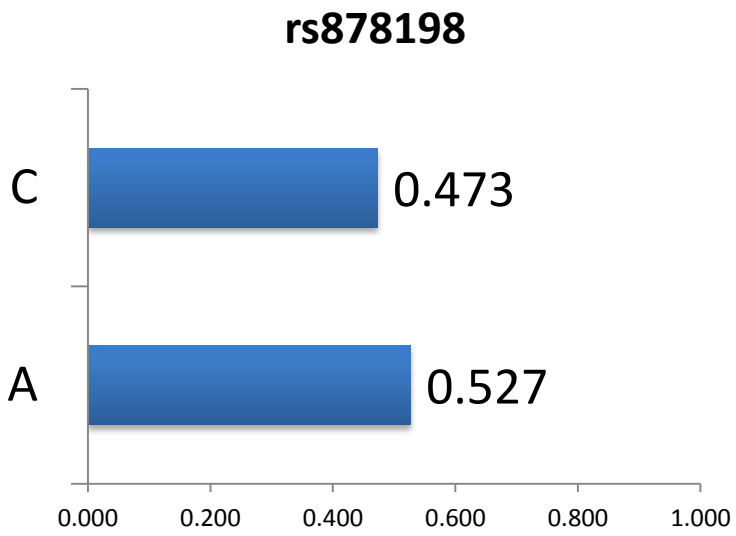
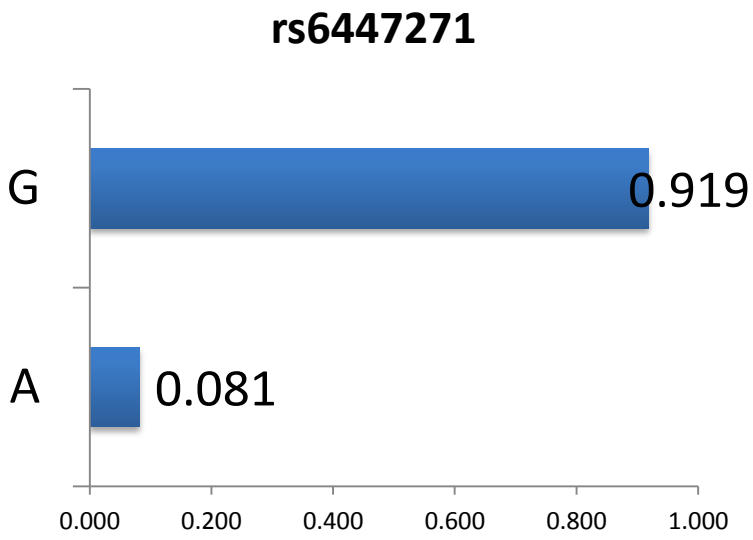
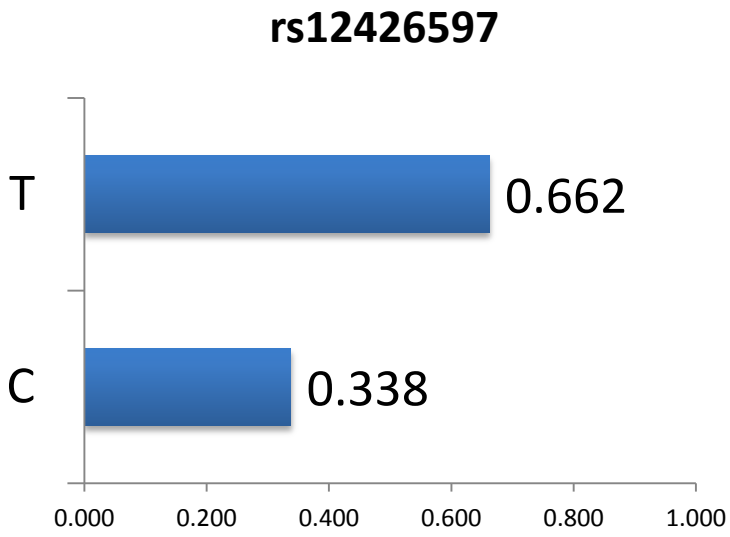
Linkage

