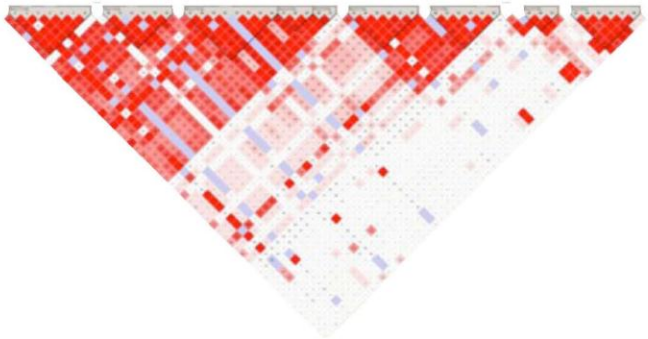
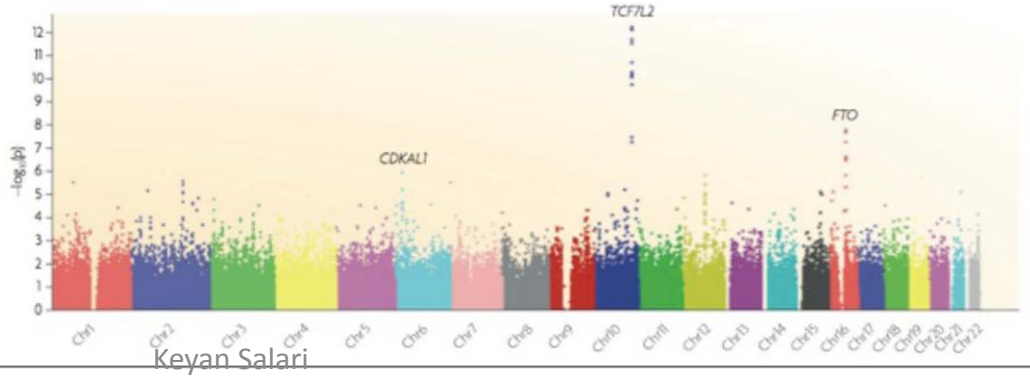
1. Natural variation in the human genome

2. Genetic Association & Linkage Disequilibrium



3. Genome-wide association studies



Hum Genet (2002) 111 : 75–87
DOI 10.1007/s00439-002-0763-x

ORIGINAL INVESTIGATION

Stephanie M. Fullerton · Andrew G. Clark
Kenneth M. Weiss · Scott L. Taylor · Jari H. Stengård
Veikko Salomaa · Eric Boerwinkle
Deborah A. Nickerson

Sequence polymorphism at the human apolipoprotein AII gene (*APOA2*): unexpected deficit of variation in an African-American sample

Sequence APOA2 in 72 people

Look at patterns of polymorphisms

Chimp		Site no.[a]													
SNP haplotype no.	Sequence haplotype no.	1 5 5 C	2 0 1 G	1 8 7 2 T	1 2 1 8 G	2 6 7 1 ?	2 0 3 8 G	2 0 8 5 C	2 1 1 5 G	2 2 3 3 C	2 8 1 8 C	2 8 6 8 C	3 9 9 4 C	3 0 2 7 T	3 0 9 2 A	2 0 8 G

Find polymorphisms at these positions.

Reference sequence is listed.

Chimp		Site no.[a]															Sample			
SNP haplotype no.	Sequence haplotype no.	1 5 5 C	2 0 1 G	1 8 7 2 T	1 2 1 8 G	2 6 7 1 ?	2 0 3 8 G	2 0 8 5 C	2 1 1 5 G	2 2 3 3 C	2 8 1 8 C	2 8 6 8 C	3 9 9 4 C	3 0 2 7 T	3 0 9 2 A	2 0 8 G	J	N	R	T
Core re-sequenced samples																				
	S9	G		C		20	●		A	●	●	●	●	●	●	●		0	0	1	1

Sequence of the first chromosome.

Circle is same as reference.

