Computational Biology 2
Lecture 3

Special Human Genome issues

Pawan Dhar

BII

Feb 2001
Q. What is the Human Genome Project?

A massive research effort to determine complete sequence of 3 billion bases long human DNA and discover all the ~ 30,000 to 35,000 genes

Started in 1990

Q. What is the current status of HGP?

Rough draft of the human genome completed in June 2000. To complete the finished high quality sequence by 2003!!
A quick look...

**Q. How far the project has gone?**

The Draft sequence (4x - 5x) was published in Feb 2001. Length of fragments 10,000 bases, approx. location known. Finished sequence: 63%.

**Q. What is the difference between draft sequence and high quality sequence?**

Draft sequence: gaps and errors, 4-5 time coverage. Finished sequence: 1 error / 10,000 bases, 8x-9x coverage.
Q. What are the goals of the Human Genome project?

Short term
1. Generate working draft of 90% of human genome
2. Obtain high quality sequence by 2003
3. Make all data publicly available

Long term
1. Create novel sequencing techniques
2. Develop rapid identification tool for DNA variants
3. Characterize genes through functional genomics
4. Initiate large scale comparative genomics
5. Identify ELSI of human genomics
6. Develop new databases and tools for data generation, capture and functional studies
Q. What happens when HGP is completed?

The next step would be to determine

* Gene number, exact locations, and functions
* Gene regulation
* DNA sequence organization
* Chromosomal structure and organization
* Non-coding DNA types, amount, distribution, information content, and functions
A quick look…

continued …

* Coordination of gene expression, protein synthesis, and post-translational events
* Protein - protein Interaction
* Predicted vs experimentally determined gene function
* Evolutionary conservation among organisms
A quick look...

concluded

* Protein conservation (structure and function)
* Proteomes (total protein content and function)
* Correlation of SNPs with health and disease
* Disease-susceptibility prediction based on gene sequence variation
* Genes involved in complex traits and multigene diseases
* Complex systems biology
* Developmental genetics, genomics
Q. When is a genome completely sequenced?

Case studies (finished sequence)
- Drosophila: 120 Mb / 180 Mb
- Human #22: 33.5 Mb / 56 Mb

Regions that cannot be sequenced:
telomeres, centromeres

Human chromosome 22:
60% sequenced region contains 97% euchromatin
Key accomplishments that sparked Human Genome Sequencing Project

(i) 1977-82: Sequencing of bacterial virus ØX74, lambda phage, SV40 and human mitochondrion

(ii) 1980: Creation of human genetic map (Botstein)

(iii) Mid 80s: Creation of physical maps of yeast & worm genomes chromosomal position (Olson and Sulston)

(iv) Development of shotgun sequencing (Schimmel), automated DNA sequencers (Leroy Hood etc.)
Human Genome Project

Pharmaceuticals
Agriculture
Forensics
Anthropology
Bioremediation

Biofuels
Industrial Processes
Q. What are potential benefits of HGP?

1- Molecular Medicine

* Improved diagnosis of disease
* Genetic predispositions to disease
* Rational drug design
* Gene therapy
* Pharmacogenomics "custom drugs"
Potential benefits of HGP (continued…)

2- Microbial Genomics

* New energy sources (biofuels)
* Environmental monitoring to detect pollutants
* Protection from biological and chemical warfare
* Safe, efficient toxic waste cleanup
* Understanding disease vulnerabilities and revealing drug targets
Potential benefits of HGP (continued…)

3- Risk Assessment

* Assess health damage and risks caused by radiation exposure, including low-dose exposures

* Assess health damage and risks caused by exposure to mutagenic chemicals and cancer-causing toxins

* Reduce the likelihood of heritable mutations
4- Bioarchaeology, Anthropology, Evolution and Human Migration

* Study evolution through germline mutations in lineages

* Study migration of different population groups based on female genetic inheritance

* Study mutations on the Y chromosome to trace lineage and migration of males compare breakpoints in the evolution of mutations with
Potential benefits of HGP (continued…)

