Big Data Seminar Series: OMICS Data

Nov 5, 2019

Katerina Kechris, Lauren Vanderlinden, Harry Smith

Department of Biostatistics and Informatics

Colorado School of Public Health

Center for Innovative Design & Analysis

colorado school of public health

Outline

1. Current technologies available for the various omics types (Kechris)

2. Insights available using current omics analysis methods (Smith)

3. Tips for handling the common statistical themes in omics data analysis (Vanderlinden)

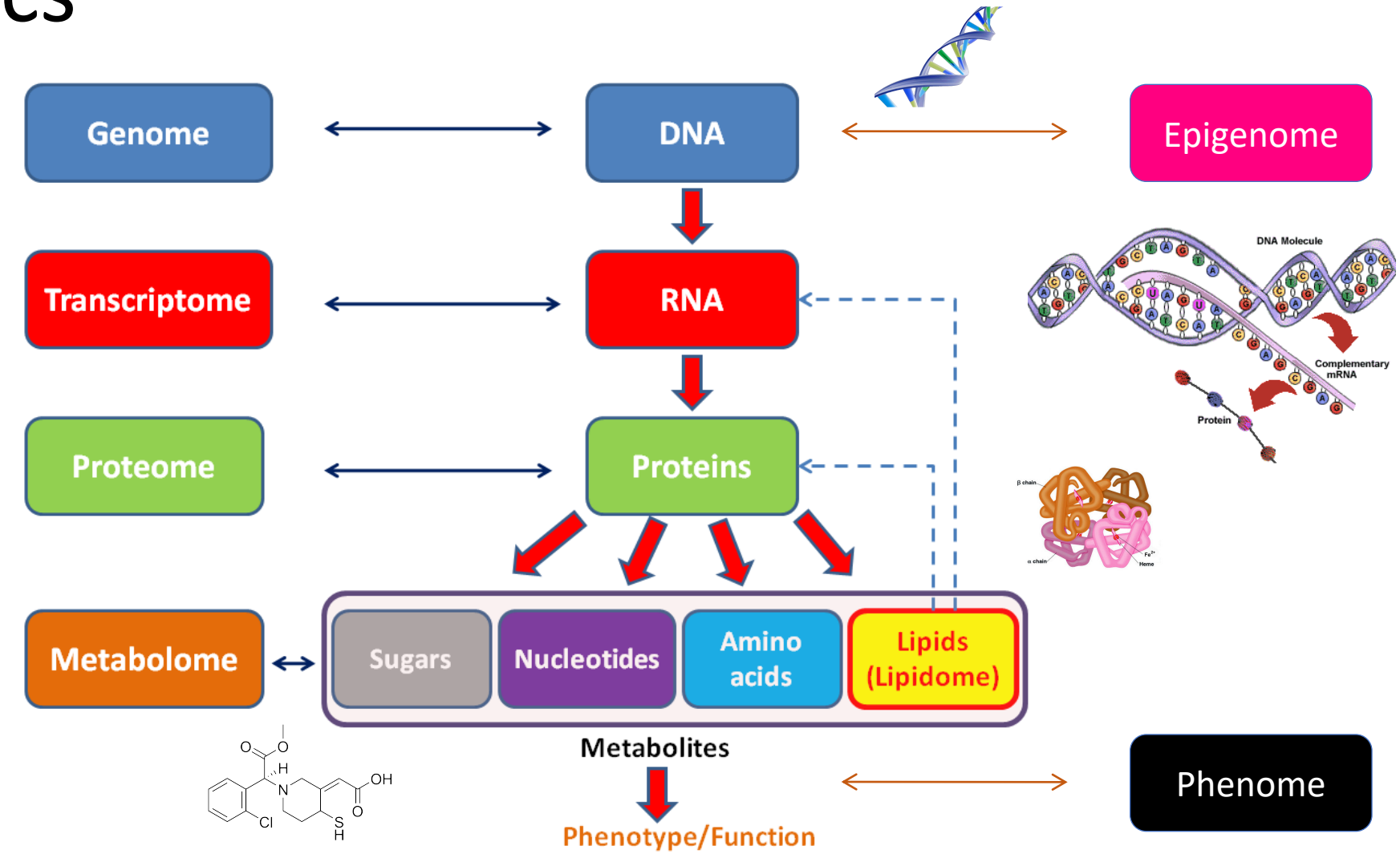
4. Questions and discussion to plan your omics study

Part 1: Background

Katerina Kechris

Center for Innovative Design & Analysis
colorado school of public health

Omics

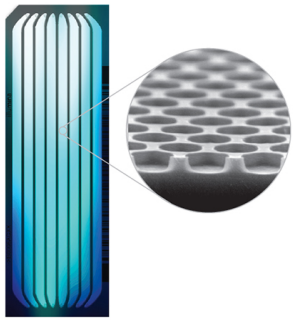


Adapted from http://www.sciencebasedmedicine.org http://www.scientificpsychic.com/fitness/transcription.gif
http://themedicalbiochemistrypage.org/images/hemoglobin.jpg http://upload.wikimedia.org/wikipedia/commons/c/c6/Clopidogrel_active_metabolite.png
http://creatia2013.files.wordpress.com/2013/03/dna.gif

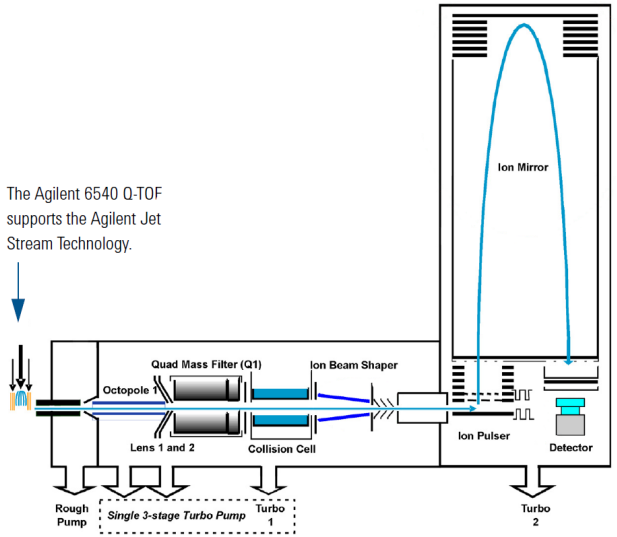
Technologies

1. Microarrays (RNA/DNA)

2. Sequencing (RNA/DNA)



3. Mass-spectrometry (proteins/metabolites)



DNA

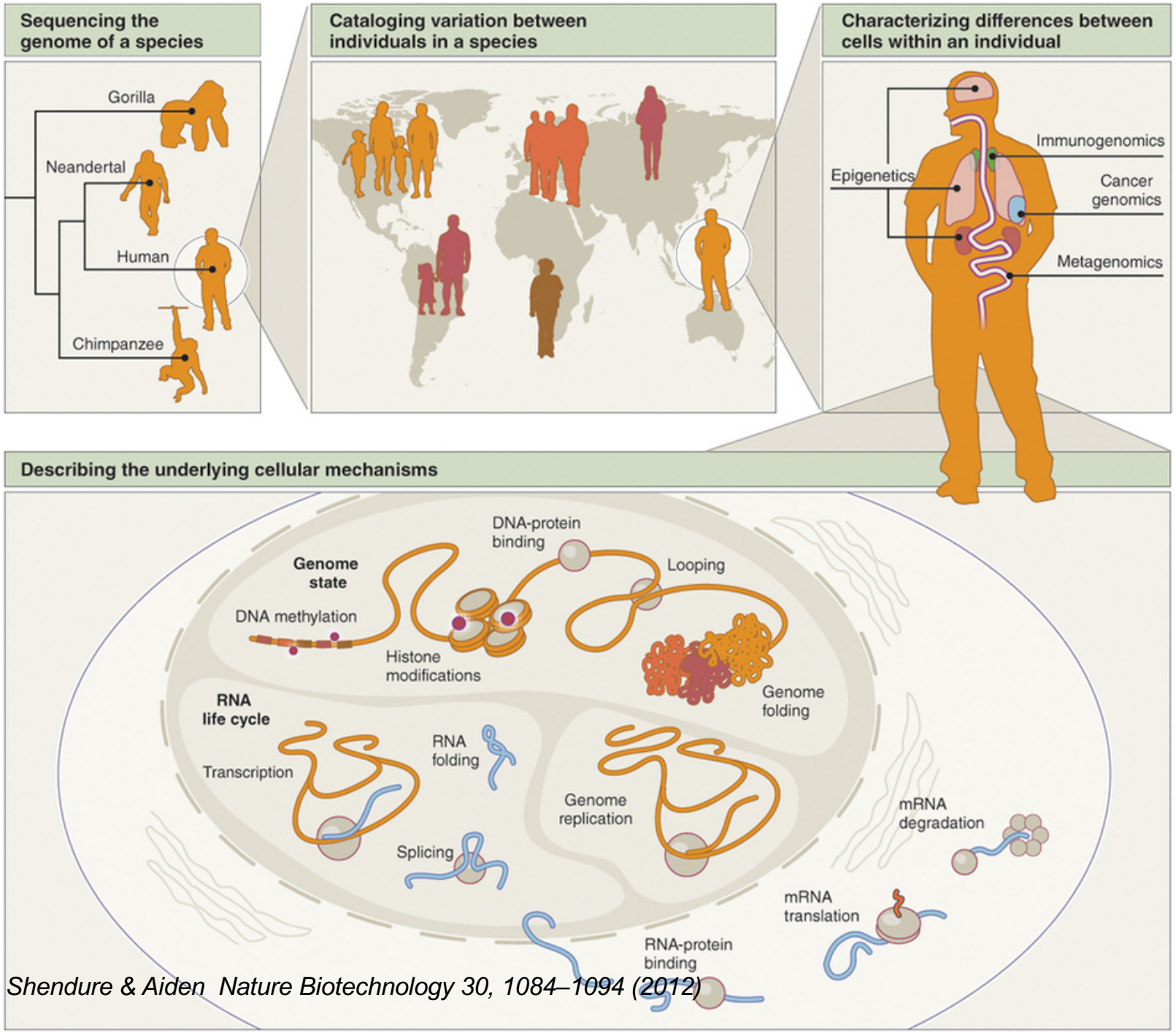
- Genome
  - Across species
  - Within population
- Exome
- Single nucleotide polymorphisms
- Chromosome conformations

DNA Modifications & Interactions

- DNA methylation (epigenome)
- Histone modifications (epigenome)
- DNA binding proteins (e.g., transcription factor)

RNA

- mRNA (transcriptome)
- Other species
  - miRNA, lncRNA 16s rRNA (microbiome)
- RNA binding proteins (e.g., splicing factors)
- Methylation RNA (epitranscriptome)
- Single-cell



Shendure & Aiden Nature Biotechnology 30, 1084–1094 (2012)

Proteins

- Abundance
- Structure
- Protein-protein interactions
- Post-translation modifications (e.g., phosphoproteomics, glycoproteomics)

Metabolites

- Types of small molecules
  - Lipids – lipidomics
  - Exogenous factors– exposome
  - Diet/drugs - nutrigenomics
- Toxicology (changes due to chemical)
- Metabolic reactions (e.g., fluxomics)
- Nuclear magnetic resonance (NMR) (metabonomics)

Large-scale Projects & Databases

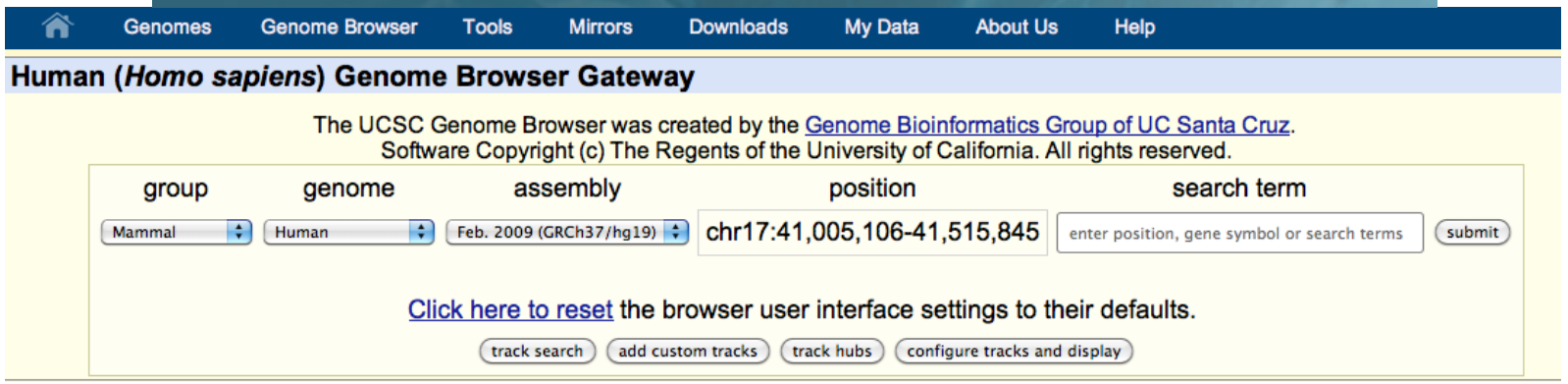
ENCODE

mod ENCODE

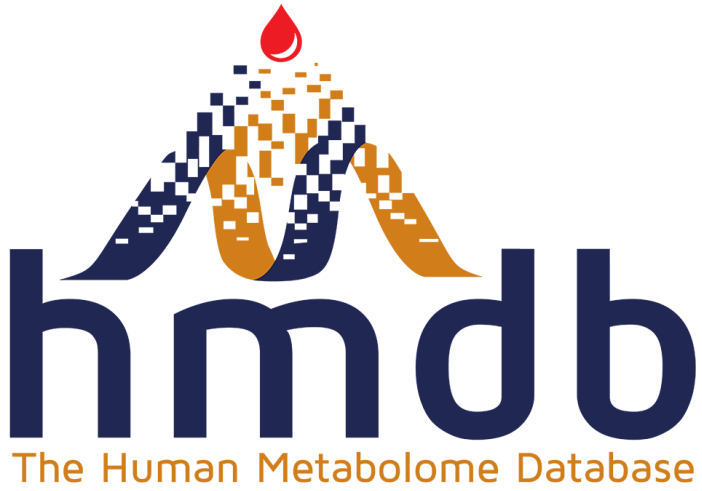
CellMiner™

NCI 60 Database

THE CANCER GENOME ATLAS



Large-scale Projects & Databases



translating
the code of life

Center for Innovative Design & Analysis
colorado school of public health

Multiple-Cohorts & Populations

NIH National Institutes of Health
All of Us Research Program

COMETS COnsortium of METabolomics Studies

PACE
Pregnancy And Childhood Epigenetics

NIH National Institutes of Health

Home » Research & Training

ENVIRONMENTAL INFLUENCES ON CHILD HEALTH OUTCOMES (ECHO) PROGRAM

Resources @ AMC

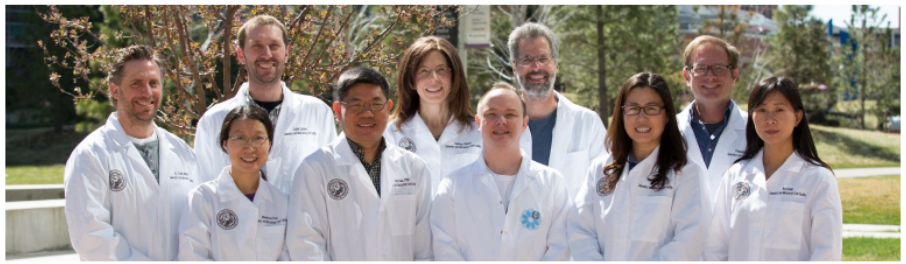
SCHOOL OF MEDICINE
RNA Bioscience Initiative
UNIVERSITY OF COLORADO **ANSCHUTZ MEDICAL CAMPUS**

