Molecular Psychiatry (2012) 17, 108–114
© 2012 Macmillan Publishers Limited All rights reserved 1359-4184/12
www.nature.com/mp

ORIGINAL ARTICLE

The cost of large numbers of hypothesis tests on power, effect size and sample size

LC Lazzeroni[1,2] and A Ray[1]

[1]*Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA and* [2]*Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA*

Advances in high-throughput biology and computer science are driving an exponential increase in the number of hypothesis tests in genomics and other scientific disciplines. Studies using current genotyping platforms frequently include a million or more tests. In addition to the monetary cost, this increase imposes a statistical cost owing to the multiple testing corrections needed to avoid large numbers of false-positive results. To safeguard against the resulting loss of power, some have suggested sample sizes on the order of tens of thousands that can be impractical for many diseases or may lower the quality of phenotypic measurements. This study examines the relationship between the number of tests on the one hand and power, detectable effect size or required sample size on the other. We show that once the number of tests is large, power can be maintained at a constant level, with comparatively small increases in the effect size or sample size. For example at the 0.05 significance level, a 13% increase in sample size is needed to maintain 80% power for ten million tests compared with one million tests, whereas a 70% increase in sample size is needed for 10 tests compared with a single test. Relative costs are less when measured by increases in the detectable effect size. We provide an interactive Excel calculator to compute power, effect size or sample size when comparing study designs or genome platforms involving different numbers of hypothesis tests. The results are reassuring in an era of extreme multiple testing.

Molecular Psychiatry (2012) **17,** 108–114; doi:10.1038/mp.2010.117; published online 9 November 2010

Keywords: GWAS; large-scale data; multiple testing; sample size calculator; simultaneous inference; statistical power

Introduction

The numbers of hypothesis tests in science and genomics, in particular, are increasing at an ever-expanding rate. Total studies and hypothesis tests per study have both increased exponentially since the 1920s when the conventional 0.05 significance level was first adopted.[1,2] Simultaneously, as technological advances have provided the means to easily measure, store and manipulate huge quantities of data, the need for stronger *a priori* testing of one hypothesis has become more complex to justify. With almost inevitable large numbers of hypothesis tests in a single experiment comes the well-recognized need to use some type of statistical correction for multiple testing to avoid generating ever-increasing numbers of false-positive results.[3,4,5] The more stringent level of evidence required necessarily reduces the power to identify a true-positive finding. As a consequence, some experts now advocate larger and larger sample sizes for genomic studies[6,7] that are impractical for many diseases and, even when practical, may require broader, more heterogeneous phenotype definitions and less costly, more imprecise phenotypic measurements to accomplish.

Current genotype microarray technology is now approaching a capacity of five million single-nucleotide polymorphisms (SNPs) per study.[8] As whole genome sequencing becomes widely available, the number of tests will continue to increase.[9] Investigation of a variety of phenotype definitions, genetic models and subsets of individuals also increases the total number of hypothesis tests actually underlying any reported finding. Other fields (for example, neuroimaging or data mining of medical records) face a similar burden of multiplicity of tests.[10,11] Cross-disciplinary studies, such as genomic analyses of neuroimaging or other high-dimensional phenotypes, will only further expand the problem. There are reasonable arguments as to when it is appropriate to statistically adjust for multiple tests and how best to do so.[12–14] The universe of tests subject to adjustment, the level of stringency, underlying assumptions and statistical methods are all subjects of debate.

Whatever adjustment is used, extensive multiple testing inevitably leads to some loss of power or the

Correspondence: Dr L Lazzeroni, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Room 3100, 401 Quarry Road, Stanford, CA 94305-5723, USA.
E-mail: lazzeroni@stanford.edu

Received 16 August 2010; accepted 28 September 2010; published online 9 November 2010

need for a compensatory increase in the detectable effect size and/or sample size. This study examines the statistical 'cost' defined in terms of power, targeted effect size or sample size imposed by large and ever-increasing numbers of hypothesis tests, for the broad class of tests based on statistical estimates that follow a Normal distribution in large samples: (1) We develop a novel formulation of this relationship and (2) discuss its implications for genomic and other studies in the era of high-throughput technology. (3) Explicit numerical results and a user-friendly calculator for study design and comparison of alternative high-throughput technologies are also provided.

