UNIVERSITÀ DEGLI STUDI ROMA TRE

Department of Engineering
Master Degree Course in Computer Engineering

Single Nucleotide Polymorphism

Thesis of the course of Biomedical Informatics

Authors
D. Tosoni , D. Sicignani, U. Buonadonna

Academic Year
2013/2014

Chapter 1

Introduction

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

In 1953, two biochemists, the American *James Watson* and the English *Francis Crick* show that the structure of the DNA molecule is comparable to that of a spiral staircase; a sort of spiral-shaped double helix.

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

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 *al-*

ternating sugar-phosphate backbone. According to base pairing rules (A with T and C with G), hydrogen bonds bind the nitrogenous bases of the two separate polynucleotide strands to make double-stranded DNA.

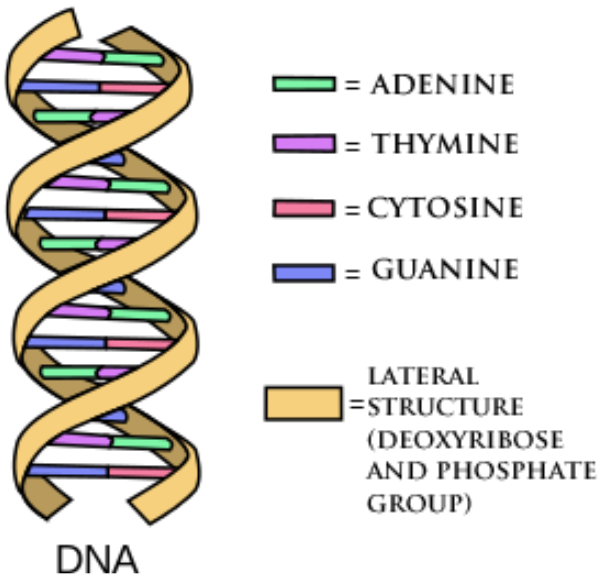

Figure 1.1: DNA structure

DNA is **well-suited for biological information storage**. The DNA backbone is resistant to cleavage, and both strands of the double-stranded structure store the same biological information. Biological information is **replicated** as the two strands are sepa-rated. A significant portion of DNA (more than 98 per cent for humans) is non-coding, meaning that these sections do not serve a function of encoding proteins.

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.

In this essay we will discuss a particular case of genomic mutation, the **Single Nucleotide Polymorphism**.

Chapter 2

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.

For example, if we have two sequenced DNA fragments from different individuals (see *Figure 2.1*):

- AAGCCTA
- AAGCTTA

The second one contain a difference in a single nucleotide. In this case we say that there are two alleles. Almost all common SNPs have only two alleles.

The genomic distribution of SNPs is not homogenous; SNPs occur in non-coding regions more frequently than in coding regions.

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

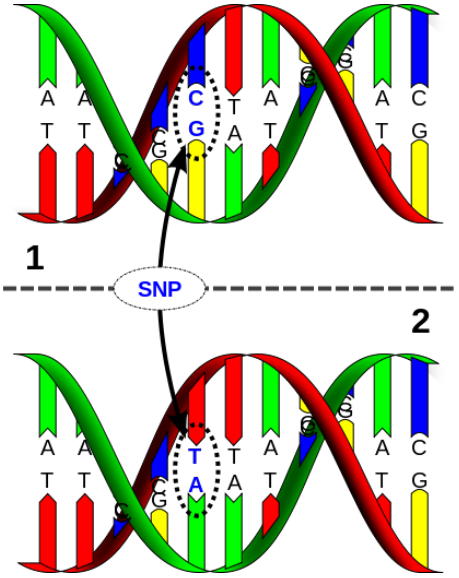

Figure 2.1: SNP example

2.1 The different possible types of SNPs

As previously seen, Single Nucleotide Polymorphisms may fall within *coding* sequences of genes, *non-coding* regions of genes, as well as in the *intergenic* regions (regions between genes).

2.1.1 What is a coding?

To understand the difference between SNPs' types, we have to see what a coding is.

