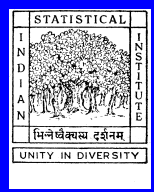
MODEL-FREE LINKAGE AND ASSOCIATION MAPPING OF COMPLEX TRAITS USING QUANTITATIVE ENDOPHENOTYPES

Saurabh Ghosh

Human Genetics Unit

Indian Statistical Institute, Kolkata



Most common diseases are caused by a combination of genes and environment

Stroke

Breast cancer

Diabetes

Obesity

Inflammatory Bowel Disease

Manic-depression

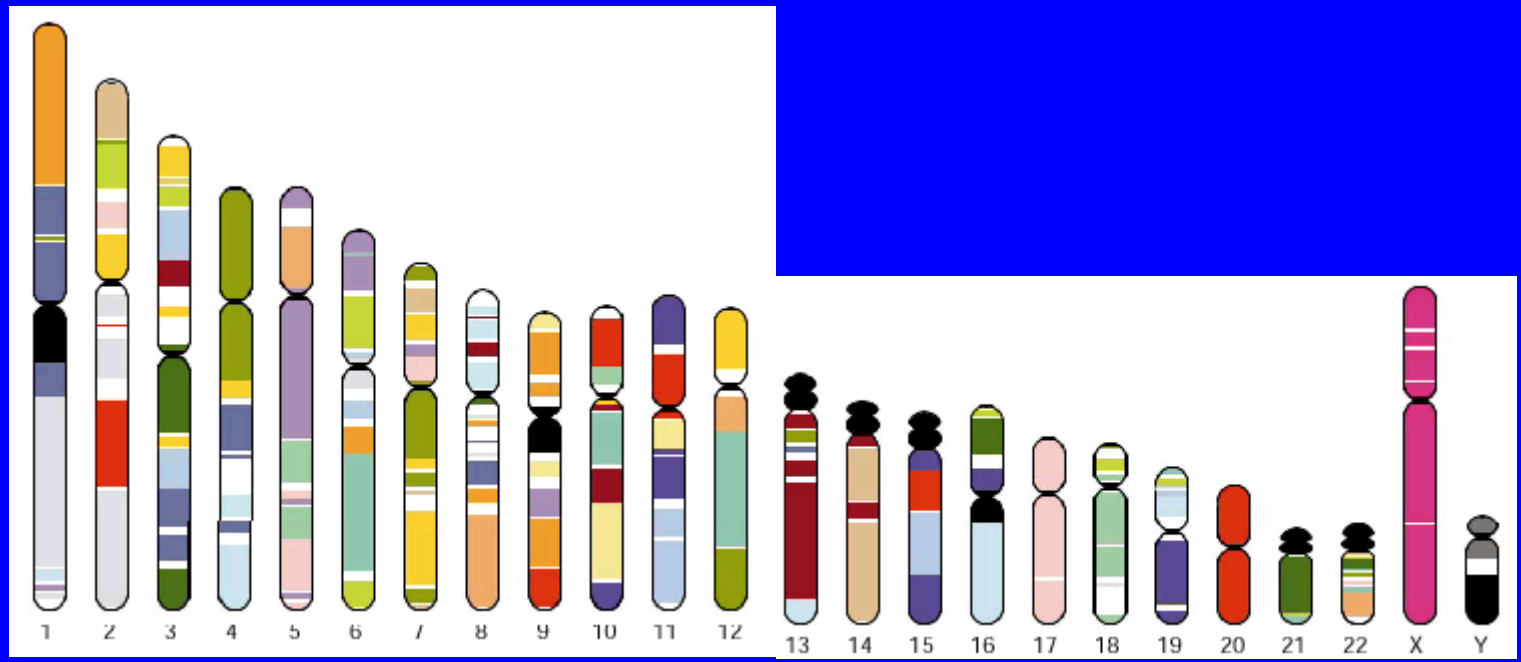
Myocardial Infarction

Hypertension

High Cholesterol

Schizophrenia

Where are those genes?



QUANTITATIVE TRAITS and COMPLEX TRAITS

Complex traits are controlled by multiple loci, some with minor gene effects, and genetic variation at any one locus does not completely determine the trait.

Examples: diabetes, coronary artery disorder, schizophrenia.

