Network Based Data Analysis - Lauria

authorName

telegram: @authorName

Github: mainRepoLink

March 28, 2022

Contents

Part I

Introduction to omics

Chapter 1

High-throughput biological data

High-throughput techniques, ofter referred to as **omic techniques**, are a way to collect extensive, comprehensive, quantitative and large scale data on a certain aspect of a biological system. The main omics data types are genomics, transcriptomics, proteomics, metabolomics; each of them possesses specific individual technologies and methods.

1.1 Required biological background

It is assumed that you posses some amount of knowledge regarding the following topics:

- Molecular biology of the cell (cell types and characteristics)
- Molecular components of the cell (mainly proteins and nucleic acids)

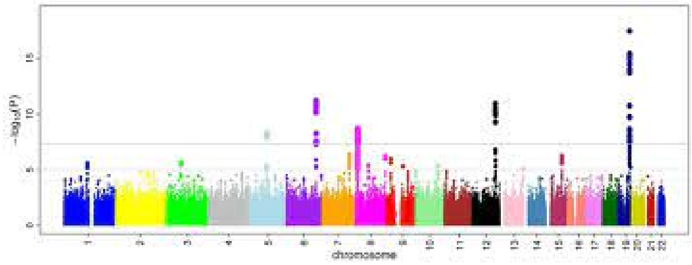
1.2 Genomics

Genomics refers to the study of the genome, which is the entire DNA (or RNA in some bacteria and viruses) content of a cell. The focus is generally on the variation of the genome of an individual (or a group of individuals) compared to a reference genome (a sequence that is shared by most individuals of a scpecies). Particularly important is the study of **SNPs** (Single Nucleotide Polymorphisms) which allows to retrieve information such as philogenetics, disease predisposition, forensic applications (individual recognition, paternity testing...) and many others. SNPs can be studied either via **DNA microarrays**, which are economic since they target only specific regions of the genome where the presence of SNPs is known, or via **whole genome sequencing**, which is more expensive but gives information.

1.2.1 Genomic wide association studies

Genome Wide Association Studies (GWAS) are studies where many genetic markers are studied across the genome a population in orded to try and correlate them to specific conditions (mainly predisposition and presence of disease).

Figure 1.1: Manhattan plot: Mahnattan plots are a common way to represent GWAS results



1.3 Transcriptomics

The **transcriptome** is the collection of mRNA species/transcripts in a cell at a given time. It can provide a lot of information on cell state, growth conditions and many others, which is why it is heavily studied; moreover transcriptomics is **robust**, relatively **cost effective** and **user friendly**.

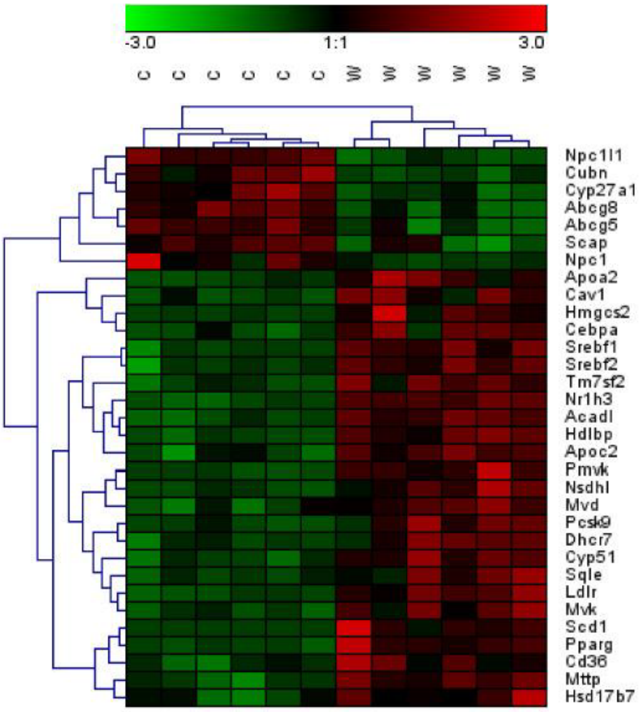
1.3.1 Microarrays

Microarrays can be used to measure levels of mRNA in a high-throughput fashion. Roughly, a microarry workflow can be summarized as:

- Extract RNA content from cells
- Obtain fluorescence-marked cDNA
- Hybridize the cDNA with the DNA probes on the chip
- Wash chip to remove non-hybridized cDNA
- Record output fluorescent signal
- Normalize the result:
  - If using a 2 color microarray (red-marked test sample, green marked control sample), you obtain the fold expression of the transcript compared to the reference (red = overexpressed, yellow = same expression level, green = underexpressed)
  - If using a single color microarray (only marked test sample) the output can be normalized in various ways, generally substracting the signal of aspecific probes (of the 20-25 bases a couple of them are mismatched, to test for aspecific interaction)

  Further normalization steps may be required
- Analyze the output

Figure 1.2: Microarray output: Genes with different expressions can be visualized, each column is a sample and each row is one of the 25 genes. The colour represents the level of expression



