Genome Sequencing

Jutta Marzillier, Ph.D
Lehigh University
Biological Sciences
The human body has about 100 trillion cells. Each cell harbors the same genetic information in its nucleus in form of DNA. Depending on cellular and developmental stage of a cell only a subset of genes is expressed.
What is a Genome?

- A genome is an organism’s complete set of DNA.
- Genomes vary in size: a bacterium contains about 600,000 building blocks called nucleotides, while human and mouse genomes have about 3 billion.
- Genes encoded by the DNA comprise the entire set of hereditary, instructions for building, running and maintaining an organism, and passing life onto the next generation.
What is DNA?

- The genetic information is stored in the nucleus of each cell in form of chromosomes.
- Chromosomes are long, tightly packed strings of DNA that harbor the genetic information.
- This genetic information is needed to construct other components of cells, such as proteins or RNA molecules.

The building blocks of DNA are called nucleotides. Only four nucleotides make up the DNA. These nucleotides are organized in the DNA double helix discovered by Watson and Crick.
Why is Genome Sequencing Important?

- To understand how the genome as a whole works – how genes work together to direct growth, development and maintenance of an entire organism
- Understand how gene expression is regulated in a particular environment
- 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)
How is Genome Sequencing Done?

**Clone by Clone**

Create a crude physical map of the whole genome before sequencing with restriction enzymes.

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

**Shotgun sequencing**

Break genome into random fragments, sequence each of the fragments and assemble fragments based on sequence overlaps.
DNA sequencing

- Biochemical method to determine the sequence of the nucleotide bases that make up the DNA
DNA sequencing, the ‘cradle’ of genomics…….

Seven Cheers for Technology

In 1986, the year in which the automated DNA sequencer was invented, GenBank held a scant 9.6 million bases. Yet discussion began on the feasibility of sequencing all three billion bases of the human genome. It was estimated at the time that such an undertaking would take “30,000 person-years of effort and upward of $3 billion.” Thanks to rapid improvements in the technology, however, the project was completed in just 7 years.

In this issue of The Scientist, we celebrate the DNA sequencer, as well as other key technologies that have, and are, transforming life science research. Each of our chosen seven—the others are the BLAST algorithm, the DNA microarray, the yeast two-hybrid assay, the MALDI-TOF mass spectrometer, the lab-on-a-chip, and the optical trap—is in its own way shaking the foundations of life science research.

As a group, they tell the story of the past and future of molecular biology. While each technology followed a very different path, the accounts share the virtues of brilliance, endeavor, perseverance and, tellingly, the cross-fertilization of ideas. Technology development is arguably the single most important factor in the rapid pace of science.

We start by extolling the sequencer (p. 25). By the end of the Human Genome Project in 2003 this workhorse of the genomics revolution had produced some 28.5 billion nucleotides, and thanks to BLAST (p. 21), the scientific community had bioinformatics tools to sieve all that data.

Next, we salute three enablers of functional genomics. There’s the DNA microarray (p. 27), which has been key to revealing which genes turn on and off in response to disease, pharmaceuticals, or developmental signals on a genome-wide scale. The yeast two-hybrid assay (p. 32) and MALDI-TOF (p. 37) have each enabled proteomics, the former by enabling researchers to identify proteins without first purifying them, the latter by helping researchers to map protein-protein interaction maps in yeast, fruit flies, nematodes, and man.

Closing out the seven are commendations for microfluidics (p. 43) and the optical trap (p. 48). These represent a new chapter in scientific research, an era of miniaturization, in which chemical libraries are rapidly and inexpensively screened for new drug candidates, complex biochemical tests are performed while you wait in the doctor’s office, and the conformational acrobatics of proteins are understood at the molecular level.

Throughout the pages in this issue we explore the history and evolution of each of our seven chosen technologies. We include a series of colorful “how it works” illustrations, showing what goes on under the hood of exemplars of each of the highlighted technologies. Ever wonder what happens when you stick an Affymetrix GeneChip array into a GeneChip Reader, or how an optical trap works? Wonder no more, you can find out on pages 30 and 50.

