Bioinformatic resources

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

July 7, 2022

Contents

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 algorithms 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.

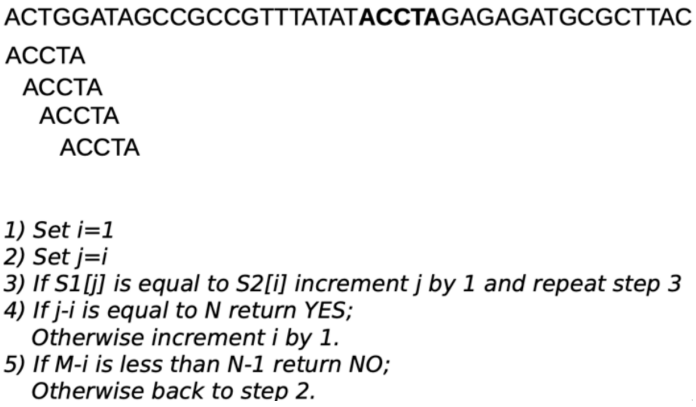

Figure 1.1: Naïve implementation of a substring search algorithm

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 the main problems related with biological databases 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.

3.2 Popular databases

3.2.1 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.

3.2.2 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.

3.2.3 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.

3.2.4 Others

Other examples of derived databases are:

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

3.2.5 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.

3.2.5.1 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.

3.2.5.1.1 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.

3.2.5.2 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 interface 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).

3.2.5.2.1 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 sites and 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 hidden Markov models.

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 sequences. The elements in the matrix represent all possible bases at each position. Example of these matrices are:

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

PFM

	1	2	3	4	5	6	7	8	9
A	0	6	0	3	4	0	6	5	1
C	0	0	1	0	1	0	0	0	2
G	0	0	0	3	0	0	0	1	2
T	6	0	5	0	1	6	0	0	1

PPM

	1	2	3	4	5	6	7	8	9
A	0	1	0	0.5	0.67	0	1	0.83	0.17
C	0	0	0.17	0	0.17	0	0	0	0.34
G	0	0	0	0.5	0	0	0	0.17	0.34
T	1	0	0.83	0	0.17	1	0	0	0.17

Figure 4.1: Example of PFM and PPM

4.2.2.1 Populating a position frequency matrix

In PFM, columns represent the position of the sequences and on the rows the nucleotide we expect to find on that position. So, when populating the matrix each column will contain the number of

bases in the position that are counted. In this way the frequency at which a specific base at each specific position of multiple aligned sequences is counted. 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.
- δ is an indicator function, such that:

$$\delta(X_{i,j} = k) = \begin{cases} 1 & if X_{i,j} = k \\ 0 & otherwise \end{cases}$$

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. PPM are useful because they are a normalization of the PFMs, making different matrices comparable with each other. Its cells are computed as:

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

Given:

- k the set of all symbols in the alphabet.
- N the number of aligned sequences.
- j iterates over the length of the sequence.
- δ is an indicator function, such that:

$$\delta(X_{i,j} = k) = \begin{cases} 1 & if X_{i,j} = k \\ 0 & otherwise \end{cases}$$

4.2.2.3 Assessing the probability that a sequence belong to a PPM

In PPMs, probabilities are calculated for each position independently. So, PPMs make the assumption that there is no statistical dependence between positions in the pattern. Dependence is not base specific but transcription factor specific. According to this PPMs can be considered models of a pattern that refers to a specific transcription factor. Meaning that searching if a function belongs to a PPM is equivalent to say how close the sequence is to that model. To assess the probability for a sequence S to belong to a PPM the probabilities for each base i found at each position j are multiplied:

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

According to this equation if a sequence contains a base not read yet found, the probability would be zero. This would be incorrect, so the matrix needs to be corrected so that the probability of observing another base in a position is no longer zero.

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 integer or real amount added to the number of observed cases in order to change the expected probability.

$$p_{i,empirical} = \frac{x_i}{N} \qquad p_{i,\alpha\text{-}smoothed} = \frac{x_i + \alpha}{N + \alpha d}$$

Where d is the number of observations added during Laplace smoothing.

4.2.2.4.2 Adding a background model Another way to correct PPMs is to add a background model. A background model reports the probabilities of observing a specific base at a specific position. Generally, in any position of the sequence each sequence has an expected probability of being observed of 25%. However, background models can vary, for example amino-acids instead of nucleotides or the GC content of the organism of the sequence can be exploited to build a more accurate model. With this a new matrix is computed as:

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

Where b represent the background model. 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 Assessing sequences' scores in a PWM matrix

To assess how close an input sequence is to the model implemented to a PWM a score should be computed. This is done by summing up the scores found in the columns for each base. The score indicates how much the sequence is different from a random sequence. This score is usually the measure used to search for positions where a putative binding for a transcription factor can be found. Indeed, for positions where the score, it might suggest that in that specific region a specific TF, embedded in the model, can bind to. Again, we assume we have pseudocounts embedded in our matrices to avoid zero scores.

	1	2	3	4	5	6	7	8	9
A	-inf	1.38	-inf	0.69	0.99	-inf	1.38	1.20	-0.39
C	-inf	-inf	-0.39	-inf	-0.39	-inf	-inf	-inf	0.31
G	-inf	-inf	-inf	0.69	-inf	-inf	-inf	-0.39	0.31
T	1.38	-inf	1.20	-inf	-0.39	1.38	-inf	-inf	-0.39

Figure 4.2: Example of a PWM

4.3 Hidden Markov model - HMM profile

Hidden Markov Models (HMM) are widely used in bioinformatics, as they provide for a useful tool when dealing with observational data. They are used for example to simulate signalling pathways. A Markov chain is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules, and it represents the theoretical base for HMM. The possible future states are fixed and not based on how the process arrived at its present state. Meaning, the model has no memory of its past states and it does not matter how much time is spent one one state or the other. In a Markov chain, the state is directly visible to the observer: state transition probabilities are the only parameter. In a HMM instead, the state remains transparent, while the output, dependent on the state, is easily obtainable.

To formalize these concepts: 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 $T = P(X|i)$, the probability of X emission in state i, the emission is what the user sees.

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

The simplest way to build a model of a pattern representing a motif using HMM, is to build states that represent the different positions in the sequence and to associate some emission probabilities given each state. Given a set of sequences a set of transition, with each one representing a position in the sequence have to be modelled. In each state each nucleotide with probabilities proportional to the probabilities calculated in the position probability matrix are found, as shown in figure 4.4.

