Integrative meta analyses to combine transcriptomics studies

Florian Rohart[1], A. Eslami[2], S. Bougeard[3], C. Wells[1,4], K-A. Lê Cao[5]

[1] Australian Institute for Bioengineering and Nanotechnology (The University of Queensland),
[2]Canada,
[3]France,
[4]University of Glasgow,
[5] The University of Queensland Diamantina Institute

Outline

Outline

Motivation

Heaps of publicly available data that have been under used; frequently used in only one publication with low sample size.

What can we do with a lot of data?

Motivation

Heaps of publicly available data that have been under used; frequently used in only one publication with low sample size.

What can we do with a lot of data?

Combine studies that focus on the same question, 2 ways:

Motivation

Heaps of publicly available data that have been under used; frequently used in only one publication with low sample size.

What can we do with a lot of data?

Combine studies that focus on the same question, 2 ways:

- **meta-analysis:** combines the results obtained on each single study.

  In the context of Differentially Expressed Genes (DEG), a gene is differentially expressed if it is so in every single study => Venn Diagram

Motivation

Heaps of publicly available data that have been under used; frequently used in only one publication with low sample size.

What can we do with a lot of data?

Combine studies that focus on the same question, 2 ways:

- **meta-analysis:** combines the results obtained on each single study.

  In the context of Differentially Expressed Genes (DEG), a gene is differentially expressed if it is so in every single study => Venn Diagram

- **integrative-analysis:** combines the studies to obtain new results.

  DEG analysis on the concatenated data. Increased sample size, which should increase power

Outline

Heaps of data - example used throughout

- **Fibroblasts (Fib):** main connective tissue cells present in the body;
- **human Embryonic Stem Cells (hESC):** pluripotent cells and can become all cell types of the body;
- **human induced Pluripotent Stem Cells (hiPSC):** genetically reprogrammed to an hESC-like state by being forced to express genes and factors important for maintaining the defining properties of embryonic stem cells

Classification framework.
Fibroblasts sit away from hESCs/hiPSC; hESCs and hiPSCs share similarities.

Heaps of data - example used throughout

Training set

Experiment	platform	Fib	hESC	hiPSC
Bock et al., 2011	Affymetrix HT-HG-U133A	6	20	12
Briggs et al., 2013	Illumina HumanHT-12 V4	18	3	30
Chung et al., 2011	Affymetrix HuGene-1.0-ST V1	3	8	10
Ebert et al., 2009	Affymetrix HG-U133 Plus2	2	5	3
Guenther et al., 2010	Affymetrix HG-U133 Plus2	2	17	20
Maherali et al., 2008	Affymetrix HG-U133 Plus2	3	3	15
Marchetto et al., 2010	Affymetrix HuGene-1.0-ST V1	6	3	12
Takahashi et al., 2014	Agilent SurePrint G3 GE 8x60K	3	3	3
total	8 datasets / 5 platforms	43	62	105

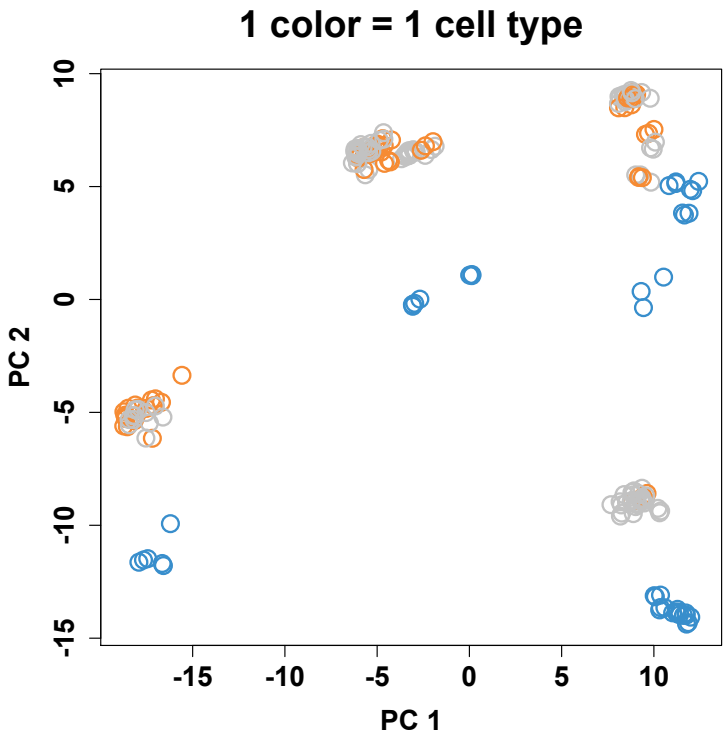
Test set

Experiment	platform	Fib	hESC	hiPSC
Andrade et al., 2012	Affymetrix HuGene-1.0-ST V1	3	6	15
Hu et al., 2011	Affymetrix HG-U133 Plus2	1	5	12
Kim et al., 2009	Affymetrix HG-U133 Plus2	1	1	3
Loewer et al., 2010	Affymetrix HG-U133 Plus2	4	2	7
Si-Tayeb et al., 2010	Affymetrix HG-U133 Plus2	3	6	6
Vitale et al., 2012	Illumina HumanHT-12 V4	8	3	18
Yu et al., 2009	Affymetrix HG-U133 Plus2	2	10	16
total	7 datasets / 3 platforms	22	33	77

Raw data available at www.stemformatics.org. Classical pre-processing: background correction, log2 transform, mapping to Ensembl ID and YuGene normalisation (Lê Cao, Rohart et al. (2014)). Around 15,000 genes

