The Genetic and Evolutionary Computation Conference (GECCO'12)

BIOINFORMATICS

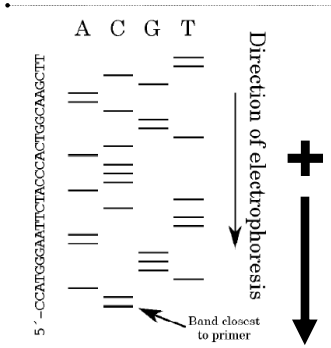
William S. Bush, Ph.D.

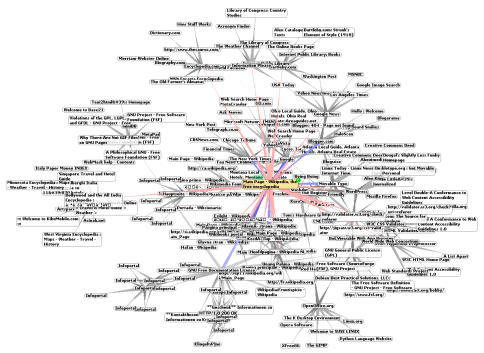
Assistant Professor of Biomedical Informatics
Center for Human Genetics Research
Vanderbilt University, Nashville, Tennessee USA
http://www.sigevo.org/gecco-2012/
william.s.bush@vanderbilt.edu
gettinggeneticsdone.com
chgr.mc.vanderbilt.edu/bushlab

GECCO'12 Companion, July 7–11, 2012, Philadelphia, PA, USA.
ACM 978-1-4503-1178-6/12/07.

ABOUT THE INSTRUCTOR

- William S. Bush is an Assistant Professor of Biomedical Informatics in the Center for Human Genetics Research at Vanderbilt University. Using a combination of bioinformatics, basic statistical approaches, and more advanced data mining and machine learning techniques, he studies how patterns of genomic variation influence the function of both individual genes and entire biological systems.





BIOINFORMATICS (1990)

Bioinformatics

Mark S Boguski

National Center for Biotechnology Information, Bethesda, USA

Computer databases, networks and software tools are essential materials and methods for biomedical research and are involved in almost every aspect of disease gene mapping and positional cloning. Public databases of DNA and protein sequences and genetic and physical map information are increasing rapidly in size and complexity and are also improving in quality, comprehensiveness, interoperability and access. A new generation of software tools for navigating through the biomedical literature has become available. Programs for sequence homology searching and genetic map construction have become more sophisticated, yet easier to use. Global computer networks are bringing state-of-the-art capabilities to all.

Current Opinion in Genetics and Development 1994, 4:383–388

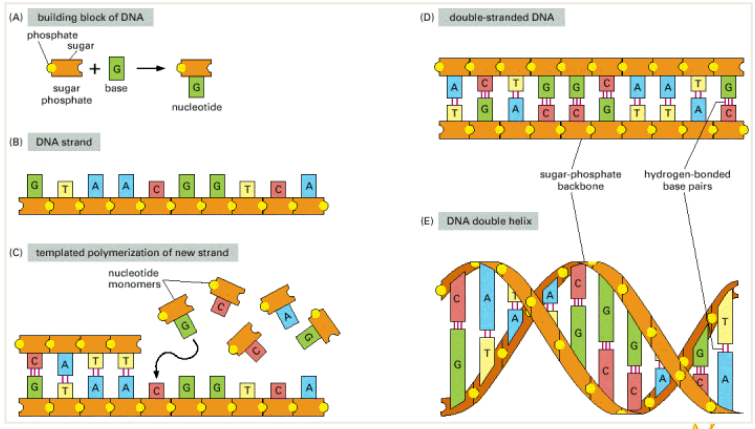
Table 1. Selected World-Wide Web (WWW) and Gopher servers.

Resource	WWW Uniform Resource Locator or Gopher address	Features and comments
World-Wide Web	http://info.cern.ch/	WWW project background; pointers to the world's on-line information; information on WWW software products; frequently asked questions (and answers).
NCSA Mosaic	http://www.ncsa.uiuc.edu/	Starting points for Internet exploration; Internet resources meta-index; "What's New" with NCSA and Internet.
Virtual Library: Biosciences	http://golgi.harvard.edu/	Comprehensive list and links to WWW resources for biology and medicine.
ExPASy	http://expasy.hcuge.ch/	Comprehensive library of documents describing e-mail servers, databases and software for molecular biology.
NCBI GenBank®	http://ncbi.nlm.nih.gov/	Information for submitting and updating sequences; GenBank® release notes; homology and text searching of sequence databases; *Entrez* MEDLINE browser.
Genome Data Base (GDB)	http://gdbwww.gdb.org/	Search GDB and On-line Mendelian Inheritance in Man (OMIM)
CEPH/Généthon	http://www.genethon.fr/	Human genome genetic and physical mapping data; search for information on YAC clones and STSs; QUICKMAP software for viewing CEPH map.
Cooperative Human Linkage Center (CHLC)	gopher://gopher.chlc.org/	Markers, genotype data and integrated maps.
UK Human Genome Mapping Project	gopher://menu.crc.ac.uk/	Primers, probes and chromosome abnormality database; cell lines from patients with genetic disorders or cytogenetic abnormalities.
MIT Genome Center	http://www-genome.wi.mit.edu/	Quarterly data releases from the human physical mapping and mouse genetic mapping projects.
Jackson Laboratory	http://www.jax.org/	Locus catalog and genetic maps of the mouse.
Dan's Favorite BioGopher	gopher://gopher.gdb.org/	Library catalogs around the world; world-wide campus phonebooks; search for people by name, location, research interest or funding agency.

Genotype to Phenotype

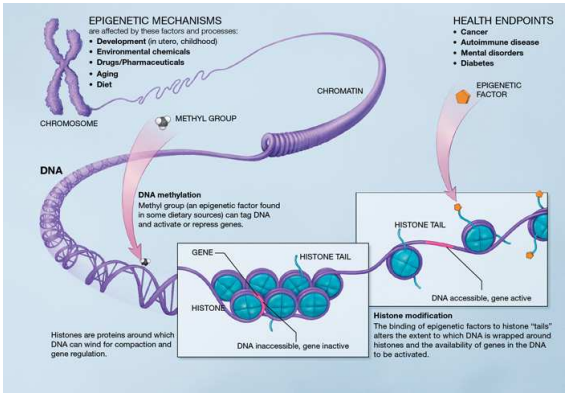
DNA

Alberts et al. 2002



Chromatin

NIH Common Fund http://commonfund.nih.gov/epigenomics/figure.aspx



Prediction of Chromatin States

Ernst et al. 2011

ARTICLE

doi:10.1038/nature09906

Mapping and analysis of chromatin state dynamics in nine human cell types

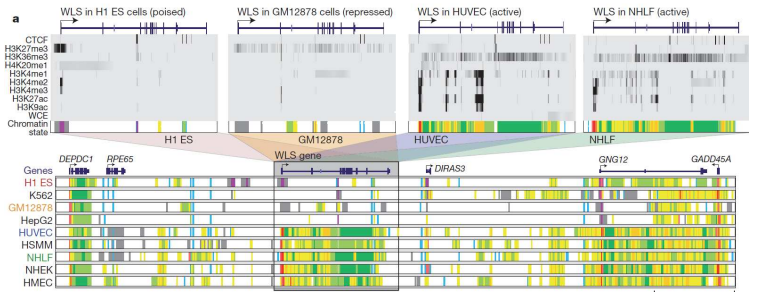
Jason Ernst[1,2], Pouya Kheradpour[1,2], Tarjei S. Mikkelsen[1], Noam Shoresh[1], Lucas D. Ward[1,2], Charles B. Epstein[1], Xiaolan Zhang[1], Li Wang[1], Robbyn Issner[1], Michael Coyne[1], Manching Ku[1,3,4], Timothy Durham[1], Manolis Kellis[1,2]* & Bradley E. Bernstein[1,3,4]*

Candidate state annotation

- Active promoter
- Weak promoter
- Inactive/poised promoter
- Strong enhancer
- Strong enhancer
- Weak/poised enhancer
- Weak/poised enhancer
- Insulator
- Transcriptional transition
- Transcriptional elongation
- Weak transcribed
- Polycomb repressed
- Heterochrom; low signal
- Repetitive/CNV
- Repetitive/CNV

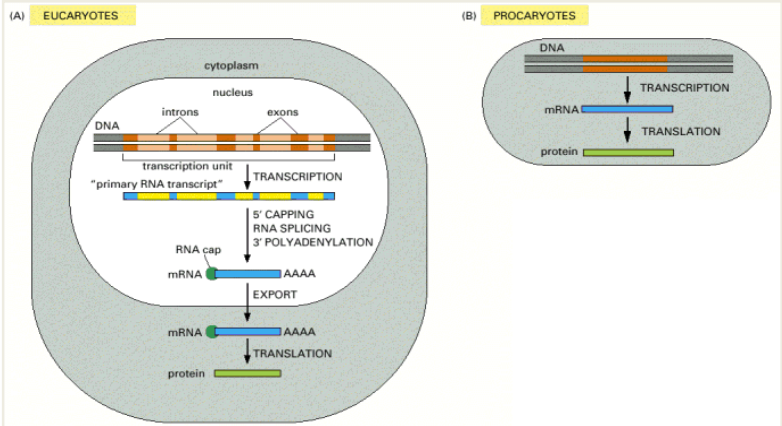
Prediction of Chromatin States

Ernst et al. 2011



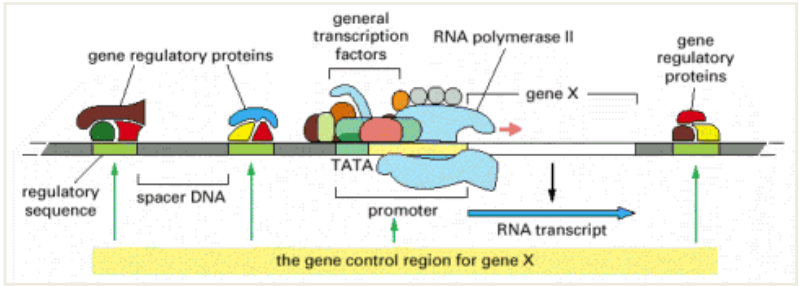
Transcription

Alberts et al. 2002



Regulation of Transcription

Alberts et al. 2002



Transcriptional Complexity

26 JUNE 2009 VOL 324 **SCIENCE**

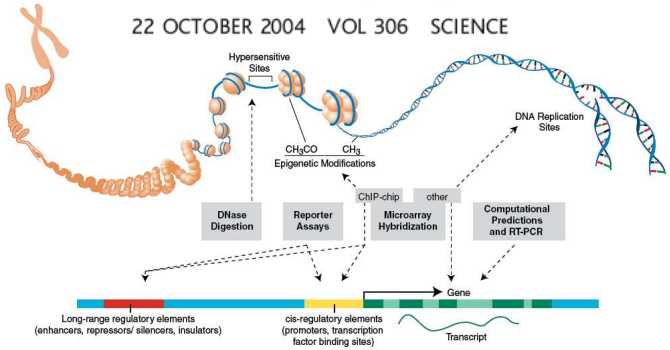
Diversity and Complexity in DNA Recognition by Transcription Factors

Gwenael Badis,[1*] Michael F. Berger,[2,3*] Anthony A. Philippakis,[2,3,4*] Shaheynoor Talukder,[1,5*] Andrew R. Gehrke,[2*] Savina A. Jaeger,[2*] Esther T. Chan,[5*] Genita Metzler,[6] Anastasia Vedenko,[7] Xiaoyu Chen,[1] Hanna Kuznetsov,[6] Chi-Fong Wang,[8] David Coburn,[1] Daniel E. Newburger,[2] Quaid Morris,[1,5,9,10] Timothy R. Hughes,[1,5,10]† Martha L. Bulyk[2,3,4,11]†

