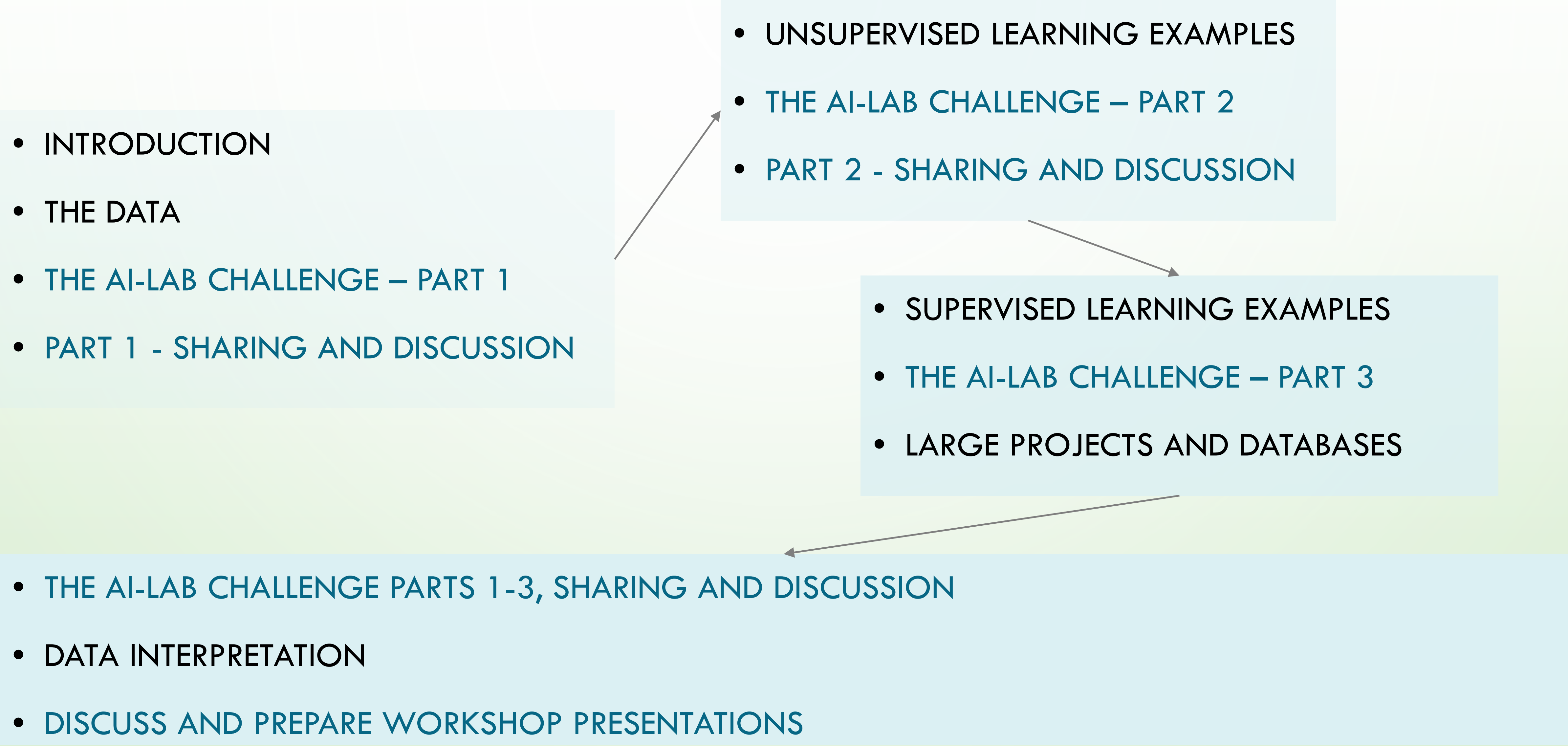
AI LAB

2.1 - UNSUPERVISED LEARNING

FRANCESCA M. BUFFA

LAB STRUCTURE



- INTRODUCTION
- THE DATA
- THE AI-LAB CHALLENGE – PART 1
- PART 1 - SHARING AND DISCUSSION

- UNSUPERVISED LEARNING EXAMPLES
- THE AI-LAB CHALLENGE – PART 2
- PART 2 - SHARING AND DISCUSSION

- SUPERVISED LEARNING EXAMPLES
- THE AI-LAB CHALLENGE – PART 3
- LARGE PROJECTS AND DATABASES

- THE AI-LAB CHALLENGE PARTS 1-3, SHARING AND DISCUSSION
- DATA INTERPRETATION
- DISCUSS AND PREPARE WORKSHOP PRESENTATIONS

UNSUPERVISED LEARNING APPLICATIONS

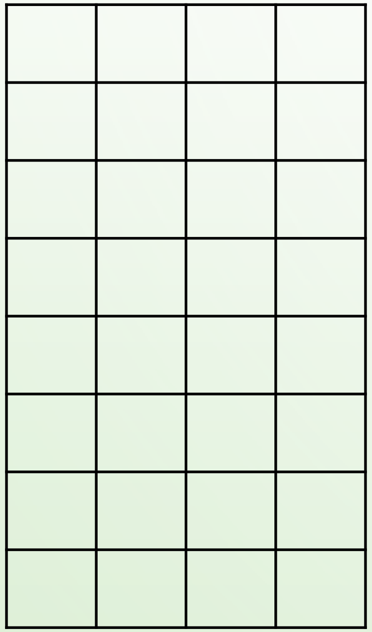
- EXPLORATORY DATA ANALYSIS
- QUALITY CONTROL
- CLASS DISCOVERY
- DIMENSIONALITY REDUCTION

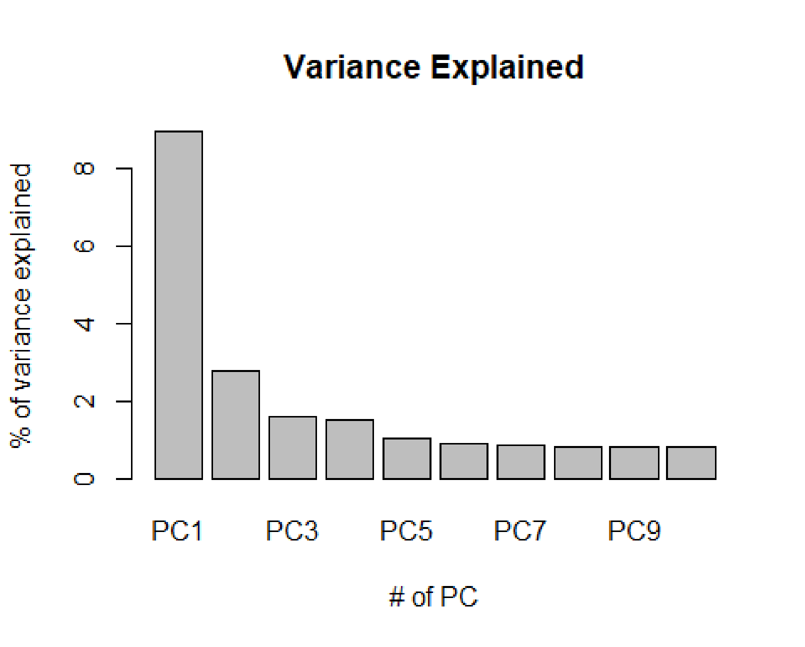
EXAMPLE: QC IN SINGLE CELL SEQUENCING EXPERIMENT