We do not presume to list the seven technologies that have transformed the life sciences, but rather our pick of key technologies. If you are outraged, or even just astonished, by what we’ve left off, we’d love to hear from you—there could well be scope for “Seven More Technologies that are Transforming the Life Sciences.”

In the meantime, please join us in celebrating our magnificent seven.

Richard Gallagher, Editor (rgallagher@the-scientist.com)
Jeffrey M. Perkel, Senior Editor (jperkel@the-scientist.com)

Life Sciences Thrive Through the Collaboration between Biologists, Engineers, and Computer Scientists
DNA: chemically linked chain of nucleotides which each consists of a phosphate, a sugar (deoxyribose), and a nucleobase (adenine cytosine, guanine, thymine).
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
  - deoxynucleoside-triphosphates are the 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

Watosn et al., MGB, 2008
Reaction Mechanism for DNA synthesis

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
DNA sequencing

Maxam A.M. and W. Gilbert, 1977
A new method for sequencing DNA.

Sanger et al., 1977
DNA sequencing with chain terminating inhibitors.
Proc. Natl. Acad. Sci 74:5463
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.

Frederick Sanger
Nobel Prize (1980)

Watson et al., MBG, 2008
Dideoxy Method of Sequencing (Sanger, 1975)

“Normal” DNA synthesis:
DNA strand as template
Primer
Deoxynucleotides
Polymerase enzyme
Use several cycles to amplify

- DNA synthesis is carried out in the presence of limiting amounts of dideoxynucleoside triphosphates that results in chain termination
- Through chain termination fragments of distinct sizes are generated that can be separated by gel electrophoresis
- Original method used radio-labeled primers or dideoxynucleotides

![Diagram of dideoxy method of sequencing](image-url)
Dideoxy Method of Sequencing

Original method used radio-labeled primer or dideoxynucleotides

This method required four separate DNA synthesis reactions to be separated by electrophoresis in four parallel lanes. The gel needs to be dried, exposed to film, developed and manually read.

Approx. 150 bases read length

Summary – Sequencing Method Established

- Need of four reactions in parallel
- Heat labile polymerase
- Use of radioactivity
- Low resolution on gels
- Approx. 150 nucleotides read length
- Time consuming

Improvements

- Use of fluorescently labeled dideoxynucleotides
- One-lane electrophoresis
- Introduction of capillary electrophoresis to increase resolution (up to 1,000 ntes)
- Use of heat stable polymerase (Taq)
- Automation
Taq-Polymerase


Mullis integrated the Taq polymerase into the thermo-cycling process of PCR (1988, Nobel Prize, 1993).

Automated DNA Sequencing Methods Involving Polymerase Chain Reaction

Fluorescently labeled ddNTPs used for sequencing reaction

ddNTPs used for automated sequencing are labeled with different fluorescent dyes representing each of the 4 bases.

Left image: dideoxyribonucleoside triphosphate prevents strand extension at 3’ end.

Right diagram: Single-stranded DNA to be sequenced is subjected to enzymatic reaction with ddNTPs. Electrophoresis using laser to activate the fluorescent dideoxy nucleotides and a detector to distinguish the colors allows separation of DNA products in the same lane of gel.
Capillary Gel Electrophoresis

Sequence ladder by radioactive sequencing compared to fluorescent peaks

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.

Advantage
High resolution
Read length up to 1,000 nucleotides
Automated DNA Sequencing Using Capillary Gel Electrophoresis

ABI 310
DNA sequencer, one-capillary instrument
Capillary and electrode: Negatively charged DNA migrates towards anode.

