Bioinformatics - Summary NEW

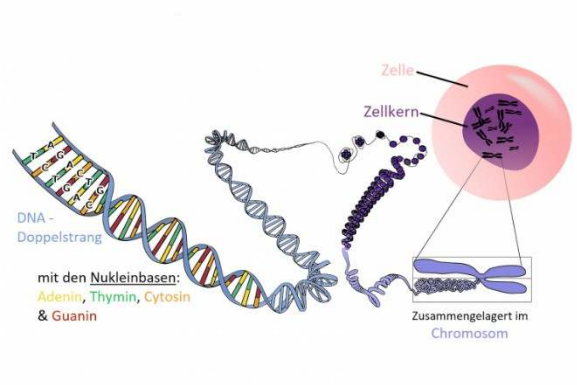
Computational methods in bioinformatics (Chalmers tekniska högskola)

MVE510 Introduction to bioinformatics

Inhaltsverzeichnis

L1: A first introduction to bioinformatics

- In the nucleus of the cell we have the DNA consisting of 23 pairs of chromosomes
- DNA carries information (can be ordered in a number of different ways)
- Chromosomes are representing the DNA sequence that is built up on our standard bases A, C, G, T
- Genes are certain patterns in the DNA that define how the proteins are translated
- The amount of information that are carried by a cell is quite large



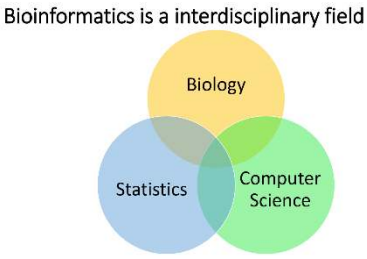
Sequencing of the human genome

Definitions

- **Genome Sequencing** = process of determining the entirety, or nearly the entirety, of the DNA sequence of an organism's genome at a single time; Characterize the exact order of ACGT -> aim to identify mutations in relation to reference
- **Transcriptome Sequencing** = process of determining the genetic codes contained in the transcriptome, and their relative proportions
- **Transcriptome** = set of genes which are transcribed; the full range of messenger RNA, or mRNA, molecules expressed by an organism; RNA is copied from pieces of DNA and contains information to make proteins and perform other important functions in the cell
- **Genome** is the collection of all DNA present in the nucleus and the mitochondria of a somatic cell. The initial product of genome expression is the **Transcriptome**, a collection of RNA molecules derived from those genes
- An organism's genes are expressed, that is transcribed from the genome's DNA code into messenger RNA (mRNA) code, and subsequently, may be translated into proteins that function within the organism's cells

- Sequencing or reading DNA: translate it into information that we can process

Bioinformatics

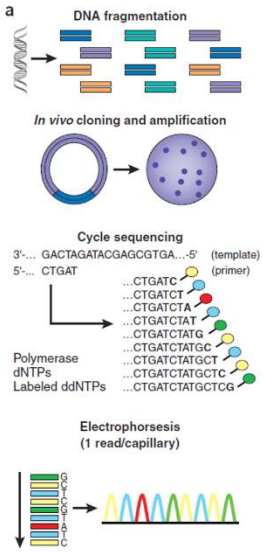


- **Bioinformatics** = analysis of molecular data
  - How should data be generated?
  - How should data be analyzed?
  - What biological conclusions can I draw from the data?
- Many large datasets generated by high-throughput measurement techniques
- Challenges: curse of dimensionality ->if you look at very high dimensional data -> challenge is to identify the correct patterns in a proper way

L2: Next generation DNA sequencing (NGS)

The history of DNA sequencing – from serial to 'massively parallel'

Sanger sequencing



- "first generation" sequencing ('rapid' sequencing)
- Originally only short fragments (80 bases fragments)
- It was the first major sequencing technique
- Shot gun sequencing: pick random DNA fragments from a longer DNA
    - Reason: sequencing technique has a limited read length
- Sequence = Read

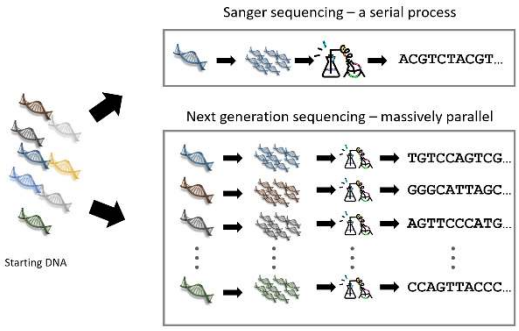
Characteristics

- Serial process – only one fragment is sequenced in one reaction
- Has a limited output and expensive
- Has a high accuracy – sequencing errors are relatively uncommon
- 'golden standard'

Process

- Done by cloning -> let bacteria grow (divides itself and produces many copies); takes a long time
- Chain-termination: start with a template; recreate the other DNA strand by using this classically pairing
- Two types of nucleotides: have a specific form -> sort according to size -> based on the color we can make out a DNA sequence
- Two issues when sequencing the DNA fragments:
    - We start with a low amount of DNA, we have quite a few of it, so in order to read them we need to amplify them
    - Do reading itself

Next generation sequencing platforms



- Second generation
- Introduced in 2006 and is today widely used
- From serial to parallel: multiple DNA sequences are 'read' in one chemical reaction
- Several platforms on the market – each have their own chemistry

Definition

Reads = the output information from a DNA sequencer

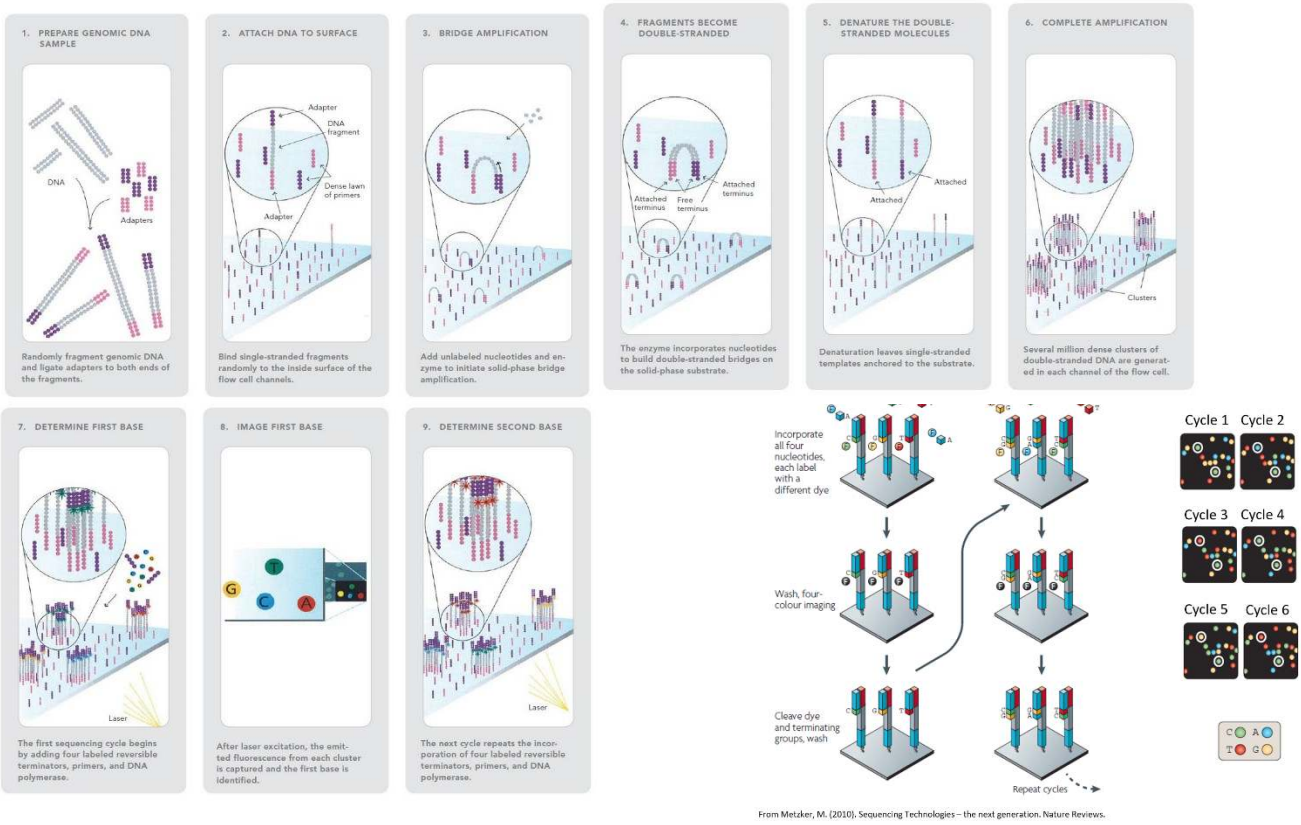
Illumina

- Introduced 2006
- High throughput, can sequence billions of fragments in a single run
- Short read lengths (up to 350 base pairs)
- Currently the most used platform

Illumina sequencing -> **EXAM: Describe briefly how Illumina sequencing works**

- Fragements are put on surface -> the surface has been prepared to that they can attach
- They will distribute on the surface with a certain distance
- Quite sparsly distributed on the surface

- Amplification process around the fragments
- Forming bridges
- Clusters of copies of the DNA fragments -> all are identical
- Clusters are separated ->optimization process: as many clusters on the surface as possible, but at the same time prevent them from bleeding into each other/interfer with each other
- They all produce the same number of copies, but not in practice
- Some sequences are underrepresented in the output due to amplification (issues with PCR)
- Use light to indicate the order of the nucleotide fragements

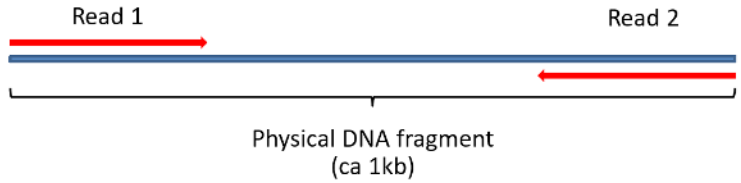


From Metzker, M. (2010). Sequencing Technologies – the next generation. Nature Reviews.

- Only exactly one nucleotide can be incorporated each cycle
- Depending on what fragment you have, you have different nucleotides that are attached here
- You get only one nuclotide attaching to each cluster
- Each bar corresponds to a cluster and not a single fragment
- Once everything is done, we take a picture of the entire surface
- Each cluster will no represent a color depending on what nucleotide was incorporated
- The more cylces you repeat, the more of the DNA fragment you sequence
- Challenge in Illumina:
    - Where do we have the clusters?
    - Identify the specific order in which the different colors appear
- Images for each cylce: look at the same cluster and its color -> come up with a aequence with that time-series
- hundreds of thousands high-resolution images are analyzed during a sequencing run
- processing of these images (terabyted of data) requires a small computer cluster

Paired-end sequencing

- Illumina sequencing can be done ‘paired-end’ where both ends of the same DNA fragement is sequenced
- This provides information about both ends of a single physical DNA fragment
- Simplifies the data anlaysis, for example the reconstruction of genome



- Sequence both ends of the same DNA fragment
- You will get data that are connected -> one read in the first part and one read in the second part of the same physical molecule
- You can get 100 of nuclotides in each end
- Paired-end simplifies a lot of the analysis -> you know that the reads are connected with a distance

Summary

Advantages	Disadvantages
- High throughput, cost efficient, especially when sequencing large volumes of DNA - Low error rate compared to other NGS platforms - Paired-end reads	- Short sequence reads (up to 350 bases, often shorter) - High error rates in certain genomic contexts (e.g. high GC-content)

EXAM: Select one long read sequencing platform and describe how it works

Pacific bioscience (PacBio)

Third generation sequencing

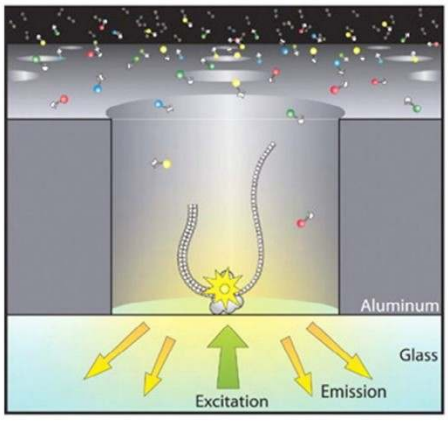
- Introduced 2011
- Can sequence long contiguous DNA fragments, typically often longer than 10,000 bases
- Many third-generation sequencing techniques do not use any DNA amplification step. Sequencing is done of individual DNA molecules
- Throughput is still lower than second generation sequencing (but the difference is decreasing). Still much higher than Sanger sequencing

Pacific Bioscience – SMRT sequencing

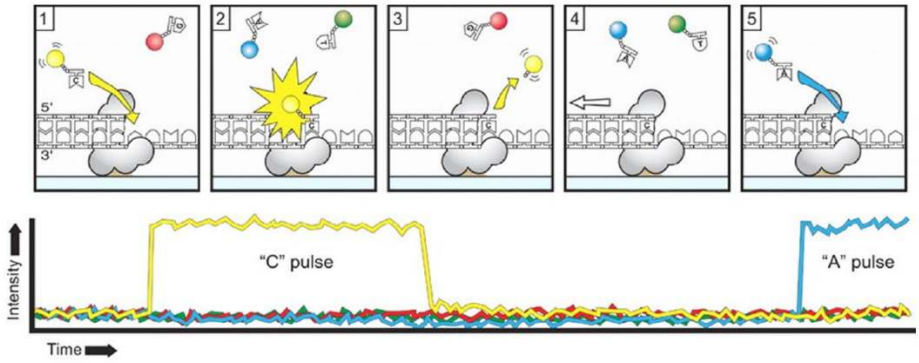
- SMRT = Single Molecule Real Time Sequencing
- Single-molecule sequencing in 'real time'
- Has a relatively low throughput but can generate very long reads (up to 100,000 bases)
- You don't have any amplification
- You sequence a single DNA molecule

Process

- You have a small well
- Amplify light for certain wavelengths
- **Polymerase** = any of several enzymes that catalyze the formation of DNA or RNA from precursor substances in the presence of preexisting DNA or RNA acting as a template
- Bottom: specific polymerase/complex that builds the opposite strand of a DNA fregment
- Polmerase is excited by lasers -> you shoot it with photons
- By integrating nucleotides that flow array you will get flash of a specific wavelength once the nucleotide is incorporated -> this is decided by the label



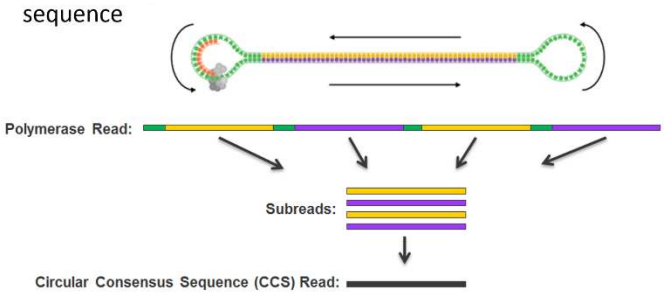
- You can follow the incorporation of nucleotides in real time
- Time series with pulses at different wavelengths depending on what nuleotides you have inserted:



-

Pacific Bioscience – HiFi reads

- PacBio can sequence the DNA fragments multiple time to increase accuracy
- The resulting 'subreads' can then be combined into a consensus sequence



Summary

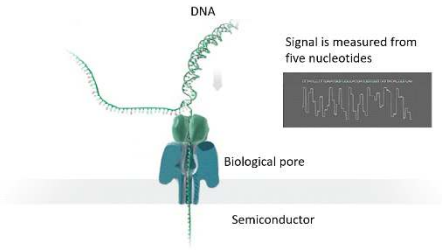
Advantages	Disadvantages
- Long sequence reads - Single molecules – no PCR amplification necessary - Error patterns are more random	- Lower throughput, relatively expensive - Error prone (>10% seqencing errors) - Requires DNA of very high quality

Oxford Nanopore – MinION

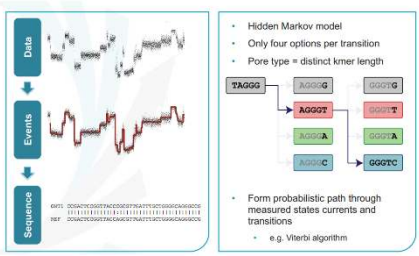
- Introduced in 2015; newest
- Single molecule 'real-time' sequencing by pulling the DNA strand through a biological pore
- The MiniION sequencers weights 90 grams and is connected to the USB port of a computer
- Adaptive Sampling: select the fragements you want to sequence. Done by reading the fragments in real-time and reverse the process if has already been sequenced before

Process

- In the semiconductor you have a biological pore where you can pull a single strand of DNA through
- By putting a potential this results in currents that are induced and measured by the semiconductor



- Current differs depending on the nucleotides that is pulled through the pore
- The big challenge is to get resolution: you can not measure differences in current based on a single nucleotide, you can only measure it based on five nucleotides, but if you have the patterns of five nucleotides you can go back and recreate the DNA sequence in almost all case



From https://konradpaszkiewicz.wordpress.com/2014/04/10/nanopore/

Summary

Advantages	Disadvantages
- Long sequence reads (up to 100kb and beyond) - Single molecules – no PCR amplification necessary - Easy sample preparation - Protable – sequence DNA anywhere?	- Lower throughput, relatively expensive - Error prone. Problems with homopolymers - Requires DNA of high quality

Summary of NGS platforms

- Compared to Sanger sequencing ('golden standard', traditional DNA sequencing), the next generation sequencing platform has higher throughput, but the data is generally more prone to errors
- The most common short read platform is Illumina, which also has the highest throughput and the lowest cost
- Third generation platforms, especially PacBio and Oxford Nanoprobe, are rapidly growing in popularity and offers read lengths > 100kb

Second generation DNA sequence platforms

- High throughput
- Short reads
- Low error rate (still higher than Sanger sequencing)
- Requires amplification (process by which a nucleic acid molecule is enzymatically copied to generate a progeny population with the same sequence as the parental one)

Third generation DNA sequence platforms

- Lower throughput
- Long reads
- High error rate
- Real time and single molecules

EXAM: Describe the main differences between traditional DNA sequencing (e.g. Sanger sequencing) and next generation DNA sequencing (e.g. Illumina sequencing). Describe the characteristics and error patters of sequence reads generated by the Illumina platform.

	Sanger sequencing	Next generation sequencing
Read Length	Only short fragments Shot gun sequencing: pick random DNA fragments from a longer DNA (Reason: sequencing technique has a limited read length)	Short read lengths
Serial vs Parallel Process	Serial process – only one fragment is sequenced in one reaction	From serial to parallel: multiple DNA sequences are 'read' in one chemical reaction
Throughput	Has a limited output	High throughput, can sequence billions of fragments in a single run and cost efficient
Cost	Expensive	Cost efficient, especially when sequencing large volumes of DNA

Errors	Has a high accuracy – sequencing errors are relatively uncommon	High error rates in certain genomic contexts (e.g. high GC-content)

EXAM: Give two examples of application areas where next generation sequencing has had a major impact. Describe also why. (3p)

Applications of next generation sequencing

- Genome sequencing
- Transcriptomics (RNA-seq)
- Metagenomics

Drastically reduced the cost of sequencing, provides high speed and throughput that can produce and enormous volume of sequences, determination of the sequence data from single DNA fragments of a library that are segregated in chips, avoiding the need for cloning in vectors prior to sequence acquisition

EXAM: Describe the difference between short and long reads. Name one sequencing platform that produces short reads and one that produces long reads. Name one advantage and one disadvantage with long reads./ Summarize the advantages and disadvantages of using long reads generated by current sequencing platforms in relation to using short reads.

- **Short reads:** fragments up to 350 nucleotides, a single read will only partly cover a gene/exon/intron
  - First generation sequencing platforms: Sanger sequencing (80 base fragments)
  - Second/Next generation sequencing platforms: Illumina
- **Long reads:** do not cover complete chromosomes, but typically often longer than 10,000 bases -> no PCR amplification necessary
  - Third generation sequencing platforms
    - Pacific Bioscience (PacBio)
    - Oxford Nanopore – MinION
  - Advantage
    - No DNA/PCR amplification step – sequencing is done of individual DNA molecules
    - Real time and single molecules
    - Long sequence reads
  - Disadvantage
    - Lower throughput, relatively expensive
    - Error prone
    - Requires DNA of very high quality

EXAM: Describe also at least one application where long reads are preferred over short reads.

- assembly of complete genomes out of complex microbial communitites,
- scientists need reads long enough to span whole genes or complex genomic elements such as disease-causing repeat expanisions or structural variants

L3: Sequencing errors and Preprocessing of NGS data

Data from next generation sequencing (NGS) / Challenges in the analysis of NGS data

- Next generation data is often *highly fragmented*
- Short read fragments are up to 350 nucleotides. A single read will thus only partly cover a gene/exon/intron. Long reads will not cover complete chromosomes
- **Shotgun sequencing:** the reads are randomly selected from the input DNA
  - Random genomic region
  - Random DNA strand
- Multiple reads are thus necessary to full describe a genomic region of interest
- Next generation sequencing generates *large volumes* of data!
- NGS is *error-prone* and contains many forms of inaccuracies:
  - Substitutions
  - Insertions/deletions (indels)
  - Duplicated reads
  - Adapter contamination
- The type of error depend on the *sequencing chemistry*. Different sequencing platforms therefore produce different forms of errors.
- As many error as possible needs to be removed before the data can be trusted

Error patterns

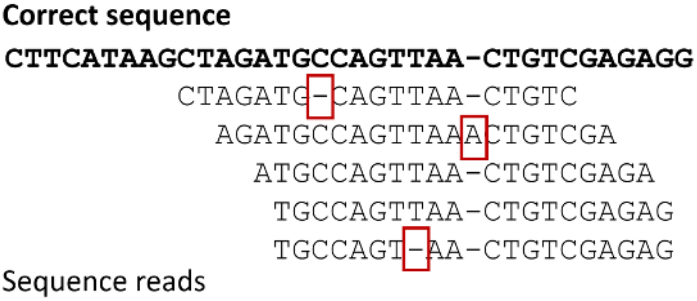
EXAM: Describe the meaning of the following types of errors: 'substitution', 'indel' and 'duplicate

Subsitutions

Real nucleotides have been substituted for incorret ones

Insertions and deletions ('indels')

Defined from the relation of the read

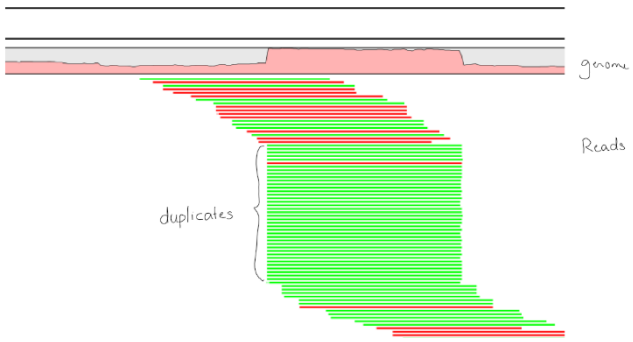


Duplicates

- Duplicates are caused by sequencing the same physical DNA fragment multiple times. These reads all come from the same DNA molecule and do not describe the true diversity in the sample

- Duplicates are typically caused by biases in the amplification where certain DNA fragments are amplified with higher efficiency. More amplification often means more duplicates -> you have multiple copies of the same DNA fragement in multiple reads
- Reason: lack of diversity/variability in the starting material
- In many applications, duplicates are important to remove to avoid incorrect and misleading results#

- Reads have been matched to the genome
- Various starting positions for the reads -> shotgun
- Red/green: one strand of the DNA, two different strands
- Duplicates: other errors can seem to appear more common



Adapter contamination

- Most sequencing platforms adds adapters to the ends of the reads
- Typically, one of the adapters are sequenced and needs to be removed before analysis
- However, if the DNA fragment is too short, the sequencing process can start to also sequence a part of the other adapter

EXAM: What is 'adapter contamination'? Why is this an issue and where does it come from?

Adapter contamination is when parts of the adapter are still present in the reads. Removal of the first 5' adapter is straight-forward. When the DNA fragments are very short, there is a risk that we also read parts of the 3' adapter. This is much hard to remove and may result in adapter contamination.

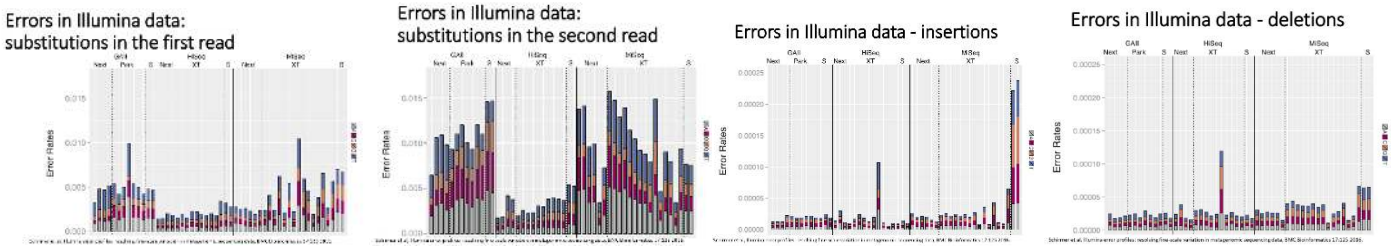
Adapter contamination

- Result from sequencing DNA fragments that are too short.
- The read will then include parts of the adapter in the end.
- Can be a large issue when working with degraded DNA.
- Needs to be removed in order to avoid errors in the down-stream analysis

EXAM: Describe the characteristics and error patterns of sequence reads generated by Illumina platform

Errors in Illumina data

- Illumina sequencing has an average error rate up to 1%:
- high error rate for substitutions and a low error rate for indels. GC-rich regions has the highest error rate.
- The error rate is dependent on
    - The position of the read. The probability for an error increase for each sequenced base pair
    - The genomic context. T has a higher error rate than A, C and G. GC-rich patterns, such as XGG has a higher error rate
    - In paired-end sequencing, the first read has a lower error rate then the second read
- Duplicates can also be common in Illumina data if the sample preparation is not done properly (or if the starting material is limited)



Substitutions in the first read (Paired-end sequencing)

- Mix different bacteria
- Investigate actual error rate in Illumina
- 3 different Illumina plattforms
- 3 regions: each bar corresponds to a unique dataset
- Total error rate -> multiply by 100 to get percentage
- Huge differences in the errors for the different nucleotides

Substitutions in the second read (Paired-end sequencing)

- Trend that blue one is bigger, followed by gray, G, orange one is the smallest

Insertions

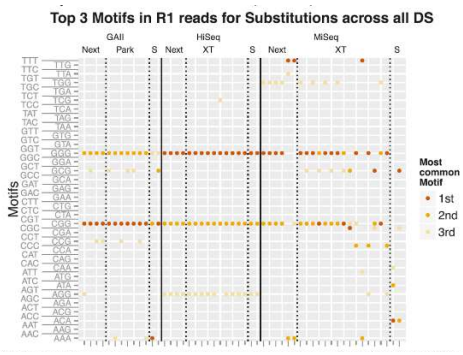
- Super rare error in Illumina data

Deletions

- Substitutions are a problem, they increase on the second read while insertions and deletions are much rare, appear to a much less degree
- T are more problematic
- Contxt also has an influence

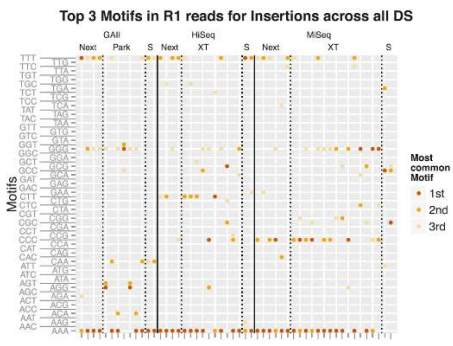
EXAM: Site-specific errors (SSEs) can have a significant impact on the performance of a mutation caller. What are SSEs and how can they influence the results?

