Genomics at a glance – Part 1/2

Recently, Genomics has been getting spotlights upon the completion of the Human Genome Project in 2003 and a series of innovations in sequencing technologies. As NGS technology matures and the cost of sequencing is dropping fast, there is an explosion in the amount of data, and it elevates the necessity of computer infrastructures. Clearly, there is no doubt that NGS technology brought the "Gene-omics" revolution to the field and opened up a whole new HPC market. However, from a Biologist's viewpoint, there are some important aspects missed by media and Sci-fi movies.

Even if we were done with Genomics, but that is not all …

Life forms arose from the edge of Chaos. This is a statement that well describes the uncertainty and complexity of life forms. Unlike many other scientific fields, especially quantitative research areas, Biology is a science where the analysis and predictions were probabilistic. Most outcomes of biological research depend on hypothesis, speculations and their controlled environment. Nonetheless, the field of Genomics equipped with mighty sequencers still provide only a part of the entire story. As you can see from the central dogma illustration, Figure 1, genetic information flow does not stop at DNA or RNA level. The ultimate goal is to gain detailed insight on proteins, the final products of gene expression as well as work-horses for maintaining life. There are complex controlling mechanisms between each step in Figure 1. Although proteins are the final products, they also regulate transcription and translation at DNA and RNA levels. It is like a feedback-loop in a simple view, but the complexity of genomic system is way beyond a single line feedback-loop. They are actually multilayer feedback-networks intertwined such as transcriptional, post-transcriptional and translational regulatory networks. We still are quite far from the completion!

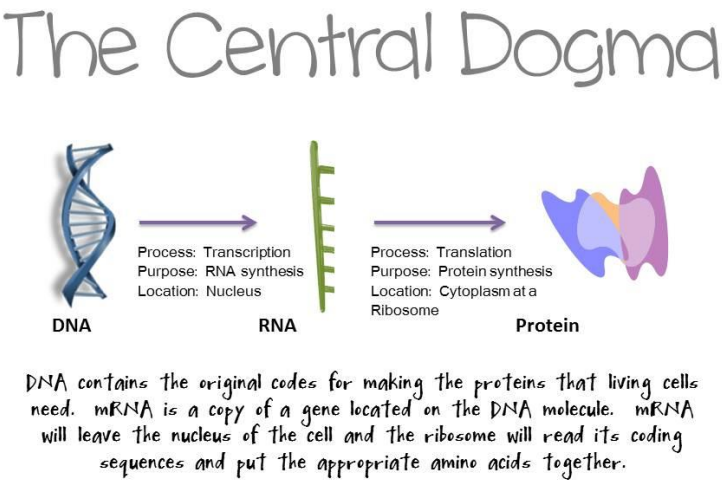

Figure 1 Central Dogma (Obtained from http://www.dbriers.com/tutorials/wp-content/uploads/2013/12/Central_Dogma-GODS-wikimedia.jpg)

All the recent excitement about NGS is reminiscent of gene therapy from the late 80s. The rush to apply gene therapy led to a series of failures caused tragic deaths and unexpected cancers and put the field into dark days. Without having complete and accurate knowledge, manipulating genes should be taken cautiously since we do not know yet what we do not know now.

The role of Next Generation Sequencing in Genomics and Medicine

Knowledge in DNA sequences is not only important in basic life science research, but also indispensable in biotechnology, medical diagnosis and forensic biology. NGS technology brings tremendous advances in those fields consuming DNA sequences as a key information: by lowering sequencing cost, shortening sequencing time and increasing its accuracy.

