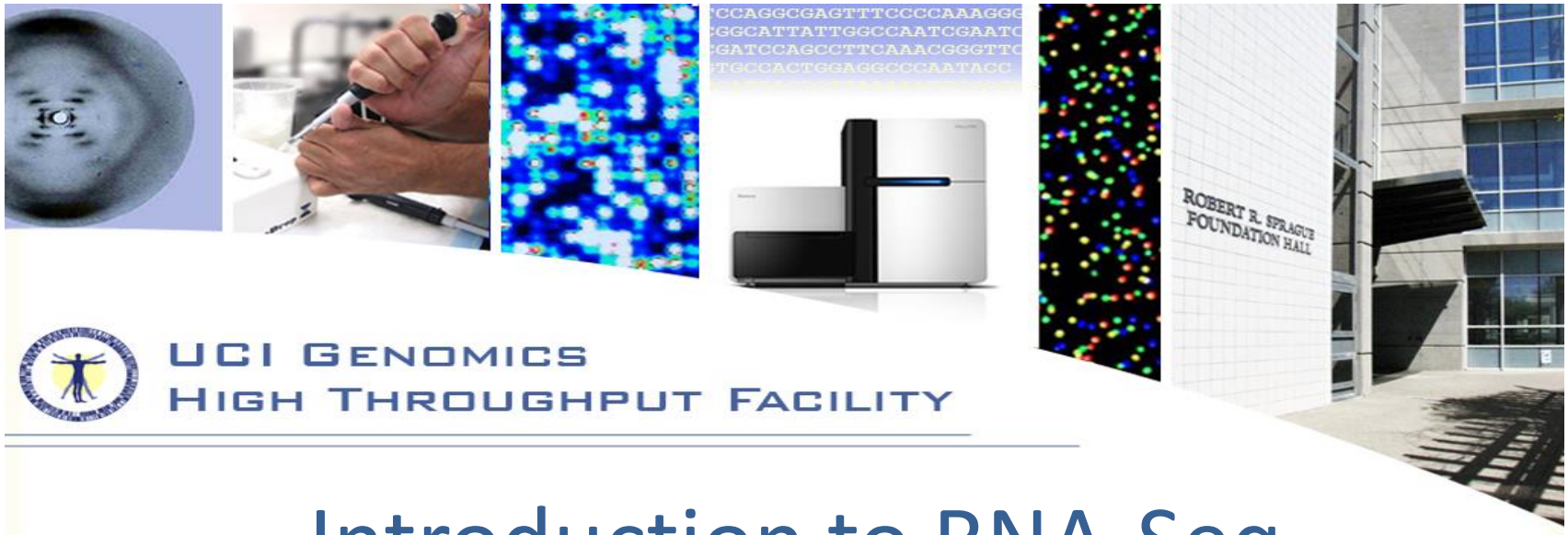


UCI Genomics
High Throughput Facility

Introduction to RNA-Seq Data Analysis

Jenny Wu

Director of Bioinformatics

Genomics Research and Technology Hub

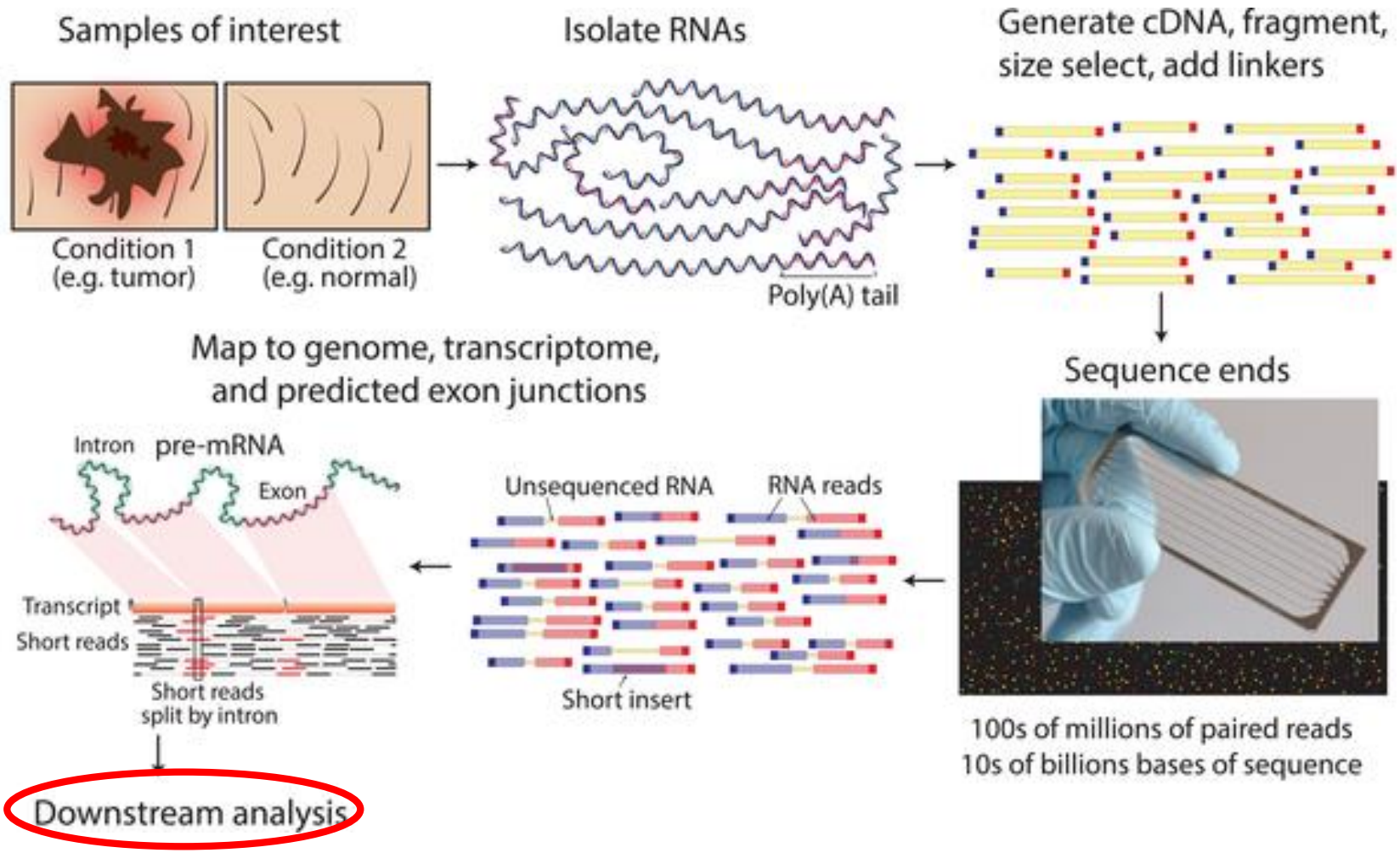
Chao Family Comprehensive Cancer Center

UC Irvine

Outline

- Introduction: RNA-Seq data generation and its applications
- RNA-Seq Data Analysis
  - ✓ Experimental Design
  - ✓ General workflow and Data Analysis Pipeline
    - Preprocessing to count matrix
    - Data normalization
    - Exploratory Data Analysis
    - Statistical analysis of differential gene expression (afternoon)