Site specific errors (SSE)



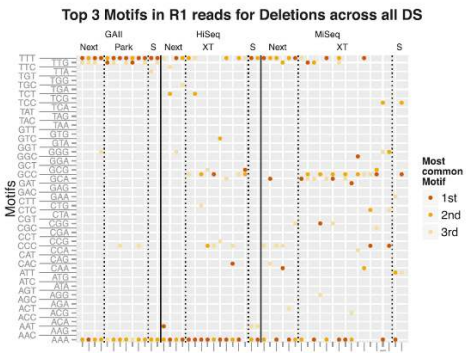
Substitutions

- Y-axis: all possible triplets of nucleotides before the error
- For each column we have three dots -> shows which of the triplets have the highest error rate
- G's can result in sequencing errors for Illumina
- TGG, GGG, CGG, AGG



Insertions

- issues are A's
- A has lower binding energy than G
- AAA



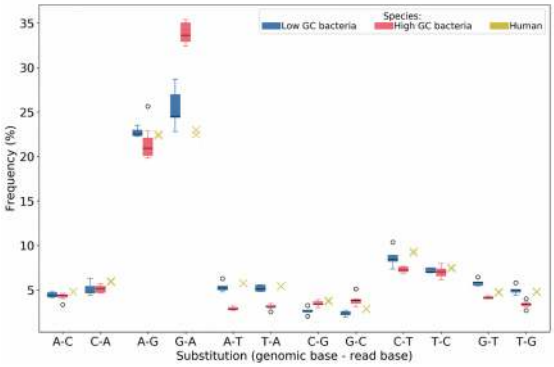
Deletions

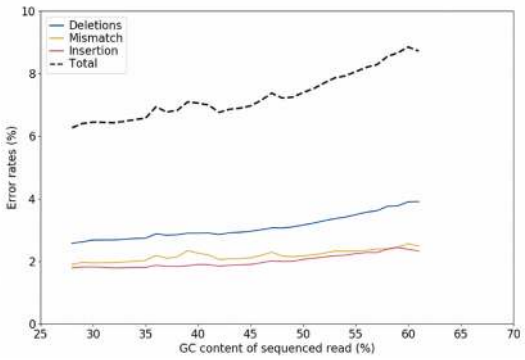
- Correltaions with the context when it comes to the errors
- TTG, AAA

Errors in PacBio data

- The error rate in PacBio data is high, up to 15%
- The errors do not dependent on the context and uniformly distributed over the reads
- Sequencing the same region many times is therefore an efficient way to remove errors in PacBio data

Errors in Nanopore data





Preprocessing of NGS data: quality and filtering

Pre-processing is the first step used to 'clean' NGS data

- Identifies erroneous reads and base pairs
- Cleans data by removing errors: remove reads and bases that are not of sufficient quality
- Important to ensure a correct down-stream analysis

Qualiy scores (Phred scores)

Probability of an error at a specific nucleotide

Describes the probability of errors. If a base *i* has an error probability *p* the quality score *q* is given by

$$q = -10 \log10(p)$$

If a base *i* has a quality score *q* the error probability is given by

$$p = 10^{-q/10}$$

$$q = -10 \log10(p)$$
$$p = 10^{-q/10}$$

Error probability (*p*)	Accuracy	Quality score (*q*)
0.25	75%	6
10^{-1}	90%	10
10^{-2}	99%	20
10^{-3}	99.9%	30
10^{-4}	99.99%	40

- The quality scores are encoded to save disk space
- **Encoding rule:** 33+quality score
- ASCII table: describes the underlying numbers
- Make quality scores into numbers
- Produce a character list where one single character corresponds to a quality score which in turn corresponds to an error probability

EXAM: Describe two ways of removing errors from short read data from the Illumina platform. Discuss also their strengths and weaknesses.

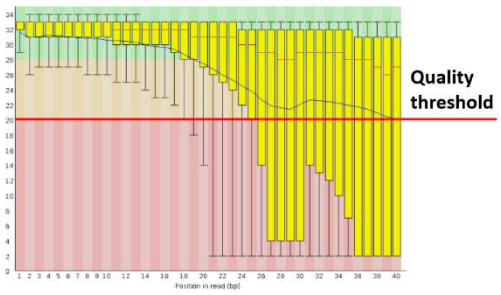
Pre-processing: Removal of sequencing errors

- Three main approaches
  - **Filtering:** completely remove bad reads
  - **Trimming:** remove problematic parts of the reads
  - **Correcting ('denoising'):** correct errors encountered in the reads
- Other forms of errors can also be corrected
  - Adapter contamination
  - Removal of read duplicates

EXAM: What is the difference between trimming and filtering of reads? What is the end result of trimming and filtering, respectively?

Filtering -> completely remove bad reads

- The overall score of a read is calculated. Reads with an overall score below a pre-specified threshold are removed from the analysis.
- Common thresholds:
  - Minimum score over the read (or a proportion of the read),
  - Average score over the read,
  - Average or minimum score over a 'sliding window' (e.g. 50 bp).
- Many reads have only a low quality in certain regions. Filtering of reads may therefore throw away good data



Trimming -> remove problematic parts of the read

- Removal of regions that are bad
- For Illumina data, trimming is done from the end of the read
- The read is trimmed until a quality level is achieved

→ There is a trade-off between data volume and quality. The stricter we filter/trim, the more reads will be discarded but the pre-processed data will have a higher quality

Summary Pre-processing

- Many sequencing errors in Illumina data still have a high score!
- Substitutions
  - >50% of the errors have a high error score (>30).
- Indels
  - >10% of the errors have a high error score.
- Pre-processing of data is thus not a guarantee that the data is error free!
- NGS data contains a lot of errors and needs to be pre-processed to remove incorrect reads and bases
- The pre-processing is based on the quality score, which estimates the probability that a specific nucleotide is incorrect
- Common pre-processing approaches are filtering and trimming of reads
- There is a trade-off between data volume and quality. The stricter we filter/trim, the more reads will be discarded but the pre-processed data will have a higher quality

EXAM: What is pre-processing? What is often necessary to perform before further analysis of the data?

Pre-processing aims to remove sequencing errors and other artifacts from the data. These are introduced in the sequencing process, often at a higher rate than traditional Sanger sequencing. Pre-processing will thus improve the overall quality of the data, which is necessary to avoid errors and misinterpretations in the down-stream analysis.

EXAM: Name one method that is used for pre-processing of DNA sequence data. Describe how it works and what it aims to achieve. (2p)

One method used for pre-processing is the trimming of reads. Since many sequencing techniques (e.g. Illumina) has a lower quality in the end of the reads, the last nucleotide can be removed ('trimmed') to increase the overall quality of the read. The trimming procedure typically works by removing nucleotides until the quality is sufficiently high, e.g. on the quality of the following nucleotide is above a preset cut-off. A drawback of this approach is that the trimmed reads will be shorter in length.

EXAM: Describe the meaning of the following types of errors: 'substitution', 'indel' and 'duplicate'

Substitutions means that a nucleotide has be incorrectly substituted by another nucleotide. An indel means that a nucleotide has either been deleted or inserted. A duplicate is multiple sequencing of the same physical DNA molecule that gives rise to identical reads.

Substitution: One nucleotide is incorrectly exchanged for another one

Indel: A nucleotide is incorrectly inserted or deleted

Duplicate: The same DNA fragment is sequenced several times. Often a result from biases in the amplification before sequencing. Errors in the amplification step can appear in a subset of the duplicates and can be mistaken for mutations

EXAM: Site-specific errors (SSEs) can have a significant impact on the performance of a mutation caller. What are SSEs and how can they influence the results?

Site-specific errors are errors that depend on the genetic context, i.e. the DNA sequence around the error. Some sequencing techniques, such as Illumina, have issues with site-specific errors, where the error rate can reach high levels in association with specific sequencing patterns (e.g. GGX-motifs). If not taken into account, SSE can have a negative impact on the results. In particular, they can, due to the high error rate, be misclassified as true mutation and can thus lead to false positives and, potentially, incorrect biological interpretation.

EXAM: Scientists have identified a strain of rye that is more tolerant to drought and they use next generation sequencing to to look for mutations. Rye is a diploid organism with a 8 gigabases long genome. A reference genome is available.

(a) Describe a suitable experimental setup for this project. What sequencing platform do you recommend? What coverage should be used? (3p)

- Illumina platform
- 72.3-fold coverage, 14 to 15.4-fold coverage

(b) Desribe how you would analyze the generated data. Make sure that your methods match the selected sequencing platform (3p)

L4: Genome sequencing

Genome evolution – SNPs, indels and structural variation

EXAM: What is the difference between germline and somatic mutations? Which type would be of main interest in the study?

- The genome of an organism is evolved through random mutations
- For humans and many other multicellular organism mutations can be either germline or somatic
- **Germline mutations**
  - Inherited from the parents
  - Passed to the offspring (present in germ cells)
- **Somatic**
  - Mutations that are not inherited
  - Not passed to the offspring

SNPs and indels

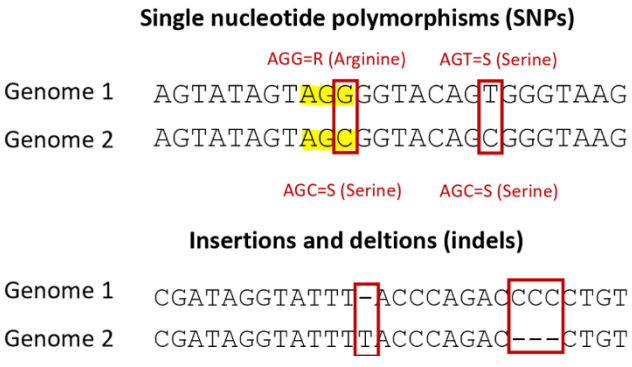
Definition

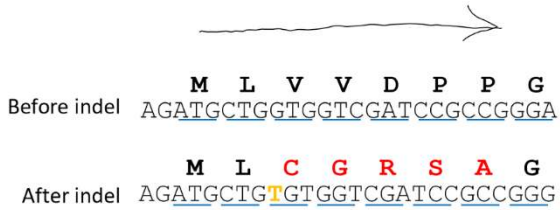
Single nucleotide polymorphisms (SNPs) = substitutions affecting a single nucleotide position

Indels = insertion or deletions affecting one or a few nucleotides

- If a SNP is in a coding region and results in an amino acid change it is called *non-synonymous* otherwise it is called *synonymous*
  - Synonymous= located in such a way that it does not change how your genes work due to the redundancy in your genetic code
- Indels can result in frameshifts which can impair a complete gene
  - Frameshift = shift coding sequence of a gene; amino acids are coded based on triplets/nucleotides and if you insert one single nucleotide in this sequence this triplet will become out of phase or it will end in a different frameshift
  - the protein that is coded there looks completely different
  - Frameshift can remove the function of a gene

- SNPs
  - Two different shifts marked with red boxes
  - Left: nonsynonymous mutation, change of protein -> change behaviour of cell
  - Right: changes only the DNA but not any active building blocks, encodes for same amino acid
  - Change in last part in code -> synonymous
  - In the beginning -> more likely to be nonsynonomous
- Indels
  - Left: one insertion
  - Right: three deletions
- Frame shifts
  - T is inserted -> shift of the frame, from here on the sequence is shifted, coding will be completely different, changes the function of the proteine





- One single adding or deletion can make a huge difference

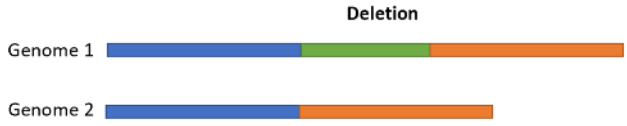
Mutations: Strucutural variants

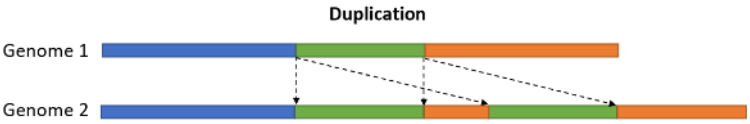
Copy number variation

- A chromosomal region that is **duplicated** or **deleted**
- Can be caused by many mechanisms including impaired DNA replications mechanisms (e.g. in cancer)
- Used by e.g. bacteria to regulate **gene expression** – more copies of a genes means higher expression
- Important in evolution: the genes in the new region can evolve into completely novel biochemical functions (neofunctionalization)

Genome evolution

- **Deletion:** green part has been removed
- **Duplication:** two exact copies of the green part



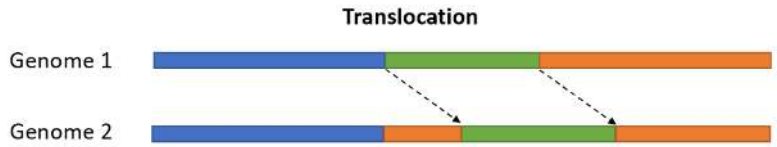


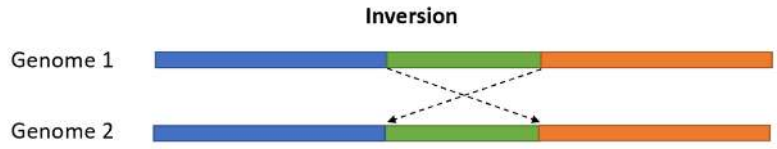
Chromosomal rearrangements

- A chromosomal region that has been **removed** and **inserted** at another place in the genome
- Can be caused by e.g. impaired DNA replication mechanisms
- Also important in evolution. Can give rise to **fusion genes** i.e. two different genes that have been fused together. If functional, the new product can beneficial or harmful

Genome evolution

- **Translocation:** One piece of DNA is simply moved to another position
- **Inversion:** You have a piece of DNA and you simply shift it around -> you get a completely different sequence





EXAM: A polyploid organim has two or more copied of their chromosomes. Explain why mutation calling is often more challenging in a polypoid organism compared to haploid organism (organism with one chromosomal copy)

Frequency of mutations

Haploid organisms (e.g. bacteria and many single cell eukaryotes)

- *One copy of the chromosome:* Mutations are present (100% of the DNA) or absent (0% of the DNA)

Diploid organisms (e.g. humans)

- *Two copies of the chromosome:* Mutations are present (100%, homozygous), present in one chromosomal copy (50%, heterozygous -> random, can be problematic even if it only affects one chromosome) or absent (0%) of the DNA

Polyploid organisms (>2 copies)

- *Many copies of the chromosome*: A wide range of mutation frequencies are possible (fern has for example 630 chromosome copies!!)

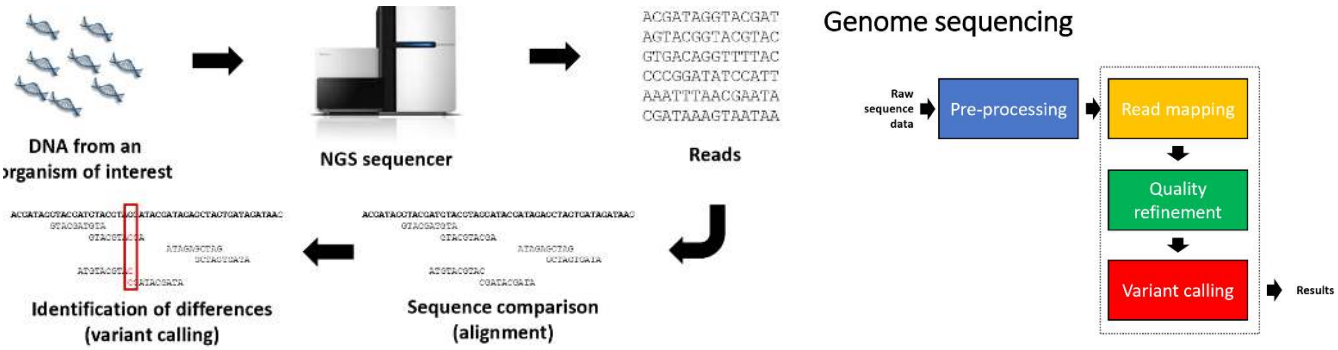
Note that germline mutations are (almost always) present in all the cells while somatic mutations are present in a subset of the cells. In a sample with many cells, somatic mutations has thus typically a lower frequency.

Analysis of data from genome sequencing

➔ Dependent on three main steps: Read mapping, quality refinement, SNP calling

Genome sequencing

- SNPs, Indels and structural variants can be identified by whole genome sequencing
- Genome sequencing aims to characterize DNA in an organism to study the presence of mutations
- The process is reference-based where the DNA is compared against a reference sequence
- Reads: represent random parts of the DNA
- Match genes towards the reference: Once they have been matched we look for differences -> indicate mutations compared to the reference



- **Pre-processing:** make data as high quality as possible
- **Read mapping:** where do the reads actually match the reference
- **Quality refinement:** handling errors
- **Variant calling:** make a decision -> is the change you see an actual mutation or due to a sequencing error, statistical question, so we need a statistical method; Where do we see mutations and where noise?

Read Mapping

- Aims to identify where each sequence read match the reference genome
  - Chromosome and position
  - Strand
- Take differences between the reads and the reference into account
  - Biological differences between the sequenced genome and the reference

- Sequencing errors in the reads

Coverage, quality refinement, score recallibration

EXAM: How is the calling of SNPs affected by the coverage? What is a suitable coverage?

Coverage

Definition

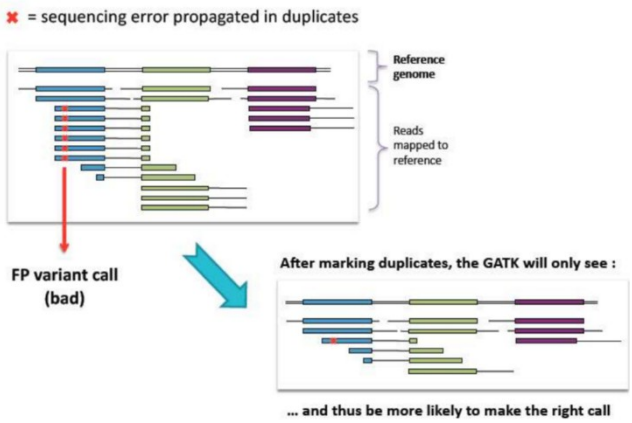
Coverage = number of times a nucleotide in the reference is 'covered' by reads
Average coverage = average coverage of all nucleotide positions in the reference

- Higher coverage means that there is more information and thus higher accuracy in identifying mutations
- The coverage depends on the experimental design, the amount of sequencing data generated, quality of the sequencing data etc, but is typically in the range 15x-200x

Quality refinement

- There are several errors in NGS data that need to be removed to reduce the number of false positives.
- Errors that are particularly problematic in genome sequencing include
  - Duplicates
  - Incorrectly aligned reads
- The **quality refinement** step aims to remove errors in the data and errors introduced in the read mapping.

- Duplication errors
- If there is an error all the copies will also have it and if we sequence it multiple times we get reads with many errors
  - -> can easily be identified as mutations
  - -> should be only counted once
- Correspond to one physical DNA fragment
- Identify duplicates and merge them together



Sensitive read matching can remove 'artificial' mutations

Quality Score recalibration

- Correct quality scores are critical for downstream analysis. Systematic biases may contribute to false results when calling variants.
- The quality scores are therefore often recalibrated. This is done by first dividing the reads into groups based on its sample, sequencing run, sequencing lane, dinucleotide context, etc.
- The recalibrated quality score is then calculated based on the actual mismatch frequency. Common SNPs are removed in this analysis.

EXAM: One important analysis step in genome sequencing is the calling of single nucleotide polymorphisms (SNPs). Provide a brief description of calling of SNP and discuss briefly the main statistical challenges. (3p)

Variant calling of SNPs

- **Variant calling** aims to identify SNPs in the sequenced genome compared to the reference
- SNP calling is done by statistical modelling of the read data and its errors
- This is done by analyzing each nucleotide position in the reference and compare it to the data (the reads)
  - We go to every single position and decide if we have a mutation there or not

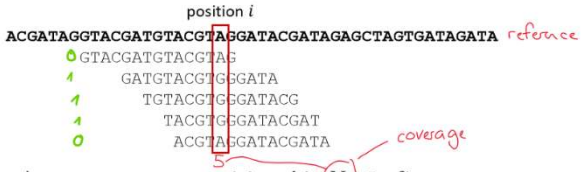
- The aim is to distinguish between true mutations and sequencing errors
- A good caller should have a high sensitivity (find all true mutations) and a high specificity (ignore all false positives)

Challenges

- Sequencing error rate is high, especially in certain genomic regions
- The coverage is varying over the genome and may be low in some regions
- A large number of positions needs to be analyzed
- Are the changes a result of a true mutation or only the results of many reads with sequencing errors?
    - Curse of dimensionality: mistake randomness for true patterns

A naïve approach: a naïve variant caller for SNPs

We will use a binomial test to detect SNPs in a genome. Assume that we are interested in analyzing position i and that the reference has an 'A' at this position.



Assume that the coverage at position i is N_i. Define

$$X_{i,j} = \begin{cases} 1 & \text{if read } j \text{ at position } i \text{ an not an A,} \\ 0 & \text{if read } j \text{ at position } i \text{ is an A.} \end{cases}$$

- Binominal test to identify SNP -> which position does actually have a mutation
    - All nuclotides are changed -> mutation
    - A few -> sequencing error
- Compare if they match to the reference
    - 1-> mismatch
    - 0 -> correct

Let

$$Y_i = \sum_{j=1}^{N_i} X_{i,j}\,.$$

Y_i is the total number of reads that does not match the reference at position i.

If the reads and their errors are independent, it follows that

$$Y_i \sim \text{Bin}(N_i, p_i)$$

where p_i is the probability of observing another base than the reference (i.e. not an "A") at position i.

- Summing all mismatches up to the total coverage, all reads at that position
- Sum of number of reads that does not match the reference - > if we have a lot of agreement, Y will be small
- Errors between each read are independent (assumption)
- How big is pi?
    - If pi is small it speaks in favor of sequencing errors
    - If pi is big we have a mutation

We can assess if there is a mutation at position i by using a statistical tests:

$$H_0: p_i = p_{\text{error}}$$
$$H_1: p_i > p_{\text{error}}$$

Here, p_{error} is the probability of a sequencing error, which is the lowest value p_i can take (if there is no mutation we only observe sequencing errors).

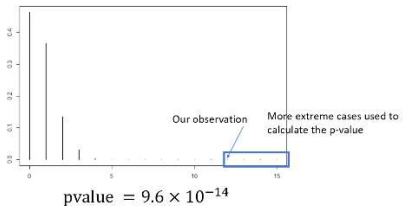
Under the null hypothesis, a p-value for the test can be calculated by

$$\text{p-value} = \text{Prob}(Y_i \geq y_i) = \sum_{j=y_i}^{N_i} \text{Prob}(Y_i = j)$$

- Mutation: pi should be higher than perror
- **Naive:** p_{error} does not change for different positions, it actually depends on the context, same error rate at each position
- We will calculate how extreme is the observation, given the fact that pi is equal to the error rate
- P-value: probability that our y value is larger than what we observed, more extreme

Example

Assume that the sequencing error for a specific position is $p_{\text{error}} = 0.05$ (5%) and the coverage is 15. If 3 reads are matching the reference while reads 12 have a mutation (say a "C" instead of an "A").



$$\text{pvalue} = 9.6 \times 10^{-14}$$

- Observation: 12
- **P-value:** how extreme the value is compared to what we expect under the null hypothesis that we only have sequencing errors here, we sum up this observation and all more extreme observations (blue box)
    - Small p-value: very unlikely to see this observation if it is only caused by sequencing errors -> mutation
    - Probability of Y taken a specific value
    - How far are we that this is only caused by sequence error

<u>Why is this caller naïve?</u>

- The assumptions about the errors are not true
    - The error rate differs considerably between positions
    - The errors are not independent – they depend on the context
- We do not use any information about the quality of the sequenced nucleotides
- The caller does not distinguish between homo- and heterozygosity which will make it inefficient in diploid genomes

compensate by using an error rate that is a bit higher
-> test will be a bit more conservative
-> more data is needed to say if you have mutations
-> easier to handle situations where you have a higher error rate

EXAM: Describe one statistical approach to identify SNPs and discuss its advantages and weaknesses.

The GATK unified genotyper

GATK (Genome analysis toolkit) contain the unified genotyper, which is a more advanced mutation caller.

The unified genotyper calculates

$$\text{Prob}(G_i|D_i)$$

for each position i. Here G_i set to any of the possible genotypes (AA, AC, AG, AT, CC, CG, CT, GG, GT and TT for a diploid organism). D_i denotes the data available at the position i.

- Genotypes: all possible combinations that you can have between your chromosomial copies
- Diploid -> 10 possibilities
- Pick one with highest probability -> bayesian approach instead of test

Using Bayes theorem, this can be rewritten as

Model

$$\text{Prob}(G_i|D_i) = \frac{\text{Prob}(D_i|G_i) \times \text{Prob}(G_i)}{\text{Prob}(D_i)}$$

Prior knowledge

where

$$\text{Prob}(D_i) = \sum_i \text{Prob}(D|G_i) \times \text{Prob}(G_i)\,.$$

The sum is calculated over all 10 genotypes.

The unified genotyper assumes that

$$\text{Prob}(D_i|\,G) = \prod_{r \in \{\text{good read bases at } i\}} \frac{\text{Prob}(r|H_1)}{2} + \frac{\text{Prob}(r|H_2)}{2}$$

where $G = H_1H_2$ and, with ε_i=error probability of position i,

$$\text{Prob}(r|H) = \begin{cases} 1-\varepsilon_i & \text{if r} = H \\ \varepsilon_i & \text{if } r \neq H \end{cases}$$

ε_i is derived from a site-specific error model.

- The unified genotyper has several advantages compared to the naïve caller.
  - Only reads of sufficient quality are included
  - A more sophisticated error model is used. The quality values are taken into account
  - The probability of each possible genotype is estimated
- GATK also offers information on 'best practices' in mutation calling (description how exactly you should do your bioanalysis)

Post-processing – filtering of variants

- Many genome sequencing experiments results in a very long list of variants that may need to be filtered before it can be interpreted.
- Criteria for filtering typically includes
    - **Strand bias**, i.e. variant bases only in one read direction
    - **Clustered position**, e.g. variant bases always at the end of reads
    - **Poor mapping**, i.e. variant bases only in reads with lower mapping quality

- All mutations in the end of the reads -> false positive, structure issue
- Different strand have different error probabilities -> Direction matters in sequencing to determine the error
- All the mutations on one single strand

- It is also possible to filter variants based on their biological function and impact
- Such criteria include
    - Synonymous/non-synonymous
    - Commonness among healthy individuals in the populations
    - Its predicted biochemical impact: is it located in a protein domain that is likely to have a impact on the protein function
    - Previous knowledge, e.g. its association to other diseases

SUMMARY

SNP calling results in a huge data reduction. From billions of observations to, in some case, a handful of significant positions. Careful filtering to remove errors in the different steps is therefore vital.

EXAM: One important analysis step in genome sequencing is the calling of single nucleotide polymorphisms (SNPs). Provide a brief description of calling of SNP and discuss briefly the main statistical challenges. (3p)

Calling of SNPs is used to identify the position of the gene where we have a true mutation. In this process, we use the result of all read covering the position and make a decision whether any deviances from the reference is caused by sequencing errors or true mutations. The decision is typically based on a statistical model, which describes both the error pattern (i.e. sequencing errors). There are several statistical challenges in SNP calling. First, the model needs to accurately describe the error patterns, which are often complex. Most genome are also large, which means that a large number of tests needs to be performed. If this multiple testing is not controlled, there is a risk for a large number of false positives. Finally, SNP calling in non-haploid organisms is especially hard since only a proportion of the reads may contain the mutation (half of the reads for a heterozygous mutation).

EXAM: What is mutation calling? Describe briefly how a method for calling mutations works.

Mutation calling is the process where mutation in the genome is identified. An important part is to distinguish between true mutations and sequencing errors.
The process includes four main steps:
1) preprocessing,
2) mapping the reads to the reference,
3) quality refinement to remove artifacts and sequencing errors and
4) identification of mutations

EXAM: How is the calling of SNPs affected by the coverage? What is a suitable coverage?

