

# Estimating Gene Position from Genetic Linkage Studies: Use of Covariates

Dave Glidden

K-Y Liang, Y-F Chiu and AE Pulver

June 9, 2003

## Outline

- Complex Genetic Diseases  
*Ex: Schizophrenia*
- Localizing genes
- Can localization be more formal, efficient?
- Exploiting covariate data
- Data Example
- Simulation Studies

Slide 1

## Schizophrenia

- Debilitating psychiatric disease  
*affects 1% of the population*
- Possible risk factors  
*not breast fed, perinatal infection  
birth trauma, Rh mismatch*
- Strong family clustering  
*MZ twin concordance 40-50%*
- No clear inheritance pattern
- Where are the genes?

1

Slide 2

### Complex Genetic Diseases

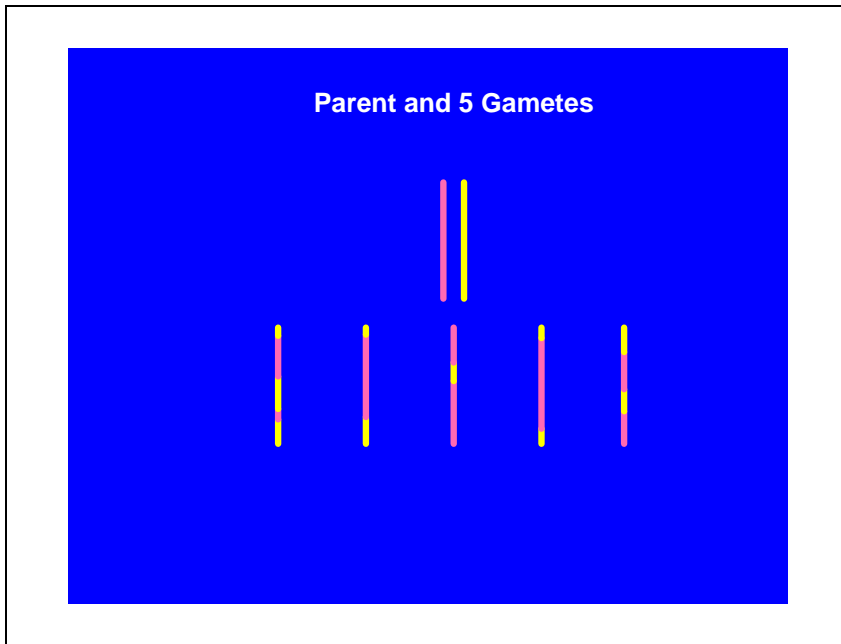
- Relatively Common Diseases
- Environmental risk factors
- Gene-environment interaction
- Multiple genetic loci
- No clear inheritance pattern

Slide 3

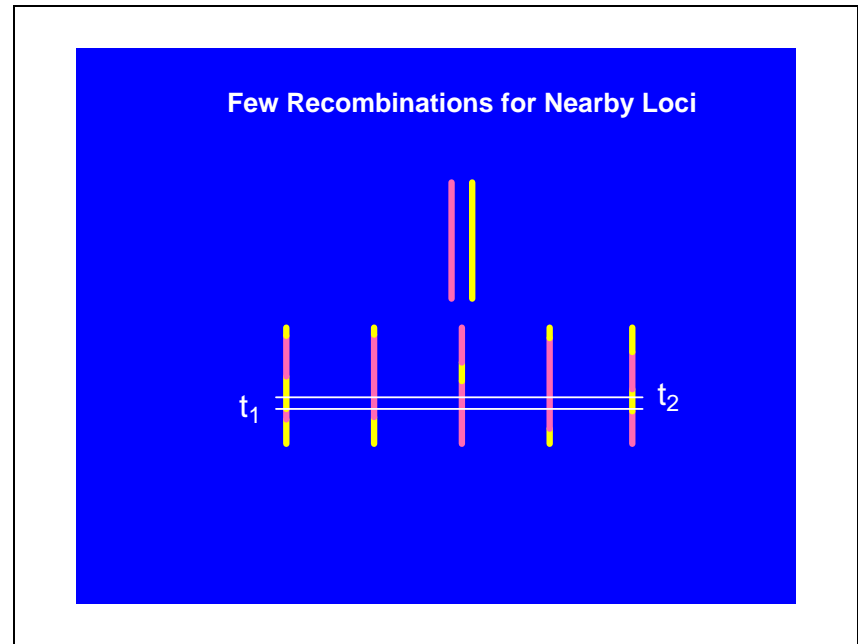
### Inheritance

- DNA organized into 23 chromosome pairs
- Each locus: two variants (alleles)  
*randomly selected from their 2 alleles*
- One from each parent  
*randomly selected from their 2 alleles*
- Chromosomes inherited in blocks  
*blocks defined by recombination*
- Nearby loci are inherited together  
*said to be **linked***

