

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_illuminahisq_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)

The screenshot shows the homepage of the TCGA Data Coordination Center (DCC). At the top, there is a banner with the text "Log in on tcga-data.nci.nih.gov." Below the banner, the NIH logo and the "THE CANCER GENOME ATLAS" logo are displayed. A navigation bar with links to "Home", "Download Data", "Tools", "About the Data", and "Publication Guidelines" is present. The "Download Data" link is highlighted with a dropdown menu showing options like "Data Matrix", "Bulk Download", "Open-Access HTTP Directory", "Controlled-Access HTTP Directory", "File Search", and "PanCan Atlas Downloads". A large blue button labeled "Download Data" is located below the dropdown menu. To the right of the dropdown, there is a main content area with a heading "TCGA Data Portal" and a brief description of the portal's purpose. Below this, there is a table titled "Available Cancer Types" listing various cancer types with their corresponding case counts and last update dates. To the right of the main content area, there is a sidebar titled "Announcements" containing two entries: "11/16/2015 - Upcoming XSD 2.7.0 update" and "10/30/2015 - Software release". Both announcements provide details about software updates and operational changes. At the bottom of the sidebar, there is a link to "See all announcements".

Log in on tcga-data.nci.nih.gov.

NIH THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

Home Download Data Tools About the Data Publication Guidelines

TCGA Data Portal

The Cancer Genome Atlas (TCGA) is a platform for researchers to search, download, and analyze data sets from genome-wide sequence analysis.

Please note some data is controlled-access. Please visit the [Access Tiers](#) page for more information.

The TCGA Data Portal (TCGA-DP) is a system for sequencing, cataloging, and accessing BAM files and metadata for sequencing projects.

[Download Data](#)

Choose from four ways to download data

Available Cancer Types	# Cases Shipped by BCR*	# Cases with Data*	Date Last Updated (mm/dd/yy)
Acute Myeloid Leukemia [LAML]	200	200	12/14/15
Adrenocortical carcinoma [ACC]	80	80	02/12/16
Bladder Urothelial Carcinoma [BLCA]	412	412	02/01/16
Brain Lower Grade Glioma [LGG]	516	516	01/25/16
Breast invasive carcinoma [BRCA]	1100	1097	01/25/16
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	308	307	01/25/16
Cholangiocarcinoma [CHOL]	36	36	01/25/16

Announcements

11/16/2015 - Upcoming XSD 2.7.0 update

Starting with the next round of clinical and biospecimen XML file updates, the BCR will begin a migration of the clinical and biospecimen XML to a new XSD, 2.7.0. This is not backwards compatible with the current 2.6 XSD and these changes may break existing parsers. Users can expect this update to begin as early as November 20th and take 2-3 weeks to complete.

Existing 2.6 XML will remain available on the TCGA file system (https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftusers/anonymous/tumor).

If you have any questions or concerns, contact TCGA-DCC-BINL-L@LIST.NIH.GOV.

10/30/2015 - Software release

The Data Portal release scheduled for today has been completed and the system returned to normal operations.

If you notice any difficulty, contact TCGA-DCC-BINL-L@LIST.NIH.GOV.

[See all announcements](#)

More TCGA Information

More information about The Cancer Genome Atlas program can be found by following the links below:

[TCGA Home](#) | [Contact Us](#) | [For the Media](#)

TCGA Data Coordination Center (DCC)

 THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

TCGA Home | Contact Us | For the Media

Home Download Data Tools About the Data Publication Guidelines

Home > Download Data > Data Matrix

In This Section

Data Matrix

The Data Matrix only provides the latest revision of each archive; older revisions are available through bulk download or HTTP access. Also, it does not allow for querying across multiple disease studies.

Select initial matrix filter settings. To view all data, click [here](#) or click "Apply" without choosing any settings. (Note: unfiltered matrix is large and can take some time to load.)

Filter Settings

Select a disease: GBM - Glioblastoma multiforme

Data Type:
All
CNV (CN Array)
CNV (SNP Array)
Clinical

Center/Platform:
All
BCGSC (IlluminaHiSeq_miRNASeq)
BCM (ABI)
BI (ABI)

Access Tier:
 All
 Protected
 Public

Tumor/Normal:
 Tumor - matched
 Tumor - unmatched
 Normal - matched
 Organ-Specific Control
 Cell Line Control

Batch Number:
All
Batch 1
Batch 2
Batch 3

Sample:
ID Matches:
TCGA: --

Paste Sample List:

Upload Sample List:
 No file selected.

