Disentangling the genetic causes of autism

María M. Abad-Grau[1]

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

By suspecting that the scarce results obtained by several genome wide association studies, even in highly heritable diseases, could be due to wrong assumptions inferred from reference genomes but not true in affected individuals, we re-analyzed data from the Autism Genome Project (AGP) Consortium starting from the genotype calling phase and trying to made the minimum assumptions possible, and identified patterns of individual raw genotypes associated to the disease, mainly lack of hybridization and cross hybridization. In light of it, we re-designed the polygenic model of individual risk and were able to explain most of the variance – AUC larger than 0.9 - in an independent dataset of whole exome sequencing provided by the Autism Sequencing Consortium. The outstanding analytical results pointed out to several highly significant positions, most of them having the autistic individuals as the group with the unknown variants. By performing functional annotation clustering, the most significant cluster (5.6 enrichment score) put together three categories related to the nuclear pore complex (NPC) interacting protein. We also analyzed sequenced genotypes of Aspergers individuals and found out that they are much closer to individuals with typical development than to strict autistic ones.

1 Introduction

Since the end of the *Human Genome Project (HGP)*, more than 20 years of work of countless people and billion of dollars have been invested in *Genome Wide Association Studies (GWAS)* trying to find loci associated or in linkage with associated loci to complex diseases. Frequently, *de novo* variants are commonly presumed as possible explanations of results barely explaining a limited proportion of the variance, in studies that keep growing in sample size with the hope to increase the amount of significant results. While according to risk prevalence and disease heritability in several complex diseases, inherited variants should be found able to explain most of the cases, most of current results are far worse of what they should be. As an example, in Multiple Sclerosis (MS), by using a formula that related the *Area Under the ROC Curve (AUC)* –a measure of accuracy in risk predictors– with disease heritability and risk prevalence [1], maximum AUC should be around 0.95 for all European countries, from the one with the lowest prevalence (Spain) to those ones with the largest prevalences (Iceland and Sweden). However, genetic risk scores barely reached 0.70 of AUC [2].

In the case of Autism, a disease with a very high disease heritability (in fact, if an identical twin has autism, the likelihood that the other twin will also have is higher than 90%), it would be 0.997 in Bangladesh, the world country with the lowest risk prevalence and 0.999 in Singapore, the world country with the largest prevalence. After realizing how far are we to build genetic risk scores with these AUC values, we had the intuition that the clue could arise by questioning the first preprocessing steps, mainly genotype calling and quality control.

In fact, we found genotype calling methods usually removed a true signal by building prior models based on common healthy populations, mainly using data from the *International Hapmap Project (IHMP)* [3] or the *1000 Genomes Project* [4]. We also discovered quality control procedures were so stringent that true signal was sometimes disregarded. By trying to study the pattern of missing and Mendelian inconsistent (MI) loci in different complex diseases, we found that these patterns were far a way of being at random, so they should be considered as part of the genetic risk models. When using a case dataset to build prior models in a genotype calling method instead of a control or reference data set, there was a largest proportion of missing calls in a control dataset [5]. We focused on autism because of its very large inheritance and risk prevalence, in order to reduce environmental confounders as much as we could have.

We were aware of the existence of at least two different sources of missingness, that should be differentiated: one is an uncertain call when raw intensities lies in between two of the three classic genotypes (homozygous wild type, homozygous mutant and heterozygous) and the other is an uncertain call because both-alleles raw intensities are too low to sustain underlying hypothesis of an individual to have one of the three possible genotypes (non-hybridized probes). Genotype call algorithms usually does not make this difference and most GWAS ignore those missing genotypes identified by the calling algorithm. Other problem related to the state-of-the-art genotype call algorithms, was their per-loci approach (see Figure 1). By using this approach and, even worst, computing prior distributions based on reference genomes, a high bias towards control genotypes is usually introduced. Moreover, a considerable amount of loci do not show the common three-clusters pattern (AA/AB/BB genotypes) or they are very close to each other to perform a correct

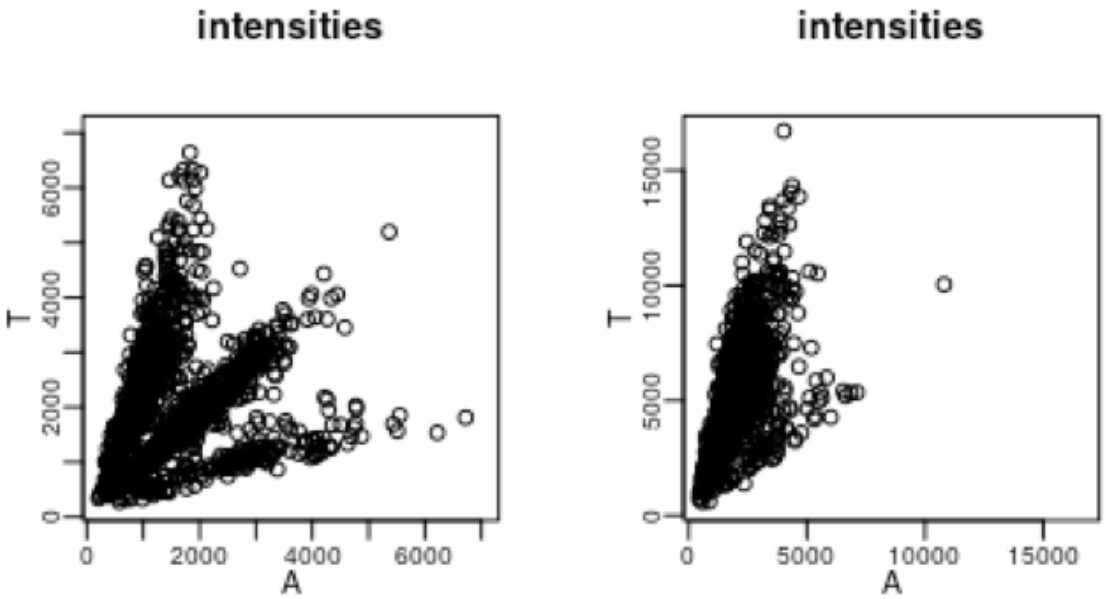

Figure 1: Scatter plots from raw intensities, θ_A, θ_B, the two alleles, one per axis, of two snps. The plot at the right shows that raw intensities are already too close of each other to perform a correct genotype call.

classification (see as an example the scatter plot at the right in the same figure).

2 Results and methods

Genotype calling For our goal of detecting non-hybridized snps that can point out to an alternative variant, it would be more important to have low bias towards the reference genome than low error rates. Therefore, we developed *Individual Genotype Call (IGC)*, a genotype call procedure able to identify as missing only genotypes compatible with non-hybridized probes. To avoid biased genotype calls, genotypes are inferred by focusing on only one sample at a time. Although this way,

we would not need to use prior models for intensities, we observed very different log-raw-intensity scatter plots (we will refer to them as gentprints) from different individuals and we did not always succeed in clustering identification by assuming different statistical distributions neither by using computing vision methods for object recognition. As an example, we tried to perform genotyping from a 3D gentprint using frequencies as the z axis and a Hessian matrix (second derivative) to propagate from cluster peaks to valleys (see in Figure 3 (a) a 3D gentPrint map for a case where most genotypes are easily recognized by view as belonging to AA, AB or BB group). However, the inclusion of a missing group makes clustering more difficult. In order to understand which points in a gentprint should be classified as missing due to both intensities too low, we make coloured gentprints and used a different color for autosomes (black), chromosomes X (orange), Y (cyan) and PAR-XY (magenta). As females do not have Y chromosome, cyan dots in female gentprints can give an idea of where non-hybridized points in any other chromosome should be placed within the gentprint (see Figure 2 for a trio from CEPH IHMP and array Illumina 1MDuo). Gentprints in this figure also give an idea of the large plot variability within the same family and support the idea of using an individual-based approach instead of the position-based usual approach with a normalization procedure. In fact some genotypes were wrongly classified as missing (Figure 3 (b)).

