

Introduction to Genomics

Atul Butte, MD

atul_butte@harvard.edu

Children's Hospital Informatics Program

www.chip.org

Children's Hospital • Boston

Harvard Medical School

Massachusetts Institute of Technology

Introduction

- Molecular biology for the bioinformaticist * [Long](#)
- Microarrays [Long Med Short](#)
- Gene measurement * [Long](#)
- Fold-difference calculations [Link](#)
- Measurement noise [Link](#)
- Reproducibility [Long Short](#)
- Using microarrays is not hypothesis-free [Link](#)

Analytic methods

- Multiple-chip analysis methods [Long Med Short](#)
- Relevance Networks * [Link](#)
- Advantages of Relevance Networks [Link](#)
- Model-independence [Long Short](#)
- Causality (real data) [Link](#)

Real data and relevance networks

- Cancer Pharmacogenomics * [Link](#)
- CardioGenomics [Link](#)
- Muscular Dystrophy * [Link](#)
- Laboratory / Phenotypic [Long Short](#)

Bio+medical informatics

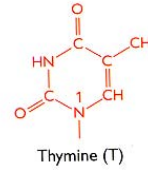
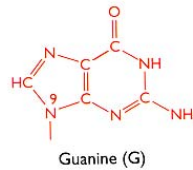
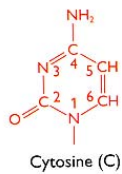
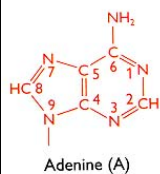
- Data types in bioinformatics [Link](#)
- Parallels between medical and bioinformatics * [Link](#)
- Developing diagnostic tests * [Link](#)

Advanced analysis and future directions

- Differential analysis (real data) [Link](#)
- Publicly available tools [Link](#)
- Web-based microarray tools * [Link](#)
- Linking results to findings with Unchip [Link](#)
- PGA Multi-center integration [Link](#)
- Visualization * [Link](#)
- How this will change medicine * [Link](#)
- Conclusion and our team [Link](#)

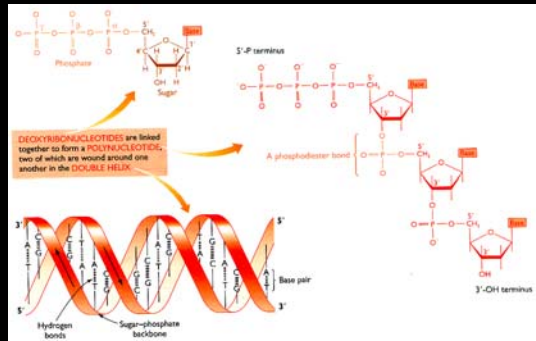
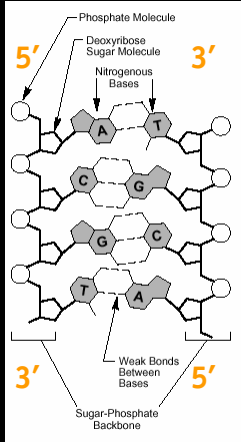
Basic Biology

- Organisms need to produce proteins for a variety of functions over a lifetime
 - Enzymes to catalyze reactions
 - Structural support
 - Hormone to signal other parts of the organism
- Problem one: how to encode the instructions for making a specific protein
- Step one: nucleotides



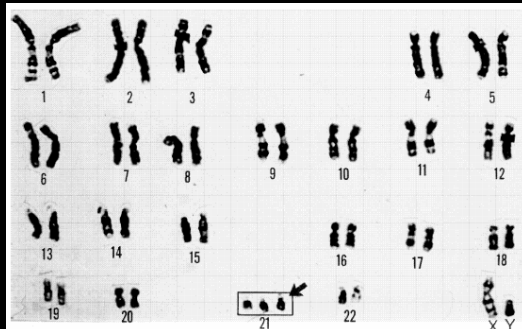
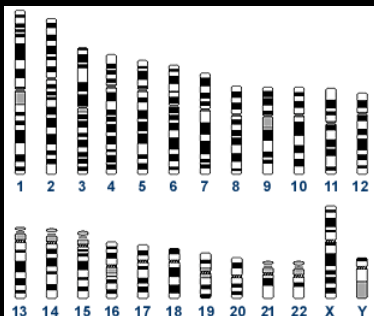
Basic Biology

- Complementary nucleotides form base pairs
- Base pairs are put together in chains (strands)
 - Naturally form double helixes
 - Redundant information in each strand



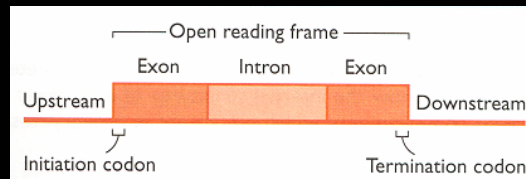
Chromosomes

- We do not know exactly how strands of DNA wind up to make a chromosome
- Each chromosome has a single double-strand of DNA
- 22 human chromosomes are paired
- In human females, there are two X chromosomes
- In males, one X and one Y



