BMC Genetics

METHODOLOGY ARTICLE Open Access

An entropy test for single-locus genetic association analysis

Manuel Ruiz-Marín*[1], Mariano Matilla-García[3], José Antonio García Cordoba[1], Juan Luis Susillo-González[2], Alejandro Romo-Astorga[2], Antonio González-Pérez[2], Agustín Ruiz[2] and Javier Gayán[2]

Abstract

Background: The etiology of complex diseases is due to the combination of genetic and environmental factors, usually many of them, and each with a small effect. The identification of these small-effect contributing factors is still a demanding task. Clearly, there is a need for more powerful tests of genetic association, and especially for the identification of rare effects

Results: We introduce a new genetic association test based on symbolic dynamics and symbolic entropy. Using a freely available software, we have applied this entropy test, and a conventional test, to simulated and real datasets, to illustrate the method and estimate type I error and power. We have also compared this new entropy test to the Fisher exact test for assessment of association with low-frequency SNPs. The entropy test is generally more powerful than the conventional test, and can be significantly more powerful when the genotypic test is applied to low allele-frequency markers. We have also shown that both the Fisher and Entropy methods are optimal to test for association with low-frequency SNPs (MAF around 1-5%), and both are conservative for very rare SNPs (MAF<1%)

Conclusions: We have developed a new, simple, consistent and powerful test to detect genetic association of biallelic/SNP markers in case-control data, by using symbolic dynamics and symbolic entropy as a measure of gene dependence. We also provide a standard asymptotic distribution of this test statistic. Given that the test is based on entropy measures, it avoids smoothed nonparametric estimation. The entropy test is generally as good or even more powerful than the conventional and Fisher tests. Furthermore, the entropy test is more computationally efficient than the Fisher's Exact test, especially for large number of markers. Therefore, this entropy-based test has the advantage of being optimal for most SNPs, regardless of their allele frequency (Minor Allele Frequency (MAF) between 1-50%). This property is quite beneficial, since many researchers tend to discard low allele-frequency SNPs from their analysis. Now they can apply the same statistical test of association to all SNPs in a single analysis., which can be especially helpful to detect rare effects.

Background

The etiology of complex diseases is due to the combination of genetic and environmental factors, usually many of them, and each with a small effect. The identification of these small-effect contributing factors is still a demanding task, often requiring a large budget, thousands of individuals, and half-a-million or more genetic markers. Even so, success is not guaranteed. In the last decade, genetic association tests have become widely used, since they can detect small genetic effects. The current availability of genome-wide genotyping tools, combined with large collections of affected and unaffected individuals, has allowed for association analysis of the entire genome with the intention to detect even those small genetic effects (i.e., Odds-Ratios (OR) around 1.2) that influence common complex diseases.

We have seen recently a proliferation of genome-wide association (GWA) analyses, some of which are identifying even genes with only small or modest effect sizes ([1] for a review). Nonetheless, the genetic factors found so far do not explain the total heritability of these diseases. Perhaps, the genetic architecture of these diseases is more complex than previously thought, involving many more genes, each with a small effect, and interacting among them and with environmental factors in complex ways. There is also the

* Correspondence: manuel.ruiz@upct.es
[1] Department of Quantitative Methods, Technical University of Cartagena, Paseo Alfonso XIII, 50, 30203, Cartagena, Spain
Full list of author information is available at the end of the article

possibility of a large background of rare mutations, each possibly having a relatively large effect, but at a very low frequency [2]. Clearly, there is a need for more powerful tests of genetic association, and especially for the identification of rare effects. This need will probably be exacerbated when low-cost whole genome sequencing becomes available, uncovering a large amount of rare variants in humans [3].

Although Information Theory was originally applied in the context of communication and engineering problems [4], entropy-based approaches have been also successfully applied to gene mapping. Specifically, there are information-theory-based tests implemented for population-based association studies using genotypic tests in case-control analysis and QTL analysis [5,6], gene-centric multimarker [7] and haplotype-based association studies [8] or epistasis analysis [9-12]. Moreover an entropy-based Transmission Disequilibrium Test (TDT) has also been described to conduct genome-wide studies in family trios [13].

In spite of these achievements, there is a scarce amount of simple and user-friendly computer programs to analyze and prioritize genome wide signals using entropy-based algorithms. Furthermore, a general entropy-based allelic test has not been described, studied and implemented in software to date. We have created a new genetic association test based on entropy that provides a general tool to conduct whole genome association studies. It is a new, simple, consistent and powerful test to detect genetic association of biallelic/SNP markers in case-control data, by using symbolic dynamics and symbolic entropy as a measure of gene dependence. Furthermore, we have implemented these algorithms in a software freely available to the scientific community. Using this computer program, named Gentropia, we have applied this entropy test, and a conventional test, to simulated and real datasets, to illustrate the method and estimate type I error and power of the test.

Results

To illustrate the method we used data from the SNP Resource at the NINDS Human Genetics Resource Center DNA and Cell Line Repository http://ccr.coriell.org/ninds/. The original genotyping was performed in the laboratory of Drs. Singleton and Hardy (NIA, LNG), Bethesda, MD USA [14]. We have used data on 270 patients with Parkinson's disease and 271 normal control individuals who were genotyped for 396,591 SNPs in all 22 autosomal chromosomes using the Illumina Infinium I and Infinium II assays. Cases were all unrelated white individuals with idiopathic Parkinson's disease and age of onset between 55-84 years (except for 3 young-onset individuals). The control sample was composed of neurologically normal, unrelated, white individuals. To explore the properties of the entropy test, and compare it to an equivalent conventional chi-square test, we simulated and analyzed datasets with specific properties. To simulate the specific effect size of a genetic variant, we wrote an algorithm that fixes the odds-ratio (OR) attributed to the SNP, and either fixes or sets randomly the minor allele frequency (MAF) in controls. Subsequently it estimates the MAF in cases necessary to generate the desired OR. Then, specific genotypes are generated for cases and controls according to the estimated allele frequencies in each group, and assuming Hardy-Weinberg equilibrium. Most datasets include 500 cases and 500 controls, and SNP marker genotypes were simulated under different genetic models (OR equal to 1 (no effect), 1.25, 1.5 and 2), and different marker allele frequencies (MAF equal to 0.05, 0.2, and 0.4). Type I error of the statistical tests was evaluated in a dataset where 10,000 SNPs were simulated under the null hypothesis, with allele frequencies chosen randomly between 0 and 0.5, but each SNP had a similar MAF in the case and control groups. For the power analysis, each dataset contains 100 SNPs with specific OR and MAF. Finally, to evaluate low allele-frequency markers in more detail, we simulated datasets of 5000 cases and 5000 controls, and 1000 SNPs with a variety of effect sizes (OR equal to 1, 1.5 and 1.8) and allele frequencies (MAF = 0.01, 0.03 and 0.06).

