Lecture 1 : Course Highlights

Monday 28 March 2016

Scribed by Anja Brandon and revised by the course staff

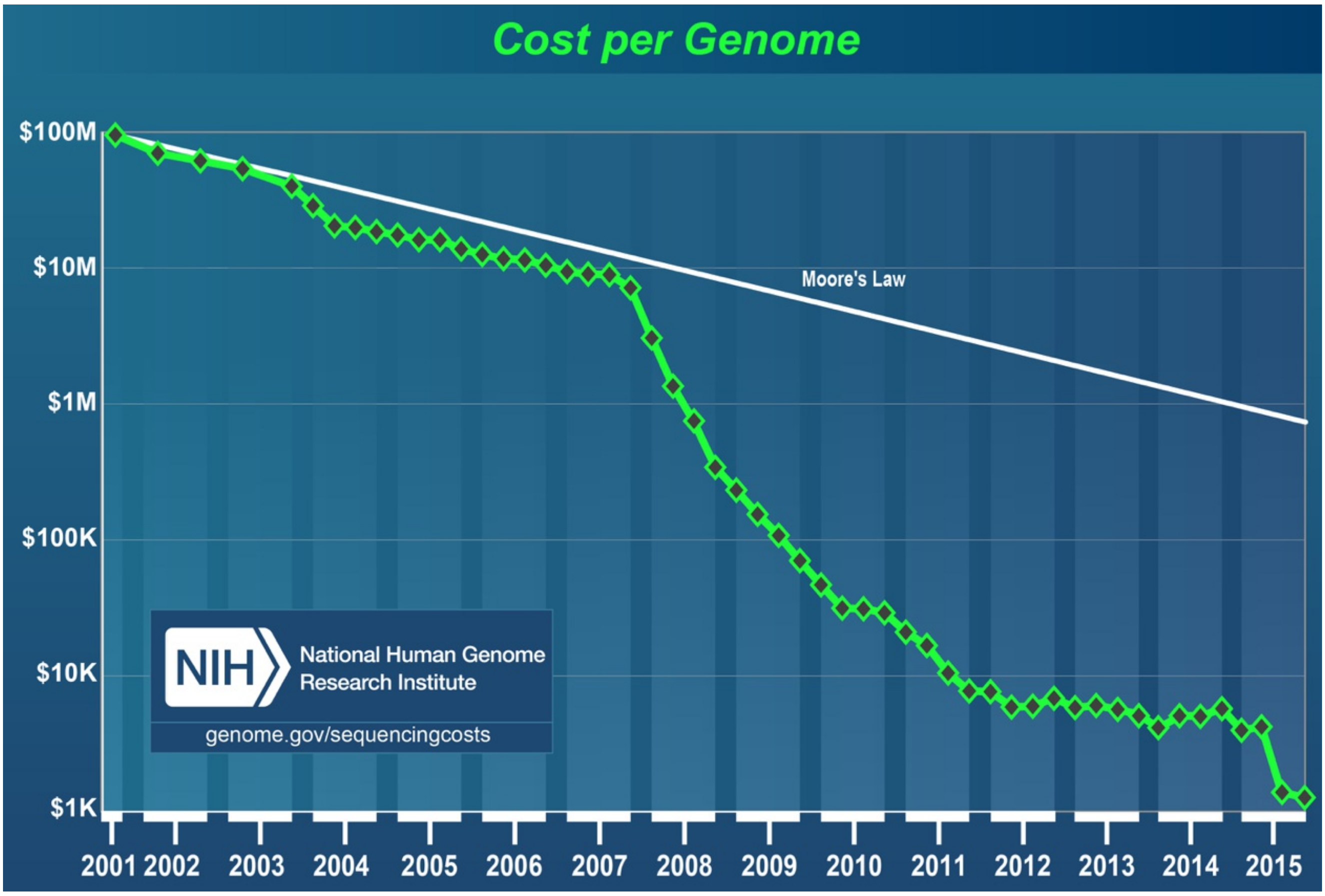
These notes are still quite unpolished. We'll be updating it in the next couple of days.

Need to understand high-throughput sequencing

The main object of interest in this course is thegenome of a organism. The genome of an organism is its genetic material which is usually made of deoxyribonucleic acid. All computational methods we discuss in this course will try to solve a problem which is related to deducing genome or some property that is quite close to the genome.

