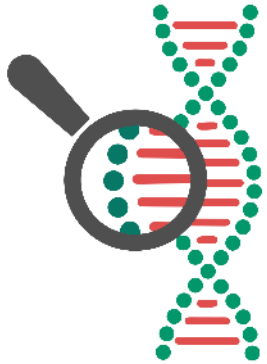
THETA
Statistical Genetics Group

Magda Mielczarek, PhD
THETA Biostatistics Group
Wroclaw University of Environmental and Life Sciences
magda.mielczarek@upwr.edu.pl

→ Analysis and interpretation of next-generation sequencing data



Projects in the field of:

- **Genomics**
  - Exploring genomic mutations underlying diseases with composite phenotypes (based on the **human** trio)
    → various diseases with overlapping symptoms consisting of ***endocrinopathy*** (thyroid, pituitary, and adrenal glands disorders) humoral ***immunity deficiency*** food, and inhalant ***allergies*** ***metabolism disorders, microsomia***)
  - DNA sequence features underlying large-scale duplications and deletions in **human**
  - Genetic background of clinical mastitis in Holstein-Friesian **cattle**
  - Genomic variability in species such as **cattle**, **bison, wolf, pigs**
  - An effect of large-scale deletions and duplications on transcript expression (multiomics).
  - etc.

- **Transcriptomics**
  - DIfferential Gene Expression Studies in:
    - **honey bee** in different intensities of electric field
    - ***Arabidopsis thaliana*** mutants (Elo)
  - Long non-coding RNAs variability in **porcine** skeletal muscle
  - The effect of transcriptomic annotations in breast cancer differential gene expression study (**human**)
  - etc.

- **Other omics**

THETA
Statistical Genetics Group

BIOSTATISTICS GROUP
WROCLAW, POLAND

The effect of transcriptomic annotations in **breast cancer** differential gene expression study

M. Mielczarek, B. Czech, R. Stępień, J. Szyda

Wrocław University of Environmental and Life Sciences

Objectives

- **Motivation**
  - In NGS data analysis, including differential gene expression (DGE) study, the most recent reference genome/transcriptome (**GRCh38**) is regarded as the annotation standard
  - However, **GRCh37** is also considered relevant

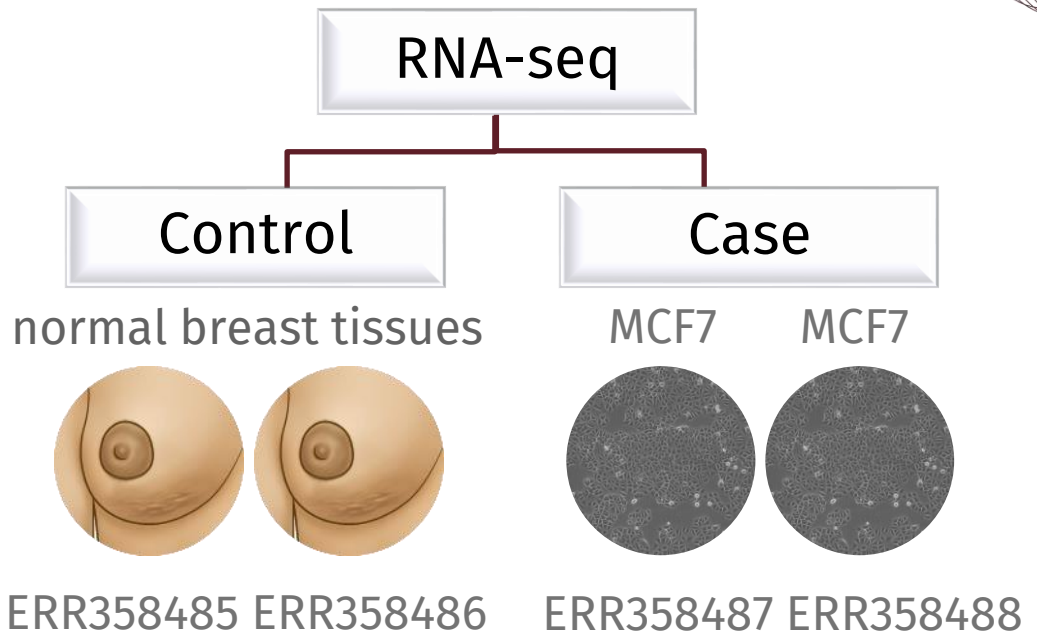
- **Goals**
  - Investigation of the transcriptome annotation version effect on:
    - DGE of all genes
    - genes related to survival prognosis in breast cancer