Methods

This study extends and expands the basic principles of power analysis found in introductory biostatistics texts. We consider the class of hypothesis tests based on a statistic, B, such that the standardized z-statistic, $Z = (B - b)/(s/\sqrt{N})$, has a standard Normal distribution in large samples, where b is the expected value of B and $s/\sqrt{N}$ is its standard error in a sample of size N. Owing to the Central Limit Theorem[15] and its extensions, this situation encompasses most commonly used statistical tests. For example, B might be a coefficient in a regression model, a sample proportion, a difference of two means, the log of an odds ratio (OR) or another maximum likelihood estimate. One-degree-of-freedom χ^2 tests are also covered owing to the relationship between the Normal and χ^2 distributions.

Suppose we want to use B to test whether its expected value $b = 0$. Let α be the significance level or Type I error probability and β be the Type II error probability, the probability that the null hypothesis is accepted for an alternative value of $b \neq 0$. By definition, the power of the test is 1–β. A single two-sided hypothesis test is illustrated in Figure 1a. The unit of measurement for the x axis is the standard error for a sample of size N. The curve on the left, centered at zero, is the sampling distribution of B under the null hypothesis in large samples (that is, B has a Normal distribution with mean 0 and standard error $s/\sqrt{N}$). The critical values are $\pm C$ standard errors, where C depends on α and is taken from the cumulative Normal distribution. For $\alpha = 0.05$, $C = 1.96$. The null hypothesis will be rejected if B is more than 1.96 standard errors from zero (that is, in either shaded region). The curve on the right centered at b is the sampling distribution of B under the alternative hypothesis that the expected value of $B = b$. (Throughout, we use 'effect size' to refer to the expected value of B under the alternative. Our results may not hold for other definitions that are used in the literature.) The power to reject the null hypothesis is the percentage of area under the alternative distribution inside the bold-outlined area above the upper critical value C. The distance between the critical value C and the effect size b, denoted by D, is also a multiple

