Differential Gene Expression II – quantifying differences

Biol4230 Tues, April 3, 2018

Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

- When is a difference significant I
  - modest numbers of counts: Fisher's Exact Test
  - means and standard deviations: Student's t-test
- The signal and the noise - normalization
- When are differences significant II
  - multiple test correction: Bonferroni
  - False discovery rates (FDR, q-value)

To learn more:

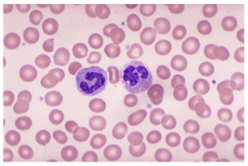
1. Pevsner, Chapter 8 pp. 331-373
2. Draghici, Soren (2012) "Statistics and data analysis for microarrays using R and Bioconductor" Chapman and Hall
3. http://bioinformatics.ucdavis.edu/docs/2015-march-workshop/_downloads/Thursday_BDJ_stats.pdf

Differential Gene Expression

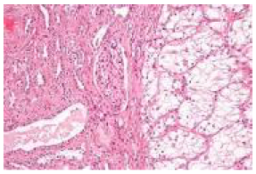
- Large quantity of data (>20,000 genes)
  - Affychip data has ?20 replicates per gene
  - RNAseq has counts (FPKM: Fragments per Kilobase per Million mapped reads)
  - but a small number of biological replicates
- Ideally, identify modest change (1.5x or larger) for modest levels of transcription
  - 10 or fewer transcripts may account for 90% of reads, so 5,000 transcripts for < 10% of reads
  - If technical replicates vary more than 2x, how do you measure 1.5x change?
- Large numbers of tests: how to correct?
  - Family-wide-error-rate (FWER) Bonferroni correction (used for similarity search E()-values)
  - False-discovery-rate (FDR, qvalue)

Cells in different tissues are different

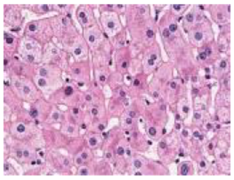
red blood cell



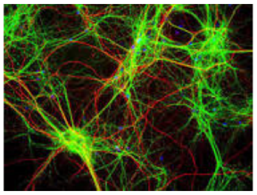
kidney



liver

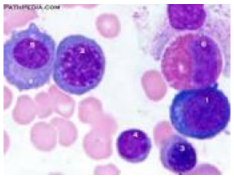


brain



muscle

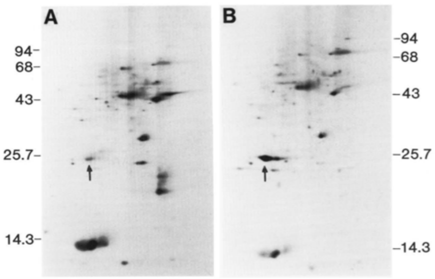
lymphocyte



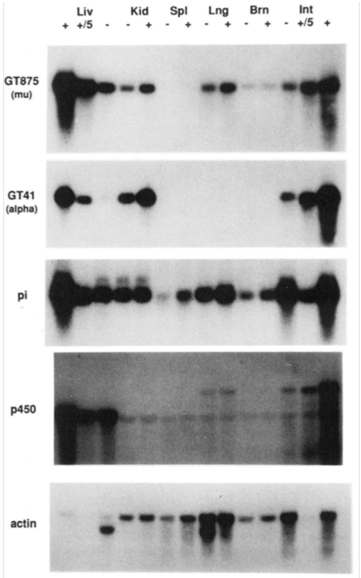
because they express different proteins from different mRNAs

induction of detoxification gene mRNAs

Liver Proteins



Pearson, W. R. et al *J Biol Chem* **258,** 2052–2062 (1983).



Pearson, W. R. et al. *J Biol Chem* **263,** 13324–13332 (1988).

Microarrays vs RNAseq

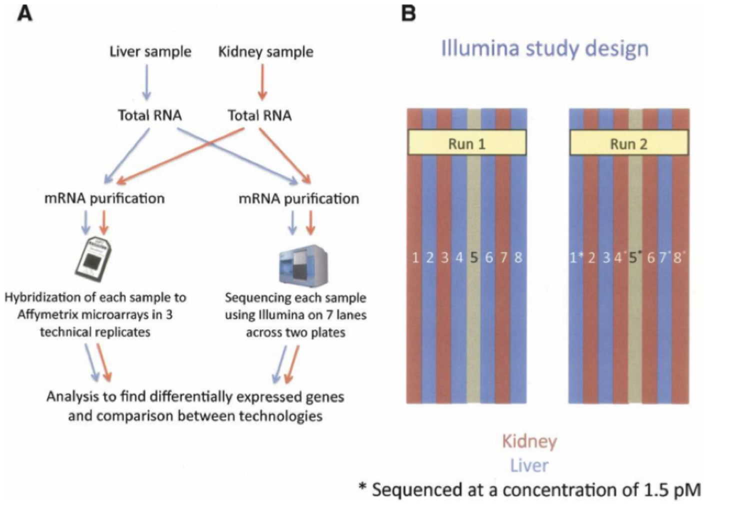

A
Liver sample
Kidney sample
Total RNA
Total RNA
mRNA purification
mRNA purification
Hybridization of each sample to Affymetrix microarrays in 3 technical replicates
Sequencing each sample using Illumina on 7 lanes across two plates
Analysis to find differentially expressed genes and comparison between technologies
B
Illumina study design
Run 1
Run 2
1 2 3 4 5 6 7 8
1* 2 3 4 5* 6 7* 8*
Kidney
Liver
* Sequenced at a concentration of 1.5 pM

Figure 1. Graphical representation of the study design. (*A*) Summary of the experimental design. (*B*) The lanes in which each sample was sequenced across the two runs. In each run, the control sample was sequenced in lane *5*. Samples were sequenced at two concentrations: 1.5 pM (indicated by an asterisk) and 3 pM (no asterisk).

Marioni et al. Genome Res. **18,** 1509–1517 (2008).

Microarrays vs RNAseq

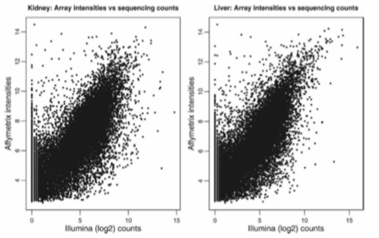


Figure 3. Comparing counts from Illumina sequencing with normalized intensities from the array, for kidney (*left*) and liver (*right*). In each panel, the average ($\log_2$) counts for each gene are plotted on the *X*-axis, and the corresponding normalized intensities from the array are shown on the *Y*-axis. To avoid taking the log of 0, we added 1 to each of the average counts prior to taking logs.

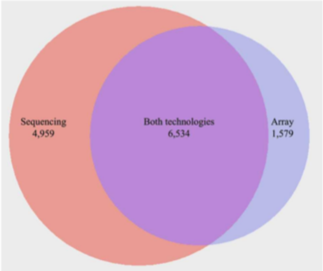


Figure 5. A Venn diagram summarizing the overlap between genes called as differentially expressed from the (*left* circle) sequence data and from the (*right* circle) array. The number of genes called by both technologies is indicated by the overlap between the two circles.

