Bioinformatics

Jason H. Moore, Ph.D.

Frank Lane Research Scholar in Computational Genetics
Associate Professor of Genetics
Adjunct Associate Professor of Biological Sciences
Adjunct Associate Professor of Community and Family Medicine
Dartmouth College

Affiliate Associate Professor of Computer Science
University of New Hampshire

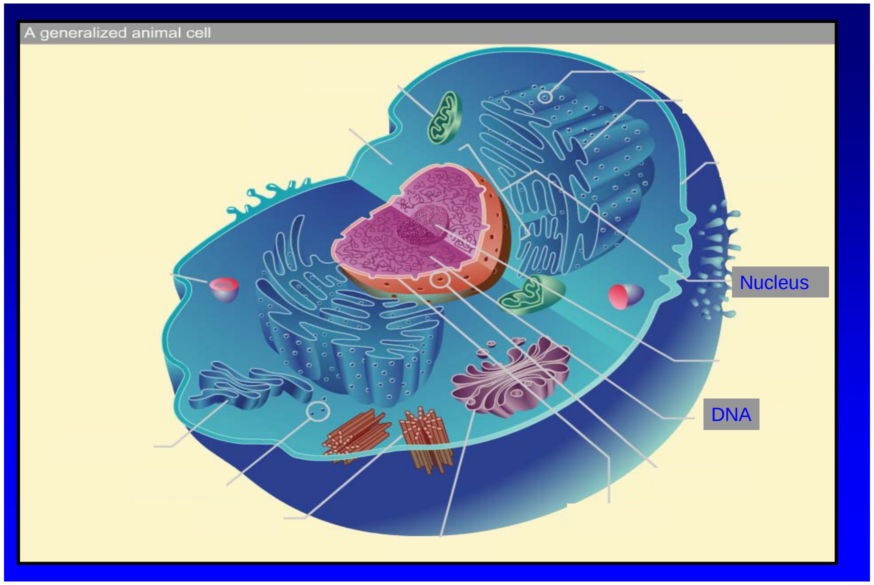
Adjunct Associate Professor of Computer Science
University of Vermont

www.epistasis.org
jason.h.moore@dartmouth.edu

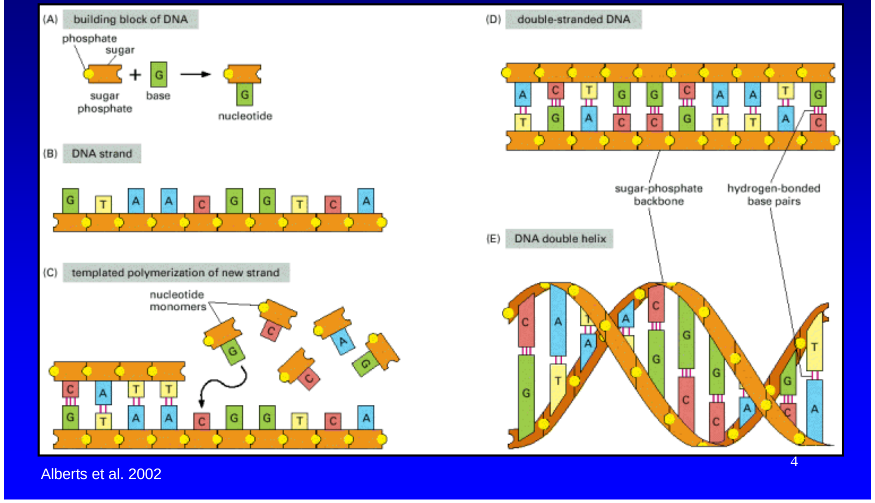
1

Genotype -> Phenotype

2



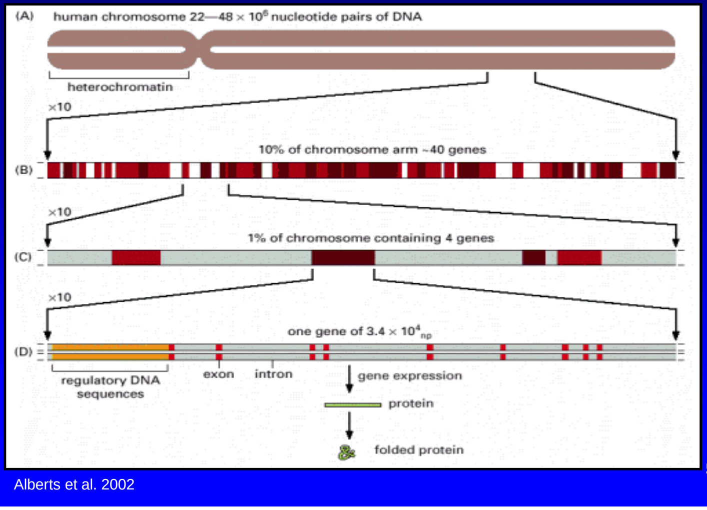
DNA



Alberts et al. 2002

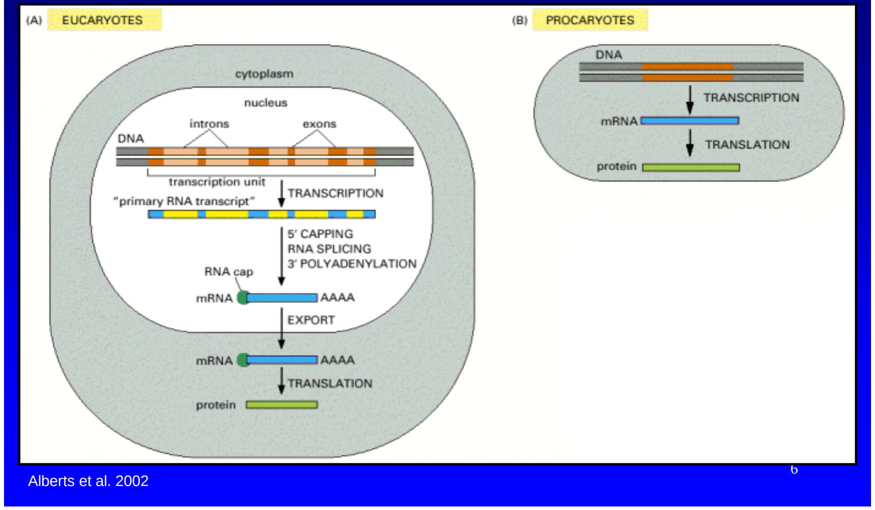
4

Chromosome Structure


(A) human chromosome 22—48 × 10^6 nucleotide pairs of DNA
heterochromatin
×10
10% of chromosome arm ~40 genes
(B)
×10
1% of chromosome containing 4 genes
(C)
×10
one gene of 3.4 × 10^4 np
(D)
regulatory DNA sequences
exon
intron
gene expression
protein
folded protein

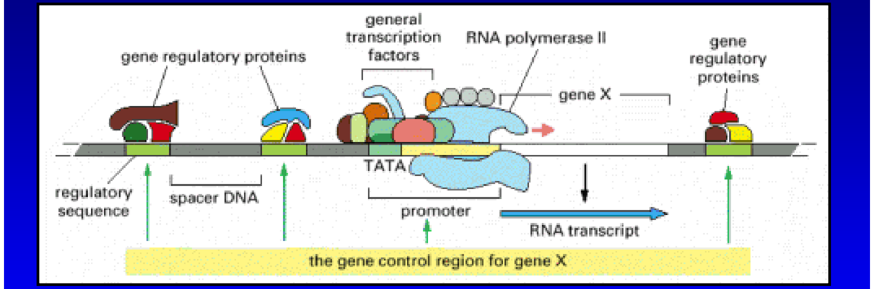
Alberts et al. 2002

Transcription



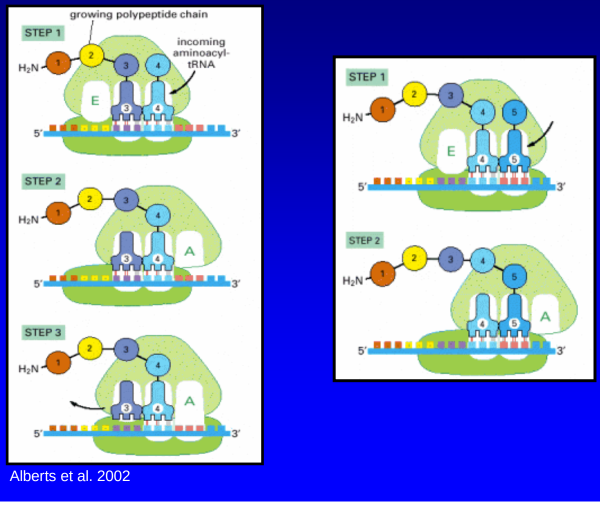
Alberts et al. 2002

Regulation of Transcription

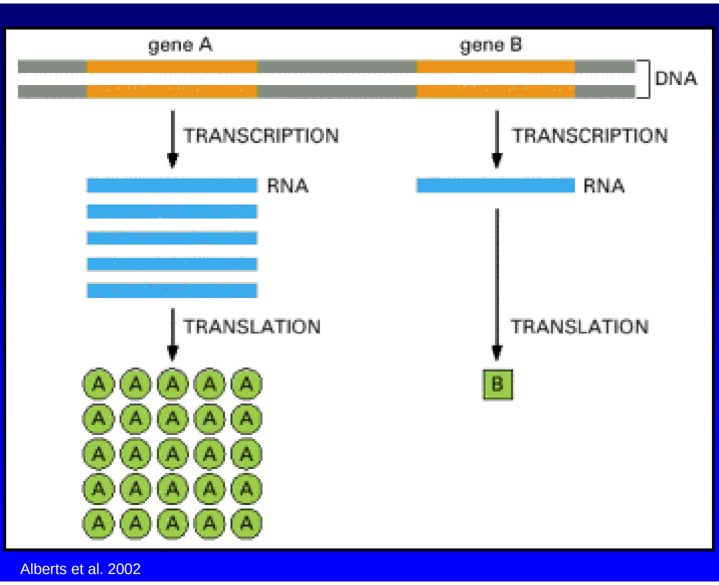


Alberts et al. 2002

Translation

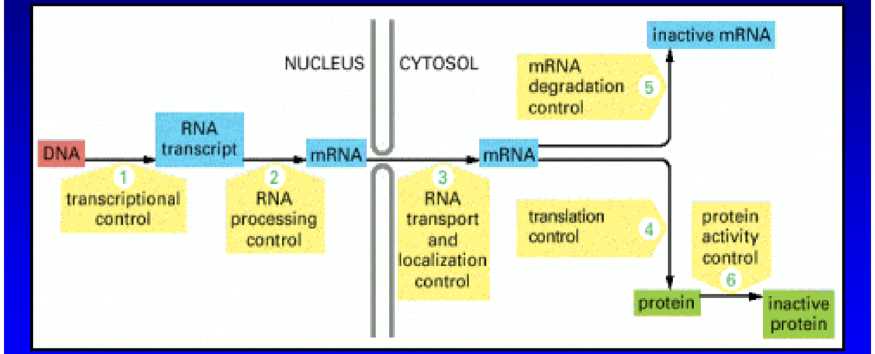


Alberts et al. 2002



Alberts et al. 2002

Control of Gene Expression



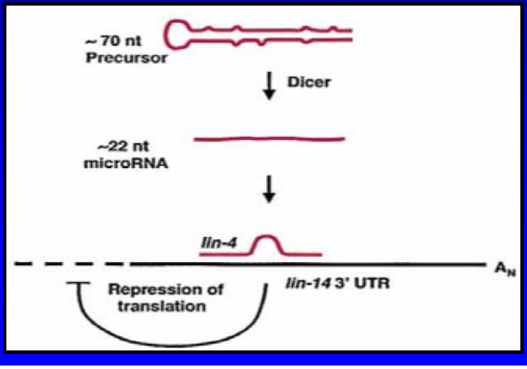
Alberts et al. 2002

10

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

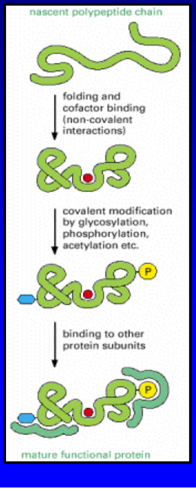
microRNAs: Tiny Regulators with Great Potential

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



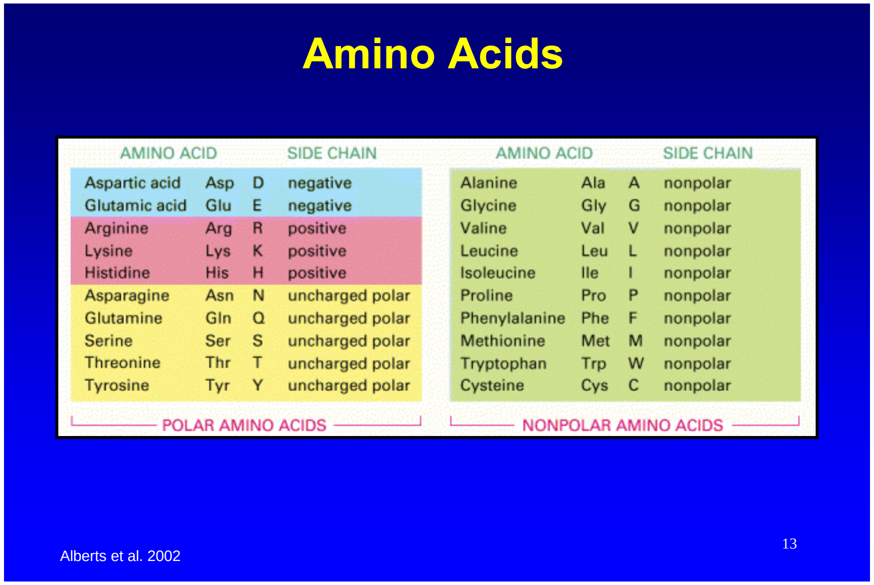
11

Protein Folding


nascent polypeptide chain
folding and cofactor binding (non-covalent interactions)
covalent modification by glycosylation, phosphorylation, acetylation etc.
P
binding to other protein subunits
mature functional protein