Home > Research > Shared Resources > **Genomics**

- Home
- Services
- Facility and Platforms
- Data Analysis
- Quote Request
- Sample Submissions and Forms
- Contact Us

Address:

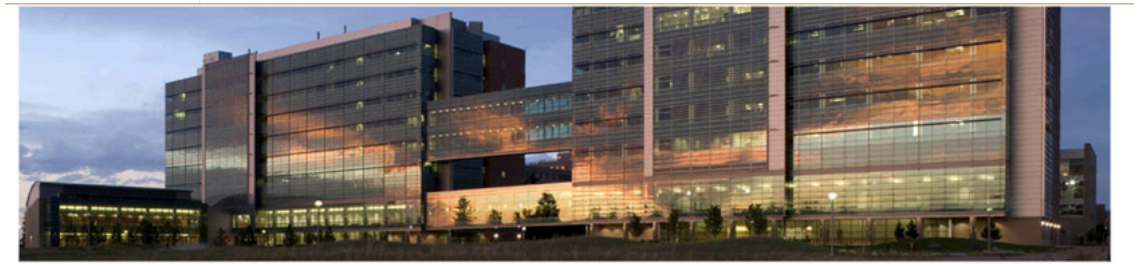
Location/Fed Ex
Genomics and Microarray Core
Anschutz Medical Campus
RC-2, Room 9400
12700 E. 19th Ave.
Aurora, CO 80045
Fax: 303-724-6046



Genomics Shared Resource Home Page

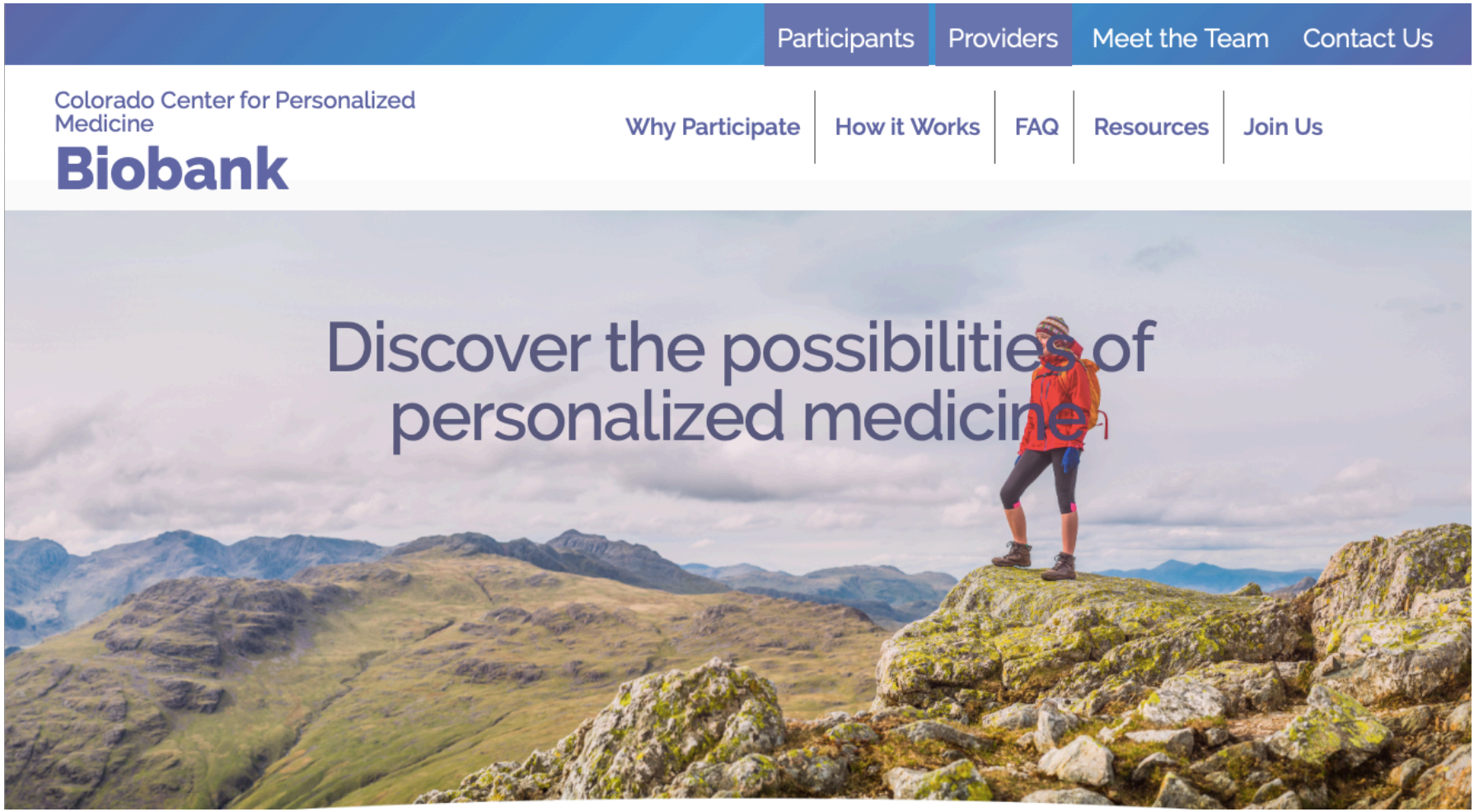
The Genomics and Microarray Shared Resource at University Of Colorado Denver Cancer Center is an advanced, state-of-the-art DNA and Protein microarray and Next Generation (NextGen) DNA sequencing technology center providing crucial research support for investigators interested in using:

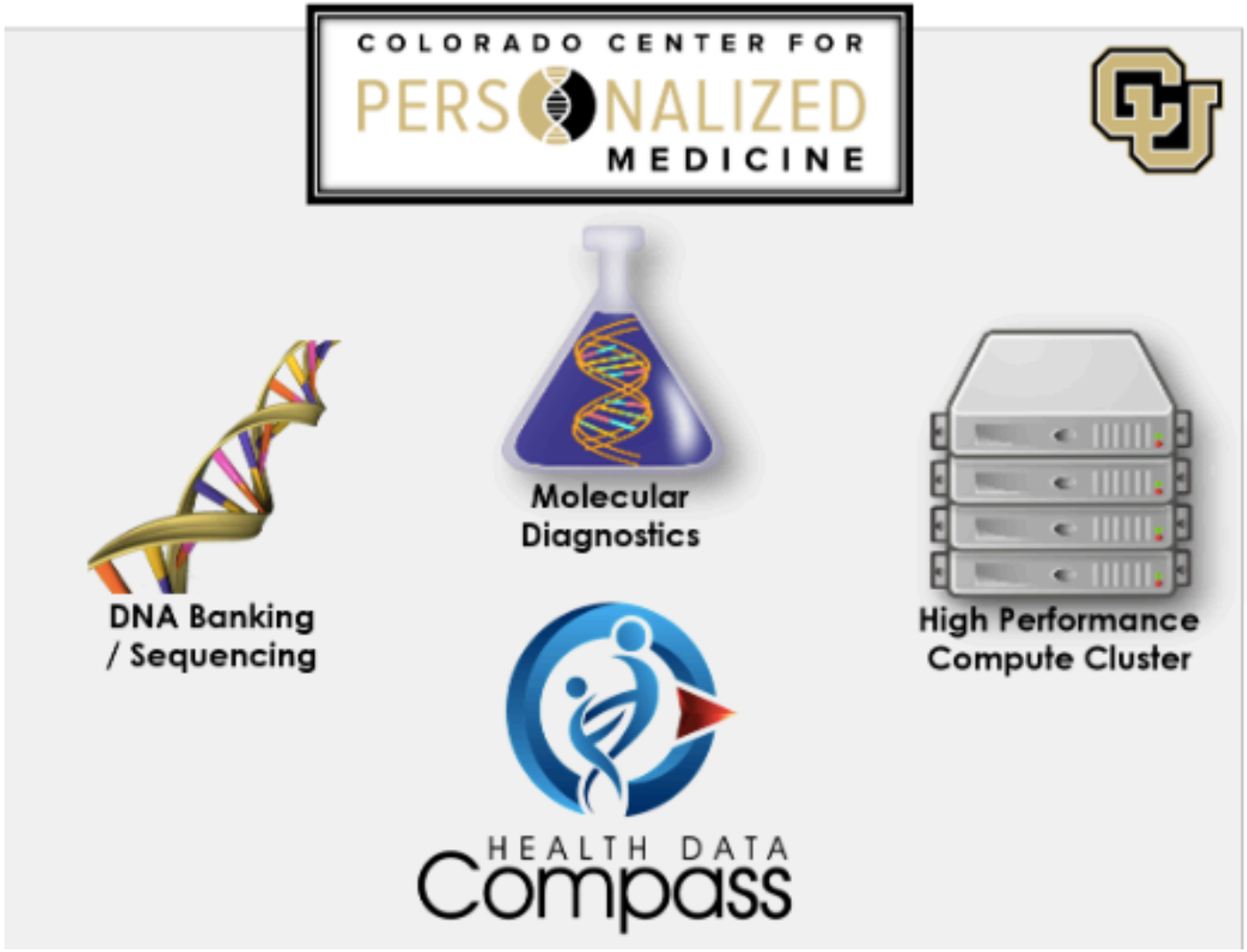
- Next Generation Sequencing:
  - Illumina HiSeq 2500/4000 sequencing
  - Illumina MiSeq sequencing
  - LifeTech IonPGM sequencing
- DNA Microarray:
  - Illumina BeadArrays
  - Agilent Microarrays



University of Colorado School of Medicine Biological Mass Spectrometry Facility

Center for Innovative Design & Analysis
colorado school of public health


Participants
Providers
Meet the Team
Contact Us
Colorado Center for Personalized Medicine
Biobank
Why Participate
How it Works
FAQ
Resources
Join Us
Discover the possibilities of personalized medicine


COLORADO CENTER FOR
PERSONALIZED
MEDICINE
Molecular
Diagnostics
DNA Banking
/ Sequencing
High Performance
Compute Cluster
HEALTH DATA
Compass

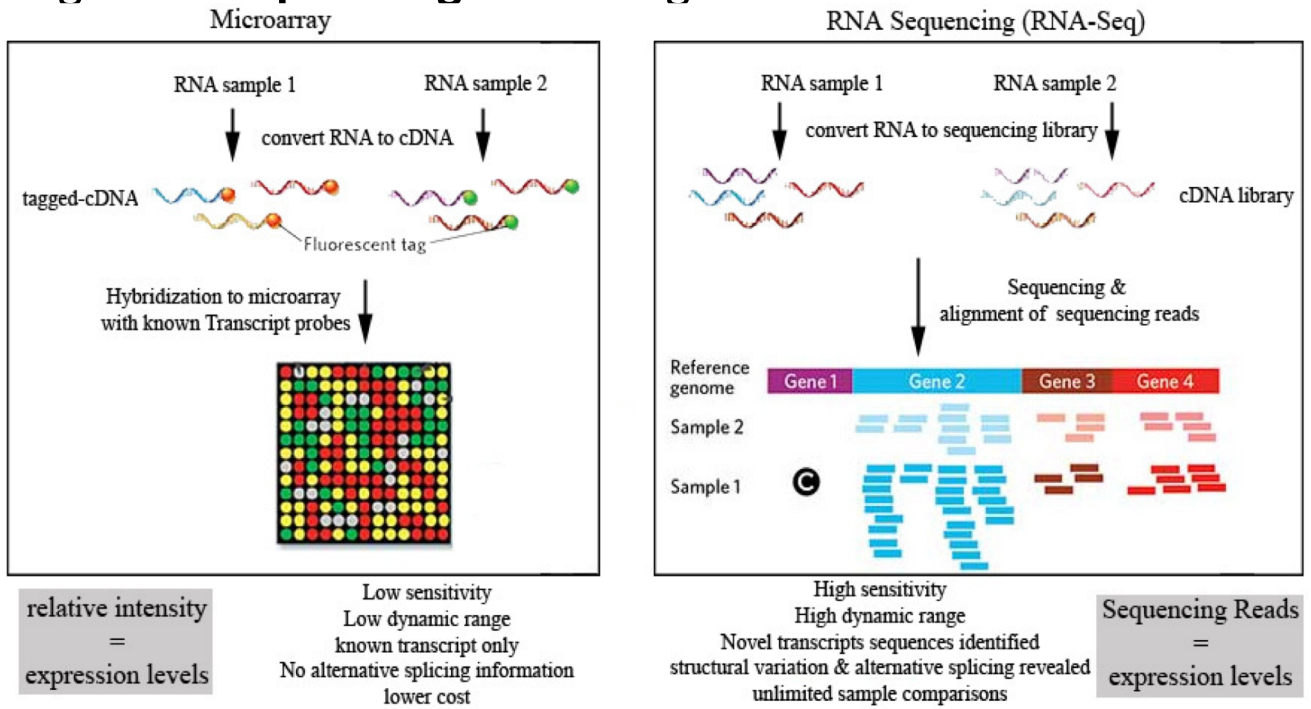
Part 2:
What questions can you answer with omics data?