When assessing the probability that a sequence is generated by a HMM the probability of each state the sequence leads are summed up. This simple HMM can be extended with a background HMM. HMM for motif analysis become useful when dealing with indels or missing information, a very common scenario when performing multiple sequence alignment. Indels are modelled with edges connecting non-sequential states. The probability represented by the edge is proportional to the read count supporting the indel.

4.3.2 Match a sequence to a HMM profile

If the sequence S is ATG, how many ways ways we can navigate this profile assuming match, insertion and deletion to match the seuence to the model? The simplest way is to assume all matches

$$P(S|Model) = 0.8 * 0.1 * 0.7 * 0.2 * 0.6 * 0.1 * 0.9 = 0.0006048 \texttt{ BMMME}$$

But it is not the only possibility, for example:

$$P(S|Model) = 0.8 * 0.1 * 0.2 * 0.4 * 0.25 * 0.75 = 0.0012 \texttt{ BMDMIE}$$

After generating the model, the probability that a sequence is generated by a HMM can be computed as:

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

Where:

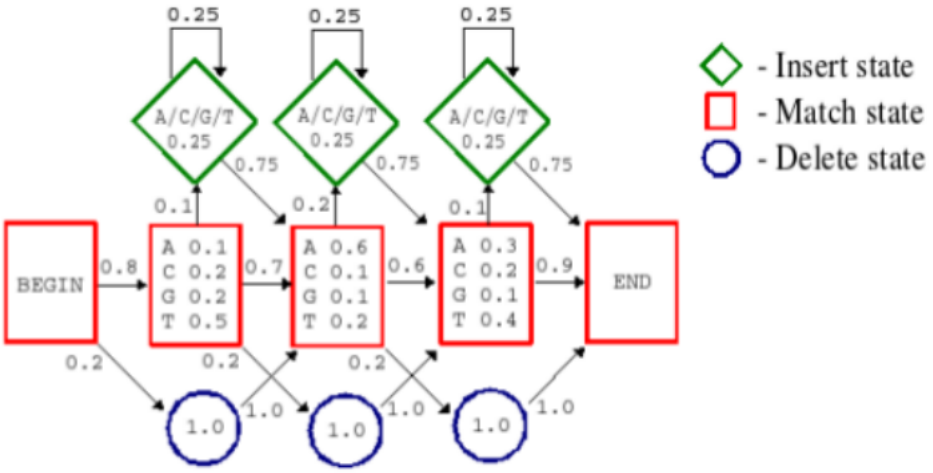

Figure 4.3: Match a sequence to a HMM profile

- S is the sequence.
- w are the probabilities parameters.
- π are all possible paths.

Finding the path with the highest probability means to find the best alignment to the HMM profile. Computing all paths in a brute-force way is clearly inefficient, but efficient algorithms to compute this probability like the Forward-Backward and the Viterbi exist.

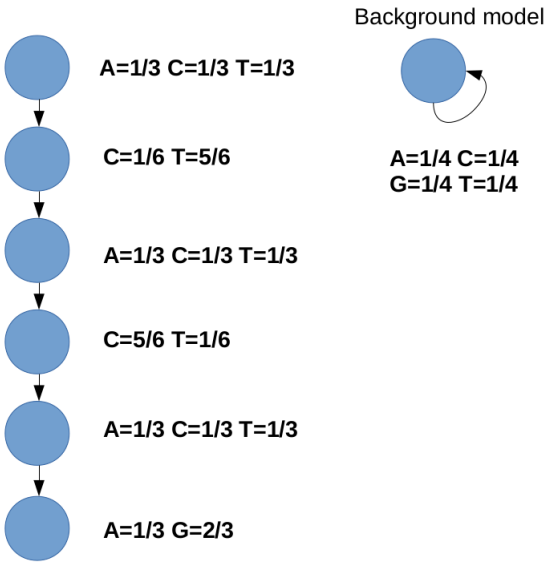

Figure 4.4: Example of a HMM and background HMM models

4.4 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

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

4.5 Motif identification

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

4.5.1 Finding known motifs - pattern matching

Pattern matching is the problem of finding known motifs, for example seeing if a binding of a protein X to an upstream region of a gene Y 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.

4.5.1.1 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.

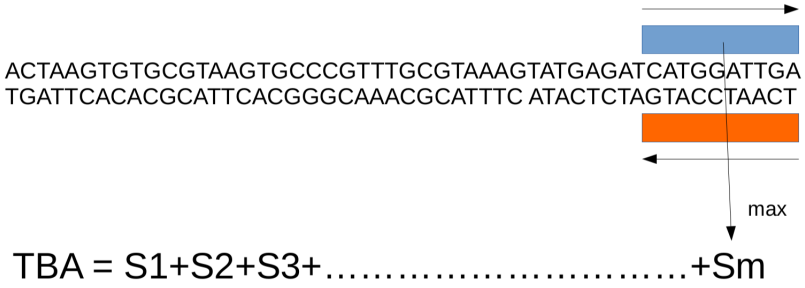

Figure 4.5: TBA method

4.5.2 Finding de novo motifs - pattern discovery

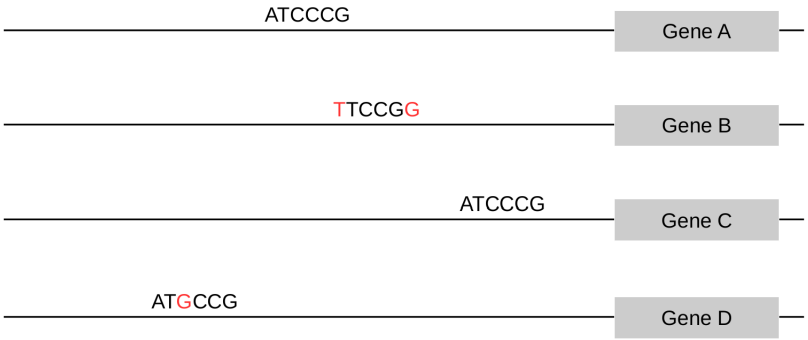

Figure 4.6: General scenario of a *de novo* motif identification

Pattern discovery is the problem of finding de novo motif, for example finding the motifs upstream of a specific gene Y and which is the structure of these motifs. Given a set of sequences, the objective is to find the most represented motifs. Using the MEME suite it is possible to identify new sequences and through Jaspar (or other databses) 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.

4.5.2.1 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.