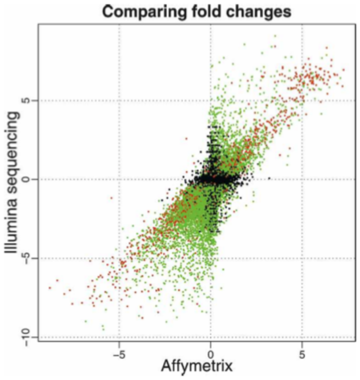


Figure 4. Comparison of estimated $\log_2$ fold changes (liver/kidney) from Illumina (*Y*-axis) and Affymetrix (*X*-axis). We consider only genes that were interrogated using both platforms and genes where the mean number of counts across lanes was greater than 0 for both the liver and kidney samples. (Red and green dots) Genes called as differentially expressed based on the Illumina sequencing data at an FDR of 0.1%, with a mean number of counts greater than (red) or less than (green) 250 reads in both tissues. (Black dots) Genes not called as differentially expressed based on the Illumina sequencing data. The set of differentially expressed genes that show the strongest correlation between the two technologies seems to be those that are mapped to by many reads (red), while the correlation is weaker for differentially expressed genes mapped to by fewer reads (green).

Marioni et al. Genome Res. **18,** 1509–1517 (2008).

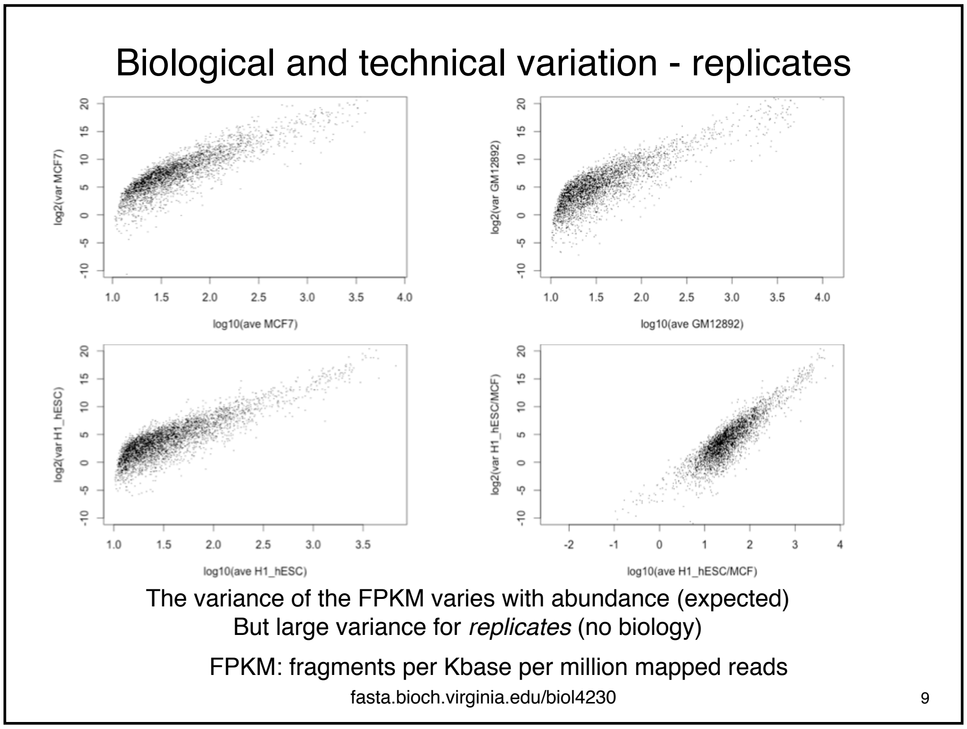
Measuring differences – sources of variation

Technical

- RNA isolation
- cDNA synthesis
- hybridization (AffyChip)
- PCR amplification
- G+C content
- sequencing depth
- location on AffyChip/ sequencing "lane"

Biological

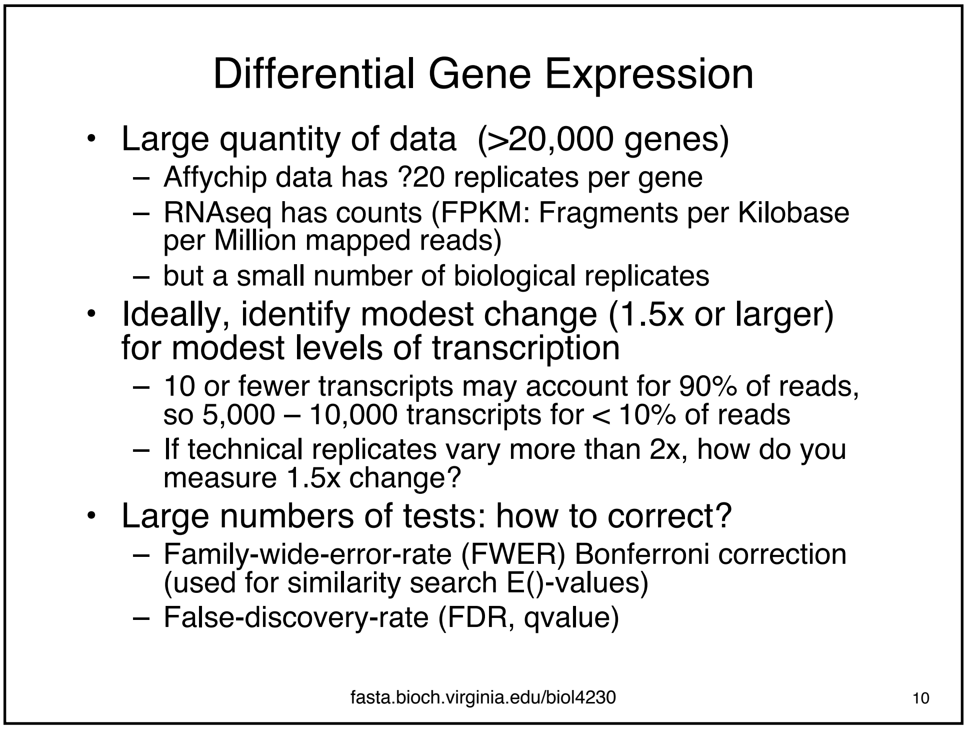
- genetic background
- sex
- last meal/sleep/exercise
- dividing/quiescent
- cell type within tissue type
- …



Biological and technical variation - replicates

The variance of the FPKM varies with abundance (expected)
But large variance for *replicates* (no biology)

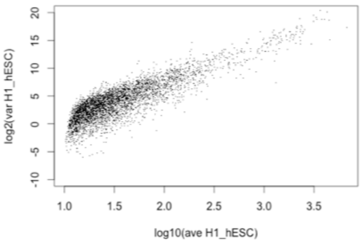
FPKM: fragments per Kbase per million mapped reads

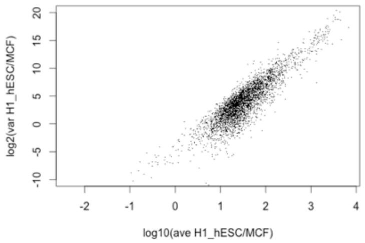


Differential Gene Expression

- Large quantity of data (>20,000 genes)
  - Affychip data has ?20 replicates per gene
  - RNAseq has counts (FPKM: Fragments per Kilobase per Million mapped reads)
  - but a small number of biological replicates
- Ideally, identify modest change (1.5x or larger) for modest levels of transcription
  - 10 or fewer transcripts may account for 90% of reads, so 5,000 – 10,000 transcripts for < 10% of reads
  - If technical replicates vary more than 2x, how do you measure 1.5x change?
- Large numbers of tests: how to correct?
  - Family-wide-error-rate (FWER) Bonferroni correction (used for similarity search E()-values)
  - False-discovery-rate (FDR, qvalue)

Biological and technical variation - replicates

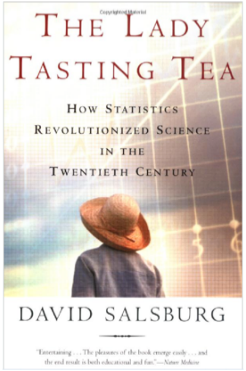




The variance of the FPKM varies with abundance (expected)
But large variance for *replicates* (no biology)