    - Downstream Pathway and network analysis
- Summary

RNA-Seq Data Generation



Griffith et al, 2015

RNA-Seq Applications

- ✓ Differential gene expression (DGE)
- ✓ Differential alternative splicing
- ✓ Transcript discovery
- ✓ Genome annotation (de novo transcriptome assembly)
- ✓ Allele specific expression
- ✓ Differential polyadenylation
- ✓ RNA editing, fusion discovery, variant detection..

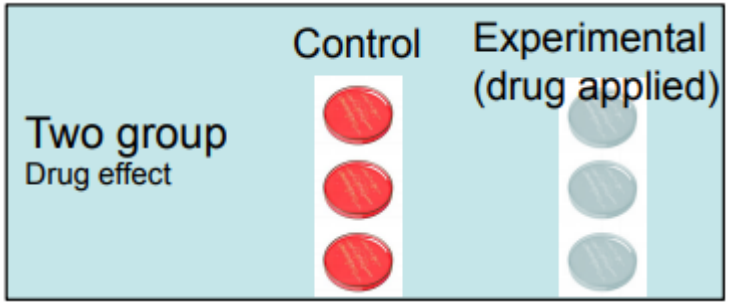
Experimental Design

- ✓ Sample size estimation and power analysis
- ✓ Library type : polyA enriched or ribo reduction, stranded or not, ERCC
- ✓ Sequencing: read length, paired end or single read, sequencing coverage

