

Supplementary Dataset S2: Examining the role of transcription and post-transcriptional modifications on mediating intragenomic conflict in honey bees

Lindsay Clark, HPCBio, Roy J. Carver Biotechnology Center, University of Illinois, Urbana-Champaign
Sean Bresnahan, Huck Institutes of the Life Sciences, Pennsylvania State University

2022-12-23

Contents

Study design	1
Sample processing	2
Identification and differential expression of lincRNAs	2
QC on read counts	3
Normalization and filtering	5
Sample clustering	9
Differential gene expression	10
Heatmaps for differential lincRNA expression	11
Analysis of isoform switching	13
QC and differential gene expression	13
Quality control	13
Mitochondrial reads	16
Normalization and filtering	18
Sample clustering	23
Differential gene expression	24
Heatmaps for differential gene expression	25

Study design

This study uses whole genome resequencing data produced by Illumina DNA sequencing and transcriptome data produced by Oxford Nanopore direct RNA sequencing to understand behavioral differences between non-aggressive and aggressive honey bees. Two crosses in a reciprocal cross design between Africanized honey bees (AHB) and European honey bees (EHB) were made. Cross A had AHB as the female parent and EHB as the male parent. Cross B had EHB as the female parent and AHB as the male parent. Three or more aggressive bees and three or more non-aggressive bees (labeled T and C, respectively) were collected from each cross. Twelve head and abdomen samples, spanning multiple batches over the course of a year, were ultimately analyzed, with three samples per experimental group. RNA samples were sequenced with Oxford Nanopore Technology direct RNA sequencing on a FLO-MIN106 flowcell.

We test two main hypotheses:

1. Aggression is associated with an enrichment for paternal allele-biased transcription,
2. Allele-biased transcription is associated with allele-biased RNA m6A

Additionally, we assess whether either allele-biased transcription or allele-biased RNA m6A are associated with differential gene expression and/or differential isoform usage.

All scripts and code for the full analyses are available at
<https://github.com/sbresnahan/allele-specific-transcription-and-m6A>.

Sample processing

FAST5 files from all samples were processed with Guppy 5.0.16 using the `rna_r9.4.1_70bps_hac.cfg` configuration (Oxford Nanopore Technologies) for base calling on a GPU. FASTQ files labeled “pass” were concatenated together to generate one FASTQ per sample. Minimap 2.21¹ was used for alignment to the *Apis mellifera* HAv3.1 reference genome (plus common viral genomes) in splice-aware mode with a kmer length of 14, using the published gene annotation to guide alignment. SAMtools 1.12² was used to sort and index the resulting BAM files. Flair 1.5³ was then used to convert BAM files to BED format and correct intron-exon junctions based on the published annotation, then call isoforms across all samples. These isoforms included published transcripts and genes as well as novel transcripts and genes.

Finally, EpiNano⁴ was used to quantify read coverage and RNA m6A probability at parent SNPs within transcripts. combined with a custom R script (`src/combine_results_2022-08-03.R`) to subset the EpiNano output to the A sites at the center of RRACH motifs. Read coverage and m6A probability at each position within published and novel transcripts, separated by allele, were then combined to generate read count and m6A probability matrices.

Identification and differential expression of lincRNAs

Of the 31,475 transcripts output by Flair⁵, 451 were already annotated as long non-coding RNAs (lncRNAs) in the NCBI GCF_003254395.2_Amel_HAv3.1 annotation, while 12 were for genes annotated as ribosomal RNA (rRNA) or small nuclear RNA (snRNA), and 17,880 were for protein-coding genes. Although snoRNAs, miRNAs, tRNAs, and guide RNAs were annotated in the reference genome, Flair did not output transcripts for any of these.

Transcripts were defined as “intergenic” if they had no overlap with any annotated protein-coding, rRNA, miRNA, snRNA, snoRNA, tRNA, or guide RNA gene; 3,908 transcripts were considered intergenic by this measure. Additionally, 77 transcripts were on the mitochondrial genome (including transcripts matching gene NC_037641.1:1186000). Transcripts were retained for further analysis if they were non-mitochondrial, intergenic, and were at least 200 nucleotides in length, leaving 3806 transcripts for further analysis, including 6 transcripts that were later removed for being viral in origin.

The `findORFs` function in the ORFhunterR⁶ R package was used to identify putative open reading frames (ORFs) and their lengths. However, 142 of the 165 annotated lncRNAs that were analyzed had putative ORFs using this method, so this information was ignored in favor of other approaches for determining if transcripts

¹Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34:3094-3100. doi:10.1093/bioinformatics/bty191

²Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, Volume 10, Issue 2, giab008, <https://doi.org/10.1093/gigascience/giab008>

³Tang, A.D., Soulette, C.M., van Baren, M.J. et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* 11, 1438 (2020). <https://doi.org/10.1038/s41467-020-15171-6>

⁴Liu, H., Begik, O., Lucas, M.C. et al. Accurate detection of m6A RNA modifications in native RNA sequences. *Nat Commun* 10, 4079 (2019). <https://doi.org/10.1038/s41467-019-11713-9>

⁵Tang, A.D., Soulette, C.M., van Baren, M.J. et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* 11, 1438 (2020). <https://doi.org/10.1038/s41467-020-15171-6>

⁶Grinev V, Yatskou M, Skakun V, Chepeleva M, Nazarov P (2021). ORFhunteR: Predict open reading frames in nucleotide sequences. R package version 1.2.0.

were protein-coding. Using CPC2⁷, 61 out of 3806 transcripts were determined to have coding potential. Additionally, BLASTX (from BLAST+ v2.10.1) was used to align transcripts to the Uniprot and Swiss-Prot combined database (from a June 2020 mirror) under default parameters. Using a cutoff E value of 0.001, 120 transcripts aligned to a known protein sequence. Lastly, gene sequences for all annotated snRNA, snoRNA, miRNA, tRNA, rRNA, and guide RNAs were extracted from the GCF_003254395.2_Amel_HAV3.1 reference sequence. Since there were no annotated 7SL RNA genes in the reference, I used sequences NR_001992.2 and NR_037753.2 from *Drosophila melanogaster* to represent those. I then aligned the putative lincRNA transcripts to all of these other non-coding RNAs using discontinuous megablast (from BLAST+ v2.10.1) under default parameters, finding six transcripts aligning very strongly to rRNA genes.

Altogether, 3642 transcripts were retained as lincRNAs for being non-mitochondrial, non-viral, intergenic, at least 200 nt long, not predicted to have coding potential by CPC2, not aligning to known proteins, and not aligning to other non-coding RNAs. This is an increase over the 3048 lincRNA transcripts identified prior to fixing the Flair bug. Tallies of annotated and novel lincRNAs and their corresponding genes are listed in Table 1.

Table 1: Tallies of annotated and unannotated lincRNAs after filtering.

Gene_type	Genes	Annotated_lincRNAs	Annotated_pseudogene_RNAs	Novel_RNAs
Annotated	219	158	1	116
Novel	2970	0	0	3367

QC on read counts

Out of the 3642 lincRNA transcripts identified, Flair output read counts for 3640 of them. The proportion of reads mapping to lincRNAs varied among samples, with T samples having slightly lower proportions than C samples ($P = 0.055$). At a gross scale the total number of lincRNA reads followed the total library size (Fig. 2; see library sizes in previous reports).

⁷Kang Y. J., Yang D. C., Kong L., Hou M., Meng Y. Q., Wei L., Gao G. 2017. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Research 45(Web Server issue): W12–W16.

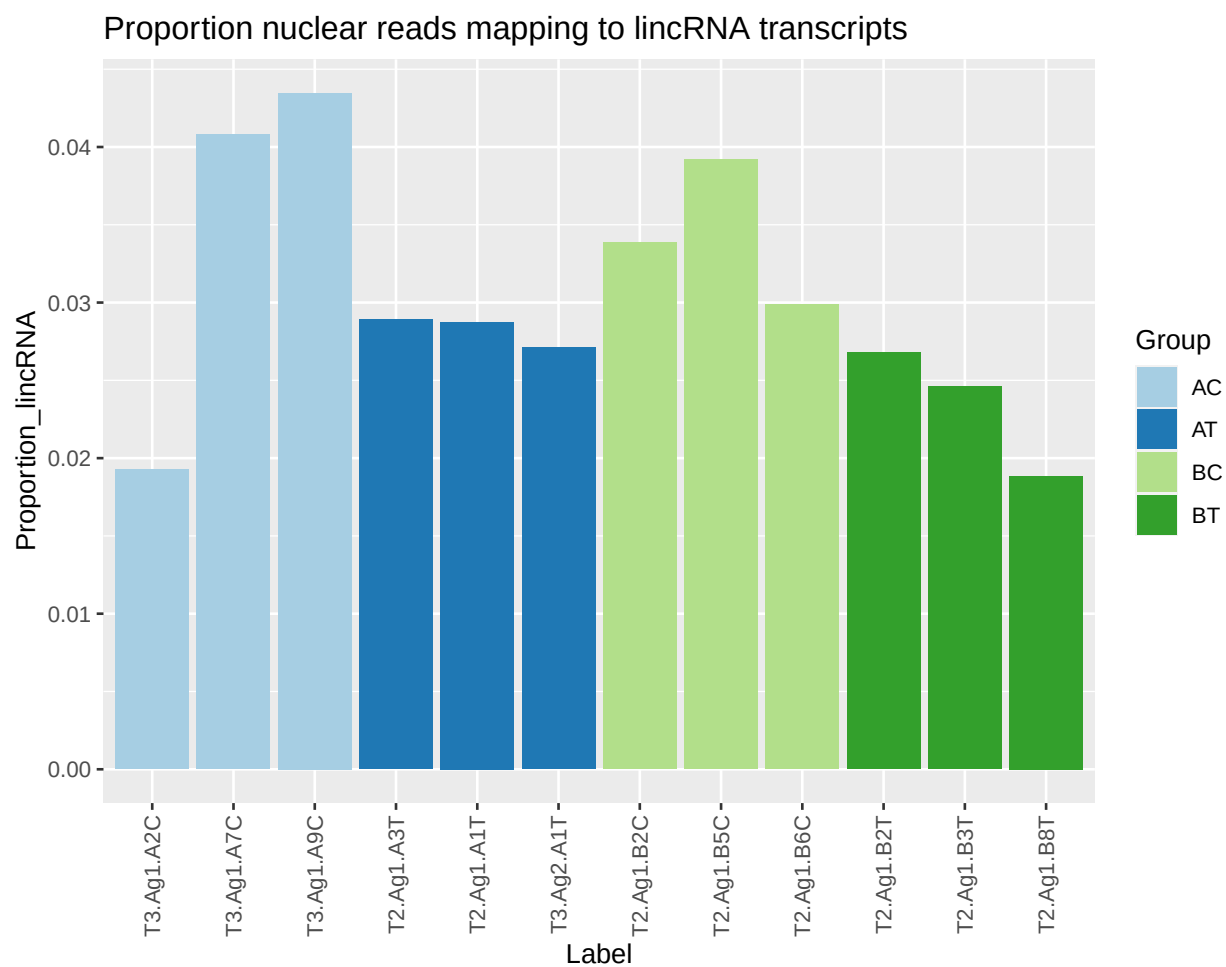


Figure 1: Proportion of reads mapping to lincRNAs, out of total reads mapping to the nuclear genome. Reads mapping to mitochondrial or viral genomes were excluded.

