VISG Virtual Institute of Statistical Genetics

Experimental design for genome-wide case-control studies

Roderick D. Ball, Scion†, P.B. 3020, Rotorua, New Zealand. rod.ball@scionresearch.com

SUMMARY: A Bayesian method for design of experiments for genome-wide case-control studies is given. A closed-form expression for an approximate Bayes factor is used to determine sample sizes required for power to detect effects with a given Bayes factor, B, where B is chosen sufficiently large to obtained respectable posterior odds, as needed to robustly detect effects. Dominant, recessive and additive models are considered for SNP markers in linkage disequilibrium with a putative causal locus. Power calculations are illustrated with respect to a typical scenario and recent genome-wide association studies (GWAS). We conclude that even current thresholds of $P < 5 \times 10^{-8}$ may not be sufficient and that much larger sample sizes are needed to robustly detect effects in the range of effect sizes and minor allele frequencies previously considered detectable.

KEYWORDS: Experimental design, power, Bayes factor, genome-wide association studies, case-control studies, linkage disequilibrium, association tests, QTL, QTN, genome scan.

Introduction

Genome-wide association studies require strong evidence for marker-trait associations due to the need to overcome low prior odds for any given marker to be in sufficiently high LD with a practically significant causal effect.

Despite initial optimism, many published genomic associations were spurious (Altshuler *et al.* 2000; Emahazion *et al.* 2001; Terwilliger and Weiss 1998; Ball 2007). More recent studies had much larger sample sizes and used more stringent thresholds: *e.g.* WTCCC (2007), had $n = 2000$ cases and $n = 3000$ controls and a threshold of $P < 5 \times 10^{-7}$. Our Bayesian calculations (Ball 2008) showed that of 24 WTCCC effects with $P < 5 \times 10^{-7}$ about half (those effects near the threshold) correspond to posterior odds less than 1. The required threshold continues to decrease and is currently $P < 5 \times 10^{-8}$, yet even this may not be sufficient. What threshold to use is critical yet the optimal threshold in any situation is not clear. At the root of the problem is that fact that a low p-value does not necessarily correspond to strong evidence that an effect is real (Berger and Berry 1988).

To overcome these problems we propose designing experiments with power to detect effects with a sufficiently large Bayes factor to obtain respectable posterior odds. We extend our previous method for quantitative traits (Ball 2004, 2005), to case-control studies.

Genome-wide case-control studies

- interest in complex diseases
- sample of cases and controls
- additive, dominant, or recessive model
- estimate odds ratio or relative risk ⇒ need strong evidence
- many small effects
- large sample sizes, many SNPs ($p \ll n$ problem)
- low prior odds

p-values, Bayes factors, and strength of evidence

- No relationship between p and $\Pr(H_1 \mid y)$ indept. of sample size, expt. design etc.
- Direct interpretation of Bayes factor as strength of evidence:

$$B = \frac{\Pr(y \mid H_1)}{\Pr(y \mid H_0)} \tag{1}$$

$$\text{posterior-odds} = B \times \text{prior-odds} \tag{2}$$

- smaller p needed for given B with increasing sample size

⇒ use B to specify strength of evidence

⇒ design experiments with power to obtain sufficiently large B

A typical scenario

- genome scan with 500k SNPs, ⇒ prior odds 1:50000
- 10 causal loci expected,
- posterior odds desired 20:1
- extent of LD $\sim 60kb$, ⇒ Bayes factor required 10^6.

