Journal of Biomedical Informatics 77 (2018) 50–61

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Heritable genotype contrast mining reveals novel gene associations specific to autism subgroups

Matt Spencer[a], Nicole Takahashi[b], Sounak Chakraborty[c], Judith Miles[b,d], Chi-Ren Shyu[a,e,f,*]

[a] Informatics Institute, University of Missouri, 241 Naka Hall, Columbia, MO 65211, USA
[b] Thompson Center for Autism & Neurodevelopmental Disorders, University of Missouri, 205 Portland St, Columbia, MO 65211, USA
[c] Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, USA
[d] Department of Child Health, School of Medicine, MA204 Medical Sciences Building, University of Missouri, Columbia, MO 65212, USA
[e] Department of Electrical Engineering and Computer Science, University of Missouri, 201 Naka Hall, Columbia, MO 65211, USA
[f] School of Medicine, University of Missouri, MA204 Medical Sciences Building, Columbia, MO 65212, USA

ARTICLE INFO

Keywords:
Data mining
Autistic disorder
Genetics
Frequent pattern mining

ABSTRACT

Though the genetic etiology of autism is complex, our understanding can be improved by identifying genes and gene-gene interactions that contribute to the development of specific autism subtypes. Identifying such gene groupings will allow individuals to be diagnosed and treated according to their precise characteristics. To this end, we developed a method to associate gene combinations with groups with shared autism traits, targeting genetic elements that distinguish patient populations with opposing phenotypes. Our computational method prioritizes genetic variants for genome-wide association, then utilizes Frequent Pattern Mining to highlight potential interactions between variants. We introduce a novel genotype assessment metric, the Unique Inherited Combination support, which accounts for inheritance patterns observed in the nuclear family while estimating the impact of genetic variation on phenotype manifestation at the individual level. High-contrast variant combinations are tested for significant subgroup associations. We apply this method by contrasting autism subgroups defined by severe or mild manifestations of a phenotype. Significant associations connected 286 genes to the subgroups, including 193 novel autism candidates. 71 pairs of genes have joint associations with subgroups, presenting opportunities to investigate interacting functions. This study analyzed 12 autism subgroups, but our informatics method can explore other meaningful divisions of autism patients, and can further be applied to reveal precise genetic associations within other phenotypically heterogeneous disorders, such as Alzheimer's disease.

1. Introduction

Autism is defined by the presence of a core set of symptoms involving behavioral, social, and cognitive deficits [1]. These phenotypes are measured during diagnosis, often employing sub-scores specific to the individual phenotypes. Diagnostic sub-scores differ greatly between individuals with autism, demonstrating the disorder's extensive phenotypic variation. The diversity within individuals bearing the same diagnosis is concerning, and suggests that autism is too broad of a classification [2]. Overly broad grouping is problematic for association studies, as testing for associations over a diverse population severely reduces the power of the test. Thus, the autism research community has endeavored to study subgroups of children with shared attributes [2–9]. In this work, we further emphasize this focus by partitioning children into pairs of opposing subgroups and specifically searching for the major genetic differences between them.

The complex genetic etiology of autism rivals its phenotypic variability, supporting the contemporary consensus that autism is a collection of etiologically distinctive disorders that cause a consistently recognizable phenotype. *De novo* mutations [10–12], inherited variation [13–15], and environmental factors [16–18] have all been linked to autism development. Familiar genetic disorders (ex. Fragile X syndrome) appear in approximately 10% of autism cases [6]. This multiplicity suggests that the onset of autism, and even the onset of specific autism subtypes, is likely due to a combination of factors, possibly including both genetic and environmental elements [6,19,20]. Our knowledge about interactions between risk factors during the development of autism is insufficient, stressing the need for further examination of the interactions of multiple genotypes [21].

It is estimated that around half of the genetic contribution of autism

* Corresponding author at: Department of Electrical Engineering and Computer Science, University of Missouri, 201 Naka Hall, Columbia, MO 65211, USA.
E-mail addresses: mcsgx2@mail.missouri.edu (M. Spencer), takahashin@health.missouri.edu (N. Takahashi), chakrabortys@missouri.edu (S. Chakraborty), mileshj@health.missouri.edu (J. Miles), shyuc@missouri.edu (C.-R. Shyu).

https://doi.org/10.1016/j.jbi.2017.11.016
Received 1 September 2017; Received in revised form 15 November 2017; Accepted 28 November 2017
Available online 29 November 2017

development stems from common variants [22] Furthermore, gene-gene interactions are often responsible for the complexity of disease susceptibility, to the point where gene interactions are considered to be more important than the independent effects of the genes in some cases [23]. Associations approximating these genetic interactions are often measured at the nucleotide level, which can be accomplished by examining the joint association of Single Nucleotide Polymorphisms (SNPs) to a disease. However, SNP microarrays often comprise millions of genotypes, making it impossible to examine every combination of these SNPs. The important question becomes: how will we decide which combinations of SNPs (denoted as "SNP-sets") to test?

Most prior attempts to address this challenge have concentrated on the selection of SNPs based on prior knowledge, such as SNPs from genes closely related to an autism phenotype [24–30]. These studies tend to consider a limited number of loci (< 100 SNPs) and are hypothesis-specific, reducing the potential for novel discoveries. Existing methods, such as the popular genome association analysis toolkit PLINK [31], are capable of testing epistasis effects, but are limited to pairs of SNPs and test them exhaustively, which is inefficient and time-consuming. By searching all genes for significant interactions, we can discover novel autism candidate genes and interactions between novel and known autism candidate genes. This was previously impossible due to the sheer magnitude of SNPs within genes. Fortunately, recent advances in computational power using distributed computing techniques makes it possible for the autism research community to have a more comprehensive view of SNP-sets associated with the disorder.

To discover genotype combinations relevant to the differentiation of specific autism subgroups while avoiding the limitations of manually selecting genotypes to combine, we developed a novel procedure called Heritable Genotype Contrast Mining (HGCM). This multi-disciplinary method integrates data mining techniques with traditional bioinformatics strategies to address the combinatorial problem of testing combinations of SNPs while searching for associations. HGCM avoids the limitations of pre-selecting SNPs by utilizing genome-wide SNP prioritization, facilitating the discovery of novel associations with autism. The data are partitioned into autism subgroups early in the process, allowing our method to highlight combinations of SNPs that are abundant in specific subgroups using a Frequent Pattern Mining algorithm. Opposing subgroups are then contrasted to reveal differentiating SNP-sets and identify genetic associations specifically relevant to the subgroups. HGCM was developed with focuses on testing combinations of SNPs and comparing disease subgroups, and can be applied to investigate the etiology of any disease with complex heritable genetic contribution and subtype structures.

2. Material and methods

The data analyzed in this study were obtained from the Simons Foundation Autism Research Initiative (SFARI) – Simon's Simplex Collection (SSC) [32]. SSC contains copious data from 2591 simplex families, where simplex refers to families with exactly one child diagnosed with autism (the "proband"). This amounts to SNP microarray genotypes and phenotypic data describing 11,560 individuals including probands age 4–17 and their parents and unaffected siblings. SSC includes families from USA and Canada, but otherwise has no specific geographical restriction.

The SSC genotype dataset was too large to be analyzed by normal means, particularly since our aim was to examine the combinatorial search space of many SNPs. To address this, we utilized a research computing environment comprising 192 cores and 960 GB of memory spread over eight compute nodes. Data storage was managed by Apache Hadoop [33], a framework that supports data distribution and analysis over multiple machines. Resource-demanding computations were executed using Apache Spark [34], an in-memory distributed-computing framework.

2.1. Procedure overview

We describe the full HGCM procedure (Fig. 1) with details about each step in the following sections. As a preprocessing step, missing genotypes are imputed. Pairs of opposite subgroups are formed using existing autism subtype classifications or autism behavior scores from the widely used Social Responsiveness Scale [35]. SNPs are tested for primary association with each subgroup using a genome-wide prioritization procedure, and the most significant SNPs are selected. Our extended Frequent Pattern Mining implementation identifies combinations of these selected SNPs that are prevalent in the subgroups and evaluates these prevalent SNPs for their potential to contribute to autism development in the specific genetic context of the individual families. The prevalence of SNP combinations is contrasted within opposite subgroups, and high-contrast genotype combinations are tested for association with the corresponding subgroup.

2.2. Missing genotype imputation

SSC data include genotypes measured by three genotyping arrays [36]. The number of loci consistent with dbSNP build 147 [37] measured by each array is shown in Fig. 1. Genotypes were preprocessed using a step described by Verma, et al. [38] to standardize the genotype measurements from the different arrays, a process common to meta-analyses called missing genotype imputation [39]. The bioinformatics tool Beagle (version 4.1) [40] was used to infer genotypes that were missing due to the differences in array measurements, resulting in 2,950,235 SNP genotypes for all individuals.

