Bioinformatic resources

Giacomo Fantoni
telegram: @GiacomoFantoni
Ilaria Cherchi
telegram: @ilariacherchi
Github: https://github.com/giacThePhantom/thesis_notes

July 4, 2022

Contents

1 Introduction **4**
- 1.1 Bioinformatics . . . 4
  - 1.1.1 The human genome . . . 4
- 1.2 Involvement of computer science . . . 4
  - 1.2.1 Databases . . . 4
  - 1.2.2 Program . . . 5
  - 1.2.3 Algorithm . . . 5

2 Scientific literature **6**
- 2.1 Literature sources . . . 6
  - 2.1.1 Primary literature . . . 6
  - 2.1.2 Secondary literature . . . 6
- 2.2 Structure of a scientific article . . . 6
- 2.3 Impact measures . . . 7
  - 2.3.1 Impact of a journal . . . 7
  - 2.3.2 Personal impact . . . 7
  - 2.3.3 Peer review . . . 7

3 Biological databases **8**
- 3.1 Introduction . . . 8
  - 3.1.1 Classification of databases . . . 8
  - 3.1.2 Data sources . . . 8
  - 3.1.3 Nomenclature . . . 8
  - 3.1.4 Reference genome . . . 8
  - 3.1.5 Popular databases . . . 9
  - 3.1.6 GenBank . . . 9
  - 3.1.7 RefSeq . . . 9
  - 3.1.8 UniProt . . . 9
  - 3.1.9 Others . . . 9
  - 3.1.10 Genome browser . . . 9

4 Motif analysis **11**
- 4.1 Introduction . . . 11
  - 4.1.1 Definition . . . 11
  - 4.1.2 Functions . . . 11
- 4.2 Motif search . . . 11
  - 4.2.1 Consensus sequence . . . 12

4.2.2 Positional matrix . . . 12
4.2.3 Hidden Markov model . . . 13
4.2.4 Sequence logos . . . 13
4.3 Motif identification . . . 14
4.3.1 Finding known motifs - pattern matching . . . 14
4.3.2 Finding de novo motifs - pattern discovery . . . 15

5 Expression analysis **17**
5.1 Introduction . . . 17
5.1.1 Expressed genes . . . 17
5.1.2 Differential gene expression . . . 17
5.1.3 Databases . . . 18
5.2 Microarrays . . . 18
5.2.1 Introduction . . . 18
5.2.2 Fabrication . . . 19
5.2.3 Reading signal . . . 19
5.2.4 Image analysis . . . 19
5.2.5 Batch effect . . . 19
5.2.6 Data pre-processing . . . 19
5.2.7 Gene expression microarray . . . 21
5.3 RNA-sequencing . . . 24
5.3.1 Introduction . . . 24
5.3.2 Illumina's pipeline . . . 24

6 Gene set enrichment **26**
6.1 Introduction . . . 26
6.1.1 Functional groups characterized by gene expression change . . . 26
6.1.2 Gene-set enrichment analysis . . . 26
6.1.3 Ontology . . . 26
6.1.4 Controlled vocabulary . . . 26
6.2 Gene ontology . . . 27
6.2.1 Concepts hierarchy . . . 27
6.2.2 Ontology structure . . . 27
6.2.3 The gene ontology project . . . 28
6.3 Gene sets . . . 28
6.3.1 Sources and types . . . 28
6.3.2 Differences between pathways and processes . . . 28
6.3.3 Enrichment test . . . 29
6.3.4 Whole-distribution - GSEA enrichment . . . 30
6.3.5 Gene set filter . . . 30
6.3.6 Redundancy problem . . . 30

7 Network analysis **32**
7.1 Introduction . . . 32
7.1.1 Network definition . . . 32
7.1.2 Networks in system biology . . . 32
7.2 Graphs . . . 32
7.2.1 Definition . . . 32

7.2.2 Magnitude of a graph . . . 32
7.2.3 Degree of a graph . . . 33
7.2.4 Weighted graphs . . . 33
7.2.5 Complete graphs . . . 33
7.2.6 Paths . . . 33
7.2.7 Bipartite graphs . . . 33
7.2.8 Graphs connections . . . 33
7.2.9 Subgraphs . . . 34
7.2.10 Trees . . . 34
7.2.11 Clique . . . 34
7.2.12 Isomorphisms . . . 35
7.2.13 Representing a graph . . . 35
7.3 Networks properties . . . 35
7.3.1 Analysis of single elements . . . 35
7.3.2 Analysis of groups . . . 37
7.3.3 Analysis of network . . . 37
7.3.4 Small world effect . . . 37

8 Genome wide association studies **38**
8.1 Types of studies . . . 38
8.1.1 Descriptive studies . . . 38
8.1.2 Analytical studies . . . 38
8.1.3 Cohort studies . . . 38
8.1.4 Case-control study . . . 39
8.2 GWAS . . . 40
8.2.1 Objective of GWAS . . . 40
8.2.2 Main application of GWAS . . . 40
8.2.3 GWAS methodology . . . 41
8.2.4 Single nucleotide polymorphisms . . . 41

Chapter 1

Introduction

1.1 Bioinformatics

Bioinformatics can be defined as the development of new algorithms and statistical methods that allow to establish relations between members of huge sets of data. It can also be described as the analysis and interpretation of different data types including nucleotide and amino acid sequences, protein domains and protein structures. The development and implementation of these programs allow efficient access and management of different types of information.

1.1.1 The human genome

The human genome contains around 3.2 billion base pairs. About 80% of it is associated with a biochemical function. Of particular interest is non-coding DNA, which doesn't code for proteins but is mainly involved in:

- Protection of the genome.
- Gene switches.
- Gene expression regulation.
- Transcription factor binding sites.
- Operators.
- Enhancers.
- Promoters.
- Silencers.

1.2 Involvement of computer science

Computer science plays a fundamental part in bioinformatics, providing the algorithms necessary to exploit data collected in experiment to reach a significant conclusion.

1.2.1 Databases

Databases or data banks are collections of correlated data utilized to represent a portion of the real world. They are structured in a way to allow data organization and management in terms of:

- Insertion.
- Update.
- Search.
- Deletion.

1.2.2 Program

A program codifies an algorithm into a programming language. It is used to test and realize a proposed solution. Computer science can be defined as the science of the automatic elaboration of information, with algorithm as its central focus.

1.2.3 Algorithm

Algorithms (from the name of the Persian mathematician *Muhammad ibn Musa al-Khwarizmi*) can be defined as a system of well-defined rules and procedures that lead to the solution of a problem with a finite number of steps. They can be described in pseudo-code.

1.2.3.1 Substring search algorithm

A substring search algorithm is an algorithm that searches the occurrence of a string in another, allowing to understand if the former is contained in the latter. A naïve implementation is described in figure 1.1.

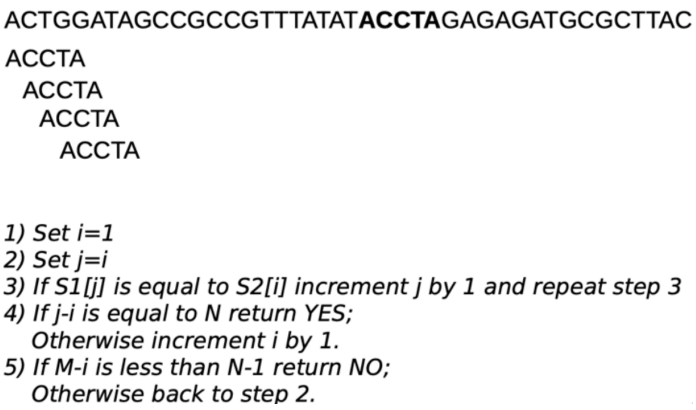

Figure 1.1: Naïve implementation of a substring search algorithm

Chapter 2

Scientific literature

2.1 Literature sources

All bioinformatics works are based on literature. Different sources of literature can be found.

2.1.1 Primary literature

Primary literature is defined as original materials. It is authored by researchers, contains original research data and is usually published in a peer-reviewed journal. Primary literature works can be:

- Journal articles or conference proceedings, which are usually the first formal appearance of a result.
- Original articles: the original research conducted by the authors, including methods and resources used.
- Letters or communications: short reports of original research focused on an outstanding finding whose importance means that it will be of interest to scientists in other fields.

2.1.2 Secondary literature

Secondary literature is the summary or review of the theories and results of original scientific research. Secondary literature works can be:

- Open letters.
- Comments.
- News.
- Reviews.
- Correspondence.
- Opinions.
- Protocols.

