Informative missing genotypes: hopes for the craving revolution of genome-wide risk models

María M. Abad Grau[1]*, Víctor Potenciano[1], Mohamed Abd Allah Makhlouf[2], Serafín Moral[3], Andrés Masegosa[4], Sergio Torres Sánchez[1], Eladio Garví[5], María I. García Sánchez[6], Guillermo Izquierdo Ayuso[6], Antonio Alcina[7] & Fuencisla Matesanz[7]

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

[2]*Faculty of Computers & Informatics, Suez Canal University, Ismailia, Egypt*

[3]*Department of Computer Science and Artificial Intelligence, ETSIIT, Universidad de Granada, Granada, Spain*

[4]*Department of Mathematics, CITIC, Universidad de Almería, Almería, Spain*

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

[6]*Unidad de Gestión Clínica de Neurología, Unidad de Esclerosis Múltiple, Hospital Virgen Macarena, Sevilla, Spain*

[7]*Instituto de Parasitología y Biomedicina López Neyra, Department Cell Biology and Immunology, Granada, Spain*

After more than 15 years since the first human genome was sequenced, genomic profiling for complex diseases still remains a big challenge. In this work we show that missing genotypes and genotyping errors may be the clue to the current inability to build accurate genetic predictors of complex diseases. In some trio data sets of affected offspring we have observed a

*Justice, O Lord, is on your side; we are shamefaced (Dan 9:7).

sound informative missing pattern associated with the trait –i.e. there are more missing in offspring than in their parents– and mostly biased against the call of what it is the low-risk allele according to the known genotypes. We show how under this informative missing pattern, the way we impute missing genotypes may completely change predictor outcomes from being lowly sensitive to being lowly specific, as in many variants with small but true effects on the disease, the low-risk allele may become the high-risk allele. This is only a very hard and time-consuming preliminary work with a very long history of defeats, very little answers and many new open questions and conjectures. However, it may point out to the need of a more careful look to the genotyping procedures as the only way to ever succeed in the prediction of individual risk to complex diseases. We have also described the history of this research by writing a research notebook, supplied as supplementary material. Together with the implications this work may have in the way to face genome-wide studies and the problem of genomic profiling, we believe the research notebook may give some hope to nonconformist scientists. In fact, many researchers of any background are tired of an untrustable way to do science, in which there are many publications with almost exactly opposite conclusions. Through the research notebook the first author tells how this "under control" experimental research became such an obscure, uncrontrollable and full of mistakes work that it throwed up more confusion than light and completely blocked a personal life. Only when she started questioning herself as the main author of the work and listened to her heart with the help of the Church, she was able to put an end to this endless work.

Since genome-wide association studies started, building genetic predictive models of individual risk to highly polygenic diseases has became a very challeging task. As an example, in

Multiple Sclerosis (MS) the state-of-the-art predictive Area Under the ROC Curve (AUC), around 0.64 [1], was achieved by using the common weighted Genetic Risk Score (wGRS), while it should be around 0.95 by considering risk prevalence and disease heritability (see the end of page 4 at the supplementary material).

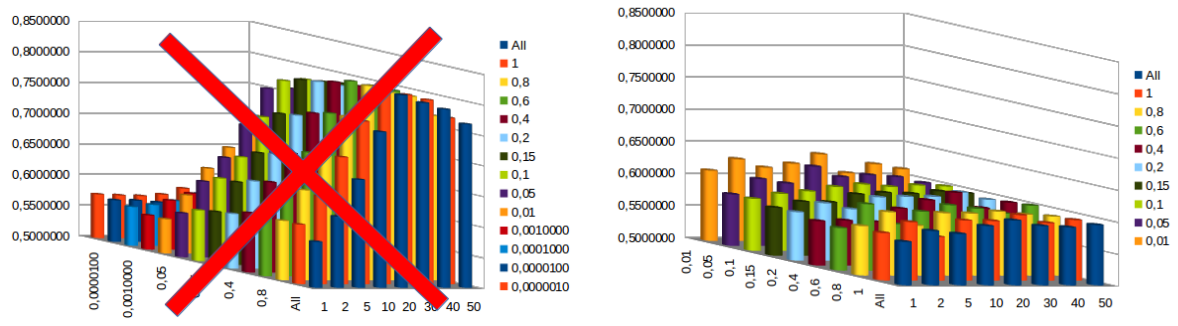
Optimistic opinions about genomic profiling have often flourished to dry soon under the evidence that only under simulations the individual risk can be predicted as expected given the heritability and prevalence of the disease under study. This work is the history of another optimistic beginning that, after years of tries and defeats, hope has not lost but the results are far away from being considered a positive answer. However, from so many failures we, by pure chance, discovered an informative missing pattern in three out of the five data sets analyzed, being these three data sets the only three ones genotyped by Affymetrix (now Thermo Fisher Scientific) and the only three ones related to an autoinmune disease, two about MS, the other about asthma. Whether these are true factors for the existence of this informative missing pattern or not remains an open question to us.