Linkage, part 1 – Allele Frequencies





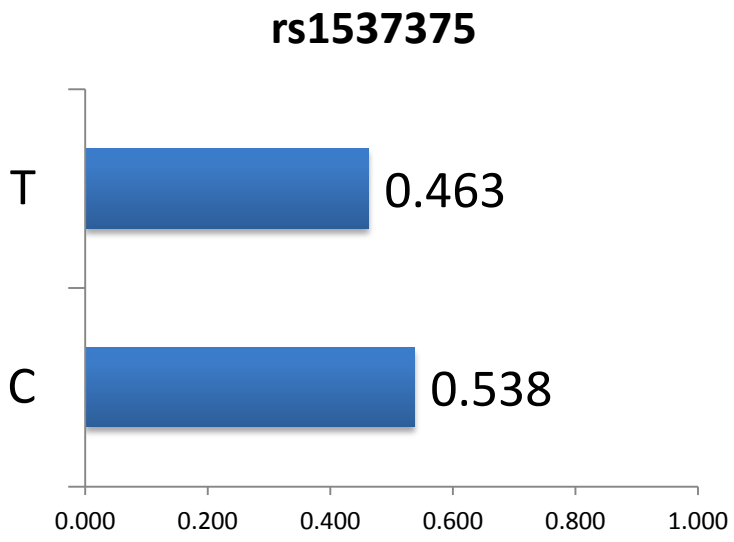


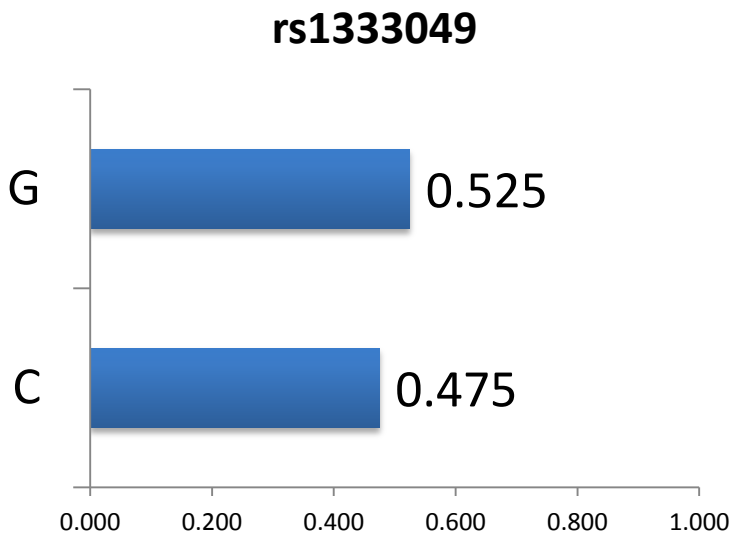
Linkage, part 1 – Haplotypes

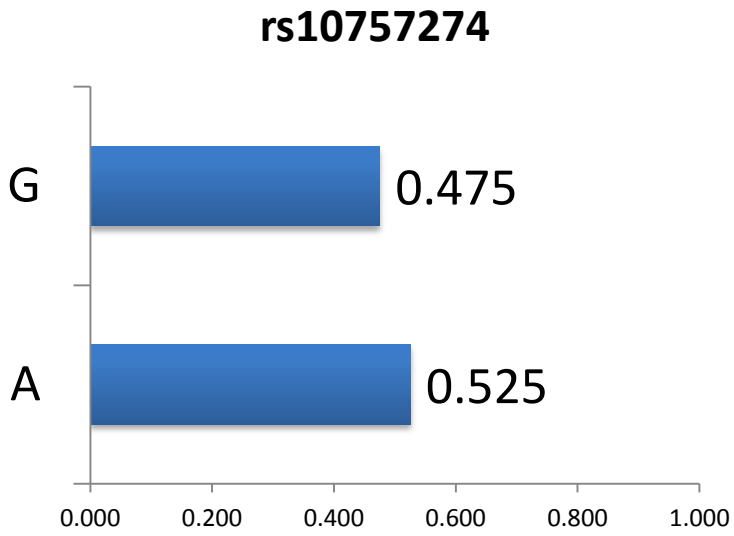
SNP order: rs878198, rs6447271, rs12426597

Haplotypes	Expected	Obs Freq
AA,AA,CC	0.0002	0
AA,AA,CT	0.0008	0
AA,AA,TT	0.0008	0
AA,AG,CC	0.0047	0
AA,AG,CT	0.0185	0
AA,AG,TT	0.0181	0
AA,GG,CC	0.0268	0
AA,GG,CT	**0.1050**	**0.0270**
AA,GG,TT	**0.1028**	**0.1622**
AC,AA,CC	0.0004	0
AC,AA,CT	0.0015	0
AC,AA,TT	0.0014	0
AC,AG,CC	0.0085	0
AC,AG,CT	**0.0332**	**0.0811**
AC,AG,TT	**0.0325**	**0.0270**
AC,GG,CC	**0.0481**	**0.0811**
AC,GG,CT	**0.1884**	**0.3514**
AC,GG,TT	**0.1845**	**0.1351**
AC,AA,CC	0.0002	0
CC,AA,CT	0.0007	0
CC,AA,TT	0.0006	0
CC,AG,CC	0.0038	0
CC,AG,CT	0.0149	0
CC,AG,TT	**0.0146**	**0.0541**
CC,GG,CC	**0.0216**	**0.0270**
CC,GG,CT	0.0846	0
CC,GG,TT	**0.0828**	**0.0541**

