DOI:
10.1159/000371579

Document Version
Peer reviewed version

Published in final edited form as:
Hum Hered. 2015 ; 79(1): 5–13. doi:10.1159/000371579.

Prioritizing Rare Variants with Conditional Likelihood Ratios

Weili Li[a,b], **Sara Dobbins**[c], **Ian Tomlinson**[d], **Richard Houlston**[c], **Deb K. Pal**[e], and **Lisa J. Strug**[a,b]

[a]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto

[b]Program in Genetics and Genome Biology, The Hospital for Sick Children Research Institute, Toronto, Ont., Canada

[c]Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton

[d]Wellcome Trust Centre for Human Genetics, Oxford

[e]Department of Clinical Neuroscience, Institute of Psychiatry, Kings College London, London, UK

Abstract

Background—Prioritizing individual rare variants within associated genes or regions often consists of an ad hoc combination of statistical and biological considerations. From the statistical perspective, rare variants are often ranked using Fisher's exact p values, which can lead to different rankings of the same set of variants depending on whether 1- or 2-sided p values are used.

Results—We propose a likelihood ratio-based measure, maxLRc, for the statistical component of ranking rare variants under a case-control study design that avoids the hypothesis-testing paradigm. We prove analytically that the maxLRc is always well-defined, even when the data has zero cell counts in the 2×2 disease-variant table. Via simulation, we show that the maxLRc outperforms Fisher's exact p values in most practical scenarios considered. Using next-generation sequence data from 27 rolandic epilepsy cases and 200 controls in a region previously shown to be linked to and associated with rolandic epilepsy, we demonstrate that rankings assigned by the maxLRc and exact p values can differ substantially.

Conclusion—The maxLRc provides reliable statistical prioritization of rare variants using only the observed data, avoiding the need to specify parameters associated with hypothesis testing that can result in ranking discrepancies across p value procedures; and it is applicable to common variant prioritization.

Keywords

Evidential paradigm; Genetic association; Exact testing; Likelihood ratio

Dr. Lisa Strug The Hospital for Sick Children 555 University Avenue Toronto, ON M5G 1X8 (Canada) lisa.strug@utoronto.ca.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Introduction

It is well appreciated that the single-marker analysis approach, which has been successful in identifying common disease-associated variants in genome-wide association studies, has limited power to detect association with rare variants (minor allele frequency, MAF, <5%). To increase statistical power, various methods have been proposed to combine genetic information across multiple rare variants within a gene or region of interest [1–9]. While these 'combined methods' have been shown to be effective as a first step in detecting associated regions, the ultimate goal is localizing individual causal variants within the identified regions that directly affect disease presentation. There is limited statistical literature addressing the difficult task of fine mapping amongst rare variants within previously detected associated genes or regions. Ionita-Laza et al. [10] proposed a sliding-window approach to scan genes or regions of interest to further detect causal rare variant-enriched sub-regions. A similar sliding-window-based method was also published by Brisbin et al. [11]. However, as with any initial identification of associated regions of interest, fine mapped subregions may still include a large number of individual rare variants that require further refinement. For example, in Ionita-Laza et al. [10], the authors applied one of the combined methods, the variable threshold approach [4], in autism spectrum disorder and identified two significantly associated genes, one of which spans an ~235-kb region on chromosome 2. Even with further fine mapping, the significantly associated sub-region covers a 26-kb region. Therefore, in order to select the most promising variants for follow-up biological experiments, further prioritization on the individual variant level is necessary and statistical information can assist with this additional refinement.

The most commonly implemented approach to prioritize *individual* variants in case-control studies is to rank them based on p values. For rare variants specifically, Fisher's exact p values are often computed since low MAFs can lead to sparse and skewed data. However, the computation of Fisher's exact p values requires the investigator to choose whether the p values will be 1- or 2-sided, and which 2-sided method to implement. The p value rankings for the same set of rare variants can differ substantially across these choices, due to the asymmetry of the discrete hypergeometric distribution underlying Fisher's exact test. Therefore, two investigators with different beliefs or knowledge of the direction of association may select different rare variants for follow-up based on statistical considerations, a problem that we encountered when deciding how to prioritize rare sequence variants from a linkage region in a study of rolandic epilepsy (see Materials and Methods).

Although prior information such as expert opinions, biological knowledge or findings from previous studies may improve prioritization, results heavily depend on what type of prior information is used and how this information is incorporated into the prioritization. Therefore, a prioritization method based solely on the observed data and independent of a-priori beliefs on the direction of association could provide a common objective starting point from which to work.

The evidential paradigm (EP) [12–15], an alternative to Frequentist and Bayesian paradigms for statistical inference, uses likelihood ratios (LRs) to measure statistical evidence in a

given body of data and does not require one to specify hypothesis-testing characteristics or prior probability distributions. The evidential paradigm has favorable operational characteristics, where interpreting evidence via the LR for two simple hypotheses will not lead one to favor an incorrect hypothesis with high probability [16–18], i.e. P_0 (LR > k) for k > 1, referred to as the probability of misleading evidence, M, is bounded by 1/k. The EP has been applied to genetic linkage and association studies [17–20], and here we propose an EP method to prioritize rare variants within an associated region. The calculation of the LR as a measure of evidence in the EP requires the specification of two simple hypothesized values for the parameter of interest, e.g. a null and an alternative value. While the null value is often chosen to represent no difference, for example, no association or no linkage, many strategies are available to choose the alternative value depending on the study context. In clinical trials, the alternative value can be chosen to represent the minimum clinically important difference, or a 2-fold change in studies of differential gene expression [14]. For prioritization in a region of association, choice of a simple alternative across all markers in the region is not obvious. The maximum likelihood estimator, the parameter value that is best supported by the data, would be a convenient choice, although the probability of misleading evidence M is not bounded [14] when the LRs are constructed using the maximum likelihood estimator as the alternative value. However, when the goal of the study is to prioritize variants within an associated region, one is less concerned with elevated probabilities of falsely observing evidence for association versus none. Here we define new operational characteristics tailored to variant prioritization.