We began this work after realizing the limits of the state-of-the-art approaches for genomic profiling, such as wGRS [1,2]. We worked under two generalizations of those common approaches: (1) to fix no limit in the number of SNP loci used in order to take into account very small effects and (2) to use genome-wide haplotypes instead of just genotypes to keep real genetic transmission and transcription patterns. For the first generalization approach we considered different p-value cutoffs that were measured with the Transmission Disequilibrium Test (TDT). For the second generalization approach we decided to build predictors using sliding windows of different sizes instead of

only just 1-snp markers as the input variables of the predictor. To measure p-values in these multi-snp variables we proposed a TDT generalization, $2G - TDT$ [3], wich uses a training-test sampling approach to avoid overfitting due to the exponential increase in variable values with the increase of the window size. To accurately resolve whole-chromosome haplotypes of each individual we used family trios. The method was described in 2012 [2,3]. To graphycally understand the difference of the method with a wGRS, see Figures 1 and 2 at the supplementary material for a wGRS and the method we defined, haplotype-based weighted Genetic Risk Score (hwGRS), respectively. In summary, the method builds a risk predictor for a genome-wide haplotype. To compute individual risk, the two genome-wide haplotypes an individual has, half of their genome inherited from the father and the other half from the mother, are combined assuming a recessive genetic model.

The first data set we tried came from the association study performed on family trios by the International Multiple Sclerosis Genetic Consortium (IMSGC) in 2007 [4] and it seemed that AUC drastically improved. However, we were not able to replicate these results in any of all the other diseases we tried: caries, ADHD and autism, those trio data sets that, by that time, were available at NIH dbGaP. We then decided at least to replicate results with another trio data set of MS_ES with only 80 trios, as we could not afford a larger sample due to our reduced fundings. Negative results again. And what it was worse, we repeated the experiment with another version of the IMSGC data set a year ago and could not replicate the first results (see Figure 1).

We then realized our good results in the first IMSGC data set were returned because of a unconcious data transformation we performed right from the beginning in the IMSGC data set. The transformation was imputing missing genotypes, and it was doing it in offspring in a different



(a) (b)

Figure 1: AUC for different p value (from $2G$-TDT on the training data set) thresholds ("all" means all SNPs were used) and different haplotype sizes (1 to 50). Model was learned from 500 trios randomly chosen (training data set) and AUC was estimated from the remaining 431 trios (test data set).

(a) Results from ped files provided directly by IMSGC before they were released and put in dbGaP. These results were wrongly computed, as it will be explained later. (b) Results from ped files in dbGaP. It has to be noted that source data used for both plots were almost the same, the different was in the software changes that were done by that the time, being unaware of them.

way from their parents. It consisted on imputing missing data by maximizing the transmission of the minor-tdt (mt) allele, i.e., the one with less transmissions in the data set with the missing data or, in terms of the TDT, the allele used to compute the c in the TDT equation: $(b-c)^2/b+c)$. We call this approach the cT approach. This transformation required a different way to compute AUC for it to be correct but we did not know by that time as we were not aware of being doing any transformation at all. In fact, to measure specificity we could not use parents, as there will be less missing snps imputed as mt alleles than in offspring. Once we correctly computed AUC, it significantly drop due to a strong decrease in predictor specificity so that the method was useless (see Figure 2 in which sensitivity and specificity are separately shown). It can be observed that the more sensitive a predictor is, the less specific, with average values being too low.

It can also be observed that sliding windows were of no help, as only with 1-snp windows (black lines) we can observe the changes in predictive outcomes with the change in p-value cutoffs, as it should be expected in complex diseases. There is also a similar behaviour when using an additive approach (see Figure 4), which means that our haplotype-based model was not important at all, at least for the current state-of-the-art in which the main problem is the existence of an informative missing pattern.

