Differential expression analysis with limma and SAM

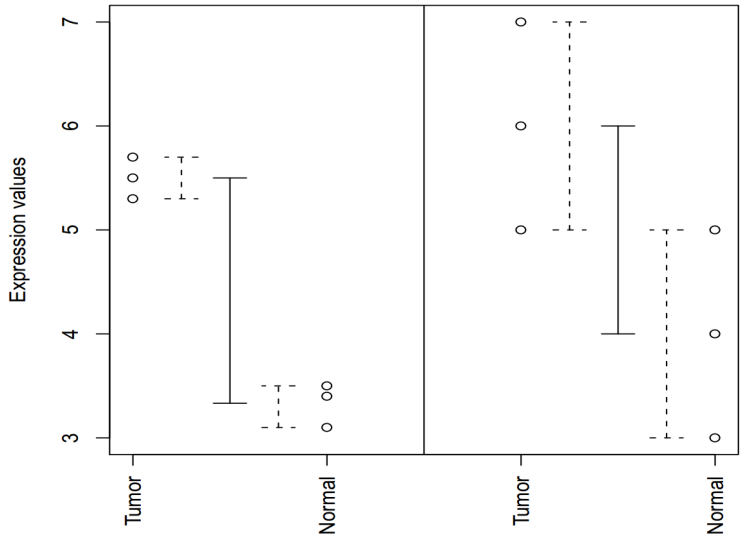
Cory Giles - January 17 - RGCB

Differential expression calculation

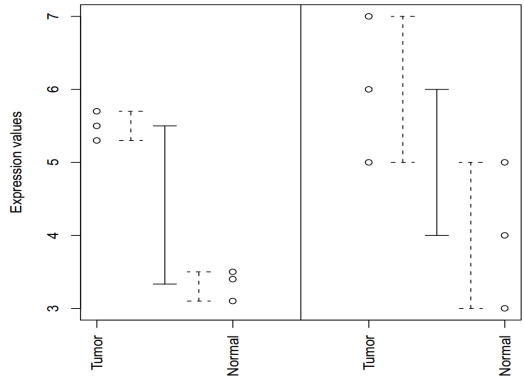
- Identifying differentially expressed genes between two conditions or groups of conditions
- Identifying **significant** changes
- **Many** genes, and many more genes than observations (arrays)
- Multiple hypothesis testing
- Can be cast as a ranking problem or a significance testing problem

Variability and gene expression

Simplest method, fold change, does not take gene variability into account.



Variability and gene expression



T-test:

$$t = \frac{\mu_1 - \mu_2}{s^2 \left(n_1^{-1} + n_2^{-1} \right)}$$

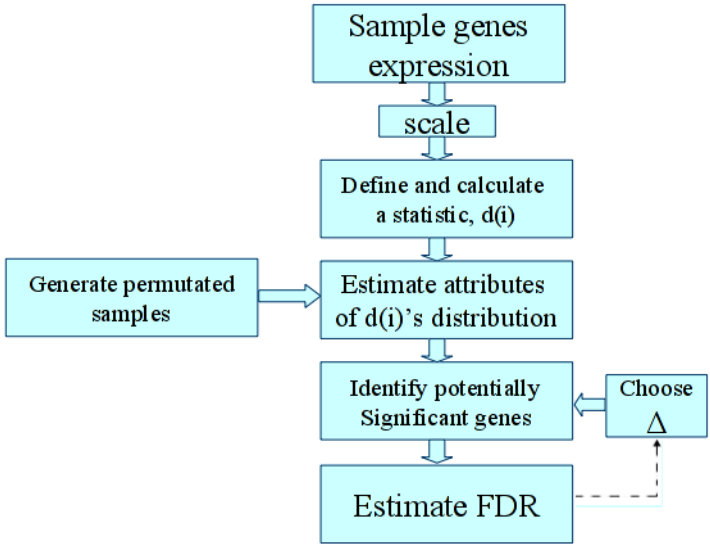
With *pooled* sample variance:

$$s^2 = \frac{\sum^{i}(x_i - \mu_1)^2 + \sum^{j}(x_j - \mu_2)^2}{n_1 + n_2 - 2}$$

T-test problems

- T-test statistic (and p-value) crucially depends on difference in means and variance, but...
- Hard to estimate variance with small sample size
- **No multiple hypothesis testing**
    - Can be done with, e.g, Bonferroni/Holm/Benjamini correction, but with large loss of power

SAM: Significance Analysis of Microarrays



SAM: Significance Analysis of Microarrays

How does SAM improve on T-test?

