$$X_1, \ldots, X_K = x_K, I_k = i)$$

$$\sum_{i_{k-1}, i_{k+1}} \Pr(X_1 = x_1, \ldots, X_K = x_K, I_k = i, I_{k-1} = i_{k-1}, I_{k+1} = i_{k+1})$$

$$\sum_{i_{k-1}} \phi_{k-1}(x_1, \ldots, x_{k-1}; i_{k-1}) \pi^{(k-1)}_{i_{k-1} i}$$

$$\times g^{(k)}(x_k; i)$$

$$\times \sum_{i_{k+1}} \pi^{(k)}_{i i_{k+1}} \psi_{k+1}(x_{k+1}, \ldots, x_K; i_{k+1})$$

www.csiro.au

Cherry Bud Workshop 25-28 March 2008

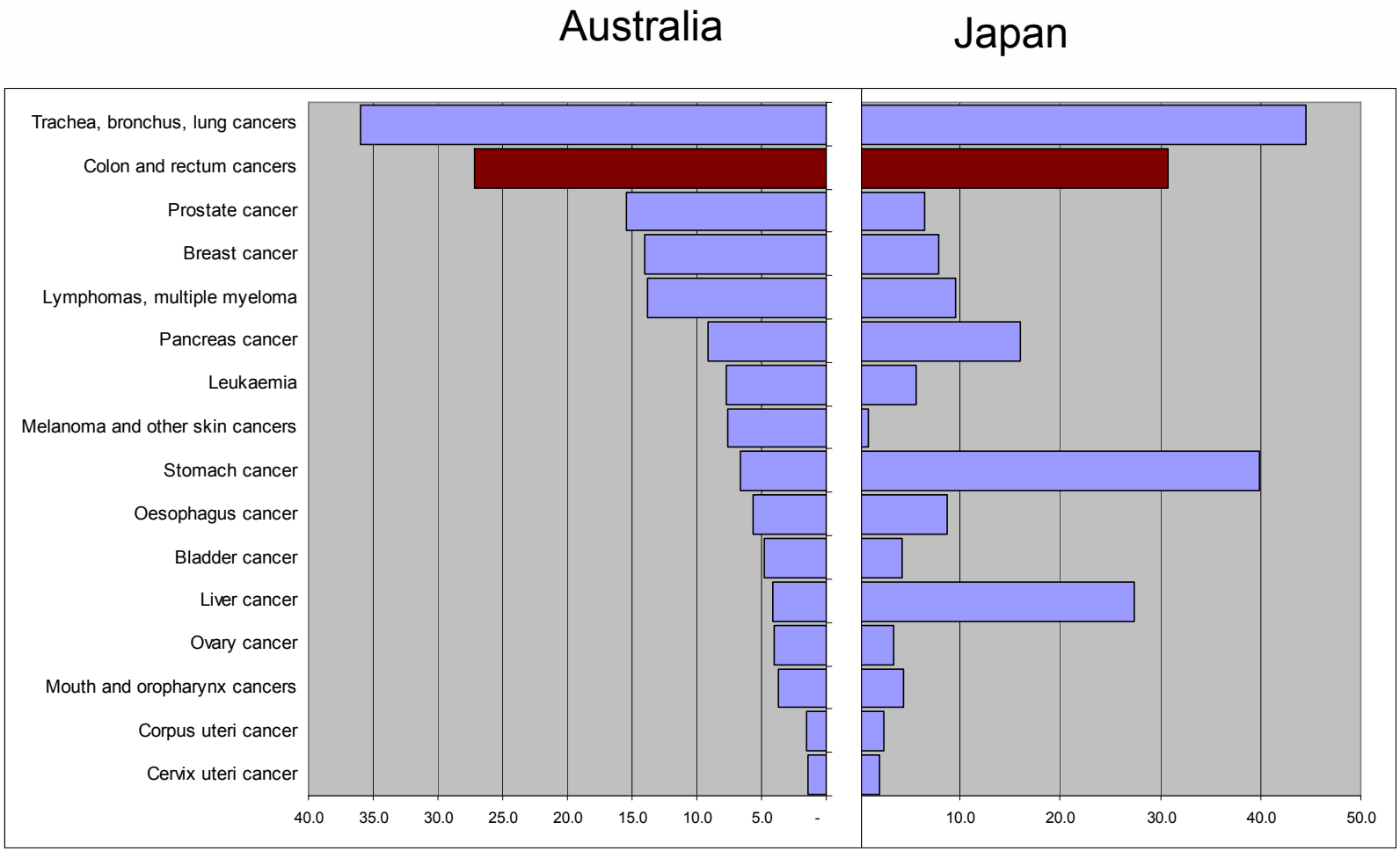
Analysing high-density SNP marker data for linkage with colorectal cancer

Ian W. Saunders

Preventative Health National Research Flagship Program
CSIRO Mathematical and Information Sciences
Adelaide, South Australia

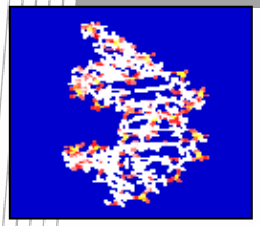
CSIRO

Cancer Death Rates



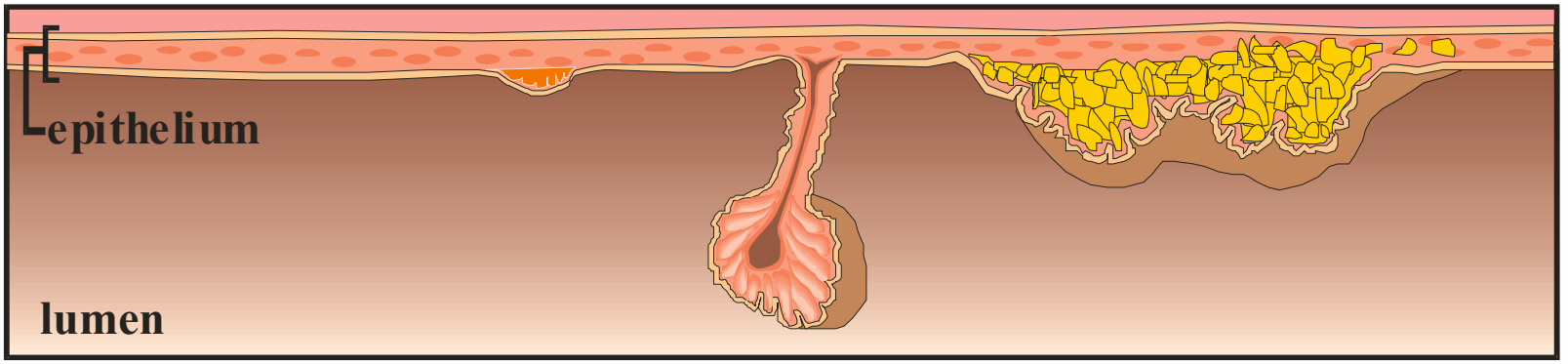
Source: World Health Organisation

Our goal: CRC-specific early diagnosis.



Risk assessment **Early detection** **Diagnosis** **Monitoring**

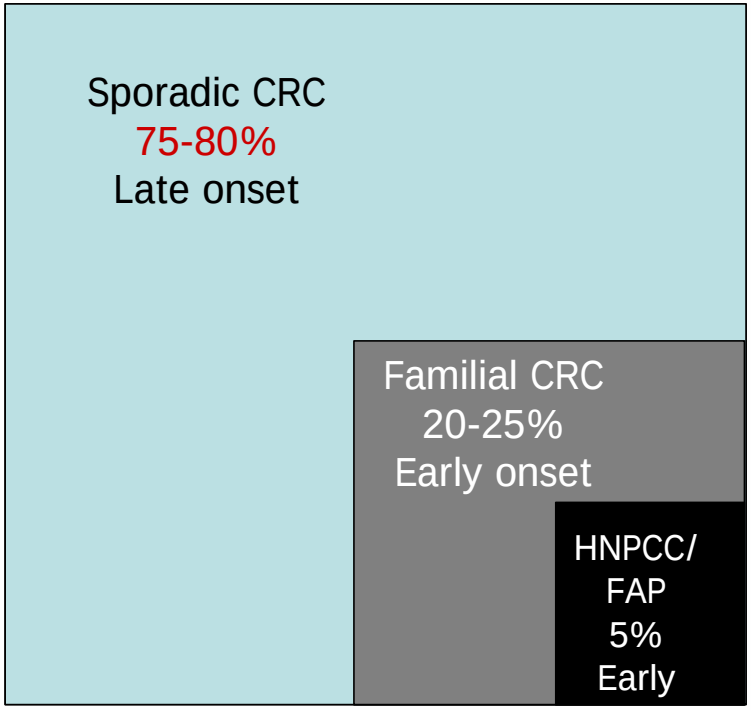
Normal → Hyperplasia → Adenoma → Adenocarcinoma →



(Kinzler & Vogelstein, Cell, 87: 159)