Data Level:
 Level 1
 Level 2
 Level 3

Availability:
 Available
 Pending
 Not Available

Preservation:

Submitted Since (Date):

Submitted Up To (Date):

Only show samples with data available for all columns

Get web service URL for this filter

TCGA Data Portal Home | Site Map | Report a Problem
TCGA Home | Contact Us | Web Site Policies | Accessibility | RSS | Publication Guidelines

TCGA Data Coordination Center (DCC)

Firebrowse

 FIREBROWSE beta

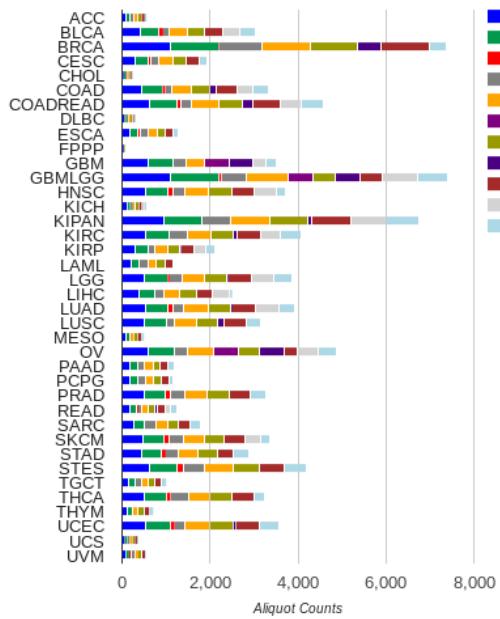
HOME BROAD GDAC WEB API TUTORIAL RELEASE NOTES ANALYSES GRAPH FAQ CONTACT

[View Expression Profile](#)   [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

TCGA data version 2015_11_01 



The chart displays the number of aliquots for each analysis type across various cancer types. The x-axis represents 'Aliquot Counts' from 0 to 8,000. The y-axis lists cancer types: ACC, BLCA, BRCA, CESC, CHOL, COAD, COADREAD, DLBC, ESCA, FPPP, GBMLGG, GBM, HNSC, KICH, KIPAN, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, STES, TGCT, THCA, THYM, UCEC, UCS, UVM. The legend indicates the analysis types: Clinical (blue), SNP6 CopyNum (green), LowPass DNaseq CopyNum (red), Mutation Annotation File (grey), methylation (orange), miR (purple), miRseq (yellow-green), mRNA (dark purple), mRNASeq (dark red), raw Mutation Annotation File (light grey), and Reverse Phase Protein Array (light blue).

Aliquot Counts

 ©2015 Broad Institute of MIT & Harvard. Downloading data from this site constitutes agreement to [TCGA data usage policy](#).

<http://firebrowse.org/>

Firebrowse

The screenshot shows the FireBrowse interface for TCGA data version 2015_11_01 for CESC. On the left, a sidebar lists various analysis types: Clinical Analyses, CopyNumber Analyses, Correlations Analyses, Methylation Analyses, miRNA Analyses, mRNA Analyses, mRNASeq Analyses, Mutation Analyses, Pathway Analyses, and RPPA Analyses. The main area displays a horizontal bar chart titled "CESC mRNASeq Archives". The chart has a legend at the top indicating "Clinical" (blue), "Primary" (green), "Auxiliary" (light green), and "SDRF/Mage" (yellow). A red arrow points from the text "Illuminahiseq_rnaseqv2-RSEM_genes_normalized (MD5)" to the corresponding blue bar, which has a value of 307. Other bars in the chart include "mRNA" (grey) with a value of 0, "mRNASeq" (dark red) with a value of 304, "raw Mutation Annotation File" (grey) with a value of 0, and "Reverse Phase Protein Array" (teal) with a value of 173. The x-axis is labeled "Aliquot Counts" and ranges from 0 to 400. A tooltip at the bottom of the chart states: "Downloading data constitutes agreement to [TCGA data usage policy](#)".

Cervical squamous cell carcinoma and endoc...

TCGA data version 2015_11_01 for CESC

Clinical

307

295

Illuminahiseq_rnaseqv2-RSEM_genes_normalized (MD5)

307

0

304

0

173

Aliquot Counts

CESC mRNASeq Archives

Primary Auxiliary SDRF/Mage

mRNA Analyses

mRNASeq Analyses

Pathway Analyses

RPPA Analyses

Download data constitutes agreement to [TCGA data usage policy](#)

BROAD INSTITUTE ©2015 Broad Institute of MIT & Harvard. Downloading data from this site constitutes agreement to [TCGA data usage policy](#).