Transcriptional Complexity

The ENCODE (ENCyclopedia Of DNA Elements) Project

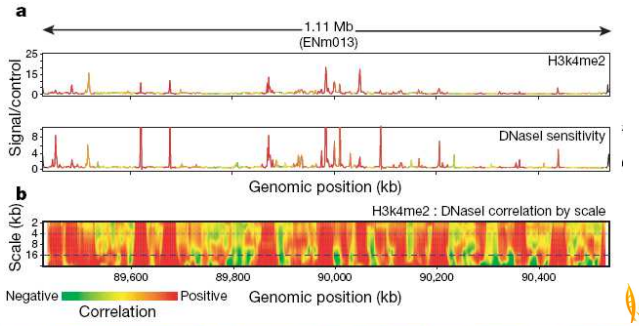
22 OCTOBER 2004 VOL 306 SCIENCE



Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

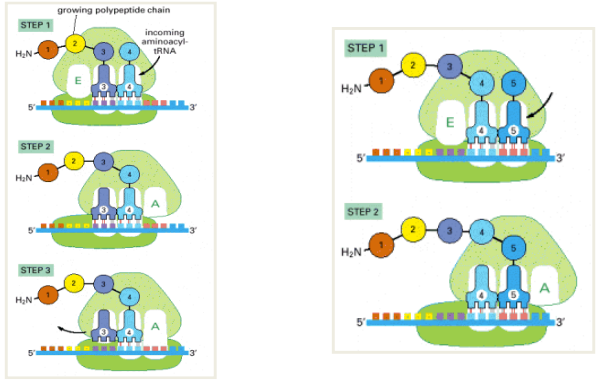
NATURE|Vol 447|14 June 2007

The ENCODE Project Consortium*



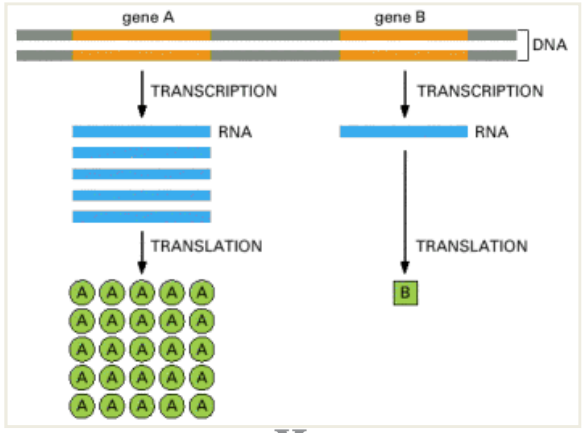
Translation

Alberts et al. 2002



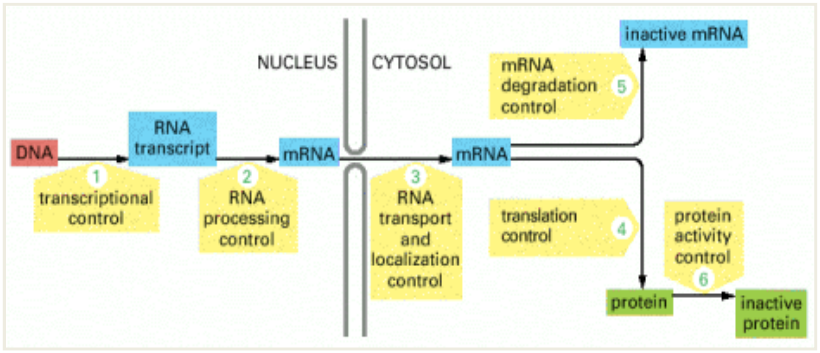
Translation

Alberts et al. 2002



Control of Gene Expression

Alberts et al. 2002

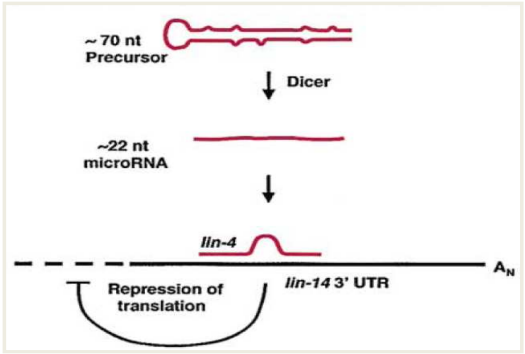


Alberts et al. 2002

Cell, Vol. 107, 823–826, December 28, 2001, Copyright ©2001 by Cell Pr

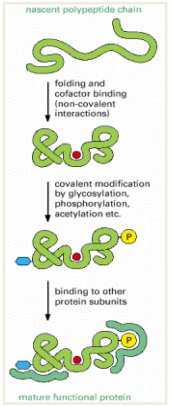
microRNAs: Tiny Regulators with Great Potential

Victor Ambros[1]
Department of Genetics
Dartmouth Medical School
Hanover, New Hampshire 03755



Protein Folding

Alberts et al. 2012



Amino Acids

Alberts et al. 2012

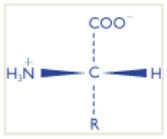
AMINO ACID			SIDE CHAIN	AMINO ACID			SIDE CHAIN
Aspartic acid	Asp	D	negative	Alanine	Ala	A	nonpolar
Glutamic acid	Glu	E	negative	Glycine	Gly	G	nonpolar
Arginine	Arg	R	positive	Valine	Val	V	nonpolar
Lysine	Lys	K	positive	Leucine	Leu	L	nonpolar
Histidine	His	H	positive	Isoleucine	Ile	I	nonpolar
Asparagine	Asn	N	uncharged polar	Proline	Pro	P	nonpolar
Glutamine	Gln	Q	uncharged polar	Phenylalanine	Phe	F	nonpolar
Serine	Ser	S	uncharged polar	Methionine	Met	M	nonpolar
Threonine	Thr	T	uncharged polar	Tryptophan	Trp	W	nonpolar
Tyrosine	Tyr	Y	uncharged polar	Cysteine	Cys	C	nonpolar
POLAR AMINO ACIDS				NONPOLAR AMINO ACIDS			

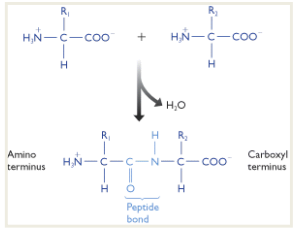
CHGR VANDERBILT UNIVERSITY WILLIAM S. BUSH PHD MS

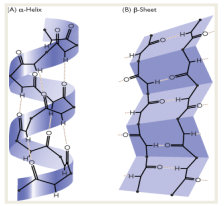
Alberts et al. 2002

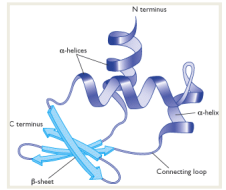
Amino Acids Peptides Proteins

Brown 2002







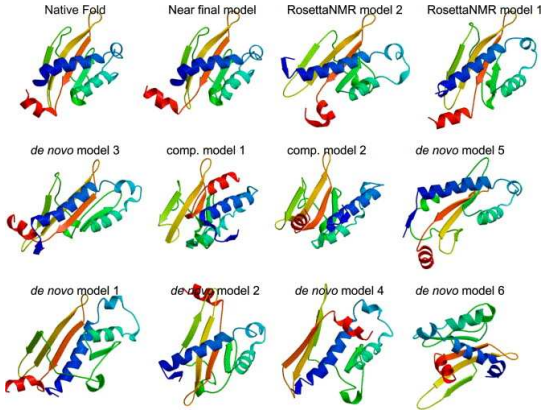


CHGR VANDERBILT UNIVERSITY WILLIAM S. BUSH PHD MS

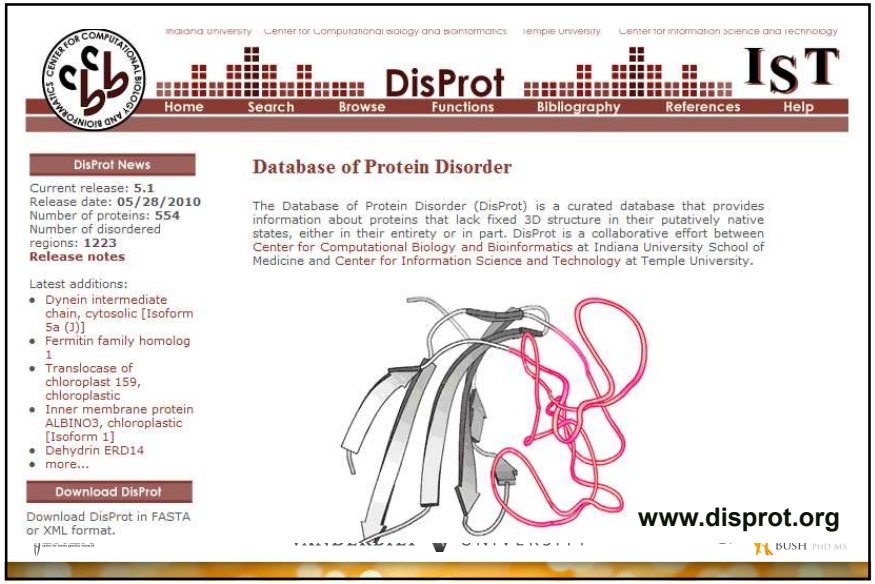
Brown 2002

Protein Structural Prediction

Meiler and Baker 2005



CHGR VANDERBILT UNIVERSITY WILLIAM S. BUSH PHD MS

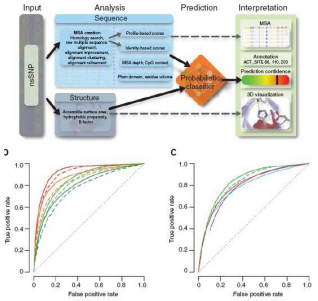


Impact of Protein Changes

Adzhubei et al, Ng and Henikoff.

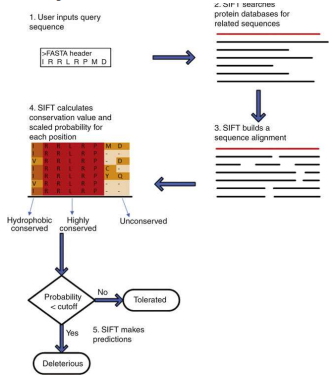
A method and server for predicting damaging missense mutations

Ivan A Adzhubei[1,7], Steffen Schmidt[2,7], Leonid Peshkin[3,7], Vasily E Ramensky[4], Anna Gerasimova[5], Peer Bork[6], Alexey S Kondrashov[5] & Shamil R Sunyaev[1]



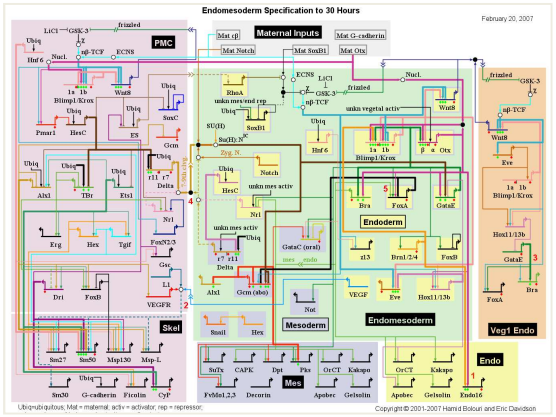
Predicting Deleterious Amino Acid Substitutions

Pauline C. Ng[1,2] and Steven Henikoff[1,3,4]