Chimp		Site no.[a]															Sample			
SNP haplotype no.	Sequence haplotype no.	15 5 C	20 1 G	18 7 2 T	12 1 8 G	26 7 1 ?	20 3 8 G	20 8 5 C	21 1 5 G	22 3 3 C	28 1 8 C	28 6 8 C	39 9 4 C	30 2 7 T	30 9 2 A	20 8 G	J	N	R	T
Core re-sequenced samples																				
	S9	G		C		20	●		A	●	●	●	●	●	●	●	0	0	1	1
	S9a	G		C		18	●		A	●	●	●	●	●	●	●	0	1	0	1
	S2	G		C		19	●		●	●	●	●	●	●	●	●	15	10	12	37
	S2a	G		C		20	●		●	●	●	●	●	●	●	●	0	2	3	5
	S2b	G		C		18	●		●	●	●	●	●	●	●	●	0	2	1	3
	S2c	G		C		21	●		●	●	●	●	●	●	●	●	1	0	1	2
	S1d	G		●		19	●		●	●	●	●	●	●	●	●	5	0	0	5
	S1	G		●		16	●		●	●	●	●	●	●	●	●	17	19	14	50
	S1a	G		●		18	●		●	●	●	●	●	●	●	●	5	1	0	6
	S1b	G		●		15	●		●	●	●	●	●	●	●	●	2	0	0	2
	S1c	G		●		17	●		●	●	●	●	●	●	●	●	1	0	0	1
	S6	●		●		16	●		●	●	●	●	●	●	●	●	1	2	0	3
	S5	●		●		14	●		●	T	●	A	●	●	●	●	1	4	2	7
	S3	●		●		14	●		●	T	●	A	●	C	G	A	0	3	6	9
	S7	●		●		13	C		●	●	T	●	●	●	●	●	0	2	0	2
	S8	●		●		13	C		●	●	T	●	●	C	G	●	0	1	1	2
	S4	●		●		13	C		●	●	T	●	T	C	G	●	0	1	6	7
	S4a	?		●		14	C		●	●	T	●	T	C	G	●	0	0	1	1

Chimp		Site no.[a]														Sample			
SNP haplotype no.	Sequence haplotype no.	15 5 C	20 1 G	18 7 2 T	12 1 8 G	26 7 1 ?	20 3 8 G	20 8 5 C	21 1 5 G	22 3 3 C	28 1 8 C	28 6 8 C	39 9 4 C	30 2 7 T	30 9 2 A	20 8 G	J	N	R	T
Core re-sequenced samples																			
	S9	G		C		20	●		A	●	●	●	●	●	●	●	0	0	1	1
	S9a	G		C		18	●		A	●	●	●	●	●	●	●	0	1	0	1
	S2	G		C		19	●		●	●	●	●	●	●	●	●	15	10	12	37
	S2a	G		C		20	●		●	●	●	●	●	●	●	●	0	2	3	5
	S2b	G		C		18	●		●	●	●	●	●	●	●	●	0	2	1	3
	S2c	G		C		21	●		●	●	●	●	●	●	●	●	1	0	1	2
	S1d	G		●		19	●		●	●	●	●	●	●	●	●	5	0	0	5
	S1	G		●		16	●		●	●	●	●	●	●	●	●	17	19	14	50
	S1a	G		●		18	●		●	●	●	●	●	●	●	●	5	1	0	6
	S1b	G		●		15	●		●	●	●	●	●	●	●	●	2	0	0	2
	S1c	G		●		17	●		●	●	●	●	●	●	●	●	1	0	0	1
	S6	●		●		16	●		●	●	●	●	●	●	●	●	1	2	0	3
	S5	●		●		14	●		●	T	●	A	●	●	●	●	1	4	2	7
	S3	●		●		14	●		●	T	●	A	●	C	G	A	0	3	6	9
	S7	●		●		13	C		●	●	T	●	●	●	●	●	0	2	0	2
	S8	●		●		13	C		●	●	T	●	●	C	G	●	0	1	1	2
	S4	●		●		13	C		●	●	T	●	T	C	G	●	0	1	6	7
	S4a	?		●		14	C		●	●	T	●	T	C	G	●	0	0	1	1

Commonly Used Descriptors

- Haplotype Frequencies
  - The frequency of each type of chromosome
  - Contain all the information provided by other summary measures
- Commonly used summaries
  - D
  - D'
  - r^2 or Δ^2

Haplotype Frequencies

Linkage equilibrium expected for distant loci

Linkage equilibrium expected for nearby loci

Fill out this table.
X11 is number of times that haplotype is seen.

