Some Statistics for RNA-Seq

Paul Eilers

Erasmus Medical Center, Rotterdam
and
Biometris, Wageningen

RNA-seq Course, Wageningen, June 2013

My subject today: RNA-Seq

- Let others find the SNPs: it is "stable" knowledge
- Copy numbers might be interesting
- But results are rather unstructured
- Gene expression is a well-defined subject
- And we can learn a lot from it
- My experience and inspiration: mostly human data
- But also a plant example: Soybean

Why do you want sequencing?

- You have too much money?
- You love large hard disks?
- You crave for long computations times?
- You don't want to look old-fashioned?
- No, you expect more bang for your bucks!
- Let's see

Overview

- Basically, sequencing is counting (c)DNA fragments
- Statistics for counts are special
- You will get a little theory and three case studies
- Yeast: two samples, two technical replicates
- Kidney and liver: two samples, 7 technical replicates
- HapMap cell lines, 10 biological samples

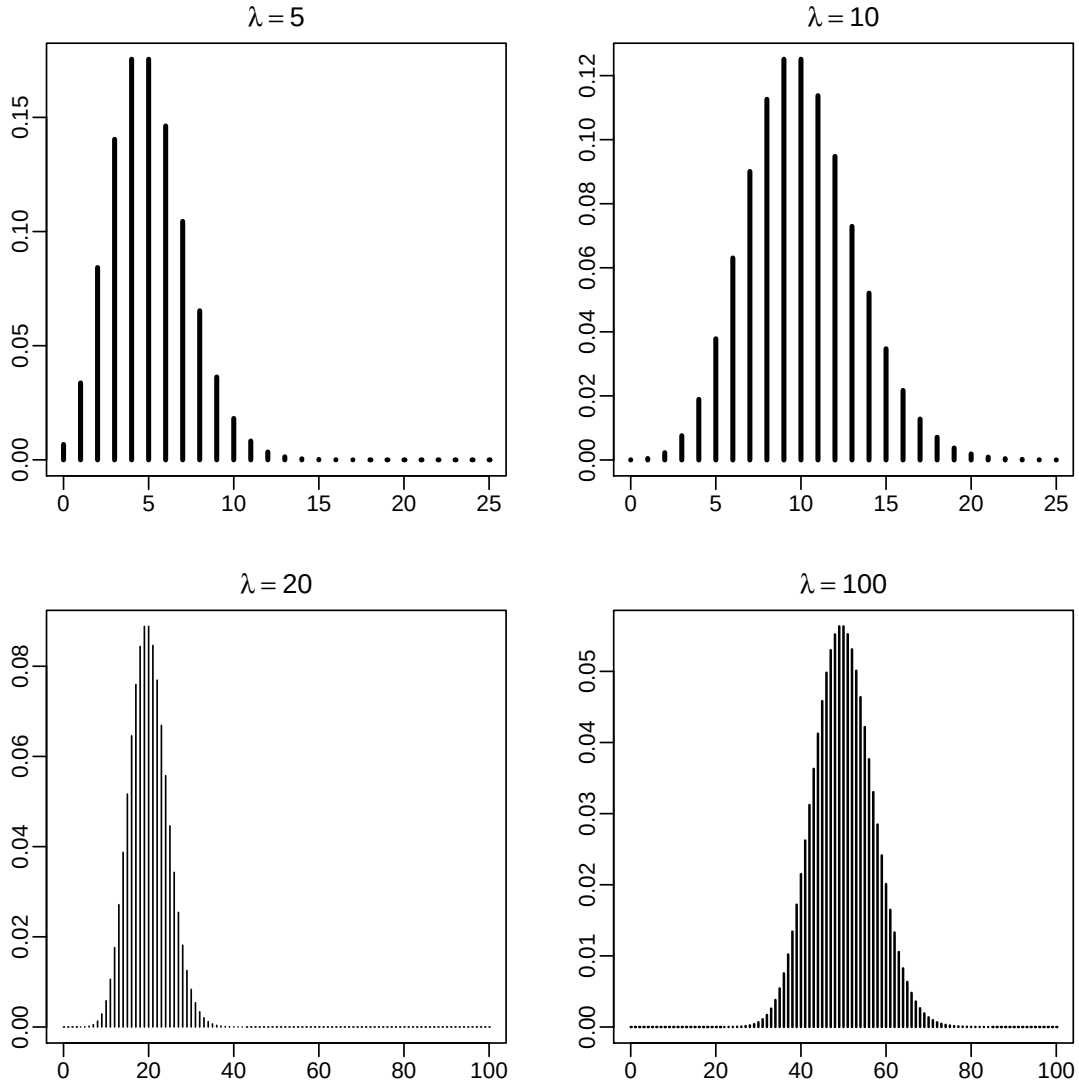
Counting

- We count reads per gene (per exon)
- On average 20 - 100 reads per gene
- Number of genes in examples: 30k (human) or 7k (yeast)
- Consider one arbitrary gene
- Probability that one arbitrary read maps to this gene is quite low
- It is 1/30000 or 1/7000, say 0.01%
- But we have many reads, say one million
- This is the province of Poisson statistics

The Poisson distribution

- If p is the probability of a hit (a read mapped to a gene)
- And n is the total number of reads
- Then you expect $\lambda = pn$ hits
- You will never get exactly λ hits
- There will be a range of possible values near λ
- Each value has a certain probability of occurring
- There is a formula for it, the Poisson distribution
- It only depends on λ

Poisson pictures



Poisson formulas

- Formula for the Poisson distribution ($k \geq 0$):

$$\Pr(y = k) = \lambda^k e^{-\lambda} / k!$$

- R has a function for it: `pr = dpois(k, lambda)`
- The average is μ
- The standard deviation is $\sigma = \sqrt{\mu}$
- This is remarkable
- If you know μ, you also know the uncertainty
- Compare to the normal distribution, where μ and σ are unrelated

