Informative missing genotypes: A second research notebook

María M. Abad Grau[1,12]

[1]*Department of Computer Languages and Systems, ETSIIT, Universidad de Granada, Granada, Spain*

[2]*Bioinformatics Group, CITIC ACASES Unit, Universidad de Granada, Granada, Spain*

This work is a continuation of the first research notebook supplied as supplementary material of an unpublished work [1].

Granada, August 1st and 2nd 2017

Introduction Since 2004 my research interests were strongly condensed in only one idea: to be able to predict individual risk to complex diseases from the human genome. June 30th 2017 was the deadline of the second grant I had in order to reach this highly challenging task. As through the many years this idea became an obsession and I and other people around were suffering a lot from that obsession, I decided to put an end to this work at exactly the grant deadline. By that day I and the other co-authors have accepted that the results of the study would not be published. They were far away worse than I expected but at least we were able to find some clues to explain why, after many years and several labs hardly working since the end of the Human Genome project, the dreamed genomic profiling for complex diseases seemed to be unreachable. However, there was another reason we thought this work would be very difficult to be published. The reason, we

believed, was that I wrote a researcher notebook in which I told the research steps given together with personal and religious experiences that, in my opinion, moved me to give any single step in the research work. Because of this we only tried in Nature, a journal strong enough to take a hard risk, and after they refused even to send the work to the reviewrs, I just uploaded the document to my lab website (http://bios.ugr.es) [1] and submitted it to a website for unplished works (still pending of an answer). To be honest, as I was the first one dissapointed with the work because I had much better hopes, I had accepted to have the last results unpublished.

Since that work I have decided to write, everytime I lead a work, a researcher notebook in which I describe the research steps and, if I feel they are related to, personal, relational and religious experiences as well, as now I believe everything is connected. I understand that if results are good, it should not matter whether I use a researcher notebook as the only supplementary material. However, I think the era of the positivism started many years ago and science should take more risks and allow to be influenced by a more recent philosophy so that new paradigms may evolve. In fact, I believe it may be more trustable a research notebook, i.e., a history of a work in which we can understand the thinking the researchers had in each moment they made every experiment than an explanation done after final results are obtained, together with a bunch of numbers and plots they need to provide in order to convince the reviewers they are trustable. For this to be true it is very important to be very systematic and write the notebook as it used to be before the electronic appliances appeared. My commitment is, once an entrance to the notebook is written, never to change the older entrances, except for grammatical and typographical corrections.

The first experiment worked! A few days after the end of the HaploRisk study that could not be published, I knew one of the co-authors was very dissapointed. He was in Spain for 6 months working with me in the improvement of our current results related with a predictor of Multiple Sclerosis (MS) and he was able to outperform AUC results and wanted to publish them. When he left Spain in 2016 I told him to wait for the end of our study before publishing his work, as I suspected there was something wrong in what we were doing under the HaploRisk study and therefore in the source data I gave to him. In fact, by that time the data we were using had missing genotypes that reported AUCs wrongly upward biased [1]. He was very upset when I told him about a month ago he could not publish his work. I could not stop thinking how to solve a situation I felt responsible of. I even thought about trying to publish his work telling that it improved AUC even if AUC was wrongly computed given the way we completed missing data – this was actually the longest "headache" of the HaploRisk study, from 2008 until almost the end of the project, before finding out the cause of the bias and ending up admitting our inability to reach an acceptable AUC given disease prevalence and heritability of MS [1] once we correctly computed it –. However, it sounded to me of very little interest to improve an already biased solution without removing the bias. Soon I suddenly thought about a new experiment, after the many experiments we tried during years under the HaploRisk study. It seemed to me very stupid that I had not came up with it during the last year of the grant. The idea was to use a different training set that the one I have always used in all the experiments tried. In short, the predictive model, an haplotype-based weighted Genetic Risk Score (hwGRS) [2], was learned using a training set from a family trio (parents and one affected offspring) SNP data set with affected offspring. From now on we will call this data set a case trio dataset. An hwGRS computes, for each of the two genome-wide haplotypes an individual has, the

probability of being of high risk for a given trait [2] and it combines them to estimate the individual risk (one of the ways the two probabilities are combined are equivalent to the widely-used wGRS [2,3]). To learn the model, the phase is accurately solved using family genotypes and the training data set is recoded as a set of transmitted (high risk) and untransmitted (low risk) genome-wide haplotypes from parents to affected offspring [2]. It has to be noticed that the main conclusion of the HaploRisk study was that, for a succesful prediction under the current microarray technologies used in our models, the missing data had to be appropiately handle. This conclusion was based on the fact that we found an informative missing pattern in the two Affymetrix arrays and at the same time the only two autoimmune diseases we analyzed [1]. This pattern showed a larger number of missing genotypes in affected (almost all were offspring) than in unaffected (most of the parents) individuals, which was consistent for all the chromosomes.

However, we were not able to distinguish informative missing due to the disease from missing at random [4]. My conjecture inspiring the current work was that, because of this, we could not build accurate risk predictors. In fact, the predictors we built were very little specific, as they classified almost every individual as affected, including most of the offspring from a "control trio data set" (all members being unaffected) that we used to measure specificity. Under this new approach, and in order to help distinguishing informative missing from missing at random, the training data set would not entirely come from the case trio data set but from both the case and the control trio data sets. In fact, from the case training data set we would only select the subset of transmitted (high risk) genome-wide haplotypes and we will disregard the low risk haplotypes (those untransmitted by the parents). For low risk haplotypes we would use the transmitted haplotypes from the control trio data set. Although parents from the case trio data set were all unaffected but one, the

way we completed the missing data was different in parents that in offspring. In fact, we used the cT approach, which favours transmissions of the c or low risk allele, being this allele defined, for each SNP, as the one with the lowest count of transmissions from an independent data subset (that we called "TrainingForHaploRisk" data subset). For this reason it seems to me now evident that we could not have good specificity results, as the training low risk haplotypes came from the parents – at the case trio data set – while the test low risk haplotypes came from the offspring – at the control trio data set –.

I have implemented this approach for MS with the same data set used in the last experiment in the HaploRisk study (see last entrance at the researcher notebook [1], although I did not have the strength to describe it given the negative results). The data set was composed of 854 trios out of the 931 family trios that passed quality control in the IMSGC study [5]. Only 854 were selected for the only reason that we could not recover from our old server the three CEL files of the 77 remaining trios and we needed CEL files in order to obtain genotypes from the scratch. We performed our own calling process because (1) we needed to obtain a model from which any other family trio can be called; (2) we had to make sure the procedure works without normalization, as it should work for single family trios instead of for a data set with a bunch of them; (3) we wanted to make sure no quality control procedure was necessary to make the calling, as we are not performing a Genome-Wide Association Study (GWAS) but a predictive model and the learning machine itself will learn the best model and get automatically rid of low quality SNPs and errors in whole individuals affecting the final outcomes. This way we have a complete procedure for genomic profiling on the shelf, so that the trait risk can be inferred from a new individual only with DNA from the individual and their parents, i.e., calling raw intensities and phasing genotypes

can be performed without the need of a complete new validation data set. Therefore, to obtain individual risk to a given trait we propose to follow the steps that are described in Table **??**. I had started to describe it here but I am using now a Table because this is a very important information that I need to carefully describe and may suffer corrections in the future due to my bad memory. I am doing the same to describe the steps to learn and test a predictor (now Table **??**). [Added in Granada on September 13th, 12:30pm: it may be interesting to describe the last experiment under the HaploRisk study in which we already used our own calling process as described above and why now I think to understand why it turned out to be a new failure. The experiment was just trying to compute (conventional) specificity, i.e., to use unaffected individuals not related with the affected ones. We used the 30 offspring from HapMap CEPH. However, the predictor was still learned using as low risk haplotypes those coming from the unaffected parents from the case trio training data subset, from the IMSGC GWAS, imputing missing genotypes using the cT approach. This approach complete missing genotypes from parents either by the c or the b allele, while offspring are always completed with the c allele, and specificity was "trained" using parents and was tested using offspring from the control data subset. But now I believe there may be another reason, perhaps more important that is not related with missing data: unaffected parents of affected offspring may also have more risk variants in their non transmitted haplotype than unaffected from an independent control data set. This is related with the fact that genetic background may be mandatory to have MS or another complex disease but the environment may be crucial for the disease to appear. This is why to predict within-family specificity shoud be also much more difficult than to predict (conventional) specificity. To use the non-transmitted haplotypes from the parents in the "case trio data set" may introduce a high bias in the predictor.]

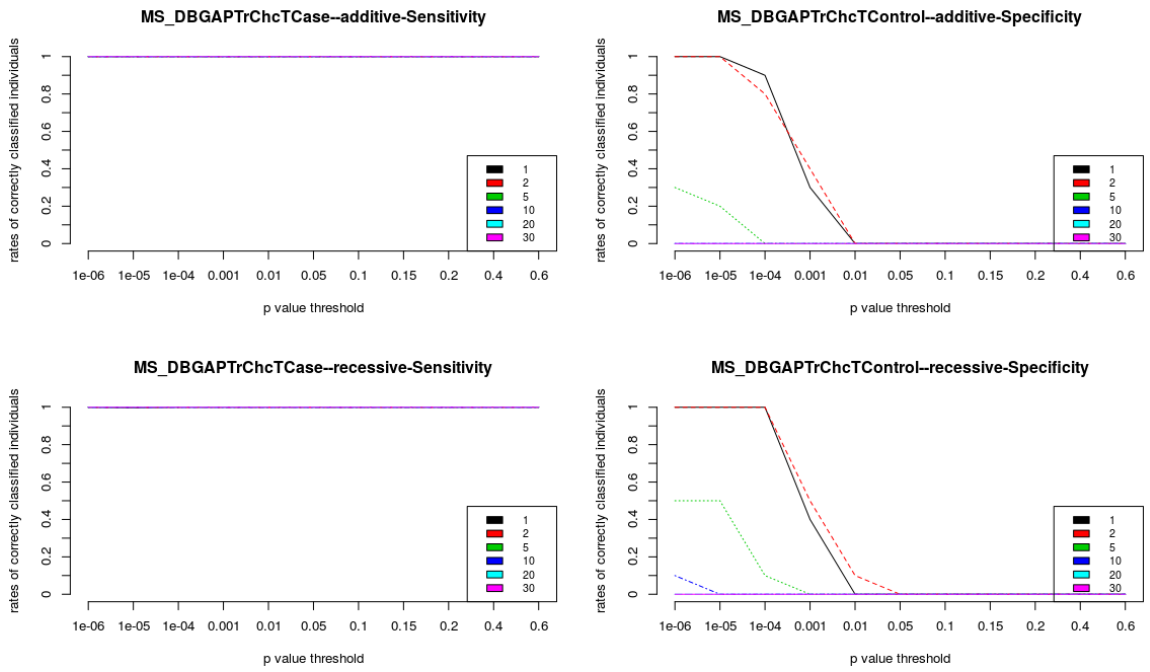

Figure S1: Sensitivity (left plots) and specificity (right plots) from a predictor learned using transmitted haplotypes from a control training data subset. Genotype calling always uses the offspring from the TrainingForHaploRisk data subset. Additive approach is shown on the first row, while the recessive approach is shown on the second row. Rates of correctly classified individuals are shown (y-axis) for different p value threshods (x-axis) and different window lengths (coloured lines).

Results for both, sensitivity and specificity in, respectively, a case and control trio test data sets are shown in Figure1.

I will also use another table to describe the steps I had to add to measure accuracy in a case validation data set that is genotyped with a different array, as it happened with the validation set from an Spanish sample of MS trios we obtained under the HaploRisk study. I have already started following the steps to get the sensitivity of this validation data set.

Calahonda (Granada), August 5th 2017

I have validated the MS predictors learned with the new approach, for both additive and recessive genetic models, with the Spanish MS trio data set, MS_ES we genotyped under the HaploRisk study. Again the conjecture I holded for so many years, that the recessive model should outperform the additive one is not confirmed, as the additive model is never worst and sometimes lightly overcomes the recessive model. Anyway the genetic models reported very good sensitivity results, with almost always the 100% of children (all affected) being correctly classified (sensitivity=1). The worst results was 0.95 for the recessive model. It was obtained with the predictor using sliding windows of size 2 and p value threshold $1e-7$ (Figure 2). [Addition on August 8th: It has to be noted that genotype calling in this data set could not be done using the model learned with the TrainingForHaploRisk case training data subset from IMSGC, as the array used was a more recent one, Affy SNP 6.0 and the genotype call cannot be done with the BRLMM algorithm [6] but with Birdseed (http://www.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx)].

Next experiment I will try will be to validate the predictors using the parents of the Spanish set as control individuals. For this to be correctly set up, their missing genotypes should also be completed as in children, i.e., maximizing the c allele in those positions which remain missing once we complete missing using the other two members of the trio. I believe this is the way children are complete under the cT approach but there may be some exceptions. Therefore, to make sure results I have just obtained for children completing their genotypes under the cT approach, I will generate a new data set without missing genotypes from the original data set with missing data and I will compute sensitivity using children, specificity using parents (as all parents are unaffected)

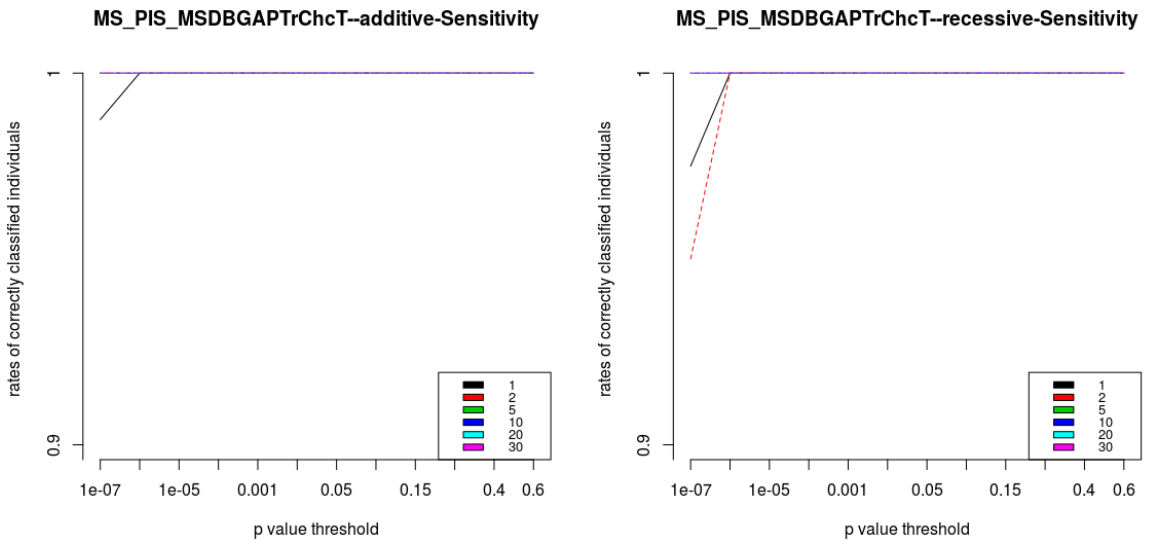

Figure S2: Sensitivity obtained from the validation MS_ES data set. Additive approach is shown on the left, while the recessive approach is shown on the right. Rates of correctly classified individuals are shown (y-axis) for different p value thresholds (x-axis) and different window lengths (coloured lines).

and overall accuracy and AUC using the whole sample.

