The Human Genome Project

Biol4230  Tues, March 21, 2017
Bill Pearson  wrp@virginia.edu  4-2818  Jordan 6-057

• A brief history of DNA and genomes
• The Human Genome Project
• The Draft Human Genome (2001)
  – history and strategy
  – quality metrics
  – human biology
  – viewing genomes
  – computing on genomes
• Next Generation Genomes

To learn more:

1. Pevsner, Chapter 19 pp. 791 – Human Genome
2. Pevsner, Chapter 18 pp. 729 – Eukaryotic Genomes
The human genome sequence

- Assembled from pieces
  - PFP clone by clone, Celera Whole Genome Shotgun
  - Some regions hard to clone, some regions (repeats) hard to assemble
  - not complete, not perfect
- Determined from multiple individuals
  - an initial set of SNPs (single nucleotide polymorphisms) that can track variation
- Gene prediction (ab initio) is useless
  - virtually all gene predictions based on earlier evidence
  - no new gene types
  - many new genes (additional paralogs, duplications)

What is in a genome?


<table>
<thead>
<tr>
<th></th>
<th>E. coli</th>
<th>Plas.</th>
<th>Yeast</th>
<th>Plant (ARATH)</th>
<th>Homo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size(Mb)</td>
<td>4.64</td>
<td>22.8</td>
<td>12.5</td>
<td>115</td>
<td>3289</td>
</tr>
<tr>
<td>Genes</td>
<td>4288</td>
<td>5268</td>
<td>5770</td>
<td>25.5K</td>
<td>~25K</td>
</tr>
<tr>
<td>kb/Gene</td>
<td>0.95</td>
<td>4.34</td>
<td>2.09</td>
<td>4.53</td>
<td>27</td>
</tr>
<tr>
<td>%coding</td>
<td>87.8</td>
<td>52.6</td>
<td>70.5</td>
<td>28.8</td>
<td>1.3</td>
</tr>
<tr>
<td>introns</td>
<td>0</td>
<td>7406</td>
<td>272</td>
<td>107K</td>
<td>53K</td>
</tr>
<tr>
<td>repeat%</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>2.4</td>
<td>15</td>
<td>46</td>
</tr>
</tbody>
</table>

Pevsner, Table 16-1

fasta.bioch.virginia.edu/biol4230
A history of genomes

Landmarks in genetics and genomics


fasta.bioch.virginia.edu/biol4230

5


fasta.bioch.virginia.edu/biol4230

6
The Human Genome Project

Sequencing capacity (2004)
- 40 million lanes/year ~ 8 billion bases – Whitehead Inst. (1 bacterial genome/day)
- >40 billion bases/year, world-wide
  (1000 bacterial genomes/year; 1 mammalian genome/year)

Sequencing capacity (2011) – Illumina sequencing 200 billion bases/week/machine, ~30,000 human genomes/year

Sequencing capacity (2015) – at least 300,000 human genomes/year
Sequencing Timeline

1.8 Mbp
Bacterial genome sequencing
H. influenzae
E. coli
S. cerevisiae sequencing
C. elegans sequencing
D. melanogaster sequencing
A. thaliana sequencing

120 Mbp

PFP (Publicly Funded Project) Hierarchical Sequence Strategy


Hierarchical shotgun sequencing

Genomic DNA

BAC library

Organized mapped large clone contigs

BAC to be sequenced

Shotgun clones

Shotgun sequence

Assembly

fasta.bioch.virginia.edu/biol4230
Celera’s general approach:

Random Shotgun of 27.27 Million reads, ave length 543bp.

16 Libraries from five donors
   2, 10, and 50 kb libraries

Total of 5.1X Coverage of genome. (3.6X from one donor)
Mult. Capillary Sequencers (ABI 3700): 175,000 reads/day

Used Genbank BAC as of Sept 2000:
   4.3Gbp of 20% finished and 75% rough draft sequence.
   Created a 3X coverage from data using a random shredding program that yielded 550 bp reads.

Combined 8X dataset.


Celera Approach #1:
WGA (Whole-genome assembly Schematic)

**Celera Approach #2:**
**CSA (Compartmentalized Shotgun Assembly)**

- Build scaffolds
- Partially sequenced
- Aligned BAC from PFP
- Shotgun clones from 3 different size libraries

**Distribution of Scaffold sizes**
(Not the same as sequence Contigs)

CSA versus PFP for Chr 8 over 1Mbp


CSA versus PFP for Chr 22 (Finished)
Determining Gene Number in Genome is Hard

Developed a homology/evidence based system called Otto.

