Applied Genomics

Saul Pierotti

June 4, 2020

Introduction

- Genomics is the study of genome structure and function
- The genome is the entire genetic content of an organism
- Applied genomics is the use of technologies, tools and experimental designs to analyse genome and extract information form them
- Genetics studies differences: we cannot track things that are not different among individuals
- A reference genome of a species is the basis used for analyzing the genome of an individual
  - In some cases if I do not have a reference genome I can use that of a similar species
- We have about 2 nuclear genomes per cell, but even thousands of mithocondrial genomes
- Mithocondrial genomes can be not all equal: heteroplasmy
- The human nuclear genome is around 3 Gb, the mithocondrial genome 16.7 Kb
- Population genetics is important for this course
- Small population are susceptible to high levels of inbreeding
- Differences between population arise when there are reproductive barriers
- Effective population size is the number of individual that originated a population
  - It is a measure of inbreeding
- Sex determination can be mediate by sex chromosomes, temperature, ploidy
- Phenotype is influenced by the environment
- A phenotype is an observable charachteristic
- Comparative genomics is the study of genomic differences between species
  - It is really helpful for genome annotation
- The first draft of the human genome was completed in 2001, and the HGP was started in 1990
- 3% of human DNA is coding
- Repetitive sequences are problematic for assembling genomes
- Nuclear DNA is 99.99% identical among individuals, while mitochondrial genome is more similar
- The simplest definition of gene is "coding region"
- We can predict the phenotype of an animal just looking at the genotype (!)
- To do applied genomics I need a reference genome
- If I do not have a reference genome for my species of interest, I need to construct it or I can use one of a closely-related species
- Genomics produces around 10 Zb of data per year
  - We cannot store everithing: we must select what is worth storing and what is not
  - It is interesting to look at portions that differ from the reference genome
- The cost of sequencing is dropping in a way similar to Moore's law
  - Around 2008 the drop was much faster than Moore's law, thanks to NGS
- The shotgun approach does not have a particular target, it sequences everything
- Genomic data are typically stored in the cloud
- Hardy-Weinberg equilibrium
  - $\begin{cases} p^2+q^2+2pq = f(AA)+f(Aa)+f(aa) = (p+q)^2 = 1 \\ p+q = 1 \end{cases}$
  - The allele frequencies refer to the current generation, while the genotype frequencies refer to the

    next generation
  - It holds in absence of genetic drift, non-random mating, selection, migration, mutation
- Mendel's first law: alleles segregate with other alleles
- Mendel's second law: independent assortment
- Mendel's third law: some alleles are dominant on others
- We reviewed PCR, agarose gel electrophoresys and Sanger sequencing basics
- Reference genomes can be found in the Ensemble database
- Penetrance is the proportion of individual with a given genotype that manifest the associated phenotype

Genome structure and variability in vertebrates

- LINEs are autonomous repetitive sequences of 6-8 kb
  - The LINE1 family is the most abundant
  - There are around 50k LINEs in a genome
- SINEs are depend on LINEs for transposition and are 100-300 bp long
  - They are derived from the 7SL RNA
  - The 7SL RNA is involved in the signal recognition particle that guides protein translation to the ER
- Alu is a SINE and it is the most abundant repeat in primates (1M copies)
- LINEs and SINEs are retrotransposons, transposons that move via an RNA intermediate
- MIRs(mammalian interspersed repeats) are a type of SINE found in mammals
- LTRs are retroviral elements and they are 1.5-3 kb long
- DNA transposons are 2-3 kb long and code for a trasposase
- They can be spotted with repeat masker
- This tool can mark SINE, LINE, Alu and will mask it in my sequence
- Masking means to substitute a sequence with a stretch of NNNN of the same length
- Pseudogenes can be processed or non processed (with introns) and they are not recognised by repeatmasker
- Cot curves are obtained by melting the genome and observing the re-annealing process
  - Before melting the genome is sheared in 1kb chunks
  - The rate limiting step of reassociation is the collision of complementary strands, a second order kinetic
- The copy number of a sequence influences the time needed for re-annealing
- In simple genomes the cot curves are sigmoids, while in eucariots they are complex
- Eukariotic cot curves can usually be resolved in 3 sections
  - An highly repetitive portion of DNA re-anneals quickly
  - A moderately repetitive region
  - Unique sequences
- To put in perspective in a human genome
  - Coding regions represent 1.5% of the genome
  - Conserved regions represent 3%
  - Non conserved unique regions 44%
  - Transposons are 45%
  - Constitutive heterochromatine 6.6%
  - Microsatellites 2%
- Constitutive heterochromatine is highly repetitive with short tandem repeats
  - It is typically centromeric or on the short arm of acrocentric chromosomes, where it forms constrictions
- Satellite DNA is that portion of the genome that when it is centrifuged it forms thin bands that are lighter than the bulk genome
  - Sequence density depends only on GC content
- Minisatellites are 10-30 bp long and are usuallyu near telomeres
  - Some of them are hypervariable (VNTRs), so they are useful for the identification of individuals

