PEP-NET

European Commission

Predictive Epigenetics: Fusing Theory and Experiment

Why do we need bioinformatics?

written by Andrea Hita Ardiaca (ESR14) with input from Mihaela Pruteanu

---

The power of data

With every passing day, our world generates more data as all aspects of life are becoming digitized. The massive datasets, among many other purposes, help us understad back the world, its dynamics, our behavior…

It is also the case in Biology that massive datasets are being collected from many sources: patients with multiple diseases, organs during development, cells organizing in a body tissue and much more.

In epigenetics, our PEP-NET core field, a particular technology called “Next Generation Sequencing” has opened many doors to explore and characterize from large-scale data how epigenetics regulates cell function.

Next Generation Sequencing (NGS)

NGS is a technology which allows us to read the code from DNA fragments. Thus, if DNA from cells or tissues is isolated and chopped into fragments, we can determine its code! This is referred to as sequencing.

By adding a little creativity into the NGS technology and how we isolate our DNA fragments, many possibilities to capture epigenetic activity emerge. Do we want to capture the structure of the chromatin; when this is open or closed? We isolate the fragments of DNA associated to one status. e.g., open chromatin and determine what is the DNA code in chromatin open fragments assuming the non-present DNA code corresponds to closed regions. Do we want to study histone modifications or proteins bound to DNA? Again, we can isolate only the fragments of DNA that are around a particular histone modification, or a protein bound to DNA with specific sample treatments. Do we want to look at gene transcription? We can convert RNA to DNA and then sequence its code.

With all these possibilities, NGS experiments have become the gold standard for many applications such as gene expression and epigenetic modifications profiling.

Datasets resulting from NGS are large in volume and consist of short pieces of genetic code that need to be converted to meaningful quantitative information before drawing any insight. This task demands computer algorithms with capabilities to process such volumes of biological information, in a robust, accurate and semi-automatized manner (Figure 1). The need for specific data science algorithms oriented to biology has led to the emergence of Bioinformatics.

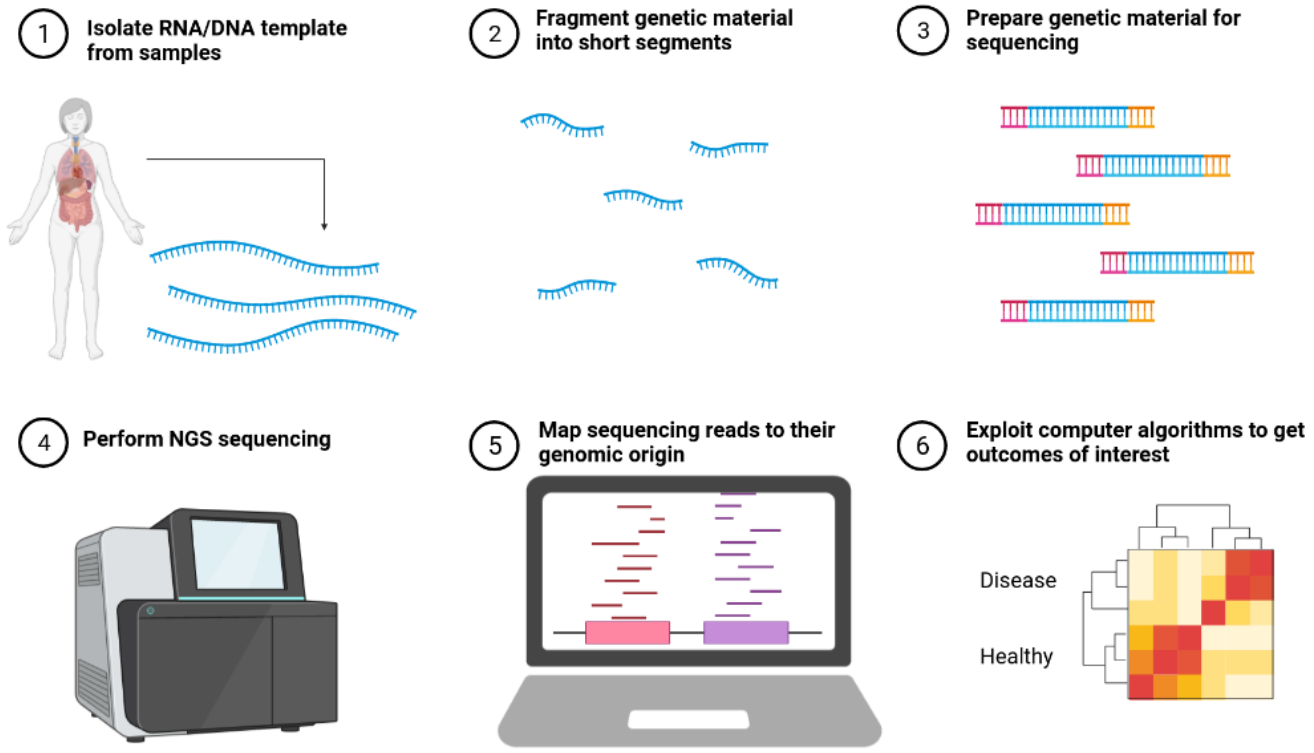

Figure 1. Next generation sequencing (adapted from “RNA sequencing”, by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates).

Bioinformatics

We refer to Bioinformatic methods as those who have the ability to exploit the power of computers to perform analytical tasks in large volumes of biological data.
Bioinformatics fuses three areas:

- statistics and probability theory, which live at the core of algorithms who process data;
- computer programming, which is the indispensable language to talk to the computer servers who actually perform the computational job;
- knowledge of the application-field, in our case epigenetics, which is key to build meaningful algorithms and interpret its results.

Bioinformaticians use these three ingredients to develop computer programs capable of turning large datasets like NGS data into information. In the case of NGS, data processing typically requires determining back the origin of DNA in the reference genome of the specie and then quantifying the level of DNA fragments in the position of interest. Beyond initial data processing steps, bioinformaticians have developed algorithms to answer many specific analytical questions such as:

- What cell populations are in our tissue sample?
- Can we define an epigenetic signature which can predict the prognosis of a certain cancer type?
- What epigenetic regulatory elements are involved in the conversion of a cell type to another?

In our PEP-NET network, we are interested in understanding epigenetics from a quantitative perspective. Data is a key ingredient to both quantitatively characterize a system of interest and to validate mathematical models with real evidence. For this reason, Bioinformatics is a pillar in our research!