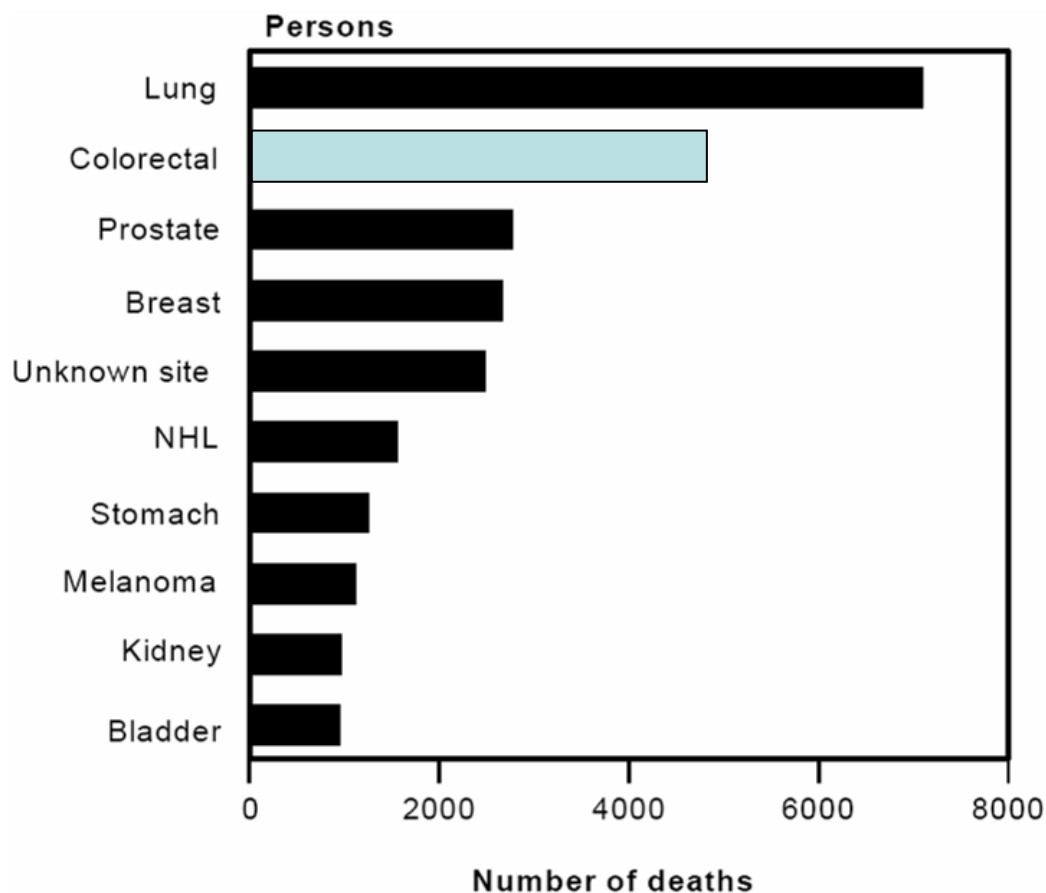


## Locating genetic links to disease using SNP Chips

Ian Saunders

CSIRO Preventative Health Research Flagship  
CSIRO Mathematical and Information Sciences

# Colorectal cancer

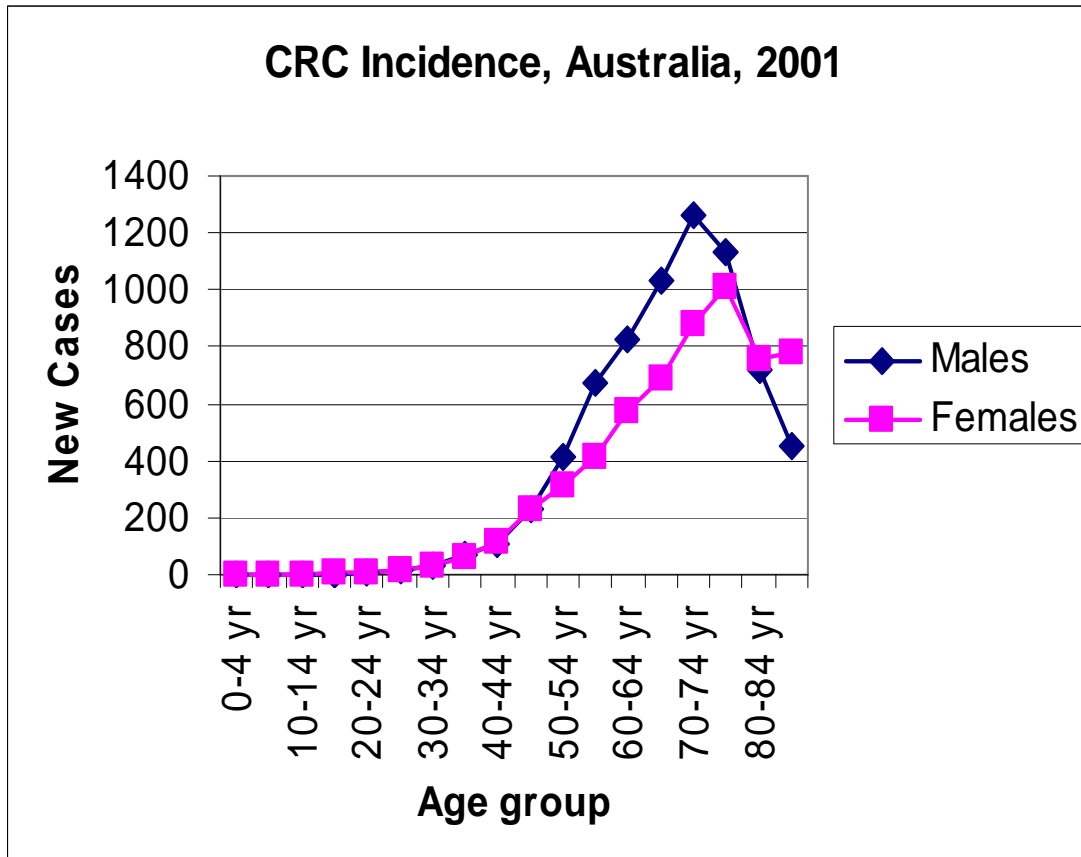


- CRC is the second highest cause of cancer-related death in Australia (3<sup>rd</sup> world wide).

