



Children's Hospital
Informatics Program


Harvard
Medical School

Computation in the New Medicine


Isaac S. Kohane



Children's Hospital
Informatics Program


Harvard
Medical School

Overview


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



Children's Hospital
Informatics Program


Harvard
Medical School

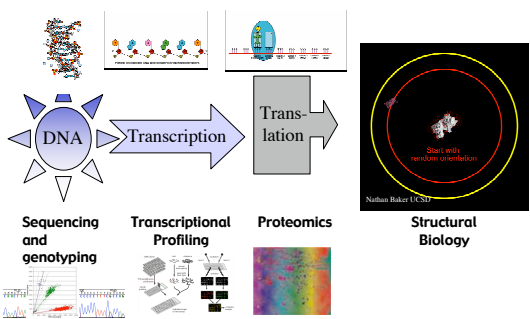
Overview

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


Children's Hospital
Informatics Program


Harvard
Medical School

The Long Path from Genotype to Function




Sequencing and genotyping


Transcriptional Profiling

Proteomics

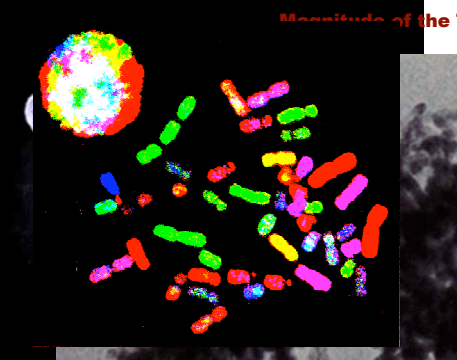
Structural Biology


Nathan Boker (UCSD): "Start with random variations"


Children's Hospital
Informatics Program


Harvard
Medical School

Complexity of the Task

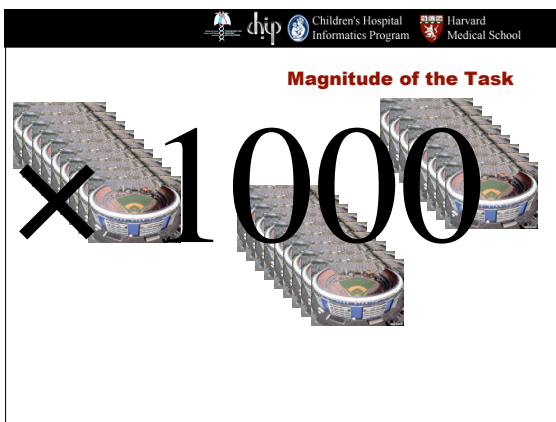
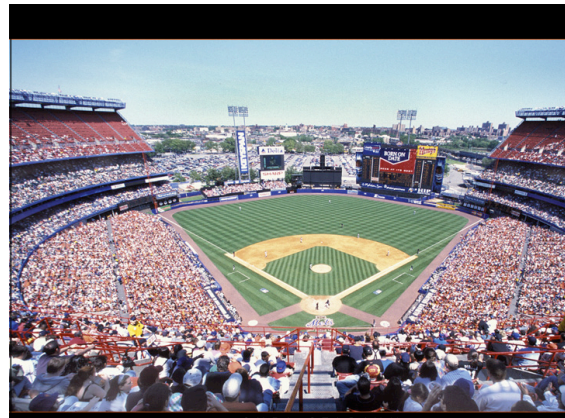
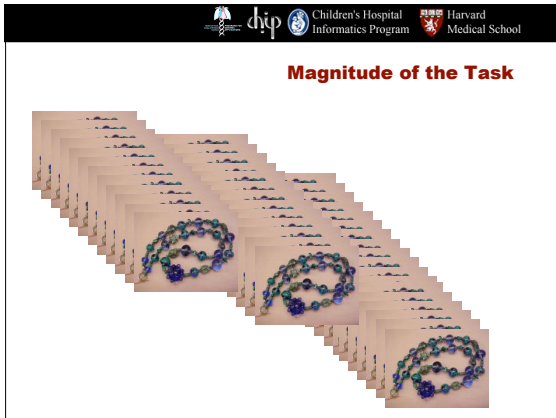



