RNA-Seq clustering and differential expression

Reanne Bowlby

February 17, 2016

Where to find TCGA data

Hierarchical clustering using pheatmap

Consensus clustering using ConsensusClusterPlus

Differential gene expression using SAM

What you need for this tutorial

R packages

- pheatmap
- ConsensusClusterPlus
- samr

Data

- CESC.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.data.txt.formatted
- CESC_covariates.formatted.subset.txt

Where to find TCGA data

CGHub – bams https://cghub.ucsc.edu/

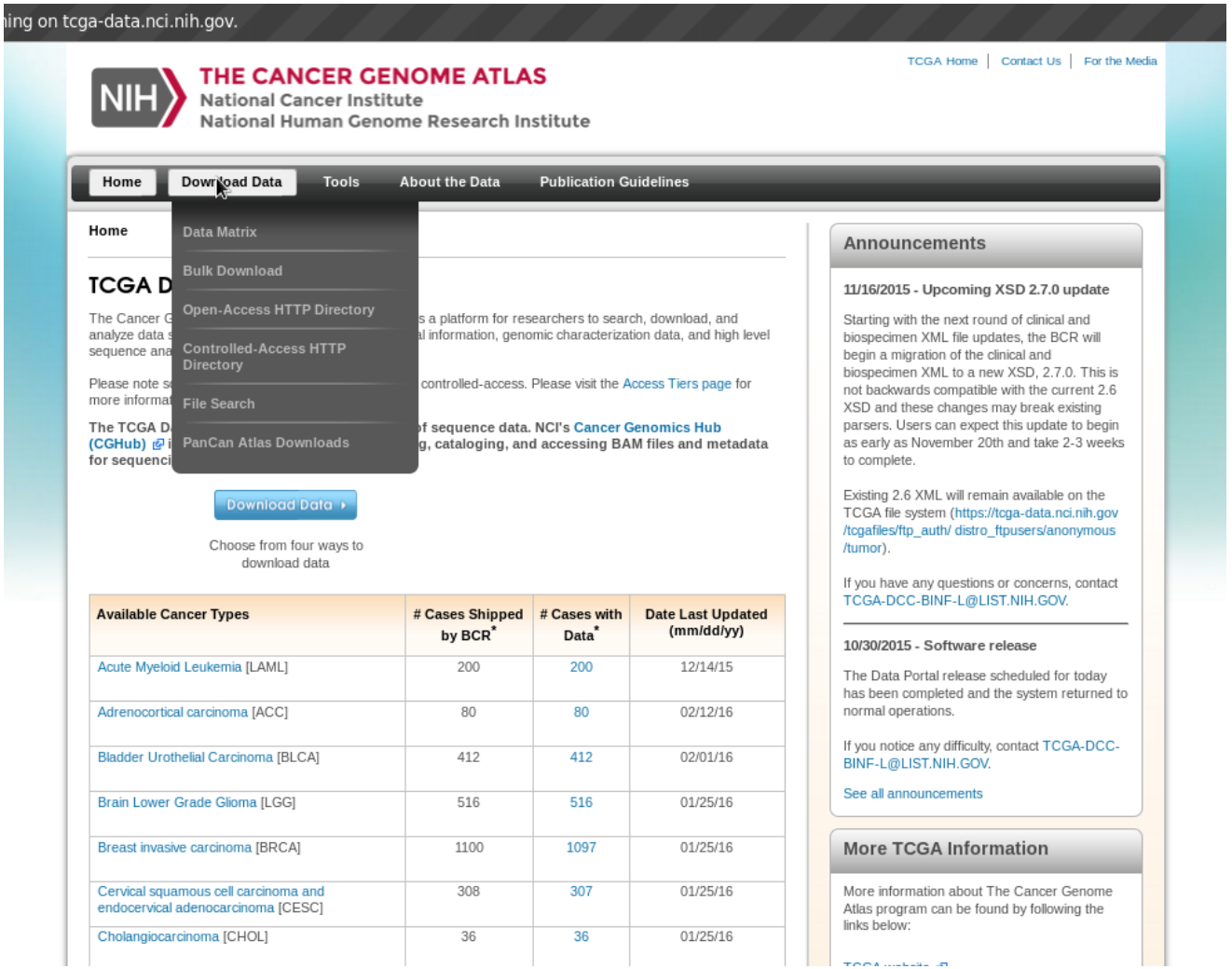
Data coordination center – processed data https://tcga-data.nci.nih.gov/tcga/