Alberts et al. 2002

12

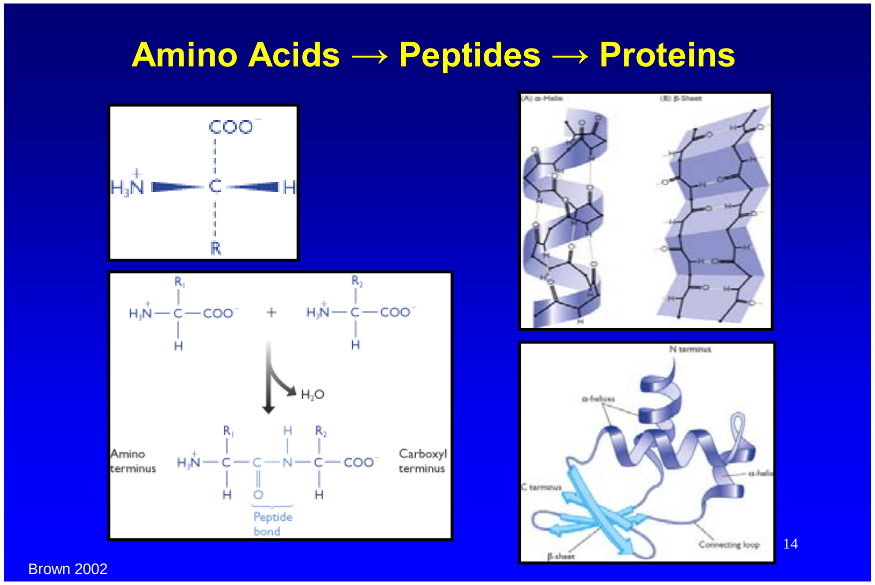


Amino Acids

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			

Alberts et al. 2002

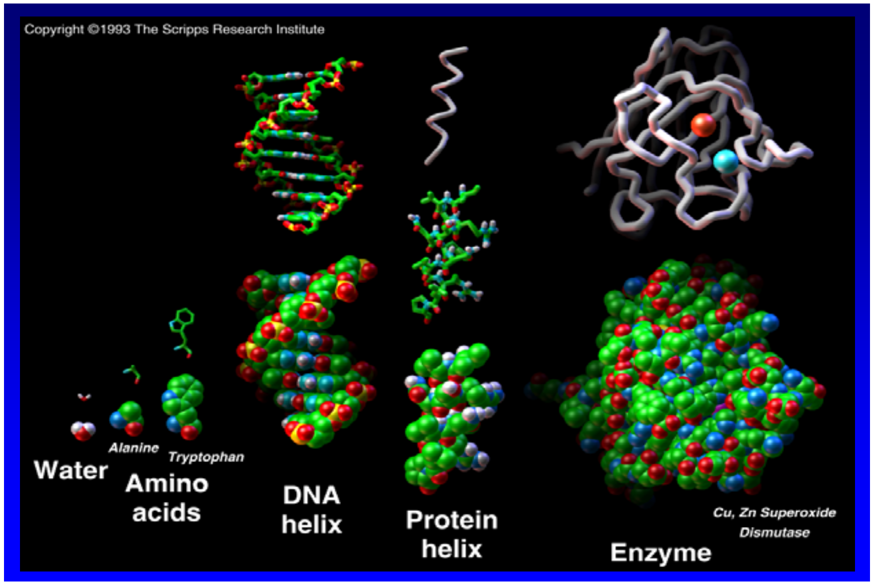
13



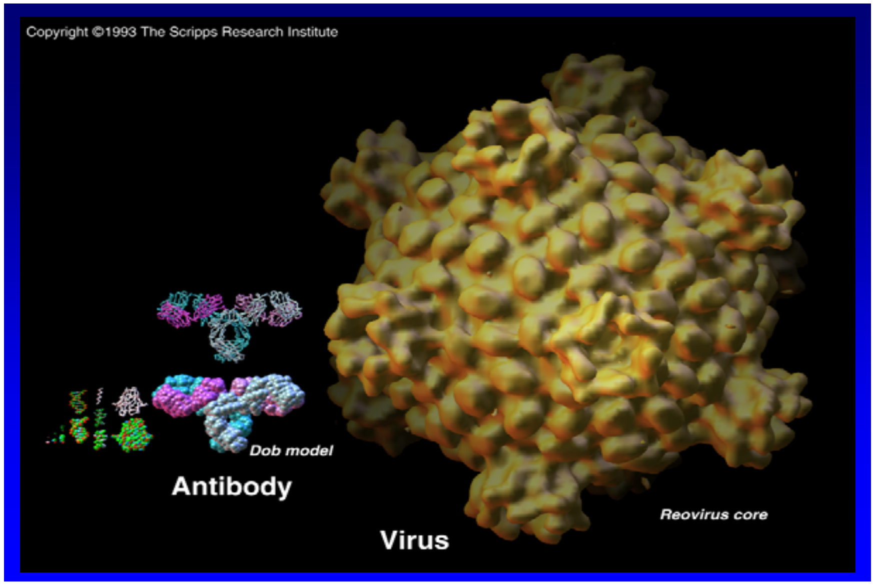
Amino Acids → Peptides → Proteins

Brown 2002

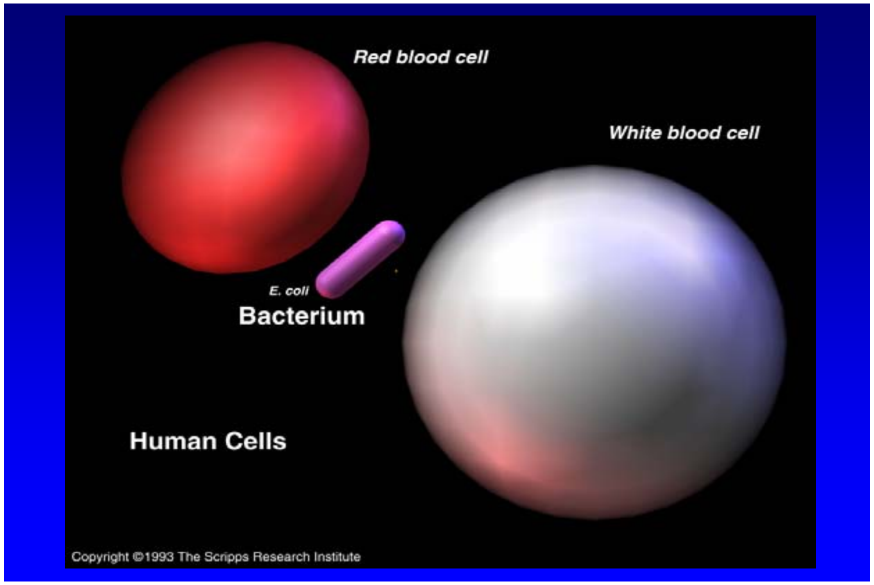
14



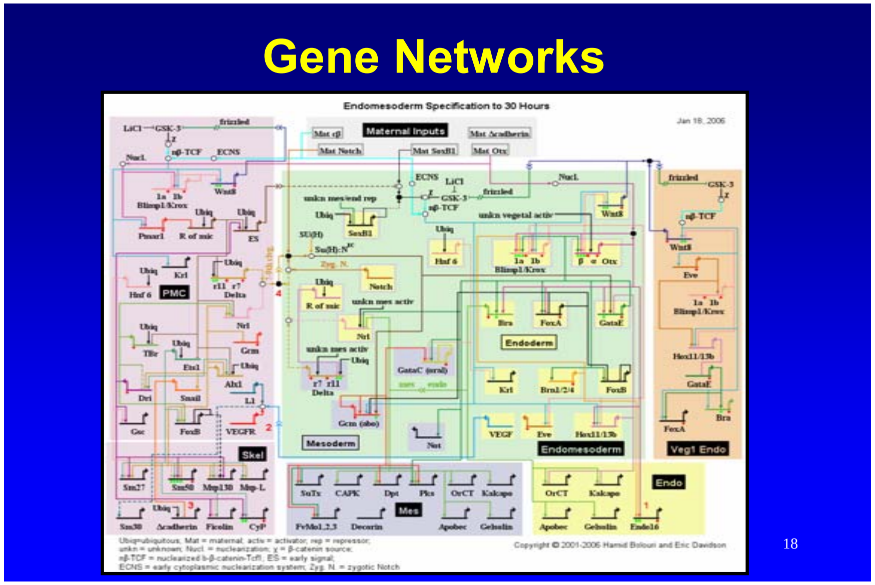
Copyright ©1993 The Scripps Research Institute



Copyright ©1993 The Scripps Research Institute



Copyright ©1993 The Scripps Research Institute



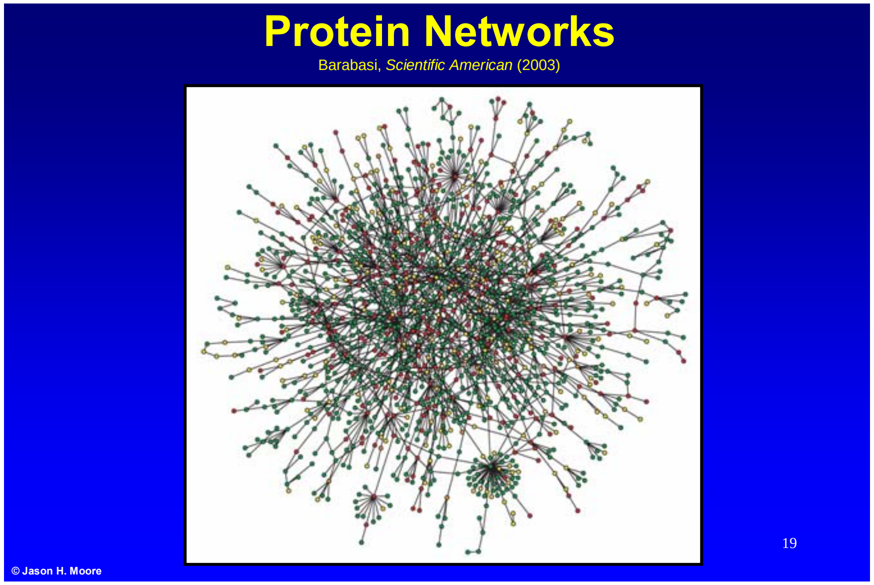
Gene Networks

Endomesoderm Specification to 30 Hours

Jan 18, 2006

Copyright © 2001-2006 Hamid Bolouri and Eric Davidson

18

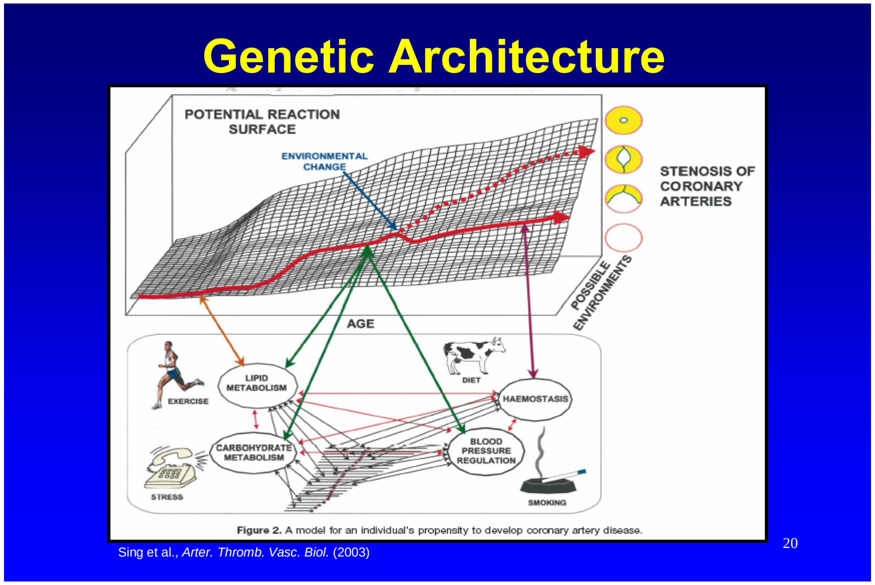


Protein Networks

Barabasi, *Scientific American* (2003)

19

© Jason H. Moore



Genetic Architecture

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

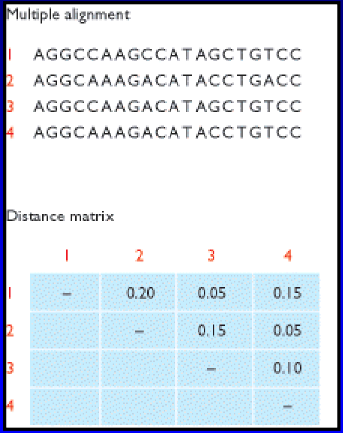
Sing et al., *Arter. Thromb. Vasc. Biol.* (2003)

20

The Tree of Life

http://tolweb.org/tree/

Molecular Phylogenetics

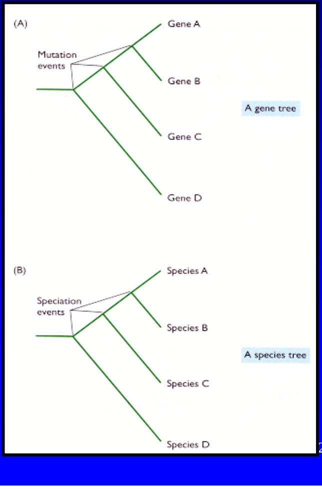


Multiple alignment

1	AGGCCAAGCCATAGCTGTCC
2	AGGCAAAGACATACCTGACC
3	AGGCCAAGACATAGCTGTCC
4	AGGCAAAGACATACCTGTCC

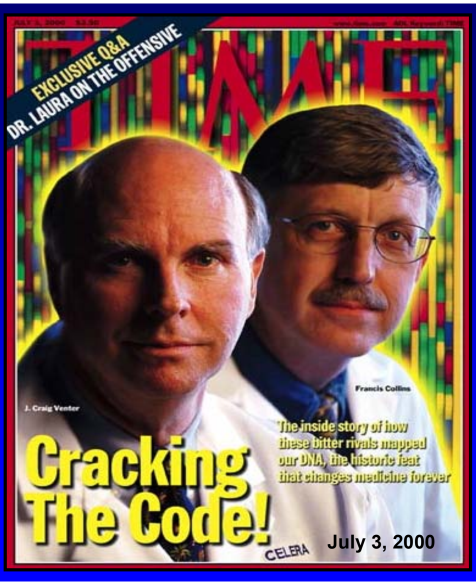
Distance matrix

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



Brown 2002

Measuring DNA


TIME
EXCLUSIVE Q&A
DR. LAURA ON THE OFFENSIVE
J. Craig Venter
Francis Collins
Cracking The Code!
The inside story of how these bitter rivals mapped our DNA, the historic feat that changes medicine forever
CELERA