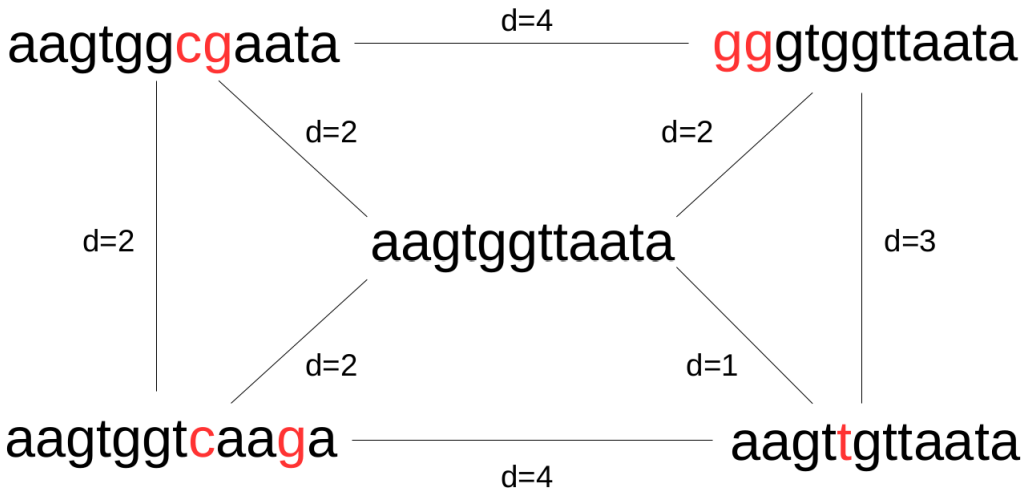

Figure 4.7: Consensus sequence

4.5.2.2 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-mer 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.

4.5.2.3 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.

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 analyzed to uncover characteristics of diseases, development and dynamic responses to stimuli.

5.1.2 Differential gene expression (DGE)

Gene expression profiles are extremely heterogeneous, since they vary based on the individual, tissue, condition and cells of origin. Variability is mainly due to alternative splicing and different regulation. 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.

For example, a basic experiment can be done by simply measuring two expression profiles, subtract the overlap, obtain the difference and interpret it. Two things to pay attention to are negative and positive controls and the range of variability within samples.

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 to have trustful data.
3. Running the experiment.
4. Image processing done by a machine.
5. Low-level analysis: data pre-processing with normalization.
6. High-level analysis: actual data 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

Even if the cons of high-throughput methods are important obstacles for performing complete, trustful analysis, these methods are becoming always better (less errors), cheaper and the computational side of the experiment is also improving to make the best out of the output data.

5.1.3 Main technologies

- Microarray technology: used more in the past.
- RNA-seq technology: today's best approach to DGE analysis, as it is becoming more accessible and allows for transcriptome sequencing and quantification of mRNA transcripts.

5.1.4 Databases

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

5.1.4.1 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.

5.1.4.2 ArrayExpress

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

5.1.4.3 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 were the first high throughput technology. They are useful to investigate a variety of omics data, for example:

- Genomic and transcriptomic profiles.
- DNA-protein interactions.
- The methylome.
- The microbiome.

Data interpretation is subject to specific computational analyses. Microarrays are considered high-throughput because they 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, the microarray. Microarrays are basically a glass surface hosting spots containing redundant information about the DNA.

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, to ensure a strong hybridization.

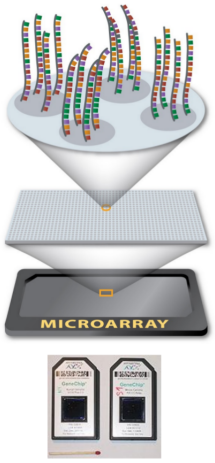

Figure 5.1: Structure of a microarray

Moreover microarrays can have different numbers of channels:

- 2 channels: test and control samples are labeled with different fluorophores. It is a comparative technique implemented at the experiment stage of the analysis. After hybridization to the microarray the two samples have different colours, for example red for cancer and green for normal cells), which will be be read by a reading system.
- 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 (figure 5.2).

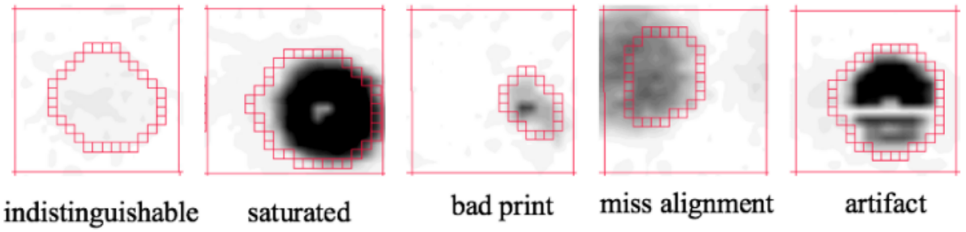

Figure 5.3: Result of segmentation and analysis of the result

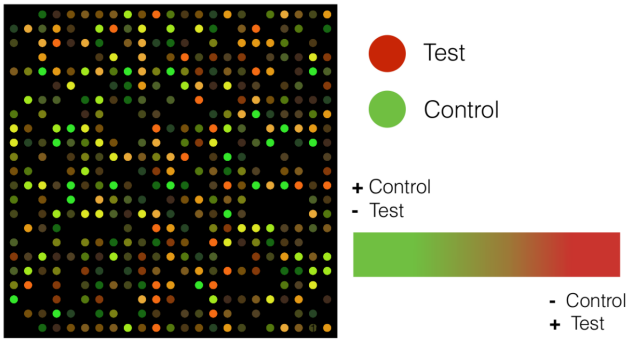

Figure 5.2: The image is in grey scale but is usually represented in a red/green scale that represents the light emitted by the two fluorophores

If each of the spot contains probes from both samples, the resulting color will be the combination of the quantity of DNA from the two conditions that hybridized in that specific spot. So, the resulting color in each spot 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, after position the images on a grid, is a **segmentation** analysis: the shape and patterns inside the data is analysed to assess the signal quality for each spot (figure 5.3).

If the quality assessed by the segmentation is fine, the analysis of the flourescent signal is performed. The background and foreground are identified to correct for noise generated by the former to better determine the actual data signal. One of the standard method for signal correction 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 (IQR) on the distribution in order to find feature, exclusion zone and background. Finally, each microarray is provided with a specific design file. After signal correction, the fluorescence of each spot is estimated and the relative expression for each gene (spots can be viewed as genes or portion of the genome) is interpreted thanks to annotation information.