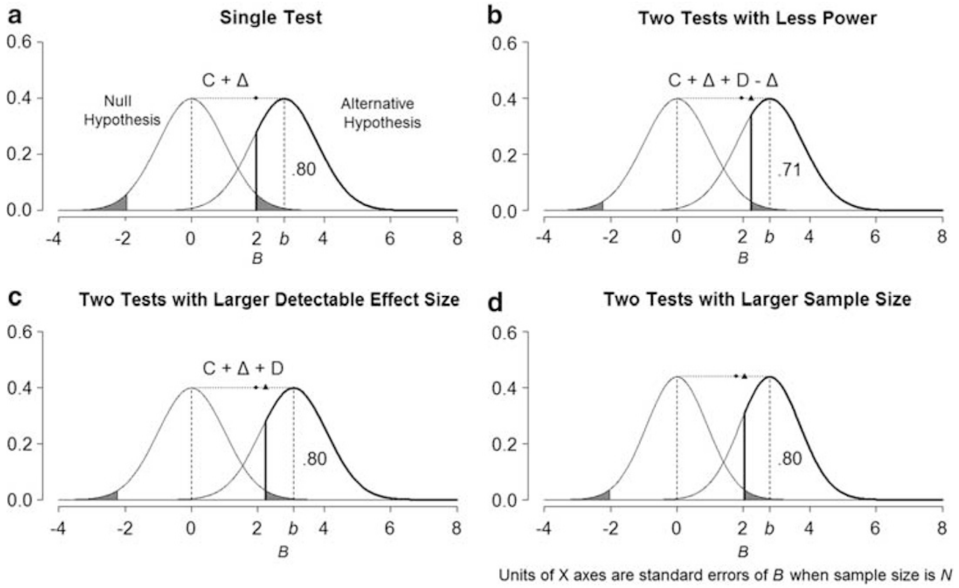

Figure 1 Distribution of B under the null and alternative hypotheses illustrating the power analysis for a single test (**a**) and two tests (**b–d**). (**a**) A single 0.05 significance test uses critical value 1.96 (circle on horizontal line). It has 80% power (area outlined in bold) to detect an effect of size $b = 2.80$. (**b**) For two tests, the critical value is increased by $\Delta = 0.28$ standard errors to 2.24 (triangle on horizontal line). Distance, D, from the mean of the alternative distribution, is reduced by Δ and the power (outlined in bold) is reduced to 71% at the original effect size. (**c**) To accommodate the increased critical value, the alternative curve can also be shifted to the right by Δ standard errors to maintain 80% power at a larger effect size $b = 1.96 + 0.28 + 0.84 = 3.08$. (**d**) Finally, the sample size can be increased to make the densities narrower and taller. This reduces the overlap between the null and the alternative, maintaining 80% power at original effect size 2.80.

of the standard error obtained from the Normal distribution. For 80% power, $D = 0.84$. Thus, 80% power is reached for a 0.05 significance test when the effect size is $(1.96 + 0.84)\times(s/\sqrt{N})$ or 2.80 standard errors away from zero. For 90% power, $D = 1.28$ and the effect size is 3.24 standard errors. Formally, the effect size is $(C + D)\times(s/\sqrt{N})$, where C is the $(1 - \alpha/2)\times$ 100th percentile of the Normal distribution and D is the $(1 - \beta)\times$100th percentile of the Normal distribution.

To evaluate the statistical cost of multiple tests, we extend the analysis to $H > 1$ hypothesis tests with a Bonferroni adjustment.[16] This conservative approach sets the per-test significance level to α/H, which guarantees for an experiment comprised of H tests that the probability of one or more false positives (known as the family-wise error rate) is no more than α. This is true for any possible relationship among the tests. (See Discussion for implications for other less conservative multiple-testing corrections.) For $\alpha = 0.05$ and $H = 2$, the Bonferroni critical value is 2.24, measured in standard errors. Let $\Delta = 2.24–1.96 = 0.28$ denote the increase in the critical value. Increasing the critical value reduces the bold-outlined area under the alternative distribution, reducing power from 80 to 71% (Figure 1b). Note that the cost of going from a single test with 80% power to two Bonferroni-adjusted tests is always a reduction of power to 71% if the initial effect size and sample size are maintained. The number does not depend on the type of test, statistical model, effect size or sample size.

The cost of multiple tests can also be quantified in terms of the effect size that can be detected with the original, single-test power and sample size. The new detectable effect size is found by moving the alternative hypothesis to the left by Δ units (Figure 1c) to accommodate the increase in the critical value. For $H = 2$, the minimum detectable effect size at 80% power is therefore $C + \Delta + D = 1.96 + 0.28 + 0.84 = 3.08$ standard errors. In general, the detectable effect size for H tests relative to that for a single test is

$$\frac{C + \Delta + D}{C + D} = 1 + \frac{\Delta}{C + D} \qquad (1)$$

which depends only on H, α and β and not other aspects of the test, the sample size or the original effect size. Note that the impact of H is only through the critical value and not power, as power is evaluated on a per-test basis.