The main concept to analyse is the **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; a group of three nitrogenous bases is called a **codon**. The code defines how codons specify which amino acid will be added next during protein synthesis. 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**: this is the key concept that we will need in the following.

To understand better, here is the table that shows, for each amino acid (20 in total + START and STOP), the sequences that can generate it:

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

2.1.2 SNPs in the coding sequences

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

1. Synonymous
2. Nonsynonymous
    - Missense
    - Nonsense

First ones does not result in a change in the protein sequence, because the "original" sequence and the real sequence of bases both code the same amino acid.

Second ones, instead, change the amino acid sequence of protein. In their turn, they can be of two types: *Missense*, in which a single nucleotide change results in a codon that codes for a different amino acid (that can render the resulting protein non-functional), and *Nonsense*, that results in a premature stop codon, or a nonsense codon and then in a truncated, incomplete, and usually non-functional protein product.

Let us see an example of *Missense mutation*:

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	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 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

This is, instead, an example of *Nonsense mutation*:

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**			

Nonsense mutation are jointly responsible for many diseases; they can cause a genetic disease by damaging a gene responsible for a specific protein (for example *dystrophin* in *Duchenne muscular dystrophy*).

Examples of diseases in which nonsense mutations are known to be among the causes include:

- **Cystic fibrosis**
- **Duchenne muscular dystrophy** (dystrophin)
- **Beta thalassaemia** (beta-globin)
- **Hurler syndrome**

On the other hand, cancer associated Missense mutations can lead to drastic destabilisation of the resulting protein.

2.1.3 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*).

Chapter 3

How to found SNPs: DNA Sequencing

In order to understand, according to what already said, what a SNP can cause, we must first of all *identify the SNP in the DNA* of the subject under consideration.

There are many possible types of analysis that can be performed (DNA sequencing, capillary electrophoresis, mass spectrometry, electrochemical analysis, ...); in this essay we will look at the most common one: **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."

3.1 History of DNA Sequencing

Over the years, since the discovery of DNA by *Miescher* in 1870, the problem of DNA sequencing has been addressed in an increasingly thorough (see *Figure 3.1*). This also

because, in 1940, *Avery* realized, by means of an experiment, that the so-called *transforming principle* (the carrier of genetic information) discovered in 1928 by Griffith was DNA.