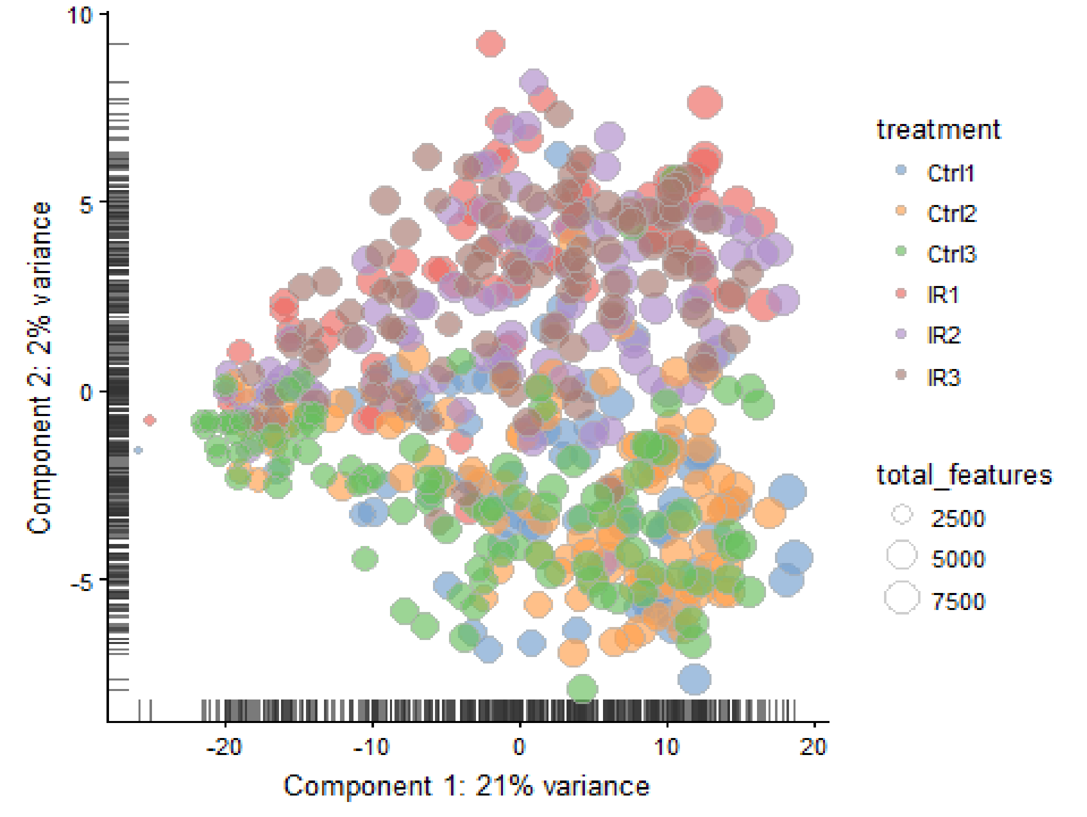
#Cells 528

# Samples (Cntrl vs Treatment)

#Features/Genes 2637

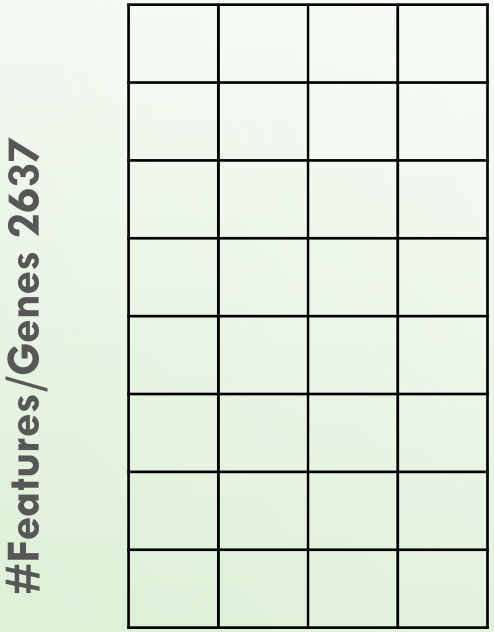


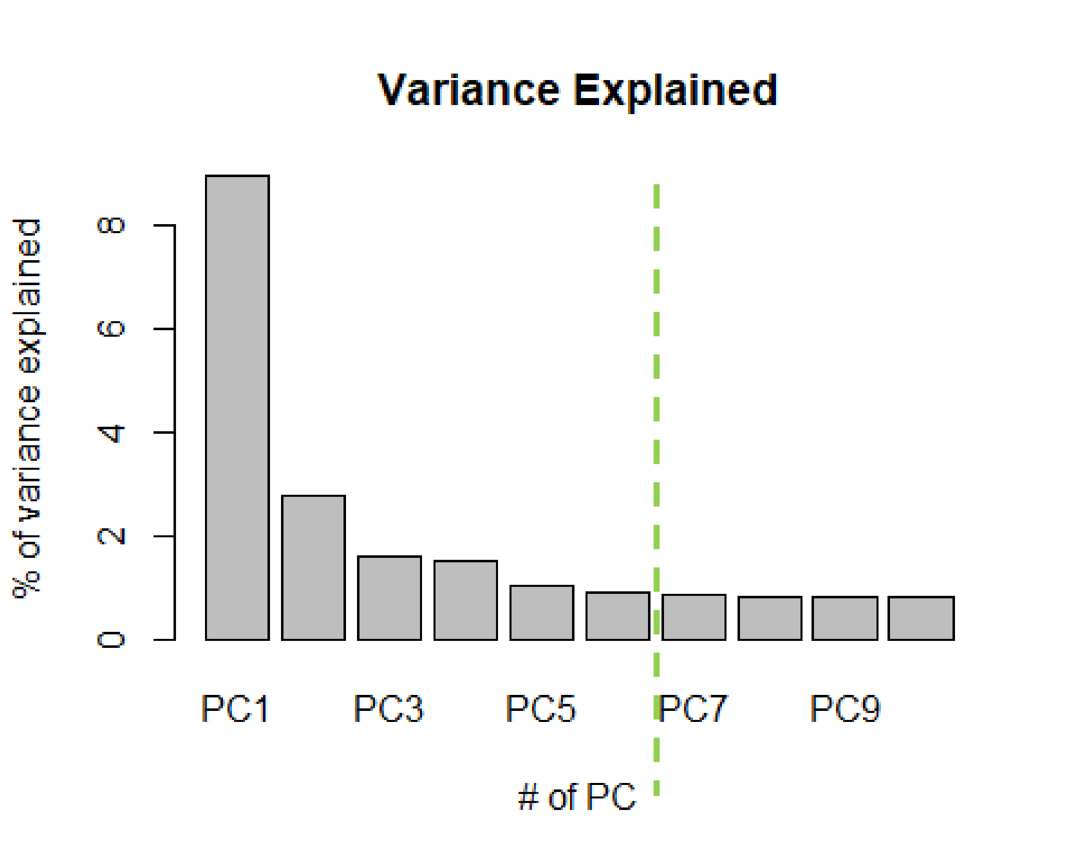




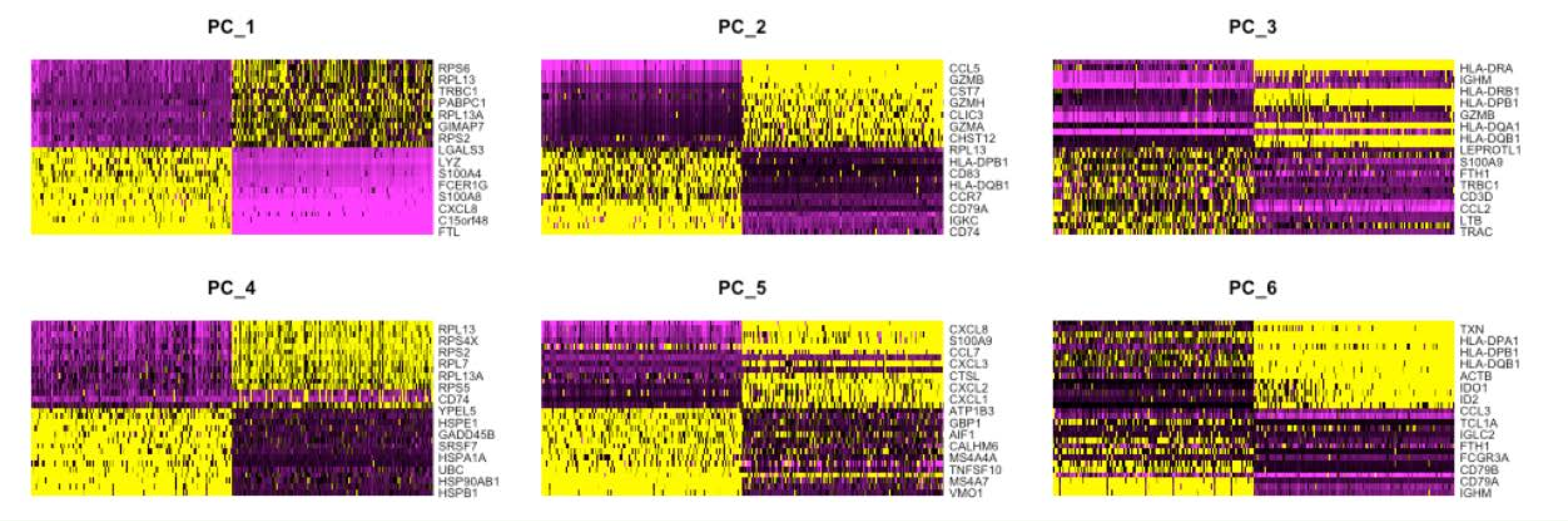
PCA DIMENSIONALITY REDUCTION

#Cells 528
# Samples (Cntrl vs Treatment)