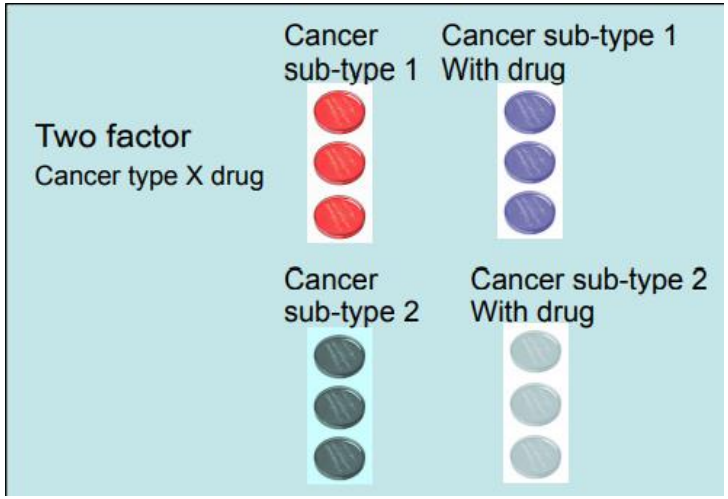
Our most popular setup: ribo reduction, stranded, PE100, 25-50M/sample, 3+ biological replicates per condition.

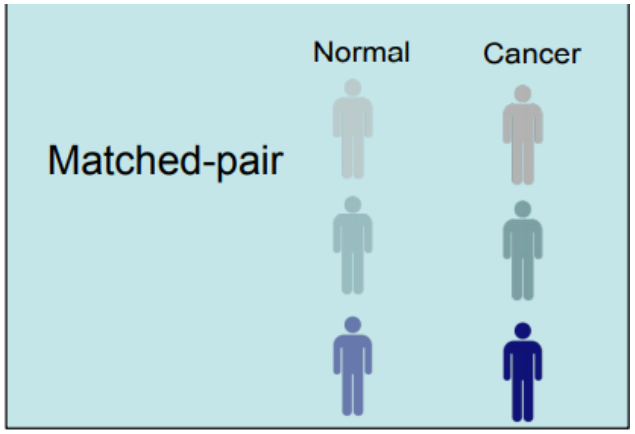
Experimental Design

✓Simple Design



✓Complex Design

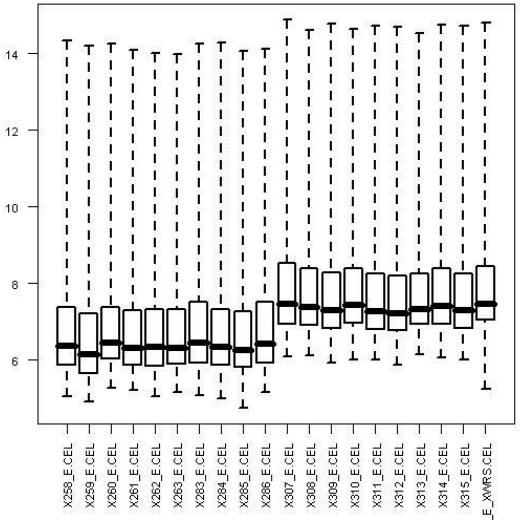


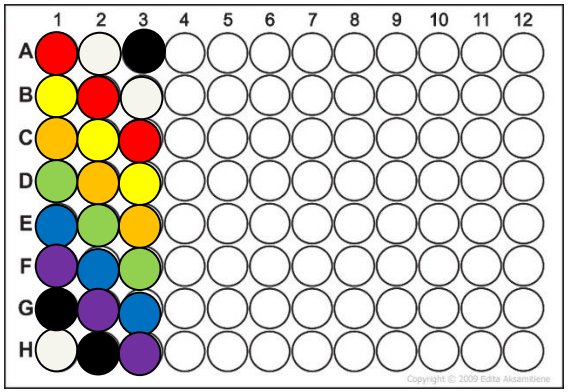


Statistical analysis of data with complex design can be done using R based tools such as DESeq2 and limma

Batch Effect

- Randomization and Replication
  - **Don't** do all of one factor level together
  - Arrange the samples randomly



