Journal of Proteomics & Bioinformatics - Open Access

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000159
JPB/Vol.3 Issue 12

Genetic Inheritance and Genome Wide Association Statistical Test Performance

Philip Cooley[1]*, Robert Clark[1], Ralph Folsom[1] and Grier Page[2]

[1]RTI International, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709
[2]RTI International Atlanta, Koger Center, Suite 119, 2951 Flowers Road South, Atlanta, Georgia, 30341-5533

Abstract

The choice of a statistical method significantly affects the power profiles of Genome Wide Association (GWA) predictions. Previous simulation studies of a single synthetic phenotype marker determined that the gene model or mode of inheritance (MOI) was a major influence on power. In this paper, the authors compare the power profiles of GWA statistical methods that combine MOI specific methods into multiple test scenarios against individual methods that may or not assume a MOI gene model consistent with the marker that predicts the association. Combining recessive, additive and dominant individual tests, and using either the Bonferroni Correction method or the MAX test (Li et al., 2008) has power implications with respect to single test GWA-based methods. If the gene model behind the associated phenotype is not known, a multiple test procedure could have significant advantages with respect to single test procedures.

Our findings do not provide a specific answer as to which statistical method is best. The best method depends on the MOI gene model associated with the phenotype (diagnosis) in question. However, our results do indicate that the common assumption that the MOI of the locus associated with the diagnosis is additive has consequences.

Our results indicate that researchers should consider a multi-test procedure that combines the results of individual MOI-based core tests as a statistical method for conducting the initial screen in a GW study. The process for combining the core tests into a single operational test can occur in a number of ways. We identify two: the Bonferroni procedures and the MAX procedure, each of which produce very similar statistical power profiles.

Introduction

In this study, we examine statistical methods used to perform genome wide association studies (GWAS). GWAS usually apply univariate statistical tests to each gene marker or single nucleotide polymorphism (SNP) as an initial step. This SNP based test is statistically straightforward and the testing is done with standard methods (e.g. χ^2 tests, regression) that have been studied outside of the GWAS context. A recent paper by Kuo and Feingold (2010) described the most commonly used methods, and the authors note that a compound procedure which combined two or more statistical tests is used.

The literature contains a number of papers that make statistical power comparisons among subsets of these methods, including Sasieni (1997) and Freidlin et al. (2002), but the question of which method is best suited to univariate scanning in a GWAS remains an open issue. The choice of method depends on the match between the true genetic model underpinning the association and the type of model assumed by the method.

We used a multiple test procedure that combines the most promising of the methods identified in the literature and apply them to a set of synthetic marker data with known properties. Our goal is to identify marker properties that can be linked to optimal methods (with reference to statistical power) for predicting associations in GWAS. We know from prior studies that the statistical procedure a researcher chooses influences GWAS prediction accuracy, and that there are specific properties of the underlying markers that determine the optimization of the procedure choice, see Kuo and Feingold (2010). We included the important properties that influence the association prediction accuracy into our synthetic marker data via a Monte Carlo simulation process, and we link the properties to the influencing marker to study their individual and collective contributions to association prediction. A synthetic marker dataset allows us to assess the performance of different statistical methods in a GWAS context. We apply a number of statistical methods to the simulated data and use their statistical power profiles to evaluate the performance of the methods. We also quantify the relationships between locus traits and prediction accuracy.

Our study identifies a number of these properties and quantifies the loss in power if non-optimal methods are used. Similar results have been reported in earlier studies by Sasieni (1997) and Freidlin et al. (2002). These studies reinforce the view that the major influence on prediction accuracy is the gene model of the locus associated with the diagnosis.

We are particularly interested in assessing the consequences of applying a statistical method that assumes an inherent additive mode of inheritance (MOI) property to SNP data that is non-additive. Our motivation for this is twofold. First, the additive MOI model is commonly employed in GWAS, and second, the answer to the question: "what statistical methods should be used to conduct GWA type studies?" does not have a definitive answer. The best method typically depends on what MOI gene model has been associated with the associated diagnosis.