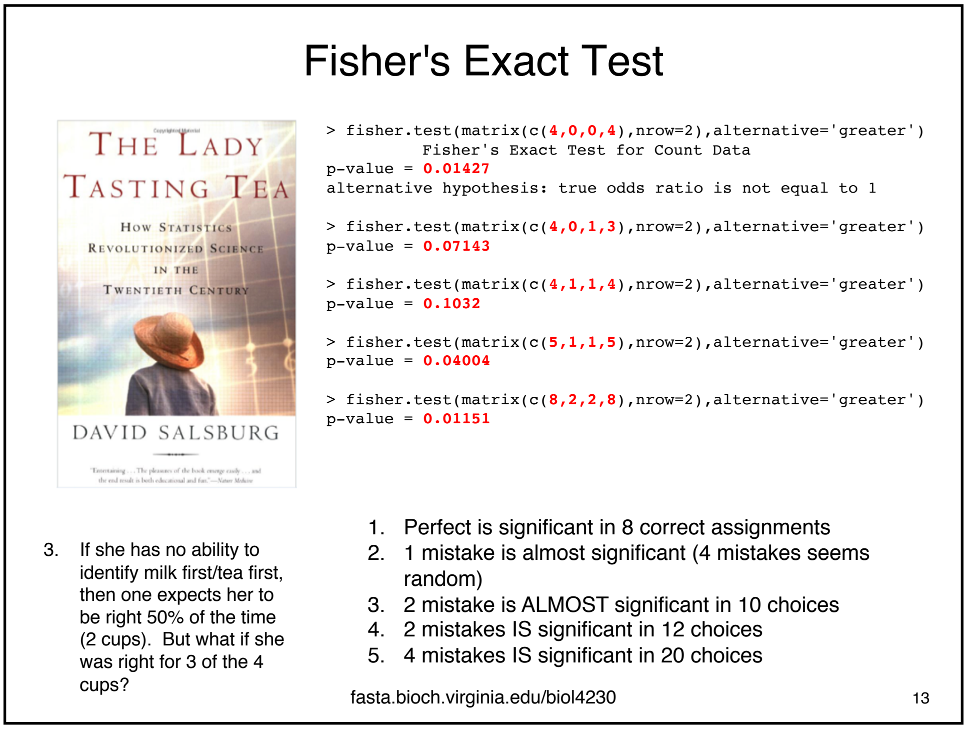
Goal: to identify differential expression
Separate between sample differences
from within sample differences

The significance of differences: Fisher's Exact Test


THE LADY TASTING TEA
HOW STATISTICS REVOLUTIONIZED SCIENCE IN THE TWENTIETH CENTURY
DAVID SALSBURG

1. Around 1930, Muriel Bristol claimed, in a conversation with R. A. Fisher, that she could tell when milk was poured into tea, which was much preferable to tea being poured into milk.
2. Fisher choose to test this hypothesis by preparing 8 cups of tea, 4 tea first, 4 milk first, and asking Ms. Bristol to identify the 4 cups with tea first.
3. If she has no ability to identify milk first/tea first, then one expects her to be right 50% of the time (4 cups). But what if she was right for 6 of the 8 cups?

```
> fisher.test(matrix(c(4,0,0,4),nrow=2),
+             alternative='greater')
        Fisher's Exact Test for Count Data
data:  matrix(c(4, 0, 0, 4), nrow = 2)
p-value = 0.01427
alternative hypothesis: true odds ratio is not equal to 1
```



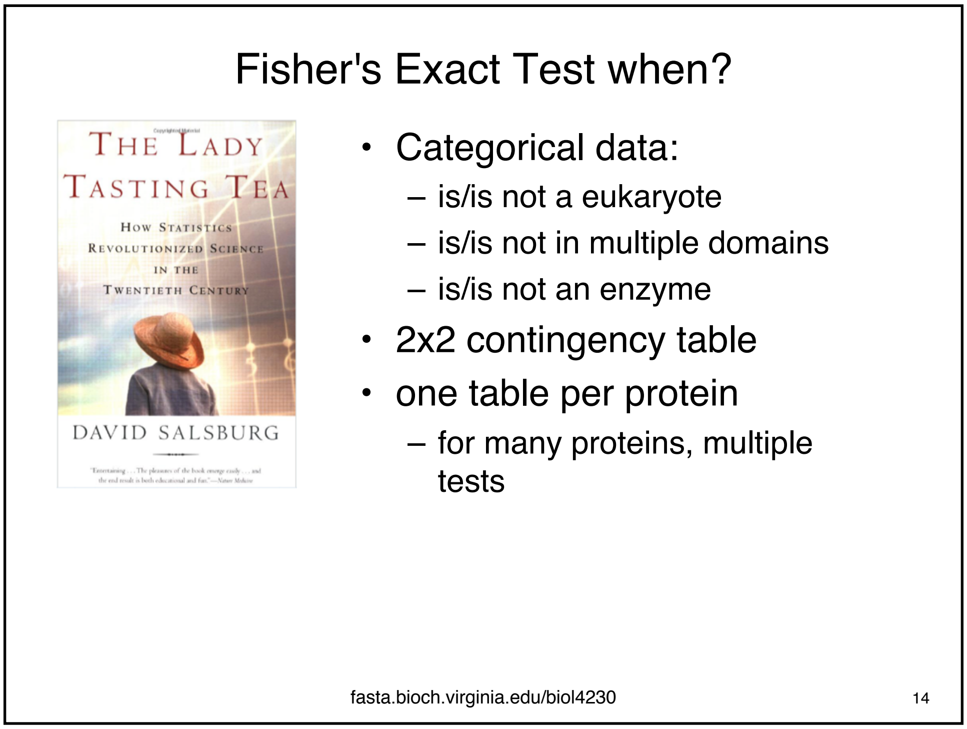
Fisher's Exact Test

THE LADY TASTING TEA
How Statistics Revolutionized Science in the Twentieth Century
DAVID SALSBURG

```
> fisher.test(matrix(c(4,0,0,4),nrow=2),alternative='greater')
          Fisher's Exact Test for Count Data
p-value = 0.01427
alternative hypothesis: true odds ratio is not equal to 1

> fisher.test(matrix(c(4,0,1,3),nrow=2),alternative='greater')
p-value = 0.07143

> fisher.test(matrix(c(4,1,1,4),nrow=2),alternative='greater')
p-value = 0.1032

> fisher.test(matrix(c(5,1,1,5),nrow=2),alternative='greater')
p-value = 0.04004

> fisher.test(matrix(c(8,2,2,8),nrow=2),alternative='greater')
p-value = 0.01151
```

3. If she has no ability to identify milk first/tea first, then one expects her to be right 50% of the time (2 cups). But what if she was right for 3 of the 4 cups?

1. Perfect is significant in 8 correct assignments
2. 1 mistake is almost significant (4 mistakes seems random)
3. 2 mistake is ALMOST significant in 10 choices
4. 2 mistakes IS significant in 12 choices
5. 4 mistakes IS significant in 20 choices



Fisher's Exact Test when?