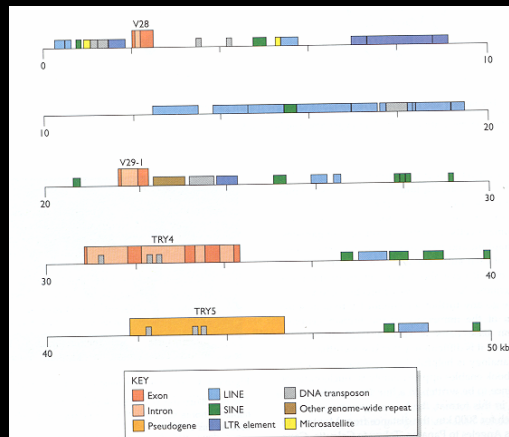
What does a gene look like?

- Each gene encodes instructions to make a single protein
- DNA before a gene is called upstream, and can contain regulatory elements
- Introns may be within the code for the protein
- There is a code for the start and end of the protein coding portion
- Theoretically, the biological system can determine promoter regions and intron-exon boundaries using the sequence syntax alone



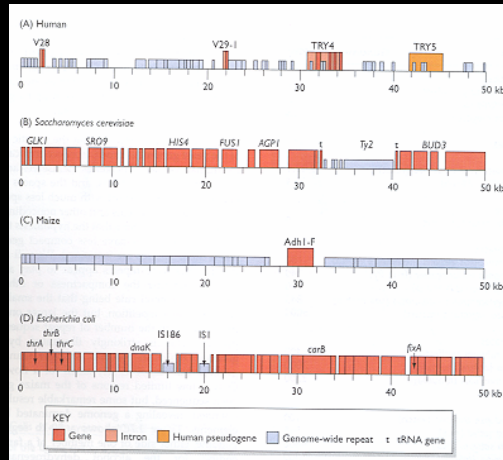
Area between genes

- The human genome contains 3 billion base pairs (3000 Mb) but only 35 thousand genes
- The coding region is 90 Mb (only 3% of the genome)
- Over 50% of the genome is repeated sequences
 - Long interspersed nuclear elements
 - Short interspersed nuclear elements
 - Long terminal repeats
 - Microsatellites
- Many repeated sequences are different between individuals



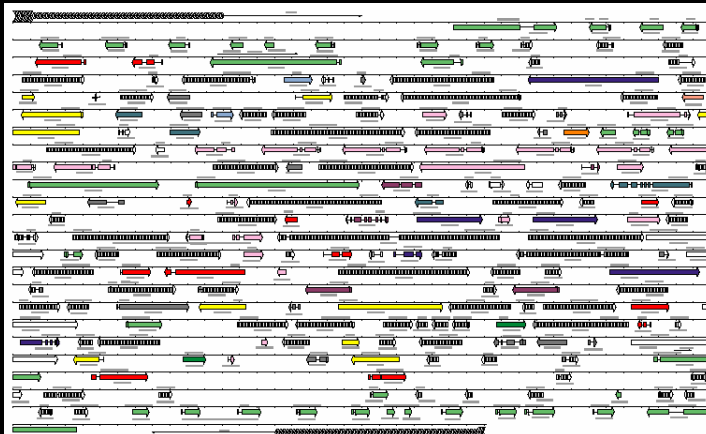
Genome size

- We're the smartest, so we must have the largest genome, right?
- Not quite
- Our genome contains 3000 Mb (~750 megabytes)
- E. coli has 4 Mb
- Yeast has 12 Mb
- Pea has 4800 Mb
- Maize has 5000 Mb
- Wheat has 17000 Mb



Genomes of other organisms

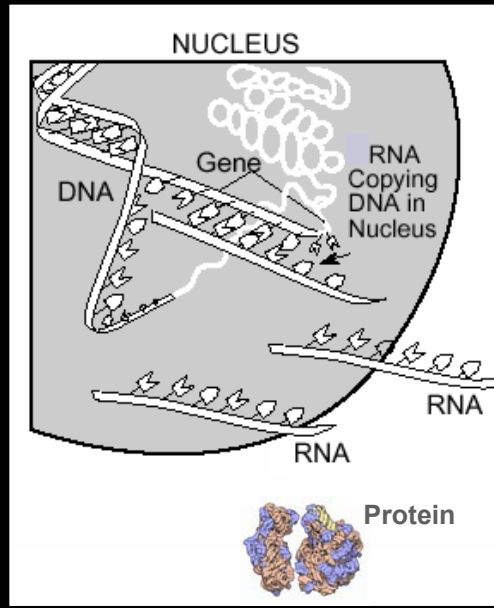
- *Plasmodium falciparum* chromosome 2



Gardner M, et al. Science; 282: 1126 (1998).