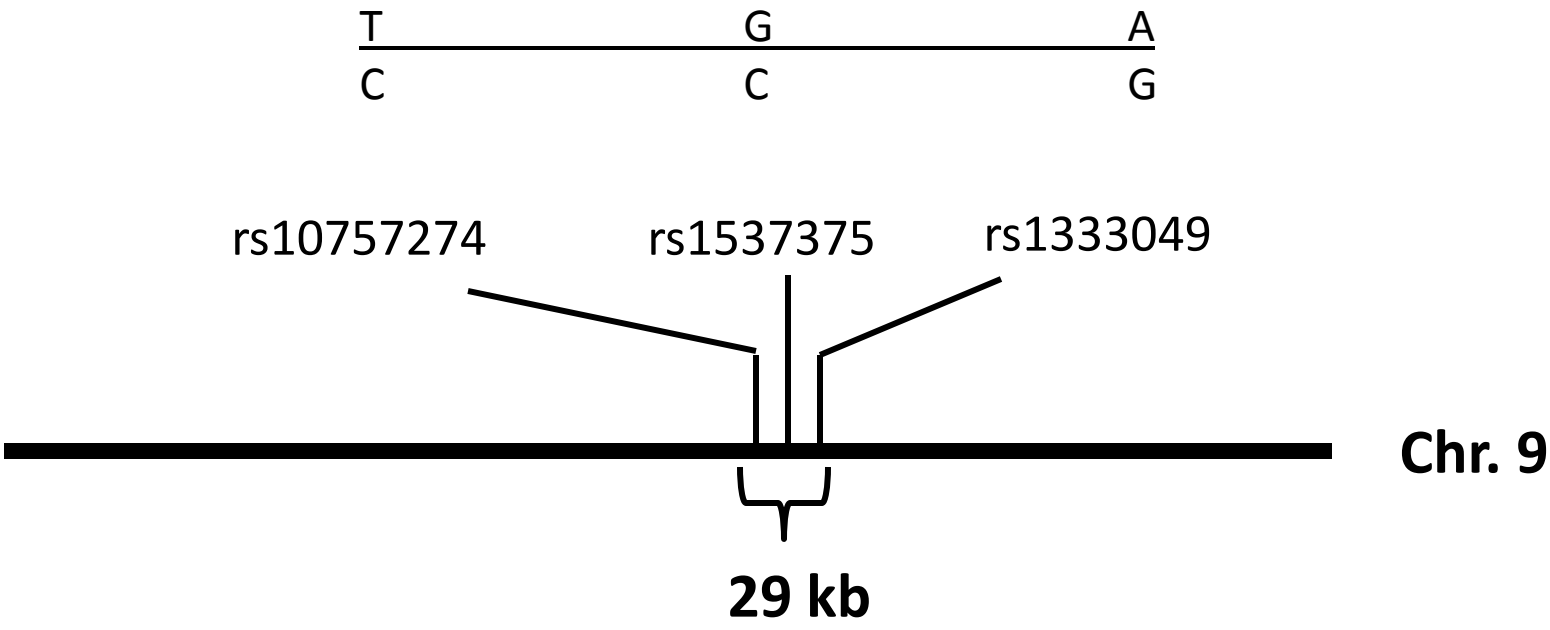
Linkage, part 2 – Allele Frequencies







Genetic Linkage 2



Linkage, part 2 – Haplotypes

SNP order: rs1537375 rs1333049, rs10757274

Haplotypes	Expected	Obs Freq
CC,CC,AA	0.0180	0
CC,CC,AG	**0.0326**	**0.025**
CC,CC,GG	**0.0147**	**0.225**
CC,CG,AA	0.0398	0
CC,CG,AG	**0.0720**	**0.025**
CC,CG,GG	**0.0326**	**0.025**
CC,GG,AA	0.0220	0
CC,GG,AG	0.0398	0
CC,GG,GG	0.0180	0
CT,CC,AA	0.0310	0
CT,CC,AG	**0.0561**	**0.025**
CT,CC,GG	0.0254	0
CT,CG,AA	0.0685	0
CT,CG,AG	**0.1239**	**0.350**
CT,CG,GG	0.0561	0
CT,GG,AA	**0.0378**	**0.075**
CT,GG,AG	**0.0685**	**0.025**
CT,GG,GG	0.0310	0
CT,CC,AA	0.0133	0
TT,CC,AG	0.0241	0
TT,CC,GG	0.0109	0
TT,CG,AA	0.0295	0
TT,CG,AG	0.0533	0
TT,CG,GG	0.0241	0
TT,GG,AA	**0.0163**	**0.225**
TT,GG,AG	0.0295	0
TT,GG,GG	0.0133	0

CCG / CCG →

TGA / CCG →

TGA / TGA →

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 T	1 2 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

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

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	

	2818 C	2818 T	
3027 T	125/146	2/146	127/146 T alleles
3027 C	9/146	10/146	19/146 C alleles
	134/146 C Allele	12/146 T allele	

