Computational gene finding in the human genome: how many genes do we have?

Steven Salzberg
Director, Center for Bioinformatics and Computational Biology
Horvitz Professor, Dept. of Computer Science
University of Maryland
http://cbcb.umd.edu
We’ve been trying for a long time to determine how many genes we have

letters to nature

Nature 201, 847 (22 February 1964); doi:10.1038/201847a0

A Preliminary Estimate of the Number of Human Genes

F. VOGEL

Institut für Anthropologie und Humangenetik, University of Heidelberg, Germany.

RECENT results of molecular genetics enable us to estimate the number of human genes, if certain assumptions are made. The following data are available: (1) The α-chain of human haemoglobin contains 141, the β-chain contains 146 amino-acids, corresponding to a molecular weight of about 17,000 each. Assuming a triplet code this means that the α- and β-chains are determined by 423 and 438 nucleotide pairs, respectively. According to Svedberg's law, many proteins consist of sub-units of the same order of magnitude (molecular weight of about 17,500). Hence, the assumption seems to be warranted that one average structural gene might have a length of about 450 nucleotide pairs. (2) The weight of one haploid human chromosome set in human spermatozoa is about $2.72 \times 10^{-12}$ g. Granulocytes contain about $6.23 \times 10^{-12}$ g; lymphocytes contain about $5.84 \times 10^{-12}$ g (ref. 5). Extensive examinations have shown that the DNA content is constant in all resting cells of one species, which have the same number of chromosome sets, and depends on the degree of polyploidy. The assumption seems to be justified that most of the DNA works as genetic material, even if in some cells minor fractions with other functions might possibly be present. In the following calculations the total amount of DNA in a haploid human chromosome set is estimated to be about $3 \times 10^{-12}$ g. (3) Usually the genetic variants of human haemoglobins differ in one amino-acid substitution only. One structural gene can only produce one single type of genetically determined polypeptide chain. As much as we know, this applies for other genetically determined proteins as well. This means that the genetic information for these structural genes can only be present once. Any degree of polyteny for these loci in the germ cells is highly unlikely. As has been mentioned, however, the DNA content of diploid cells is about twice the content of (haploid) spermatozoa. We assume that the total genetic information is only present once.
1,000,000 genes? 100,000 genes?

A Gene Map of the Human Genome


The human genome is thought to harbor 50,000 to 100,000 genes, of which about half have been sampled to date in the form of expressed sequence tags. An

PONCEDENCES

Predicting the total number of human genes

Francisco Antequera^1 & Adrian Bird^1

^1Institute of Cell and Molecular Biology, Darwin Building, University of Edinburgh, Scotland EH9 3JR, UK

How many genes in the human genome?

Chris Fields^1, Mark D. Adams^2, Owen White^1 & J. Craig Venter^1

^1The Institute for Genomic Research, 952 Clopper Road, Gaithersburg, Maryland 20878, USA

ARTICLE

Number of CpG Islands and Genes in Human and Mouse

F Antequera and A Bird

Estimation of gene number in mammals is difficult due to the high proportion of noncoding DNA within the nucleus. In this study, we provide a direct measurement of the number of genes in human and mouse. We have taken advantage of the fact that many mammalian genes are associated with CpG islands whose distinctive properties allow their physical separation from bulk DNA. Our results suggest that there are ~45,000 CpG islands per haploid genome in humans and ~37,000 in the mouse. Sequence comparison confirms that about 20% of the human CpG islands are absent from the homologous mouse genes. Analysis of a selection of genes suggests that both human and mouse are losing CpG islands over evolutionary time due to de novo methylation in the germ line followed by CpG loss through mutation. This process appears to be more rapid in rodents. Combining the number of CpG islands with the proportion of island-associated genes, we estimate that the total number of genes per haploid genome is ~80,000 in both organisms.

