Children's Hospital Informatics Program | Harvard Medical School

Computing and Genomic Medicine

Isaac S. Kohane

http://www.chip.org

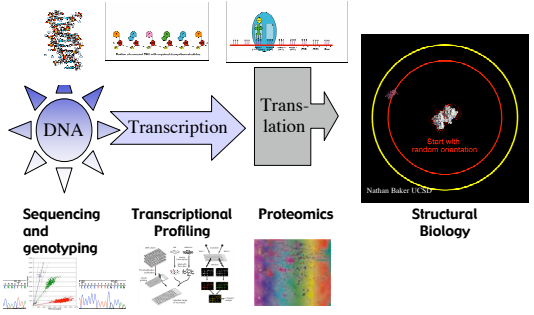
Children's Hospital Informatics Program | Harvard Medical School

Overview

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

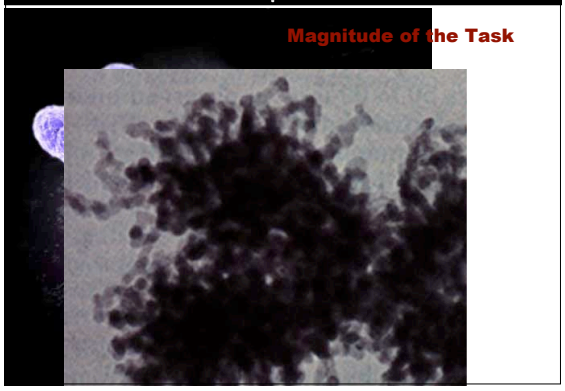
Children's Hospital Informatics Program | Harvard Medical School

The Long Path from Genotype to Function



Children's Hospital Informatics Program | Harvard Medical School

Magnitude of the Task

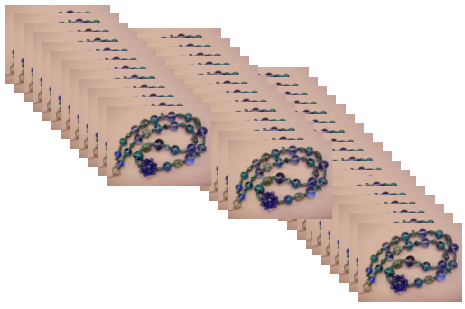


Children's Hospital Informatics Program | Harvard Medical School

Magnitude of the Task

Children's Hospital Informatics Program | Harvard Medical School

Magnitude of the Task

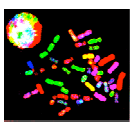


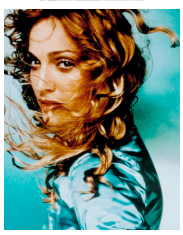
Children's Hospital Informatics Program — Harvard Medical School

Magnitude of the Task

Children's Hospital Informatics Program — Harvard Medical School

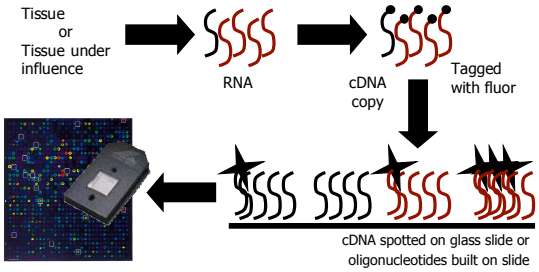
Myhrvold's Madonna Complex





Children's Hospital Informatics Program | Harvard Medical School

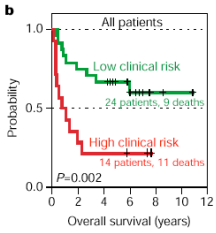
RNA/DNA expression detection chips



Schena M, et al. Proc Natl Acad Sci USA; 93: 10614 (1996).
Entire issue. Nature Genetics, 21: supplement (Jan 1999).

Children's Hospital Informatics Program | Harvard Medical School

The Promise: The New Diagnostics



Alizadeh et al.