Material

- **RNA-seq:**
  - MCF7 breast cancer samples and normal breast tissues
  - NCBI BioProject ID: PRJEB4829
  - Illumina HiSeq 2000, 2 x 100 bp
  - Library size: 73,974,766 - 97,983,949 reads

- All known transcripts (coding and non-coding) annotated to **GRCh37** and **GRCh38**

- **Naderi breast cancer prognosis** down- and upregulated **genes**

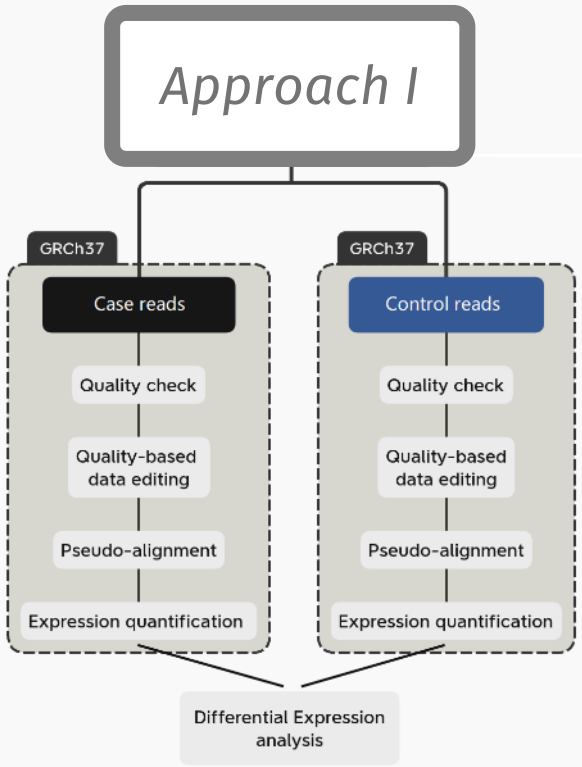


A gene-expression signature to predict survival in breast cancer across independent data sets

A Naderi, A E Teschendorff, N L Barbosa-Morais, S E Pinder, A R Green, D G Powe, J F R Robertson, Aparicio, I O Ellis, J D Brenton & C Caldas