Motivated by the EP, we introduce a LR-based measure, maxLRc, for ranking rare variants that does not require one to specify a direction of association or choose among multiple methods to compute 2-sided exact p values. Distinct from the combined methods that are designed to detect associated genes or regions, our goal here is to provide an objective measure of the strength of statistical evidence to use for prioritizing individual rare variants within an associated region. We show analytically that the maxLRc is based on the same underlying model as Fisher's exact test. Via simulation, we show that 1-sided Fisher's exact p values, calculated assuming the same directions of effect for all variants, performed substantially worse than 2-sided exact p values and the max-LRc, unless all causal variants are simulated with the same directions of effects. In addition, we show that the max-LRc outperforms 2-sided Fisher's exact p values in most simulation settings.

Materials and Methods

With a case-control study design, a 2×3 contingency table can be constructed to compare genotype frequencies in cases and controls at each rare variant. Due to the low MAF, the probability of observing the homozygous genotype for the minor allele is extremely low, and it is customary to collapse the data into a 2×2 table by case-control status and presence or absence of the minor allele.

Maximum Conditional Likelihood Ratio

We propose to rank rare sequence variants using a LR-based measure, the maxLRc. Let Y_i denote the case-control status for a random sample of i = 1, …, N individuals, with $Y_i = 1$

for cases and $Y_i = 0$ for controls, and X_i, a dichotomous variable with $X_i = 1$ for carriers of the minor allele and $X_i = 0$ for non-carriers. Assuming a logistic regression model, logit $\{P(Y_i = 1)\} = \beta_0 + \beta_1 X_1$, the parameter of interest is β_1, the log odds ratio, which measures the effect of the rare variant on the case-control status. In a logistic regression model, the exposures are regarded as fixed quantities while the case-control status is considered random. Although in a typical genetic association study design, subjects are selected based on their case-control status, the logistic model provides correct estimates of β_1 invariant to study design [21]. So here we consider the number of carriers and non-carriers of the minor allele, denoted by n_1 and n_2, respectively, as fixed under the logistic regression model. For a 2×2 table, let

$$t_0 = \sum_{i=1}^{N} Y_i$$

reflect the total number of cases and

$$t_1 = \sum_{i=1}^{N} X_i Y_i$$

the total number of cases who carry the minor allele (table 1). Then the full likelihood of the observed data is given by

$$L(\beta_0, \beta_1; Y_i) = \begin{pmatrix} n_1 \\ t_1 \end{pmatrix} \left\{ \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} \right\}^{t_1} \left\{ \frac{1}{1+\exp(\beta_0+\beta_1)} \right\}^{n_1-t_1} \times \begin{pmatrix} n_2 \\ t_0 - t_1 \end{pmatrix} \left\{ \frac{\exp(\beta_0)}{1+\exp(\beta_0)} \right\}^{t_0-t_1} \left\{ \frac{1}{1+\exp(\beta_0)} \right\}^{n_2-t_0+t_1}.$$

To re-express the likelihood as a function of β_1 alone, we condition the full likelihood on the sufficient statistic for β_0, t_0, and the resulting conditional likelihood has the following form

$$L_c(\beta_1 | t_0; t_1) = \frac{\begin{pmatrix} n_1 \\ t_1 \end{pmatrix} \begin{pmatrix} n_2 \\ t_0 - t_1 \end{pmatrix} \exp(\beta_1 t_1)}{\sum_{\mu=\max\{0,t_0-n_2\}}^{\min\{t_0,n_1\}} \begin{pmatrix} n_1 \\ \mu \end{pmatrix} \begin{pmatrix} n_2 \\ t_0 - \mu \end{pmatrix} \exp(\beta_1 \mu)}.$$

Note that conditioning on the sufficient statistic t_0 is equivalent to fixing the total number of cases and, in turn, the total number of controls. Since the number of carriers and non-carriers are also fixed under the model assumption, both margins of the 2×2 table are considered fixed. This is the same assumption underlying Fisher's exact test, and therefore, the conditional likelihood has exactly the same formulation as Fisher's non-central hypergeometric distribution. Let $\hat{\theta}_{MCLE}$ denote the maximum conditional likelihood estimate of the odds ratio. We define

$$max\ LRc = \frac{L_c\left(\hat{\theta}_{MCLE}\right)}{L_c\left(1\right)},$$

and we propose the simple idea of prioritizing rare sequence variants by the maxLRc. Essentially, the maxLRc contrasts the conditional likelihood at the odds ratio value that is best supported by the data and the value of 1, representing no association. Fully specifying the conditional likelihood, we have

$$\begin{aligned} max\ LRc &= \frac{L_c\left(\hat{\theta}_{MCLE}\right)}{L_c(1)} \\ &= \frac{\binom{n_1}{t_1}\binom{n_2}{t_0-t_1}\hat{\theta}^{t_1}_{MCLE}}{\sum_{\mu=\max\{0,t_0-n_2\}}^{\min\{t_0,n_1\}}\binom{n_1}{\mu}\binom{n_2}{t_0-\mu}\hat{\theta}^{\mu}_{MCLE}} \Big/ \frac{\binom{n_1}{t_1}\binom{n_2}{t_0-t_1}}{\sum_{\mu=\max\{0,t_0-n_2\}}^{\min\{t_0,n_1\}}\binom{n_1}{\mu}\binom{n_2}{t_0-\mu}} \\ &= \hat{\theta}^{t_1}_{MCLE}\sum_{\mu=\max\{0,t_0-n_2\}}^{\min\{t_0,n_1\}}\binom{n_1}{\mu}\binom{n_2}{t_0-\mu} \Big/ \sum_{\mu=\max\{0,t_0-n_2\}}^{\min\{t_0,n_1\}}\binom{n_1}{\mu}\binom{n_2}{t_0-\mu}\hat{\theta}^{\mu}_{MCLE}. \end{aligned}$$