Case-control studies, dominant model

Table 1. Observed counts and expected proportions for cases and controls by genotype class in the dominant model.

	observed counts		expected proportions	
	aa	Aa or AA	aa	Aa or AA
Case	n_{11}	n_{12}	p_{11}	p_{12}
Control	n_{21}	n_{22}	p_{21}	p_{22}

Test statistics and variance

$$\hat{\eta} = \log\left(\frac{n_{11}n_{22}}{n_{12}n_{21}}\right) = \log\hat{p}_{11} - \log\hat{p}_{12} - \log\hat{p}_{21} + \log\hat{p}_{22} \tag{3}$$

$$\sigma_1^2 = \frac{1}{c}\left(\frac{1}{p_{11}} + \frac{1}{p_{12}}\right) + \frac{1}{1-c}\left(\frac{1}{p_{21}} + \frac{1}{p_{22}}\right) \tag{4}$$

$$Z_n = \frac{\hat{\eta}}{\hat{\sigma}_1} \sim N\left(\frac{\eta}{\sigma_1}, \frac{1}{n}\right) \qquad \text{(sampling distn.)} \tag{5}$$

Approximate Bayes factor and power calculation

Asymptotic approximation to Bayes factor (*cf.* Ball Jan. 2007, Wakefield Jul. 2007):

$$\pi(z) \sim N\left(0, \frac{1}{a}\right) \qquad \text{(prior} \sim a \text{ sample pts.)} \tag{6}$$

$$g(z \mid Z_n) \cong N\left(\frac{n}{n+a}Z_n, \frac{1}{n+a}\right) \qquad \text{(Bayes' theorem)} \tag{7}$$

$$B = \frac{\pi(z=0)}{g(z=0 \mid Z_n)} \qquad \text{(Savage–Dickey)} \tag{8}$$

$$B \approx \frac{\sqrt{a}}{\sqrt{n+a}} \exp\left(\frac{n^2 Z_n^2}{2(n+a)}\right) \tag{9}$$

The power ($\mathcal{P}$) is given by:

$$Z_B^2 = \frac{n+a}{n^2}\log\left(\frac{n+a}{a}B^2\right) \tag{10}$$

$$\mathcal{P} \approx \Pr(|Z_n| > Z_B \mid \eta, a, n, \nu) \approx 1 - \Phi\left(\sqrt{n}\left(Z_B - \frac{\eta}{\sigma_1}\right)\right) \tag{11}$$

Incorporating LD

- test statistic at marker locus as above
- partial derivs. to adjust prior at trait locus to prior at marker locus
- minor allele frequencies (MAF) p, q at marker, causal locus (resp.)
- case-control genotype frequencies p_{ij}, q_{ij} at marker, causal locus (resp.)

$p_{ij} = f_{ij}(q_{ij}, D)$ *e.g.* (dominant model):

$$p_{11} = \frac{q_{11}\,(D + (1-p)\,(1-q))^2}{(1-q)^2} + \frac{q_{12}\left(2\,((1-p)\,q - D)\,(D + (1-p)\,(1-q)) + ((1-p)\,q - D)^2\right)}{q^2 + 2\,(1-q)\,q}, \tag{12}$$

$$p_{12} = \frac{q_{11}}{(1-q)^2} \times \left[2\,(p\,(1-q) - D)\,(D + (1-p)\,(1-q)) + (p\,(1-q) - D)^2\right] + \frac{q_{12}}{q^2 + 2\,(1-q)\,q} \times \Big[(D + p\,q)^2 + 2\,(p\,(1-q) - D)\,(D + p\,q) + (D + (1-p)\,(1-q))\,(D + p\,q) + 2\,((1-p)\,q - D)\,(D + p\,q) + (p\,(1-q) - D)\,((1-p)\,q - D)\Big], \tag{13}$$

$$p_{21} = \frac{q_{21}\,(D + (1-p)\,(1-q))^2}{(1-q)^2} + \frac{q_{22}\left(2\,((1-p)\,q - D)\,(D + (1-p)\,(1-q)) + ((1-p)\,q - D)^2\right)}{q^2 + 2\,(1-q)\,q}, \tag{14}$$

$$p_{22} = \frac{q_{21}}{(1-q)^2} \times \left[2\,(p\,(1-q) - D)\,(D + (1-p)\,(1-q)) + (p\,(1-q) - D)^2\right] + \frac{q_{22}}{q^2 + 2\,(1-q)\,q} \times \Big[(D + p\,q)^2 + 2\,(p\,(1-q) - D)\,(D + p\,q) + (D + (1-p)\,(1-q))\,(D + p\,q) + 2\,((1-p)\,q - D)\,(D + p\,q) + (p\,(1-q) - D)\,((1-p)\,q - D)\Big] \tag{15}$$