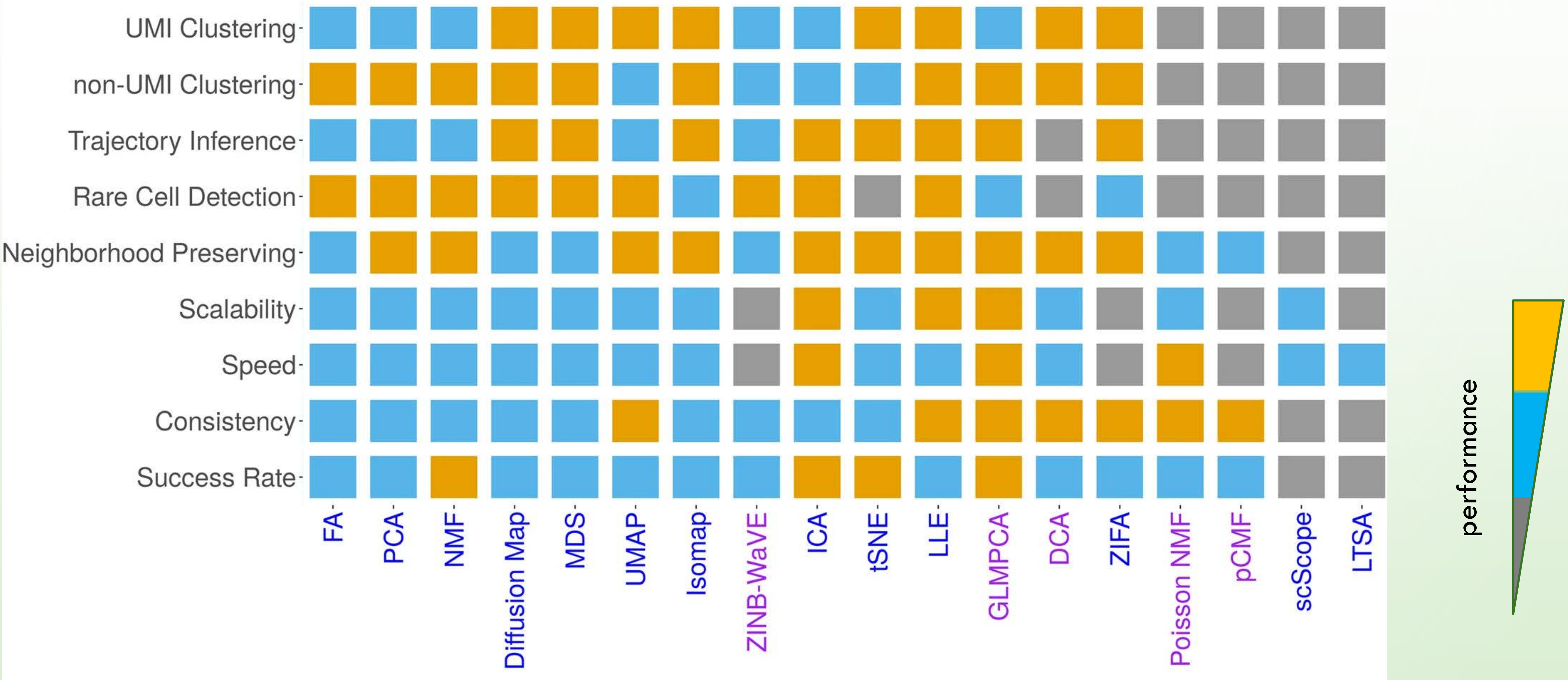




PCA DIMENSIONALITY REDUCTION



DIMENSIONALITY REDUCTION



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1898-6

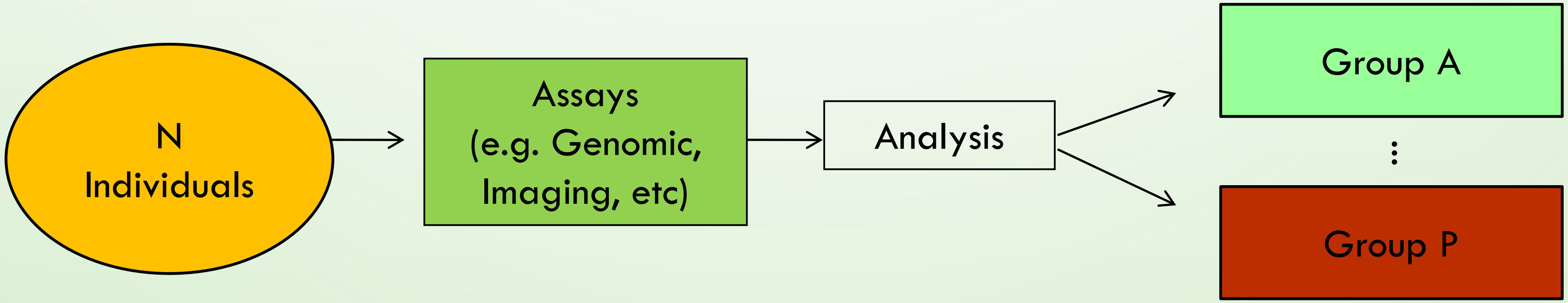
CLUSTERING APPLICATIONS

ARE THERE GROUPS OF SIMILAR DISEASES/CELLS?

“UNSUPERVISED APPROACH”

CLASS DISCOVERY

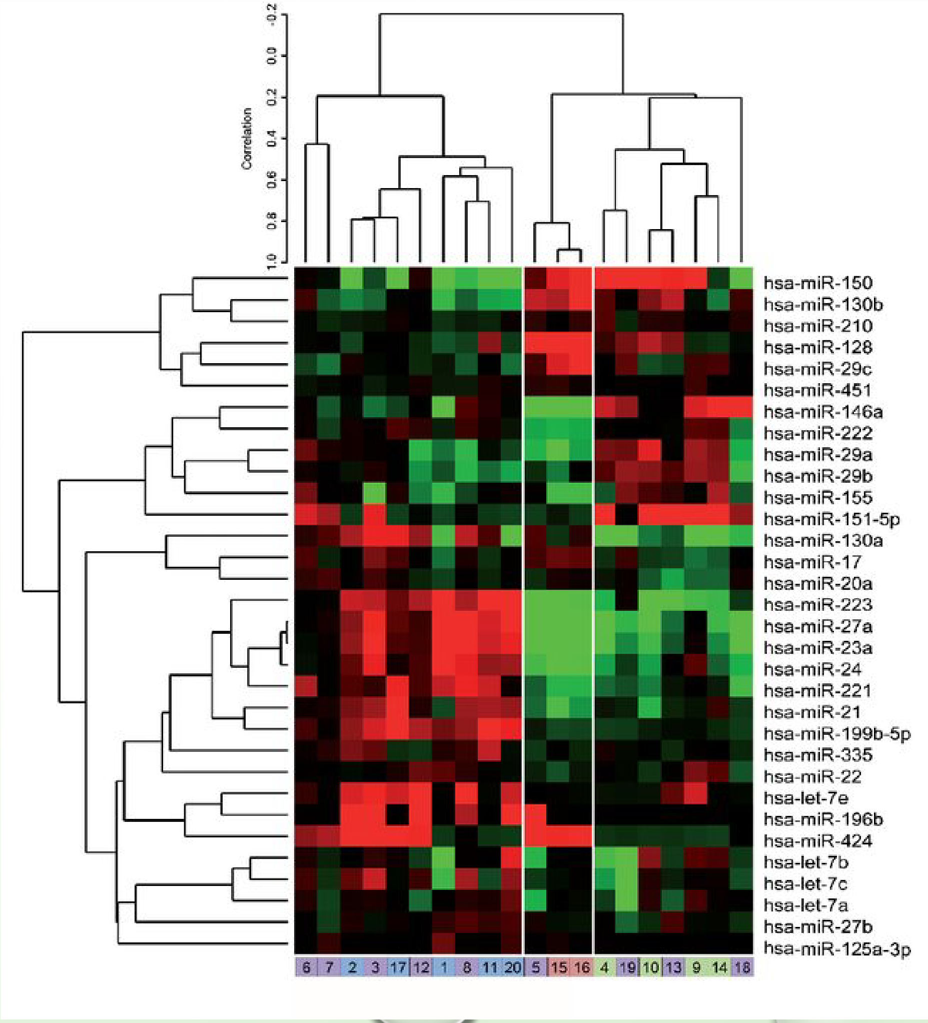
FIND GROUPS OF SIMILAR CASES OR SIMILAR FEATURES