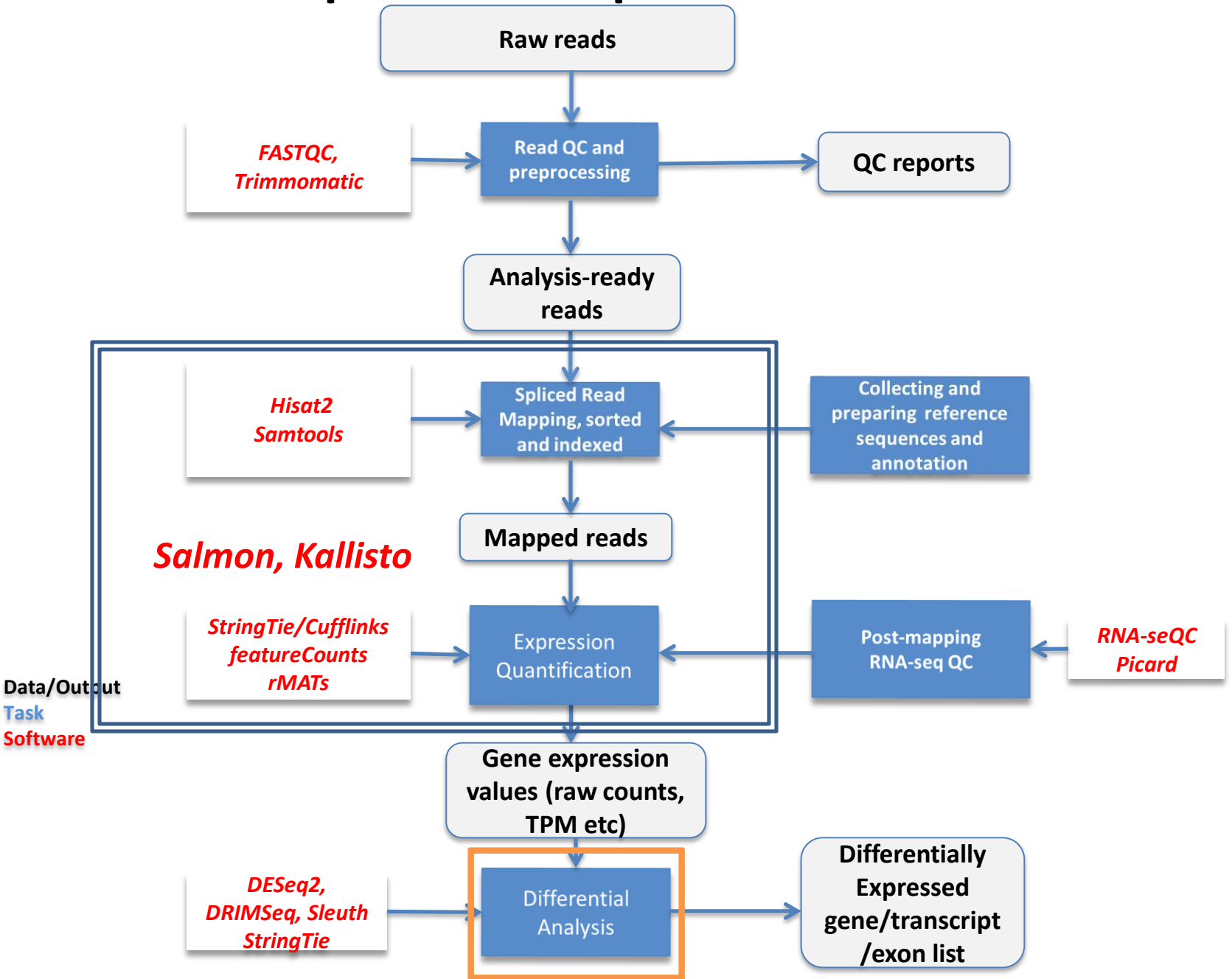


Have one replicate in each row and each column!

  - If batch effects are spread evenly over factor levels, they can be accounted for statistically (blocking)

http://www.cellsignet.com/media/templ.html

RNA-Seq Data Pipeline for DGE



Transcript Counting vs. Classic Approach

- Novel transcripts, mutation identification
- Performance: execution time, RAM usage, multi-reads
- Downstream analysis: Gene length and GC content correction

Expression Quantification: Count Table

	Sample 1	Sample 2	…	Sample N
Gene 1	K_{11}	K_{12}	…	K_{1N}
Gene 2	K_{21}	K_{22}	…	K_{2N}
…			…	
Gene p	K_{p1}	K_{p2}	…	K_{pN}

- ✓ K_{ij} is discrete positive, skewed, large dynamic range.
- ✓ p$\gg N$ small number of replicates for bulk RNA-seq.
- ✓ In scRNA-seq, each sample is a cell. So N can be bigger than p.