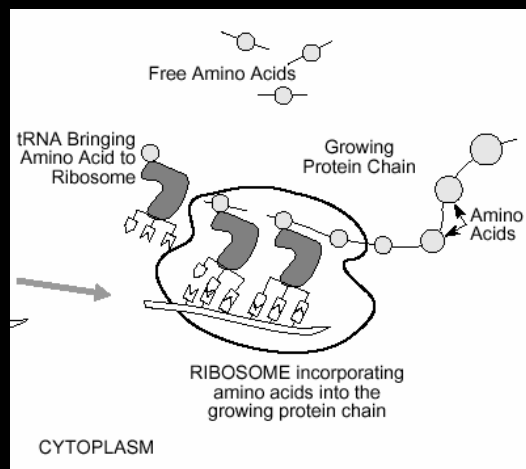
mRNA is made from DNA

- Genes encode instructions to make proteins
- The design of a protein needs to be duplicable
- mRNA is transcribed from DNA within the nucleus
- mRNA moves to the cytoplasm, where the protein is formed



Digitizing amino acid codes

- Proteins are made of 20 (21) amino acids
- Yet each position can only be one of 4 nucleotides
- Nature evolved into using 3 nucleotides to encode a single amino acid
- A chain of amino acids is made from mRNA



Genetic Code

Phe	[171 UUU } AAA 0 203 UUC } GAA 14	Ser	[147 UCU } AGA 10 172 UCC } GGA 0	Tyr	[124 UAU } AUA 1 158 UAC } GUA 11	Cys	[99 UGU } ACA 0 119 UGC } GCA 30
Leu	[73 UUA } UAA 8 125 UUG } CAA 6	stop	[118 UCA } UGA 5 45 UCG } CGA 4	stop	[0 UAA } UUA 0 0 UAG } CUA 0	stop	[0 UGA } UCA 0 Trip - 122 UGG } CCA 7
Leu	[127 CUU } AAG 13 187 CUC } GAG 0 69 CUA } UAG 2 392 CUG } CAG 6	Pro	[175 CCU } AGG 11 197 CCC } GGG 0 170 CCA } UGG 10 69 CCG } CGG 4	His	[104 CAU } AUG 0 147 CAC } GUG 12	Arg	[47 CGU } ACG 9 107 CGC } GCG 0 63 CGA } UCG 7 115 CCG } CCG 5
Ile	[165 AUU } AAU 13 218 AUC } GAU 1 71 AUA } UAU 5	Thr	[131 ACU } AGU 8 192 ACC } GGU 0 150 ACA } UGU 10 63 ACG } CGU 7	Asn	[174 AAU } AAU 1 199 AAC } GAU 33	Ser	[121 AGU } ACU 0 191 AGC } GCU 7 113 AGA } UCU 5 110 AGG } CCU 4
Met	[221 AUG } CAU 17	Lys	[248 AAA } UUU 16 331 AAG } CUU 22				
Val	[111 GUU } AAC 20 146 GUC } GAC 0 72 GUA } UAC 5 288 GUG } CAC 19	Ala	[185 GCU } AGC 25 282 GCC } GGC 0 160 GCA } UGC 10 74 GCG } CGC 5	Asp	[230 GAU } AUC 0 262 GAC } GUC 10 301 GAA } UUC 14 404 GAG } CUC 8	Gly	[112 GGU } ACC 0 230 GGC } GCC 11 168 GGA } UCC 5 160 GGG } CCC 8

Nature; 409: 860 (2001).

Molecular Biology

Nucleotides



Double helix



Chromosome

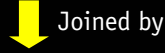


Gene/DNA

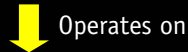


Genome

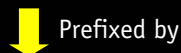
tRNA



Ribosome



mRNA



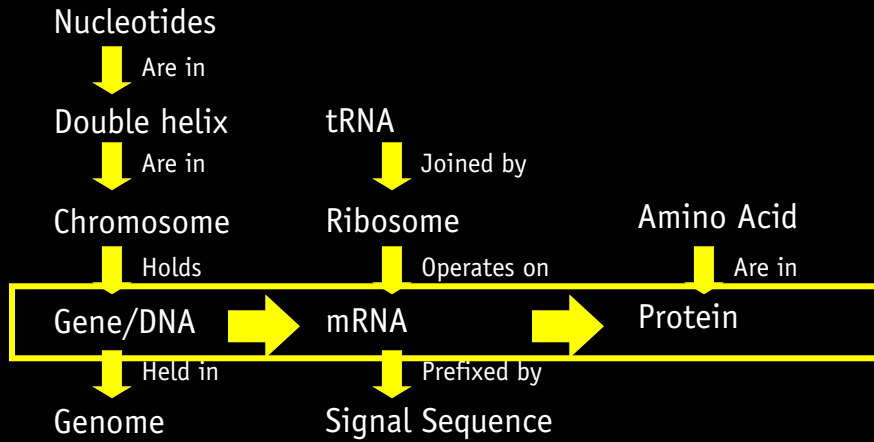
Signal Sequence

Amino Acid



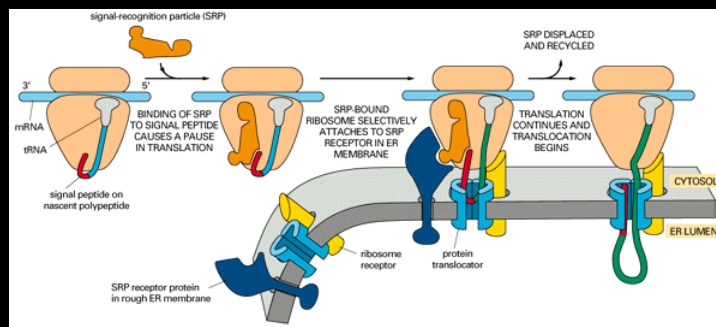
Protein

Central Dogma



Protein targeting

- The first few amino acids may serve as a signal peptide
- Works in conjunction with other cellular machinery to direct protein to the right place



