Your Genes and Your Health
http://bio84.stanford.edu/

The Human Genome Project

Doug Brutlag, Professor Emeritus
Biochemistry and Medicine (by courtesy)
brutlag@stanford.edu
The Human Genome Project: Should we do it?

  - Not hypothesis driven.
  - Fishing expedition or stamp collecting.
  - Eliminate funds from investigator initiated science.

  - Limit sequencing to 1.5% of genome that codes proteins.
  - Do not sequence intergenic regions “genetic wastelands”.
  - Do not sequence repeated regions (centromeres, telomeres and heterochromatin).

  - Technology of the time permitted 500 bp per day per person.
  - Move from radioactively labeled sequencing to fluorescent sequencing permitted complete automation up to 1 gigabyte per year.
Genome Sizes

Human Genome
Mouse Genome

Fruit Fly Genome

Nematode Genome

Yeast Genome

E. coli Genome

~3,000,000,000 bp
~160,000,000 bp
~100,000,000 bp
~15,000,000 bp
~5,000,000 bp
The Human Cytogenetic Map
Human Genome (~3000 Mb)

Human Chromosome (~130 Mb)

YAC (~0.5-1.0 Mb)

BAC (~0.1-0.2 Mb)

Courtesy Eric Green
Clone-Based Physical Mapping

Chromosome

Clones

Contigs

Courtesy Eric Green
Subclone Construction

1. Prepare Multiple Copies
2. Randomly Fragment
3. Subclone Fragments

BAC DNA
Shotgun Sequencing Strategy

BAC

© Doug Brutlag 2016

Courtesy
Eric Green
Clone-Based Shotgun Sequencing

Construct clone map and select mapped clones

Generate several thousand sequence reads per clone

Assemble
Whole-Genome Shotgun Sequencing

Generate tens of millions of sequence reads

Assemble
Whole Genome Shotgun versus Clone Sequencing

Let's sequence the human genome with the shotgun strategy.

That is impossible, and a bad idea anyway.

1997

Gene Myers

Phil Green

Thanks to Seraf in Batzoglou
Hierarchical Sequencing Vs. Whole Genome Shotgun Sequencing

(from Gibson & Muse, A Primer of Genome Science)
M13mp18 Sequencing Vector

http://www.mikeblaber.org/oldwine/bch5425/lect33/lect33.htm

M13 Circular DNA With LacZ gene

Insert 1 kb Human DNA Segment

LacZ Gene
DNA Synthesis by DNA polymerases
DNA Sequencing by Chain Termination

Single-stranded DNA to be sequenced

5' C T G A C T T C G A C A A

Add:
DNA polymerase I
dATP
dGTP
dCTP
dTTP
plus limiting amounts of fluorescently labeled
ddATP
ddGTP
ddCTP
ddTTP

Electrophoresis using laser to activate the fluorescent deoxy nucleotides and a detector to distinguish the colors

So the sequence of the template strand is

3' G A C T G A A G C T G A C T G A A A A A
5' C T G A

Larger fragments

Smaller fragments
Analyzing Fluorescent DNA Sequencing Data

Computer Analysis
Fluorescent DNA Sequencing Results
Sequence Assembly

(a) Sequence reads

Read 1: CACATACACATGG
Read 2: TCAATGGGGCTAA
Read 3: AGCACGGACTTGTCACTACACATG
Read 4: ACACATGGAAATA
Read 5: GGGCTAATGATTGTCAC
Read 6: TGATTGTCACATA
Read 7: ATTCATGAAGCACGGA
Read 8: GTCACTACACATGATCAATGGGG

Use computer to assemble sequence reads

(b)

<table>
<thead>
<tr>
<th>Number</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>ATTCATGAAGCACGGA</td>
</tr>
<tr>
<td>3</td>
<td>AGCACGGACTTGTCACTACACATG</td>
</tr>
<tr>
<td>8</td>
<td>GTCACTACACATGATCAATGGGG</td>
</tr>
<tr>
<td>2</td>
<td>TCAATGGGGCTAA</td>
</tr>
<tr>
<td>5</td>
<td>GGGCTAATGATTGTCAC</td>
</tr>
<tr>
<td>6</td>
<td>TGATTGTCACATA</td>
</tr>
<tr>
<td>1</td>
<td>CACATACACATGG</td>
</tr>
<tr>
<td>4</td>
<td>ACACATGGAAATA</td>
</tr>
</tbody>
</table>

