Gene Set Enrichment Analysis

Palle Duun Rohde, Stefan McKinnon Høj-Edwards, Izel Fourie Sørensen & Peter Sørensen

2022-04-29

Contents

1 Introduction

From association studies it have been shown that the markers associated with trait variation are not uniformly distributed throughout the genome, but are enriched in genes that are connected in biological pathways [Allen et al., 2010, Lage et al., 2012, Maurano et al., 2012, O'Roak et al., 2012]. This knowledge could be used to construct a statistical modelling framework which quantifies the joint effect of a set of markers located within *e.g.* genes, sequence ontology, biological pathways, protein interactions, or any other type of externally, prior biological knowledge. Such SNP sets can be termed **genomic features**.

The methodology that collectively test a set of genome-wide genetic markers for association with phenotypic variation is known as **gene set enrichment analyses** (GSEA) [Wang et al., 2007, Listgarten et al., 2013]. The idea of aggregating smaller units into larger sets, was originally inspired by gene expression microarray analyses, where individually differential expressed genes are of minor interest, instead the focus is on identifying patterns of differentially expression by aggregating genes exhibiting similarity in their functional annotation [Goeman et al., 2004, Subramanian et al., 2005, Goeman and Bühlmann, 2007]. Various GSEA approaches have been developed through the years, and have been reviewed extensively, see *e.g.* [Wang et al., 2010, Fridley and Biernacka, 2011, Mooney et al., 2014, Leeuw et al., 2016]. Common for all approaches is the test for association between trait variation and the joint contribution of multiple genetic variants aggregated within predefined sets, *i.e.*, genomic features.

Genomic features are collections of genetic variants grouped together based on common biological- or molecular functions, or other characteristics. The aggregation of genetic variants rely on prior biological knowledge from external sources such as protein-protein interactions (*e.g.*, STRING [von Mering et al., 2005]), biological pathways (*e.g.*, KEGG [Kanehisa and Goto, 2000]), gene functions (*e.g.*, gene ontology (GO) terms [Ashburner et al., 2000]), sequence ontologies (*e.g.*, introns, exon and binding sites [Eilbeck et al., 2005]), drug targets (*e.g.*, drug bank [Wishart et al., 2006]), genome-wide expression patterns (*e.g.*, GTEx [Consortium, 2015]), or prior trait associations (*e.g.*, human GWAS catalog [Buniello et al., 2019]). In addition, feature sets can be created from other types of omic data such as metabolomic, proteomic or epigenetic variation.

Naturally, only genetic markers located within the genomic features can be considered in the analysis, thus, an important step is the mapping of variants to the genomic features (Figure 1). This is typically done by grouping all markers within known gene regions. To capture regulatory regions for each gene, upstream and downstream regions are often included, and potentially also any regions in linkage disequilibrium (LD) with the gene. Therefore, some markers may be linked to multiple feature sets.