Otto searches scaffold sequences for homology against known protein, (RefSeq,) EST, and runs 3 de novo gene prediction programs to see if areas of homology are consistent with a gene transcript.

De novo sequences include all gene-prediction transcripts from GRAIL, Genscan, and FgenesH sorted on the basis of matches to EST, protein, or other mouse rat libraries.

Predicted Genes [26,588 - ~39k]

<table>
<thead>
<tr>
<th>evidence</th>
<th>&gt;=1</th>
<th>&gt;=2</th>
<th>&gt;=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Otto</td>
<td>17,968</td>
<td>17,501</td>
<td>15,877</td>
</tr>
<tr>
<td>De Novo</td>
<td>21,350</td>
<td>8,619</td>
<td>4,947</td>
</tr>
</tbody>
</table>

Distribution of transcripts with varying exon number

Fig. 9. Comparison of the number of exons per transcript between the 17,968 Otto transcripts and 21,350 de novo transcript predictions with at least one line of evidence that do not overlap with an Otto prediction. Both sets have the highest number of transcripts in the two-exon category, but the de novo gene predictions are skewed much more toward smaller transcripts. In the Otto set, 19.7% of the transcripts have one or two exons, and 5.7% have more than 20. In the de novo set, 49.3% of the transcripts have one or two exons, and 0.2% have more than 20.

Exon Length

Intron Length

Lander Nature 409, 860–921 (2001), Fig. 35
Segmental duplication versus Retrotransposition

Retrotransposition of mRNA's into the genome converts a gene with introns into an intronless gene flanked by direct repeats and containing a poly A at the 3' end.

Most of these genes will be reverse transcribed badly
=> become inactivated genes (pseudogenes.)
  901 found, 97 appear to be functional

Segmental duplication is a duplication on the DNA level within or between chromosomes.
Detailed view of Duplication events on Chr18 and and 20

**Fig. 13.** Segmental duplications between chromosomes in the human genome. The 24 panels show the 1077 duplicated blocks of genes, containing 10,310 pairs of genes in total. Each line represents a pair of homologous genes belonging to a block; all blocks contain at least three genes on each of the chromosomes where they appear. Each panel shows all the duplications between a single chromosome and other chromosomes with shared blocks. The chromosome at the center of each panel is shown as a thick red line for emphasis. Other chromosomes are displayed from top to bottom within each panel ordered by chromosome number. The inset (bottom, center right) shows a close-up of one duplication between chromosomes 18 and 20, expanded to display the gene names of 12 of the 64 gene pairs shown.
Comparison of Repeats from Celera and PFP

<table>
<thead>
<tr>
<th>Class of Repeat</th>
<th>Celera (PFP)</th>
<th>PFP (Celera)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alu</td>
<td>9.9%</td>
<td>9.9%</td>
</tr>
<tr>
<td>Mammalian interspersed repeat (MIR)</td>
<td>2.3%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Medium reiteration (MER)</td>
<td>1.7%</td>
<td>1.7%</td>
</tr>
<tr>
<td>Long Terminal Repeat (LTR)</td>
<td>5.3%</td>
<td>5.3%</td>
</tr>
<tr>
<td>Long interspersed nucleotide element (LINE)</td>
<td>16.1%</td>
<td>16.1%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>35%</strong></td>
<td><strong>35%</strong></td>
</tr>
</tbody>
</table>

### Classes of interspersed repeat in the human genome

<table>
<thead>
<tr>
<th>Class</th>
<th>Length</th>
<th>Copy number</th>
<th>Fraction of genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINEs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autonomous ORF1</td>
<td>6-8 kb</td>
<td>800,000</td>
<td>21%</td>
</tr>
<tr>
<td>ORF2 (pol) AAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-autonomous</td>
<td>100-300 bp</td>
<td>1,500,000</td>
<td>13%</td>
</tr>
<tr>
<td>SINEs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autonomous gag pol (ren)</td>
<td>6-11 kb</td>
<td>450,000</td>
<td>8%</td>
</tr>
<tr>
<td>Non-autonomous</td>
<td>1.5-3 kb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retrovirus-like elements</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autonomous transposase</td>
<td>2-3 kb</td>
<td>300,000</td>
<td>3%</td>
</tr>
<tr>
<td>Non-autonomous</td>
<td>80-3,000 bp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA transposon fossils</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lander Nature **409**, 860–921 (2001), Fig. 35  
fasta.bioch.virginia.edu/biol4230

SNP (Single Nucleotide Polymorphism) frequency varies in the genome more than predicted.