Calahonda, morning August 6th, 2017

Yesterday evening I had already started with the next experiment I told about in the last notebook entrance. I was first trying to make sure that missing genotypes of offspring completed under the cT approach was equivalent to the c approach. I already suspected there were some differences because of past experiments and now I wanted to know what exactly in which these differences consisted on, and mainly, whether some differences may be due to programming bugs. Very soon I found the first bug. When I was using the cT approach I have decided to remove Mendelian inconsistencies, i.e. to change them by missing genotypes. The problem is that the computer program I made ("completeGenotyesUsingTrios.cpp") removed them before applying the cT approach. Therefore, a trio genotype could completely change. This problem did not exist with the c approach, as the program I wrote to apply this approach was different ("completeGenotypesUsingExternalAlleles.cpp"). To give an example of the bug, in the validation Spanish data set, chromosome 22 and SNP at position 17, the original genotypes for the family trio with family id $\#1$, was $GG/TT/TT$ (father, mother and offspring genotypes), which a Mendelian inconsistency from the father to the offspring. Under the approach cT, the trio genotype was first changed as missing ($??/??/??$) and later replaced by $GT/GT/GG$ following that approach, given that the c allele was G (the c allele is always estimated using the TrainingForHaploRisk data subset from the original case trio data set, IMSGC in the MS predictor). i was trying to patch everthing very fast, as I always badly do and something happened (this time it was not I lost internet connection, I really cannot remember what happened) that made me think I was working against God's will and

moved me to stop, firstly upset and “complaining” by myself but at the moment I started praying I was gicing thanks for these things to happen, after so many years being an agnostic and working as crazy, almost always messing up the research work and mainly my personal life. It had happened a few minutes ago that my mother, a 93 years old widow, had asked me to pick her up in about 30 minutes ad give her a walk and I was tempted to refuse it and keep working. After the few seconds prayer I decided to stop what I was doing and go with her. I still ahd about 15 minutes free, I realized I had not made the plots to show the results I told about in the last entrance, I did them and went to pick up my mom. We actually had a great night, it was very hot, with a beautiful almost full moon and very calmed sea. We ended up at having a light dinner at the beach and I had a very pleasant bath.

This morning when I went to Mass I was thinking at what I was doing yesterday and, in fact, I went too far changing two scripts. At the same time I was changing them in order to skip the bug without changing the cpp program I decided without a calmed thinking (this happens very often to me) to remove one of the 9 and 11 input parameters they had, related with the action to do in the case of a Mendelian error (I was just trying to perform always the same and easiest action: to remove Mendelian inconsistencies replacing the trio gemotype by a missing one). As I use a control version software, the plan is to recover the version I had and perform the changes in the ”completeGenotyesUsingTrios.cpp” program, which is the one which has the bug.

Calahonda, August 6th, 2017, 7:55pm

I have corrected the computer program "completeGenotyesUsingTrios.cpp". When trying to make sure that offpsring under the cT approach were completed the same way as under the c approach I detected another bug. This time it was at "completeGenotypesUsingExternalAlleles.cpp". For years I have felt very guilty of introducing so many bugs in my software, this was one of the main reasons I forced myself to write a research notebook and encourage to every researcher to do the same (please refer to the first entrance of the researcher notebook given as supplementary material [1] to know more about it). I founded an important bug: the computer program just replaced the missing genotypes for the c allele and, in the case a mendelian inconsistencies arised, it went an step back and left the genotype as missing. As an example, with the c allele being T, a trio genotype $CT/CC/??$ remained this way instead of $CT/CC/CT$, i.e., first completing the alleles with no ambiguity (the offspring inherits a C from its mother) and second completing missing using the c allele (T). Once I have corrected the two programs, offspring have exactly the same genotypes either under the cT and under the c approach. Now I believe sensitivity in the validation data set(offspring) should be reduced. But what it worse, the training data set used to learn the predictor (the one with missing genotypes completed using the cT approach) should be generated again, and it should give also a lightly worse predictive capacity. So I will complete again missing genotypes in all data subset under the cT approach, learn the predictor again and test it with the test data set used in the first experiment and with the validation Spanish MS data set, MS_ES.

Calahonda, August 8th, 2017, 1:49am

I could not sleep thinking about this work and I thought it could be of interest to show differences in the total amount of SNPs that were selected under some p value thresholds between the predictor learned without performing missing imputation – a predictor that did not work at all [1,2] – and the predictor under the cT approach. When I started making a table to show these results I realized that I did not have learned the predictor without missing imputation using the genotype calling proposed in this work: to obtain one model (three cluster centers –one for each genotype– and the variance-covariance matrix in the case of the BRLMM Affymetrix algorithm used for the IMSGC MS data set) at each SNP using only the training data subset (TrainingForHaploRisk case data subset) and to use this model to make genotype calls for this data subset and all other data subset we need (TestForHaploRisk case data subset, test case data subset and control case data subset – the 30 CEPH family trios from HapMap –). For the validation (Spanish) data subset, MS_ES, this is not possible, as they were genotyped using a different array, Affy SNP 6.0, and we cannot use the BRLMM algorithm but birdseed, a newer one.

Now the only server that still works is learning that predictor. It will take a couple of days to have results required to make the table I wanted. I will also show specificity and sensitivity from this predictor, although it should be similar, I believe, to the one obtained from the predictor learned without performing missing imputation as well but using genotype calls from the GWAS conducted by IMSGC [5] and the International HapMap Project (IHMP) [7] and selecting only the SNPs both studies had in common after quality control.

Calahonda, August 14th 2017, 9:49am

For a few days I could not work. I did not have internet connection. I was very relaxed since my faith has grown up. I do not complain anymore when things do not work with not a clear cause. I understood it was not time to work and I focused on other things. I had actually to attend my family and it was good. This early morning, as everyday, I checked internet connection. As today it worked again I sent the last scripts to run in background and I will have results very soon.

Calahonda, August 14th 2017, 15:17

I already have results. Results under the additive model are better than when using the recessive model. I think they are also much better than I expected. I am very confusing. Thinking again that I need to take more time to control the new parameters I keep adding to the new experiments. Now I am thinking myself whether these results are better than those obtained before [1] only just because the predictor was learned using cases and controls offspring or because I solved genotype calling by using always the same model learned from the case training subset.

So from this result I should at least to perform a new experiment in which the predictor will be learned from the case trio training subset, as I made before [1] but with the genotype calling made using the same model for all subsets (the one learned from the case training subset).

No missing imputation		cT approach
P value threshold $1e-7$		
Chrom#	total SNPs selected Chom#	total SNPs selected
6	16	
12	1	
total	17	

No missing imputation		cT approach
P value threshold $1e-6$		
Chrom#	total SNPs selected	
3	1	
6	22	
12	1	
17	1	
total	25	

No missing imputation		cT approach
P value threshold $1e-5$		
Chrom#	total SNPs selected	
1	1	
3	2	
5	2	
6	30	
10	3	
12	2	
14	1	

Granada, September 8th 2017

today I finally got sometime to continue this work. I connect to the server to try remember what i was doing last time I worked on it by August middle. I got lost. I read the last entrances of this notebook. As I was already confused with the results when I wrote them, I have just decided to make a new plan again. It should be the last one because I already have good results and I want to put an end of this work. I think results so far are good enough to be publish. I believe they are actually great results that open an amazing research line in genetic profiling of complex diseases. I will stop writing now, I will think about the plan an,d once I make sure it is a good plan to make the last experiments in order to present the results and be able to answer the main and more urgent questions, i will go ahead. I hope to be able to finish this work and to send it to a journal along next week. So far table is incomplete, perhaps I do not complete it of perhaps I do (I will write a new one to be faithful to the rules I propose to present results by using a research notebook in which once an entrance is written I will never change it except for grammatical or typographical issues.

Granada, September 8th 2017, 17:59 (about an hour after the last entrance)

I have realized that since almost a month without working on this project (nothing like this had happened before, since 2004 that I started this project) I have forgotten several things and mainly I do not believe anymore in anyone of all my old conjectures. I think this is very good and I will take advantage of this so that researcher bias can be significantly reduced. I am not going to think about any conjecture for now. I will just go ahead with a plan I have just made up and once I got all results I will try to find an explanation for them. The plan is to obtain performance results from

predictors learned changing the configuration of the following variables: genetic model (additive and recessive), missing imputation approach (no missing imputation, c approach and cT approach) and the training and test data sets and patterns used: learning and test from the case trio data set, learning from the case trio data set and test from case and control data sets using only children and learning and test from case and control data sets using only children. I will start with no missing imputation, i will continue with the c approach and will finish with the c approach. For all of them I will use the two genetic models and all the traning and test configurations allowed by condidering the missing imputation approach used.

Granada, September 8th 2017, 18:53

Once I was preparing the batch script I have realized that I obtained promising results for the Autism disease and I forgot to write about it. These results cut down a past conjecture [1] that only with Affymetrix arrays we observed a significant pattern of higher missing rates for children in case trios (affected offspring, mostly unaffected patterns). In fact, I have forgotten to make a test in the Autism trio data sets we had. I did this test last month and I already got a plot in which we can observe the pattern in all the chromosmes (see Figure 3). I will not try to learn a predictor from this data set. I believe it will take some time to handle genotype calling in a correct way because the genotyping technique is different. In the case the remaining experiments support the thesis that a correct imputation of missing genotypes are essential to build sucesful predictors, it will be better to put aside this task and to approach it as part of a different work. Anyway, I think it is very good news to know that the genotyping made by Illumina does not get rid of this pattern of different missing rates. Perhaps results in CARIES [1], in which this pattern is only true for

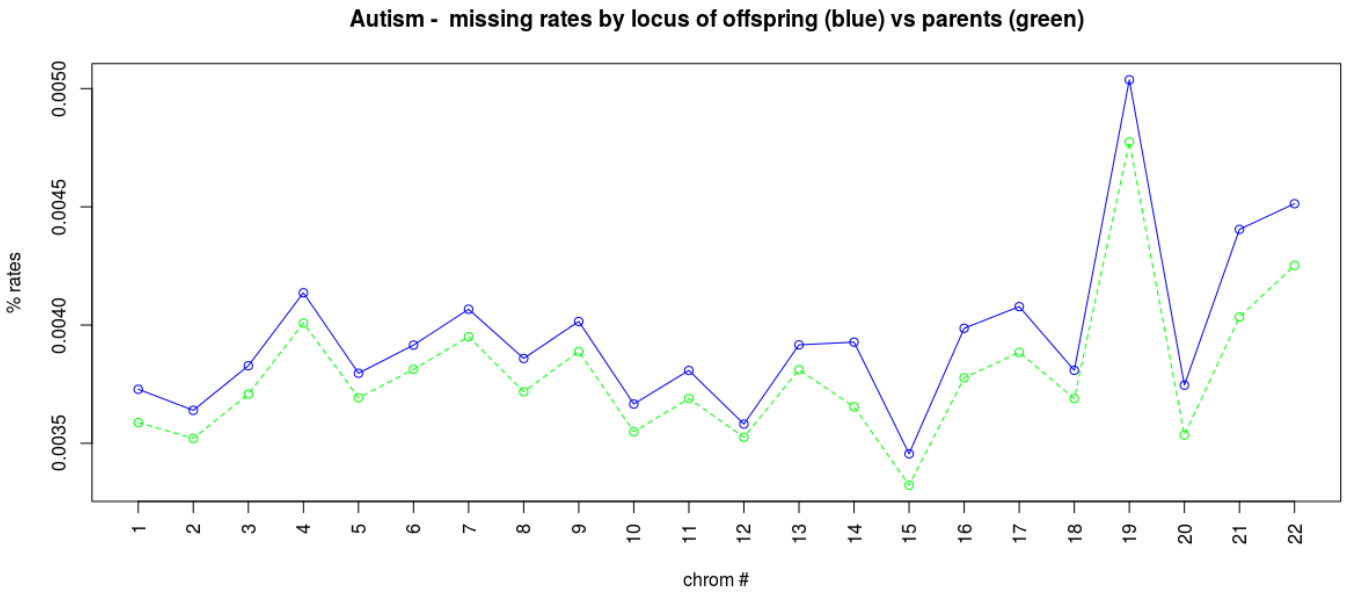

Figure S3: Line plots by chromosomes for parents and offsprings in Autism data set.

some chromosomes, may show that this trait has only those chromosomes as its genetic causes. I tried to made the same study with the 7 diseases of the Wellcome Trust Case Control Consortium (WTCCC) 1 [8] but the final files with the calls had no missing rates and, when I tried to use the raw data I did not success because a lack of time, computer external memory and human resuources to handle it.