Harry Smith

Center for Innovative Design & Analysis
colorado school of public health

Question 1

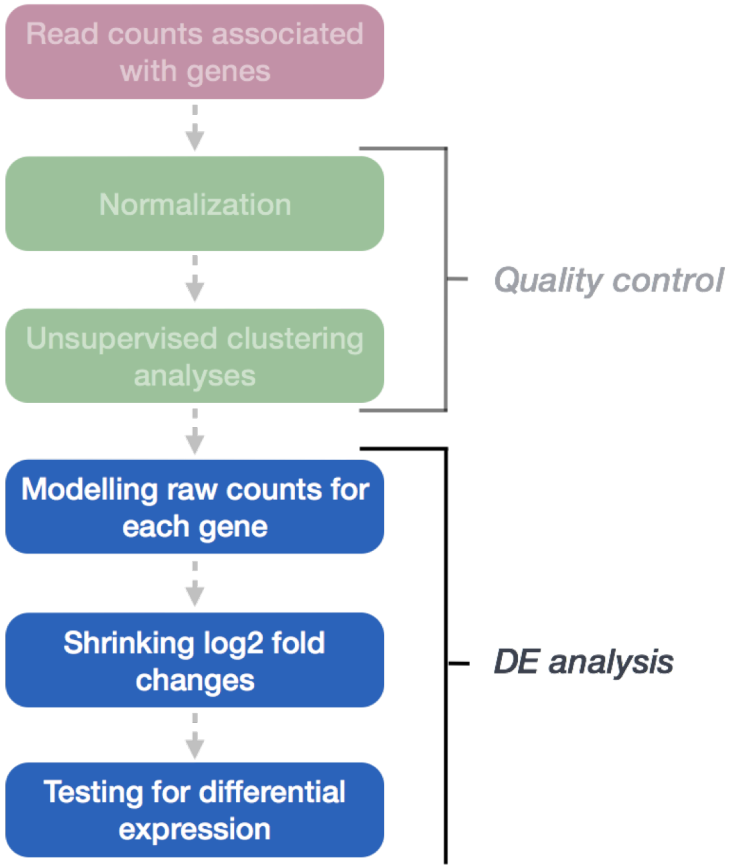
- **Question:** Are gene/transcripts expressed at different levels between two experimental groups?
- **Solution 1:** Differential Expression analysis using **RNA-Sequencing** technologies and DESeq2.

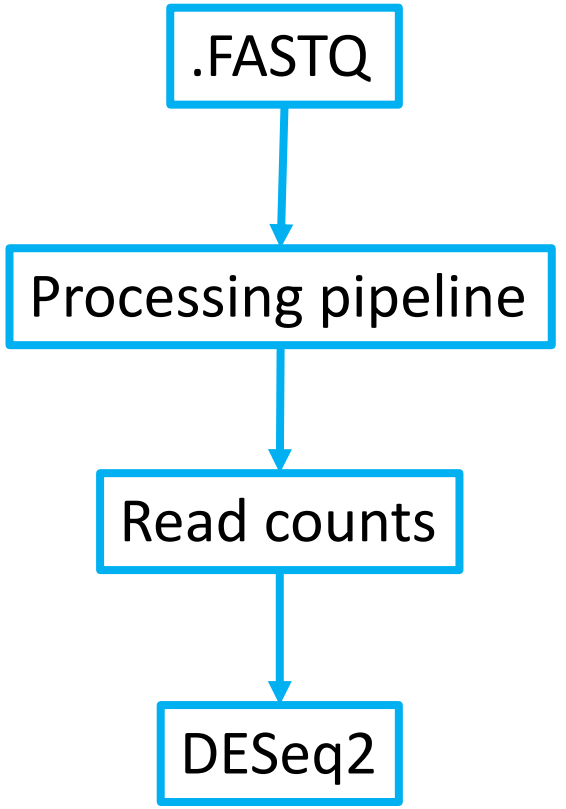


https://www.otogenetics.com/rna-sequencing-vs-microarray/

Question 1

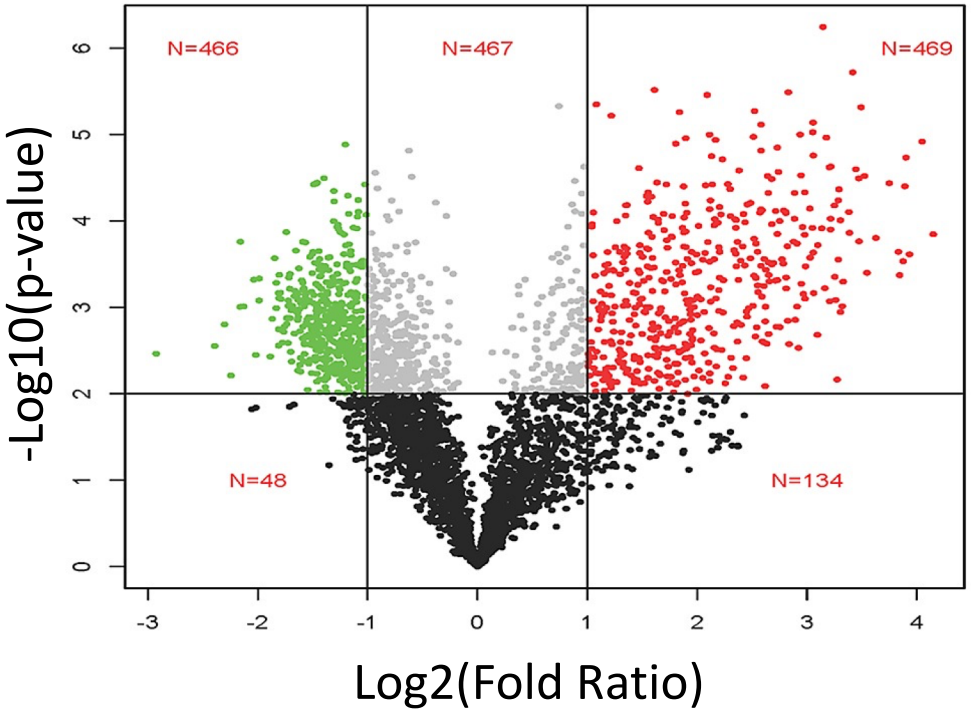
- Example





https://hbctraining.github.io/DGE_workshop/lessons/05_DGE_DESeq2_analysis2.html

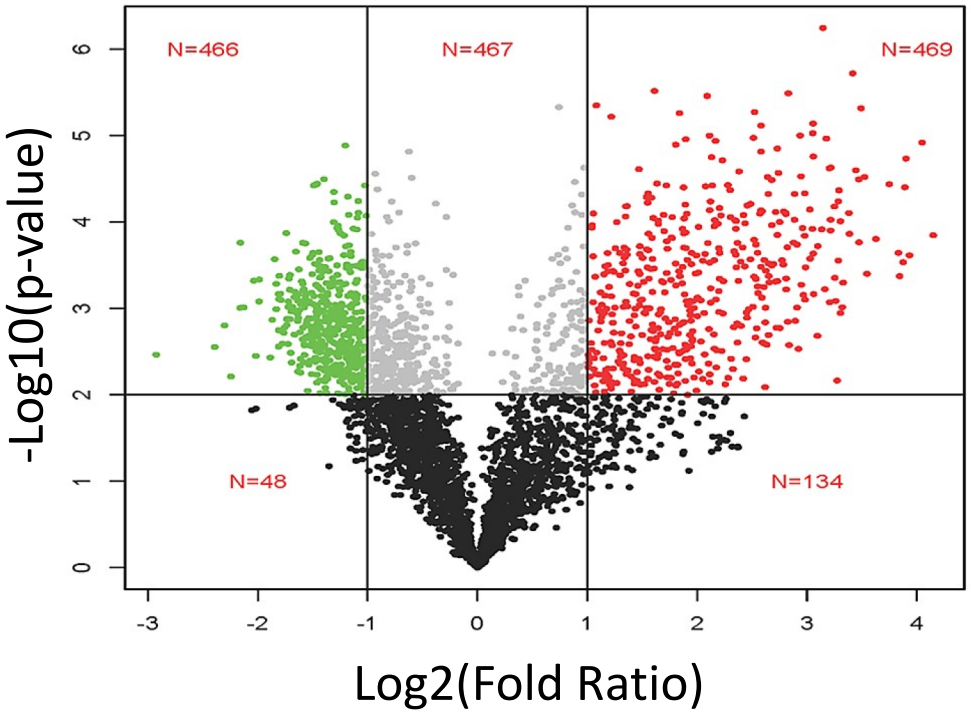
Question 1



Identify differentially expressed genes

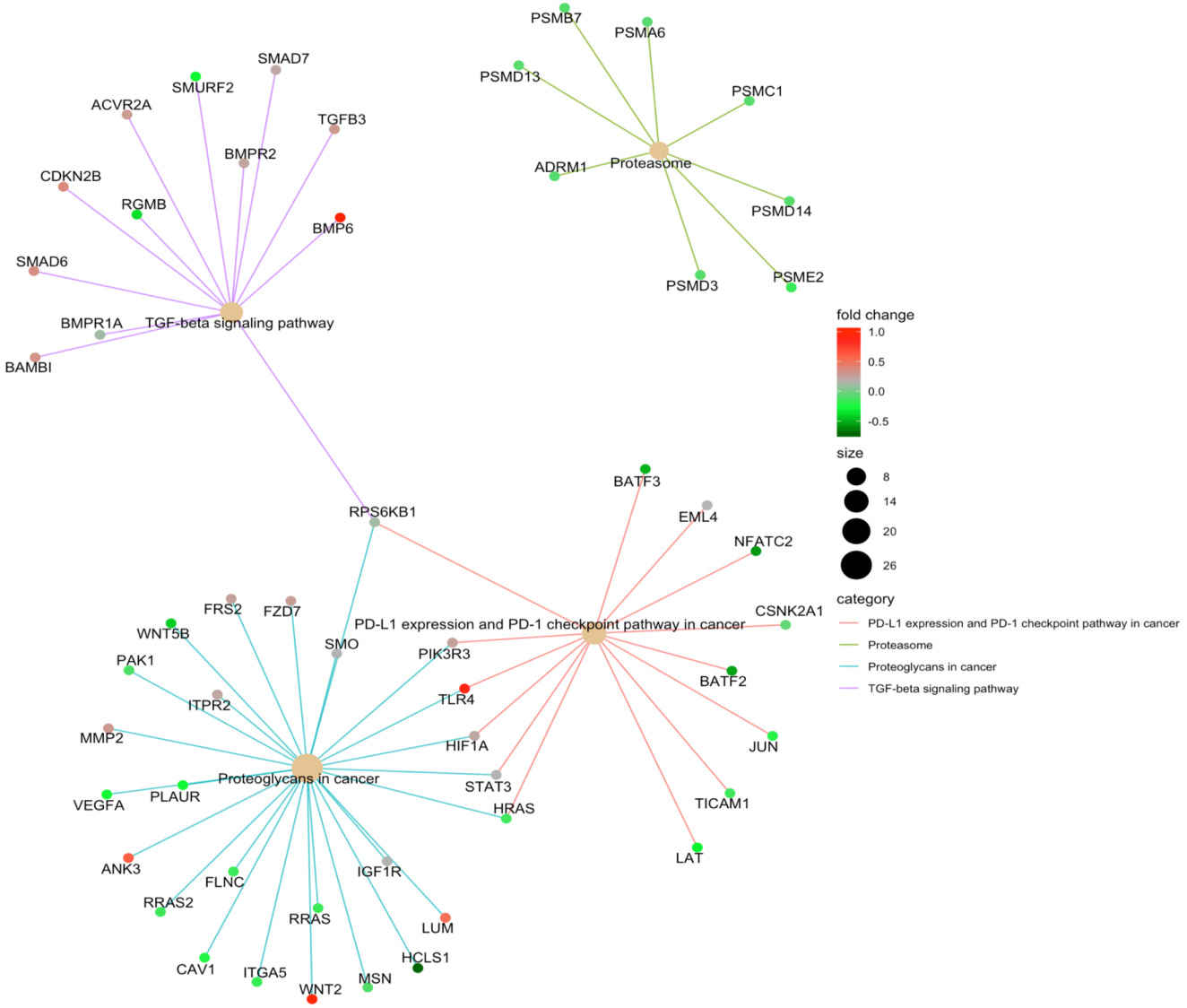
https://figshare.com/articles/_Volcano_plot_used_in_differential_expression_analysis_/448380/1

Question 1



Identify differentially expressed genes

Identify enriched biological pathways based on DE genes



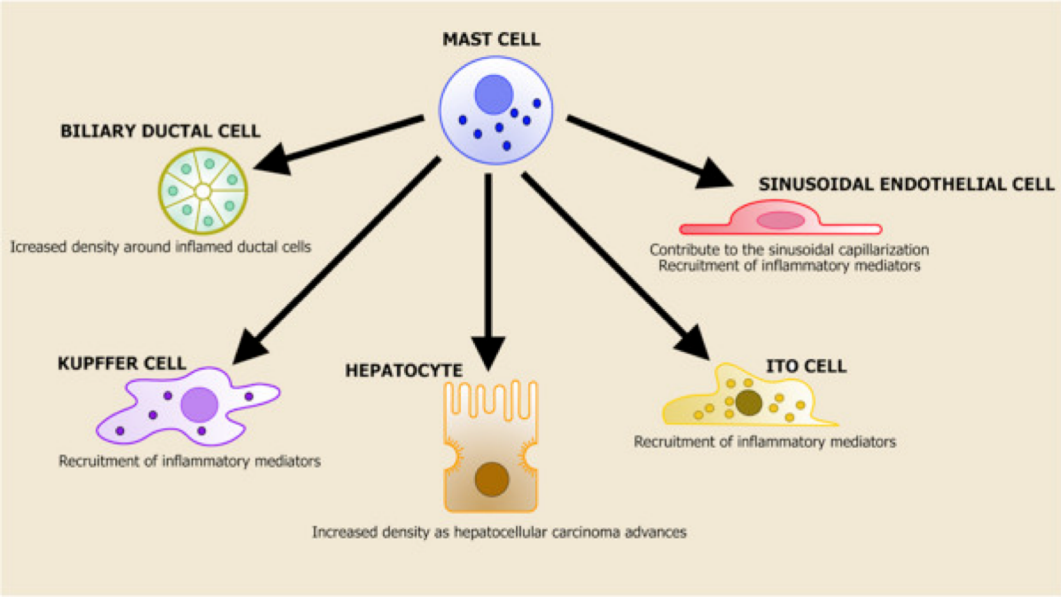
Center for Innovative Design & Analysis
colorado school of public health

https://figshare.com/articles/_Volcano_plot_used_in_differential_expression_analysis_/448380/1

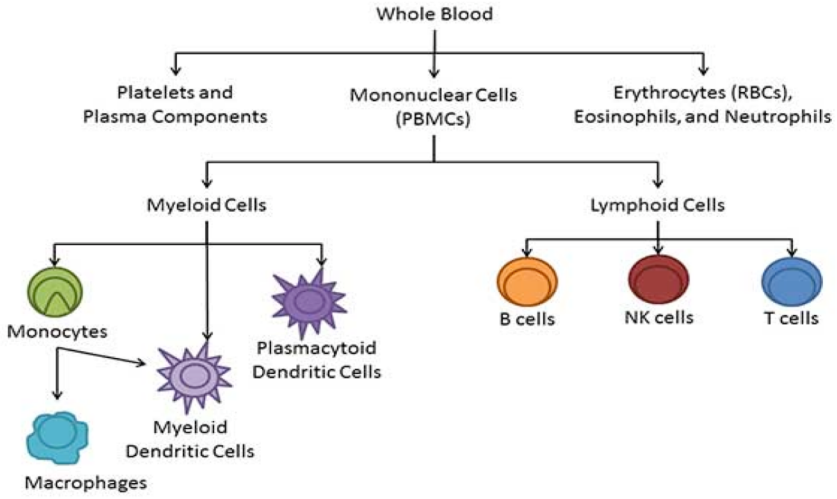
Question 2

- **Question:** Can omics be used to reveal complex and rare cell populations, uncover regulatory relationships between genes, and track the trajectories of distinct cell lineages in development?
- **Solution:** Identify complex and rare cell populations and uncover regulatory relationships between genes using single-cell RNA-Sequencing technologies.

