Testing for Linkage Replication

Michael C. Neale
Seth B. Roberts

Virginia Institute for Psychiatric and Behavioral Genetics

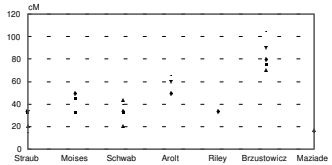
Linkage replication is important

If not essential!

- Basic desirable property in science
- Traditionally high significance levels
  - Bayesian 23 chromosome
  - Type 1 error rate
  - Lod > 3.3, p < .00074
- Relaxation of these levels
  - Complex traits
  - Genotyping errors?
  - Biotech
  - Larger Type I error rate

Variation in peak location exists

Schizophrenia 6p LOD>1.5



Other disorders too

some non-psychaitric

- Bipolar 21q 5 studies 30cM
- IDDM 14q 2 studies 70cM
- Multiple sclerosis 5p 2 studies 60cM
- Psoriasis 4q, 20p 2 studeies 40cm

Published studies

Simulation and analytic methods

- Theoretical
  - Likelihood ratio support interval (Ott 91)
  - One LOD unit either side of peak 95%CI
- But:
  - False positives
  - Biased if samples are 'small'

Simulation

Horvatta et al 1998 Mol Psychiat

- 5 susceptibility loci
  - allele freq .05
  - prop var 5% (75% E)
  - 1cM map
  - within 25cM scan
  - 100 - 1000 sib pairs
- Mean distance of peak from true QTL
  - 10.4 100 sib pairs
  - 2.6 1000 sib pairs

Theoretical

Lander & Kruglyak 1995 Nat Genet

- Take random walk from QTL
  - Interval: LOD < t from maximum
  - zL: prop alleles IBD in ASP's
  - Lambda: locus specific relative risk
- As zL or Lambda decrease N gets huge
  - To get 1cM CI:
    - 170 FI meioses for zL=.975
    - 400 FI meioses for zL=.855-.975
    - 1500 zL=.75
    - 2800 zL=.67
    - 7600 zL=.60
    - 37000 zL=.55

Animal work by Darvasi also relevant

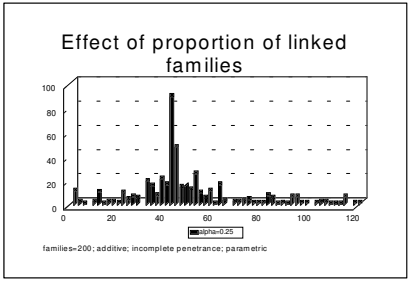
Roberts et al Simulation Study

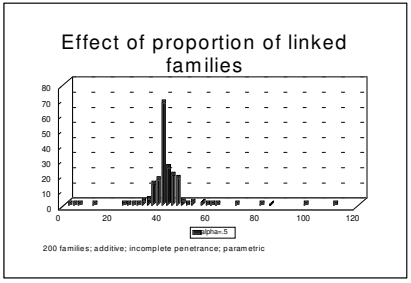
Effects of phenocopies etc

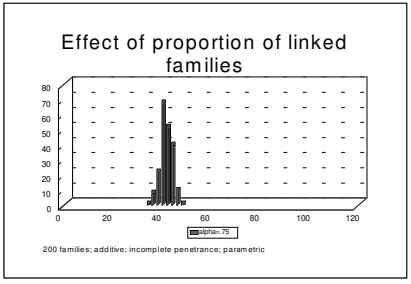
- 13 Markers equally spaced 5 or 10cM
- Nuclear fams: ASPs and their parents
- N 200-1600 families
- Two disease loci, various gene actions
- Proportion linked (alpha) .25-1.00
- Prevalence approx 3%
- Parametric & Non-parametric

