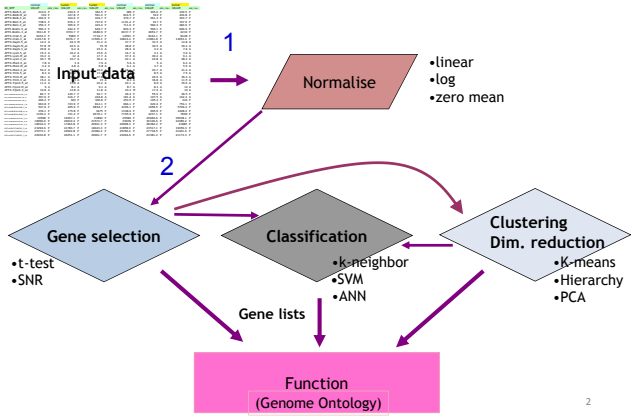
COSC 348: Computing for Bioinformatics

Lecture 17:
Microarray Data Analysis: gene selection

Lubica Benuskova

http://www.cs.otago.ac.nz/cosc348/

Streamlined microarray analysis



Differential gene expression

- The Experiment measures gene expression in rats:
  - Two groups: (WT: wild-type rat, KO: gene knock-out rat)
  - Question: Which genes are affected by the treatment? How significant is the effect? **We must compare each pair of genes**.

ID	WT_1_R	WT_2_R	WT_3_R	WT_4_R	KO_1_R	KO_2_R	KO_3_R	KO_4_R
93173_at	242.3	240.1	292.9	216.3	180.1	172.6	147.3	152.4
101937_s_at	316.7	346.7	438.3	228.5	133.7	201.3	253.3	287.4
104272_s_at	286.2	351.9	354.6	339.1	180.6	432.7	210.2	53.6
98590_at	1,066	748.8	1,011.4	607.7	584.5	791.8	355.8	530
102425_at	264.7	241.4	450	134.3	138.3	242	212.6	125.4
96608_at	1,979.8	1,913.2	2,367	1,616	1,270.5	1,191.6	1,401.2	1,330.9
94407_at	339.3	360.4	283	309.1	236.9	329.3	196.8	89.4
161149_r_at	1,947.7	1,179.4	1,708	1,251	1,297.1	594.3	1,070.5	1,055.8
100144_at	4,821.6	3,639.6	4,415.5	3,846	3,268.5	2,438.5	2,799	2,537.4
95134_at	498.6	853.1	881.2	582.8	255.1	859.3	288.7	457.8
96921_at	746.1	410.6	858.8	667.4	534.8	444	475.4	320.3
94689_at	534	438	456.2	555.2	466.6	404.3	295.2	146.4
160268_at	737.7	1,099.2	1,138.4	978.8	806.5	978	587.8	245.3
96180_at	609.5	516.9	540.1	312.8	344.8	191.8	427.9	347.1
92618_at	4,888.8	4,234.2	4,703.7	2,994.9	4,093.1	2,938.9	2,150.2	1,969.2
93203_f_at	111.8	186.8	112.9	158.1	100.8	67	119.9	90.6
102574_at	171.3	81.7	230.9	123.3	107.9	50.6	112.3	132.4
160966_at	221.2	310	454.3	242.5	238	196.2	330.7	50.8
160827_at	294.5	341.1	360.4	170.3	231.6	289.4	196.4	58.1
104116_at	1,836.3	829.3	1,258.7	1,561	722.3	810.4	943.9	1,172.1
95434_at	1,207.8	1,294.8	1,314.6	1,513.8	878.2	773.9	715.8	1,181.5

We need multiple samples

- In order to determine whether a gene has undergone differential expression between two (or more) conditions, multiple observations in each condition are required (i.e multiple rats, patients, etc.).
- That is, many samples with the same condition must be measured because there are individual variations in gene expressions and also experimental errors in producing and processing microarrays.
- The task is to distinguish whether the variation in gene expression between two (or more) conditions is due to the condition itself or due to a natural variation among subjects in the same group or due to the experimental errors.

Statistics: what difference is significant?

