Improving reproducibility on tree based multimarker methods: TreeDTh

José Javier Moreno-Ortega and Nuria Medina-Medina and Rosana Montes-Soldado and María Mar Abad-Grau

Abstract Tree-based transmission-disequilibrium tests are valuable tools to perform fine-mapping in the search of genetic factors for complex diseases, as they use evolutionary information to relate haplotypes affecting the disease. However, the number of different haplotype trees exponentially increases with the number of markers used, leading to spurious associations due to sample overfitting. If the usual Bonferroni correction is applied to avoid those spurious associations, true risk variants may also be missed. In this work we considered a different solution to avoid sample overfitting of haplotype trees. It consists of dividing the data set into at least two parts and using one of them to choose the haplotype tree which models the disease, and the other one to assess the statistical significance. As a practical example to evaluate the performance of our proposal, we modified the *TreeDT* algorithm and observed a significant improvement in reproducibility while reducing the type I errors.

José Javier Moreno Ortega
Department of Computer Languages and Systems - CITIC - University of Granada. e-mail: jjmoreno@ugr.es

Nuria Medina-Medina
Department of Computer Languages and Systems - CITIC -University of Granada. e-mail: nmedina@ugr.es

Rosana Montes-Soldado
Department of Computer Languages and Systems - CITIC -University of Granada. e-mail: rosana@ugr.es

María Mar Abad-Grau
Department of Computer Languages and Systems - CITIC -University of Granada. e-mail: mabad@ugr.es

1 Introduction

Genome-Wide Association Studies (GWAS) are a first step currently performed in the search of genetic mutations that increase susceptibility to complex diseases. Fine-mapping is a second step that has to be performed after GWAS have found markers – usually Single-Nucleotide Polymorphisms (SNPs) –, in association with a complex disease in the majority of the available genotype data sets. The first wide selection of candidate genes is usually done by analysing their linkage disequilibrium with the actual susceptibility disease gene, which may have not been sequenced [12] [9]. Fine-mapping can shred more light on to where to find the exact location of that gene, or at least narrow down the selection of candidates found in the first step, increasing the chances of replicating the association found in a different data set [4].

Perhaps the most sensible fine map is the one which considers relationships among haplotypes, such as how they departure from each other because of mutation and recombination. However, as fine maps cover small regions in the genome, recombinations are usually ignored [10] and to infer a haplotype tree representing how mutations took place in the population is a simple yet powerful approach to perform fine mapping.

Several authors have explored this idea in order to extend the classic Transmission/Disequilibrium Test (TDT) [12] for multiple markers. The basic biallelic bimarker TDT only measures differences in transmission of one allele. As the number of different markers increases, the number of different models, defined by combining haplotypes, also increases and many generalizations of the simple TDT may be defined, as those using haplotype trees. ET-TDT [10] uses an unrooted evolutionary tree as the basis for grouping haplotypes. The groups reduce the complexity of the model while capturing the information of the genetic transmission. In Treescan [13], a haplotype tree is estimated using maximum parsimony [5]. The clades of the tree are treated as simple alleles, using the F-statistic from a standard one-way ANOVA to measure the association. ALTree [1], which also uses parsimony to build a phylogeny from the haplotypes, chooses a chi-square test as the statistical analysis method and achieves an improvement when two susceptibility sites are involved. Durrant et al. [4] proposed to create a cladogram using simple hierarchical group averaging techniques based on a distance metric between haplotypes. Then a logistic regression model is applied. In TreeDT [11], genealogical trees are estimated to the left and right of the location of interest. The construction of the tree is based on the prefixes shared by the haplotypes. All subtree sets, up to a limited cardinality, are finally explored using a Z statistic.

All methods explained above can be broadly depicted in two steps. In the first one, the tree explaining the evolution of haplotypes in the population is inferred using the whole data set. In the second one, statistical significance is computed, again, using the entire data set. This scenario leads to a common problem: as the number of different trees exponentially increases with the

number of markers, so it does the chance of sample overfitting due to the fact of learning and testing the model on the same data set. If the number of markers were very low, the problem could be ignored. Thus, in the very extreme situation of only one biallelic marker there would be only one tree with two subtrees, one for each allele, regardless of the algorithm used. In fact, in that situation, the solution would be equivalent to the TDT. The above mentioned sample overfitting is the reason why these methods are hardly reproducible, and only associations found involving one or two markers may be confirmed in different data sets [3].