However, it was interesting to observe how under both models, genotype-based additive and haplotype-based recessive models, without missing imputation the predictor was biased to type II errors (large specificity and short sensitivity) while when using the cT approach to impute missing SNPs the predictor move to the other extrem, with a majority of type I errors so that the predictor was highly sensitive and lowly specific (see Figures 2 (recessive model) and 4 (additive model) for

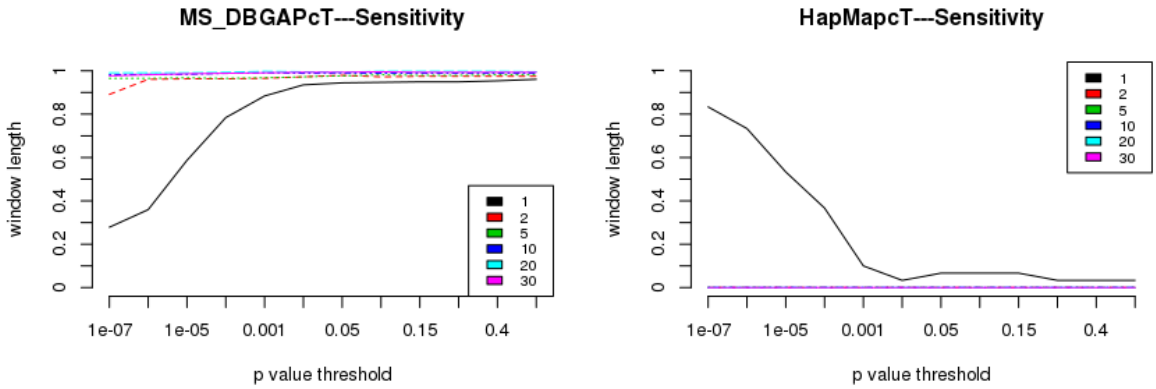


Figure 2: Generalization accuracy of MS predictor: sensitivity (left plot) and specificity (right plot) using the cT approach under the recessive genetic model. Training and test data sets and p-values were obtained as for the experiments whose results are displayed in 1.

sensitivity and specificity under the cT approach and Figures 3 (recessive model) and 5 (additive model) with no data transformation made, i.e. keeping missing data).

In fact, the transformation we made by chance gave us the opportunity to understand better the data we have in our hands. We observed an informative missing pattern consisting on higher missing rates in affected individuals before any transformation and this fact may throw some light into the way to approach the problem (see Figure 6). The pattern can be observed in all the chromosomes.

We have also recently confirmed the same pattern in MS_ES (see Figure 7) and in a trio dataset of asthma collected by the GALA I study [5] (data not shown, see post scriptum at the researcher notebook).

However, the pattern could not be observed in the other trio data sets we had with affected

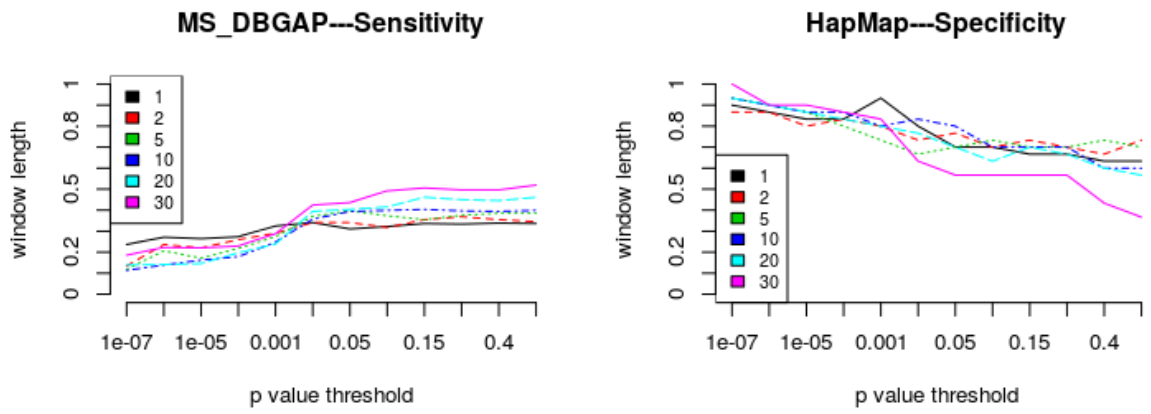


Figure 3: Generalization accuracy of MS predictor: sensitivity (left plot) and specificity (right plot) without imputing missing genotypes under the recessive genetic model. Training and test data sets and p-values were obtained as for the experiments whose results are displayed in 1.