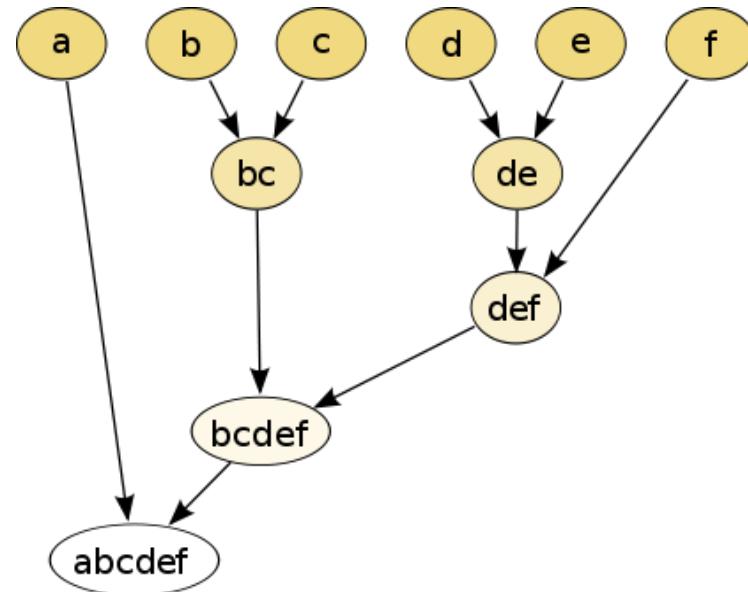
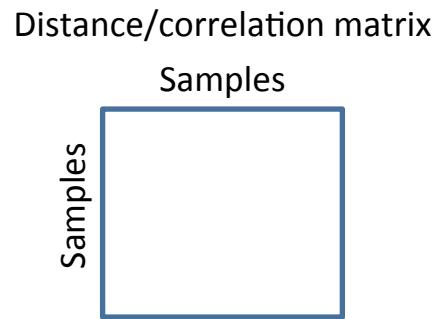
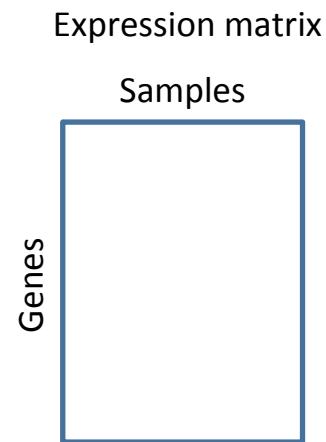
Illuminahiseq_rnaseqv2-RSEM_genes_normalized

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

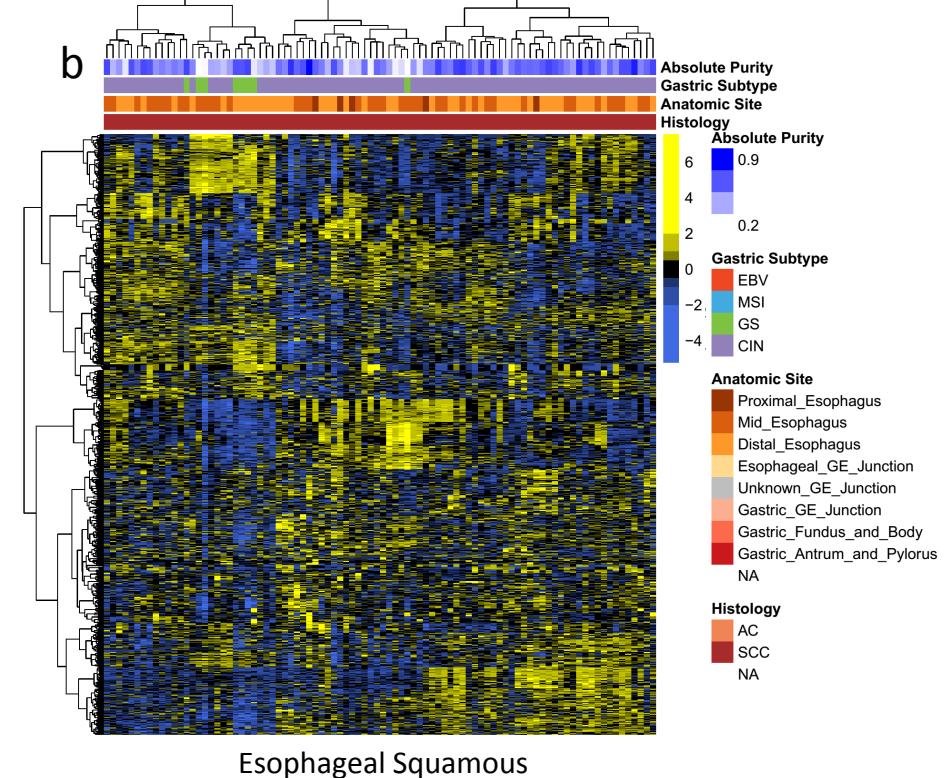
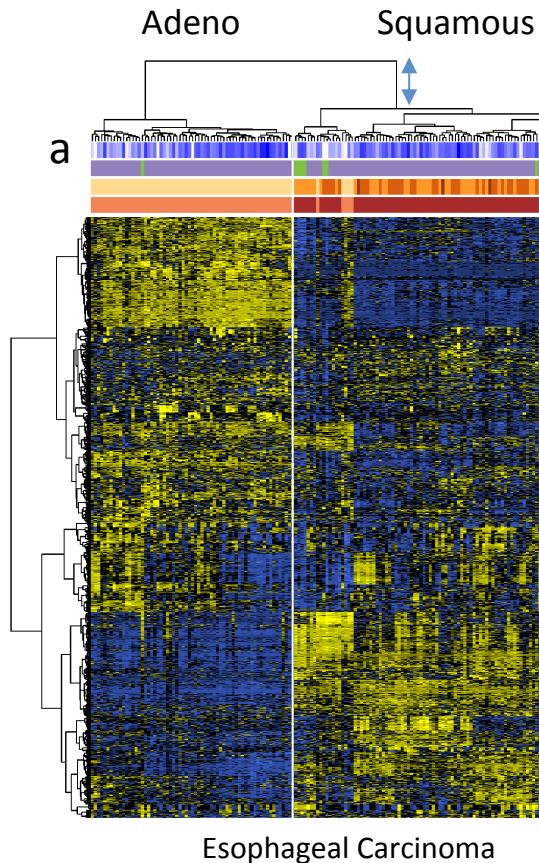
Hierarchical clustering – pheatmap