- Incidence of CRC has risen slightly over the past 40 years.

- Deaths from CRC have decreased slightly over the past 20 years.

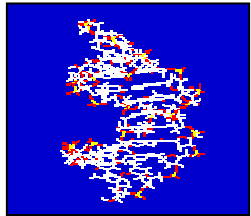
Source: AIHW (2004), *Cancer in Australia, 2001*



- Australia has a higher incidence of CRC than most developed countries: (NZ>Aust>US>Can>UK)
- 12500 new cases p.a.
- CRC is a disease of the affluent.
- Risk of CRC increases dramatically after age 50



Our goal: CRC-specific early diagnosis.



Risk

assessment



Early detection

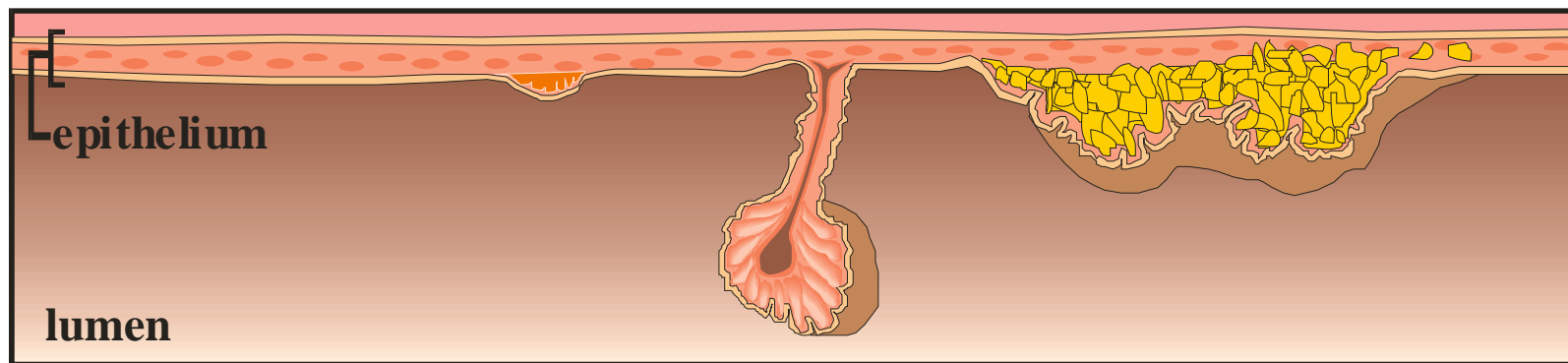


Diagnosis

Monitoring



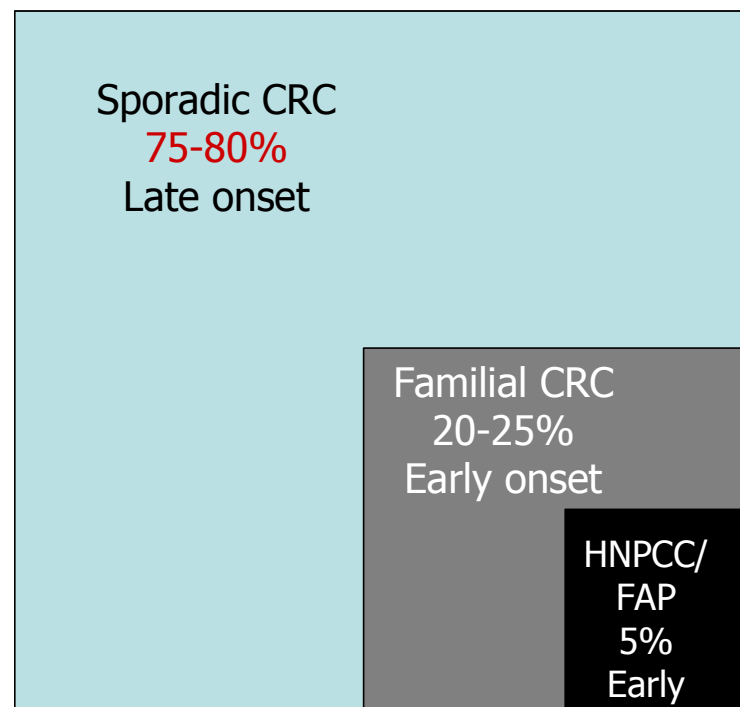
Normal → Hyperplasia → Adenoma → Adenocarcinoma →



(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.



## 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...

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]aaaatggttactgtgg	GSYM:MOT8; AC	0.48

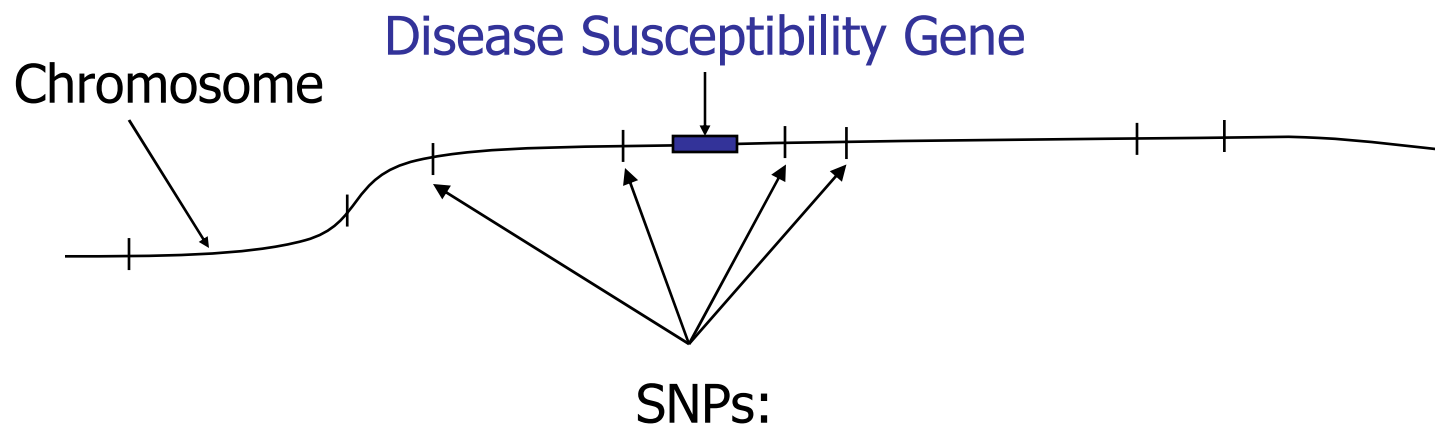
- Millions of well characterised SNPs are now available – 1,467,365 on Chr 1 (2005)

