Big Data Seminar Series: OMICS Data

January 20, 2021

Katerina Kechris & Lauren Vanderlinden

Departments of Biostatistics & Informatics and Epidemiology

Colorado School of Public Health

Center for Innovative Design & Analysis
colorado school of public health

Outline

1. Current omics technologies (Kechris)

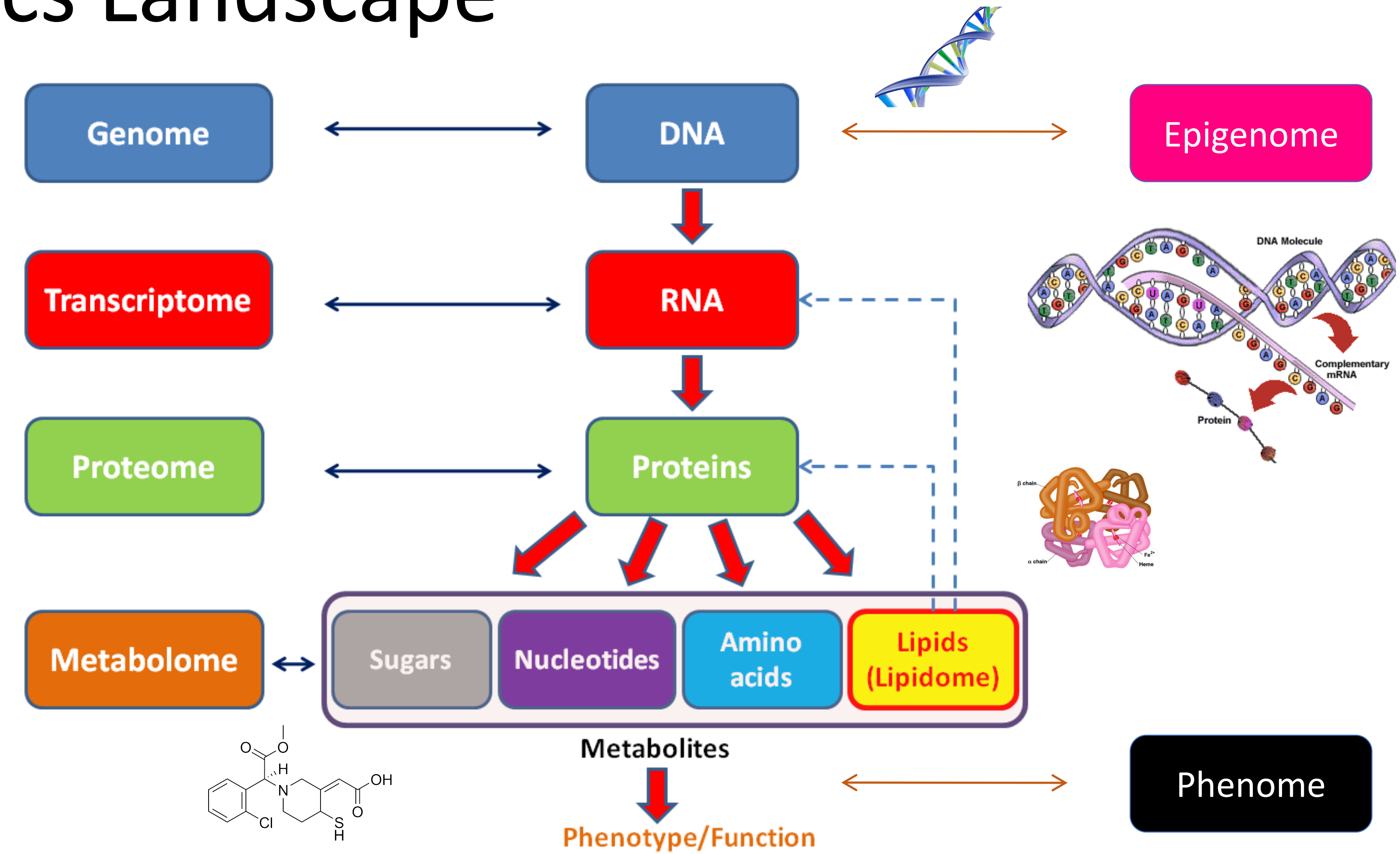
2. Examples of analyses (Kechris)

3. Common statistical themes in omics data analysis (Vanderlinden)

4. Questions and discussion to plan your omics study

Part 1:
Technologies

Omics Landscape



Genome

DNA

Epigenome

Transcriptome

RNA

Proteome

Proteins

Metabolome

Sugars

Nucleotides

Amino acids

Lipids (Lipidome)

Metabolites

Phenotype/Function

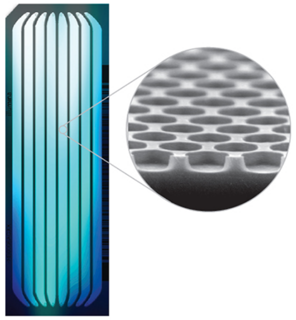
Phenome

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

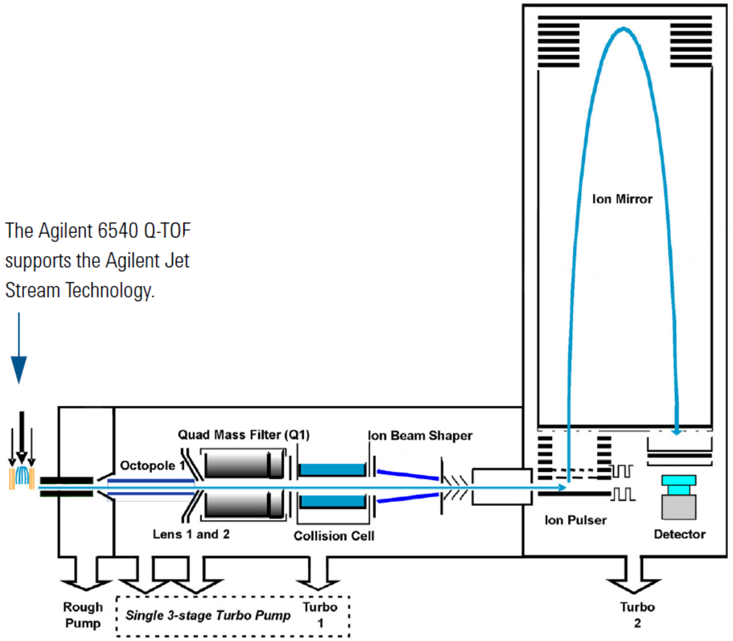
1. Microarrays (RNA/DNA)

2. Sequencing (RNA/DNA)



3. Mass-spectrometry (proteins/metabolites)

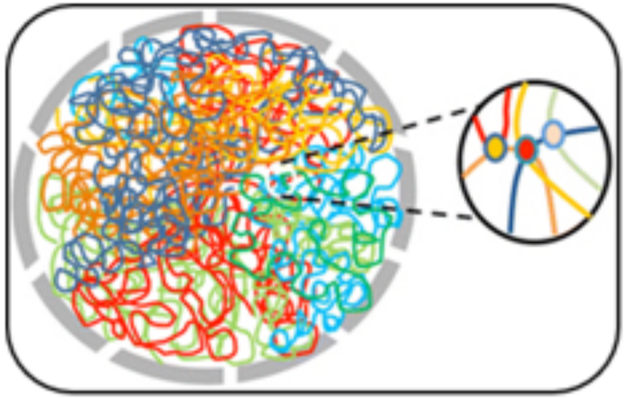
The Agilent 6540 Q-TOF supports the Agilent Jet Stream Technology.



https://www.thermofisher.com; https://www.illumina.com; https://www.creative-proteomics.com

DNA

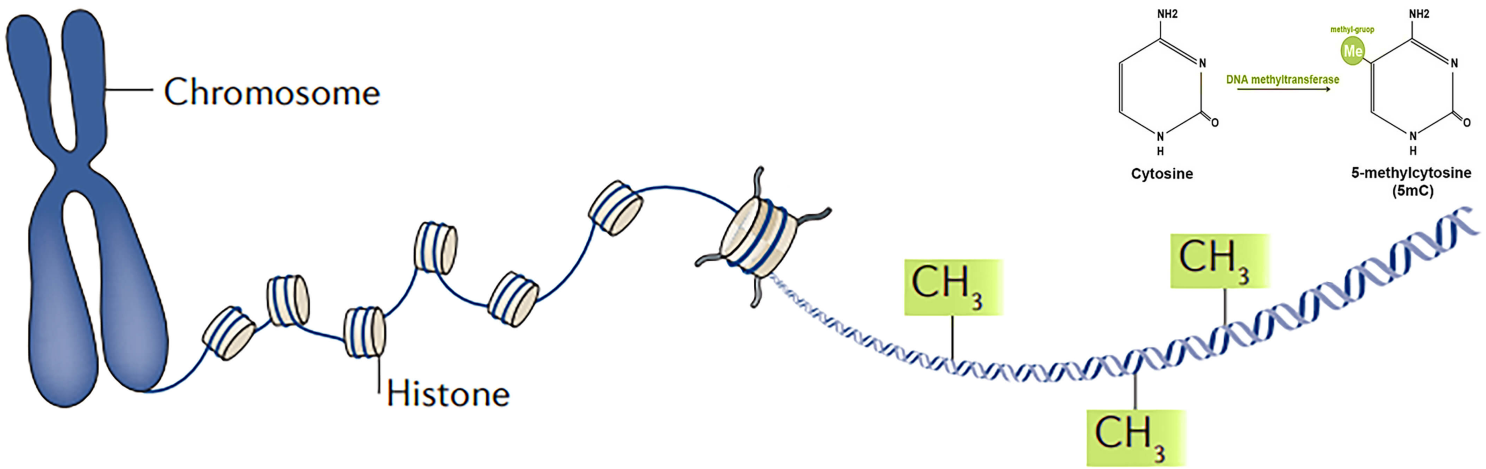
- Genome (whole genome sequencing, WGS)
  - Within and across population
  - Across species
- Exome
- Single nucleotide polymorphisms (SNPs)
- Chromosome conformations (3C/Hi-C)



https://www.creativebiomart.net/

DNA Modifications & Interactions

- DNA methylation (epigenome) (methyl-Seq)
- Histone modifications (epigenome) (ChIP-Seq)
- DNA binding proteins (e.g., transcription factor)
- Chromosome accessibility (ATAC-Seq)

