Neuromac project on the differential gene expression in germ free mices

Bérénice Batut

Contents

Presentation of the biological context

Questions

- What are the differentially expressed genes of the microglya cells between germ free (GF) and conventional (SPF) mice at different ages?
- Is there a difference between the ages?
- Is there a difference between the genders?
- Which genes? Which pathways?

Hypotheses

- Differential gene expression between SPF and GF already known for 8 weeks mices (article)
- Expected gender bias in 104w SPF

Data

- RNA-seq data of microglya cells of mices: (Figure 1 and details
  - 2 types
    - Conventional (SPF)
    - Germ free (GF)
  - 3 ages
    - 8 weeks (8w)
    - 52 weeks (52w)
    - 104 weeks (104w)
  - 2 genders
    - Male (m)
    - Female (f)
- Library preparation and sequencing
  - Library preparation with Illumina Nextera XT Sample Preparation
  - First strand
  - Single-end data

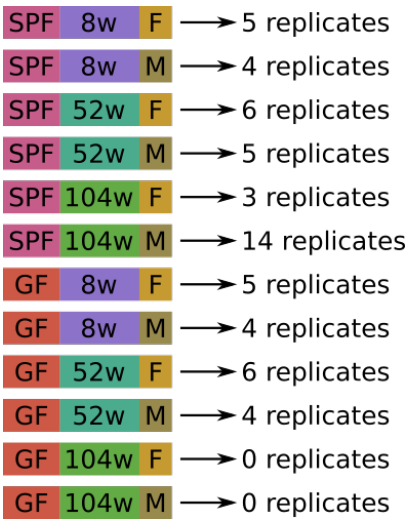

Figure 1: Repartition of the replicates in the different groups

- Sequencing with HiSeq 1000

Analyses

Workflow applied on each dataset: Figure 2

Quality control and trimming

Quality control

Objectives

- Check the quality of the raw sequences

Details

- FastQC on every datasets
- MultiQC (Ewels et al. 2016) report to aggregate the FastQC reports

Results: (details in the Google doc in "FastQC report")

- Small sequences; 50 bp
- Between 22.4M and 52.2M of reads
- Good average quality score per sequence (Figure 3)
- No big reduction of quality at the end of the sequences (Figure 4)
- Per Base Sequence Content: 52 samples with a warning and 2 with an error (non homogenous proportion of each base at the beginning of the sequences)
- Per Sequence GC Content (Figure 5): all samples with a warning (small increase of %GC around 10-20 bp, likely related to the overrepresentation of C and G sequences at the begining seen in the previous check)
- Per Base N Content (Figure 6): 10 samples with a warning
- Numerous read duplications (Figure 7): expected for RNA seq data
- Overrepresented sequences (Figure 8): 55 samples with a warning and 1 with an error
  - Need to check with BLAST what are the overrepresented sequences