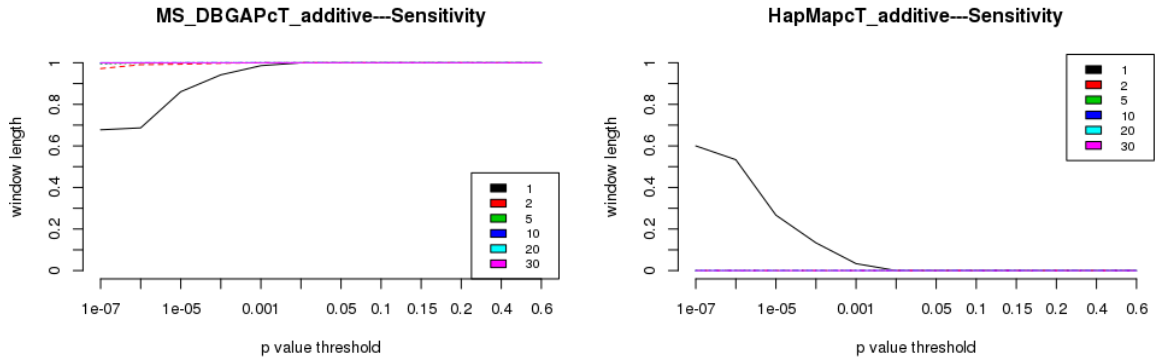


Figure 4: Generalization accuracy of MS predictor: sensitivity (left plot) and specificity (right plot) using the cT approach under the additive genetic model. Training and test data sets and p-values were obtained as for the experiments whose results are displayed in 1.

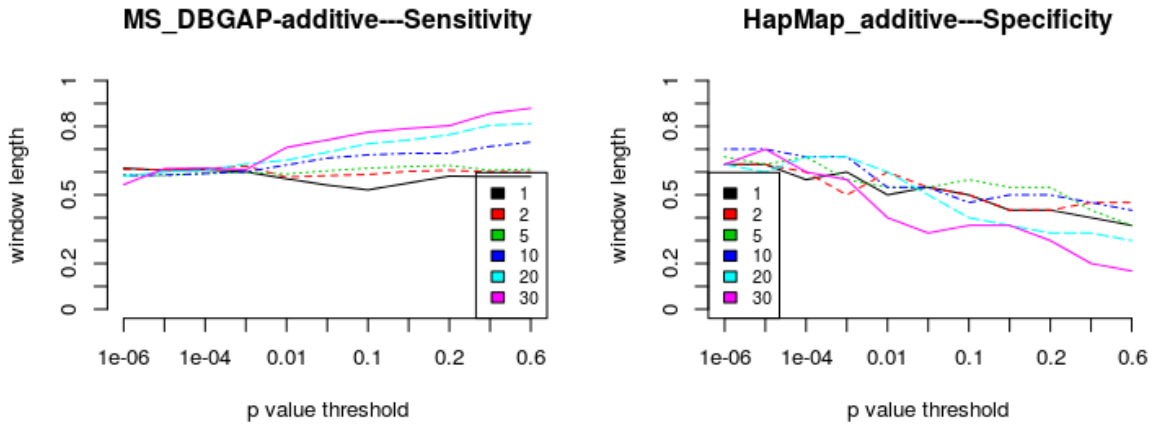


Figure 5: Generalization accuracy of MS predictor: sensitivity (left plot) and specificity (right plot) without imputing missing genotypes under the additive genetic model. Training and test data sets and p-values were obtained as for the experiments whose results are displayed in 1.

