The Human Genome Project –
Its History and Advancements into New Research
and Technology

Jutta Marzillier, Ph.D
Lehigh University
Biological Sciences

September 5th, 2014
Objectives

Techniques that enabled genome sequencing

Process of Human Genome Sequencing

New DNA sequencing Technologies

Some research spinoffs
Guess, who turned 60 last year !!!
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

W is it possible to suggest a structure for the salt of deoxyribonucleic acid (DNA)? This structure has novel features which have considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey. They claim that their model cannot be improved on in advance of publication. Their model contains three intertwined strands, with the phosphates near the outside, and the bases on the inside. In our model, this structure is maintained for two reasons:

1) We believe that the material which gives the X-ray diagram is the base, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other.

2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Perutz in the model. In his model the phosphates line on the outside and the bases are linked by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribonucleic acid. This structure has two helical chains each wound around the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphorus-dioxide groups joined by 3'-deoxyribose residues with 3'-5' linkage. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions.

Each chain loosely resembles Perutz's model No. 1, that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms through it is close to Pauling's standard configuration, the sugar being roughly perpendicular to the attached bases. There is no intercalation of each chain every 3.4 A. in the structure. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are parallel to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pairs must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 6; purine position 2 to pyrimidine position 5.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms that is, with the keto rather than the enol configuration, it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and cytosine (purine) with guanine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on those chains the other member must be thymine: similarly for guanine and cytosine. The sequence of bases on a single chain, does not appear to be restricted in any way. However, if only specific pairs of bases can be formed it follows that if the sequence of bases on one chain is given, then the sequence of the other chain is automatically determined.
Guess, who turned 10 last year !!!
Human Genome Project

2001 Draft Human Genome Sequence
2003 Finished Human Genome
   (50 years after DNA structure solved)

Two techniques published in 1977 by

- Sanger et al.
  DNA sequencing by chain termination or dideoxysequencing
- Maxam & Gilbert
  DNA sequencing by chemical modification

The original method of Sanger sequencing and multiple improvements regarding chemistry and computation lead to complete sequencing Human Genome (and many others).

Sanger sequencing is expensive and Next-Generation-Sequencing (NGS) technology took its place.
Development Sequencing Time Line

- 1953: Discovery of the DNA double helix by Watson & Crick
- 1956: Development of the "chain termination" method by Sanger
- 1975: Sanger sequencing bacteria
- 1982: Nature publish the automated DNA sequencer
- 1990: The Human Genome Project begins
- 1995: Haemophilus influenzae sequenced
- 1996: S. cerevisiae sequenced
- 1997: Human Genome Project and Celera publish drafts of the human genome
- 2000: D. melanogaster and Arabidopsis thaliana sequenced
- 2001: Human genome sequenced
- 2002: Mouse genome sequenced
- 2003: Human Genome Project finishes
- 2005: GenBank exceeds 1 billion bases
- 2006: C. elegans sequenced
How many base pairs (bp) are there in a human genome?

How much did it cost to sequence the first human genome?

How long did it take to sequence the first human genome?

When was the first human genome sequence complete?

Why is the Human Genome information important?

to serve as a reference for human disease
How many base pairs (bp) are there in a human genome?

~ 3 billion bp (haploid)

How much did it cost to sequence the first human genome?

How long did it take to sequence the first human genome?

When was the first human genome sequence complete?

Why is the Human Genome information important?
How many base pairs (bp) are there in a human genome?

~ 3 billion bp (haploid)

How much did it cost to sequence the first human genome?

~ $3 billion

How long did it take to sequence the first human genome?

When was the first human genome sequence complete?

Why is the Human Genome information important?
How many base pairs (bp) are there in a human genome?

~ 3 billion bp (haploid)

How much did it cost to sequence the first human genome?

~ $3 billion

How long did it take to sequence the first human genome?

~ 13 years

When was the first human genome sequence complete?

Why is the Human Genome information important?
How many base pairs (bp) are there in a human genome?

~ 3 billion bp (haploid)

How much did it cost to sequence the first human genome?

~ $ 3 billion

How long did it take to sequence the first human genome?

~ 13 years

When was the first human genome sequence complete?

