Gene Set Enrichment Analysis (GSEA)

Nicolas Delhomme, Bastian Schiffthaler

What to do with your list of genes?

- Mostly, you will select a few candidates and do functional characterisation in the lab, right?
- So you went ballistic and did a whole genomic analysis and you dare look at only a couple genes?

There must be another way

- Well, sort of. GSEA is one.
- GSEA aims at answering the question: is my list of genes (the gene set) associated with experimental condition
  - e.g. are there unusually many de-regulated genes in my gene list
  - e.g. is my DE gene list enriched for some functional processes

GSEA methods

- Reviewed in Kharti et al., 2012
  - Over-representation analysis (ORA) - are differentially expressed (DE) genes in the set more common than expected?
  - Functional class scoring (FCS) - summarize statistic of DE of genes in a set, and compare to null
  - Pathway topology (PT) - include pathway knowledge in assessing DE of genes in a set

GSEA you know?

GSEA you know?

- Gene Ontology Annotation
- KEGG
- reactome
- …

GSEA methods (1)

- Competitive methods (Goeman and Bühlmann, 2007): depend on a competitive null hypothesis which assumes the genes in a set do not have a stronger association with the experimental condition compared to randomly chosen genes outside the set.
- ORA, GAGE, Camera, GSVA, …

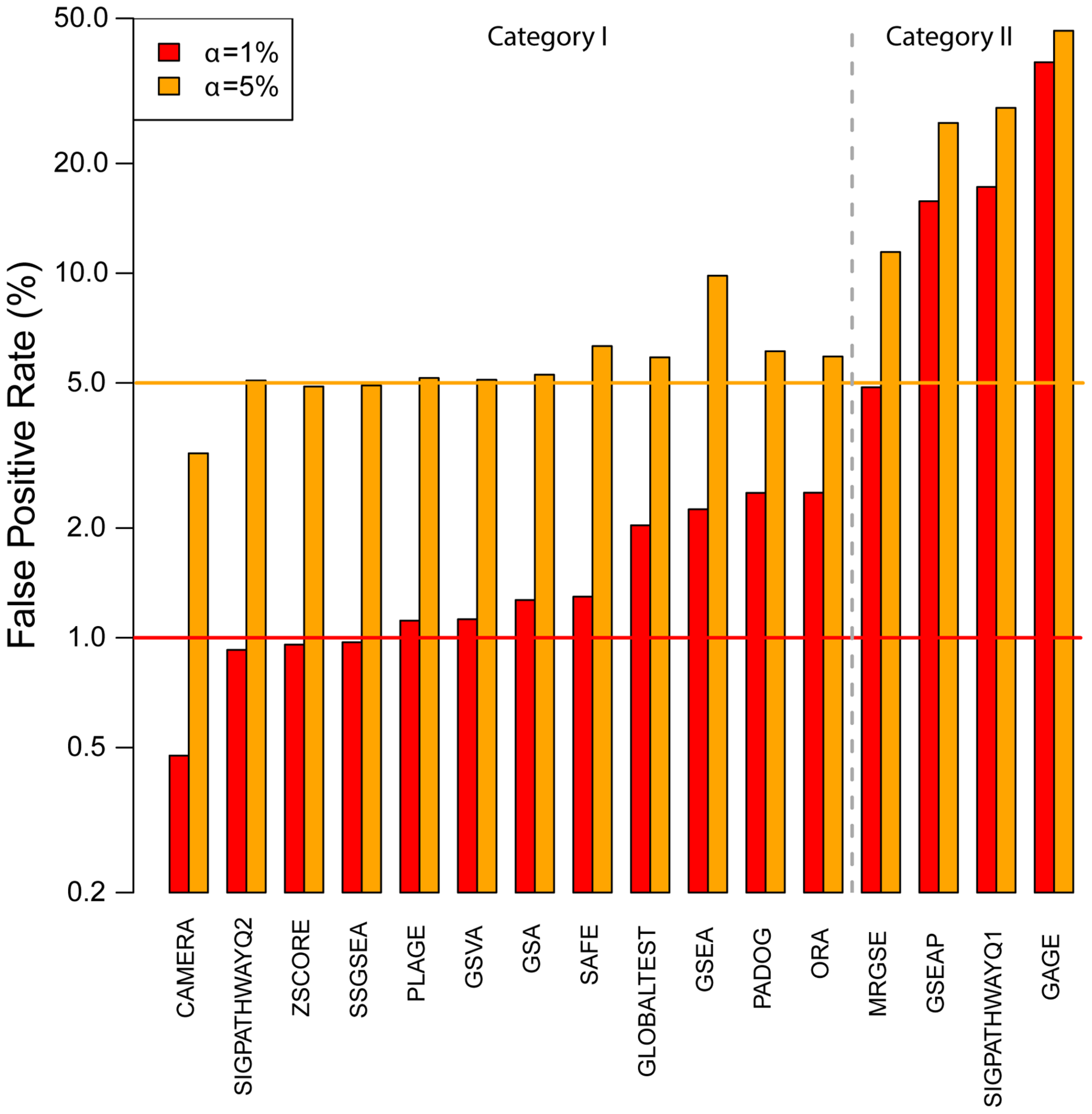
GSEA methods (2)