Tissue: Liver

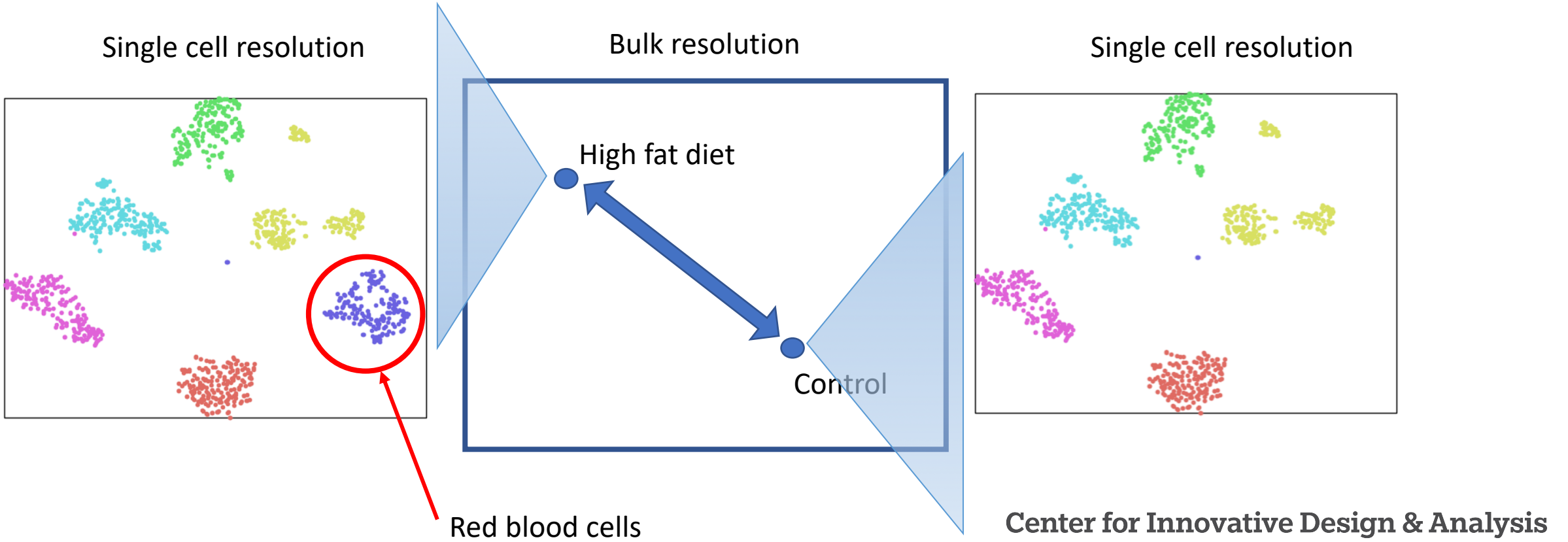


Tissue: Blood



Question 2

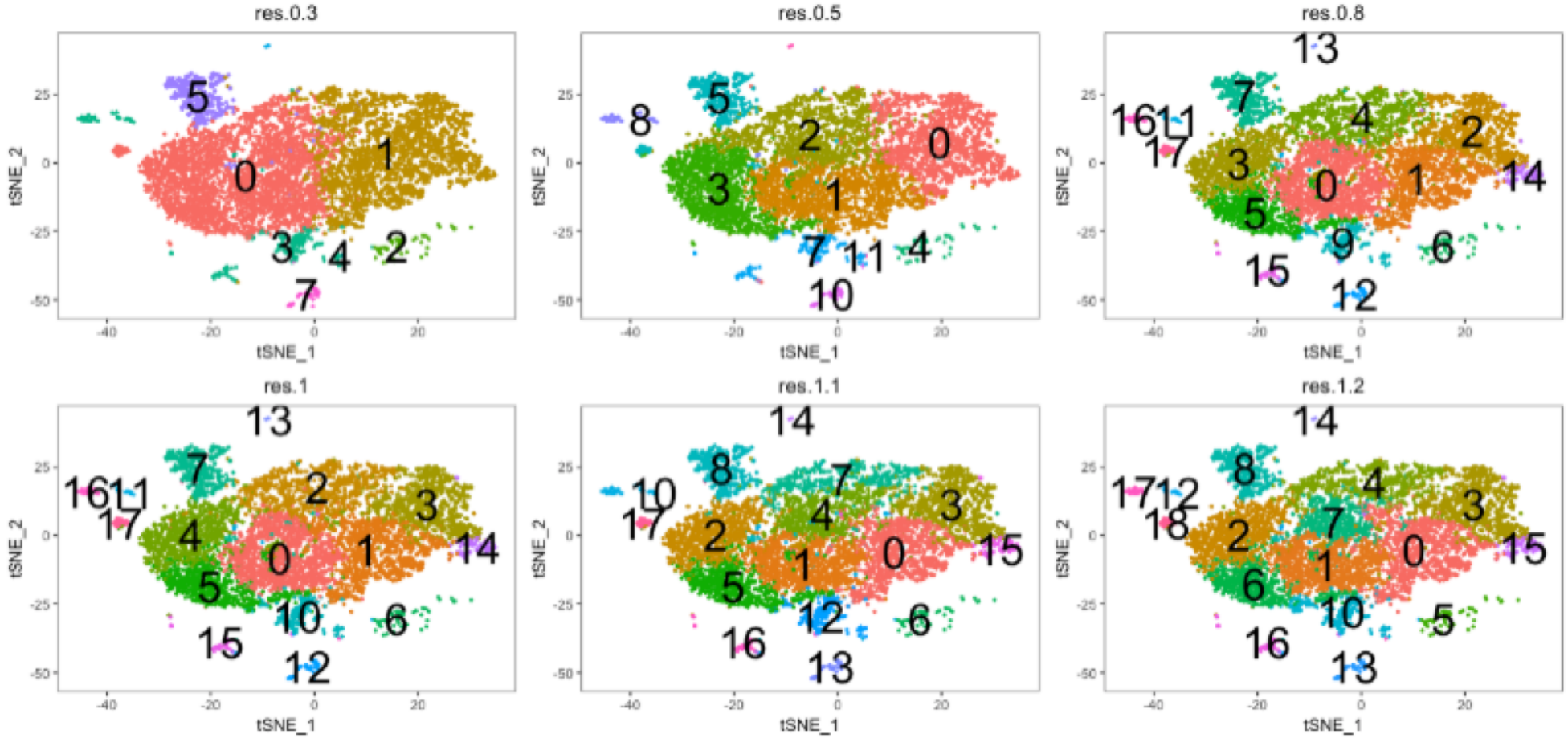
- With single cell resolution you can get one profile for each individual cell in the sample



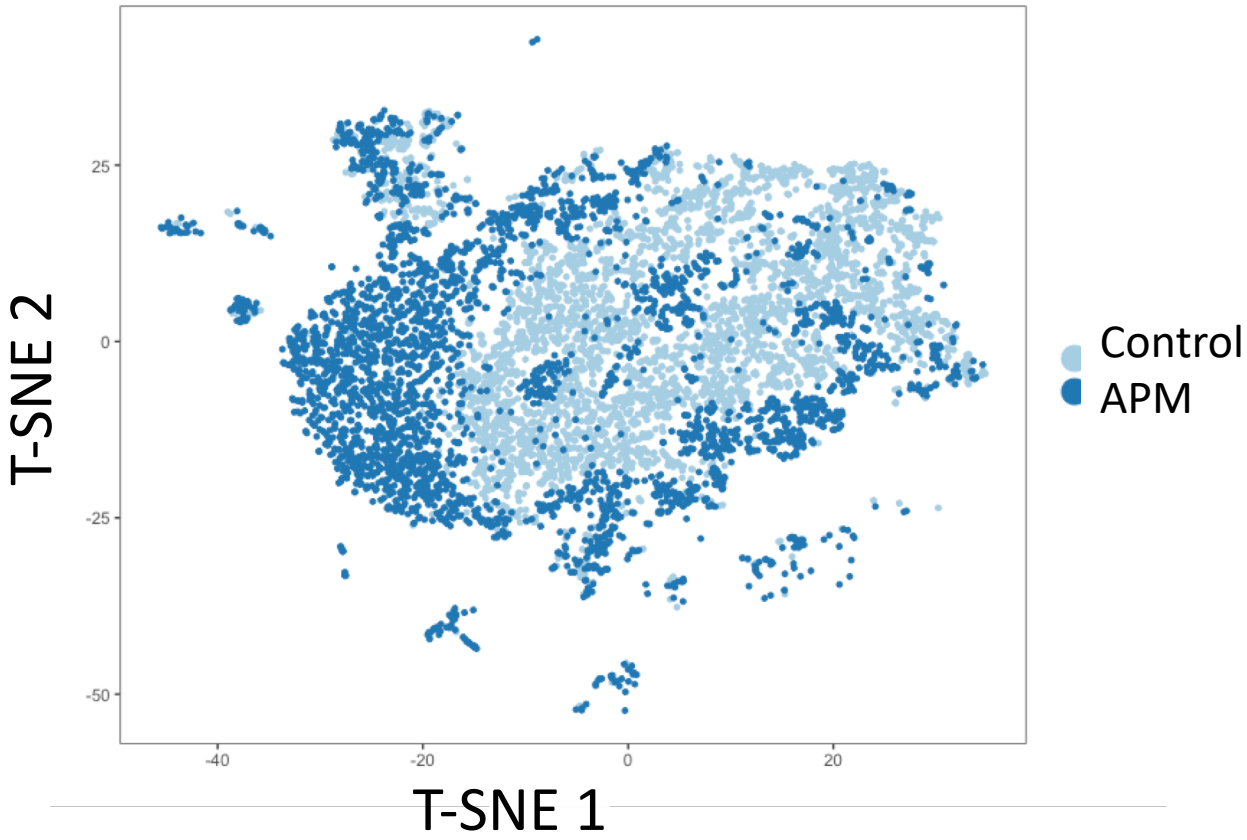
Question 2: Experimental background

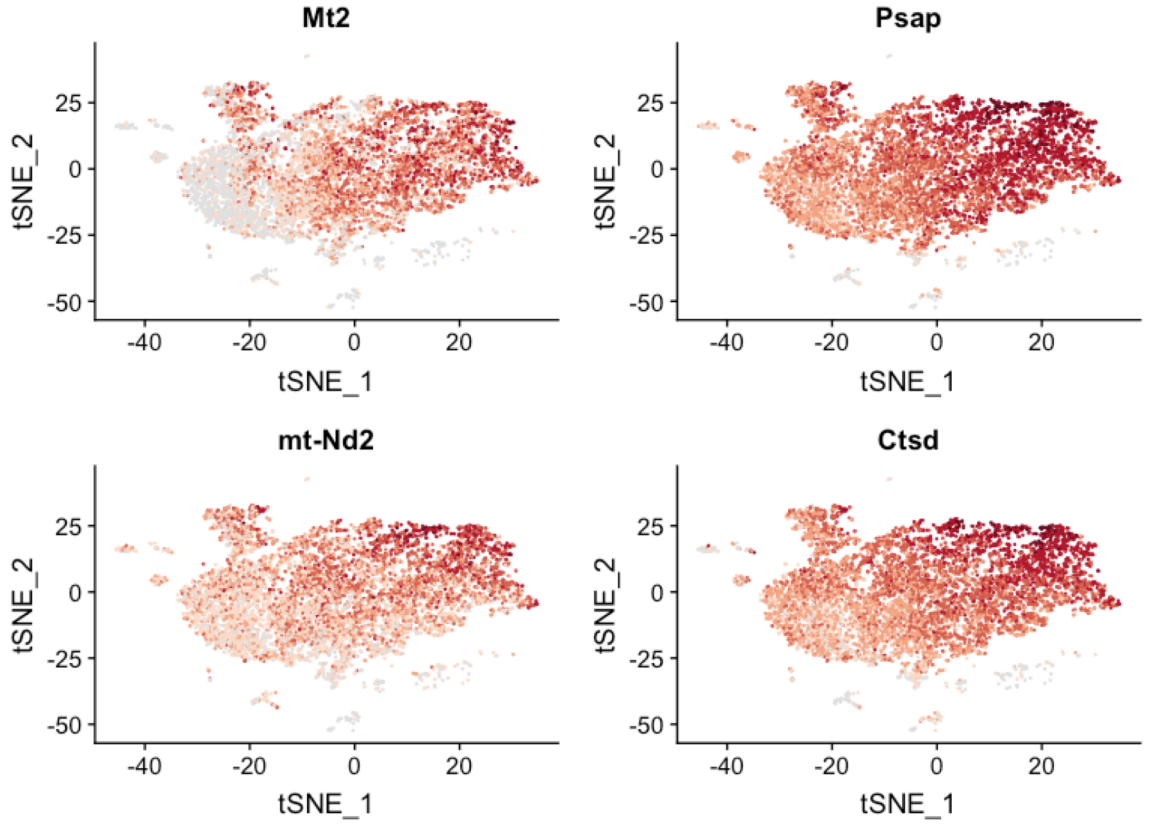
- Researchers were interested in how the transcriptomic profile of lung tissue was affected by an exposure at the single-cell level.
- Bronchoalveolar lavage cells
- Mouse model
- Two groups
  - Control
  - Exposed
- One time point

