Protein Cell 2011, 2(7): 519–522
DOI 10.1007/s13238-011-1059-5

Protein & Cell

MINI-REVIEW

Multiple phenotypes in genome-wide genetic mapping studies

Jurg Ott✉**, Jing Wang**

Key Laboratory of Mental Health Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
✉ Correspondence: ottjurg@psych.ac.cn

Received May 17, 2011 Accepted May 23, 2011

ABSTRACT

For many psychiatric and other traits, diagnoses are based on a number of different criteria or phenotypes. Rather than carrying out genetic analyses on the final diagnosis, it has been suggested that relevant phenotypes should be analyzed directly. We provide an overview of statistical methods for the joint analysis of multiple phenotypes in case-control association studies.

KEYWORDS multiple phenotypes, genetic mapping, case-control, statistical method

INTRODUCTION

For many complex traits, a good number of phenotypes (endophenotypes, covariates, and biological variables) are recorded and some of them flow into the definition of *affected* and *unaffected*. As a simple example, hypertension is defined by thresholds on two measurements, systolic and diastolic blood pressure. A more complicated example is schizophrenia, for which books (DSM-IV; SCID, http://www.scid4.org/) provide guidance on how to define this trait and other major mental disorders. Genetic mapping studies by linkage or association analysis may now be carried out on the basis of the dichotomy, *affected* (cases) versus *unaffected* (controls), but the reduction of dimensionality from multiple phenotypes to just one dimension may involve a loss of information and efforts have been made to base genetic studies on multiple relevant phenotypes. This outline summarizes statistical approaches to working with the multiplicity of phenotypes in a meaningful manner.

A telling example of the value of considering multiple phenotypes rather than a conventional disease definition is hypertension in Lyon hypertensive rats. It was shown that two different blood pressure measurements, diastolic and pulse pressure, are controlled by different genes on different chromosomes (Dubay et al., 1993). Thus, it appears likely that an analysis based on the conventional definition of "hypertension" might have missed both of these genes.

THE ONE-BY-ONE APPROACH

The simplest way of handling multiple phenotypes is to analyze each of them separately. Each is likely to shed light on a given trait from a somewhat different angle. Depending on sample size, we may expect that one or more of the phenotypes will result in a statistically significant outcome. In practical application, for a given (quantitative) phenotype and a given SNP with genotypes AA, AB, and BB, we want to test whether phenotype means are the same at the three SNP genotypes. This is usually carried out by a one-way analysis of variance (ANOVA) resulting in an F statistic with 2 and $n - 3$ degrees of freedom (df), where n is the number of observations (individuals). For multiple phenotypes, multiple F tests are carried out.

As each phenotype may only capture one aspect of the trait, the one-by-one approach presumably is suboptimal. There is also the question of how to handle the multiple p-values resulting from the analysis of multiple phenotypes; this question will be addressed below under the heading of *multiple testing*. Thus, various ways of analyzing multiple phenotypes in a combined manner have been proposed and applied.

PRINCIPAL COMPONENTS

In genetic linkage analysis on family pedigree data, various multivariate approaches to considering multiple phenotypes have been developed, particularly in the earlier literature. For example, a linear combination (a weighted sum) of phenotypes that maximizes the linkage to a marker locus was shown to be much more powerful than standard linkage analysis of one phenotype at a time (Allison et al., 1998). However, in this outline we focus on genetic association studies.

One approach of working with several phenotypes jointly has been to transform phenotypes into principal components (PCs). A principal component is a linear combination of variables, and principal components are constructed in such a way that the first PC extracts from the variables most of the variability, with the second PC having second largest variance, and so on. Furthermore, PCs are designed to be independent of each other. In principle, there are as many PCs as there are variables, but the main aim of constructing PCs is to work with only a small number of them so that together they account for, say, 80% of the variance in the data. That is, a relatively small number of PCs reflect most of the information in the data. This statistical technique is elegant and efficient but one must bear in mind that PCs are a statistical construct, which imposes a specific structure on the data. Nonetheless, PCs have been used for many years to condense phenotype information into a small number of variables (PCs), more recently also to condense *genotype* information for large numbers of single-nucleotide polymorphisms (SNPs) into a very small number of dimensions (Patterson et al., 2006; Price et al., 2006). As an aside, a common way of combining variables in social studies is to create a *scale*, that is, a collection of items combined into a single score (DeVellis, 2003). For example, a scale is often a simple sum of responses to questions and, thus, is similar to a principal component except that it is unweighted.