- **Penalizes low-expressed (unreliable mean and variance) genes:**
  - Adds a constant "exchangeability factor" $s0$ to the denominator of its test statistic
  - $s0$ is the same for all genes

.

$$d_i = \frac{\mu_1 - \mu_2}{s_i + s_0}$$

- d_i - Test statistic of gene i:
- s_i - Pooled standard deviation of gene i

- **Larger $|d_i|$ means stronger (normalize differential expression**

Tusher, Tibshirani, and Chu, PNAS, 2001.

LIMMA - LInear Models for Microarray Analysis

- Fits a linear model to each gene.
- "Borrows" information about variability across genes using empirical Bayes methods.

Requires from the user:

- The expression matrix
- "Design" matrix, which summarizes the experimental design (different treatments or combinations of treatments)

- "Contrast" matrix - identifies the "contrast" of interest (case vs control, B cell vs T cell, etc.)

For a given analysis, 1 design matrix, possibly multiple contrast matrices for different biological questions.

BAD ways to calculate DE

- **Order by FC or FC cutoff**
  - doesn't take variance into account
- **T-test**
  - **Estimates gene variance for each gene individually**
    - With small sample sizes, a high probability that variance will be seriously underestimated for some genes

- Prone to false positives on genes with low variance
- Low "power"

Modern approaches to DE calculation

Homoscedastic methods assume that each treatment group has the same variance:

- ANOVA (not recommended), RVM, limma, VarMixt

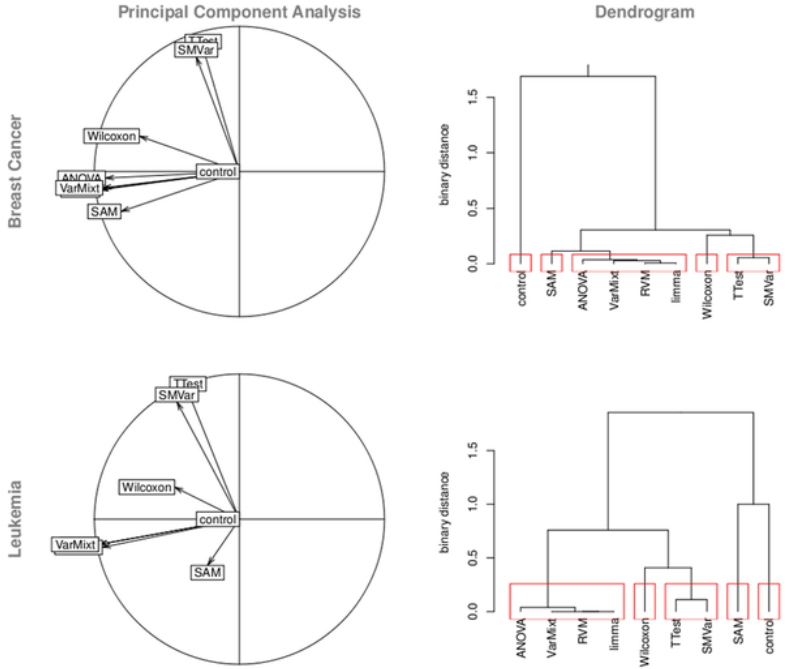
Heteroscedastic methods do not make this assumption (and must estimate the variance for each group):

- Welch t-test, SMVar

Nonparametric methods do not assume any particular probability distribution:

- Significance analysis of microarrays (SAM), Wilcoxon rank-sum

Similar assumptions -> similar results



Jeanmougin et al, 2010, PloS One.