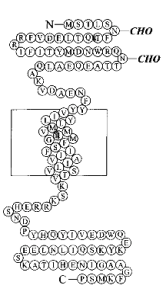
Children's Hospital Informatics Program | Harvard Medical School

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

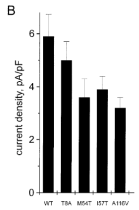
Children's Hospital Informatics Program | Harvard Medical School

Changes in Use of "Every day" Medications



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*¶



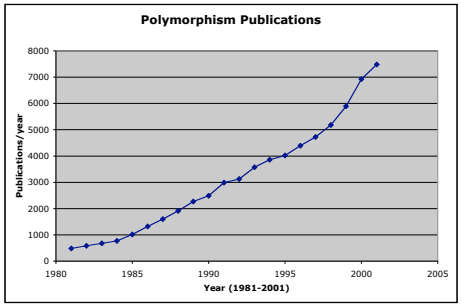
Children's Hospital Informatics Program | Harvard Medical School

Table S3: *EGFR* mutation status in gefitinib-treated lung cancer

Gefitinib sensitivity	Sample	Histology	Source	Gender	Exons	Sequence alteration	Nucleotide
Y	IR1T	adenocarcinoma	U.S.	M	19	Del-3	2239_2247delTTAAGAGAA, 2
Y	P003	adenocarcinoma	U.S.	M	19	Del-3	2239_2247delTTAAGAGAA, 2
Y	IR4T	bronchioloalveolar carcinoma	U.S.	F	19	Del-4	2240_2257delTAAGAGAAGCA
Y	IR2T	adenocarcinoma	U.S.	F	19	Del-5	2238_2255delATTAAGAGAAGC 2237A>T
Y	IR3T	adenocarcinoma	U.S.	F	21	Substitution	2573T>G
Y *in vitro*	IRG H3255	adenocarcinoma	U.S.	F	21	Substitution	2573T>G
N	IR5	adenocarcinoma	U.S.	F	18-24	None detected	n/a
N	IR6	adenocarcinoma	U.S.	M	18-24	None detected	n/a
N	IR8	adenocarcinoma	U.S.	F	18-24	None detected	n/a
N	IR9	NSCLC	U.S.	F	18-24	None detected	n/a

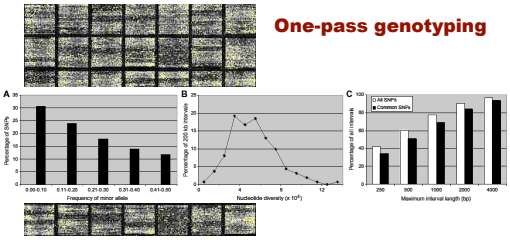
Children's Hospital Informatics Program | Harvard Medical School

Growth in Our Knowledge of Variants



Children's Hospital Informatics Program — Harvard Medical School

One-pass genotyping



•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.

Patil, Science 2001

Children's Hospital Informatics Program — Harvard Medical School

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

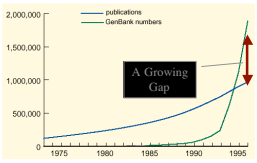


Fig. 1 Cumulative growth of molecular biology and genetics literature (blue) compared with DNA sequences (green). Articles in the 'GS' (molecular biology and genetics) subset of MEDLINE are plotted alongside DNA sequence records in GenBank over the same time period. The former data was obtained with the help of R.M. Woodsmall of NCBI and the latter data is available (ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt). No attempt has been made to eliminate data redundancy among either the DNA sequence records or information contained in the literature.

Systems

Ermolaeva O, et al. Nature Genetics; 20: 19 (1998).

Children's Hospital Informatics Program — Harvard Medical School

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!"

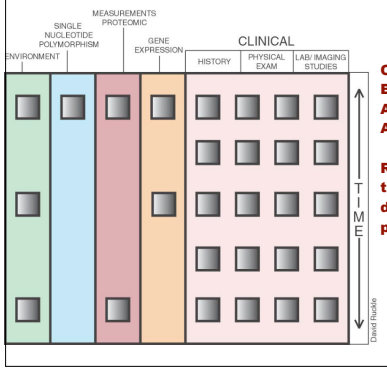
Finally, the statistician says, "It is not statistically significant!"