- Microsatellites (SSRs) are 2-5 bp and they are found everywhere in the genome
- Genes are probably around 20k, most of them protein coding
- Histone genes don't have introns
- More than 99% of genes is represented by introns
- Exons are around 200 bp on average
- Intron size is really variable, from 100 bp to several Mb
- There are portions of mithocondrial DNA integrated in the nuclear genome
  - These are called NUMTS and they are mostly pseudogenes, but maybe some of them are functional
  - They are still being integrated, so they tend to be quite variable
  - The ones integrated most recently tend to be really similar to the mithocondrial sequences
- Gene families can be in tandem or interspersed
- Instersped genes could have been moved by transposons

Sanger sequencing

- Sanger sequencing was developped by F. Sanger in 1977
- It uses DNA polymerase to extend a primer using genomic DNA as a template
- ddNTPs are incorporated in the reaction mixture in a controlled ratio
  - ddNTPs lack a 3'-OH and thus cannot be used for extending the DNA chain
  - The ration ddNTP/dNTP is typically around 1/100, but it depends on the lenght of the desired sequence
  - When a ddNTP is incorporated the extension reaction is interrupted
- In the original Sanger publication and in the first approaches radioactive labels were used
  - Either the primers were labeled, or a single dNTP (not ddNTP!) in the mixture
  - 4 different reaction were performed and separately processed, each with a single ddNTP type
- In the current approach, dye-terminator sequencing, the 4 different ddNTPs are labeled with different fluorophores
  - The strand itself and the primers are not labeled
  - A single reaction is done with all the 4 labeled ddNTPs and run together on the capillary gel
- The ssDNA fragments are run in a capillary electrophoretic apparatus
  - ssDNA hairpins are a serious probelm for lenght resolution
  - A denaturing polyacrilamide-urea gel is used
- In dye-terminator sequencing an electropherogram is produced, showing fluorescent intensity peaks for each of the 4 channels
- Sanger sequencing primers are designed according to the region that I want to sequence
  - In general I use Sanger when I have a reference!

Next generation sequencing

- NGS platforms: Illumina, Ion torrent (Thermo fisher), PacBio, Nanopore, 454
  - PacBio was going to be acquired by Illumina, but the antitrust opposed and the merger was canceled
  - We have short reads, therefore assembly is difficult
  - 454 (La Roche, pirosequencing) is practically dead today
- The depth of coverage is the number of unique reads that contain a specifc nucleotide in the assembly
- Sequencing a mammalian genome at 50x costs around 2000€+VAT in China or South Korea
  - BGI (Bejing genome institute) is the largest sequencing provider and it is chinese
  - NOVOGENE is from South Korea
  - If I chose to use these services, I need to consider shipping restrictions, costs, and product degradation
- In sequencing, if we are not sure about a variant we exclude it
- When I do genotyping by sequencing, the regions of interest have a very high depth of coverage so I can trust the results

- We have tools for allignment of reads to a reference like bowtie
    - They produce a BAM file
- There are tools for calling mutations, Indels, ecc.
- Fastq is similar to fasta but it has additional information on it
    - It uses ASCII symbols to code a quality score (PHRED score, from the homonimous software) in a separate line from the one where the bases are stored
        - PHRED uses hard-code lookup tables of peak charachteristics to estimate quality
    - The highest quality is 93 for fastq
    - The quality score is the ASCII code of the charachter (!)
        - ASCII 33 to 126 are used, encoding scores from 0 to 93
    - The quality score rarely exceeds 60 in raw data, but can be higher in assemblies
    - The threshold quality score now accepted for base calling is 30
- Illumina reads file looks like a fastq, but quality scores have a different scale
    - If my file has scores higher than 90, it is an Illumina file
- Allignments are saved in .sam format, a tab-delimited text file that can be converted in a binary .bam file
    - The sam file contains the sequence of reads, their genomic alignment coordinate, contig, mate read name
    - samtools is used for working with sam files
- Sequencers are actually high-performance PCs with Intel Xeon CPUs and 48 Gb of RAM!
- The NGS platforms that are not portable are so because the optics cannot be miniaturized
- Paired-end reads are obtained by reading a short fragment from both directions
    - If the fragment is short I just sequence it 2 times in opposing directions
    - If the fragment is longer I get information about 2 sequence at a known distance from each other
- Mate pairs are obtained from long fragments (2-5 kb)
    - I biotinylate the ends of the fragment
    - I make the fragment circular by pairing the biotinylated ends
    - I Break the circular fragment in 200-600 bp pieces
    - I select the fragment containing the biotyn (and so the opposing ends of the original 5 kb fragment)
    - I sequence it in pared-end reads: I know 2 sequences that are 5 kb apart

Ion torrent

- There are many sequencing chips, with different throughputs
- The sequencing device is a semiconductor chip with millions of nano-wells
    - Each well is represented as a pixel
- DNA fragments are clonally amplified on acrylamide beads that are poured on the chip and go in the wells, one for each well
- The chip is sequentially flodded with the 4 nucleotides, allowing a stepwise progression of DNA synthesis
- The addition of a nucleotide releases a proton, changing the pH of the well
- The drop in pH is recorded as a base call for the well
- I have clonal amplification on positively charged spheres
- During the addition of a nucleotide, a proton is released
- If I add nucleotides one at a time, I can sense the pH change due to the many protons released by the clones
- If I have multiple nucleotides of the same type in a row, I get a stronger signal
- Regions with a stretch of the same nucleotide are called omopolymeric
    - It is difficult to exactly count the number of nucleotides in these regions
- The raw data produced is called ionogram
- We use universal adapters with a specific portion to amplify the DNA fragments
- The machine is called ion or proton torrent
- In the preparation step we obtain thousands of template molecules
- The first step of the workflow is library preparation
- Library preparation depends on the kind of samples

