Biology 644

Old Title: Bioinformatics for Molecular Biologists

Potential New Title: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

Outline

- Pearson's and Spearman's Correlations
- Microarrays
  - Genome expression analysis
  - ChIP-chip
- R Lab
  - Problem Set 1
  - Chapter 2
  - Chapter 2 Supplemental

Pearson's Correlation Coefficient ρ

- Measure of the linear correlation (dependence) between two variables X and Y
- Takes a value between +1 and −1 inclusive
  - 1 = total positive correlation
  - 0 = no correlation
  - −1 = total negative correlation.
- When applied to a population, the Pearson's correlation coefficient is commonly represented by the Greek letter ρ (rho) and is referred to as the population correlation coefficient or the population Pearson correlation coefficient.
- The formula for ρ is:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

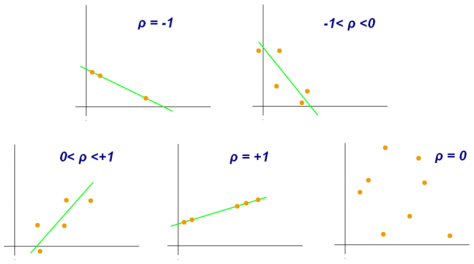
- where cov is the covariance, σ_x is the standard deviation of X , μ_x is the mean of X , and E is the expectation.
- Cannot capture the slope of a linear relationship, nor many aspects of a nonlinear relationship

Pearson's Correlation Coefficient r

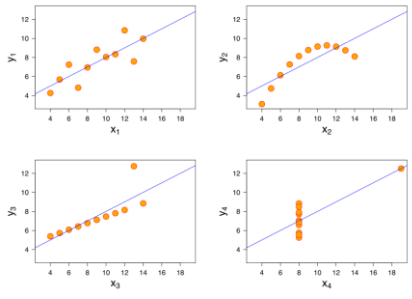
- Measure of the linear correlation (dependence) between two variables X and Y
- Takes a value between +1 and −1 inclusive
  - 1 = total positive correlation
  - 0 = no correlation
  - −1 = total negative correlation.
- When applied to a sample, Pearson's correlation coefficient is commonly represented by the letter r and is referred to as the sample correlation coefficient or the sample Pearson correlation coefficient.
- We obtain a formula for r by substituting estimates for the population covariances and variances based on a sample:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

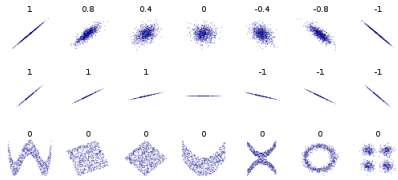
Pearson Correlations ρ and r



Pearson Correlations ρ and r



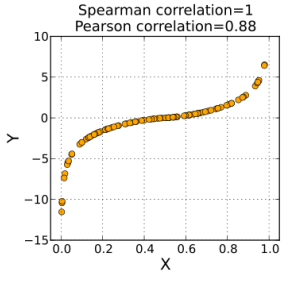
Pearson Correlations ρ and r



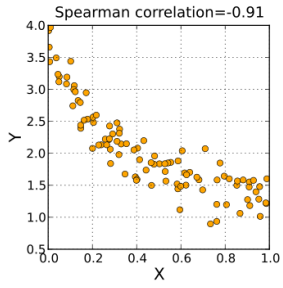
Spearman's Correlation Coefficient ρ

- A nonparametric measure of statistical dependence between two variables using the relative rankings of the values
- Assesses how well the relationship between two variables can be described using a monotonic function.
  - If X and Y are monotonically related then all data-points with greater x-values than that of a given data-point will have greater y-values as well.
- If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.
- Defined as the Pearson correlation coefficient between the ranked values
- Identical values (rank ties or value duplicates) are assigned a rank equal to the average of their positions in the ascending order of the values.
  - Example: If in an ordered ranking entries 4 and 5 have the same value then both are given rankings of 4.5
- More resistant to outliers compared to Pearson's Correlation Coefficient