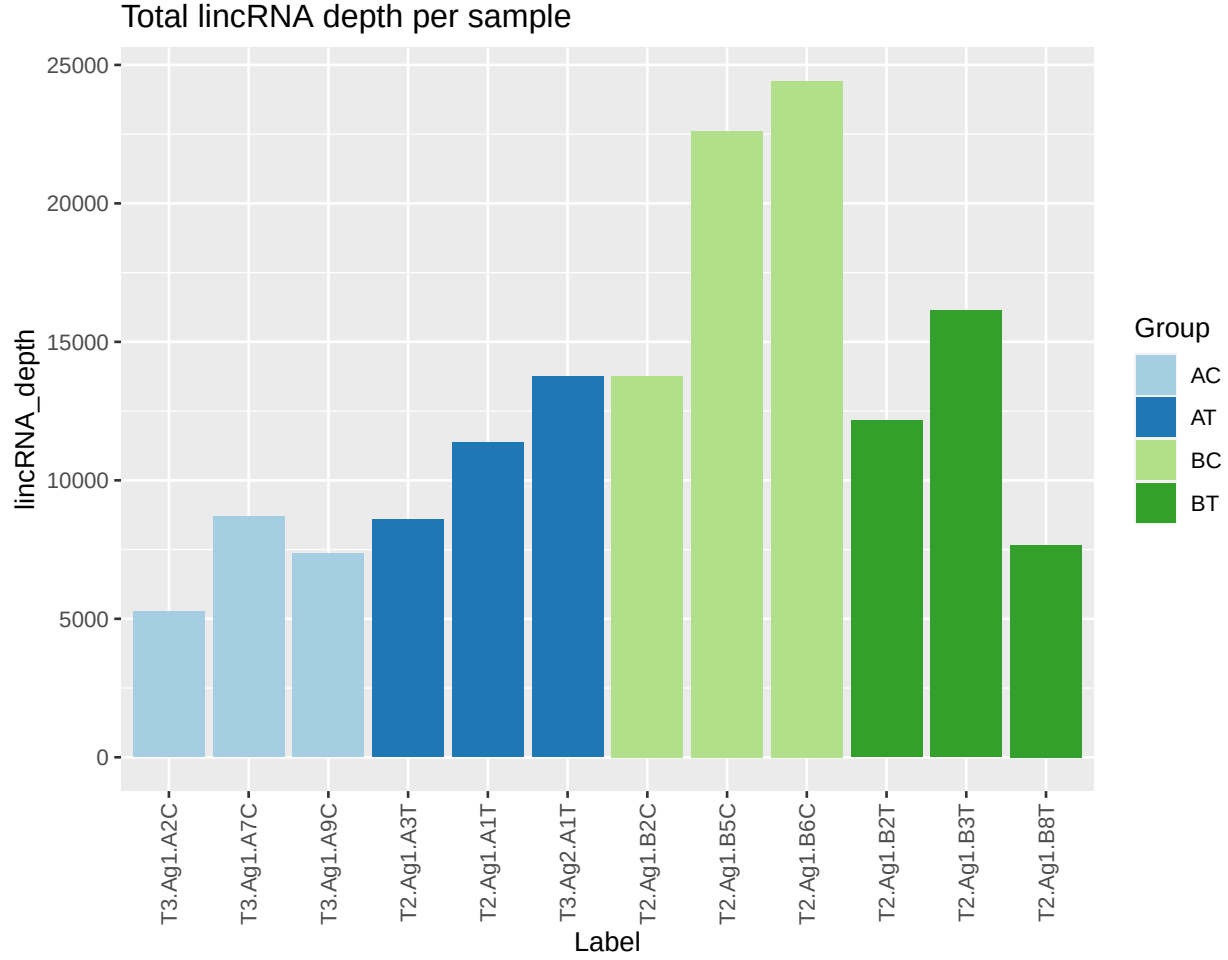


Figure 2: Total number of reads mapping to lincRNAs in each sample.

Normalization and filtering

When comparing expression levels, the numbers of reads per transcript need to be normalized not only because of the differences in total number of reads, but because there could be differences in RNA composition such that the total number of reads would not be expected to be the same. The TMM (trimmed mean of M values) normalization⁸ in the edgeR package⁹ uses the assumption of *most genes do not change* to calculate a normalization factor for each sample to adjust for such biases in RNA composition. In this dataset, TMM normalization factors fluctuated between 0.84 and 1.3. TMM normalization factors are multiplied by the library size in order to generate the effective library size, where smaller TMM normalization factors indicate more overrepresented genes in the sample. Normalization factors varied among samples but were not particularly associated with experimental group (Fig. 3).

⁸Robinson MD, Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11, R25.

⁹Robinson MD, McCarthy DJ, Smyth GK (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” *Bioinformatics*, 26(1), 139-140.

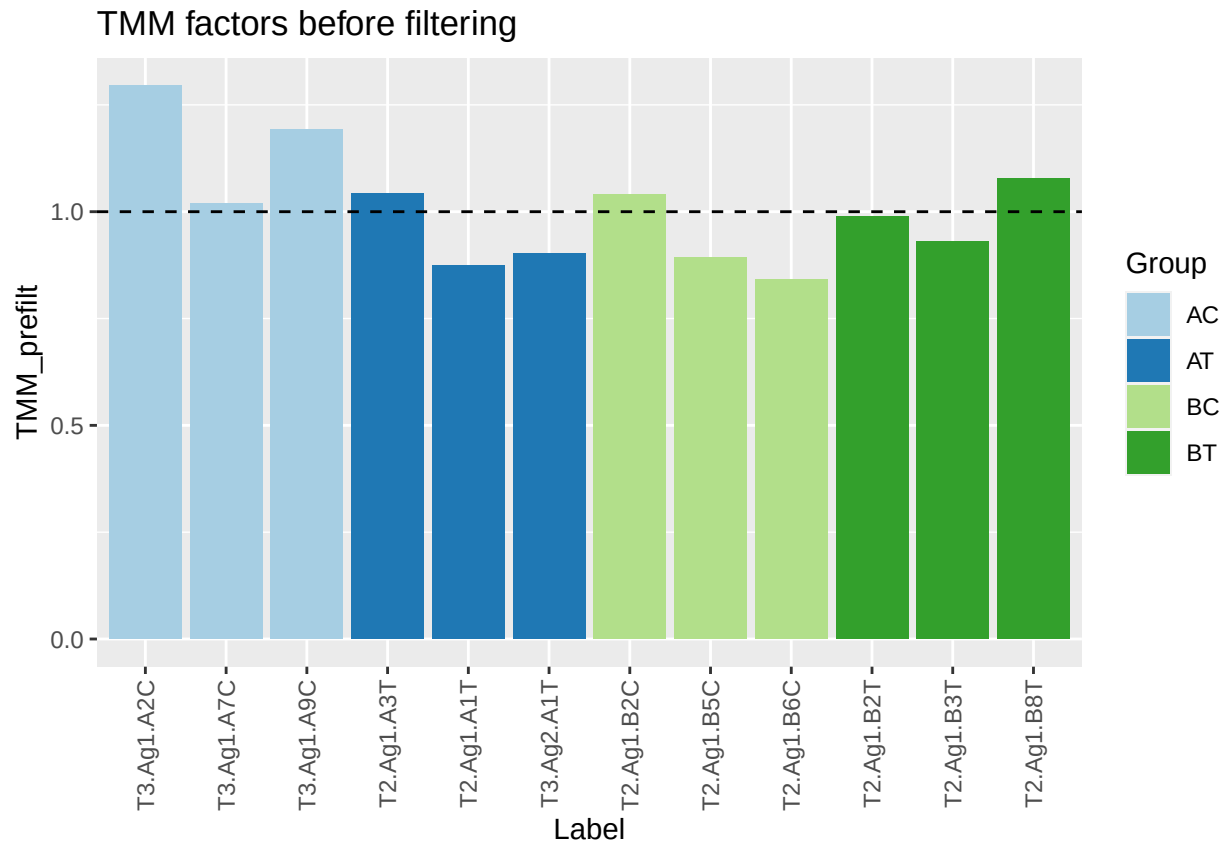


Figure 3: Normalization factors before filtering

While the sequence analysis identified a total of 3,640 lincRNA transcripts, many of these might not have detectable expression in enough samples for differential expression to be detectable. Therefore, we set the detection threshold at 250 cpm (counts per million) in at least 3 samples, which resulted in 1,225 lincRNAs being filtered out, leaving 2,415 lincRNAs to be analyzed for differential expression, containing 96.54% of the reads. This threshold resulted in a fairly even number of genes being detected across samples, considering the variability in library size (Fig. 4). In a typical Illumina RNA-seq experiment, we might use 1 cpm as the filtering threshold. Given the lower read depth of this experiment due to the use of Oxford Nanopore sequencing and filtering down to lincRNAs, a higher threshold was needed to differentiate between lincRNAs that could be detected in few samples vs. lincRNAs that could be detected in most samples. I arrived at a threshold of 250 by examining the detection rates at a variety of thresholds ranging from 200 to 500, where 164 CPM corresponded to zero reads in the raw data (results not shown).

After filtering, TMM normalization was performed again and normalized log2-based count per million values (logCPM) were calculated using edgeR's¹⁰ `cpm()` function with `prior.count = 2` to help stabilize fold-changes of extremely low expression genes. TMM normalization factors after filtering are shown in Fig. 5, and compared to pre-filtering values in Fig. 6.

¹⁰McCarthy DJ, Chen Y, Smyth GK (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." *Nucleic Acids Research*, 40(10), 4288-4297.

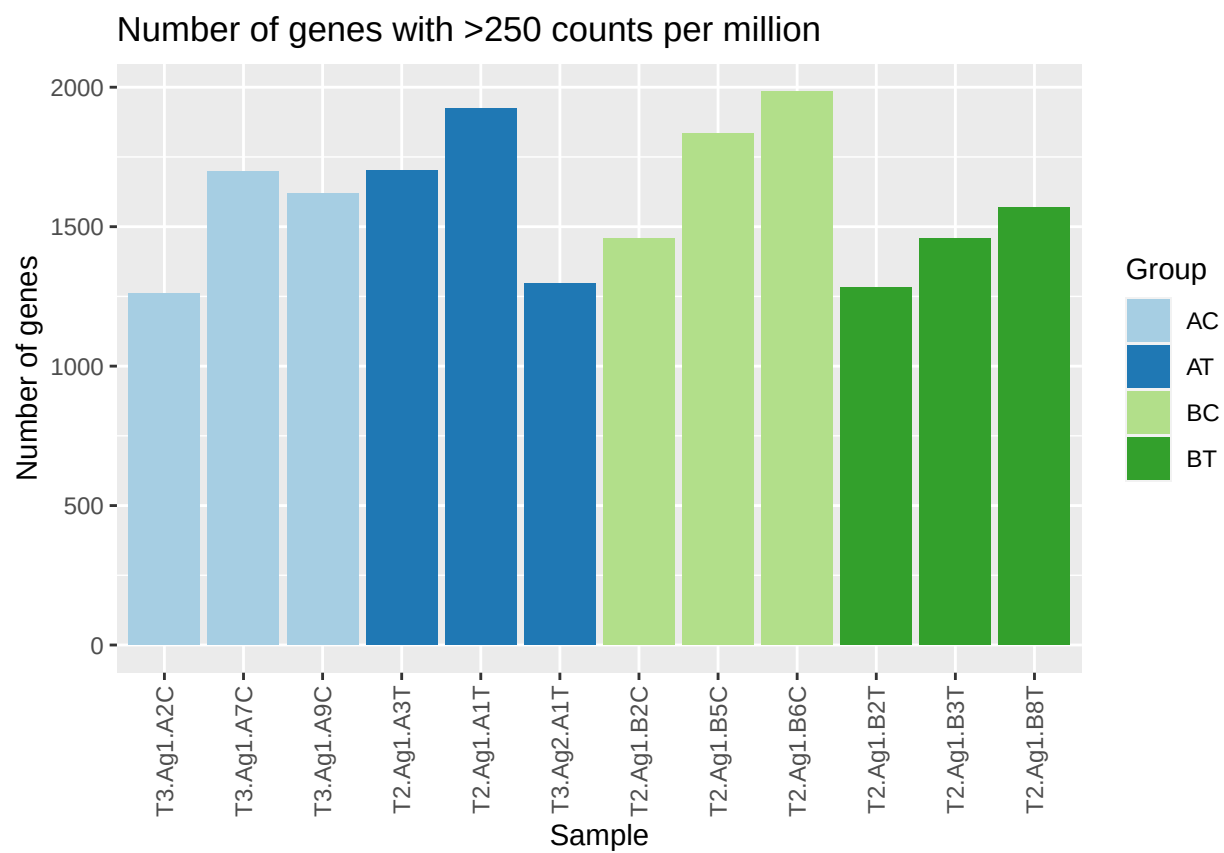


Figure 4: Number of detected genes passing the filtering threshold in each sample.