Outline

Unwanted variation/batch effect appears clearly on PCA



Unwanted variation/batch effect appears clearly on PCA

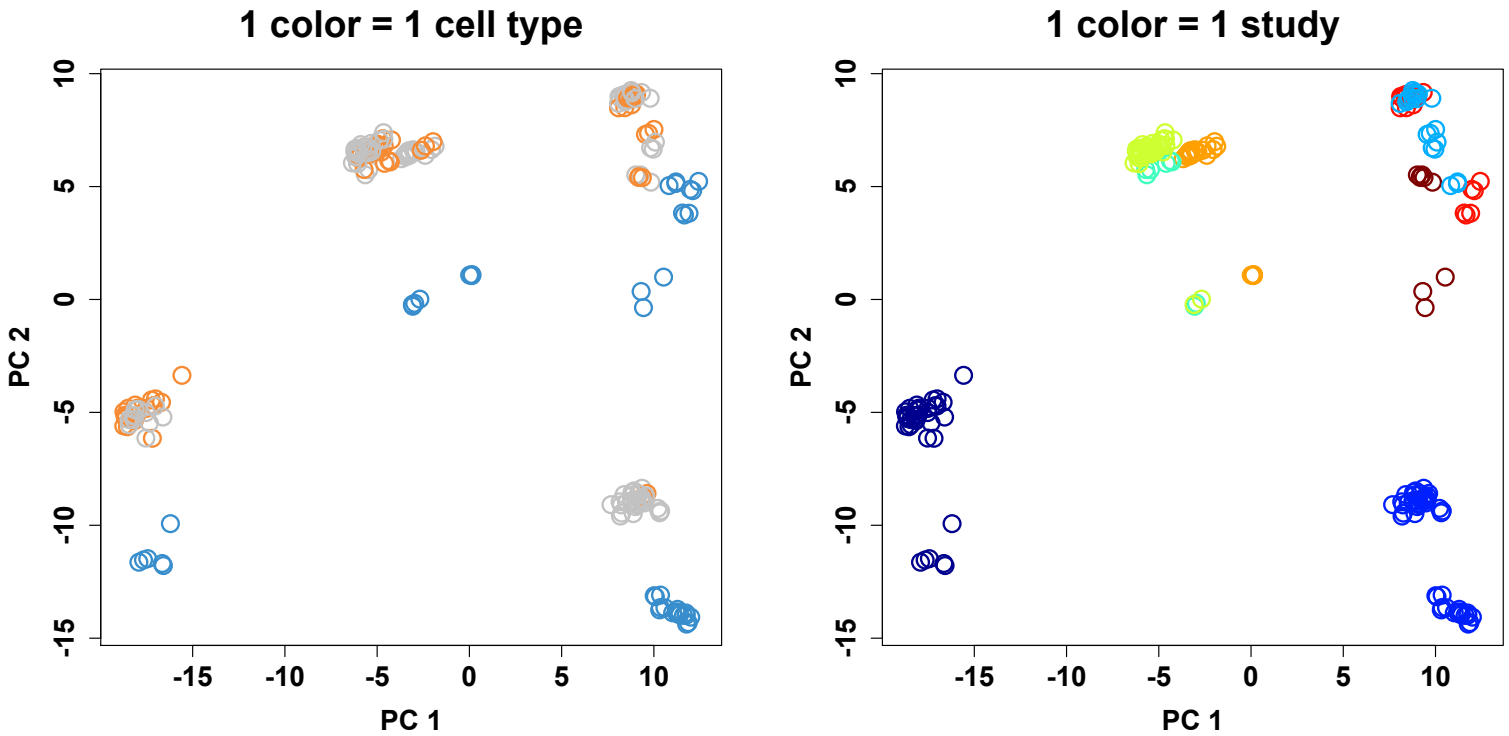


Figure: Between group variance is higher than within group variance. 3 cell types, 8 studies

Outline

Deal with unwanted variation/batch effect

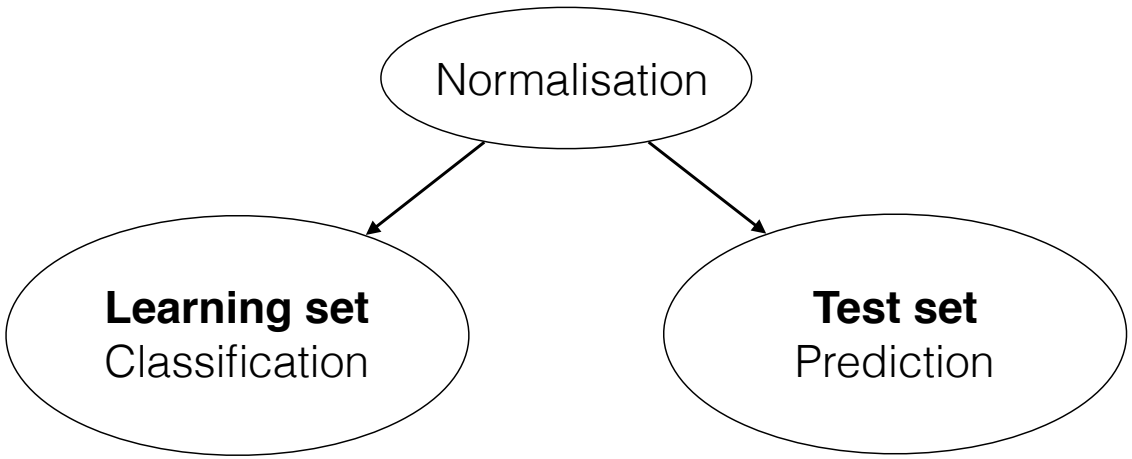
Methods to accommodate batch effects:

- Quantile normalisation (Bolstad et al., 2003),
- batch mean-centering (Sims et al., 2008; Luo et al., 2010),
- ComBat (Johnson, Li, and Rabinovic, 2007),
- YuGene (Lê Cao, Rohart et al., 2014),
- linear model (batch as fixed effect),
- LMM-EH-PS (Listgarten et al., 2010),
- RUV-2 (Gagnon-Bartsch and Speed, 2012),
- ...

Outline

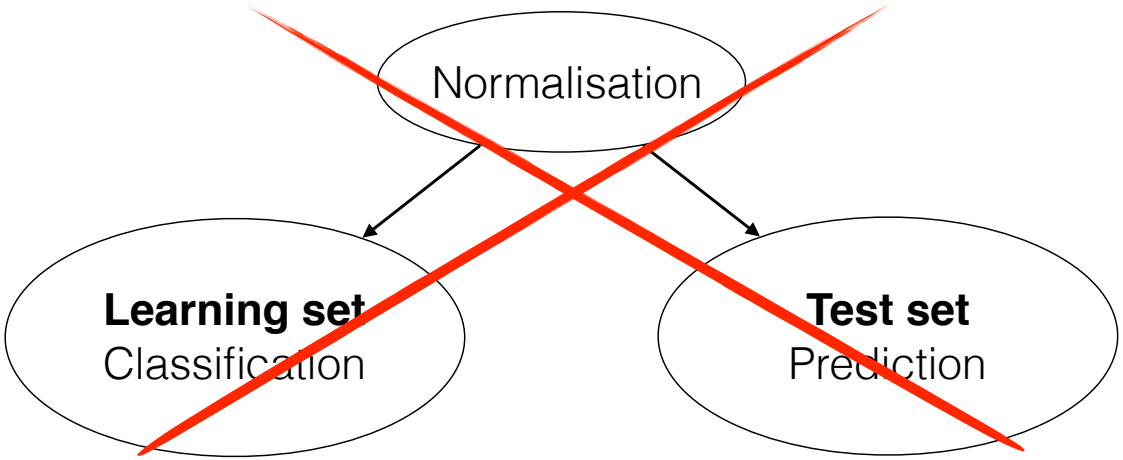
Testing prediction accuracy is problematic - overfitting/bias?