2.2 Structure of a scientific article

Scientific articles tend to have a well defined structure, composed, in order, of:

1. Title.
2. Abstract.
3. Keywords.
4. Introduction or background.
5. Methods or experiments.
6. Results or analysis.

7. Discussion.
8. Conclusion.
9. References or bibliography.
10. Figures and tables.
11. Supplementary materials.

2.3 Impact measures

An impact measure is used to define the goodness of a research or if it had a big impact in the community.

2.3.1 Impact of a journal

A measure of impact of a journal measure the impact of the publication of a journal. It can be measured in different ways:

- Impact factor (IF): a measure that reflects the average number of citations of articles published in a science journal. It can be biased due to self-citations, journal-forced citations and it does not take into account negative citations. It is computed as:

$$IF_y = \frac{Citations_y}{Publications_{y-1} + Publications_{y-2}}$$

- Journal of Citation reports JCR.
- Scimago Journal Rank SJR.

2.3.2 Personal impact

The personal impact measure the impact of a researcher. I can be measured as:

- H-index: an index that attempts to measure both the productivity and the impact of the published work of a scientist or scholar. A scholar with an index of h has published h papers each of which has been cited by others at least h times. It serves as an alternative to more traditional journal impact factor metrics in the evaluation of the impact of the work of a particular researcher.
- Web of Science WOS.
- Scoups.
- Google Scholar.

2.3.3 Peer review

Peer-reviewed articles are also called refereed articles. Peer review allows to:

- Independently verify theories and assumptions.
- To screen for the works ethic.
- Asses appropriateness for publication.
- Check for transparency of research.
- Assess the quality of the research.

Depending on the journal or publisher this process can takes from weeks to months.

Chapter 3

Biological databases

3.1 Introduction

3.1.1 Classification of databases

A huge number of biological databases are available and they can be distinguished as:

- Primary databases containing sequences of nucleotides and amino acids.
- Derived and specialized databases containing protein domains and motifs, protein structures, genes, transcripts, expression profiles, variations, pathways and many other informations.

Each database is characterized by a central biological element which constitutes the object around which the principal entry of the database is constructed.

3.1.2 Data sources

Data in these databases is derived from:

- Literature.
- In-vitro and in-vivo analysis.
- In-silico analysis.

3.1.3 Nomenclature

One of tha main problems related with biological database is nomenclature. There can be different name for the same gene or different genes with the same name. To uniquely identify genes and proteins and manage the large amount of information related, primary data banks assign an accession number to each element they store.

3.1.4 Reference genome

A reference genome is a digital sequence of nucleic acids assembled to be a representative sequence for a given species. It is assembled from DNA sequencing of a set of donors. An example of reference genome is *GRCh38* from which *hg38* is derived aggregating many donor informations.

3.1.5 Popular databases

3.1.6 GenBank

GenBank contains nucleotide sequences. The aim of the database is to store and archive historically important but redundant nucleotide sequences. Data can be submitted singularly or in a batch manner.

3.1.7 RefSeq

RefSeq is a curated and non redundant collection of DNA, RNA and protein sequences. Each RefSeq entry represents a sings molecule in a particular organism. Its basis is compiled with a process of collaboration, extraction and computation from GenBank. Each molecule is annotated reporting the name of the organism, the correct gene symbol for that organism and informative names of proteins when possible.

3.1.8 UniProt

UniProt is a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. It provides protein sequences, domains and structural information like subcellular location for many species. It also includes some alignment and mapping tools.

3.1.9 Others

Other examples of derived databases are:

- dbGap.
- Structure.
- Gene.
- Biosystems.

3.1.10 Genome browser

A genome browser is a database containing reference sequence assemblies for one or more genomes. It allows to browse data at various detail levels, from chromosome to gene, down to a single exon or intron. It also allows for the comparison between species and data extraction.

3.1.10.1 UCSC genome browser

The UCSC genome browser contains the genome of about 100 species, but it does not provide a browser for all of them. It integrates informations like SNPs, sequence conservation, regulatory elements (ENCODE) and others.

3.1.10.1.1 UCSC table browser The UCSC table browser allow to extract data from the database tables without the need for a graphical interface. It can also align sequences, annotate SNPs and convert data between genome versions. It is a flexible tool that can retrieve data for one or more genes in a variety of formats. When submitting heavy task it will redirect them to Galaxy, an online workflow system.

3.1.10.2 Ensembl genome browser

The Ensembl genome browser is the European genome browser. It focuses on vertebrate genomes. It includes genomic variants, both somatic and structural, and regulatory elements data. It offers an interfcace to access data directly BioMart. All Ensemble transcripts are based on proteins and mRNAs contained in the databases:

- UniProt/Swiss-Prot (manually curated).
- UniProt/TrEMBL (not reviewed).
- NCBI RefSeq (manually curated).

3.1.10.2.1 Biomart Biomart is a data mining platform which is able to address complex queries on ENSEMBL. It is similar to the USCS table browser, while being more powerful as it can retrieve both annotation and sequences.

Chapter 4

Motif analysis

4.1 Introduction

4.1.1 Definition

A DNA motif is a pattern of nucleotide sequences. They are usually associated to DNA-protein binding site and so to regulatory regions. They are a small pattern, usually between 5 and 30bp that can recur many times in the genome and many times in the same gene. Motifs can be:

- Standard.
- Palindromes.
- Gapped.

4.1.2 Functions

DNA motif functions include:

- Sequence specific binding sites reached by transcription factors, nucleases and ribosomes.
- mRNA processing:
  - Splicing: exonic splicing enhancer ESE.
  - Editing: protospacer adjacent motif PAM, a DNA sequence that immediately follows the target DNA sequence of the Cas9 nuclease in the CRISPR system.
  - Polyadenilation.
  - Transcription termination.

4.1.2.1 Degenerate motifs

Motifs in regulatory regions are often similar but variable: they are degenerate. Transcription factors are often pleiotropic, meaning that they regulate a lot of genes, but they need to be expressed at different levels. Degenerate motifs cause non-specific binding: a protein can bind genomic position different with respect to the one corresponding to the expected functional state.

4.2 Motif search

The objectives of motif search are to identify:

- Over-represented motifs in the genome.
- Motifs conserved in ortholog sequences.
- Sequences that can be candidates for transcription factor binding.

Motifs can be represented as a consensus sequence or as profiles like positional matrices or HMMs.

4.2.1 Consensus sequence

A consensus sequence represents the result of multiple sequence alignments with the goal of finding recurrent motifs across the sequences. This sequence can be potentially different from all input sequence: it presents only the most conserved sequences for each position. It is built such that it minimizes the distance from each input sequence at each position. It can be also written following a IUPAC notation.

4.2.2 Positional matrix

A positional matrix is an alternative way to represent a motif than the consensus sequencs. The elements in the matrix represent all possible bases at each position. Example of these matrices are:

- Position frequency matrix PFM or PSWM.
- Position weight matrix PWM or PSSM.
- Position probability matrix PPM or PFM.

4.2.2.1 Populating a position frequency matrix

A position frequency matrix is computed as:

$$M_{k,j} = \sum_{i=1}^{N} \delta(X_{i,j} = k)$$

Given k the set of all symbols in the alphabet. N the number of aligned sequences. j iterates over the length of the sequence.

4.2.2.2 Populating a position probability matrix

A position probability matrix is very similar with respect to the position frequency matrix, with the exception that each cell represent the probability that in that sequence position a particular base will be found. Its cells are computed as:

$$M_{k,j} = \frac{1}{N} \sum_{i=1}^{N} \delta(X_{i,j} = k)$$

4.2.2.3 Assessing the probability that a sequence belong to a PPM

To assess the probability for a sequence to belong to a PPM the probabilities for each base i found at each position j are multiplied:

$$P(seq \in PPM) = \prod_{j=1}^{R} M_{seq_j,j}$$

4.2.2.4 Correcting PPMs

4.2.2.4.1 Laplace smoothing Laplace smoothing introduces pseudocounts to allow to estimate probabilities in case of too few observations. A pseudocount is an amount added to the number of observed cases in order to change the expected probability.

4.2.2.4.2 Adding a background model Another way to correct PPMs is to add a background model. With this a new matrix is computed as:

$$M_{k,j} = \log_2 \frac{M_{k,j}}{b_k}$$

Where b represent a background model and can vary across nucleotides for organisms with high GC content. It is typically computed as:

$$b_k = \frac{1}{|k|}$$

And so is 0.25 for nucleotides and 0.05 for amino acids.