Methods

- Four approaches to assess DGE



Fastqc

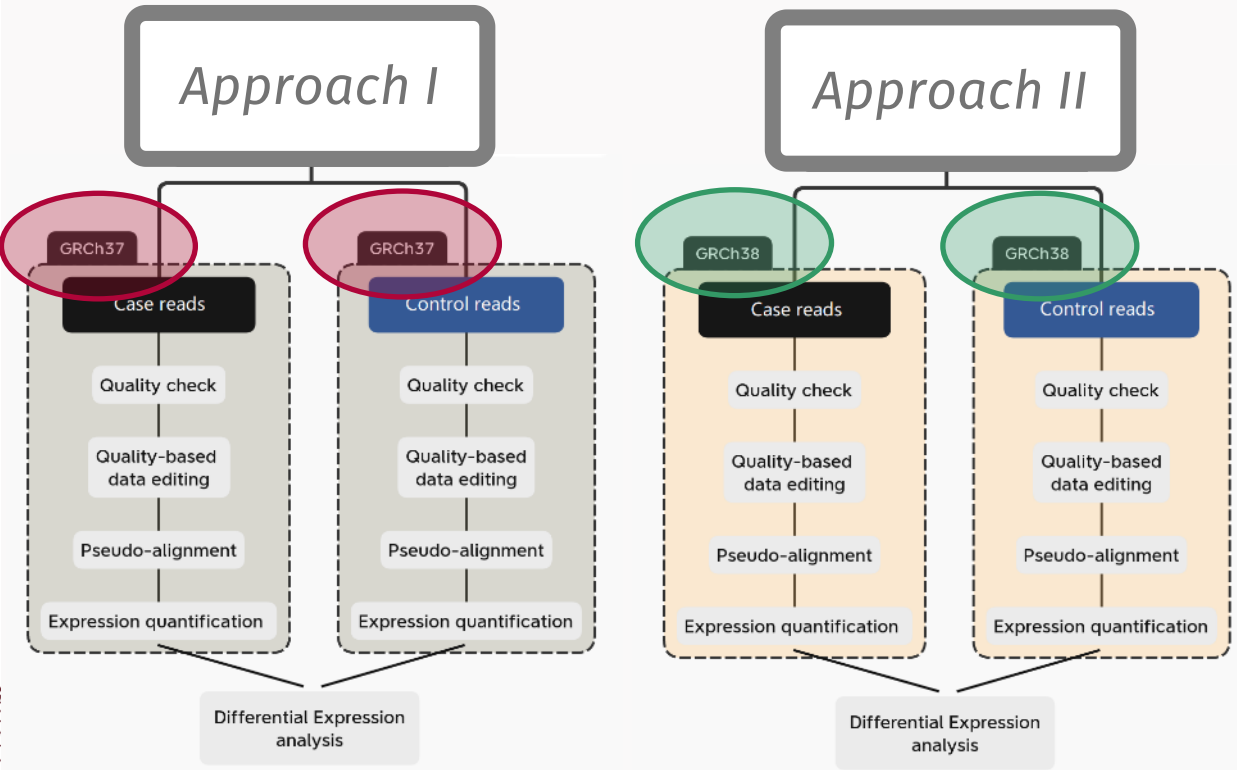
Trimmomatic

Kallisto

DESeq2

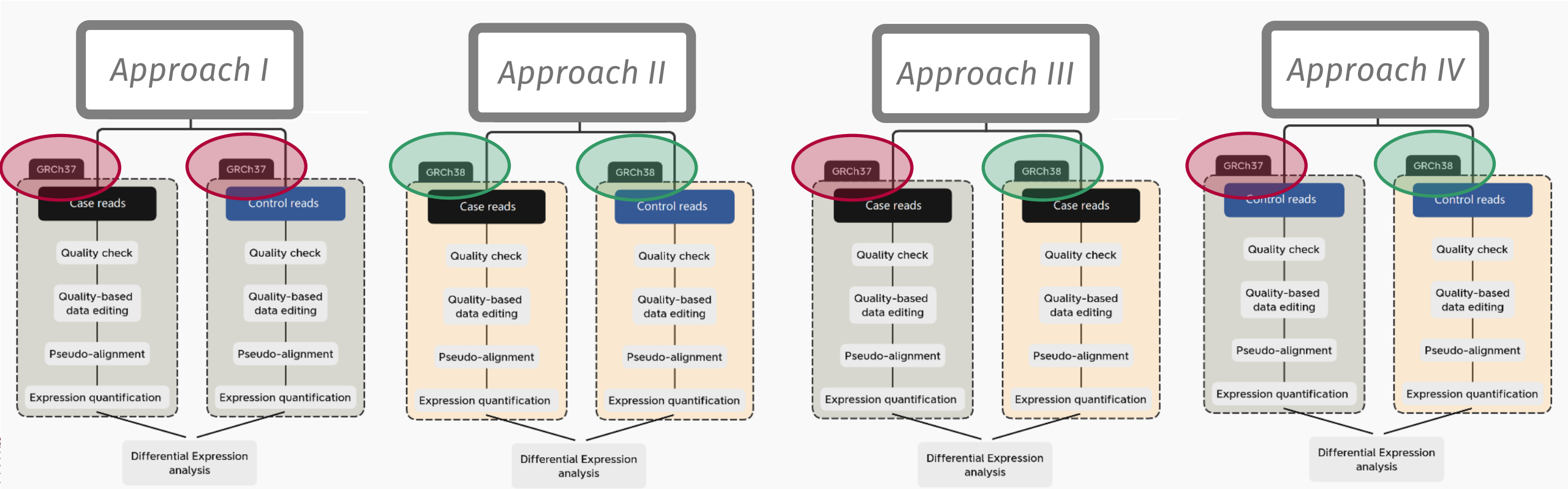