The entropy allelic and genotypic tests can be compared to association tests used commonly in the field of Human Genetics. For a biallelic SNP marker, a test of association between the SNP and a disease can be computed by comparing the allelic or the genotypic frequencies in cases and controls. The conventional allelic test is a chi-square test statistic with 1 degree of freedom, while the conventional genotypic test is a chi-square test with 2 degrees of freedom.

Null simulations

A simulated dataset, consisting of 500 cases and 500 controls, was analyzed with both conventional chi-square and entropy-based association tests. A total of 10,000 SNPs, with different allele frequencies ($0 < MAF < 0.5$), were simulated under the null hypothesis, that is, to have no effect on the trait. This analysis can reflect whether the new entropy-based association tests conform to the theoretical distribution. We counted the number of test-statistics that had values above the critical values of the expected distribution, to estimate the type I error for each test.

The conventional chi-square and the entropy-based tests, in both its genotypic and allelic versions, yield approximately the expected number of false positives (see Table 1), suggesting they all conform to the expected theoretical distributions (χ^2 with 1 or 2 degrees of freedom). The entropy-based test statistic was always equal or larger in value than the conventional chi-square test. On average, the entropy-based genotypic method increased the test-statistic in 0.047 chi-square units (a 1.7 percent), while the entropy allelic

Table 1: Type I Error for Conventional and Entropy tests.

Expected	Conventional Allelic	ALi	Conventional Genotypic	GEi
0.0500	0.0475	0.0478	0.048	0.0514
0.0100	0.0109	0.0111	0.009	0.0104
0.0010	0.0012	0.0013	0.001	0.0010

AL = Entropy allelic test. GE = Entropy Gentotypic test.

test exhibited an average increase of 0.003 chi-square units (a 0.1 percent).

Power analysis

To estimate the power of both conventional and entropy-based tests, we carried out an analysis of simulated dataset of 500 cases and 500 controls. Sets of 100 SNPs were simulated under different alternative hypothesis, with different effect sizes (odds-ratios of 1.25, 1.5, and 2) and minor allele frequencies (0.05, 0.2 and 0.4). The entropy-based test statistic was always equal or larger in value than the conventional chi-square test, and therefore its power was also always equal or larger (see Tables 2 and 3). This increase in power is small, and it is more pronounced for the genotypic than for the allelic test. The gain in power with the genotypic entropy test tends to become apparent for larger chi-square values, or especially in markers with low allele frequency. For the allelic test, the entropy test is also sensibly more powerful for the $OR = 2$ and $MAF = 0.05$ simulation.

Because the gain in power is correlated with the size of the chi-square statistic, we computed a "proportional power gain", that is, the difference between the entropy and the conventional chi-squares, divided by the conventional chi-square. This proportional gain allows us to compare the gain across the different simulated scenarios. As can be seen in Table 4, the average increase in power is only small, ranging between 0.1 and 9.7%, except for the genotypic test on low allele frequency SNPs, for which the power gain range is much larger (5.2-9.7%). In general, the gain in power increases when the OR increases and when MAF decreases.

Results show that the entropy test is similar or more powerful than the conventional chi-square test. The gain in power is small, and in some cases not different from the false-positive increase under the null hypotheses. Nonetheless, the entropy tests are an improvement over genotypic tests, for reasons discussed in the Discussion, and may become useful when power is limited, and especially, for the analysis of low allele-frequency SNPs.

Low-allele frequency markers

To study in more detail the performance of the genotypic conventional and entropy tests in low-allele frequency markers, we simulated datasets of 5000 cases and 5000 controls, so there would be enough power to detect these rare effects. Each dataset included 1000 SNPs simulated under an specific effect size (*OR* equal to 1, 1.5 and 1.8) and allele frequency (MAF equal to 0.01, 0.03 and 0.06).

The analysis of the null-effect markers (OR = 1), reveals that both tests conform approximately well to the hypothetical null distribution for allele frequencies of 0.06 and 0.03. However, both tests are too conservative for very rare alleles, with minor allele frequencies around 1% (Table 5).

Table 6 confirms that both tests behave similarly when the study has enough statistical power, that is, for allele frequencies above 5%, and even for markers with lower frequency and large effect (MAF = 3% and OR = 1.8). In contrast, it is evident that the genotypic entropy test is more powerful than the genotypic conventional test for markers with rare (MAF = 0.01), or low allele-frequency (MAF = 0.03).

It is important to note here that the Fisher exact test is often used as a test of association for rare SNPs. For this reason, we have also compared the Fisher and Entropy tests, in their allelic and genotypic versions. For low frequency SNPs (MAF = 0.03) the results suggest that all four tests conform to their theoretical distribution (Tables 7 and 8). We find that Fisher and Entropy are quite similar for the allelic test, with the Entropy test being slightly more powerful (but also slightly more liberal) than Fisher. We conclude that both tests are essentially equivalent for the allelic test. However, both allelic tests are more powerful than any of the genotypic tests (Table 7).

Tables 8 and 9 describe the Genotypic test, for MAFs of 0.03 and 0.01. When comparing Fisher versus Entropy genotypic tests with low frequency SNPs (MAF = 0.03), power is also very similar, slightly better for Fisher than Entropy (Table 8).

Table 2: Power (%) for conventional (CA) and entropy (AL) allelic tests for different Minor Allele Frequencies (MAF) and Odds-ratios (OR).