~ 2000-2003

Why is the Human Genome information important?
How many base pairs (bp) are there in a human genome?

~ 3 billion bp (haploid)

How much did it cost to sequence the first human genome?

~ $3 billion

How long did it take to sequence the first human genome?

~ 13 years

When was the first human genome sequence complete?

~ 2000-2003

Why is the Human Genome information important?

to serve as a reference for human disease
Why is Genome Sequencing Important?

- To obtain a ‘blueprint’ – DNA directs all the instructions needed for cell development and function
- DNA underlies almost every aspect of human health, both, in function and dis-function
- To study gene expression in a specific tissue, organ or tumor
- To study human variation
- To study how humans relate to other organisms
- To find correlations how genome information relates to development of cancer, susceptibility to certain diseases and drug metabolism (pharmacogenomics)
- Outlook: Personalized Genomics (David Church, Harvard)
The establishment of various sequencing technologies gave rise to many new fields in biology, medicine and engineering

- Identification of genes contributing to disease
- synthetic biology
- personalized medicine
- genomic screening
- ........
Outline

DNA Sequencing

Biological Dogma and Principle of DNA synthesis

Sanger sequencing and its improvements

Next-Generation-Sequencing (NGS) Technologies

3rd Generation Technologies

Projects and ‘Spinoffs’
The human body has about 100 trillion cells with more than 200 different cell types. Each cell harbors the same genetic information in its nucleus in form of DNA containing chromosomes. Depending on cellular, developmental, and functional stage of a cell only a subset of genes is expressed.

Why is the Knowledge about the Human Genome interesting?

http://www.genomenewsnetwork.org/articles/06_00/sequence_primer.shtml

http://www.pharmainfo.net/files/images/stories/article_images/
What is DNA?

DNA constitutes the heritable genetic information that forms the basis for the developmental programs of all living organisms.

A genome is an organism’s complete set of DNA.

The DNA is made up of four building blocks called nucleotides.

DNA sequence is the particular side-by-side arrangement of bases along a DNA strand.

DNA sequencing is a biochemical method to determine the sequence of the nucleotide bases that make up the DNA.
Principle of DNA Synthesis

Arthur Kornberg demonstrated DNA replication in a cell-free (*in vitro*) bacterial extract (Nobel prize, 1959)

- Discovered *DNA polymerase* (Pol1) to facilitate DNA synthesis
- Unraveled the *mechanism* of DNA synthesis
  - nucleotide building blocks
  - a single DNA strand serves as a template
  - can only extend a pre-existing chain (*primer*)
  - a free 3’ Hydroxyl end is required

Watson et al., MGB, 2008
The 3’ hydroxyl group of the primer attacks the α-phosphoryl group of the incoming nucleotide thereby forming a phosphodiester bond ($S_N2$ reaction).

Watson et al., MBG, 2008
Dideoxy Sequencing according to Sanger

Both nucleotide types can be incorporated into growing DNA chain.

Presence of dideoxy-cytosine in growing chain blocks further addition of incoming nucleotides.

Watson et al., MBG, 2008
How was Sanger Genome Sequencing Done?

Clone by Clone

Create a crude physical map of the whole genome by restriction mapping before sequencing.

Break the genome into overlapping fragments and insert them into BACs and transfec into *E.coli*.

• need known flanking region to anneal primer

Shotgun sequencing

Break genome into random fragments, insert fragments into vector, sequence each of the fragments and assemble fragments based on sequence overlaps.
Dideoxy Method of Sequencing (Sanger, 1975)

DNA synthesis is carried out in the presence of limited amounts of fluorescently labeled dideoxyribonucleoside triphosphates that results in chain termination. Through chain termination, fragments of distinct sizes are generated, which can be separated by gel electrophoresis.

fasciaworld.com.23/dideoxy-chain-termination
Separation of Sequencing Fragments by Capillary Gel Electrophoresis

Capillary gel electrophoresis: Samples passing a detection window are excited by laser and emitted fluorescence is read by CCD camera. Fluorescent signals are converted into basecalls.