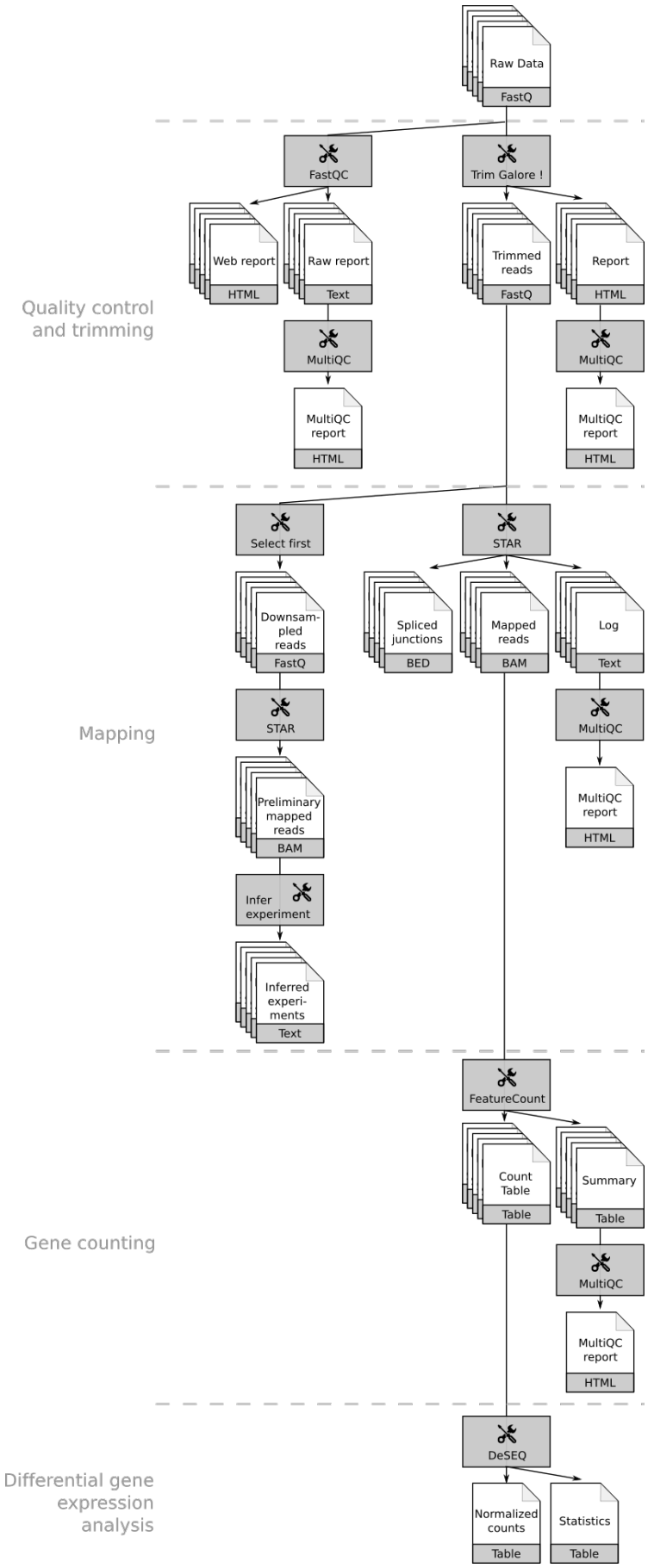

Figure 2: Workflow applied on each dataset

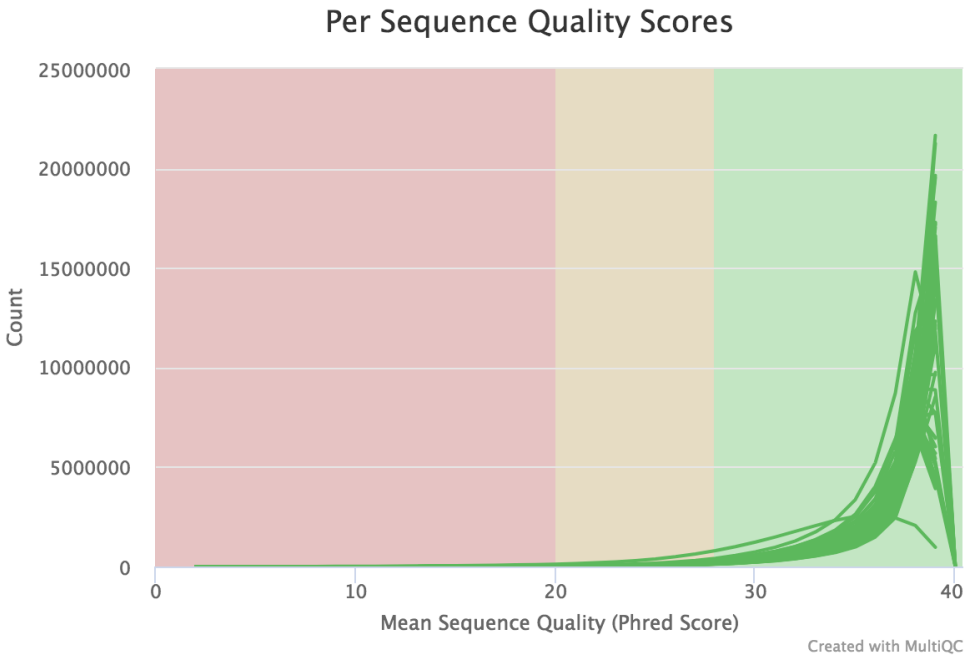

Figure 3: Per Sequence Quality Scores (generated with FastQC and MultiQC)

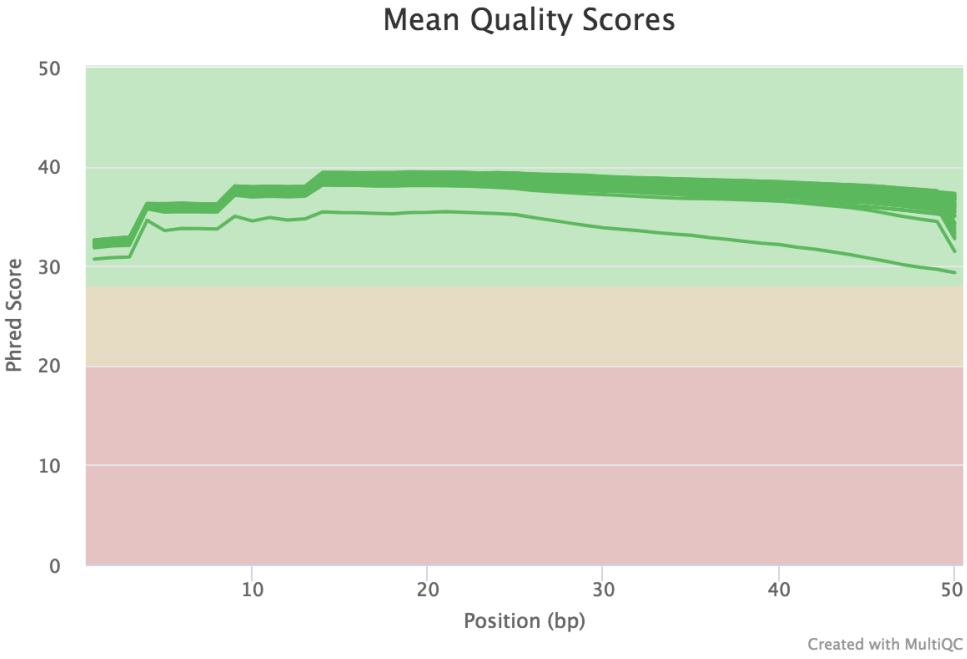

Figure 4: Sequence Quality Histograms (generated with FastQC and MultiQC)