	MAF	**0.05**		**0.2**		**0.4**	
OR = 1.25	α	CA	AL_i	CA	AL_i	CA	AL_i
	0.001	1	1	12	12	20	20
	10^{-4}	0	0	2	3	2	3
	10^{-5}	0	0	1	1	1	1
	10^{-6}	0	0	1	1	1	1
	10^{-7}	0	0	0	0	1	1
OR = 1.5							
	0.001	12	12	71	71	88	88
	10^{-4}	6	6	40	40	72	72
	10^{-5}	2	3	20	20	51	51
	10^{-6}	0	0	10	10	30	30
	10^{-7}	0	0	5	5	17	19
OR = 2							
	0.001	77	79	100	100	100	100
	10^{-4}	48	49	100	100	100	100
	10^{-5}	25	26	96	96	100	100
	10^{-6}	19	19	90	91	100	100
	10^{-7}	10	11	82	82	98	98

Nonetheless, for very rare alleles (MAF<0.01), both tests are extremely conservative, more so the Entropy test, which consequently shows lower power for association than the Fisher test. In summary, it seems that both tests, Fisher and Entropy, are optimal to test for association with low-frequency SNPs (MAF around 1-5%), and both are conservative for very rare SNPs (MAF<1%).

These results alltogether suggest that symbolic entropy based tests are valid for testing for association, and do not create a significant bias under the null hypothesis. Moreover, the entropy tests are more stable than the conventional

Table 3: Power (%) for conventional (CG) and entropy (GE) genotypic tests for different Minor Allele Frequencies (MAF) and Odds-ratios (OR).

	MAF	**0.05**		**0.2**		**0.4**	
OR = 1.25	α	CG	GE_i	CG	GE_i	CG	GE_i
	0.001	0	1	8	9	10	10
	10^{-4}	0	0	2	2	3	3
	10^{-5}	0	0	2	2	1	1
	10^{-6}	0	0	0	0	1	1
	10^{-7}	0	0	0	0	0	0
OR = 1.5							
	0.001	8	9	51	51	81	81
	10^{-4}	5	6	28	28	59	59
	10^{-5}	0	2	11	11	35	37
	10^{-6}	0	0	7	8	21	22
	10^{-7}	0	0	3	4	7	9
OR = 2							
	0.001	61	67	100	100	100	100
	10^{-4}	32	37	98	98	100	100
	10^{-5}	20	21	93	93	100	100
	10^{-6}	12	14	84	84	99	100
	10^{-7}	4	8	63	66	97	97

and Fisher exact tests regardless the allelic frequency. In addition, entropy tests are less expensive in computational terms than Fisher exact test.

Parkinson disease

To illustrate the analysis method in a real dataset, we have analyzed a sample of 270 Parkinson disease patients and 271 controls, genotyped for 396,591 SNPs across the genome. This dataset includes SNPs with a wide variety of

Table 4: Allelic and genotypic entropy-tests gain power (%) for different Minor Allele Frequencies (MAF) and Odds-ratios (OR).

		Allelic Power	Genotypic Power
OR	**MAF**	**Gain (Mean%)**	**Gain (Mean%)**
1.25	0.4	0.1	0.3
1.5	0.4	0.2	0.8
2	0.4	0.6	2
1.25	0.2	0.1	0.8
1.5	0.2	0.4	1.2
2	0.2	1.1	2.2
1.25	0.05	0.3	9.7
1.5	0.05	0.7	5.3
2	0.05	1.7	5.2

characteristics, such as different allelic and genotypic frequencies.

As we saw in the simulated datasets, the entropy tests are generally more powerful than the conventional tests. For these real data, we find some SNPs for which the entropy chi-square is lower than the conventional chi-square. However, these markers have low call rates in cases (lower than 45%), suggesting the presence of genotyping errors, and therefore would generally be excluded from association analysis. For the genotypic test, chi-square values (2 df) range between 0 and 41.95 for the conventional test, and 0-44.95 for the entropy test. On average, the entropy chi-square is 1.9% larger than the conventional test. If we consider the top-100 chi-square values for each test, there is 92% concordance in the SNPs that appear in these two rankings (irrespective of order within the ranking). The 8 SNPs chosen only by the conventional test still appear in the top 112 SNPs for the entropy test, revealing that the entropy test agrees well with the conventional test. Nonetheless, the 8 SNPs chosen only by the entropy test appear in ranks 105-359 in the conventional test. These SNPs far down the ranking of the conventional test have a common characteristic, they have a low frequency for the rare genotype (0-2 individuals only). As we saw for the null simulations, for low allele/genotype frequencies, the genotypic entropy test statistic is larger than the conventional chi-square, suggesting that the entropy test can help detect genetic effects in low allele/genotype frequency SNPs. For the allelic test, chi-square values (1 df) range between 0-30.35 for the conventional test, and 0-32.29 for the entropy test. Both tests agree well in chi-square size, with only a 0.1% difference on average. Both tests also agree on 96% of the SNPs in their top-100 ranking, and the 4 SNPs in disagreement, are ranked no lower than 106th in the other ranking. These tests are nearly identical, with the entropy test slightly more powerful.

Discussion

Several entropy-based tests have been recently developed for population-based and family-based genetic association studies to perform gene mapping of complex diseases. However, to our best knowledge a simple and computationally feasible allelic entropy-based test useful for GWA studies is not available yet. Allelic and genotypic methods represent the gold-standard statistical test to start the prioritisation of markers during GWAS. The development of

Table 5: Type I error for genotypic test with low minor allele frequencies (MAF).

	Expected	CG	GEi
MAF = 0.01			
	0.05	0.022	0.019
	0.01	0.000	0.000
MAF = 0.03			
	0.05	0.044	0.052
	0.01	0.008	0.013
MAF = 0.06			
	0.05	0.057	0.062
	0.01	0.013	0.013

CG = Conventional genotypic test. GE = Entropy genotypic test.