1.3.2 Next generation sequencing

Next generation sequencing (NGS) is another way to analyze the transcriptome; each mRNA species in each sample is broken into fragments, and all fragments are sequenced base by base. The current technologies manage to sequence small pieces of RNA, around 250 bases. By using overlapping ends of the fragments, we can reassamble RNAs. A higher amount of fragments grants an higher coverage. This method allows **massively parallel sequencing**, **sequencing of both known and unknown transcipt** (no probe requirement), detecting transcripts in a **high dynamic range** (up to 10^6 copies)

1.3.3 RNA-seq pipeline

After obtaining the raw sequencing results, the RNA-seq pipeline uses some other resources, namely input files (reference genome, gene annotation) and programmes (some accessed through cloud to speed up data analysis), to obtain polished information and results.

1.3.4 Microarrays vs sequencing

Overall, microarrays are more economic but can only give information on transcripts for which a specific probe exists on the chip. They allow to obtain differential expression studies but not absolute

Figure 1.3: NGS flowchart

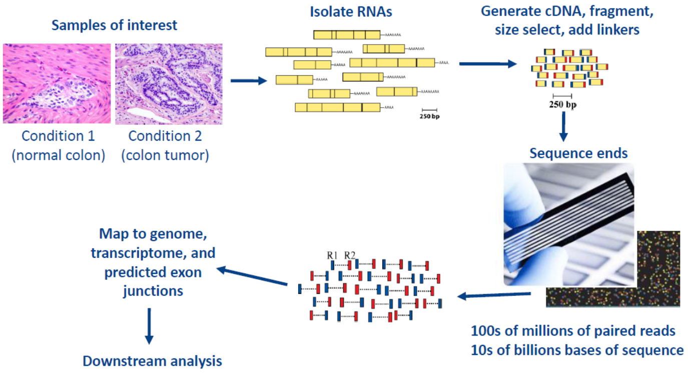
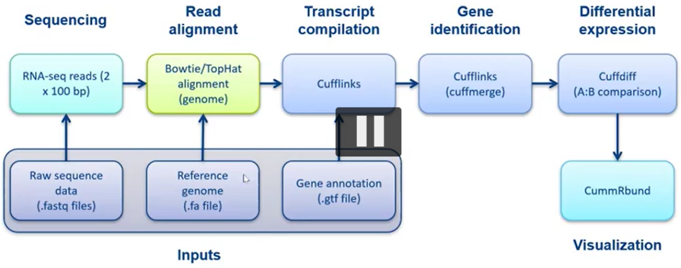


Figure 1.4: RNA-seq pipeline



quantification. Sequencing is more costly, but unbiased and higher-throughput; moreover it allows further analysis, such as mapping and absolute quantification.

1.4 Proteomics

The **proteome** is the set of all proteins produced under a given set of conditions. The analysis of the proteome, called proteomics, has countless applications. Proteomic techniques allow for high-throughput analysis of protein content, yet again the throughput is significantly lower than transcriptomic techiques: this is because proteins are more complex than nucleic acids, both in structure and sequence, and they do not have a convenient 1:1 pairing pattern. For this reasons, proteomic techniques usually rely on other physical characteristics, mainly **mass** and **charge**.

1.4.1 2D gel electrophoresis

Standard gel electrophoresis separates proteins based on their size (varying the crosslinking rate of the polyacrylamide net); this method has very limited resolution (it is influenced by many factors, such as denaturation, SDS coating, fragmentation...). 2D gel electrophoresys partially increments the resolution since it separate proteins based on both weight and charge. The standard protocol for 2D gel electrophoresis consists of:

- Protein extraction, purification and usually denaturation
- First dimension electrophoresis: proteins migrate in a low density gel in which molecules have been placed to create a pH gradient. The proteins distribute along the gradient solely because of their charge (regardless of mass)
- Second dimension electrophoresis: the first dimension gel is placed at the edge of a regular polyacrylamide gradient gel, which allows to separate the molecules by weight.
- Coloration and acquisition of gel signal
- The result is a sort of map, with increasing pH value from left to right and decreasing weight from top to bottom.

These maps can be then compared either with maps from other samples (to identify differences) or with databases to identify the proteins spots. Notice that this procedure does not technically identify univocally the protein and thus, if required, you might have to cut the protein spot from the gel and procede with further analysis. Another significant problem with this technique is the need for large amounts of samples to have significant signal levels.