Usually



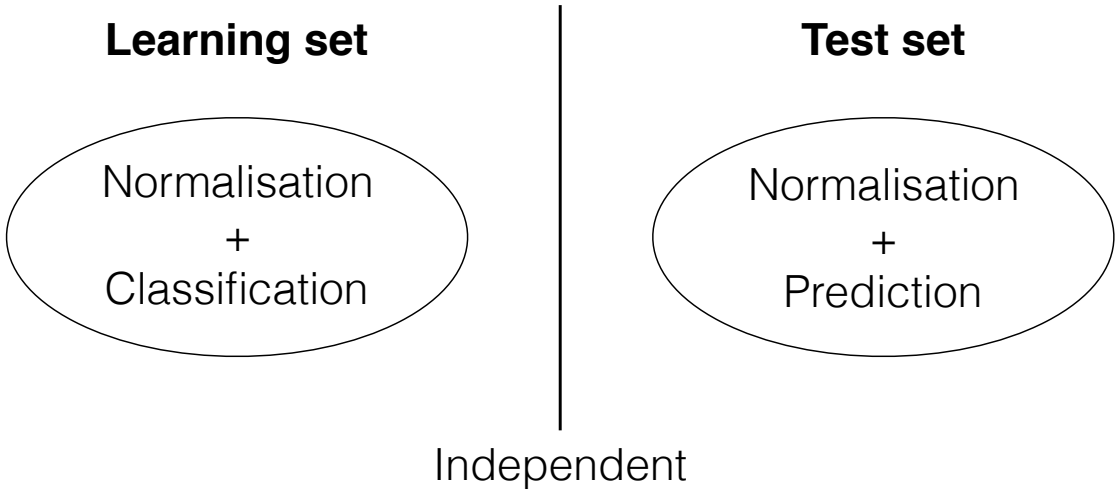
Testing prediction accuracy is problematic - overfitting/bias?

But biased



Testing prediction accuracy is problematic - overfitting/bias?

What should be done



Testing prediction accuracy is problematic - overfitting/bias?

ComBat; state of the art, known to efficiently remove batch effect, **but**

- normalises all data together (CV are biased)
- sensitive to adding/removing samples/datasets
- limited ways to assess downstream efficiency on independent test samples/datasets: no prediction tools except normalising a dataset with the training (Hughey and Butte, 2015) (can be dodgy)

Testing prediction accuracy is problematic - overfitting/bias?

ComBat; state of the art, known to efficiently remove batch effect, **but**

- normalises all data together (CV are biased)
- sensitive to adding/removing samples/datasets
- limited ways to assess downstream efficiency on independent test samples/datasets: no prediction tools except normalising a dataset with the training (Hughey and Butte, 2015) (can be dodgy)

Linear (mixed) models

- mostly no way to assess downstream efficiency on independent test datasets
- no prediction tools for new dataset after normalising by linear (mixed) models

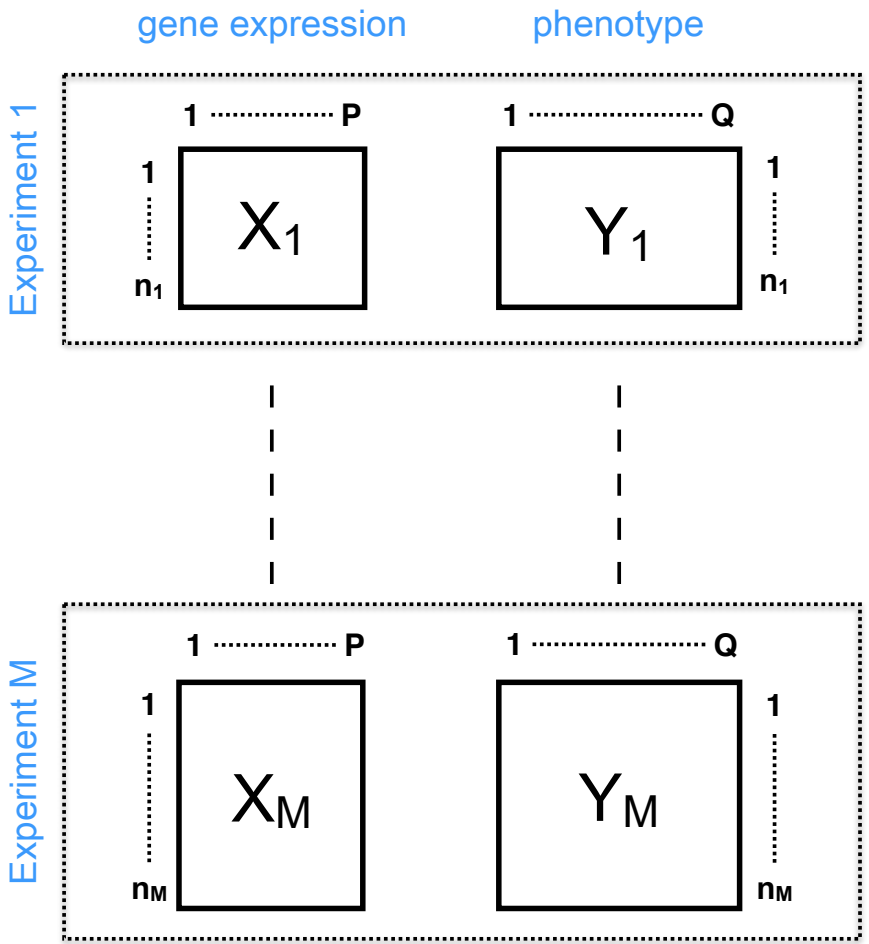
Aims

Because normalization should be done with downstream analysis in mind,

we propose a new method that simultaneously aims to

- Classify samples from several datasets
- Use only a small subset of variables
- Be applicable, available and useable

Design



X is used to explain Y

Outline

Outline

It's complicated...