DNA fragments labeled with different fluorescent dyes migrate according to their size past a laser.
Sequencing Electropherogram

- Laser excites dyes, causing them to emit light at longer wavelengths
- Emitted light is collected by a CCD camera
- Software converts pattern of emissions into colored peaks

Plot of colors detected in sequencing sample scanned from smallest to largest fragment

www.contexo.info/DNA_Basics/dna_sequencing.htm
High-Throughput Whole Genome Sequencing

Analysis of 384 sequencing reactions in parallel

George Church, Scientific American, January 2006, pp47-54
JGI Sequencing Facility
(Joint Genome Institute, US Department of Energy)

Assume more than 20 384-capillary sequencers running simultaneously
approx. 700 bp per capillary run
approx. 3 hours per run

Approx. 40 Million bases per day in one facility
The Human Genome Project

Initial sequencing and analysis of the human genome
**International Human Genome Sequencing Consortium**

The Sequence of the Human Genome
Venter et al. (Celera Genomics)

Finishing the euchromatic sequence of the human genome
**International Human Genome Sequencing Consortium**
The Human Genome

- Current genome (build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps (2004)
- Overall error rate is less than 1 error per 100,000 bases
- Encodes for approx. 25,000 protein-coding genes
- Genes make up about 1-2 % of the total DNA

![Diagram showing the percentage distribution of different DNA elements such as LINEs, SINEs, introns, protein-coding regions, and unique DNA.](https://example.com/diagram.png)
Who’s sequence was used for the ‘Human Genome Project’ and Why is it Important?

- Different sources of DNA were used for original sequencing
  - Celera: 5 individuals; HGSC: ‘many’
- 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
- Between any two humans the amount of genetic variation is about 0.1%
- Recognize gene-regulatory components important in understanding human biology and disorders, such as heart disease, cancer, and diabetes.
- Understand the mechanisms of genetic heredity and how different organisms relate to each other

Development Sequencing Time Line

1953: Discovery of the DNA double helix by Watson & Crick

1975: Frederick Sanger invents the "chain termination" method of DNA sequencing

1982: GenBank launches with 606 sequences and 680,338 bases

1986: Nature publishes, and Applied Biosystems commercializes, the automated DNA sequencer

1990: The Human Genome Project begins

1991: Haemophilus influenzae is the first free-living organism sequenced (1.8 million bases)

1993: Applied Biosystems releases the first sequencer based on capillary electrophoresis

1995: Yeast chromosome III sequenced (315,000 bp)

1996: Saccharomyces cerevisiae sequenced (14 million bp)

1997: GenBank exceeds 1 billion bases

1998: Caenorhabditis elegans sequenced (97 million bp)

1999: First human chromosome (22) sequenced (33.4 million bp)

2001: Human Genome Project and Celera publish drafts of the human genome (3 billion bp)

2003: Human Genome Project finishes

2004: GenBank holds 44.5 billion base pairs, and 40.6 million sequences

2005: Mouse genome sequenced (2.5 billion bp)

2007: Drosophila melanogaster (180 million bp) and Arabidopsis thaliana (120 million bp) sequenced

August 29, 2005
Development Sequencing Time Line

1953:
Discovery of the DNA double helix by Watson & Crick

2005

2006
The duck-billed platypus: part bird, part reptile, part mammal – and the genome to prove it

- The Platypus is our most distant mammal relative
- Has retained a large overlap between two very different classifications
- The platypus shares 82% of its genes with human, mouse, dog opposum and chicken
- Decoding the platypus genome helps to understand the origins of mammal evolution

The Race for the $1,000 Genome

Human Genome Project (2001, initial draft):
$2-3 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

Wanted: “!!!!!! The $1,000 Genome !!!!!!!!”
- low cost
- high-throughput
- high accuracy
2nd (or Next) Generation Sequencing Technologies

**Motive:**
- Greater sequencing throughput
- More economical sequencing
- High fidelity
- Long read length

**Strategy:**
- Reduce reaction volume
- Extend the numbers of sequencing reactions
- Massively parallel sequencing

Roche/454: pyrosequencing

Solexa (Illumina)

SOLiD (ABI)

NATURE METHODS, Vol.5, No1, January 2008
Comparison of Sequencing Technologies

Choose the technology that fits your project! Benefit from our experience to tailor complete sequencing solutions and programmes to meet your individual requirements!