Such traits are usually binary in nature.

Heritable quantitative characters, possibly correlated, generally are precursors of complex traits.

Example: Blood pressure and total cholesterol are precursors of coronary artery disease.

WHY STUDY QUANTITATIVE TRAITS?

To study a complex trait.

Often genetic studies of a binary trait (e.g., hypertension) use dichotomization of a correlated quantitative variable (e.g., s.b.p) based on some pre-determined threshold(s) [e.g., s.b.p > 140].

This leads to loss of information on trait variability and statistical power.

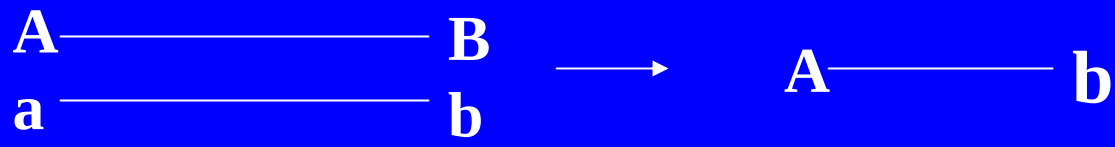
For genetic analysis of a binary complex trait, it may be more powerful to consider a quantitative trait correlated with the end-point trait.

Genes found to control the quantitative trait may also be involved in the pathogenesis of the complex trait.

THE TWO PARADIGMS OF GENE-MAPPING

1. LINKAGE

measured in terms of RECOMBINATION FRACTION (θ), the probability of a crossover between two loci.



From biological considerations, $0 \leq \theta \leq 0.5$. $\theta=0.5$ implies that two loci are unlinked, $\theta=0$ implies complete linkage.

2. ASSOCIATION

measured in terms of LINKAGE DISEQUILIBRIUM (δ), the deviation of probability of joint occurrence of alleles at two-locus from random occurrence of the alleles.

$$\delta = P(AB) - P(A)P(B)$$

LINKAGE ANALYSIS OF QUANTITATIVE TRAITS

AIM: make inferences about the recombination fraction between a hypothetical QTL and a marker locus.

Is the QTL "physically close" to the marker locus?

For a sib-pair, the QTL paradigm is that if the quantitative trait values of the sibs are close to each other, the sibs have "similar" allelic composition at the QTL. If their trait values differ significantly, the allelic compositions at the QTL are different.

If the marker locus is LINKED (physically close) to the QTL, this property will also be exhibited by the marker locus. If the marker locus is UNLINKED, there will NOT be any relation between the trait values and allelic compositions.

THE BASIC QTL MODEL

A quantitative trait Y is controlled by an autosomal, biallelic QTL with alleles (A,a).

$E(Y|AA) = \alpha$ $\quad$ $Var(Y|AA) = \sigma^2$

$E(Y|Aa) = \beta$ $\quad$ $Var(Y|Aa) = \sigma^2$

$E(Y|aa) = -\alpha$ $\quad$ $Var(Y|aa) = \sigma^2$

If $\beta = 0$, we say that Y has no dominance effect at the QTL.

π_t is the i.b.d. (identity-by-descent) score at the QTL (unobservable).

π_m is the i.b.d. score at the marker (sometimes can be exactly computed, often necessary to be estimated as P_m).

AIM: to find relationship between "suitable" functions of Y and P_m for relative-pairs.

POPULAR QTL MAPPING METHODS

1. **VARIANCE COMPONENTS METHODS**
   (Amos et al. 1990, Almasy and Blangero 1998)

Random Effects Model: $Y = \Sigma G_i + E$; $i=1,2,\ldots,k$

where G_i is the genotypic effect of the i^{th} QTL assumed to have a normal distribution;

E is the environmental effect with a normal distribution.

Data: Quantitative traits and marker genotypes of pedigrees.

Quantitative traits within a pedigree are assumed to follow a **MULTIVARIATE NORMAL** DISTRIBUTION.

Test for linkage at the i^{th} QTL is a likelihood-ratio test for the variance of $G_i = 0$ versus being positive.