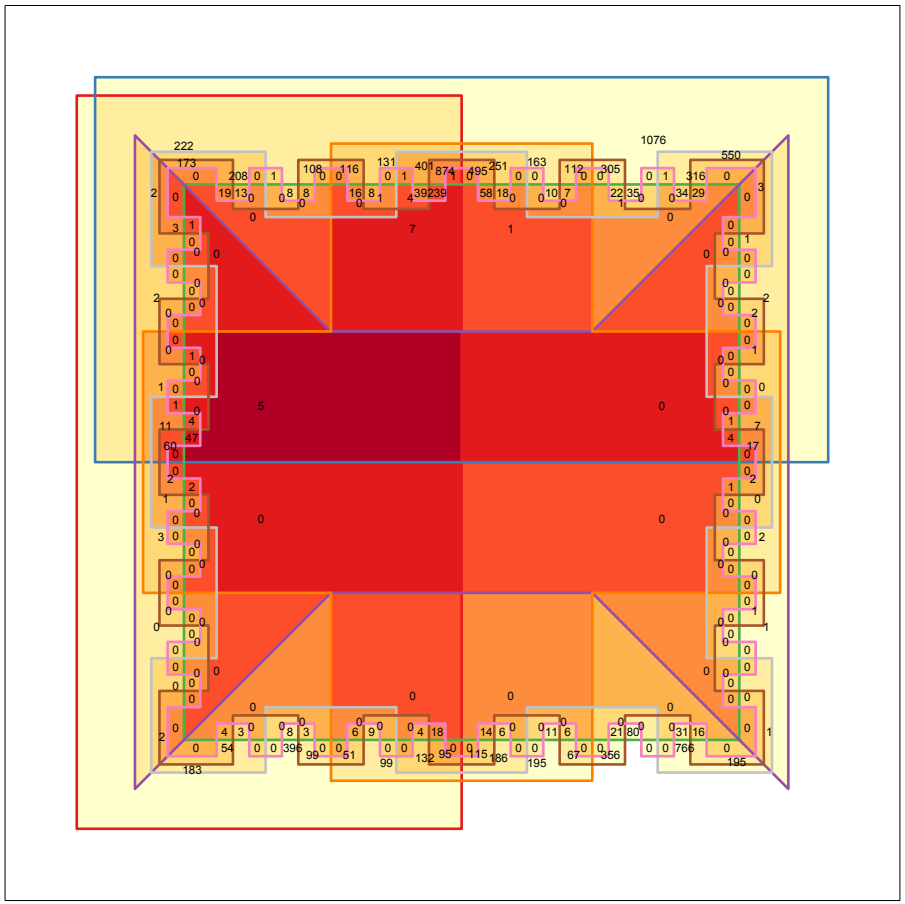
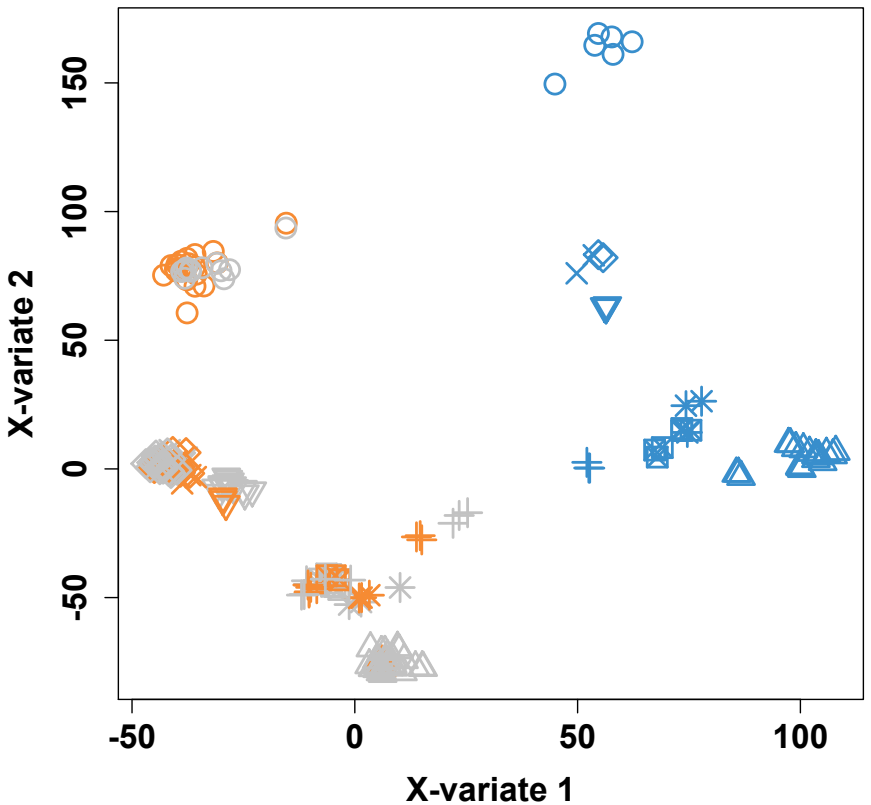


Figure: Venn Diagram of the genes declared as Differentially Expressed with a FDR$< 10^{-5}$.
5 Genes in common.