Transcriptional Regulation

- Amount of protein is roughly governed by RNA level
- Transcription into RNA can be activated or repressed by transcription factors

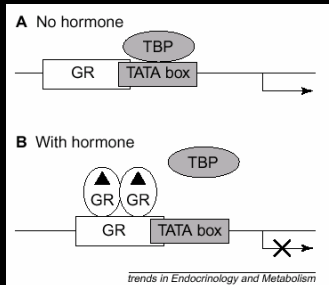


Figure 2. Displacement of the TBP by the hormone-bound GR leads to repression of the osteocalcin gene (B). With no hormone present, the gene is not repressed (A). Abbreviation: GR, glucocorticoid receptor; TBP, TATA binding protein. Triangles represent hormone.

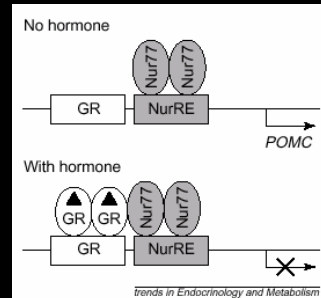
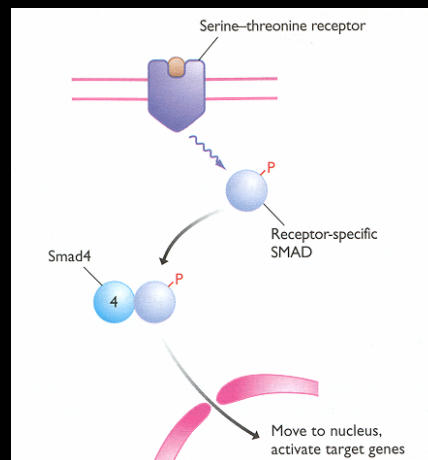


Figure 4. Interaction of the GR with Nur77 on DNA leads to the repression of *POMC*. Abbreviations: GR, glucocorticoid receptor; NurRE, Nur-response element; *POMC*, gene encoding proopiomelanocortin. Triangles represent hormone.

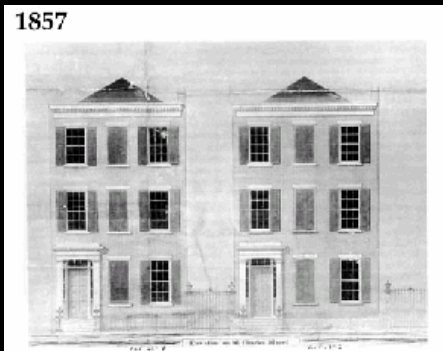
What starts the process?

- Transcriptional programs can start from
 - Hormone action on receptors
 - Shock or stress to the cell
 - New source of, or lack of nutrients
 - Internal derangement of cell or genome
 - Many, many other internal and external stimuli



Temporal Programs

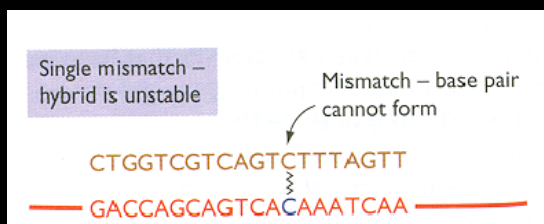
- Segmentation versus Homeosis: same two houses at different times



Scott M. Cell; 100: 27 (2000).

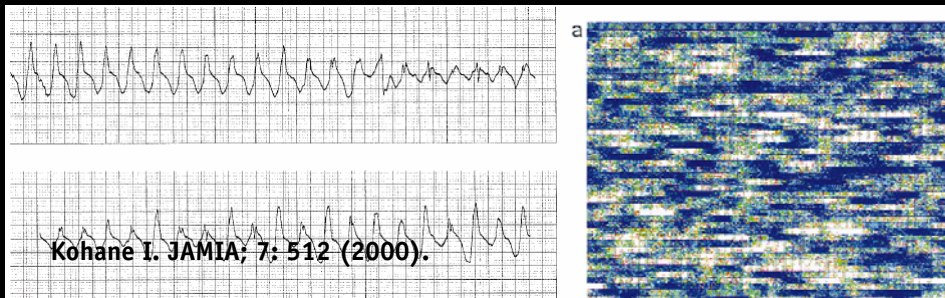
mRNA

- mRNA can be transcribed at up to several hundred nucleotides per minute
- Some eukaryotic genes can take many hours to transcribe
 - Dystrophin takes 20 hours to transcribe
- Most mRNA ends with poly-A, so it is easy to pick out
- Can look for the presence of specific mRNA using the complementary sequence



