Methods for Analyzing Gene Expression Data

Mikhail Dozmorov

2021-03-15

Taxonomy of Data Analysis Methods

Supervised Learning

- Class or group labels are known a priori and the goal of the statistical analysis pertains to identifying differentially expressed genes (AKA feature selection) or identifying combinations of genes that are predictive of class or group membership.

Taxonomy of Data Analysis Methods

Unsupervised Learning

- The statistical analysis seeks to find structure in the data without knowledge of class labels.

Supervised Learning

- **Class comparison/ Feature selection**
  - T-test / Wilcoxon rank sum test
  - F-test / Kruskal-Wallis test
  - Adjustment for multiple comparisons
- **Class Prediction**
  - K nearest neighbors
  - Compound Covariate Predictors
  - Classification trees
  - Support vector machines
  - etc.

Hypothesis testing

- The hypothesis that two means μ_1 and μ_2 are equal is called a null hypothesis, commonly abbreviated H_0.
- This is typically written as $H_0 : \mu_1 = \mu_2$
- Its antithesis is the alternative hypothesis, $H_A : \mu_1 \neq \mu_2$

Hypothesis testing

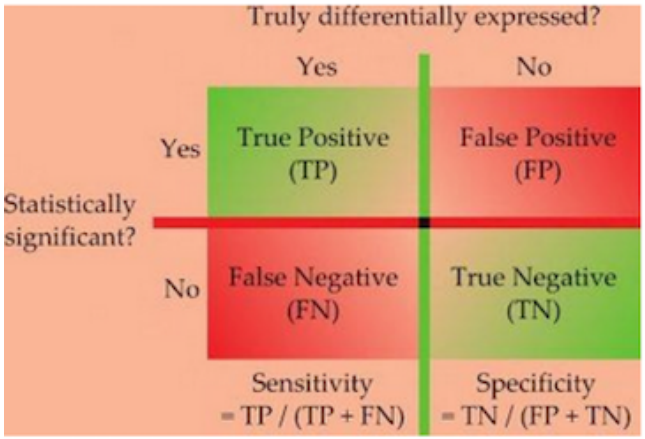
- A statistical test of hypothesis is a procedure for assessing the compatibility of the data with the null hypothesis.
  - The data are considered compatible with H_0 if any discrepancy from H_0 could readily be due to chance (i.e., sampling error).
  - Data judged to be incompatible with H_0 are taken as evidence in favor of H_A.

Hypothesis testing

- If the sample means calculated are identical, we would suspect the null hypothesis is true.
- Even if the null hypothesis is true, we do not really expect the sample means to be identically equal because of sampling variability.
- We would feel comfortable concluding H_0 is true if the chance difference in the sample means should not exceed a couple of standard errors.

Hypothesis Testing

- **Type I error**: The probability of rejecting a null hypothesis when it is true. (e.g., a gene is declared to be differentially expressed when it is not.)
- **Type II error**: The probability of accepting a null hypothesis when it is false. (e.g., a gene is declared to not be differentially expressed when it actually is.)



P-value

- The p-value for a hypothesis test is the probability, computed under the condition that the null hypothesis is true, of the test statistic being at least as extreme as the value of the test statistic that was actually obtained.
  - A large p-value (close to 1) indicates a value of t near the center of the t-distribution.
  - A small p-value indicates a value of t in the far tails of the t-distribution.

Hypothesis testing

- The **mean** μ_X of a random variable X is a measure of central location of the density of X.
- The **variance** of a random variable is a measure of spread or dispersion of the density of X.
- $Var(X) = E[(X - \mu)^2] = \sum \frac{(X-\mu)^2}{(n-1)} = \sigma^2$
- Standard deviation $= \sqrt{Var(X)} = \sigma$

Two-sample comparison

Let us consider the simplest case: two-sample comparison. Our goal is to find the list of genes that are differentially expressed. Suppose we have:

- n_1 samples in group 1
- n_2 samples in group 2
- For each gene, $n_1 + n_2$ expression levels are recorded for all the samples