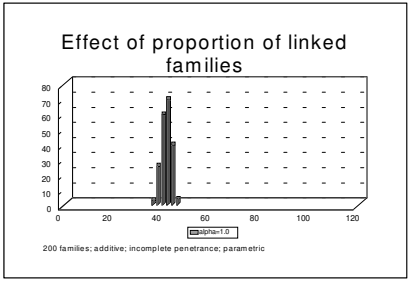
Method

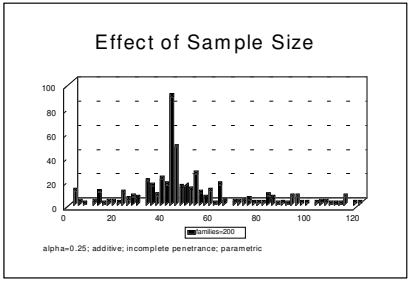
- GASP (Wilson et al 1996) simulate genotype & phenotype data
- SAS - penetrance & phenotypes > dx
- Genehunter (Kruglyak et al 1996)
  - Multipoint LODs every 2cM
  - Markers every 10cM

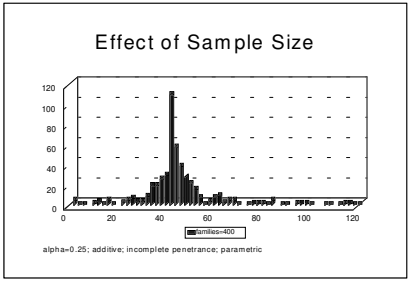

Effect of proportion of linked families
100
80
60
40
20
0
0
20
40
60
80
100
120
alpha=0.25
families=200; additive; incomplete penetrance; parametric


Effect of proportion of linked families
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120
alpha=.5
200 families; additive; incomplete penetrance; parametric

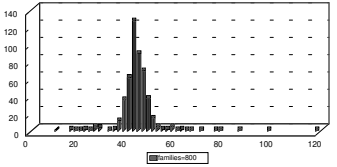

Effect of proportion of linked families
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120
alpha=.75
200 families; additive; incomplete penetrance; parametric


Effect of proportion of linked families
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120
alpha=1.0
200 families; additive; incomplete penetrance; parametric


Effect of Sample Size
100
80
60
40
20
0
0
20
40
60
80
100
120
families=200
alpha=0.25; additive; incomplete penetrance; parametric

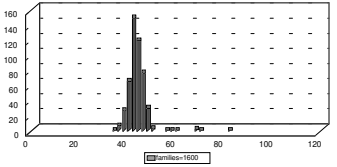

Effect of Sample Size
120
100
80
60
40
20
0
0
20
40
60
80
100
120
families=400
alpha=0.25; additive; incomplete penetrance; parametric

Effect of Sample Size



alpha=0.25; additive; incomplete penetrance; parametric

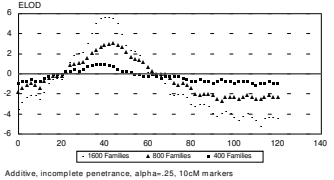
Effect of Sample Size



alpha=0.25; additive; incomplete penetrance; parametric

Expected LOD score

Effect of changing sample size



Additive, incomplete penetrance, alpha=.25, 10cM markers

Test of homogeneity

Combining data from two or more studies

- Support intervals overlap?
  - Biased in small-moderate samples
- Meta-analysis (Li & Rao 96; Gu et al 98)
  - Usually for vs against linkage
  - Summary statistics
- Formal test for heterogeneity (Roberts et al 1999; Roberts 1999)

Formal test

Likelihood Ratio Chi-squared

- Obtain raw data from k studies
- lnLi = log likelihood of data for parameter estimates theta i
- G0 = sum of ln Li
- G1 = sum of ln Li when location estimates are constrained to be equal
- 2(G0 - G1) ~ chi-squared with k-1 df

Computing LRT for heterogeneity

In practice

- Obtain multipoint curves for each of the k datasets
- Sum maximum LOD for each 'LODu'
- Sum multipoint curves across datasets and find maximum 'LODc'
- LRT = 2 ln 10(LODu - LODc)

Considerations

- Same locus -> same gene action?
- Do trait-relevant loci cluster genomically?
- Usual limitations of linkage studies x k:
  - genotyping errors
  - phenotyping errors

Conclusions

Be careful out there

- Heterogeneity adds enormously to location error
  - Narrow phenotypic definitions?
- Sample sizes / design could be better
  - Big sibships
- Watch out for false positives
  - QTL effect sizes expected to be biased upwards
- Replicate or be damned