Applied Genomics

Saul Pierotti

January 25, 2020

Course structure

- Population genetics
- Genome structure and variability in vertebrates (we may mention plants and bacteria)
- High throughput genomic platforms
- Applications of NGS
- Array comparative genome hybridization
- PLINK, genetic data analysis. How to use this software and apply some design using this tool
- Linkage analysis and genetic mapping
- QTL analysis

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)

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
- 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
- Penetrance is the proportion of individual with a given genotype that manifest the associated phenotype

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
- Wen 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.
- If I do not have enough money to sequence any 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
- 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
  - The quality score is the ASCII code of the charachter (!)
  - The highest quality is 90 for fastq
  - The quality score rarely exceeds 60 in row data, but can be higher in assemblies
  - The threshold quality score now accepted for base calling is 30
- Allignments are saved in .sam format, a tab-delimited text file that can be converted in a binary .bam file
  - samtools is used for working with sam files

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 beads that are poured on the chip and go in the wells, one for each well
- The chip is sequentially flod 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
- In emulsion PCR I use a bead, different from the one used in sequencing
- The ideal case is that in a droplet I have a bead and a single DNA fragment
- The bead is used to retrieve my sample after the PCR
- 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 real throughput of my sequencing system is lower than the theoretical one
- To increase my output, I can regolate my flow (nucleotides added) considering gc content of my target
- 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
- 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 part of interest, for instance the exome
  - In order to sequence the exom 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
- 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
- HiSeq can be used for WGS (3-4 lanes per genome) and exome capture (4 samples per lane)

Roche 454

- It works in similar way to Ion Torrent, but it senses the release of pyrophosphate during elongation
- 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)
  - 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

Illumina

- Adapters are ligated to my fragments
- In a flowcell, I have many oligos that can anneal with the adapters
- After bridge-amplification, I get clonal clusters of fragments
- In the elongation step I add all the nucleotides together, marked with fluorophores
- 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
- There is 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

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

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 obtain by photolitography
- The chip is made so to have a grid pattern of sticky spots
- 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 adapter sequence
  - An anchor probe binds to the adapter
  - Fluorescent nucleotides are incorporated by ligase
- This company 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 error rate is quite high, and probably it cannot be reduced under 5%
- It is costly, 40k€ for 10x in mammalian genomes
- It is a golden standard for new sequencing projects, usually matched with Illumina

  - PacBio facilitates assembly, Illumina gives a low error rate
- PacBio was going to be bought by Illumina, but the anti-trust opposed it and themerger was canceled

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
- DNA passes trough an hemolysine pore altering the ion flow thorugh the pore
- Interpretation of the raw data is difficult, because the meaning of reads depends on the sequence context
  - 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

Aplotypes

- 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

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 belowe 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

Plink

- A pedigree is a representation of individuals 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 a file.ped and 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 from the output of a genotyping platform
- A polymorphism 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 given a threshold
  - I want to exclude low-frequency alleles: my focus is the population, not the individual
  - I can exclude SNPs that violate HW
  - 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

De novo sequencing

- The human genome is repeat rich
- The main approaches are whole genome shotgun and hierarchical shotgun approach (BAC based)
- Hierarchical shotgun allows to resolve repetitive regions by building bigger contigs (!)

- 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
- 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
- The alternative to hierarchical shotgun is whole genome shotgun
- 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

Repeated sequences

- 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 repeat-masker

DNA chips

- 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

Copy number variation

- 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 (!)
- 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
- 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 %
    - I need to have a certain GC % 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

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

Detect inbreeding

- 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