ZBI ZENTRUM FÜR BIOINFORMATIK

UNIVERSITÄT DES SAARLANDES

Computational Methods for Epigenome Analysis – RNAseq group 1

Midhuna Immaculate Joseph Maran

Siwen Chen

ZBI ZENTRUM FÜR BIOINFORMATIK

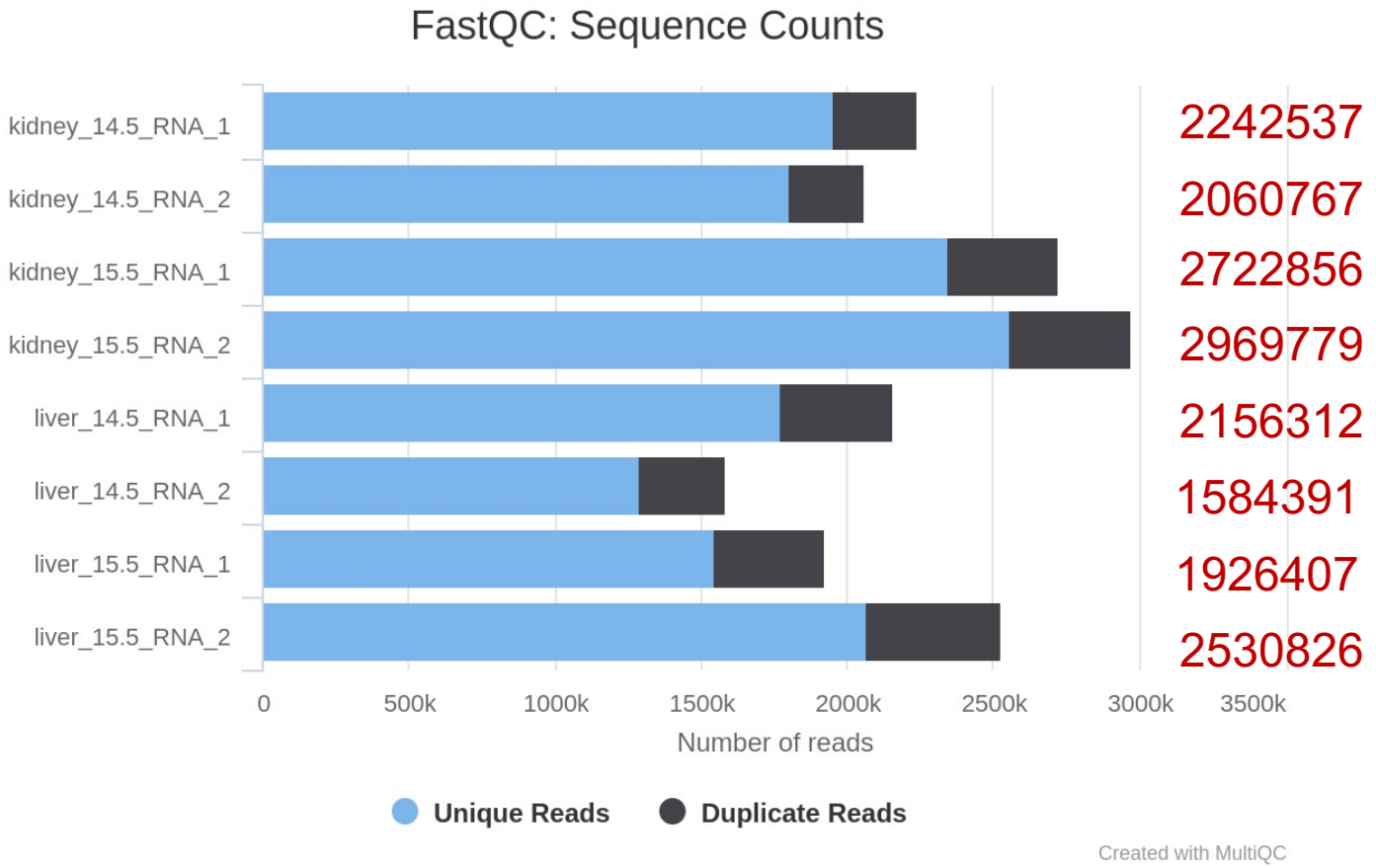
UNIVERSITÄT DES SAARLANDES

Overview

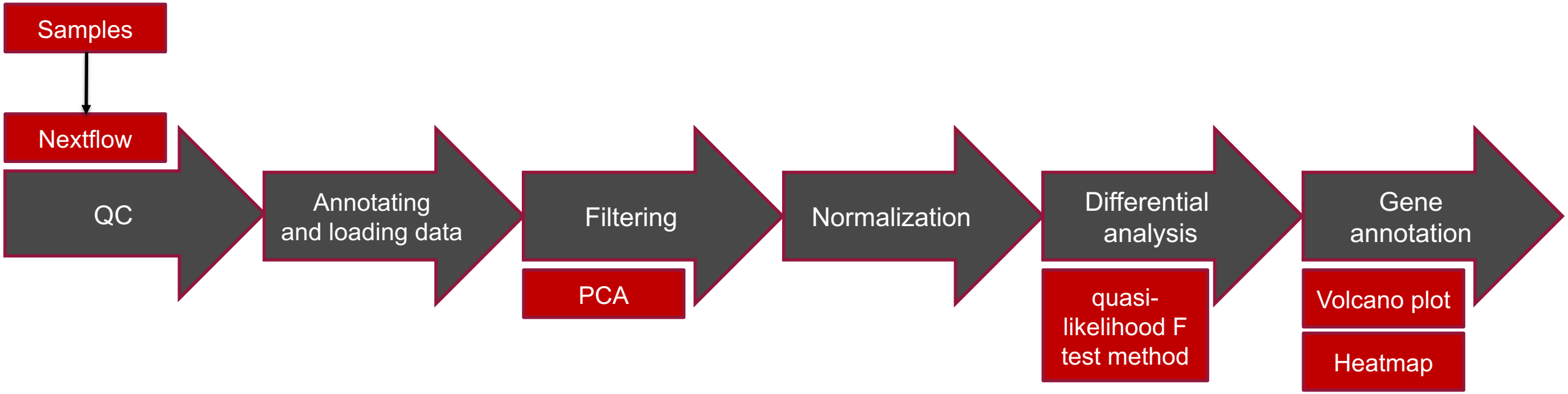
- Introduction
  - Datasets
  - Workflow
- Results
  - Quality Control
  - Differential Analysis
  - Exploratory Analysis & Integrative Data Exploration
  - Integrative Analysis
- Challenges and Limitations
- Conclusion

Introduction

- Dataset: 8 samples
  - 2 organ-specific samples: Liver and Kidney
  - 2 timepoints: 14.5 and 15.5 hrs
  - 2 replicates



Workflow



R-libraries:

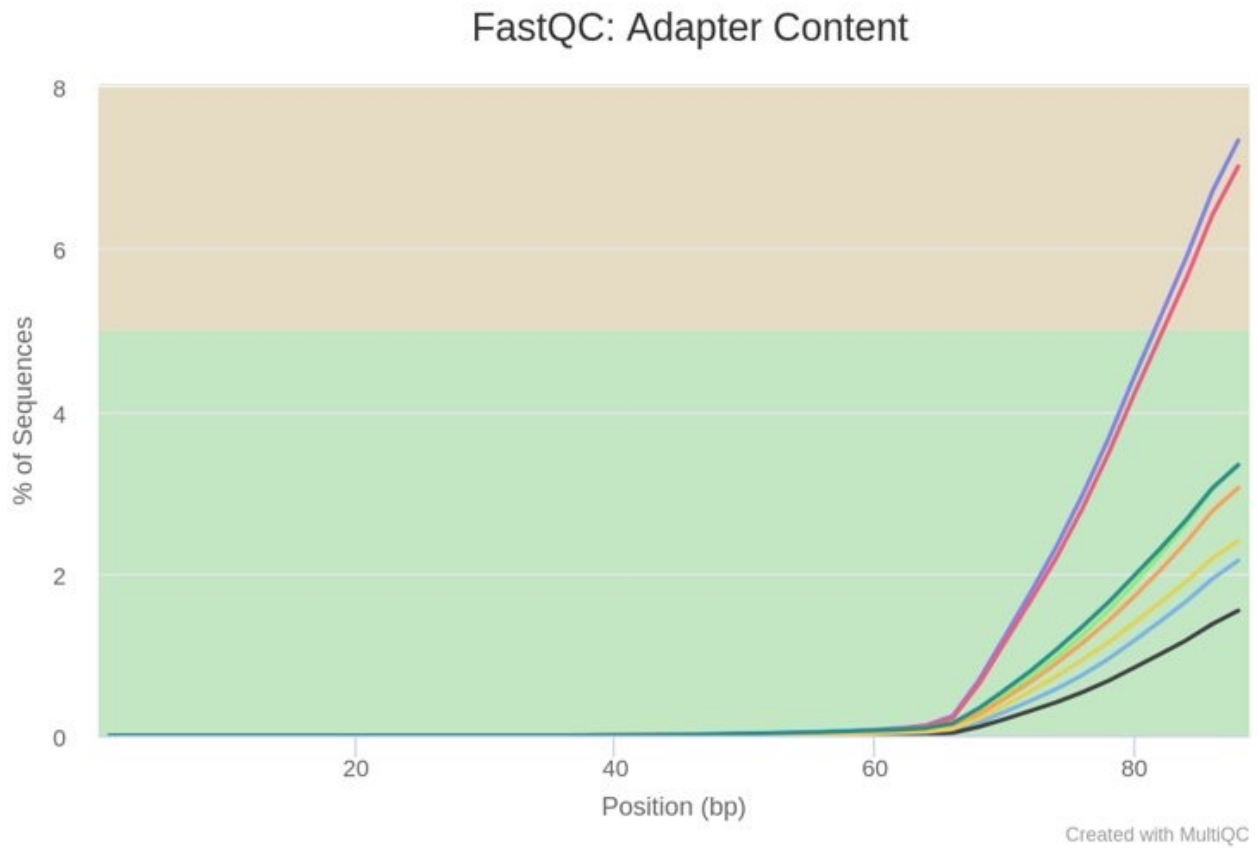
- tximport
- edgeR
- biomaRt

Quality Control: MultiQC

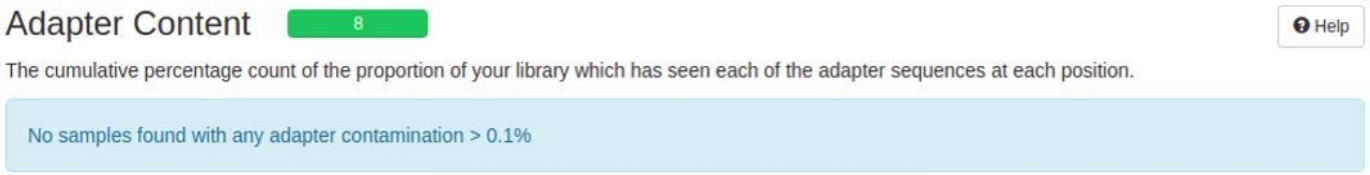
Raw Data

Trimmed Data

Quality Control: MultiQC



Raw Data

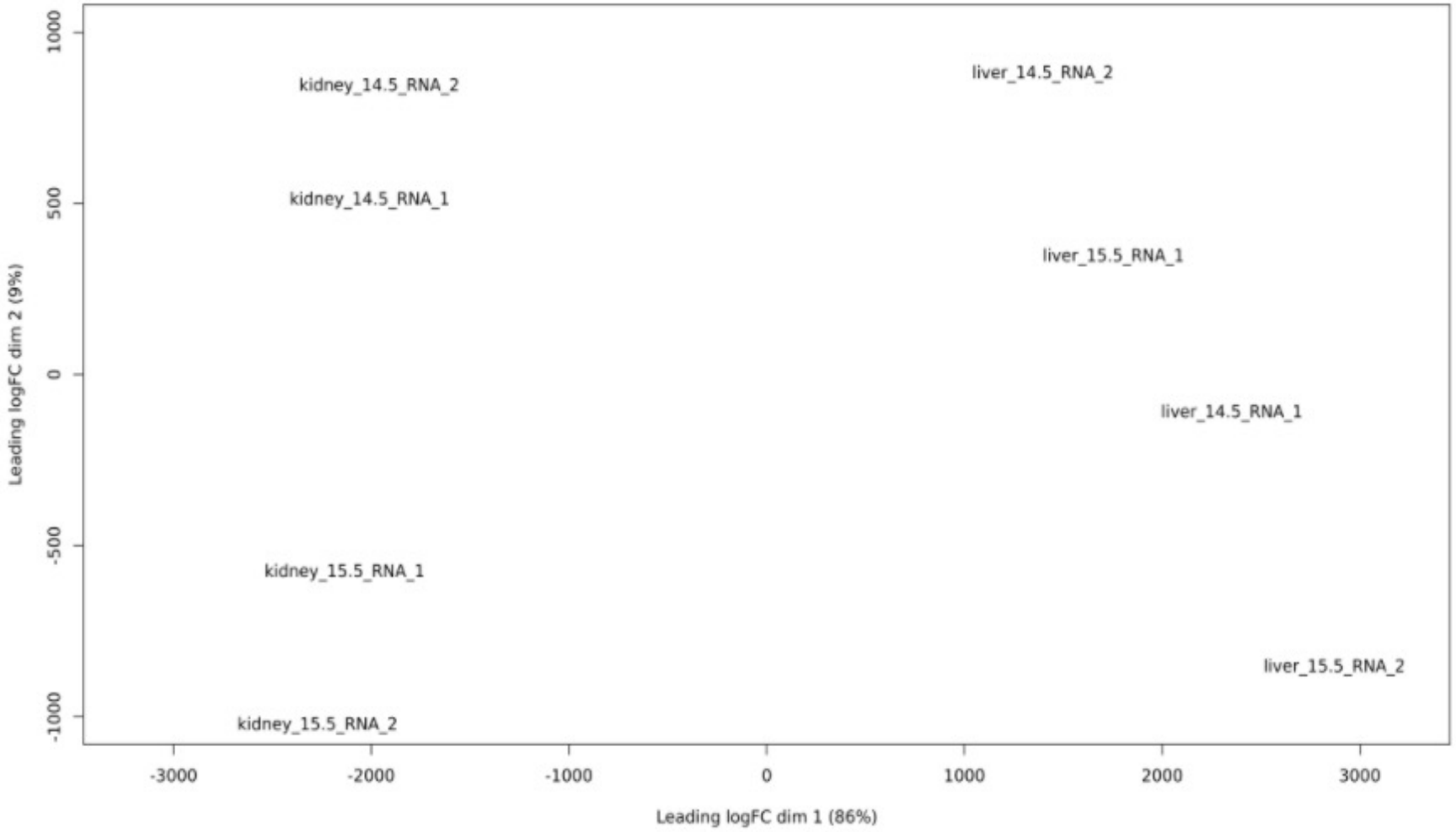


Trimmed Data

Differential Gene Expression Analysis

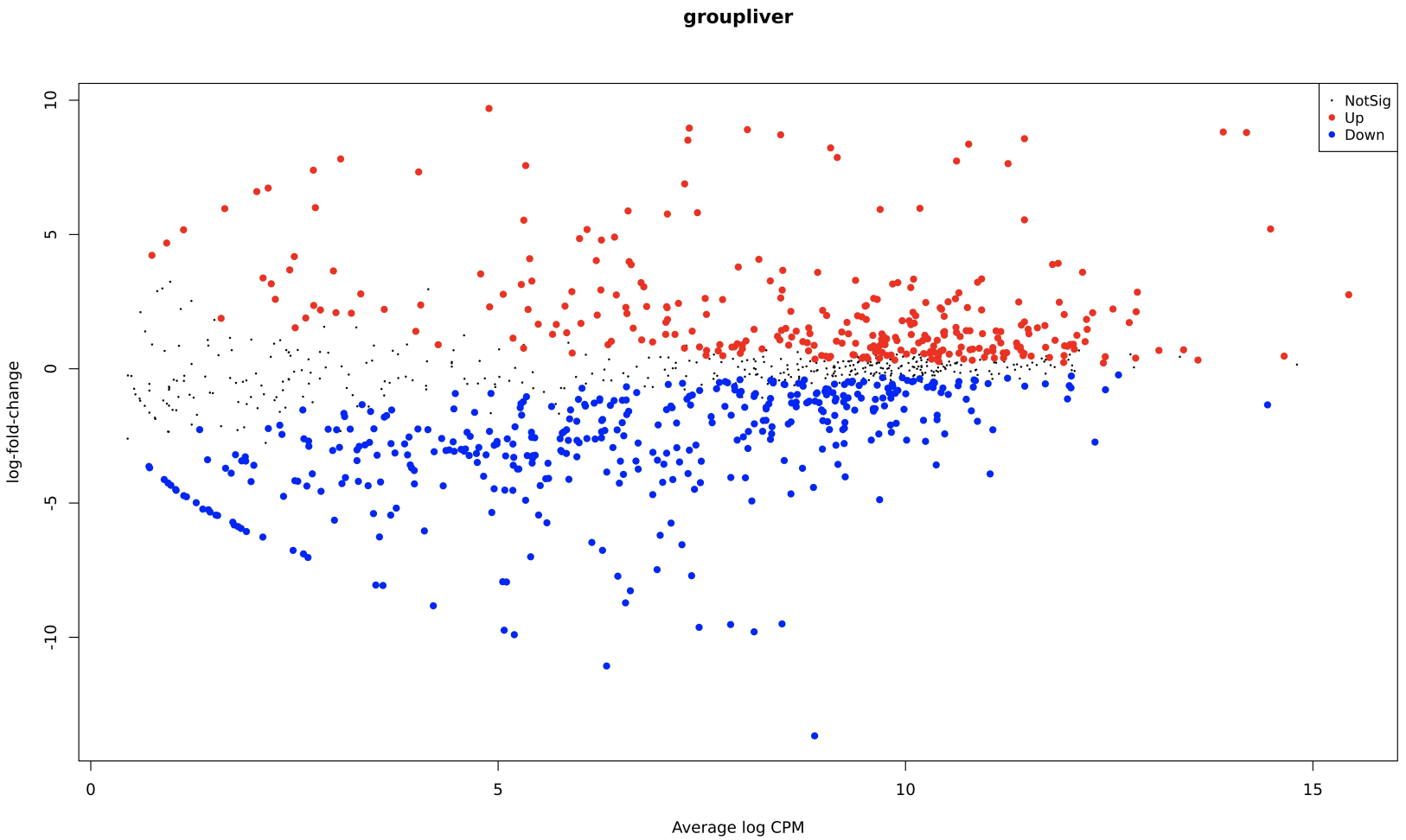
- Identify differential gene expression between two sample groups (Kidney vs. Liver)
- R package – edgeR
  - ✓ Input – gene count matrix
    - Genes, isoforms, sampleID, timepoint, replicate, group
  - ✓ Output – table of log fold changes and FDR for each gene (MA plot)
    - Log(2) fold changes – up/down-regulated gene expression
    - FDR – false discovery rate
- Previously,
  - ✓ Design matrix, DGEList object (matrix)