A third way to quantify the cost of multiple tests is the increase in sample size needed to maintain the original level of power at the original effect size. If the sample size is increased by a factor m, it has the effect of dividing the standard error by $\sqrt{m}$, narrowing the widths of both the null and alternative curves accordingly (Figure 1d). To offset the addition of Δ, the distance between the two curves expressed in terms of the new standard error $(s/\sqrt{mN})$ for the larger sample size and H tests should equal the distance expressed in the original standard errors $(s/\sqrt{N})$ for a single test. Thus, m should satisfy

$$(C + \Delta + D)\times\frac{s}{\sqrt{m\times N}} = (C + D)\times\frac{s}{\sqrt{N}}$$

or

$$m = \left(1 + \frac{\Delta}{C + D}\right)^2. \qquad (2)$$

For two tests, $\alpha = 0.05$ and 80% power, the sample size should be multiplied by $m = 1.2$ to maintain the original power at the original effect size. Again, m depends only on H, α and β and not the effect size or original sample size. Note that the sample size multiplier is the square of the effect size multiplier.

More generally, suppose a study design with H tests at unadjusted significance level α and power $1-\beta$ is to be compared to a second study design with H^* tests at unadjusted significance level α^* and power $1-\beta^*$, then the detectable effect size for the second design relative to the effect size of the first design is

$$\frac{C^* + \Delta^* + D^*}{C + \Delta + D} \qquad (3)$$

where the numbers in the denominator are based on the initial design and the numbers in the numerator are based on the second. Furthermore, a sample size multiplier for the second design relative to the first is given by

$$\left(\frac{C^* + \Delta^* + D^*}{C + \Delta + D}\right)^2 \qquad (4)$$

Numerical results were calculated using the Normal distribution and expressions (1) and (2) in R.[17] All tests are two tailed and, if not otherwise specified, at significance level $\alpha = 0.05$. Logs, unless otherwise indicated, are base 10. A Cost of Multiple Tests calculator that implements equations (1)–(4) in Microsoft Excel 2003 for any choices of H, α and β is also provided (Supplementary Table 1).

Results

For a fixed targeted effect size and fixed sample size, power decreases as the number of tests and corresponding critical value increase (Table 1, Figure 2a). If the power for a single test is 80%, the power is approximately 50% for 10; 10% for 1000; and 1% for 100 000 Bonferroni-adjusted tests. To avoid a drop in nominal power without increasing sample size, an investigator may target larger effect sizes (Table 1, Figure 2b) using equation (1). For one million tests, the effect size multiplier is 2.25 at 80% power and 2.08 at 90% power. Suppose it has been determined that a sample of 100 yields 80% power to detect a difference in group means when the mean in group 1 is 10 and the mean in group 2 is 12. The original sample size of 100 would also yield 80% power to detect a mean difference of $2.25\times 2 = 4.50$, with Bonferroni adjustment for one million tests. The effect size multiplier works on the scale of the underlying Normally distributed test statistic. For

Table 1 Cost of multiple tests for 80 and 90% power and $\alpha = 0.05$

Number of tests (H)	*Critical value* ($C + \Delta$)	*Power for a single test*					
		80%			*90%*		
			Multipliers			*Multipliers*	
		Power (%)	*Effect size*	*Sample size*	*Power (%)*	*Effect size*	*Sample size*
1	1.96	80.0	1.00	1.000	90.0	1.00	1.000
5	2.58	58.8	1.22	1.488	74.6	1.19	1.416
10	2.81	49.7	1.30	1.690	66.7	1.26	1.588
50	3.29	31.3	1.47	2.161	48.1	1.41	1.988
100	3.48	24.9	1.54	2.372	40.6	1.47	2.161
500	3.89	13.8	1.69	2.856	25.8	1.60	2.560
1K	4.06	10.4	1.75	3.062	20.7	1.65	2.722
5K	4.42	5.3	1.88	3.534	11.9	1.76	3.098
10K[a]	4.56	3.9	1.93	3.725	9.4	1.80	3.240
50K	4.89	1.8	2.05	4.202	5.0	1.90	3.610
100K	5.03	1.3	2.10	4.410	3.7	1.95	3.802
300K[a]	5.23	0.8	2.17	4.709	2.3	2.01	4.040
500K	5.33	0.6	2.20	4.840	1.8	2.04	4.162
560K[a]	5.35	0.5	2.21	4.884	1.7	2.05	4.202
1M	5.45	0.4	2.25	5.062	1.4	2.08	4.326
1.2M[a]	5.48	0.4	2.26	5.108	1.3	2.09	4.368
1.8M[a]	5.55	0.3	2.28	5.198	1.0	2.11	4.452
2.5M[a]	5.61	0.2	2.30	5.290	0.9	2.13	4.537
5M[a]	5.73	0.2	2.35	5.523	0.6	2.16	4.666
10M	5.85	0.1	2.39	5.712	0.5	2.20	4.840
50M	6.11	0.0	2.48	6.150	0.2	2.28	5.198
100M	6.22	0.0	2.52	6.350	0.1	2.31	5.336
500M	6.47	0.0	2.61	6.812	0.1	2.39	5.712
1B	6.57	0.0	2.65	7.022	0.0	2.42	5.856
1T	7.53	0.0	2.99	8.940	0.0	2.72	7.398

