Differential expression

- Identify genes whose expression levels are associated with a response or covariate of interest
  - clinical outcome such as survival, response to treatment, tumor class;
  - covariate such as treatment, dose, time.
- Estimation: estimate effects of interest and variability of these estimates.

  E.g. Slope, interaction, or difference in means.
- Testing: assess the statistical significance of the observed associations.

Multiple hypothesis testing

- Define an appropriate Type I error or false positive rate.
- Apply multiple testing procedures that
  - control this error rate under the true unknown data generating distribution,
  - are powerful (few false negatives),
  - take into account the joint distribution of the test statistics.
- Report adjusted p-values for each gene which reflect the overall Type I error rate for the experiment.
- Use resampling methods to deal with the unknown joint distribution of the test statistics.

multtest package

- Multiple testing procedures for controlling
  - Family-Wise Error Rate (FWER): Bonferroni, Holm (1979), Hochberg (1986), Westfall & Young (1993) maxT and minP;
  - False Discovery Rate (FDR): Benjamini & Hochberg (1995), Benjamini & Yekutieli (2001).
- Tests based on t- or F-statistics for one- and two-factor designs.
- Permutation procedures for estimating adjusted p-values.
- Fast permutation algorithm for minP adjusted p-values.
- Documentation: tutorial on multiple testing.

Data Reduction in Microarray Experiments

Images

↓

Intensities (normalization)

↓

Expression measures (normalization)

↓

Score

↓

Choose a cut off: report a list of differentially expressed genes and error rate

Differential gene expression

- Identify genes whose expression levels are associated with a response or covariate of interest
  - clinical outcome such as survival, response to treatment, tumor class;
  - covariate such as treatment, dose, time.
- Estimation: In a statistical framework, assigning a score can be viewed as estimating an effects of interest (e.g. difference in means, slope, interaction). We can also take the variability of these estimates into account.
- Testing: In a statistical framework, deciding on a cut-off can be viewed as an assessment of the statistical significance of the observed associations.

Example: Two populations

A common problem is to find genes that are differentially expressed in two populations.

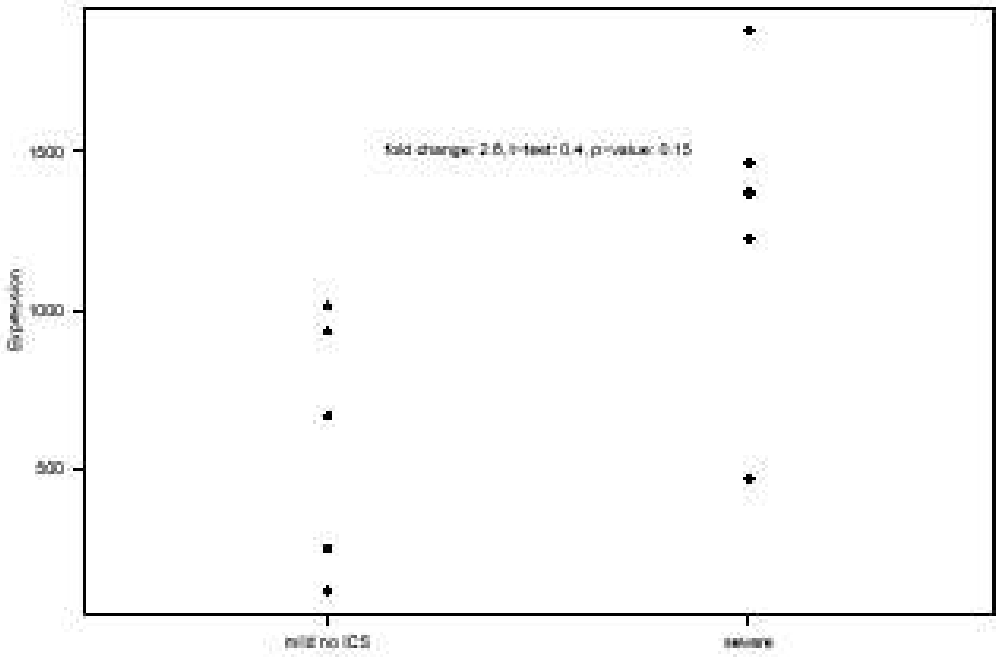
Many method papers appear in both statistical and molecular biology literature.