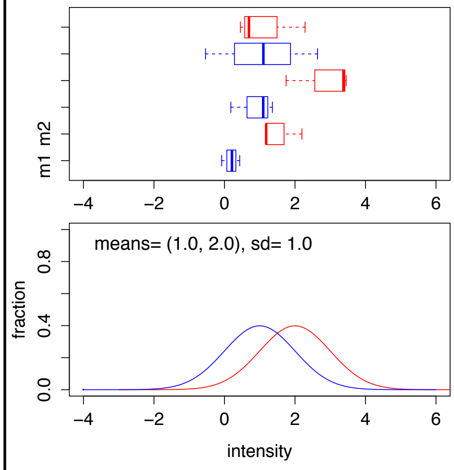
THE LADY TASTING TEA
How Statistics Revolutionized Science in the Twentieth Century
DAVID SALSBURG

- Categorical data:
  - is/is not a eukaryote
  - is/is not in multiple domains
  - is/is not an enzyme
- 2x2 contingency table
- one table per protein
  - for many proteins, multiple tests

Differential gene expression

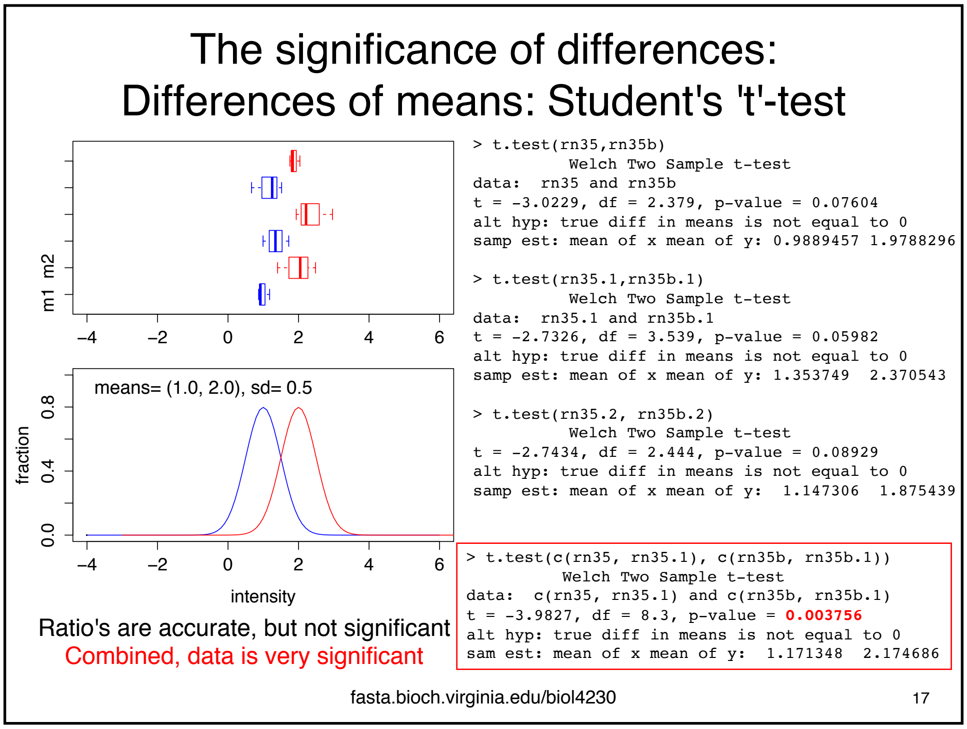
- mRNA levels affect protein levels
  - no mRNA, no protein
  - little mRNA, sometimes lots of protein (long half-life)
  - lots of mRNA, often lots of protein
- RNA abundance:
  - most RNA is ribosomal RNA (rRNA)
  - 10 – 50 mRNA species account for >90% of mRNA abundance
  - sensitive methods detect < 1 molecule/cell (but not with single cells)
- which changes matter?
  - fold differences
    - 100X, from 1:100 molecules/cell?
    - 5X, from 50,000 to 250,000 molecules/cell?
  - mostly high abundance? mostly low abundance?

The significance of differences: Differences of means: Student's 't'-test



```
data:  rn3 and rn3b
t = -4.6426, df = 2.283, p-value = 0.0335
alt hyp: true diff in means is not equal to 0
sample est: mean(x) mean(y)   0.1886128 2.8588774

data:  rn3.1 and rn3b.1
t = 0.4594, df = 2.536, p-value = 0.6824
alt hyp: true diff in means is not equal to 0
sample est: mean(x) mean(y)   1.518745  1.069586

data:  rn3.2 and rn3b.2
t = -0.3909, df = 3.342, p-value = 0.7195
alt hyp: true diff in means is not equal to 0
sample est: mean(x) mean(y)  0.8793091 1.1442473
```

Ratio's are accurate, one significant



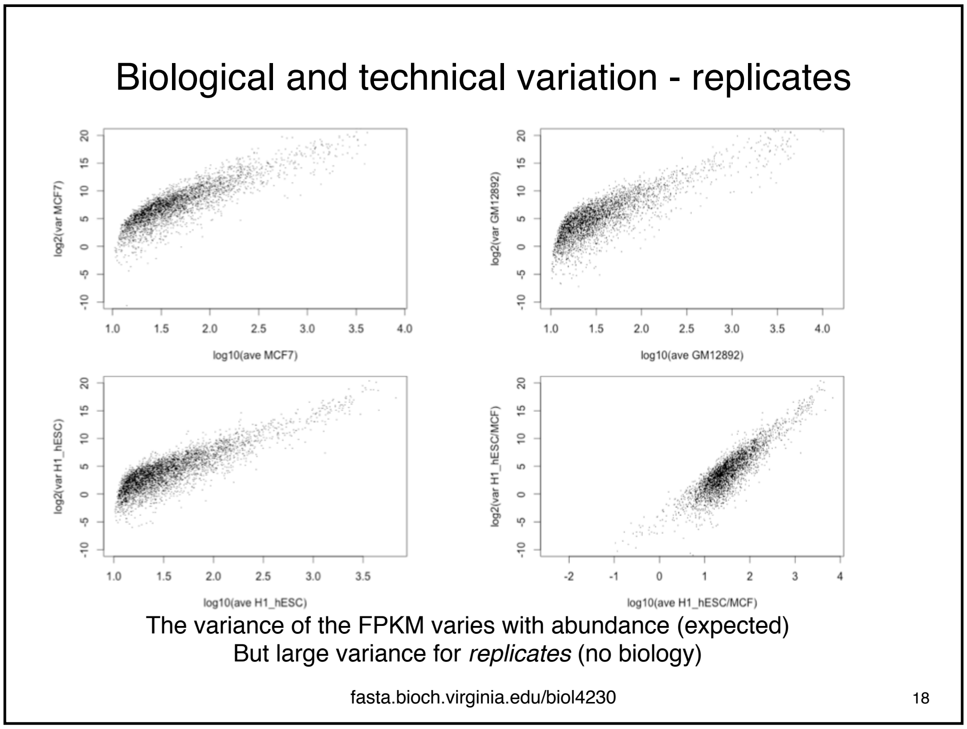
The significance of differences: Differences of means: Student's 't'-test

```
> t.test(rn35,rn35b)
          Welch Two Sample t-test
data:  rn35 and rn35b
t = -3.0229, df = 2.379, p-value = 0.07604
alt hyp: true diff in means is not equal to 0
samp est: mean of x mean of y: 0.9889457 1.9788296

> t.test(rn35.1,rn35b.1)
          Welch Two Sample t-test
data:  rn35.1 and rn35b.1
t = -2.7326, df = 3.539, p-value = 0.05982
alt hyp: true diff in means is not equal to 0
samp est: mean of x mean of y: 1.353749  2.370543

> t.test(rn35.2, rn35b.2)
          Welch Two Sample t-test
t = -2.7434, df = 2.444, p-value = 0.08929
alt hyp: true diff in means is not equal to 0
samp est: mean of x mean of y:  1.147306  1.875439
```

```
> t.test(c(rn35, rn35.1), c(rn35b, rn35b.1))
          Welch Two Sample t-test
data:  c(rn35, rn35.1) and c(rn35b, rn35b.1)
t = -3.9827, df = 8.3, p-value = 0.003756
alt hyp: true diff in means is not equal to 0
sam est: mean of x mean of y:  1.171348  2.174686
```