Assembly of sequence reads

(c)

ATTCATGAAGCACGGAACCTTGTCACTACACATGATCAATGGGGCTAATGATTGTCACATACACATGGAAATA

Contig
February, 2001 Draft Sequence

International Human Genome Sequencing Consortium (2001)

Venter et al. (2001)

Courtesy Eric Green
The genome finishers

Dedicated scientists are working hard to close the gaps, fix the errors and finally complete the human genome sequence. Elie Dolgin looks at how close they are.
Shotgun Sequencing Strategy

BAC

“Finishing”

“Working Draft” Sequence

Finished Sequence
Polymerase Chain Reaction Overview: Exponential Amplification of DNA
PCR Requirements

DNA
- Need to know at least the beginning and end of DNA sequence
- These flanking regions have to be unique to strand interested in amplifying
- Region of interest can be present in as little as one copy
- Enough DNA in 0.1 microliter of human saliva to use PCR

DNA Polymerase Enzyme
- DNA polymerase from *Thermus aquaticus*—Yellowstone
- Alternatives: *Thermococcus litoralis, Pyrococcus furiosus*

Thermocycler
PCR Applications

Forensics
- assessment/reassessment of crimes
- 13 FBI CODIS markers

Archaeology
- determine gene sequences of ancient organisms
  - Neandertals
  - Denisovans
  - Otzi
- rethinking the past, human origins

Molecular Biology
- Cloning genes
- Sequencing genes
- Finishing genome sequences
- Amplification of DNA or RNA

Medicine
- Diagnostics for inherited disease
- Diagnostics for gene expression
- Diagnostics for epigenetics
Finishing the euchromatic sequence of the human genome

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (build 35) contains 2,866,000 nucleotides interrupted by only 341 gaps. However, 99% of the euchromatic genome is accurate to an error rate of 1 per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require further work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome now encodes only 20,000–21,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

The Human Genome Project (HGP) was launched in 1990 with the goal of obtaining a highly accurate sequence of the vast majority of the euchromatic portion of the human genome. The initial work followed a two-pronged approach: (1) the mapping of the human and mouse genomes to allow the study of inherited disease and provide a crucial scaffold for genome assembly; and (2) the sequencing of organisms with smaller, simpler genomes to serve as a testbed for method development and assist in interpreting the human genome. With success along both paths, the sequencing of the human genome itself eventually became feasible. The International Human Genome Sequencing Consortium (IHGSC), an open collaboration involving twenty centers in six countries, was formed to carry out this component of the HGP.

In February 2001, the IHGSC and Celera Genomics each reported draft sequences providing a first overall view of the human genome. These sequences allowed systematic analysis of both the genome itself, including identification of genes, combinatorial architecture of proteins, regional differences in genome composition, distribution and history of transposable elements, distribution of polymorphisms and relationship between genetic recombination and physical distance. Moreover, systematic knowledge of the human genome has enabled new tools and approaches that have markedly advanced biomedical research.

Both draft sequences, however, had important shortcomings. The IHGSC sequence, for example, contained ~10% of the euchromatic genome; it was interrupted by ~350,000 gaps and the order and orientation of many segments within local regions had not been established. The IHGSC thus turned to the challenge of completing the sequence of the euchromatic genome. Operationally, a finished sequence was defined as having an error rate of at most one per 10^5 bases, and the goal for completion was coverage in finished sequence of at least 95% of the euchromatic genome, with only one gap per 10,000 bases or the shallowest of available technologies (see http://www.gnomon.org/1000929). The goal was challenging because the human genome is complex with numerous dispersed repeats and large segmental duplications, which greatly complicate the determination of genome structure and sequence. In fact, near complete sequences have been obtained so far only for three multicellular organisms: the nematode C. elegans, the mouse and the fly D. melanogaster. The IHGSC sequence is ~10-fold smaller than the human genome, and have much simpler structure.