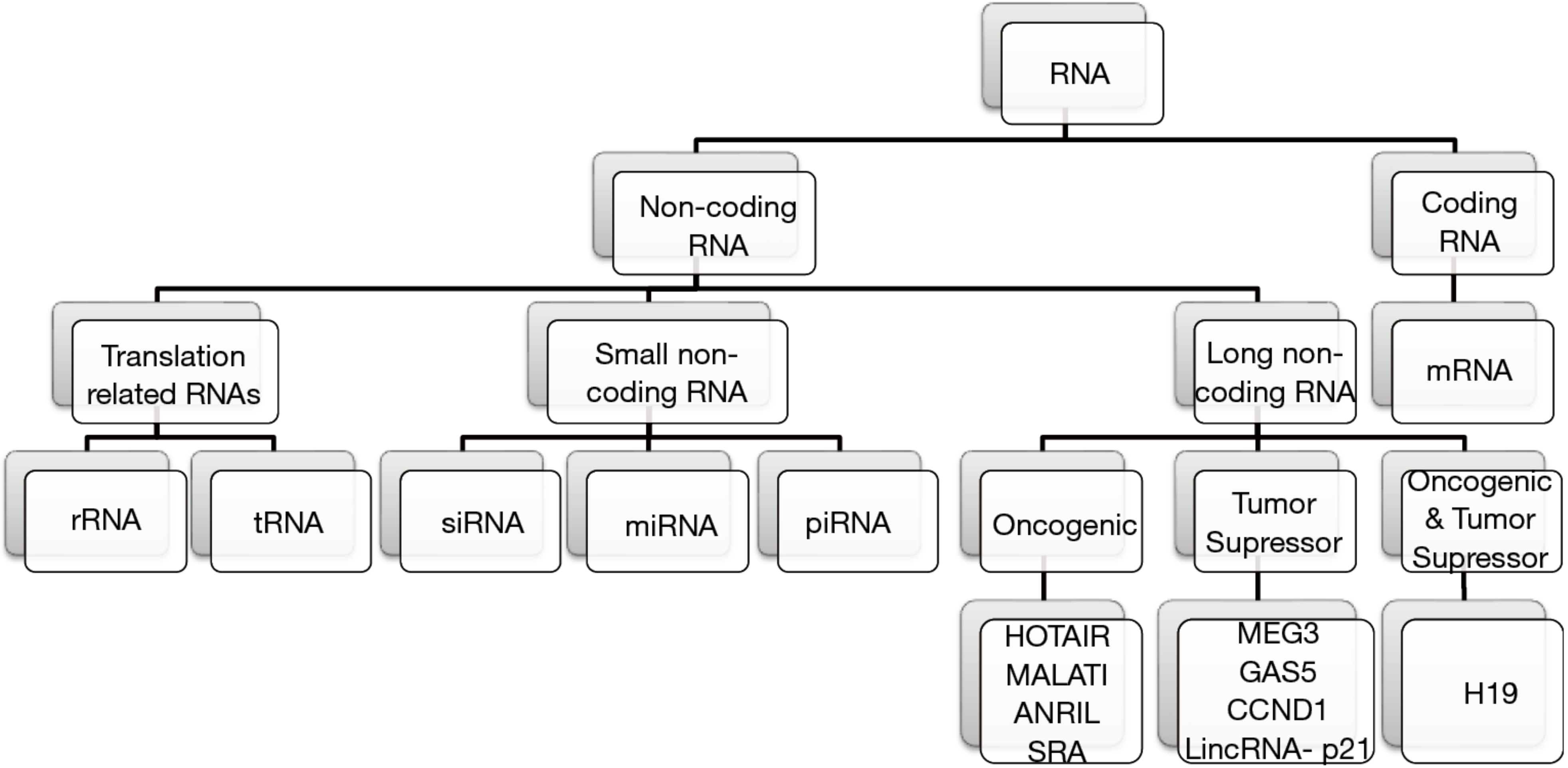


https://www.creativebiomart.net/

RNA

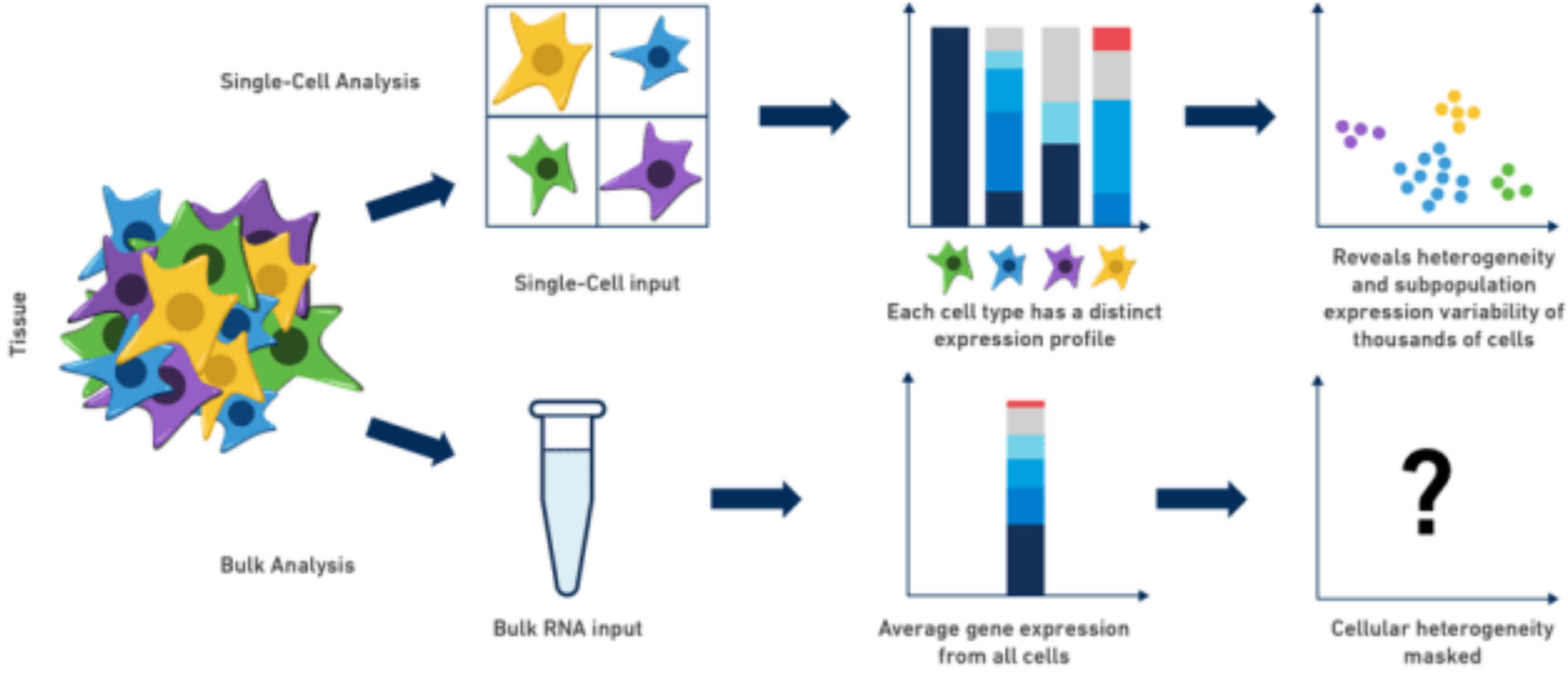
- mRNA (transcriptome) (RNA-Seq)
- RNA binding proteins (e.g., splicing factors) (CLIP-Seq)
- Methylation RNA (epitranscriptome) (MeRIP-Seq)
- Other types
  - miRNA, lncRNA, etc
  - 16s rRNA (microbiome)

RNA

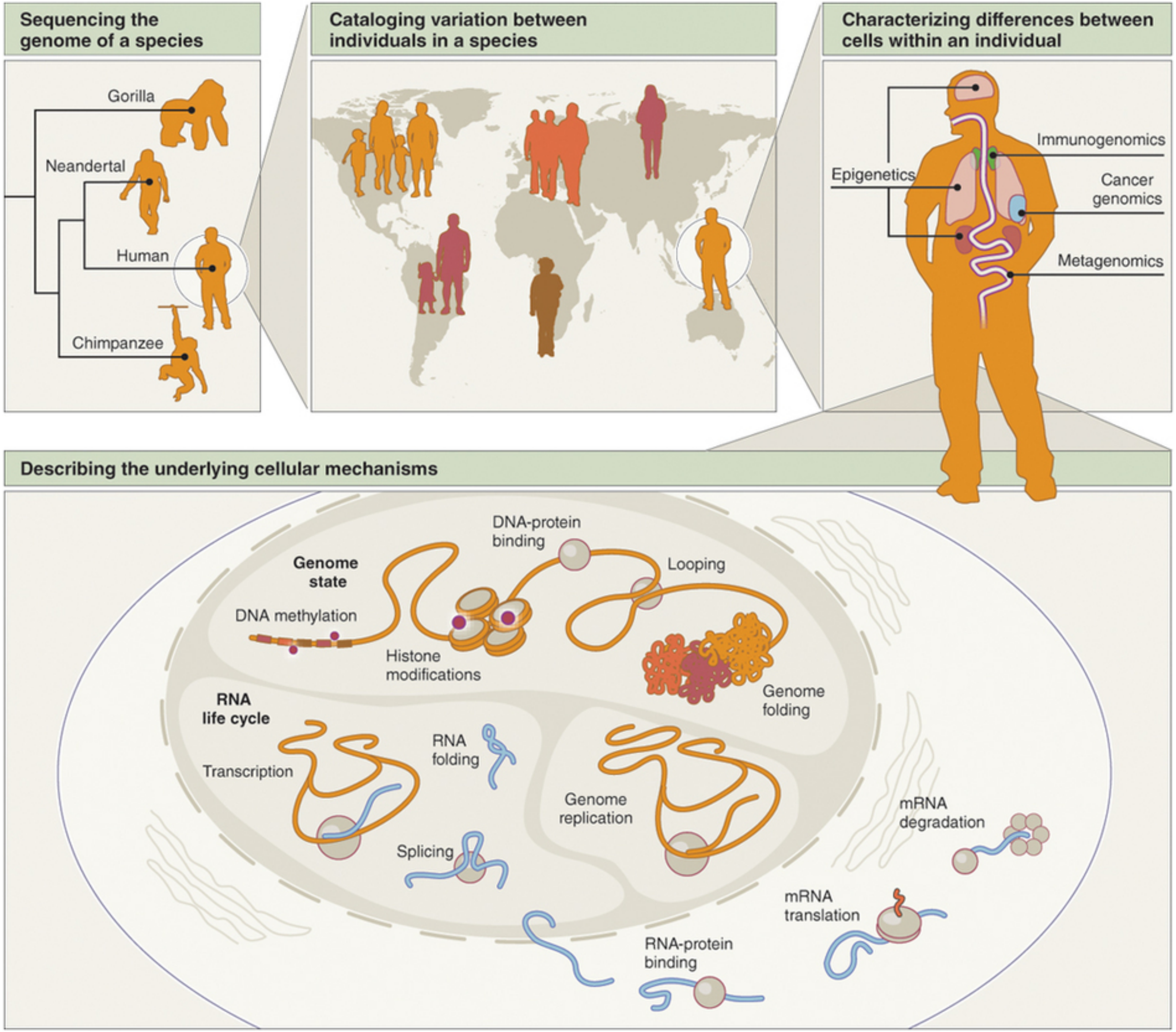


Diamantopolous et al., 2018 ATM

Single-cell vs Bulk Cell


Single-Cell Analysis
Tissue
Single-Cell input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells
Bulk Analysis
Bulk RNA input
Average gene expression from all cells
?
Cellular heterogeneity masked

