Introduction to Large-Scale Biomedical Data Analysis

A grand overview of the course

- Introductory course created for the BIG (Bioinformatics, Imaging and Genetics) concentration.
- Purpose of the course: introduce modern high-dimensional biomedical data analysis from:
  - Bioinformatics and computational biology.
  - Biomedical imaging.
  - Statistical genetic.
  - Microbiome.

Contents of the course

- Focus on:
  - Scientific background: questions and motivations.
  - Technologies.
  - Data and their characteristics.
  - Brief overview of some statistical methods, opportunities and challenges for statisticians.

  There will be a lot of materials and new terminologies!
- **Not** covered in this course: detailed statistical theories and methods for data analyses.
- This is a knowledge-centric class, big picture and concepts are more important.

Format of the course

- Co-taught by multiple instructors.
- Students are evaluated by 2 reading assignments: you need to write reading reports!

Introduction to high-throughput omics data analysis

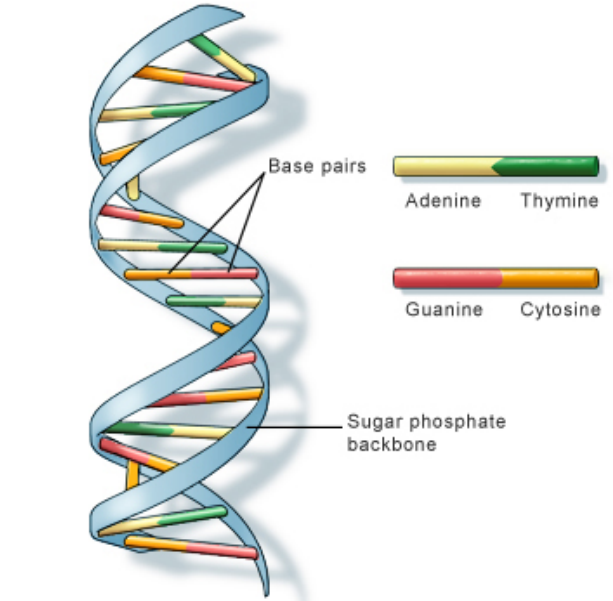
Outline

- Biological Backgrounds: DNA and DNA sequencing.
- High-throughput technologies and application.
- Feature selection from high-throughput data.

Background: DNA and sequencing

DNA (DeoxyriboNucleic Acid)

- A molecule contains the genetic instruction of all known living organisms and some viruses.
- Resides in the cell nucleus, where DNA is organized into long structures called **chromosomes**.
- Most DNA molecule consists of two long polymers (**strands**), where two strands entwine in the shape of a double helix.
- Each strand is a chain of simple units (**bases**) called **nucleotides**: A, C, G, T.
- The bases from two strands are complementary by **base pairing**: **A-T, C-G.**



DNA sequence

- The order of occurrence of the bases in a DNA molecule is called the **sequence** of the DNA. The DNA sequence is usually store in a big text file:

```
ACAGGTTTGCTGGTGACCAGTTCTTCATGAGGGACCATCTATCACAACAG
AGAAAGCACTTGGATCCACCAGGGGCTGCCAGGGGAAGCAGCATGGGAGC
CTGAACCATGAAGCAGGAAGCACCTGTCTGTAGGGGGAAGTGATGGAAGG
ACATGGGCACAGAAGGGTGTAGGTTTTGTGTCTGGAGGACACTGGGAGTG
GCTCCTGGCATTGAAACAGGTGTGTAGAAGGATGTGGTGGGACCTACAGA
CAGACTGGAATCTAAGGGACACTTGAATCCCAGTGTGACCATGGTCTTTA
AGGACAGGTTGGggccaggcacagtggctcatgcctgtaatcccagcact
```

- Some interesting facts:
  - Total length of the human DNA is **3 billion bases**.
  - Difference in DNA sequence between two individuals is less than 1%.
  - Human and chimpanzee have 96% of the sequences identical. Human and mouse: 70%.

Genome size (total length of DNA)

Organism	Genome size (bp)	# genes
E. coli	4.6M	4,300
S. cerevisiae (yeast)	12.5M	5,800
C. elegans (worm)	100M	20,000
A. thaliana (plant)	115M	28,000
D. melanogaster (fly)	123M	13,000
M. musculus (mouse)	3G	23,800
H. sapiens (human)	3.3G	25,000

DNA sequencing

- Technologies to determine the nucleotide bases of a DNA molecule.
- Motivation: decipher the genetic codes hidden in DNA sequences for different biological processes.
- **Genome projects**: determine DNA sequences for different species, e.g., human genome project.
- **Genomic research** (in a nutshell): study the functions of DNA sequences and related components.

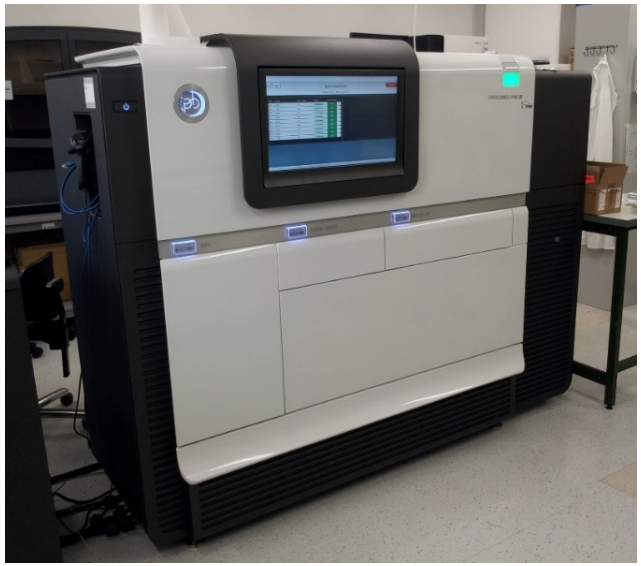
Sequencing technologies

- Traditional technology: **Sanger sequencing**.
  - Slow (low throughput) and expensive: it took Human Genome Project (HGP) 13 years and $3 billion to sequence the entire human genome.
  - Relatively accurate.
- New technology: different types of **high-throughput sequencing**.

Next generation sequencing (NGS)

- Aka: high-throughput sequencing, second-generation sequencing.
- Able to sequence large amount of short sequence segments in a short period:
  - Quick: hundreds of millions sequences in a run.
  - Cheap: sequence entire human genome costs one thousand dollars now.

Next generation sequencing technologies



Applications of NGS

- NGS has a wide range of applications.
  - DNA-seq: sequence genomic DNA.
  - RNA-seq: sequence RNA products.
  - ChIP-seq: detect protein-DNA interaction sites.
  - Bisulfite sequencing (BS-seq): measure DNA methylation strengths.
  - A lot of others.