Consequences of low counts

- Let's look at a comparison of two counts
- Rough approximation: 4σ difference gives a p-value below 0.05
- Compare counts y to given λ
- Fold change needed: $f = y/\lambda > 1 + 4/\sqrt{\lambda}$
- Example $\lambda = 5, 10, 20, 50, 100$
- Then $f \approx 2.8, 2.3, 1.9, 1.6, 1.4$
- Only large fold changes can be detected with low counts
- And that is for a very optimistic p-value

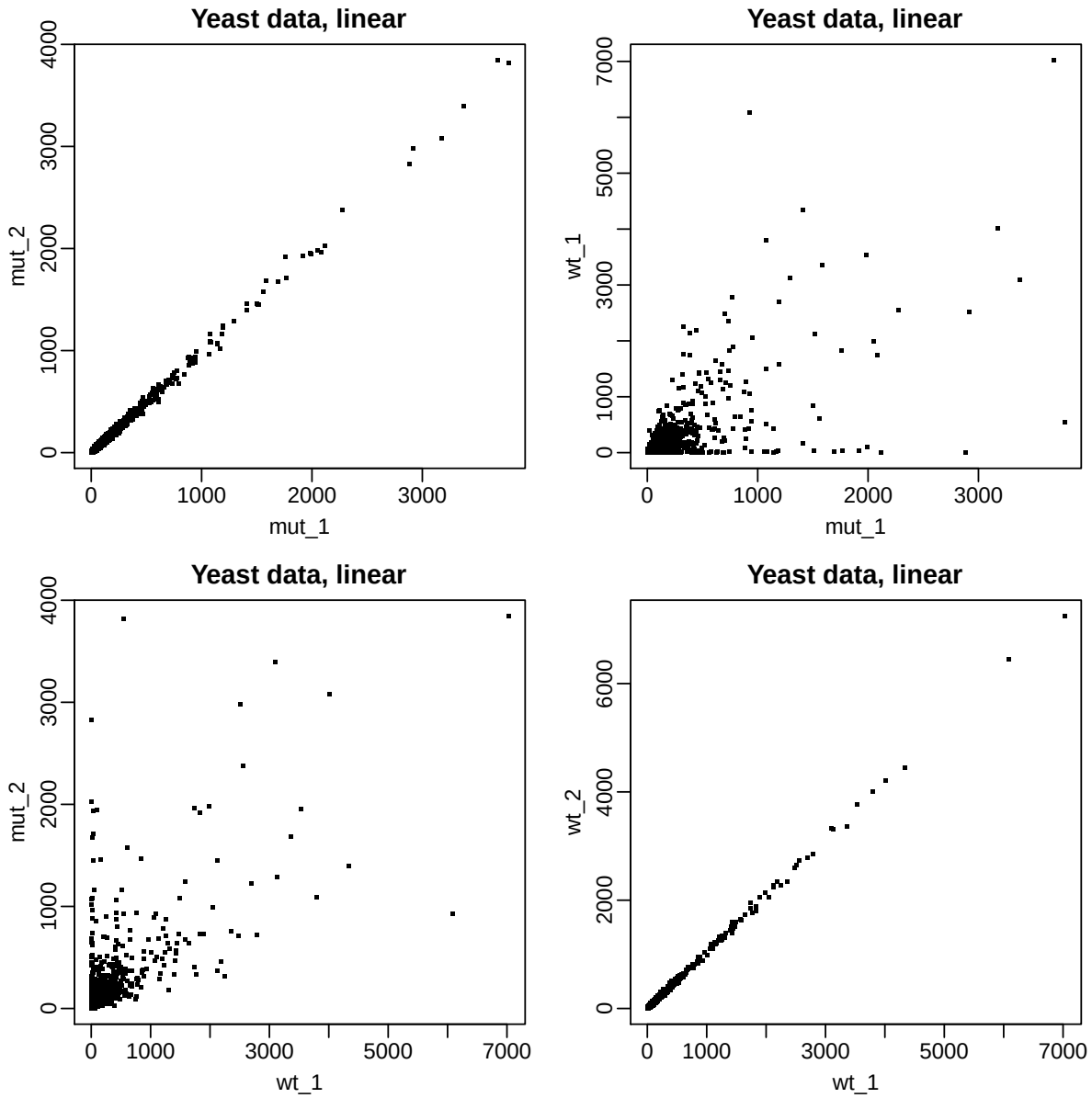
Poisson in real life

- The Poisson distribution is an idealization
- Do we see it in real life?
- Surprisingly (to statisticians), yes!
- We are used to seeing over-dispersion: $\sigma > \sqrt{\mu}$
- Examples: the yeast and kidney/liver data sets