Convert to fractions

	2818 C	**2818 T**	
3027 T	.86	.013	.87 T alleles
3027 C	.061	.068	.13 C alleles
	.92 C Allele	.08 T allele	

Disequilibrium Coefficient D_{AB}

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

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

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

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

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

Calculate D_{AB}

$$D_{AB} = P_{AB} - P_A P_B$$
$$= .86 - (.87)(.92)$$
$$= .86 - .80$$
$$= .06$$

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

R – correlation coefficient

$$R = \frac{D_{AB}}{SQR(P_A \times P_a \times P_B \times P_b)}$$

Calculate R

$R = D_{AB} / SQR(P_A P_a P_B P_b)$

$= .06 / SQR(.87 * .13 * .92 * .08)$

$= .06 / SQR(7.2 \times 10^{-3})$

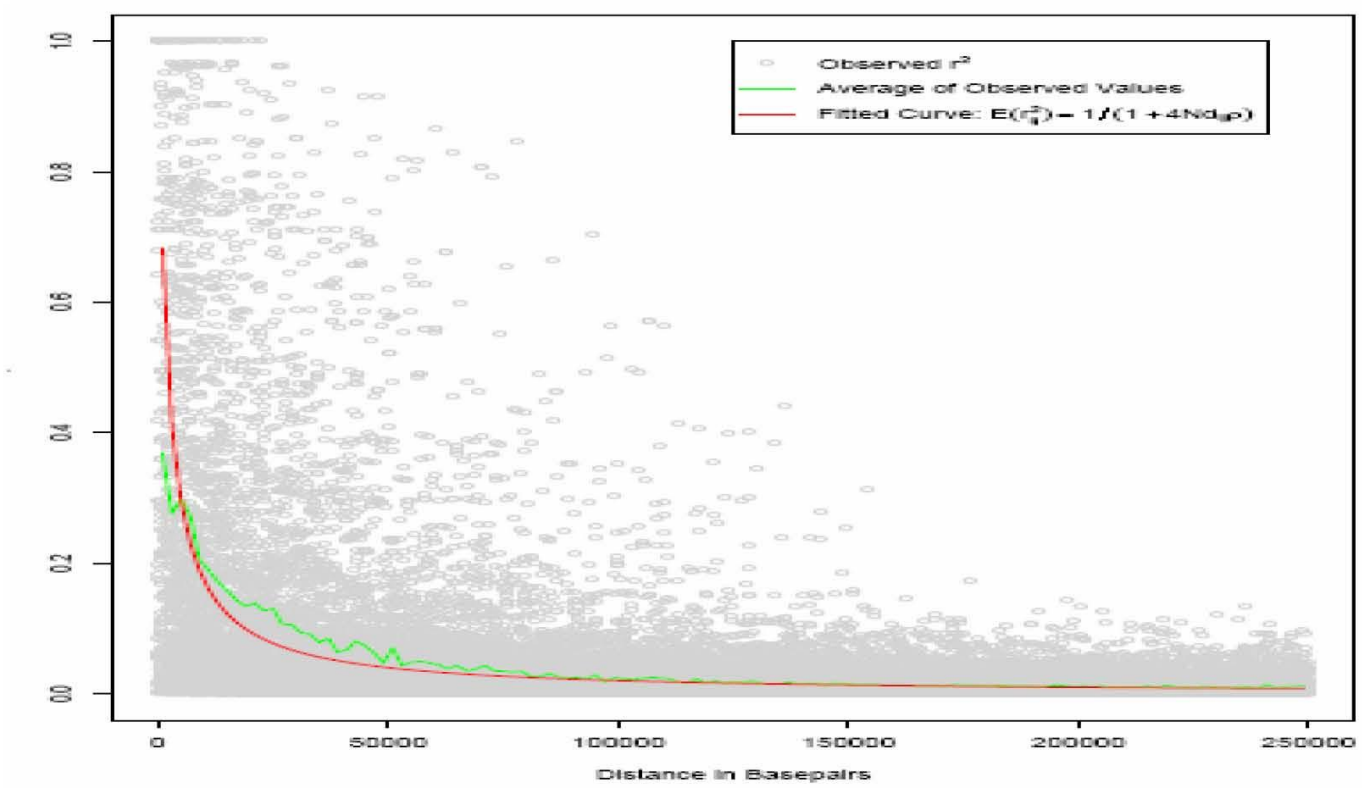
$= .06 / .085 = .706$

$R^2 = .497$

More on r^2

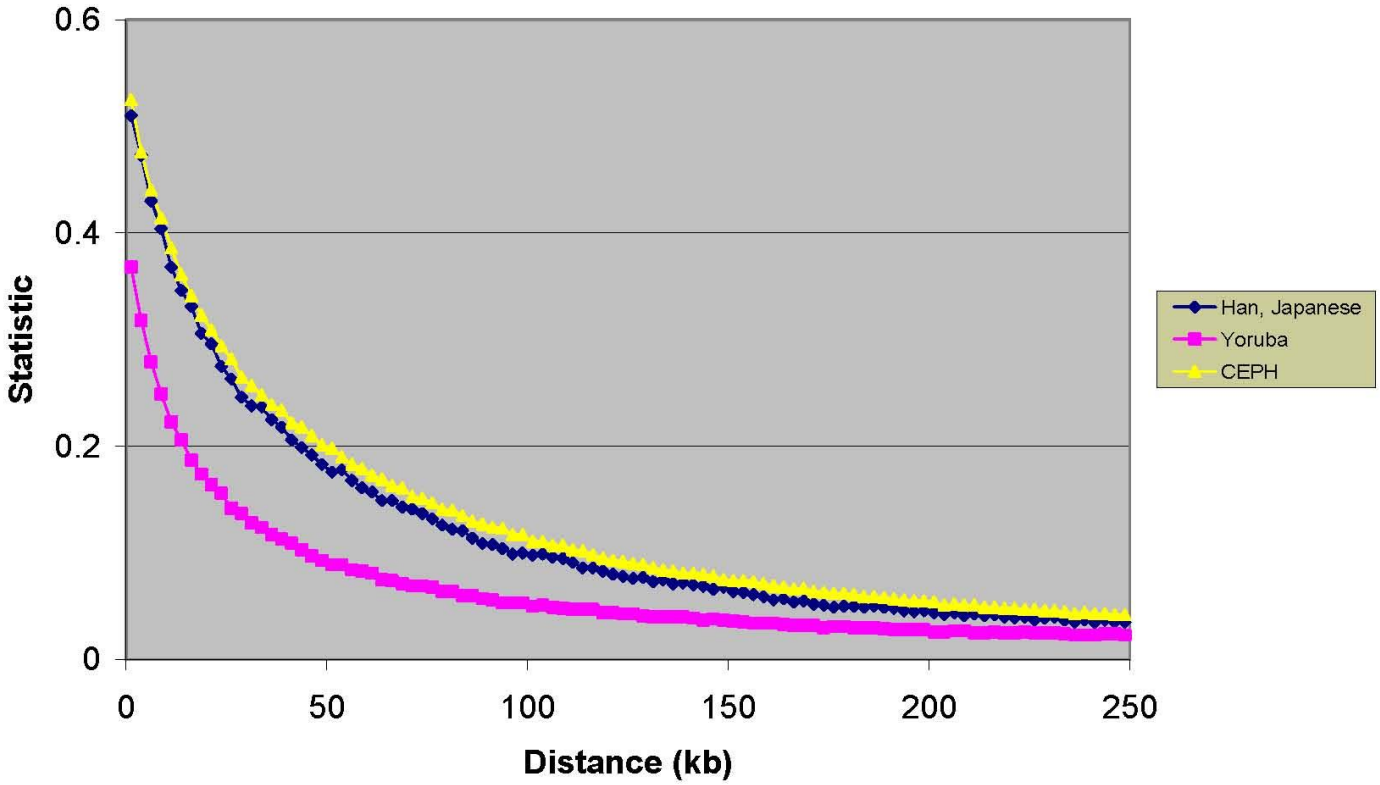
- $r^2 = 1$ implies the markers provide exactly the same information
- The measure preferred by population geneticists
- Measures loss in efficiency when marker A is replaced with marker B in an association study
  - With some simplifying assumptions (e.g. see Pritchard and Przeworski, 2001)

