The Fruits of the Genome
Sequences for Society
David Botstein

Lewis-Sigler Institute for Integrative Genomics
Princeton University

---

**Genome Sizes and Gene Numbers**

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Size</th>
<th>Genes (for Proteins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td>12 megabases</td>
<td>5,800</td>
</tr>
<tr>
<td>Worm</td>
<td>100 megabases</td>
<td>19,400</td>
</tr>
<tr>
<td>Fly</td>
<td>120 megabases</td>
<td>13,400</td>
</tr>
<tr>
<td>Plant</td>
<td>115 megabases</td>
<td>25,500</td>
</tr>
<tr>
<td>Human/Mouse</td>
<td>3300 megabases</td>
<td>22,000</td>
</tr>
</tbody>
</table>

The basic cellular functions of all eukaryotes are carried out by proteins (and RNAs) whose **structure and function** are conserved.
Associating Biological Information with DNA Sequence

Molecular Biology: sequencing & analysis

Gene

Genetics: study of mutations and variants

Protein

Biochemistry

Function


The Amino Acid Sequence of a Protein

MDSEVAALVIDNGSMCKAGFAGDDAPRAVPPSIV
GRPRHQQIMVMQKDSYVGDEAQSKRGILTLYP
IEHGIVTNWDDMEKIWHHTFYNELRVAEEHPVLL
TEAPMNPKNREKMTQIMFETFNVPFVYSIQAVL
SLYSSGRTTGVLDSGDGVTHVPIYAGFSLPHAI
LRIDLAGRDLTDYLMKILSERGYSFSTTAEREIVR
DIKKEKLCYVALDFEQEMQTAQQSSSSIESYELPDG
QVITIGNERFAPEALFHSVGLGESAIGDQTTRYN
SINKCDVDVRKELYGNIVMSGGTMMFGIAERMQK
EITALAPSSMVKIIAPPERRYSVWIGSILASLT
TFQQMWSIKQEQYDESGPSIVVHKCF*
Sequence Similarity Between Yeast and Human Actin

Yeast: 27
KFGILRTIEGVTMNDQTHMHTTFINELKAVPEEHPVYLLTEAPFMNFNSREKMT
74
KFGILRT+VFIHEGVTMNDQTHMHTTFINELKAVPEEHPVYLLTEAPNFSNSREKMT
96
Human: 1
KFGILRTIEGVTMNDQTHMHTTFINELKAVPEEHPVYLLTEAPFMNFNSREKMT
27
KFGILRT+VFIHEGVTMNDQTHMHTTFINELKAVPEEHPVYLLTEAPNFSNSREKMT
50

Yeast: 87
QIMFSTTVPHAFYVSQ1AVSLV3SGLQTVT74G1+LJFHLALLLD
204
QIMFSTTVPHAFYVSQ1AVSLV3SGLQTVT74G1+LJFHLALLLD
146
Human: 1
QIMFSTTVPHAFYVSQ1AVSLV3SGLQTVT74G1+LJFHLALLLD
204
QIMFSTTVPHAFYVSQ1AVSLV3SGLQTVT74G1+LJFHLALLLD
120

Yeast: 257
ELPD0QVITIGEQL6RASLFYSLVEAGIDPTTSHS1HMCDVQREKLYGIVMS
324
ELPD0QVITIGEQL6RASLFYSLVEAGIDPTTSHS1HMCDVQREKLYGIVMS
264
Human: 1
ELPD0QVITIGEQL6RASLFYSLVEAGIDPTTSHS1HMCDVQREKLYGIVMS
306
ELPD0QVITIGEQL6RASLFYSLVEAGIDPTTSHS1HMCDVQREKLYGIVMS
246

Yeast: 327
GOTMFPQGIAENRQKETIALAPSMKVEXITAPERYSVWIGGILALSMTFGQMNISQ
384
GOTMFPQGIAENRQKETIALAPSMKVEXITAPERYSVWIGGILALSMTFGQMNISQ
360
Human: 1
GOTMFPQGIAENRQKETIALAPSMKVEXITAPERYSVWIGGILALSMTFGQMNISQ
360

