# Applied Genomics

## Intro

It is important to understand that DNA is in living creatures, in cells, so if you want to access DNA, you need to plan experiments accordingly. It is important to keep in mind how data is created.

First part (Fontanesi) of the course will be mainly theoretical, second part (Bovo) will be mainly practical. 16 lectures.

Final exams: 3 parts:

1. Written exams with 25 close answers and 5 short answers. Auto evaluation test will be given to students before written exam.
2. Preparation of a novel genomic project. Novel means that nobody did it before. (you have to look a lot of literature). Genomic project will be given to the prof the same day of the written test. In material and methods there will be a breakdown of the costs of the project.
3. Interview mainly on the project submitted and two questions according to the main objectives.

## Genomic project

There is a virtual budget that will be randomly assigned (needs to be respected more or less). The students select the desired organism, but it needs to be complex (eukaryotic). Budget should be spent for everything concerning the project, including money spent for reaching the desired location (ex sahara desert). Human studies are more expensive. Important to “sell” the project” and make sense of it. To produce data you need machines, (NGS etSc.) so there is a cost associated with their use (however machines will not be bought). A budget balance should be provided to demonstrate that all money is spent. Important to what could be the results. Develop the idea for the genomic project along the lectures.

Important to give the exam as soon as possible because professors will be better and score higher and exam less difficult.

In slides list of topics for master thesis (12 founded projects).

## Foundational concepts in genetics

Genetics is the science of biological diversity. Genomics is the tool that gives the geneticists the possibility of studying genetics. We will also be studying comparative genetics.

4 main branches of genetics:

1. Classical genetics (transmission of genetics or formal genetics)
2. Molecular genetics
3. Population genetics
4. Quantitative genetics

The projects will be probably focusing on population genetics and quantitative genetics.

**Classical genetics**

The fundamental approach of transmission genetics is the genetic cross (the breeding of two selected individuals and understand how traits are passed through the generation).

It is important to know the different systems of sex determination in animals when studying different organism to construct pedigrees.

The building of genetic maps based on frequencies of recombinants are sometimes still used to validate genome assembly.

**Molecular genetics**

Molecular genetics techniques like PCR can generate errors, and when collecting data for genomic work we have always to consider the error rate.

**Population genetics**

Studies genetic composition of groups of individuals and how it changes in time and geography. It is basically the study of evolution. Population genetics founders are Hardy and Weinberg who created the rules for Hardy-Weinberg equilibrium (slides)

**Quantitative genetics**

Also referred as the genetics of complex traits, that are traits that are defined by a large number of genes and have a continuous distribution in the population. Also environmental factors influence this traits, so it is very complicated. Quantitative genetics determine the level of heritability of genetic traits.

## NGS technologies

For NGS we have different technologies:

1. ThermoFisher (IonTorrent)
2. Illumina
3. Roche (roche its dying)
4. PacBio
5. Oxford Nanopore

**Ion Torrent technology**

Ion Torrent technology: this tech produces H+ ions during the sequencing and the machine senses difference of pH. The chip for this technology is very compact, and the size of the chip sets the limit for the sequencing. With the biggest chip you can sequence 25 Gb (giga bases), and the biggest can read up to 100-130 M reads.

Ion torrent technology works like this: the chip is constructed using nanowells, which are the points in which the reaction occurs. The DNA is allocated in a rigid sphere that hosts more molecules of the same sequence, and in one there are only one type of sequence. We put one nucleotide at a time in the well, and when the voltage changes, we know we put the right nucleotide. After we put one nucleotide in the well, we wash the well to remove non reacted nucleotides.

**Ion Torrent Workflow**

First of all I need to get DNA and fragment it to obtain sequences of about 200-300 bp. Then some adaptors are put at one end of the sequences. Those adapters are used to couple the sequence with a known primer. The amplification takes place in an emulsified PCR setting (each droplet is considered as a well containing one DNA fragment) (each droplet needs to contain equimolar quantity of dNTPs, polymerase ecc.) then, a microsphere is used with a specific sequence that can bind to the adapter that we put on the sequence at the beginning. Now you can have bubble with one sphere, bubbles with no spheres, bubble with one sphere but two different sequences, or bubbles with more than one sphere.

More than one DNA can be sequenced in one setting. This is possible because when generating the library, we can put a “barcode tag” consisting of a known sequence that I can retrieve later.

After the sequencing primer there is the key sequence. It consists of 4 bases that are used as quality indicator for the sequence. This sequencing produces an ionogram, and if the ionogram of the key sequence is not readable, then the sequencing is discarded. The height of bars in the ionogram is related to the number of nucleotides incorporated during a flow. After the key sequence there is a blank space in the ionogram.

After the sequencing there is the data analysis. The Torrent Suite software gives you preliminary information on the wells used and the quality of the run (polyclonal and monoclonal wells), and the length of the reads. If the barcode is not readable you discard the read. Before data analysis you have to remove the tail and primer sequence. To assemble your sequence, usually you align it to a reference genome. After the mapping, you obtain the FASTQ file, which contains all reads and quality score. Then there is the BAM file, which also contains alignment data, it contains an index to easily access data. Then there is the VCF file, which contains mutation (SNPs and structural) information.

The reads are filtered based on different parameters: length (small ones are removed), quality (ex polyclonal), but for some application you want to keep small reads.

FASTA format just contains sequence, FASTAQ also contains quality information. Q score is the score used for quality. Higher Q scores indicate a smaller probability of error, low Q score indicate that a significant portion of the reads are unusable. A score of 20 represent n error rate of 1 in 100 (1 base wrong for 100 bases). Q score is computed: , where e is the probability of the base being wrong.

Ion torrent has an average quality score of 20 (not optimal).

**454 Roche**

It’s the “mother” of the ion torrent technology. What is detected is not the ion H+ but the pyrophosphate, which is coupled with Sulfurylase and Luciferase (which also uses ATP) and light is produced. This technology is no longer used because Ion Torrent is much cheaper (light systems are usually expensive). This technology uses larger chips with larger wells, and there are less wells in one chip. You need more copies of DNA per well to have a big enough signal to be detected.

**ABI SOLID**

It’s a type of sequencing based on ligation. This technology is almost dead. Its main limitation is that reads are very short, around 50 bp.

The library prep offers two options: fragment libraries or single DNA libraries. With fragment libraries two sequences of DNA that are a known distance apart are ligated together and they compose one template. Fragment libraries allow paired end sequencing.

In both libraries adapters are ligated at the end of the target sequence (in fragment libraries an adapter is also put in between the two target DNAs). DNA sequences are then amplified on beads, and the beads are positioned in glass slides. Slides are of 1 , 4 , 8 formats and barcoding is an option to maximize the number of samples analyzed.

The system works with the use of probes constituted of 2 bases, a number of universal bases that bound to every sequence, and a number of universal bases that are bounded with fluorescent dye, and that are cleaved every cycle releasing the dye. There are 4 dyes for 16 possible nucleotide couples. A primer is attached to the adapter, and the probes anneal to the sequence and when ligation occurs, the dye is released. When the sequence is completed, another primer binds to the adapter, but this time the bases are shifted in respect to the first primer by 1. This system is repeated 5 times (5 primers) and sequencing is possible. This system ensures unmatched accuracy, but is also the most expensive.

**Illumina**

This technology is constantly improving and is the dominant technology. The steps of the workflow are similar to that of IonTorrent. You have to prepare the library and add some sequences. Then there is cluster generation, which is used to obtain a lot of DNA of the same sequence, then there is sequencing and data analysis.

The library prep consists of your sequence, then the primer is added on both sides of the sequence, then there is a capture sequence (on both sides), which is a sequence used to anchor the DNA to the wells in the machine. Also, you can add to the sequence the barcode sequence.