Technology	Brief description
ChIP-seq	Locate protein-DNA interaction or histone modification sites.
CLIP-seq	Map protein-RNA binding sites
RNA-seq	Quantify expression
SAGE-seq	Quantify expression
RIP-seq	capture TF-bound transcripts
GRO-seq	evaluate promoter-proximal pausing
BS-seq	Profile DNA methylation patterns
MeDIP-seq	Profile DNA methylation patterns
TAB-seq	Profile DNA hydroxyl-methylation patterns
MIRA-seq	Profile DNA methylation patterns
ChiRP-seq	Map lncRNA occupancy
DNase-seq	Identify regulatory regions
FAIRE-seq	Identify regulatory regions
FRT-seq	Quantify expression
Repli-seq	Assess DNA replication timing
MNase-seq	Identify nucleosome position
Hi-C	Infer 3D genome organization
ChIA-PET	Detect long distance chromosome interactions
4C-seq	Detect long distance chromosome interaction
Sono-seq	Map open-chromatin sites
NET-seq	determine *in vivo* position of all active RNAP complexes.
NA-seq	Map Nuclease-Accessible Sites

DNA-seq

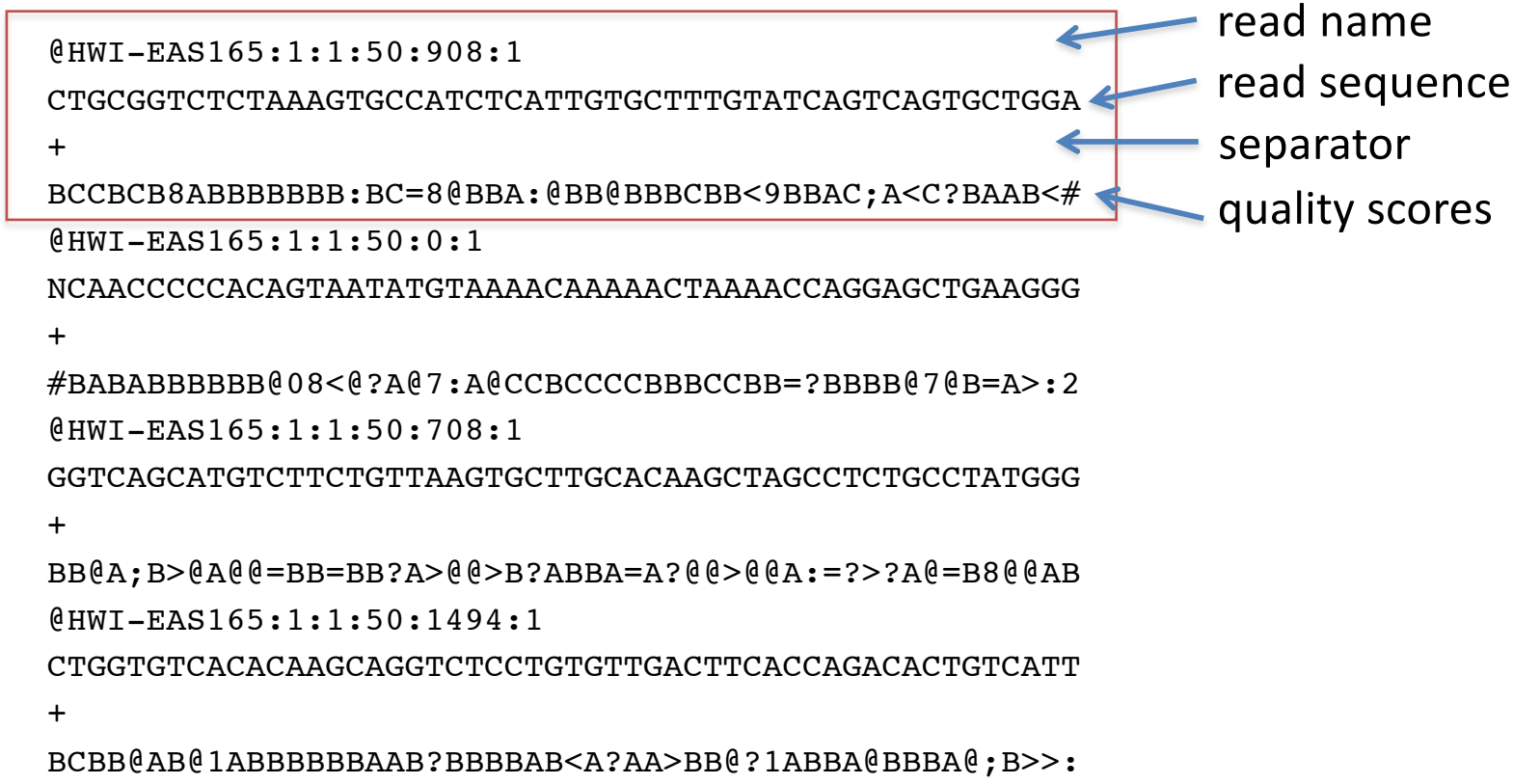
- Sequence the untreated genomic DNA.
  - Obtain DNA from cells, cut into small pieces then sequence the segments.
- Goals:
  - **Genome re-sequencing**: compare to the **reference genome** and look for genetic variants:
    - Single nucleotide polymorphisms (SNPs)
    - Insertions/deletions (indels),
    - Copy number variations (CNVs)
    - Other structural variations (gene fusion, etc.).
  - ***De novo assembly*** of a new unknown genome.

RNA-seq

- Sequence the “transcriptome”: the set of RNA molecules.
- Goals:
  - Catalog RNA products.
  - Determine transcriptional structures: alternative splicing, gene fusion, etc.
  - Quantify gene expression: the sequencing version of gene expression microarray.

Raw data from NGS

- Large text file (millions of lines) with simple format.



Sequence Alignment

- Sequence Alignment
  - Use the known genome (called “reference genome”) as a blue print.
  - Determine where each read is located in the reference genome.
- Need: sequence reads file and a reference genome.
- It is basically a string search problem: where is the short (50-letter) string located within the reference string of 3 billion letters.
- Brute-force searching is okay for a single read, but computationally infeasible to alignment millions of reads.
- Clever algorithms are needed to preprocess the reference genome (indexing), which is beyond the scope of this class.

Popular alignment software

- Bowtie: fast, but less accurate.
- BWA (Burrows-Wheeler Aligner): same algorithm as bowtie, but allow gaps in alignments.
  - about 5-10 times slower than bowtie, but provide better results especially for paired end data.
- Maq (Mapping and Assembly with Qualities): with SNP calling capabilities.
- ELAND: Illumina's commercial software.
- A lot of others. See http://en.wikipedia.org/wiki/List_of_sequence_alignment_software for more details.

Once you have the reads aligned

- Downstream analyses depend on purpose.
- Often one wants to manipulating and visualizing the alignment results. There are several useful tools:
  - file manipulating (format conversion, counting, etc.): samtools/Rsamtools, BEDTools, bamtools, IGV tools.
  - Visualizing: IGV (Java GUI).

Feature selection from high-throughput data

High-throughput data characteristics

- Large size
- Simple structure
- Noisy, low signal to noise ratio
- Prone to technical artifacts
- A lot of high-throughput data analyses are some type of “feature selection”.

Why feature selection?

- There are large number of features in high-throughput data.
- Most of them are not related to the outcome of interest.
- Using a small number of “informative” features provides more precise targets and parsimonious model for prediction.