2.3. Frequent pattern mining

Frequent Pattern Mining (FPM) is a data mining technique that excels at identifying combinations of features that occur repeatedly (i.e. frequent patterns) [41,42]. For this study, our goal was to utilize FPM to ascertain the prevalence of SNPs and SNP-sets within autism populations. FPM requires data to be translated into binary "items", with the two states indicating the presence or absence of the item in a person. To satisfy this constraint, bi-allelic SNPs, which have three states (homozygous for the major allele, heterozygous, or homozygous for the minor allele), must be condensed into a two-state representation. HGCM does this by combining genotypes containing the putative major allele into the "absent" state, while the "present" state contains only the homozygous minor allele genotype.

This binary construction accounts fully for the case where a genotype has a recessive effect. It will still account for dominant and additive effects of the minor allele, since both of these cases would lead to an enrichment of the homozygous genotype (and thus the present item state). Although these genotypes are handled less elegantly, their effects are likely to be detected. Furthermore, the consequences of this strategy are mitigated because the resulting SNP-sets are subsequently tested for association in a manner that makes no assumptions about the most likely genetic model utilized by any variant.

Once genotypes are converted into items, the population prevalence of each item is calculated; this is called the "support" of the item (Fig. 2A). This metric can be extended to groups of items, or SNP-sets: the support of a SNP-set is the proportion of people with all the items in the set. The support of a SNP-set will never exceed the support of its subsets. Thus, FPM automatically rejects supersets of SNP-sets that do not meet a specified support threshold, or "minimum support". This allows for drastic reductions in the number of SNP-sets that must be examined (Fig. 2B).

The memory requirement of storing many combinations of items is a challenge of FPM, as the number of combinations grows exponentially when analyzing tens of thousands of items. We impose a heavy burden on the FPM algorithm by including SNPs that are representative of the entire genome, to the point where the analysis exceeds the capabilities

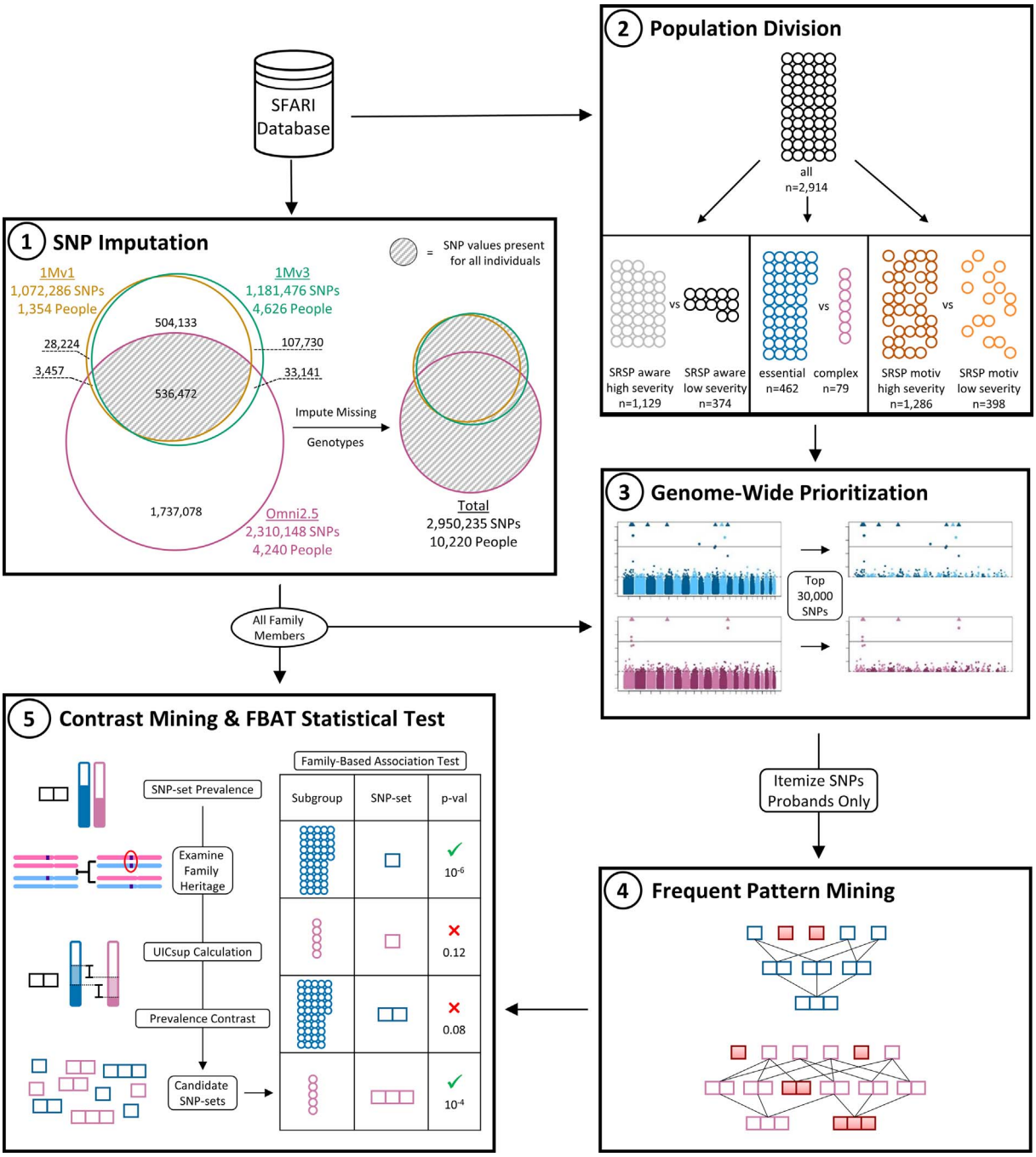

Fig. 1. Overview of the Heritable Genotype Contrast Mining procedure. The major operations are indicated in boxed groups, numbered in order of occurrence in the workflow. Abbreviations: SFARI = Simon's Foundation Autism Research Initiative, SNP = Single Nucleotide Polymorphism, SRSP = Social Responsiveness Scale – Parent Report, FBAT = Family-Based Association Test.

of typical computing systems. Furthermore, we aimed to develop a system that incorporates as many of the available data as possible while searching for associations, leading to our creation of the Unique Inherited Configuration support metric described below. Thus, for the HGCM procedure we developed a customized Spark-enabled implementation of FPM, integrating our novel metric into the algorithm and utilizing a distributed in-memory computing environment capable of analyzing the large dataset.

2.4. Population division

FPM tools avoid examining all possible combinations of items by utilizing minimum support thresholds (Fig. 2B). Although this is computationally beneficial, it is limiting from a knowledge discovery standpoint. When a small homogeneous subgroup is diluted in a large diverse population, patterns specific to the subgroup are often eliminated by FPM. However, examining the subgroup in isolation allows these patterns to emerge (see Fig. 2C). Thus, HGCM features a procedure known as Contrast Mining [43–45] which divides the autism cohort into subgroups, performs FPM on each subgroup individually, and compares opposite subgroups to identify SNP-sets which appear frequently in one subgroup but rarely in the other.

SSC families were divided into subgroup pairs based on characteristics of the proband (Table 1). One subgroup pair was formed using a

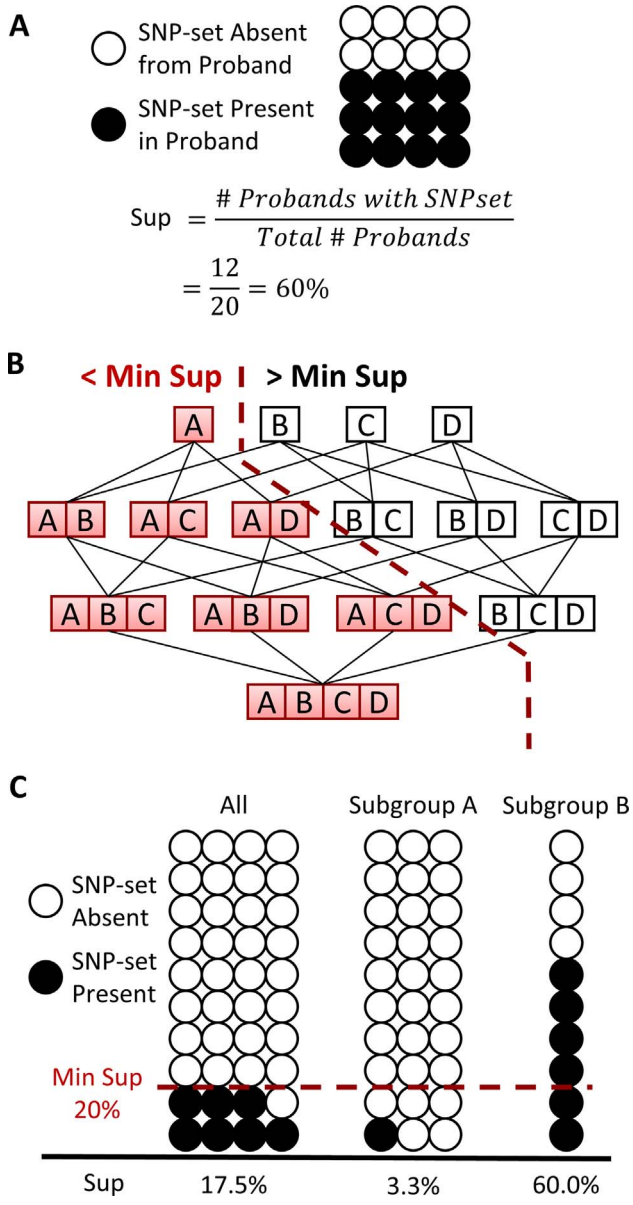

Fig. 2. Overview of Frequent Pattern Mining and Contrast Mining. (A) Example calculation of the support of a SNP-set. (B) When a SNP-set is infrequent (support is less than the specified minimum support, indicated by red shading), so are all its supersets. One single-SNP set "A" being infrequent guarantees that all SNP-sets including "A" are also infrequent, significantly reducing the combinatorial search space. (C) A SNP-set that is eliminated from a FPM analysis of the whole population may pass the Min Sup threshold when examining a specific subgroup. Abbreviations: Sup = support, Min Sup = minimum support.