	2818 C	**2818 T**	
3027 T	X11	X21	# 3027 T alleles
3027 C	X12	x22	#3027 C alleles
	# 2818 C Allele	# 2818 T allele	

Allele Counts

Haplotype frequencies

Disequilibrium Coefficient D_{AB}

D_{AB} is hard to interpret

- Sign is arbitrary …
  - A common convention is to set A, B to be the common allele and a, b to be the rare allele
- Range depends on allele frequencies
  - Hard to compare between markers

D’ – a scaled version of D

More on D'

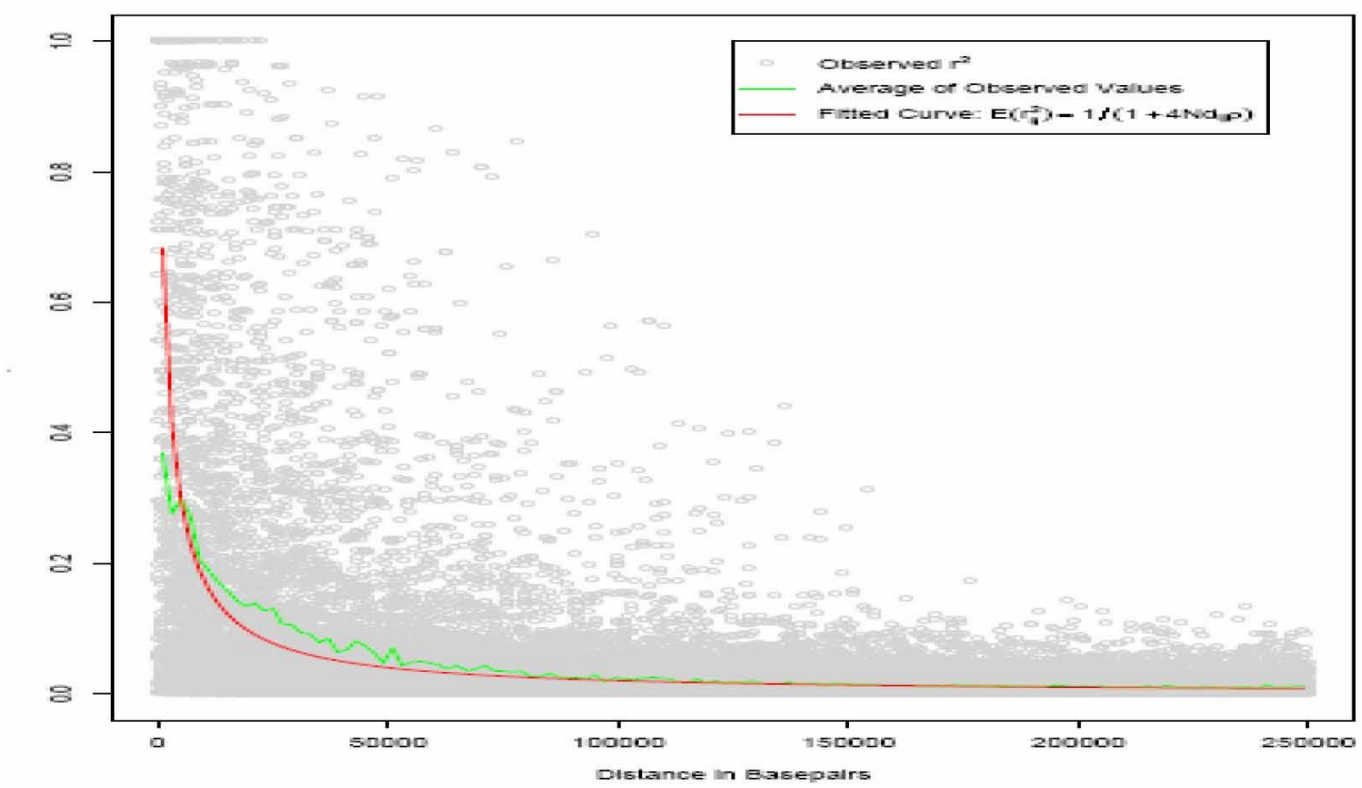
- Pluses:
  - D' = 1 or D' = -1 means no evidence for recombination between the markers
  - If allele frequencies are similar, high D' means the markers are good surrogates for each other
- Minuses:
  - D' estimates inflated in small samples
  - D' estimates inflated when one allele is rare