**ABI 3730XL**
(Applied Biosystems/Sanger)
up to 1,100 bases/read
96 reads/run
approx. 1 MB/day and machine

First choice for finishing projects; full length cDNA sequencing; single sample sequencing.

**GS FLX/454**
(Roche Diagnostics)
up to 250 bases/read
up to 400,000 reads/run
up to 100 MB/run/7.5 hours

Optimal for in-depth analysis of whole transcriptomes; bacterial genomes; small eukaryotic genomes.

**Genetic Analyzer/Solexa**
(Illumina)
up to 50 bases/read
up to 60 Mio reads/run (paired-end)
up to 2,000 MB/run/6.5 days
Sequencing by synthesis

Highly attractive for resequencing projects of e.g. production strains; small RNA, SAGE and ChIP; ultra-deep sequencing of SNPs or mutations.

**SOLID DNA Sequencer**
(Applied Biosystems)
up to 35 bases/read
up to 85 Mio reads/run (paired-end)
up to 3000 MB/run/6 days
Sequencing by ligation

Highly attractive for resequencing projects of e.g. production strains; small RNA, SAGE and ChIP; ultra-deep sequencing of SNPs or mutations.
Next-Next Generation Sequencing:
Helicos - true Single Molecule Sequencing (tSMS)

- Sequences directly single molecules of DNA
- No amplification
- Increases speed
- Lowers cost

Within a flow cell billions of Single DNA molecules are captured on surface. Per cycle only one fluorescent dideoxynucleotide is added. Wash, measure fluorescence, cleave label and continue process with each of the other three base 25-base read length

www.helicosbio.com
Recent achievements

May 2007: James Watson’s genome deposited
  454 technology (pyrosequencing)
  2 month, $ 1-2 million

Sept 2007: Craig Venter’s genome deposited
  Sanger technology, 4 years, 70 million

*Diploid genomes*

*Approx. 3.5% of Watson’s genome could not be matched to the reference genome*

*Venter’s genome had 4.1 million DNA variants comprising 12.3 Mb
  - 3.2 million SNPs
  - non-SNP variants

*These variants include a potentially increased risk of alcoholism, coronary heart disease, obesity, Alzheimer’s disease, antisocial behavior and conduct disorder.*
Summary

Recent advances in DNA sequencing catapulted Life Sciences into the ‘Genomic Era’

DNA sequencing based on improved Sanger technology enabled sequencing of many whole genomes, including that of the roundworm, yeast, mouse, human, dog, and others

Through long base reads Sanger technology is a powerful tool to generate reference genomes

The 2\(^{nd}\) generation of sequencing (named ‘Method of the Year 2007’) allows high-throughput, lower cost sequencing useful for genome comparisons, personalized genomics and potential clinical diagnostics
Pyrosequencing - the 454 Technology

DNA fragments are ligated to adapters of known sequence
Amplification of individual DNA fragments on beads in emulsion PCR
Addition of one unlabeled nucleotide at a time
Upon nucleotide incorporation pyrophosphate is released, converted to ATP which fuels the luciferase driven generation of light emission
Read length is 100-150 nucleotides
Solexa Technology

Ligate DNA fragments to adapters
Bind to solid surface and amplify to create clusters of approx. 1,000 copies of single stranded DNA
For sequencing add four differently labeled, reversible terminator nucleotides
After incorporation read fluorescence, remove fluorophore and terminator and repeat cycle
Read length is 30-35 nucleotides
SOLID Technology (Applied Biosystems)

DNA fragments are ligated to adapters and amplified on beads by emulsion PCR
A sequencing primer is hybridized to adapter
A mixture of octamer oligonucleotides compete for ligation to primer
The bases in 4th and 5th position on the oligonucleotides are encoded by one of four color labels.
Ligated oligo is cleaved after position 5 which removes label and the cycle is repeated
Whole process is repeated by off-set primer n-1
Nucleotide read length is 30-35
Race for the $1,000 Genome

Goal: - personalized patient treatment strategies

Triggers basic research:
- how is activation of genes regulated?
- understanding genetic links to cancer
- facilitate genetic engineering
- hunt for “disease” genes and mutations
- diagnosis and treatment of diseases?