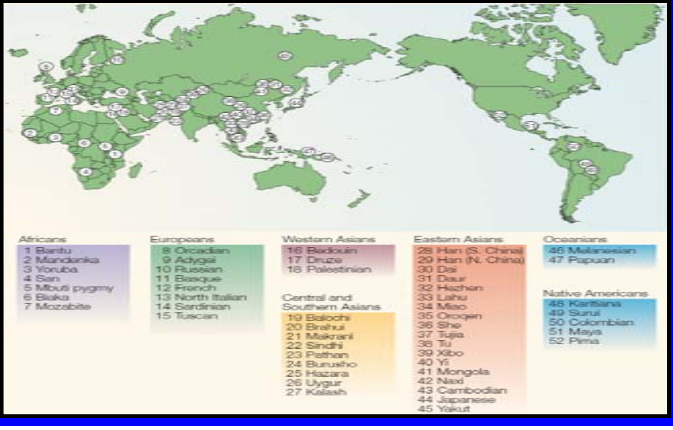
July 3, 2000

OPINION

NATURE REVIEWS | GENETICS
VOLUME 6 | APRIL 2005 | 333

The Human Genome Diversity Project: past, present and future

L. Luca Cavalli-Sforza

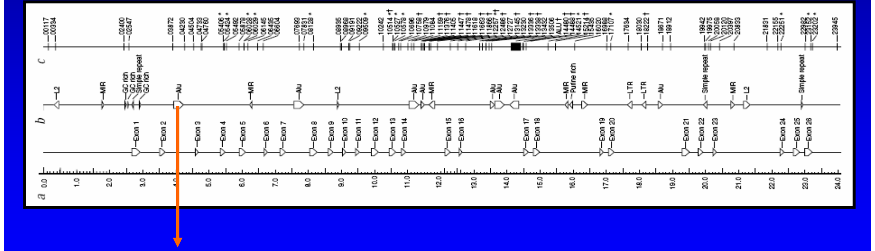


Transposon Repeat Polymorphism

Sequence variation in the human angiotensin converting enzyme

nature genetics • volume 22 • may 1999

Mark J. Rieder[1], Scott L. Taylor[1], Andrew G. Clark[2] & Deborah A. Nickerson[1]



```
  1 ggctgggcgt ggtggctcaa gcctgtaatc ccagcacttt gggaggctga ggtgggcgca
 61 tcgcttgagc ccaggagttc aagaccagcc tggccaacat ggcaaaacct cgtctctaca
121 aaaaaaaaat agctgggctt ggtggtgcgt gcacctacag tcccagctac tcttgaaact
181 gagggggaag gatcacctga gcccaggagg tcaaggctac agtgagctgt gattgcacta
241 ctgcacccca gcctgcgtga cagagtgaga cctcccccca aaaaaaagag agagagaaaa
```

SNPs

Single Nucleotide Polymorphisms

Subject #1	Subject #2	Subject #3
-- AGGTCA -- -- AGGTCA --	-- AGGTCA -- -- AGCTCA --	-- AGCTCA -- -- AGCTCA --

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

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

SNPs

Single Nucleotide Polymorphisms

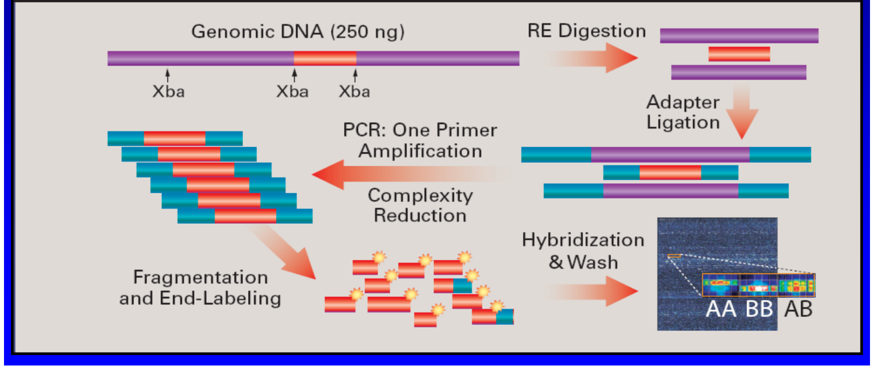
- ~ 1 SNP every 100 bp
- ~ 30 million SNPs
- ~500,000 SNPs in coding DNA
  - Synonymous (silent)
  - Nonsynonymous
    - Deleterious effect
    - Beneficial effect
    - No effect

AFFYMETRIX

Data Sheet

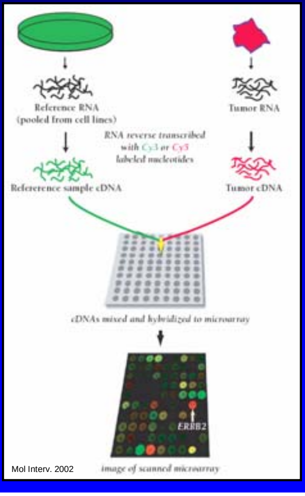
GeneChip® Human Mapping 500K Array Set

Figure 1: GeneChip® Mapping Assay Overview.

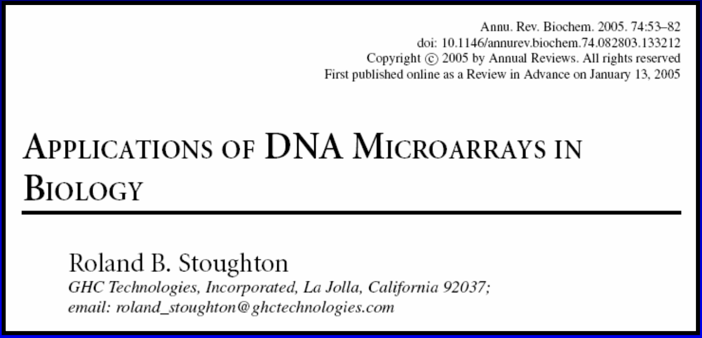


Measuring RNA

cDNA Microarrays


Reference RNA (pooled from cell lines)
Tumor RNA
RNA reverse transcribed with Cy3 or Cy5 labeled nucleotides
Reference sample cDNA
Tumor cDNA
cDNAs mixed and hybridized to microarray
ERBB2
image of scanned microarray

Mol Interv. 2002

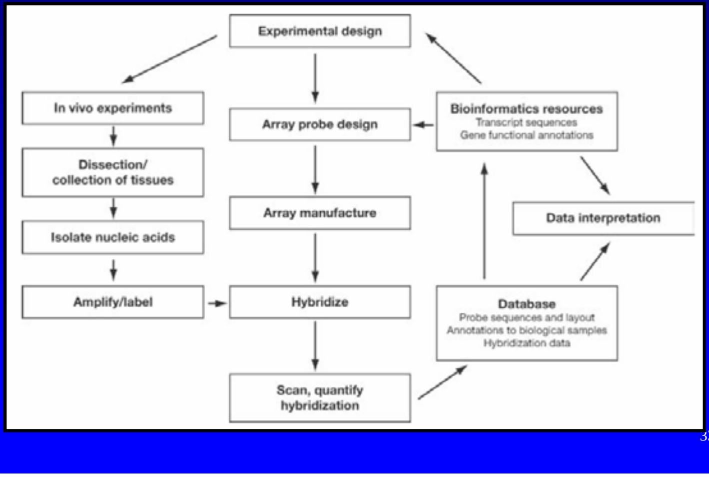


Annu. Rev. Biochem. 2005. 74:53–82
doi: 10.1146/annurev.biochem.74.082803.133212
Copyright © 2005 by Annual Reviews. All rights reserved
First published online as a Review in Advance on January 13, 2005

Applications of DNA Microarrays in Biology

Roland B. Stoughton
GHC Technologies, Incorporated, La Jolla, California 92037;
email: roland_stoughton@ghctechnologies.com

Process Flow for Microarrays



Experimental design

In vivo experiments

Dissection/ collection of tissues

Isolate nucleic acids

Amplify/label

Array probe design

Array manufacture

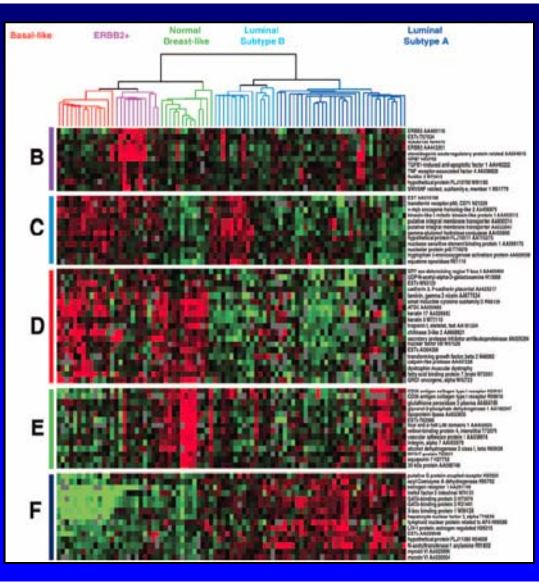
Hybridize

Scan, quantify hybridization

Bioinformatics resources
Transcript sequences
Gene functional annotations

Data interpretation

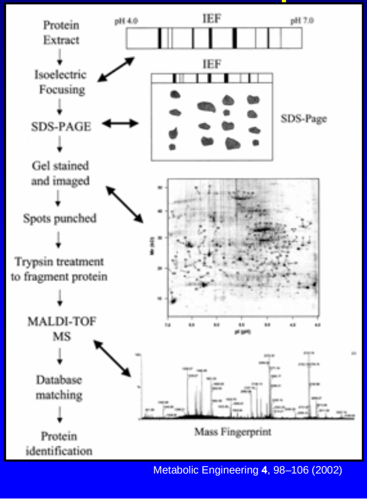
Database
Probe sequences and layout
Annotations to biological samples
Hybridization data



Mol Interv. 2002

Measuring Proteins

2D Gels and Mass Spectrometry



Protein Extract

Isoelectric Focusing

SDS-PAGE

Gel stained and imaged

Spots punched

Trypsin treatment to fragment protein

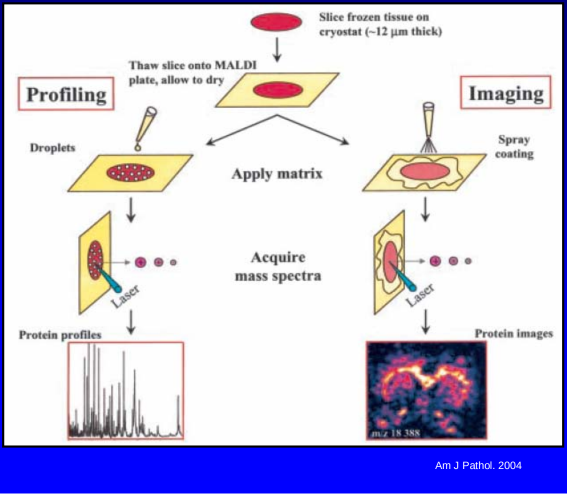
MALDI-TOF MS

Database matching

Protein identification

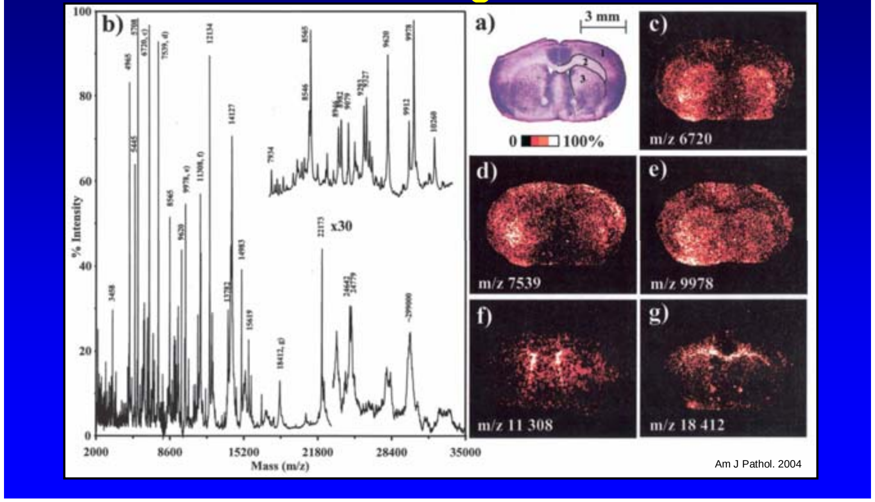
Metabolic Engineering **4**, 98–106 (2002)

Protein Profiling in Tissues



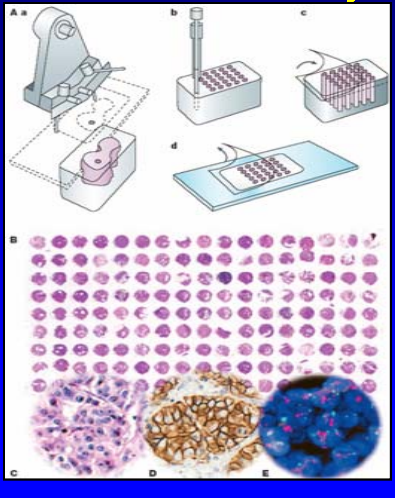
Am J Pathol. 2004

Protein Profiling in Tissues



Am J Pathol. 2004

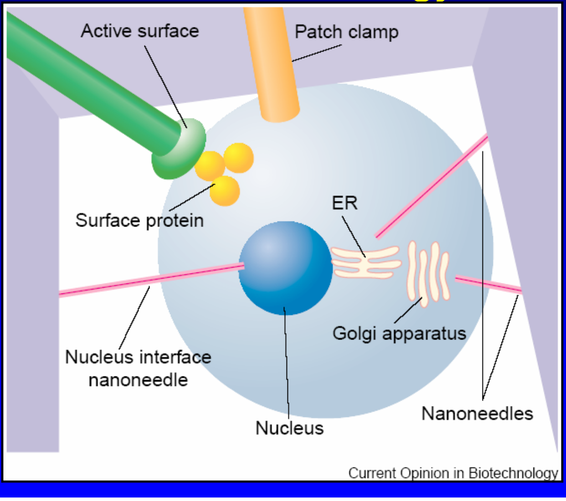
Tissue Microarrays



Nat Rev Drug Discov. 2003

Emerging Technologies

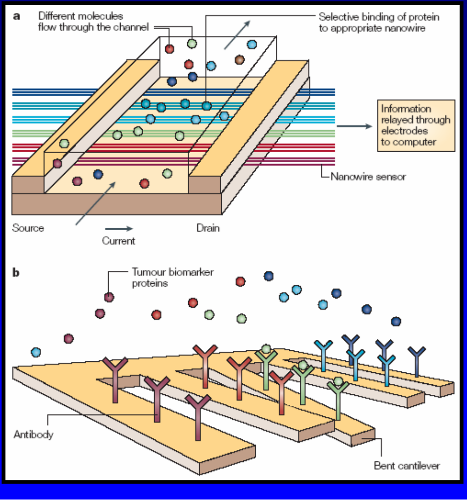
Nanotechnology



Current Opinion in Biotechnology

41

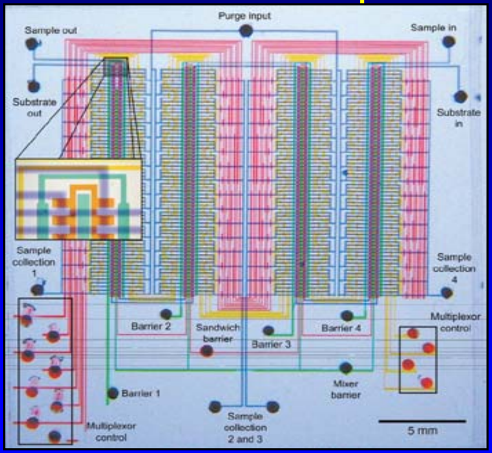
Nanotechnology


a
Different molecules flow through the channel
Selective binding of protein to appropriate nanowire
Information relayed through electrodes to computer
Nanowire sensor
Source
Current
Drain
b
Tumour biomarker proteins
Antibody
Bent cantilever