Clustering: guilty-by-association

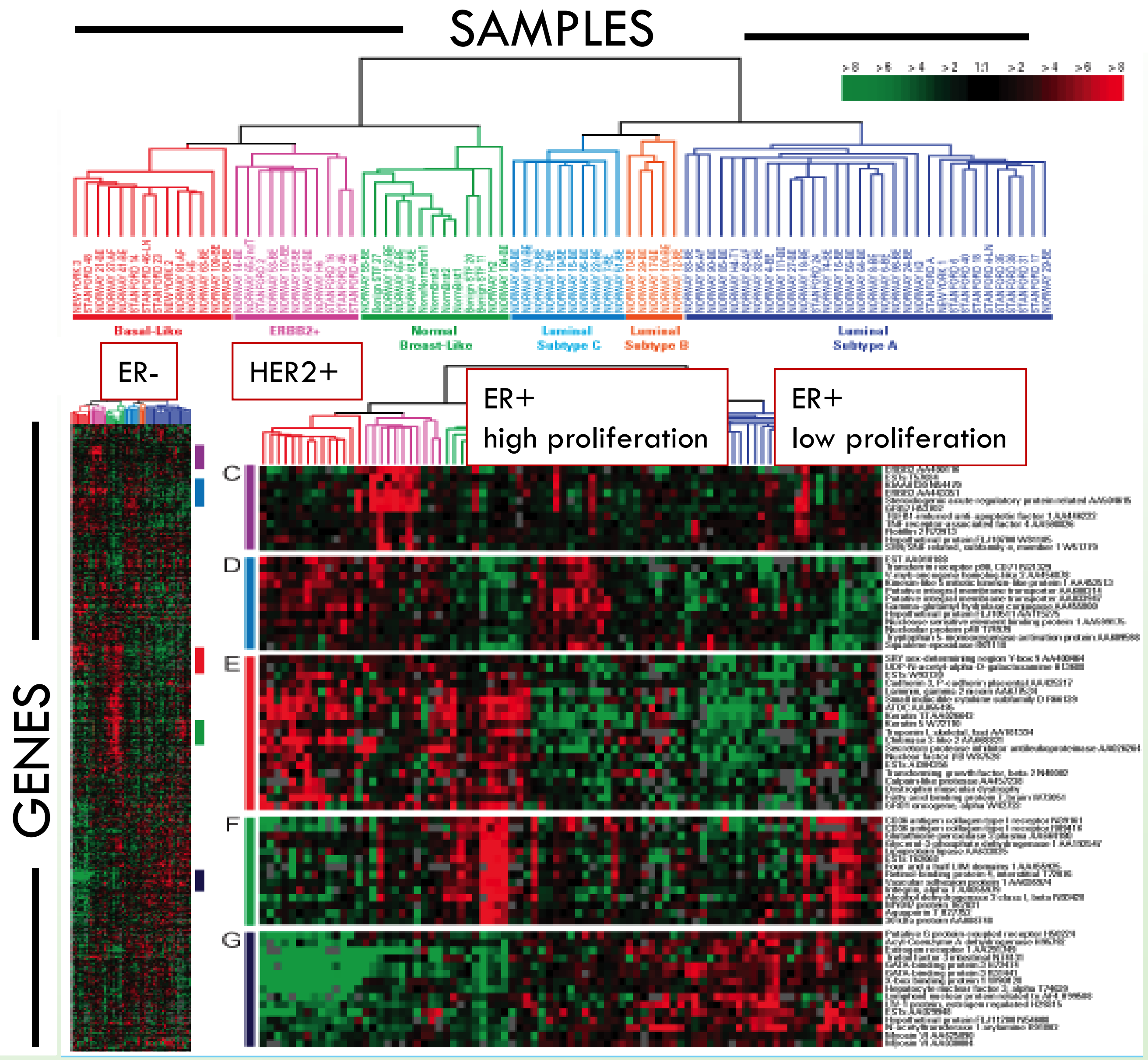
Are my samples similar?
Samples with similar genomic profile might for example have a similar prognosis or response to treatment.
Or in single cell might come from the same cell population.

Are my genes similar?
Suppose genes A and B are grouped in the same cluster. This mean they are expressed under the same conditions. Then we can hypothesize that genes A and B are involved in similar pathways/share function.



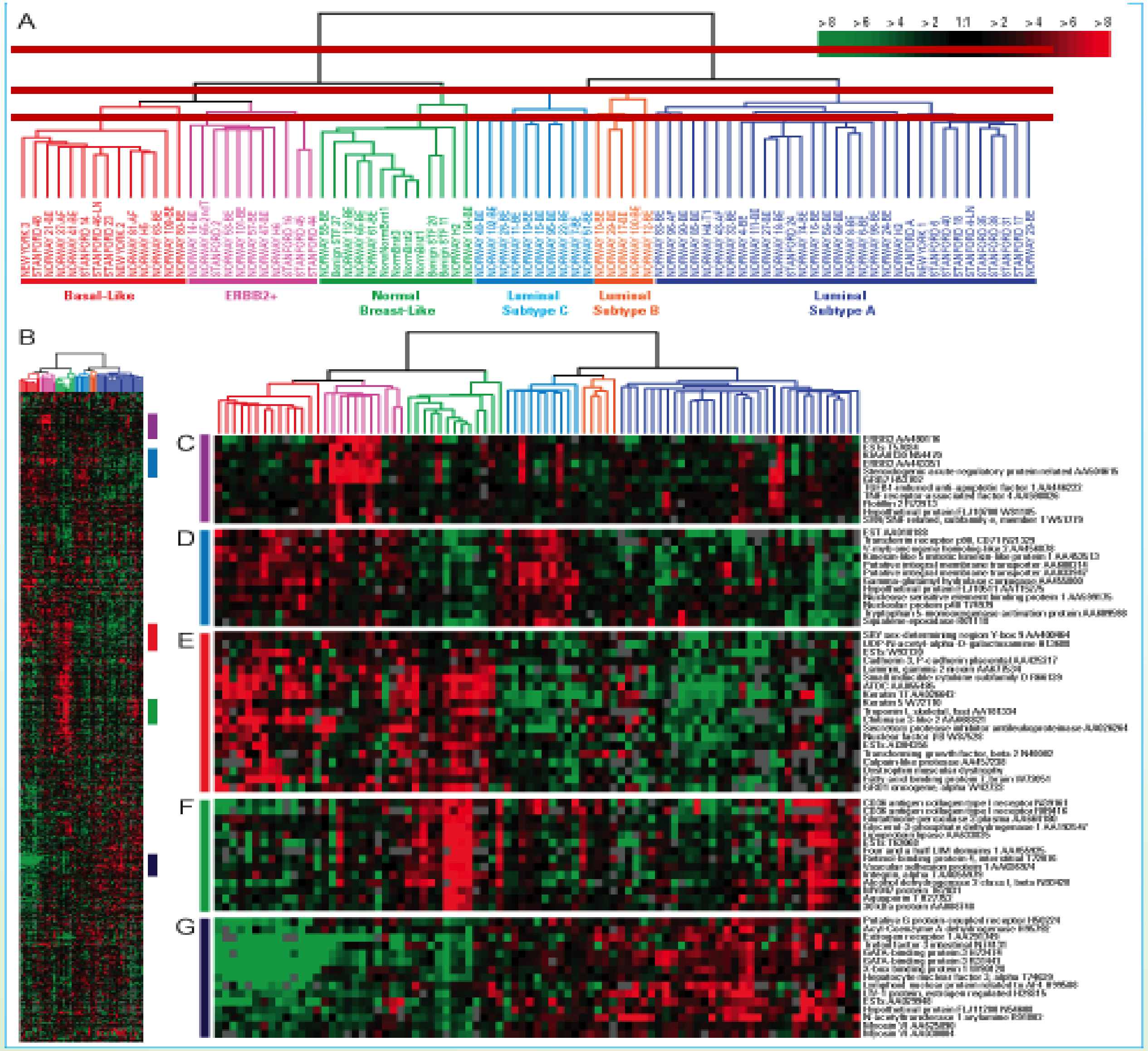
Unsupervised: Breast Cancer Subtypes (PAM50)

- Unsupervised approach: outcome not considered
- Classification based on Expression of Breast Cancer Intrinsic Genes
- Gene expression microarrays
- Hierarchical clustering used to represent distance between samples
- Groups identified matching existing clinical knowledge



Perou et al, Nature 2000
Sorlie et al, PNAS 2001

How many clusters?