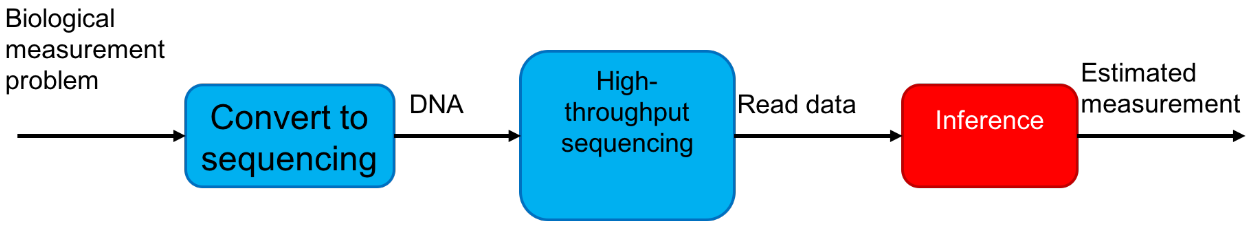
High-throughput sequencing is the technology to sequence the genome these days. Only 15 years ago, the the sequencing technologies were around 6 orders of magnitude slower than they are today (and more expensive by around 6 orders of magnitude too!!).

Interest in sequencing was first sparked by the Human Genome Project. This was a big consortium that got together to sequence the human genome in 1990. They released their first draft in 2003. It cost $2.7 billion, and 13 years of work by labs around the world to sequence the genome. In 2015, it cost around $1000 to sequencing a genome. This is testament to how far the technology has evolved. As shown below, the cost of DNA sequencing has been falling at a rate faster than Moore's law over the last 15 years.



Cost of DNA sequencing over the years.

DNA is a very important biological molecule, but it’s only one of many important biological molecules. Other important biological molecules include ribonucleic acids and proteins. Some innovative bio-chemistry has allowed the use of DNA sequencing technology to measure properties of various other biological molecules (and there are proposals to detect dark matter using this). The basic paradigm is to reduce the estimation problem of interest to the a DNA sequencing problem, which is then sequenced using high-throughpt sequencing. This is similar in principle to the reduction used to solve many mathematical problems, or to show NP-hardness of various problems. This is illustrated below.



The -seq paradigm: Convert problem of interest to DNA sequencing problem and solve that.
One analogy would be to think of high-throughput sequencing to be a piece of equipment (like a microscope, say) that can be used to measure a variety of quantities. The challenge to the bio-chemists is to covert the problem of interest to them to a problem which can be measured using high-throughput sequencing (just like biologists have to design experiments such that the results can be observed under a microscope). The challenge to the computational biologist is to do the relevant type of inference on the data observed using high-throughput sequencing. Some important sequencing assays are:

- RNA-Seq: RNA important intermediate product for producing protein from DNA. Every cell in a human has the same DNA but very different RNA. RNA in cells also shows a lot of variation over time, and the environment. RNA-Seq is an assay to "measure" RNA.
  This was the first assay in which high-throughput sequencing was used to measure a molecule other than DNA. It was developed in 2008 by Mortazavi *et al*.

- ChIP-Seq: The difference in RNA across cells is to a large extent due to the fact that DNA in cells are bound to proteins called histones. In different cells different parts of the genome are bound to histones. Only parts of the DNA that are not bound to histones get converted to RNA. This is an assay which was developed measure the regions of the genome that are bound to histones in cells. This was developed in 2007 by Johnson *et al*. Another recent assay calledATAC-seq measures regions of the genome that are *not* bound to histones.

- Hi-C-Seq: This is an assay used to measure measure the 3D structure of molecules. This was developed by Belton *et al* in 2012.

One of the most interesting and important problems is predicting the phenotype (physical characteristics like a person's height, or if he/she likes the colour red) from the person's genotype (DNA sequenced). This is important in medicine to predict susceptibility to diseases. A big success-story here is the discovery of that presence of a particular mutation in the gene BRCA1 increases the risk of breast cancer to around 45%.

Another important application of high-throughput sequencing iscancer. Cancer is a "disease of the genome". It is caused by rearrangements of the genome (which are sometimes very large). By sequencing cancer cells, one gets information about the nature of the cancer-causing mutation, and tailor treatment to that.

Non-invasive pre-natal testing for genetic birth defects is another very interesting application of high-throughput sequencing. There are traces of foetal DNA in the maternal blood. The main idea here is to sequence the maternal blood, and try to infer foetal genetic birth defects from it. This has been used successfully to detect Down's syndrome.

What is High-throughput sequencing?

