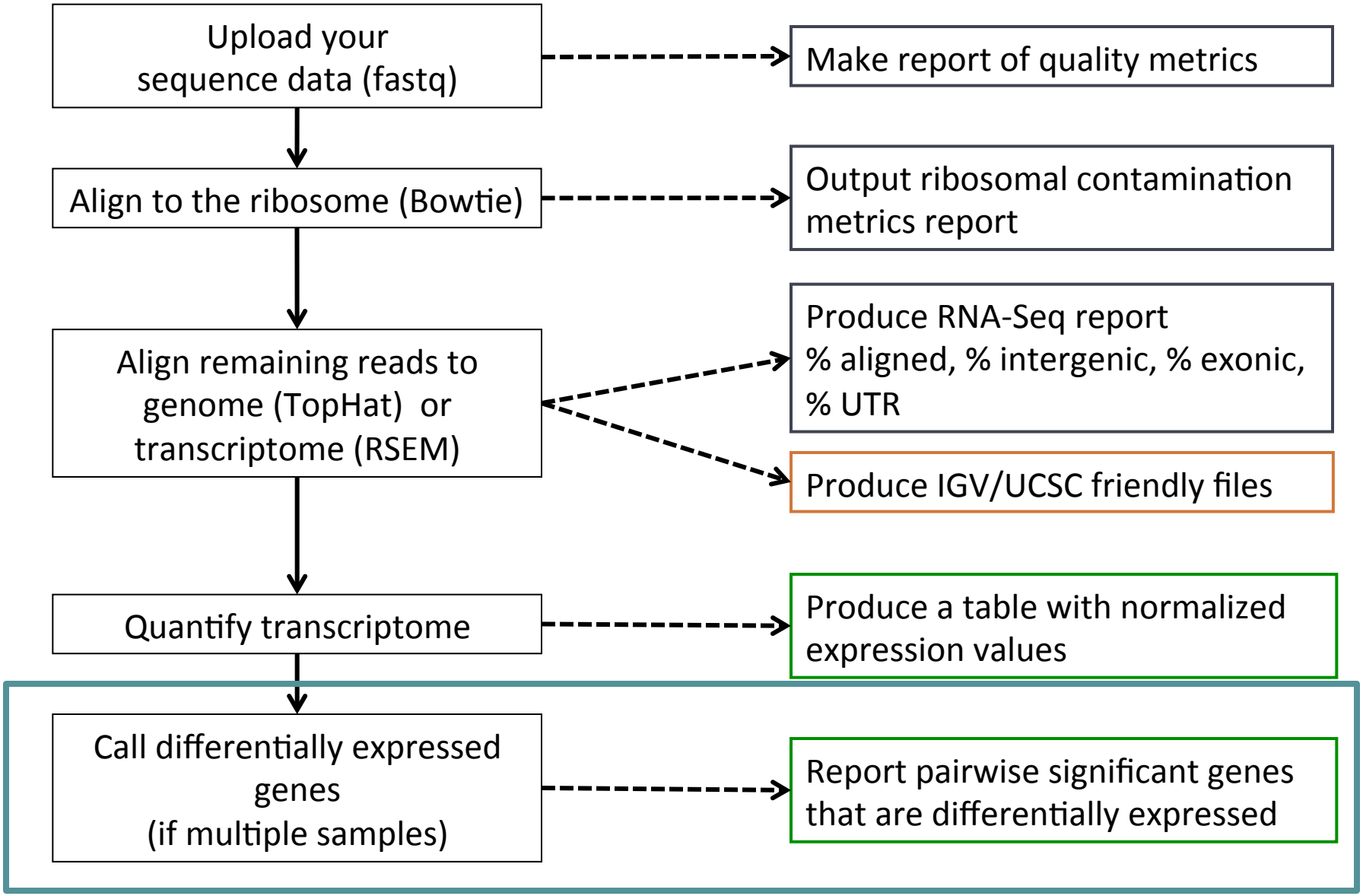
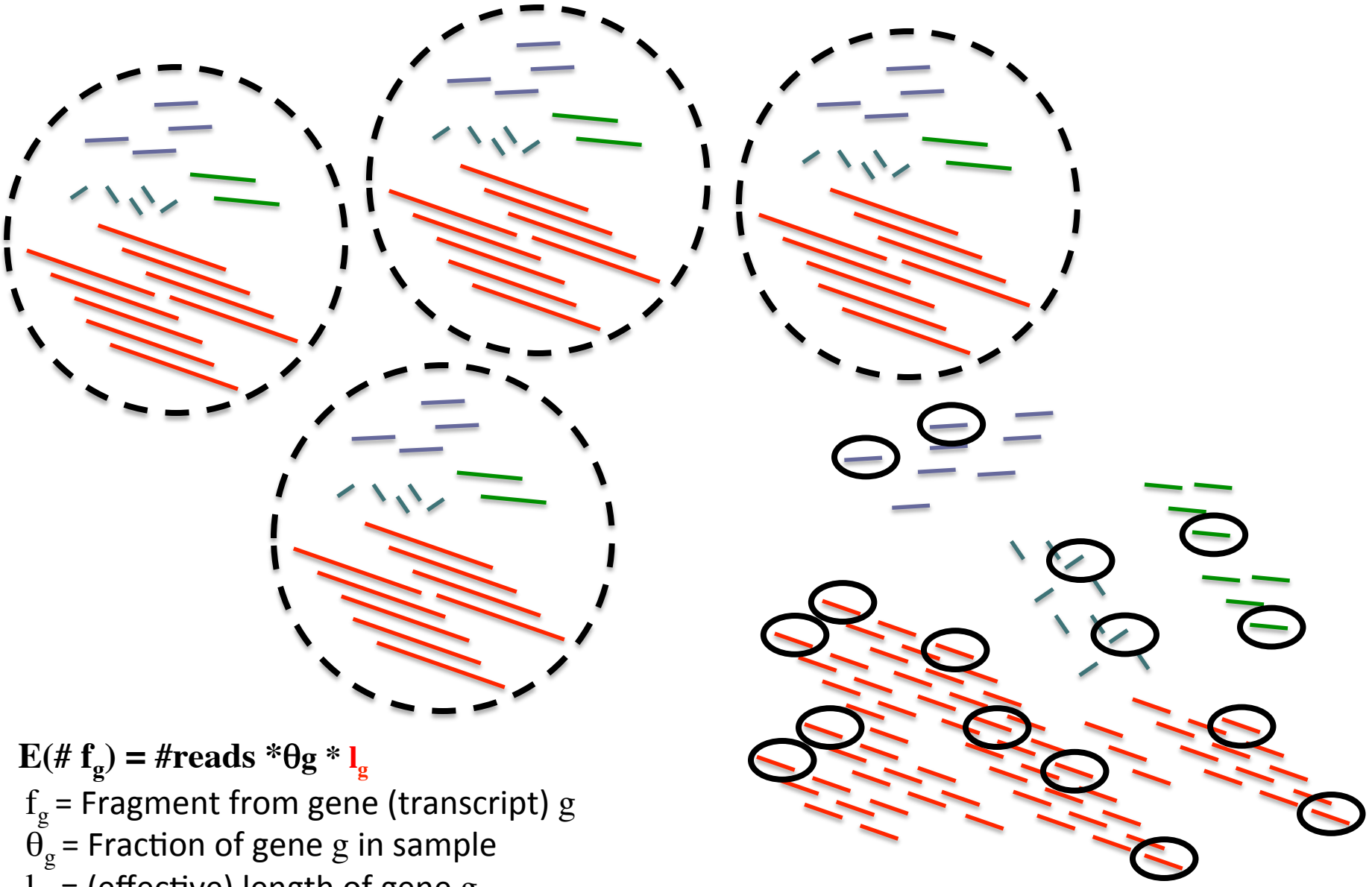
Analysis of count data

Our typical RNA quantification pipeline



Were we left….



$\mathbf{E(\#\ f_g) = \#reads * \theta g * l_g}$

f_g = Fragment from gene (transcript) g

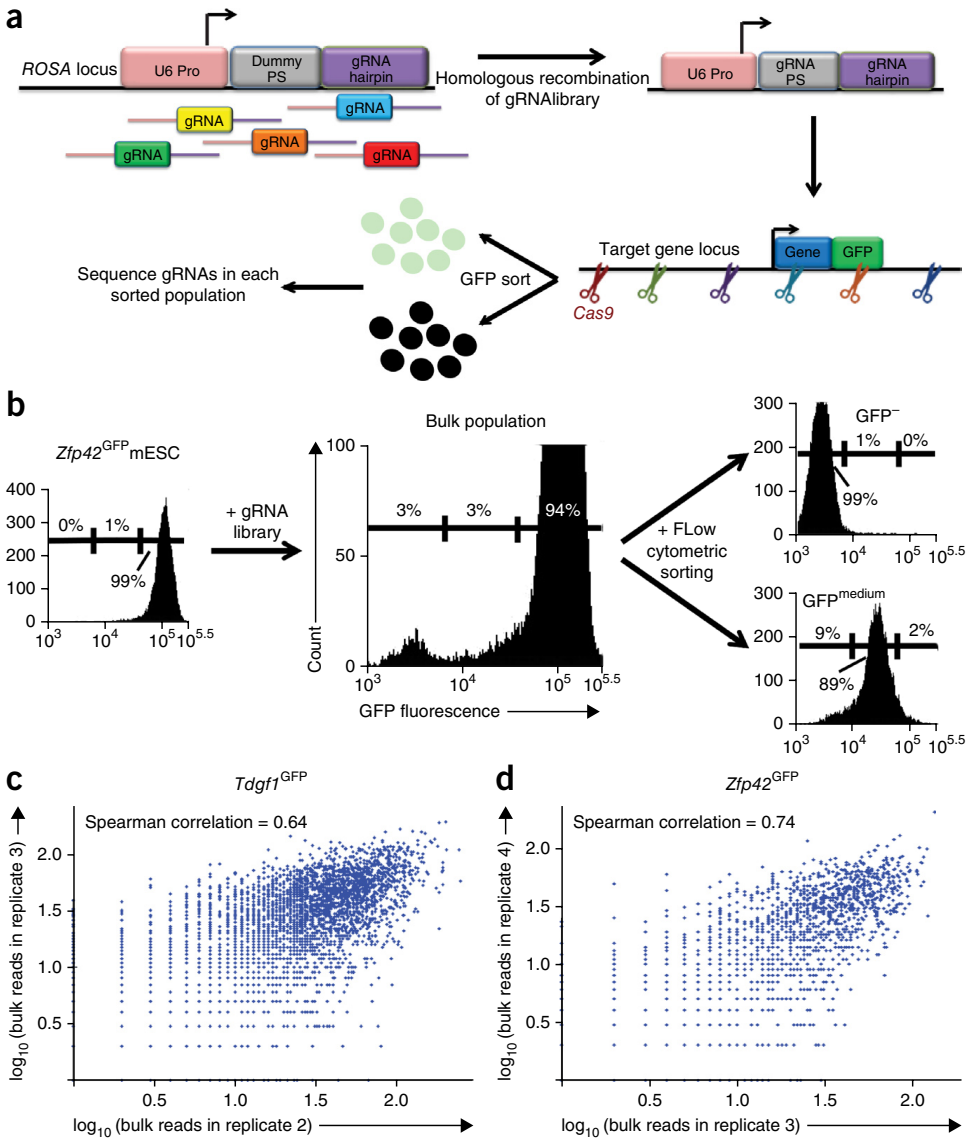
θ_g = Fraction of gene g in sample

l_g = (effective) length of gene g

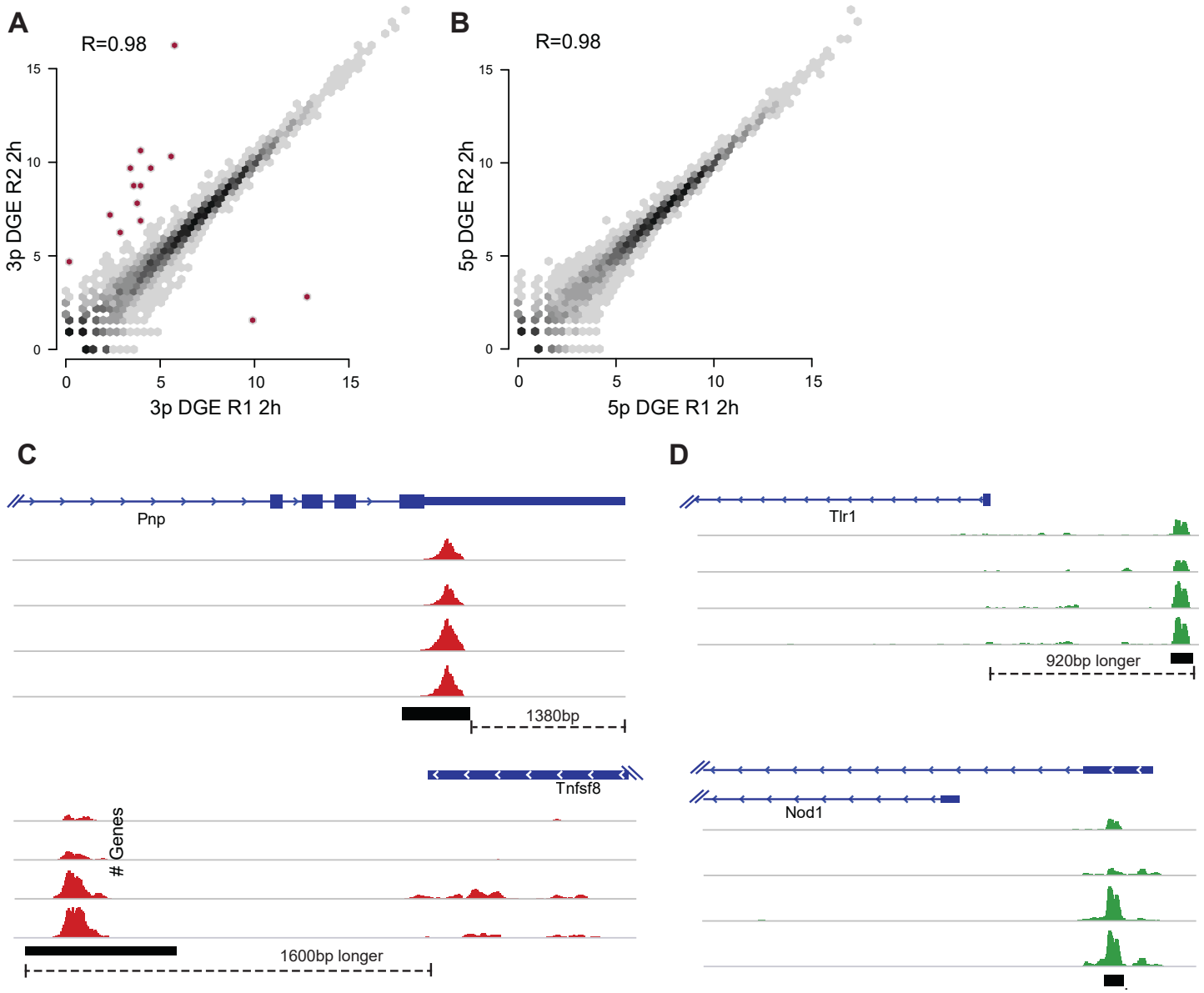
Revisiting correlation

High-throughput mapping of regulatory DNA

Nisha Rajagopal[1], Sharanya Srinivasan[1,2], Kameron Kooshesh[2,3], Yuchun Guo[1], Matthew D Edwards[1], Budhaditya Banerjee[2], Tahin Syed[1], Bart J M Emons[2,4], David K Gifford[1] & Richard I Sherwood[2]