A potential drawback of PCs is that they combine variables purely on the basis of their variance. In genetics, however, it would be desirable to preferentially consider variables with high heritability. Thus, so-called principal components of heritability (PCH) have been derived (Ott and Rabinowitz, 1999; Klei et al., 2008). The first PCH is that linear combination of variables with the highest heritability, the second PCH has second-highest heritability, and so on. As appealing and intuitive as this concept is, it does not seem to have been applied much in practice.

MULTIVARIATE PHENOTYPE

Rather than condensing a multiplicity of phenotypes into a small number of PCs, one might consider using the phenotypes directly, without modification, if their number is not too large. Thus, a single, one-dimensional phenotype used in the one-by-one approach is replaced by a vector of phenotypes, and one (multivariate) analysis is carried out on this vector, that is, on all phenotypes. For example, in diabetes research, two important phenotypes, serum insulin release (I/G30) and insulin sensitivity (HOMA-IR), have been used as one bivariate phenotype to test for combined differences in means between genotypes (Holmkvist et al., 2009). Generally, testing for simultaneous differences in means of several phenotypes between three SNP genotypes would be carried out in a multivariate analysis of variance (MANOVA). This approach was used, for example, to test for simultaneous mean differences of personality traits among the three genotypes of an SNP in the promoter region of the DRD4 gene (Bookman et al., 2002).

An approximation to multivariate analysis has been proposed as follows (Manly, 2007). Consider genome-wide association testing involving large numbers of SNPs and a number k of phenotypes relevant for a complex trait. For a given SNP, association with each of the k phenotypes is tested and the largest test statistic is retained for this SNP, provided that all k test statistics have the same null distribution, for example, chi-square with 2 df. Alternatively, test results need to be converted to p-values and the smallest p-value is retained for the given SNP. This approach is carried out for each SNP, where the smallest p-value is retained each time at whatever phenotype this minimum p-value occurred. The single, genome-wide test statistic for the association with all phenotypes is then the overall smallest p-value, $p_{\min}$, occurring at any of the SNPs. How to interpret $p_{\min}$ will be described in the next section.

MULTIPLE TESTING

For the moment, consider multiple SNPs but only a single phenotype. At each SNP, an association test is carried out resulting in a nominal significance level, p_i, which is equal to the probability of obtaining a test result at least as extreme as the one found given no association. Often, a result is called significant when $p_i \leqslant 0.05$. With large numbers of SNPs, however, this criterion will furnish 5% "significant" results by chance alone. Thus, we need to make the criterion stricter and do this by defining a new significance level, α = probability that one or more test results are extreme just by chance (Zhang and Ott, 2009), and then want to keep $\alpha \leqslant 0.05$. If p is a given threshold for nominal significance and all m tests are independent then $\alpha = 1 - (1 - p)^m$. Thus, for a fixed small α, we need to set $p = 1 - (1 - \alpha)^{1/m} \approx \alpha/m$, which is known as the Bonferroni correction for multiple testing, and α is called the genome-wide or experiment-wise significance level.

Particularly for dependent tests, for example, association tests involving large numbers of SNPs, the Bonferroni correction is very conservative and reduces power although many researchers feel that a Bonferroni-corrected p-value represents the only legitimate claim for significance. As a possible solution, an alternative way of defining significance has been proposed, that is, the probability that a test result is false given it is significant. This quantity is known as the false discovery rate, FDR (Benjamini, 2010). Particularly FDR levels obtained by the Benjamini-Hochberg method (Benjamini et al., 2001) are easy to apply and well-known. However, we will not discuss the FDR here further.

As a potential solution to the problem of large numbers, *m*, of dependent tests, it has been proposed that one should find the corresponding number, *m*', of independent tests (*m*' < < *m*), and apply the Bonferroni correction with *m*' rather than *m* (Cheverud, 2001). While this is an appealing concept it turns out that computing *m*' as originally proposed is not straightforward but a simpler solution and corresponding software have been developed (Nyholt, 2004).

With *k* phenotypes and *m* SNPs, the total number of tests can be as high as *k* × *m*. At least the phenotypes are often strongly correlated, so subjecting that many tests to the Bonferroni correction tends to be rather conservative. A solution to this problem consists of randomization testing discussed in the next section.