Our results show the major factors that influence association predictions. They also indicate that a strategy based on predicting associations using multiple statistical methods can be more accurate

***Corresponding author:** Philip Cooley, RTI International, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709, Tel: (919) 541- 6509; Fax: (919) 541- 6178; E-mail: pcc@rti.org

Received November 19, 2010; **Accepted** December 15, 2010; **Published** December 17, 2010

Citation: Cooley P, Clark R, Folsom R, Page G (2010) Genetic Inheritance and Genome Wide Association Statistical Test Performance. J Proteomics Bioinform 3: 321-325. doi:10.4172/jpb.1000159

(much more accurate if the governing marker is recessive) than those that assume a single (additive) mode. The multiple test procedure proposed here uses a combination of the recessive, additive and dominant MOI-optimal statistical methods, all of which are derived from the well known Cochran-Armitage test. We also examined different procedures for combining the tests.

Methods

We examine the accuracy of association detection by generating synthetic data with properties that are known to influence statistical power. We used a Monte Carlo method to generate the data from a set of random variables described below. The main purpose of the synthetic data is to act as a "truth set" to assess the performance of commonly used statistical methods used in a GWAS context.

Generating the data

Our method for generating the synthetic marker data is derived from a study by Iles (2002) and is based on Mendelian concepts of inheritance. We include autosomal dominant and autosomal recessive patterns that are single gene inheritance patterns. We also incorporate additive and multiplicative inheritance patterns to represent the actions of multifactorial inheritance patterns.

A description of the data generating process begins from the notion of disease penetrance. Penetrance in genetics is the proportion of individuals carrying a particular variation of a gene (allele or genotype) that also express an associated trait. We designate A as the risk allele, and **a** as the allele without risk. By using the relationships between penetrance and relative risk as defined in Table 1, for different MOI categories, generating the synthetic gene dataset was straightforward. Specifically the steps were:

1. Preload the details that define the factor combinations per MOI category (we used 864 combinations for this study). The factors are:

1.1 n_j = the target number of cases and controls in a given experiment (100, 250, 500,1000, 2000,4000, 8000, 9500),

1.2 dp_j = the disease penetrance, (.3, .4, .5),

1.3 $ErrP_j$ = the misclassification error rate contained in the phenotype data, (0, 2, 5%),

1.4 $ErrG_j$ = the misclassification error rate contained in the genotype data, (0, 2, 5%),

1.5 Φ_j = the relative risk, (1.0, 1.15, 1.3, 1.45),

2. Draw a genotype distribution (at random from the master set of genotype distributions obtained from real distribution data (i.e., the Schymick et al. (2007) data). At this stage, Chan et al. (2009) recommends that a minor allele frequency (MAF) threshold not be applied. They argue that filtering MAFs out of the process because of low frequencies or to maintain Hardy–Weinberg equilibrium (HWE) deviation has little effect on the overall false positive rate and in some cases, filtering MAF only serves to exclude SNPs. The effect of this step is to select a specific genotype distribution (at random) from the master distribution.

3. Use Table 1 to assign a case (1) or a control (0) based on the selected genetic relative risk (GRR), penetrance (P) and MOI factors. This step converts the GRR ratio value into the probability that the case occurs for the MOI gene model of interest. This process can be represented by the following logic that was derived from Iles (2002):

Major Homozygote (aa): Assume that the aa genotype is selected. The probability of a case given this selection is equal to the disease penetrance P, or. $\Psi_{aa} = P$.

Minor Homozygote (AA) – liability increasing allele: Assume the aa genotype is selected. GRR can be expressed as a ratio of two probabilities: the probability of a case for a minor homozygote divided by the probability of a case for a major homozygote, i.e.,

$$\Psi_{AA} = \text{Prob(case/AA)} / \text{Prob(case/aa)} = x/P. \quad (1)$$

From (1) the probability of a case given the minor genotype $= x = \Psi_{AA} * P$ (2)

where Ψ_{AA} = one of the assigned risk factors and P is one of the assigned penetrance factors.