Revisiting correlation



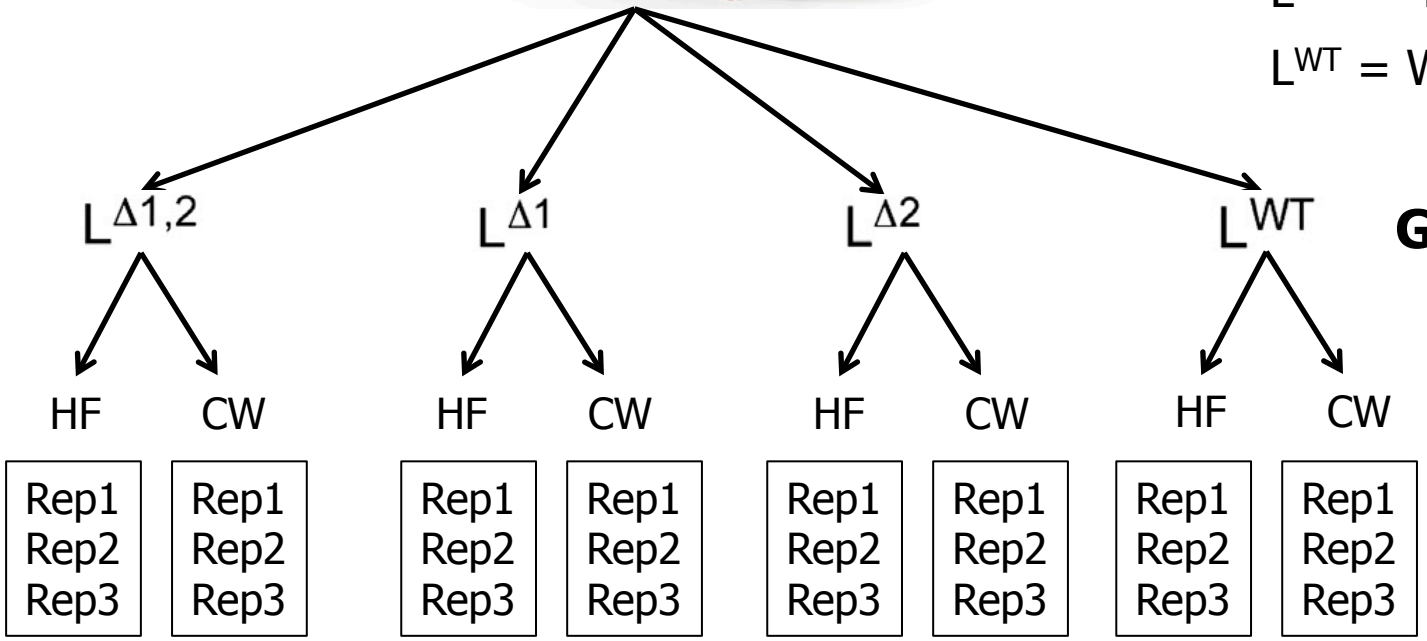
Experimental approach

$L^{\Delta 1}$ = Jnk1 KO

$L^{\Delta 2}$ = Jnk2 KO

$L^{\Delta 1,2}$ = Double KO

L^{WT} = Wildtype



24 samples: 4 Genotypes x 2 Diets x 3 replicates

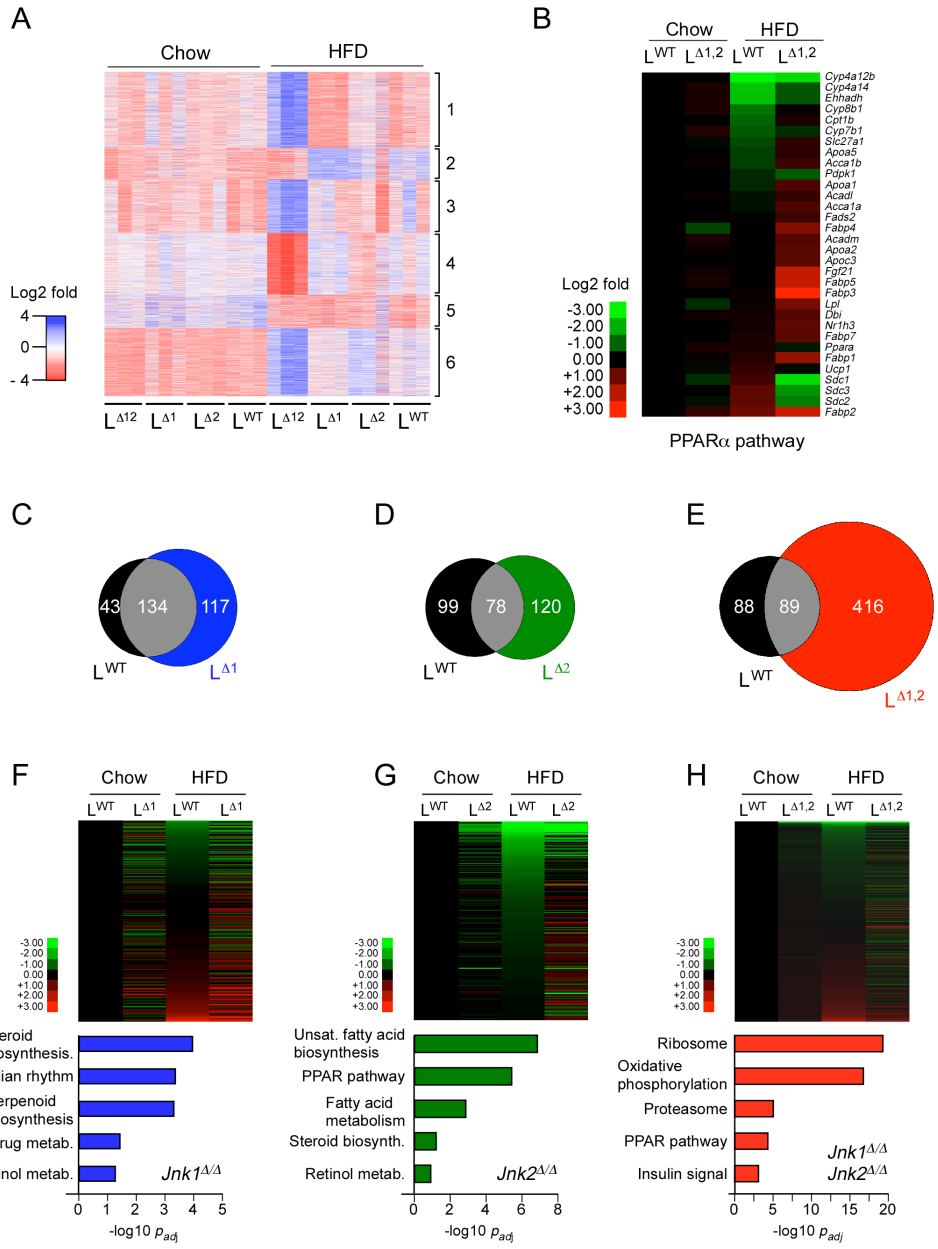


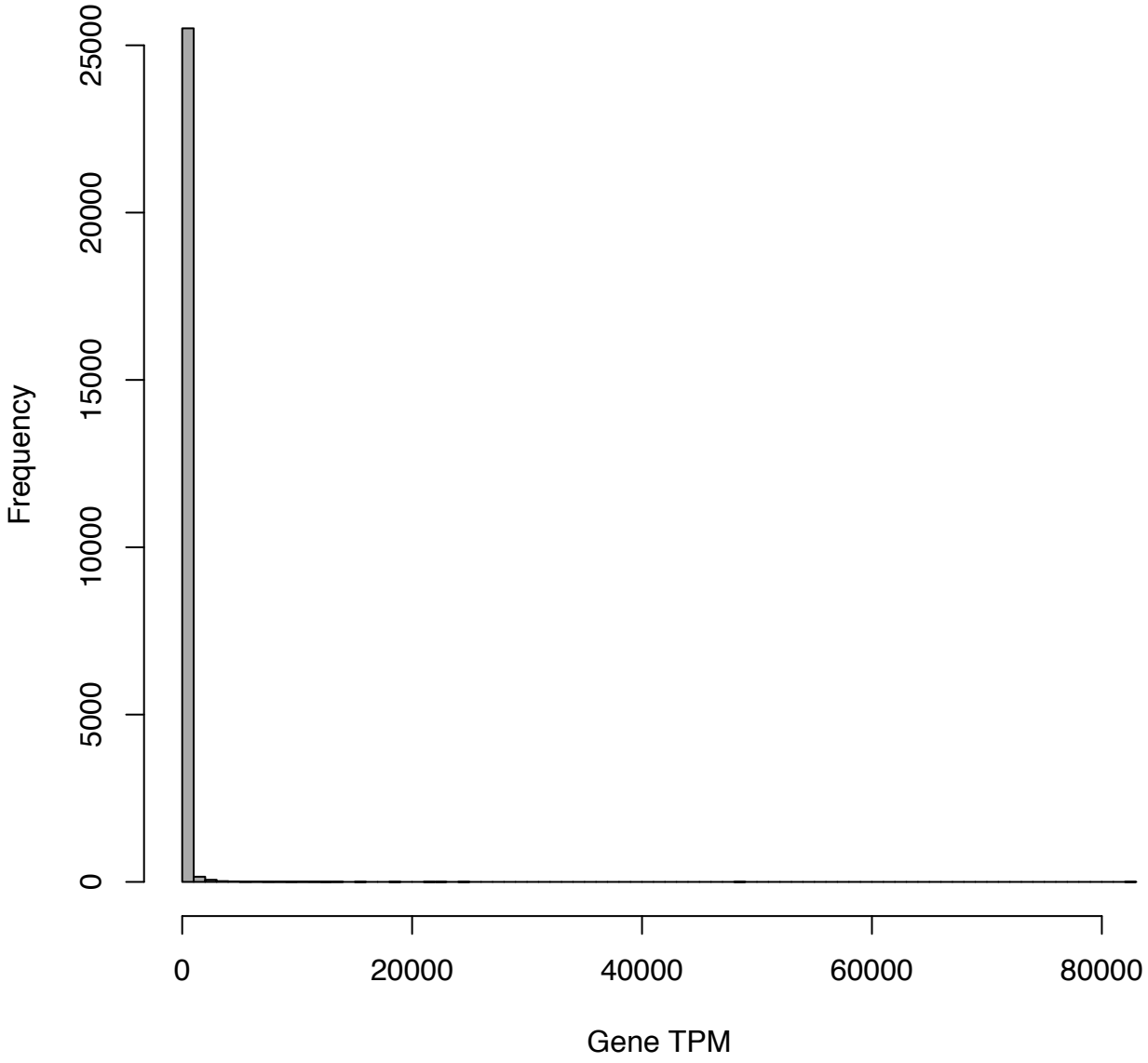
Figure S3. Analysis of hepatic genes differentially regulated by high fat diet in control and liver-specific JNK-deficient mice, Related to Figure 3.