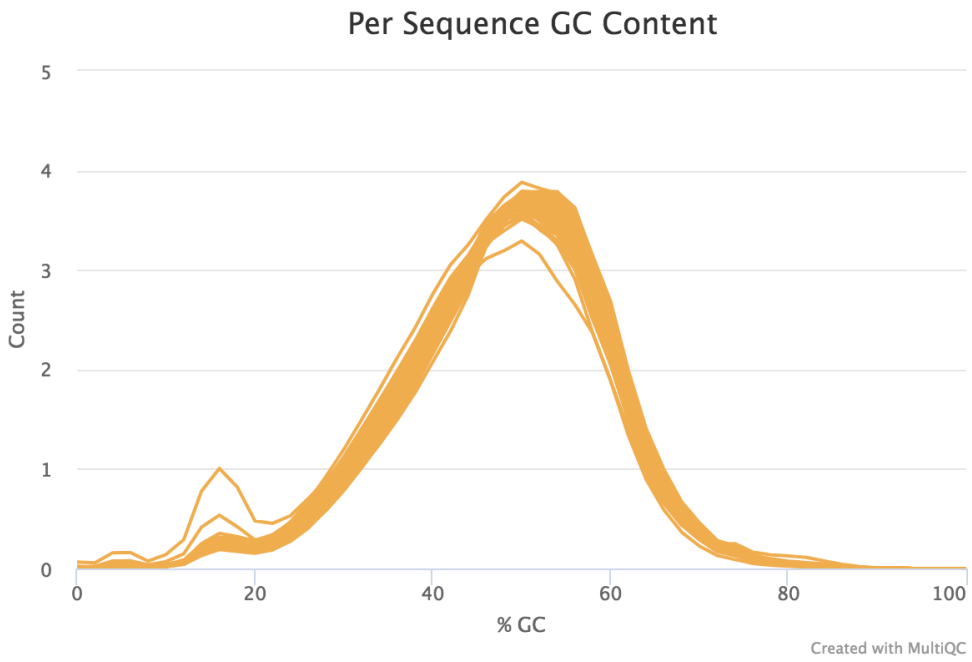

Figure 5: Per Sequence GC Content (generated with FastQC and MultiQC)

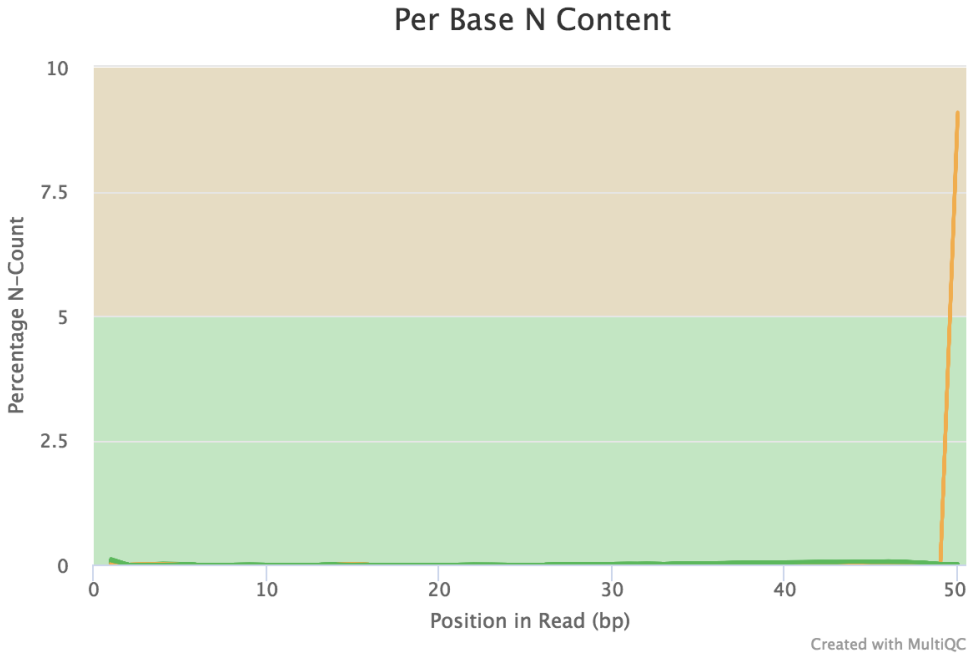

Figure 6: Per Base N Content (generated with FastQC and MultiQC)

