Power, Sample Size and Replication

David Evans[1,2,3]

1 Institute for Molecular Bioscience, University of Queensland

2 University of Queensland Diamantina Institute

3 MRC Integrative Epidemiology Unit, University of Bristol

Outline

1. Aims
2. Statistical power
3. Estimate the power of association analysis
   - Analytically
   - Empirically
4. Multiple Testing
5. Replication

1. Aims

1. Know what type-I error and power are

2. Know that you can/should estimate the power of your association analyses (analytically or empirically)

3. Know that there a number of tools that you can use to estimate power

4. Be aware that there are many factors that increase type-I error and decrease power

5. Be able to understand strategy and criteria for replication

2. Statistical power

H_0: There is NO association between a marker and a trait

H_1: There is association between a marker and a trait

		In reality…	
		H_0 is true	**H_1 is true**
We decide…	**H_0 is true**	$1 - \alpha$	β Type-2 error
	H_1 is true	α Type-1 error	$1 - \beta$ Power

Power: probability of detecting association when H_1 is true.

Definitions

- Power

  The probability of detecting a given size effect in a population from a sample size N, using significance criterion alpha

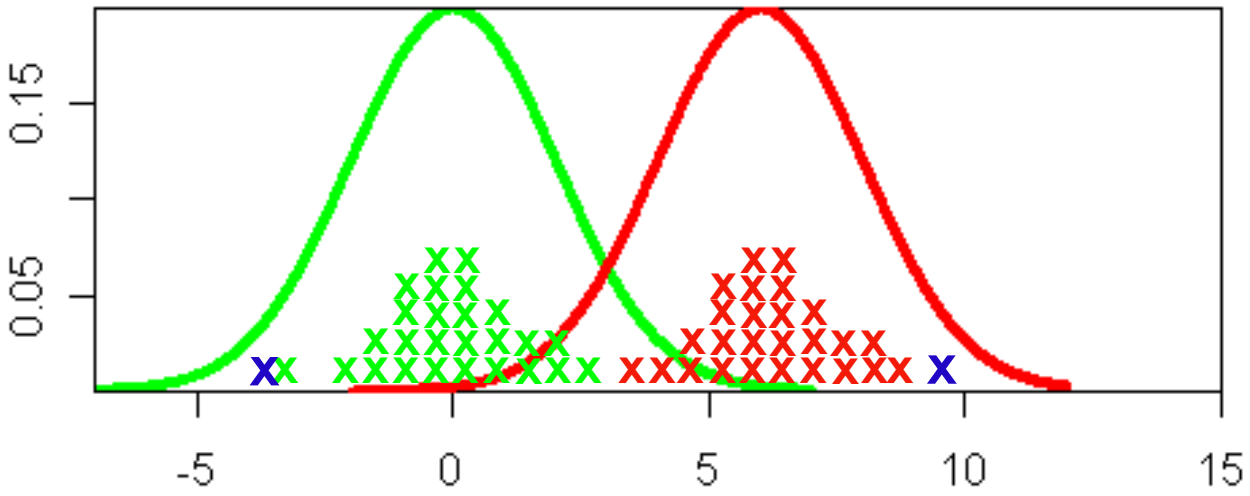
- Type I error

  The probability of incorrectly rejecting the null hypothesis of no association

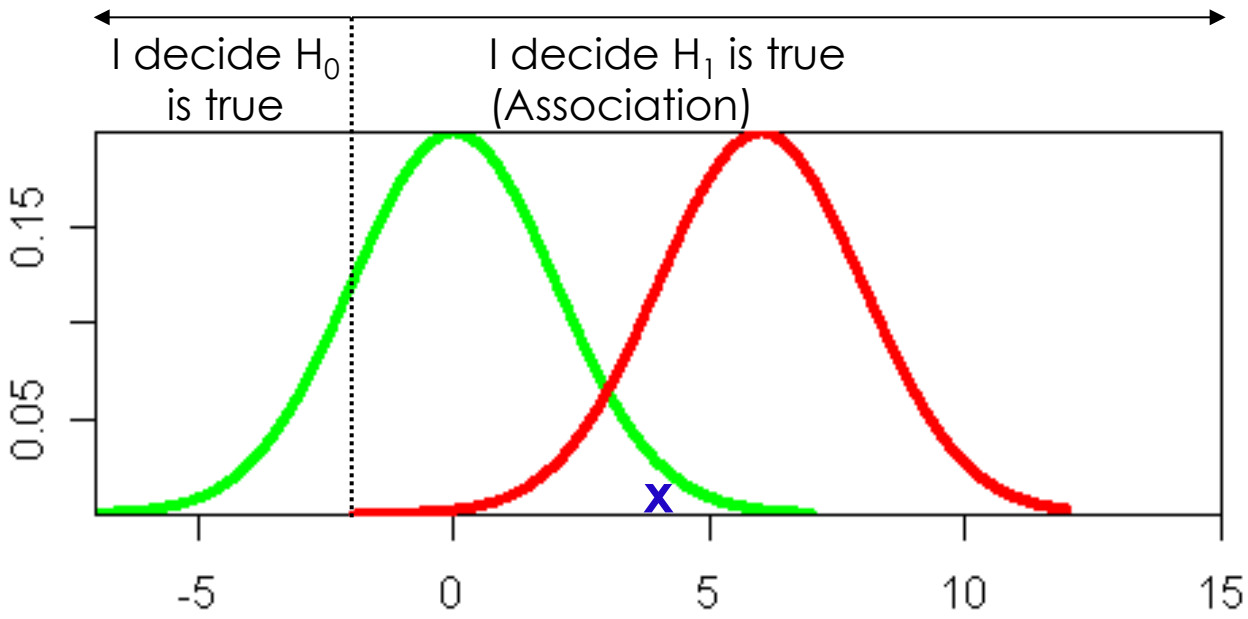
H_0: There is NO association between a marker and a trait

H_1: There is association between a marker and a trait

Association test statistic has different distributions under H_0 and H_1

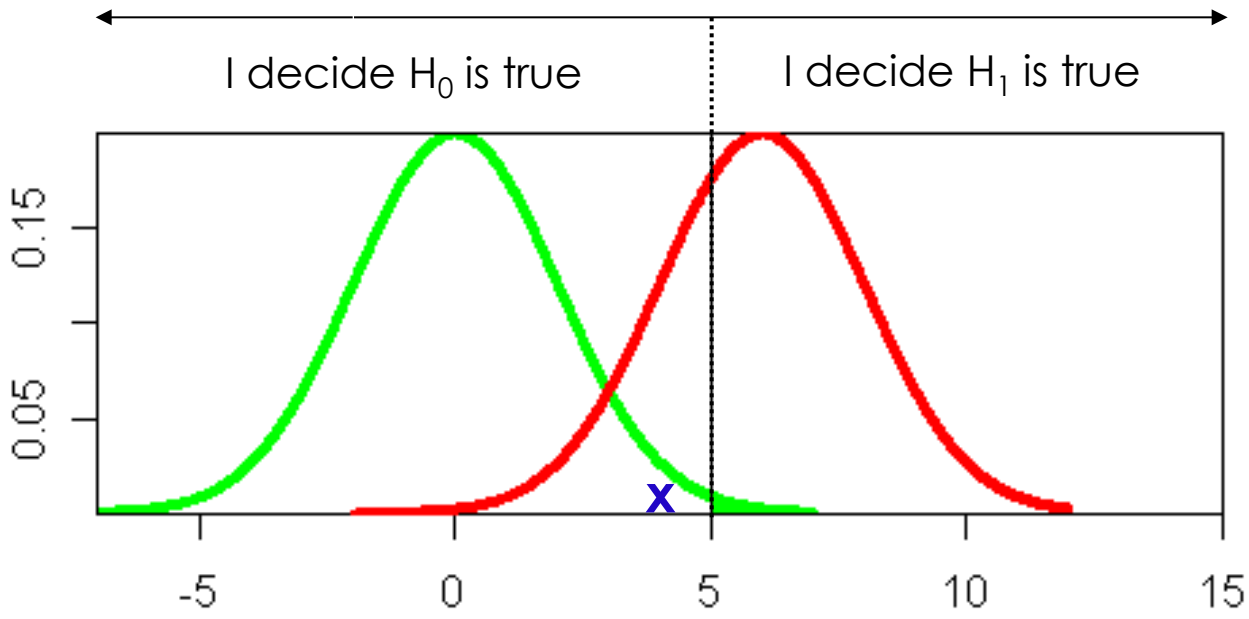


Where should I set the threshold to determine significance?



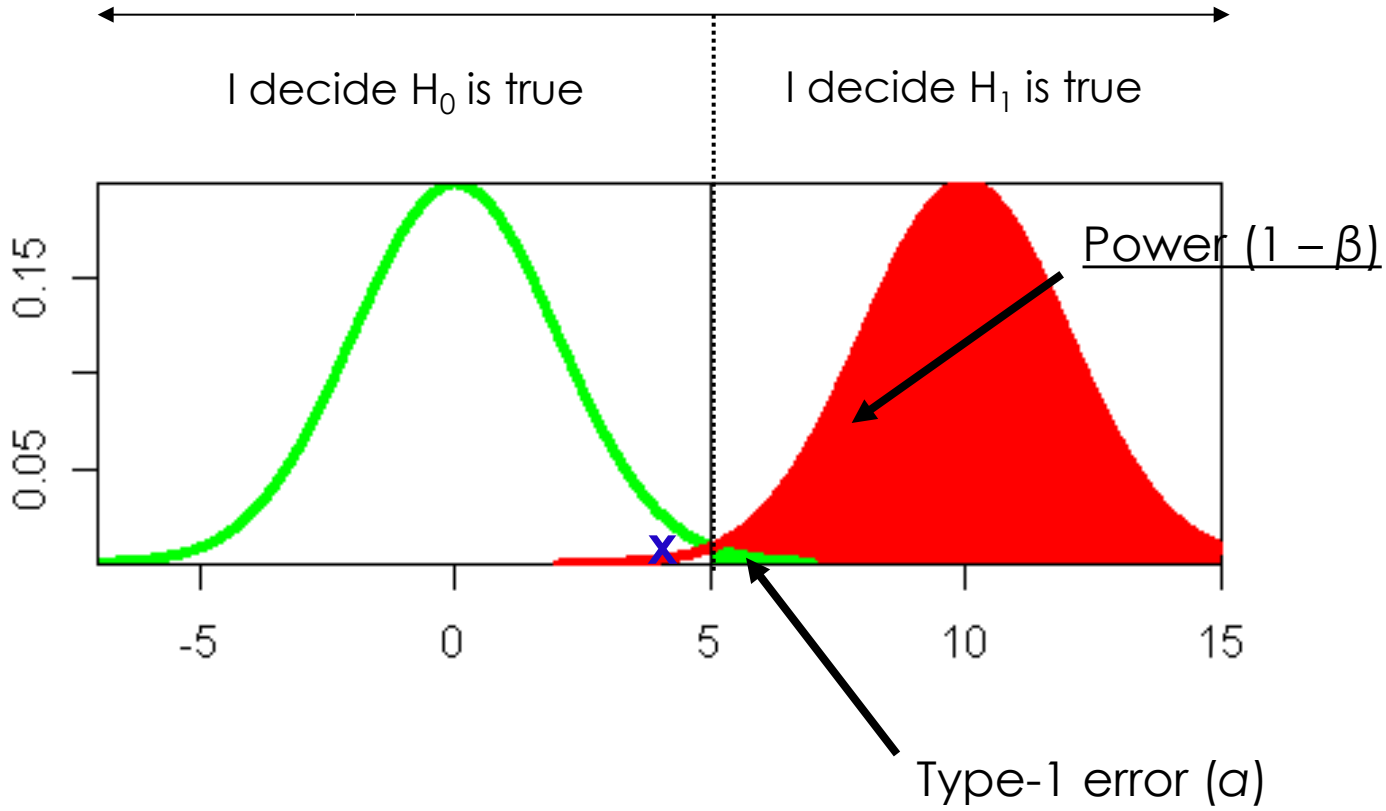
Threshold	Power ($1 - \beta$)	Type-1 error (α)
Too low	**High**	**High**

Where should I set the threshold to determine significance?