As an alternative, we decided to use an artificial vision approach to build a model. A human expert did the first-phase training, by performing genotype clustering by view of the log-raw-intensities individual maps of 40 samples (see an example of an "irregular" gentprint on Figure 4 (a)) . In a second-phase during model training, the model was applied on other 180 samples and

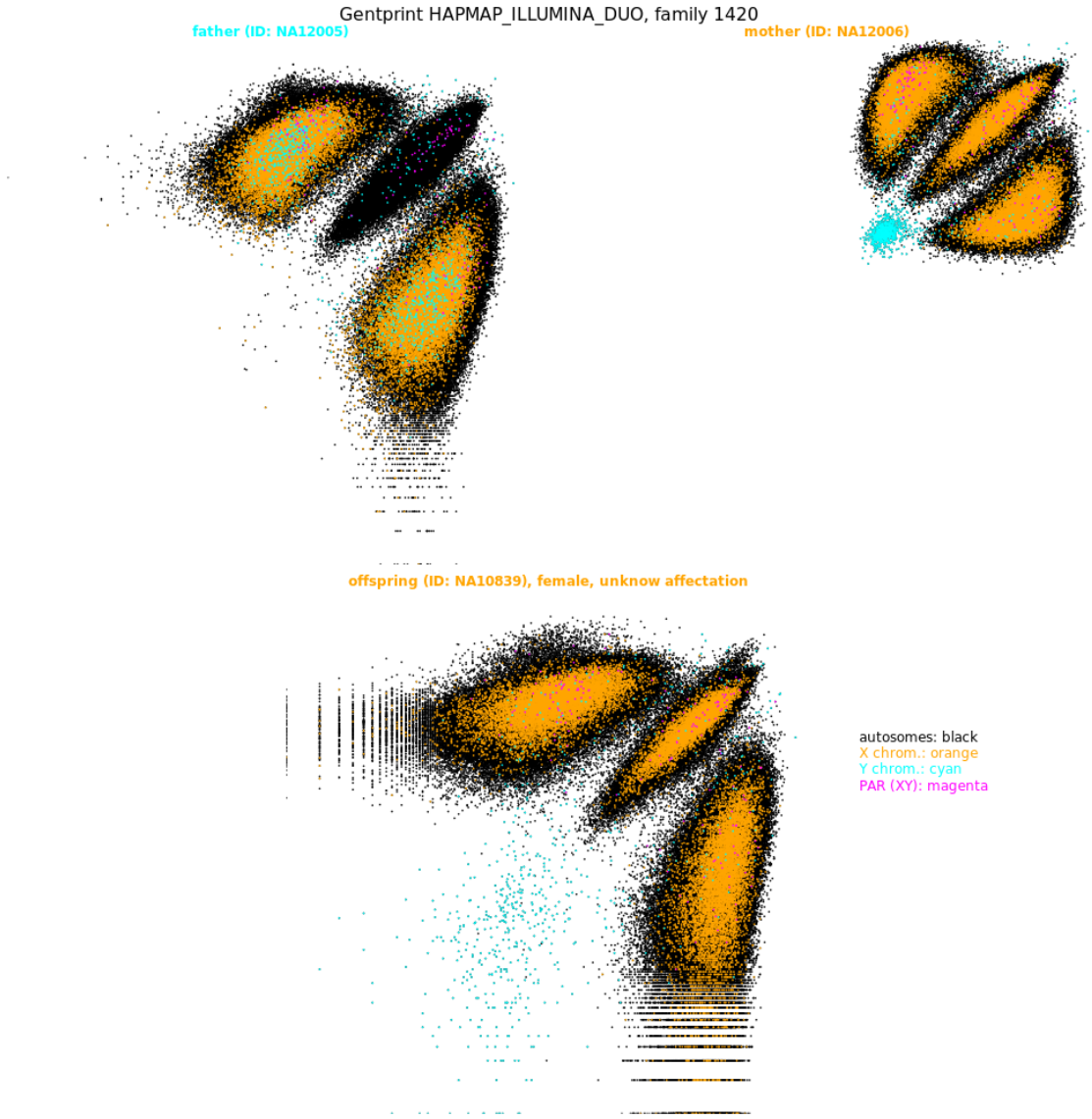

Figure 2: Gentprints (log of raw genotype intensities) from a daughter (at the bottom) and her parents (father on the left, mother on the right). Axis range are the same for the three family members. It can be observed large differences among them, not only in intensity ranges but in shapes.

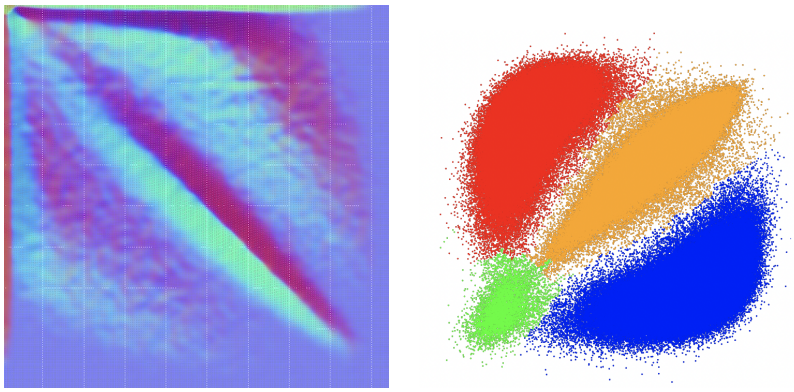

Figure 3: GentPrints to show a computing vision method for genotype clustering. (a) A 3D gentPrint map (Y axis is inverted because it is a graphical image, not a statistical plot) where it can be observed the intuition about the genotype calling method. (b) A gentprint with already classified genotypes as blue (AA), orange (AB), red (BB) and green (missing). Some genotypes classified as missing seem to belong to known alleles.

the human expert just accepted or corrected the clustering of each new image (it was a yolov8n model trained with Roboflow instance segmentation with 96.1% sensitivity and 96.1% precision in the test data subset). The model must be applied on new data in order to have the four irregular polygons on the new gentprint (an example of model inference can be seen in Figure 4 (b)). The computer vision watershed algorithm is utilized afterwards, considering the points inside polygons as local minima, in order to classify all the genotypes (see Figure 4 (c)). It has to be observed that samples used to build the model should not bias genotyping outcomes the way they do when using the classic by-locus approach. In fact, by using a instance-based approach instead of a per-locus approach, the model focuses on each individual and is much less influenced by other individuals, helping to identify four regions by adapting itself to very different maps. It just learns

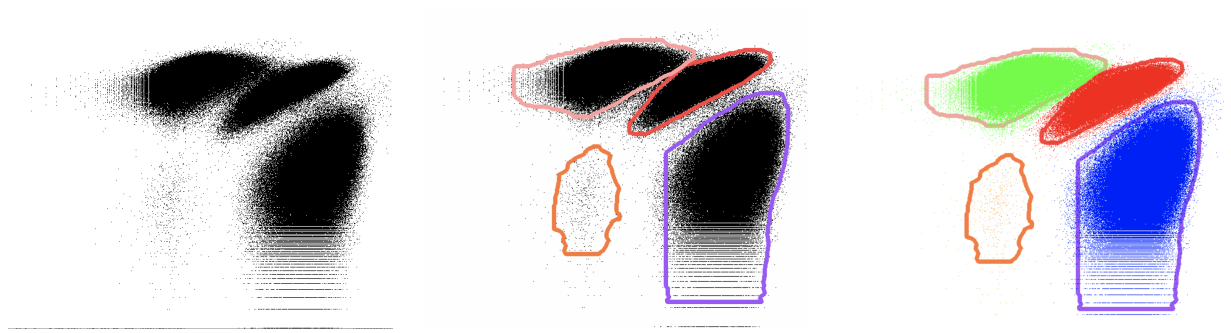

Figure 4: A summary of the IGC procedure to perform genotype calling on a given individual. First we produce an individual log-raw intensities map for all positions and chromosomes (a). Then we infer clusters using yolov smart polygons (b). And finally we classify the points outside de polygons as well by using the wathershed algorithm.

to identify always 4 labels, even if there are very little points in some of the clusters (usually the missing genotype) and it classifies as non-missing those points with ambiguous genotypes using the Euclidean distance to initial clusters.

Another advantage of performing genotype call for one array at a time would be the reproducibility of any result (there could be some differences if we improve the visual model but is should not be trait-biased) and the feasibility of a direct development of a disease profiling test just requiring the genotypes of the offspring or their parents, ascertained one by one.

Missingness-based Polygenic Risk Models In the attempt of improving estimation accuracy of individual risk to complex diseases, different *Polygenic Risk Models (PRM)* have been used. Most of them apply logistic regression on a *Polygenic Risk Score (PRS)*, a measure based on computing accumulated risk through different loci [6]. A frequent way to reduce computational time when com-

puting the score, is by selecting loci or by assuming independent contribution of loci or both. The first approach may disregard important loci for a different trait, the second one may upward-bias effect sizes in presence of linkage disequilibrium (LD). There are several problems to reproduce results in a different dataset when using a PRS that selects loci by considering LD, overall heritability, or allele weights from a reference genome: from gross errors due to differences in allele coding that may be carried out from the genotype call step, to problems derived from population stratification when using datasets belonging to different ethnic groups. A PRM based on machine learning may have the same gross errors, but there are many approaches to improve computational efficacy and generalization capacity (results reproducibility) without the need, and the consequent bias increment, of using external information. They also can avoid overfitting by being parsimonious. Some examples are naïve Bayes (which is actually equivalent to a low-biased PRS able to compute weights without using external information [7]), among Bayesian networks, or J48 among decision trees. The main advantage of J48, and decision trees in general, is to be white box models, i.e. the model learned can be easily understood by humans. Still all of these models are far away to achieve the expected AUC that they should reach when considering disease heritability and risk prevalence in many complex diseases, much less they will generalize in a validation dataset from, perhaps, a different population.