- Self-contained methods: null hypothesis that only considers genes within a set and again assumes that they have no association with the experimental condition
- SAFE, ZSCORE, SSGSEA, …

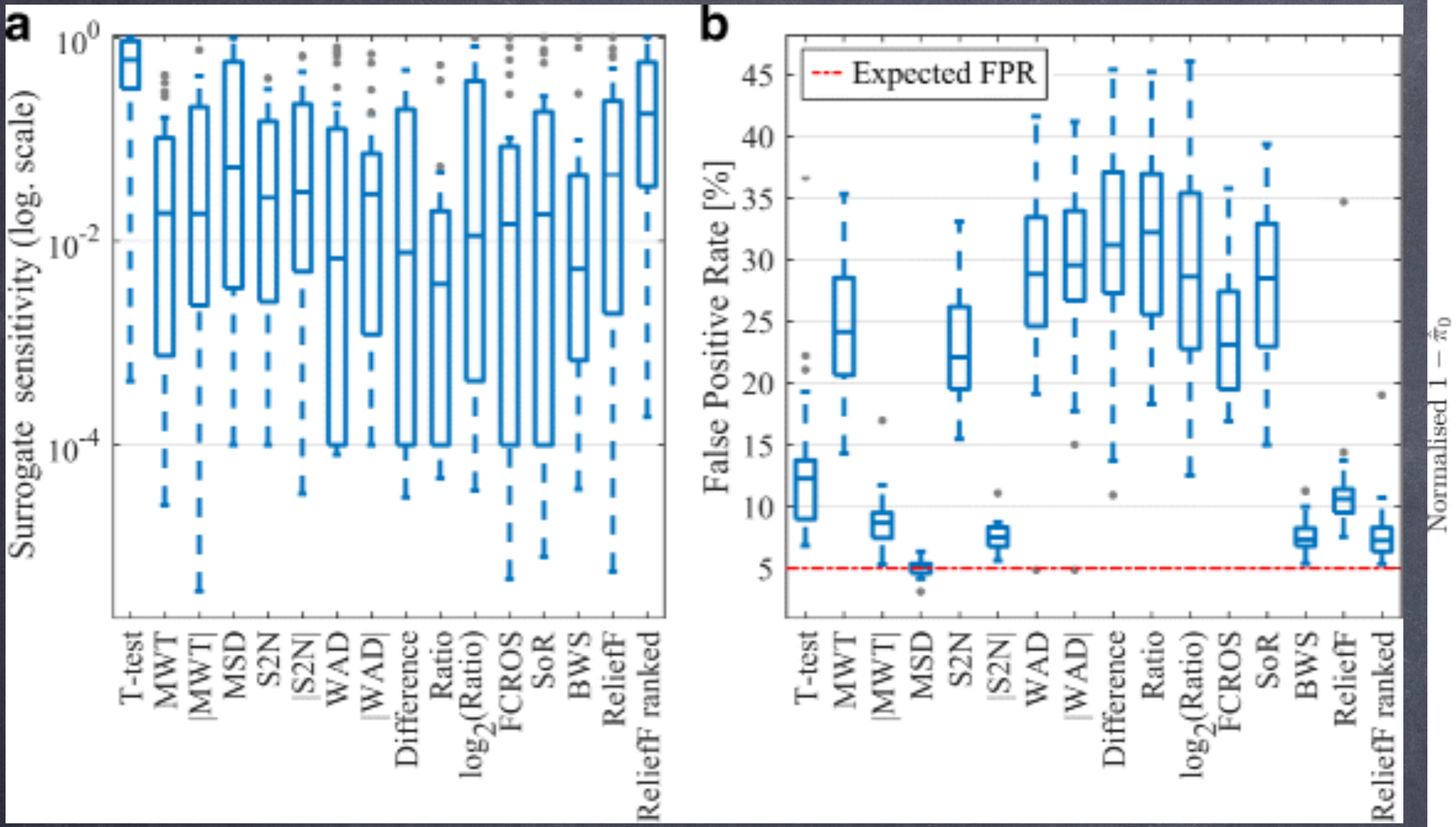
What to pick?