5 - DNA Forensics (Identification)

* identify potential suspects
* exonerate persons wrongly accused of crimes
* identify crime and catastrophe victims
* establish paternity and other family relationships
* identify endangered and protected species
* detect bacteria and other organisms that may pollute air, water, soil, and food
* match organ donors with recipients in transplant programs
* determine pedigree for seed or livestock breeds
* authenticate consumables such as caviar and wine ages of populations and historical events
Potential benefits of HGP (continued…)

6 - Agriculture, Livestock Breeding, and Bioprocessing

* disease-, insect-, and drought-resistant crops
* healthier, more productive, disease-resistant farm animals
* more nutritious produce
* biopesticides
* edible vaccines incorporated into food products
* new environmental cleanup uses for plants like tobacco
Q. What will future genomic scientists look like?

Experts in biology, computer science, engineering, mathematics, physics, chemistry and management !!!!
Salient features of human genome

1. Large variation in the distribution of genes, transposable elements, GC content, CpG islands and recombination rate. HOX genes: most repeat poor genes in the human genome

2. 30,000 - 40,000 human genes
   Genes spotted so far: 22,000
   Arabidopsis: 25,498, Worm: 19,099
   More complex (i.e., more alternative splicing)
Salient features …

3. The human proteome is huge and complex

4. 100s of genes (bacterial transfer)
   Dozens of genes (transposable elements)
5. About half of human genome derives from transposable elements there is a marked decline in the overall activity of such elements in the hominid lineage.

6. The pericentromeric and subtelomeric regions of chromosomes have large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in the yeast, fly or worm.
7. Alu analysis: there may be strong selection in favor of preferential retention of Alu elements in the GC rich regions and that these ‘selfish’ elements may benefit their human hosts.

8. The mutation rate is about twice as high in males as in female meiosis. Thus, most mutation occurs in males.

9. Large GC-poor regions are strongly correlated with ‘dark G-bands’ in karyotypes.
10. Recombination rates tend to be much higher in the distal regions (around 20 Mb) of chromosomes and on shorter chromosome arms. In general, in a pattern that promotes the occurrence of at least one crossover per chromosome per arm in each meiosis.

11. >1.4 million SNPs have been identified.
The unfinished story...

Genome sequencing is still unfinished, lots of gaps remain - especially with respect to heterochromatic segments which confound even the best sequencing machines.

~88% genome has been successfully cloned and sequenced.

Church in Barcelona
By: Gaudi
Some interesting facts

~ 35% of all human genes may be read in several ways. Thus the human genome can encode five times as many proteins as fruitfly or roundworm.

The human genome is a museum of viral infection suffered by humanity

Genes encoding ~223 proteins seem to have come from bacteria

“In places genome looks like a sea of reverse transcribed DNA with a small admixture of genes” - David Baltimore
Conclusion: All humans are Genetically Modified Organisms!
Some interesting facts

Human genome contains 200 times more DNA than yeast but 200 times less DNA than amoeba!

Q: Why is it hard to detect human genes?

Answer: 1. Gene density

Gene density / million bases

<table>
<thead>
<tr>
<th>Number of Genes</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Human</td>
</tr>
<tr>
<td>117</td>
<td>Fruit fly</td>
</tr>
<tr>
<td>197</td>
<td>Roundworm</td>
</tr>
<tr>
<td>221</td>
<td>Arabidopsis</td>
</tr>
</tbody>
</table>
Some interesting facts

Answer 2: Human genes are highly fragmented (small exons and longer-than-average introns)
Introns: 87 - 3300 bases, Exons: lower limit 19 bases

Largest human gene: Dystrophin (2.4 million bases)
Encodes the muscle protein 'dystrophin'
Most of the gene is non-coding.

Record-holder for coding sequence: 'titin'
Length: 80,780 bases, divided into 178 exons,
Largest titin-exon: 17,106 bases
Genome sequencing finished till 2001

599 viruses and viroids
205 naturally occurring plasmids
185 organelles
31 eubacteria
7 archaea
1 fungus
2 animals and
1 plant.
Whole genome or hierarchical?