4.2.3 Hidden Markov model

A Markov chain is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules. The possible future states are fixed and not based on how the process arrived at its present state. The state is directly visible to the observer: state transition probabilities are the only parameter. The state remain transparent, while the output is easily obtainable. A HMM of the first order is defined as:

- A finite set of states S.
- A discrete alphabet of symbols.
- A matrix of transition probabilities $T = P(i|j)$, the probability of transition from state j to i.
- A matrix of emission probabilities (the probability of skipping a state) $T = P(X|i)$, the probability of X emission in state i.

4.2.3.1 Assessing the probability that a sequence is generated by a HMM

The probability that a sequence is generated by a HMM can be computed as:

$$P(S|w) = \sum_{\pi} P(S, \pi|w)$$

Where S is the sequence, w are the probabilities parameters, π are all possible computationally inefficient paths. Efficient algorithm to compute this probability are Forward-Backward and Viterbi.

4.2.4 Sequence logos

Sequence logos are visual representation of positional matrices and simple HMM profiles. The height of each character in a sequence logo is proportional to its information content: 2 bit if 1 base occurs in all input sequences, 1 if two bases occur and 0 if all bases occur equally. The higher the variability, the lower the height of a specific base. In particular the height of base b at position l is computed as:

$$f(b,l)R_{sequence}(l)$$

Where:

$$R_{sequence}(l) = 2 - (H(l) + e(n))$$

Such that $H(l)$ is the Shannon entropy and is computed as:

$$H(l) = -\sum_{b=a}^{t} f(b,l)\log_2 f(b,l)$$

And

$$e(n) = \frac{1}{\ln 2} \cdot \frac{4-1}{2n}$$

4.3 Motif identification

There are two types of motif identification: pattern matching and pattern discovery.

4.3.1 Finding known motifs - pattern matching

Pattern matching is the problem of finding known motifs, for example seeing if a binding of a protein to an upstream region of a gene is significant. In order to find out whether a transcription factor matches a promoter the PFM matrix is used to compute a score for each sliding window. This scores can be plotted against a threshold, so as to identify regions able to support a putative binding.

4.3.1.1 Total binding affinity

Total binding affinity TBA is a cutoff-free method. The TBA is a method used to describe the affinity of a DNA sequence for a transcription factor described by a PFM with a single score. It takes into account binding sites of all possible affinities and considers the whole sequence, keeping into account both high and low affinity sites. For a sequence is computed as:

$$a_{rw} = \sum_{i=1}^{L-l+i} \max\left(\prod_{j=1}^{l} \frac{P(w_j, r_{i+j-1})}{P(b, r_{i+j-1})}, \prod_{j=1}^{l} \frac{P(w_{l-j+1}, r'_{i+j-1})}{P(b, r'_{i+j-1})}\right)$$

Where:

- r is the sequence.
- w is the PFM.
- l is the length of w.
- L is the length of r.
- r_i is the nucleotide at position i.
- r'_i is the nucleotide at position i on the other strand.
- $P(w_j, r_i)$ is the probability to observe the given nucleotide r_i at position j of w.
- $P(b, r_i)$ is the background probability to observe the same nucleotide r_i.

4.3.2 Finding de novo motifs - pattern discovery

Pattern discovery is the problem of finding de novo motif, for example finding the motifs upstream of a specific gene. Given a set of sequences, the objective is to find the most represented motifs. In particular the hidden sequence can be considered of length N and with a slightly different pattern in each sequence. Using the MEME suit it is possible to identify new sequences and through Jaspar they can be compared to already characterized transcription factors. Methods can be:

- Exact: give optimal solution given specific parameters.
- Approximated: give suboptimal solution decreasing the computational burden. They are MULTIPROFILER, CONSENSUS, MEME, Gibbs sampler and MotifSampler for example.

4.3.2.1 Distance between a real motif and the consensus

The distance between a real motif and the consensus is generally less that that for two real motifs. The consensus sequence must be guessed and a scoring function to compare different guesses and choose the best one must be chosen.

4.3.2.2 Elements of the problem

The problem of finding de novo motifs can be formalized considering the following elements:

- n the length of each sequence.
- DNA, an array of size $t \times n$.
- l, the length of the motif or l-mer.
- s_i, the starting position of an l-her in sequence l.
- $s = (s_1, s_2, \ldots, s_t)$, an array of motifs starting position.

If the starting positions s are given, finding the consensus is easy. When those are not given, finding the best motif is solving the median string problem.

4.3.2.3 The median string problem

Given a set of t DNA sequences the objective is to find a pattern that appears in all t sequences with the minimum number of mutations. The Hamming distance is used, such that:

$$d_h(v, w) = \#\ nucleotide\ pairs\ that\ do\ notmatch\ when\ v\ and\ w\ are\ aligned$$

Then, for each DNA sequence I, all $d_h(v, x)$ are computed, where x is an l-mer with starting position s_i. Then the minimum $d_h(v, x)$ among all l-mers of the sequence. The $TotalDistance(v, DNA)$ is the sum of the minimum Hamming distances for each DNA sequence I, so

$$TotalDistance(v, DNA) = \min_s d_h(v, s)$$

Where s is the set of starting positions.

4.3.2.4 Planted motif search

Planted motif search PMS is another algorithm that can be used to find de novo motifs. Given t sequences of length n it outputs a motif M of length l. Variants of interest have an hamming distance of d from M. From the input sequences the algorithm generates all possible l-mers. Let C_i be the collection of this l-mers. Then this are sorted and the duplicates deleted. Then for all sequences the motif common for all lists of motifs is found, and that outputs the planted motif sets.

Chapter 5

Expression analysis

5.1 Introduction

5.1.1 Expressed genes

The expressed genes are those genes that have been transcribed. A gene expression profile of a cell is the snapshot of which genes are expressed in that cell at the time the sample was taken. Knowing which genes are expressed in a cell allows the identification of new genes or transcripts and the comparison of expression profiles between samples. Variability in gene expression is mainly due to alternative splicing and different regulation. It can be analysed to uncover characteristics of diseases, development and dynamic responses to stimuli.

5.1.2 Differential gene expression

During a differential gene expression experiment the expression profile of genes is compared between samples. Comparison can be done between:

- Different cells.
- Different tissues.
- Different disease states.
- Different developmental stages.
- Different culture conditions.

Negative and positive controls and the range of variability within samples must be taken into account.

5.1.2.1 Differential gene expression workflow

A typical differential gene expression analysis workflow consists of:

1. Formulation of the biological question.
2. Experimental design: choice of platform, control and replicates.
3. Running the experiment.
4. Image processing done by a machine.
5. Low-level analysis: data pre-processing with normalization.
6. High-level analysis.
7. Obtaining biological conclusions and interpretation of results.

5.1.2.2 High throughput methods

To perform differential gene expression analysis high throughput methods can be used. Their pros and cons are described in table 5.1.

Pros	Cons
Fast	Difficult to filter non coding RNA
Comprehensive (entire genomes)	Not enough attention to design
Easy	Artefacts
Getting cheaper	Cannot afford controls or replicates

Table 5.1: High throughput methods pros and cons

5.1.3 Databases

Repositories of array and NGS data mainly contain expression data. All of these databases can be interrogated with Bioconductor packages in R.

5.1.3.1 Gene expression omnibus

The gene expression omnibus or GEO is a public repository for the archiving and distribution of gene expression data submitted by the scientific community. It is a curated, online resource for gene expression data browsing, query, analysis and retrieval. It is convenient for the deposition of gene expression data as required by funding agencies and journals. Submitted data needs to include:

- Platform.
- Sample.
- Series.
- Dataset.

GEO is connected to repositories specifically tailored to store raw data like BioProject or SRA.

5.1.3.2 ArrayExpress

ArrayExpress or EBI si another online repository of array expression data.

5.1.3.3 Gene expression Atlas

The gene expression atlas GXE NCBI provides information on gene expression patterns. The raw data is re-analysed with common pipelines.

5.2 Microarrays

5.2.1 Introduction

Microarrays have been introduced at the beginning of the 2000s and are the first high throughput technology. They are useful to investigate, for example:

- Genomic profiles.
- Transcriptomic profiles.
- The methylome.
- DNA-protein interactions.