Correlation coefficient R

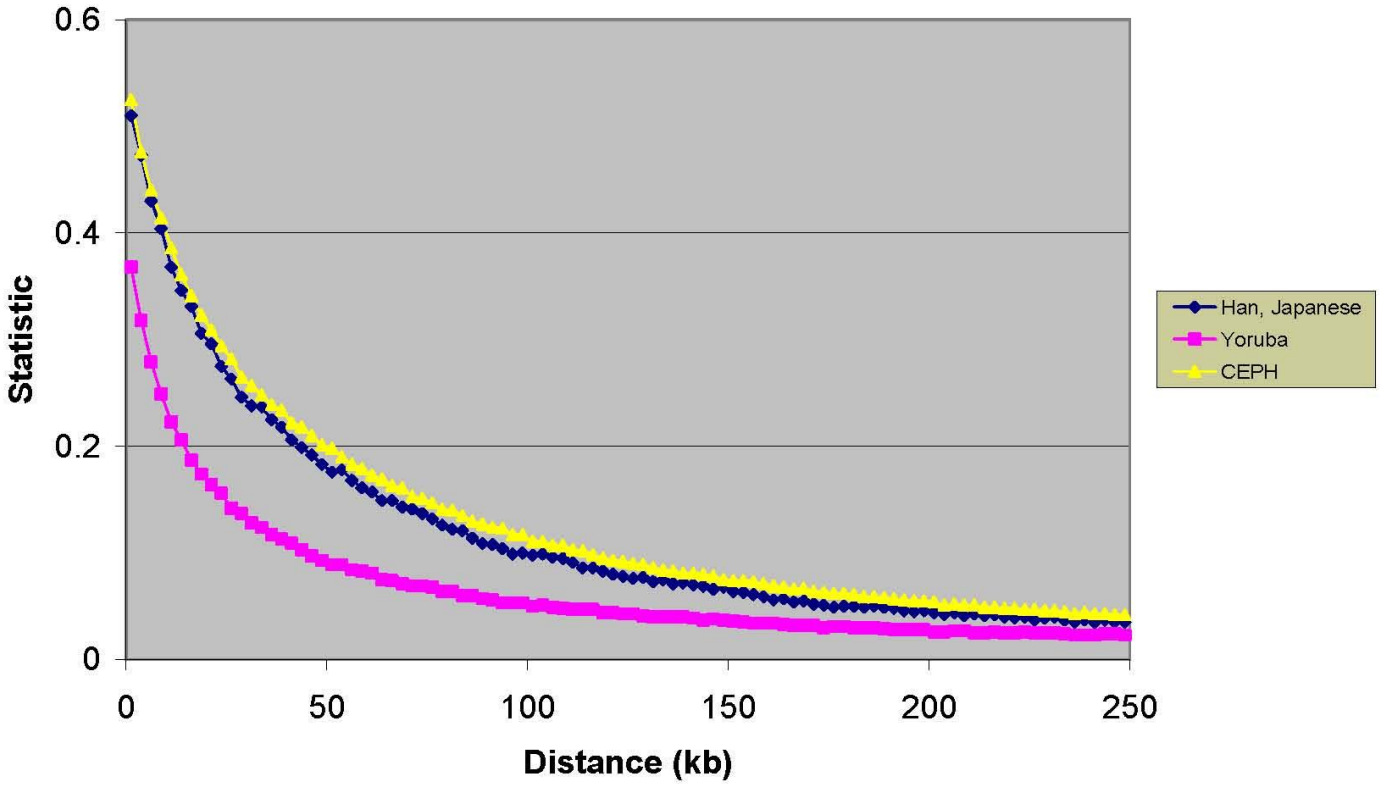
More on r^2

- $r^2 = 1$ implies the markers provide exactly the same information
- The measure preferred by population geneticists
- Measures loss in efficiency when marker A is replaced with marker B in an association study
  - With some simplifying assumptions (e.g. see Pritchard and Przeworski, 2001)

