Harvard-MIT Division of Health Sciences and Technology
HST.512: Genomic Medicine
Prof. Isaac Samuel Kohane

Children's Hospital
Informatics Program

Harvard
Medical School

Genomic Medicine

Lecture 1
Block 1
Isaac S. Kohane

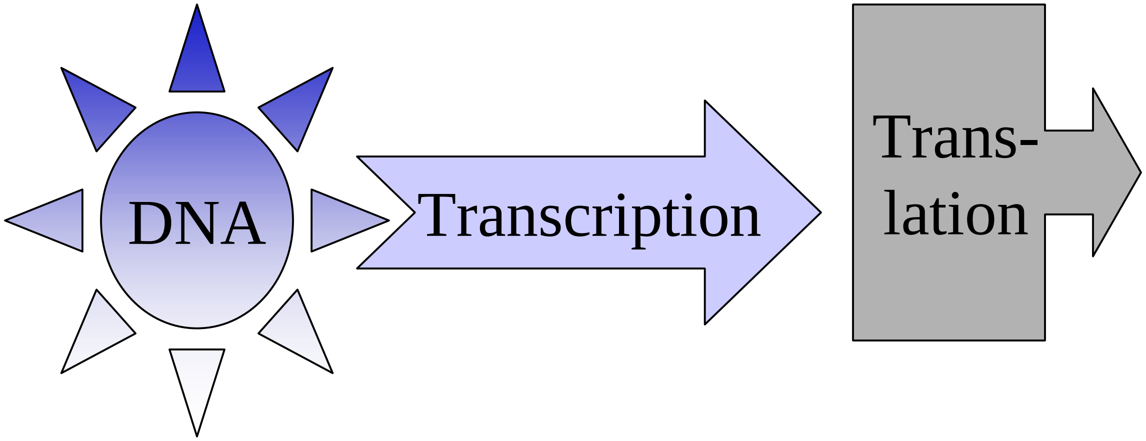
Overview

- The future is now
- Genomic vs genetic
- Heredity
- Resequencing of the diagnostic process
- Accelerating consumer activation

Overview

- **The future is now**
- Genomic vs genetic
- Heredity
- Resequencing of the diagnostic process
- Accelerating consumer activation