The coverage is the number of times a specific position is covered by reads. A high coverage is necessary for accurate calling of SNPs. If the coverage is low, there is a risk that too few reads are present, which will make the decision whether any deviations are true mutations or sequencing errors much harder. Thus, a low coverage will result in lower statistical power. The coverage should be as high as possible. However, in most cases, costs set the limit. It is, however, important to set the coverage based on the fact that reads are randomly places throughout the genome

and some areas will, due to pure chance, have a lower coverage. It is often recommended that the average coverage should be at least 50x-100x. A coverage <15x is often problematic since some regions may then end up with a coverage <10x.

EXAM: What is the difference between germline and somatic mutations? Which type would be of main interest in the study?

Germline mutations are inherited by the offspring from its parents. These mutations are therefore present in all cells in the human body. Somatic mutations are mutations that appear over time and are only present locally and often in a small subset of the cells. Germline mutations are the main focus in this study.

EXAM: Describe the main bioinformatic steps necessary to identify SNPs from sequence reads generated by next generation DNA sequencing.

As always, the data is first pre-processed. After that, the reads are aligned against a reference sequence, e.g. the full human genome. Since the reads will be relatively close to the reference, the alignment is often done with fast and less sensitive index-based methods such as BWA. After that, the alignment is quality assessed. This can include recalculation of the quality scores and removal of duplicates. Regions with high dissimilarity can also be re-aligned with more sensitive methods. Next, mutations are called. This is typically done using a suitable statistical method which tests whether the difference in nucleotide distribution between the reference and observed data are caused by sequencing errors or by an underlying mutation. Finally, the list of putative mutations often needs to be quality assessed and artifacts are removed. Sometimes, the list of mutations is also filtered based on their biological function (synonymous/nonsynonymous, location within the gene, the function of the gene etc).

EXAM: A polyploid organim has two or more copied of their chromosomes. Explain why mutation calling is often more challenging in a polypoid organism compared to haploid organism (organism with one chromosomal copy)

Haploid organisms only carry one copy of each chromosome, which means that a mutation either exist in 0 or 100% of the DNA. In polyploid organisms, which have multiple chromosomal copies, mutations can be present in a subset of the DNA. For example, in diploid organisms, mutations can either exist in 0%, 50% or 100%. Identification of mutations only present in a part of the data is much harder than mutations that are present in all parts of the data. The lower the proportion, the more likely it is that it is caused by sequencing error. Indeed, identification of smaller signals has a lower power and can be more easily be confused with sequencing errors.

EXAM: In a research project, scientists are investigating germline mutations in wheat associated with tolerance to cold. Next generation sequencing is used to sequence the genome at a coverage of 20 times.

(a) The scientists aims to extend the project and they have two alternatives: 1) increase the coverage or 2) include more biological replicates (i.e. wheat samples). Discuss the advantage of both approaches. What would you recommend? (3p)

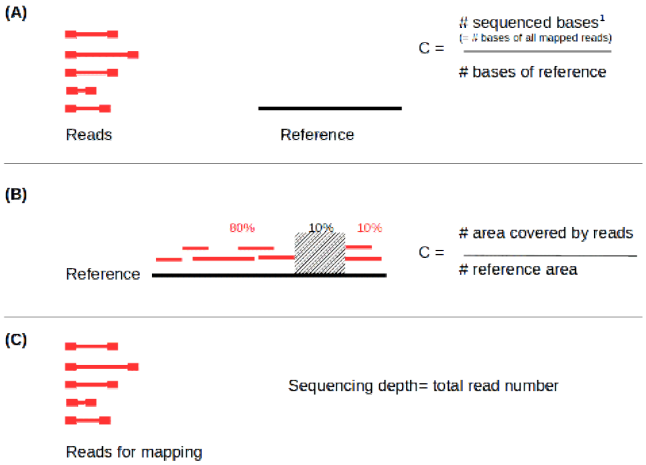
- The greater the number of replicates, the greater the precision of the experiment. It is necessary to repeat an experiment on a large number of subjects to increase the significance of an experimental result. As our sample size increase, the confidence in our estimate increases, our uncertainty decreases and we have greater precision.
- Higher coverage of a given region increases the confidence in SNP identification and variant calling for that region; that is especially important for rare variants, where even infrequent sequencing errors can lead to misleading conclusions; ensure that the genomic region of interest can be studied with high confidence

(b) The scientists want to achieve an average genome coverage of 10 times. The sequencing is performed on the Illumina platform using 100 bases long paired-end reads. How many reads and how many bases are necessary to achieve an average coverage of 10? (2p)

Lander/Waterman equation: C = LN/G paired-end: C = (2*L)N/G
C: coverage

G: haploid genome length
L: read length
N: number of reads

$$average\ coverage = \frac{read\ length * number\ of\ mapped\ reads}{genome\ size}$$



(c) What is your view on a coverage of 10 times? Is it high or low? Do you see any risks with using a coverage of 10 times? (2p)

- low coverage
- A coverage <15x is often problematic since some regions may then end up with a coverage <10x.
- If the coverage is low, there is a risk that too few reads are present, which will make the decision whether any deviations are true mutations or sequencing errors much harder. Thus, a low coverage will result in lower statistical power.

EXAM: It is possible to use RNA-seq to identify mutations. This makes it possible to analysis the gene expression and find DNA alterations from the same sequencing data.

(a) What analysis steps do you think are necessary to identify single nucleotide polymorphisms (SNPs) from RNA-seq data. Describe any many differences compared to the standard analysis of SNPs based on whole genome sequencing data. (3p)

(b) Fusion genes are formed by merging parts of two different genes and can be created by genome duplications and/or translocations. If the fusion genes are transcribed they can be found in RNA-seq data. Design a mehtod to identify fusion genes from RNA-seq data. Provide an overview of the main necessary steps and identify main challenges? (3p)

Paired-end sequencing

- Illumina sequencing can be done 'paired-end' where both ends of the same DNA fragement is sequenced
- This provides information about both ends of a single physical DNA fragment
- Simplifies the data anlaysis, for example the reconstruction of genome

In paired-end sequencing, the first read has a lower error rate then the second read; substitutions

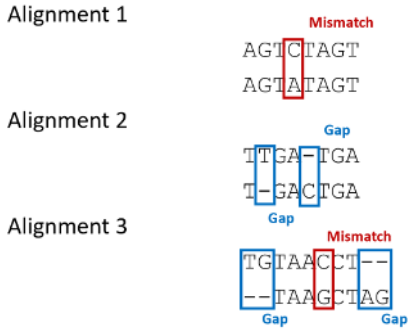
L5: Sequence alignment

Problem

- The sequence reads – many and short
- The reference – few and large

➔ How do we find where our sequence read match (‘align’) to a reference sequences? (We want the best possible alignment: the position where it actually matches our reference in the best possible way)

Pair-wise alignment of DNA sequences



<u>Two main forms of alignment</u>

- **Global alignment:** Two sequences are aligned over their full length.
- **Local alignment:** Two sequences are aligned based on their best matching subsequences.
    - Local alignments are used to match short sequence reads against long reference sequences

Score Alignment

Example

score matrix

Let $S(a,b) = \begin{cases} 5, & a = b \\ -4, & a \neq b \end{cases}$ and $d = -7$. linear gap

Score the following alignments.

Alignment 1

AGTCTAGT
AGTATAGT

S=5+5+5-4+5+5+5+5=31

Alignment 2

TTGA-TGA
T-GACTGA

S=5-7+5+5-7+5+5+5=16

Alignment 3

TGTAACCT--
--TAAGCTAG

S=-7-7+5+5+5-4+5+5-7-7=-9

Notation

Let x and y be two DNA sequences of length m and n respectively

x_i is the i^{th} nucleotide of x.

x_{ij} is the subsequence that includes the nucleotides $x_i, x_{i+1}, x_{i+2}, \dots, x_j$

Models for scoring alignments

Global alignment, no gaps ($n=m$)

We will score alignments by comparing "match models" to a "random model"

The random model (R)

Assumption: Each nucleotide $a \in \{A,C,G,T\}$ occurs randomly and independently with a probability q_a

The probability of observing an alignment then becomes $P(x,y|R) = q_{x_1} q_{x_2} \dots q_{x_m} q_{y_1} q_{y_2} \dots q_{y_n} = \prod_{i=1}^{m} q_{x_i} \prod_{i=1}^{n} q_{y_i}$

The match model (M)

Pairs of nucleotides (a,b) occurs according to a joint probability distribution p_{ab} $a,b \in \{A,C,G,T\}$

$p(x,y|m) = \prod p_{x_i,y_i}$

We will score alignment using the ratio between the probabilities of the two models, i.e.:

$$\frac{p(x,y|M)}{p(x,y|R)} = \frac{\prod_{i=1}^{m} p_{x_i,y_i}}{\prod_{i=1}^{m} q_{x_i} \prod_{i=1}^{n} q_{y_i}} = \prod_{i=1}^{m} \frac{p_{x_i,y_i}}{q_{x_i} q_{y_i}}$$

$$S = \log \frac{P(x,y|M)}{P(x,y|R)} = \log \prod_{i=1}^{m} \frac{p_{x_i,y_i}}{q_{x_i} q_{y_i}} = \sum_{i=1}^{m} \underbrace{\log \frac{p_{x_i,y_i}}{q_{x_i} q_{y_i}}}_{S(x_i,y_i)} = \sum_{i=1}^{m} S(x_i,y_i)$$

The score of the alignment is calculated from the score at each aligned pair of nucleotides

The score function $S(a,b)$ is defined over 4×4 matrix

S	A	C	G	T
A	5	-4	-4	-4
C	-4	5	-4	-4
G	-4	-4	5	-4
T	-4	-4	-4	5

We want to have a positive score if we align nucleotides → match model has a higher probability than the random model

We want to have a negative score if we do not align nucleotides → match model has a lower score than the random model

\+ match
\- mismatch

Scoring gaps

$x_1, \dots, x_i, x_{i+1}, x_{i+2}, x_{i+3}, x_{i+4}, \dots, x_m$ $\quad m \neq n$

$y_1, \dots, y_j, \underbrace{-, -, -}_{?}, y_{j+1}, \dots, y_n$

If there is a gap in sequence y of length g, the probability under the M model is assumed to be

$P(gap|M) = f(g) \prod_{k=1}^{g} q_{x_{i+k}}$ opposite: $q_{y_{j+k}}$

The contribution of the gap to the score

$$S = \log \frac{P(gap|M)}{P(gap|R)} = \log \frac{f(g) \prod_{k=1}^{g} q_{x_{i+k}}}{\prod_{k=1}^{g} q_{x_{i+k}}} = \log f(g) = \gamma(g)$$

$\gamma(g)$: gap penalty

Two common gap penalties:

$\gamma_1(g) = -gd$ linear gap score

(g: length of the gap, d: cost parameter)

$\gamma_2(g) = -d - (g-1)e$ affine gap score

(e: gap-extension parameter)

Typically: $e < d$

You can not have a gap of 1

The Needleman-Wunsch algorithm

The Needleman-Wunsch algorithm can be used to find **global alignments**. The best alignment is identified by iteratively fill an alignment matrix and backtrack from the highest value.

Finding the optimal alignment

Global alignment with gaps

Needleman-Wunsch-algorithm

Idea: Build and score the alignment iteratively

Define F as $F(i,j)$ = score of the best alignment between $x_1,\dots,x_i$ and $y_1,\dots,y_j$

1) Calculate $F(i,j)$ for all i and j

2) Use F to find the optimal alignment

$F(i,j)$ can be calculated from $F(i-1,j-1)$, $F(i-1,j)$, $F(i,j-1)$

Why?

Three scenarios:

① Aligned pair: x_i are aligned to y_j

```
          i-1
x  A C G T  xi
y  A C C T  yj
          j-1
```

$F(i,j) = F(i-1,j-1) + S(x_i, y_j)$

↳ score of the best alignment between $x_1,\dots,x_{i-1}$ and $y_1,\dots,y_{j-1}$

② Gap in y: x_i is aligned to a gap

```
          xi-1 xi
x  ... T A G C
y  ... T A G -
          yj   yj
```

$F(i,j) = F(i-1,j) - d$

③ Gap in x: y_j is aligned to a gap

```
          xi   xi
x  ... G A C -
     G A C G
          yj-1 yj
```

$F(i,j) = F(i,j-1) - d$

At each step: select the alignment that has the best score

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + S(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

Starting conditions:

$F(0,0) = 0$, $F(i,0) = -id$, $F(0,i) = -id$

Example

Align x = AGCT with y = CGCT

Where $S(a,b) = \begin{cases} 5 & a=b \\ -4 & a \neq b \end{cases}$ and $d = -5$

x

y		0	A 1	G 2	C 3	T 4
	0	0	-5	-10	-15	-20
C	1	-5	-4	-9	-5	-10
G	2	-10	-9	1	-4	-9
C	3	-15	-14	-4	6	1
T	4	-20	-19	-9	1	11

If we have gaps in y, we will progress along x not in y

(1) Starting positions

(2) Only insert gaps in y
AGCT + gap penalties
- - - -

(3) Same for x

4) 3 options:
- > align → diagonal (→ match → +5, → mismatch → -4)
- > gap in x → move position in y forward } -5
- > gap in y → move position in x forward }

Choose largest score in each position

(5) Identify next alignment: Start here and backtrack/follow the connections

AGCT
CGCT

Score: 11

Repition Global alignment

TCGGAT / TC-CAT OR TCGGAT x / TCC-AT y Score (11)

$F(i,j)$ = score of the best alignment between $x_{1,i}$ and $y_{1,j}$

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + S(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

$F(0,0) = 0$, $F(i,0) = -id$, $F(0,j) = -jd$

Note that multiple alignments can have exactly the same score
In this case, both alignments are reported

Example

Align x = TCGGAT with y = TCCAT using

$$S(a,b) = \begin{cases} 5 & a=b \\ -4 & a \neq b \end{cases}$$ and $d = -5$

using Needleman-Wunsch algorithm

Example

x

y			0	1	2	3	4	5	6
				T	C	G	G	A	T
0		0	0	-5	-10	-15	-20	-25	-30
1	T	-5	-5	0	-5	-10	-15	-20	
2	C	-10	0	10	5	0	-5	-10	
3	C	-15	-5	5	6	1	-4	-9	
4	A	-20	-10	0	1	2	6	1	
5	T	-25	-15	-5	-4	-3	1	11	

Example

			1	2	3	4	5	6
			T	C	G	G	A	T
		0	-5	-10	-15	-20	-25	-30
1	T	-5	5	0	-5	-10	-15	-20
2	C	-10	0	10	5	0	-5	-10
3	C	-15	-5	5	6	1	-4	-9
4	A	-20	-10	0	1	2	6	1
5	T	-25	-15	-5	-4	-3	1	11

The Smith-Waterman algorithm

The Smith-Waterman algorithm can be used to find **local alignments**. The best alignment is found similarly to the NW algorithm but where all negative values has been replaced by zeros and the alignment can start and end at any positions

Local alignments

Alignment of sequences of x and y

The Smith-Waterman algorithm finds the subsequences with the highest alignment score

Smith-Waterman modifies Needleman-Wunsch by

1) Alignments can start and end anywhere in the matrix

2) Scores can not be negative

3)

$$F(i,j) = \max \begin{cases} F(i-1,j-1) - S(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \\ 0 \end{cases}$$

This means that

1) All negative values become 0

2) We search for the highest value of F and backtrack as long as the score is positive

Example

Smith-Waterman

Align $x = AACTGGT$ with $y = ACGG$ using $S(a,b) = \begin{cases} 5 & a=b \\ -4 & a \neq b \end{cases}$ and $d = -4$

y		0	1 A	2 A	3 C	4 T	5 G	6 G	7 T
0		0	0	0	0	0	0	0	0
1	A	0	5	5	1	0	0	0	0
2	C	0	1	1	10	6	2	0	0
3	G	0	0	0	6	6	11	7	3
4	G	0	0	0	2	2	11	16	12

1) Replace negatives with zeros

2) Identify highest value in the matrix and start here

3) Go as deep as possible as long as the score is positive

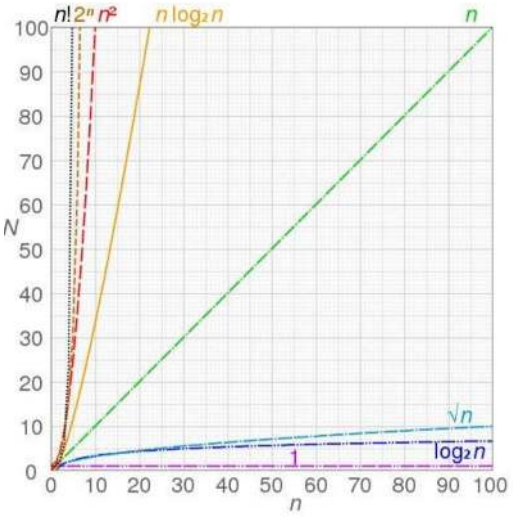
A C T G G
A C - G G

Score: 16

Computational complexity

If an algorithm is *O(n)* this means that the number of computations grows linearly with respect to *n* (which can, for example, be the number of input data).

Other examples: $O(n^2)$, *O(log n), O(nm), etc.*



N: number of computations
n: number of input data

➔ Shows how different the different functions grow

- Computational complexity: measure efficiency of computer algorithms
- How many computations are needed by the algorithm
- Scaling: how many computations do we need as a function of the input data, how does it grow as the number of input parameters grow

In our application, Smith-Waterman will have a complexity of

$$O(L_R L_G)$$

for each read. Here,

L_R= the length of the reads,

L_G= the length of the genome (reference).

This can become very slow if L_G is large (which it is in our case).

➔ Smith-Waterman are too slow and resource heavy! Can we make the alignment faster?

➔ BLAST

BLAST: The Basic Local Alignment Search Tool

EXAM: Provide an overview of the BLAST algorithm for local alignment.

BLAST can be used to match any sequence (‘query’) against a reference (‘database’).

The algorithm work as follows:

1. Create a table of words (subsequences) of size W and their location in the reference (called an index or a hash).
2. Find the position of the words present in the sequence. These positions are called ‘seeds’.
3. Extend the alignment around the seeds. BLAST is around 50 times faster than Smith-Waterman.

-> BLAST is around 50 times faster than Smith-Waterman

We go through the reference and collect all possible combinations of nucleotides and store their position, where exactly do they appear in the sequence

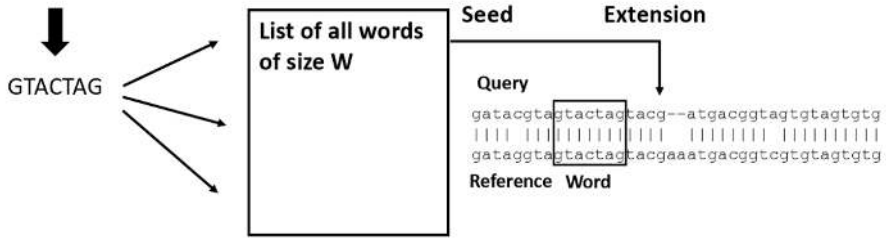
Reference ('database')
ACGAGTGAGTGCCGAGTACGTAGCGTAGGAGTGAGTTGGAGTGAGACGTGAGT

Table (hash) of words of size 7

Maximum $4^7=2^{14}=16384$ rows

Word	Position
ACGAGTG	1
CGAGTGA	2
GAGTGAG	3, 38
AGTGAGT	4, 30

Query **Word**
GATACGTAGTACTAGTACGATGACGGTAGTGTAGTGTG



- We take our read which is our query, we search the read for the specific words
- We go to the table and we look at the positions where this word actually exists
- Once we found a number of positions in the database where this word appears we use this as starting point or seed and we do the alignment in the neighborhood around the seed
- **Cost:** we require that the word is present in the query in order to an alignment
- We need to have an exact match for the word
- Cost is sensitivity

- The word size changes the sensitivity
  - Large word size means faster algorithm but lower sensitivity
  - Smaller word size means slower algorithms but higher sensitivity.
- BLAST can compare
  - Nucleotide vs nucleotide ('blastn')
  - Protein vs protein ('blastp')
  - Nucleotide vs protein ('blastx')
  - Protein vs nucleotide ('tblastn')

Summary

- BLAST uses a 'seed-and-extend' algorithm to efficiently calculate local alignments. This makes it 50 times fast than Smith-Waterman.
- BLAST is however still to slow to efficiently map reads from next generation sequencing to a reference

EXAM: Provide an overview of the BLAST algorithm for local alignment.

BLAST is an algorithm that performs local alignments but with a computational efficiency that is higher than Smith-Waterman (SW). In contrast to SW, BLAST does not evaluate all possible alignments. Instead, BLAST uses 'seed' around which the alignments are calculated. The seeds are identified by using a large lookup table where words of a pre-specified length are stored together at their positions in the reference. By identifying which words that are present in the input sequences, possible starting points for the alignments can be identified. BLAST returns the alignments that scores above a specified minimum score.

EXAM: Describe the differences between the Needleman-Wunsch and SmithWaterman algorithms. What modifications are necessary to identify optimal local, instead of global, alignments.

The Needleman-Wunsch algorithm can be used to find **global alignments**. The best alignment is identified by iteratively fill an alignment matrix and backtrack from the highest value.

The Smith-Waterman algorithm can be used to find **local alignments**. The best alignment is found similarly to the NW algorithm but where all negative values has been replaced by zeros and the alignment can start and end at any positions

Smith-Waterman modifies Needleman-Wunsch by

1) Alignments can start and end anywhere in the matrix

2) Scores can not be negative

3)

$$F(i,j) = \max \begin{cases} F(i-1,j-1) - S(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \\ 0 \end{cases}$$

EXAM: Explain the difference between a global and local alignment. Provide one example on when to use global alignments and one example on when to use local alignments.

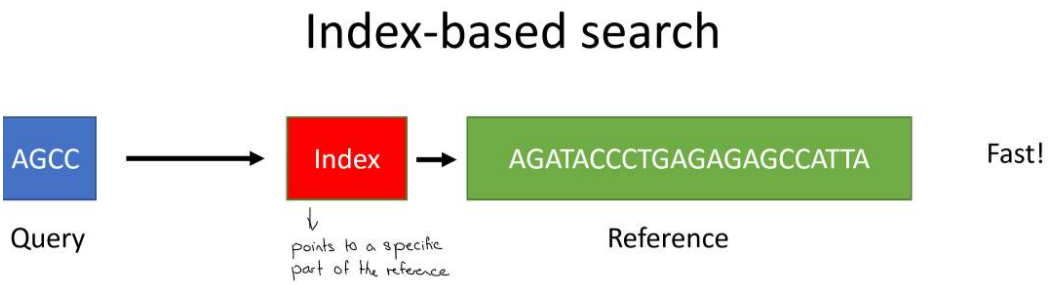
- Global alignment: two sequences are aligned over their full length
  - Needleman-Wunsch algorithm
  - Tries to find the best overall match for all sequences
  - Compare sequences in cases where we have reason to believe that the sequences are related along their entire length
  - Most useful when the sequences in the query set are smiliar and roughly equal size
- Local alignment: two sequences are aligned based on their best macthcing sequence
  - used to match short sequence reads against long reference sequences
  - Smith-Waterman algorithm
  - Looks for the best matches between two sequences; find regions of high local similarity
  - More useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequences motifs within their lareger sequence context
  - Searching for a small gene ina large chromosome
  - Detecting when a large sequence may have been rearranges

L6: Suffix trees and arrays

What is an index?

- An index is a data structure that enable fast lookup of keys (exact matches on a DNA sequence).
- In our case the data structure will built form the reference sequence and keys from our query sequence.
- Many form of indices assume that the reference is long (e.g. a genome) and the query sequence is short (e.g. a NGS read).
- Once the index is built, we can use it to lookup to find where specific read match the reference.
- We will focus on exact matches but it is also possible to include mismatches and gaps.

Index-based read search



Suffix trees

Suffix trees are created from the suffixes of a DNA sequence. By traversing the tree, we can find the position of any subsequence

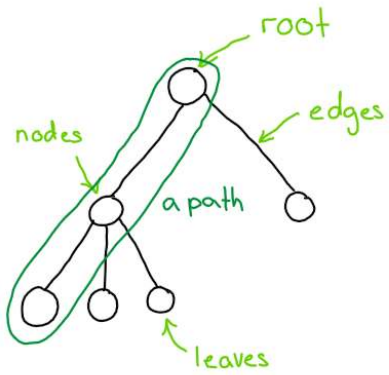
- Are highly efficient to find exact matches in a sequence. The complexity is $O(L_R)$ for each read.
- Note that L_G is gone (!!!). The size of the reference does no longer matter!

BUT…. (there is no free lunch!)

- Building the suffix trees takes time ($O(L_G)$) but we only need to do this once.
- Suffix trees are very large. The suffix tree for the entire human genome is so large that it can not be stored in the memory of a standard computers.
- If we want to include mismatches and gaps, the complexity increases fast!

Definition: Tree

A rooted tree is a datastructure that is made from nodes that are connected with edges and that does not have any cycles



The top node is called the root

The end nodes are called leaves

Definition: Suffix

A suffix of a DNA sequence x is any substring that includes the last nucleotide

Example: x = "ACGACT$"

termination character

Suffix	Starting position
ACGACT$	1
CGACT$	2
GACT$	3
ACT$	4
CT$	5
T$	6
$	7

A DNA sequence of length n has exactly n suffixes

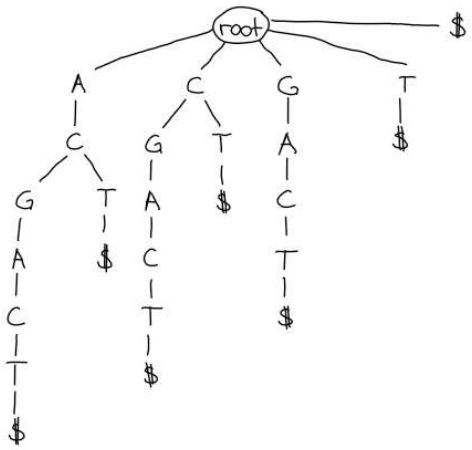
Definition: Suffix Tree → helpful in finding matches from reads to reference

A suffix tree is a tree with a one-to-one correspondence between the paths from the root to the leaves and the suffixes of a DNA sequence

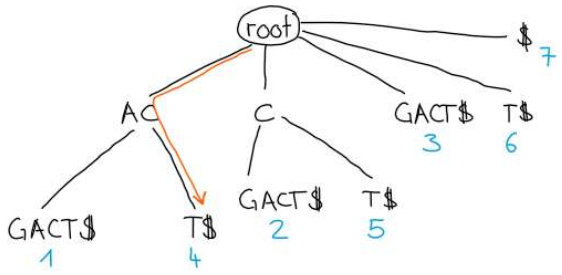
Example

Calculate the suffix tree for x = ACGACT$

Suffix	Starting position
ACGACT$	1
CGACT$	2
GACT$	3
ACT$	4
CT$	5
T$	6
$	7



The suffix tree can be written in a more compact form by concatenating nodes with a single edge.



Suffix trees can be used to efficiently find substrings.

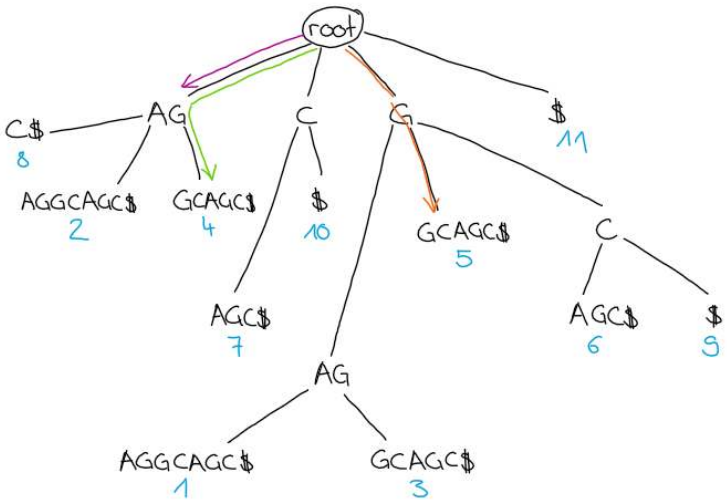
Example: Find "ACT" in ACGACT$

Position 4

To find ACT we traverse the suffix tree and compare each letter to each node in the tree

A more complex example. Find (a) GG (b) AGGC and (c) AG in GAGAGGCAGC$

Suffix	Position
GAGAGGCAGC$	1
AGAGGCAGC$	2
GAGGCAGC$	3
AGGCAGC$	4
GGCAGC$	5
GCAGC$	6
CAGC$	7
AGC$	8
GC$	9
C$	10
$	11



a) pos 5
b) pos 4
c) pos 2, 4, 8

How many comparisons did we make?

