Harvard-MIT Division of Health Sciences and Technology
HST.950J: Engineering Biomedical Information: From Bioinformatics to Biosurveillance
Course Directors: Dr. Isaac Kohane, Dr. Marco Ramoni

Children's Hospital Informatics Program
Harvard Medical School

Human Variations

Marco F Ramoni, PhD

September 15th, 2005

Biomedical Computing
6.872 / HST 950

Outline

Properties of the Genome

Basics

- 80s revolution and HGP;
- Genetic polymorphisms;
- Evolution and selection;

Genetic diseases

- Tracking genetic diseases;
- Traits and complex traits;

Genomic diseases

- Blocks of heredity;
- Tracking blocks.

The Genetic Study of the Future

Candidates identification

- Find the genes;
- Find the SNPs;

Study design

- Case/control studies;
- Pedigree studies;
- Trios, sibs and TDT;

Study analysis

- Single gene association;
- Multivariate association;
- Validation.

The Problem

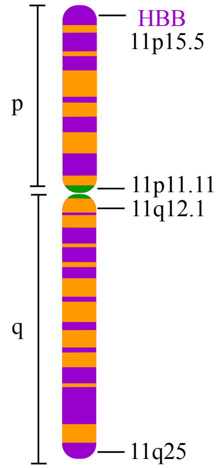
The context: Sickle cell anemia is a monogenic disorder due to a mutation on the β-globin (HBB) at 11p15.5.

The problem: SCA phenotype ranges from asymptomatic to early childhood death.

The phenotype: SCA subjects have an increased risk of stroke (6-8%) before 18 yrs.

The hypothesis: Other genes modulate this risk of stroke.

Chromosome 11



HBB Sequence in Normal Adult Hemoglobin (Hb A):

Nucleotide	CTG	ACT	CCT	GAG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Glu	Glu	Lys	Ser
	3			6			9

HBB Sequence in Mutant Adult Hemoglobin (Hb S):

Nucleotide	CTG	ACT	CCT	GTG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Val	Glu	Lys	Ser
	3			6			9

Figure by MIT OCW.

Finding Candidate Genes

Rationale: Bar a genome-wide scan you need likely culprits.

Start: OMIM (NCBI/NIH)

Extend:

- ✓ Literature;
- ✓ Regions;
- ✓ Microsatellites;
- ✓ Mechanisms of actions (pathways);

Refinement: Cast a large net and run a wide scan on a subset of patients.

See the OMIM, Online Mendelian Inheritance in Man.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

Finding The Right SNPs

Option 1. Finding the causative SNP:

- Rationale: Find the cause, select if there is a functional role.
- Drawback: What is functional? Exons, promoter, splicing, etc.

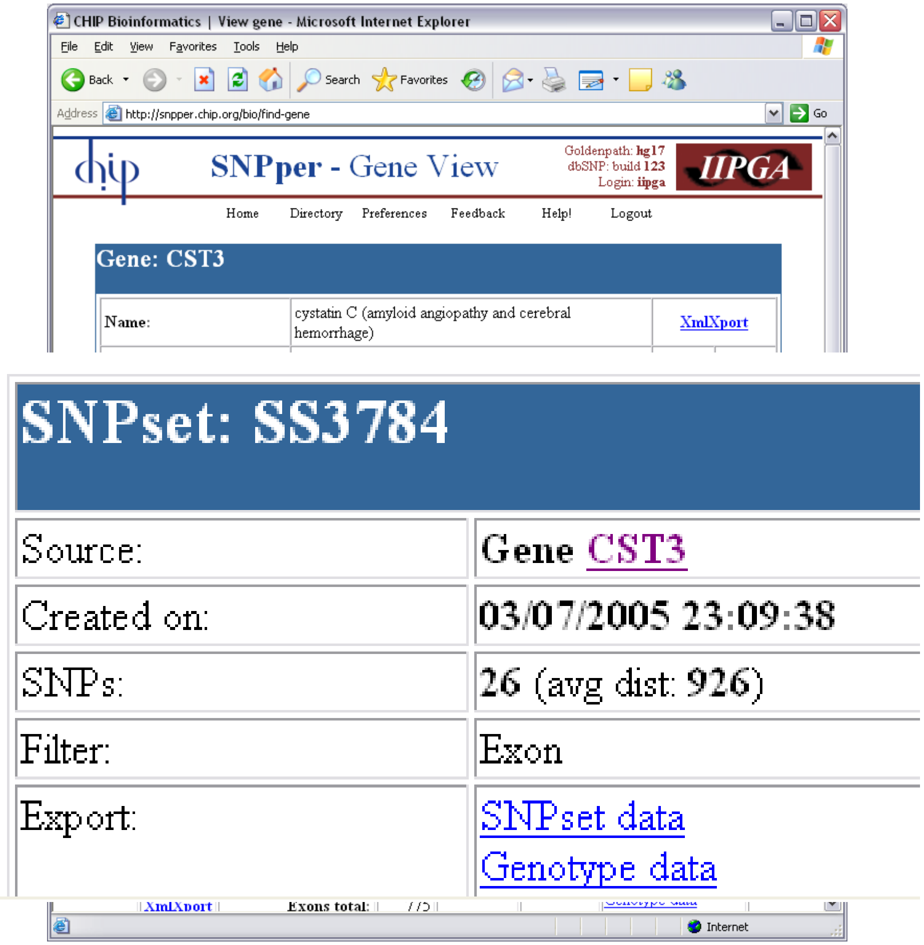
Option 2. Finding related SNPs:

- Rationale: Chose SNPs that represent the gene through LD.
- Drawback: Tough to get the causative root.

Figure removed due to copyright reasons.

Hunting Causative SNPs

- Strategy: Select the SNPs on the basis of their role.
- Options: Non synonymous, in exons, in promoter, in other regulatory region.
- Source: dbSNP (NCBI/NIH).
- Faster: Portal SNPPER.
- Bonus: Primer design.
- Example: Select all the SNPs in CST3 located on exons.
- Filtering: From 146 to 26.
- Problem: Uncovered regions.


CHIP Bioinformatics | View gene - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://snpper.chip.org/bio/find-gene
Go
chip
SNPper - Gene View
Goldenpath: hg17
dbSNP: build 123
Login: iipga
IIPGA
Home Directory Preferences Feedback Help! Logout
Internet

Gene: CST3

Name:	cystatin C (amyloid angiopathy and cerebral hemorrhage)	XmlXport

SNPset: SS3784

Source:	Gene CST3
Created on:	03/07/2005 23:09:38
SNPs:	26 (avg dist: 926)
Filter:	Exon
Export:	SNPset data Genotype data

XmlXport Exons total: 7/5

Courtesy of Dr. Alberto A. Riva. Used with permission.

Fishing Across Genes

Rationale: Find the optimal coverage for the entire gene.

Problem: We need to know how SNPs are transmitted together in the population.

Source: HapMap.org

Hapmap: Genotype of 30 trios in 4 populations every 5k bases.

Strategy: 1) Identify blocks of LD and 2) Choose the SNPs that represent these blocks.

Figure removed due to copyright reasons.

Genome Wide Scan

- Technologies for genotyping:
- By SNP (individual primer);
- By Sample/Locus;
- Genome-wide: GeneChip® Mapping 100K Set (soon 500k) using a technology similar to expression arrays.
- 500k means 1 SNP every 6, close to the resolution of the HapMap.

Figure removed due to copyright reasons.

Study Design

- Classification by sampling strategy:
  - Association: Unrelated subjects with/out phenotype.
    - Case/Control: Two sets of subjects, with and without.
    - Cohort: Natural emergent phenotype from study.
  - Pedigrees: Traditional studies focused on heredity.
    - Large pedigree: One family across generations.
    - Triads: Sets of nuclear families (parents/child).
    - Sib-pairs: Sets of pair of siblings.
- Classification by experimental strategy:
  - Double sided: Case/control studies.
  - Single sided: e.g trios of affected children.

Analysis Methods

- Study designs and analysis methods interact.
- We review five main analysis types:
  - Association studies: Case/control association.
  - Linkage analysis: Traditional analysis of pedigrees.
  - Allele-sharing: Find patterns better than random.
  - TDT: transmission disequilibrium test.
- Typically, these collections are hypothesis driven.
- The challenge is to collect data so that the resulting analysis will have enough power.

Association Studies

Method: Parametric method of association.