Children's Hospital Informatics Program | Harvard Medical School

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 geno me

Children's Hospital Informatics Program | Harvard Medical School

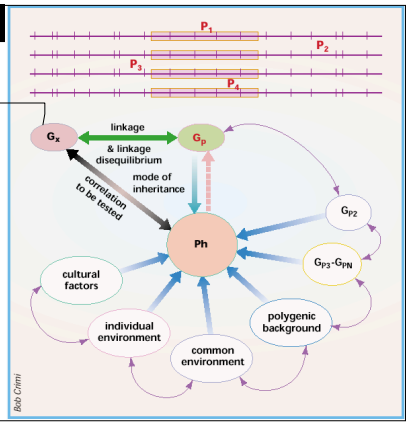


Comprehensive Bioinformatics Approach: All Data Are Grist

Reductionist methods that take into account data-type particularities

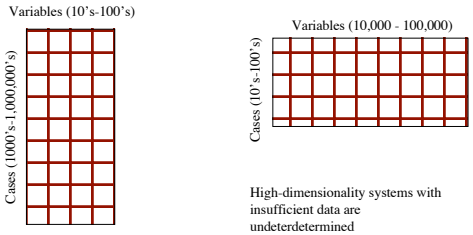
Traditional genetics

Interactions between all the "grist" is relevant to the health state



Children's Hospital Informatics Program | Harvard Medical School

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



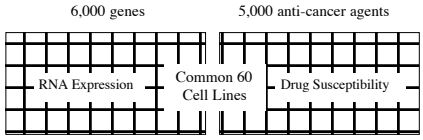
High-dimensionality systems with insufficient data are undeterdetermined

Not tractable by standard biostatistical techniques

Children's Hospital Informatics Program | Harvard Medical School

The New Pharmacology

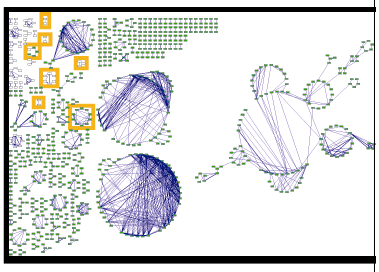
- RNA expression in NCI 60 cell lines was determined using Affymetrix HU6000 arrays
  - ✓ 5,223 known genes
  - ✓ 1,193 expressed sequence tags
- The RNA expression data set and Anti-cancer susceptibility data set were merged, using the 60 cell lines the two tables had in common



Children's Hospital Informatics Program | Harvard Medical School

Genes and Anti-Cancer Agents

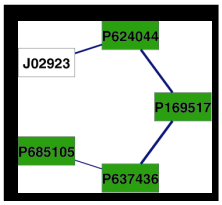
- Threshold $r^{\wedge 2}$ was 0.8
- 202 networks
- 834 features out of 11,692 (7.1%)
- 1,222 links out of 68,345,586 (.0018%)
- Only one link between a gene and anti-cancer agent



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 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

Children's Hospital Informatics Program — Harvard Medical School

Overview

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

Children's Hospital Informatics Program — Harvard Medical School

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 $V_G = V_A + V_D + V_I$

Children's Hospital Informatics Program — Harvard Medical School

Obesity

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

Children's Hospital Informatics Program — Harvard Medical School

Heritability

All those of you with… leave the room

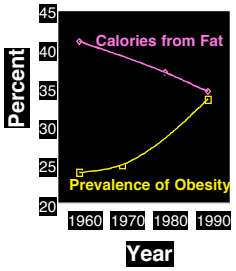
Children's Hospital Informatics Program — Harvard Medical School