EXAM: Describe the algorithm for calculating the suffix array.

Definition: Suffix Array

A suffix array is the end of suffixes sorted in lexicographical order

$ is always sorted last

Example: What is the suffix array of ACGACT?

Suffix	Position
ACGACT$	1
CGACT$	2
GACT$	3
ACT$	4
CT$	5
T$	6
$	7

Suffix array	Position
ACGACT$	1
ACT$	4
CGACT$	2
CT$	5
GACT$	3
T$	6
$	7

Can we use the suffix array to efficiently find matrices of substrings?

Example

Reference: GAGAGGCAGC$

Suffixes	Position
GAGAGGCAGC$	1
AGAGGCAGC$	2
GAGGCAGC$	3
AGGCAGC$	4
GGCAGC$	5
GCAGC$	6
CAGC$	7
AGC$	8
GC$	9
C$	10
$	11

Reference: GAGAGGCAGC$

Suffix array	Sorted position
AGAGGCAGC$	2
AGC$	8
AGGCAGC$	4
CAGC$	7
C$	10
GAGAGGCAGC$	1
GAGGCAGC$	3
GCAGC$	6
GC$	9
GGCAGC$	5
$	11

Sorted BWT = sorted original sequence

Barrows-Wheeler Transform

1) Identify all suffixes

2) Sort the suffixes lexicographically to create a suffix array

3) Create the cylinder suffix array by extending the suffixes

4) Take the last column

Example

Calculate the BW-transform for ACGACT$

Sorted suffixes	Position	Cylinder suffix array	BWT
A C G A C T $	1	A C G A C T $	$
A C T $	4	A C T $ A C G	G
C G A C T $	2	C_1 G A C T $ A	A
C T $	5	C_2 T $ A C G A	A
G A C T $	3	G A C T $ A C_1	C
T $	6	T $ A C G A C_2	C
$	7	$ A C G A C T	T

Observe that

1) The BWT is the nucleotide that preceeds the sorted sequence

2) The i^{th} occurence of nucleotide a in the BWT is equal to the i^{th} occurence of the same nucleotide in the sorted sequence

Properties of the BWT

1) All suffixes and thus the original sequence can be reconstructed from the BWT

2) The BWT can be used to find positions of substrings (Ferragina-Manzini algorithm)

Finding substrings using the BWT

1) Start from the back of the substring

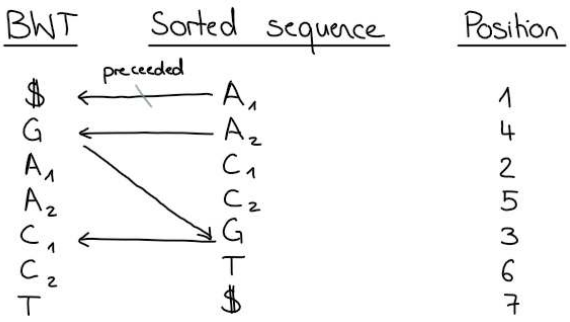
2) Find the nucleotide in the sorted substring

3) Find the corresponding nucleotide in the BWT
Note the order of the letters:

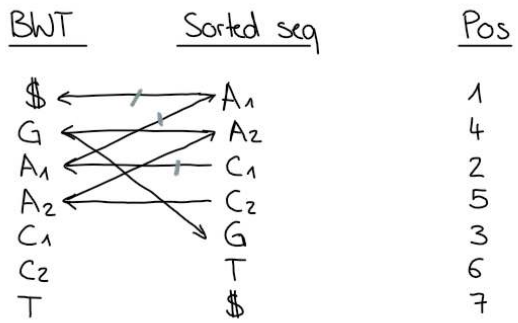
4) Find the corresponding nucleotide in the sorted sequence
(nucleotide at the same position)

5) Repeat until done

Example: Find CGA in ACGACT$



Example: Find GAC in ACGACT$

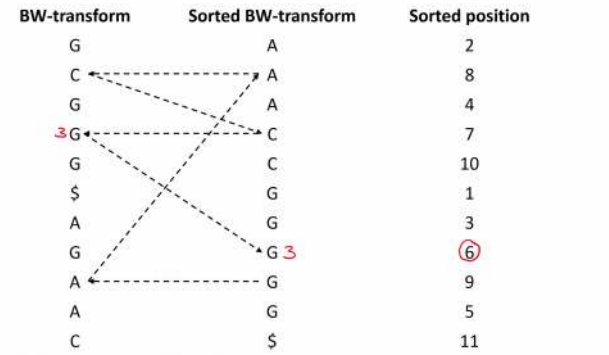


Example

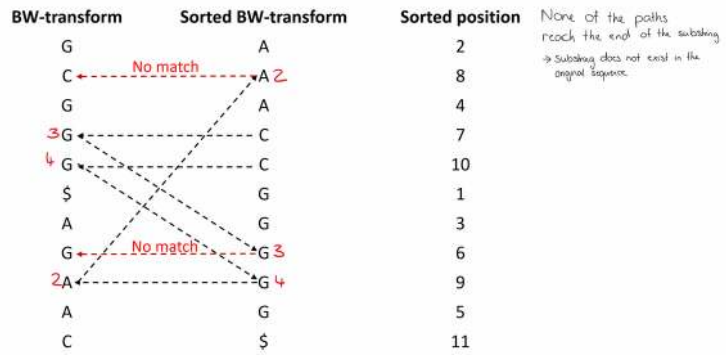
Reference: GAGAGGCAGC$

Suffix array	Sorted position	Cylinder suffix array	BW-transform
AGAGGCAGC$	2	AGAGGCAGC$G	G
AGC$	8	AGC$GAGAGGC	C
AGGCAGC$	4	AGGCAGC$GAG	G
CAGC$	7	CAGC$GAGAGG	G
C$	10	C$GAGAGGCAG	G
GAGAGGCAGC$	1	GAGAGGCAGC$	$
GAGGCAGC$	3	GAGGCAGC$GA	A
GCAGC$	6	GCAGC$GAGAG	G
GC$	9	GC$GAGAGGCA	A
GGCAGC$	5	GGCAGC$GAGA	A
$	11	$GAGAGGCAGC	C

Use the BW-transform to find GCAG in GAGAGGCAGC$.



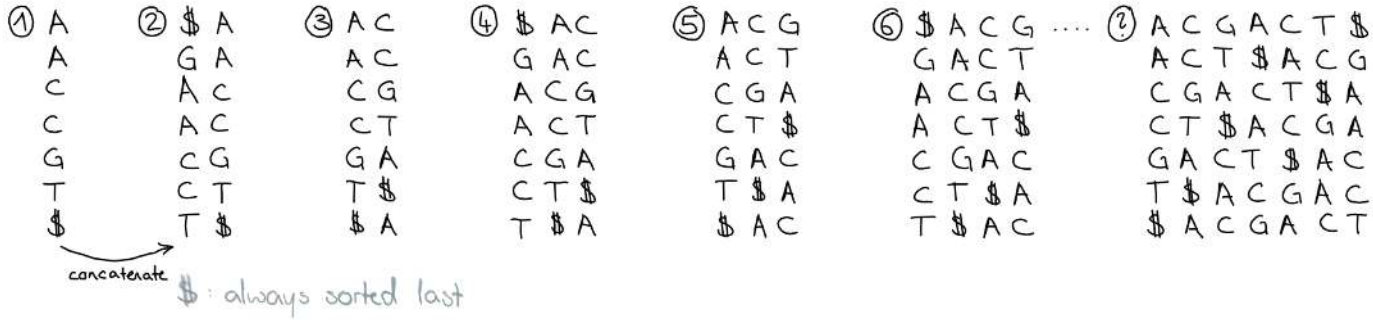
Use the BW-transform to find GAGC in GAGAGGCAGC$.



Reconstruction of the original sequence

1) Sort the BWT lexicographically
2) Concatenate the BWT to the sorted BWT
3) Sort the result
4) Go to 2 (We stop when we have as many characters as the BWT

Example: Reconstruct the sequence from the BWT "$GAACCT" (ACGACT$)



1	2	3	4	5	6
$	A	$A	AC	$AC	ACG
G	A	GA	AC	GAC	ACT
A	C	AC	CG	ACG	CGA
A	C	AC	CT	ACT	CT$
C	G	CG	GA	CGA	GAC
C	T	CT	T$	CT$	T$A
T	$	T$	$A	T$A	$AC

7	8	9	10	11
$ACG	ACGA	$ACGA	ACGAC	$ACGAC
GACT	ACT$	GACT$	ACT$A	GACT$A
ACGA	CGAC	ACGAC	CGACT	ACGACT
ACT$	CT$A	ACT$A	CT$AC	ACT$AC
CGAC	GACT	CGACT	GACT$	CGACT$
CT$A	T$AC	CT$AC	T$ACG	CT$ACG
T$AC	$ACG	T$ACG	$ACGA	T$ACGA

12	13	14
ACGACT	$ACGACT	ACGACT$
ACT$AC	GACT$AC	ACT$ACG
CGACT$	ACGACT$	CGACT$A
CT$ACG	ACT$ACG	CT$ACGA
GACT$A	CGACT$A	GACT$AC
T$ACGA	CT$ACGA	T$ACGAC
$ACGAC	T$ACGAC	$ACGACT

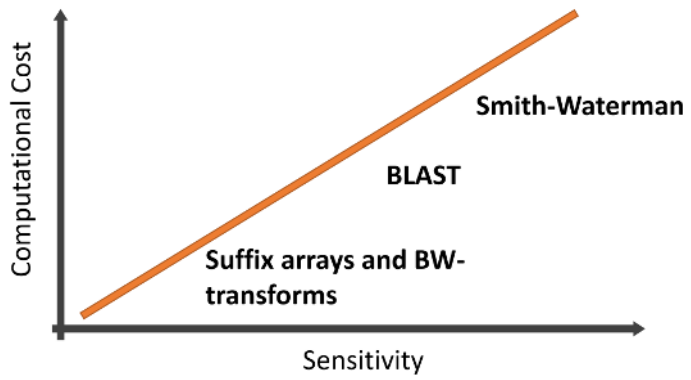
The Burrows-Wheeler Aligner (BWA)

Burrows-Wheeler transform -> oders the genome in a specific way that makes identification of subsequences highly efficient

- Based on seed-and-extend approach.
- Seeds are identified using suffix arrays and the Burrows-Wheeler transform.
- Seeds are extended into full alignments using more accurate algorithms.
- Optimized for large volumes of reads from next generation DNA sequencing.

Choosing a suitable aligner

- There is always a trade-off when it comes to sequence alignment
- The more correct and it should be able to identify mismatches, the higher the computational cost
- You do not need a really high sensitivity most of the times



EXAM: Describe the performance of Smith-Waterman, BLAST and suffix array-based methods. Which method is the most computational efficient? Why? Which method is most sensitive? Why?

Smith-Waterman

- Slow but highly sensitive.
- Software: FASTA and EMBOSS.

Seed and extend; BLAST

- Faster but have reduced sensitivity.
- Software: BLAST

Suffix-array and Burrow-Wheeler transform

- Very fast, low sensitivity. But often suitable for mapping sequence read to a reference.
- Software: BWA, Bowtie

SUMMARY

- Suffix trees are created from the suffixes of a DNA sequence. By traversing the tree we can find the position of any subsequence.
- The Burrows-Wheeler transform orders the genome in a specific way that makes identification of subsequences highly efficient
- Read alignment using suffix arrays and the Burrows-Wheeler transform are as fast as a suffix tree but does not require the entire tree to be created and stored in memory

EXAM: Describe the performance of Smith-Waterman, BLAST and suffix array-based methods. Which method is the most computational efficient? Why? Which method is most sensitive? Why?

The performance of an alignment method is typically connected to its sensitive. Methods with the highest performance often have the lowest sensitivity to the choice of aligner needs to be selected based on the specific application. Regarding Smith-Waterman, BLAST and suffix array-based methods, the latter are most efficient. Suffix arrays-based methods uses specific data-structures that are very efficient in identifying identical matches. Smith-Waterman is the least efficient but also the most sensitive. SW calculates the scores of all possible alignment which make it possible to find the optimal alignment. This requires, however, much more calculations than both BLAST and suffix array-based methods, that does not evaluate all possible solutions.

EXAM: What is the main disadvantage of using suffix trees to mapping of reads to the human genome?

suffix trees, which can be very large

EXAM: Explain the principle behind suffix arrays and the Burrows-Wheeler transform and why it computationally efficient to use when mapping reads from next generation sequencing to a reference genome. (2p)

Suffix arrays and the Burrows Wheeler transform are specific datastructures that enable us to find any subsequences in our reference genomes can be found. They work as an index, which can be seen as a table where we can quickly lookup any position of any subsequence. The use of suffix arrays + BWT is very efficient when we map reads to a reference genome. The computational complexity of the mapping itself is not dependent on the size of the reference. Also, the index itself, which takes time to compute, only needs to be derived once. In contrast to suffix trees, which can be very large, suffix arrays + BWT only requires the same storage size as the original genome sequence.

EXAM: What is the main disadvantage of using suffix trees to mapping reads to the human genome?

- Suffix trees are very large. The suffix tree for the entire human genome is so large that it can not be stored in the memory of a standard computers.
- If we want to include mismatches and gaps, the complexity increases fast!

EXAM: Describe the algorithm for calculating the Burrows-Wheeler transform.

1) Identify all suffixes
2) Sort the suffixes lexicographically to create a suffix array
3) Create the cylinder suffix array by extending the suffixes
4) Take the last column

EXAM: Describe the algorithm for finding subsequences using the suffix array and the Burrows-Wheeler transform.

1) Start from the back of the substring
2) Find the nucleotide in the sorted substring
3) Find the corresponding nucleotide in the BWT
   Note the order of the letter
4) Find the corresponding nucleotide in the sorted sequence
   (nucleotide at the same position)
5) Repeat until done

EXAM: Suffix arrays and the Burrows-Wheeler transform can be used for mapping reads against a reference. Explain why this approach makes mapping of reads more computationally efficient compared to the Smith-Waterman algorithm.

The BWT is of the same size as the reference and can thus be efficiently stored in the memory of a computer. The BWT only needs to be calculated once for a reference sequence and can then be used for matching all of the reads.

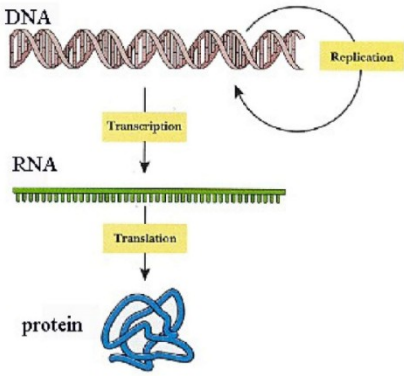
EXAM: What properties of the BWT makes it attractive for finding exact matching of sequencing reads against a reference?

The BWT enables fast lookup of exact matches. One the BWT is transformed, the complexity does not depend on the length of the reference. This makes it highly efficient when matching multiple reads to the same reference. In comparison to suffix trees, which also have these features, the BWT is of the same size as the reference and can thus be efficiently stored in the memory of a computer. Finally, the BWT only needs to be calculated once for a reference sequence and can then be used for matching all of the reads.

EXAM: Demonstrate how the original sequence can be reconstructed from the BWT.

This can be done by observing that the BTW is the predecessor of the sorted BTW. Thus, by iteratively concatenating the BTW to the sorted BTW the original sequence can be recreated (demonstration of the reconstruction is necessary for full points).

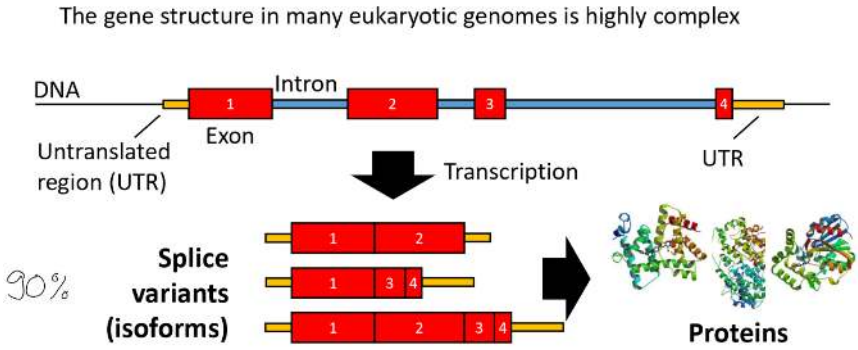
L7: Transcriptome sequencing (RNA-seq)



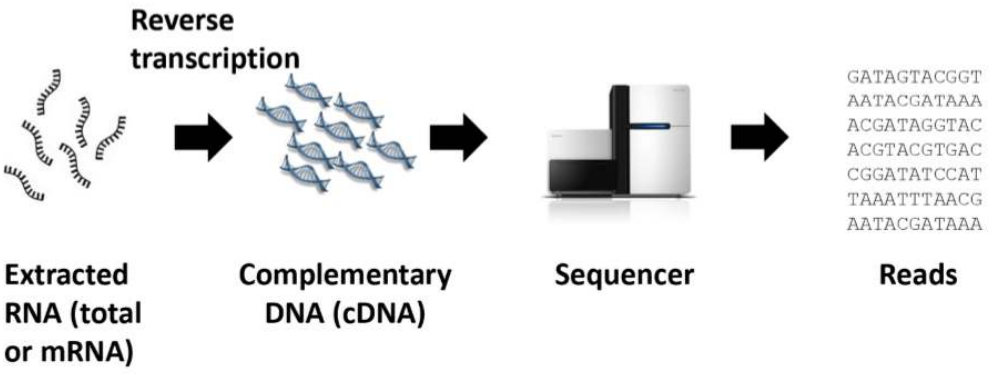
- Converted from information to activity
- RNA is used as a template to create proteins
- Protein: active building blocks of a cell
- Proteins have very different chemical properties
- Gene expression: process seen on the right

The strucutre of genes

- **DNA:** carrier of information
- **Exon:** are read into a protein; The sequence of DNA present in mature messenger RNA, some of which encodes the amino acids of a protein; any part of a gene that will form a part of the final mature RNA produced by that gene after introns have been removed by RNA splicing
- **Intron:** are transcripted, but are later removed in splicing; The sequence of DNA in between exons that is initially copied into RNA but is cut out of the final RNA transcript and therefore does not change the amino acid code.
- **UTR:** read in the beginning and end, but are never translated; A genomic DNA sequence that is not translated into an RNA sequence
- A gene can give rise to a number of different variants



RNA-seq process



- Reverse Transcription: take RNA and read it back into DNA
- Reads belong to a random piece of RNA in the original sample

Data analysis of RNA-seq data

Three main steps

1. Quantification of the gene expression

From reads to a (semi)quantitative measurements of gene expression

Mapping genes to reference and count which gene they mapped to; count = number of read matches

The higher number of reads we could count for each gene the more RNA copies we had

2. Normalization

Correction of systematic errors within and between samples

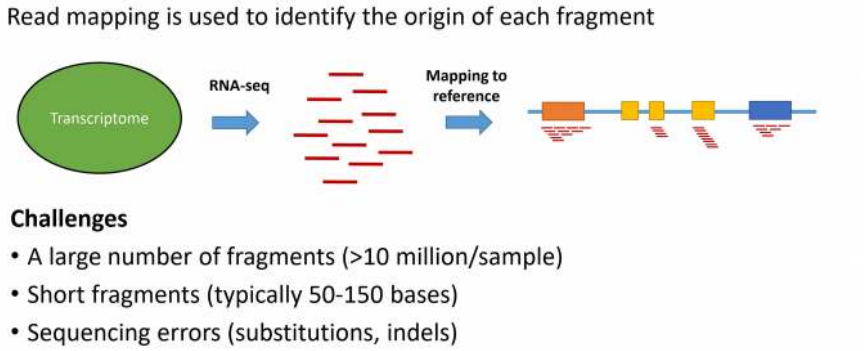
3. Identification of differentially expressed genes (DEGs)

Find genes with a significant difference in gene expression

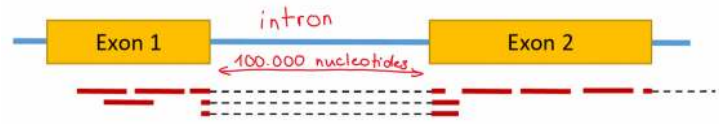
EXAM: Explain the difference between quantifying the expression of genes, exons and isoforms.

1. Gene Quantification

Quantification of gene expression is based on mapping of the sequenced fragments against a reference. The reference typically consists of a genome or a de novo assembled transcriptome.



Common references

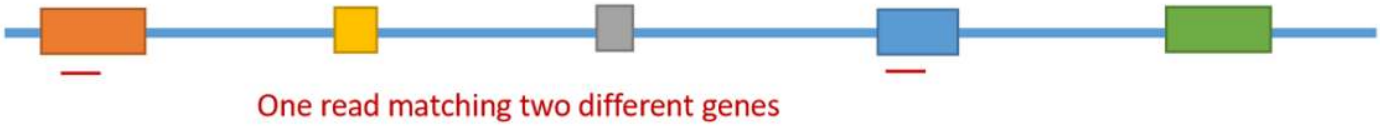
- **Genome**
  - Requires high quality genome assembly
  - The mapper needs to be able to handle splicing (splice-aware)
  - 

- **De novo assembled transcriptome**
  - Construction of the transcripts from the sequence data
  - Hard to identify splicing. Representing mainly mature spliced transcripts
  - If we do not have a genome: looking at reads that overlap and merge them together to create a reference
- **Gene quantification**
  - Count the total number of fragments for each gene
- **Exon quantification**
  - Counting the number of fragments for each exon
  - Splicing can be identified from fragments mapping junctions
- **Isoform quantification**
  - Estimates the abundance for each isoform.
  - Requires mathematical models and complementary data such as a database of known splice variants

Multiple matches

When you take the reads and start to match them back to the genome, you will realize that many of the reads from the RNA will have multiple matches -> two genes are really similar

- Multiple matches is common and caused by
  - Genes with regions that are similar within the genome (e.g. recent paralogs)
  - Repetitive regions in transcribed areas
  - Errors in the reads caused by problems in the sequencing



One read matching two different genes

- Approaches to handle reads with multiple matches includes
  - **None** – the reads are excluded from the analysis
  - **All** – assign the reads to all its matching all regions
  - **Random** – assign the read randomly to one of its machting regions -> most common
  - **Statistical predictive models** (try to guess based on the neighbors of the genome)

EXAM: Normalization is an important step in the analysis of transcriptomic data. Name two different normalization methods and describe their advantages and disadvantages.

2. Normalization

Normalization within and between samples is necessary for comparability. Common methods includes RPKM, upper quartile and TMM.

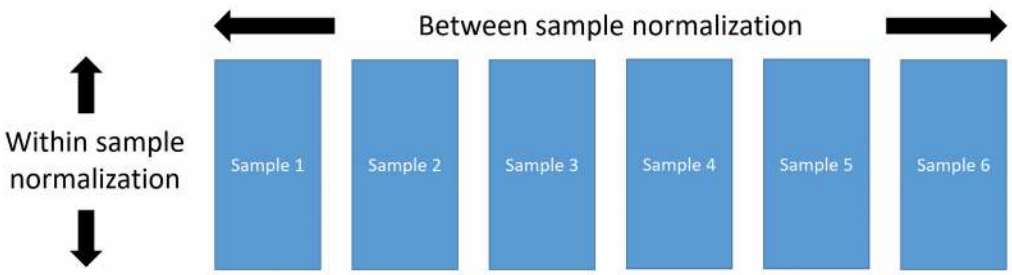
Data from different gene sand samples are not directly comparable due to systematic errors
-> Normalize data to make it comparable between samples and genes -> identify systematic errors and remove them (error that affects multiple genes or multiple samples)

Definition

Sequencing depth = The number of fragments sequenced in each sample differs
Gene lengths = Longer genes are represented by more RNA bases and will be represented by more fragments
Sample quality = Errors affecting large part of a sample is introduced by quality issues in the sample preparation and sequencing.
Other systematic errors

- Sequencing errors
- Fragment distribution over the gene
- G/C-bias

Normalization aims to reduce the noise by identifying and removing systematic effects.



Within sample normalization: compare genes	**Between sample normalization:** compare samples

Counts

Samples

Gene	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
ENSG00000000419	11	16	10	19	8	24	14	19
ENSG00000000457	28	22	17	21	18	15	21	13
ENSG00000001167	112	107	89	95	75	87	80	73
ENSG00000002016	1	1	1	1	0	0	0	0
ENSG00000002834	54	83	59	76	104	88	61	73
ENSG00000002919	1055	865	481	738	402	533	351	404
ENSG00000004700	79	162	125	124	45	118	99	146
ENSG00000004777	4	1	1	5	3	1	5	0
ENSG00000004779	256	327	227	234	278	364	214	276
ENSG00000004897	126	151	111	163	96	97	110	156
ENSG00000005175	49	59	57	37	26	47	52	53
ENSG00000005187	9	18	12	47	23	22	18	42
ENSG00000005339	895	404	422	985	393	188	613	635
ENSG00000005844	82	110	93	66	149	130	207	330
ENSG00000006007	180	173	147	196	132	186	148	208

Genes

Counts

Samples

Gene	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
ENSG00000000419	11	16	10	19	8	24	14	19
ENSG00000000457	28	22	17	21	18	15	21	13
ENSG00000001167	112	107	89	95	75	87	80	73
ENSG00000002016	1	1	1	1	0	0	0	0
ENSG00000002834	54	83	59	76	104	88	61	73
ENSG00000002919	1055	865	481	738	402	533	351	404
ENSG00000004700	79	162	125	124	45	118	99	146
ENSG00000004777	4	1	1	5	3	1	5	0
ENSG00000004779	256	327	227	234	278	364	214	276
ENSG00000004897	126	151	111	163	96	97	110	156
ENSG00000005175	49	59	57	37	26	47	52	53
ENSG00000005187	9	18	12	47	23	22	18	42
ENSG00000005339	895	404	422	985	393	188	613	635
ENSG00000005844	82	110	93	66	149	130	207	330
ENSG00000006007	180	173	147	196	132	186	148	208

Genes

CPM – Counts per million mapped reads

Estimates the relative abundance of a gene in relation to the total sequencing depth (total number of successfully mapped reads).

Reads for gene g

$$\mathrm{CPM}_g = \frac{R_g}{M/10^6}$$

Total number of mapped fragments

Between sample normalization

M: summing the reads over all genes in this particular sample; total number of mapped

RPKM – Reads per kilobase per million

Estimates the relative abundance for each gene in relation to its length and total sequencing depth (total number of successfully mapped reads).

Reads for gene g

$$\mathrm{RPKM}_g = \frac{R_g}{(L_g/10^3)(M/10^6)}$$

Length of gene g

Total number of mapped fragments

Extension of CPM

Normalization between and within samples

The longer RNA frequence you have the more likely it is to get reads from it

Remarks

Normalization based on the total number of counts may introduce systematic errors

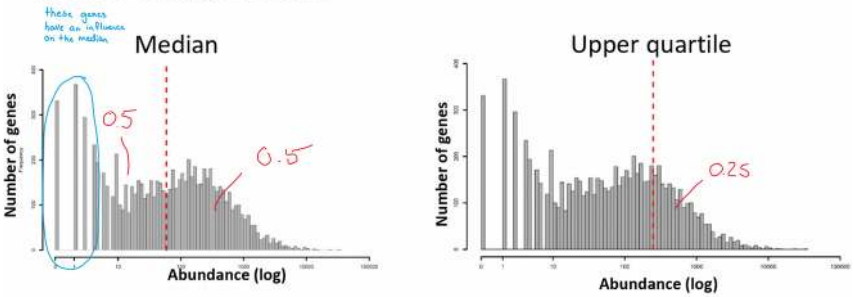
- The total RNA pool is often dominated by few highly expressed genes (e.g. actin). -> the sum will be dominated by that
- Variation in expression of the high abundant genes can therefore have big impact on the normalization.
    - If you have a high variability in these highly expressed genes, it means that the total number of reads will also have a high variability

- -> we divide each gene by sth that is highly variable
- -> decreases our ability to identify the single
- -> we introduce more noise
- -> you add an systematic effect

- Even more important, systematic changes in their expression between the experimental conditions can lead to incorrect results.