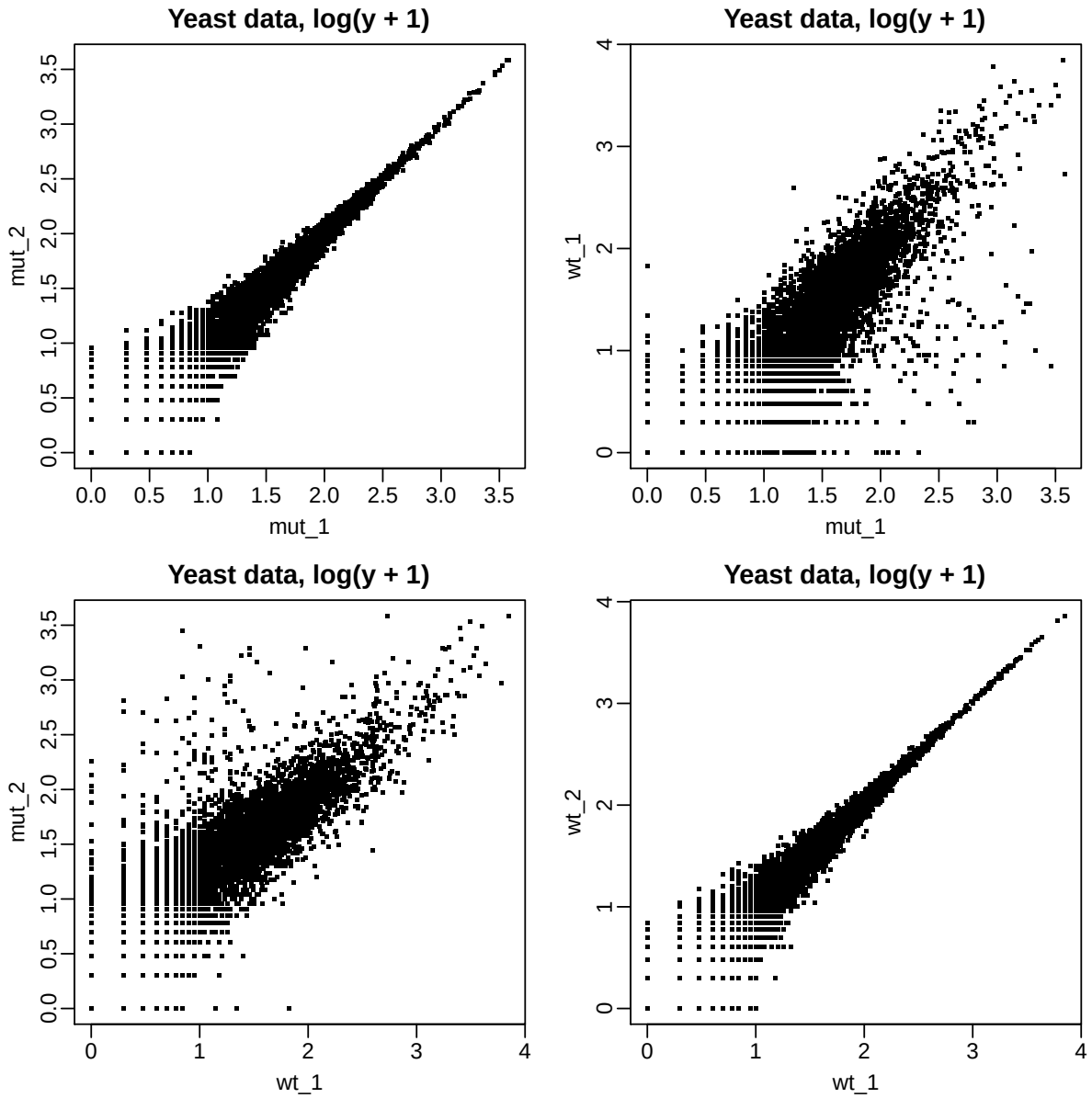
The yeast data

- Bioconductor package `yeastRNASeq` contains the reads
- Bioconductor package `Genominator` shows how to map them
- Vignette "Working with the ShortRead Package"
- Resulting table is in the materials for the practical
- Two type of yeast (wild-type and mutant), duplicate runs
- A little over 400k reads per run; 7124 Ensembl gene IDs

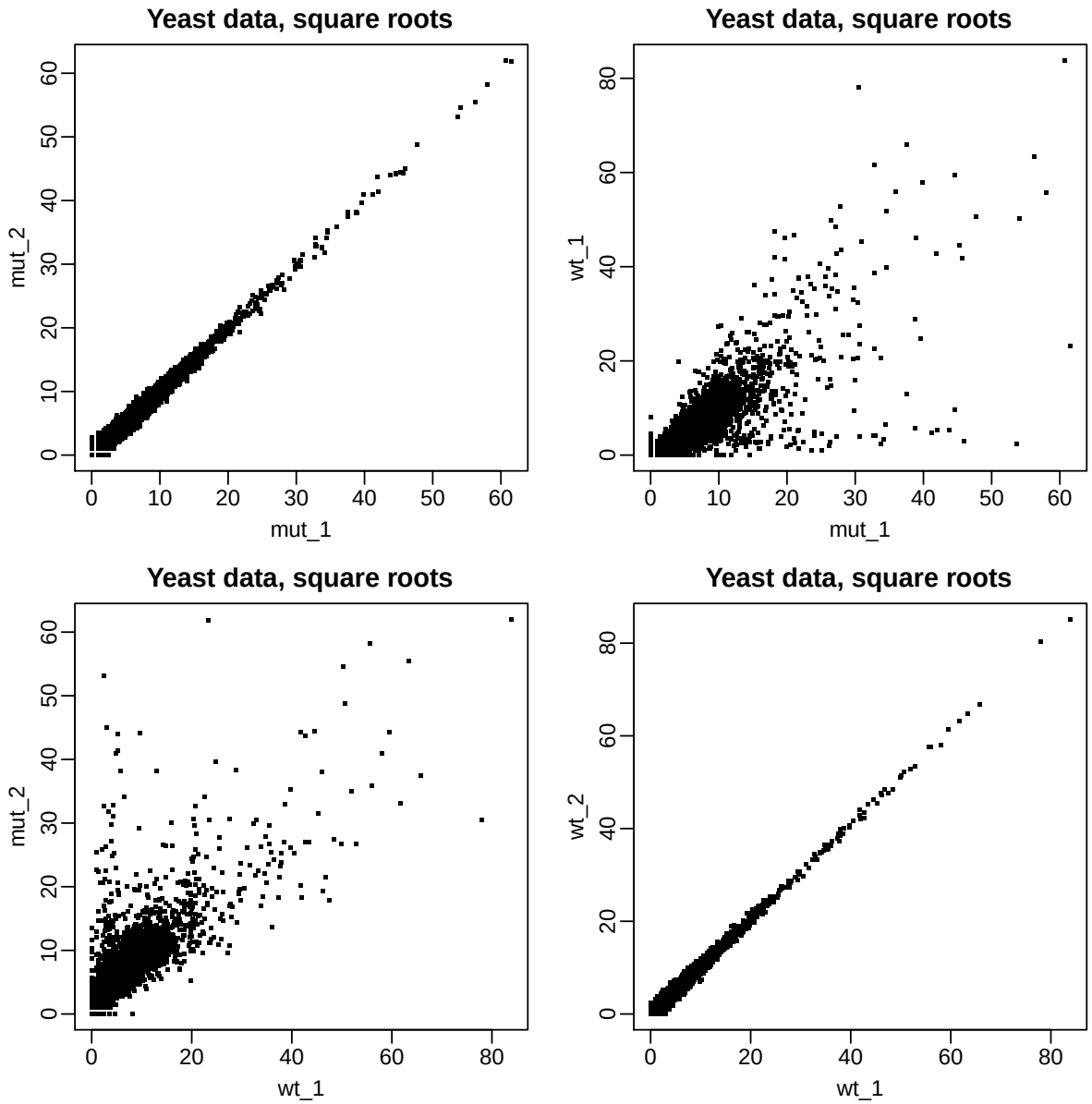
Plots of the yeast data: large counts dominate