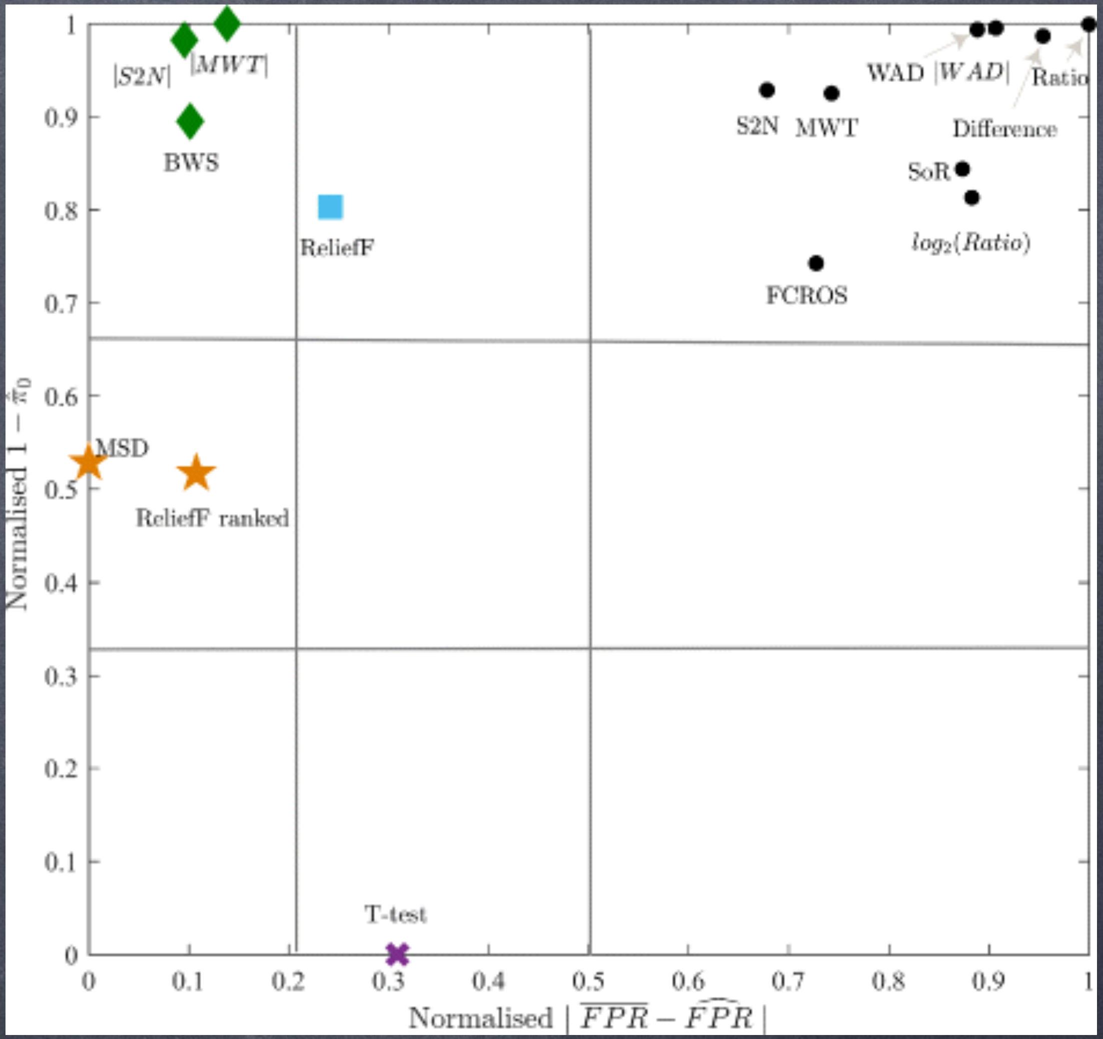
- Maciejewski, 2013
  - sample randomisation based methods are better
  - popular methods (GSEA, SAFE) do not strictly test the competitive null hypothesis: A significant result from these methods does not necessarily mean that the gene set of interest contains more genes associated with the phenotype than its complement, but it rather means that either the gene set or its complement are associated with the phenotype.