RANDOMIZATION TESTS

Consider again an association test for a given phenotype with each of *m* SNPs, which may or may not be correlated. Define the largest test result, $T_{\max}$, as our single genome-wide test statistic. To assess the significance level associated with $T_{\max}$, we need to find the null distribution of $T_{\max}$, that is, its distribution given no association. The trick is now to approximate such a distribution on the computer by Monte Carlo (computer simulation) methods. All we need to do is to create a new dataset by randomly permuting the labels *case* and *control* but leaving genotypes untouched. Clearly, in such a dataset there cannot be any association between disease and genotypes. Then we compute the largest test statistic in this new dataset. We do this a large number of times so that the resulting largest test statistics approximate the distribution of $T_{\max}$ under no association. The proportion of permutated datasets with a largest test statistic at least as large as the observed $T_{\max}$ approximates the *p*-value associated with $T_{\max}$. Such a procedure is called a permutation (randomization) test. It elegantly furnishes unbiased estimates of *p*-values whose accuracy depends on the number of permutation samples. The only drawback of the procedure might be computing time. For example, for a somewhat accurate estimation of a genome-wide significance level of $p = 0.05$, more than 1000 randomization datasets are required (Table 1).

Table 1 Number, *m*', of independent tests and 95% confidence interval for a true significance level of $p = 0.05$ based on an estimated significance level of $\hat{p} = 0.05$. Based on the binomial distribution, B (0.05, *m*')

m'	Confidence interval
100	(0.016, 0.113)
1000	(0.037, 0.065)
5000	(0.044, 0.056)
10000	(0.046, 0.054)

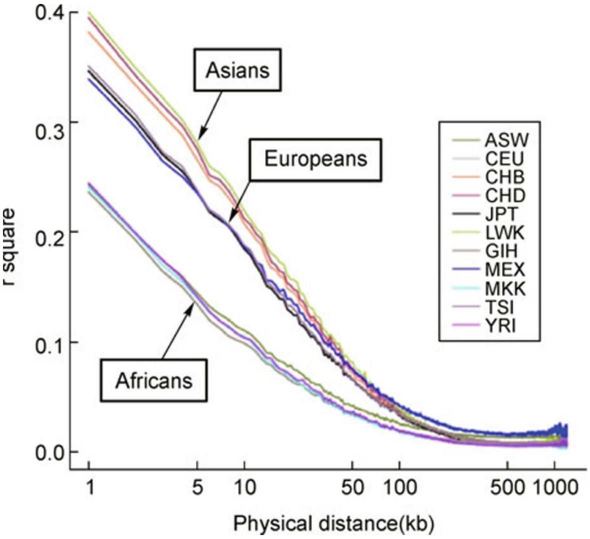

Figure 1. Decay of linkage disequilibrium (r^2) for pairs of SNPs as a function of their physical distance. The graphs show that background linkage disequilibrium is highest in Asians and lowest in Africans. Calculations carried out by Dr. Qingrun Zhang for SNPs on chromosome 22.

For multiple phenotypes observed on an individual, as mentioned above, we may define the smallest significance level, $p_{\min}$, over all phenotypes at any SNP as our genome-wide test statistic. An equivalent permutation-based smallest *p*-value is obtained in a good number of permutation datasets. Their proportion with smallest *p*-values at least as small as the observed $p_{\min}$ then represents the significance level associated with $p_{\min}$ (Manly, 2007).

Computer-based randomization tests have been implemented in a number of approaches (for example, Hoh and Ott, 2000; Hoh et al., 2001).

Randomization procedures as discussed here represent the most general approach to dealing with multiple SNPs and multiple phenotypes. For a single phenotype, the question has been raised what number *m*' of independent tests would be equivalent to a very large number of SNPs ($m \to \infty$), which, if dense enough, must be associated with each other. In human linkage analysis, such a question has previously been tackled (Lander and Kruglyak, 1995). For association analysis, it has been found that the equivalent number of independent SNPs is *m*' = 1,000,000 in Europeans and *m*' = 2,000,000 in Africans (Pe'er et al., 2008). Thus, in people of European origin, whenever a nominal significance level at a given SNP is smaller than 0.05/1,000,000 = 5×10^{-8} it can be considered significant. In Asians, as demonstrated in Fig. 1, linkage disequilibrium between markers tends to be even stronger than in people with European ancestry so that *m*' in Asians is expected to be smaller than 1,000,000. Consequently, the critical nominal significance level is increased and association results can be declared significant more easily than in individuals of European and African ancestry. However, these significance levels only apply to a single phenotype. With multiple phenotypes, the prudent solution is to apply permutation testing, which can guard against too many false positive claims of association.