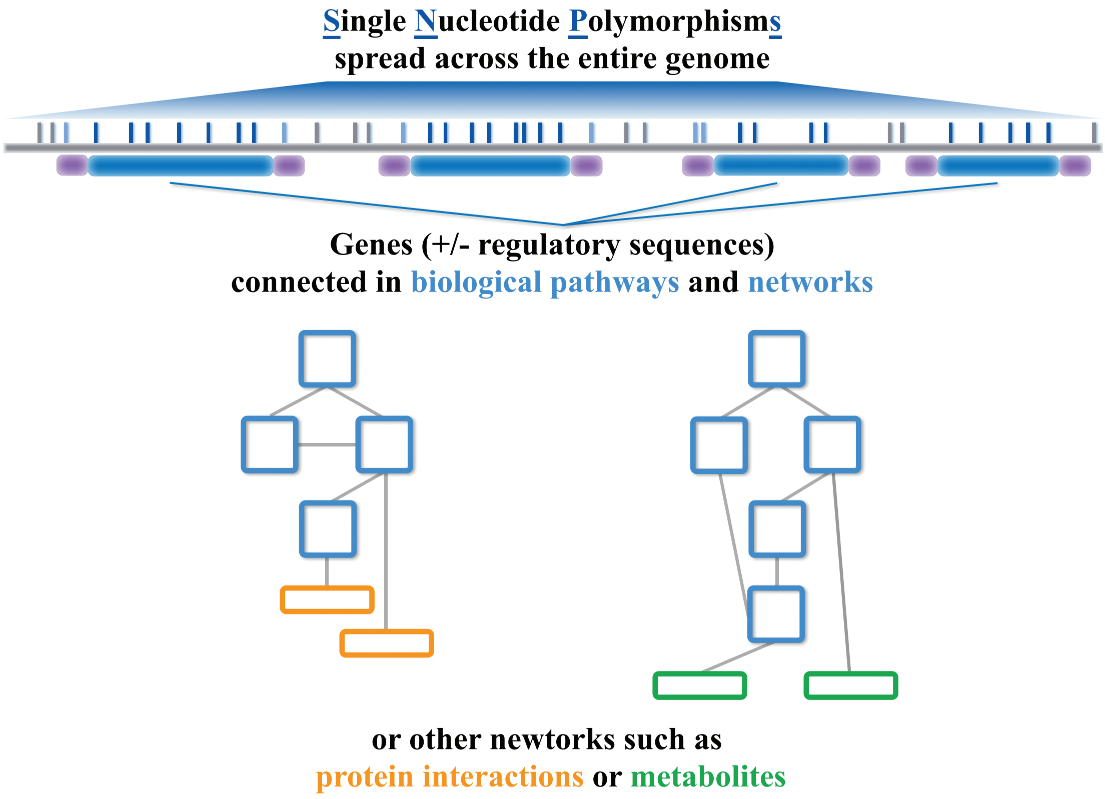
Single Nucleotide Polymorphisms
spread across the entire genome
Genes (+/- regulatory sequences)
connected in biological pathways and networks
or other newtorks such as
protein interactions or metabolites

Figure 1: Graphical representation of genomic feature classes. First, all SNPs located within the same gene region (*e.g.* the transcribed region, dark-blue SNPs) are aggregated. Gene regions can be extended such that SNPs within regulatory regions are included (light-blue SNPs). Second, genes can then be grouped based on prior biological information, *genomic features*, such as genes connected in pathways, or based on other similarities, such as protein networks or common metabolite signatures. The genomic feature classes are thus collections of SNPs that has been aggregated based on shared biological or molecular characteristics.

The degree of new knowledge obtained from the genomic feature analysis strongly depend on the quality and complexity of the genomic feature class. The more reliable the resource is, the more accurate the following results will be. However, if the degree of feature complexity is low, and the quality is high, the outcome might be of minor interests, *e.g.* chromosomal regions are a well-defined feature, but enrichment of certain chromosomes or chromosomal regions might be of less interest. It is therefore important prior to the feature analysis to clearly formulate the scope of the analysis.

Genomic feature models have the prospect to contribute with novel knowledge, but they highly depend upon the availability and specificity of the prior biological information. Unfortunately, such information is not readily available across the tree of life. For model organisms and humans, much information is available, but even for well-studies organisms, such as livestock species, the amount and level of detail is limited. However, the definitions of genomic features are constantly evolving, and new knowledge will continuously be added, also for those organisms which currently are lacking good feature information.

1.1 Different GSEA modelling approaches

The GSEA test can be categorized as belonging to either a *Single-step* or a *Two-step* approach. In the single-step approaches, a genomic feature is modeled by a single model. The estimated effects are then evaluated, either by the properties of the model (*e.g.*, score based statistics) or by comparing the model to a null hypothesis. The set of markers are modeled as a *joint* contribution to a phenotypic trait, by including them as an extra random effect. In the two-step approaches, a single model is used to calculate test statistics on all the markers' effects (*i.e.*, from linear regression or linear mixed models). The test statistics described below all attempt to determine whether a given set of genetic variants contributes to the observed phenotypic trait.

1.1.1 Null hypotheses

We distinguish between two types of null hypotheses, the *competitive* and the *self-contained* [Goeman and Bühlmann, 2007, Maciejewski, 2013]. The self-contained is the easiest and corresponds to determining whether a genomic feature, by it self, does not display any association to the phenotypic trait. This is usually done by defining that the variance component or predicted effect equals zero.

The competitive corresponds to determining whether the degree of association within a genomic feature is the same as outside the genomic feature.

Naturally, the choice of null hypothesis affects the choice of test statistic, but also the biological interpretation of the significance of a finding. The self-contained may be preferable over a competitive, as it has more power [Goeman and Bühlmann, 2007], and the biological interpretation is simpler, as it determines whether there is or there is no association.