Here is where a *missingness-based Polygenic Risk Model (missing-PRM)* comes to sense. A noticeable backward of these models is that they cannot compute any kind of allele risk, which may be very important to know gene variants related to any disease. However, loci risk is enough to find out genes associated to a disease, even if the exact associated variant cannot be directly

known. Other than that, there are many advantages of using a missing-PRM to assess individual risk in a complex disease, especially when the allele-based models did not succeed. The main one being their utilization of missing pattern information, disregarded by allele-based models, and perhaps crucial to reach AUC rates close to their expected values.

We defined the simplest missing-PRM we could think without any prior distributions: an accumulative model from binary values at each locus, with no weights, if we opted for a PRS; or a naïve Bayesian network, or a J48 decision tree in order to use a white-box model. This means only one input variable per chromosome, representing the amount of missing data an individual has at a given chromosome, or only one input variable at all if all chromosomes are collapsed.

An alternative genotype data that we can also use to build a PRM, which also avoid gross errors due to differences in allele coding, especially when the control and case datasets come from different sources, is the number of heterozygous loci per individual. We referred as "hetero-PRM" to a PRM built with this information. This test, assuming the genotype call algorithm is good enough to classify as missing non-hybridized genotypes, can throw some complementary information in order to identify loci associated to a given trait.

We computed AUC of J48 decision trees in a 5-fold cross validation approach using the 30 offspring in CEPH IHMP family trios as controls and a subset with the same size randomly chosen from genotypes of AGP as cases. To avoid array-biased results, both datasets were genotypes using the Illumina 1M Duo array. A decision tree with only one input variable was built for each chromosome (see AUC results in Figure 5). We also computed AUC from a decision tree built

with only one (accumulative) variable for all chromosomes (see Figure 6).

Exome-wide Polygenic Risk Models We used *Wide Exome Sequencing (WES)* data of individuals with autism and other Autism Spectrum Diseases (ASD), their parents and control trios from the *Autism Sequencing Consortium (ASC)* to validate our results from array data. We used the same approaches: missing-based (see Figure 7) and hetero-based PRSs (see Figure 8), which we have called "missing-exome-PRS" and "hetero-exome-PRS" respectively. For the hetero-based PRS, we considered all heterozygous genotypes, regardless the variants, some of which are not SNPs but polymorphic micro-satellite markers. In a first study, we used as control group 24 offspring under the group "typical development" and their parents. As cases we used other 24 offspring under the group "AUTISM", randomly selected from a total of 464 trios within this group.

By considering the sample size is quite small, AUC for missing-exome-PRM using only genotypes from the offspring (basic configuration), seems to be close to what it should be expected (AUC 0.9 when using all chromosomes together, and even larger values for some single chromosomes), as it occurred in missing-PRM (AUC larger than 0.9). However, results from hetero-exome-PRM in basic configuration (only offspring used) were clearly lower (AUC 0.7) than those obtained by hetero-PRM (AUC larger than 0.5). This hopeless result when using an independent WES-based dataset is in agreement with state-of-the-art genetic studies of autism and gives more value to the results obtained when considering (1) an individual-based genotype call algorithm able to perform an accurate identification of missing genotypes, and (2) missingness-based genotype models. Moreover, we confirmed our intuitions of using missing genotypes in light of a

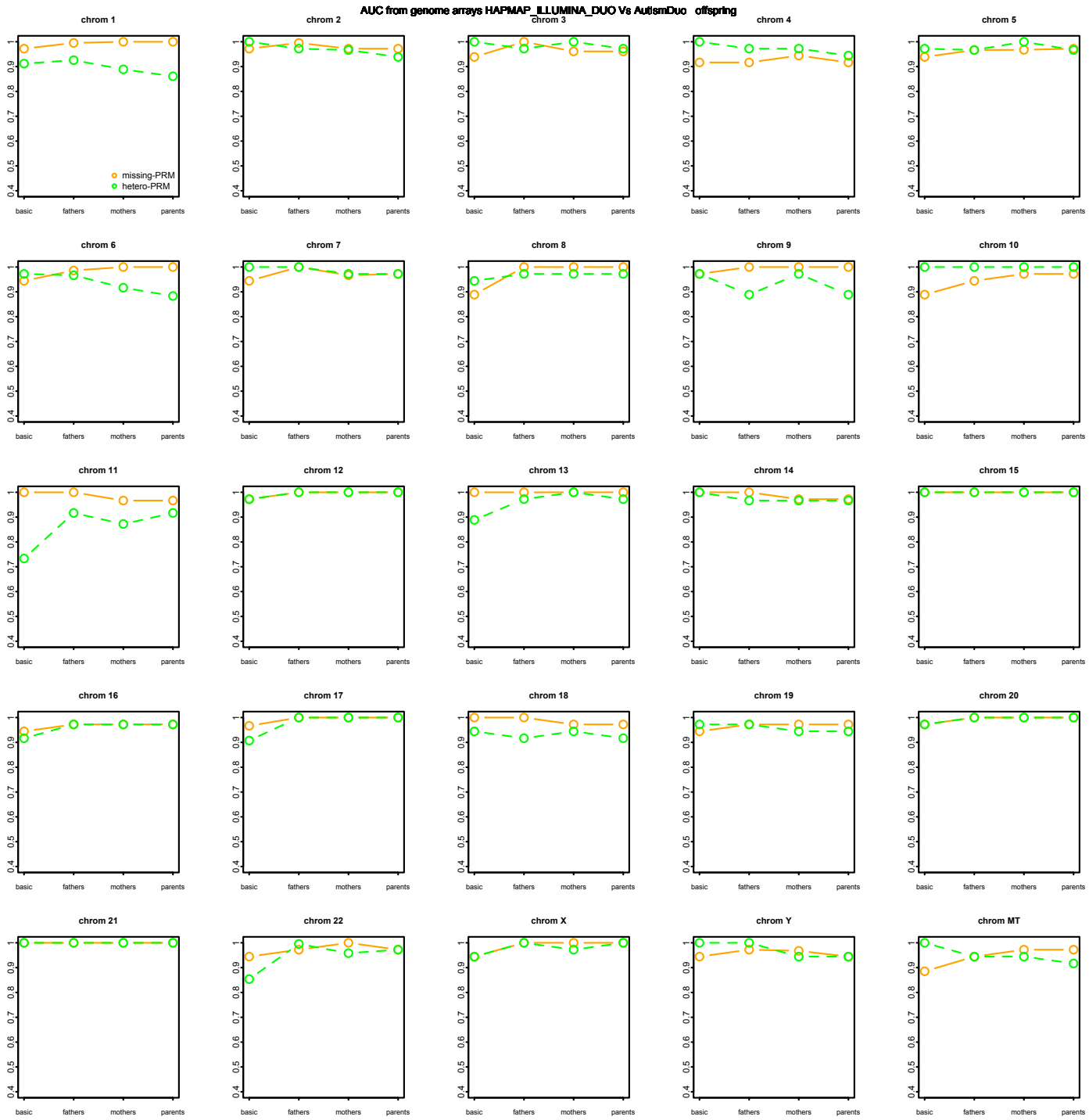

Figure 5: AUC results using a different J48 decision tree per chromosome to predict offspring affectation risk. Within each plot we built decision trees for missing-PRM (orange) and hetero-PRM (green), using counts from offspring (X-label "basic"), or integrating it (by multiplying individual risks) with other PRM from one or the two parents ("fathers", "mothers", "parents").