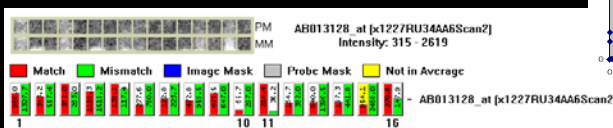
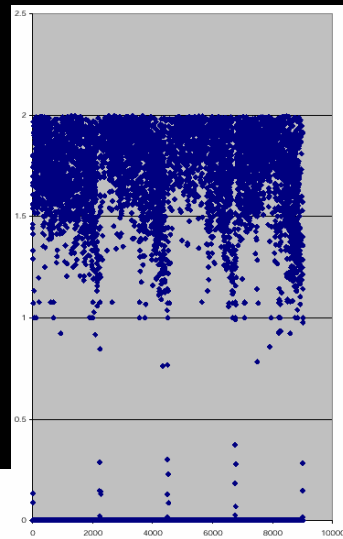
Common Challenges

- High bandwidth data collection
 - Physiological measurements with high sample rates
 - Higher density microarrays
- Data storage
 - 15% US population = 200 million multiGB images
 - Raw sequencing trace files for one human = 300 terabytes



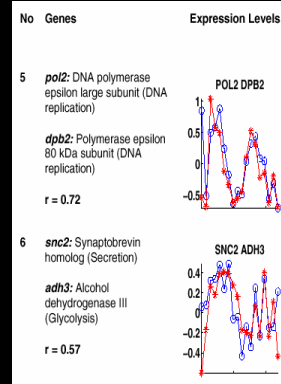
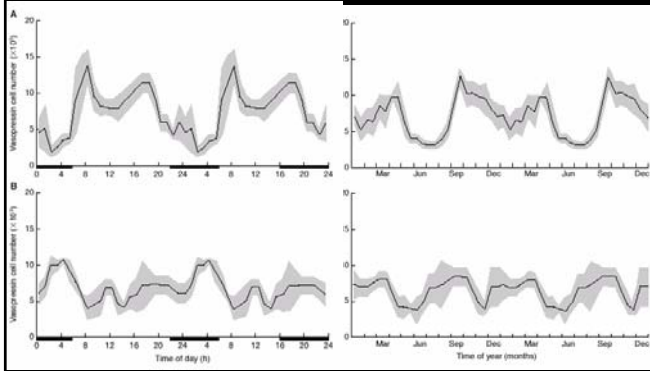
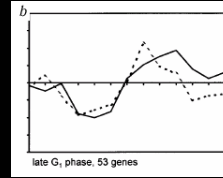
Common Challenges

- Measurement Noise
 - Artifacts in physiological measures
 - Poor expression measurement reproducibility
- Data Models
 - Lack of standards in medical records
 - HL7, HIPAA
 - Too many standards in bioinformatics
 - Gene Expression Markup Language (GEML)
 - Gene Expression Omnibus (GEO)
 - Microarray Markup Language (MAML)
 - Medical record as sample annotation



Common Challenges

- Many frequencies and phase shifts
 - Clinical endocrinology spans seconds to decades
 - What are the naturally occurring genomic frequencies?
- What is the relevant source for data?
 - What is the functional tissue for sleep apnea, hypertension, diabetes?



Common Challenges

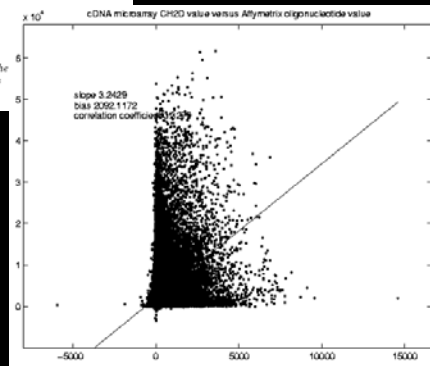
- Comparing new signals to old

003 072599801 0007
The Journal of Clinical Endocrinology & Metabolism
Copyright © 1999 by The Endocrine Society

Vol. 84, No. 4
Printed in U.S.A.

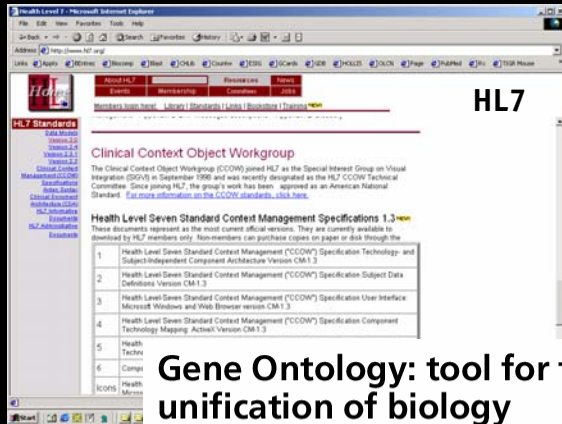
Comparison of the Low Dose Short Synacthen Test (1 μ g), the Conventional Dose Short Synacthen Test (250 μ g), and the Insulin Tolerance Test for Assessment of the Hypothalamo-Pituitary-Adrenal Axis in Patients with Pituitary Disease

