Harvard-MIT Division of Health Sciences and Technology
HST.950J: Engineering Biomedical Information: From Bioinformatics to Biosurveillance
Course Directors: Dr. Isaac Kohane, Dr. Marco Ramoni

Harvard Medical School

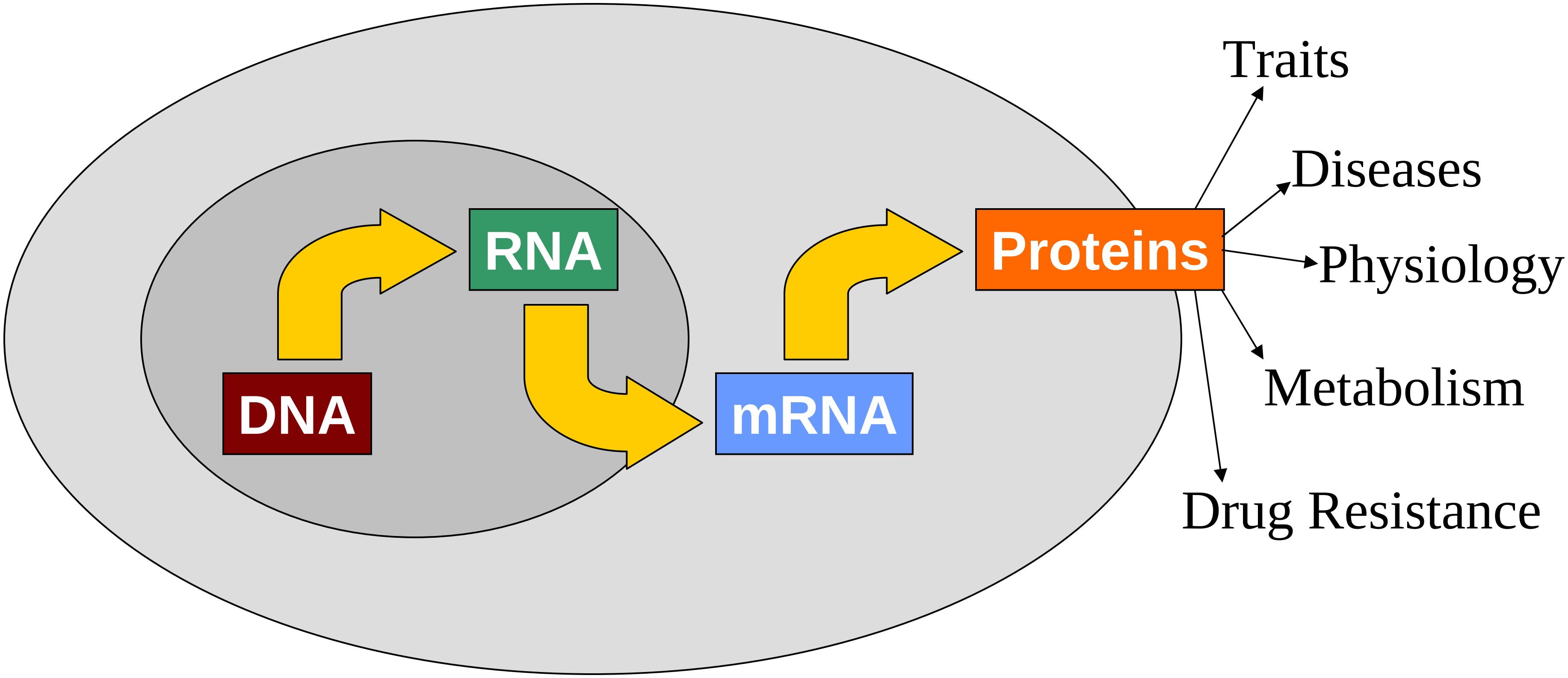
Massachusetts Institute of Technology

Bioinformatics

Marco F Ramoni, PhD
September 13th, 2005

Biomedical Computing
6.872 / HST 950

Central Dogma of Molecular Biology



The DNA

- The DNA looks like a ladder twisted into a helix.
- The sides of the "ladder" are formed by molecules of sugar and phosphate.
- The "rungs" consist of nucleotide bases joined by hydrogen bonds.
- The bases are the "letters" that spell out the genetic code: A, T, G, C.
- Base pairs are bases paired across the ladder: A pairs with T, and G pairs with C.