Similar to the unconditional likelihood estimate of the odd ratio, $\hat{\theta}_{MCLE}$ does not always exist for 2×2 tables. In general, 2×2 contingency tables can be classified into three cases [22] : (1) complete separation, where the two cell counts in the main- or off-diagonal of the 2×2 table equal zero; (2) quasi-complete separation, where only one of the four cell counts is zero, and (3) overlap, in which case there is no zero cell in the table. While the $\hat{\theta}_{MCLE}$ always exists in the overlap case, it is not well-defined when data is in complete or quasi-complete separation, i.e. $\hat{\theta}_{MCLE}=0$ or ∞; however, we show analytically that the maxLRc remains well-defined in these two cases, and it is equal to

$$\frac{\binom{N}{t_0}}{\binom{n_1}{t_1}\binom{n_2}{t_0-t_1}}\ (\text{see Appendix for proof}).$$

The above result shows that in the two separation cases, the max-LRc is equivalent to the inverse of the hypergeometric probability of the observed 2×2 table.

Fisher's 2-Sided Exact p Value

In the hypothesis testing framework, 1-sided tests may be appropriate and more powerful than 2-sided tests if additional information on the directions of association is known and correct. However, such information is rarely available for all causal and nearby variants in a region of interest. One strategy is to assume the same effect for all variants in a region [9]; however, such an assumption can lead to substantially inferior ranking performance, as observed in our simulations (see online suppl. table S1; see www.karger.com/doi/10.1159/000371579 for all online suppl. material). Therefore, we focus on comparisons of

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

prioritization performance between the maxLRc and 2-sided Fisher's exact p values. There are many strategies to compute 2-sided p values for Fisher's exact test, each with its own advantages and drawbacks. Among these methods, the most popular and widely implemented method is the 'minimum likelihood' approach. In this approach, 2-sided Fisher's exact p value is calculated by summing up the probabilities of all possible tables with the same margins as the observed one, whose associated probabilities are less than or equal to the probability of the observed table. We use this minimum likelihood approach here.

Simulation Study

We performed extensive simulations to empirically compare the prioritization performance between the maxLRc and 2-sided Fisher's exact p value. We considered the number of cases to be 100, 150, 200, 250, 500 and 1,000, with a control:case ratio of 1, 1.5 and 2, for a total of 18 sample size combinations. For each sample size (N), we simulated a total of 100 rare variants, among which, the number of *truly* associated rare variants, Q, was set to be 10 or 20. Each rare variant, regardless of causal or null, was generated with a MAF randomly selected from the uniform (0.005, 0.05) distribution. For truly associated variants, we first considered the scenario where the directions of effect were randomly assigned as positive or negative, and therefore, the proportion of deleterious or protective variants was not fixed. We then considered the scenario where 80% of the truly associated variants have deleterious effects. The effect sizes, $|\beta_1|$, for the Q truly associated variants ranged from 0.41 to 2.5, corresponding to odd ratios of 1.5–10.5 for deleterious variants or 1/10.5–1/1.5 for protective variants. All variants were simulated under Hardy-Weinberg equilibrium and all null variants were simulated assuming an underlying odds ratio of 1. Due to the low MAF, rare variants usually do not show strong linkage disequilibrium with each other [23, 24], and therefore, all variants were simulated independently of each other. For each set of parameter values, we repeated the simulation 1,000 times.

Prioritization Performance Comparison

The comparisons between the maxLRc and Fisher's exact p values were based on three metrics: (1) Kendall's correlation between the 'true' ranking underlying the simulation and the ranking assigned by the maxLRc or Fisher's exact p values; (2) the number of truly associated (NT) variants among the K selected, where K is a pre-specified number of variants to be selected for follow-up, and we considered all integer values of K ranging from 1 to 50, and (3) the average ranking of the collection of truly associated variants.

To evaluate (1), we only used the set of truly associated variants, and the 'true' ranking was defined by the ordering of the absolute values of the underlying effect sizes $|\beta_1|$. Here, a larger Kendall's correlation coefficient indicates better agreement with the 'true' ranking. For (2), we considered the scenario where there are limited resources and only K variants are to be followed-up, and we compared the NT variants in the K selected according to rankings assigned by the maxLRc or Fisher's exact p values. At each value of K considered, we plotted the median of the 1,000 NT values (obtained from the 1,000 repetitions) according to the maxLRc or Fisher's exact p values (NT plot). To compare the overall ranking performance, we summed up NT over all values of K, here denoting this sum by

Europe PMC Funders Author Manuscripts

sumK, and a larger sumK value indicates better overall performance. In (3), for each repetition, we calculated the average ranking of the truly associated variants assigned by each of the maxLRc and exact p values, and we then compared the means of the 1,000 averages.

In addition, we used simulations to compare the performance of the maxLRc and 2-sided Fisher's exact p value under the same sample size configuration as in our motivating study of rolandic epilepsy [26]. Specifically, we simulated rare variants with 27 cases and 200 controls, Q = 10 and randomly assigned directions of association.