Cluster generation occurs in a flow cell. They are generated thanks to an amplification reaction. The capture sequence can be “captured” by an adapter anchored to the bottom of the flow cell. The complementary strand is discarded after the strand connected to the adapter is synthesized. Clusters are a group of DNA strands positioned in a 1-2 micron spot. Then there is bridge amplification, in which the DNA strand binds to a different adapter in the flow cell, and another strand is synthetized. This operation is done many times to obtain a lot of DNA. The bridges are then denatured and reverse strands are removed.

Clusters can be randomly positioned in the flow cell or, in the newest technology, ordered (patterned flow cell). In the first case analysis is more complicated, and some clusters can overlap so the signal cannot be read easily by the machine. The patterned flow cell has some wells, and each well contains one cluster.

Every time a nucleotide is incorporated, light is produced based on the type of nucleotide. Each type of nucleotide has a different fluorescent molecule and terminator attached to it (this prevents the adding of two identical nucleotide one after the another, like could happen in ion torrent technology). During each cycle you add all 4 nucleotides, and of course just one is incorporated. Then the photo is taken to look at the colour of the nucleotide, then the terminator and the fluorophore are cleaved.

An illumina read is approximately 200bp long

To make the process cheaper, “less” colour sequencing was invented.

**Two colour sequencing**

With this system, A has both the green and red fluorophores, G has no fluorophore, and C and T have red and green respectively.

**Single colour sequencing**

With this system, G and C have no colour. While A and T have colour, but also, C has a tail that can attach the colour, while A has a cleavable fluorophore. A first image is taken, then a chemistry step either attaches the fluorophore to C or cleaves fluorophore to A and does nothing to G and T. After this chemistry step, another image is taken to finally discriminate between bases.

This different “less” colour systems use nucleotides that are cheaper to produce and has the same quality of reads than the four colours sequencing.

**Paired end sequencing**

With this technique is possible to sequence the DNA from its two ends. Each strand gets sequenced by its two ends, so the advantage is that you get a read that is “double”. If the left and right reads overlap, then it makes assembling the genome easier, if they don’t overlap, knowing the two limits of a read makes sequence alignment easier, since you know both end of the read and you also know that there is sequence between them. This unknown sequence can be retrieved from the reference genome.

Paired end sequencing sequences have two different adapters (one for each end), and also the flowcell has a mixture of the two complementary adapters, this allows the sequencing on both sides. (slide 69 Lecture 3) A graph showing the amount of data

Description automatically generated with medium confidence

**Illumina applications**

It is important to know all the different applications and the best machine for doing each application 🡪 for the genomic project. Machines with limited throughput cannot be used for all applications.

**Oxford nanopore technology – long reads**

Oxford nanopore reads can be up to 10s of kb, on average they are 5-10kb. An important thing to consider is that is crucial to avoid fragmentation of DNA during it’s isolation. This limits the possible samples to use. For example, you cannot sequence ancient DNA 🡪 it is crucial the preservation state of your sample.

With this technology, DNA amplification is not required nor needed, and is also possible to sequence directly RNA fragments.

The system consists of a system that moves the single strand DNA through the nanopore, and the nanopore (made of protein) is connected to a system that measures the electric current that moves through the pore. As the DNA passes, electric current is modified (decreases). Different groups of nucleotides produce different modifications, while it is not possible to measure the current produced by a single nucleotide. Since it is quite difficult to translate the modification of current to single nucleotides, the error rate is quite high. Also the DNA moves trough the pore and is “unzipped” thanks to a motor protein

The sequencer itself is very small, around the size of a cellular phone. The chip contains the membrane with the nanopore.

The advantage of this technology is that is portable, relatively cheap, can be used outside (if you have power and can prepare the library). It can be used on the spot in respect to the Nagoya agreement.

**PacBio**

This sequencing is called the SMRT (single molecule real time) sequencing.

This tech uses modified nucleotides that have a tail of 6 phosphates and a fluorophore, while the 3’ is left untouched. The system consists of wells at the bottom of which there is attached a polymerase, and when a nucleotide is attached to the growing sequence, light is produced for each nucleotide and detected by a detector above the wells. The error rate was quite high, because the polymerase can add 1000 bp/s but we don’t have a technology that can detect light that fast. So what they did was invent the CCS (circular consensus sequence) 🡪 the target DNA is made circular and light is detected only on some “spots” of the sequence, and since the molecule is circular a single spot is sequenced many times reducing the error rate. Also, it can detect modifications on DNA because when the polymerase encounters a modified base, it pauses, and this can be detected.

**CCS**

Preparation of DNA: the double stranded DNA is modified by attaching adapters that make the sequence circular. Since the sequence used as template is circular, it can be sequenced many times reducing the error rate so much that now its one of the most precise technologies. The reads are >20 Kb.

**Why long reads:**

* Pros: easier to assemble genomes --- identify structural variation --- identify phase variations: which variants are on which chromosomes --- sequencing of GC rich regions --- assembly of repetitive regions
* Cons: cost --- DNA preparation

60 Gb (illumina) 🡪 500$

60 Gb is the limit of having a good coverage for the human genome which is 3 Gb (at least 20 times 🡪 20X).

**Budget assignment / projects**

MY BUDGET IS 🡪 30’000

With a big budget you can do studies on humans, because human genomic data is more expensive. This is because its trickier to control the environmental factors. It is hard to standardize the data, so you need a lot of samples (the bigger the sample number the more costly the study is). A trick to save money in studies is to get data from individuals of the extreme tails of the phenotypic normal distribution (for a quantitative trait). Also, in humans almost everything has already been done 🡪 so its difficult to find the topic. In other organism its easier to think of a novel project.

DNA pooling: this can be done to cut costs, but is important to get equimolar DNA from all participants.

De novo sequencing: it’s an option as a project 🡪 not for me because I have too few money.

## QC, coverage, variants, example workflow for data analysis

Before analysing the data, its important to know how your data was produced, with which technology, how were the samples taken, also if multiple types of sequences from different organisms were sequenced together.

**DNA Quality**

First of all, it is important to assess the purity and quality of DNA: the 260 nm/280 nm ratio is used. This is a ratio between absorbance values. DNA has an absorbance value of 260 nm, while proteins have 280 nm, so this ratio tells you how much of your sample is consists of proteins. The higher the ratio the better, usually it is accepted as pure DNA a ratio of 1.8.

Abnormal 260/280 ratios derive from contamination of proteins or reagents of extraction like phenol.

Another ratio used is the 260 nm/230 nm. Expected values are in the range of 2 – 2.2. This is used in particular when handling DNA from plants, as their DNA samples are often contaminated with carbohydrates and this lowers the 260/230 ratio.

To asses the integrity of DNA, usually gel electrophoresis is performed.

When applying for NGS it is important to make sure that the DNA meets the requirements of quality, concentration, volume of the sample, total quantity of DNA, spectrophotometric ratios.

The company that accepts your samples, usually performs some quality assessments and gives you back the results.

DNA quality 🡪 260/280 ratio and 260/230 ratio

DNA integrity 🡪 electrophoresis

**Coverage**

When you give the company your samples, they ask you how many reads/coverage you want. You can calculate how many bp you will obtain with a given coverage (because you know the length of the genome of the organism you want to sequence). You can use your computed number to do a rapid check when you obtain the results.

Depth of coverage: the average depth of sequencing coverage can be estimated with the formula: LN / G where L is the read length, N is the number of reads, G is the haploid genome length (all measures in bp). 🡪 5X means that your DNA was sequenced 5 times. Depending on the project, you may need different coverage. For example, for investigating copy number variants a low coverage is enough, for study of SNPs, high coverage is needed.