42

Nature Reviews Cancer 2005

Lab-on-a-Chip



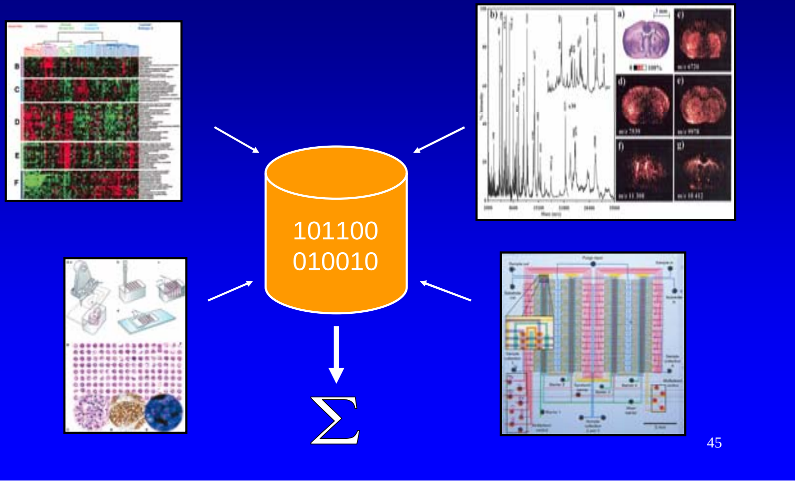
Current Opinion in Structural Biology 2003

43

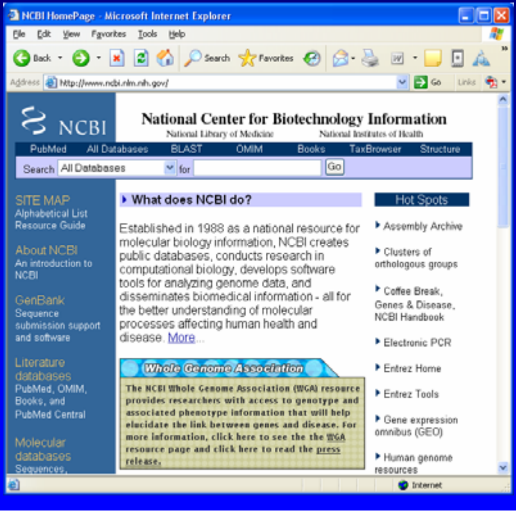
Databases

44

Bioinformatics: Databases



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


NCBI HomePage - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/
Go Links
PubMed All Databases BLAST OMIM Books TaxBrowser Structure
Search All Databases for Go
Internet

National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences,

What does NCBI do?

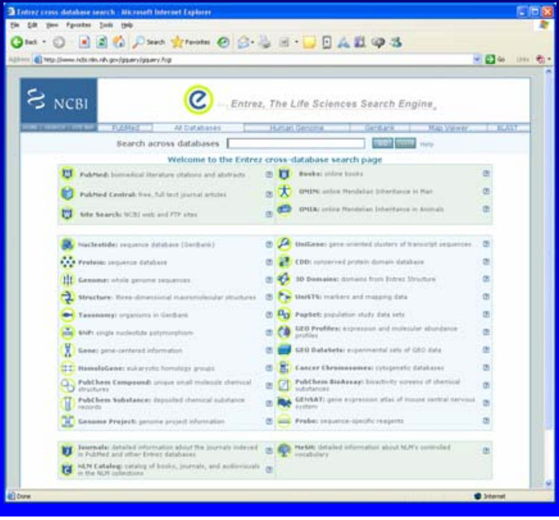
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

Whole Genome Association

The NCBI Whole Genome Association (WGA) resource provides researchers with access to genotype and associated phenotype information that will help elucidate the link between genes and disease. For more information, click here to see the the WGA resource page and click here to read the press release.

Hot Spots
- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources

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


Entrez cross-database search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi
Go Links
NCBI
PubMed All Databases Human Genome GenBank Map Viewer BLAST
Help
Done
Internet

Entrez, The Life Sciences Search Engine

Search across databases

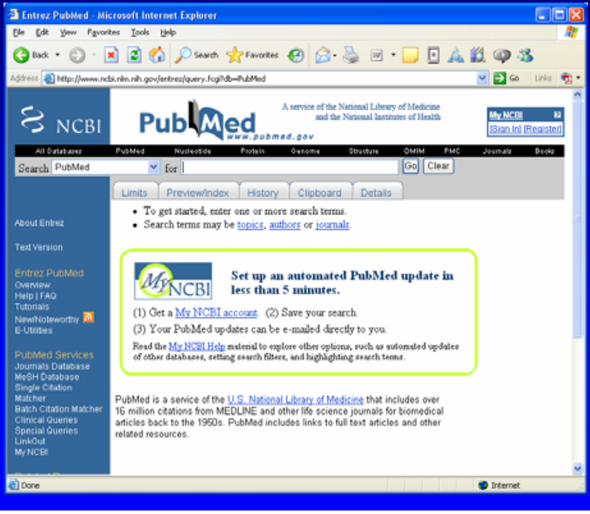
Welcome to the Entrez cross-database search page

- PubMed: biomedical literature citations and abstracts
- PubMed Central: free, full text journal articles
- Site Search: NCBI web and FTP sites
- Books: online books
- OMIM: online Mendelian Inheritance in Man
- OMIA: online Mendelian Inheritance in Animals

- Nucleotide: sequence database (GenBank)
- Protein: sequence database
- Genome: whole genome sequences
- Structure: three-dimensional macromolecular structures
- Taxonomy: organisms in GenBank
- SNP: single nucleotide polymorphism
- Gene: gene-centered information
- HomoloGene: eukaryotic homology groups
- PubChem Compound: unique small molecule chemical structures
- PubChem Substance: deposited chemical substance records
- Genome Project: genome project information
- UniGene: gene-oriented clusters of transcript sequences
- CDD: conserved protein domain database
- 3D Domains: domains from Entrez Structure
- UniSTS: markers and mapping data
- PopSet: population study data sets
- GEO Profiles: expression and molecular abundance profiles
- GEO DataSets: experimental sets of GEO data
- Cancer Chromosomes: cytogenetic databases
- PubChem BioAssay: bioactivity screens of chemical substances
- GENSAT: gene expression atlas of mouse central nervous system
- Probe: sequence-specific reagents

- Journals: detailed information about the journals indexed in PubMed and other Entrez databases
- NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections
- MeSH: detailed information about NLM's controlled vocabulary

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


Entrez PubMed - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
Go Links
NCBI
PubMed
www.pubmed.gov
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search PubMed for Go Clear
Limits Preview/Index History Clipboard Details
Done
Internet

A service of the National Library of Medicine and the National Institutes of Health

About Entrez

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI

- To get started, enter one or more search terms.
- Search terms may be topics, authors or journals.

My NCBI

Set up an automated PubMed update in less than 5 minutes.

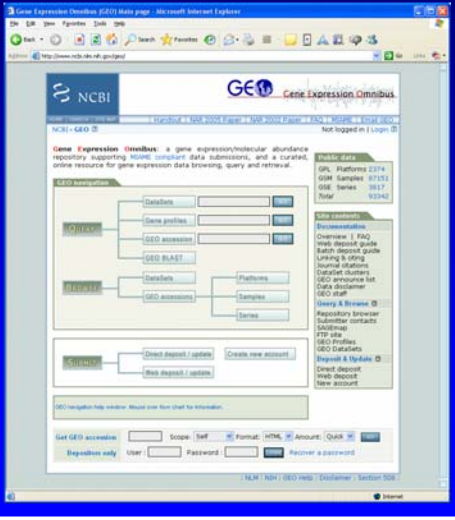
(1) Get a My NCBI account. (2) Save your search.

(3) Your PubMed updates can be e-mailed directly to you.

Read the My NCBI Help material to explore other options, such as automated updates of other databases, setting search filters, and highlighting search terms.

PubMed is a service of the U.S. National Library of Medicine that includes over 16 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. PubMed includes links to full text articles and other related resources.

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


Gene Expression Omnibus (GEO) Main page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/geo/
NCBI
GEO Gene Expression Omnibus
Handout | NAR 2005 Paper | NAR 2002 Paper | FAQ | MIAME | Email GEO
NCBI > GEO
Not logged in | Login
GEO navigation
Query
DataSets
Gene profiles
GEO accession
GEO BLAST
Browse
DataSets
GEO accessions
Platforms
Samples
Series
Submit
Direct deposit / update
Create new account
Web deposit / update
GEO navigation help window. Mouse over flow chart for information.
Get GEO accession
Scope: Self
Format: HTML
Amount: Quick
Depositors only
User :
Password :
Recover a password
| NLM | NIH | GEO Help | Disclaimer | Section 508
Internet

Gene Expression Omnibus: a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

Public data

GPL	Platforms	2374
GSM	Samples	87151
GSE	Series	3817
Total		93342

Site contents

Documentation

Overview | FAQ
Web deposit guide
Batch deposit guide
Linking & citing
Journal citations
DataSet clusters
GEO announce list
Data disclaimer
GEO staff

Query & Browse

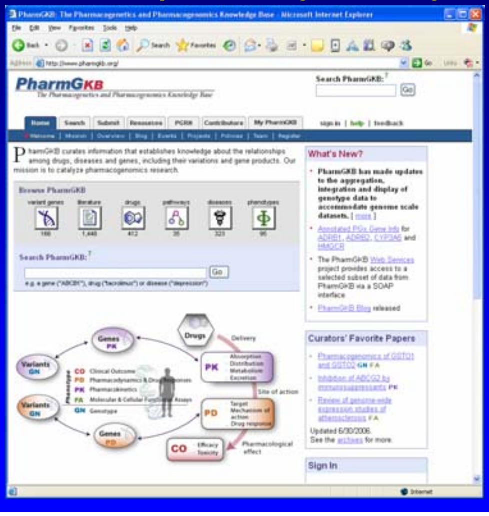
Repository browser
Submitter contacts
SAGEmap
FTP site
GEO Profiles
GEO DataSets

Deposit & Update

Direct deposit
Web deposit
New account

49

www.pharmgkb.org



PharmGKB curates information that establishes knowledge about the relationships among drugs, diseases and genes, including their variations and gene products. Our mission is to catalyze pharmacogenomics research.

Search PharmGKB:

e.g. a gene ("ABCB1"), drug ("tacrolimus") or disease ("depression")

What's New?

- **PharmGKB has made updates to the aggregation, integration and display of genotype data to accommodate genome scale datasets.** [more]
- Annotated PGx Gene Info for ADRB1, ADRB2, CYP3A5 and HMGCR
- The PharmGKB Web Services project provides access to a selected subset of data from PharmGKB via a SOAP interface
- PharmGKB Blog released

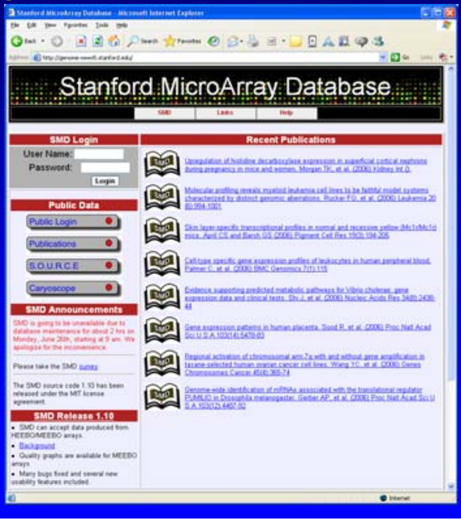
Curators' Favorite Papers

- Pharmacogenomics of GSTO1 and GSTO2 GN FA
- Inhibition of ABCG2 by immunosuppressants PK
- Review of genome-wide expression studies of atherosclerosis FA

Updated 6/30/2006.
See the archives for more.

50

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


Stanford MicroArray Database - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://genome-www5.stanford.edu/
Stanford MicroArray Database
SMD
Links
Help
SMD Login
User Name:
Password:
Login
Public Data
Public Login
Publications
S.O.U.R.C.E
Caryoscope
Internet

Recent Publications

- Upregulation of histidine decarboxylase expression in superficial cortical nephrons during pregnancy in mice and women. Morgan TK, et al. (2006) Kidney Int 0.
- Molecular profiling reveals myeloid leukemia cell lines to be faithful model systems characterized by distinct genomic aberrations. Rucker FG, et al. (2006) Leukemia 20 (6):994-1001
- Skin layer-specific transcriptional profiles in normal and recessive yellow (Mc1r/Mc1r) mice. April CS and Barsh GS (2006) Pigment Cell Res 19(3):194-205
- Cell-type specific gene expression profiles of leukocytes in human peripheral blood. Palmer C, et al. (2006) BMC Genomics 7(1):115
- Evidence supporting predicted metabolic pathways for Vibrio cholerae: gene expression data and clinical tests. Shi J, et al. (2006) Nucleic Acids Res 34(8):2438-44
- Gene expression patterns in human placenta. Sood R, et al. (2006) Proc Natl Acad Sci U S A 103(14):5478-83
- Regional activation of chromosomal arm 7q with and without gene amplification in taxane-selected human ovarian cancer cell lines. Wang YC, et al. (2006) Genes Chromosomes Cancer 45(4):365-74
- Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster. Gerber AP, et al. (2006) Proc Natl Acad Sci U S A 103(12):4487-92

SMD Announcements

SMD is going to be unavailable due to database maintenance for about 2 hrs on Monday, June 26th, starting at 9 am. We apologize for the inconvenience.