RNA-Seq: Normalization

Gene-length bias

- ✓ Differential expression of longer genes is more significant because long genes yield more reads

RNA-Seq normalization methods:

- ✓ Scaling factor based: Total count, upper quartile, median, **DESeq2, TMM in edgeR**
- ✓ **Quantile**, FPKM (cufflinks)
- ✓ **ERCC**

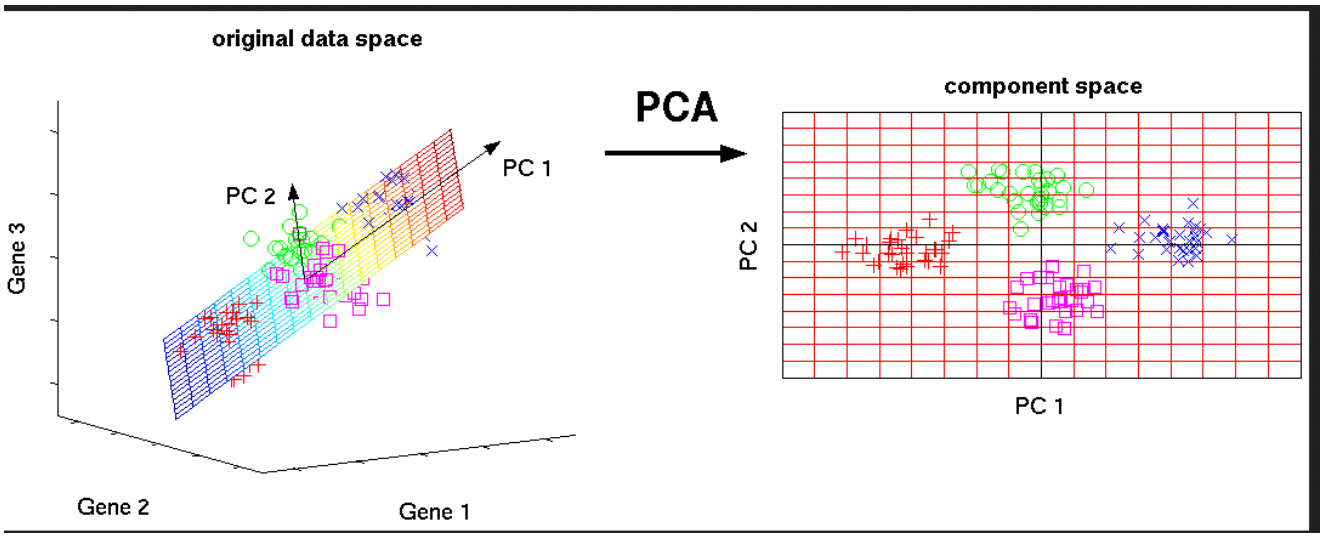
Normalize by gene length and by number of reads mapped, e.g. **RPKM/FPKM** (reads/fragments per kilo bases per million mapped reads) or **TPM** (Transcripts per million).

Commonly Used Computing Techniques in RNA-Seq

- ✓ Dimension Reduction: *PCA, t-SNE, UMAP etc*
- ✓ Feature Selection
- ✓ Clustering: *K means, hierarchical etc*
- ✓ Differential Statistics: linear and generalized linear models.

Exploratory Data Analysis:
Linear Dimension Reduction

- ✓ Principle Component Analysis (PCA) is a standard technique for visualizing high dimensional data and for data compression.