Prioritizing Rare Variants in a Study of Rolandic Epilepsy

We illustrate how rankings assigned by the maxLRc and 2-sided Fisher's exact p value could differ substantially for the same set of rare variants using next-generation sequence data from a rolandic epilepsy study. Rolandic epilepsy is the most common epilepsy syndrome in childhood [25], and it is linked to and associated with a 600-kb region on chromosome 11 [26]. To prioritize individual rare variants within this region, we used next-generation sequence data from 27 rolandic epilepsy cases sequenced on the Illumina GAIIX platform, and an independent sample of 200 individuals ascertained for a study of colorectal cancer whole-genome sequenced by complete genomics and made available to us as a control group [27]. After standard quality control analysis, 207 rare variants with a MAF <5% remained for the prioritization. We then ranked the 207 rare variants based on both maxLRc and 2-sided Fisher's exact p value.

Results

When ranking with 1-sided Fisher's exact p values, if the directions of effect for all causal variants are randomly assigned or fixed at 80% positive, 1-sided Fisher's exact p value performs substantially worse than the maxLRc and 2-sided Fisher's exact p value. Only when all causal variants are simulated to have positive effect and 1-sided p values are calculated in the positive direction, do they outperform the other two methods (see online suppl. table S1).

We now focus on ranking performance comparisons between the maxLRc and 2-sided Fisher's exact p values and present results from simulations generated with an equal number of cases and controls, although results for other control:case ratios are similar. The mean Kendall's correlation coefficients, averaged over the 1,000 repetitions, are presented in table 2 for the scenarios where Q = 10 and in online supplementary table S2 for Q = 20. In all cases, rankings assigned by the maxLRc are in better agreement with the underlying 'true' rankings, as indicated by a higher mean Kendall's correlation.

NT plots for simulations with Q = 10 are provided in figures 1 and 2; and those with Q = 20 are provided in online supplementary figures S1 and S2. When the directions of association are randomly assigned, the maxLRc selects more or an equal number of truly associated variants than 2-sided Fisher's exact p value for a given K in almost all cases. When the majority (80%) of the truly associated variants have deleterious effect, the maxLRc performs better or worse depending on the value of K; however, it has a better overall performance, as

indicated by a larger sumK value, in all sample size configurations except when the sample size is small (at N = 200; see online suppl. tables S3 and S4).

The means of the average rankings of the collection of truly associated variants are summarized in table 3 and online suppl. table S5 for Q = 10 and Q = 20, respectively. The means of the average rankings of the collection of all causal variants are smaller using the maxLRc than 2-sided Fisher's exact p value in all cases considered; i.e. on average, the truly associated variants are always collectively ranked higher by the maxLRc. The results from simulations with control:case ratios of 1.5 and 2 are similar to those presented above (data not shown).

Analysis of the rare variants identified in a 600-kb region linked to and associated with rolandic epilepsy illustrated ranking discrepancies between the maxLRc and 2-sided Fisher's exact p value. The top ten ranked variants by the maxLRc, together with their corresponding rankings by 2-sided Fisher's exact p values, are provided in table 4. It is evident that the two methods can prioritize this set of variants differently, disagreeing even on the top ranked variant. The ranking discrepancies can be substantial, for example, rs1806176, which is ranked 5th by the maxLRc, is ranked 12th by the p value approach. Simulation results suggest that the maxLRc outperforms 2-sided Fisher's exact p value under similar sample sizes, with Kendall's correlation coefficient = 0.31 and 0.23, sumK = 269 and 244 and average ranking of the collection of Q = 26.37 and 29.69, for the maxLRc and 2-sided Fisher's exact p value, respectively.

Discussion

The maxLRc does as good or outperforms 2-sided Fisher's exact p value in prioritizing rare variants in most cases that we considered. Across all simulation scenarios, rankings assigned by the maxLRc correlate better with the underlying true rankings, and the collection of all causal variants is always ranked higher by the max-LRc. When only a few variants are to be selected for follow-up, i.e. K is very small, the two methods perform similarly regardless of sample size; and as sample size gets large, the two methods are expected to have equivalent performance. The difference between the two methods fundamentally lies in how they measure the strength of statistical evidence. The maxLRc is based on the distribution of the data in the observed 2×2 table, whereas 2-sided Fisher's exact p values further incorporate the probability of more extreme tables that could have been observed. Which values are to be defined as more extreme depends on whether the investigator is conducting a 1-sided or 2-sided test, a choice the EP approach does not require.

In constructing the maxLRc, we chose to use conditional likelihood for several reasons. First, it allows for elimination of the nuisance parameter β_0. It also ensures that the maxLRc is always well-defined, even in the two separation cases. This property does not hold if a maximized LR is constructed from estimated or profile likelihoods. Finally, the derivation of the conditional likelihood requires the same assumption as Fisher's exact test, which excludes the possibility that the two methods perform differently simply due to different model assumptions.

The maxLRc is a conditional LR. Therefore, 2 × log (maxLRc) follows a χ^2 distribution with 1 degree of freedom asymptotically, under fairly general regularity conditions [28]. This implies that we could calculate asymptotic p values based on the maxLRc despite having sparse data, and the prioritization of variants based on this asymptotic p value would coincide with the rankings provided by the maxLRc. Computing asymptotic p values in the sparse data setting would, of course, be contrary to recommended standard statistical practice from a hypothesis testing perspective.

Although we proposed the maxLRc for rare variant prioritization, this method is applicable to common variants as well, without requiring the genotype categories being collapsed into minor allele carrier status (details in the Appendix). In addition, the method of conditioning on sufficient statistics, which forms the basis of exact conditional logistic regression [29], can be used to eliminate multiple nuisance parameters providing a theory for including covariates in the calculation of the maxLRc. However, as the sample size and number of covariates increase, the computational burden becomes prohibitive for practical applications. In such cases, profile likelihood could be used instead of conditional likelihood to maintain computational efficiency.