Slide 4



Slide 5



Slide 6

### Distance and Linkage

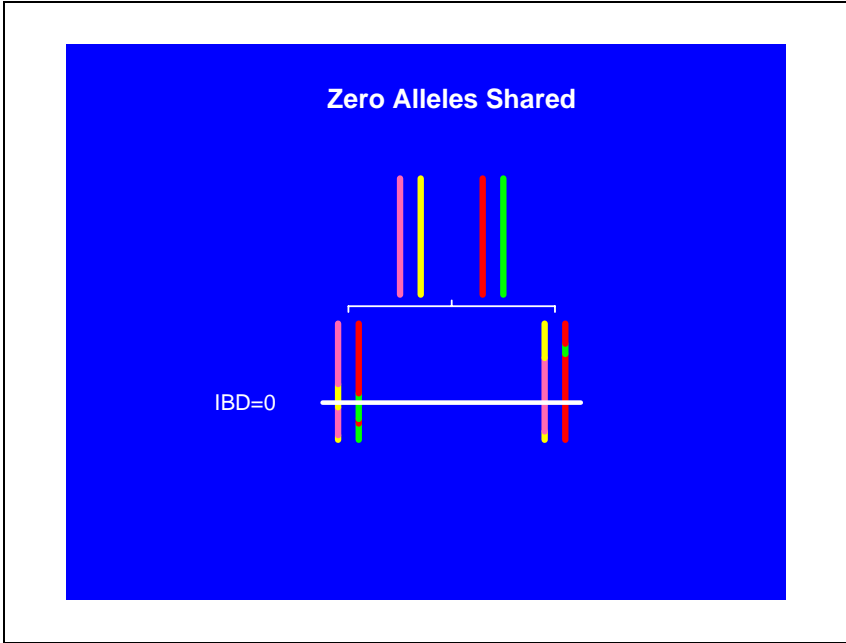
- $\theta_{t_1, t_2}$ : recombination fraction  
*Pr alleles at  $t_1$  &  $t_2$  from diff grandparents*
- If  $\theta_{t_1, t_2} < 0.50$ ,  $t_1$  and  $t_2$  are linked  
*ie, they are nearby*
- Distance (*Genetic*): Morgan = 100cM  
*1 Morgan,  $E(\# \text{ recombinations}) = 1$*
- Distance (*Physical*): Basepair = nucleotide  
*1 Mb = million basepairs*
- 1 cM is about 1 Mb (1.05 Men, 0.70 Women)