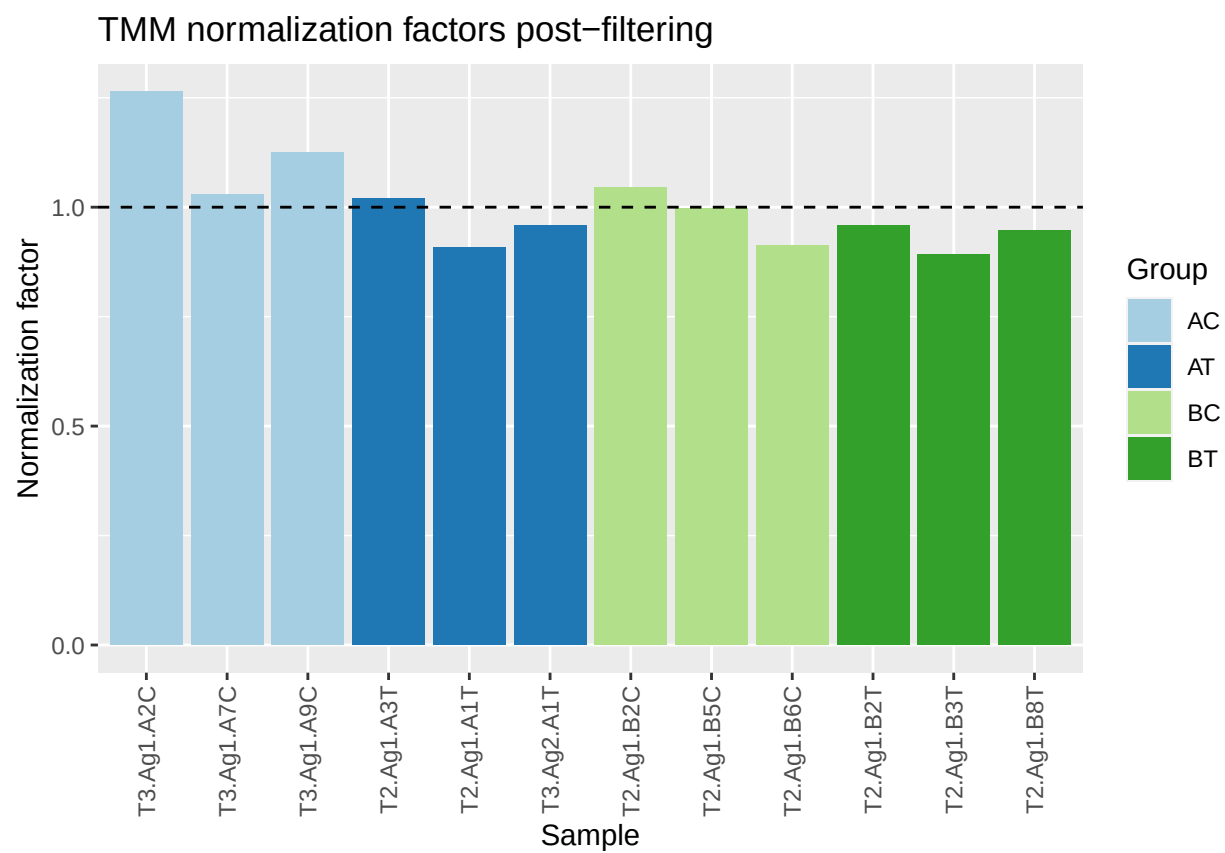


Figure 5: TMM normalization factors post-filtering.

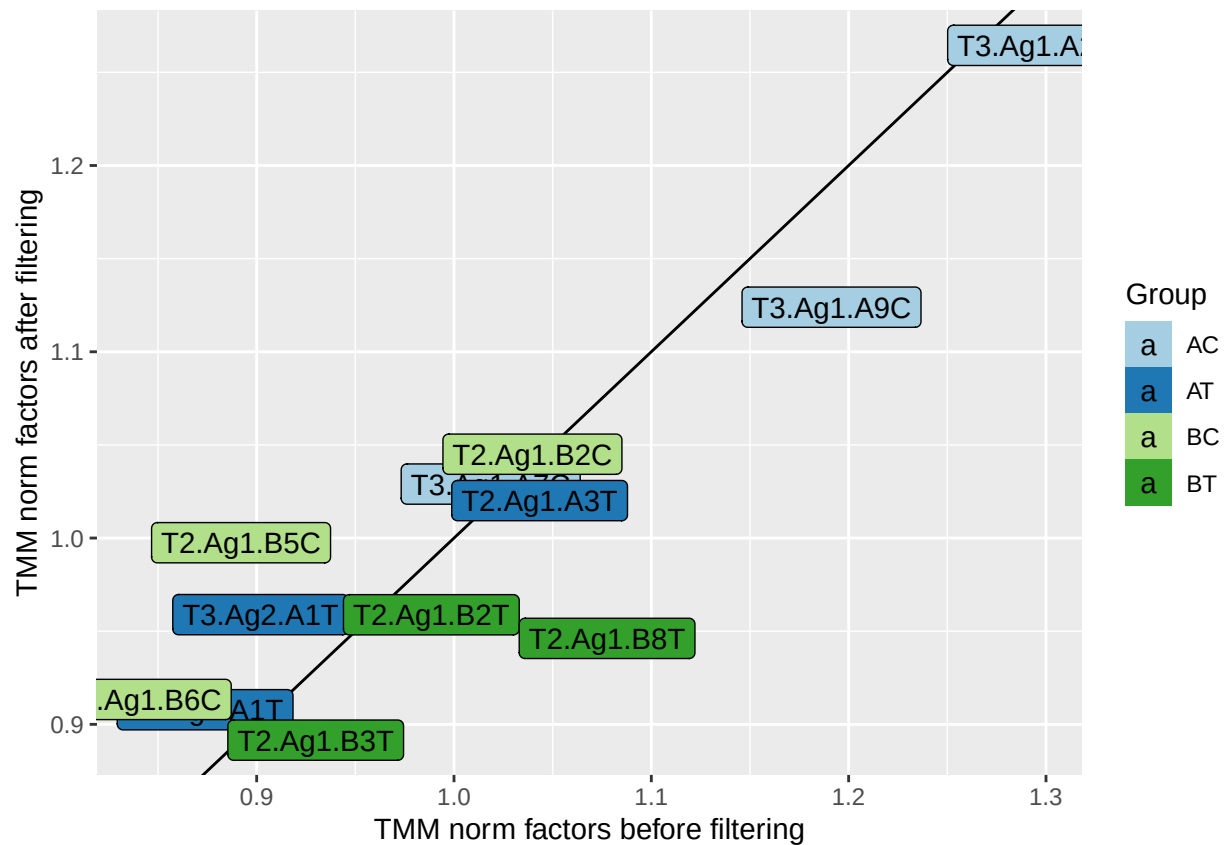


Figure 6: Comparison of TMM normalization factors before and after filtering.

Sample clustering

Multidimensional scaling in the `limma`¹¹ package was used to assess the largest effects on gene composition among samples. The normalized logCPM values from all 2415 lincRNAs between each pair of samples were chosen to construct the multidimensional scaling plot. The first axis separated the two crosses, while the second axis separated C and T samples (Fig. 7).

¹¹Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). `limma` powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43(7), e47.

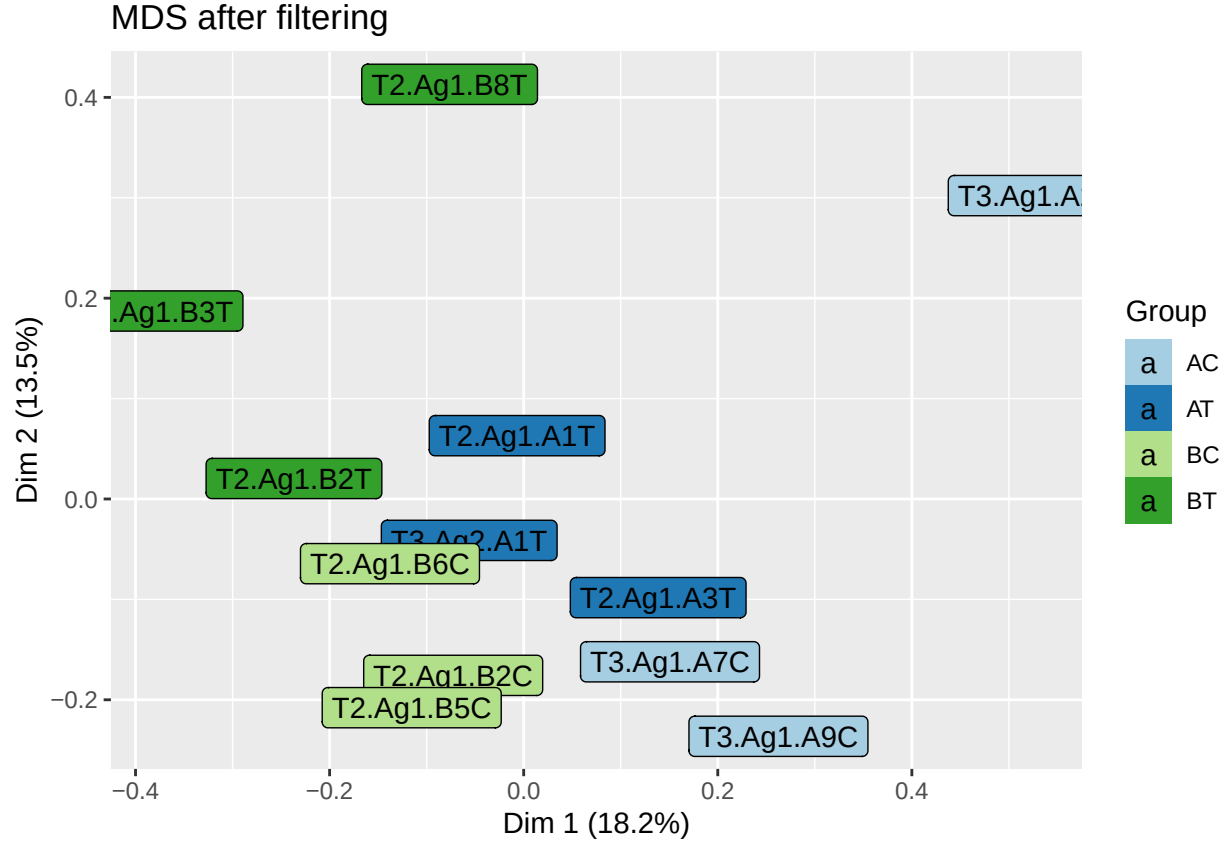


Figure 7: Multidimensional scaling of samples after filtering, using the top 5000 variable genes between each pair of samples.