Features from high-throughput data

- Basic feature:
  - DNA sequence: genetic variant
  - Gene expression data: genes
  - DNA methylation data: CpG sites
- A group of basic features: a set of variants, genes, or CpG sites.
- Higher order features:
  - Transformation of a group of features, for example, the principal components.

Purpose of feature selection

- Select “biomarkers” to:
  - Understand biology, and/or identify therapeutic targets:
    - Basic features are more informative
    - A group of features are okay
    - Higher order features are not useful (black box)
  - Construct prediction models:
    - All types of features are useful, higher order features might be the most effective.

Types of feature selection

- Supervised:
  - With known outcomes: disease status, phenotypical values, etc.
  - Look for features correlated with the outcome.
  - Examples of supervised feature selection: differential expression/methylation, GWAS/EWAS.
- Unsupervised:
  - No outcome. Try to find a small number of features representing data well.
  - Useful for clustering.

Supervised feature selection – differential expression (DE)

- Applies to GE microarray and RNA-seq.
- Goal: find genes that are expressed differently between (among) conditions.
- Procedure in a nutshell:
  - Properly normalize data.
  - Perform statistical test for each gene.
  - Correct for multiple testing, and use a threshold to call DE.

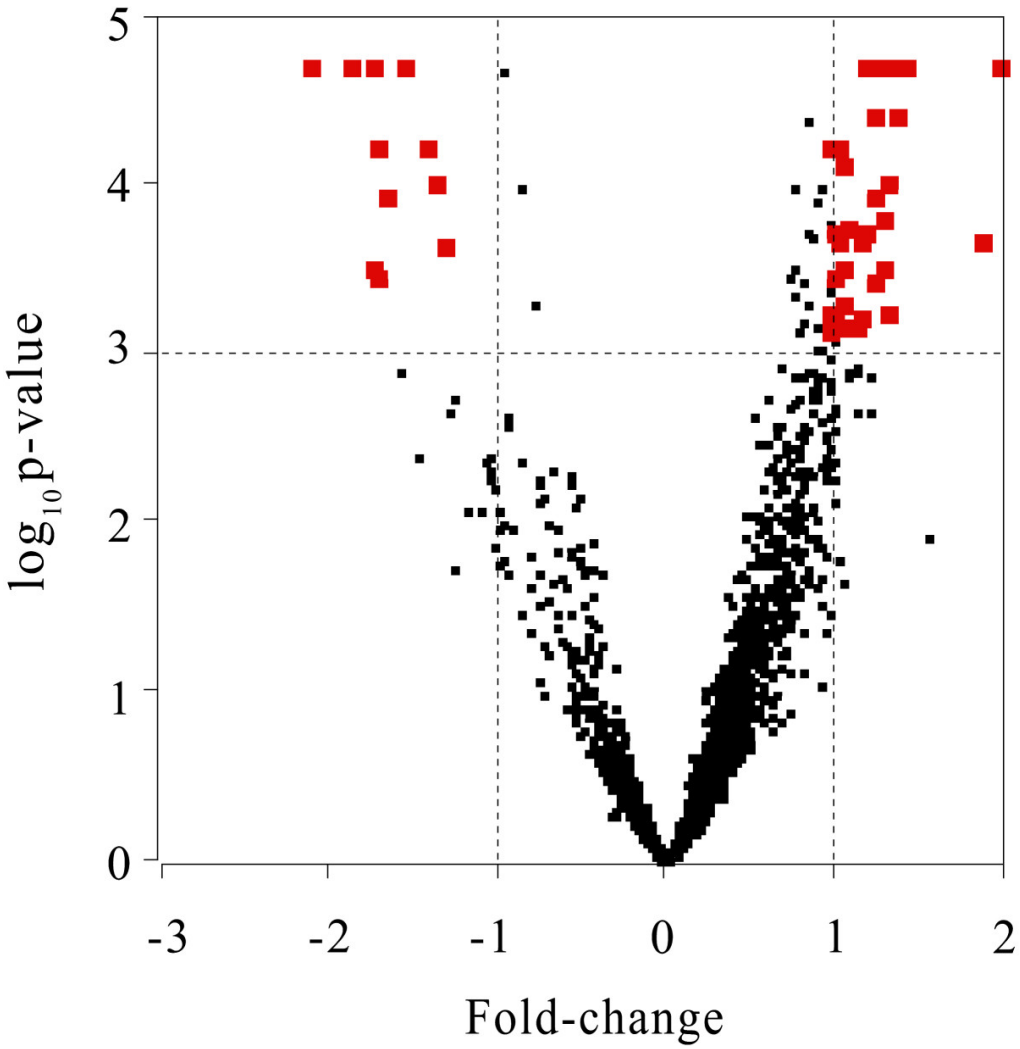
Gene expression data

	Normal	Normal	Cancer	Cancer
1007_s_at	8.575758	8.915618	9.150667	8.967870
1053_at	6.959002	7.039825	6.898245	7.136316
117_at	7.738714	7.618013	7.499127	7.610726
121_at	10.114529	10.018231	10.003332	9.809068
1255_g_at	5.056204	4.759066	4.629297	4.673458
1294_at	8.009337	7.980694	8.343183	8.025335
1316_at	6.899290	7.045843	6.976185	7.063050
1320_at	7.218898	7.600437	7.433031	7.201984
1405_i_at	6.861933	6.042179	6.165090	6.200671
1431_at	5.073265	5.114023	5.159933	5.063821
...				

DE in microarray/RNA-seq

- Large body of works, many are highly cited:
  - SAM/limma: for microarray
  - DESeq/edgeR/cufflink: for RNA-seq
- The focus is to overcome the problem of small sample size, which leads to unstable variance estimation.
  - Gene-by-gene t-test is not good.