R

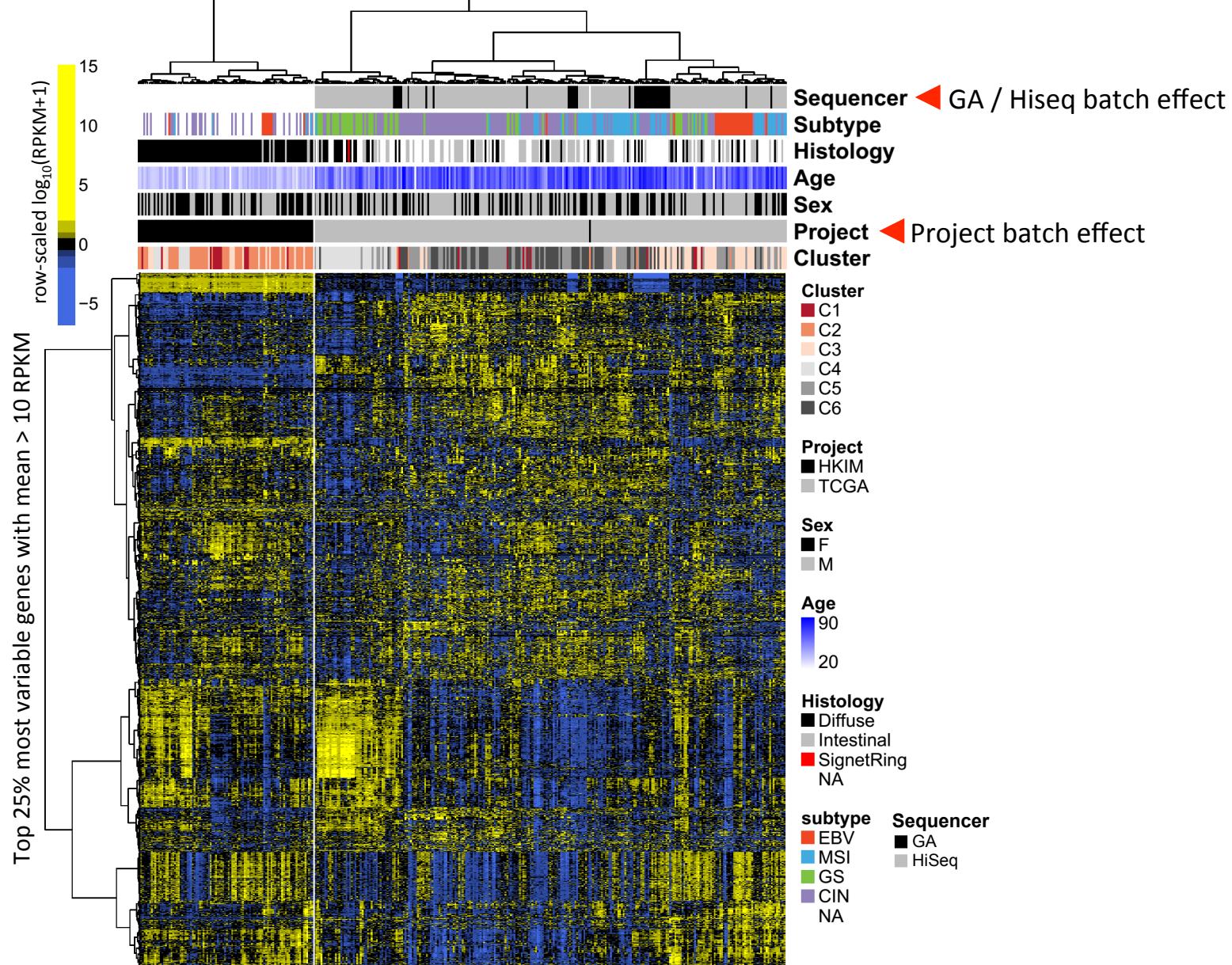
Example1: clustering TCGA ESCA

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



