VISG Virtual Institute of Statistical Genetics

VISG Experimental Designs Project 2011: Experimental Design for Genome-wide case-control studies.

Rod Ball

Scion

NZ Molecular Mapping meeting,
Havelock North, 27–28 Oct 2011

ABSTRACT

Hundreds of thousands of markers are being genotyped on thousands of individuals in genome-wide association studies, with the goal of the relating genotype to phenotype. Yet only a small proportion of heritable variation has been explained by putative SNP associations to date, raising the question: Where is the 'dark matter' in the genome?

We have previously developed an experimental design method for genome-wide case-control studies where we determine the power to detect effects with a given Bayes factor, where the Bayes factor is chosen sufficiently large to obtain a respectable level of posterior odds, for additive, dominant and recessive genetic models.

Our results show that current experiments lack power and many published effects are not robustly detected, which may explain the dark matter in the genome.

Our paper is now published in Genetics and is available at:

`http://www.genetics.org/content/early/2011/09/13/genetics.111.131698.abstract`

We will describe enhancements to the method including an improved estimator for the additive model and extension to the general model (2 d.f. test).

The method has been implemented in R and java. We will demonstrate power calculations using the R package which we propose to submit to CRAN, the comprehensive R archive network.

Talk outline

- VISG experimental designs project
- Case-control studies
- Additive model
  - Modified additive model estimator
- General model

Talk outline

- R package `ccDesign` and examples
- Java program
- Testing
- Next steps
- Conclusions

VISG experimental designs project

- Experimental design for genome-wide case control studies
  - find loci associated with (complex) diseases
  - design for power to detect effects with given Bayes factor
  - Bayes factor large enough to overcome low prior odds for genomic associations
  - ensure claimed effects are robustly detected
  - solves problem of poorly calibrated p-values, what threshold to choose

VISG experimental designs project

- paper published:

http://www.genetics.org/content/early/2011/09/13/genetics.111.131698.abstract

- R package developed
  - for submission to CRAN http://www.cran.r-project.org
  - waiting for VISG approval to release

Experimental Designs for Robust Detection of Effects in Genome-wide Case-control Studies

Roderick D. Ball*

[+] Author Affiliations

*To whom correspondence should be addressed. E-mail: rod.ball@scionresearch.com

Abstract

A method is given for the design of experiments to robustly detect associations between markers and causal loci in genome-wide case-control studies with a given strength of evidence as defined by the Bayes factor. Power calculations are derived using a generic approximate Bayes factor based on an estimate of the log-odds ratio and its standard error, leading to a closed form expression for the power. Expressions for the power are derived for bi-allelic loci for additive, dominant, or recessive modes of gene action. Robust detection of effects requires a Bayes factor of the order of 10^6 to overcome the low prior odds for individual SNP associations in a genome scan. This requires significantly larger effects or sample sizes and corresponds to a more stringent threshold than previously used. Power to detect effects with a range of allele frequencies, odds-ratios, linkage disequilibrium and strength of evidence, is demonstrated, and power to detect previously reported associations (APOE---Alzheimer's and the Wellcome Trust Case Control Consortium (WTCCC) SNPs) is re-examined. Power was low except for larger effects (e.g. odds-ratio around 2) at allele frequencies near 0.5, and where marker and causal loci had similar allele frequencies. At low minor allele frequencies (e.g. 0.05) power is low particularly in the recessive model

VISG experimental designs project

- improved estimator for additive model
- method extended to general model
- Java program developed
  - applied for summer eResearch project for GUI development

Case-control studies

- Pre-specified number of cases and controls

- Typically estimate odds ratio or relative risk

Table 1. Contingency table for a case-control study with 3 genotypic classes.