Data interpretation is subject to specific computational analyses. Microarrays monitor thousands of genes in parallel. Each spot contains multiple and identical DNA probes and thousands of spots are disposed as a matrix on a solid support.

5.2.2 Fabrication

Microarrays can be fabricated using different technologies. Probes can be:

- Oligonucleotides.
- cDNA.
- Small PCR fragments related to specific mRNA.

The probes are synthesized and placed on the support and can have different lengths, usually between 25 and 60nt. Moreover microarrays can have different numbers of channels:

- 2 channels: test and control samples are labelled with different fluorophores.
- 1 channel: one sample is loaded per time.

Depending on the technology microarrays can capture, for example:

- Exons.
- Genes.
- $3'$-ends.

5.2.3 Reading signal

A scanner allows to read the fluorescence light emitted by the fluorophores. The information is stored in 2 images for channels arrays at 16 bit resolution. The image in grey scale is represented in a red-green scale that represents the light emitted by the two fluorophores. The resulting colour will be proportional to the quantity of test and control DNA.

5.2.4 Image analysis

After having obtained the images, these have to be analysed. This is performed by a specific technology like Affymetrix. The first step is a segmentation analysis: the shape and patterns inside data is analysed to assess the signal quality for each spot. The background and foreground are identified to correct for noise generated by the former. One of the standard method is to create a signal model and fit the data to it in order to evaluate the quality of the spot, for example computing the interquartile range on the distribution in order to find feature, exclusion zone and background. Then the fluorescence of each spot is estimated and the relative expression for each gene is interpreted thanks to annotation information.

5.2.5 Batch effect

For microarrays is complex to compare different technologies as different probes and methods are used. It is always preferable to avoid performing an integrative analysis.

5.2.6 Data pre-processing

Data pre-processing is needed to reduce errors introduced during the experimental process. It consists typically of:

- Background subtraction: eliminates background noise.
- Normalization: all samples are brought into a similar range of distribution, to reduce the effect of:
  - Unequal quantity of starting sample.
  - Differences in labelling efficiency.
  - Differences in detection efficiency.
  - System biases.
- Summarization: summary of information from several spots into a single measure for each gene.
- Statistical quality control: removes low quality samples and probe sets.

5.2.6.1 Two channels array

The pre-processing pipeline for 2 channels array consists of different steps.

5.2.6.1.1 Background correction During background correction signal R_s and G_s and background estimates R_b and G_b are separated. Then the background corrected estimates R_c and G_c are computed as:

$$R_c = R_s - R_b \qquad \wedge \qquad G_c = R_s - R_b$$

Or as:

$$R_c = \max(R_s - R_b, 0) \qquad \wedge \qquad G_c = \max(G_s - G_b, 0)$$

5.2.6.1.2 Summarization and transforms Log-ratios estimates relative expression as:

$$\log \frac{R_c}{G_c}$$

5.2.6.1.3 Normalization Normalization is useful to identify systematic intensity-dependent bias in the data. The ratio of signal might depend on the average signal density measured across different channels. The function of dependance can be fitted to a polynomial regression like Loess to obtain normalization to make the plot more informative.

5.2.6.2 One channel array

Many methods have been developed to pre-process Affymetrix one channel arrays:

- Advanced methods: GCRMA, PLIER.
- Popular methods: RMA and MAS5.
- Rudimentary methods: MAS4, LOESS.

5.2.6.3 Robust multi array average

Robust multi array average is a pre-processing methods that consists of three steps.

5.2.6.3.1 Background correction Background correction removes local artefacts and noise. The probe measure data is assumed as a combination of background noise in a normal distribution and signal in an exponential distribution. Assuming strictly positive distribution of signal background, the corrected signal is positively distributed. Background correction is performed on each array separately using the observed distribution of PMs (probe measures). It estimates μ, σ and λ in:

$$PM = Signal + Background \rightarrow Signal : S \sim e^{\lambda} \wedge Background : B \sim N(\mu, \sigma^2)$$

Then an estimate $E(S|PM)$ for each PM value can be obtained.

5.2.6.3.2 Normalization Normalization is used to remove array effects, making all distributions the same. Quantile normalization is used to correct for array biases, as it compares the expression levels between arrays for various quantiles. It protects against outliers.

5.2.6.3.3 Summarization Summarization combines probe intensities across arrays to get a single intensity value tor each gene or probeset. In median polishing each chip is normalized to its median and each gene normalized to its median. The procedure is repeated until medians converge. A maximum of 5 iteration is allowed to prevent infinite loops.

5.2.7 Gene expression microarray

An expression microarray experiment is used to test differences in gene expression between two or more conditions that could be for example cancer versus normal or different treatments. Each condition can be represented by one or more samples. The null hypothesis is that there exists no difference between the gene expression in the two conditions. The comparison is done using the ratio between the test and the control samples. It should not differ in case of null hypothesis validity. These ratios are also defined as fold changes:

$$FC = \begin{cases} Ratio & Ratio > 1 \\ -\frac{1}{Ratio} & Ratio < 1 \end{cases}$$

Because ratios are not symmetric with respect to 1 the statistics are not easy to analyse, so the log-ratio is often used. The log ratio of the null hypothesis should be 0.

5.2.7.1 Replicates

Replicates are needed considering the noise of microarray data. They can be distinguished between:

- Technical replicates: experiments on more RNA samples obtained from the same biological source.
- Biological replicates: experiments on more biological sources belonging to the same condition.

Ideally each condition should be represented by more biological replicates in order to perform a statistical test. They can also be summarized as mean for each gene.

5.2.7.2 Statistical tests

Microarray correlation can be exploited to identify differentially expressed genes. A gene is called differentially expressed through a Z statistics. To find significantly differentially expressed genes a test statistic for each gene should be used. A low p-value is interpreted as evidence that the null hypothesis can be false and so a gene is differentially expressed.

5.2.7.2.1 T-Test The T-test is a parametric test to check the difference between the mean of two groups. It assumes that the variance of those two groups is the same. Is computed as:

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

5.2.7.2.2 Walch t-test Walch t-test considers different variance between two groups. It is the default implementation for R $t.test()$ function. Is computed as:

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}$$

5.2.7.2.3 Wilcoxon test The Wilcoxon test is a non parametric test to check the equality of two distributions.

5.2.7.2.4 Permutation test The permutation test generates a null distribution on an observation of interest by changing the group labels. It compares the values observed in data and the values in the generated null distribution. Is computed as:

$$p = \frac{\#\{b : |T_b| \geq T_{obs}\}}{B}$$

5.2.7.2.5 ANOVA analysis ANOVA analysis is used to control different treatments. It allows to test the null hypothesis that the differences within and between at least 3 groups are the same on average.

5.2.7.2.6 2-way ANOVA 2-way ANOVA compares the mean differences between groups that have been split on two independent variables called factors. It understand if there is an interaction between the two independent variables on the dependent one.

5.2.7.2.7 From p-value to probability of significant results Given a p-value and n the number of hypothesis to test, P the probability of having al least one significant result is computed as:

$$P = 1 - (1 - pvalue)^n$$

5.2.7.3 Correction methods

Correction methods are used to correct p-values in the presence of multiple null hypothesis. Considering n multiple hypothesis the probability of observing one significant result due to chance is:

$$1 - (1 - pvalue)^n$$

Some examples are:

- Bonferroni: very conservative, it has a significance threshold of $\frac{\alpha}{N}$. It reduces false positive while introducing false negatives.
- Benjamini-Hochenberg: it tunes false positives and false negatives.
- False discovery rates: it checks if the kth ordered p-value is larger than $\frac{k \cdot a}{N}$

5.2.7.4 Receiver operating characteristic

The receiver operating characteristic find genes that better discriminate between two conditions by plotting sensitivity and 1−specificity. Let AUC be the area under the ROC curve, then:

- $AUC = 1$: perfect classification.
- $AUC = 0.5$: random classification.

5.2.7.5 Clustering

The objective in differential gene expression is to look for genes that behave differently between samples. Once having obtained then they can be clustered according to similar expression to make the outliers more obvious.

5.2.7.5.1 Hierarchical clustering The algorithm creates a dendrogram based on on a definition of distance measure between observation pairs. The idea is:

- Starting with N observation and a distance measure for all $\frac{N(N-1)}{2}$ pairs. Each observation is a cluster.
- For $i = n$ to 2:
  - Examine all inter-clusters distances and fuse the cluster with lower distance. The distance between fused clusters represents the height of the bar in the dendrogram.
  - Calculate new inter-cluster distances between the remaining $i - 1$ clusters.