Heterozygote (aA): Assume the aA genotype is selected. By the same argument, the phenotype risk given a heterozygote is:

$$\Psi aA = \text{Prob(case/aA)} / \text{Prob(case/aa)} = y/P, \quad (3)$$

By the same arguments, the risk of a case given the heterozygote $= y = \Psi aA * P$ (4)

where ΨaA = one of the assigned risk factors and P is one of the assigned penetrance factors.

4. Using the estimate of x and y, assign a case or control at random using the four different MOI models in conjunction with equations (2) and (4) and Table 1 below. We assigned cases in proportion to x (y) and controls in proportion to 1- x (1-y) for the minor homozygote (heterozygote) genotypes respectively. For the MOI models that assume an elevated risk from the minor and the hetero genotypes, we would expect a higher proportion of cases to be more easily identified via the statistical procedures. The specification of risk depends on specific and unknown disease mechanisms. A relative risk of 1.7 is considered strong and is associated with positive replication, see Sladek et al. (2007), and a risk of 1.3 is considered by Ziegler et al. (2008) to be a realistic assumption for complex diseases. We limited our focus to a relative risk range of 1.15 to 1.45 and were particularly interested in scenarios with low relative risk.

5. Individuals are either assigned as cases or controls according to the probabilities given in Table 1.

6. Continue with the above process until *n1* cases and *n2* controls are generated (note in this example *n1* = *n2*, but that can be tailored to a specific set of *n1* – *n2* targets).

7. Apply a set of statistical methods to predict associations and record the results.

8. Generate 1,000 replicate experiments for each set of 3,456 factor combinations.

The simulated dataset that we generated has the following characteristics:

MOI	Major homozygote Ψ_{AA}	Minor homozygote[1] $\Psi_{aa} = \frac{\text{Pr(case/aa)}}{\text{Pr(case/AA)}}$	Heterozygote[2] $\Psi_{aA} = \frac{\text{Pr(case/aA)}}{\text{Pr(case/AA)}}$
Recessive	1	Φ	1
Dominant	1	Φ	Φ
Additive	1	2*Φ-1	Φ
Multiplicative	1	Φ *Φ	Φ

[1] Ψ_{aa} is the relative risk of homozygous minor to homozygous majors
[2] Ψ_{aa} is the relative risk of heterozygote to homozygous major

Table 1: Relative Risk assumptions by MOI, see Iles (2002).

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000159
JPB/Vol.3 Issue 12

Key Words	Frequency
Autosomal Dominant	3805
Autosomal Additive	12
Autosomal Multiplicative	21
Autosomal Recessive	3775

Table 2: Distribution of Genes in OMIM by MOI.

	AA	Aa	Aa	Total
Case	r0	r1	r2	R
Control	s0	s1	s2	S
Total	n0	n1	n2	N

AA – Major genotype
Aa – Heterozygote genotype
aa – Minor genotype

Table 3: Terms Defined in Equations (5) and (6).

N	CA-A	BON	MAX	CA-D	CA-R	2df-G
100	2.26	1.73	1.91	0.62	0.46	0.00
250	13.47	12.22	12.25	9.46	5.56	2.41
500	30.00	28.12	28.06	24.19	12.63	10.48
1000	49.91	48.31	48.26	45.72	25.11	25.21
2000	68.80	67.40	67.31	64.72	41.08	46.35
4000	83.00	81.86	81.94	80.25	58.24	65.96
8000	94.30	93.69	93.72	92.58	72.36	80.52
9500	96.07	95.63	95.46	95.01	75.66	83.90

Table 4: Power Results by Statistical Method and Number of Cases: Additive MOI Data.

- The proportion of cases (controls) that are major homozygotes = 50.3 (63.0)%.
- The proportion of cases (controls) that are heterozygotes = 39.2 (31.3)%.
- The proportion of cases (controls) that are minor homozygotes = 10.5 (5.7)%.
- With MOI distribution:
  - Recessive = 25%,
  - Dominant = 25%,
  - Additive = 25%, and
  - Multiplicative = 25%.