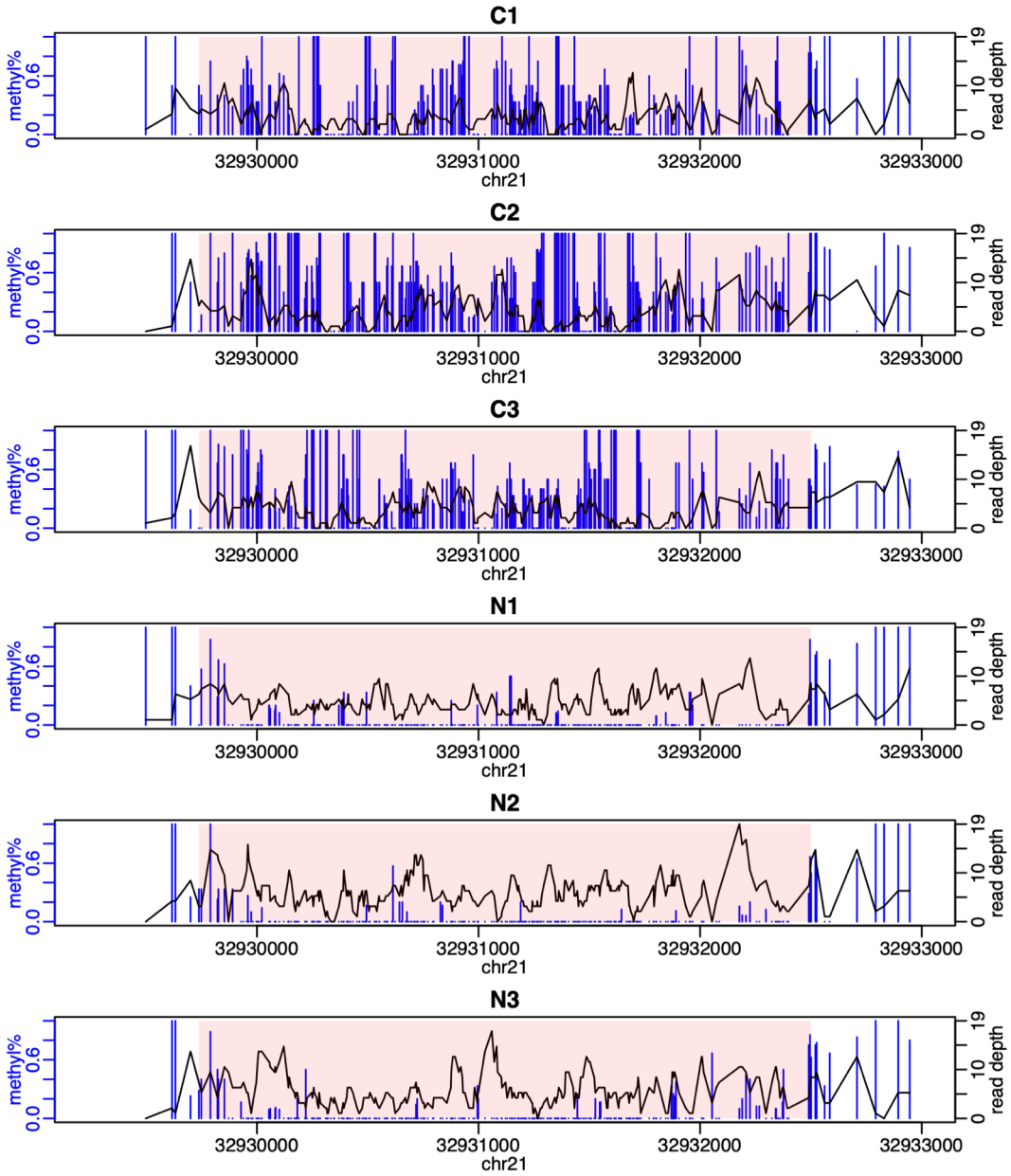
Volcano plot for DE



Supervised feature selection – differential methylation (DM)

- Goal: compare methylation levels between/among groups.
- Typical approach: perform hypothesis test on each CpG sites.
- Popular DM Methods:
  - Microarray: minfi.
    - Essentially t-test/linear regression on beta values.
  - Bisufite sequencing: bsseq, DSS
    - Smoothing on methylation levels.
    - Use beta-binomial model for counts.

An example DMR



Supervised feature selection – GWAS

- GWAS (Genome-wide association study): identify genetic variants associated with outcome of interest.
  - Genetic variants: mostly SNPs.
  - Outcome: can be continuous (weight, height, blood pressure) or categorical (disease status).
  - Typical approach: regression at each SNP, i.e., outcome~SNP+covariates

An example of disease-associated SNP

Cases:

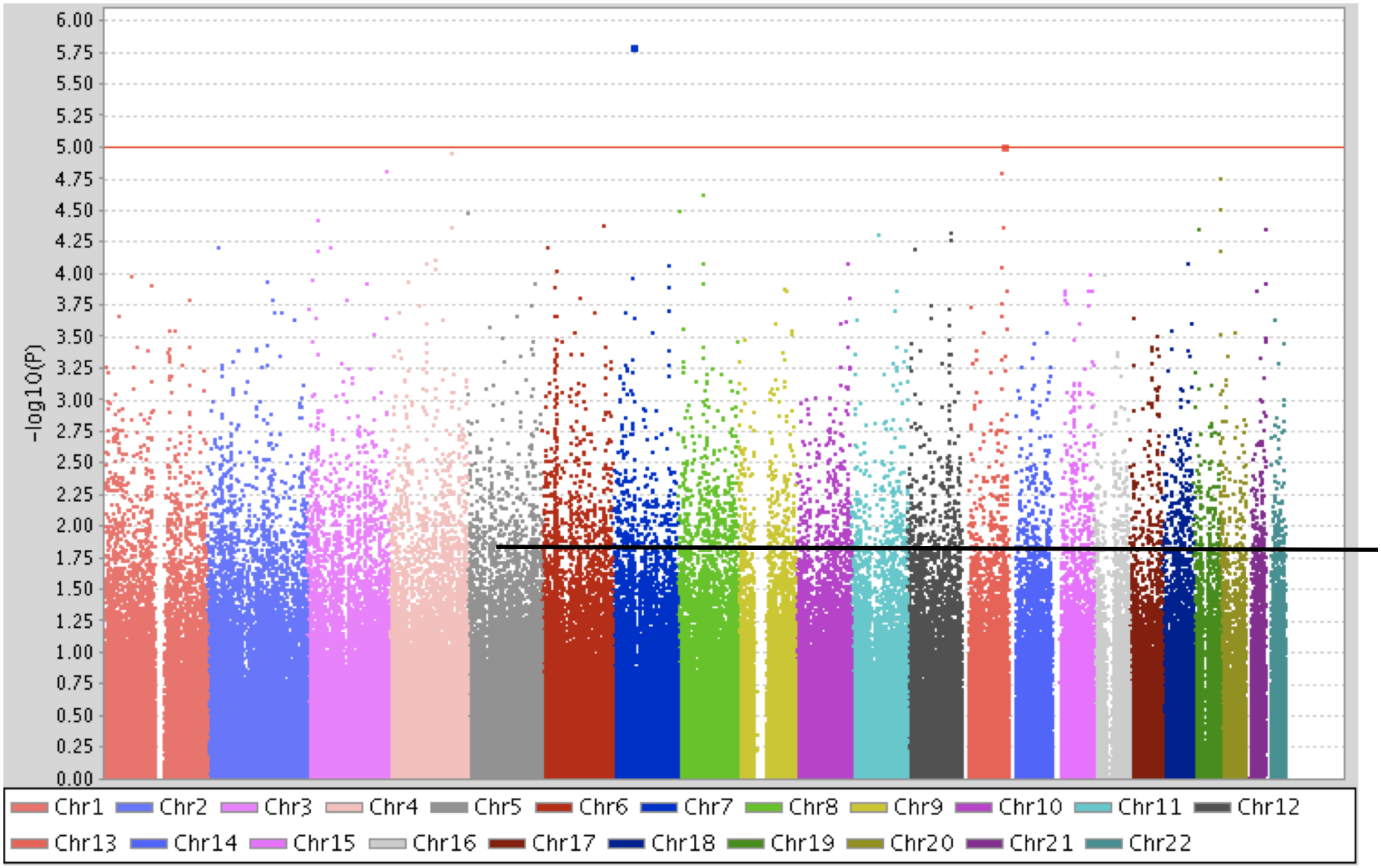
AGAGC**A**GTCGACA**G**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC**G**ACATGATAG**C**C
AGAGC**C**GTCGACA**T**GTATAG**T**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**A**GTGAGATC**G**ACATGATAG**T**C
AGAGC**A**GTCGACA**G**GTATAG**T**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC**G**ACATGATAG**C**C
AGAGC**A**GTCGACA**G**GTATAG**C**CTACATGAGATC**A**ACATGAGATC**G**GTAGAGC**A**GTGAGATC**G**ACATGATAG**C**C
AGAGC**C**GTCGACA**T**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC**A**ACATGATAG**C**C
AGAGC**C**GTCGACA**T**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**A**GTGAGATC**A**ACATGATAG**C**C
AGAGC**C**GTCGACA**G**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**A**GTGAGATC**A**ACATGATAG**T**C
AGAGC**A**GTCGACA**G**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**T**GTAGAGC**C**GTGAGATC**G**ACATGATAG**C**C