Outline

Partial Least Square (PLS-DA) on our datasets



○ Bock	◇ Guenther	■ Fibroblast
△ Briggs	▽ Maherali	■ hESC
+ Chung	⊠ Marchetto	■ hiPSC
× Ebert	✳ Takahashi	

Partial Least Square (Wold, 1966): maximise the covariance between linear combinations of X and Y

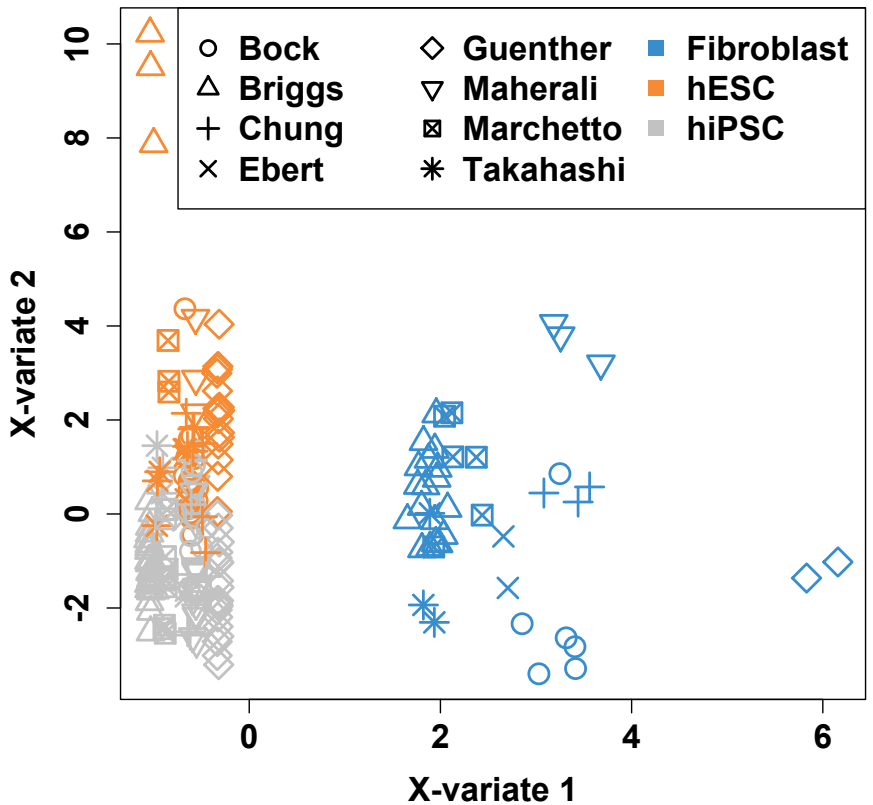
maximise the covariance

$$\max_{||a||=||b||=1} cov(Xa, Yb)$$

- a, b loading vectors
- Xa, Yb PLS-components

Outline

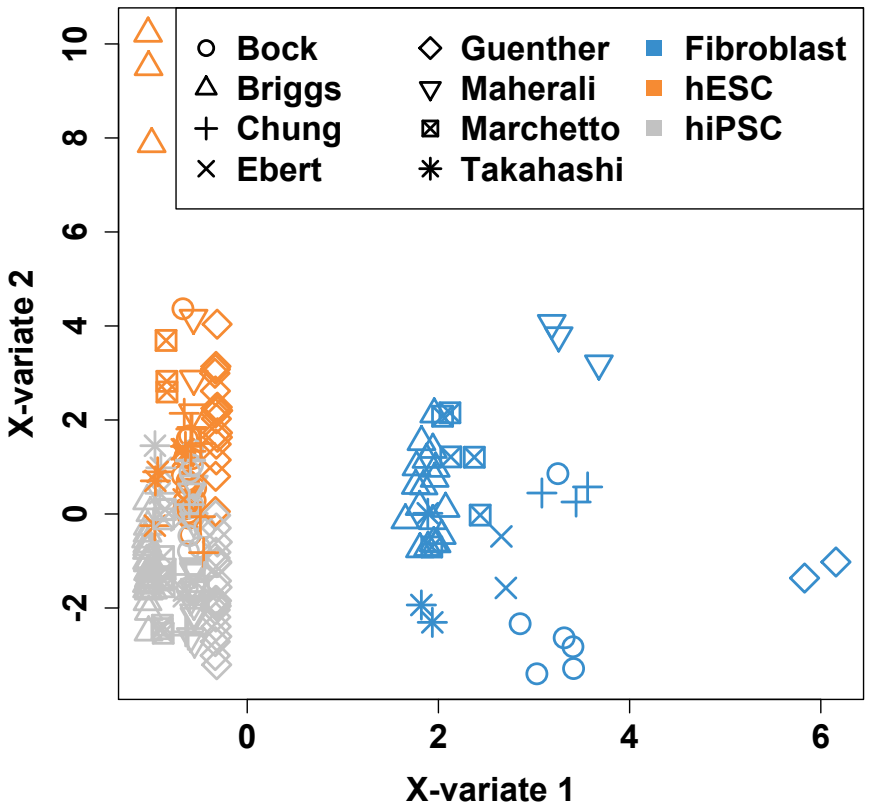
Don't forget the group structure!



Fib	hESC	hiPSC
100	91.9	86.7

Table: Classification accuracy (%), based on 2+15 genes

Don't forget the group structure!



