significantly (Adjusted p-value = 2.7e-05) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition significantly (4.5e-03) toned down the SHIME effect (grey and red coloring).

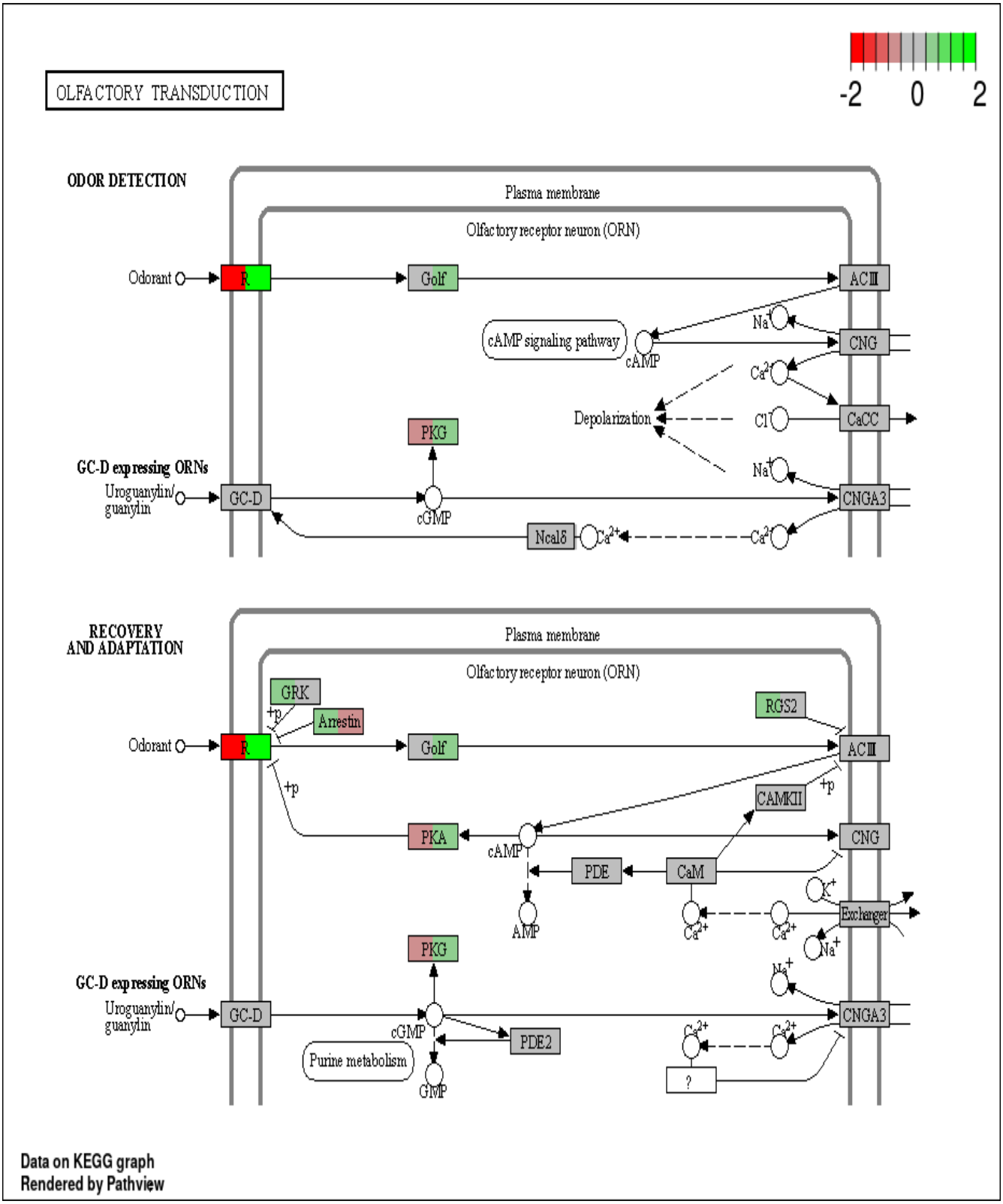

Figure 10.8: KEGG pathview of the olfactory transduction pathway, which was upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

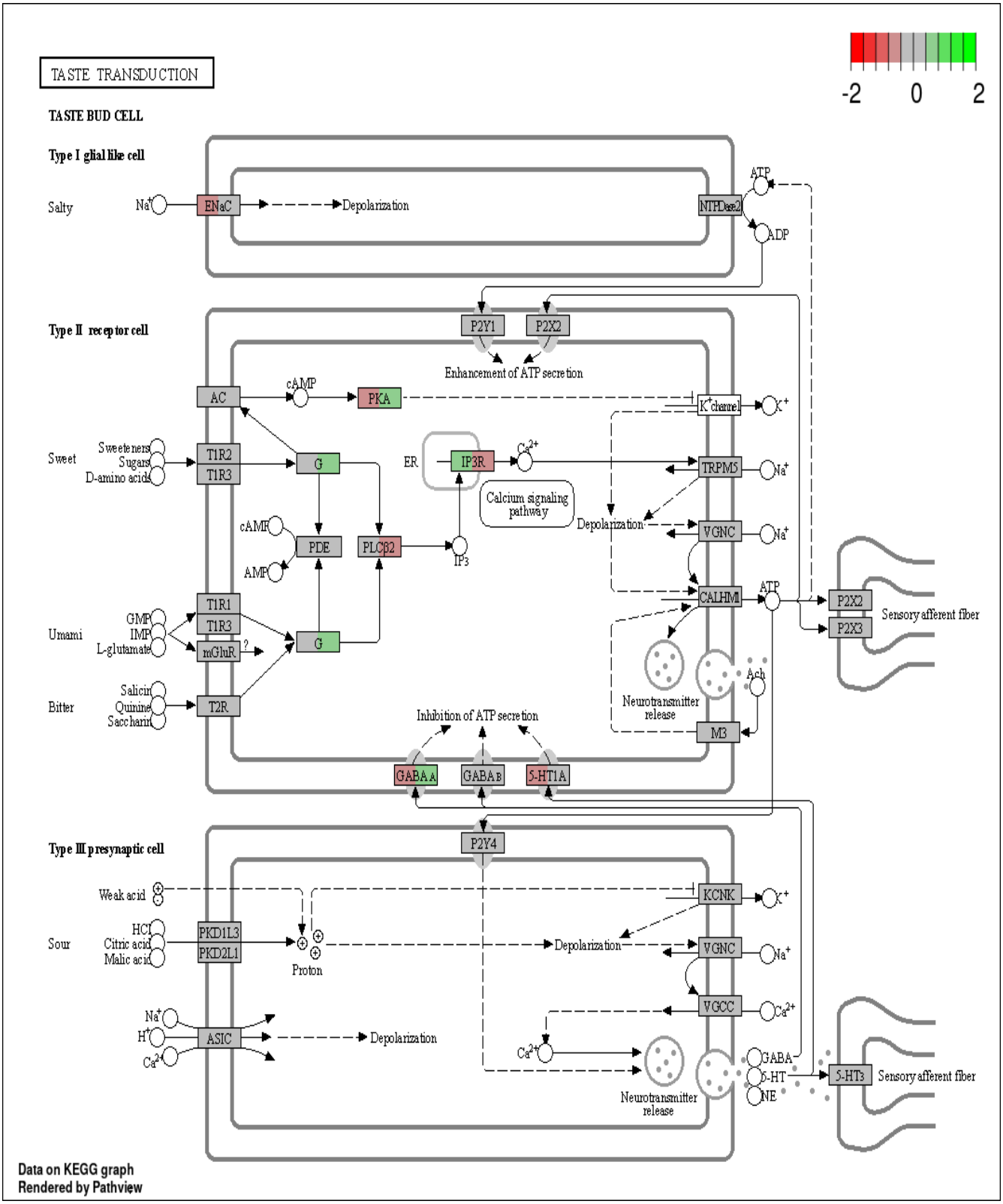

Figure 10.9: KEGG pathview of the taste transduction pathway, which was upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

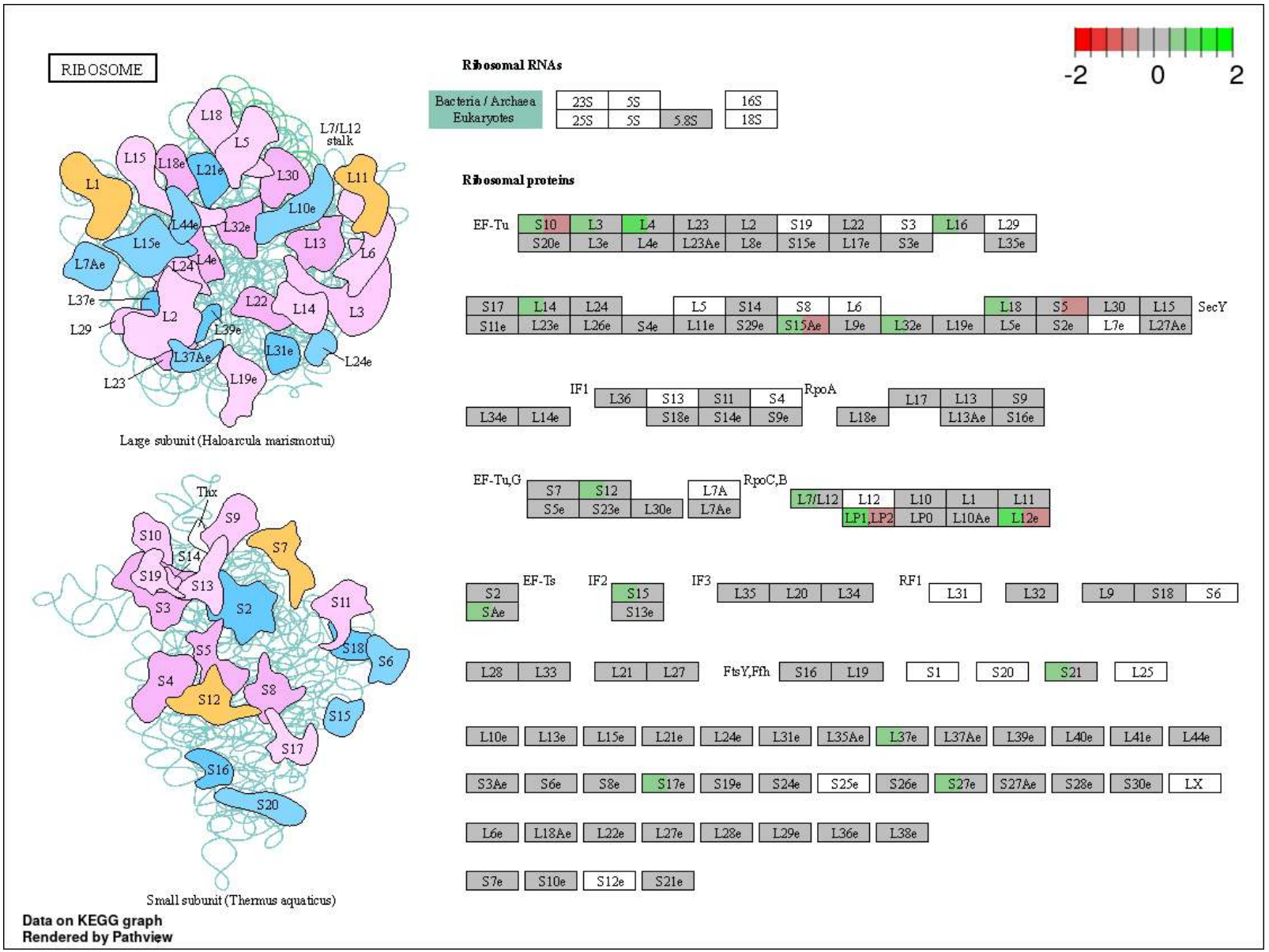

Figure 10.10: KEGG pathview of the ribosome pathway, which was significantly (Adjusted p-value = 0.00022) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