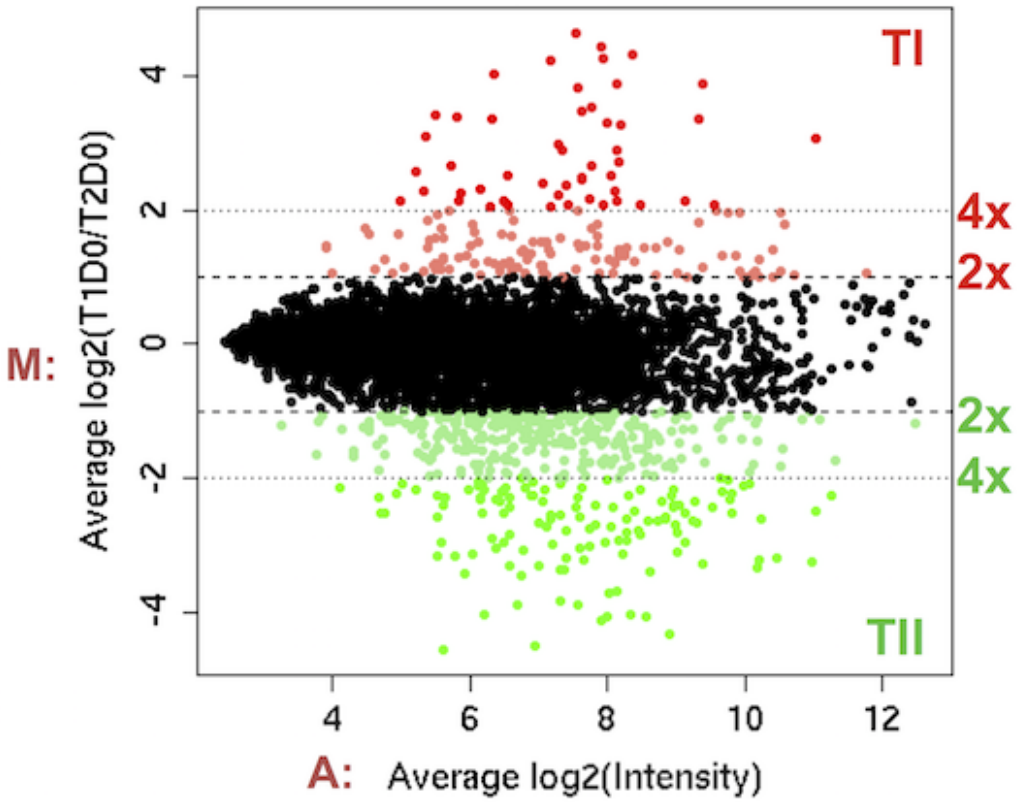
Determine which genes have differential expression between the two groups of samples.

Differential expression

- Many experiments are carried out to find genes which are **differentially expressed between two (or more) samples**.
- Initially, comparative experiments were done with few, if any, replicates, and statistical criteria were not used for identifying differentially expressed genes. Instead, simple criteria were used such as fold-change, with 2-fold being a popular cut-off.

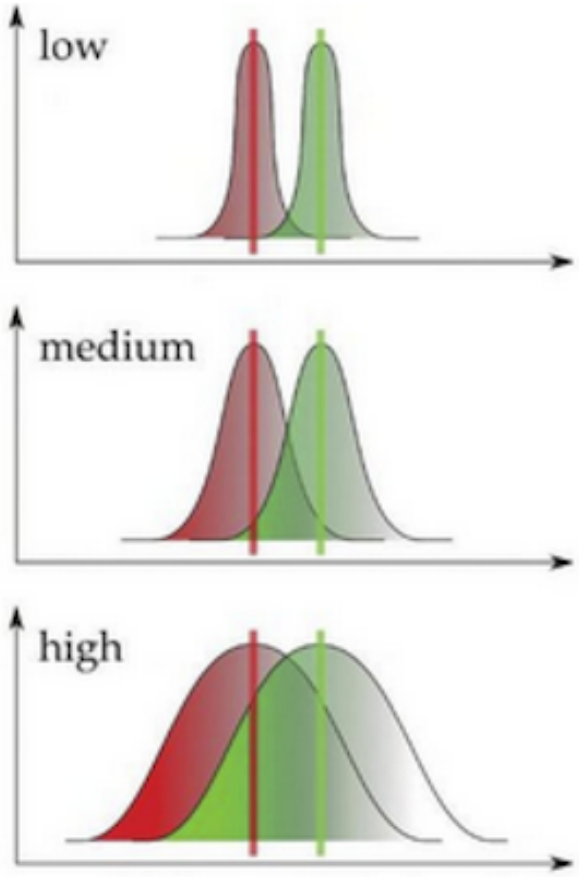
DE by Average Fold-Change (M)

- Simple fold-change rules give no assessment of statistical significance.
- Need to construct test statistics incorporating variability estimates (from replicates).



