

Some (probably naïve and
simplistic) early morning
thoughts on statistics, and
RNAseq.

Titus Brown

Sampling M&Ms from very large boxes of them.

Setup 1: you have two trunks of M&Ms

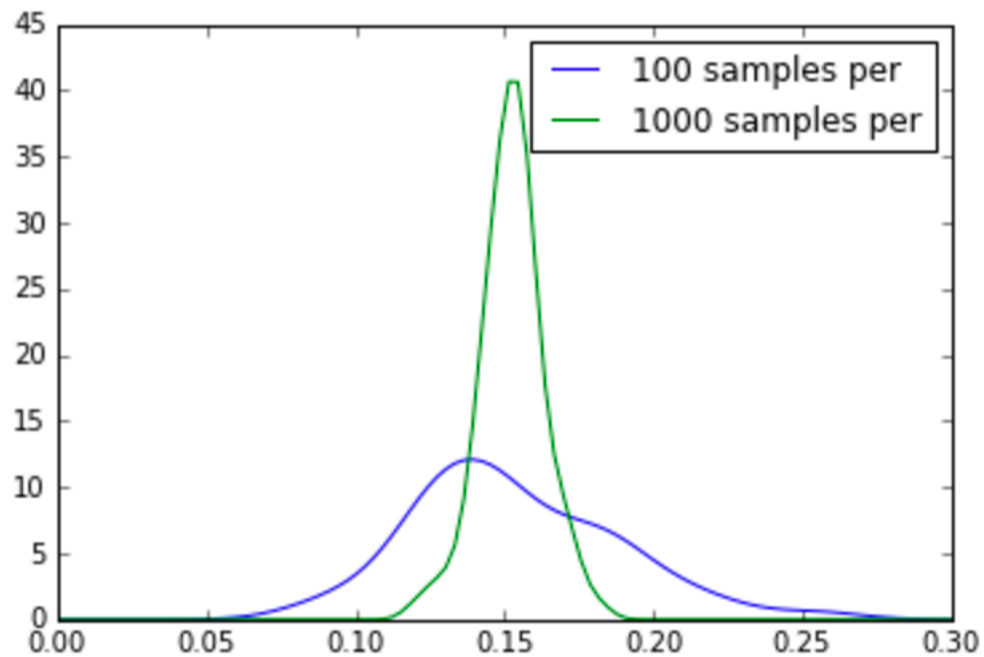
You want to know which trunk has more blue M&Ms
...but it costs you money/time to sample from each trunk.

Assumptions:

Trunks are big enough that you're sampling with replacement
Trunks are well mixed.

Conclusion from 1st set up --

The deeper you sample, the lower your relative variation across replicates.



Setup 2: you want to know which M&Ms are more or less prevalent between trunks

Now ***multiple testing*** becomes a problem –

If you have 10 colors you're looking at,

And a false-positive rate of 0.2 on your statistical test,

roughly 2 of your “different prevalence” calls will be wrong.

(Similar reasoning applies to false negative rates.)

Setup 3: different trunk-producing factories.

You can order trunks of M&Ms from two different factories.

You believe that the factories are producing different color M&Ms at different rates.

You want to know which M&Ms are being produced at different rates.

Each factory has a different "true" rate;

Each trunk is produced according to that "true" rate but each trunk is a somewhat inaccurate sampling of that true rate;

Setup 3 complicates your sampling strategy!

How deeply do you sample M&Ms from each trunk, vs how many different trunks do you sample? (since it costs extra money to get each trunk – shipping charges!)

(This is separate from how many times you sample M&Ms from each trunk! => technical replicates)

Only in the last case, are you doing science...

First case, you have an n of 1 (you're looking at only one trunk).

- What if the factory was having a weird day? How would you know?

Second case, you're ***still*** only looking at one trunk.

Third case, you can look at ***technical*** and ***experimental*** replicates both, and make a statistical argument.

You still have to worry about *batch effects*.