Question 2: Cluster identification



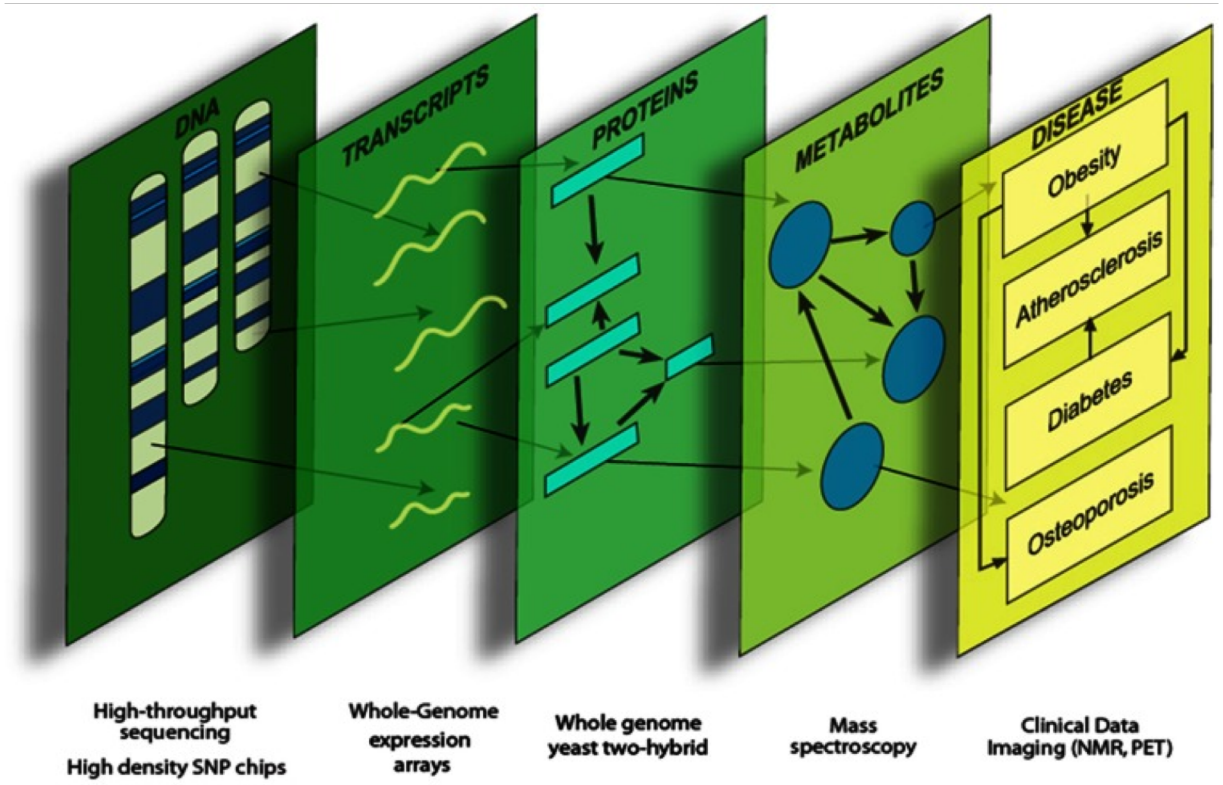
Question 2: Sample specific expression





Question 3

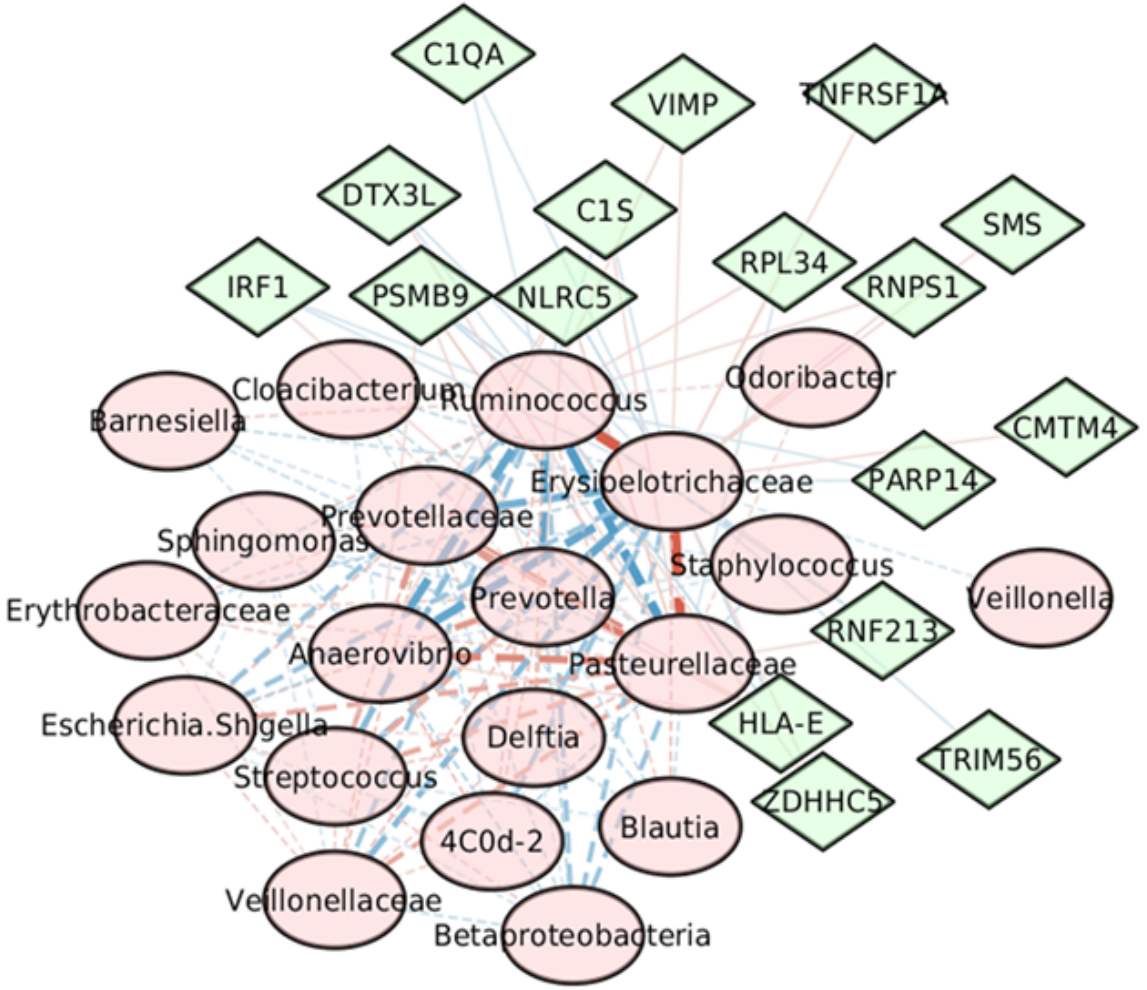
- **Question:** I have multiple omics datasets. Is there a way for me to integrate these data and generate meaningful results?
- **Solution 1:** You can use the smCCnet package (Shi. W et al., https://academic.oup.com/bioinformatics/article/35/21/4336/5430928)
- **Solution 2:** Or you can use a Systems Genetics Approach



https://www.researchgate.net/figure/Systems-genetics-analysis-Systems-genetics-integrates-genetic-variation-intermediate_fig1_237014601

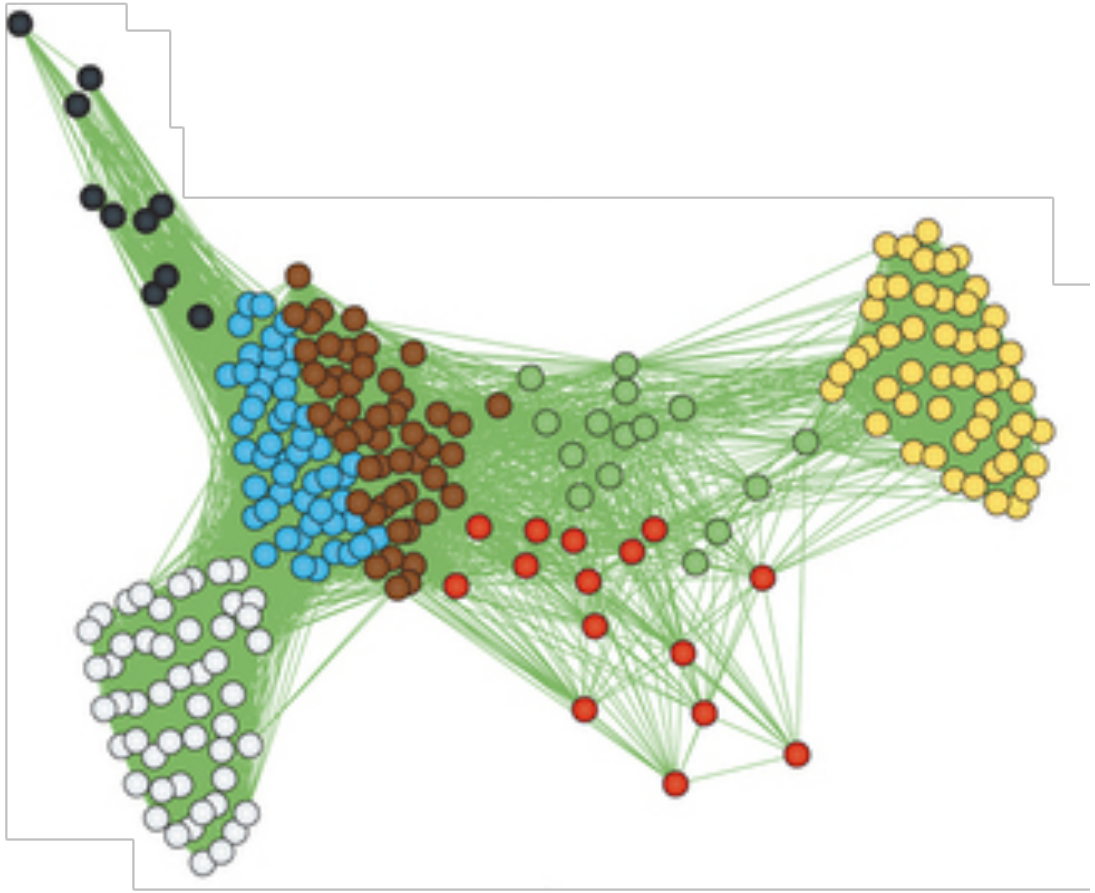
smCCnet

- Phenotype: sCD14 serum levels
- Omics data set 1: RNA-Seq
- Omics data set 2: Microbiome

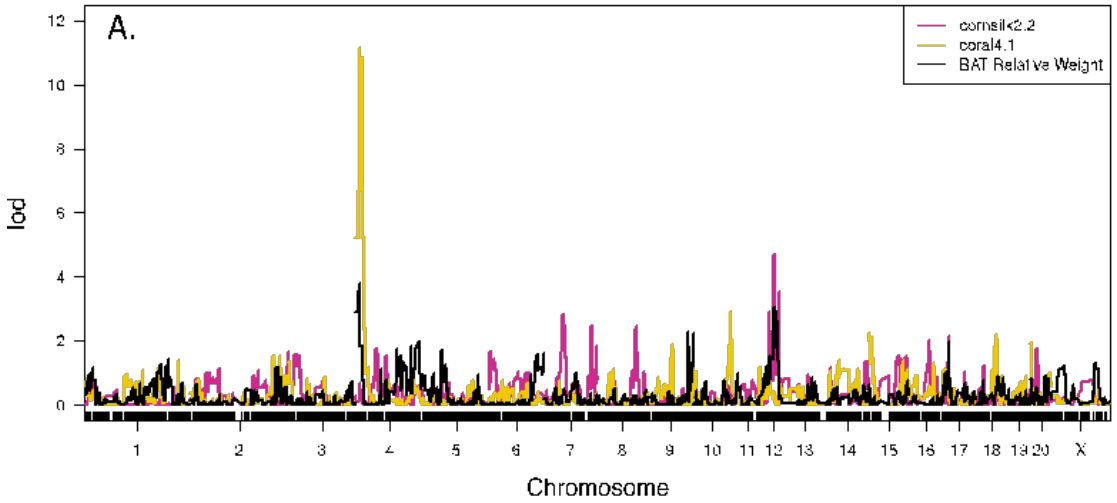


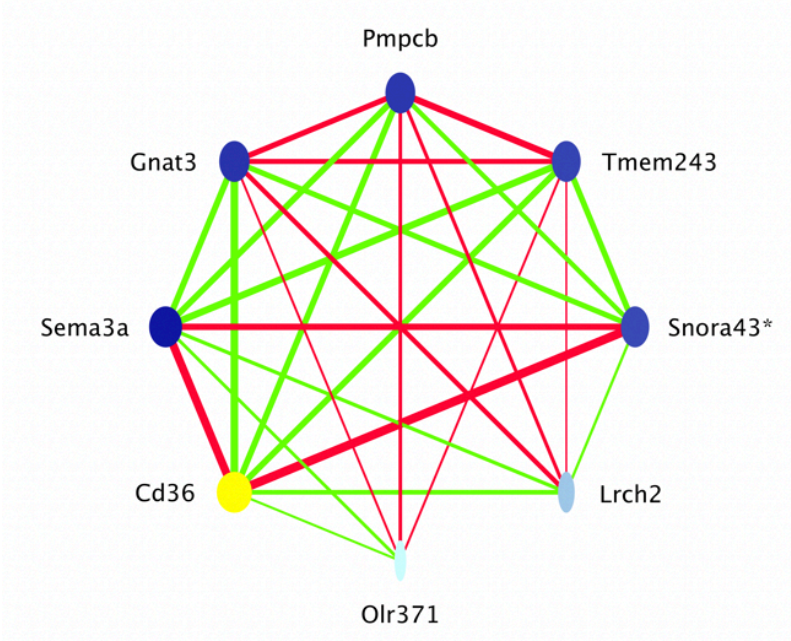
Question 3: Example

- **Terminology**
  - Quantitative trait loci (QTL)
  - Weighted gene co-expression network analysis (WGCNA)
  - Network (think large scale)
  - Co-expression module (think small scale)
  - Eigengene



Question 3: Example





Co-expression Module	Number of Genes in Module	Proportion of Variation Explained by Module Eigengene	Associated Phenotype	Phenotypic QTL*	Correlation Coefficient	P-value	Module Eigengene QTL*
Cornsilk2.2	5	0.71	BAT relative weight	Chr 12: 28.1 Mb (13.2-38.5)	0.42	0.020	Chr 12: 27.3 Mb (26.4-40.5)
Coral4.1	8	0.65	BAT relative weight	Chr 4: 13.3 Mb (0.6-14.7)	-0.43	0.018	Chr 4: 14.5 Mb (13.7-21.8)
Darkseagreen	16	0.54	Glucose incorporation into BAT lipids	Chr 2: 200.0 Mb (167.6-224.1)	0.56	0.001	Chr 2: 205.3 Mb (200.5-207.7)

