BMC
Medical Genomics

Performance of case-control rare copy number variation annotation in classification of autism

Worrawat Engchuan[1], Kiret Dhindsa[2], Anath C Lionel[3,4], Stephen W Scherer[3,4,5,6], Jonathan H Chan[1], Daniele Merico[3*]

Abstract

Background: A substantial proportion of Autism Spectrum Disorder (ASD) risk resides in *de novo* germline and rare inherited genetic variation. In particular, rare copy number variation (CNV) contributes to ASD risk in up to 10% of ASD subjects. Despite the striking degree of genetic heterogeneity, case-control studies have detected specific burden of rare disruptive CNV for neuronal and neurodevelopmental pathways. Here, we used machine learning methods to classify ASD subjects and controls, based on rare CNV data and comprehensive gene annotations. We investigated performance of different methods and estimated the percentage of ASD subjects that could be reliably classified based on presumed etiologic CNV they carry.

Results: We analyzed 1,892 Caucasian ASD subjects and 2,342 matched controls. Rare CNVs (frequency 1% or less) were detected using Illumina 1M and 1M-Duo BeadChips. Conditional Inference Forest (CF) typically performed as well as or better than other classification methods. We found a maximum AUC (area under the ROC curve) of 0.533 when considering all ASD subjects with rare genic CNVs, corresponding to 7.9% correctly classified ASD subjects and less than 3% incorrectly classified controls; performance was significantly higher when considering only subjects harboring *de novo* or pathogenic CNVs. We also found rare losses to be more predictive than gains and that curated neurally-relevant annotations (brain expression, synaptic components and neurodevelopmental phenotypes) outperform Gene Ontology and pathway-based annotations.

Conclusions: CF is an optimal classification approach for case-control rare CNV data and it can be used to prioritize subjects with variants potentially contributing to ASD risk not yet recognized. The neurally-relevant annotations used in this study could be successfully applied to rare CNV case-control data-sets for other neuropsychiatric disorders.

Background

Autism Spectrum Disorders (ASD) affect about 1% of the population, with a higher prevalence in males than females, and are characterized by impairments in social interaction and communication, as well as by repetitive and restricted behavior [1,2]. ASDs are highly heritable [3] and genomic studies have revealed that a substantial proportion of ASD risk resides in *de novo* germline and rare inherited genetic variation, ranging from chromosome abnormalities and copy number variation (CNV) [4-8] to single nucleotide variation [9-14]. Genomic studies have highlighted a striking degree of genetic heterogeneity, with variation distributed across numerous genes enriched in synaptic components as well as broader neuronal processes and neurodevelopmental pathways [15]. While numerous ASD loci have been recognized to date, they only account for a small fraction of the overall estimated heritability [16], consistent with the prediction that there are about 1000 loci underlying ASD [17].

Rare CNVs are expected to contribute to ASD risk in up to 10% of ASD subjects [15], whereas estimates for rare nucleotide substitutions and small insertions/deletions

* Correspondence: daniele.merico@sickkids.ca
[3]Program in Genetics and Genome Biology, The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario M5G 0A4, Canada
Full list of author information is available at the end of the article

range between 10 and 40% for different studies (depending on the sequencing technology, the modes of inheritance investigated and the depth of phenotypic characterization) [9-14].

Certain multigenic regions with rare yet recurrent inherited or *de novo* CNVs (genomic disorder loci such as 15q11.2-q13 duplications and 16p11.2 deletion/duplication), as well as rare variants disrupting single genes (such as *NRXN1*), have a well established contribution to ASD etiology. *De novo* CNVs are observed in up to 5-10% of screened ASD subjects; however, not all of these events have a clear contribution to ASD risk, which is thought to depend on the size of the genomic change and the gene pathways perturbed. In this sample collection, 3.0% of ASD subjects harbored a *de novo* or inherited genic CNVs classified as pathogenic according to clinical annotation guidelines [18] and consensus catalogue of ASD loci (124 genes and 55 loci) [16]; more than half of these pathogenic CNVs were *de novo*.

While univariate statistics are well suited for investigating global burden and for discovering pathways implicated in ASD, multivariate machine learning methods are best suited to achieve optimal classification performance using all available gene annotations. Data-driven classification is complementary to CNV clinical classification based on established ASD genetics, and thus helps expanding the number of "explained" subjects. Unlike other published studies focusing uniquely on pathogenic CNV classification [19], we aimed at classifying subjects based on the contribution of all rare CNVs.