Please take the SMD survey

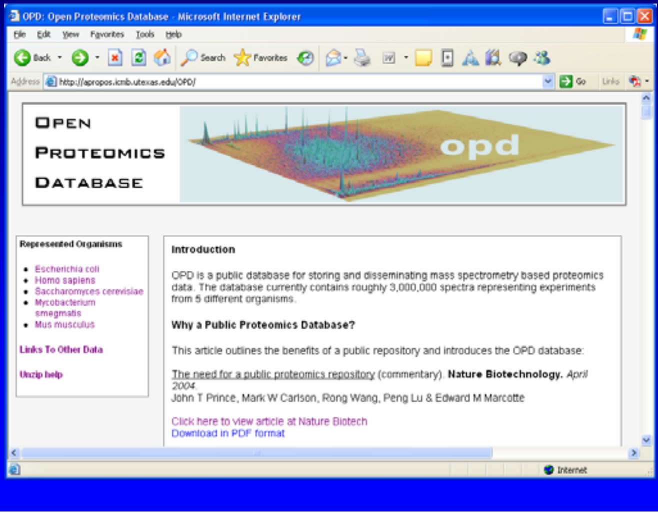
The SMD source code 1.10 has been released under the MIT license agreement.

SMD Release 1.10

- SMD can accept data produced from HEEBO/MEEBO arrays.
- Background
- Quality graphs are available for MEEBO arrays
- Many bugs fixed and several new usability features included

51

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


OPD: Open Proteomics Database - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://apropos.icmb.utexas.edu/OPD/
Go
Links
OPEN PROTEOMICS DATABASE
opd
Internet

Represented Organisms

- Escherichia coli
- Homo sapiens
- Saccharomyces cerevisiae
- Mycobacterium smegmatis
- Mus musculus

Links To Other Data

Unzip help

Introduction

OPD is a public database for storing and disseminating mass spectrometry based proteomics data. The database currently contains roughly 3,000,000 spectra representing experiments from 5 different organisms.

Why a Public Proteomics Database?

This article outlines the benefits of a public repository and introduces the OPD database:

The need for a public proteomics repository (commentary). **Nature Biotechnology.** *April 2004.*
John T Prince, Mark W Carlson, Rong Wang, Peng Lu & Edward M Marcotte

Click here to view article at Nature Biotech
Download in PDF format

52

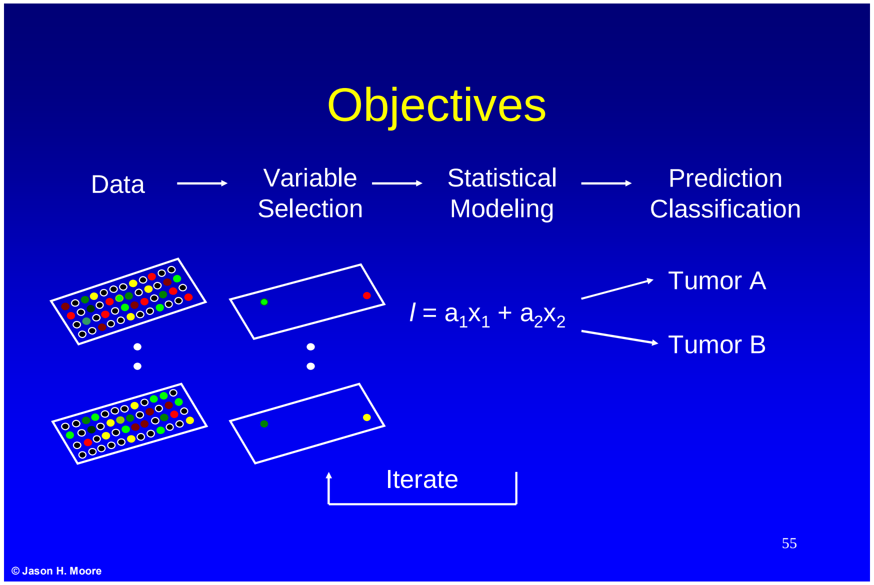
Analysis

53

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?

54



Objectives

Data → Variable Selection → Statistical Modeling → Prediction Classification

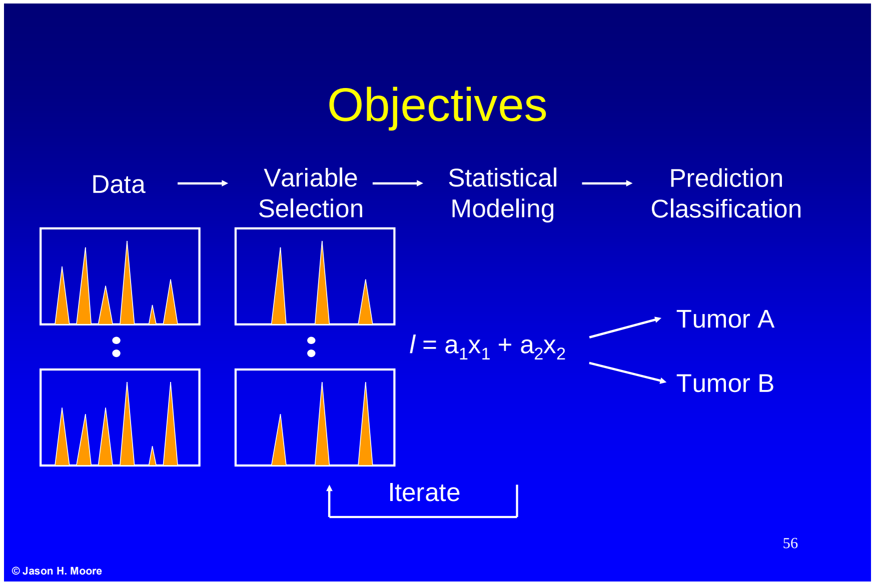
$l = a_1x_1 + a_2x_2$

Tumor A

Tumor B

Iterate

55



Objectives

Data → Variable Selection → Statistical Modeling → Prediction Classification

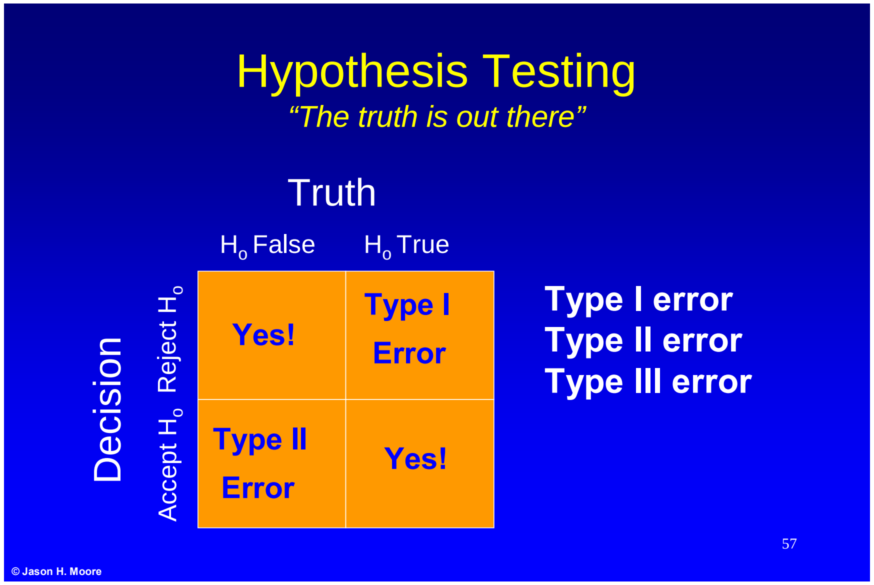
$l = a_1x_1 + a_2x_2$

Tumor A

Tumor B

Iterate

56



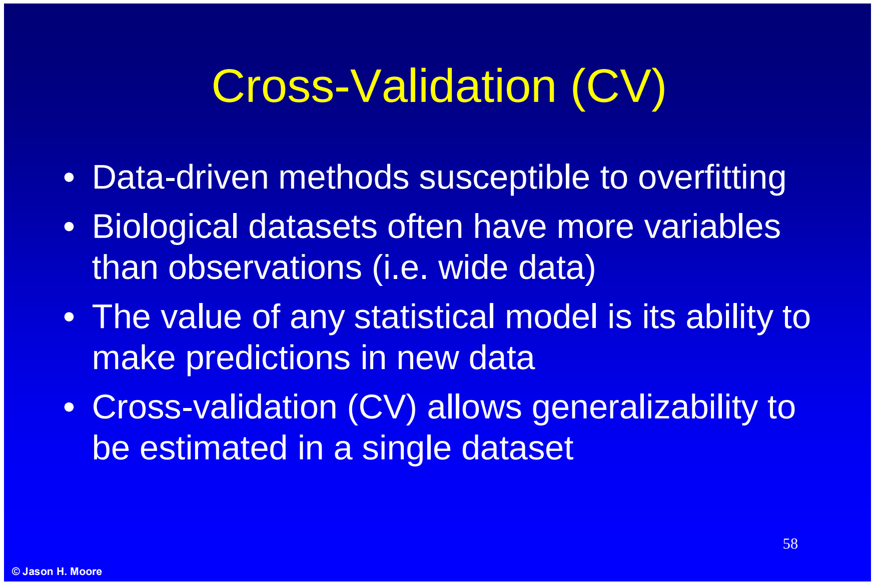
Hypothesis Testing

"The truth is out there"

Decision

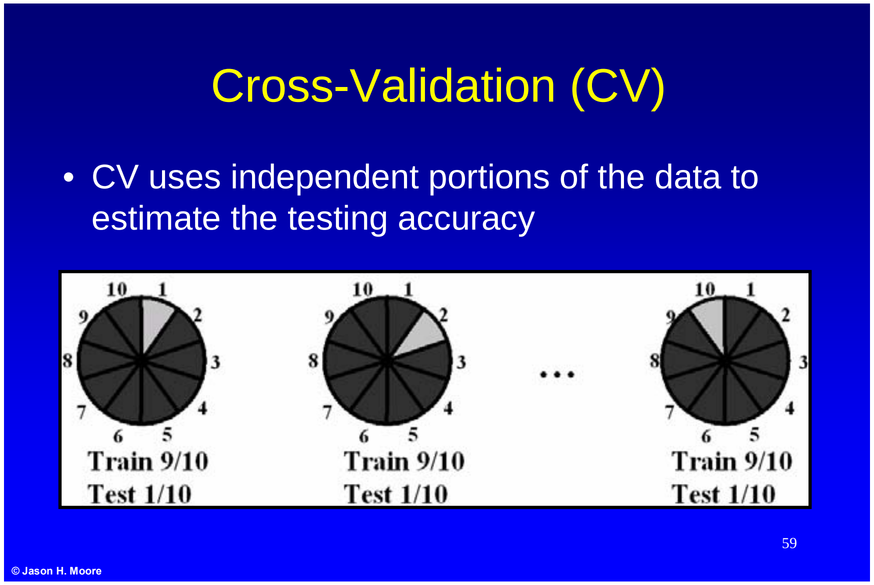
	Truth: H_o False	H_o True
Reject H_o	**Yes!**	**Type I Error**
Accept H_o	**Type II Error**	**Yes!**

Type I error
Type II error
Type III error



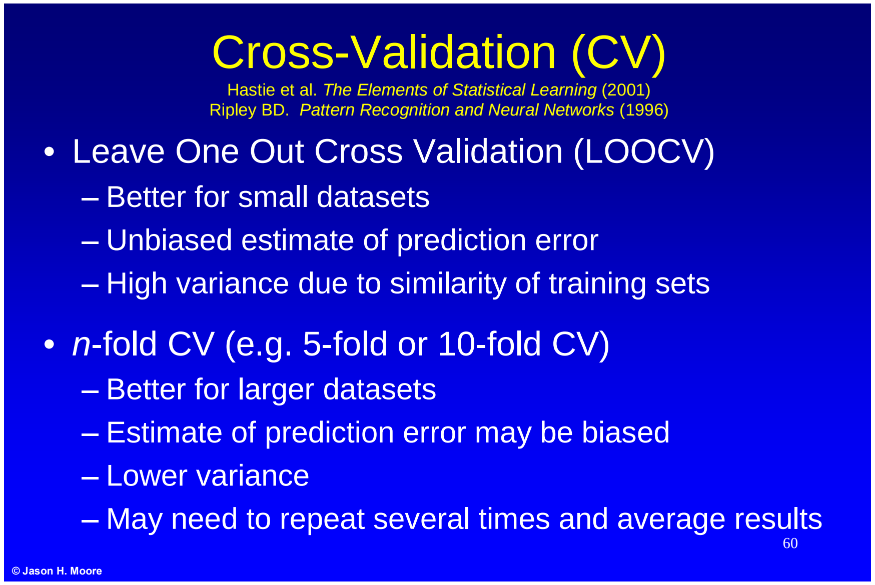
Cross-Validation (CV)

- Data-driven methods susceptible to overfitting
- Biological datasets often have more variables than observations (i.e. wide data)
- The value of any statistical model is its ability to make predictions in new data
- Cross-validation (CV) allows generalizability to be estimated in a single dataset



Cross-Validation (CV)

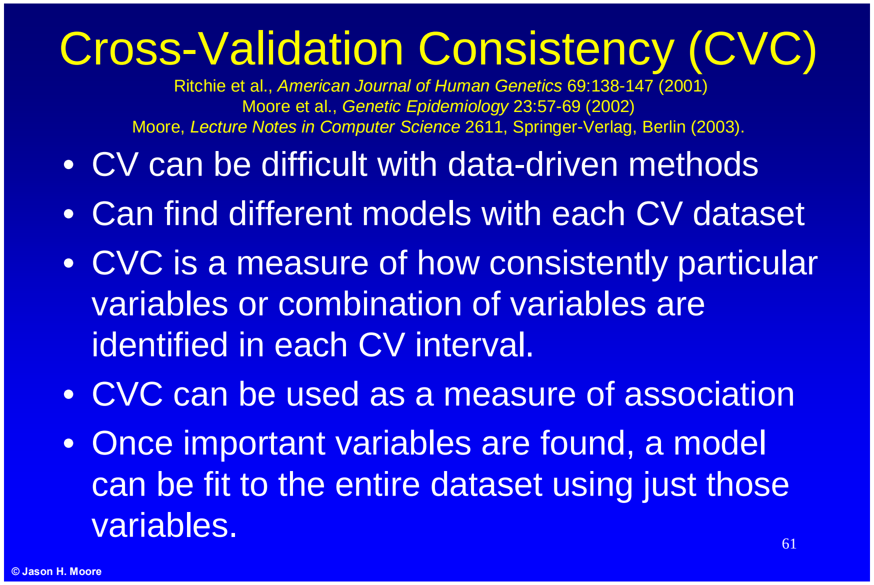
- CV uses independent portions of the data to estimate the testing accuracy