E.g. Maybe you have different people sampling the trunks on different days, and those people have biases in how they're sampling

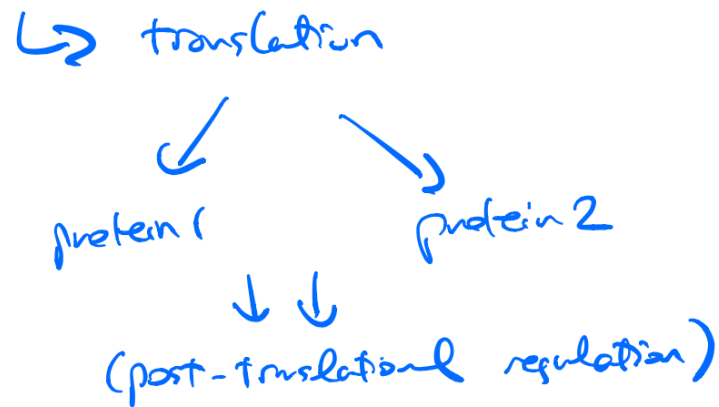
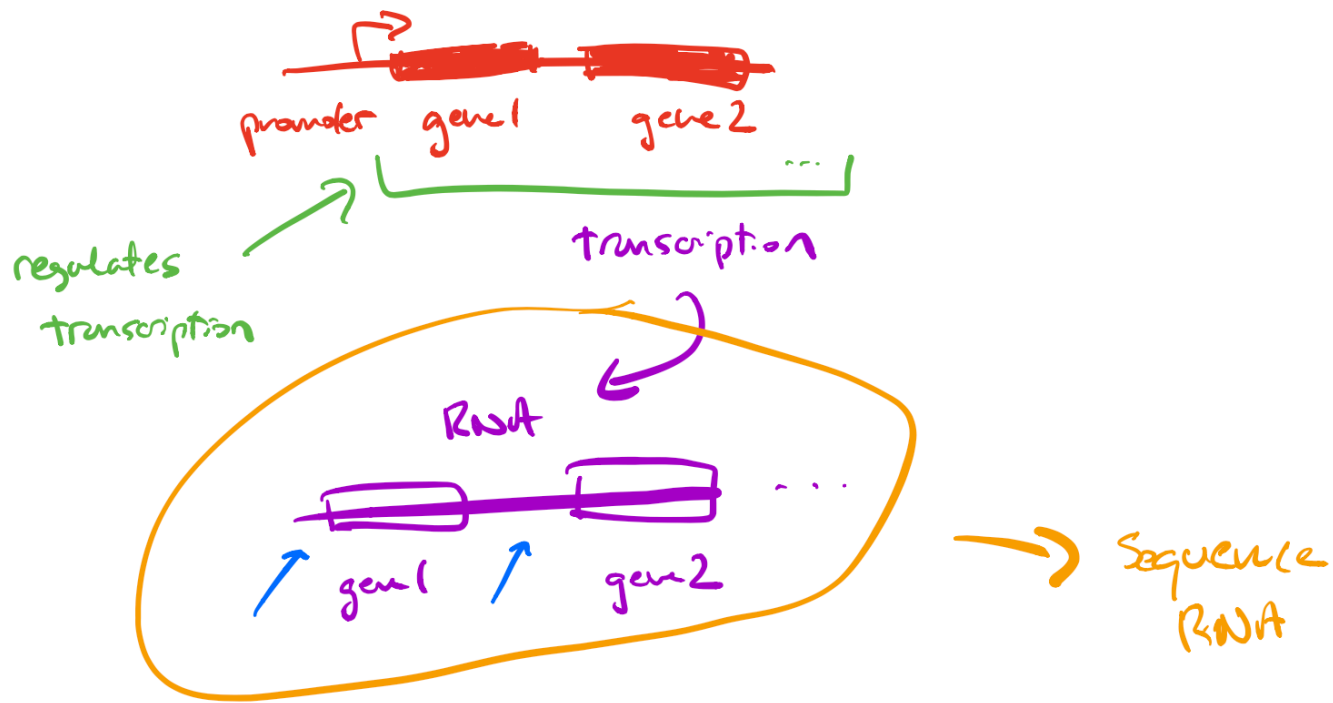
Hypothesis testing vs exploratory analysis.