Children's Hospital
Informatics Program


Harvard
Medical School

Magnitude of the Task





Children's Hospital Informatics Program Harvard Medical School

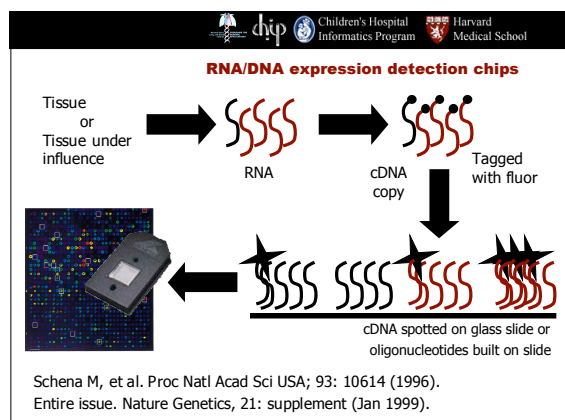
Madonna Complex

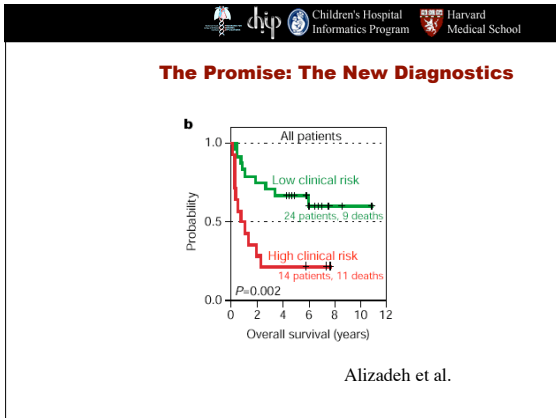
- Some say $\text{Madonna}_{\text{Music}} > \text{Madonna}_{\text{Person}}$
- 4.7 GB vs
- $(3 \times 10^9) \times 2 \text{ (bits/base)} / 8 \text{ (bits/byte)} = 0.75 \text{ GB}$

Children's Hospital Informatics Program Harvard Medical School

Madonna Complex

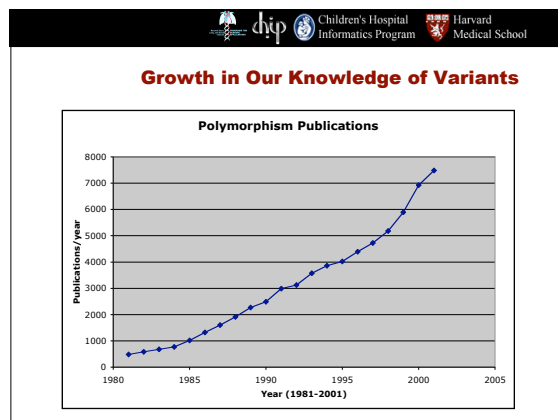
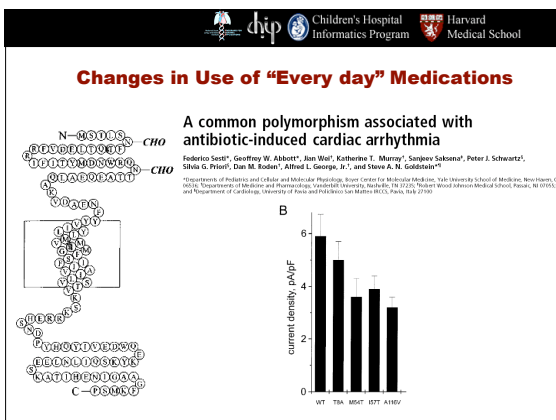
- Is Madonna, her DNA sequence?
- No, and her current state is captured by..
 - ✓ Alternative splicing ($\times 3/\text{gene}$)
 - ✓ Post-translational modification ($\times 100 - 1000$)
 - ✓ Location of gene product ($\times 10^{12}$)
- She's a little more complicated than her music (are you surprised?)