Associated SNP

Controls:

AGAGC**A**GTCGACA**T**GTATAG**T**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**A**GTGAGATC**A**ACATGATAG**C**C
AGAGC**A**GTCGACA**T**GTATAG**T**CTACATGAGATC**A**ACATGAGATC**T**GTAGAGC**C**GTGAGATC**G**ACATGATAG**C**C
AGAGC**A**GTCGACA**T**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**T**GTAGAGC**C**GTGAGATC**A**ACATGATAG**C**C
AGAGC**C**GTCGACA**G**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**T**GTAGAGC**C**GTGAGATC**G**ACATGATAG**T**C
AGAGC**C**GTCGACA**G**GTATAG**T**CTACATGAGATC**G**ACATGAGATC**T**GTAGAGC**C**GTGAGATC**A**ACATGATAG**C**C
AGAGC**A**GTCGACA**G**GTATAG**T**CTACATGAGATC**G**ACATGAGATC**T**GTAGAGC**A**GTGAGATC**G**ACATGATAG**C**C
AGAGC**C**GTCGACA**G**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**T**GTAGAGC**C**GTGAGATC**G**ACATGATAG**C**C
AGAGC**C**GTCGACA**G**GTATAG**T**CTACATGAGATC**A**ACATGAGATC**T**GTAGAGC**A**GTGAGATC**G**ACATGATAG**T**C

“Manhattan plot” for GWAS results



Sparse learning

- For methods we discussed so far, features are selected one-by-one.
- Another type of approach is to feed large number of features in a model to select.
  - This will consider co-linearity among features.
  - Better for prediction.
- Sparse learning: a class of methods for finding a sparse representation of the input data as a linear combination of predictors.

LASSO

- One type of sparse learning: regularized regression by LASSO (least absolute shrinkage and selection operator)
  - Put a large number of predictors in a regression model, penalize the regression coefficients with L1 penalty.
  - Depending on the penalty strength, some coefficients will be shrunk to 0, thus deselected.

$$\min_{\beta_0,\beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

Unsupervised feature selection

- Without outcome, or outcome information are not used.
- Often used for unsupervised sample clustering, e.g., identify cancer subtypes.
- Goal is to find a low dimension representation of high dimension data.

Approaches for unsupervised feature selection

- Most common: Select N features with largest "variation".
  - Features behave similarly among samples are not informative.
- Definition of variation can vary:
  - Sample variance (normal distribution)
  - Coefficient of variation (when there's mean-variance dependence, like count data in RNA-seq)
  - Gini index: select ones with outliers. This was used in detecting rare cell type in scRNA-seq (Jiang et al 2016 GB).

Problems in variation-based feature selection

- One wants to select features having
  - Large between-group variation
  - Small within-group variation
- Large marginal variation can be caused by large within-group variation.

Approaches (cont.)

- Another approach:
  1. Select features based on variation
  2. Generate cluster labels via clustering
  3. Transform unsupervised feature selection into supervised feature selection with these generated cluster labels
- Problem:
  - This can artificially makes cluster tighter.
  - If the initial clustering is wrong, this will make it more wrong.

From basic to high-order features

- Now we have basic features (genes, CpG sites) selected.
- We can further process the features to produce better representation of the data
  - Grouping the features to improve power.
  - Transform the features to further reduce dimension.
  - Combine features from different data modalities.

Feature groups

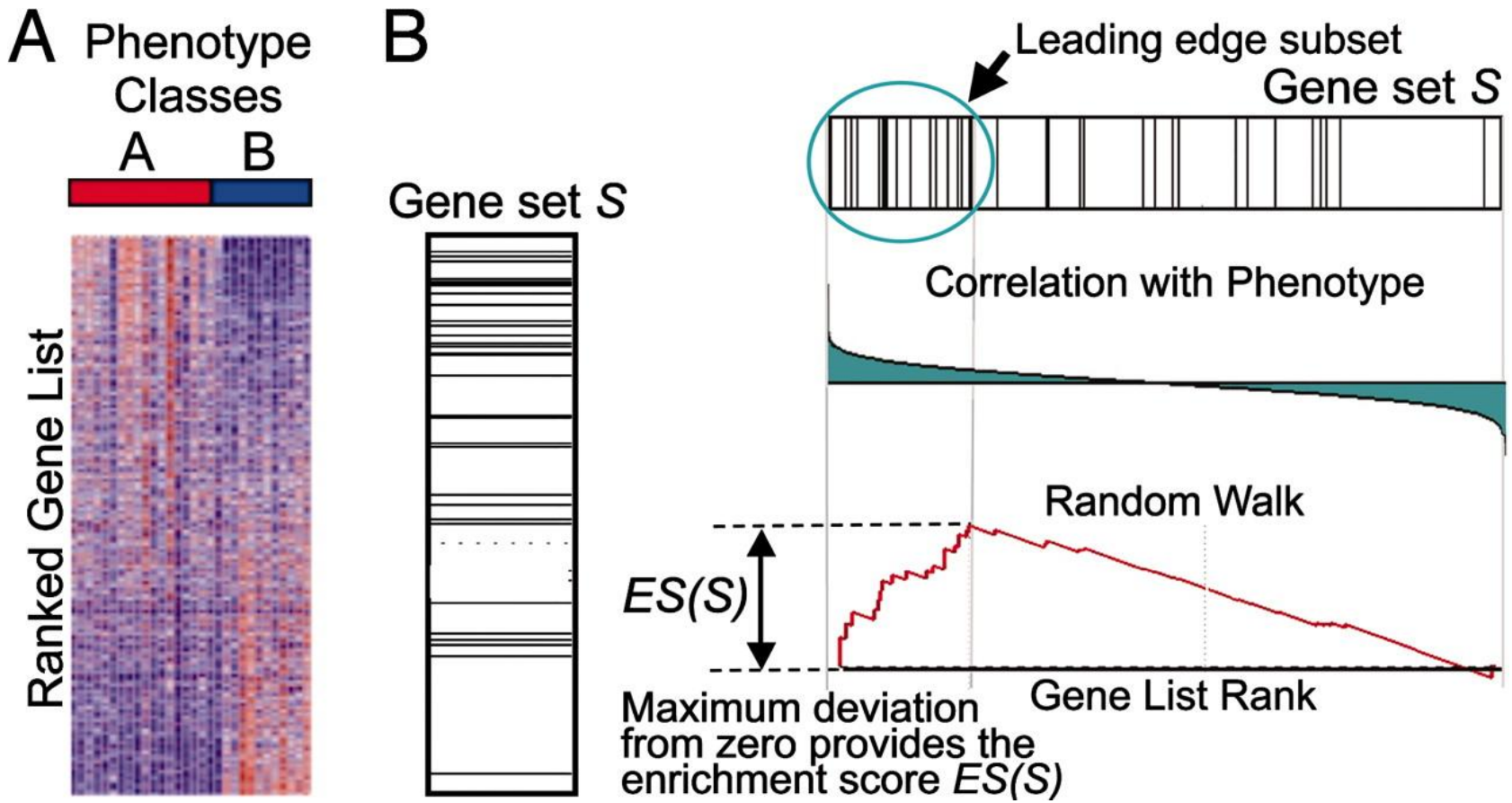
- Sometimes the statistical power is low in supervised feature selection, due to
  - Small sample size.
  - Sparse effect: rare event (small proportion of cases have aberrant SNP, GE, or methylation).
- Group features and aggregate data can boost power. Examples:
  - Expression: gene set enrichment analysis (GESA).
  - GWAS: burden test or SKAT.
  - DNA methylation: CpG clusters.