Science is basically progressed by the invention of measuring methods. High-throughput Sequencing is one such measurement tool. However, high-throughput Sequencing is different from many measurement tools in the fact that it has a significant computational component. High-throughput sequencing (also called *shotgun sequencing*) takes the DNA sequence as input, breaks it into smaller fragments and returns a noisy version of these smaller fragments, called *reads*. We note that the length of reads range from 50-50,000, while the human genome is of length 3 billion. Fortunately these small noisy subsequences also contain information about the genome. Extraction of this however needs clever computational processing. This is the flavour of the problems we will discuss in this class.

As a single read is very short, it contains very little information about the entire sequencing. However, a typical experiment generates a few million reads (and hence is called "high-throughput"). We also note that the sequencing process can be very noisy. Each of the reads can be potentially different from the original subsequence of the DNA the read came from.

The sequencing revolution is because of rapid evolution of sequencing technologies. Sequencing began with Fred Sanger who first came up with the Sanger sequencing technology. This was a very low throughput technology and was the dominant technology till the late 1990s. Second generation sequencing was pioneered by Illumina and is the dominant technology currently. Recent developments in this have allowed scientists to sequence individual cells. This is called single-cell sequencing. There has been recent developments leading to third and fourth generation sequencing that has been pioneered by companies like PacBio and Oxford Nanopore.

High-throughput sequencing is a fast changing are with new technologies emerging constantly. All these technologies give us reads, but are use very different chemical processes to generate them. There are two main properties of reads that are important from a computational perspective

1. *Read lengths* The longer the reads are the more information they contain. Ideally one would want a read to be the entire genome. Sadly, this does not seem to be achievable by the chemistry currently and in the foreseeable future. Illumina reads are around 100bp-200bp long depending upon the specific machine. PacBio reads are of lengths >10000 bp. These are much longer than the Illumina reads but yet are much shorter than the genome lengths.
2. *Error rates and types of errors* The rates of errors in the reads are also important from a computational perspective. Illumina has low error rates 1-2%, with the errors being mostly substitution errors (*i.e.* a base being replaced by some other base). PacBio reads on the other hand have error rates of 10-15% with the primary forms of errors being insertions and deletions.

The figure below shows the characteristics of different sequencing technologies.

Sequencer	Sanger 3730xl	454 GS	Ion Torrent	SOLiDv4	Illumina HiSeq 2000	Pac Bio
Mechanism	Dideoxy chain termination	Pyrosequencing	Detection of hydrogen ion	Ligation and two-base coding	Reversible Nucleotides	Single molecule real time
Read length	400-900 bp	700 bp	~400 bp	50 + 50 bp	100 bp PE	>10000 bp
Error Rate	0.001%	0.1%	2%	0.1%	2%	10-15%
Output data (per run)	100 KB	1 GB	100 GB	100 GB	1 TB	10 GB
Approx cost per GB		10,000	1000	100	10	1000

Characteristics of different sequencing technologies.

Data science of high-throughput sequencing

The success of high-throughput sequencing is mainly due to the creative use of read data to solve various problems. This involves solving many data-science problems.

Any data science problem can be divided into three stages:

1. *Processing the data* : Examples of problems of this type in high throughput sequencing include problems like:
2. Assembly or *de novo* assembly: Recovering the original genome from short noisy reads.
3. Variant calling: Individuals of the same species have very similar genomes. For example, any two humans share 99.8% of their genetic material (which means they differ in about 6 million bp). As we have a human reference, very often scientists just want to know the differences of an individual from this reference genome. The problem of inferring this is called the variant calling problem.
4. Phasing: The chromosomes in humans (and other higher animals) come in pairs. These are sequenced together. Very often scientists want to separate the sequence on the two chromosomes. This is called the phasing problem.

5. Quantification: RNA is an important biological molecule in cells, as discussed above. There are 10s of thousands of types of RNA molecules observed in a cell. Unlike DNA where there is just two copies of a molecule in the cell, there are many copies of the each RNA molecule in the cell. Scientists are interested in estimating how many copies of RNA of each type are in a cell.