The gene expression table

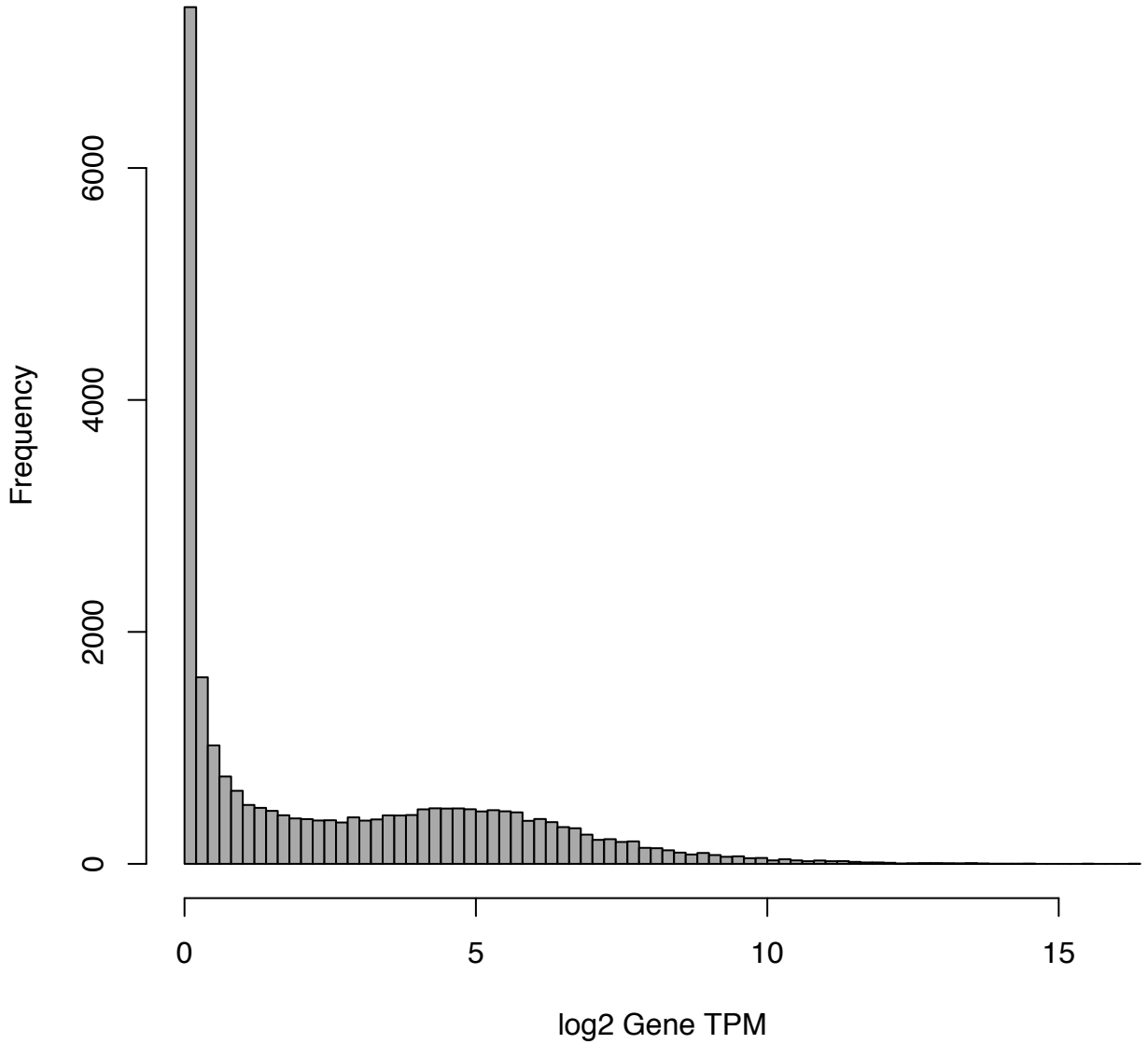
- Genes are quantified. Each gene or isoform has:
  - A TPM value
  - A (expected) fragment count vaue
- All samples were quantified in the same fashion and arranged into a table of genes (22,000) x samples (24).
  - Row i gives the expression of the gene i across all samples
  - Row j gives the expression of genes in sample j.

gene	$L^{\Delta 1,2}$.rep1	$L^{\Delta 1,2}$.rep2	$L^{\Delta 1,2}$.rep3	$L^{\Delta 1}$.rep1	$L^{\Delta 1}$.rep2	$L^{\Delta 1}$.rep3	$L^{\Delta 2}$.rep1	$L^{\Delta 2}$.rep2	$L^{\Delta 2}$.rep3
Mir301	0	0	0	0	0	0	0	0	0
Cpne2	157	158.98	88.04	69	111.99	114.33	93	208	140
Capn5	36	65	46	46	69	42	33	58	59.01
Lage3	313.06	241.23	276.23	218.9	285.19	359.65	269.7	359.04	417.47
Brd7	379	358.58	390	336	357.26	368.08	264	564.07	476
Dimt1	77	68	58	54	62	60	54	76	97.03
AK017068	0	0	0	0	0	0	0	0	0

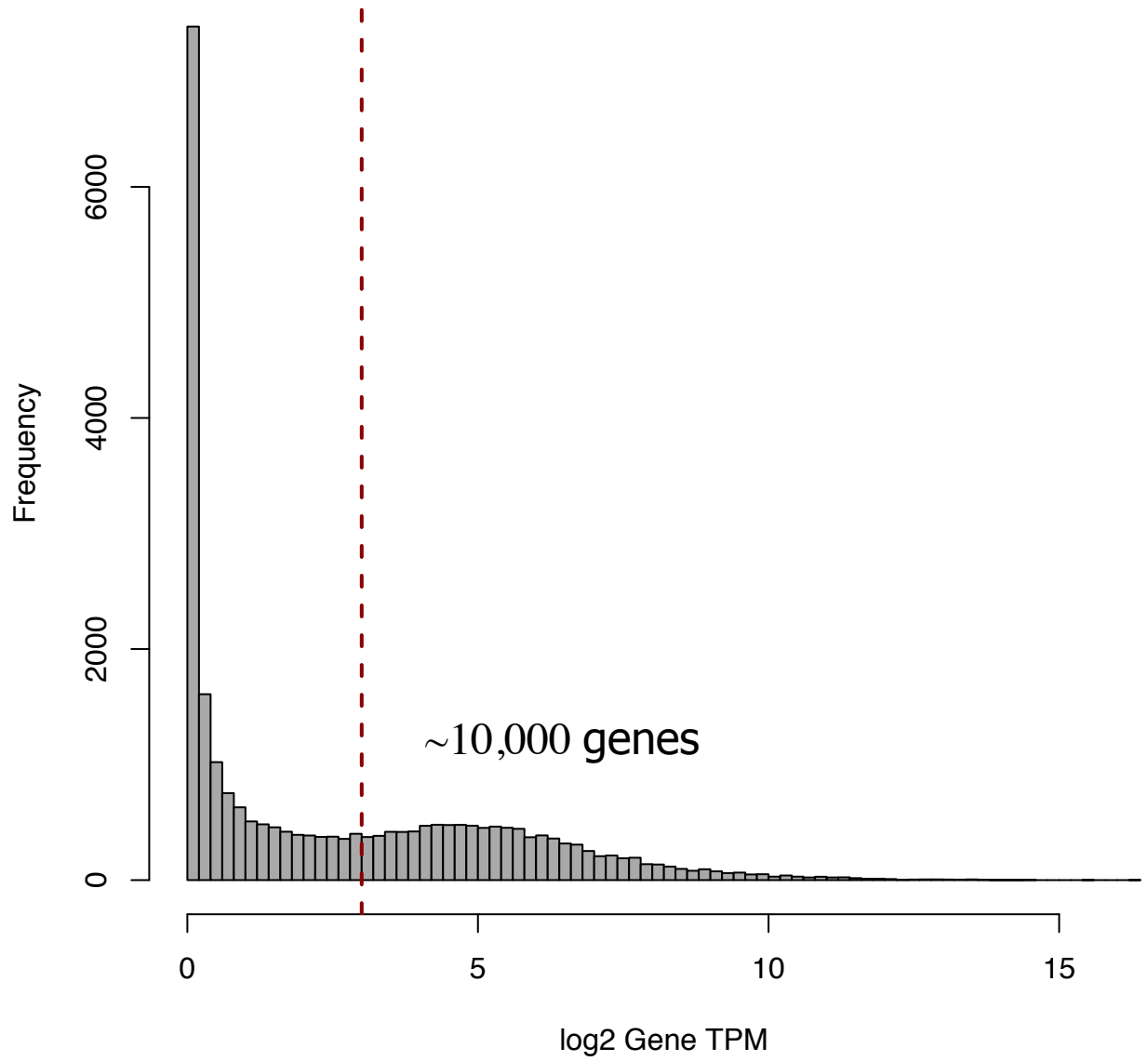
Q1: Who is expressed?



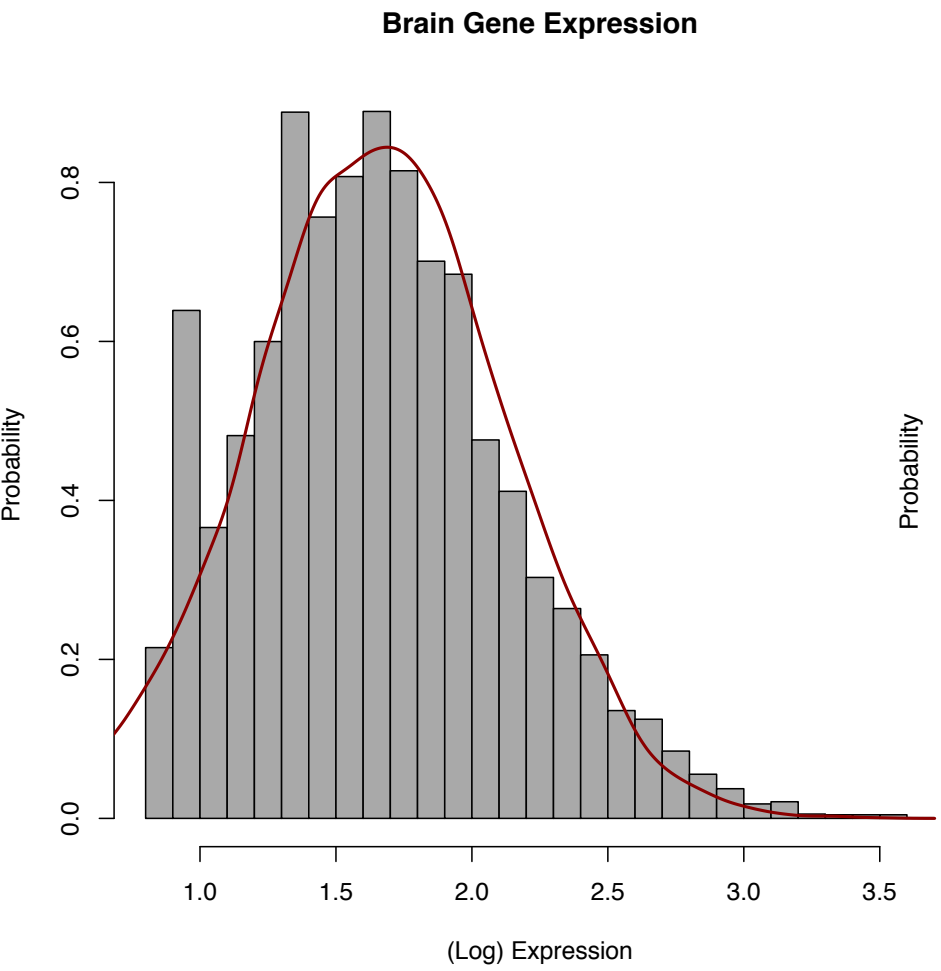
Q1: Who is expressed?

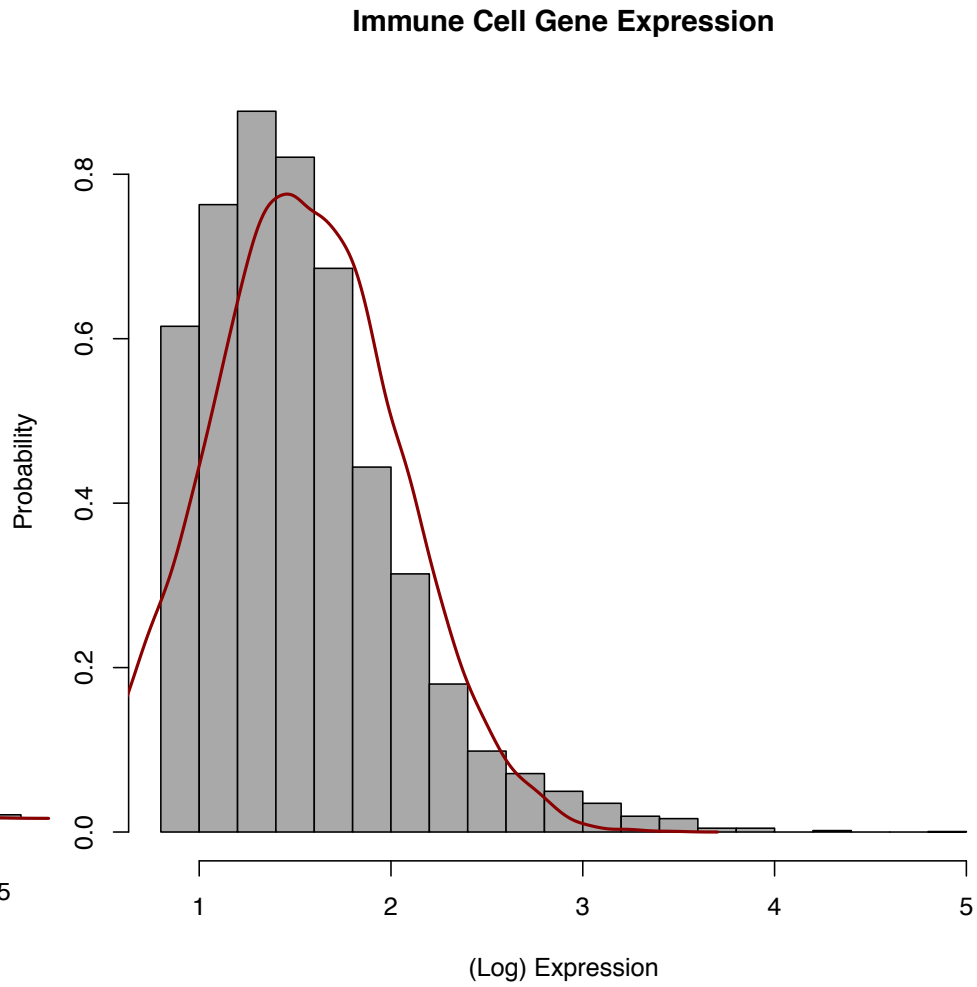


Q1: Who is expressed?



Q2: How do we model gene expression?