Are you doing reaching conclusions, or do you plan on doing followup studies to understand your results?

In the first case, you need to make a conclusive statistical argument (replicates, etc.) about the distributions of M&Ms being generated by the factory.

In the second case, you only need to say “hey, it looks like white M&Ms are more abundant in factory 2 – let’s go visit and see if we can understand why.”

(Most statisticians speak about the *former* in papers.)



Making the analogy:

Each M&M color is a gene.

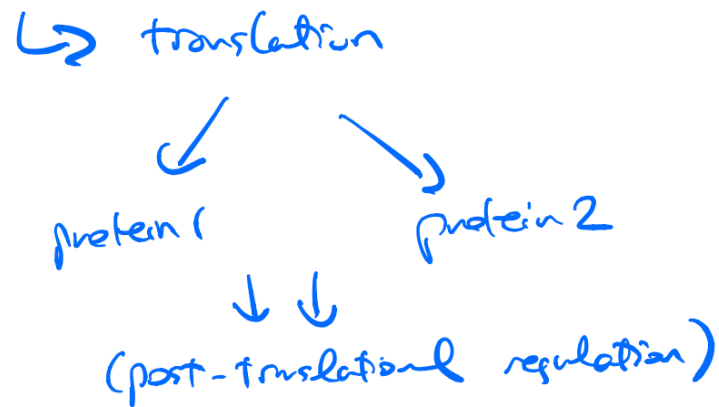
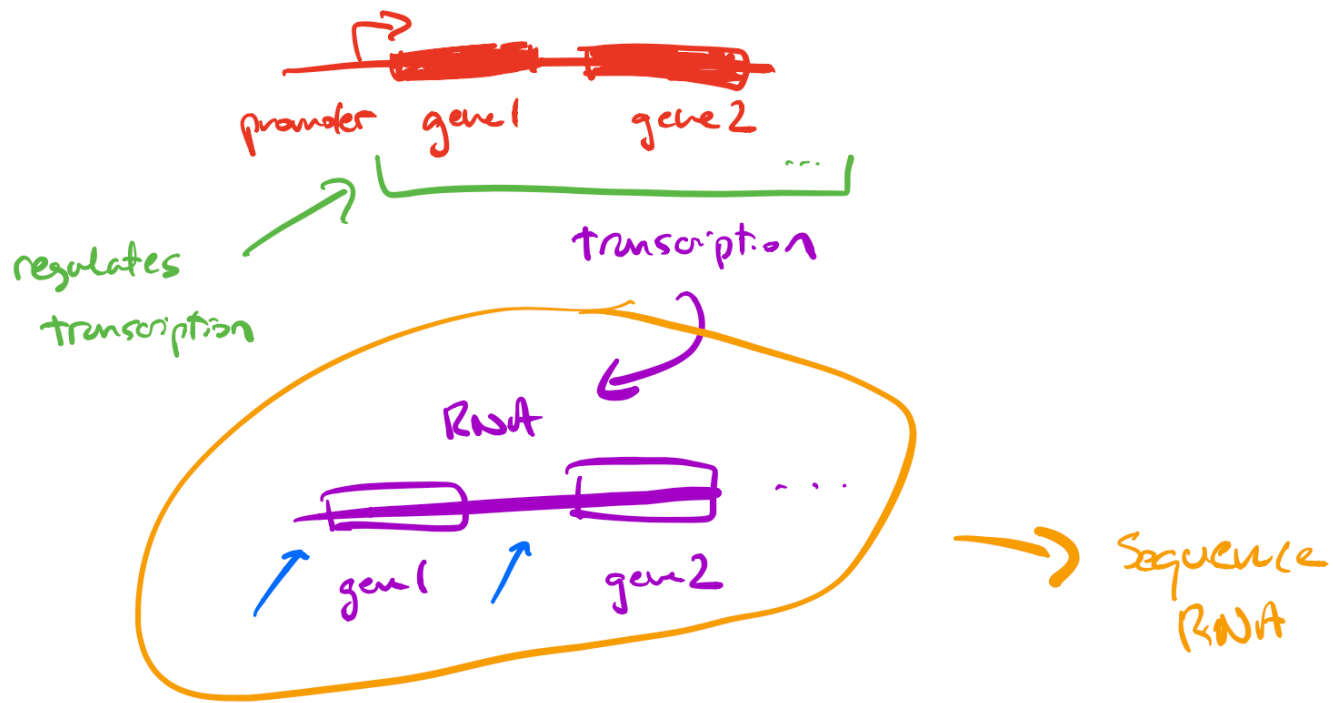
Each M&M count is a gene count.