# 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: 1,000,000 SNP array
- Staining, scanning and genotype calling fully automated



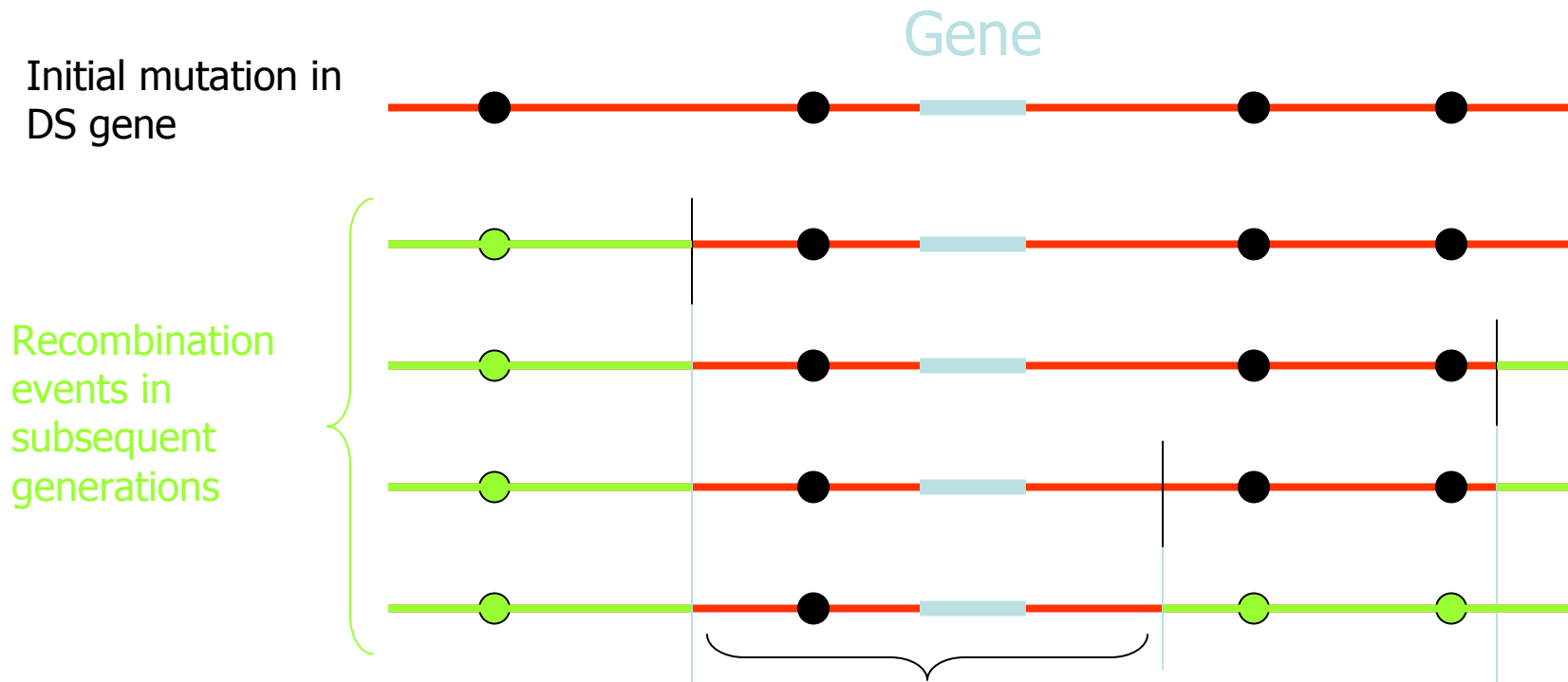
## Using SNPs in linkage studies



SNPs:  
Spots where individual nucleotides may differ between individuals

- **For 100k SNP chip: 4 SNPs per gene**
- **Functional genes make up about 1% of genome**
- **SNPs on chip probably not causative**
- **Looking for linkage/association - 'markers' - rather than a functional relationship**

# History of a mutation



- A SNP near the DS gene will remain associated with it: 'linkage'
- Range of linkage across the earth's population  $\sim 3\text{kb}$   
(cf. 18kb on Affymetrix Xba 50k SNP chip)



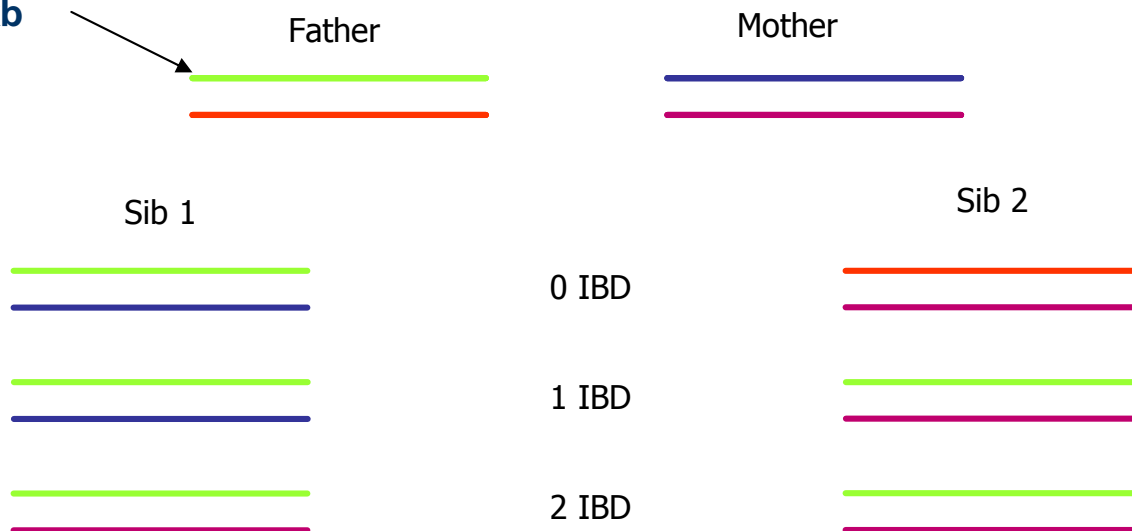
# Sib-based studies

For late onset diseases like CRC, parents often are not available for genotyping

Methods based on siblings (sibs) still feasible

Test based on probabilities of sharing 0, 1 or 2 alleles IBD in sib pairs where one or both have the disease