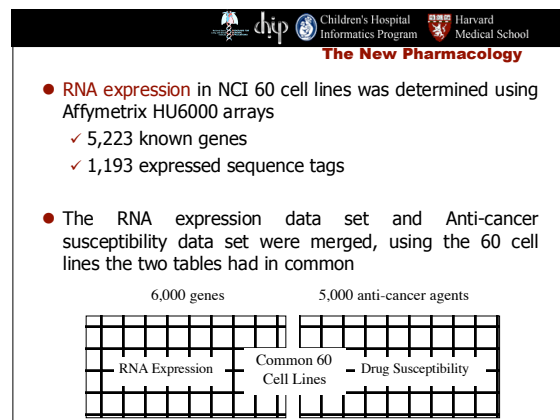
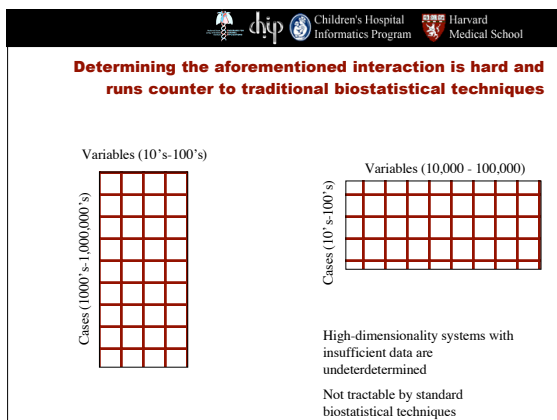
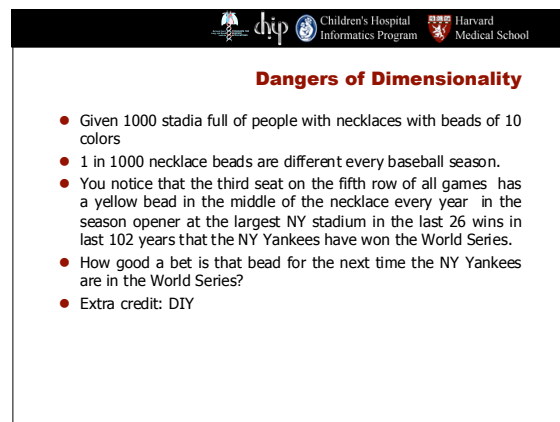
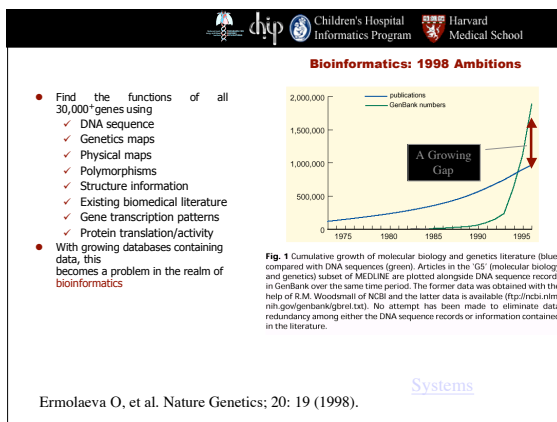
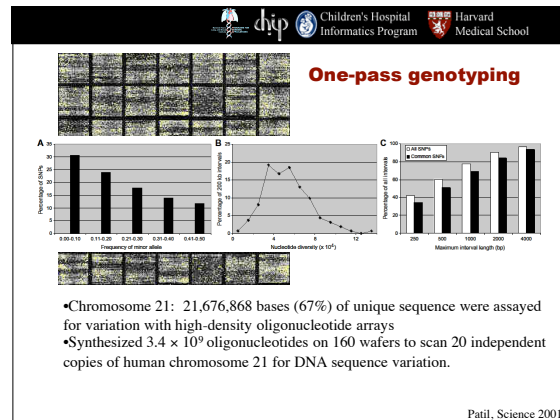
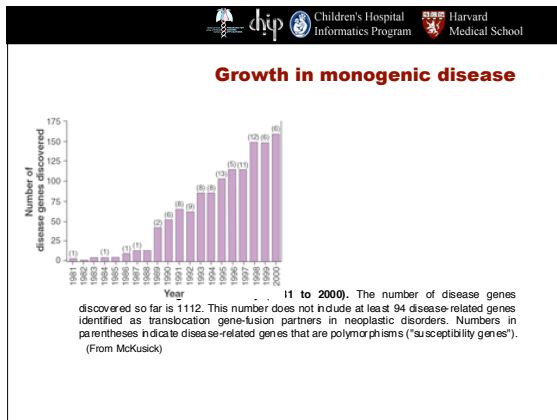