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

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



After:
Allred JB. JADA 95:417, 1995

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

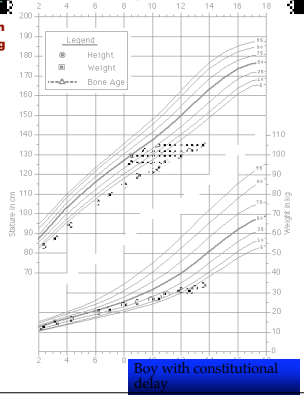
Overview

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

Children's Hospital — Harvard

Pediatric Growth Monitoring

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



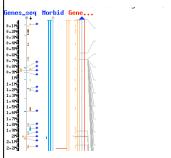
Boy with constitutional delay

Children's Hospital Informatics Program | Harvard Medical School

Work-up of Short Stature with Poor Growth

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

Children's Hospital Informatics Program | Harvard Medical School

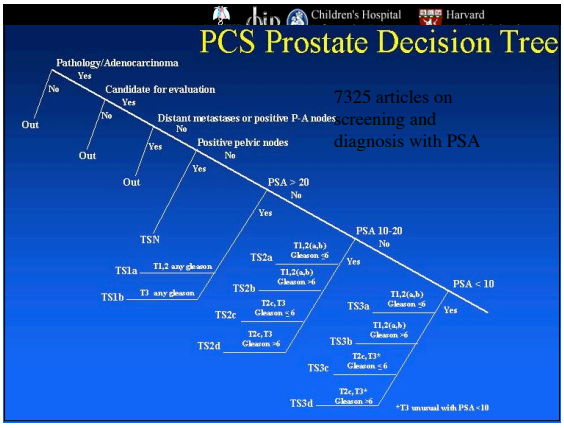


symb	links	cyto.	full name
PGPL	sv ex	Xp22.33	Pseudoautosomal GTP-binding protein-like
SHOX	- - -	Xpter-p22.32:Yp11.3	short stature homeobox
VSPA	- - -	Xp22.33	Visuospatial/perceptual abilities
PRKX	- - -	Xp22.3	protein kinase, X-linked
HYR	- - -	Xp22.3	H-Y regulator, or repressor
AGMX2	- - -	Xp22	agammaglobulinemia, X-linked 2 (with growth hormone deficiency)
AIC	- - -	Xp22	Aicardi syndrome
CFNS	- - -	Xp22	craniofrontonasal syndrome (craniofrontonasal dysplasia)
DFN6	- - -	Xp22	deafness, X-linked 6, sensorineural
MNG2	- - -	Xp22	multinodular goiter 2

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!

Children's Hospital | Harvard



7325 articles on screening and diagnosis with PSA

Children's Hospital Informatics Program | Harvard Medical School

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

Children's Hospital Informatics Program | Harvard Medical School

Re-engineering the knowledge-base

Cochrane C

1992 Found
for medical
Review Gro

7000 Collab

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\par 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 the time required.
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.

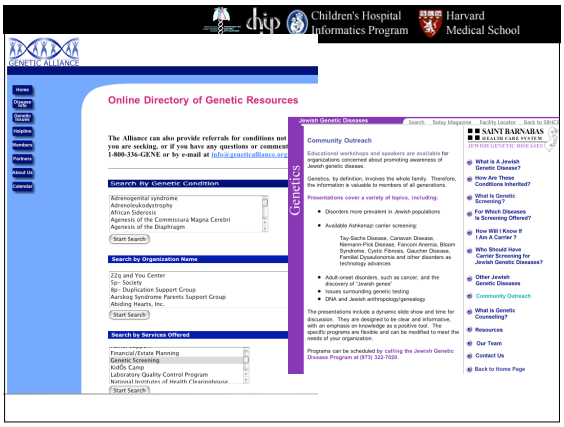
- top of this page | about the group | about the collaboration | links -

Children's Hospital Informatics Program | Harvard Medical School

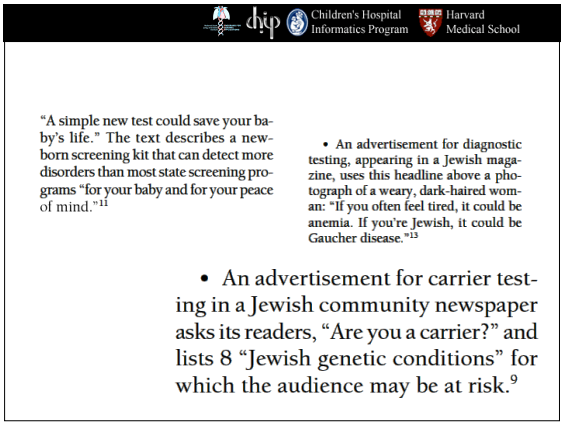
What kind of testing are doctors ordering?

- Just looking at Cancer Susceptibility Tests (CST)
  - ✓ 1251 physicians (820 primary care)
- In 12 months, approx 30% ordered or referred (only 7% directly ordered)
- Factors affecting ordering:

Northeast, feeling competent, advertising materials, and most importantly having the patient ask for it.