	aa	Aa	AA	
Control	n_{11}	n_{12}	n_{13}	n_1
Case	n_{21}	n_{22}	n_{23}	n_2

Additive model

2 odds ratios:

$$\frac{p_{11}p_{22}}{p_{12}p_{21}}, \quad \frac{p_{12}p_{23}}{p_{22}p_{13}} \tag{1}$$

assumed equal, log-odds ratio $\eta = \log \mathrm{OR}$ estimated as:

$$\begin{aligned} \widehat{\eta} &= \frac{1}{2}\left(\log \frac{p_{11}p_{22}}{p_{12}p_{21}} + \log \frac{p_{12}p_{23}}{p_{22}p_{13}}\right) \\ &= \frac{1}{2}\left(\log p_{11} - \log p_{13} - \log p_{21} + \log p_{23}\right) \qquad (2) \\ & \qquad (3) \end{aligned}$$

i.e. averaging the 2 log-odds ratios, with variance

$$\mathrm{var}\widehat{\eta} = \frac{1}{4}\left(\frac{1}{n_1 p_{11}} + \frac{1}{n_1 p_{13}} + \frac{1}{n_2 p_{21}} + \frac{1}{n_2 p_{23}}\right) \tag{4}$$

However: for low odds ratios get very low p_{23},, $\Rightarrow$ the second term has high variance.

Modified additive model estimator

Use weighted average:

$$\widehat{\eta}_{opt} = c_1 \log \frac{p_{11}p_{22}}{p_{12}p_{21}} + c_2 \log \frac{p_{12}p_{23}}{p_{22}p_{13}}, \qquad c_1 + c_2 = 1 \quad (5)$$

- choose c_1, c_2 optimally

- maximise 'non-centrality parameter' $\Rightarrow$ maximise power

General model

- 2 separate odds ratio estimates (aa vs. Aa and Aa vs. AA)

$$\eta = (\eta_1, \eta_2) \tag{6}$$

- η_1, η_2 not assumed equal $\Rightarrow$ 2 d.f. test

- Bayes factor $\leftrightarrow$ test statistic relationship from Savage-Dickey formula as in 1 d.f. case

General model

- Critical value:

$$\chi_c^2 = 2\frac{n+a}{n}\log(\frac{n+a}{a}B) \tag{7}$$

- Compute power from non-central χ_2^2.

General model

Non-centrality parameter given by:

$$V_{n_1,n_2} = \frac{1}{n_1}\begin{pmatrix} \frac{1}{p_{11}} + \frac{1}{p_{12}} & \frac{-1}{p_{12}} \\ \frac{-1}{p_{12}} & \frac{1}{p_{12}} + \frac{1}{p_{13}} \end{pmatrix} + \frac{1}{n_2}\begin{pmatrix} \frac{1}{p_{21}} + \frac{1}{p_{22}} & \frac{-1}{p_{22}} \\ \frac{-1}{p_{22}} & \frac{1}{p_{22}} + \frac{1}{p_{23}} \end{pmatrix} \tag{8}$$

$$\text{ncp} = \eta' V_{n_1,n_2}^{-1} \eta \tag{9}$$

R package `ccDesign` – INSTALLATION

```
$  R CMD INSTALL ccDesign_1.0-2.tar.gz
* installing to library    /home/rod/R/2.12/library'
*␣installing␣*source*␣package␣  c c D e s i g n  ␣...
**␣R
**␣inst
**␣preparing␣package␣for␣lazy␣loading
**␣help
***␣installing␣help␣indices
**␣building␣package␣indices␣...
**␣testing␣if␣installed␣package␣can␣be␣loaded

*␣DONE␣(ccDesign)
```