The proposed scores range from:

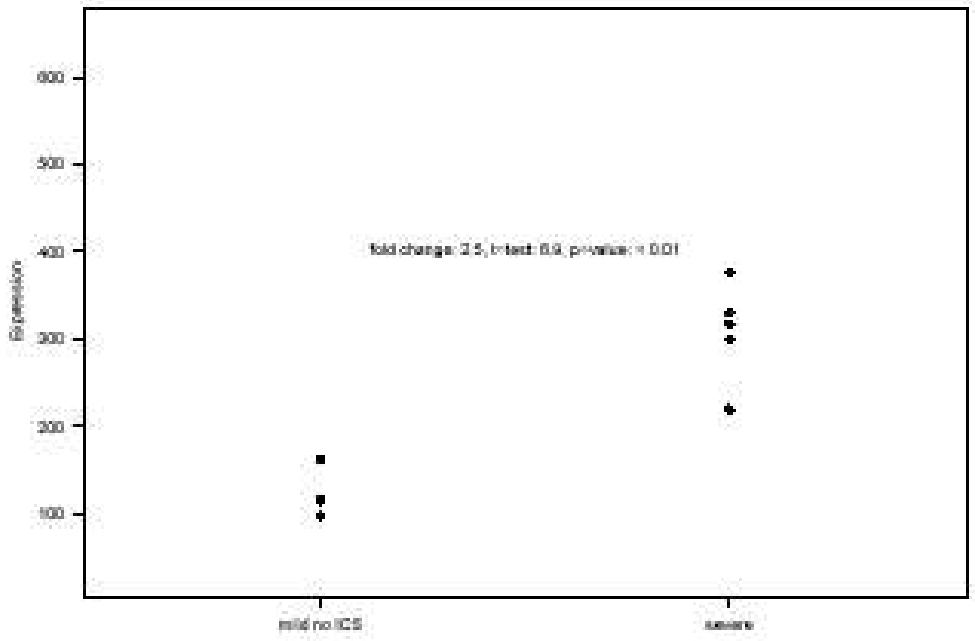
- ad-hoc summaries of fold-change,
- variantes on the t-test,
- and posterior means obtained from Bayesian or empircal Bayes methods.

What's the difference? Mainly the way in which the variation within population is incorporated

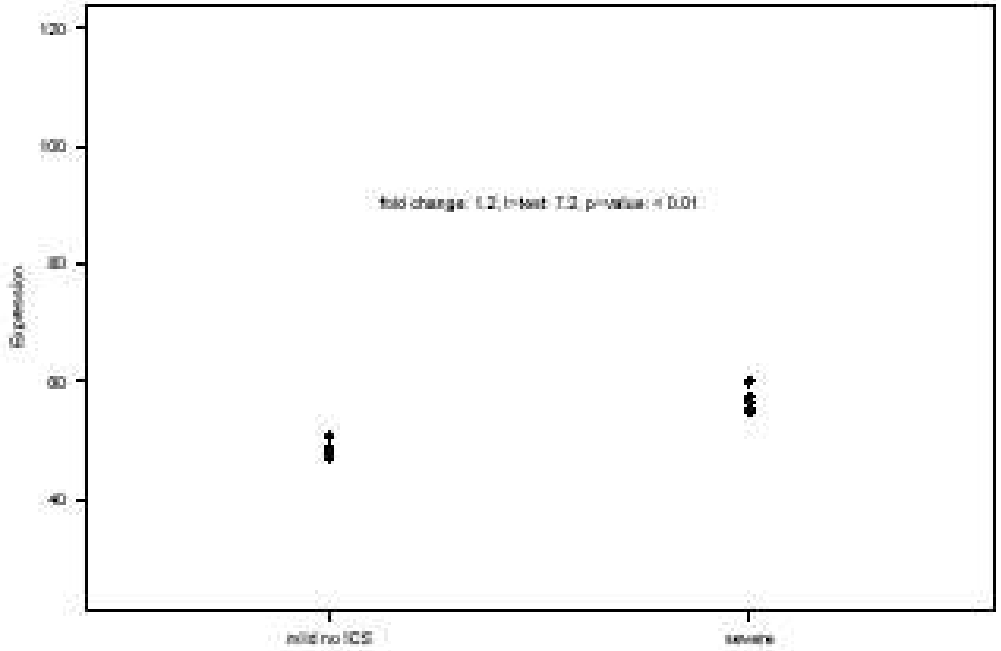
Should we consider variability of estimate?