PCA – Data Exploration



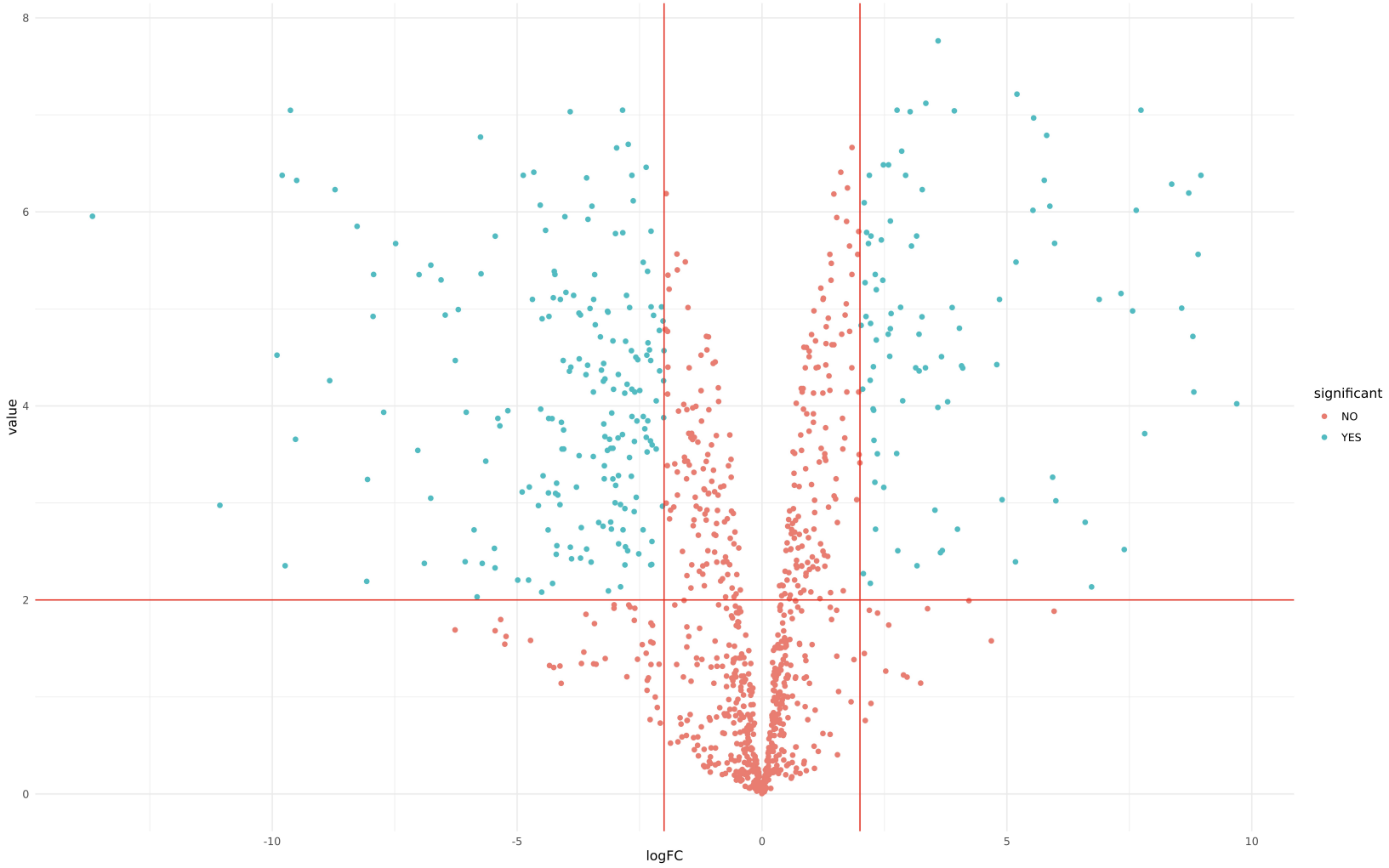
- Filtering (CPM)
- Each sample - 1392 genes
- Filter non-expressed and low abundance genes, 1127 genes
  - CPM > 0,5
  - Sum(CPM) >= 3
- PCA analysis
  - PC1 explains 86% variance
  - PC2 explains 9% variance

Quasi-likelihood F test method



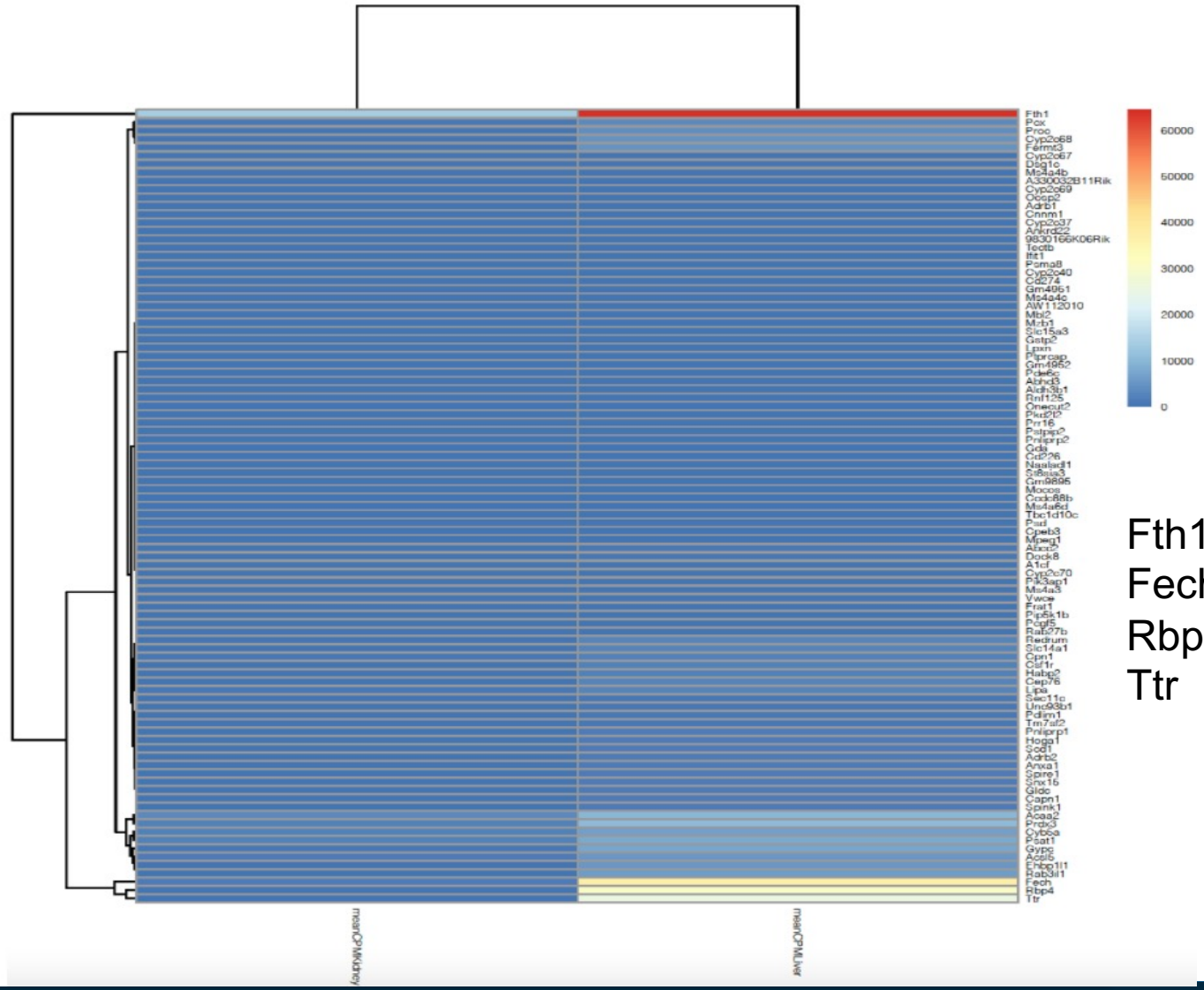
- Normalization – TMM
- Differential analysis
  - Quasi-likelihood F test method
  - Estimating the dispersion
- BiomaRt
  - Ensemble‘s ID and gene coordinates

Differential-Expression Genes – Volcano plot



- Differential Expression Genes list –
  - Chromosome
  - Start and end position
  - GeneID
  - **-log10FDR** and LogFC
  - MeanCPMKidney and MeanCPMLiver
  - EnsembleID
- Significant genes –
  - **FDR <= 0.01**
  - |logFC| >= 2

Differential-Expression Genes - Heatmap



- Differential Expression Genes list –
  - Chromosome
  - Start and end position
  - GeneID
  - -log10FDR and LogFC
  - MeanCPMKidney and MeanCPMLiver
  - EnsembleID
- Significant genes –
  - FDR <= 0.01
  - |logFC| >= 2