We propose a new approach in sample testing which benefits from the higher power that larger haplotypes usually achieve, but without detecting spurious associations due to sample overfitting. To do that we chose *TreeDT* [11] and defined *treeDT-holdout (TreeDTh)* based on it. Under the holdout approach two data sets are used, one for training and other for testing. We performed simulations under a wide range of genetic scenarios and observed a remarkable reduction in spurious associations, therefore showing a significant higher reproducibility.

2 TreeDTh

Our work in this paper focused on improving test reproducibility of tree-based TDTs. Our proposal divides the process of finding a disease variant into two independent parts. The first one deals with the creation of the trees, in the exact same manner as the original version and will be explained in section 2.1. The second part uses a new data set to infer a model based on the information gathered in the first phase (section 2.2). For the whole process to take place, we need two data sets, which we create by splitting the original data set into two subsets: S_1 and S_2.

2.1 Phase 1. Creating the best model

TreeDTh creates two trees for each location using S_1, one for the left and one for the right. A location is the potential disease susceptibility locus between two markers. For each left and right tree all possible subtrees are obtained. Then the subtrees are grouped in all possible sets of size one to three. The best set for each side (right and left) is then stored as the best model for that location, considering the best set the one minimizing the p value. Finally the best model is the one corresponding to the location which minimizes the p value between all the locations, which will be considered as the reference location.

2.2 Phase 2. Assessing model performance

To avoid overfitting, the model is updated using the second data set S_2. Its structure will remain the same, but the counts of the haplotypes in the sets will be updated with the new data. For each haplotype h_1 in the model built using S_1 the most similar haplotype in S_2 is found, h_2. Then the counts of h_1 are updated with the counts of h_2. As the similarity measure, we used the length measure [14], which computes the largest number of consecutive matching alleles. The starting marker for the comparison between two haplotypes will be determined by the reference location. The direction of the comparison will be left and right depending on the set we are updating. As a simple example, we can consider two haplotypes of length 6, $h_a = 000000$ and $h_b = 100010$. Now, assuming that the reference location is situated between the third and the fourth markers we come up with two possible comparisons. For the right direction, positions 4 to 6 are compared. The first difference is found at marker 5 and so, markers 5 to 6 are considered different. The distance for the markers on the right is therefore 2. For the left direction markers 3 to 1 are compared. The first difference is found at marker 1. The distance for the markers on the left is then 1. Once the frequencies are updated in the model, it is possible to calculate new p values using the new frequencies. Since the corresponding distributions were calculated and stored in the phase 1, it is sufficient to calculate the appropriate statistics and compare them with the distributions.

3 Data sets

We generated SNP data sets of nuclear families (the parents and a child) to test the performance of our proposal under different criteria: type I errors under population stratification and admixture, power and locus specificity. We used msHOT [6] to draw 1000 realistic populations to test type-I errors and another 100 to test power (it uses the standard of coalescent model with recombination). From these populations, each with 500 family trios, we used trioSample [7] to obtain samples from populations under different criteria as explained above. A more detailed justification of how data sets were generated is explained at the supplementary website (http://bios.ugr.es/treedth).

3.1 Data sets to test population stratification and admixture

We used the same approach considered in several previous works to test population stratification and admixture [16, 15, 7]. Populations were paired so that frequencies of disease alleles were 0.2 and 0.3 for each population at a pair, and minor allele frequencies (MAFs) were 0.5 for the first population and it was parameterizable for the second population: $q \in \{0.1, 0.3, 0.5\}$. From each pair, we generated 9 different data sets with 500 trios, by combining two variables affecting population stratification and admixture: q, the MAF for the second population at each pair, and pp, the proportion of individuals used from the first population of each pair, $pp \in \{0.5, 0.25, 0.17\}$, so that the remaining number of trios up to 500 were chosen from the second population.

3.2 Data sets to test power and locus specificity

Our approach has been previously proposed [7] as a modification from older approaches [16, 15] to allow testing locus specificity and to obtain more realistic data sets to test power by using the coalescent model with recombination to draw populations [6]. Therefore, once populations were generated, one or two disease loci were selected (MAF had to be in the interval $[0.2 - 0.4]$) and three (additive, dominant, recessive) / six (additive, dom-and-dom, rec-or-rec, dom-or-dom, threshold and modified) genetic models were respectively chosen. Relative risk RR, the probability rate of having the disease when disease alleles are carried or not, was also considered as a variable to compare results: $RR = \{1.2, 1.6, 2.0, 2.4, 2.8\}$. A set of consecutive SNPs surrounding one disease locus (recombination $\theta = 0$) was used as markers to compute the statistic and different number of SNPs were considered: $\{1, 2, 4, 6, 8, 10\}$.