Figure 1.5: Isoelectric focusing: separation of proteins in a pH gradient

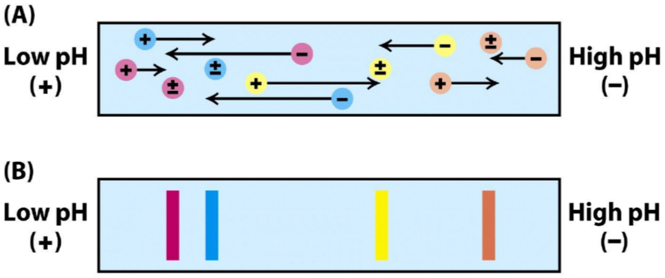


Figure 3-11
Biochemistry, Sixth Edition
© 2007 W.H.Freeman and Company

Berg, Tymoczko & Stryer, "Biochemistry", 6th edn, 2006, p. 73

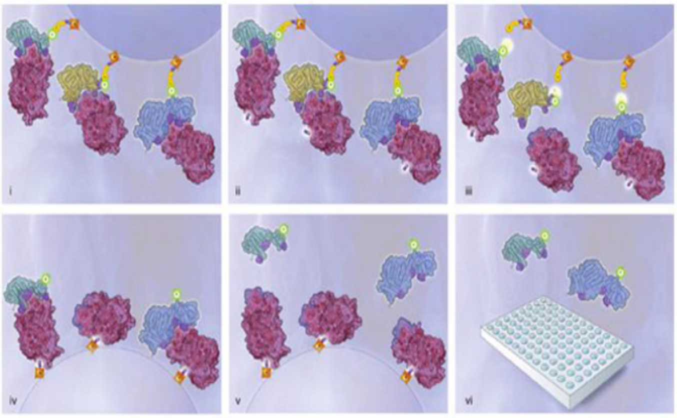
1.4.2 Liquid chromatography/mass spectrometry

Mass spectrometry consists of a plethora of techniques that allow to separate molecules based on their mass to electrical charge ratio. This technique allows for high-throughput identification of proteins (without need for further analysis). The main limitation of mass spectrometry is the need for highly purified protein extracts (and thus labour intensive procedures) since it is highly susceptible to contaminants; for this reason the protein extract is usually purified and diluted using liquid chromatography and the output is analyzed via mass spectrometry (from which the name liquid chromatography/mass spectrometry or LC-MS for short). Another big advantage of this technique is the very low amount of sample required (some μg of purified proteins are enough). Mass spectrometry does not allow for absolute quantification but only relative quantification (due to peptide volatilization and other factors). LC-MS can usually identify up to about 1000 proteins per sample (more realistically around 600-700 due to dynamic range issues). LC-MS can be used to study also post-traslation modifications (such as phosphorilation).

1.4.3 Protein arrays

Protein arrays are conceptually similar to DNA arrays, but instead of DNA probes they use **aptamers**, which are oligonucleotides or peptide molecules designed to uniquely bind to a specific molecule. Proteins bind the aptamers and a fluorescent signal is used for detection. This method is easy and cheap, but it measures a low amount of proteins and only proteins for which an aptamer was found. It is still an emergent technology.

Figure 1.6: Protein array



1.4.4 Proteomics vs transcriptomics

Both proteomics and transcriptomics are very powerful techniques. In general transcriptomics is cheaper, more robust and user-friendly, while proteomics pose some more limitations, namely due to protein purification and stability. Both techniques keep being developed since they provide different informations, for example transcriptomics allows to study non-coding RNAs while proteomics allows to study post-translational modifications (the opposite in not true).

1.5 Metabolomics

Metabolomics is a field of life science research that uses high-throughput technologies to identify and/or characterize all the small molecules or metabolites in a given cell, tissue or organism (the so called metabolome). There are two main approaches in metabolomics:

- Quantitative methods: quantitatively identify target metabolites in a sample
- Chemometric methods: profiling samples based on metabolites in them

These approaches can be useful, for instance, to early detection and diagnosis of diseases, since certain metabolites correlate do higher risk of certain pathologies.

1.6 Other high-throughput data sources

Countless other sources of high-throughput data exist and many more are becoming viable.

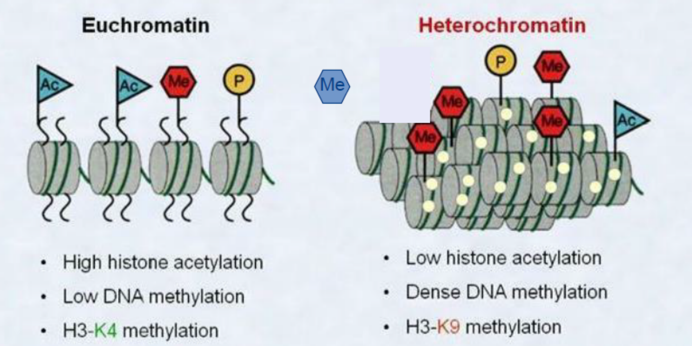
1.6.1 Microbiome

With the term **microbiome** we collectively refer to all the microbes in the human body, whether they are bacteria, fungi, protozoa, viruses or other. This heterogeneous population is of interested since it is highly represented in most regions of our body; the role of microbes is important since they contribute to some physiological functions of the organism (gut microbes), they can cause pathologies when deregulated, their populations show differences among individual hosts. (*To go into more detail, please refer to the "Computational Microbial Genomics Notes - Segata" available* ***here***)

1.6.2 Epigenomics

Epigenetics is the study of heritable changes in gene activity that are not caused by changes in the DNA sequence. Some mechanisms that produce such changes are **DNA methylation** and **histone modification**. Both mechanisms alter how genes are expressed without altering the underlying DNA sequence.