However:
- How can patient privacy be protected?
- Will insurance companies and employers use genetic information to screen out those at high risk for disease?
- genetic engineering of bioterrorism agents
Addition of dNTP releases a pyrophosphate (PPI) stochiometrically.
Sulfurase converts PPI to ATP.
ATP and Luciferase drive conversion of luciferin to oxyluciferin that generates visible light and can be measured with a CCD camera. Each light signal is proportional to the number of nucleotides incorporated.
Apyrase digests unincorporated nucleotides.
Sequencing through nanopore

Measure changes in electric current as DNA is passes through a nanopore surrounded by two pairs of tiny gold electrodes. Electrodes record electrical current perpendicular to the DNA strand. As each DNA base is structurally and chemically different, each base creates its own electronic signature.

Johan Lagerqvist
www.ucsdnews.ucsd.edu
The next biological revolution

- Genomics as a toolbox that will enable a truly personalized, predictive, and preventive medicine
- L Hood: ”All the big revolutions [in science] are technique-driven.”
Steps of genome sequencing

- Break genome into smaller fragments
- Sequence those smaller pieces
- Piece the sequences of the short fragments together

**Challenge:** sequencing methods are limited by the physical size of the gel that is used to separate the labeled fragments (approx. 700 bp per reaction)
DNA sequencing approaches

- 2 different methods used

- **Shot gun approach**
  - Useful for smaller genomes

- **Clone-by-clone**
  - Useful for sequencing genomes of higher vertebrates that contain repetitive sequences
Method of shotgun sequencing

- Genome is broken into much smaller, overlapping fragments. Each fragment is sequenced and the genome is assembled based on overlapping sequences.
- Useful for sequencing small genomes.
Repetitive sequences make up a large fraction of vertebrate genomes.

Almost 50% of the human genome consists of repetitive sequences.

More than half of the unique sequences consists of the introns and exons of genes.
Repetitive sequences make correct assembly difficult

- Repetitive sequences make up a large fraction of vertebrate genomes.

- Repetitive fragments from different parts of the DNA appear to overlap. Assembling these sequences would result in a loss of the information that lies between the original repeats.
Clone-by-clone approach- 1

- Generate a genomic library
- Endonuclease digest
- Ligation into BACs
- Transformation into *E. coli*

- Collection of cloned DNA fragments results in genomic DNA library
Clone-by-clone approach - 2

- Position individual BAC clones on the physical map of the human genome sequence on the basis of their restriction digest “fingerprints”.
# How does the human genome compare to others?

<table>
<thead>
<tr>
<th># of protein-coding genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human</strong> 20,000-25,000</td>
</tr>
<tr>
<td><em>C. elegans</em> (nematode) 19,000</td>
</tr>
<tr>
<td><em>D. melanogaster</em> (fruitfly) 15,000</td>
</tr>
<tr>
<td><em>Haemophilus</em> (bacterium) 1,738</td>
</tr>
</tbody>
</table>

- The biological complexity of an organism is not simply a function of the number of genes in its genome.
Are we only twice as complex as the roundworm and the fruit fly?

- Many genes encode for more than one protein product (*alternative splicing*)
- *Transcription factors and enhancers* probably provide greater flexibility of gene expression
Humans have many genes not found in invertebrates

- These include genes encoding
  - Antibodies and T cell receptors
  - Major histocompatibility complex
  - Cytokines
  - Molecules involved in blood clotting
  - Many mediators of apoptosis
Why do we need DNA Sequencing?

- Decipher the genetic code of living organisms
- Generate high-quality reference sequences
- Recognize gene-regulatory components important in understanding human biology and disorders, such as heart disease, cancer, and diabetes.
- Understand the mechanisms of genetic heredity and how different organisms relate to each other