Threshold	Power ($1 - \beta$)	Type-1 error (α)
Too low	**High**	**High**
Too high	**Low**	**Low**

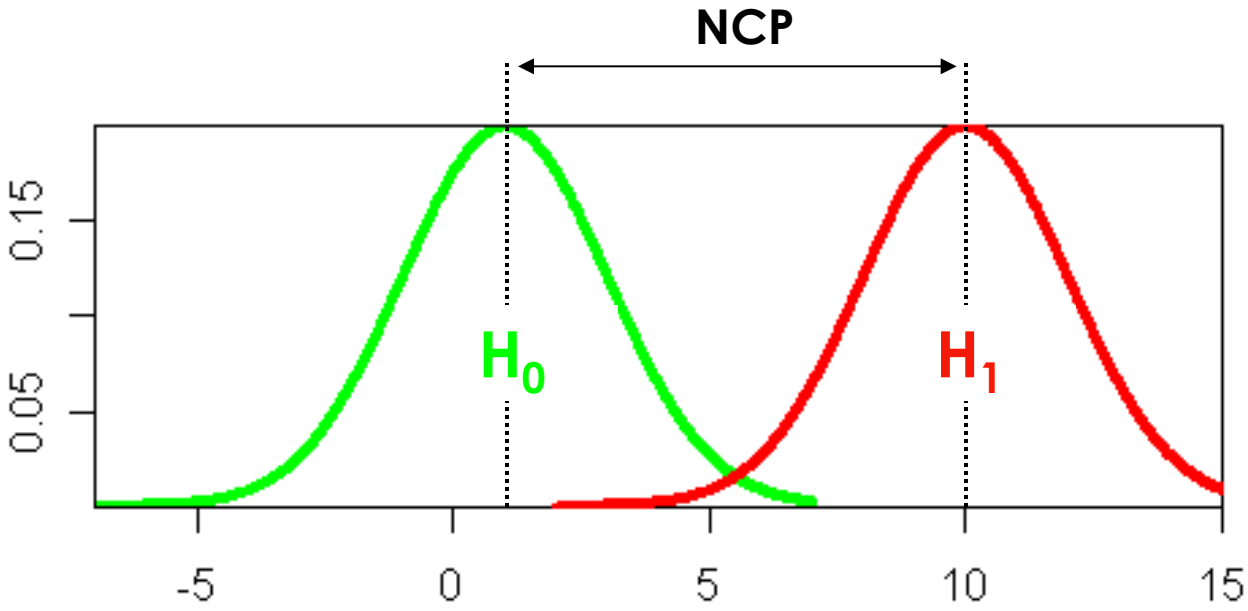
How do I maximise Power while minimising Type-1 error rate?



1. Set a high threshold for significance (i.e. results in low α [e.g. 0.05-0.00002])

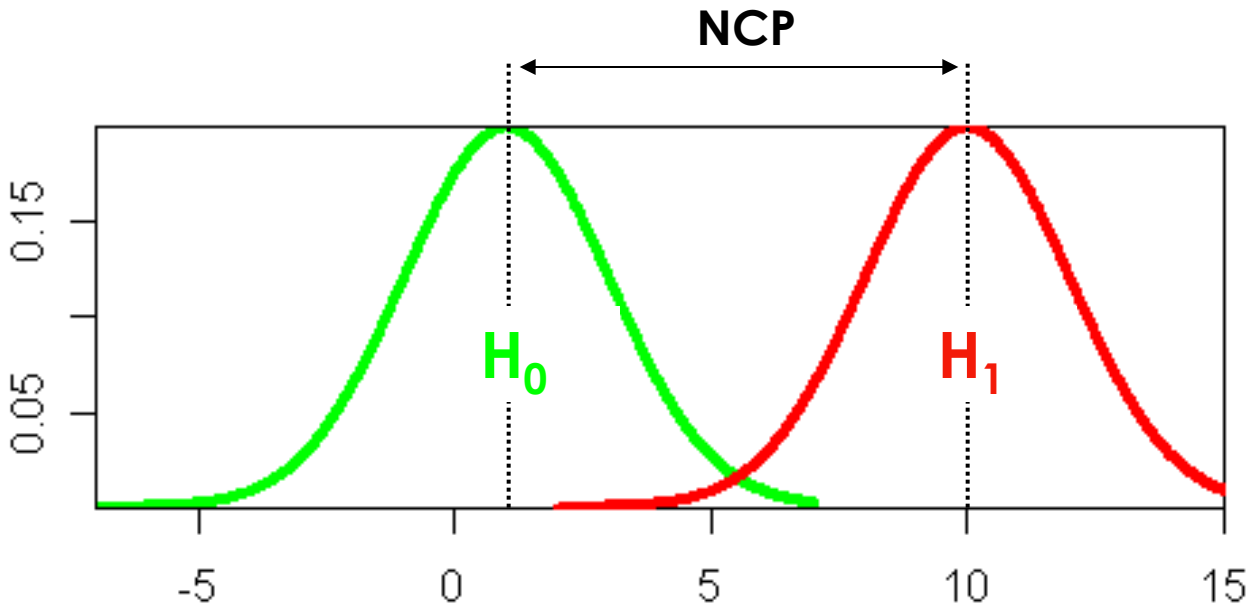
2. Try to shift the distribution of the association test statistic when H_1 is true as far as possible from the distribution when H_0 is true.

Non-centrality parameter



	Central X^2	Non-central X^2
Mean (μ)	**df**	**df + NCP**
Variance (σ^2)	**2*(df)**	**2*(df) + 4*NCP**

These distributions ARE NOT chi-sq with 1df!! Just for illustration… (Question- why do we use chi-sq?)



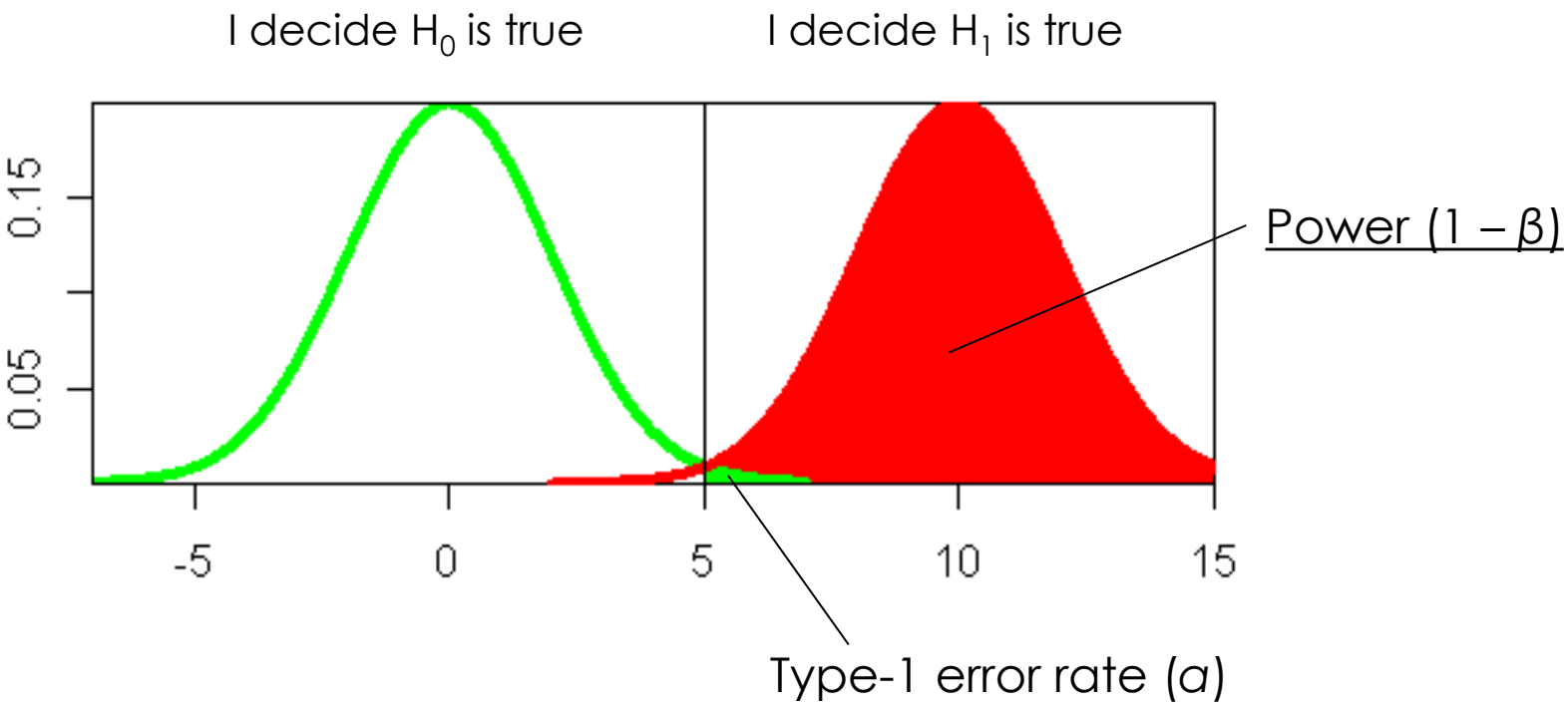
Small NCP → Big overlap between H_0 and H_1 distributions → **Lower power**

Large NCP → Small overlap between H_0 and H_1 distributions → **Greater power**

- Sample size does NOT scale linearly with Power
- But, sample size scales linearly with NCP

3. Estimate power for association

Theoretical power estimation



1. Set type 1 error rate (e.g. $\alpha = 0.05$)
2. Determine what critical value this corresponds to on the X axis
3. Work out the non-centrality parameter of the test (NCP = $E(H_1) - E(H_0)$)
4. Calculate the area to the right of the threshold under H_1

Trivial Example: OLS Linear Regression

Under H_0:

($\beta = 0$)

$$\frac{b - 0}{SE} \longrightarrow Z(0,1)$$ (*in large samples)

$$\frac{(b - 0)^2}{SE^2} \longrightarrow \chi^2_1$$ Central chi-square distribution

Under H_1:

($\beta \neq 0$)

$$\frac{b - 0}{SE} \longrightarrow Z(\frac{\beta}{SE}, 1)$$

$$\frac{(b - 0)^2}{SE^2} \longrightarrow \chi^2_1(\frac{\beta^2}{SE^2})$$ Non-central chi-square distribution

Trivial Example: OLS Linear Regression

$$\frac{(b-0)^2}{SE^2} \longrightarrow \chi^2_1\left(\frac{\beta^2}{SE^2}\right)$$

1. Set type 1 error rate (e.g. $\alpha = 0.05$)
2. Determine what critical value this corresponds to on the X axis

```
qchisq(p = 0.05, df = 1, ncp = 0, lower.tail = FALSE, log.p = FALSE)
```

```
[1] 3.841459
```

3. Work out the non-centrality parameter of the test $\frac{\beta^2}{SE^2}$

$\beta = 0.1$

$SE \approx 1/\sqrt{N}$ (Assume N = 1000)

4. Calculate the area to the right of the threshold under H_1

```
pchisq(q=3.84, df=1, ncp = 0.1^2/(1/1000), lower.tail = FALSE, log.p = FALSE)
```

```
[1] 0.8854512
```

Factors that influence power and type-1 error

	Association	
	Quantitative	Case-control
1. Disease model Effect size, MAF, disease prevalence	✓	✓
2. Genome coverage (r^2)	✓	✓
3. Sample size	✓	✓
4. Ascertainment	✓	✓
5. Deviations in trait distribution	✓✓	✓
6. Measurement error*	✓	✓
7. Genotyping errors*	✓	✓
8. Missing data*	✓	✓

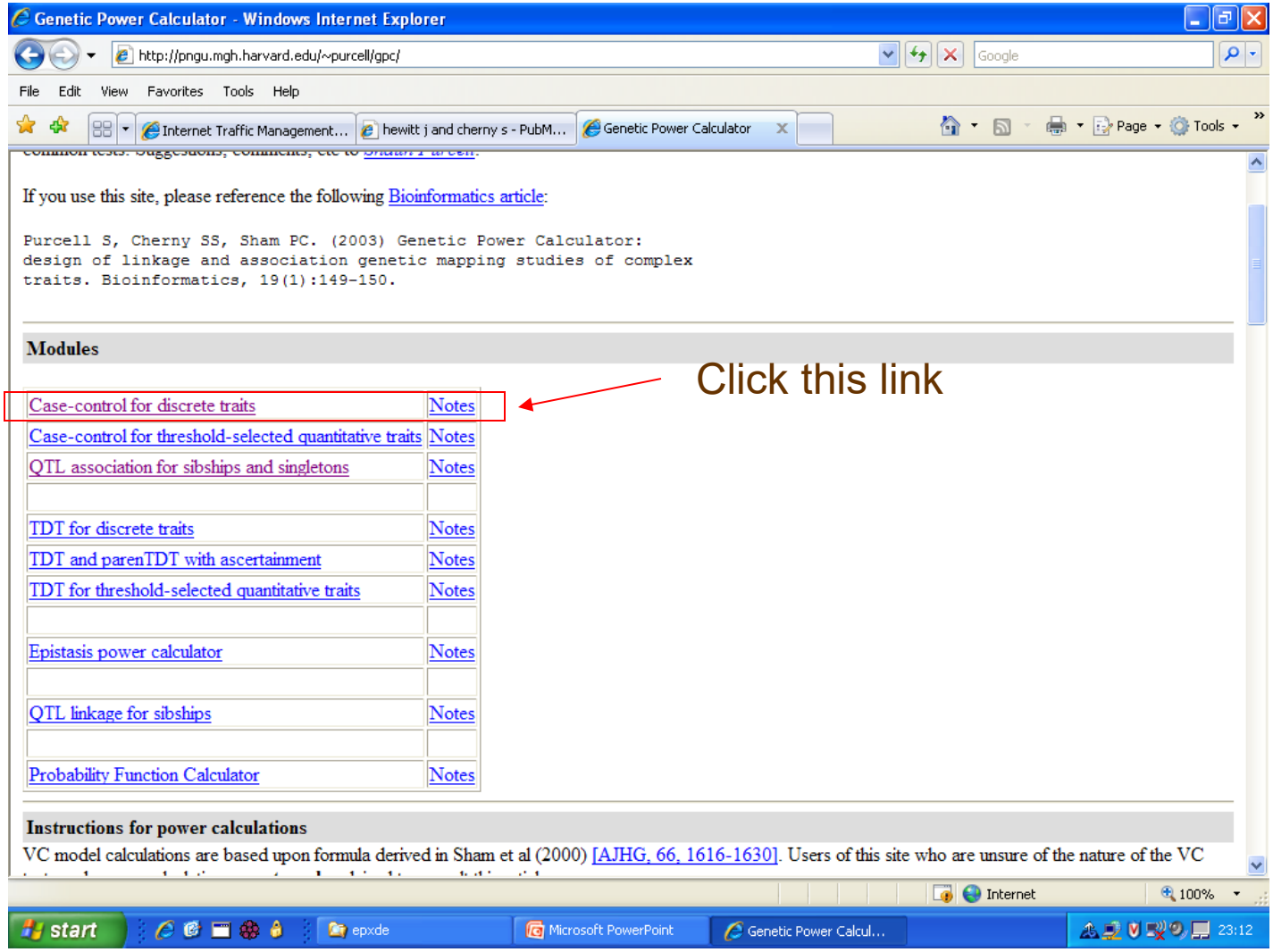
*Assume random

Theoretical power estimation: Exercise

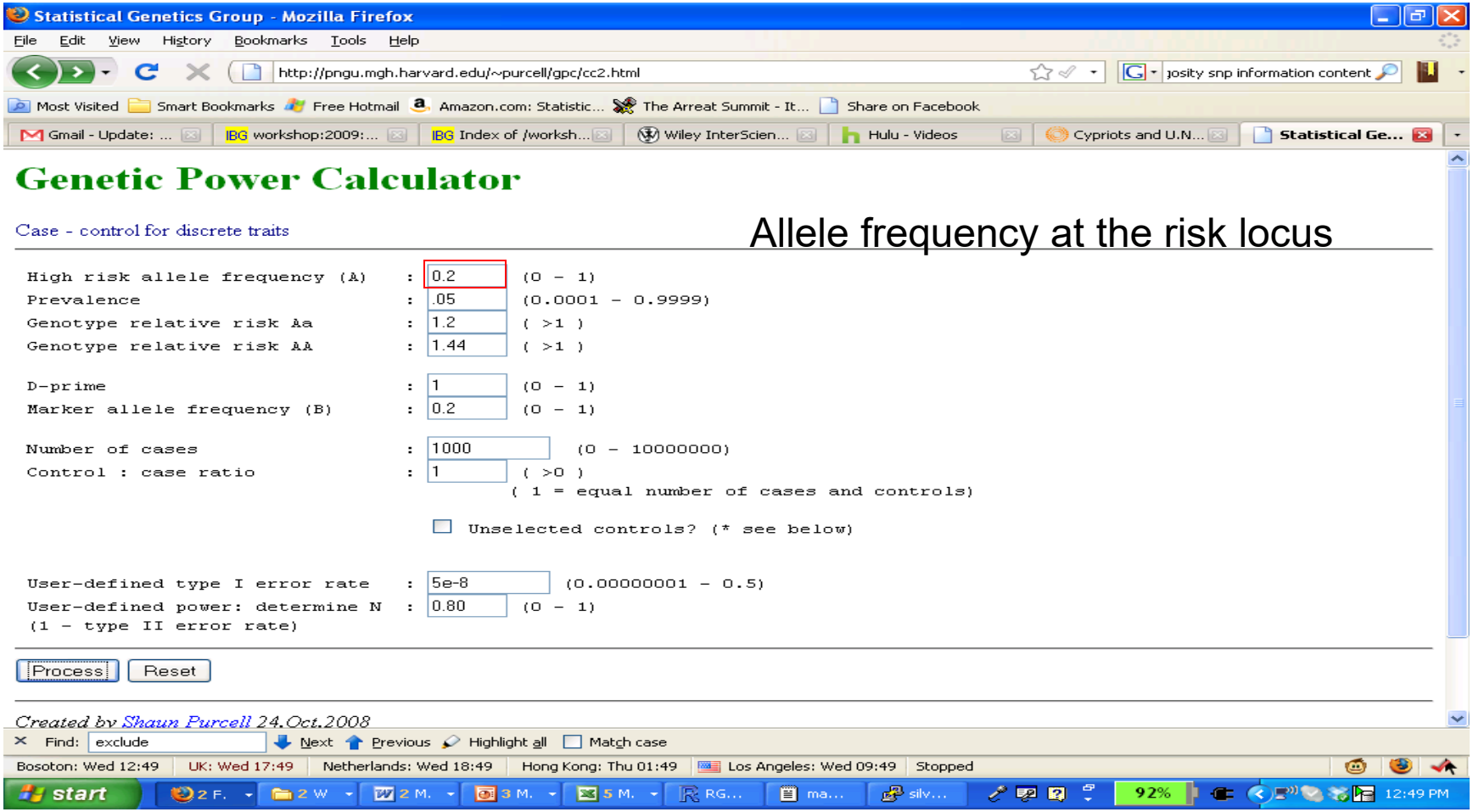
- What case control sample size do we need to achieve 80% power for genome-wide significance for an odds ratio of 1.2 in a multiplicative model and an allele frequency of 20% when we directly type the locus for a disease with 5% prevalence?