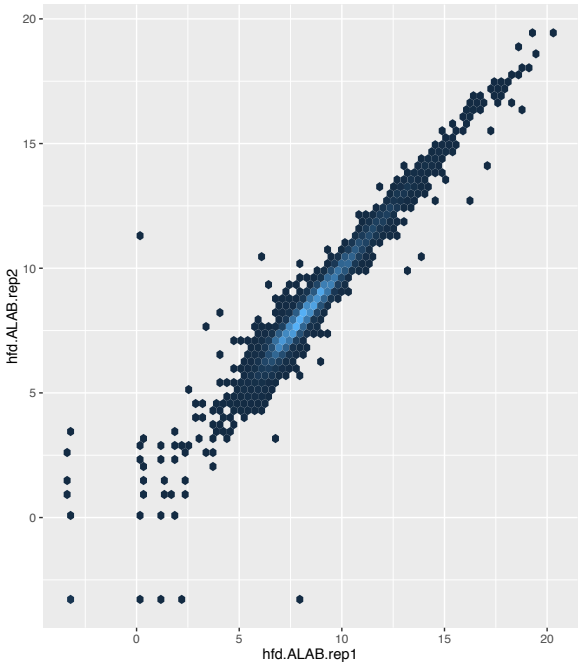


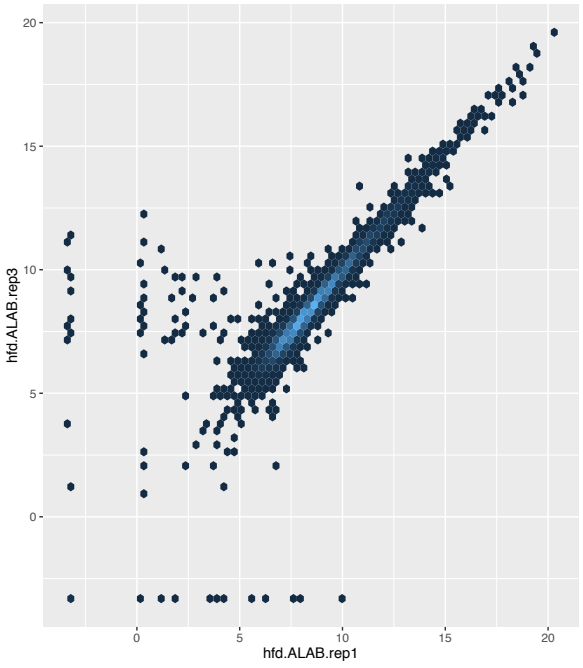


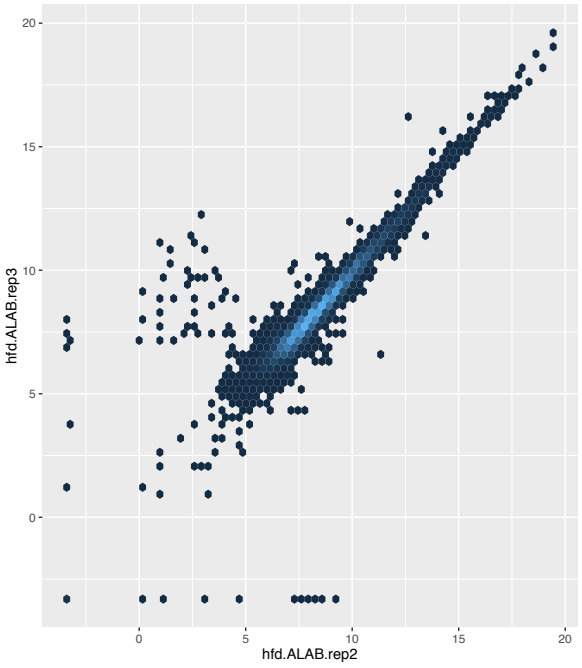
Normal distribution with mean and variance estimated from gene exp. data

Gene expression distributes roughly log-normal

Task1: How variable is my data?

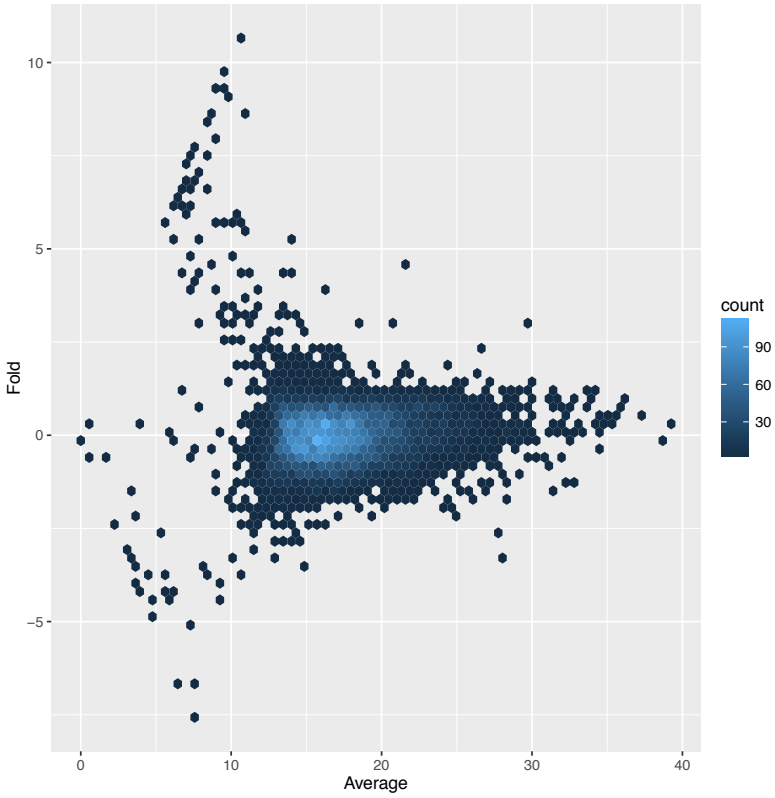




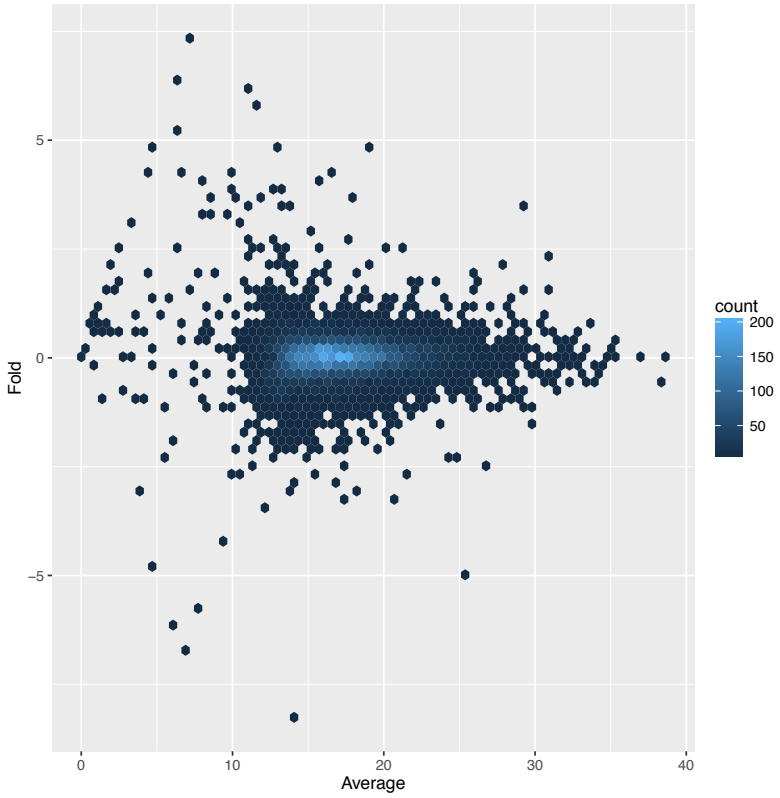


Task 2: What may be my power?

$L^{\Delta 1,2}$ vs L^{WT} High Fat diet



High Fat vs Normal diet

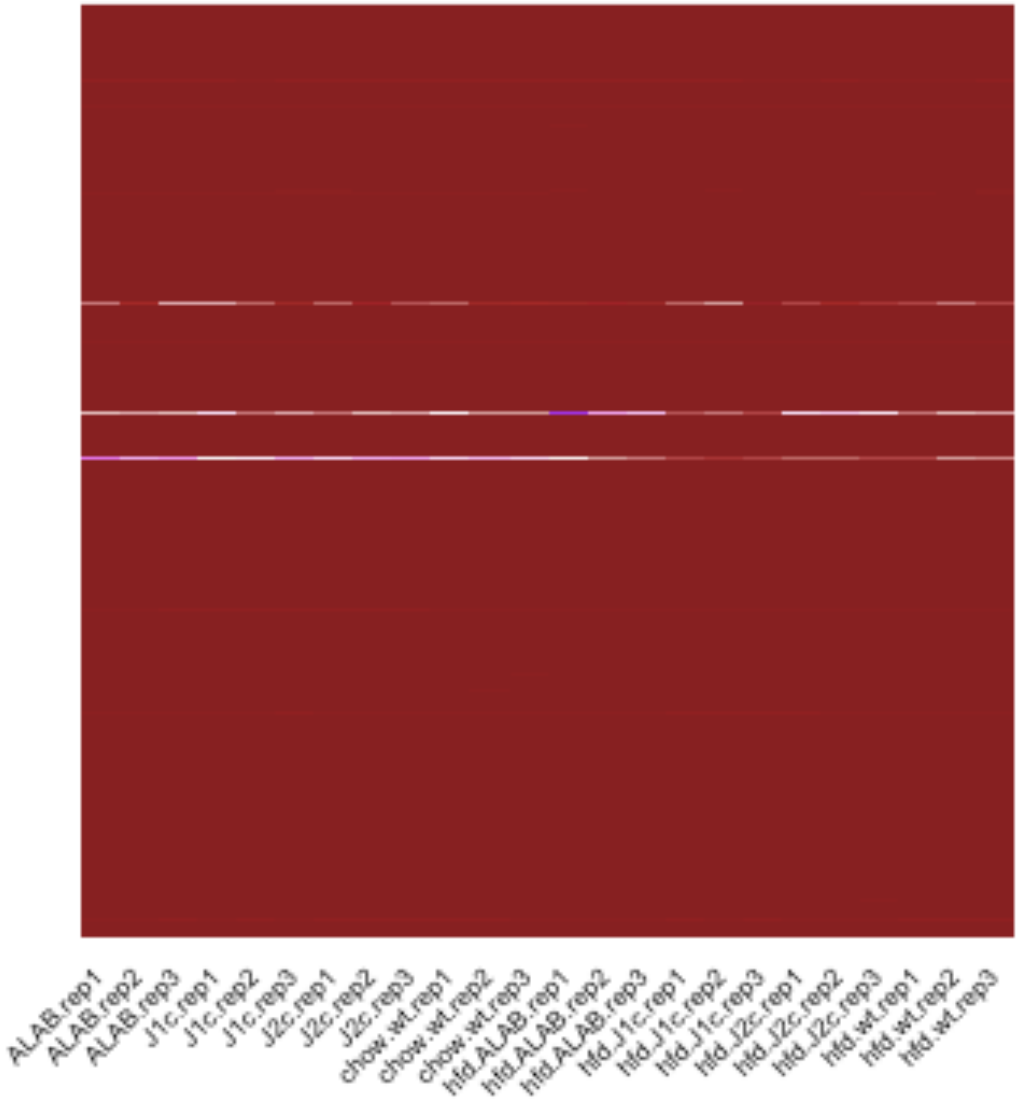


Task 3: How can I look at ALL my data at once?

- Matrices are too large, takes too much compute power:
  - Initial 25,000 x 24 = 500,000 entries
  - Filtering reduces by half: 250,000 entries
- We can focus on interesting genes. Which ones?
  - Genes that vary most are most informative, how do we find them?
    - Variance or standard deviation
    - Coefficient of variance: sd/mean!

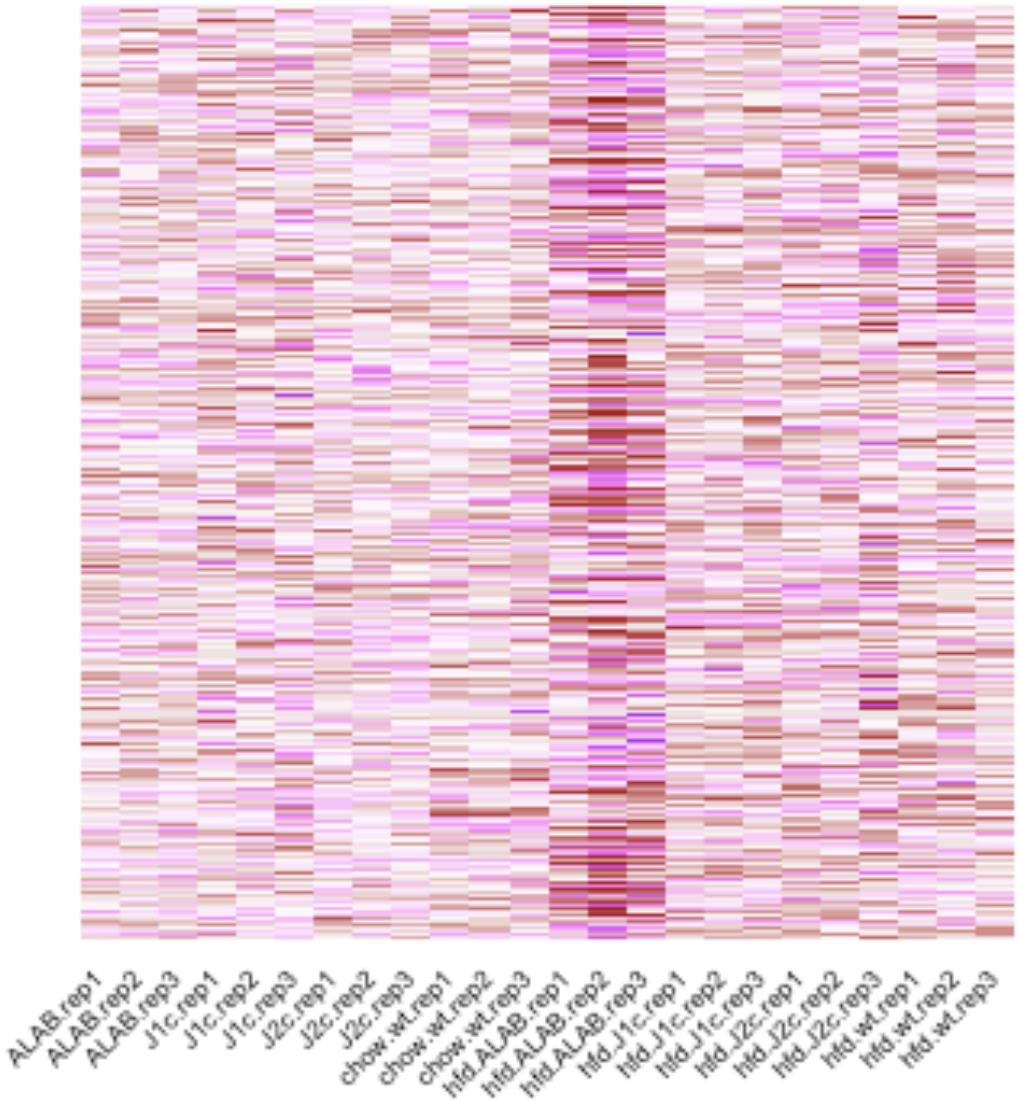
Task 3: How can I look at ALL my data at once?