Another measure is breadth of coverage which is the percentage of target bases that have been sequenced for a given number of times. 🡪 if all of your target bases were sequenced 5X and your target of coverage was 5X, then the breadth of coverage was 100%.

For sequencing for the first time a genome, you have to use high depth of coverage, and also Illumina is not the best tech. You probably want a long read technology.

Each technology can give you different outputs 🡪 slide 17 lecture 4

**Variant Discovery**

There are many different types of variants.

**Sequence variants**

* SNPs (single nucleotide polymorphism 🡪 a mutation with a frequence in the population of at least 1% )
* Insertions
* Deletions
* Indels
* Substitution (an alteration where the nucleotides in a sequence are changed but the length of the sequence remains the same)

**Structural variants**

* CNVs (variation in the number of copies of a region in the genome)
* Inversion
* Translocation
* Other structural variants (fusions of chromosomes and/or genes)

**Sequence ontology**

Relationships between sequences depicted in a non cyclic graph and the relationship are directed (example 🡪 slide 25 lecture 4). Ontologies are very useful, for example to find the function of a protein using similar proteins with known function. A lot of machine learning includes gene ontologies.

C>G Mutation is referred as S in Sanger electroferogram 🡪 there is specific nomenclature for every mutation.

**Data analysis pipeline**

The first file you obtain after sequencing is a FASTQ file, that contains all the read and the quality of each base. After alignment of reads over the reference genome you obtain a BAM file. A BAM file displays the alignment. It is the compressed binary file of a regular text file that is the SAM file, but this is usually not used since it is too heavy. The alignment (and only the alignment, NOT variant calling) is performed by different algorithms, the two most used are BWA and Bowtie, which are based on different mathematical concepts: Burrows-Wheeler Transform for the BWA and Hash tables for Bowtie. Each of this different two tools has the ability to find different variations, so variation that the BWA tool finds, might not be found by Bowtie and vice versa. After the alignment the “search” for variants begins (variant calling), and the vcf file is created, which is a specific text file format that contains information about the different variants found, and after that the last step is variant annotation. This is the most common pipeline for variant discovery 🡪 slide 31 lecture 4. For each of this steps, tons of tools have been developed.

Problems: shorter reads make it difficult to find their location in the genome. Also, the amount of data makes it challenging for the computers, so we have to compromise speed with precision.

When assembling a de novo genome, sometimes mistakes are made, and reads are put in the wrong place. So happens that when someone uses your assembled genome as reference, some errors are shown. To decide if the error is in the reference genome or in the newly assembled one, we can consult the quality score of the alignment. There is in fact a Mapping Quality (MQ) score computed for each alignment.

**Standardized pipelines:**

It is becoming increasingly common in the field to develop a pipeline, provide the environment for the data analysis with all the software. One example is the GATK pipeline (slides), that can be standardized for many different studies, and so it can easily be replicated. The idea is that if everybody uses the same software, variant discovery is independent of the software used.

## Sequencing file formats and pre-processing software

**FastQ**: this format provides the sequence of the reads and their quality. It consists of 4 lines of text:

1. The first line begins with @ character and is followed by the sequence identifier (which needs to be unique)
2. Raw sequence letters
3. This line starts with a+ character and is followed by the same sequence identifier
4. Encodes the quality by using an ASCII character, each character specifies a quality score based on the PHRED score. The quality score for a specific nucleotide is aligned to the nucleotide letter.

**Paired-end sequencing:**

In paired end sequencing, each read is sequenced two times: 1 time from one direction and 1 time in the second direction. So when analysing the reads, those two reads will have the same name and at the end /1 or /2. It is very important to consider those reads as part of the same read.

**Pre-processing:** in this phase of a project, we check the quality of the reads and perform a quality control and removal of bad quality reads. This procedure is done with tools, one of this is FASTQC. For every tool is important to know the requirements needed in terms of software and hardware.

FASTQC provides a modular set of analyses. The analysis result is a graph and the possible scores are green – orange – red (good to bad).

The first module is Basic Statistics, this file contains the name of the fastq file, file type, encoding, total sequences, sequence length (length of the reads) , the % of GC and per base sequence quality.

Per base sequence quality is one of the most important metrics. It consists of a boxplot graph that shows aggregated quality score (PHRED) statistics of each position along all reads in the file. The average is represented with a line in the middle of the boxes. Usually the last part of the reads have lower quality (very true for older technologies). Another important graph is the quality of read distribution 🡪 this is used to decide whether to trim some reads.

Usually at the beginning and middle of sequencing, the percentage of GC bases is constant, but nearing the end of the length of the read, the GC content starts fluctuating because the sequencing is not reliable anymore. There is also the GC distribution over all sequences. This distribution should be bell shaped. An usually shaped distribution could indicate a contaminated library, because different organisms have different GC content percentages.

Another important module is the sequence duplication level analysis 🡪 tells you how many times a specific sequence was sequenced. To avoid duplicated DNA, a strategy is to use a PCR-Free approach, but this is only possible if you can obtain a lot of DNA.

**Duplicated reads:**

Duplicate reads are usually removed. They come from PCR, and they fake high coverage. Duplicated reads can also trick the user to think that a specific SNP is present, however when removing duplicate reads you realize that that SNP was present only on one read, while different reads that cover the same spot don’t have such SNP.

**Trimming bases:**

One strategy is to trim the bases of the end of the sequence until you encounter a base with a defined Q score. Another strategy is to define a window (a subsample of given length of bases in the sequence) and trim the bases if the average Q score of the bases in the window is below a given value. The problem is to set the window size, a big window size can lead to removal of good quality bases.

**Homopolimeric regions**

Homopolimeric regions are tough to sequence, and sometimes we can find there an SNP, however that may be likely the result of poor sequencing, and maybe that specific base had a low phred score. 🡪 if we have AAAAA in reference, and AAATA in the sequencing, maybe that T is not a real T.

**Prinseq:**

Another important software for quality control is Prinseq, a perl based free software.

**SAM:**

has a header section and information about alignment. In the header there is information about the genome used as reference. SN stands for “reference name”, LN stands for “reference length”, @SQ stands for “chromosome of the reference” and there is one SQ for each chromosome.

In the SAM file is also reported the command line command that has been used to generate the file, and this retains information of the alignment tool used, it starts with CL.

The alignment section comes next and consists of 11 fields. The ref field indicates the chromosome, the next field indicates the start position of the sequence in the chromosome. (slide 73 lecture 4-5-6).

An interesting field is the flag field. It contains information about the read, for example if the read is duplicated, if the read is paired end, or if the read is unmapped: an unmapped read can occur from contamination from other organisms, the organism you sequenced was sick with a virus or bacteria 🡪 in this case you can recycle your data and add new information, for example that your animals are sick. This is metagenomics.

There are tools that have been developed in order to obtain viruses DNAs starting from DNA of the host organism.

There are a lot of SAM flags and there are websites used for retrieving flags and their meaning.

**CIGAR field**

Another important field is the CIGAR field, which consists of a string and each element has a meaning:

Example: 4S8M2I4M1D3M 🡪 4S means that the first 4 nucleotides don’t match with the reference. Those S nucleotides have been effectively deleted from the read in order to produce the alignment. We say that the nucleotides have been “soft-clipped”. M means that the match is perfect, it also says M when there is one mismatch. This case is different from S because the alignment is “forced” while in the S case the bases are ignored 🡪 so the cigar cannot be used for identifying SNPs. I means insertion, D means deletion.

## Variant Calling:

Variant calling depends on different parameters:

* Base quality
* Proximity to indels or homopolimeric regions
* Mapping qualities of the reads supporting the SNP (increased read length or pair end sequencing)
* Sequencing depth (coverage)
* Individual vs multi-sample calling.

