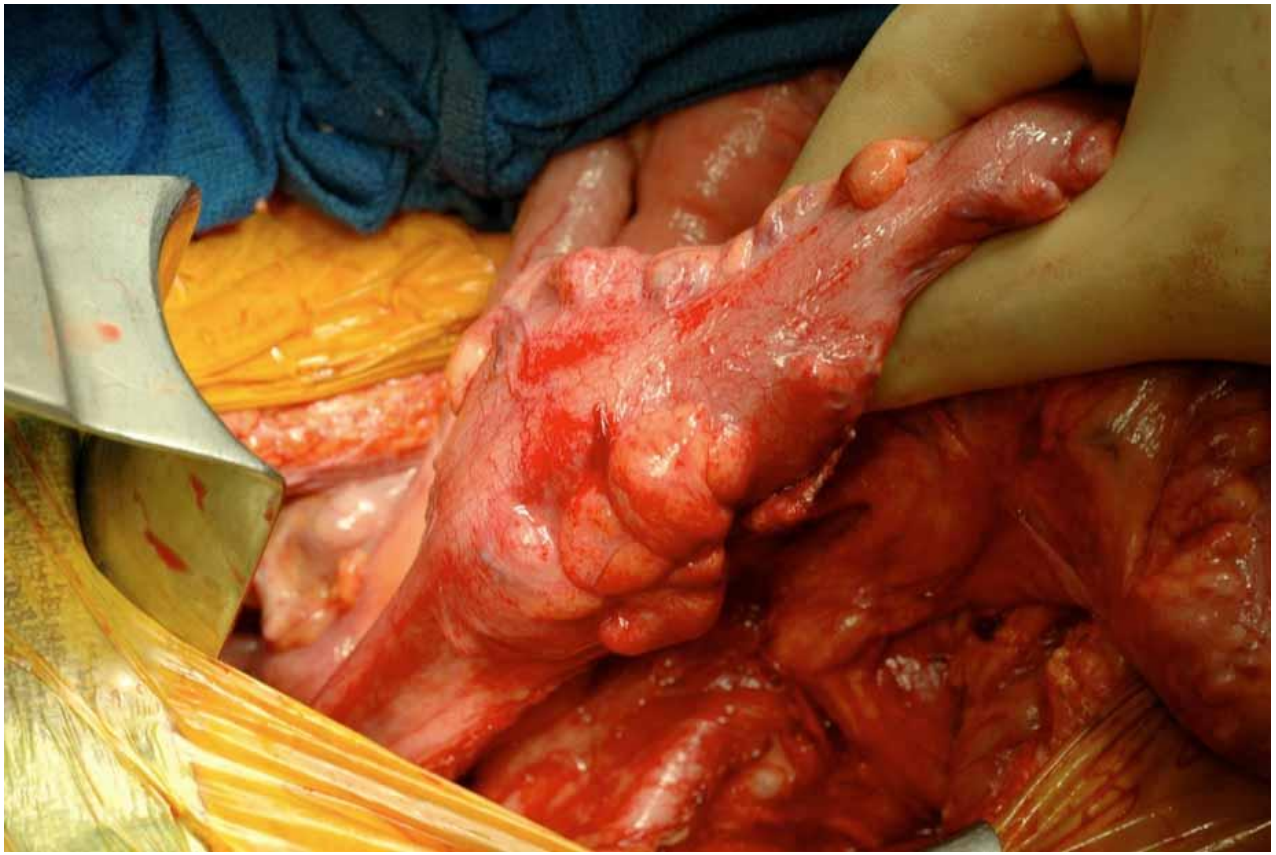
# Colorectal cancer



1057 cases
960 controls

550K SNPs

APRIL 10

# 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?

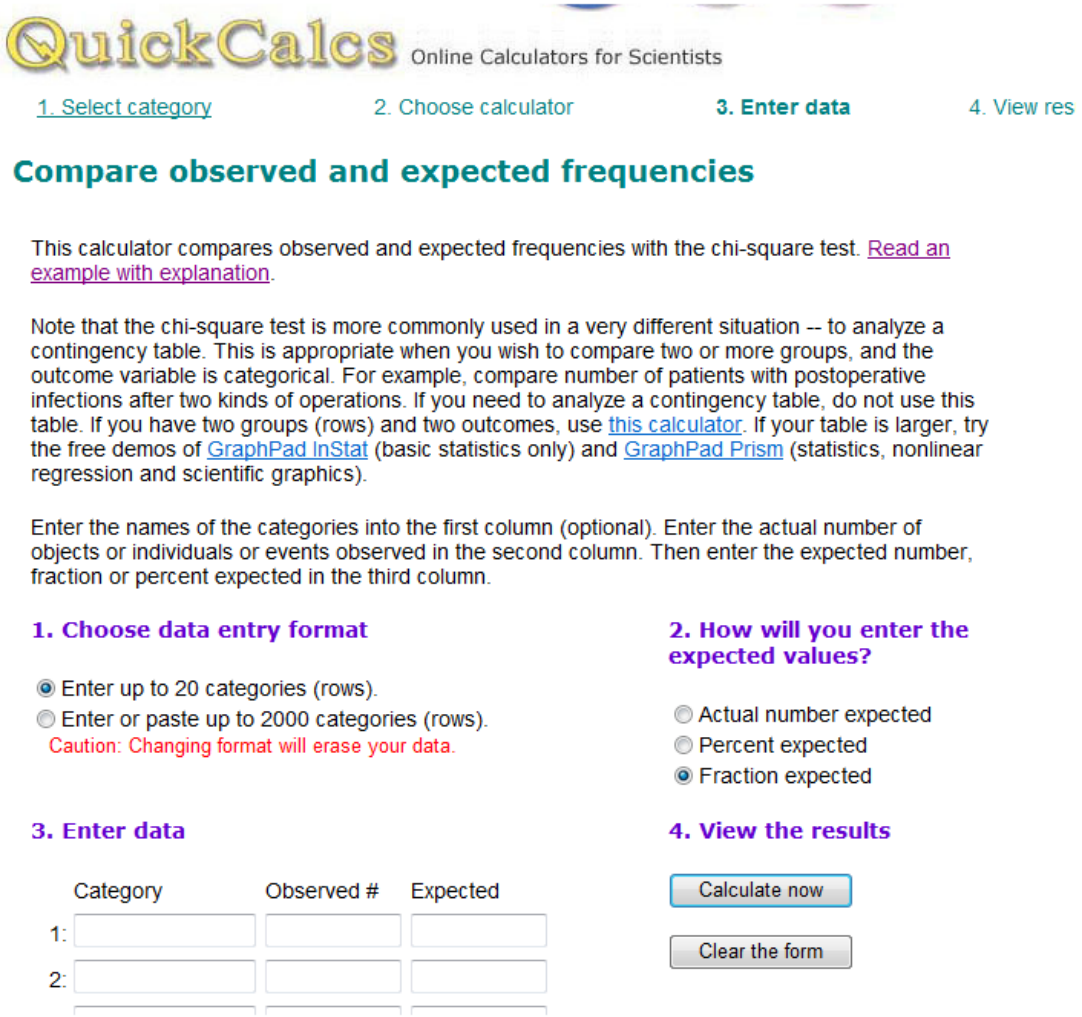
Cancer: 0.57G 0.43T