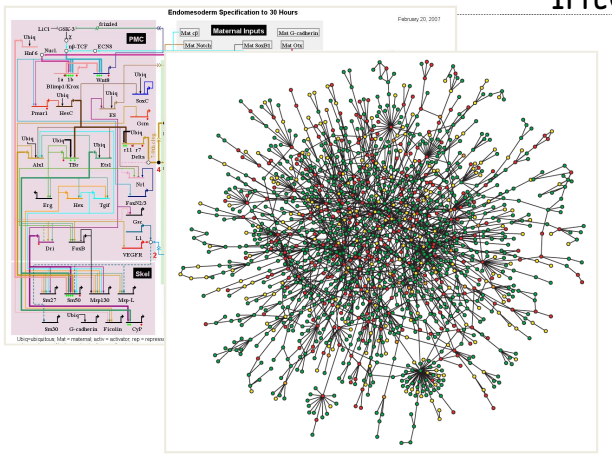


BIOMOLECULAR INTERACTIONS DRIVE BIOLOGY

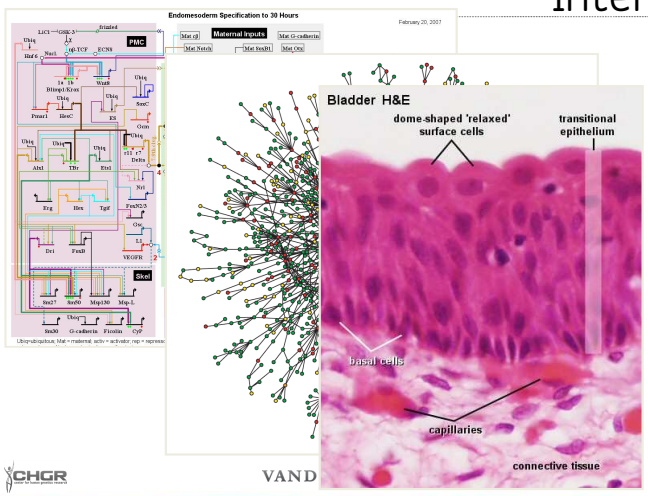
Interactome



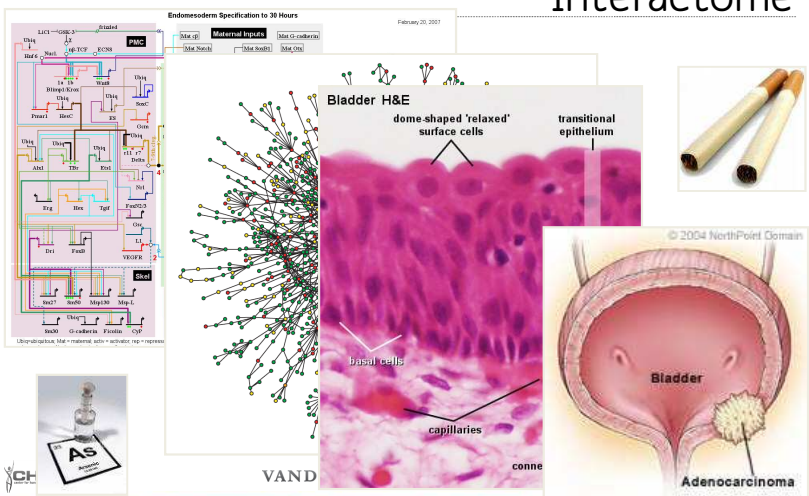
Interactome



Interactome



Interactome



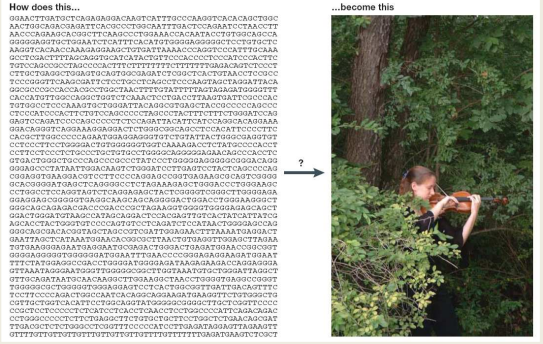
Some Big Questions

Genetics and Life

THE PHENOGENETIC LOGIC OF LIFE

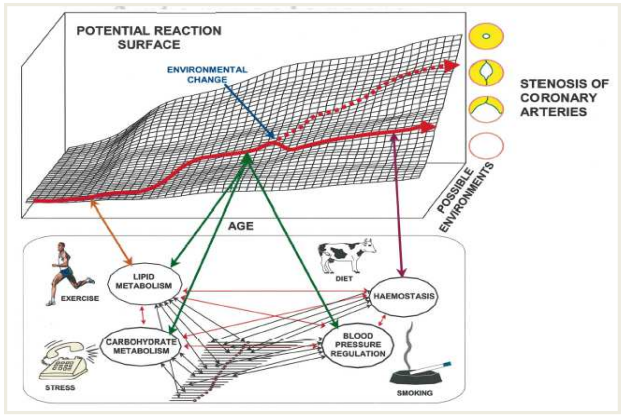
Kenneth M. Weiss

NATURE REVIEWS | GENETICS VOLUME 6 | JANUARY 2005 | 37



Genetic Architecture

Sing et al., *Arter. Thromb. Vasc. Biol.* (2003)



The Tree of Life

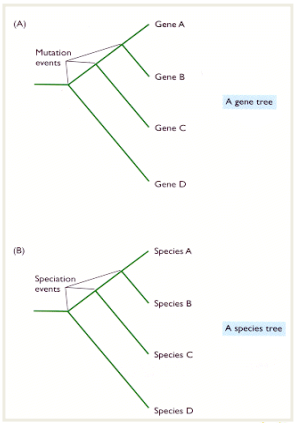
Molecular Phylogenetics

Multiple alignment

```
1  AGGCCAAGCCATAGCTGTCC
2  AGGCAAAGACATACCTGACC
3  AGGCCAAGACATAGCTGTCC
4  AGGCAAAGACATACCTGTCC
```

Distance matrix

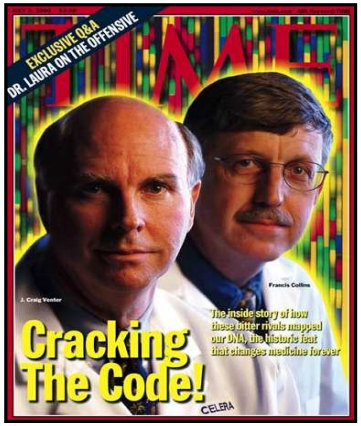
	1	2	3	4
1	–	0.20	0.05	0.15
2		–	0.15	0.05
3			–	0.10
4				–

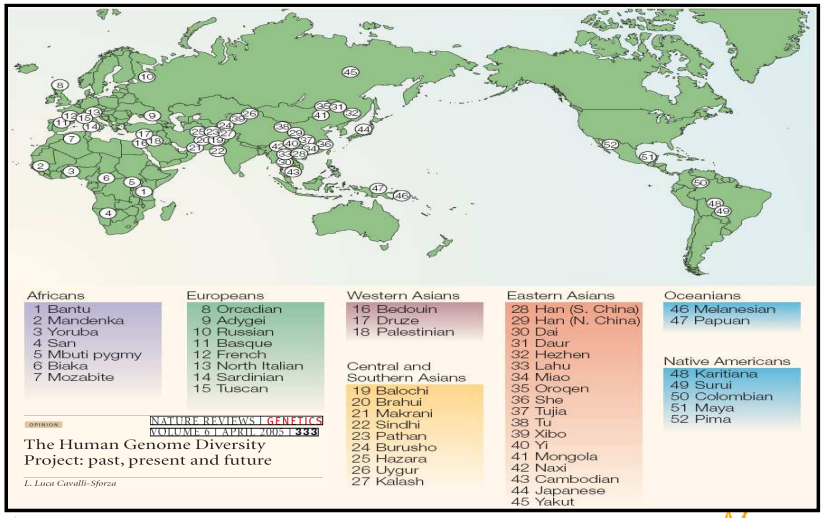


Brown 2002

MEASURING DNA

The Human Genome Project





OPINION
NATURE REVIEWS | GENETICS VOLUME 6 | APRIL 2005 | 333

The Human Genome Diversity Project: past, present and future

L. Luca Cavalli-Sforza

Single Nucleotide Polymorphisms

Subject #1	**Subject #2**	**Subject #3**
-- AG**G**TCA -- -- AG**G**TCA --	-- AG**G**TCA -- -- AG**C**TCA --	-- AG**C**TCA -- -- AG**C**TCA --

Two ***alleles*** (**G** and **C**)

Three ***genotypes*** (**GG**, **GC**, **CC**)

Single Nucleotide Polymorphisms

- ~ 1 SNP every 100 bp
- ~ 30 million SNPs
- ~500,000 SNPs in coding DNA
  - Synonymous (silent)
  - Nonsynonymous
    - Deleterious effect
    - Beneficial effect
    - No effect

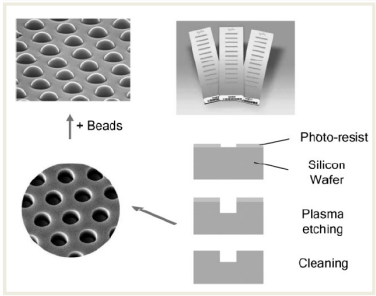
Capturing SNPs

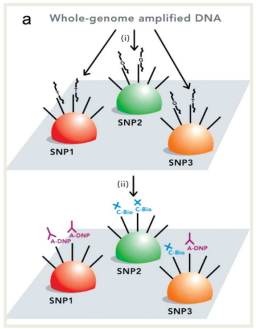
Whole genome genotyping technologies on the BeadArray™ platform

Frank J. Steemers and Kevin L. Gunderson

Illumina, Inc., San Diego, CA, USA

Biotechnol. J. 2007, 2, 41–49





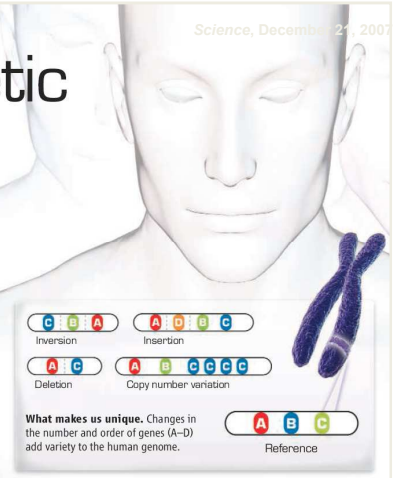
BREAKTHROUGH OF THE YEAR

Human Genetic Variation

Equipped with faster, cheaper technologies for sequencing DNA and assessing variation in genomes on scales ranging from one to millions of bases, researchers are finding out how truly different we are from one another

THE UNVEILING OF THE HUMAN GENOME ALMOST 7 YEARS AGO cast the first faint light on our complete genetic makeup. Since then, each new genome sequenced and each new individual studied has illuminated our genomic landscape in ever more detail. In 2007, researchers came to appreciate the extent to which our genomes differ from person to person and the implications of this variation for deciphering the genetics of complex diseases and personal traits.

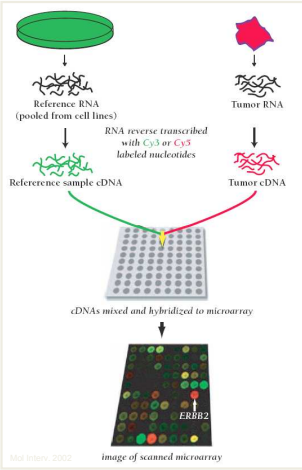
Less than a year ago, the big news was triangulating variation between us and our primate cousins to get a better handle on genetic changes along the evolutionary tree that led to humans. Now, we have moved from asking what in our DNA makes us human to striving to know what in my DNA makes me me.



What makes us unique. Changes in the number and order of genes (A–D) add variety to the human genome.

