Introduction to NGS

Simon Rasmussen
Associate Professor
DTU Bioinformatics
Technical University of Denmark
2018
Life science data deluge

- **Massive unstructured data** from several areas: DNA, patient journals, proteomics, imaging, ...
- Impacts Industry, Environment, Health
- Societal grand challenges
- Cheap sequencing technologies results in **explosion** of DNA data
DNA sequencing

Reading the order of bases in DNA fragments
Why NGS?

Transforming how we are doing biological science (and bioinformatics) by allowing experiments that could not have been done before, and perform experiments much faster.
by producing **massive** amounts of sequence data, really **fast**
First generation: Sanger

- Fragment DNA
- Clone into plasmid and amplify
- Sequence using dNTP + labelled ddNTPs (stops reaction)
- Run capillary electrophoresis/gel and “read” DNA code
- Low output, long reads (~800-1200 nt), high quality
able future at least, a high-coverage sequence of the normal genome are not. In most cancer genomes these rare germline variants far acquired variants in a cancer genome are sobering. To obtain a one high-coverage cancer genome sequence has recently been detection of subclones carrying drug-resistance mutations.

Until recently, this was an unattainable fantasy. However, the improvements in the rate of DNA sequencing over the past 30 years have been enormous. In 1977, Sanger introduced the chain-termination method for DNA sequencing, allowing the determination of a single molecule. However, the rate of sequencing was very slow, with only a few thousand to tens of thousands of base pairs generated per day. This rate increased significantly with the advent of capillary electrophoresis in the 1980s, allowing for the sequencing of a few million base pairs per day. The 1990s saw the introduction of second-generation sequencing technologies, such as pyrosequencing and capillary sequencing, which allowed for the sequencing of billions of base pairs per week. These technologies have been further improved over the years, with the arrival of third-generation sequencing methods such as single-molecule real-time (SMRT) sequencing and nanopore sequencing, which allow for the sequencing of a single molecule.

The proposal to sequence large numbers of cancer genomes has been a topic of discussion. Approximately 100,000 somatic mutations from cancer genomes have been reported in the quarter of a century since the first somatic mutation was identified. However, the full range of somatically acquired genetic alterations in cancer, including all point mutations, copy number changes, and epigenetic modifications, will require a comprehensive sequencing approach. To address this challenge, the International Cancer Genome Consortium (ICGC) was established in 2007, with the mission of coordinating the sequencing of thousands of cancer genomes to identify the complete set of somatically acquired mutations in different types of cancer. The full sequencing of 20,000 cancer genomes or more is required to identify the driver mutations and thus all the cancer genes. In a clinical context, such data may inform decisions about chemotherapy and determine the patients who are most likely to respond. The costs of sequencing are likely to be between 1 and 10 billion, but the potential translational implications are profound. These data may harmonize the product, maximize use of resources and optimize clinical care.

There is, therefore, much work to be done over the next few years. There is a need to determine the complete set of somatically acquired mutations in different types of cancer, which will require sequencing of hundreds of cases for each type. This effort will be expensive and, to some extent, we cannot predict what will be found. However, the human genome is finite. Therefore, with further technological advances in DNA sequencing that are already in sight, we have to coordinate the work internationally to maximize use of intellectual property, sample quality, clinical annotation, data quality, data storage and sequencing completion. Most importantly, given the demanding nature of the task, the ICGC will coordinate studies to minimize duplication of effort and enable the most parsimonious and harmonized approach.

To coordinate the ICGC effort, a Cancer Genome Atlas will be established for each of the 100 cancer types. For each such type, a project will be established, as described in Box 1. These projects will provide the highest possible sequence quality and will harmonize the product. This is the mission of the International Cancer Genome Consortium (ICGC), which has been established in 2007 as a multinational, collaborative initiative involving the world’s leading cancer genome research groups.

The ICGC seeks to comprehensively characterize the somatically acquired genetic alterations in cancer, including all point mutations, copy number changes, and epigenetic modifications. The aim of ICGC is to comprehensively characterize somatically acquired genetic alterations in cancer, including all point mutations, copy number changes, and epigenetic modifications. The ICGC will sequence a large number of cancer genomes to identify the complete set of somatically acquired mutations in different types of cancer. The full sequencing of 20,000 cancer genomes or more is required to identify the driver mutations and thus all the cancer genes. In a clinical context, such data may inform decisions about chemotherapy and determine the patients who are most likely to respond. The costs of sequencing are likely to be between 1 and 10 billion, but the potential translational implications are profound. These data may harmonize the product, maximize use of resources and optimize clinical care.
able future at least, a high-coverage sequence of the normal genome will outnumber the somatic mutations present. Therefore, for the foreseeable future, the parameters of experiments to catalogue all somatically acquired variants in a cancer genome are sobering. To obtain a complete catalogue of somatic mutations from an individual human cancer genome can be obtained, in billions of bases of DNA sequence per week, yields that are predicted to arrive of second-generation sequencing technologies promises a new era for cancer genomics. These platforms currently generate new era for cancer genomics. These platforms currently generate billions of bases of DNA sequence per week, yields that are predicted to arrive of second-generation sequencing technologies promises a

Improvements in the rate of DNA sequencing over the past 30 years and into the future.

Figure 3

1977 - Sanger Chain-termination method

Human genome

10X Genomics

Illumina

Solid

Ion Torrent

Pacific Biosciences

Oxford Nanopore

Complete Genomics

1st generation to NGS

Stratton et al., Nature 2009
Sequencing costs

- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady.

- DNA sequence data **is doubling faster than computer speeds**!
Human sequencing

- First draft genome of human in 2001, final 2004
- Estimated costs $3 billion, time 13 years

Today:
- 1-2000$ for one genome
- A couple of days!
Storage and analysis

Highest cost is (almost) not the sequencing but storage and analysis

A standard human (30-40x) whole-genome sequencing exp. would create 150 Gb of data

One Illumina HiSeq X Ten system: 18,000 human genomes per year!
Distributed data production

World wide
>900 centers

>60 Pb pr
year (2014)

Data transfer and storage becomes an issue

http://omicsmaps.com
The X Genomes projects

- Human population projects
  - 1000 genomes project (2500 individuals)
  - Genomics England (100k individuals)
  - US Precisions medicine (1 million individuals)
- 100K pathogens project, Earth Microbiome project, Cancer genome project, Plants and animals, Insects,…
NGS in the clinic

- Diagnostics of patients (+ fetus)
- Focused treatment of cancer patients
- Sequencing of bacterial isolates
- Country-wide projects:
  - UK, US, Qatar, Finland, China, …
  - DK: Danish regions want to sequence 100k individuals
• Personalized medicine initiative in DK

• Up to 70% of medicine does not work!

• Sequence 100,000 patients on hospitals

• Use extensive registry data

• Current: 100 mill DKK (estimated 2 bill DKK)

• National Genomics Institute
• Personalized medicine initiative in DK

• Up to 70% of medicine does not work!

• Sequence 100,000 patients on hospitals

• Use extensive registry data

• Current: 100 mill DKK (estimated 2 bill DKK)

• National Genomics Institute

Ethical debate?
NGS & Bioinformatics

- Extreme data size causes problems
  - Just transferring and storing the data
  - Standard comparisons fail (N*N)
  - Standard/old tools can not be used
  - Think in fast and parallel programs
What can we use it for?

- Whole genome re-sequencing
- Population genomics
- Diagnostics
- Cancer genomics
- Ancient genomes
- Metagenomics
- RNA sequencing
- Single cell sequencing
- Genomic Epidemiology
- Anything with DNA
How it works?