Each trunk of M&Ms is a collection of cells.

- & each sampling of a trunk is a technical replicate.

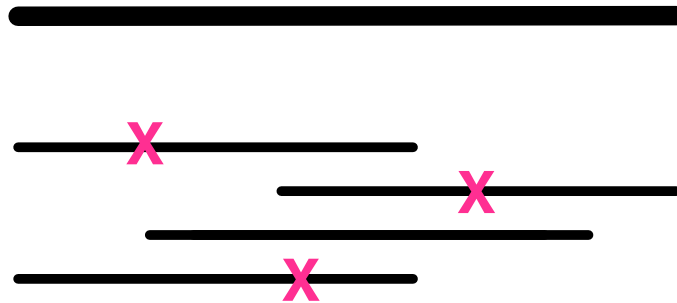
Each factory is a different experiment (biological replicate).

Batch effects: different days; different “hands”; different water; different kits; different sequencing core; ...anything *systematic*.

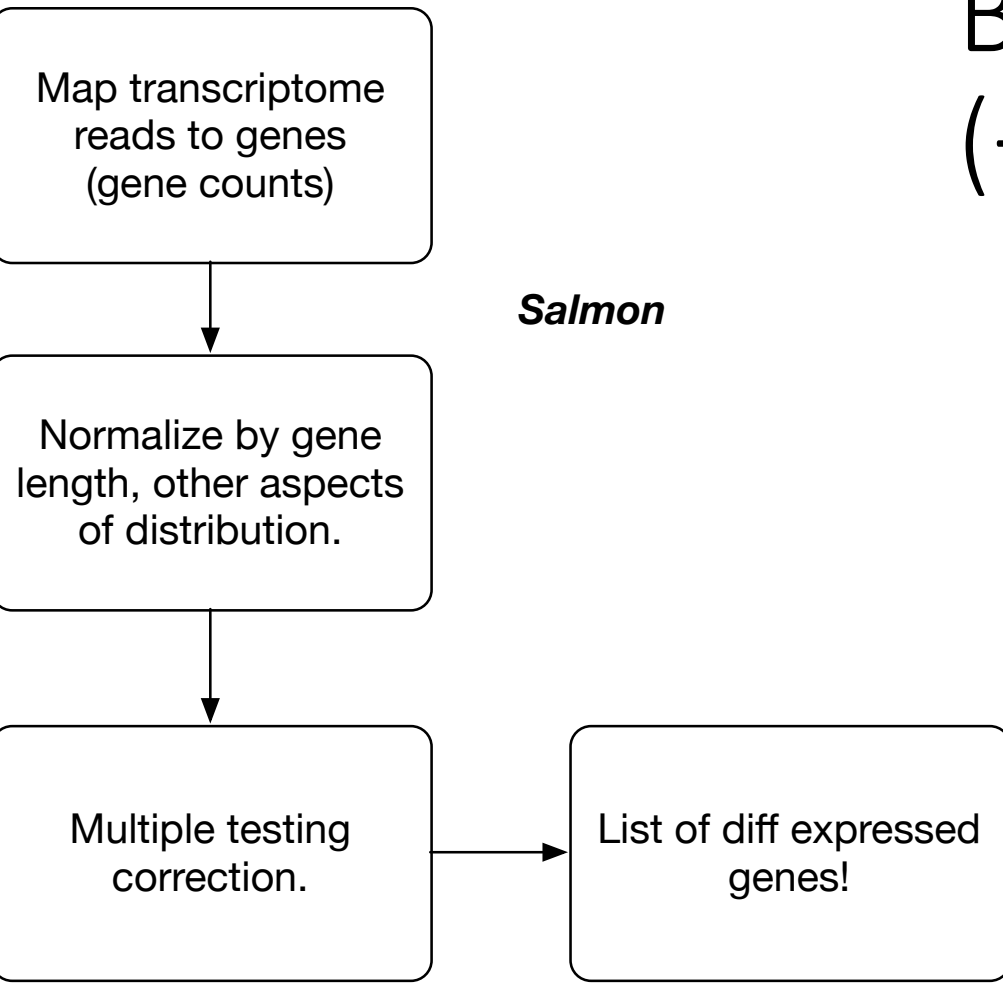


Counting with sequencing

We have a reference gene set and want to know how *much* we have. So, you “map” reads back to reference and count.



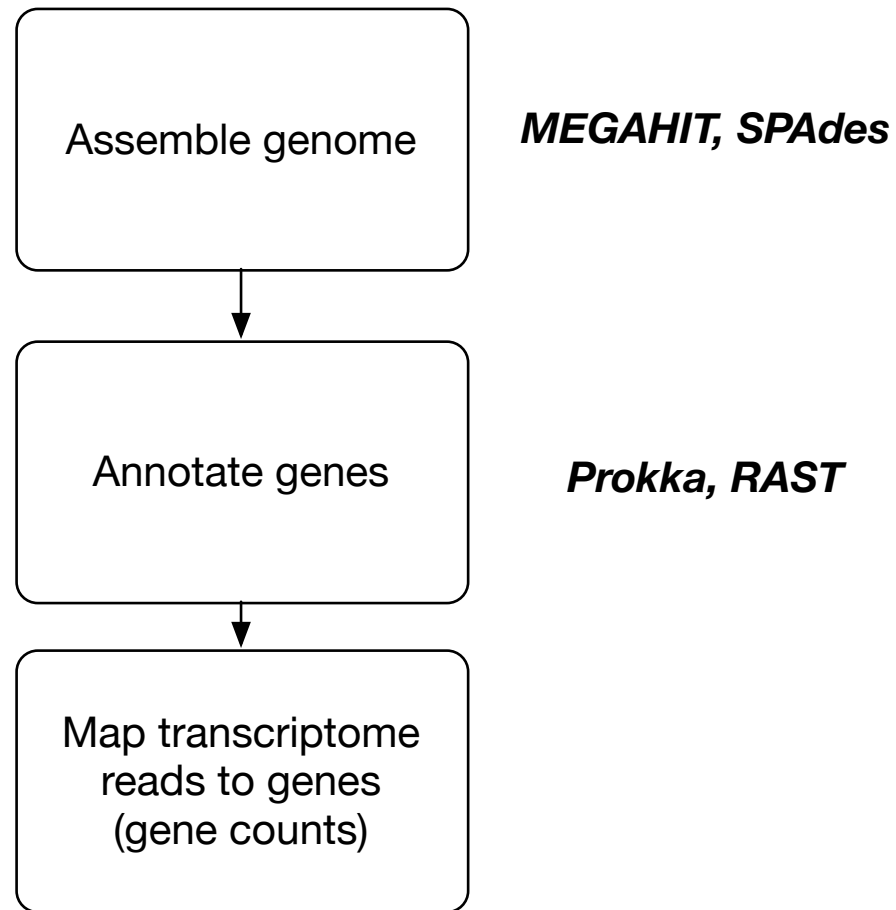
Basic RNAseq procedure (+ tools)