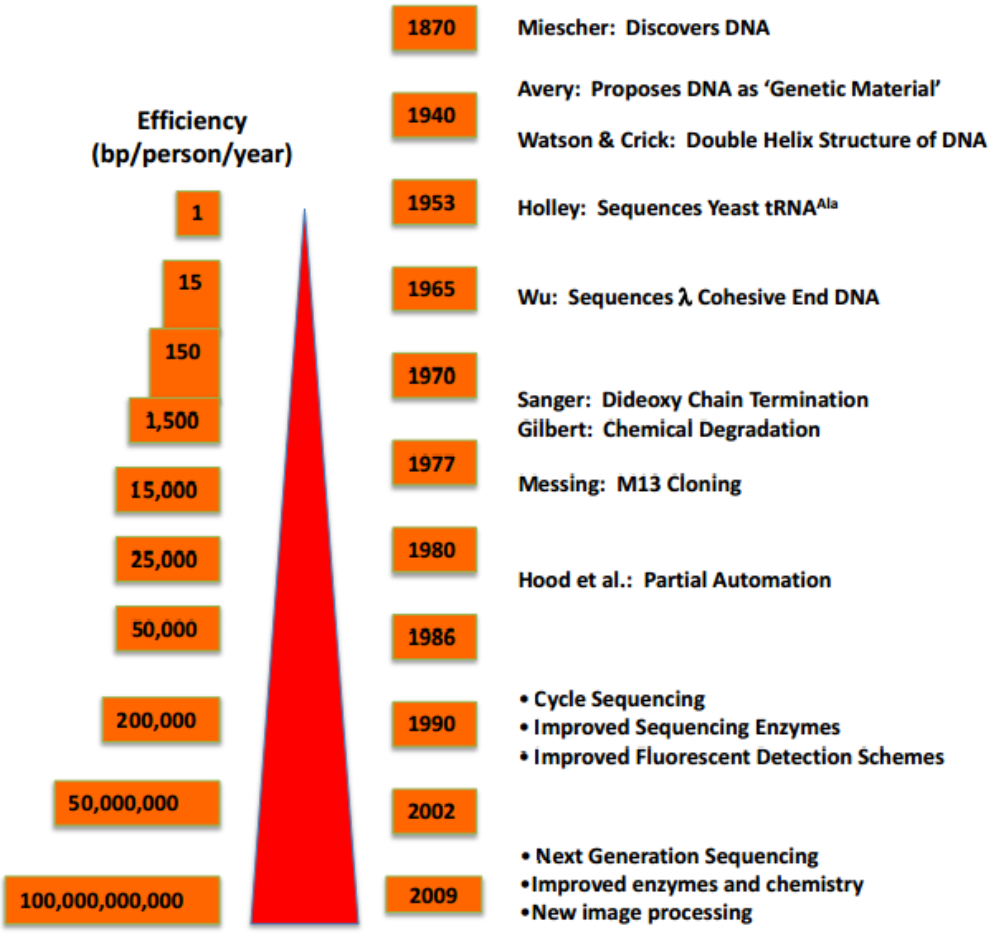

Figure 3.1: History of DNA Sequencing

3.1.1 Avery's experiment

In short, the Avery experiment was based on the Griffith experiment. Griffith used in his studies the *Streptococcus pneumoniae*. In particular, two of its strains:

- the S strain, which can cause pneumonia in guinea pigs (virulent strain)
- the R strain is not able to cause pneumonia in guinea pigs (avirulent strain)

The main result was this:

injection in mouse of type S bacteria, killed after thermal treatment, and type R live bacteria was able to cause disease and death of the animal. From the tissues of mouse could isolate live bacteria of the S strain.

So, he verified and demonstrated that in a mixture containing either S dead bacteria and R alive bacteria, were to be happened the exchange of some substance (genetic material) that would confer virulence to bacteria R (which were then transformed into S).

The experiment of Avery aimed to determine what this substance was and it was discovered that it would necessarily be DNA.

3.1.2 Progress over the years

The first sequencing occurred in 1953 by Holley.

Since then, the efficiency of sequencing increased exponentially over the years: if in 1953 a person in a year could sequence only one *bp* (base pair), in the seventies we get to more than 1,500, in the nineties to more than 200000 and few years ago, in 2009, to more than 100 billion bps!

The first full DNA genome to be sequenced was that of a bacteriophage in 1977. Medical Research Council scientists deciphered the complete DNA sequence of the Epstein-Barr virus in 1984, finding it to be 170 thousand base-pairs long.

3.1.3 Sequencing methods

Over the years, many methods have been developed for sequencing the DNA. It goes from **basic methods**, such as the *Maxam-Gilbert sequencing* and *Chain-termination methods*, to **advanced methods** such as the *Shotgun sequencing* or *PCR Bridge*, to get to the **next-generation methods** (*Massively Parallel Signature se-quencing (MPSS), Polony sequencing, 454 pyrosequencing, Illumina (Solexa) sequencing, SOLiD sequencing, Single Molecule Real Time (SMRT) sequencing, ...).*

3.1.4 Next-Generation Sequencing

Nowadays, thanks to technological progress we pushed even further forward. As can be seen from the following chart, 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.

For this reason, usually *cloud computing* services are used. They allow to take advantage of the computing power of multiple computers at the same time, parallelizing the work and thereby reducing the overall time.