Plots of logarithms: noisy at low counts



Plots of square roots: a good compromise



How to judge (technical) reproducibility?

- Consider the two wild-type samples
- Put counts in a table, $Y = [y_{ij}]$, with two columns and 7124 rows
- Compute the following
  - sums per row: a_i for row i
  - sums per column: b_j for column j
  - total of the whole table: t
  - expected values for y_{ij}: $\mu_{ij} = a_i b_j / t$
  - standardized residuals: $r_{ij} = (y_{ij} - \mu_{ij}) / \sqrt{\mu_{ij}}$
  - $s = \sqrt{\sum_i \sum_j r_{ij}^2 / (m-1)(n-1)}$.
  - (table has m rows, n columns)
- Now s should be close to 1, the Poisson limit.

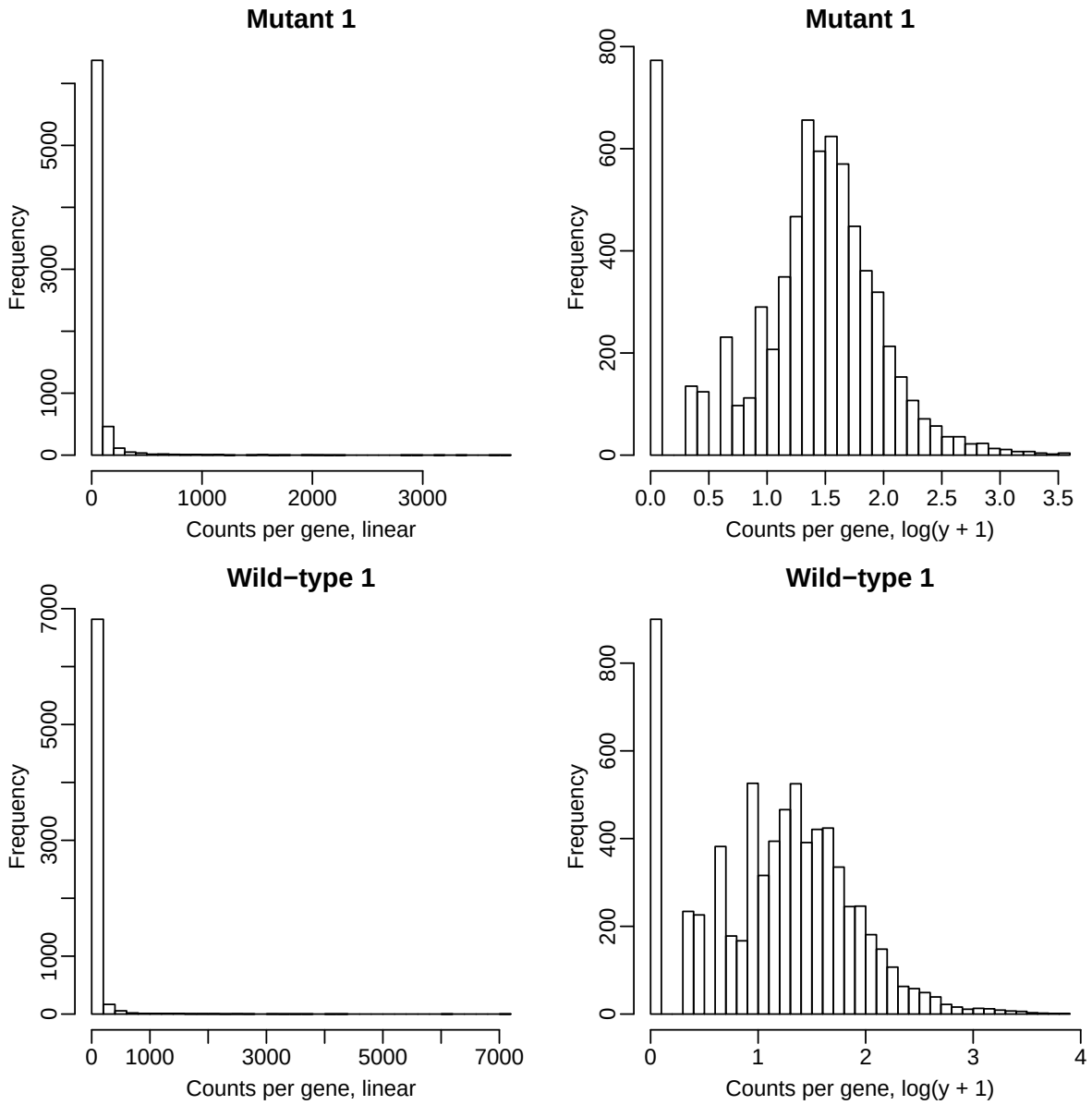
More on reproducibility

- The procedure is like the χ^2 test for contingency tables
- In fact $\chi^2 = \sum_i \sum_j r_{ij}^2$
- But the test is not quite meaningful here
- The size of s matters
- You should first remove rows of Y with all zero counts
- For wild-type yeast we get $s = 1.04$
- For the mutant it is $s = 1.03$
- Excellent results

How many genes are useful?