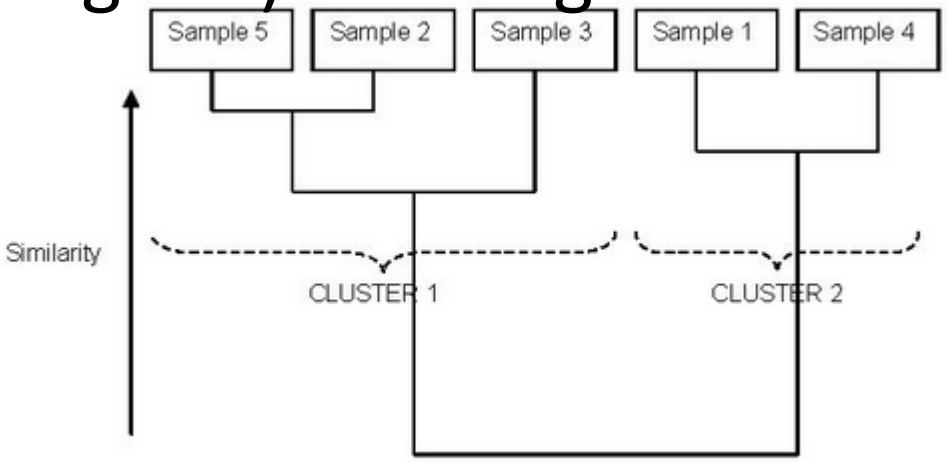
Matthias Scholz, 2015

Dimension Reduction with PCA

- ✓ Reduce p-dimensional dataset into much smaller number (2,3)
- ✓ Find a new(smaller) set of variables that retains most of the information in all samples
- ✓ Effective way to visualize multivariate data
- ✓ PCA should be applied on data that have approximately the same scale in each variable

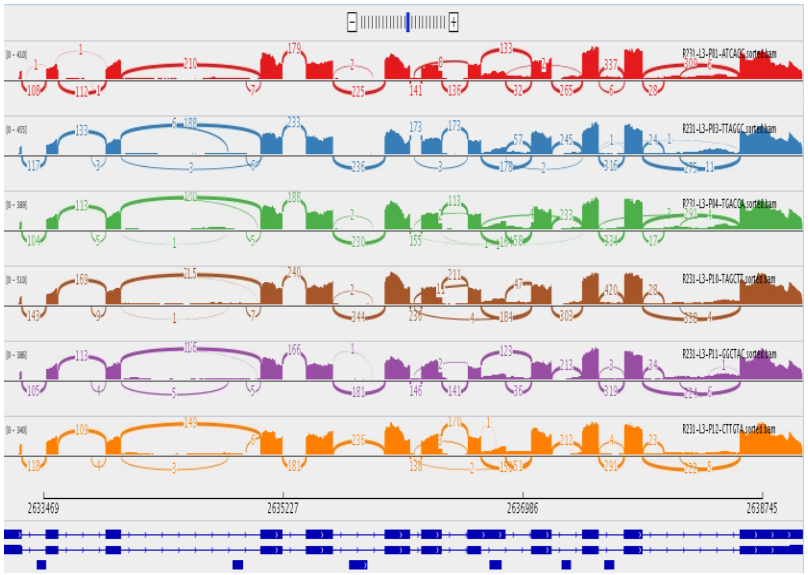
Hierarchical Clustering

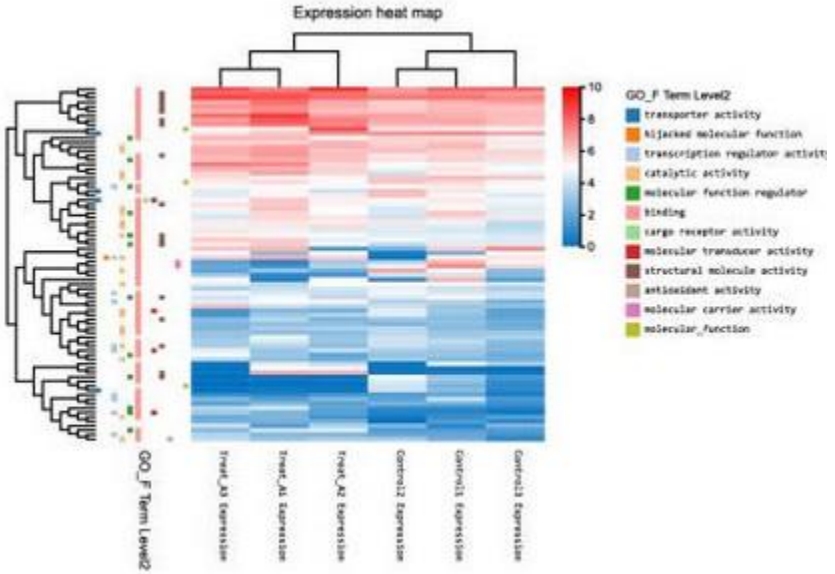
- Agglomerative method: bottom up. Starts with N groups.
- Steps:
  - Define distance (e.g. Euclidean, correlation)
  - Compute distance and merge progressively
- Results: a tree(dendrogram) showing how close things are to each other.