Practical Exercise

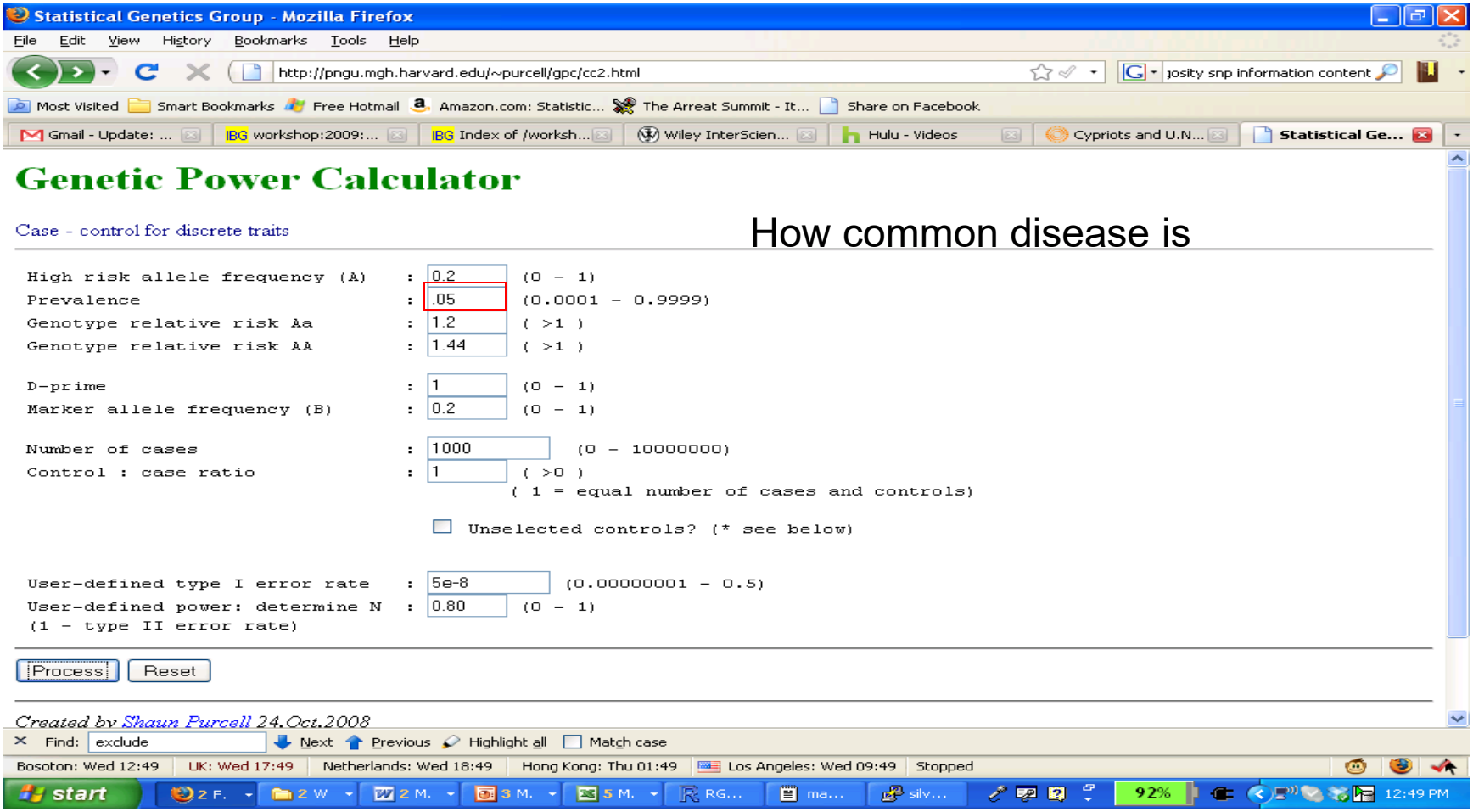
http://zzz.bwh.harvard.edu/gpc/ (Note new location!)