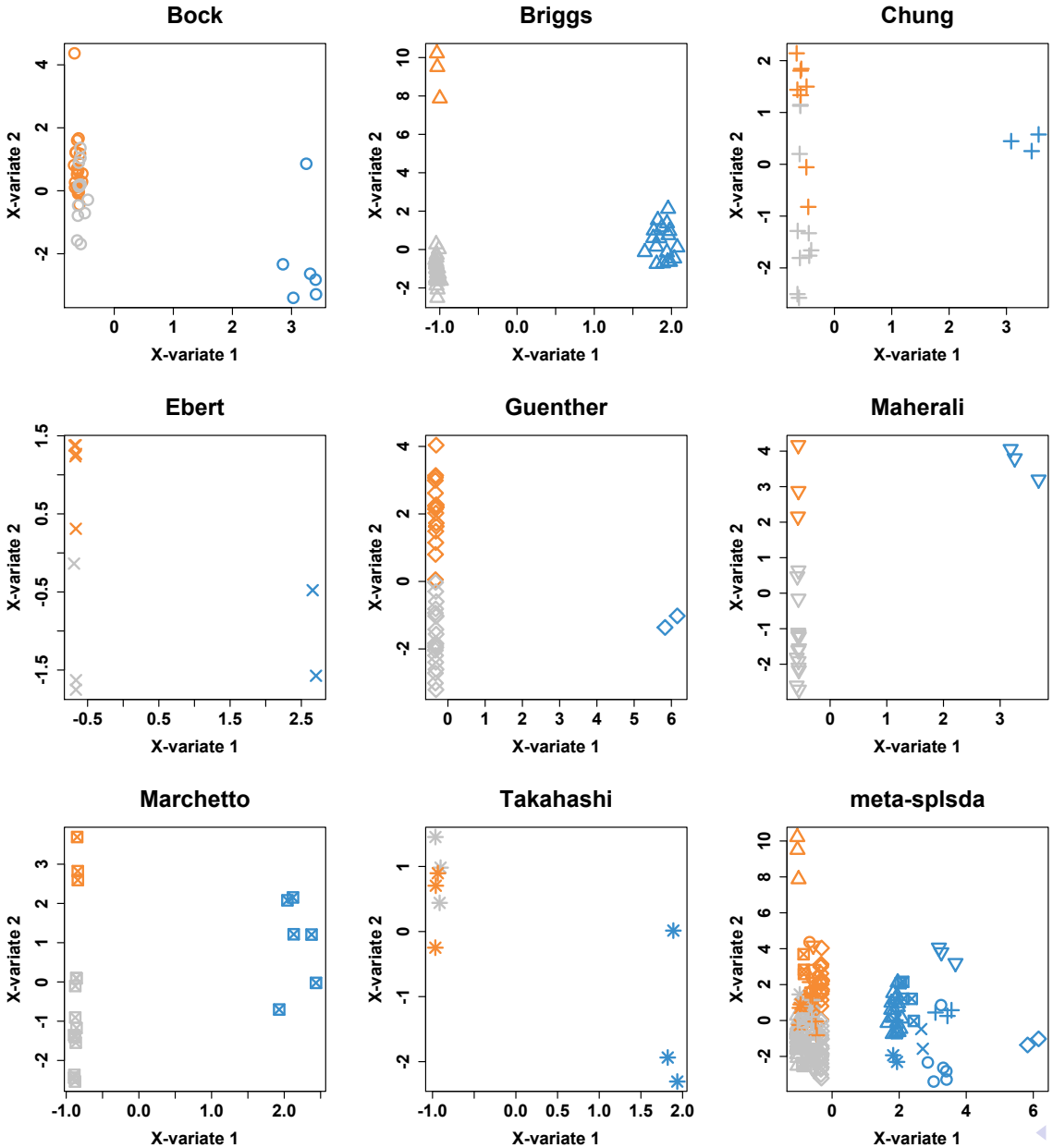
Fib	hESC	hiPSC
100	91.9	86.7

Table: Classification accuracy (%), based on 2+15 genes

meta-splsda

$$\max_{||a||_2=||b||_2=1} \sum_{m=1}^{M} n_m cov(X_m a, Y_m b) + \lambda_1 ||a||_1 + \lambda_2 ||b||_1$$

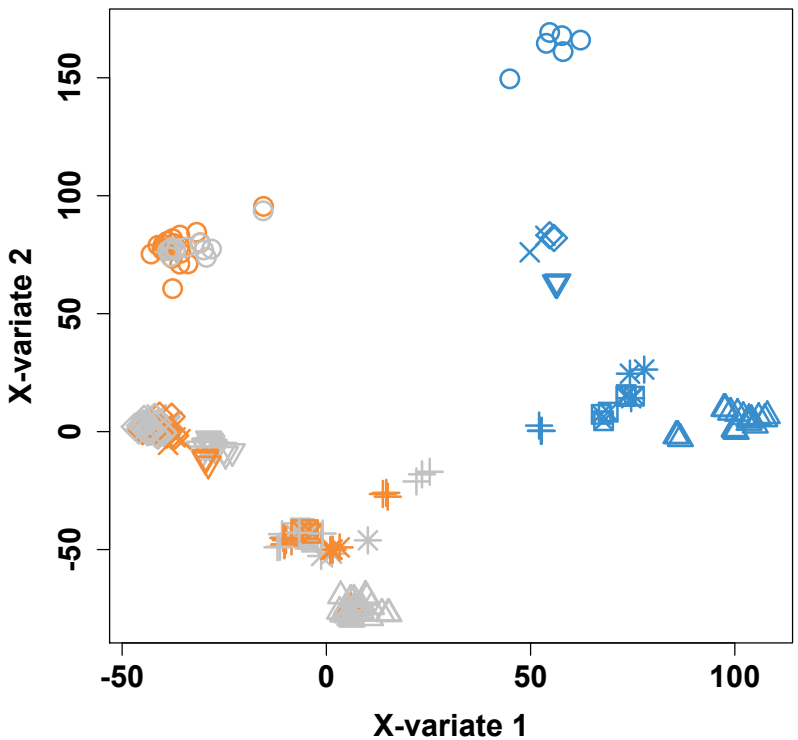
- **global** loading vectors a, b; *shared by all groups*
- **partial** PLS-components $X_m a$, $Y_m b$; *group specific*

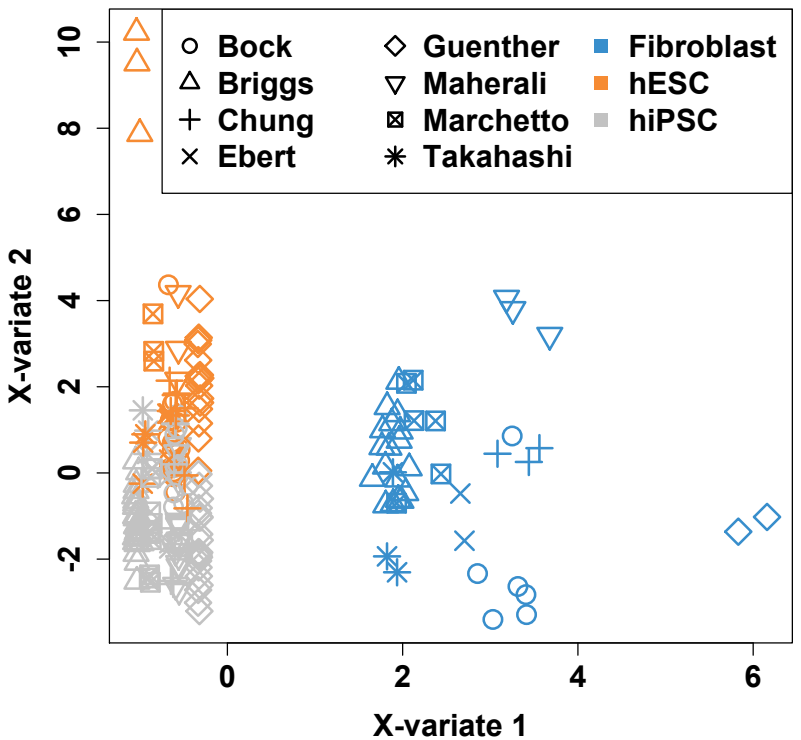


	Study	BER	Fib	hESC	hiPSC
○	Bock	22.2	100	100	33.3
△	Briggs	0.00	100	100	100
+	Chung	15.0	100	75.0	80.0
×	Ebert	11.1	100	100	66.7
◇	Guenther	2.0	100	94.1	100
▽	Maherali	11.1	100	66.7	100
⊠	Marchetto	0.00	100	100	100
✳	Takahashi	44.4	100	66.6	0.00
	overall	7.1	100	91.9	86.7