Genetics of CRC

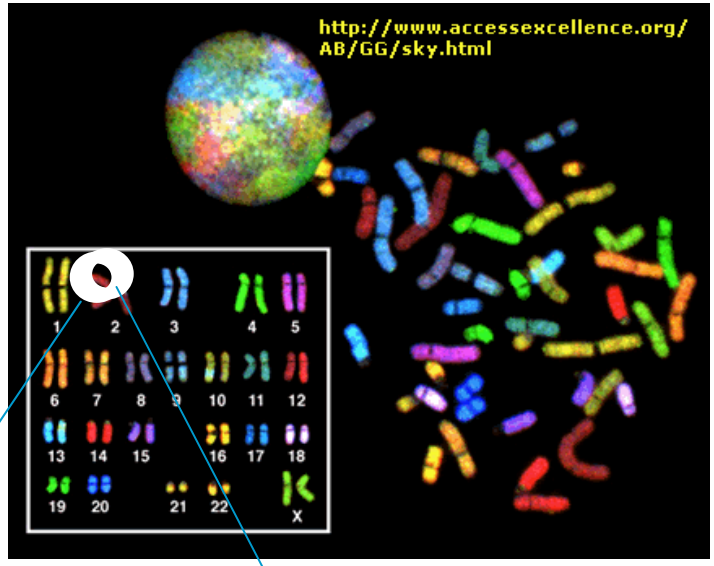
- About 25% of CRCs are in younger (<55) individuals or with a family history of CRC, suggesting a heritable susceptibility.
- Familial – high penetrance single genes, multigenic traits?
- Genotype-environment interactions affect CRC risk?
- SNPs for more sensitive genetic analysis.

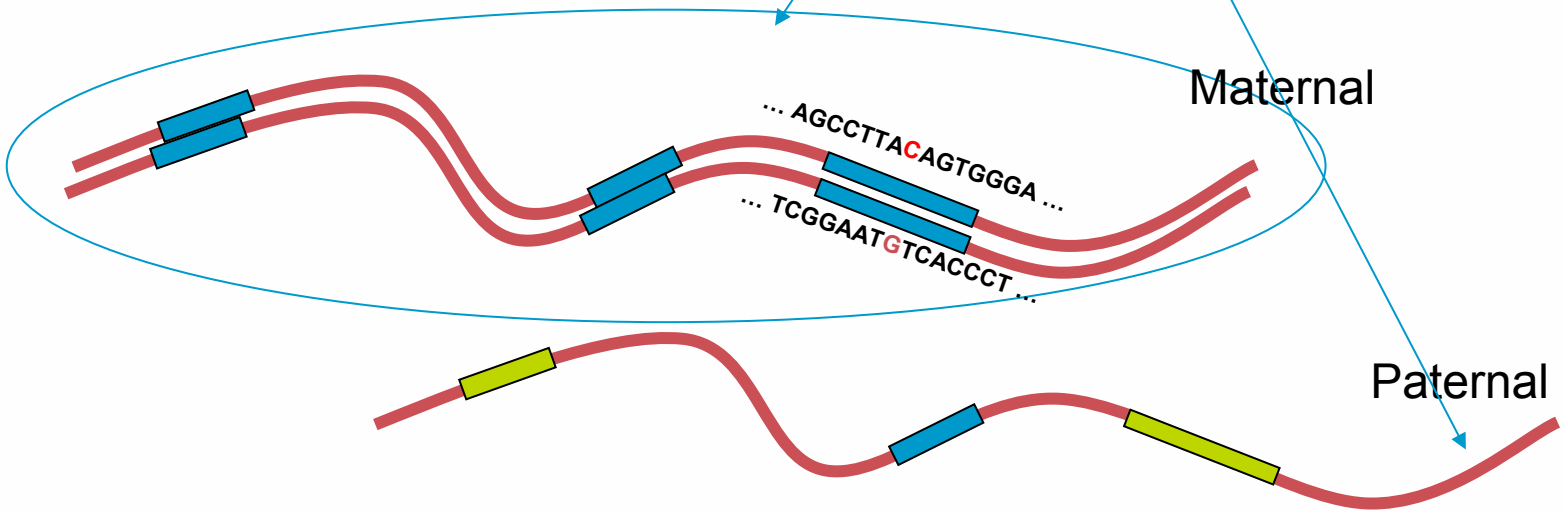


- J.P. Terdiman *et al.* (1999) AJG 94, 2344-2356.

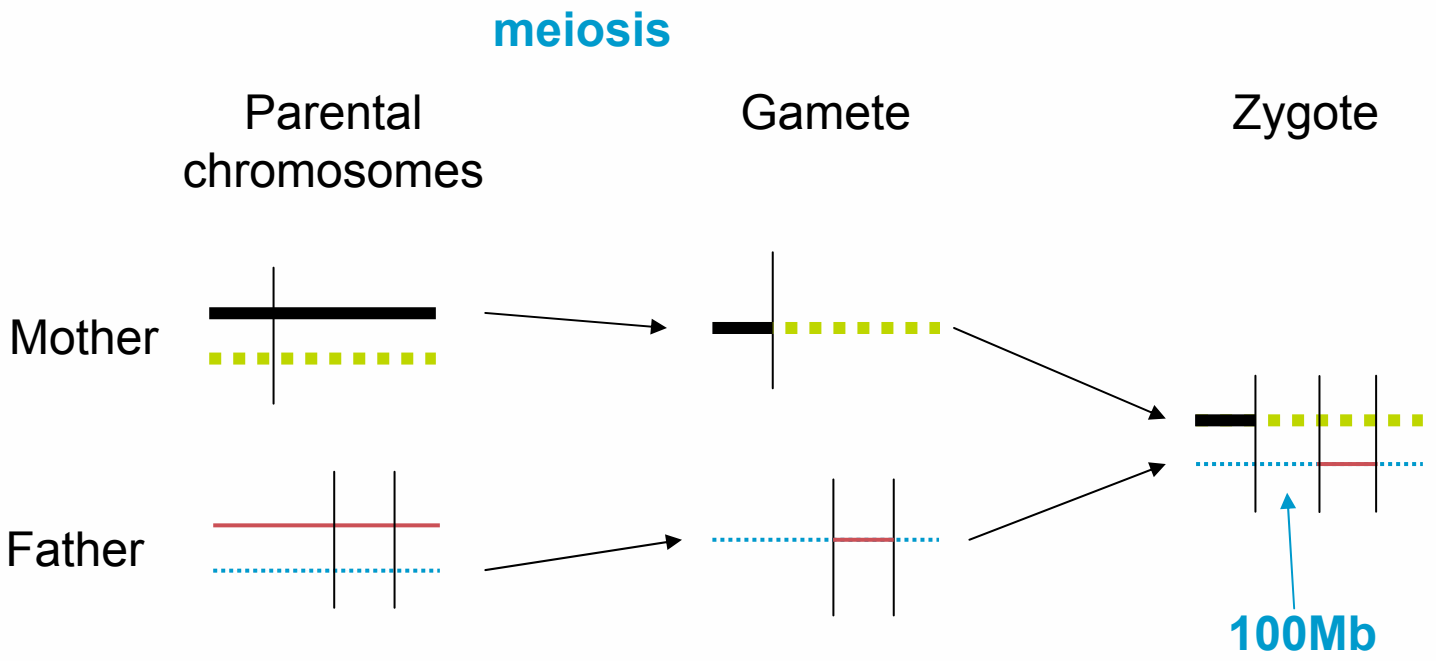
Some Biology …

- **Our cells contain DNA made up of 2 copies of each of 22 'autosomal' chromosomes**
  (plus sex chromosomes, either XX or XY)
- **Chromosomes: on average**
  - About 10^8 "base pairs" (bp) or "nucleotides"
  - About 10^3 genes of length about 10^3 bp
  - So about 1% of chromosome made up of genes





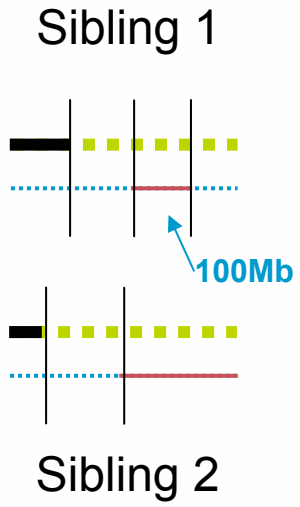
Recombination



- **In *meiosis* – production of egg and sperm – the parents' chromosomes "recombine" at about 1 or 2 points on each chromosome – an average of about 30 per meiosis; one per 100Mb**

Linkage – keeping it in the family

- Closely related individuals share large sections of their DNA
- For example, if two siblings both inherited a particular allele from their mother, they probably share 50Mb or so of DNA surrounding it as well
- So it is easier to find linkage in relatives than in “unrelated” individuals where only very short (3kb) sections are shared
- However, the actual DNA sequences will be different in different families
- We’d like to know where the sections of ‘shared’ DNA are located



Single Nucleotide Polymorphism (SNP)

- **A position in the human genome where a single nucleotide varies between chromosomes while those around it don't**