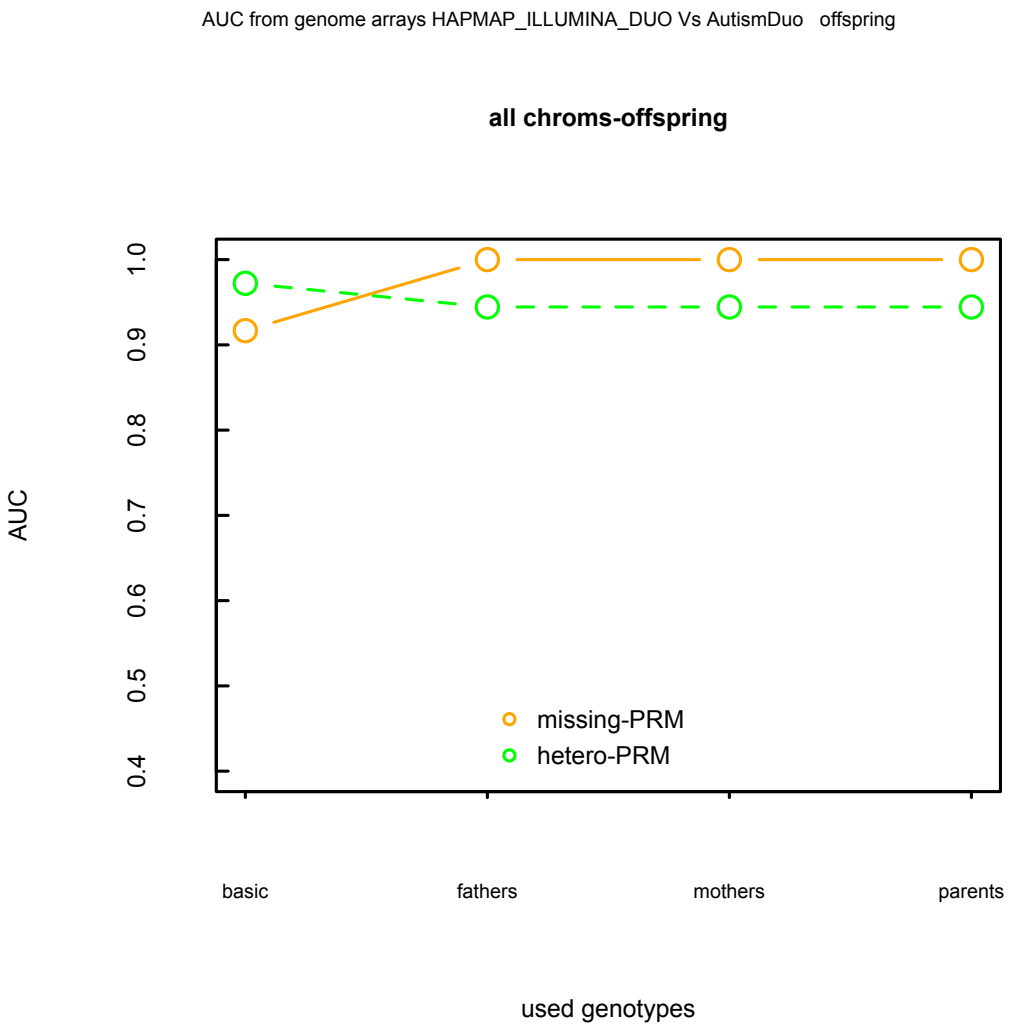

Figure 6: AUC results using only one J48 decision tree for all chromosomes to predict offspring affectation risk. We built two different decision trees: a missing-PRM (orange) and an hetero-PRM (green), using counts from offspring (X-label "basic"), or integrating it (by multiplying individual risks) with other PRM from one or the two parents ("fathers", "mothers", "parents").

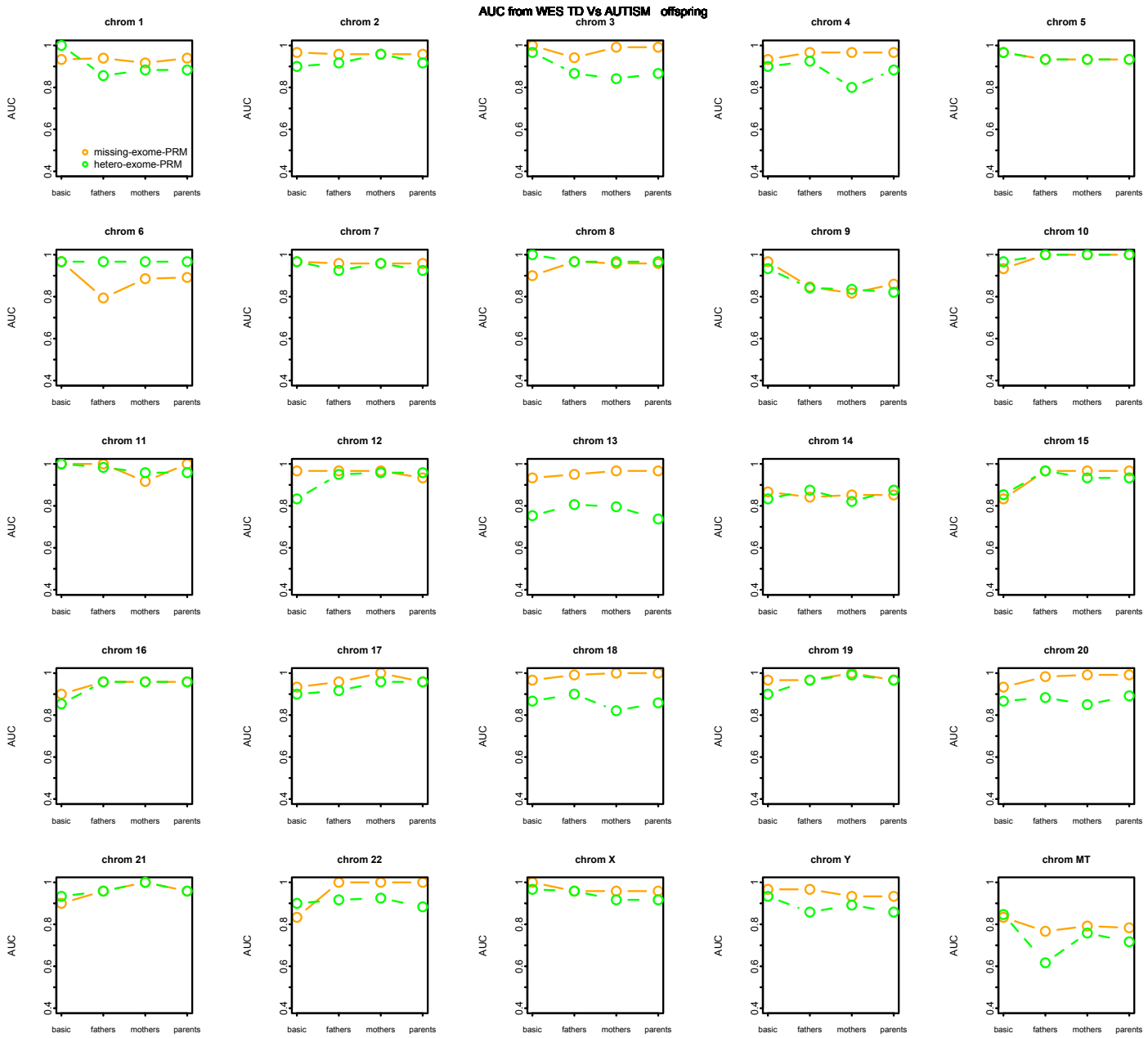

Figure 7: AUC results using a different J48 decision tree per chromosome to predict offspring affectation risk. Within each plot we built decision trees for missing-exome-PRM (orange) and hetero-exome-PRM (green), using counts from offspring (X-label "basic"), or integrating it (by multiplying individual risks) with other PRM from one or the two parents ("fathers", "mothers", "parents").