2. HASEMAN-ELSTON APPROACH AND EXTENSIONS

(Haseman and Elston 1972,
Extensions: Amos et al. 1989, Wijsman and Olson 1993, Olson 1995, Tiwari and Elston 1997, Elston et al. 2000……)

Data: quantitative trait values and marker genotypes of independent sib-pairs (and preferably marker genotypes of both parents).

The regression equation:

$$E(Z|P_m) = a + b\,P_m$$

where Z = sib-pair squared difference in trait values

$\beta = 0$ (no dominance at the QTL)

$b=0 \Leftrightarrow \theta = 0.5$ (unlinked) and $b < 0 \Leftrightarrow \theta < 0.5$ (linked)

Testing for b=0 versus b < 0: distribution-free approach

A NON-PARAMETRIC ALTERNATIVE

(Ghosh and Majumder 2000, Ghosh et al. 2003)

Non-parametric regression instead of Haseman-Elston linear regression:

$$Z = \Psi(P_m) + e$$

where Ψ is a real-valued function.

Ψ is estimated using kernel smoothing :

$$\Psi(x_j) = [\Sigma_i \kappa(\{x_i - x_j\}/h) y_i] / \Sigma_i \kappa(\{x_i - x_j\}/h)$$

The kernel function used:

$$\kappa(x) = 0.75\,(1-x^2)\, I_{|x|<1}$$

Optimal *h* selected using leave-out-one cross validation.

The test statistic:

Δ = 1 - (residual sum of squares in regression / total variation in Z)

Analogous to R^2 in linear regression.

Test for linkage is equivalent to $\Delta=0$ versus $\Delta > 0$.

Difficult to obtain asymptotic distribution of Δ .

Use permutations to obtain empirical p-value of observed Δ .

CAVEAT: Δ does not take into account direction of relationship between Z and P_m. Under null, random positive relationship will inflate rate of false positives. Test additionally for correlation between Z and P_m whenever Δ significant.

(200 sib-pairs, α=3, σ=1, p=0.5, ρ=0.5, θ=0.01)

		POWER		
Model	β	HE (sq diff)	HE (cr pr)	KS
Normal	0	0.899	0.912	0.878
	1	0.825	0.838	0.852
	2	0.685	0.702	0.769
Chi-sq	0	0.890	0.902	0.876
	1	0.817	0.823	0.849
	2	0.651	0.657	0.766
Poisson	0	0.872	0.881	0.875
	1	0.805	0.802	0.850
	2	0.613	0.604	0.764

EXTENSION OF HE TO SIBSHIP DATA

(Ghosh and Reich 2002)

Data: Quantitative traits $Y = (y_1, y_2, \ldots)$ on sibships of size ≥ 2

Contrast function: $\Sigma\, c_i y_i$ $(= c'Y)$ where $\Sigma\, c_i$ $(= c'1) = 0$

Quadratic function of marker i.b.d. scores: $c'\, \Pi_m c$
where Π_m is the matrix of i.b.d. scores with $(i,j)^{th}$ element the i.b.d. score of the i^{th} and j^{th} sibs $(i \neq j)$ and diagonals 0.

NOTE:
i.b.d. scores of all possible sib-pairs in a sibship are not independent. Given a sibship of size n, there exists an independent set of (n-1) pairwise i.b.d. scores. For example, $\{\pi_{12}, \pi_{13}, \ldots, \pi_{1n}\}$ is an independent set of i.b.d. scores. Given this set, all other pairwise i.b.d. scores are dependent.

Why choose the contrast function?

1. The classical HE can be viewed as a special case of a contrast function for n=2 with c = (1 -1).
2. The contrast function corresponds to the second principal component of quantitative trait values.

The corresponding reqression equation:

$$E\{(c'Y)^2 | c' \Pi_m c\} = a + b\, c' \Pi_m c$$

where there is no dominance at the QTL,

$$b=0 \Leftrightarrow \theta = 0.5 \text{ and } b > 0 \Leftrightarrow \theta < 0.5$$

Choice of c : {1,-1/(n-1),…,-1/(n-1)}

Consistent with HE.

Higher coefficient associated with the independent set of i.b.d. scores $\{\pi_{12},\pi_{13},\ldots,\pi_{1n}\}$ and lower coefficient with other pairs.