Tarca et al, 2013

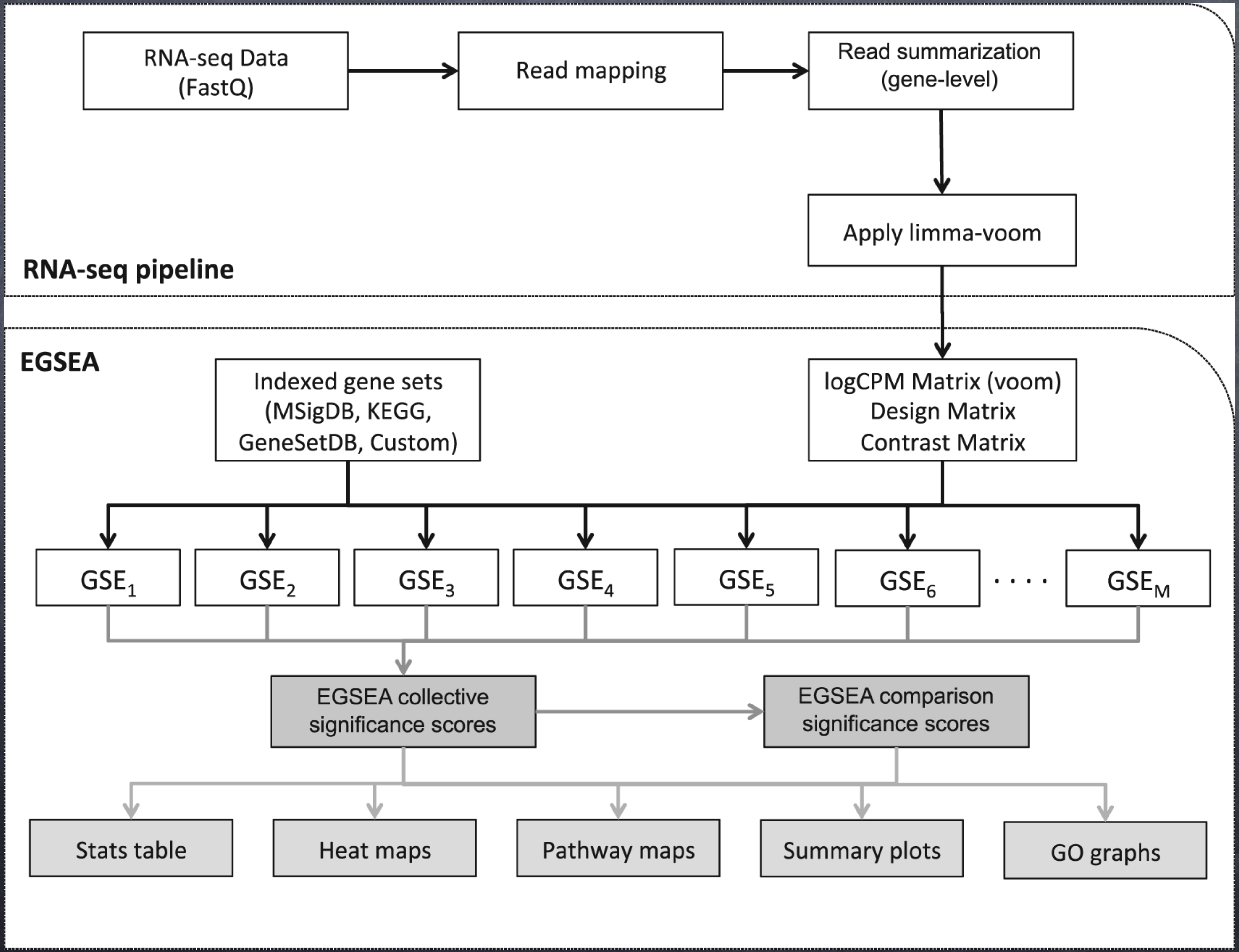


Zyla et al., 2017





Alhamdoosh et al., 2017
Ensemble GSEA



RNA-seq Data (FastQ)