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

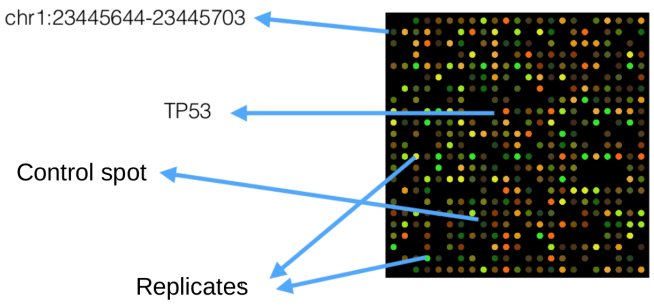

Figure 5.4: Image analysis: each microarray is provided with a specific design file

5.2.5 Expression microarrays

There are different types of arrays, but in any category, for example expression microarrays, there have sub-categories like for example, exon, gene or 3' array. Collecting all the signal coming from the different pieces or genes in different conditions,the quantity representing a transcript is identified for each condition, allowing to know if one condition will generate a differential expression of the gene. The results of micorarrays' experiments are extremely specific, making it difficult to integrate different analysis.

5.2.6 Data pre-processing

After the experiment is done, data pre-processing is needed to reduce errors introduced during the experimental process. It consists typically of:

- Background subtraction: eliminates background noise.
- Normalization (intra/inter sample): all samples are brought into a similar range of distribution, to reduce the effect of:
  - Unequal quantity of starting sample.
  - Differences in labeling efficiency.
  - Differences in detection efficiency.
  - System biases.
- Summarization: aggregation 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.

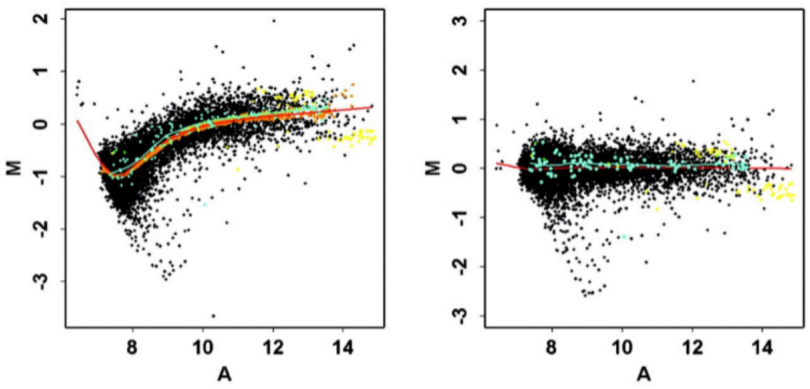

Figure 5.5: For the majority of probes, M should be equal to zero. On the right: LOESS normalization.

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 MA normalization MA normalization is useful to identify systematic intensity dependent biases in the data. MA normalization follows empirical observations known to be present in most microarray technologies. The ratio of signal might depend on the average signal intensity measured across different channels, normalization gets rid of this bias. The function of dependence can be fitted to a polynomial regression like Loess to obtain normalization to make the plot more informative.

$$Loess : y = f(A) \quad M' = M - y$$

Expected $\hat{M}$ is 0 among all observable intensities. This MA methodology is also very informative since it allows for performing basic DEG analysis.

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 (discussed below) and MAS5.
- Rudimentary methods: MAS4, LOESS.

5.2.6.3 Robust Multi-array average

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

5.2.6.3.1 Background correction Background correction removes local artifacts and noise. The probe measure data is assumed as a combination of background noise in a normal distribution and signal in an exponential distribution.

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

By assuming strictly positive distribution of signal background, the corrected signal is also positively distributed. Background correction is performed on each array separately. μ, σ and λ are estimated separately in each chip using the observed distributions of PMS. Introducing these parameters in the formula above, 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. Quantile normalization is widely used in many "inter-sample" normalization tasks.

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

After pre-processing, the downstream analysis can take place, as the previous steps pre-process a trustful matrix. 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 to test is that *there exists no difference between the gene expression in the two conditions.* The comparison between the samples is done using the ratio between the test and the control samples. The ratio should not differ in case of null hypothesis validity. These ratios are also defined as *fold changes*:

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

Fold change, at a gene level, is a ratio that represents how much the quantity of a gene changes in different conditions. Because ratios are not symmetric with respect to 1 the statistics are not easy to analyze, so the log-ratio is often used. The log ratio of the null hypothesis should be 0.

$$Ratio = [0, inf] \ \rightarrow \ Log(Ratio) = [-inf, inf]$$

5.2.7.1 Replicates

The fold change and the log-ratio are measures that give an immediate idea of what is the difference in expression at gene level. It can be extended to the level of the entire gene expression profile. However, random deviations from the null hypothesis can happen, which can lead to the introduction of errors introducing false positives or negatives. To improve statistical confidence in the results, we should have 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. Ideally a matrix containing many instances for each condition should be obtained.

5.2.7.2 Statistical tests

Many tools for statistical analyses have been developed, which need to take into consideration, among the other parameters, noise, several replicates, thousands of genes and the multiple hypothesis problem. 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. At single-gene level, basic statistics, like the T-test can be applied.

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, an extension of t-test. 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:

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

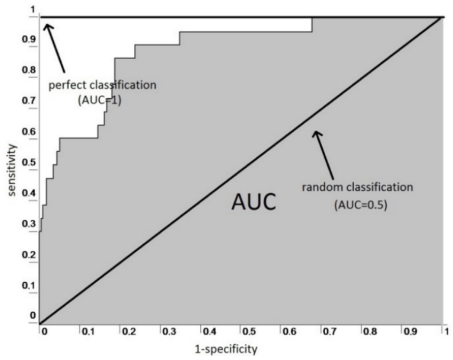

Figure 5.6: Allows to find genes that better discriminate between the two conditions

5.2.7.3 Correction methods

Correction methods are used to correct p-values in the presence of **multiple null hypothesis**. Correction for this problem is extremely important. Say we have a set of hypothesis we want to test simultaneously, like for a microarray setting. Considering n multiple hypothesis the probability of observing one significant result due to chance is:

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

Some examples of correction methods 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 $\frac{k \cdot a}{N}$

5.2.7.3.1 ANOVA analysis ANOVA analysis is a method for multiple hypothesis analysis and 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.

5.2.7.3.2 2-way ANOVA 2-way ANOVA is another multiple null hypothesis analysis method: it 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.

5.2.7.4 Receiver Operating Characteristic

At the end, of the analysis of gene expression data a ROC curve is generated. The receiver Operating Characteristic finds 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.