Differential gene expression

Differential gene expression (DE) analysis was performed using the limma-trend method¹². Multiple testing correction was done using the False Discovery Rate method¹³. The model run was:

$$Y = \text{Group}$$

where *Group* refers to the AC, AT, BC, and BT groups. The following contrasts were then assessed:

- **BT - BC**: Differences between T and C bees from the B cross
- **AT - AC**: Differences between T and C bees from the A cross
- **BC - AC**: Differences between B and A crosses, with C as the baseline
- **BT - AT**: Differences between B and A crosses in T samples
- **(BT + AT) - (BC + AC)**: Differences between T and C bees across both crosses
- **(BT + BC) - (AT + AC)**: Differences between crosses, looking at both C and T
- **(BT - BC) - (AT - AC)**: Interaction effect; differential effect of cross on differences between T and C bees

¹²Chen Y, Lun ATL and Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; referees: 5 approved]. F1000Research 2016, 5:1438 (doi: 10.12688/f1000research.8987.2)

¹³Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300.

Numbers of significantly up- and down-regulated genes are shown in Table 2.

Table 2: Number of gene significant for differential expression at $FDR < 0.05$.

	BT - BC	AT - AC	BC - AC	BT - AT	(BT + AT) - (BC + AC)	(BT + BC) - (AT + AC)	(BT - BC) - (AT - AC)
Down	1	0	1	4	3	2	0
NotSig	2409	2413	2404	2408	2401	2396	2415
Up	5	2	10	3	11	17	0

Results for all genes are listed in the file “differential_lincRNA_expression_results_2022-07-06.xlsx”.

Heatmaps for differential lincRNA expression

Across all of the experimental groups contrasts, a one-way ANOVA was run in limma, identifying 41 significant lincRNAs. These lincRNA were then plotted in a heatmap to summarize gross changes in lincRNA expression across groups. (Fig. 8).

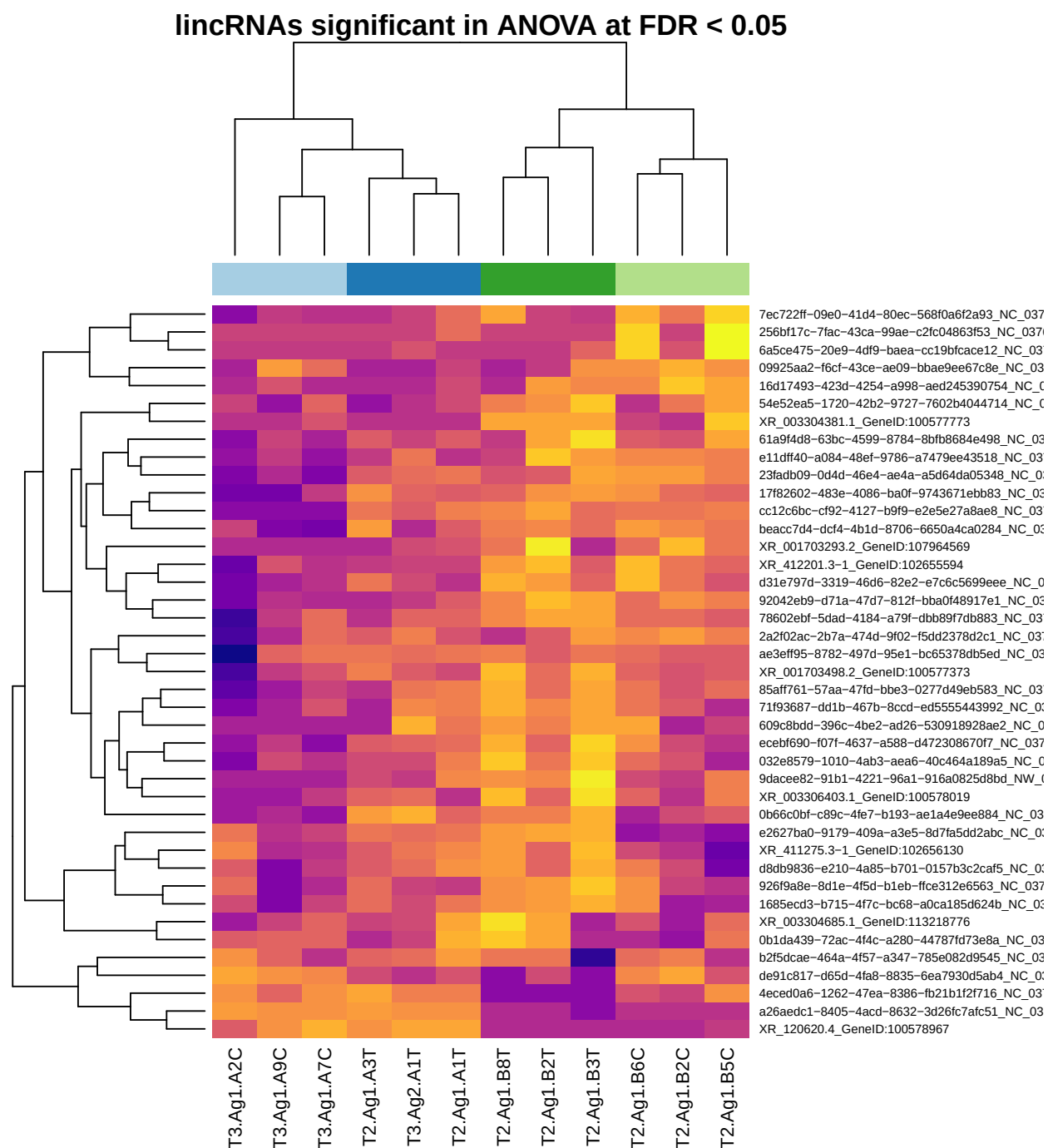


Figure 8: Heatmap of 29 lincRNAs with FDR < 0.05 in one-way anova for effect of experimental group.

Analysis of isoform switching

Few genes were significant for isoform switching, and those that were tended to be highly expressed, suggesting that more significant switches could have been detected with higher sequence read depth. However, the switches that were detected are potentially interesting, for example the switch from a coding transcript to a non-coding transcript for *Apid1* in cross B as compared to cross A, as well as group BC as compared to BT (see PDFs of switch consequences). Genes with significant switches with putative functional consequences are listed in Table 3. With respect to the analysis before fixing the flair bug, more significant switches with functional consequences were found for most contrasts.

Table 3: Genes with significant (FDR < 0.1) isoform switches with putative functional consequences.

	gene_id	gene_name	condition_1	condition_2	gene_switch_q_value
1	GeneID:406140	Apid1	AC	BC	0.0000000
3	GeneID:406140	Apid1	BC	BT	0.0000000
7	GeneID:406140	Apid1	AT	BT	0.0000000
8	GeneID:406121	Mrjp3	AT	BT	0.0000000
10	NC_001566.1:13000	NA	AC	BC	0.0000506
13	GeneID:406140	Apid1	A	B	0.0001549
2	GeneID:406121	Mrjp3	A	B	0.0002034
110	GeneID:409924	LOC409924	C	T	0.0004062
31	GeneID:409299	LOC409299	C	T	0.0008968
11	GeneID:406121	Mrjp3	AC	BC	0.0018311
4	NC_037644.1:5664000	NA	A	B	0.0025668
5	GeneID:552829	LOC552829	A	B	0.0029525
12	GeneID:725215	LOC725215	BC	BT	0.0032881
14	GeneID:409924	LOC409924	BC	BT	0.0080161
15	GeneID:727037	LOC727037	AT	BT	0.0116237
16	GeneID:410806	LOC410806	AT	BT	0.0199640
71	GeneID:724565	LOC724565	A	B	0.0335776
81	GeneID:413141	LOC413141	A	B	0.0553019
17	GeneID:413141	LOC413141	AT	BT	0.0709019
18	GeneID:727037	LOC727037	BC	BT	0.0813682
19	GeneID:551806	LOC551806	BC	BT	0.0852233
51	GeneID:413069	LOC413069	C	T	0.0905798
6	GeneID:551459	LOC551459	C	T	0.0905798
82	GeneID:551763	LOC551763	C	T	0.0905798

QC and differential gene expression

Quality control

Libraries varied considerably in overall size (Fig. 9) but were similar in quality (Fig. 10). Two samples had low levels of viral contamination (Fig. 10). Most reads mapping to the nuclear genome mapped to annotated transcripts (Fig. 10). Number of genes detected was linearly correlated with library size, indicating that saturation of gene detection was not reached (Fig. 11). The A-C group, which consisted entirely of new samples from December 2021, had the smallest library size and lowest number of genes detected (Figs. 9 and 11).

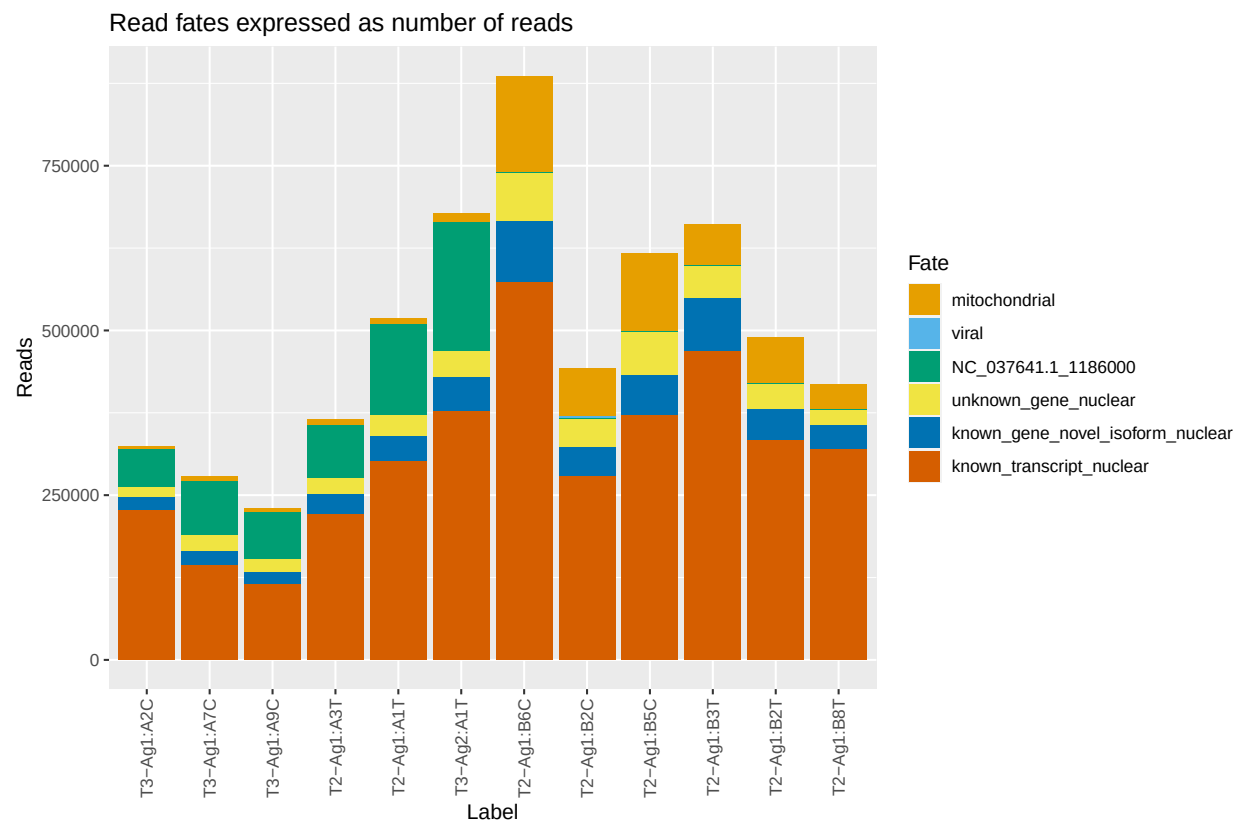


Figure 9: Library sizes expressed as number of aligned reads.