Read mapping

Read summarization (gene-level)

Apply limma-voom

RNA-seq pipeline

EGSEA

Indexed gene sets (MSigDB, KEGG, GeneSetDB, Custom)

logCPM Matrix (voom) Design Matrix Contrast Matrix

GSE$_1$ GSE$_2$ GSE$_3$ GSE$_4$ GSE$_5$ GSE$_6$ . . . . GSE$_M$

EGSEA collective significance scores

EGSEA comparison significance scores

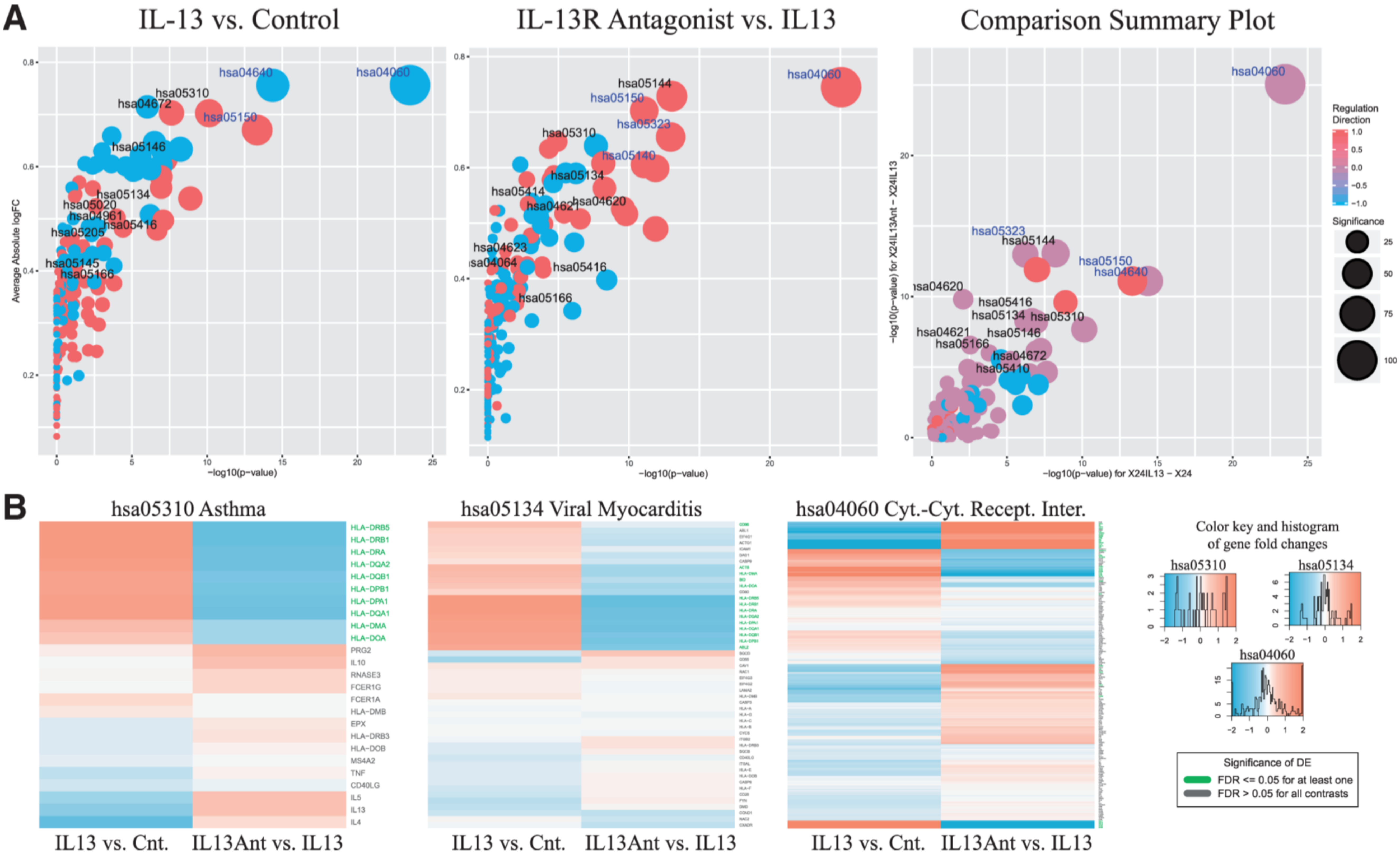
Stats table

Heat maps

Pathway maps

Summary plots

GO graphs


A
IL-13 vs. Control
IL-13R Antagonist vs. IL13
Comparison Summary Plot
Average Absolute logFC
−log10(p−value)
hsa04640
hsa04060
hsa05310
hsa04672
hsa05150
hsa05146
hsa05134
hsa05020
hsa04961
hsa05416
hsa05205
hsa05145
hsa05166
hsa05144
hsa05323
hsa05140
hsa05414
hsa04621
hsa04620
hsa04623
hsa04064
−log10(p−value) for X24IL13Ant − X24IL13
−log10(p−value) for X24IL13 − X24
Regulation Direction
1.0
0.5
0.0
−0.5
−1.0
Significance
25
50
75
100
hsa05410
B
hsa05310 Asthma
HLA-DRB5
HLA-DRB1
HLA-DRA
HLA-DQA2
HLA-DQB1
HLA-DPB1
HLA-DPA1
HLA-DQA1
HLA-DMA
HLA-DOA
PRG2
IL10
RNASE3
FCER1G
FCER1A
HLA-DMB
EPX
HLA-DRB3
HLA-DOB
MS4A2
TNF
CD40LG
IL5
IL13
IL4
IL13 vs. Cnt.
IL13Ant vs. IL13
hsa05134 Viral Myocarditis
CD86
ABL1
EIF4G1
ACTG1
ICAM1
DAG1
CASP9
ACTB
HLA-DMA
BID
HLA-DOA
CD80
HLA-DRB5
HLA-DRB1
HLA-DRA
HLA-DQA2
HLA-DPA1
HLA-DQA1
HLA-DQB1
HLA-DPB1