To test locus specificity, SNP markers were chosen at different recombination fractions (genetic linkage) from one disease locus: $5e-05, 0.0001, 0.00015$ and 0.0002.

4 Results

4.1 Population stratification and admixture

Once a tree is constructed from a group of haplotypes, the different subtrees generated are a meaningful way of grouping haplotypes. The complexity of the tree, and hence the number of subtrees, increases with the number of markers used. Since *TreeDT* explores all possible subtrees, the problem of

multiple testing must be addressed. The solution proposed in the original algorithm is to use the minP method of Westfall and Young [8], but our experiments have proved it not to be enough, as the number of false positives (type I error) is higher than it should be expected by chance. For example, the proportion of associations found in 1000 executions for data sets generated with parameters $q = 0.3$ and $pp = 0.75$ using an α nominal value to reject the hypothesis of 0.001 was 0.001 in the case of *TreeDTh* and 0.027 in the case of *TreeDT*. The meaning of this result is that *TreeDT* generates false positives, casting doubt on the reliability of its power. Our approach in *TreeDTh* solves this problem. Tables 1 and 2 at the supplementary website contain the results for data sets simulating different situations of population stratification and admixture under the conditions of the null hypothesis.

4.2 Power, locus specificity and reproducibility

Both methods reached similar power when a small number of markers was used, as the model created was quite simple. Figures S1, S7 and S13 at the supplementary website show how power (association at recombination fraction $\theta = 0$) is practically the same for *TreeDT* and *TreeDTh*. However, *TreeDT* (red line) continued to detect association even when the distance to the disease susceptibility locus was increased. In contrast, *TreeDTh* (green line) rapidly dropped association rates as we moved away from the disease susceptibility locus and it reached the nominal α value when testing markers not in linkage with the disease locus. The more markers were used, the more complex the model became and consequently the risk of sample overfitting increased. It is in this scenario that the differences between both methods became more apparent, but it has to be noted that association rates reached by TreeDT at $\theta = 0$ are due not only to power, but also to false positives originated by model overfitting (see supplementary figures S6, S12 and S18 for window size 10 and different genetic models). Therefore, a better way to measure the ability of the tests to detect association is to check its behaviour in a different data set, that is, its reproducibility. The associations found by our method are practically always confirmed whereas the ones found with the original method are not. Figure 1 shows the proportion of associations confirmed in a second data set by both methods. Figures S19 to S36 for the remaining haplotype lengths, relative risks and genetic models can be accessed at the supplementary website.

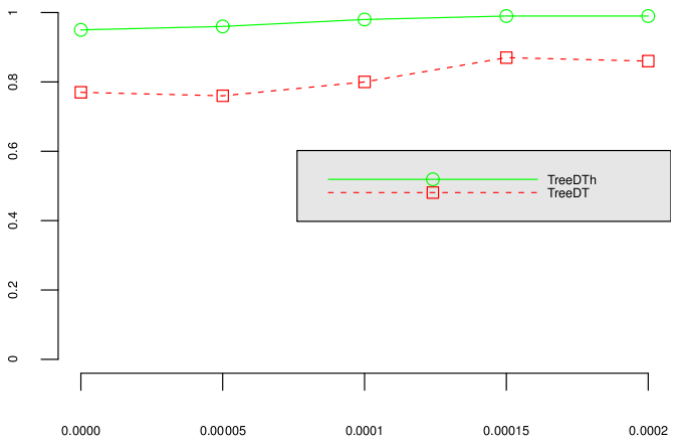

Fig. 1 Comparison of the reproducibility of *TreeDT* (red line) versus *TreeDTh* (green). One locus recessive genetic model, window size 10 and relative risk 2.0.