Figure 1.7: Examples of epigenetic modifications



1.6.3 Micro RNAs

Micro RNAs, or miRNAs, are a family (around 1000 in humans) of short (20-22 bases) non-coding RNAs which affect mRNA translation and therefore protein expression. They are produced as precursors in the nucleus, then when they find and pair with their target sequence, they recruit cellular machinery that activates them and causes transcript degradation. miRNAs can be isolated from total RNA and can be profiled to get information on genes that are currently regulated.

1.6.4 Interactome

The **interactome** of a protein is the set of molecules (generally other proteins) that interact with that specific protein (therefore we usually talk about protein-protein interaction). Studying the interactome of a protein can help understand its function and possible ways to modulate its activity. There are databases for interactomes. Interactomes are generally shown in 3D graphs, with nodes representing proteins and lines representing connections between them.

Chapter 2

Clinical trial study design

2.1 Required background

Some basic knowledge regarding cell lines, immortalized cell lines, model organisms, human studies and their pros and cons are required. *This section might be expanded more clearly in the future.*

Chapter 3

Second Week

3.1 Central Dogma

3.2 Measure biomolecules

Proteins

- Western blot
- ELISA
- Northern blot
- Enzyme assay

RNA

- **DNA microarray**: it is an old technique
- RT-PCR
- **RNA sequencing** based on next generation sequencing

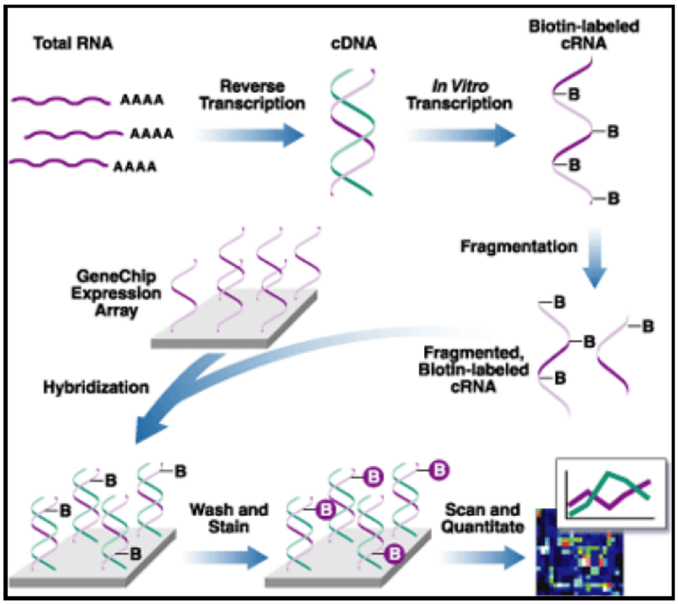
3.2.1 DNA microarrays

points on a glass slide, technology invented in 1990s, they were developed after macroarrays. The technology can be used to quantify miRNA, SNP excetera. It includes all the genes in the genome, including the splice variants and it is reliable for gene expression quantifications. It is easy to analyze, thanks to the bioinformatics resources that are now available.

Affymatrix is a device that permits to do everything on a chip, it was developed for a series of species.

Competitor of Affimetrix is Illumina BeadChip, beads are ligated to oligonucleotides. The beads are inside wells. Illumina makes two colours cheaps (in terms of fluorescence). The possible source of errors are depicted in the image: The first thing to do is to extract RNA from cells, which could be contaminated, then the extraction of RNA occurs, and a low extraction is possible. Degradation of RNA could also happen. During the reverse transcription, and fluorescent labelling, you have to

Figure 3.1



ensure that the reverse transcription happen for each molecule, and each has to be labled. After, during hybridization has to occur equally for all the molecules, it could also happen cross-hybridization. Longer fragments can help in this. Scanning errors could also take place at the end.

Other types of errors, can be due to non-specific fluorescence. Binding could also happen not specifically. Some methods exist to reduce the noise (error), for example: the Affy matrix, which uses two probes, one with Perfect Matches and a similar probe, but with a single different nucleotids (mismatch). If you subtract the two fluorescences, you should eliminate the noisy fluorescence (due to non-specific hybridizations).

At the end, it is obtained an **absolute quantification**.

identical Probe, and one which distinguish itself for a single mismatch. It is tried to remove the fluorescence which is not due to a specific hybridization. It is not very effective.

Affymatrix give data for quality control check. what is RMA?

spy probes hwich are not converted to a specific gene, If I'm analyzing human RNA, I can add mouse RNA.