- Zero counts are useless: about 9% (of the genes)
- Counts below 10 are hardly useful
- For mutant: 29%, for wild-type: 44%
- Counts above 100 are reasonably useful
- For mutant: 12%, for wild-type: 12%
- So you lose a lot of genes

Histograms of yeast counts



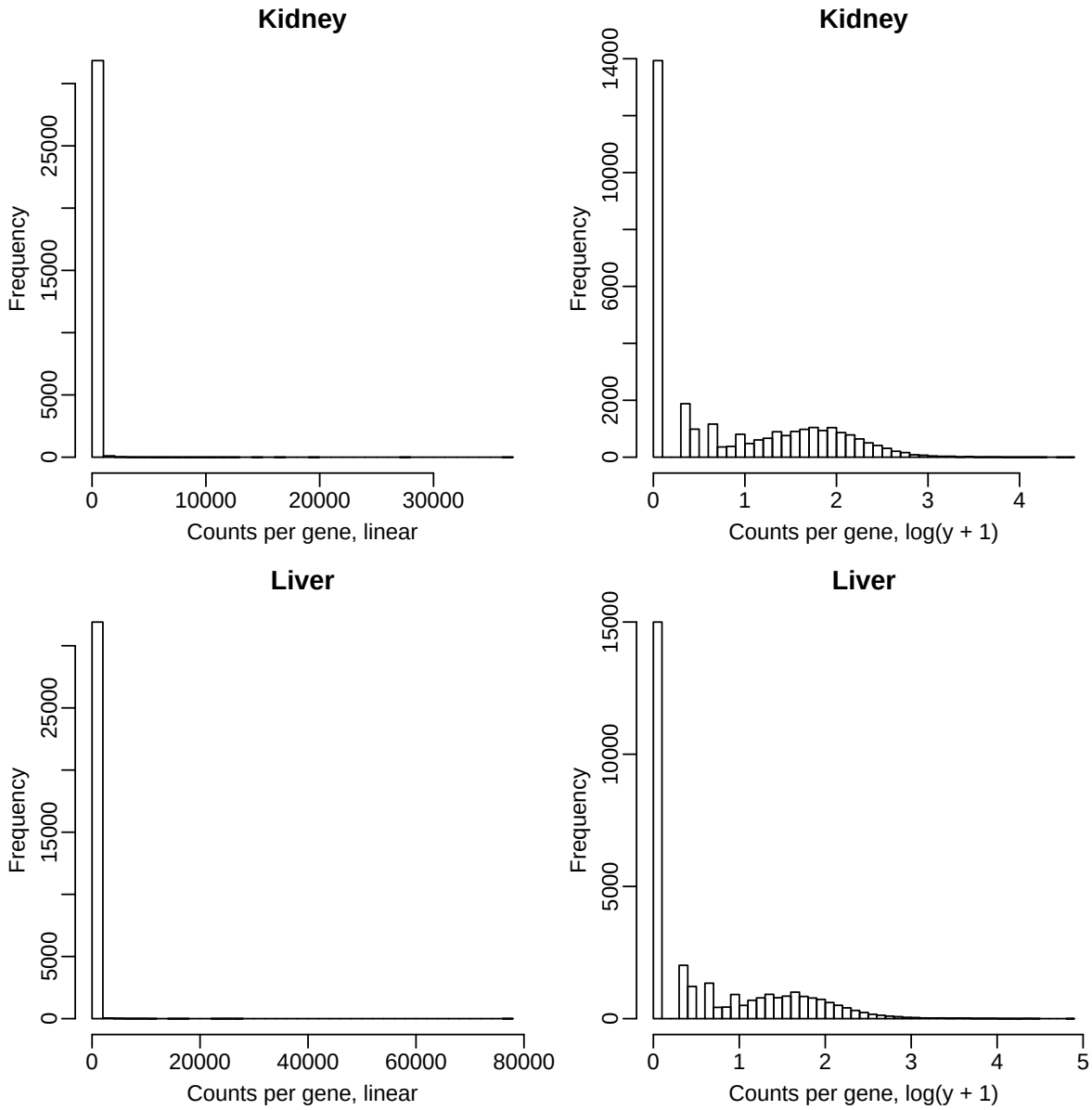
Groups with multiple samples

- The previous reasoning was based on single samples
- Often you will have groups with multiple samples
- If the Poisson assumption is true ...
- Add the counts (per gene) of all samples per group
- Treat them as single “super-samples”
- Check the assumption as for reproducibility
- With a table of counts per group

The Marioni et al. data

- Paper: Marioni et al. *Genome Res.* 2008 18: 1509-1517
- One kidney sample, one liver sample
- Seven technical replicates of each
- Illumina Genome Analyzer
- Counts per gene given as supplementary file
- You will find it in the practice material