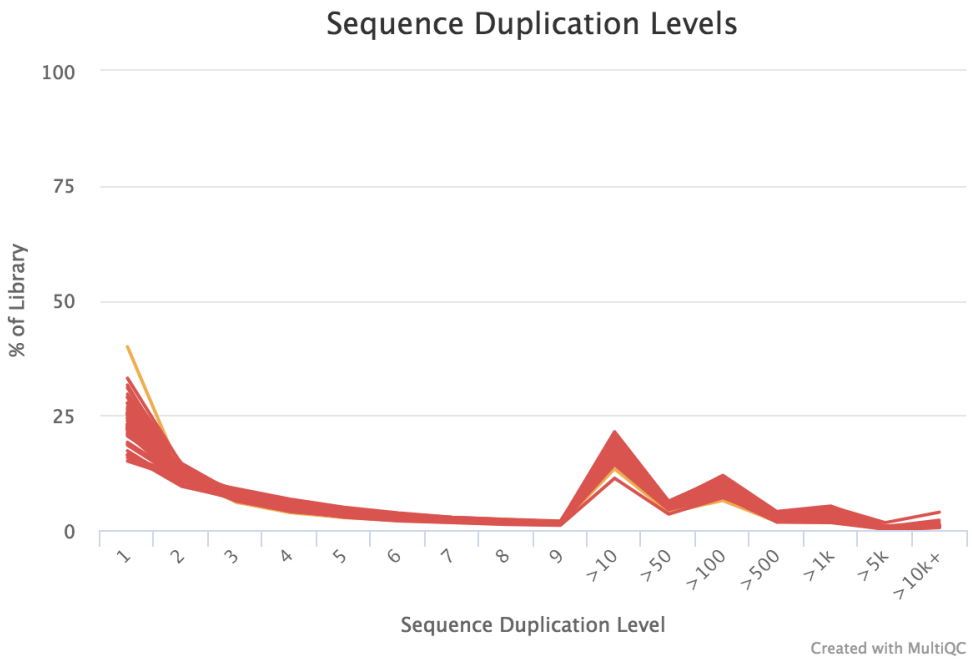

Figure 7: Sequence Duplication Levels (generated with FastQC and MultiQC)

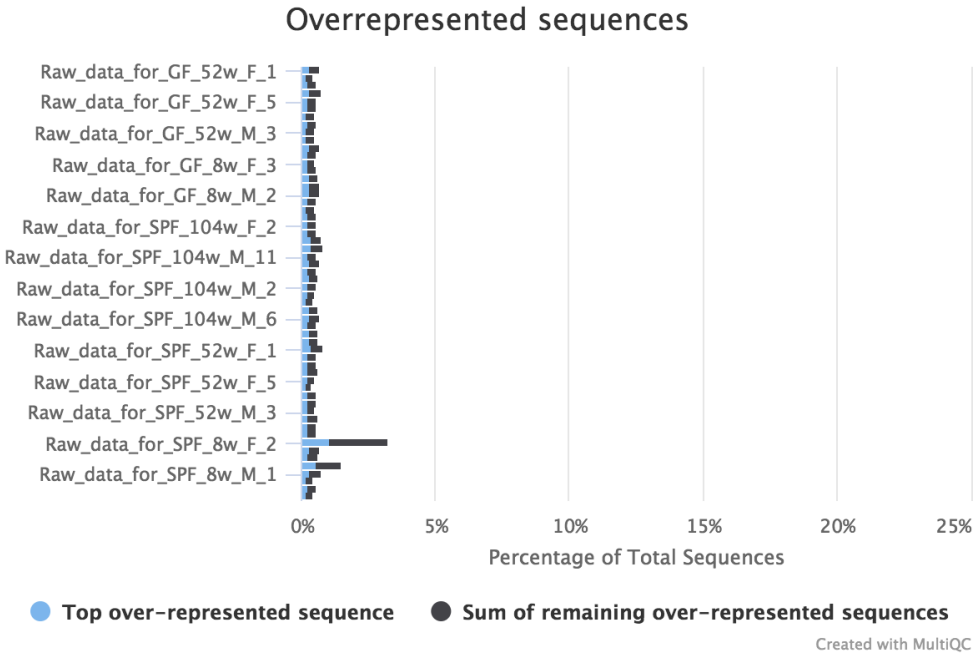

Figure 8: Overrepresented sequences (generated with FastQC and MultiQC)

- Few remaining adapters at the end of the sequences (Figure 9)

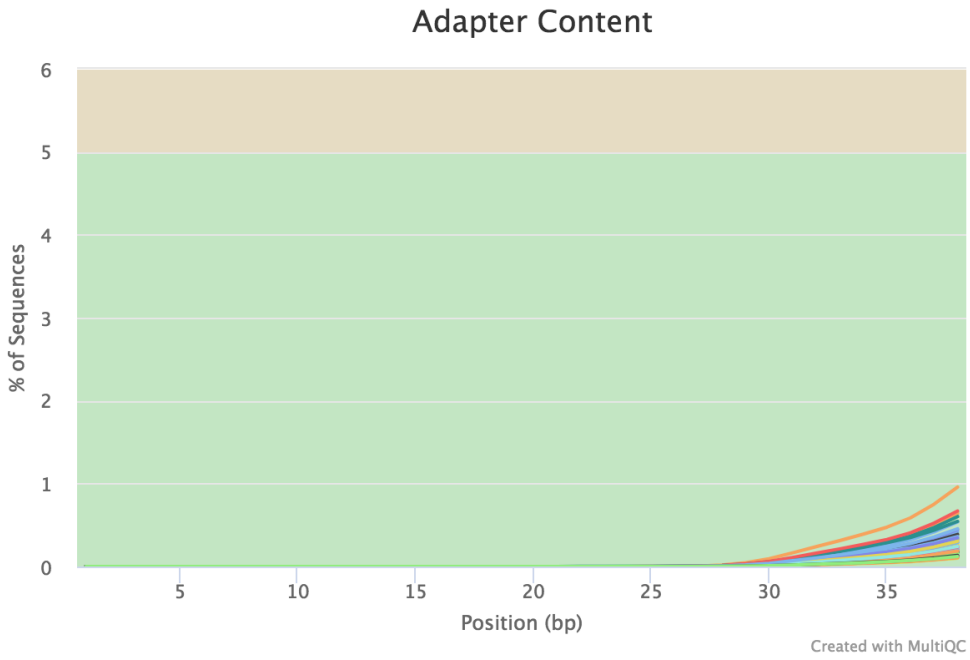

Figure 9: Adapter Content (generated with FastQC and MultiQC)