Firebrowse - http://firebrowse.org/

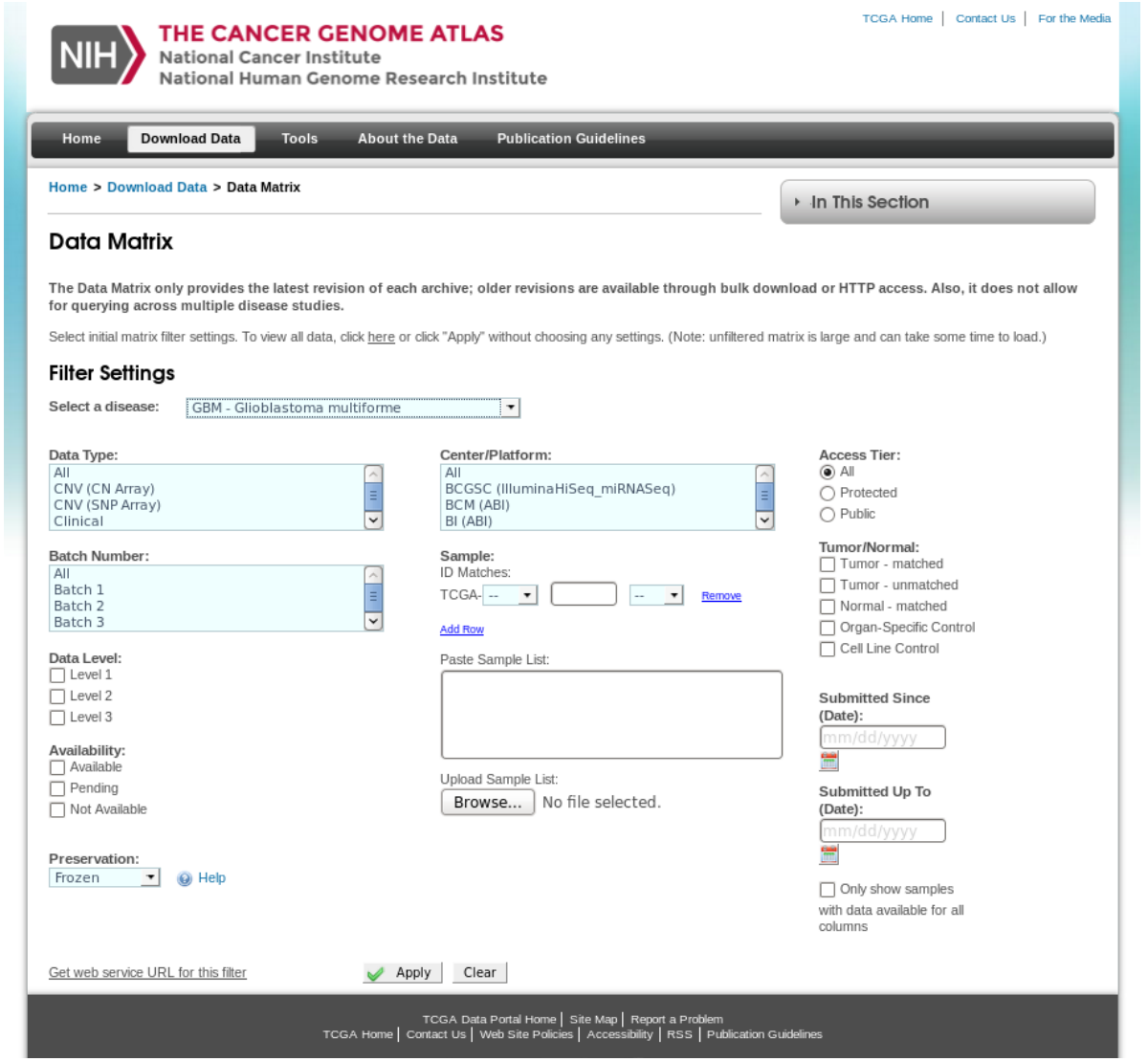
Cbioportal - http://www.cbioportal.org/

Regulome Explorer - http://explorer.cancerregulome.org/all_pairs/

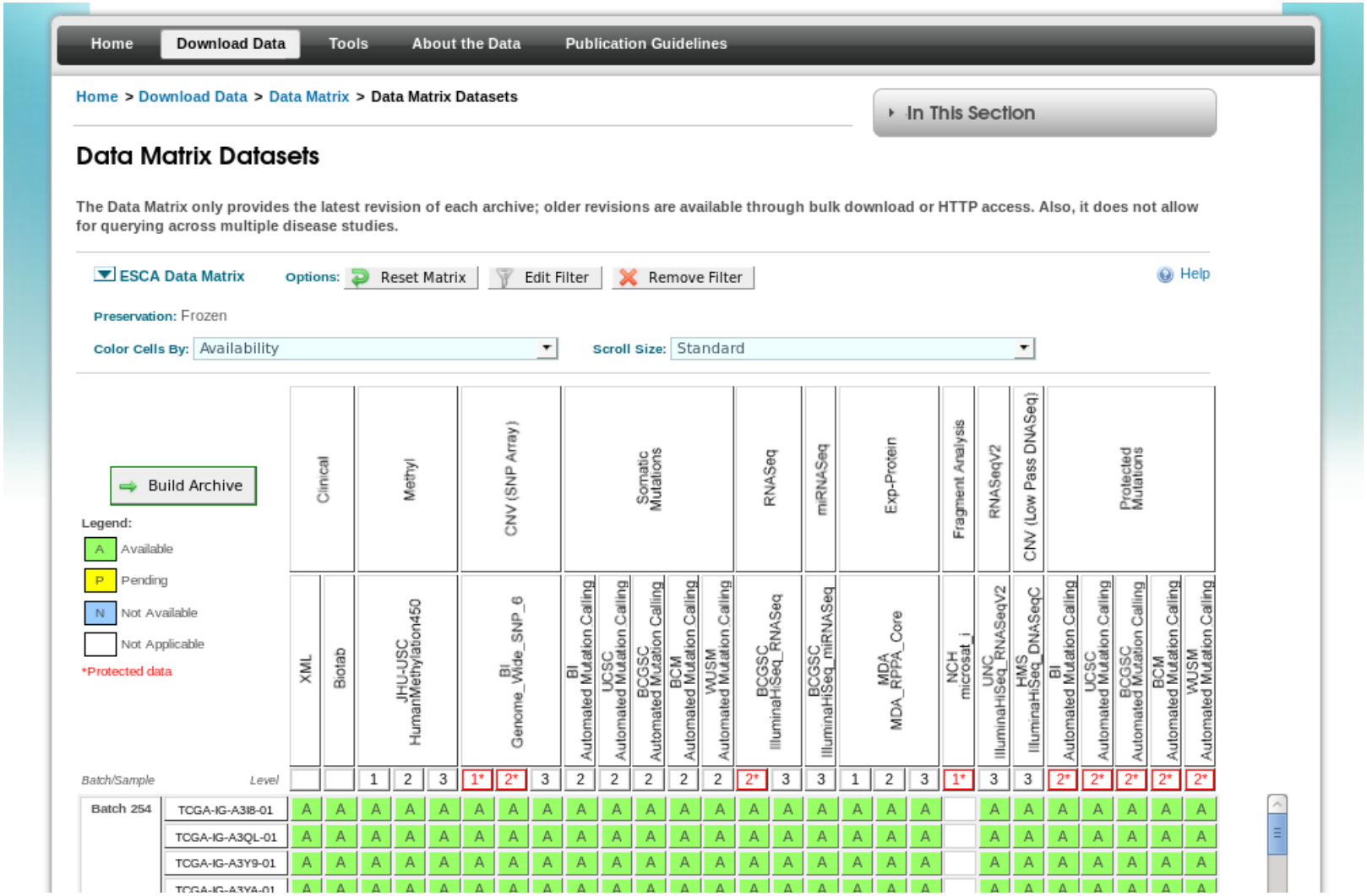
TCGA Data Coordination Center (DCC)



TCGA Data Coordination Center (DCC)



TCGA Data Coordination Center (DCC)