Robust alternatives

There are several robust alternative that can be used instead of the total number of mapped reads:



It is also possible to normalize in relation to a set of reference genes.

- Robust to changes in a proprortion of the gene -> Alternatives of the total sum
- **Median:** if you take all the genes and sort them from the lowest to the highest and take the middle one; 50% of the data, 50% of the data; middle point that separates the distribution in two equal parts
  - -> no single or single set of genes influences this median a lot
- **Upper quartile:** median of the upper part from the median
  - -> you move away from the low abundant genes
  - -> you make an estimate that is slightly more robust

RNA-seq data is semiquantitative

- **Up-regulated:** the cell starts to produce more from this particular gene in this particular sample -> change in abundance (=The total number of individuals of a species or type present in a given area, in a given ecosystem or within a particular habitat)
- If one gene increases, the others need to decrease -> we can not measure the total abundance of RNA molecules, we only measure the relative abundance between genes
- We only want to identify the primary changes (Gene1; up-regulated) and not the secondary changes (other genes) due to the semiquantitativeness of the data
- Normalization: remove secondary effects, but keep primary effects

Trimmed mean of M-values

Let Y_{ij} be the number of counts for gene i and sample j and N_j be the total number of reads for sample j. Select a sample (sample r) as the reference.

For a sample j define

$$M_i = \log_2 \frac{Y_{ij}}{N_j} / \frac{Y_{ir}}{N_r} = \log_2 \frac{Y_{ij}}{N_j} - \log_2 \frac{Y_{ir}}{N_r}$$

$$A_i = \log_2 \sqrt{\frac{Y_{ij}}{N_j} \times \frac{Y_{ir}}{N_r}} = \frac{1}{2}\left(\log_2 \frac{Y_{ij}}{N_j} + \log_2 \frac{Y_{ir}}{N_r}\right)$$

TMM assumes that only a small proportion of the genes are differentially expressed. This means that M_i is close to zero the vast majority of the genes.

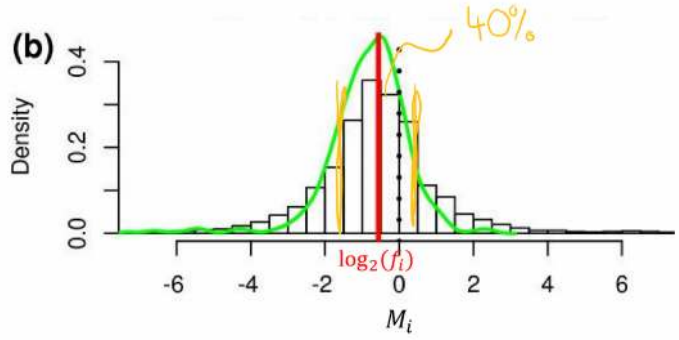
An adjustment factor for sample j, f_j is calculated as

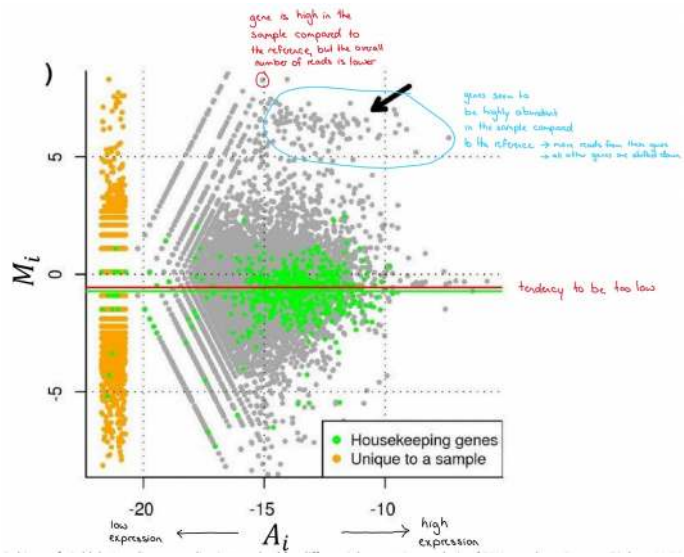
$$\log_2(f_j)=\text{TrimmedMean}(M_i)$$

In the trimmed mean, the largest 30% and the lowest 30% of the M_i values are not included. This mean that only the middle 40% of the values are included in the mean.

A normalization factor is then calculated as $\widetilde{N}_j = f_j \times N_j$. (Total number of reads for sample j)

- Try to correct the data to remove the secondary effect while keeping the primary effect
  - Small differences effect a larger number of genes is not of primary interest while big effects effecting a smaller poriton of the genes is the signal that we are after
- Take a sample, normalize the sample against a reference ->Compare every single sample against the reference
- **M**: how each genes differ between sample and reference
  - M value is positive if the differences of proportion is higher in the sample compared to the reference or negative in case the proportion is higher in the reference compared to the sample
- **A**: measures the sum; how many reads we have
  - A value is high if we have a high proportion in the sample and low if we have overall a low proportion in the sample



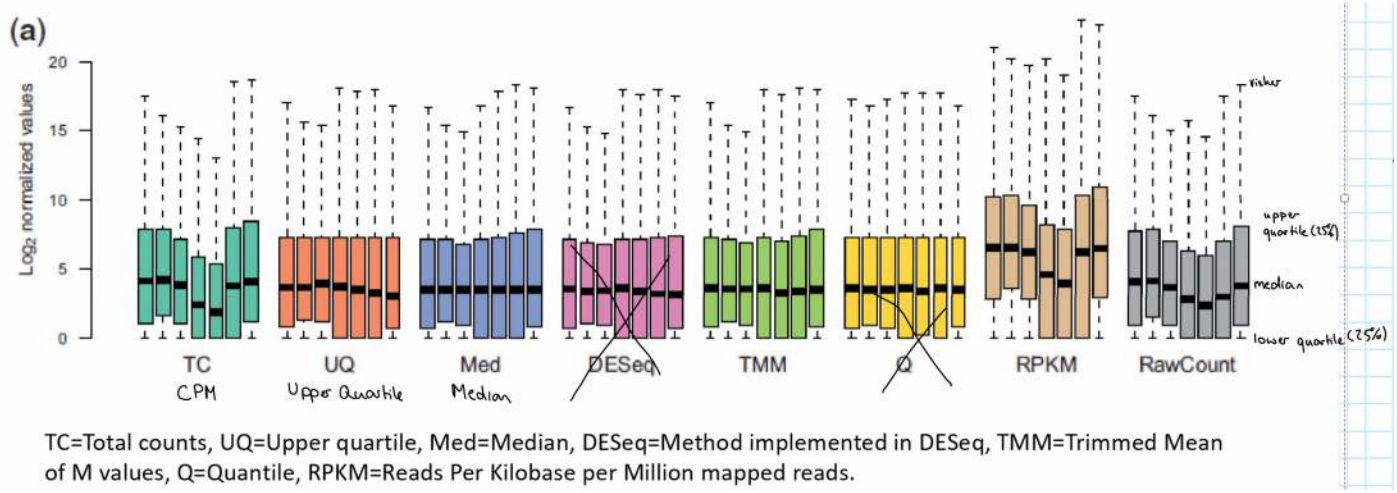


- M:
  - negative -> lower in the sample compared to reference
  - Positive -> higher in the sample compared to reference
- High expression = high number of reads
- Housekeeping genes (green): extremely stable, mean is quite close to the estimated red line, not affected by normalization -> the red line is right, the majority of the genes that do not change seem to be at the right place
- We make the explicit assumption that only a small portions of the genes change and we correct the samples based on that assumption to make that the vast majority of the genes has an average value of zero or in general no change in average

Normalization – general remarks

Remark 1: Within sample normalization is often not necessary! Systematic effects that are similar between samples may be canceled out when estimating the difference in expression.

Remark 2: Many methods for identification of differentially expressed genes works with counts. Some normalization methods modifies data in way that the counts are lost. It is therefore important to use a normalization method that is appropriate for the downstream statistical analysis.



TC=Total counts, UQ=Upper quartile, Med=Median, DESeq=Method implemented in DESeq, TMM=Trimmed Mean of M values, Q=Quantile, RPKM=Reads Per Kilobase per Million mapped reads.

Shows after normalization

- Is there are difference before and after normalization?
- Variability between samples seems to decrease
- Difference between CPM and RPKM is not so big (almost identical)
- **Median:** we have normalized with respect to the median -> median is the same across the samples
- **UQ:** upper quartile has been fixed
- Having both is not possible
- **TMM** is in between both -> TMM is mostly preferred
- Purpose of normalization is not to make it identical -> remove variability instead
- Improper normalization can reduce performance

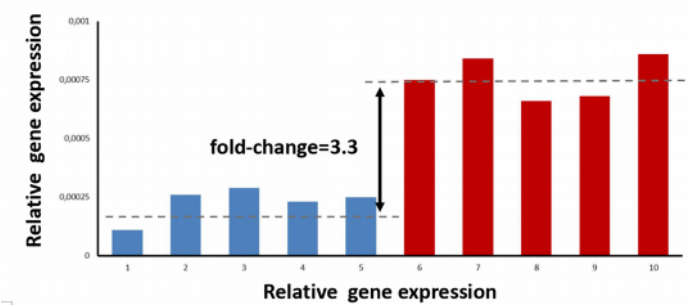
3. Identification of differentially expressed genes

Aim: Identify genes with altered gene expression level between experimental conditions.
Strategy: Examine and analyze each gene separately

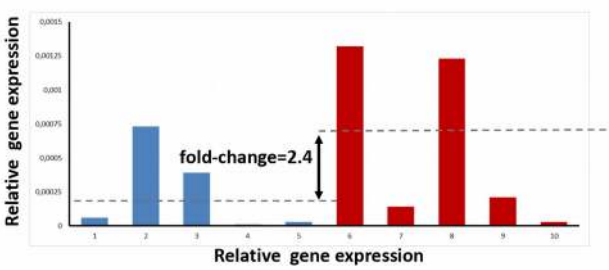
Advantages	Disadvantages
• Straight-forward and limits complexity of the analysis • Provides gene-level information	• Many genes implies many comparisons • Information for each gene is limited (e.g. few samples)

When can we consider the expression to be altered?

Example 1:



Example 2:



The difference needs to be interpreted with respect to the variability!

- Look how many times the average value has changed
- Two different conditions
- Expression in red is generally higher than in genes in blue

- Not so interested in high variability, the fold change is still quite high
- BUT: Increase is not consistent between the samples
- Only looking at genes that have a high fold change will mean that we mix these two scenarios, but we are only interested in the first one, because the change is not consistent in the second one

Technical variability

- Sample quality
- Sample preparation, e.g. RNA extraction and cDNA preparation
- Sequencing errors and quality, errors from the mapping

Biological variability

- Genotype and other effects on individual level
- Effects introduced by sampling and/or experimental setup
- Transcription is a stochastic process with spatial and temporal differences

Typically biological variability is higher than technical variability -> noisy data

Challenges in the statistical analysis

1. Data consists of count (discrete)
2. Large number of genes (> 10.000)
3. Typically few replicates (replicates (samples) << genes)
4. Large gene-specific variability

Statistical approach includes

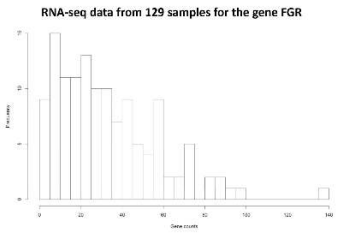
- A model describing the data and its variance structure
- A test for identification of differentially expressed genes

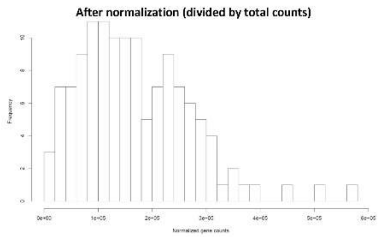
Three main approaches

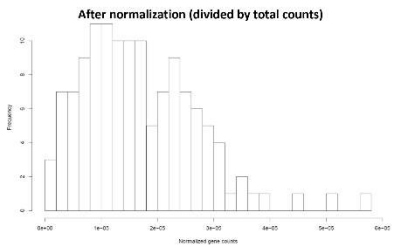
1. Methods based on normal assumptions
2. Methods based on non-parametric methods
3. Methods based on count distributions

1. Methods based on normal assumptions -> most used

- Use a variance-stabilizing transformation to make the distribution of the data a close to a normal distribution as possible. (Variance-stabilizing: disentangels the variance from the expected value (are always connected normally); so it does not depend on the mean value)
- Common transformation are the square-root and log transformations.
- It is practically impossible to select a transformation that is perfect for all genes!
- Normally distributed data enabels the use of more flexible statistical tools such as linear models (the topic of the next lecture!)







➔ Becomes more symmetric
➔ Fittet best normal distribution (approx.); Normal approx are sufficiently good
➔ Red curve: normal distribution
➔ You can use transformation to get closer to a normal distribution

2. Methods based on non-parametric methods

- Makes no explicit distributional assumptions
- Common non-parametric methods include the Wilcoxon-MannWhitney test (Wilcoxon rank sum), the Kruskal-Wallis test and Fisher's exact test.
- Can be sensitive to ties, i.e. values that are identical between samples. This can occur frequently for genes with low abundance and thus low counts.

3. Methods based on count distributions

- Assumes that the data follows a specific count distribution. Common distributions include
  - Poisson distribution
  - Binomial distribution
  - Negative binomial distribution
- Often overdispersed count-models are required. These models add extra variability than present in the standard models.

Statistical fact 1: Count data has a mean-variance dependence
Statistical fact 2: The variance of count data can be divided into two parts

$$\text{Total variance} = \text{Sampling variance} + \text{Overdispersion}$$

Random selection of DNA fragments

Dominated by biological variation

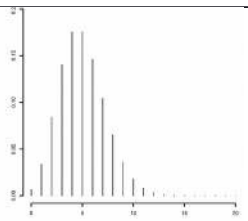
The Poisson model

- A basic model for describing counts
- The variability depends on the expression level

X_{gi} = counts of gene g in sample i

$\text{Exp}[X_{gi}] = \mu_g$

$\text{Var}[X_{gi}] = \mu_g$

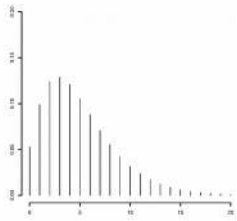


The Poisson model assumes that two genes with the same expression level have the exact same variability!

Basic form of count model
If we increase the gene abundance, we also have to increase the variability
This argument is not true -> simple to work with, but it is not enough, does not cover all the variability we see

The negative binomial model

- Provides more variability than the Poisson model
- Variance divided into two parts
  1. Poisson noise from random sampling of DNA fragments
  2. Gene-specific technical and biological noise - *overdispersion*

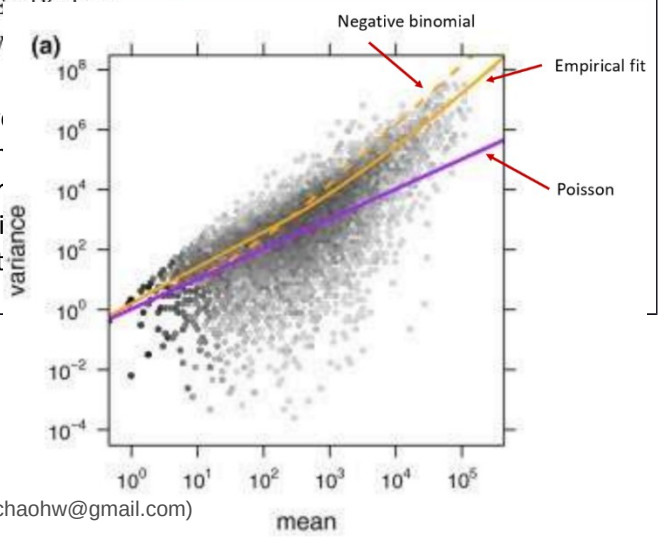


X_{gi} = counts of gene g in sample i

E

V

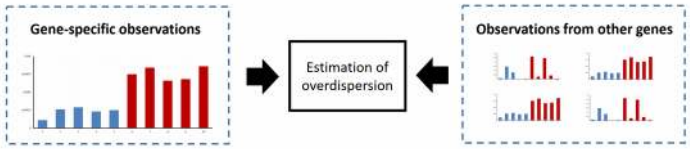
Ov
is h
var
Phi
ext



- Each point is a gene
- There is a trend
- Three lines try to capture this trend
- The negative bionomial is a better model than the poisson for this data

Estimation the gene-specific overdispersion is important. Sharing of variability between genes significantly improves the performance in datasets with few samples.

- Estimation of the gene-specific overdispersion is hard when few samples are available.
- Sharing of information between genes generates more robust estimates.



- This is called a shrinkage model. Often implemented by using Bayesian statistics.

Differentially expression

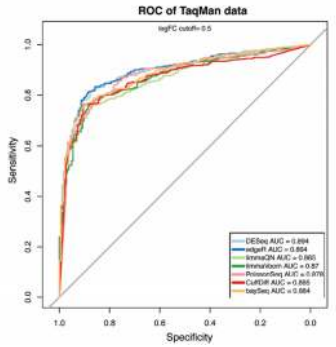
Identification of differentially expressed genes is done by hypothesis testing. Statistical models are used to describe the data and its variability.

- The differentially expression is assessed fro each gene using a hypothesis test.
- The p-value provide information about rejecting H0.
- Multiple testing: many tests – many p-values. More on this later!

For each gene g,
H_0: Gene g is not differentially expressed
H_A: Gene g is differentially expressed

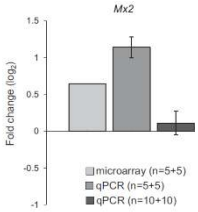
Methods for RNA-seq data

Method	Model	Sharing of overdisp between genes?	Multifactorial designs	Reference
Fisher's exact test	Hypergeometric	No	No	Bullard et al 2010
PoissonSeq	Overdisp. poisson	No	No	Li et al 2012
baySeq	Negative binomial	Yes, hierarchical Bayesian	No	Hardcastle et al 2010
edgeR	Negative binomial	Yes	Yes	Robinson et al 2010, McCarthy et al 2012
DESeq2	Negative binomial	Yes, empirical Bayes	Yes	Love et al 2014
voom	Approx. normal	Yes, empirical Bayes	Yes	Law et al 2014



Interpretation of gene lists

- Gene lists should be interpreted with care
  - P-values are often biased due to incorrect model assumptions
- False positives are common!
- PCR verification of selected genes is highly recommended. Preferably on an independent and larger set of samples.



Experimental Design

Biological replication is essential. More biological replicates means higher power. Pooling should be avoided!

EXAM: In an RNA-seq experiments, the expression of human genes are investigated in liver tissue from five patients. The abundance of each gene is quantified by mapping the RNA-seq reads to a reference and counting the number matches for each gene. The result for two different genes is shown in the table below together with the total number of reads for each patient. The lengths of the genes are show in the last column.

Gene	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Gene length
Gene 1	1000	1500	250	150	100	1000
Gene 2	5000	8000	1200	12000	8000	2000
Total	5×10^6	1×10^7	1×10^6	3×10^6	2×10^6	

(a) Describe the normalization method RPKM (reads per kilobase per million). In what way does it reduce the variability in the data? (2p)

(b) Use RPKM to calculate normalized values for the genes in the table above. (2p)

(c) Assume that samples 1,2 and 3 are from healthy individuals and samples 4 and 5 are from sick individuals. Calculate the average fold-change between these two groups. Interpret the results. Are the genes up or down-regulated? (2p)

(a) SOLUTION:

RPKM – Reads per kilobase per million

Estimates the relative abundance for each gene in relation to its length and total sequencing depth (total number of successfully mapped reads).

$$\mathrm{RPKM}_g = \frac{R_g}{(L_g/10^3)(M/10^6)}$$

Reads for gene g

Length of gene g

Total number of mapped fragments

(b) SOLUTION:
Gene 1:200, 150, 250, 50, 50
Gene 2:500, 400, 600, 2000, 2000

(c) SOLUTION:
Gene 1: Fold-change: 200/50 = 4. Down-regulated in sick individuals.
Gene 2: Fold-change: 500/2000 = 1/4. Up-regulated in sick individuals.

EXAM: Briefly describe how reads from RNA-seq can be used to estimate gene expression (i.e. RNA abundance)./ Each gene is typically test for differential expression, for example between two groups of patients. What does this mean and how is it done?

RNA-seq works by randomly selecting RNA-fragments, which are reverse-transcribed and sequenced. The resulting reads correspond thus to random positions of the RNA expressed in the sample. For a particular gene, the number of reads is a measure of its abundance. This is derived by mapping the reads to a reference and counting the number of reads matching each read. Gene expression is estimated by mapping reads to a reference, typically the full genome of the investigated the organism. By counting the number of matches to each gene, we can estimate its expression.

EXAM: Explain the difference between quantifying the expression of genes, exons and isoforms.

- In RNA-seq, the quantification can be done at different resolution. When quantifying the overall expression of a gene, we count all reads matching any part of the gene.
- It is also possible to quantify individual exomes. In this case, we calculate the number of reads for each exome. This enables identification of specific splicing patterns.
- It is finally possible to quantify the expression of specific isoforms. An isoform is a protein variant created by combining specific exons. By using a database of known isoforms, it is possible to base on information where

the reads are mapping, predict abundance of the individual isoforms. This process is however complex and not possible to do accurately in species that lack information about the isoforms

In short:
Quantification of genes means that we count all reads matching to the gene.
Quantification of exons means that we count all reads matching to the individual exons.
Quantification of isoforms means that we try to estimate the abundance of each possible gene product (isoform).

EXAM: Data from RNA-seq is often said to be semiquantitative. Explain what this means and its implications, i.e. what can we not measure using RNA-seq because of that the data is semiquantitative?

RNA-seq data is semiquantitative, which means that even if we can quantify gene abundances, we can only do this in relation to the abundances of other genes. Information about the absolute abundance, i.e. how many RNA molecule present in the sample, is not measured.

This has important consequences in how we interpret RNA-seq data. In particular, an increase of gene expression identified in RNA-seq can either be a result on a large number of RNA molecules for that specific gene or, alternatively, a decrease of RNA molecules of one of several other genes. Similarly, a decrease in measured gene expression could corresponds to a lower number of RNA molecules for the specific gene or an increase in RNA-molecules in one or several other genes. This means that we cannot relate the differences that we see to absolute changes in RNA-molecules, which makes RNA-seq data hard to interpret.

In short:
The data from RNA-seq is semiquantitative in the sense that it is quantitative but only in relation to other genes. This means that we do not have information about the total number of RNA molecules in the sample. Thus, RNA-seq data cannot distinguish between an increase of expression in one gene and the decrease in expression of all other genes.

EXAM: In transcriptomics, high-throughput sequencing is used to quantify the abundance of mRNA from individual genes. What sequencing platforms are suitable for transcriptomics? Name at least two platforms and describe their advantages and disadvantages when used in transcriptomics.

Two suitable platforms are Illumina and Oxford Nanopore.

- The advantages of the Illumina sequencing platforms are the low price of the data, the low error rate and the large number of reads that can be generate within short time.
- The disadvantage is the short read length.
- The advantage of the Oxford Nanopore (minION) is the long reads, which enable us to sequence entire transcripts. This facilitates the analysis of splicing.
- The disadvantage is the relatively high price and the high error rate.

EXAM: What does sequencing depth mean in this context? Why is a sufficiently high sequencing depth important? Can the sequencing depth become too high?

Higher sequencing depth means that more reads match the genes and thus more information about their abundance. This will increase our statistical power to detect differentially abundant genes. Also, a higher sequencing depth enable us to identify low-expressed genes. If the sequencing depth is too low, many genes may be come undetectable and thus overlooked. The sequencing depth can, in principle, never be too high. More data will, however, cost more and require more effort when analyzed.

EXAM: Normalization is the process used to remove systematic bias. Describe why this is important. Give at least two examples of what could happen if transcriptome data is not properly normalized.

Removal of systematic bias is important to make samples and genes comparable. This reduces the overall variability of the data, which makes it easier to identify the true biological differences. If the transcriptomics data is not

normalized many things can happen. If we compare genes without a proper within-sample normalization, differences in gene length may affect our results. Longer genes will generate more reads and these genes will thus seem to be more expressed.

If we compare the expression of a gene between samples without any normalization, the variability will likely be larger and our statistical power to detect whether it is differentially expressed will be lower. This increases the risk that we miss important differentially expressed genes. Finally, gene expression data is semi-quantitative so the up/down-regulation of one gene will affect all other genes. If these secondary effects are not removed, they may end up as false positives, thus making the end-result hard to interpret.

EXAM: RNA-seq is sometimes used to find mutations. The advantage is that this enables both analysis of gene expression and identification of various DNA alterations from the same data.

(a) How would you design an RNA-seq experiments which aims both to analyze gene expression and to identify mutations? For example, what sequencing platforms would you recommend and what aspects needs to be taken into account when deciding the sequencing depth? (2p)

Sequencing can be done similar to a standard RNA-seq experiment. A commonly used platform is Illumina. Note however that the short read length of Illumina will make it harder to identify larger alterations. Also, the sequencing depth will depend on the mRNA abundance of the genes. The coverage, and thus the ability to identify mutations, will vary between genes. If we want to be able to find mutations in the majority of the genes, the sequencing depth must be enough to also result in a high coverage of the low abundant genes.

(b) Describe briefly the analysis steps that you think are necessary to identify mutations from RNA-seq data. (3p)

The analysis can be done as follows. First the data is pre-processed to remove bad reads. Next, the reads are aligned to a reference. If we are using the human genome, a fast index-based aligner should be satisfactory. Note that the aligner should be spice-aware in order to align reads over exon junctions. Note that we will have low coverage for gene with low expression and quality assessment and removal of artifacts is therefore important in order to avoid false positives. Next, mutations can be called using similar approaches as used in whole genome sequencing. The identified mutations should be filtered to remove artifacts caused by e.g. duplicates.

(c) Discuss what forms of mutations that you expect to find using this approach. Are there any mutations that may be missed in the RNA-seq data that could have been identified from whole genome sequencing data? (3p)

This approach will be able to identify SNPs and short indels present in transcribed regions. This includes SNPs that causes a change in the amino acid sequence of the corresponding genes and indels that causes frame-shifts (both have the potential to alter protein function). It is also possible to identify larger structural variations that results in fusion genes. Mutations outside coding regions will however be missed. Also mutations in regions that are transcribed but spliced away (e.g. mutations in intron) will most likely not be identified. Mutations is genes that are low abundant will be harder to find.

EXAM: In transcriptomic analysis using RNA-seq, random mRNA fragments are extracted and sequenced.

(a) Describe the process of gene quantification. How do we go from reads to gene counts?

Genes in RNA-seq are quantified by counting reads. First reads are aligned to a reference genome where all genes are annotated. After that, we examine the position of the alignments and calculate the number of reads matching every gene. Note that the quantification can happen both on gene level (counting all exones), exone level but also for each possible isoform (typically much harder).

(b) What is the difference between normalization techniques counts per million (CPM) and trimmed mean of M-values (TMM)?

CPM normalization aims to remove the dependence on sequencing depth by dividing each gene count with the total count of the sample. In contrast to CPM, TMM aims to remove secondary effects caused by the semiquantitativeness

of RNA-seq data. In particular, TMM uses a reference sample and calculates the difference between each sample and the reference for all genes (by calculating M-values defined as the logarithm of the ratio of counts). TMM assumes that the large majority of all genes are not differentially expressed and thus, the M-value for most genes should be close to zero. A trimmed mean value is used to calculate the mean of the M-values and the data is then adjusted so that this mean is equal to zero.

EXAM: The normalization methods 'counts per million reads' (CPM) and 'reads per kilobase per million' (RPKM) are based on the total counts in a sample. State at least one reason why this can introduce errors into the data.