- Sort by CV plot top variable genes



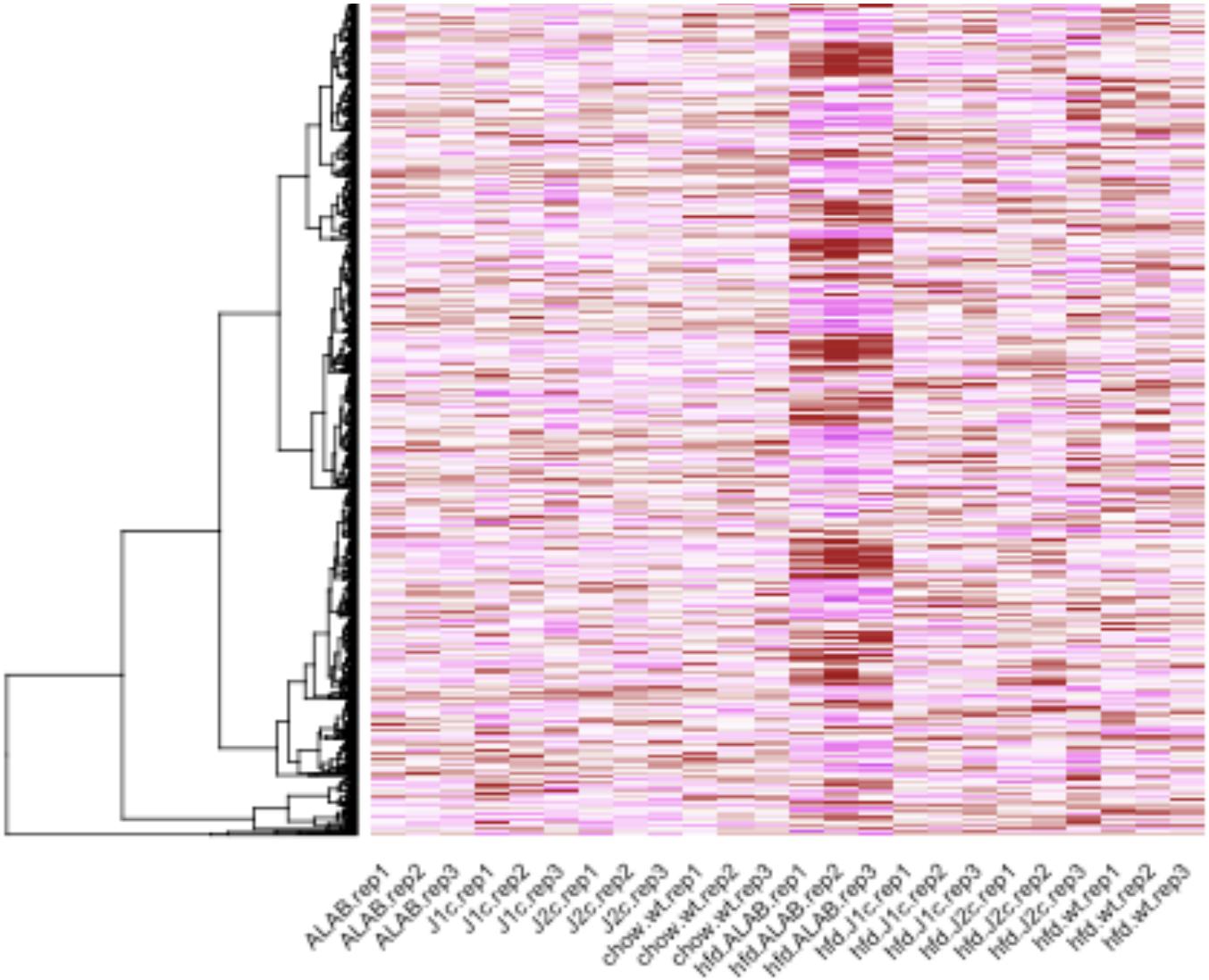
Task 3: How can I look at ALL my data at once?

- Sort by CV plot top variable genes, but also "scale" values



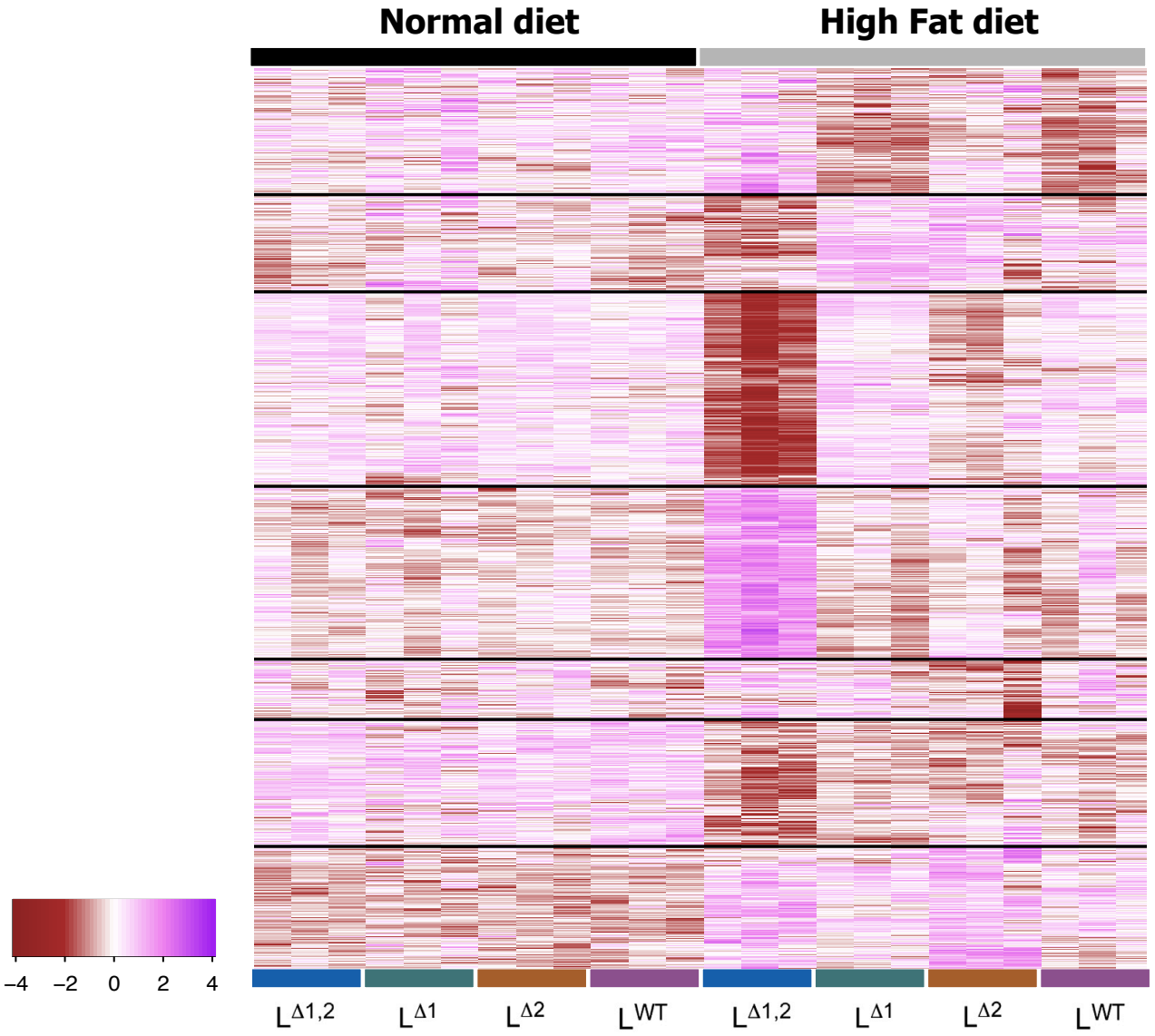
Task 3: How can I look at ALL my data at once?

- Sort by CV plot top variable genes, but also "scale" and "cluster" values



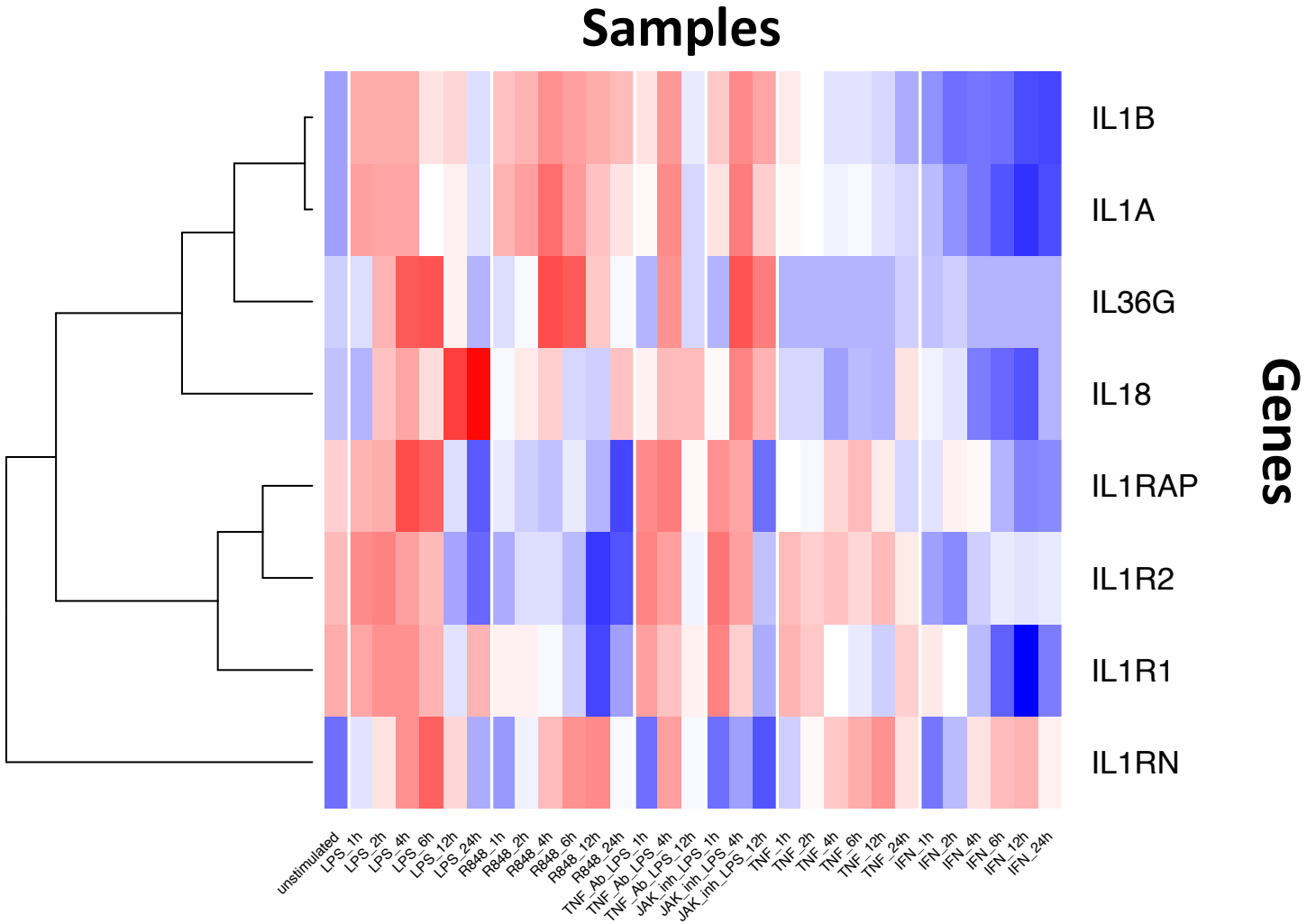
Task 3: How can I look at ALL my data at once?

- Sort by CV plot top variable genes, but also "log-scale" and "k-cluster" values



What is clustering?

Clustering – Similar patterns



Hierarchical clustering – when are vector similar?

Gene	Cond1	Cond2	Cond3	Cond4
g_1	2.5	5	7.5	10
g_2	0.1	0.5	0.8	1.1
g_3	0.2	0.3	0.4	11
g_4	2.5	8	8	9

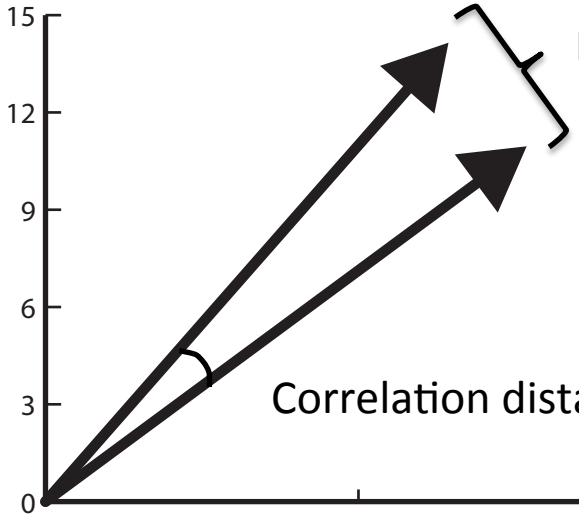
Clustering is about similarity:

- Between two rows (specified by a distance function)
- Between two sets of rows (specified by the linkage method)

What is a distance?