T. A. M. ABDU, T. A. ELHADD, K. NEARY, AND R. N. CLAYTON
Departments of Endocrinology and Clinical Chemistry (R.N.), North Staffordshire Hospitals; and the Department of Medicine, School of Postgraduate Medicine, Keele University, Staffordshire, Stoke on Trent, United Kingdom ST4 6QG



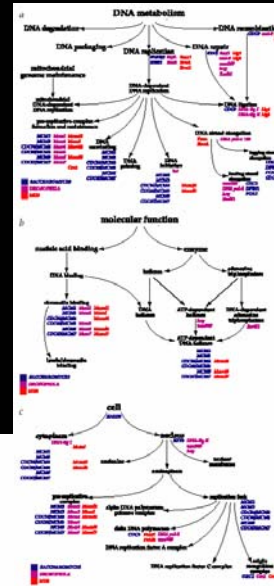
Common Challenges

- Continued development of controlled vocabularies



Gene Ontology: tool for the unification of biology

The Gene Ontology Consortium*



Common Challenges

- Security

nature genetics volume 20 september 1998

screening. However, during the 36-year history of mandatory newborn screening, states have paid relatively little attention to protecting the security of the tissue sample (a dried blood spot or filter paper disc which many programs retain for an extend period). The time is ripe for comprehensive reassessment of regulations governing newborn screening. How will we acquire, analyze and store these data? How will we use this information to help people stay well or ameliorate disease? How will we ensure that information is not misused?

BMJ VOLUME 322 28 APRIL 2001 bmj.com

Ethical requirements for genetic databases

- Follow respectful protocols in approaching people and eliciting medical histories and information about relatives
- Secure informed consent to broad, perhaps open ended, study, and also maybe commercial application of findings
- Manage anonymisation interlinking of databases, and other privacy issues
- Establish confidentiality and security safeguards
- Develop defensible responses to requests for personal data by public health authorities, police, courts, employers, lenders, insurers, and subject relatives
- Devise sound data access, ownership, and intellectual property policies
- Be clear about whether and how individuals will be informed of findings that might be medically helpful to them
- Arrange supervision by research ethics and privacy protection bodies

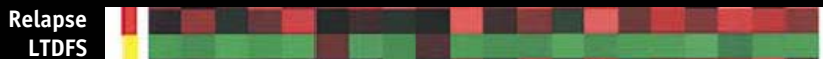
- Privacy
- Ethics

Genetic Testing in Life Insurance "TO BE, OR NOT TO BE"

Genetic testing has many controversial issues surrounding it. The way it affects life and disability insurance is only one area of its reach. However with this area there are two very differing viewpoints. There is the group that believe genetic testing is necessary to prevent insurers from insolvency due to adverse selection. Alternatively, there is the group who believes it infringes on the public's privacy and leads to discrimination. This group wants to protect the public from genetic testing's reign. Which group is right? That question may never be answered, but in any case, there is a lot to be considered.

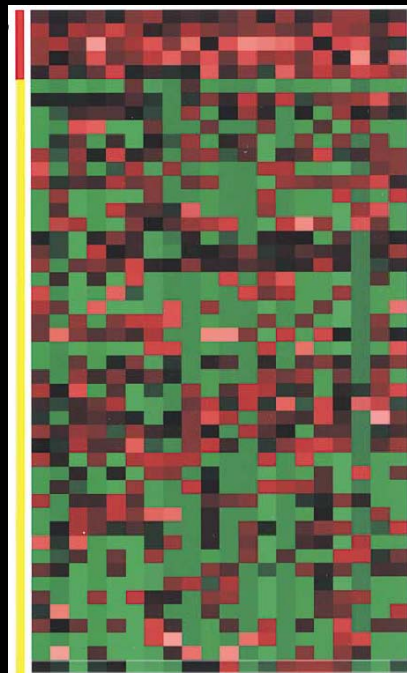
How many samples do we need?

- To prove an 8% difference in event-free survival, is it easier to use 10 patients or 100 patients?
- To make a list of genes that differentiate patients with early relapse from LTDFS, is it easier to use 1 sample of each, or 100 samples of each?



Yeoh, et al. Cancer Cell 2002, 1: 133.

With
microarray
diagnostics,
sample size
is less about
power...



...and
much
more
about
modeling
the
variation
of the
condition

How do we avoid overfitting?

- In other words, with too few samples, it is too easy to overfit the measurements, especially when measuring 20 to 30 thousand genes
- We have techniques like support vector machines that even further expand the number of features
- And even the ones we get wrong, we later find they're been misclassified, or define a new subgroup...