Summarizing Disequilibrium



slide created by Goncarlo Abecasis

Comparing Populations …

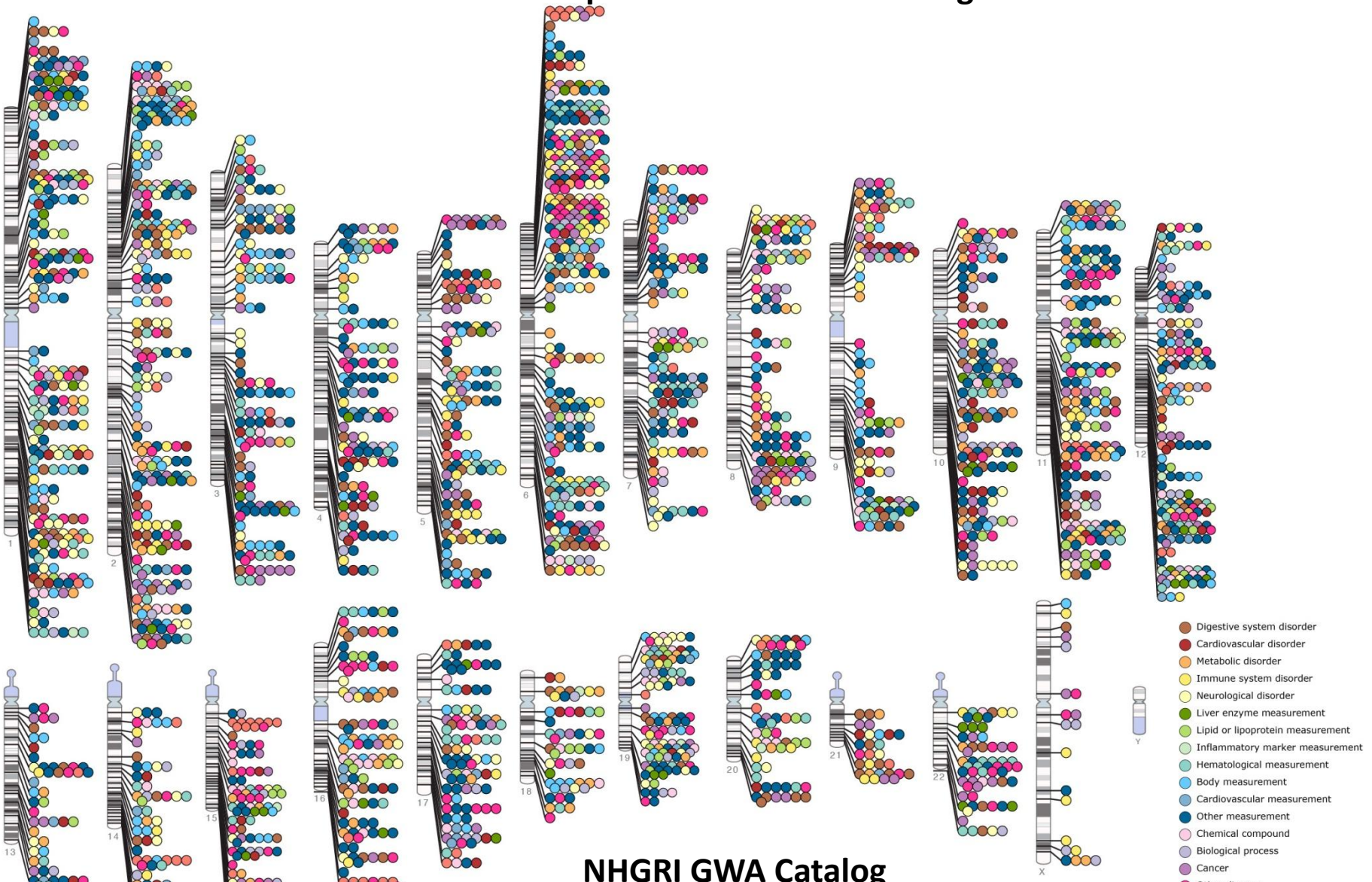


LD extends further in CEPH and the Han/Japanese than in the Yoruba

GWAS readings

**How to Interpret a Genome-wide Association Study
http://stanford.edu/class/gene210/files/readings/Pearson_JAMA_2008.pdf
Finding the missing heritability of complex diseases
http://stanford.edu/class/gene210/files/readings/Manolio_Nature_2009.pdf
How to Use an Article About Genetic Association: A:
http://stanford.edu/class/gene210/files/readings/GWAS1_JAMA_2009.pdf
How to Use an Article About Genetic Association: B:
http://stanford.edu/class/gene210/files/readings/GWAS2_JAMA_2009.pdf
How to Use an Article About Genetic Association: C
http://stanford.edu/class/gene210/files/readings/GWAS3_JAMA_2009.pdf

Published Genome-Wide Associations through 07/2012
Published GWA at $p \leq 5X10^{-8}$ for 18 trait categories



NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

nature International weekly journal of science

ARTICLE PREVIEW

NATURE | ARTICLE

日本語要約

Seventy-five genetic loci influencing the human red blood cell

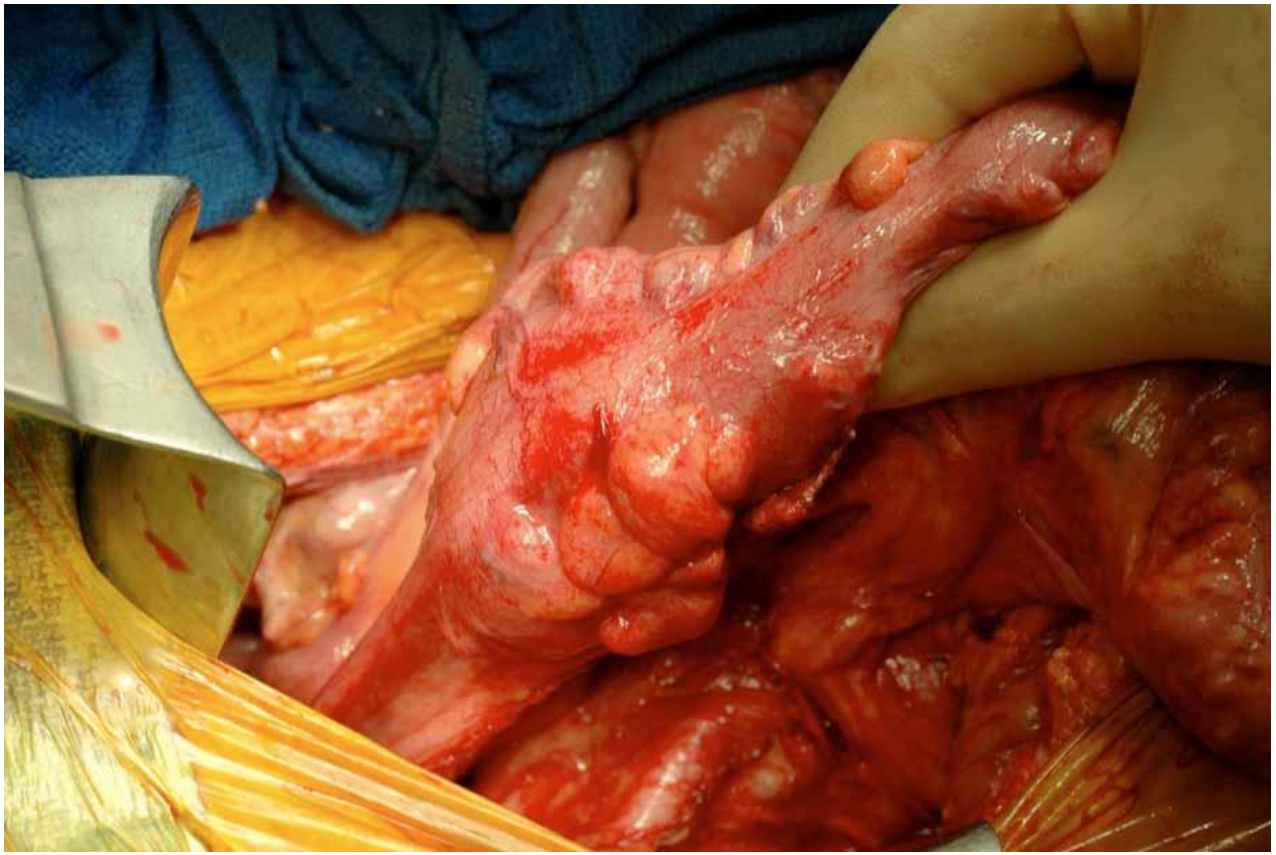
Pim van der Harst, Weihua Zhang, Irene Mateo Leach, Augusto Rendon, Niek Verweij, Joban Sehmi, Dirk S. Paul, Ulrich Elling, Hooman Allayee, Xinzhong Li, Aparna Radhakrishnan, Sian-Tsung Tan, Katrin Voss, Christian X. Weichenberger, Cornelis A. Albers, Abtehale Al-Hussani, Folkert W. Asselbergs, Marina Ciullo, Fabrice Danjou, Christian Dina, Tõnu Esko, David M. Evans, Lude Franke, Martin Gögele, Jaana Hartiala + *et al.*

Affiliations | Contributions | Corresponding authors

Nature **492**, 369–375 (20 December 2012) | doi:10.1038/nature11677

Received 06 February 2012 | Accepted 15 October 2012 | Published online 05 December 2012 | Corrected online **19 December 2012**

Colorectal cancer



1057 cases
960 controls

550K SNPs

Colorectal cancer data from rs6983267

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

			Genotype			Frequency	
Panel	Group	Total	GG	GT	TT	G	T
A	All affected individuals	1,027[a]	352	486	189	0.579	0.421
	Cancers only	620	202	302	116	0.[illegible]	0.431
	Adenomas only	407[a]	150	184	73	0.595	0.405
	Controls	960	235	471	254	0.490	0.510
B	Colorectal cancer	1,261	1,334	2,216	811	[illegible]	0.440

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
A	All affected individuals	1,027[a]	352	486	189	0.579	0.421
	Cancers only	620	202	302	116	0.569	0.431
	Adenomas only	407[a]	150	184	73	0.595	0.405
	Controls	960	235	471	254	0.490	0.510
B	Colorectal cancers	4,361	1,334	2,216	811	0.560	0.440

Are these different?