previously defined subtype classification characterized by morphological categories. The Autism Dysmorphology Measure classifies probands as "dysmorphic" or "nondysmorphic" (equivalently, complex or essential) according to a decision tree based on the presence of explicit physical abnormalities [7]. Five more subgroup pairs were formed representing the range of severities for the sub-scores of the Social Responsiveness Scale (SRS): Parent Report [35] (chosen due to high response rate). These sub-scores measure five categories: the ability to notice social cues (awareness), the ability to interpret social cues (cognition), the ability to communicate expressively (communication), the motivation to engage in social behavior (motivation), and the expression of stereotypical behaviors or restricted interests (mannerisms). Appendix A details subgroup inclusion criteria using SSC terminology.

2.5. Genome-wide SNP prioritization

We return to the question: how do we decide which SNP-sets to test? FPM algorithms answer this by utilizing the minimum support threshold: SNP-sets will be filtered according to their prevalence in the affected population. However, this approach fails to account for linkage disequilibrium, a known phenomenon causing associations between SNPs. Frequently co-occurring SNPs will often be those that have a physical association, rather than an association with autism, and most of these SNP-sets will have no association with autism when comparing cases and controls. However, the FPM algorithm does not account for controls - the analysis examines a group in isolation (note that cases and controls are compared in other stages of HGCM).

To overcome this, HGCM identifies SNPs with some evidence of a primary effect for the disorder using Bioconductor's GWASTools [46] package. Minor allele frequencies within probands and unaffected family member controls are used to perform a logistic regression analyses, determining the association of each SNP with the affected population. Generally, it is important to choose a p-value cutoff that corrects for the testing of millions of loci [47]. However, our purpose for finding the strength of single-locus associations in HGCM is not to make statistical claims, but rather to select SNPs to combine into SNP-sets. Thus, HGCM selects the 30,000 most significant SNPs, as this roughly corresponds to a p-value cutoff of 0.05 for most subgroups (as opposed to the stringent Bonferroni threshold of $1.67e-8$ typically used for GWAS).

2.6. Contrast Mining utilizing the UICsup

The direct application of FPM in this context would consider only the data gathered for the autism probands, but the data additionally include valuable information within the genotypes of immediate family members. Many of the autistic probands with a prevalent genotype have unaffected family members with the same genotype, providing evidence against the genotype's contribution to autism development in the specific proband. Such a comparison between the genotypes of probands and their unaffected family members allows genotypes to be considered for their potential contribution to autism development in the context of the unmeasured genetic landscape of the family and in relatively similar environmental conditions.

To generate clinically relevant association candidates, it is necessary to consider the available information provided by the genotypes of close family members. Thus, we extended the Frequent Pattern Mining algorithm to calculate not only the support of the SNP-sets, but also an adjusted version of the support that accounts for inheritance patterns. We call this novel metric the "Unique Inherited Configuration support" (UICsup) because it calculates the proportion of probands that are the only member of their nuclear family bearing all the items in the SNP-set, thus the proband has a *unique configuration* of the *inherited* variants. Our incorporation of this new metric into the data mining procedure allows us to generate SNP-set candidates with higher potential for strong disease association, since UICsup is stricter than the unmodified support and accounts for more of the available data.

Fig. 3A depicts some patterns of SNP inheritance where all the displayed probands are included in the support for the depicted SNP-set, but not all of them would contribute to UICsup. HGCM integrates this novel metric accounting for heritable genotypes into the traditional FPM procedure by calculating the UICsup of each SNP-set (Fig. 3B) along with the support of the SNP-set. UICsup and the unaltered support are both considered during the Contrast Mining calculations to identify the SNP-sets with high-contrast between opposing subgroups. SNP-sets with a major increase of prevalence in either the support or the UICsup from one subgroup to another are highlighted as high-contrast candidates that will be tested for significant association.

In our application, candidates are generated when the prevalence between subgroups exceeds 50%, to ensure that contrasting genes have

Table 1
Names and descriptions for the examined subgroups, displayed as opposing subgroup pairs.

Subgroup 1				Subgroup 2		
Name	Description	Size		Name	Description	Size
Dysmorphic	Significant dysmorphology	79	vs	Nondysmorphic	No significant dysmorphology	462
High-severity awareness	High SRS-Parent Report Social Awareness sub-score	1129	vs	Low-severity awareness	Low SRS-Parent Report Social Awareness sub-score	374
High-severity cognition	High SRS-Parent Report Social Cognition sub-score	1860	vs	Low-severity cognition	Low SRS-Parent Report Awareness sub-score	171
High-severity communication	High SRS-Parent Report Social Communication sub-score	1786	vs	Low-severity communication	Low SRS-Parent Report Social Communication sub-score	201
High-severity mannerism	High SRS-Parent Report Autistic Mannerisms sub-score	1912	vs	Low-severity mannerism	Low SRS-Parent Report Autistic Mannerisms sub-score	202
High-severity motivation	High SRS-Parent Report Social Motivation sub-score	1286	vs	Low-severity motivation	Low SRS-Parent Report Social Motivation sub-score	398