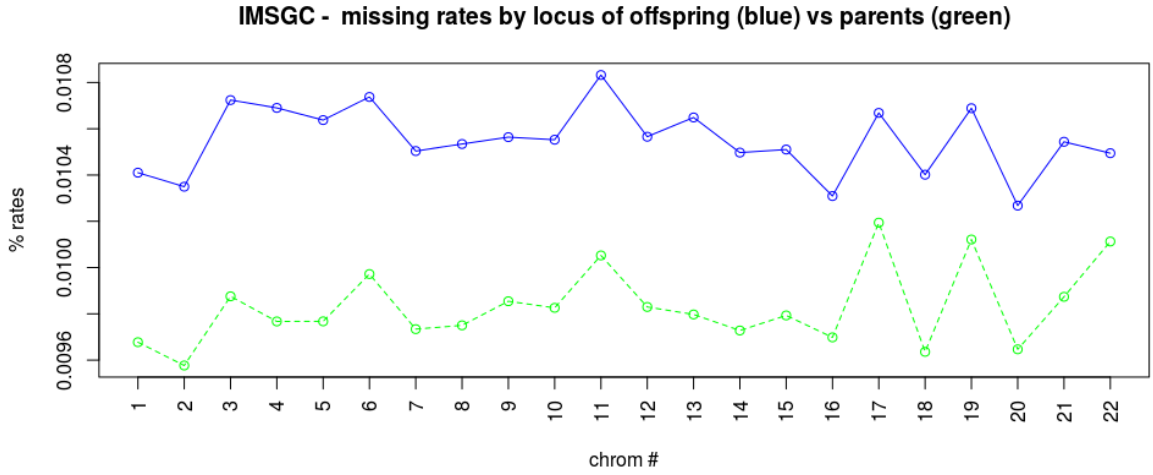


Figure 6: Line plots with average missing rates by chromosomes for parents and offspring in IMSGC MS data set. Plots were produced by using the whole data set with 931 trios.

Figure 7: Line plot with average missing rates by chromosomes for parents and offspring in MS_ES data set. Plots were produced by using the whole data set with 80 trios.

offspring, or it was not constant for all the chromosomes. In fact, in ADHD there is a very light inverse pattern of larger missing rates in parents for all the chromosomes (see Figure 8). In caries there is not a clear pattern, as rates are higher in parents for some chromosomes and lower for others (see Figure 9).

We used an Affymetrix trio data set with all control members to make sure the informative missing pattern favouring offspring missing rates was not there. In fact, we used the HapMap CEPH trio data set from the International HapMap Project [6] (see Figure 10). Results were surprising for us, as the pattern was exactly the opposite one: a sound higher rate of missing genotypes in parents. But the reason seems to be clear: parents only share half of their genomes with other trio member (the offspring) while offpring share all their genotypes (half with the father and half with the mother) which means that all rare variants in offspring will appear in some parent but not the other way around, so that it turns out that rare variants are in average, less rare if carried by