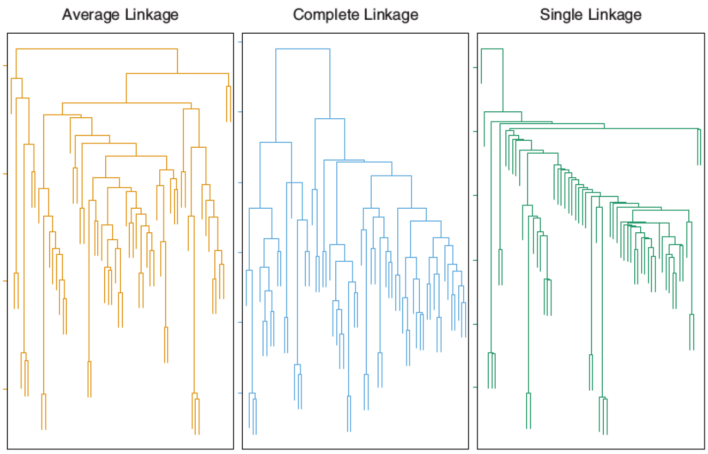

Figure 5.7: Different linkage methods. The length of the branches represents the distance between samples.

5.2.7.5 Clustering

The objective in differential gene expression is to look for genes that behave differently between samples, either up- or down- regulated. In clustering, the deregulated genes are aggregated together, to see if they form some patterns, maybe they come from the same pathways which is overall deregulated. 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 the definition of a distance (dissimilarity) measure between observation pairs. The algorithm works as follows:

- Starting with N observations and a distance measure like euclidean distance or Pearson correlation factor for all $\frac{N(N-1)}{2}$ pairs. At the beginning, 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 (containing many observations), as the distance measure is only used between single samples (clusters with one instance).

5.2.7.5.2.1 Complete linkage In complete linkage maximal intercluster dissimilarity is reached. Computes all pairwise dissimilarities between the observations in cluster A and in cluster B, and record the **largest** of these dissimilarities.

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

RNA sequencing is a next-generation sequencing approach that sequences the cDNA from the mRNA component. RNA-seq is considered a high-throughput technology because the whole transcriptome can be sequenced in one shot. The whole transcriptome can be compared against the whole transcriptome of another sample. It is very cheap, sequencing ~ 400 gigabases per flow cell, but it has an error rate up to 1%, issues with AT and GC regions and long sequencing items.

5.3.1 Issues

RNA-seq shares some litigations with microarray technology. For example, one needs availability of biological replicates (at least 3) and to take care of batch effects to derive confident conclusions. RNA-seq also requires the selection of a sequencing protocol after library preparation. For example, the choice of PE over SE, controlling the read length (100/150 bp) and sequencing depth (20-50M per sample for differential expression).

5.3.2 Illumina's protocol

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, stored in a FASTQ file, 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. Splice-aware alignment is nowadays pretty fast, it is implemented though heuristic methods which converges to satisfying results. Alternatively, reads can be assembled de novo through a tool like Trinity.

5.3.2.6 Quantifying reads per gene

The aim of read quantification is to count sequence reads per gene (parallel to quantify the fluorescence signal in microarray). Several decision and precautions must be made when mapping reads to the genome. For example filtering out rRNA, tRNA and mitRNA is necessary, while non-coding RNA can be included or excluded depending on the type of analysis performed. Alternative splicing, overlapping genes and pseudogenes are dealt with. Moreover, different type of counts can be employed:

5.3.2.6.1 Counts used by differential expression methods Counts used for differential expression analysis are:

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

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

- **RPKM**: reads per kilobase of exon per million reads mapped. It is called FPKM for fragments. It's the most common type of count used.

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

  The typical length of a gene and the typical number of reads generated by a RNA-seq experiments are taken into account by the constants.

5.3.2.6.2 Intra-sample normalization The most common count used for intra-sample normalization is **TPM** It is the measurement of the proportion of transcripts in the pool of RNA. It takes into account the length of the transcript, more reads that fall into that gene. Let $\tilde{l}_i$ the length of transcript i, then:

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

5.3.2.7 Inter sample 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

The aim is to go from differentially expressed genes found in previous analyses to biological functions. How does my data relate to known biological functions? Are there specific functions that are characterized by gene expression changes?

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. It will be discussed in section 6.4
- **Network**: they can be just visual, identify modules satisfying some joint gene expression and topology requirement.
- **Pathways**: they can be just visual, or they are scored exploiting gene expression and topology.

Before diving into the concepts od GSE, we will first have a look at Gene Ontology.

6.2 Gene Ontology

An **ontology** formally 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. A **controlled vocabulary** provides a way to organize knowledge for subsequent retrieval, but does not allow reasoning about the entities. 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 databases structure and search queries.

6.3 Gene ontology

A gene ontology is a way to capture biological knowledge for individual gene products in a written and computable form, in the sense that it has a formal structure such that it can be used by a

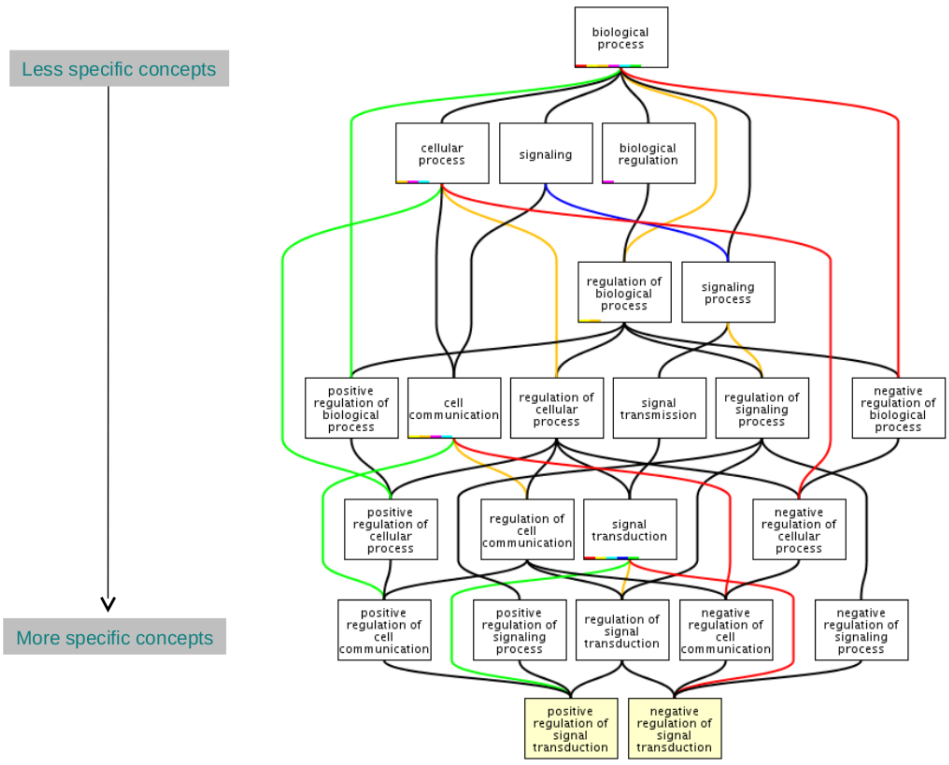

Figure 6.1: Acyclic graph representation of Gene Ontology

machine. A GO can also be defined as a set of concepts and their relationships to each other is arranged as a hierarchy. An ontology is structured as an acyclic graph where terms can have more than one parent (figure 6.1). Terms are linked by directed relationships like: "is part of", "regulates". Each node can have multiple parents, but no cycles.