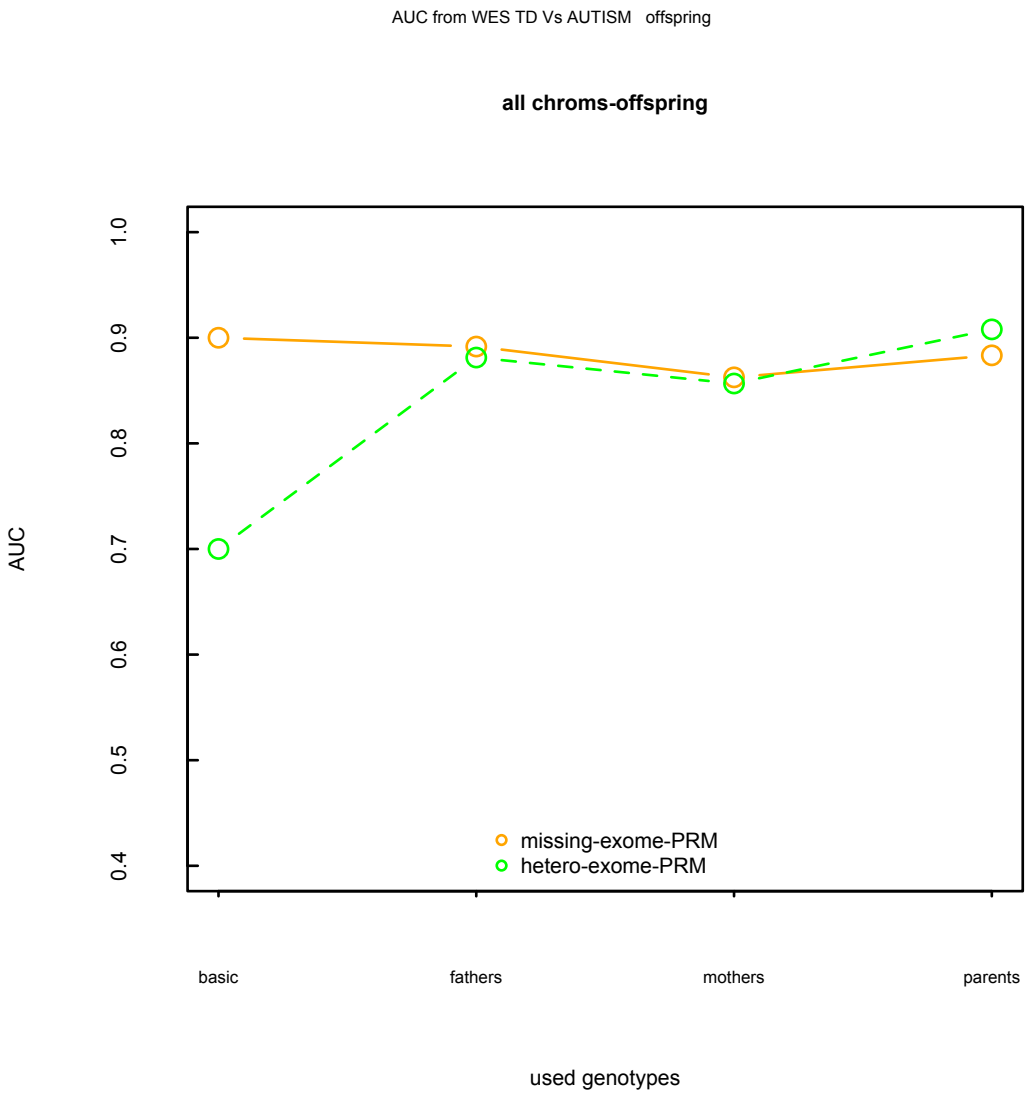

Figure 8: AUC results using only one J48 decision tree for all chromosomes to predict offspring affectation risk. Within each plot we built decision trees for missing-exome-PRM (orange) and hetero-exome-PRM (green), using counts from offspring (X-label "basic"), or integrating it (by multiplying individual risks) with other PRM from one or the two parents ("fathers", "mothers", "parents").

preliminary study [5].

Aspergers versus autism As the ASC provides genotype data from different ASD groups, we selected Aspergers individuals in order to check whether differences with controls can be ascertained reaching similar levels of AUC. We run analogous models and observed a negative answer whe using only the offspring genotypes (see Figures 9 and 10 for one decision tree by chromosome and one decision tree for all chromosomes, respectively).

This result shows that Aspergers individuals are genetically closer to controls compared with autistic ones. However, if we observe AUC when using parents, we found AUC values larger than 0.9 in missing-exome-PRM. Perhaps some missing variants in parents were repaired in offspring, although we would need a deeper study to probe this.

To test whether Aspergers are closer to TD individuals or to autistic ones, we built equivalent PRMs considering Aspergers and autism individuals as the two groups (see Figures 11 and 12).

The first important result is that AUC is larger than 0.9 for both missing and hetero exome-PRMs, which means that the two traits, Aspergers and autism, are genetically differentiable.

If we compare these results with those obtained from PRM between TD and Aspergers, with AUC 0.7, we can conclude that Asperger is a trait genetically closer to typical development (TD) than to autism, which is in agreement with the idea of Aspergers not to be considered a disease.

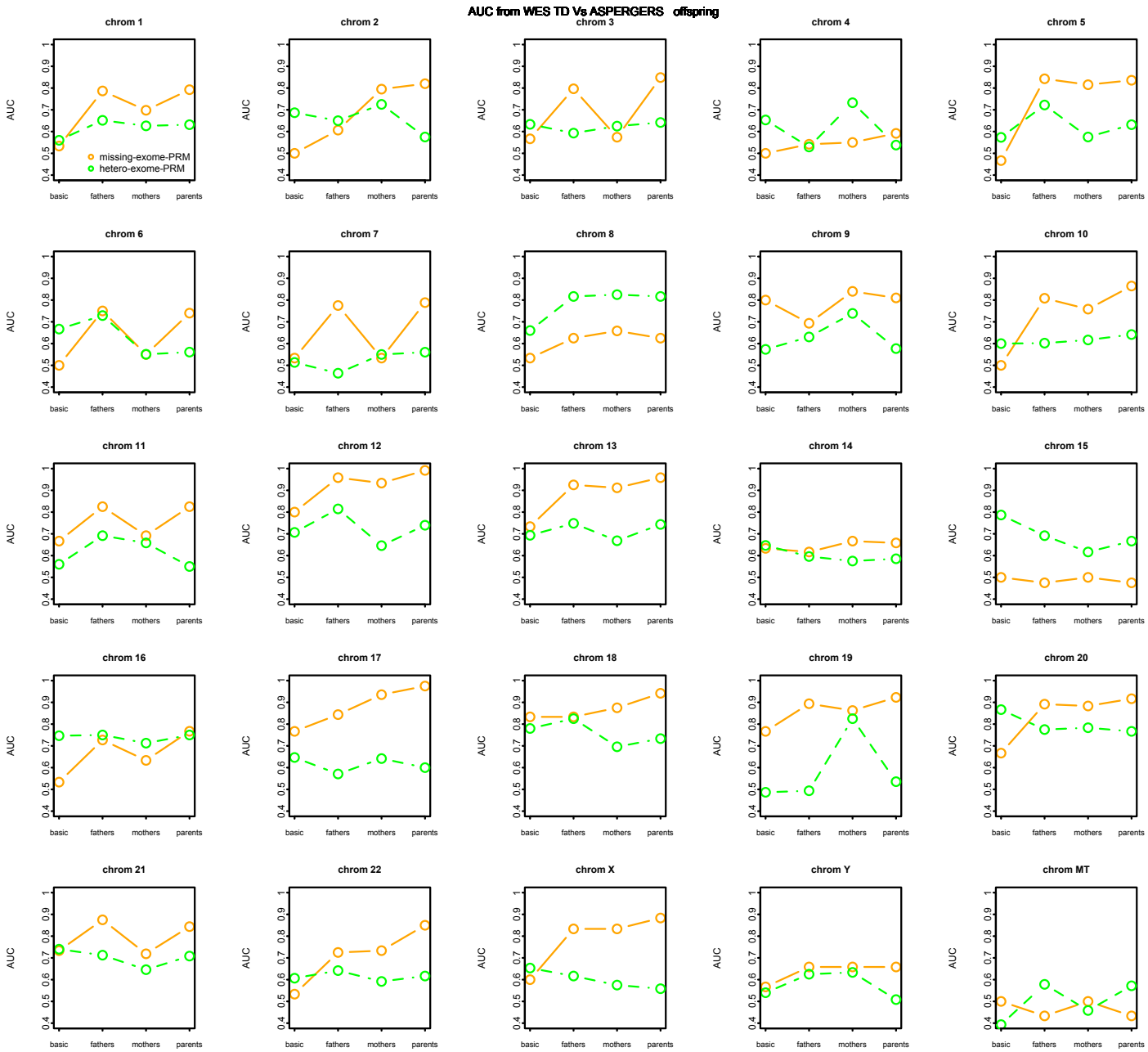

Figure 9: AUC results using a different J48 decision tree per chromosome to predict offspring Aspergers affectation risk. Within each plot we built decision trees for missing-exome-PRM (orange) and hetero-exome-PRM (green), using counts from offspring (X-label "basic"), or integrating it (by multiplying individual risks) with other PRM from one or the two parents ("fathers", "mothers", "parents").