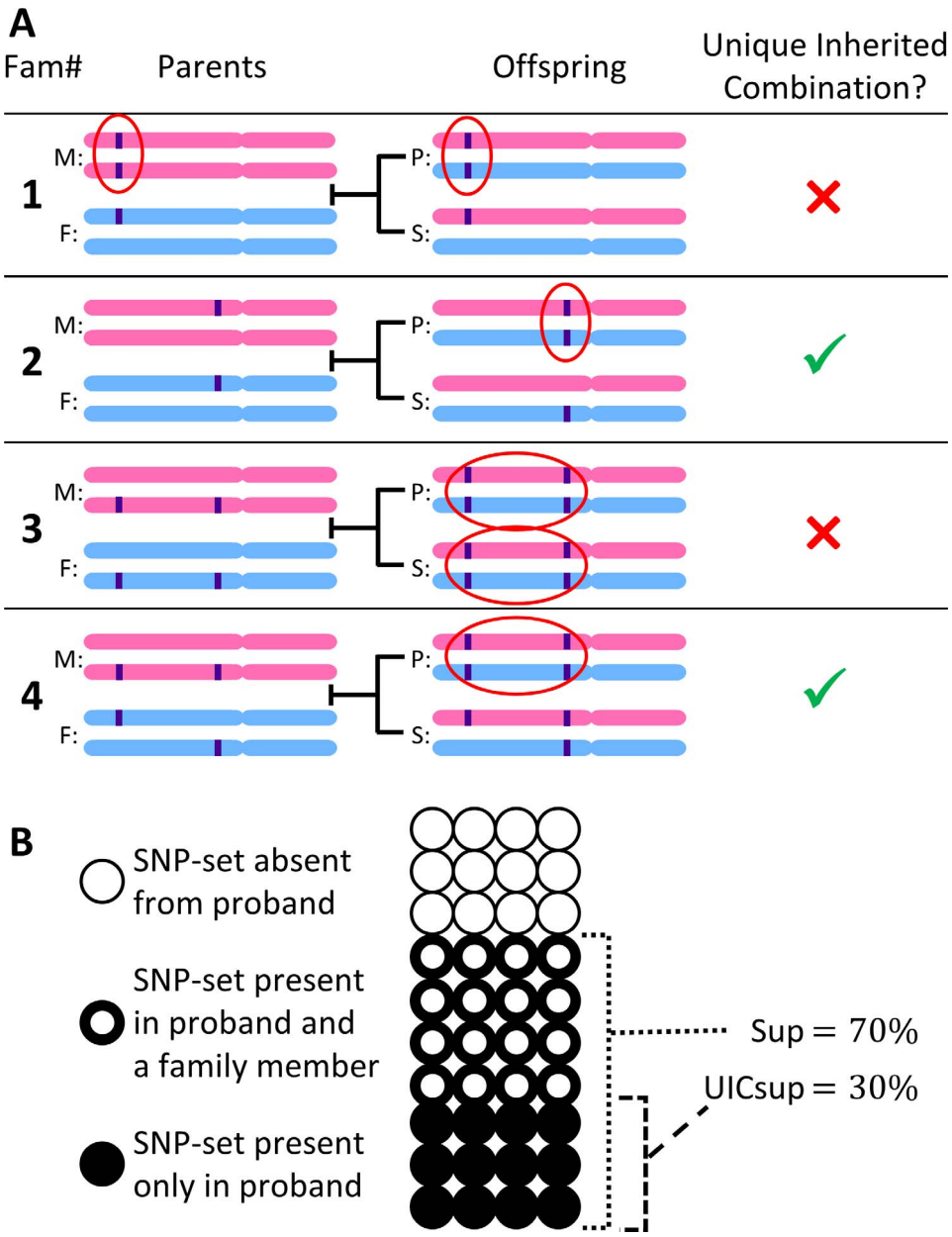

Fig. 3. Demonstration of the Unique Inherited Configuration support (UICsup) metric. (A) Various examples of SNP inheritance patterns from mothers ("M") and fathers ("F") to probands ("P") and their unaffected siblings ("S"). Bars across chromatids represent a SNP minor allele. Circles highlight when SNPs are homozygous for all the minor alleles of the SNP-set in question, equivalently indicating when a person would be assigned the item for the SNP-set. Note that all these probands have the depicted items. The right column indicates whether the proband has a unique configuration of inherited variants, and therefore is included in the UICsup. Families 1 and 2 show single-SNP inheritance patterns; families 3 and 4 show how the metric is applied for 2-SNP inheritance patterns. Family 1 and 3 probands do not have a unique configuration of the inherited variants because a non-proband family member has the item. Family 2 and 4 probands do have unique configurations of the inherited variants because they are the only family members with all the items in the SNP-set. (B) Example calculation of the support and UICsup of an SNP-set.

sufficient discriminatory power between subgroups. The specific value chosen for this threshold presents a choice between increasing the burden on the statistical testing and increasing the precision of the generated candidates. Reducing the threshold will generate more candidates, but a higher proportion of them will be deemed insignificant by subsequent testing. Furthermore, we consider it important that any discovered results apply to a meaningful proportion of the examined groups. In our experience with the SSC dataset, a prevalence difference of 50% generated enough candidates to yield interesting results while preventing the need for unjustified burden on the statistical testing. This value can be adjusted in future applications to accommodate the situation.

2.7. FBAT statistical testing

SNP-set candidates are tested for statistical association using the Family-Based Association Test (FBAT) [48]. We adjust for the testing of multiple hypotheses using the Benjamini-Hochberg correction [49]. With a simplex family sample, individuals within each family have much stronger genetic relationships than the average relationship between sampled individuals. In this case the standard association tests used for unrelated individuals becomes biased. This problem of bias due to mixed relatedness is avoided by FBAT by using within-family comparisons compatible with several different family structures. Only the SNP-sets containing SNPs on different chromosomes are tested, as this statistical procedure does not adequately account for physical linkage. Genes are considered to be associated with a subgroup when at least one SNP within the gene is significantly associated with the subgroup.

2.8. Examination of discoveries

We examined the discovered genes by seeing how many genes are selected during the HGCM procedure steps, and by comparing with previous autism literature. The two major selection criteria are (1) high-contrast SNPs with major differences in subgroup prevalence and (2) SNPs significant using the FBAT. Genes containing at least one selected SNP were considered to pass the HGCM selection process. We compared with previous literature to determine which genes are already believed to be relevant to autism. The genes passing the selection processes were cross-referenced with AutDB 3.0 [50], a reference for genes associated with autism also known as SFARI Gene, which contains candidate genes with varying levels of confidence from studies with varying specific research settings. We also implemented a broad search to identify HGCM genes that could be found in NCBI PubMed abstracts related to autism research using the search criteria "(autism OR asd) AND (gene AND < *gene name* >)", where the name of each gene was inserted.

3. Results

3.1. Associations with autism subgroups

Significant genetic associations were found within 10 autism subgroups, including a maximum of 172 genes associated with the dysmorphic subgroup (Fig. 4, blue[1] bars). In total, 286 distinct genes are associated to at least one autism subgroup (adj. $p < .05$). 193 of these are potentially novel candidate autism genes (red bars), as they were not present in AutDB or found in the PubMed abstract search. For each subgroup contrast, multiple novel genes distinguished each subgroup from its counterpart.

This results section primarily focuses on the 286 associated protein-coding genes, as these have direct functional implications. Additionally, 84 non-coding RNAs (predominantly lincRNAs) have significant associations, and 56.8% of the significant SNPs (689 SNPs) are not within any known genes.

3.2. Comparison with previous literature

We calculated the number of genes within which SNPs passed the major selection processes of HGCM, these are listed in Table 2. The 894 genes with high subgroup contrast were selected from a pool of 11,339 genes that were prevalent in at least one subgroup. Thus, contrast mining eliminated 92% of the pool of genes; of these selected genes, 32% survived the FBAT. This indicates that the contrast mining procedure accomplished its intended purpose of identifying strong candidate genes and gene combinations from the combinatorial search space of many SNPs.

We searched AutDB and PubMed for previous documentation of associations with the significant genes. The majority of the AutDB genes were not selected due to similar prevalence in opposing subgroups, presumably because they are relevant to autism in general and less specific to autism subtypes. Recall that AutDB is a broad collection of autism-related work with varying degrees of certainty from unsupported to high-confidence, and tracks genes under investigation for their potential relevance to autism, so it is not expected that any one method would reproduce a majority of these genes. 49 of our associated genes are present in AutDB – these are genes previously found to be related to autism in general that we found to be specifically relevant to autism subgroups. We note that the results of our method included none of the AutDB genes labelled as "unsupported" and only one labelled as "high-confidence". Most of these two categories are significant (or not) regardless of the subgroup being examined, so they do not reveal genetic distinctions between subgroups. Most of our significant findings that overlap with AutDB likely have intermediate levels of support due to previous attempts to associate these genes with a broad population of autism, rather than the more homogeneous subgroups examined in this study.

The broader PubMed abstract search found that 44 additional genes, identified by HGCM but not included in AutDB, have been reported in publications related to autism (see Table 2 and Fig. 4, green bars). Thus, the remaining 193 HGCM genes are considered novel for autism studies (Fig. 4, red bars). We note that the proportion of previously known *significant* gene candidates is larger than the proportion of previously known *high-contrast* gene candidates. This suggests that the FBAT further improved the quality of selected genes, highlighting associations that are both statistically significant and clinically relevant for autism subgroup distinctions.

It is important to note that the PubMed search is limited by the fact that not all autism studies report significant findings in the abstract, particularly when many associations are found. A notable example is Yuen et al.'s recent study [51], which identified 18 new candidate genes. A comparison with the full text revealed that our HGCM method also found one of these genes, *PCDH11X*, to be associated with the dysmorphic, high-severity cognition, high-severity communication, and high-severity motivation subgroups.

3.3. Method validation

We examine specific procedure elements for their contribution to the HGCM method. Assessing the contribution of the genotype imputation, we notice that although only 18.2% of the SNPs were directly measured by all three microarrays (Fig. 1). 54.3% of the significant SNPs identified by the HGCM method were from this non-imputed fraction. However, 51.0% of HGCM-identified genes was supported by at least one significant imputed SNP, and 36.7% of the genes were discovered strictly due to the inclusion of imputed genotypes, showing an important contribution from this procedural step.

To similarly evaluate the contribution of the UICsup metric, we separated the significant genes that would not have been identified if

[1] For interpretation of color in Fig. 4, the reader is referred to the web version of this article.