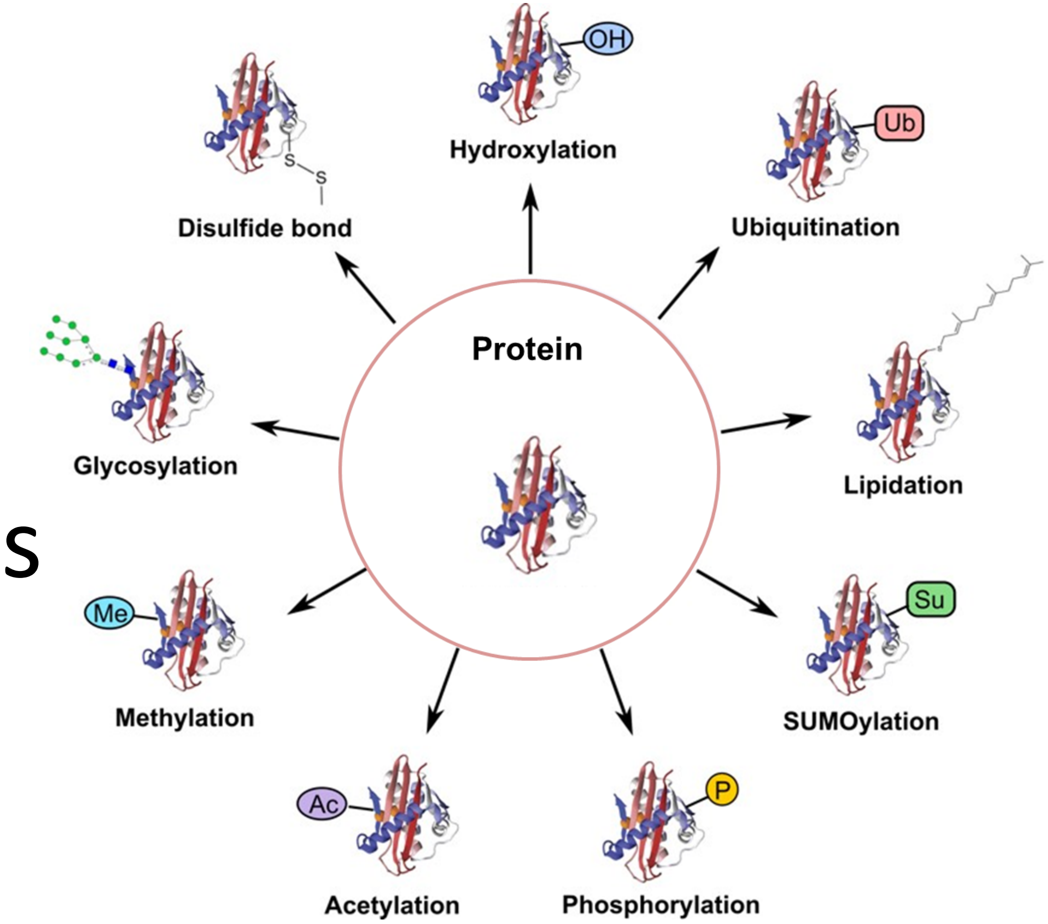
https://www.10xgenomics.com/



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

Proteins

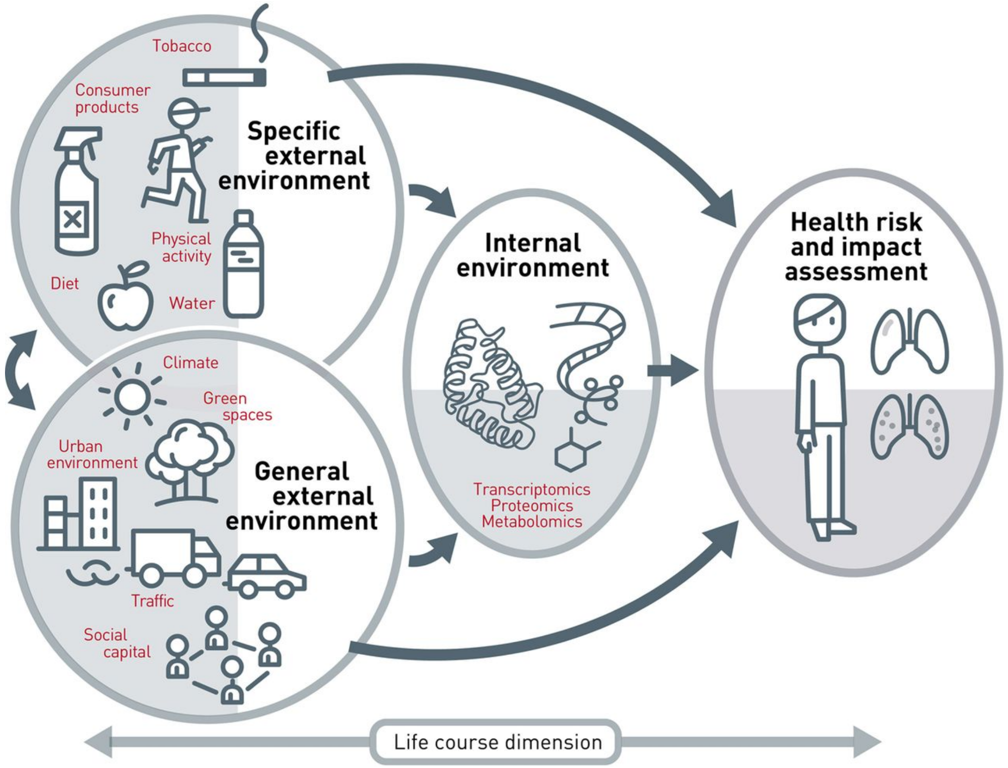
- Abundance
- Structure
- Protein-protein interactions
- Post-translation modifications (e.g., phosphoproteomics, glycoproteomics)



https://www.polyquant.com/discovery-proteomics/ https://www.creative-proteomics.com/

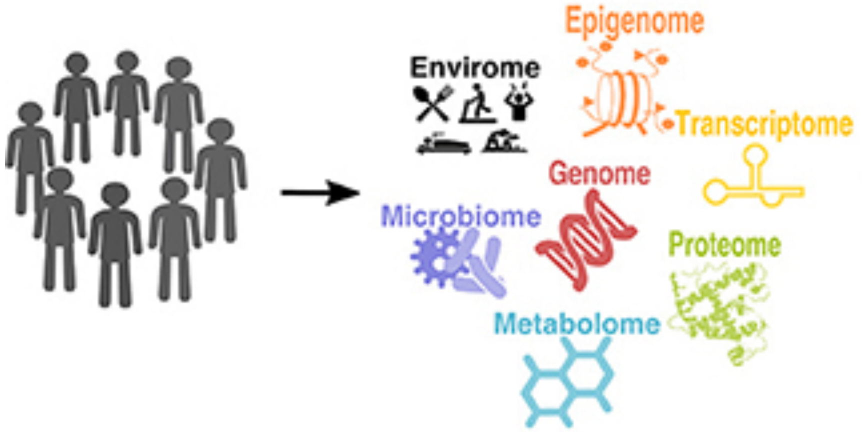
Metabolites

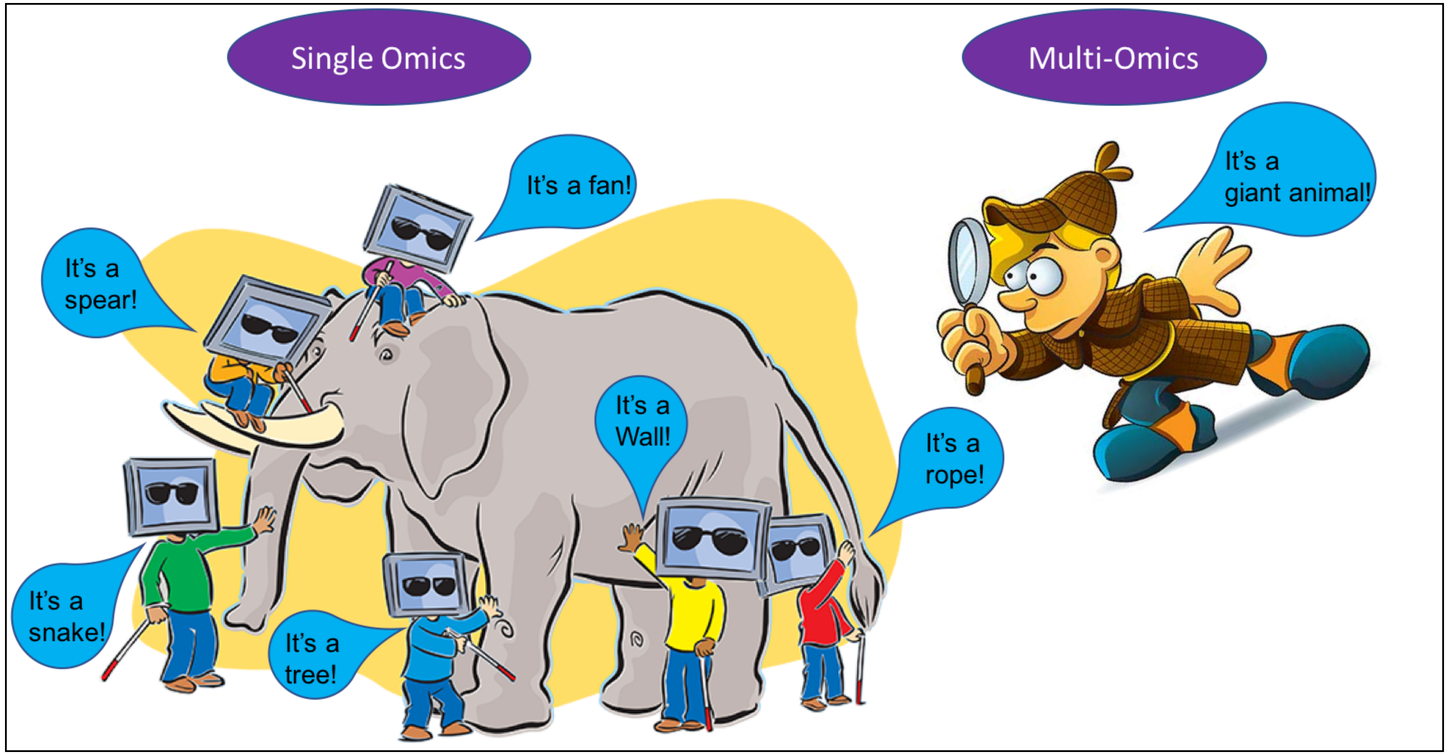
- Types of small molecules
  - Lipids – lipidomics
  - Exogenous factors– exposome
  - Diet/drugs - nutrigenomics
- Toxicology (changes due to chemical)
- Metabolic reactions (e.g., fluxomics)



https://thorax.bmj.com/content/69/9/876

Multi-Omics

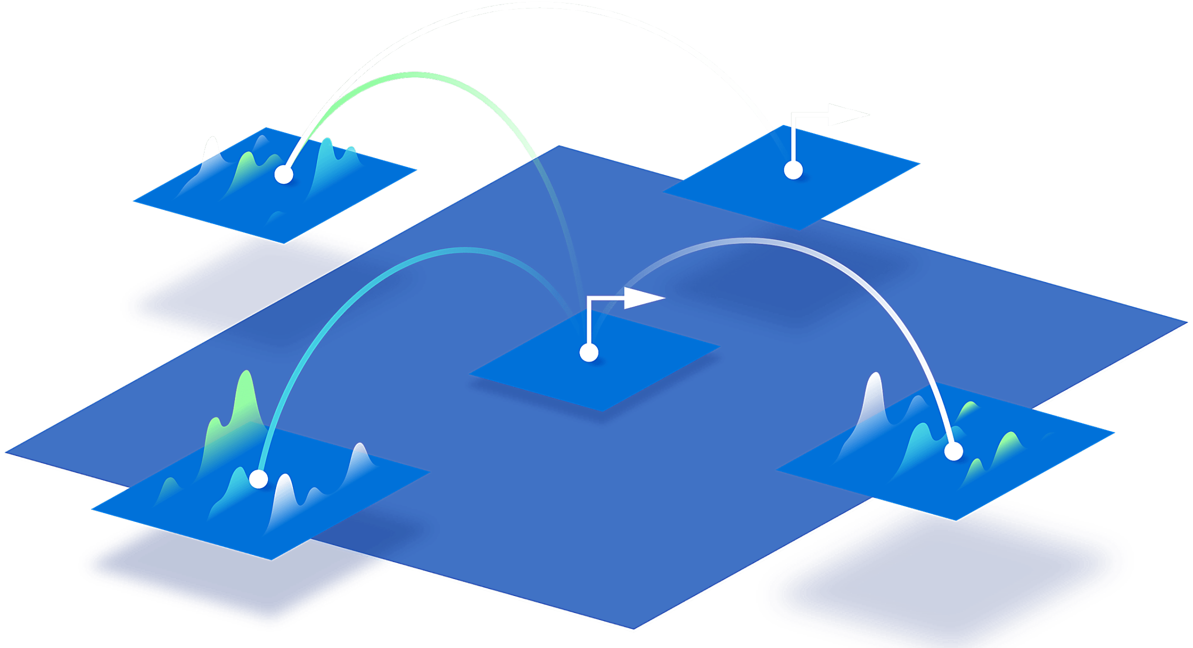




http://melgen.org/multi-omics-approach/ Vilne & Shunkert 2018

Multi-Omics

From same cell, simultaneous detection of mRNA & chromatin accessibility (e.g., Multiome 10X Genomics)