Fig. 14. SNP density in each 100-kbp interval as determined with Celera-PFP SNPs. The color codes are as follows: black, Celera-PFP SNP density; blue, coalescent model; and red, Poisson distribution. The figure shows that the distribution of SNPs along the genome is nonrandom and is not entirely accounted for by a coalescent model of regional history.


fasta.bioch.virginia.edu/biol4230
Gene duplication in complete protein clusters (Lek)

Human Chromosomes colored by mouse chromosomes

Figure 46 Conserved segments in the human and mouse genome. Human chromosomes, with segments containing at least two genes whose order is conserved in the mouse genome as colour blocks. Each colour corresponds to a particular mouse chromosome. Centromeres, subcentromeric heterochromatin of chromosomes 1, 9 and 16, and the repetitive short arms of 13, 14, 15, 21 and 22 are in black.

Lander Nature 409, 860–921 (2001), Fig. 35
Retrieving the data: Genome Browsers (UCSC)

Genome Browsers (UCSC)
Genome data (UCSC table browser)

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form. The Users Guide for general information and sample queries, and the OpenHelix Table Browser Tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clad: | 
---|---
Human | 1

Genome data (UCSC table browser)

| UCSC Genes (knownGene) Summary Statistics |
|---|---|
| Item count | 30 |
| Item bases | 48,074 (40.49%) |
| Item total | 133,625 (112.39%) |
| Smallest item | 30 |
| Average item | 4,454 |
| Biggest item | 19,211 |
| Block count | 131 |
| Block bases | 15,231 (12.80%) |
| Block total | 35,507 (29.84%) |
| Smallest block | 26 |
| Average block | 271 |
| Biggest block | 3,238 |

Region and Timing Statistics

<table>
<thead>
<tr>
<th>Region</th>
<th>chr1:109631271-109750270</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bases in region</td>
<td>119,900</td>
</tr>
<tr>
<td>Bases in gaps</td>
<td>0</td>
</tr>
<tr>
<td>Load time</td>
<td>0.02</td>
</tr>
<tr>
<td>Calculation time</td>
<td>0.00</td>
</tr>
<tr>
<td>Free memory time</td>
<td>0.00</td>
</tr>
<tr>
<td>Filter</td>
<td>off</td>
</tr>
<tr>
<td>Intersection</td>
<td>off</td>
</tr>
</tbody>
</table>

fasta.bioch.virginia.edu/biol4230 33
Genome data (UCSC table browser)

<table>
<thead>
<tr>
<th>chr</th>
<th>hg38_refGene</th>
<th>at_codon</th>
<th>tx_st</th>
<th>gene_id</th>
<th>transcript_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>hg38_refGene</td>
<td>CDS</td>
<td>109620269-109620271</td>
<td>0.000000</td>
<td>+</td>
</tr>
<tr>
<td>chr1</td>
<td>hg38_refGene</td>
<td>exons</td>
<td>109625303</td>
<td>0.000000</td>
<td>2</td>
</tr>
<tr>
<td>chr1</td>
<td>hg38_refGene</td>
<td>CDS</td>
<td>109625662-109625792</td>
<td>0.000000</td>
<td>+</td>
</tr>
<tr>
<td>chr1</td>
<td>hg38_refGene</td>
<td>exons</td>
<td>109626160</td>
<td>0.000000</td>
<td>1</td>
</tr>
<tr>
<td>chr1</td>
<td>hg38_refGene</td>
<td>CDS</td>
<td>109626726-109626912</td>
<td>0.000000</td>
<td>+</td>
</tr>
<tr>
<td>chr1</td>
<td>hg38_refGene</td>
<td>exons</td>
<td>109627175</td>
<td>0.000000</td>
<td>2</td>
</tr>
<tr>
<td>chr1</td>
<td>hg38_refGene</td>
<td>CDS</td>
<td>109627429-109627518</td>
<td>0.000000</td>
<td>+</td>
</tr>
<tr>
<td>chr1</td>
<td>hg38_refGene</td>
<td>exons</td>
<td>109627774</td>
<td>0.000000</td>
<td>1</td>
</tr>
</tbody>
</table>

GFF/GTF format

fasta.bioch.virginia.edu/biol4230

www.ensembl.org
The human genome sequence

• Assembled from pieces
  – PFP clone by clone, Celera Whole Genome Shotgun
  – Some regions hard to clone, some regions (repeats) hard to assemble
  – not complete, not perfect
• Determined from multiple individuals
  – an initial set of SNPs (single nucleotide polymorphisms) that can track variation
• Gene prediction (ab initio) is useless
  – virtually all gene predictions based on earlier evidence
  – no new gene types
  – many new genes (additional paralogs, duplications)