'Short' section of DNA –  
no recombinations in ONE  
generation - ~1000kb





## IBD Probabilities for sib pairs conditional on disease state

For a random pair of sibs at a randomly chosen locus

$$\Pr(0 \text{ IBD}) = 1/4$$

$$\Pr(1 \text{ IBD}) = 1/2$$

$$\Pr(2 \text{ IBD}) = 1/4$$

The numbers IBD at a point near a disease gene will be different in sib pairs with both affected

$$\Pr(2 \text{ affected} | i \text{ IBD})$$

$$= \sum_{\text{genotypes of siblings}} \Pr(2 \text{ affected} | \text{genotypes}) \Pr(\text{genotypes} | i \text{ IBD})$$

$$= \Pr(2 \text{ affected} | AA, AA) \Pr(AA, AA | i \text{ IBD})$$

$$+ \Pr(2 \text{ affected} | AA, AB) \Pr(AA, AB | i \text{ IBD}) + \dots$$

$$= f(AA) f(AA) \Pr(AA, AA | i \text{ IBD})$$

$$+ 2 f(AA) f(AB) \Pr(AA, AB | i \text{ IBD}) + \dots$$

$$f(g) = \Pr(\text{affected} | \text{genotype } g)$$

And then  $\Pr(i \text{ IBD} | 2 \text{ affected}) = \Pr(2 \text{ affected} | i \text{ IBD}) \Pr(i \text{ IBD}) / \Pr(2 \text{ affected})$

**Table 1. Probability of genotype given IBD status**

$\Pr(G_1, G_2   I, p)$	$I$		
	0	1	2
$\{G_1, G_2\}$			
0,0	$p^4$	$p^3$	$p^2$
0,1	$4p^3q$	$2p^2q$	0
0,2	$2p^2q^2$	0	0
1,1	$4p^2q^2$	$pq$	$2pq$
1,2	$4pq^3$	$2pq^2$	0
2,2	$q^4$	$q^3$	$q^2$

$p = \Pr(\text{B allele})$   $q = \Pr(\text{A allele})$



## Prob(IBD|Affected status)

The formulae don't simplify in any useful way.

Results for a possible model for CRC incidence and penetrance

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

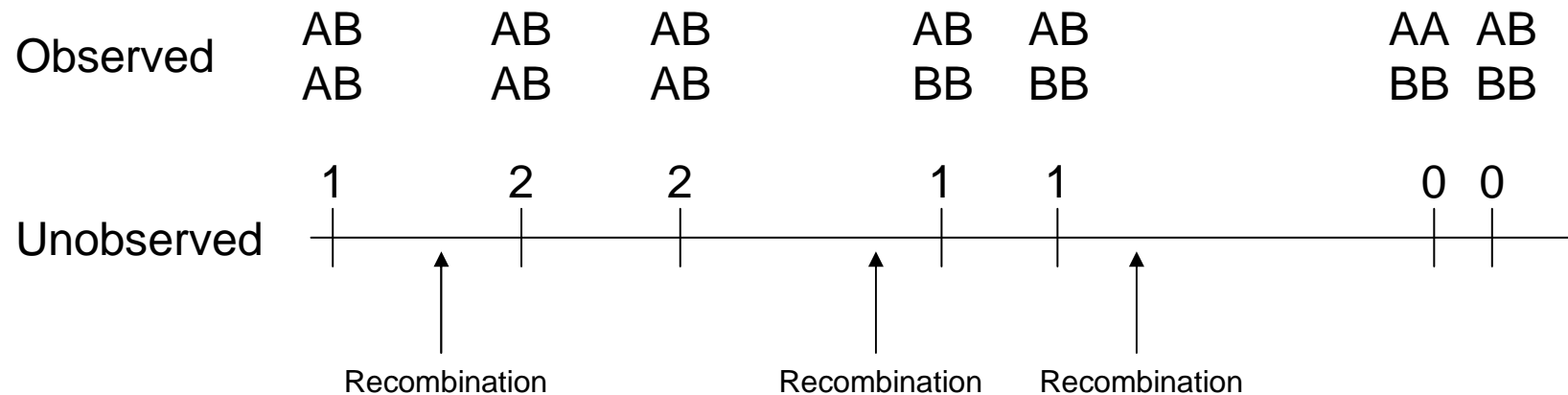
LR test statistic at SNP  $k$  for difference from no linkage is a linear combination  $Y_k = w'l$  of counts of number of sib pairs in each IBD class

But IBD status is not directly observable. How do we deduce it from SNP genotypes of the sibs?



## Determining IBD status

The IBD status between two siblings at a sequence of SNPs forms a nonstationary Markov chain with transition matrix determined by the recombination rate

