

Single Nucleotide Polymorphism

D. Tosoni, U. Buonadonna, D. Sicignani

Biomedical Informatics, 2014

Introduction

DNA

In 1870, the Swiss chemist Miescher discovered inside the nucleus of a cell a giant molecule: **deoxyribonucleic acid**.

Deoxyribonucleic acid (DNA) is a molecule that encodes the genetic instructions used in the development and functioning of all known living organisms and many vi-ruses. DNA is a nucleic acid; together with proteins and carbohydrates, nucleic acids compose the three major macromolecules essential for all known forms of life.

DNA components

Most DNA molecules consist of *two biopolymer strands coiled around each other to form a double helix*. The two DNA strands are known as **polynucleotides** since they are composed of simpler units called **nucleotides**. Each nucleotide is composed of a **nitrogen-containing nucleobase**—either **guanine** (G), **adenine** (A), **thymine** (T), or **cytosine** (C)—as well as a monosaccharide sugar called **deoxyribose** and a **phosphate** group.

The nucleotides are joined to one another in a chain by *covalent bonds between the sugar of one nucleotide and the phosphate of the next*, resulting in an *alternating sugar-phosphate backbone*.

Rules: A with T and C with G.

DNA structure

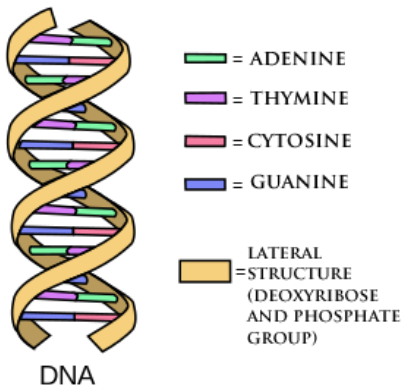


Figure 1: DNA structure

Mutations

That said, it is easy to understand how DNA is important for life. For this reason, even a small mutation (a change of the nucleotide sequence of the genome of an organism) can be decisive and cause diseases.

We will discuss a particular case of genomic mutation, the **Single Nucleotide Polymorphism**.

Single Nucleotide Polymorphism

Single Nucleotide Polymorphism: what is it?

A **Single Nucleotide Polymorphism (SNP)** is a DNA sequence variation occurring commonly within a population (e.g. 1 per cent) in which a Single Nucleotide — A, T, C or G — in the genome differs between members of a biological species or paired chromo-somes.

SNP example

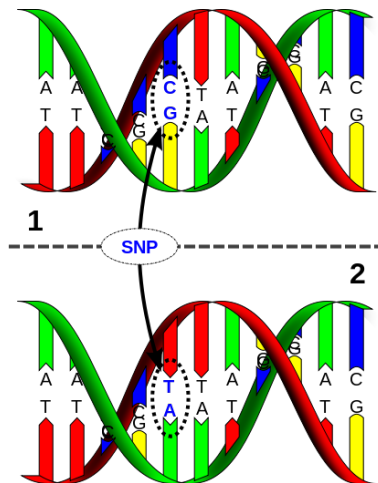


Figure 2: SNP example

What can cause a SNP?

The main causes of a SNP are:

- ① natural selection, acting and fixating the allele of the SNP that constitutes the most favorable genetic adaptation
- ② like genetic recombination
- ③ mutation rate

The different possible types of SNPs

What is a coding?

Genetic Code: it is the *set of rules* by which information encoded within genetic material (DNA or even mRNA sequences) is *translated* into proteins by living cells.

During the translation, the sequence of nitrogenous bases is treated in groups of three at a time (**codon**). The code defines how codons specify which amino acid will be added next during protein synthesis.

The “original sin”

Generally, three-nucleotide codon in a nucleic acid sequence specifies a single amino acid. On the other hand, **a single amino acid can be specified by more than one codon.**

Amino acid	Codons
Ala/A	GCT, GCC, GCA, GCG
Arg/R	CGT, CGC, CGA, CGG, AGA, AGG
Asn/N	AAT, AAC
Asp/D	GAT, GAC
Cys/C	TGT, TGC
Gln/Q	CAA, CAG
Glu/E	GAA, GAG
Gly/G	GGT, GGC, GGA, GGG
His/H	CAT, CAC
Ile/I	ATT, ATC, ATA
Leu/L	TTA, TTG, CTT, CTC, CTA, CTG
Lys/K	AAA, AAG
Met/M	ATG
Phe/F	TTT, TTC
Pro/P	CCT, CCC, CCA, CCG
Ser/S	TCT, TCC, TCA, TCG, AGT, AGC
Thr/T	ACT, ACC, ACA, ACG
Trp/W	TGG
Tyr/Y	TAT, TAC
Val/V	GTT, GTC, GTA, GTG
START	ATG
STOP	TAA, TGA, TAG

Figure 3: Amino acids

Types of SNPs

SNPs may fall within *coding* sequences of genes, *non-coding* regions of genes, as well as in the *intergenic* regions (regions between genes).