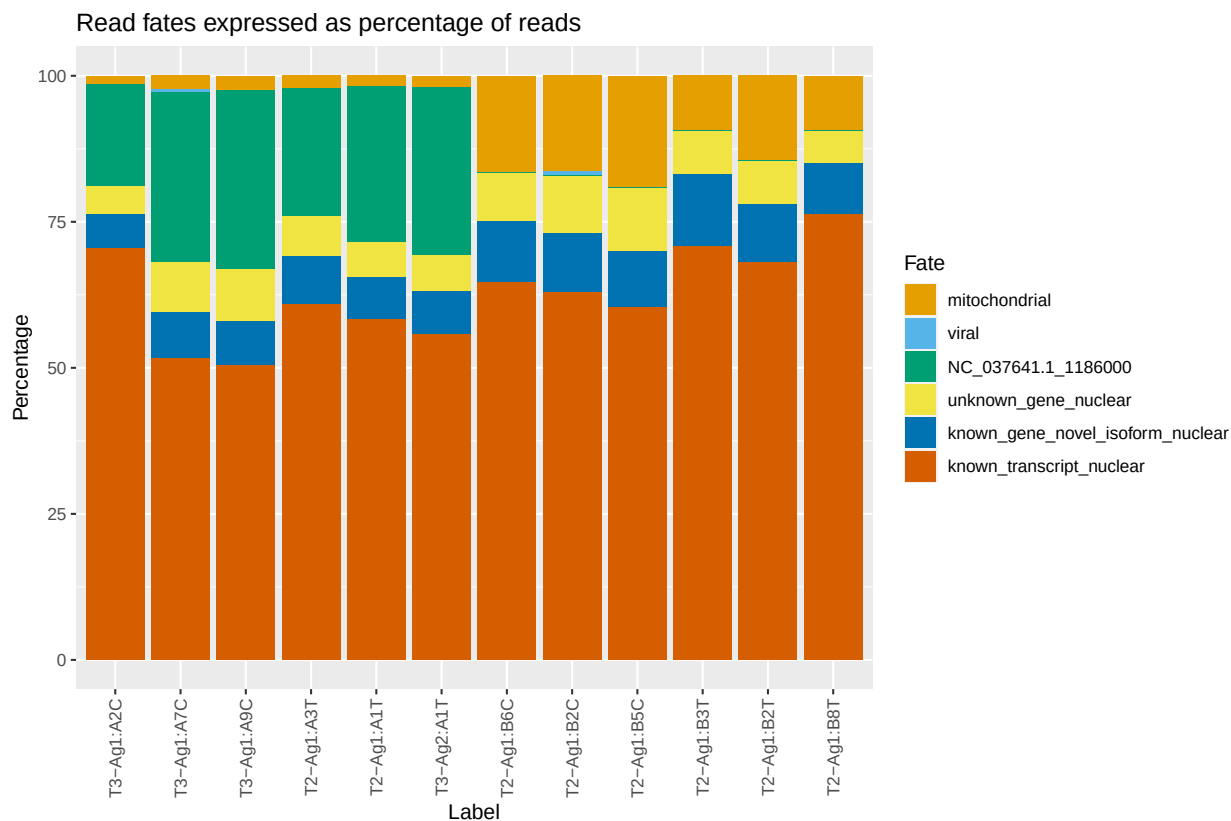


Figure 10: Percentages of reads aligned to categories of transcripts.

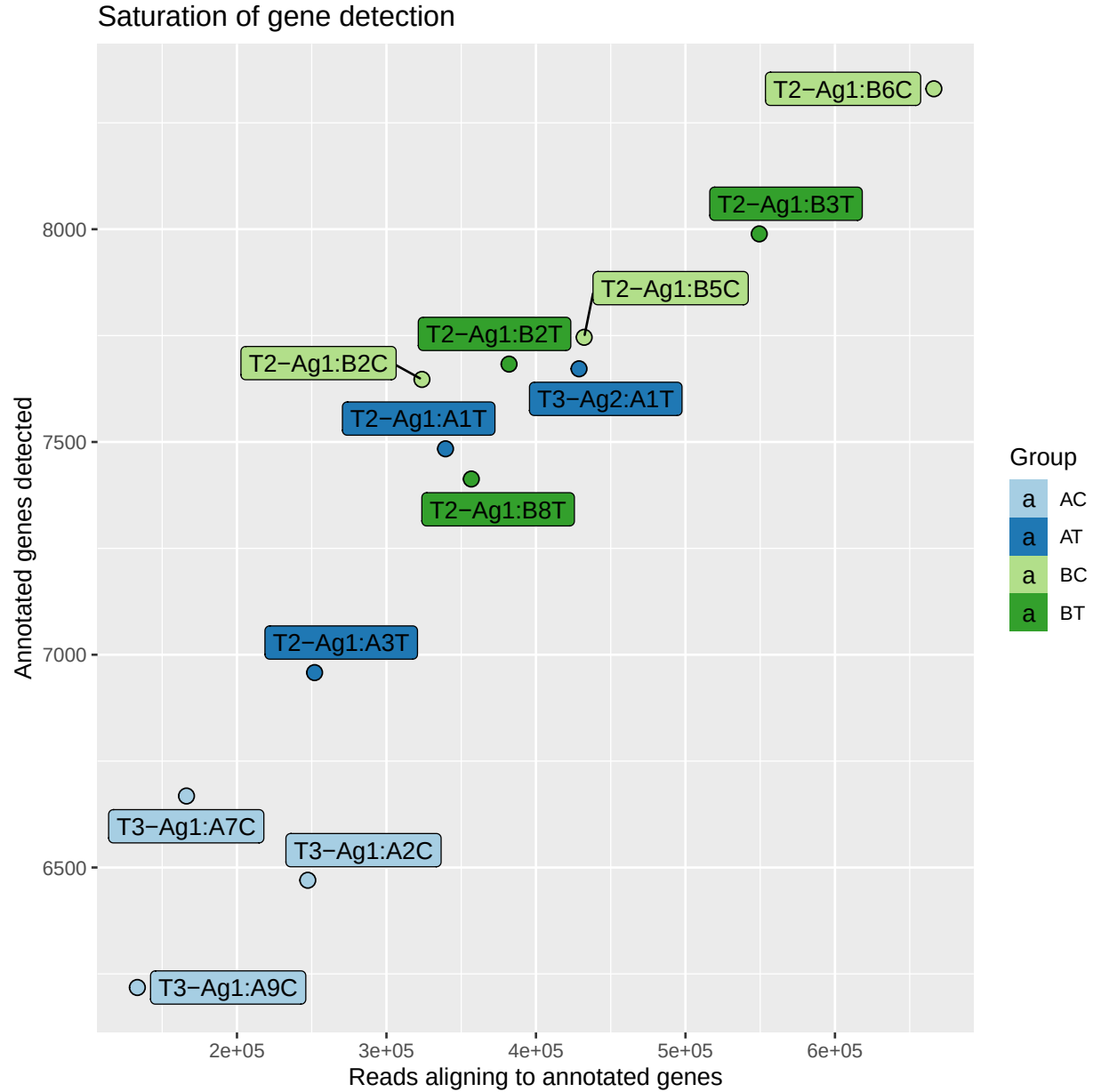


Figure 11: Number of genes detected vs. number of reads aligning to genes. Only genes from the published annotation are considered.

Mitochondrial reads

The vast majority of reads mapping to mitochondrial transcripts mapped correctly to the mother in both crosses (Figs. 12 and 13). Because of the mismapping of mitochondrial reads to the nuclear genome (NC_037641.1:1186000) in cross A, I excluded from downstream analysis all mitochondrial reads including those mapping to NC_037641.1:1186000. I found that this filtering reduced variance in TMM normalization factors (data not shown).

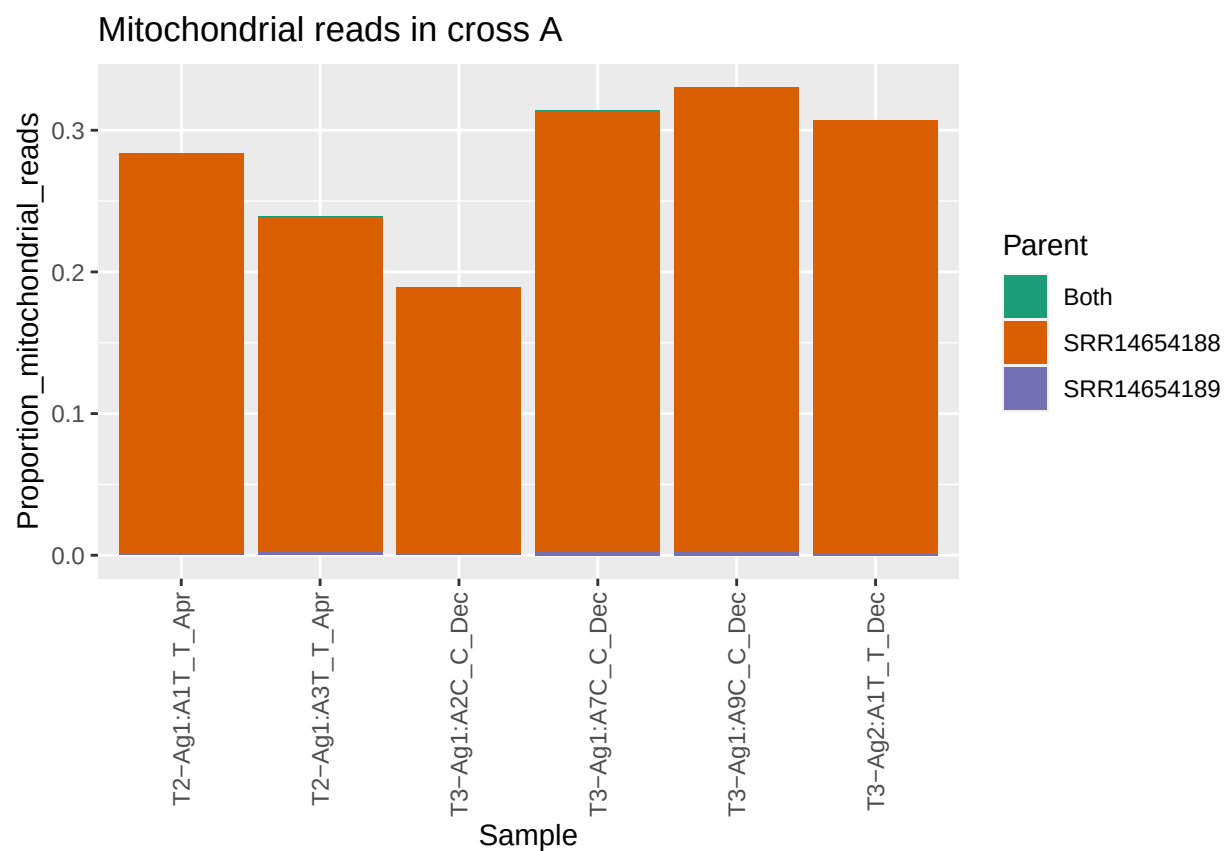


Figure 12: Proportion of read assigned to the maternal mitochondrion (SRR14654188), paternal mitochondrion (SRR14654189) or mitochondrion of unknown parent (Both) in cross A. Reads assigned to gene NC_037641.1:1186000 are counted as mitochondrial.