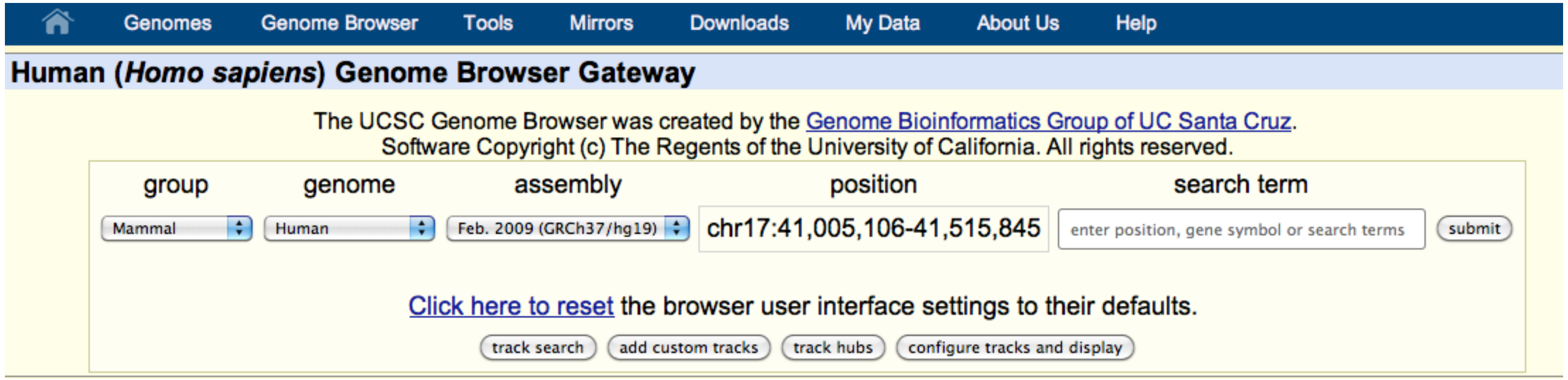


Large-scale Projects & Databases

ENCODE

mod ENCODE

THE CANCER GENOME ATLAS



Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

group	genome	assembly	position	search term
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr17:41,005,106-41,515,845	enter position, gene symbol or search terms

Click here to reset the browser user interface settings to their defaults.

Large-scale Projects & Databases

hmdb

The Human Metabolome Database

HUPO

HUMAN PROTEOME ORGANIZATION

translating
the code of life

Multiple-Cohorts & Populations

NIH National Institutes of Health
All of Us Research Program

NHLBI
TOPMed

COMETS
COnsortium of METabolomics Studies

PACE
Pregnancy And Childhood Epigenetics

NIH National Institutes of Health

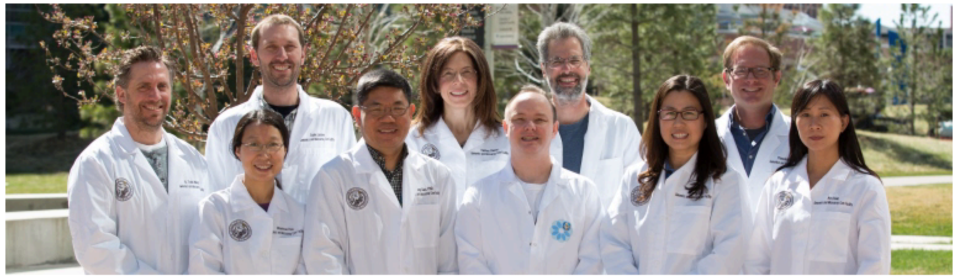
Home » Research & Training

ENVIRONMENTAL INFLUENCES ON CHILD HEALTH OUTCOMES (ECHO) PROGRAM

Resources @ AMC

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

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



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:

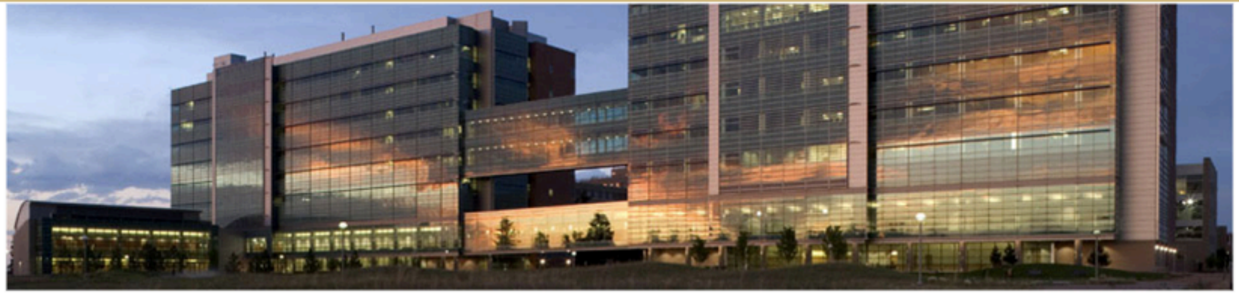
- Next Generation Sequencing:
  - Illumina HiSeq 2500/4000 sequencing
  - Illumina MiSeq sequencing
  - LifeTech IonPGM sequencing
- DNA Microarray:
  - Illumina BeadArrays
  - Agilent Microarrays

Address:

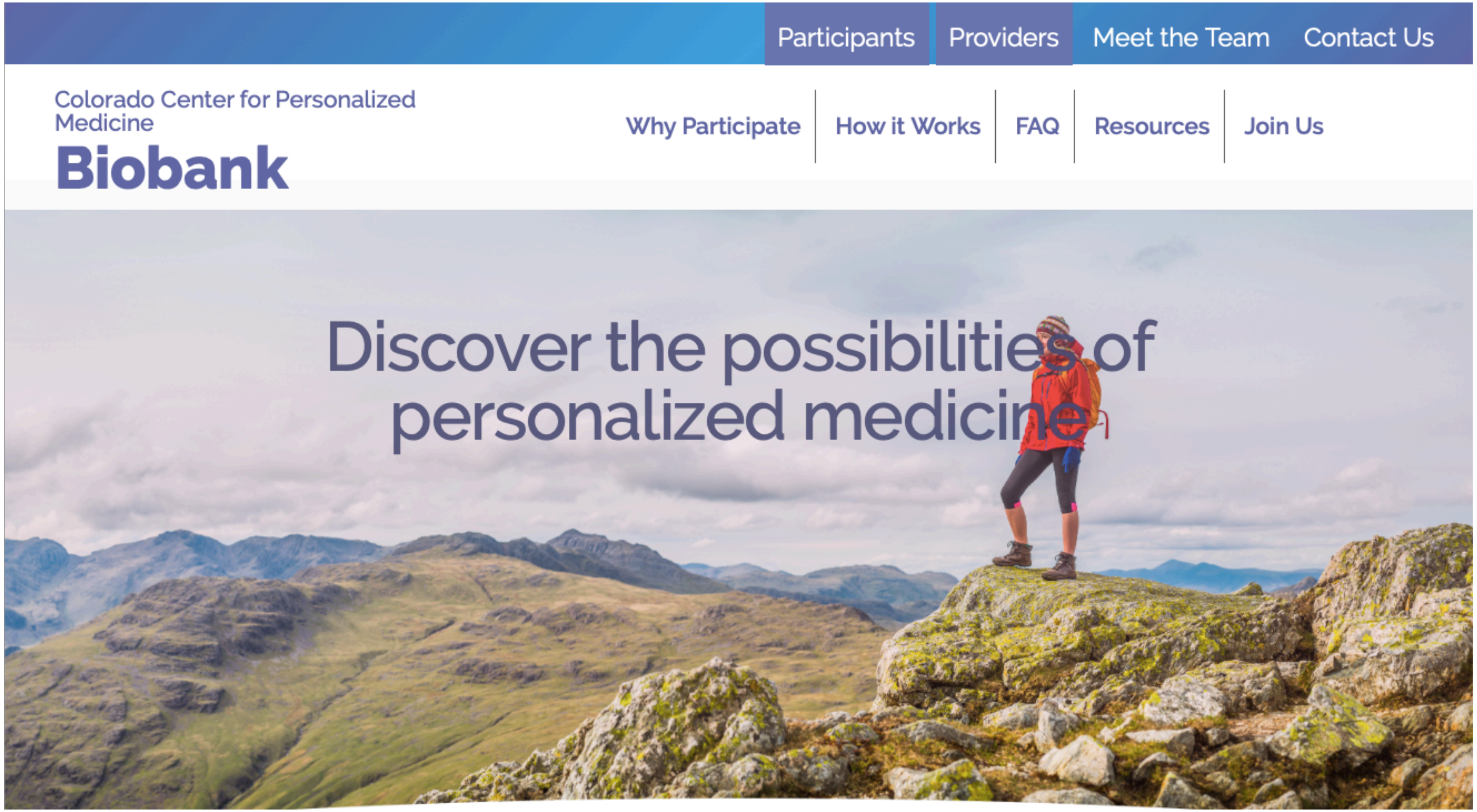
Location/Fed Ex

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

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



University of Colorado School of Medicine Biological Mass Spectrometry Facility



Colorado Center for Personalized Medicine

Biobank

Discover the possibilities of personalized medicine

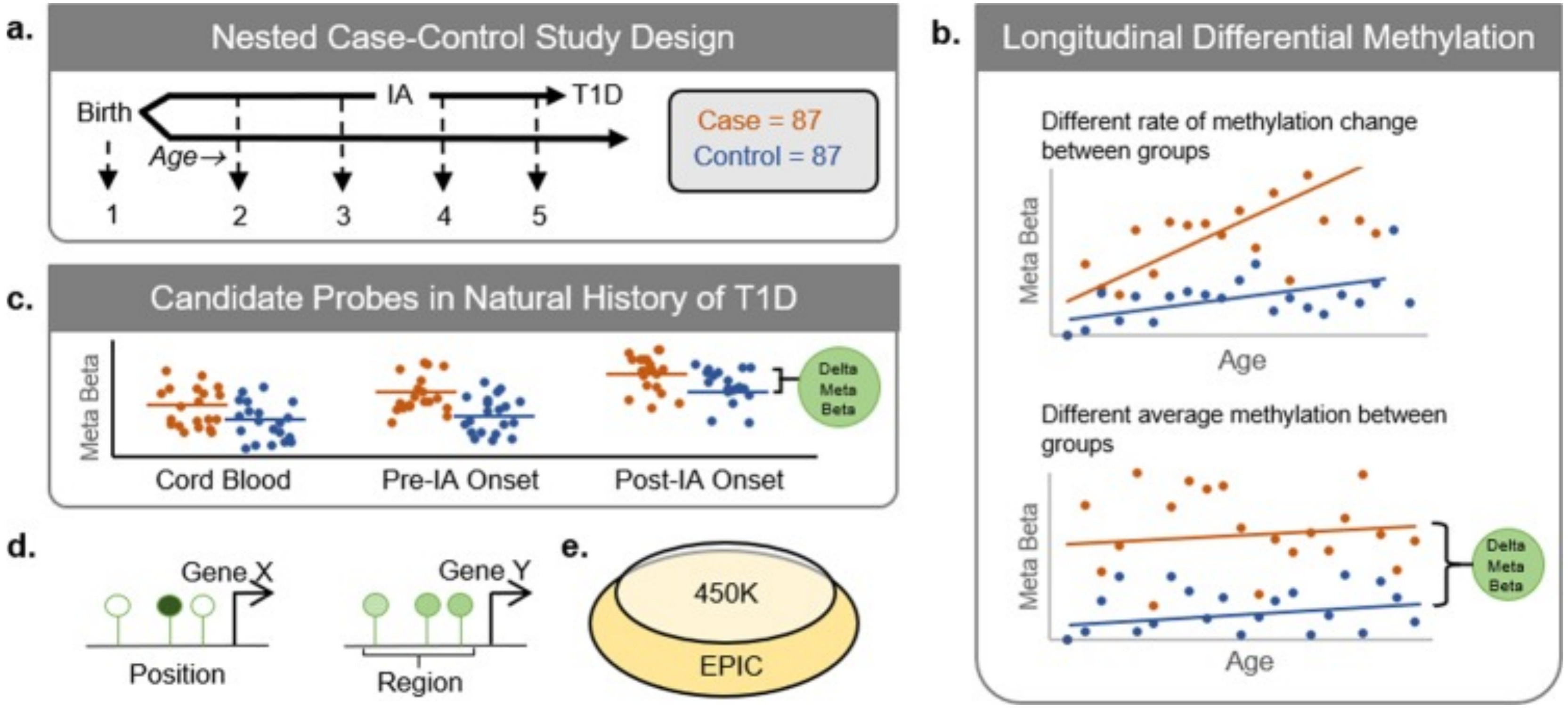
Part 2:
Examples

Study 1: Epigenetics & Type 1 Diabetes (T1D)

with Jill Norris (Epi, CSPH)