The Long Path from Genotype to Function



DNA

Transcription

Trans-
lation

Structural
Biology

Sequencing
and
genotyping

Transcriptional
Profiling

Proteomics

Magnitude of the Task

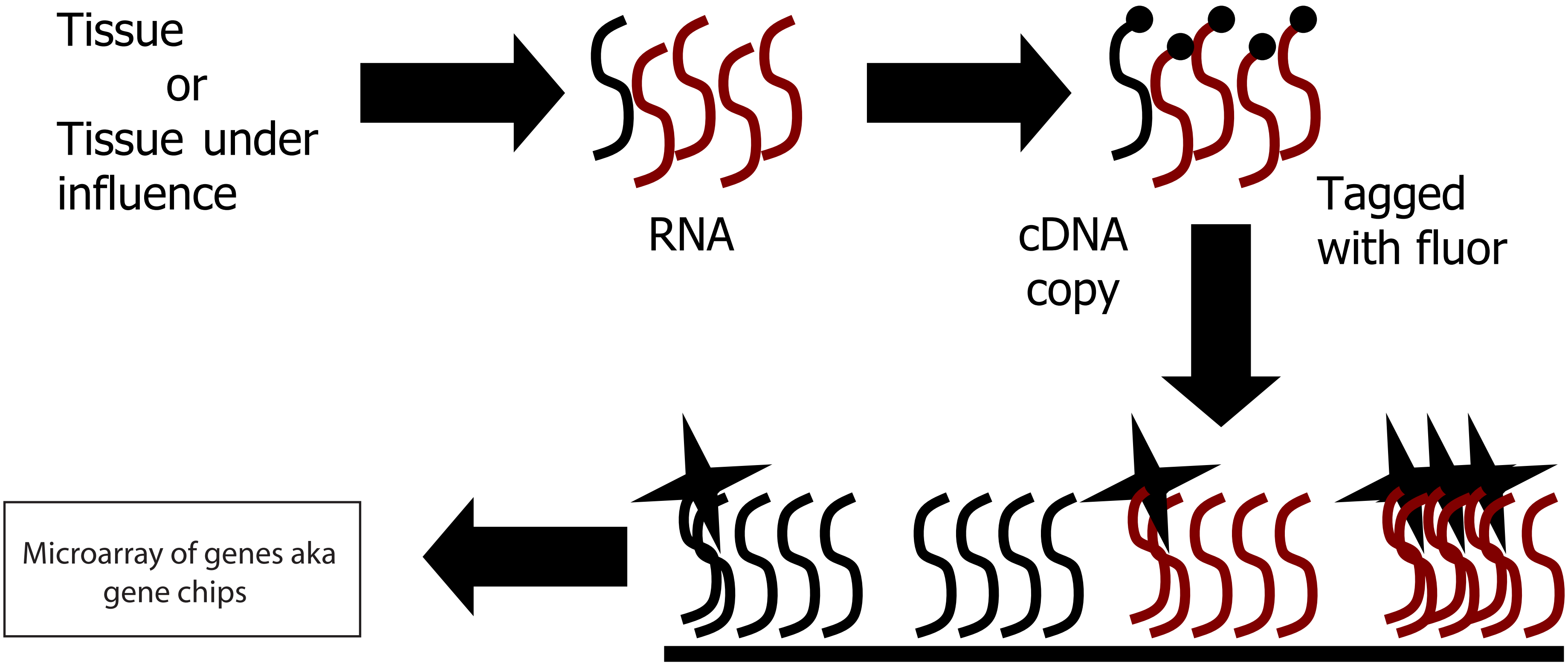
Magnitude of the Task

× 1000

Madonna Complex

- Some say $\text{Madonna}_{\text{Music}} > \text{Madonna}_{\text{Person}}$
- 4.7 GB vs
- $(3 \times 10^9) \times 2$ (bits/base) / 8 (bits/byte) = 0.75GB
- Is Madonna, her DNA sequence?
- No, and her current state is captured by..
  - ✓ Alternative splicing (x 3/gene)
  - ✓ Post-translational modification (x 100 -1000)
  - ✓ Location of gene product (x 10^{12})
- She’s a little more complicated than her music (are you surprised?)

RNA/DNA expression detection chips



Tissue
or
Tissue under influence

RNA

cDNA copy

Tagged with fluor

Microarray of genes aka gene chips

cDNA spotted on glass slide or oligonucleotides built on slide

Please refer to following papers:

Schena M, et al. Proc Natl Acad Sci USA; 93: 10614 (1996).

Entire issue. Nature Genetics, 21: supplement (Jan 1999).

The Promise: The New Diagnostics

Please see Nature. 2000 Feb 3;403(6769):503-11.

Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.

Alizadeh AA, et al.

Retreading the cancer chemotherapeutic protocol

- Cancer and Leukemia Group B (CALGB)
- CALGB has grown into a national network of 29 university medical centers, over 185 community hospitals and more than almost 3000 physicians who collaborate in clinical research studies aimed at reducing the morbidity and mortality from cancer
- Dozens of new protocols (breast cancer, prostate cancer, renal cancer) that use genome-wide
  - ✓ Which genes best predict survival?
  - ✓ Which adjuvant improves surgical outcome the best?
  - ✓ Can we find expression measure proxies for Stage, Grade and Cell Type

New Taxonomy of Human Disease

- Clinicians may have moved on from calling ‘fever’ a disease, but they still rely on phenotypic criteria to define most diseases,
- Yet these may obscure the underlying mechanisms and often mask significant heterogeneity.
- Thomas Lewis pointed out in 1944, diagnosis of most human disease provides only “insecure and temporary conceptions”
- Of the main common diseases, only the infectious diseases have a truly mechanism-based nomenclature.