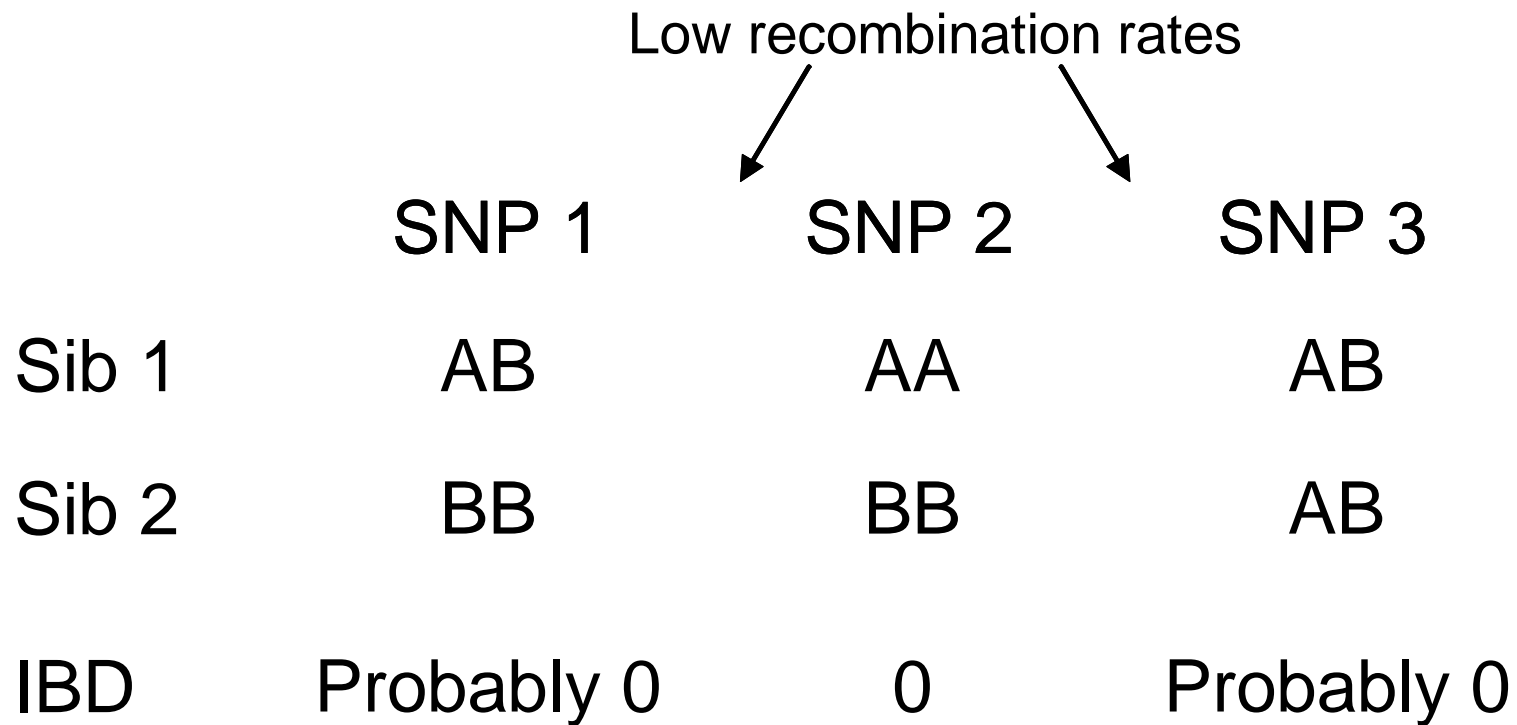


The genotype we observe is a random function of the unobserved IBD state so the MC is “hidden” – a HMM

We can infer back from the observed genotype to the hidden IBD state – close to 100% accuracy



## Basic idea





## Notation

Denote by  $X_k$  the pair of genotypes for the two siblings at SNP  $k$

And by  $I_k$  the IBD status at SNP  $k$

We want to find  $\Pr(I_k=i|X_1=x_1, \dots, X_K=x_K)$

Then write

$$g^{(k)}(x; i) = \Pr(X_k = x | I_k = i).$$

$$\pi_{ij}^{(k)} = \Pr(I_{k+1} = j | I_k = i)$$

$$\phi_k(x_1, \dots, x_k; i) = \Pr(X_1 = x_1, \dots, X_k = x_k, I_k = i)$$

$$\psi_k(x_k, \dots, x_K; i) = \Pr(X_k = x_k, \dots, X_K = x_K | I_k = i)$$

(Note that  $\phi$  is a joint probability while  $\psi$  is a conditional probability)



## Transition matrix, $\pi_{ij}(\theta)$

**Table 1. Transition probabilities of Markov Chain  $I_k$**

$\pi_{ij}(\theta)$	$j=0$	$j=1$	$j=2$
$i=0$	$(1-\theta)^4 + 2\theta^2(1-\theta)^2 + \theta^4$	$4\theta(1-\theta)^3 + 4\theta^3(1-\theta)$	$4\theta^2(1-\theta)^2$
$i=1$	$2\theta(1-\theta)^3 + 2\theta^3(1-\theta)$	$(1-\theta)^4 + 6\theta^2(1-\theta)^2 + \theta^4$	$2\theta(1-\theta)^3 + 2\theta^3(1-\theta)$
$i=2$	$4\theta^2(1-\theta)^2$	$4\theta(1-\theta)^3 + 4\theta^3(1-\theta)$	$(1-\theta)^4 + 2\theta^2(1-\theta)^2 + \theta^4$

$$\begin{aligned}
 & \Pr(X_1 = x_1, \dots, X_K = x_K, I_k = i) \\
 &= \sum_{i_{k-1}, i_{k+1}} \Pr(X_1 = x_1, \dots, X_K = x_K, I_k = i, I_{k-1} = i_{k-1}, I_{k+1} = i_{k+1}) \\
 &= \sum_{i_{k-1}} \phi_{k-1}(x_1, \dots, x_{k-1}; i_{k-1}) \pi_{i_{k-1}i}^{(k-1)} \\
 &\quad \times g^{(k)}(x_k; i) \\
 &\quad \times \sum_{i_{k+1}} \pi_{ii_{k+1}}^{(k)} \psi_{k+1}(x_{k+1}, \dots, x_K; i_{k+1})
 \end{aligned}$$

$$\begin{aligned}
 \phi_k(i) &= \Pr(X_1 = x_1, \dots, X_k = x_k; I_k = i) \\
 &= \sum_{i_{k-1}} \Pr(X_1 = x_1, \dots, X_{k-1} = x_{k-1}; I_{k-1} = i_{k-1}) \\
 &\quad \times \Pr(I_k = i \mid I_{k-1} = i_{k-1}) \\
 &\quad \times \Pr(X_k = x_k \mid I_k = i) \\
 &= g^{(k)}(x_k; i) \sum_{i_{k-1}} \phi_{k-1}(x_1, \dots, x_{k-1}; i_{k-1}) \pi_{i_{k-1}i}^{(k-1)}
 \end{aligned}$$

$$\begin{aligned}
 \psi_k(i) &= \Pr(X_k = x_k, \dots, X_K = x_K \mid I_k = i) \\
 &= \sum_{i_{k+1}} \Pr(X_{k+1} = x_{k+1}, \dots, X_K = x_K \mid I_{k+1} = i_{k+1}) \\
 &\quad \times \Pr(I_{k+1} = i_{k+1} \mid I_k = i) \\
 &\quad \times \Pr(X_k = x_k \mid I_k = i) \\
 &= g^{(k)}(x_k; i) \sum_{i_{k+1}} \pi_{ii_{k+1}}^{(k)} \psi_{k+1}(x_{k+1}, \dots, x_K; i_{k+1})
 \end{aligned}$$