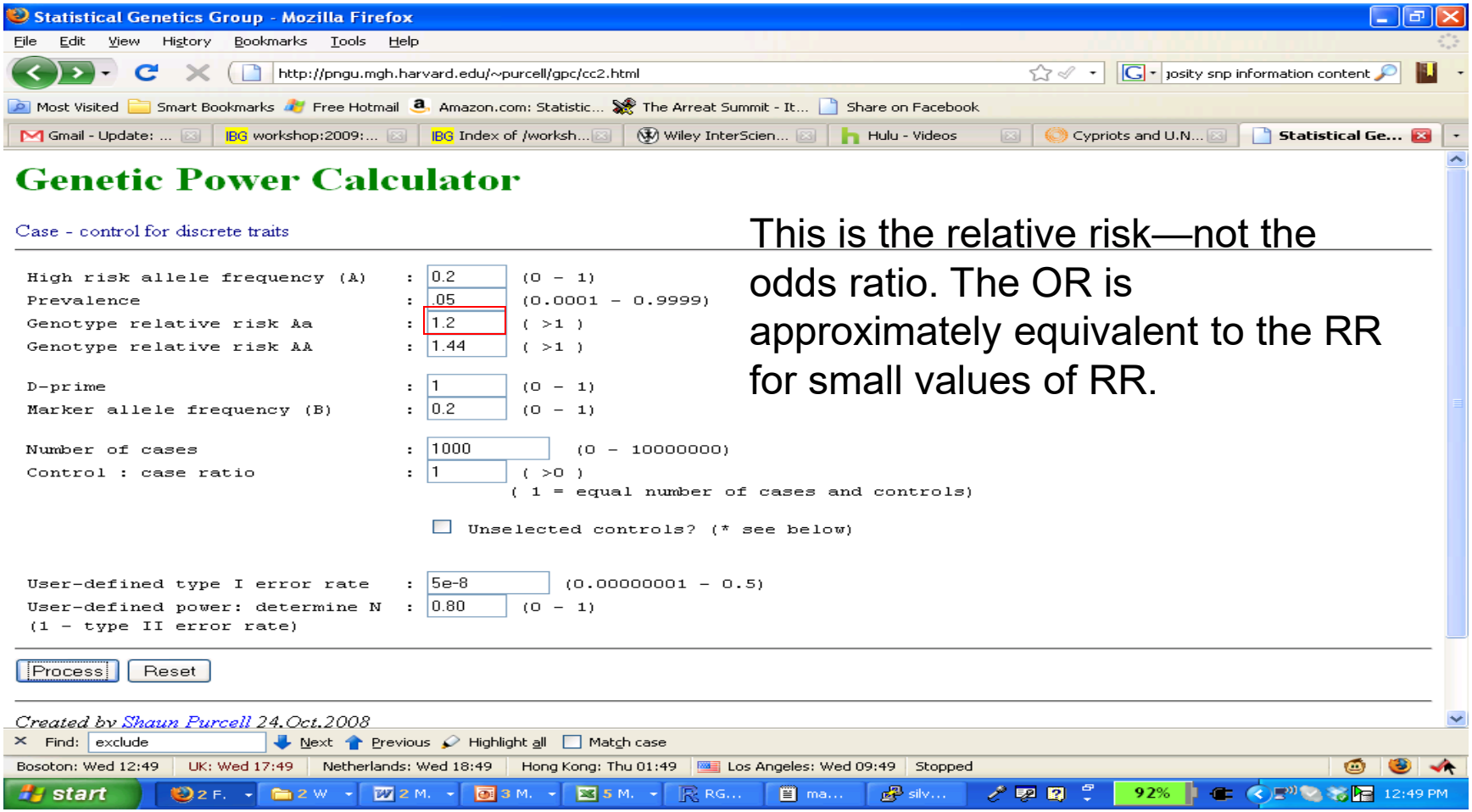
Practical Exercise



Practical Exercise

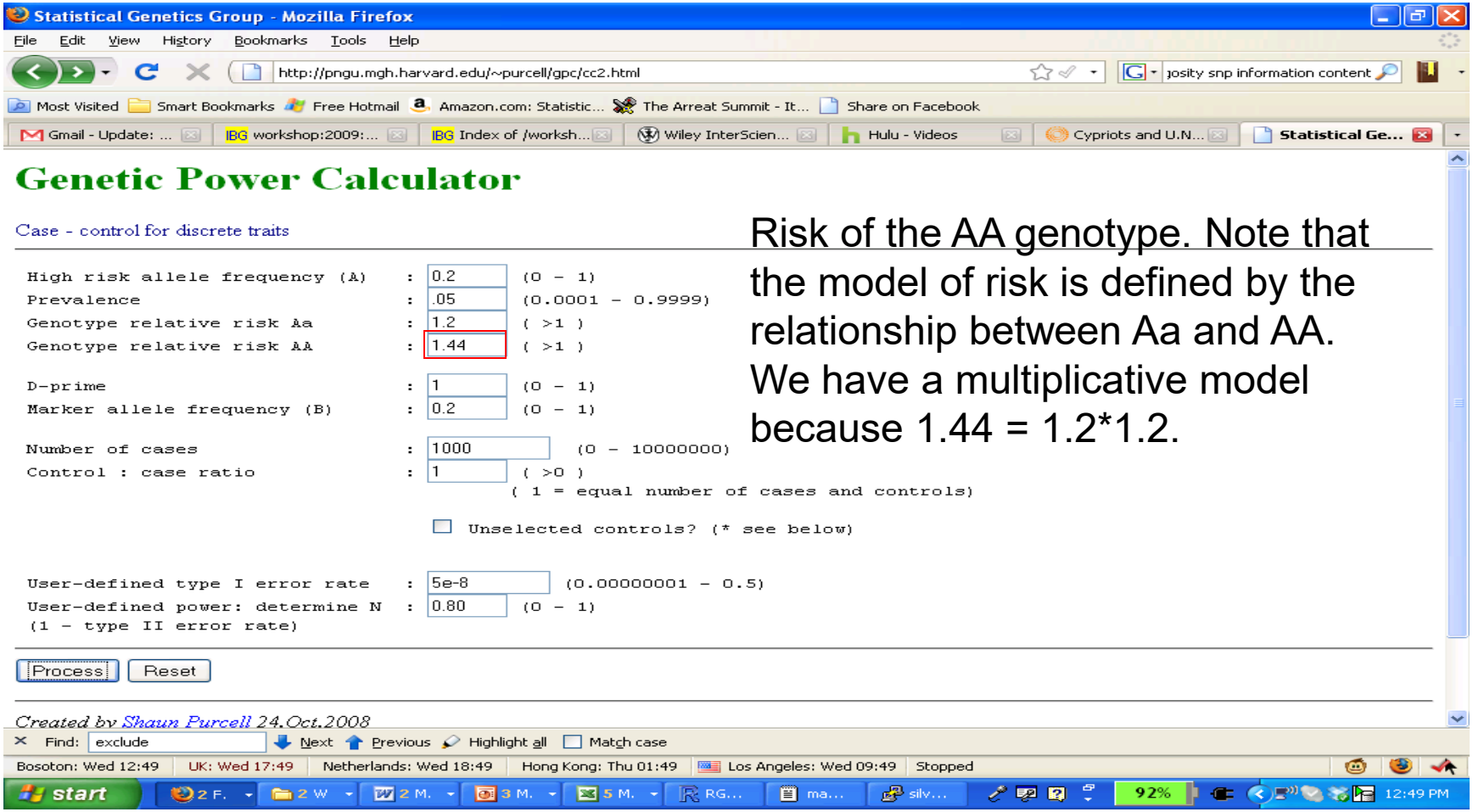


Practical Exercise



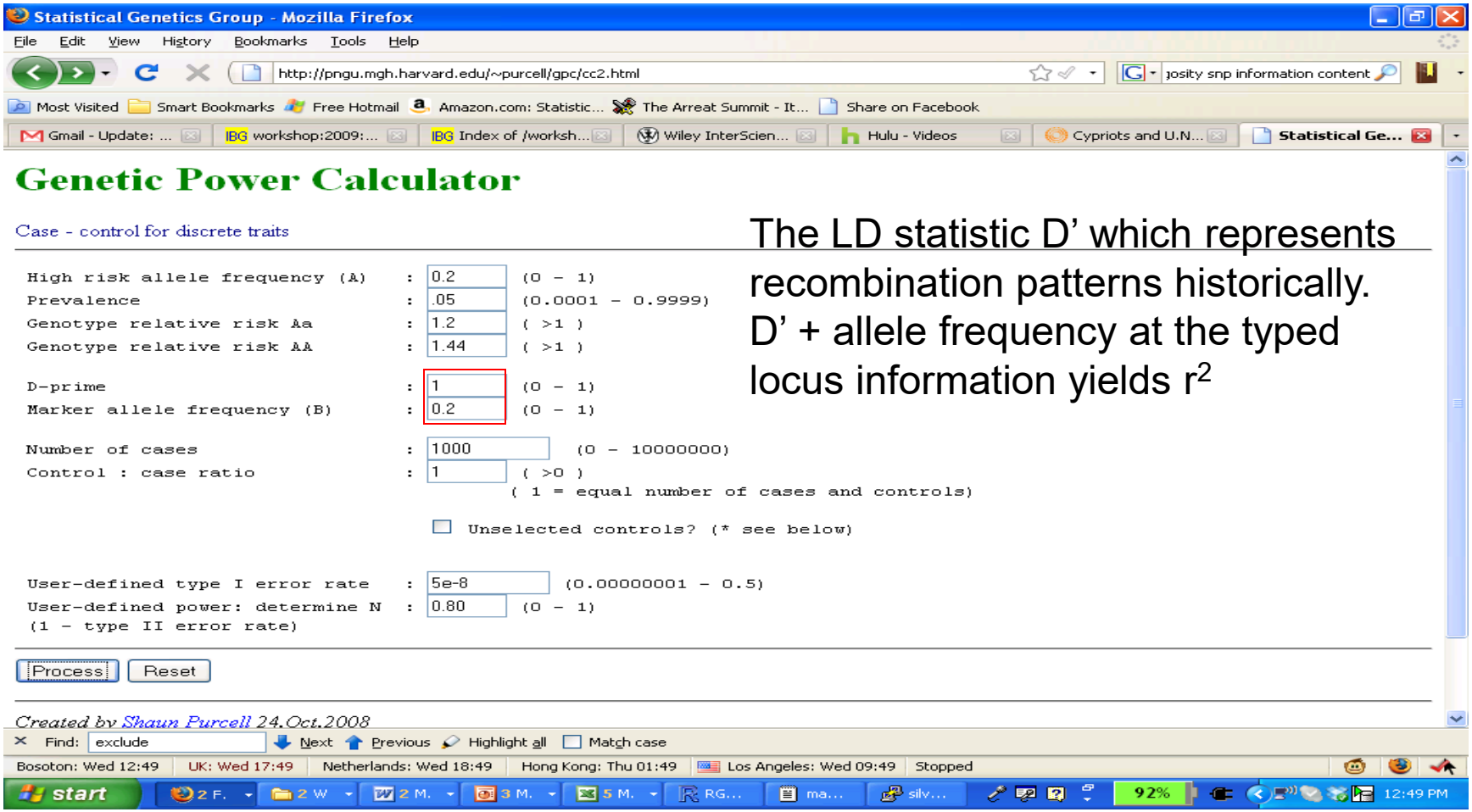
This is the relative risk—not the odds ratio. The OR is approximately equivalent to the RR for small values of RR.

Practical Exercise



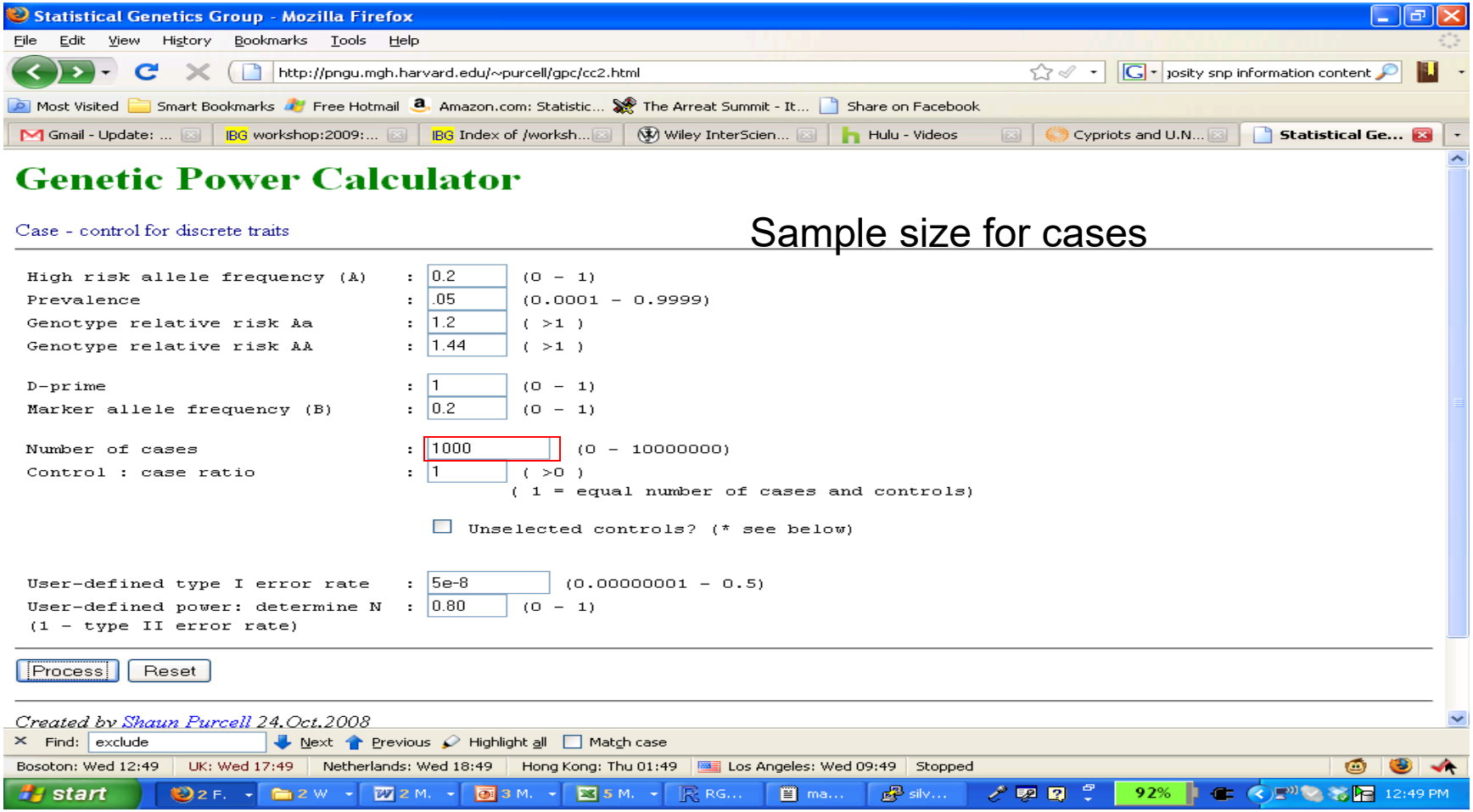
Risk of the AA genotype. Note that the model of risk is defined by the relationship between Aa and AA. We have a multiplicative model because 1.44 = 1.2*1.2.

Practical Exercise

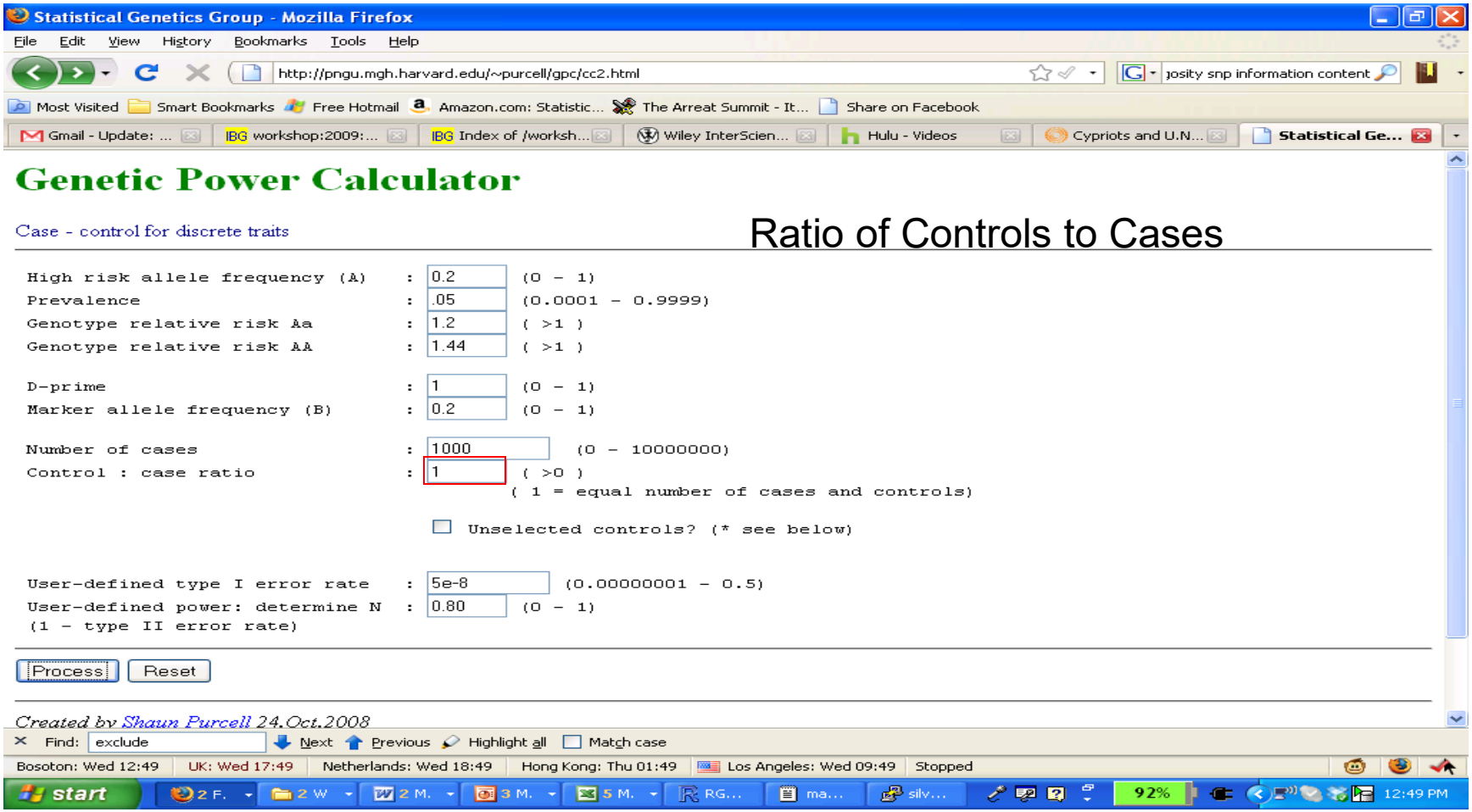


The LD statistic D' which represents recombination patterns historically. D' + allele frequency at the typed locus information yields r^2