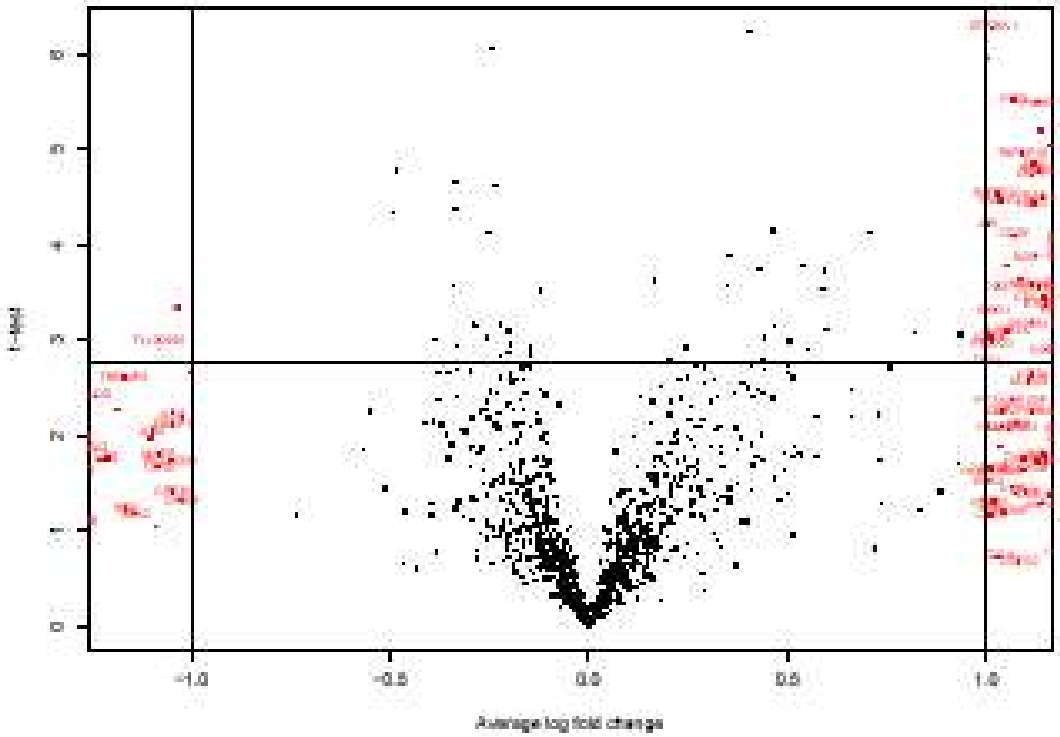
Should we consider variability of estimate?



Should we consider variability of estimate?



Volcano Plots



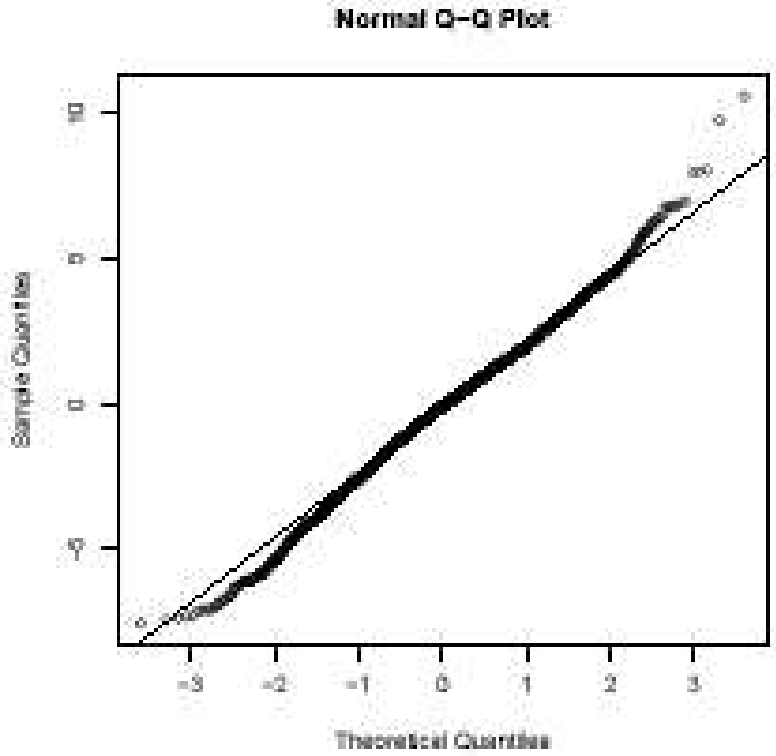


Figure 1: Normal Q–Q plot of t–statistics for leukemia data.