6.3.1 Concepts in GO

GO is organized in three main hierarchies:

- Molecular Function: an elemental activity, task or job (e.g., protein kinase activity);
- Biological Process: a commonly recognized series of events (e.g., cell division);
- Cellular Component: where a gene product is located (e.g. mitochondrion).

Each one of these has its own acyclic graph which contains a detailed description of the biological domain focusing on these three different abstraction levels.

6.3.1.1 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 (more specific).

6.3.1.2 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 dependencies that are required to describe a pathway. Some examples are:

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

It might be difficult to distinguish between molecular functions and biological processes. A general rule is that a process should include multiple distinct passages.

6.3.1.3 Cellular component

A cellular component is linked to a component of a cell with the condition that is part of a larger object and can be part of an anatomic structure. Some examples are:

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

6.3.2 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 (or is involved in a particular biological process, or is located within a certain cellular component), determined by a particular method and described in a reference.

6.4 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.

6.4.1 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).

The main resources for this type of data are:

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

6.4.2 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.

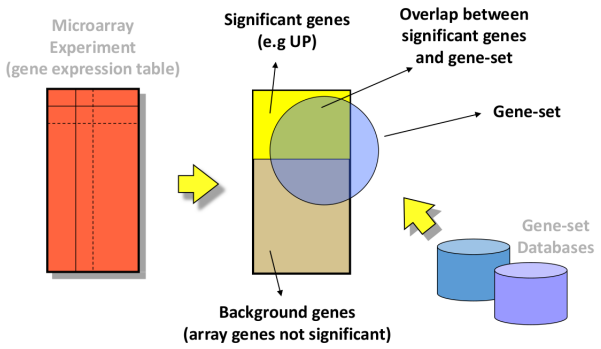

Figure 6.2: In this microarray two-class experimental design, significant genes are selected by a threshold. The circle represents the overlap between significant genes and gene-set. Fisher's test will test if this overlap is larger than expected by random sampling the captured genes?

6.4.3 Enrichment methods

An enrichment test combines a gene expression matrix 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. Other designs can be implemented, for example an expression matrix can be obtained with multiple conditions employing for example ANOVA.
After choosing which genes are down- and up-regulated form the gene expression table (or matrix), the actual enrichment test can be performed. 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. If by repeating the random sampling we never find a bigger overlap than the one we selected, it means that the gene-set is not casual.

6.4.3.1 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, b, c, d are the size of the four subsets.:

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

With "significant" meaning down- or up-regulated, based on our experimental design. 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.

6.4.4 Whole-distribution - GSEA enrichment

Enrichment test can either be threshold-dependent, e.g., Fisher's test, but can also be a **whole-distribution** test, e.g. GSEA, which considers all the genes in the experiment.

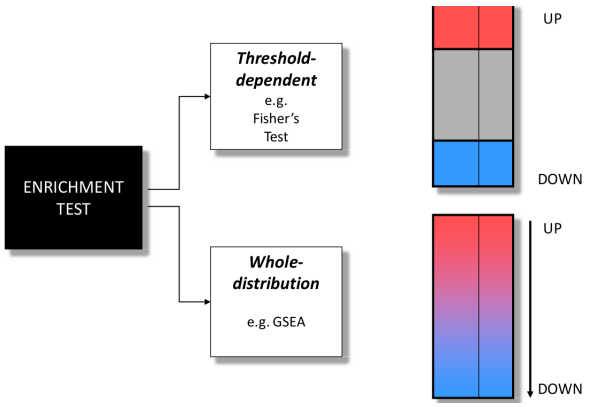

Figure 6.3: Difference between an enrichment test with genes selected by a threshold *vs* GSEA

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.

GSEA is useful as it tackles three major threshold methods' problems:

- There is no "natural" value for the threshold;
- Different results at different threshold settings;
- Loss of information due to thresholding (No resolution between significant signals with different strengths, weak signals neglected)

GSEA can work on two different experimental designs. It can either use the expression matrix of a two-class comparison and rank the genes based on fold change, log-ratio, t-test and SAM; but it can also use the expression matrix of **correlation to phenotype**: in this case the genes will be ranked based on the Pearson correlation. We will be focusing on the two-class comparison experimental design.

6.4.4.1 Process

Calculate the ES score First the enrichment score is calculated: genes are ranked according to a specific measure like the fold change. It tells us where are the gene-set genes located in the ranked list and if the distribution is random, or if there's an enrichment in either end.

Then the chance of observing a gene in a position towards the upregulated or down-regulated 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.

The final ES score is the maximum running ES score.

So, an high enrichment score means high local enrichment, and vice versa.

To state that there's an enrichment we need to compare the ES againts a null hypothesis. In Fisher's test the test itself provided a p-value, in this case the p-value is calculated using a permutation method.

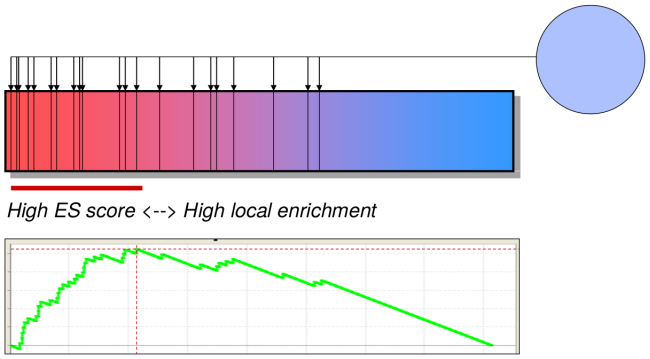

Figure 6.4: Top bar: Indicates where the gene-set genes are located in the ranked list. We need to find out whether the distribution random, or if there's an enrichment in either end. Every present gene (black vertical bar) gives a positive contribution, while every absent gene (no vertical bar) gives a negative contribution. Bottom graph: The bottom curve is the running ES score.

Empirical p-value estimation We need a way to estimate the empirical p-value for every gene-set.