5 Conclusions

Organizing haplotypes into complex structures like trees based on their genetic information is a very powerful approach to fine mapping. However, the problem of multiple testing because of the huge number of different trees compromises test reproducibility as the model usually overfits to the data used to infer it. The problem of multiple testing and therefore sample overfitting increases with the number of markers used, as a consequence of the raise in the number of different models and their complexity. The result is an increment in association rates which can be explained by two factors: (1) an increase in power because more markers in linkage with the susceptibility locus may better capture association [2, 15, 7] and (2) sample overfitting in which case associations found are not verified on a new data set. However, with our approach, we control sample overfitting so that increases in association rates are only a consequence of truly genetic factors, i.e., power. Therefore, in this paper we have proposed a way to obtain a powerful test without compromising its reproducibility. The *TreeDTh* idea can be extended to other tree based algorithms. Moreover, instead of the holdout approach, multisample techniques such as cross validation may be used to avoid overfitting.

Web resources

A supplementary website has been created for this work at http://bios.ugr.es/treedth, where Figures S1-S36, Tables T1-T2, the software *trioSample* used to obtain data sets upon which to perform simulations (scripts for linux and software

in c++) and *TreeDTh*, the software used to implement the method, are available.

Acknowledgements The authors were supported by the Spanish Research Program under project TIN2007-67418-C03-03, the Andalusian Research Program under project P08-TIC-03717 and the European Regional Development Fund (ERDF).

References

1. Bardel, C., Danjean, V., Hugot, J.P., Darlu, P., Gnin, E.: On the use of haplotype phylogeny to detect disease susceptibility loci. BMC Genetics **6** (2005). ALTree
2. Clayton, D.: A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. American Journal of Human Genetics **65**, 1170–77 (1999)
3. Clayton, D., Jones, H.: Transmission/disequilibrium tests for extended marker haplotypes. American Journal of Human Genetics **65**, 1161–1169 (1999)
4. Durrant, C., Zondervan, K.T., Cardon, L.R., Hunt, S., Deloukas, P., Morris, A.P.: Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. American Journal of Human Genetics **75**(1), 35–43 (2004). CLADHC
5. Eck, R.V., Dayhoff, M.O.: Atlas of Protein Sequence and Structure. National Biomedical Research Foundation (1996)
6. Hellenthal, G., Stephens, M.: mshot: modifying hudson's ms simulator to incorpore crossover and gene conversion hot spots. Bioinformatics **23**, 520–521 (2007)
7. MM, M.A.G., Medina-Medina, N., Montes-Soldado, R., Moreno-Ortega, J., Matesanz, F.: Genome-wide association filtering using a highly locus-specific transmission/disequilibrium test. Human Genetics **128**, 325–44 (2010)
8. P, P.W., Young, S.: Resampling-Based Multiple Testing: Examples and Methods for p-Value adjustment. New York: Wiley (1993)
9. Palmer, L.J., R, L.R.C.L.: Shaking the tree: mapping complex disease genes with linkage disequilibrium. The Lancet **366**, 1223–1234 (2005)
10. Seltman, H., Roeder, K., Devlin, B.: Transmission/disequilibrium test meets measured haplotype analysis: Family-based association analysis guided by evolution of haplotypes. American Journal of Human Genetics **68**(5), 1250–1263 (2001). ET-TDT
11. Sevon, P., Toivonen, H., Ollikainen, V.: Tree pattern mining for gene mapping. IEEE/ACM Trans. Comput. Biol. Bioinf. **3**(2), 174–85 (2001)
12. Spielman, R.S., McGinnis, R.E., Ewens, W.J.: Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (iddm). American Journal of Human Genetics **52**, 506–516 (1993)
13. Templeton, A., Maxwell, T., Posada, D., Stengrd, J., Boerwinkle, E., Sing, C.: Tree scanning: A method for using haplotype trees in phenotype/genotype association studies. Genetics **169**(1), 441–453 (2005). TREESCAN
14. Tzeng, J.Y., Devlin, B., Wasserman, L., Roeder, K.: On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. American Journal of Human Genetics **72**(4), 891–902 (2003)
15. Yu, K., Gu, C.C., Xiong, C., An, P., Province, M.: Global Transmission/Disequilibrium tests based on haplotype sharing in multiple candidate genes. Genetic Epidemiology **29**, 223–35 (2005). DOI 10.1002/gepi.20102
16. Zhang, S., Sha, Q., Chen, H., Dong, J., Jiang, R.: Transmission/Disequilibrium test based on haplotype sharing for tightly linked markers. American Journal of Human Genetics **73**, 566–79 (2003)