q_{ij} in terms of q, q_0 (baseline risk), ν (prevalence). *e.g.* :

$$q_{11} = \frac{(1-q)^2\,(1-q_0)}{1-\nu}, \tag{16}$$

$$q_{12} = \frac{\left(q^2 + 2\,(1-q)\,q\right)\,(1-q_0\,r)}{1-\nu}, \tag{17}$$

$$q_{21} = \frac{(1-q)^2\,q_0}{\nu}, \tag{18}$$

$$q_{22} = \frac{\left(q^2 + 2\,(1-q)\,q\right)\,q_0\,r}{\nu} \tag{19}$$

Results

Table 2. Bayes factors ($B_a, a = 1$) corresponding to various thresholds, in case-control studies with $n = 5000$ ($n_{cases} = 2000$, $n_{controls} = 3000$) or $n = 50000$ ($n_{cases} = 20000$, $n_{controls} = 30000$).

	α $(n = 5000)$			α $(n = 50000)$		
	5×10^{-7}	5×10^{-8}	1.5×10^{-9}	5×10^{-7}	5×10^{-8}	1.5×10^{-9}
B	4460	41900	$1.29 \times 10^{6*}$	1370	12700	386000

*$B > 10^6$

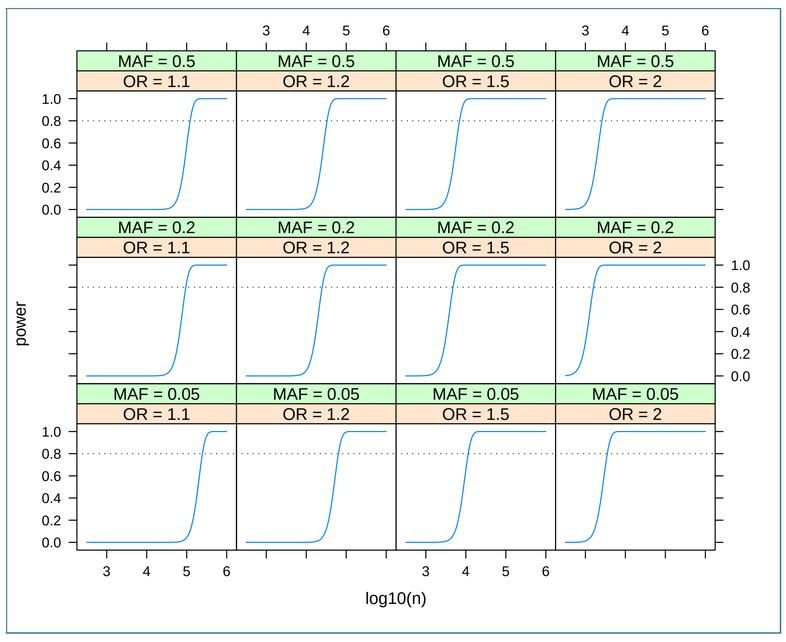

Figure 1: Dominant model, power for $B = 1,000,000$ versus $\log_{10} n$.