new and more powerful association tests can aid in the identification of small or rare effects, which may be widespread in the etiology of complex diseases, as shown in the recent GWAS [1]. To cover this need, we have developed a new likelihood ratio test of genetic association for biallelic markers such as SNP markers that is based on symbolic analysis and the relevant concept of entropy. Other authors, [8,5] and [6] among others, have used the concept of entropy for case/control association studies. In [8], the authors develop a statistic, namely T_{PE}, that asymptotically follows a χ^2 distribution. In order to obtain the asymptotic distribution of T_{PE}, they require entropy to be continuously differentiable with respect to the frequencies of the haplotypes, which represent a problem when the frequency of an haplotype is zero either in cases or in controls. In such a case then the haplotypes need to be grouped with others haplotypes which yield a decrease in statistical power. Moreover the T_{PE} statistic also requires the estimation of an inverse matrix. Since this is not always possible, this inverse matrix has to be approximated by its generalized inverse, possibly introducing a bias in the statistic. Also, the computation of T_{PE} is more expensive in computational running times than our entropy-based test GE_i [6]. provides a measure for linkage disequilibrium (LD) between a marker and the trait locus, that is based on the comparison of the entropy and conditional entropy in a marker in extreme samples of population. Nevertheless, the authors do not give the distribution of the constructed measure, and hence it is not possible to assign a statistical significance to the procedure. Finally, [5], in the context of clusters of genetic markers, uses multidimensional scaling in conjunction with the Mutual Information (MI) between two discrete random variables. They use the fact that under the null of no association, MI can be approximated by means of a second order Taylor series expansion to a Gamma distribution. These entropy-based methods provide a tool to test for allelic and genotypic association between a marker and a qualitative phenotype. In these papers, the empirical size and power of the tests has not been computed nor compared in power with conventional tests.

The entropy test has several advantages over conventional tests: (1) It has been proved that the test is consistent. This is a valuable property since the test will asymptotically reject any systematic deviations between the distributions of cases and controls. (2) Importantly, the test does not require prior knowledge of parameters, and therefore can not be biased by potential decisions of the user. These properties, together with the fact that the test is simple, intuitive and fast in computational terms, make this test a theoreti-

Table 6: Genotypic-Test Power (%) for different low minor allele frequencies (MAF) SNPs and Odds-ratios (OR).

MAF	**0.01**		**0.03**		**0.06**	
α	CG	GE_i	CG	GE_i	CG	GE_i
OR = 1.5						
0.05	65.6	81.4	100	100	100	100
0.01	49.9	64.3	99.3	99.3	100	100
10^{-3}	24.3	33.5	95.1	95.3	100	100
10^{-4}	9.6	14.4	87.4	87.5	100	100
10^{-5}	3.0	5.2	74.6	75.1	100	100
10^{-6}	0.9	1.2	55.9	57.4	98.9	99.0
10^{-7}	0.3	0.4	39.5	40.9	95.8	95.8
OR = 1.8						
0.05	87.4	99.1	100	100	100	100
0.01	85.1	96.8	100	100	100	100
10^{-3}	77.2	88.6	100	100	100	100
10^{-4}	64.2	74.8	100	100	100	100
10^{-5}	44.7	55.4	100	100	100	100
10^{-6}	28	34.8	99.7	99.8	100	100
10^{-7}	14.4	20.5	98.9	98.9	100	100

CG = Conventional genotypic test. GE = Entropy genotypic test.

cally appealing and powerful technique to deal with the detection of genetic association.

We have shown, both in simulated and real data, that the entropy and conventional tests, both in their genotypic and allelic versions, fit well their expected null distributions, or are even conservative for the detection of rare alleles (MAF ≤ 0.01). More so, the entropy genotypic test is more powerful than the conventional test, especially for those low-frequency SNPs. This is an important property, because there is a current need for tools to detect rare genetic effects.

The Fisher exact test is often used as a test of association for rare SNPs, although it is hard to program because of the complexity of its formula, and it is also computationally intensive. To make sure that the Entropy test is efficient

Table 7: Fisher versus entropy allelic tests for different Odds-ratios (OR).

Allelic	MAF 0.03					
	Type I error		Power		Power	
	OR = 1		OR = 1.5		OR = 1.8	
alpha	Fisher	Entropy	Fisher	Entropy	Fisher	Entropy
0.05	0.052	0.057	100.0	100.0	100.0	100.0
0.01	0.011	0.013	99.8	99.8	100.0	100.0
1.E-03	0.001	0.001	97.7	98.1	100.0	100.0
1.E-04	0.000	0.000	91.4	92.2	100.0	100.0
1.E-05	0.000	0.000	81.5	83.0	100.0	100.0
1.E-06	0.000	0.000	66.3	67.3	99.9	99.9
1.E-07	0.000	0.000	49.3	50.5	99.5	99.6

MAF = Minor allele frequency.

also for rare SNPs, we have compared the Fisher and Entropy tests, in their allelic and genotypic versions. We have shown that the Entropy test is as powerful as the Fisher exact test for the analysis of low frequency SNPs (MAF between 1-5%). Therefore, this entropy-based test has the advantage of being optimal for most SNPs, only losing power respect to the Fisher test for very rare alleles (MAF<1%). This property is quite beneficial, since many researchers tend to discard low allele-frequency SNPs from their analysis. Now they can apply the same statistical test of association to all SNPs in a single analysis.

These entropy tests are easy to compute with the formulas provided in this paper, which can be incorporated into any genetic analysis tool. We are making freely available a simple software (Gentropia) to carry out these entropy-based genetic analyses. A linux version of the software can be downloaded from the following Website: http://www.neocodex.com/en/Gentropia.zip. The analysis is quite fast. For example, an association analysis of 1,000 SNPs on 10,000 individuals takes only 4 seconds on a 2.4 Ghz CPU; A genome-wide association analysis of 400,000 SNPs on 550 individuals takes 84 seconds, which is quite satisfactory

Conclusions

In summary, this is an application of symbolic analysis and entropy to carry out a genome-wide association analysis. We have implemented this simple and fast method in a freely available software http://www.neocodex.com/en/Gentropia.zip. This entropy-based method to detect genetic association is more powerful than conventional tests, and can be especially useful in the detection of rare effects due to low-frequency genotypes. The method can be improved to include other tests of association (dominance, recessive, etc.), and covariates. Moreover, the method can be extended for the detection of epistasis.

Methods

Entropy Model

First we give some definitions and introduce the basic notation.

Let P be the population to be studied. Denote by C the set of cases with a particular disease in P and by C^c the complementary, that is, the set of controls. Let N_{ca} and N_{co} be the cardinality of the sets C and C^c respectively and let $N = N_{ca} + N_{co}$ be the total amount of individuals in the population.

Table 8: Fisher versus entropy genotypic tests tests for different Odds-ratios (OR).