Feature group in gene expression – GSEA

Subramanian *et. al.* (2005) PNAS

- Motivation:
  - DE analysis has low power, no or few DE genes detected.
  - Combine cumulative effects from many slightly altered genes.
  - "*An increase of 20% across all genes encoding members of a metabolic pathway...may be more important than a 20-fold increase in a single gene*"
- Approach:
  - Given a set of genes ***S*** and the whole gene list ***L*** ranked by significance from DE test.
  - Question: is ***S*** randomly distributed in ***L***.
  - Solution: Kolmogorov-Smirnov test.
  - Software can do GSEA: DAVID, Enrichr, MSigDB, etc.
  - Gene set can be defined by pathways or functional groups.

GSEA


A
Phenotype
Classes
A
B
Ranked Gene List
B
Gene set S
Leading edge subset
Gene set S
Correlation with Phenotype
Random Walk
ES(S)
Gene List Rank
Maximum deviation
from zero provides the
enrichment score ES(S)

Feature group in GWAS

- Mainly used to deal with rare variant problem.
- Burden test: Li and Leal (2008) *AJHG*, Madsen and Browning (2009) *Plos Genetics*
  - Group variants to improve power.
  - Variant group is pre-determined, for example, on the same gene
  - Collapsing data using weighted sums.
- SKAT (Sequence Kernel Association Test): Wu *et al.* (2011) *AJHG*
  - Aggregates individual score test statistics of SNPs in a SNP set to get SNP-set level p-values.
  - Based on a linear mixed model.

High-order features

- We can transform a set of basic features to get "high-order" features.
  - High-order features are functions (can be nonlinear) of the basic features.
  - Popular methods for such transformation: PCA, ICA, tSNE, etc.
- Pros and cons
  - Further reduce data dimension.
  - Can be more robust in prediction.
  - Lost biological meaning.

Combining different types of features - Multimodal feature fusion

- Different data type has different number of features, e.g., 25k genes, 28 million CpG sites.
- It's not easy to construct correspondence between different types of features.
- Method for feature fusion:
  - Simplest one: stack up the features.
  - Based on factor analysis: CCA (canonical correlation analysis).
  - Based on machine learning methods, e.g., support vectors.

Artifacts in high-throughput data

- High-throughput data are noisy
- Data need to be carefully preprocessed and normalized before feature selection.
- Sometimes data transformation (i.e., log) is helpful.
- There are many data normalization methods (we didn't touch that in this lecture):
  - QN, RMA, and GCRMA for GE microarray.
  - TMM, CQN for RNA-seq.
  - scran, scNorm, etc. for scRNA-seq.

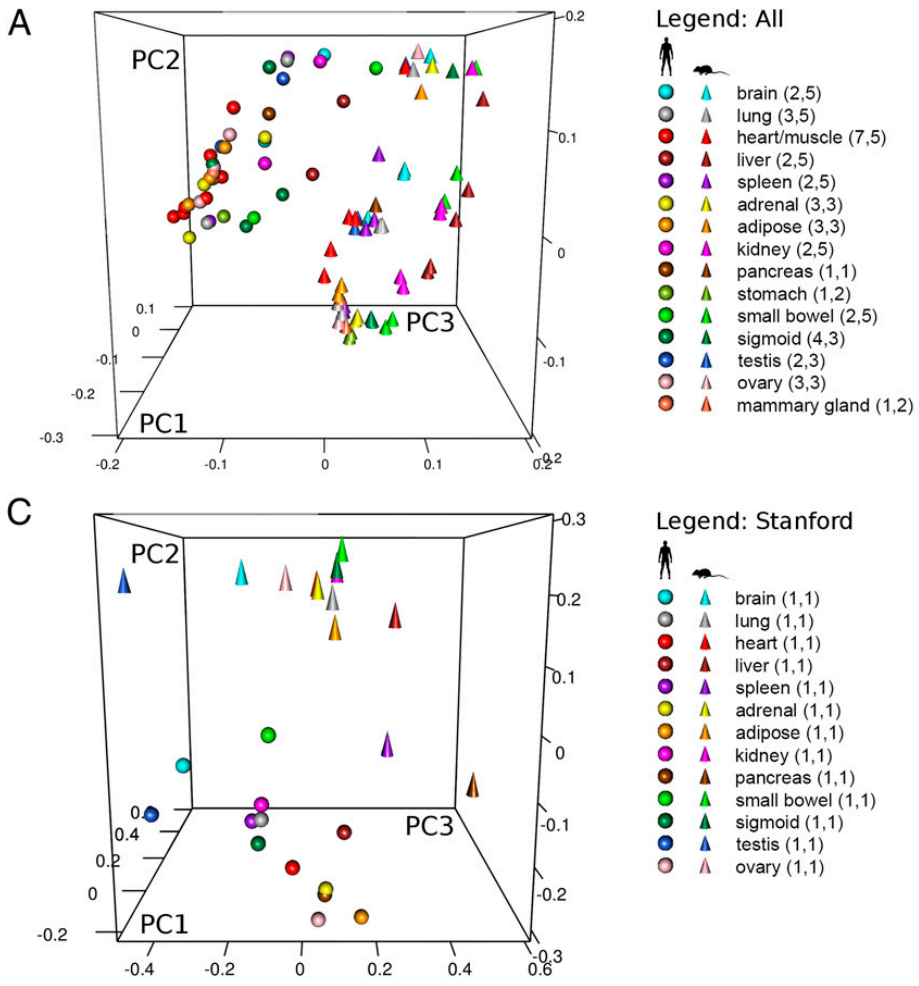
Technical artifact – batch effect

- HT experiments are very sensitive to experimental conditions:
  - Equipment, agents, technicians, etc.
- Data generated from different “batches” (lab, time, etc.) can be quite different, but data from the same batch tend to be more similar.
- Methods for identifying and removing batch effects is under continuous developments.