Cross-Validation (CV)

Hastie et al. *The Elements of Statistical Learning* (2001)
Ripley BD. *Pattern Recognition and Neural Networks* (1996)

- Leave One Out Cross Validation (LOOCV)
  - – Better for small datasets
  - – Unbiased estimate of prediction error
  - – High variance due to similarity of training sets
- *n*-fold CV (e.g. 5-fold or 10-fold CV)
  - – Better for larger datasets
  - – Estimate of prediction error may be biased
  - – Lower variance
  - – May need to repeat several times and average results



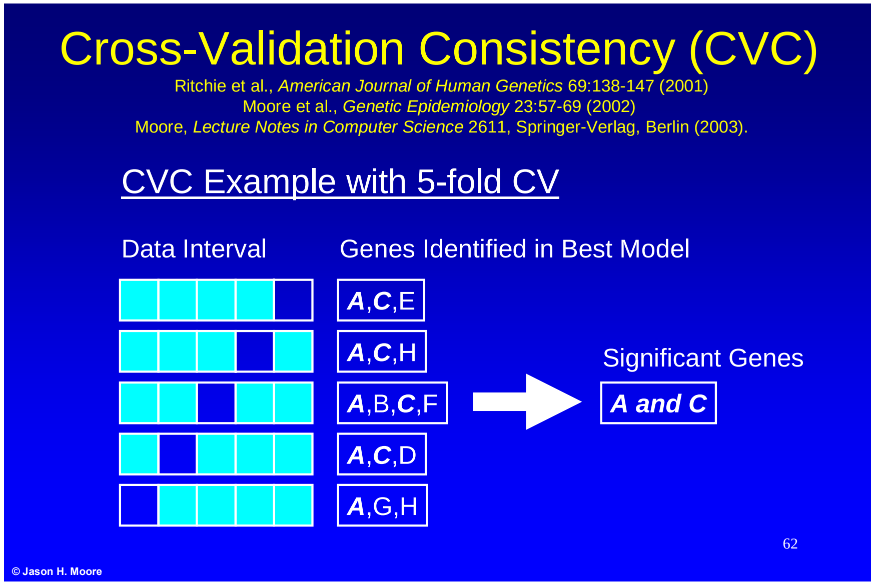
Cross-Validation Consistency (CVC)

Ritchie et al., *American Journal of Human Genetics* 69:138-147 (2001)
Moore et al., *Genetic Epidemiology* 23:57-69 (2002)
Moore, *Lecture Notes in Computer Science* 2611, Springer-Verlag, Berlin (2003).

- CV can be difficult with data-driven methods
- Can find different models with each CV dataset
- CVC is a measure of how consistently particular variables or combination of variables are identified in each CV interval.
- CVC can be used as a measure of association
- Once important variables are found, a model can be fit to the entire dataset using just those variables.

61

© Jason H. Moore



Cross-Validation Consistency (CVC)

Ritchie et al., *American Journal of Human Genetics* 69:138-147 (2001)
Moore et al., *Genetic Epidemiology* 23:57-69 (2002)
Moore, *Lecture Notes in Computer Science* 2611, Springer-Verlag, Berlin (2003).

CVC Example with 5-fold CV

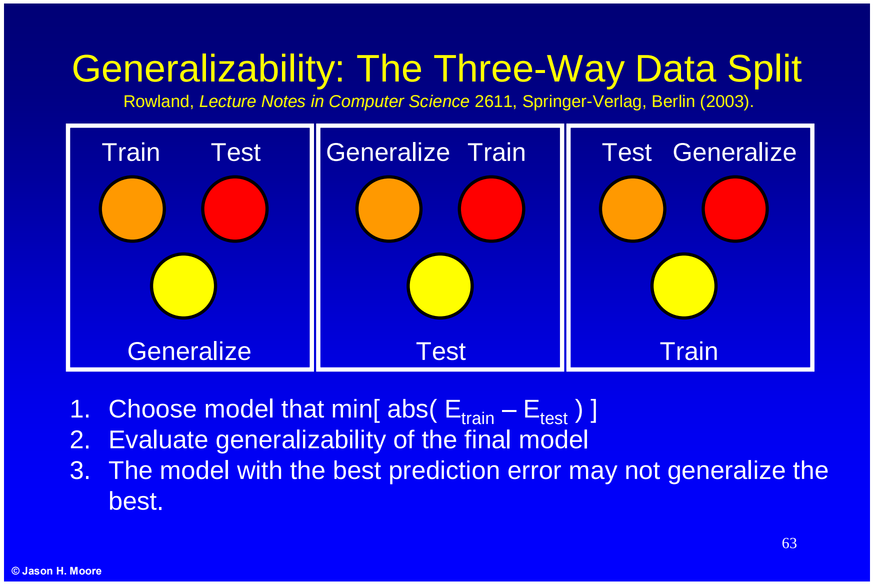
Data Interval	Genes Identified in Best Model
■■■■□	***A***,***C***,E
■■■□■	***A***,***C***,H
■■□■■	***A***,B,***C***,F
■□■■■	***A***,***C***,D
□■■■■	***A***,G,H

Significant Genes

A and C

62

© Jason H. Moore



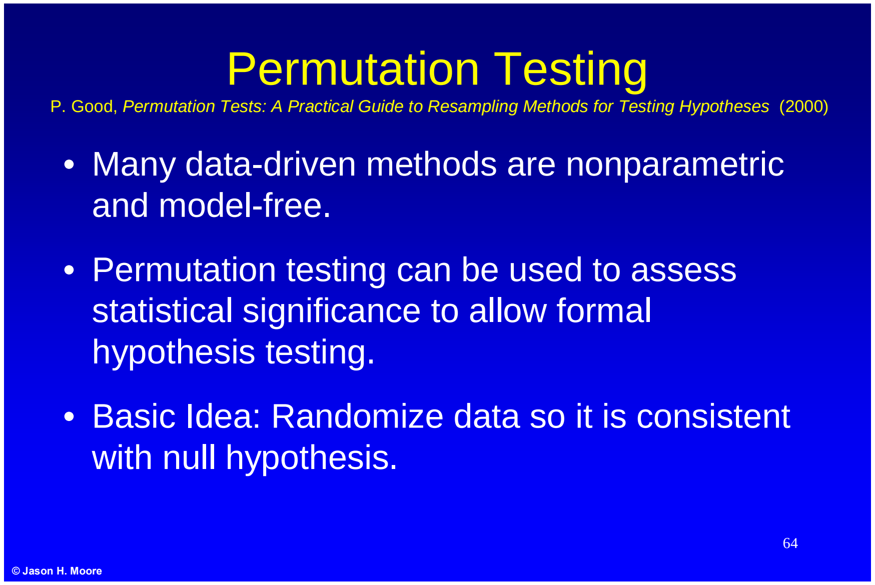
Generalizability: The Three-Way Data Split

Rowland, *Lecture Notes in Computer Science* 2611, Springer-Verlag, Berlin (2003).

1. Choose model that min[abs($E_{train} - E_{test}$)]
2. Evaluate generalizability of the final model
3. The model with the best prediction error may not generalize the best.

63

© Jason H. Moore



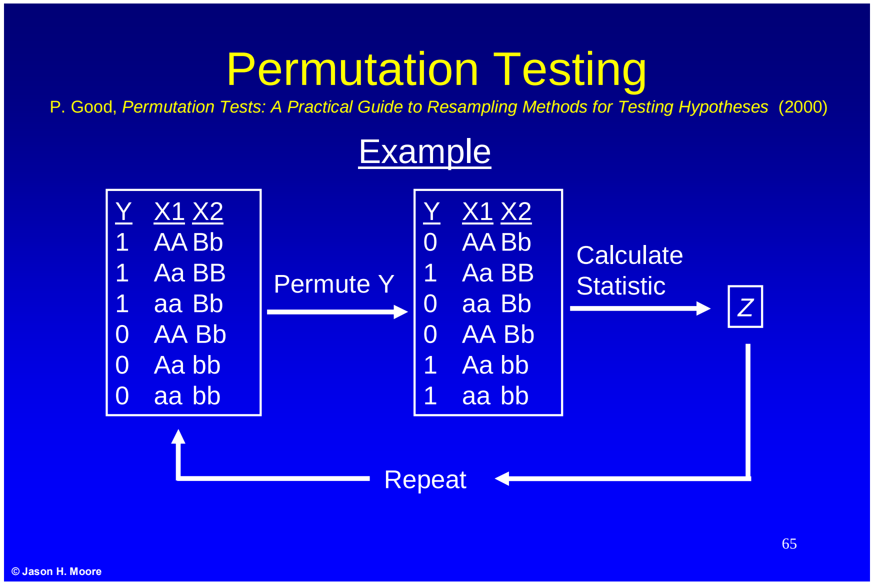
Permutation Testing

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.

64

© Jason H. Moore



Permutation Testing

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

Example

Y	X1	X2
1	AA	Bb
1	Aa	BB
1	aa	Bb
0	AA	Bb
0	Aa	bb
0	aa	bb

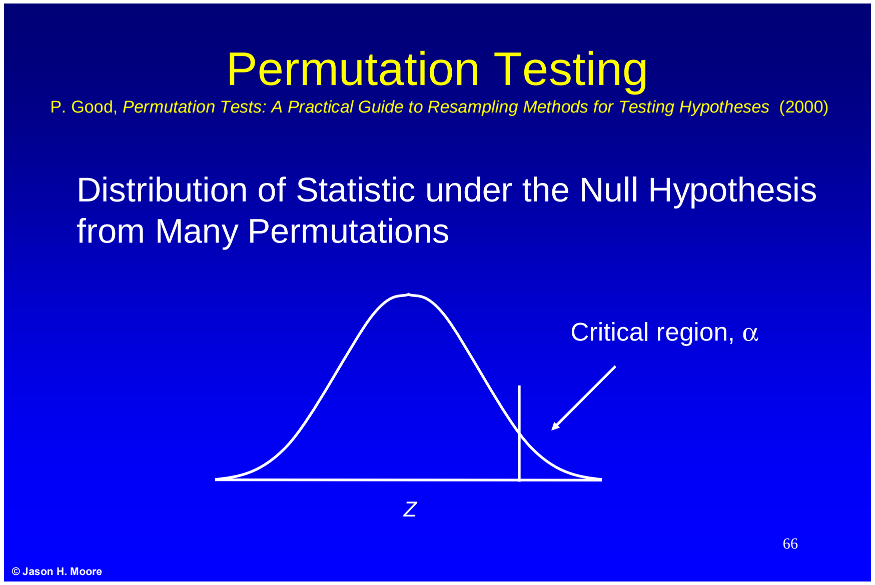
Permute Y

Y	X1	X2
0	AA	Bb
1	Aa	BB
0	aa	Bb
0	AA	Bb
1	Aa	bb
1	aa	bb

Calculate Statistic

Z

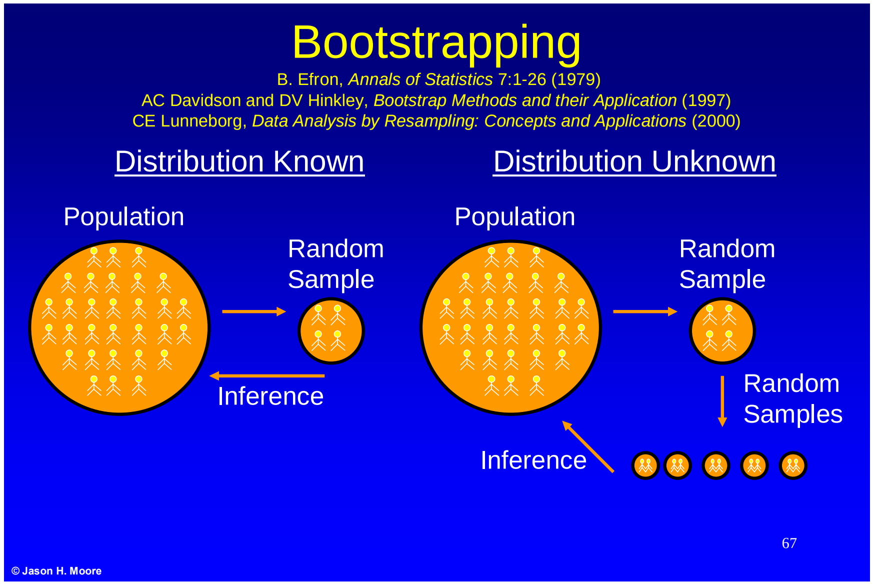
Repeat



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)

Distribution Known

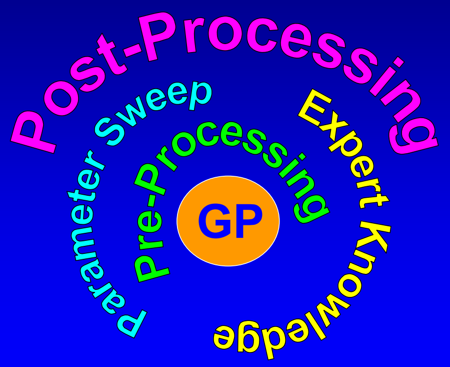
Distribution Unknown

Genetic Programming

Is GP an appropriate computational tool for solving complex biological problems??

Vanilla GP - NO!

Modern GP - YES!