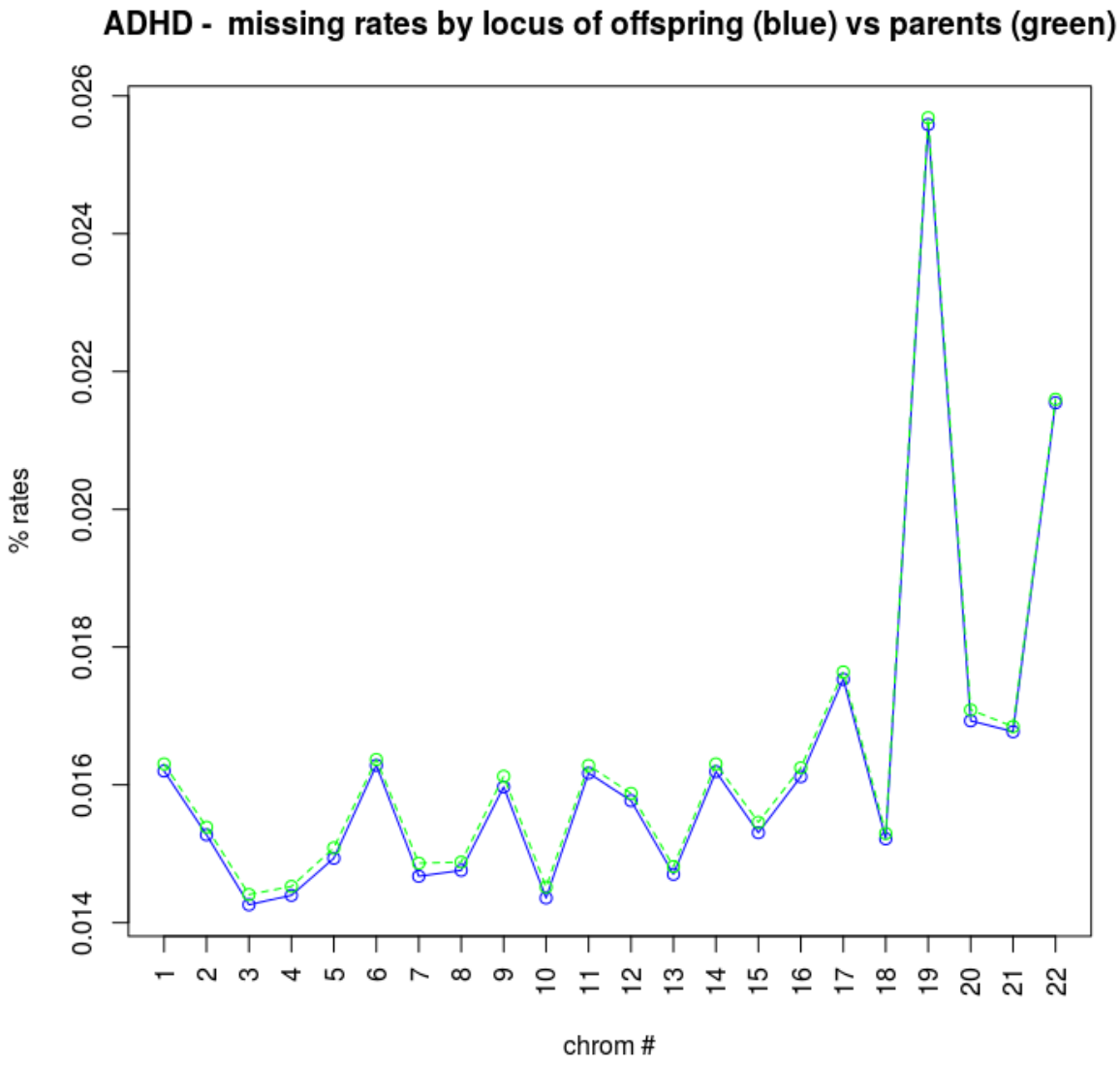


Figure 8: Line plots with average missing rates by chromosomes for parents and offspring in ADHD data set. Plots were produced by using the whole data set with 896 trios.

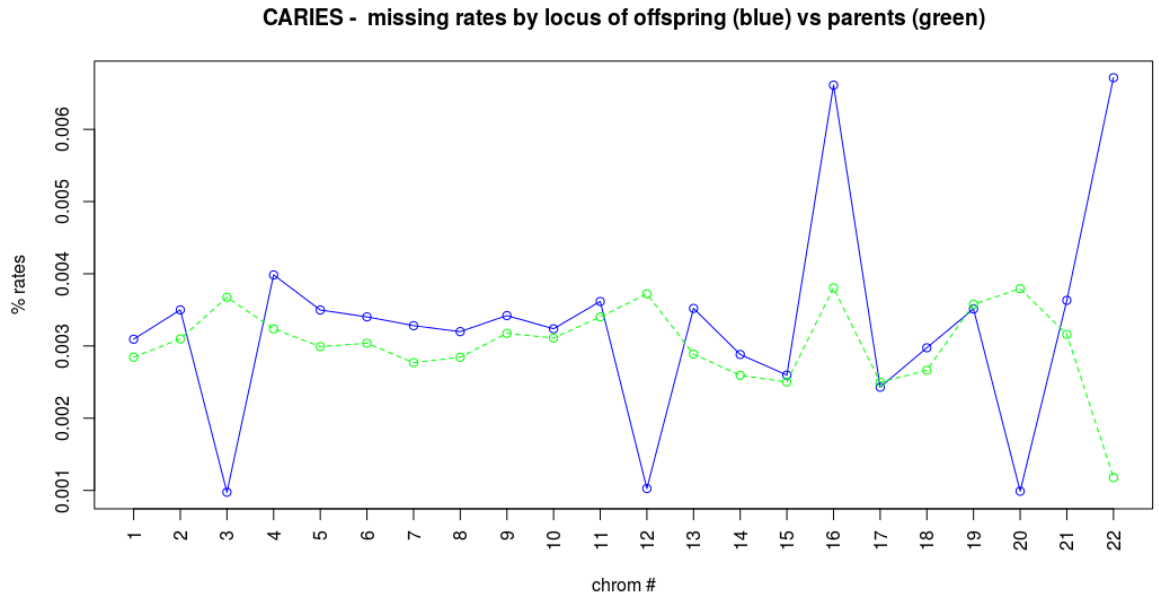


Figure 9: Line plots with average missing rates by chromosomes for parents and offspring in Caries data set. Plots were produced by using the whole data set with 178 trios.

offpsring than if not. As rare variants are more difficult to be called and Affymetrix technology seems to be very cautious incalling uncertain genotypes there will be more missing genotypes in parents than in offspring. For this same reason it should be more rare variants correctly called in offspring than in parents in control trio data sets (just an hypothesis, confirmation not even tried).