In summary, the maxLRc provides reliable statistical prioritization of sequence variants and outperforms the standard method of prioritizing by Fisher's exact p values in the majority of settings we considered. Although the difference in some cases was minimal, using the maxLRc to prioritize variants avoids the need to make arbitrary decisions about hypothesis testing parameters such as whether to compute 1- or 2-sided p values, and which 2-sided p value to compute. Moreover, the computational time is equivalent for the maxLRc and Fisher's exact p values, making it an attractive and easy to implement alternative. The maxLRc is applicable to both rare and common variants and can be easily implemented in R [30], with R code available at strug.research.sick-kids.ca.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Drs. Lei Sun, Jerry Lawless and Paul Corey for helpful discussions during the development of this work. This work was supported by a postgraduate scholarship from the Natural Sciences and Engineering Research Council (NSERC) of Canada and a Canadian Institute of Health Research (CIHR) training grant in Genetic Epidemiology and Statistical Genetics (to W.L.); a NSERC discovery grant, the Ontario Ministry of Research and Innovation early researcher award program and CIHR grant [MOP-258916] (to L.S.); a European Union Marie Curie International Reintegration Award of the Seventh Framework Programme, Waterloo Foundation, Epilepsy Research UK, Charles Sykes Epilepsy Research Trust, NIHR Specialist Biomedical Research Centre for Mental Health of South London and Maudsley NHS Foundation Trust (to D.P.); Cancer Research UK, Bobby Moore Fund for Cancer Research UK (to R.H. and I.T.); Research Network and the NHS via the Biological Research Center of the NIHR at the Royal Marsden Hospital NHS Trust (to R.H.) and core infrastructure support to the Wellcome Trust Center for Human Genetics (to I.T.).

Appendix

Derivation of the maxLRc under the Quasi-Complete and Complete Separation Cases

When a 2×2 table is under complete or quasi-complete separation (see table A1 for examples), the maximum conditional likelihood estimate of the odd ratio, $\hat{\theta}_{MCLE}$, is equal to ∞ (table A1a–c) or 0 (table A1d–f). Without loss of generality, we provide the proof below for the case where $\hat{\theta}_{MCLE}=\infty$.

$$
\begin{aligned}
max\ LRc &= \hat{\theta}_{MCLE}^{t_1} \sum_{\mu=\max\{0,t_0-n_2\}}^{\min\{t_0,n_1\}} \binom{n_1}{\mu}\binom{n_2}{t_0-\mu} \Big/ \sum_{\mu=\max\{0,t_0-n_2\}}^{\min\{t_0,n_1\}} \binom{n_1}{\mu}\binom{n_2}{t_0-\mu} \hat{\theta}_{MCLE}^{\mu} \\
&= \hat{\theta}_{MCLE}^{t_1} \sum_{\mu=\max\{0,t_0-n_2\}}^{t_1} \binom{n_1}{\mu}\binom{n_2}{t_0-\mu} \Big/ \sum_{\mu=\max\{0,t_0-n_2\}}^{t_1} \binom{n_1}{\mu}\binom{n_2}{t_0-\mu} \hat{\theta}_{MCLE}^{\mu} \\
&= \frac{\hat{\theta}_{MCLE}^{t_1}\left\{\binom{n_1}{\max\{0,t_0-n_2\}}\binom{n_2}{t_0-\max\{0,t_0-n_2\}}+\ldots+\binom{n_1}{t_1}\binom{n_2}{t_0-t_1}\right\}}{\binom{n_1}{\max\{0,t_0-n_2\}}\binom{n_2}{t_0-\max\{0,t_0-n_2\}}\hat{\theta}_{MCLE}^{\max\{0,t_0-n_2\}}+\ldots+\binom{n_1}{t_1}\binom{n_2}{t_0-t_1}\hat{\theta}_{MCLE}^{t_1}} \\
&= \frac{\binom{n_1}{\max\{0,t_0-n_2\}}\binom{n_2}{t_0-\max\{0,t_0-n_2\}}+\ldots+\binom{n_1}{t_1}\binom{n_2}{t_0-t_1}}{\binom{n_1}{\max\{0,t_0-n_2\}}\binom{n_2}{t_0-\max\{0,t_0-n_2\}}\frac{\hat{\theta}_{MCLE}^{\max\{0,t_0-n_2\}}}{\hat{\theta}_{MCLE}^{t_1}}+\ldots+\binom{n_1}{t_1}\binom{n_2}{t_0-t_1}} \\
&\xrightarrow[\hat{\theta}_{MCLE}=\infty]{} \frac{\binom{n_1}{\max\{0,t_0-n_2\}}\binom{n_2}{t_0-\max\{0,t_0-n_2\}}+\ldots+\binom{n_1}{t_1}\binom{n_2}{t_0-t_1}}{\binom{n_1}{t_1}\binom{n_2}{t_0-t_1}} \\
&= \frac{\binom{n_1+n_2}{t_0}}{\binom{n_1}{t_1}\binom{n_2}{t_0-t_1}} = \frac{\binom{N}{t_0}}{\binom{n_1}{t_1}\binom{n_2}{t_0-t_1}},
\end{aligned}
$$

where

$$
\binom{n_1}{\max\{0,t_0-n_2\}}\binom{n_2}{t_0-\max\{0,t_0-n_2\}}+\ldots+\binom{n_1}{t_1}\binom{n_2}{t_0-t_1}=\binom{n_1+n_2}{t_0}
$$

by Vandermonde's identity.