Granada, September 9th 2017, 00:11

I have realized, since I concluded that missing genotypes are completed in offpsring the same way in both cT and c approaches, that we do not need the cT approach anymore neither an external data set, as part of the HapMap CEPH data set I was using, in order to use the offspring haplotypes as control or low risk haplotypes in the training set. Under the c approach, the untransmitted haplotypes from parents will be used as low risk haplotypes. Moreover, for the test set we can use

also parental haplotypes, but instead of current parents, we can make up a virtual data set to test predictor specificity by randomly selecting a pair of parental haplotypes and consider both of them as the genotype of a virtual parent. Therefore, we will have n parents and n offspring in a trio data set of n trios. The idea is to check specificity in a generic situation similar to a case control data set. It would be great to have a predictor able to correctly predict as healthy those unaffected parents of affected offspring, and this will actually be a really usable predictor for genetic profiling in the clinic. I will actually measure also specificity in real parents but I believe it will not be very high. I will point out some ideas in the discussion section in order to improve specificity in parents of affected offspring.

Granada, September 11th 2017, 12:35

Iam very happy of using this researcher notebook. Reading the last entrance I have ralized how bad memory I have and how my conjectures have changed again after a month without working. But the new results are helping to refresh my memory and to understand the last improvements I had in august. I think it is wrong what I claimed in the last entrance: "we do not need [] an external data set [] in order to use the offspring haplotypes as control or low risk haplotypes in the training set". What it seems to be true is that, in the case the c approach is enough, is that we could use parents from the control data set (this way we will have double numbers of samples). Figure 4 shows results when using the old approach (to build the model only with a case trio data set). Now I have shown specificity for the parents of affected offspring (parents in the test case trio data subset), what I have called "specificity intra family" and conventional specificity from an independent control trio data set (I used the CEPH parents from Hapmap). It can be seen that, opposite to

it should be, specificity is much worse in the control data set. [Added on September 13th 2017, 13:35: It can also be observed that addtivie and recessive genetic models show an opposite pattern in this experiment: sensitivity is higher that within-family specificity under the additive approach, while it works the other way under the recessive approach. I do not have enough information for even dare to interpret this result neither why within-family specificity are better than conventional specificity. It is true that the predictor specificity was learned using the non-transmitted haplotypes of the parents in the case training data subset and within-family specificity was measured using the remaining parents of the same data set, while conventional specificity was measured using an (independent) control trio data subset.]

Now the plan is to repeat these plots under the c approach. In the case it does not work, I think about two different choices:

1. To use the cT approach to learn the predictor and:
    - To test its sensitivity using offpspring from the test case trio data subset (same genotypes as if using the c approach).
    - To test its (conventional) specificity using the parents of the control trio data set (CEPH HapMap) completed using the c approach.
    - To test within-family specificity using the parents from the test case trio data subset completed under the c approach.
2. If the first option does not work, I will try again what gave the only success so far in this work, obtained in august: To learn the predictor using as cases in the training set the off-

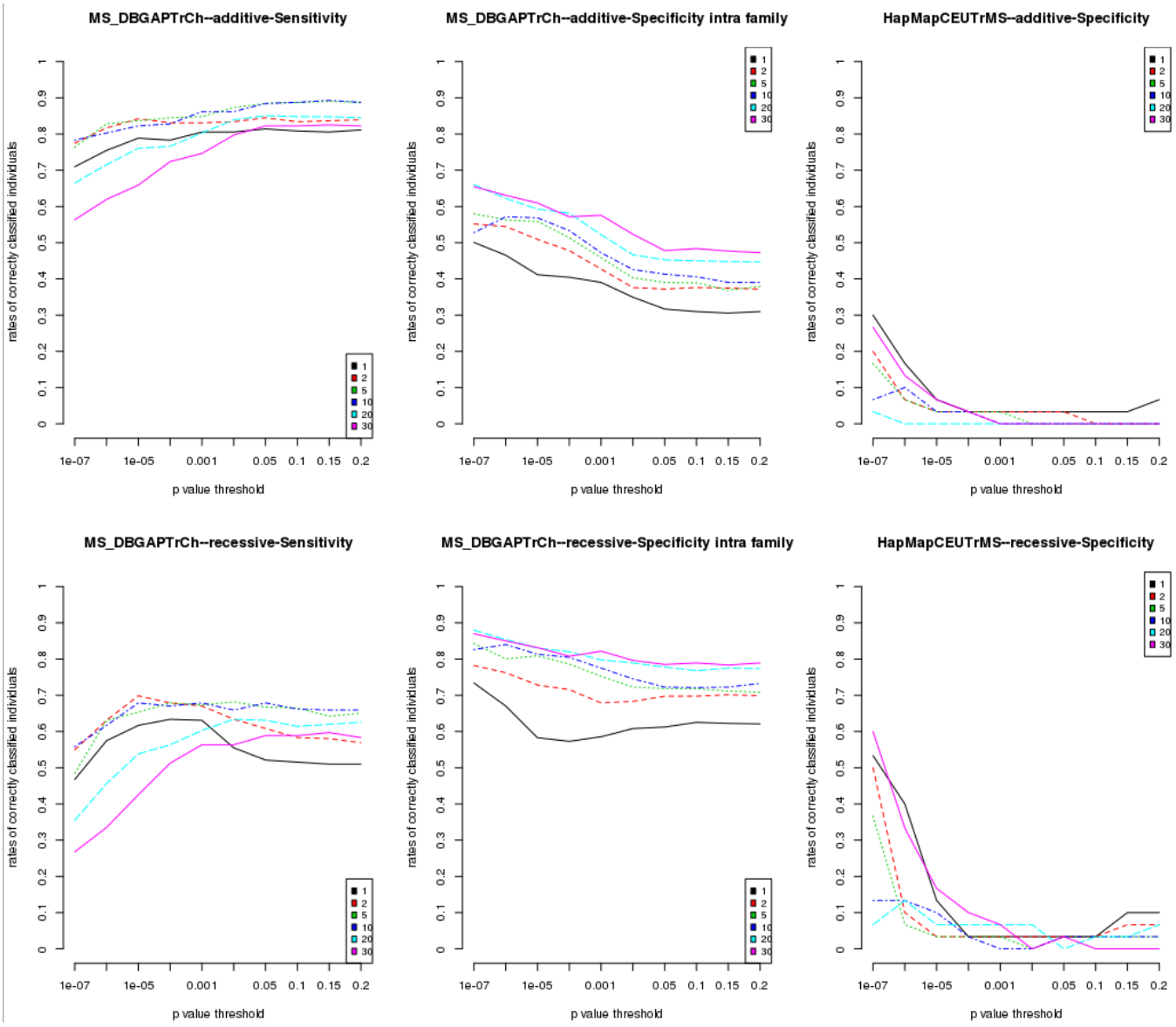

Figure S4: Sensitivy (first column), specificity within families (second column) and (conventional) specificity (right column) from MS_DBGAP with genotype calling using the offspring from the TrainingForHaploRisk data subset. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

spring from the training case trio data subset under the c approach (I actually used the cT approach, offspring must be completed the same way as I showed before that the two approaches complete offspring the same way, however the first step used to compute p values and high and low risk haplotypes for each sliding window size is not the same, so I will use again the cT approach in a third choice for the training case data subset) and as controls the parents from the control test trio data subset under the c approach (I actually used the cT approach, but by considering that I showed that offspring are completed in the same way under the two approaches, this time I will use the c apprach, as this way there will be double number of control individuals: parents can be used instead of offpsirng) and

- To test its sensitivity using offpsring from the test case trio data subset (same way as in the first choice).
- To test its (conventional) specificity using the parents of the test control trio data subset (CEPH HapMap) completed using the c approach.
- To test within-family specificity using the parents from the test case trio data subset completed under the c approach (same way as in the first choice).

3. Same as second option but using the cT approach for the case training data subset. Although offpsring are completed the same way as under the c approach, parents are also used to compute p values and low and high risk haplotypes that will be used dor the most significant sliding windows selected.

4. Same as second option but using the genotypes keeping missing data to compute p values and low and high risk haplotypes that will be used for the most significant sliding windows

selected.

5. Same as second option but using the genotypes keeping missing data for all the training and test data subsets used.

For all of them we will obtain results under the additive and the recessive genetic models.

This way we are considering all alternatives possible under three different parameters:

1. Genetic model: recessive and additive
2. Missing imputation method ($keepMissing$, c, cT)
   (a) for data sets used to compute p values and select sliding windows and high and low risk haplotypes
   (b) for training data subsets (cases and controls) used to train (build) the predictor
   (c) for test data subsets (cases and controls) used to test the predictor
3. Data subset used for training and test specificity (control samples): parents from the case trio data set or parents (or offspring if the cT approach has been used) from the control trio data set

Now that I am remembering what I did last august, what I believe it made the difference when I obtain the first clear "success" was to use data from a control trio data set to train the predictor specificity, i.e., to provide control samples from a control trio data set instead of using

the unaffected parents from the case trio data set to make up the control samples of the training set. If I am right, I will not have any success by repeating the experiment shown in Figure 4 but using the c approach. But I will try anyway as I have very bad memory and also in order to make sure about my last conjectures. I was not interested in making any conjecture to avoid researcher bias but the more I work on this project these days the more I remember.

Granada, September 12th 2017, 11:51am

As it is goint to take another day to have results using the c approach, I will first try to build a predictor using as control samples for the training set, half of the parents of the control data set (HapMap CEPH) keeping missing data. At this stage, I do not even dare to think about any conjecture related with this experiment.

Granada, September 13th 2017, 12:48pm

Now I have results using control samples to train the predictor specificity instead of using the non-transmitted haplotypes of the parents in the case training data subset from the IMSGC data set (see the table above with an explanation of the way to obtaining training and test data subsets and to solve genotype calling so that the procedure can be directly applied to a new individual providing also the genotypes of their parents). I used 40 out of the 60 parents from the HapMap CEPH data set as control training data subset and the remaining 20 parents to test (conventional) specificity. As it is shown in Figure 5, conventional specificity is very high for small p-value thresholds. However, within-family specificity is much shorter.

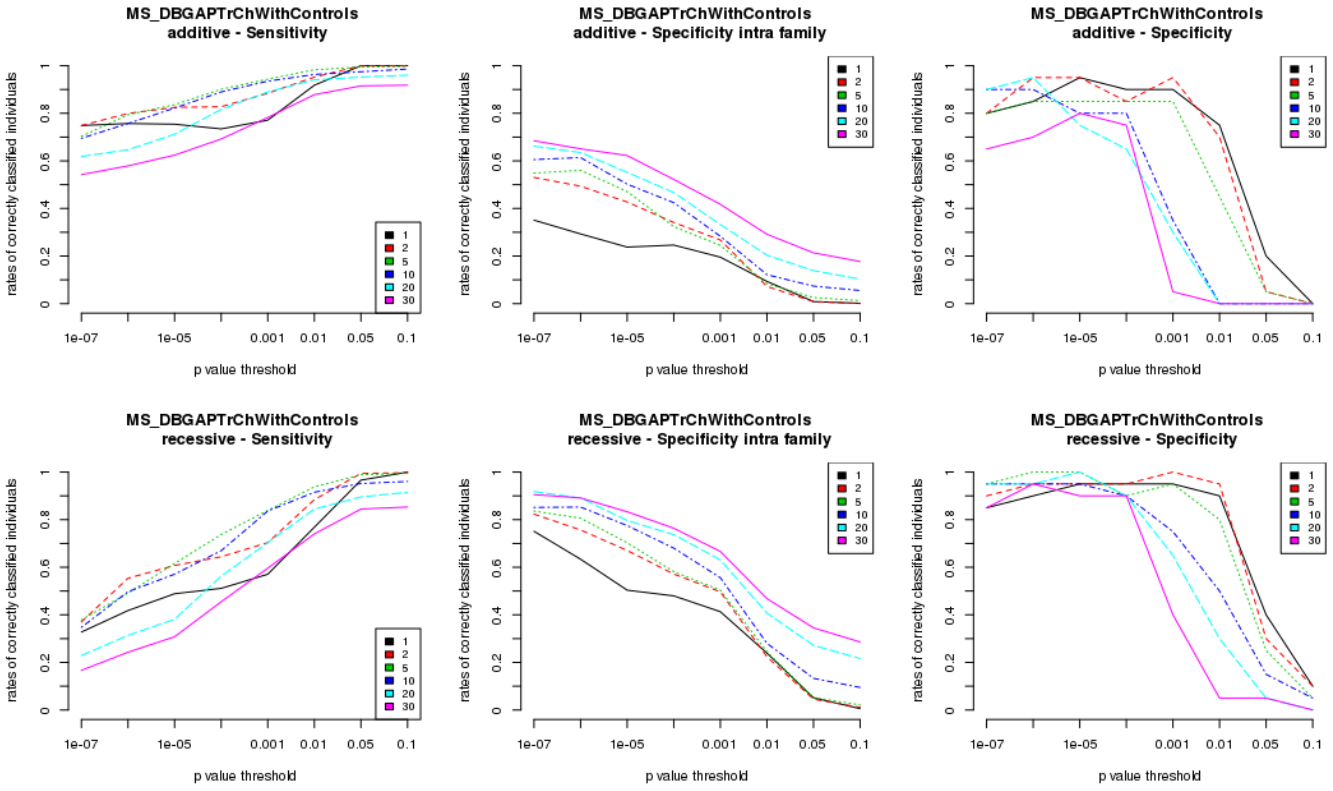

Figure S5: Sensitivy (first column), specificity within families (second column) and (conventional) specificity (right column) from IMSGC data set with genotype calling using the offspring from the TrainingForHaploRisk data subset. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

I now will continue working in order to have results under the c approach, I will also compute results for the cT approach. I am reducing the alternatives I will try to the following:

1. Genetic model: 2 options(recessive and additive)

2. Missing imputation method used to compute p values, select sliding windows and high and low risk haplotypes, and train and test de predictor: 3 options ($keepMissing$, c and, mixed $cT - c$). For the mixed configuration, the cT approach will be used to compute p values, select sliding windows and high and low risk haplotypes and the c approach will be used to train and test the predictor.

I will measure sensitivity, within-family specificity and conventional specificity for the 2×3 different configurations. For all the experiments I will use the control training and test trio data subsets from HapMap CEPH to train specificity and test (conventional) specificity. Results keeping missing genotypes is already shown in Figure 5. I will also compute sensitivity and within-family specificity using the Spanish validatation data set MS_ES with 80 trios. It has to be noted that, as this data set with genotyped with a different array (Affymetrix 6.0), genotype calling could not be done using the models for all loci learned with BRLMM algorithm from the TrainingForHaploRisk data subset made up from IMSGC data set.