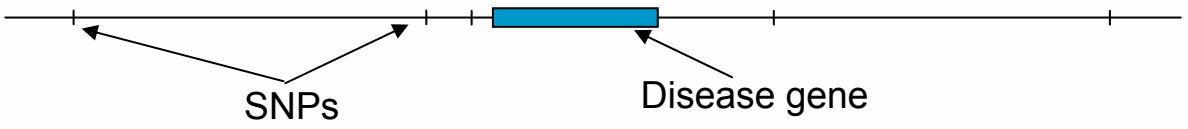
Me: …AGCCTTACAGTGGGA…
…AGCCTTACAGTGGGA…

You: …AGCCTTAGAGTGGGA…
…AGCCTTACAGTGGGA…

ACTG – "nucleotides" or "bases"

Chromosome	Chromosomal Location	Allele A	Allele B	Flanking Sequence	Associated Gene	Freq_A Cau
1	2672921	C	G	gtctatttcagcctta[C/G]agtgggagccttcagc	GSYM:PRDM16;	0.76
1	3776202	A	C	aggcctgagatgagac[A/C]aaaatgtttactgtgg	GSYM:MOT8; AC	0.48

- **Millions of well characterised SNPs are now available – 11,883,685 SNPs in dbSNP last week, but only 225,446 coding non-synonymous – potentially causative**
- **Potential to use SNPs as markers to find association – as markers of a nearby gene, not causes**

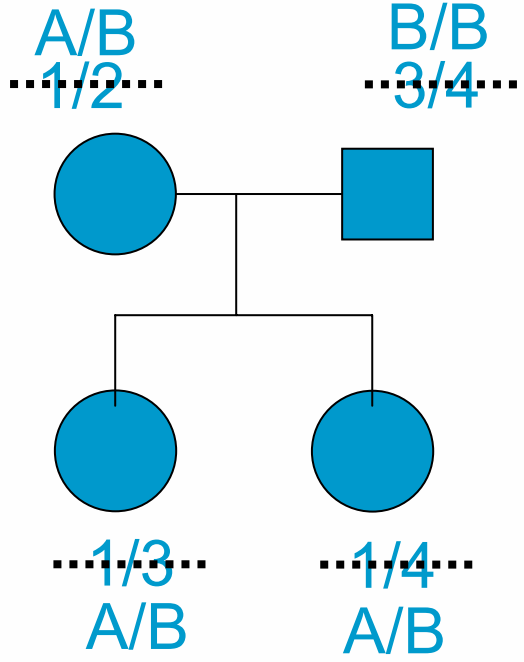


Affymetrix SNP Genotyping Platform

- Platform technology to perform full genome SNP analysis
- Rapidly increasing density of SNP analysis.
- Affymetrix:
  - 2003: 10,000 SNP array
  - 2004: 100,000 SNP array (2 x 50k)
  - 2005: 500,000 SNP array
  - 2007: 900,000 SNP array
- Staining, scanning and genotype calling fully automated

Using high density markers 1. Checking relationships

- **Analysis of the data depends on the pairs actually being siblings, so it's good to check.**
- **Note that the "children" are generally in their 60's so it's often not practical to genotype parents or check memories relating to adoption etc**
- **However, the high density of SNP data allows us to determine relationships with confidence**
- **A simple method uses the number of SNPs where the two siblings have the same genotype**

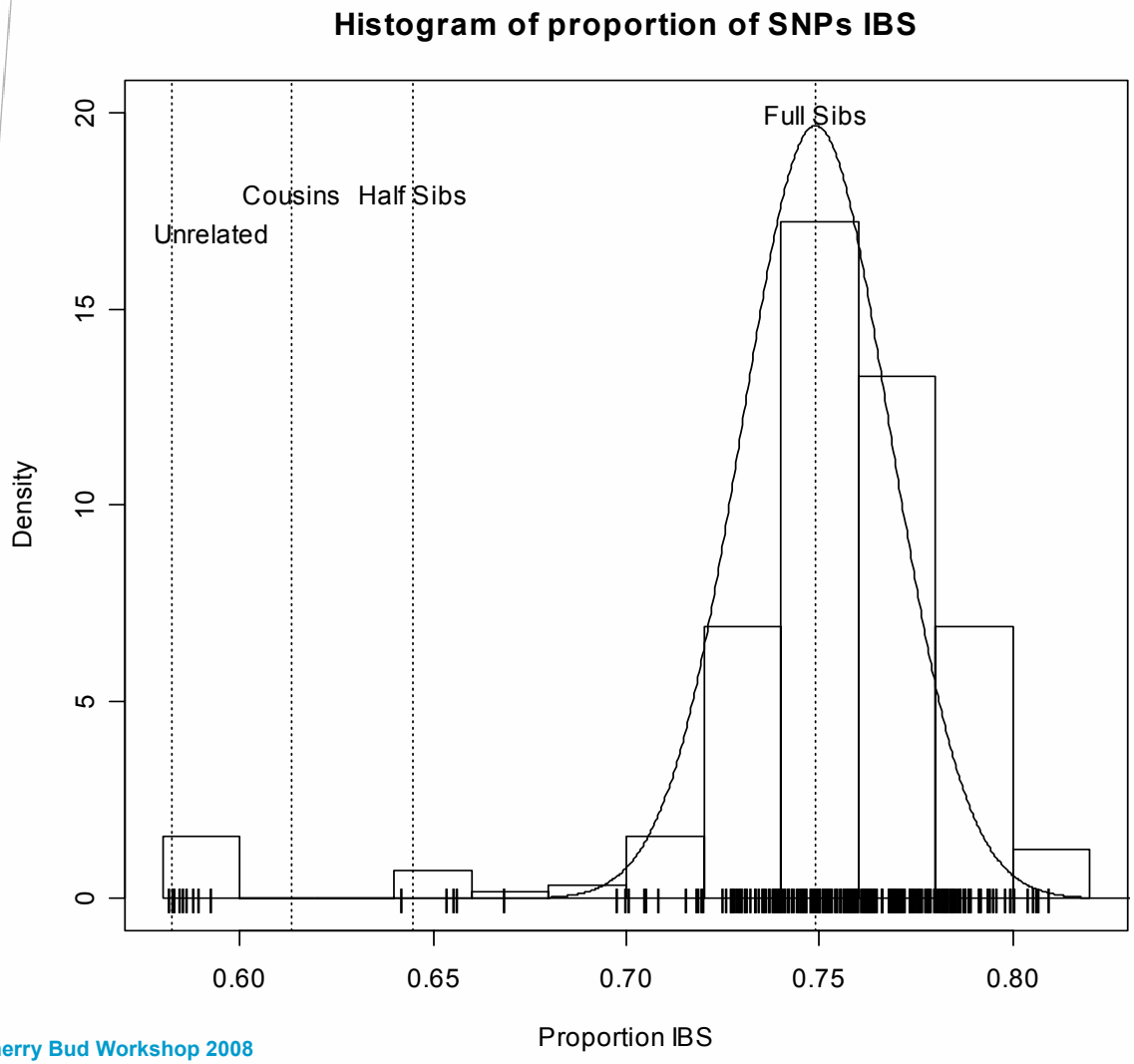


Probabilities for various relationships

Relationship	Prob(Same genotype)	Average for Xba chip
Parent/Child	$p^2+(1-p)^2$	0.7068
Full siblings	$\frac{1}{4}(p^4+4p^2(1-p)^2+(1-p)^4)+\frac{1}{2}(p^2+(1-p)^2)+\frac{1}{4}$	0.7490
Half siblings	$\frac{1}{2}(p^4+4p^2(1-p)^2+(1-p)^4)+\frac{1}{2}(p^2+(1-p)^2)$	0.6446
Uncle/nephew etc	$\frac{1}{2}(p^4+4p^2(1-p)^2+(1-p)^4)+\frac{1}{2}(p^2+(1-p)^2)$	0.6446
First cousins	$\frac{3}{4}(p^4+4p^2(1-p)^2+(1-p)^4)+\frac{1}{4}(p^2+(1-p)^2)$	0.6136
Unrelated	$p^4+4p^2(1-p)^2+(1-p)^4$	0.5825

p = frequency of A allele

Results for 136 “sibling” pairs

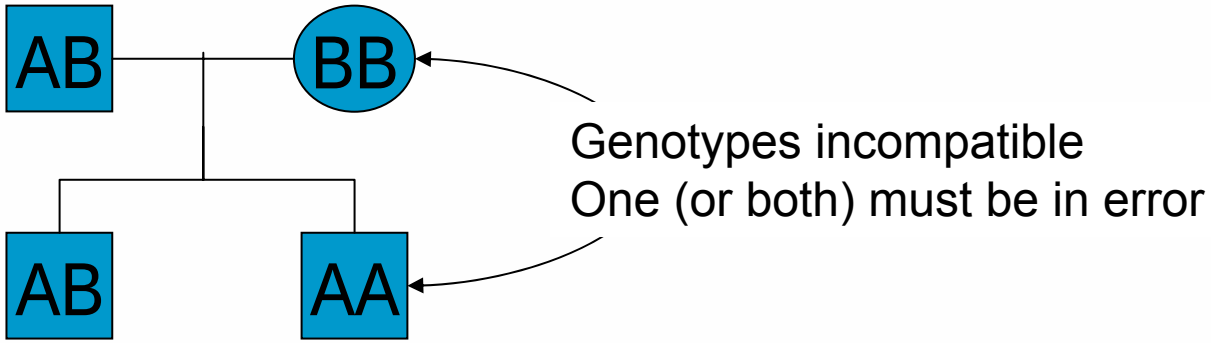


- **We can check the relationships in a family from observable IBS data**
- **(Ethical issues)**