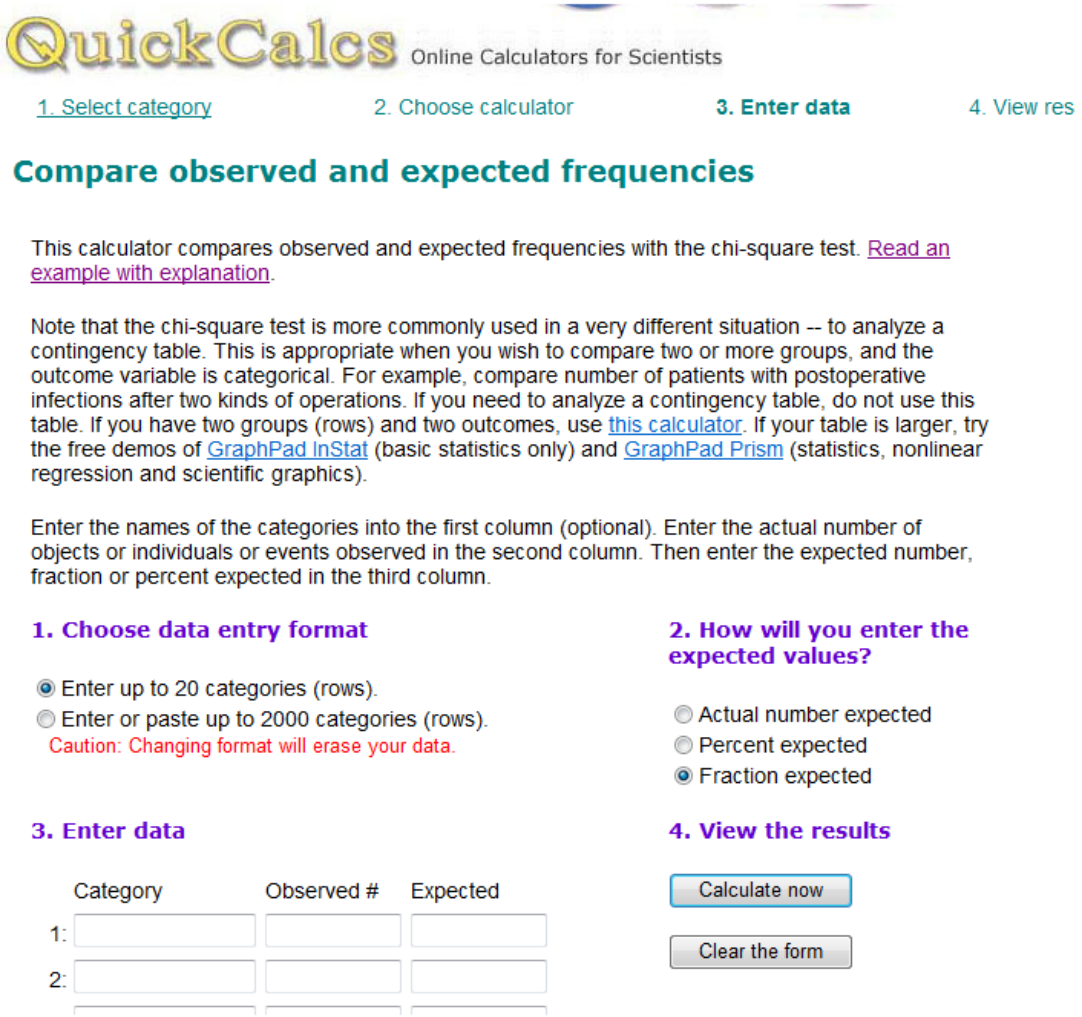
Cancer: 0.57G 0.43T

controls: 0.49G 0.51T

Chi squared

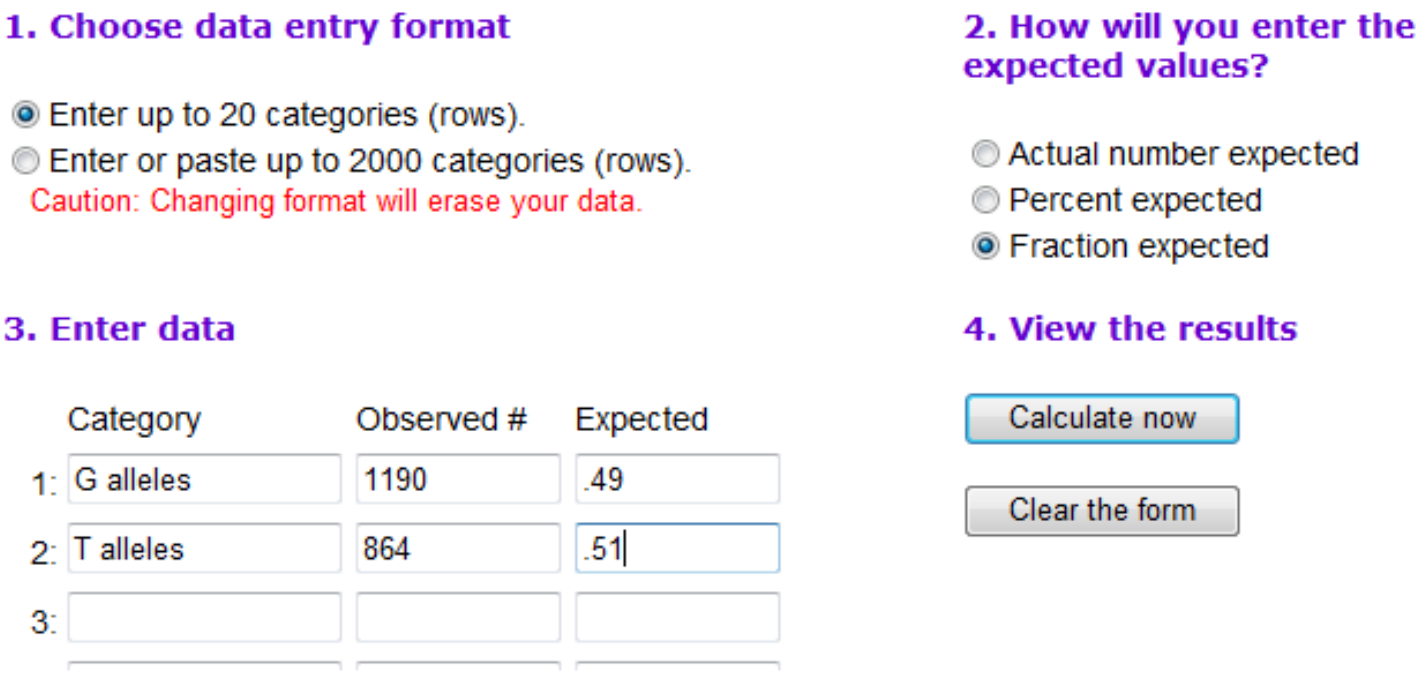
Chi squared

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



Chi squared

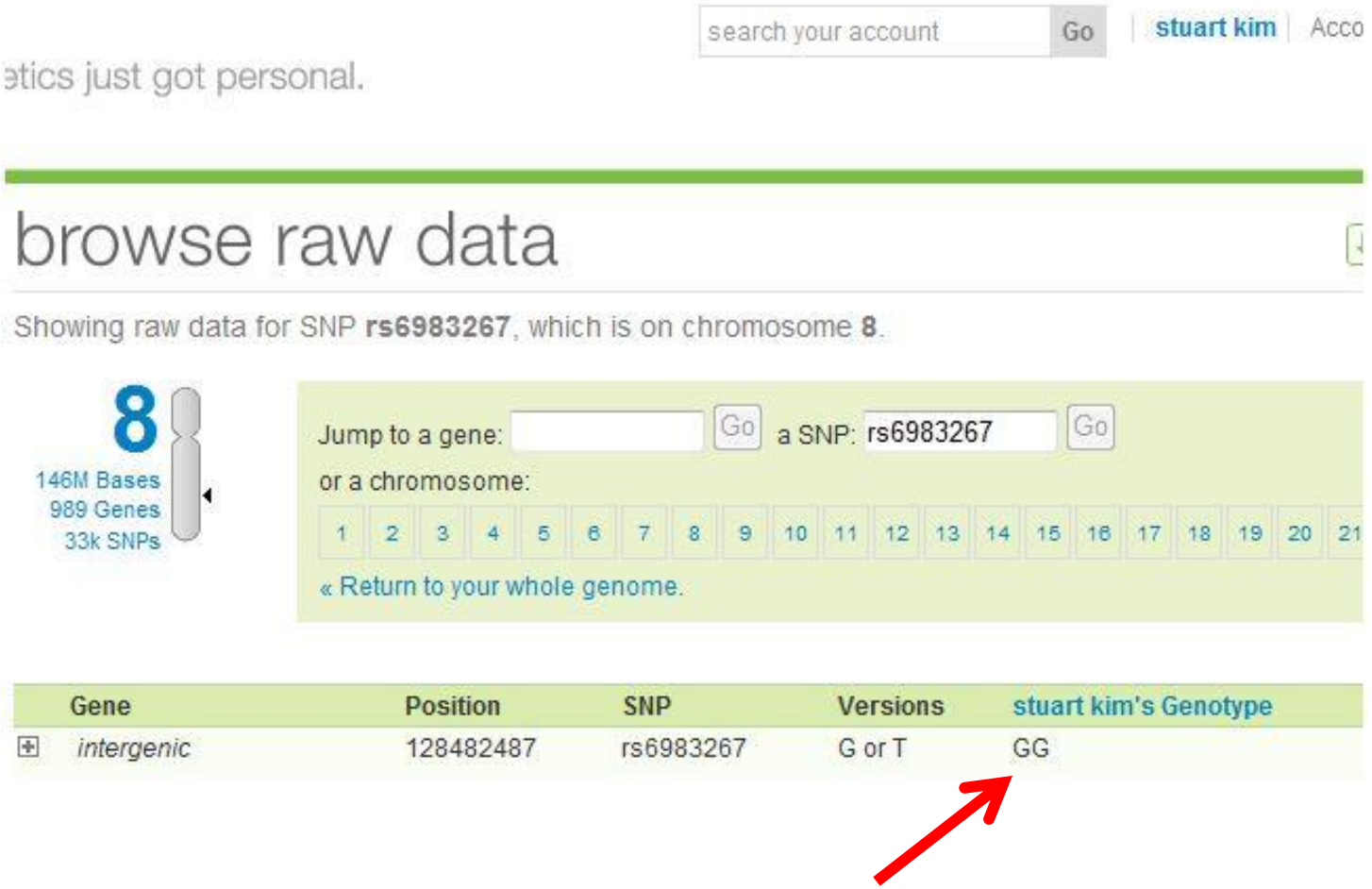
http://www.graphpad.com/quickcalcs/chisquared1.cfm



Chi squared = 31

P values = 10^{-7}

Stuart's genotype



Homozygous bad allele ☹

Other models

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

			Genotype		
Panel	Group	Total	GG	GT	TT
A	All affected individuals	1,027[a]	352	486	189
	Cancers only	620	202	302	116
	Adenomas only	407[a]	150	184	73
	Controls	960	235	471	254
B	Colorectal cancers	4,361	1,334	2,216	811

Dominant: Assume G is dominant.
GG or GT vs TT

	GG or GT	TT
Cases	838	189
Controls	706	254

Other models

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

			Genotype		
Panel	Group	Total	GG	GT	TT
A	All affected individuals	1,027[a]	352	486	189
	Cancers only	620	202	302	116
	Adenomas only	407[a]	150	184	73
	Controls	960	235	471	254
B	Colorectal cancers	4,361	1,334	2,216	811

Recessive: Assume G is recessive.
GG vs GT or TT

	GG	GT or TT
Cases	352	675
Controls	235	725

Other models

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

			Genotype		
Panel	Group	Total	GG	GT	TT
A	All affected individuals	1,027[a]	352	486	189
	Cancers only	620	202	302	116
	Adenomas only	407[a]	150	184	73
	Controls	960	235	471	254
B	Colorectal cancers	4,361	1,334	2,216	811

additive: GG > GT > TT

Do linear regression 3 genotype x 2 groups

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

Cancer .57 G/.43 T = 1.32

Control .49 G/ .51T = 0.96

Allelic Odds Ratio = 1.32/0.96 = 1.37

How different is this SNP in the cases versus everyone?

Allelic odds ratio*: ratio of the allele ratios in the cases divided by the allele ratio in the entire population

(need allele ratio from entire population to do this)

Likelihood ratio: What is the likelihood of seeing a genotype given the disease compared to the likelihood of seeing the genotype given no disease?