Trimming

Objectives

- Eliminate the bad quality ends, even if globally quality reports are good. Not so much removed

Details

- Trim Galore! on every datasets
- MultiQC (Ewels et al. 2016) report to aggregate the Trim Galore! reports

Results

- Few 3'-ends bases eliminated
- Few sequences eliminated (too small after trimming)
- To do
  - Add table with raw results
  - Fix MultiQC report aggregation

Mapping

Objectives

- Map the reads on the reference genome of Drosophila

Preliminary mapping

Objectives

- Infer if the library is strand specific or not by
  - Extraction of some reads
  - Mapping them on the reference genome
  - Use an annotation file

  - Infer on gene of which strand the mapping reads fit on

Details

- Downsampling of the dataset: Extraction of 200,000 reads with "Select first" tool
- Mapping with
  - STAR (Dobin et al. 2013)
  - mm10 as reference genome
- Infer the strand with "Infer Experiment" of RSeQC (L. Wang, Wang, and Li 2012)

Results

- Unstranded library
- To do
  - Add table with raw results
  - Add MultiQC report aggregation

Actual mapping

Objectives

- Map the trimmed reads on the reference genome to annotate them

Details

- Mapping with
  - STAR (Dobin et al. 2013), a splice aware mapper
  - mm10 as reference genome
  - mm10_UCSC_07_15_genes as gene model for splice junctions

Results: (Figure 10 and details in the Google doc in "STAR report")

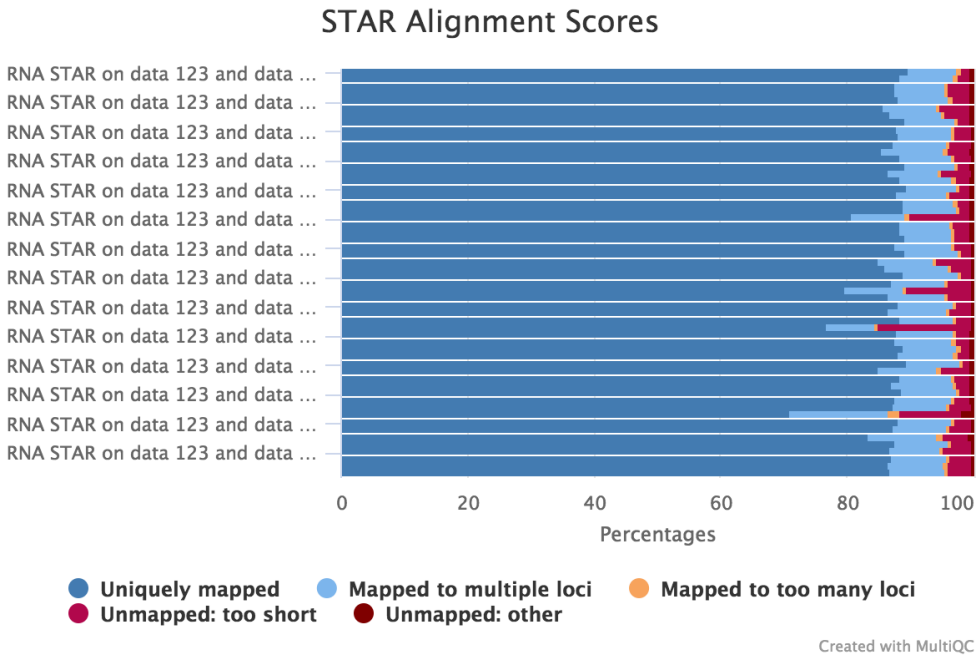

Figure 10: Alignment Scores (generated with STAR and MultiQC)

- Good percentage of reads that are uniquely mapped (> 70.9%)
- Good percentage of reads that are mapped (uniquely + multi-mapped, > 85%)
- Relatively low percentage of unmapped reads
- Keep an eye on

  - SPF_8w_F_2
  - SPF_52w_F_6
  - SPF_52w_F_1
  - SPF_8w_M_1

Gene counting

Objectives

- Count the number of reads that are mapped on genes

Details

- Counting with
  - FeatureCounts (Liao, Smyth, and Shi 2013)
  - mm10_UCSC_07_15_genes as Gene annotation file
  - Unstranded protocol

Results (Figure 11 and details in the Google doc in "FeatureCounts report")