Slide 7

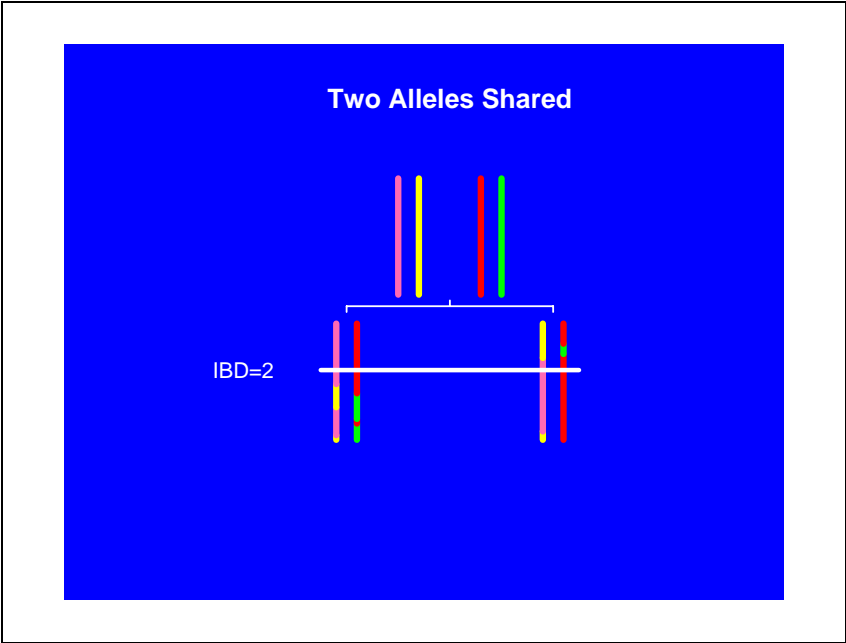
### Affected Sibling Pairs

- Siblings share 0,1, or 2 alleles  
 *$pr(0)=0.25, pr(1)=0.50, pr(2)=0.25$*
- Mean # shared = 1
- ASP greater sharing at susceptibility locus (SL)  
*how much greater? complex genetics*
- Locus linked to SL: greater sharing  
*less than at susceptibility locus*
- Excess marker sharing: nearby gene?

Slide 8



Slide 9



Slide 10

### Schizophrenia Study

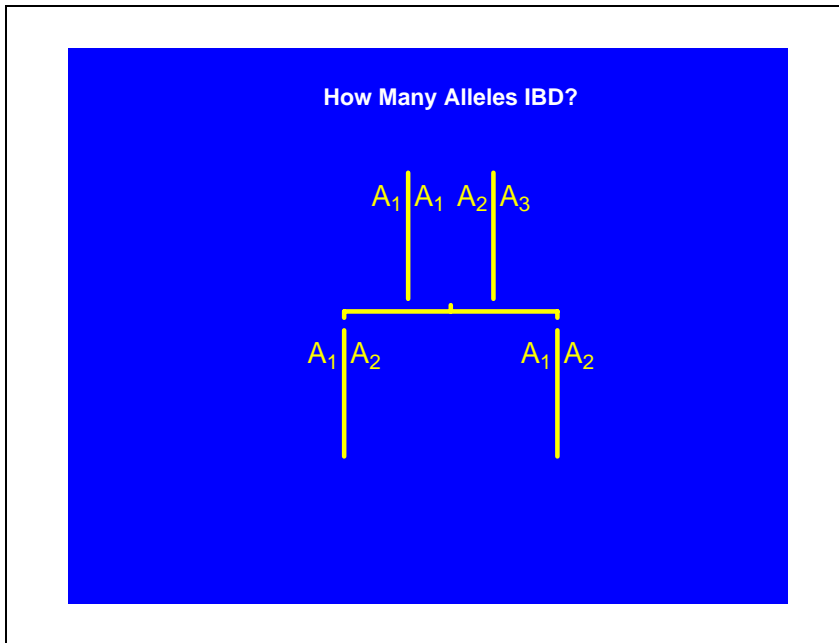
- Pulver group *Nature Genetics* 1998
- 64 affected sibs with schizophrenia
- Examined sharing at 452 markers
- $\tilde{S}(t)$ : number alleles shared at  $t$
- Test:  $\sqrt{2/n} \sum_{i=1}^n \{\tilde{S}_i(t) - 1\}$   
*correct test for genome-wide significance*

Slide 11

### Region Localized

- Many significant tests in region  
*chromosomes 13 and 8*
- Interesting region is large  
*many genes, physically large*
- Where is susceptibility locus  
*most significant marker?*
- Some markers aren't fully informative
- Data:  $S(t) = E\{\tilde{S}(t)|\text{Marker information}\}$
- Liang: map based on  $S(t)$