6. *Managing the data* : These involve handling large data bases which raise problems like:

7. Privacy

8. Compression

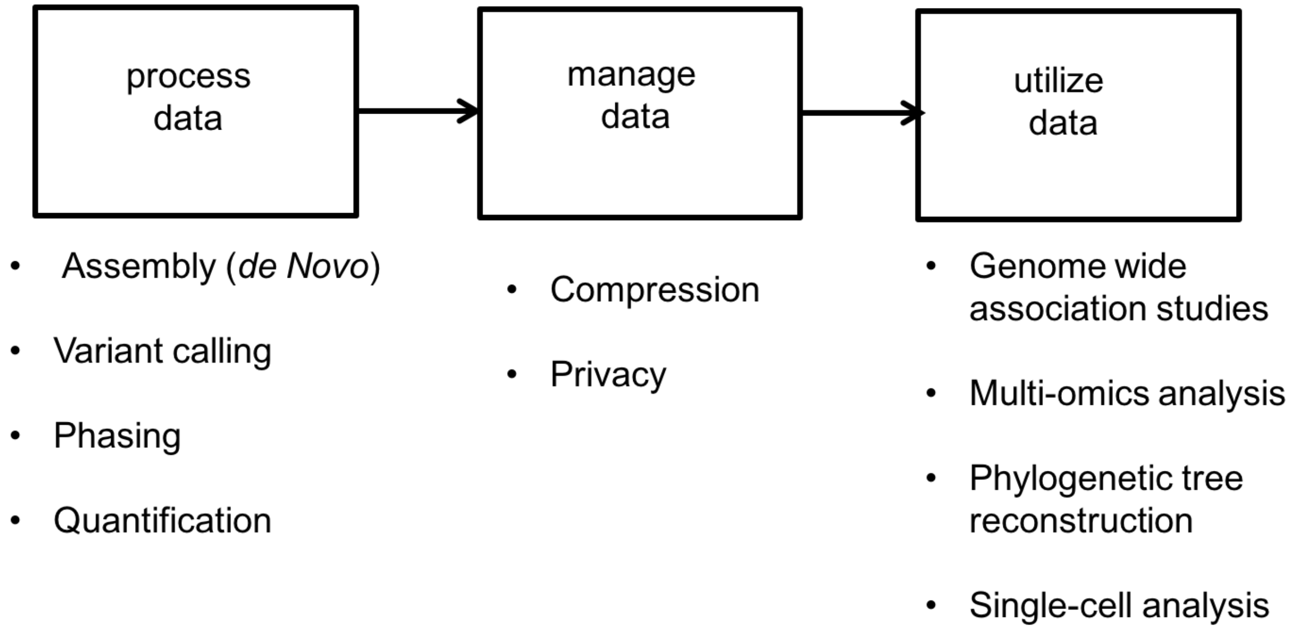
9. *Utilising the data* : This is basically using the data to make useful infrence. This includes problems like:

10. Single-cell analysis: This involves analysing single-cell data to infer properties like diversity in cell populations.
11. Genome Wide Association Studies (GWAS): This basically involves obtaining association between

genomes and various characteristics of individuals.

12. Multi-omics data analysis: This involves using data from DNA, RNA, and protein measurements to make predictions on characteristics of individuals.

These different problems are illustrated below:



Data science of High-throughput sequencing.

Tools used

- Combinatorial algorithms: Problems like genome assembly involve working on combinatorial objects like graphs and use of many interesting combinatorial algorithms.
- Statistical Signal Processing: This is necessary for dealing with noise in data.
- Information Theory: As one has to do a lot of inference, knowing how much data is necessary to be able to have good estimates is of paramount importance.
- Machine Learning: There are lots of interesting connections between processing high-throughput sequencing data and machine learning.