Granada, September 13th, 19:11

I have already obtained plots for predictor validation using the Spanish data set when missing genotypes are kept (Figure 6). The right column shows results when virtual parents with both

genome-wide haplotypes are low risk haplotypes. 80 virtual parents have been created from the 160 original parents by randomly coupling their low risk haplotypes. I have not taken into account gender information, therefore a parent may be created with haf of their genome from a mother and the other one from a father. These parents should theoretically have genotypes closer to those obtained from a control data set. However, results are very different to those shown for conventional specificity (right column) in Figure 5. Right now I am not able to interpret these results. I suspect it may be due just to an error in the script I have used to obtain the genotypes of these virtual parents. However, I will go ahead, as I hope to finish this work very soon if results using the c or the mixed $cT - c$ approaches are good, at least for sensitivity and conventional specificity (I already obtained then under the cT approach in august and it worked, now I just need to repeat the experiment to gain some certainty about those results).

Granada, September 14th, 10:05

I have results under the c approach. Sensitivity and conventional specificity are very high right from the smaller p value thresholds (see Figure 7). Within-family specificity is really low.

Granada, September 14th, 15:11

Results for the validation data set (the Spanish trio data set) under the c approach is shown in Figure 8. It seems that every single individual, including all unaffected parents are classified as affected. There is only an exception for sliding window size 1 when using virtual coupling in order to test something closer to conventional specificity. However, this results is so different to the other that

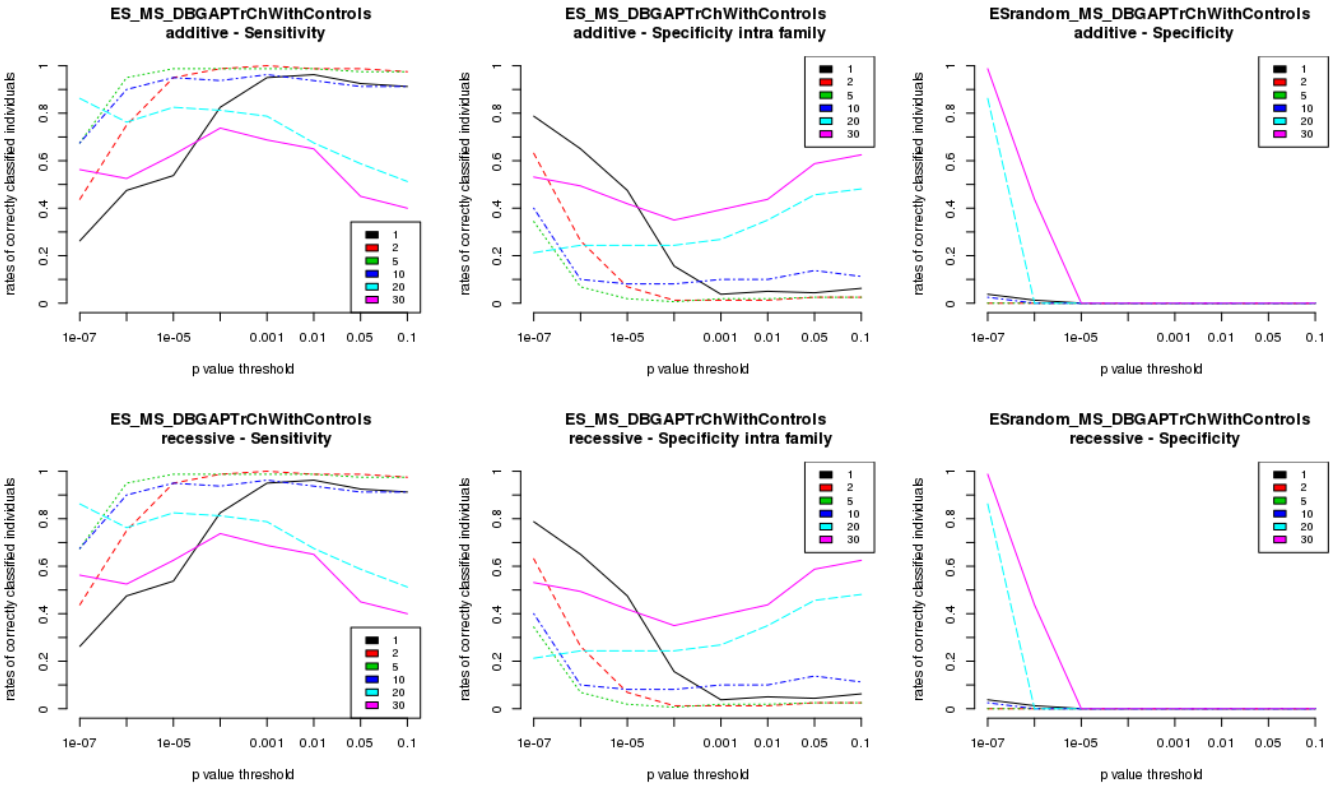

Figure S6: Sensitivy (first column), specificity within families (second column) and almost conventional specificity (right column) from MS_ES with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

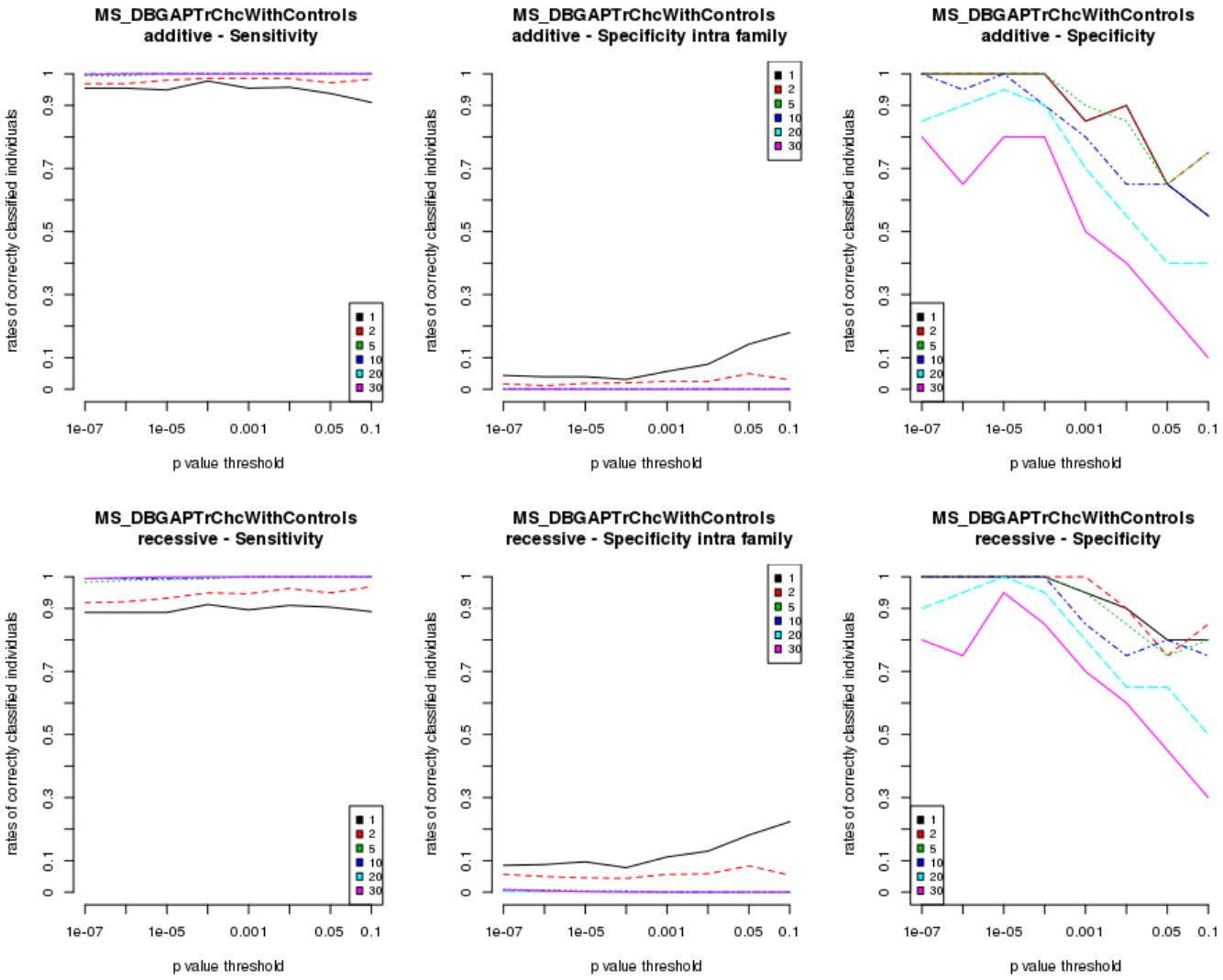

Figure S7: Sensitivy (first column), specificity within families (second column) and almost conventional specificity (right column) from IMSGC data with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set and missing genotypes completed using the c approach. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

it may be due to some software bug.

Granada, September 15th, 12:32

I have finished the experiment under the cT approach (it is actually only the cT approach to compute los and high risk haplotypes and TDT-2G p values to select positions for each predictor, and c to perform learning and testing, this is why I have called it lately mixed $cT - c$ approach). Figure 9 shows results. I removed results for sliding windows of size 30 because I had some bugs and I do not want to lose time running again the script for now, there was nothing extrange in the results I already had for that window size. The experiment was the one I did in august but I wanted to repeat it again. Now I have added within-family specificity (middle column) and, as it an be seen, it is an absolute defeat.

Granada, September 15th 2017, 13:20

I was launching the script to validate these results again with the Spanish validation data set, I already launched it but I could not stop thinking in two different ideas. One of them started yesterday. It was about using a second disease. I currently only have dbGaP permit to use the MS IMSGC data set and the Autism data set. This last one was genotype with Illumina. Although I cannot use the procedure explained and used for Affymetrix, I could just use the calling genotypes available at dbGaP. I would also need a trio control data set. I check out and HapMap 3 included CEPH genotypes from the Illumina Human 1M array, a reduced version of the Illumina Human 1M Duo used for the Autism data set. It will take several days but I think it will be very important

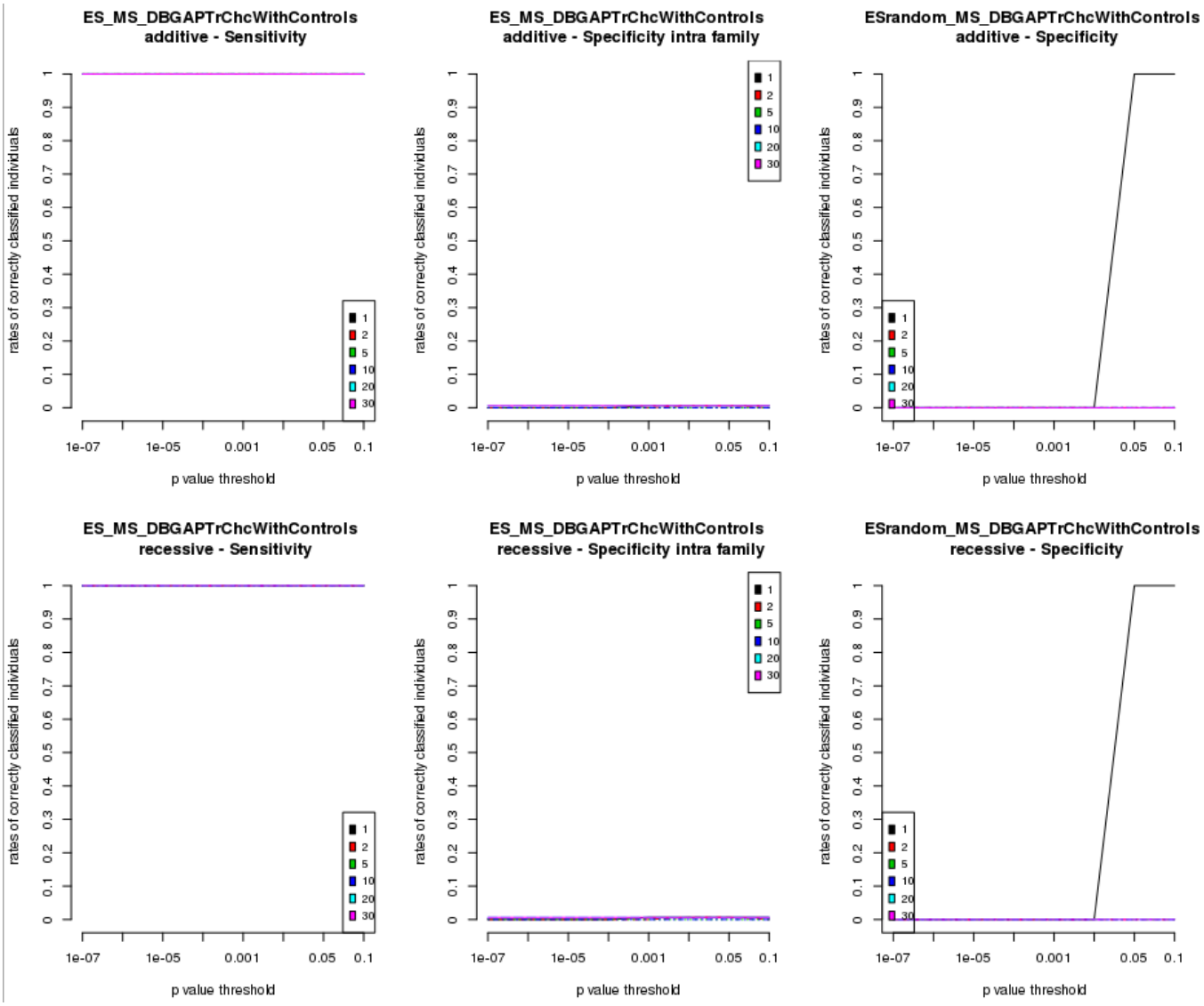

Figure S8: Sensitivy (first column), specificity within families (second column) and almost conventional specificity (right column) from MS_ES with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