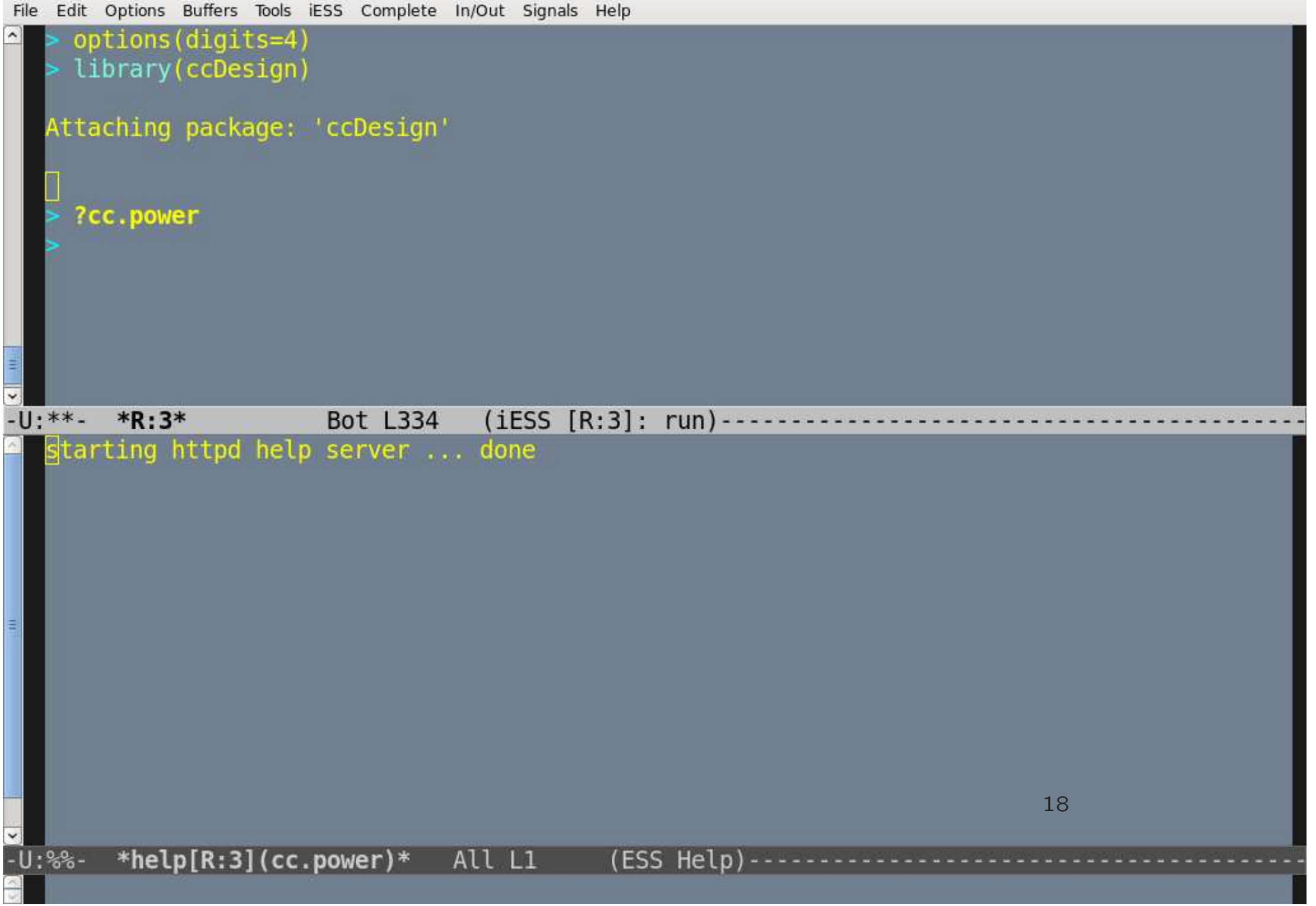
Can also install from within R (once package is on CRAN).

```
File  Edit  Options  Buffers  Tools  iESS  Complete  In/Out  Signals  Help
> options(width=132)
> library(help=ccDesign)

                Information on package 'ccDesign'

Description:

Package:        ccDesign
Version:        1.0-2
Title:          Design of experiments for detection of linkage disequilibrium in genome-wide case-control studies
Author:         Rod Ball <rod.ball@scionresearch.com>
Maintainer:     Rod Ball <rod.ball@scionresearch.com>
Description:    R package for design of experiments for genome-wide case-control studies. Determines sample size or
                power for detecting associations with a given Bayes factor. The Bayes factor should be chosen large
                enough to give respectable posterior odds. This requires Bayes factors of the order of 10^6 in
                genome-wide case-control studies where prior odds are low. Sample sizes needed to get this strength of
                evidence are substantially higher than those from traditional power calculations. The corresponding
                threshold for p-values is substantially lower than commonly used.
Depends:
License:        GPL version 2 or newer
URL:            mailto:rod.ball@scionresearch.com www.scionresearch.com/
Packaged:       2011-09-02 04:56:03 UTC; rod
Built:          R 2.12.0; ; 2011-10-20 01:51:27 UTC; unix

Index:

cc.power                Function to determine power of case-control study to
                        detect linkage disequilibrium with a functional locus
                        with a given Bayes factor.
cc.design               Function to determine sample size (number of cases and
                        controls) necessary to detect a locus with a given Bayes
                        factor with given power.
calc.B.ABF              Function to calculate approximate Bayes factor from Z-statistic
                        with sampling distribution N(0,1/n).
calc.Zc.ABF             Function to calculate critical value, Zc for given Bayes factor,
                        prior, and sample parameters.
calc.Zalpha.ABF         Function to calculate Z value corresponding to significance
                        level alpha.
calc.alphaB.ABF         Calculate the alpha value corresponding to a given Bayes factor
                        for given prior and sample parameters.
gpc.power               Function to get power of case control study to detect linkage
                        disequilibrium with a functional locus with a given significance
                        level. Uses the 'Genetic Power Calculator' website.
```