- I can sequence amplicons, genomes, RNA libraries
- I can only sequence small fragments: I need a fragmentation step
- Fragmentation can be done by sonication or with aspecific DNases
- Playing with the time of fragmentation, I can modulate the length of the fragments
- Frequently I need to try in different ways (!)
- It is a random process!
- I have to amplify all my fragments by PCR
- It will take forever with standard PCR, so I do emulsion PCR where every drop harbours a reaction
- I then do an electrophoresis to get only the fragments of a certain size
- NGS can typically sequence from 25 up to 400 nucleotides, but the highest throughput is around 100 BP per read
- The ideal case is that in a droplet I have a bead and a single DNA fragment
- The DNA attached to the acrylamide beads after emulsion PCR is retrieved with the ion sphere particle enhancement technique
  - I use magnetic beads with streptavidin to capture the amplified acrylamide beads
  - The beads bind biotin-labeled nucleotides on the beads, so this will select only sucessfully amplified beads
  - After magnetic pull I denature the streptavidin and release the acrylamide template beads
- To increase my output, I can regolate my flow (nucleotides added) considering gc content of my target
  - One flow is the addition of 1 nucleotide to the chip
- The real throughput of my sequencing system is lower than the theoretical one
  - If 2 different fragments are amplified together I get mixed reads, and they give me false sequences as output
  - I need to ignore the mixed reads, but they will waste some of my sequencing wells
  - The same for 2 beads with the same fragment: duplicate reads
- The ionogram is expected to show on average a peak every 4 flows
  - If I see too many peaks close to each other, I probably have mixed reads
- I have a reference sequence known, and if this sequence reaches a threshold signal I keep my read, otherwise I discard it
- In missed reads I have too many empty spaces in each read, more than statistically reasonable
- Ion torrent reads are longer than Illumina, around 400 bp, and it runs in a shorter time (2 to 7 hours)
- A single run produces 50 Mb to 15 Gb raw
- Ion torrent is now specialising in clinical applications with the PGM machine
- Usually it does not support paired-end sequencing, but the PGM sequencer does

Roche 454

- It works in similar way to Ion Torrent, but it senses the release of pyrophosphate during elongation
  - NUcleotides are added in flows like for ion torrent
  - PPi and adenylil-sulfate are converted to ATP and sulfate by sulfurylase (Sulfate adenylyltransferase)
  - ATP is used by luciferase releasing a photon
  - APyrase continously removes excess NTPs that are not incorporated
- It was the first NGS to be developped, but also the first one to become obsolete
- Like Ion Torrent, it uses beads on a chip and the target is amplified by emulsion PCR
- PP_i is used by sulphurylase to synthetize ATP, ATP is used by luciferase to produce light
- Light emission is sensed by a CCD camera producing a pyrogram
- This technique was abandoned because it is too expensive
  - The cost is mainly due to the many enzymes used (sulphurylase, luciferase, apyrase)
  - The CCD camera is expensive
- It has revolutionized bacterial taxonomy because it allowed to sequence the rRNA 16s
  - This is because it can produce longer reads than other NGS techniques
- It can produce 500 Mb per run with a read lenght of 400-600 bp

Illumina

- Library preparation
  - DNA is fragmented by sonication
  - Overhangs are blunted or repaired
    - I use T4 polymerase and Klenow fragment for repair and blunting
    - Polymerase activity fills the 5' overhangs and exonuclease activity removes the 3' overhangs
  - All the 5' ends are poshporylated, since their status is unknown after sonication
    - I use T4 PNK
  - A single A is added at the 3' of the fragments to avoid ligation of different fragments in the blunt ligation step
    - Adapter on the contrary have a single T 3' overhang to facilitate ligation with the fragments
  - Adapters are ligated to fragments with ligase
  - PCR is used for selectively amplifying fragments that have been correctly ligated to adapters
    - The primers anneal to the end of the adapter and have a tail that adds sequences used for the library amplification in the flowcell
- In a flowcell, I have many oligos that can anneal with the inserted sequences at the end of the adapters
- Bridge amplification amplifyes the fragments in the flowcell
- After bridge-amplification, I get clonal clusters of fragments
- In the elongation step I add all the nucleotides together, marked with fluorophores and reversibly inhibited at their 3'
  - These modified nucleotides are called reversible dye terminator chemistry
  - The primers used anneal to the adapter that is not part of the flowcell
- The elongation is stepwise because there is a block in 3' that inhibits elongation
  - I can easily deal with homoplimeric regions (!)
  - Because of this the error rate is much lower
- After nucleotide incorporation I excite the flowcell with laser and capture the resulting fluorescence with a CCD camera
- An enzimatic step that cleaves the 3' block and the fluorophore
- I can sequence both ends of my fragments, and this is really useful for the assembly step
  - I can play with fragment size to obtain my contigs
- When I sequence a genome, I need to consider sequencing depth and coverage
  - Sequencing depth is the average number of times that a nucleotide in my reference genome is represented in a read
- In order to reduce cost, I can run more samples in the same lane by using a barcode attached to my fragments
- MySeq can be used for metagenomics (16S sequencing, 24 samples per lane) and for microbial WGS
  - It can procduce 3-7 million reads per lane
- HiSeq can be used for WGS (3-4 lanes per genome) and exome capture (4 samples per lane)
  - It can procduce 100-160 million reads per lane