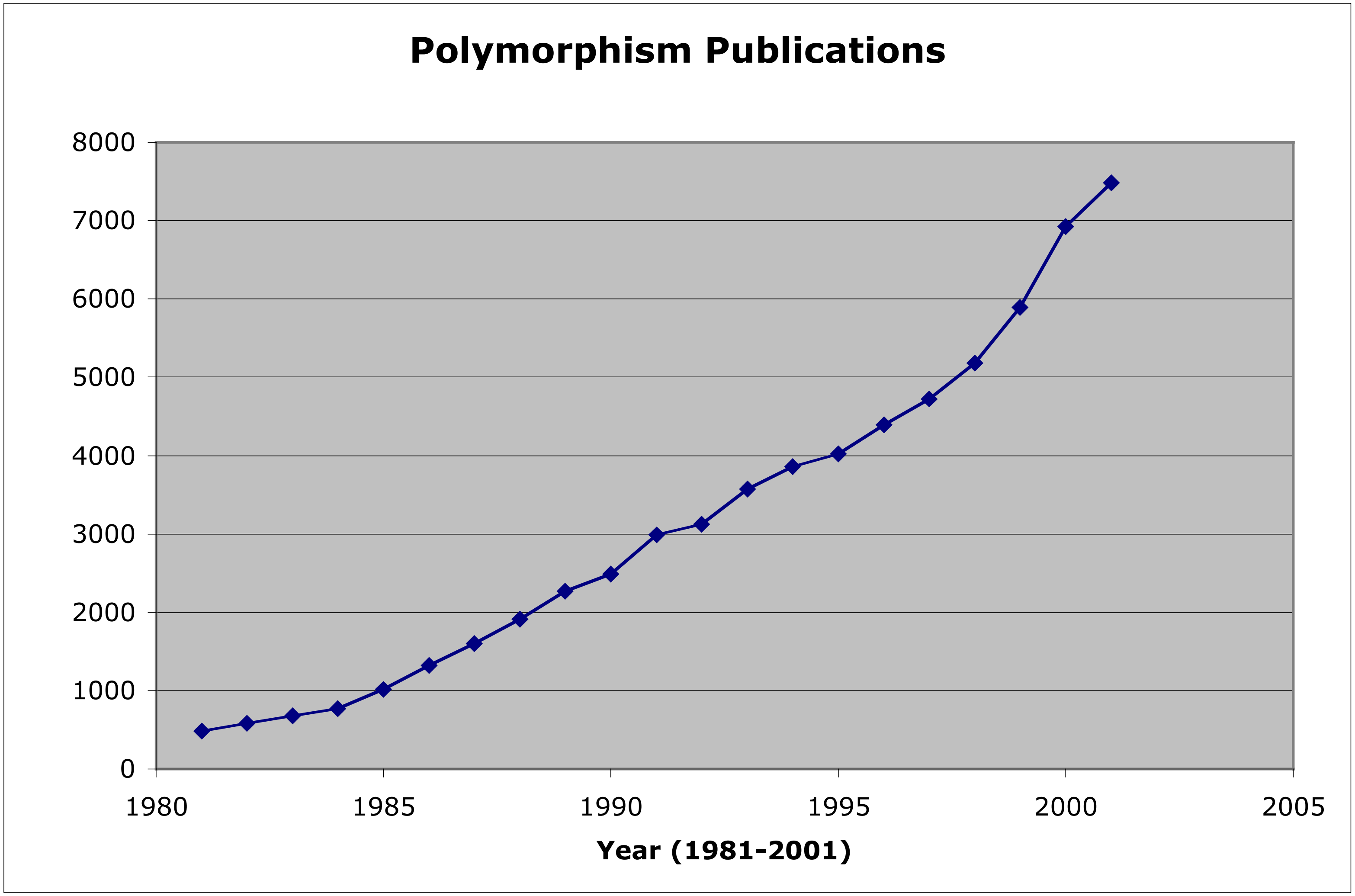
Changes in Use of “Every day” Medications

Please see Proc Natl Acad Sci U S A. 2000 Sep 12;97(19):10613-8.

A common polymorphism associated with antibiotic-induced cardiac arrhythmia.

Sesti F, et al.

Growth in Our Knowledge of Variants



Growth in monogenic disease

- **Pace of disease gene discovery (1981 to 2000).** The number of disease genes discovered so far is 1112. This number does not include at least 94 disease-related genes identified as translocation gene-fusion partners in neoplastic disorders. Numbers in parentheses indicate disease-related genes that are polymorphisms ("susceptibility genes").

  (From McKusick)

One-pass genotyping

Please see Figure 1 of Science. 2001 Nov 23;294(5547):1719-23.

Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21

Patil N, et al.

- Chromosome 21: 21,676,868 bases (67%) of unique sequence were assayed for variation with high-density oligonucleotide arrays
- Synthesized 3.4 × 10^9 oligonucleotides on 160 wafers to scan 20 independent copies of human chromosome 21 for DNA sequence variation.