R package `ccDesign`

Two main functions

```
cc.power()    : compute the power for given sample size,
                design parameters
cc.design()   : compute the sample size (n.cases,
                n.controls) for given power
```


File Edit Options Buffers Tools iESS Complete In/Out Signals Help
> options(digits=4)
> library(ccDesign)
Attaching package: 'ccDesign'
> ?cc.power
>
-U:**- *R:3* Bot L334 (iESS [R:3]: run)
starting httpd help server ... done
-U:%%- *help[R:3](cc.power)* All L1 (ESS Help)

File Edit View History Bookmarks Tools Help

http://127.0.0.1:24672/library/ccDesign/html/cc.design.html

R: Functions for design of experi...

cc.design {ccDesign} R Documentation

Functions for design of experiments to detect linkage disequilibrium in genome-wide case-control studies

Description

Find the sample sizes (number of cases and number of controls) required to detect linkage disequilibrium with a given Bayes factor, with a given power, or find the power of experimental designs to detect linkage equilibrium with a given Bayes factor.

Usage

```
cc.design(B, OR, D, p, q, power, baseline.risk, Dprime=NULL, R=NULL,
          prevalence=NULL, n.cases, n.controls, model=c("additive",
          "dominant","recessive","general"), a=1, sigma2.eta=NULL, verbose=FALSE,
          amalgamate.cells=FALSE, pmin=0.1, pmax=0.99, ninterp=20, print.power.curve=TRUE)
cc.power(B, OR, D, p, q, baseline.risk, Dprime=NULL, R=NULL, prevalence=NULL,
         n.cases, n.controls, model=c("additive","dominant","recessive","general"),
         a=1, sigma2.eta=NULL, verbose=FALSE, amalgamate.cells=FALSE,show.attributes=FALSE)
```

Arguments

B	Bayes factor
OR	Odds ratio
D	Linkage disequilibrium coefficient
p	Bi-allelic marker allele frequency
q	Bi-allelic QTL allele frequency (for risk allele)
power	`cc.design`: Power, or probability of detecting an effect with Bayes factor greater than B
baseline.risk	Baseline risk, i.e. the probability of being a case for a genotype with no risk alleles
Dprime	D': i.e. linkage disequilibrium as proportion of the maximum (or minimum, if negative); need to give `D` _or_ `Dprime`
R	Relative risk: in the additive model: relative risk per copy of the risk allele, in dominant or recessive models relative risk for the high risk genotype(s) compared with the low risk genotype(s) in the 2x2 contingency table; need to give `R` _or_ `OR`.
prevalence	disease prevalence, i.e. the probability of being a case in the population; need to give `prevalence` _or_ `baseline.risk`