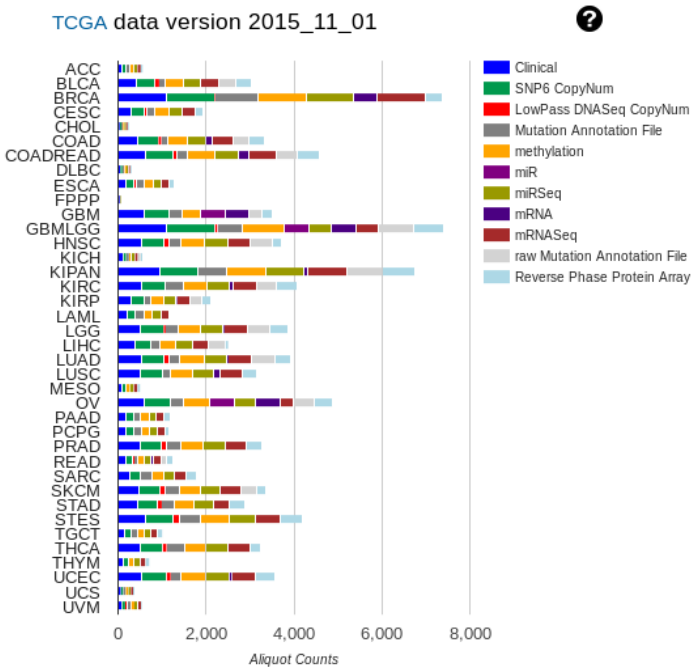
Firebrowse

HOME BROAD GDAC WEB API TUTORIAL RELEASE NOTES ANALYSES GRAPH FAQ CONTACT

View Expression Profile | Enter gene name | Enter cohort abbrev | View Analysis Profile

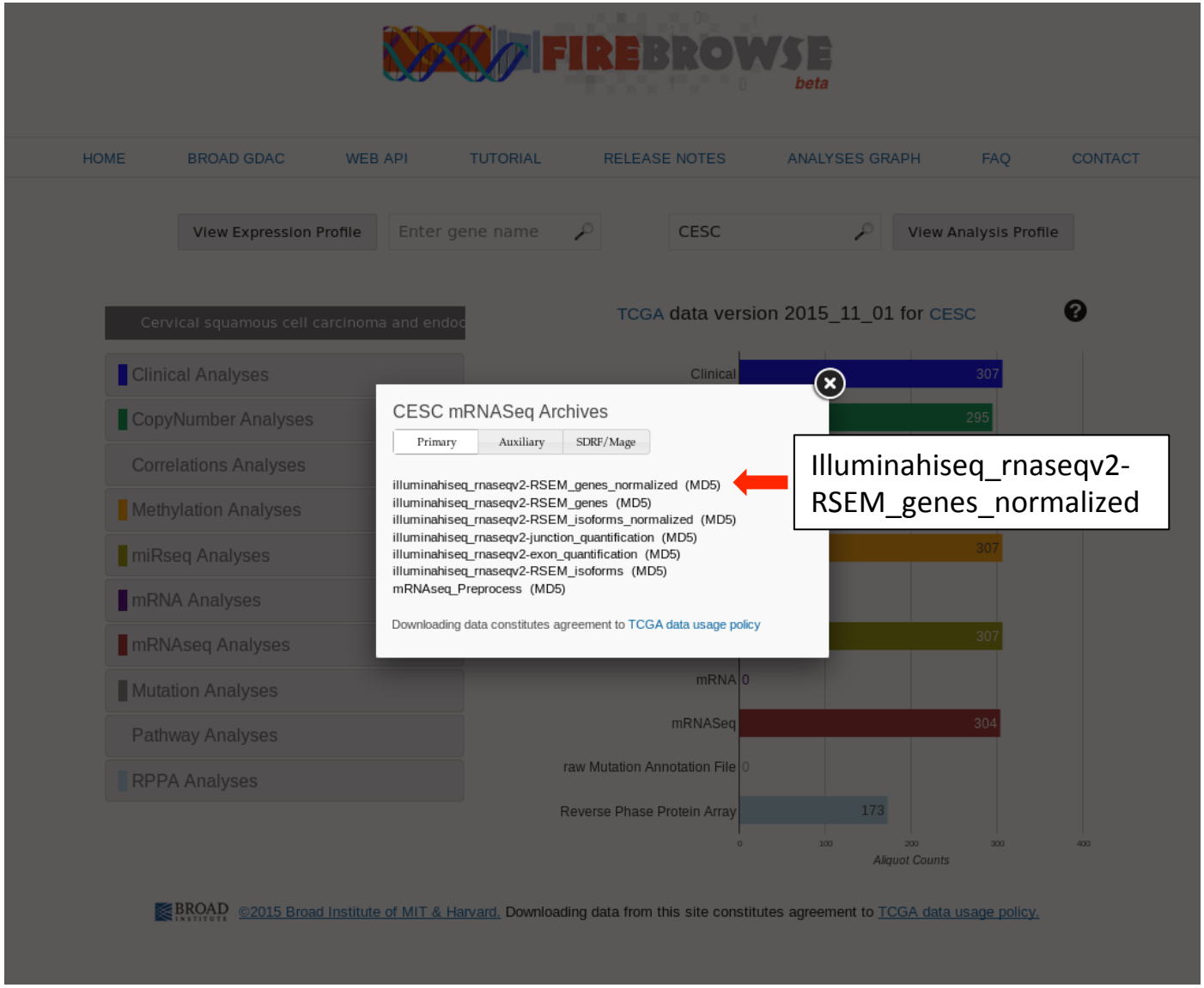
SELECT COHORT

- Clinical Analyses
- CopyNumber Analyses
- Correlations Analyses
- miR Analyses
- miRseq Analyses
- mRNA Analyses
- mRNAseq Analyses
- Mutation Analyses
- Pathway Analyses
- RPPA Analyses