- For gene k we have two vectors of expression values
  - Condition 1: $X_k = \{x_{k1}\, x_{k2}\, x_{k3} \ldots x_{ki} \ldots x_{kn(x)}\}^T$
  - Condition 2: $Y_k = \{y_{k1}\, y_{k2}\, y_{k3} \ldots y_{ki} \ldots y_{kn(y)}\}^T$
- Question 1: if we see a difference, are we actually observing differential expression, or is it due to something else (individual variation and/or experimental error)?
- Question 2: how big a change do we need to see for us to think we are observing differential expression? (i.e., what counts as significant differential expression?)

Gene selection based on statistical significance

ID	WT_1_R	WT_2_R	WT_3_R	WT_4_R	KO_1_R	KO_2_R	KO_3_R	KO_4_R
93173_at	242.3	240.1	292.9	216.3	180.1	172.6	147.3	152.4
101937_s_at	316.7	346.7	438.3	228.5	133.7	201.3	253.3	287.4
104272_s_at	286.2	351.9	354.6	339.1	180.6	432.7	210.2	53.6
98590_at	1,066	748.8	1,011.4	607.7	584.5	791.8	355.8	530
102425_at	264.7	241.4	450	134.3	138.3	242	212.6	125.4
96608_at	1,979.8	1,913.2	2,367	1,616	1,270.5	1,191.6	1,401.2	1,330.9
94407_at	339.3	360.4	283	309.1	236.9	329.3	196.8	89.4
161149_r_at	1,947.7	1,179.4	1,708	1,251	1,297.1	594.3	1,070.5	1,055.8
100144_at	4,821.6	3,639.6	4,415.5	3,846	3,268.5	2,438.5	2,799	2,537.4
95134_at	498.6	853.1	881.2	582.8	255.1	859.3	288.7	457.8
96921_at	746.1	410.6	858.8	667.4	534.8	444	475.4	320.3
94689_at	534	438	456.2	555.2	466.6	404.3	295.2	146.4
160268_at	737.7	1,099.2	1,138.4	978.8	806.5	978	587.8	245.3
96180_at	609.5	516.9	540.1	312.8	344.8	191.8	427.9	347.1
92618_at	4,888.8	4,234.2	4,703.7	2,994.9	4,093.1	2,938.9	2,150.2	1,969.2
93203_f_at	111.8	186.8	112.9	158.1	100.8	67	119.9	90.6
102574_at	171.3	81.7	230.9	123.3	107.9	50.6	112.3	132.4
160966_at	221.2	310	454.3	242.5	238	196.2	330.7	50.8
160827_at	294.5	341.1	360.4	170.3	231.6	289.4	196.4	58.1
104116_at	1,836.3	829.3	1,258.7	1,561	722.3	810.4	943.9	1,172.1
95434_at	1,207.8	1,294.8	1,314.6	1,513.8	878.2	773.9	715.8	1,181.5

For each gene compare the **average** value of the expression between two conditions.

For each gene calculate the significance of the difference – are the averages the same or different?

Identify and select genes with high statistical significance for further analysis.

Population and sample

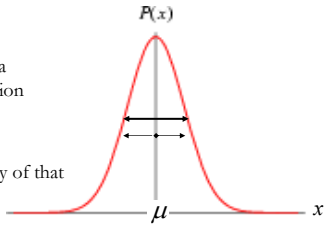
- The basic idea of statistics is this: we want to extrapolate from the data we have collected to make general conclusions about everybody.
- There is a large population of data out there, and we have **randomly** sampled parts of it. Random: each unit has an equal chance to be selected. We analyze our *sample* to make inferences about the *population*.
- Clinical studies
  - **Sample**: Subset of patients who were tested in our hospital.
  - **Population**: All similar patients all over the world.
- Laboratory research
  - **Sample**: The data we actually collected.
  - **Population**: All the data we could have collected if we had repeated the experiment infinitely may times the same way on all mice in the world.