Methods

- Four approaches to assess DGE



Methods

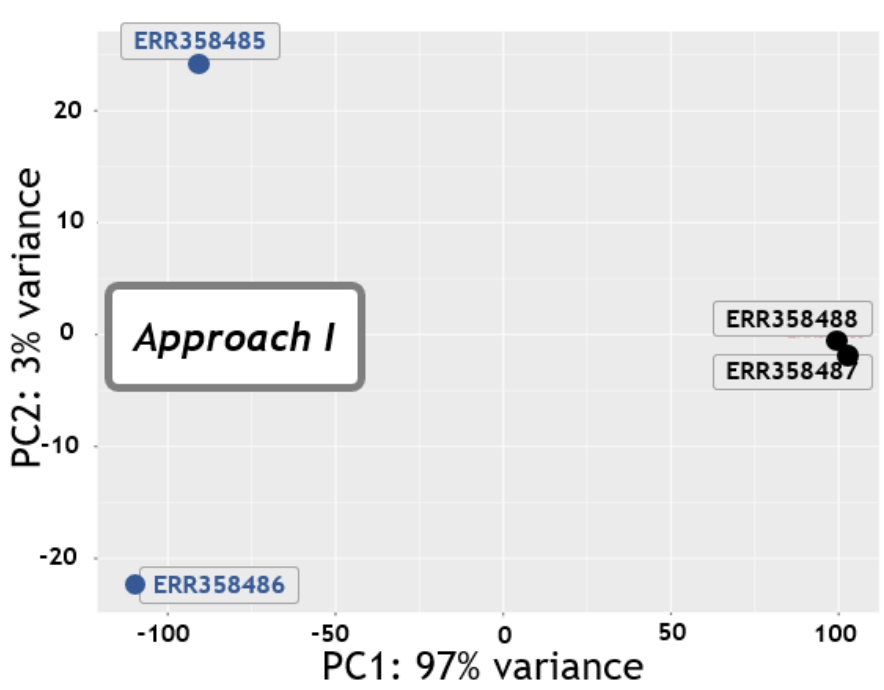
- Four approaches to assess DGE

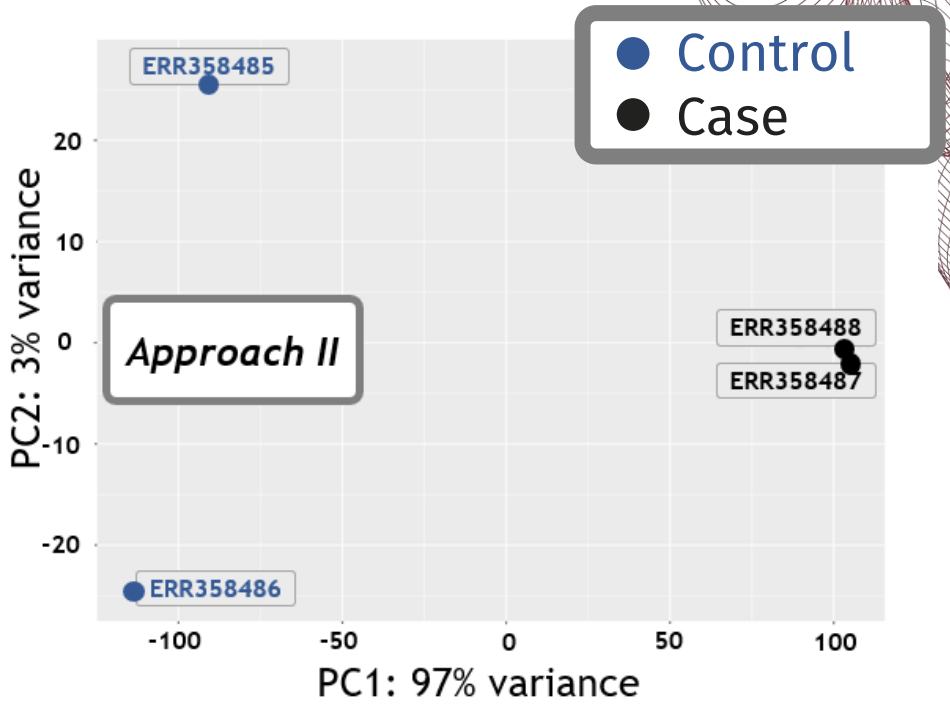