Normalization based on the total number of counts may introduce systematic errors

- The total RNA pool is often dominated by few highly expressed genes (e.g. actin). -> the sum will be dominated by that
- Variation in expression of the high abundant genes can therefore have big impact on the normalization.
- Even more important, systematic changes in their expression between the experimental conditions can lead to incorrect results.

EXAM: Describe the difference between 'within sample normalization' and 'between sample normalization'. What kind of systematic effects do they remove? (3p)

- **Within sample normalization:** compare genes
- **Between sample normalization:** compare samples

EXAM: Name one advantage of using median normalization instead of RPKM. (1p)

Robust to changes in a proprortion of the gene -> Alternatives of the total sum

EXAM: Describe the different steps of the normalization method 'trimmed mean of M-values' (TMM).

1. Select a sample as the reference
2. Define M: how each gene differs between sample and reference, A: sum how many reads we have
3. An adjustment factor for sample j is calculated
4. In the trimmed mean the largest 30% and lowest 30% of the M values are not included; this means that only the middle 40% of the values are included in the mean
5. A normalization factor is calculated

I

EXAM: Differentially expressed genes, i.e. genes that differ in expression between group of samples, are identified using statistical methods. Explain one method that is based on a normal assumption and one method that takes advantage of the discrete structure of RNA-seq data. Discuss the strengths and weaknesses of each method. (4p)

1. Methods based on normal assumptions
   - Can use the flexibility of models based on the normal distribution
   - Will always results in approximation and thus sub-optimal performance
   - Hard to find a transformation that is suitable for all genes
2. Methods based on non-parametric methods
   - Robust against model assumptions
   - Requires a larger number of samples
   - Can be highly sensitive against ties, which are common for gene with few counts
3. Methods based on count distributions
   - The correct statistical nature of the data is described
   - Complicated models that requires assumption about the mean-variance relationship
   - Computationally more expensive (but usually manageable)

L8: Linear models

Linear models: describe certain measured data given a set of conditions/parameters that we have set to meet those measurements

Linear models

Let Y_j be the outcome of an experiment

$j = 1, \dots, m$ m is the number of samples (or experiments)

$x_{1j}, \dots, x_{pj}$ a set of p covariates (variables) that influences Y_j

The linear model assumes that $Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_p x_{pj} + \varepsilon_j$ (error term)

$\beta_0, \beta_1, \beta_p$ are parameters (coefficients) → how x_1 influences y

Y_j is called the dependent variable

$x_{1j}, \dots, x_{pj}$ are called independent variables or covariates

ε_j is the error term. We assume that $\varepsilon_j \sim N(0, \sigma^2)$ → Variance does not change → variability does not change

Normal distributed, variability does not change based on the repetitions

σ^2 is the variance

$\varepsilon_1, \dots, \varepsilon_m$ are independent

Special case 1: If $p = 1$ and x_{1j} is continuous, the linear model is reduced to

$Y_j = \beta_0 + \beta_1 x_{1j} + \varepsilon_j$

X_1j takes values along the x-axis

Simple linear regression

Special case 2: Let $p = 1$ and x_{1j} be a categorical variable taking two different values indicating whether sample j belongs to category 1 or 2 (1 can be healthy and 2 can be sick)

$Y_j = \beta_0 + \beta_1 x_{1j} + \varepsilon_j$ (constant if we have a sick sample)

Same equation as before

If we test if $\beta_1 = 0$ vs $\beta_1 \neq 0$ the linear model becomes equivalent to a t-test (assuming equal variance)

If x_{1j} takes k categorical values, the linear model is equivalent to an ANOVA

Parameter estimation

In linear models the parameters can be estimated from the likelihood function

Select parameters that maximize the function

$L(\beta_0, \beta_1, \dots, \beta_p, \sigma^2) = \prod_{j=1}^{m} f(y_j)$ $y_1, \dots, y_j$ are our observations

$f(y_j) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_j - (\beta_0 + \beta_1 x_{1j} + \dots + \beta_p x_{pj}))^2}{2\sigma^2}}$

Density function

Maximizing L results in ML point estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p, \hat{\sigma}_\varepsilon$

ML = maximum-likelihood

$\hat{\beta}_k \sim N(\beta_k, \sigma^2_{\hat{\beta}_k})$

If we replace $\sigma^2_{\hat{\beta}_k}$ with $S^2_{\hat{\beta}_k}$, then $T = \frac{\hat{\beta}_k}{S_{\hat{\beta}_k}} \sim t_{\underbrace{n-(p+1)}_{\text{degrees of freedom}}}$ (number of parameters)

T-distribution -> are parameters zeroed?

The distribution of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ is known,

$\hat{\beta}_k \sim \text{Normal}(\beta_k, \sigma^2_{\hat{\beta}_k})$

We can use this to test

$H_0: \beta_k = 0$

$H_1: \beta_k \neq 0$

Evaluation of linear models

SST (sum of squares total)

Measure of total variability

$SST = \sum_{j=1}^{m} (y_j - \bar{y})^2$ total variability

Second Y: mean

$\bar{y} = \frac{1}{n} \sum_{j=1}^{m} y_j$ $\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \dots + \hat{\beta}_p x_{pj}$

SST-SSE: variability explained by the model

$SSE = \sum_{j=1}^{m} (y_j - \hat{y}_j)^2$ error sum of squares

Error variability -> should be low: can predict very good

$R^2 = 1 - \frac{SSE}{SST} = \frac{SST - SSE}{SST}$ how much is left if the error is removed

$R^2_{adj} = 1 - \frac{n-1}{n-(p+1)} \frac{SSE}{SST}$

R^2: number between 0 and 1

-> 1: model fits the data nicely

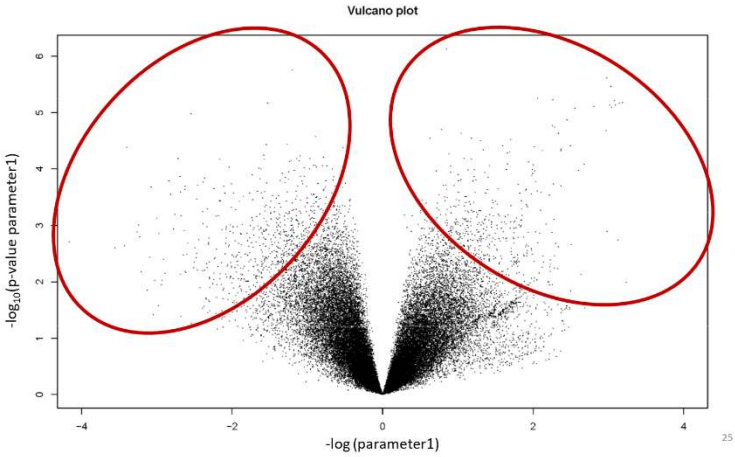
-> close to zero: there is a lot of variability in the data that is not explained by the model

-> how big is the portion of variability that the model explains

R^2_{adj}: takes number of parameters into account

-> adds a correction factor

-> a model with more parameters is required to fit the requirements will be higher



- Combine two histograms -> Combined in vulcano plot
- Minus added (typically done)
- Red circles: interesting genes; genes that differ, p value significantly high
- right: increase in agrressive group
- left; decrease in agressive group
- Each dot is a gene

Taking additonal parameters

- The extended model describes more/better, overall R score has increased
- R -> how big part of the variability is explained by the model
- Genes with a lower p-value (more significant) in the extended model

SUMMARY

- Linear models are highly flexible statistical tools that can be used to analyze gene expression data from RNA-seq experiments.
- Linear models are implemented in R under the lm function.
- Each gene is analyzed separately – this results in a large number of linear models, coefficients and p-values.
- Visualization of the results for all gene simultaneously can be used to verify the model and identify genes that are differentially expressed.
- Adding additional factors may increase the fit of the model, reduce the variability and increase the number of significant genes.

EXAM: In a research study, gene expression analysis is performed on colon cancer biopsies to identify differentially expressed genes involved in tumor development. The study contains two groups of patients, one with patients that have malignant tumors and one with patients that have benign tumors. The gene expression is measured for each gene (*n* = 25*;* 000 genes) using RNA-seq and the resulting data is normalized and log-transformed. A linear model is used to identify the differentially expressed genes.

(a) Formulate a linear model that can be used to identify genes that are differentially expressed between the groups. Introduce suitable variable names and model assumptions. (3p)

(b) In the study, the age and gender of each patient is documented. Extend the linear model in a) by including these variables. Why may the extended model be better for finding differentially expressed genes? (2p)

(c) What statistical assumptions are done in the analysis in a) and b)? Are these assumptions true? What may happen if the assumptions are violated? (2p)

EXAM: In a scientific study transcriptomics was used to investigate the gene expression patterns in breast cancer. Two groups of patients were included in the study, one group consisting of patients diagnosed with cancer (biopsies were taken from the tumor) and one control group consisting of healthy patients (biopsies were taken from normal tissue). The gene expression was measured for each gene (n = 20; 000) using RNA-seq and the resulting data was normalized and log-transformed.

(a) Formulate a linear model that can be used to identify genes that are differentially expressed between the groups. Introduce suitable variable names and model assumptions.

(b) How many parameters are needed to be estimated in order to fit the model to all of the genes in the dataset?

(c) Assume that each patient group can be further divided into the categories 'old' (older than 50 years) and 'young' (younger than 50 years). Update the model in (a) to include this new information.

(d) Compare the model in (a) and (c). What are the advantages of the model in (c)? Can you see any advantages of the model in (a)?

L9: Multiple testing

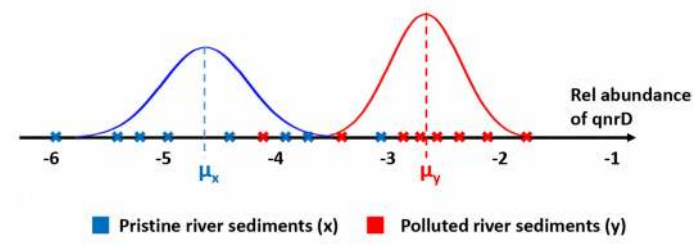
In bioinformatics we are working with high-dimensional data -> **High-dimensional data = many tests**

- **Genome sequencing:** Large number of positions in the genome where each tested for the presence of mutations.
- **Transcriptomics**: Expression of thousands of genes are measured where each gene is tested for differentially expression.
- **Metagenomics:** Bacterial communities contains tens of thousands of species where each species is tested for differentially abundance.

Statistical hypothesis test

1. Formulate assumptions of the data
2. Formulate statistical hypothesis: Null hypothesis, Alternative hypothesis
3. Apply a decision rule that decides when null hypothesis should be rejected

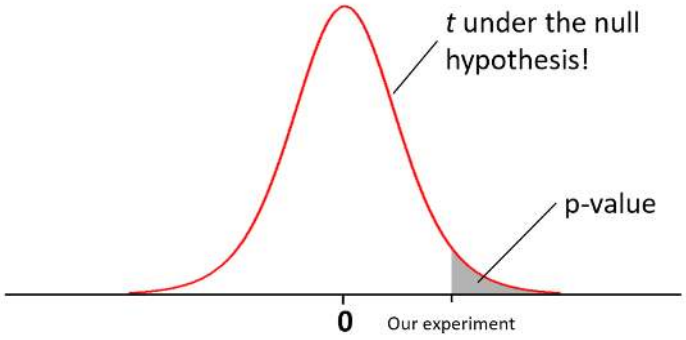
The t-test



H_0: qnrD has the same abundance in pristine and polluted sediments ($\mu_y=\mu_x$)

H_A: qnrD is more abundant in polluted sediments ($\mu_y>\mu_x$)

$$t = \frac{\bar{y} - \bar{x}}{\sqrt{\mathrm{SE}_y^2 + \mathrm{SE}_x^2}}$$



- **p-value:** probability that we see something more extreme given that the null hypothesis is true
  - measures the probability of more extreme cases -> measures the deviation from the null hypothesis
  - The smaller the pvalue the more extreme our observation is in the light of the null hypothesis
- **Gray area to the right:** probability that we get a bigger deviation

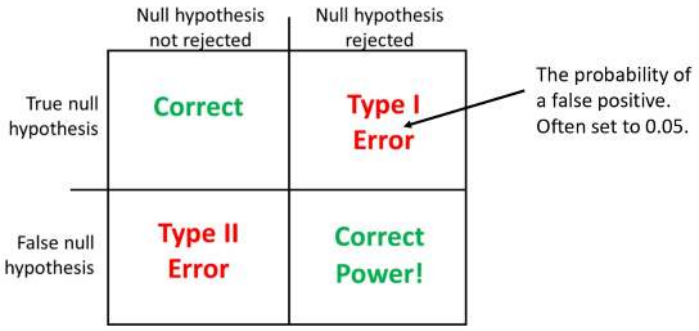
Facts about the p-value

1. Measures deviation from the null hypothesis ("significance")
2. When the null hypothesis is true, the p-value is a random value between 0 and 1
3. Summaries the entire testing procedure

The absence of evidence is not the evidence of absence

Nonsignificant pvalue -> large value -> does not imply that the null hypothesis is true -> we just failed to show that we have a significant deviation from the null hypothesis

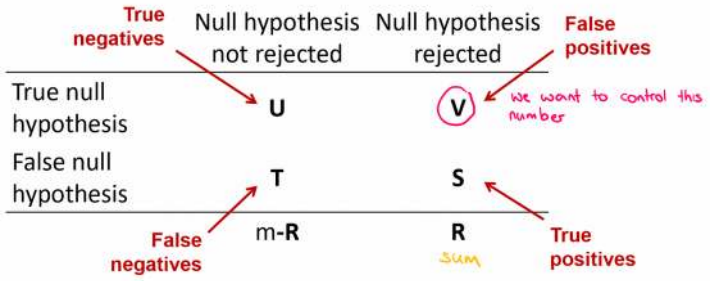
The outcome of a test



- Type 1 error: probability you typically want to control = same as the p-value (measures deviation from the null hypothesis); should be small enough
- The smaller you make your type 1 error, the larger gets the type 2 error and vice verse

Multiple Testing

Assume that we perform m tests.



- We do not do one test, we do multiple tests that end up in one of the four squares
- Count how many of our tests end up in the 4 categories
- Probabilities are replaced with numbers -> Matrix of numbers
- Each outcome of the test will be random

Performing statistical tests on thousands genes/positions can result in a large number of false positives

Assume that we perform m tests.

Performing m tests at a level α where H_0 is true result in $m \times \alpha$ false positives (in average).

Correcting for multiple testing means controlling the number of false positives (i.e. **V**)!

Correcting the results for multiple testing means that we control ('keep track') on the number of false positives

Family-wise error rate (FWER)

The probability of at least one false positive, i.e.

$$\text{FWER} = \text{Prob}(V > 0) = P(\text{at least one false positive})$$

Under the null hypotheses (H_0 is true for the m tests)

$$\begin{aligned}\text{Prob}(V > 0) &= \text{Prob}(\text{at least one false positive}) \\ &= 1 - \text{Prob}(\text{no false positives}) = 1 - (1-\alpha)^m\end{aligned}$$

- If FWER is small it means that the probability that V is exactly equal to zero is large which means that we do not have a single false positive -> good
- If we do one single test, the FWER will basically coincide with the type 1 error or the level of the test alpha
- As soon as we increase the number of tests, this probability will increase quite rapidly

The Bonferroni correction

Bonferroni controls the probability of at leat one false positive (Family-wise error rate). It is often considered to be too conservative

Algorithm: Bonferroni correction

Divide the significance level α (p-value cut-off) by the number of performed test (i.e. m).

A Bonferroni adjusted p-value can be calculated by multiply each p-value with the number of performed tests (i.e. m).

It can be shown that Bonferroni corrected p-values always control the Family Wise Error Rate (using Boole's inequality).

Interpretation: If we perform m tests and use the significance cut-off α/m, then $\text{Prob}(V > 0) \leq \alpha$.

- make sure that the FWER does not increase
- If we can keep the FWER under a certain level regardless of how many tests we make, we will limit the probability we get for false positives -> Bonferroni correction
- The FWER is at most alpha
- If we are stricter to alpha we loose more power
- It is a very conservative method -> we loose more power

$$p_{adj} = \min(1, p \times m)$$
$$p_{adj} < 0.05$$

A stricter p-value cut-off comes at the cost of reduced statistical power!

Bonferroni and controlling the FWER is often considered to be too conservative! in bioinformatics

- We require bigger deviations from the null hypothesis -> we need more evidence to say that the null hypothesis is not true -> if it is indeed the case that the null hypothesis is not true, it is hard to reach that conclusion, especially if you have little evidence -> will directly affect the statistical power
- Argument against Bonferroni: B is quite conservative, it is often ok to have a few false positives especially if that also includes to have more true positives
- If you have a less strict approach also some more true tests are included
- One typical scenario when it is not too conservative: If you have a certain number of tests and in order to reach the argument from point A to point B you need all of thes tests to be true, then you need to do Bonferroni correction and it is not too conservative
- But as soon as your interested to test a bit separately it is ok if one of your tests is false posititves

EXAM: Describe the concept of false discovery rate and how it is calculated. How can the false discovery rate be used to control the number of false positives? (3p)

False discovery rate (FDR)

The false discovery rate (FDR) controls the expected proportion of false positive among the signifcant tests

The number of false positives in relation to the total number of rejected null hypotheses (significant tests).

$$\mathrm{FDR} = \mathrm{Exp}[\mathrm{V}/\mathrm{R}]$$

- Is a bit more relaxed, can allow several false positives, but the rate will still be small

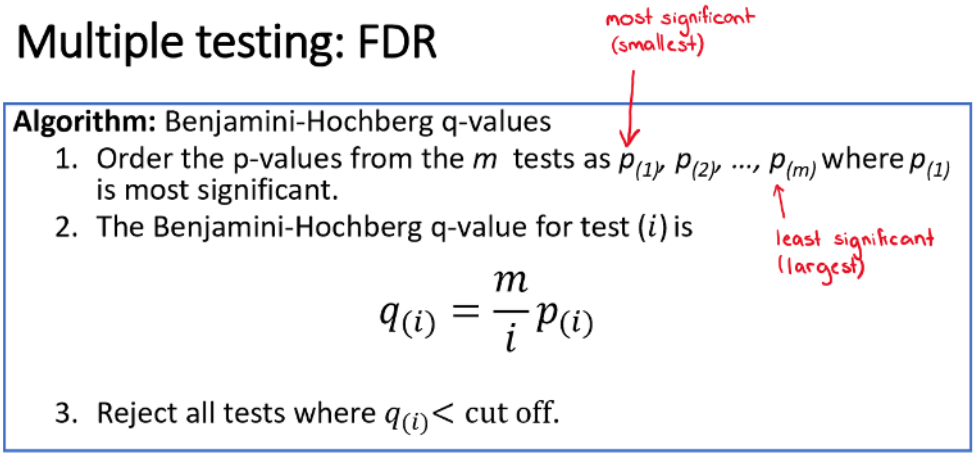
EXAM: Describe the Benjamini-Hochberg method of false discovery rate (FDR). How can it be used to solve the multiple testing problem

Benjamini-Hochberg correction

- Benjamini-Hochberg correction controls the false discovery rate (FDR). Proof is complicated and based on
- Assumptions:
  - The individual tests are independent.
  - P-values from tests where the null hypothesis is true are uniformly distributed (i.e. that the statistical distributional assumptions are true).
- BH is more robust -> relies on less assumptions
- The FDR cut-off can be set to the proportion of false positives that you can handle (0.05-0.20 is common in genomics)
  - 0.1 -> 10 percent are estimated to be false, 90 percent are estimated to be true
  - 0.2 -> you will allow more significant tests, because the correction will not be as severe, a bigger proportion will be false

- The more relaxed cut-ooff/false positives you allow, the more significant tests you will get, the more the statistical power will increase -> set how big proportion of false positive you can handle
- Family wise error rate is more conservative than the BH because the pvalues after the correction are larger

Multiple testing: FDR



- q: adjusted p-values
- i: position of the pvalue
- Largest p- value has i=m and thus will be not adjusted at all
- Everything less than the cut-off will be called significant
- The FDR is estimated to be less than the cut off

Multiple testing – a few tips

For high-dimensional genomics data

- Correction for multiple testing should always be done to ensure that there is not too many false positives. In many situtations, correction for multiple testing is a requirement.
- Bonferroni is often too strict and a small proportion of false positives is typically acceptable. Estimation of the false discovery rate is therefore often preferred.
- It is good to minimize the number of comparisons. Only test what you need to know!

EXAM: In bioinformatics, a large number of tests is often performed simultaneously. This gives rise to the multiple testing problem.

(a) Describe the multiple testing problem. What can happen if it is not properly addressed? (2p)

The multiple testing problem arise when several statistical hypotheses is tested. For each test, the probability for false positives will increase. In fact, if □ hypotheses are test at a significance level of □ we expect to have □ × □ false positives in average. The false positives can, if not properly handled, lead to incorrect interpretation of the results

The multiple testing problem arise when testing many hypotheses. If we, for example, test 100 tests, which at a level of 0.05 we will get 5 significant tests even when all null hypotheses are true. The challenge with multiple testing is thus to differentiate between false positives and true positives. See lecture notes for full details.

(b) What is the family-wise error rate (FWER)? Also, describe the Bonferroni method and how it can be used to control the FWER (3p).

The familiy-wise error rate (FWER) is the probability that we have at least one false positive. By controling the FWER (make sure that it is below a certain value), we will limit the number of false positives.

The Bonferroni correction method controls the FWER. The method can either modify the significance level by dividing it by the number of tests. Alternatively, the p-values can be adjusted by multiplying them with the number of tests. The Bonferroni method ensures that the FWER is then below the significance level. Note, however, that this method often is considered to be too conservative.

In short

FWER is the probability that at least one of the tests results in false positives. The Bonferroni method adjusts the pvalues by multiplying them with the number of performed tests. This will control the FWER.

(c) Assume that we search for mutations in a genome consisting of 10,000,000 positions. The five most significant genes have the pvalues 7×10-12, 2×10-11, 3×10-9, 3×10-8 and 7×10-7. Use the Bonferroni method to adjust the p-values. How many positions are significant if we require the FWER to be less than 0.01? (2p)

Original p-values:
7×10^{-12}, 2×10^{-11}, 3×10^{-9}, 3×10^{-8} and 7×10^{-7}.

Number of tests: 10,000,000

Adjusted p-values:
7×10^{-5}, 2×10^{-3}, 3×10^{-2}, 3×10^{-1} and 1.

Number of significant tests after adjustment: 2

After adjustments, i.e. multiplication by 10,000,000, two tests are significant (i.e. have p-values<0.01)

EXAM: Assume that we search for differentially expressed genes in a transcriptomic dataset. The dataset contains in total 20,000 genes. The five most significant genes have p-values 1.5×10-6, 4×10-6, 2.5 × 10-5, 6 × 10-5 and 8 × 10-5. Use the Benjamini-Hochberg method to correct the p-values. What is the false discovery rate if the significance threshold is set after the fifth gene? Explain the result. (2p)

SOLUTION: $FDR = 8 \times 10^{-5} \times 20000/5 = 0.32$

EXAM: In the analysis of differentially expressed genes, multiple tests are typically performed. This can result in a number of false positives. One way to control for multiple testing is to use Bonferroni method. Describe the Bonferroni method and how it limits the number of false positives. Is Bonferroni a suitable approach for this particular application? Why? Why not?

Bonferroni controls for multiple testing by controlling the family-wise error rate (FWER), which is defined as the probability that we have at least one false positive among our tests. To perform a Bonferroni correction, the significance threshold is modified by divided it by the total number of tests. Alternatively, adjusted p-values can be calculated by multiply all p-values with the total number of tests. The Bonferroni method is often too conservative for RNA-seq and can significantly reduce the statistical power. Often, we can allow a proportion of false positive as long as it is not too high.
Thus, alternative approaches, such as the false discovery rate (FDR) is often preferred.

EXAM: What are the main statistical issues associated with multiple testing? How does Bonferroni adress the issues?

- Performing statistical tests on thousands of Fgenes/positions can result in a large number of false positives
- Correcting for multiple testing means controlling the number of false positives
- Bonferroni: controls the probability of at least one false positive (family-wise error rate), often considered to be too conservative

EXAM: Assume that you search for differentially expressed genes in a transcriptomic dataset. The dataset contains in total 10,000 genes. The three most significant gene have the p-values 3.75 × 10-7, 7.32×10-7 and 2.12×10-6. What are the corresponding Bonferronicorrected p-values? How many genes would be significant if the cut-off for the corrected p-values were 0.01? (2p)

Gene	P-value
Gene 7321	6.5×10^{-10}
Gene 6236	2.3×10^{-9}
Gene 18345	7.8×10^{-9}
Gene 64	1.1×10^{-8}
Gene 1098	9.3×10^{-7}
Gene 14037	4.5×10^{-6}
Gene 2254	2.3×10^{-6}
Gene 845	1.2×10^{-5}
Gene 16346	5.6×10^{-5}
Gene 1658	8.5×10^{-5}

L10: Unsupervised data exploration

Supervised methods

- Supervised methods rely on metadata, i.e. data that provides information about the data. This can be any form of medical, biological or technical information about the samples.
- Linear models are a supervised method. Here we specifically look for differences associated with a covariate (i.e. metadata). This includes, for example, identification of differentially expressed genes between different groups of patients.

Unsupervised methods

- Unsupervised methods does not utilize metadata, i.e. no additional information about the samples are used.
- Unsupervised methods focus on the identification of patterns in the data. In bioinformatics this typically means patterns between samples or genes.
- Unsupervised methods are explorative and does not rely on any specific hypotheses. This means that we can identify almost any form of patterns.
- Another common application of unsupervised methods is dimension reduction, where high-dimensional data is visualized in e.g. two dimensions.

Two main methods:

Clustering:

- aims find groups of samples or genes ('clusters') that are 'similar'.
- Two important methods: Hierarchical clustering and K-means clustering.

Principal component analysis (PCA):

- used for exploration and visualization of high-dimensional data
- identifies the dimensions with highest variability. These are used to represent the data at a lower dimension (e.g. in two dimensions).

Hierarchical clustering

Builds a hierarchy describing the similarity between clusters

1. **Agglomerative:** All samples start as individual clusters which are merged together (aggregated) according to
   a. A distance measure describing the separtation between data points
   b. A linkage criterion measuring the distance between clusters
2. **Divisive:** All samples start in a single cluster which is then successively divided

The number of clusters are then defined based on a minimum similarity threshold

Distances

All clustering techniques are based on a distance metric that measures the similarity between two data points

A few notations

We will use $x = (x_1, \dots, x_p)$ and $y = (y_1, \dots, y_p)$ to denote two data points defined in $\mathbb{R}^p$. Let $d(x, y)$ denote the distance between x and y.

Let X and Y denote two clusters with k and l members each, i.e. $X = \{x^{(1)}, x^{(2)}, \dots, x^{(k)}\}$ and $Y = \{y^{(1)}, y^{(2)}, \dots, y^{(l)}\}$. Let $D(X, Y)$ denote the distance between X and Y.

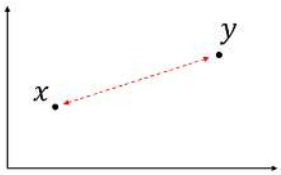
The distance matrix

- Contains all pair-wise distances between the data points.
- The distance matrix is symmetric since $d(x, y) = d(y, x)$.
- This matrix can be really big since it contains n^2 elements (only $n(n-1)/2$ elements are actually needed).
- Some algorithms works instead with similarity matrices.

	1	2	3	4	5	6	7	8	9	10
1	0.00	5.74	5.04	4.60	4.11	4.63	4.13	5.70	4.59	3.33
2	5.74	0.00	3.49	6.29	5.96	5.30	6.62	7.22	5.68	4.60
3	5.04	3.49	0.00	5.68	5.70	4.26	4.62	6.25	5.94	4.94
4	4.60	6.29	5.68	0.00	3.73	4.98	5.95	6.44	3.71	3.18
5	4.11	5.96	5.70	3.73	0.00	5.45	5.50	4.97	4.33	3.76
6	...									