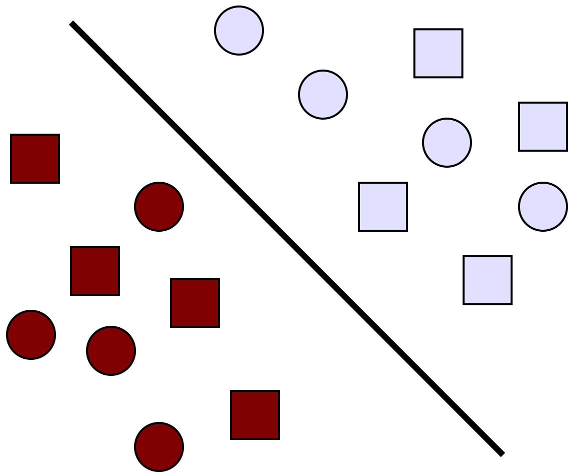
Strategy: Traditional case vs control approach.

Test: Various tests of association.

Sample: Split group of affected/unaffected individuals.

Caveats: Risk of stratifications (admixtures) - case/control split by populations.

Advantages: Easily extended to complex traits and ideal for exploratory studies.



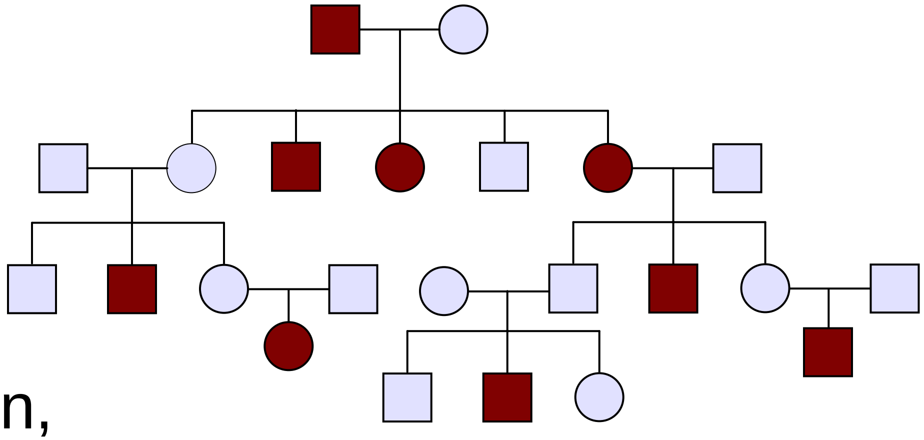
Linkage Analysis

Method: Parametric model building.

Strategy: Compare a model with dependency between phenotype and allele against independence model.

Test: Likelihood ratio - or lod score log(LR).

$$LR = \frac{p(Data \mid M_1)}{p(Data \mid M_0)}$$



Sample: Large pedigree or multiple pedigrees.

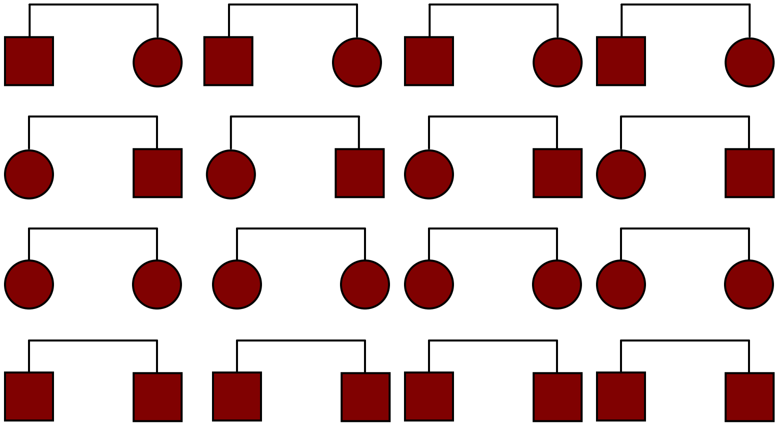
Caveats: Multiple comparison, hard for complex traits.

Allele Sharing

Method: Non parametric method to assess linkage.

Test: An allele is transmitted in affected individuals more than it would be expected by chance.

Sample: It uses affected relatives in a pedigree, counts how many times a region is identical-by-descent (IBD) from a common ancestor, and compares this with expected value at random.



Caveats: Weak test, large samples required.

Transmission Disequilibrium Test

Method: Track alleles from parents to affected children.