Multi-sample calling requires a lot of memory and space, the GATK pipeline released a way to make multi sample calling less demanding: if you have a variant called in one group, you can add a group of sequenced individuals and see if the variant is in their genome or not.

**VCF File:**

It has an header and data, it has different fields and the header describes what is in the field (slide 85 lecture 4-5-6).

Each variant has an identifier (RS identifier) that has been standardized and retrievable in databases, one of those is dbSNP.

The FILTER field is used to tell you if the variant passed the filters you decided to use. For example you only wanted variants with a quality of at least 20. The software finds you the variants that also have a quality below that threshold, but the FILTER field is of course different. It is useful in case you change your mind and want to retrieve all variants.

FORMAT column: here is stored information about the genotype of the organism, for example homozygous for the variant or reference allele, or heterozygous. 0 = reference, 1 = first alternative allele , 2 = second alternative allele.

**IGV Tool:**

It lets you visualize the result of alignment and variant calling: first you inport the reference genome, and then you visualize the reads assembled to the reference genome and the SNPs called on the reads. You can also see deletions, insertions etc. It lets you see the depth of sequencing of the region you are inspecting, and by looking at the depth you can determine whether the deletion or insertion is homozygous or heterozygous.

## Genome Assembly and Annotation

Genome Assembly is the practice of assembling a de-novo genome and is very computationally demanding. Genome annotation is done after genome assembly and is the process of inferring structure and function to assembled sequences.

**Genome Assembly**

Historically the way of assembling genome was the Hierarchical shotgun sequencing.

1. BAC library creation
2. Cloning
3. Assembling by shotgun sequencing (the sequences inside the bacs are fragmented and sequenced, and at the end aligned and the genome was assembled)

**How assembly is done**

First of all you need to select an individual representative of the species. A lot of DNA is needed. Also, RNA can be very helpful for genome annotation (using RNA-seq) 🡪 if there is the corresponding RNA, then that piece of DNA is a gene. Is important also the choice of the sequencing technology, because each technology comes with its own assembly tools 🡪 this determines what kind of computational resources you will need.

After assembling the genome, you have to perform quality checks, to see if your assembled genome is good. You will always end up with gaps in the genome, so you have to do gap filling at one point.

**Things to know before assembly**

Important things to consider are:

* Genome size
* Ploidity level
* Heterozygosity
* Repeated sequences

Estimating DNA size:

Doing a genome assembly you need a lot of coverage (>60x). If you don’t know the size of your genome, you can estimate it by using flow cytometry studies (c-value). C-value is the mass in picograms of a haploid set of chromosomes. 1 pg = 978 Mb.

An alternative to flow cytometry to determine the size of your sequence is K-mer frequency in your sequenced genome (not yet assembled). A k-mer is a unit of subsequence of DNA consisting of k-length. Ex: 3-mer CTG. To each k-mer can be assigned a value for coverage depending on the number of times it occurs in your sequenced genome. K-mers can be arranged in a curve that will have a normal distribution. K-mers need to be long enough to map uniquely in the genome. However, you can have k-mers that never map to the genome, and k-mers that map multiple times, and this may often be because they are part of repeated sequences.

Problem of repeats:

Second generation sequencing machines struggle with repeated sequences, this is because they produce smaller reads. Third generation sequencing machines don’t have this problem because they produce longer reads.

Information about heterozygosity:

You can have two reads from the same part of the genome that have different sequence due to different alleles. If the difference is only 1 base, then there is no big problem, but the difference between alleles consists of multiple bases, the assembly algorithm can make error and lead to read discard (final assembly is haploid). Large population sizes tend to have more heterozygosity, so in this case is recommended to sequence inbred individuals. Diploid tissues are manageable, while tetraploidity and above are very problematic and result in fragmented assembly.

GC content:

GC content can be a problem for Illumina sequencing, and result in low or no coverage. This can be compensated by using third generation technology or an increased coverage.

DNA quality:

For a de-novo sequencing third generation technologies are preferred because they produce long reads, but in order to have long reads we need to have high quality DNA.

When few cells are available, PCR is needed and this can lead to artifacts and uneven coverage.

Is also important to minimize contaminants like mitochondrial DNA, chloroplasts DNA, pathogens and symbionts.

Strategies:

You can decide to use different technologies: short reads, long reads (better than short reads) or also an hybrid approach, using both short reads and long reads.

End of assembly:

Regardless of the approach used, you will end up with gaps. To fix those, there are different technologies available to improve contiguity.

* Optical mapping methods (BioNano)
* Linked-read technologies (10X Genomics Chromium system)
* Chromatin association (Hi-C)

Computational resources:

To do an assembly of a diploid genome of about 1Gb, you need at least 96 CPUs and 1 TB of RAM, along with 3 TB of local storage and 10 TB of shared storage 🡪 all of this requires working in a server and you have to pay for the time you use.

Algorithms:

* Greedy algorithms
* overlap layout consensus (OLC)
* string graph
* de Bruijn graph (DBG)
* hybrid algorithm

Graph Theory:

A graph is a set of elements linked together. We have vertices (V) that are elements linked together, then we have edges (E) that are the links. Also, directions of those linkages are also important. Graphs that do not have orientations are called undirected graphs. Node’s indegree are the number of incoming edges, nodes outdegree is the number of outcoming edges.

A node is balanced if indegree equals outdegree, or semi balanced If the difference is only one.

In genome assembly, graphs are used to link the different reads 🡪 a read is connected to the other unidirectionally if the end of the first read is equal to the prefix of the second read.

Greedy algorithm:

They take 1 read and compare it to all the others. Those are no longer used for NGS data analysis.

Brujin graph:

This is the most used algorithm because it uses less computational resources and is faster.

This is an algorithm that uses k-mers as substrings of reads. A de Bruijn graph is a graph whose vertices correspond to k-1-mers and edges represent k-mers.

Example (k =3)

ATGCG → 3-mers → ATG, TGC, GCG

GCGTG→ 3-mers → GCG, CGT, GTG

GTGGC→ 3-mers → GTG, TGG, GGC

TGGCA→ 3-mers → TGG, GGC, GCA

To form the graph we only consider the (k-1)-mers, and in particular the distinct ones. The graph is built using vertices that correspond to (k-1)-mers (and since those are unique, there are unique vertices), edges rapresent k-mers, and have direction. We then want to find a path that visits one edge just one time (but can visit nodes more than one time) 🡪 this is done with additional graph theory and is called Eulerian walk.

**How to construct a De Bruijn graph**

<https://www.youtube.com/watch?v=TNYZZKrjCSk>

<https://www.youtube.com/watch?v=FCDJIx-W7C8> 🡨 when Eulerian path is not unique

<https://www.youtube.com/watch?v=f5kgmqcwb8M> 🡨 practical building of De Bruijn graph

In real life some bases have errors and this leads to fake k-mers. If we choose low length k-mers, we put more error containing k-mers in the graph, while if we use longer k-mers maybe only one k-mer has the error. Also reducing the size of the k-mers leads to more complexity and possibility of errors, so it is better to use longer k-mers. With small k-mers it is more probable to have more than one Eulerian walk (slide 56 lecture 6).

A problem that may arise is that sometimes an Eulerian path cannot be identified, because you have a branched structure due to repetitive regions in the genome. This regions can be studied by building a second library using paired-end sequencing or mate pair-end🡪 but of course is more costly. In real life, k-mers used are around 30-40 bases (for short reads technology).