AB SOLiD

- It is dead by now, but could be potentially great because it gives the highest throughput
- Its reads are really short (30 bp) so it is computationally heavy to asseble the reads and it is impossible to use with repetitive regions
- It is based on sequencing by oligo ligation detection
- DNA is fragmented and ligated to P1 and P2 adapters
  - I can prepare simple or mate-paired libraries
  - Mate-paired libraries use also an internal adapter
- A clonal population of beads is created from the fragments by emulsion PCR
- Beads are covalently attached to a glass slide
- The slide is incubated with an universal sequencing primer, di-base probes and DNA ligase
- The primer has its 5'P oriented towards the fragment (away from the bead)

- There are 16 (all the possible 4^2 2-mers) different dibase probes, marked with 4 different fluorophores
  - Each dye could correspond to 4 of the 16 di-base probes
  - The probes have the di-base specific sequence at the 3' and 6 additional aspecific positions at their 5'
- The fluorescent signal is acquired
- 3 bases at the 5' of the probes are cleaved, removing the dye
- Phosphatase is also used so to cap eventual unextended fragments
- The step is repeated until needed by ligating another di-base probe at the 5' of the previous one
- Since each probe is 8 nucleotides long but 3 nucleotides are cleaved, the system interrogates 2 bases every 5 (2 interrogated and 3 no, then 2 interrogated and so on)
- Everithing melted and repeated with different primers that have an offset of -1, -2 , -3, and -4 bases
  - In this way I get a reading for the missing positions
  - I get 2 reads from 2 different probes at each position
- There is an unique conbination of bases that can give the combination of reads, so I can reconstruct the original sequence
  - Each ordered combination of colors corresponds to a base
- It can produce more than 100 Gb per run!
- Unfortunately read lenght is around 50-100 bp

Complete genomics

- It is used for re-sequencing common genomes
- Fragments are made circular and then amplified by rolling circle amplification, obtaining DNA nanoballs (DNBs)
- DNBTMs are around 200 nm in size
- They are anchored in a chip obtained by photolitography
- The chip is made so to have a grid pattern of sticky spots
  - A sticky spot is just a small well in the surface of the chip
- Each sticky spot recives exactly one nanoball, thanks to proprietary technology
- Sequencing is done by combinatorial probe-anchor ligation (cPAL)
- The DNB contains genomic DNA and an initial adapter sequence
- 1 of 4 labeled anchor probes made of a specific position and 8 aspecific positions binds to the adapter and it is ligated with ligase (similar to AB SOLiD)
- Complete genomics was acquired by BGI and it does not sell sequencing equipment
- You send samples to them and they produce data in ~100 days

PacBio

- It is really a promising technology
- Reads are long, up to a 6-10 kb, but troughput is low
- The long reads make assembly easy and don't require really high depth of coverage
- The error rate is quite high (15%), and probably it cannot be reduced under 5%
- It is costly, 40k€ for 10x in mammalian genomes
- It does not require amplification, so no bias is introduced
- It is a golden standard for new sequencing projects, usually matched with Illumina
  - Assembling PacBio reads can be difficult due to the high error rate
  - Single Illumina reads are around 1% error, so assembling Illumina reads to PacBio reads is much easier
  - I can use PacBio to resolve repetitive regions and difficult-to-amplify regions, and Illumina for depth of coverage and accuracy
- PacBio was going to be bought by Illumina, but the anti-trust opposed it and themerger was canceled
- It is based on Single Molecule Real Time (SMRT) Sequencing
- Hairpin adapters are ligated to both ends of dsDNA, so to create a circular molecule
- A sequencing primer and DNA polymerase are added and bind the DNA molecules, but do not extend

it since nucleotides are not added

- The DNA fragments with bound primer and polymerase are inserted in a SMRT cell
- The SMRT cell contains millions of wells on its surface, called Zero-mode waveguides (ZMWs)
- Each ZMW hosts a single DNA molecule with a single DNA polymerase and primer
- The DNA polymerase is fixed at the botton of the ZMW, making the single-molecule signal detectable
- Labeled nucleotides are added and their addition is detected in each ZMW
  - The systhesis is continuous, not stepwise like in Illumina
  - Detection is possible since the nucleotide pauses for some time at the bottom of the ZMW while it is added
  - AFter addition the dye is cleaved and diffuses away
- 2 modes of operation are possible: Circular consensus Sequencing (CCS) and Continuous Long Reads (CLR)
- In CCS the circular DNA is read again and again generating an high-fidelity (HiFi) read, with 99% accuracy
- In CLR the molecule is as long as possible and it is read for as many nucleotides as possible, so to generate long reads
- It is teoretically possible to detect modified bases with PacBio but this is not still done in practice
  - Modified bases tend to spend more time in the polymerase