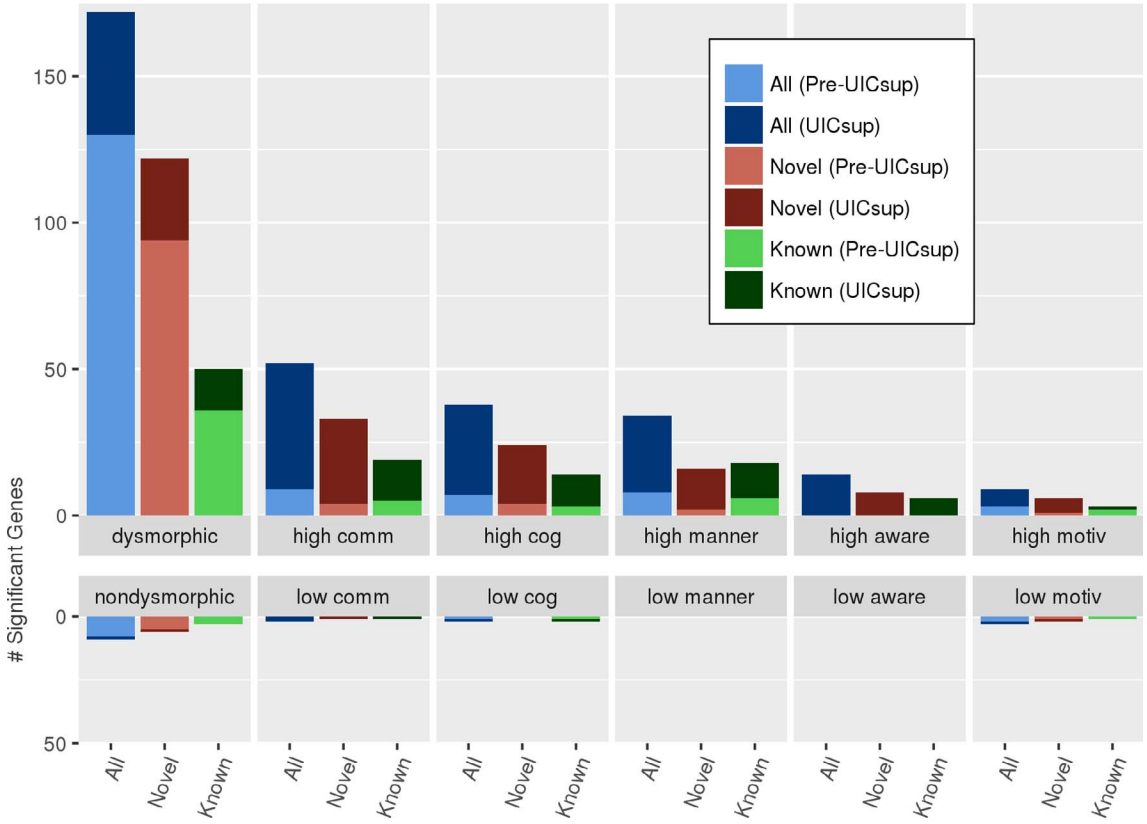

Fig. 4. The number of genes HGCM found to be significantly associated with each subgroup. Subgroup pairs are aligned vertically. Gene counts are partitioned according to the presence of the gene in the AutDB gene database and PubMed abstract search (known) and absence from prior literature (novel). Bars labelled "Pre-UICsup" indicate the number of genes that would have been found without the implementation of the UICsup metric, whereas "UICsup" bars indicate the additional genes that were found due to the incorporation of UICsup. Abbreviations: high = high-severity, low = low-severity, comm = communication, cog = cognition, manner = mannerisms, aware = awareness, motiv = motivation.

Table 2
Categorized AutDB genes remaining after HGCM selection operations.

HGCM major selection criteria	Genes	Genes present in AutDB	Genes found in PubMed search	Novel autism candidate genes
High subgroup contrast (difference in prevalence > 150%)	894	118 (13.2%)	214 (23.9%)	646 (72.3%)
Significant via FBAT (adj. p-value < .05)	286	49 (17.1%)	82 (28.7%)	193 (67.5%)

not for UICsup, shown in Fig. 4 (dark bar segments). UICsup was responsible for identifying 49.7% of HGCM genes, including 52.7% of HGCM genes classified as "known". Crucially, UICsup was responsible for over 80% of the genes associated with the SRS diagnostic sub-score subgroups.

3.4. Comparison with existing method

We demonstrate the contribution of our method by comparing to the capabilities of PLINK [31]. HGCM has several advantages over PLINK from a theoretical standpoint. PLINK supports testing for epistasis by testing all pairs of included SNPs in a case-control comparison. As this is an exhaustive examination of the combinations (and since the tool is not parallelized), this process takes longer than our FPM-based method, but many comparisons can still be completed in a reasonable timeframe. However, this limits potential epistatic effects tested to pairwise interactions, whereas our method has no methodological restriction on the number of SNPs that can form combinations. Additionally, PLINK is primarily designed for population-based samples and supports limited integration of family information. Basic family-based association testing for disease traits can be performed, but this option is not compatible with epistasis testing. Finally, PLINK lacks functionality which would contrast opposing subgroups as we do in this work.

To the best of our ability given the differences in capabilities between the applications, we performed an analogous analysis using PLINK to identify SNP combinations associated with the dysmorphic subgroup. Starting with the 30,000 most significant SNPs identified by our genome-wide SNP prioritization step, PLINK's epistasis test was used to identify SNP pairs associated with the dysmorphic subgroup in a case-control comparison (note that the analysis does not utilize pedigree information). The significant SNP pairs were measured in dysmorphic and nondysmorphic subgroups, and the prevalence of these SNP pairs were contrasted using the same criterion employed by HGCM: at least 50% prevalence difference. PLINK identified eight gene pairs specific to the dysmorphic subgroup, compared with the 69 gene pairs highlighted by HGCM. Three of the genes in PLINK associations overlap with genes found by HGCM. This, in addition to the unique results found by the inclusion of UICsup, demonstrates the importance of considering pedigree data in genome association analyses, when it is available.

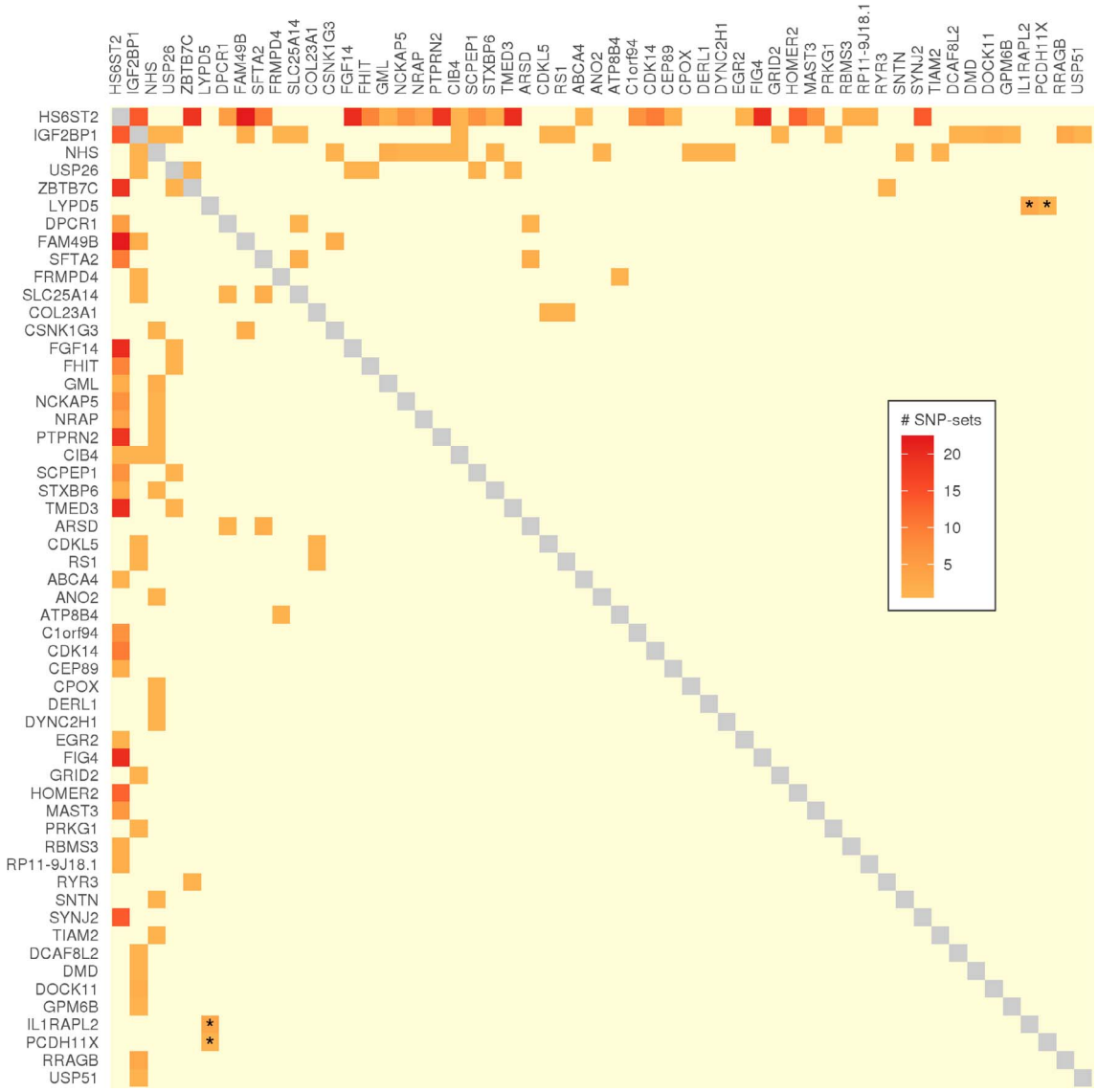

Fig. 5. The number of significant SNP-sets supporting each gene pair found to be associated with a subgroup by HGCM. Only pairs of protein-coding genes are shown. The asterisks (*) indicate gene pairs associated with the high-severity motivation subgroup – all other gene pairs are associated with the dysmorphic subgroup. The upper and lower triangles of the matrix are equivalent.