Find: Previous Next Highlight all Match case

R package ccDesign, cc.power() arguments

```
> args(cc.power)
function (B, OR = NULL, D, p, q, baseline.risk, Dprime = NULL,
    R = NULL, prevalence = NULL, n.cases, n.controls,
    model = c("additive","dominant","recessive","general"),
    a = 1, sigma2.eta = NULL, verbose = FALSE,
    amalgamate.cells = FALSE, show.attributes = FALSE)
```

```
B                    : Bayes factor
D or Dprime          : LD coefficient
p, q                 : marker and QTL allele frequencies
OR or R              :  odds ratio or relative risk
baseline.risk        :  baseline risk or prevalence
  or prevalence
a or sigma2.eta      : prior variance
```

```
File  Edit  Options  Buffers  Tools  iESS  Complete  In/Out  Signals  Help
R version 2.12.0 (2010-10-15)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
.
.
.
Type 'q()' to quit R.

[Previously saved workspace restored]

> options(STERM='iESS', editor='emacslient')
> options(digits=4)
> library(ccDesign)

Attaching package: 'ccDesign'


> cc.power(B=1e6, OR=1.5, D=0.25, p=0.5, q=0.5, baseline.risk=0.1, n.cases=1000,
+          n.controls=2000, model="additive", a=1)
[1] 0.9993
> cc.power(B=1e6, OR=c(1.5,1.5), D=0.25, p=0.5, q=0.5, baseline.risk=0.1, n.cases=1000,
+          n.controls=2000, model="general", a=1)
[1] 0.8125
>
```



```
-U:**-  *R:3*          Bot L332    (iESS [R:3]: run)----------------------------------------
```

R package `ccDesign`

`cc.design()` function: calculate the sample size needed for a given power.