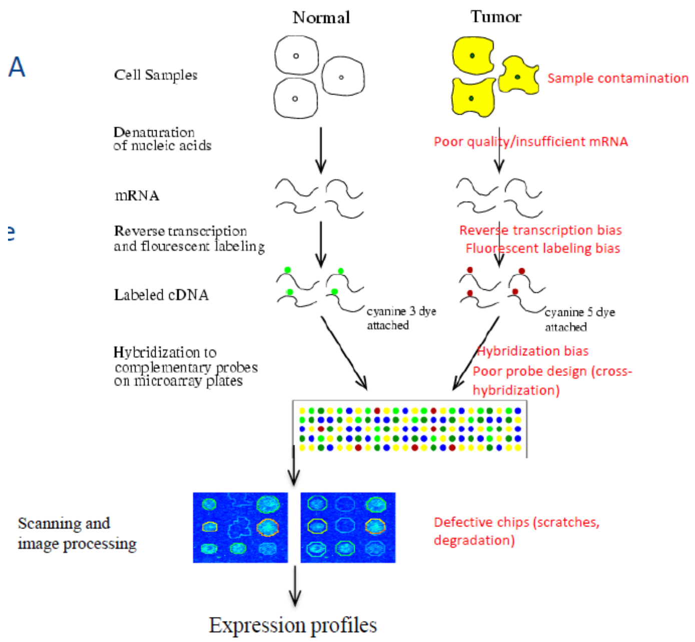
Bioconductor?

3.3 Normalization

In addition to background correction, it can be done a normalization. It is made a boxplot, where each column represents a profile of the dataset, it represents the distribution of values in the dataset. Compare the distribution of values. It tries to realign the data, compute the avagare and subtract the avarage.

Quantile normalization is more drastic. Other techniques are then possible.

Figure 3.2


Normal
Tumor
A
Cell Samples
Sample contamination
Denaturation of nucleic acids
Poor quality/insufficient mRNA
mRNA
Reverse transcription and flourescent labeling
Reverse transcription bias
Fluorescent labeling bias
Labeled cDNA
cyanine 3 dye attached
cyanine 5 dye attached
Hybridization to complementary probes on microarray plates
Hybridization bias
Poor probe design (cross-hybridization)
Scanning and image processing
Defective chips (scratches, degradation)
Expression profiles

3.4 Other problems with transcriptomics

- certain ranges of values, if the expression of a gene is over this rang, it will be not possible to have a response.
- The gene might not be on the chip,
- Can't differentiate splice variants well. Some chips are designed to visualize only a part of a gene. Theere are some specialized chips to do this, but normally this doesn't happen. YOu need to have the promoter of the gene.
- The gene might be below detection limit
- Can't differentiate RNA synthesis and degradation
- Can't tell us about post translational events
- Bioinformatics can be difficult

Affymetrix GenexChip generates several type of files, such as • DAT (image file), CEL (raw data file), CDF (chip definition file) All platforms have different data formats.

microarray standards like MIAME or MAGE-ML are STANDARDS ...

Bioconductor is a open source software project for the analysis and comprehension of genomic data. Provide a wide range of powerful statistical andgraphical tools.

Microarrays can be found in repositories, like Array express, sited in UK, Gene Expression Omnibus, founded by NIH in USA, the japanese CIBEX, and other databases like NetAffx, Ensembl, TIGR, Stanford.... ArrayExpress and GEO have not overlapping pieces of data. Recount2 has small number of datasets, and most of them are also of GEO

3.5 RNAseq

it's an alternative way of measuring the level of expression of some genes, it uses next generation sequencing techniques. How this? it represents the toal technologies which permit to sequence an entire human genome in one week only, and of course with lower costs, if compared to the amount of money that was spent for the human genoe project. The speed and the cost are nowadays lowering.

At the beginning, long stratches of DNA sequenced, with this one instead the reads are really shorter (200 bases long around), but in much larger quantities. The idea is, if you want to quantify the messanger RNA of a cell, and sequence them, you have to split the nucleic acid: stretches thatare around 250 bases long. To reassable, it is possible to take different samples, and for each one of them to perform a random fragmantation. Through computations, it is possible to reunify the reads. It is possible to quantify the messanger RNA molecules, to quantify the different splicing products. For the riassambly, a software is used through these sequencings, it is possible to identify the reads.

RNA seq sequencing everything which is inside the cells. Requires computer with big rams and big storage to manage all the number of reads. With cloud computing it is possible to have thses problems solved. Recount provides datasets from GEOs. Raw counts are obtained from the first steps, preprocessing. It is produced a matrix, each row is a gene, each col a sample. Integer numbers. Different from the microarrays since here there are integers.

Several type of files, FastQ is a text-file where each row is a sequence, produced by illumina sequencing machines. Different types of FastaQ files are produced. The BAM is a text-file with the read and next to it the coordinates with reference to the genome of sequencing. At the end the raw counts

now it is needed preprocessing, counts are treaky to handle. Confront samples, confront levels of expressions, inter-sample or intra-sample. Since counts, compare numbers? no, each run doesn't produce the same amount of reads. This means that it is not possible to compare the number presents. Normalize with respect total number of reads needed

The bigger gene larger amounts of gene, normalize length reads and of genes. Some normalization processes are figured in the table below. RPM stands for Read per Million, normalize count per the total n of reads, RPKM normalizes gene length and mapped reads. Depending on hte order of normalization, one of the 2 results. TPM is slightly better.