The human genome – initial insights

1. There were reported to be about 30,000 to 40,000 predicted protein-coding genes in the human genome. Currently, ENSEMBL reports 20,300 protein coding genes. Similar to Arabidopsis (plant, 26,000 genes) and pufferfish (21,000 genes), and marginally more genes than are found in many nematode and insect genomes.
2. A small number (~100) of genes may have been acquired "laterally", not directly, from bacteria or other organisms.
3. More than 98% of the human genome does not code for genes. Much of this genomic landscape is occupied by repetitive DNA elements such as long interspersed elements (LINEs) (20%), short interspersed elements (SINEs) (13%), long terminal repeat (LTR) retrotransposons (8%), and DNA transposons (3%). Thus half the human genome is derived from transposable elements.
4. Segmental duplication is frequent, particularly in pericentromeric and subtelomeric regions. More common in humans than in yeast, fruitfly, or worm genomes.
5. There are several hundred thousand Alu repeats in the human genome. These have been thought to represent elements that replicate promiscuously. However, their distribution is nonrandom: they are retained in GC-rich regions.
6. The mutation rate is about twice as high in male meiosis than in female meiosis. This suggests that most mutation occurs in males.
7. More than 1.4 million single nucleotide polymorphisms (SNPs) were identified. SNPs are single nucleotide variations that occur once every 100 to 300 base pairs (bp). 36 million in Oct., 2014
Sequencing capacity (2011) – Illumina sequencing 200 billion bases/week/machine, ~30,000 human genomes/year
Sequencing capacity (2015) – at least 300,000 human genomes/year

DNA sequencing technology pre-1998

1977 – 1985
radioactive

1985 – 1995
dye-terminators

fasta.bioch.virginia.edu/bio4230
Adapted from Richard Wilson, School of Medicine, Washington University, "Sequencing the Cancer Genome". http://tinyurl.com/5f3alk

<table>
<thead>
<tr>
<th>Genome size:</th>
<th>3 Gb == 3000 Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Req'd coverage:</td>
<td>6 12 24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>3730</th>
<th>454 FLX</th>
<th>HiSeq</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp/read</td>
<td>600</td>
<td>500</td>
<td>200</td>
</tr>
<tr>
<td>Reads/run</td>
<td>96</td>
<td>1,000,000</td>
<td>180,000,000</td>
</tr>
<tr>
<td>bp/run</td>
<td>57,600</td>
<td>500,000,000</td>
<td>4E+10</td>
</tr>
<tr>
<td>#:runs req'd</td>
<td>312,500</td>
<td>72</td>
<td>2</td>
</tr>
<tr>
<td>Cost per run</td>
<td>$48</td>
<td>$7,500</td>
<td>$5,000</td>
</tr>
<tr>
<td>Total cost</td>
<td>$15,000,000</td>
<td>$540,000</td>
<td>$10,000</td>
</tr>
</tbody>
</table>

source: Francis Ouellette, OICR
fasta.bioch.virginia.edu/biol4230

Next (2\textsuperscript{nd}, 3\textsuperscript{rd}, 4\textsuperscript{th}) Generation Technologies

fasta.bioch.virginia.edu/biol4230
Next (2\textsuperscript{nd}, 3\textsuperscript{rd}, 4\textsuperscript{th}) Generation Technologies

Table 3 | Sequencing platform comparison

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Illumina SOLiD</th>
<th>Roche 454</th>
<th>454 Life Sciences RSO (HiSeq)</th>
<th>Pacific Biosciences RS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library amplification method</td>
<td>emPCR(^*) on bead surface</td>
<td>emPCR(^*) on bead surface</td>
<td>Enzymatic amplification on glass surface</td>
<td>NA (single molecule detection)</td>
</tr>
<tr>
<td>Sequencing method</td>
<td>Polymerase-mediated incorporation of unlabelled nucleotides</td>
<td>Ligation-mediated addition of 2-base encoded fluorescent oligonucleotides</td>
<td>Polymerase-mediated incorporation of end-blocked fluorescent nucleotides</td>
<td>Polymerase-mediated incorporation of terminal phosphate labelled fluorescent nucleotides</td>
</tr>
<tr>
<td>Detection method</td>
<td>Light emitted from secondary reactions initiated by release of PPi</td>
<td>Fluorescent emission from ligated dye-labelled oligonucleotides</td>
<td>Fluorescent emission from incorporated dye-labelled nucleotides</td>
<td>Real time detection of fluorescent dye in polymerase active site during incorporation (PPi removed as part of PPi release on nucleic acid incorporation)</td>
</tr>
<tr>
<td>Post incorporation method</td>
<td>NT (unlabelled nucleotides are added in base-specific fashion, followed by detection)</td>
<td>Substitution errors rare, insertion/deletion errors at homopolymer 400bp variable-length read pairs</td>
<td>End of read substitution errors 150bp/100-1000bp</td>
<td>Random insertion/deletion errors &gt;1,000bp</td>
</tr>
</tbody>
</table>