HGP’s choice

Hierarchical
Celera’s choice

Whole genome shotgun sequencing

Why?

1. Lower cost
2. Faster results
3. Incorporated IHGSC results
4. Accuracy?
Technology for large scale sequencing

Key innovations

Wet bench
1. 4 color fluorescence based sequence detection
2. Improved fluorescent dyes
3. Dye labeled terminators
4. Polymerases specifically designed for sequencing
5. Cycle sequencing
6. Capillary gel electrophoresis

Dry

PHRED and PHRAP software package
Phred reads DNA sequencer trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to output files.

1. It predicts peak locations using the assumption that fragments should be locally evenly spaced. This helps determine the correct number of bases in a region where peaks are not well resolved, noisy or displaced.

2. Observed peaks are identified in the trace and matched to the predicted peak locations. Some are omitted and some are split, yielding the main base sequence.

3. Unmatched peaks are analyzed to see if they represent bases, and if so is inserted into the sequence.
a program for assembling shotgun DNA sequence data

1. Masking likely `garbage' sequence at the beginning and end of reads. Phrap looks for regions consisting almost entirely of a single base; such regions are likely due to poor data quality.

2. Identifying all potentially overlapping pairs of sequences. Two overlapping sequences must have at least one exact match of length \textit{minmatch} (typically 14 bases), and the alignment between the sequences must have a score of at least \textit{minscore} (default: 30).
3. Modified quality scores are calculated for each base in each read, taking into account confirmation by overlapping reads as well as orientation and sequencing chemistry. "LLR" scores are computed for each pairwise alignment, measuring of overlap length and quality. High quality discrepancies that potentially indicate different copies of a repeat lead to low LLR scores. Potential problem clones like chimeras are also identified.

4. Merge reads into contigs, starting at the pairwise overlaps with the highest LLR scores. Likely chimeras are not used.

5. A consensus sequence is extracted from the merged sequence, based on voting from the highest adjusted quality scores at any base.
Automated sequencing centers

100,000 sequencing reactions / 12 hours
URL for protocols

http://www.nhgri.nih.gov/genome_hub
Sequencing progress
Key steps in assembling individual sequenced clones into draft genome
Quality control

* Nucleotide accuracy determined by PHRAP score.

* Finished portion of genome sequence: error rate of less than 1 per 10,000 bases (PHRAP SCORE > 40)

<table>
<thead>
<tr>
<th>PHRAP score</th>
<th>Percentage of bases in the draft genome sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–9</td>
<td>0.6</td>
</tr>
<tr>
<td>10–19</td>
<td>1.3</td>
</tr>
<tr>
<td>20–29</td>
<td>2.2</td>
</tr>
<tr>
<td>30–39</td>
<td>4.8</td>
</tr>
<tr>
<td>40–49</td>
<td>8.1</td>
</tr>
<tr>
<td>50–59</td>
<td>8.7</td>
</tr>
<tr>
<td>60–69</td>
<td>9.0</td>
</tr>
<tr>
<td>70–79</td>
<td>12.1</td>
</tr>
<tr>
<td>80–89</td>
<td>17.3</td>
</tr>
<tr>
<td>&gt;90</td>
<td>35.9</td>
</tr>
</tbody>
</table>

PHRAP scores are logarithmically based representation of the error probability. A PHRAP score of X corresponds to an error probability of $10^{-X/2}$. Thus, PHRAP scores of 20, 30 and 40 correspond to accuracy of 99%, 99.9% and 99.99%, respectively. PHRAP scores are derived from quality scores of the underlying sequence reads used in sequence assembly. See http://www.genome.washington.edu/UWGC/analysistools/phrap.htm.
Assembling problems

1. Errors in the initial sequence contigs persist

2. GigAssembler may fail to merge some overlapping sequences because of poor data quality

3. Allelic differences or misassemblies of the initial sequence contigs

Overall artifactual duplication is about 100 Mb