CMIS Technical Report 07/37

Using high density markers 2: Genotyping errors

- **In a family where we have genotypes from both parents and two sibs we found 64 SNPS out of 57241 had 'Mendelian Errors' – genotypes incompatible with Mendelian inheritance**



Genotyping error rate

- **It is not too hard to compute the expected number of SNPs with Mendelian errors for a given rate π of genotyping errors**
- **For families with 2 parents and m children genotyped:**

$$\pi P_{ME}(p_A, m) = (m+2)\pi - 2\pi\left\{p_A^2 + p_B^2 + \left(\tfrac{1}{2}\right)^{m-1} p_A p_B + 4\left[\left(\tfrac{3}{4}\right)^m - \left(\tfrac{1}{2}\right)^m\right] p_A^2 p_B^2\right\} - m\pi p_A p_B \left(3p_A^2 + 4p_A p_B + 3p_B^2\right)$$

- **For families with 1 parent and m children genotyped**

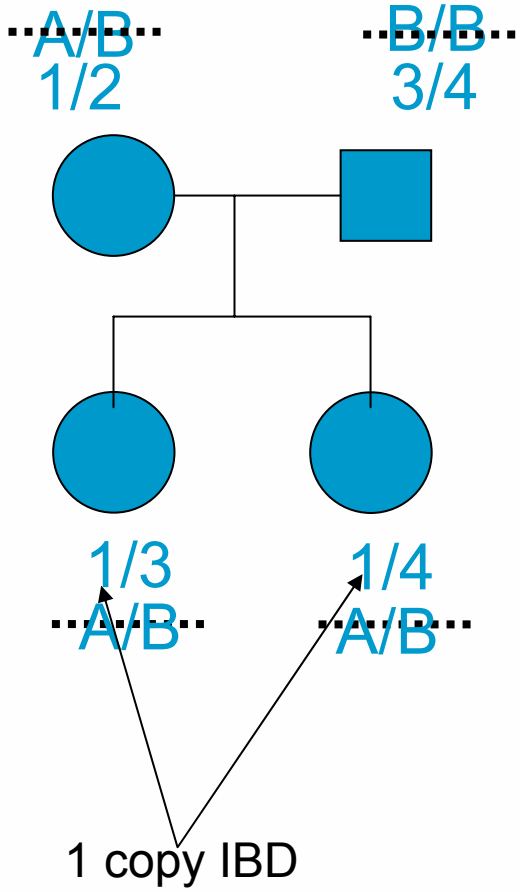
$$\pi P_{ME}^{(1)}(p_A, m) = \pi p_A p_B (2 - (1 - \tfrac{1}{2} p_B)^m - (1 - \tfrac{1}{2} p_A)^m + \tfrac{1}{2} m)$$

- **Leads to an estimate from our families of π =0.13%**
- **CEPH trios: π =0.17%**
- **Can cause loss of information in data analysis, so useful to make corrections:**
  - Small change to linkage algorithm to allow for “observed genotype” differing from “true genotype”
  - Included in subsequent analysis
  - Saunders et al, *Genomics* (2007)

Detecting linkage: Identical by Descent

- **Sharing of DNA between relatives is measured by the number of copies (0, 1 or 2) they inherited from a common ancestor "identical by descent" – "IBD"**
- **IBD probabilities for a disease gene between siblings**

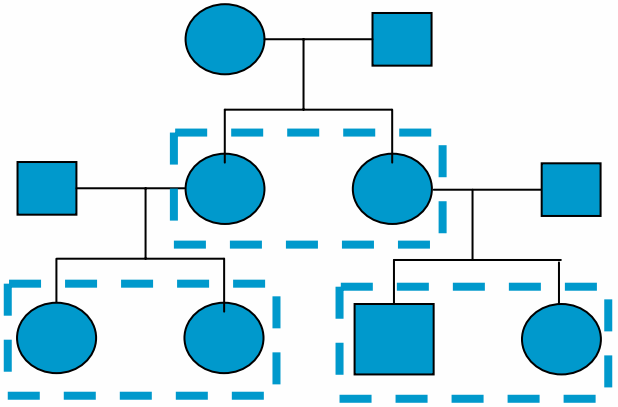
Shared alleles	No linkage	Number affected in pair		
		Both	One	Neither
0	25%	6%	35%	24%
1	50%	49%	50%	50%
2	25%	45%	15%	26%



- **LR test statistic for linkage at SNP *k* is a linear combination $Y_k = w'I$ of counts of number of sib pairs in each IBD class**
- **(IBD status not observable but we can deduce it from genotypes with high accuracy)**
- **Saunders *et al.* (2007) Genetic Epidemiology**

Data for our study

- **Affymetrix Xba chip: 57241 autosomal SNPs**
- **Trial data**
  - 40 individuals
  - 28 pairs of siblings
  - = 11 pairs with both siblings affected
    + 17 pairs with only one affected
  - (small numbers so unlikely to find effect)



- **Major study**
  - 1700 individuals in 110 families
  - 350 genotyped
  - 203 sib pairs: 46 2-affected + 157 1-affected
  - Used sib pair information only
  - Yuki Sugaya investigating use of complete pedigrees

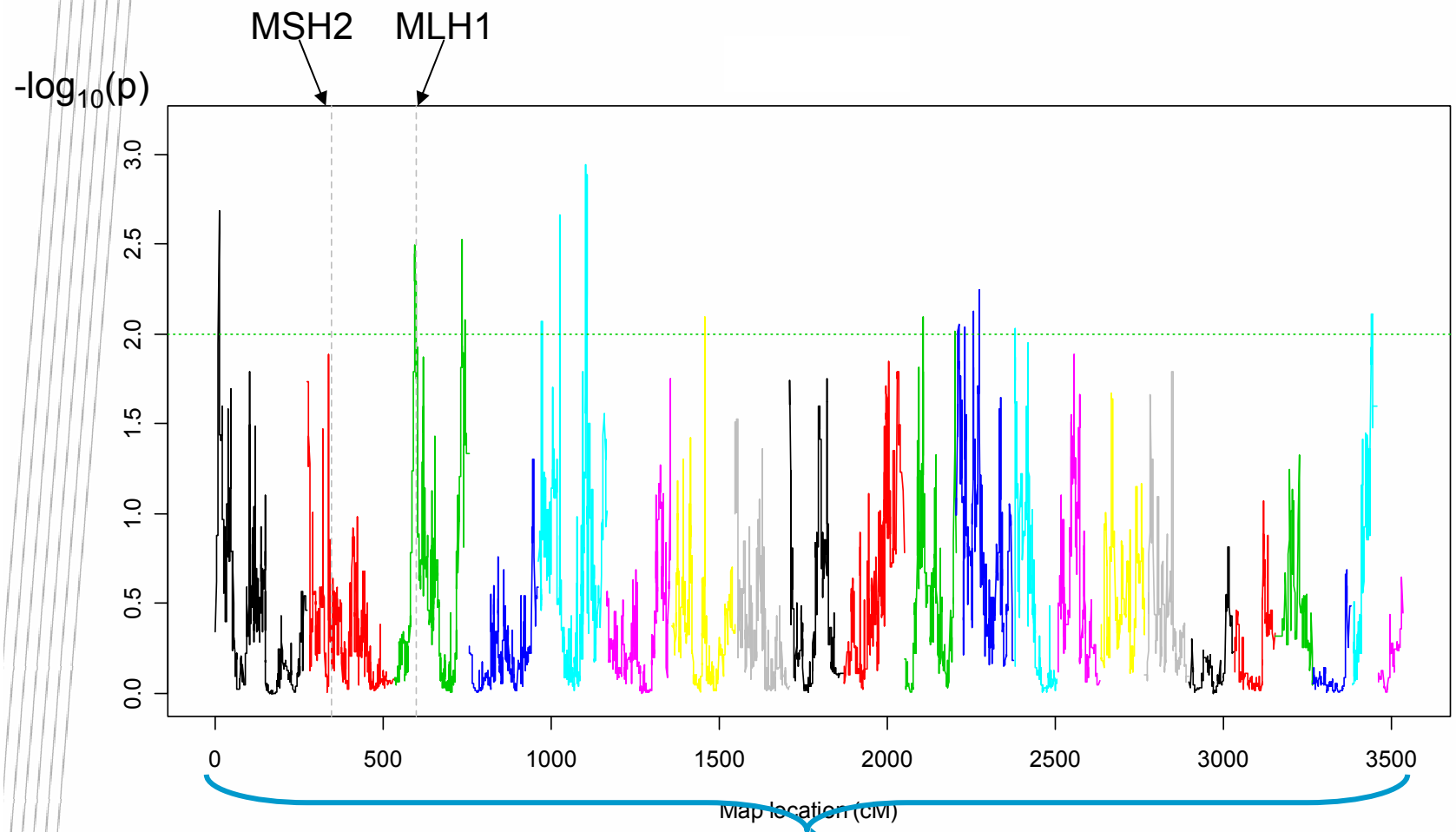
Some results

- **Results for 28 sib pairs known to have MLH1 or MSH2 mutation. Genes near SNPs 4550 and 8523.**