Bioinformatics: 1998 Ambitions

- Find the functions of all 30,000$^+$genes using
  - ✓ DNA sequence
  - ✓ Genetics maps
  - ✓ Physical maps
  - ✓ Polymorphisms
  - ✓ Structure information
  - ✓ Existing biomedical literature
  - ✓ Gene transcription patterns
  - ✓ Protein translation/activity
- With growing databases containing data, this becomes a problem in the realm of bioinformatics

Please see Figure 1 of Nat Genet. 1998 Sep 20(1):19-23.

Data management and analysis for gene expression arrays.

Ermolaeva O, et al.

Madonna Complex

- Some say $\text{Madonna}_{\text{Music}} > \text{Madonna}_{\text{Person}}$
- 4.7 GB vs
- $(3 \times 10^9) \times 2$ (bits/base) / 8 (bits/byte) = 0.75GB
- Is Madonna, her DNA sequence?
- No, and her current state is captured by..
  - Alternative splicing (x 3/gene)
  - Post-translational modification (x 100 -1000)
  - Location of gene product (x 10^{12})
- She’s a little more complicated than her music (are you surprised?)

Overview

- The future is now
- **Genomic vs genetic**
- Heredity
- Resequencing of the diagnostic process
- Accelerating consumer activation

An engineer, a physicist, a mathematician, a computer scientist, and a statistician are on a train heading north, and had just crossed the border into Scotland. They look out the window and see a black sheep for the first time.

The engineer exclaims, “Look! Scottish sheep are black!”

The physicist yells, “No, no. Some Scottish sheep are black.”

The mathematician looks irritated and says, “There is at least one field, containing at least one sheep, of which at least one side is black.”

The computer scientist says, “Oh, no, a special case!”

Finally, the statistician says, “It is not statistically significant!”

Genomic vs genetic

Genetic Medicine	Genomic Medicine
Low frequency of ~1000 of usually high penetrance genes	The genetic risk for common diseases will often be due to disease-producing alleles with relatively high frequencies (>1%). All genes may be disease causing.
1000’s of relatively uncommon diseases (1/300 for most common)	Common disorders due to the interactions of multiple genes and environmental factors
Mostly assessed indirectly & focused On single genes	Direct experimental access to the entire genome