*QTL peak position with 1.5 LOD Support Interval

Part 3: Common Themes Across Omics Types

Lauren Vanderlinden

Center for Innovative Design & Analysis
colorado school of public health

Common Themes Among All Omics Datasets

1. Data Storage
2. Processing Data
   - Normalization
   - QC plots
3. Multiple Testing Comparisons
4. Enrichment Analysis
5. Validation
6. Questions to keep in mind

Data Storage

- Depends on core/company generating the data
- Raw data backup
- Software can now perform on a compressed file (e.g. fastq.tar.gz)
- Allow 3-4x the amount of the raw data as empty space computing
- Plan for where analysis will be conducted:
  - Local Server
  - Cloud computing
  - Galaxy

RNA-Seq Fastq

Size = # reads * (100 + 2*readLength)

Example: 100 million reads with a read length of 150 = 40G

Methylation Array Idat

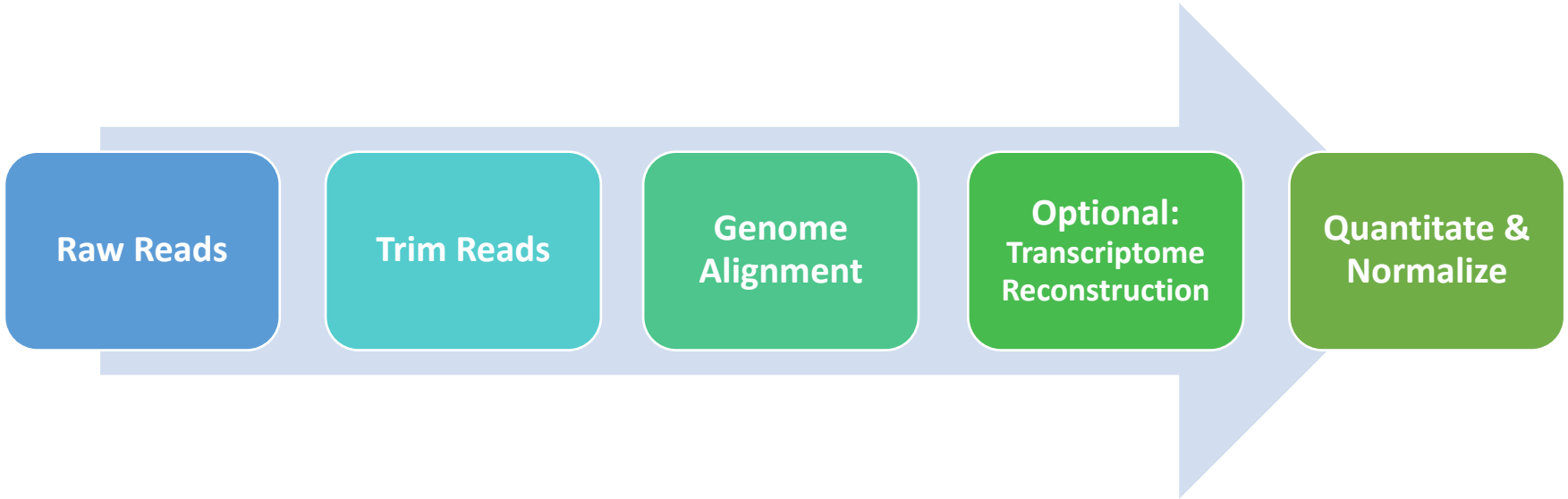
450K ~ 7MB

EPIC ~ 11MB

2 files per sample

Processing Data

- Much more processing time than traditional data
- Raw data is provided as 1 (or 2) files/sample and not a pretty matrix
- Example of RNA-Seq pre-processing steps:



Normalization

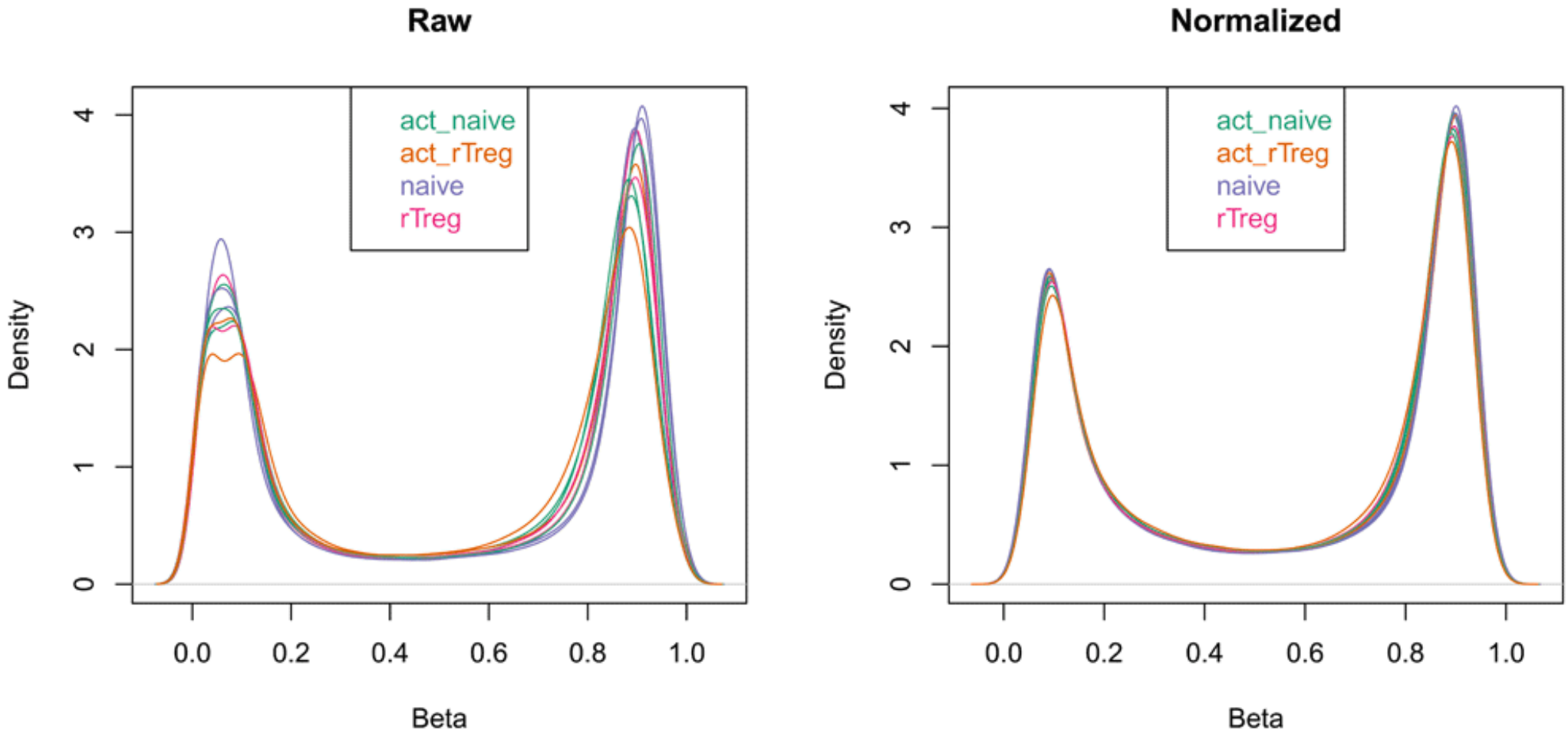
Process of removing (or minimizing) non-biological variation

- RNA-Seq
  - Reads/Fragments Per Kilobase per Million (RPKM/FPKM)
  - Transcripts per Million (TPM)
  - Quantile
  - Weighted Trimmed Mean of Log Expression Ratios (M values) (TMM)
  - DESeq Median of Ratios (geometric mean & scaling factor)
  - Removal of Unwanted Variation (RUV)
  - Surrogate Variable Analysis (SVA)
- Metabolomics (MS):
  - Locally estimated scatterplot smoothing (LOESS)
  - Systematic Error Removal using Random Forest (SERRF)
  - Median
  - Quantile
  - Cross-Contribution Compensating Multiple Standard Normalization (CRMN)
  - SVA
  - RUV
  - R/MSprep evaluates best method for metabolomics MS data
- Methylation Arrays:
  - subset-quantile within array normalization (SWAN)
  - normal-exponential using out-of-band probes (Noob)
  - single-sample Noob (ssNoob)
  - Functional normalization (Funnorm)
- Microarrays:
  - Robust Multichip Average (RMA)
  - Guide to Probe Logarithmic Intensity Error (PLIER)

R/Normalyzer:
A Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets

No Standard Method!

QC Density Plots – Methylation Array Example

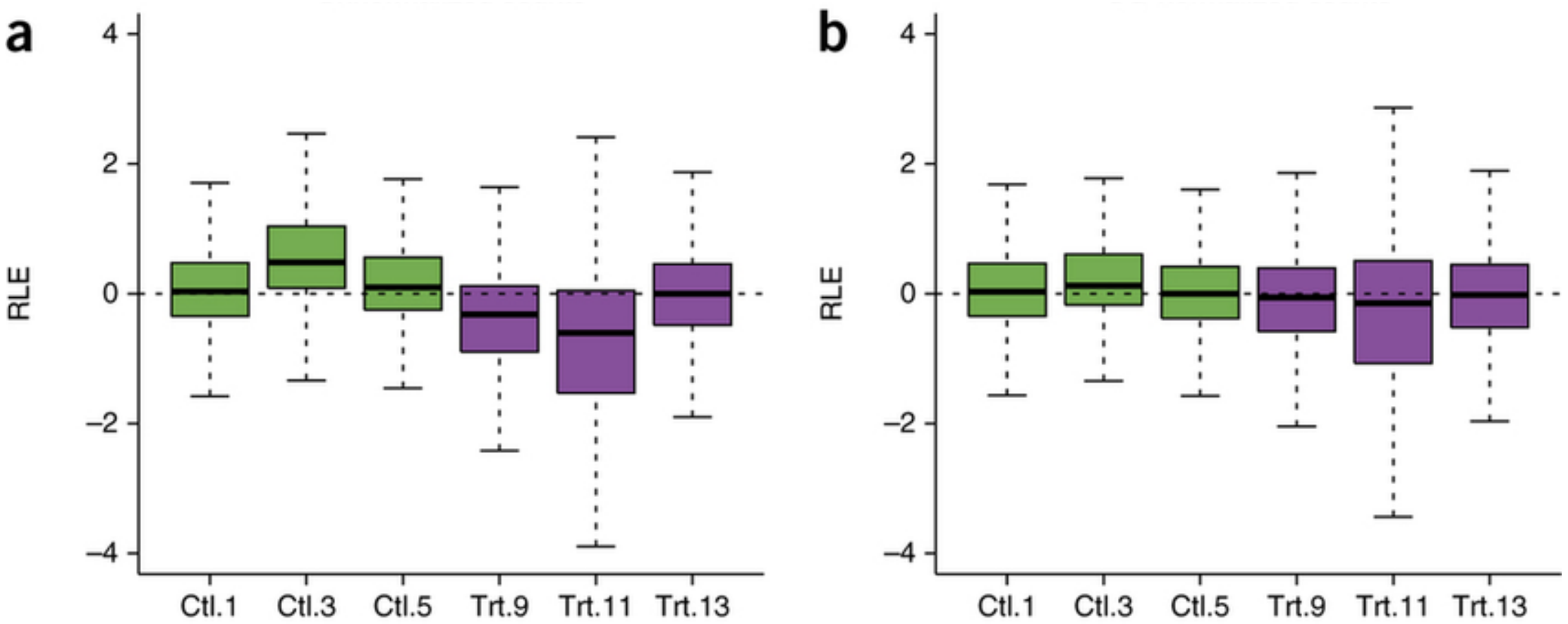


Maksimovic J, Phipson B and Oshlack A. A cross-package Bioconductor workflow for analysing methylation array data [version 3]. F1000Research 2017, 5:1281 (doi: 10.12688/f1000research.8839.3)

QC RLE Plots: Relative Log Expression

RAW Counts

Normalized Counts



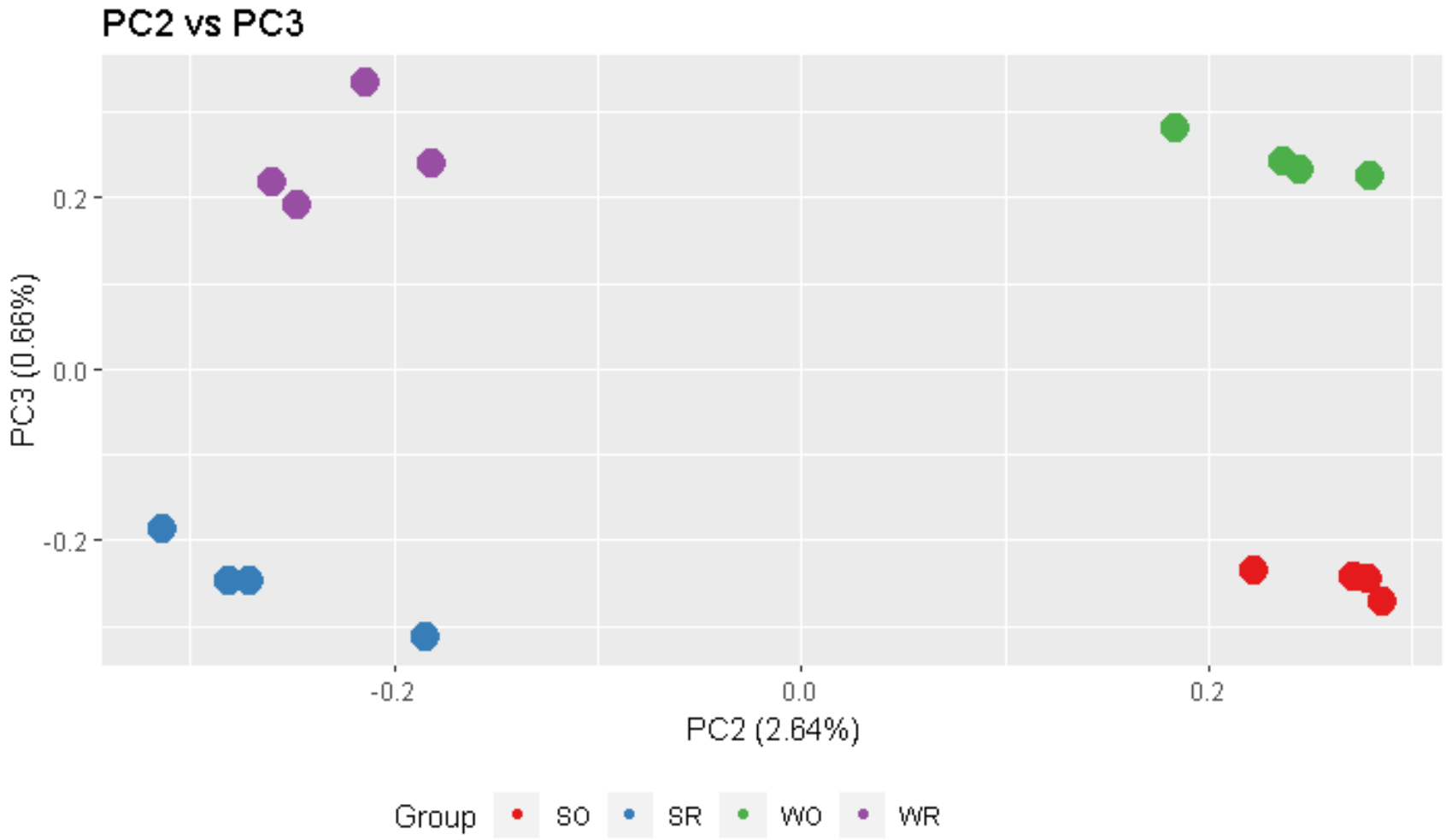
Risso D, Ngai J, Speed T, Dudoit S (2014). "Normalization of RNA-seq data using factor analysis of control genes or samples." *Nature Biotechnology*, **32**(9), 896–902.