High resolution
Read length up to 1,000 nucleotides
Use of heat-stable *Taq* Polymerase Enabled Automation of Sequencing Reaction

Kary Mullis invented the Polymerase Chain Reaction
High-Throughput Whole Genome Sequencing

Analysis of 384 sequencing reactions in parallel

George Church, Scientific American, January 2006, pp47-54
Assume more than 20 384-capillary sequencers running simultaneously approx. 1000 bp per capillary run approx. 3 hours per run

Approx. 60 Million bases per day in one facility
> Sequence,
TGCGGCTGCCAGATTTTGTAACGGGTTTGGAAATCGACGGGAGAAGACAGCGCGG
GCCTAGAAGGCCCCCGTAATGCCCCCTCTAGAAGGCACCGGACGAGACGGCAACACGGTG
CGGATCGATAGATGGCACCGGAGACAAGCGAAGACGGCCGCAGAGCCGTCGC
CGGCTGACGCCCCTGAGAGCCCCGTAGACGGACGAACGGTGCGGATCGATAGATGGCACCGGAGACAAGCGAAGACGGCCGCAGAGCCGTCGC
TGAAACGCGAAAGTGGAAGGTTCCTTACCTATGGAGGGGTAAGGGAGCGAGCT
CCAGCGAGCGACCACCAGATAGGTTCTTGTCCGGGATGACTGAACGGAG
AGAGACTACCCCTTATTAGCGACCTCCGGCTCGCCAGGTAGGTACCGAACGATGA
GAGAGGTACCTAGACCCGTACAGGCCGGGGTTTATCCCCCGGCCGTTACAGCATG
GTCATTTTGGGTAGGTACGTTATCGTAAGCATCCTACTCAACCAACAGAACCAGATGGTT
ACGTAACCGGGGTACGTTACGTAACGTGAACTAGATCGTAACGAAATGGGAAAGATC
GCCGCCCAGGCGGCGCCCGCAGCGGGMGTACGTTTTGTCAGGTATGTCACTAGGG
AGGGGTGAATCTAGATCTAAGGAGCATCAGAAGGAGCTTCTTCGTCTACTGCTC
TGCTAGATGGATAGCTGATGACGATAGCTGACGCTACTCTCGAAGGTGGAGACCATC
CACCTCATCCAGTCCTTCCAGCGGAGAACTCGATAAGTGAGCTGGAGATCATC
TGACCAGCGCGATCGTCTGCCGACGACTCCGCATCCGCCGCGAGGGTTTC
TGCCGGGGCTGGCCACCGCTGCCAGATCCGCTACCGCGACGAAGGCGTCCCATC
AGGTGTACGTGCTCGACGAGGCTAGC

Primary sequencing reads:

cagacctgtcgtcgtcgcgtatatactgagctagtcgact
tagctagccggatatattaccagacgtgtcgtcgcgtactgata
cctgatcgtatctgcgtatagctagccggatatattaccagacgt

Align sequences for homologies

Goal for Genome Sequencing:
Sanger: 8-fold coverage
454: 30 fold coverage
What did we learn from HGS?

- Less than 2% of the Human Genome codes for protein
- Encodes for approx. 20,000 protein-coding genes
- The human genome sequence is almost exactly the same (99.9%) in all people
- The genome size does not correlate with the number of estimated genes

Humans have only about twice the number of genes as a fruit fly and barely more genes than a worm!

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Size</th>
<th>Est. Gene #</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>4.6 Mb</td>
<td>4,400</td>
</tr>
<tr>
<td>Yeast</td>
<td>12.1 Mb</td>
<td>6,200</td>
</tr>
<tr>
<td>Roundworm</td>
<td>97 Mb</td>
<td>19,700</td>
</tr>
<tr>
<td>Fruit Fly</td>
<td>180 Mb</td>
<td>13,600</td>
</tr>
<tr>
<td>Rice</td>
<td>389 Mb</td>
<td>37,500</td>
</tr>
<tr>
<td>Human</td>
<td>3200 Mb</td>
<td>25,000</td>
</tr>
</tbody>
</table>
The Human Genome Project Sequence Represents a ‘Composite’ Genome

- Different sources of DNA were used for original sequencing
- The term ‘genome’ is used as a reference to describe a composite genome
- The many small regions of DNA that vary among individuals are called polymorphisms:
  - Mostly single nucleotide polymorphisms (SNPs)
  - Insertions/deletions (indels)
  - Copy number variation
  - Inversions

SNPs: the human genome has at least 38 million SNPs, most of these SNPs contribute to human variation, some of them may influence development of diseases, susceptibility to certain drugs, toxins, infectious agents
  - 1.4 million short stretches of insertions or deletions
  - 14,000 large DNA deletions
Causes of Sickle Cell Anemia and Cystic Fibrosis have been pinpointed to specific mutations in their Protein-Coding DNA.