COMBINING CONTRAST AND MEAN FUNCTIONS

Mean value of quantitative values correspond to first PC.

$$E(\bar{y}^2 | 1' \Pi_m 1) = a + b\,(1' \Pi_m 1)/n^2$$

where b is same as in the contrast function equation.

Mean function and contrast function are uncorrelated.

Combined least squares minimization:

$$\Sigma_j\{u_j - \beta_0 - \beta_1 x_j\}^2 + \Sigma_j\{v_j - \beta_2 - \beta_1 z_j\}^2$$

u_j, v_j: sq contrast and mean functions; x_j,y_j: i.b.d. regressors

Test for linkage based on least squares estimate of β_1.

Alternative: Combined non-parametric regressions of u_j on x_j and y_j on z_j using kernel smoothing.

(200 sibships with 4 sibs each, α=3, σ=1, p=0.5, ρ=0.5, θ=0.01)

		POWER		
Model	β	KS	HE (cr pr)	CM
Normal	0	0.912	0.918	0.931
	1	0.875	0.860	0.868
	2	0.798	0.742	0.749
Chi-sq	0	0.903	0.909	0.915
	1	0.856	0.832	0.836
	2	0.754	0.717	0.711
Poisson	0	0.880	0.887	0.893
	1	0.826	0.811	0.813
	2	0.735	0.661	0.651

MULTIVARIATE PHENOTYPES

Complex traits are characterized by correlated quantitative variables constituting a multivariate phenotype.

The end-point trait is usually binary in nature. Since quantitative traits carry more information on variation within genotypes, it may be more prudent to use a vector of correlated phenotypes for linkage analysis of the complex trait.

Analyzing individual components of a multivariate phenotype vector separately leads to the statistical problem of multiple comparisons.

Analyzing a "genetically relevant" phenotype is more powerful than a blind multivariate analysis with all the components (Ott and Raboniwitz 1999).

Existing Methods For Mapping Multivariate Phenotypes

Variance Components Methods (e.g., Amos et al. 1990):

- **Require modeling of covariance structure of the components of the multivariate phenotype vector.**
- **Generally, multivariate normality assumed: problems with robustness, difficulty to verify assumptions in high dimensions.**

Data Reduction Techniques (e.g., Elston et al. 2000):

- **Principal components analysis: a linear combination of the components is defined as the "new" phenotype.**
- **Model-free methods (e.g., Haseman-Elston) possible, but linkage results difficult to interpret.**

Quantitative and Qualitative Phenotypes in a Multivariate Phenotype vector

Example: EEG (quantitative),
ERP (quantitative),
depression (qualitative)
may comprise a multivariate phenotype vector for studying the genetic risk of alcoholism.

Variance components methods (assuming multivariate normality) are incompatible even if only one of the components of the vector is binary/qualitative.

Principal components obtained by decomposing the correlation matrix comprising both quantitative and qualitative phenotypes are unreliable.

The Proposed Method

Constructing the multivariate phenotype vector

$(Y_1, Y_2, \ldots, Y_k)$ is a multivariate vector of quantitative traits, Z is the binary end-point trait.

A logistic regression of Z on $(Y_1, Y_2, \ldots, Y_k)$ will identify those traits significantly correlated with Z. The multivariate phenotype comprise only these traits.

The regression

$G_1, G_2, \ldots, G_m$ are genotypes at *m* marker loci.

Most linkage methods model the dependence of $(Y_1, Y_2, \ldots, Y_k)$ on $G_1, G_2, \ldots, G_m$. Since the response variable is some function of $Y_i s$, inferences are sensitive to the distribution of $Y_i s$.

Alternative formulation: Reverse the response and explanatory variables (e.g., single trait version of Sham et al. 2002) in a linear regression set-up.

The regression model:

$$\Pi_i = \Sigma_j \beta_j d_{ij} + e_i$$

where, Π_i is the estimated i.b.d. score at a marker locus, d_{ij} is the squared difference of trait values corresponding to the i^{th} sib-pair and the j^{th} trait component of the multivariate phenotype vector, and e_i is a random component.

Linkage for the complex trait can be tested in terms of β_j :

$$H_0: \beta_j = 0 \text{ for all } j \text{ versus } H_1: \cup \beta_j < 0$$

The test is performed by a log-likelihood-ratio statistic. Since H_1 is constrained, the asymptotic distribution is difficult to obtain and is determined empirically using permutation principles.