SNP		4549	4550	4551	…	8522	8523	8524	
1 Affected:	0 IBD	3	3	3		3	3	3	
	1 IBD	9	9	9		8	8	8	
	2 IBD	5	5	5		6	6	6	
2 affected:	0 IBD	2	2	2		1	1	1	
	1 IBD	6	6	6		3	3	3	
	2 IBD	3	3	3		7	7	7	
	$-\log_{10}$(p-value)	0.41	0.41	0.41		2.1	2.1	2.1	

- **Deviation from 25%/50%/25% suggests linkage with the disease. Measure of deviation based on likelihood ratio.**

Results for test data (28 sib pairs)



57241 individual tests – is this result "surprising"?

Joint properties of the sequence of test statistics

- **We can calculate pointwise test statistics, but we now have 57241 of them with strong correlation.**
- **It turns out that the sequence of statistics Y_k can be approximated by an autoregressive (Markov) process which does not depend strongly on the alternative disease model.**

$$\operatorname{cov}(Y_i, Y_j) = (1 - w_1^2)e^{-4|\lambda_i - \lambda_j|} + w_1^2 e^{-8|\lambda_i - \lambda_j|}$$

$$\approx e^{-4(1+w_1^2\}|\lambda_i - \lambda_j|}$$

- **So that**

$$Y \sim N(0, \Sigma)$$

- **Where Σ is the above covariance matrix**

Joint properties of the sequence of test statistics

- **The presence of a disease susceptibility genes at G alters the distribution of Y_G, and hence the joint distribution**

$$Y \sim N(\mu_* s(x_*), \Sigma + (\sigma_*^2 - 1) s(x_*) s(x_*)')$$

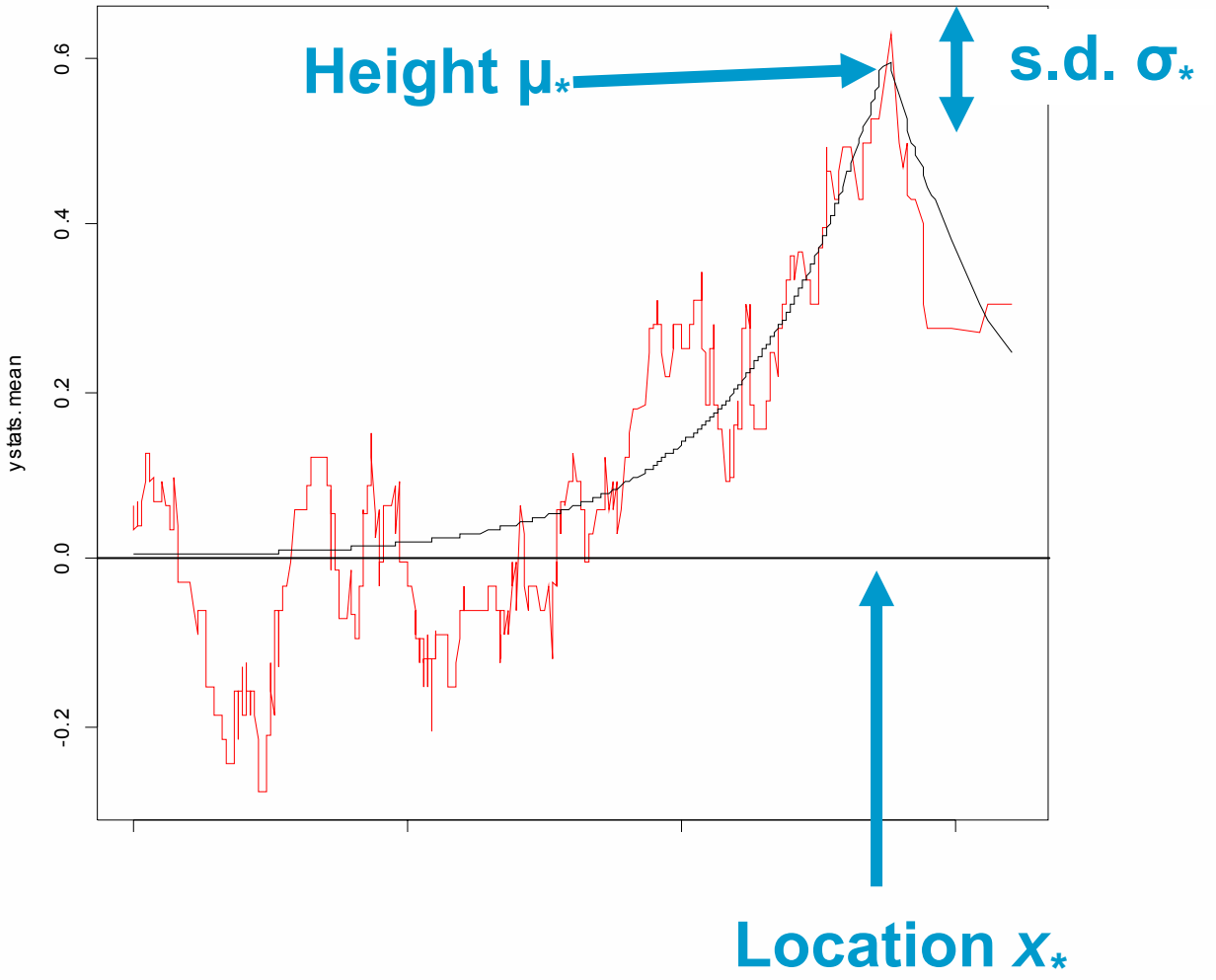
$$s(x_*) = e^{-4(1+w_1^2)|x - x_* 1|}$$

- **Simulation of the joint distribution is easy.**
- **Computation of the likelihood requires inversion of the covariance matrix**

$$L(x_*, \mu_*, \sigma_*^2) = \frac{1}{(2\pi)^{Kn/2} \left\{\det(\Sigma_*)\right\}^{n/2}} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{n} (y_i - \mu_* s(x_*))' \Sigma_*^{-1} (y_i - \mu_* s(x_*))\right)$$