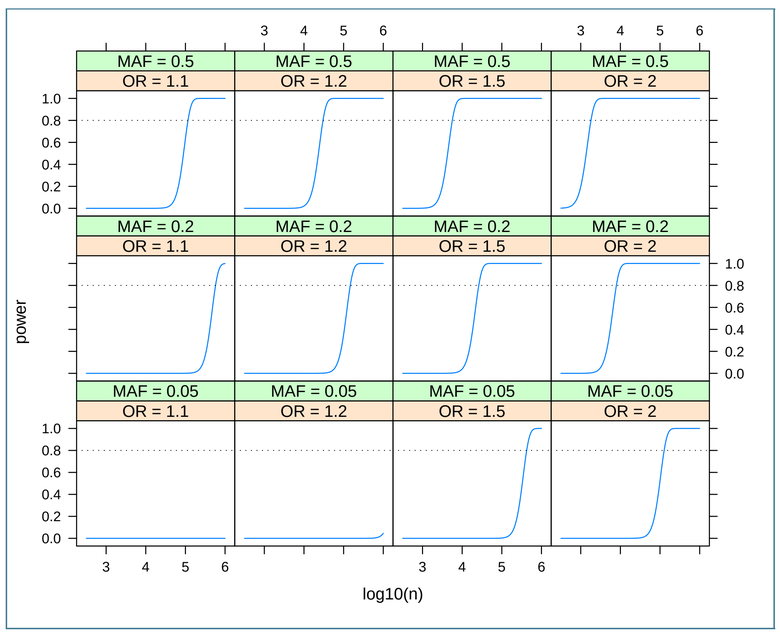

Figure 2: Recessive model, power for $B = 1,000,000$ versus $\log_{10} n$.

Table 3. Minimum odds ratio detectable with power $\mathcal{P} = 0.8$ to obtain a Bayes factor $B \geq 10^3, 10^6$ in case-control studies with $n_{cases} = 2000$, $n_{controls} = 3000$†.

		$B = 1000$			$B = 10^6$		
q	p/q	$D' = 0.5$	$D' = 0.8$	$D' = 1$	$D' = 0.5$	$D' = 0.8$	$D' = 1$
0.5	1.00	1.54	1.329	1.27	1.71	1.414	1.34
0.2	1.00	1.96	1.567	1.45	2.24	1.714	1.56
0.1	1.00	3.12	2.193	1.93	*	2.594	2.22
0.5	0.50	2.38	1.767	1.60	2.85	2.016	1.77
0.2	0.50	*	2.479	2.13	*	3.068	2.53
0.1	0.50	*	*	*	*	*	*

*min. OR > 3.5; † additive model.

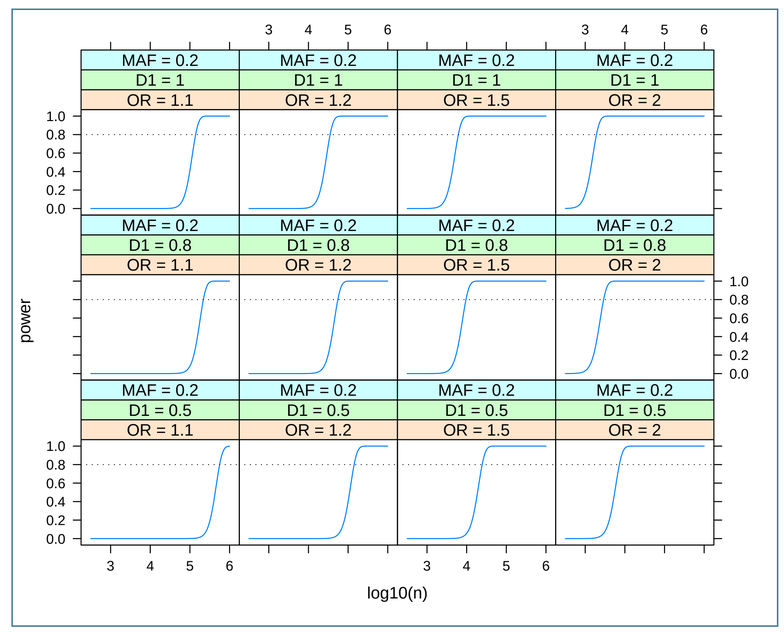

Figure 3: Additive model, power for $B = 1,000,000$ versus $\log_{10} n$.