Abbreviations: B, billion; K, thousand; M, million; T, trillion.
[a]Existing or proposed GWAS genotyping platform.

example, effect size multiplication for an OR should be carried out after conversion to the natural log scale. Suppose 500 total cases and controls provide 90% power to detect an OR = 1.2, the effect size multiplier for 90% power and 1000 tests is 1.65. Since $\log_e(1.2) = 0.18$, the same sample of 500 yields 90% power in 1000 tests to detect a $\log_e(\mathrm{OR}) = 1.65 \times 0.18 = 0.30$ or, equivalently, an $\mathrm{OR} = \exp(0.30) = 1.35$.

To compensate for a greater number of tests, a more realistic strategy may be to increase the sample size (equation 2). Table 1 and Figure 2c give sample sizes needed to maintain the original power at the original targeted effect size. For one million tests, the sample size multiplier is 5.06 for 80% power and 4.33 for 90% power, using equation (2). In the first example above, 506 subjects would be sufficient to reach 80% power to detect that the means differ by 2 in one million Bonferroni-adjusted tests. Although it might appear counterintuitive, the sample size multiplier is smaller for 90% power because the initial sample size is larger. In the same example, 132 subjects would be needed to reach 90% power for one test. For 90% power for one million tests, $4.33 \times 132 = 572$ subjects are needed. Noting the nearly linear relationship in Figure 2c, we also obtained an approximate rule-of-thumb for the sample size multiplier by fitting zero-intercept linear regression models to the results in Figure 2c. The estimated slopes show that m is approximately $1 + \gamma \times \log_{10} H$, where $\gamma = 1.2$ for 50% power, 0.68 for 80% power, 0.55 for 90% power and 0.38 for 99% power.

The rate at which the critical value and, consequently, the effect size and sample size multipliers increase becomes slower and slower as the number of tests becomes larger (Figure 2d), owing to the exponential decline in the tails of the Normal density. For example, effect size multipliers for one million vs ten million tests at 80% power are 2.25 and 2.39, respectively. Sample size multipliers are 5.06 and 5.71. At 80% power, ten million tests require only a 6% increase in the targeted effect size or a 13% increase in the sample size when compared to one million tests. In contrast, 10 tests require a 30% increase in the targeted effect size or a 70% increase in the sample size as compared with a single test. For