QC PCA Plots

Clustering by different factors:

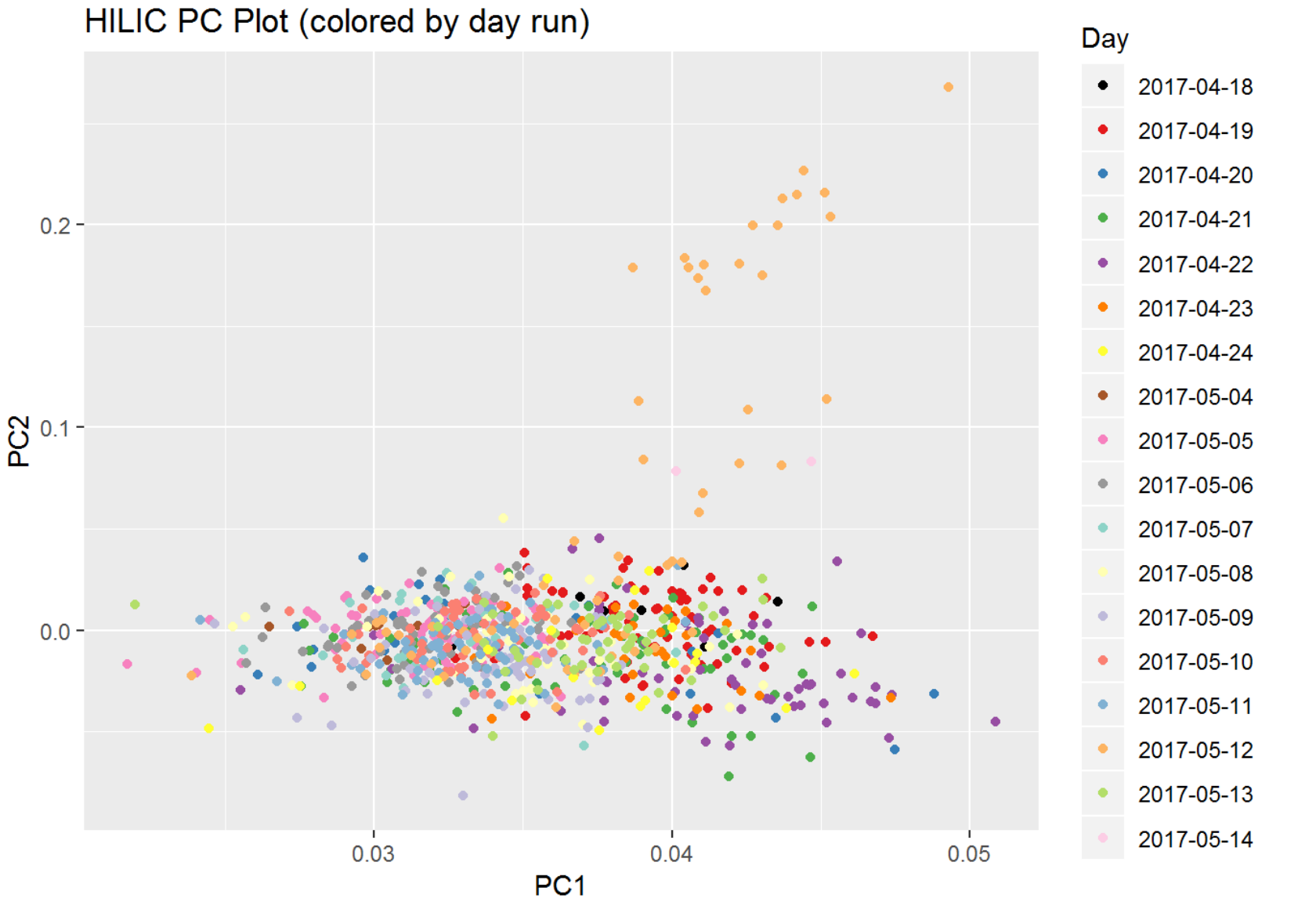
X-axis separating by tissue

Y-axis separating by strain



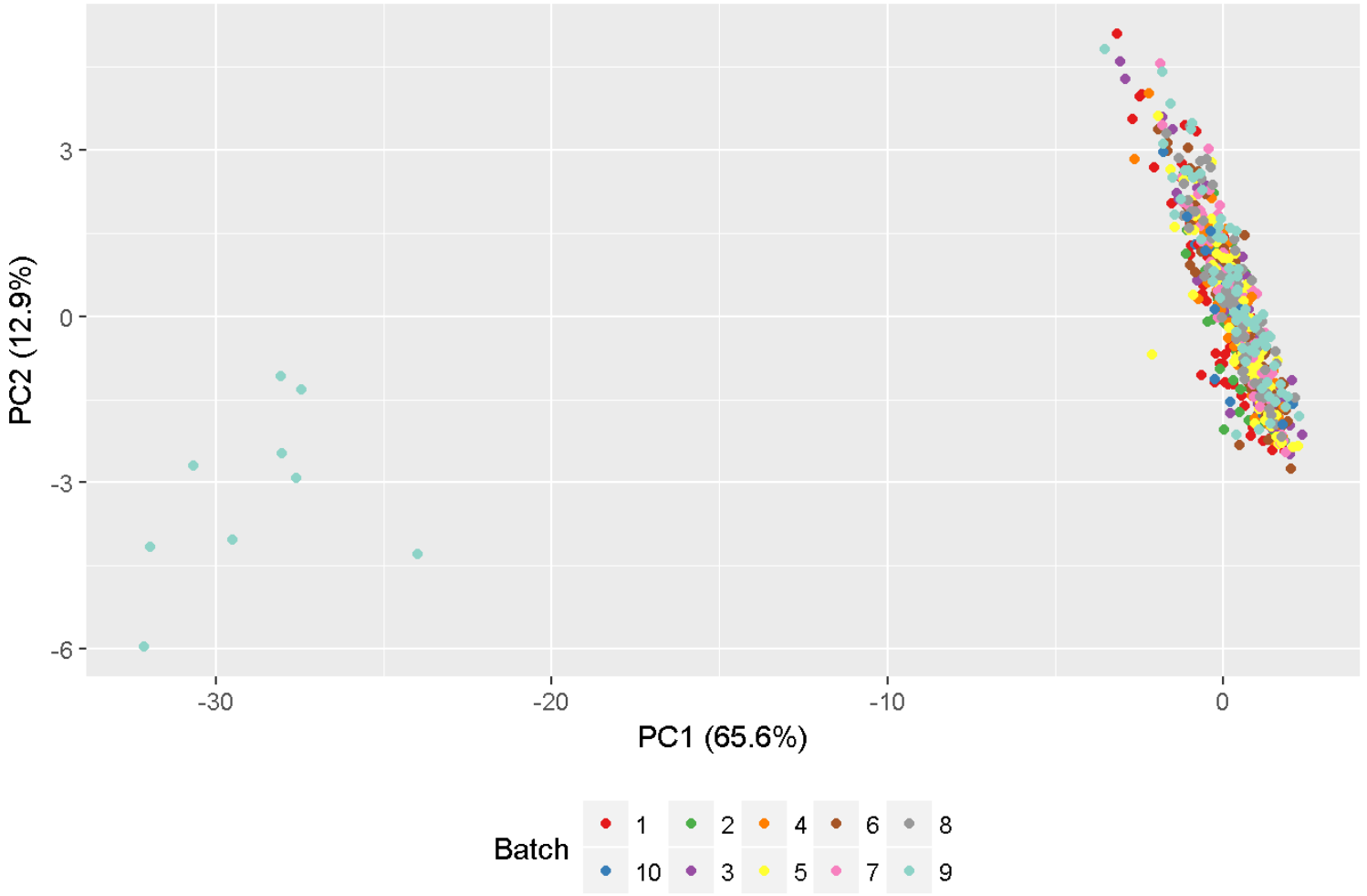
Batch Effects

Samples colored
by batch
(date run)