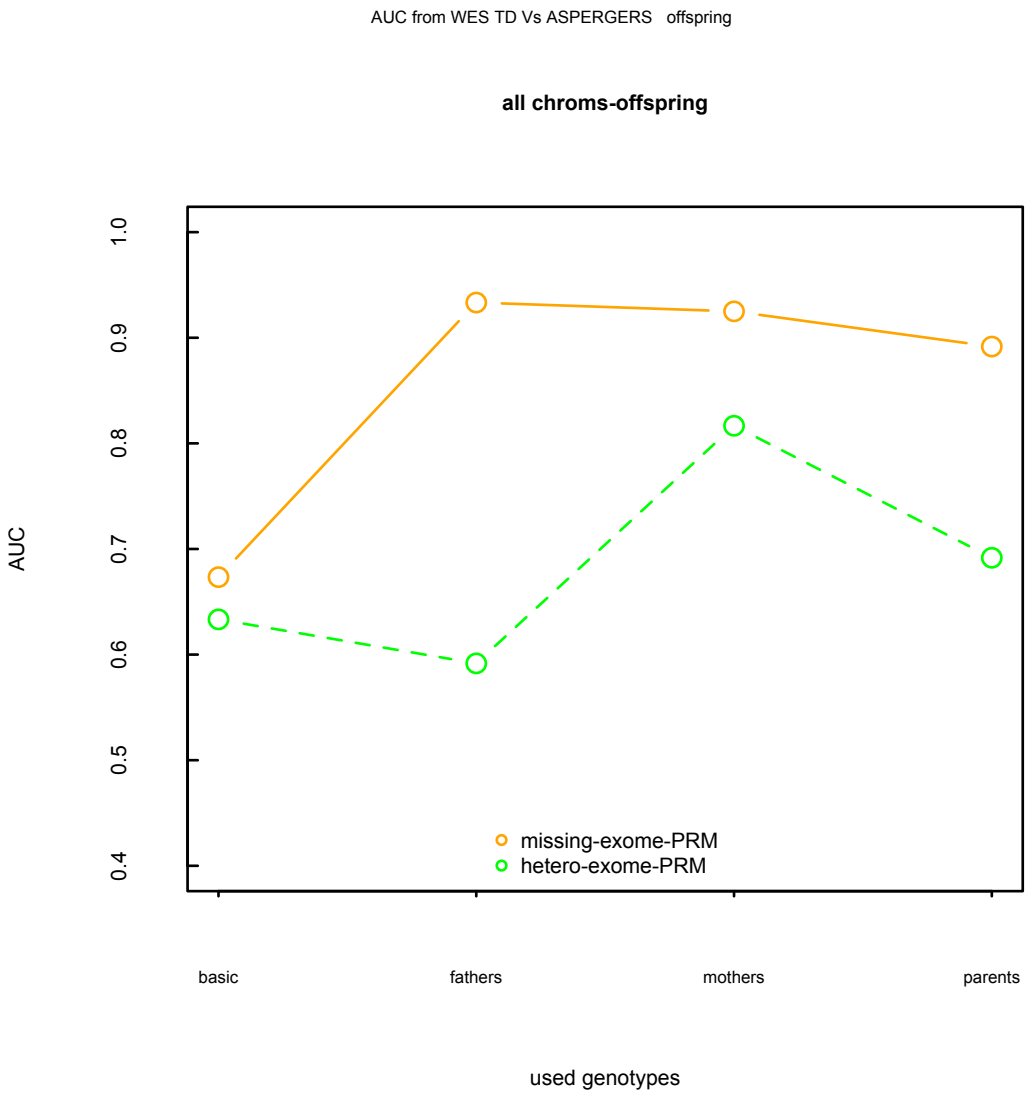

Figure 10: AUC results using only one J48 decision tree for all chromosomes to predict offspring Aspergers affectation risk. Within each plot we built decision trees for missing-exome-PRM (orange) and hetero-exome-PRM (green), using counts from offspring (X-label "basic"), or integrating it (by multiplying individual risks) with other PRM from one or the two parents ("fathers", "mothers", "parents").

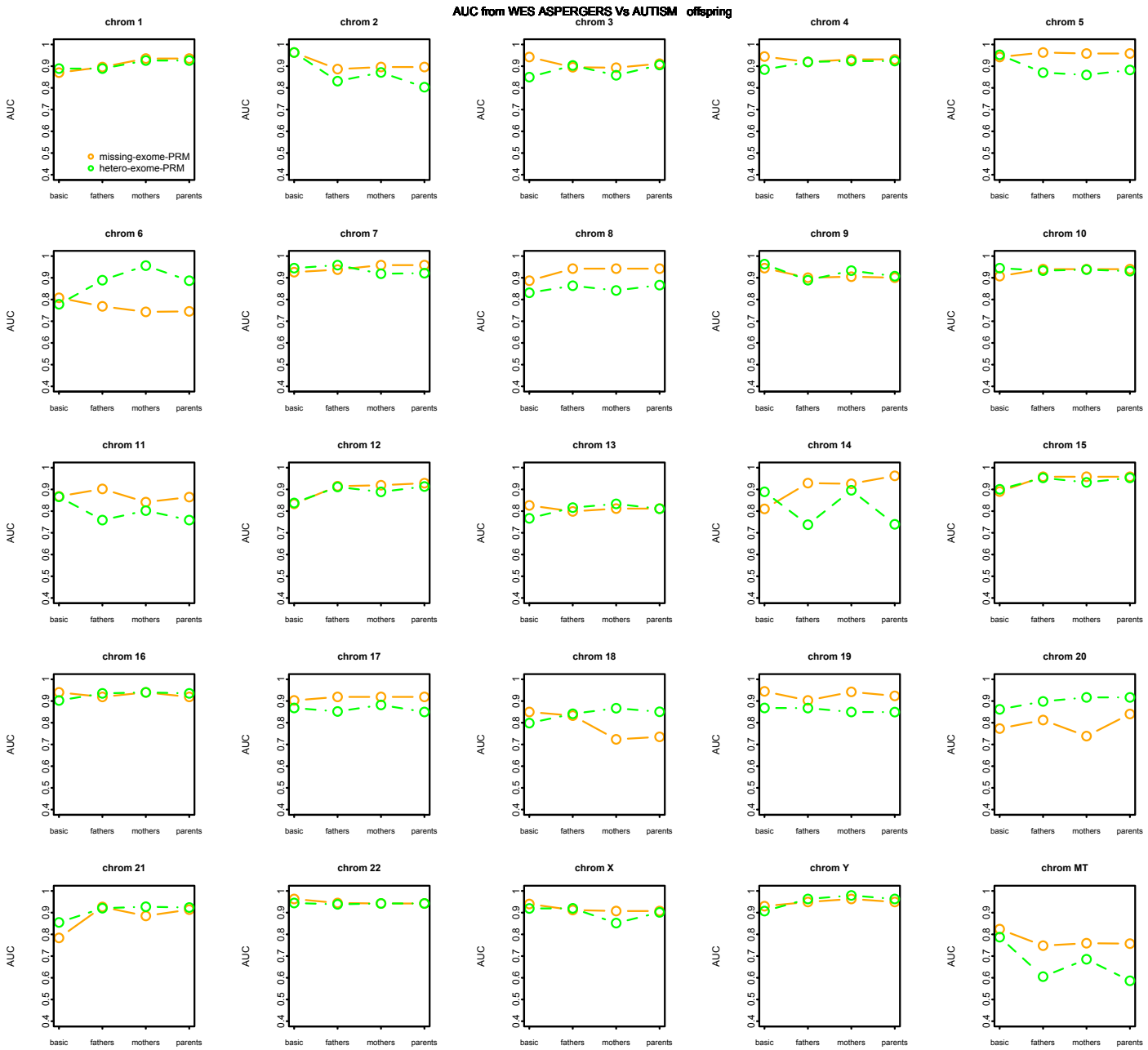

Figure 11: AUC results using a different J48 decision tree per chromosome to predict offspring Aspergers affectation risk. Within each plot we built decision trees for missing-exome-PRM (orange) and hetero-exome-PRM (green), using counts from offspring (X-label "basic"), or integrating it (by multiplying individual risks) with other PRM from one or the two parents ("fathers", "mothers", "parents").