Histograms of Marioni data



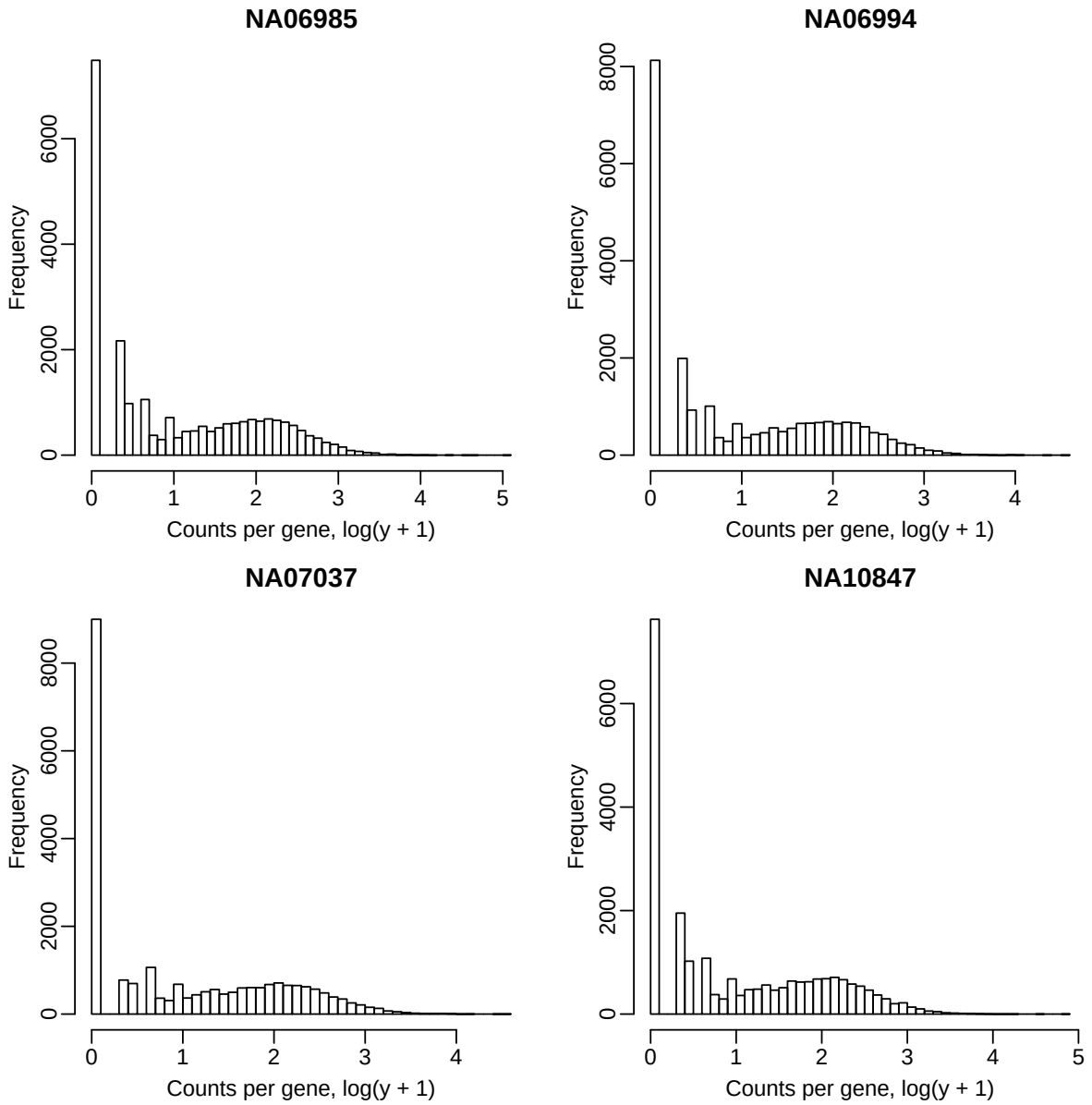
Summary of Marioni data

- 32000 genes, 1.7M reads on average
- Very good technical reproducibility
- Not many useful genes
- Over 40% has zero counts
- About 20% has more than 100 counts

Montgomery et al. data

- Montgomery et al (2010) *Nature* 464: 773–777
- Comparison of Affy arrays and sequencing
- HapMap (NA) cell lines
- Hansen et al. (2012) used part of data for Bioconductor `cqn` package
- The package contains a table: 23522 genes and 10 subjects
- It is also in the practice materials