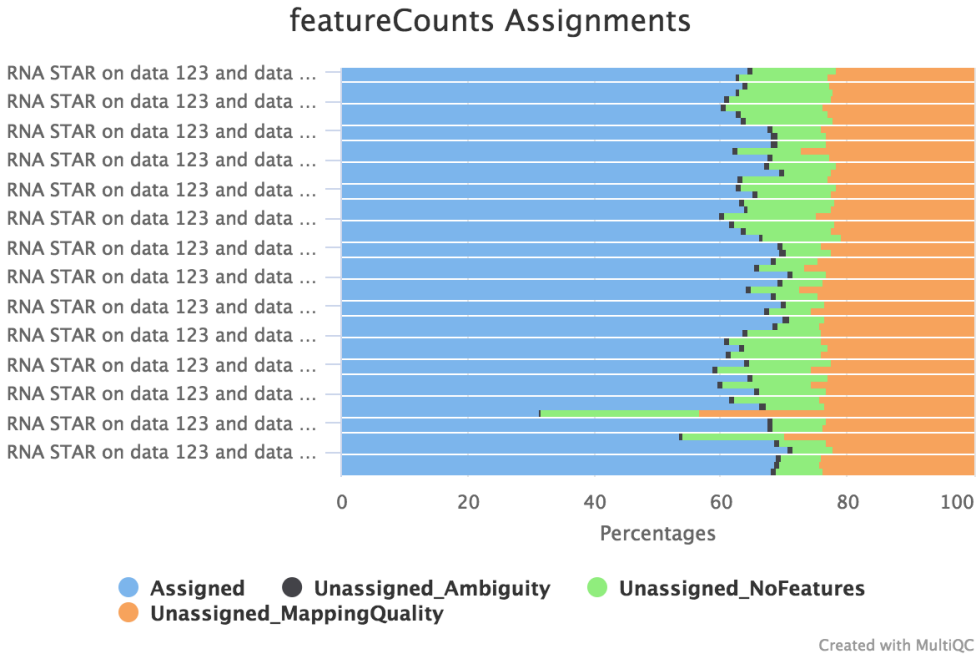

Figure 11: Reads assigned (generated with FeatureCounts and MultiQC)

- Good percentage of assigned reads to genes (> 60 % except 2 samples)
- Keep an eye on
  - SPF_8w_F_2 (already issue for the mapping)
  - SPF_8w_F_5

Differential expression analyses

Progressive complexification of the analyses

Understanding the impact of age: for each gender and each mouse type (GF & SPF)

Questions

- What are the differences between the ages?

- Which genes and pathways are differentially expressed?
- In which proportion?

Objectives

- Get a better idea of the basal differences between the ages

Details

- Design (Figure 12)

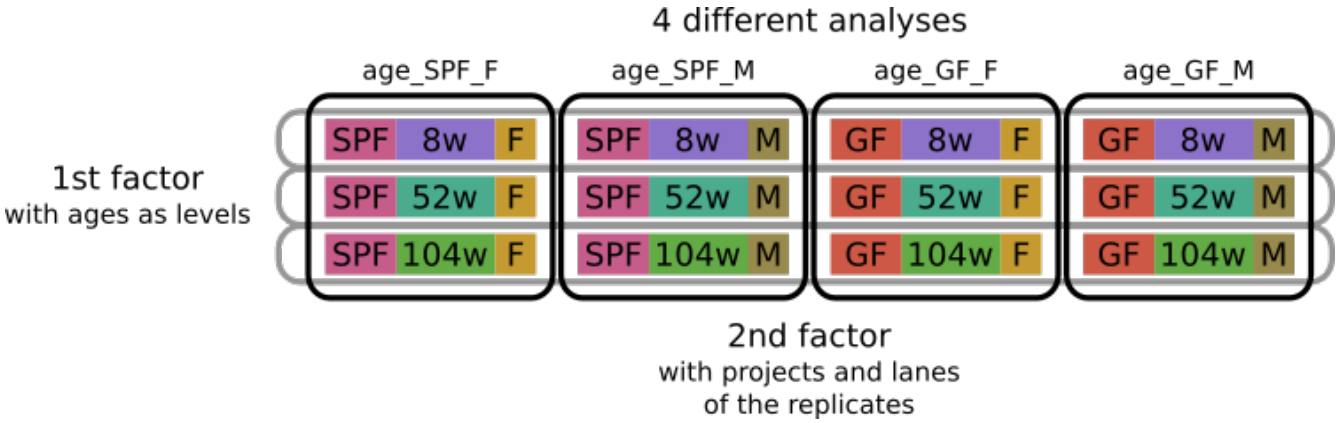

Figure 12: Analyses to understand the impact of age on gene expression

- Analyses with
  - DESeq2 (Love, Huber, and Anders 2014)
  - Factor 1: age
  - Factor 2: lane with sequencing project (*e.g.* `project_s195_7`) if different from the age factors

Results (Figure 13)

Understanding the impact of gender: for each age and each type

Questions

- How complex are the differences between the gender?
- Can we have the both gender in a global analysis?

Objectives

- Get a better idea of the impact of gender on the gene expression, if it is different between the ages and to know if the impact of the gender if more important than the type of mices (GF/SPF)

Details

- Design (Figure 14)
- Analyses with
  - DESeq2 (Love, Huber, and Anders 2014)
  - Factor 1: gender
  - Factor 2: lane with sequencing project if different from the age factors

Results (Figure 15)

Understanding the impact of type

Questions

- What is the impact of being GF on the expression levels?