Oxford nanopore

- Long reads, but high error rate and low thorughput
- The reads can potentially be very long, up to 100kb depending on library preparation
- Libraries are prepared with double strand fragments in which are joined at one end by an hairpin adaptor
- DNA passes trough a modified hemolysine pore altering the ion flow thorugh the pore
  - The pore is around 1 nm in diameter (half the widdht of DNA!)
- A motor protein pre-loaded on an adapter oligo regulates the speed at which DNA moves through the pore and acts as an helicase
- After all the molecule has passed through the pore, since the adaptor is an hairpin the reverse strand is sequenced
  - In this way I have 2 reads per fragment
- Many DNA molecule pass in parallel through many nanopores (around 500 in MinION, 2000 in GridION) placed on a membrane
- Each nanopore has its own sensor and it is placed in a nanowell
- Bases are read in 4-mers and then the signal of subsequent reads is interpolated by a Recurrent Neural Network (RNN)
- Interpretation of the raw data is difficult, because the meaning of reads depends on the sequence context
  - Error rate is around 4%
  - I can recognise not only ATCG, but also U, 5mC, and other modified bases
  - Machine learning!
- The platform is cheap (~5000€, sometimes given for free) but adapters and accessories are expensive
- It is invaluable when I have to work on-site, since it is small and portable
- MinION can produce 150 Mb per run with 48kb reads
- Oxford Nanopore can be used not only for sequencing, but also for detecting protein-DNA interactions, small molecules

Other NGS techniques

- Intelligent BioSystems Mini20 is a sequencing by synthesis platform designed for clinical use
  - It has 100 nt long reads, but it is expected to be able to compete with Sanger sequencing
  - It is sensitive to repeats
  - Full costs are still not clear, but the instrument costs 120k $ and the disposable flow cell 150 $
- Genia Technologies is an early stage announcment for a system that should combine Ion Torrent and

Oxford Nanopore
    - It claims a sensitivity 1-2 orders of magnitude greater than Oxford Nanopore, with as many as 100k pores per chip, with as many as 100k pores per chip
    - Planned sample cost less than 100 $

Applications of NGS

- A genome assembly can be done in chromosomes or in scaffolds
- Scaffolds are assembled from contigs
- Sometimes it is not possible to assemble entire chromosomes
- The quality score of an assembly (n50) is the minimum size of scaffolds that contain 50% of the assembled genome
- A human chromosome is on average 80-100 Mb
- ChiPseq (chromatine immunoprecipitation) is a method used to analyse DNA-protein interactions
    - The output is a library of sequences that bind the protein of interest
    - The first step is to fix the proteins with DNA using formaldehyde
    - Subsequently, cells are lised and DNA fragmented
    - The sequences of interest are recovered with Ab against the protein of interest
    - I reverse the DNA-protein binding and sequence the fragments
- If I want to reduce cost, I can sequence only the exome
    - The exome is around 1% of the genome in humans!
    - In order to sequence the exome I need a capturing system
    - If not commercially available I have to evaluate if developping a capturing system is worth it
    - In order to enrich for the exome, I need to have specific probes that bind to exon regions, either in solution (on beads) or in microarrays
    - Probes are typically obtained from cDNA
- If I do not have enough money to sequence every individual, I can pool DNA samples in group (i.e. breed) and do a sequencing for each group
- A reduced representation library is obtained by digestion of the genome
    - I run the digest on agarose and retrieve only a specific subset of MW
    - If I see definite bands in the gel, these probably come from repeated regions that are cut at the same lenght
        - I want to exclude this (!)
    - In the digestion, I can choose a restriction enzyme with a long target sequence if I want longer fragments (cut site less probable!) and vice versa
- RNA-seq is an NGS application used for revealing the presence and quantity of RNA in a biological sample at a given moment
    - It can be used to asses alternative splicing events, post-trascriptional modifications
    - It can be used to identify exon-intron boundaries
    - In general, RNA is isolated and eventually enriched for the species of interest (mRNA, tRNA, . . .)
        - rRNA usually needs to be depleted since it represent 90% of the total rRNA content
    - cDNA is synthesised from the RNA library, fragmented and size-selected
        - Biases can be introduced at this step
        - Direct RNA sequencing has been tried by several companies
- NGS has several applications in cancer research
    - It can be used for detecting not only SNPs but also large indels, by looking at differential depth of coverage
    - I can detect cromosomal translocation by looking for reads that span 2 chromosomes
    - I can detect viruses and aother pathogen's genomes
- A CNV is a 1 kb or longer DNA segment present at variable copy number
- They can be discovered by analizing the depth of coverage of the region
    - This does not tell me in which allele the copies are (!)
- Array competitive genomic hybridization (aCGH) was once a golden standard for CNVs, now it is not

  - It is used for the identification of tumors
  - It is performed on a DNA microarray
  - Single probes are 50-75 nucleotides long and they are syntetized
    - They are selected so to be spaced around 20 kb apart and to have a specific GC content
    - I need to have a certain GC content so to be able to do the annealing step for all the microarray at the same temperature
    - I do not want probes on repeated sequences
  - I do the hybridization with a reference DNA and the sample mixed and marked with different fluorophores
  - I measure the log_2 of the ratio of the intensities in order to call CNVs
    - 0 means that I have the same number of copies, 1 that I have the double number of copies
  - If I want to decrease the noise I can decide to call only more than 5 (es) sequential calls at the same level
    - In this way I loose resolution (!)
  - Note that if I compare the X chromosome in males and females, I get double the reads in females (!)
  - It is a good complement for cytogenetics
- High density SNP arrays are the most common method used for CNV detection
  - CNV alters the signal intesity of certain probes due to the differential amount of DNA present
  - I can also detect a concerted pattern of intensity alteration in neighboring SNPs
  - Various algorithms are availabe for calling CNVs from SNP array data
- ATAC-seq (Assay for Transposase-Accessible Chromatin) uses a transposase to generate fragments in open chromatine regions, outside of nucleosomes
- Bisulphite sequencing is used to detect methylated regions by converting C but not 5mC to T with bisulphite