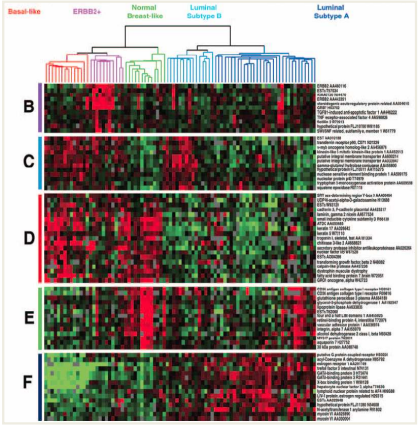
MEASURING RNA

cDNA Microarrays



Heatmaps

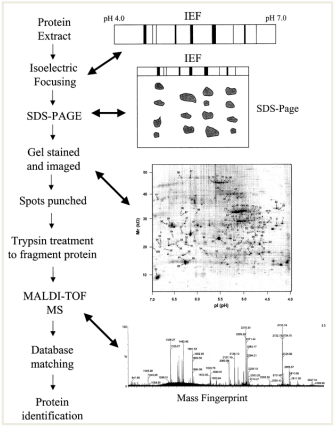
Mol Interv. 2002



MEASURING PROTEINS

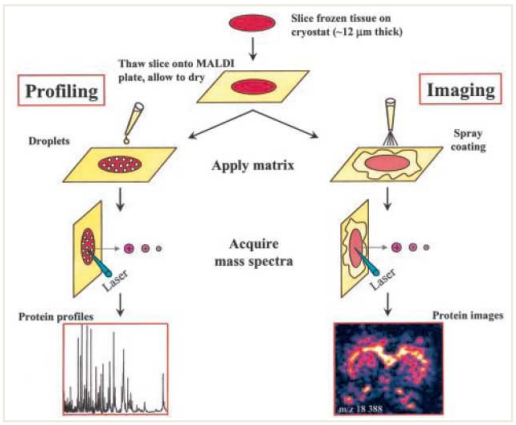
2D Gels and Mass Spectrometry

Metabolic Engineering 4, 98–106 (2002)

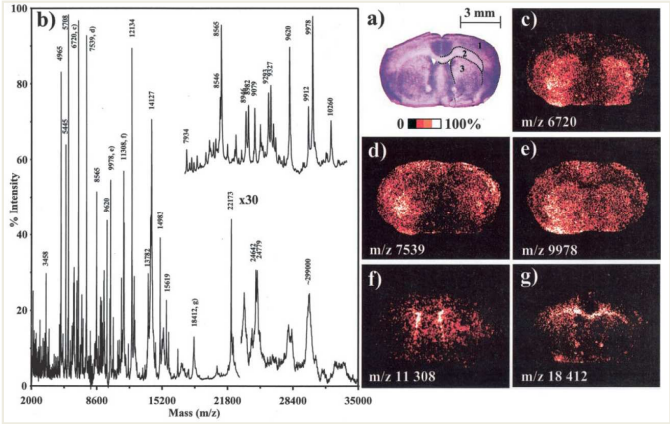


Protein Profiling in Tissues

Am J Pathol. 2004

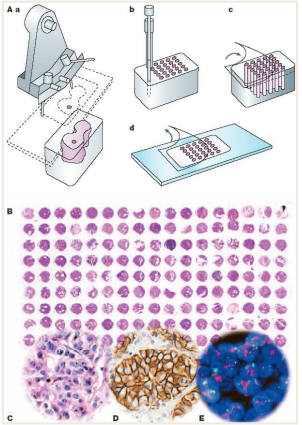


Protein Profiling in Tissues



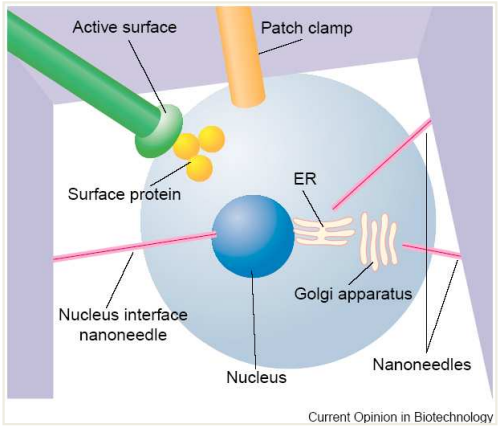
Tissue Microarrays

Nat Rev Drug Discov. 2003



EMERGING TECHNOLOGIES

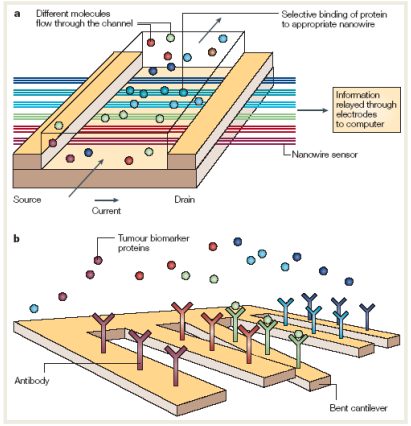
Nanotechnology



Current Opinion in Biotechnology

Nanotechnology

Nature Reviews Cancer 2005



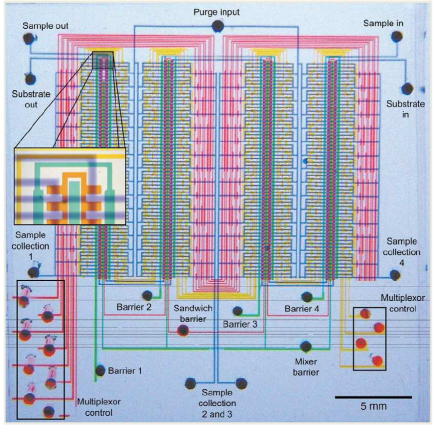
CHGR

VANDERBILT UNIVERSITY

WILLIAM S. BUSH PHD MS

Lab-on-a-Chip

Current Opinion in Structural Biology 2003



CHGR

VANDERBILT UNIVERSITY

WILLIAM S. BUSH PHD MS

The $1000 Genome

J. Craig Venter INSTITUTE

Home | Research | DNA Sequencing | Conferences | Education | Press | Jobs | Contact

Press Release

FOR IMMEDIATE RELEASE

J. CRAIG VENTER SCIENCE FOUNDATION ANNOUNCES $500,000 TECHNOLOGY PRIZE FOR ADVANCES LEADING TO THE $1,000 HUMAN GENOME

ROCKVILLE, MD (September 23, 2003). The J. Craig Venter Science Foundation announced today a $500,000 Genomic Technology Prize. The prize, to be awarded one time only, is aimed at stimulating the scientific and technology research community to significantly advance automated DNA sequencing so that a human genome can be sequenced for $1,000 or less as soon as possible. The prize was announced during New Frontiers in Sequencing Technology session at the 15th annual Genome Sequencing and Analysis Conference (GSAC) in Savannah, Georgia.

100 Genomes in 10 Days

ARCHON GENOMICS X PRIZE

DONATE | X PRIZE FOUNDATION

ARCHON X PRIZE FOR GENOMICS | TEAMS | NEWS & EVENTS | TAKE ACTION | DISCOVER | ABOUT

Revolution Through Competition. ▸ TAKE ACTION

ARCHON X PRIZE FOR GENOMICS

- Introduction
- Why Genomics?
- Prize Overview
- The Promise of Personalized Medicine
- Frequently Asked Questions

"The X PRIZE Foundation is taking a bold and important step — incentivizing innovation in technology that will alleviate a bottleneck in genomic sciences. We, people who suffer from genetic conditions, will benefit so much from

PRIZE OVERVIEW

A $10 MILLION PRIZE FOR THE FIRST TEAM TO SUCCESSFULLY SEQUENCE 100 HUMAN GENOMES IN 10 DAYS

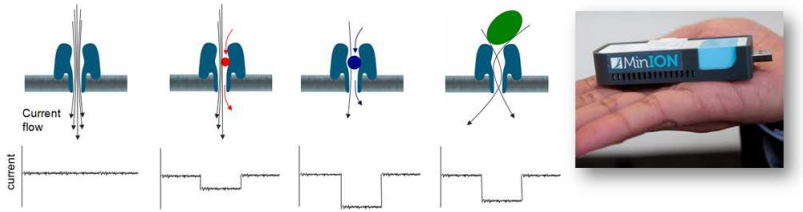
What Inspired this X PRIZE?

In 2000, Dr. J. Craig Venter led the first private team to successfully sequence a complete human genome. In the preceding decades combined governmental and private funding efforts spent $100s of millions to develop the instrumentation required. It took the Venter team $100 million and nine months to achieve their historic accomplishment

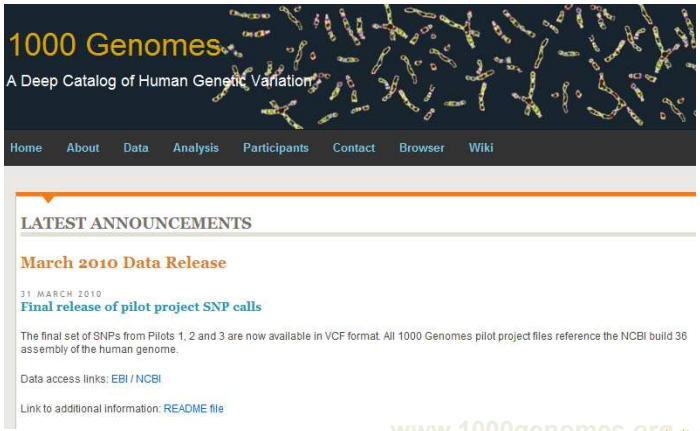
The J. Craig Venter Science Foundation offered the $500,000 Innovation in Genomics Science and Technology Prize in September 2003 aimed at stimulating development of less expensive and faster sequencing technology. To attract even more resources to this exceptionally worthy goal, Dr. Venter joined forces with the X PRIZE Foundation, wrapping his competition and prize purse into the Archon X PRIZE for Genomics.

Nanopore Technology

http://www.nanoporetech.com



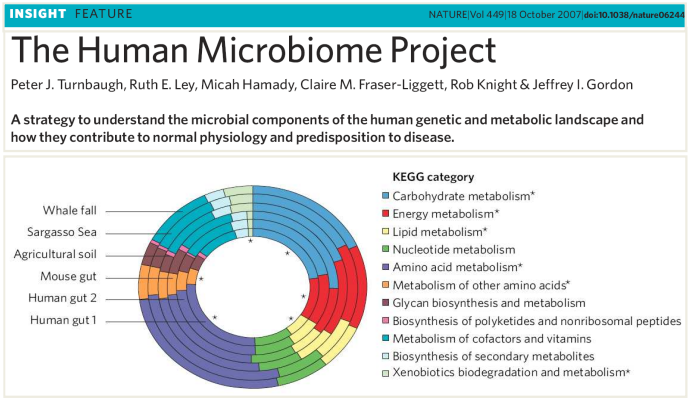
1000 Genomes Project



www.1000genomes.org

Our other genomes

http://nihroadmap.nih.gov/hmp/



INSIGHT FEATURE NATURE|Vol 449|18 October 2007|doi:10.1038/nature06244

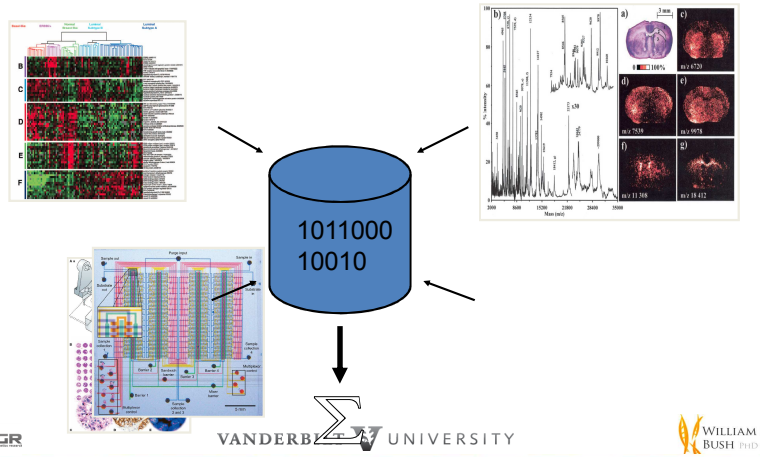
The Human Microbiome Project

Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight & Jeffrey I. Gordon

A strategy to understand the microbial components of the human genetic and metabolic landscape and how they contribute to normal physiology and predisposition to disease.