Table 1. ALL subgroup prediction accuracies using support vector machine (SVM)

Subgroups	Training set ^a		Test set ^b	
	Apparent accuracy ^c	True accuracy ^d	Sensitivity ^e	Specificity ^f
T-ALL ^g	100%	100%	100%	100%
E2A-PBX1	100%	100%	100%	100%
TEL-AML1	98%	99%	100%	98%
BCR-ABL	94%	97%	83%	98%
MLL rearrangement	100%	100%	100%	100%
Hyperdiploid >50	93%	96%	100%	93%

^aThe training set consisted of 215 samples.
^bThe blinded test set consisted of 112 samples.
^cApparent accuracy was determined by leave-one-out crossvalidation.
^dTrue accuracy was determined by class prediction on the blinded test set.
^eSensitivity = (the number of positive samples predicted)/(the number of true positives)
^fSpecificity = (the number of negative samples predicted)/(the number of true negatives)
^gThe distribution of cases in the training and test sets are: T-ALL (28 cases, 15 cases); E2A-PBX1 (18, 9); TEL-AML1 (52, 27); BCR-ABL (9, 6); MLL (14; 6); hyperdiploid >50 (42, 22).

Yeoh, et al. Cancer Cell 2002, 1: 133.

Cross-validation

- Random permutation and cross-validation are commonly used in evaluating strategies for picking diagnostic genes
- These can help reduce the danger of overfitting
- **But only additional samples will allow algorithms to learn the variation in disease**
- This reduces false positives

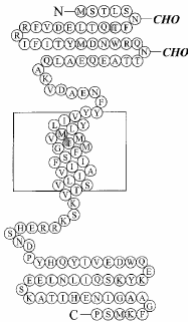
Using Genomics to Treat

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*¶

*Departments of Pediatrics and Cellular and Molecular Physiology, Boyer Center for Molecular Medicine, Yale University School of Medicine, New Haven, CT 06536; †Departments of Medicine and Pharmacology, Vanderbilt University, Nashville, TN 37235; ‡Robert Wood Johnson Medical School, Passaic, NJ 07055; and §Department of Cardiology, University of Pavia and Policlinico San Matteo IRCCS, Pavia, Italy 27100

Edited by Vincent T. Marchesi, Yale University School of Medicine, New Haven, CT, and approved July 6, 2000 (received for review May 16, 2000)



- Genes will help us determine which drugs to use in particular disease subtypes
- Genes will help us predict those who get side-effects

Sesti F. PNAS 97:10613, 2000

Using Genomics to Find New Drugs

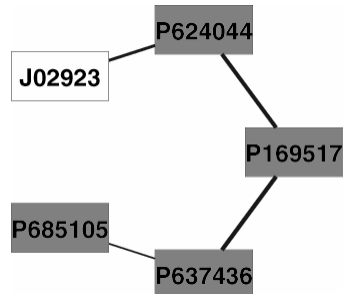
article

© 2000 Nature America Inc. • <http://genetics.nature.com>

A gene expression database for the molecular pharmacology of cancer

Uwe Scherf^{1,8}, Douglas T. Ross², Mark Waltham¹, Lawrence H. Smith¹, Jae K. Lee¹, Lorraine Tanabe¹, Kurt W. Kohn¹, William C. Reinhold¹, Timothy G. Myers⁴, Darren T. Andrews¹, Dominic A. Scudiero⁵, Michael B. Eisen³, Edward A. Sausville⁶, Yves Pommier¹, David Botstein³, Patrick O. Brown^{2,7} & John N. Weinstein¹

- The human genome project and genomics will help us find new drugs
- The entire pharmaceutical industry currently targets 500 cellular targets; this will grow to 3,000 to 10,000



Scherf, U. Nature Genetics 24:236.
Butte, AJ. PNAS 97:12182.

Many physicians do not know how to use the genome

NATIONAL
Science/Health

The New York Times

Home

Editorial

Science

Health

Arts

World

Jeopardizing Your Future?

September 19, 2000

[E-Mail This Article](#)

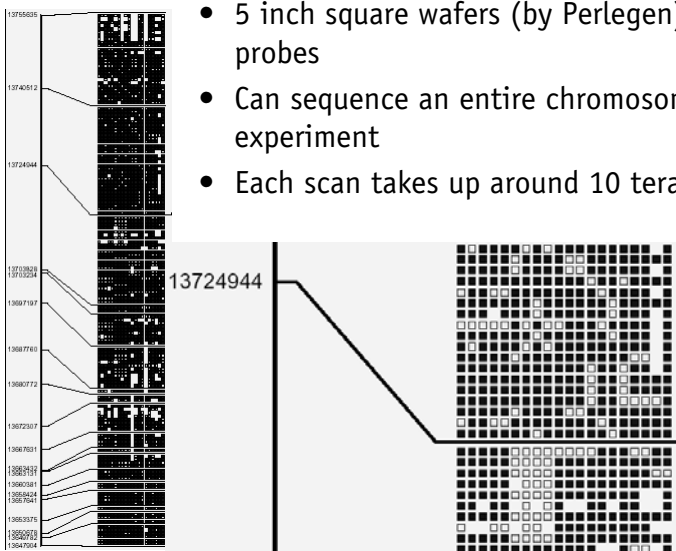
VITAL SIGNS

IN PRACTICE: Genetics: Blind Spot in Medical Training

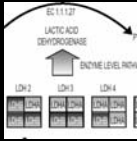
The explosion of knowledge about what role genetics plays in disease has altered the way medical care is being provided at the most basic levels, and many health professionals may not be up to the task.