Genetic vs Genomic

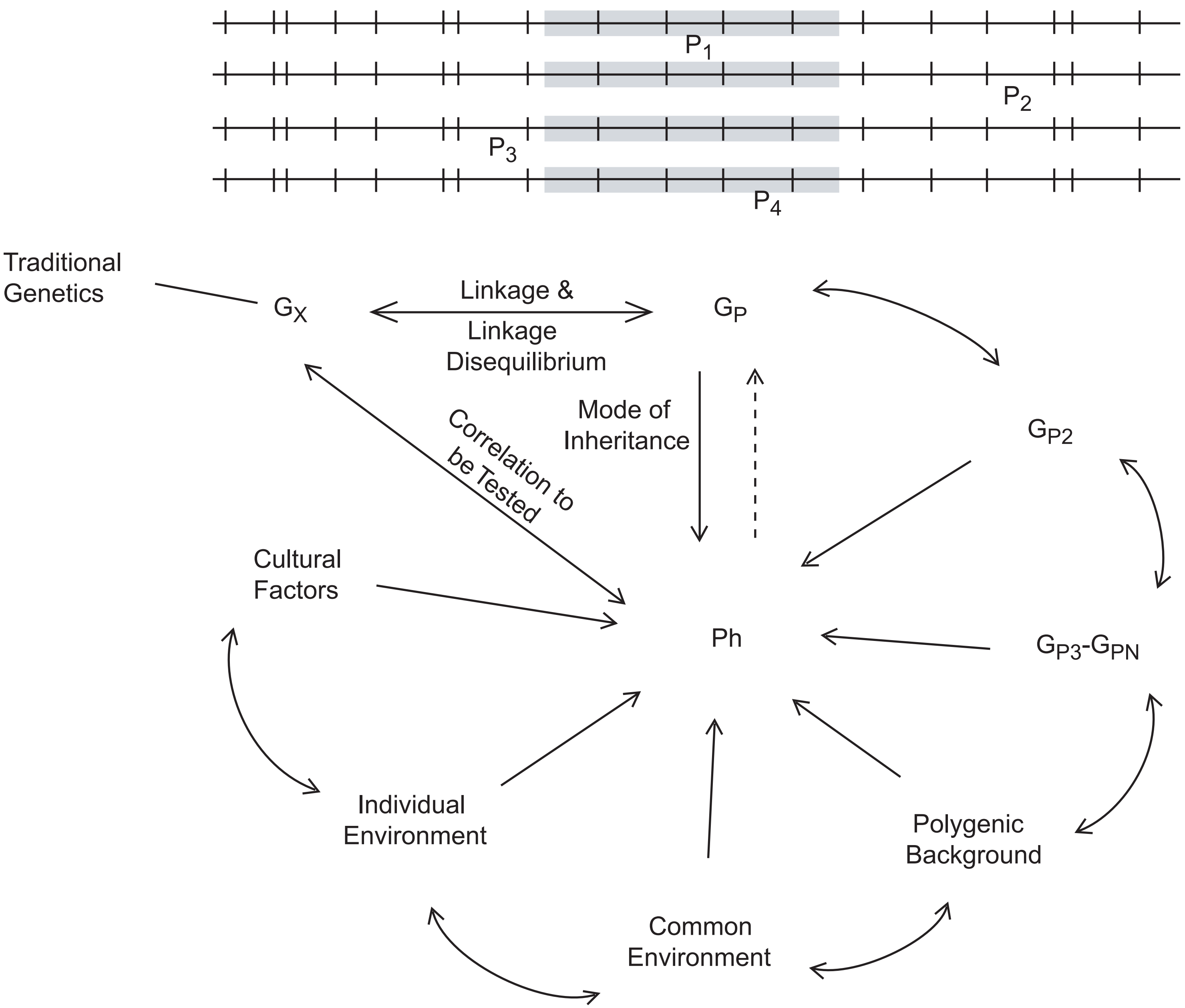
Genetic	*Genomic*
Structural genomics	Functional genomics
Genomics	Proteomics
Map-based gene discovery	Sequence-based gene discovery
Monogenic disorders	Multifactorial disorders
Specific DNA diagnosis	Monitoring of susceptibility
Analysis of one gene	Analysis of multiple genes in gene families, pathways, or systems
Gene action	Gene regulation
Etiology (specific mutation)	Pathogenesis (mechanism)
One species	Several species

Adapted from McKusick and Peltonen

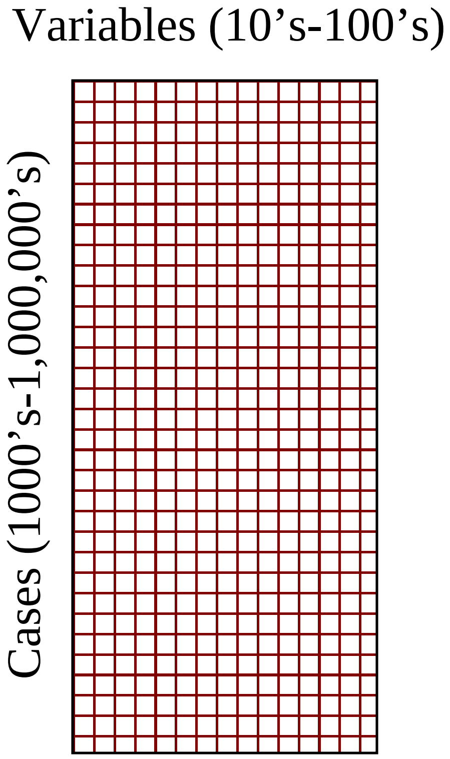
Comprehensive Bioinformatics Approach:
All Data Are Grist

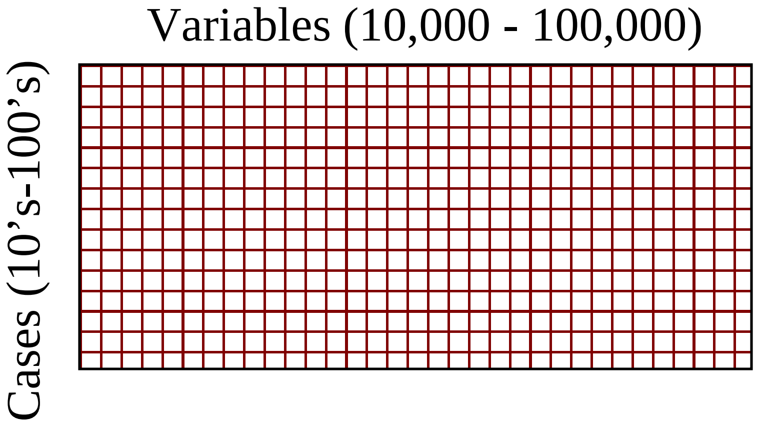
Reductionist methods that take into account data-type particularities