$$\Sigma_* = \Sigma + (\sigma_*^2 - 1) s(x_*) s(x_*)'$$

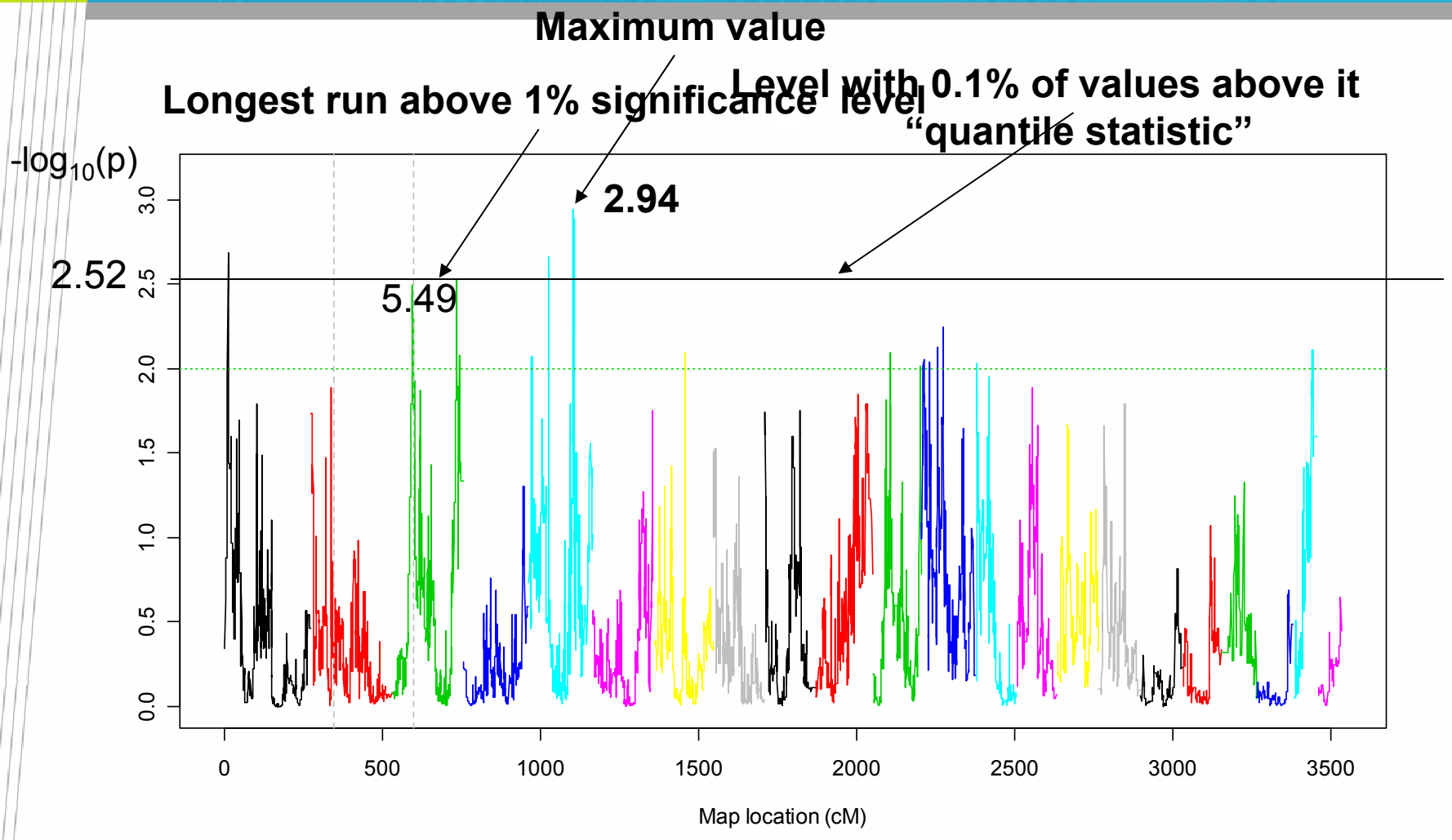
Linkage Model and data



Properties of genome-wide summary statistics

- **To determine if the 57241 pointwise LR values are together enough to indicate the presence of an effect requires us to define a "genome-wide summary" (GWS) statistic and see whether this statistic is bigger than expected by chance**
- **Simulating the null distribution lets us determine the critical points for any required GWS statistic, then simulating the alternative for a given genetic link gives the power to detect that link.**

Genome-wide summary (GWS) statistics



Power of GWS statistics

- **Power estimates based on 10000 simulation runs**

n_1	n_2	Maximum run above		Quantile Statistic			
		1%	0.1%	10%	1%	0.1%	Max
500	0	75.2%	81.5%	30.1%	78.6%	85.0%	81.9%
0	100	85.3%	90.6%	31.6%	87.5%	92.4%	91.0%

1% or 0.1% quantile statistics generally give the greatest power

Single DS gene – 0.1% quantile statistic

Confidence level 95%

n	n1	n2	Power
300	200	100	90%
300	100	200	100%
300	0	300	100%
500	300	200	100%
500	200	300	100%
500	0	500	100%
1000	500	500	100%
2000	1000	1000	100%
4000	2000	2000	100%
10000	5000	5000	100%

Two DS genes – 0.1% quantile statistic

Confidence level 95%

n	n1	n2	Power
300	200	100	40%
300	100	200	71%
300	0	300	89%
500	300	200	76%
500	200	300	92%
500	0	500	99%
1000	500	500	100%
2000	1000	1000	100%
4000	2000	2000	100%
10000	5000	5000	100%

Five DS genes – 0.1% quantile statistic

Confidence level 95%

n	n1	n2	Power
300	200	100	13%
300	100	200	19%
300	0	300	26%
500	300	200	19%
500	200	300	30%
500	0	500	47%
1000	500	500	56%
2000	1000	1000	92%
4000	2000	2000	100%
10000	5000	5000	100%

GWS significance results

1. Test data

- With a 11 2-affected and 17 1-affected pairs:
- Total run length 19.1 (79 for 5% significance)
- 0.1% quantile statistic 2.75 (3.538 for 5% significance)

- **Not surprisingly – nonsignificant genome-wide**

2. Latest results for major study (~200 pairs)

	Model 3 0.1% QS	Mean IBD 0.1% QS	Model 3 Maximum
5% Critical level	*3.58*	*3.58*	*4.01*
Data value	**4.02**	**3.96**	**4.67**

- **Clear evidence of significant linkage**
- **But where in the genome?**

Bayesian estimation of gene location

- **The model is specified in terms of three parameters**
  - μ_* – the strength of the association with disease – mean of Y at disease locus
  - σ_* – the variation between individuals in the strength of association – sd of Y at disease locus
  - x_* – the location of the gene
- **μ_* and σ_* are determined by the penetrance and allele frequencies of the disease susceptibility (DS) gene**
- **μ_* is 0 if there is no DS gene**
- **The analysis gives probability distributions for the three parameters which can be plotted to graphically illustrate their possible values**

Posterior distribution of parameters

- **Likelihood for *n* sib pairs**

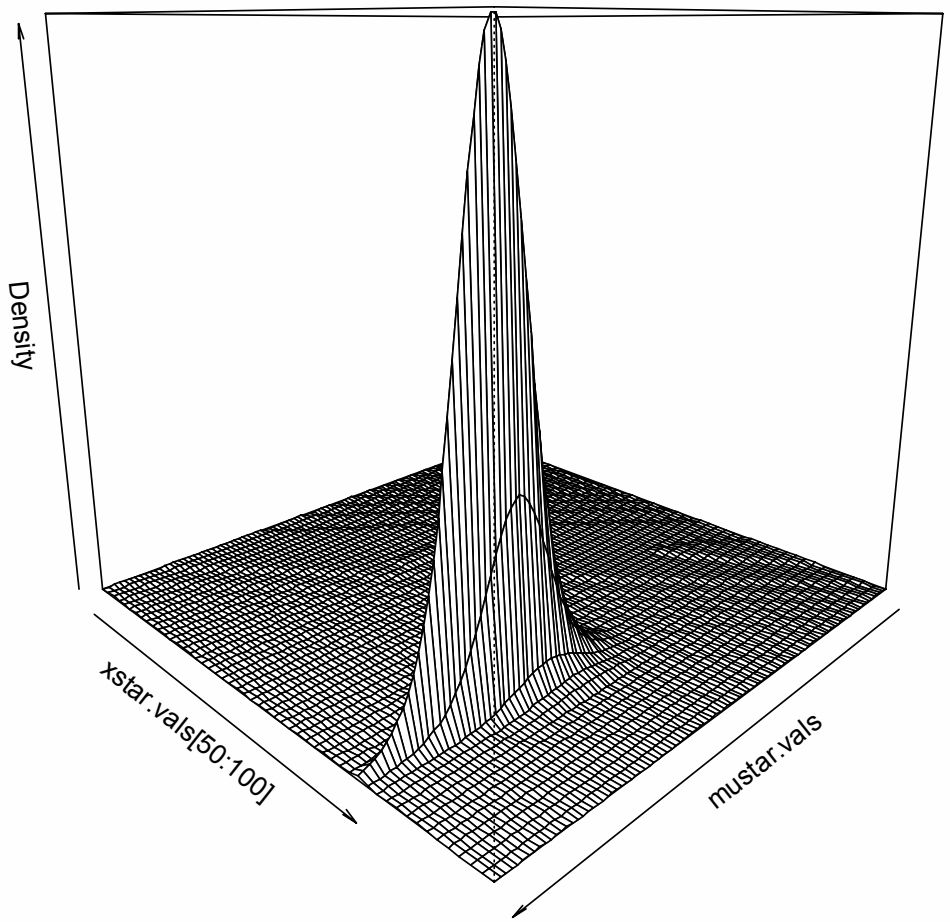
$$L(x_*, \mu_*, \sigma_*^2) = \frac{1}{(2\pi)^{Kn/2}\{\det(\Sigma_*)\}^{n/2}} \exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}(y_i - \mu_* s(\lambda_*))' \Sigma_*^{-1} (y_i - \mu_* s(\lambda_*))\right)$$

- **Given a prior distribution *p* for the parameters, the posterior density is proportional to**