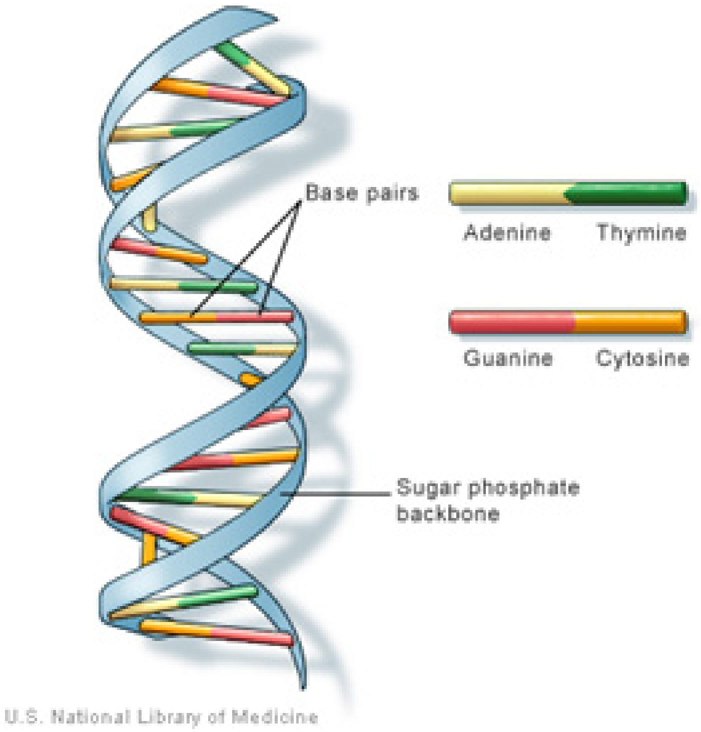


Image courtesy of the U.S. National Library of Medicine.
http://www.nlm.nih.gov/

The Gene

Gene: Physical and functional unit of heredity passed from parent to offspring.

Function: A DNA segment containing the information to produce specific proteins.

Intron: Gene part spliced out of the coding sequence.

Exon: Gene part actually coding for proteins.

Alternative splicing: Not all exons are always included in the final coding sequence.

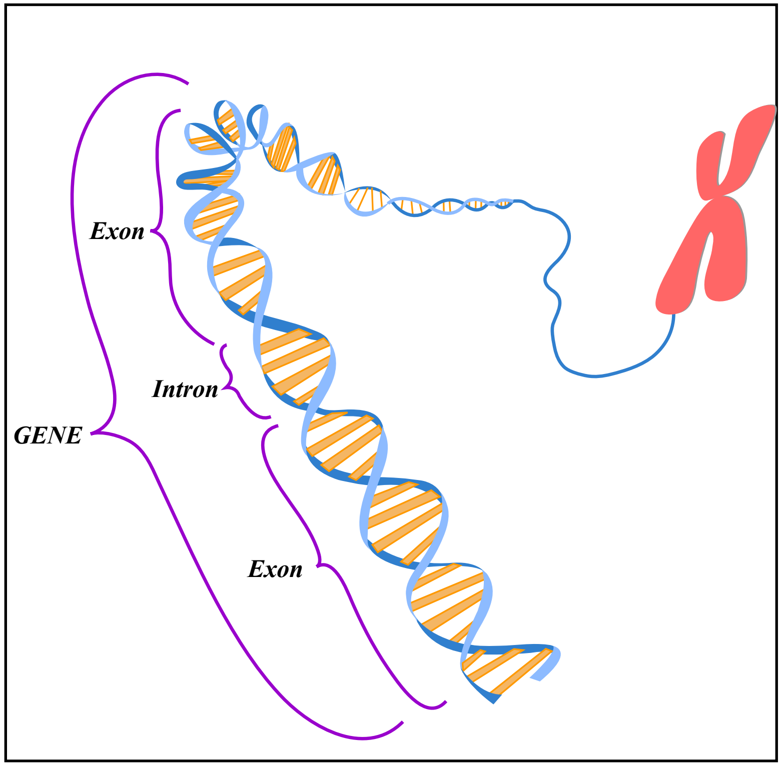


Figure by MIT OCW.

Proteins

Definition: A large complex molecule made up of one or more chains of amino acids.

Role: Proteins are the expression of genes.