- 28,000 - 34,000
- Based on alignments to pufferfish (*Tetraodon nigroviridis*)

Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence

Hugues Roest Crollius, Olivier Jaillon, Alain Bernot, Corinne Dasilva, Laurence Bouneau, Cécile Fischer, Cécile Fizames, Patrick Wincker, Philippe Brottier, Francis Quétier, William Saurin & Jean Weissenbach

- Based on expressed sequence tag (EST) alignments to human chromosome 22

- Based on assemblies of ESTs
- Extrapolated to whole human genome

- Correction to avoid over-counting immunoglobulin ESTs
- 81,000 genes, based on assemblies of ESTs
- 57,000 genes, extrapolating from Chr 21 and 22
The gene count guessing game

*Nature* 19 May 2000:
Vol. 288, no. 5469, pp. 1146 – 1147
DOI: 10.1126/science.288.5469.1146

NEWS OF THE WEEK

HUMAN GENOME PROJECT:
And the Gene Number Is ...?
Elizabeth Pennisi

COLD SPRING HARBOR, NEW YORK—Even though a draft sequence of the human genome is nearing completion, biologists still don’t know how many genes it contains. Indeed, the range of estimates seems to be growing rather than shrinking. The question lies at the core of our understanding of genetic complexity. If genomes are the books of life, then genes are the words that tell the story of each organism. Biologists have long assumed that microorganisms are short stories and complex organisms such as humans, great tomes.

Place your bet. Uncertainty over the number of human genes has sparked a debate—and a betting pool.
Human genome paper I:  
*Nature* 409(15 Feb 2001), 860-921

- 30,000 - 40,000 genes
- Large degree of uncertainty about total
- Number of distinct transcripts and proteins even less certain
Human genome paper II: Science 291(16 Feb 2001), 1304-51

- 26,588 genes
- 12,000 additional “likely” genes based on similarity to mouse or other evidence
Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

*A list of authors and their affiliations appears in the Supplementary Information

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 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused 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 seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

• 20,000 - 25,000 genes
How do we find genes?

- *Ab initio* gene finding
- Expressed sequence tags (ESTs)
- Full-length cDNA sequencing
- Alignment of protein sequences to genomic DNA
- Combining all the evidence together
Training a Gene Finder

\[
\theta_{\text{max}} = \arg\max_{\theta} \sum_{(S, \phi) \in K} P(\phi | S, \theta) = \arg\max_{\theta} \frac{\sum_{(S, \phi) \in K} P_t(q_0, q_L) \prod_{i=1}^{L} P_e(x_i | q_i) P_t(q_i | q_{i-1})}{P(S | \theta)}
\]
GlimmerHMM
Eukaryotic Gene-Finding System

OVERVIEW

GlimmerHMM is a new gene finder based on a Generalized Hidden
Markov Model (GHMM). Although the gene finder conforms to the
overall mathematical framework of a GHMM, additionally it incorporates
splice site models adapted from the GeneSplicer program and a decision
tree adapted from GlimmerM. It also utilizes Interpolated Markov Models
for the coding and noncoding models. Currently, GlimmerHMM’s GHMM
structure includes introns of each phase, intergenic regions, and four
types of exons (initial, internal, final, and single). A basic user manual can
be consulted here.

SYSTEM REQUIREMENTS

GlimmerHMM is released as source code and was tested on Linux
RedHat 6.x+, Sun Solaris, and Alpha OSF1, but should work on any Unix
system.

ACCURACY

GlimmerHMM has been trained on several species including
Arabidopsis thaliana, Coccioides species, Cryptococcus neoformans,
and Brugia malayi. New: trainings for C. elegans and Danio rerio (zebrafish)
are now available!