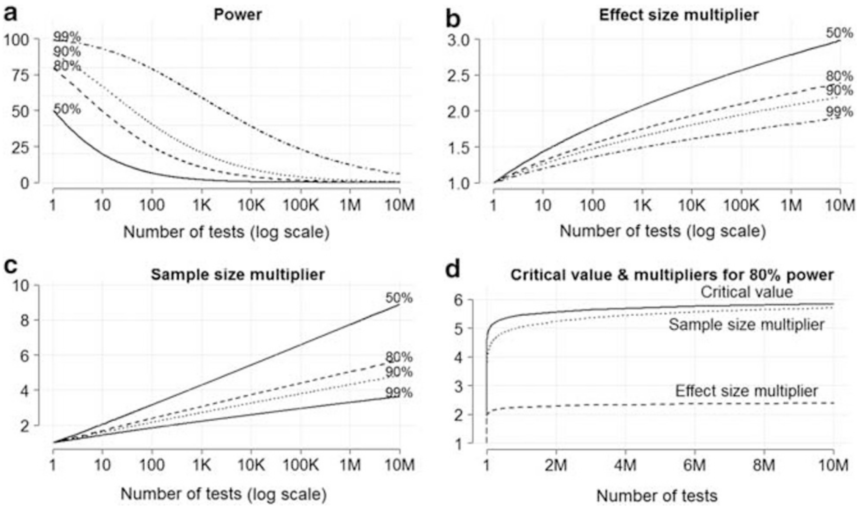

Figure 2 Power (**a**), effect size multiplier (**b**) and sample size multiplier (**c**) as a function of the number of tests on the log scale up to ten million tests, where power for a single test is 50, 80, 90 or 99%. The effect size multiplier is the number by which the effect size for a single test must be multiplied to maintain the same power for the same sample size at the specified number of tests. The sample size multiplier is the number by which the sample size for a single test must be multiplied to maintain the same power for the same effect size at the specified number of tests and is nearly linear with respect to the log of the number of tests. (**d**) Critical value, effect size multiplier and sample size multiplier for 80% power, with the number of tests on the raw (unlogged) scale. As the number of tests increases, the rate of increase in all three decreases dramatically.

80% power and one billion or one trillion tests, the required sample sizes, respectively, are approximately 7 or 9 times that needed for a single test. See Table 1 for some numerical results and the provided Excel calculator (Supplementary Table 1) to explore unreported results and specific study designs.

The sample size and effect size multipliers can be used in other ways. For example, different numbers of tests can be compared by taking the ratio of two sample size multipliers or the effect size multipliers. Consider alternatives to a genotyping platform with 1.2 million SNPs at 80% power. The detectable effect size for a larger array with 2.5 million SNPs would yield equivalent power for a 2% bigger effect size or, alternatively, a 4% bigger sample size. A smaller array with 560 000 SNPs would yield equivalent power for a 2% smaller effect size or a 4% smaller sample size. Following a data analysis involving multiple tests, the sample size multiplier can also be used to find the effective sample size for a single test *post hoc*. For example, suppose 1000 hypothesis tests have been carried out by permutation test in a sample of 120 subjects with negative results. The 1000 test sample size multiplier at 80% power is 3.062. Thus, the effective sample size is $120/3.062 = 39$.

Discussion

The observation that the relationship between the number of hypothesis tests and the detectable effect size or required sample size depends only on the choice of significance level and power covers most commonly used statistical tests. This relationship is independent of the original effect size and sample size and other specific characteristics of the data, the statistical model and test procedure. Our results show that most of the cost of multiple testing is incurred when the number of tests is relatively small. The impact of multiple tests on the detectable effect size or required sample size is surprisingly small when the overall number of tests is large. When the number of tests reaches extreme levels, on the order of a million, doubling or tripling the number of tests has an almost negligible effect. This is reassuring in light of continuing developments in methods of high-throughput data collection and the trend toward data mining and exploration of a variety of statistical models entailing greater numbers of tests and comparisons. In addition, we used our results to create a power, effect size and sample size calculator to facilitate the comparison of alternative large-scale study designs.

Underlying the results is the fact that the critical value in the context of extreme multiple testing is not affected much by typical changes in the number of tests. Thus, even a rigorous application of the Bonferroni correction that accounts for every one of a large number of tests is unlikely to change actual hypothesis test results. Consequently, multiple analyses of alternative genetic or statistical models impose minimal costs in a statistical sense in the context of a one million SNP genome-wide association study. Nonetheless, very large sample sizes are still required for large numbers of tests if, as many have proposed, it is necessary to target very small effect sizes that would require a sizable sample even for a single test.