1 Discussion

There are two common properties in the data sets where the pattern was observed, they are autoinmune diseases and they were genotyped with Affymetrix technology (see Table 1). We suggest that the informative pattern of missing genotypes have to be deeply investigated as predictive results may completely change. In fact, on one hand we may impute missing data using the cT approach, i.e. favouring which are low-risk variants in the incomplete data set. As in complex diseases there

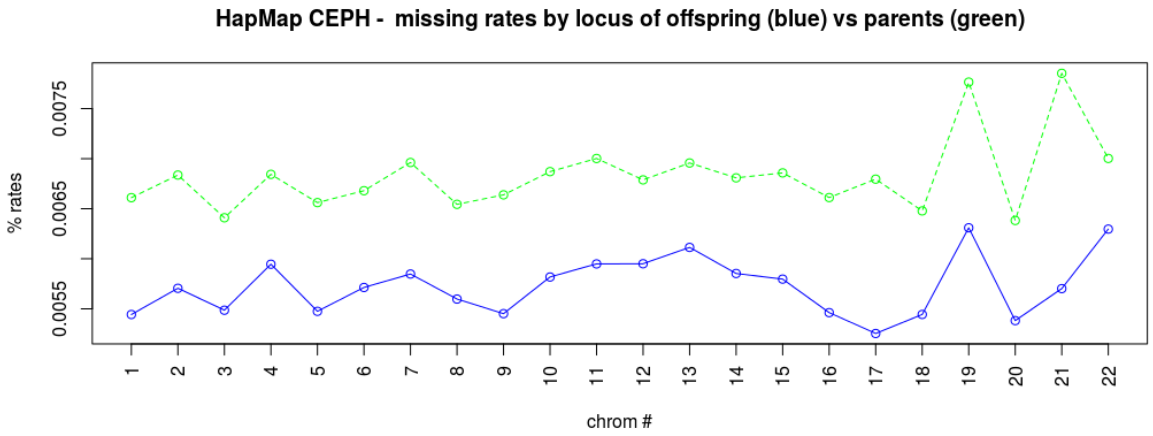


Figure 10: Line plots with average missing rates by chromosomes for parents and offspring in HapMap CEPH data set. Plots were produced by using the whole data set with 30 trios.

trait	# trios	source	platform	number of SNPs
ADHD	896	Perlegen	$PERLEGEN-600K$	599171
asthma (GALA I)	489	Affymetrix	$AFFY\_6.0$	934940
Autism	1323	Illumina	$ILLUMINA\_Human\_1M$	1069796
Caries	178	Illumina	$Human610\_Quadv1_B$	601273
IMSGC	931	Affymetrix	$Affymetrix500K$	262264 ($Mapping250K\_Nsp$) 238304 ($Mapping250K\_Sty$)
MS_ES	80	Affymetrix	$AFFY\_6.0$	934940

Table 1: Data sets used with sample size after passing QC, genotyping arrays and number of SNPs.

may be thousand variants with very small effect, this approach may cause a low-risk variant of a small effect marker to become a high-risk variant of a larger effect marker, in such a way that we will have a trend to raise type I errors, with higher power and lower specificity. On the other hand, if we do not specially consider missing genotypes or we use other genotyping technologies without this informative missing pattern, detected risk variants may be biased towards the major allele, which is not a problem as far as it is also the high-risk variant. However, whenever a high-risk variant is associated with the mutant allele, as it is more difficultly ascertained and there are more missing calls, its effect may be underbiased. Therefore, predictors built on data with missing genotypes or with algorithms that forced genotype calling towards the major allele, may have a trend to mostly type II errors, showing very low power and high specificity. This situation is in agreement with the state-of-the-art genomic profiling of complex diseases, in which AUC is much lower than it should be and it is caused by a lack of power. A similar trend against the high risk allele whenever it is also the minor allele, may occur in algorithms used to perform missing imputation and haplotype reconstruction. As genetic causes of complex diseases may come from hundred thousands of variants with very small effects, we also wonder whether current procedures for quality control may not only remove noise but also a true signal. From our results there is very little to conclude, except to questioning the current genotyping technologies and to give hopes to the research about genomic profiling if we were able to correctly call rare variants instead of giving wrong calls or missing ones.