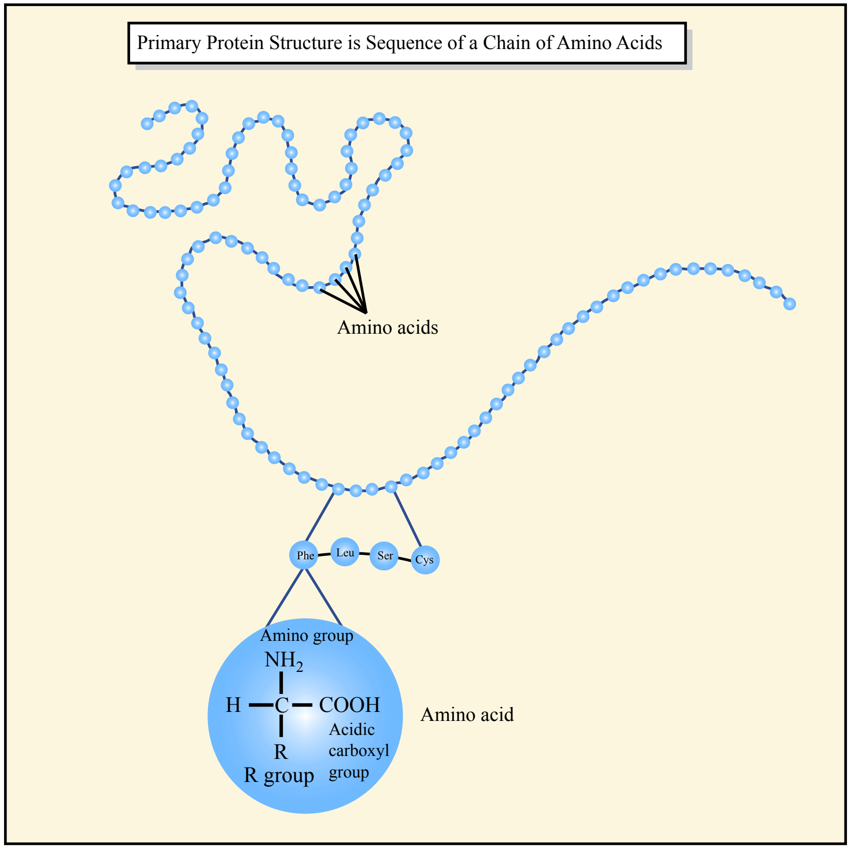


Figure by MIT OCW.

Why is it so difficult?

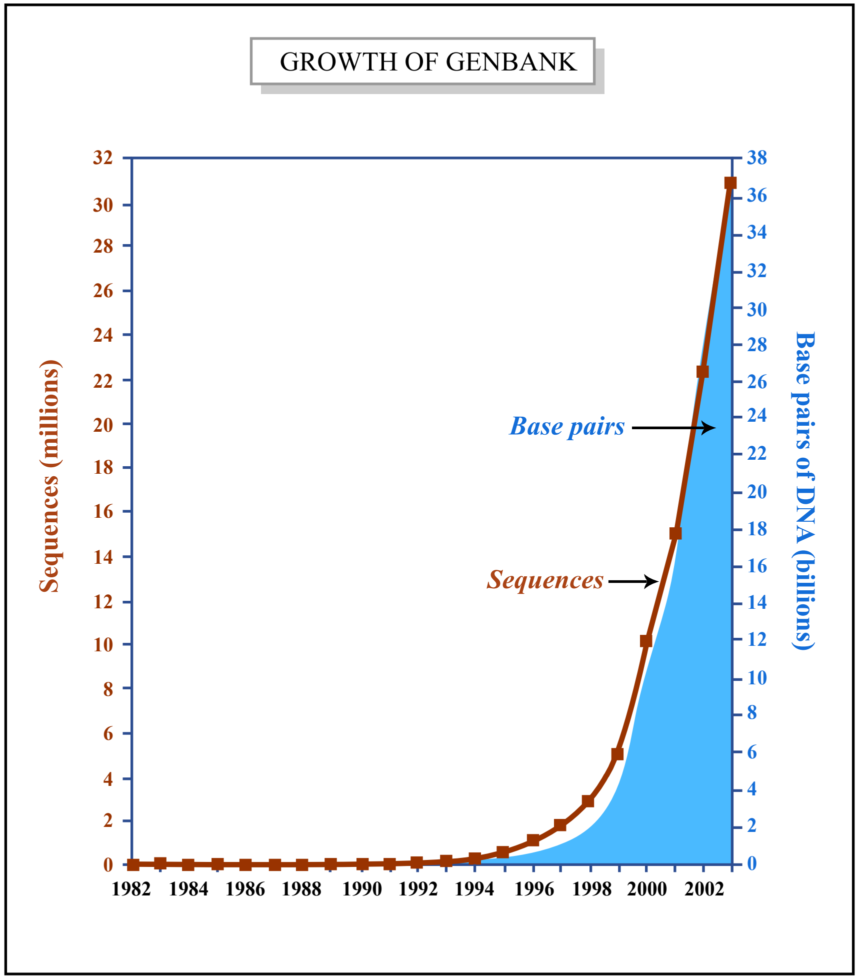


Figure by MIT OCW.

Graphs illustrating the Number of disease genes discovered from years 1981-2000 and the Number of Entries in Mendelian Inheritance in Man from years 1965-2000 have been removed due to copyright reasons.

Structural Genomics

Intuition: We can find genetic bases of observable characters (like diseases) without knowing how actual coding works.

Origins: Sturtevant (1913) finds traits-causing genes.

Outcast: Ernst Mayr called it “Bean bag genetics”.

Reasons: No markers to identify coding regions.

Markers: Botstein (1977) showed that naturally occurring DNA contains markers identifying genomic regions of polymorphisms.

The Future: Prominent item on the of the HGP first draft:

> A SNP map promises to revolutionize both mapping diseases and tracing human history.

SNPs: Single Nucleotide Polymorphisms, subtle variations of the human genome across individuals.

Single Nucleotide Polymorphisms

Key: Most frequent polymorphism.

Nature: Single base variation.

Frequency: >1% of the population.

Occurrence: 1 every 1000 bases.