Statistical hypothesis testing

- For **each** gene
  - Pose *Null Hypothesis* (H_o) that gene is not affected
  - Pose *Alternative Hypothesis* (H_a) that gene is affected
  - Use statistical techniques to calculate the probability the gene is NOT affected (calculation of the so-called p-value)
  - If p-value < some critical value α reject H_o and accept H_a
- The issues:
  - Assumption of normal (Gaussian) distribution of data
  - Assumption of equality of variance. Use *moderated variance,* i.e. calculated based on the distribution of variances across all genes to make it equal for all genes
  - Multiple testing: ~10 000 genes per experiments

Normal (Gaussian) probability distribution

- Many continuous variables follow a normal distribution, and it plays a special role in statistical tests.
- The *x*-axis represents the values of a particular variable (i.e. gene expression values)
- The *y*-axis represents the probability of that *x* value *P*(*x*).
- *P*(*x*) is calculated by diving the proportion of individuals of the population that have the *x* value of the variable by the total number of individuals *n*.



$$P(x) = \frac{n(x)}{n}$$

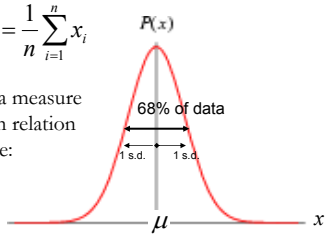
Normal (Gaussian) probability distribution

- Mean = average value of expression of the gene within a population (μ):

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

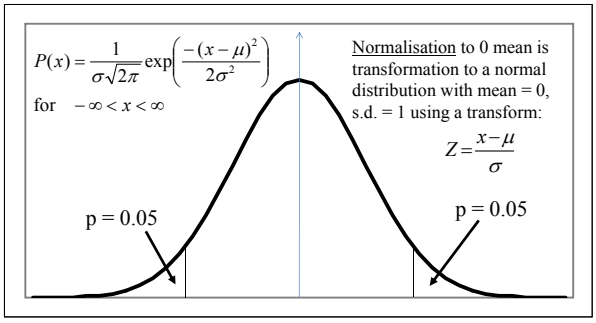
- Standard deviation (s.d. or σ) is a measure of how much the values *x* vary in relation to the mean. σ^2 is called variance:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$



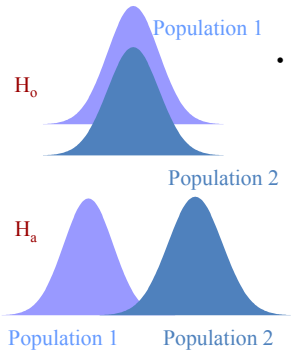
- 68% of the normal distribution lies within one s.d. of the mean (distribution is symmetrical about the mean).

Normalisation to zero mean



0.05 = p-value: probability of getting a result this extreme or more extreme given the null hypothesis is true.

Student's t-test: are the means equal or not?



- For **each** gene
  - Pose the null hypothesis (H_o) that gene is not affected, i.e. the two means are equal
  - Pose Alternative Hypothesis (H_a) that gene is affected, i.e. the two means are not equal
  - Calculate t-statistics and the p-value
  - If p-value < some critical value α reject H_o and accept H_a

Independent group t-test

- Used to compare the means of two **independent** groups.
- **Assumptions:** Subjects are randomly assigned to one of two groups. One group receives treatment. The distribution of the values being compared are normal with approximately equal variances.
- Test: The hypotheses for the comparison of two independent groups are:
  - H_o: $\mu_1 = \mu_2$ (means of the two groups are equal)
  - H_a: $\mu_1 \neq \mu_2$ (means of the two group are not equal)
- A low p-value for this test (less than 0.05 for example) means that there is evidence to reject the null hypothesis H_0 in favour of the alternative hypothesis H_a.

Paired t-test

- **Assumptions:** The observed data are from the same subject or from a matched subject and are drawn from a population with a normal distribution.
- **Characteristics:** Same subjects are often tested in a before and after situation (across time, with some intervention such as a therapy), or subjects are paired such as with twins, or with subject as alike as possible.
- **Test:** The paired t-test is actually a test that the difference between the two observations is 0. So, if d represents the difference between observations, the hypotheses are:
  - H_o: $d = 0$ (the difference between the two observations is 0)
  - H_a: $d \neq 0$ (the difference is not 0)

Calculating *t*-statistic

- First calculate t statistic and then find the p value
- **For the paired** t-test, t is calculated using the following formula:

$$t = \frac{mean(d)}{\sigma(d)/\sqrt{n}}$$

Differences d_i: $d_i = x_i - y_i$

n is the number of pairs being tested.