SNPs in the coding sequences

SNPs that fall in this category can be divided into two subcategories:

- ① Synonymous
- ② Nonsynonymous
 - Missense
 - Nonsense

Missense mutation - Example

Original DNA code for the amino acid sequence:

C	A	T	C	A	T	C	A	T	C	A	T	C	A	T	C	A	T
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Resulting amino acids:

His	His	His	His	His	His	His
-----	-----	-----	-----	-----	-----	-----

If we had, for example, a replacement of the eleventh nucleotide:

C	A	T	C	A	T	C	C	T	C	A	T	C	A	T	C	A	T
---	---	---	---	---	---	---	----------	---	---	---	---	---	---	---	---	---	---

Resulting amino acids will be:

His	His	His	Pro	His	His	His
-----	-----	-----	------------	-----	-----	-----

Nonsense mutation - Example

Original DNA code for the amino acid sequence:

A T G	A C T	C A C	C G A	G C G	C G A	A G C
-------	-------	-------	-------	-------	-------	-------

Resulting amino acids:

Met	Thr	His	Arg	Ala	Arg	Ser
-----	-----	-----	-----	-----	-----	-----

If we had, for example, a replacement of the tenth nucleotide:

A T G	A C T	C A C	T G A	G C G	C G A	A G C
-------	-------	-------	--------------	-------	-------	-------

Resulting amino acids will be:

Met	Thr	His	Stop			
-----	-----	-----	-------------	--	--	--

SNPs not in coding regions

SNPs that are not in protein-coding regions may still affect:

- 1 gene splicing
- 2 transcription factor binding
- 3 messenger RNA degradation
- 4 ...

Gene expression affected by this type of SNP is referred to as an **eSNP** (*expression SNP*).

How to found SNPs: DNA sequencing

DNA sequencing

“DNA Sequencing is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of the four bases — adenine, guanine, cytosine, and thymine — in a strand of DNA.”

History of DNA Sequencing

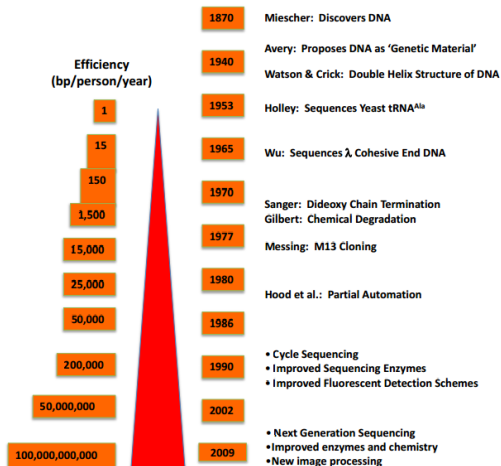


Figure 4: History of DNA Sequencing

Sequencing methods

Over the years, many methods have been developed for sequencing the DNA:

- 1 **basic methods**, such as the *Maxam-Gilbert sequencing* and *Chain-termination methods*
- 2 **advanced methods**, such as the *Shotgun sequencing* or *PCR Bridge*
- 3 **next-generation methods**, such as *Massively Parallel Signature se-quencing (MPSS)*, *Polony sequencing*, *454 pyrosequencing*, *Illumina (Solexa) sequencing*, *SOLiD sequencing*, *Single Molecule Real Time (SMRT) sequencing*, ...

Next-Generation Sequencing

Nowadays, thanks to technological progress we pushed further forward. It is possible to sequence **more than 100 million base pairs in about a week** (generating a very high amount of data).

This is called the **Next-Generation Sequencing**.

However, the higher the speed of sequencing, the more there is a problem: **interpretation**. It often represents a real bottleneck; a single computer is not able to interpret a sequencing at the same speed of which it is presented to him.

Solution: *cloud computing*.