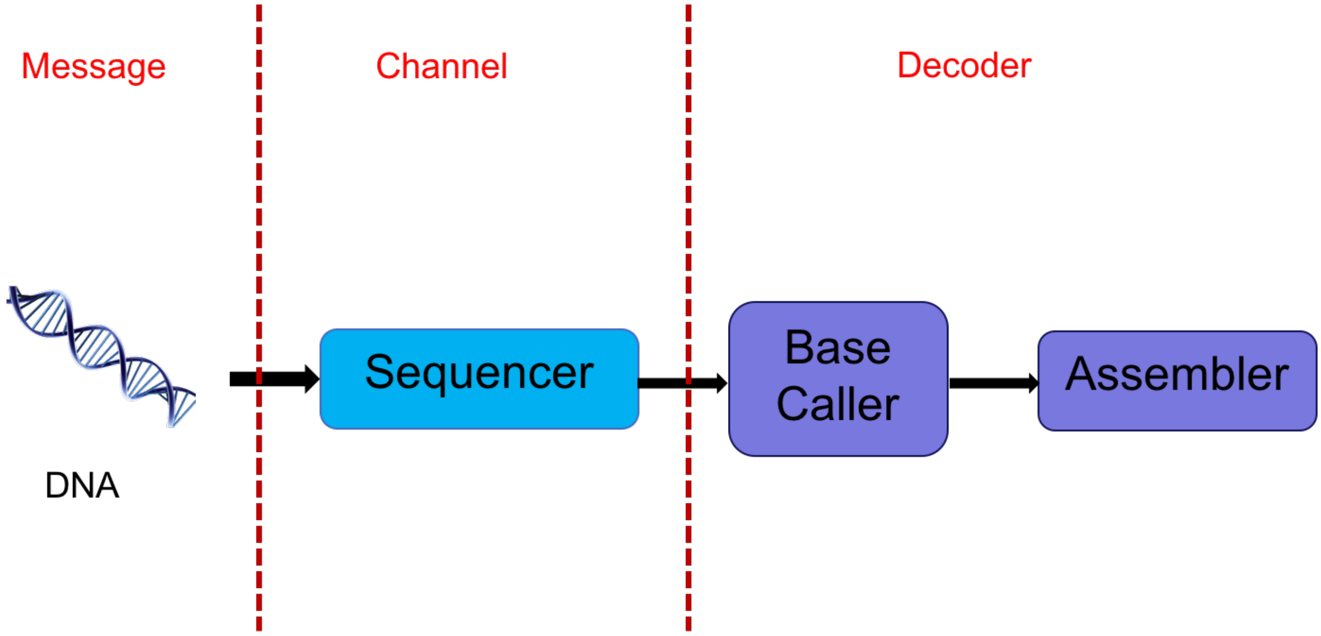
The main method of working with high-throughput sequencing data is first trying to model the data. This usually involves many assumptions which are not true in practice. Then these models are used to come up with interesting algorithms. These algorithms are then tried on real data. As real data does not satisfy these assumptions, it takes some effort to get these algorithms working on real data, even when the modelling is reasonable.

Two Examples

In this section, we discuss two representative problems that will be covered in this course.

DNA-assembly

The DNA sequencing machine outputs an analog signal (this may be light signals, or electric signals depending upon the technology). We want to process this signal to get the genome. In essence, one could think of the DNA as a message, the sequencer as a bad channel, and the base caller and assembler as the decoder. This abstraction is shown below:



DNA assembly as a message decoding problem.
This gives us multiple levels of thinking about problems here. At the level of analog signals (from which we want to extract digital information), we have a statistical signal processing problem. We have to estimate discrete bases from analog signals. This involves various stochastic models with many parameters (which may need to be estimated). Further one also has to often have to deal with signal from adjacent bases interfering with each other. This inter-symbol interference is also a signal processing problem.

We can also think of the problem of assembling the genome from the reads obtained after processing the analog signals. The first order of business here would be to get an estimate of the number of reads necessary to be able to assemble with reasonable accuracy. This involves using tools from information theory, which identify bottlenecks and give us design principles to deal with them. Then one has to design efficient algorithms to overcome these bottlenecks. By efficient here we mean $$\Omega(n^2)$$, where $$n$$ is the number of reads. In general, the scale of data makes any super-linear algorithm unfeasible in most cases. However, there are cases where smart algorithm design, and low level optimised software allows one to use upto $$O(n^2)$$ algorithms.

Single-cell RNA quantification and analysis

As discussed above, RNA is another important biological molecule. One can think there existing around 10,000 RNA sequences observed in cells, each of which are ~1,000-10,000 bp long. There are many copies of the same RNA molecule (known as a transcript) in a cell. Biologists are interested in estimating the number of RNA transcripts in a cell.

Biologists and chemists have figured out ways to convert RNA back into DNA (mainly using an enzyme *reverse transcriptase*), and then sequence the DNA to get reads using shotgun sequencing. The computational problem is trying to estimate the number of transcripts of each type from these reads.

One often uses the reference of known transcripts observed in an organism, known as a *transcriptome*. Even with this, the problem is interesting as many transcripts have common subsequences. Hence, one can not always be sure of where a read originates from. A very nice algorithm for solving this problem is the expectation-maximisation. This is a tool used to solve this problem.

In a bulk experiment, biologist take a tissue (100s of millions of cells), crush it and get shotgun sequencing reads from the mixture of the RNA of all the cells. So the transcript counts (or abundances) obtained are an estimate of sum over all cells. Recently, it has been possible to use single-cell sequencing to get reads of RNA from different single cells. Biologists are interested in estimating diversity of types of cells in the tissue using this. This leads to a lot of interesting questions, like if processing single-cell reads to get single-cell counts and if clustering using them is the right thing to do.

Slides