Conclusions

- Bayes factor B of the order of 10^6, needed for good posterior odds in common scenarios.
- Current experiments underpowered to detect effects with this level of evidence.
- $P < 5 \times 10^{-7} \leftrightarrow B \sim 4400 \quad (a = 1)$ not sufficient for good posterior odds. This value *twice* would be sufficient. $P < 5 \times 10^{-8} \leftrightarrow B \sim 42000$. Still not sufficient. $P < 1.5 \times 10^{-9}$ sufficient for the typical scenario with conservative prior assumptions.
- Effects not robustly detected if power low (*cf.* Visscher 2008), and detected effects tend to be over-estimated (selection bias) (*cf.* Ball 2001).
- Power very low if MAF $= 0.05$ or OR $= 1.2$.
- Power low for OR < 2.0, MAF < 0.2 under the recessive model. A significant proportion of moderate to large size effects may be undetected.
- Maximum LD and hence power, is reduced if marker and trait frequencies (p, q) are disparate (*cf.* Yang *et al.* 2010)
- Failure of GWAS to explain much of the genetic variation may be due to failure of the "common disease common variant" hypothesis, but may also be due to lack of power.
- Sample sizes as high as $n = 10^6$ needed for good power to detect some of the effects claimed.

References

Altshuler, D., J. N. Hirschhorn, ..., L. Groop, and E. S. Lander 2000: The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nature Genetics 26: 76–80.

Ball, R. D. 2001: Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian Information Criterion. Genetics **159:** 1351–1364.

Ball, R. D. 2004: ldDesign—Design of experiments for detection of linkage disequilibrium. http://cran.r-project.org/src/contrib/Descriptions/ldDesign.html

Ball, R. D. 2005: Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. Genetics 170: 859–873.

Ball, R. D. 2007 (Jan.): Statistical analysis and experimental design Chapter 8, In: Association mapping in plants. N. C. Oraguzie *et al.* editors, Springer Verlag, ISBN 0387358447. (69pp)

Ball, R. D. 2008: Case studies in association mapping—frequentist and Bayesian measures of evidence. Poster presentation, International Society for Bayesian analysis conference, Hamilton Island, July 21–25, 2008; and New Zealand Statistical Association annual conference, University of Waikato, Hamilton, New Zealand, Sept 1–2, 2008.

Berger, J. O. and Berry, D. A. 1988: Statistical analysis and the illusion of objectivity. American Scientist. 159-165.

De Silva, H. N. and Ball, R. D. 2007: "Linkage disequilibrium mapping concepts", Chapter 7, In: Association mapping in plants. N. C. Oraguzie *et al.* editors, Springer Verlag, ISBN 0387358447. (31pp)

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University and Novartis Institutes for biomedical research 2007: Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331–1345.

Dickey, J. M. 1971: The weighted likelihood ratio, linear hypotheses on normal location parameters. Ann. Math. Statist. 42: 204–223.

Dudbridge, F. and Gusnanto, A. 2008: Estimation of significance thresholds for genomewide association scans. Genetic Epidemiology 32:227–234.

Emahazion, T., Feuk, L., Jobs, M., Sawyer, S. L., Fredman, D., *et al.* 2001: SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. Trends Genet. 17: 407-413.

Terwilliger, J. D., and K. M. Weiss 1998: Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr. Opin. Biotechnol. 9: 578–594.

The Wellcome Trust case Control Consortium (WTCCC) 2007: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.

Visscher, P. M. 2008: Sizing up human height variation. Nature Genetics 40: 489–490.

Wakefield, J. 2007 (Aug.): A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am. J. Hum. Genet. 81: 208–227.

Yang, J. Beben, B., McEvoy, D. P., Gordon, S.Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. 2010: Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42: 565–569, 608.

Acknowledgements

This work was funded by the New Zealand Foundation for Research Science and Technology through a contract with the Virtual Institute for Statistical Genetics (VISG). The author would like to thank Dr. Tony Merriman and Professor Peter Visscher for useful comments on this work.

For further information contact: R. D. Ball, Scion† P. B. 3020, Rotorua, New Zealand. Email: rod.ball@scionresearch.com. URL: www.scionresearch.com.

†Scion is a trading name of the New Zealand Forest Research Institute Limited.