Cystic fibrosis patients have a deletion of three base pairs in the CFTR sequence. Protein folds incorrectly and is marked for degradation.

Red blood cells carrying mutant hemoglobin are deprived of oxygen.
The Race for the $1,000 Genome

Human Genome Project (2001, initial draft):
> $ 3 billion (includes development of technology)
  “raw” expenses estimated at $300 million

Rhesus macaque (2006)
$ 22 million

By end of 2007:
$ 1-2 million for full mammalian genome sequence
(Jim Watson using pyro-sequencing technology)

Wanted: “!!!!!! The $ 1,000 Genome !!!!!!!!”
- low cost
- high-throughput
- high accuracy
Next Generation Sequencing (NGS) Focuses on Miniaturization and Parallelization

Sanger

Cyclic Array Sequencing (454 Pyrosequencing, Life Technologies SOLID Illumina Hi-Seq 2000)

- Easier library preparation
- Fragment amplification on solid surfaces
- Nucleotide-to-nucleotide sequence determination
- Sequence from both ends

Generate dense planar array of DNA features

Apply cycles of enzymatic driven biochemistry

Imaging-based data collection
DNA Preparation for High Throughput Sequencing

1. DNA is sheared into small fragments and each fragment is attached to a single bead.

2. DNA fragments on beads are amplified and placed into individual reaction chambers; Amplification needed to achieve required signal strength.

Life Technologies, Inc.
Reaction Mechanism for DNA synthesis

The 3’ hydroxyl group of the primer attacks the $\alpha$-phosphoryl group of the incoming nucleotide thereby forming a phosphodiester bond ($S_{N}2$ reaction).

Watson et al., MBG, 2008
Add one kind of dNTP per cycle. A nucleotide complementary to the template strand generates a light signal. The light signal is recorded by a CCD camera. The signal strength is proportional to the number of nucleotides incorporated.

Sequencing is recorded ‘live’. Read length up to 400 bases.

~ 1.7 Million reactions in parallel.

Diameter of single reaction chamber is approx. 44 µm.
Image Capturing

Example:
Addition of a C nucleotide

Each bright dot signals the incorporation of a C nucleotide at this position

Brighter dots mean more C nucleotides have been incorporated into the growing DNA chain

Flowgram: conversion of light intensity signals
Solexa/ Illumina Sequencing Technology

Use a set of deoxynucleotides that carry
- each a fluorescent label that can be cleaved off
- a reversibly terminating moiety at the 3’ hydroxyl position
BGI – The Sequencing Factory  
Beijing Genome Institute

Purchased 128 HiSeq2000 sequencers from Illumina in January 2010 each of which can produce 25 billion base pairs of sequence a day
Illumina Update (2013)

HiSeq X

A million dollar machine capable of sequencing 1800 human genomes a year, i.e. 3-4 genomes a day

BUT: buyer must order 10 machines and agree only to use for sequencing Human genomes

Cost will sink to $1000 per genome
Ion Torrent Sequencing

DNA to be sequenced is placed in wells of semiconductor chip
Add one nucleotide type per reaction cycle
When incorporated, a hydrogen ion is released and change in pH is recorded and converted into voltage
Oxford Nanopore Sequencing (MinION)

Targets $1000 genome sequencing

Hand-held disposable sequencer

Pull a single strand DNA through nanopore

As base go through they interrupt an ionic current that reveals each base’s identity

Should allow long reads without the need to piece together many short reads

Science, 21 Feb. 2014
## Costs associated with Genome Sequencing

Tracking costs associated with genome sequencing is facilitated by the National Human Genome Research Institute (NHGRI) via [this link](http://www.genome.gov/sequencingcosts/).