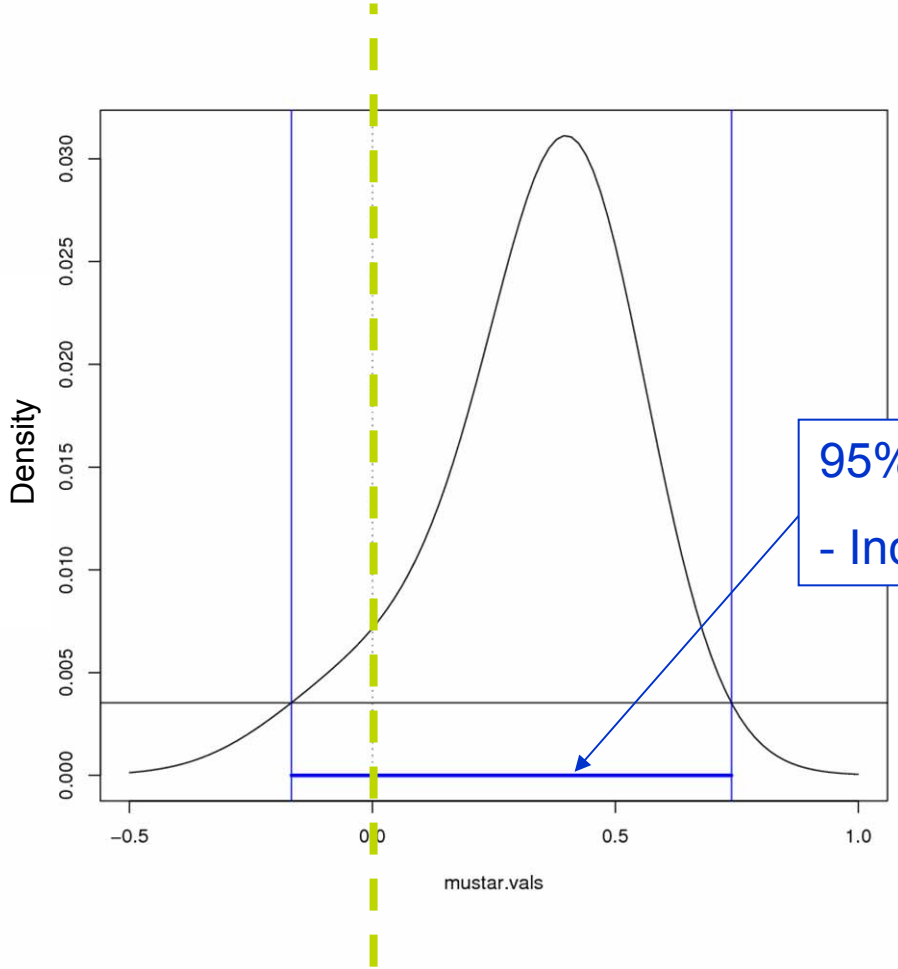
$$L(x_*, \mu_*, \sigma_*^2)\, p(x_*, \mu_*, \sigma_*^2)$$

- **If the likelihood can be computed on a sufficiently dense grid, covering most of the likely range of the parameters, the posterior density can be obtained simply by dividing the computed values by their sum**
- **This allows the computation of posterior probability intervals for individual parameters and also joint distributions of pairs of parameters**
- **For simplicity, used uniform prior – no information**

Joint distribution of location and “strength”



Chromosome A: marginal distribution of “strength”

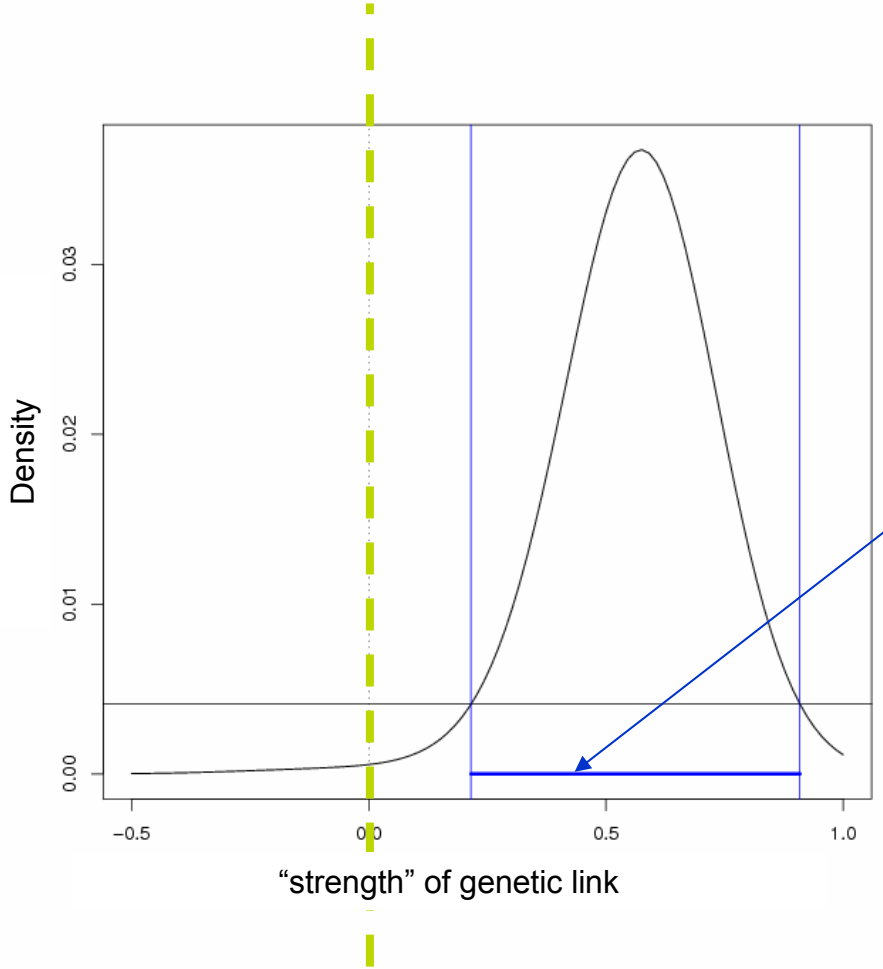


95% probability interval

- Includes zero

No evidence of DS gene on Chromosome A

Chromosome B : marginal distribution of “strength”

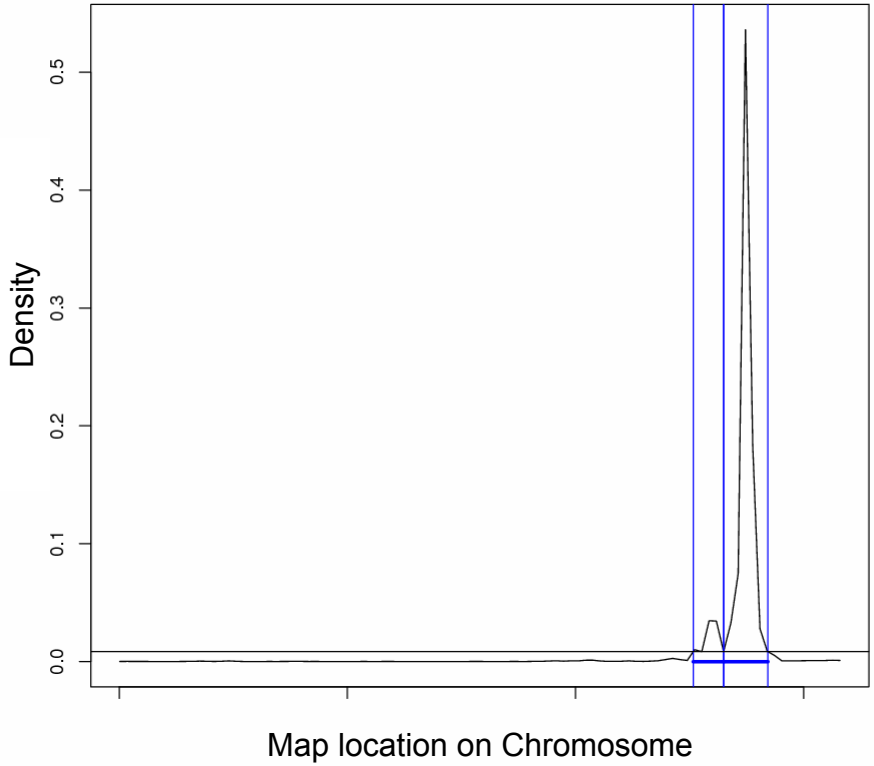


Strong evidence of DS gene on Chromosome B

95% probability interval

- Excludes zero

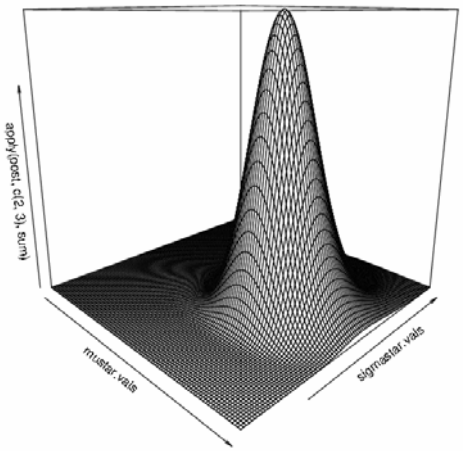
Chromosome B

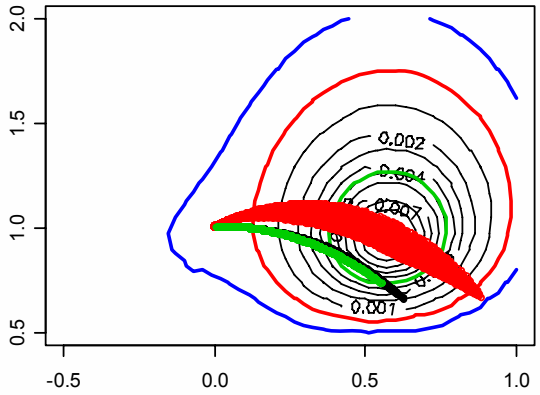


- **Strongest evidence for gene**
- **Most likely location within about 5 centimorgans**