Summarizing Disequilibrium



Comparing Populations …



LD extends further in CEPH and the Han/Japanese than in the Yoruba

Colorectal cancer

1057 cases
960 controls

550K SNPs

Table 1 Risk of colorectal neoplasia associated with the 8q24 SNP rs6983267

Panel	Group	Total	Genotype			Frequency		Allele χ^2	*P* value	Allelic OR (for G against T)	95% c.i.
			GG	GT	TT	G	T				
A	All affected individuals	1,027[a]	352	486	189	0.579	0.421	31.79	1.72×10^{-7}	1.43	1.26–1.63
	Cancers only	620	202	302	116	0.569	0.431	18.96	1.34×10^{-5}	1.38	1.19–1.59
	Adenomas only	407[a]	150	184	73	0.59[illegible]	0.405	25.01	5.70×10^{-7}	1.53	1.29–1.81
	Controls	960	235	471	254	0.490	0.510				
B	Colorectal cancers	4,261	1,334	2,216	811	0.560	0.440	29.71	5.02×10^{-8}	1.19	1.12–1.26

1027 Colorectal cancer

960 controls

Cancer: 0.57G 0.43T

controls: 0.49G 0.51T

Table 1 Risk of colorectal neoplasia associated with the 8q24 SNP rs6983267

			Genotype			Frequency			
Panel	Group	Total	GG	GT	TT	G	T	Allele χ^2	*P* value
A	All affected individuals	1,027[a]	352	486	189	0.579	0.421	31.79	1.72×10^{-7}
	Cancers only	620	202	302	116	0.569	0.431	18.96	1.34×10^{-5}
	Adenomas only	407[a]	150	184	73	0.595	0.405	25.01	5.70×10^{-7}
	Controls	960	235	471	254	0.490	0.510		
B	Colorectal cancers	4,361	1,334	2,216	811	0.560	0.440	29.71	5.02×10^{-8}

Are these different?