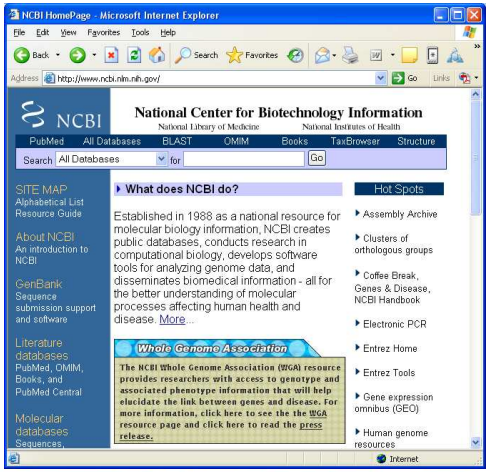
DATABASES

Bioinformatics: Databases

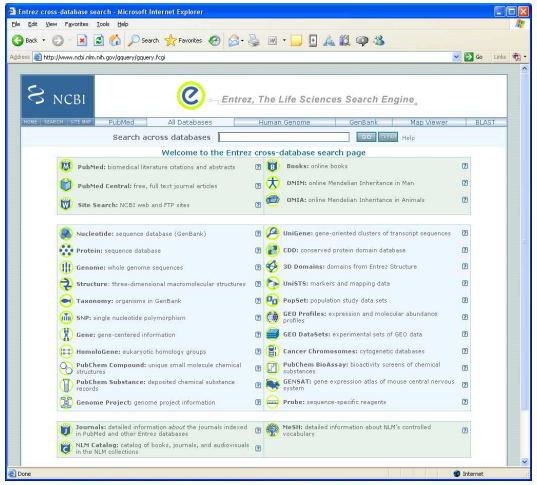


VANDERBILT UNIVERSITY

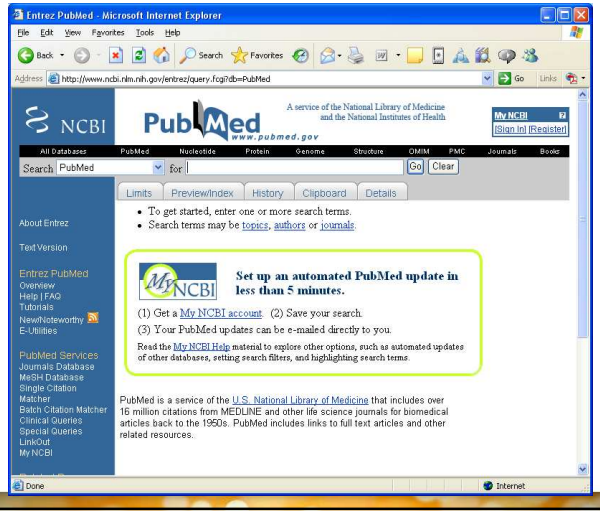
http://www.ncbi.nlm.nih.gov



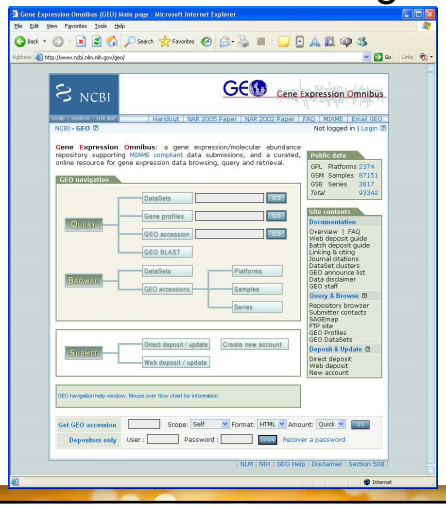
http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi



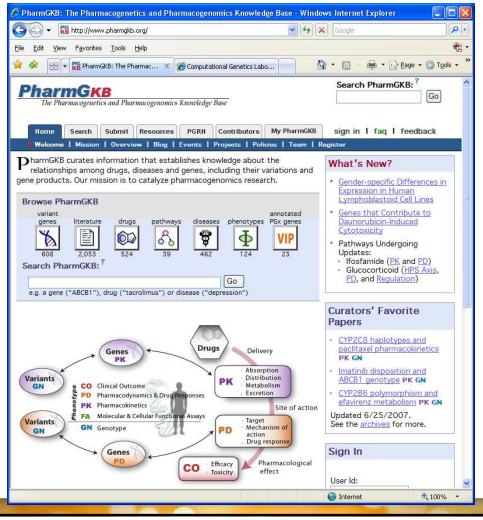
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMe



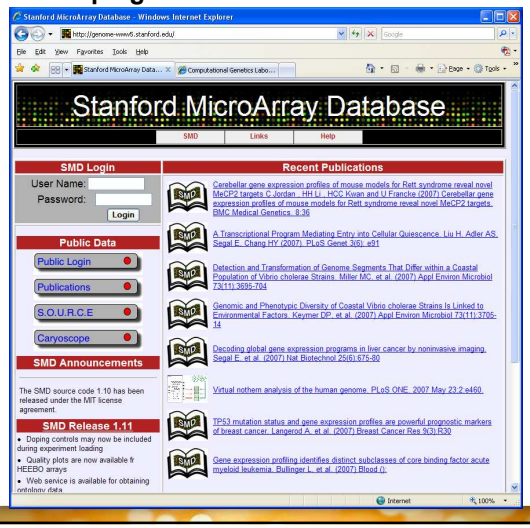
http://www.ncbi.nlm.nih.gov/geo/



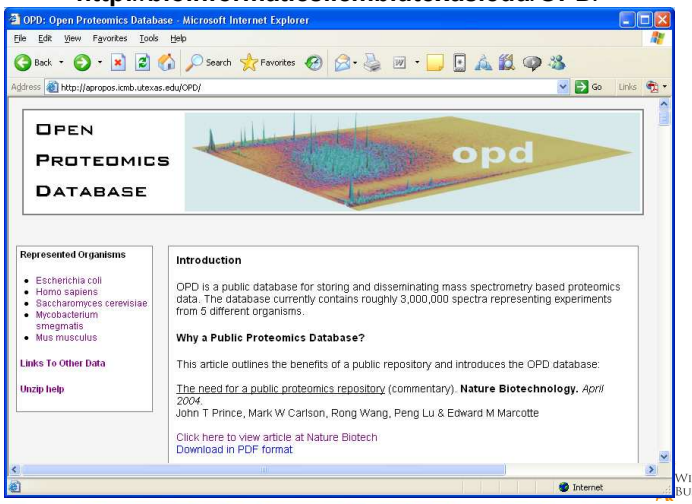
www.pharmgkb.org



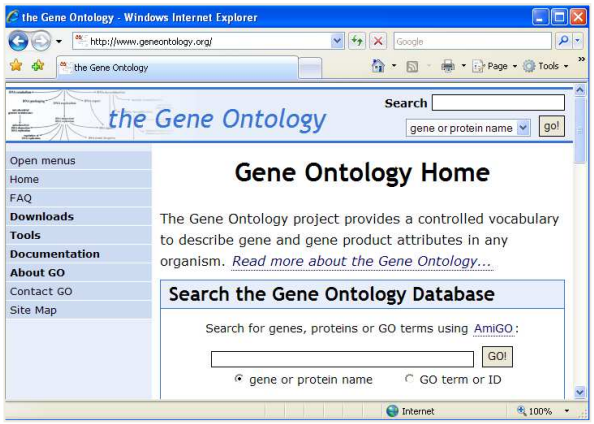
http://genome-www5.stanford.edu/



http://bioinformatics.icmb.utexas.edu/OPD/



http://www.geneontology.org

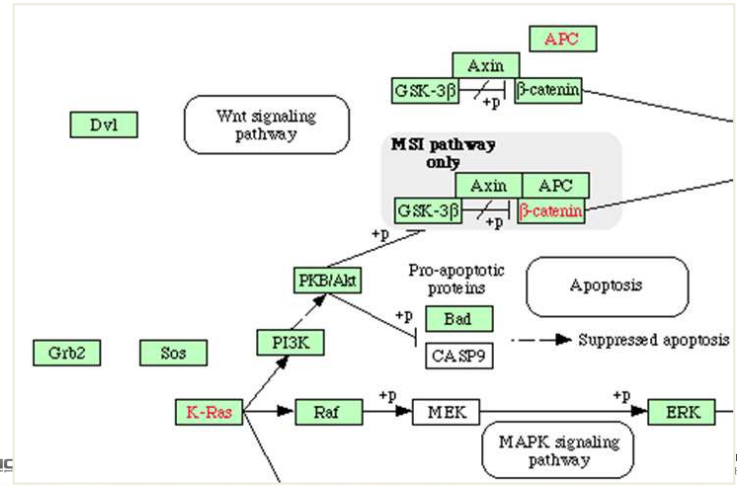


http://www.geneontology.org

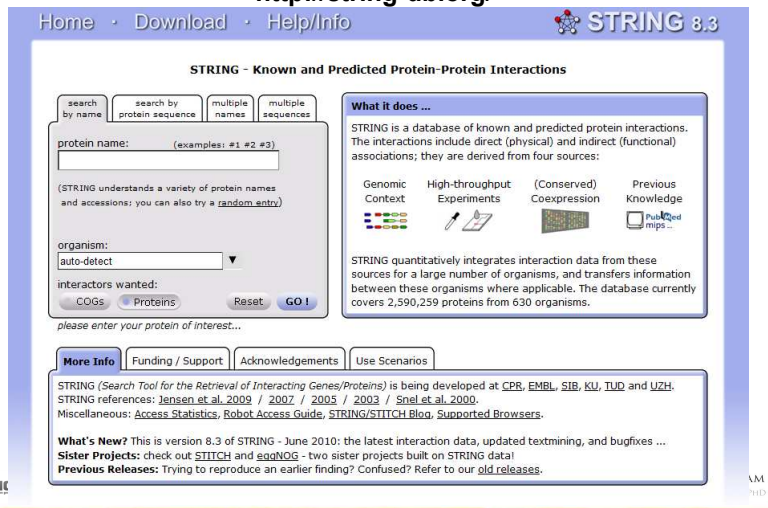
- all : all [218539]
  - GO:0005575 : cellular_component [152301]
    - GO:0005623 : cell [107815]
      - GO:0044464 : cell part [107776]
        - GO:0005622 : intracellular [86035]
          - GO:0044424 : intracellular part [85270]
            - GO:0005737 : cytoplasm [72113]
              - GO:0044444 : cytoplasmic part [66867]
                - **GO:0005794 : Golgi apparatus [1618]**

http://www.genome.jp/kegg/

http://www.genome.jp/kegg/



http://string-db.org/

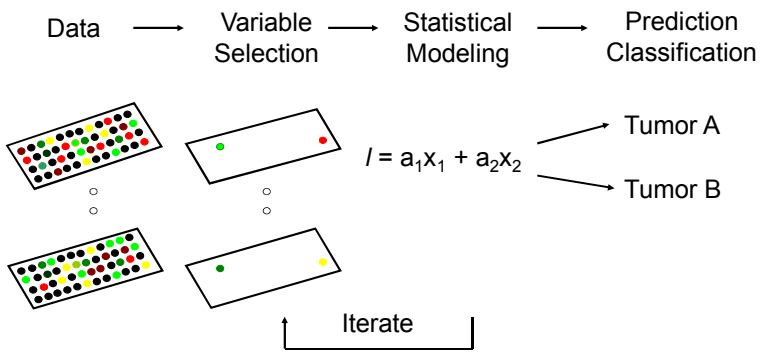


Analysis

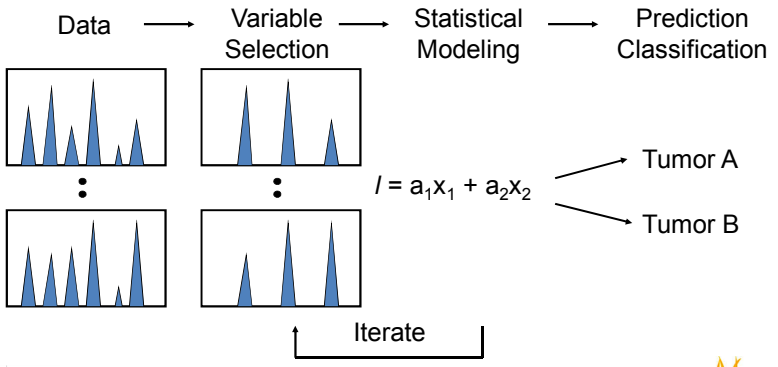
Mining Biomolecular Patterns

- Can we classify and/or predict biological and clinical endpoints using genetic, genomic, and/or proteomic data?
- Which biomolecules are important?
- What is their pattern or statistical relationship?