Many classification algorithms have been proposed to solve different classification problems. Random Forest (RF) is an ensemble learning method that can be used for binary classification as well as for regression analysis. A RF classifier is composed of a large number of decision trees, each trained using a random subset of samples (bagging) and constructed by selecting the best splitting features from random and independent sampled subsets of features [20,21]. RF has been used in different fields, including computational biology; it is robust, accurate, with limited or no overfitting issues, and with transparent feature utilization. Despite its robustness, the relevance measures used by RF to select the best splitting features have been shown to be positively biased towards features with a larger number of categorical values; additional shortcomings occur in presence of highly correlated variables [22]. To avoid these issues, the Conditional Inference Forest (CF) classification method was proposed as a modification of RF that utilizes statistical inference testing for feature selection during tree construction [23].

Support Vector Machines (SVM) [24] and Artificial Neural Networks (NN) [25] are two popular classification methods, based on different mathematical models than Random Forest. SVM uses its kernel to project the data points from the feature space to a higher dimensional space, where a hyperplane is sought to maximize the separation between the nearest data points from each class (also called the support vectors). Artificial Neural Networks, in particular Multilayer Perceptrons (here called simply Neural Networks, or NNs) are built up of three or more layers of nodes and weighted connections between them; weights are set maximizing the rate at which the output layer correctly identifies the true class of given examples.

In this work, we investigated classification performance of different methods and the corresponding fraction of ASD subjects who can be reliably classified as cases.

Methods

ASD sample collection

Samples were collected as part of the Autism Genome Project (AGP), an international consortium with more than fifty contributing sites in North America and Europe; informed consent was obtained from all participants. ASD was diagnosed based on the ADOS (Autism Diagnostic Observation Schedule) and ADI-R (Autism Diagnostic Interview - Revised) [26,27]. Patients with karyotypic abnormalities, Fragile X syndrome or other genetic syndromes causing congenital malformations were excluded. Only samples from subjects with European ancestry were used for this analysis, as determined by multidimensional scaling analysis (MDS) of known SNPs (single nucleotide polymorphisms).

Unrelated controls were assembled from three large studies in which subjects had no previous psychiatric history: SAGE (Study of Addiction Genetics and Environment), Ontario Colorectal Cancer study, HABC (Health Aging and Body Composition) [28-31]. Controls were genotyped on the same array platforms (Illumina 1M single or duo arrays) as ASD subjects and parents, applying the same quality control procedures and CNV calling algorithms.

Part of the samples (stage 1) was analyzed and published in 2010 [8], whereas the remainder (stage 2) is being published in a separate paper [32].

CNV calling and clinical classification

Only samples meeting quality thresholds were used for CNV analysis. CNVs (of size 30 kb or greater) were detected using an analytical pipeline optimized for Illumina 1M arrays [8,33]. All *de novo* CNVs were experimentally validated. Samples with copy number variation greater than 7.5 MB were excluded.

The clinical classification of *de novo* and inherited CNVs as pathogenic, uncertain or benign was established according to the American College of Medical Genetics guidelines [18], on the basis of genetic loci known to be implicated in ASD (124 genes and 55 loci) [16]; large and

very rare CNVs were also classified as pathogenic, even if they were not reported before. Experimental CNV validation details can be found in [33].

CNV gene annotations and gene-set construction

Rare CNVs were mapped to genes whenever at least one transcript overlapped the CNV; transcript coordinates were based on RefSeq hg18 (the same build as for the BeadChip arrays). Similar performance results were obtained using the more stringent exonic mapping (i.e. requiring at least one exon to be overlapped by a CNV).

Curated neurally-relevant gene-sets captured (i) predicted haploinsufficiency [33], (ii) brain expression levels, based on the spatiotemporal Brainspan RNA-seq data-set [35], as well as brain expression specificity compared to other tissues, based on the Novartis Tissue Atlas microarrays [36], (iii) experimentally-determined synaptic complex membership [37] and regulation by *FMR1*, a key modulator of mRNA translation required for synaptic plasticity, based on two different methods [38,39], (iv) implication in neurological or neuropsychiatric disease in humans (according to HPO, the Human Phenotype Ontology) [40], (v) implication in abnormal nervous system or abnormal behavior in mice, according to Mammalian Phenotype Ontology (MPO) annotations provided by Mouse Genome Informatics (MGI) [41], and (vi) neuronal or nervous system function based on Gene Ontology (GO) [42] annotations or pathway membership. A more detailed description is provided in the supplementary methods and results [see Additional file 1].