http://firebrowse.org/

Firebrowse


FIREBROWSE
beta
HOME
BROAD GDAC
WEB API
TUTORIAL
RELEASE NOTES
ANALYSES GRAPH
FAQ
CONTACT
View Expression Profile
Enter gene name
CESC
View Analysis Profile
Cervical squamous cell carcinoma and endoc
TCGA data version 2015_11_01 for CESC
Clinical Analyses
CopyNumber Analyses
Correlations Analyses
Methylation Analyses
miRseq Analyses
mRNA Analyses
mRNAseq Analyses
Mutation Analyses
Pathway Analyses
RPPA Analyses
CESC mRNASeq Archives
Primary
Auxiliary
SDRF/Mage
illuminahiseq_rnaseqv2-RSEM_genes_normalized (MD5)
illuminahiseq_rnaseqv2-RSEM_genes (MD5)
illuminahiseq_rnaseqv2-RSEM_isoforms_normalized (MD5)
illuminahiseq_rnaseqv2-junction_quantification (MD5)
illuminahiseq_rnaseqv2-exon_quantification (MD5)
illuminahiseq_rnaseqv2-RSEM_isoforms (MD5)
mRNAseq_Preprocess (MD5)
Downloading data constitutes agreement to TCGA data usage policy
Illuminahiseq_rnaseqv2-RSEM_genes_normalized
Clinical
307
295
307
307
mRNA 0
mRNASeq
304
raw Mutation Annotation File 0
Reverse Phase Protein Array
173
0
100
200
300
400
Aliquot Counts
BROAD
©2015 Broad Institute of MIT & Harvard. Downloading data from this site constitutes agreement to TCGA data usage policy.

Illuminahiseq_rnaseqv2-RSEM_genes_normalized

Hybridization REF	TCGA-2W-A8YY-01A-11R-A37O-07	TCGA-4J-AA1J-01A-21R-A38B-07	TCGA-BI-A0VR-01A-11R-A10U-07	TCGA-BI-A0VS-01A-11R-A10U-07	TCGA-BI-A20A-01A-11R-A14Y-07	TCGA-C5-A0TN-01A-21R-A14Y-07	…
~~gene_id~~	~~normalized_count~~	~~normalized_count~~	~~normalized_count~~	~~normalized_count~~	~~normalized_count~~	~~normalized_count~~	
?|100130426	0	0	0	0	0	0	
?|100133144	8.6373	7.6079	24.022	25.8961	14.303	5.5414	
?|100134869	27.4857	29.1286	31.7792	24.7455	18.7904	11.8219	
?|10357	118.5443	180.7091	165.6901	210.1524	303.6883	330.4378	
?|10431	1072.961	1014.951	954.1212	716.2174	785.9684	943.5694	
?|136542	0	0	0	0	0	0	
?|155060	169.1702	265.2713	119.8546	119.5418	194.5892	46.1367	
?|26823	0.7153	0.8543	0	0.689	0.6619	0	
…							

RSEM genes normalized – quantile normalization
divide by the 75th percentile and multiply by 1000