Interactions between all the “grist” is relevant to the health state



Determining the aforementioned interaction is hard and runs counter to traditional biostatistical techniques





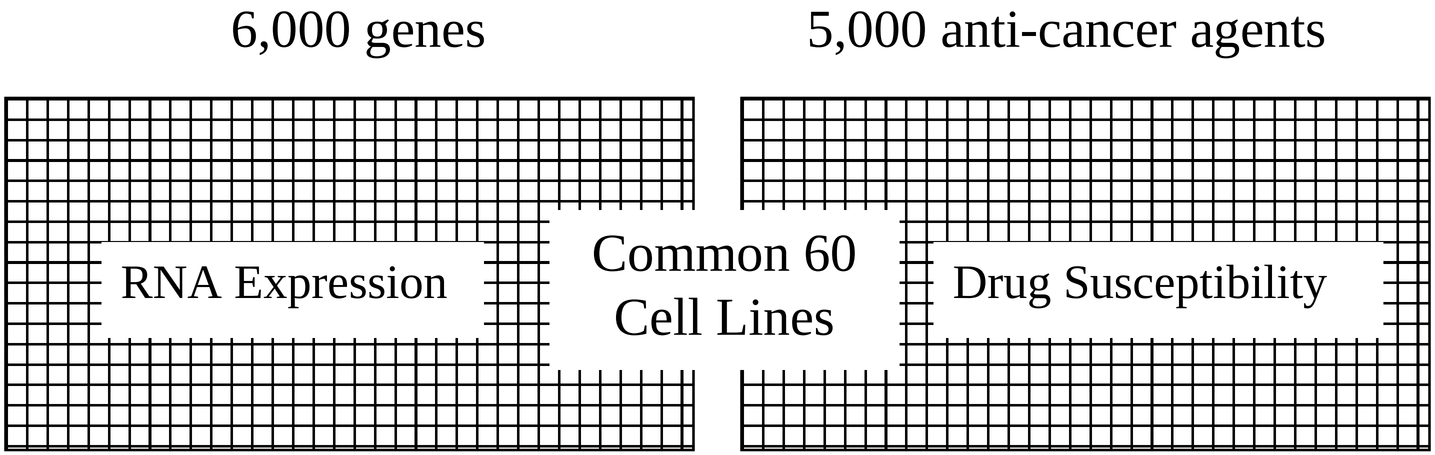
High-dimensionality systems with insufficient data are undeterdetermined