Example2: clustering TCGA STAD + an external cohort

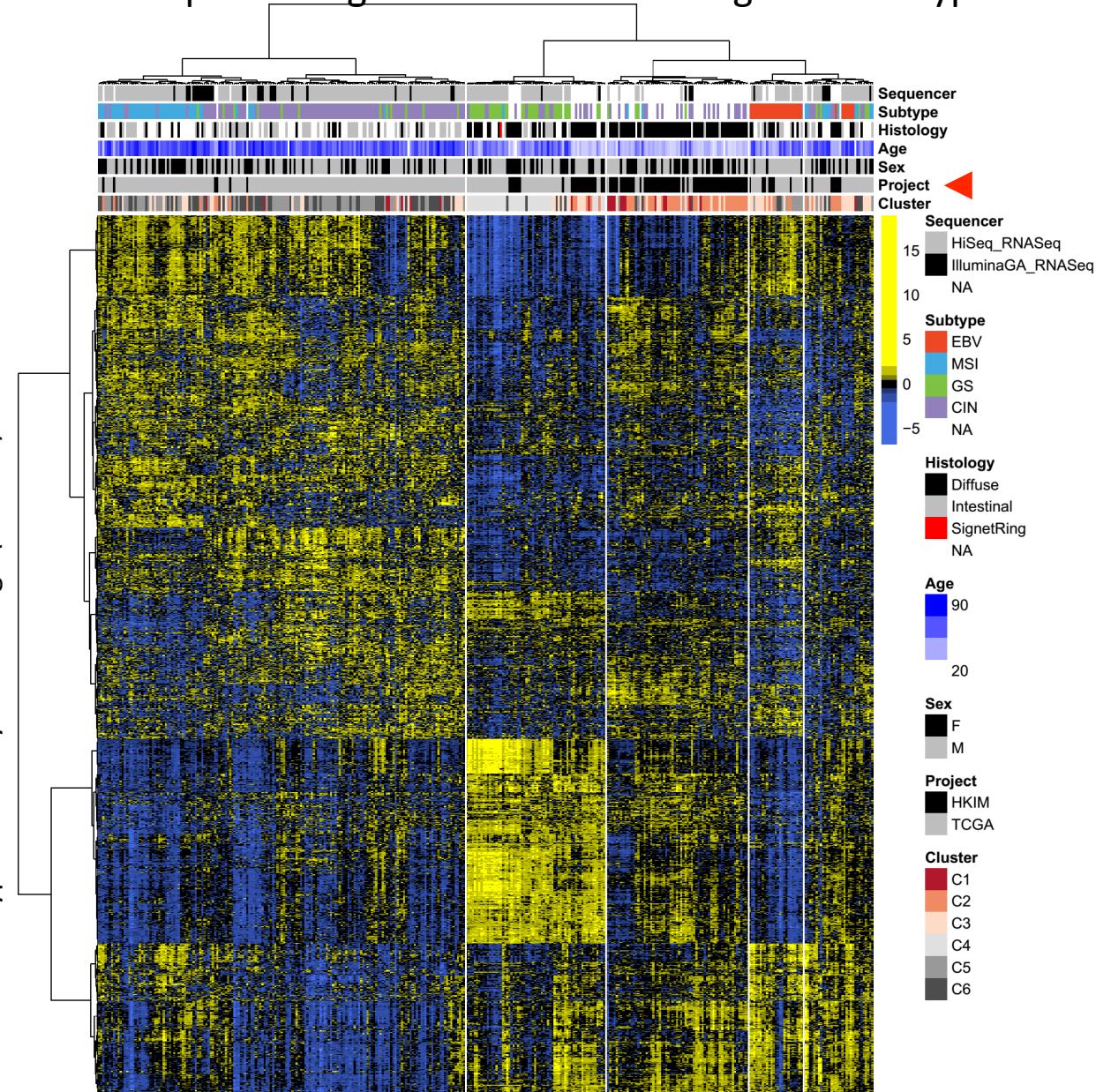
The genes selected for clustering can greatly impact the result