3.5. Gene pair associations

The HGCM method is designed to produce candidate SNP-sets representing interactions between any number of genes. However, the FBAT statistical test found all larger SNP-sets to be insignificant, resulting in single SNPs and pairs of SNPs associated with subgroups. HGCM gene pairs were generated when at least one SNP in each gene composed a significant SNP-set. In total, 71 gene pairs were associated with a subgroup; 55 distinct genes were part of at least one such gene pair. Most of these gene pairs were supported by one or two SNP-sets, but 11 gene pairs were supported by at least 10 SNP-sets connecting the two genes (Fig. 5). In fact, these highly supported gene pairs all included the *HS6ST2* gene and were associated with the dysmorphic subgroup. Variation in *HS6ST2* has been observed in previous autism studies, but it is not known to be a significant contributing gene [52–54]. Two gene pairs were associated with the high-severity motivation subgroup, and the rest with the dysmorphic subgroup; this is unsurprising as this subgroup also has the largest number of individually significant genes.

3.6. Case study for deeper understanding

Last, we examined the significant SNPs within the *DMD* gene. This gene was notable due to its association with several subgroups including the dysmorphic, nondysmorphic, high-severity communication, and low-severity communication subgroups – opposing sides of two subgroup contrasts. This presents a situation that is unique to a method like this which focuses on differences between paired subgroups. To understand this phenomenon, we visualized the positions of the associated SNPs in the context of the *DMD* exons (Fig. 6). The SNPs localized to two regions: 5 SNPs in a 100 kb region towards the start of the gene (ChrX:32950000-33050000), and 9 SNPs in a 500 kb region towards the end (ChrX:31600000-32100000). The localization of SNPs associated with more severe phenotypes suggests the existence of two regions within the gene which, when damaged, lead to severe phenotypes. SNPs associated with less severe phenotypes are near these regions, but somewhat isolated, indicating that variation further away from these critical regions affects the phenotype less seriously.

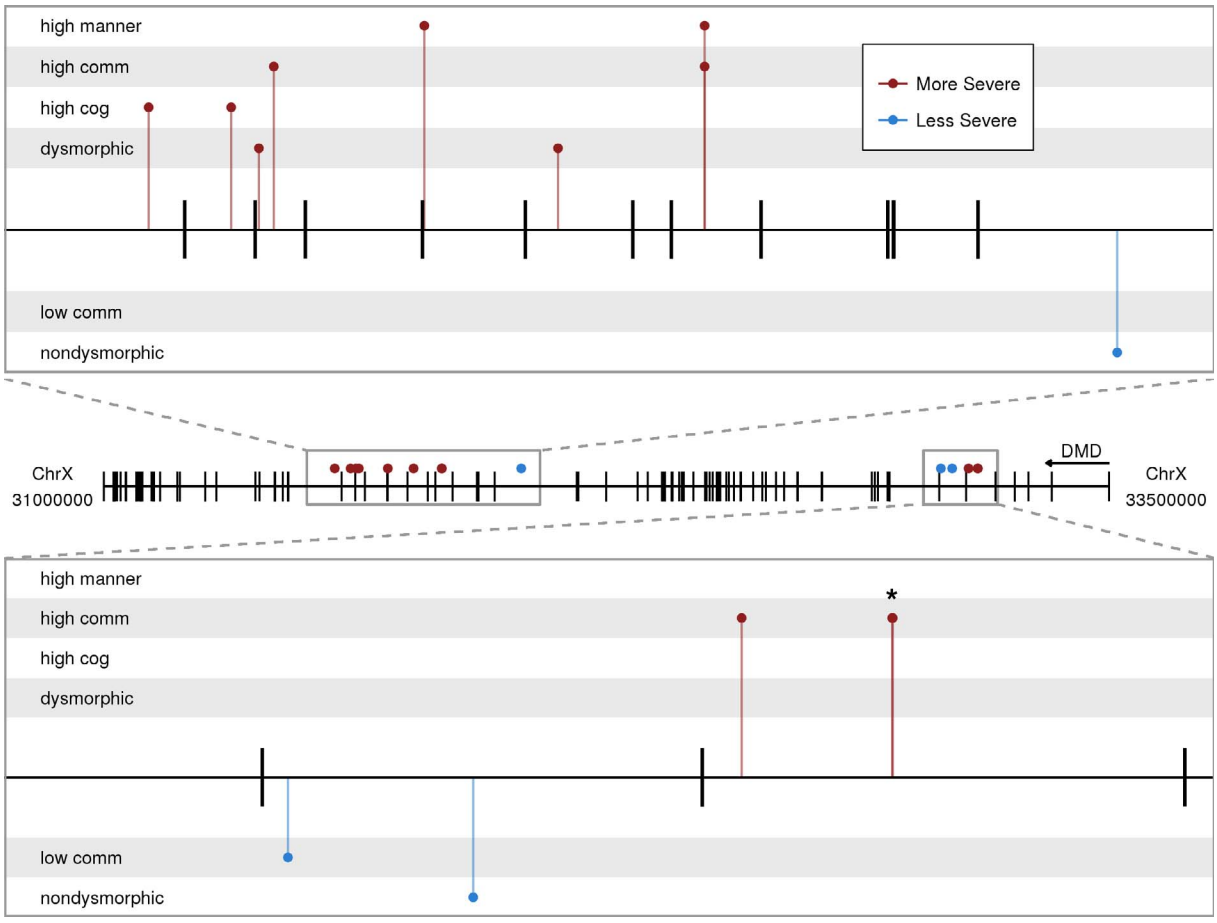

Fig. 6. The location of SNPs within the DMD gene (ENSG00000198947) associated with autism subgroups divided into more and less severe categories. Note that the dysmorphic/nondysmorphic subgroup pair is based on the presence or absence of specific physical features rather than high/low severity of a phenotype; however, patients classified as dysmorphic tend to have higher severity in behavioral measures than nondysmorphic probands. Black bars represent DMD exons, obtained using Ensembl BioMart (GRCh38.p7). Note that DMD is transcribed from the reverse strand, from right to left in this figure. The top panel magnifies the chromosomal region ChrX:31550000-32070000 and the bottom panel magnifies ChrX:32925000-33090000. The asterisk (*) indicates two adjacent SNPs (79 bp apart), both associated with the high-severity communication subgroup. Abbreviations: high = high-severity, low = low-severity, manner = mannerisms, comm = communication, cog = cognition.

3.7. Availability of discovered genes

Our informatics approach provides many candidate autism genes for future investigation, and highlights genes that are potentially relevant to specific subgroups and autism phenotypes. As a guide to improving the understanding of these subgroups, we provide a summary of the most significant findings for each examined subgroup in Table 3. A comprehensive dataset detailing our findings, including the discovered genes associated with subgroups, the involved SNPs, and a summary of the prevalence of these SNPs in the examined subgroup pairs, is provided in Appendices B and C. It also contains details about significant findings regarding non-protein-coding genes and SNPs in non-genic DNA to facilitate the growth of knowledge of these genetic regions.

4. Discussion

Our Heritable Genotype Contrast Mining procedure is a data-driven, *de novo* method for detecting gene-subgroup associations which reveal many novel autism candidates and correlate known autism factors to specific phenotypes. HGCM exclusively selects high-contrast genotypes, emphasizing the discovery of genes that may make a critical distinction in the development of precise autism subtypes. We utilized this method to reveal many genetic associations with autism patients grouped by explicit shared characteristics.

Hundreds of genes are estimated to be involved in autism development, and our informatics study reveals many candidates that can be studied with specific focuses on the associated subgroups and phenotypes. In the results of this study (Appendices B and C), we provide abundant information to prioritize these candidates in different ways, such as genes with the highest subgroup contrast or the most significant p-values. Our future prioritization strategy will involve identifying important pathways containing these associated genes, and to investigate groups of candidates related to specific functions for deeper understanding, as several key insights have been revealed from the linking of genes to functionality in autism [55,56].

Many SNP-sets prevalent in each low-severity subgroup are also prevalent in the corresponding high-severity subgroup, and were not selected for association testing due to our focus on high-contrast subgroup differences. This leads to the consistent enrichment of significant SNP-sets in the high-severity subgroups over their low-severity counterparts, seen in Fig. 4. This enrichment pattern is explained by the existence of baseline SNP-sets that induce the low-severity phenotype and additional SNPs that combine with these SNP-sets to increase the phenotype severity. This supports our hypothesis that combinations of SNPs are a driving force for autism subtype etiology. It explains the pattern seen in the five subgroup pairs defined by SRS phenotype severity scores, but recall that the dysmorphic and nondysmorphic subgroups are defined by explicit physical traits, and not by phenotype severity. The two subgroups in this pair are more phenotypically distinct than the other five subgroup pairs, leading to less overlap in SNP-sets prevalent in the two subgroups. In fact, we believe the higher phenotypic distinction between the dysmorphic and nondysmorphic

Table 3
The three most significant genes and gene pairs associated with each subgroup.