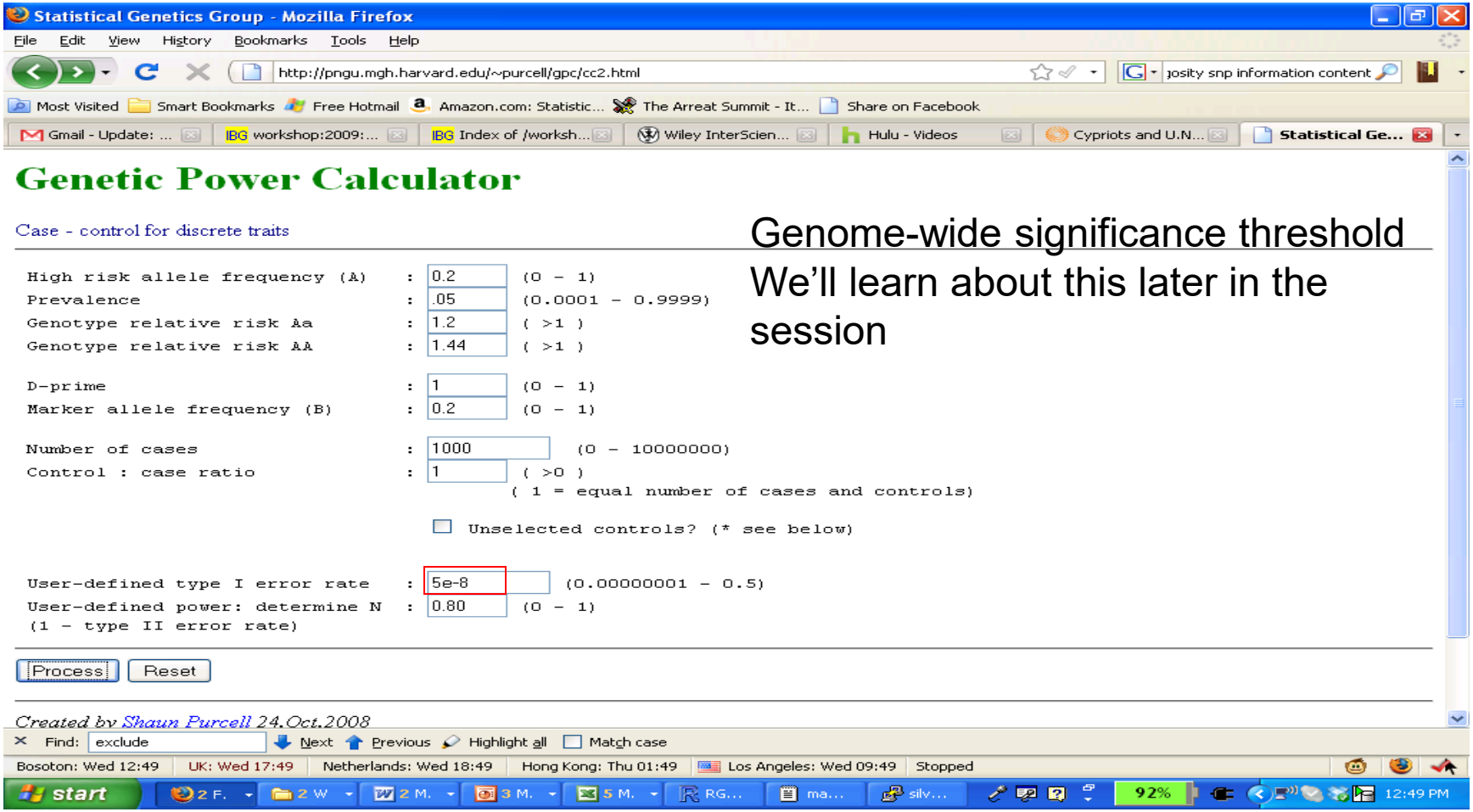
Practical Exercise



Practical Exercise



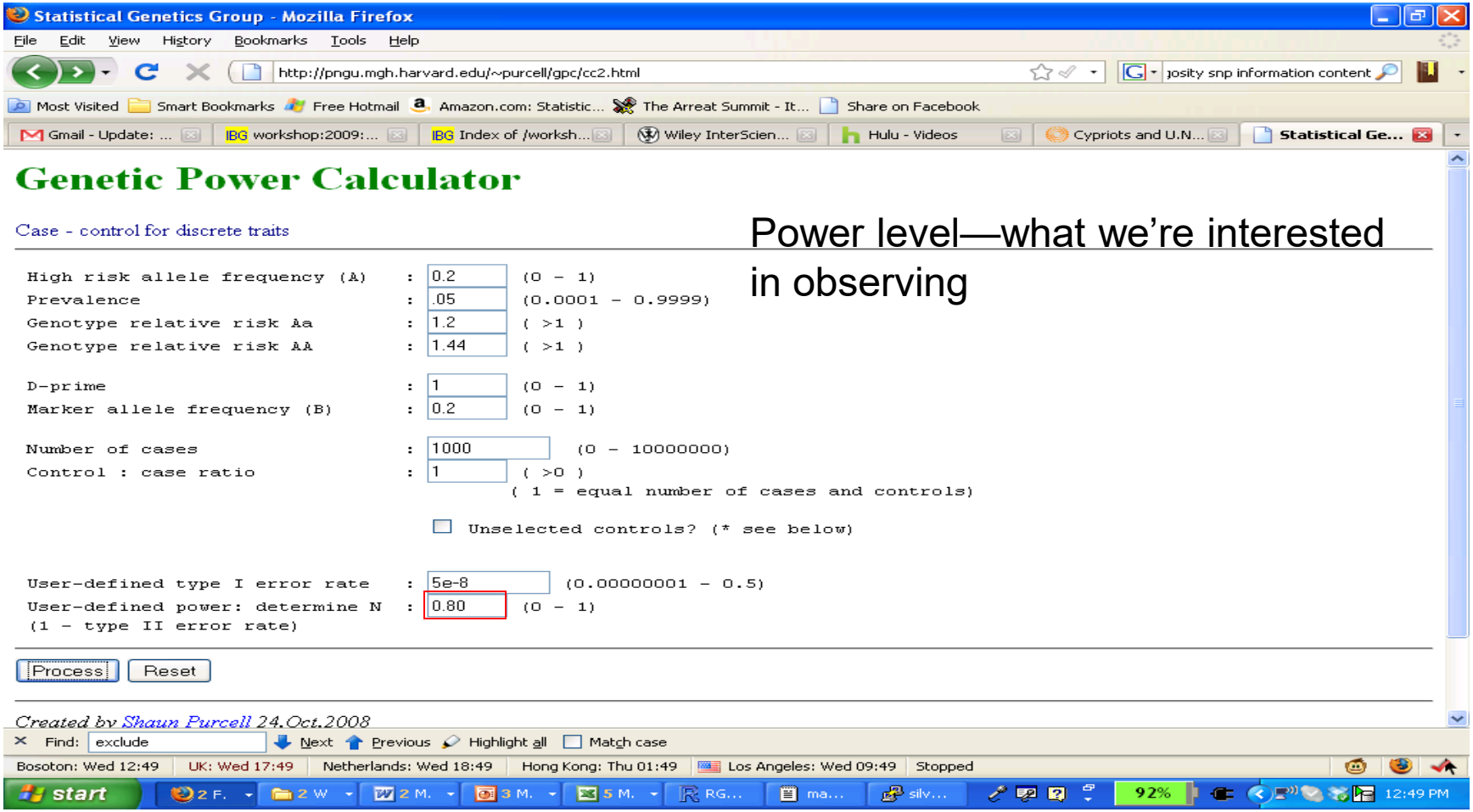
Practical Exercise



Genome-wide significance threshold

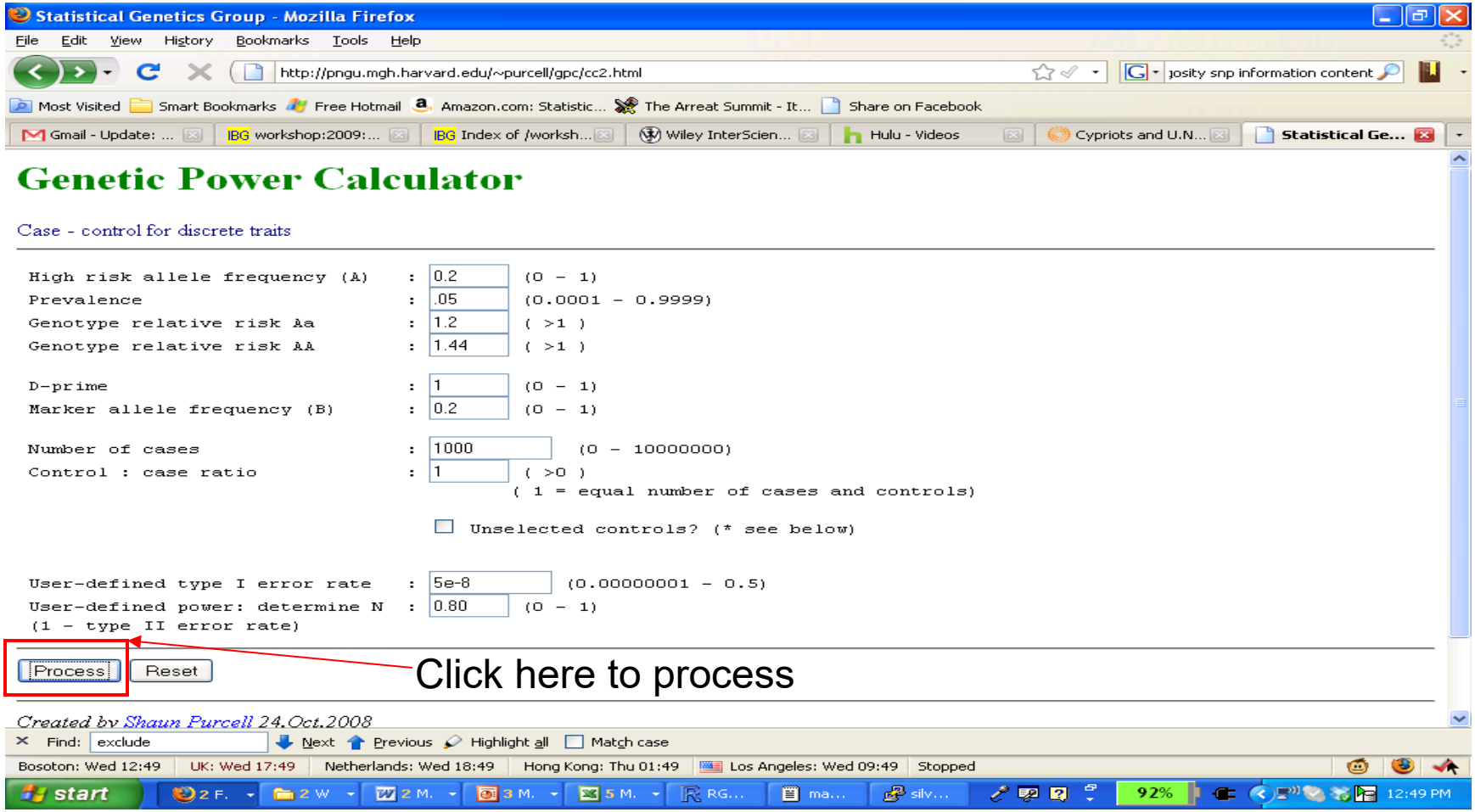
We'll learn about this later in the session