Slide 12

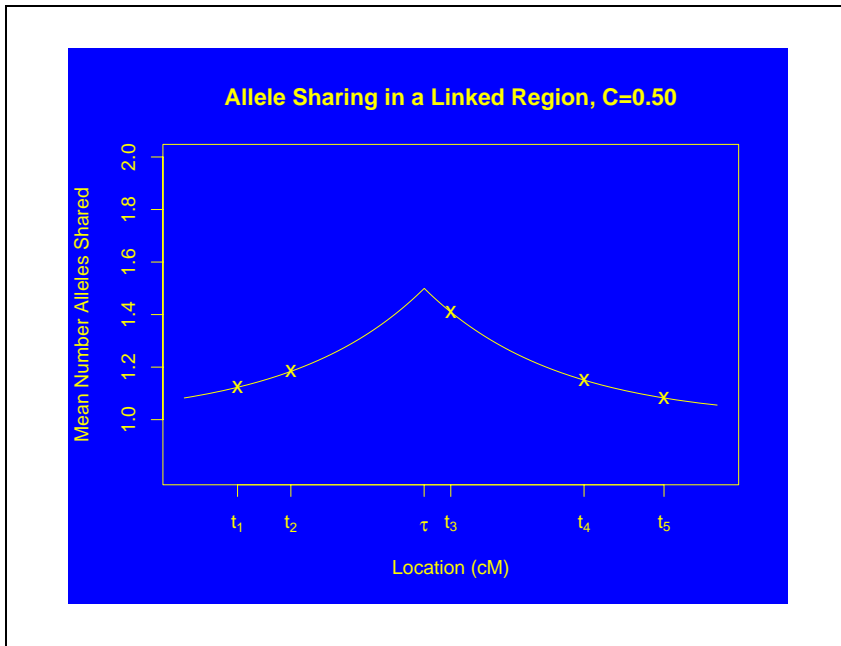


Slide 13

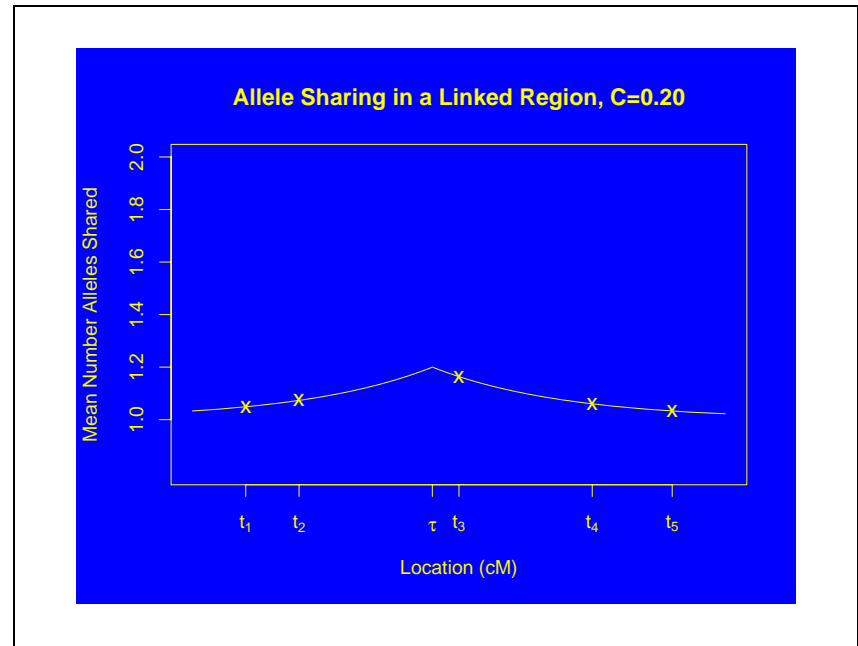
**Liang et al. (2001)**

- $\tau$ : trait locus in cM
- $E\{S(t) - 1|\Phi\} = (1 - 2\theta_{t,\tau})^2 C$
- $\Phi$ : ascertainment event (affected sibs)
- $C = E\{S(\tau) - 1|\Phi\}$
- Using the Haldane mapping function
- $E\{S(t) - 1|\Phi\} = \exp(-0.04|t - \tau|)C$
- Analogy with longitudinal data
- Generalized estimating equations to estimate  $\tau$

Slide 14



Slide 15



Slide 16

### Statistical Issues

- Dependence  $\{S(t_1), \dots, S(t_M)\}$  hard to specify  
*recombination interference*  
*GEE appealing*
- Oops  $|t - \tau|$ : not differentiable at  $\tau = t$
- Replace with  $\|t - \tau\|_\epsilon =$   
 $|t - \tau|$  if  $|t - \tau| > \epsilon$   
 $(t - \tau)^2/2\epsilon + \epsilon/2$  if  $|t - \tau| \leq \epsilon$
- $\epsilon$ : arbitrary  
*chosen to be small*  
*small effect on  $\hat{\tau}$*

Slide 17

### Heterogeneity

- Complex Diseases: multiple? genes, etiologies
- Represented by: disease subtypes  
disease severity, age at onset
- Can this information refine linkage?
- Useful for BRCA1, prostate cancer
- Pulver data: age at onset