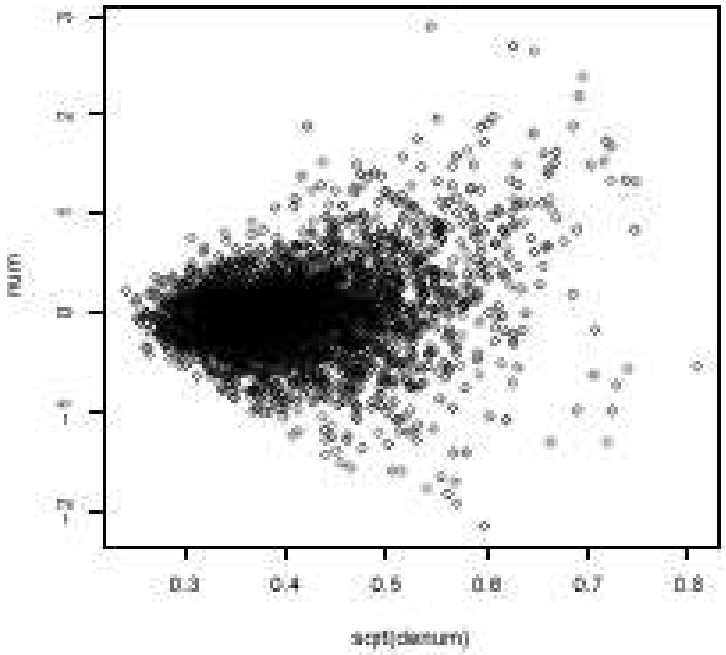


Figure 2: Numerator vs. square root of denominator of the t-statistics for the leukemia data.

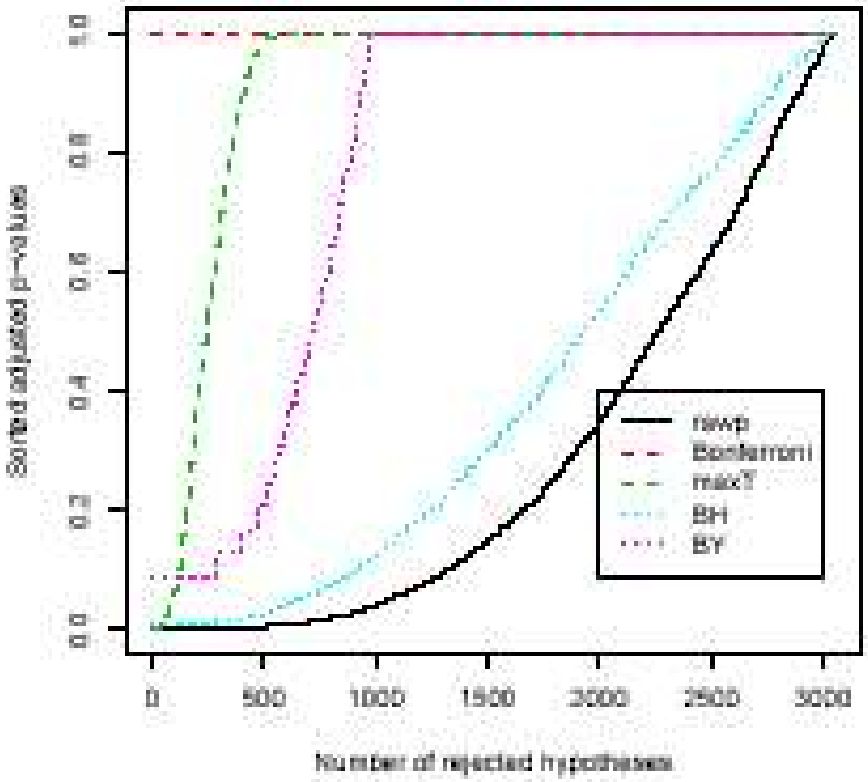


Figure 3: Sorted adjusted p–values for the leukemia data.

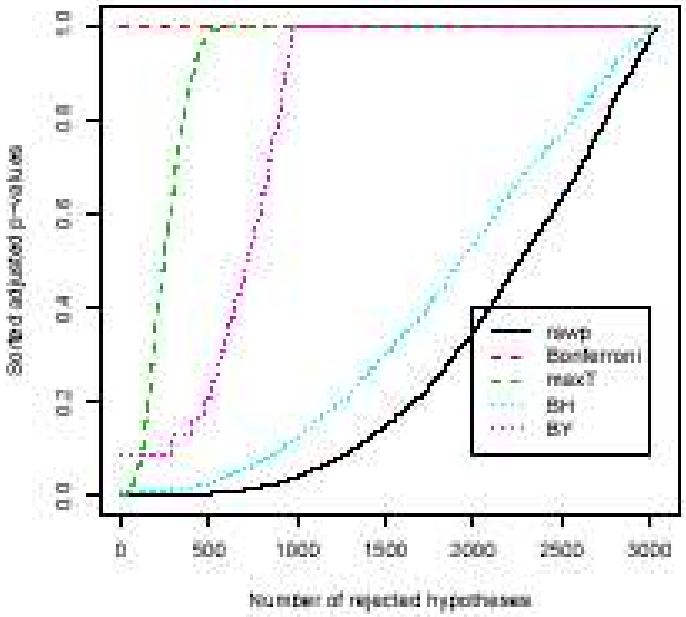


Figure 3: Sorted adjusted p–values for the leukemia data.