Differential expression genes, t-test, which genes are affected by the changes. When the distance between averages is significant. To do a t-test, some other methods are to be applied.

Description of the experiment gives us informations about the modality of analysis that should be considered to make the analysis, two groups, a control and a test group, minimally. Average experiments have replicates, to reduce the amount and the probability of errors, altough more expensive.

Guidelines:

- choose the data
- preliminary analysis
- normalization
-

In order to make better decision, is to make interesting biological questions.

boxplot, column output of one microarray normalization: to make values comparable, basically you compute median for each array and subtract the mean for that array. next divide all the values for the standard deviation, boxes nicely alligned and same space. This because assume the large majority of genes are not changing from a sample to another. Normalization tries to eliminate that error. Several file formats, the databases considered are GEO, attempt to standardize format of files, MIAME for example.

Recount2 takes dataset foundable in GEO, preprocess them

RNAseq is a different approach to quantify RNA. RNA molecules are cut with restriction enzymes, chop original RNA. Then, the library can be fed to the sequencing machine, which produces a very long file with the list of the fragments and their sequences. Re-assamble those fragments in order to obtain the original RNA. Re-assembly is easier thanks to a reference genome, sequence several version of the library: chopping in different ways, generation of different fragments. This augments the probability of overlapping. Computational intensive (Gb). Recount permits to have the raw counts, a matrix obtained after performing several action, each row a gene and column a sample, n of fragments associated to that gene. Idea of the level of expression of that gene. Difference with microarray matrix: integrals, relative count (on 10000000 reads, n fragments where seen to be related to that particular gene). The total number of reads have to be equal in each sample, otherwise normalization needed. The number of fragments augment if gene longer. Several methods to make comparisons between different samples. Depending on the normalization that I want, I need a specific package.

microarray: limitation of the chip, design RNA sequencing: no limitation, as everything is sequenced. extra computations.

microarray gives direct quantification of the mRNA.

3.6 Features of the selected database

RNA sequencing data are those selected. mRNA is not able to tell us about the proteomics of a cell, as the degradative path is not weel studiable.

dataset with two groups maybe, max 3 groups. A sufficient number of replicates have to be used to reduce errors. limited number of groups, large number of samples. Very large dataset, it can be taken a number of samples from the total dataset. At least 10 samples per group. Take or a dataset from Geo, or Array or structure.

Geo or ArrayExpress give preprocessed data. processed data in case of RNAseq are the raw counts. Other modifications are necessary. Why not giving the final results? Each step requires a decision about the algorithm to choose.

Assays are the samples

series file from GEO, open in excel. The number represents the probe ID, in fact, different probes can be used to find a gene. It is necessary to go from the probe to the gene. Different IDs represent different probes to find the gene. If you use microRNA datasets, you have to convert the microRNA to its genes.

Recount3 contains tumor data,

SRA is a repository for RNAseq data. It has only raw data, not preprocessed data.

Series matrix file, check the data is in there. Geo could fail in case of absence of data. *getGEO* can be used to find a file in the local directory.

If the RNAseq data are not shown in GEO, but there is an SRA reference, try to find the SRA ID in recount3

Part II

Introduction to PCA

Chapter 4

Third week

4.1 General strategy to analyze a dataset

4.2 Principal Component Analysis

Data from a big table, Starting point is a pxn matrix with p is equal ot the number of genes and n to the number of samples. The number of samples has to be between 40 to 100, if smaller than 20 analysis become difficult. If bigger, it is possible to use a subset of the dataset. p is normally larger than n, as of course genes are more. Source of errors are introduced, making noise. Generally a lot of non informative data, or even wrong, most often it is necessary to transform the data and normalize it.

Formulate a question and try to answer to that, about the aim, which genes are differentially expressed between different groups. Is it possible to group samples with similar expression profile?. Think about the statistical test then. Once done the question, open data file, data normalization, transformation and PCA. Questions through the description of the experiment.

For **class comparison**, it is possible to use a **t-test**, what are the genes whose expression is different between different groups. Statisticall significance has to be detected. T-test to understand if H_0 is true or the alternative hypothesys is True. The test gives a p-value. Calculate the $T_{observed}$. Difference obtained by chance or not? Are the data extracted from the same distribution (in this case mean of the two groups should be similar)? A solution is to use a correction method, like Bonferroni. The problem of **threshold** is not already solved, normally 0.05%. The percentage remains arbitrary. 0.05 means that you have a mistake as result only for the 5% of the times. The ones with smaller p-value are the most interesting. It is possible to sort the genes and take those with lower values of p-value.

Estimates of mean and standard deviation, very different values in different experiments. Especially if number of samples is low. Larger number of samples make the estimates more robust.

The **multiplicity problem** has to be considered. For a given gene and a given type I error rate ($\alpha = 5\%$), we know that this gene has a 5% probability to be a false positive. Thus, when doing one test at $\alpha = 5\%$ for each gene, we know that the number of false positives will be 5% times the number of tests (5 FP for 100 tests, 50 FP for 1000 tests,. . . , 2200 FP for 44000 tests). The T-test produces more reliable results with normal distributed data.