Children's Hospital Informatics Program
Harvard Medical School
GENETIC ALLIANCE
Online Directory of Genetic Resources
Search by Genetic Condition
Search by Organization Name
Search by Services Offered
Genetics
Community Outreach
SAINT BARNABAS HEALTH CARE SYSTEM
JEWISH GENETIC DISEASES


Call Us Toll Free! 1-888-404-GENE
View Your Cart
About Us | Find Your Test | Questions | Affiliates | Language
GeneTree
DNA TESTING CENTER
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
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
Save now!
Take advantage of this limited-time offer and save $30
*Offer ends February 15, 2003
YES! I Want To Order My Prepaid Kit!


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]

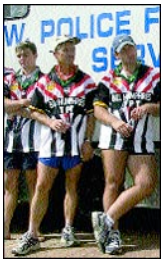
Children's Hospital Informatics Program | Harvard Medical School

Current trends in criminal inventory and state screening

Children's Hospital Informatics Program | Harvard Medical School

The privacy challenge is now

How many variants define a person uniquely ?

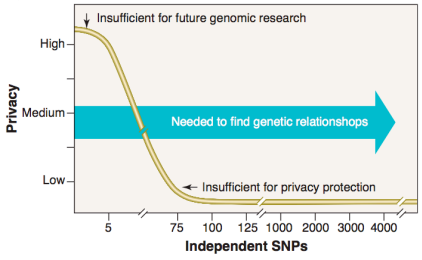


The men of Wee Waa

Children's Hospital Informatics Program | Harvard Medical School

Tension between research and privacy

Altman, Science 2004



Trade-offs between SNPs and privacy.

Children's Hospital Informatics Program | Harvard Medical School

History

- The modern history of medical controlled vocabularies begins with the U.S. Army General Surgeon who petitioned Congress to fund a medical library.
- The position eventually became "The US Surgeon General" and the library the National Library of Medicine
  - ✓ http://www.nlm.nih.gov/

Children's Hospital Informatics Program | Harvard Medical School

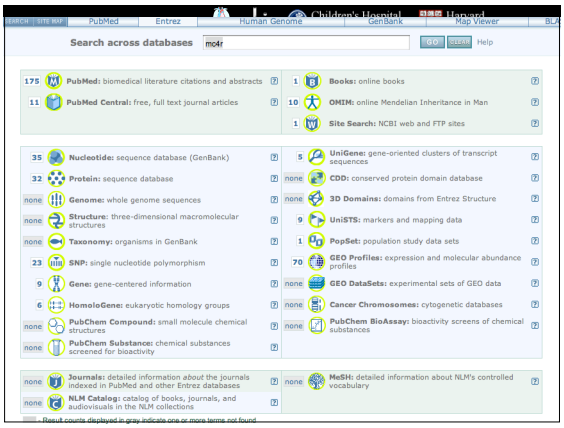
History

- Library collection was indexed with Index Medicus (created by NLM) which is published in book form.
- Index Medicus was extended to index medical literature articles.
- Index Medicus was extended further to provide on-line indexing (1960). This became the Medical Subject Headings (MeSH).
- Exponential growth has continued for a century, well past the capacity of paper records
- Today: 15 million citations for biomedical articles back to the 1950's.

Children's Hospital Informatics Program | Harvard Medical School

As the molecular aspects of disease became clear

- 15 million citations for biomedical articles back to the 1950's.
- 1998, National Center for Biotechnology Information (NCBI) formed as a complement to NLM.
- Databases and standardized nomenclatures.
- One NCBI database is GenBank, the nucleic acid sequence database that contains sequence information from more than 100 000 different organisms
- Now constructed with the understanding of the fundamentally interwoven character of biological information.
  - ✓ Closes we have to a unified thesaurus if not a unified biological theory.



Children's Hospital Informatics Program — Harvard Medical School

Your assignment

- A 60 year old man has been found face down, and dehydrated. You are one of a team of assistants working with a the genomic diagnostician. He gives you the following sequence.

- MLGQSTEELRVRLASHLRKLKRLLRDADDLQKRLAVYQAG
- Your assignment is to determine if this sequence might inform her about the cause of the presentation.
- In answering the question:
  - ✓ What part of the genome is this?
  - ✓ What diseases are associated with it?
    - ☞ Is it an unusual variant?
  - ✓ In which other organisms is this kind of sequence found?
  - ✓ Are there model organisms to study this variant?
- Document how you arrive at these answers.