- DNA methylation link between genetic susceptibility & environmental exposure in T1D
- Most studies on individuals already diagnosed with T1D
- Goal: Study pre-disease DNA methylation changes associated with later development of T1D

Study Design: DNA methylation measured prior to onset of clinical T1D from Diabetes Autoimmunity Study in the Young (DAISY) cohort (n=174)
Platform: Illumina BeadChip Array
Analysis: longitudinal mixed model, meta-analysis, region-based analysis


a.
Nested Case-Control Study Design
Birth
IA
T1D
Age→
1
2
3
4
5
Case = 87
Control = 87
b.
Longitudinal Differential Methylation
Different rate of methylation change between groups
Meta Beta
Age
Different average methylation between groups
Meta Beta
Delta Meta Beta
Age
c.
Candidate Probes in Natural History of T1D
Meta Beta
Delta Meta Beta
Cord Blood
Pre-IA Onset
Post-IA Onset
d.
Gene X
Position
Gene Y
Region
e.
450K
EPIC

Johnson et al., (2020) Longitudinal DNA methylation differences precede type 1 diabetes *Scientific Reports*

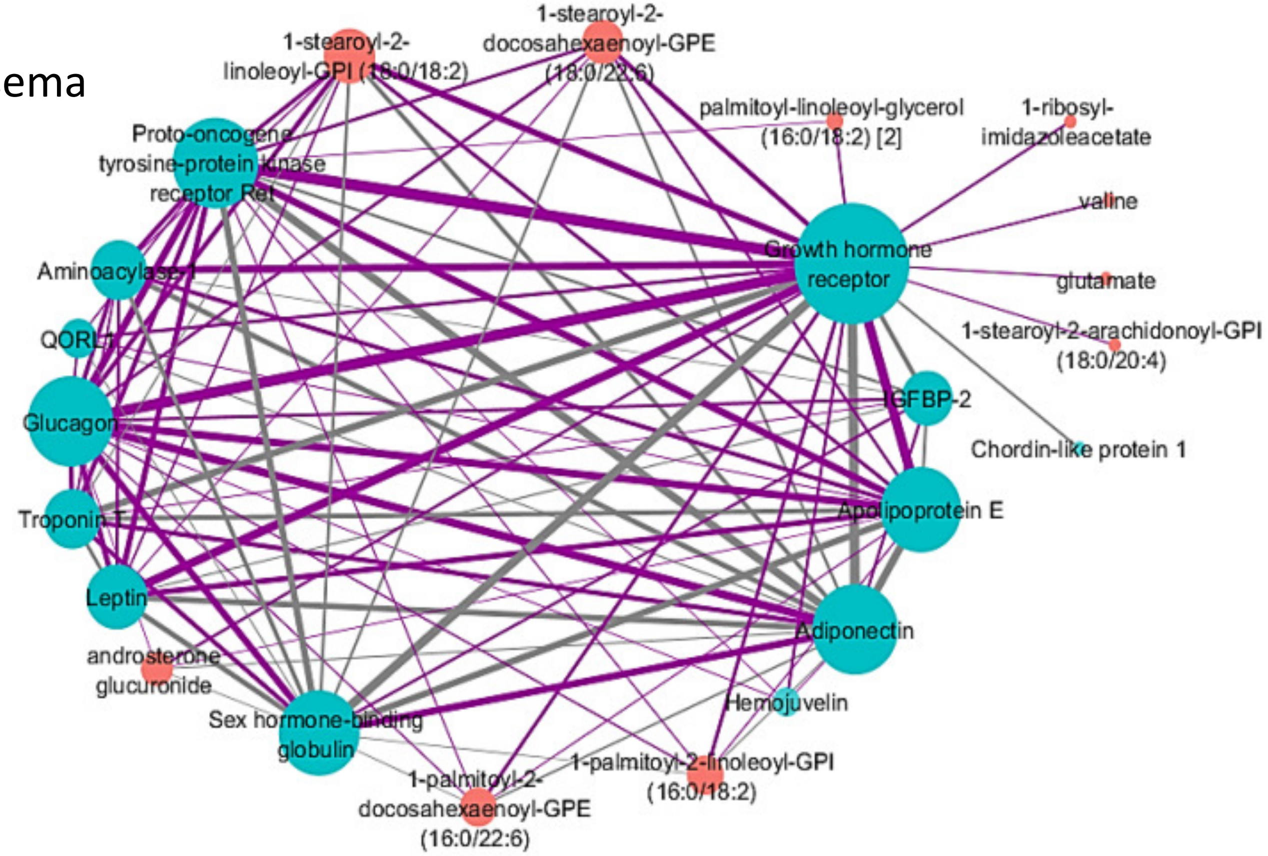
Study 2: Protein-Metabolite Networks in Chronic Obstructive Pulmonary Disease (COPD)

with Russ Bowler (NJH)

- Most biomarker studies focus on single molecules, but panels have shown to improve prediction
- Examine proteins & metabolites to find phenotype specific networks as candidate biomarkers

Outcome:
% emphysema

Study Design: proteins and metabolites measured in blood on COPDGene cohort subjects (n=1008)
Platform: Metabolon, SOMAScan
Analysis: sparse canonical correlation analysis, adjusting cell counts



1-stearoyl-2-linoleoyl-GPI (18:0/18:2)
1-stearoyl-2-docosahexaenoyl-GPE (18:0/22:6)
palmitoyl-linoleoyl-glycerol (16:0/18:2) [2]
1-ribosyl-imidazoleacetate
Proto-oncogene tyrosine-protein kinase receptor Ret
valine
Growth hormone receptor
Aminoacylase-1
glutamate
QORL1
1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)
Glucagon
IGFBP-2
Chordin-like protein 1
Troponin T
Apolipoprotein E
Leptin
Adiponectin
androsterone glucuronide
Hemojuvelin
Sex hormone-binding globulin
1-palmitoyl-2-linoleoyl-GPI (16:0/18:2)
1-palmitoyl-2-docosahexaenoyl-GPE (16:0/22:6)

Mastej et al., (2020) Identifying Protein-metabolite Networks Associated with COPD Phenotypes. *Metabolites*

Study 3: Role of miRNA in Alcohol Related Behaviors

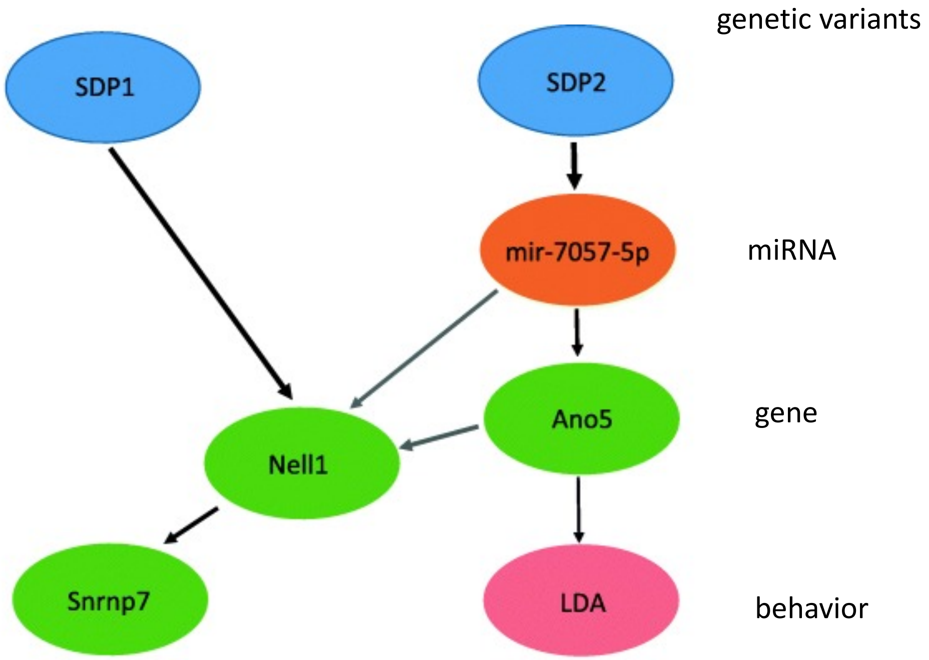
with Laura Saba, Boris Tabkaoff (SSPPS), Paula Hoffman (SOM)

- Increasing role of miRNA in alcohol related behaviors
- Role of miRNAs as mediators of the genetic effect on behaviors is not fully understand

Study Design: expression measured in brain of recombinant inbred panel in mice; genotypes, behavioral phenotypes, and gene expression in brain available in panel
Platform: small RNA sequencing
Methods: Bayesian Network Analysis

Rudra et al., (2018) Predictive modeling of miRNA-mediated predisposition to alcohol-related phenotypes in mouse. *BMC Genomics*

Low Dose Activation (measure of sensitivity to low dose of ethanol)



Part 3:
Common Themes

Common Themes Among All Omics Projects

1. Study Design and Planning
2. Data Storage
3. Processing Data
   - Normalization
   - QC plots
4. Multiple Testing Comparisons
5. Enrichment Analysis
6. Validation
7. Discussion

1. Study Designs

- Simple 2-group comparisons
    - E.g. differential expression/abundance
    - Easiest processing, many investigators can do by themselves
- 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)
- Talk to CIDA for designs outside of a simple 2-group comparison

Data Collection Questions: RNA-Seq example

- Communicate with the core/company collecting data is key to figure out best technology for your needs
- Do you want bulk or cell-specific level?
  - Single cell vs bulk
- 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
  - 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)

2. 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
- Long term storage

RNA-Seq Fastq

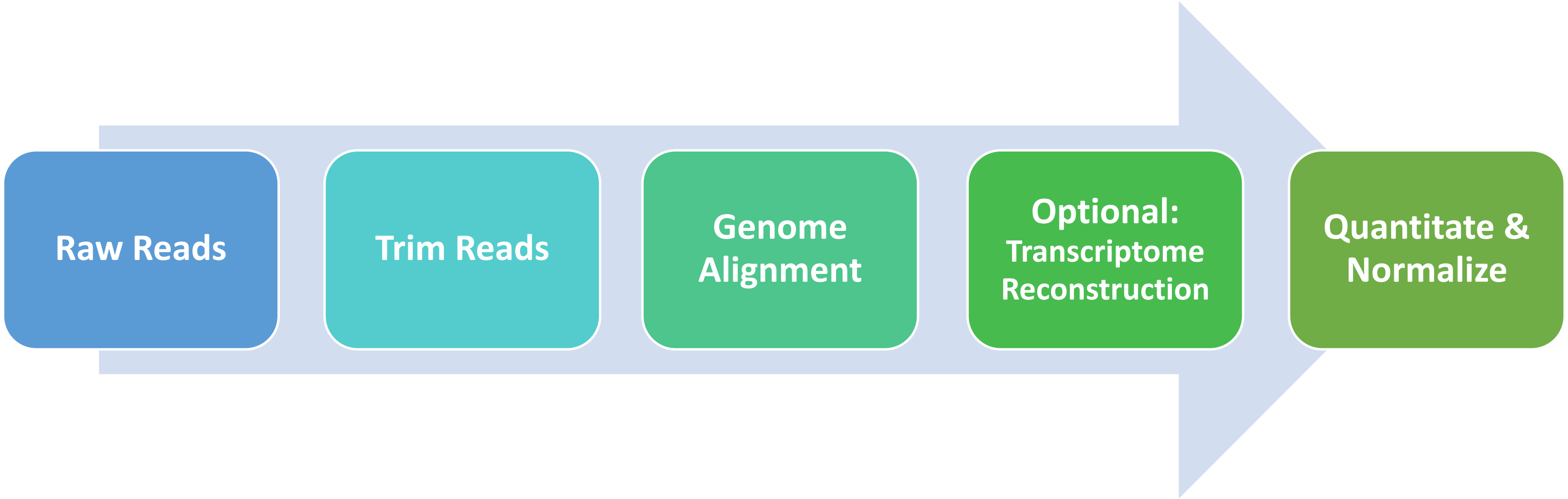
Size = # reads * (100 + 2*readLength)
Example: 100 million reads with a read length of 150 = 40G