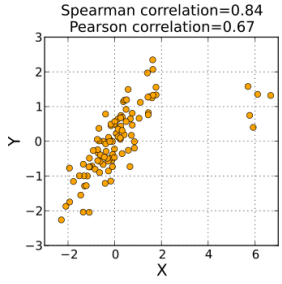
Spearman's Correlation Coefficient ρ



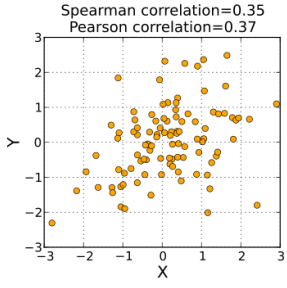
Spearman's Correlation Coefficient ρ



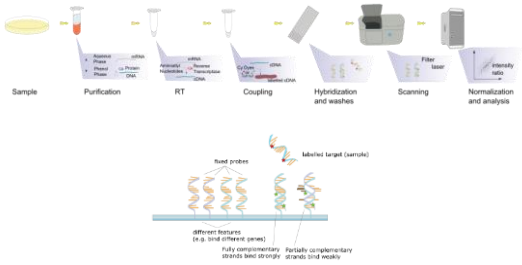
Spearman's Correlation Coefficient ρ



Spearman's Correlation Coefficient ρ

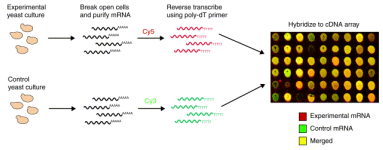


1-Color Microarray

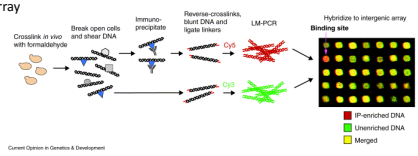


2-Color Microarray

Expression Array

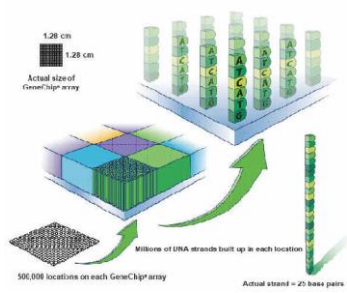


Tiling Array



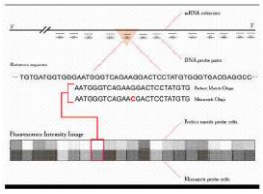
Affymetrix GeneChip

Affymetrix Microarray



Oligonucleotides (oligos), usually 25-mers (25 bases long), are directly synthesized onto a glass wafer. Each array contains up to 900,000 different oligos and each oligo is present in millions of copies.

Affymetrix Probe Set



- Probe Set - Each gene is represented on the array by a series of different oligonucleotide probes.
- Probe Pair - Consists of a perfect match oligonucleotide and a mismatch oligonucleotide.
- The perfect match probe has a sequence exactly complimentary to the particular gene and thus measures the expression of the gene.
- The mismatch probe differs from the perfect match probe by a single base substitution at the centre base position, disturbing the binding of the target gene transcript.
- This helps to determine the background and nonspecific hybridization that contributes to the signal measured for the perfect match oligo. The GeneChip Operating Software MAS algorithm subtracts the hybridization intensities of the mismatch probes from those of the perfect match probes to determine the absolute or specific intensity value for each probe set.

Terms used in Affymetrix GeneChips

- **Target -** Fragmented, biotinylated anti-sense cRNA prepared from mRNA to be analysed. Target molecules are hybridized to the probe array and the levels of hybridization are measured with the GeneArray scanner after the array is stained with biotin - streptavidin-phycoerythrin (SAPE).
- **Probe -** Single-stranded DNA oligonucleotide synthesized directly on the surface of the GeneChip array using photolithography and combinatorial chemistry. The 25 base oligonucleotide is designed to be complementary to a specific gene transcript.
- **Probe Cell -** Single square-shaped feature on an array containing probes with a unique sequence. The size can vary depending on the array type, typically 20 µm or 18 µm. Each probe cell contains millions of probe molecules.
- **Perfect Match (PM) -** Probes that are designed to be complementary to a reference sequence.
- **Mismatch (MM) -** Probes that are designed to be complementary to a reference sequence except for a homomeric mismatch at the central position (e.g., 13th position of 25 base probe, A->T or G->C). Mismatch probes serve as a control for cross-hybridization.
- **Probe Pair -** Two probe cells, a PM and its corresponding MM. On the probe array, a probe pair is arranged with a PM cell directly above a MM cell.
- **Probe set -** A set of probes designed to detect one transcript. A probe set usually consists of 11-20 probe pairs. For example, an 11 probe pair set is made up of 11 PM probes and 11 MM probes for a total of 22 probe cells. Newer array designs from Affymetrix, e.g., HG-U133, contain probe sets with 11 probe pairs. Older designs have average probe set numbers of 16 or 20 probe pairs.
- **Target Sequence -** The portion of a transcript reference sequence that is interrogated by a probe set on the array. The target sequence extends from the first base of the most 5' probe to the last base of the most 3' probe.

Gene Expression Heat Map