- From randomized data a null-hypothesis distribution is generated. See permutation setting in subsection 6.4.4.2!
- The empirical p-value is computed observing how many random data have enrichment score greater than the read ones. The p-value is estimated for each gene set for which we calculated the ES.

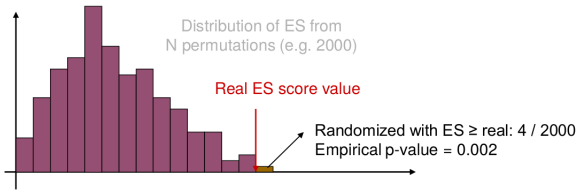

Figure 6.5: Estimation of the empirical p-value for each gene-set

Calculate the FDR - False Discovery Rate correction Then the FDR correction of the p-value is computed. The standard ranking metrics is:

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

N.B.: The p-value only depends on the single gene-set performance, while the FDR depends on the performance of all gene-sets.

6.4.4.2 Permutations

Permutations setting have important implications. **Gene permutations** can be used when biological replicates are very similar within classes and classes 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.

6.4.4.3 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.

6.4.5 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 (e.g., regulation of biopolymer catabolism). Small gene-sets are undesired as their statistics are noisy and may decrease the FDR of other sets.

6.4.6 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 color represent differentially expressed genes.
- The size of the node is the number of 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, in particular since the advent in systems biology. Genome and genomics, proteome and proteomics are represented by networks. The objectives of system biology are:

- Comprehension at the level of system, representing it using formalisms like networks.
- Analysis of individual components.
- Analysis of interactions.
- Analysis of potential emerging properties.

Network (graph theory) allows us to recognize properties seen only at system level and not individual, need to capture the whole system dynamics. In biology, networks are used to model Signal transduction networks, gene regulatory networks, protein-protein interaction.

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 analysis

Comparing two big networks is not as easy. For example, defining isomorphisms is simply not feasible. A field of research focuses on network properties and works on three main abstraction levels: analysis of **single elements**, analysis at the level of **groups** and **global** analysis.

7.3.1 Analysis of single elements

In the analysis of single elements the more important nodes are identified. **Centrality measures** are a class of measures that indicate of the importance of a node in a graph or network. Different centrality measures are available, the main ones being **degree**, **closeness** and **betweenness** centrality. But other are also used, like the eigenvector centrality.

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. E.g., in protein-protein interaction, the failure (or absence) of a hub could produce a dramatic cascade effect.

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'. Can be seen a the ratio between the number of nodes and a notion of t*total distance* between the node and all the others, calculated as shortest path.

Nodes with high closeness centrality can access quickly other nodes and have rapid cascade effects on other nodes. Closeness us to rank the nodes in order of importance.

7.3.1.3 Eccentricity centrality

Eccentricity centrality is computed as:

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

The eccentricity is a measure of the centrality index.

Is calculated 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 Eigenvector 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, which means diving the graph by its characteristics. First a similarity function between nodes is defined in terms of network topology. For example, in the enrichment analysis (gene ontology) the higher the overlap, the closer the terms 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 $N(n)$, the neighbours of n, divided by the number of potential links between them:

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

For directed graphs the 2 factor is eliminated.

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 Observational studies

Observational studies can be divided in two categories: **descriptive** and analytical. The main idea is to observe a sample of population and extract information about environmental and/or genetic characteristics wrt specific genetic or phenotype marks (often, diseases).

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 studied. They are often the first step or initial inquiry into a new topic event, disease or condition. They typically estimate the frequency and the magnitude of the event analyzed.

8.1.2 Analytical studies

Analytical studies take the descriptive studies a step forward. 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.
Analytical studies can be ulteriorly divided into cohort and case studies.

8.1.2.1 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 or pollutant.

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.

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.

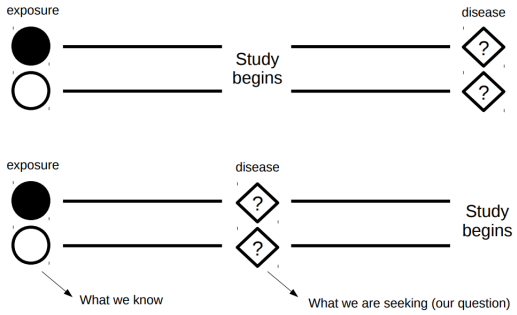

Figure 8.1: Prospective and retrospective cohort studies.

8.1.2.2 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

From such a table we can define three main measures of association:

Strength of association: Relative Risk (RR) and Risk Excess (RE) First, from this table the percentage of individuals exposed harboring the disease can be computed:

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

Also the percentage of individuals not exposed and not harboring 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: incidence of developing the disease is much lower when the population is exposed to an event.
- $RR \sim 1$ indicates an absence of risk.
- $RR > 1$ indicates a risk factor.

Precision of association: Confidence Interval (CI) Remember, the risk is calculated always in a sample of the population. It is a range of values, on the basis of the sample data, in which the population value (or true value) may lie.
Formal definition of CI: *If the measurement of the estimate could be replicated many times, the correct value is inside the interval* 95% *(or* 90% *or* 80%*...) of the time.* In practice, we can be reasonably confident that the correct value (which is the one we are observing in a population) is inside the confidence interval.
Usually the relative risk indicates the amount of random error around the point estimate. The formula for calculating the RR the becomes something like:

$$RR = pointestimate\,(lowerconfidencelimit - upperconfidencelimit)$$

Where the subtraction in the parenthesis represent the confidence interval.

Significance of association: p-value (P) From the table also a significance of association can be computed. It requires a p-value, that determine how unlikely it is that the events observed arise by chance.

8.1.3 Case-control study

The purpose of a case-control study is typically to study rare diseases or multiple exposures (or genetic factors) that may be related to a single outcome. Participants are selected based on **outcome** status and therefore divided in two categories, case- and control-subjects:

- 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.

8.1.3.1 Measure of association

The same table as a cohort study is built (as in table 8.1), but the measure of strength of association is different.

Strength of association: Odds Ratio (OR) First the odds[1] 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:

[1] An odd is a ratio of probabilities, ratio between the probability that an event will happen and the probability that an event will not happen.

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

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.

OR can be interpreted like RR:

- $OR < 1$: protective factor.
- $OR = 1$: absence of risk.
- $OR > 1$: risk factor.

Significance of association: p-value (P)

Precision of association: Confidence Interval (CI)

Relation between RR and OR

RR and OR are two different measures, derived from two different types of studies, one searching for the exposure and one searching for the outcome. Some mathematical relation exists between these two measures, that tells us that when the disease is rare, mathematically the $RR = OR$. A graphical representation of this relation is depicted in figure 8.2. The difference is that the RR is *immediate*, meaning that if a person has probability 2 of developing a disease, it is the same as saying that it has double the risk of developing it of a person not exposed. With OR instead, this reasoning is not immediate.