Variability and gene expression

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



Two-sample comparison, T-test

Let the mean and standard deviation expression levels for samples in two groups be

$\bar{x_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$, and $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})^2$

The two-sample pooled t-statistics is given by

$$t = \frac{\bar{x_2} - \bar{x_1}}{s_p \sqrt{1/n_1 + 1/n_2}}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

is the pooled estimate of the standard deviation.

T-test assumptions

- Data must be independent random samples from their respective populations
- Sample size should either be large or, in the case of small sample sizes, the population distributions must be approximately normally distributed.
- When assumptions are not met, non-parametric alternatives are available (Wilcoxon Rank Sum/Mann-Whitney Test)

Welch's t-test

Does not assume equal variances for each group

$$t_g^{Welch} = \frac{\bar{y_{g}}1. - \bar{y_{g}}2.}{\sqrt{\frac{S_{g1}^2}{n_1} + \frac{S_{g2}^2}{n_2}}}$$

The variances S_{g1}^2 and S_{g2}^2 are then estimated independently in both groups for each gene

When there are few replicates...

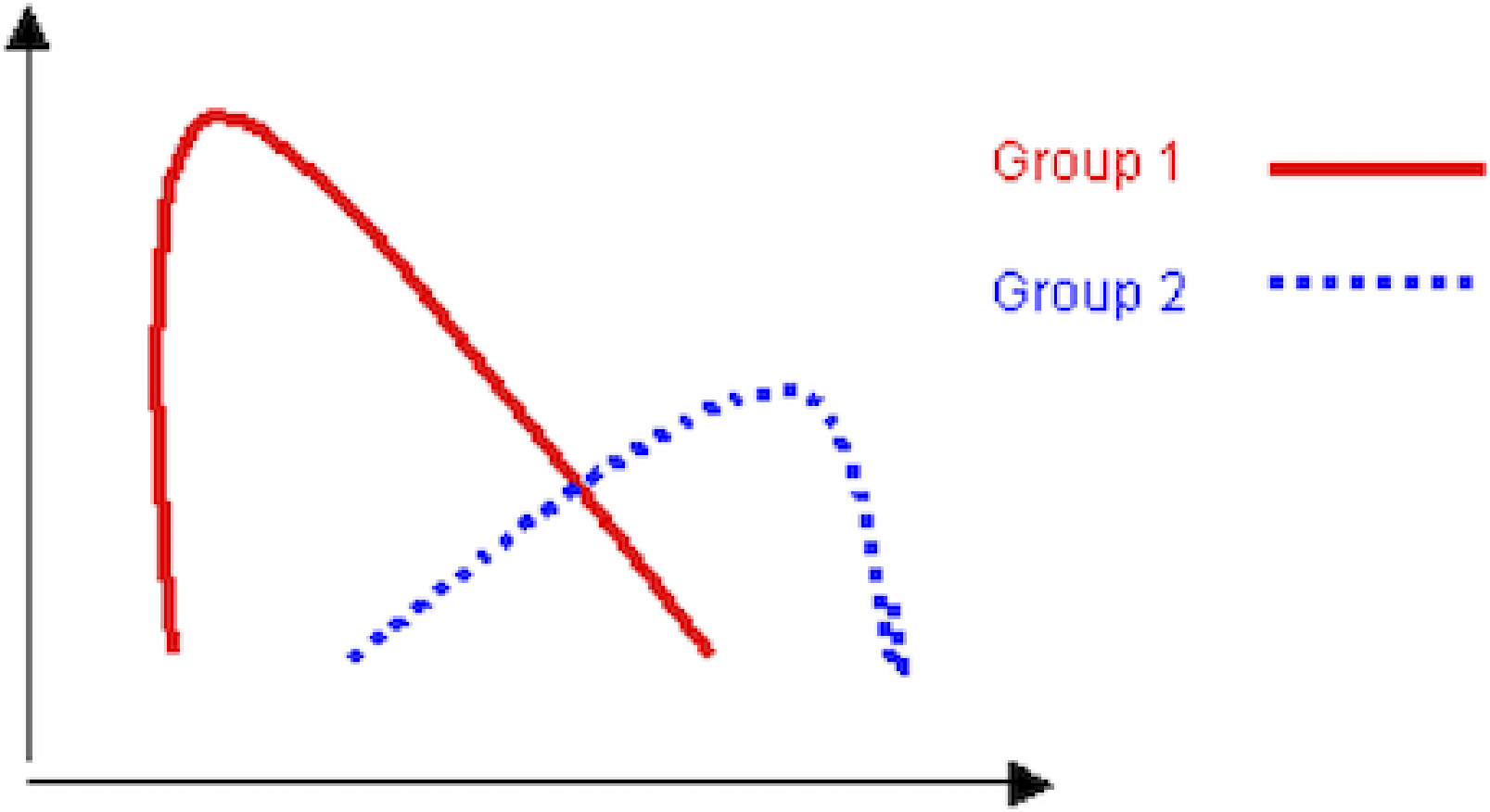
- Fold change using averages $\bar{M}$ can be driven by **outliers**
- T-statistics $\frac{\bar{M}}{se(\bar{M})}$ can be driven by **tiny variances**
- Solution: "robust" version of t-statistic
- Replace mean by **median**
- Replace standard deviation by **median absolute deviation**