Computing the maxLRc for Common Variants

Let Y_i = 1 or 0 represent the case-control status, and X_i = 0, 1 or 2 denote the number of minor alleles that subject i carries, therefore, assuming an additive genetic effect. For easier representation, let us further define n_1, n_2 and n_3 as the number of subjects carrying 0, 1 or 2 copies of the minor allele and r_1, r_2 and r_3 as the number of cases carrying 0, 1 or 2 copies of the minor allele, respectively. The sufficient statistic for β_0 is

$$
S_0=\sum_{c=1}^{3} r_c=\sum_{i=1}^{N} Y_i,
$$

the total number of cases; and the sufficient statistic for β_1 is

$$S_1{=}0 \times r_1{+}1 \times r_2{+}2 \times r_3{=}\sum_{i=1}^{N} X_i Y_i.$$

The conditional likelihood for the data, as a function of β_1 alone, is then

$$L_c\left(\beta_1 | S_0;\mathrm{S}_1\right) = \frac{\binom{n_1}{r_1}\binom{n_2}{r_2}\binom{n_3}{r_3}\exp\left(\beta_1 S_1\right)}{\sum_{\gamma^* \in \Gamma}\binom{n_1}{r_1^*}\binom{n_2}{r_2^*}\binom{n_3}{r_3^*}\exp\left\{\beta_1\left(0 \times \gamma_1^* {+} 1 \times \gamma_2^* {+} 2 \times \gamma_3^*\right)\right\}}.$$

Where Γ includes all possible combinations of $r^* = (r_1^*, r_2^*, r_3^*)$ such that $r_1^* {+} r_2^* {+} r_3^* {=} S_0$ and $0 \times r_1^* {+} 1 \times r_2^* {+} 2 \times r_3^* {=} S_1$.

The maxLRc is then given by

$$\frac{L_c\left(\hat{\beta}_{1\ MCLE} | S_0;\mathrm{S}_1\right)}{L_c\left(1 | S_0;\mathrm{S}_1\right)},$$

with $\hat{\beta}_{1\ MCLE}$ representing the maximum conditional likelihood estimate of β_1.

Table A1

Example 2 × 2 tables under complete separation (a and d) or quasi-complete separation (b, c, e and f); A, B, C, D > 0

	Carrier	Non-carrier		Carrier	Non-carrier		Carrier	Non-carrier
a			**b**			**c**		
Case	A	0	Case	A	0	Case	A	B
Control	0	C	Control	C	D	Control	0	D
d			**e**			**f**		
Case	0	B	Case	A	B	Case	0	B
Control	C	0	Control	C	0	Control	C	D

References

1. Li B, Leal SM. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. PLoS Genet. 2009; 5:e1000481. [PubMed: 19436704]
2. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5:e1000384. [PubMed: 19214210]
3. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010; 34:188–193. [PubMed: 19810025]
4. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010; 86:832–838. [PubMed: 20471002]

5. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. PLoS Genet. 2011; 7:e1001322. [PubMed: 21408211]
6. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89:82–93. [PubMed: 21737059]
7. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rie der MJ, Nickerson DA, Team NGESP-ELP. Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012; 91:224–237. [PubMed: 22863193]
8. Turkmen AS, Lin S. Blocking approach for identification of rare variants in family-based association studies. PLoS One. 2014; 9:e86126. [PubMed: 24465912]
9. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83:311–321. [PubMed: 18691683]
10. Ionita-Laza I, Makarov V, ARRA Autism Sequencing Consortium. Buxbaum JD. Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. Am J Hum Genet. 2012; 90:1002–1013. [PubMed: 22578327]
11. Brisbin A, Jenkins GD, Ellsworth KA, Wang L, Fridley BL. Localization of association signal from risk and protective variants in sequencing studies. Front Genet. 2012; 3:173. [PubMed: 22973297]
12. Royall, RM. Statistical Evidence: A Likelihood Paradigm. Chapman and Hall; London: 1997.
13. Blume JD. Likelihood methods for measuring statistical evidence. Stat Med. 2002; 21:2563–2599. [PubMed: 12205699]
14. Bickel RD. The strength of statistical evidence for composite hypotheses: inference to the best explanation. Statistica Sinica. 2012; 22:1147–1198.
15. Zhang Z, Zhang B. A likelihood paradigm for clinical trials. J Stat Theory Pract. 2013; 7:157–177.
16. Royall RM. On the probability of observing misleading statistical evidence. J Am Stat Assoc. 2000; 95:760–780.
17. Strug LJ, Hodge SE. An alternative foundation for the planning and evaluation of linkage analysis. II. Implications for multiple test adjustments. Hum Hered. 2006; 61:200–209. [PubMed: 16877867]
18. Strug LJ, Hodge SE. An alternative foundation for the planning and evaluation of linkage analysis. I. Decoupling ‘error probabilities’ from ‘measures of evidence’. Hum Hered. 2006; 61:166–188. [PubMed: 16865000]
19. Strug LJ, Hodge SE, Chiang T, Pal DK, Corey PN, Rohde C. A pure likelihood approach to the analysis of genetic association data: an alternative to Bayesian and frequentist analysis. Eur J Hum Genet. 2010; 18:933–941. [PubMed: 20424645]
20. Hodge SE, Baskurt Z, Strug LJ. Using parametric multipoint lods and mods for linkage analysis requires a shift in statistical thinking. Hum Hered. 2011; 72:264–275. [PubMed: 22189469]
21. Breslow, NE.; Day, NE. Statistical Methods in Cancer Research, vol I: The Analysis of Case-Control Studies. IARC Scientific Publications; Lyon: 1980.
22. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika. 1984; 71:10.
23. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 2001; 69:124–137. [PubMed: 11404818]
24. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant… or not? Hum Mol Genet. 2002; 11:2417–2423. [PubMed: 12351577]
25. Shinnar S, O’Dell C, Berg AT. Distribution of epilepsy syndromes in a cohort of children prospectively monitored from the time of their first unprovoked seizure. Epilepsia. 1999; 40:1378–1383. [PubMed: 10528932]
26. Strug LJ, Clarke T, Chiang T, Chien M, Baskurt Z, Li W, Dorfman R, Bali B, Wirrell E, Kugler SL, Mandelbaum DE, Wolf SM, McGoldrick P, Hardison H, Novotny EJ, Ju J, Greenberg DA,

Russo JJ, Pal DK. Centrotemporal sharp wave EEG trait in rolandic epilepsy maps to Elongator Protein Complex 4 (ELP4). Eur J Hum Genet. 2009; 17:1171–1181. [PubMed: 19172991]

27. Derkach A, Chiang T, Gong J, Addis L, Dobbins S, Tomlinson I, Houlston R, Pal DK, Strug LJ. Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. Bioinformatics. 2014; 30:2179–2188. [PubMed: 24733292]
28. Andersen EB. The asymptotic distribution of conditional likelihood ratio tests. J Am Stat Assoc. 1971; 66:4.
29. Cox, DR.; Snell, EJ. Analysis of Binary Data. ed 2. Chapman and Hall/CRC; London; New York: 1989.
30. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna: 2010.