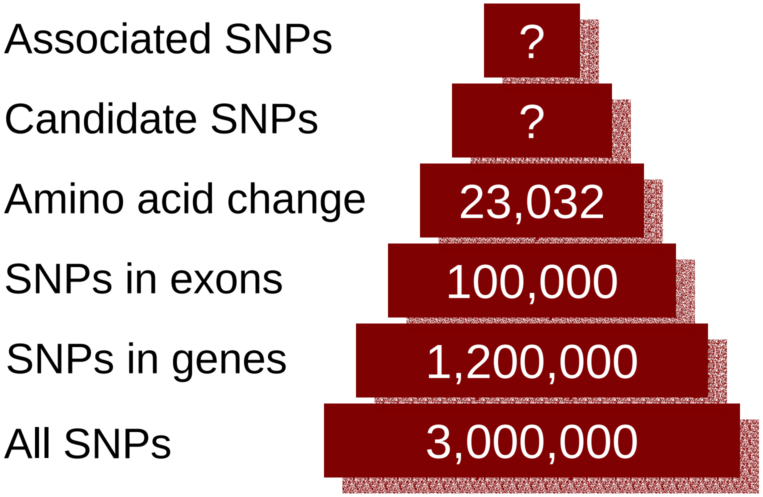
Total: 3,000,000.

Feature: Occur in coding regions.

Types: According to position.

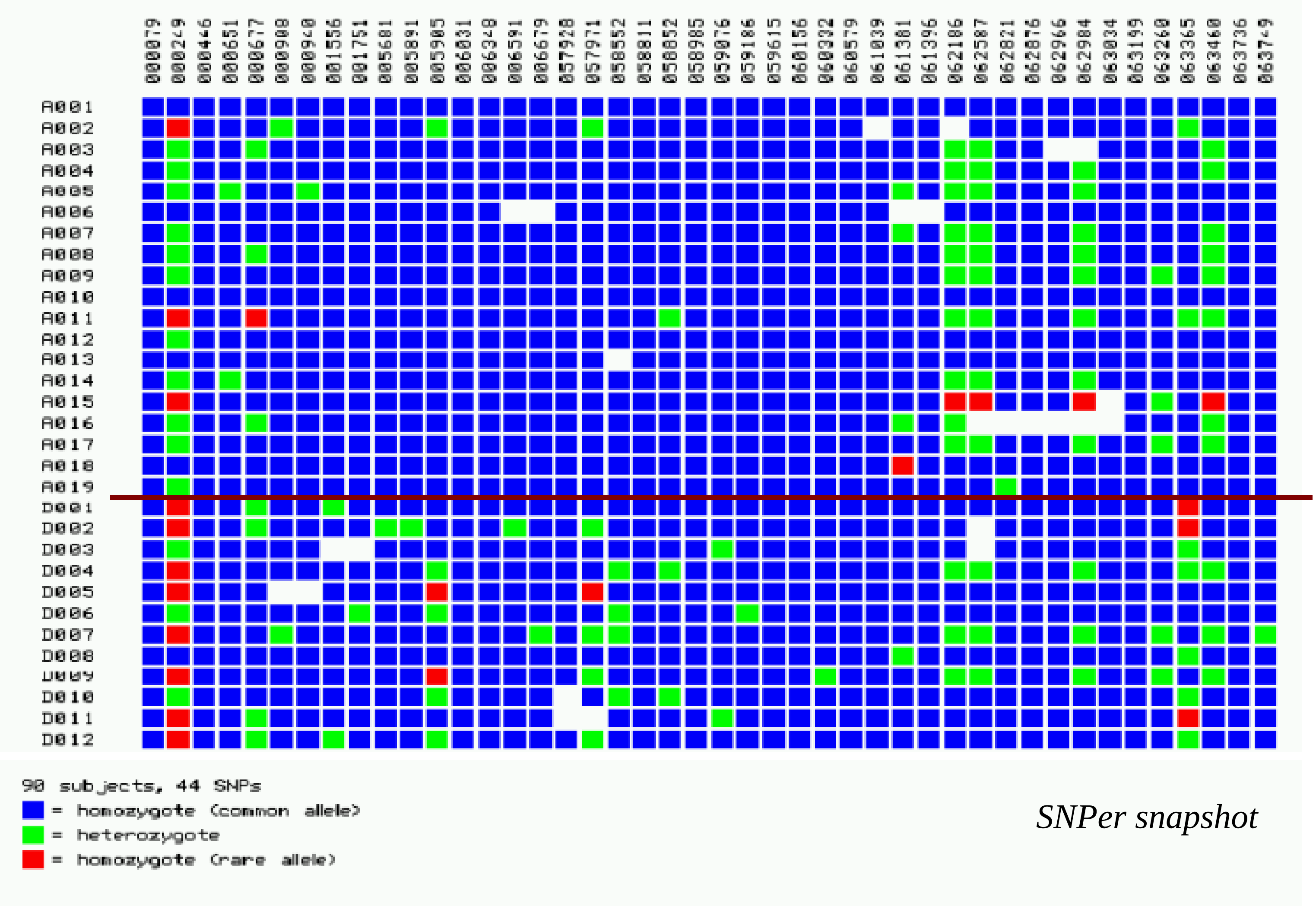
- cSNP: in coding region.
- rSNP: in a regulatory region.
- sSNP: no change in A acid.