5.2.7.5.2 Linkage Linkage defines the distance between clusters.

5.2.7.5.2.1 Complete linkage In complete linkage maximal intercluster dissimilarity is reached. All pairwise dissimilarities between the observations in cluster A and in cluster B are computed and the largest is recorded.

5.2.7.5.2.2 Single linkage In single linkage minimal interclsuter dissimilarity is reached. All pairwise dissimilarities between the observations in cluster A and in cluster B are computed and the smallest is recorded. It can result in extended, trailing clusters in which single observations are fused one at a time.

5.2.7.5.2.3 Average linkage In average linkage te mean intercluster dissimilarity is reached. All pairwise dissimilarities between the observations in cluster A and in cluster B are computed and the average is recorded.

5.2.7.5.2.4 Centroid linkage In centroid linkage the dissimilarity between the centroid for cluster A (a mean vector of length p) and for cluster B is computed and recorded. It can result in undesirable inversions.

5.3 RNA-sequencing

5.3.1 Introduction

RNA sequencing is a next-generation sequencing approach that sequences the cDNA from the mRNA component. The whole transcriptome can be compared against the whole transcriptome of another sample. It is very cheap, sequencing $\sim$ 400 gigabases per flow cell, but it has an error rate up to 1%, issues with AT and GC regions and long sequencing items.

5.3.2 Illumina's pipeline

Illumina's RNA-sequencing pipeline consists of different steps.

5.3.2.1 Sample preparation

In sample preparation RNA is extracted and sheared into 300-600np fragments through sonication or enzymatic digestion.

5.3.2.2 Library preparation

During library preparation adapter sequences are ligated to the fragments. Barcoding is possible allowing for sample multiplexing.

5.3.2.3 Cluster generation

During cluster generation the library is amplified through PCR and loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters.

5.3.2.4 Sequencing

Sequencing is done by synthesis: sequencing reagents, including fluorescently labelled nucleotides are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated n times to create a read length of n bases.

- Paired ends are used for duplicates, splicing analysis and discovery of novel isoforms.
- Single ends are used for gene expression analysis.

5.3.2.5 Alignment

Reads are aligned to a reference genome or transcriptome or to a genomic region. Splice-aware aligner such as TopHat or STAR should be used. The alignments are refined according to coding sequences using known and predicted splice junctions.

5.3.2.6 Quantifying reads

Reads per gene are quantified. non-coding RNA are filtered out. Alternative splicing, overlapping genes and pseudogenes are dealt with. Different type of counts can be employed:

- Standard count: number of reads for transcript.
- CPM: counts scaled by the number of fragments sequenced N times one million. It allow to compare each transcript across different samples.

$$CPM_i = \frac{X + i}{N} \cdot 10^6$$

- TPM: measurement of the proportion of transcripts in the pool of RNA. It takes into account the length of the transcript. It is not used for direct comparison across samples, but for intra-sample normalization. Let $\tilde{l}_i$ the length of transcript i, then:

$$TPM_i = \frac{X_i}{\tilde{l}_i}\left(\frac{1}{\sum\limits_j \frac{X_j}{\tilde{l}_j}}\right)10^6$$

- RPKM: reads per kilobase of exon per million reads mapped. It is called FPKM for fragments.

$$RPKM = \frac{X_i}{\left(\frac{\tilde{l}_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\tilde{l}_i N}10^9$$

5.3.2.7 Normalization

To perform inter-sample normalization quantile normalization is used to make the distribution identical. The best normalization methods for differential expression are coupled with sophisticated approaches as very low expressing genes are tricky, for example those with $FPKM < 1$.

Chapter 6

Gene set enrichment

6.1 Introduction

6.1.1 Functional groups characterized by gene expression change

Functional groups characterized by gene expression change are:

- Gene sets: the set is scored depending on the expression level of its member genes.
- Network: modules satisfying some joint gene expression and topology requirement are identified.
- Pathways: they are scored exploiting gene expression and topology.

6.1.2 Gene-set enrichment analysis

Gene-set enrichment analysis is the breakdown of cellular functions into gene sets. Every set of gene is associated to a specific cellular:

- Function.
- Process.
- Component.
- Pathway.

Microarray or RNA-seq data can be related to gene sets in order to mine its functional meaning, to find which gene sets summarize at best gene expression patterns.

6.1.3 Ontology

Ontology normally represents knowledge as a set of concepts within a domain and the relationships among those concepts. It can be used to reason about the entities within that domain and may be used to describe the domain.

6.1.4 Controlled vocabulary

A controlled vocabulary provides a way to organize knowledge for subsequent retrieval but does not allow reasoning about the entities.

6.1.4.1 Ontology and controlled vocabulary uses

Ontologies and controlled vocabularies are heavily used in biological databases as they allow the organization of data within a database, providing a meaningful link between database structure and search queries.

6.2 Gene ontology

A gene ontology is a way to capture biological knowledge for individual gene products in a written and computable form.

6.2.1 Concepts hierarchy

A set of concepts and their relationships to each other is arranged as a hierarchy.

6.2.1.1 Molecular function

The molecular function describes activities that happen at a molecular level like catalytic or binding activity. This category includes the activities rather than the entities that are involved in an action and do not specify where, when or in which context the actions happen. Molecular functions can be executed by single gene products or complexes of gene products. Some examples are:

- Catalytic activity.
- Transport activity.
- Binding.
- Toll receptor binding.

6.2.1.2 Biological process

A biological process is a series of events resulting from multiple ordered groups of molecular functions. A biological process is different from a pathway, as gene ontology does not report the dynamics or the dependences that are required to describe a pathway. Some examples are:

- Cellular physiological processes.
- Signal transduction.
- Metabolic process of pyrimidine.
- Glucose transport.

6.2.1.3 Cellular component

A cellular component is linked to a component of a cell with the condition that is part of a larger object. Some examples are:

- Ribosome.
- Nucleus.
- Neuron parts.
- Internal nuclear membrane.

6.2.2 Ontology structure

An ontology is structured as an acyclic graph where terms can have more than one parent. Terms are linked by directed relationships like:

- Is part of.
- Regulates.

6.2.3 The gene ontology project

Originally in the gene ontology or GO project the hierarchies were completely independent, without links between them. From 2009 biological processes and molecular functions are linked, as biological processes are ordered assemblies of molecular functions. GO is required as there are inconsistencies in the human language: different concepts can have the same name. Furthermore it enables to interpret quickly large datasets. The aim of the GO project is to:

- Compile ontologies.
- Annotate gene products using ontology terms.
- Provide a public resource of data and tools for annotation.

A gene ontology annotation is a statement that a gene product has a particular molecular function, determined by a particular method and described in a reference.

6.3 Gene sets

6.3.1 Sources and types

Other than gene ontology there are other sources and types of gene-sets:

- Pathways (KEGG).
- Protein families and domains (PFAM).
- Predicted target of regulators like mRNA and transcription factors (MSigDB-c3).
- Protein-protein interaction modules.
- Gene expression:
  - Up and down regulation after treatment or in relation to disease (MSigDB-c2).
  - Co-expression across many conditions (MSigDB-c4).
- Genotype-phenotype association (DiseaseHub).
- Genomic position (MSigDB-c1).

6.3.1.1 Main resources

The main resources for these type of data are:

- Bioconductor.
- DiseaseHub.
- MSigDB.
- PathwayCommons.
- WhichGenes.

6.3.2 Differences between pathways and processes

From a biological perspective the difference between pathways and processes is philosophical. It is still worth speculating in a bioinformatics perspective because a gene is annotated for a GO biological process if the curators deem it significantly contributes to the process according to a number of evidences. Pathway include the wiring of genes and gene products, hence they rely on a more

intensive curation process. Some pathways include large ubiquitous actors such as the proteasome that may confound enrichment analysis, whereas the are usually absent from GO processes.

6.3.3 Enrichment test

An enrichment test combines a gene expression table with gene-set databases to build an enrichment table. In the enrichment table each gene set is associated with a p-value, that gives use the probability that the differentially expressed genes are part of the gene set. This is a two class design: genes can be ranked according to different statistics like fold change, log ratio or t-test and a selection by threshold can be performed. Also an expression matrix can be obtained with multiple condition employing for example ANOVA. A gene set can both overlap with significant gene and background genes. To test whether this overlap is significant it must be compared with random sampling of captured genes.

