So you have data that looks like this…

	Group 1			Group 2		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
Gene 1	524	345	134	13	4	20
Gene 2	52	78	35	43	55	30
Gene 3	5	15	0	3	9	25
⋮						

Common Question: Is there underlying difference between the groups (differential expression) or is any difference just **natural variability**?

Statistics: Do the groups have the same ***expected value*** or ***mean***?

A familiar (?) comparison

- Microarrays:
  - gene expression is (summary) of fluorescent intensity of spot after normalization
  - On log 2 scale, usually measurements range 4-14
  - Common to take a t-test of log of expression of gene expression

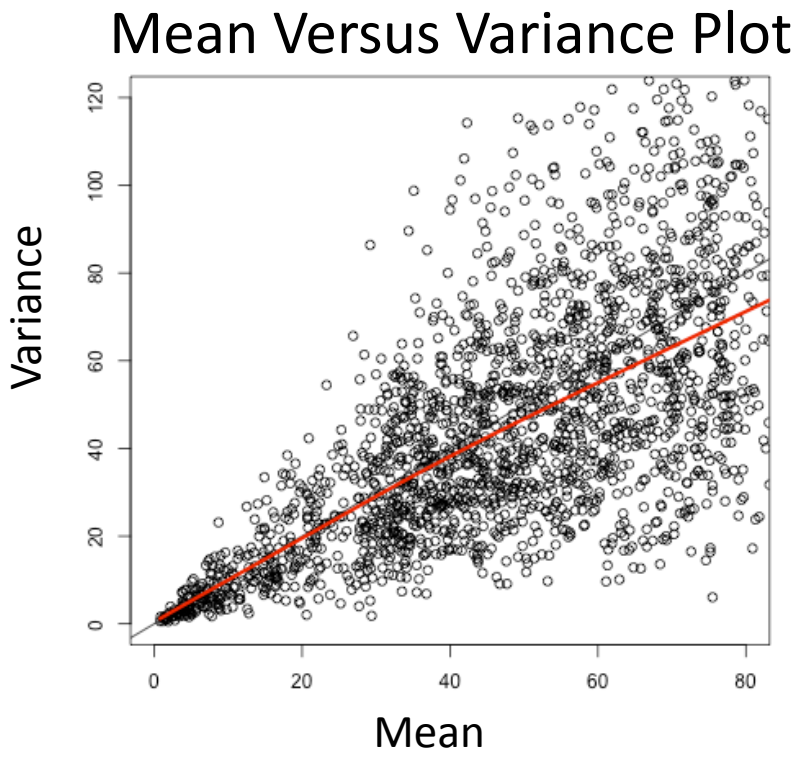
Are RNA-Seq results different?

Well, yes …

Two issues to consider

- Still need to normalize – can see raw counts not comparable because of different scales
- Data are counts – not appropriate for t-test
  Question of what is the 'right' measure of natural variability

	Group 1			Group 2		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
Gene 1	524	345	134	13	4	20
Gene 2	52	78	35	43	55	30
Gene 3	5	15	0	3	9	25
⋮						
TOTAL	16M	40M	10M	20M	18M	23M



Each point a ‘gene’ with 10 samples

Poisson distribution: Common ‘distribution’ or description of how multiple measurements behave

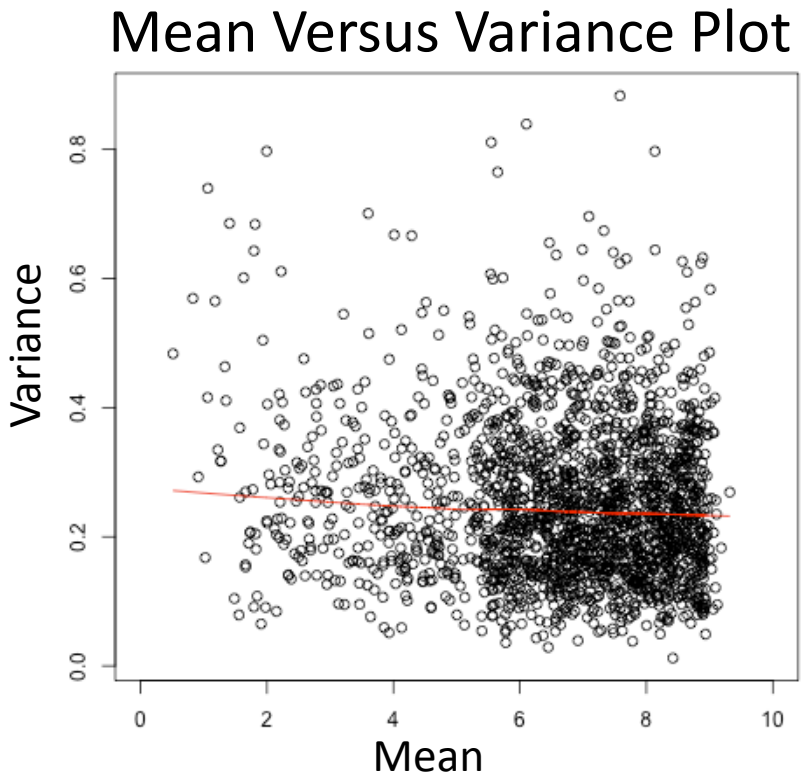
Variability of measurement equals mean (multiplicative error)