(need data from entire population to do this. We can do this in the class GWAS. For cancer vs controls, the two groups were separate and so we do not know the genotype frequencies of the population as a whole.)

Increased Risk: What is the likelihood of seeing a trait given a genotype compared to overall likelihood of seeing the trait in the population?

(need data from entire population to do this. We can do this in the class GWAS. For cancer vs controls, the two groups were separate and so we do not know the genotype frequencies of the population as a whole.)

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.

SNPedia

The SNPedia website
http://www.snpedia.com/index.php/SNPedia

A thank you from SNPedia
http://snpedia.blogspot.com/2012/12/o-come-all-ye-faithful.html

Class website for SNPedia
http://stanford.edu/class/gene210/web/html/projects.

List of last years write-ups
http://stanford.edu/class/gene210/archive/2012/projects_2012.html

How to write up a SNPedia entry
http://stanford.edu/class/gene210/web/html/snpedia.html

SNPedia

Summarize the trait
Summarize the study
- How large was the cohort?
- How strong was the p-value?
- What was the OR, likelihood ratio or increased risk?

Which population?
What is known about the SNP?
- Associated genes?
- Protein coding?
- Allele frequency?

Does knowledge of the SNP affect diagnosis or treatment?

Class GWAS

Go to genotation.stanford.edu
Go to “traits”, then “GWAS”
Look up your SNPs
Fill out the table
Submit information

Class GWAS

Class should split into ~10 groups
Each group should calculate association of all 5 SNPs with one trait.
Do a chi-squared test and write down the p value.
We will make a table from the class results

Class GWAS

Which SNP is best associated with which trait?
Is this significant? What should be the significance cut off?
Are there other significant associations?
What might cause the other significant associations?

Class GWAS

Split into groups of 4-5
Each group should select on trait/SNP
Calculate:

- allelic odds ratio
- likelihood ratio