Plink

- A pedigree is a standardised representation of individuals in a population and relationships among them
  - It can be represented in plane text or in binary form
- Plink is an important tool for working with reference genomes
  - It can work with text files (`--file` parameter, without extension for homonimous .ped and .map files)
  - It can work with binaries (`--bfile` option)
- PED and MAP file work in pairs: I typically have my_file.ped and my_file.map with the same root name and referring to the same data
- The PED (pedigree) file is a text file with a row for each individual
  - It stores the pedigree of the population
  - This format is standard and it is used by different tools
  - It is Tab-separated and there are fields for the father, mother, sex, family, phenotype, SNPs
  - Missing data are usually reported with 0
- The MAP (map on the genome) file is a text file that has a line for each SNP
  - It reports chromosome number, SNP ID, position, distance from other SNPs
  - It is produced processing the raw output of a genotyping platform
- A polymorphism is such if it has a frequency higher than 1%
- Before doing data analysis, check your data (!)
  - I want to exclude faulty individuals and faulty loci
  - Plink can filter out data at a given threshold
  - I want to exclude low-frequency alleles: my focus is the population, not the individual
  - I can exclude SNPs that violate the HW equilibrium
  - I can exclude mendelian errors: genotypes that are impossible given the parents
- Basic usage
  - `--freq` gives the frequency of a SNP

  - If I don't trust the data provider about the sexes, I can check for absolute homozigosity at X loci: in this case I have a male
  - I can want to filter out duplicates due to sampling errors
  - If I am working with non-human or I have partially assembled scaffolds, I need to specify `--allow-extra-chromosomes` or the species, if available in plink (es. `--sheep`)
  - `--out` specifies the root filename of the output
  - `--noweb` is usually required otherwise it checks forever for updates

Genome assembly

- The main approaches for sequencing large repeat-rich genomes are whole genome shotgun and hierarchical shotgun (BAC based)
  - The human genome is repeat rich
- Hierarchical shotgun allows to resolve repetitive regions by dividing the genome in 100-200 kb chunks and sequencing these separately
  - This makes long-range assembly errors unlikely and it reduces the incidence of short-range errors
  - The single chunks are then sequenced by shotgun
- It is possible that some of these chunks suffer rearrangement in the library preparation process
- At the time of the first human genome, sequencing was expensive so we could not sequence BACs and then assemble them, we needed to select non-duplicate BACs beforehand
- I start from a gene in a known position in a chromosome, and check which BACs contain it by PCR
  - This links my assembly to the physical chromosome
- Reads are assembled in contigs, which are joined in scaffolds
- At the scaffold level I can now the gap size among scaffolds thanks to paired reads
- Genetic maps are linkage maps, and they can be used for assembling genomes
- Physical maps refer to the position of a gene in the chromosome
- A strategy to select overlapping BACs is to digest them with restriction enzymes and search for common fragments among different BACs
- The main problems of hierarchical shotgun are that it is slow and assembly is problematic if some BACs contain chimeric DNA
  - Chimeric DNA is a fragment that is created by the association of fragments from different chromosomes during the construction of the library
- N50 is a statistics that defines assembly quality in terms of contiguity
  - It is the lenght of the sorterst contig that allows to surpass 50% coverage of the genome
- The state of the art is to do a first PacBio sequencing to get a rough map to which I can attach subsequent precise Illumina paired-end reads
  - I want to use more than one Illumina run, with different lengths, so to discriminate repetitive regions and to correct errors in the PacBio phase
- Radiation hybrid maps now can be used for refining an assembly
  - I form an hybridome between an immortalized cell from a different species and a normal cell from the organism that I want to sequence
  - The hybrid will lose most of the genome of the normal cell, and it will retain a random fragment
  - In this way I can get a library (!)
  - The evaluation of the retained fragment is done by cariotyping thanks to banding patterns
  - I can then test by PCR to locate specific tags
  - By cross-referencing caryotype and PCR I can get a rough map of in which chromosome genes are (not so useful now, used in the pre-sequencing era)
  - If before the formation of the hybrid I irradiate the normal line, I break its DNA and get small fragments
  - In this case I want to have a very big library, where each clone has a small fragment
  - I can test by PCR in order to understand which markers from which chromosomes I get from each clone
  - If in my assembly I have a contig that I cannot locate, I design PCR primers for that region

  - I test all the library with the primers, and I select the clones that contain my tag
  - I check those clones for other markers of known position, and I check the ones that are more frequently associated with the tag of unknown position
  - In this way, I can say that the unallocated contig is physically linked to a tag of known position
  - The distance from between tags defined in this way is defined in cRay