<table>
<thead>
<tr>
<th>Species</th>
<th>Nuc Sens</th>
<th>Nuc Spec</th>
<th>Nuc Accur</th>
<th>Exon Sens</th>
<th>Exon Spec</th>
<th>Exact Genes</th>
<th>Size of test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. rerio</td>
<td>93%</td>
<td>78%</td>
<td>86%</td>
<td>77%</td>
<td>69%</td>
<td>24%</td>
<td>549 genes</td>
</tr>
<tr>
<td>C. elegans</td>
<td>96%</td>
<td>95%</td>
<td>96%</td>
<td>82%</td>
<td>81%</td>
<td>42%</td>
<td>1886 genes</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>97%</td>
<td>99%</td>
<td>98%</td>
<td>84%</td>
<td>89%</td>
<td>60%</td>
<td>809 genes</td>
</tr>
<tr>
<td>Cryptococcus</td>
<td>96%</td>
<td>99%</td>
<td>98%</td>
<td>86%</td>
<td>88%</td>
<td>53%</td>
<td>350 genes</td>
</tr>
<tr>
<td>Coccioides</td>
<td>99%</td>
<td>99%</td>
<td>99%</td>
<td>84%</td>
<td>86%</td>
<td>60%</td>
<td>503 genes</td>
</tr>
<tr>
<td>Brugia</td>
<td>93%</td>
<td>98%</td>
<td>95%</td>
<td>78%</td>
<td>83%</td>
<td>25%</td>
<td>477 genes</td>
</tr>
</tbody>
</table>

GlimmerHMM has been recently trained on human. The table below
presents its performance compared to Genscan on 963 human RefSeq
genes selected randomly from all 24 chromosomes, non-overlapping
with the training set. The test set contains 1000 bp of untranslated
sequence on either side (5’ or 3’) of the coding portion of each gene.

<table>
<thead>
<tr>
<th>Method</th>
<th>Nuc Sens</th>
<th>Nuc Spec</th>
<th>Nuc Acc</th>
<th>Exon Sens</th>
<th>Exon Spec</th>
<th>Exon Acc</th>
<th>Exact Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GlimmerHMM</td>
<td>86%</td>
<td>72%</td>
<td>75%</td>
<td>72%</td>
<td>62%</td>
<td>67%</td>
<td>1.7%</td>
</tr>
<tr>
<td>Genscan</td>
<td>86%</td>
<td>68%</td>
<td>77%</td>
<td>65%</td>
<td>60%</td>
<td>63%</td>
<td>1.3%</td>
</tr>
</tbody>
</table>

Done
How do we find genes?

- *Ab initio* gene finding
- Expressed sequence tags (ESTs)
- Full-length cDNA sequencing
- Alignment of protein sequences to genomic DNA
- Combining all the evidence together
Putting it all together manually

• View *ab initio* predictions and sequence alignments within a genome editor:
Putting it all together automatically with JIGSAW
Evaluating methods on 1% of the human genome: ENCODE

The ENCODE (ENCyclopedia Of DNA Elements) Project

The ENCODE Project Consortium*

The ENCYclopedia Of DNA Elements (ENCODE) Project aims to identify all functional elements in the human genome sequence. The pilot phase of the Project is focused on a specified 30 megabases (~1%) of the human genome sequence and is organized as an international consortium of computational and laboratory-based scientists working to develop and apply high-throughput approaches for detecting all sequence elements that confer biological function. The results of this pilot phase will guide future efforts to analyze the entire human genome.

With the complete human genome sequence now in hand (1–3), we face the enormous challenge of interpreting it and learning how to use that information to understand the biology of human health and disease. The ENCYclopedia Of DNA Elements (ENCODE) Project is predicated on the belief that a comprehensive catalog of the structural and functional components encoded in the human genome sequence will be critical for understanding human biology well enough to address those fundamental aims of biomedical research. Such a complete catalog, or “parts list,” would include protein-coding genes, non-protein-coding genes, transcriptional regulatory elements, and non-coding RNAs.