Clustering RNA-Seq data

Hierarchical clustering – pheatmap

Expression matrix

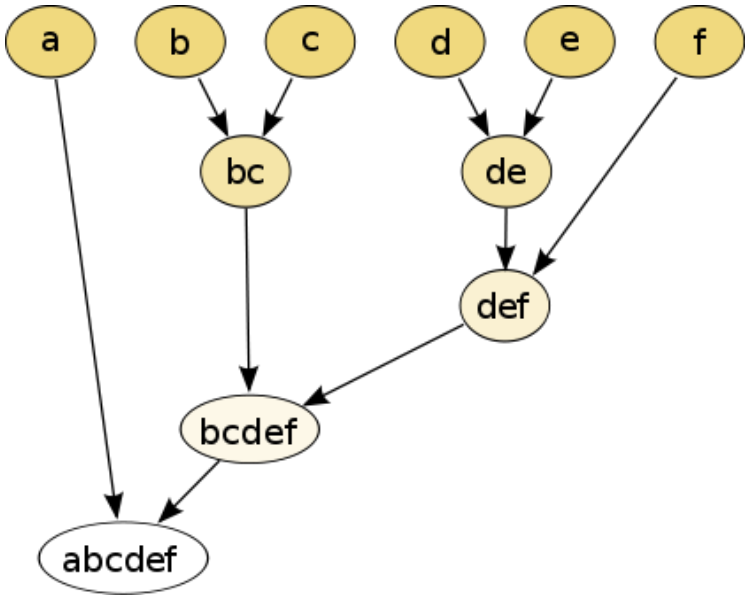
Samples

Genes

Distance/correlation matrix

Samples

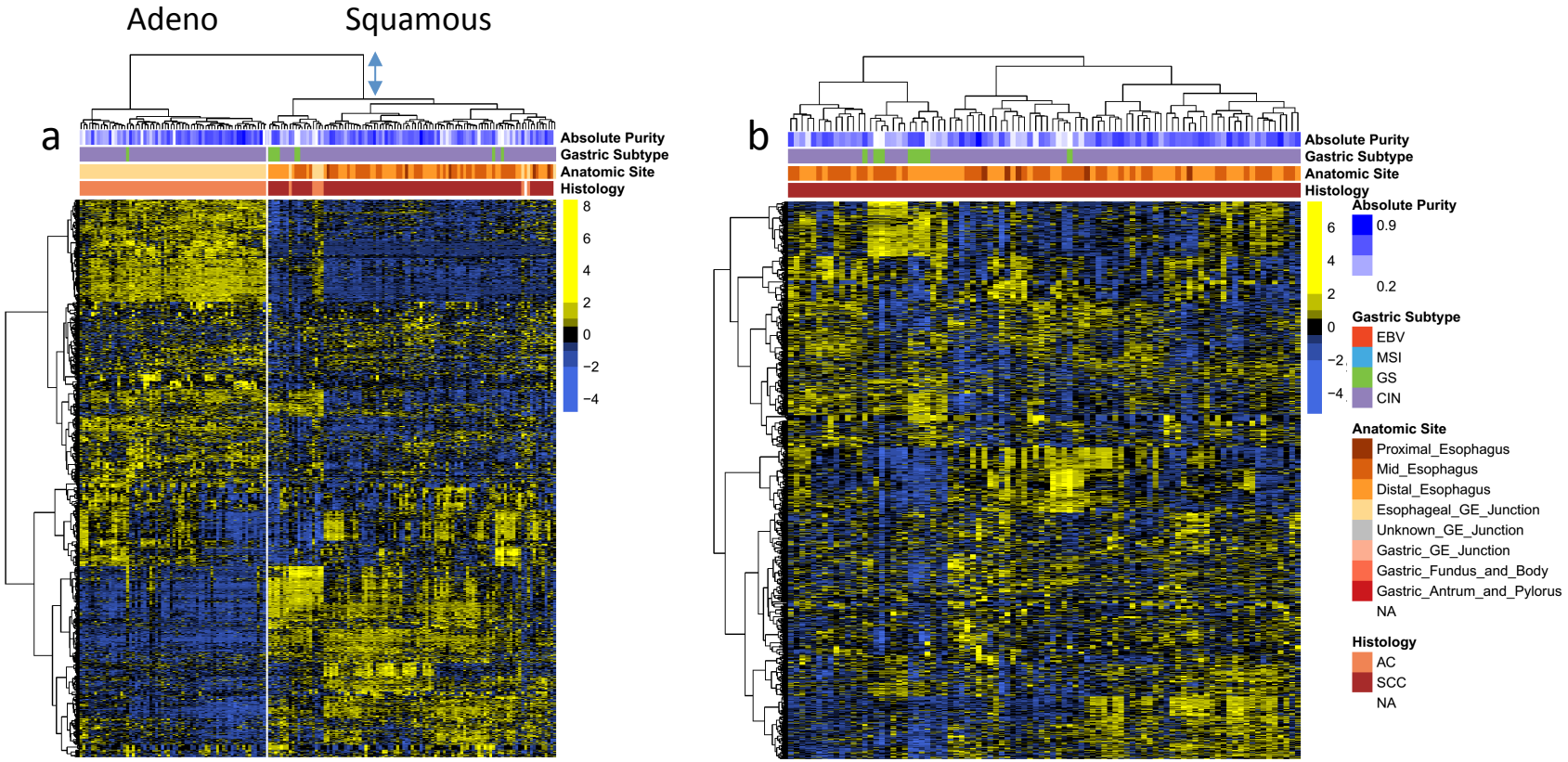
Samples



R

Example1: clustering TCGA ESCA

Use the dendrogram heights to decide on the number of clusters
Find subtypes within a subtype