Methylation Array Idat

450K ~ 7MB
EPIC ~ 11MB
2 files per sample

3. 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)

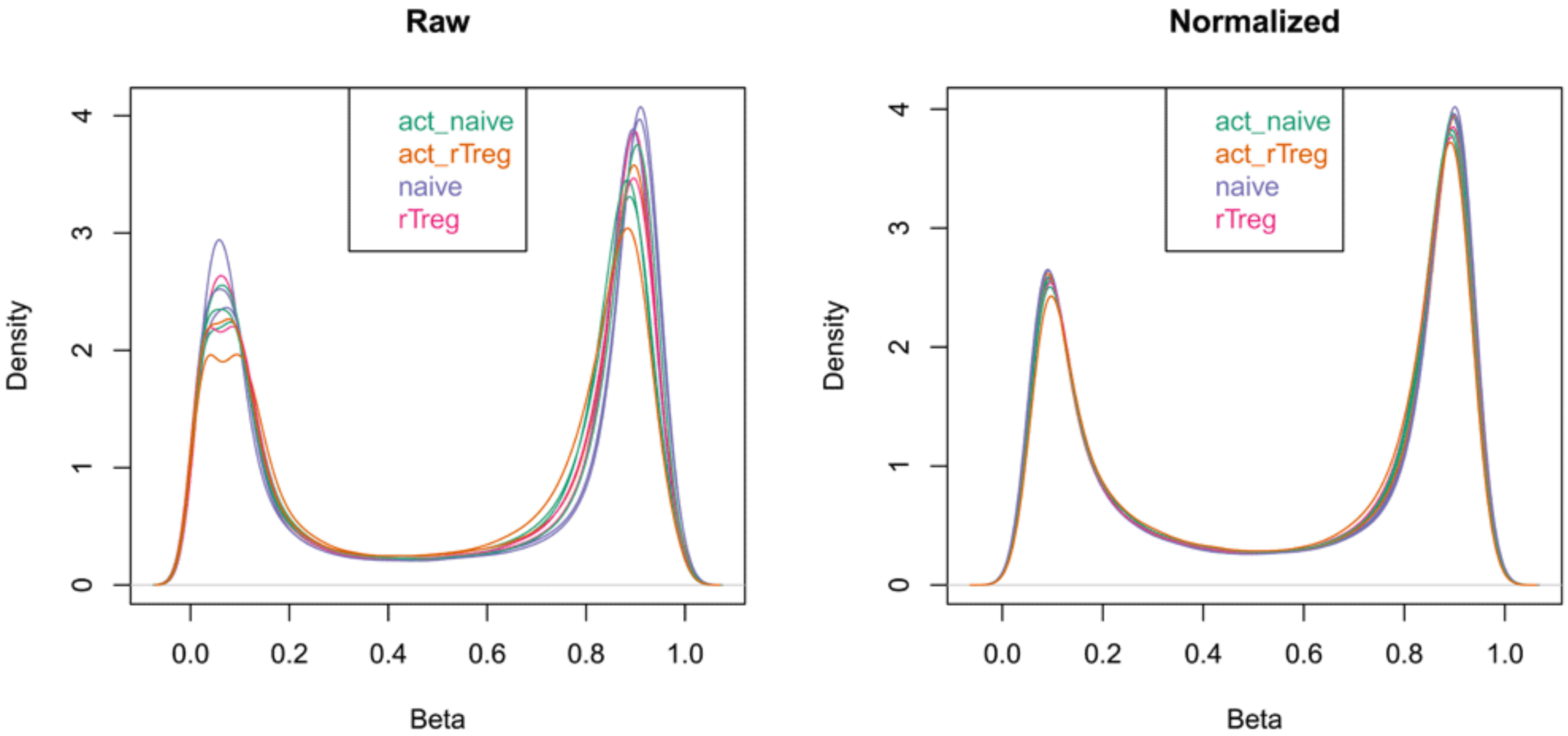
- Metabolomics (MS):
  - Locally estimated scatterplot smoothing (LOESS)
  - Systematic Error Removal using Random Forest (SERRF)
  - Median
  - Quantile
  - Cross-Contribution Compensating Multiple Standard Normalization (CRMN)
  - SVA
  - RUV
  - R/MSprep evaluates best method for metabolomics MS data

- 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 Visualization: Evaluating Normalization Density Plots – Methylation Array Example

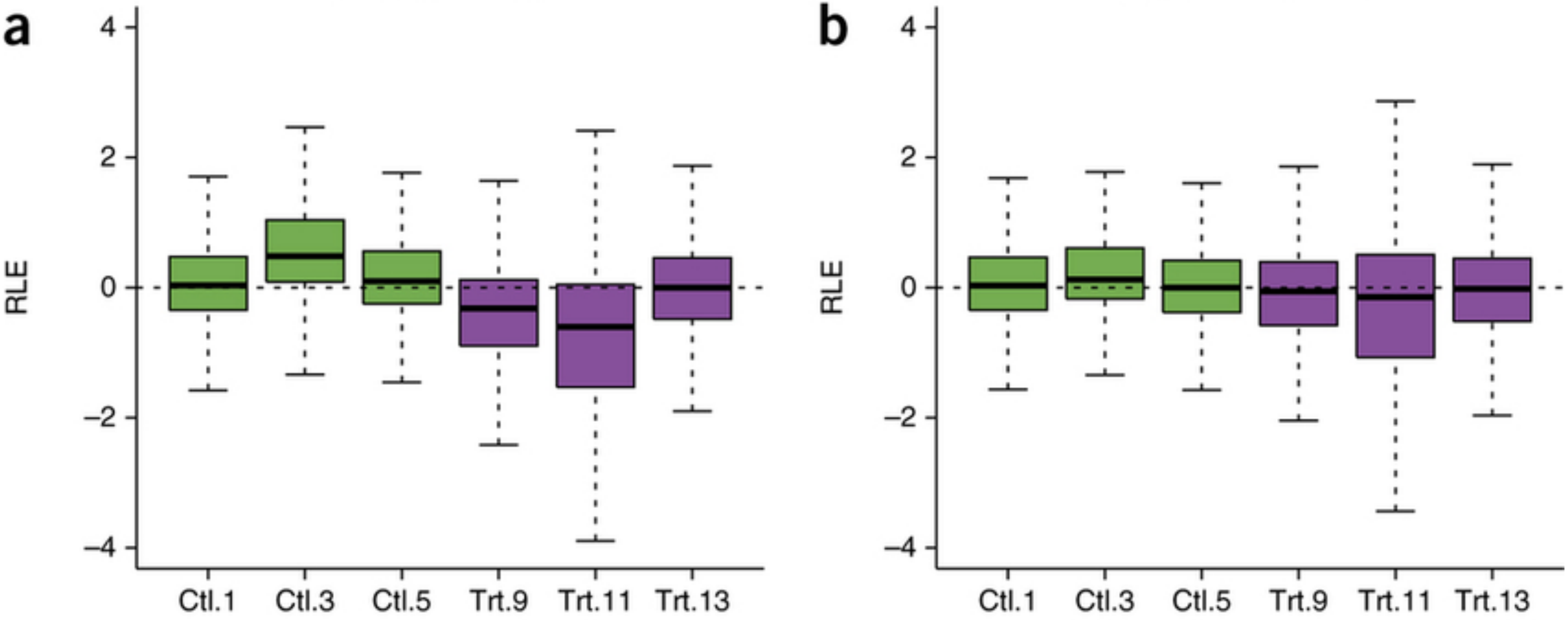


Maksimovic J, Phipson B and Oshlack A. A cross-package Bioconductor workflow for analysing methylation array data [version 3]. F1000Research 2017, 5:1281 (doi: 10.12688/f1000research.8839.3)

RLE Plots: Relative Log Expression

RAW Counts

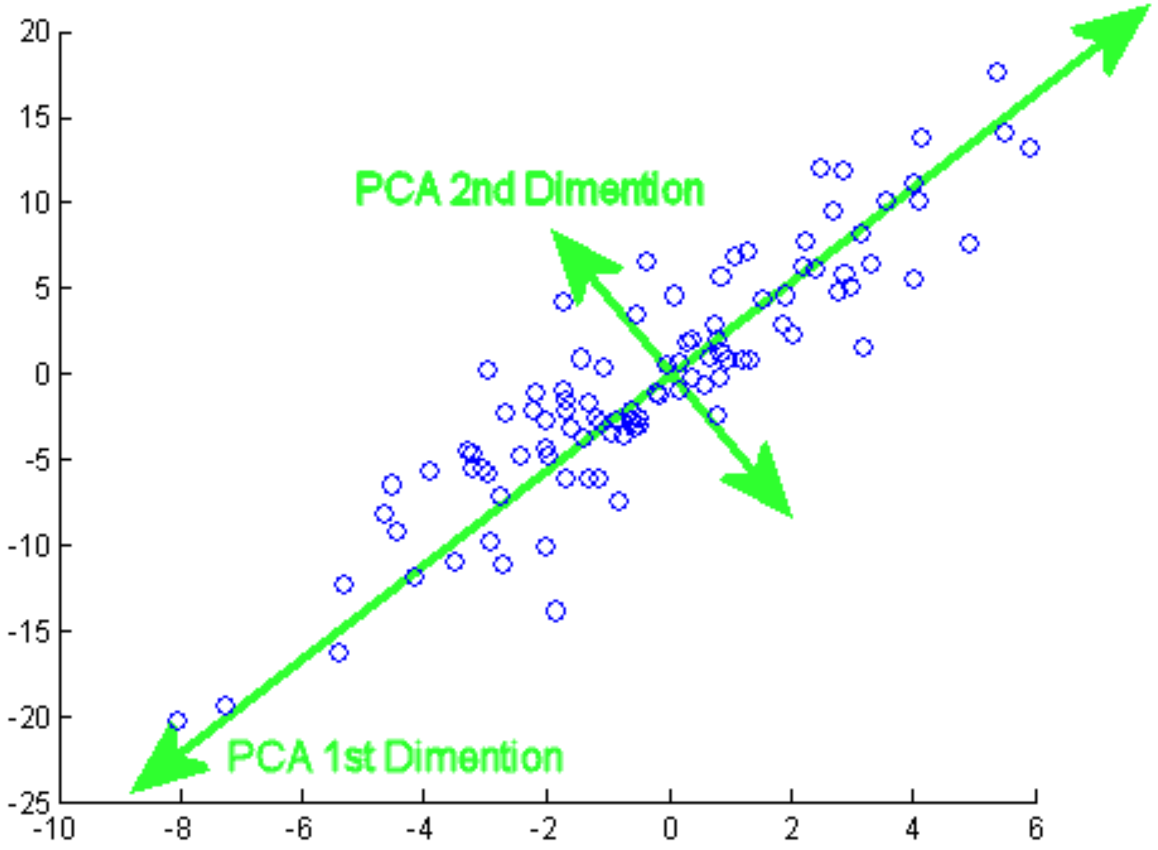
Normalized Counts



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.