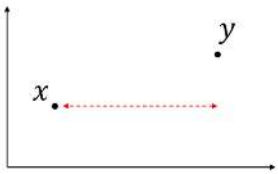
Euclidean distance

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$



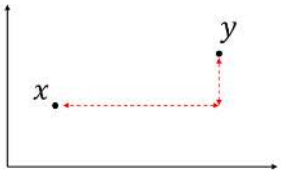
Maximum distance (L^∞)

$$d(x,y) = \max_i |x_i - y_i|$$



Manhattan distance

$$d(x,y) = \sum_{i=1}^{p} |x_i - y_i|$$



Correlation distance

$$\text{corr}(x,y) = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)\text{var}(y)}} \quad \text{and} \quad d(x,y) = \frac{1 - \text{corr}(x,y)}{2}$$

where $\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\text{var}(x) = \text{cov}(x,x)$.

Agglomerative hierachical clustering

Algorithm: Agglomerative hierarchical clustering

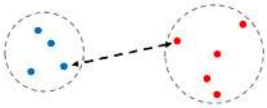
1. Calculate a distance matrix containing the pair-wise distance between all data point
2. Let each data point be a cluster
3. Identify the most similar clusters X and Y according to a *linkage criterion.*
4. Merge X and Y into a new cluster. Update the distance matrix.
5. Goto 3 if the number of clusters are more than 2.

Linkage Criteria

The linkage criterion calculates the similarity between clusters.

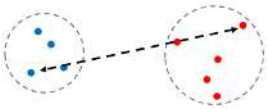
Single linkage

$$D(X,Y) = \min_{x \in X, y \in Y} d(x,y)$$



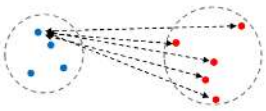
Complete linkage

$$D(X,Y) = \max_{x \in X, y \in Y} d(x,y)$$



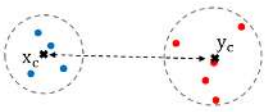
Average linkage

$$D(X,Y) = \frac{1}{kl} \sum_{\substack{\text{all pairs } (x,y), \\ x \in X, y \in Y}} d(x,y)$$



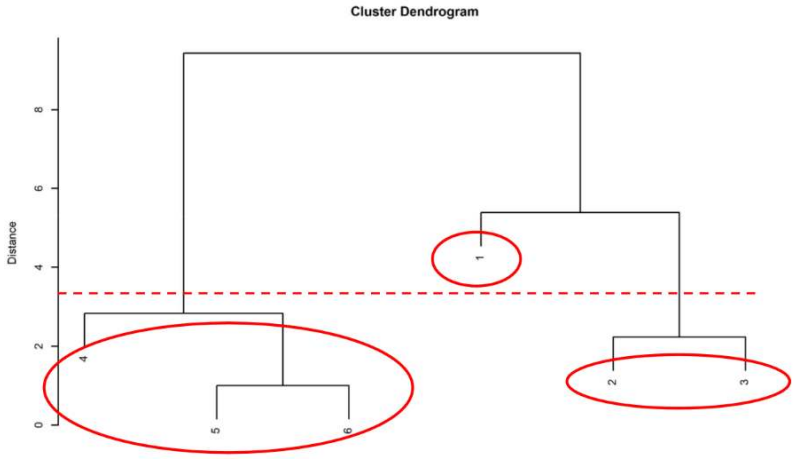
Centroid linkage

$$D(X,Y) = \text{d}(\text{x}_\text{c}, \text{y}_\text{c})$$



Dendogram

- Dendogram describes the different merge
- Data points at the leaves
- Y-axis: distance -> Distance from distance matrix
- Hierarchy shows where the different clusters are merged
- Depending on which size of groups we look at/how many clusters we want, we can cut off accordingly



K-means clustering

- The k-means clustering algorithm will always converge (is equivalent to the Expectation-Maximization algorithm).
- A drawback with K-means clustering is that it is often hard to decide the number of clusters ().
- There are more advanced initiation methods that provides faster convergence.

- A method for clustering n observations in to s clusters. Note that s is pre-specified!
- We want assign each sample to one of the s clusters such that the variance within each cluster is minimized. This means that we want to minimize

$$\sum_{j=1}^{s} |c_j| Var(c_j).$$

Cluster size

- K-means clustering implicitly assumes a Euclidean distance between points.

The variance within a cluster can be calculated by

$$Var(c_j) = \sum_{x \in c_j} \sum_{i=1}^{n} (x_i - \mu_i)^2$$

where μ_i is the *mean point* of c_j. Note that minimizing $Var(c_j)$ is equivalent to minimizing the Euclidian distance between all pairs of data points in the cluster, i.e.

$$\frac{1}{2} \sum_{x,y \in c_j} \sum_{i=1}^{n} (x_i - y_i)^2$$

The naïve algorithm

Start with an initial set of k randomly selected cluster means $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$.

Step 1: Assignment

Assign each data point to the closest cluster based on the Euclidean distance to the cluster mean.

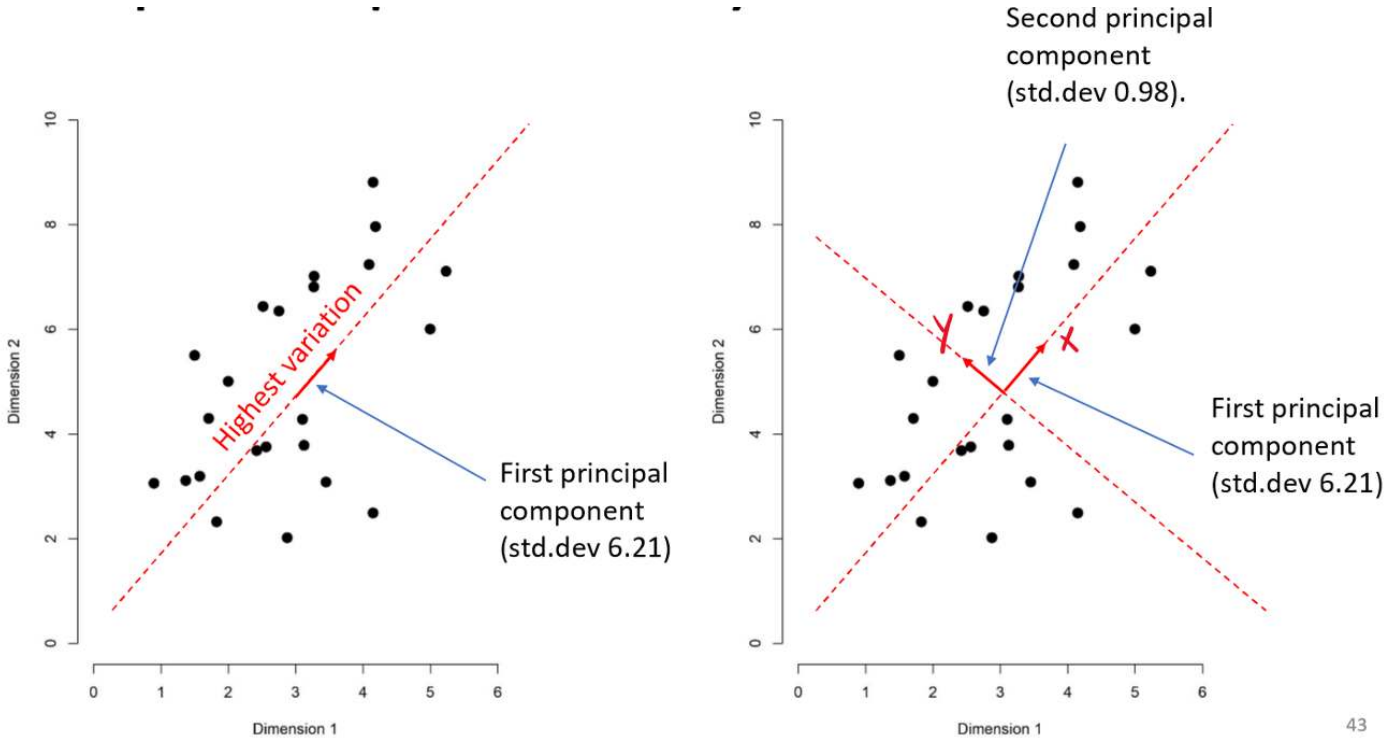
Step 2: Update

Recalculate the cluster mean values.

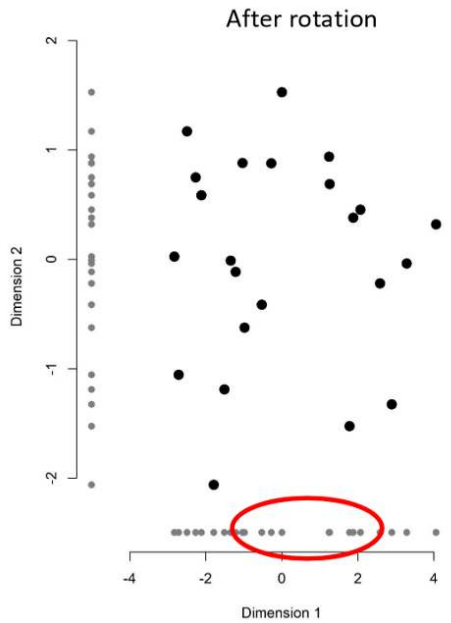
Repeat these steps until no more assignments are done.

Principal component analysis (PCA)

- Method used for exploration and visualization of high-dimensional data
- The main principle: Identify the dimensions of the data that explains as much variability (data points that are clearly separated; search for dimensions in which the data is spread the most) as possible. These dimensions are defined by principal components.
- The data can then be projected on a set of the to principal components (often two) which results in a reduction of the dimensionality
- PCA is often used to identify patterns in the data. However, PCA does not assign each data point to a cluster.
- Number of principal components = number of dimensions



There are gaps that we cannot see because of high-dimensionality which appear once we find out the PC -> Plot the data and see separation



Some notation

We will, as before, use $x = (x_1, \ldots, x_p)$ to denote a data point. Note that we will here regard x as a random variable.

Let $y_1, y_2, \ldots y_n$ denote n linear combinations of the coordinates of x, i.e.

$$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = a_1^T x$$
$$y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{1n}x_n = a_2^T x$$
$$\ldots.$$
$$y_n = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = a_n^T x$$

The principal components will be identified by selecting $\{a_{ij}\}$ according to the following rules

The first principal component is the linear combination $y_1 = a_1^T x$ that maximizes the variability $\mathrm{Var}(y_1)$ under the condition that $a_1^T a_1$=1.

The second principal component is the linear combination $y_2 = a_2^T x$ that maximizes the variability $\mathrm{Var}(y_2)$ under the condition that $a_2^T a_2$=1 and is independent of y_1.

The *nth* principal component is the linear combination $y_n = a_n^T x$ that maximizes the variability $\mathrm{Var}(y_2)$ under the condition that $a_n^T a_n$=1 and is independent of $y_1, y_2, \ldots, y_{n-1}$.

The principal components corresponds to the eigenvectors of the covariance matrix. They are therefore often calculated by singular value decomposition (SVD).

A principal component analysis results

1. The principal components $y_1, \ldots, y_n$ constitutes a new orthogonal basis for the data. The number of principal components is the same as the dimension of the data (or, if lower, the number of samples).
2. The rotation matrix, which is formed from vectors $a_1, a_2, .., a_n$. This matrix describe how data can be rotated to the new basis. The individual elements are sometimes called loadings. Note that the rotation matrix is not unique – it can always be multiplied by -1.
3. The estimated variance for each principal component. These are always in decreasing order.

Clustering and PCA – some remarks

- Clustering and PCA are explorative methods. They can be used to visualize data and formulate hypotheses but they does not result in any statistical test (and thus no measure of the significance).
- Many methods automatically center and scale the data. You may need to manually turn such features off.
- There is a wide range of different tools for clustering and PCA. The command heatmap can for example cluster and display the results as a heatmap and a dendrogram.

SUMMARY

- Unsupervised methods can be used to explore and visualize highdimensional data.
- Clustering is a technique used to find groups of data points ('clusters'). Common clustering methods includes hierarchical clustering and k-means clustering. A distance metric and linkage criterion is needed to perform hierarchical clustering.
- Principal component analysis can be used to reduce the dimension of the data. The principal components are identified based on the dimension on where the variability of the data is as high as possible

EXAM: Unsupervised methods are often used to find patterns in high-dimensional data.

(a) What is the main difference between clustering and principal component analysis (PCA)? Give examples on when to use clustering and when to use PCA.

- Clustering and PCA are both unsupervised methods.

- The aim of clustering is to identify clusters, i.e. group of genes or samples that shows similar patterns in the data.

- The aim of PCA is to produce a lowdimensional representation of the data. This can, for example, be used to visualize highdimensional data is 2d- or 3d- plots. Since high variability sometimes corresponds separation of data-points, PCA is useful to identify heterogeneity of the data. Note however, that PCA does not explicitly identify clusters.

- One example when clustering is used is to identify subgroups of samples in cancer research based on transcriptomics data. Here, each subgroup can correspond to different properties of the tumors.

- One example where PCA is used is to analyze patterns from amplicon sequence data. Here, the PCA can be used to visualize the relationship between samples and thus demonstrate there relationships

In short
Clustering is the process where we group samples into clusters based on their similarity. PCA is a method to visualize high-dimensional data in lower dimensions by picking dimensions that describe most of the variability. Clustering is thus used when we want to group samples while PCA is used to explore patterns in the data.

(b) How does hierarchical cluster analysis work? Describe the different steps of the algorithm and explain meaning of a 'distance metric' and a 'linkage criterion'.

Hierarchical clustering can be either agglomerative or divisive. In agglomerative clustering, each data-point starts as its own cluster. These are then merged successively. The distance between data-points are measured using a distance metric (e.g. Euclidean or correlation). The distance between clusters is measured using a linkage criterion (e.g. single linkage or complete linkage).

First a distance matrix, describing the pair-wise distances between all clusters are calculated. In each step, the algorithms identifies the clusters with the lowest distance. These clusters are then merged and the distance matrix is updated. This is repeated until all data-points are in one single cluster. The process can be visualized as a dendrogram, which can be used to select a suitable number of clusters.

In short
Hierarchical clustering works by successively merge samples into larger and larger clusters. The distance metric defines how to measure the distance between two samples. The linkage criterion defines how to measure the distance between two clusters. The algorithm works by merging the closest clusters until all samples are in one single

clusters. The results are typically visualized in a dendrogram describing at what distances the different clusters merged. See lecture notes for full details.

EXAM: Describe the difference between supervised and unsupervised data analysis. Describe a scenario where unsupervised data analysis would be of interest.

In supervised data analysis, there is a predefined organization of the samples that is utilized in the analysis. For example, if samples are collected from sick and health individual (e.g. samples for RNA-seq) we typically perform a supervised analysis where we compare the genes between the predefined groups 'sick' and 'healthy'. In contrast, unsupervised analysis we do not take advantage of any pre-defined structure of the data. Instead, we use analysis methods to see if the samples are grouped or organized according to a specific pattern. One example where unsupervised analysis can be used is to see whether samples taken from the tumour of a set of individuals separate into distinct sub-groups or if they are all from one large group.

Supervised analysis are based on metadata, such as information about the samples (e.g. 'sick' and 'health' individuals). Unsupervised analysis use no metadata and is instead used to find patterns in the data. Supervised analysis is typically used to pin-point specific genes/mutations/species or other features between pre-specified groups. So as long has you have a specific hypothesis or question and sufficient information about your samples, supervised analysis is preferred. Unsupervised analysis is used to generate new questions from the data. It recommended when you want to discover new patterns or trends in the data or to analyze samples where metadata is lacking

EXAM: Describe the main difference between hierarchal clustering and k-means clustering

Hierarchical clustering builds a hierarchy between clusters using a pre-defined distance metric. The most common form is the agglomerative clustering where each observation starts in each own clusters, which are then successively merged into larger clusters. The merging of clusters, as a function of the distance metric, can be visualized as a dendrogram. The merging of clusters can be done using approaches such as complete linkage and single linkage. The number of formed cluster will vary depending on the similarity cut-off. k-means clustering aims to divide the observations into k separate clusterings. In contrast to the hierarchical clustering, k is here pre-specified. This is done forming the k-clusters in a way that each observation is a member of the cluster for which it has the shortest distance to its mean.

EXAM: What is principal component analysis (PCA) and how can it be useful in the analysis of count data from tanscriptomics or metagenomics?

Principal component analysis is a method to analyze highdimensional data by representing in fewer, typically two, dimensions. The method works by identifying the dimension which explains most of the variability of the high dimensional data. These dimensions are called principal components. By representing the data along these dimensions, as much of the information as possible can be visualized in two dimensions. Principal component analysis, thus, enables identification of patterns in the data that cannot be easily seen by examining all of the dimension of the data simultaneously.

EXAM: Describe the difference between PCA and hierarchical agglomerative clustering? When do you use clustering?

Hierarchical agglomerative clustering is used to divide the data (i.e. all genes or samples) into groups (clusters). PCA does not do any clustering analysis and does thus not provide any clustering in the end. Instead, PCA identifies dimensions of the data that consists of most of the variability – and thus is likely to contain a large part of the information.

It is, however, possible to identify a clustering of the data by examining the principal components explaining of the variability. Indeed, if the data is structured in separated clusters, this often give rise to a large variability, which can

thus be captured by the principal components. PCA will, however, not provide data points with any formal assignment into the different clusters.

EXAM: How does the choice of distance metric influence the outcome of an hierarchical clustering? What difference do you expect when the correlation distance is used instead of the Euclidean distance?

The choice of distance measures is a critical step in clustering. It defines how the similarity of two elements (x, y) is calculated and it will influence the shape of the clusters.

Correlation-based distance considers two objects to be similar if their features are highly correlated, even though the observed values may be far apart in terms of Euclidean distance. The distance between two objects is 0 when they are perfectly correlated.

If we want to identify clusters of observations with the same overall profiles regardless of their magnitudes, then we should go with *correlation-based distance* as a dissimilarity measure. This is particularly the case in gene expression data analysis, where we might want to consider genes similar when they are “up” and “down” together.

If Euclidean distance is chosen, then observations with high values of features will be clustered together. The same holds true for observations with low values of features.

EXAM: The linkage criterion defines how the distance between clusters are measured. What is the difference between using complete linkage and single linkage? Give one example where it is suitable to use complete linkage and one example where it is suitable to single linkage.

- Single linkage: looks at all possible distances and takes the smallest distance
    - You link two clusters based on the minimum distance between 2 elements. A drawback of this method is that it tends to produce long thin clusters since you make the link based on only 2 points.
- Complete linkage: looks at the maximum distance -> sets an upper limit on the distance -> much mores conservative
    - You link two clusters based on the max distance between 2 elements. Opposite problem with single link. Clusters tend to be overly conservative.
    - produces tighter clusters than single-linkage
- Average link: Instead of making a decision based on a single pair of elements, you take the distance between every pair of elements.

EXAM: Describe how a dendrogram can be used to visualize the result of the hierarchical clustering and how it can aid in deciding the number of final clusters. (2p)

A dendrogram is a tree-structured graph used in heat maps to visualize the result of a hierarchical clustering calculation. The result of a clustering is presented either as the distance or the similarity between the clustered rows or columns depending on the selected distance measure.

The key to interpreting a hierarchical cluster analysis is to look at the point at which any given pair of cards “join together” in the tree diagram. Cards that join together sooner are more similar to each other than those that join together later.

The key to interpreting a dendrogram is to focus on the height at which any two objects are joined together.

In the dendrogram above, the height of the dendrogram indicates the order in which the clusters were joined.

Once the dendrogram has been constructed, we slice this structure horizontally. All the resulting child branches formed below the horizontal cut represent an individual cluster at the highest level in your system and it defines the associated cluster membership for each data sample.

L11: Metagenomics

Microorganisms

- Present in every habitat and are integral members in almost every ecosystem on Earth
- Microorganism = Organism that you can not see without a microscope; it is too small to be seen by the eye
    - Bacteria
    - Eucaryotes
- The vast majority of microorganisms are unculturable, i.e. they can not easily be cualtivated in a lab using standard protocols

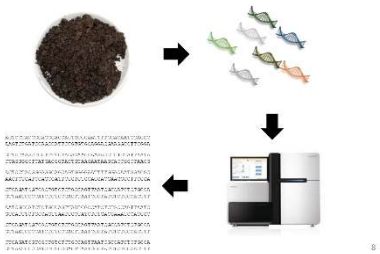
Metagenomics

Definition

Metagenomics = study of the metagenome, which is the collective genome in a microbial community
Abundance = The total number of individuals of a species or type present in a given area, in a given ecosystem or within a particular habitat

- Microorganisms are often organized in complex communities
- In metagenomics, DNA is randomly sampled from the metagenome which is used to derive information about the present microorganisms and their biological functions.

1. Anything that contains bacteria/microoragnism
2. Pick out DNA from everything that lives
3. Put it through sequence scanner -> fragments
4. Get gene sequence



Fragments can come from different species, but we do not know from which species they come from
Challenge: what does it mean, where it comes from

Two approaches

Who are there?

- Analysis of the present microorganisms and their taxonomic affiliation (species, genus, order, etc)
- Estimation of abundance and diversity – how many species are there, how common are they and do they differ between samples?

What are they doing? ('gene-centric analysis')

- Analysis of the present genes and pathways
- Estimation of gene abundances and functional analysis of their biochemical role.

Two sequencing techniques

Amplicon sequencing ('metabarcoding')

- Sequencing of a specific genomic region of interest ('barcode') that are amplified by PCR (amplicon)
- The region is first PCR-amplified using 'general primers' that works for a wide range of species
- Before we sequence we add an amplifcation step where we focus on a specific genetic marker (is especially good in identifying special species)

Shotgun metagenomic sequencing

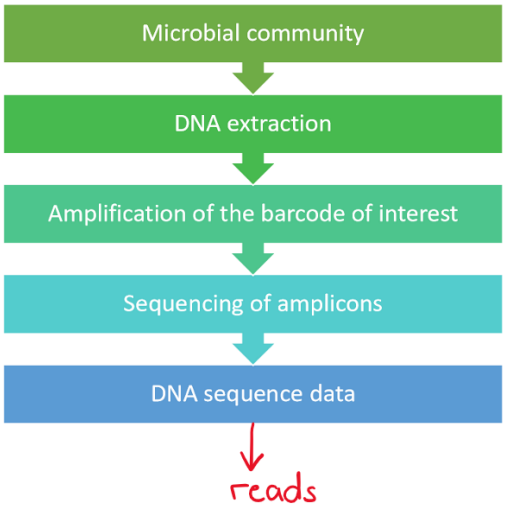
- Sequencing of random fragments from the entire metagenome (shotgun)

- Total DNA from a sample is sequenced
- Random fragment from the entire metagenome, including all of the genes in the present microorganisms

Amplicon sequencing

Amplicon data is analyzed by clustering the reads into OTUs which are then annotated

- Technique to address the question **"who is there?"**
- General idea: Focus only on a specific genetic marker that provides information about the organism. The marker is sometimes referred to as a barcode.
- Sequencing of a large number of markers from a sample provides a picture of the total microbial community



- Analysis starts from the total DNA present in a microbial community.
- The output is a large number of reads from the same marker (i.e. the same region) but from different microbial cells
- Amplification and sequencing of a specific barcode reduces the generated data volumes

14

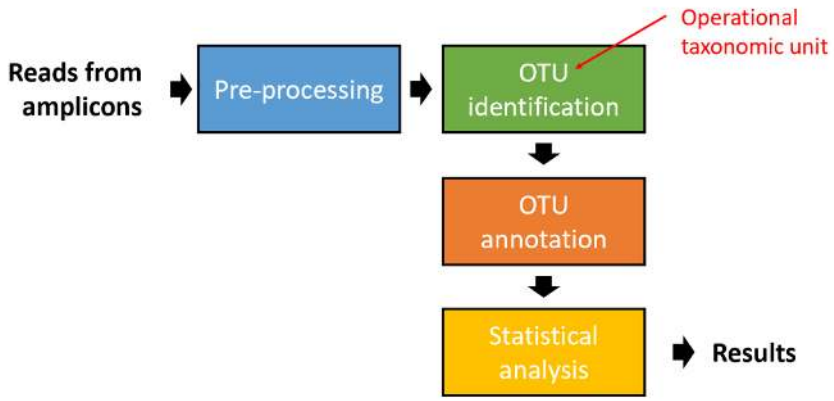
Marker selection

- Varies between genomes from different organisms
- Are flanked by conserved regions so that it can be easily amplified in as many species as possible
- The length matches the sequencing technique

- If it varies too little, too few differences between species, we loose the sensitivity
- If it is identical between two species we can not tell from which species it comes from
- If it varies too much (different between different cells within the same species), we can not match it -> at one point we need to compare it to a reference sequence /database
- If we get a lot of variability there we might not be able to say if it is the same species

--> variability of this region is important

Data analysis



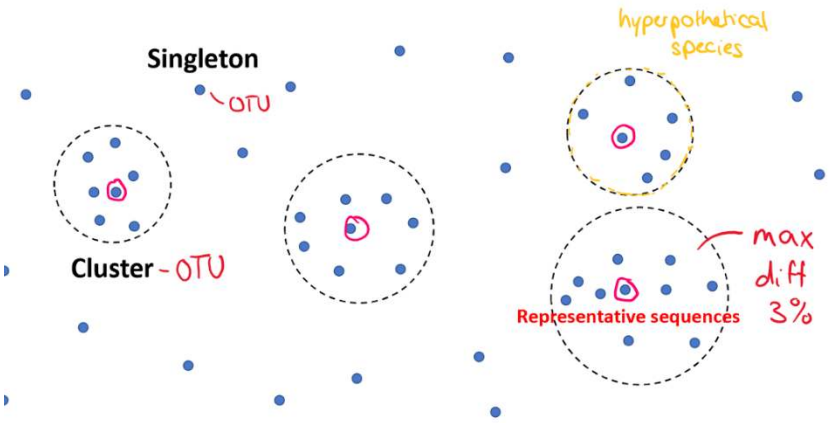
OTU = artificial species that are created by looking into the data; operational definition used to classify groups of closely related individual

Identification of OUT

The OTU abundance distribution can be used for comparisons between samples and for estimation of the biodiversity

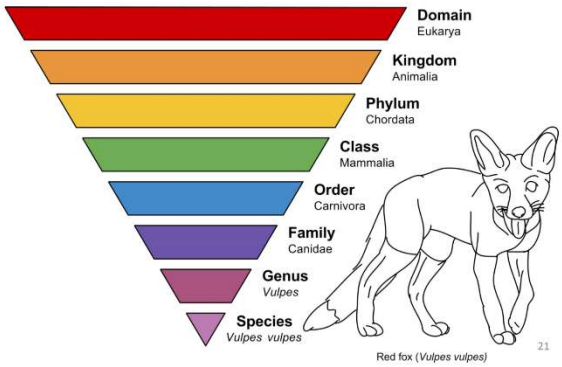
- An **operational taxonomic unit** (OTU) is a putative species formed by clustering sequences from amplicons (a piece of DNA or RNA that is the source and/or product of amplification or replication events).
- Sequences that are sufficiently similar are clustered together and assumed that they come from the same type of organism.
- Sequences that does not cluster with any other sequence are called **singletons**. These sequences are OTUs but are, in many cases, discarded since they are only observed once.
- A commonly used similarity threshold used for 16S is 97%. Those sequences that have a similarity >97% are clustered together and are thus assumed to be from the same OUT

- Form clusters around reads that are very similar
- Reads -> group them together into hyperpothetical species
- Number of reads into each cluster tells us something about the hyperpothetical species and its abundance
- -> the more reads -> the more members we have in the cluster
- -> the more common that species is in our data/sample
- Annotate the data



Taxonomic affiliation

- Hierachical system: Divide everything that lives on earth
- Theory and practice of grouping individuals into species, arranging species into larger groups, and giving those groups names, thus producing a classification



-

Diversity

Alpha diversity

The diversity on the local level (in a habitat). This include, for example, diversity at a specific geographical site or in a specific individual.

Beta diversity

Diversity between habitats. This include, for example, the diversity between geographical sites or between individuals.

Alpha diversity

Let p_i be the proportion of OTU $i = 1, \dots, n$.

Richness

Unique number of OTUs. Can easily be calculated by counting the number of OTUs where $p_i > 0$.

Evenness

Shannon diversity index. Estimates the evenness of a distribution by calculating the entropy,

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

A higher value indicates a more uniform distribution.