6.3.3.1 Fisher's exact test

Fisher's exact test calculates the exact probability of the table of observed cell frequencies in a table given that:

- The null hypothesis of independence is true.
- The marginal totals of the observed table are fixed.

It does not require to perform random sampling as it is based on the theoretical null-hypothesis distribution: the hypergeometric distribution. In particular it gives the probability that the overlap between significant genes and gene-set is greater than the one expected by random sampling the captured genes. Fisher's table is built such that:

Up regulated genes	Gene Set: No	Gene Set: Yes	
Yes	a	b	$a+b$
No	c	d	$c+d$
	$a+c$	$b+d$	

Table 6.1: Fisher's exact test table

Considering the table 6.1:

- a depicts the significant gene not in overlap.
- b the significant ones in overlap.
- c the background genes not in overlap.
- d the background gene in overlap.

Then the exact probability of the table is:

$$\frac{(a+b)! \cdot (c+d)! \cdot (a+c)! \cdot (b+d)!}{n! \cdot a! \cdot b! \cdot c! \cdot d!}$$

The p-value is calculated by summing all probabilities less than or equal to the probability of the observed table. Fisher's exact test can be used to evaluate the overlaps between gene-sets from databases. It is usually employed in threshold-dependent scenario and suffers from some limitations.

6.3.4 Whole-distribution - GSEA enrichment

Whole distribution methods like GSEA, gene set enrichment analysis, have been shown to be more stable and statistically powerful. Instead of excluding genes they are ranked according to a measure. This type of test can be used to compare 2 classes of genes and they can be correlated to a phenotype with a Pearson correlation test.

6.3.4.1 Process

First the enrichment score is calculated: genes are ranked according to a specific measure like fold change. Then the chance of observing a gene in a position towards the upregulated or downregulated region is computed. Based on the position in which genes are found, it can be established whether the event is random or not. Every gene present in a gene list gives a positive contribution, while every absent one a negative one. So, an high enrichment score means high local enrichment. After this the enrichment score distribution is generated for the null hypothesis using permutation and the empirical p-value is computed observing how many random data have enrichment score greater than the read ones. Then the FDR correction of the p-value is computed. The standard ranking metrics is:

$$\log_2 FC \cdot -10 \log_{10} pvalue$$

6.3.4.2 Permutations

Permutations setting have important implication and can be used only when biological replicates are very similar within classes or are well separated. When biological replicates tend to be dissimilar or stratified according to hidden experimental factors other whole-distribution enrichment methods need to be used.

6.3.4.3 Other uses of GSEA

The GSEA tool allow to perform more general types of analysis specifying designs:

- .GMT data format contains GO ID, description and information of the genes contained in the GO term.
- The gene expression table contains CDM or RPKM values for each gene.
- The expression phenotype file .cls determine which sample belongs to which class.

6.3.5 Gene set filter

Gene set for enrichment analysis are usually filtered by size. Large gene-sets are undesired if they are derived from gene ontology or other functional resources as the usually correspond to uninformative concepts. Small gene-sets are undesired as their statistics are noisy and may decrease the FDR of other sets.

6.3.6 Redundancy problem

There are many redundant gene-sets. For example gene ontology has a very large number of gene-sets, often with slight differences. Moreover different pathway databases have different but overlapping definitions of pathways. Globally it is useful to grasp the overlap relations between enriched

gene-sets. To do so a visualization framework beyond the enrichment table is needed. The redundancy problem can be handled by correcting for inter-redundancy and prioritizing the most enriched gene-sets or by visualizing gene-set overlap as a network with the EnrichMap tool. Considering up and down regulation a map can be built such that:

- Each node is a gene set.
- The size of the node is the number of genes.
- The colour represent differentially expressed genes.
- The edge thickness the overlap or similarity degree.

The network can be then clustered based on similarity values in order to identify redundancy sets.

Chapter 7

Network analysis

7.1 Introduction

7.1.1 Network definition

A network is a series of interconnected components, systems or entities. They can be used to describe a large variety of physical or abstract phenomena. Nodes can represent different entities and arcs any kind of interaction.

7.1.2 Networks in system biology

Networks are a relevant part of system biology, ome and omics. The objectives of system biology are:

- Comprehension at the level of system.
- Analysis of individual components.
- Analysis of interactions.
- Analysis of potential emerging properties.

7.2 Graphs

7.2.1 Definition

The analysis of a network structure should be done using appropriate mathematical methods. Graph theory is the tool able to extract information from the networks. A graph is a mathematical object defined by a set of nodes and arcs. It is denoted $G = (V, E)$, such that $V = \{v | v \ nodes\}$ and $E = \{(i, j) | i, j \in V\}$.

7.2.2 Magnitude of a graph

The magnitude of a graph is characterized by:

- The number of nodes $|V|$ or order of G.
- The number of arcs $|E|$ or size of G.

7.2.3 Degree of a graph

The degree of a node in a graphs is the number of arcs that are incident with that node. In a direct graph the out degree is the number of arcs going out of a node, while the in degree is the number of arcs directed into a node.

7.2.4 Weighted graphs

A weighted graph is a graph which arcs have associated a weight, generally defined as a weighting function:

$$w : E \to \mathbb{R}$$

7.2.5 Complete graphs

A complete graph is a direct or indirect graphs in which each pair of nodes is adjacent. If (u, v) is an arc in G then v is adjacent to node y. So, for a complete graph:

$$(u, v) \in E \forall u, v \in V$$

7.2.6 Paths

A path is a sequence of nodes $(v_1, v_2, \ldots, v_n)$ such that:

$$\{(v_1, v_2), (v_2, v_3), \ldots, (v_{n-1}, v_n)\} \subseteq E$$

7.2.6.1 Simple paths

A simple path is a path without repeated nodes:

$$P = \{(v_1, v_2), (v_2, v_3), \ldots, (v_{n-1}, v_n)\} \subseteq E \wedge \forall v_i, v_j \in P, v_i \neq v_j$$

7.2.6.2 Cycles

A cycle is path such that:

$$P = \{(v_1, v_2), (v_2, v_3), \ldots, (v_{n-1}, v_n)\} \subseteq E \wedge \forall v_i, v_j \in P, v_i \neq v_j \wedge v_1 = v_n$$

A graph is called cyclic if it contains a cycle, otherwise it is called acyclic.

7.2.7 Bipartite graphs

A bipartite graph is an indirect graph $G = (V, E)$ such that:

$$(u, v) \in E \Rightarrow u \in V_1 \wedge v \in V_2 \vee v \in V_1 \wedge u \in V_2$$

7.2.8 Graphs connections

An indirect graphs is connected if each pair of nodes is connected by a path.

7.2.8.1 Weakly connected graphs

A directed graph is weakly connected if for each pair of nodes (u, v) it exists a directed path from u to v or from v to u.

7.2.8.2 Strongly connected graphs

A directed graph is strongly connected if a directed path between each pair of nodes exists.

7.2.8.3 Sparse graphs

A graph is sparse if:

$$|E| \sim |V|$$

7.2.8.4 Dense graphs

A graph is dense if:

$$|E| \sim |V|^2$$

7.2.9 Subgraphs

A graph $G' = (V', E')$ is a subgraph of $G = (V, E)$ if:

$$V' \subseteq V \wedge E' \subseteq E$$

7.2.10 Trees

Trees are complete, acyclic graphs. A tree T spans $G = (V, E)$ if $T = (V, E')$ and $E' \subseteq E$.

7.2.10.1 Understanding if a graph is a tree

Let $G = (V, E)$ an indirected graph, then the following statements are equivalent:

- G is a tree.
- Each pair of nodes in G is connected by a unique single path.
- G is connected, but if a node is removed from E, the resulting graph is not connected.
- G is connected and $|E| = |V| - 1$.
- G is acyclic and $|E| = |V| - 1$.
- G is acyclic but if an edge is added to E, the resulting graph contains a cycle.

7.2.11 Clique

A clique in an indirect graph $G = (V, E)$ is a subset V' of the set of nodes V such that for each two nodes in V' it exists a unique arc that connects them. So the subgraph induced by v' is complete.

7.2.12 Isomorphisms

An isomorphism between two graphs G and G' is a biunivocal correspondence $f : V(G) \to V(G')$ such that u and v in G are adjacent if and only if $f(u)$ and $f(v)$ are adjacent in G'.

$$G = (V, E) \wedge G = (V', E'), f : V \to V' : (u, v) \in E \Leftrightarrow (f(u), f(v)) \in E'$$

If an isomorphism between two graphs can be built they are isomorphic.