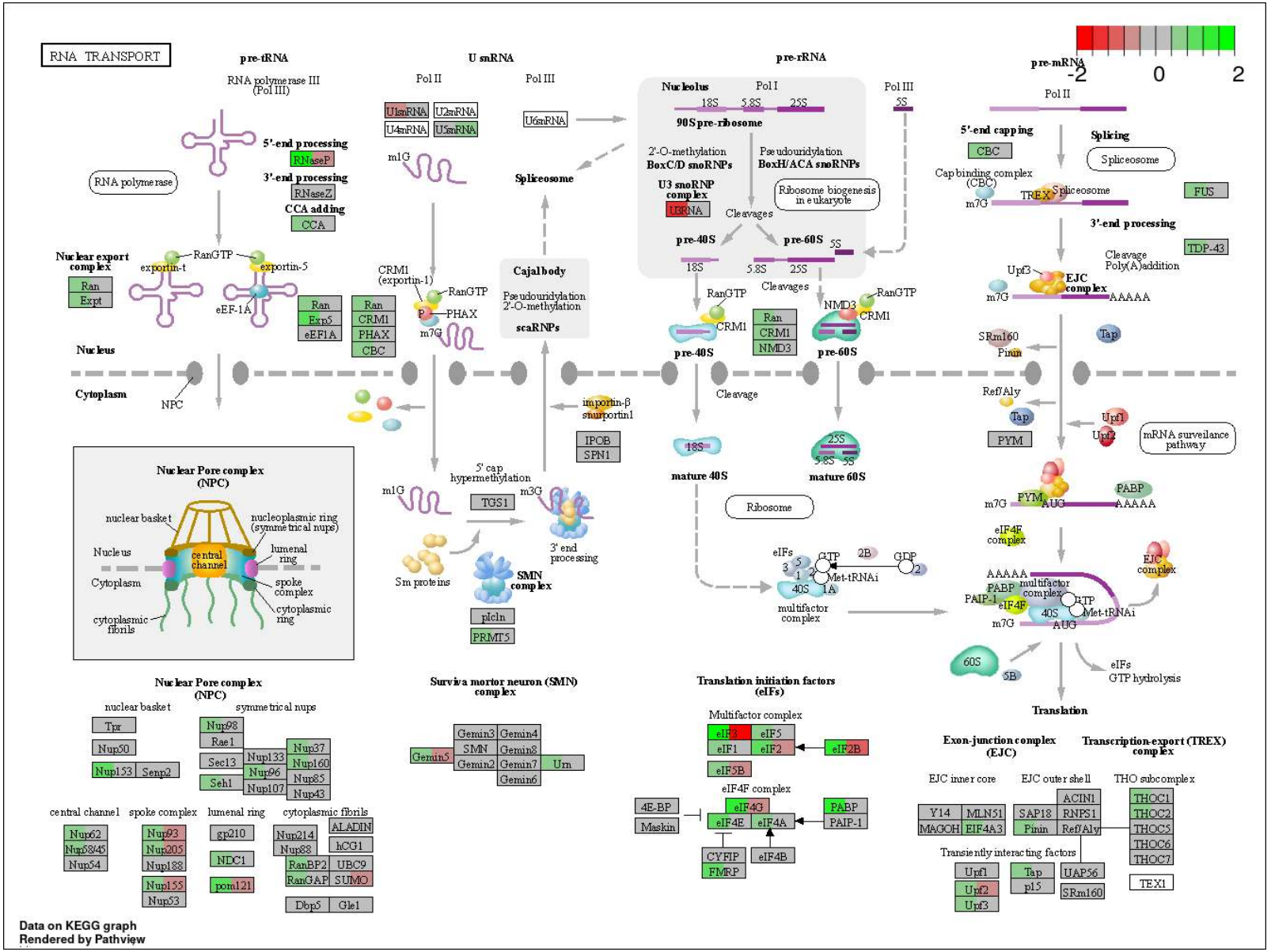

Figure 10.11: KEGG pathview of the RNA transport pathway, which was significantly (Adjusted p-value = 1e-05) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

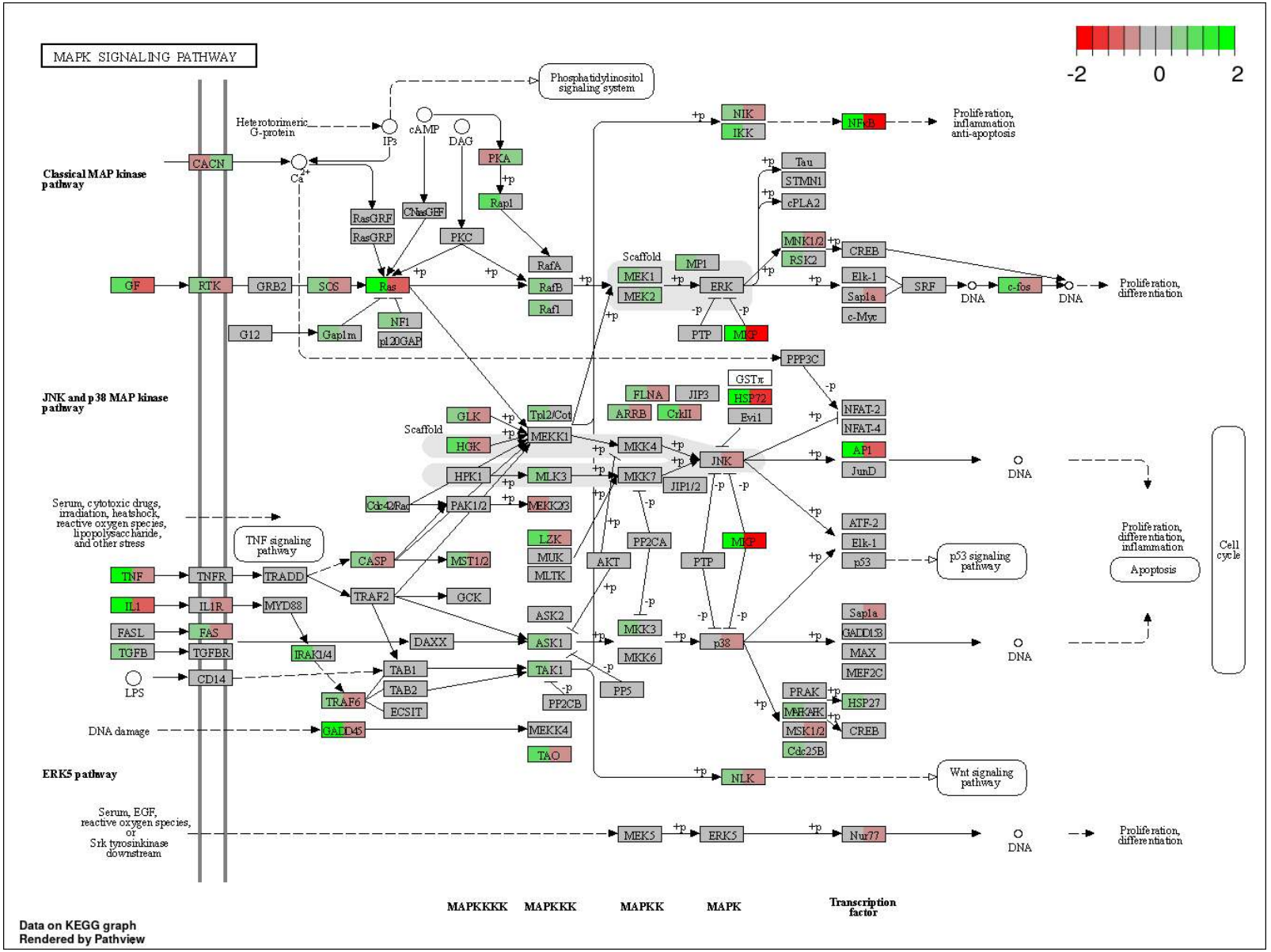

Figure 10.12: KEGG pathview of the MAPK signaling pathway, which was significantly (Adjusted p-value = 8.5e-06) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition significantly (p = 9.9e-03) toned down the SHIME effect (grey and red coloring).

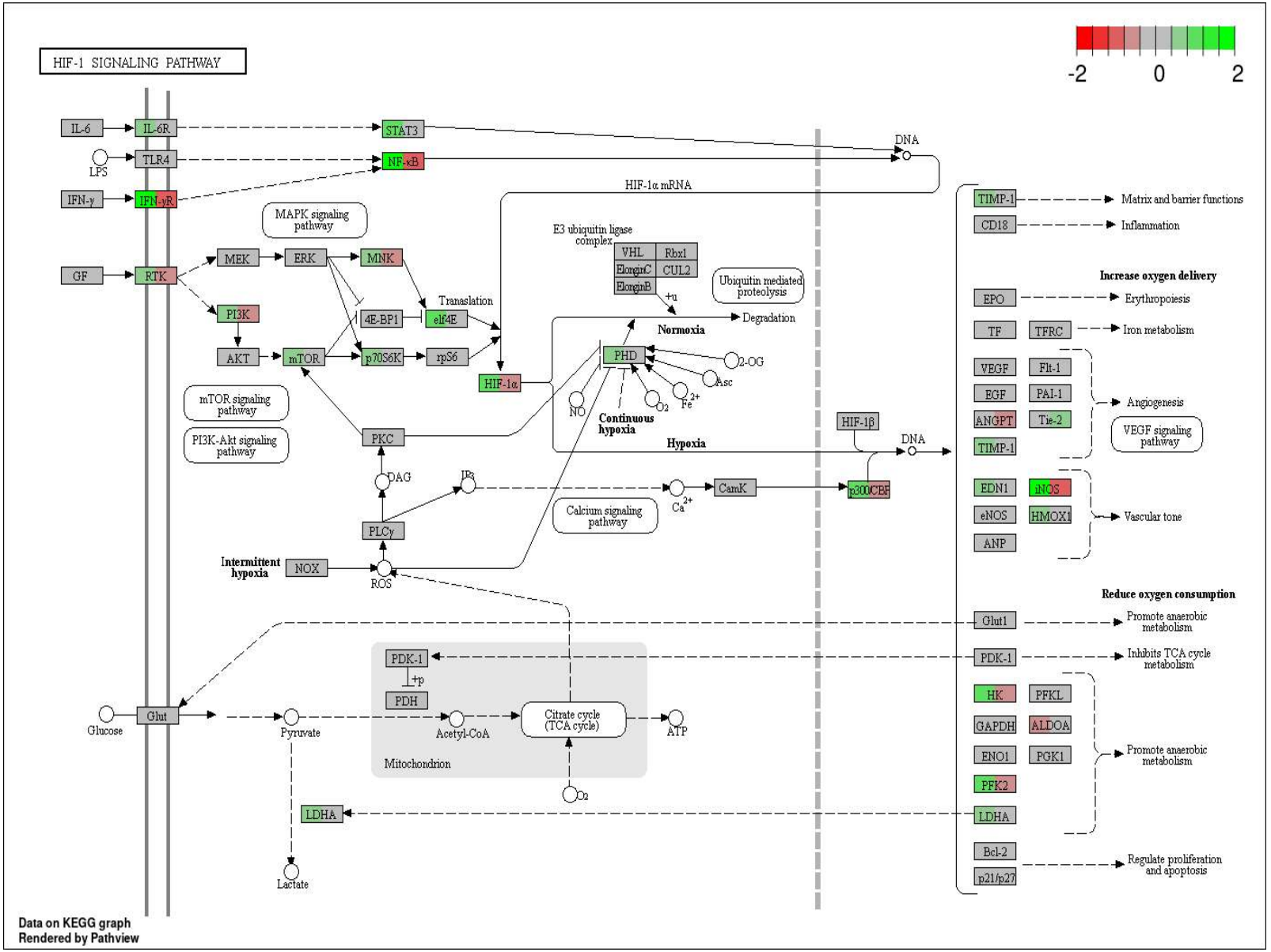

Figure 10.13: KEGG pathview of the HIF-1 signaling pathway, which was significantly (Adjusted p-value = 0.0052) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