Cross-validation strategy

Stratified three-fold cross-validation was used to avoid overfitting [43]. Subjects with at least one rare CNV were randomly divided into three equal subsets, each with the same proportion of cases and controls. The union of two subsets was used to train the model, while the remaining subset was used as the test set, to assess the classification performance. This process was repeated three times without re-dividing the data-set, so that each subset was used once as test set and twice as training set. Twenty independent cross-validation iterations were performed to estimate the mean and standard deviation of the area under the curve (AUC), to model the classification stochasticity. Absence of overfitting was further assessed by replacing real classification features with randomized features based on gene identity permutation.

Classifier implementation details

For RF, we used the implementation provided by the R/CRAN package 'randomForest' version 4.6-7 [44], which follows the original algorithm proposed by Breiman 2001 [20]. We used default settings unless otherwise

specified. For CF, we used the implementation provided by the R/CRAN package 'party' version 1.0-9 [45]. We used default settings unless otherwise specified. R 2.15.2 was used for all RF and CF analyses.

For the linear SVM, the libSVM package was used [46] in MATLAB R2013a. The cost parameter was kept at default as 1 and class weights were kept even. Each feature was independently normalized and rescaled to a 0-1 interval prior to being input into the classifier.

The Neural Network was built with two middle layers of 100 and 50 nodes each, a learning rate of 0.005 (which affects the rate of connection weight adaptation) with a 0.9 momentum (which affects the acceleration of connection weight adaptation). The network was trained through back-propagation, and unlike the SVM, it did not require feature normalization or scaling. The NN was also implemented in MATLAB R2013a.

Feature relevance metrics and feature selection

Two feature relevance metrics were utilized for RF: Mean Decrease Accuracy (MDA) and Mean Decrease Gini, as implemented in the randomForest package. MDA is calculated by permuting the value of the feature in OOB (out of bag) samples and comparing the accuracy of prediction before and after permutation tree by tree. Mean Decrease Gini is a measure of how each feature contributes to the homogeneity of the nodes and leaves.

MDA was also used for CF, as implemented by the party package function 'varimp', using the standard version (unless stated otherwise); in fact, the conditional version, which adjusts for correlations between predictor variables, is extremely computationally demanding.

For MDA-based selection, we ranked features based on MDA, and selected a predefined number of features. The procedure used for step-wise decorrelation and the detailed implementation of Minimum Redundancy Maximum Relevance Feature Selection (MRMR) [46] are described in detail in the supplementary methods and results [see Additional file 1].

Percentage of correctly classified ASD subjects

The percentage of correctly classified ASD subjects was calculated as the number of ASD subjects correctly predicted in at least 15 out of 20 iterations divided by the study total (1,892); the prediction probability cutoff was chosen ensuring that the percentage of correctly predicted ASD subjects without pathogenic or *de novo* CNVs exceeds the percentage of incorrectly predicted controls by more than 1.5x. Carriers of pathogenic and *de novo* CNV are easier to classify, thus the ratio described above is a more conservative estimate of the true positive to false positive ratio. Results did not change substantially when requiring subjects to be classified in at least 10 out of 20 iterations.

Results and discussion

Feature construction

We analyzed 1,892 ASD subjects (1623 males and 270 females) and 2,342 platform-matched controls (1093 males and 1250 females) with at least one rare CNV (frequency 1% or less); all subjects are of Caucasian ethnicity. Rare CNVs were mapped to gene transcripts; *de novo* and inherited rare CNVs were labeled as pathogenic, uncertain or benign following clinical annotation guidelines. Univariate burden and pathway analysis, as well as details on the clinical CNV classification, can be found in a separate publication [32].

Classification features were constructed for every subject as gene counts for each gene-set, i.e. counting how many genes participating in a gene-set harbor a rare genic gain or loss; in particular, separate features were constructed for gains and losses. For CNV mapping, we decided to focus on transcripts of known genes as there is a wealth of information that can be used to predict their implication in autism, and their boundaries are well characterized in the human genome; upstream transcription start site regulatory motifs, or other non-coding sequence could be the object of another paper.

Only subjects harboring at least one rare genic CNV were used for classification, as features would be constantly zero for the other subjects, but all subjects were considered when reporting percentage "explained" statistics. This resulted in a subset of 1,570 ASD subjects (80.8%) and 1,916 controls (81.8%), of which 958 ASD subjects and 1,113 controls had at least one genic loss, whereas 1,132 ASD subjects and 1,363 controls had at least one genic gain. 78 ASD subjects harbored at least one genic *de novo* CNV (30 gains, 51 losses); 57 ASD subjects subjects harbored at least one *de novo* or inherited CNV recognized as "pathogenic" according to clinical significance annotation [16,18] (21 gains, 37 losses). All subsets presented a gender composition similar to the full data-sets.