Perou et al, Nature 2000
Sorlie et al, PNAS 2001

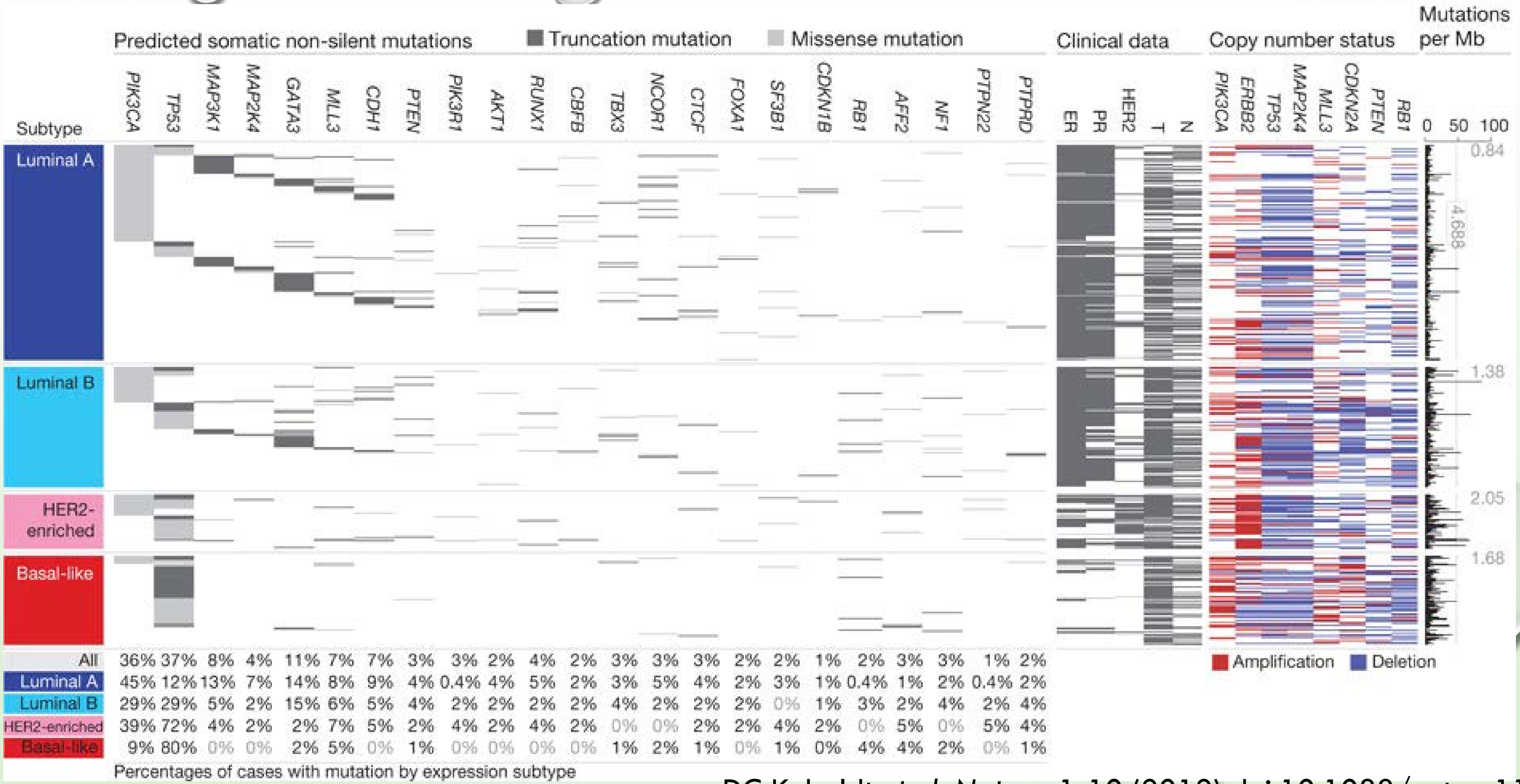
REPRODUCIBILITY

WEIGELT, B ET AL, LANCET ONCOLOGY, 2010

WHEN USING DIFFERENT METHODS TO ASSIGN PATIENTS TO EACH CLUSTER:

- BASAL-LIKE CANCERS WERE CONSISTENTLY CLASSIFIED.
- ASSIGNMENT OF INDIVIDUAL CASES TO LUMINAL A, LUMINAL B, HER2, AND NORMAL BREAST-LIKE SUBTYPES WAS DEPENDENT ON THE METHOD USED.
- THE SIGNIFICANCE OF ASSOCIATIONS WITH OUTCOME OF EACH MOLECULAR SUBTYPE, OTHER THAN BASAL-LIKE AND LUMINAL A, VARIED DEPENDING ON THE METHOD USED.

Link between gene expression clusters and genomic features

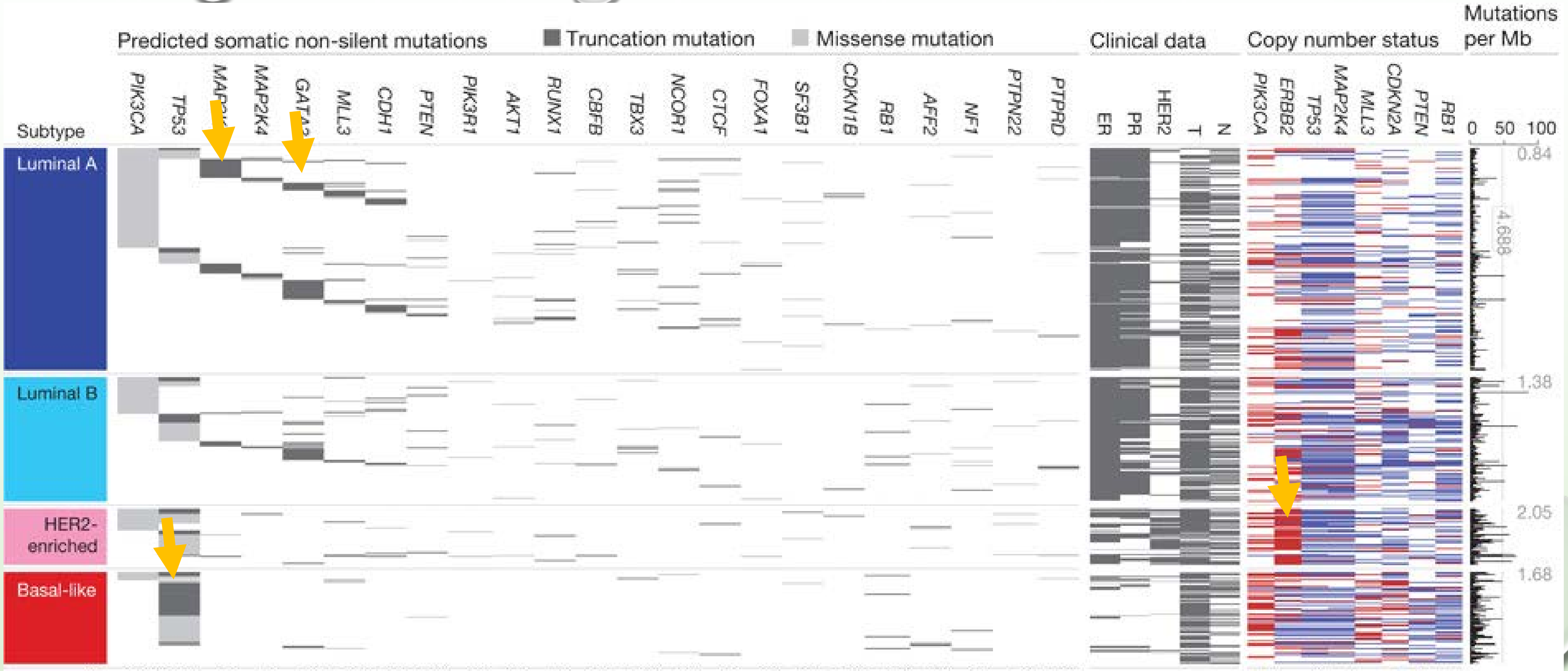


Subtype	PIK3CA	TP53	MAP3K1	MAP2K4	GATA3	MLL3	CDH1	PTEN	PIK3R1	AKT1	RUNX1	CBFB	TBX3	NCOR1	CTCF	FOXA1	SF3B1	CDKN1B	RB1	AFF2	NF1	PTPN22	PTPRD
All	36%	37%	8%	4%	11%	7%	7%	3%	3%	2%	4%	2%	3%	3%	3%	2%	2%	1%	2%	3%	3%	1%	2%
Luminal A	45%	12%	13%	7%	14%	8%	9%	4%	0.4%	4%	5%	2%	3%	5%	4%	2%	3%	1%	0.4%	1%	2%	0.4%	2%
Luminal B	29%	29%	5%	2%	15%	6%	5%	4%	2%	2%	2%	2%	4%	2%	2%	2%	0%	1%	3%	2%	4%	2%	4%
HER2-enriched	39%	72%	4%	2%	2%	7%	5%	2%	4%	2%	4%	2%	0%	0%	2%	2%	4%	2%	0%	5%	0%	5%	4%
Basal-like	9%	80%	0%	0%	2%	5%	0%	1%	0%	0%	0%	0%	1%	2%	1%	0%	1%	0%	4%	4%	2%	0%	1%