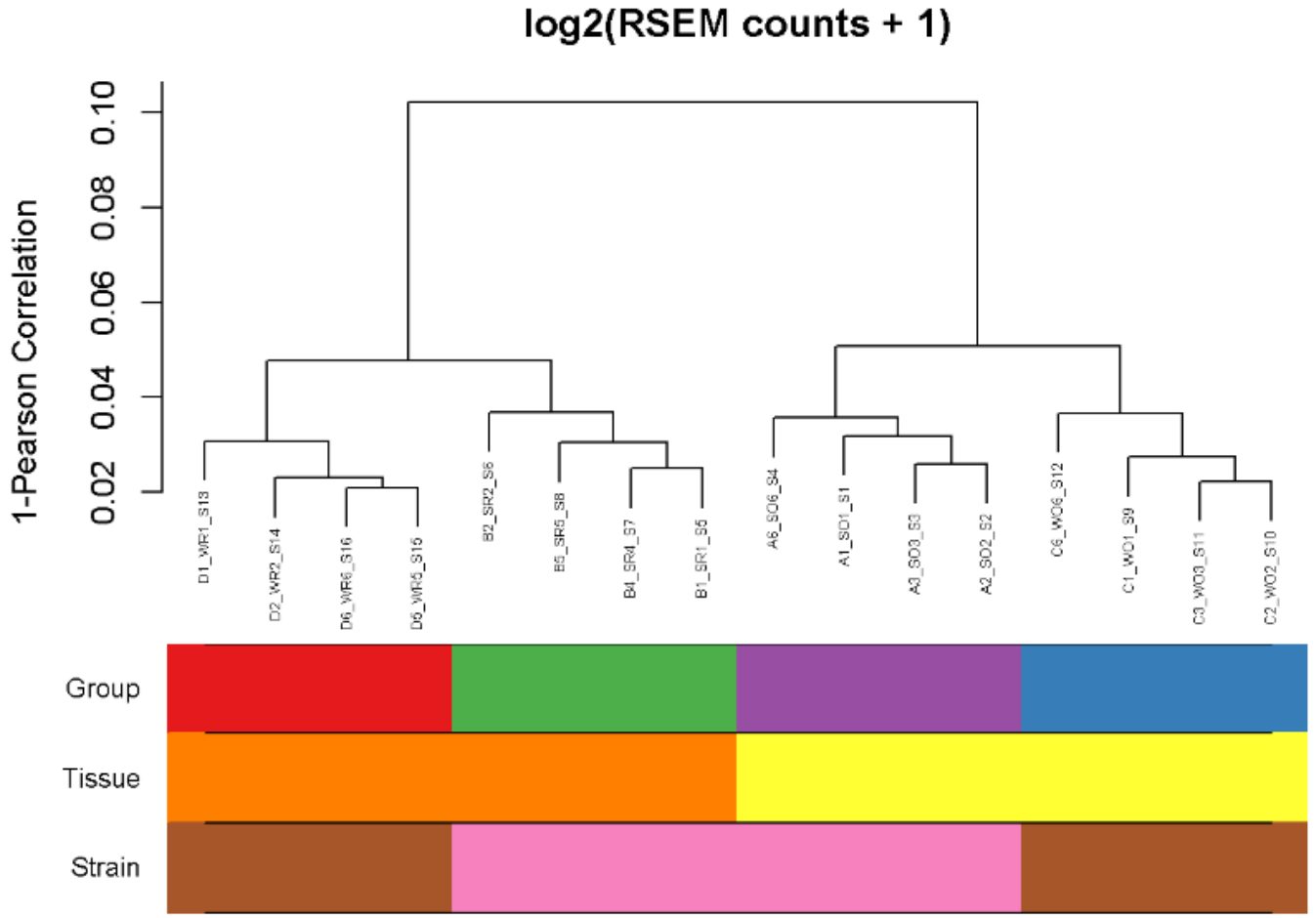


Sample Level QC

Can help identify poor quality samples

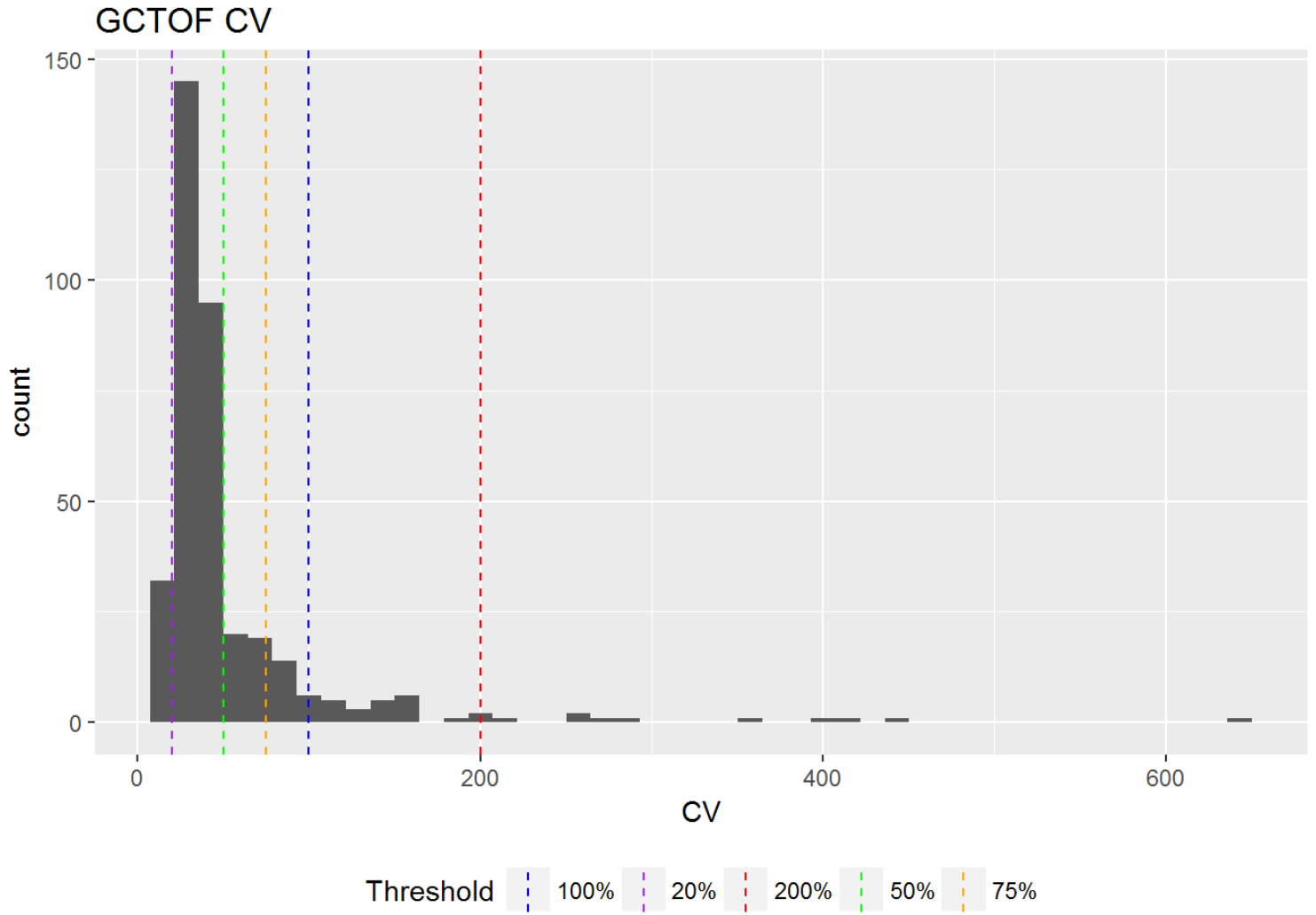


QC Dendrograms



Feature Level QC

- Detection above background threshold
- Coefficient of variation (CV) threshold

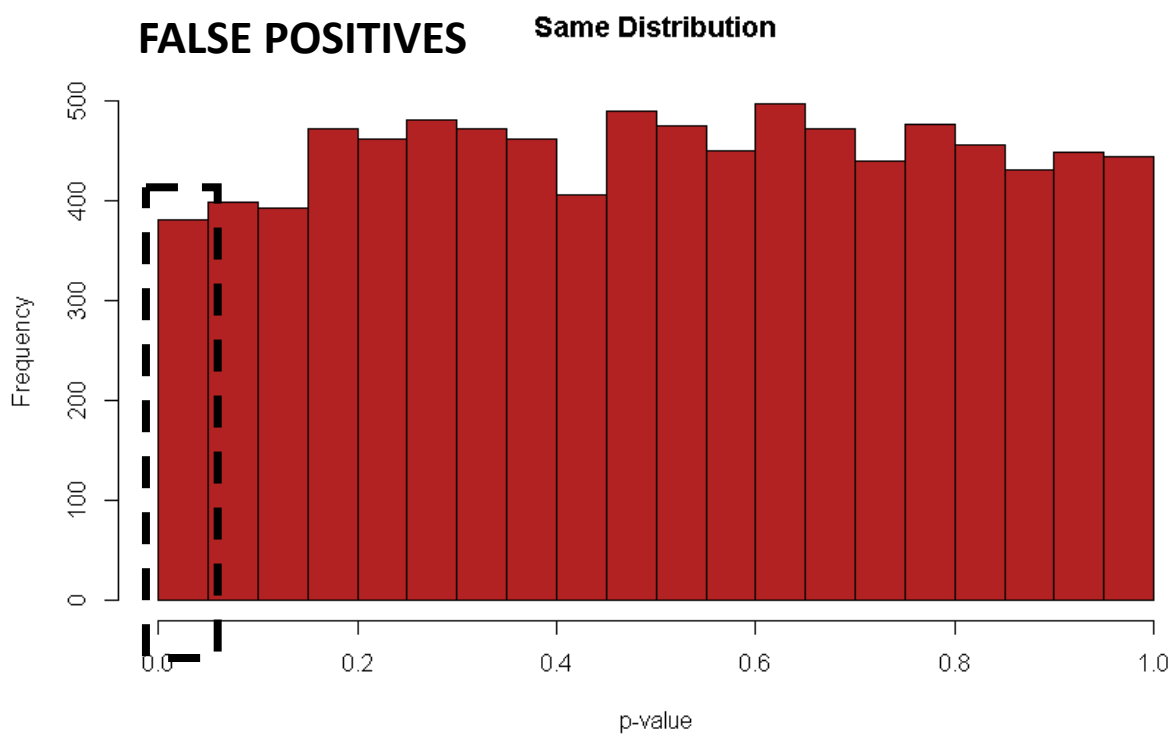


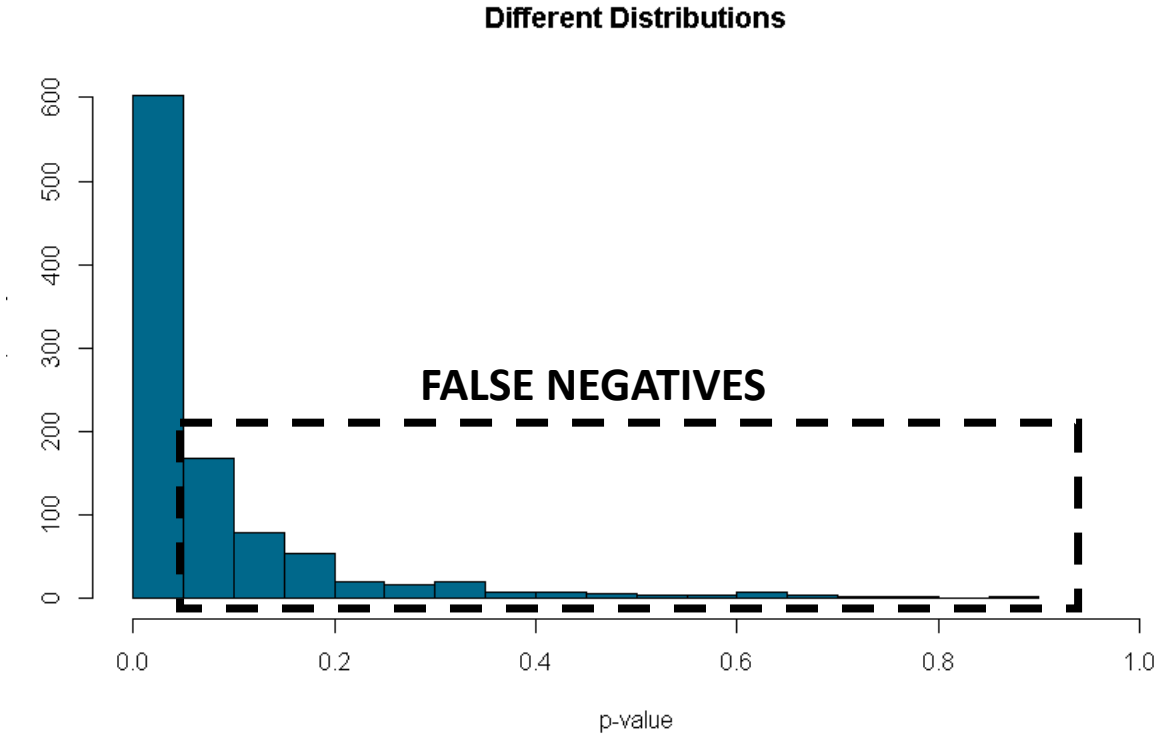
Multiple Testing

- Same statistical model on every feature
  - Example: 20,000 genes, then you have 20,000 tests
  - If you leave alpha = 0.05 you would expect 1,000 false positive results (Yikes!)
- Perform correction for multiple testing
- All methods are assuming all tests are independent
- Bonferroni
  - Multiple the p-value by the # of tests performed
  - Most conservative and considered too harsh

False Discovery Rate (FDR)

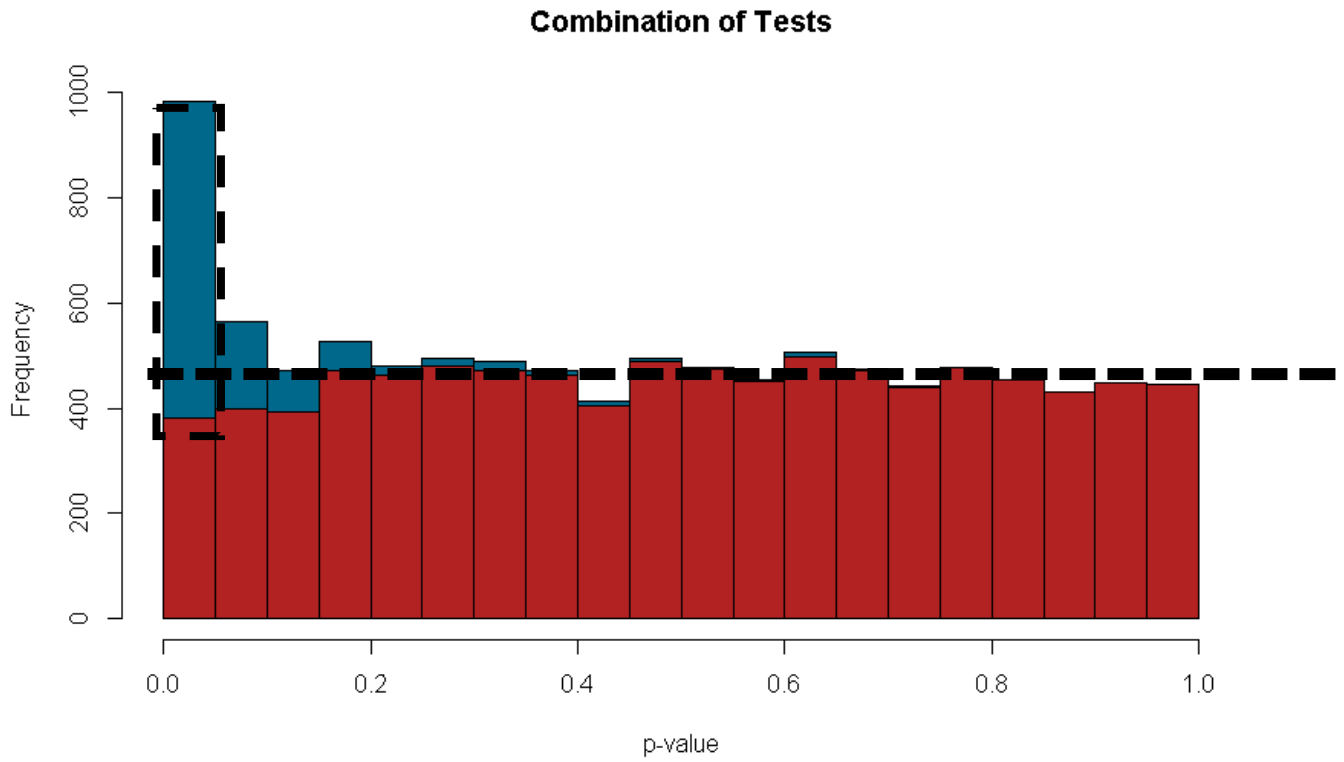
- Adjusts each p-value differently depending on rank





False Discovery Rate (FDR)

- Tries to estimate your distribution of non-significant p-values (makes power analyses difficult)



Enrichment & Over-representation Analysis

- Big picture of system level
- Static (Over-representation)
- Fluid (Enrichment)
  - Gene Set Enrichment Analysis (GSEA)

GENEONTOLOGY
Unifying Biology

KEGG PATHWAY Database
Wiring diagrams of molecular interactions, reactions and relations

PANTHER
Classification System

DSigDB Drug SIGnatures DataBase
Collection of Annotated Drug / Compound Gene Sets

	Candidates	Genome (background)
In Pathway		
Not in Pathway		

Background Set is Important

- What is present in study sample type
  - Example: if looking at lung tissue you would not expect all genes to be expressed in the lung regardless of study design
- Arrays certain genes are over-represented
  - Various number of probes/gene
  - Example: Illumina's EPIC array there is a range of 1 to 1,487 probes/gene, with a median of 20 probes per gene
    - R/missMethyl takes into account how many probes are designed on array

	Candidates	Genome (background)
In Pathway		
Not in Pathway		

Validation

- Reproduce quantitation:
  - High-throughput methods are not the gold standard in quantitation
  - Gene expression: qRT-PCR
  - Methylation: Pyrosequencing
  - Metabolomics: Targeted or internal standard
- Functional validation:
  - Gene knock-down or knock-out methods
  - Use different dataset (publically available) show this effect
- Multi Omics Integration:
  - Gene candidate in both ChIP-Seq and RNA-Seq
  - Correlation among methylation and gene expression

Know Your Biology Question Prior to Conducting Omics Experiment: RNA-Seq example

- Do you want bulk or cell-specific level?
  - Single cell sequencing vs bulk sequencing
- What type(s) of RNA do you want to look at?
  - mRNA only (polyA selection or possibly Tag-Seq)
  - Long non-coding and other longer types (total RNA)
  - miRNA and other smaller RNAs (small RNA processing different than the others and these need to be measured on separate sequencing runs)
  - Rare RNA types like fusion genes? (longer paired-end reads)
- What level are you looking on quantitating your data on?
  - Gene level only
  - Isoform specific level
  - Reconstruct your own transcriptome (need deep sequencing)

Known Your Analysis Type Prior To Conducting Omics Experiment

- Simple differential expression at a gene-centric level
  - Easiest processing
- More complex models
  - More processing time
- Data driven network analysis
  - Need a higher sample size
  - WGCNA suggests at a MINIMUM 20 samples
- Machine learning
  - Needs the highest sample size (hundreds)

Discussion: Starting your study

1. Talk to core to plan experiment & discuss
   - Technology
   - Protocol options
   - Timeline
   - Sample handling and prep

2. Plan for computing needs (software, hardware) & data storage

Discussion: Starting your study

3. Work with biostatistician/bioinformatician especially if
   - More complex study design (e.g., multiple time points, biological/treatment groups)
   - More complex analyses (e.g., alternative splicing, transcriptome reconstruction, gene fusion)
4. Budget time and effort for data analysis (biggest bottleneck)

Discussion