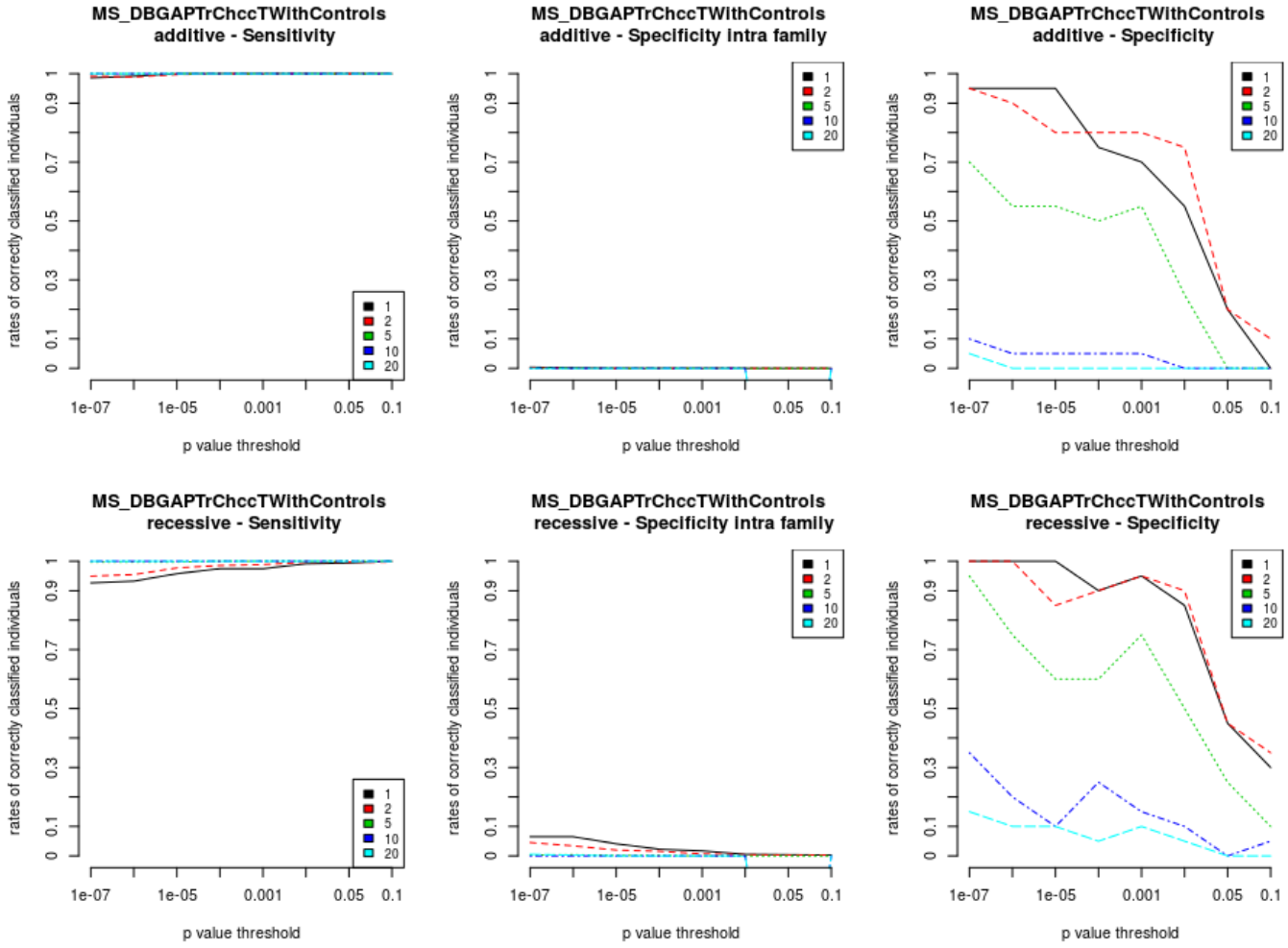

Figure S9: Sensitivity (first column), specificity within families (second column) and conventional specificity (right column) from IMSGC data set under the mixed $cT-c$ approach and with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

if an accurate predictor may be built, even if it is accuracy cannot be measured as that from MS as the genotyping calls, normalization and quality control are supposedly performed for the whole data set. I claimed that they should be solved separatedly. Even more, the training subset should be used to learn a genotype model for each SNP and the test subset and any other new individual should be called using those models. This way we can perform genotype calling with only one individual as we will have a predictor ready to be used. Although I said that we would need also their parents, it must be noted that results for the additive genetic model are also quite good, and this model does not require phased genotypes. The second issue was about a new experiment. Why not use a mixed $cT - keepMissing$ approach? I thought about it because the cT approach (actually, mixed $cT - C$ approach reports very good results for single SNPs (sliding windows of size 1) and the shortest p value threshold. The fact that parents of affected indivifuals are almost always wrongly classified as affected could be reduced if no missing imputation were performed. It seems that cT is detecting significant associations due to rare variants and the c approach may produce several type I errors in parents.

Granada, September 15th 2017, 13:55

I have thought again about the MUT approach to impute missing data [1] and I really think it has more sense than the cT approach. I have decide to perform an experiment under this approach using the proposal claimed in this work (HaploRisk2), i.e., using control trios also in the training subset in order to learn the predictor. I also have realized that, in the case the last experiment proposed in the entrance before this gave good within-family specificity results, perhaps I would not need to use a control trio data set. Let's see ...

Granada, September 16 2017, 10:17

I already have results for the mixed $cT - c$ approach in the validation data set MS_ES (Figure 10). It seeme that every individual, affected or not, is classified as affected, regardless the genetic model used, thje sliding window size and the p value threshold. No conjectures or attempt to find an explanation for now.

I am now changing names to result files trying to be more faithful to what they represent. This is a reason of many mistakes, as I already explained [1]. I was about to call "AccuracyCaseControlTrainingMSDBGAPTrChc_ES.png" to the file with the last plots and then I realized this file already existed to show results under the c approach, while now I had results for the mixed $cT - c$ approach. Therefore I named it as "AccuracyCaseControlTrainingMSDBGAPTrChccT_ES.png". Because of this, I decided to change the names to the text files with accuracies to me more clear. But then, how would I name the files under the mixed $cT - keepMissing$ approach? This is just an example about the mess in the files I keep. Some of the first plots may not be produced again because perhaps I lost the result files or rewrote them. Therefore, I always keep repeating the whole experiment. This help me also to think about the whole experiment again and perhaps to change my conjectures. Although this issue has provoke me tremendous nightmares for years, now I have learned this is not that bad, because helps me to be more open and avoid sometimes researcher biases. My really bad memory may be also an advantage for the same reason. The whole problem is so challenging and the chances of errors are so high, that I think the more I repeat experiments or do similar ones, the better.

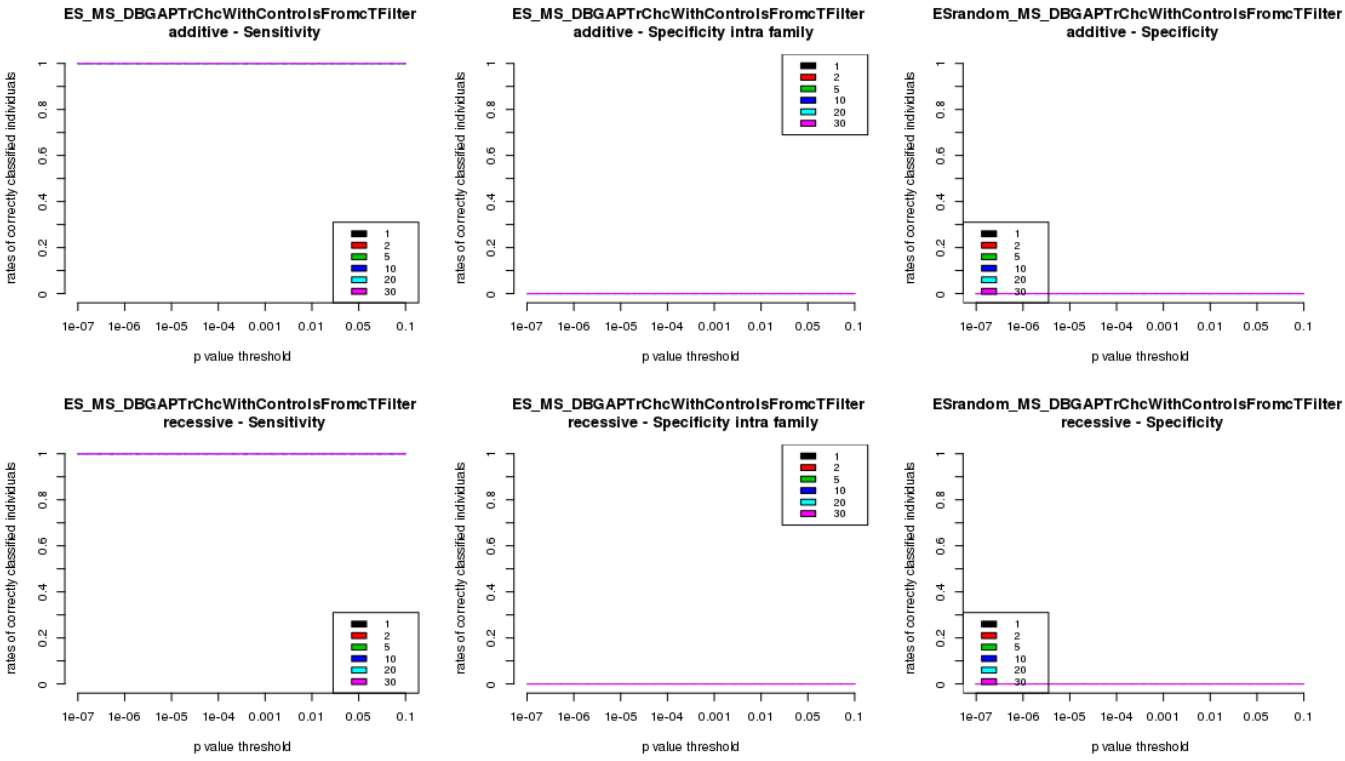

Figure S10: Sensitivy (first column), specificity within families (second column) and almost conventional specificity (right column) from MS_ES under the mixed $cT - c$ approach and with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

Granada, September 16th, 12:34

I think I have an error in the results for the validation data set. I think I was using as predictor one built from the Test data subset in IMSGC, while the model should be built with the training data subset composed of the cases from the TestForHaploRisk data subset and the controls from the training control data subset. I found this error trying to launch the script to test the $cT - keepMissing$ approach and what I called "almost conventional specificity" by making up virtual IMSGC parents with two non-transmitted genome-wide haplotypes randomly coupling from the real parents. I was looking for the last time I launched that script and it was with the validation data set, and I found out this error. I have to modified the script.

Granada, September 16th, 13:17

I have plot results under the mixed $cT - keepMissing$ approach testing sensitivity (left column), within-family specificity (middle column) and conventional specificity (right column) in Figure 11.

For short haplotypes conventional specificity are similar to that obtained under the $keepMissing$ approach (see Figure 5). However, these results show both much higher sensitivity in small p value thresholds for all sliding window sizes. This sensitivity improvement comes together with a higher conventional specificity for single SNPs or small sliding window sizes. Therefore, it seems that single SNPs and very small p value thresholds give the largest AUC, as when using the mixed $cT - c$ approach.

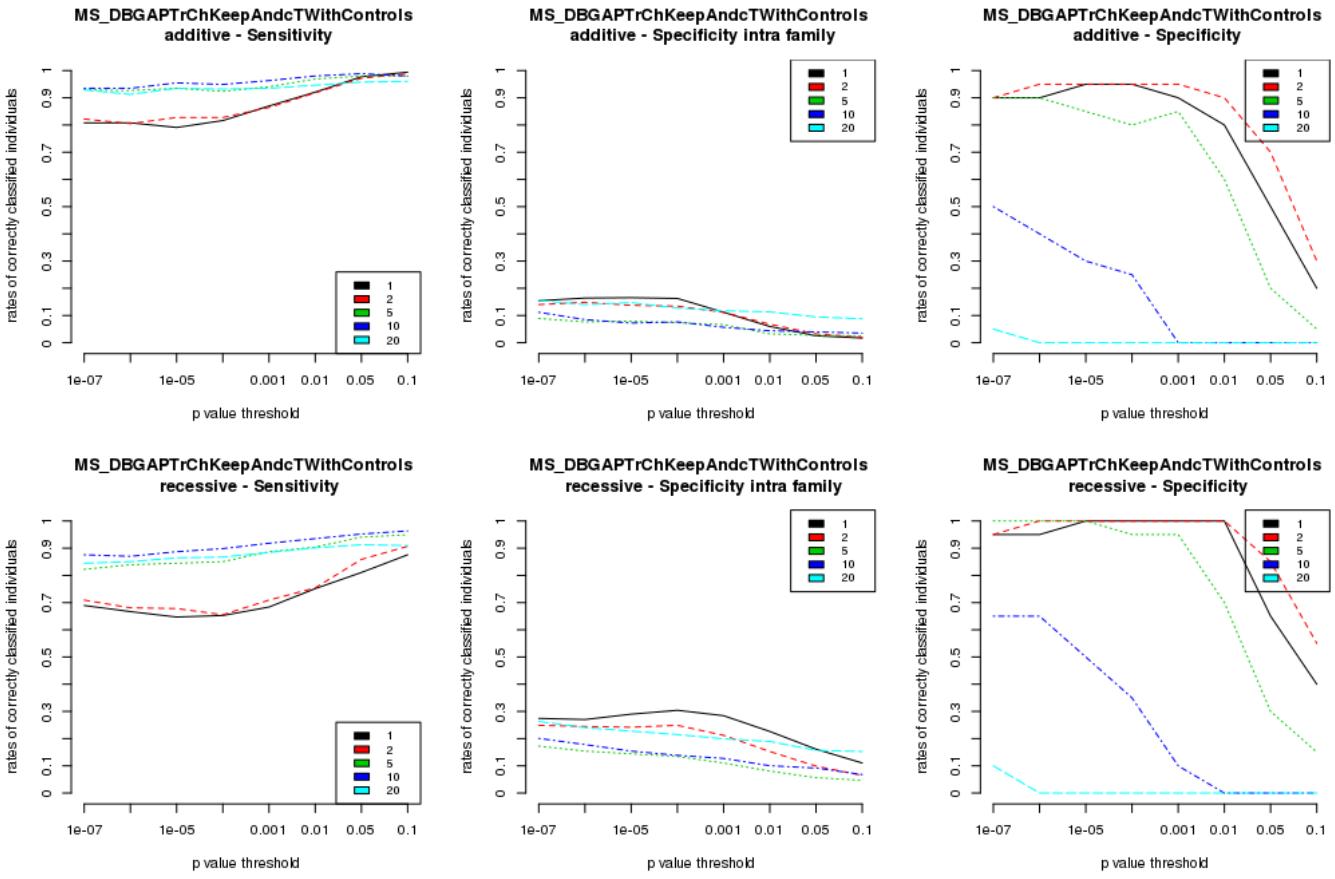

Figure S11: Sensitivity (first column), specificity within families (second column) and conventional specificity (right column) from IMSGC data set under the mixed $cT - keepMissing$ approach and with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

See AUC and accuracy results for this experiment (mixed $cT - keepMissing$ approach). I do not understand well why accuracy increases a little bit for the larger threshold p values, but AUC strongly decrease. Anyway, they both agree to be good enough for the shorter window sizes and threhold p values.