Example2: clustering TCGA STAD + an external cohort

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

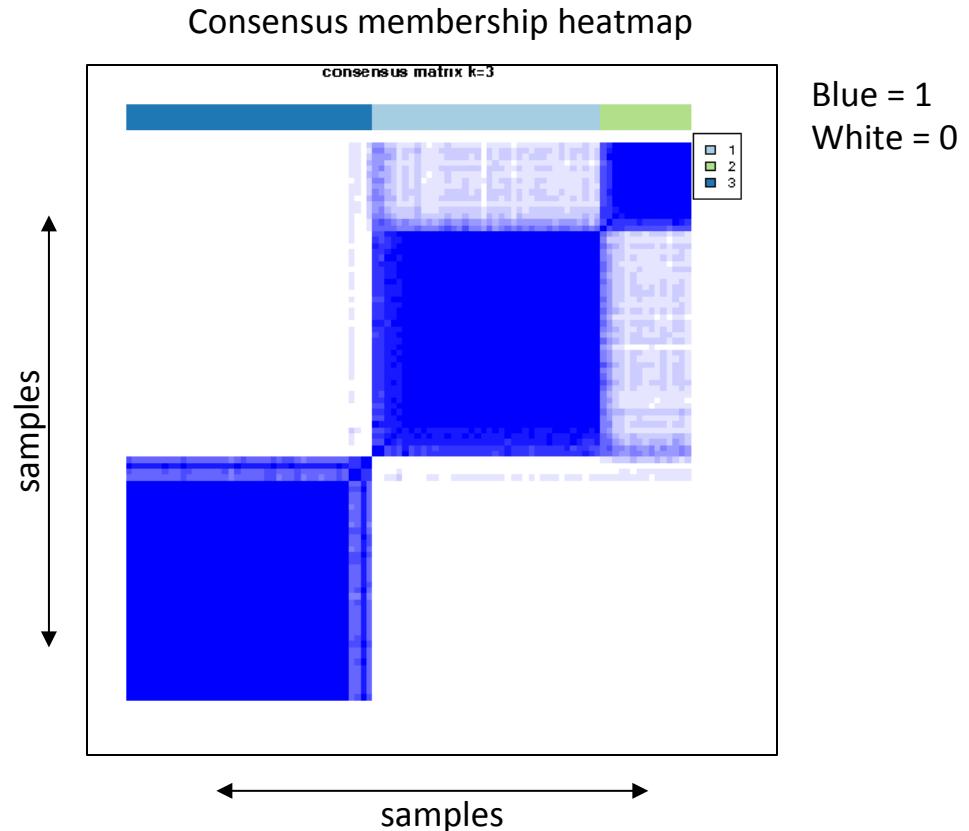
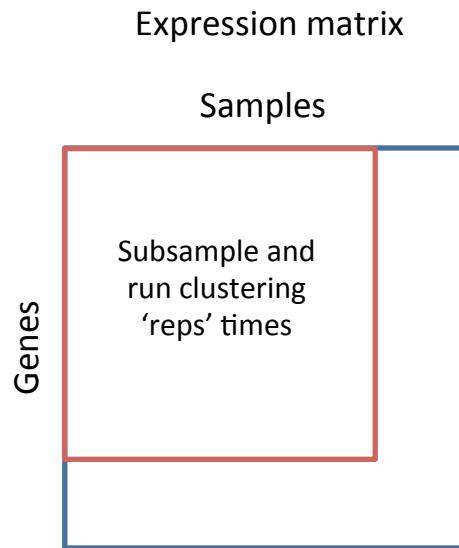
Top 500 differentially expressed genes in each gastric subtype sorted by fold change (n=1485)