After microarrays comes wafers...

- Chromosome 21 has 21 million base-pairs
- 5 inch square wafers (by Perlegen) hold 3.4 billion probes
- Can sequence an entire chromosome in one experiment
- Each scan takes up around 10 terabytes



Take Home Points




- Not all pathways will be reverse engineered by microarrays



- With microarrays, sample size plays a larger role in accuracy rather than power




- Due to rapidly changing information, one is never truly finished analyzing a microarray data set

 The Harvard-MIT
Division of
Health Sciences
and Technology

**Bioinformatics
Functional Genomics**

*training a new generation
of quantitative scientists
in bioinformatics and
functional genomics*

PhD Degree Program



Bioinformatics and
Integrative Genomics
big.chip.org

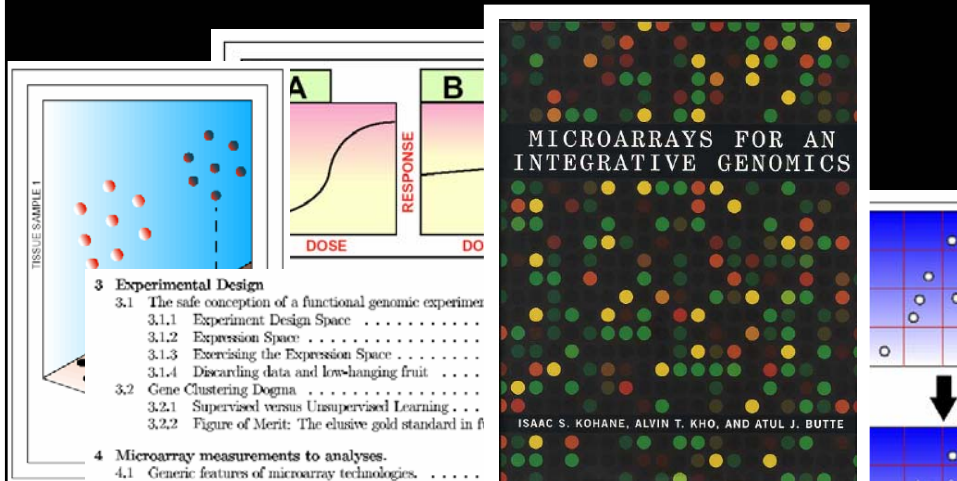
NIH Funded
New PhD training
program in
bioinformatics for
quantitative
individuals

Includes training in wet-
and dry-biology,
clinical medicine

First class Fall 2002

Microarrays for an Integrative Genomics

- The first text-book on microarray analysis and experimental design
- Barnes and Noble, Borders, Amazon: US\$32-40



Collaborators and Support

- Collaborations
 - Scott Weiss / Channing Laboratory
NHLBI Program of Genomics Applications
Nurses Health Study
Physicians Health Study
Normative Aging Study
 - Seigo Izumo / Beth Israel
NHLBI Program of Genomic Applications
Framingham Heart Study
 - David Rowitch / Dana Farber
NINDS Innovative Technologies
 - Dietrich Stephan / Children's National Medical Center
Leukemia Diagnostics
 - Towia Libermann / Beth Israel
NIDDK Biotechnology Center
- Victor Dzau / Brigham and Women's
Angiotensin signaling
- Terry Strom / Beth Israel
NIAID Immune Tolerance Network
- Louis Kunkel / Children's Hospital
Muscular Dystrophy
- C. Ron Kahn and M. E. Patti / Joslin Diabetes Center
Diabetes Genomic Anatomy Project



- Support
 - NIH: **NLM, NINDS, NHLBI, NIDDK, NIAID, NHGRI, NCI, NIGMS**
 - Lawson Wilkins NovoNordisk Award
 - Merck / MIT Fellowship
 - Genentech Foundation Fellowship
 - Endocrine Fellow Foundation

**Bioinformatics at the
Children's Hospital Informatics Program
www.chip.org**

Staff

- Isaac Kohane,
Director
- Atul Butte
- Steven Greenberg
- Peter Park
- Marco Ramoni
- Alberto Riva
- Yao Sun
- Zoltan Szallagi

Fellows

- Ashish Nimgaonkar
- Sunil Saluja
- Dominic Alloco

Post-doctoral fellows

- Zhaohui Cai
- Sangeeta English
- Alvin Kho
- Voichita Marinescu
- Eric Tsung
- Alex Turchin

Students

- Kyungjoon Lee
- Jinyun Chen

Alumni

- Ling Bao
- Aaron Homer
- Janet Karlix
- Ju Han Kim
- Winston Kuo
- Mark Whipple
- Maneesh Yadav

Atul Butte, MD
atul_butte@harvard.edu