Ratio's are accurate, but not significant
Combined, data is very significant

fasta.bioch.virginia.edu/biol4230 17

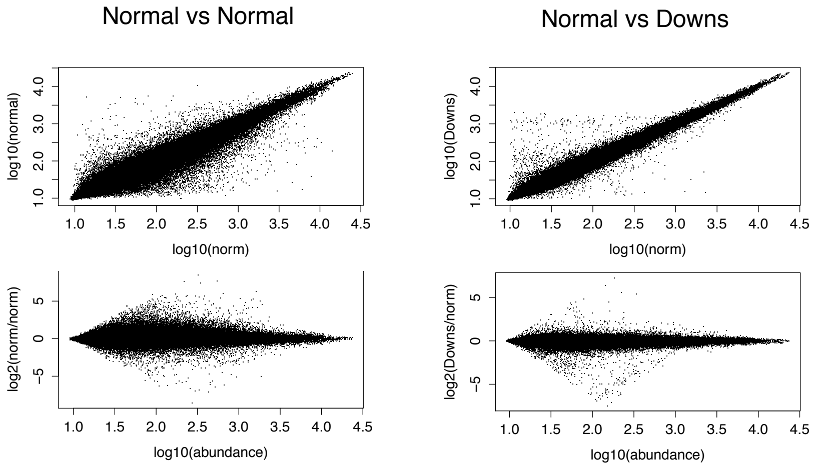


Biological and technical variation - replicates

The variance of the FPKM varies with abundance (expected)
But large variance for *replicates* (no biology)

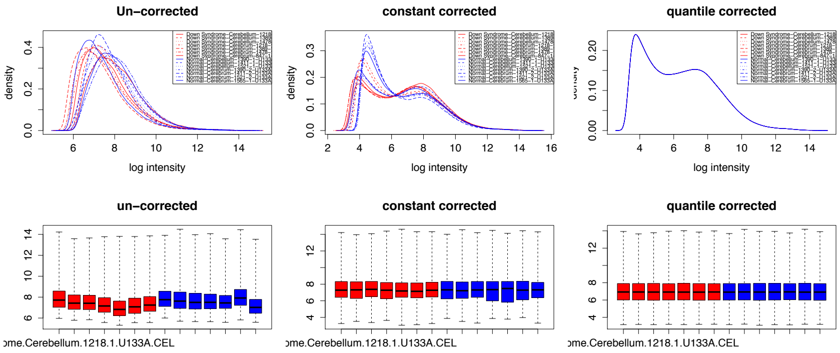
fasta.bioch.virginia.edu/biol4230 18

The significance of differences: normalization



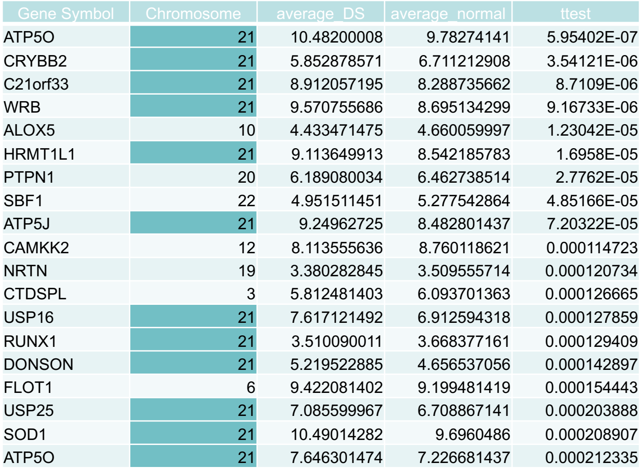
Why are the replicates different?
Should the bulk properties differ?

The significance of differences: normalization



Why are the replicates different?
Should the bulk properties differ?
Should individual genes differ?
Should blue (normal) and red (Downs) differ?

Differential Gene expression

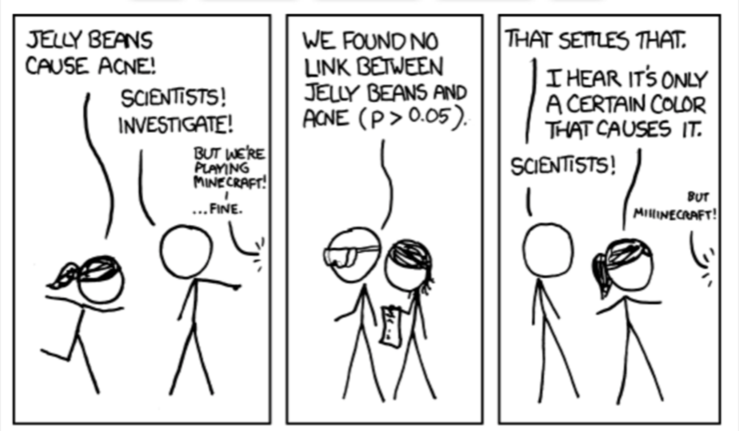


Gene Symbol	Chromosome	average_DS	average_normal	ttest
ATP5O	21	10.48200008	9.78274141	5.95402E-07
CRYBB2	21	5.852878571	6.711212908	3.54121E-06
C21orf33	21	8.912057195	8.288735662	8.7109E-06
WRB	21	9.570755686	8.695134299	9.16733E-06
ALOX5	10	4.433471475	4.660059997	1.23042E-05
HRMT1L1	21	9.113649913	8.542185783	1.6958E-05
PTPN1	20	6.189080034	6.462738514	2.7762E-05
SBF1	22	4.951511451	5.277542864	4.85166E-05
ATP5J	21	9.24962725	8.482801437	7.20322E-05
CAMKK2	12	8.113555636	8.760118621	0.000114723
NRTN	19	3.380282845	3.509555714	0.000120734
CTDSPL	3	5.812481403	6.093701363	0.000126665
USP16	21	7.617121492	6.912594318	0.000127859
RUNX1	21	3.510090011	3.668377161	0.000129409
DONSON	21	5.219522885	4.656537056	0.000142897
FLOT1	6	9.422081402	9.199481419	0.000154443
USP25	21	7.085599967	6.708867141	0.000203888
SOD1	21	10.49014282	9.6960486	0.000208907
ATP5O	21	7.646301474	7.226681437	0.000212335

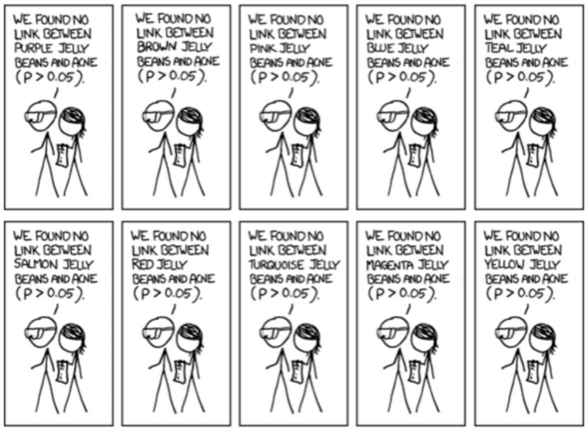
Differential Gene Expression

- Large quantity of data (>20,000 genes)
  - Affychip data has ?20 replicates per gene
  - RNAseq has counts (FPKM: Fragments per Kilobase per Million mapped reads)
  - but a small number of biological replicates
- Ideally, identify modest change (1.5x or larger) for modest levels of transcription
  - 10 or fewer transcripts may account for 90% of reads, so 5,000 transcripts for < 10% of reads
  - If technical replicates vary more than 2x, how do you measure 1.5x change?
- Large numbers of tests: how to correct?
  - Family-wide-error-rate (FWER) Bonferroni correction (used for similarity search E()-values)
  - False-discovery-rate (FDR, qvalue)

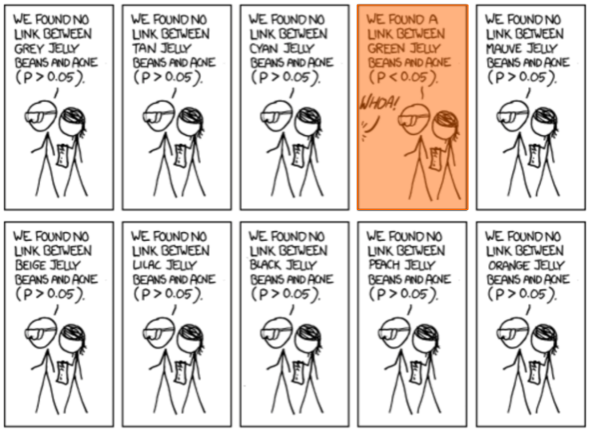
So many tests, what is significant?