Slide 18

### Covariate

- X: covariate for pair  
 $E\{S(t) - 1|X, \Phi\} = (1 - 2\theta_{t,\tau})^2 C(x)$
- $C(x) = E\{S(\tau) - 1|X, \Phi\}$
- How to model  $C(x)$ ?  
*simple categories*  
 $\text{logit}\{C(x)\} = \alpha + \beta x$
- Splines can augment modeling
- Use GEE for enriched model

Slide 19

### Pulver Study

- Pulver group *Nature Genetics* 1998
- Genome-wide scan for schizophrenia SL
- Affected proband + 1st or 2nd degree relatives  
*schizophrenia, schizoaffective: DSM IV*
- 64 affected sibpairs
- 452 micro satellite markers  
*avg 7.6 cM distance*

Slide 20

### Results

- Linkage to chromosome 13q32
  - $p = 2.2 \times 10^{-5}$
  - significant** *by genome-wide criteria*
- Linkage to chromosome 8p21-22
  - $p = 1.9 \times 10^{-4}$
  - suggestive** *by genome-wide criteria*
- Where might the susceptibility loci be?

Slide 21

### Analysis (No Covariate)

- Examined 64 sibpairs from Pulver study
- 182cM region for Ch8 (35 markers)
- 127cM region for Ch13 (28 markers)
- Apply Liang method (no covs)
- Locus on Ch13:  $\hat{\tau}=113.4\text{cM}$  ( $\hat{C} = 0.34$ )  
*95% CI, 108.7cM–118.1cM*
- Locus on Ch8:  $\hat{\tau} = 45.2\text{cM}$  ( $\hat{C} = 0.36$ )  
*95% CI, 39.3cM–51.1cM*

Slide 22

### Analysis w/Covariate

- Age onset of schizophrenia
- Early if  $\leq 21$ , Late if  $> 21$   
*cut at median*
- 20 pairs both early-onset (EE)
- 17 Early/Late (EL), 27 Late/Late (LL)

$$E\{S(t)|\Phi, X\} = 1 + \exp(-0.04|t - \tau|)C_l$$

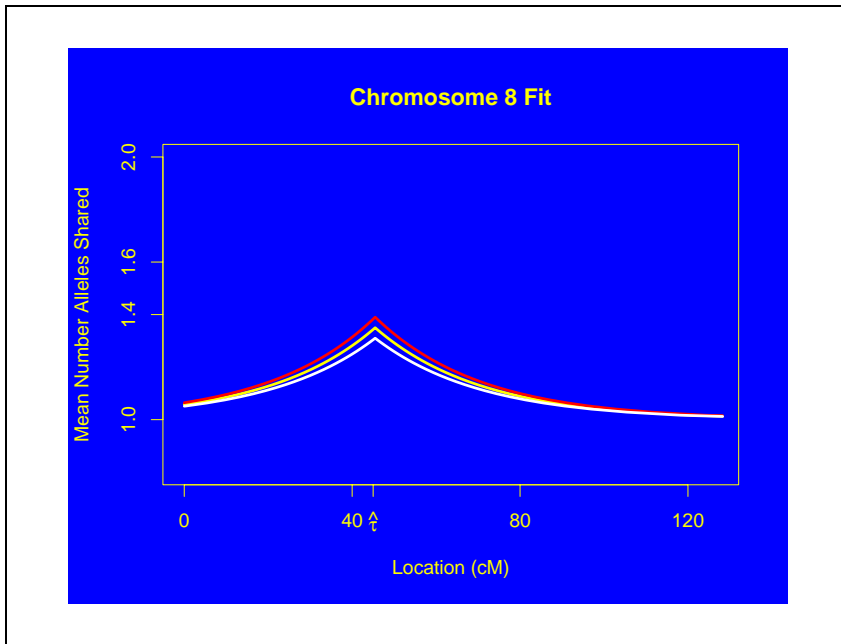
- $C_l$ , ( $l = 1, \dots, 3$ ) for EE,EL,LL respectively
- Sharing at  $\tau$  depends on age at onset