## Some typical results

### A short section of Chromosome 6 for a single pair of siblings

Sibling 1	BB	AA	AA	AA	BB	BB	AA	AA	AA	AA	BB	AA
Sibling 2	BB	AA	AA	AA	AB	BB	BB	AB	AA	AA	BB	AA
PrA	0.58	0.88	0.88	0.95	0.26	0.31	0.92	0.89	0.55	0.44	0.16	0.70
theta	0.2	0	110	0.6	161	104	3.3	49.8	11.9	0.8	0.8	
Pr(IBD=0)	0.00	0.00	0.00	0.00	0.00	0.55	1.00	0.75	0.08	0.01	0.01	0.01
Pr(IBD=1)	0.00	0.00	0.00	0.89	1.00	0.67	0.00	0.25	0.89	0.95	0.95	0.95
Pr(IBD=2)	1.00	1.00	1.00	0.11	0.00	0.00	0.00	0.00	0.03	0.04	0.04	0.04
<b>IBD</b>	<b>2</b>	<b>2</b>	<b>2</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>0</b>	<b>0</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>1</b>

**Recombinations**

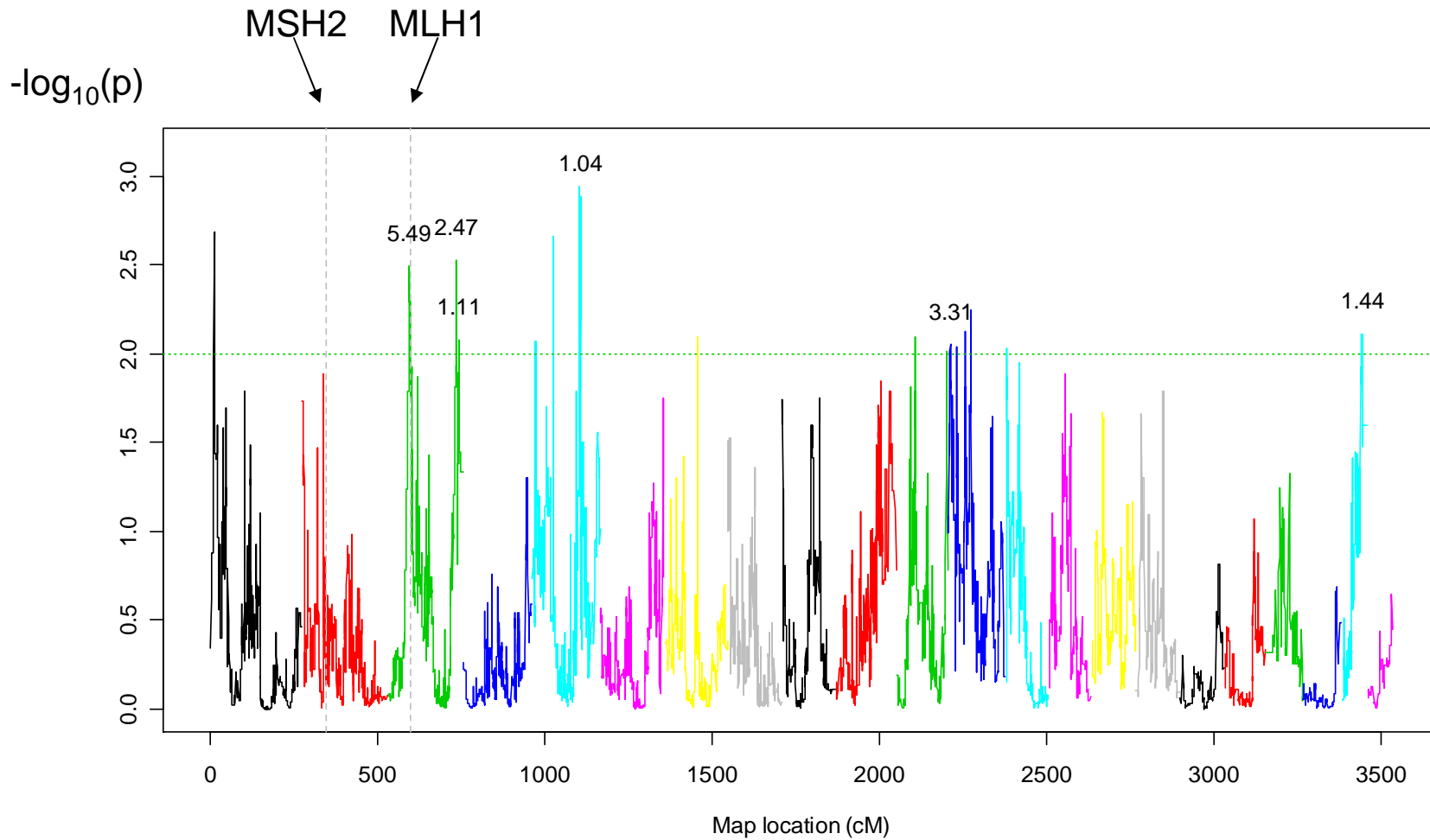
**Results for 28 sib pairs known to have MLH1 or MSH2 mutation.  
Genes near SNPs 4550 and 8523.**

SNP		4549	4550	4551	...	8522	8523	8524	
<b>1 Affected:</b>	<b>0 IBD</b>	3	3	3		3	3	3	
	<b>1 IBD</b>	9	9	9		8	8	8	
	<b>2 IBD</b>	5	5	5		6	6	6	
<b>2 affected:</b>	<b>0 IBD</b>	2	2	2		1	1	1	
	<b>1 IBD</b>	6	6	6		3	3	3	
	<b>2 IBD</b>	3	3	3		7	7	7	
	<b><math>-\log_{10}(\text{p-value})</math></b>	<b>0.41</b>	<b>0.41</b>	<b>0.41</b>		<b>2.1</b>	<b>2.1</b>	<b>2.1</b>	

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



# Results for 28 (of planned 300) sib pairs





# Genotypes uninformative

Results for the 40 individuals from the HNPCC sib pairs (+ some others)

MSH2



MLH1



SNP		4546	4547	4548	4549	4550	...	8521	8522	8523	8524	8525
Aff	AA	12	4	10	16	15		0	4	0	2	12
	AB	7	10	8	3	3		0	9	0	5	5
	BB	0	5	1	0	1		19	6	19	12	2
Unaff	AA	17	4	8	17	13		0	2	0	2	15
	AB	1	12	11	4	7		0	11	0	4	4
	BB	3	5	2	0	1		21	8	21	15	2
$-\log_{10}(p)$		0.20	0.02	0.36	0.10	0.02		0.00	0.23	0.00	0.07	0.07