Europe PMC Funders Author Manuscripts

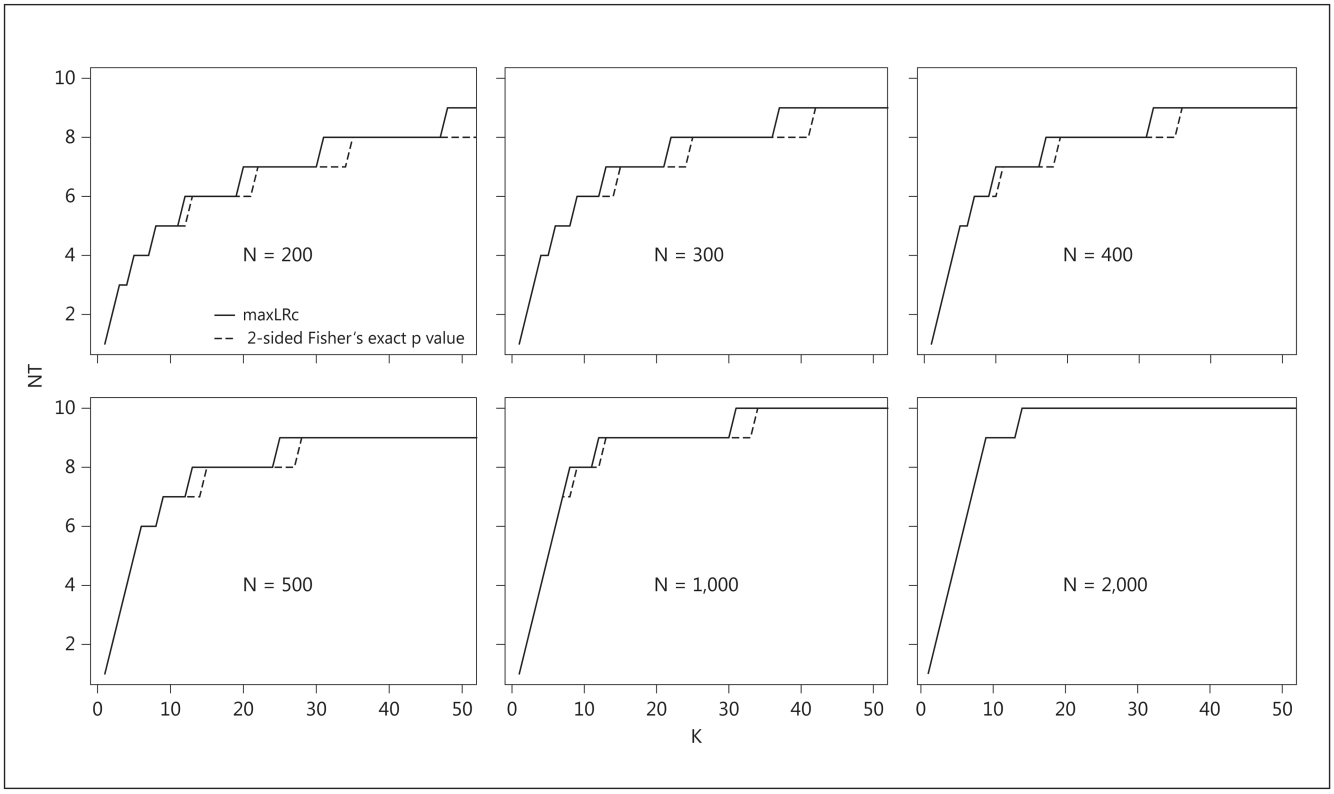

Fig. 1.

NT plots for rare variants simulated with MAF ∈ [0.005, 0.05], Q = 10 and randomly assigned directions of effect. For all sample size configurations, the maxLRc uniformly selected more or an equal number of truly associated variants than 2-sided Fisher's exact p value across all values of K.