Genotypic	MAF 0.03					
	Type I error		Power		Power	
	OR = 1		OR = 1.5		OR = 1.8	
alpha	Fisher	Entropy	Fisher	Entropy	Fisher	Entropy
0.05	0.046	0.051	100.0	100.0	100.0	100.0
0.01	0.010	0.013	99.4	99.3	100.0	100.0
1.E-03	0.001	0.000	95.8	95.3	100.0	100.0
1.E-04	0.000	0.000	88.5	87.5	100.0	100.0
1.E-05	0.000	0.000	77.7	75.1	100.0	100.0
1.E-06	0.000	0.000	61.1	57.4	99.8	99.8
1.E-07	0.000	0.000	45.1	40.9	99.2	98.9

MAF = Minor allele frequency.

Each SNP_i in each individual e 8 P can take only one of the three possible values, AA_i, Aa_i or aa_i. Let $S_i = \{AA_i, Aa_i, aa_i\}$. Moreover, each individual e 8 P belongs to either C or C^c, therefore we can say that a SNP_i takes the value (X_i, ca) if e 8 C or (X_i, co) if e 8 C^c, for X_i 8 S_i. We will call an element in $S_i \times \{ca,co\}$ a *symbol*. Therefore we can define the following map

$$f_i : P \to S_i \times \{ca, co\}$$

defined by $f_i(e) = (X_i, t)$ for X_i 8 S_i and t 8 $\{ca,co\}$, that is, the map f_i associates to each individual e 8 P the value of its SNP_i and whether e is a control or a case. We will call f_i a *symbolization map*. In this case we will say that individual e is of (X_i, t) -type. In other words, each individual is labelled with its genotype, differentiating whether the individual is a control or a case.

Denote by

$$n_{X_i} = \#\{e \in P \mid f_i(e) = (X_i, ca)\}, \quad (1)$$

and

$$m_{X_i} = \#\{e \in P \mid f_i(e) = (X_i, co)\}, \quad (2)$$

that is, the cardinality of the subsets of P formed by all the individuals of (X_i, ca) -type and (X_i, co) -type respectively. Therefore $n_{X_i} + m_{X_i}$ is the number of individuals of X_i-type.

Also, under the conditions above, one could easily compute the relative frequency of a symbol (X_i, t) 8 $S_i \times \{ca, co\}$ by:

$$p_{X_i} = \frac{\#\{e \in P \mid e \text{ is of } (X_i, ca)-\text{type}\}}{N} \quad (3)$$

and

$$q_{X_i} = \frac{\#\{e \in P \mid e \text{ is of } (X_i, co)-\text{type}\}}{N}. \quad (4)$$

Table 9: Fisher versus entropy genotypic tests tests for different Odds-ratios (OR).

Genotypic	MAF 0.01					
	Type I error		Power		Power	
	OR = 1		OR = 1.5		OR = 1.8	
alpha	Fisher	Entropy	Fisher	Entropy	Fisher	Entropy
0.05	0.036	0.022	86.0	81.4	99.9	99.1
0.01	0.030	0.000	71.7	64.3	98.0	96.8
1.E-03	0.000	0.000	41.8	33.5	91.9	88.6
1.E-04	0.000	0.000	20.1	14.4	80.1	74.9
1.E-05	0.000	0.000	8.5	5.2	64.0	55.5
1.E-06	0.000	0.000	2.3	1.2	43.8	34.8
1.E-07	0.000	0.000	0.7	0.4	25.1	20.5

MAF = Minor allele frequency.

Hence the total frequency of a symbol X_i is $s_{X_i} = p_{X_i} + q_{X_i}$.

Now under this setting we can define the *symbolic entropy* of a SNP_i. This entropy is defined as the Shannon's entropy of the 3 distinct symbols as follows:

$$h(S_i) = -\sum_{X_i \in S_i} s_{X_i} \ln(s_{X_i}). \quad (5)$$

Symbolic entropy, $h(S_i)$, is the information contained in comparing the 3 symbols (i.e., the 3 possible values of the genotype) in S_i among all the individuals in P.

Similarly we have the symbolic entropy for cases, controls and case-control entropy by

$$h(S_i, ca) = -\sum_{X_i \in S_i} p_{X_i} \ln(p_{X_i}), \quad (6)$$

$$h(S_i, co) = -\sum_{X_i \in S_i} q_{X_i} \ln(q_{X_i}), \quad (7)$$

and

$$h(C, C^c) = -\frac{N_{ca}}{N} \ln\left(\frac{N_{ca}}{N}\right) - \frac{N_{co}}{N} \ln\left(\frac{N_{co}}{N}\right), \quad (8)$$

respectively.

Construction of the entropy test

In this section we construct a test to detect gene effects in the set C of cases with all the machinery defined in Section 1. In order to construct the test, which is the aim of this paper, we consider the following null hypothesis:

$$H_0 : SNP_i \text{ distributes equally in } C \text{ than in } C^c, \quad (9)$$

that is,

$$H_0 : q_{X_i} = \frac{N_{co}}{N_{ca}} p_{X_i} \text{ for } i = 1, 2, 3 \quad (10)$$

against any other alternative.

Now for a symbol (X_i, t) 8 $S_i \times \{ca, co\}$ and an individual e 8 P we define the random variable $Z_{(X_i,t)e}$ as follows:

$$Z^t_{X_i e} = \begin{cases} 1 & \text{if } f_i(e) = (X_i, t) \\ 0 & \text{otherwise,} \end{cases} \quad (11)$$

that is, we have that $Z^t_{X_i e} = 1 = 1$ if and only if e is of (X_i, t) -type, $Z_{(X_i,t)e} = 0$ otherwise. Therefore, given that an individual e is a case, $t = ca$, (respectively e is a control $t = co$), the variable $Z_{(X_i,t)e}$ indicates whether individual e has genotype X_i (taking value 1) or not (taking value zero).