Non-parametric tests

- Non-normally distributed data
- More robust to outliers
- Less power
- Used when t-test assumptions cannot be met

Non-parametric tests

- **Mann-Whitney test (or Wilcoxon rank-sum test)**
  - differences in the sums of ranks between 2 populations
  - even if the medians are the same, there can be a statistically significant difference from the distribution of ranks



Anova: Analysis of Variance

Doing multiple two-sample t-tests would result in an increased chance of committing a Type I error.

For this reason, ANOVAs are useful in comparing (testing) three or more means (groups or variables) for statistical significance.

Classes of ANOVA models

1. **Fixed-effects model**: a statistical model that represents observed quantities as non-random
2. **Random-effects model**: used when the treatments are not fixed
3. **Mixed model**: contains both fixed and random effects

Anova: Analysis of Variance

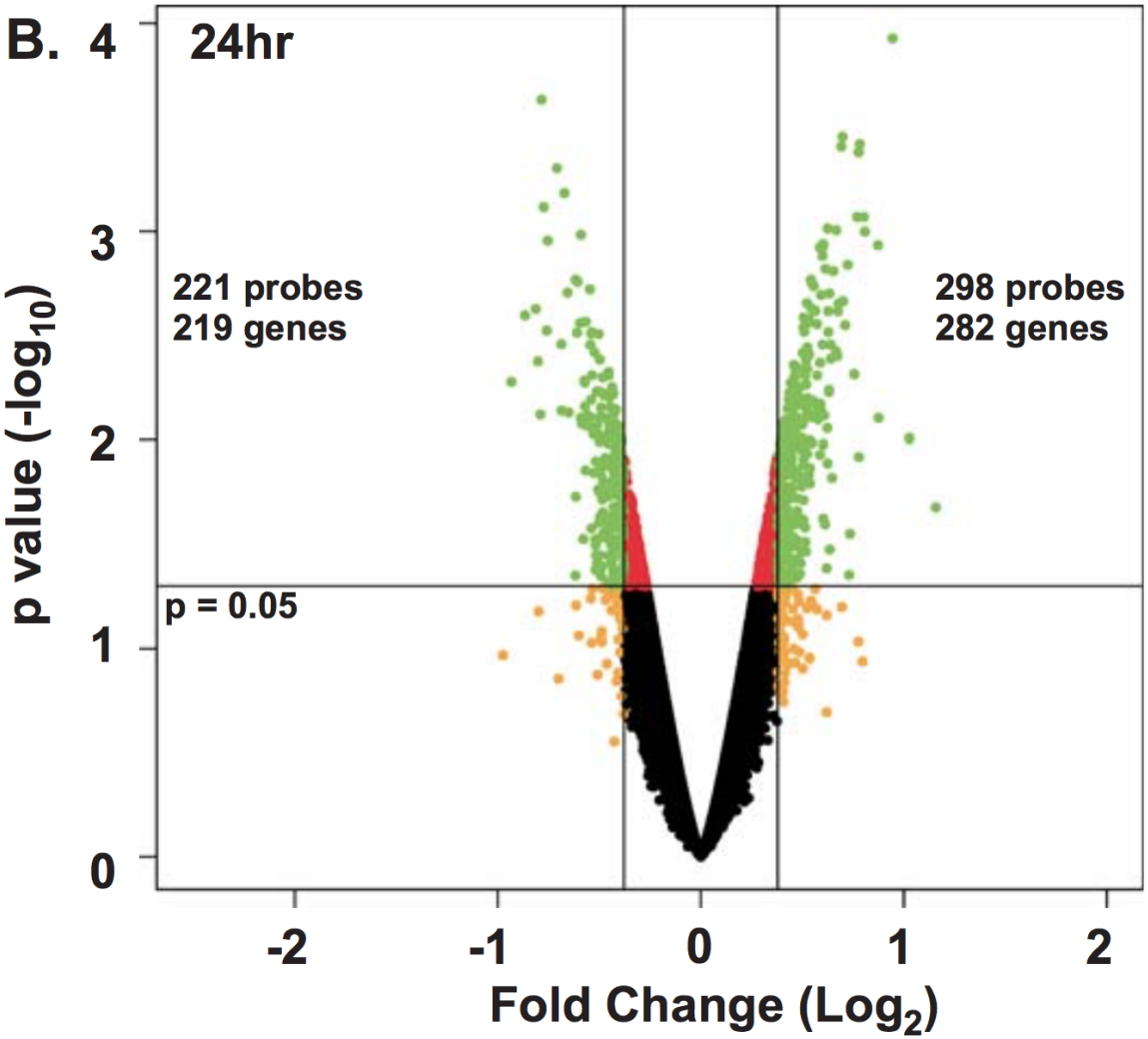
Common Designs and Tests

- **One-way ANOVA** is used to test for differences among two or more independent groups (means). When there are only two means to compare, the t-test and the ANOVA F-test are equivalent; the relation between ANOVA and t is given by $F = t^2$.