Objectives

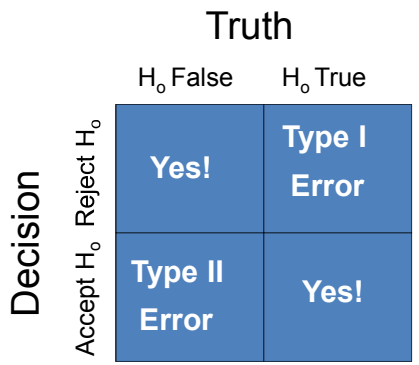


Objectives



Hypothesis Testing

"The truth is out there"



Type I error
Type II error
Type III error

Permutation Testing

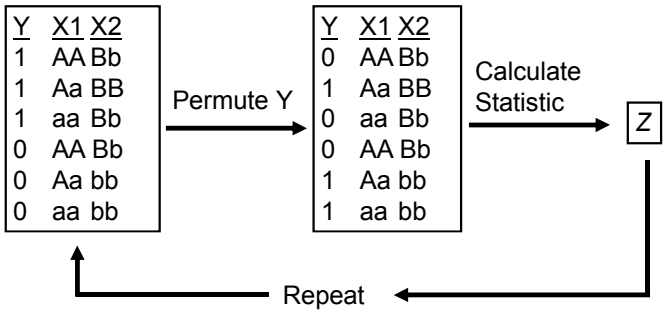
P. Good, Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses (2000)

- Many data-driven methods are nonparametric and model-free.
- Permutation testing can be used to assess statistical significance to allow formal hypothesis testing.
- Basic Idea: Randomize data so it is consistent with null hypothesis.

Permutation Testing

P. Good, Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses (2000)

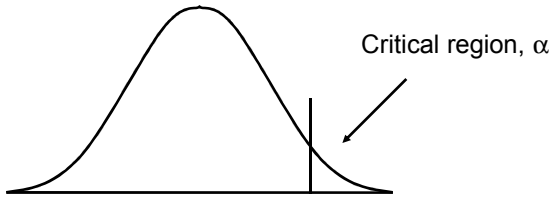
Example



Permutation Testing

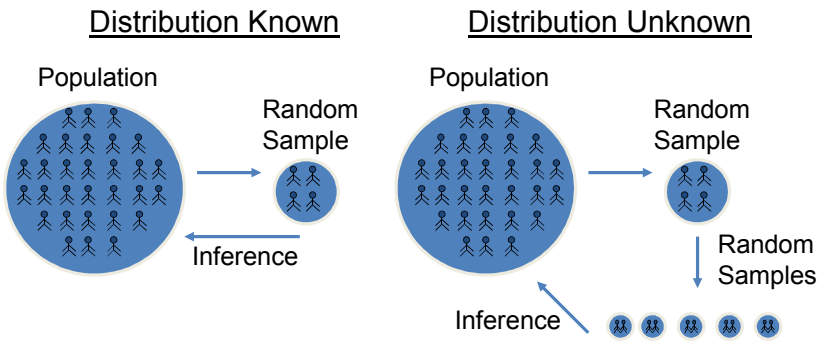
P. Good, Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses (2000)

Distribution of Statistic under the Null Hypothesis from Many Permutations



Bootstrapping

B. Efron, *Annals of Statistics* 7.1-26 (1979)
AC Davidson and DV Hinkley, *Bootstrap Methods and their Application* (1997)
CE Lunneborg, *Data Analysis by Resampling: Concepts and Applications* (2000)



Health is a Complex System

Sing et al., *Arter. Thromb. Vasc. Biol.* (2003)

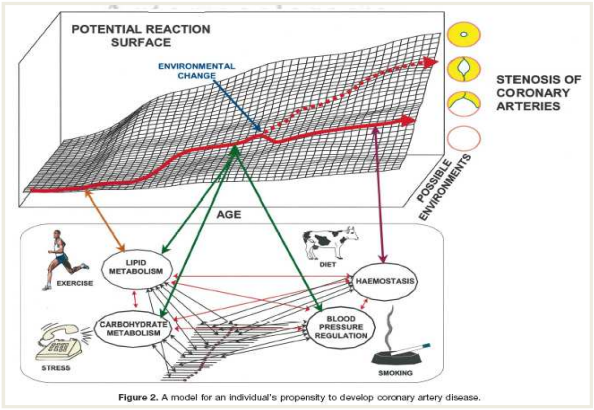


Figure 2. A model for an individual's propensity to develop coronary artery disease.

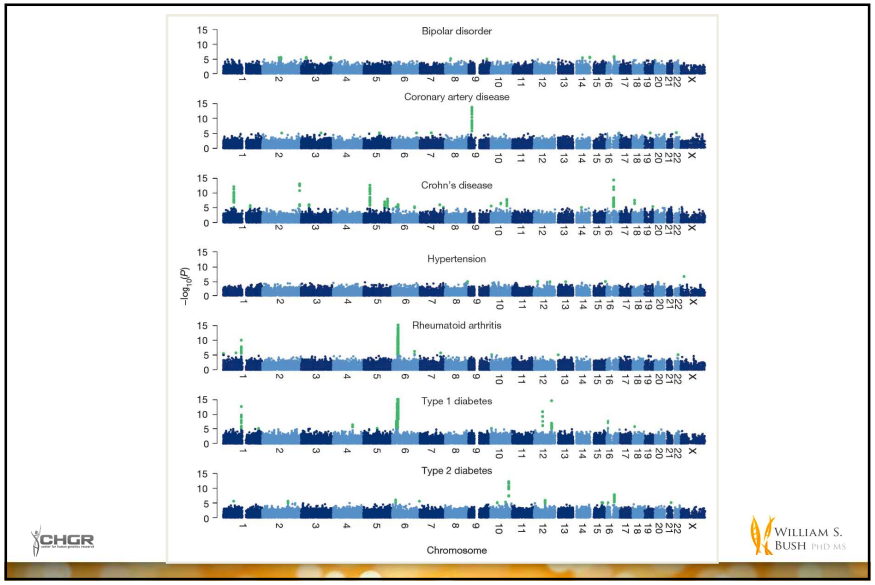
Vol 447|7 June 2007|doi:10.1038/nature05911 nature

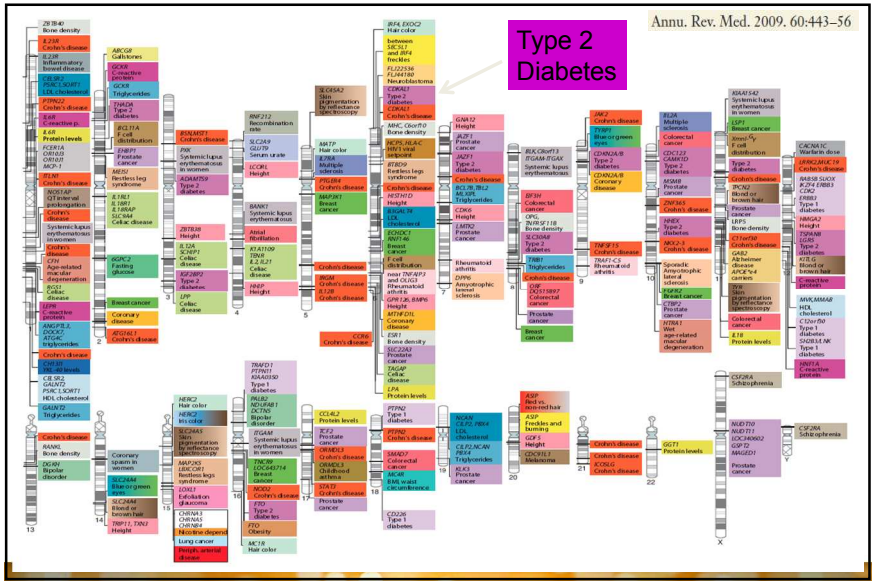
ARTICLES

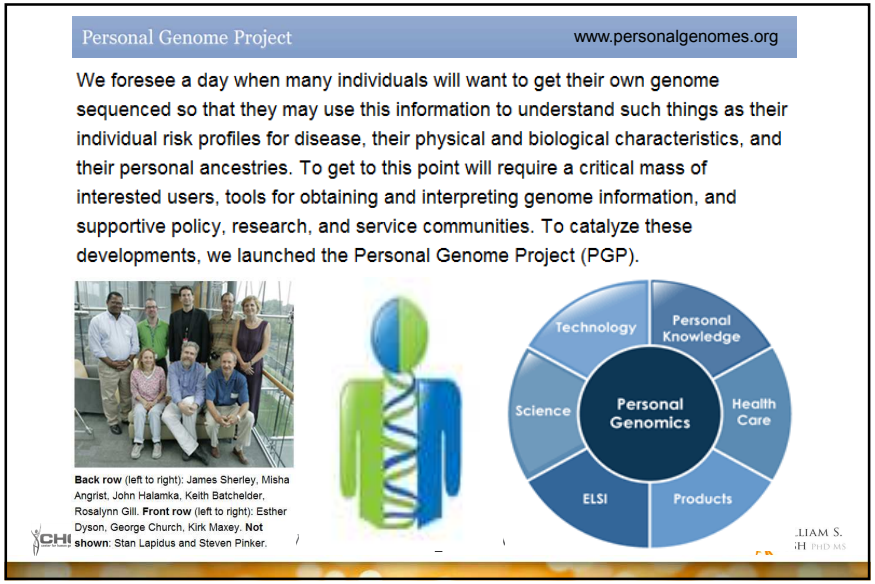
Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

- 50 Research Groups
- 14,000 cases and 3,000 shared controls
- 500,000 SNPs
- Seven complex human diseases:
  - bipolar disorder (BD)
  - coronary artery disease (CAD)
  - Crohn's disease (CD)
  - hypertension (HT)
  - rheumatoid arthritis (RA)
  - type 1 diabetes (T1D)
  - type 2 diabetes (T2D)


Bipolar disorder
Coronary artery disease
Crohn's disease
Hypertension
Rheumatoid arthritis
Type 1 diabetes
Type 2 diabetes
$-\log_{10}(P)$
Chromosome
CHGR
WILLIAM S. BUSH PHD MS


Type 2 Diabetes
Annu. Rev. Med. 2009. 60:443–56

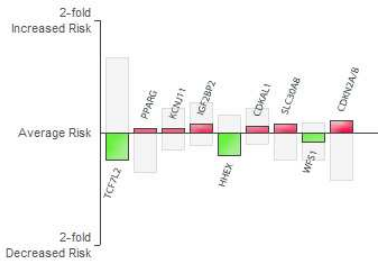

Personal Genome Project
www.personalgenomes.org
We foresee a day when many individuals will want to get their own genome sequenced so that they may use this information to understand such things as their individual risk profiles for disease, their physical and biological characteristics, and their personal ancestries. To get to this point will require a critical mass of interested users, tools for obtaining and interpreting genome information, and supportive policy, research, and service communities. To catalyze these developments, we launched the Personal Genome Project (PGP).
Back row (left to right): James Sherley, Misha Angrist, John Halamka, Keith Batchelder, Rosalynn Gill. Front row (left to right): Esther Dyson, George Church, Kirk Maxey. Not shown: Stan Lapidus and Steven Pinker.
Technology
Personal Knowledge
Science
Personal Genomics
Health Care
ELSI
Products


Personal Genome Project
www.personalgenomes.org
Public Profile
Vital Signs
Traits
Medications, Vitamins, Supplements
Data Sets
Tissues
CHGR
VANDERBILT UNIVERSITY
WILLIAM S. BUSH PHD MS

Type 2 Diabetes

Clinical Report on 9 reported markers, updated February 2nd, 2009.

26 % Attributable to Genetics



Jason Moore

20.1 out of 100

people of European ethnicity who share Jason Moore's genotype will get Type 2 Diabetes between the ages of 20 and 79.

Average

21.9 out of 100

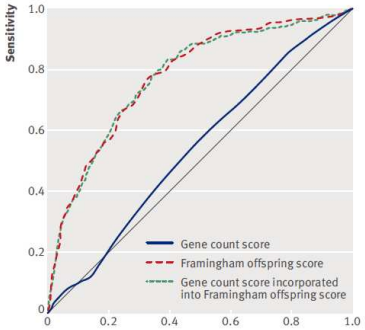
people of European ethnicity will get Type 2 Diabetes between the ages of 20 and 79.

CHGR VANDERBILT UNIVERSITY WILLIAM S. BUSH PHD MS

BMJ 2010 **RESEARCH**

Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study

Philippa J Talmud, professor of cardiovascular genetics,[1] Aroon D Hingorani, professor of genetic epidemiology,[1,2] Jackie A Cooper, biostatistician,[1] Michael G Marmot, MRC professor of epidemiology and public health,[2] Eric J Brunner, reader in epidemiology and public health,[2] Meena Kumari, senior research fellow in population health,[2] Mika Kivimäki, professor of social epidemiology,[2] Steve E Humphries, BHF professor of cardiovascular genetics[1]



CHGR WILLIAM S. BUSH PHD MS

A Genetics Company Fails, Its Research Too Complex

By NICHOLAS WADE
Published: November 17, 2009

The New York Times

DeCode Genetics, a pioneering company that used the Icelandic population as its guinea pigs in detecting disease-causing mutations, filed for bankruptcy on Tuesday.