DISCUSSION

Psychiatric traits like schizophrenia exhibit a multitude of characteristics that deviate from “normal.” The “endophenotype approach” in schizophrenia research attempts to make use of molecular biology and neurobiology to identify specific brain dysfunctions (Braff et al., 2007). Here we have provided an overview of statistical approaches to make simultaneous use of endophenotypes in the search for genetic risk factors underlying psychiatric traits. Our purpose has been to increase power and to decrease the occurrence of false positive results, with the latter being a particularly important problem in genetic association studies (Ott, 2004). We plan to implement relevant approaches in user-friendly computer programs and make them generally available to researchers.

ACKNOWLEDGEMENTS

Support from the National Natural Science Foundation of China (Grant No. 30730057) is gratefully acknowledged. We thank Dr. Qingrun Zhang for carrying out the calculations on linkage disequilibrium leading to Figure 1.

REFERENCES

Allison, D.B., Thiel, B., St Jean, P., Elston, R.C., Infante, M.C., and Schork, N.J. (1998). Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. Am J Hum Genet 63, 1190–1201.

Benjamini, Y. (2010). Discovering the false discovery rate. J R Stat Soc, B 72, 405–416.

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125, 279–284.

Bookman, E.B., Taylor, R.E., Adams-Campbell, L., and Kittles, R.A. (2002). DRD4 promoter SNPs and gender effects on Extraversion in African Americans. Mol Psychiatry 7, 786–789.

Braff, D.L., Freedman, R., Schork, N.J., and Gottesman, I.I. (2007). Deconstructing schizophrenia: an overview of the use of endophenotypes in order to understand a complex disorder. Schizophr Bull 33, 21–32.

Cheverud, J.M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87, 52–58.

DeVellis, R.F. (2003). Scale development: theory and applications, 2nd ed. Thousand Oaks, Calif.: Sage Publications, Inc. viii, 171.

Dubay, C., Vincent, M., Samani, N.J., Hilbert, P., Kaiser, M.A., Beressi, J.P., Kotelevtsev, Y., Beckmann, J.S., Soubrier, F., Sassard, J.,*et al.* (1993). Genetic determinants of diastolic and pulse pressure map to different loci in Lyon hypertensive rats. Nat Genet 3, 354–357.

Hoh, J., and Ott, J. (2000). Scan statistics to scan markers for susceptibility genes. Proc Natl Acad Sci U S A 97, 9615–9617.

Hoh, J., Wille, A., and Ott, J. (2001). Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res 11, 2115–2119.

Holmkvist, J., Banasik, K., Andersen, G., Unoki, H., Jensen, T.S., Pisinger, C., Borch-Johnsen, K., Sandbaek, A., Lauritzen, T., Brunak, S.,*et al.* (2009). The type 2 diabetes associated minor allele of rs2237895 KCNQ1 associates with reduced insulin release following an oral glucose load. PLoS One 4, e5872.

Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet Epidemiol 32, 9–19.

Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11, 241–247.

Manly, B.F.J. (2007). Randomization, bootstrap, and Monte Carlo methods in biology, 3rd ed. Boca Raton, FL: Chapman & Hall/CRC, 455

Nyholt, D.R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74, 765–769.

Ott, J. (2004). Association of genetic loci: Replication or not, that is the question. Neurology 63, 955–958.

Ott, J., and Rabinowitz, D. (1999). A principal-components approach based on heritability for combining phenotype information. Hum Hered 49, 106–111.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet 2, e190.

Pe’er, I., Yelensky, R., Altshuler, D., and Daly, M.J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol 32, 381–385.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904–909.

Zhang, Q., and Ott, J. (2009). Multiple Comparisons/Testing Issues. In: Handbook on Analyzing Human Genetic Data: Computational Approaches and Software. S. Lin, and H. Zhao, eds. Berlin: Springer. 277–287.