UNIVERSITÄT DES SAARLANDES

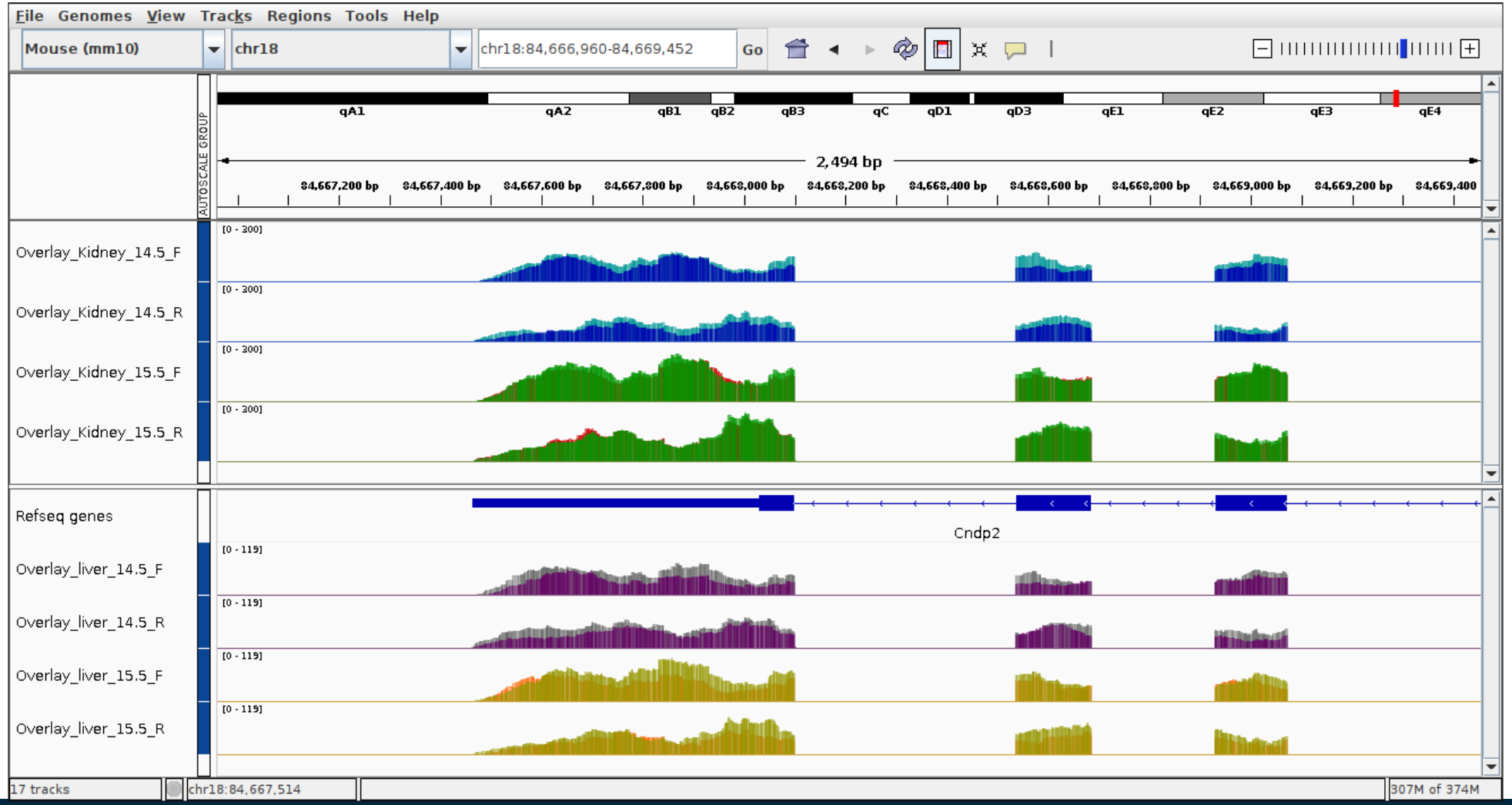
Integrated Analysis

Exploratory Analysis

Gene CNDP2:

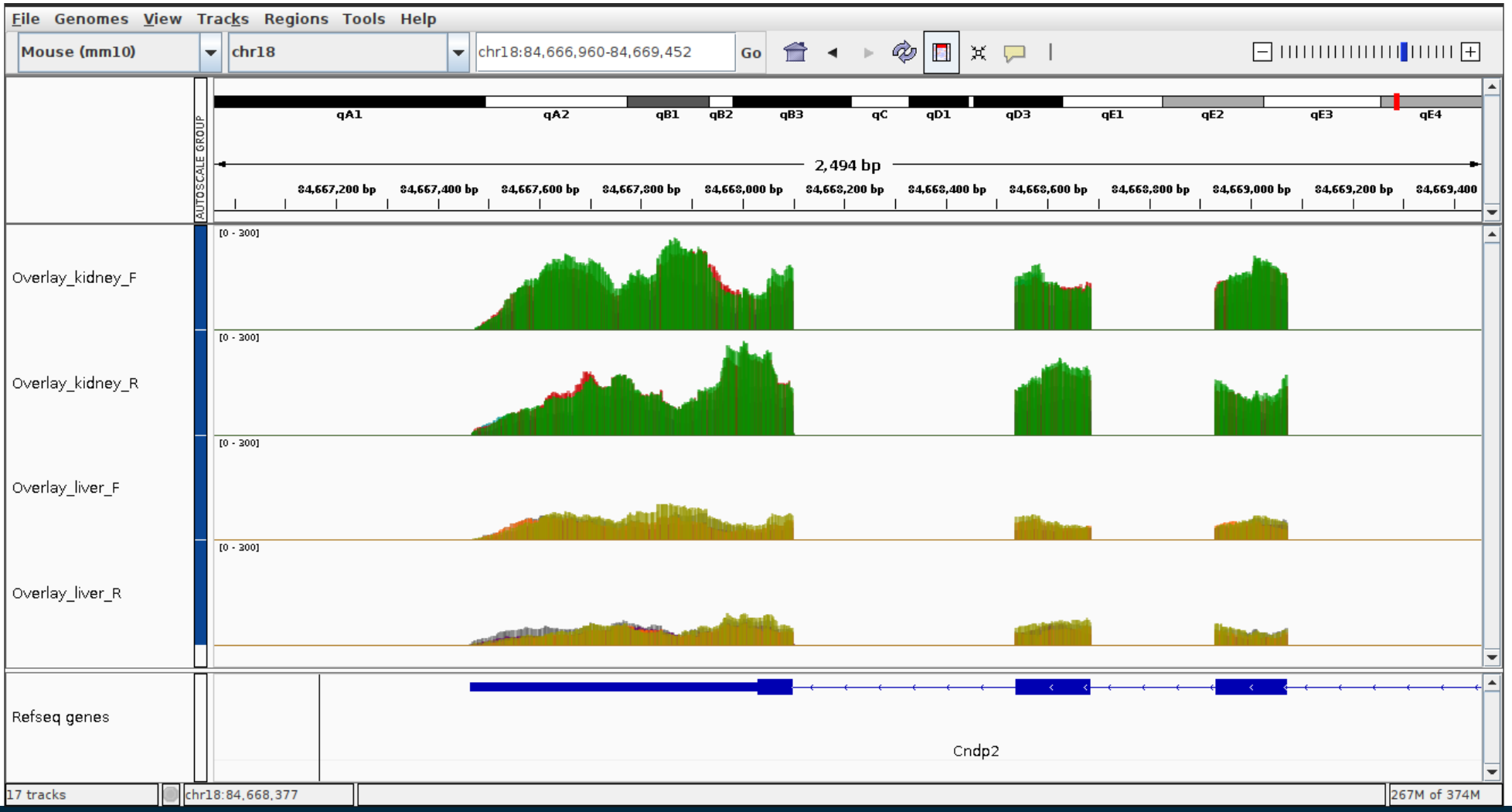
- Cytosolic non-specific dipeptidase
- Hydrolyzes dipeptides