Common ‘fix’: transform the data

- Take square-root or log of counts

  (similar to microarrays)

Variance stays same as mean changes



If expected counts high, won’t be too bad of fix
But, if counts low, not be great fix …

Better: Use tests meant for count data

- Common tests you may read about
  - Fisher's Exact Test
  - Chi-Square Test of Independence
  - Binomial test
- All pool samples from same group (i.e. sum counts) to test whether the same mean

	Group 1			Group 2		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
Gene 1	524	345	134	13	4	20
Gene 2	52	78	35	43	55	30
Gene 3	5	15	0	3	9	25
⋮						
TOTAL	16M	40M	10M	20M	18M	23M

	Group 1	Group 2
Gene 1	1003	37
Gene 2	330	128
Gene 3	20	37
⋮		
TOTAL	66M	61M

This misses key variability between samples – loss of information

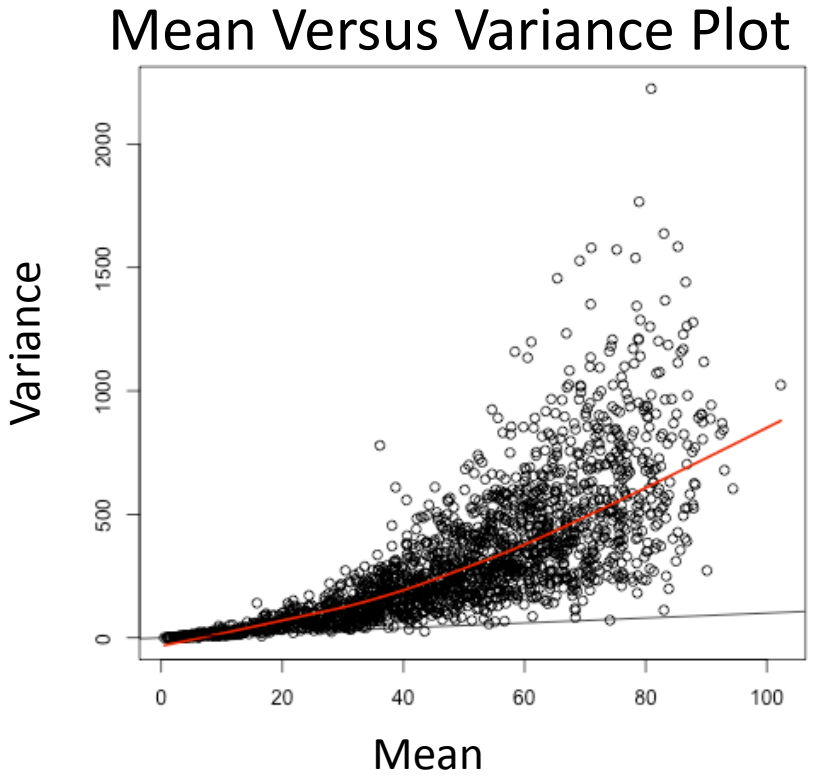
- Statistically, this is because of the distribution (Poisson) we assumed, mean=variance.
  Practically, that means there is no unexplained differences between samples
- T-test, on other hand, estimates variability in each group
  Statistics: each group a normal with mean and variance that can be anything

Fix: Alternative distribution Negative Binomial

- Main difference:

  variance = mean + ϕ mean^2

  ϕ: addition parameter allows unexplained variability



Still have variability grow with mean, which is good, because that's what real data looks like!

Is there a single ϕ, or a different one for each gene ϕ_g?

Probably different for each gene

- But there's a tension here… estimating lots of ϕ_g with small sample sizes (e.g. 4-8)
- Similar problem in microarrays, e.g. where very low variation in a gene makes small differences look significant
- 'Empirical Bayes': Estimate ϕ_g for each gene *g.* Then ***shrink*** the estimates ϕ_g so more like a common *prior* value ϕ.
- Stabilizes the estimates