Clustering RNA-Seq data

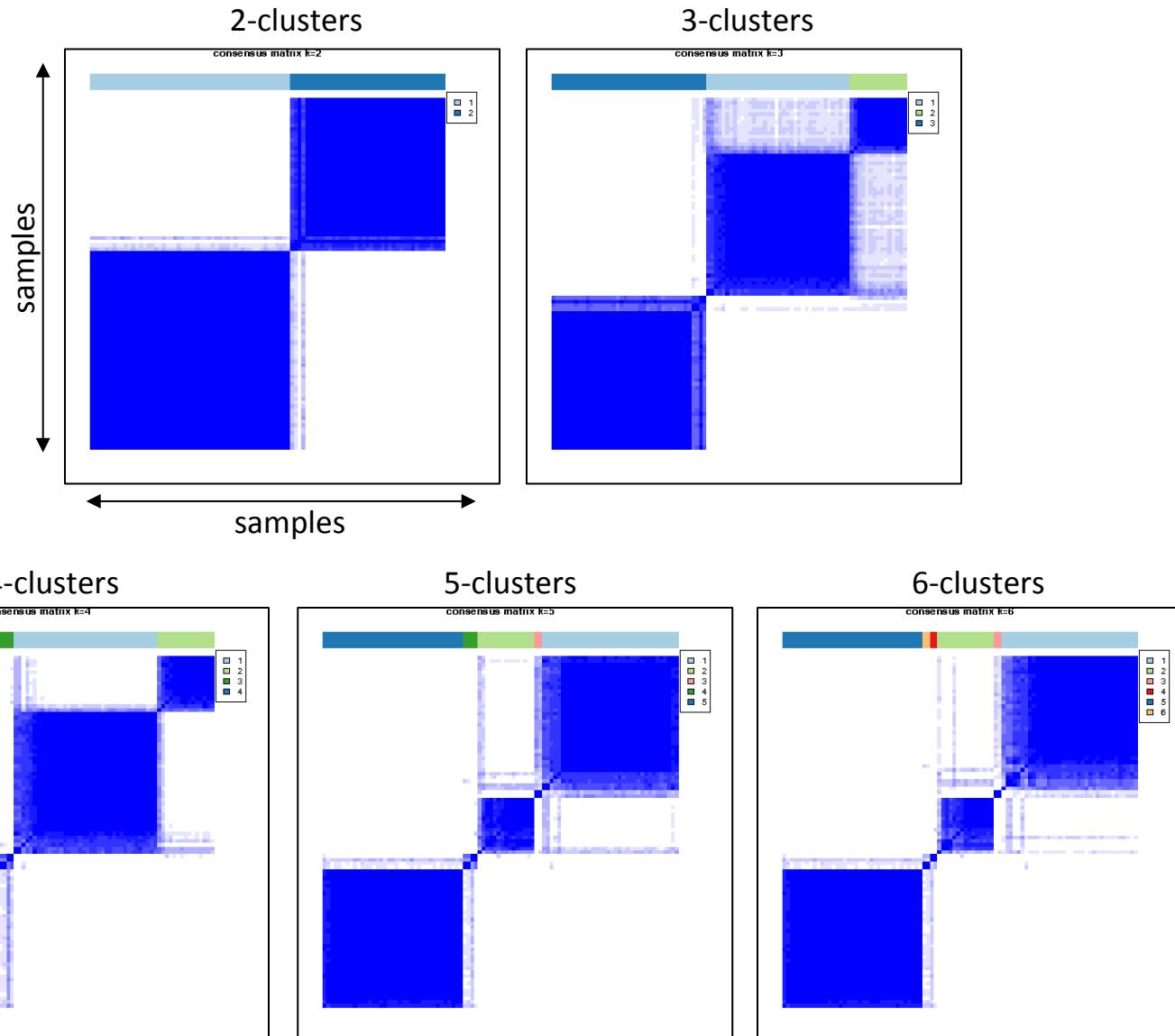
Consensus clustering – ConsensusClusterPlus

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.