Percentages of cases with mutation by expression subtype

DC Koboldt *et al. Nature*, 1-10 (2012) doi:10.1038/nature11412

Link between gene expression clusters and genomic features

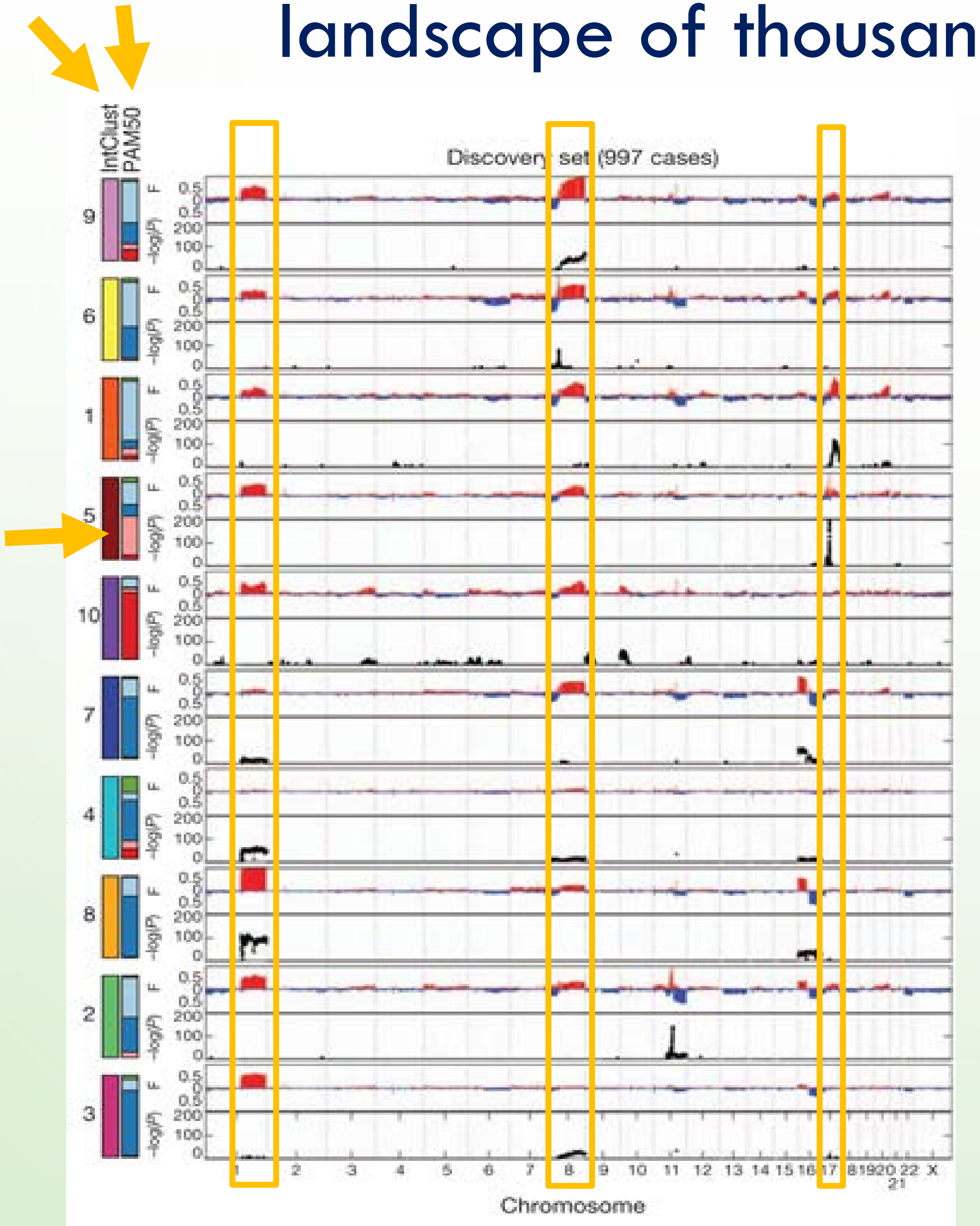


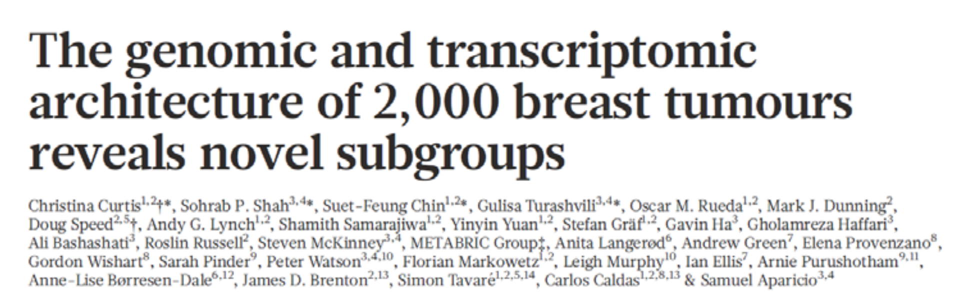
	PIK3CA	TP53	MAP3K1	MAP2K4	GATA3	MLL3	CDH1	PTEN	PIK3R1	AKT1	RUNX1	CBFB	TBX3	NCOR1	CTCF	FOXA1	SF3B1	CDKN1B	RB1	AFF2	NF1	PTPN22	PTPRD
All	36%	37%	8%	4%	11%	7%	7%	3%	3%	2%	4%	2%	3%	3%	3%	2%	2%	1%	2%	3%	3%	1%	2%
Luminal A	45%	12%	13%	7%	14%	8%	9%	4%	0.4%	4%	5%	2%	3%	5%	4%	2%	3%	1%	0.4%	1%	2%	0.4%	2%
Luminal B	29%	29%	5%	2%	15%	6%	5%	4%	2%	2%	2%	2%	4%	2%	2%	2%	0%	1%	3%	2%	4%	2%	4%
HER2-enriched	39%	72%	4%	2%	2%	7%	5%	2%	4%	2%	4%	2%	0%	0%	2%	2%	4%	2%	0%	5%	0%	5%	4%
Basal-like	9%	80%	0%	0%	2%	5%	0%	1%	0%	0%	0%	0%	1%	2%	1%	0%	1%	0%	4%	4%	2%	0%	1%

Percentages of cases with mutation by expression subtype

DC Koboldt *et al. Nature*, 1-10 (2012) doi:10.1038/nature11412

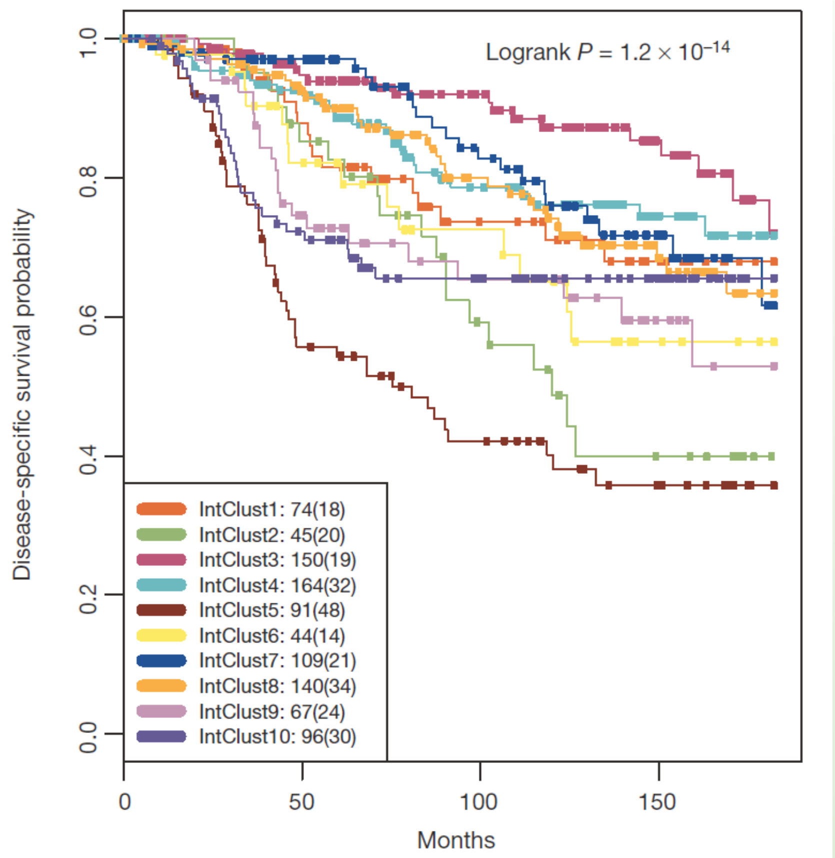
Clustering on Copy Number Changes and transcriptional landscape of thousands of tumours (*IntClust*)



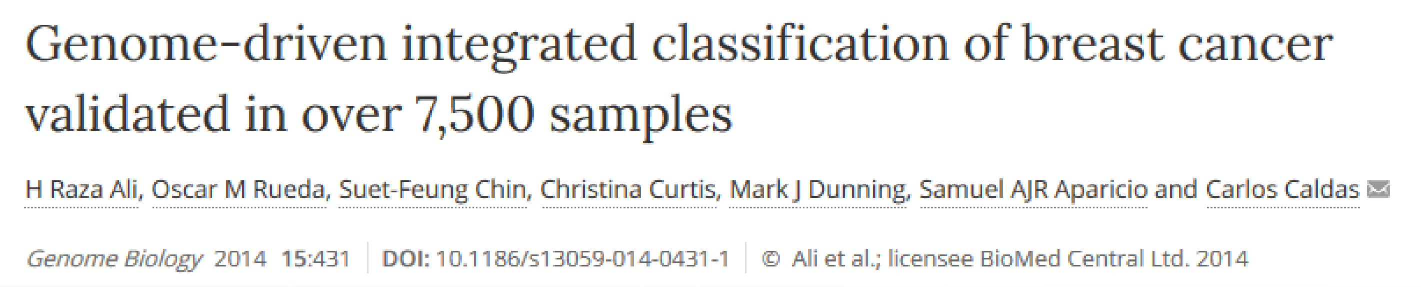


The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups

Christina Curtis[1,2]†*, Sohrab P. Shah[3,4]*, Suet-Feung Chin[1,2]*, Gulisa Turashvili[3,4]*, Oscar M. Rueda[1,2], Mark J. Dunning[2], Doug Speed[2,5]†, Andy G. Lynch[1,2], Shamith Samarajiwa[1,2], Yinyin Yuan[1,2], Stefan Gräf[1,2], Gavin Ha[3], Gholamreza Haffari[3], Ali Bashashati[3], Roslin Russell[2], Steven McKinney[3,4], METABRIC Group‡, Anita Langerød[6], Andrew Green[7], Elena Provenzano[8], Gordon Wishart[8], Sarah Pinder[9], Peter Watson[3,4,10], Florian Markowetz[1,2], Leigh Murphy[10], Ian Ellis[7], Arnie Purushotham[9,11], Anne-Lise Børresen-Dale[6,12], James D. Brenton[2,13], Simon Tavaré[1,2,5,14], Carlos Caldas[1,2,8,13] & Samuel Aparicio[3,4]