The goal of clustering is to group together samples that are “similar”.

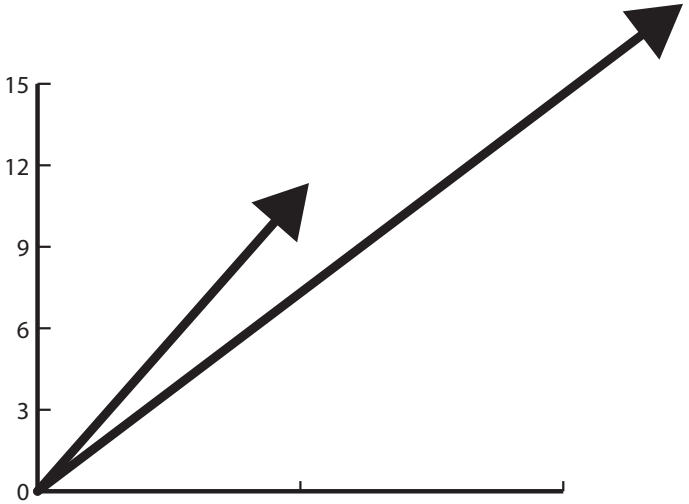
- When are two expression profiles “similar”?
- We consider each expression profile as a large “vector”. Each gene being a “dimension”

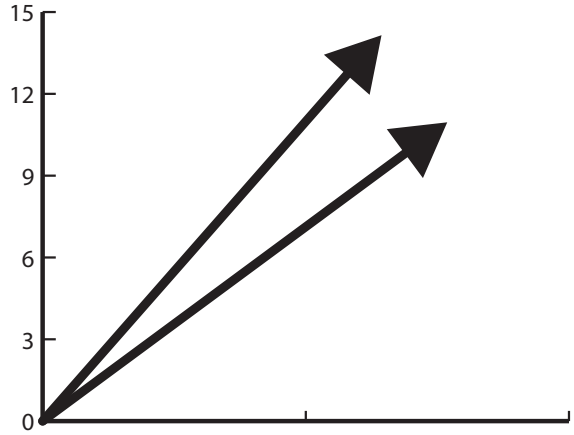


Euclidean distance $d(\vec{x},\vec{y}) = ||\vec{x}-\vec{y}|| = \sqrt{\sum_i (x_i - y_i)^2}$

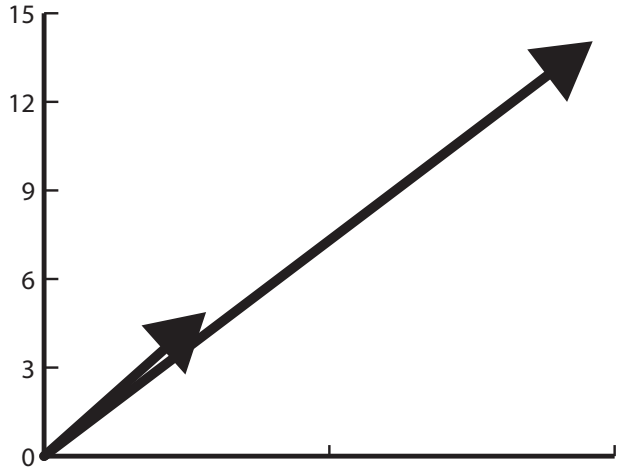
Correlation distance $d(\vec{x},\vec{y}) = \frac{1 - cor(\vec{x},\vec{y})}{2}$

What do difference distance care for?





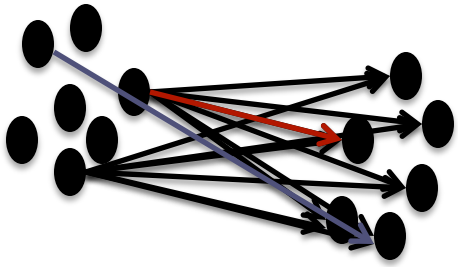
Similar correlation distance, very different euclidean distance



Correlation distance almost 0

Similarity between groups of points

Linkage: Distance between two sets (`d(R,S)`)



Single Linkage $min\{d(r,s), s \in S, r \in R\}$

Complete Linkage $max\{d(r,s), s \in S, r \in R\}$

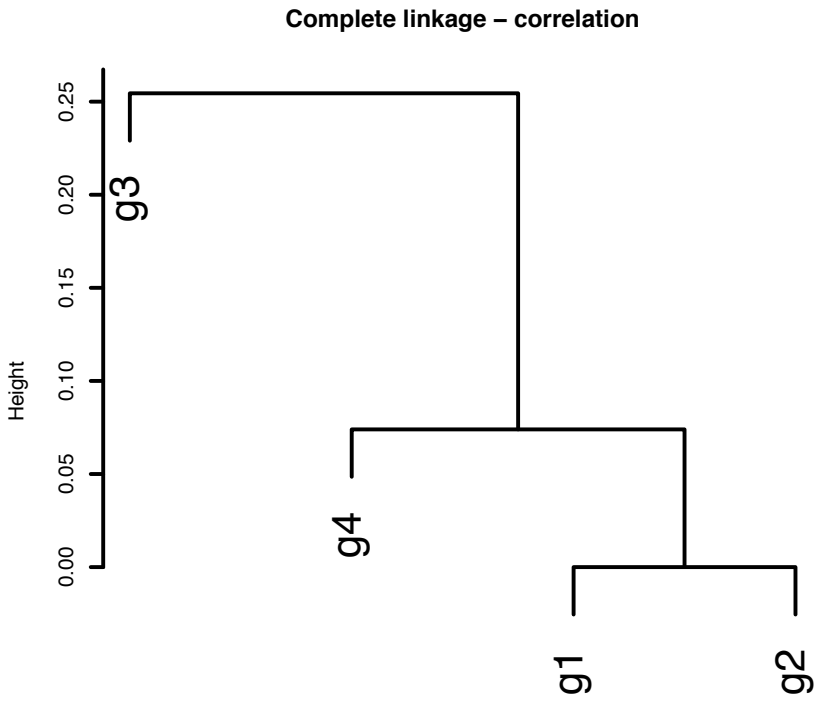
Average Linkage $mean\{d(r,s), s \in S, r \in R\}$

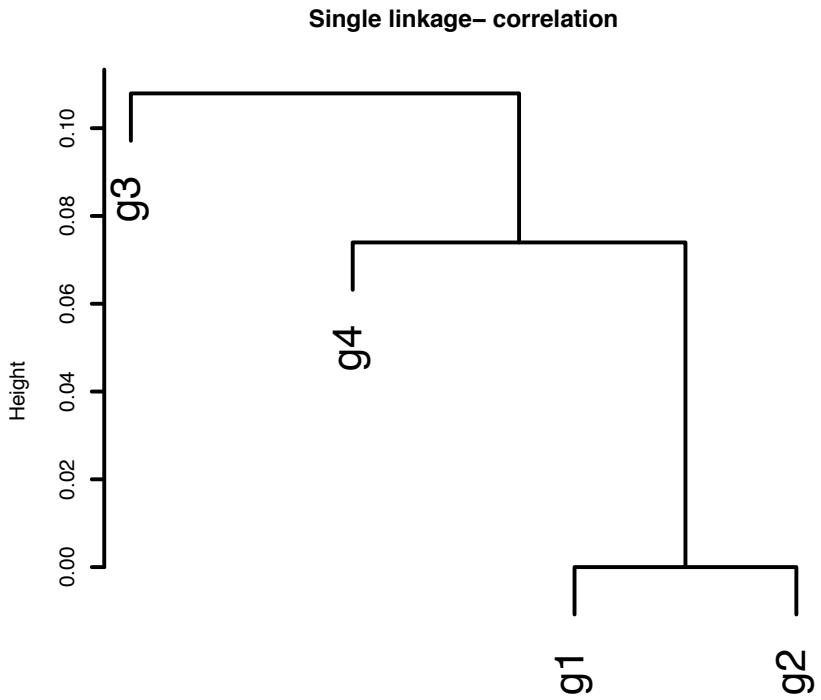
Similarity between groups of points

- Linkage: Distance between two sets (`d(R,S)`)
  - Complete: $max\{d(r,s), s \in S, r \in R\}$
  - Average: $mean\{d(r,s), s \in S, r \in R\}$
  - Single: $min\{d(r,s), s \in S, r \in R\}$
- Distance: Correlation, geometric (euclidean)

Gene	Cond1	Cond2	Cond3	Cond4
g_1	2.5	5	7.5	10
g_2	0.	0.5	0.8	1.1
g_3	0.2	0.3	0.4	11
g_4	2.5	8	8	9

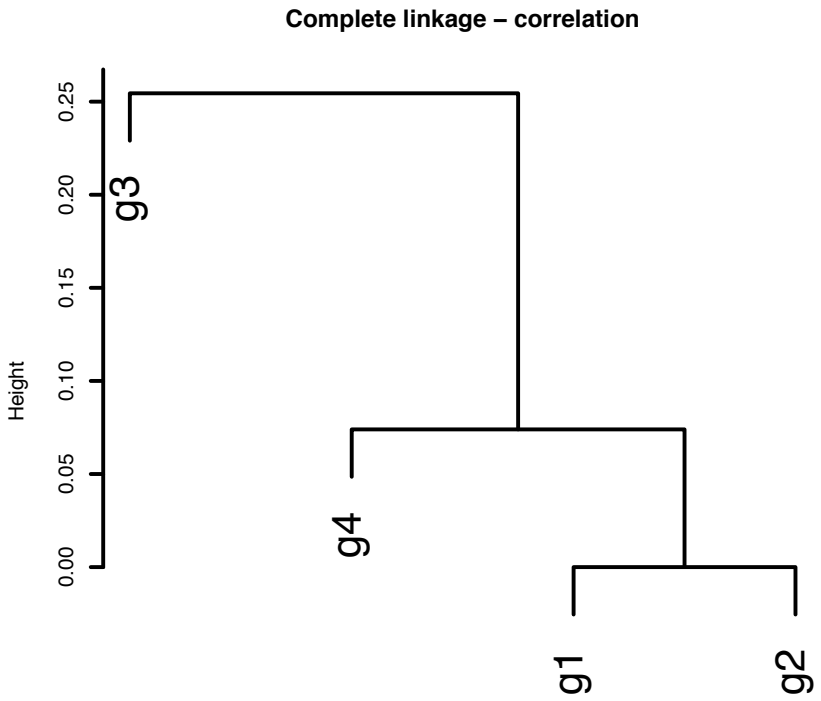
The effect of the linkage method

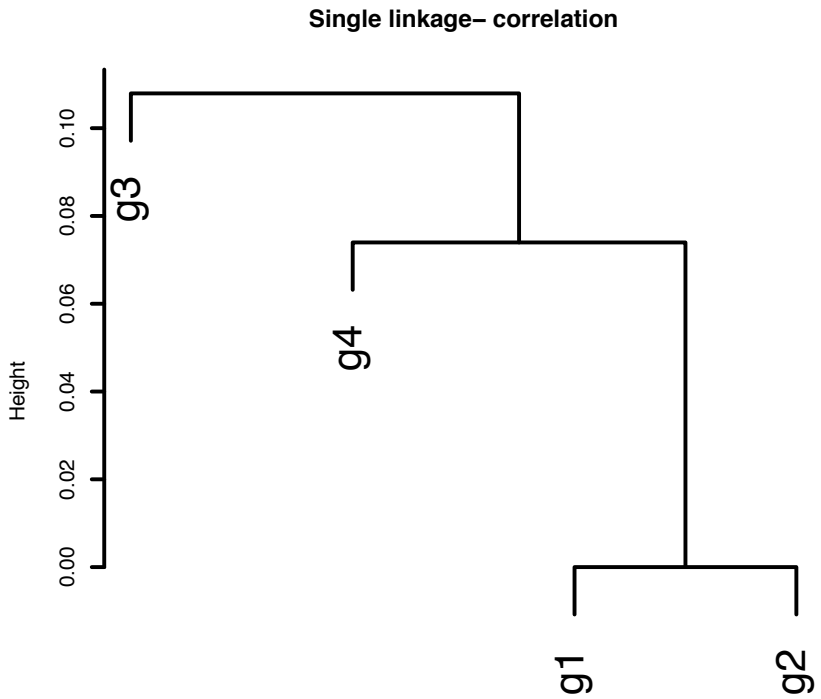




Gene	Cond1	Cond2	Cond3	Cond4
g_1	2.5	5	7.5	10
g_2	0.1	0.5	0.8	1.1
g_3	0.2	0.3	0.4	11
g_4	2.5	8	8	9