Post-Processing

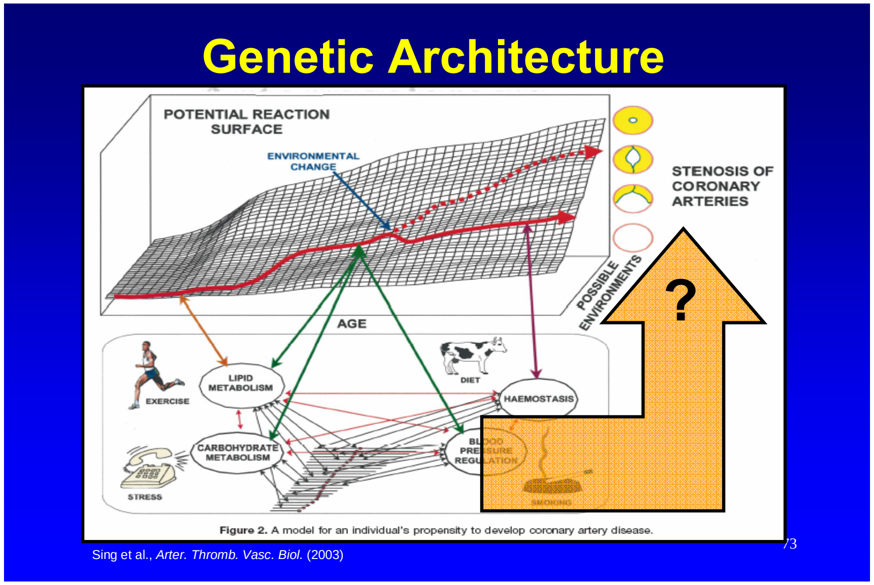
Parameter Sweep

Pre-Processing

GP

Expert Knowledge

Case Study: Symbolic Modeling

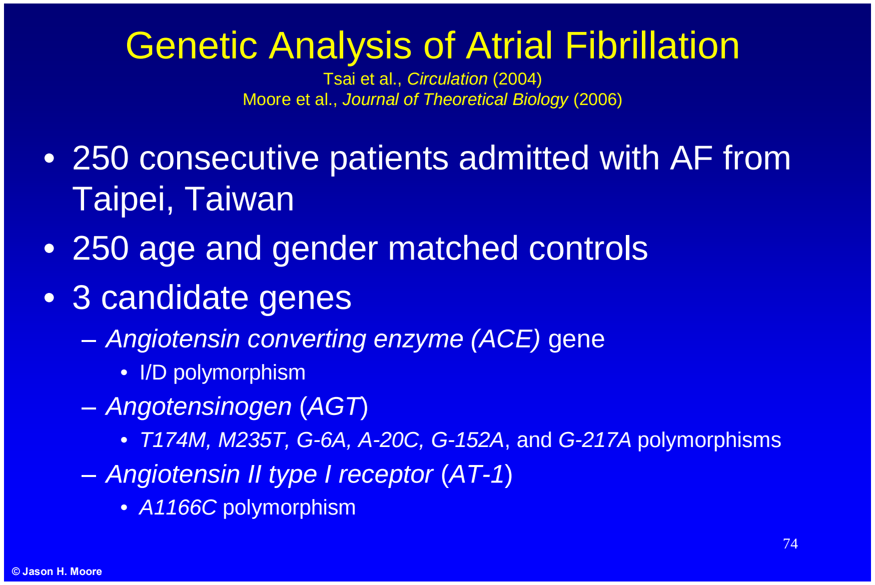


Genetic Architecture

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

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

73

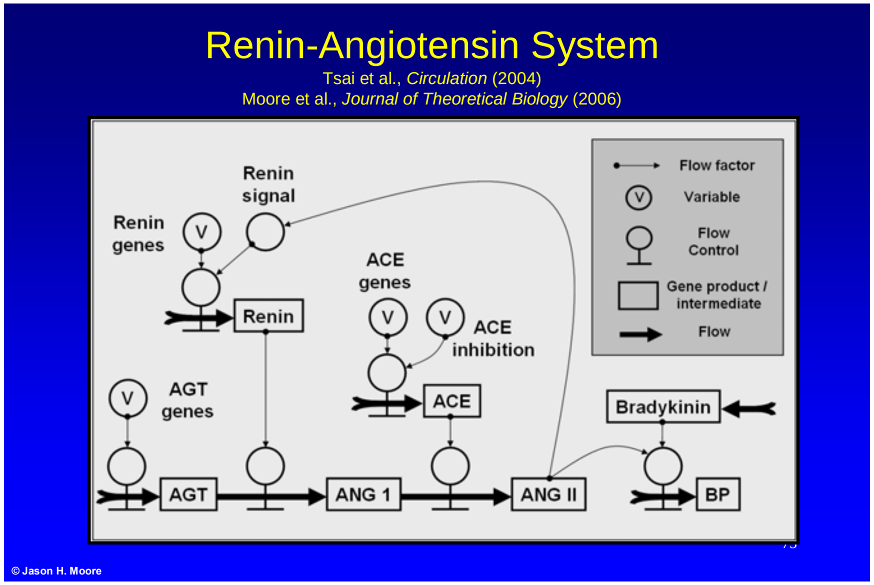


Genetic Analysis of Atrial Fibrillation

Tsai et al., *Circulation* (2004)
Moore et al., *Journal of Theoretical Biology* (2006)

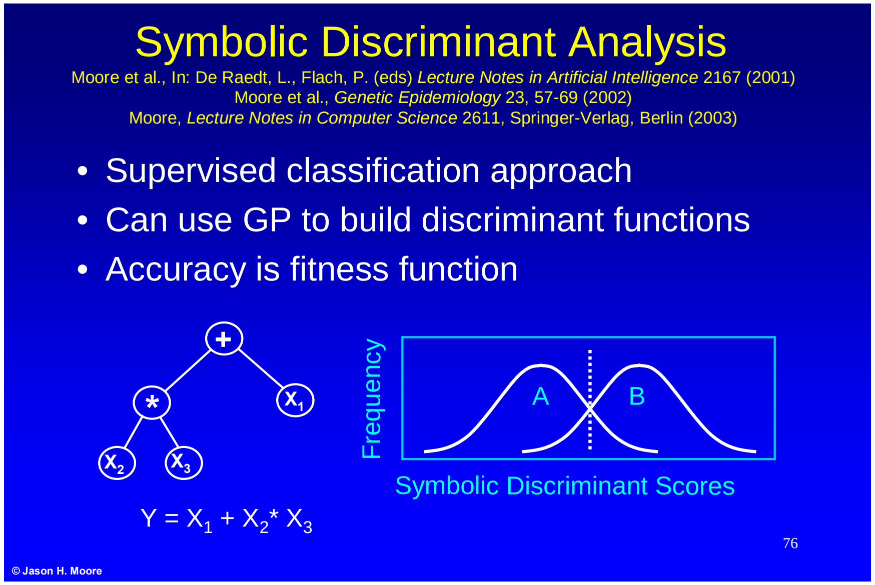
- 250 consecutive patients admitted with AF from Taipei, Taiwan
- 250 age and gender matched controls
- 3 candidate genes
  - *Angiotensin converting enzyme (ACE)* gene
    - I/D polymorphism
  - *Angotensinogen (AGT)*
    - *T174M, M235T, G-6A, A-20C, G-152A*, and *G-217A* polymorphisms
  - *Angiotensin II type I receptor (AT-1)*
    - *A1166C* polymorphism

74



Renin-Angiotensin System

Tsai et al., *Circulation* (2004)
Moore et al., *Journal of Theoretical Biology* (2006)



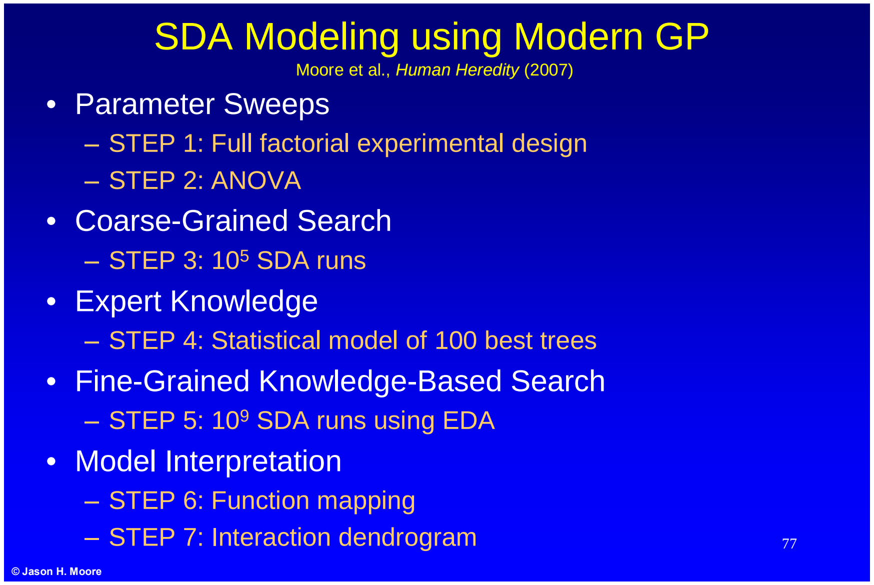
Symbolic Discriminant Analysis

Moore et al., In: De Raedt, L., Flach, P. (eds) *Lecture Notes in Artificial Intelligence* 2167 (2001)
Moore et al., *Genetic Epidemiology* 23, 57-69 (2002)
Moore, *Lecture Notes in Computer Science* 2611, Springer-Verlag, Berlin (2003)

- Supervised classification approach
- Can use GP to build discriminant functions
- Accuracy is fitness function

$Y = X_1 + X_2 * X_3$

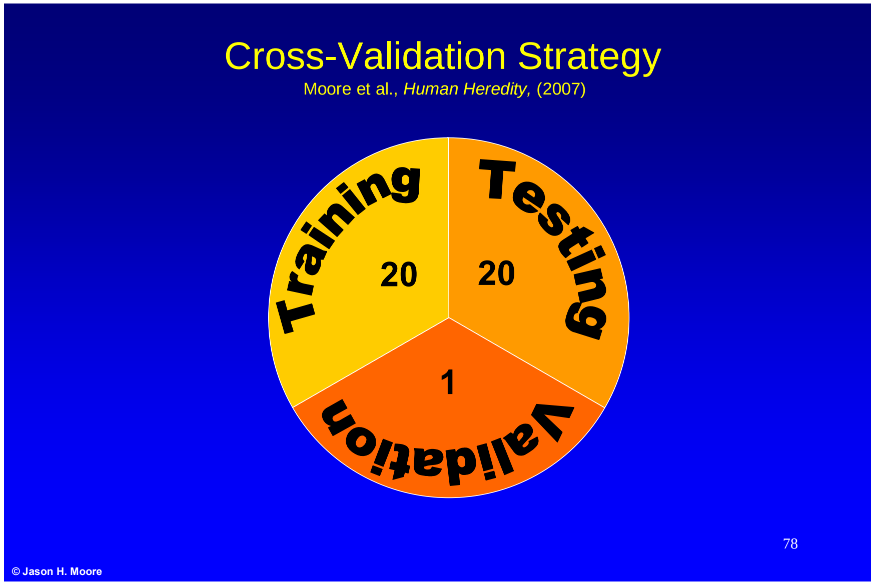
76



SDA Modeling using Modern GP

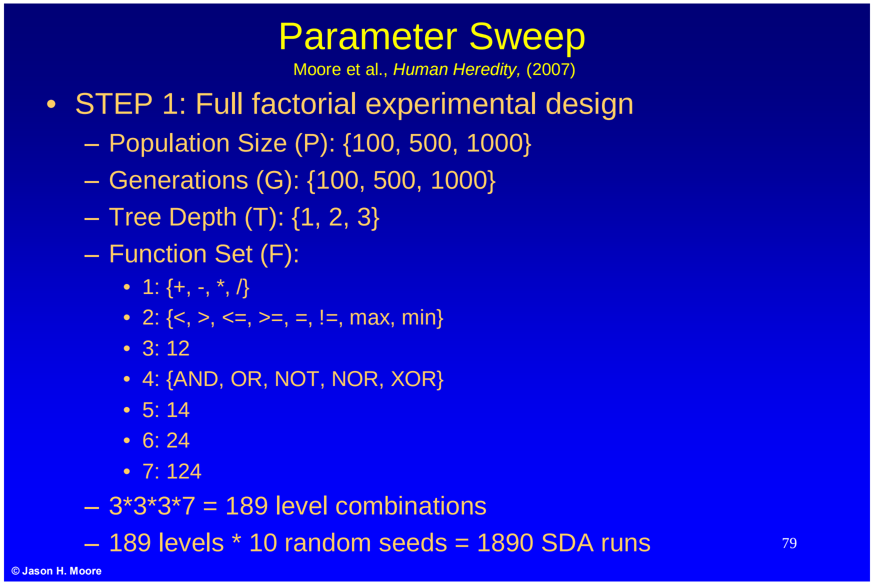
Moore et al., *Human Heredity* (2007)

- Parameter Sweeps
  - STEP 1: Full factorial experimental design
  - STEP 2: ANOVA
- Coarse-Grained Search
  - STEP 3: 10^5 SDA runs
- Expert Knowledge
  - STEP 4: Statistical model of 100 best trees
- Fine-Grained Knowledge-Based Search
  - STEP 5: 10^9 SDA runs using EDA
- Model Interpretation
  - STEP 6: Function mapping
  - STEP 7: Interaction dendrogram



Cross-Validation Strategy

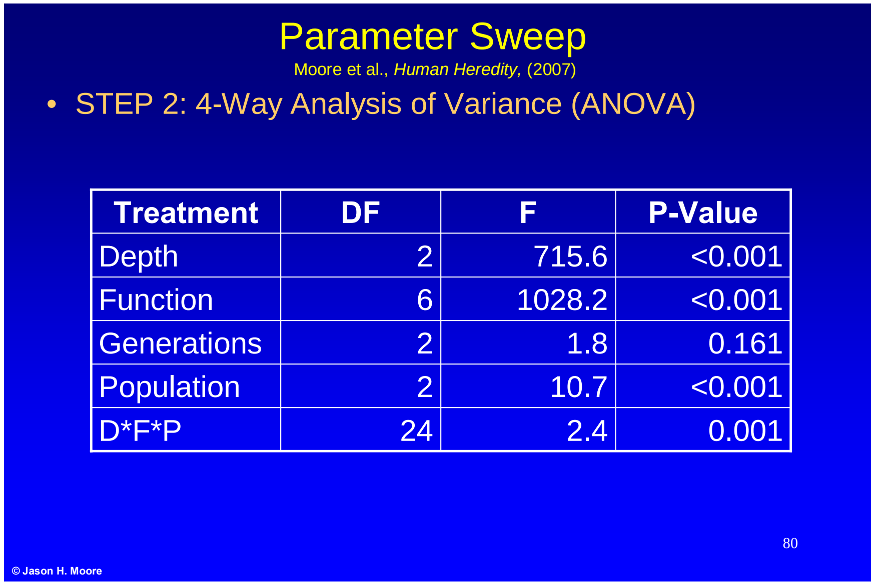
Moore et al., *Human Heredity,* (2007)