Our results also show that, for the range of contemporary genotyping platforms, there is little difference with respect to detectable effect size or required sample size. At 80% power, the detectable effect size for a 2.5 million SNP platform is only 1.8% bigger than for 1.2 million tests and 4.1% bigger than for 560 000 tests. This relatively small increase is potentially offset by the fact that, for any risk-related variant in the genome, denser chips are more likely to include a tagging SNP with a larger effect size. Thus, consideration of alternative platforms should be primarily based on monetary costs and not the possibility of having 'too many tests.' Whole genome sequencing theoretically provides data on approximately three billion base pairs, but might reasonably be expected to generate no more than 100 million variant sites. The sample size required to maintain equivalent power for that many tests is only 30% greater than that required for a 560K SNP chip. Newer, denser SNP chips do include lower frequency variants than older chips, whereas whole genome sequencing theoretically enables identification of variants appearing at any frequency in the sample. The effect sizes of these lower frequency variants are likely to differ on average from those of more common polymorphisms, an issue that should also be considered when designing a study.

Our results also suggest that the Bonferroni correction may not be as conservative as sometimes thought, when compared with multiple testing methods based on the effective number of tests, either explicitly or implicitly through permutation testing or other means.[5,18–23] In a one million SNP genome-wide association study, a precise estimate of the 'effective number of independent tests' as, say, 900 000 would not much improve the detectable effect size over the Bonferroni correction and significant findings are likely to be the same. In practice, the effective number of independent tests will usually be larger. In this paper, we have not considered significance levels other than 0.05. For a very large number of tests, it may be appropriate to use a less stringent standard for family-wise error. Such cases can be addressed using Supplementary Table 1.

This study is limited in that the calculations are based on the Normal approximation and rely on its accuracy. Although this is usual for analytic discussions of power, this caveat should be considered before applying our formulae to small sample sizes. Our results have also not been shown to apply to hypothesis tests based on other statistical distributions, such as the F distribution or the χ^2 distribution with more than one degree of freedom. We have also followed common practice in ignoring any possible dependence of the standard error on the true value of *b*. Lastly, there is a practical limit to the size of data set that can be conveniently analyzed and interpreted even if samples can be generated that theoretically provide adequate power.

Our theoretical results do not address problems that can arise when conclusions are based on extremely small *P*-values. Extreme results sometimes reflect rare data errors, ascertainment biases, confounding or other problems of study design or implementation. Large sample sizes do not reduce bias; rather, large samples increase the chance of a false positive finding in the presence of bias.[24] Furthermore, a *P*-value is a random variable, subject to variation from sample to sample. When tiny *P*-values are required, only one or two differences in the sample data can change a result from true positive to false negative or false negative to true positive.

This study also assumes that samples are drawn from the same population and effect sizes are the same for different numbers of tests. These assumptions may not always be realistic. Larger samples may require less expensive data collection protocols to offset increased costs in time and money. Furthermore, subjects may be recruited from a wider variety of sources and be examined at different sites in order to increase their numbers. Broader inclusion/exclusion criteria may also be used. Such strategies can introduce greater heterogeneity and effectively reduce the average effect size in larger samples. Choices among alternative chip or sequencing technologies can also implicitly correspond to different effect sizes. For example, a denser SNP chip may have a SNP closer to or in stronger linkage disequilibrium with a causal variant. New, denser chips and sequencing also permit the assessment of lower frequency variants, which may individually account for less of the population variance of a trait, although they also can have stronger effects on individual risk if lower frequencies are the end result of evolutionary selection. None of these limitations are unique to this study.