SIGN IN TO RECOMMEND
TWITTER
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

CYRUS JULY 9

Enlarge This Image

deCode

A lab at deCode Genetics, which found that the genetic nature of human disease was far more complex than anyone thought.

The company's demise suggests that the medical promise of the human genome may take much longer to be fulfilled than its sponsors had hoped. Based in Reykjavik, Iceland, it was founded in 1996 by Dr. Kari Stefansson, a research neurologist who worked at the University of Chicago and at Harvard. After the human genome sequence was achieved in 2003, Dr. Stefansson quickly realized that Iceland's excellent medical records, combined with the genealogical information available on its close-knit population, provided a fine test bed for seeking the roots of genetically complex diseases like cancer, diabetes and schizophrenia.

Related

Times Topics: deCODE Genetics, Incorporated

CHGR WILLIAM S. BUSH PHD MS

VIEWPOINT

NATURE REVIEWS | GENETICS
VOLUME 11 | JUNE 2010

Missing heritability and strategies for finding the underlying causes of complex disease

Evan E. Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore and Joseph H. Nadeau

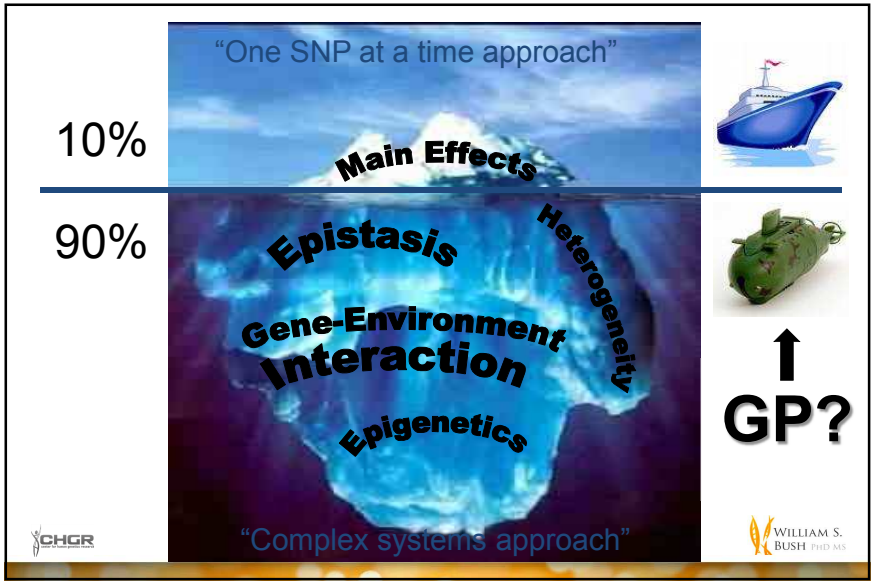
How should we solve the problem of 'missing heritability' in complex diseases?

Jason H. Moore. The case of the missing heritability for common human diseases should not be a mystery to anyone given the inherent complexity of the relationship between genotype and phenotype.

The time is now to philosophically and analytically retool for a complex genetic architecture or we will continue to underdeliver on the promises of human genetics.

Indeed, life, and thus genetics, is complicated[46] and some will soon ask, as seismologists have[47], whether we are trying to predict the unpredictable.

CHGR VANDERBILT UNIVERSITY WILLIAM S. BUSH PHD MS



Challenges

- Complexity – nonlinearity, heterogeneity
- Dimensionality – multiple genetic risk factors
- Scale – millions of attributes

Using Expert Knowledge in GP

- Multi-Objective Fitness (GPTP'06)
  - Fitness = F(Accuracy + Knowledge)
- Recombination (EvoBIO'07)
  - Recombine trees with good building blocks
- Mutation (PRIB'07)
  - Mutate trees with poor building blocks
- Sensible Initialization (CEC'09)
  - Initialize trees with good building blocks

GPTP'07
Does Complexity Matter?

GUIDELINES

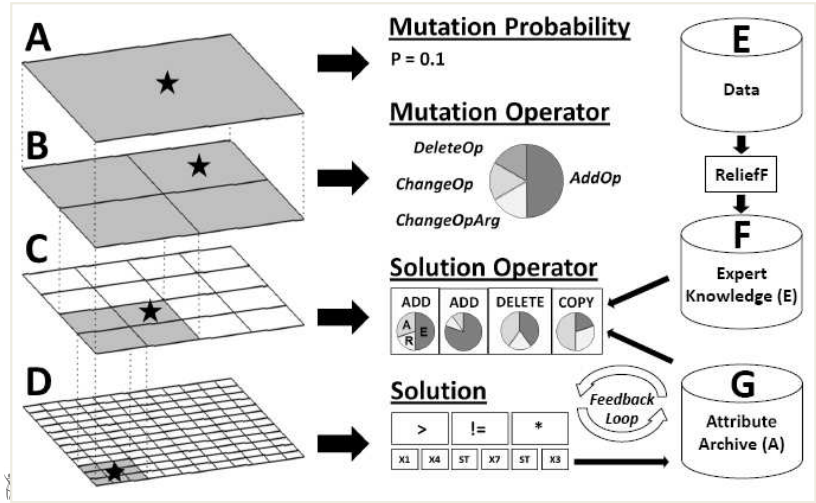
From artificial evolution to computational evolution: a research agenda

Wolfgang Banzhaf, Guillaume Beslon, Steffen Christensen, James A. Foster, François Képès, Virginie Lefort, Julian F. Miller, Miroslav Radman and Jeremy J. Ramsden

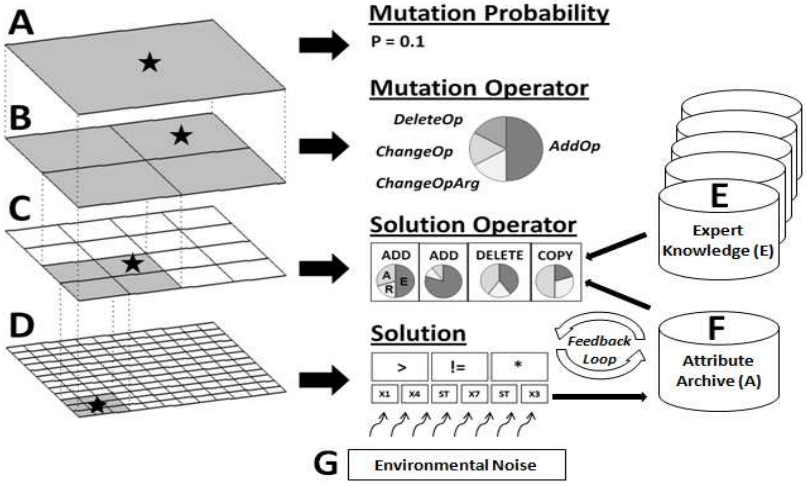
NATURE REVIEWS | GENETICS VOLUME 7 | SEPTEMBER 2006

CHGR VANDERBILT UNIVERSITY WILLIAM S. BUSH PHD MS

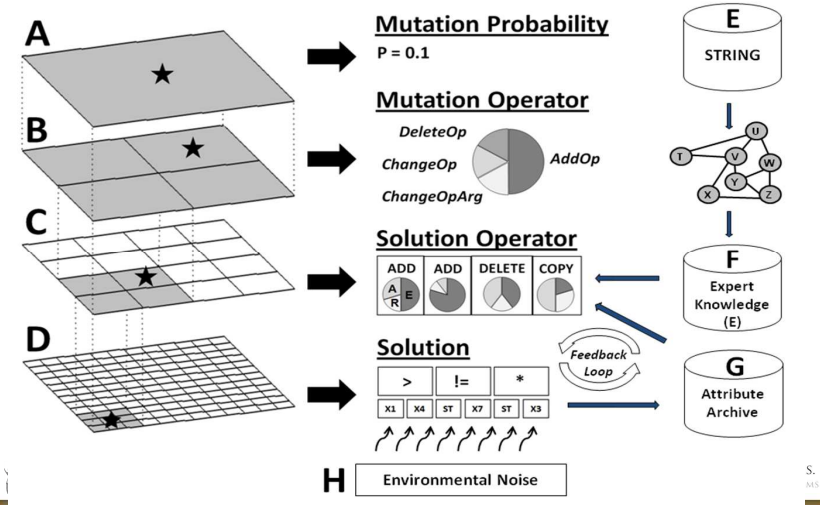
A Computational Evolution System – GPTP'08



A Computational Evolution System – GPTP'09



A Computational Evolution System – GPTP'10



Hum Genet (2008) 124:19–29
DOI 10.1007/s00439-008-0522-8

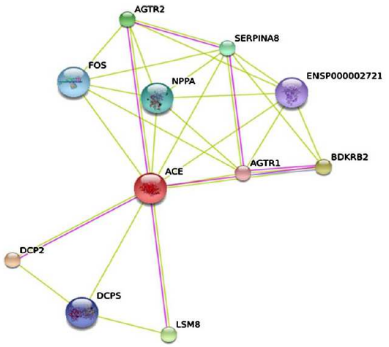
REVIEW

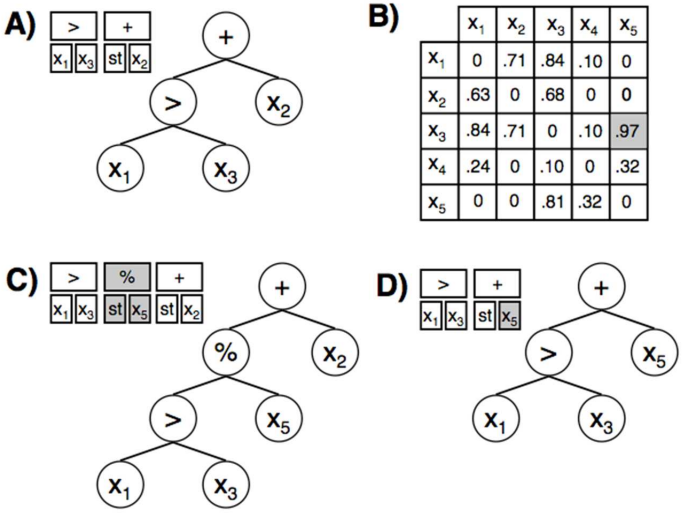
Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases

Kristine A. Pattin · Jason H. Moore

STRING

- 2.5 million proteins
- 630 organisms



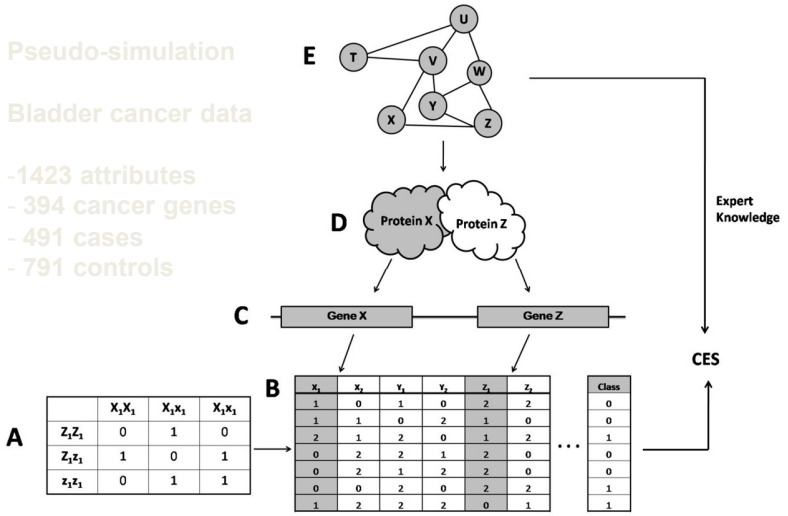


	x_1	x_2	x_3	x_4	x_5
x_1	0	.71	.84	.10	0
x_2	.63	0	.68	0	0
x_3	.84	.71	0	.10	.97
x_4	.24	0	.10	0	.32
x_5	0	0	.81	.32	0