Practical Exercise

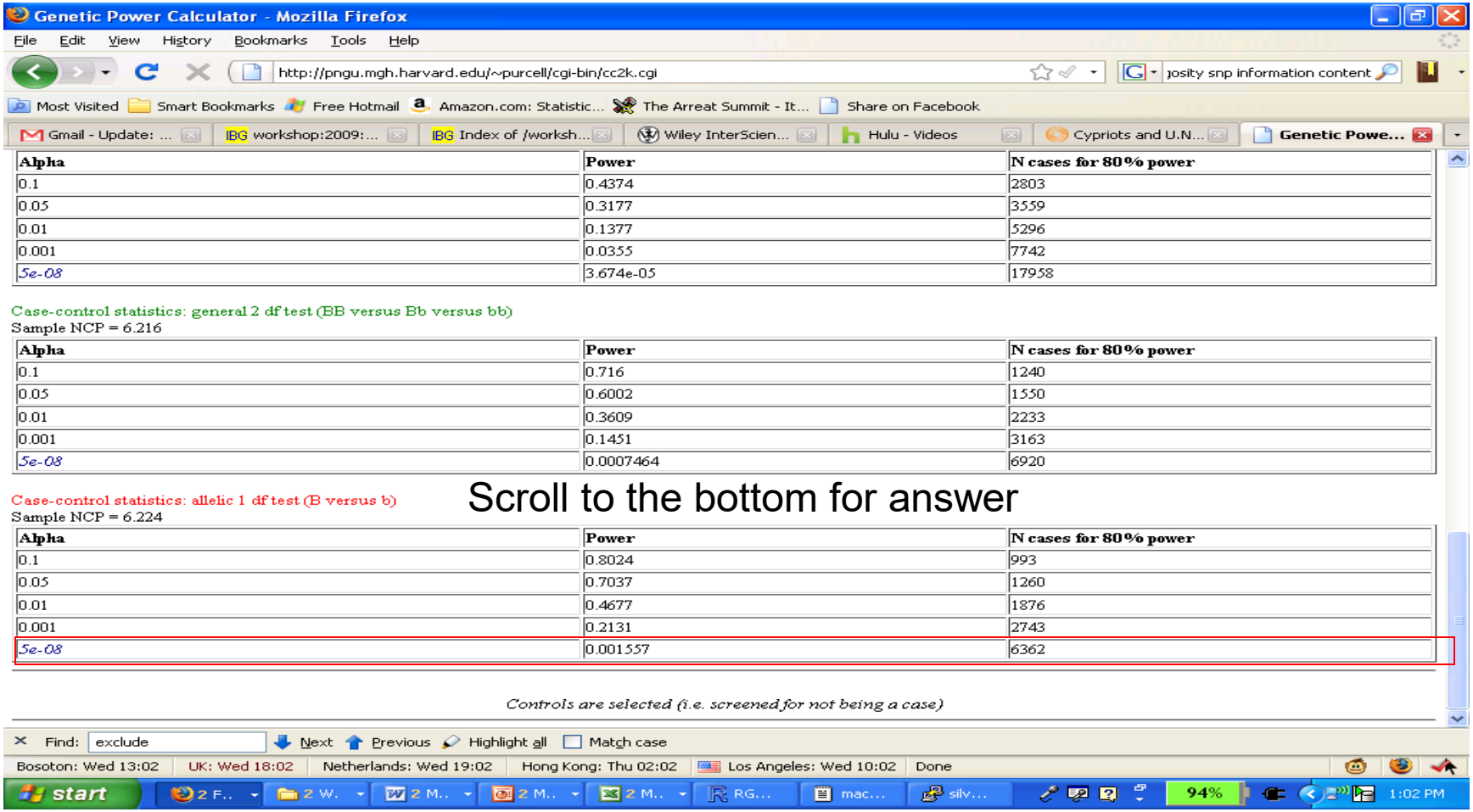


Power level—what we're interested in observing

Practical Exercise

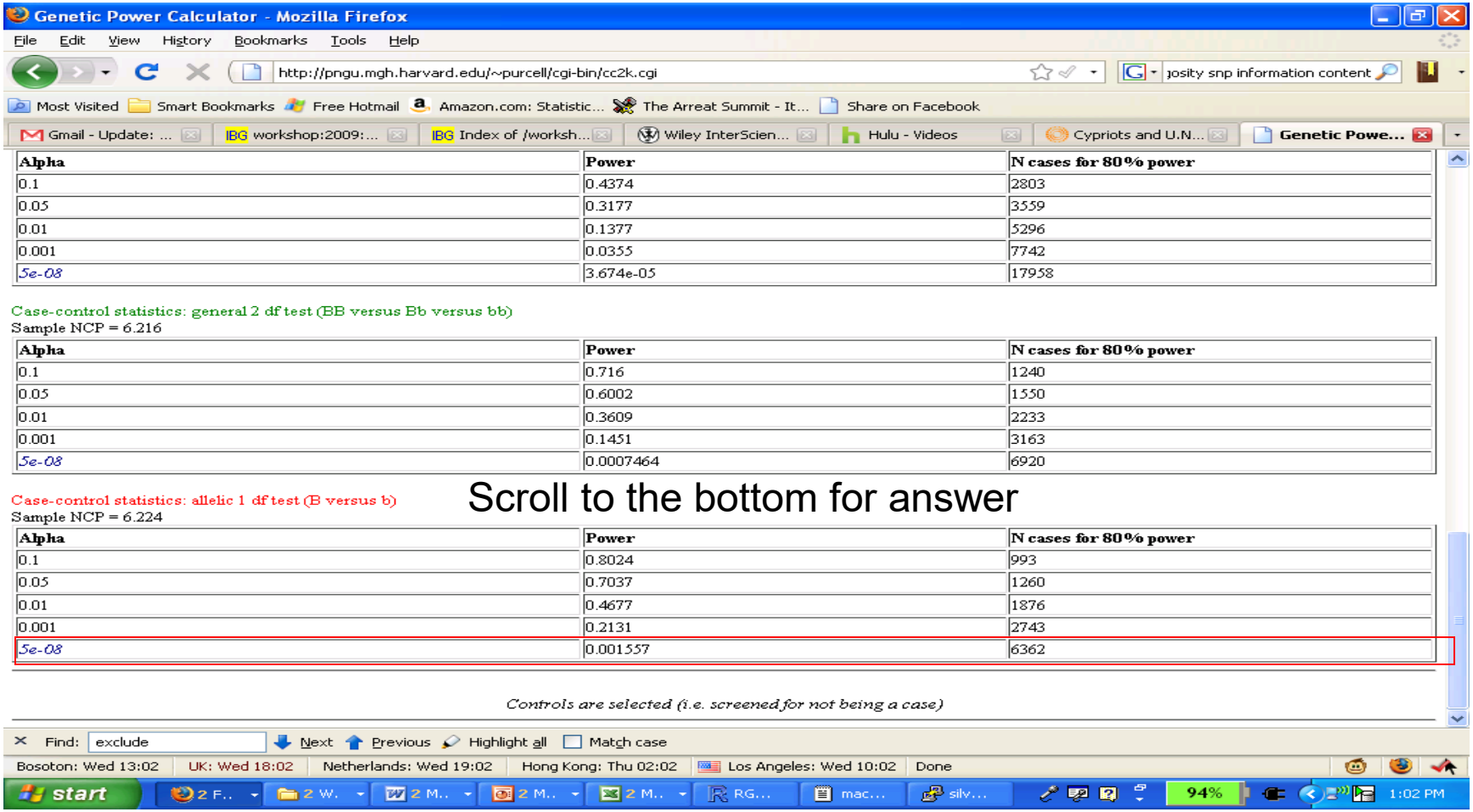


Practical Exercise



Scroll to the bottom for answer

Practical Exercise



Scroll to the bottom for answer

6,362 case samples required: total sample size 12,724

Practical Exercise

- What power do we have to detect a locus with an odds ratio of 1.5 at genome-wide significance assuming a multiplicative disease model, a disease allele frequency of 5%, a disease prevalence of 0.5% and 3000 cases and controls?

Genetic Power Calculator

Case - control for discrete traits

High risk allele frequency (A)	:	0.05	(0 - 1)
Prevalence	:	0.005	(0.0001 - 0.9999)
Genotype relative risk Aa	:	1.5	(>1)
Genotype relative risk AA	:	2.25	(>1)
D-prime	:	1	(0 - 1)
Marker allele frequency (B)	:	0.05	(0 - 1)
Number of cases	:	3000	(0 - 10000000)
Control : case ratio	:	1	(>0)

(1 = equal number of cases and controls)

☐ Unselected controls? (* see below)

User-defined type I error rate	:	0.00000005	(0.00000001 - 0.5)
User-defined power: determine N (1 - type II error rate)	:	0.80	(0 - 1)

Process Reset

Created by Shaun Purcell 24.Oct.2008

***Note** : unselected controls indicates a true random population sample (e.g. for a 1% disease, 1% of controls would also, by chance, have the disease); if this is unchecked (the*

0.01	0.2124	11085
0.001	0.06518	16206
5e-08	0.0001204	37587

Case-control statistics: general 2 df test (BB versus Bb versus bb)
Sample NCP = 28.11