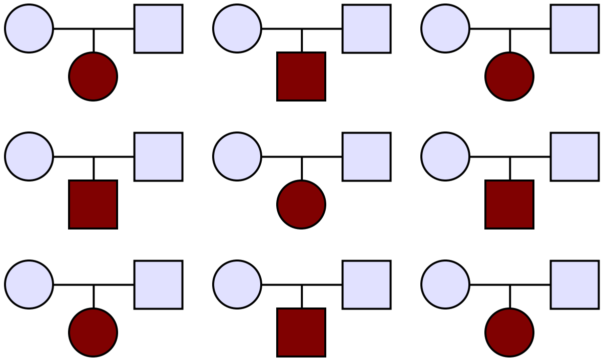
Strategy: Transmitted=case / non transmitted=controls.

Test: Transmission disequilibrium test (TDT).

Sample: Triads of affected child and parents.

Caveats: Test is not efficient and is prone to false negatives.

Advantages: Powerful test and stratification not an issue.



Stroke Study Design

Design: Nation-wide cohort study of over 4000 African American in 26 centers.

Subjects: 1392 SCA subjects with at least one complication from SCA (92 with stroke, 6.2%).

Genes: 80 candidate genes involved in vaso-regulation, inflammation, cell adhesion, coagulation, hemostasis, proliferation, oxidative biology and other functions.

SNPs: Coverage selected with bias to function (256).

Risk factors: α-thalassemia, history, age, gender.

Filtering: Missing data and Hardy-Weinberg on unaffected reduces the set to 108 SNPs on 80 genes.

Single Gene Association

Method: One SNP at the time.

Analysis: Test statistics (like we had an hypothesis).

Style: Observational by pseudo hypothesis-driven.

Results: A list of SNP/genes.

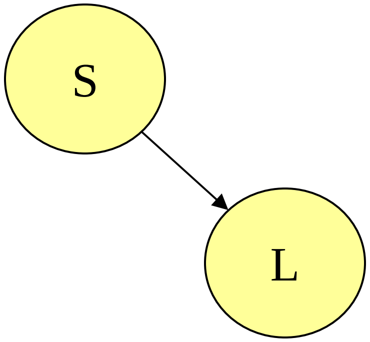
Validation: Replication.

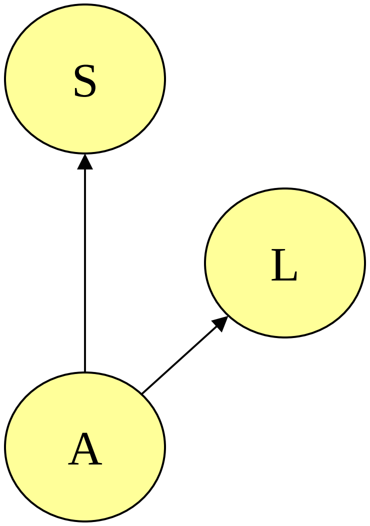
Table removed due to copyright reasons.
Please see:

Table 2 in Hoppe, C., et al. "Gene interactions and stroke risk in children with sickle cell anemia." *Blood* 103, no. 6 (Mar 15, 2004): 2391-6. *Epub* (Nov 13, 2003.)

Spurious Association/Confounding

- Association of shoe size (S) and literacy (L) in kids.
- If I act on S, I will not change L: If you buy bigger shoes, will your kids learn more words?
- No: age (A) make S and L conditionally independent.

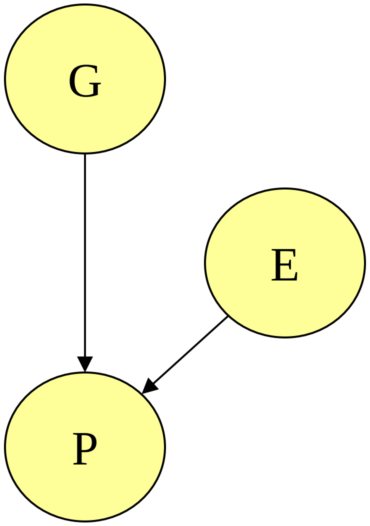




Missed Associations

Gene environment interaction:

No association between genotype and phenotype



Association appears conditional on an environmental factor

Bayesian Networks

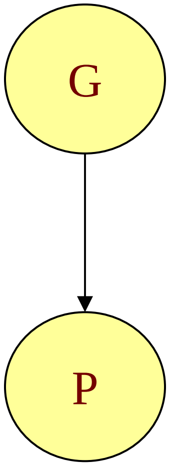
Definition: Direct acyclic graph (DAG) encoding conditional independence/dependence.