1.1.2 Evaluating the test statistics

Once a test statistic has been calculated, it needs to be evaluated to determine whether the genomic feature of interest is significant. This is done by finding the test statistic's position within a distribution, allowing us to evaluate the probability of finding a test statistic of the given magnitude by chance

We distinguish between three types of distributions; the exact, the approximate, and the empirical found distribution.

The *exact distributions* (e.g. hypergeometric test) are derived from the test statistic itself. They might seem to be the preferred, but only if the test statistic actually does describe the desired property being tested.

The *approximate distributions* (e.g. χ^2) relies on that some distributions approximate each other under certain conditions. We can then replace an intangible expression with a simpler, but when being applied to actual data, the conditions are 'bent' into place.

The *empirical distributions* are the brute-force 'when-all-else-fails' solutions we attend to, when the other distributions are too computational demanding, or the conditions for approximating seem to strongly bent. Usual methods for obtaining these are bootstrapping or permutation routines, but caution should be taken under which conditions the routines are performed.

2 Single-step approaches

In the single-step approaches, the data is fitted to the model with AI-REML to obtain estimates of the variance components ($\hat{\theta}$) and the likelihood, which is used to compare nested models with a likelihood ratio test (LRT).

The LRT, Wald's test, and Rao's Score test, can be referred to as 'The Holy Trinity' [Rao, 2009], and are all related to the likelihood and the first and second derivation (Figure 2). The first derivative gives the slope of the function [1], and the second derivative is related to the uncertainty of the estimated variance component.

[1] The first derivative of the likelihood is also referred to as the 'score', which is the basis of the Score based statistic.

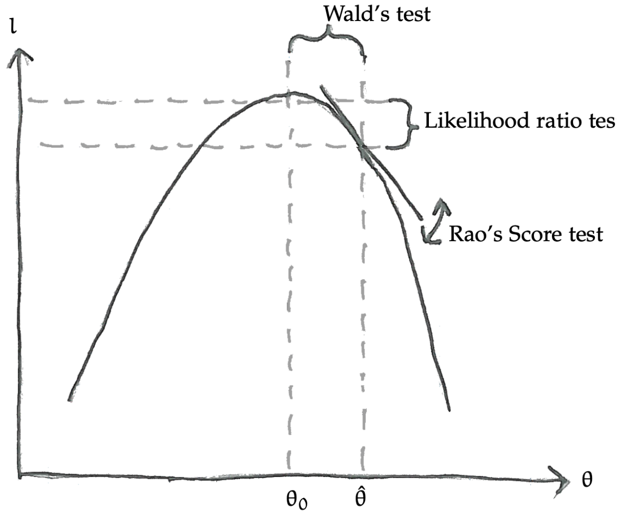

Figure 2: The Holy Trinity of Likelihood Ratio, Wald's and Rao's Score test. The graph displays likelihood as a function of the variance components, maximised at the true value, $\hat{\theta}$.

The LRT compares the model fit between the full model and the reduced model. In Wald's test, the model parameters are fitted using the full model, and test if the estimated variance component is significantly different from a particular value (usually zero). Rao's Score test uses the reduced model (*i.e.*, null model), and estimates the size of improvement in model fit, if an additional variance component was added to the model. Both the Wald and the Rao's score tests are asymptotically equivalent to the LRT, that is, as the sample size becomes infinitely large, the values of the Wald and Rao's score test statistics will become increasingly close to the test statistic from the LRT. A few additional details on the different approaches are given below.

2.1 Likelihood Ratio Test

The likelihood ratio tests (LRT) are used to assess whether a reduced model fits the data better than the full model by comparing the likelihoods of the two models. A high LR indicates that the full model is better at explaining the observed genomic variance than the reduced model with one less variance component. The reduced model has to be nested within the full model, and when REML is used, the two models being compared has to have the same fixed effects, otherwise the two likelihoods are not comparable. The LRT statistic can be derived as:

$$T_{\mathrm{LRT}} = 2\ln\left[\frac{L(\hat{\theta}|\mathbf{y})}{L(\hat{\theta}_r|\mathbf{y})}\right] = -2\left[l(\hat{\theta}_r|\mathbf{y}) - l(\hat{\theta}|\mathbf{y})\right], \tag{1}$$

where $l(\hat{\theta}|y)$ is the log-likelihood for the full model, and $l(\hat{\theta}_r|y)$ is the log-likelihood for the reduced model. When the sample size is sufficiently large, the LRT statistic is χ^2 distributed with κ degrees of freedom, where κ is the difference in parameters between the two comparable models.