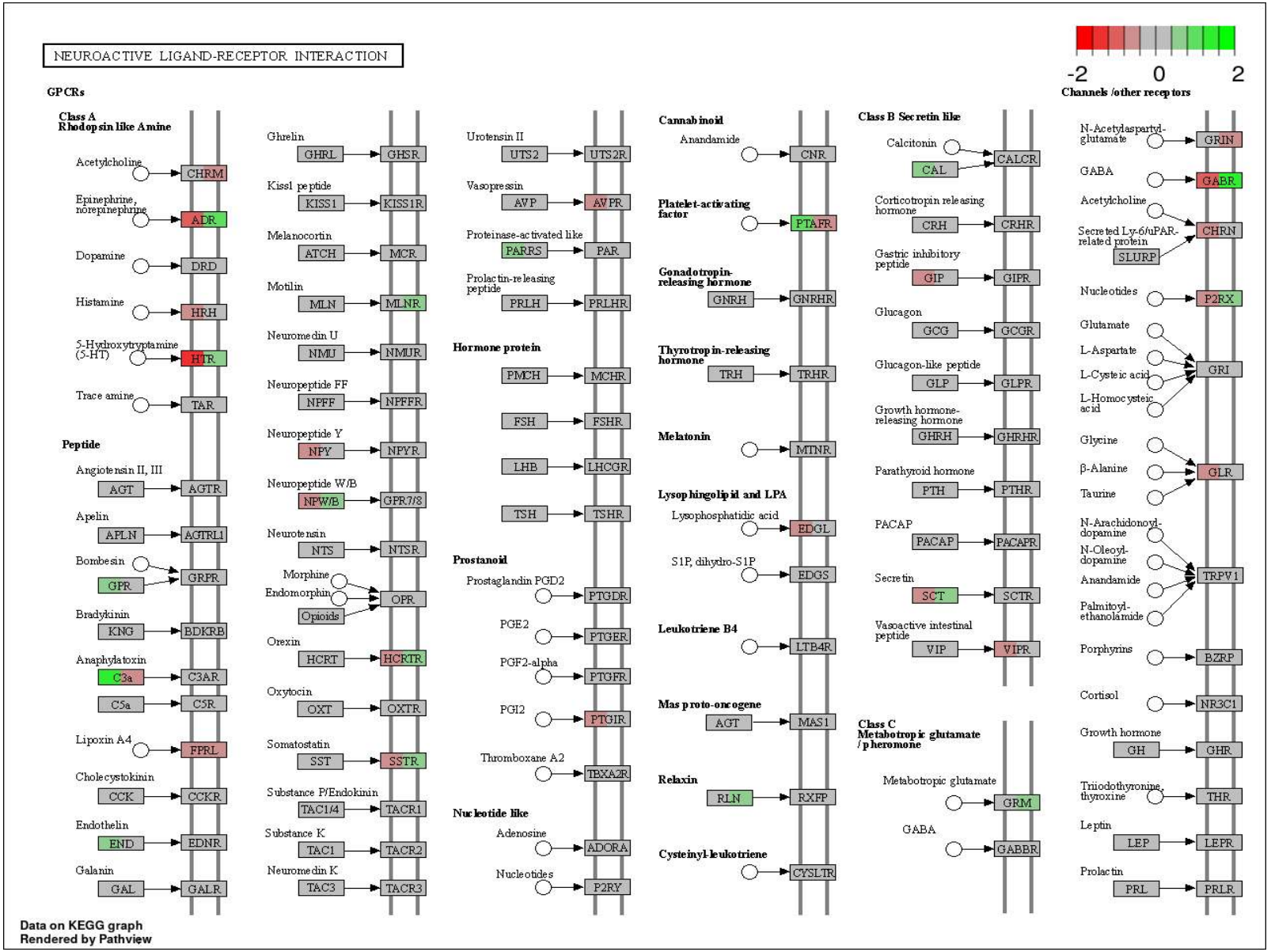

Figure 10.14: KEGG pathview of the neuroactive ligand-receptor pathway, which was upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

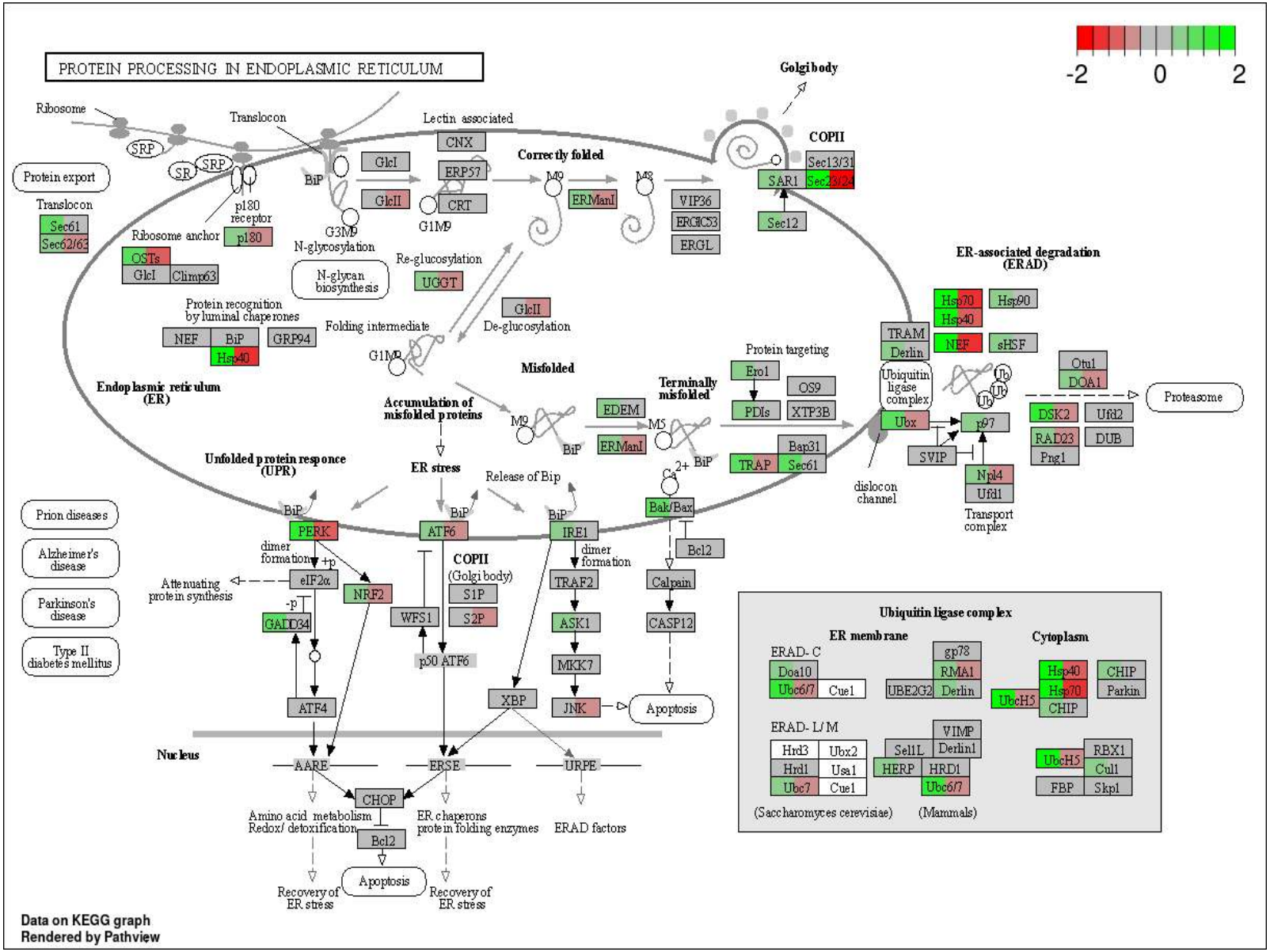

Figure 10.15: KEGG pathview of the protein processing in endoplasmic reticulum, which was significantly (2.0e-06) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition significantly (p = 6.6e-04) toned down the SHIME effect (grey and red coloring).

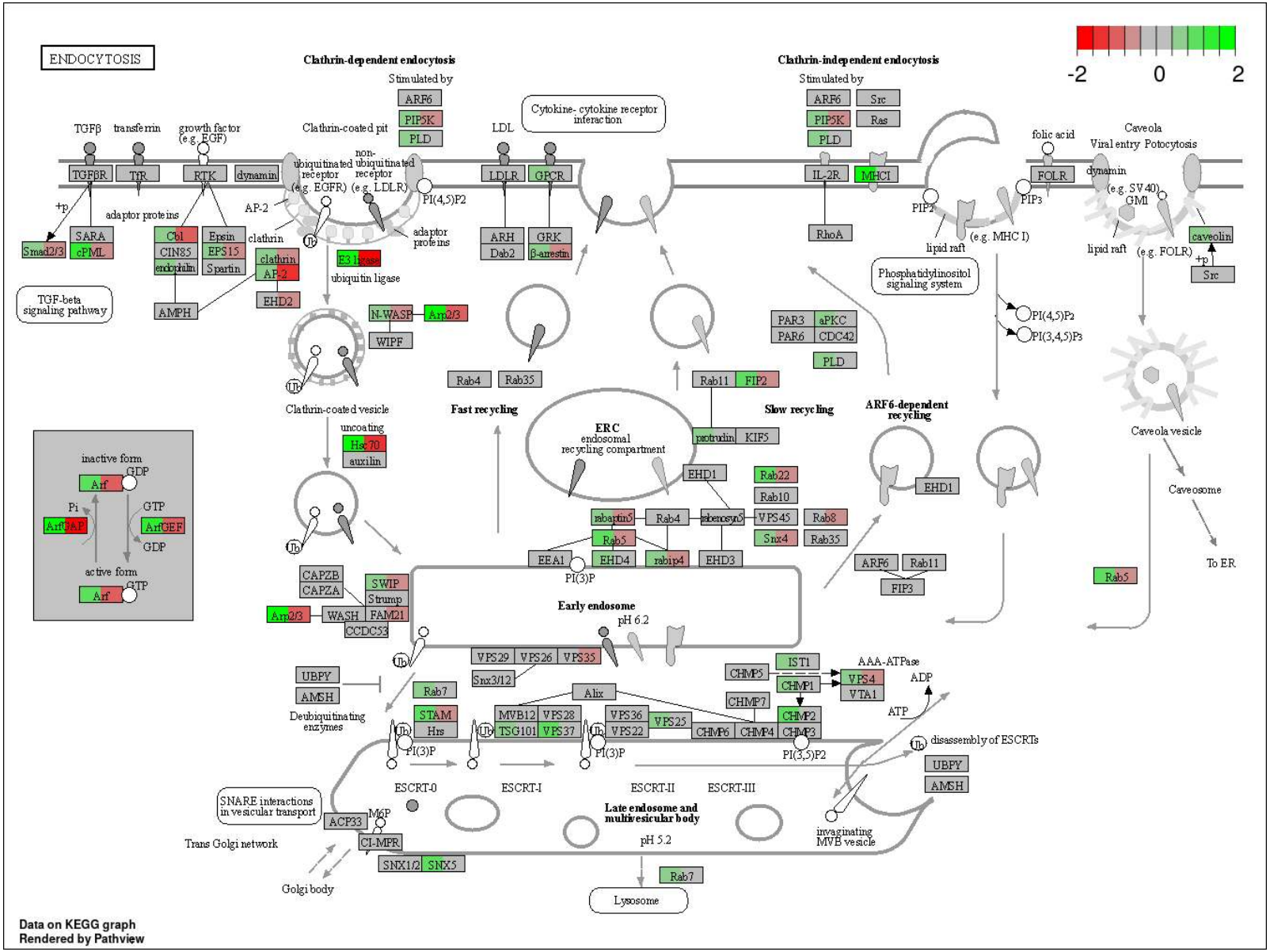

Figure 10.16: KEGG pathview of endocytosis, which was significantly (0.00024) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition significantly ($p = 0.00046$) toned down the SHIME effect (grey and red coloring).