The effect of the linkage method

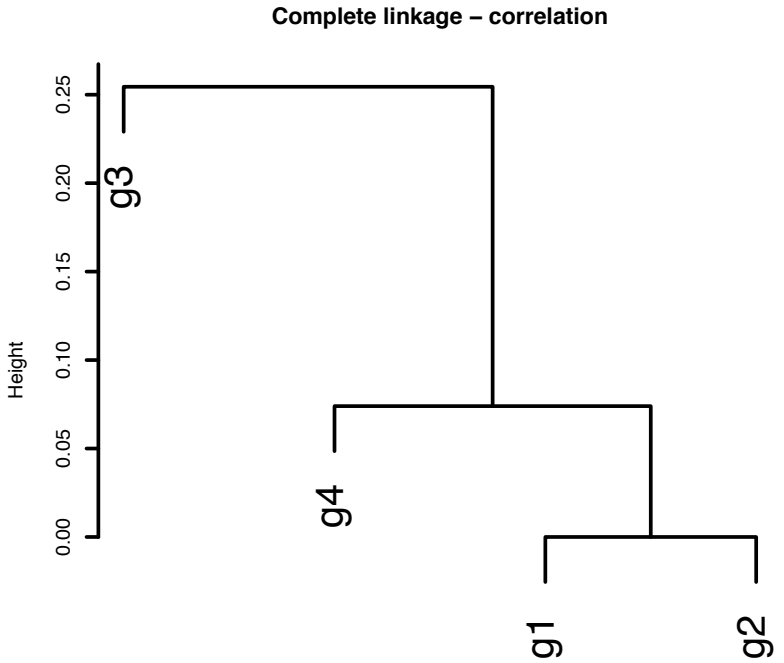


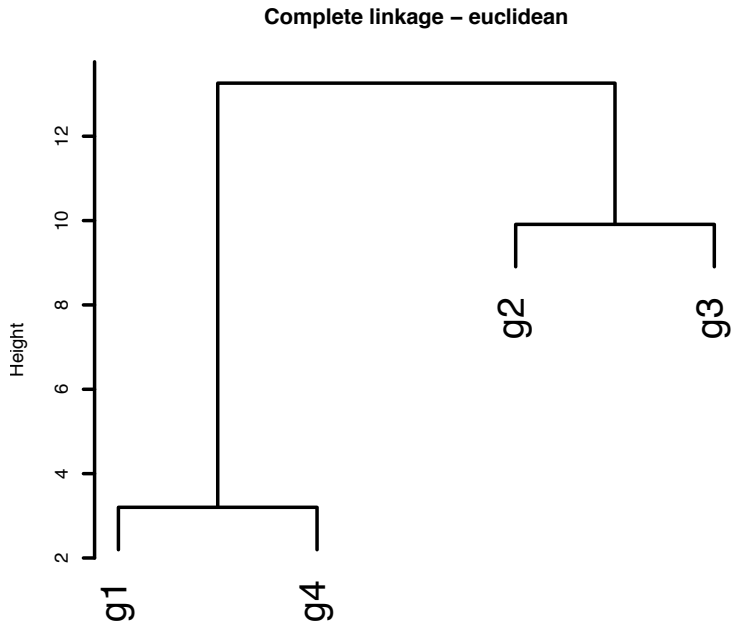


Correlation distance matrix

Column1	g1	g2	g3	g4
g1	0	0.00137174	0.10792118	0.07396763
g2	0.00137174	0	0.12430885	0.05598953
g3	0.10792118	0.12430885	0	0.25448339
g4	0.07396763	0.05598953	0.25448339	0

Effect of the distance!





Column	g1	g2	g3	g4
g1	0	0.0014	0.1079	0.0740
g2		0.0000	0.1243	0.0560
g3			0.0000	0.2545
g4				0

Correlation distance matrix

Column1	g1	g2	g3	g4
g1	0.00	12.25	8.88	3.20
g2		0.00	9.91	13.28
g3			0.00	11.24
g4				0.00

Geometric (Euclidean) distance matrix

Playing with clustering

```
#Define the toy matrix#
#######################
m =  rbind (c(2.5,5,7.5,10), c(0.1,0.5,0.8,1.1), c(0.2,0.3,0.4,11), c(2.5,8,8,9))

#Give column and row names#
###########################
rownames(m) = c("g1","g2","g3","g4");
colnames(m) = c("c1","c2","c3","c4");


#Compute the correlation distance matrix#
########################################
submat.dist = as.dist( (1 - cor(t(m)) )  /2 );


#Plot clustering with the three main methods#
#############################################
plot( hclust(submat.dist, method="complete",members=NULL), main="Complete linkeage -
correlation", sub="", xlab="", lwd=3);
plot( hclust(submat.dist, method="average",members=NULL),  main = "Average Linkeage -
correlation", sub="", xlab="", lwd=3);
plot( hclust(submat.dist, method="single",members=NULL),  main = "Single Linkeage- correlation",
sub="", xlab="", lwd=3);

#Plot clustering with the three main methods, using the euclidean distance#
############################################################################
plot( hclust(dist(m), method="complete",members=NULL), main="Complete linkeage - euclidean",
sub="", xlab="", lwd=3);
plot( hclust(dist(m), method="average",members=NULL),  main = "Average Linkeage -  euclidean",
sub="", xlab="", lwd=3);
plot( hclust(dist(m), method="single",members=NULL),  main = "Single Linkeage - euclidean",
sub="", xlab="", lwd=3);
```

Convinced of power – direct comparison

- Compare groups of samples to tease out genes that are (even subtly) different between samples. Ideally we should:
  - Incorporate inter group variability
  - Changes in mean
- Goal: Define a statistic that capture both group mean and variance.
  - Power: For a given gene, can use the statistic to reject the null hypothesis that the mean gene expression is the same for the two groups?
- Two approaches:
  - Parametric:
    - Find a distribution that fits gene expression values
    - Fit the distribution for each gene
    - Statistic: likelihood ratio (different means vs same mean). The t-test is a special case when we assume values are Gaussian distributed.
  - Non parametric:
    - Define a statistic i.e. a t-test
    - Permute samples to estimate values under randomness

Task 4: Finding the genes that differ between conditions

- Recall that intuitively gene expression follows a simple binomial model

$$P(n_g \mid \theta_g) = \binom{N}{n_g} \theta_g^{n_g} (1-\theta_g)^{N-n_g} \approx \frac{e^{-\theta_g N} (\theta_g N)^{n_g}}{n_g!}$$

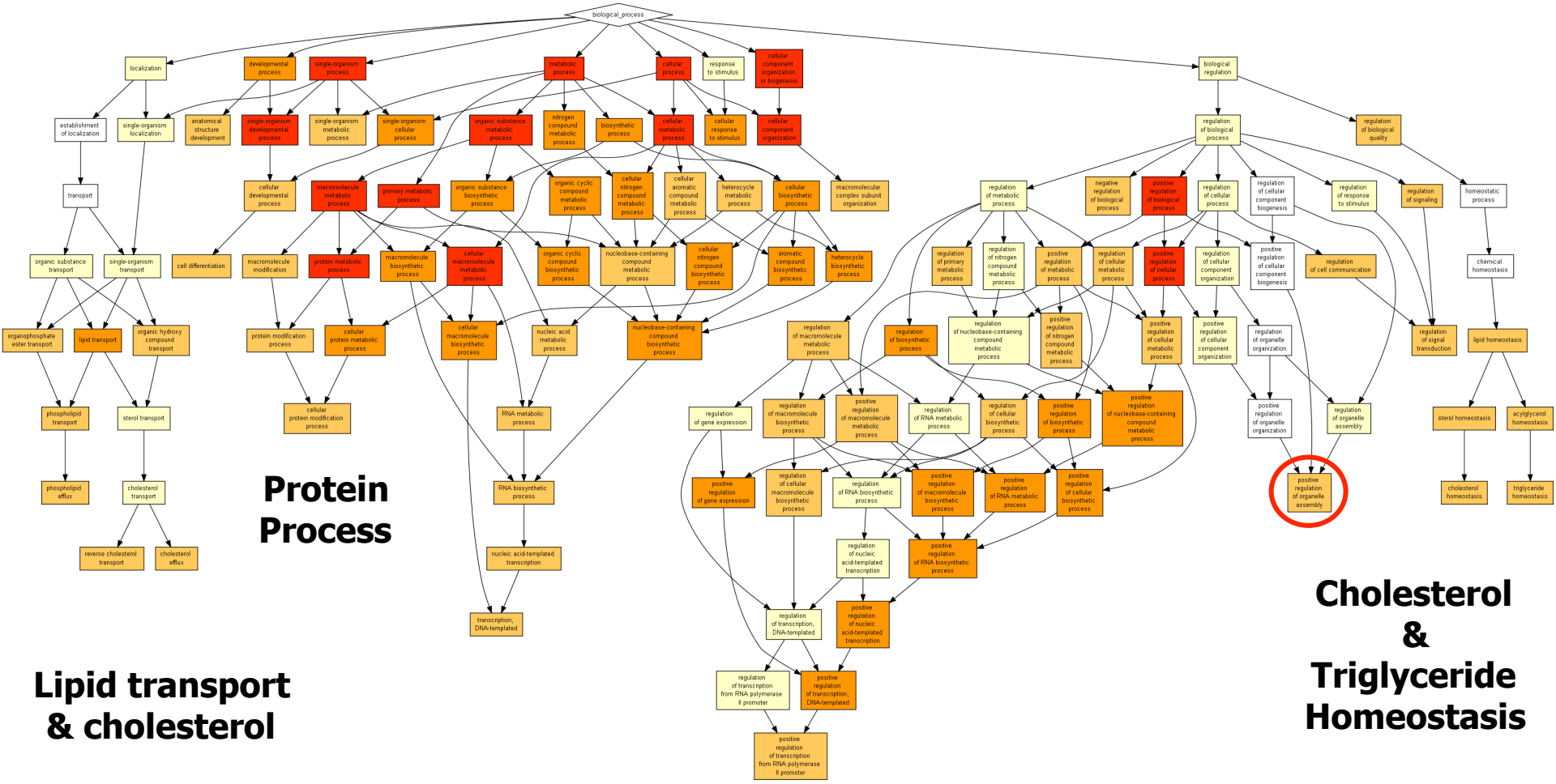
- In practice, the observed variance exceeds the mean, a sign that this is not the right model. A different distribution, the "negative binomial" explains better gene expression across replicates. The negative binomial is determined by its mean and variance.
- Current methods estimate the mean for each gene across replicates then estimate variance for genes with similar expression mean to increase the estimate robustness.
- Test: Likelihood ratio test fitting two different mean and variance for each group of replicates vs a model that uses the same mean and variance across all replicates.

Task 4: Finding the genes that differ between conditions

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
Gm10680	1706.444506	3.112741296	0.307812566	7.67590916	1.64249E-14	2.59131E-11
Gm15441	570.059034	2.51195109	0.305538932	5.766699111	8.08392E-09	2.55075E-06
BC044868	84.33739286	2.489946733	0.23813335	7.306606714	2.73973E-13	2.35766E-10
Gpnmb	159.988699	2.448488099	0.31026759	5.474268509	4.39323E-08	1.22313E-05
Cgref1	91.22187245	2.386996101	0.304450438	5.376888639	7.5784E-08	2.04963E-05
Ppp1r3g	59.45598745	2.353683977	0.322862231	4.967084475	6.7967E-07	0.000126152
Apoa4	27008.03585	2.177590962	0.303513773	4.703545893	2.55682E-06	0.000410217
Bcl2l14	27.57604462	2.090578545	0.322425406	4.157794392	3.21335E-05	0.002982115
Lamb3	384.2696995	2.033628639	0.175630137	7.308703732	2.69732E-13	2.35766E-10
D330041H03 Rik	32.45424455	1.96146267	0.30178968	4.014261424	5.96322E-05	0.005085393
Krt23	87.52892053	1.951166495	0.295386495	4.066423196	4.77402E-05	0.004145952
Igfbp2	13252.29789	1.928489158	0.238558789	4.940036634	7.81079E-07	0.000142186

Resulting in 528 genes at log2FoldChange > 0.75 and pvalue < 0.05

Task 6: What are these genes?



Protein Process

Lipid transport & cholesterol

Transcription

Cholesterol & Triglyceride Homeostasis

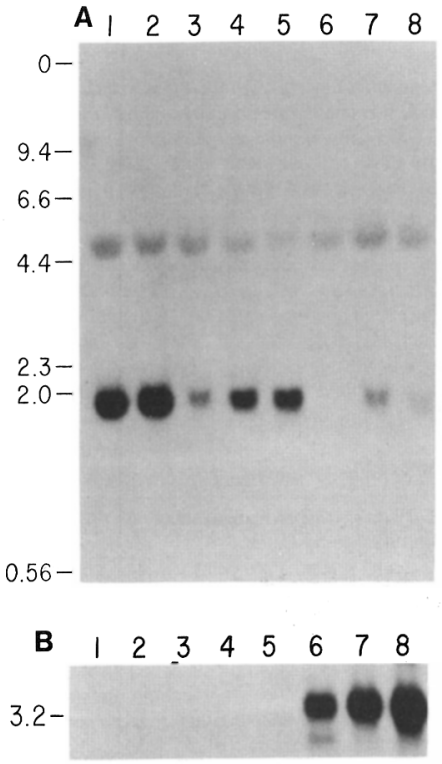
Cholesterol homeostasis:
5.98 (15170, 56, 453, 10)
$P < 5.31 \times 10^{-6}$

Why do we "adjust" the p-value?

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
Gm10680	1706.444506	3.112741296	0.307812566	7.67590916	1.64249E-14	2.59131E-11
Gm15441	570.059034	2.51195109	0.305538932	5.766699111	8.08392E-09	2.55075E-06
BC044868	84.33739286	2.489946733	0.23813335	7.306606714	2.73973E-13	2.35766E-10
Gpnmb	159.988699	2.448488099	0.31026759	5.474268509	4.39323E-08	1.22313E-05
Cgref1	91.22187245	2.386996101	0.304450438	5.376888639	7.5784E-08	2.04963E-05
Ppp1r3g	59.45598745	2.353683977	0.322862231	4.967084475	6.7967E-07	0.000126152
Apoa4	27008.03585	2.177590962	0.303513773	4.703545893	2.55682E-06	0.000410217
Bcl2l14	27.57604462	2.090578545	0.322425406	4.157794392	3.21335E-05	0.002982115
Lamb3	384.2696995	2.033628639	0.175630137	7.308703732	2.69732E-13	2.35766E-10
D330041H03 Rik	32.45424455	1.96146267	0.30178968	4.014261424	5.96322E-05	0.005085393
Krt23	87.52892053	1.951166495	0.295386495	4.066423196	4.77402E-05	0.004145952
Igfbp2	13252.29789	1.928489158	0.238558789	4.940036634	7.81079E-07	0.000142186

Where are we?