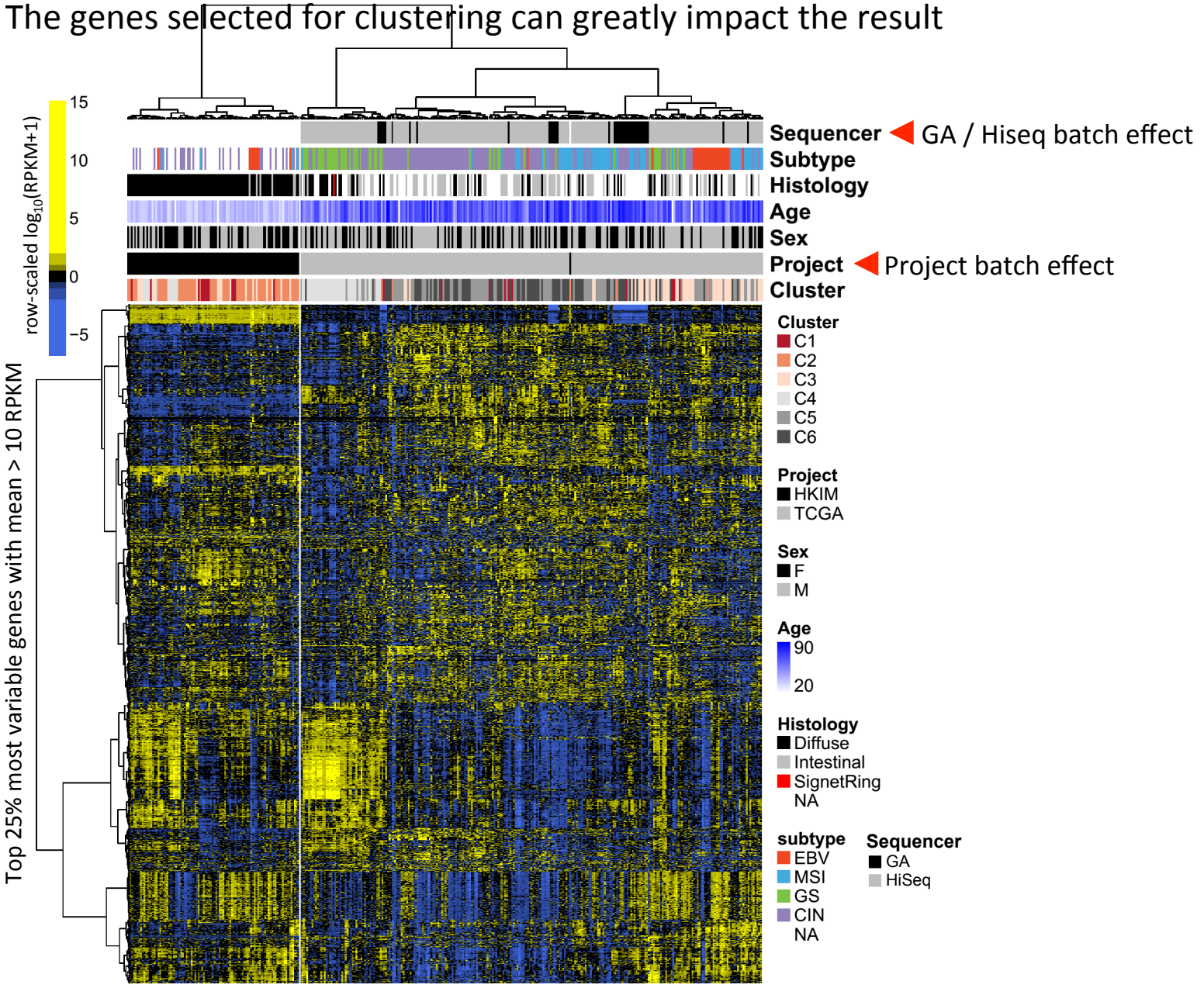


Esophageal Carcinoma

Esophageal Squamous

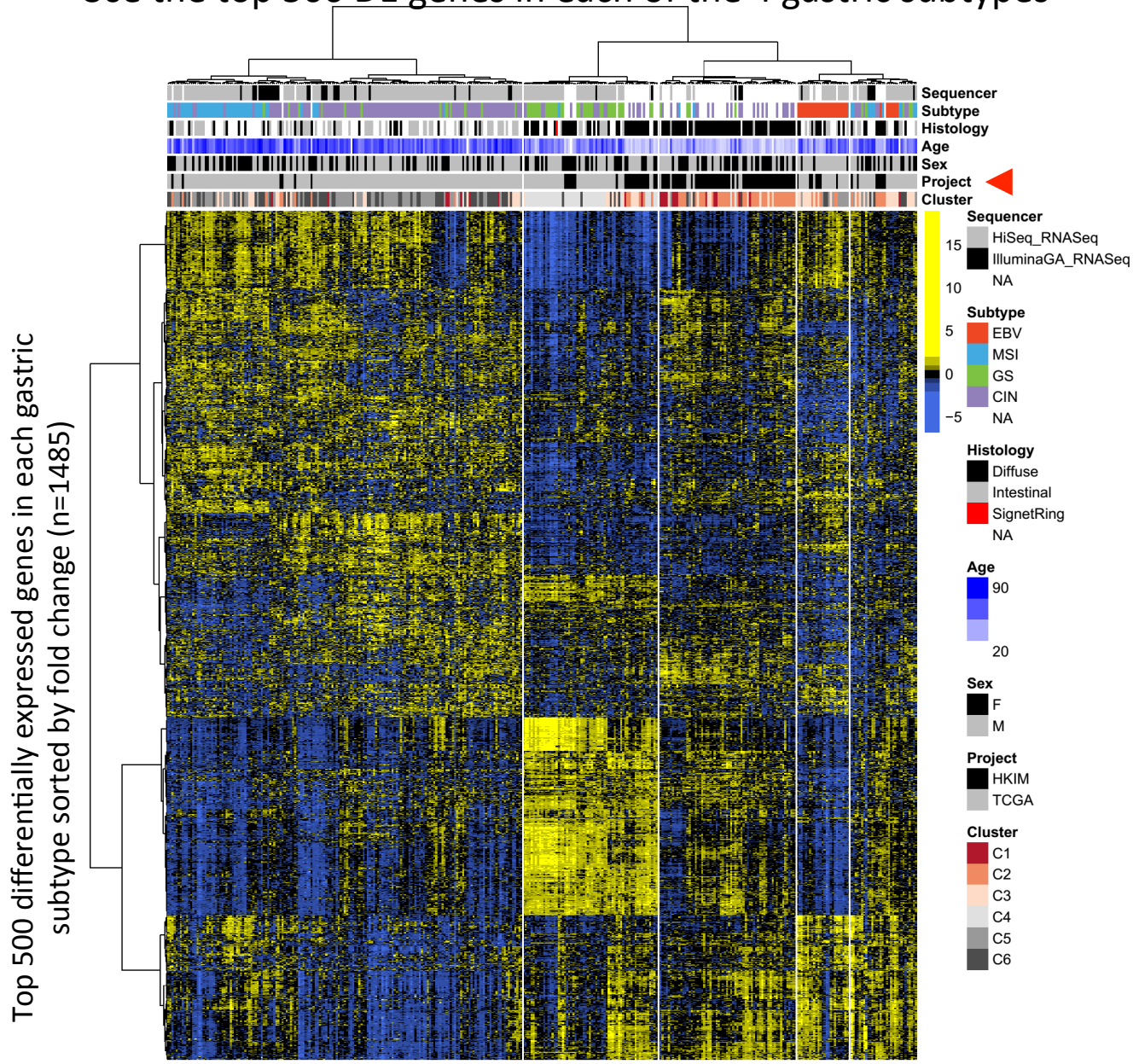
Example2: clustering TCGA STAD + an external cohort

The genes selected for clustering can greatly impact the result



Example2: clustering TCGA STAD + an external cohort

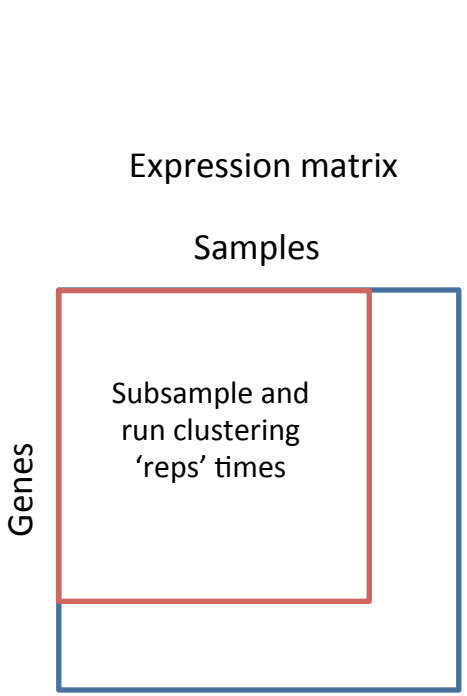
Use the top 500 DE genes in each of the 4 gastric subtypes