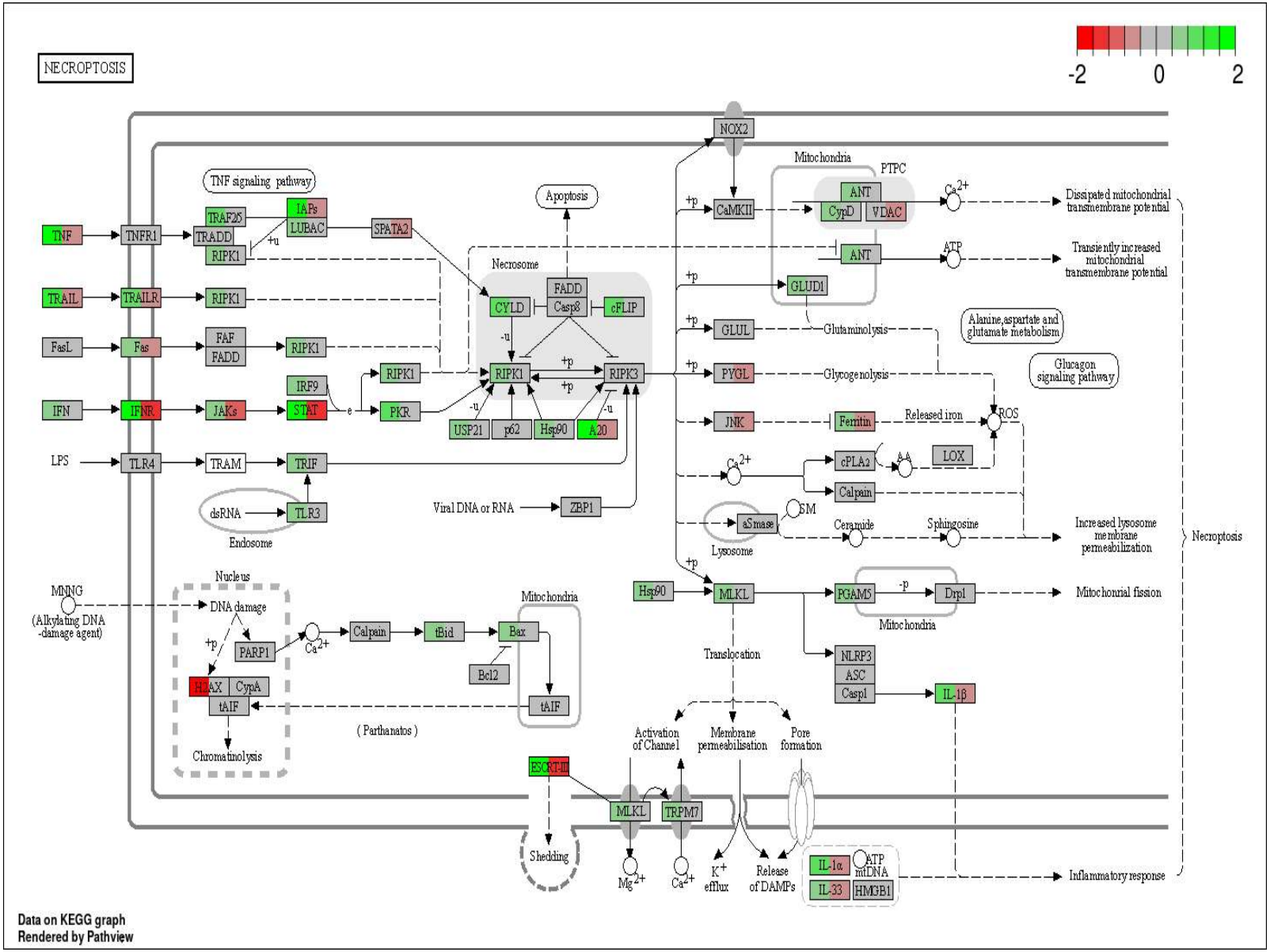

Figure 10.17: KEGG pathview of necroptosis, which was significantly (1.2e-07) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition significantly ($p = 7.4e-03$) toned down the SHIME effect (grey and red coloring).

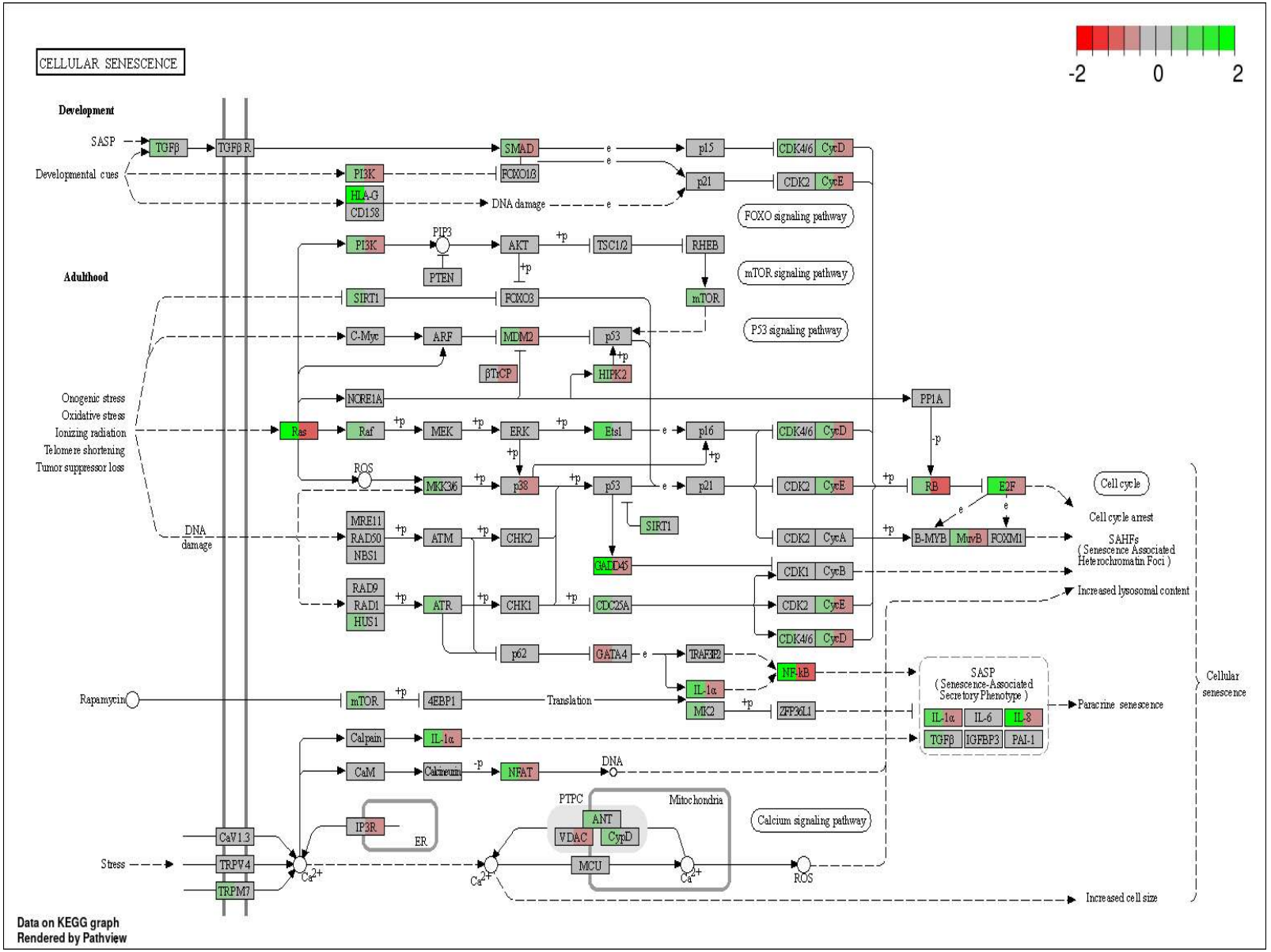

Figure 10.18: KEGG pathview of the cellular senescence pathway, which was significantly (0.0029) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

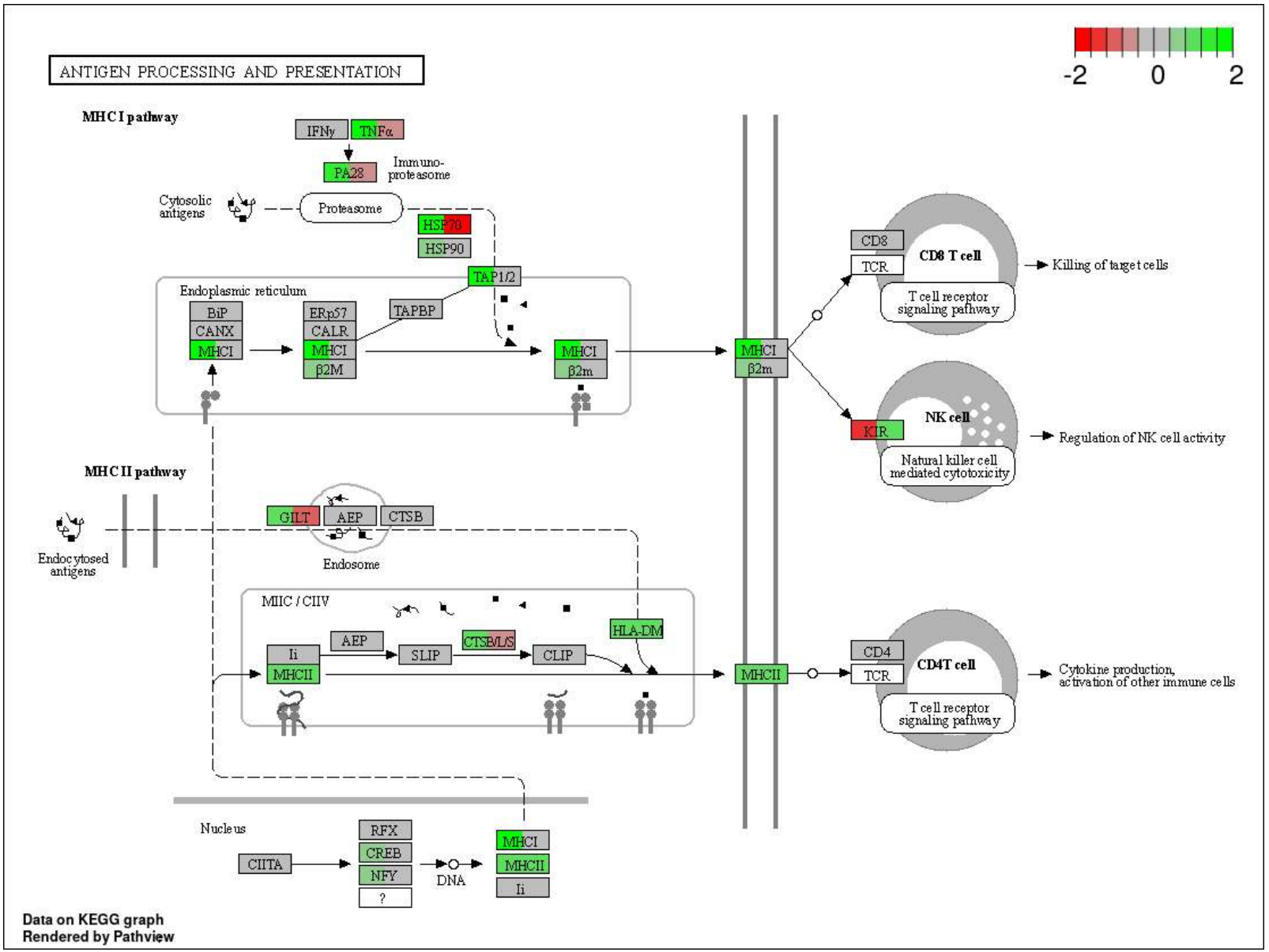

Figure 10.19: KEGG pathview of the antigen processing and presentation pathway, which was significantly (0.0028) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