<table>
<thead>
<tr>
<th>Date</th>
<th>Cost per Mb</th>
<th>Cost per Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep-01</td>
<td>$5,292.39</td>
<td>$95,263,072</td>
</tr>
<tr>
<td>Mar-02</td>
<td>$3,938.64</td>
<td>$70,175,437</td>
</tr>
<tr>
<td>Sep-02</td>
<td>$3,413.80</td>
<td>$61,448,422</td>
</tr>
<tr>
<td>Mar-03</td>
<td>$2,396.20</td>
<td>$53,751,684</td>
</tr>
<tr>
<td>Oct-03</td>
<td>$2,230.98</td>
<td>$40,157,554</td>
</tr>
<tr>
<td>Jan-04</td>
<td>$1,508.91</td>
<td>$28,780,376</td>
</tr>
<tr>
<td>Apr-04</td>
<td>$1,135.70</td>
<td>$20,442,576</td>
</tr>
<tr>
<td>Jul-04</td>
<td>$1,107.46</td>
<td>$19,934,346</td>
</tr>
<tr>
<td>Oct-04</td>
<td>$1,028.85</td>
<td>$18,519,312</td>
</tr>
<tr>
<td>Jan-05</td>
<td>$974.16</td>
<td>$17,534,370</td>
</tr>
<tr>
<td>Apr-05</td>
<td>$897.76</td>
<td>$16,159,639</td>
</tr>
<tr>
<td>Jul-05</td>
<td>$890.90</td>
<td>$16,180,224</td>
</tr>
<tr>
<td>Oct-05</td>
<td>$766.73</td>
<td>$13,801,124</td>
</tr>
<tr>
<td>Jan-06</td>
<td>$699.20</td>
<td>$12,585,659</td>
</tr>
<tr>
<td>Apr-06</td>
<td>$651.81</td>
<td>$11,732,535</td>
</tr>
<tr>
<td>Jul-06</td>
<td>$636.41</td>
<td>$11,455,315</td>
</tr>
<tr>
<td>Oct-06</td>
<td>$581.92</td>
<td>$10,474,556</td>
</tr>
<tr>
<td>Jan-07</td>
<td>$522.71</td>
<td>$9,408,739</td>
</tr>
<tr>
<td>Apr-07</td>
<td>$502.61</td>
<td>$8,947,003</td>
</tr>
<tr>
<td>Jul-07</td>
<td>$495.96</td>
<td>$8,327,342</td>
</tr>
<tr>
<td>Oct-07</td>
<td>$397.09</td>
<td>$7,147,571</td>
</tr>
<tr>
<td>Jan-08</td>
<td>$102.13</td>
<td>$3,063,820</td>
</tr>
<tr>
<td>Apr-08</td>
<td>$15.03</td>
<td>$1,352,982</td>
</tr>
<tr>
<td>Jul-08</td>
<td>$8.36</td>
<td>$752,080</td>
</tr>
<tr>
<td>Oct-08</td>
<td>$3.81</td>
<td>$342,502</td>
</tr>
<tr>
<td>Jan-09</td>
<td>$2.59</td>
<td>$232,735</td>
</tr>
<tr>
<td>Apr-09</td>
<td>$1.72</td>
<td>$154,714</td>
</tr>
<tr>
<td>Jul-09</td>
<td>$1.20</td>
<td>$108,065</td>
</tr>
<tr>
<td>Oct-09</td>
<td>$0.78</td>
<td>$70,333</td>
</tr>
<tr>
<td>Jan-10</td>
<td>$0.52</td>
<td>$46,774</td>
</tr>
<tr>
<td>Apr-10</td>
<td>$0.35</td>
<td>$31,512</td>
</tr>
<tr>
<td>Jul-10</td>
<td>$0.35</td>
<td>$31,125</td>
</tr>
<tr>
<td>Oct-10</td>
<td>$0.32</td>
<td>$29,092</td>
</tr>
<tr>
<td>Jan-11</td>
<td>$0.19</td>
<td>$20,963</td>
</tr>
<tr>
<td>Apr-11</td>
<td>$0.12</td>
<td>$16,712</td>
</tr>
<tr>
<td>Jul-11</td>
<td>$0.09</td>
<td>$10,497</td>
</tr>
<tr>
<td>Oct-11</td>
<td>$0.09</td>
<td>$7,743</td>
</tr>
<tr>
<td>Jan-12</td>
<td>$0.06</td>
<td>$5,866</td>
</tr>
<tr>
<td>Apr-12</td>
<td>$0.07</td>
<td>$5,901</td>
</tr>
<tr>
<td>Jul-12</td>
<td>$0.07</td>
<td>$5,985</td>
</tr>
<tr>
<td>Oct-12</td>
<td>$0.07</td>
<td>$6,618</td>
</tr>
<tr>
<td>Jan-13</td>
<td>$0.06</td>
<td>$5,861</td>
</tr>
<tr>
<td>Apr-13</td>
<td>$0.06</td>
<td>$5,826</td>
</tr>
<tr>
<td>Jul-13</td>
<td>$0.06</td>
<td>$5,550</td>
</tr>
<tr>
<td>Oct-13</td>
<td>$0.06</td>
<td>$5,550</td>
</tr>
<tr>
<td>Jan-14</td>
<td>$0.04</td>
<td>$4,008</td>
</tr>
<tr>
<td>Apr-14</td>
<td>$0.05</td>
<td>$4,320</td>
</tr>
</tbody>
</table>
Single DNA Strand Sequencing (Pacific Biosciences) – 3rd Generation Sequencing