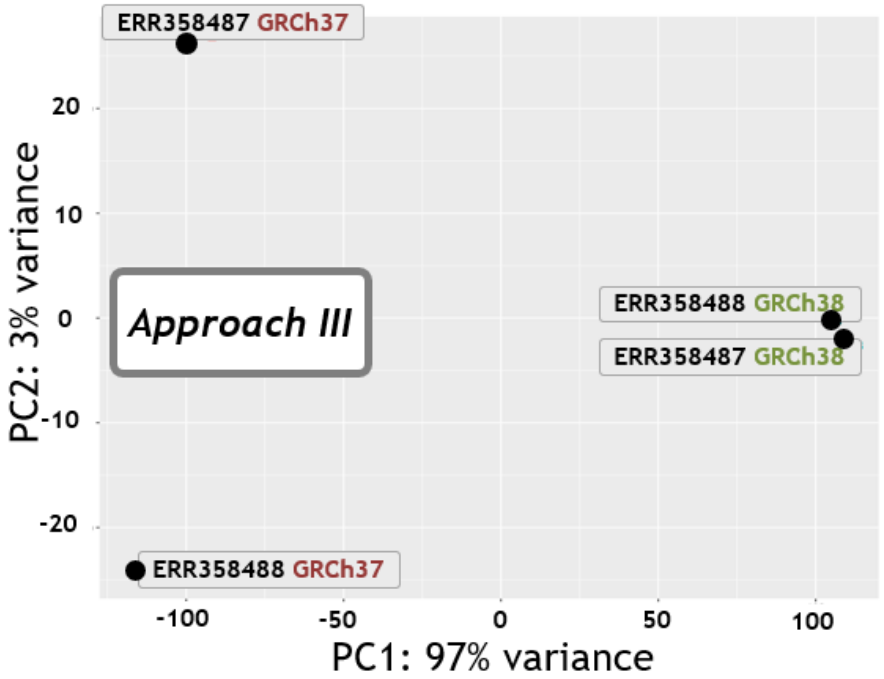
Methods

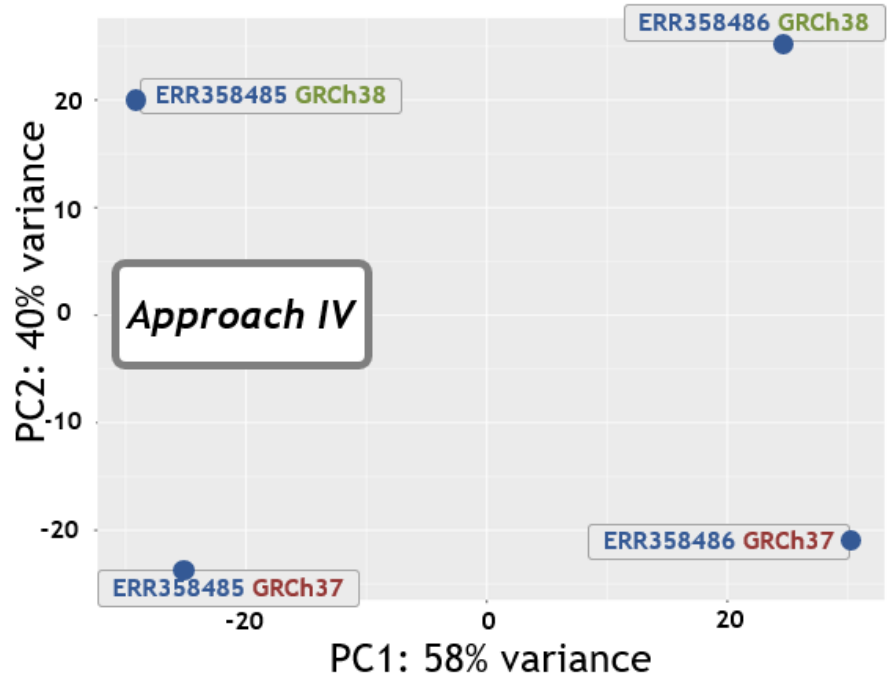
- One differential gene expression (DGE) and four approaches
- All genes:
  - PCA
- Naderi breast cancer prognosis down- and upregulated genes
  - T-test
- DEGs:
  - Top20
  - Gene Set Enrichment Analysis (GSEA)