Building longer reads with Illumina – Mate Pair-end sequencing:

It is a variant of standard paired end sequencing, but this time instead we generate long insert paired end DNA libraries with fragments up to 15Kb. Large inserts span over regions that are problematic to assemble, such as repetitive elements. First of all we select the regions that we want to sequence and biotinylate their ends, then we circularize the fragment and digest away the non-circularized DNA. Then we fragment the circularized DNA and only keep the fragments that contain the biotin to form the library. 🡪 then we perform Illumina sequencing. In this way we know that between the two pair end reads, there is a big insert of N Kb and we can map the reads accordingly.

**Assembly quality and gap-filling**

To check assembly quality there are different parameters:

* N50
* Completeness

N50 is the length of the smallest contig after they have been ranked from longest to shortest, such that the sum of contig lengths up to it covers 50% of the total size of all contigs.

* The N50 value is the length of the last and hence shortest contig whose addition makes the combined length greater than 50% of the total (slide 68 lecture 6).

The N50 in general is a metric used for assessing an assembly’s contiguity.

Scaffolding and gap-filling:

In scaffolding, assembled contigs are stitched together based on information from paired short reads. The unknown sequence between the contigs will be filled with Ns. If matching reads are instead used to join contigs together, for example long reads, actual sequence will fill in the gaps, and this is referred to as gap filling. So to fill the gaps, we can perform a second sequencing using long reads based technology or mate pair end sequencing.

There are also supporting technologies, most of which are used to improve the contiguity of genome assemblies (the scaffold N50 length by at least five fold). These include:

* optical mapping methods (e.g., BioNano).
* linked-read technologies (e.g., 10X Genomics Chromium system)
* chromatin association (Hi-C)

Scaffolding and gap-filling using genetic linkage maps:

We can use genetic linkage maps to know if two contigs (for example) are very far away ore close to each other.

**Statistics for genome annotation**

Statistics: it is important for genome annotation to have a good genome 🡪 how good a genome is can be checked by looking at the N50. A good N50 should be around the medial length of a gene. This means that 50% of the genes will be contained in a single contig.

In general, bigger genomes have bigger genes. If the size of genome is known, then is possible to obtain a rough estimate of gene lengths and estimate a minimum N50.

Another important statistics is the average gap size of the scaffold.

Also of course genome coverage should be 90-95% 🡪 cannot be 100% due to repeats (to estimate how big the genome is we can use c-value) and gene coverage should be complete.

BUSCO: a tool used to check genome completeness is BUSCO, a software that has a set of genes and searches those genes in the assembly. In this set of genes there are genes that have been recorded in all species and in related species, only one time per genome 🡪 if BUSCO finds those genes more than one time, then there are errors in the assembly, and if BUSCO doesn’t find one or more of those genes, then it means that the assembly is not good.

**Genome annotation**

Annotate means to give one or more roles to a sequence in the genome.

Steps are:

1. Annotation of repetitive elements (and masking)
2. Structural annotation 🡪 annotation of genes
3. Functional annotation 🡪 annotation of gene functions
4. Data submission, maintenance, refinements, updates ecc.

**Annotation of repetitive elements**

Transposable elements: transposable elements have recurrent structures 🡪 this means that we can build algorithms that recognize those structures and annotate them. Usually those are HMM-based algorithms

Repetitive elements: these are different to identify since their borders are not well defined, they can be inside other repetitive element and there are multiple classes. There are dedicated tools to identify those and there are two classes of tools that are used in conjunction:

* Homology based tools.
* De novo tools.

Homology based tools – library-based approaches:

Repetitive sequences are searched by comparing input data to a set of reference sequences contained in a library. The library can be homemade by the user, and tailored to the requirements of the job, or you can use a generalist library. The most commonly used is REPBASE.

RepeatMasker: is a library based program used for searching repetitive sequences. It can be used with custom curated libraries or with Dfam, which is an open database of multiple sequence alignments, each contains a set of representative members of a specific transposable element family. It contains also some HMMs of sequence profiles for identifying repeats and transposable elements.

De novo tools:

One of those is TEdenovo.

Masking of repetitive elements:

Masking repetitive elements means transforming every nucleotide identified as a repeat to an N. The masking of repetitive elements makes it easier to annotate the genome 🡪 for example, sometimes transposons are similar to ORFs , so masking those is very helpful.

**Annotation of genes:**

The two steps are:

* structural annotation: how do genes look like and where they are?
* functional annotation: assign a biological information to genes: it is usually done with machine learning methods and functions are computationally inferred based on similarity.

Prokaryotes annotation is much easier than in eukaryotes, since there is no splicing, there are recurrent consensus sequences, and the intergenic regions are a small percentage of the genome. Another important aspect to take in mind is codon bias, that is different for each species.

In eukaryotes, is important to check for Exon-Intron boundaries. Those boundaries are not always clear, the only certain sequence is the GT base pair as the first bases in the intron.

Structural annotation: once again there are diverse approaches (different algorithms types)

* *Intrinsic or Ab-Initio*: is uses mathematical methods (instead of external evidence) to search genes within the sequence itself, using parameters like coding potential and splice site prediction ecc. So, every model built is specific to a species, since the parameters for estimating genes are different. Those algorithms can be very accurate. With this method, ORF scanning is performed: we check for initiation codons and termination codons. The problems of this algorithms is that, when searching for ORFs, you have to scan 6 reading frames (three in one direction, three in the other direction 🡪 slide 31 lecture 7). Also for prokaryotes annotation with this method is much easier than in eukaryotes. Another important aspect to take in mind is codon bias, that is different for each species. One example of a tool that works this way is AUGUSTUS.
* *Extrinsic*: this paradigm uses external evidence for finding genes in the genome. It uses, for example, RNA-seq to find transcripts, and then finds genes from similarity. Also, other polypeptide sequences from other species can be used 🡪 we use polypeptide sequences because those are more conserved than the underlying nucleotide sequences.

It is crucial to have high quality evidence for genome annotation 🡪 RNAseq is an integral part of genome projects to have accurate information about exons and existence of genes.

Annotation of other elements:

Annotation is not only done for genes. Long-non-coding RNAs, microRNAs, pseudogenes, non protein coding genes ecc also require annotation.

Manual curation:

It is important to really validate a new assembly and annotation to have a team of curators to look at the project 🡪 and this is important because all annotation is a computationally generated.

**Annotation Formats**

Most of the time the format is a GFF file, other formats are GTF , BED , GENBANK , EMBL

GFF file (general feature format):

1. seqname - name of the chromosome or scaffold;

2. source - name of the program that generated this feature, or the data source (database or project name)

3. feature - feature type name, e.g. Gene, Variation, Similarity

4. start - Start position\* of the feature, with sequence numbering starting at 1.

5. end - End position\* of the feature, with sequence numbering starting at 1.

6. score - A floating point value.

7. strand - defined as + (forward) or - (reverse).

8. frame - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on.

9. attribute - A semicolon-separated list of tag-value pairs, providing additional information about each feature

There are 9 columns for each line, each line corresponds to a different feature. There are optional track definition lines.

BED file:

Very similar to the GTF file, with one feature per line and 3-12 columns of data, plus optional track definition lines.

**Data submission**

There are different databases such as Genbank, European Nucleotide Archive ecc. You also have to submit metadata for example: what were the specimens from which the samples were taken, what is the sequencing technology used ecc.

## Application of NGS – Different strategies

**Pool Seq**

Pool-seq consists of sequencing DNA from multiple individuals instead of just one. In a VCF file normally you also obtain the genotype of the individual, but of course in pool-seq this is not doable. 🡪 in pool-seq the genotype refers to the frequency of the allele (slide 4 Lecture 9). The allele frequency is then useful for analysis of a population.

Disclaimer: this kind of sequencing only tells you information about genetic variability, and nothing more, since you cannot derive from which individual an allele is from. Also, it is crucial to use equimolar DNA from every individual.