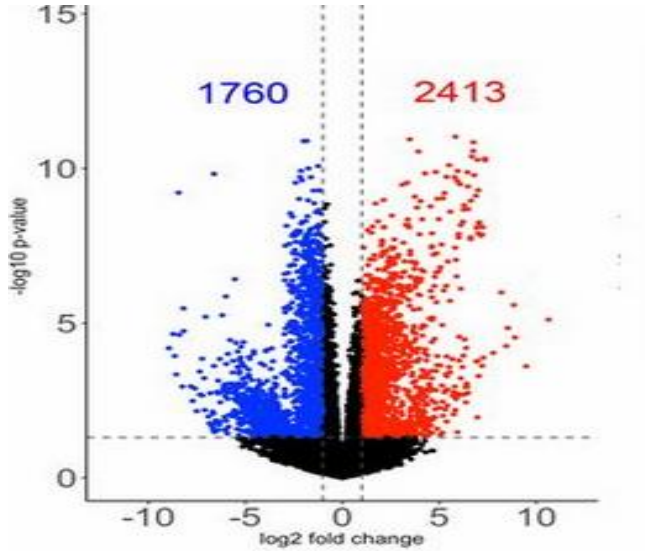


RNA-Seq Data Visualization

- IGV, Sashimi plots
- Volcano plots, heat maps, PCA







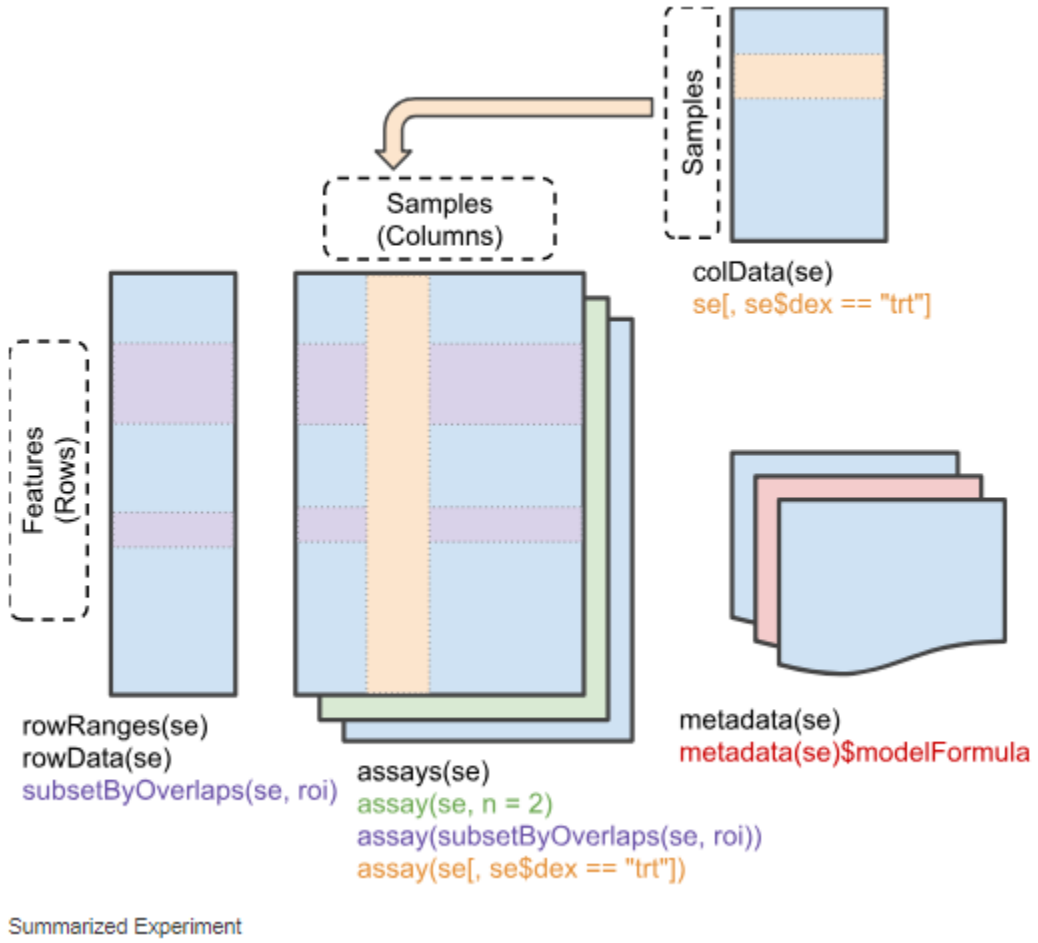
Visualizing RNA-Seq mapping with IGV

http://www.broadinstitute.org/igv/UserGuide

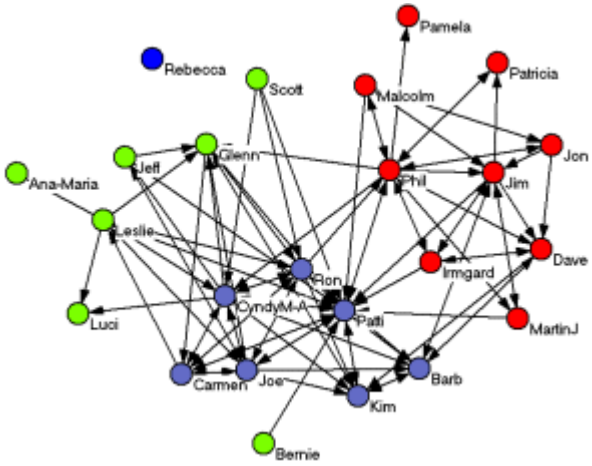
Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Thorvaldsdóttir H et al. Brief Bioinform. 2013

Gene Expression Data Structure in R/Bioconductor:

SummarizedExperiment/ExpressionSet



Why Pathway Analysis



- Logical next step in any high throughput experiment
- Goal: to characterize biological meaning of joint changes in gene expression
- Why? Often sets of genes doing related functions are changed

Pathway and Network Analysis