Comparison of commercially available next generation platforms (Illumina SOLiD, 454 Life Sciences and Roche 454) and a single molecule platform (Pacific Biosciences), illustrating the similarities and differences in these technologies, according to several metrics. NA, not applicable; PPi, pyrophosphate.

\(^*\) emPCR (emulsion PCR) is a bulk amplification process whereby library fragments are combined with beads and PCR reagents in an emulsion that allows emulsion amplification of millions of bead DNA combinations in a single tube.


Cost per Raw Megabase of DNA Sequence

Moore's Law

NIH National Human Genome Research Institute
genome.gov/sequencingcosts

fasta.bioch.virginia.edu/biol4230 43
NGS applications

• genome (re)sequencing
  – de novo genomes: 454 in Bact, small Euks, PacBio
  – SNP discovery and genotyping (barcoded pools), population genetics: Illumina
  – targeted, “deep” gene resequencing
  – metagenomics
• structural/copy-number variation
  – Tumor genome SV/CNV: Illumina/PET
• RNA-seq: transcriptomics
• ChIP/CLIP/etc-seq: DNA/RNA-binding, or DNA/RNA-modification
• Chromatin conformation capture (3C, Hi-C, etc.)

three phases of analysis

• primary
  – conversion of raw machine signal into sequence and qualities, QC, filtering, trimming, etc.
• secondary
  – read alignment to reference genome or transcriptome
  – or de novo assembly of reads into contigs
• tertiary
  – SNP discovery/genotyping
  – transcript clustering/quantification (RNA)
  – peak discovery/quantification (ChIP)
primary analyses

- Illumina GA Pipeline:
  - Firecrest: raw image -> clusters
  - Bustard: clusters -> sequence reads
  - Gerald/Eland -> raw alignment, sequence updates
- 454/IonTorrent: convert flowgram to FASTQ
- PacBio: decode video images as FASTQ, etc.
- core labs do these primary analyses for you
- “raw” image/video files are huge, and not stored
  - new primary analysis tools can’t be re-run on old data

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATA
+HWUSI-EAS100R:6:73:941:1973#0/1
!''*(((****))%%%++)(%%%)(%**%
```
**Phred-scaled** base qualities

\[
Q_{\text{Phred}} = -10 \log_{10} \left( P_{\text{err}} \right)
\]

\[
Q_{\text{Solexa}} = -10 \log_{10} \left( \frac{P_{\text{err}}}{1 - P_{\text{err}}} \right)
\]

\[
Q_{\text{Sanger}} = 33 + \min(Q_{\text{Phred}}, 40)
\]

---

**conversion of sol/ill qualities**

- **“maq” package contains** `sol2sanger` and `ill2sanger` **utilities to convert to standard Phred-scaled quality encoding**

- Or, use `galaxy.org “FASTQ Groomer”`
read filtering/trimming/QC

• newer Illumina pipelines deliver unfiltered reads, with “chastity” filter tags:
  @EAS139:136:FC706VJ:2:5:1000:12850 1::Y:18:ATCACG
  - pipeline version dictates whether Y means “bad” (1.8+, recent) or “good” (pre-1.8)
• chimeric reads containing adapters, primers, etc. should be trimmed (sickle, scythe)
• barcoding, merging, data manipulations
• FASTQC

secondary analysis

• alignment back to the reference
  - computationally demanding – can’t use BLAST
  - many algorithms (Maq, BWA, bowtie, Mosaik, NovoAlign, SOAP2, SSAHA, …)
    • sensitivity to seq. errors, polymorphisms, indels, rearrangements?
    • heuristic tradeoffs in time vs. memory vs. performance
The human genome sequence

• Assembled from pieces
  – PFP clone by clone, Celera Whole Genome Shotgun
  – Some regions hard to clone, some regions (repeats) hard to assemble
  – not complete, not perfect

• Determined from multiple individuals
  – an initial set of SNPs (single nucleotide polymorphisms) that can track variation

• Next Generation Sequencing puts genome data in experimenters hands