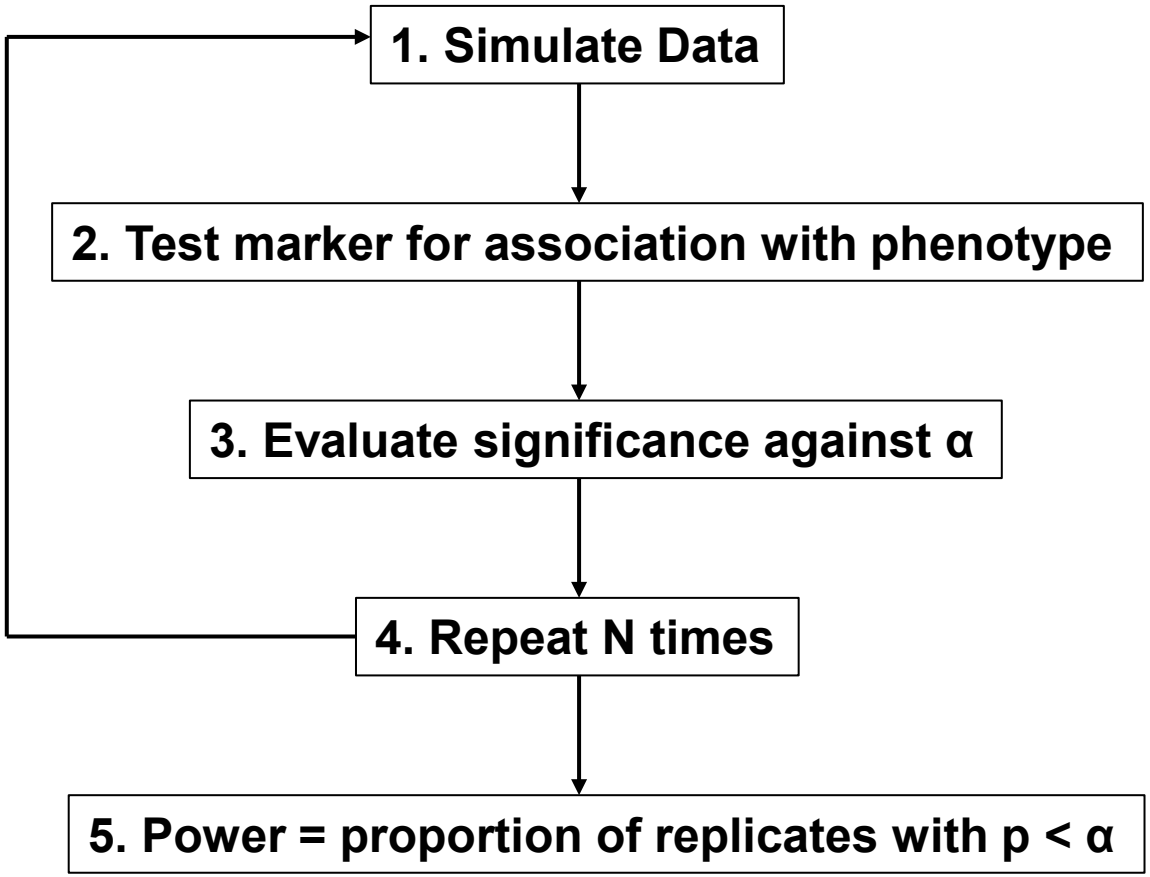
Alpha	**Power**	**N cases for 80% power**
0.1	0.9995	822
0.05	0.9986	1028
0.01	0.9916	1481
0.001	0.9553	2098
5e-08	0.3425	4590

Case-control statistics: allelic 1 df test (B versus b)
Sample NCP = 28.18

Alpha	**Power**	**N cases for 80% power**
0.1	0.9999	658
0.05	0.9996	835
0.01	0.9969	1243
0.001	0.9782	1818
5e-08	0.443	4216

Controls are selected (i.e. screened for not being a case)

Empirical power estimation



```
rm(list=ls())
set.seed(12345) #Set seed to enable repeatability
p <- 0.5 #(Decreaser) Allele frequency
q <- 1 - p #Other allele
Nsample <- 1000 #Set sample size
Nrep <- 1000 #Set number of replications
Nsig <- 0 #Count variable for number of significant replicates
threshold <- 0.05 #Threshold for declaring significance
beta <- 0.1 #Effect size
a <- sqrt(1/(2*p*q)) #Additive value
residvar <- 1 - beta^2 #Residual variance
for (i in seq(1,Nrep)) { #Loop over replicates
#Simulate genotypes assuming HWE
  x <- sample(x=c(-a,0,a) ,size=Nsample, replace=TRUE, prob=c(p^2,2*p*q,q^2))
  y <- beta*x + rnorm(n=Nsample, mean=0, sd=sqrt(residvar))
  if(summary(lm(y~x))$coefficients[2,4] < threshold) {Nsig=Nsig+1}
 }
power = Nsig/Nrep #Power is proportion of significant replicates
Power #0.885
```

Trivial Example: OLS Linear Regression

$$\frac{(b-0)^2}{SE^2} \longrightarrow \chi^2_1\left(\frac{\beta^2}{SE^2}\right)$$

1. Set type 1 error rate (e.g. $\alpha = 0.05$)
2. Determine what critical value this corresponds to on the X axis

```
qchisq(p = 0.05, df = 1, ncp = 0, lower.tail = FALSE, log.p = FALSE)
```

```
[1] 3.841459
```

3. Work out the non-centrality parameter of the test $\frac{\beta^2}{SE^2}$

$\beta = 0.1$

$SE \approx 1/\sqrt{N}$ (Assume N = 1000)

4. Calculate the area to the right of the threshold under H_1

```
pchisq(q=3.84, df=1, ncp = 0.1^2/(1/1000), lower.tail = FALSE, log.p = FALSE)
```

```
[1] 0.8854512
```

4. Multiple Testing

Genome-wide Association

High throughput genotyping

Other Multiple Testing Considerations

- Genome-wide association is really bad
  - At 1 test per SNP for 500,000 SNPs
  - 25,000 expected to be significant at $p<0.05$, by chance alone

Other Multiple Testing Considerations

- Genome-wide association is really bad
  - At 1 test per SNP for 500,000 SNPs
  - 25,000 expected to be significant at $p<0.05$, by chance alone
- To make things worse
  - Dominance (additive/dominant/recessive)
  - Epistasis (multiple combinations of SNPs)
  - Multiple phenotype definitions
  - Subgroup analyses
  - Multiple analytic methods

Bonferroni Correction

- For testing 500,000 SNPs
  - 5,000 expected to be significant at $p<0.01$
  - 500 expected to be significant at $p<0.001$
  - ……
  - 0.05 expected to be significant at $p<0.0000001$
- Suggests setting significance level to $\alpha = 10^{-7*}$
- Bonferroni correction for m tests
  - set significance level for p-values to $\alpha = 0.05 / m$
  - (or adjust the p-values to $m \times p$, before applying the usual $\alpha = 0.05$ significance level)

- *See Risch and Merikangas 1999

Genome-wide Significance

- Multiple testing theory requires an estimate of the number of 'independent tests'
- Risch and Merikangas 1996 estimated a threshold of 10^{-6} = (0.05/(5*10,000))
- HapMap 2005 estimate 10^{-8} based on encode deep sequencing in ENCODE regions
- Dudbridge and Gusnato, and Pe'er et al. 2008 Genetic Epidemiology estimate based on 'infinite density' like Lander and Kruglyak 1995 generate $5x10^{-8}$

5. Replication

Replication

- Replicating the genotype-phenotype association is the “gold standard” for “proving” an association is genuine
- Most loci underlying complex diseases will not be of large effect
- It is unlikely that a single study will unequivocally establish an association without the need for replication

Winner's Curse

If the location of a variant and its phenotypic effect size are estimated from the same data sets, the effect size will be over-estimated, in many cases substantially. Statistical significance and the *estimated* magnitude of the parameter are highly correlated.

H Göring et al. Am J Hum Genetics 2001;69:1357-69

Guidelines for Replication

Replication studies should be of sufficient size to demonstrate the effect

Replication studies should conducted in independent datasets

Replication should involve the same phenotype

Replication should be conducted in a similar population

The same SNP should be tested

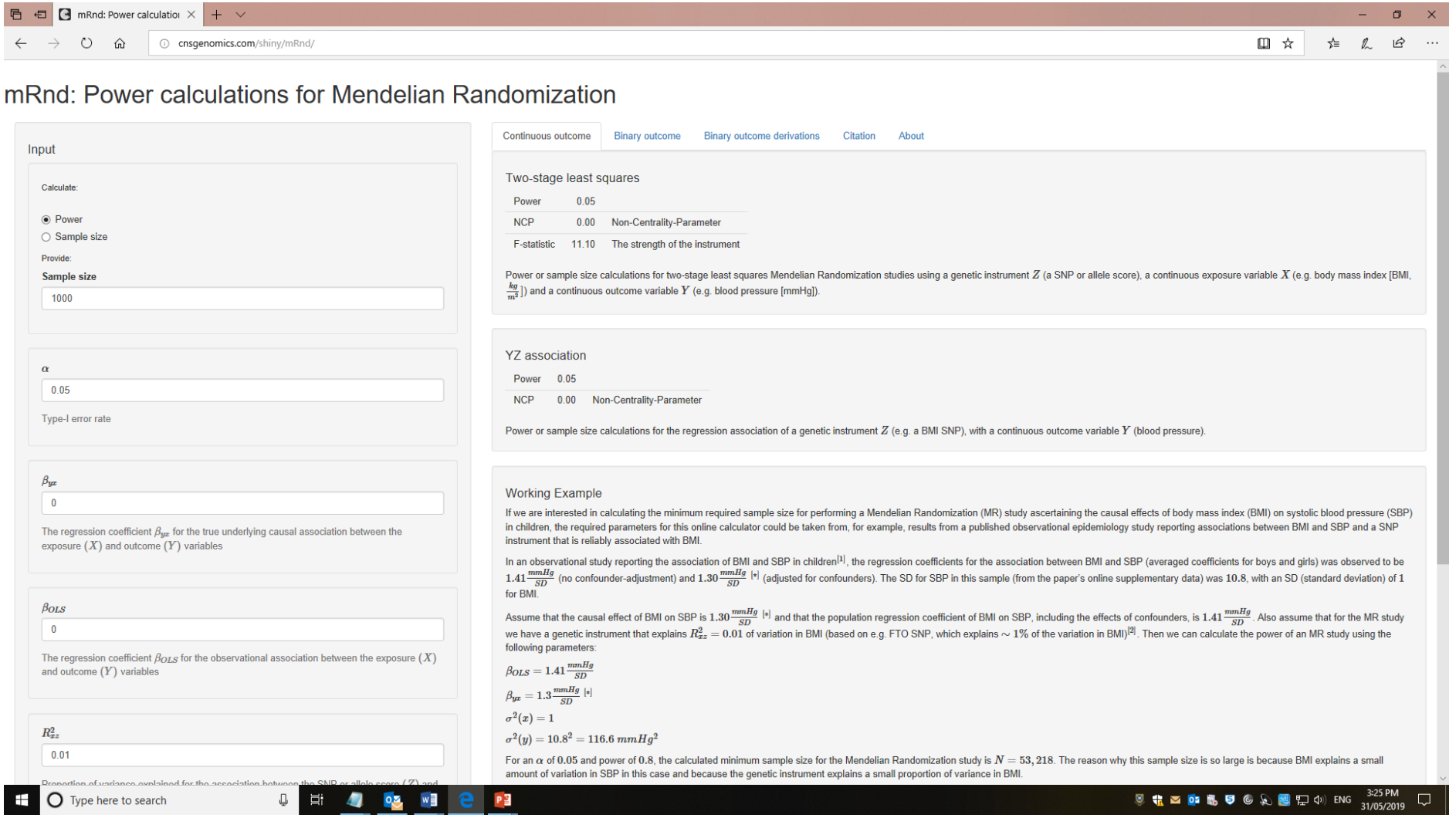
The replicated signal should be in the same direction

Joint analysis should lead to a lower p value than the original report

Well designed negative studies are valuable

Mendelian randomization Power

http://cnsgenomics.com/shiny/mRnd/



Mendelian randomization Power

http://cnsgenomics.com/shiny/mRnd/