BER= average of the proportion of wrong classification in each class

Summary, PLS-DA vs meta-splsda





Outline

Combination of methods

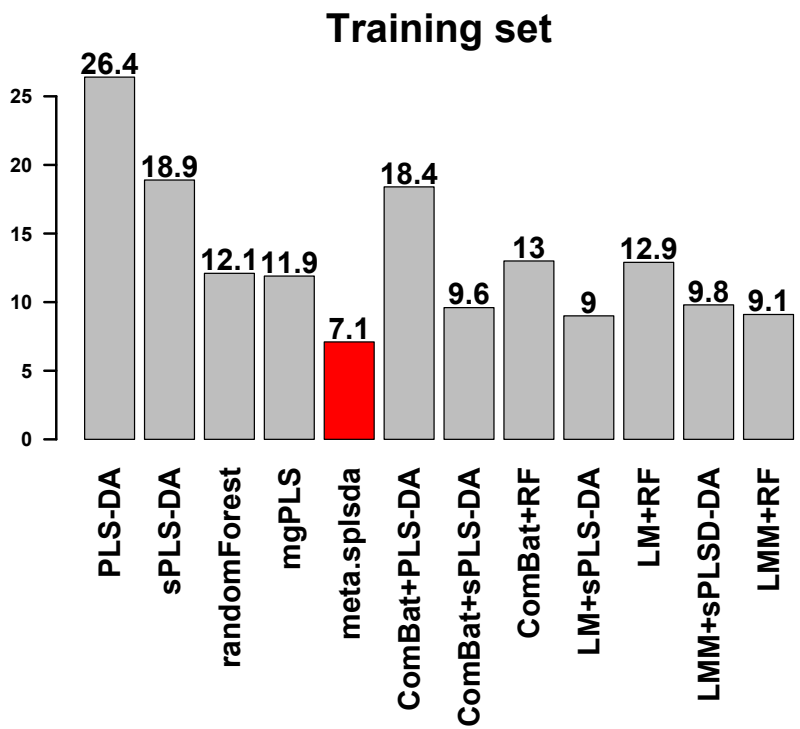
Normalization method:

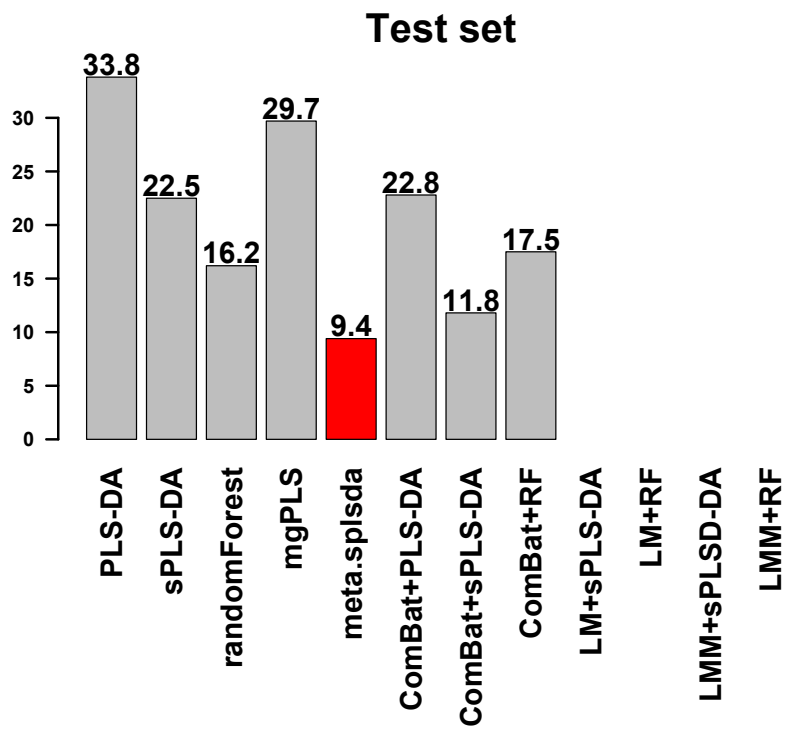
- nothing
- ComBat
- Linear models (LM)
- Linear mixed models, study effect as random (LMM)

Classification/variable selection method:

- PLS-DA
- sPLS-DA
- RandomForest (RF)

Results - Balanced Error Rate (BER)