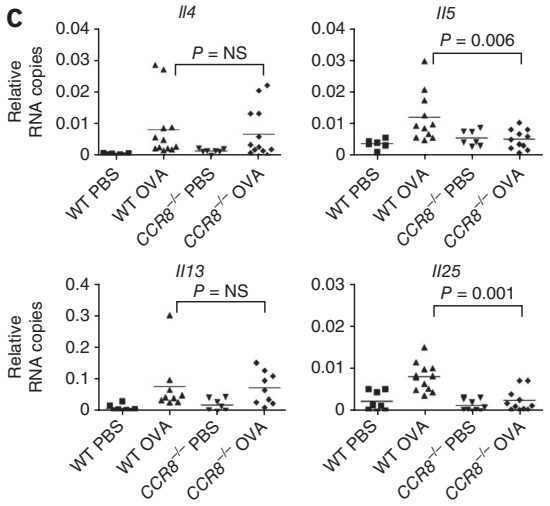
80s



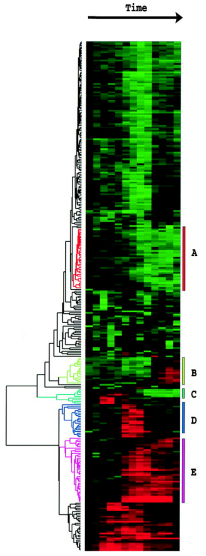
Cell, Vol. 47, 667–674, December 5, 1986, Copyright © 1986 by Cell Press

Handful of datapoints

90s



Islan et al Nat. Imm. 2011



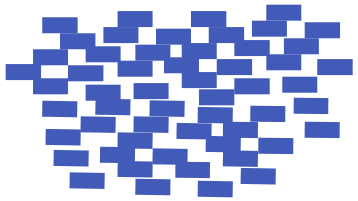
Thousands of datapoints

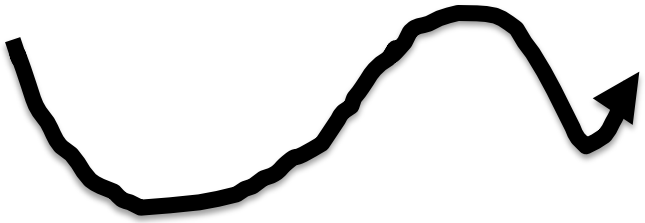
Michael B. Eisen et al. PNAS 1998

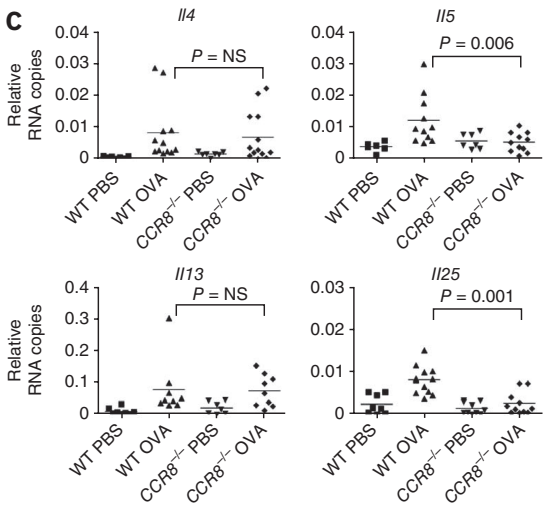
Biology slowly becoming a “big data” science

2010s

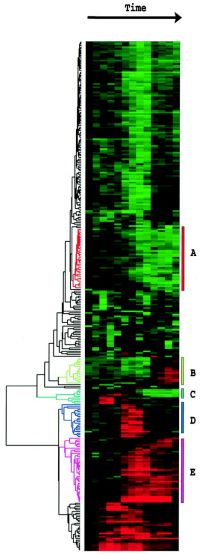
Sequenced reads







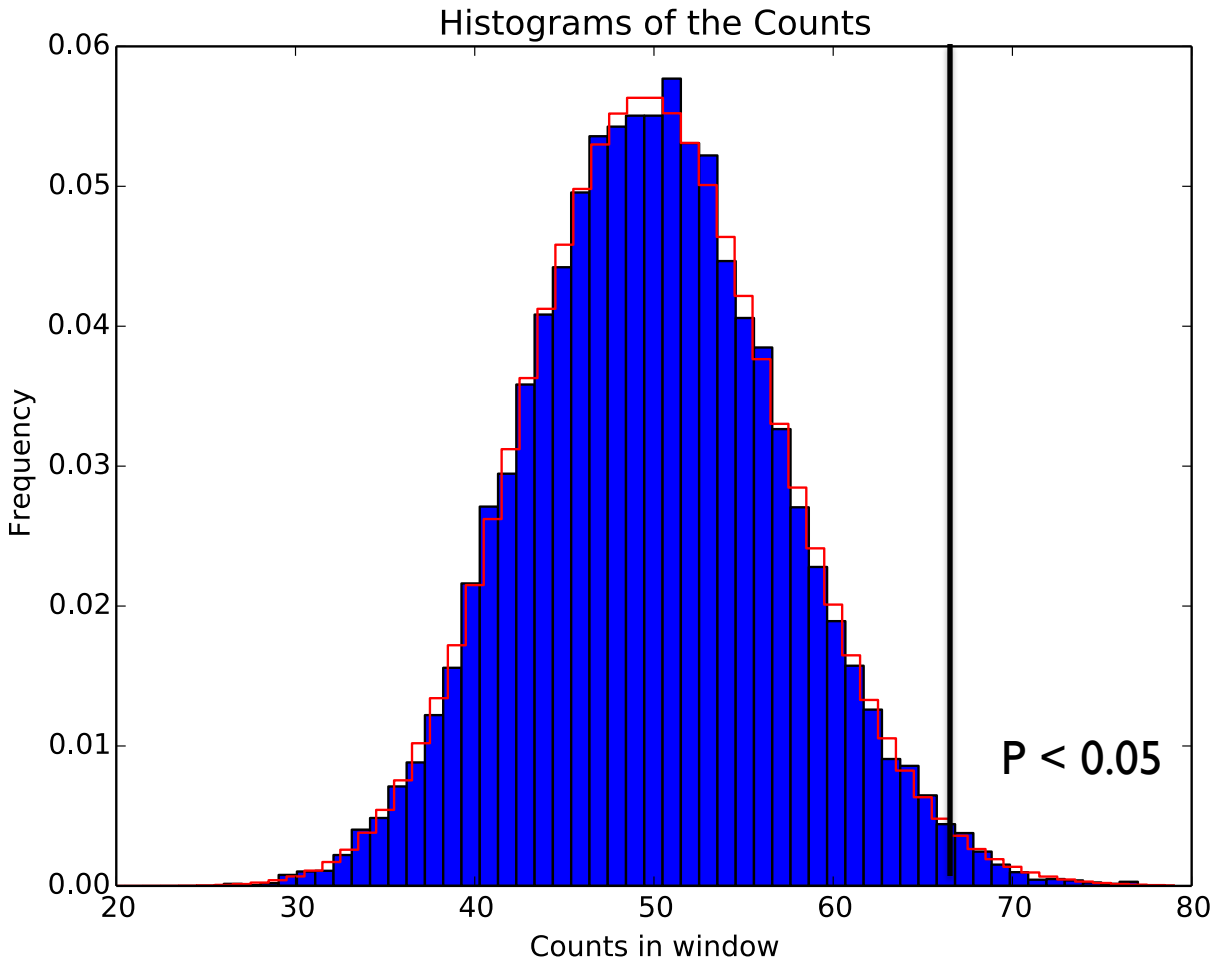
Millions-billions datapoints



Statistical methods are deeply embedded – two concepts

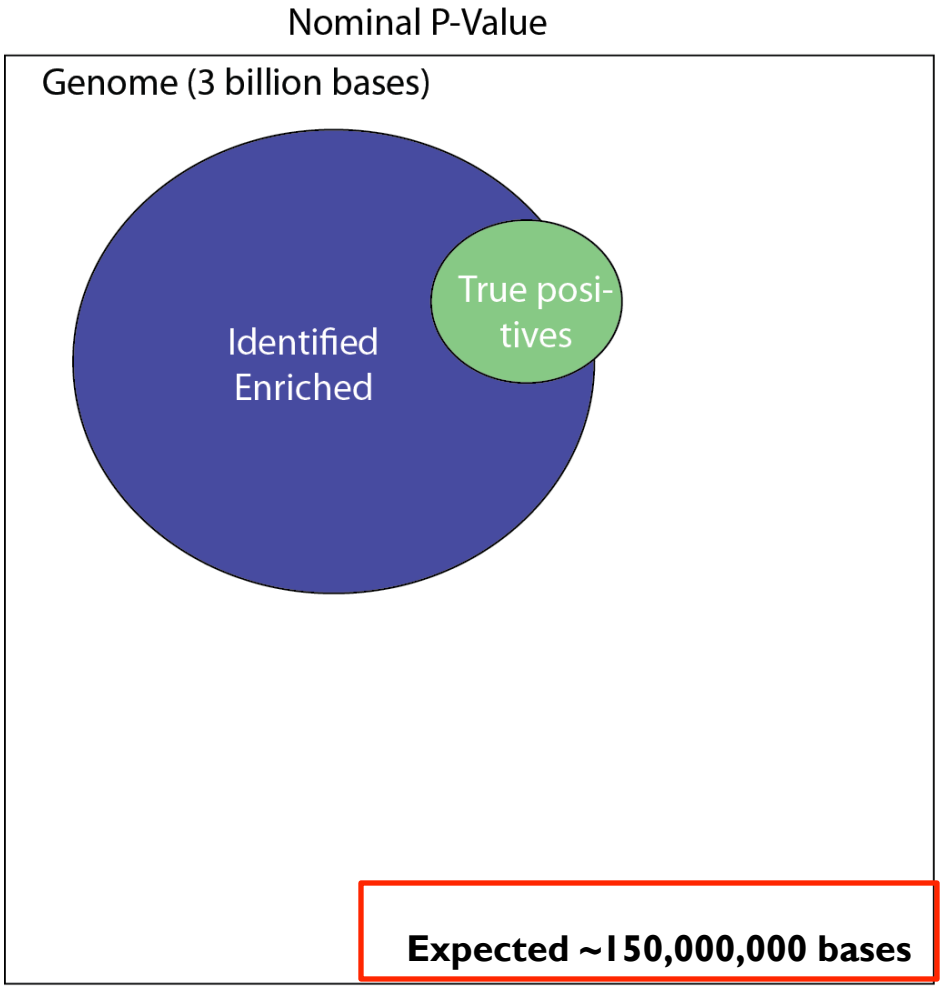
Multiple testing problems
Modeling count data

We can’t use a nominal p-value any longer



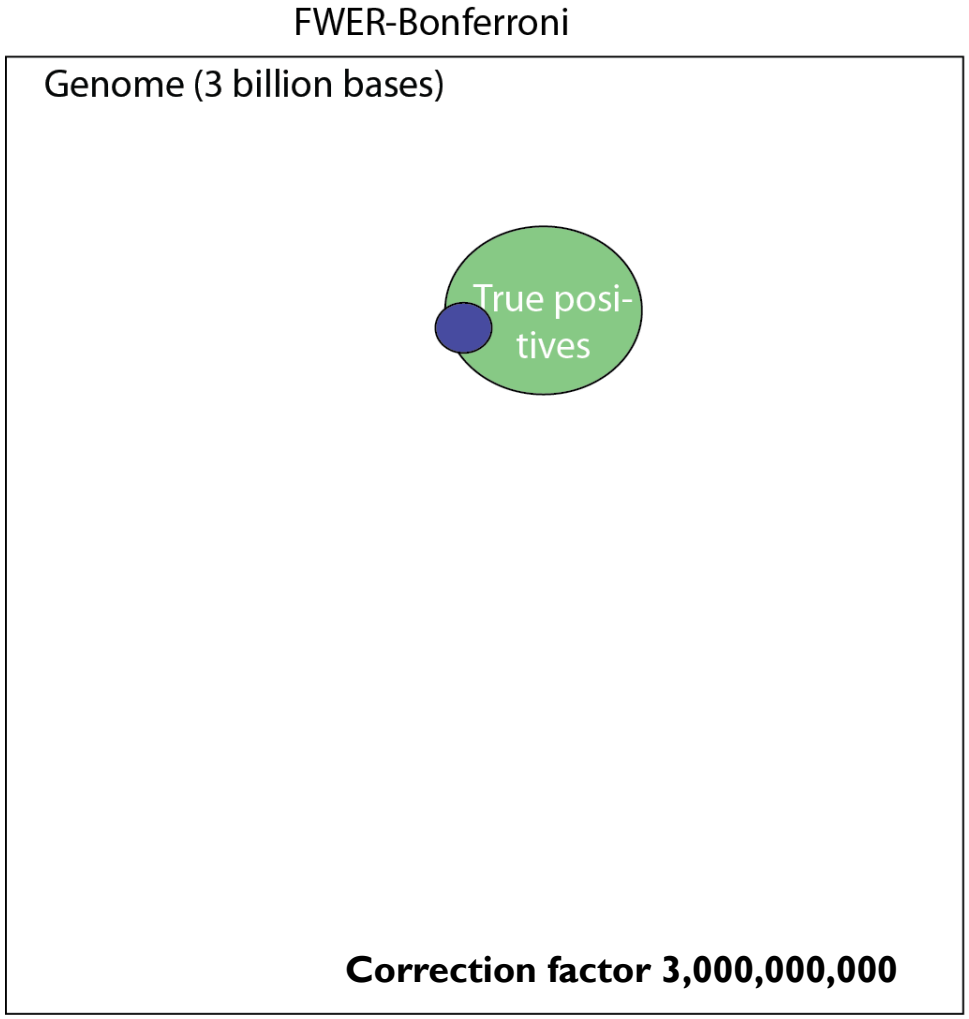
All will be noise!

The genome is large, many things happen by chance



We need to correct for multiple hypothesis testing

The Bonferroni correction



Bonferroni corrects the number of hits but misses many true hits because its too conservative – How do we get more power?

Can we correct for multiple testing a bit more subtly?

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 289–300

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel

[Received January 1993. Revised March 1994]

Downloadable from: http://garberlab.umassmed.edu/bootcamp.2015/BH.pdf