BGI sequencing 🡪 illumina but chinese.

**Targeted DNA-sequencing**

Targeted sequencing is a way of detecting known and novel variants in selected sets of genes or genomic regions. Usually, selection of the regions of the genome to sequence is done by PCR amplification or hybridization capture methods.

The Ion AmpliSeq gene panels can be used to select the genes that you want to sequence. Those can be pre-built gene panels, that for example contain genes involved in cancer, or you can decide to build your own panels.

For small targeted sequences, Sanger sequencing can be used after amplifying the desired genome portion.

Another option is Illumina amplicon sequencing, that also uses a PCR step to amplify the desired regions and then sequence them.

**Whole Exome Sequencing**

Sequencing the exomes is very cost-effective since in a complex genome genes are just a small percentage of the genome. Also, most of the interesting and phenotype linked variants are found in exomes. The library preparation takes as little as 1 day and can be expanded to include UTRs and microRNA genes.

To only sequence the exomes, you have to capture them first:

You fragment the genome 🡪 you use biotinylated-probes to capture the exomes 🡪 you washed non bounded fragments 🡪 sequencing.

Exome sequencing can be very useful for finding genes related to diseases (since most mutations causing disease occur in the exomes). In fact, it can be used for de-novo variant discovery, or the variants can be filtered against a set of polymorphism available in public databases.

**Methyl-seq**

With this technique you want to discover the methylation pattern of the DNA 🡪 in fact most cytosines are methylated, in particular in the CpG islands, that are genomic regions with a unique composition (Slide 34 Lecture\_9). In this technique you treat the DNA with Bisulfite. This reagent transforms normal cytosines in uracils, while methylated cytosines remain the same. Uracils are converted to Thymines during PCR are amplification. So when sequencing is performed, every Cytosine that is found is a methylated one.

The problem here is how to identify C>T SNPs. To do this, is important to know that Guanines in the filament opposing the Cytosines are not affected by conversion. This strand-specificity has been used to distinguish bisulfide converted Cytosines from Tymines.

**ChIP-seq**

Chromatin Immunoprecipitation followed by sequencing. ChiP-Seq starts with crosslinking of DNA-protein complexes, then samples are fragmented and treated with exonucleases to trim unbound nucleotides (nucleotides that interact with the protein cannot be digested by exonucleases). Then protein-specific antibodies are used to immunoprecipitate the DNA-protein complex. Then the DNA is extracted and sequenced.

**RNA-seq**

RNA seq is the sequencing of RNA. You can perform transcriptome studies, by sequencing mRNAs. Differently from the genome, the transcriptome is plastic 🡪 changes in base of the environment the organism/cell finds itself in. The different procedures of RNA-seq are:

1. Oligo(dT) primers
2. Random primers 🡪 the RNA is extracted, fragmented, and transformed in cDNA using random primers.

It is important to know that the protocols for RNA-seq are plenty and very strict.

After sequencing, a part of data analysis consists of reconstructing the transcript, and then aligning on the genome, or do the opposite. 🡪Align-then-assemble / Assemble-then-align. (Align on the reference genome). Also, another information that we get from re-assembling transcripts is to determine the most abundant transcript 🡪 so the most expressed. This procedure is not that easy and we have to:

1. Map reads on a genome
2. Compute number reads assigned to all genes in different experiments (counting matrix)
3. Normalize the data using different measures, one to remember is RPKM (Reads Per Kilobase of exons per Million mapped reads)

## Population genetics and genotyping by sequencing

**Things to know: Hardy-Weinberg law**

In population genetics, the Hardy–Weinberg principle, also known as the Hardy–Weinberg equilibrium, states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences. These influences include *genetic drift, mate choice, assortative mating, natural selection, sexual selection, mutation, founder effect, inbreeding.*

If a population is in Hardy-Weinberg equilibrium, the allele frequencies can predict the genotype frequencies.

The formula for the genotypes is:

🡪 where p and q are the two alleles. This formula can be extended for more than two alleles.

**Classification of genome variation:**

Genome variants can be molecular (SNPs and indels) or structural (CNV, translocations, inversions).

## High throughput genotyping tools

High-throughput genotyping studies are crucial for generating a large volume of genotyping data for the identification of association with phenotypes.

**Microarrays 🡪 https://www.youtube.com/watch?v=NgRfc6atXQ8**

this technology is ideal for large-scale genotyping studies. Can be used to detect single nucleotide polymorphism and other variations across the genome. You can choose the alleles that you want to detect in your study. Microarrays are a platform consisting of a chip in which there are multiple wells, and in this wells there are probes that hybridize to specific sequences. If the sequence is present in the sample, then it will hybridize, else it will not hybridize. Microarrays so do not need a reference genome in order to function.

Microarrays outputs usually 2 files: one file containing genotype information and the other containing map information 🡪 that uses a reference genome. (file containing genotype information slide 7 Lecture population\_genomics)

It is important to choose the right reference genome. In fact, there are different versions of reference genomes because those are constantly improving.

**Application of high throughput genotyping in agricultural species**

Genome wide association studies (GWAS) and QTL: GWAS and QTL mapping can identify markers that are associated to production traits and that explain a relevant fraction of the genetic variability

Genomic selection: Genomic selection aims to improve quantitative traits in large breeding populations through the use of whole-genome molecular markers. Genomic prediction combines marker data with phenotypic and pedigree data, when available, to increase the accuracy of predicting breeding and genotypic values.

Marker-assisted selection (MAS) and marker-assisted breeding (MAB): MAS and MAB efficiently select for desirable traits during breeding, where a trait of interest is selected based on a marker linked to that trait, rather than based on the trait itself.

Parentage: DNA parentage testing is a valuable tool that allows breeders to confidently select elite animals, knowing their ancestry is correct. It can be used to fast-track genetic progress or confirm parentage, and may determine if an animal carries defective genes—helping to avoid defects in future offspring.

**Probes for genotyping (usable in microarrays)**

To genotype, we use probes, which are single strand sequences that are attached to beads (or another type of support). This probe sequence stops right before the polymorphic sequence.

ATGGCGTC(G) 🡪 if G is the SNP, the probe sequence is ATGGCGTC

Adding the sequence to genotype to the system, this sequence hybridizes to the probe. Then the DNA pol adds just one base 🡪 the polymorphic base, but fluorescent nucleotides are added. Depending on the colour of the nucleotide added, we know what the base in the SNP is. If the individual is homozygous, only one colour is emitted. If the individual is heterozygous, we obtain a mixed signal instead. This system can fail if the probe is not a unique sequence in the genome. If the probe is designed in a repetitive sequence, then there is a mess. To avoid this, masking repetitive sequences is essential. Other problems that can occur is that the probe is biased toward one variant, or that the probe does not capture DNA very well ecc.

This is just one technology used for genotyping assay. Another technology can be the Axiom genotyping technology.

**Custom genotyping:**

There are companies that can create custom probes / assays for your own problem 🡪 example: design a high density SNPchip for the X chromosome.

**Genotyping by sequencing**

In this type of sequencing, you only sequence some regions of the genome with high coverage, to make sure you capture all the possible alleles. To only sequence the selected regions, you perform an enrichment based on hybridisation methods or REs digestion methods or PCR (slide 30 Lecture population\_genomics)

Genotyping by sequencing can be done without a reference genome, since you can align all the sequences with themselves and compare the SNPs this way.

**Design of a SNP genotyping assay (useful study https://doi.org/10.1371/journal.pone.0006524)**

Designing a genotyping assay you have to capture all the most informative alleles 🡪 so you have to sequence good representatives individuals of a population. An important parameter is MAF (minor allele frequencies, considering only two alleles, the max number of MAF is 0.5)

A genotyping tool should be useful for all 🡪 all breeds for example. A genotyping tool for a pig should be usable for all breeds of pigs, since you have to capture the genetic variability of the animal.