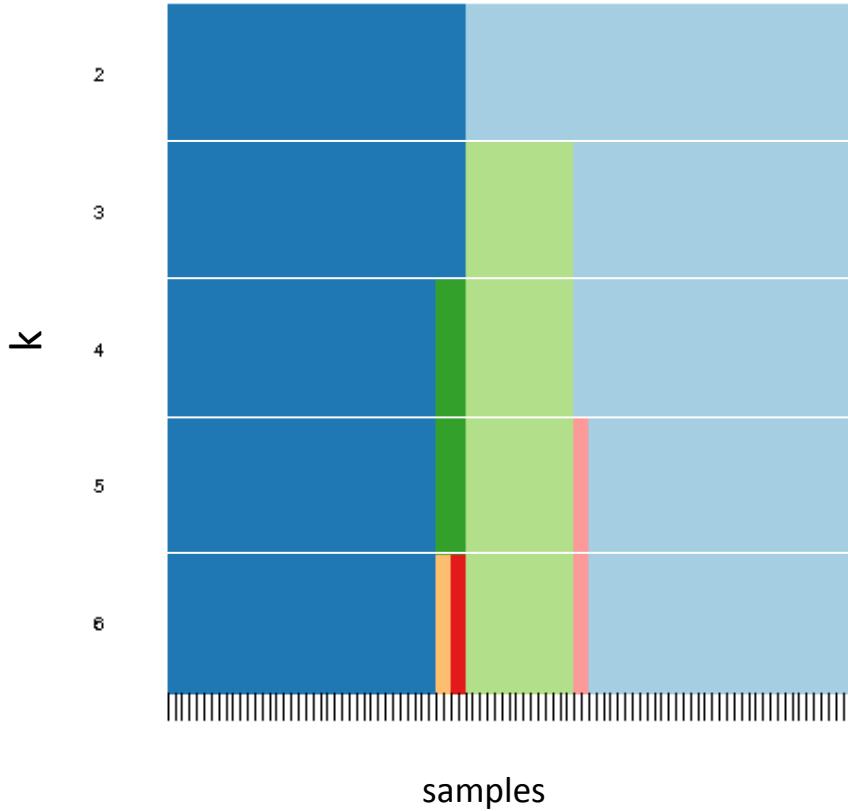
R

Consensus Cluster Plus – consensus membership heatmaps



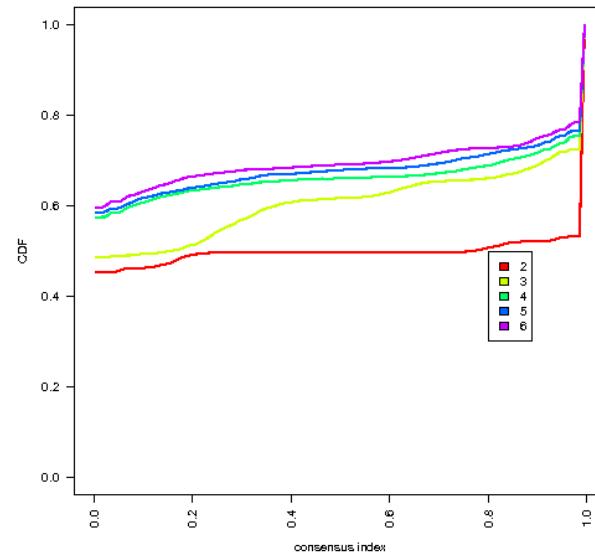
Consensus cluster plus output

tracking plot

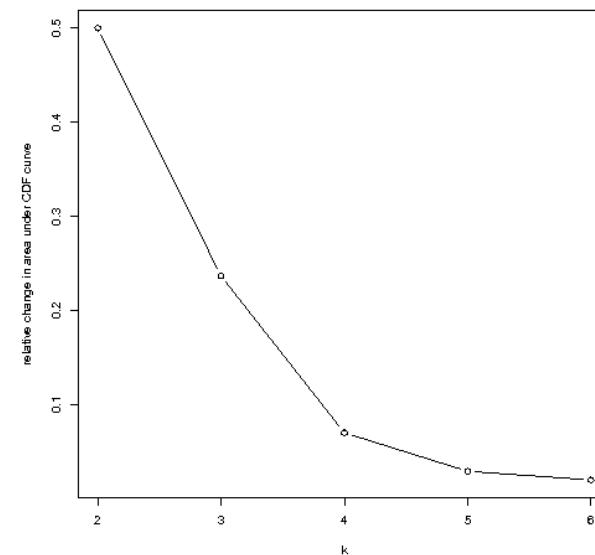


samples

consensus CDF



Delta area



R

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

Differential Expression

SAMR

Deseq

EdgeR

R

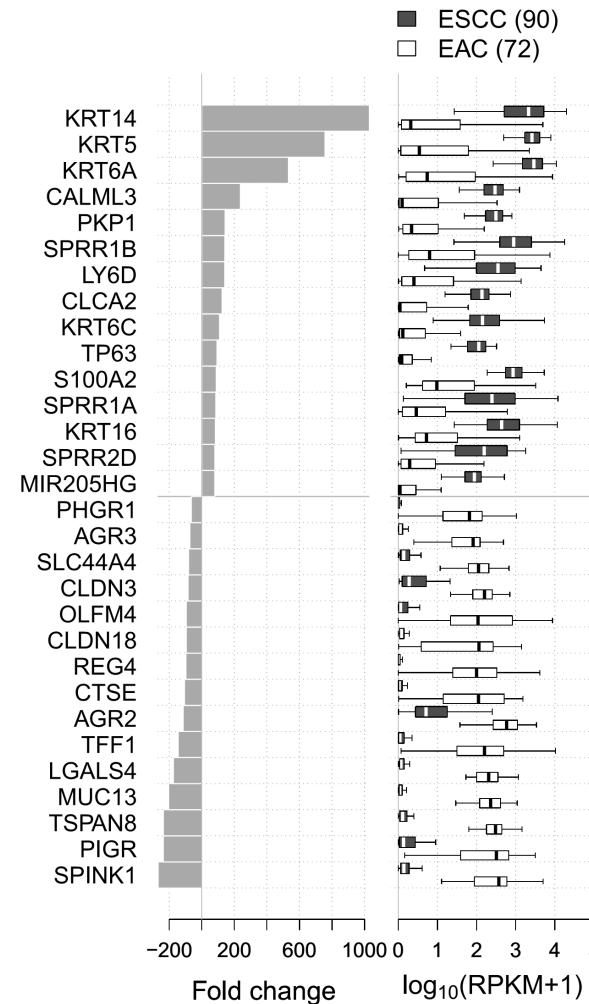
Differential Expression

You can filter the genes by:

- Mean expression
- Fold change
- Wilcoxon test

Use for pathway analysis

- ingenuity
- gene enrichment
- DAVID



Questions?