Gene-sets (and corresponding classification features) were organized in three groups: (a) 20 curated gene-sets of neurobiological relevance, capturing brain expression, synaptic components, neuro-phenotypes in human and mouse and predicted haploinsufficiency (Table 1), (b) gene-sets corresponding to GO annotations, (c) gene-sets corresponding to pathways (KEGG, Reactome, NCI Pathway Interaction Database, Biocarta databases) [48-53]. GO and pathway gene-sets were filtered to remove exceedingly large or small sets, resulting in 1425 GO sets (out of 5657) with 100 to 3000 genes, and 519 pathway sets (out of 1763) with 50 or more genes. The total gene count, regardless of gene-set membership, was also used as a classification feature.

For different classifiers and parameter settings, we classified all ASD subjects, or only the ASD subjects carrying a pathogenic or *de novo* rare CNV. This was

Table 1 Curated gene-sets description and gene number

Gene-set ID	Gene-set Description	Gene N#
hi015	Predicted haploinsufficiency (most inclusive)	8862
hi035	Predicted haploinsufficiency	4136
hi055	Predicted haploinsufficiency (most stringent)	2214
ExpsNov_BrainFeAd_sp	Specific expression in human adult or fetal brain (Novartis Tissue Atlas)	1285
Synapse_GrantFull	Post-synaptic density components	1407
FMR1_Targets_Darnell	FMR1 targets (Darnell et al)	840
FMR1_Targets_Ascano	FMR1 targets (Ascano et al)	927
thrEXPR_log2rpkm	Expressed in brain (BrainSpan)	13802
thr4.86_log2rpkm	Expressed in brain, very high (BrainSpan)	4595
thr3.32_log2rpkm	Expressed in brain, high/medium (BrainSpan)	4604
thr0.84_log2rpkm	Expressed in brain, medium/low (BrainSpan)	4603
thr.MIN_log2rpkm	Not expressed in brain (BrainSpan)	4600
PhHs_NervSys_ADX	Human nervous system phenotype (HPO), autosomal dominant or X-linked	620
PhHs_NervSys_All	Human nervous system phenotype (HPO)	784
PhHs_MindFun_ADX	Higher mental function phenotype (HPO), autosomal dominant or X-linked	395
PhHs_MindFun_All	Higher mental function phenotype (HPO)	687
MmHs_Neuro_All	Mouse neuro phenotype (MGI/MPO)	3479
MmHs_Extend_All	Mouse developmental phenotype (MGI/MPO)	4314
NeuroF_large	Neurobiological function, inclusive	2601
NeuroF_small	Neurobiological function, stringent	1088
Total	Total gene count	18203

helpful to evaluate the performance of our classification approach for more obviously implicated genes and loci. In addition, we classified either all ASD subjects with any rare genic CNV, or only the carriers of rare genic losses (using only loss-based features) or rare genic gains (using only gain-based features). This was helpful to evaluate the predictive power of the two CNV types.

Classification results: RF and CF

For all classifiers, we used a robust cross-validation approach and tested randomized features to ensure the absence of significant overfitting (Figure 1).

Random Forest (RF) was our original choice for its resilience to overfitting and its capability to handle a large feature space [20]. However RF has been criticized for its positive selection bias towards features with more categorical values [22]; for this reason, Conditional Inference Forest (CF) has been suggested as an alternative to RF [23].

The performance of RF and CF was compared for the 20 curated neurally-relevant features plus the total gene count. We found CF to have slightly higher AUC than

RF when classifying all subjects (all CNV types, loss-only, gain-only), whereas it had slightly lower AUC than RF for *de novo* and pathogenic subjects (Table 2). We inspected the feature relevance metrics for both classifiers, using the Mean Decrease Accuracy and Mean Decrease Gini for RF, and Mean Decrease Accuracy, with or without correlation adjustment, for CF. Both RF metric showed greater relevance for gain-based features when classifying all subjects, in contradiction with the fact that classification based uniquely on losses displayed a better AUC than the one based on gains. On the other hand, CF relevance metrics showed the opposite pattern, with loss-based features more relevant than gain-based features, in accordance with expectations (Figure 2). For this reason, we used CF in place of RF for the rest of the analyses.

The AUC achieved by CF using the 20 curated neurally-relevant features and the total count was greater by several standard deviation (sd) units than the AUC achieved using the total gene count alone; it was also greater than the AUC achieved by the total gene count plus 20 matched randomized features obtained by permuting the gene identities and re-computing the gene-set counts for gains and losses. This was particularly the case when classifying all subjects; on the other hand, when classifying only *de novo* or pathogenic CNV carriers versus controls, gains displayed an AUC close to the AUC obtained using the total gene count alone. This can be interpreted in relation to the larger size of pathogenic and *de novo* gains compared to control gains. In addition, as expected, adding the 20 randomized features to the total gene count did not lead to a remarkable increase of the AUC (i.e. lower than or within one sd unit) (Table 3).