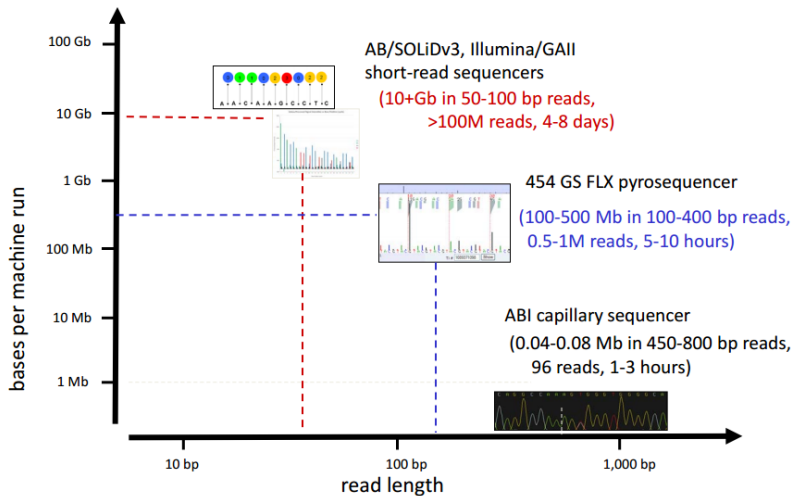


Figure 5: Nowadays Sequencing

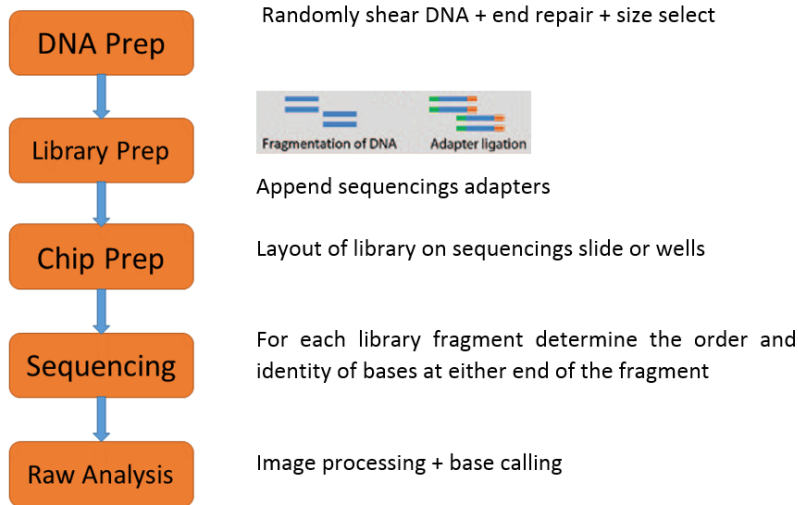


Figure 6: NGS work-flow

Whole Exome Sequencing

- The type of sequencing used to obtain the data that our web-app manages.
- The **Whole Exome Sequencing** test is a highly complex test that is newly developed
- In contrast to “common” sequencing tests that analyze one gene or small groups of related genes at a time, the WES test analyze the *exons or coding regions of thousands of genes simultaneously* using next-generation sequencing techniques.

Exome: portion of the human genome that contains functionally important sequences of DNA that direct the body to make proteins essential for the body to function properly

Whole Exome Sequencing

- It is known that **most of the errors that occur in DNA sequences that then lead to genetic disorders are located in the exons.** Therefore, sequencing of the exome is thought to be an efficient method of analyzing a patient's DNA to discover the genetic cause of diseases or disabilities.
- WES includes a **mitochondrial genome sequencing.** (Mitochondria: structures within cells that convert the energy from food into a form that cells can use).

Data Format

CSFASTA

The sequencer that generates the data managed by our web-app provides the results of the sequencing using format **CSFASTA**

>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken

```
ADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSLGQNPTEAELQD  
MINEVDADGNGTID FPEFLTMMARKMKD TDSEEEIREAFRVFD-  
KDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA  
DIDGDGQVNYEEFVQMMTAK*
```

SNP Databases

Why databases?

Because SNPs are expected to facilitate large-scale association genetics studies, there has recently been great interest in SNP discovery and detection. For this reason databases can to serve as a central repository. Once discovered, polymorphisms could be used by additional laboratories, using the sequence information around the polymorphism and the specific experimental conditions.

Most important DBs

- 1 **dbSNP**, a SNP database from the *National Center for Biotechnology Information (NCBI)*
- 2 **SNPedia**, a wiki-style database supporting personal genome annotation, interpretation and analysis

Support DBs

Furthermore, there are various support database that allow, for example, to bind a SNP to the disease that causes:

- 1 **OMIM** database describes the association between polymorphisms and diseases (e.g., gives diseases in text form)
- 2 **Human Gene Mutation Database** provides gene mutations causing or associated with human inherited diseases and functional SNPs
- 3 **GWAS Central** allows users to visually interrogate the actual summary-level association data in one or more genome-wide association studies
- 4 ...