2.2 Wald's test

The Wald's test is a parametric test that compares an estimated variance component to some particular value, θ_0, based on some null hypothesis:

$$\frac{(\hat{\theta} - \theta_0)^2}{\mathrm{Var}(\hat{\theta})} \tag{2}$$

The test statistic is assumed χ^2-distributed with one degree of freedom. If the null hypothesis was that the $\{i^{\text{th}}\}$ variance component was equal to zero, the above can be expressed as a quadratic form by

$$T_{\text{Wald}} = (\hat{\theta}_i - 0)' \left[\mathbf{I}_E(\hat{\theta})^{-1}\right]^{ii} (\hat{\theta}_i - 0) \tag{3}$$

where $\left[\mathbf{I}_E(\hat{\theta})^{-1}\right]^{ii}$ is the $\{i^{\text{th}}\}$ diagonal element of the inverse expected information matrix. Wald's test has the advantage, that it only requires fitting and estimating the parameters under the full model. If the test fails to reject the null hypothesis, this suggests that removing the corresponding variance component from the model will not substantially harm the fit of that model.

Wald's test is computed as the parameter estimate divided by its asymptotic standard error. The asymptotic standard errors are computed from the inverse of the second derivative matrix of the likelihood with respect to each of the covariance parameters. The Wald's test is valid for large samples, but it can be unreliable for small data sets. When used on correlated variance components, I_E might not be full rank and therefore not invertible.

2.3 Rao's Score test

Rao's score test requires estimating only a single model that does not include the parameter(s) of interest. Thus, one can test if adding the variance component to the model will result in a significant improvement in model fit, without fitting additional models. The test statistic is based on the slope (or score) of the likelihood function, using model parameters estimated under the null model. If the null model is true, then the slope of the likelihood function is close to zero. If the null model is not true, fixing a variance component to a value will penalise the likelihood.

Instead of calculating likelihoods for both the null and the full model, the first and second derivatives is used to get an indication of the produced change. The Rao's Score test statistic can be formulated as:

$$T_{\text{Rao}} = \left(l'(\theta_i = 0, \hat{\theta}_{-1})\right)' \left[\mathbf{I}_E(\theta_i = 0, \hat{\theta}_{-1})^{-1}\right]^{ii} \left(l'(\theta_i = 0, \hat{\theta}_{-1})\right), \tag{4}$$

where $\left(l'(\theta_i = 0, \hat{\theta}_{-i})\right)$ is the first derivative of the *full model's* likelihood function, calculated using the parameters estimated with the *null model* and the parameter of interest (θ_i) fixed cf. null model. $\left[\mathbf{I}_E(\theta_i = 0, \hat{\theta}_{-1})^{-1}\right]^{ii}$ is the $\{i^{\text{th}}\}$ diagonal element of the inverse expected information matrix, under same conditions as the first derivative. It is possible to use the average between the expected and observed information matrix, i.e. the average information matrix, as it may be easier to compute [Freedman, 2007, Johnson and Thompson, 1995, Madsen et al., 1994, Jensen et al., 1997].

The Rao's Score test has an asymptotic distribution of χ^2 with number of parameters in $\hat{\theta}_i$ as degrees of freedom when the null hypothesis is true. Some issues related to the test statistic may occur if the information matrix is not positive definite which can happen if the null hypothesis is true [Freedman, 2007].

2.4 Score based statistics

There are several different score-based statistics that also are derived from the first derivative of the likelihood. The score statistic can be written as

$$T_{\text{Score}} = \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Z}\mathbf{g}_i\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{5}$$

which under the null hypothesis $H_0 : \sigma_i^2 = 0$ should be close to zero. If the parameters are estimated under the null model,the score statistic for a group of markers i is:

$$T_{\text{Score}} = \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}_0^{-1}\mathbf{Z}\mathbf{g}_i\mathbf{Z}'\hat{\mathbf{V}}_0^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{6}$$

Utilizing that $\hat{\mathbf{P}}\mathbf{y} = \hat{\mathbf{V}}_0^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\mathbf{e}}$, T_i can be computed as:

$$T_{Score} = \frac{1}{2}\hat{\mathbf{e}}'\mathbf{Z}\mathbf{G}_i\mathbf{Z}'\hat{\mathbf{e}} = \frac{1}{2}\hat{\mathbf{e}}'\mathbf{Z}\frac{\mathbf{W}_i\mathbf{W}_i'}{m_i}\mathbf{Z}'\hat{\mathbf{e}}, \tag{7}$$

where the latter expansion is done for the subset of markers. This is computational simple, and also easy to derive an empirical distribution of the score statistic under both the competitive and self-contained null hypothesis.