Exploratory Analysis



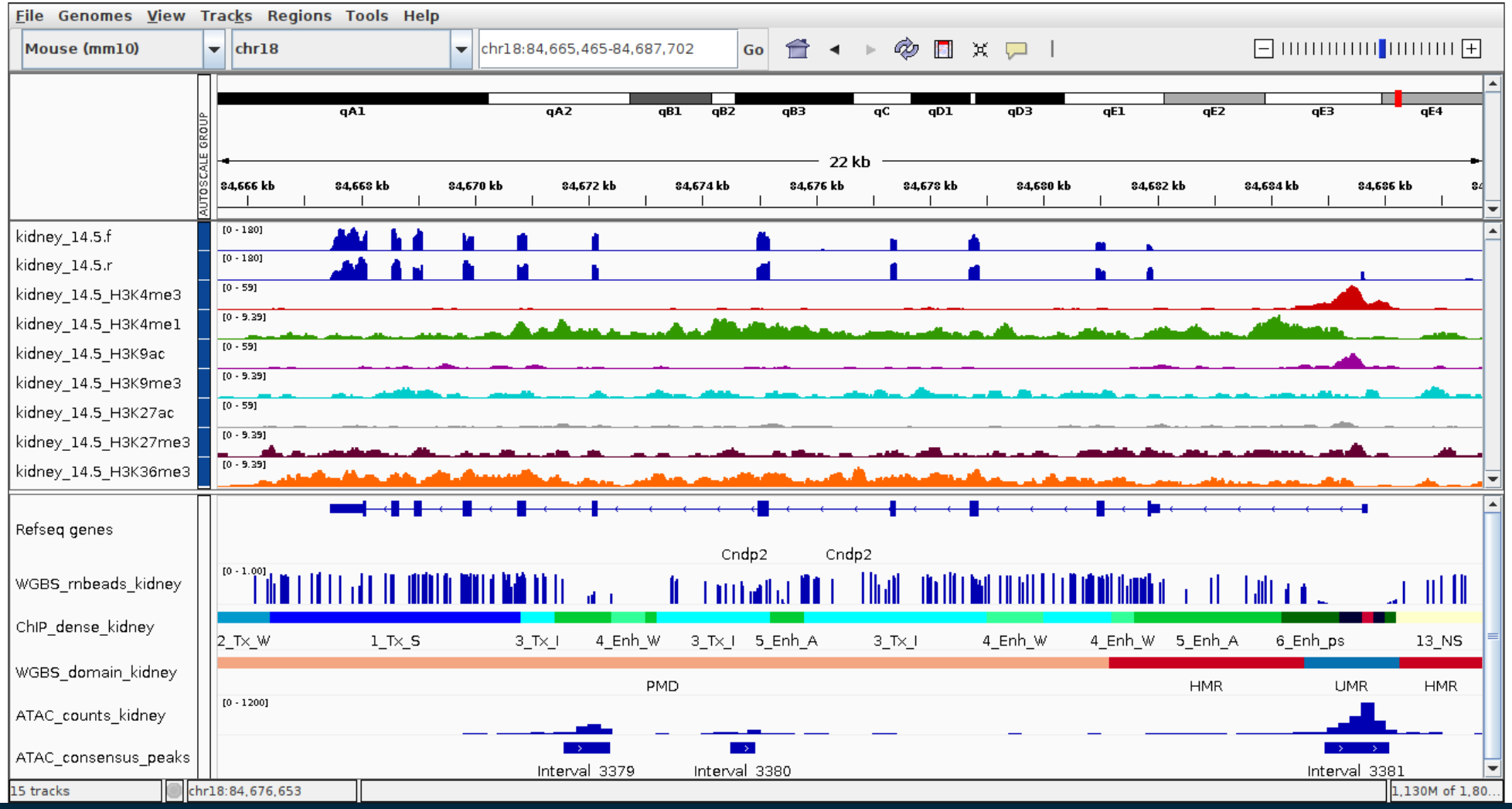
- Overlap: replicates

Exploratory Analysis



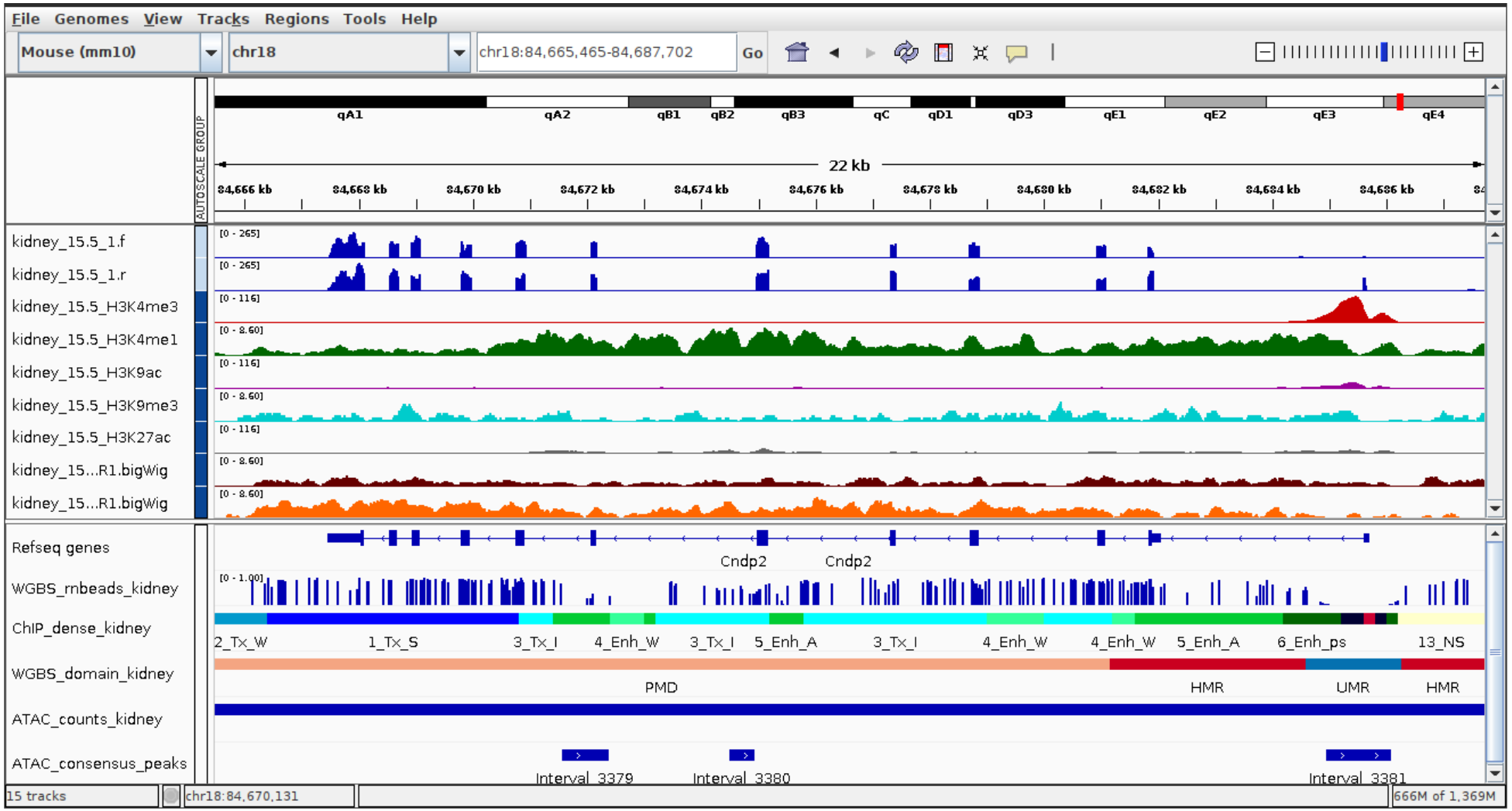
- Organ-specific overlay

Integrative Data Exploration



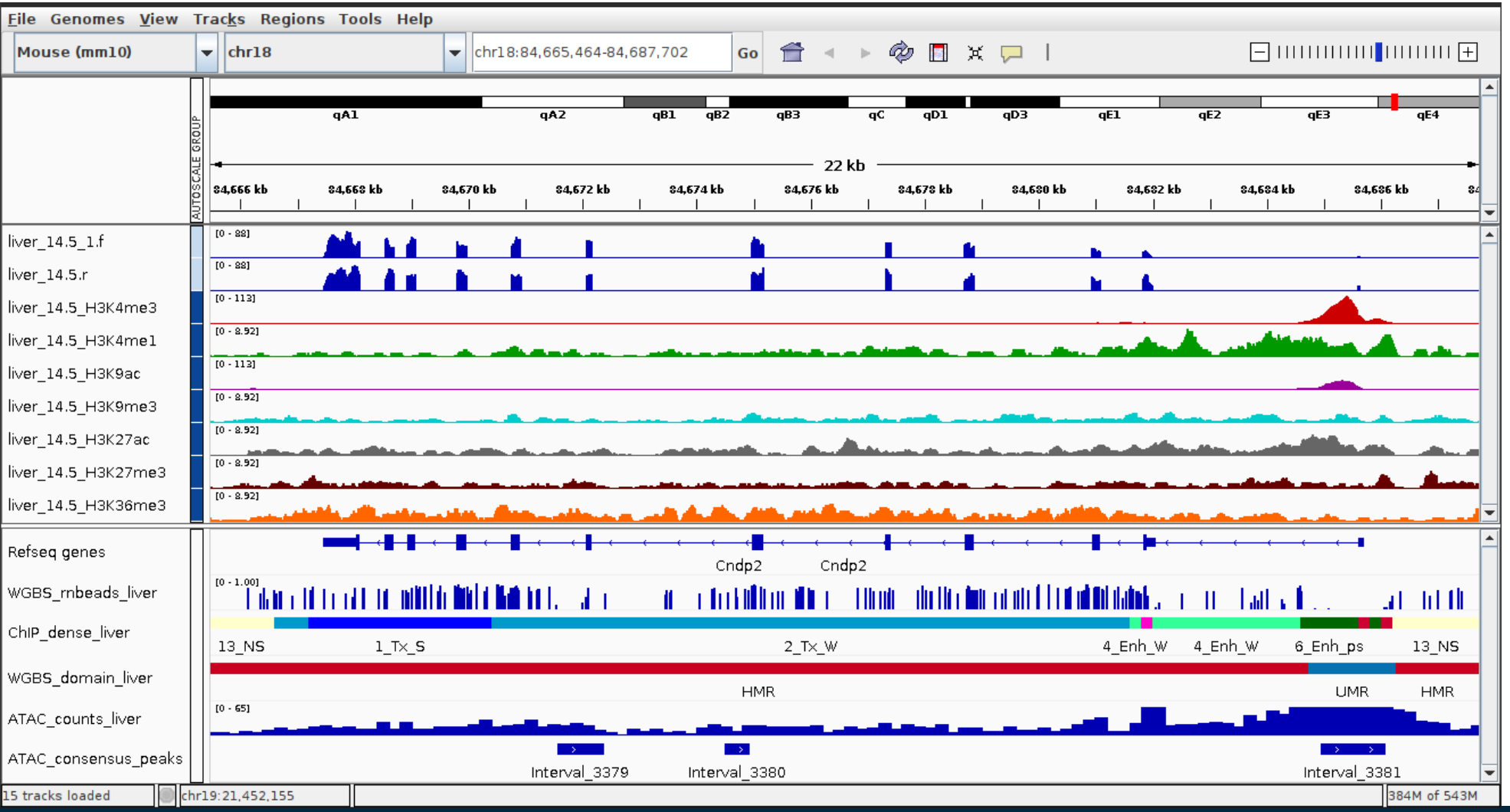
- Kidney_14.5 hr (cndp2)
- Narrow peaks:
  - H3k4me3
  - H3k9ac
- Broad peaks:
  - H3k4me1
  - H3k36me3