Results: PCA

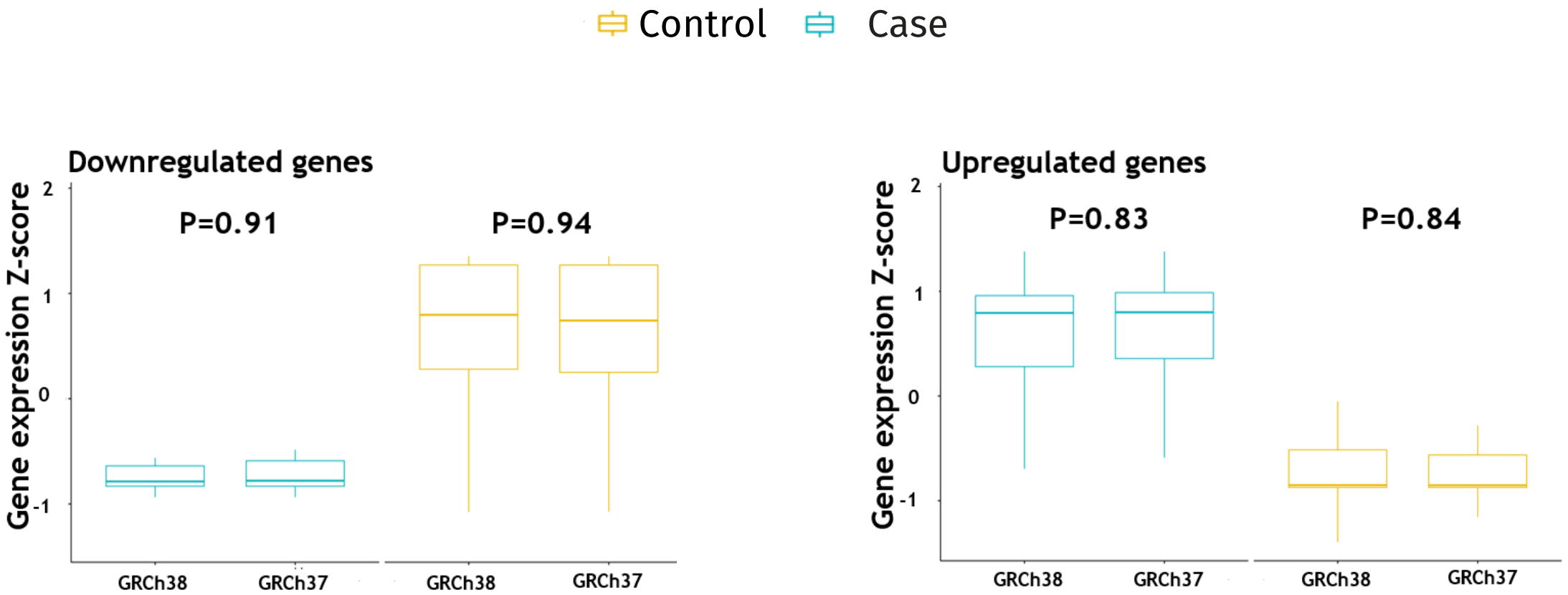

ERR358485
ERR358488
ERR358487
ERR358486
Approach I
PC2: 3% variance
PC1: 97% variance
20
10
0
-10
-20
-100
-50
0
50
100


Control
Case
ERR358485
ERR358488
ERR358487
ERR358486
Approach II
PC2: 3% variance
PC1: 97% variance
20
10
0
-10
-20
-100
-50
0
50
100


ERR358487 GRCh37
ERR358488 GRCh38
ERR358487 GRCh38
ERR358488 GRCh37
Approach III
PC2: 3% variance
PC1: 97% variance
20
10
0
-10
-20
-100
-50
0
50
100

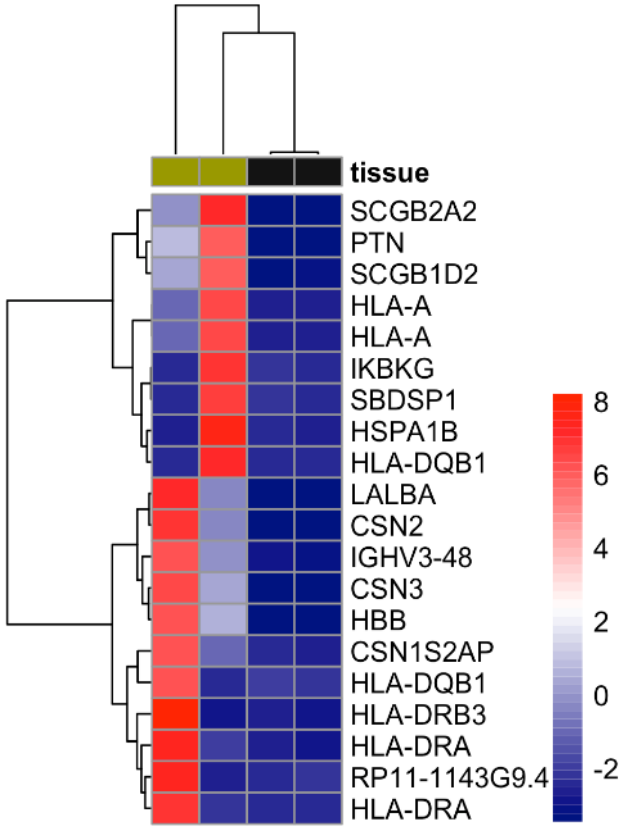

ERR358486 GRCh38
ERR358485 GRCh38
ERR358486 GRCh37
ERR358485 GRCh37
Approach IV
PC2: 40% variance
PC1: 58% variance
20
10
0
-10
-20
-20
0
20