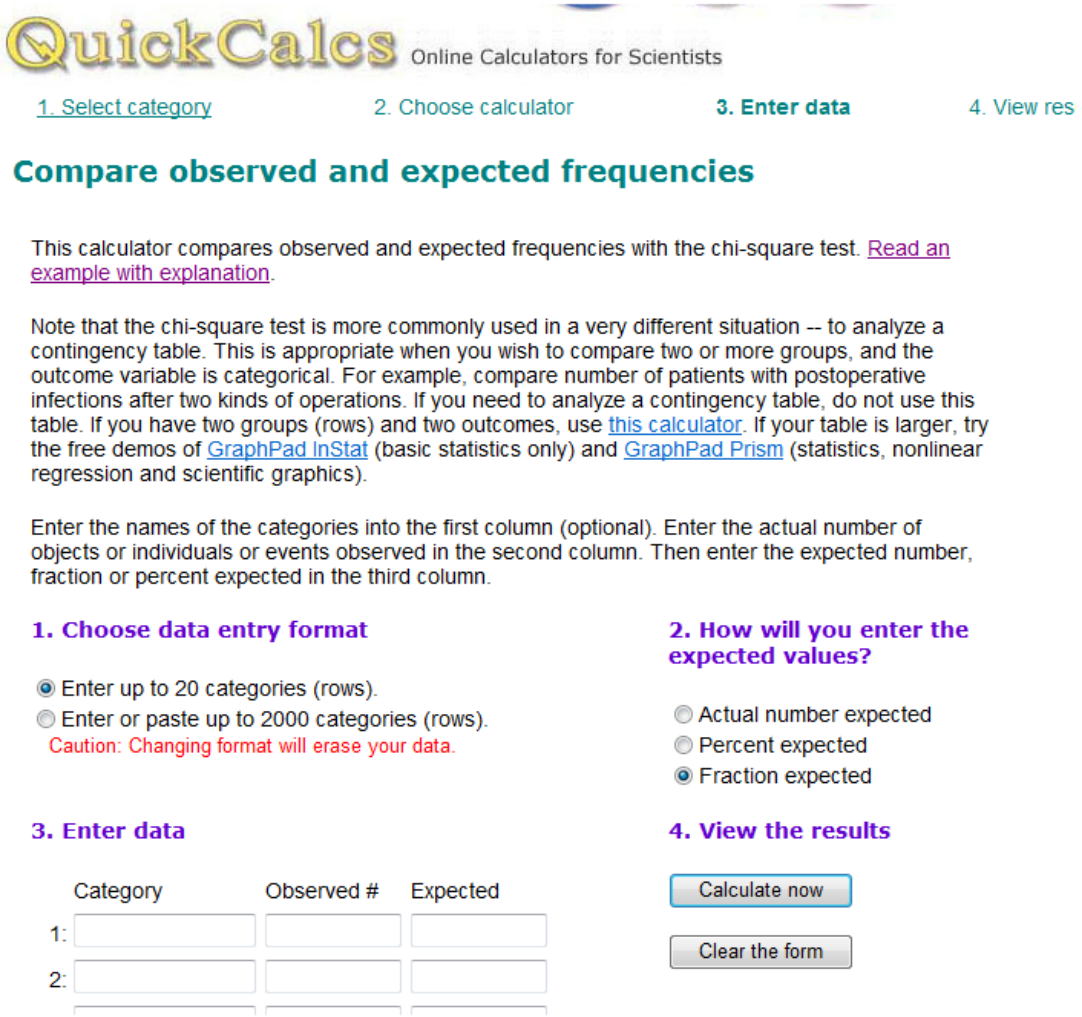
Cancer: 0.57G 0.43T

controls: 0.49G 0.51T

Chi squared

Chi squared

http://www.graphpad.com/quickcalcs/chisquared1.cfm



Chi squared

http://www.graphpad.com/quickcalcs/chisquared1.cfm

Table 1 Risk of colorectal neoplasia associated with the 8q24 SNP rs6983267

			Genotype			Frequency			
Panel	Group	Total	GG	GT	TT	G	T	Allele χ^2	*P* value
A	All affected individuals	1,027[a]	352	486	189	0.579	0.421	31.79	1.72×10^{-7}
	Cancers only	620	202	302	116	0.569	0.431	18.96	1.34×10^{-5}
	Adenomas only	407[a]	150	184	73	0.595	0.405	25.01	5.70×10^{-7}
	Controls	960	235	471	254	0.490	0.510		
B	Colorectal cancers	4,361	1,334	2,216	811	0.560	0.440	29.71	5.02×10^{-8}

Cancer: 352GG + 486GT = 1190 G alleles

486GT + 189 TT = 864 T alleles

Controls: 0.49G

0.51T

Chi squared

http://www.graphpad.com/quickcalcs/chisquared1.cfm

1. Choose data entry format

- (•) Enter up to 20 categories (rows).
- ( ) Enter or paste up to 2000 categories (rows).

Caution: Changing format will erase your data.

2. How will you enter the expected values?

- ( ) Actual number expected
- ( ) Percent expected
- (•) Fraction expected

3. Enter data

	Category	Observed #	Expected
1:	G alleles	1190	.49
2:	T alleles	864	.51
3:			