<table>
<thead>
<tr>
<th>Yeast/Mammalian Protein</th>
<th>Sequence Identity (%)</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ubiquitin</td>
<td>96</td>
<td>yes</td>
</tr>
<tr>
<td>Actin</td>
<td>89</td>
<td>yes</td>
</tr>
<tr>
<td>ADP-Ribosylation Factor</td>
<td>77</td>
<td>yes</td>
</tr>
<tr>
<td>Beta-tubulin</td>
<td>75</td>
<td>partial</td>
</tr>
<tr>
<td>Alpha-tubulin</td>
<td>74</td>
<td>partial</td>
</tr>
<tr>
<td>Heat Shock HSP70</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>YPT1/Rab1</td>
<td>71</td>
<td>yes</td>
</tr>
<tr>
<td>HMG-CoA Reductase</td>
<td>67</td>
<td>yes</td>
</tr>
<tr>
<td>Transcription Init. Factor IID</td>
<td>65</td>
<td>yes</td>
</tr>
<tr>
<td>Cytochrome C</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>KAR2/BiP</td>
<td>62</td>
<td>yes</td>
</tr>
<tr>
<td>Calmodulin</td>
<td>60</td>
<td>yes</td>
</tr>
<tr>
<td>RAS1/N-ras, RAS2/K-ras</td>
<td>60</td>
<td>yes</td>
</tr>
<tr>
<td>CDC28/CDC2</td>
<td>59</td>
<td>yes</td>
</tr>
<tr>
<td>SEC18/NSF</td>
<td>46</td>
<td>yes</td>
</tr>
<tr>
<td>Co-metallothionein</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Dihydrofolate Reductase</td>
<td>32</td>
<td>yes</td>
</tr>
<tr>
<td>Profilin</td>
<td>28</td>
<td>yes</td>
</tr>
<tr>
<td>P-glycoprotein/MDR</td>
<td>26</td>
<td>yes</td>
</tr>
<tr>
<td>Glucose Transporter</td>
<td>25</td>
<td>yes</td>
</tr>
</tbody>
</table>

[Protein and Fink, 1985 (updated)]
The Intellectual Impact of the Genomic View

- The "grand unification" of biology: the functional parts of all living things are related by lineage.

"Once we understand the biology of E. coli, we will understand the biology of the elephant"
---Jacques Monod, ca.1960

- The challenge for the future is to understand not just mechanisms at the individual process level, but also the interactions among all the processes and their mechanisms.

- Genomics makes possible experiments and analysis at the "systems" level. This requires highly parallel experimental methods and computation-intensive analysis.

Fruits of the Genome

- Quantitative understanding of evolution from sequence.

- Comparative Genomics: "grand unification" of biology.

- The many uses of DNA sequence polymorphism: from forensics to disease gene identification.

- Functional Genomics: defining diseases through gene identities and genome-scale patterns of gene expression.

- DNA Diagnostics: detecting disease, disease progression and predisposition to disease.
Darwin's Great Intuitive Insight

Charles Darwin (1837)

"Universal" Unrooted Phylogenetic Tree of Life
Out of Africa: The evolutionary path of the human species

Rooted Phylogenetic Tree of Life
Multiple Sequence Alignment of *mutS* Homologs


Distinguishing Orthologs and Paralogs from a Gene Family by Parsimonious Assignment of Gene Duplications and Losses

MutS Homologs Evolve Diverged Functions

[Image: Diagram showing relationships between MutS homologs.]


Extracting Functional Information from the Human Genome Sequence

- Finding and Characterizing Human Disease Genes
  - DNA polymorphisms (SNPs & haplotypes)
  - Simple mendelian (ca. 3000) & complex (very few)
  - Complex disorders (a handful, maybe)
- Comparative Genomics: associating human genes with their functional equivalents in experimental model systems
  - Using the evolutionary information:
    - orthologs and paralogs
    - Genetic alterations, RNAi and other gene-based interventions
Extracting Functional Information from the Human Genome Sequence

- Patterns of Gene Expression
  DNA microarrays & Quantitative PCR
  Immediately useful for diagnosis
  (e.g. cancer subtypes)

- Systems Biology: understanding at a different level?
  Signal transduction, pathways, interactions

DNA Hybridization: Complementary Sequences Find Each Other to form Double Helices
Mapping Human Genes using DNA Polymorphisms

[Botstein, White, Skalnick & Davis, 1980]

DNA Polymorphisms Can Map Human Disease Genes by Linkage

[Wyman and White, 1980]
Thousands of Inherited Disease Genes have been Found

[Glazier Nadeau & Aikman, 2006]

Today, OMIM lists 2,799 of a total of 4,466 Mendelian phenotypes (mostly inherited diseases) have been associated with specific genes.

Gene Identification through Linkage Mapping Provides Basic Mechanistic Information for Inherited Diseases

Huntington’s Disease --- class of amplification of trinucleotide repeat diseases (myotonic dystrophy, fragile X, spinocerebellar ataxia, etc.)

Amyotrophic Lateral Sclerosis --- understanding of the critical issues around reactive oxygen species in the brain.

Ataxia-telangiectasia and BRCA1 --- implication of cell cycle checkpoints and DNA repair in the etiology of cancer.

Retinoblastoma --- realization that cancer can be caused by loss of function as easily as by inappropriate gain of function.
**DNA Evidence is Ubiquitous in Crime Fiction**

Watching these shows, it becomes clear that most (if not quite all) plots involve DNA evidence.