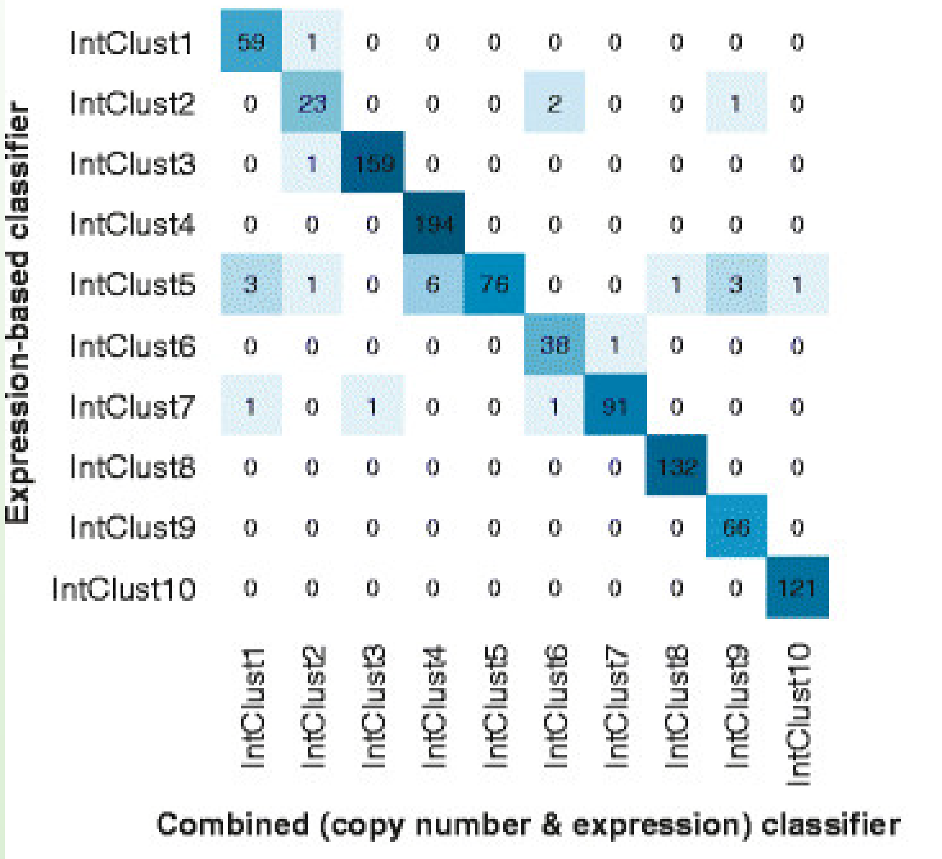
Gene expression-based approach for classifying breast tumors into the ten IntClust subtypes



Genome-driven integrated classification of breast cancer validated in over 7,500 samples

H Raza Ali, Oscar M Rueda, Suet-Feung Chin, Christina Curtis, Mark J Dunning, Samuel AJR Aparicio and Carlos Caldas

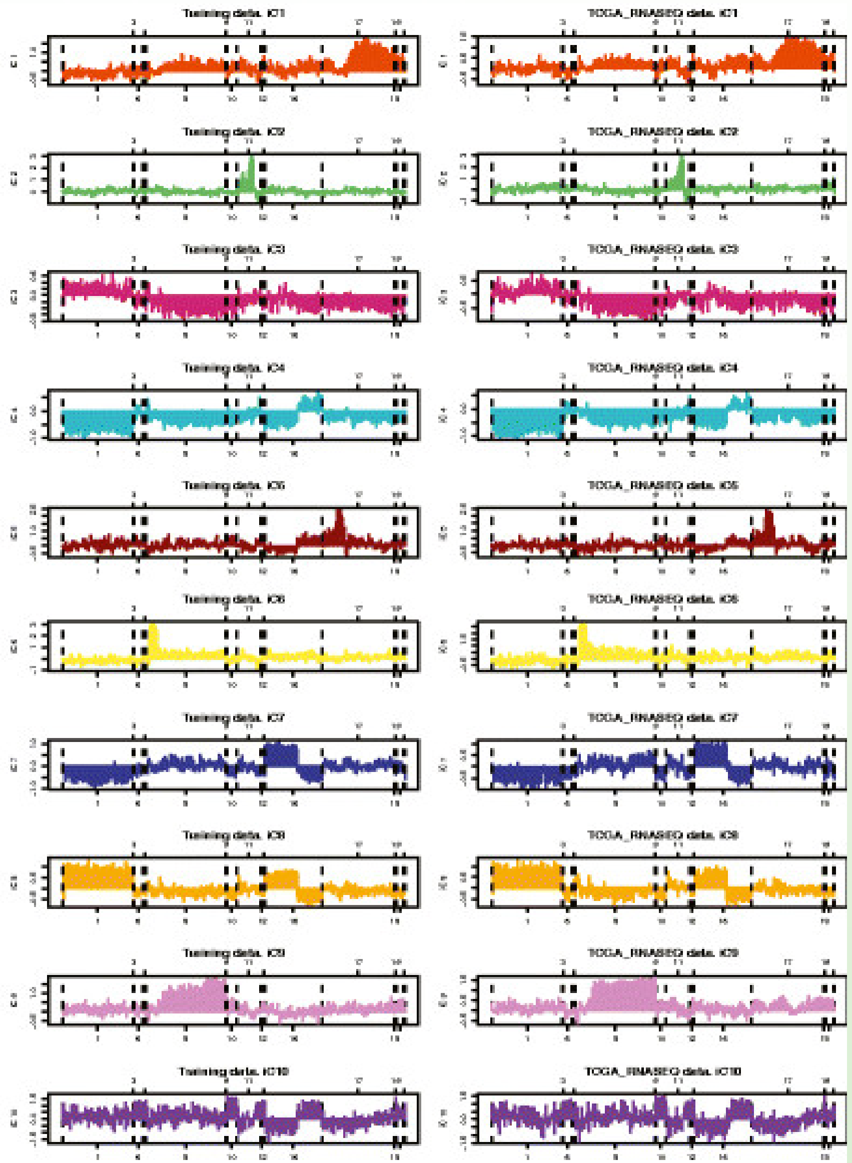
Genome Biology 2014 15:431 | DOI: 10.1186/s13059-014-0431-1 | © Ali et al.; licensee BioMed Central Ltd. 2014