Granada, September 17th, 11:45

Results for specificity using virtual parents of the case offspring (right columns) are shown for mixed $cT - keepMissing$ and $cT - c$ approaches in Figures 13 and 14 respectively.

This specificity, that I called before "almost conventional specificity" is clearly very bad, and very different from the conventional specificity obtained when using real controls from the CEPH HapMap data set. To me, it could mean that parents of affected MS individuals also carry on the genetic basis to develop te disease but for some environmental reasons they remained healthy.

Granada, September 17, 13:36

Now I am working on the Autism data set. i have decided to test the method in this data set. The main issue was to find a control trio data set, but I have realized that we do not need a control trio data set, it will be enough to have unrelative samples. Therefore, I will use the WTCCC2 British Cohort 1958BC control data set genotyped with Illumina 1.2 array. I have the SNP annotation file and also the one used for the Autism GWAS, which was genotyped with Illumina Human 1M array. This way, I will flip (change strands) whenever they were not used in the same direction following the annotation files.

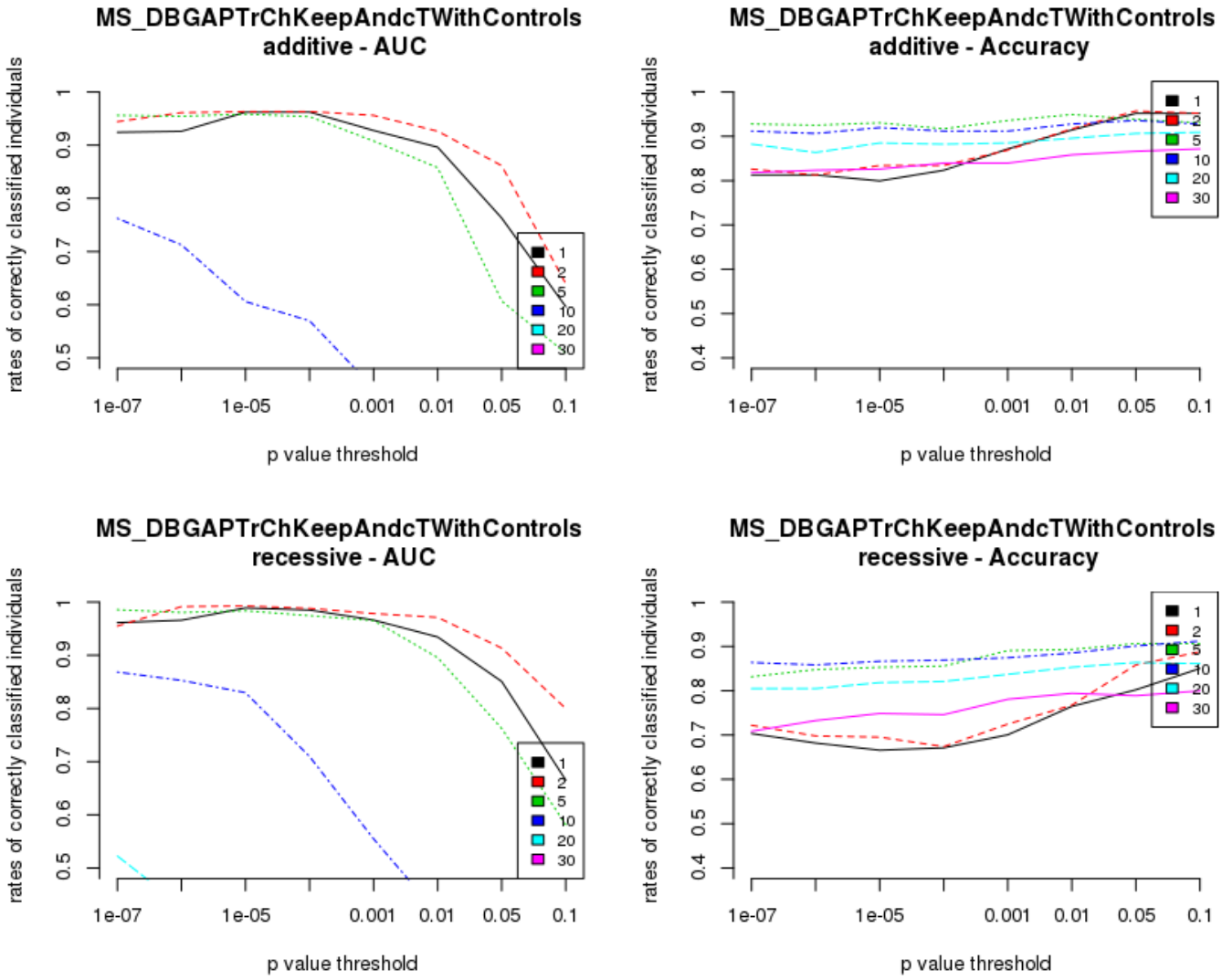

Figure S 12: AUC (left column) and accuracy (right column) from IMSGC data set under the mixed $cT - keepMissing$ approach and with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model. Values were computed using affected offspring – as when computing sensitivity – and control parents – as when computing conventional specificity –.

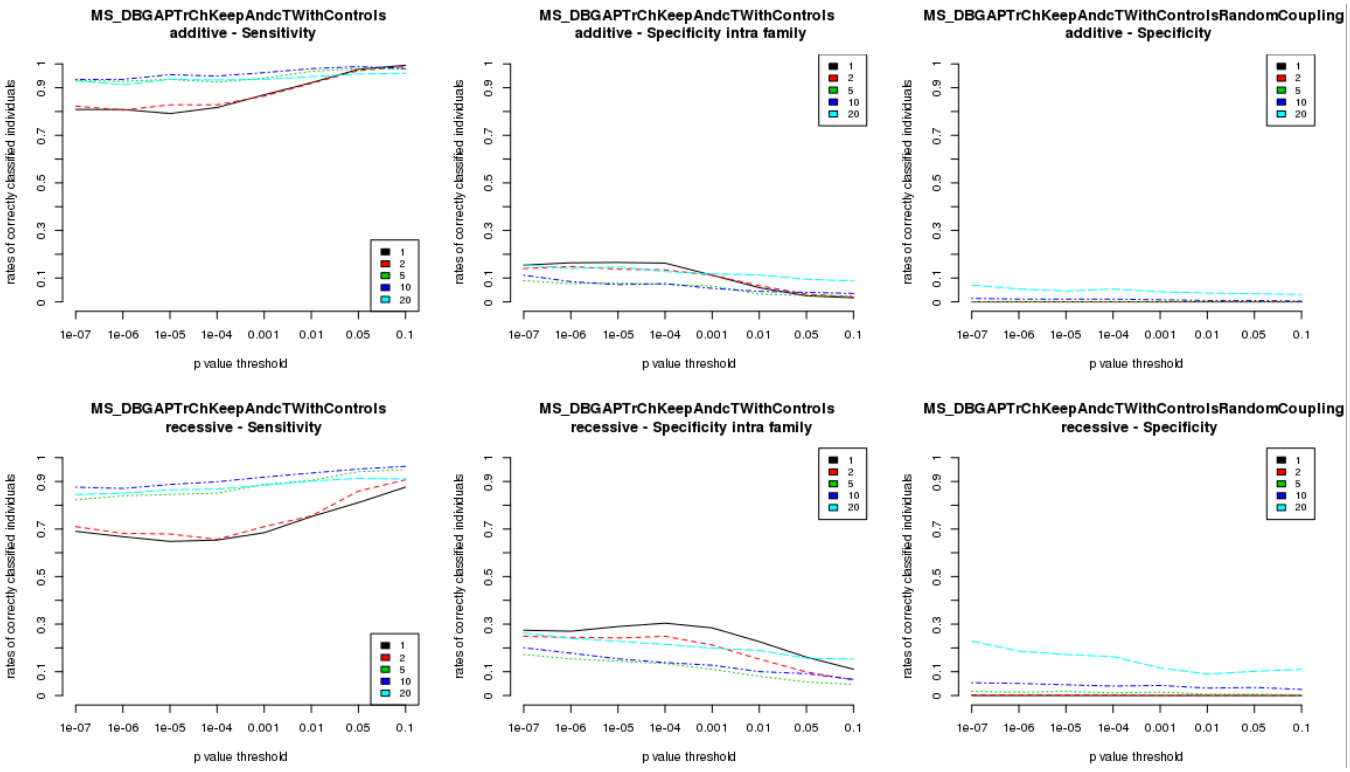

Figure S 13: Sensitivity (first column), specificity within families (second column) and "almost conventional specificity" (right column) from IMSGC data set under the mixed $cT-keepMissing$ approach and with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

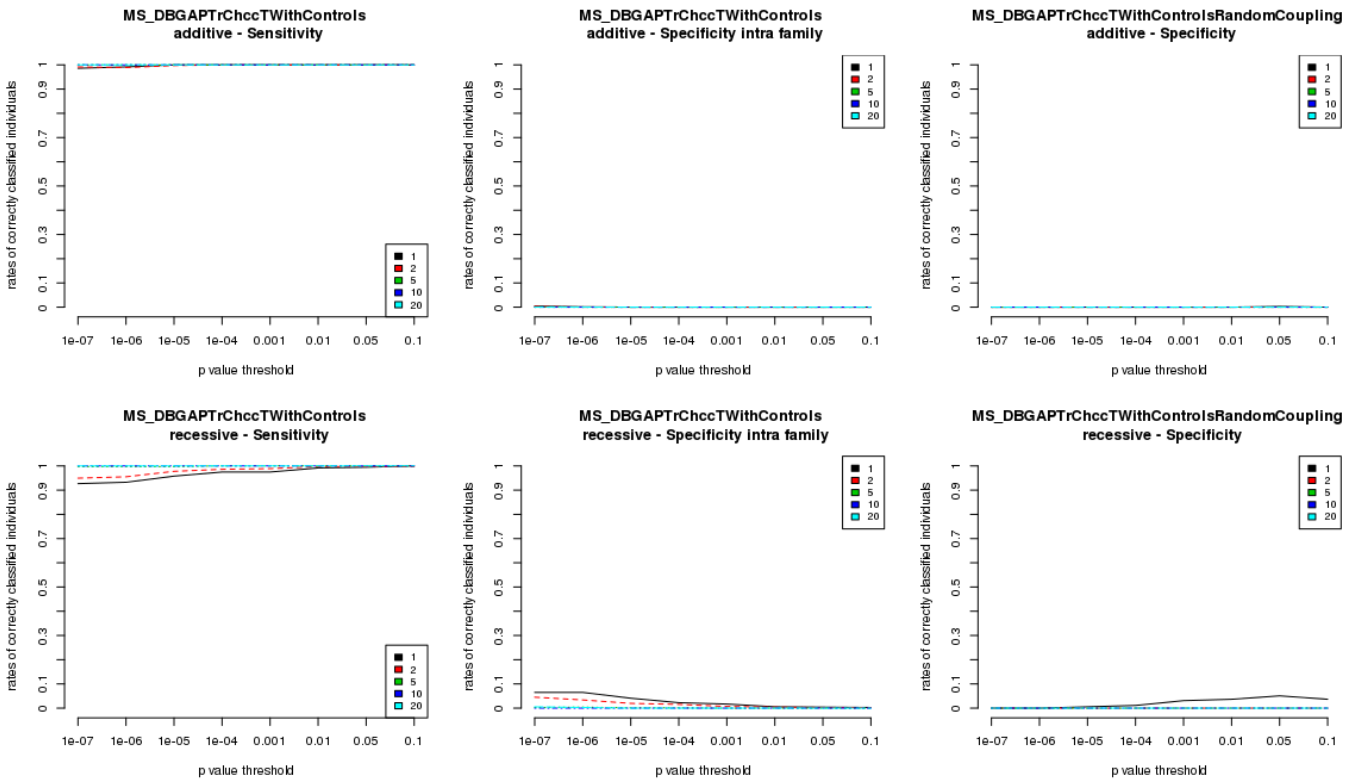

Figure S 14: Sensitivity (first column), specificity within families (second column) and "almost conventional specificity" (right column) from IMSGC data set under the mixed $cT-c$ approach and with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

Moreover, I will use the MUT approach to complete missing genotypes. I want to compare its results with the cT approach. This time I will remove computation for sliding windows of size 30: I will save a significant computational time but, what it is more important due to the almost full external memory, I will save disk space.

Granada, September 18, 11:52

Yesterday evening I was thinking about the cT and MUT approaches and thought about the H approach and that it would be interesting to try it again. It has be noted that I tested all these approaches under the HaploRisk study but I want to test them under this new stud (HaploRisk2) in which I having good results under the cT approach and the basic changes are not related with the approach (I am using the oldest ones) but with the way to build and test a predictor, by adding true control individuals from a control data set. So I will check the H approach using this new way to build and test a predictor. I may have serious disk spaces problems. If this is the case, I will remove some results obtained for the already tested approaches. If I needed them again I will have to compute those results again. This has turned out to be a normal way to work for me for years. The major advantage is that, as I am always detecting and correcting bugs, I trust always the more recent versions.

Granada, September 22, 13:27

Now I do not get very surprised to see how many times I get results against my conjectures, even after such a large experience in (failed) experiments ... I have just got results for the MUT

approach (it is actually a mixed approach $MUT - MU$ for the same reason I had to use a mixed $cT - c$ approach) and they seem to be much better that those results obtained when using the $cT - c$ approach (see Figure 15).