**DNA Polymorphisms are Abundant in the Human Genome**

The original RFLP

[Wyman and White, 1980]

Markers from a commercial DNA Forensics laboratory

[Ryan Forensic website]
The FBI has Settled on a Standard Set of Multiallelic Markers

CODIS: Combined DNA Index System (FBI)

Non-Inherited Dinucleotide Repeat Polymorphisms Appear in Colon Tumor Cells

[Aaltonen et al., 1993]
Isolation of Yeast *msh2* and *mlh1* Mutations, with a Hypothesis, September 1993

Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair

Micheline Strand*, Tomas A. Prolla†, R. Michael Liskay‡ & Thomas D. Petes*  

Finally, we note that the phenotype of the mutation involved in one type of familial colorectal cancer (decreased stability of simple repeats) is that predicted for a mutation affecting DNA mismatch correction. Such a mutation could represent a functional homologue of *PMS1, MLH1* or *MSH2* or another component of the mismatch repair system (for example, a DNA helicase or single-strand binding protein).

Nature 365:274 (September 16, 1993)

---

The Human MSH2 Ortholog Predisposes to HNPCC (Human Non-Polyposis Colon Cancer)

The Human Mutator Gene Homolog MSH2 and Its Association with Hereditary Nonpolyposis Colon Cancer

Richard Fishel,* Mary Kay Lescoe,* M. R. S. Rao,† Neal G. Copeland,† Nancy A. Jenkins,‡ Judy Garber,§ Michael Kane,§ and Richard Kolodner†

Today, it is known that ca. 90% of all familial HNPCC families have mutations in either the human MSH2 or MLH1 homologs
Genome-Wide Gene Expression Patterns Determined Using Hybridization to DNA Microarrays

Extracting Data
Hierarchical Clustering

Bringing Together Similar Patterns of Gene Expression

[Elise et al., 1998]
Randomized Data

Rows Ordered by Hierarchical Clustering
Rows Ordered by Hierarchical Clustering with Nodes Flipped to Optimize Ordering

440 human cell and tissue samples (out of more than 20,000)

A new kind of map of the human genome...

Pat Brown
Mike Eisen
Max Diehn
Xin Chen
Jon Pollack
Chuck Perou
Therese Sorlie
Mitch Garber
Marcia Scherber
Matt van de Rijn
Gavin Sherlock
Mike Fero
Molecular Portraits of Cancer

Molecular Portraits of Breast Tumors: Norway/Stanford Cohort
Molecular Portraits of Breast Tumors: Dutch Cohort

(Data from van t’Veer et al, 2002)

Correlation of Subtype with Outcome in Different Cohorts
A genomic hypothesis test

Hypothesis: the four breast cancer subtypes represent fundamentally different diseases arising from different cell types and/or by different pathways of oncogenesis.

If so, then women who inherit genes predisposing to breast cancer, and who thereby have a many-fold increased risk, should all have the same tumor subtype.

Test: Assess the patterns of gene expression of breast tumors in BRCA1 or BRCA2 carriers.

BRCA1 mutations predispose to tumors of the “Basal” subtype
(Data from van ’t Veer et al, 2002)
Clinical Applications of Microarray Information

• Better diagnosis: definition of more biologically and clinically homogeneous cancer subtypes. Greater power to test efficacy in trials.

• Earlier detection: identification of secreted molecules that can be detected in blood tests

Clinical Applications of Microarray Information

• New therapeutic targets: identification of molecules expressed in tumors that can be aimed at
  • membrane proteins as antibody therapy targets e.g. Her2/ERBB2 (Herceptin)
  • receptor tyrosine kinases as small molecule targets e.g. specific antagonists of Abl or Kit (Gleevec)

• Monitoring and predicting response: finding the appropriate therapy, old or new, for each individual tumor
Examples of Human Cancer-Causing Genes

These genes have been implicated in cancer (*) targets of successful drugs.

Chronic Myelogenous Leukemia Patients Treated with Specific Antagonist (Gleevec) Directed Against the Product of the ABL Gene
Breast Cancer Patients Treated with an Antibody Drug (Herceptin) Directed Against the Product of the HER2 Gene

Results of a randomized trial in which women were treated after removal of the primary tumor: the effect is about 2-fold improvement in survival, and highly significant statistically.

Issues for the Future

- Personal genome as predictor of health: confronting the reality that we have no robust theory or understanding of the relationship between genotype and complex diseases (as opposed to single-gene Mendelian ones).
Issues for the Future

- How to reconcile interpretation of DNA sequence by doctors and patients (or somebody else-- a statistical geneticist?) with the probabilistic nature of the connections between sequence and disease:
  -- The case of Huntington’s (no therapeutic options today)
  -- The case of HNPCC (heightened surveillance, by colonoscopy, of obvious survival value)
  -- The case of HER2 amplification in breast tumors (an effective drug, trastuzumab (Herceptin) available)