72kb      56kb

19kb      476kb

No association based on genotypes



## Simulation of sequence of test statistics

**We can calculate pointwise test statistics, but we now have 58000 of them with strong correlation.**

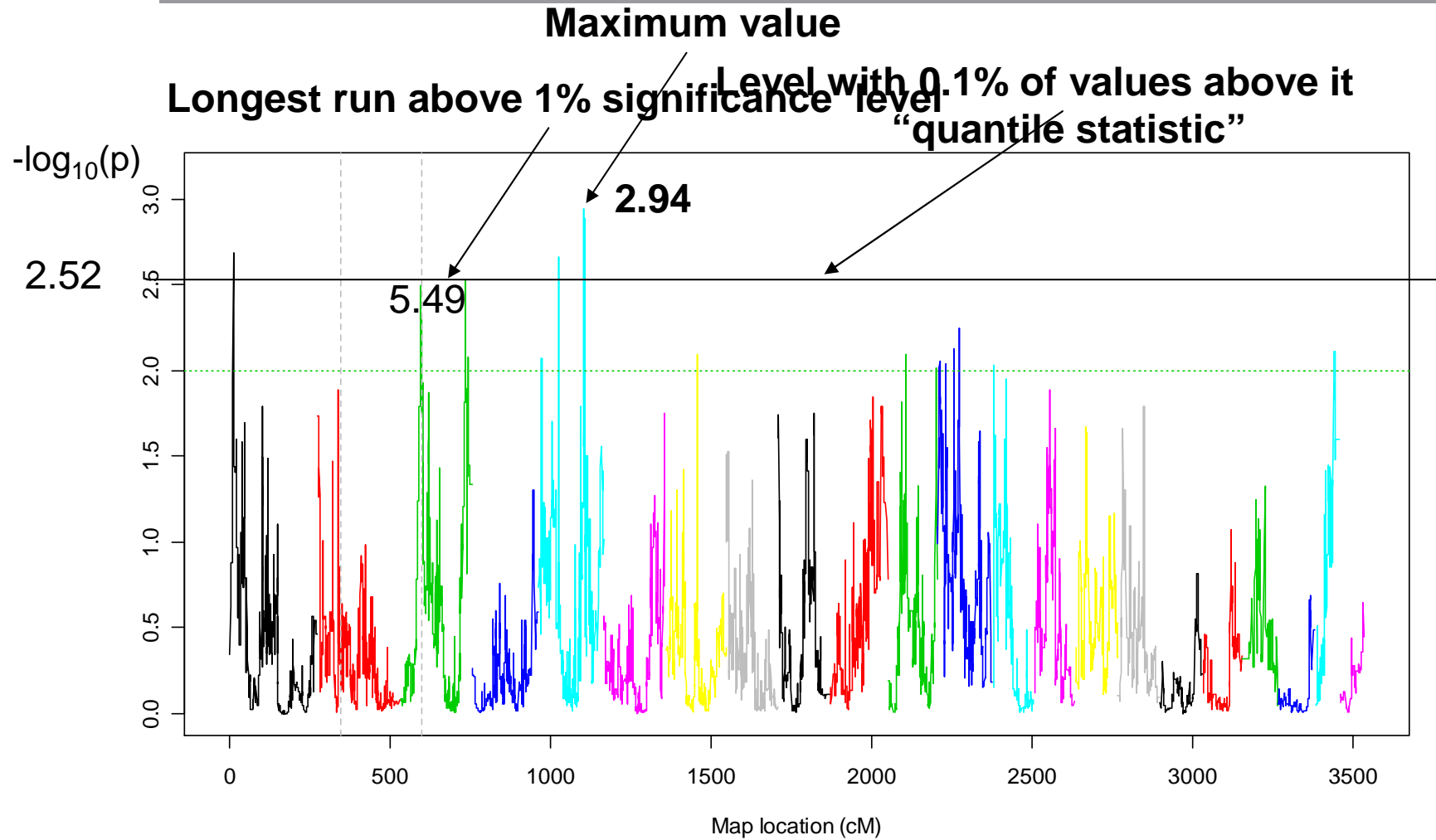
**It turns out that the sequence of statistics  $Y_k$  can be approximated by an autoregressive (Markov) process which does not depend strongly on the alternative disease model.**

**The presence of a disease susceptibility genes at  $G$  alters the distribution of  $Y_G$ , but the dependence remains the same, so we just have to simulate an AR process conditional on its value at one or more points.**

**Simulating the null distribution lets us determine the critical points for any required test statistic, then simulating the alternative for a given genetic link gives the power to detect that link.**