Granada, September 22, 15:35

And now I have plots for the H approach. See Figure 16. These results were more in agreement with what I could expect, as the H approach filters SNPs without completing data in support of the transmission of some allele. This way I expected to have results similar to those using the c approach. It would be interesting to think about a mixed $HT - H$ approach but right now I cannot think about if this approach would even have sense.

Granada, October 14 2017

I have decided to finish the work by now. I was trying to use the Autism data set but I am having many technical problems with the server and I believe it is time to finish by now. I have also changed my mind aboout the journal to submit it. My first idea was to try Nature again but now I think I have more important issues in my personal life and I need to publish these results anyway. I will try PLoS ONE. It is a very open-mind journal and I hope the reviewers are OK with the idea of this researcher notebook.

———

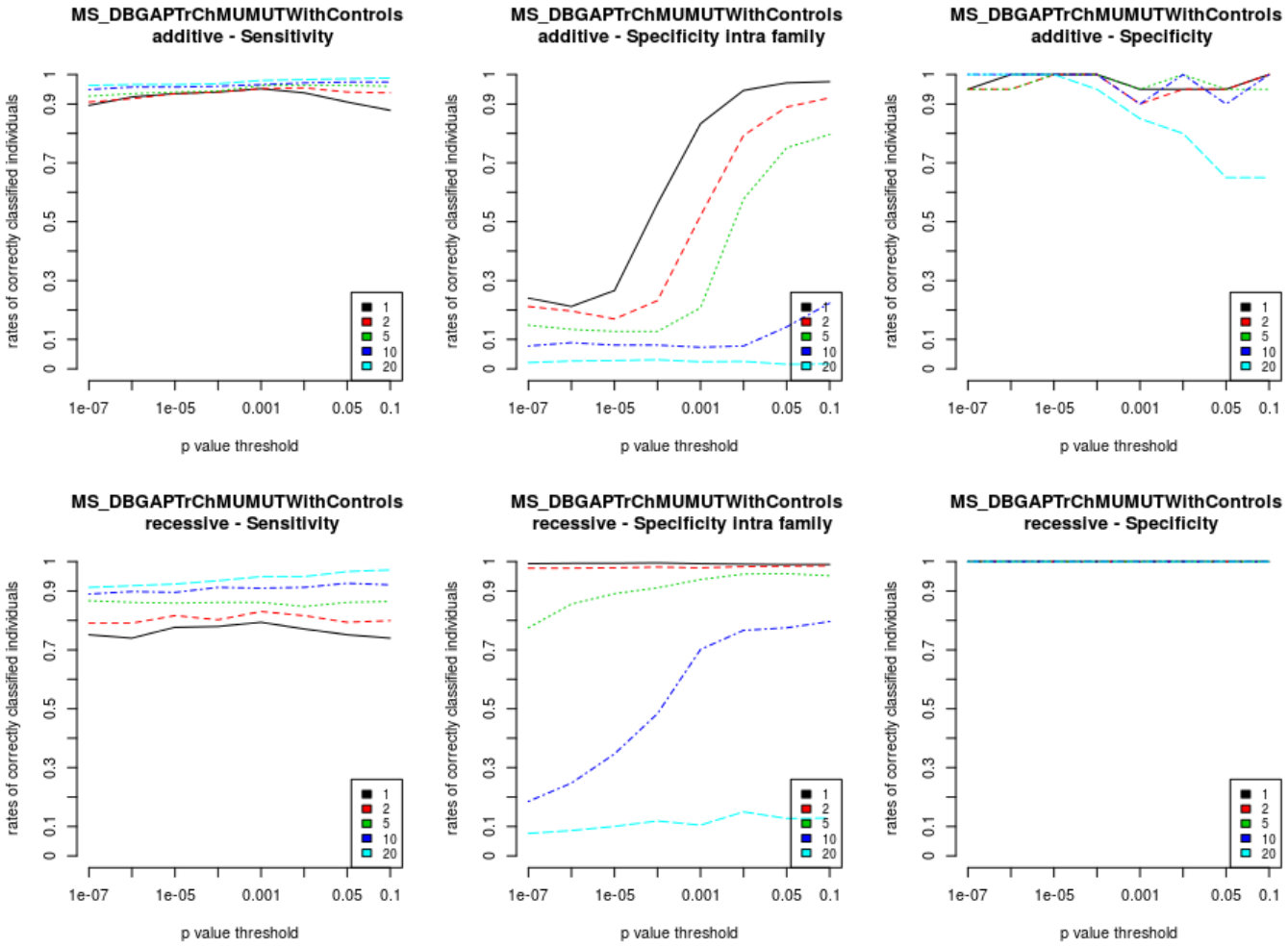

Figure S 15: Sensitivity (first column), specificity within families (second column) and conventional specificity (right column) from IMSGC data set under the mixed $MUT - MU$ approach and with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

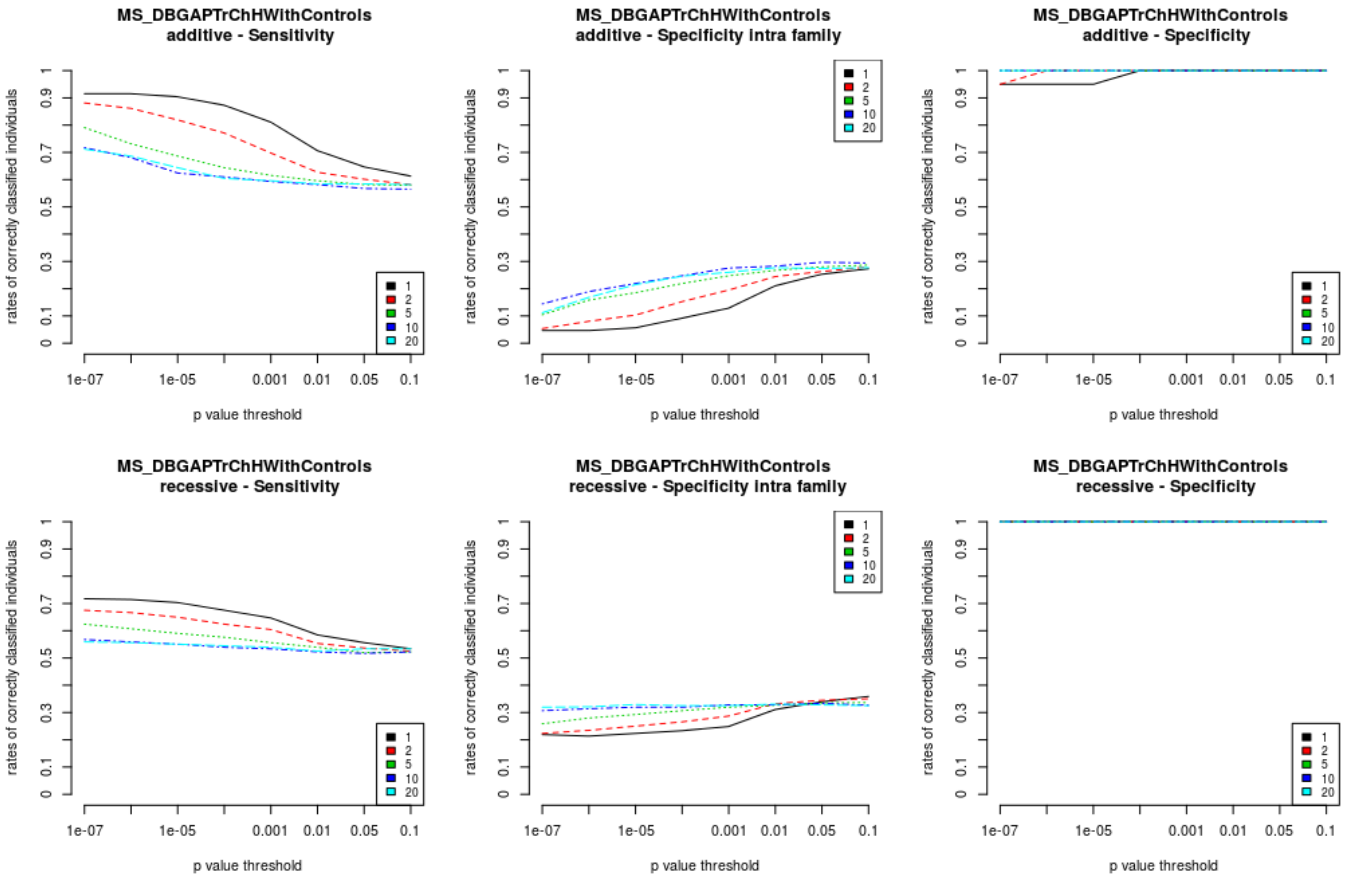

Figure S16: Sensitivity (first column), specificity within families (second column) and conventional specificity (right column) from IMSGC data set under the mixed H approach and with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

Granada, October 15th, 15:56

I thought I had already the most important results to start writing a paper and yesterday I started and I made the plots now I am showing here (Figure 17):

After making these plots I realized that I did not have results with the validatation data set. I got those results for the $cT - c$ approach (Figure 10) and I was quite relaxed thinking they should be ok because with the c approach they were similiar to those from the test data subset made from the IMSGC data set. However today I was making these plots for the MUMUT approach and I got very weird results (see Figure 19).

As it is being very difficult to understand why sensitivity is almost 0, i.e., every individual are considered with no risk for MS, I am repeating again results under the $cT - c$ approach, as every time I redo an experiment I can have some differences.

Now I have started repeating the experiment. My plan of finishing this work by October 15th is already not fulfilled.

Granada, October 15th, 17:34

While repeating experiments for the $cT - c$ approach I have realized I had introduced an error in the sequence of intermediate files between bash scripts I was using and, because of these, results did not make any sense: I had almost zero sensitivity and, at the same time, almost zero specificity. Now I will repeat the procedure for the $MUT - MU$ approach as well.

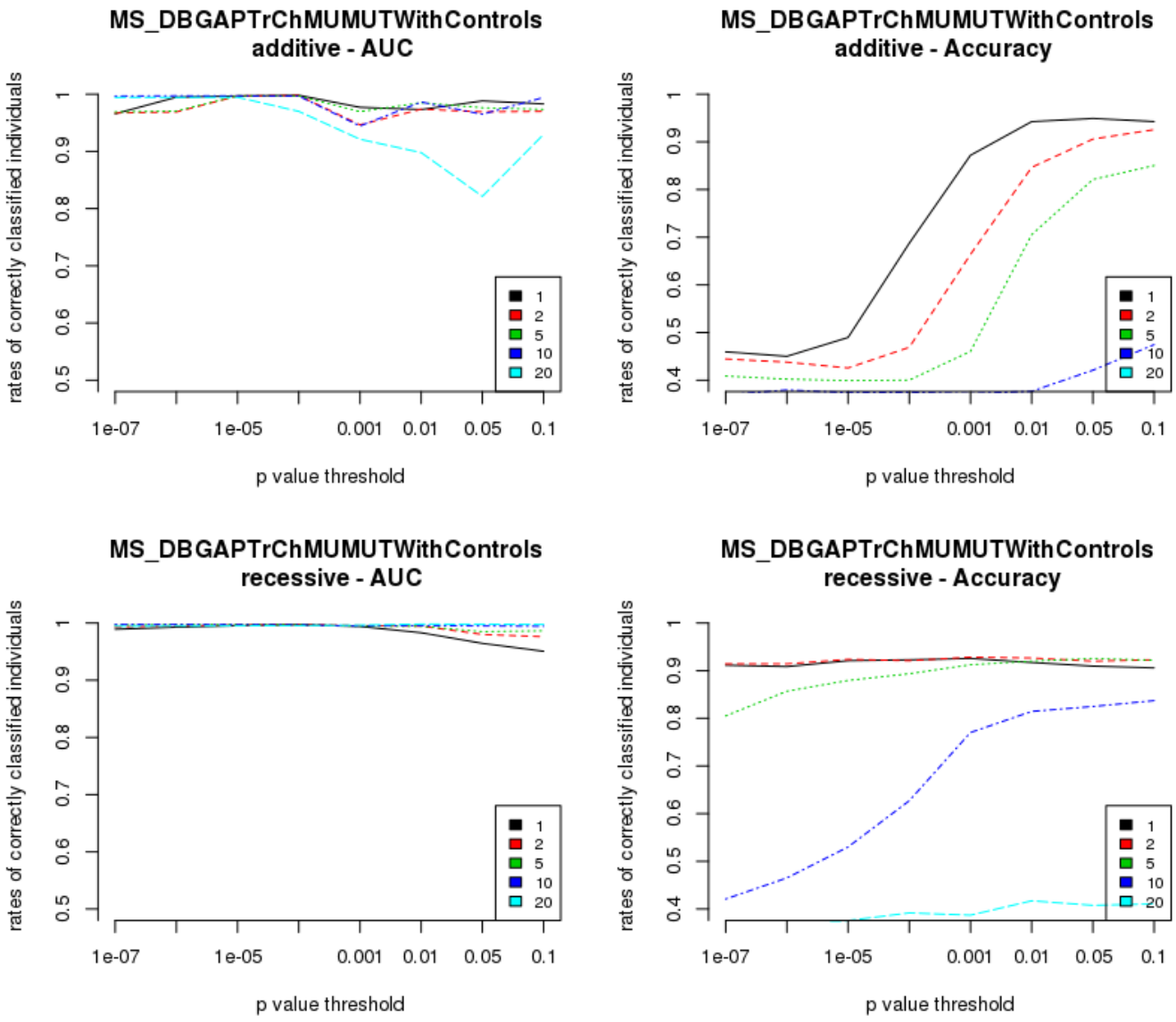

Figure S17: AUC (left column) and accuracy (right column) from IMSGC data set under the mixed $MUT - MU$ approach and with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model. Values were computed using affected offspring – as when computing sensitivity – and control parents – as when computing conventional specificity –.

