Human Genome Sequencing and Its Impact on Science and Medicine

Jutta Marzillier, Ph.D
Lehigh University
Biological Sciences

November 16th, 2011
In June 2000 at the White House:
First rough draft of the “book of life” announced

Clinton:
Completion of the Genome Project would
“...revolutionize the diagnosis, prevention and treatment of
most, if not all, human diseases......”

Collins:
“...personalized medicine is likely
to emerge by the year 2010.....”
In February 2001 the ‘First Draft’ of the Human Genome is Published

Venter et al., Celera, Science, 2001

DNA Sequencing

• Human Genome DNA Sequencing is based on **chain termination** or **dideoxysequencing** method developed by Frederick Sanger in 1977
• At the same time Maxam & Gilbert published a DNA sequencing method by **chemical modification** of nucleotides

The original method of Sanger sequencing and multiple improvements regarding chemistry and computation lead to complete sequencing of the more than 3 billion basepairs containing Human Genome (and many others).

Sanger sequencing is expensive and **Next-Generation-Sequencing (NGS)** technology took its place.
The human body has about 100 trillion cells. 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/
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)
- Personalized Genomics (David Church, Harvard)
Milestones to DNA Sequencing and Annotation

1953: Watson and Crick described the double helix DNA structure

1959: Arthur Kornberg identified all components to synthesize DNA in vitro

1966: Nirenberg and colleagues describe genetic code

1977: Frederick Sanger Publishes the method of di-deoxysequencing

1986: Introduction of the Polymerase Chain Reaction by Kary Mullis using the heat-stable Taq-Polymerase enables automation

1990: Introduction of fluorescent label and capillary gel electro-Phoresis made launch of the HGPfeasible

http://www.wellcome.ac.uk/en/fourplus/DNA.htm
What Challenge Was it to Sequence the Human Genome?

It took more than 6 years to determine the complete base sequence of the *E. coli* genome (4.6 Mb). The Human Genome has 3200 Mb.

At this rate it would have taken a single lab more than 4,000 years to sequence the entire human genome.

The average fragment read length is about 500 to 800 bases.

It would take a minimum of six million (3 billion/500) reactions to sequence the human genome (no overlap, 1-fold coverage).

Develop procedural and computational algorithms and efficient database management.

http://t2.gstatic.com/images/
Challenges of Whole Genome Sequencing:

The Human Genome comprises 3 billion base pairs

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*

Shotgun sequencing

Break genome into random fragments, insert fragments into vector, sequence each of the fragments and assemble fragments based on sequence overlaps

• need known flanking region to anneal primer
Challenges of Whole Genome Sequencing:
Sanger Method of DNA Sequencing generates 800 bp reads

Generates labeled DNA fragments of various sizes
DNA fragments are separated by gel or capillary electrophoresis and nucleotides are ‘called’
These reads call about 800 nucleotides


- need known flanking region to anneal primer
Capillary Gel Electrophoresis increases resolution

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

http://dnasequencing.wordpress.com/
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

http://snhs-plin.barry.edu/images/JGI_Sequencing.jpg
Genome Assembly

Primary sequencing reads:

- cagacgtgtcagtcgactcgataataactgagctagtgcgact
- tagctagccggatagttattaccagacgtgtcagtcgactcgata
- ccagatcgatcgtcgattgccgatagctagccggatagttattaccagacgt

Align sequences for homologies:

- Green: 3-fold coverage
- Yellow: 2-fold coverage

Fragment assembly to contig:

- Green: 3-fold coverage
- Yellow: 2-fold coverage

Goal for Genome Sequencing:
- Sanger: 8-fold coverage
TGCGGCTGCCAGATTTTGTACGGGGTTTGAAGTCGACGGAGAGAACAGCGCGGGCCTAGAAGGCCCCGTAATGCCCCCTGAGAGCCCCGTAGACCGGACGAAAGCCTCGGACCGGATTCAGATAGTAGGACACC
GGAGACAAGCGAAGACGGCAGCCGAGAGCCGTCCGGCAGCTGACGCCCGCGTAGGAAGATATTCGTGTGAAGTGCGTCACATTCTACGGGTGAAACGCGAAAGTGGAAGGTTCCTTACCTATGGAGGG
GTAAGGGAGCGAGCTCCAGCGAGCGACCCGACCCCCCGACATAGGTTTCTTTTGTCCGGGTTGATGCTGAACGAGAGAGACAGGCTTTTACGATACCGTACAGGCCGGGGTTTATCCCCCGGCCGACCATGTAACCGGAGAGATGTTACGTAAGGAGGAGGCAGCTCCAGCGAGCGACCGCACCCCGACATAGGTTCTTGTCGGGGTAGTCGAACGAGAGAGACTACCCCTTTTAGCAGCCGGTCGCCACAGGTACCGAACGATGAGAGAGGTACCTAGACCGTACAGGCCGGGGTTTATCCCCCGGCCGATACAGCATGGTCATTTTGGGTAAGGTACGTTACGTAAGCATCACTCACTCAAACAGAACCAGATGTTACGTAACCCCGGGTTACGTAACGTAACAGAACCACTAACCCGTGGCCGCCCCGAGGGCGGCCAGCAGGGTTACGT TGTTTGTCAGGTATGCTACTAGGGGAGGGTGAACATATGAAATCTCAAGGAGCATCAGAAAGGAGCTTCTTCGTCTACTGCTCGTGTCTGAGTCGATGACGTATGAGCGTGCCTACTCGAAGGTGGAGACCATCCTGCACCTCATCCAGTCCTTCGACGCGGAGAACTCGATAAGTGAGCTGGGAGTCATCTGACCCGGCGCGATCGTCTGCCGACCGACTGGCCTCGCATCCGCCGCGAGGTTCTGCGGGCGGCTGGCACCGCTGCCAGATCTGCTACGCGGACATCTGCACAGGGATGGCTACCGAGGTTGATCACGTCGCTACCGCGACGAGGCGTCACCTGCAGGTGTCGTGCAGACCGTGCCATGCGCGGAAGTcccGCGATGGAAGGCGTTGCTCAGCGTGCGAAGCTGCGCGCGATGAAGAAGCGGCCGCCGCCCCGCCACCCGGGGCGTAGAAGCAACTAGGAGGGACCAGGCGTCCCCGAGCCCAGGAGGCGTCAGCCGGGCTCCAGTGCCCAAGCGATCGGACGAACGCGTCCGGCGCAAATCGCATACGAGTGAGCGCGAGGCTAGC
Genome Annotation: Searching for Open Reading Frames by Identifying all possible Start and Stop Codons

Start – ATG

Stop – TAA, TAG, TGA

AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGATCAC
TGTCCTTCTGCCATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCG
CTGCTGGCCCTCTGCCATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCG

Attach biological information to it

http://www.brooklyn.cuny.edu/bc/ahp/BioInfo/graphics/GP.GeneticCode.GIF
What did we learn from HGS?

- Less than 2% of the Human Genome codes for protein
- The human genome encodes for approx. 21,000 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 describing Human Variation

- 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

SNPs: the human genome has at least 10 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
Genomic Achievements since the Human Genome Project:
2nd or (NGS) Generation Sequencing
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

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

Shendure & Ji, nature biotechnology, 2008

800 bases read 35–70 bases read
Cell Lysis and DNA Preparation

Cells are lysed in lysis buffer which is passed over a column.

Silica particles bind specifically to DNA.

After several washing steps DNA can be eluted in water.
DNA fragmentation (Nebulization)

Shearing double-stranded DNA into fragments ranging from about 400 to 1000 base pairs. This population of smaller-sized DNA species, generated from a single DNA sample, is referred to as a “library”.

The Nebulizer is a small plastic device used to atomize liquids-and compressed air to shear large DNA into smaller fragments.

http://dna.uga.edu/docs/GS-FLX-Titanium-General-Library
Library Fragment Size Selection

DNA sample is loaded onto an agarose gel, which separates the fragments by size.

Gel is stained with Ethidium bromide and visualized under UV.

DNA fragments of size interest are excised.