- Similar arguments to `cc.power()`, also give `power`.
- Input values `n.cases, n.controls` used as initial values and to establish `case:control` ratio.

```
> args(cc.design)
function (B, OR, D, p, q, power, baseline.risk, Dprime = NULL,
          R = NULL, prevalence = NULL, n.cases, n.controls,
          model = c("additive","dominant","recessive","general"),
          a = 1, sigma2.eta = NULL, verbose = FALSE,
          amalgamate.cells = FALSE, pmin = 0.1, pmax = 0.99,
          ninterp = 20, print.power.curve = TRUE)
```

```
> # OR=1.5, 50% MAF, additive model, 100%LD
> cc.design(B=1e6, OR=1.5, D=0.25,p=0.5,q=0.5, baseline.risk=0.1,
+           n.cases=1000,  n.controls=2000, model="additive", a=1,
+           power=0.8, ninterp=12)
Power curve:
      n.controls n.cases power
 [1,]        502     251 0.100
 [2,]        554     277 0.145
 [3,]        611     306 0.206
 [4,]        674     337 0.283
 [5,]        744     372 0.377
 [6,]        821     411 0.485
 [7,]        906     453 0.599
 [8,]       1000     500 0.710
 [9,]       1103     552 0.809
[10,]       1218     609 0.887
[11,]       1344     672 0.941
[12,]       1483     742 0.974
[13,]       1637     819 0.990
1093 controls and 547 cases for power 0.8
         n n.controls    n.cases
      1640       1093        547
```

```
> # OR=1.2, 5% MAF, additive model, 100% LD
> cc.design(B=1e6,OR=1.2,Dprime=1,p=0.05,q=0.05,baseline.risk=0.1,
+           n.cases=1000,n.controls=1000,model="additive",a=1,
+           power=0.8,ninterp=12)
Power curve:
      n.controls n.cases power
 [1,]      13697   13697 0.100
 [2,]      15067   15067 0.146
 [3,]      16574   16574 0.207
 [4,]      18232   18232 0.285
 [5,]      20055   20055 0.380
 [6,]      22061   22061 0.488
 [7,]      24268   24268 0.602
 [8,]      26695   26695 0.713
 [9,]      29365   29365 0.811
[10,]      32302   32302 0.888
[11,]      35533   35533 0.942
[12,]      39087   39087 0.974
[13,]      42996   42996 0.990
29017 controls and 29017 cases for power 0.8
         n n.controls    n.cases
     58034      29017      29017
```

```
> # 5% MAF, general model, 100% LD
> cc.design(B=1e6,OR=c(1.2,1.2),Dprime=1,p=0.05,q=0.05,
+           baseline.risk=0.1,n.cases=1000,n.controls=1000,
+           model="general",a=1,power=0.8,ninterp=12)
Power curve:
      n.controls n.cases power
 [1,]      18585   18585 0.100
 [2,]      20256   20256 0.147
 [3,]      22078   22078 0.210
 [4,]      24064   24064 0.290
 [5,]      26228   26228 0.386
 [6,]      28586   28586 0.495
 [7,]      31157   31157 0.610
 [8,]      33959   33959 0.720
 [9,]      37013   37013 0.816
[10,]      40341   40341 0.891
[11,]      43969   43969 0.943
[12,]      47923   47923 0.974
[13,]      52233   52233 0.990
36431 controls and 36431 cases for power 0.8
         n n.controls    n.cases
     72862      36431      36431
```

```
> # 5% MAF, additive model, 50% LD
> cc.design(B=1e6,OR=1.2,Dprime=0.5,p=0.05,q=0.05,
+           baseline.risk=0.1,n.cases=1000,n.controls=1000,
           model="additive",a=1,power=0.8,ninterp=12)
Power curve:
      n.controls n.cases power
 [1,]      53148   53148 0.100
 [2,]      58454   58454 0.146
 [3,]      64288   64288 0.207
 [4,]      70705   70705 0.285
 [5,]      77763   77763 0.380
 [6,]      85525   85525 0.488
 [7,]      94062   94062 0.603
 [8,]     103450  103450 0.714
 [9,]     113777  113777 0.811
[10,]     125133  125133 0.888
[11,]     137624  137624 0.942
[12,]     151361  151361 0.974
[13,]     166469  166469 0.990
112415 controls and 112415 cases for power 0.8
         n n.controls    n.cases
    224830     112415     112415
```

```
> # 5% MAF, p=2q
> cc.design(B=1e6,OR=1.2,Dprime=1,p=0.1,q=0.05,baseline.risk=0.1,
+           n.cases=1000,n.controls=1000,model="additive",a=1,
+           power=0.8,ninterp=12)
Power curve:
      n.controls n.cases power
 [1,]      26012   26012 0.100
 [2,]      28686   28686 0.145
 [3,]      31635   31635 0.206
 [4,]      34887   34887 0.283
 [5,]      38474   38474 0.378
 [6,]      42429   42429 0.485
 [7,]      46791   46791 0.600
 [8,]      51601   51601 0.711
 [9,]      56906   56906 0.809
[10,]      62756   62756 0.887
[11,]      69207   69207 0.941
[12,]      76322   76322 0.974
[13,]      84168   84168 0.990
56328 controls and 56328 cases for power 0.8
         n n.controls    n.cases
    112656      56328      56328
```

```
> # 5% MAF, recessive model
> cc.design(B=1e6,OR=1.2,Dprime=1,p=0.05,q=0.05,baseline.risk=0.1,
+           n.cases=1000,n.controls=1000,model="recessive",a=1,
+           power=0.8,ninterp=12)
Power curve:
      n.controls n.cases power
 [1,]     603395  603395 0.100
 [2,]     660075  660075 0.147
 [3,]     722080  722080 0.209
 [4,]     789910  789910 0.289
 [5,]     864111  864111 0.385
 [6,]     945282  945282 0.494
 [7,]    1034079 1034079 0.608
 [8,]    1131216 1131216 0.719
 [9,]    1237479 1237479 0.815
[10,]    1353723 1353723 0.891
[11,]    1480886 1480886 0.943
[12,]    1619996 1619996 0.974
[13,]    1772172 1772172 0.990
1218546 controls and 1218546 cases for power 0.8
         n n.controls    n.cases
   2437092    1218546    1218546
```