- Children's Hospital Informatics Program Harvard Medical School
- ### Retreading the cancer chemotherapeutic protocol
- Cancer and Leukemia Group B (CALGB)
 - CALGB has grown into a national network of 29 university medical centers, over 185 community hospitals and more than almost 3000 physicians who 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

- Children's Hospital Informatics Program Harvard Medical School
- ### 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.





Children's Hospital Informatics Program

Harvard Medical School

Genes and Anti-Cancer Agents

- 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

Children's Hospital Informatics Program

Harvard Medical School

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 tumorigenicity 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

Children's Hospital Informatics Program

Harvard Medical School

What does this have to do with biology and the life of a cell?

Layer	Description
7. Applications	User and application level tasks (e.g. resource sharing)
6. Presentation	Display, storage form etc.
5. Session	Establishes session: OS to transport abstraction
4. Transport	Guarantees end-to-end control within a session (who's calling, answering etc)
3. Network	Addressing, routing through nodes
2. Data link	Error free path between nodes
1. Physical	Electrical and mechanical

Children's Hospital Informatics Program

Harvard Medical School

What are the major lessons you learned in 6.001?

Children's Hospital Informatics Program

Harvard Medical School

Ant Levels of

(A) Organization of the IGH locus

(B) Construction of an immunoglobulin gene by genome rearrangement

Children's Hospital Informatics Program

Harvard Medical School

Overview

- The future is now and it computes
- 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

Malignant maths (From The Economist)

Overview

- The future is now and it computes
- 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^2)
 - ✓ 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

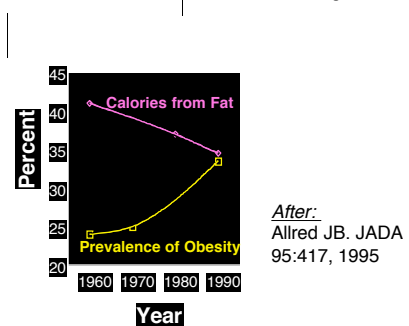
- Don't some people just eat and not get fat?
- Isn't it in their genes?



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



Children's Hospital Informatics Program Harvard Medical School

Heritability is defined with respect to environment

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

Children's Hospital Informatics Program Harvard Medical School

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

Children's Hospital Informatics Program Harvard Medical School

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

Children's Hospital Informatics Program Harvard Medical School

Overview

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

Children's Hospital Informatics Program Harvard Medical School

Pediatric Growth Monitoring


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


Boy with constitutional delay

Children's Hospital Informatics Program Harvard Medical School


Work-up of Short Stature with Poor Growth

- T4, IGF-1, ESR, CBC, anti-gliadin 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?






Center for
Human Genome
Research



Harvard
T.H. Chan School of
Public Health

Harvard
Medical School

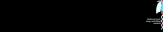


Gene	Symbol	Link	Cyto.	Full name
SHOX	SHOX	4p16.3	Xp22.33	Pseudoautosomal GTP-binding protein-like
SHOX	SHOX	4p16.3	Xp22.33	Short stature homeobox
SHOX	SHOX	4p16.3	Xp22.33	Visuospatial/perceptual abilities
SHOX	SHOX	4p16.3	Xp22.33	protein kinase, X-linked
SHOX	SHOX	4p16.3	Xp22.33	H-Y regulator, or repressor
SHOX	SHOX	4p16.3	Xp22.33	agamaglobulinemia, X-linked 2 (with growth hormone deficiency)
SHOX	SHOX	4p16.3	Xp22.33	Aicardi syndrome
SHOX	SHOX	4p16.3	Xp22.33	craniofrontonasal syndrome (craniofrontonasal dysplasia)
SHOX	SHOX	4p16.3	Xp22.33	deafness, X-linked 6, sensorineural
SHOX	SHOX	4p16.3	Xp22.33	multinuclear goiter 2

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

They respond to Growth Hormone therapy!

