What is the human genome?

- The **human genome (DNA)** is the complete sequence of 3 billion letters.
- Genome is same in all our cells.
- **Genes** encode instructions to make **proteins**.
- Proteins are the key functional and structural elements in life.
- DNA for genes copied into **RNA**, moves out and serves as blueprint for protein.
- The genome is 99.9% identical between us.
Basic Base-pairing

- Complementary nucleotides form base pairs
- Base pairs are put together in chains (strands)
  - Naturally form double helixes
  - Redundant information in each strand

**Diagram:**
- Phosphate Molecule
- Deoxyribose Sugar Molecule
- Nitrogenous Bases
- Weak Bonds Between Bases
- Sugar-Phosphate Backbone
- A phosphodiester bond
- DEOXYRIBONUCLEOTIDES are linked together to form a POLYNUCLEOTIDE, two of which are wound around one another in the DOUBLE HELIX
Single Nucleotide Polymorphisms

• A single nucleotide position in the genome sequence for which there are two or more alternative alleles
• If the genome is a book, then these are alternative spellings
  – color versus colour
  – behavior / color versus behaviour / colour
• SNP rate 1 per 1300 bases
• dbSNP has over 2.5 million “public” SNPs
• Typical gene has 2-4 common variants

• Consistent with small founding population under 10,000
• Left Africa 5,000 years ago with this distribution
• Any village represents 70% of variation in the world
• Common disease-common variant hypothesis: common diseases occur due to segregation of common alleles, not occurrence of rate mutations
SNP Maps

- Maps detailing where chromosomes differ among individuals

SNPs already impact clinical care

A common polymorphism associated with antibiotic-induced cardiac arrhythmia

Federico Sesti*, Geoffrey W. Abbott*, Jian Wei†, Katherine T. Murray†, Sanjeev Saksena‡, Peter J. Schwartz§, Silvia G. Priori§, Dan M. Roden¶, Alfred L. George, Jr.¶, and Steve A. N. Goldstein**

*Departments of Pediatrics and Cellular and Molecular Physiology, Boyer Center for Molecular Medicine, Yale University School of Medicine, New Haven, CT 06520; †Departments of Medicine and Pharmacology, Vanderbilt University, Nashville, TN 37235; ‡Robert Wood Johnson Medical School, Piscataway, NJ 07055; and §Department of Cardiology, University of Pavia and Policlinico San Matteo IRCCS, Pavia, Italy 27100

Edited by Vincent T. Marchesi, Yale University School of Medicine, New Haven, CT, and approved July 6, 2000 (received for review May 16, 2000)

- Genes will help us determine which drugs to use in particular disease subtypes
- Genes will help us predict those who get side-effects
RNA expression detection chips

Tissue or Tissue under influence

RNA

Tagged with fluor

cDNA spotted on glass slide or oligonucleotides built on slide


Tell Me More
NCI 60 data set

NCI 60 is set of 60 cancer cell lines against which the National Cancer Institute has tested over 50,000 chemicals to find anti-cancer agents.

NCI recently provided us with consistent, validated data for 5,084 agents, representing the concentration of each agent needed to cause 50% growth inhibition compared to control (GI50) for each cell line in NCI 60.

Data was provided as $-\log_{10}(\text{GI50})$.

- Lower number means higher GI50, or less sensitivity to an agent.

Unfortunately, few of these anti-cancer agents have documented mechanisms of action (or even a name listed).

http://www.dtp.nci.gov
RNA expression data set

- RNA expression in NCI 68 cell lines was determined using Affymetrix HU6000 arrays
  - 5,223 known genes
  - 1,193 expressed sequence tags

- The RNA expression data set and Anti-cancer susceptibility data set were merged, using the 60 cell lines the two tables had in common

<table>
<thead>
<tr>
<th>6,000 genes</th>
<th>5,000 anti-cancer agents</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA Expression</td>
<td>Common 60 Cell Lines</td>
</tr>
<tr>
<td></td>
<td>Drug Susceptibility</td>
</tr>
</tbody>
</table>
• Threshold $r^2$ was 0.8
• 202 networks
• 834 features out of 11,692 (7.1%)
• 1,222 links out of 68,345,586 (.0018%)
• Only one link between a gene and anti-cancer agent

Genes and Anti-Cancer Agents

- Elevated levels of J02923 (lymphocyte cytosolic protein-1, LCP1, L-plastin, pp65) is associated with increased sensitivity to agent 624044
- Agent 624044 is 4-Thiazolidinecarboxylic acid, 3-[[6-[2-oxo-2-(phenylthio)ethyl]-3-cyclohexen-1-yl]acetyl]-2 thioxo-, methyl ester, [1R-[1a(R*),6a]]- (9CI)
- LCP1 is an actin-binding protein involved in leukocyte adhesion
- LCP1 is thought to play a role in tumorigenicity
- LCP1 is expressed in most human cancer cell lines
- Other thiazolidine carboxylic acid derivatives are known to inhibit tumor cell growth
Lab Test Relevance Networks

- Matrix of 642 lab tests with 28,566 overlapping results
- Threshold $r^2$ was 0.6, $n$ was 50

48 lab tests out of 642 (7%), 36 links out of 205,761 (.017%)