Disease model

- The joint distribution of μ and σ gives information about the likely disease model
- It can be presented as a contour plot where the peak of the “mountain” represents the most likely values
- The coloured “banana” shapes represent a range of possible models – the red area represents recessive models, green is additive and black is dominant.
- It can be seen that recessive models are more consistent with the data



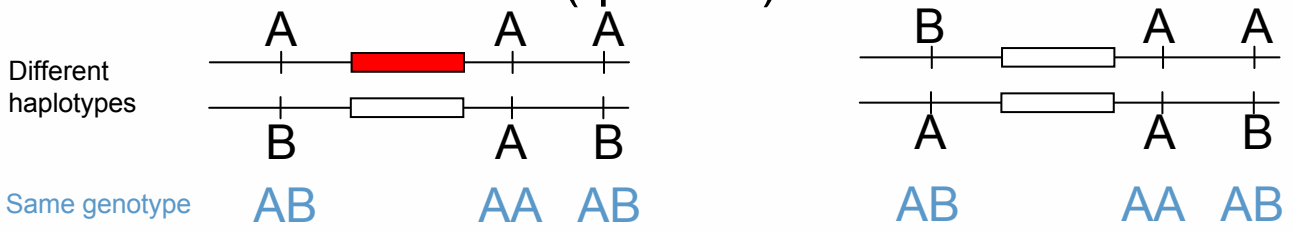


Results of gene location modelling

- **The analysis has identified a number of promising regions for further study by fine mapping**
- **The strongest signals are on Chromosomes … and these should be given priority.**
- **There is a suggestion that recessive inheritance is more likely than dominant or additive models.**

Where to from here … ?

- **Association studies:**
  - The higher SNP densities make it likely that disease genes will be associated with SNP patterns – with 1000000 SNPs the average separation is 3kb
  - To do this effectively requires knowing which SNP patterns are on which of the two copies of the chromosome (“phase”)



  - Associated SNP sequence ‘masked’ by other copy

  - Ongoing research in this area – Huwaida Rabie

Where to from here … ?

- **Issues for higher densities still**
  - The first complete genome – all 3Gb – sequenced for less than $US 1m has just been released (James Watson).
  - Forget “coding”, “nonsynonymous” etc – just get everything!
  - 11,883,685 SNPs and lots of more complex forms of variation
  - Phase may still be a problem
  - False positives???
  - Generic methods – Bonferroni correction, False Discovery Rate, etc will risk losing signal among the noise
  - Need to incorporate other biological knowledge
    - Effects are mediated by proteins working in complex metabolic processes which are partly understood
    - Changes in coding bases affect behaviour of proteins in partly known ways
  - How can such partial prior knowledge be modelled in a way that will allow it to be built in to analysis?
    - Bayesian methods or equivalent penalised frequentist methods

Where to from here … ?

- **Issues for data integration**
  - This has been about “genomics”, but there seems to be a new “-omics” invented every day: proteomics, metabolomics, transcriptomics, interactomics, metagenomics, YF-omics each with massive databases of varying accuracy
  - Integrating these has many complex issues:
    - Modelling or data mining?
    - Gene expression and genotype – cis- and trans-acting genes
    - Multiple data levels - …, cell, …, tissue, …, organism, … with different experiments and technologies collecting data at each level
    - Highly nonlinear processes: “kinase kinase kinase”
    - Study design – levels of variation and replication
    - Integration between different research teams
    - Coping with “observational” data where experiments are not possible (eg human studies)
- **Many of these are statistical rather than biological or computational issues**

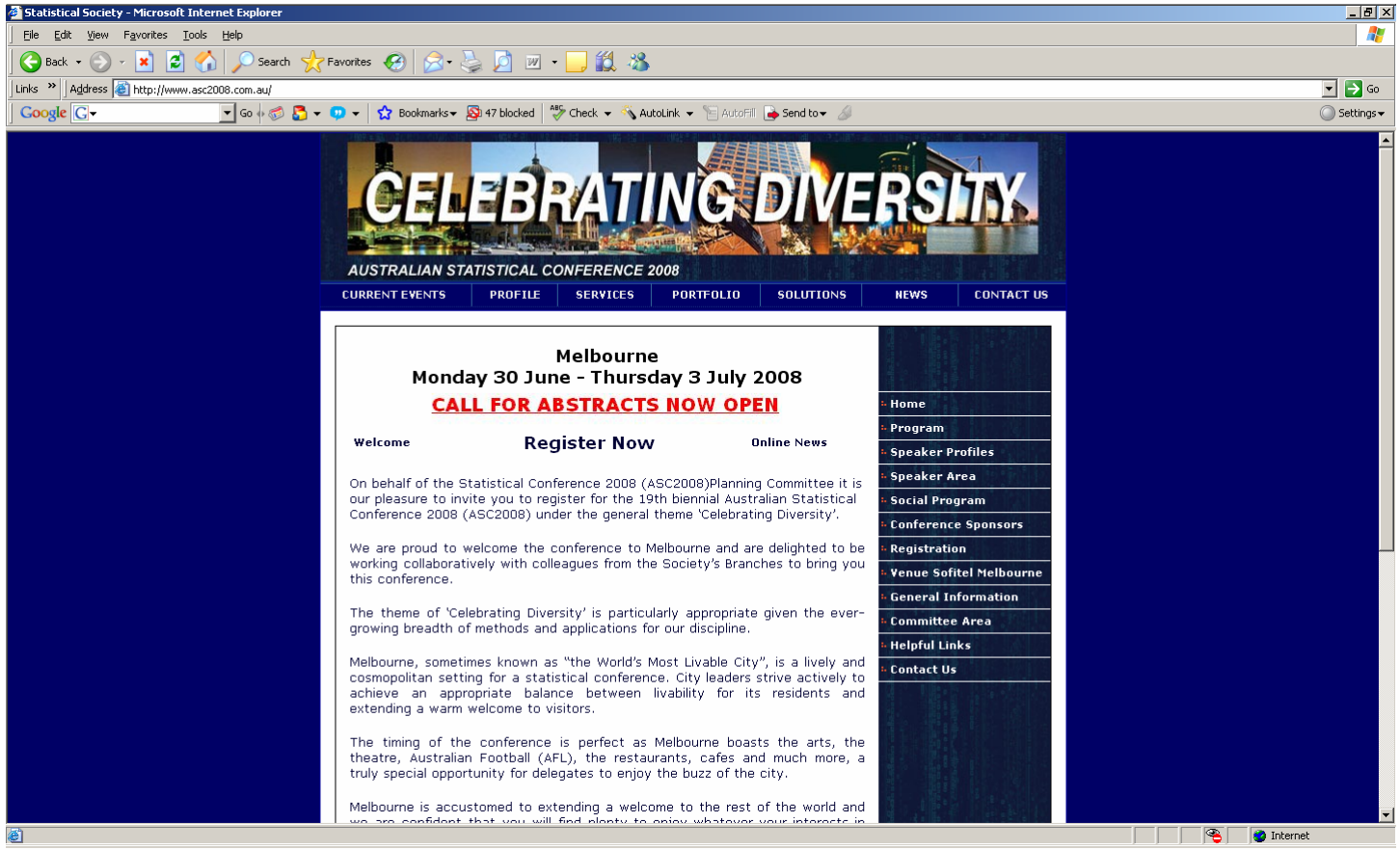
Where to from here … ?

- **Issues for data integration**
  - This has been about “genomics”, but there seems to be a new “-omics” invented every day: proteomics, metabolomics, transcriptomics, interactomics, metagenomics, YF-omics each with massive databases of varying accuracy
  - Integrating these has many complex issues:
    - Modelling or data mining?
    - Gene expression and genotype – cis- and trans-acting genes
    - Multiple data levels - …, cell, …, tissue, …, organism, … with different experiments and technologies collecting data at each level
    - Highly nonlinear processes: “kinase kinase kinase”
    - Study design – levels of variation and replication
    - Integration between different research teams
    - Coping with “observational” data where experiments are not possible (eg human studies)
- **Many of these are statistical rather than biological or computational issues**

Final comments

- **Modern biological research is critically dependent on management and analysis of large amounts of complex data**
- **The processes underlying the data and the interactions between them are also complex, but there is growing understanding of them**
- **Integration of information and studies is key**
- **Many of the issues are statistical**
- **There is lots of fun to be had for statisticians both in the analysis and in the mathematical developments**

Australian Statistical Conference
Melbourne 30 June – 3 July 2008



http://www.asc2008.com.au/

Acknowledgements

Colleagues:

- **University of Melbourne**
  - **John Hopper**
  - **Mark Jenkins**
  - **Melissa Southey**
- **Flinders Medical Centre**
  - **Graham Young**
- **Royal Melbourne Hospital**
  - **Finlay Macrae**
- **CSIRO Preventative Health Research Flagship**
  - **Garry Hannan**
  - **Jesper Brohede**
  - **Jason Ross**
- **CSIRO Mathematical and Information Sciences**
  - **Huwaida Rabie**
  - **Mike Buckley**
- **Keio University**
  - **Yuki Sugaya**