ENCODE Gene finding ASsessment Project (EGASP)
EGASP results: Exon prediction accuracy among 28 different methods

From Guigo et al., Genome Biology 2006, 7(Suppl 1):S2
### EGASP results: overall gene accuracy

<table>
<thead>
<tr>
<th>Genefinding Method</th>
<th>Sens</th>
<th>Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUGUSTUS-any</td>
<td>47.9</td>
<td>35.5</td>
</tr>
<tr>
<td>FGENESH++</td>
<td>69.9</td>
<td>42.0</td>
</tr>
<tr>
<td>JIGSAW</td>
<td>72.6</td>
<td>65.9</td>
</tr>
<tr>
<td>PAIRAGON-any</td>
<td>69.5</td>
<td>61.3</td>
</tr>
<tr>
<td>AUGUSTUS-abinit</td>
<td>24.3</td>
<td>17.2</td>
</tr>
<tr>
<td>GENEMARK.hmm-A</td>
<td>15.2</td>
<td>3.2</td>
</tr>
<tr>
<td>GENEMARK.hmm-B</td>
<td>16.8</td>
<td>7.9</td>
</tr>
<tr>
<td>GENEZILLA</td>
<td>19.5</td>
<td>8.8</td>
</tr>
<tr>
<td>ACEVIEW</td>
<td>63.5</td>
<td>48.6</td>
</tr>
<tr>
<td>AUGUSTUS-EST</td>
<td>47.6</td>
<td>37.0</td>
</tr>
<tr>
<td>ENSEMBL</td>
<td>71.6</td>
<td>67.3</td>
</tr>
<tr>
<td>EXOGEAN</td>
<td>63.1</td>
<td>80.8</td>
</tr>
<tr>
<td>EXONHUNTER</td>
<td>21.9</td>
<td>6.3</td>
</tr>
<tr>
<td>PAIRAGON+NSCAN_EST</td>
<td>69.5</td>
<td>61.7</td>
</tr>
<tr>
<td>AUGUSTUS-dual</td>
<td>26.0</td>
<td>18.6</td>
</tr>
<tr>
<td>DOGFISH</td>
<td>10.8</td>
<td>14.6</td>
</tr>
<tr>
<td>MARS</td>
<td>33.4</td>
<td>24.9</td>
</tr>
<tr>
<td>NSCAN</td>
<td>35.4</td>
<td>36.7</td>
</tr>
<tr>
<td>SAGA</td>
<td>4.3</td>
<td>3.4</td>
</tr>
<tr>
<td>GENSCAN</td>
<td>15.5</td>
<td>10.1</td>
</tr>
<tr>
<td>KNOWNgene</td>
<td>77.0</td>
<td>72.7</td>
</tr>
<tr>
<td>TWINSSCAN</td>
<td>22.3</td>
<td>20.2</td>
</tr>
</tbody>
</table>
So, where are we now?
Ensembl genes
www.ensembl.org

- 21,774 “Known genes”
- 1,036 Novel genes
- 3,994 RNA genes
- 69,185 Genscan gene predictions

Let’s not forget pseudogenes
- 27,130 pseudogenes
- www.pseudogene.org
Current NCBI gene counts
www.ncbi.nih.gov

- **Entrez Gene:**
  - 38,621 genes
  - But this includes pseudogenes

- **RefSeq:**
  - 28,961 genes
  - 31,784 transcripts
**Lesson:** Science isn’t decided by voting

<table>
<thead>
<tr>
<th>CCDS Totals</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCDS IDs</td>
<td>14,795</td>
</tr>
<tr>
<td>Gene IDs</td>
<td>13,142</td>
</tr>
<tr>
<td>Sequence IDs</td>
<td>31,724</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence IDs by Organization</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NCBI RefSeq</td>
<td>15,496</td>
</tr>
<tr>
<td>EBI, WTSI Records</td>
<td>16,228</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene ID</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes with &gt;1 CCDS ID</td>
<td>1,205</td>
</tr>
</tbody>
</table>

Table as of March 2, 2005

Feb. 26, 2007
So, we still don’t have a gene count

…and for many genes, we aren’t yet sure of their exon-intron structure…

…and there are > 1000 other genomes already complete or under way…

…and so we aren’t giving up!
Thanks…

Mihaela Pertea

Jonathan Allen (LLNL)

Art Delcher

Bill Majoros (Duke Univ.)

Brian Haas (Broad Institute)
Evaluating Gene Predictions

Fig. 3 from Guigo et al., Genome Biology 2006, 7(Suppl 1):S2.