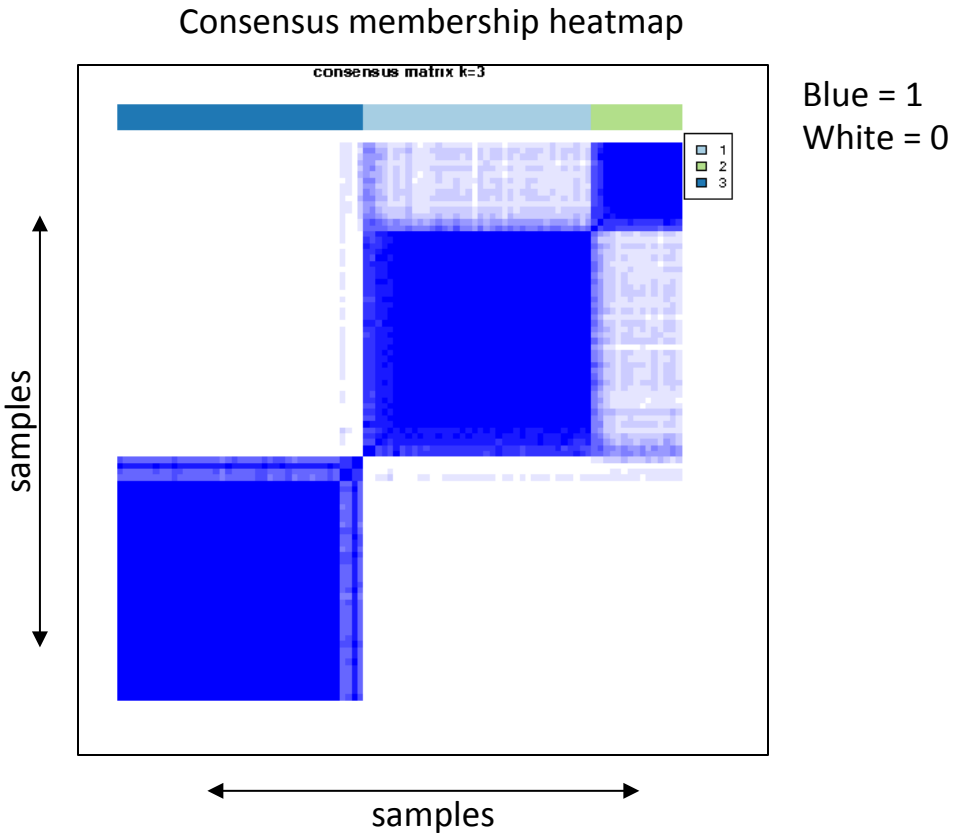


Clustering RNA-Seq data

Consenesus clustering – ConsensusClusterPlus

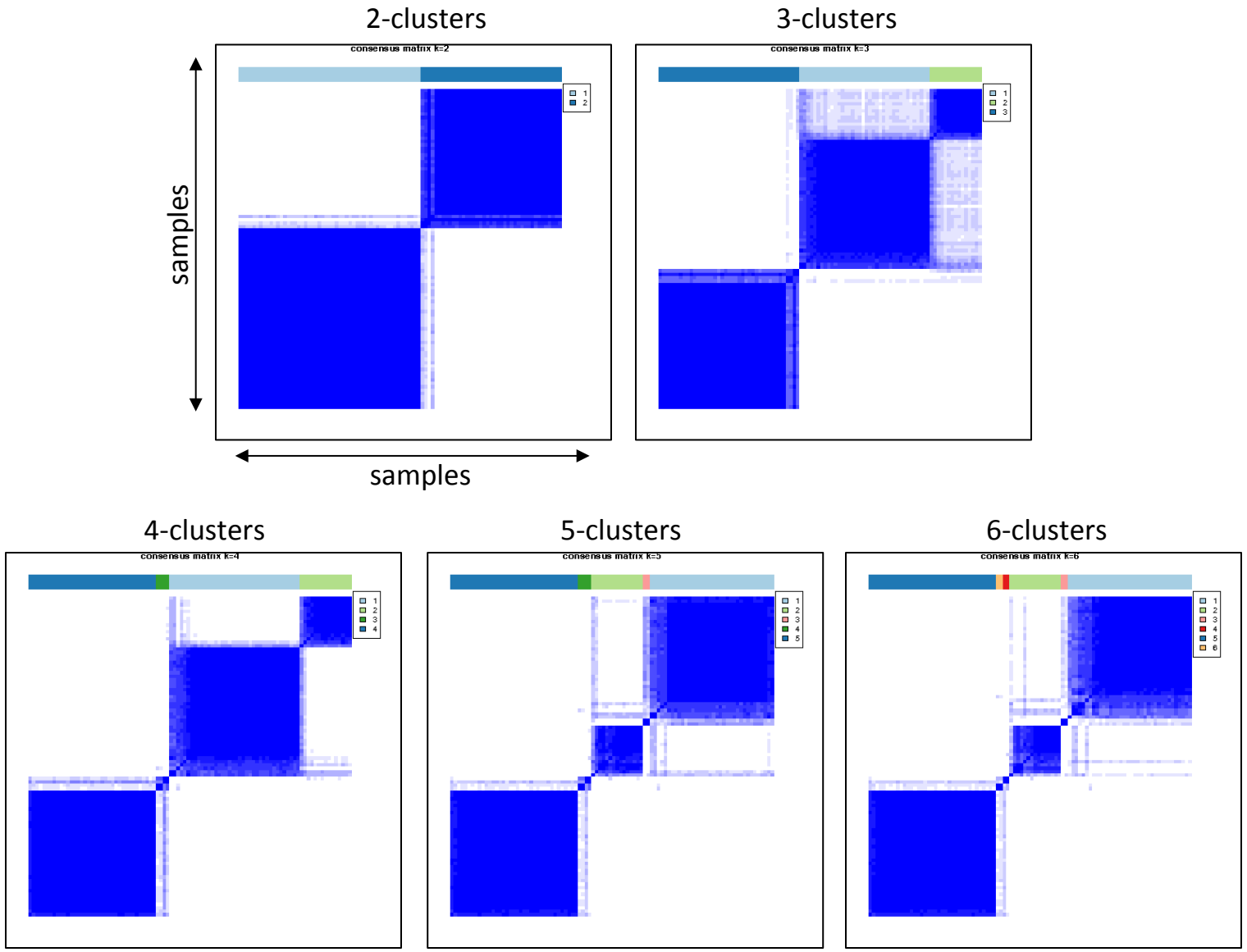
The consensus membership heatmap shows the proportion of the time two samples occupy the same cluster out of the number of times they occur in the same subsample.