3 Two-step approaches

In the first step, a test statistic for the association (*e.g.*, t-statistics) of individual markers with the trait phenotype is obtained from traditional single-marker regression (can also be from a mixed model or a Bayesian linear regression). In the second step, for each set of markers being tested, a summary statistic is obtained. For each set an appropriate summary statistic measuring the degree of association between the set of markers and the phenotypes is computed.

3.1 Gene set statistics

Determination of association of individual markers is based on a single marker test statistic such as the t-statistics and a threshold for this statistic.

Let m denote the total number of markers tested, m_F is the total number of markers belonging to the set of interest, m_A is the number of associated markers, and m_{AF} is the number of associated markers belonging to the feature. Thus m, m_A, and m_{AF} are fixed.

We consider two properties of a marker; 1) to be associated to the phenotypic trait, and 2) belong to the genomic feature of interest. Let H_0 denote the null hypothesis, that the two properties of a marker are independent, or equivalently that the associated markers are picked at random from the total population of tested markers. Rivals et al. [2007] show that this can be formulated and tested in a number of ways.The different tests can be evaluated using an exact (Hypergeometric), approximate (χ^2), or empirical distribution (T_{Sum}) under the null hypothesis.

3.1.1 Hypergeometric test

The total number of markers that belong to the genomic feature of interest and that are associated to the trait phenotype can be computed as

$$T_{\text{Count}} = m_{AF} = \sum_{i=1}^{m_F} \mathbf{I}(t_i > t_0) \tag{8}$$

where t_i is the $\{\text{i}^{\text{th}}\}$ single marker test statistics, t_0 is an arbitrary chosen threshold for the single marker test statistics, and I is an indicator function that takes the value 1 if the argument ($t_i > t_0$) is satisfied.

The number of associated markers that belong to a genomic feature, m_{AF}, can be modelled using a Hypergeometric distribution that has a discrete probability distribution that describes the probability of m_{AF} successes in m_F draws without replacement (can only be drawn one time) from a finite population of size m containing exactly m_A successes. Thus if the null hypothesis is true (associated markers are picked at random from the total population of tested markers), then the observed value m_{AF} is a realization of the random variable M_{AF} having a hypergeometric distribution with parameters m, m_A, and M_F, which we denote by $M_{AF} \sim \textsc{Hyper}(m, m_A, m_F)$.

However, the hypergeometric test assumes that the markers being sampled are independent, a rather strong assumption in genetic data. Therefore, the hypergeometric test might not correctly identify significant association, but instead associated markers that are strongly correlated [Goeman and Bühlmann, 2007].

3.1.2 χ^2 test

The second summary statistic is based on a χ^2 test. Let the observed data be presented in a contingency table where each observation is allocated to one cell of a two-dimensional array of cells according to the values of the two outcomes:

Table 1: Contingency table for χ^2 test in two-step approach.

	In feature	Not in feature	Total
Associated	m_{AF}	m_{AnF}	m_A
Not associated	m_{nAF}	m_{nAnF}	m_{nA}
Total	m_F	m_{nF}	m

Let again H_0 denote the null hypothesis that the property to belong to the genomic feature of interest, and that to be associated, are independent. If the occurrence of these two outcomes are statistically independent, we expect the number in the $\{\text{ij}^{\text{th}}\}$ cell to be $f_{ij} = \frac{m_i m_j}{m^2}$. Based on this expectation we can compute the following summary statistic:

$$T_{\chi^2} = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(m_{ij} - m \cdot f_{ij})^2}{m \cdot f_{ij}} \tag{9}$$

where f_{ij} is the observed frequency in the contingency table. This is called the χ^2 test for independence and it has been shown that the T_{χ^2} variable is asymptotically χ^2 distributed with one degree of freedom [Wackerly et al., 1996, Rivals et al., 2007]. The alternative hypothesis corresponds to the variables having an association or relationship, where the structure of this relationship is not specified.