IN A FLASH

New DNA sequencers watch an enzyme called DNA polymerase as it uses fluorescently tagged bases to synthesize DNA. Each base is identified by a distinguishing colour that flashes as the base is incorporated into the DNA strand.

Reportedly generating sequences an average of 1,500 bp long
Examples of Projects and ‘Spinoffs’ derived from the HGP:

UCSC Genome Browser collects genome sequences

http://genome.ucsc.edu/
Build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

More than 80% of the human genome has at least one biochemical activity. Many regulatory regions transcribed into RNA with regulatory function.
Launched in 2008

Catalog of human genetic variation

Genomic knowledge will contribute to the fields of genetics, medicine, pharmacology, biochemistry, and bioinformatics
GWAS – Genome Wide Association Study

What is GWAS?

Examination of many common genetic variants associated with a trait

Association between single nucleotide polymorphisms (SNPs) and onset of major diseases, such as diabetes or heart disease, compared to healthy individuals.

Open access repository
The Cancer Genome Atlas (TCGA)

Reveal commonalities between cancer types, investigate molecular abnormalities, and define mutations that are confined to specific tumors
Development of prognostic, diagnostic and therapeutic strategies
Single Cell Sequencing – Method of the Year 2013

Single-cell sequencing can enable the discovery of clonal mutations, cryptic cell types or transcriptional features that would be diluted or averaged out in bulk tissue.

Rare cell types heterogeneous samples, phenotypes associated with mosaicism or variability.

Technologies for single-cell amplification and sequencing are maturing. As the cost and ease of examining individual cells improves, the approach will enter the hands of more researchers as a standard tool for understanding biology at high resolution.
Summary_1

Sanger sequencing provided the basis to initiate and complete the Human Genome Project and many other genomes.

Next Generation Sequencing (NGS) with high throughput and less costs took its place for genome sequencing (Pyrosequencing, Illumina, Ion Torrent).

2000: 4 eukaryotic genomes
2012: > 250 eukaryotic genomes, > 1000 human genomes

The various methods of sequencing have their advantages and disadvantages (e.g. cost, accuracy, fragment length, etc.)
• Data sharing (interdisciplinary approach between biology, engineering, and bioinformatics)

• Fewer protein-coding genes than thought

• Many regulatory RNA elements

• Studying of genetic variation and evolutionary origins

• Obtaining a genome ‘blueprint’ is not sufficient to explain the occurrence or susceptibility to disease, the combination of many factors needs to be taken into account

• The combination of various approaches (GWAS, 1000 Genomes, ENCODE, TCGA) and others will provide leads to the origin and treatment of human diseases.