... ATGCGATACTCGATCTCCCGA ...
... ATGCGATACGCGATCTCCCGA ...



Associated SNPs	?
Candidate SNPs	?
Amino acid change	23,032
SNPs in exons	100,000
SNPs in genes	1,200,000
All SNPs	3,000,000

Reading SNP Maps



90 subjects, 44 SNPs

🟦 = homozygote (common allele)

🟩 = heterozygote

🟥 = homozygote (rare allele)

SNPer snapshot

Courtesy of Dr. Alberto A. Riva. Used with permission.

Complex Traits

Problem: Traits don’t always follow single-gene models.

Incomplete Penetrance: Some individuals with genotype do not manifest trait. Breast cancer / BRCA1 locus.

Age	40	55	80
Carrier	37%	66%	85%
Non Carrier	0.4%	0.3%	8%

Genetic Heterogeneity: Mutation of more than one gene can cause the trait. Difficult in non experiment setting.

Retinitis pigmentosa: from any of 14 mutations.

Polygenic cause: Require more than one gene.

Hirshsprung disease: needs mutation 13c and 21c.

Genetic Markers

Task: Find basis (genotype) of diseases (phenotype).

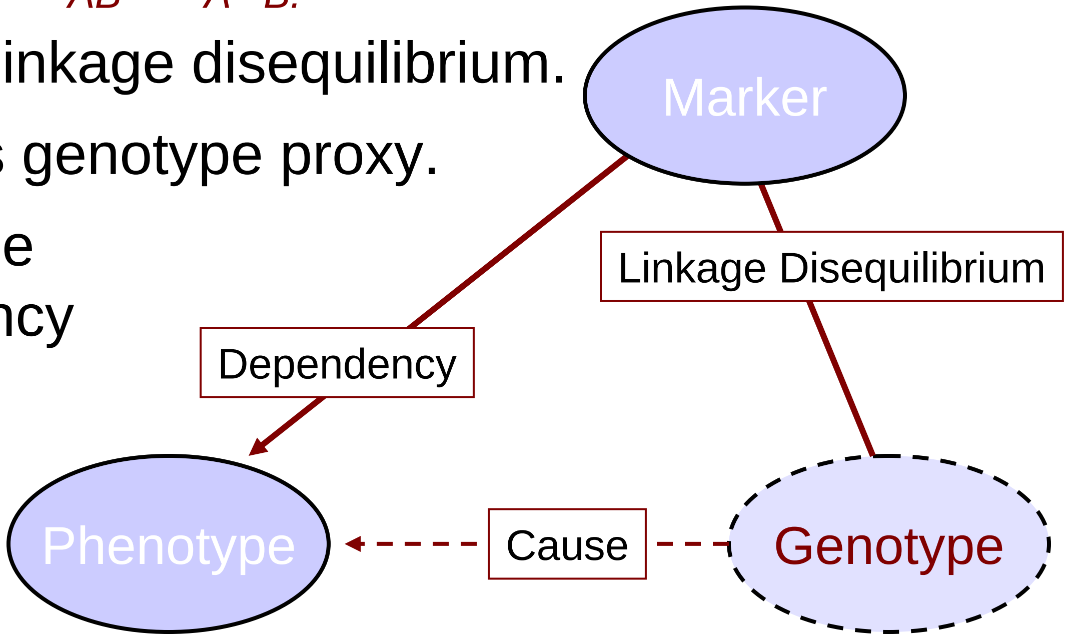
Linkage equilibrium: Loci Aa and Bb are in equilibrium if transmission probabilities π_A and π_B are independent.

$$\pi_{AB} = \pi_A \pi_{B.}$$

Marker: Flag regions in linkage disequilibrium.

Strategy: Use marker as genotype proxy.

Dependency: Observable measure of dependency between marker and phenotype.



Gene Expression

Genetic code: The instructions in a gene defining how to make a specific protein.

Amino acid: A 3-letter “word” combining A, T, G, C.

Protein production: Proteins are produced in two steps.

- Transcription: Gene code is transcribed into mRNA.
- Translation: mRNA is transformed and transported out of the nucleus to be translated into proteins.

Expression: mRNA amount produced from the information contained in a gene.

Functional Genomics

- Goal: Elucidate functions and interactions of genes.
- Method: Gene expression is used to identify function.
- Tools: Characteristic tools of functional genomics:
  - ✓ High throughput platforms.
  - ✓ Computational and statistical data analysis.
- Style: The intellectual style is different:
  - ✓ Research is no longer hypothesis driven.
  - ✓ Research is based on exploratory analysis.
- Issue: Functional genomics is in search of a sound and accepted methodological paradigm.