Evenness

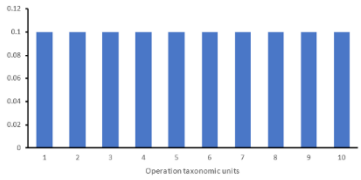
Simpson's diversity index. Calculates evenness by estimating how likely that two randomly picked OTUs are from the same species.

$$\lambda = \sum_{i=1}^{n} p_i^2$$

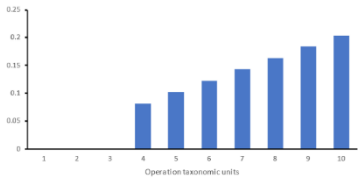
The Simpson index is a value between 0 and 1. A higher value indicates a less uniform distribution (max value achieved if $p_i = 1$ for some i).

First example: completely uniform
Second example: lacks, 10 dominates

Not the number is of interest but how they differ



Richness, unique OTUs: 10
Evenness, Shannon: 2.30
Evenness, Simpson: 0.10



Richness, unique OTUs: 7
Evenness, Shannon: 1.90
Evenness, Simpson: 0.15

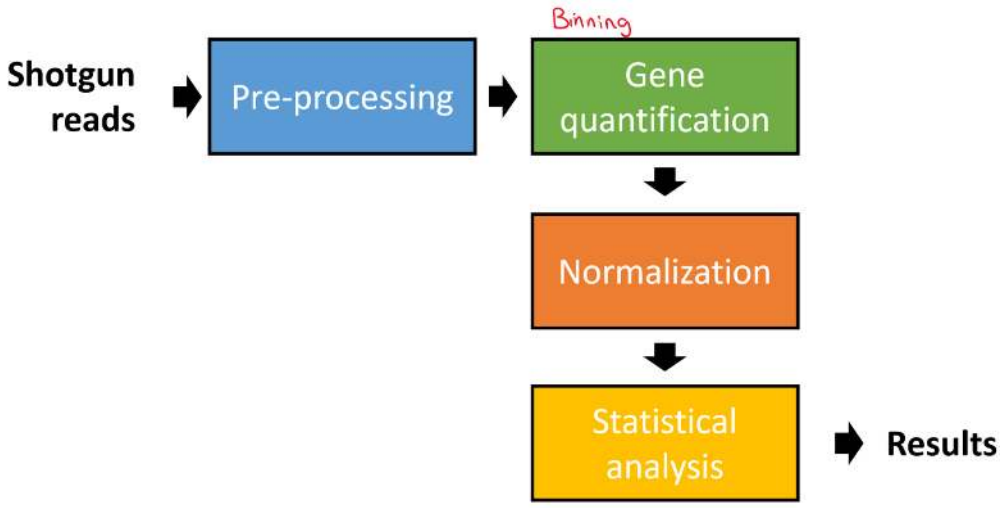
Diversity – rarefaction

- The diversity indices are dependent on the sequencing depth.
- Examples
  - A higher sequencing depth means more detected OTUs and therefore a higher richness.
  - More detected OTUs means a higher Shannon index (maximum is log(N) where N is the number of OTUs)
- In order to make indices between samples comparable they need to be rarefied, i.e. subsampled to the same sequencing depth.
  1. Select a sequencing depth n.
  2. For each sample, randomly sample n fragments without replacement.
  3. Count new OTU abundances for the rarefied data.

Shotgun metagenomic sequencing

- Total DNA from a sample is sequenced
- Random fragment from the entire metagenome, including all of the genes in the present microorganisms.
- Shotgun metagenomics can be used to address the question 'what are they doing'. This is done by analyze the present genes.

Data Analysis

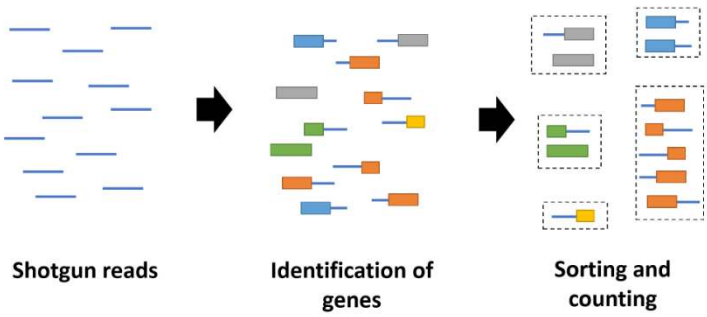


Gene quantification: we take all the reads and put them into bins that decribe their function

Gene quantification

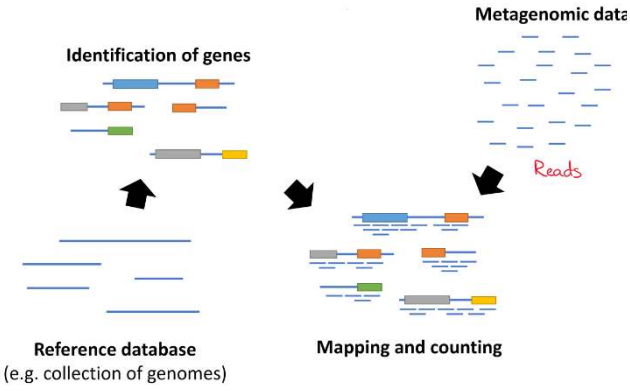
Direct binning of metagenomic reads

- Search each metagenomic fragments for the presence of genes.
- A vast number of the microbial genes are not present in the databases. The search therefore requires sensitive aligners and approximate matches are often accepted.
- Requires relatively long reads and not generally possible to do for short reads. Can be done for genes that are very well-conserved.
- 'Bins' are finally formed by counting the number of reads for each type of gene.



Reference-guided binning -> shorter reads

- Guided binning uses an annotated reference database that contains the genomes of the microorganisms present in the sample.
- Each metagenomic fragment is mapped against the reference database.
- 'Bins' are formed by counting the number of reads matching each type of gene present in the genomes.
- Typically done for data with short reads.
- Commonly used reference databases
    - Genomes
    - Gene catalogues
    - De novo assembly of the metagenome



- Multiple matches is also present in mapping of metagenomic data to the reference
    - Genes with regions that are evolutionary conserved between species,
    - Genes that are horizontally transferred between bacterial cells,
    - Errors in the reads caused by problems in the sequencing.
- Common solutions includes to handle reads with multiple matchesincludes
    - None – the reads are excluded from the analysis,
    - All – assign the reads to all its matching all regions,
    - Random – assign the read randomly to one of its matching regions

Normalization

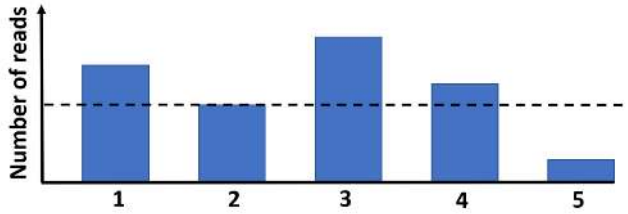
Data from different metagenomic samples are not directly comparable

- Differences in sequencing depth
- Systematic effects caused by
  a) Variation in sampling and preparation
  b) The quality sequencing
  c) Systematic errors introduced in the sequencing (e.g. GC-bias)
  d) Taxonomic composition

Normalization is necessary to make samples comparable.

Down-sampling (rarefying)

Each sample is down-sampled to a specified number of DNA fragments.



Advantage: Makes counts comparable.
Disadvantage: Discards valuable data. Samples with very low sequencing depth may need to be excluded.

28

Commonly methods for normalization are

• Counts per million reads (CPM)
• Reads per kilobase per million (RPKM)
• Median
• Upper quartile
• TMM
• Reference gene (e.g. 16s)

Methods developed for RNA-seq data may not work as reliable on shotgun metagenomic data.

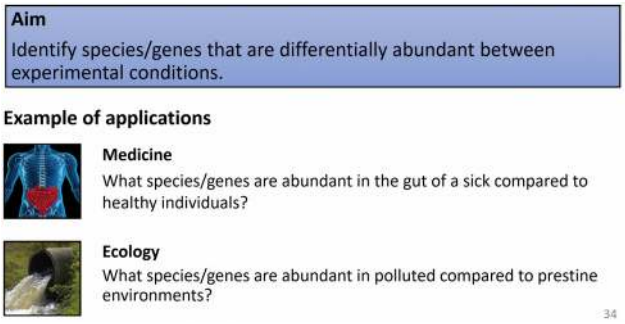
Amplicon and shotgun metagenomic sequencing: Statistical analysis

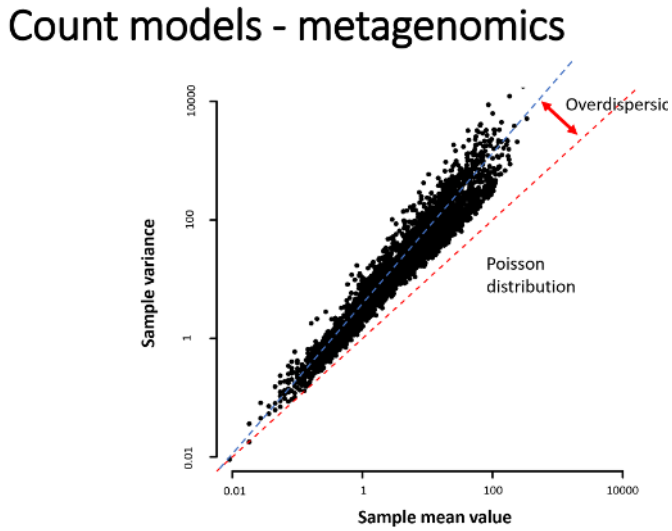
The nature of metagenomic data

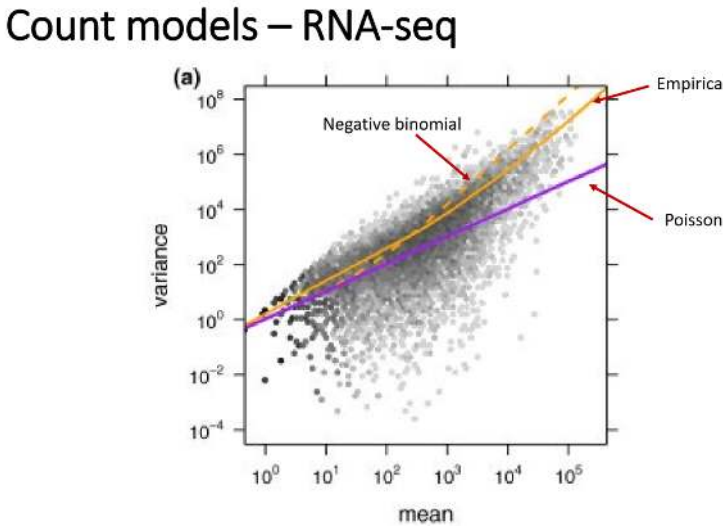
- **High dimensional:** many species/genes present within a single community
- **Few samples:** due to costs associated with sequencing and sample preparation; sequencing depth is prioritized over biological replication

- **Very high variability:** due to randomly selecting DNA fragments + technical noise (DNA estraction + sample preparation, sequencing errors, annotation/binning errors) + biological noise (variation between microbial communities: variation in species composition, variation between genotypes)
    - Variability is typically higher than in RNA-seq data

Comparative analysis of metagenomes







Many methods used in RNA-seq can be applied to shotgun metagenomics and vice versa

EXAM: In metagenomic amplicon sequencing, a specific genomic region ('barcode') is sequenced and analyzed to assess the taxonomic composition of a microbial community.

(a) Provide an overview of the analysis of data from amplicon sequencing. Which are the necessary steps and why are they important? / How does amplicon sequencing work? Describe the process, the necessary bioinformatics analysis and what biological insights amplicon sequencing can provide

The main steps are pre-processing, OTU clustering, OUT annotation and Statistical analysis.

The analysis of amplicon sequencing data includes 1) preprocessing, 2) OTU identification, 3) OTU annotation and 4) statistical analysis.

1. The preprocessing is used to remove and/or trim reads with low quality. Reads with a high number of sequencing errors can otherwise be interpreted as new species.

2. In the OTU identification step, reads are clustered based on similarity to form OTUs which constitutes potential species. A cut-off of 3% sequence similarity is often used and sequences that have a similarity of 97% or more are thus considered to come from the same species. Reads that do not cluster are called singletons. A representative sequence (e.g. the longest sequence) is used to represent each cluster.

3. In the third step, the OTUs are annotated by comparing them (e.g. the representative sequence) against a reference database. The reference database consists of sequences with a known taxonomic affiliation, i.e. we have

some knowledge from which species the sequence come from. Dependent on the match to the reference we can annotate each OTU to various taxonomic ranks. Some OTUs may not have a match and are therefore considered to be unknown.

4. In the final steps, the data is statistically analyzed. This can include comparison of the counts of OTUs between samples (comparative metagenomics) or to investigate the diversity in one or several samples. This steps requires that the data is normalized to make it comparable (e.g. using rarefaction).

(b) One way to analyze amplicon sequence data is to estimate the diversity. What is the difference between richness and evenness and what do they measure? Describe one way to estimate richness and one way to estimate evenness

Richness measures how many species that are present in a sample and evenness how even their abundances are distributed. A sample can have high richness and low evenness and, vice versa, low richness but high evenness. Richness can be estimated by counting the number of different species (OTUs) that are found in a sample. Evenness can be estimated using Shannons index (measure the 'distance' to a uniform distribution). A high Shannon means a more uniform abundance distribution. Another way to measure evenness is to use the Simpson index, which measures the probability that we get the same species if we take two reads. Note that all these diversity indices are influenced by sequencing depth and rarefication of the data is therefore necessary

In short
Richness estimates the number of species present in a community while evenness measures how even their abundance are. Richness can be estimated by counting the number of unique OTUs in sample. Evenness can be estimated by Shannons or Simpsons index.

EXAM: Why do we sometimes need to rarefy the data before estimating the richness?

EXAM: Shotgun metagenomics is used to study bacterial communities by sequencing random fragments of their collective genome (metagenome).

(a) What does binning mean? What is the difference between direct and reference-guided binning? Which approach would you recommend if you have very short sequence reads? Why?(3p)

The purpose of binning is to quantify gene abundances. This is done by placing reads in different 'bins' depending on their gene or gene function. The number of genes in each bin is a measure of the gene abundance.

Direct binning tries to identify genes directly from the reads. This approach is only applicable to reads that are of substantial length so they contain a substantial part of the gene.

Reference-guided binning uses a reference. The approach is similar to RNAseq, where each read is mapped to the reference. By using an annotation of the reference, we can see what gene it contains and, based on the number of matching reads, quantify the abundances of these genes in the metagenome.

If you have very short reads, reference-guided binning should be used.

In short
Direct binning tries to identify genes directly from the reads. Reference-guided binning maps the reads to a reference, which has been annotated for the presence of genes. If you have short reads, the reference-guided binning should be used since it is very hard to identify genes directly from the reads. See lecture notes for full details.

(b) Why do we normalize data from shotgun metagenomics? Also, mention at least one type of systematic error that may be present in metagenomic data but is not present in transcriptomic data.

Shotgun metagenomic data is affected by a wide range of systematic bias that make it hard to compare gene abundances, both between and within samples. Such systematic errors include for example, differences in

sequencing depth and gene length. Normalization is used to remove these biases and thereby reduce the overall data variability. This makes it easier to identify differentially abundant genes.

One type of systematic error present in metagenomic data but typically not in transcriptomic data is differences in taxonomic composition. The proportion of eukaryotes, viruses and bacteria can differ between samples. These differences in taxonomic composition, will then affect the relative abundance of all bacterial genes. If not removed, this can significantly increase the variability thus reducing the statistical power

In short
Data from shotgun metagenomics is affected by a wide range of systematic errors and needs to be normalized before samples can be compared. One type of systematic errors that is present in metagenomic data and not transcriptomic data is differences in the taxonomic composition where e.g. the proportion of bacteria can differ significantly between samples. See lecture notes for full details.

EXAM: In amplicon metagenomic sequencing specific barcode regions (e.g. 16s) are sequenced and used to assess what type of microorganisms that are there ('who is there'?).

(a) Amplicon sequences are often annotated by comparing them to a reference database. However, a large proportion of the amplicon sequences can typically only be annotated to a high taxonomic rank (i.e. above species level) or not be annotated at all. Discuss the reasons why some amplicon sequences are hard or impossible to annotate.

Amplicons are annotated by comparing the representative sequence from their OTU against a reference database. There are several reasons why some amplicons may be hard to analyze

1) The species from which the amplicon comes from is not present in the reference database. In this case we will end up with an imperfect match which means that the amplicon can not be annotated or not annotated down to species-level.
2) There are sequencing errors in the amplicon. In this case, the correct species may be present but due to the errors, no satisfactory match is possible. Many of the amplicons with many errors clusters into singletons and by excluding this, the risk for this scenario is reduced.
3) The 16s region is too similar between the species in this part of the taxonomic tree. Using the 16s region is not perfect and for some species the region is identical or at least very similar. In these cases, it can be hard to say exactly from which species the amplicon comes from. In this scenario, it is typically possible to annotate the sequences to a higher taxonomic level (e.g. genus).

(b) It is also possible to use shotgun metagenomics to investigate what microorganisms that are present in a samples. This can be done by identifying the barcode region (e.g. 16s) in the individual shotgun reads and then analyze such reads similarly to amplicon sequence data. Discuss the advantages and disadvantages with such an approach. Can you see any reason for generating amplicon metagenomic data for samples for which shotgun metagenomic data is already available?

It is possible to identify the 16s region from shotgun metagenomic data. This has both advantages and disadvantages. A main advantage is that we only need to generate one data type, which significantly lowers the cost and the complexity of the experiments. Another advantage is that the shotgun data will cover the entire 16s region – not only the region targeted by the amplicons. There are however several disadvantages. One may disadvantage is that the number of reads form the 16s region will be relatively low. Only a small part of the bacterial genome comes from the 16s region and low-abundant species will therefore be overlooked. Furthermore, if short-read data is used, it may be problematic to identify the 16s region properly (it will often only be partially present in a read). Due to the issues with the limited sequencing depth of individual 16s regions in shotgun data, it is often complemented with 16s amplicon sequencing. This enable us to answer the questions 'who is there' and 'what are they doing' at a sufficiently high sensitivity

EXAM: In gene-centric analysis of shotgun metagenomic data, the abundance of genes are compared between samples from different experimental conditions.

(a) Describe briefly the bioinformatical process of gene quantification ('binning'). What is the difference between direct and guided binning? (3p)

Gene quantification of metagenomic data is done as follows. For guided binning, the sequence reads are aligned to a reference. This can include a collection of microbial genomes, a gene catalogue or contigs that have been assembled from the data. The reference needs to be annotated so the start and end of the genes are known. The alignment is typically done with relatively sensitive aligners to that differences in strains of the organisms in the community are taken into account. Next, the reads are grouped based on what gene they match. This can be done using both at a more higher functional level or at a more specific gene level. The gene abundances is then estimated based on how many reads that ends up in each group. For direct binning, a reference is not used. Instead the genes are identified directly on the reads, which are then binned based on the gene. This requires however reads of sufficient length.

(b) Name at least two statistical approaches that can be used to identify differentially abundant genes. What are their strengths and weaknesses?/ In metagenomics, genes are often quantified by counting reads. This leads to data that does not follow the normal distribution. Describe at least two statistical approaches that does not rely on normal assumptions and that can be used to identify differentially abundant genes in metagenomic count data.

One statistical approach is to model the gene abundance by using discrete overdispersed count models. This approach will describe the nature of the data accurately but a drawback is the complexity of these models and that they need to be fit using numerical methods which can make them slow. Also, modeling the dependence between the expected value and the variance is non-trivial. Another approach is to transform the data to a normal distribu tion and use standard methods such as linear models. This can be done using a variance stabilizing transform such as the square-root and the logarithm. An advantage is that methods based on the normal distribution are plentiful, flexible and fast. A drawback is that the transformation will never be perfect - or even good – for all genes. The normal assumption will therefore be not be valid which results in lower statistical performance. A final option is to use non-parametric methods, which does not rely on an underlying distributional assumption. This makes them robust but, as a consequence, they require more sample to reach a sufficiently high statistical power. Non-parametric methods are also sensitivity to many ties, which may be common in low-abundant count data

EXAM: In metagenomics, shotgun sequencing can be used to investigate the biological function of microbial communitites in a smaple

(a) Provide an brief outline of the anaysis of shotgun metagenomic data.

Data from shotgun analysis is typically analyzed using a gene-centric approach. First, the data is pre-processed. After that, the gene abundances are quantified, a process sometime called 'binning', either by aligning all reads to a reference or search for genes directly on the reads. Gene are quantified based on the number of matching reads. After that, the gene count data is normalized and then analyzed statistically.

(b) Discuss the choice of reference when mapping reads from shotgun meatgenomic data. What options are available? What are their advantages and disadvantages?

The choice of a reference is important and can affect the end result. One option is to create a large databases with as many genome as possible. It can, however, be hard since only a small proportion of all microorganisms have been sequenced to date. If references genomes are not available, it is possible to use a gene catalogue, i.e. collection of genes. These are often very large but can cover more than the reference genome. A disadvantage is that they only cover genes and not other parts of the genome. It is, finally, possible to use a de novo assemble reference, created directly from the metagenomes. A challenge is that this process is computationally heavy and fast from perfect.

© Describe two statistical approaches for the identification of differentially abundant genes. Describe at least one advantage and one disadvantage for each approach.

One method for statistical analysis is to use methods based on normal distributions. These methods are easy to use and very flexible. A drawback is that it is hard to transform the data so that all genes follows approximately normal distribution. Another method is to use count models. An advantage is that these methods describe the discrete nature of the count data and have thus, typically, the highest statistical power. A drawback is that they are generally more complex than methods based on normal distributions and may require significantly more computation to derive results.

EXAM: Resistance to the fluoroquinolone antibiotics is caused by mutations in the gyrase A gene (gyrA). In a research study, researchers aimed to investigate the effect that traveling has on the abundance of mutations in gyrA in the human gut microbiome. The hypothesis was that traveling to countries where antibiotic bactera are more prevalent may increase the abundance of mutated gyrA genes in the gut microbiome.

(a) How would you design the study given that you have unlimited resources? Describe the choice of study design, which subjects to include, what samples to collect, sequencing methodology to use, and the principal steps in the data analysis. (3p)

(b) Which are the main complications in analyzing mutations in a microbiome? How do these complications differ compared to analyzing mutations in the genome of a single organism? (2p)

Since a single human cell contains approximately 1,000 times more DNA than a single bacterial cell (approximately 6 billion bp versus 4-5 million bp), even a low level of human cell contamination within a microbiome sample can substantially complicate the sample processing and sequencing.

EXAM: Relate amplicon sequencing to shotgun sequencing. What kind of information can you extract from data generated by shotgun sequencing that is not present in data generated by amplicon sequencing?/ Discuss shotgun metagenomics in relation to amplicon metagenomics. What are the benefits of using shotgun metagenomics? What are the main challenges?

shotgun data will cover the entire 16s region – not only the region targeted by the amplicons.

Unlike 16S rRNA sequencing, shotgun metagenomic sequencing can read all genomic DNA in a sample, rather than just one specific region of DNA. For microbiome studies, this means that shotgun sequencing can identify and profile bacteria, fungi, viruses and many other types of microorganisms at the same time.

Challenges:

- Requires computer processing power beyond what an ordinary laboratory would possess.
- Can introduce errors in the assembly process: Sequencing errors in combination with repeated regions cause major problems in shotgun sequencing, mainly due to the failure of assembly programs to distinguish single base differences between repeat copies from erroneous base calls.
- Requires a reference genome.
- May not be able to assemble repetitive sequences.

EXAM: When designing an shotgun metagenomics experiments, the number of samples that are going to be sequenced and the sequencing depth per sample needs to be decided. If we assume that we have a fixed budget for the experiment, a large number of included samples will result in a lower sequencing depth and, vice versa, a higher sequencing depth will limit the number of samples that can be included. Discuss these parameters and their importance. What could happen if an experiment includes too few samples? What could happen if a too low sequencing depth is used?

- Sample size: obtaining a desired statistical power
- Sequencing depth: achieve a higher accuracy at a lower cost; increase confidence, minimze the probability of fault results, detect low abundance variants
    - Accuracy
    - Number of detected genes and expression levels
    - Proportion of rare variants
- If sample size is too low, yields inreliable and irrepoducible results
- If sequencing depth is too low we do not get accurat results and might get a lot of fault results and do not detect low abundance variants

EXAM: In metagenomics, sequencing of the 16s rRNA gene is used to estimate the taxonomic diversity in a sample. (a) What is a 'hypervariable' region and why are they important when parts of the 16s rRNA gene are sequenced using short read platforms? (2p)

- Hypervariable regions, portions in the genome or proteome of a species with much higher levels of variation than other similar areas, are found in all kinds of organisms, from viruses to higher eukaryotes. They are usually associated with important functions related to interactions with other organisms.
- Species-specific sequences within a given hypervariable region constitute useful targets for diagnostic assays and other scientific investigations.
- These approaches are usually only able to query short DNA sequences; therefore, it is important to identify the regions within the target gene that supply the most taxonomic information in the smallest stretch of nucleotides. Additional benefits of small amplicon size may include increased assay sensitivity and applicability to archival specimens

(b) Provide an overview of the bioinformatical analysis necessary to estimate the taxonomic diversity from 16s amplicon data. What steps are essential? (3p)

RNA gene sequencing/Amplicon sequencing

- DNA isolation.
- Heating to separate the strands and specific primers.
- Primer extension with DNA polymerase.
- Repeat the above steps to obtain multiple copies of the 16S rRNA gene.
- Run agarose gel, checking for the accurately sized product.
- Purification and sequencing of PCR product.

EXAM: In shotgun metagenomics, DNA fragments are randomly selected from a microbial community and then sequenced. The sequencing is done with either short or long read sequencing techniques.

(a) Do you see any advantages of using long reads instead of short reads? Do you see any disadvantages? (2p)

- **Long reads:** do not cover complete chromosomes, but typically often longer than 10,000 bases -> no PCR amplification necessary
    - Advantage
        - No DNA/PCR amplification step – sequencing is done of individual DNA molecules

        - Real time and single molecules
        - Long sequence reads
        - Long reads can generate more accurate assemblies than short-read technologies, especially when there is no reference genome to check against (de novo assembly) or in repetitive sections of the genome and regions with complex genetic rearrangements
    - Disadvantage
        - Lower throughput, relatively expensive
        - Error prone
        - Requires DNA of very high quality
        - A downside to long-read sequencing is that the accuracy per read can be much lower than that of short-read sequencing

(b) Would you analyze shotgun metagenomics data from short and long reads the same way? If yes, describe why. If no, specify how you would make the analysis different. (2p)

Reference-guided binning -> shorter reads
Direct binning of metagenomic reads -> long reads

(c) It is possible to generate both short and long read sequencing data from the same sample. Discuss how shotgun metagenomics can be improved by combining short and long read data. (4p)

Our results provide an effective strategy for combining long and short-read DNA sequencing data to explore and distill the maximum information out of soil metagenomics.
By taking advantage of the strengths of both long and short reads, hybrid-read assembly with StringTie is more accurate than long-read only or short-read only assembly, and on some datasets it can more than double the number of correctly assembled transcripts, while obtaining substantially higher precision than the long-read data assembly alone

EXAM: Gene set enrichment analysis is a bioinformatical concept to improve the biological interpretation of gene lists generated in transcriptomics and metagenomics.

(a) Explain the concept of integrative analysis. What is require to perform integrative analysis and what insights can it provide? (2p)

(b) Assume that you have performed a transcriptomic analysis of 20,000 genes of which 1,000 were found to be significant (had a false discovery rate less than 0.05). Assume further that you are interested to assess the overrepresentation of the Gene Ontologyterm 'response to stress' (GO:0006950). Of the 20,000 measured genes, 750 are annotated with the 'response to stress' GO-term, of which 250 were significant. Describe how you can use a hypergeometric test (also known as the Fisher's exact test) to statistically assess whether 'response to stress' is overrepresented among the significant genes. Note: You do not need to calculate the p-value, only describe how you do it. (3p)