Comparison of the transcriptional landscapes between human and mouse tissues

Shin Lin[a,b,1], Yiing Lin[c,1], Joseph R. Nery[d], Mark A. Urich[d], Alessandra Breschi[e,f], Carrie A. Davis[g], Alexander Dobin[g], Christopher Zaleski[g], Michael A. Beer[h], William C. Chapman[c], Thomas R. Gingeras[g,i], Joseph R. Ecker[d,j,2], and Michael P. Snyder[a,2]

- One major conclusion is that tissues are more similar within a species, compared with the same tissue across species.



RESEARCH ARTICLE

A reanalysis of mouse ENCODE comparative gene expression data [version 1; referees: 3 approved, 1 approved with reservations]

Yoav Gilad, Orna Mizrahi-Man

Department of Human Genetics, University of Chicago, Chicago, IL, 60637, USA

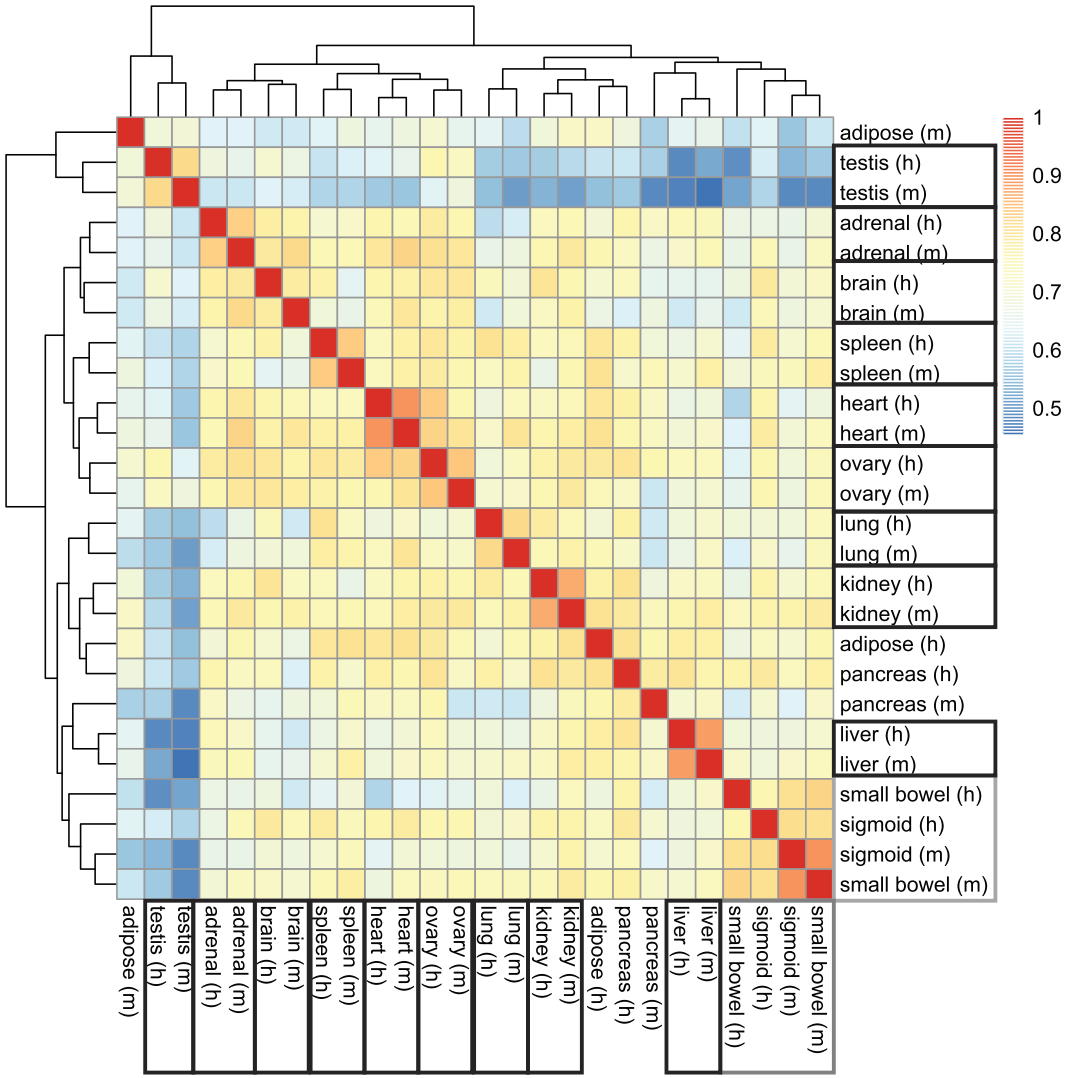
- Experimental design: data are from 5 batches.

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

After correcting for batch effects

- Tissues tend to cluster together more.