Then $Z^t_{X_i e} = 1$ is a Bernoulli variable with probability of "success" either p_{X_i} if $t = ca$ or q_{X_i} if $t = co$, where "success" means that e is of (X_i, t) -type. Then we are interested in to know how many e's are of (X_i, t) -type for all symbol (X_i, t) 8 $S_i \times \{ca, co\}$. In order to answer the question we construct the following variable

$$Y^t_{X_i} = \sum_{e \in P} Z^t_{X_i e} \quad (12)$$

The variable $Y^t_{X_i}$ can take the values $\{0,1,2,..., N\}$. Therefore, it follows that the variable $Y^t_{X_i}$ is the Binomial random variable

$$Y^{ca}_{X_i} \approx B(N, p_{X_i}) \quad \text{or} \quad Y^{co}_{X_i} \approx B(N, q_{X_i}). \quad (13)$$

Then the joint probability density function of the 6 variables

$$P(Y^{ca}_{AA_i} = a_1, Y^{ca}_{Aa_i} = a_2, Y^{ca}_{aa_i} = a_3, Y^{co}_{AA_i} = a_4, Y^{co}_{Aa_i} = a_5, Y^{co}_{aa_i} = a_6) \quad (14)$$

is:

$$\frac{(a_1+a_2+a_3+a_4+a_5+a_6)!}{a_1!a_2!a_3!a_4!a_5!a_6!} p^{a_1}_{AA_i} p^{a_2}_{Aa_i} p^{a_3}_{aa_i} q^{a_4}_{AA_i} q^{a_5}_{Aa_i} q^{a_6}_{aa_i} \quad (15)$$

where $a_1 + a_2 + a_3 + a_4 + a_5 + a_6 = N$. Consequently the joint distribution of the 6 variables $(Y^{ca}_{AA_i}, Y^{ca}_{Aa_i}, Y^{ca}_{aa_i}, Y^{co}_{AA_i}, Y^{co}_{Aa_i}, Y^{co}_{aa_i})$ is a multinomial distribution.

The likelihood function $L(p_{AA_i}, p_{Aa_i}, p_{aa_i}, q_{AA_i}, q_{Aa_i}, q_{aa_i})$ of the distribution (15) is:

$$\Gamma \cdot p^{n_{AA_i}}_{AA_i} p^{n_{Aa_i}}_{Aa_i} p^{n_{aa_i}}_{aa_i} q^{m_{AA_i}}_{AA_i} q^{m_{Aa_i}}_{Aa_i} q^{m_{aa_i}}_{aa_i} \quad (16)$$

where $\Gamma = \frac{N!}{n_{AA_i}! n_{Aa_i}! n_{aa_i}! m_{AA_i}! m_{Aa_i}! m_{aa_i}!}$. Also, since

$$p_{AA_i} + p_{Aa_i} + p_{aa_i} + q_{AA_i} + q_{Aa_i} + q_{aa_i} = 1$$

it follows that the logarithm of this likelihood function remains as

$$\ln(L(p_{AA_i}, p_{Aa_i}, p_{aa_i}, q_{AA_i}, q_{Aa_i}, q_{aa_i})) = \ln(\Gamma) + \sum_{X_i \in S_i} n_{X_i} \ln(p_{X_i}) + m_{AA_i} \ln(q_{AA_i}) + m_{Aa_i} \ln(q_{Aa_i}) + m_{aa_i} \ln(1 - p_{AA_i} - p_{Aa_i} - p_{aa_i} - q_{AA_i} - q_{Aa_i}).$$

In order to obtain the maximum likelihood estimators $\hat{p}_{X_i}$ and $\hat{q}_{X_i}$ of p_{X_i} and q_{X_i} respectively for all $i = 1,2,3,$, we solve the following equations

$$\frac{\partial \ln(L)}{\partial p_{X_i}} = 0 \quad \frac{\partial \ln(L)}{\partial q_{X_i}} = 0 \quad (17)$$

to get that

$$\hat{p}_{X_i} = \frac{n_{X_i}}{N} \quad \hat{q}_{X_i} = \frac{m_{X_i}}{N}, \quad \hat{s}_{X_i} = \frac{n_{X_i} + m_{X_i}}{N} \quad (18)$$

Then, under the null H_0, we have that $q_{X_i} = \frac{N_{co}}{N_{ca}} p_{X_i}$ and thus,

$$s_{X_i} = p_{X_i} + q_{X_i} = p_{X_i} + \frac{N_{co}}{N_{ca}} p_{X_i} = \frac{N}{N_{ca}} p_{X_i}.$$

Therefore the likelihood ratio statistic is (see for example [15]):

$$\lambda_i(Y) = \frac{p^{n_{AA_i}}_{AA_i} p^{n_{Aa_i}}_{Aa_i} p^{n_{aa_i}}_{aa_i} q^{m_{AA_i}}_{AA_i} q^{m_{Aa_i}}_{Aa_i} q^{m_{aa_i}}_{aa_i}}{\hat{p}^{n_{AA_i}}_{AA_i} \hat{p}^{n_{Aa_i}}_{Aa_i} \hat{p}^{n_{aa_i}}_{aa_i} \hat{q}^{m_{AA_i}}_{AA_i} \hat{q}^{m_{Aa_i}}_{Aa_i} \hat{q}^{m_{aa_i}}_{aa_i}} \quad (19)$$

and thus, under the null H_0 we get that $\lambda_i(Y)$ remains as:

$$\frac{\left(\frac{N_{co}}{N_{ca}}\right)^{N_{co}}\left(\frac{N_{ca}}{N}\right)^{N} s_{AA_i}^{n_{AA_i}+m_{AA_i}} s_{Aa_i}^{n_{Aa_i}+m_{Aa_i}} s_{aa_i}^{n_{aa_i}}}{\prod_{X_i\in S_i}\left(\frac{n_{X_i}}{N}\right)^{n_{X_i}} \prod_{X_i\in S_i}\left(\frac{m_{X_i}}{N}\right)^{m_{X_i}}} \tag{20}$$

On the other hand, $GE_i = -2\ln(\lambda_i(Y))$ asymptotically follows a Chi-squared distribution with 2 degrees of freedom (see for instance [15]). Hence, we obtain that the estimator $\widehat{GE}_i$ of GE_i is:

$$\begin{aligned}\widehat{GE}_i = -2N[&\frac{N_{co}}{N}\ln\left(\frac{N_{co}}{N_{ca}}\right)+\ln\left(\frac{N_{ca}}{N}\right)+\\ &\sum_{X_i\in S_i}\frac{n_{X_i}+m_{X_i}}{N}\ln\left(\frac{n_{X_i}+m_{X_i}}{N}\right)-\\ &\sum_{X_i\in S_i}\frac{n_{X_i}}{N}\ln(\frac{n_{X_i}}{N})-\sum_{X_i\in S_i}\frac{m_{X_i}}{N}\ln(\frac{m_{X_i}}{N})=\\ &= 2N[\hat{h}(C,C^c)+\hat{h}(S_i)-\hat{h}(S_i,ca)-\hat{h}(S_i,co)]\end{aligned} \tag{21}$$

Therefore we have proved the following theorem.