We acknowledge that the distribution of MOI traits above is not representative of how inheritance traits are distributed in humans. The Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/omim) provides the best source of information on the MOI distribution (Table 2). However, OMIM is disproportionally populated by genes linked to single Mendelian disorders. Therefore, genes associated with multifactorial disorders are under-represented in OMIM. Because polygene influences are assumed to be a major source of additive and multiplicative SNP behavior, the distribution in Table 2 is likely biased. Accordingly, we populated SNPs in our data with equal MOI representation and acknowledge that it is not representative of the true distribution.

The three "optimal" MOI specific methods are the three variations of the Cochran-Armitage (CA) trend test described in Zheng and Gastwirth (2006). We also included a fourth "individual" method, the 2df genotype test, which is a commonly used method. Using the notation in Table 3 to define the 2x3 table of case-control counts stratified by genotype , a test statistic ($T^2(x)$) for the three variations of the CA trend methods is defined as:

$$T^2(x) = n\,[\Sigma_{0,1,2}\ \{x_i\,(s\,r_i - r\,s_i)\}]^2 / [r\,s\,(\Sigma_{0,1,2}\ n\,\{x_i\,x_i\,n_i\} - \{\Sigma_{0,1,2}\ (x_i\,n_i)^2\})]. \quad (5)$$

The values represented in equations (5) and (6) are shown in Table 3 below and the value of x_i defines the specific test x0 = 0, x2 = 1 and x1 = {0 – recessive, .5 – additive, 1 – dominant).

Under the null hypothesis of no association, $T^2(x)$ has an asymptotic χ^2 distribution with 1 degree of freedom.

As an alternative to equation (5), it is also possible to use a normally distributed test statistic per Li et al. (2008):

$$N(x) = n^{1/2}\,[\Sigma_{0,1,2}\ \{x_i\,(s\,r_i - r\,s_i)\}] / [r\,s\,(n\,\Sigma_{0,1,2}\,\{x_i\,x_i\,n_i\} - [\Sigma_{0,1,2}\ \{x_i\,n_i\}^2]^{1/2}. \quad (6)$$

Under the null hypothesis of no association, N(x) has an asymptotic normal distribution N(0,1), which suggests a one-sided test because the synthetic data assumes that the minor allele conveys the risk of phenotype.

We use eight sample size assumptions with equal numbers of cases and controls to perform our analysis, with N defined as the number of cases = 100, 250, 500, 1,000, 2,000, 4,000, 8,000, or 9,500. We estimate statistical power by statistical method and N using a significance threshold of $\alpha = 10^{-7}$. In a GWA study, researchers usually perform a single marker analysis as a starting point to identify SNPs for additional and more comprehensive analysis. This initial pass creates a large number of statistical tests as well as a high potential for false-positive predictions, which has caused researchers to perceive the need for a very low threshold. Accordingly, recent studies have used type-I threshold levels on the order of 10^{-7} as in Iles (2002), Ziegler et al. (2008) and Van Es et al. (2008).

The multi-test statistical methods we use in our comparisons are:

1. The Bonferroni (BON) method, shown in Holm (1979). A simple form of the Bonferroni correction results when using n **methods to test for an association outcome. The correction involves dividing the alpha level by n**. For example, if the association of a given SNP involves using three different statistical methods, the corrected alpha level (α) would be $\alpha/3$. This would ensure that the overall chance of making a Type I error is still less than α.
2. A MAX method from Li et al. (2008) that departs from the Bonferroni method. Bonferroni assumes that the individual tests are mutually independent, while Li et al. (2008) assumes that the individual tests are correlated and incorporates an approximation of the joint distributions into the method.