Not tractable by standard biostatistical techniques

The New Pharmacology

- RNA expression in NCI 60 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



6,000 genes

5,000 anti-cancer agents

RNA Expression

Common 60 Cell Lines

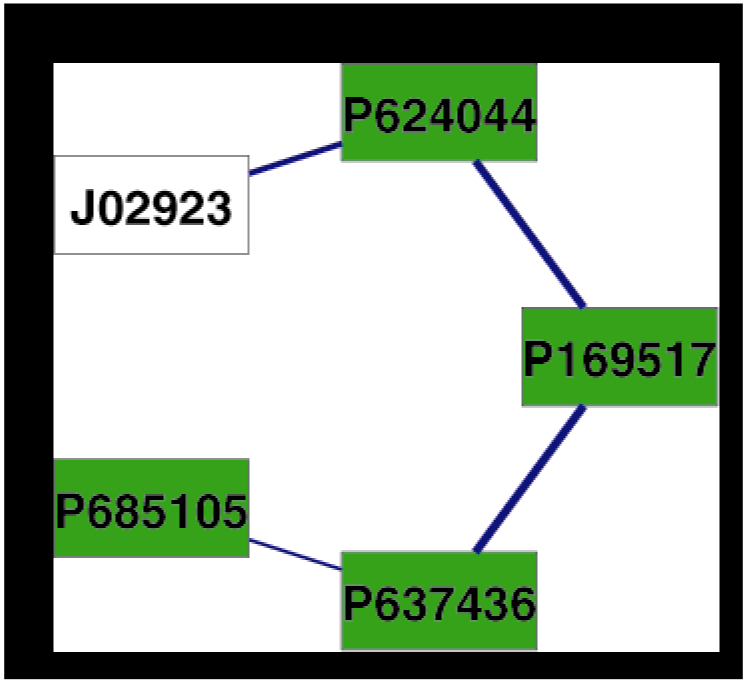
Drug Susceptibility

Genes and Anti-Cancer Agents

- Threshold $r^{\wedge 2}$ 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 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
- A role for LCP1 in tumorogenicity has been previously postulated
- Low level expression of LCP1 is thought to occur in most human cancer cell lines
- Other thiazolidine carboxylic acid derivatives are known to inhibit tumor cell growth



Butte et al. PNAS 2000

Overview

- The future is now
- Genomic vs genetic
- **Heredity**
- Resequencing of the diagnostic process
- Accelerating consumer activation

Heritability: the way a population geneticist would think of it.

- Heritability in the Broad Sense (H)
  - ✓ This measure of heritability includes all genetic influences on the phenotype, whether due to additive, dominance, or interactive effects.
  - ✓ $H^2 = V_G / V_P$, where $V_G = V_A + V_D + V_I$

Obesity

- Don’t some people just eat and not get fat?
- Isn’t it in their genes?

Heritability
All those of you with… leave the room

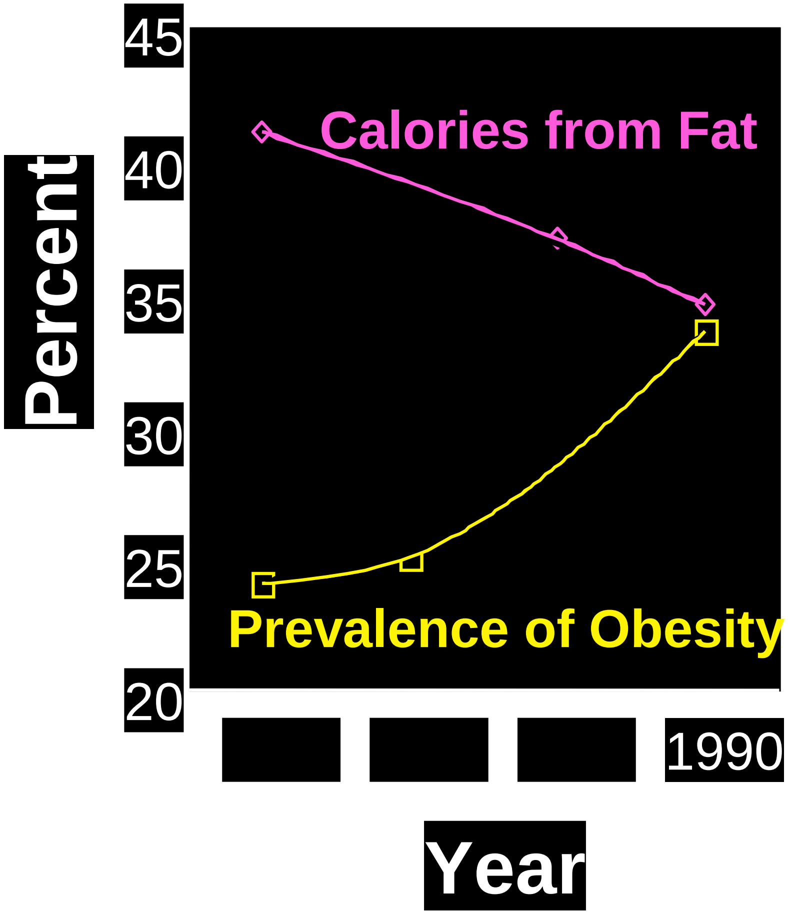
OBESITY

- **National Center for Health Statistics:**
  Over 50% of US adults have BMI > 25
  About 22% of US adults have BMI > 30

- **National Health & Nutrition Examination Survey III**:
  20% of U.S. children overweight

- **Behavioral Risk Factor Surveillance System (CDC)**
  Prevalence of obesity up by 50% from 1991 - 1998

Prevalence of Obesity Compared to Percent Calories from Fat Among US Adults



<u>*After:*</u>
Allred JB. JADA
95:417, 1995

Heritability is defined with respect to environment

- How do we define environment?
  - ✓ Diet
  - ✓ Daily habits
  - ✓ Environmental insults
  - ✓ Medical care
  - ✓ ….