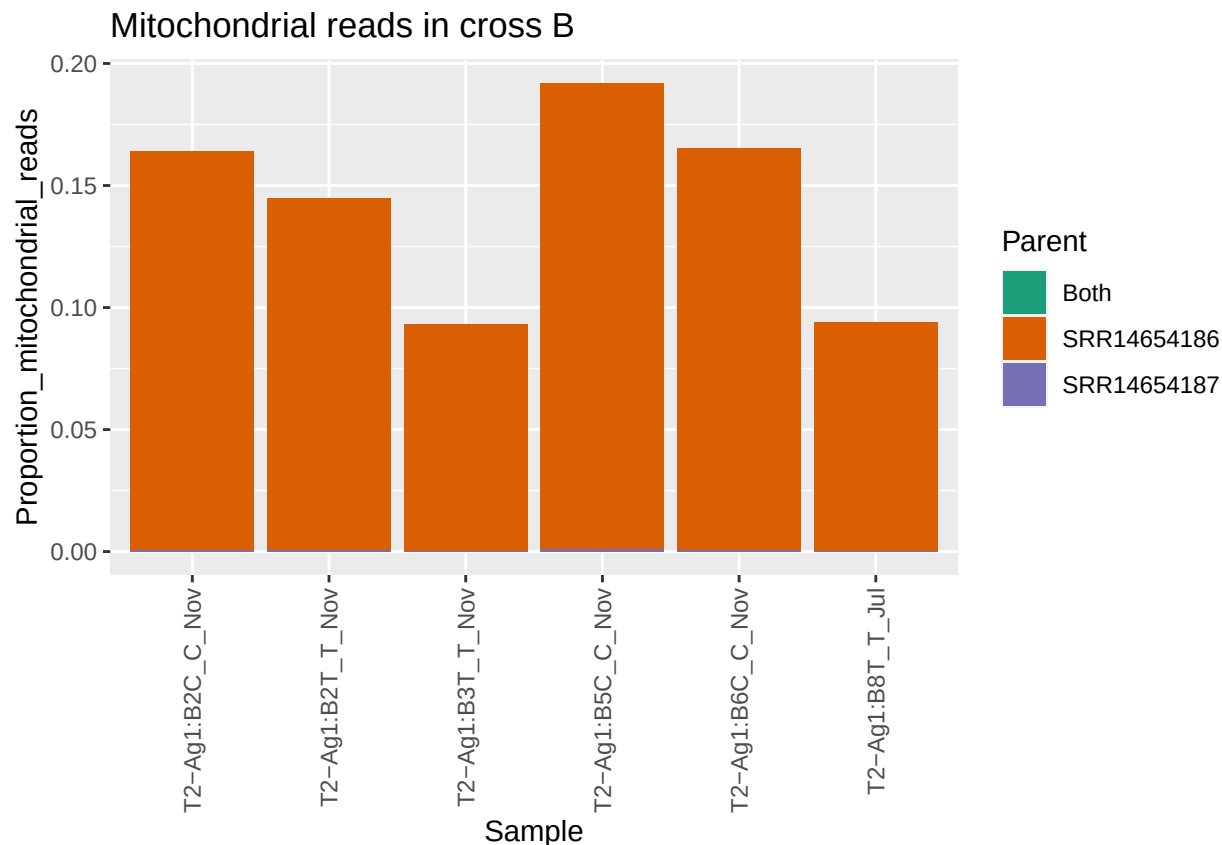


Figure 13: Proportion of read assigned to the maternal mitochondrion (SRR14654186), paternal mitochondrion (SRR14654187) or mitochondrion of unknown parent (Both) in cross B. Reads assigned to gene NC_037641.1:1186000 are counted as mitochondrial.

Normalization and filtering

When comparing expression levels, the numbers of reads per gene need to be normalized not only because of the differences in total number of reads, but because there could be differences in RNA composition such that the total number of reads would not be expected to be the same. The TMM (trimmed mean of M values) normalization¹⁴ in the edgeR package¹⁵ uses the assumption of *most genes do not change* to calculate a normalization factor for each sample to adjust for such biases in RNA composition. In this dataset, TMM normalization factors fluctuated between 0.79 and 1.21. TMM normalization factors are multiplied by the library size in order to generate the effective library size, where smaller TMM normalization factors indicate more overrepresented genes in the sample. Normalization factors varied among samples but were not particularly associated with experimental group (Fig. 14).

¹⁴Robinson MD, Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11, R25.

¹⁵Robinson MD, McCarthy DJ, Smyth GK (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” *Bioinformatics*, 26(1), 139-140.

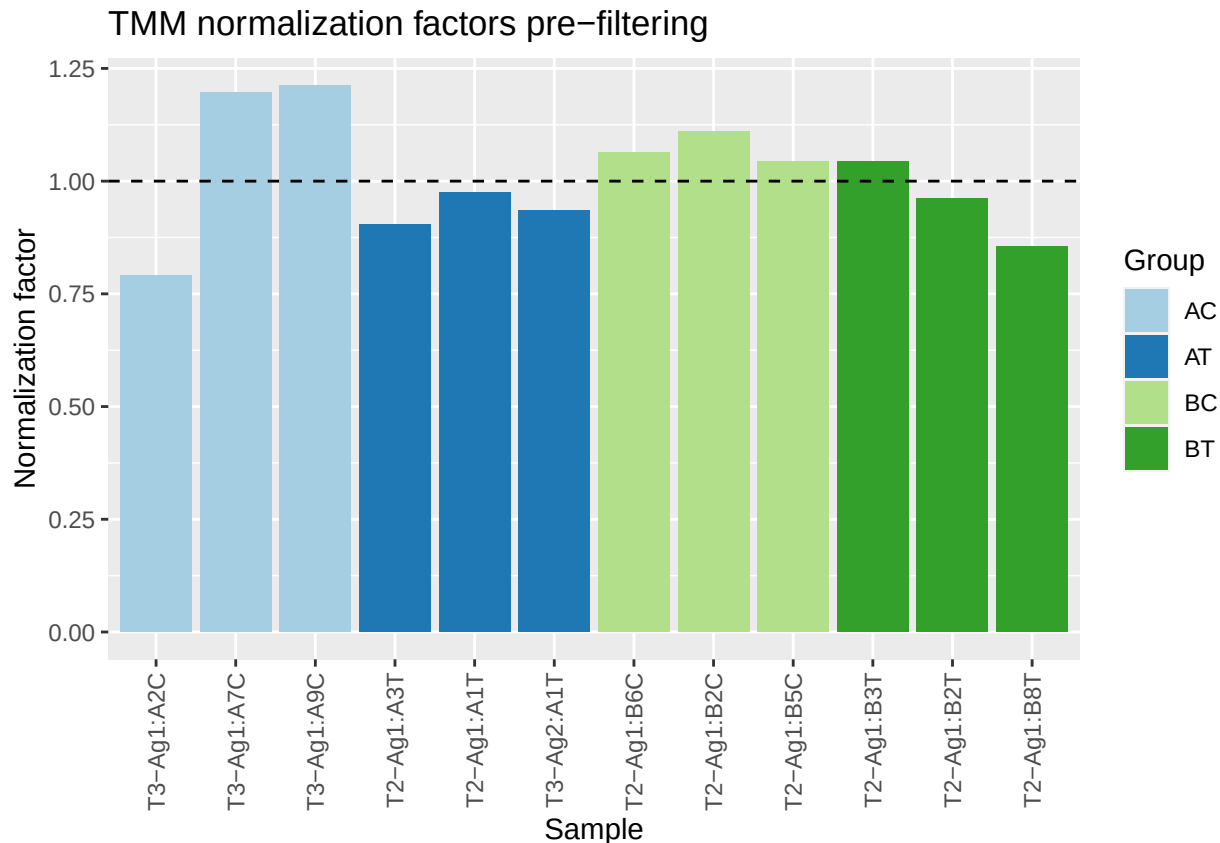


Figure 14: TMM normalization factors pre-filtering.

While the `flair collapse` output a total of 19,902 genes (after filtering mitochondrial genes), many of these might not have detectable expression in these samples. Therefore, we set the detection threshold at 8 cpm (counts per million) in at least 3 samples, which resulted in 5,440 genes being filtered out, leaving 14,462 genes to be analyzed for differential expression, containing 99.42% of the reads. Of these genes, 8058 were from the published annotation, representing 90.30% of reads prior to filtering. This threshold resulted in a fairly even number of genes being detected across samples, considering the variability in library size (Fig. 15). In a typical Illumina RNA-seq experiment, we might use 1 cpm as the filtering threshold. Given the lower read depth of this experiment due to the use of Oxford Nanopore sequencing, a higher threshold was needed to differentiate between genes that could be detected in few samples vs. genes that could be detected in most samples. I arrived at a threshold of 8 by examining the detection rates at a variety of thresholds ranging from 4 to 12 (results not shown).

After filtering, TMM normalization was performed again and normalized log2-based count per million values (logCPM) were calculated using `edgeR`'s¹⁶ `cpm()` function with `prior.count = 2` to help stabilize fold-changes of extremely low expression genes. TMM normalization factors after filtering are shown in Fig. 16, and compared to pre-filtering values in Fig. 17.

The overall distribution of logCPM values differed by batch (Fig. 18). The peak on the left of Fig. 18 represents undetected genes, and thus is higher in batches later than November 2020 because those library sizes were smaller.

¹⁶McCarthy DJ, Chen Y, Smyth GK (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." *Nucleic Acids Research*, 40(10), 4288-4297.