Measuring Expression

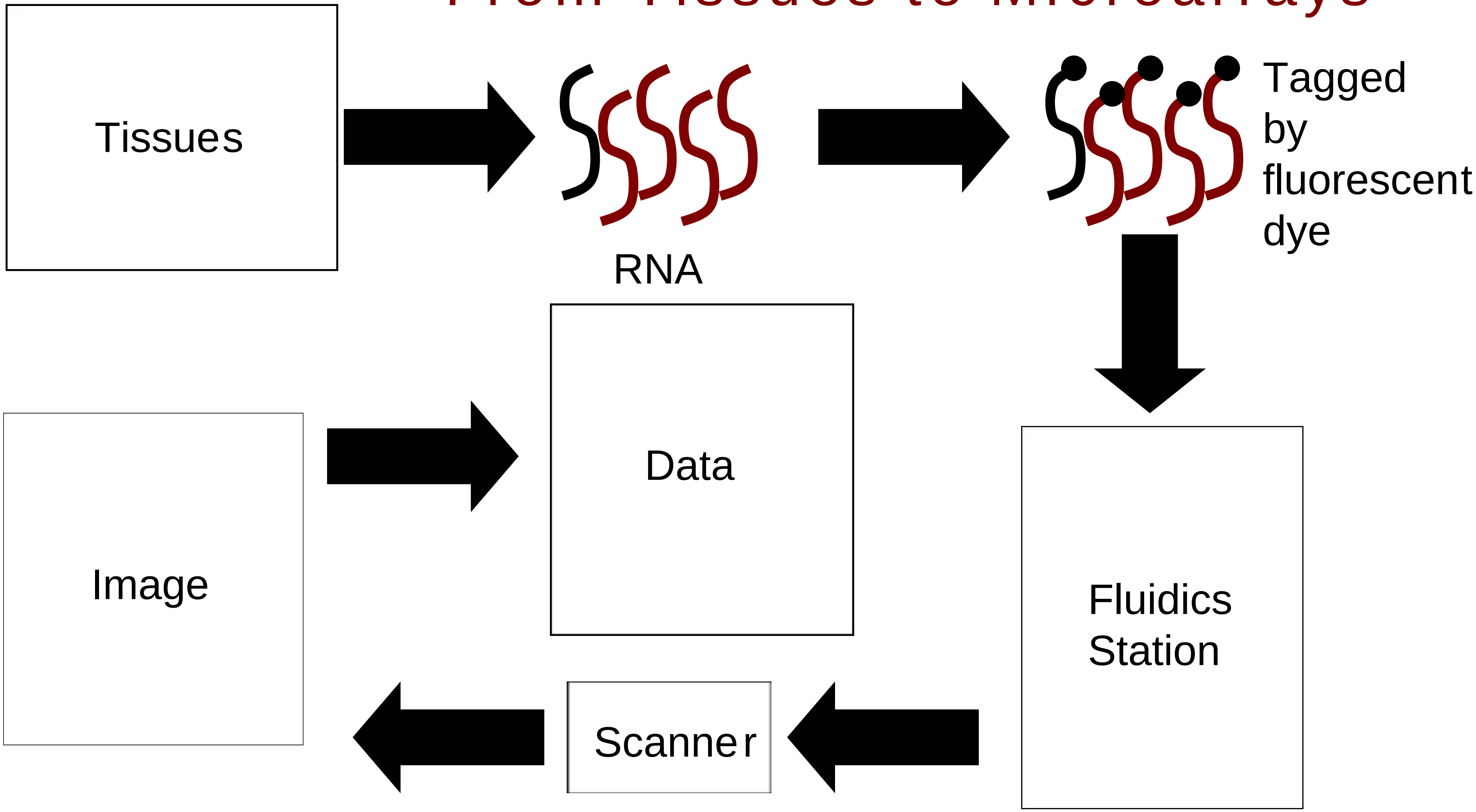
Rationale: Measurement of gene expression reverses the natural expression process.

Hybridization: Process of joining two complementary strands of DNA or one each of DNA and RNA to from a double-stranded molecule.

Artificial process: Backward the mRNA production.

- ✓ DNA samples (probes) are on the microarray.
- ✓ Put cellular labeled mRNA on the microarray.
- ✓ Wait for the sample to hybridize (bind).
- ✓ Scan the image and, for each point, quantify the amount of hybridized mRNA.