Slide 23

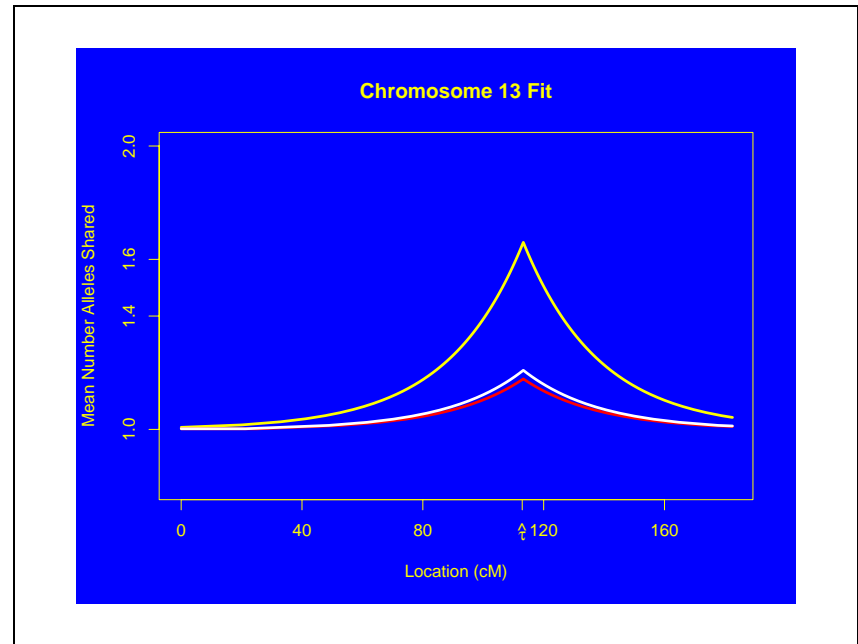
### Covariate Results

- Chromosome 13:  
 $\hat{\tau}=112.9\text{cM}$   $\hat{C} = (0.66, 0.18, 0.21)$   
*95% CI: 109.8cM–116.0cM*  
*CI width: shrunk by 29%*
- Chromosome 8:  
 $\hat{\tau} = 45.0\text{cM}$   $\hat{C} = (0.35, 0.39, 0.31)$   
*95% CI: 39.2cM–51.2cM*  
*CI width: increased by 5%*