Study of genomes

Linkage disequilibrium

- Linkage disequilibrium (LD) is the nonrandom association of alleles at different loci
- I can define the LD coefficient D_{AB} as the difference among the frequency of gametes carrying the combination of alleles AB, and the product of the independent frequencies of gametes with alleles A and B
$$D_{AB} = p_{AB} - p_A * p_B$$
- If $D = 0$ I am in linkage equilibrium (LE): the fequency of occurence of the AB aplotype is that expected from allele frequencies
- D decreases in generations at a rate that depends on the frequency of recombination among the 2 loci
$$D_{AB}(t+1) = (1-c) * D_{AB}(t)$$
- LE will always be reached, albeit usually really slowly
  - Even in loci in unlinked loci, D only halves each generation
- The normalised version of D, called D', is the ration with its maximum possible value
$$D' = D/D_{max}$$
- Another often used metric is r^2, which is related to D and is the correlation coefficient of the 2*2 genotype matrix
$$r^2 = \frac{D^2}{p_A(1-p_A) * p_B(1-p_B)}$$
- We can detect crossing-over by looking for the association of genetic markers
- An aplotype is a cluster of genes that are usually eredited toghether
- The probability of CO between 2 genes is measured in cM
  - 1 cM is a genetic distance such that in 100 meiosis I expect 1 CO
  - It is around 10^6 nucleotides for mammals
- If I have a simple dominant trait, I am certain only about the allele frequency of the recessive
  - I can recover it by $recessive\ allele = \sqrt{recessive\ phenotype}$
  - Doing this, I am assuming that the population is infinite, there is no mutation, no selection, no genetic drift, no migration, random mating
- If the observed genotype frequencies are different from the ones expected from HW equilibrium, It means that there are factors at play that perturbate the equilibrium
  - There can also be genotyping problems (my region is difficult to sequence and I do not get the right sequence)
- Two loci are in linkage disequilibrium if they do not occur randomly with respect to each other
- Aplotypes are patterns of genetic variation in populations
- The genotype is not sufficient for predicting the aplotypes
  - I cannot differentiate if a variation is in one chromosome or the other (!)
  - We need information on aplotype frequencies or on the parents
- PHASE is a website for analyzing aplotypes
- I cannot determine the aplotype by only looking at the genotype: I need data on the population
- In human the average linkage disequilibrium is low, around 1kb
- When effective population size is low, likage disequilibrium is large
  - This is true for lifestock

- In DNA sequencing chips, I detect a series of SNPs distanced about the linkage disequilibrium
  - If 2 SNPs are close enough, I can infer that the sequence in between is what I would expect from the aplotype

Signatures of selection

- A selective sweep is the reduction or elimination of variation among the nucleotides in neighboring DNA of a mutation as the result of recent and strong positive natural or artificial selection
  - The DNA sequence is analysed in a 40 kbp sliding window in pooled sequence data
  - The number of reads corresponding to the 2 alleales is evaluated for each SNP
    - I count the number of reads with the major allele (n_{MAJ}) and with the minor allele (n_{MIN})
  - This leads to estimating an heterozygosity score for each window position for each pool

$$H_p = 2\sum n_{MAJ} * \sum n_{MIN}/(\sum n_{MAJ} + \sum n_{MIN})^2$$

    - $\sum n_{MAJ}$ and $\sum n_{MIN}$ are respectively the sums of the number of reads for each category for all the SNPs in the window
    - If I have an homozigous region all the SNPs there will have their major allele more frequent!
    - $H_p = 1$ is when the 2 alleles have equal frequency ($\sum n_{MAJ} = \sum n_{MIN}$)
    - $H_p = 0$ when $\sum n_{MIN} = 0$
  - The H_p value is then normalized as

$$Z_{H_p} = (H_p - \mu_{H_p})/\sigma_{H_p}$$

  - I can set a Z_{H_p} threshold and call a selective sweep when the threshold is passed for a window position
  - Data is usually presented with a Manhattan plot with th Z_{H_p} score of each window position plotted along the genome axis
    - Colors are used to distinguish where a cromosome ends and another start

Quantitative trait loci (QTLs)

- A quantitative trait locus (QTL) is a locus (section of DNA) that correlates with variation of a quantitative trait in the phenotype of a population of organisms
- They are mapped by identifying which molecular markers are correlated to the observed phenotype
- Expression quantitative trait loci (eQTLs) are genomic loci that explain variation in expression levels of mRNAs

Genotyping

- Genotyping means to determine the genotype at one locus
- I can perform high throughput genotyping with beadchips
  - I have beads with primers that anneal in different positions in the genome, so to be evenly spaced and below the linkage disequilibrium lenght
  - The output of a beadchip is essentially a .map file with additional experimental information (signal intensity for the SNP)
  - The position of some probes in the genome is unknown, so the row of their SNP starts with 0 (chromosome) and ends with 0 (position)
- The main genotyping platforms are from Illumina and Affimetrix
- The probe is designed so to
  - Bind to a unique region (it has to be long enough!)
  - It has to have standard GC content, so I can melt all the chip at the same temperature
- The specific fragments to be genotyped are detected by primer extension
  - I have a primer right in front of a SNP
  - I add the 2 possible nucleotides for the SNP labeled with different fluorophores and blocked

    - I see what happens
- The minor allele frequency (MAF) is the frequency of the rarer variant of a SNP
    - It can go from 0 to 0.5
- I do not need to genotype all the SNPs
    - I can take advantage of linkage disequilibrium to detect aplotypes
    - Polimorphic sites are more informative than sites with rare variants, so I tend to focus on them for determining an aplotype
- Genotyping by sequencing (GBS) allows to detect unknown SNPs and it is typically done with pooled reduced representation libraries
- Illumina can produce customized genotyping chips