Java program

- takes input file
- do many designs in one run
- optional columns to specify D or D', baseline risk or relative risk, odds ratio(s) or relative risk(s)
- summer eResearch proposal with BestGrid for GUI development

```
$ cat design8.in
    B     D    p      q    OR  OR2 prevalence  model n_cases n_controls  power  a
  1e6   0.1  0.3    0.2 1.72  1.5    0.13   general    2000       2000    0.8  1
  1e6   0.1  0.3    0.2 1.72  1.5    0.13  additive    3000       3000    0.8  1
```

```
$ ./ccdesign_java design8.in
called with 1 argument(s)
design8.in:
using input data file: design8.in
File headers:
    B     D    p      q    OR  OR2 prevalence  model n_cases n_controls  power  a
Data line:
[ 1] 1e6 0.1 0.3  0.2 1.72  1.5     0.13   general   2000       2000    0.8  1
GENERAL model inputs:
B = 1.00000e+06, Dprime =  0.71429, D =  0.10000, p =  0.30000, q =  0.20000,
OR1 =  1.72, OR2 =  1.50, R1 =  1.59867, R2 =  1.38402, q0 =  0.10483,
nu =  0.13000, a =  1.00000, sigma2_eta = -1.00000

Power curve:
  i   n_controls      n_cases        power
  0         1484         1484      0.10019
  1         1572         1572      0.12701
  2         1666         1666      0.15950
.
.
.
 13         3155         3155      0.80016
.
```

```
.
.
 20          4737          4737      0.99001
GENERAL model summary
n_controls =     2000, n_cases =     2000, power =  0.30122
    3155 controls and     3155 cases for power  0.80016

Data line:
[ 2] 1e6 0.1 0.3  0.2 1.72 1.5      0.13  additive   3000        3000    0.8  1

additive model, solving for relative risk

q =  0.20000, OR =  1.71920, prevalence =  0.13000, model = ADDITIVE
baseline_risk =  0.10364, relative_risk =  1.59995
ADDITIVE model inputs:
B = 1.00000e+06, Dprime =  0.71, D =  0.10, p =  0.30, q =  0.20,
OR =  1.72, R =  1.60, q0 =  0.10, nu =  0.13, a =  1.00,
sigma2_eta = -1.00

Power curve:
  i   n_controls      n_cases        power
  0          1186         1186      0.10011
  1          1259         1259      0.12532
  2          1337         1337      0.15585
.
.
.
 13          2585         2585      0.78998
```

```
 14          2745          2745       0.84244
.
.
.
 20          3932          3932       0.99000
ADDITIVE model summary
n_controls =     3000, n_cases =     3000, power =  0.90474
    2614 controls and     2614 cases for power  0.80030
```

Results summary ($\mathrm{OR} = 1.2, \mathrm{MAF} = 0.05$**):**

- $n = 58000$ for additive model at max LD, ($p = q = 0.05, D' = 1$)
- approximately 25% higher sample size for general model
- approximately $2\times$ higher sample size for $p = 2q, q = 0.05, D' = 1$
- approximately $42\times$ higher sample size for recessive model

Testing

- Many combinations tested

- Compared with Genetic Power calculator
  - `gpc.power()` function calls GPC from R and gets results
  - input appropriate α threshold corresponding to Bayes factor
  - similar non-centrality parameters and power (as expected) for dominant, recessive models

Next steps

- GUI for java program?
- R package vignette (in progress).
- upload to CRAN
- 2-stage designs?

Conclusions

- WTCCC sample size adequate to detect OR $= 1.5$ in ideal conditions, *not* OR $= 1.2, \text{MAF} = 0.05$ as claimed.

- loose power: lower OR, $D' < 1$, $p > q$, recessive model

- much higher sample sizes for OR $= 1.2, \text{MAF} = 0.05$ especially in recessive model

- General model requires $\sim 25\%$ higher sample size than additive when true model is additive

Conclusions

- larger B, n, needed to robustly detect effects e.g. $B \sim 10^6$
- many putative associations not robustly detected
- even some quite large effects likely to be undetected
- explaining the 'dark matter in the genome' ?

ccDesign: the genomic power calculator

VISG Virtual Institute of Statistical Genetics

Acknowledgements

Thanks to Gail Timmerman-Vaughan for testing and feedback on the R package.

This project was funded by the New Zealand Foundation for Research, Science, and Technology through a grant to the New Zealand Virtual Institute for Statistical Genetics (VISG).

University of Otago New Zealand · Plant & Food Research · SCION Next generation biomaterials · ViaLactia · ESR · agresearch · The University of Auckland New Zealand