Results: T-test of Naderi genes

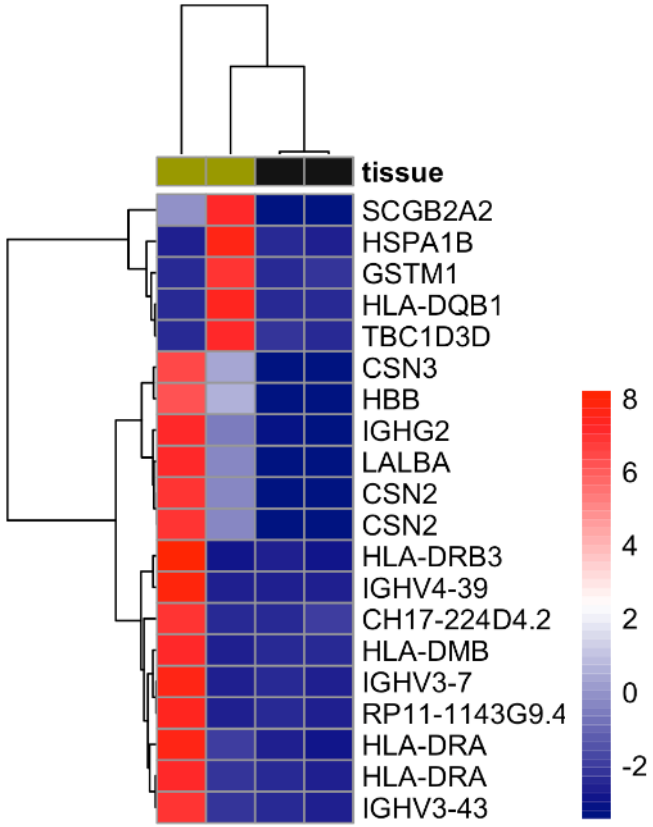


Results: Top20 DE genes

Approach I



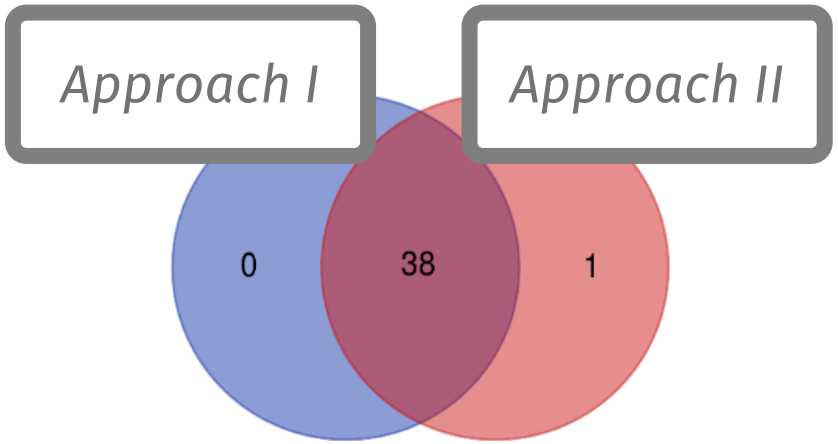
Approach II



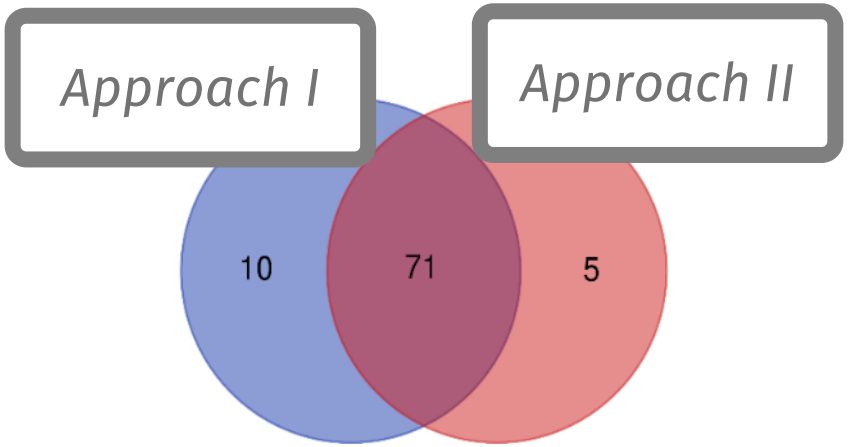
→ overlap in 11 genes

Results: GSEA of DEGs

Hallmark pathways



KEGG pathways



Conclusion

- The choice of the transcriptomic annotation version does not affect differential gene expression estimates nor the prognosis

This it the end, thank you!