Advantages :

- does not require modeling of covariance structure of component phenotypes. Robust with respect to violations in distributional assumptions.

- Linkage results are easier to interpret than analyses based on phenotypes defined by data reduction techniques.

- can incorporate both quantitative and qualitative traits.

- using i.b.d. scores as the response variable allows for (*n-1*) independent observations for a sibship of size *n* [not true if functions of quantitative traits are response variables].

SIMULATIONS

Example 1: $(X,Y,Z) \sim N_3(\mu,\Sigma)$

Example 2: (X,Y,Z) such that $(X,Y) \sim N_2(\mu,\Sigma)$ and $Z \sim$ chi-square and correlated with (X,Y)

Example 3: (X,Y,Z) such that X is normal, Y is chi-square and Z is Poisson; pairwise correlated

U is a binary trait defined by $\Phi_3(X,Y,Z)$ standardized.
In each case, there is a common QTL.

Example 4: $(X,Y,Z) \sim N_3(\mu,\Sigma)$ such that (X,Y) is controlled by a common QTL, Z is not. U is defined as $\Phi_2(X,Y)$
Logistic regression gave significance "correctly".

200 sibships (60 of size 2, 80 of 3, 60 of 4), major gene-effect for each trait =3, allele frequencies at trait locus=(0.7,0.3).
Power at θ=0.01 based on 1000 replications.

Model	β (dom)	HE (PC)	RR (PC)	RR (Mult)	HE (Bin)	RR (Bin)
1	0	.802	.854	.899	.721	.735
	1.5	.751	.778	.815	.643	.670
2	0	.766	.817	.845	.718	.723
	1.5	.714	.751	.779	.626	.621
3	0	.683	.702	.770	.608	.628
	1.5	.624	.665	.713	.577	.582
4	0	.695	.699	.731	.438	.457
(all 3)	1.5	.617	.632	.690	.375	.361
4	0	.754	.784	.816	.609	.624
(first 2)	1.5	.700	.728	.776	.543	.558

COLLABORATIVE STUDY ON THE GENETICS OF ALCOHOLISM (COGA)

Multicenter project to detect and map susceptibility genes for alcoholic dependence (SUNY, Wash U, IUPUI, U Iowa, UCONN, Howard U, Rutgers U, SFBR)

262 families, COGA proband and having at least three full sibs and both parents/larger sibships.

405 markers with average inter-marker distance 10.9 cM

Qualitative and Ordinal Traits:
COGA (DSM-III-R + Feighner)
DSM -IV
ICD-10

LINKAGE ANALYSES ON COGA DATA

No promising linkage finding with binary clinical phenotypes like COGA, DSM-IV and ICD-10.

Alternative strategy was to analyze quantitative precursors.

Quantitative Traits:
Maximum number of drinks in a 24 hour period (a measure to grade non-alcoholic individuals)

 # externalizing symptoms (symptom count associated with alcoholism)

Electroencephalogram (EEG) Waves (data collected at 31 bipolar electrode positions)

MULTIVARIATE PHENOTYPES IN COGA

Univariate studies have provided linkage evidence on Chromosome 4 (near the *ADH* gene cluster) for:

 # Max-drinks in a 24 hour period [Saccone et al. 2000]
Beta 2 EEG [Porjesz et al. 2002, Ghosh et al. 2003]
 # Externalizing Symptoms [Ghosh et al. 2008]

Binary trait: COGA diagnosis

Logistic regression: All 3 quantitative traits significantly correlated (p-value < 0.001)

Scan on Chromosome 4 using reverse regression:

Peak at 118 cM (p-value < 0.0001)

CLOSING REMARKS

- **A single quantitative phenotype correlated with a binary end-point trait may not be a sufficiently good surrogate for the end-point trait. Rather, a multivariate phenotype comprising both quantitative and binary phenotypes may be more optimal.**

- **Using i.b.d. scores as the response variable and components of a multivariate phenotype as explanatory variables in a regression set-up has both analytical and interpretational advantages.**

No single gene-finding method is sufficient or complete.
Multiple roads should lead to Rome.

Our Human Genetics Unit Team