Pseudo-simulation

Bladder cancer data

-1423 attributes
- 394 cancer genes
- 491 cases
- 791 controls



Experimental Design and Analysis

Datasets = 100 per model
- Strong, medium and weak PxP interaction

Runs = 100 per dataset

Generations = 1000

Grid Size = 18x18

Control 1: no expert knowledge

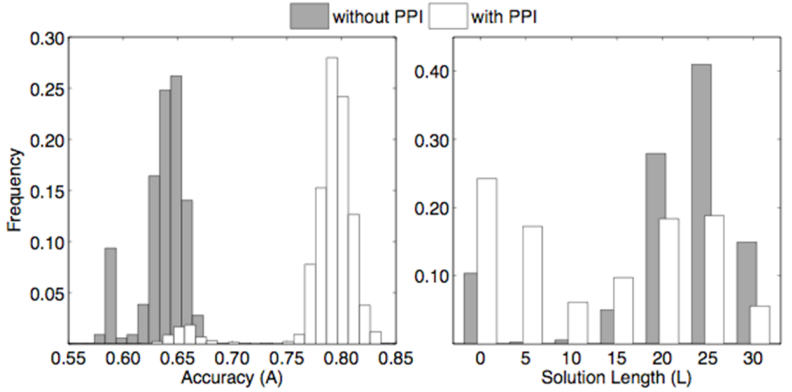
Control 2: attributes with weak PxP interaction

Measured success in finding correct attributes

Results

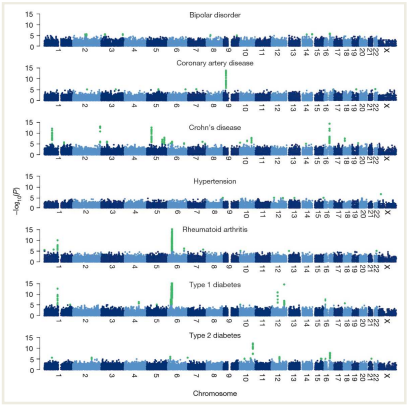
Table 1-1. Percentage of datasets in which CES successfully identified the correct SNP pairings as the most frequent, for the four confidence score scenarios considered.

Confidence Score of PPI	With Expert Knowledge	Control
0.998	100%	0%
0.963	100%	0%
0.933	96%	0%
0.916	80%	0%



Summary

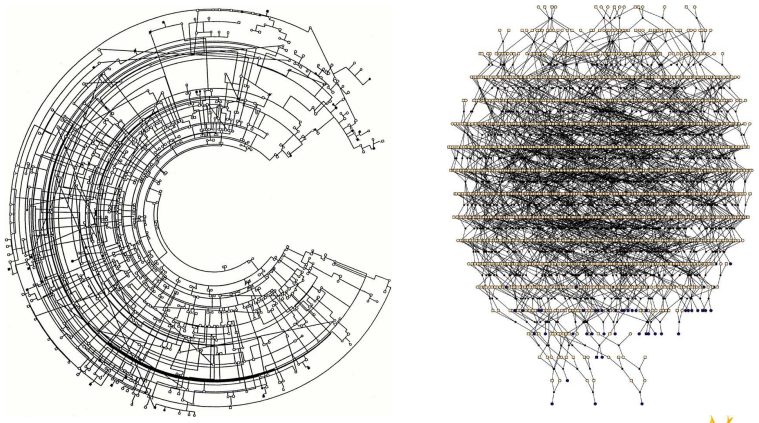
- Our CES is able to learn to exploit protein-protein interactions as a source of domain-specific knowledge.
- Can this scale to 10^6 or more attributes?



PROBLEM REPRESENTATION

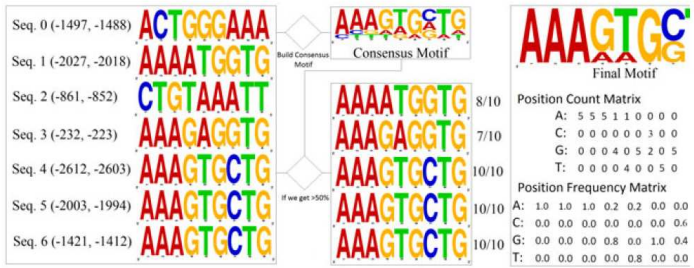
Graph Partitioning

Racette et al. Neurology 2002, Lie et al, 2008



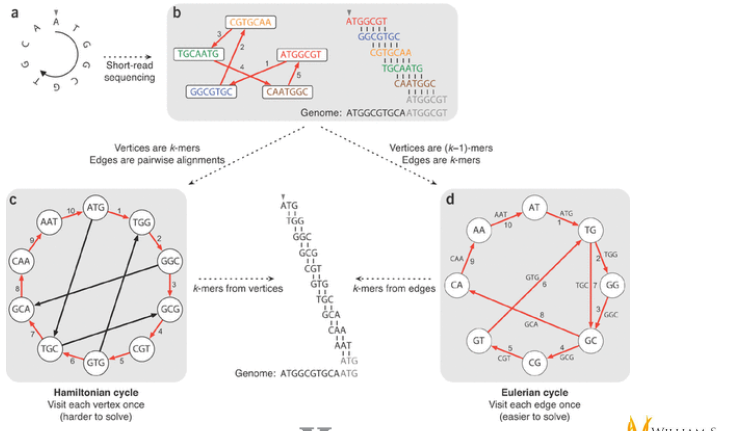
Motif Discovery

Gonzalez-Alvarez et al. EvoBIO 2012



Sequence Alignment

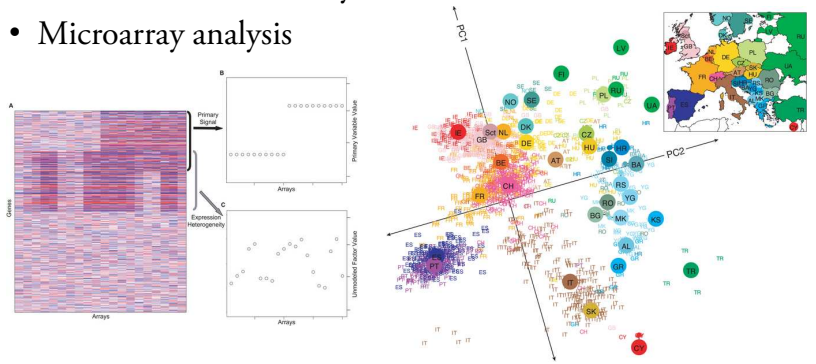
Compeau et al. Nature Biotech 2011



Latent Variable Discovery

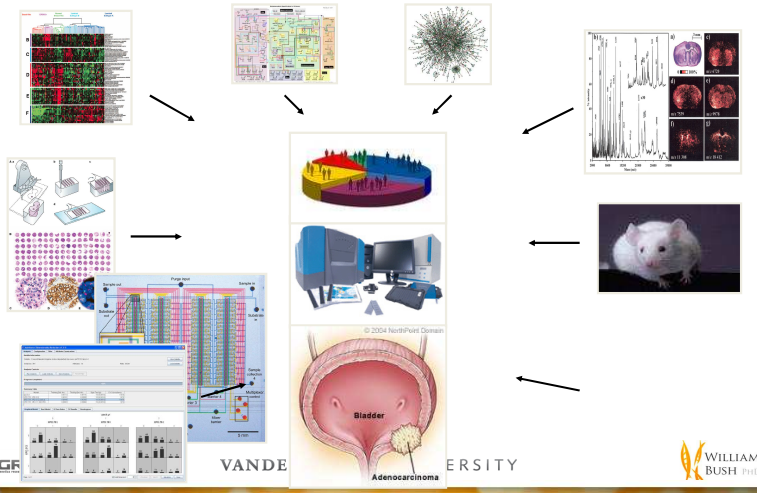
Leek and Storey, PLoS Genetics 2007, Novembre et al, Nature 2008

- PCA for ethnic ancestry
- Microarray analysis

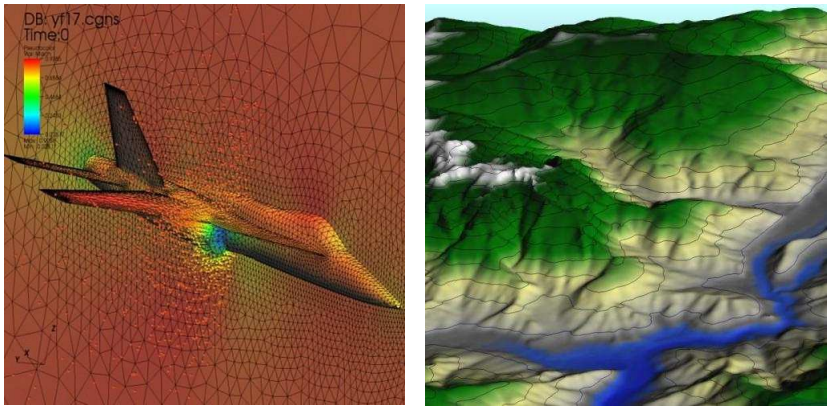


The Future

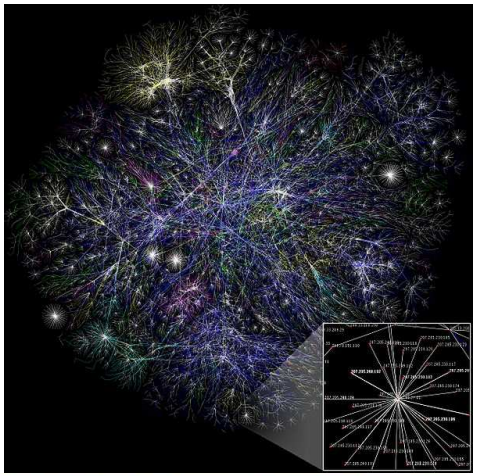
Collaborative Approach Needed



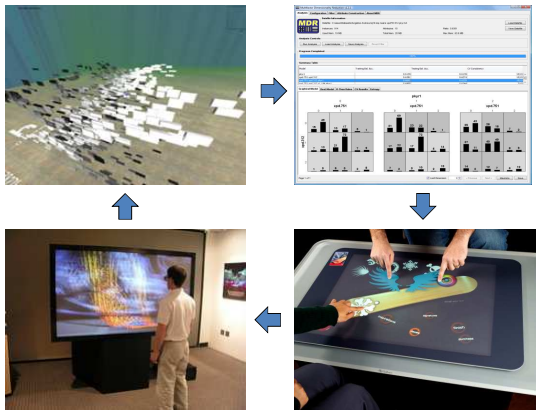
Scientific Visualization



Information Visualization



Visual Analytics



EDUCATION

SPECIAL SECTION

MATHEMATICS IN BIOLOGY

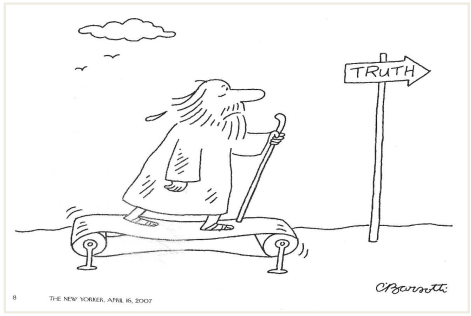
VIEWPOINT

Introductory Science and Mathematics Education for 21st-Century Biologists

William Bialek[1,3] and David Botstein[2,3]*

Galileo wrote that "the book of nature is written in the language of mathematics"; his quantitative approach to understanding the natural world arguably marks the beginning of modern science. Nearly 400 years later, the fragmented teaching of science in our universities still leaves biology outside the quantitative and mathematical culture that has come to define the physical sciences and engineering. This strikes us as particularly inopportune at a time when opportunities for quantitative thinking about biological systems are exploding. We propose that a way out of this dilemma is a unified introductory science curriculum that fully incorporates mathematics and quantitative thinking.

These traditions have resulted in a deep bifurcation in culture and quantitative competence among the scientific disciplines. On one branch are mathematics, the physical sciences, and engineering. Scientists educated along this branch achieve a high level of quantitative expertise: They generally have some mastery over and comfort with not only multivariate calculus and differential equa-



- chgr.mc.vanderbilt.edu/bushlab
- www.gettinggeneticsdone.com
- william.s.bush@vanderbilt.edu