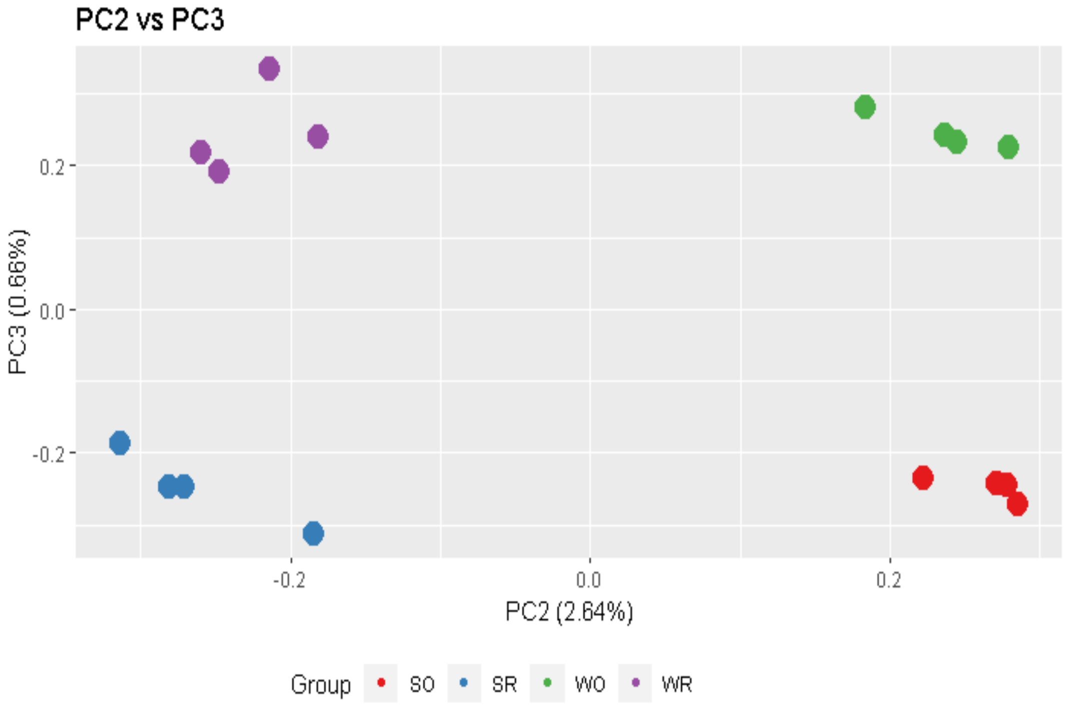
Data Reduction for QC Purposes

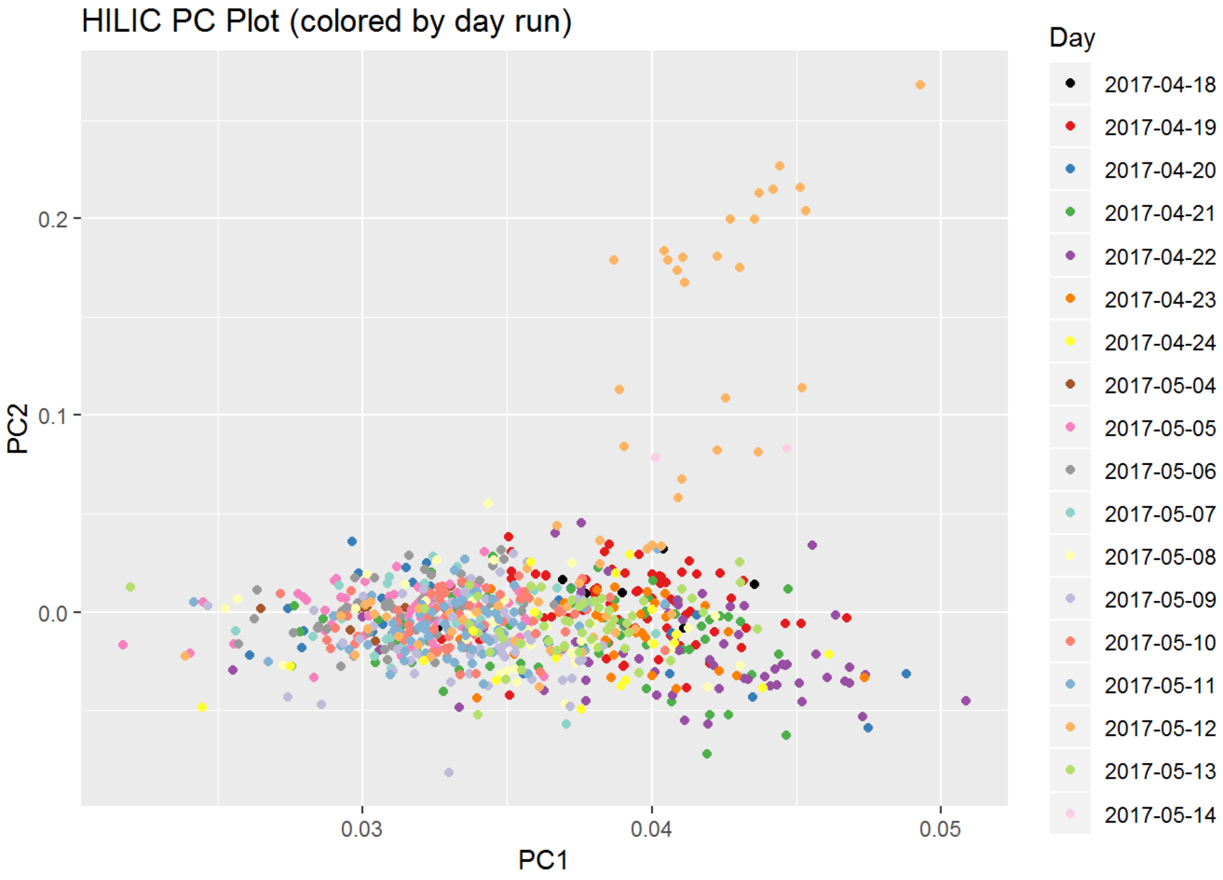
- Principal Component Analysis (PCA)
- Factor Analysis
- Singular Value Decomposition (SVD)
- Independent Component Analysis (ICA)



https://dataconomy.com/

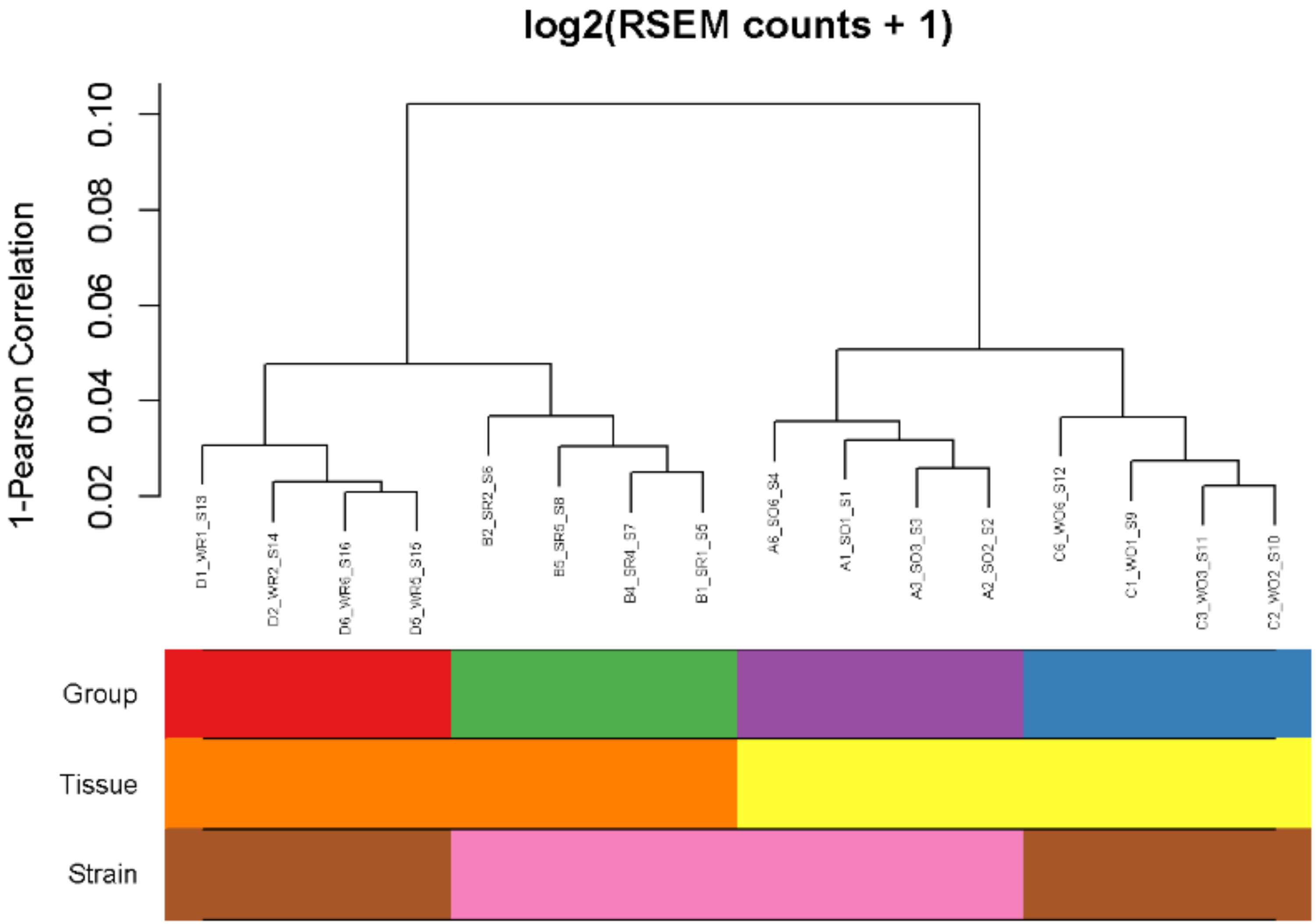
QC PCA Plots





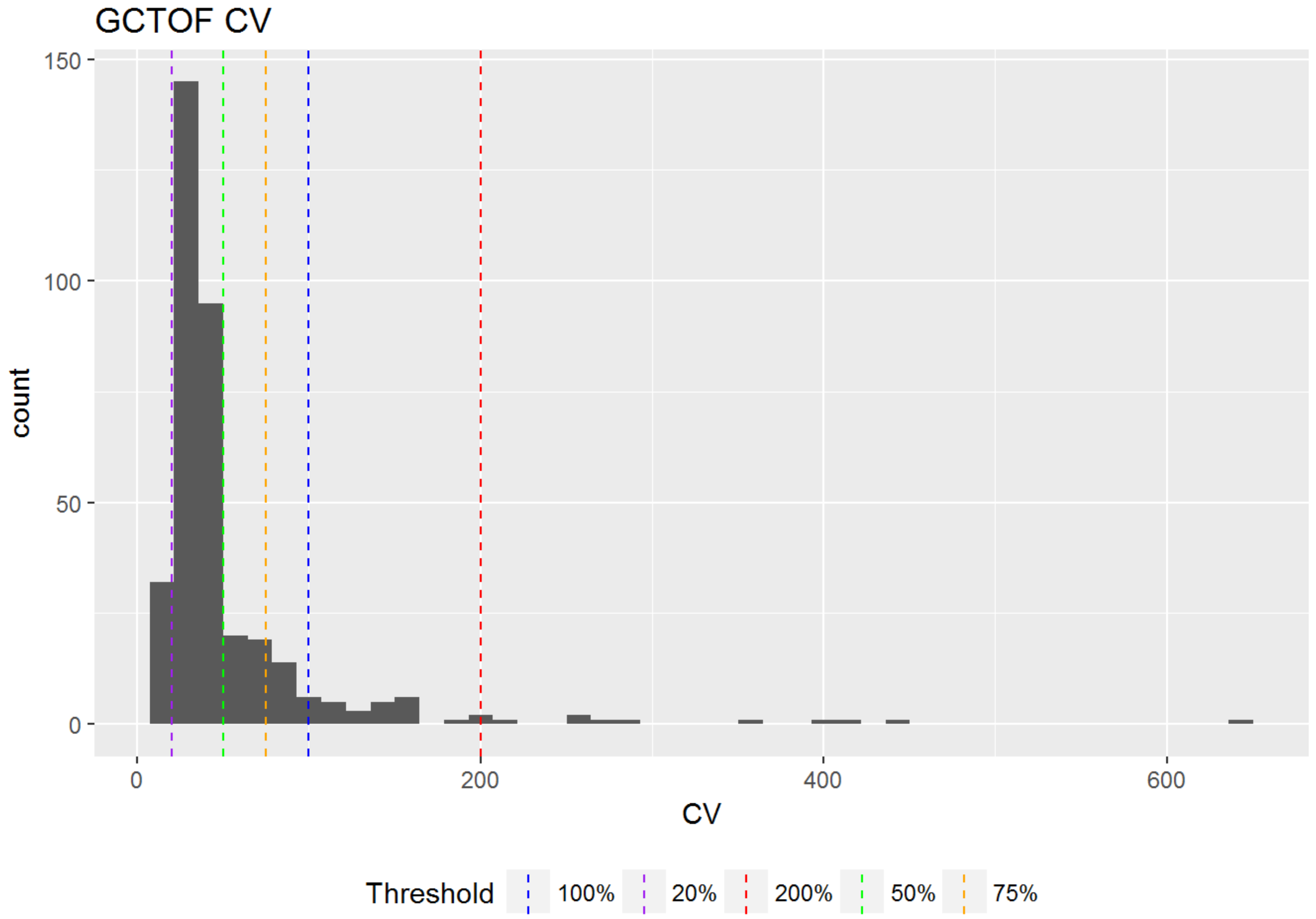
Clustering by biological or technical factors
Depending on study design