Integrative Data Exploration



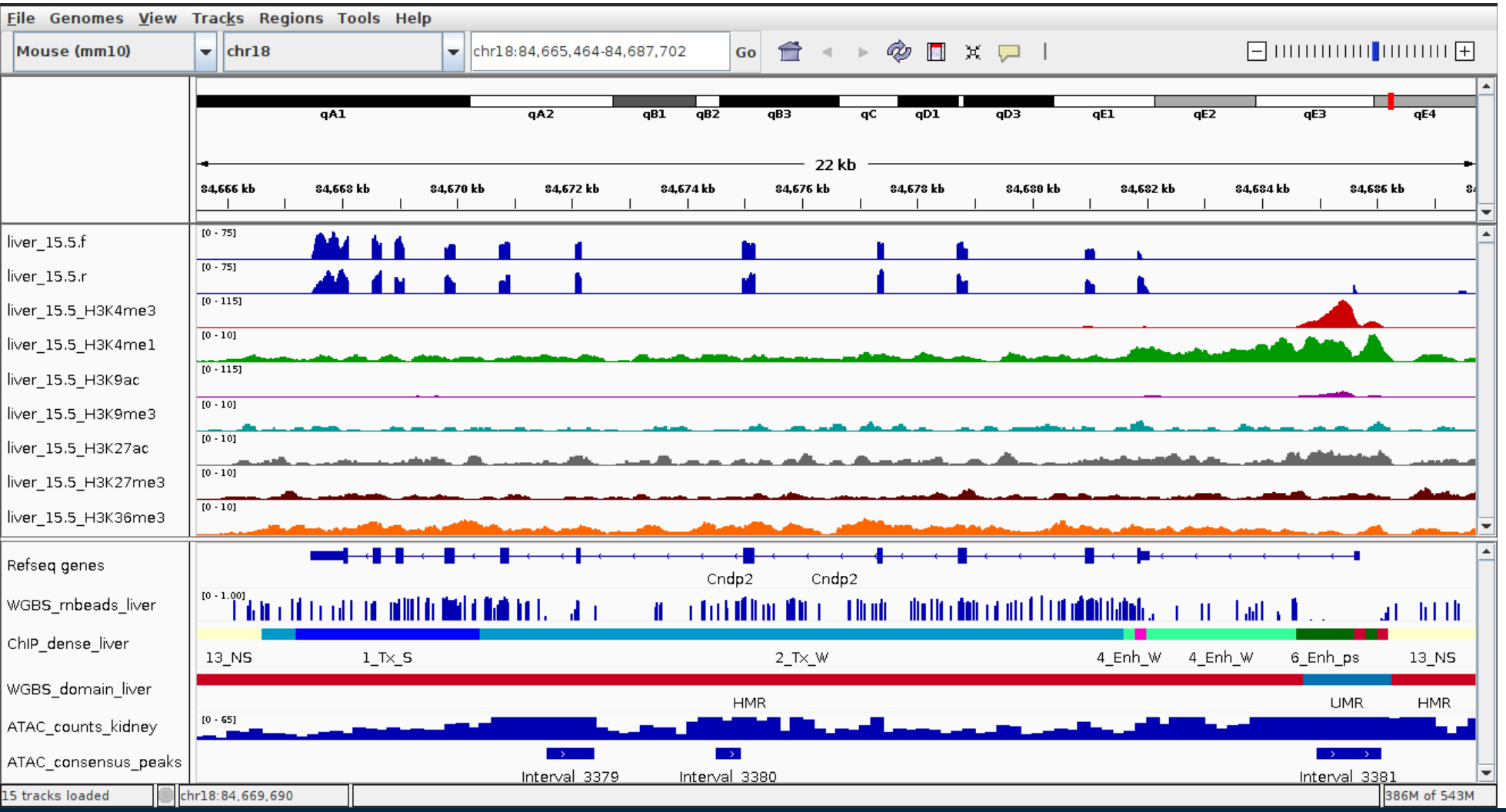
- Kidney_15.5 hr
- Narrow peaks:
  - H3k4me3
- Broad peaks:
  - H3k4me1
  - H3k36me3

Integrative Data Exploration



- liver_14.5 hr
- Narrow peak:
  - H3k4me3
- Broad peaks:
  - H3k4me1
  - H3k27ac
  - H3k36me3

Integrative Data Exploration



- Liver_15.5 hr
- Narrow peak:
  - H3k4me3
- Broad peaks:
  - H3k4me1
  - H3k36me3

Integrative Data Exploration

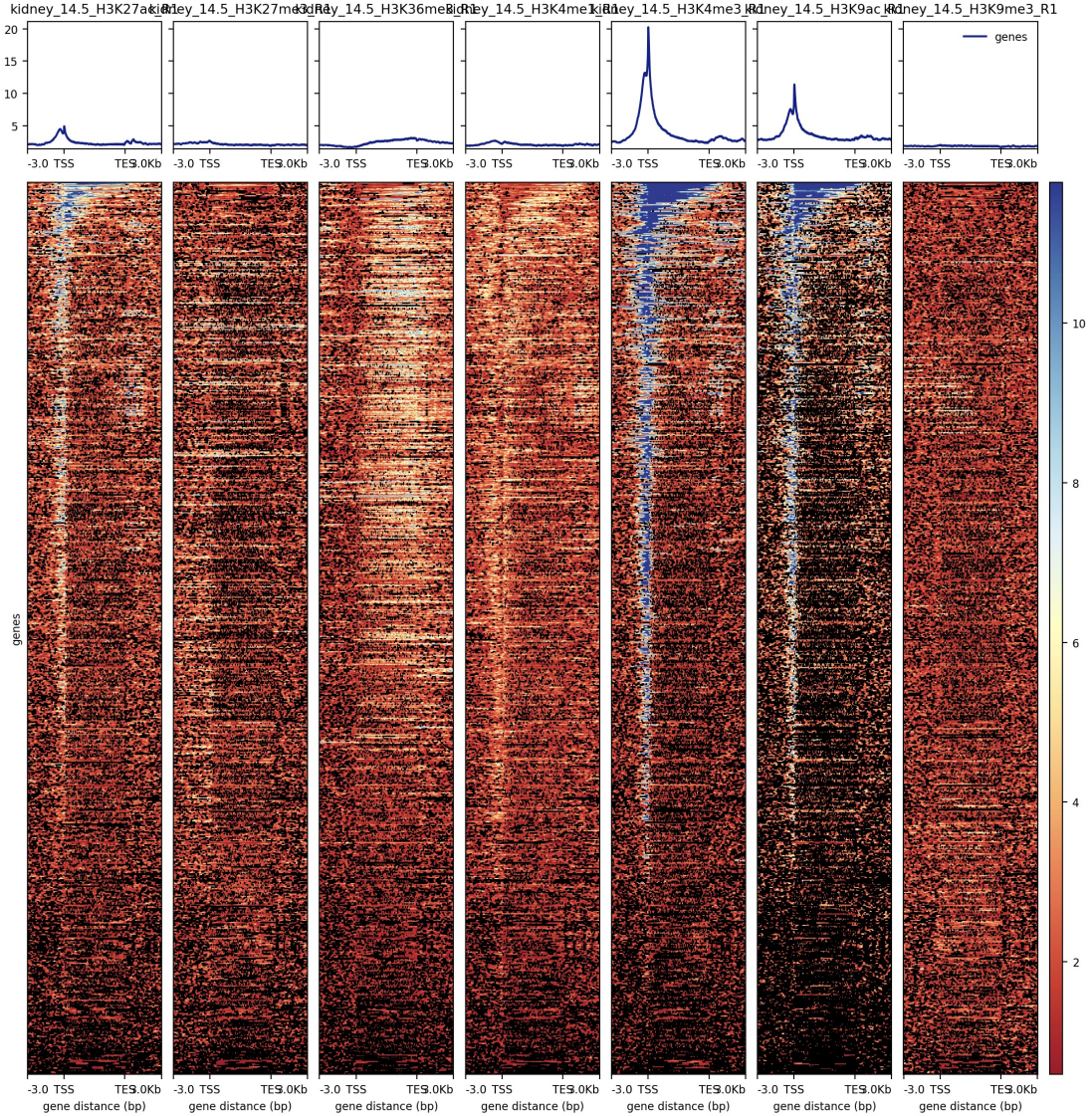
- DEGene: GDA (Guanine Deaminase)
- Kidney_14.5 hr
- Narrow peaks:
  - H3k4me3
  - H3k9ac
- Broad peaks:
  - H3k4me1
  - H3k36me3
  - H3k27me3

Integrated Analysis

Chromatin states and gene expression –

- Each gene –
  - Chromosome
  - genomic coordinates
  - geneID
  - **TPM**
  - **strand**
- Gene expression and histone marks (ChIP seq)
  - 7 histone marks
    - H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9ac, H3K9me3

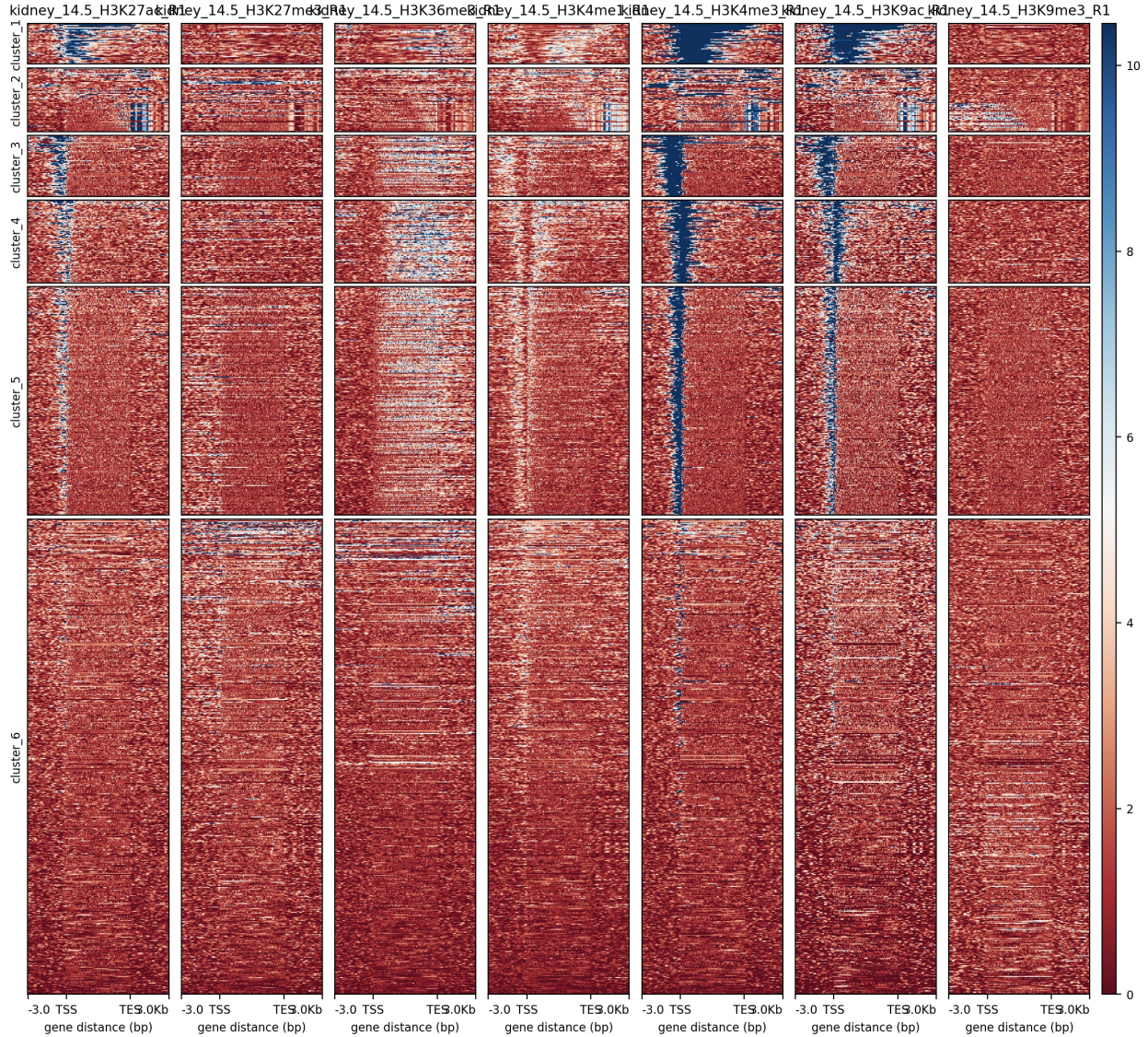
Integrated Analysis – Histone mark coverage