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

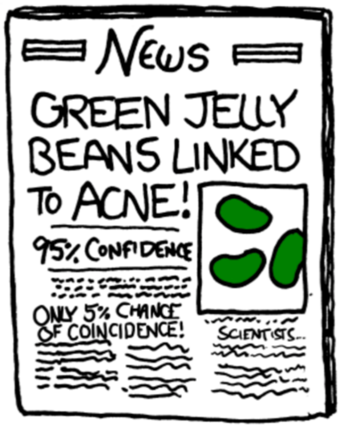
So many tests, what is significant?


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

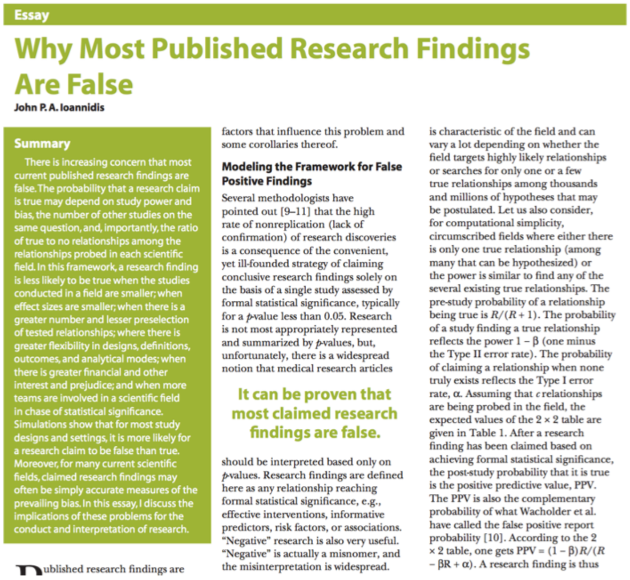
So many tests, what is significant?


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

So many tests, what is significant?


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Ioannidis, J. P. A. *PLoS Med.* **2,** e124 (2005).



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are

factors that influence this problem and some corollaries thereof.

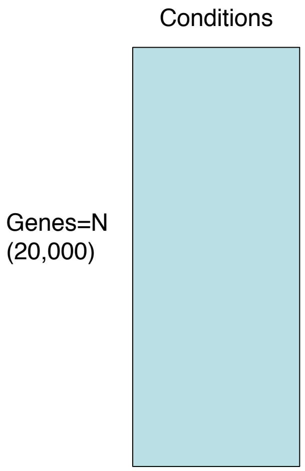
Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread.

It can be proven that most claimed research findings are false.

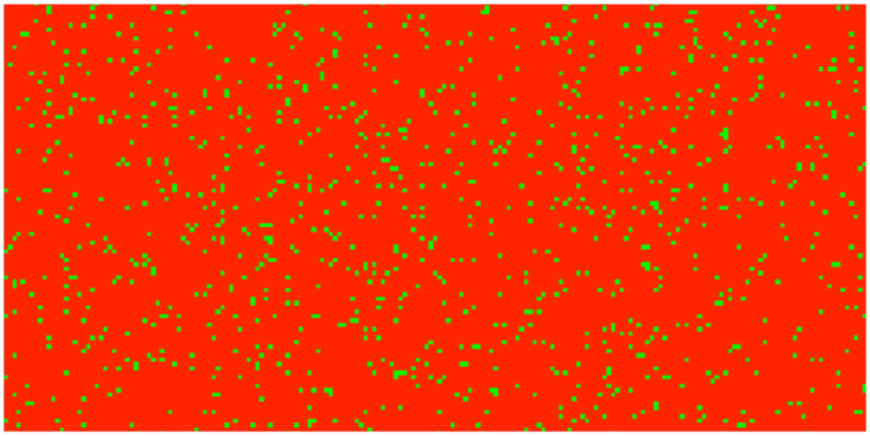
is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $\text{PPV} = (1-\beta)R/(R - \beta R + \alpha)$. A research finding is thus

How many tests?



At least N (~20,000) simultaneous tests

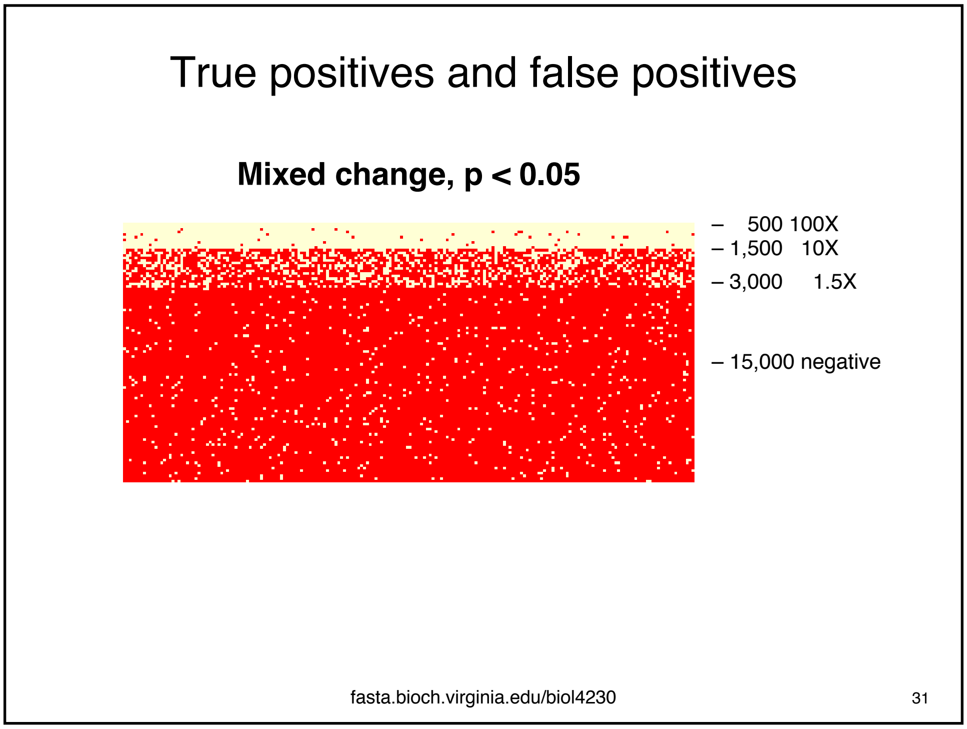
How many tests?



20,000 simultaneous t-tests on random normal data from the same distribution.There are 1,009 green points (false positives), making up 0.05 of the comparisons (at α = 0.05).