- **For an unpaired** (independent group) t-test, the following formula is used:

$$t = \frac{mean(x) - mean(y)}{\sqrt{\sigma^2(x)/n(x) + \sigma^2(y)/n(y)}}$$

Where $\sigma(x)$ is the standard deviation of x and $n(x)$ is the number of elements in X.

Calculating p values

- We need to know the value of p:
  - We have access to a function, which calculates p for a given critical value of t and df (degrees of freedom)
  - or alternatively have a table of critical t values indexed by t_p and df = n–1.

df	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$ ← p-value
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
∞	1.282	1.645	1.960	2.326	2.576

Critical values of t-statistic for given df

Values p and threshold α

- Once we have calculated a gene-specific t-test statistic, we determine the p-value for each gene, p_k.
- The p-value = 0.01 means that random sampling from identical populations (if they were identical) would lead to a difference smaller than we observed in 99% of experiments and larger than we observed in 1% of experiments.
- If p-value < α, reject H_o and accept H_a.
- We speak about statistically significant difference in gene expression between two conditions only when the corresponding p-value is small enough. Question: what does small mean?

Threshold of significance α

- We have to decide how small a p-value needs to be for us to think that the difference we are observing cannot be explained solely by chance (i.e. noise).
- When we test *a single* hypothesis, it is common to fix a type I error rate of $\alpha \leq 0.05$ (called level of significance).
- Type I error: reject null hypothesis when it is true (i.e., say a gene is differentially expressed when it really isn't).
- Type II error: fail to reject the null hypothesis when it is false (i.e., say a gene is not differentially expressed when it really is).

Frequency of type I errors

- Using a type I error rate of $\alpha = 0.05$ means that we are willing to make a type I error in 5% of our hypothesis tests.
- That is, if $\alpha = 0.05$, 5% of the time that the H_o is true, we will say that it's false.
- So, for every 20 hypothesis tests we perform, on average we expect 1 type I error.
- What if we are performing =10,000 hypothesis tests for 10,000 genes?
- The results will be 500 TYPE I ERRORS!

Control of frequency of type I errors

- Adjusting a type I error rate of $\alpha = 0.001$ means we will have just 1 error per 10,000 hypotheses tests, which is acceptable. But this a value is a way too strict, and maybe no gene will meet it.
- There are other more sophisticated methods of control of frequency of type I error available, called **Multiple Comparison Procedures**.
- These procedures guarantee that "family-wise error" $\leq \alpha$, where a "family-wise error" is defined to be the occurrence of a single type I error in the entire family (set) of hypotheses being tested.
  - The most popular methods are the Bonferoni correction, Holm correction and the False Discovery Rate introduced by Benjamini and Hochberg (Bonferoni correction is part of SPSS).

Signal to noise ratio (SNR)

- Another very simple and popular gene selection measure.
- Signal to Noise ratio (SNR) is a measure used in science and engineering to quantify how much a signal has been corrupted by noise.
- It is defined as the ratio of signal power to the noise power corrupting the signal.
- A ratio higher than 1:1 indicates more signal than noise.
- While SNR is commonly used for electrical signals, it can be applied to any form of signal.

Signal to noise ratio (SNR)

- A general definition of SNR is the reciprocal of the coefficient of variation, i.e., the ratio of mean to standard deviation of a signal. Indices 1 and 2 apply for condition 1 and 2, respectively.
- Signal to Noise ratio (SNR) =

$$\frac{(mean_1 - mean_2)}{(\sigma_1 + \sigma_2)} \geq cut\ off\ value$$

- NO assumptions about normality or variances
- The bigger the cut off value the better SNR
- Used by NeuCom (software for data processing and classification) developed at AUT you will use in labs.

Actual gene selection

- Gene selection based on *t*-test: rank the genes by p value and select the top $n = 50$ genes based on their p.
- Gene selection based on signal to noise ratio (SNR): order the genes from largest SNR to lowest and select the top $n = 50$ genes.
- The goal is to select genes, which have the biggest differential expression between two conditions, that is those genes that would be the best predictors of difference between the different conditions.

What's next?

- After the genes have been selected:
- Clustering and principal component analysis
  - Data Exploration
  - finding patterns
- Classification
  - Classify samples based on particular genetic profile
  - predict treatment outcome / select the best treatment
- Functional analysis: compare/evaluate functions of genes