In summary, under the null hypothesis that the probability of a marker belonging to a genomic feature is independent of being associated to the trait phenotype (i.e. $p_{AF} = p_{nAF}$), the exact distribution of M_{AF} is the hypergeometric distribution $M_{AF} \sim \textsc{Hyper}(m, m_A, m_F)$. This distribution can, if m is large, be approximated with the bionomial distribution $M_{AF} \sim Bi(m_A, m_F/m)$. If the two samples, are large, it is also possible to exhibit an approximately normal variable Z or its square $D^2 = Z^2$, the latter being hence approximately χ^2 distributed with one degree of freedom.

One of the differences between the hypergeometric and χ^2 test statistic is that the latter implicitly distinguishes between over- or under-representation, i.e. the squared difference between the expected and observed counts for all the 4 cells contribute to the T_{χ^2} test statistic. It is possible to test for both over-representation ($p_{AF} > p_{nAF}$) or under-representation ($p_{AF} < p_{nAF}$).

Both tests are potentially of interest for understanding the genetic basis of complex traits. If the number of associated markers is very small in the genomic feature then it may be interpreted as selection/highly conserved region. If the number of associated markers is large in the genomic feature then this may indicate we have identified an important feature underlying the genomic variance of the trait.

In cases where both over-representation and under-representation of genomic features are of interest then it is generally most appropriate to consider a two-sided test. It is also possible to define more detailed and specific hypothesis such as testing whether the associated markers contribute negatively or positively to the trait of interest.

However, there is the arbitrariness of the threshold for determining 'significantly associated', no matter how it is chosen and markers whose test statistics differ by a tiny amount may be treated completely differently.

By design this test will have high power to detect association if the genomic feature harbour markers with large effects, but it will not detect a situation where there are many markers with small to moderate effects [Newton et al., 2007]. In this case, it is more powerful to use a summary statistic such as the mean or sum of the test statistics for all markers belonging to the same genomic feature.

3.1.3 $T_{\mathbf{Sum}}$

As noted above, if the phenotypic trait of interest is governed by many markers with small to moderate effects, counting 'significantly associated' markers neglects a lot of information. We therefore consider the third summary statistic

$$T_{\text{Sum}} = \sum_{i=1}^{m_F} t_i \tag{10}$$

where t_i is a test statistic for the $\{\text{i}^{\text{th}}\}$ marker. There are number of choices for t_i such as likelihood ratio, the score based statistic, or the predicted marker effects, and they might be transformed by e.g.squaring. The nature of T_{Sum} is therefore difficult to describe in terms of exact or approximate distributions, and is included here as an intuitive example where empirical distributions are useful.

3.2 Permutation versus exact and asymptotic test

If we can derive an exact distribution of test statistic under the null hypothesis then we can use this to determine the level of statistical significance for the observed test statistic. The advantage of this is that it is computationally fast and that it works better if the sample size (i.e. n number of observation) is small. However, many of the test statistics are derived based on an asymptotic distribution. If the sample size is small the asymptotic formula's used to calculate the p-value may not be correct. In this case a different approach could be to find the p-value using a permutation method.

A drawback of the permutation method is that it is hard to demonstrate very low p-values. Showing that a p-value is lower than 10^{-7} for example, needs at least 10^7 permutations. Often if the sample size is small, the total number of permutations is not large enough to attain very low significance levels.

The manner of which we permute the data is not arbitrary, but depends on the nature of the null hypothesis being tested. Goeman and Bühlmann [2007] classified the null hypotheses as either *self-contained* or *competitive.*

A self-contained null hypothesis assumes that the marker, or set of markers, is not associated to the phenotypic trait, or has an effect without comparison to other markers of sets. I.e. the similarity between observations and genetics is incidental. To obtain an empirical distribution of the test statistic under a self-contained null hypothesis, we can shuffle the observations thus breaking the link between observations and genetics. This can be referred to as a subject-randomisation approach [Goeman and Bühlmann, 2007], but we refer to it as a 'permutation' approach. However, if using models with multiple random effects, where the association between only one of the random effects is in question, shuffling the observations would break the link for all random effects, rendering the permutations useless. In this case, care should be taken to permute the link between the observations and the random effect in question.