Theorem 1. Let SNP_i be a single nucleotide polymorphism. For a particular disease denote by N the number of individuals in the population, N_{ca} the number of cases and by N_{co} the number of controls. Denote by $h(C,C^c)$ the case-control entropy and by $h(S_i)$, $h(S_i, ca)$ and $h(S_i, co)$ the symbolic entropy in the population, in cases and in controls respectively, as defined in (5, 6 and 7). If the SNP_i distributes equally in cases than in controls, then

$$GE_i = 2N[h(C,C^c)+h(S_i)-h(S_i,ca)-h(S_i,co)] \tag{22}$$

is asymptotically χ_2^2 distributed.

Let α be a real number with $0 \le \alpha \le 1$. Let χ_α^2 be such that

$$P(\chi_2^2 > \chi_\alpha^2) = \alpha. \tag{23}$$

Then the decision rule in the application of the GE_i test at a 100(1-α)% confidence level is:

$$\begin{aligned}&\text{If } 0 \le GE_i \le \chi_\alpha^2 \text{ Do not Reject } H_0\\ &\qquad \textit{Otherwise} \text{ Reject } H_0\end{aligned} \tag{24}$$

Furthermore, an entropy allelic test can be developed in a similar manner. More concretely, let now define the set $A_i = \{A_i, a_i\}$ formed by the two possible alleles that form the SNP_i.

Let

$$p_{A_i} = \frac{2n_{AA_i}+n_{Aa_i}}{2N} \quad p_{a_i} = \frac{2n_{aa_i}+n_{Aa_i}}{2N} \tag{25}$$

$$q_{A_i} = \frac{2m_{AA_i}+m_{Aa_i}}{2N} \quad q_{a_i} = \frac{2m_{aa_i}+m_{Aa_i}}{2N}. \tag{26}$$

Denote by $s_{A_i} = p_{A_i} + q_{A_i}$ and $s_{a_i} = p_{a_i} + q_{a_i}$ the total allele frequency. Then we can easily define the allele entropies of a SNP_i by

$$\begin{aligned}h(A_i) &= -\sum_{X_i\in A_i} s_{X_i}\ln(s_{X_i}) = -s_{A_i}\ln(s_{A_i}) - s_{a_i}\ln(\\ h(A_i,ca) &= -\sum_{X_i\in A_i} p_{X_i}\ln(p_{X_i}) = -p_{A_i}\ln(p_{A_i}) - p_{a_i}\,l\\ h(A_i,co) &= -\sum_{X_i\in A_i} q_{X_i}\ln(q_{X_i}) = -q_{A_i}\ln(q_{A_i}) - q_{a_i}\,l\end{aligned} \tag{27}$$

Now, with this notation and following all the steps of the proof of Theorem 1, we get the following result.

Theorem 2. Let $A_i = \{A_i, a_i\}$ be the alleles forming a single nucleotide polymorphism SNP_i. For a particular disease denote by N the number of individuals in the population, N_{ca} the number of cases and by N_{co} the number of controls. Denote by $h(C,C^c)$ the case-control entropy and by $h(A_i)$, $h(A_i, ca)$ and $h(A_i, co)$ the allele entropy in the population, in cases and in controls respectively. If the allele A_i distributes equally in cases than in controls, then

$$AL_i = 4N[h(C,C^c)+h(A_i)-h(A_i,ca)-h(A_i,co)] \tag{28}$$

is asymptotically χ_1^2 distributed.

Consistency of the entropy test

Next we prove that the GE_i test is consistent for a wide variety of alternatives to the null. This is a valuable property since the test will reject asymptotically that the SNP_i distributes equally between cases and controls whenever this assumption is not true. The proof of the following theorem can be found in Appendix section. Since the proof is similar for both statistics we only prove it for GE_i.

Theorem 3. Let SNP_i be a single nucleotide polymorphism. If the SNP_i does not distribute equally in cases than in controls, then

$$\lim_{N\to\infty} \Pr(GE_i > C) = 1 \ (\text{resp. } \lim_{N\to\infty} \Pr(AL_i > C) = 1)$$

for all real number $0 < C < \infty$.

Since Theorem 3 implies $GE_i \to +\infty$ with probability approaching 1 always SNP_i does not distribute equally in cases than in controls, then upper-tailed critical values are appropriated.

Appendix: Proof of consistency

Proof of Theorem 3 First notice that the estimators $\hat{h}(S_i)$, $\hat{h}(S_i, ca)$ and $\hat{h}(S_i, co)$, of $h(S_i, ca)$, $h(S_i, co)$ and $h(S_i)$ respectively, are consistent because $p\lim_{N\to\infty} \hat{p}_{X_i} = p_{X_i}$, $p\lim_{N\to\infty} \hat{q}_{X_i} = q_{X_i}$ and $p\lim_{N\to\infty} \hat{s}_{X_i} = s_{X_i}$ Denote by $H_i = h(C, C^c) + h(S_i) - h(S_i, ca) - h(S_i, co)$ and notice that H_i can be written as

$$H_i = \sum_{X_i\in S_i} p_{X_i} \ln\left(\frac{p_{X_i}}{\left(\sum_{Y_i\in S_i} p_{Y_i}\right)(p_{X_i}+q_{X_i})}\right) + \sum_{X_i\in S_i} q_{X_i} \ln\left(\frac{q_{X_i}}{\left(\sum_{Y_i\in S_i} q_{Y_i}\right)(p_{X_i}+q_{X_i})}\right)$$

Hence, since $-\ln(x) > 1 - x$ for all $x \neq 1$ we get that

$$H_i > \sum_{X_i\in S_i} p_{X_i}\left(1 - \frac{\left(\sum_{Y_i\in S_i} p_{Y_i}\right)(p_{X_i}+q_{X_i})}{p_{X_i}}\right) + \sum_{X_i\in S_i} q_{X_i}\left(1 - \frac{\left(\sum_{Y_i\in S_i} q_{Y_i}\right)(p_{X_i}+q_{X_i})}{q_{X_i}}\right) = \sum_{X_i\in S_i}(p_{X_i}+q_{X_i}) - \left(\sum_{X_i\in S_i}(p_{X_i}+q_{X_i})\right)\left(\sum_{X_i\in S_i}(p_{X_i}+q_{X_i})\right) = 1 - 1 = 0$$

always

$$\frac{p_{X_i}}{\left(\sum_{Y_i\in S_i} p_{Y_i}\right)(p_{X_i}+q_{X_i})} \neq 1 \text{ or } \frac{q_{X_i}}{\left(\sum_{X_i\in S_i} q_{X_i}\right)(p_{X_i}+q_{X_i})} \tag{29}$$

for some $X_i \in S_i$.