More details about implementation

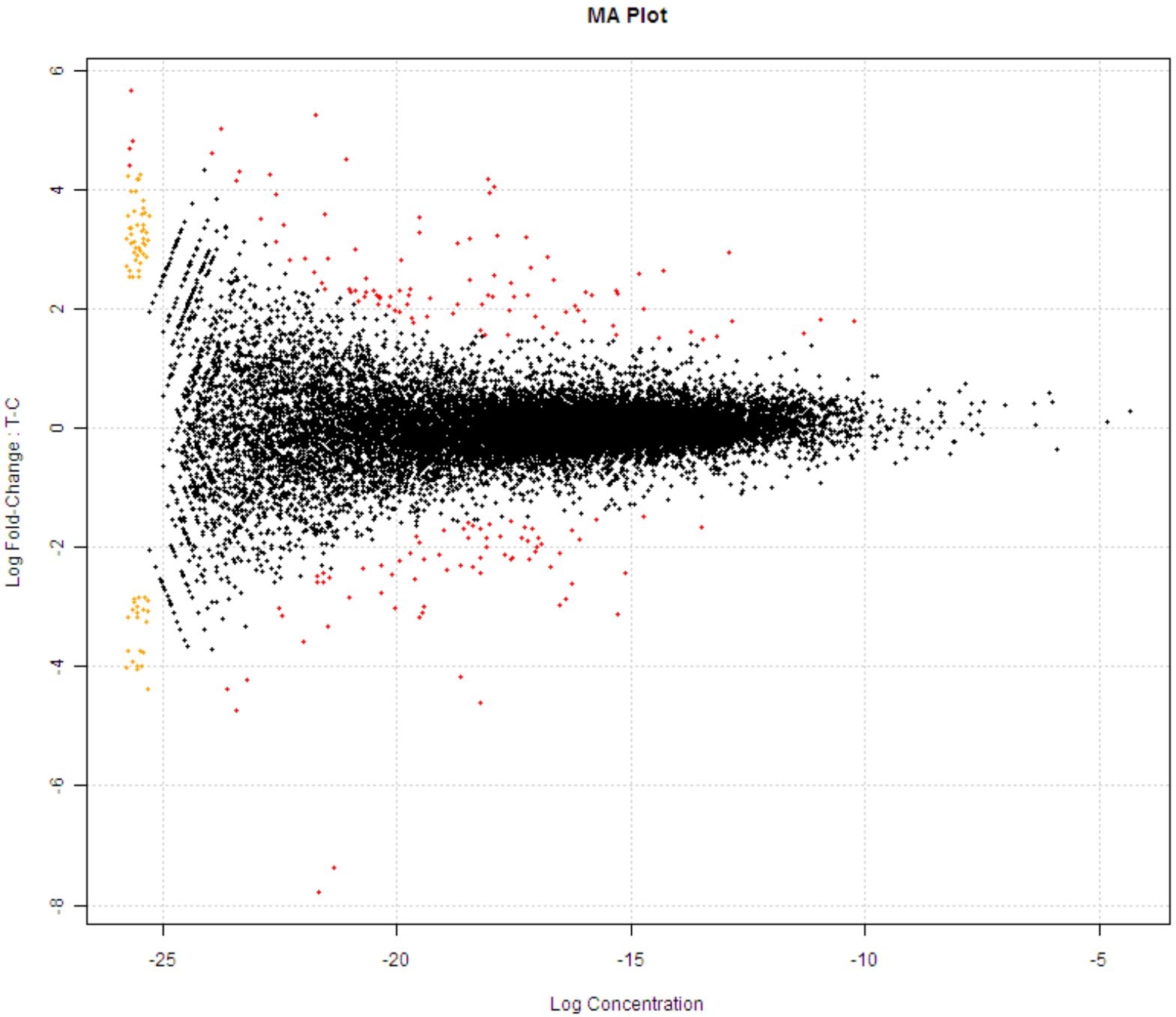
- Implemented in two different R packages: edgeR and DESeq.
  - Differences in how estimate
  - edgeR more mature, handles more situations
- More complicated situations edgeR handles (we won't talk about)
  - Multiple groups and even many different factors
  - Alternative offsets, useful for comparing #skip intron versus #overlap intron, for example.

Poisson

- Natural variability is only due to sequencing
  'sample size' is depth of sequencing
- Tend to find many things different
- Will preferentially find DE in genes with large number of counts and small differences between groups

Negative Binomial

- Natural variability from the biology and sequencing
  'sample size' is number of replicates and depth
- If dispersion is large, difference between groups must be greater to say significant


MA Plot
Log Fold-Change : T-C
6
4
2
0
-2
-4
-6
-8
-25
-20
-15
-10
-5
Log Concentration

Is pooling samples ever right?

Depends on what is the sample:

- Pure replication of sequencing, e.g. across lanes, looks very Poisson (which indicates good quality). Reasonable to sum the counts.
- Generally reasonable across flowcells/sequencing runs (but beware of possible differences if runs done at very different times, different technicians, different library preps, etc)
- Generally NOT true for any real replication of the experiment

Note, for other questions, may also be reasonable (e.g. is there any expression of this exon on any condition)

Normalization: make counts comparable

- **R**eads **P**er **K**ilobase of exon model per **M**illion mapped reads

$$RPKM(\text{sample i}) = 10^9 \text{ x } \frac{\#\,\text{counts}}{L_{region} N_i}$$

- Accounts for both differences in sequencing depth, and differences in length of region

$$\text{For differential expression} = \frac{\#\,\text{counts}}{N_i}$$

- But N_i not great for normalizing between samples

Different normalization factors

- Basically pick a different constant based on distribution of data
  Bullard, Purdom, et al (2011) *BMC Bioinformatics*
  Robinson & Oshlack (2011) *Genome Biology*
- Practical importance for DE: regardless method, can't just use adjusted values with count based techniques like edgeR; wrong measure of variability
- Simple example: if multiply all by 100, totally different results…