Europe PMC Funders Author Manuscripts

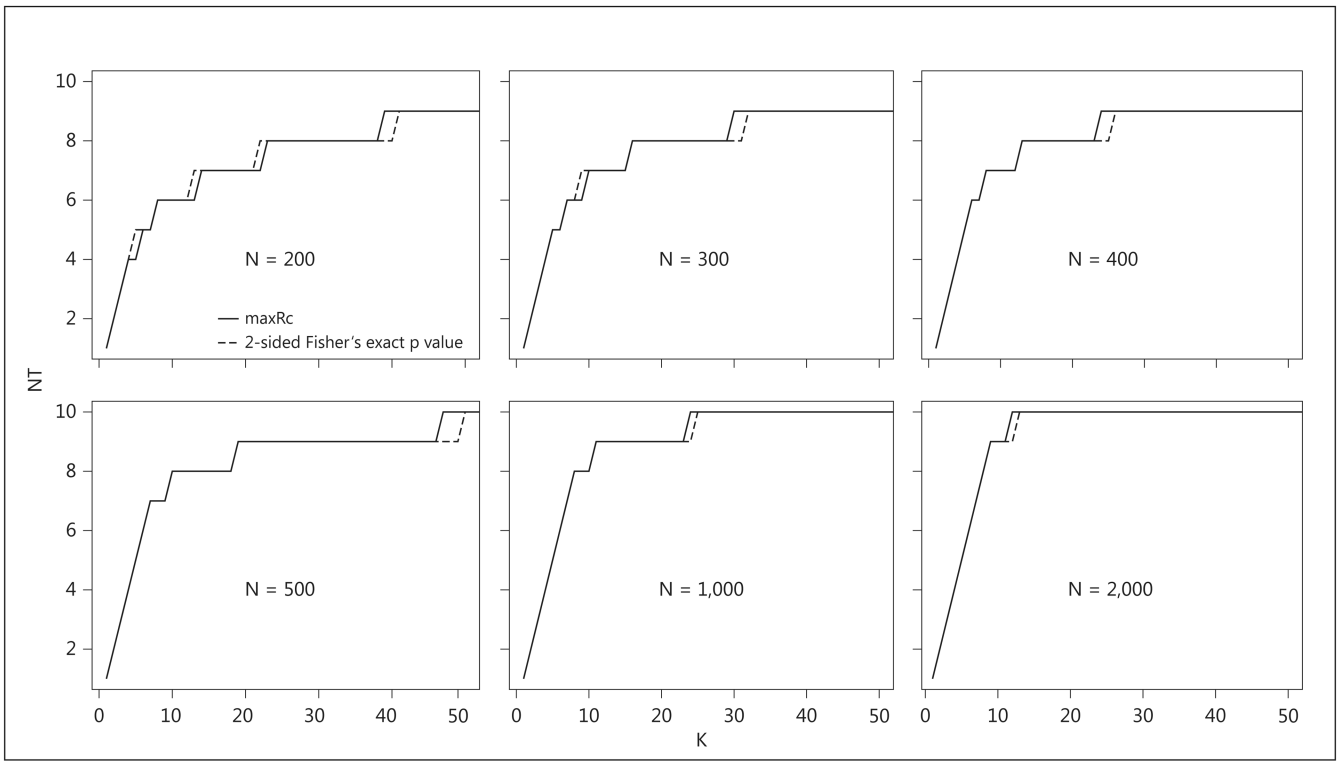

Fig. 2.

NT plots for rare variants simulated with MAF ∈ [0.005, 0.05], Q = 10 and with 80% of the casual variants having deleterious effect. For all sample size configurations, the maxLRc selected more or an equal number of truly associated variants than 2-sided Fisher's exact p value in most cases.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

An example of a 2 × 2 contingency table classified by disease status and minor allele carrier status

	Carrier of minor allele	**Non-carrier of minor allele**	
Case	t_1	$t_0 - t_1$	t_0
Control	$n_1 - t_1$	$n_2 - t_0 + t_1$	$n_1 + n_2 - t_0$
	n_1	n_2	N

Europe PMC Funders Author Manuscripts

Table 2

Mean Kendall's correlation coefficients

Sample size	100 variants (10 causal, 90 null)			
	randomly assigned direction of effect		8 deleterious and 2 protective effects	
	maxLRc	2-sided Fisher's p value	max LRc	2-sided Fisher's p value
100:100	0.39	0.35	0.48	0.46
150:150	0.42	0.40	0.51	0.49
200:200	0.43	0.41	0.53	0.51
250:250	0.43	0.41	0.55	0.54
500:500	0.47	0.46	0.58	0.57
1,000:1,000	0.50	0.50	0.59	0.59

Kendall's correlation coefficients were calculated across 1,000 repetitions, for rare variants generated with a MAF $\in$ [0.005, 0.05], Q = 10 and an equal number of cases and controls. The direction of effect is given for the causal variants.

Europe PMC Funders Author Manuscripts

Table 3

Mean of the average rankings of the collection of truly associated variants

Sample size	100 variants (10 causal, 90 null)			
	randomly assigned direction of effect		8 deleterious and 2 protective effect	
	maxLRc	2-sided Fisher's p value	maxLRc	2-sided Fisher's p value
100:100	19.32	20.63	16.71	16.87
150:150	16.97	18.19	14.59	14.89
200:200	15.25	16.28	13.23	13.55
250:250	13.42	14.27	11.75	12.09
500:500	9.64	10.04	8.62	8.79
1,000:1,000	7.16	7.30	6.79	6.87

The average rankings were calculated across 1,000 repetitions, for rare variants generated with MAF $\in$ [0.005, 0.05], Q = 10 and an equal number of cases and controls. The direction of effect is given for the causal variants

Europe PMC Funders Author Manuscripts

Table 4

Prioritization of rare sequence variants from 27 rolandic epilepsy cases and 200 colorectal cancer controls

Chr	Variant	Base-pair position	MAF	maxLRc	2-sided Fisher's exact p value	Ranking by	
						maxLRc	2-sided Fisher's p value
11	rs6484529	31,724,195	0.003	189	0.0053	1	2
11	rs180775607	31,463,255	0.005	88.08	0.0114	2	4
11	rs11031419	31,605,896	0.028	55.83	0.0052	3	1
11	rs558508	31,800,907	0.029	34.49	0.0081	4	3
11	rs1806176	31,842,323	0.017	28.71	0.0348	5	12
11	rs78174119	31,735,627	0.020	23.10	0.0133	6	5
11	rs4359181	31,759,404	0.033	18.21	0.0154	7	6
11	rs182818125	31,388,793	0.007	17.61	0.0265	8	8
11	rs182363098	31,565,443	0.007	17.61	0.0265	8	8
11	31393303	31,393,303	0.007	16.15	0.0290	10	10

The top ten rare variants, ranked by the maxLRc, are shown together with their corresponding rankings assigned by 2-sided Fisher's exact p value.