A competitive null hypothesis assumes that the marker, or set of markers, is not *more* associated than any other marker or set of markers. An empirical distribution for a competitive null hypothesis is then obtained by sampling random sets of markers. However, all parameters that might influence the test statistic must be the same. I.e. if the number of markers influence the test statistic, the same number of markers must be sampled repetitively to form the random sets. And if there is an inherent structure between the markers in the set, this structure should be present for the random sets.

References

Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I. Berndt, Michael N. Weedon, Fernando Rivadeneira, Cristen J. Willer, Anne U. Jackson, Sailaja Vedantam, Soumya Raychaudhuri, Teresa Ferreira, Andrew R. Wood, Robert J. Weyant, Ayellet V. Segre, Elizabeth K. Speliotes, Eleanor Wheeler, Nicole Soranzo, Ju Hyun Park, Jian Yang, Daniel Gudbjartsson, Nancy L. Heard-Costa, Joshua C. Randall, Lu Qi, Albert Vernon Smith, Reedik Magi, Tomi Pastinen, Liming Liang, Iris M. Heid, Jian'An Luan, and Gudmar Thorleifsson. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467:832–838, 10 2010. ISSN 00280836. doi: 10.1038/nature09410.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A Harris, David P Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C Matese, Joel E Richardson, Martin Ringwald, Gerald M Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology the gene ontology consortium*. *Nature Genetics*, 25:25–29, 2000. URL http://www.flybase.bio.indiana.eduhttp://fruitfly.bdgp.berkeley.eduhttp://genome-www.stanford.eduhttp://www.informatics.jax.org.

Annalisa Buniello, Jacqueline A.L. Macarthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L. Whetzel, Ridwan Amode, Jose A. Guillen, Harpreet S. Riat, Stephen J. Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A. Hindorff, Fiona Cunningham, and Helen Parkinson. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47:D1005–D1012, 1 2019. ISSN 13624962. doi: 10.1093/nar/gky1120.

The GTEx Consortium. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348:648–660, 2015. URL https://www.science.org.

Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. Open access the sequence ontology: a tool for the unification of genome annotations. *Genome Biology*, 6:R44, 2005. doi: 10.1186/gb-2005-6-5-r44. URL http://genomebiology.com/2005/6/5/R44.

David A Freedman. How Can the Score Test Be Inconsistent? *Am. Stat.*, 61(4):291–295, November 2007. ISSN 0003-1305. doi: 10.1198/000313007X243061. URL http://dx.doi.org/10.1198/000313007X243061.

Brooke L Fridley and Joanna M Biernacka. Gene set analysis of SNP data: benefits, challenges, and future directions. *Eur. J. Hum. Genet.*, 19(8):837–43, August 2011. ISSN 1476-5438. doi: 10.1038/ejhg.2011.57. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3172936&tool=pmcentrez&rendertype=abstract.

J. J. Goeman, S. a. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, December 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg382. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btg382.

Jelle J. Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, February 2007. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btm051. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btm051.

Just Jensen, Esa A. Mantysaari, Per Madsen, and Robin Thompson. Residual Maximum likelihood Estimation of (Co) Variance Components in Multivariate Mixed Linear Models Using Average Information. *J. Indian Soc. Agr. Stat.*, 49:215–236, 1997. URL http://isas.org.in/jisas/jsp/abstract.jsp?title=Residual.

D.L. Johnson and Robin Thompson. Restricted Maximum Likelihood Estimation of Variance Components for Univariate Animal Models Using Sparse Matrix Techniques and Average Information. *J. Dairy Sci.*, 78(2):449–456, February 1995. ISSN 00220302. doi: 10.3168/jds.S0022-0302(95)76654-1. URL http://www.journalofdairyscience.org/article/S0022-0302(95)76654-1/abstract.

Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. doi: 10.1093/nar. URL http://nar.oxfordjournals.org/content/28/1/27.abstract.