4. View the results

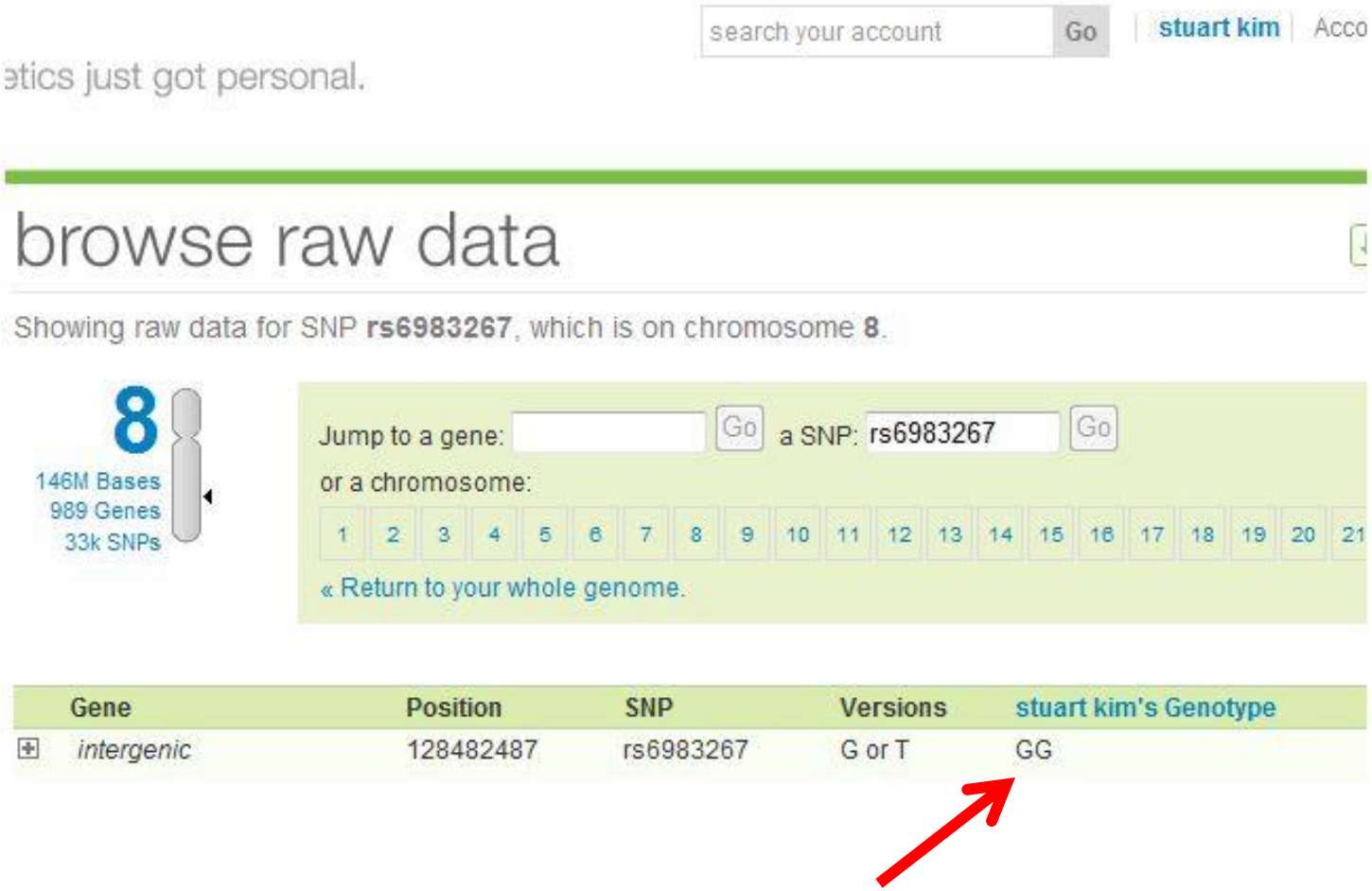
Calculate now

Clear the form

Chi squared = 31

P values = 10^{-7}

Stuart's genotype



Homozygous bad allele ☹

How different is this SNP in the cases versus the controls?

Allelic odds ratio: ratio of the allele ratios in the cases divided by the allele ratios in the controls

Likelihood ratio: Given a genotype, how much more likely are you to show a trait compared to the general population

Multiple hypothesis testing

"Of the 547,647 polymorphic tag SNPs, 27,673 showed an association with disease at $P < .05$."

- $P = .05$ means that there is a 5% chance for this to occur randomly.
- If you try 100 times, you will get about 5 hits.
- If you try 547,647 times, you should expect $547,647 \times .05 = 27,382$ hits.
- So 27,673 (observed) is about the same as one would randomly expect.