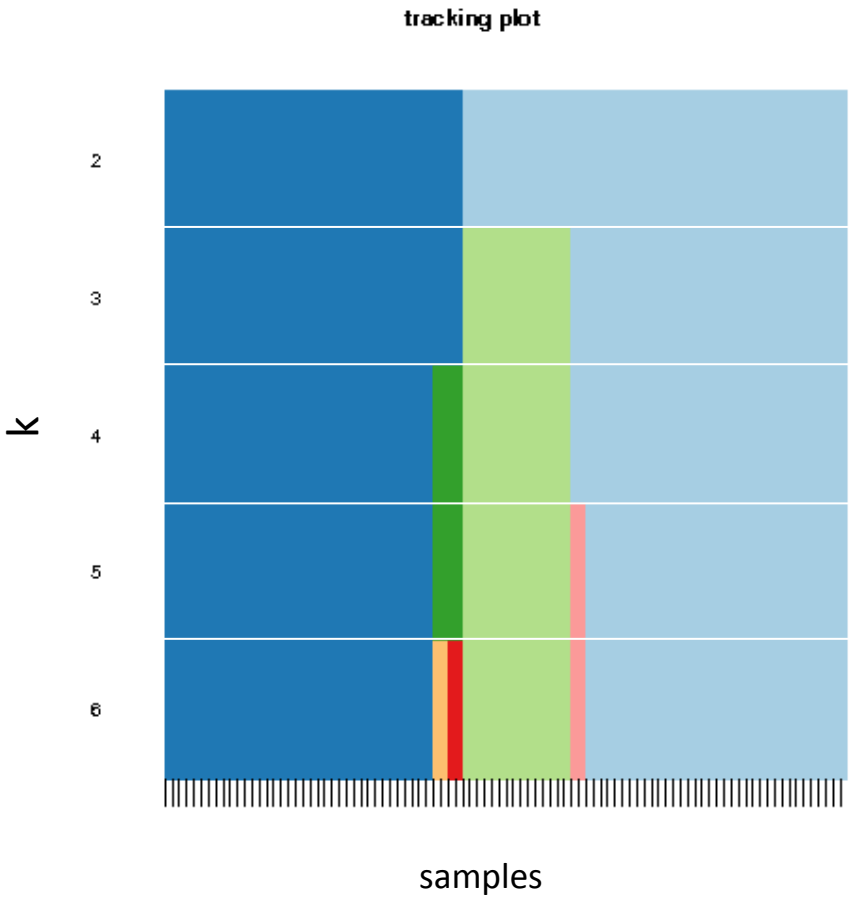


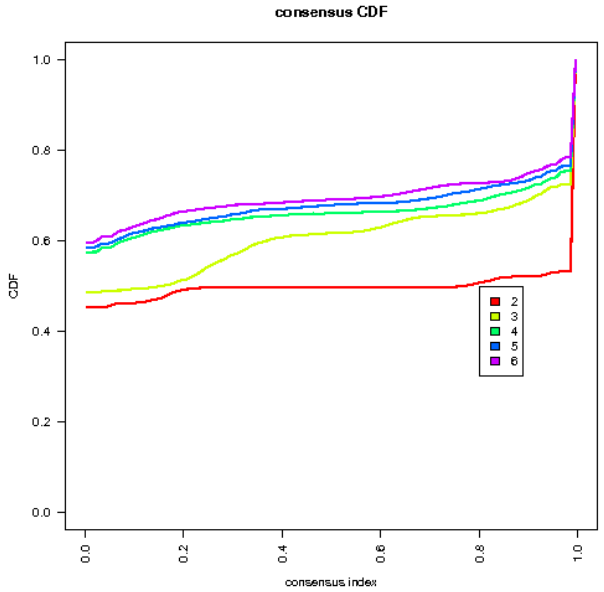
R

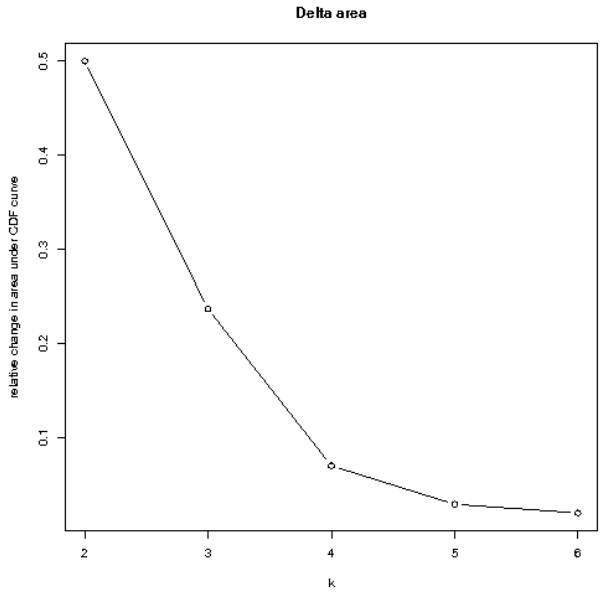
Consensus Cluster Plus – consensus membership heatmaps



Consensus cluster plus output







R

The problem with clustering algorithms is that they ALWAYS cluster the data

Differential Expression

SAMR

Deseq

EdgeR

R

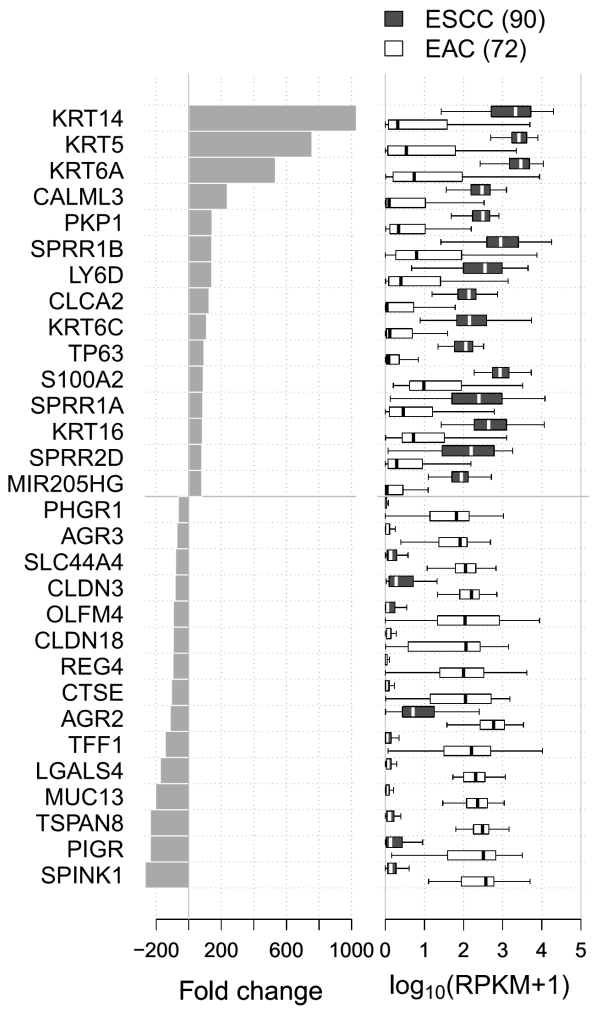
Differential Expression

You can filter the genes by:

- Mean expression
- Fold change
- Wilcoxon test

Use for pathway analysis

- ingenuity
- gene enrichment
- DAVID



Questions?