In conclusion, we have shown that feasible sample sizes can address surprisingly large numbers of hypothesis tests, even when care is taken to control the number of false positives by using an appropriately stringent statistical correction. In fact, it might be theoretically possible to design a study that achieved a 'science-wide' multiple test correction for, say, one trillion tests. Despite this finding, large samples are still needed to target the smaller effect sizes that are often anticipated to be associated with genetic variation.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

We would like to thank Professor David L Braff, University of California, San Diego, for a thoughtful review of the manuscript. This work was funded and supported by the Consortium on the Genetics of Schizophrenia (R01MH086135, R01MH065571, R01MH065554, R01MH065707, R01MH065578, R01MH065558). LCL received additional support from the following: RC2HL101748, R01DA023063, R01MH073914, R01MH083972, Department of Veterans Affairs Cooperative Studies Program Study No.

478 and VISN-21 Mental Illness Research, Education and Clinical Center.

References

1 Stigler S. Fisher and the 5% level. *Chance* 2008; **21**: 12.
2 Fisher RA. *Statistical Methods for Research Workers*, 1st edn. Oliver & Boyd Ltd.: Edinburgh and London, 1925.
3 Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995; **310**: 170.
4 Shaffer JP. Multiple Hypothesis Testing. *Ann Rev Psych* 1995; **46**: 561–584.
5 Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of *P*-values for multiple correlated tests. *Am J Hum Genet* 2007; **81**: 1158–1168.
6 Todd JA. Statistical false positive or true disease pathway? *Nat Genet* 2006; **38**: 731–733.
7 Wang WYS, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005; **6**: 109–118.
8 Human Omni 5.0, Illumina GWAS roadmap. http://www.illumina.com/applications.ilmnwhole_genome_genotyping_and_copy_number_variation_analysis/.
9 Baker M. Next-generation sequencing: adjusting to data overload. *Nat Methods* 2010; **7**: 495–499.
10 Genovese CR, Lazar NA, Nichols T. Thresholding of Statistical Maps in functional neuroimaging using the false discovery rate. *NeuroImage* 2002; **15**: 870–878.
11 Benjamini Y, Draj D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* 2001; **125**: 279–284.
12 Bender R, Lange S. Adjusting for multiple testing—when and how? *J Clin Epidemiol* 2001; **54**: 343–349.
13 Veazie PJ. When to combine hypotheses and adjust for multiple tests. *Health Serv Res* 2006; **41**: 804–818.
14 Storey JD, Tibshirani R. Statistical significance for genome-wide studies. *Proc Natl Acad Sci* 2003; **100**: 9440–9445.
15 Billingsley P. *Probability and Measure*. John Wiley & Sons: New York, 1995.
16 Miller RG. *Simultaneous Statistical Inference*, Second edition. Springer Verlag: New York, 1981, pp 6–8.
17 R Development Core Team. 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria; ISBN 3-900051-07-0, http://www.R-project.org.
18 Cheverud JM. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* 2001; **87**: 52–58.
19 Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 2004; **74**: 765–769.
20 Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 2005; **95**: 221–227.
21 Westfall PH, Young SS. *Resampling-Based Multiple Testing*. John Wiley & Sons: New York, 1993.
22 Lazzeroni LC, Lange K. A Conditional Inference Framework for Extending the Transmission/Disequilibrium Test. *Hum Hered* 1998; **48**: 67–81.
23 Selwood SP, Parvathy S, Cordell B, Ryan HS, Oshidari F, Vincent V *et al.* Gene expression profile of the PDAPP mouse model for Alzheimer's disease with and without Apolipoprotein E. *Neurobiol Aging* 2009; **30**: 574–590.
24 Dow DJ, Lindsey N, Cairns NJ, Brayne C, Robinson D, Huppert FA *et al.* 2 macroglobulin polymorphism and Alzheimer disease risk in the UK. *Nat Genet* 1999; **22**: 16–17.

Supplementary Information accompanies the paper on the Molecular Psychiatry website (http://www.nature.com/mp)