Sample Level QC: Dendrograms & PC Plots



Feature Level QC

- Detection above background threshold
- Coefficient of variation (CV) threshold
- No set feature QC for any technology

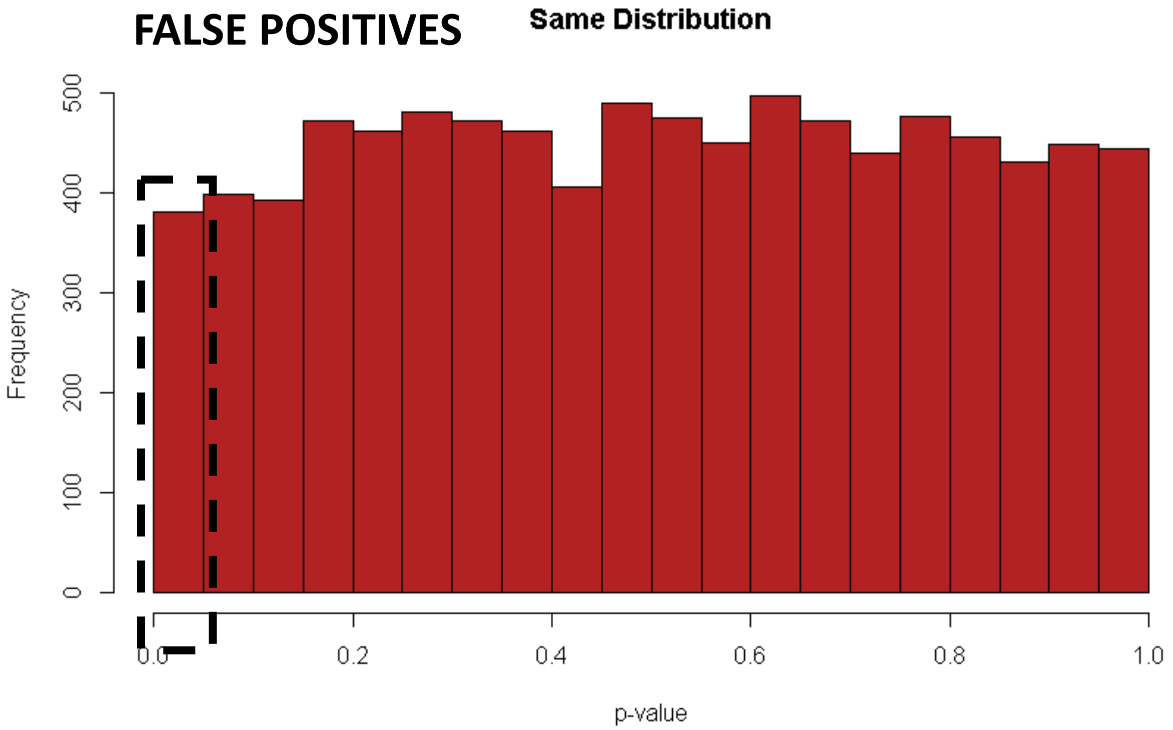


4. Multiple Testing

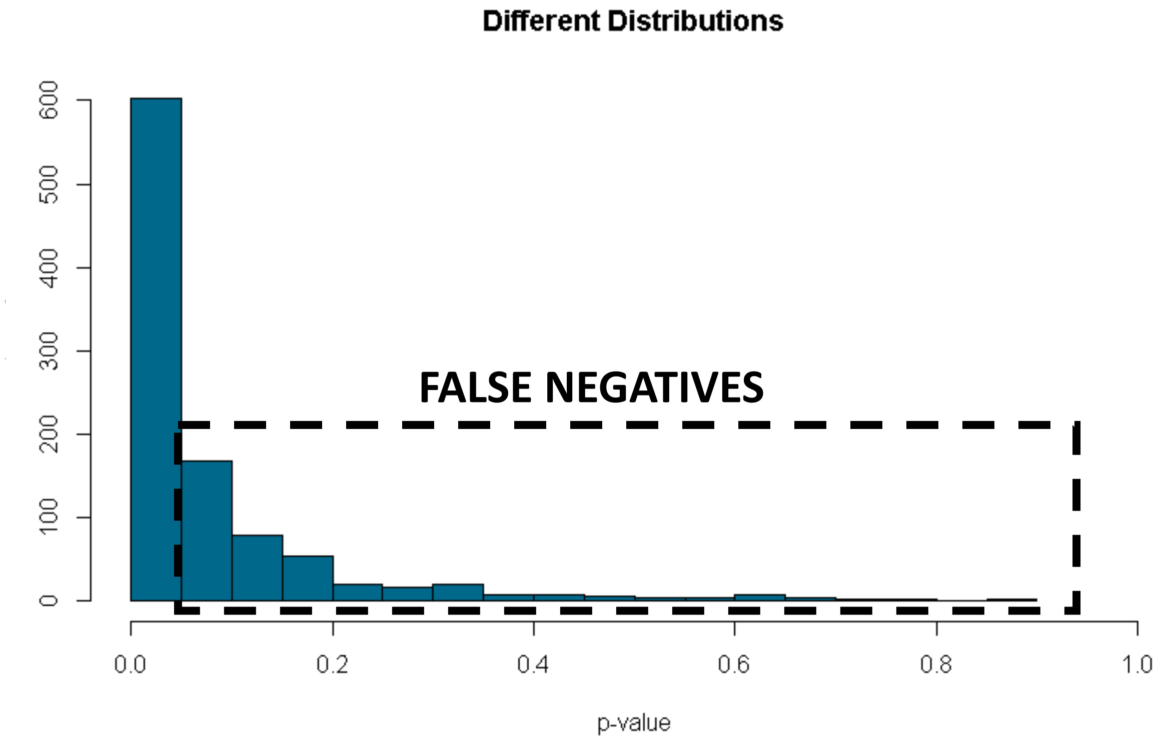
- Same statistical model on every feature
  - Example: 20,000 genes, then you have 20,000 tests
  - If you leave alpha = 0.05 you would expect 1,000 false positive results (Yikes!)
- Perform correction for multiple testing
- All methods are assuming all tests are independent
- Bonferroni
  - Multiple 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



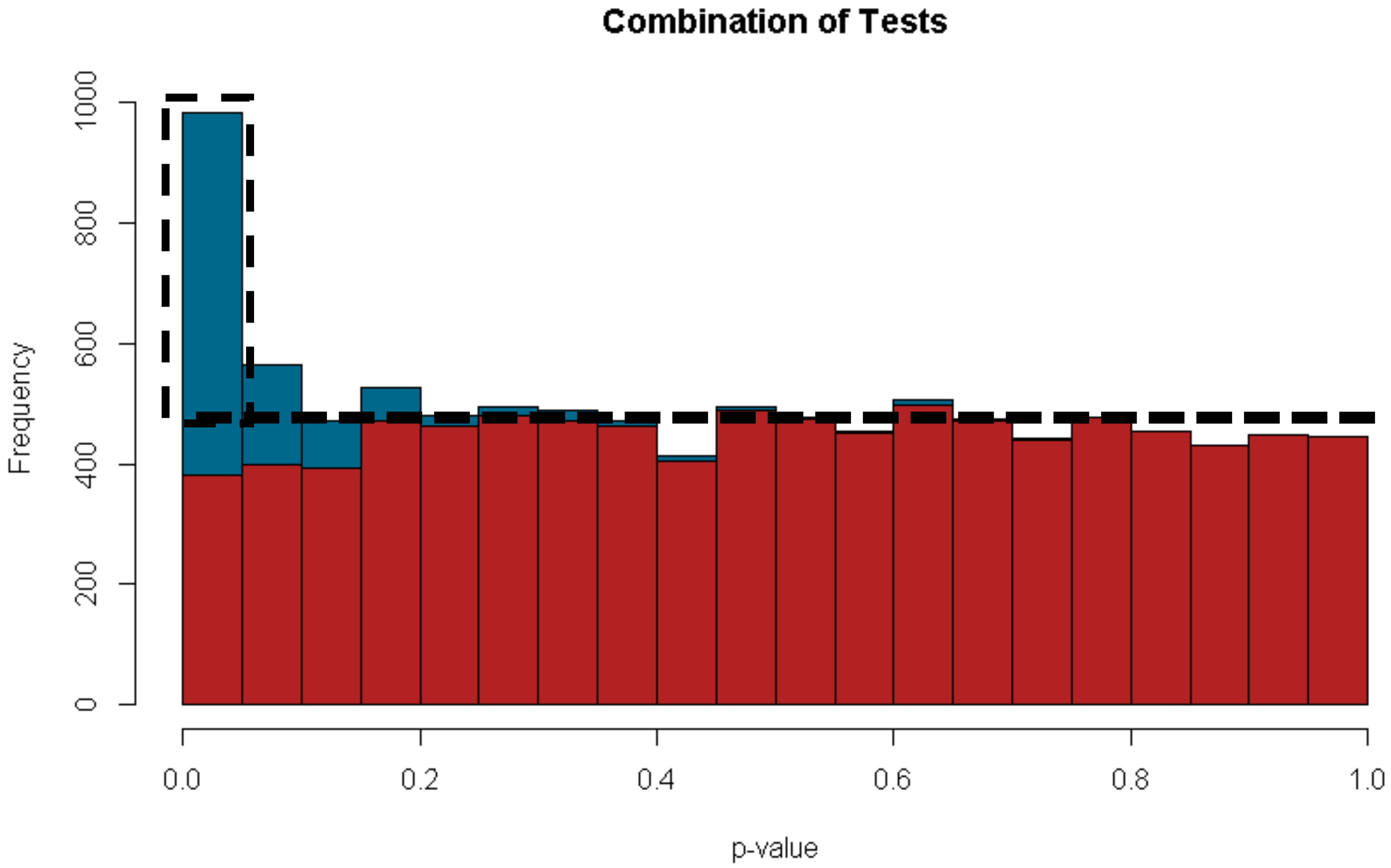
FALSE POSITIVES



FALSE NEGATIVES

False Discovery Rate (FDR)

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



5. 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		

6. 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
- Journals wanting more validation

7. 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 CIDA
   - More complex study design (e.g., multiple time points, biological/treatment groups)
   - More complex analyses (e.g., alternative splicing, transcriptome reconstruction, gene fusion)
   - CIDA provides not only analysis support, but also grant support
4. Budget time and effort for data analysis (biggest bottleneck)

Discussion