Multiple hypothesis testing

"Of the 547,647 polymorphic tag SNPs, 27,673 showed an association with disease at $P < .05$."

- Here, have 547,647 SNPs = # hypotheses
- False discover rate = q = p x # hypotheses. This is called the Bonferroni correction.
- Want q = .05. This means a positive SNP has a .05 likelihood of rising by chance.
- At q = .05, p = .05 / 547,647 = $.91 \times 10^{-7}$
- This is the p value cutoff used in the paper.

Multiple hypothesis testing

"Of the 547,647 polymorphic tag SNPs, 27,673 showed an association with disease at $P < .05$."

- The Bonferroni correction is too conservative. It assumes that all of the tests are independent.
- But the SNPs are linked in haplotype blocks, so there really are less independent hypotheses than SNPs.
- Another way to correct is to permute the data many times, and see how many times a SNP comes up in the permuted data at a particular threshold.

Fill out this table.
Convert all numbers to frequencies.

	2818 C	**2818 T**	
3027 T	X11	X21	# 3027 T alleles
3027 C	X12	x22	#3027 C alleles
	# 2818 C Allele	# 2818 T allele	

Calculate D and D’

	2818 C	2818 T	
3027 T	X11	X21	# 3027 T alleles
3027 C	X12	x22	#3027 C alleles
	# 2818 C Allele	# 2818 T allele	

D = x11 – p1q1

D_{max} is given by the smaller of p_1q_2 and p_2q_1

D’ = D/Dmax

Calculate r^2

	2818 C	2818 T	
3027 T	X11	X21	# 3027 T alleles
3027 C	X12	x22	#3027 C alleles
	# 2818 C Allele	# 2818 T allele	

$r^2 = D^2/p1p2q1q2$

Haplotype Frequencies

		Locus B		Totals
		B	b	
Locus A	A	p_{AB}	p_{Ab}	p_A
	a	p_{aB}	p_{ab}	p_a
Totals		p_B	p_b	1.0

Linkage Equilibrium Expected for Distant Loci

$$p_{AB} = p_A p_B$$

$$p_{Ab} = p_A p_b = p_A(1 - p_B)$$

$$p_{aB} = p_a p_B = (1 - p_A) p_B$$

$$p_{ab} = p_a p_b = (1 - p_A)(1 - p_B)$$

Linkage Disequilibrium Expected for Nearby Loci

$$p_{AB} \neq p_A p_B$$

$$p_{Ab} \neq p_A p_b = p_A(1 - p_B)$$

$$p_{aB} \neq p_a p_B = (1 - p_A) p_B$$

$$p_{ab} \neq p_a p_b = (1 - p_A)(1 - p_B)$$

Disequilibrium Coefficient D_{AB}

$$D_{AB} = p_{AB} - p_A p_B$$

$$p_{AB} = p_A p_B + D_{AB}$$

$$p_{Ab} = p_A p_b - D_{AB}$$

$$p_{aB} = p_a p_B - D_{AB}$$

$$p_{ab} = p_a p_b + D_{AB}$$

D' – A scaled version of D

$$D'_{AB} = \begin{cases} \dfrac{D_{AB}}{\min(p_A p_B, p_a p_b)} & D_{AB} < 0 \\ \dfrac{D_{AB}}{\min(p_A p_b, p_a p_B)} & D_{AB} > 0 \end{cases}$$

- Ranges between –1 and +1
  - More likely to take extreme values when allele frequencies are small
  - ±1 implies at least one of the observed haplotypes was not observed

Δ^2 (also called r^2)

$$\Delta^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)}$$

$$= \frac{\chi^2}{2n}$$

- Ranges between 0 and 1
  - 1 when the two markers provide identical information
  - 0 when they are in perfect equilibrium
- Expected value is $1/2n$

Cancer: 0.57G 0.43T

Cancer: G:T ratio = 0.57/.43 = 1.32

controls: 0.49G 0.51T

controls: G:T ratio = .49/.51 = .96

Allelic odds ratio = 1.32/.96 = 1.37