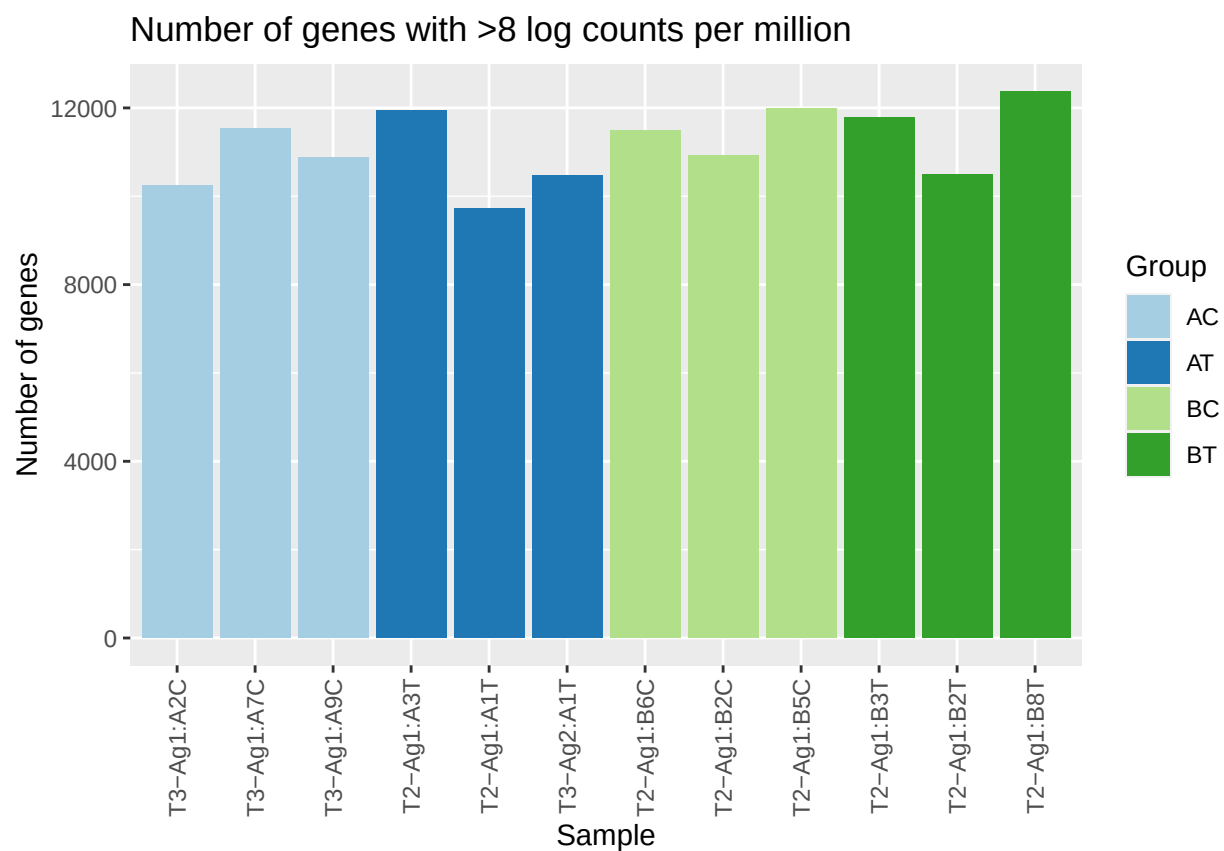


Figure 15: Number of detected genes passing the filtering threshold in each sample.

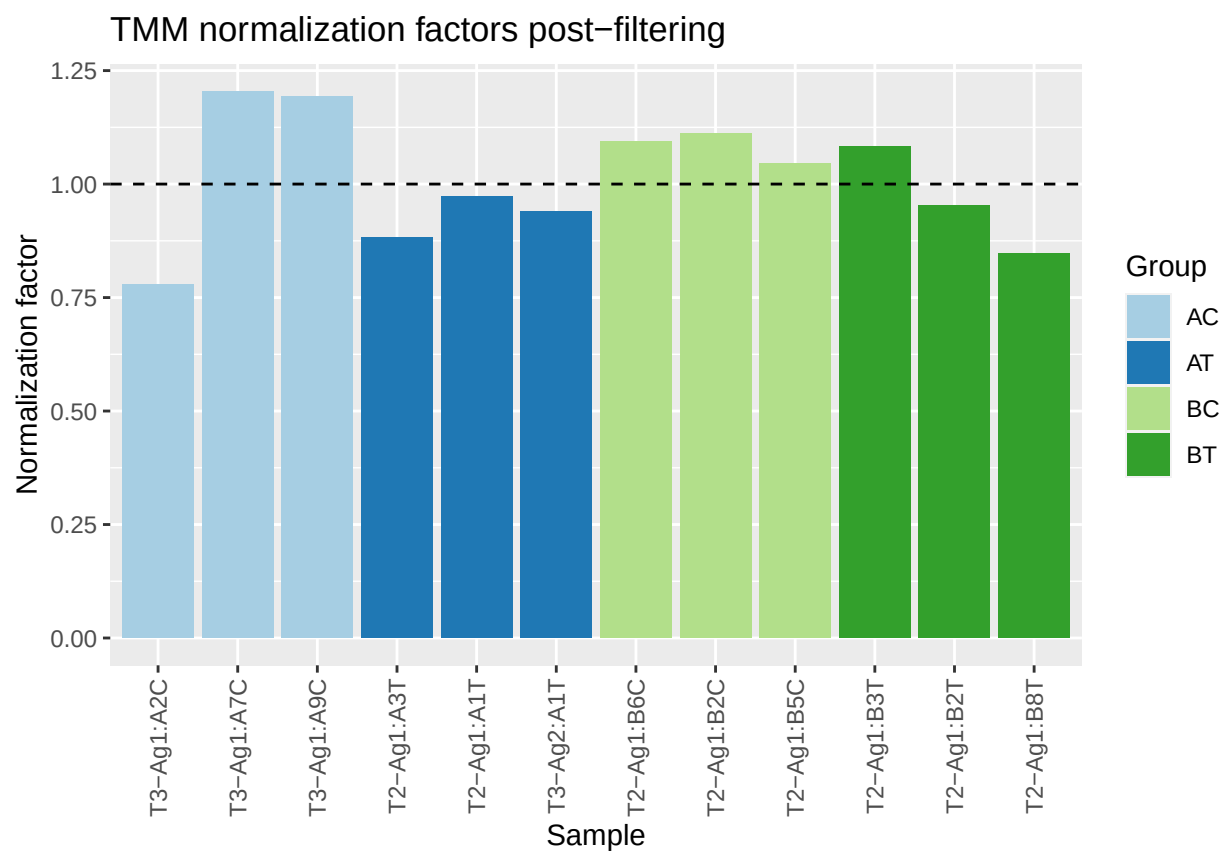


Figure 16: TMM normalization factors post-filtering.

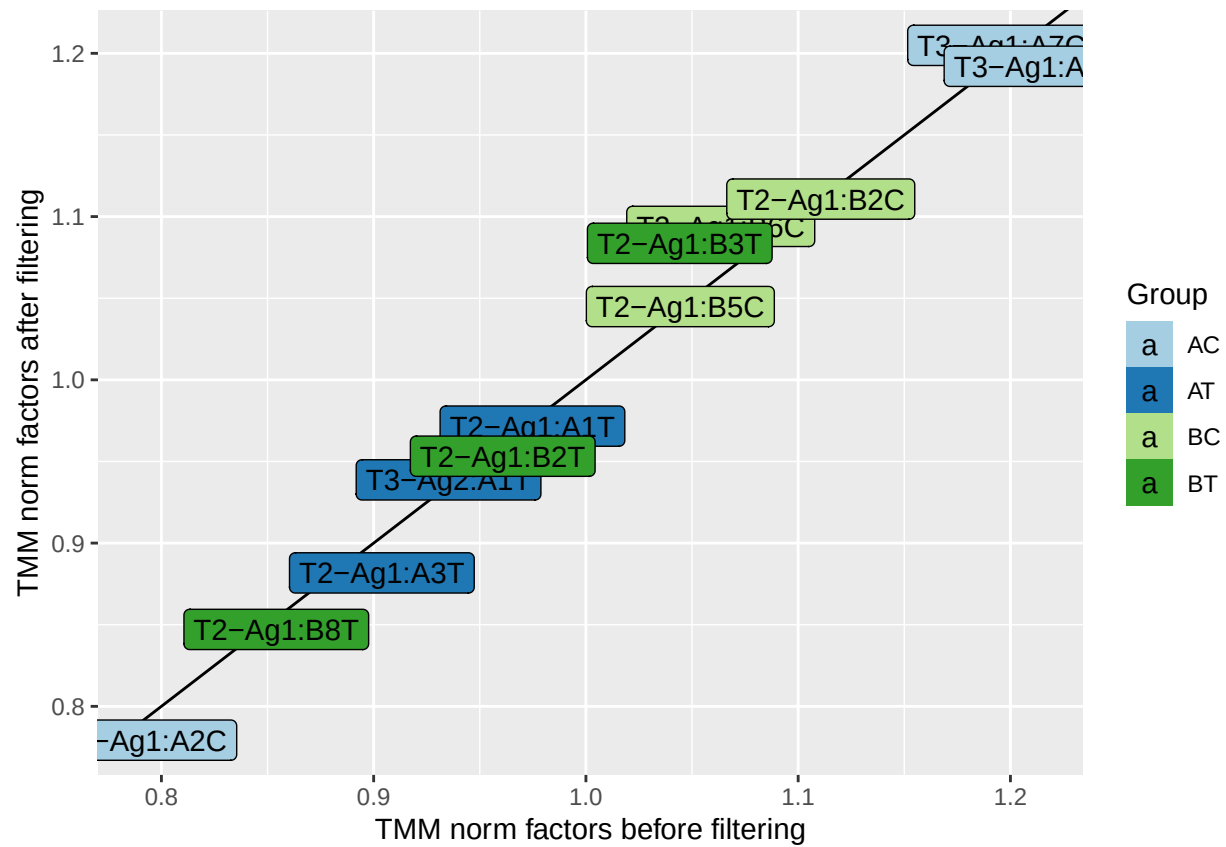


Figure 17: Comparison of TMM normalization factors before and after filtering.