Histogram of counts



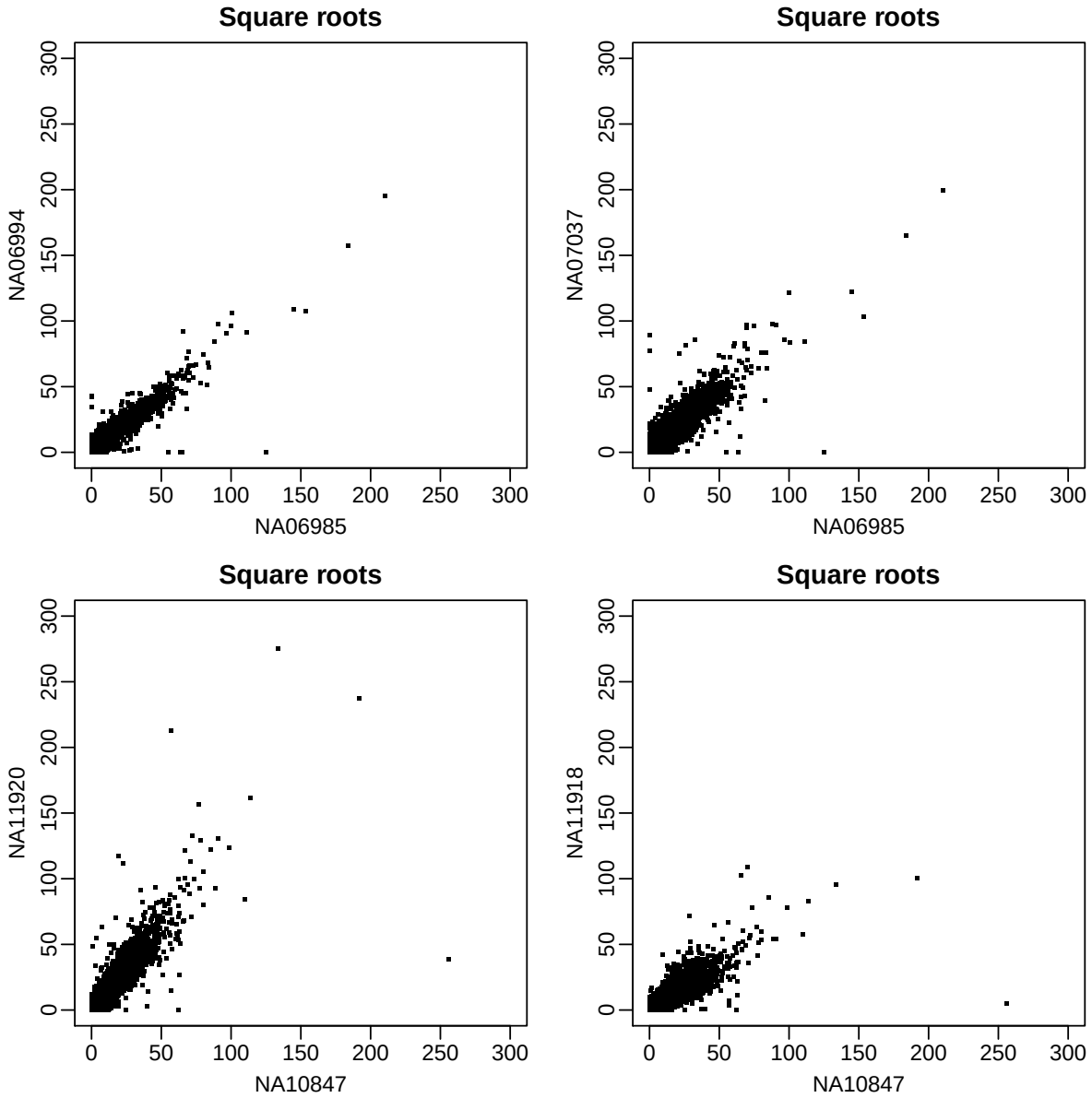
Genes with top ten largest counts

	NA11920	NA11918	NA11931	NA12003	NA12006	NA12287
ENSG00000211899	1490	24	127366	3657	54321	954
ENSG00000019582	75796	9167	52518	56673	91131	27518
ENSG00000167658	56289	10146	56159	78627	61390	64243
ENSG00000211895	4575	10482	722	52121	6648	12752
ENSG00000211890	45407	53	9	29	86	15
ENSG00000075624	26068	6905	26267	16048	37149	16675
ENSG00000136167	14872	7403	29946	14018	34726	20874
ENSG00000099875	8648	2967	10543	32190	32795	13347
ENSG00000100345	7094	3290	9109	11726	28651	5014
ENSG00000102962	3981	314	7826	20046	26670	2460

What do we see?

- Enormous variations in counts
- Ratios of 1000 to 1 between individuals
- These are cell lines
- Biological variation cannot be that large
- Reproducibility computation gives $s = 5.8$
- Many times larger than what Poisson says
- This is really worrisome
- About 68% (37%) of counts larger than 10 (100)