GWAS

- I want to find the association between a phenotype and a genomic locus
- I can genotype individuals with a SNPs array and see if there is association with the phenotype
    - I check allele frequencies that differ in the different cohorts
- The result is a Manhattan plot
    - I have the chromosome lenght on the x axis (coordinate of the SNPs)
    - In the y axis I have the -log of the p-value for the association
        - Lower p-values are on the highest part (!)
- I am doing a lot of multiple testing so my threshold must be really high (!)
    - I use the Bonferroni correction or false discovery rate
- Continuous traits tend to be normally distributed
    - On a SNPs A/G I can have 3 possible genotypes: AA, AG, GG
    - I measure the genotype of each individual and its continuous trait
    - I take the means of the groups for each genotype and I perform a statistical test on means, like ANOVA
- In order to maximize the differences between pools, I can take samples from the extreme ends of the phenotype distribution
    - This is called extreme phenotype study

Runs of Homozigosity

- The inbreeding coefficient indicates the probability that random positions among 2 individuals are equal by descent
    - It is calculated by tracing a close path on the pedigree of an individual
- Runs of homozigosity (ROH) refer to stretches of chromosome which are completely homozygus
    - This could mean that the 2 stretches are identical by descent (!)
    - The ROH % is equivalent to the coefficient of inbreeding

Selected papers

Strong signatures of selection in the domestic pig genome - Rubin et al. 2012

- Selective sweep showed selection signature in QTLs related to elongation of the back and number of vertebrae
    - Wild boars have 19 vertebrae, domestic pigs 21-23
- The domestic pig evolved from several divergent subspecies of wild boar
- They used WGS on the draft pig genome assembly for udentifying loci under seleciton since pig domestication
- They looked for allele frequency differences among wild boar and pig populations
- 2 different datasets
    - Mate pair reads from 8 pools of pigs and wild boars at 5x depth per pool
    - Paired-end reads from 37 individual pigs and 11 wild boars at 10x coverage each

- Identified an homoziguous region in the X chromosome, that is however present also in wild boars
- Selective sweep at the melanocortin 4 receptor locus, related to food intake
- Selective sweep at the NR6A1 locus, related to the number of vertebrae
- Selective sweep candidates at PLAG1, LCORL, related to body lenght
    - They are related to height also in humans
- Selective sweep at OSTN, related to the type of muscle and bone development
- Observed excess of non-synonimous substitutions in derived mutations in domestic pig and scarcity of non-sense mutations
- 8 kb duplication in a CASP10 intron in domestic pigs
- Structural variants at the KIT locus related to white spotting

High-throughput SNP discovery in the rabbit (*Oryctolagus cuniculus*) genome by next-generation semiconductor-based sequencing - Fontanesi et al. 2014

- Sequenced 2 RRLs for SNP discovery using IonTorrent Personal Genome Machine
- Genomic DNA of 10 rabbits from different breeds pooled and digested separately with HaeIII and RsaI
- Sequenced 280 Mb from the first RRL and 417 Mb from the second
- Reads combined covered 15.82% of the rabbit genome
- 62k SNPs were called
- Some SNPs were validate by Sanger

Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology - Ramos et al. 2009

- 19 RRLs derived from 4 pig breeds and a wild boar population, using AluI, HaeIII and MspI
- Sequencing with Illumina with 36 nt reads
- Identified more than 372k SNPs
- These and other SNPs (549k in total) were used for designing the Illumina Porcine 60k+ SNP beadchip, which included 64k SNPs

Whole-genome resequencing reveals loci under selection during chicken domestication - Rubin et al. 2010

- 44.5x coverage of the chicken genome using pooled DNA from 8 chicken populations and a red jungle fowl population (major wild ancestor)
- Sequencing with SOLiD
- They distinguished broilers and layers populations
- Reported 7M SNPs, 1300 deletions, and some selective sweeps
- Selective sweep in all the domestic population at the TSHR locus (metabolic regulation and photoperiod control of reproduction)
- Selective sweeps in broilers associated with growth, appetite and metabolic regulation
- No much evidence of loss-of-function mutations in chicken evolution

Examination mode

- Final exam has 2 levels
    - Preparation of a genomic project
        - A text should be written including an appropriate introduction to the problem/question that the experiment or project would like to analyse or answer, aim of the project, a section with materials and methods, expected results and impact
        - The project should be submitted to the professor one week before the interview
        - We should specify what is the aim of the project and what I'd like to solve with it

      - If it makes sense, we can undergo a discussion with him
    - The project is based on money: we'll have a budget
  - Interview based on the project submitted and other two questions
    - Only students that are positively evaluated at the first level are admitted at the second level
    - Evaluation of basic knowledge
- We get one extra point if we pass at the first attempt
- It is important to follow him
- We'll have an example of a project, the topic of the project it's up to us
- We need to choose a complex genome/organism
- Each one will have a different budget
- It's better to do the project according to what we discuss in the lectures
- It has to be something new
- The first date would be in February after Winter School and another one in March
- Near to the end of the course we'll have a test with 30 questions to test our level (it won't count for the final score)