On the other hand, H_0 is equivalent to

$$p_{X_i} + q_{X_i} = \frac{p_{X_i}}{p_{AA_i} + p_{Aa_i} + p_{aa_i}}. \tag{30}$$

Therefore under the alternative, H_1, condition (29) is always satisfied and hence $H_i > 0$.

Let $0 < C < \infty$ be a real number and take N large enough such that

$$\frac{C}{2N} < H_i. \tag{31}$$

Then it follows that

$$\Pr[G_i > C] = \Pr\left[H_i > \frac{C}{2N}\right] \tag{32}$$

Therefore, by (31) and (32) we get that

$$\lim_{N\to\infty} \Pr(G_i > C) = 1 \tag{33}$$

as desired.

Authors' contributions

MRM, MMG, and JAG conceived and designed the novel statistical test. MRM, AR, AGP and JG developed the analysis tool. JLSG and ARA implemented the software. MRM, AGP, AR and JG acquired and generated the datasets, analyzed the data and interpreted the results. MRM, AGP and JG wrote the paper. All authors read and approved the final manuscript

Acknowledgements

This study used data from the SNP Resource at the NINDS Human Genetics Resource Center DNA and Cell Line Repository http://ccr.coriell.org/ninds/. We thank the participants and the submitters for depositing samples at the repository.

Funding: This work was supported in part by Agencia IDEA, Consejería de Innovación, Ciencia y Empresa (830882); Corporación Tecnológica de Andalucía (07/124); Ministerio de Educación y Ciencia (PCT-A41502790-2007 and PCT-010000-2007-18); Ministerio de Ciencia e Innovación and FEDER (Fondo Europeo de Desarrollo Regional), grants MTM2008-03679 and MTM2009-07373. Programa de Ayudas Torres Quevedo del Ministerio de Ciencia en Innovación (PTQ2002-0206, PTQ2003-0549, PTQ2003-0546, PTQ2003-0782, PTQ2003-0783, PTQ2004-0838, PTQ04-1-0006, PTQ04-3-0718, PTQ06-1-0002) and Farmaindustria.

Author Details

[1]Department of Quantitative Methods, Technical University of Cartagena, Paseo Alfonso XIII, 50, 30203, Cartagena, Spain, [2]Department of Structural Genomics, Neocodex, Avenida Charles Darwin, 6, 41092 Sevilla, Spain and [3]Department of Quantitative Economy I, UNED, Senda del Rey 11, 28040, Madrid, Spain

Received: 23 June 2009 Accepted: 23 March 2010
Published: 23 March 2010

References

1. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges.** *Nat Rev Genet* 2008, **9(5):**356-69.
2. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: **Multiple rare alleles contribute to low plasma levels of HDL cholesterol.** *Science* 2004, **305(5685):**869-72.
3. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452(7189):**872-6.
4. Shannon CE: **A Mathematical Theory of Communication.** *Bell SystemTechnical Journal* 1948, **27(379-423):**623-656.
5. Dawy Z, Goebel B, Hagenauer J, Andreoli C, Meitinger T, Mueller JC: **Gene mapping and marker clustering using Shannon's mutual information.** *IEEE/ACM Trans Comput Biol Bioinform* 2006, **3(1):**47-56.
6. Li YM, Xiang Y, Sun ZQ: **An entropy-based measure for QTL mapping using extreme samples of population.** *Hum Hered* 2008, **65(3):**121-8.
7. Cui Y, Kang G, Sun K, Qian M, Romero R, Fu W: **Gene-Centric Genomewide Association Study via Entropy.** *Genetics* 2008, **179:**637-650.
8. Zhao J, Boerwinkle E, Xiong M: **An entropy-based statistic for genomewide association studies.** *AJHG* 2005, **77:**27-40.
9. Dong C, Chu X, Wang Y, Jin L, Shi T, Huang W, Li Y: **Exploration of gene-gene interaction effects using entropy-based methods.** *EJHG* 2008, **16:**229-235.
10. Kang G, Yue W, Zhang J, Cui Y, Zuo Y, Zhang D: **An entropy-based approach for testing genetic epistasis underlying complex diseases.** *J Theor Biol* 2008, **250(2):**362-74.
11. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC: **A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility.** *J Theor Biol* 2006, **241(2):**252-61.
12. Moore JH: **Bases, bits and disease: a mathematical theory of human genetics.** *Eur J Hum Genet* 2008, **16(2):**143-4.
13. Zhao J, Boerwinkle E, Xiong M: **An entropy-based genome-wide transmission/disequilibrium test.** *Hum Genet* 2007, **121(3-4):**357-367.
14. Fung HC, Scholz S, Matarin M, Simón-Sánchez J, Hernández D, Britton A, Gibbs JR, Langefeld C, Stiegert ML, Schymick J, Okun MS, Mandel RJ, Fernández HH, Foote KD, Rodríguez RL, Peckham E, De Vrieze FW, Gwinn-Hardy K, Hardy JA, Singleton A: **Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data.** *Lancet Neurol* 2006, **5(11):**911-916.
15. Lehmann EL: **Multivariate Linear Hypothesis.** In *Testing statistical hypothesis* 2nd edition. John Wiley & Sons, Inc, New York; 1986.

doi: 10.1186/1471-2156-11-19
Cite this article as: Ruiz-Marín *et al.*, An entropy test for single-locus genetic association analysis *BMC Genetics* 2010, **11**:19