Kasper Lage, Steven C. Greenway, Jill A. Rosenfeld, Hiroko Wakimoto, Joshua M. Gorham, Ayellet V. Segrè, Amy E. Roberts, Leslie B. Smoot, William T. Pu, Alexandre C. Pereira, Sonia M. Mesquita, Niels Tommerup, Sø ren Brunak, Blake C. Ballif, Lisa G. Shaffer, Patricia K. Donahoe, Mark J. Daly, Jonathan G. Seidman, Christine E. Seidman, and Lars A. Larsen. Genetic and environmental risk factors in congenital heart disease functionally converge in protein networks driving heart development. *Proc. Natl. Acad. Sci. U. S. A.*, 109(35):14035–40, August 2012. ISSN 1091-6490. doi: 10.1073/pnas.1210730109. URL http://www.pnas.org/content/109/35/14035.

Christiaan A. De Leeuw, Benjamin M. Neale, Tom Heskes, and Danielle Posthuma. The statistical properties of gene-set analysis. *Nature Reviews Genetics*, 17:353–364, 6 2016. ISSN 14710064. doi: 10.1038/nrg.2016.29.

Jennifer Listgarten, Christoph Lippert, Eun Yong Kang, Jing Xiang, Carl M Kadie, and David Heckerman. A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics (Oxford, England)*, 29(12):1526–1533, May 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt177. URL http://bioinformatics.oxfordjournals.org/content/29/12/1526.abstract.

Henryk Maciejewski. Gene set analysis methods: statistical models and methodological differences. *Brief. Bioinform.*, pages bbt002–, February 2013. ISSN 1477-4054. doi: 10.1093/bib/bbt002. URL http://bib.oxfordjournals.org/content/early/2013/02/09/bib.bbt002.full.

Per Madsen, Just Jensen, and Robin Thompson. Estimation of (co)variance components by REML in multivariate mixed linear models using average of observed and expected information. In *5th WCGALP*, pages 455–462, Guelph, Canada, 1994.

Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195, September 2012. doi: 10.1126/science.1222794. URL http://www.sciencemag.org/content/337/6099/1190.abstract.

Michael A. Mooney, Joel T. Nigg, Shannon K. McWeeney, and Beth Wilmot. Functional and genomic context in pathway analysis of gwas data, 2014. ISSN 13624555.

Michael A. Newton, Fernando A. Quintana, Johan A. den Boon, Srikumar Sengupta, and Paul Ahlquist. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat.*, 1(1):85–106, June 2007. ISSN 1932-6157. doi: 10.1214/07-AOAS104. URL http://projecteuclid.org/euclid.aoas/1183143730.

Brian J. O'Roak, Laura Vives, Santhosh Girirajan, Emre Karakoc, Niklas Krumm, Bradley P. Coe, Roie Levy, Arthur Ko, Choli Lee, Joshua D. Smith, Emily H. Turner, Ian B. Stanaway, Benjamin Vernot, Maika Malig, Carl Baker, Beau Reilly, Joshua M. Akey, Elhanan Borenstein, Mark J. Rieder, Deborah A. Nickerson, Raphael Bernier, Jay Shendure, and Evan E. Eichler. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397):246–250, May 2012. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature10989http://www.nature.com/nature/journal/v485/n7397/abs/nature10989.html#supplementary-information.

C. Radhakrishna Rao. Rao score test. *Scholarpedia*, 4(10):8220, 2009. doi: 10.4249/scholarpedia.8220. Revision #121946.

Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23(4):401–7, February 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btl633. URL http://www.ncbi.nlm.nih.gov/pubmed/17182697.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, 2005. URL www.pnas.orgcgidoi10.1073pnas.0506580102.

Christian von Mering, Lars J. Jensen, Berend Snel, Sean D. Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A. Huynen, and Peer Bork. String: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33, 1 2005. ISSN 03051048. doi: 10.1093/nar/gki005.

Dennis D. Wackerly, William Mendenhall, III, and Richard L. Scheaffer. *Mathematical Statistics with Applications*. Duxbury Press, Belmont, 5 edition, 1996. ISBN 0-534-20916-5.

Kai Wang, Mingyao Li, and Maja Bucan. Pathway-based approaches for analysis of genomewide association studies. *American Journal of Human Genetics*, 81:1278–1283, 2007. ISSN 00029297. doi: 10.1086/522374.

Kai Wang, Mingyao Li, and Hakon Hakonarson. Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.*, 11(12):843–854, December 2010. ISSN 1471-0056. doi: 10.1038/nrg2884. URL http://dx.doi.org/10.1038/nrg2884;http://www.nature.com/doifinder/10.1038/nrg2884.

David S. Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34, 2006. ISSN 13624962. doi: 10.1093/nar/gkj067.