The adjustment is made to have a new p-value.

It is possible to **rank** the genes based on the p-value, those with the lowest p-value are those more interesting. take for example the first thirty, draw a heat-map, where column is a sample and row is a gene. The t-test makes assumption on data, it is not the best oprion for class comparison. You have to set a threshold, and establish which are interesting and which not. The choice of threshold is always arbitrary.

It has to be started the analysis. Open the file, read the data, maybe to many $0s$. It is done a data normalization and a PCA, a way to reduce dimensionality of the dataset.

4.2.1 Normalization

Ranges of values are presented through box-plots. The middel line is the median. Points out of whiskers are outliers. Whiskers represent 1.5 times the interquartile range The range of values are really disalligned, for no reason, variation in the data that has no biological explanation. Reallign the boxes through normalization.

The variation can be sistematic, same for all the samples, or it can be random. The best way to deal with radom is to do several replicates. Random variations have a 0 mean.

The sources of variation :

Dye bias: differences in heat and light sensitivity, efficiency of dye incorporation Differences in the amount of labeled cDNA hybridized to each channel in a microarray experiment (here channel is used to refer to a particular slide/dye combination.) Variation across replicate slides Variation across hybridization conditions Variation in scanning conditions Variation among technicians doing the lab work etc.

The general strategy for normalization is to use House keeping genes, considered not to change on average across samples and conditions. Remove residual variation of house keeping genes taking a pool of these genes.

The simplest method: subtract the median from each values of the profile ⇒ all the samples are alligned on 0, through the shifting up and down of the boxes.

Variants of normalization: batch effect: source of error is known, for example different processing (very unlikely the same results), some packages are available; microarray done in different days.

Once done the alignment to 0 of the median, boxplots with the median alligned. The amplitude of the boxes is not equal, scale normalization can be done: it can be done deviding by the standard deviation for that profile for each profile, each value. To do both the normalization processes,

it can be used the **scale** built in function of R.

The t-test should be done on the final version of the data.

MAD stands for "*median standard deviation*". The possible presence of outliers justify the use of the median, off range, so extreme that can be produced by errors. Remove or mantain? it influences enormously the calculus of the mean and of the standard deviation. Median is not affected instead by outliers.

Other more sufisticated methods are available to normalize the data.

4.2.2 Data transformation

There has to be a reason for that. Replace the numerical values with log of the values. This changes the shape distribution of the data, maybe obtaining data more normally distributed. Ranks are also usable, where each value is substitued with its position after that it is sorted from the lowest to the highest. Z-scores are obtained after standardization of values.

The log-transformation is made on graphs not normally distributed. T-test can be used on normal distributions, and this is considered as True when performing the test.

4.2.3 Principal component analysis (PCA

Also known as latent vectors, latent variates, principal axes, principal factors, reduce dimensionality. **Reduce dimensionality**, each sample 30000 values collection, dataset can be seen as a series of points with 30000 dimensions. Obviously, the number of dimensions has to be reduced. PCA plots are generally 2-dimensionals or 3-dimensional.

Given a sample of n observations on a vector of p variables, with p variables/dimensions equal to 30000

$$x_1, x_2, \dots, x_n \in \Re^p \tag{4.1}$$

it is possible to go from a space of p dimensions to a space of n dimensions. transformation from the space of the xs to the ys.

Using the linear regression, we are going to obtain n coordinates transformed (transformation), where n is equal to the number of samples. We choose the best dimension which describes the most part of the variability of our data (y_1). The second dimension has to be ortoghonal to the first one (y_2). Once we have identified those directions, the point will have new coordinates.

eigenvectors are arrays of n values, eigenvalue. we obtain the directions that we like, how much the stress also. Only 2 or 3 are representative of the variability of the data

each x represent a gene, ys are said metagenes, imaginary genes, whose level of expression is given by the linear combination of all the genes.

$$y_{i,j} = a_{j,1}x_{i,1} + a_{j,2}x_{i,2} + \cdots + a_{j,p}x_{i,p} \tag{4.2}$$

A few metagenes are needed to describe variability of data. Which metagenes to use? Usully the first, the second and the third are used.

In the process $S = B^T B$ it is the covariance matrix. If the data was standardized, correlation and covariance are the same. Correlation matrix vs covariance matrix

in R...: *prcomp* is built-in into R. Array of colours to be used. one gene per row and a sample per column, in some cases, as for *prcomp*, it is needed the inverse arrangement.

In new set of coordinates, directions should be orthogonal, the first ax should be the one over which the cloud of points is more stratched.

If data is a ball, data obtained by random. The way to obtain the coefficients so that it is satisfied ... is to use the covariance matrix, obtained through the multiplication of the matrix per itself transposed. Take the eigenvectors, each one is a row of that matrix, while instead the eigenvalues provide an estimate of the percentage of the variability of the corresponding axes is out of the total variance (how spread along that direction).

The covariance matrix is symmetric and positive. As a consequence, all the eigenvalues are positive.