Parameter Sweep

Moore et al., *Human Heredity,* (2007)

- STEP 1: Full factorial experimental design
  - Population Size (P): {100, 500, 1000}
  - Generations (G): {100, 500, 1000}
  - Tree Depth (T): {1, 2, 3}
  - Function Set (F):
    - 1: {+, -, *, /}
    - 2: {<, >, <=, >=, =, !=, max, min}
    - 3: 12
    - 4: {AND, OR, NOT, NOR, XOR}
    - 5: 14
    - 6: 24
    - 7: 124
  - 3*3*3*7 = 189 level combinations
  - 189 levels * 10 random seeds = 1890 SDA runs

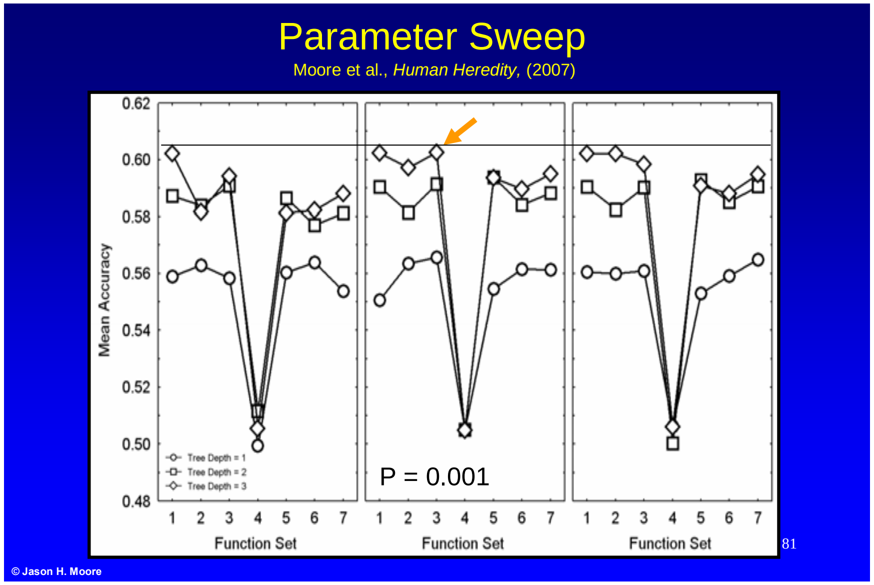


Parameter Sweep

Moore et al., *Human Heredity,* (2007)

- STEP 2: 4-Way Analysis of Variance (ANOVA)

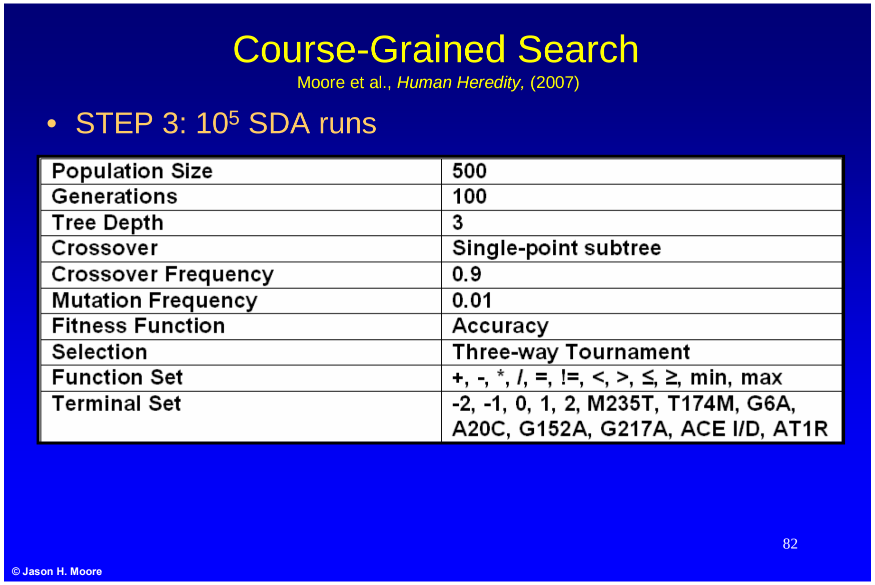
Treatment	DF	F	P-Value
Depth	2	715.6	<0.001
Function	6	1028.2	<0.001
Generations	2	1.8	0.161
Population	2	10.7	<0.001
D*F*P	24	2.4	0.001



Parameter Sweep

Moore et al., *Human Heredity,* (2007)

81



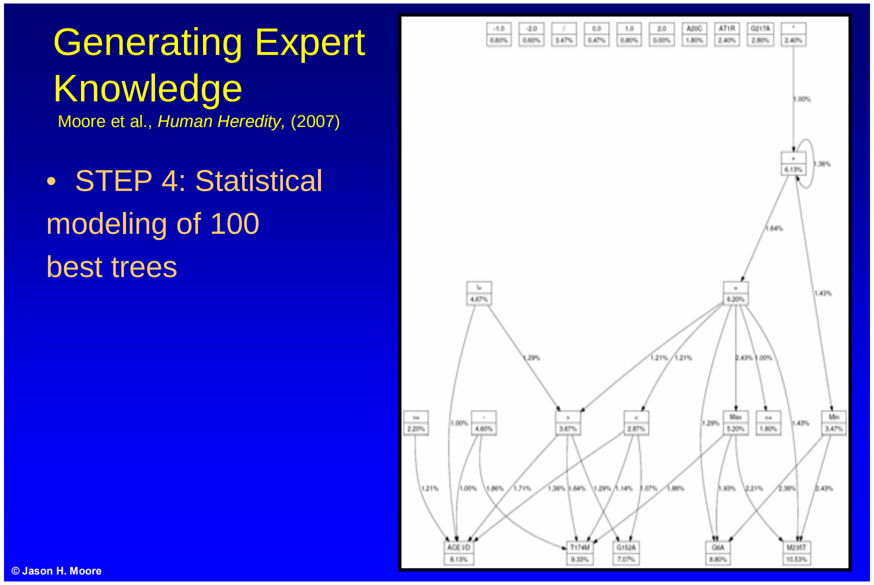
Course-Grained Search

Moore et al., *Human Heredity,* (2007)

- STEP 3: 10^5 SDA runs

Population Size	500
Generations	100
Tree Depth	3
Crossover	Single-point subtree
Crossover Frequency	0.9
Mutation Frequency	0.01
Fitness Function	Accuracy
Selection	Three-way Tournament
Function Set	+, -, *, /, =, !=, <, >, ≤, ≥, min, max
Terminal Set	-2, -1, 0, 1, 2, M235T, T174M, G6A, A20C, G152A, G217A, ACE I/D, AT1R

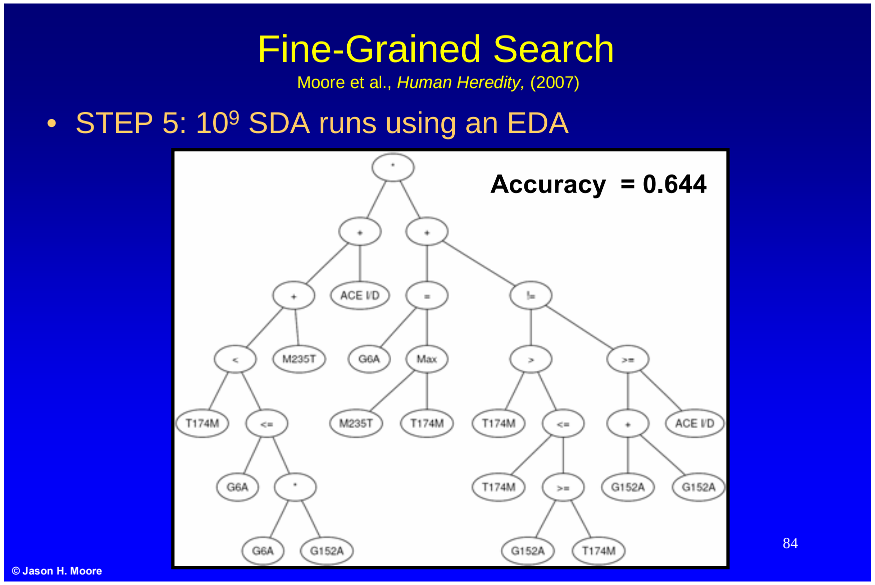
82



Generating Expert Knowledge

Moore et al., *Human Heredity,* (2007)

- STEP 4: Statistical modeling of 100 best trees



Fine-Grained Search

Moore et al., *Human Heredity,* (2007)

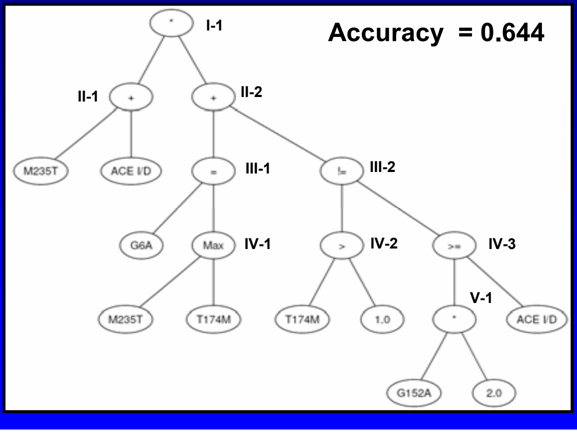
- STEP 5: 10^9 SDA runs using an EDA

84

Fine-Grained Search

Moore et al., *Human Heredity,* (2007)

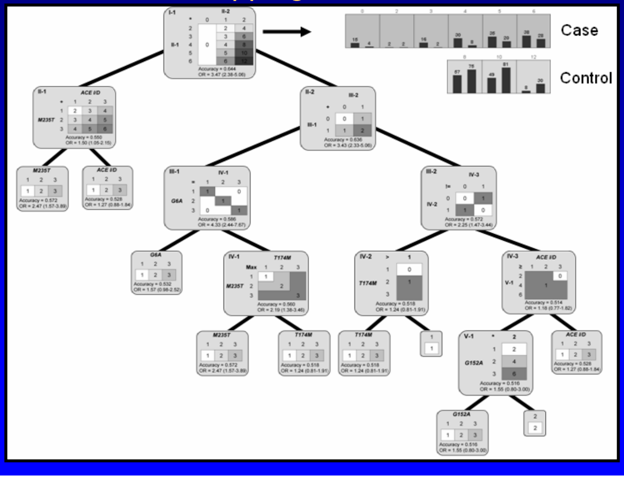
- STEP 5: 10^9 SDA runs using an EDA



Interpretation

Moore et al., *Human Heredity,* (2007)

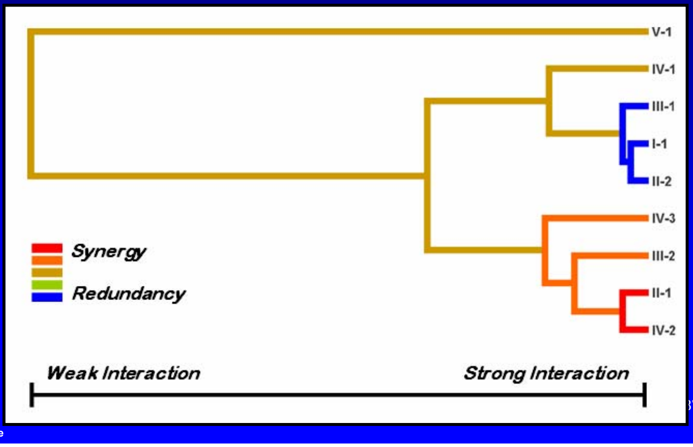
- STEP 6: Function Mapping



Interpretation

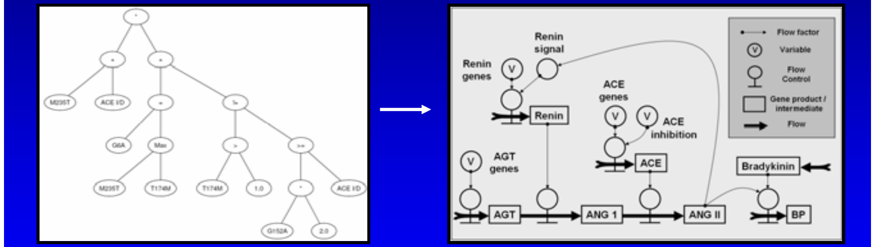
Moore et al., *Human Heredity,* (2007)
Moore et al., *Journal of Theoretical Biology* (2006)

- STEP 6: Interaction Dendrogram



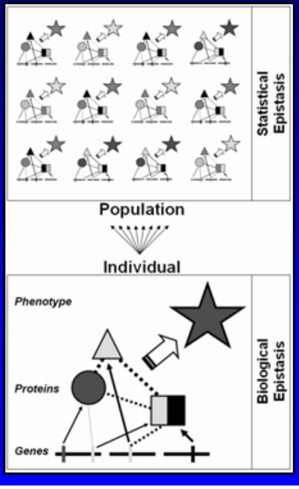
Interpretation

Moore et al., *Human Heredity,* (2007)

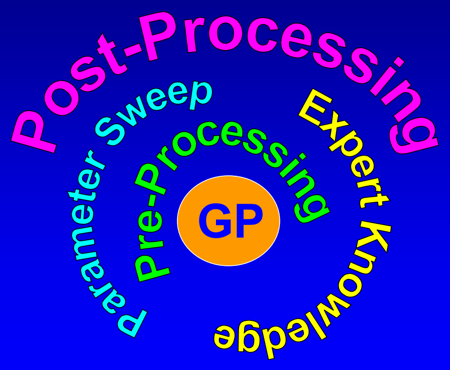


Interpretation

Moore, *Nature Genetics* (2005)
Moore and Williams, *BioEssays* (2005)



Modern GP - YES!


Post-Processing

Parameter Sweep

Pre-Processing

GP

Expert Knowledge

Additional Examples

- Moore, J.H., White, B.C. Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge. In: Genetic Programming Theory and Practice IV, in press, Springer (2006).
- Moore, J.H., White, B.C. Exploiting expert knowledge in genetic programming for genome-wide genetic analysis. Lecture Notes in Computer Science, in press (2006).
- Moore, J.H. Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics. In: Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data, IGI, in press (2006).

Acknowledgments

Nate Barney
Bill White

NIH R01 AI59694, LM009012

www.epistasis.org
compgen.blogspot.com
jason.h.moore@dartmouth.edu