Applied Genomics

Saul Pierotti

November 25, 2019

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
- A reference genome of a species is the basis used for analyzing the genome of an individual
- 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, 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
- 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
- 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
- Mendel's second law: independent assortment
- We reviewed PCR, agarose gel electrophoresys and Sanger sequencing basics

Next generation sequencing

- NGS: Illumina, Ion torrent (Thermo fisher), PacBio, Nanopore, 454
  - PacBio is going to be acquired by Illumina
  - We have short reads, therefore assembly is difficult
  - 454 (La Roche, pirosequencing) is practically dead today

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, called omopolymeric, it is difficult to exactly count the number of nucleotides
- fastq is a fasta file with additional information attached
- The raw data produced is called ionogram

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
  - Coverage is

Plink

- Plink is an important tool for working with reference genomes

To be organised

- 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)