Results

Results - statistical method assessment

Table 4 presents power estimates by statistical methods and sample size, and is based on a fixed alpha threshold ($\alpha = 10^{-7}$). All tests are one-sided and the tests included in this table are:

- The additive $\chi 2$ version of the CA (CA-A) test, which was the best method for both additive and multiplicative gene models but not particularly effective when applied to recessive MOI data.
- The Bonferroni test (BON), which uses (combines) the $\chi 2$ version of the recessive, additive, and dominant MOI specific tests (CA-R, CA-A, CA-D) and improves on the test performance of the three individual tests when the MOI gene model is not known.
- The MAX test due to Li et al. (2008) which uses (combines) the Normal version of the CA-R, CA-A, CA-D tests to improve on the test performance of the three individual tests if the MOI gene model is not known.
- The dominant χ^2 version of the CA test (CA-D), which was the best

N	CA-A	BON	MAX	CA-D	CA-R	2df-G
100	0.05	0.07	0.40	0.13	0.00	0.00
250	2.09	2.94	2.88	3.79	0.00	0.00
500	12.89	15.01	15.02	16.52	0.01	1.21
1000	32.75	34.75	34.84	36.94	0.19	12.79
2000	54.80	56.33	56.55	57.97	2.42	32.75
4000	71.01	73.48	73.46	74.90	11.24	54.50
8000	85.15	86.95	87.18	88.09	26.71	71.98
9500	88.27	90.15	90.10	91.23	31.55	76.39

Table 5: Power Results by Statistical Method: Dominant MOI Gene Data.

N	CA-A	BON	MAX	CA-D	CA-R	2df-G
100	0.00	0.00	0.38	0.00	0.00	0.00
250	0.02	0.11	0.68	0.00	0.19	0.00
500	0.42	1.70	1.77	0.00	2.11	0.01
1000	2.71	7.88	7.89	0.01	8.95	1.04
2000	9.97	19.91	19.99	0.13	21.40	6.50
4000	21.84	34.29	34.39	1.50	35.95	18.45
8000	34.94	49.35	49.67	7.17	51.01	33.14
9500	38.04	53.07	53.11	9.32	54.62	36.73

Table 6: Power Results by Statistical Method: Recessive MOI Gene Data.

N	CA-A	BON	MAX	CA-D	CA-R	2df-G
100	4.00	3.35	3.67	0.98	1.65	0.01
250	17.16	15.60	15.71	11.07	8.82	4.74
500	35.65	33.81	33.77	27.25	19.4	12.50
1000	53.84	52.08	51.78	48.11	33.65	30.72
2000	71.59	70.31	70.27	66.44	49.68	50.69
4000	85.12	84.03	83.96	81.40	64.32	67.66
8000	95.10	94.58	94.58	93.36	77.30	82.84
9500	96.36	95.99	96.08	95.30	80.27	86.42

Table 7: Power results by Statistical Method: Multiplicative MOI Gene Data.

N	CA-A (D)	MAX (D)	CA-A (R)	MAX (R)
100	0.05	0.40	0.00	0.38
250	1.45	2.88	0.02	0.68
500	12.89	15.02	0.42	1.77
1000	32.75	34.84	2.71	7.89
2000	54.80	56.55	9.97	19.99
4000	71.01	73.46	21.84	34.39
8000	85.15	87.18	34.94	49.67
9500	88.27	90.10	38.04	53.11

CA-A (D) – Cochran Armitage method – additive model version applied to dominant MOI SNP data
CA-A (R) – Cochran Armitage method – additive model version applied to recessive MOI SNP data

Table 8: Power results by CA-A and MAX Methods, for different MOI gene models.

method for the dominant gene models.

- The recessive χ^2 version of the CA test (CA-R), which was the best method for recessive gene models but the least effective when applied to non-recessive MOI data.
- The 2df genotype test (2df-g), which is never an optimal method for any of the scenarios we examined; in every scenario, a more powerful alternative can be identified (see Table 4).

Our results indicate that the best method in terms of statistical power is CA-A, but that little is lost if the MAX method is used instead. Similarly, the results in Tables 5, 6 and 7 indicate that the best method in terms of statistical power for identifying dominant MOI loci is CA-D and CA-R for recessive MOI loci. For multiplicative MOI loci, the best method is CA-A. In all four scenarios, little is lost if the MAX (or the BON) methods are used as replacements.

However, if we use the CA-A method, which is advocated by many as an initial GWAS pass and the locus in question is recessive or dominant, a power loss will occur. Table 8 indicates that while the CA-A method is the optimal choice (by as much as 2%), if the MOI of the locus is additive or multiplicative, there is a risk of as much as a 4% power loss if the locus MOI is dominant and as much as 15% loss if the MOI of the locus is recessive (see Table 8).