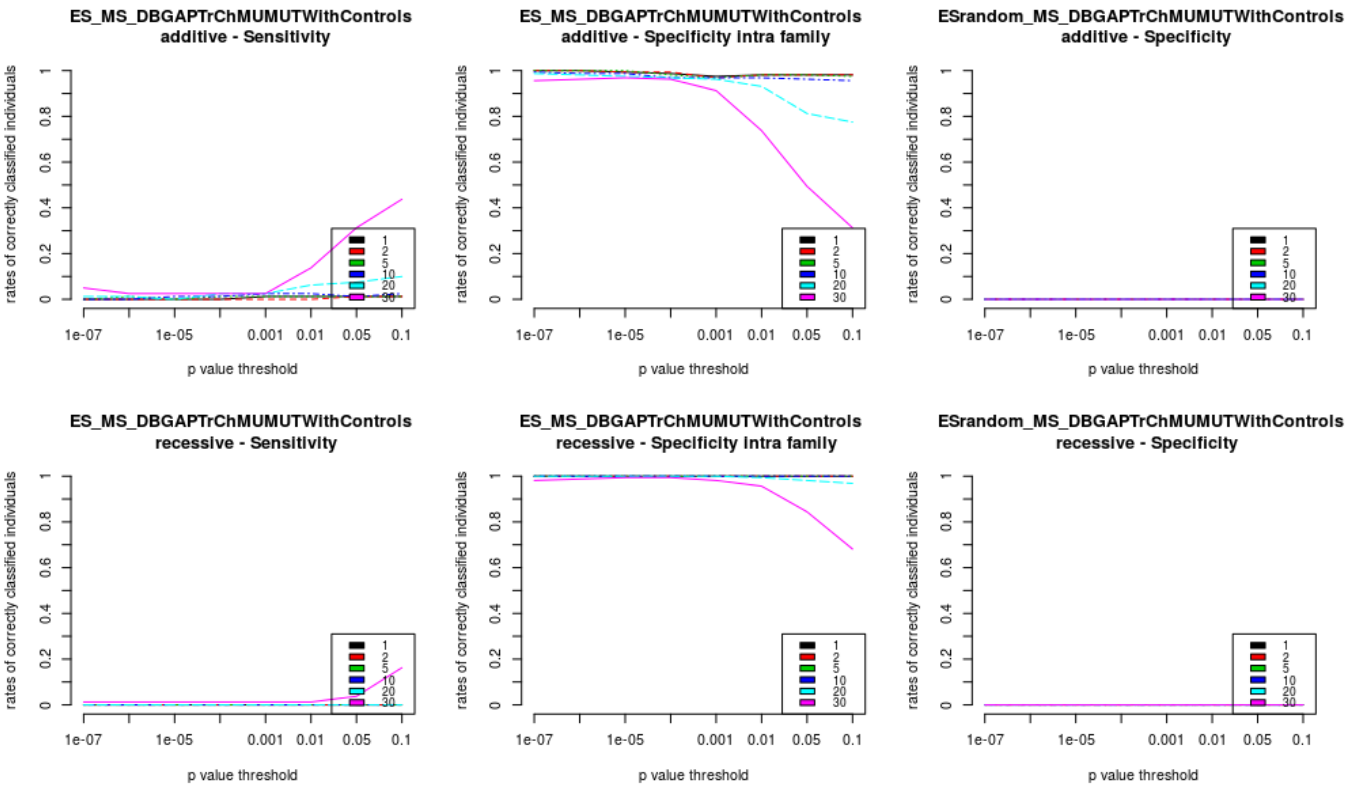

Figure S18: Sensitivy (first column), specificity within families (second column) and almost conventional specificity (right column) from MS_ES under the mixed $MUT - MU$ approach and with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

Granada, October 17th, 00:25

Late yesterdat I obtained a completely different result for the $MUT - MU$ approach (see Figure 19). I felt I had to stop working and start praying.

After leaving some time off to be able to pray and later think about this work, I realized that the very common pattern of wrong results [1] consisting of completely opposite values of sensitivity and specificity may be due to a huge problem with the data. It actually does not have any sense such a large difference. It was as if there were no differences at all in the variants between the affected and the unaffected individuals. Then I looked at the data and observed a lot of missing data due just to unphased genotypes. This occurs because of the phasing algorithm I am using: solve only certain phase given the trio family. Therefore, every locus with the thre family members heterozygous will be coded as missing. I have decided to use a phasing algorithm that solve all the loci even if there are some phasing errors. I will use beagle considering trios. for triple heterozygous loci, I will use the TrainingForHaploRisk data subset as the reference data set. It will take some time to get results again for IMSGC but I think I need to eliminate missing information due to unknown phase because it may be an important amount of data that may completeley change risk prediction.

Tables

In order to test a risk predictor, the test data subset must be coded as it was done with

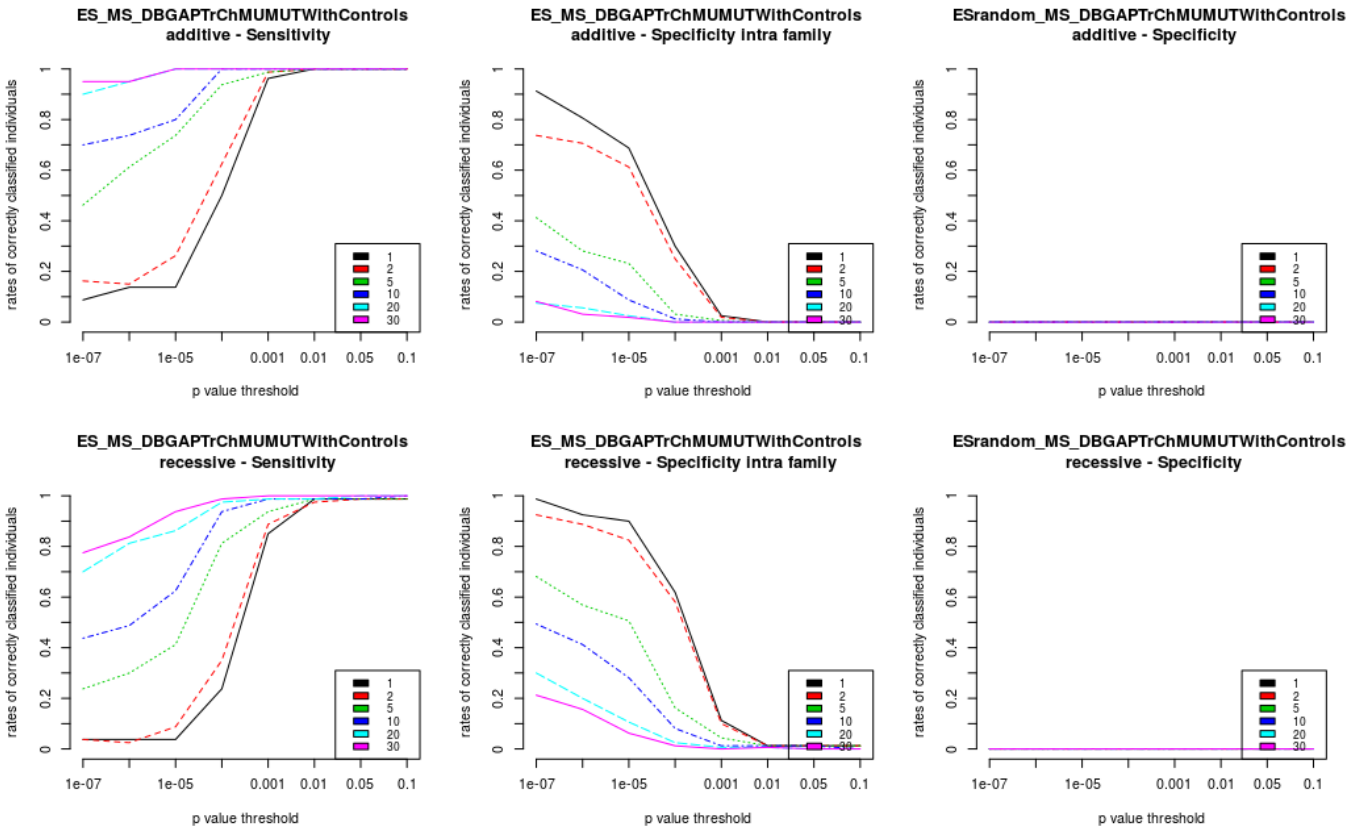

Figure S19: Sensitivy (first column), specificity within families (second column) and almost conventional specificity (right column) from MS_ES under the mixed $MUT - MU$ approach and with genotype calling using the offspring from the TrainingForHaploRisk data subset of IMSGC data set. Results for the additive genetic model are shown in the first row, while the second row shows results for the recessive genetic model.

- Input data: Raw genotype intensities for an individual and their parents by using preferentially the same array used to build the risk predictor.

- Procedure:

    1. To perform the calling (to solve the two alleles each individual has at each SNP locus) by using the model learned when calling was performed to obtain the genotypes for the TrainingForHaploRisk case data subset.

    2. To complete missing genotypes under the cT approach by using the c allele learned from the TrainingForHaploRisk data subset.

    3. To solve the phase using the parents and leaving those triple heterozigotic loci unsolved (algorithms to solve these positions such as Beagle [9] did not outperformed our results [1]).

    4. To obtain the individual risk using the recessive approach [2].

- Output data: Individual risk to the given trait.

Table S 1: Procedure to obtain individual risk to a trait given a trait predictor and the individual and parents raw genotypes.

- Input data:

  Raw genotype intensities of a case trio data set (affected indifiduals and their parents) with n trios using a given array.

  Raw genotype intensities of a control trio data set (offspring and parents, all unaffected) with m trios using the same array.

- Procedure:

  1. To randomly divide the case trio data set into two parts, the training and the test data subsets (the test subset will be used to measure sensitivity).
  2. To randomly divide the control trio data set into two parts: the training and test data subsets (the test subset will be used to measure specificity).
  3. To randomly divide the training case trio data subset into two equally-sized data subsets, the TrainingForHaploRisk and the TestForHaploRisk data subsets.
  4. To perform the genotype calling for the TrainingForHaploRisk data subset. For Affymetrix IMSGC data set I used a very basic configuration of the brlmm calling algorithm (implemented under the Affymetrix software apt-probeset-genotype): no normalization was performed, intensities were transformed by their polar coordinates R vs θ (RvT) [1], no default model neither priors were used. The model (the centers and variances for each one of the three genotypes at each locus) was written in a .models file to be used with the other data sets.
  5. To perform genotype calling for TestForHaploRisk, test case trio, and control trio data sets using only the model generated during the calling of the TrainingForHaploRisk data subset.

- Input data:

  TrainingForHaploRisk, TestForHaploRisk with phased genotypes and missing genotypes handled.

  A window size $i, i \in \{1, 2, 5, 10, 20, 30\}$

- Procedure:

  For each autosome from 1 to 22:

  For each sliding window of size i from the first position with offset of 1:

  To compute high and low risk haplotypes for the given sliding window using TDT-2G with the TrainingForHaploRisk data subset.

  To compute TDT-2G p values using the TestForHaploRisk data subset [10] given the high and low risk sets of haplotypes.

- Output data:

  A list of high and low risk haplotypes of size i for all the sliding windows within all the chromosomes.

  P values of TDT-2G for all the sliding window of size i within all the chromosomes.

Table S3: Procedure to compute association with the trait for each sliding window of size $i, i \in \{1, 2, 5, 10, 20, 30\}$ so that predictors can be built using different p value thresholds.

- Input data:

  A trio data set with phased genotypes and missing genotypes handled.

  Sliding window size $i, i \in \{1, 2, 5, 10, 20, 30\}$

  P value threshold $j, j \in \{e-0.6, e-0.5, e-0.4, e-0.3, 0.01, 0.05, 0.01\}$

- Procedure:

  For each individual in the data subset

  1. For each autosome from 1 to 22:

     For each sliding window of size i from the first position with offset of 1 and whose TDT-2G value is lower than j:

     To code each one of the 2 individual haplotypes in parents (transmitted/non transmitted) as low or high risk (binary code).

  2. To merge each one of the 2 chromosome-wide individual haplotypes among all the chromsomes, in order to have 2 genome-wide haplotypes for each individual (transmitted/non transmitted).

  3. To code the class of the transmitted haplotypes in parents as high risk haplotypes, and the non-transmitted haplotypes as low risk haplotypes.

- Output data: A genome-wide haplotypes data set with risk variants (input variables) written using a binary code and class using also a binary code.

Table S4: Procedure to code genome-wide haplotypes from a trio data set given a sliding window size i and a p value thresold j.

the TestForHaploRisk data subset, using the same lists of hiwh and low risk haplotypes already learned. The individual risk will be obtained by the combination of the genome-wide haplotype risk of each of the two haplotypes an individual has using either the recessive or the additive genetic model [2].

1. Abad-Grau, M. M. *et al.* Informative missing genotypes: hopes for the craving revolution of genome-wide risk models. URL: http://bios.ugr.es/HaploRisk/index.php (2017).

2. Abad-Grau, M. M., Medina-Medina, N., Masegosa, A. & Moral, S. Haplotype-based classifiers to predict individual susceptibility to complex diseases-an example for multiple sclerosis: 12 biostec-bioinformatics conference, Vilamoura, Algarve, Portugal, February 1-4, 2012. Proceedings. In *BIOINFORMATICS 2012 - Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms, Vilamoura, Algarve, Portugal, 1 - 4 February, 2012.*, 360–366 (2012).

3. Jager, P. D. *et al.* The role of the cd58 locus in multiple sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 5264–69 (2008).

4. Little, R. J. A. & Rubin, B. *Statistical Analysis With Missing Data* (Wiley, New York, 1987).

5. 'International Multiple Sclerosis Genetics Consortium', D. H. *et al.* Risk alleles for multiple sclerosis identified by a genomewide study. *New England Journal of Medicine* **357**, 851–62 (2007).

6. Affymetrix. BRLMM: an improved genotype calling method for the GeneChip human mapping 500K array set. Tech. Rep., Affymetrix, Santa Clara, CA: Affymetrix, Inc. (2006).

7. HapMap-Consortium, T. I. The international hapmap project. *Nature* **426**, 789–796 (2003).

8. The-Wellcome-Trust-Case-Control-Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–78 (2009).

9. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 210–223 (2009).

10. Abad-Grau, M., Medina-Medina, N., Montes-Soldado, R., Matesanz, F. & Bafna, V. Sample reproducibility of genetic association using different multimarker tdts in genome-wide association studies: Characterization and a new approach. *PLoS ONE* **7**, e29613 (2012).