7.2.12.1 Automorphism

An automorphism is an isomorphism on a graph onto itself.

7.2.12.2 Building an isomorphism

Building an isomorphism is an important problem in computer science with a complexity to be defined. Isomorphisms are useful when comparing the structure of two graphs.

7.2.13 Representing a graph

7.2.13.1 Adjacency matrix

A graph can be represented by and adjacency matrix. Let $G = (V, E)$, then the adjacency matrix is a matrix $|V| \times |V|$ such that $A_{uv} = w$ if $(u, v) \in E$ and $A_{uv} = 0$ if $(u, v) \notin E$. It grows quadratically with the number of nodes and each arc is represented two times, so it will be symmetric for indirect graphs. It is not efficient for sparse graphs.

7.2.13.2 Adjacency lists

The adjacency list of a graph $G = (V, E)$ is an array of lists. Each node has a list of the nodes to which is adjacent. The space is proportional to $|V| + |E|$ and each arch is represented two times. It is not efficient for dense graphs.

7.3 Networks properties

Theory of graphs or networks is focused on three different abstraction levels.

7.3.1 Analysis of single elements

In the analysis of single elements the more important nodes are identified. Centrality measures are indicators of the importance of a node in a graph or network. Different centrality measures are available.

7.3.1.1 Degree centrality

Degree centrality is measured as:

$$DC(n) = degree(n)$$

Nodes with high DC are defined as hubs and usually they have important roles in a network.

7.3.1.2 Closeness centrality

Closeness centrality is measured as:

$$CC(n) = \frac{|V|}{\sum_{n'} d(n, n')}$$

Where $d(n, n')$ is the length of the shortest path between n and n'. Nodes with high closeness centrality can access quickly other nodes and have rapid cascade effects on other nodes.

7.3.1.3 Eccentricity centrality

Eccentricity centrality is computed as:

$$C_s(n) = \frac{1}{\max\{dist(u, n) : u \in V\}}$$

The eccentricity is a measure of the centrality index. Is calcualted by computing the shortest path between the node v and all other nodes in the graph and then considering the longest shortest path. Higher eccentricity means that the node is proximal to other nodes.

7.3.1.4 Betweenness centrality

Betweenness centrality is computed as:

$$BC = \sum \frac{\sigma_{st}(n)}{\sigma_{st}}$$

Where σ_{st} is the number of shortest paths between s and $t>$ Nodes with high betweenness centrality can have greater control in the propagation of an effect in a whole network.

7.3.1.5 Subgraph centrality

Subgraph centrality is computed as:

$$SC(n) = \sum_{k=0}^{\infty} \frac{\mu_k(n)}{k!}$$

It accounts for the participation of a node in all sub graphs of the networks. $\mu_k(v)$ is the number of closed walks of length k starting and anding in node v.

7.3.1.6 Eivenvector centrality

Eigenvector centrality is computed as:

$$EC(n) = \frac{1}{\lambda} \sum_{t \in M(n)} EC(t) = \frac{1}{\lambda} \sum_{t \in V} a_{v,t} EC(n)$$

Where λ is a constant, a is the adjacency matrix and $M(v)$ are the neighbours of v. It can be written in vector notation as $A\vec{x} = \lambda\vec{x}$. It is a measure of the influence of a node in a network. The hogh-scoring nodes contribute more to the score of the node in question. A high eigenvector score means that a node is connected to many nodes who themselves have high scores.

7.3.2 Analysis of groups

During the analysis of groups groups or nodes are identified that have cohesion characteristics. A typical analysis is network clustering. First a similarity function between nodes is defined in terms of network topology. Then a method to group the nodes in terms of their similarity is applied.

7.3.3 Analysis of network

During the analysis at the level of network topological properties that are global to the network are identified.

7.3.3.1 Clustering coefficient

The clustering coefficient measures the degree at which nodes in a graph tend to e connected.

7.3.3.1.1 Local clustering coefficient A local clustering coefficient indicates how much the neighbours of a node are distance from being a clique. The local clustering coefficient $LCC(n)$ of a node n is given by the number of links between the members of $N(n)$ divided by the number of potential links between them:

$$LCC(n) = \frac{2|\{(u,v)|u \wedge v \in N(n) \wedge (u,v) \in E\}|}{|N(n)|(|N(n)|-1)}$$

7.3.3.1.2 Global clustering coefficient A global clustering coefficient GCC is the mean of all LCC computed across nodes in a graph and is called average clustering coefficient. The same measure can be calculated counting the number of closed triplets in the network divided by the total number of triples. $GGC = LCC$ coincide when using the weighted mean. It gives the intensity of the phenomena in the graph

7.3.3.2 Average diameter

The distance between two nodes is the least number of arcs that should be crossed to go from one node to another. The shortest path is the path that satisfies this criteria. The average diameter AD of a graph is the average shortest path computed across all pair of nodes of a graph.

7.3.3.3 Degree distribution

Let $P(k)$ be the percentage of nodes with degree k in a graph. The degree distribution is the distribution of $P(k)$ computed on all k. It can be defined as the probability of a node to have a degree k. Different distributions indicate different network topology, for example random networks or scale free networks. In scale free networks multiple hubs are present and is a hierarchical structure with an exponential degree distribution.

7.3.4 Small world effect

Small world networks have low AD and high GCC. Comparing different graphs mean to combine their different properties like randomness, modularity and heterogeneity. In particular in a regular graph each node has the same number of neighbours. In a random graph there is low AD. In a small world graph GCC tends to be similar to a regular graph or to a bigger random graph and AD is similar to a random graph.

Chapter 8

Genome wide association studies

8.1 Types of studies

8.1.1 Descriptive studies

Descriptive studies are studies in which an hypothesis is generated. Then the patterns of disease occurrence in relation to variables such as person, place and time are studies. They are often the first step or initial enquiry into a new topic event, disease or condition. They typically estimate the frequency and the magnitude of the event analysed.

8.1.2 Analytical studies

An analytical study is one in which action will be taken on a cause system to improve the future performance of the system of interest. The focus is to test an hypothesis to produce predictive data. In particular they are used to identify factors that are associated with a disease or to quantify the risk of these factors.

8.1.3 Cohort studies

Cohort studies are a type of analytical studies that involve a cohort. A cohort is a well-defined group of individuals who share a common characteristic or experience. For example individual exposed to a drug, vaccine and pollutant.

8.1.3.1 Prospective cohort studies

Prospective cohort studies potential exposure has already occurred while outcomes have yet to occur. Participants are grouped according to past or current exposure and a follow-up in the future determine whether the predicted outcome occurs.

8.1.3.2 Retrospective cohort studies

Retrospective cohort studies both exposure and outcomes have already occurred. Participants are grouped according to past exposure and certain characteristics and are compared for a particular outcome.

8.1.3.3 Measures of associations

In cohort study a 2×2 table can be built to determine, for example, the effect of exposure to a certain event on disease presence. This type of table is described on table 8.1.

	Disease	
	Yes	No
Exposed	a	b
Not exposed	c	d

Table 8.1: Cohort study's table example

Then from this table the percentage of individuals exposed harbouring the disease can be computed:

$$I_e = \frac{a}{a+b}$$

Also the percentage of individuals not exposed and not harbouring the disease:

$$I_{ne} = \frac{c}{c+d}$$

From this two measure the risk excess RE and the relative risk RR can be computed:

$$RE = I_e - I_{ne} \qquad RR = \frac{I_e}{I_{ne}}$$

The risk excess determine how the exposure betters, or worsen the chance of presenting a disease. The relative risk instead determine the nature of the exposure's event:

- $RR < 1$ indicates a protective factor.
- $RR \sim 1$ indicates an absence of risk.
- $RR > 1$ indicates a risk factor.

8.1.3.3.1 Statistical confidence From the table also a significance of association and a precision of association can be determined. The first requires a p-value, that determine how unlikely it is that the events observed arise by chance. The second requires a confidence interval CI, which can be with a certain level of confidence α, usually around 80% and 95%, that determine how many times the correct value will be inside the interval when the measurement or the estimate will be replicated many times. Usually the relative risk indicates the amount of random error around the point estimate.

8.1.4 Case-control study

The purpose of a case-control study is typically to study rare diseases or multiple exposures that may be related to a single outcome. Participants are selected based on outcome status:

- Case subjects have outcome of interest.
- Control subjects do not have outcome of interest.

