RNAseq: Differential Expression

Julin Maloof

Goals

Find genes that have different transcript levels:

- When comparing genotypes (IMB211, R500)
- When comparing treatments (low density (sun) and high density (shade))
- That respond differently to shade in the two genotypes.

Using RNAseq to quantify expression

- In theory: simple.
- Genes expressed at higher levels in the plants should generate more sequence fragments.
- Therefore we can count the number of reads mapping to a particular gene.
    - If IMB211 has more counts than R500 for a particular gene, that may indicate higher expression

Whiteboard…

Challenges

- RNAseq count data is not normally distributed
  - Cannot use Gaussian statistics because this is count data
  - Cannot use Poisson distribution because the data is “over dispersed”
  - Use a negative binomial distribution instead
- What if you have more overall reads in one library as compared to the other?
- What if one gene is expressed so highly in one sample that it dominates the read counts?
  - Need to normalize
- Small number of replicates (~ 3 per sample type), large number of tests (30,000 - 40,000)
  - use Bayesian methods to “borrow” information among genes
  - Also need multiple testing correction

Don’t be tempted by RPKM/FPKM

One common method of RNAseq quantification is Reads (or Fragments) Per Kilobase of gene length per Million reads mapped

Don't use it (at least for statistical analysis)

Whiteboard...

Problems:

- 1 read in a 100bp gene and 10 reads in a 1000bp gene both have the same RPKM. (Why is this a problem?)
- Assumes that an uniform normalization is appropriate

Outline

- Count number of reads in each gene for each sample (RSubread)
- Normalize read counts (EdgeR)
- Determine appropriate statistical models (EdgeR)
- Find differentially expressed genes (EdgeR)
- Examine results

Additive and Interaction effects

Whiteboard...

Model Matrix

In edgeR a `model matrix` is created to describe the possible experimental effects on gene expression.

Our additive model matrix will look like this:

	(Intercept)	**gtR500**	**trtDP**
IMB211_DP_1	1	0	1
IMB211_DP_2	1	0	1
IMB211_DP_3	1	0	1
IMB211_NDP_1	1	0	0
IMB211_NDP_2	1	0	0
IMB211_NDP_3	1	0	0
R500_DP_1	1	1	1
R500_DP_2	1	1	1
R500_DP_3	1	1	1
R500_NDP_1	1	1	0
R500_NDP_2	1	1	0
R500_NDP_3	1	1	0

Model Matrix

The 1s and 0s specify which effects are present in each sample and are used in the statistical model:

$$expression \sim intercept + gt + trt$$

	(Intercept)	**gtR500**	**trtDP**
IMB211_DP_1	1	0	1
IMB211_DP_2	1	0	1
IMB211_DP_3	1	0	1
IMB211_NDP_1	1	0	0
IMB211_NDP_2	1	0	0
IMB211_NDP_3	1	0	0
R500_DP_1	1	1	1
R500_DP_2	1	1	1
R500_DP_3	1	1	1
R500_NDP_1	1	1	0
R500_NDP_2	1	1	0
R500_NDP_3	1	1	0

Model Matrix

The 1s and 0s specify which effects are present in each sample and are used in the statistical model:

$$expression \sim intercept + gt + trt$$

Which is shorthand for

$$expression \sim intercept + gtR500\_Effect * gtR500 + trtDP\_Effect * trtD$$

	(Intercept)	gtR500	trtDP
IMB211_DP_1	1	0	1
IMB211_DP_2	1	0	1
IMB211_DP_3	1	0	1
IMB211_NDP_1	1	0	0
IMB211_NDP_2	1	0	0
IMB211_NDP_3	1	0	0
R500_DP_1	1	1	1
R500_DP_2	1	1	1
R500_DP_3	1	1	1
R500_NDP_1	1	1	0
R500_NDP_2	1	1	0

	(Intercept)	gtR500	trtDP
R500_NDP_3	1	1	0

Testing model terms

To test whether a factor is an important determinant of gene expression:

Compare models with and without that term.

For example, to test if genotype is important, we would compare

(full model) $expression \sim intercept + gt + trt$

to

(reduced model) $expression \sim intercept + trt$

If the full model fits the data significantly better than the reduced model, then we conclude that genotype is important.