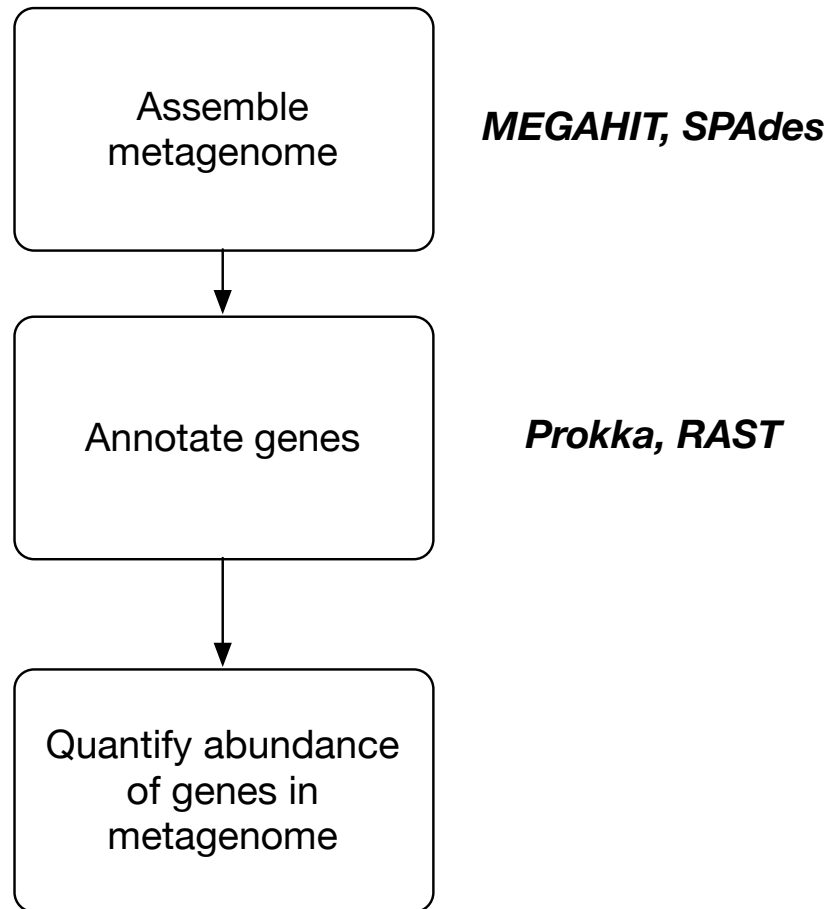
Salmon

DESeq2, Limma/voom, edgeR

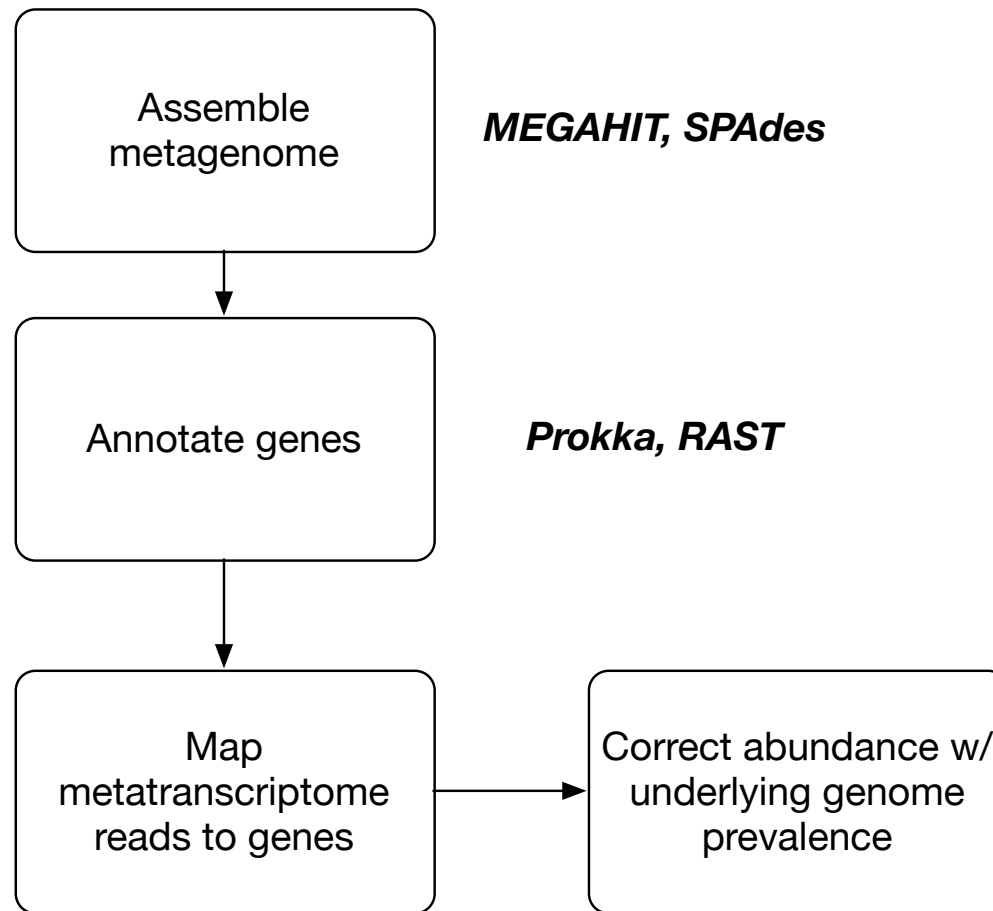
..doing RNAseq on a new genome:



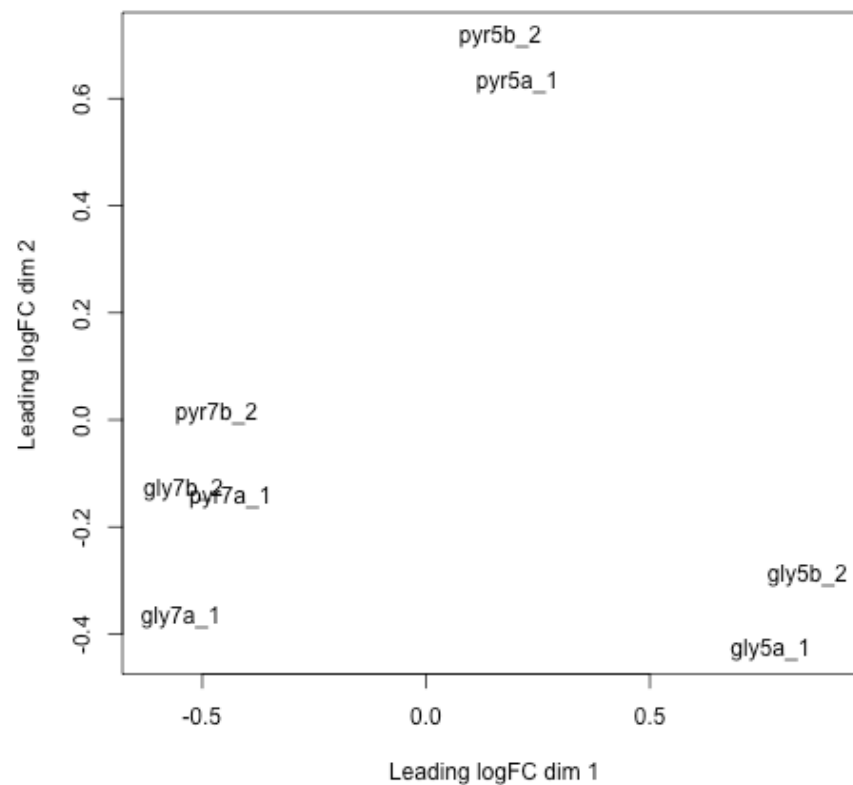
..counting genes in a metagenome:



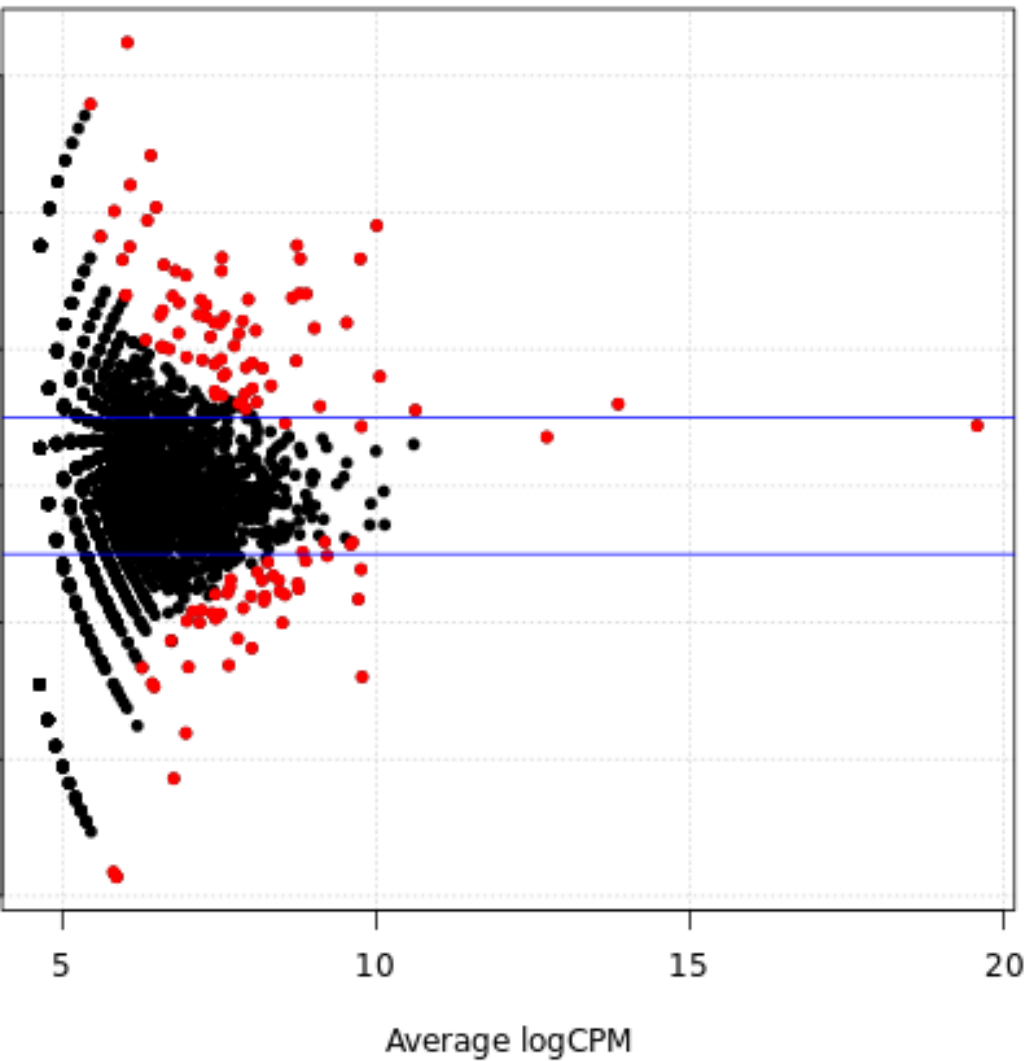
..doing RNAseq on a metatranscriptome:



Examine your sample groupings – do an MDS plot (~PCA, tSNE. etc – “ordination”)



<https://2015-sep-microbial.readthedocs.io/en/latest/diffexp>



Standard plot: “MA p

<https://2015-sep-microbial.readthedocs.io/en/latest/diffexp>

Interpreting differential expression spreadsheets

Pick an FDR cutoff and live with it; don't cherry-pick "low" FDR genes

Don't sort by magnitude of differential expression unless you're pretty sure it matters biologically for the gene(s) you're looking for.

Normalization considerations

Have to take into account many things when comparing:

Usually end up with different read amounts per sample.

Changes in shape of distribution, # of expressed genes, etc.

- Most differential expression toolkits assume less than ~20% of genes changing... probably NOT true in many microbiology situations.

re QPCR – housekeeping genes are less reliable than ALL the genes...

Other thoughts

Unless doing single-cell RNAseq, you are averaging across multiple cells. But we know that gene expression is often punctate (ref DKN).

Assumption that RNA \sim protein abundance which is manifestly not true in at least some situations.

Your computing is one part of the whole workflow – think about your experimental design and your controls :)

References

“How many biological replicates are needed in an RNAseq experiment...?” – Schurch et al., 2016 – PMC 4878611.

An older pipeline:

<https://2015-sep-microbial.readthedocs.io/en/latest/>

A more modern (but not microbial) set of tutorials:

<https://angus.readthedocs.io/en/2017/>

Metagenomics tutorials:

<http://2017-dibsi-metagenomics.readthedocs.io>