0.8% agarose gel
Paired-End Reads and mate pairs improve *de novo* assembly

Generate genomic DNA fragments of 20, 8, or 3 kb size
Add adapters to each end
Circularize fragment
Break circularized constructs with enzyme
Enrich for fragments containing internal adapter
Add adapters for sequencing

Wiley et al., Current Protocols, 2009
Emulsion PCR (Polymerase Chain Reaction)

- Adapter carrying library DNA
- Mix DNA Library & capture beads (limited dilution)
- + PCR Reagents
- + Emulsion Oil
- Create “Water-in-oil” emulsion
- Micro-reactors
- “Break micro-reactors” Isolate DNA containing beads
- Perform emulsion PCR

- Generation of millions of clonally amplified sequencing templates on each bead
- No cloning and colony picking
Illumina Technology

454 PicoTiter (Roche) plate holds 1.6 Mio reactions in parallel

Fluorescent readout of incorporated bases

 Depositing DNA templates in a highly parallel manner
Comparison of Sequencing Technologies

Compromise between cost, read lengths, accuracy, speed and output
What can a sequencer do today?

In 8 days...

“Paired end” 100 bp reads

1 to 2 billion read-pairs

300 gigabases (Gb) of total output

1 in 1,000 error rate for most base-calls

(initial draft of human genome based on 23 Gb)
Sudden and Profound Out-Pacing of Moore's Law

Transition from Sanger Sequencing to Next-Generation Sequencing technologies

**Moore's Law**, which describes a long-term trend in the computer hardware industry that involves the doubling of 'compute power' every two years

[www.genome.gov/sequencingcosts](http://www.genome.gov/sequencingcosts)
Beijing Genome Institute: BJI – The Sequencing Factory

Purchased 128 HiSeq2000 sequencers from Illumina in January 2010 each of which can produce 25 billion base pairs of sequence a day.
Genomic Achievements since the Human Genome Project: 2\textsuperscript{nd} or (NGS) Generation Sequencing
1000 Genomes Project

The 1000 Genomes Project is the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation.

http://www.1000genomes.org/
Genome Wide Association Studies (GWAS) to detect Common Gene Variants

Assumption:
To predict human diseases or susceptibility common DNA variations (~5% frequency) would be at a fault.

Many SNP alleles have been uncovered related to specific diseases. however, these variants have only accounted for a small fraction of disease risk.

The HapMap project is looking for common patterns of human DNA sequence variation and it’s potential correlation to disease.
Establish a Cancer Genome Atlas: Decoding of a Acute Myeloid Leukemia Genome

* AML is a cancer of white blood cells, which accumulate in the bone marrow and interfere with the production of normal white and red blood cells.
* Sequenced and compared the genome of cancerous bone marrow cells with normal skin cells from the same patient using the next generation technologies.
* Found a total of 10 genes carrying mutations in the cancerous cells.
* Two of those (FLT3 and NPM1) had already been implicated in the process of progression to AML.

Proof of concept that whole genome sequencing of tumor cells may be a useful strategy to uncover genes implicated in the disease.
Twin study surveys genome for cause of Multiple Sclerosis

MS causes the body’s own immune cells to attack the myelin sheath around nerve cells

MS has a genetic component

No clear genetic reason found to explain why one twin developed MS while the other did not

Both twins carry genetic variants that are linked to a higher risk getting MS

But those genetic factors seem to have been insufficient to cause disease of their own

Environmental triggers, epigenetics?

Baranzini et al.
April 29th, 2010
Genomics Research – from Bench to Bedside

Green & Guyer, nature, Vol 470, 2011
Lessons learned

• Data sharing

• 2000: 4 eukaryotic genomes
  • 2010: 250 eukaryotic genomes, many hundreds of human genomes

• Fewer protein-coding genes than thought

• Studying of genetic variation and evolutionary origins

• Cost of sequencing has fallen 100,000 fold; high throughput sequencing

• HapMap project: Catalog more than 20 million SNPs

• Cancer genome atlas to map genomic changes observed in every major type of human cancer

• However, 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