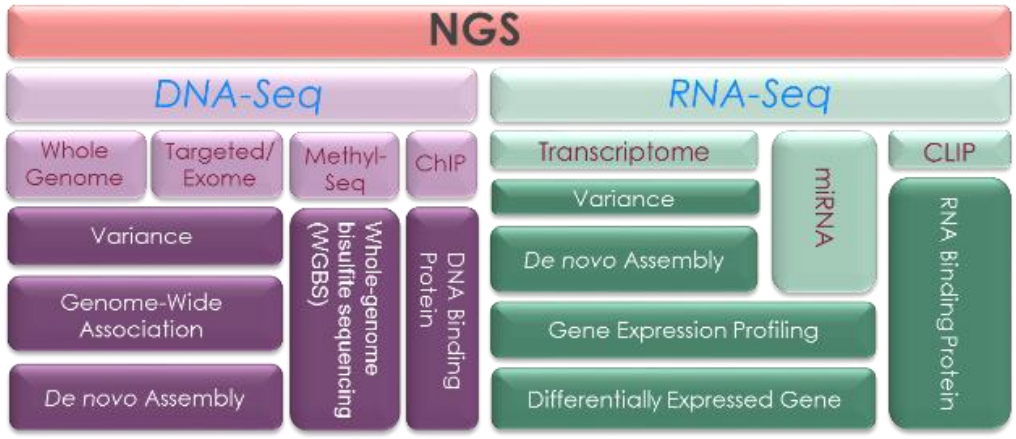

Figure 2 NGS usages in Genomics

In Figure 2, typical NGS applications based on the sequencing targets are listed. It is natural that these applications fall into two major groups, DNA-Seq or RNA-Seq. Most of the DNA-Seq studies are focused on genomic mutations or Single Nucleotide Polymorphism (SNP) in chromosomes while RNA-Seq studies are obtaining snapshots of the entire gene expressions at a given point in time. RNA-Seq rapidly replaces microarray technology and becomes a standard procedure for differentially expressed gene and RNA-Profiling studies. The ability to perform DNA-Seq and RNA-Seq at the same time for the same sample opened up a new opportunity in translational research area (Figure 3). Identifying mutations from genes