We describe here the results of a major effort by the IHGSC towards the goal of a complete human sequence. The number of gaps has been reduced 400-fold to only 341, most of which are associated with segmental duplications and will require new methods for resolution. The assembled euchromatic genome sequence has an error rate of only ~1 per 100,000 bases it contains 2,866,000 nucleotides and covers ~99% of the euchromatic genome. This paper describes the current genome sequence and the process used to produce it; examines the accuracy and completeness of the sequence; and illustrates biological analyses made possible by the sequence. We do not attempt here a comprehensive analysis of the contents of the human genome. An initial analysis was previously reported; a series of papers is being written describing the individual chromosomes 1–22, including association of genes and other features.
Figure 4 Segmental duplications across the genome. a, Segmental duplications and sequence gaps across the genome. Segmental duplications are indicated below the chromosomes in blue (length ≥ 10 kb and sequence identity ≥ 95%). Large duplications are shown to approximate scale; smaller ones are indicated as ticks. Sequence gaps are indicated above the chromosomes in red. Large gaps (>300 kb) are shown to approximate scale; smaller gaps are indicated as ticks with those that are 50 kb or smaller shown as shorter ticks. Unfinished clones are indicated as black ticks. b, Percentage of
Single Nucleotide Polymorphisms (SNPs) and Short Insertion/Deletions in The Human Genome
Illumina Solexa Sequencing Technology

Preparation of genomic DNA sample:
- Randomly fragment genomic DNA
- Ligate adapters to both ends of the fragments
Illumina Solexa Sequencing Technology

**Sequencing-By-Synthesis Demo**

- Adapter
- DNA fragment
- Dense lawn of primers
- Adapter

Attach DNA to surface
- Bind single stranded fragments randomly to the inside surface of the flow cell channels.
Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

Attached terminus
Free terminus

Fragments become double stranded
Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

Attached

Denature the double stranded molecules
Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

Clusters

Completion of amplification

On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.
Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

Laser
Illumina Solexa Sequencing Technology

Sequencing-By-Synthesis Demo

Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed.
Illumina Solexa Sequencing Technology

Second Chemistry Cycle: determine second base

To initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.
Image of second chemistry cycle is captured by the instrument. After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
Illumina Solexa Sequencing Technology

**Sequencing-By-Synthesis Demo**


Sequence read over multiple chemistry cycles. Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.
Pacific Biosciences SMRT Sequencing
Figure 6. ZMW with DNA polymerase and phospholinked nucleotides

Phospholinked nucleotides are added into the ZMW at the high concentrations required for proper enzyme functioning.
Circular Templates Gives Redundant Sequencing and Accuracy
Ion Torrent Sequencing

Semiconductor Sequencing for Life™

fastest next-gen workflow
10X more throughput
fastest-selling sequencer
all in six months

Ion Torrent
Ion 316™ -- Everything moves faster
when The Chip is the Machine™

Watch the video >
Read more >

Publications

An integrated semiconductor
device enabling non-optical
genome sequencing

Application Note

The Ion PGM™ sequencer exhibits
superior long-read accuracy
Learn more >

Life Grand Challenges

August 2011
Combating Superbug...more >

Ion Torrent long read accuracy,
download app note & data...more >

View more Ion News...go >
Ion Torrent Sequencing

Semiconductor Sequencing for Life™

www.iontorrent.com

1 877 SEQUENCE = 1 877-378-3623
Ion Torrent Sequencing
Ion Torrent Promises $1,000 Genome

http://uk.reuters.com/article/2012/01/10/us-dna-reader-idUKTRE8090B820120110

Insight: New DNA reader to bring promise

By Sharon Begley
NEW YORK | Tue Jan 10, 2012 7:06pm GMT

(Reuters) - After years of predictions that the "$1,000 genome" - a read-out of a person's complete genetic information for about the cost of a dental crown - was just around the corner, a U.S. company is announcing Tuesday that it has achieved that milestone and taken the technology several steps ahead.
Illumina Announces $1,000 Genome
J.P. Morgan Tech Show 1-16-2014

Population power. Extreme throughput. $1,000 human genome.

The HiSeq X Ten is a set of ten ultra-high-throughput sequencers, purpose-built for large-scale human whole-genome sequencing.

The First $1000 Genome
Discover how HiSeq X Ten breaks the $1000 genome barrier for human whole-genome sequencing.
Learn more »

Population Scale Studies
Learn how the HiSeq X Ten can benefit communities by enabling them to sequence their entire population.
Read blog post »
Cost per Raw Megabase of DNA Sequence

Moore's Law

NIH National Human Genome Research Institute

genome.gov/sequencingcosts
The Human Genome
How fast is the cost going down?

- 2006: $50 million
- 2008: $500,000
- 2009: $50,000
- 2010: $20,000
- 2011: $5,000
- 2012: $4,000
- 2013: $3,000
- 2014: $1,400
- 2015: $1,000

Thanks to Seraf in Batzoglou