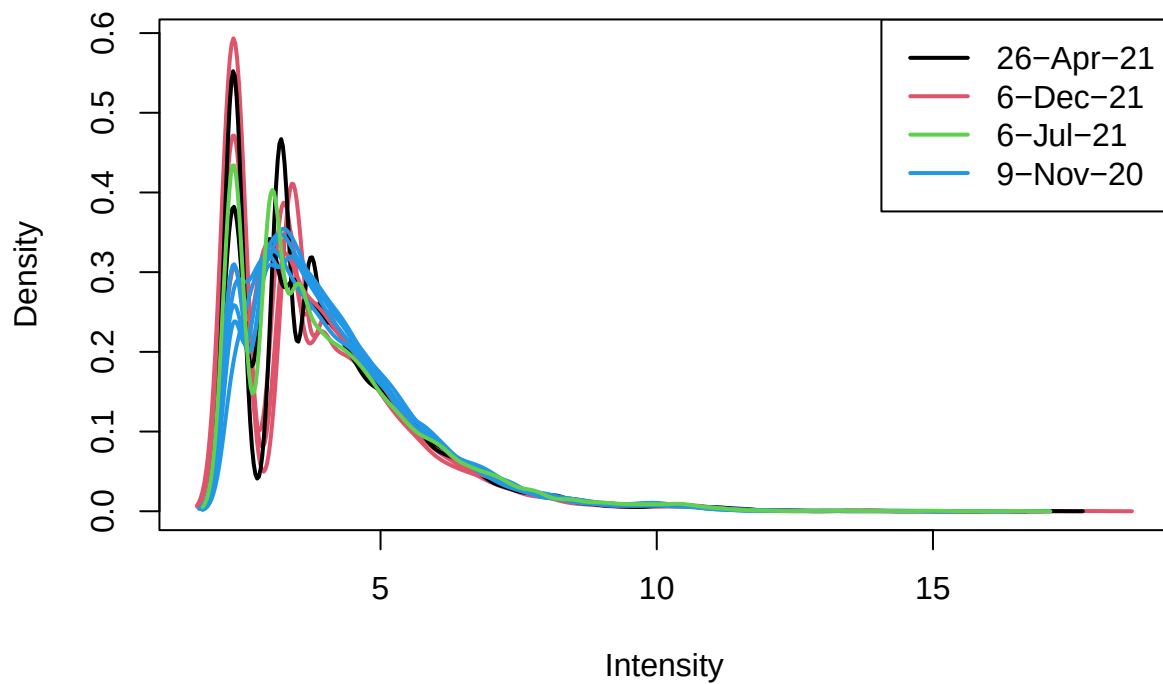


Figure 18: Distribution of logCPM values after filtering.

Sample clustering

Multidimensional scaling in the `limma`¹⁷ package was used to assess the largest effects on gene composition among samples. The normalized logCPM values of the top most 5,000 variable genes between each pair of samples were chosen to construct the multidimensional scaling plot. The first axis separated the two crosses, while the second axis separated sample T3-Ag1:A2C from the rest and to a lesser extent separated C and T samples (Fig. 19).

¹⁷Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). `limma` powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43(7), e47.

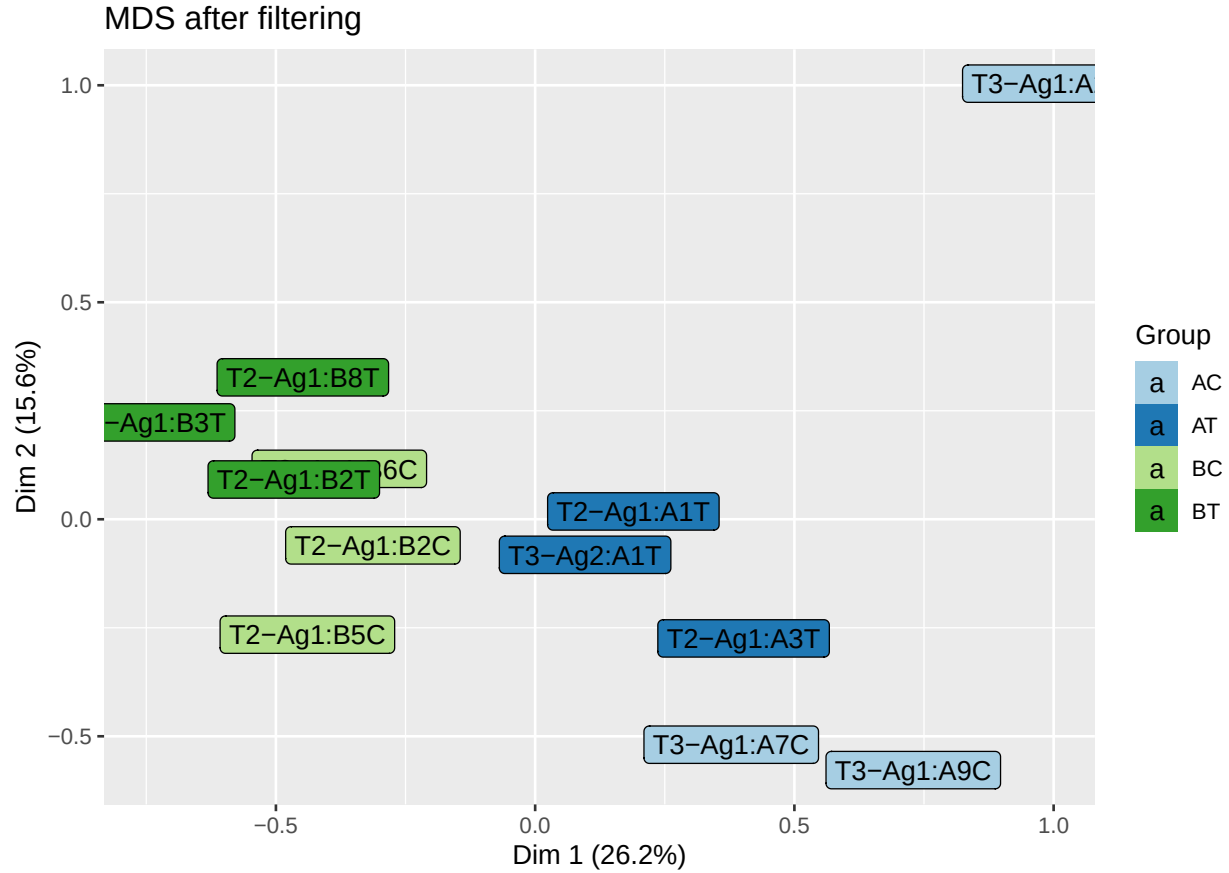


Figure 19: Multidimensional scaling of samples after filtering, using the top 5000 variable genes between each pair of samples.

Differential gene expression

Differential gene expression (DE) analysis was performed using the limma-trend method¹⁸. Multiple testing correction was done using the False Discovery Rate method¹⁹. The model run was:

$$Y = \text{Group}$$

where *Group* refers to the AC, AT, BC, and BT groups. The following contrasts were then assessed:

- **BT - BC**: Differences between T and C bees from the B cross
- **AT - AC**: Differences between T and C bees from the A cross
- **BC - AC**: Differences between B and A crosses, with C as the baseline
- **(BT + AT) - (BC + AC)**: Differences between T and C bees across both crosses
- **(BT + BC) - (AT + AC)**: Differences between crosses, looking at both C and T
- **(BT - BC) - (AT - AC)**: Interaction effect; differential effect of cross on differences between T and C bees

¹⁸Chen Y, Lun ATL and Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; referees: 5 approved]. F1000Research 2016, 5:1438 (doi: 10.12688/f1000research.8987.2)

¹⁹Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300.

Numbers of significantly up- and down-regulated genes are shown in Table 4. Fewer genes were significant after fixing the Flair bug as compared to before fixing the bug. However, nearly all significant genes were for differences between crosses, which could have been artificially inflated by the Flair bug, because alignment quality would have been better for some parents than for others depending on how many variants had been used to edit the transcript sequence.

Table 4: Number of gene significant for differential expression at $FDR < 0.05$.

	BT - BC	AT - AC	BC - AC	(BT + AT) - (BC + AC)	(BT + BC) - (AT + AC)	(BT - BC) - (AT - AC)
Down	0	0	133	0	306	0
NotSig	14461	14461	13922	14462	13465	14462
Up	1	1	407	0	691	0

Heatmaps for differential gene expression

Across all of the experimental groups contrasts, a one-way ANOVA was run in limma, identifying 619 significant genes. These genes were then plotted in a heatmap to summarize gross changes in gene expression across groups. Most of the significant genes displayed consistent expression differences between cross A and cross B (Fig. 20).

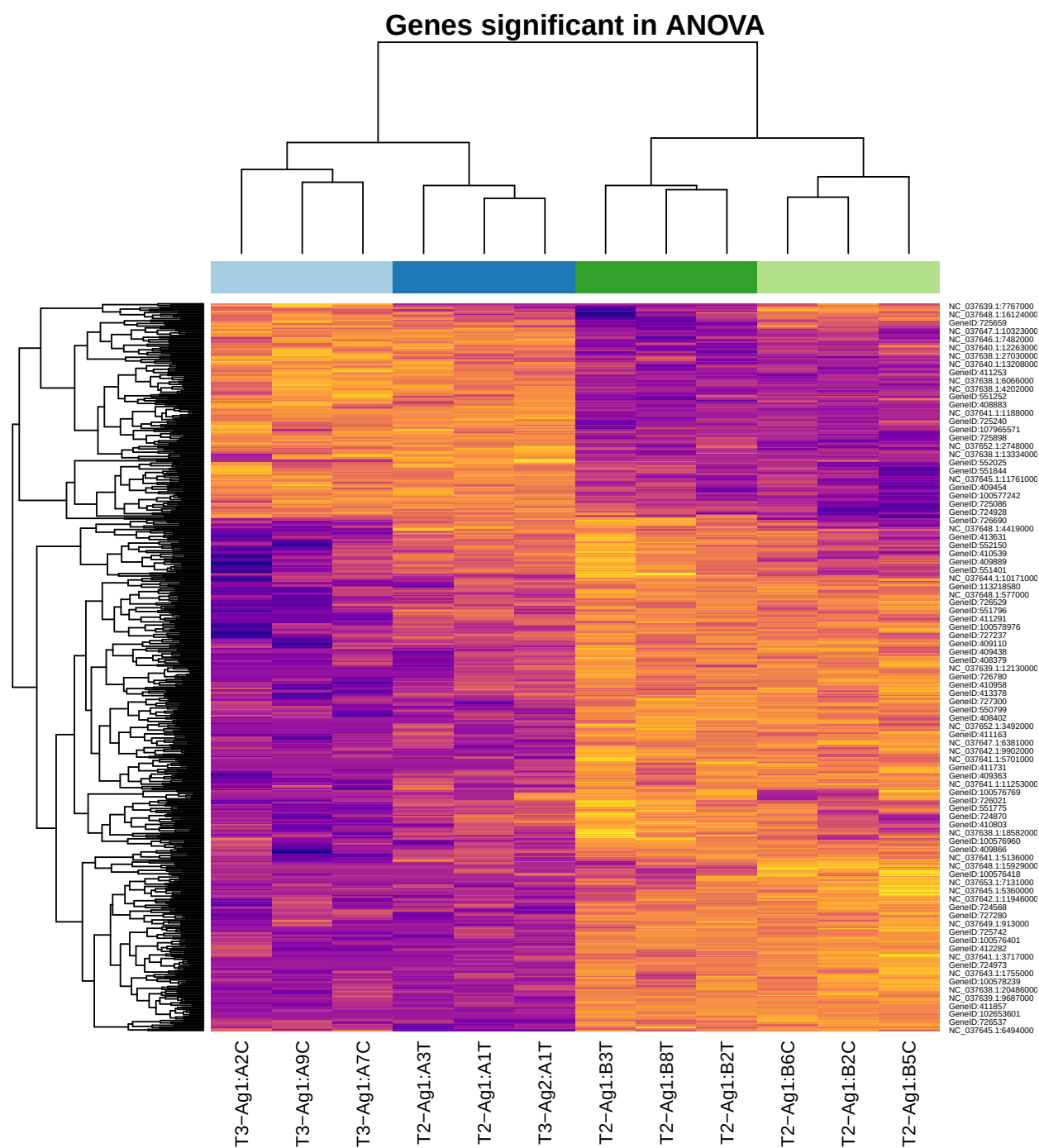


Figure 20: Heatmap of 582 genes with $FDR < 0.05$ in one-way anova for effect of experimental group. Unadjusted expression values are shown.