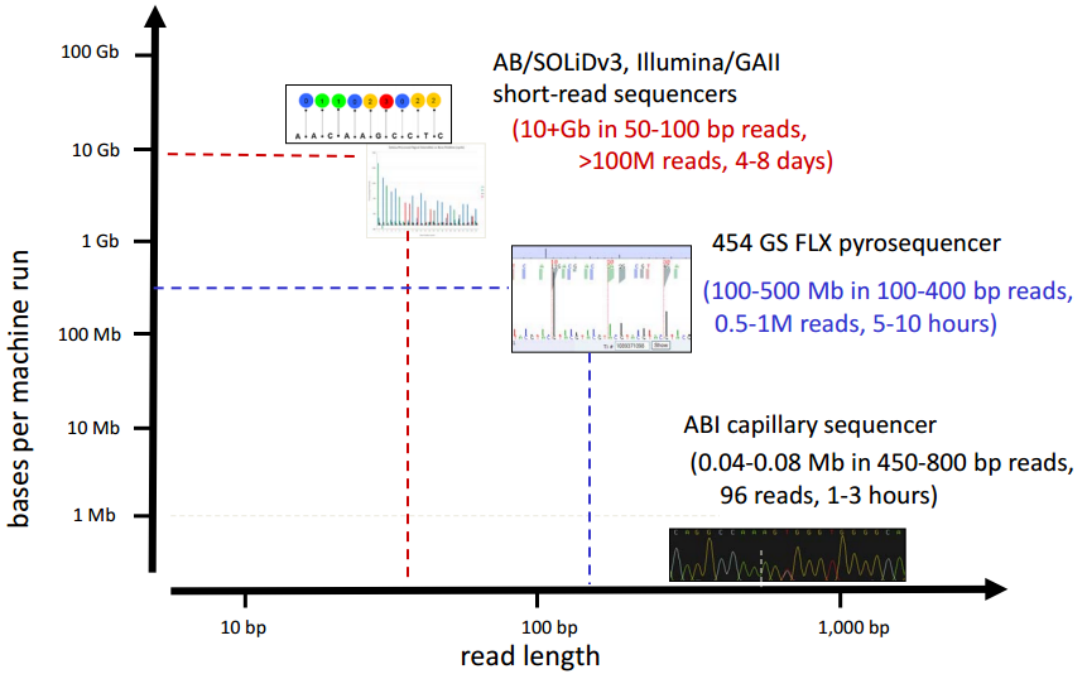

Figure 3.2: Nowadays Sequencing

The high demand for low-cost sequencing has also driven the development of high-throughput sequencing (or next-generation sequencing) technologies that parallelize the sequencing process, producing thousands or millions of sequences concurrently. High-throughput sequencing technologies are intended to *lower the cost of DNA sequencing* beyond what is possible with standard methods.
In ultra-high-throughput sequencing as many as 500,000 sequencing-by-synthesis operations may be run in parallel.

Although each next-generation sequencing platform is unique in how sequencing is accomplished, there is a similar base methodology that includes preparation, sequencing, and data analysis. Within each generalized step, the individual platforms have unique aspects.

The common work-flow is the following:

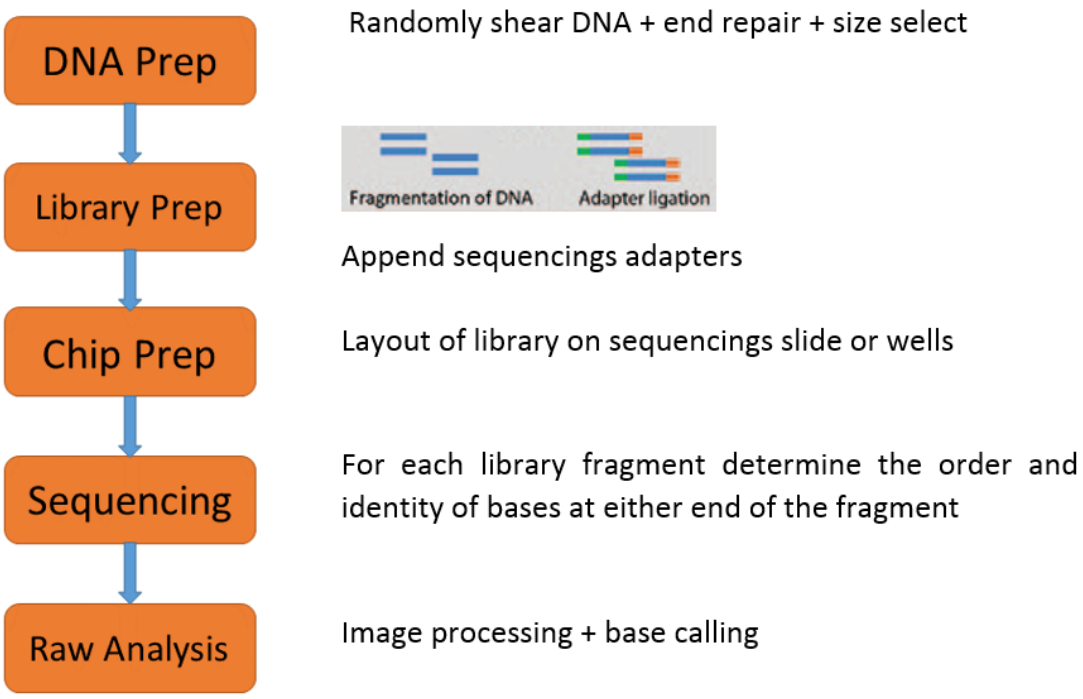

Figure 3.3: NGS work-flow

3.1.5 Whole Exome Sequencing

The type of sequencing used to obtain the data that our web-app manages is the **Whole Exome Sequencing**.

The Whole Exome Sequencing test is a highly complex test that is newly developed for the identification of changes in a patient's DNA that are causative or related to their medical concerns. In contrast to "common" sequencing tests that analyze one gene or small groups of related genes at a time, the Whole Exome Sequencing test analyze the *exons or coding regions of thousands of genes simultaneously* using next-generation sequencing techniques.

The **exome** refers to the *portion of the human genome that contains functionally im-*

portant sequences of DNA that direct the body to make proteins essential for the body to function properly. There are approximately 180000 exons in the human genome; these exons are arranged in about 22000 genes.
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.
Additionally, the WES includes a **mitochondrial genome sequencing**. Mitochondria are structures within cells that convert the energy from food into a form that cells can use. Although most DNA is packaged in chromosomes within the nucleus, mitochondria also have a small amount of their own DNA. This genetic material is known as **mitochondrial DNA**. In humans, mitochondrial DNA represents a small fraction of the total DNA in cells, but many genetic conditions are related to changes in particular mitochondrial genes.

The key principle of the test is:

"*to sequence nucleotide by nucleotide, the human exome of an individual to a depth of coverage necessary to build a sequence with high accuracy. This sequence is then compared to standards and references of what is normal in the population and the result is interpreted by laboratory directors and clinicians*"

By sequencing the exome of a patient and comparing it to normal reference sequence, variations in an individual's DNA sequence can be identified and related back to the individual's medical concerns in an effort to discover the cause of the medical disorder.

3.2 DNA Sequencing Data format

Once the sequencing is done, the data collected must be described using **well-known data formats** so that they can subsequently be used by other people without any problems.
However, as often happens, even for the sequencing there isn not a single universal format, but there are many differents (e.g.: FASTQ, SAM, BAM, etc.).

The sequencer that generates the data managed by our web-app provides the results of the sequencing using format **CSFASTA**, a *textual format* that also allows for sequence names and comments to precede the sequences.
A sequence in CSFASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. The word following the ">" symbol is the identifier of the sequence, and the rest of the line is the description (both are optional). It is recommended that all lines of text be shorter than 80 characters. The sequence ends if another line starting with a ">" appears; this indicates the start of another sequence.

An example of sequence in CSFASTA format is the following:

```
1 >MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
2 ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
3 FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
4 DIDGDGQVNYEEFVQMMTAK*
```

3.3 SNPs 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.

There are several databases that, nowadays, are used. The most important are:

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

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. ...