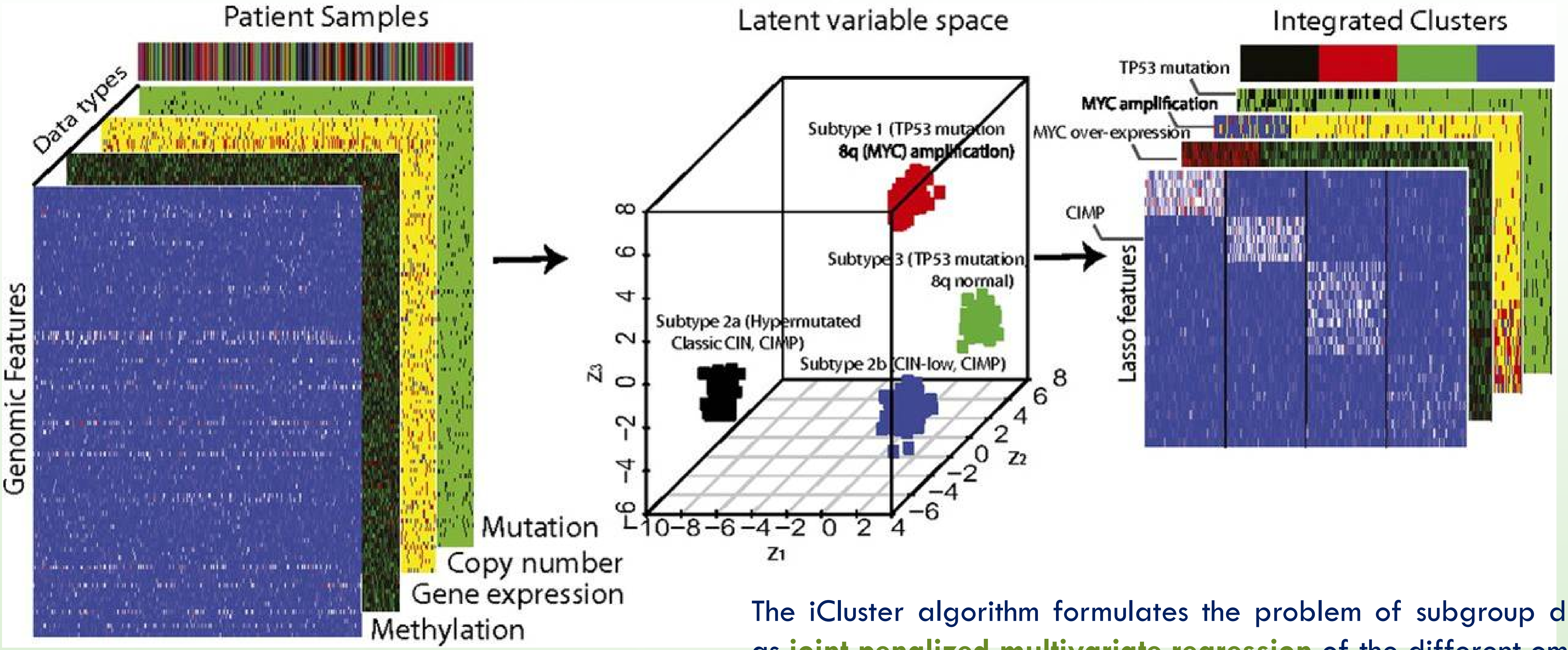
	IntClust1	IntClust2	IntClust3	IntClust4	IntClust5	IntClust6	IntClust7	IntClust8	IntClust9	IntClust10
IntClust1	59	1	0	0	0	0	0	0	0	0
IntClust2	0	23	0	0	0	2	0	0	1	0
IntClust3	0	1	159	0	0	0	0	0	0	0
IntClust4	0	0	0	194	0	0	0	0	0	0
IntClust5	3	1	0	6	76	0	0	1	3	1
IntClust6	0	0	0	0	0	38	1	0	0	0
IntClust7	1	0	1	0	0	1	91	0	0	0
IntClust8	0	0	0	0	0	0	0	132	0	0
IntClust9	0	0	0	0	0	0	0	0	66	0
IntClust10	0	0	0	0	0	0	0	0	0	121

Expression-based classifier

Combined (copy number & expression) classifier



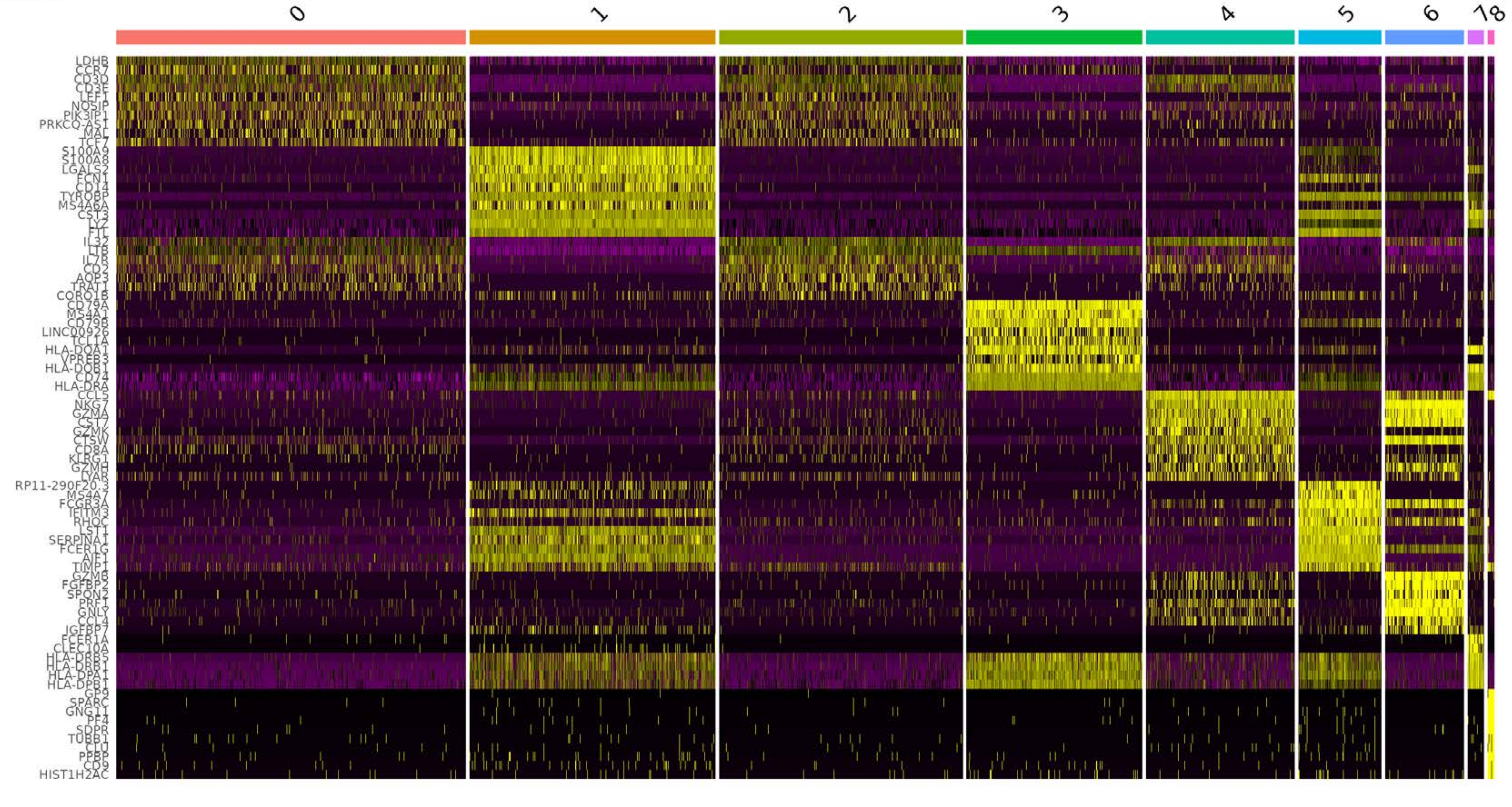
EXAMPLE OF INTEGRATED CLUSTERING METHOD: ICLUSTER



https://sites.google.com/site/ronglais/icluster

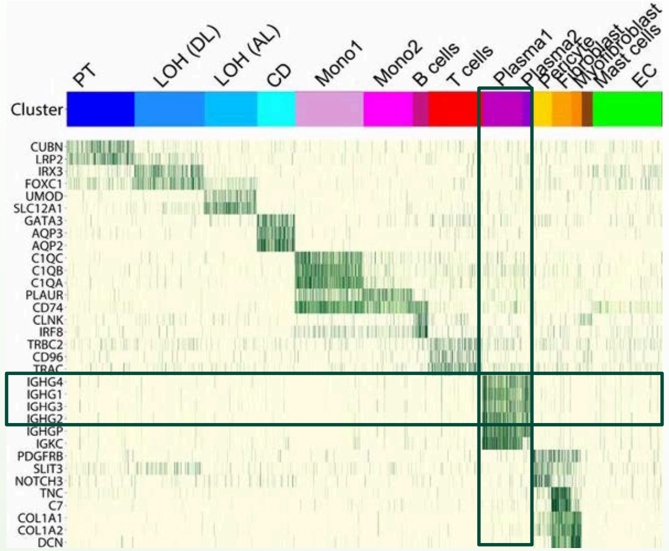
The iCluster algorithm formulates the problem of subgroup discovery as **joint penalized multivariate regression** of the different omics data types with **reference to a set of common latent variables**, which represent the underlying tumor subtypes Gaussian joint latent variable model.

CLUSTERING SINGLE CELLS RNASEQ

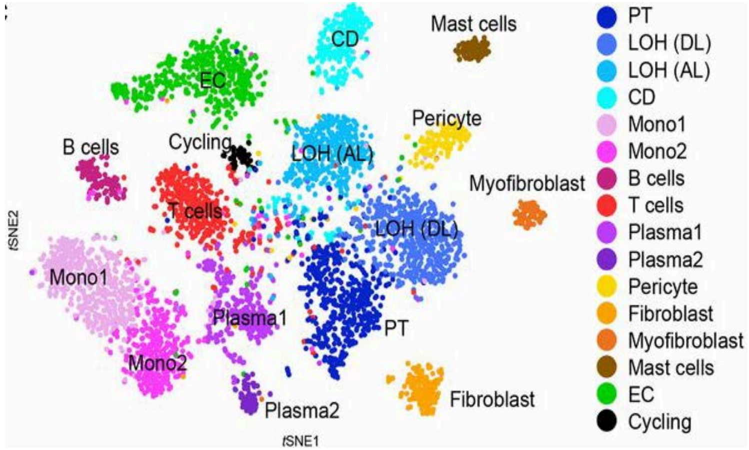


DISCOVERY OF DIFFERENT CELL PHENOTYPES

Identifying cell-type marker genes



Discovering sample heterogeneity



Uncovering tissue dynamics

Images from: Wu et al (2018) JASN and Farbehi et al (2019) Elife