- **Factorial ANOVA** is used when the experimenter wants to study the interaction effects among the treatments.
- **Repeated measures ANOVA** is used when the same subjects are used for each treatment (e.g., in a longitudinal study).

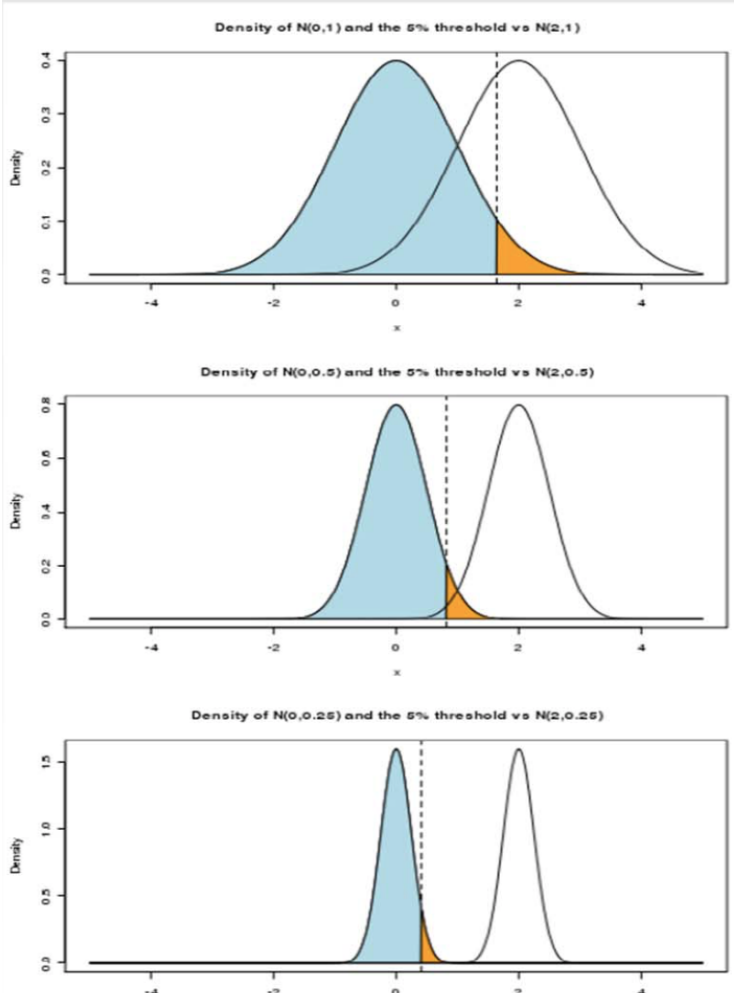
Volcano plot

- A diagnostic plot to visualize the test results.
- Scatter plot of the statistical significance (log p-values) vs. biological significance (log fold change).
- Ideally the two should agree with each other.

Volcano plot



Sample size and power calculations



- Effects of changing: variance, and mean difference (or effect size).
  Given a fixed sample size, n.
  - Larger variance $\rightarrow$ Lower Power
  - Smaller effect size $\rightarrow$ Lower Power
- *Given variance, effect size.*
  Increase sample size $\rightarrow$ Increase Power
- Can also change power by manipulating the "trade-off" between Type I and II error
  - Larger $\alpha \rightarrow$ Larger Power
  - Larger $\beta \rightarrow$ Lower Power