Chromatin states and gene expression

- kidney_14.5_RNA_1
- Enrichment/depletion
  - **H3K27ac**
  - H3K27me3
  - **H3K36me3**
  - H3K4me1
  - **H3K4me3**
  - **H3K9ac**
  - H3K9me3
- Gene order
  - Descent enrichment value

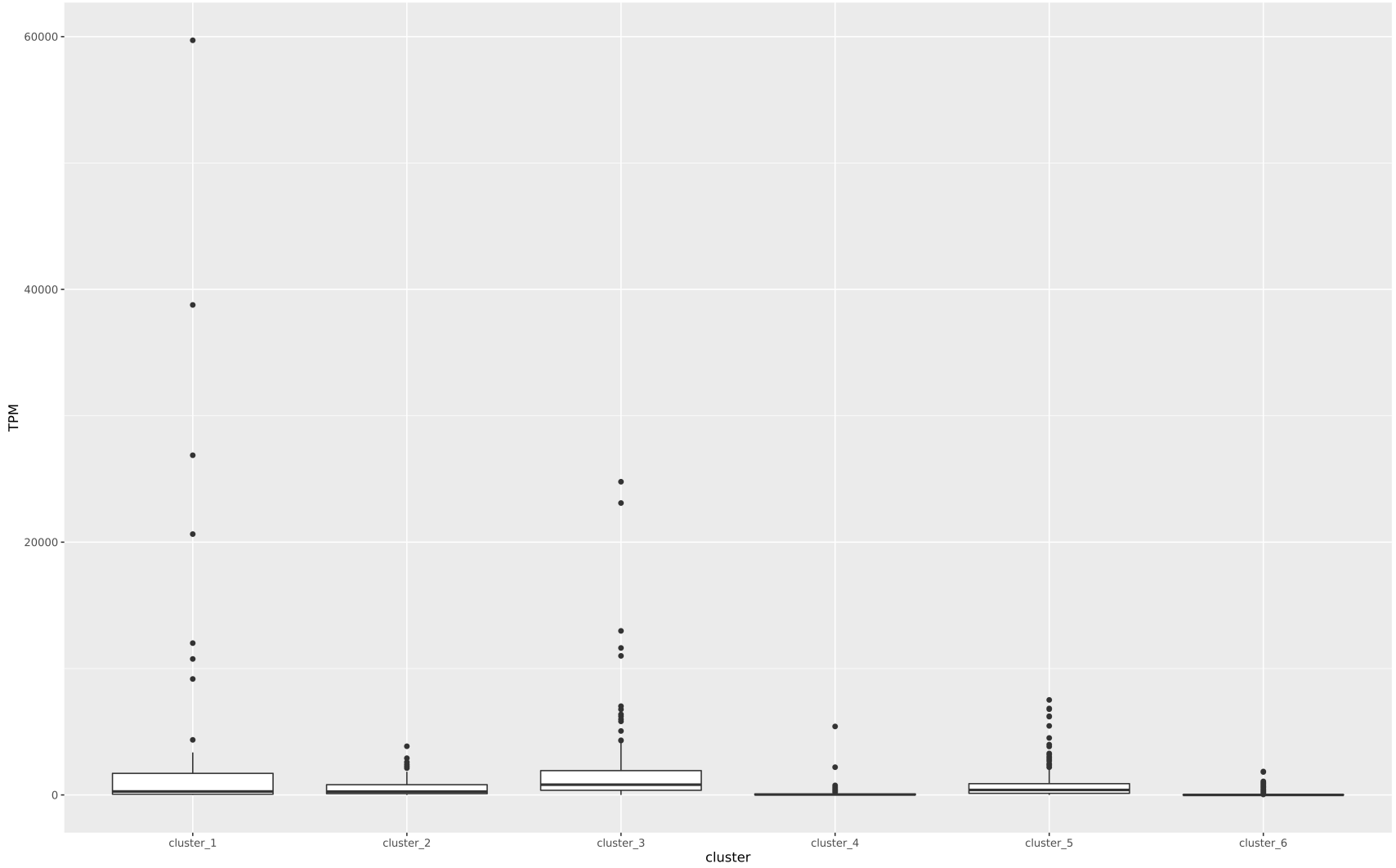
Integrated Analysis – kmeans clustering



Chromatin states and gene expression

- Kmeans – 6 clusters
  - cluster 1 – H3K27ac
- Enrichment/depletion
  - **H3K27ac**
  - H3K27me3
  - **H3K36me3**
  - H3K4me1
  - **H3K4me3**
  - **H3K9ac**
  - H3K9me3

Integrated Analysis - boxplot



Chromatin states and gene expression

- Enrichment/depletion
  - H3K27ac
  - H3K27me3
  - H3K36me3
  - H3K4me1
  - H3K4me3
  - H3K9ac
  - H3K9me3
- TPM
  - Highest – c3
  - Lowest – c4
  - Prominant - H3K4me3

Challenges and Limitations

RNA sequencing -

- Highly sensitive on low-abundance transcripts
- Higher dynamic range of expression levels
- Lower technical variation

However -

- Accurate sequence annotation and data interpretation required
- Gives only transcript abundance information
- Preference on steady-state gene expression level (vs. GRO-seq)
- Single cell RNA-seq analysis -
  - Massive cell-to-cell variability
  - Zero inflation ...