ABL2
SGCD
CD55
CAV1
RAC1
EIF4G3
EIF4G2
LAMA2
HLA-DMB
CASP3
HLA-A
HLA-G
HLA-C
HLA-B
CYCS
ITGB2
HLA-DRB3
SGCB
CD40LG
ITGAL
HLA-E
HLA-DOB
CASP8
HLA-F
CD28
FYN
DMD
CCND1
RAC2
CXADR
hsa04060 Cyt.-Cyt. Recept. Inter.
Color key and histogram of gene fold changes
hsa05310
hsa05134
hsa04060
Significance of DE
FDR <= 0.05 for at least one
FDR > 0.05 for all contrasts

Let's give it a shot

- http://bioconductor.org/help/workflows/EGSEA123/
- This is still an unrefined workflow (to say the least, but it does look promising)
- Open the R script EGSEAWorkflow.Rmd in RStudio (you copied it earlier)

References

- Some slides inspired from Martin Morgan (Bioconductor) - https://www.bioconductor.org/help/course-materials/2015/CSAMA2015/lect/L16a-gene-set-enrichment-theory-morgan.pdf
- Alhamdoosh et al, 2017, Bioinformatics, 10.1093/bioinformatics/btw623
- Glass and Girvan, 2014, Scientific Reports, 10.1038/srep04191
- Goeman & Bühlmann, 2007, Bioinformatics 23.8: 980-987.
- Grossmann et al., 2007, Bioinformatics, 10.1093/bioinformatics/btm440
- Khatri et al., 2012, PLoS Comp Biol 8.2: e1002375.
- Maciejewski, 2013, Brief. in Bioinf., 10.1093/bib/bbt002
- Tarca et al, 2013. PLOS ONE, 10.1371/journal.pone.0079217
- Zyle et al., 2017, BMC Bioinformatics 10.1186/s12859-017-1674-0