b

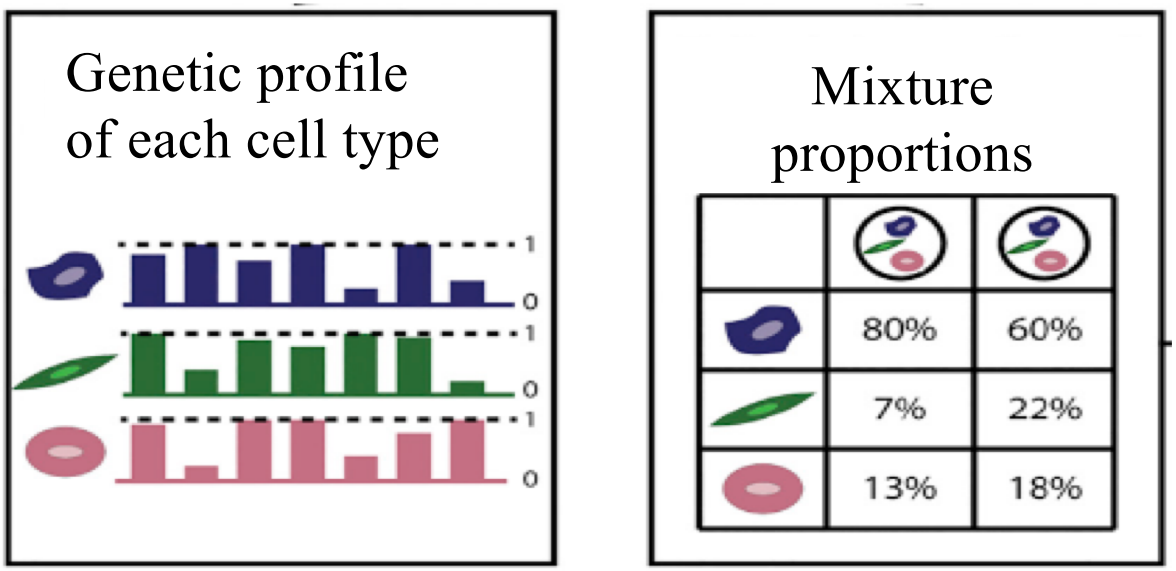


Methods to remove batch effects

- Based on linear model: batches cause location/scale changes, e.g., Combat (Johnson et al. 2007 *Biostatistics*).
- Based on dimension reduction technique: SVD, PCA, factor analysis, etc., e.g., sva (Leek et al. 2007 *PloS Genet*.).
  - The singular vectors/PCs/factors that are correlated with batch are deemed from batch effects.
  - Remove batch effects from data, leftovers are biological signals.
- The key is to find a good "baseline" for normalization:
  - Feature selection also plays important role: one wants to find features not correlated with batch: RUV (Gagnon-Bartsch and Speed 2012 *Biostatistics*, Risso et al. 2014 *Nature Biotech*)

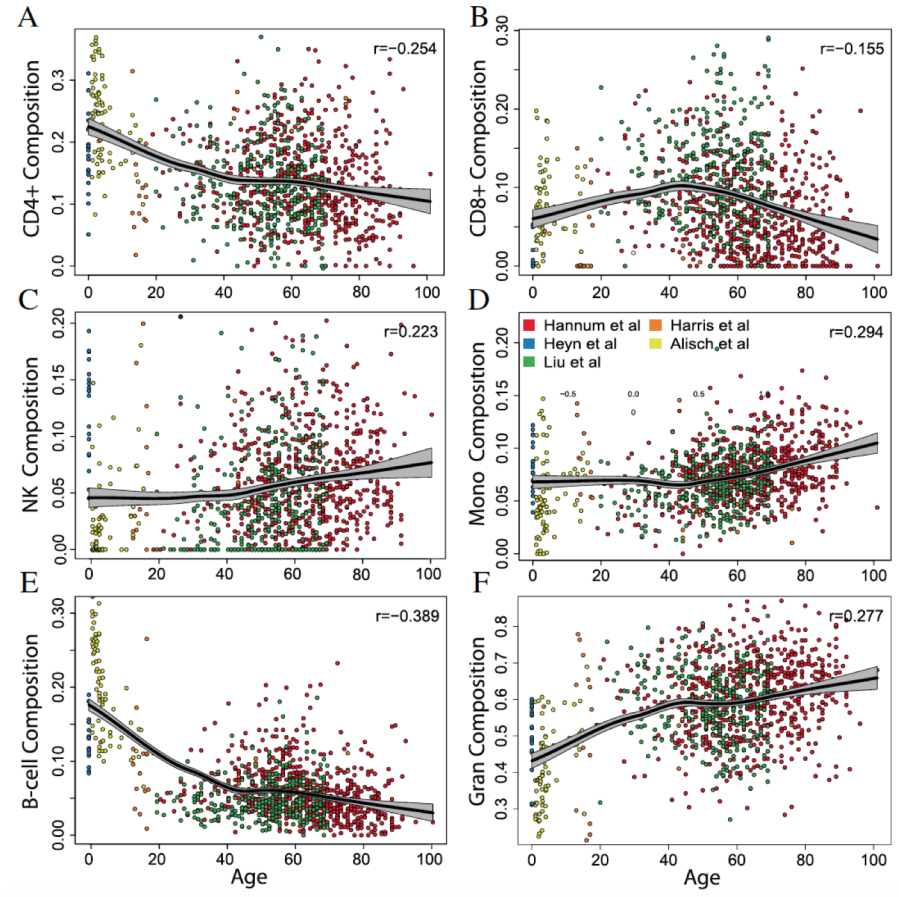
Biological artifact – tissue heterogeneity

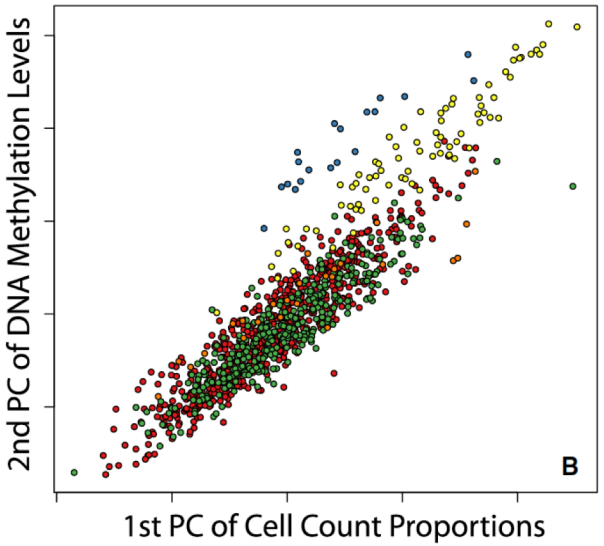
- Tissue sample is often a mixture of different cell types.
- Data collected are mixed signals.



An example: EWAS in aging study

- Cellular composition changes with age.
- Cellular composition is a major source of variability in DNA methylation datasets in whole blood.





Jaffe and Irizarry GB(2014)

Existing signal deconvolution methods

- **Reference-based** methods (some type of regression):
  - Require cell type specific signature: Abbas et al. 2009; Clarke et al. 2010; Gong et al. 2011; Lu et al. 2003; Wang et al. 2006; Vallania et al. 2018; Du et al. 2018;
  - Requires mixture proportions: Erkkila et al. 2010; Lahdesmaki et al. 2005; Shen-Orr et al. 2010; Stuart et al. 2004.
- **Reference free** methods (some type of factor analysis):
  - Gaujoux et al. 2011; Kuhn et al. 2011; Repsilber et al. 2010; Roy et al. 2006; Venet et al. 2001; Houseman et al. 2012, 2014, 2016; Rahmani et al. 2016, 2018; Lutsik et al. 2017; Xie et al. 2018;

Method to adjust for cell proportion

- In EWAS, add proportion as covariate in the model:
- More rigorous statistical modeling for DE/DM with sample mixture has been a popular topic recently, and a number of methods are developed:
  - csSAM: Shen-Orr et al. 2010, *Nature methods*
  - CellDMC: Zheng et al. 2018, *Nature Methods*
  - TOAST: Li et al. 2019, *Bioinformatics*

Rule of thumbs for genomic feature selection

- Understand your data:
  - Supervised vs unsupervised.
  - Data normalization to remove artifacts.
- Understand your goal:
  - To understand mechanism or look for drug target – identify basic features (DE, DM, GWAS, EWAS), or group of features (GSEA)
  - For outcome prediction – group features or high-order features.
- Choose proper tool(s) to achieve your goal.

Summary

- Goals of feature selection in high-throughput genomics data
  - To identify biomarkers for treatment
  - Find predictors for diagnostic model
- Methods
  - Feature-by-feature test: find ones correlated with outcome
  - Groups of features
  - Higher order features
- Other considerations: artifacts in the data