From Tissues to Microarrays



Tissues

RNA

Tagged by fluorescent dye

Fluidics Station

Scanner

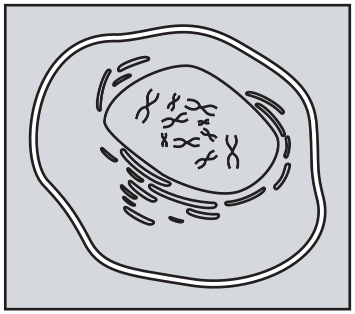
Image

Data

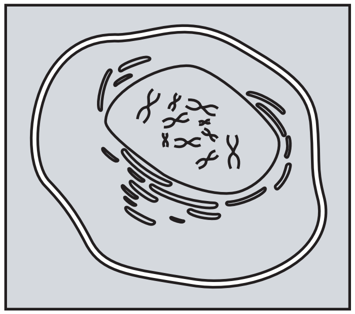
Comparative Experiments

Healthy cell

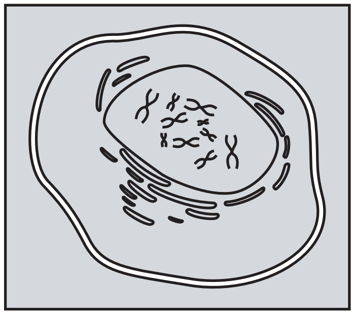
Tumor cell



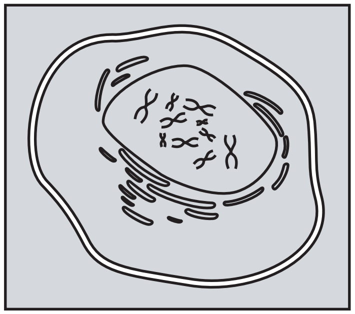
Sample 1



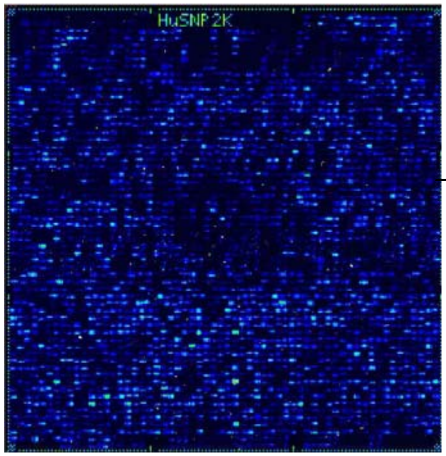
Sample 2



Sample 3

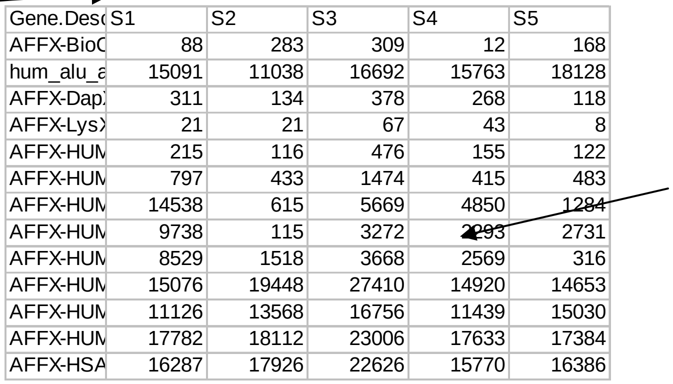


Sample 4



Samples k=1,…,n_i

genes
g=1,…,G



Gene.Des	S1	S2	S3	S4	S5
AFFX-BioC	88	283	309	12	168
hum_alu_a	15091	11038	16692	15763	18128
AFFX-Dap	311	134	378	268	118
AFFX-Lys	21	21	67	43	8
AFFX-HUM	215	116	476	155	122
AFFX-HUM	797	433	1474	415	483
AFFX-HUM	14538	615	5669	4850	1284
AFFX-HUM	9738	115	3272	2293	2731
AFFX-HUM	8529	1518	3668	2569	316
AFFX-HUM	15076	19448	27410	14920	14653
AFFX-HUM	11126	13568	16756	11439	15030
AFFX-HUM	17782	18112	23006	17633	17384
AFFX-HSA	16287	17926	22626	15770	16386

y_{gik}

Identify genes that are differentially expressed in two conditions i=A,B.

Transcriptomic Diagnosis

Example: Acute lymphoblastic leukemia (27) vs acute myeloid leukemia (11).

Method: Correlate gene profiles to an “extreme” dummy vector of 0s and 1s.

Results: 50 genes on each side.

Figures removed due to copyright reasons.
Please see:

Figure 3b of Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Calligiuri, C. D. Bloomfield, and E. S. Lander. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science* 286, no. 5439 (Oct 15, 1999): 531-7.

Transcriptomic Prognosis

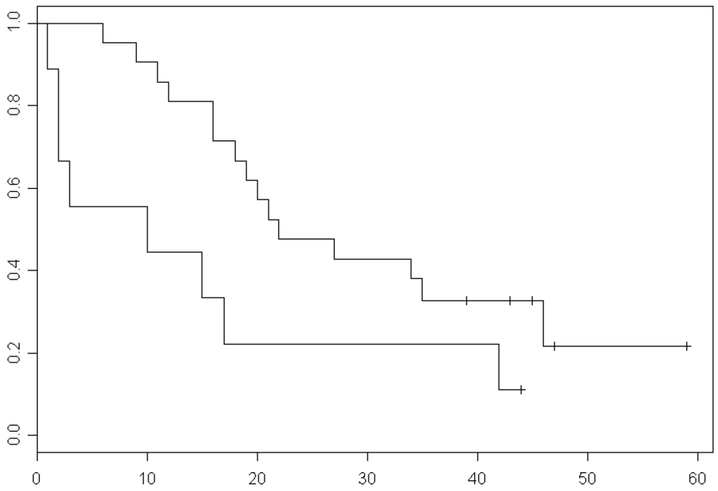
Problem: Ovarian cancer recurrence.