***Pathway Analysis Methods*:**

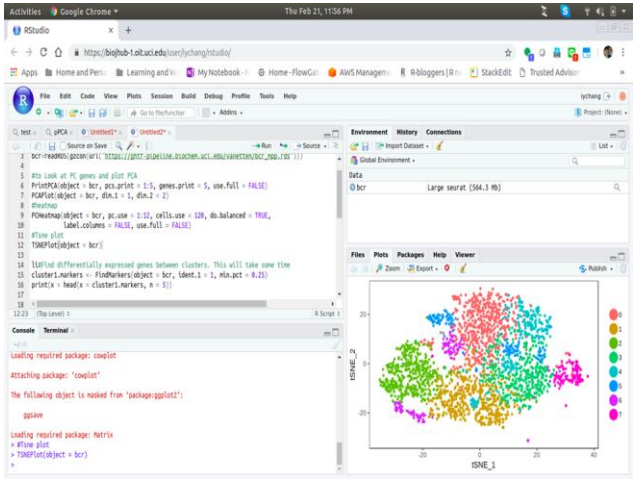
- **Functional category over representation:** discrete test for significance (*GOrilla, EnrichR, IPA, David etc*)
- Continuous test (*GSEA, PAGE*)
- Signaling Pathway Impact Analysis (*iPathway Guide*)

Network Analysis: (*WGCNA, Cytoscape* etc)

Provenance and Reproducibility

- Software containerization
- Jupyter notebook environment

User Authentication

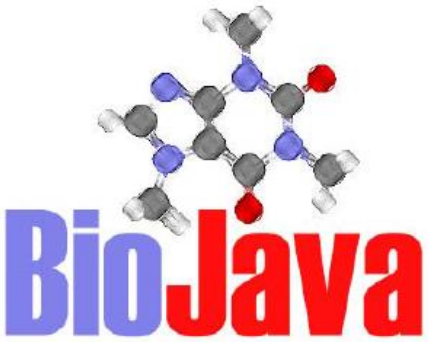


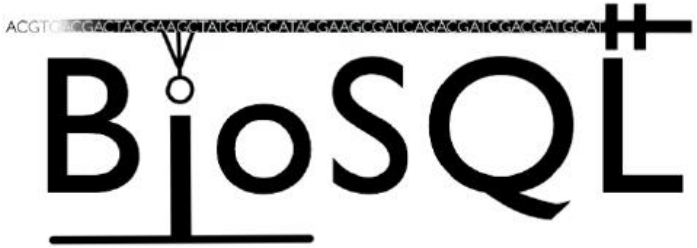
Languages in Bioinformatics

BioRuby

Open source bioinformatics library for Ruby

BioPerl





Resources in Data Analysis

Questions and Discussion

- Questions?

Thanks for your time and attention!