Figure 13: Comparison of the differential expressed genes between the ages

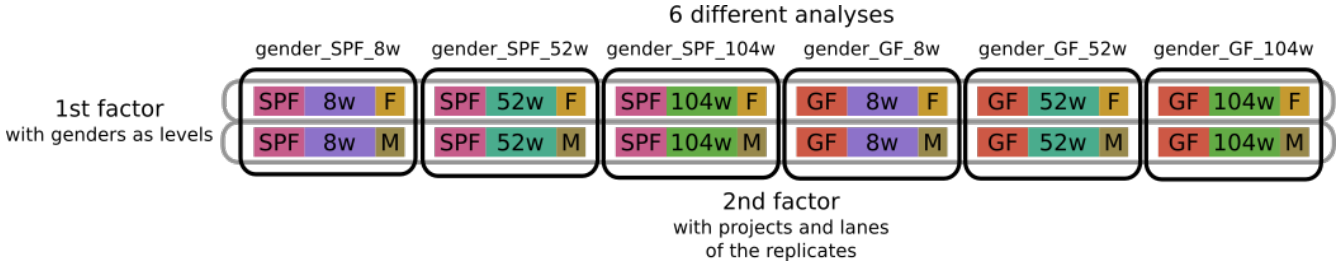

Figure 14: Analyses to understand the impact of gender on gene expression

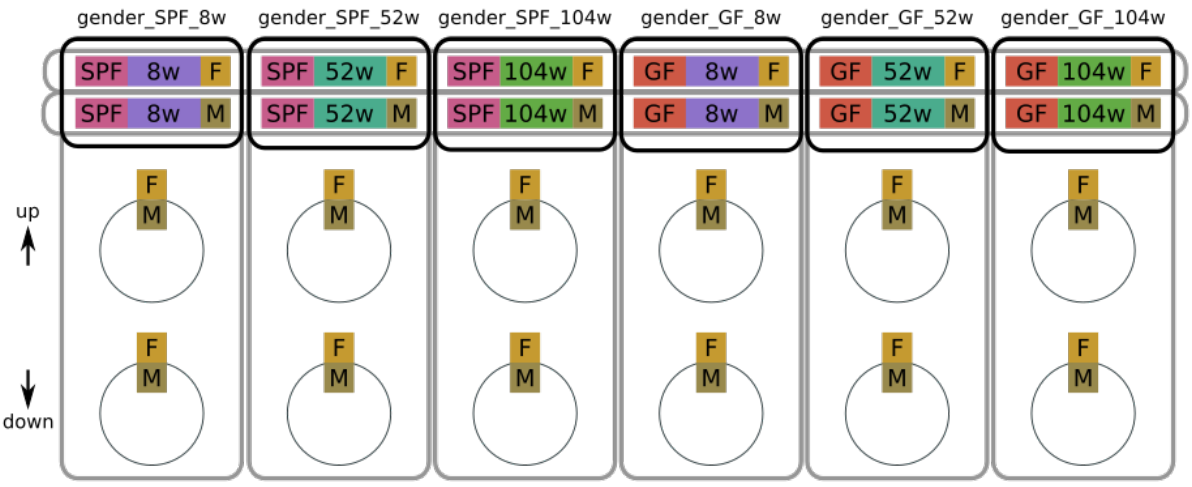

Figure 15: Comparison of the differential expressed genes between the genders

For each gender and age

Questions

- What are the differentially expressed genes of the microglya cells between GF and SPF mice at different ages?

Objectives

- Have limited cofounding factors (age/gender)
- Get an idea of the impact of SPF/GF on the gene expression

Details

- Design (Figure 16)

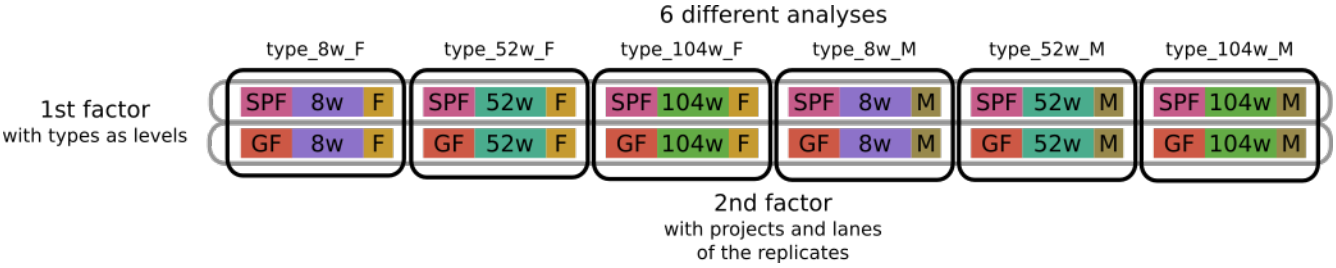

Figure 16: Analyses to understand the impact of GF on gene expression, with limitation of the cofounding factors

- Analyses with
  - DESeq2 (Love, Huber, and Anders 2014)
  - Factor 1: gender
  - Factor 2: lane with sequencing project if different from the age factors

Results (Figure 17)

For each gender

Questions

- What are the levels of differential expression between GF and SPF between the ages?