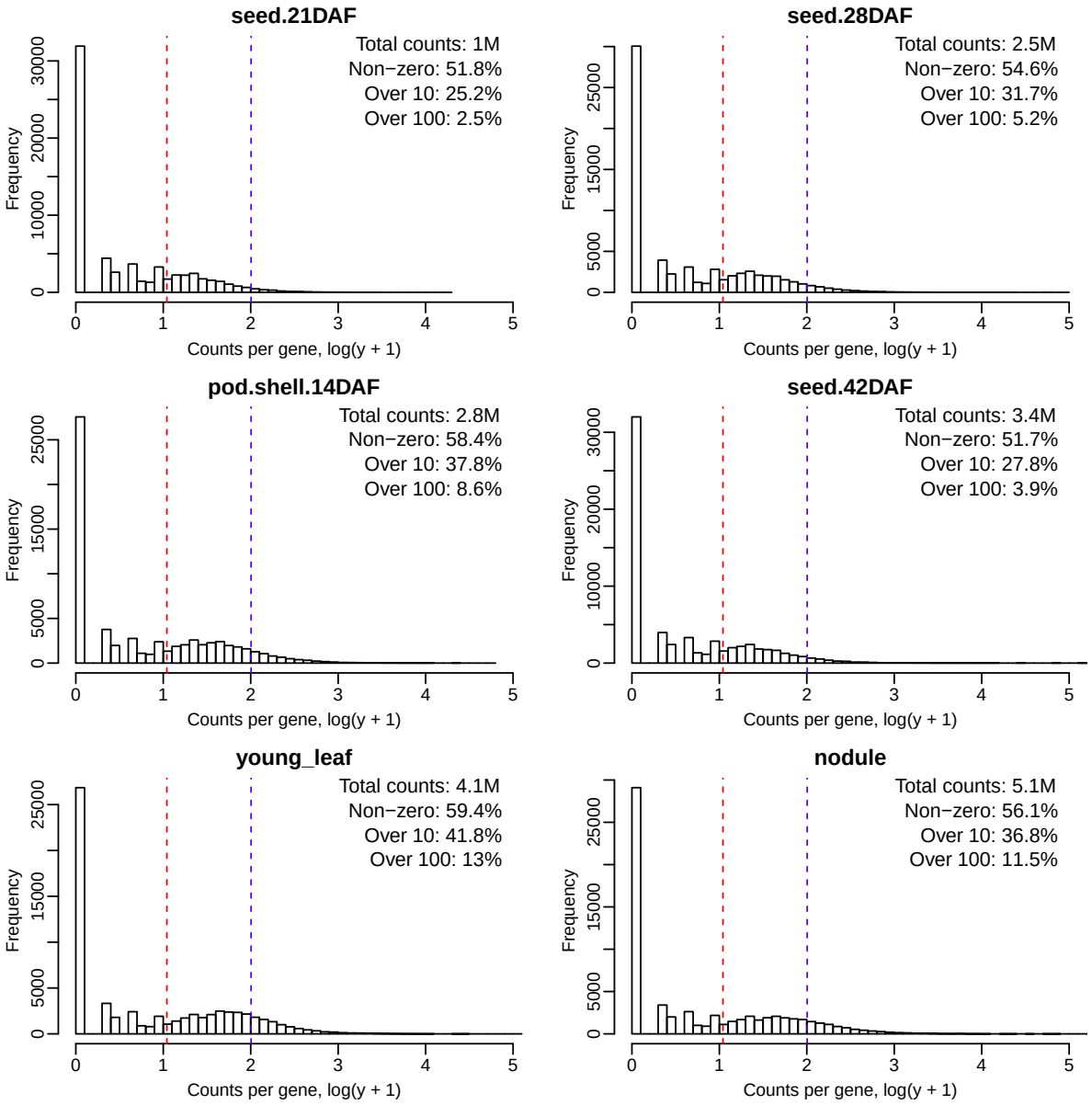
Some comparisons between samples



The Soybean data

- Paper by Severin et al. *BMC Plant Biology* (2010)
- "RNA-Seq Atlas of *Glycine max*: A guide to the soybean transcriptome
- All parts (14) of the plants, as well as seeds
- No replications
- Illumina Genome Analyzer II
- One to 5 million useful reads

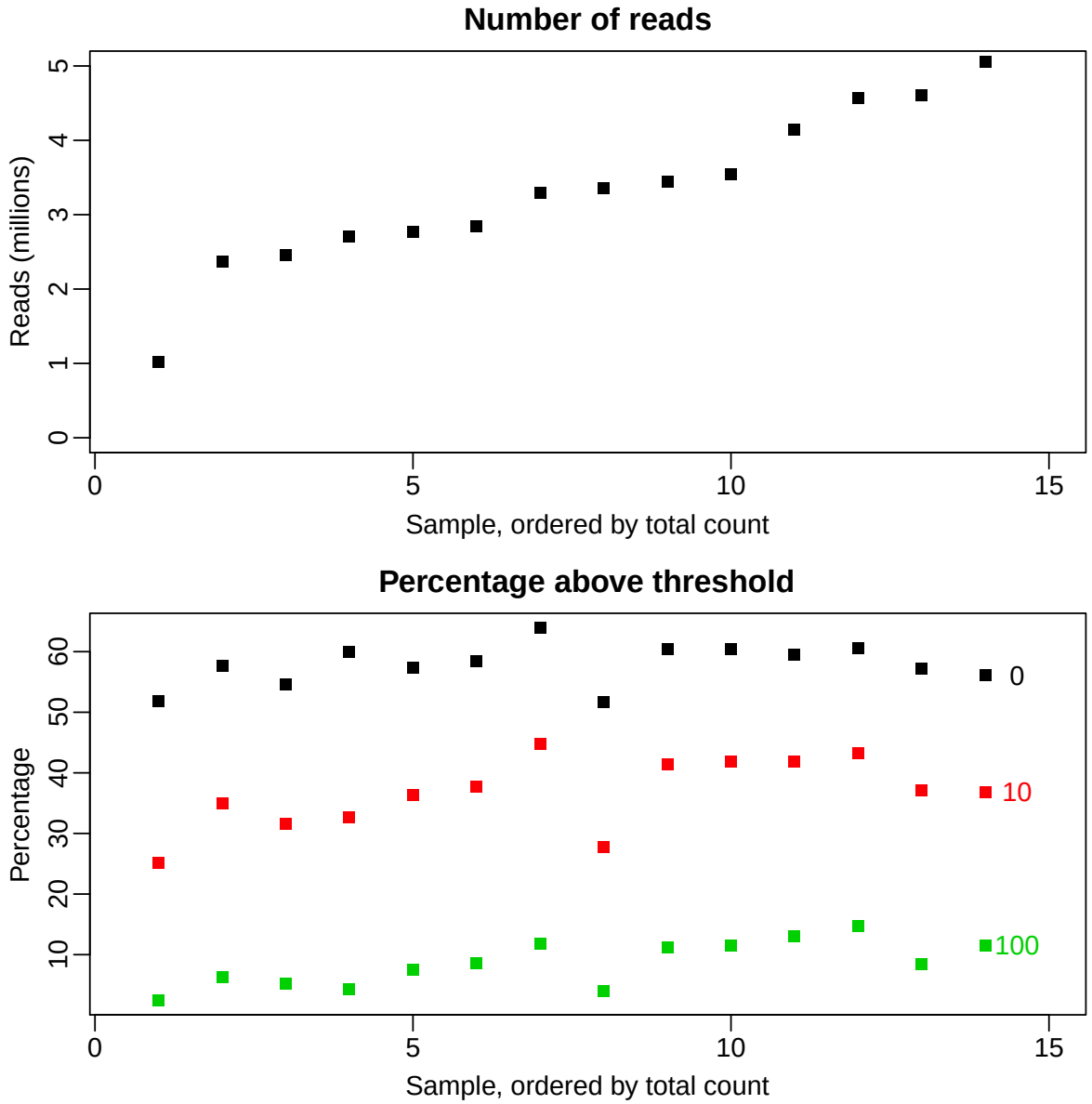
Some typical Soybean histograms



Soybean summary

- Low counts are the rule
- More than half are zero
- Only 10% or less above 100
- Remarkable and consistent distribution at low counts

Impression of the useful fraction



The high dynamic range myth

- Severin et. al. (abstract): "... dynamic range of over six orders of magnitude ..."
- This is nonsense
- Largest and second largest count: 1,102,650 and 300,802
- Only 0.002% of the counts is larger than 100,000
- And 0.02% is larger than 10,000
- A tiny minority

Conclusions

- Technical reproducibility looks very good
- Biological reproducibility (cell lines) is worrisome
- Low yield of genes with large enough reads
- My conclusion: RNA-Seq is hyped
- It might offer less then you expect