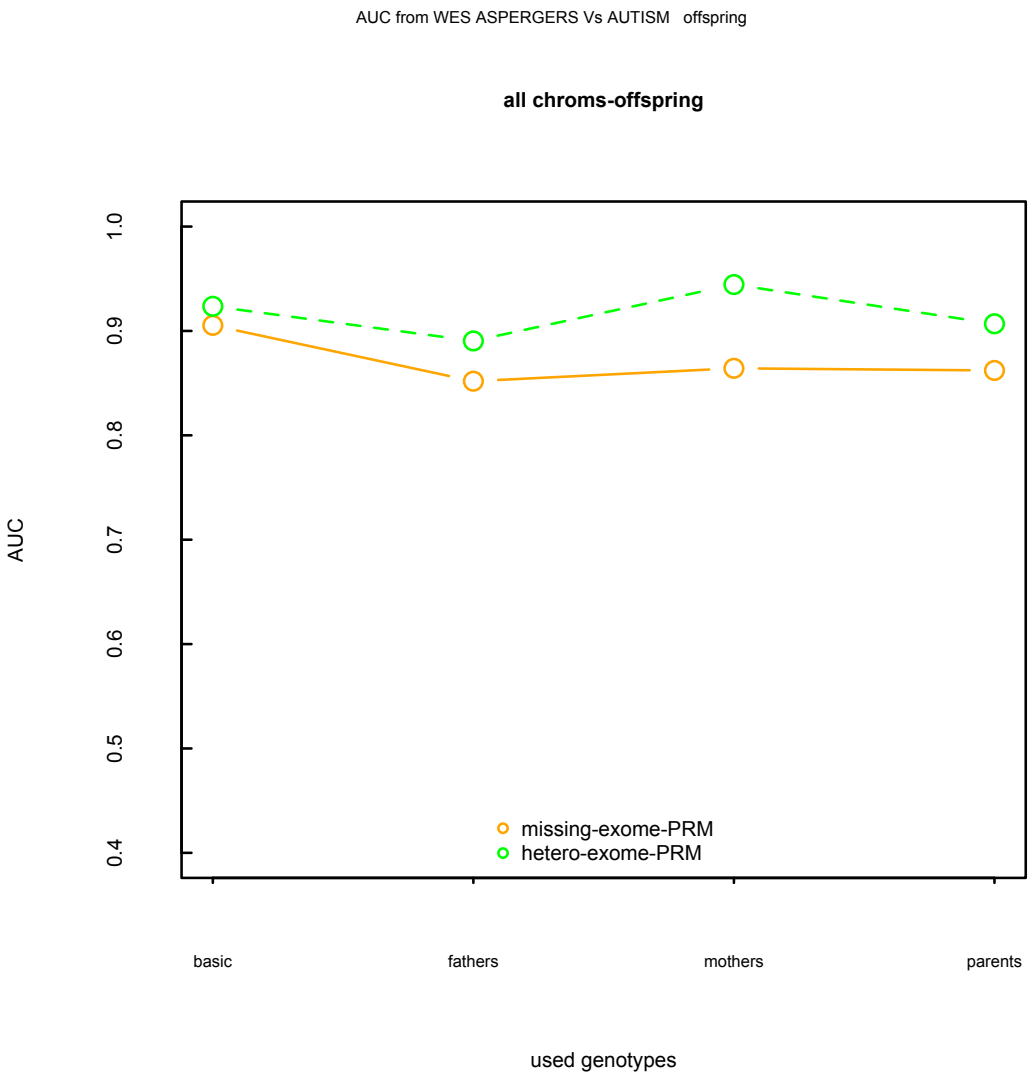

Figure 12: AUC results using only one J48 decision tree for all chromosomes to predict offspring Aspergers affectation risk. Within each plot we built decision trees for missing-exome-PRM (orange) and hetero-exome-PRM (green), using counts from offspring (X-label "basic"), or integrating it (by multiplying individual risks) with other PRM from one or the two parents ("fathers", "mothers", "parents").

Functional annotation clustering Given a PRM accurate enough to distinguish between cases and controls and confirmed in a second (WES) dataset, we performed a functional annotation clustering with those genes showing the highest levels of disease association. In a first step, we performed a χ^2 test using all autistic offspring as cases and all TD offspring as controls, and the number of missing and non missing offspring for a given locus within the two groups. We observed that there were $15,956$ out of the total $3,541,420$ genetic positions in the WES study, with a very low p value after Bonferroni correction. We then annotated these positions with the gene closer to each position, and differentiated those genes having within their ranges some of the selected positions from those just close to any of the significant positions. In the case of p value 0, the number of positions within a gene was 6545 and the number of genes was 631 (there were about 10 positions with missing disease-associated genotypes within the same gene). We also repeated the process by first differentiating positions according to whether the number of missing autistic offspring ($T2M$) and non-missing TD offspring ($T1N$) was larger than the number of non-missing autistic offspring ($T2N$) and missing TD offspring ($T1M$): $(T1N + T2M) > (T1M + T2N)$, and found only 84 genes in the first group and 547 genes in the second group, a large difference that confirms again the hypothesis of more missing genotypes in affected individuals. See Figures 13 and 14 for karyoplots with selected genes (red for those related to positions with more affected individuals missing and blue for those related to positions with more control individuals missing).

We performed a functional annotation clustering with genes that have some locus under their range with p val 0 (which means the whole variance is explained because of any of these genes), and with larger proportion of missingness in affected individuals (red color in the karyoplots),

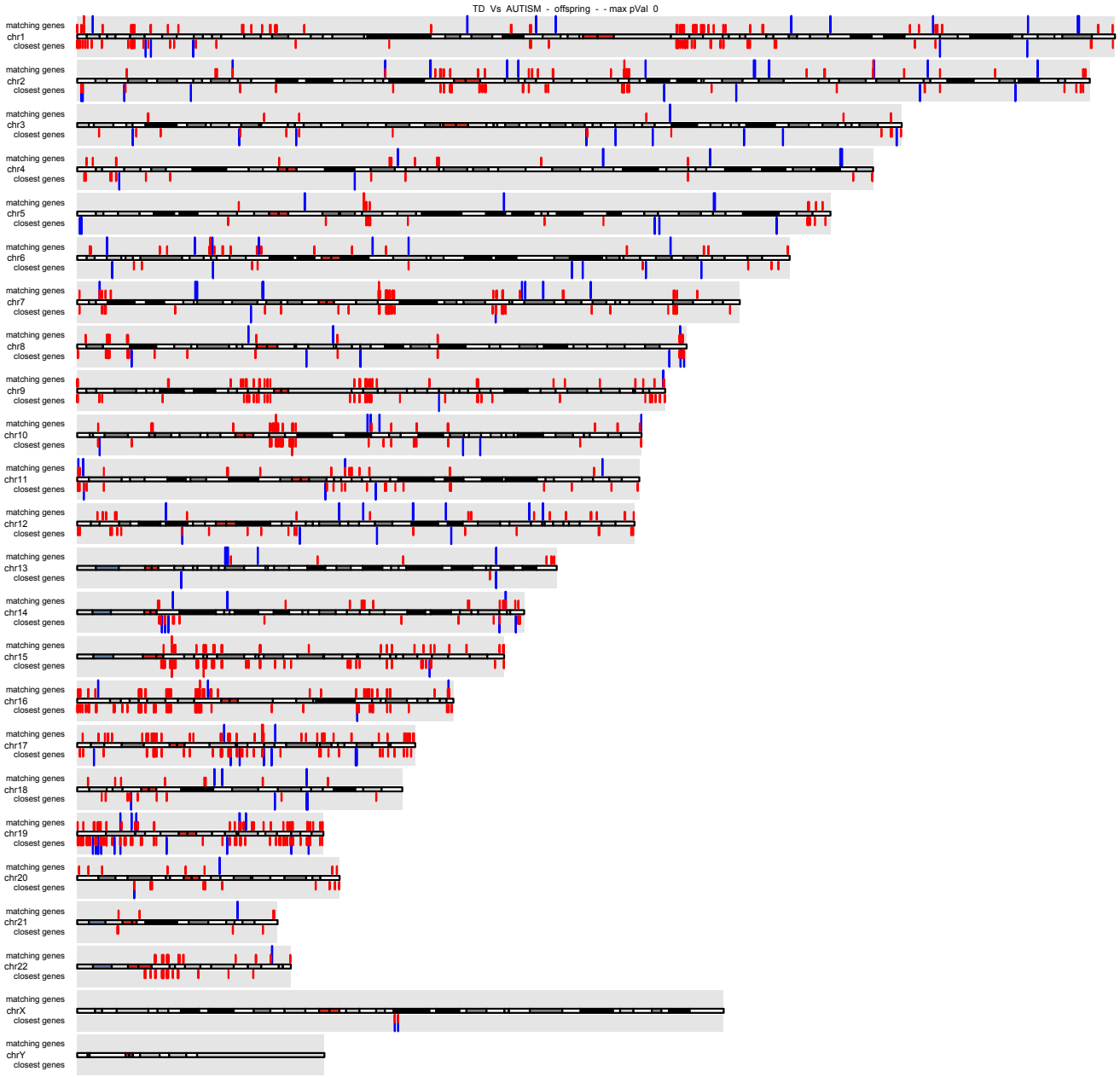

Figure 13: Karyoplot showing genes related to positions with p-val 0 (after Bonferroni) in missing-based χ^2 test for autism versus typical development offspring from WES genotypes. Genes with larger proportion of missingness in affected individuals are coloured in red. Genes with larger proportion of missingness in control individuals are coloured in blue. Genes with significant positions within its range are shown in the upper track (matching genes), while genes with significant positions close to them are shown in the lower track (closest genes).