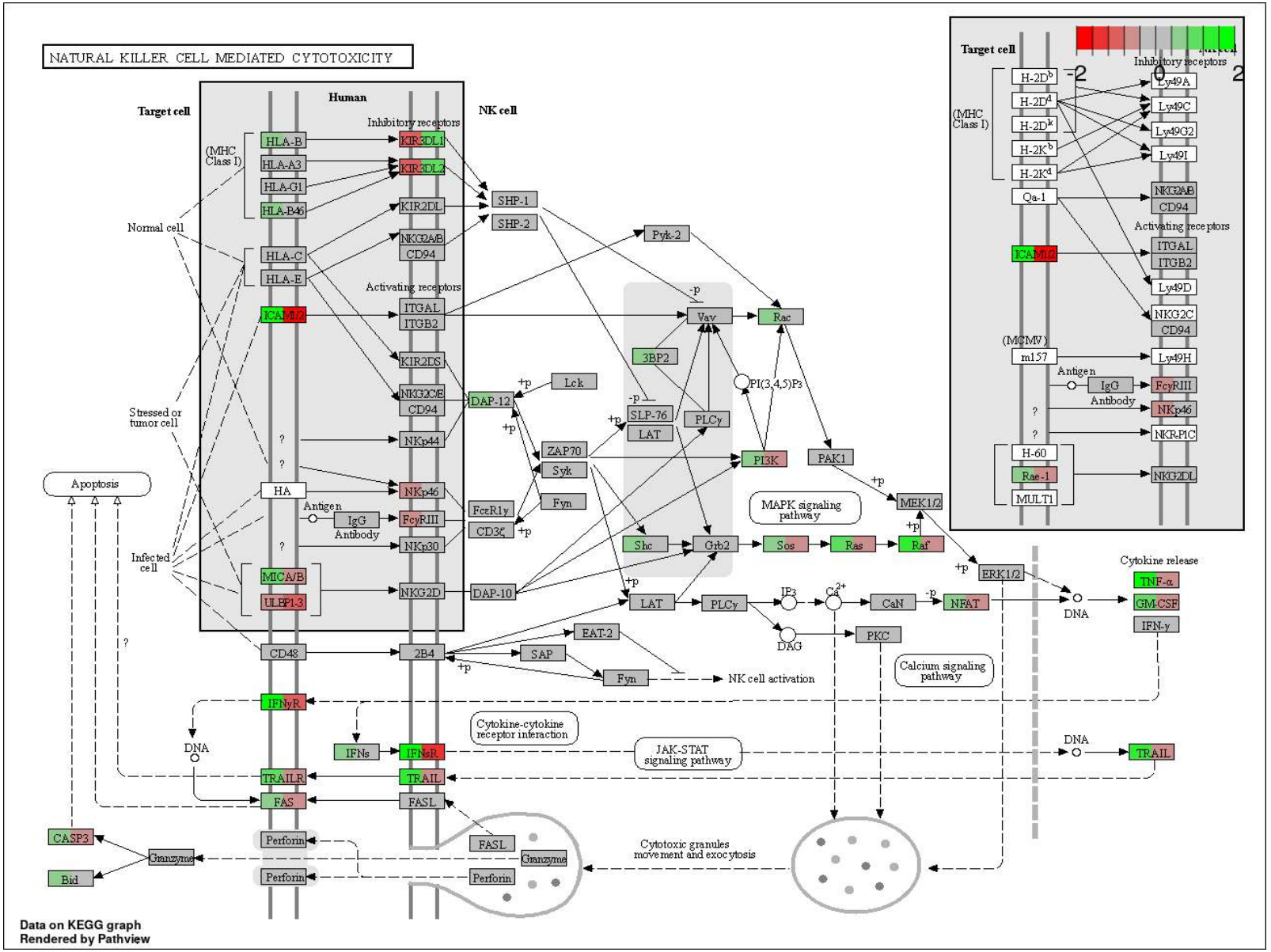

Figure 10.20: KEGG pathview of the natural killer cell mediated cytotoxicity pathway, which was significantly (0.0028) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition toned down the SHIME effect (grey and red coloring), though not significantly.

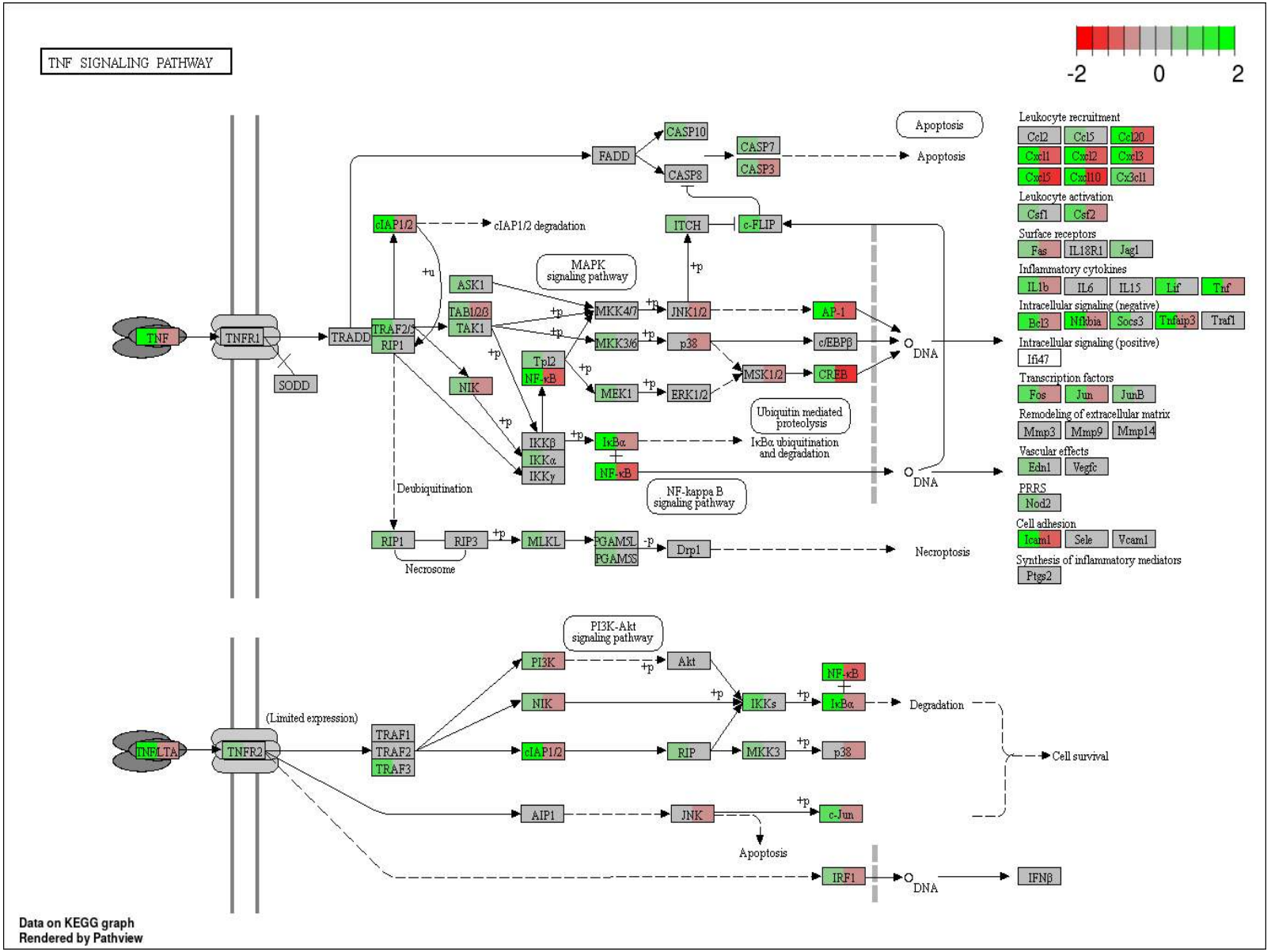

Figure 10.21: KEGG pathview of the TNF signaling pathway, which was significantly (3.3e-14) upregulated after exposure of the triple coculture cell model to the distal colon SHIME sample. The difference in gene expression between a SHIME and Blank sample and the difference in gene expression between a SHIME sample amended with LGG and the SHIME sample without LGG were mapped onto the pathway with a color coding in boxes which were divided in two states. Green coloring revealed up-regulation of several genes by SHIME stimulation. LGG addition significantly (7.6e-05) toned down the SHIME effect (grey and red coloring).

10.2 Gene ontology based gene-set analysis

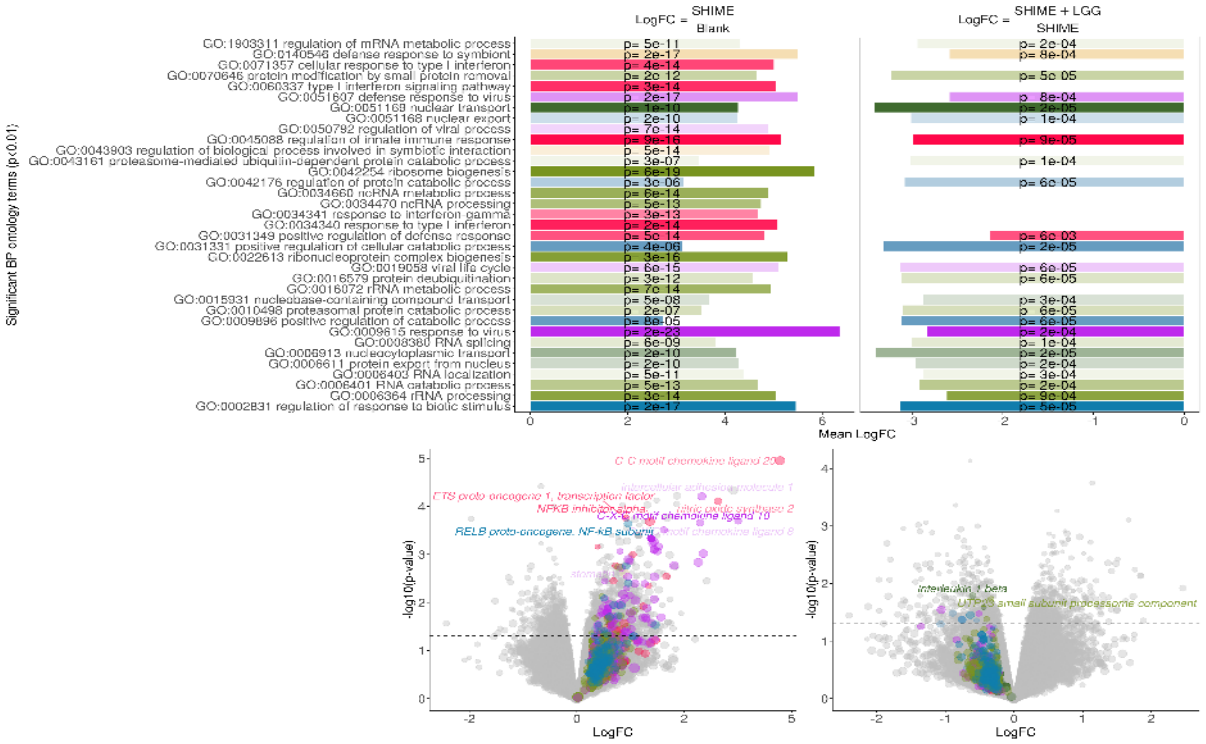

Figure 10.22: Significantly (p<0.01) up- (positive LogFC) and downregulated (negative LogFC) GO BP terms as assessed by gene-set analysis (GSA) in GAGE. P-values are adjusted for multiple testing with Benjamini & Hochberg FDR correction. Bars are colored by the GO term category: pro-inflammatory response (pink), viral interaction (purple), RNA (green) and protein (yellow) processing, signaling patways (teal), cellular transport (brown) and cellular differentiation (blue).