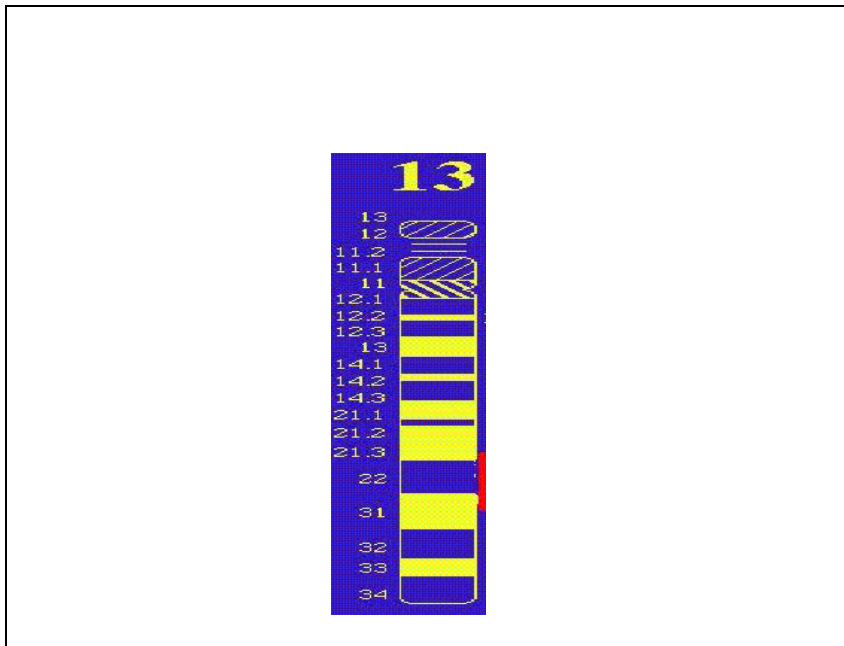
Slide 24



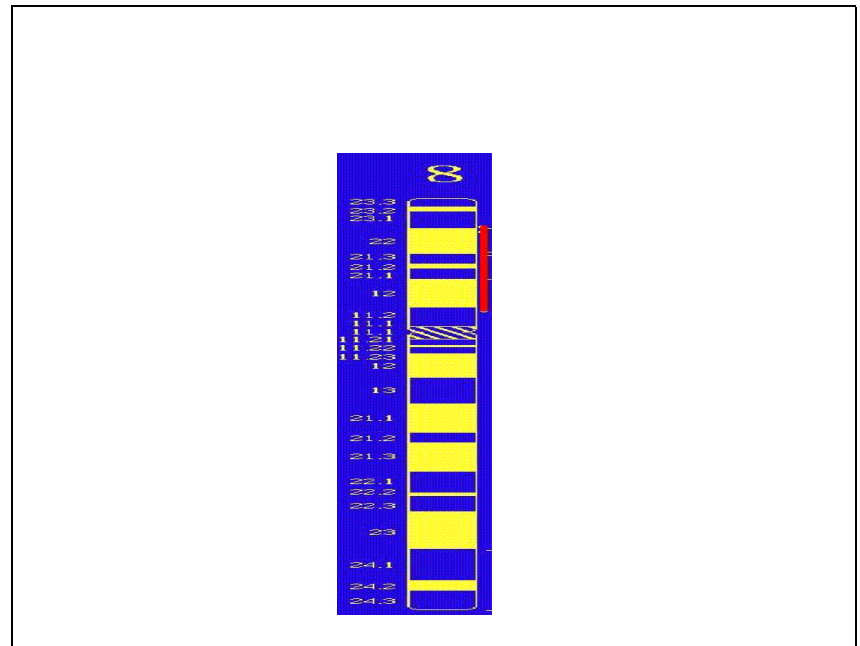
Slide 25



Slide 26



Slide 27



Slide 28

### Simulation Studies I

- Weibull penetrance (Li & Hsu, 2000)  
*disease gene freq = 0.05*  
*each copy increases risk 3.5-fold*
- ASP with recorded age at onset
- Ascertained if onset  $\leq 45$  years old
- Predictor: log mean age onset

Slide 29

### More on Simulations

- Markers at (0,10,20,...,80,90,100) cM
- $\tau = 45$ cM
- Study sizes 250, 500 sibpairs
- 50,000 simulations
- Compare covariate v. no covariate method  
*mean, variability (SE), ratio of SE*

Slide 30

**Simulation: Age at Onset**

No. Sib Pairs		Mean	SE	Relative SE
250	No Cov	45.1	5.03	-
	Cov	44.9	3.79	0.75
500	No Cov	45.1	3.18	-
	Cov	45.0	2.37	0.75

Slide 31

**Simulation Studies**

- Markers at (0,10,20,...,80,90,100) cM
- $\tau = 45\text{cM}$
- $X$  continuous predictor
- Mean allele sharing = 1.20 at  $\tau$   
*allele sharing independent of  $X$*
- Study sizes 250, 500 sibpairs
- 50,000 simulations

Slide 32

**Simulations: Useless Covariate**

No. Sib Pairs		Mean	SE	Relative SE
250	No Cov	45.0	5.29	-
	Cov	45.1	5.63	1.06
500	No Cov	45.0	3.26	-
	Cov	45.0	3.41	1.05

Slide 33

**Summary**

- Possible to improve efft of  $\hat{\tau}$   
*even when mismodelled*
- Methods examine heterogeneity  
*through observed covariate*
- General approach  
*many possible extensions*
- Avoids discarding data

Slide 34

### Future Work

- Other estimation approaches  
*Likelihoods, Alternative maps*
- Extend to general pedigrees  
*nuisance parameters increase*
- Compare to other covariate methods
- Examine two-gene methods  
*Biernacka and Bull*
- Map misspecification issues

Slide 35

### Further Info

- Liang, K.-Y., Chiu, Y.-F., Beaty, T.H. *Hum. Hered.* **51**: 64-78, 2001.
- Glidden, D.V., Liang, K.-Y., Chiu, Y.-F., Pulver, A.E. *Genet. Epidemiol* **24**: 107-117, 2003.
- [www.epibiostat.ucsf.edu/dave/talks.html](http://www.epibiostat.ucsf.edu/dave/talks.html)
- <http://www.biostat.jhsph.edu/wmchen/gf.html>

Slide 36

### **Acknowledgments**

- Kung-Yee Liang
- Yen-Feng Chiu
- Ann Pulver
- Chuck McCulloch
- Mark Segal

**Slide 37**