Genotype does not capture the individual patient state

- We need to capture and quantify the environmental influences.
- We need to capture the effect of the genotype and environmental effects on the phenotype.
- These two comprise
  - ✓ **History**
  - ✓ Physical
  - ✓ Laboratory Studies
  - ✓ Imaging

The New Medicine

A More Expansive Reductionism

- More to the state description than genome
- Given necessity to capture both environment, genomic state and their interaction.
  - ✓ Only then we can elucidate V_E and V_G and V_{GE}
    - ☞ Clinical informatics and genomic/bioinformatics
  - ✓ Required for effective new therapies
  - ✓ Required for deeper understanding of mechanism
- Requires capturing the aforementioned interactions
  - ✓ The less we capture, the more *undetermined* the system

Overview

- The future is now
- Genomic vs genetic
- Heredity
- **Resequencing of the diagnostic process**
- Accelerating consumer activation

Pediatric Growth Monitoring

- Data:
  - heights, weights
  - family history
  - bone ages
  - pubertal data, stages
- Disorders show characteristic patterns on growth chart.

Work-up of Short Stature with Poor Growth

- T4, IGF-1, ESR, CBC, anti-gliaden Ab…
- Insulin Tolerance Test/Glucagon GH Test
  - ✓ 6 hours in the hospital
  - ✓ IV insulin with symptomatic hypoglycemia
  - ✓ Glucagon with nausea
  - ✓ $1000-$2000
- Interpretation remains controversial
  - ✓ Significant false positive rate: Why?
  - ✓ Significant false negative rate: Why?

1. What is the most common chromosomal cause of short stature?
2. 2.5% of idiopathic short stature children (including males) have SHOX mutations
3. Mutants are not growth hormone deficient but…

They respond to Growth Hormone therapy!

7325 articles on screening and diagnosis with PSA

From SMA-12 to SMA-30000

- Q_i= (< $T_{sel(1)}$,$R_{sel(1),}$$k$ $_{sel(1)}$ >, ..., < $T_{sel(i)}$,$R_{sel(i),}$$k$ $sel(i)$ >)

$$P_{H_j|Q_i} = \frac{P_{Hj|Qi} P_{H_j}}{\sum_{k=1}^{n} P_{Qi|Hk} P_{H_k}}$$

- Peforming i of m possible tests,
  - ✓ we can choose $_mP_i$ (= m! / (m-i)!) test sequences
- If every test has ***r*** possible results, then there will be $r_i mP_i$ possible test histories after ***i*** tests
- sum over test histories of every length and multiply by the number of hypotheses, ***n***
- n=10 hypotheses, m=5 binary tests (r=2)
  - ✓ the analysis requires 63,300 conditional probabilities

Re-engineering the knowledge-base

Cochrane Collaborative

1992 Founded to gather, edit evidence for medical practice; 50 Collaborative Review Groups

7000 Collaborators, internationally

Overview

- The future is now
- Genomic vs genetic
- Heredity
- Resequencing of the diagnostic process
- **Accelerating consumer activation**

The privacy challenge is now

Course Administrivia

- Both MIT and Harvard Spring Breaks observed
- Problem sets:
  - 2 total
- Final project (up to 2 persons per project)
  - Project selected by March 15th

Course Overview

- Biology Refresher
- Genomic Measurement Techniques
- Functional Genomics and Microarrays
- Limits of the Technologies: Noise
- Information Science at the Center of Genomic Medicine
- Informational Resources
- Modeling, Reverse Engineering
- The Importance of Data Representation

Course Review

- Machine Learning Approach
- Association with Markers
- Case Hx: Complex Trait Analysis
- Complex Traits: What to believe
- Microarray Disease Classification I
- Direct Prediction Outcome/Mortality
- Histopathology Case History
- Microarray Disease Classification II
- Practical Genomic Medicine: Today's Practice
- Individualized Pharmacology

Course Review

- Finding new drugs
- Ethical and Social Considerations
- Commercial and Regulatory Barriers
- Newborn Testing
- The New Microbiology