Chapter 11

Integrative analysis

Results from the transcriptomics differential expression analysis were combined with functional data from mucus layer thickness, cytotoxic stress and cytokine production, flow cytometry cell counts and SCFA concentration measurements through DIABLO (Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies) from the mixOmics framework 6.16.0 [9]. DIABLO is a latent variable multivariate dimension reduction method that aims to identify the correlated variables that best explain the categorical outcome variable of interest (triple coculture cell model treatment: Blank, SHIME or SHIME + LGG). DIABLO thereto extends sparse generalized canonical correlation analysis, which uses singular value decomposition to select co-expressed variables from several omics datasets, to a classification framework.

The DIABLO model consisted of a design matrix containing the transcriptomics expression data and functional host-microbe interaction data, constructed with a link of 0.1. Global model performance was assessed by 5-fold cross validation, repeated 10 times, and 2 components were selected to fit the final model based on the overall and balanced error rates for the centroids distance criterion (Figure 11.1).

Next, smart feature selection was applied and the optimal number of variables was determined based on a 10x3-fold cross validation. The performance of the final model with 2 components, each with 6 functional respectively 10 transcriptomics variables was verified using the balanced classification error rate with majority vote and centroids distance.

The correlation between variables from the transcriptomics expression data and metadata was successfully maximized (0.95 and 0.55 for the first and second component; Figure 11.2-11.3). While a two-component model was sufficient to discriminate the blank, SHIME and SHIME+LGG cell models in both datasets, the discriminatory power was higher for the transcriptomics than the microbe-host functional data (Figure 11.4-11.5).

The variables contributing most to the two components in both data types were displayed in Figure 11.6.

Finally, the correlations between and within variables from each dataset were computed using a similarity score that is analogous to a Pearson correlation coefficient and represented in a Circos plot (Figure 11.7) [2].

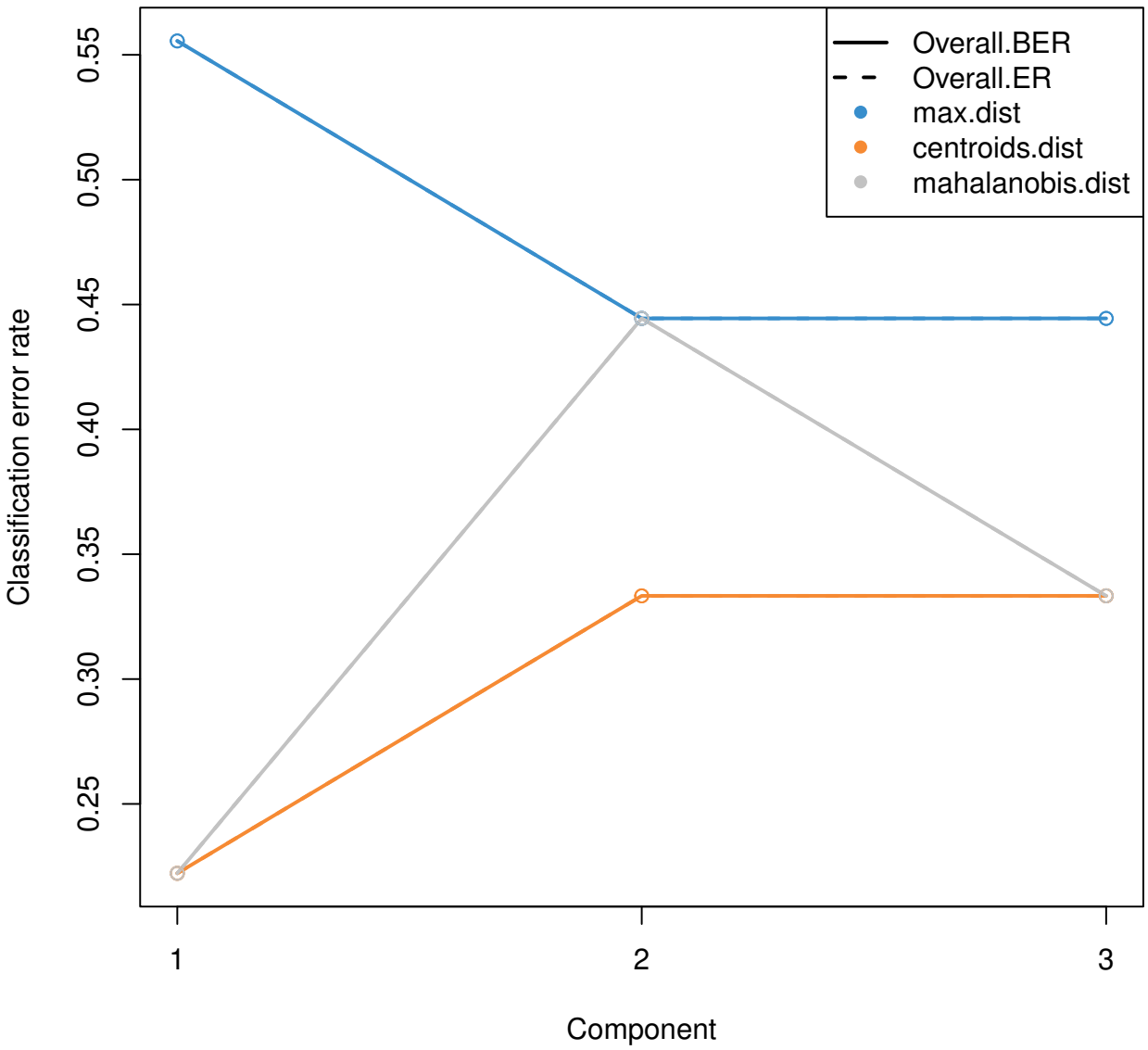

Figure 11.1: Overall and balanced classification error rate (BER) using Maximum, Centroids or Mahalanobis distances

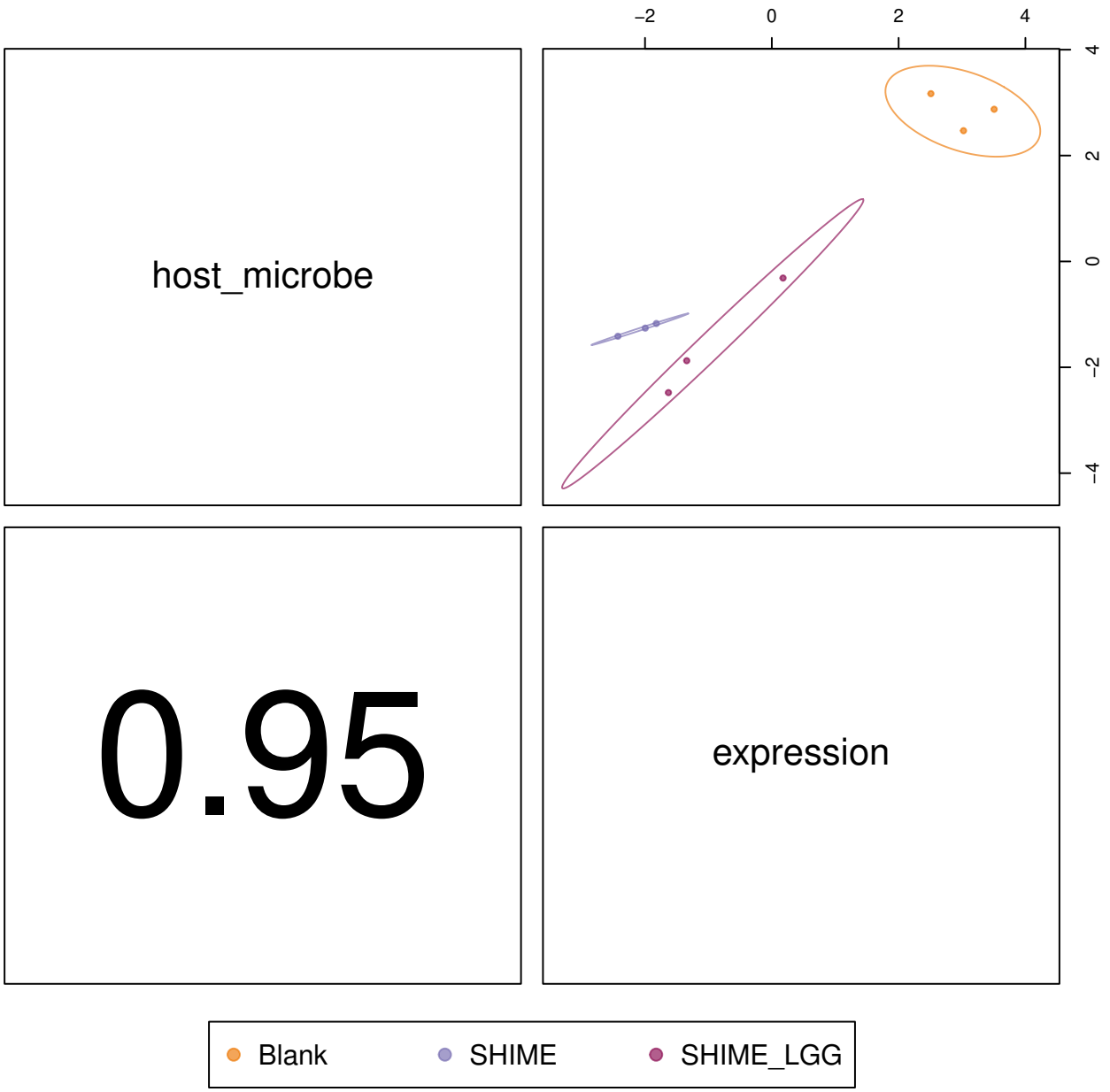

Figure 11.2: Correlation between features in the expression and metadata along the first component.