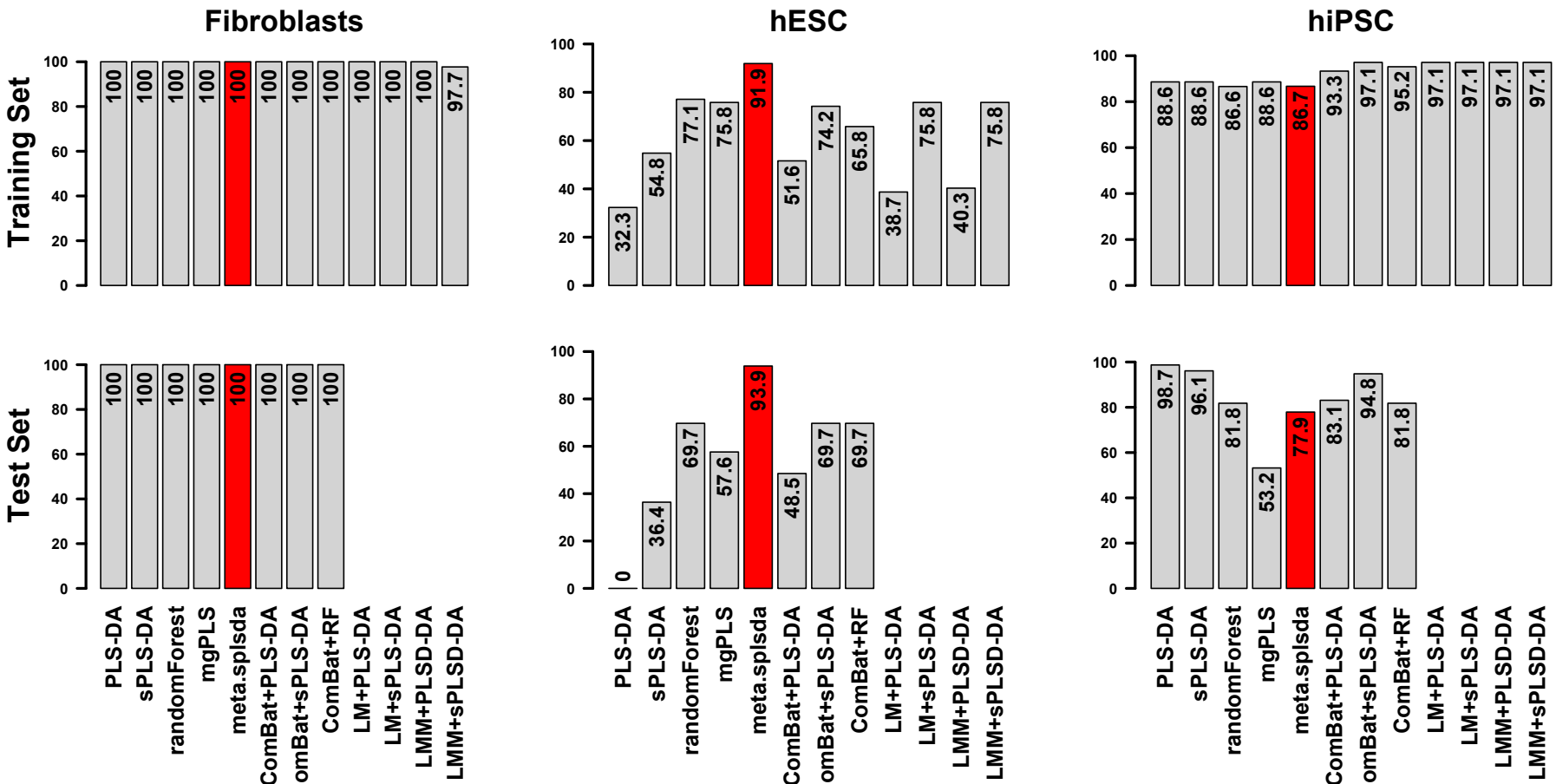




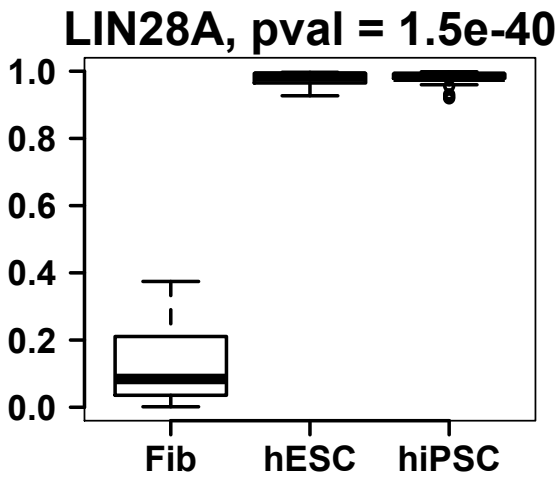
BER: average of the proportion of wrong classification in each class
LM: linear models
LMM: linear mixed models
RF: randomForest

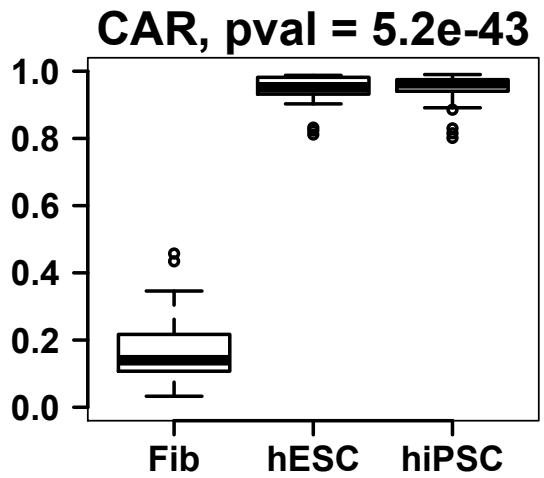
FYI Prediction with ComBat is done as in Hughey and Butte, 2015

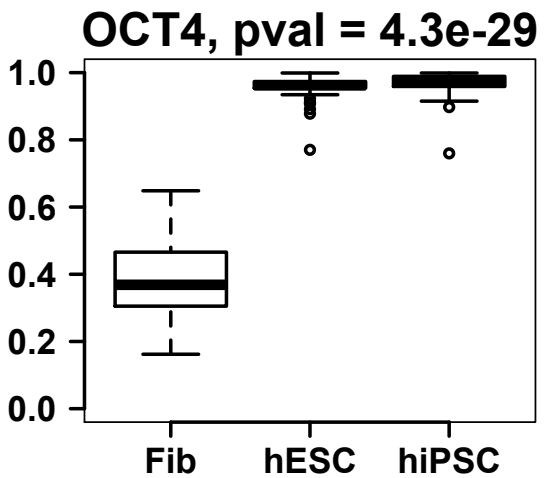
Results - Per cell type



Results - Selected genes. 2 on Comp1, 15 on Comp2







Outline

Conclusions

One single method to:

- accommodate batch effect
- classify samples
- identify biomarkers
- give study-specific graphical outputs

available soon in mixOmics R-package (http://mixOmics.org)

Conclusions

Some remarks

- the studies must share the same characteristics as in a meta analysis: won't work if one level of Y is missing in one study
- better to use more than 3 samples per study
- no p-values
- better to pre-process all studies in a similar way to limit unwanted variation

Thanks

Thanks everyone
(Wells Lab-Stemformatics team, co-authors, and you)