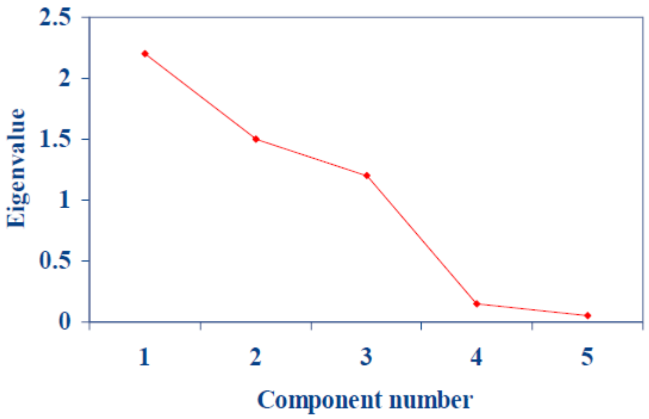
computations of eigenvalues and eigenvectors are made through a function. Coordinates of the points in the new set of coordinates as output.

The skree diagram make us see which components are relevant

Scores

- **Inspect dataset:** check data in there, not too many NAs.
- **Normalization**: The sistematic source of variation can be eliminated by using a box-plot, which means great amount of variation. normalization is made by subtracting the median. Align the size of the boxed by aligning the standard deviation
- **PCA**:

Figure 4.1: Scree diagram



4.2.4 Series Matrix of GEO

on the left column, there are present informations regarding the authors, date, platform (GEO assigns it), the type. Each line starts with !, which indicates this is a piece of metadat. The second section is specific to each sample. The third section contains the data.

Part III

Machine learning

analysis: classification: find a way to assign a label to each sample of a dataset. can provide a list of genes to perform classification.

The classification problem was solved in different ways tries. None of the methods works best for every possible dataset. depends on the nature of the data. select the best performing one.

Many of the methods depend on machine learning: learn from data. Statistical tools can be used combined with machine learning. search engines, natural language processing. Medical diagnosis.

supervised against unsupervised learning. learning extract features using a small dataset. the algorithm can so be used to other samples. Unsupervised methods are instead used without ...

4.3 Unsupervised learning

data clustering, is used also as a synonym to unsupervised learning. TWo methods we will see. it consists in understanding how info points could be clustered. Clustering among almost every field. Main aspects:

- **Distance function**: a way to measure similarity or dissimilarity
- **Clustering quality**: it has to be maximized. Intra-clusters distance could also be done. ...

clustering algorithms could be partitional or hierarchical.

4.3.1 K-means clustering

it is a partitional clustering, using a distance function, it minimizes the intra-cluster distance. K is the expected number of clusters, and it has to be given to the algorithm. sometimes the number of clusters cannot be said certainly. Leads to the user the decision about the number of clusters. Takes K seeds, which arethe initial centroids, cluster centers. Assign each node to the closest centroid. Recompute the centroids and another time it is done the process. A reasonable criteria to stop the algorithm: no reassignment of data points to different clusters, minimum change of the centroids, minimum decrease in the SSE: when this distance goes under a threshold, the algorithm stops.

$$\sum_{j=1}^{k}\sum_{x\in}$$

The ingredient is the distance function, the Euclidean distance in particular, application of the Pitagoran theorem.

$$formulamaybe$$

You can think about other possible methods

- **Strengths**: very intuitive, the complexity is ..., it is a linearly complex algorithm, it is the most popular clustering algorithm.

- **Weaknesses**: you have to know k, outliers affect the decision of the algorithm. Solution: reduce the dataset excluding the outlier. Run the algorithm different times on subsets of the dataset. It is also sensitive to the initial seed, solution: aggregate results of multiple runs. It can be used in case of simple structures, round, for difficult shapes, the algorithm doesn't work well.

4.3.2 Outliers

It is difficult to find a rule to identify them, it is subjective, you decide where to cut. Interquartile range, multiply per 1.5. The 1.5 parameter can be modified, it is arbitrary, it's controversial the number usage, more robust to outliers or not.

4.3.3 Hierarchical clustering

it is another clustering method based on a distance matrix. It is based on a dendogram, a tree. A way of clustering raws and columns of a matrix. It can be constructed top-bottom or bottom-up.

The divisive clustering abd agglomerate clustering are possible.

each point generates a node of the tree, a cluster. merge 2 clusters at every level,
for each node you have to compute distance with all the other nodes. The final step is to obtain the total cluster (4 in figure **??**). The result is the dendogram.

Where are the final clusters? depends on the level of the "cut", imagine drawing an horizontal line on the dendogram graph.

How to measure distance between two sets of points. Single link method: the distance between two clusters is the distance between two closest data points. Complete link method: the distance between the furthest points Average link: a compromise Centroid method the distance between two clusters in hte distance between their centroids.
distance matrix needed. Due to complexity, it is hard to use for large datasets.

h larger than two puts more enphasys.

Part IV

Introduction to LDA

Part V

Lasso/Ridge regression

Part VI

Rank-based signatures

Part VII

Functional enrichment analysis

Part VIII

Network analysis