Subgroup	Most Significant Genes (adj. p-value) Most Significant Gene Pairs (adj. p-value)	Subgroup	Most Significant Genes (adj. p-value) Most Significant Gene Pairs (adj. p-value)
Dysmorphic	DYNC2H1 (0.0046)[*] AC008271.1 (0.0116)[nc] IGF2BP1 (0.0135)[*] DYNC2H1|HMGB1P32 (0.0135)[*|nc] NRAP|HS6ST2 (0.0148)[*|*] CSNK1G3|HMGB1P32 (0.0148)[*|nc]	Nondysmorphic	MXRA5Y (2.0e−06)[nc] IL1RAPL2 (0.0001)[a,p] MAMLD1 (0.0003)[*] HS6ST2|OFD1P6Y (0.0080)[*|nc] USP26|OFD1P6Y (0.0111)[*|nc] –
High-severity awareness	NR6A1 (0.0027)[*] ZNF559 (0.0027)[a] ZNF559-ZNF177 (0.0027)[*] – – –	Low-severity awareness	– – – – – –
High-severity cognition	DRP2 (1.0e−48)[p] PTCHD1-AS (1.0e−48)[nc] PCDH11X (3.26e−16)[p] GRIA3|TTTY5 (0.0079)[p|nc] PTCHD1-AS|TTTY5 (0.0119)[nc|nc] DRP2|TTTY5 (0.0197)[p|nc]	Low-severity cognition	UPF3B (0.0368)[a,p] SHROOM4 (0.0368)[p] – – – –
High-severity communication	HUWE1 (1.0e−48)[a,p] PCDH11Y (7.0e−47)[p] RP11-158 M9.1 (7.6e−18)[nc] DMD|TTTY5 (0.0042)[a,p|nc] RP13-126P21.2|TTTY5 (0.0044)[nc|nc] SLC25A14|TTTY5 (0.0071)[a,p|nc]	Low-severity communication	DMD (0.0349)[a,p] CA5B (0.0484)[*] – – – –
High-severity mannerism	NLGN4X (2.5e−11)[a,p] GRP173 (1.2e−10)[*] RP11-268G12.1 (1.2e−10)[nc] PTCHD1-AS|TTTY5 (0.0012)[nc|nc] TMEM164|TTTY5 (0.0039)[*|nc] TMEM164|TTTY11 (0.0244)[*|nc]	Low-severity mannerism	– – – – – –
High-severity motivation	LYPD5 (1.3e−11)[*] TTTY5 (3.1e−07)[nc] TTTY11 (5.7e−06)[nc] LYPD5|IL1RAPL2 (0.0058)[*|a,p] LYPD5|PCDH11X (0.0148)[*|p]	Low-severity motivation	TAF7L (4.8e−07)[*] PDK3 (0.0003)[*] DRP2 (0.0003)[p] – – –

[*] Gene novel to this study (not present in AutDB or PubMed abstract search).
[a] Gene present in AutDB.
[p] Gene found in PubMed abstract search.
[nc] Non-coding gene – various RNAs that are expressed but not translated.

subgroups leads to the larger quantity of significant SNP-sets separating these groups.

The major strength of this method is the ability to highlight genetic distinctions between cohort subgroups. We demonstrate this strength here by comparing broad categories of autism patients defined by a single distinction, regardless of how diverse the resulting subgroups are. However, HGCM is also capable of examining differences between much more specific pairs of subgroups, such as ones that have nearly identical characteristics other than a single-trait distinction. This would provide genetic associations that are much more specific to the state of the targeted trait, but would suffer from a reduction in sample size and, thus, statistical power. Our preliminary attempts for understanding more precise traits in this manner involve dividing and examining one of these analyzed subgroups. We have noticed that the more specific subgroups can have just as many associated genes as the original subgroup before it was partitioned, but the specific subgroups have surprisingly few gene associations in common with their corresponding broad subgroups.

While this study successfully identified several significant combinations of genes associated with autism subgroups, it does not sufficiently account for the interactions of multiple genotypes. We expect that many autism subtypes have contributory gene-gene interactions, but our method only identified significant multi-gene associations with two subgroups. We attribute this limitation to the FPM algorithm, which imposes a stringent mechanism for choosing the combinations of SNPs to test for association: the minor alleles for specific SNPs in each gene must be enriched in the subgroups. FPM also generated very few gene combinations with more than two genes, though we suspect that there are gene trios and larger gene combinations that are important for autism development in some subgroups. The few larger gene combinations that were generated as candidates were insignificant using the FBAT statistical test. A more forgiving method would utilize haplotypes or consider multiple SNPs in a region while searching for multi-gene associations. The discovered SNP-set associations must be quite significant indeed to have fulfilled this method's precise candidate generation requirements.

Several genes are associated with multiple subgroups. This is evidence that these genes are broadly associated with autism, and that this association extends to the subgroups. It is more difficult to interpret situations when a gene is associated with each member of a subgroup pair, such as the *DMD* example shown in Fig. 6. It is tempting to disregard these results as anomalies, but in this case our closer examination revealed a potential explanation, demonstrating why seemly contradictory associations should not be immediately dismissed.

Our HGCM method provides quality candidates for autism subgroup association, and these results should be taken into consideration during the genotype measurement for future studies so that the data exist to confirm or reject the highlighted genes as contributing to the development of autism. Additionally, the HGCM method is not disorder-specific and can be applied to other disorders with potential subgroups, in addition to contrasting additional autism subgroup pairs.

5. Conclusions

In this paper, we describe Heritable Genotype Contrast Mining (HGCM), a novel method for the discovery of genes and gene-pairs associated with disorder subgroups. HGCM integrates bioinformatics tools, distributed data mining algorithms, and a novel family inheritance metric to generate candidate SNPs and combinations of SNPs in *a priori* fashion, and these are tested for association with the subgroups. We utilized this method to contrast six autism subgroup pairs, including one pair of previously characterized subgroups defined by morphological features and five pairs defined by the scores of a widely-used diagnostic test. 286 genes were discovered to be associated with a subgroup, including 193 potentially novel autism candidates. We conclude that HGCM produced valuable associations between genes and autism subgroups, and can improve precision medicine practices by identifying genetic associations within other disorders that may comprise distinct subtypes. In particular, recent discoveries about novel subtyping of patients with Alzheimer's makes this disease a suitable target for this method [57]. The autism candidate genes identified in this study should be incorporated into future data collection to verify the significance of these associations and to identify the mechanisms through which these genes contribute to the development of autism subtypes.

Contributors

MS and CS designed the informatics procedure and methods to compare results to prior work. MS performed data preprocessing, analysis, and visualization. SC provided statistical expertise during the experimental design and while determining an appropriate statistical test. NT assisted in the process of requesting the data from the Simons Foundation and provided insights in the interpretation and parsing of the data sets. JM provided expertise on the morphological subgroups and genetics perspectives during the interpretation of results. MS draft the manuscript with substantial contributions from all co-authors. CS supervised the project development and manuscript writing. All authors approved the final manuscript and agree to all aspects of the work.

Funding

This work was supported by the National Institutes of Health [Grant Nos. 5T32GM008396, 5T32LM012410-02]; the Shumaker Endowment for Biomedical Informatics; the National Science Foundation [Grant No. CNS-1429294]; and the Simons Foundation [award number #26021565-08C000066].

Competing interests

None.

Acknowledgements

The authors would like to thank Dr. Stephen Kanne for many discussions on the interpretation and use of behavior data scored in the Simons collection, Dr. Zohreh Talebizadeh for the fruitful discussions on subgroup-focused research and potential follow-up directions for this study, and Dr. Michael Phinney for sharing his data mining software and for abundant technical advice.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2017.11.016.