controls: 0.49G 0.51T

Chi squared

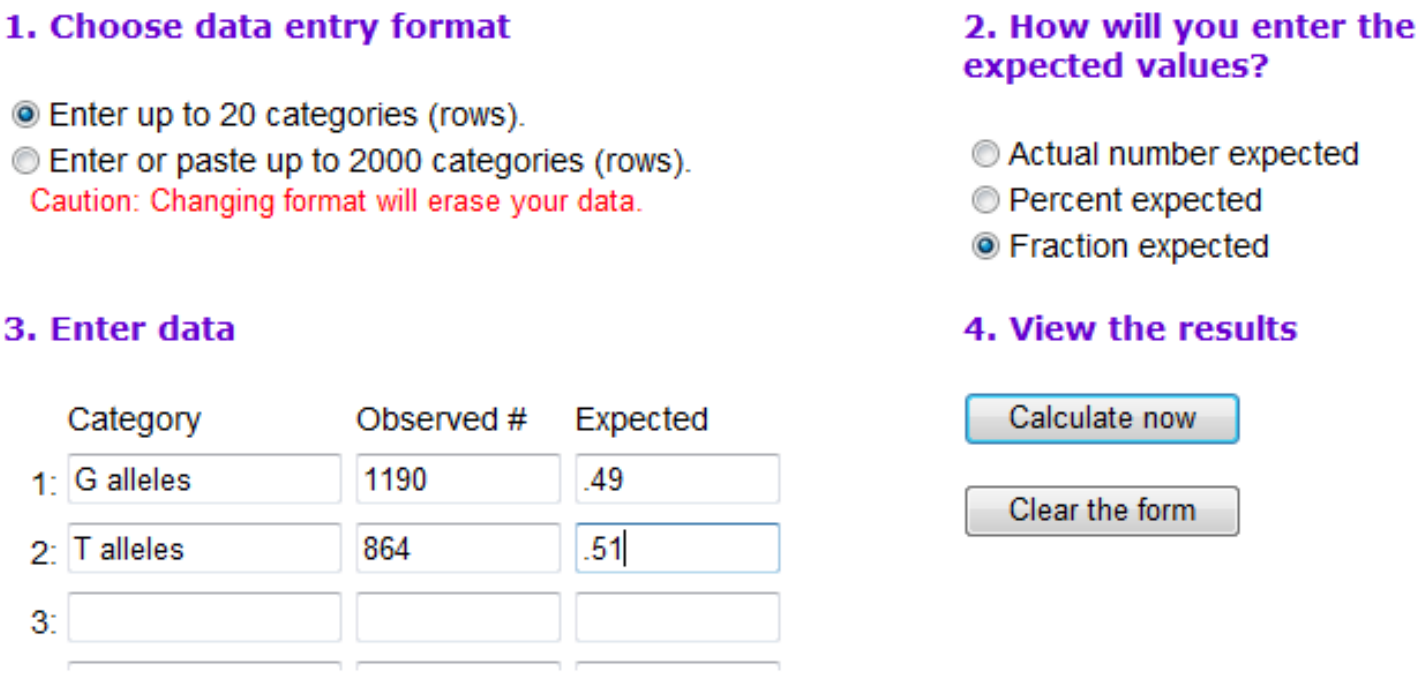
# Chi squared

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



# Chi squared

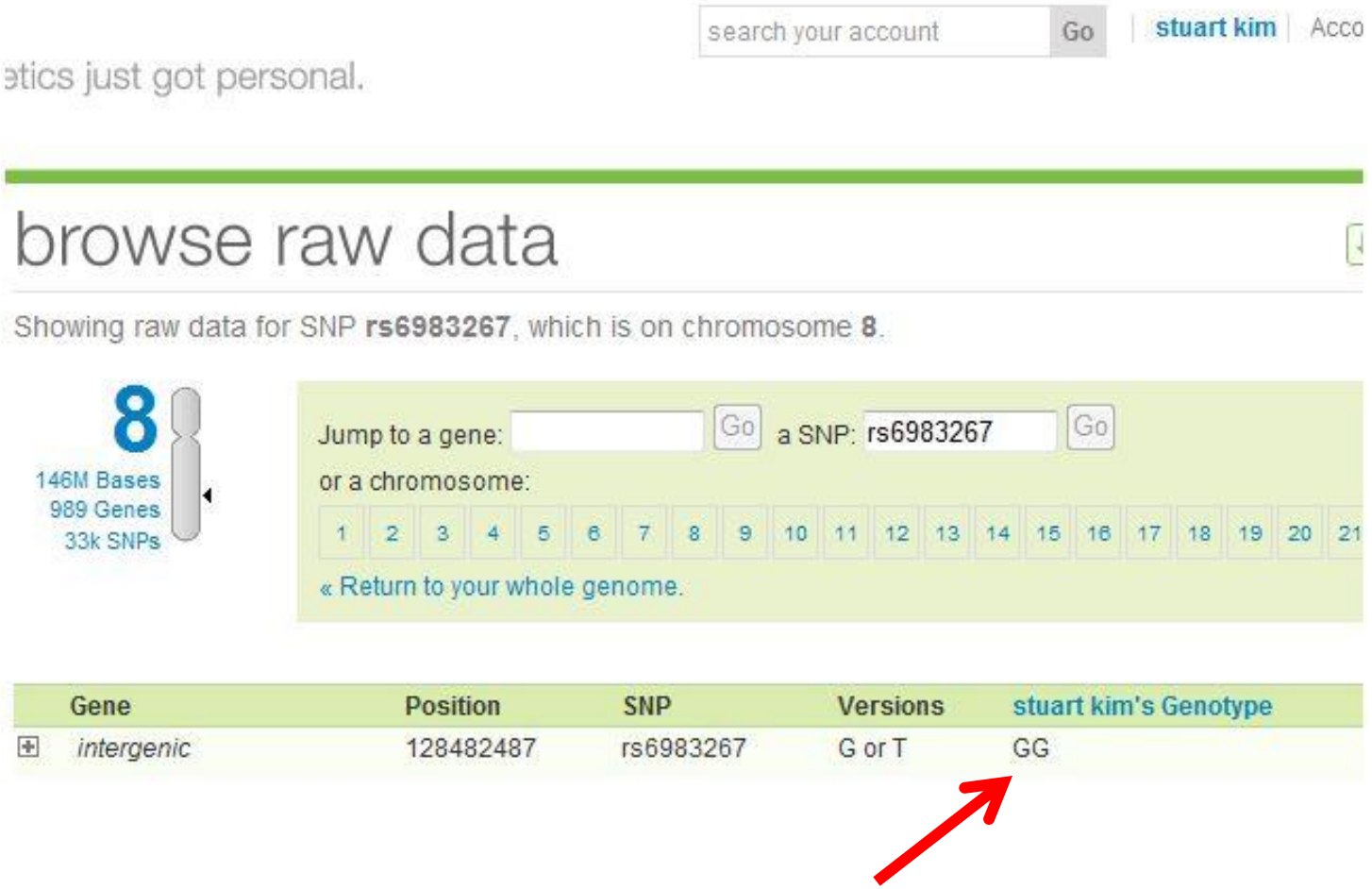
http://www.graphpad.com/quickcalcs/chisquared1.cfm




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

— = 1.37

CONTROL .49 G / .51 T = 0.96

# 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

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?