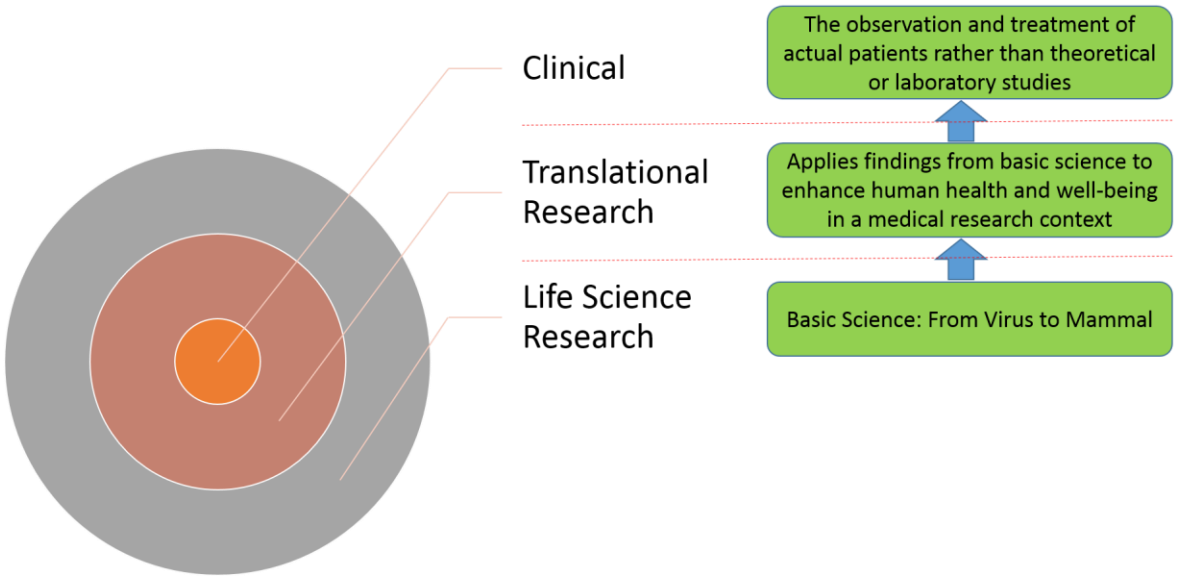

Figure 3 From basic life science to clinical

in chromosomes and knowing how those mutations change the gene expressions brings a tremendous advantage to completing the puzzle from 'the central dogma'. Although we still do not have comprehensive protein information, it is much easier to look into a small number of target proteins instead of testing large number of proteins. Currently, there are various projects to label genomic variations in humans according to their disease type, phenotypic characters, drug treatments, clinical outcomes and so on. This data integration efforts will provide the basis for precision medicine and personalized medicine in the near future.

NGS technology has brought a broad impact on the different layer of studies (see Figure 3), but the computational infrastructure requirements for these areas have not been correctly assessed. By the nature of basic life science and translational study, these two areas require the most flexible, powerful, and customizable infrastructure in terms of analysis phase in Figure 4. Large amount of samples and existing data are constantly regrouped and reanalyzed to elucidate biological phenomena in basic life science research while experts in translational research area seek any applicable knowledge toward to medical area. The data and workloads should be similar in basic life science research and translational research.

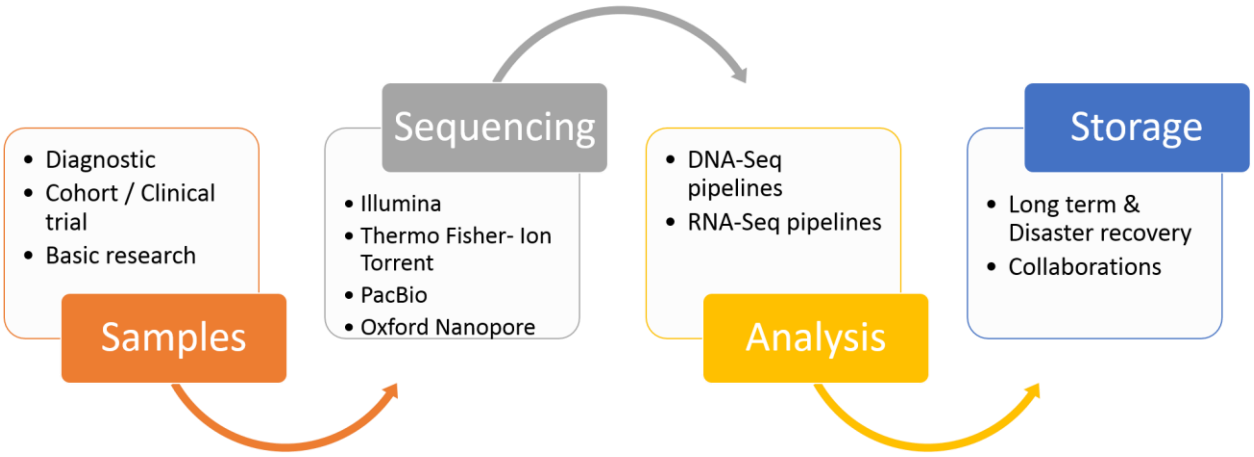

Figure 4 Overview of NGS data flow

However, translational research brings simple and consolidated methods which can be applied at the clinical area, to help making treatment decisions, or to provide any useful process that can improve human health and well-being. For an example, a set of key genes, say 10 genes, are identified as breast cancer progression in basic life science area. Experts in translational research pick that knowledge to develop method to suppress those identified genes, to stop or slow down cancer progression efficiently in different individuals through clinical trials or large cohort studies. As the results of the study, experts might figure out there are different subtypes of breast cancers which respond differently to various drug treatments based on the expression of those 10 key genes. This will allow development of a gene panel to detect such a subtype to go through different drug treatment options. Dealing with only 10 genes out of 20,000 genes will be a lot more applicable. As you can guess from this scenario, there are different

computational requirements among the different layer of studies. While high throughput is an important metric in basic life science and translational research areas, the speed of single sample processing time is more important in the clinical field.