When doing a genotyping assay, DNA pooling is an option (you get a number of samples from different individuals, then you pool the samples considering that each individual should contribute the same amount of DNA in the pool 🡪 each individual contributed equimolar DNA).

Reduced representation libraries: in the cited work, reduced libraries were used to cut costs. This time reduced libraries were created with the intent to remove repeated sequences. DNA was digested with specific restriction enzymes. After doing an electrophoresis, you obtain a ladder 🡪 in each step of the ladder you have sequences with the same molecular weight 🡪 hence they can be considered equal, and probably repeats, and you discard them.

* For a genotyping tool to be viable, you have to include all possible SNPs. So, for a mammalian genome (3 Gb), knowing you have a SNP every 30 Kb, you will have roughly 100’000 SNPs in total in the genome.

RAD – sequencing: restriction site associated sequencing is a system that uses REs to cut into the DNA and generate fragments, then, some of these fragments are collected and sequenced, instead of sequencing all the genome. This cuts costs.

## Copy number variation (CNVs) analysis

A CNV is a DNA segment that is >= 1 kb and present at variable copy number in comparison with a reference genome. Tools used to analyse CNVs are:

* Array Comparative Genomic Hybridization
* Comparative intensity analysis of SNP genotypes (like the ones seen before)
* Next-Gen sequencing platforms

**Array Comparative genome hybridisation**

This needs a reference genome, because you have to build probes. Essentially you compare two genomes. The probes are designed to cover all chromosomes. You need a reference DNA and the Test DNA. The two DNAs are labelled with two different colours. Then, you put equimolar samples of those DNA on the wells containing the probes. You then measure the ratio of the two colour signals (ratio between test DNA and reference DNA) 🡪 if the two individuals have the same amount of sequence, the ratio is 1.

The ratio itself is not used, but it is instead used . If the ratio is 1, the log is 0, then if the ratio is 2, then the log is 1 ecc.

**Identifying CNVs with NGS**

Read-pair: with pair end, if there is an insertion or deletion, the length between the two paired reads is different than normal.

Split read: sometimes happens that one of the pairs of paired end sequencing falls in a deleted area, and this causes the read to not map on the genome. The same happens if there is an insertion. (slide 47 Lecture population \_genomics)

Read depth: if there is a duplication or a loss, of course there will be a difference in the read depth that map into that area.

The tools for identifying CNVs using NGS data are plenty. One of those is PennCNV, which is a free software for CNVs calling.

If you have no money, but you only have genotyping data, you can perform a qPCR to see the level of amplification of a specific region/sequence in the DNA. First you need primers for the desired sequence to be amplified, then you need two probes, one for each SNP you want to use to quantify the haplotypes. 🡪 for qPCR: https://www.youtube.com/watch?v=iu4s3Hbc\_bw

## Population Genomics

**Concepts:**

Genetic Drift: in small population, lower frequency alleles tent to disappear (drift) and the resulting population after many generations will be homozygous. Linked to this concept is the concept of Bottleneck effect: an healthy population suddenly becomes small, since the population is now small, the allele frequencies are now very different than before. The population grows again, but the allele frequency does not come back as before (slide 61 Lecture population\_genomics).

The genetic drift is an effect of both big and small populations, but in big population the effect is very small.

Fitness: there are loci associated with fitness 🡪 in this locus, the rate of heterozygosity is very low compared to the mean of the population. This difference in rate of heterozygosity can also be used to identifying the loci associated with fitness. The alleles strongly associated with fitness are called Selective sweeps / Signature of selection.

Linkage disequilibrium:

Genes on the same chromosome are called “linked” since they are in the same chromosome. A set of gene in a chromosome is therefore called an haplotype. Gene very close to each other rarely recombine, so they are inherited together, while genes on different chromosomes recombine often. Also, if a signature of selection is present in a loci, then regions around that position rarely recombine 🡪 this is linkage disequilibrium, the fact that genes on the same chromosome (or in selective sweeps) recombine less than the expected frequency of recombination for genes on different chromosomes.

Based on the level of linkage disequilibrium we can also compute how long ago the key mutation that caused LD in the region occurred. LD is also relevant in GWAS, since in a long time even region close to each other recombine (slide 10 Lecture 13\_bovo) and markers are broken apart.

**Inbreeding**

The level of inbreeding of an individual is simply how closely related the two relatives are. It is measured by the inbreeding coefficient F-ped 🡪 ped is for pedigree

This coefficient is the probability that genes at a randomly chosen location in DNA are identical by descent.

Inbreeding depression: high level of inbreeding in the population is deleterious. It causes reduction in population variability and frequency of recessive and deleterious alleles in the population.

The F-ped has some limits:

* It does not account the true relatedness of the founder animals of the base population (as it assumes that all animals of the base population are unrelated which could not be true)
* It needs complete pedigree registration for both paternal and maternal lineages to fully account for the relationships of the animals between and within lineages
* It assumes that all pedigree registrations are correct, which is difficult to verify, especially in extensive production systems in which mating events cannot be precisely recorded
* It does not consider the stochasticity of the recombination events occurring during meiosis through the generations
* It does not consider the potential biases derived by the selection on some genomic regions

Runs of homozygosity:

Runs of homozygosity (ROH) are continuous regions of chromosomes showing homozygosity at all loci. The % of genome covered by ROH can be used to identify the genomic inbreeding coefficient F-roh (different but equivalent of the F-ped, but this time the pedigree is not used).

To identify ROH, we can use the window between SNPs (so genotyping is necessary) but of course there are problems. First we have to decide the minimum number of bases that make up a ROH 🡪 the minimum window size. This depends on the density of the genotyping tool we use. Also we can lose some SNPs if they are not represented in the tool we use (for example a genotyping microarray).

ROHs can be used to find signature of selections.

In a ROH, the important parameters are the number of ROH detected (nROH) and the average length of the ROH (lROH), the sum of all ROH segments by individuals (sROH). The proportion of the autosomal genome in ROH is .

Detecting ROHs with PLINK:

PLINK tool can be used for detecting ROHs 🡪 there are different parameters (like minimum window length) that we can modify.

Fixation index – F statistics:

It is a parameter to confront populations. It is computed with the formula: ,

Where the variations are computed with the allele frequencies. An of 1 means that the two populations are completely separated. In the second formula, HT and HS represent heterozygosity frequencies (for a given locus) between populations and within populations respectively.

If we are computing FST for a locus in two populations, and that FST is close to 1, it means that in one of the two populations that allele is almost fixed, while not in the other.

## Genome Wide Association Studies GWAS

Genome-wide association studies (GWAS) involve testing genetic variants across the genomes of many individuals to identify genotype–phenotype associations 🡪 so in GWAS we work with individuals, instead of working with populations.

In GWAS individuals are firstly genotyped, then phenotyped.

**Manhattan Plot**

it is a graphical representation for SNPs fixation 🡪 you have in the x axis each SNP, and in the y axis a statistical value, like p-value or FST (FST tells you how fixed is an allele in different populations). If the FST is close to 1, that means that in one of the population this allele is fixed, while it is not true for other populations.

* Look at Manhattan plot in slide 16 Lecture 13\_bovo

**Common disease / common variant hypothesis**

This hypothesis states that if common genetic variants influence disease, the penetrance of this variants must be very small relative to that found for rare disorders. Also, 🡪 in this case more than one allele influence the disease (disease is polygenic).

The implication of this theory is that:

* If a variant correlate with disease, but is very common in the general population 🡪 the variant has a small effect on the phenotype.
* If a variant correlate with disease, but is present only in people with disease and is very rare in the general population 🡪 the variant has a high effect on the phenotype.