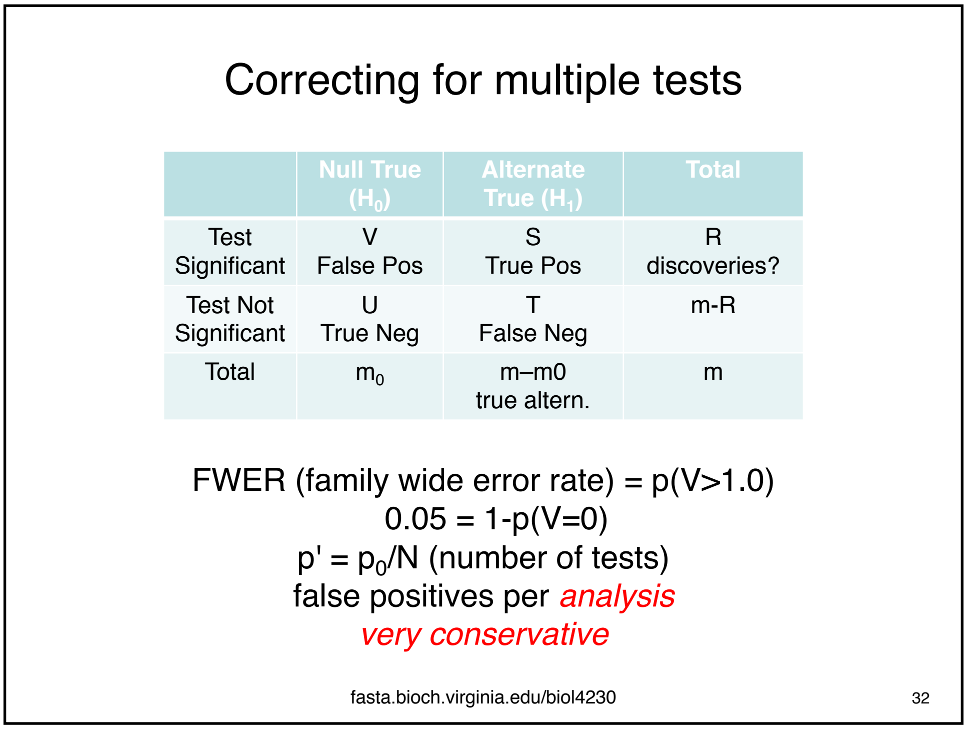
Correcting for multiple tests:

- Bonferroni:
  - E() = P D (similarity search)
  - calculate expectation as probability of result x number of tests
  - Family Wide Error Rate (FWER)
  - Ensures < 1.0 false positive among all results (<1.0 false positive after 20 studies with E<0.05)
- Q-value (False discovery rate, FDR)
  - sets a rate of false positives AMONG the set found to be significant
  - q-value < 0.01 says that one of the 100 "significant" results will occur by chance (10 of the 1000 significant)
  - which one?
    - One with least signal?
    - One with least fold change?



True positives and false positives

Mixed change, p < 0.05

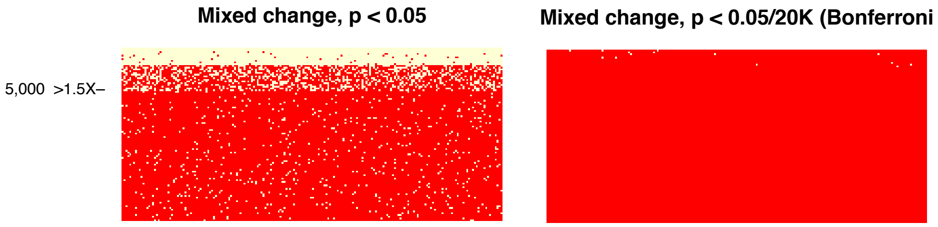


Correcting for multiple tests

	Null True (H_0)	Alternate True (H_1)	Total
Test Significant	V False Pos	S True Pos	R discoveries?
Test Not Significant	U True Neg	T False Neg	m-R
Total	m_0	m–m0 true altern.	m

FWER (family wide error rate) = p(V>1.0)
0.05 = 1-p(V=0)
p' = p_0/N (number of tests)
false positives per *analysis*
very conservative

True positives and false positives



FWER (family wide error rate) = $p(V>1.0)$
$0.05 = 1-p(V=0)$
$p' = p_0/N$ (number of tests)

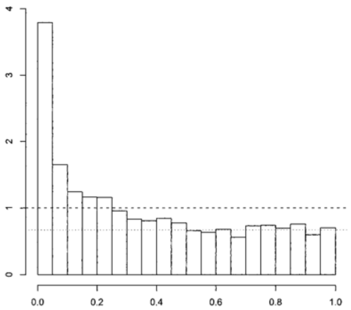
very conservative

Correcting for multiple tests

	Null True (H_0)	Alternate True (H_1)	Total
Test Significant	V False Pos	S True Pos	R discoveries?
Test Not Significant	U True Neg	T False Neg	m-R
Total	m_0	m–m0 true altern.	m

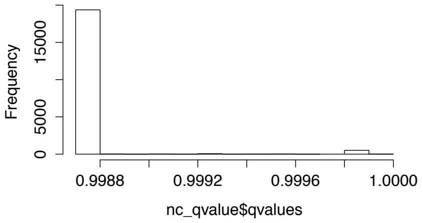
FDR (false discovery rate) = p(V/R)
Approx FDR *False* discoveries
among all discoveries
false positives per *discovery/true positive*

False-discovery rate (FDR)

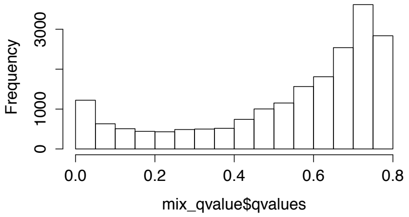


A density histogram of the 3,170 *p* values from the Hedenfalk *et al.* (14) data. The dashed line is the density histogram we would expect if all genes were null (not differentially expressed). The dotted line is at the height of our estimate of the proportion of null *p* values.

Histogram of nc_qvalue$qvalues



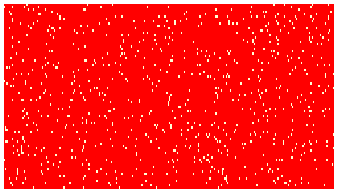
Histogram of mix_qvalue$qvalues



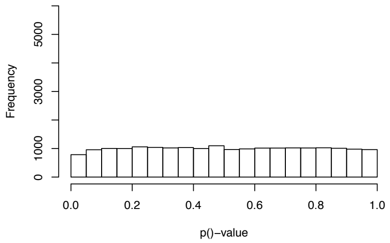
Storey (2003) PNAS 100:9440, Fig. 1

False discovery rate (FDR)

no change, p < 0.05

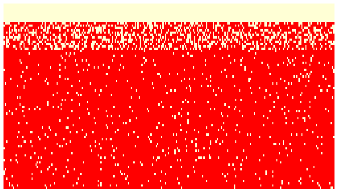


no change (p–values)