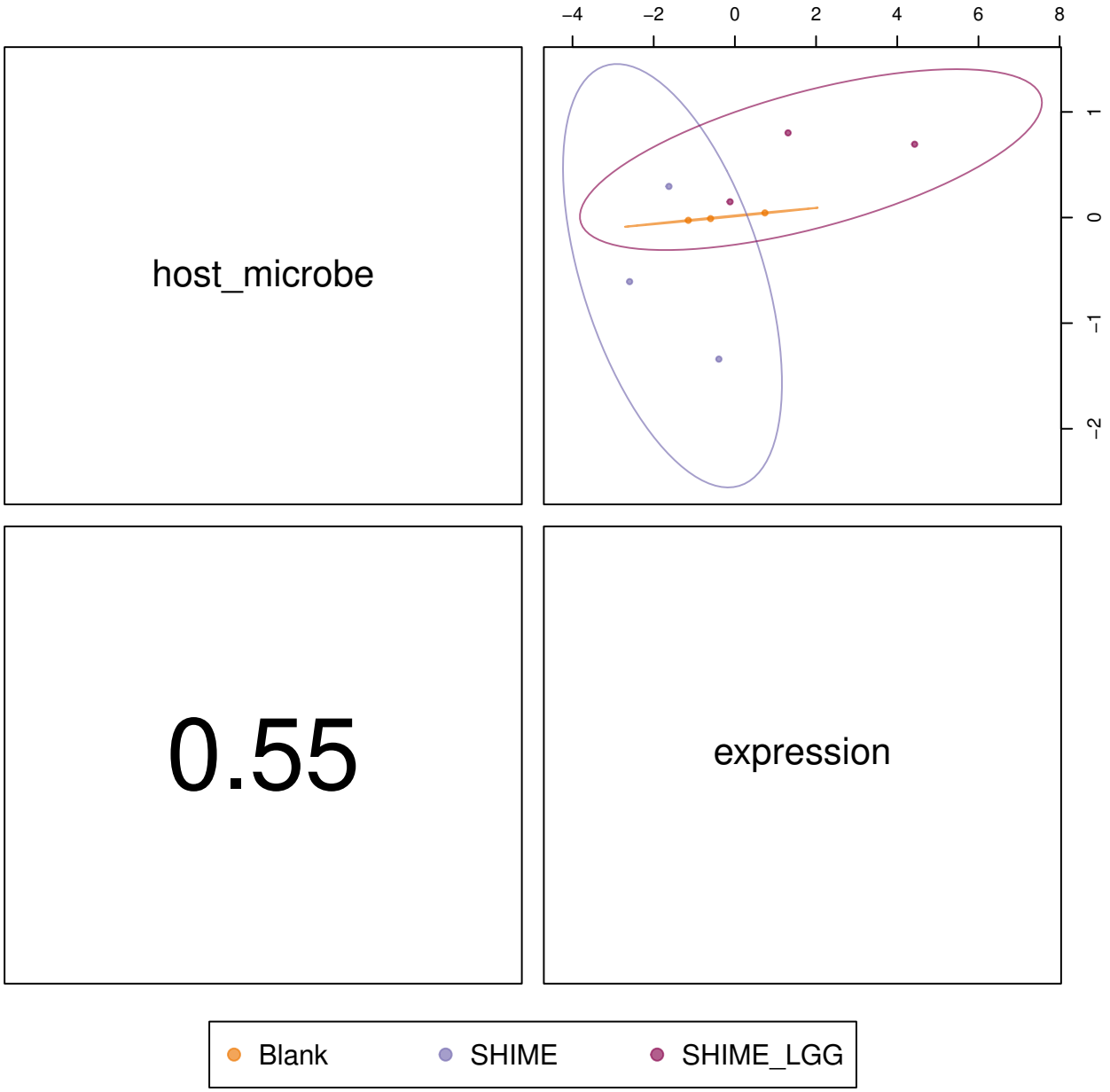

Figure 11.3: Correlation between features in the expression and metadata along the second component.

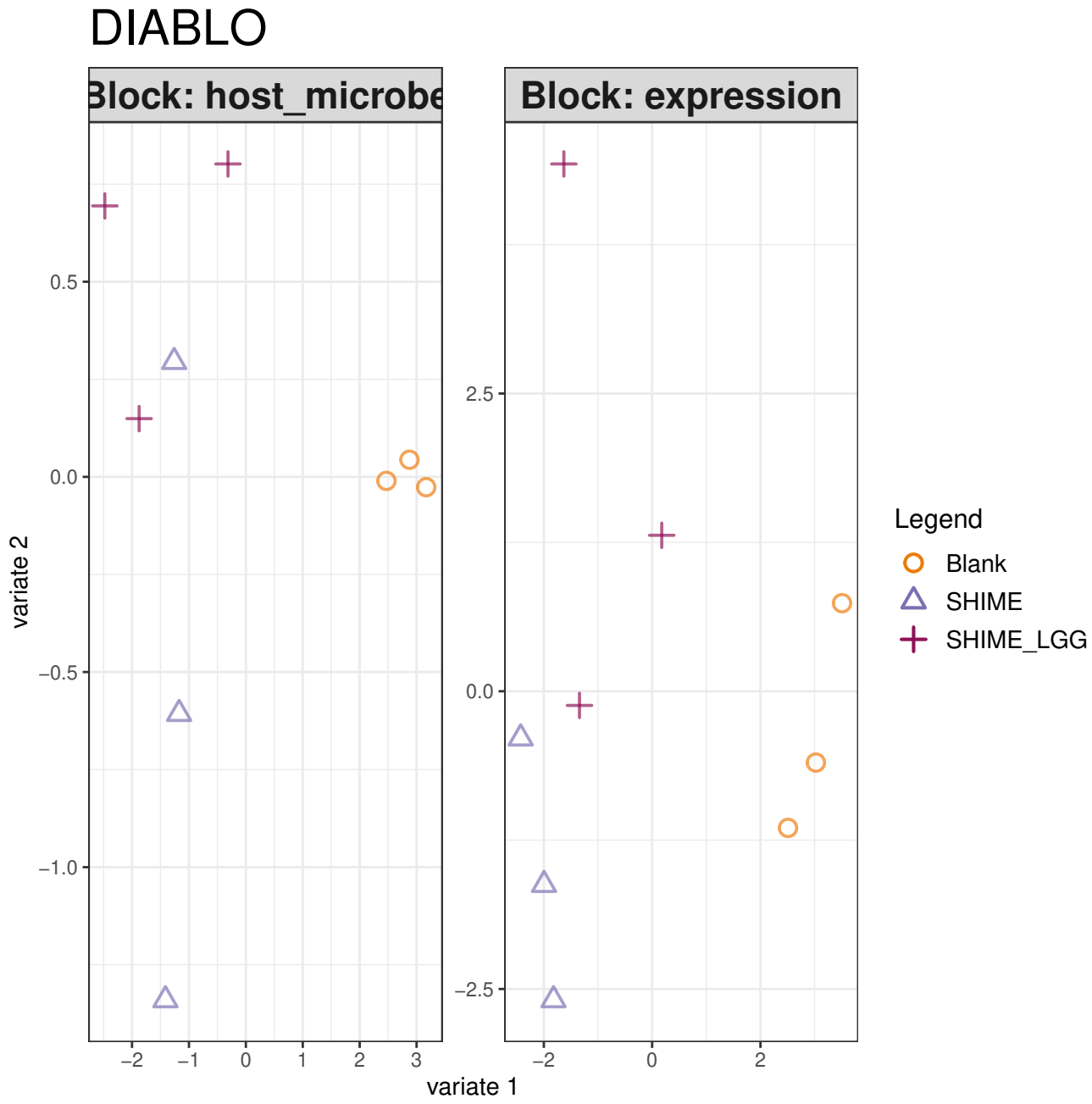

Figure 11.4: Sample projection for the separate datasets

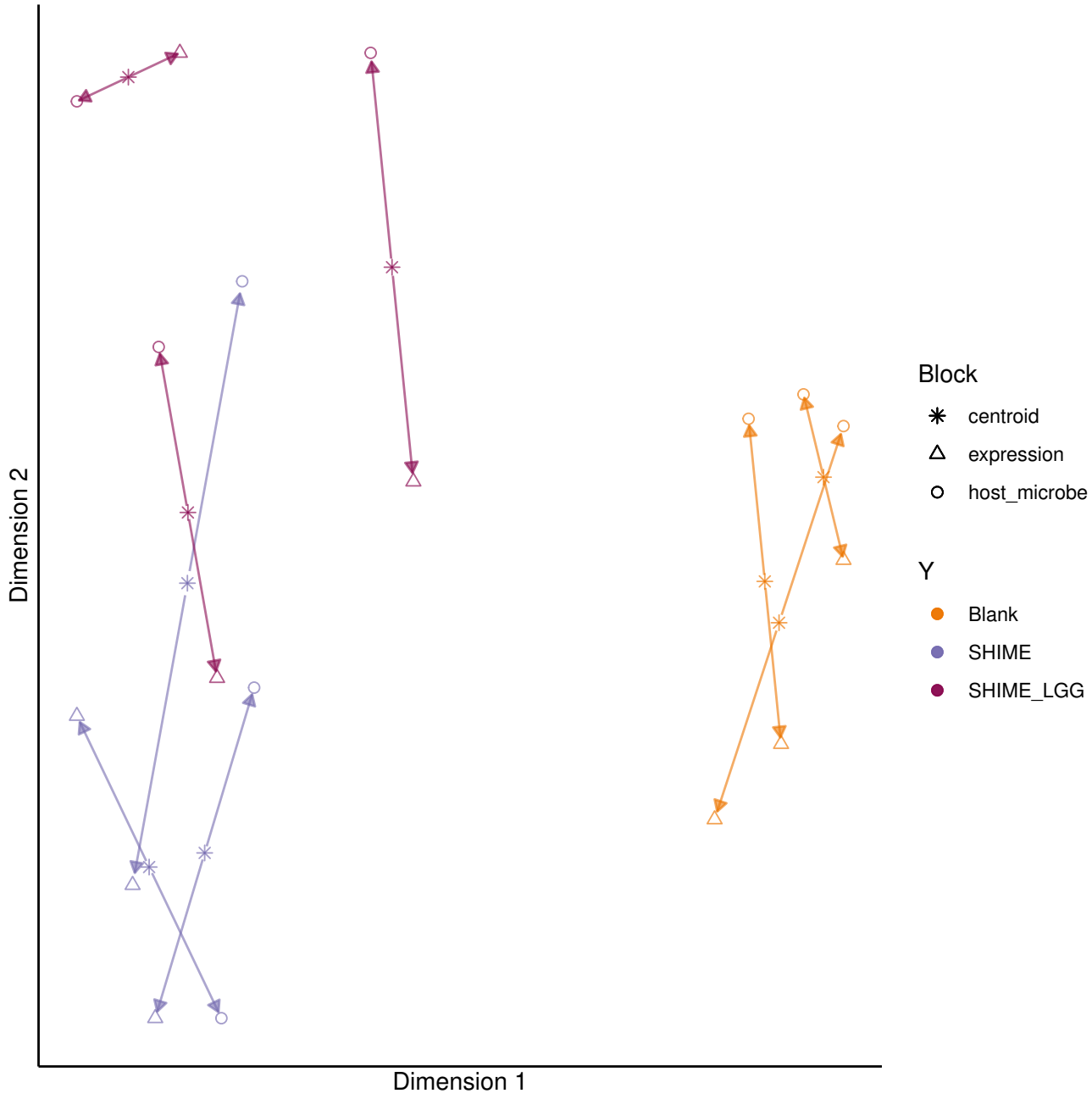

Figure 11.5: Sample projection with arrows indicating the location of each sample in the expression and metadata confirms the agreement between datasets at sample level