To test the common disease/common variant hypothesis for a phenotype, a systematic approach is needed to interrogate much of the common variation in the human genome.

1. First, the location and density of commonly occurring SNPs is needed to identify the genomic regions and individual sites that must be examined by genetic studies.
2. Secondly, population-specific differences in genetic variation must be catalogued so that studies of phenotypes in different populations can be conducted with the proper design. A study may give results for a specific population, but may not work in another population due to genetic variation. So usually studies are specific for populations/ethnicity
3. Finally, correlations among common genetic variants must be determined so that genetic studies do not collect redundant information.
4. The International HapMap Project was designed to identify variation across the genome and to characterize correlations among variants.

**Preparing a GWAS**

The main idea of an association study is to investigate the differential allele frequency of SNPs between case and control groups 🡪 the difference needs to be statistically checked.

A GWAS should consider the following issues in the design of the study:

1. Phenotype definition - determining the specific trait for investigation. A disease is a very complex phenotype, and the more complex the phenotype is, the less the sample size. 🡪 A phenotype can be binary (affected and unaffected) called **case and control** or quantitative, and in this case you need measurements. There can be semi-quantitative traits in which the trait is not continuous, but you have categories. Also another important thing is to correctly define phenotype 🡪 for example lung cancer can be defined as NSCLC or SCLC 🡪 performing a GWAS on both categories can give bad results.
2. Structure of common genetic variation in the population under investigation – LD
3. Sample size. It should be very big, in the thousands. For less complex phenotype the number can be reduced 🡪 for example if we want to find variants that correlate to the level of a specific molecule in the blood, sample size can be reduced since the level of that molecule depends only on one enzyme, so we have to capture variability only in one gene.
4. Population structure - stratification
5. Genome-wide significance and correction for multiple testing -Bonferroni correction
6. Replication 🡪 to give meaning and “strength” to the results, the experiment must be replicated.

The result of a GWAS is a Manhattan plot, where you have variants in the x axis and -log10p-value in the y axis. You then obtain plots of associations. Each dot represents a variant. Usually you obtain one very associated variant and on the same chromosome a lot of other variants associated 🡪 this is due to linkage disequilibrium, because if the phenotype causing variant is in a specific chromosome, other regions in that chromosome may look like they are correlated too.

To really pinpoint the right variant causing a phenotype, we have to look at the position of the variants and consider their relative linkage. Also, you have also to check the function of the genes in which the variants are inside.

**Linkage disequilibrium**

The international HapMap consortium demonstrated that common SNPs are arranged in block of strong linkage disequilibrium within populations, maintained by low levels of recombination and separated of hotspots of crossover activity 🡪 within a block genetic variation can be arranged on small numbers of common haplotypes.

The statistic to measure LD is: D = q12 – q1q2 , where q12 denotes the frequency of the haplotype, while q1 and q2 denotes the frequency of the two alleles. A positive LD means that the two alleles are inherited together more than expected.

If LD is 0 🡪 q12 = q1q2 and D = 0. Generally D takes values from -1 and +1.

This statistic however is dependent on the population allele frequencies, therefore is not good when comparing LD between alleles with different frequencies. To reduce this dependence two measures of LD have been proposed:



Both coefficients take values from [0 – 1]

D’ is the normalized LD statistic 🡪 the D is divided by the theoretical maximum (if D > 0) or minimum (if D < 0) difference between the observed and expected allele frequencies, this is done because the value of D is constrained by the allele frequencies. After normalization, D′ reflects the proportion of the theoretical maximum D that is achieved.

An alternative to D’ is the correlation coefficient between pairs of loci (r or Pearson correlation coefficient) usually expressed as its square. Q1 is the frequency of an allele, 1-q1 is the frequency of the second allele 🡪 here we consider just 2 alleles per loci. This coefficient essentially tells us how well one allele predicts the other.

**Indirect association**

Knowledge of the patterns of LD throughout the genome has helped in the design of efficient GWAS genotyping products. These products focus on a small number of “tag SNPs” from which we can recover much of the information about common genetic variation across the genome. Tag SNPs have indirect association with the trait under investigation, for example a genetic disease 🡪 instead of looking for the mutation causing disease, we look at the tag SNP that is strongly associated with the locus.

**Haploview**

It is a software that can show you the linkage between markers, so you can visualize better the LD of two variants.

**Sample size**

In GWAS, SNP panels are used a lot, and are considered the gold standard. But those panels, especially those with high coverage, can be very expensive if you have a sample size of 1000 individuals. A big sample size is needed to give significance to the findings.

With large populations, p-values can be very low, and therefore association can be very good.

**Bonferroni correction and false discovery rate**

When performing an experiment, there is a chance of obtaining a false-positive, and this can be quantified. A false positive means non-rejecting of the null hypothesis.

Statistical theory states that for test at a significance alfa (lets say 0.05), we expect 100\*0.05% of SNPs to yield a significant association by chance (false positives) . One method to correct for multiple testing is the Bonferroni correction. Under this approach, each test is considered independent, and the SNP wise significance level is adjusted to achieve and overall experiment-wise false positive rate of 100\*alfa%.

Example: with significance set at 0.05, if we perform 10 tests, we have a probability of false positives. This is way too much, so we adjust the significance by doing : dividing significance by the number of tests. This brings down false positives rate to 4.7%.

When testing for N SNPs 🡪 significance level is alfa / N at each SNP.

The Bonferroni correction is too conservative. For large sample size, you obtain a very small number for the significant p-value. Also, this correction was established considering that the experiments are independent, but this is not true since the SNPs are linked due to LD.

To combat the fact that with a lot of SNPs the significance will we too low, a threshold has been established for considering SNP significance and is p < 5\*10^-8.

Another problem, is that while the Bonferroni correction decreases false positives rates, it increases false negatives.

**Population stratification**

Population in a sample can be stratified 🡪 this means that there are subpopulations that have different allele frequencies, for example there is a sub-population consisting of a specified ethnicity. This, if not accounted for, can lead to masking true associations of false positives. We can construct a Multi-Dimensional Stratification 🡪 This method calculates the genome‐wide average proportion of alleles shared between any pair of individuals within the sample to generate quantitative indices (components) of the genetic variation for each individual. Individual components scores can be plotted to explore genetic variation of the individuals. In the presence of genetic structure, not accounted for in the analysis, we would expect many more signals of association, throughout the genome, than we would expect by chance.

There is the genomic control inflation factor (Ʌgc 🡪 “lambda gc) that can be computed and tells you if your population is stratified.

If your population is stratified, you can see if it is relevant for your study using the QQ-plot.

After a GWAS, we can try to correlate Phenotype (so we have to give a score to the phenotypes), and the Genotypes. The correlation should be linear in additive models (slide 49 Lecture 13\_bovo).

In GWAS the additive model is the most common, but there are different ones.

LEGGERE ARTICOLI PAG 56 LEZ13\_BOVO.

## Pedigree chart and files

Genotype data can be stored in different formats:

* .ped: contains information about phenotypes of individuals
* .map: contains information about DNA marker position

There are also binary files formats like .fam , .bim , .bed

**Ped file and Map file**

This file is essentially a tab with 6 mandatory columns:

1. Family ID
2. Individual ID
3. Paternal ID
4. Maternal ID
5. Sex
6. Phenotype

The .ped file does not contain information about the genotype 🡪 the map file does.

It also consists of a tab with 4 columns:

1. Chromosome
2. #rs or SNP identifier
3. Genetic distance (morgans)
4. Base-pair position (bp units)

Each line describes a single marker. Since in the .map file we don’t have information about the individual, it is always used in conjunction with the .ped file.