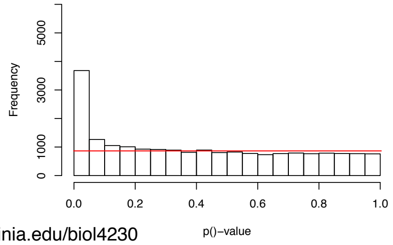


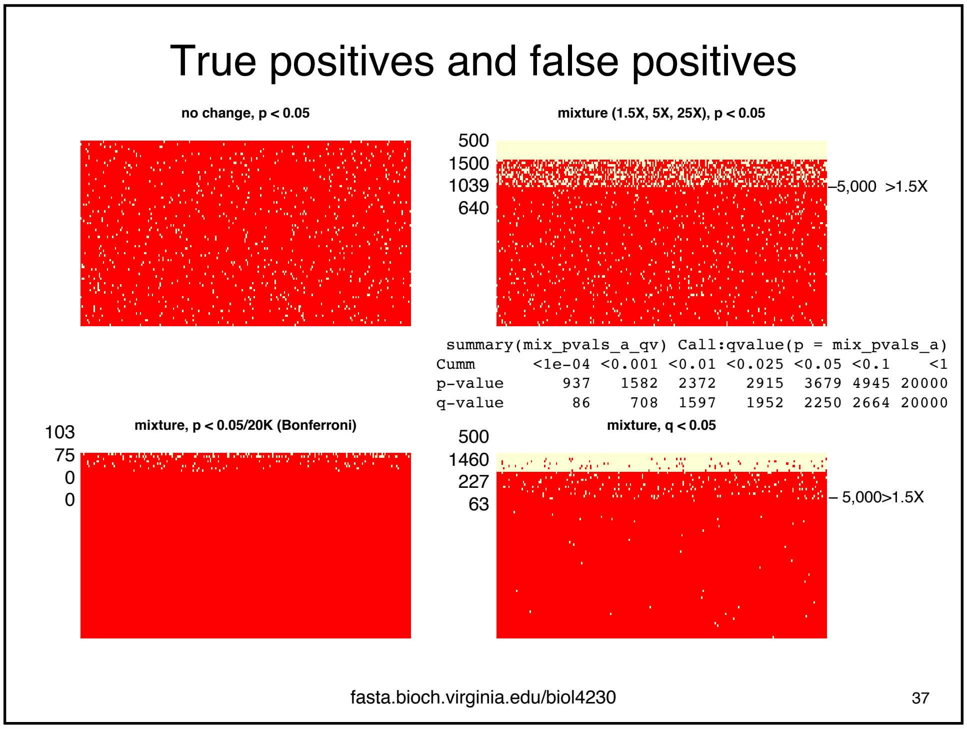
mixture (1.5X, 5X, 25X), p < 0.05

25X	500
5X	1500
1.5X	1039
nc	640



mixture (p–values)

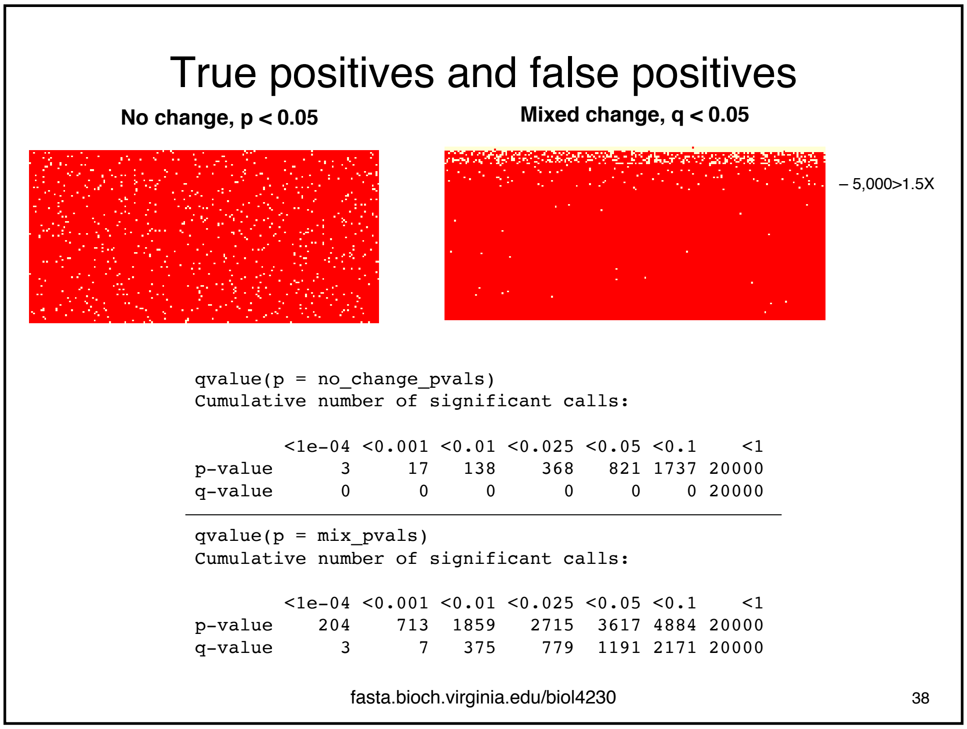




True positives and false positives

```
 summary(mix_pvals_a_qv) Call:qvalue(p = mix_pvals_a)
Cumm        <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1    <1
p-value        937   1582  2372   2915  3679 4945 20000
q-value         86    708  1597   1952  2250 2664 20000
```

fasta.bioch.virginia.edu/biol4230



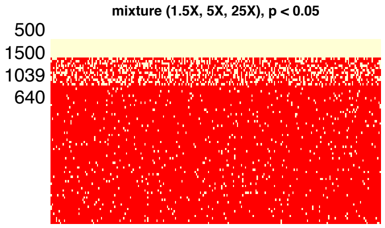
True positives and false positives

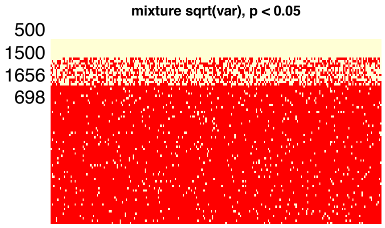
```
qvalue(p = no_change_pvals)
Cumulative number of significant calls:

        <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1    <1
p-value      3     17   138    368   821 1737 20000
q-value      0      0     0      0     0    0 20000
```

```
qvalue(p = mix_pvals)
Cumulative number of significant calls:

        <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1    <1
p-value    204    713  1859   2715  3617 4884 20000
q-value      3      7   375    779  1191 2171 20000
```

fasta.bioch.virginia.edu/biol4230

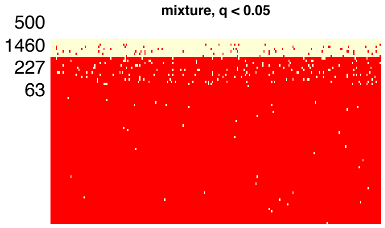
Reducing variance improves detection

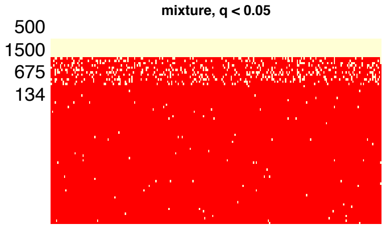




```
summary(mix_pvals_a_qv)
 Cumm      <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1
p-value       937   1582  2372   2915  3679 4945
q-value        86    708  1597   1952  2250 2664
```

```
qvalue(mix_pvals_b
 <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1
p  1853   2121  2826   3529  4354 5599
q  1381   1906  2176   2420  2809 3496
```





Differential Gene Expression

- Large quantity of data (>20,000 genes)
  - Affychip data has ?20 replicates per gene
  - RNAseq has counts (FPKM: Fragments per Kilobase per Million mapped reads)
  - but a small number of biological replicates
- Ideally, identify modest change (1.5x or larger) for modest levels of transcription
  - 10 or fewer transcripts may account for 90% of reads, so 5,000 transcripts for < 10% of reads
  - If technical replicates vary more than 2x, how do you measure 1.5x change?
- Large numbers of tests: how to correct?
  - Family-wide-error-rate (FWER) Bonferroni correction (used for similarity search E()-values)
  - False-discovery-rate (FDR, qvalue)