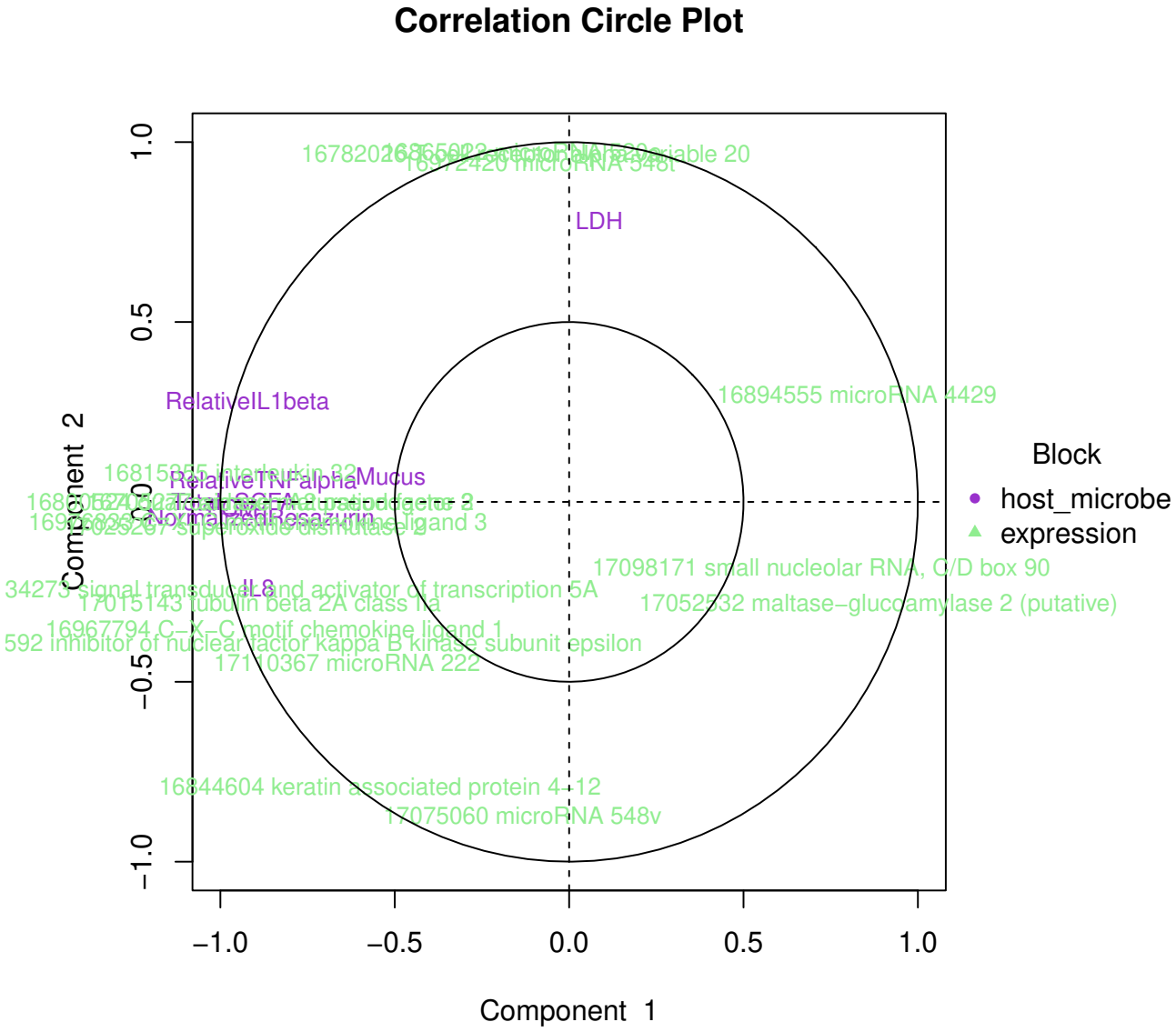

Figure 11.6: Correlation circle plot highlighting the contribution of each selected variable to each component.

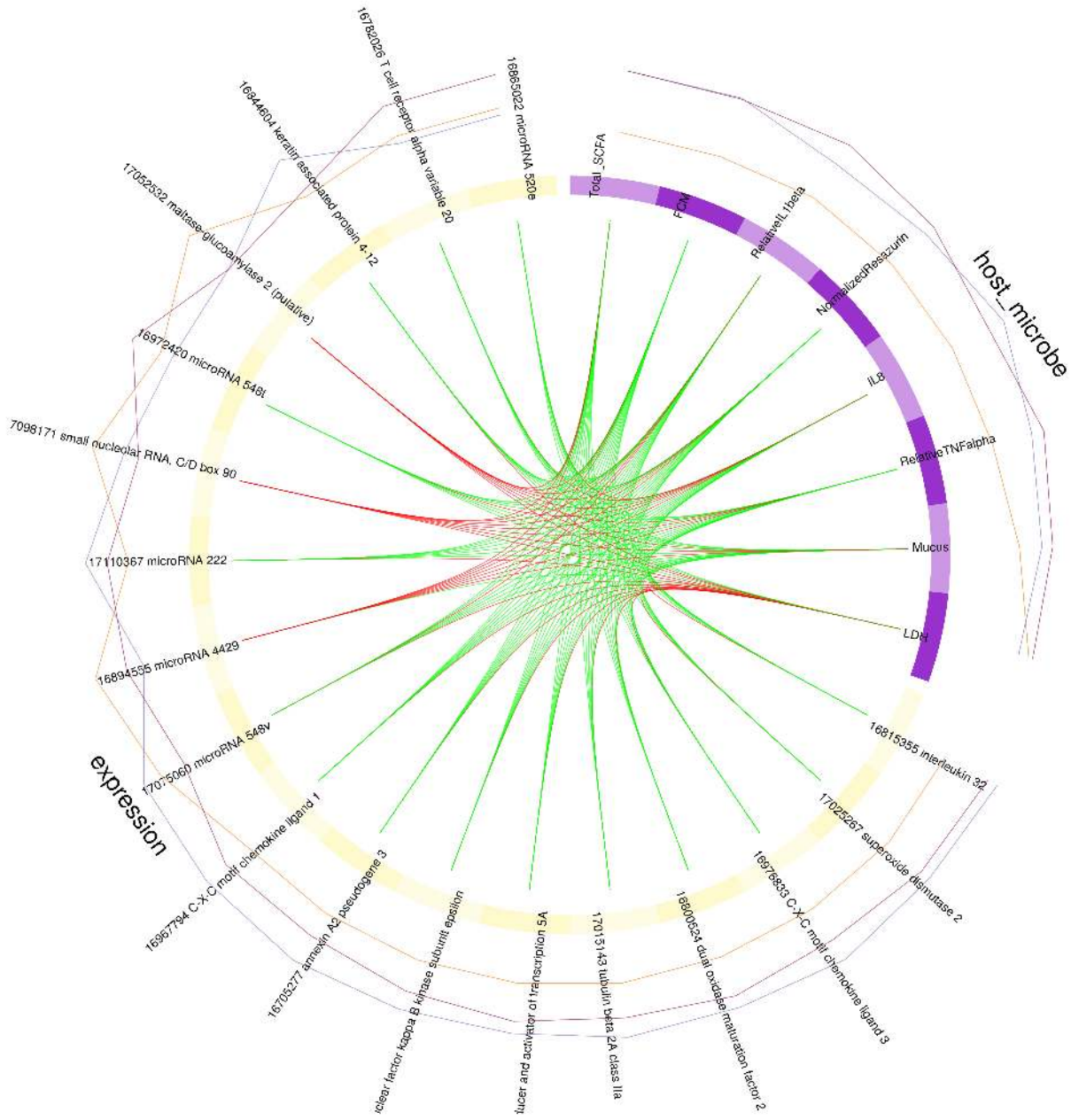

Figure 11.7: Circos plot displaying correlations (inner circle: red-negative and green-positive) among the selected functional and transcriptomics variables most predictive for triple coculture cell model treatment (outer lines:Blank-orange, SHIME-purple, SHIME+LGG - violetred).

Chapter 12

Exported results

All differential expression results obtained with Limma, topGO and GAGE were exported to a supplementary excel file.

Chapter 13

Online resources

- https://wiki.bits.vib.be/index.php/Analyze_your_own_microarray_data_in_R/Bioconductor
- http://tools.thermofisher.com/content/sfs/brochures/exon_gene_arrays_qa_whitepaper.pdf
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6347406/
- https://www.bioconductor.org/packages/release/bioc/vignettes/yaqcaffy/inst/doc/yaqcaffy.pdf
- https://storage.googleapis.com/plos-corpus-prod/10.1371/journal.pone.0029059/1/pone.0029059.s008.pdf?X-Goog-Algorithm=GOOG4-RSA-SHA256&X-Goog-Credential=wombat-sa%40plos-prod.iam.gserviceaccount.com%2F20210401%2Fauto%2Fstorage%2Fgoog4_request&X-Goog-Date=20210401T153107Z&X-Goog-Expires=3600&X-Goog-SignedHeaders=host&X-Goog-Signature=3dfed70bed1ee74dac38de0e018d71ae71720508f6e5319ae73858e89af94804a91e89e3a85d8258fbff
- https://www.fda.gov/science-research/bioinformatics-tools/microarraysequencing-quality-control-maqcseqc
- https://mgimond.github.io/ES218/Week11a.html
- https://www.r-bloggers.com/2015/12/tutorial-rna-seq-differential-expression-pathway-analysis-with-sailfish-deseq2-gage-and-pathview/
- https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-07337-9
- https://varemo.github.io/GSA_tutorial/functional_annotation.html

Bibliography

[1] Bolstad., B. M. *Low Level Analysis of High-Density Oligonucleotide Data: Background, Normalization and Summarization.* Phd thesis, University of California, Berkeley, 2004.

[2] Gonzalez, I., Le Cao, K. A., Davis, M. J., and Dejean, S. Visualising associations between paired 'omics' data sets. *Biodata Mining 5* (2012).

[3] Klaus, B., and Reisenauer, S. An end to end workflow for differential gene expression using affymetrix microarrays. *F1000Research 5* (2016), 1384.

[4] Lippa, K. A., Duewer, D. L., Salit, M. L., Game, L., and Causton, H. C. Exploring the use of internal and externalcontrols for assessing microarray technical performance. *BMC research notes 3* (2010), 349.

[5] Luo, W. J., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. Gage: generally applicable gene set enrichment for pathway analysis. *Bmc Bioinformatics 10* (2009).

[6] Moffitt, R. A., Qiqin, Y. G., Stokes, T. H., Parry, R. M., Torrance, J. H., Phan, J. H., Young, A. N., and Wang, M. D. cacorrect2: Improving the accuracy and reliability of microarray data in the presence of artifacts. *Bmc Bioinformatics 12* (2011).

[7] Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., and Smyth, G. K. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Annals of Applied Statistics 10*, 2 (2016), 946–963.

[8] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y. F., Law, C. W., Shi, W., and Smyth, G. K. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research 43*, 7 (2015).

[9] Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., and Le Cao, K. A. Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics 35*, 17 (2019), 3055–3062.

[10] Song, J. S., Maghsoudi, K., Li, W., Fox, E., Quackenbush, J., and Shirley Liu, X. Microarray blob-defect removal improves array analysis. *Bioinformatics (Oxford, England) 23*, 8 (2007), 966–71.

[11] Upton, G. J., Sanchez-Graillet, O., Rowsell, J., Arteaga-Salas, J. M., Graham, N. S., Stalteri, M. A., Memon, F. N., May, S. T., and Harrison, A. P. On the causes of outliers in affymetrix genechip data. *Briefings in functional genomics & proteomics 8*, 3 (2009), 199–212.