However, we believe the real interest of this work can only be understood when reading the research notebook supplied as supplementary material. We consider it as a profound experience of the backwards of the positivist science, wich is understood by many experimental or exact

scientists, as the only science and not just as one side in the search of knowledge in experimental, exact and technological sciences and engineering.

Many scientists educated under a positivist paradigm, are loosing hope by considering some results of experimental science. They claim that this "science" is often not able to question current paradigms and propose new ones, as researchers have a strong risk of being highly conditioned by economical or productivity interests. Through the time this work took to be finished, the first author completely changed the approach to do research. She started as in other experimental works, considering herself as the author of her ideas and willings and following her thinkings proudly and deeply convinced she had the skills needed to find a groundbreaking solution. She just was able to give some credit to other people when she understood the need of colaboration with other researchers, firstly biologists and mathematicians. For years she could not stop working with no progress at all and throwing away her personal life. At some point and thanks to an umpredictable life change, she started praying and listening to her "heart" as a way to discern God's will and used her mind only to follow the heart, instead of considering her mind as a real god. The main risk of using her only prayer to get this knowledge was self-deception or what, in terms of positivist science, uses to be called research bias. As she experimented the no sense of self-deception several times, she started discerning through a "lab" called Church with "referees" called priests. Unbelievers may smile to this opinion but we all deeply know what means to listening to our heart instead of lying on our mind as the only way to connect with the Truth or to reach some truths. We all also know the consequences of being stuck to an idea without listening to other opinions. A few years after she started working at the University of Granada in 1990 and until five years ago she lived denying and being really ungrateful with God, just because she thought her

faith could cause her problems to her crazy professional ambitions in such a positivist environment. As a help for those unbelievers who, opposite to her, are not really responsible of denying their believes and are really open to the Other, she would just quote Blaise Pascal: ”Kneel down ... and you will believe”.

Acknowledgements This work has been supported by the Spanish Research Programs “Fondo de Investigacin Sanitaria (FIS)-Instituto de Salud Carlos III (ISCIII)” [grant numbers P12/00555, PI13/01527, PI13/02714 and PI13/01466] and SAF [grant number SAF 2016-80595-C2-1-P] (Ministerio de Economía y Competitividad); the Andalusian Research Programs CTS [grant number CTS2704] and “Proyectos de Excelencia” [grant number P08-TIC-03717] (Junta de Andalucía). All of them have been cofunded by the European Regional Development Fund (ERDF). We want to particularly acknowledge the patients and the control subjects in this study for their participation and to the Nodo Biobanco Hospitalario Virgen Macarena (Biobanco Sistema Sanitario Público de Andaluca) for its help and support in the gifts of clinical samples used in this work. The Biobank is integrated in the Spanish Biobanks Network (RetBioH;www.redbiobancos.es), and supported by Instituto de Salud Carlos III, integrated in the national I+D+i 2013-2016 and co-funded by European Union (ERDF/ESF, Investing in your future) (Grant n PT13/0010/0041). We acknowledge Prof Esteban González Burchard as principal investigator of the GALA I project, for giving us access to the family trios data set of genotypes. We also acknowledge Dr Chris Gignoux and Dr Fernando Molina for their help in this work. The first author is specially grateful to Prof Dr. Ildefonso Camacho SJ who taught me that research and every other thing we do in life is only truly worth when we do them being in peace with God and ourselves; to Dr. Marco Ramoni † and Prof. Paola Sebastiani, who introduced her into this work; to Antonio Pertìñez, who was her husband when she started this work, for suffering the consequences of it without any choice; and to some other priests, mostly at the community of jesuits in downtown Granada, the one Ildefonso also belongs to, and especially to Dr. P. Gonzalo Villagrán

SJ and P. Manuel Díaz SJ who also teach at the Faculty of Theology in Granada. Without them we would have never been able to put an end to this work.

Disclaimer note Some of the co-authors of this work may not be in agreement about the personal thoughts and way to explain results the first author has shown in this paper.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to M.A.G. (email: mabad@ugr.es).

1. 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).

2. 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).

3. 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).

4. '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).

5. Pino-Yanes, M. *et al.* Genome-wide association study and admixture mapping reveal new loci associated with total ige levels in latinos. *Journal of Allergy and Clinical Immunology* **135**, 1502–1510 (2015).

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