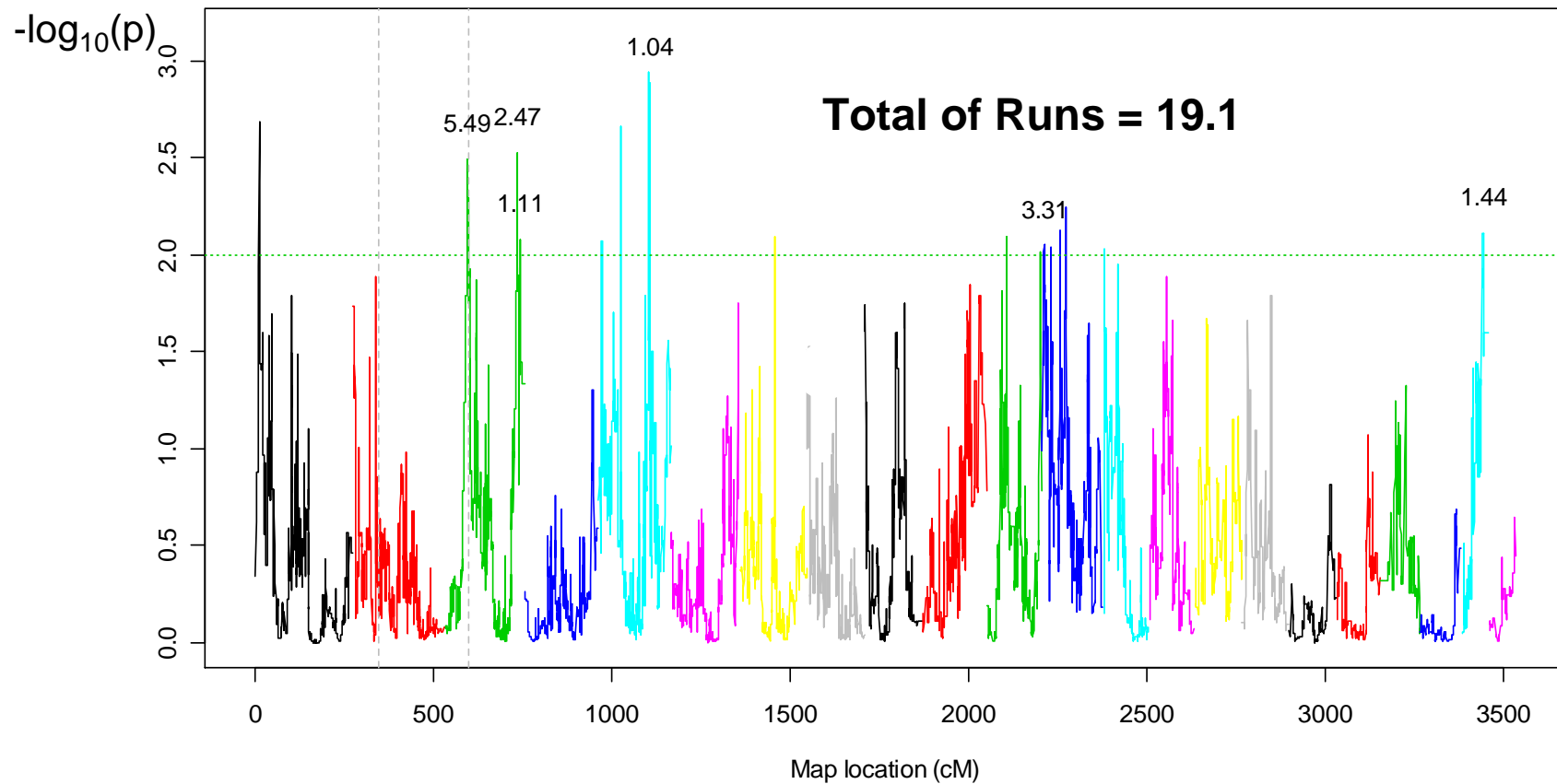


# Test statistics





# Test statistics





## Which statistic?

---

### Maximum run length (centimorgans) for genome-wide significance

---

Genome-wide Significance level	Pointwise significance		
	.01	.005	.001
10%	26.69	18.82	6.18
5%	33.97	25.12	10.51
1%	50.96	39.15	20.61

---



---

### Quantile of empirical distribution for genome-wide significance

---

Genome-wide Significance level	Quantile				
	10%	5%	1%	0.1%	Max
10%	1.537	1.932	2.696	3.334	3.660
5%	1.614	2.035	2.857	3.538	3.865
1%	1.770	2.242	3.215	3.950	4.268



## Which statistic?

$n_1$	$n_2$	Maximum run		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%

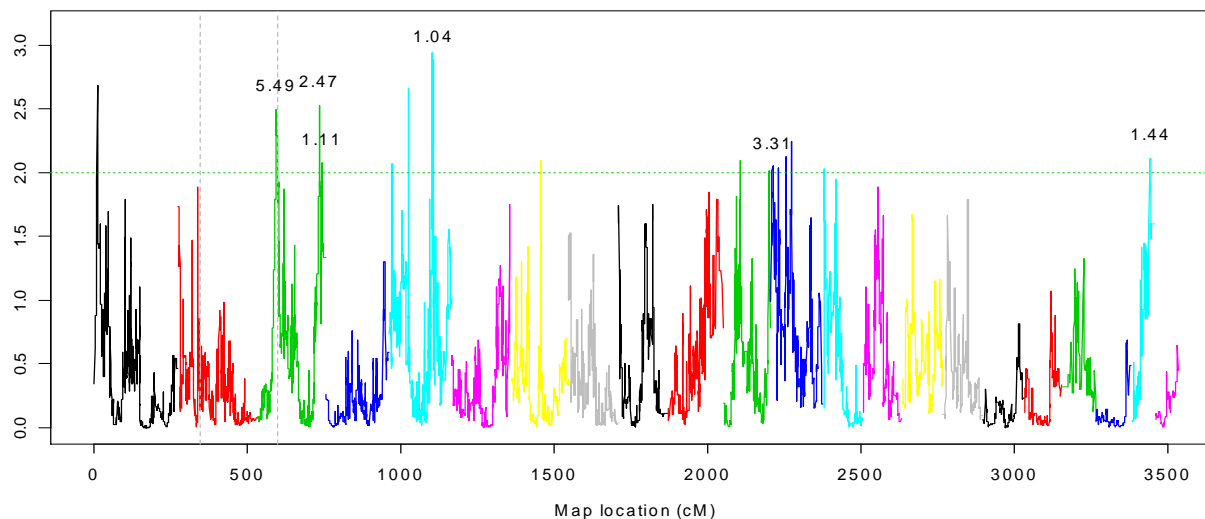
n	n1	n2	Power
300	200	100	13%
300	100	200	19%
300	0	300	26%
500	300	200	19%
500	200	300	30%
500	0	500	47%
1000	500	500	56%
2000	1000	1000	92%
4000	2000	2000	100%
10000	5000	5000	100%

## Results for test study

**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)**



**If the counts had been twice as large:**

**Total run length 198 (79 for 5% significance)**

**0.1% quantile statistic 3.89 (3.538 for 5% significance)**



## Summary

**A simple model give understanding of a complex situation**

**Other summary statistics or more complex models are easily included once the basic situation is understood**

**Even relatively small datasets can contain useful information**



# Acknowledgements

## Colleagues:

**University of Melbourne**  
**John Hopper**  
**Mark Jenkins**

## **IARC**

**Melissa Southey**

**Flinders Medical Centre**  
**Graham Young**

**Royal Melbourne Hospital**  
**Finlay Macrae**

**- collecting the samples**

**CSIRO Preventative Health Research Flagship**  
**Garry Hannan**  
**Jesper Brohede**

**- running the machinery**

**So that I just had to analyse the numbers!**