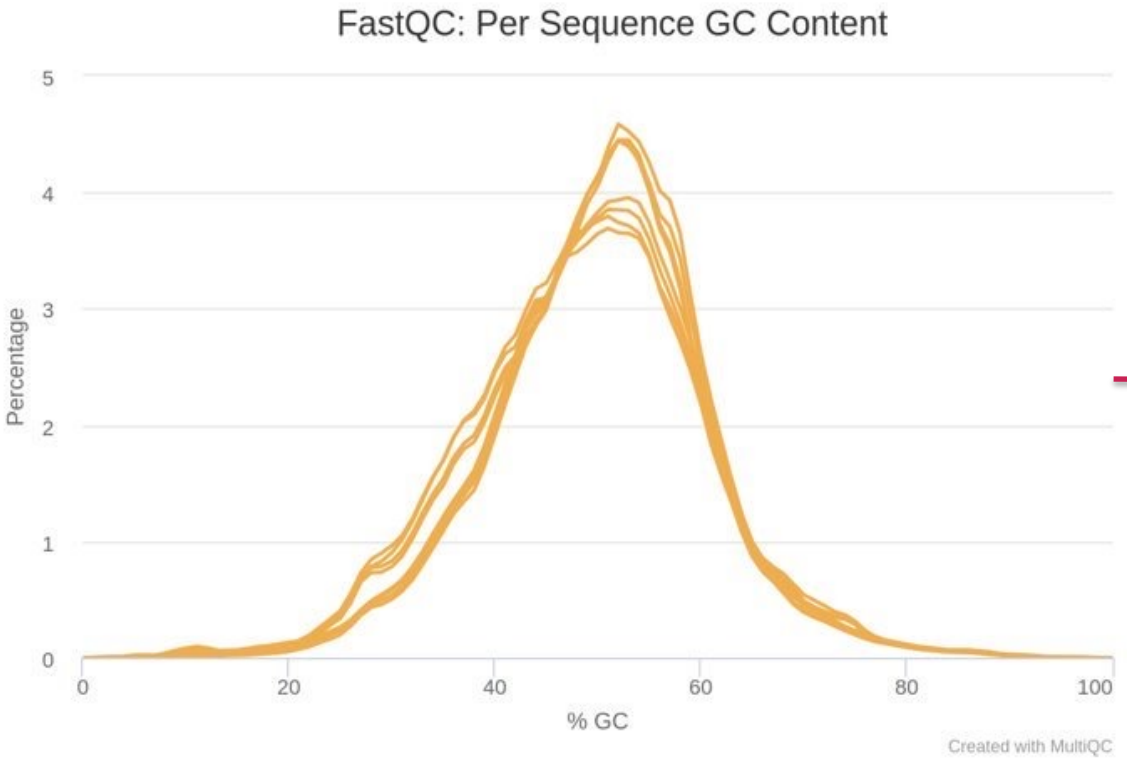
Conclusion

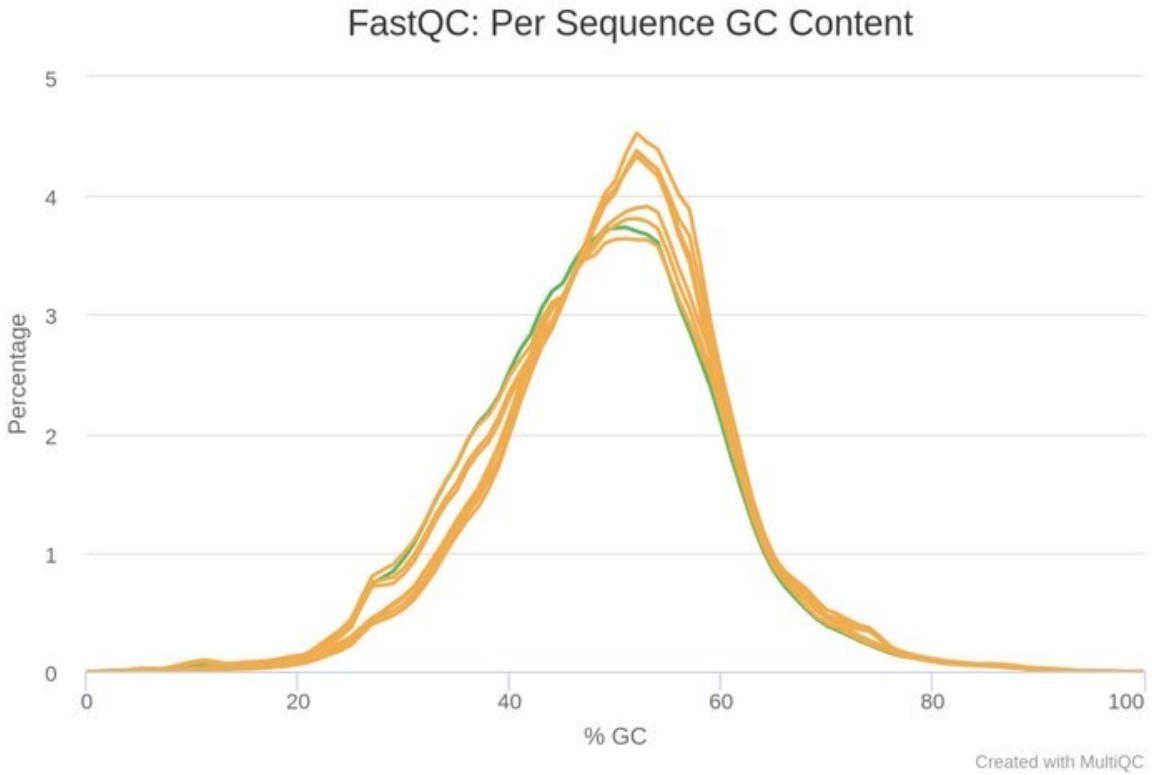
- 8 Samples
- Gene CNDP2:
  - More expression in 15.5 hr in both organ samples
  - Highly expressed in Kidney than liver
    - Possible explanation: HMD in gene body of liver samples
  - DEGene: GDA
    - H3K27me3: Possible polycomb repression

UNIVERSITÄT
DES
SAARLANDES

Thank you for your attention.

Quality Control: MultiQC





Raw Data

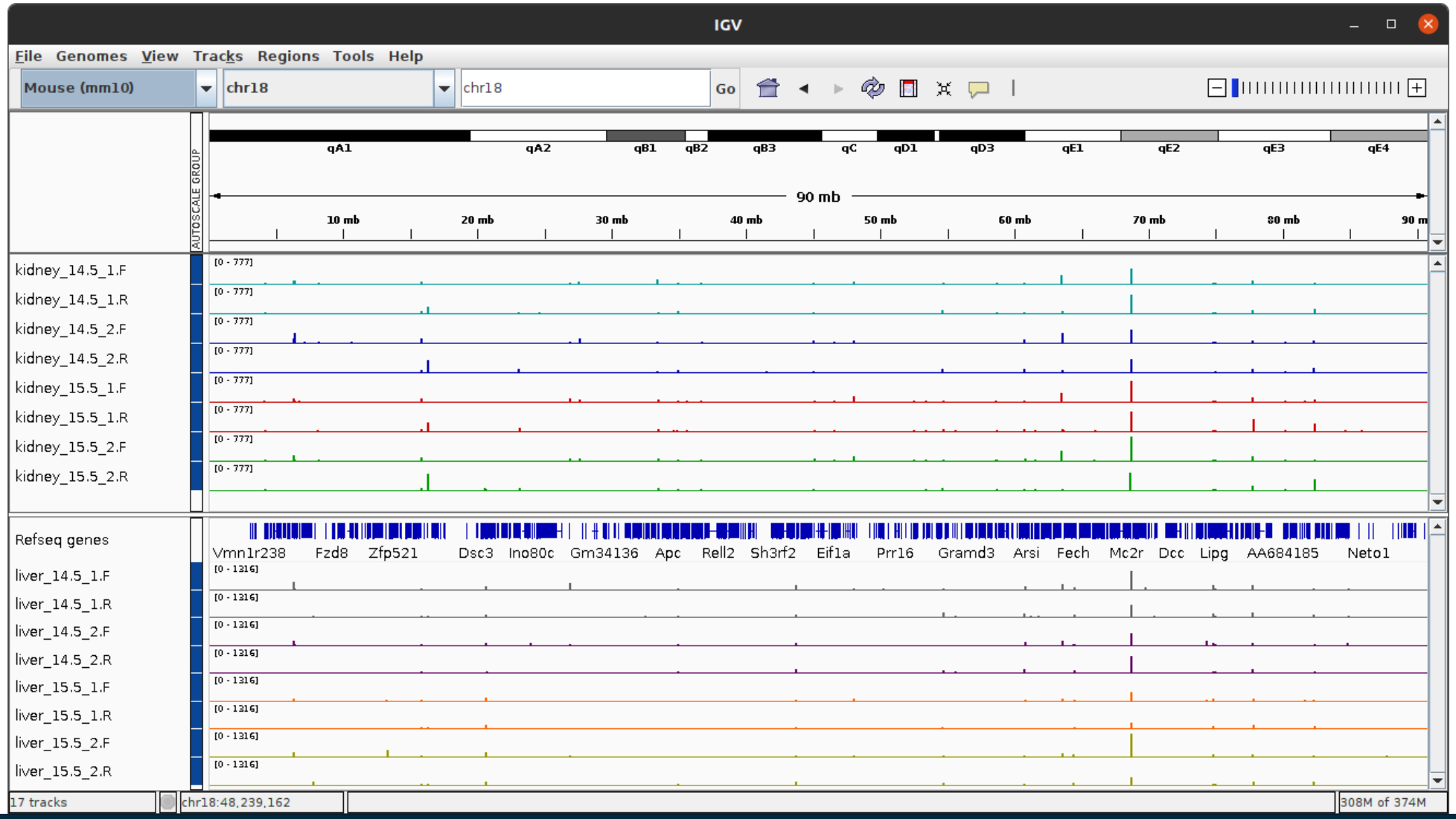
Trimmed Data

Quality Control: MultiQC

Raw Data

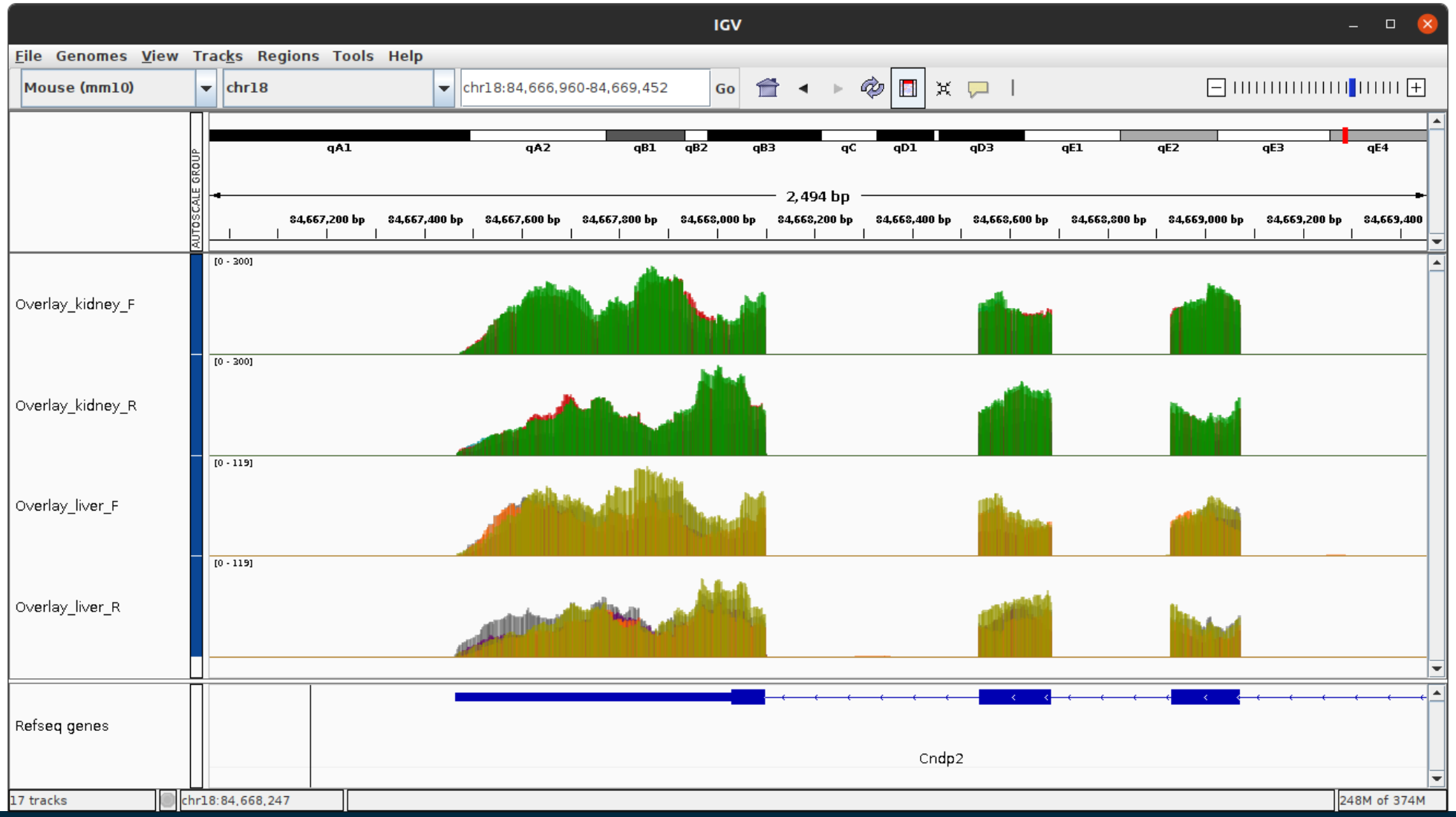
Trimmed Data

Exploratory Analysis



- Chromosome 18

Exploratory Analysis



- Timepoint overlap