Qualitative:

- Node: stochastic variables (SNPs, phenotypes, etc).
- Arcs: Directed stochastic dependencies between parents and children.

Quantitative:

- CPT: Conditional probability tables (distributions) that shape the dependency.



G		
AA	*Aa*	*aa*
0.3	0.6	0.1

G	P	
	True	*False*
AA	0.3	0.7
Aa	0.5	0.5
aa	0.9	0.1

Learning Networks

Processes: Data are generated by processes.

Probability: The set of all models is a stochastic variable $\mathcal{M}$ with a probability distribution $p(\mathcal{M})$.

Selection: Find the most probable model given the data.

$$p(M \mid \Delta) = \frac{p(\Delta, M)}{p(\Delta)} = \frac{p(\Delta \mid M)\, p(M)}{p(\Delta)}$$

Estimation: Probabilities can be seen as relative frequencies:

$$p(x_j \mid \pi_i) = \frac{n(x_j \mid \pi_i)}{\sum_j n(x_j \mid \pi_i)} \qquad p(x_j \mid \pi_i) = \frac{a_{ij} + n(x_j \mid \pi_i)}{\sum_j a_{ij} + n(x_j \mid \pi_i)}$$

Figure removed due to copyright reasons.

Prognostic Modeling

Prediction: The method used for the predictive validation can be used to compute the risk of stroke given a patient's genotypes.

Prognosis: We can build tables of risks for patients and predict the occurrence of stroke in 5 years.

Extension: How about this risk scheme as a model of stroke in the general population?

Risk	**ANXA2.6**	**BMP6.10**	**BMP6.12**	**SELP.14**	**TGFBR3.10**	**ERG.2**	**N**
	hCV26910500	*rs267196*	*rs408505*	*rs3917733*	*rs284875*	*rs989554*	
0.007 (0;0.03)	AG	TT	TT	CT	CT	AG	1
0.06 (0;0.38)	AG	TT	TT	CT	CC	AG	4
0.185 (0.09;0.30)	AA	TT	CT	CC	CC	AA	50
0.727 (0.61;0.83)	AA	TT	CC	CC	CC	AA	64
0.868 (0.70;0.97)	GG	TT	CC	CC	CC	AA	21
0.968 (0.79;1)	GG	TT	CC	CT	CC	AA	8

Predictive Validation

Cross Validation: 98.8%.

Validation: Stroke prediction of 114 subjects in different population (not the cohort).

Accuracy: 98.2%: TPR=100%; TNR=98.1% (2 errors).

Logistic regression: Identify regressors at p-value < 0.05.

Model: 5 (SELP/BMP6) & HbF.

Accuracy: 88% accurate: TPR: 0.57% (3 errors); TNR: 0.9% (10 errors).

Figure removed due to copyright reasons.

A Holistic System

- Why we do not find the causes for complex traits?
- Because we look at one gene at the time.
- Genes work together (need more than one gene to get the phenotype) but also in a redundant way (phenotype through alternative paths).
- Long distance disequilibrium, reveals more complex structures in the population.
- Prediction is necessary.

Gene Symbol	**Position**	**Single Gene**	
		Accuracy	*Cont*
ADCY9	16p13.3	71.93%	2%
ANXA2	15q22.2	43.86%	2%
BMP6	6p24.3	83.33%	5%
CSF2	5q23.3	50.88%	1%
ECE1	1p36.12	13.15%	0.2%
ERG	21q22.2	42.98%	1%
MET	7q31.2	23.68%	1%
SCYA	17q11.2	55.14%	1%
SELP	1q24.2	80.70%	7%
TEK	9p21.2	8%	1%
TGFBR3	1p22.1	50.88%	2%
HbF.P		72.81%	1%

Take Home Messages

There are two types of science:
physics and stamp collecting.
Ernest Rutherford

Revolution: The –omic scale changes the way of biomedical sciences, makes it predictive/quantitative.

Discovery: The genome is too complex for simple hypothesis, hypotheses have to be discovered.

Proof: The burden of proof has to be based on prediction, as we expect from good science.

Potential: The potential of this changes goes beyond the still fantastic power to understand and heal.