Objectives

- Be able to compare the gene expression of GF/SPF between the different ages

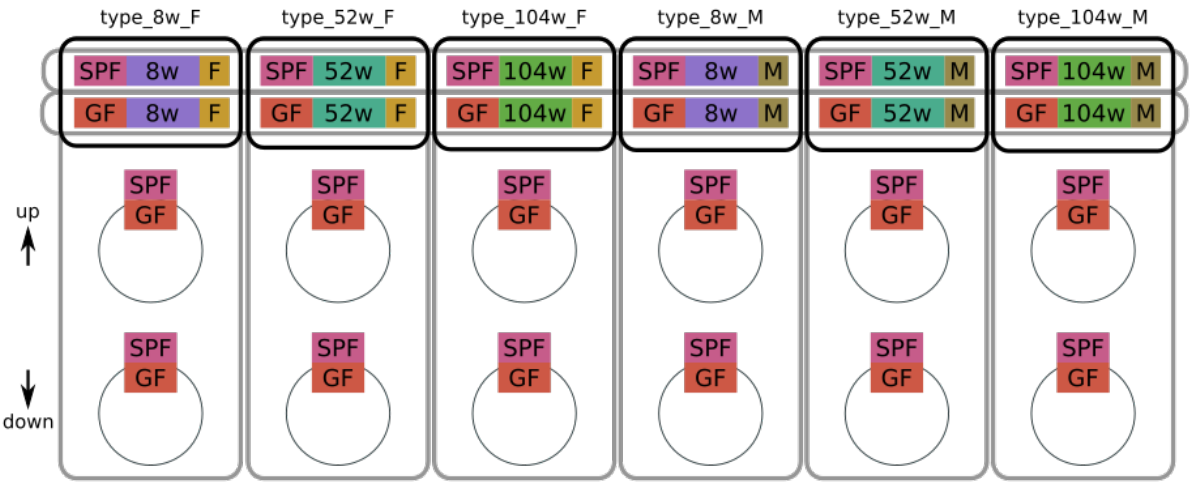

Figure 17: Comparison of the differential expressed genes between the GT and SPF

Details

- Design (Figure 14)

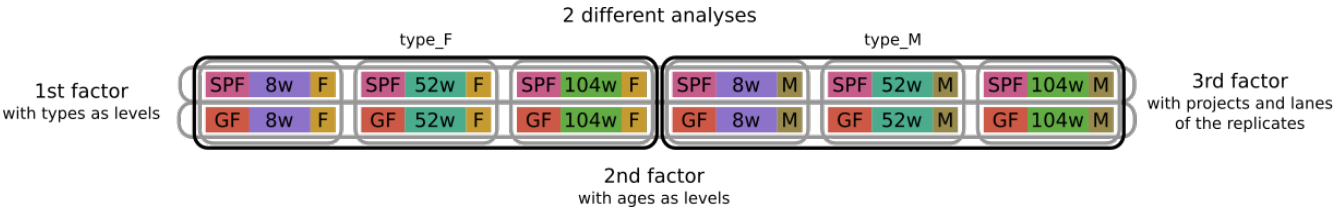

Figure 18: Analyses to understand the impact of GF on gene expression

- Analyses with
  - DESeq2 (Love, Huber, and Anders 2014)
  - Factor 1: GF or SPF
  - Factor 2: age
  - Factor 3: lane with sequencing project

Results (Figure 19)

Post differential expression analyses

To keep in mind

- Comparison of changes between the ages after normalization by the values in 8 weeks
- Comparison of the levels of differential expression between gender comparison and GF/SPF to check if the first ones are not higher than the latter ones to have a big analysis

References

Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. 2013. "STAR: Ultrafast Universal Rna-Seq Aligner." *Bioinformatics* 29 (1). Oxford University Press: 15–21.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: Summarize Analysis Results

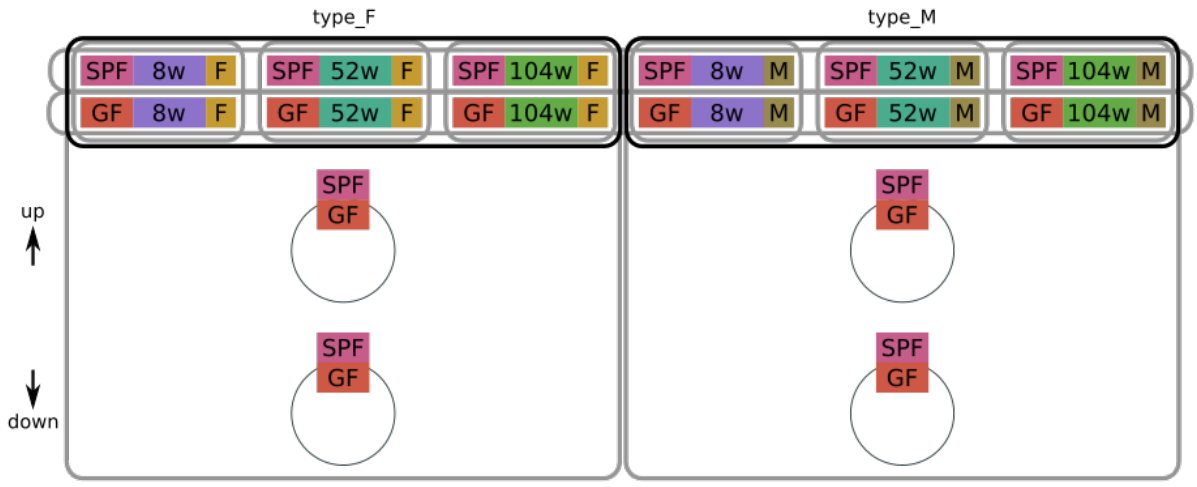

Figure 19: Comparison of the differential expressed genes between the GT and SPF

for Multiple Tools and Samples in a Single Report." *Bioinformatics* 32 (19). Oxford University Press: 3047–8.

Liao, Yang, Gordon K Smyth, and Wei Shi. 2013. "FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics* 30 (7). Oxford University Press: 923–30.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for Rna-Seq Data with Deseq2." *Genome Biology* 15 (12). BioMed Central: 550.

Wang, Liguo, Shengqin Wang, and Wei Li. 2012. "RSeQC: Quality Control of Rna-Seq Experiments." *Bioinformatics* 28 (16). Oxford University Press: 2184–5.