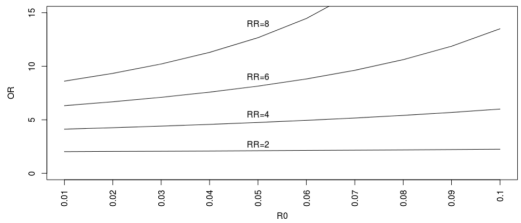

Figure 8.2: The more a disease is rare, the stroger the correlation is. On the contrary, if the disease in a population is not that uncommon, we can see a divergence between OR and R0, where R0 is the frequency of the disease in the not exposed population.

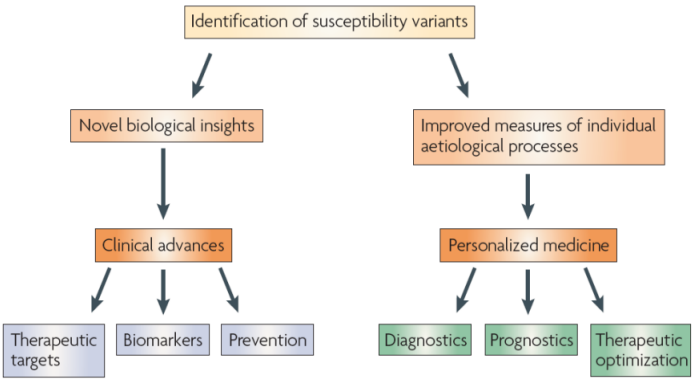

Figure 8.3: Potential of GWAS and main possible applications

8.2 GWAS

8.2.1 Objective of GWAS

The objective of a GWAS is to find connections between a phenotype (height, type-I diabetes, etc.,) known to be heritable and whole-genome genotype. GWAS were developed in 2004, mainly thanks to the HapMap project, which unraveled the existence of linkage disequilibrium blocks, which allowed the exploitation of tag SNPs. Particularly important was the realization, after the discovery of the haplotype blocks, that not all of human genetic variation (millions of SNPs) had to be genotypes to find associate variants, but only a fraction of those (**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 defense.
- Build statistical models to predict phenotype from genotype.

8.2.2 Main applications of GWAS

An overview of the possibilities provided by GWAS are reported in figure 8.3.

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

In general, genetic polymorphisms are genetic variants that have a prevalce greater than 1% in the population. There are different type of polymorphisms:

- **Single Nucletoide Polymorphisms (SNPs)**, which will be our main focus;
- Small-scale insertions and deletions;
- Microsatellite variations;
- Copy Number variations.

A SNP is defined as a single base variation in a DNA sequence. They are classified according to the minor allele frequency MAF[2]

- 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, change only the phenotype;
- Harmful, associated with a multitude of diseases;
- Latent, can become detrimental only in particular conditions (genetic-environmental cooperation).

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:

[2]Minor allele frequency (MAF) is the frequency at which the second most common allele occurs in a given population. They play a surprising role in heritability since MAF variants which occur only once, known as "singletons", drive an enormous amount of selection. (Wikipedia)

- 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.

8.2.4.5 SNP genotyping

SNP genotyping represents each SNP in the dimension of its A and B allele intensity. Not all SNPs have a clear representation.

8.2.4.5.1 Birdseed A method to call the genotype is Birdseed, a clustering algorithm that first construct training models for each SNP in the array and then compute SNPs genotyping on data of interest using the training models. Each SNP is a bird, where its wing points are genotypes AA and BB and the body is AB. Birds are computed for all SNPs. Then Birdseed estimates cluster centres and covariance matrices. Its confidence is computed as:

$$Confidence = 80\%E_1 + 20\%E_2$$

Where E_1 is the posterior to the second closest peak over the posterior to the closest and E_2 is the deviation penalty from the closest peak. Then the quality score is computed as:

$$QS = -\log_{10}(confidence + 0.00001) \cdot 2000$$

8.2.4.5.2 SNP genotyping from NGS data To genotype SNP from NGS the pileup is used. After the reads are aligned and the coverage computed the pileup is the count of the number of times the reference and the alternative reads appear. From this the allelic fraction can be computed.

8.2.4.5.3 Quality control To perform quality control of SNP genotyping:

- Subjects with missing call rate less than 1 or 5% are filtered out because of poor DNA quality.
- SNP with genotype call rates less than 5% are filtered out because of errors, noise and non-specific probes.
- SNPs with minor allele frequencies less than 1% are filtered out because they are error prone and unpowered.
- SNPs are filtered out by testing for HW equilibrium: a deviation can indicate genotyping errors, batch effects or a change in the genetic structure.

8.2.4.6 Independence of individuals

Independence among samples is a fundamental assumption of case-control studies, so related individuals are excluded, or their association taken into account during association analysis. Genomic distance between samples can be computed based on SNP profiles through genetic fingerprinting or through SPIA:

$$D(CL_1, CL_2) = \frac{1}{vN_{SNPs}} \sum_{i=1}^{N_{SNPs}} \neg\delta(CL_{1i}, CL_{2i})$$

8.2.4.7 Genetic structure

Not everyone in the population has the same genetic background: some people are more genetically similar than others. Admixed population are particularly interesting and many SNPs in the genome have different distribution between different population due to random drift. Many traits have strong population association, so the genetic background of individuals of a cohort must be taken into account to avoid structure and stratification of the samples. The degree of admixture and population of origin identification is important. This can be done through SNP-based or model-based methods.

8.2.4.7.1 Logistic regression Logistic regression is a method used to assess the associations of SNPs and outcomes. The better the fit the stronger the association. It is computed as:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots \beta_m x_m$$

Where x_i are the SNP genotype and β the covariates like age, sex and population structure. Then genetic structure correction is performed through QQ plots, where deviation from the diagonal indicates structure. The data is fitted to the diagonal as to reduce structures effects.

8.2.5 Standard cutoff in GWAS

Considering evidence for $1M$ haplotype blocks in the genome, the accepted p-value cutoff in GWAS studies is between 10^{-6} and 10^{-8}. This is due to the fact that each SNP is tested at the traditional significance level and multiple comparison is an important consideration in a GWAS analysis and must be handled properly.

8.2.6 Independent replication

Independent replication is used to evaluate systemically whether or not the discovered SNPs in initial GWAS are spurious signals. Samples are collected the same way as the original study and the association is computed with the same genetic model. The effect sizes of marker should show the same signs in both the replicated and original study.