If we knew the distribution of the MOI property, we could assess the overall risk of using an additive method such as CA-A for GWAS. However, without a reliable estimate, one should exercise caution and apply a procedure that limits the risk of making the wrong assessment of the MOI inherent to the locus inducing diagnosis.

Discussion

In the literature, many statistical methods that have been used to perform GWAS assume a MOI specific hypothesis. Our results confirm the work of many others (see Iles, 2002): that is, in the context of a single marker scenario, the best method for predicting associations in recessive SNPs was the CA-R method; the best method for dominant MOI SNPs was the CA-D method; and the best method for additive and multiplicative SNPs was the CA-A method. We also show that the 2df genotype method used in many studies (for example, see Schymick et al. (2007)) is never an optimal method because there are always other methods that provide greater statistical power. This statement holds regardless of whether the MOI is known *a priori* or not. We also show that in the context of a general method to use in the initial GWAS pass, researchers may encounter adverse consequences if, for example, the MOI of the operating locus is not consistent with the assumption employed by the statistical method used. Thus, using 2df does not appear to be appropriate for GWA studies in any circumstances.

Consequently, we examined the possibility of employing an alternative procedure that incorporates the three core tests defined above into two multi test procedures: the Bonferroni procedure and the MAX test procedure due to Li et al. (2008). These procedures are opposites in many respects, in that they assume different underlying distributions of the test statistics. The MAX method assumes that the three separate tests have dependencies that can be accounted for, whereas the Bonferroni method assumes that they are mutually independent. We note that despite these differences, the two methods produce similar power profiles.

We generated our results using 1,000 replicates per parameter combination. Our standard error estimate of power varies from .262 to .315. Consequently, our 95 percent confidence interval around the mean will be approximately plus or minus .019. While we recognize that a larger number of replicates will improve power precision, we believe that our conclusion will remain as stated.

In summary, researchers should consider a multi-test procedure combining the results of individual MOI-based core tests as a possible statistical method for conducting GWAS rather than a 2df test. Combining individual methods and comparing the individual and combined results may help identify the MOI character of the gene. The actual process of combining the core tests into a single operational test can be done in a number of ways, all of which produce very similar statistical power profiles.

Acknowledgements

The authors wish to thank Craig R. Hollingsworth of RTI for technical writing and editing assistance.

References

1. Chan EKF, Hawken R, Reverter A (2009) The combined effect of SNP-marker and phenotype attributes in genome-wide association studies. Anim Genet 40: 149-156.
2. Freidlin B, Zheng G, Li Z, Gastwirth JL (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered 53: 146-152.

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000159
JPB/Vol.3 Issue 12

3. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Statistics 6: 65-70.

4. Iles MM (2002) Effect of mode of inheritance when calculating the power of a transmission/disequilibrium test study. Hum Hered 53: 153-7.

5. Kuo CL, Feingold E (2010) What's the Best Statistic for a Simple Test of Genetic Association in a Case-Control Study? Genet Epidemiol 34: 246-253.

6. Li Q, Zheng G, Li Z, Yu K (2008) Efficient approximation of p-value of the maximum of correlated tests, with applications to genome-wide association studies. Ann Hum Genet 72: 397-406.

7. Sasieni PD (1997) From genotypes to genes: doubling the sample size. Biometrics 53: 1253-1261.

8. Schymick JC, Scholz SW, Fung H, Britton A, Arepalli S et al. (2007) Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurology 6: 322-328.

9. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A Genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881-885.

10. Van Es MA, Van Vught PWJ, Blauw HM, Franke L, Saris CG, et al. (2008) Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. Nat Genet 40: 29-31.

11. Zheng G, Gastwirth JL (2006) On estimation of the variance in Cochran–Armitage trend tests for genetic association using case–control studies. Statist Med 25: 3150-9.

12. Ziegler A, Inke RK, Thompson JR (2008) Biostatistical aspects of genome-wide association studies. Biom J 50: 1-21.