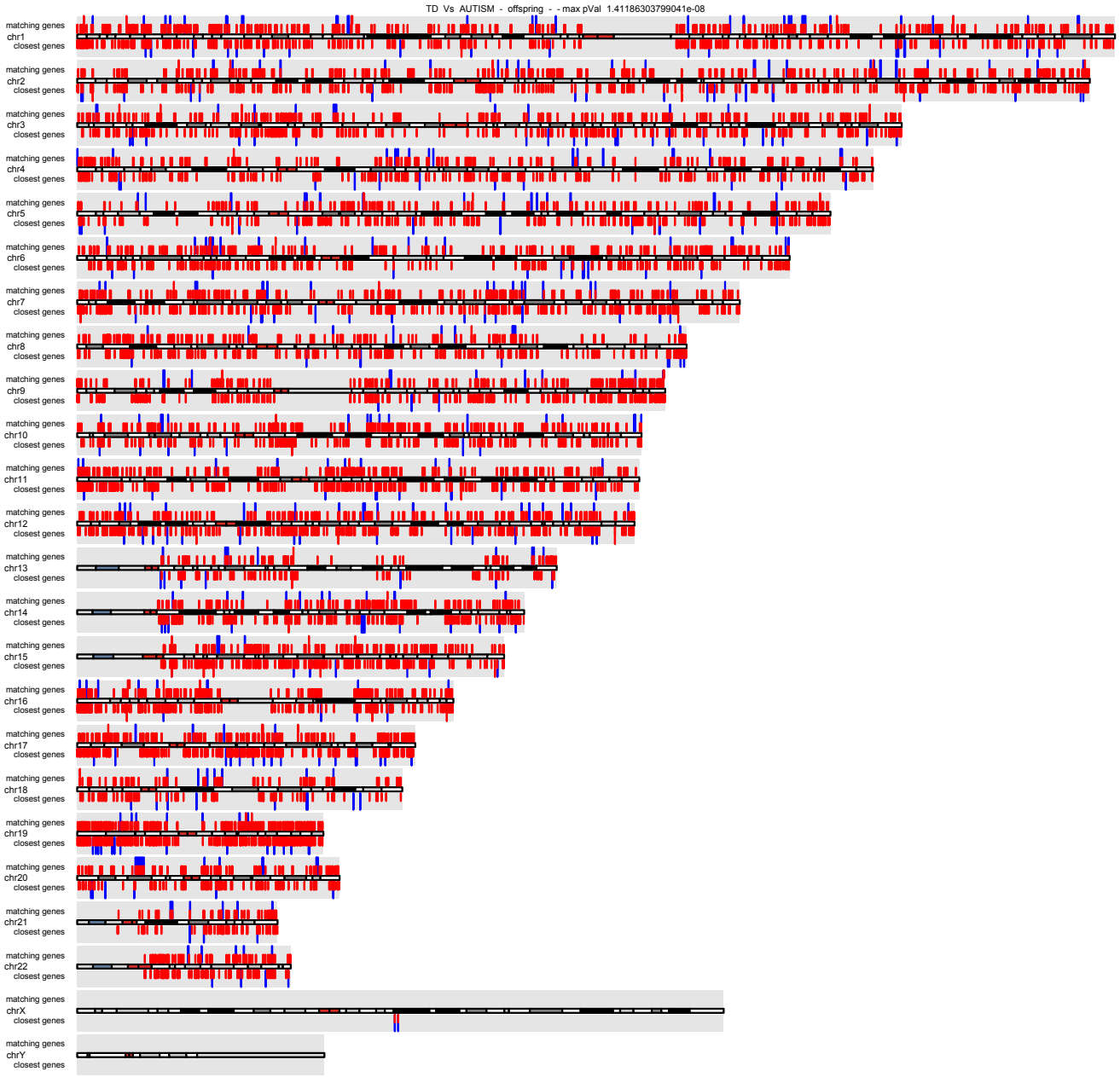

Figure 14: Karyoplot showing genes related to positions with p-val $1.4 - e08$ (after Bonferroni) in missing-based χ^2 test for autism versus typical development offspring from WES genotypes. Genes with larger proportion of missingness in affected individuals are coloured in red. Genes with larger proportion of missingness in control individuals are coloured in blue. Genes with significant positions within its range are shown in the upper track (matching genes), while genes with significant positions close to them are shown in the lower track (closest genes).

using the *Database for Annotation, Visualization, and Integrated Discovery (DAVID)* [8]. The goal of functional annotation clustering is to make biology clear to be read, by putting together similar annotations, i.e., annotations with similar gene members. 38 clusters with the highest classification stringency were found. The one with the larger enrichment score (5.6), an score related with homogeneous high p value annotations within elements in the cluster, has three elements (9.8e-4, 9.3e-4 and 9.3e-4 after Benjamini correction). Results seem to be highly significant. The common genes in that cluster are related to the nuclear pore complex (a complex that regulates genome access) interacting protein (NPIP), well conserved since African apes although not among those forming the core of the nuclear pore complex (NPC) [9].

We performed a similar analysis with the 84 genes with a larger proportion of controls than affected having missing variants. Only 5 clusters were found. The most relevant one had an enrichment score of 1.08 with three element inside (all being Kegg pathways). P values after Benjamini correction were equal to 1.

3 Discussion

We believe the results here presented open a significantly different new approach to find genetic causes of autism and other complex diseases in general. We have needed to develop new methods from the very first step: genotype calling, as, after many years of defeats [10], we were becoming aware that in the current approaches we were losing the main signal [5]. There are many questions that remain open at this point, but we need now to wait for others within the scientific community,

specially molecular biologists, to confirm our findings in order to support this approach. If the strong results related to the NPIP were confirmed, this may become the main cause of autism development, and perhaps the disease can be cured with gene therapy. There is also a issue that we need to handle, regarding short-term practical applications: the development of a risk predictor of autism, using either genome arrays or genome sequencing, may come soon to reality. It will be very important to understand, at least at light of these results, that Aspergers are genetically closer to TD than to autism, which means that Aspergers should not be considered as a disease, but just as a peculiar trait. The same must be true for other ASD sub-groups, such as Pervasive Developmental Disorder - Not Otherwise Specified (PDD-NOS). While performing similar tests with PDD-NOS versus TD and versus autism, is a pending task, it may be also possible to differentiate strict autism from PDD-NOS or other ASDs. The development of tests able to distinguish between strict autism and other ASDs will become ethically very important, specially if a couple wants to know the risk of having an autistic child by testing their respective genomes. And specially because people with these peculiar traits may have a highly satisfactory and independent life. Anyway, a child with strict autism, although will guarantee to their parents a very challenging life, may help them to become more humans and sensitive to other human beings.

4 Acknowledgment

The datasets used for the analysis described in this manuscript were obtained from the IHPM [3] and from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number, phs000267.v5.p2 (AGP consortium) and phs000298.v4.p3 (ASC). Submission of the data, phs000267.v5.p2, to db-

GaP was provided by Dr. Bernie Devlin on behalf of the Autism Genome Project (AGP). Collection and submission of the data to dbGaP were supported by a grant from the Medical Research Council (G0601030) and the Wellcome Trust (075491/Z/04), Anthony P. Monaco, P.I., University of Oxford.

1. Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. The genetic interpretation of area under the roc curve in genomic profiling. *PLoS Genetics* **6**, e1000864 (2010).

2. 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).

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

4. Fairley, S., Lowy-Gallego, E., Perry, E. & Flicek, P. The international genome sample resource (igsr) collection of open human genomic variation resources. *Nucleic Acids Research* **48**, D941–D947 (2019). URL `https://doi.org/10.1093/nar/gkz836`. `https://academic.oup.com/nar/article-pdf/48/D1/D941/31697918/gkz836.pdf`.

5. Abad-Grau, M. M. The unreachable genomic profiling of complex diseases: genotype missingness matters. Tech. Rep., Department of Computer Languages and Systems. University of Granada (2018). URL `http://bios.ugr.es/HaploRisk/HaploRiskAbadGrau.pdf`.

6. Igor, R., Kinzy, T. & Bailey, J. C. Genetic risk scores. *Curr Protoc Hum Genet.* **104**, 337–338 (2019).

7. Sebastiani, P., Solovieff, N., Puca, A. & Sun, J. X. Naïve bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: not so different after all! *Front. Genet.* **3** (2012).

8. Sherman, B. T. *et al.* David: a web server for functional enrichment analysis and functional annotation of gene lists. *Nucleic Acids Research* **50**, W216–W221 (2022). URL `https://doi.org/10.1093/nar/gkac194`. `https://academic.oup.com/nar/article-pdf/50/W1/W216/44376821/gkac194.pdf`.

9. Johnson, M., L Viggiano, L., Bailey, J. *et al.* Positive selection of a gene family during the emergence of humans and african apes. *Nature* 514–19 (2001).

10. Janssens, A. & van Duijn, C. Genome-based prediction of common diseases: advances and prospects. *Human Molecular Genetics* **17**, R166–R173 (2008).