Phenotype: Survival curve.

Method: Maximize expression and survival difference.

Validation: Independent sample.

Memorial Survival



$p < 0.032$

Figure removed due to copyright reasons.
Please see:

Spentzos, D., et al. "Gene expression signature with independent prognostic significance in epithelial ovarian cancer." *J Clin Oncol* 22, no. 23 (Dec 1, 2004): 4700-10. *Epub* (Oct 25, 2004.)

Unsupervised Methods

- Differential experiments usually end up with:
  - ✓ A list of genes changed across the two conditions;
  - ✓ A “stochastic profile” of each condition.
- Useful to identify diagnostic profiles and prognostic models.
- They are not designed to tell us something about regulatory mechanisms, structures of cellular control.
- With supervised methods, we look only at relations between gene expression and experimental condition.
- Unsupervised methods answer different experimental questions.
- We use unsupervised methods when we are interested in finding the relationships between genes rather than the relationship between genes and a training signal (eg a disease).

One Dimensional Clustering

Strategy: Compute a table of pair-wise distances (eg, correlation, Euclidean distance, information measures) between genes.

Clustering: Use permutation tests to assess the cut point.

Relevance networks: Create a network of correlated genes and remove the links below the chosen threshold.

	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
Gene 1	1	0.2	0.8	-0.3	0.5	0.7	0.1
Gene 2		1	0.5	0.6	-0.2	-0.5	0.3
Gene 3			1	0.2	0.1	-0.2	0.1
Gene 4				1	0.9	0.4	0.3
Gene 5					1	0.1	-0.4
Gene 6						1	0.1
Gene 7							1

Figure removed due to copyright reasons.
Please see:

Figure 2 of Butte, A. J., P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane. "Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks." *Proc Nati Acad Sci U.S.A.* 97, no. 22 (Oct 24, 2000): 12182-6.

Hierarchical Clustering

Figures removed due to copyright reasons.
Please see:

Iyer, V. R., et al. "The transcriptional program in the response of human fibroblasts to serum." *Science* 283, no. 5398 (Jan 1, 1999): 83-7.

Two Dimensional Clustering

- We want to discover an unknown set of patient classes based on an unknown set of gene functional classes.
- A two-dimensional optimization problem trying to simultaneously optimize distribution of genes and samples.
- Survival time (KL curves) were used as independent validation of patient clusters.

Figures removed due to copyright reasons.
Please see:

Figures 1 and 5 of Alizadeh, A. A., et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." *Nature* 403, no. 6769 (Feb 3, 2000): 503-11.

Proteomics

Opportunity: Global view of protein actions/interactions.

Challenge: We do not have a *blue print*, like we have for SNPs and gene expression.

Proteome technology: Surface-Enhanced Laser Desorption Ionization – Time of Flight (SELDI-ToF) mass spec.

Data: For each sample, it produces peaks identified by a mass/charge index.

Goal: Find the protein at a peak.

Image removed due to copyright reasons.

With G Alterovitz

Biomics

Image removed due to copyright reasons.

Pharmacogenomics of Oral Cancer

Goal: Identify gene markers of cheotherapy susceptibility in cancer tissues.

Yield: 902 (21 microarray and 254 structural) studies for a bit more than 1000 genes.

Agreement: Good agreement on 120 genes.

Focus genes: Repeatedly reported in functional and structural studies (180 in all).

Structures: Graph theoretic score of “vulnerability”.

Image removed due to copyright reasons.

With Y Yu and S Sonis

Scale Free Networks

Q: Are these findings useful?

A: Yes, if we can learn something about the global structure of the network.

Scale free network: Natural interactions create robust substructures.

Method: Allow us to analyze global properties of a graph:

- ✓ Hubs/Authorities;
- ✓ Critical paths;
- ✓ Islands and holes.

Image removed due to copyright reasons.

Building a Hit-list

Data: Cancer cell lines (NCI60):

- ✓ Microarray (60);
- ✓ Response to ~3000 anti-cancer drugs.

Note: NO oral cancer cell lines.

Compound: For each compound, select the lines in the tail of the distribution of response.

Analysis: Compute the probability of a gene to be more expressed in resistant than sensitive lines (*Sebastiani et al*, 2004).

Results: A genomic signature for each compound.

Method: Rank compounds by their "damage" to the network.

Image removed due to copyright reasons.

A Random Compound

Image removed due to copyright reasons.

An Infective Compound

Image removed due to copyright reasons.

An Effective Compound

Image removed due to copyright reasons.

An Even Stronger Compound

Image removed due to copyright reasons.

Summary

- Structural Genomics
  - Sequencing
  - Genome-wide Genomics
- Functional Genomics
  - Genomic Medicine
  - Expression Microarrays
- Proteomics
- Interactome
- Annotation Databases
- Genomic Privacy
- Pharmacogenomics