This type of study is usually preferred when funding is limited.

8.1.4.1 Measure of association

The same table as a cohort study is built (as in table 8.1), but the measure of association is different. First the odds of exposure in cases is computed:

$$\frac{\frac{a}{a+c}}{\frac{c}{a+c}} = \frac{a}{c}$$

Then the odds of exposure in control:

$$\frac{\frac{b}{b+d}}{\frac{d}{d+b}} = \frac{b}{d}$$

Finally the ratio of this two measure is the odds ratio OR, the measure of association for a case control study:

$$OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}$$

OR is in relationship with RR following the equation:

$$OR = \frac{RR(1 - R_0)}{1 - RR \cdot R_0}$$

Where R_0 is the frequency of the disease in the not exposed population.

8.2 GWAS

8.2.1 Objective of GWAS

The objective of a GWAS is to find connections between a phenotype known to be heritable and whole-genome genotype. GWAS were developed in 2004, mainly thanks to the HapMap project, which unravelled the existence of linkage disequilibrium blocks, which allowed the exploitation of tag SNPs. Specific goals are distinct:

- Identification of statistical connections between points or areas in the genome and the phenotype. The hypotheses are driven for biological studies of specific genes or regions in specific contexts.
- Generation of insights on genetic architecture or phenotype. In fact a phenotype could be due to many small genetic effects dispersed across the genome or due to few large effects concentrated in one area. An example in the second case is the MHC or major histocompatibility complex, a group of genes involved in the mechanism of immune defence.
- Build statistical models to predict phenotype from genotype.

8.2.2 Main application of GWAS

The main application of GWAS are:

- Identification of susceptibility variants for novel biological insights like:
  - Therapeutic targets.
  - Biomarkers.
- Improved measures of individual aetiological (the manner of causation of a disease or condition) processes for personalized medicine.

8.2.3 GWAS methodology

A typical GWAS methodology can be described as:

- Collect n subjects with known phenotype, usually $n \in [10^3; 10^4]$.
- Measure each one in m genomic locations representing common variation in the whole genome. Typically these are SNPs. Usually $m \in [10^5; 10^6]$, but recently with whole genome sequencing $m = 3 \cdot 10^9$.
- The data can be thought as a matrix X of dimension $n \times m$ with subject as rows and SNPs as columns. This matrix is built such that $X_{ij} \in \{0, 1, 2\}$, representing the genotype at a single column. Moreover a vector of phenotypes Y_n can be given.

Having collected the data and having computed the matrix X and the vector Y the first task is association testing: finding SNPs (column of X) that are statistically associated with Y. This can be thought of as m separate statistical tests run on the matrix X.

8.2.4 Single nucleotide polymorphisms

A SNP is defined as a single base variation in a DNA sequence. They are classified according to the minor allele frequency MAF:

- Common SNPs have $MAF \geq 1\%$.
- Rare SNPs have $MAF < 1\%$.

8.2.4.1 SNPs frequency

In the human genome SNPs compose the 0.1% and are what makes human unique. These variants can be:

- Harmless.
- Harmful.
- Latent.

They can lie in coding regions, but the majority of them are found in non-coding one. Thy are present between in 1 every 1000 bases or 1 every 100-300. The abundance of SNPs and the ease with which they can be measured make them very important. Two thirds of SNPs modification are from a C to a T. They are typically found in non-coding regions and are found less in less conserved regions. In coding regions synonymous SNPs (that don't change the structure of the coded protein) are more common.

8.2.4.2 SNPs effects

SNPs can have different effect on the genome:

- When they are found near a gene they can act as marker for that gene.
- SNPs in regulatory regions can modify transcription influencing the binding of transcription factors.
- SNPs in coding regions can modify the stricture of codified protein.

8.2.4.3 Nucleotide diversity

Nucleotide diversity measures the degree of polymorphism in DNA sequences or in a population. It is defined as the average number of nucleotide differences per site between two DNA sequences in all possible pairs in the same population and is denoted by π It is estimated as:

$$\hat{\pi} = \frac{n}{n-1} \sum_{ij} x_i x_j \pi_{ij} = \frac{n}{n-1} \sum_{i=2}^{n} \sum_{j=1}^{i=1} 2 x_i x_j \pi_{ij}$$

Where:

- x_i and x_j are the respective frequencies of the ith and jth sequences.
- n is the number of sequences in the sample.
- π_{ij} is the number of nucleotide differences per nucleotide site between the ith and jth sequences.
- $\frac{n}{n-1}$ is a normalization factor that makes the estimator independent on how many sequences are sampled.

8.2.4.3.1 Hardy-Weinberg equilibrium In a population with genotypes BB, bb and Bb, if:

- $p = freq(B)$.
- $q = freq(b)$.

The frequencies of the genotypes are then:

- $freq(BB) = p^2$.
- $freq(bb) = q^2$.
- $freq(Bb) = 2pq$

In a condition of equilibrium and will not change considering:

- No mutations.
- Population of infinite size.
- Random coupling.
- No emigrations.
- No selective pressure.

8.2.4.4 Linkage disequilibrium

Two genetic loci are said to be in linkage disequilibrium LD when there is a non-random association of alleles at different loci in a given population. It usually indicates that two alleles are near and in mammalians LD is usually lost at around $100Kbp$. Let:

- p_A be the frequency of an allele A in a genomic locus.
- p_B be the frequency of an allele B in another genomic locus.

The association between allele A and allele B is random when:

$$p_{AB} = p_A p_B$$

8.2.4.4.1 Measuring linkage disequilibrium The coefficient D is a measure of linkage disequilibrium. It is defined for two biallelic loci with alleles A and a at the first locus and B and b at the second one as:

$$D_{AB} = p_{AB} - p_A p_B \qquad D_{Ab} = -D_{AB} \qquad D_{ab} = D_{AB}$$

Being LD a property of two loci and not of their alleles, it is the magnitude being of interest, not the sign. The magnitude does not depend on the choice of the allele, and the range of D changes with allele frequency. Knowing that p_{AB} is smaller than p_A and p_B and that the frequencies cannot be negative:

$$-p_A p_B \wedge -p_a p_b \leq D_{AB} \leq p_a p_B \wedge p_A p_b$$

The possible values of D depend on the allele frequencies and as such is difficult to interpret. Because of this it is normalized in D':

$$D'_{AB} = \begin{cases} \frac{D_{AB}}{\max(-p_A p_B, -p_a p_b)} & D_{AB} < 0 \\ \frac{D_{AB}}{\min(p_a p_B, p_a p_b)} & D_{AB} > 0 \end{cases}$$

8.2.4.4.1.1 Measuring LD with r^2 To measure LD with r^2 two random variables are defined:

- X_A such that $X_A = 1$ if allele at locus 1 is A and $X_A = 0$ if the allele is a.
- X_B such that $X_B = 1$ if allele at locus 2 is B and $X_B = 0$ if the allele is b.

Or:

$$X_A = \begin{cases} 1 & allele = A \\ 0 & allele = a \end{cases} \qquad X_B = \begin{cases} 1 & allele = B \\ 0 & allele = b \end{cases}$$

Then the correlation between the two random variables can be defined as:

$$r_{AB} = \frac{Cov(X_A, X_B)}{\sqrt{Var(X_A)Var(X_B)}} = \frac{D_{AB}}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$$

And:

$$r^2_{AB} = \frac{D^2_{AB}}{p_A(1-p_A)p_B(1-p_B)}$$

This measure is usually employed as it is always a positive value.

8.2.4.4.1.2 Classifying LD LD can be classified according to the D' and r^2 values:

- When $D' = 1$ there is complete LD.
- When $r^2 = 1$ there is perfect LD.

Perfect LD implies complete LD. There are situations in which $D' = 1$ and r^2 is low, so usually both measures are reported.

8.2.4.4.2 Haplotypes An haplotype is a set of linked SNPs on the same chromosome. Genotypes don't report informations about the connections of alleles at different SNPs loci, so there could be several possible haplotypes for the same genotype. An haplotype block is defined as a cluster of SNPs in linkage disequilibrium and an haplotype boundary as sequences of blocks with strong internal linkage disequilibrium but no linkage disequilibrium between them. They usually reflect genetic recombination hotspots.

8.2.4.4.3 Tag SNPs Tag SNPs are a set of SNPs that captures most variations in haplotypes, removing redundancy.