[illegible]







Chair of Informatics
Chair of Information Systems
University of Applied Sciences

From SMA-12 to SMA-30000

- $Q = (< T_{sel(1)}, R_{sel(1)}, k_{sel(1)} > , \dots , < T_{sel(l)}, R_{sel(l)}, k_{sel(l)} >)$

$$P_{H_i|Q_i} = \frac{P_{HQ_i} P_{H_i}}{\sum_{j=1}^n P_{Q_i H_j} P_{H_j}}$$

- Performing 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 $r m P_i$ 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

Children's Hospital
Informatics Program

Harvard
Medical School

Re-engineering the knowledge-base




The scope of the CFGD Group is as follows:

1. **Cystic Fibrosis**
2. **Inborn errors of metabolism** (eg Gaucher's disease, phenylketonuria, galactosaemia). All interventions concerned with the prevention, early detection and treatment of inborn errors of metabolism.
3. **Haemoglobinopathies** (eg sickle cell disease and thalassaemias). All interventions concerned with the prevention, early detection and treatment of haemoglobinopathies.
4. **Inherited coagulopathies** (eg haemophilia). All interventions concerned with the prevention, early detection and treatment of inherited/pair coagulopathies. There are genetic disorders which do not clearly fall into these areas.

Other reviews in genetic disorders will be considered for inclusion by the editorial team and discussed with potential reviewers. A decision about inclusion will depend on: a) whether it might be more appropriately placed in another registered or emerging group b) an analysis of this paper.

If the review is likely to use considerable administrative or editorial resources (searching time, for example), we need to check the review topic is acceptable and a priority to the body that funds the Editorial Base.

- top of this page | about the group | about the collaboration | links -



MIT

Children's Hospital
Informatics Program

Harvard
Medical School

Overview

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

[illegible]

GeneTree
DNA TESTING CENTER

Call Us Toll Free! 1-888-404-GENE

About Us | Find Your Test | Questions | Affiliates | Language

Save Now by Ordering a Prepaid Kit!

Save \$20 on the purchase of your Personal Paternity Test results and get **FREE** FedEx shipping (a \$10 value) by purchasing a prepaid test kit now.

Simply follow the **easy, convenient** steps to collect your samples and send them back to us for **quick, 99.9% accurate** results.

GeneTree DNA tests are:

- **Confidential** – your complete privacy guaranteed
- **Accurate** – 99.9% reliability
- **Low Priced** – services start at \$225.
- **Fast** – results in 3-7 days
- **Convenient** – at-home cheek swab collection
- **Easy** – telephone support (888) 404-GENE

Take advantage of this limited-time offer and **save \$30**
*Offer ends February 15, 2003

YES! I Want To Order My Prepaid Kit!

Top Five Sellers

1. Free Test Kit
2. Personal Paternity DNA Test
3. Legal Paternity DNA Test
4. Siblingship DNA Analysis
5. Y-Chromosome Comparative Analysis

Search Website

Search

Advanced Search

Client Testimonials

"Thank to GeneTree... I can finally look into her eyes and feel at ease knowing that I did the right thing."


Children's Hospital Informatics Program | Harvard Medical School

"A simple new test could save your baby's life." The text describes a newborn screening kit that can detect more disorders than most state screening programs "for your baby and for your peace of mind."¹¹

- An advertisement for diagnostic testing, appearing in a Jewish magazine, uses this headline above a photograph of a weary, dark-haired woman: "If you often feel tired, it could be anemia. If you're Jewish, it could be Gaucher disease."¹³
- An advertisement for carrier testing in a Jewish community newspaper asks its readers, "Are you a carrier?" and lists 8 "Jewish genetic conditions" for which the audience may be at risk.⁹

Children's Hospital Informatics Program | Harvard Medical School

Current trends in criminal inventory and state screening



I AM NOT A PIECE OF YOUR INVENTORY.

Children's Hospital Informatics Program | Harvard Medical School

The privacy challenge is now



The men of Wee Waa

Children's Hospital Informatics Program | Harvard Medical School

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

Children's Hospital Informatics Program | Harvard Medical School

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