References

[1] American Psychiatric Association. American Psychiatric Association Diagnostic and Statistical Manual of Mental Disorders. Washington, DC1664.
[2] V.C. Wong, C.K. Fung, P.T. Wong, Use of dysmorphology for subgroup classification on autism spectrum disorder in Chinese children, J. Autism Develop. Disorders 44 (1) (2014) 9–18.
[3] V.W. Hu, T. Sarachana, K.S. Kim, et al., Gene expression profiling differentiates autism case–controls and phenotypic variants of autism spectrum disorders: Evidence for circadian rhythm dysfunction in severe autism, Autism Res. 2 (2) (2009) 78–97.
[4] A.L. Alexander, J.E. Lee, M. Lazar, et al., Diffusion tensor imaging of the corpus callosum in Autism, Neuroimage. 34 (1) (2007) 61–73.
[5] V.W. Hu, A. Addington, A. Hyman, Novel autism subtype-dependent genetic variants are revealed by quantitative trait and subphenotype association analyses of published GWAS data, Plos One 6 (4) (2011) e19067.
[6] J.H. Miles, Autism spectrum disorders—a genetics review, Genet. Med. 13 (4) (2011) 278–294.
[7] J.H. Miles, T.N. Takahashi, J. Hong, et al., Development and validation of a measure of dysmorphology: useful for autism subgroup classification, Am. J. Med. Genet. Part A 146 (9) (2008) 1101–1116.
[8] H. Ozgen, G. Hellemann, M. de Jonge, et al., Predictive value of morphological features in patients with autism versus normal controls, J. Autism Develop. Disorders 43 (1) (2013) 147–155.
[9] H. Tager-Flusberg, Defining language impairments in a subgroup of children with autism spectrum disorder, Sci. China Life Sci. 58 (10) (2015) 1044–1052.
[10] I. Iossifov, B.J. O'Roak, S.J. Sanders, et al., The contribution of de novo coding mutations to autism spectrum disorder, Nature 515 (7526) (2014) 216–221.
[11] J. Sebat, B. Lakshmi, D. Malhotra, et al., Strong association of de novo copy number mutations with autism, Science 316 (5823) (2007) 445–449.
[12] S.J. Sanders, M.T. Murtha, A.R. Gupta, et al., De novo mutations revealed by whole-exome sequencing are strongly associated with autism, Nature 485 (7397) (2012) 237–241.
[13] N. Risch, D. Spiker, L. Lotspeich, et al., A genomic screen of autism: evidence for a multilocus etiology, Am. J. Human Genet. 65 (2) (1999) 493–507.
[14] S. Ozonoff, G.S. Young, A. Carter, et al., Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study, Pediatrics 128 (3) (2011) e488–e495.
[15] D.J. Weiner, E.M. Wigdor, S. Ripke, et al., Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders, Nat. Genet. (2017).
[16] P.J. Landrigan, What causes autism? Exploring the environmental contribution, Curr. Opin. Pediatr. 22 (2) (2010) 219–225.
[17] P.M. Rodier, S.L. Hyman, Early environmental factors in autism, Mental Retardat. Develop. Disabil. Res. Rev. 4 (2) (1998) 121–128.
[18] D. Rossignol, S. Genuis, R. Frye, Environmental toxicants and autism spectrum disorders: a systematic review, Translat. Psychiat. 4 (2) (2014) e360.
[19] M. Carter, S. Scherer, Autism spectrum disorder in the genetics clinic: a review, Clin. Genet. 83 (5) (2013) 399–407.
[20] R. Deth, C. Muratore, J. Benzecry, et al., How environmental and genetic factors combine to cause autism: a redox/methylation hypothesis, Neurotoxicology 29 (1) (2008) 190–201.
[21] R.K. Yuen, B. Thiruvahindrapuram, D. Merico, et al., Whole-genome sequencing of quartet families with autism spectrum disorder, Nat. Med. 21 (2) (2015) 185–191.
[22] T. Gaugler, L. Klei, S.J. Sanders, et al., Most genetic risk for autism resides with common variation, Nat. Genet. 46 (8) (2014) 881–885.
[23] J.H. Moore, The ubiquitous nature of epistasis in determining susceptibility to common human diseases, Human Hered. 56 (1–3) (2003) 73–82.
[24] B. Anderson, N. Schnetz-Boutaud, J. Bartlett, et al., Examination of association of genes in the serotonin system to autism, Neurogenetics 10 (3) (2009) 209–216.
[25] B. Anderson, N. Schnetz-Boutaud, J. Bartlett, et al., Examination of association to autism of common genetic variationin genes related to dopamine, Autism Res. 1 (6) (2008) 364–369.
[26] A.E. Ashley-Koch, J. Jaworski, H. Mei, et al., Investigation of potential gene–gene interactions between APOE and RELN contributing to autism risk, Psychiat. Genet. 17 (4) (2007) 221–226.
[27] K. Bowers, Q. Li, J. Bressler, et al., Glutathione pathway gene variation and risk of autism spectrum disorders, J. Neurodevelop. Disorders 3 (2) (2011) 132.
[28] D.B. Campbell, C. Li, J.S. Sutcliffe, et al., Genetic evidence implicating multiple genes in the MET receptor tyrosine kinase pathway in autism spectrum disorder, Autism Res. 1 (3) (2008) 159–168.
[29] S.J. Kim, C.W. Brune, E.O. Kistner, et al., Transmission disequilibrium testing of the chromosome 15q11-q13 region in autism, Am. J. Med. Genet. Part B: Neuropsychiat. Genet. 147 (7) (2008) 1116–1125.
[30] D. Ma, P. Whitehead, M. Menold, et al., Identification of significant association and gene-gene interaction of GABA receptor subunit genes in autism, Am. J. Human Genet. 77 (3) (2005) 377–388.
[31] S. Purcell, B. Neale, K. Todd-Brown, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Human Genet. 81 (3) (2007) 559–575.
[32] G.D. Fischbach, C. Lord, The Simons Simplex Collection: a resource for identification of autism genetic risk factors, Neuron 68 (2) (2010) 192–195.
[33] K. Shvachko, H. Kuang, S. Radia, et al., editors. The hadoop distributed file system. Mass Storage Systems and Technologies (MSST), in: 2010 IEEE 26th Symposium on, 2010, IEEE.

[34] M. Zaharia, M. Chowdhury, M.J. Franklin, et al., editors. Spark: cluster computing with working sets, in: Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, 2010.
[35] J.N. Constantino, C.P. Gruber, Social responsiveness scale (SRS): Western Psychological Services Los Angeles, CA, 2007.
[36] S.J. Sanders, X. He, A.J. Willsey, et al., Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci, Neuron 87 (6) (2015) 1215–1233.
[37] S.T. Sherry, M.-H. Ward, M. Kholodov, et al., dbSNP: the NCBI database of genetic variation, Nucl. Acids Res. 29 (1) (2001) 308–311.
[38] S.S. Verma, M. De Andrade, G. Tromp, et al., Imputation and quality control steps for combining multiple genome-wide datasets, Front. Genet. 5 (2014) 370.
[39] W.S. Bush, J.H. Moore, Genome-wide association studies, PLoS Comput. Biol. 8 (12) (2012) e1002822.
[40] S.R. Browning, B.L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering, Am. J. Human Genet. 81 (5) (2007) 1084–1097.
[41] R. Agrawal, R. Srikant, editors. Fast algorithms for mining association rules, in: Proc 20th Int Conf Very Large Data Bases, VLDB, 1994.
[42] J. Hipp, U. Güntzer, G. Nakhaeizadeh, Algorithms for association rule mining—a general survey and comparison, ACM sigkdd Explorat. Newslett. 2 (1) (2000) 58–64.
[43] S.D. Bay, M.J. Pazzani, Detecting group differences: mining contrast sets, Data Min. Knowl. Disc. 5 (3) (2001) 213–246.
[44] G. Dong, J. Li, editors. Efficient mining of emerging patterns: Discovering trends and differences, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999, ACM.
[45] P.K. Novak, N. Lavrac, G.I. Webb, Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining, J. Mach. Learn. Res. (2009) 377–403.
[46] S.M. Gogarten, T. Bhangale, M.P. Conomos, et al., GWASTools: an R/Bioconductor package for quality control and analysis of Genome-Wide Association Studies, Bioinformatics 28 (24) (2012) 3329–3331.
[47] R.C. Johnson, G.W. Nelson, J.L. Troyer, et al., Accounting for multiple comparisons in a genome-wide association study (GWAS), BMC Genom. 11 (1) (2010) 724.
[48] S. Horvath, X. Xu, N.M. Laird, The family based association test method: strategies for studying general genotype-phenotype associations, Eur. J. Human Genet.: EJHG. 9 (4) (2001) 301.
[49] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. Roy. Stat. Soc. Ser. B (Methodol.) (1995) 289–300.
[50] S.N. Basu, R. Kollu, S. Banerjee-Basu, AutDB: a gene reference resource for autism research, Nucl. Acids Res. 37 (suppl 1) (2009) D832–D836.
[51] R.K. Yuen, D. Merico, M. Bookman, et al., Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder, Nat. Neurosci. (2017).
[52] A. Bremer, M. Giacobini, M. Eriksson, et al., Copy number variation characteristics in subpopulations of patients with autism spectrum disorders, Am. J. Med. Genet. Part B: Neuropsychiatr. Genet. 156 (2) (2011) 115–124.
[53] L. French, P. Pavlidis, Relationships between gene expression and brain wiring in the adult rodent brain, PLoS Comput. Biol. 7 (1) (2011) e1001049.
[54] A. Piton, J. Gauthier, F. Hamdan, et al., Systematic resequencing of X-chromosome synaptic genes in autism spectrum disorder and schizophrenia, Mol. Psychiat. 16 (8) (2011) 867–880.
[55] S.R. Gilman, I. Iossifov, D. Levy, et al., Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses, Neuron 70 (5) (2011) 898–907.
[56] A. Krishnan, R. Zhang, V. Yao, et al., Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder, Nat. Neurosci. 19 (11) (2016) 1454–1462.
[57] D.E. Bredesen, Metabolic profiling distinguishes three subtypes of Alzheimer's disease, Aging (Albany NY). 7 (8) (2015) 595–600.