CF classification using the 20 curated neurally-relevant features (gains and losses), together with the total gene count, resulted in 7.9% correctly classified ASD subjects and less than 3% incorrectly classified control; this result is reasonably close to the expected contribution of rare CNVs to ASD risk in 10% of the ASD subjects. Losses alone correctly classified 5.9% of the ASD subjects and less than 2.3% incorrectly classified controls; this suggests that gains, besides the ones already recognized to be pathogenic, have limited predictive power, a result in line with univariate burden analysis for this data-set [32].

A detailed analysis of feature relevance for all subjects, or only *de novo* and pathogenic, is presented in the supplementary methods and results [see Additional file 1].

Classification results for GO and pathways using different feature selection strategies

The set of features based on GO annotations and pathways presented the additional challenges of having many
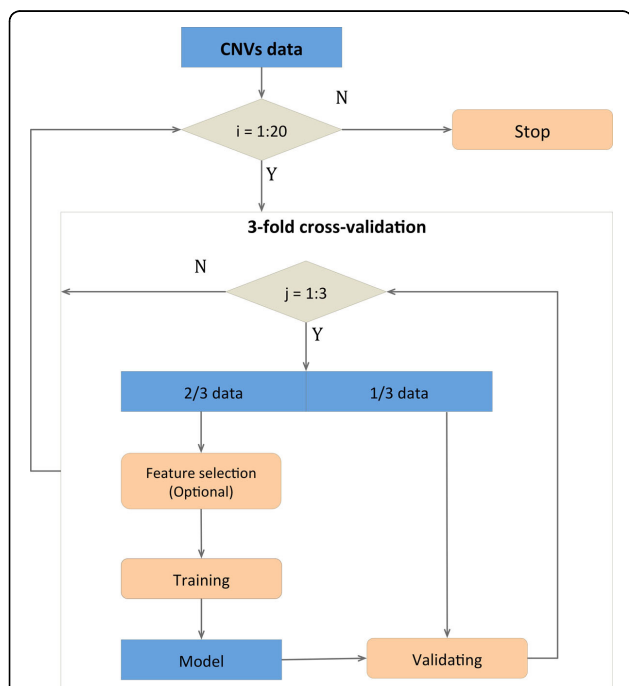
Figure 1 Cross-validation strategy. The data-set is divided into three equal subsets, each with the same propotion of ASD and control subjects. Two of the tree subsets are used as the training set the model, whereas the other subset is used as the validation set for performance quantification; this is iterated three times, so that each subset is used twice for training and once for validation. The feature selection is performed only for GO and pathway-based features. The remaining set is used as test set to assess the performance of classification. The cross-validation procedure is repeated times to estimate the mean performance and its standard deviation.

Table 2 CF and RF classification performance for 20 neurally-relevant curated features (mean ± sd)

Subject	Classifier	All CNV	Gain CNV	Loss CNV
All subjects	RandomForest	0.531±0.005	0.509±0.004	0.544±0.006
All subjects	CForest	0.533±0.004	0.513±0.005	0.546±0.003
De novo	RandomForest	0.805±0.012	0.769±0.024	0.840±0.010
De novo	CForest	0.787±0.008	0.732±0.013	0.846±0.011
Pathogenic	RandomForest	0.913±0.014	0.913±0.012	0.935±0.016
Pathogenic	CForest	0.880±0.012	0.897±0.008	0.922±0.030

features, with a high degree of mutual overlap and presence of many features with limited or no classification relevance.

To address those issues, we performed classification based on GO and pathway features in two ways: (a) using all features, (b) embedding a feature selection step in the cross-validation procedure; in particular, we used (i) Mean Decrease Accuracy (MDA) based selection, (ii) MDA based selection with stepwise decorrelation, (iii) MRMR (Minimum Redundancy Maximum Relevance Feature Selection). For each procedure, we selected the top 20, top 15% and top 40% ranking features excluding the total gene count, and then added the total gene count. It is important to point out that feature selection was based on the feature relevance metrics calculated on the data subset used for training, and performed independently for every training set, to avoid any over-fitting issues.

We assessed the classification performance of different feature selection strategies in comparison to classification without any extra feature selection step or performing manual feature selection based on previous knowledge of ASD biology.

GO without feature selection produced a suboptimal performance (the AUC was very similar to using the
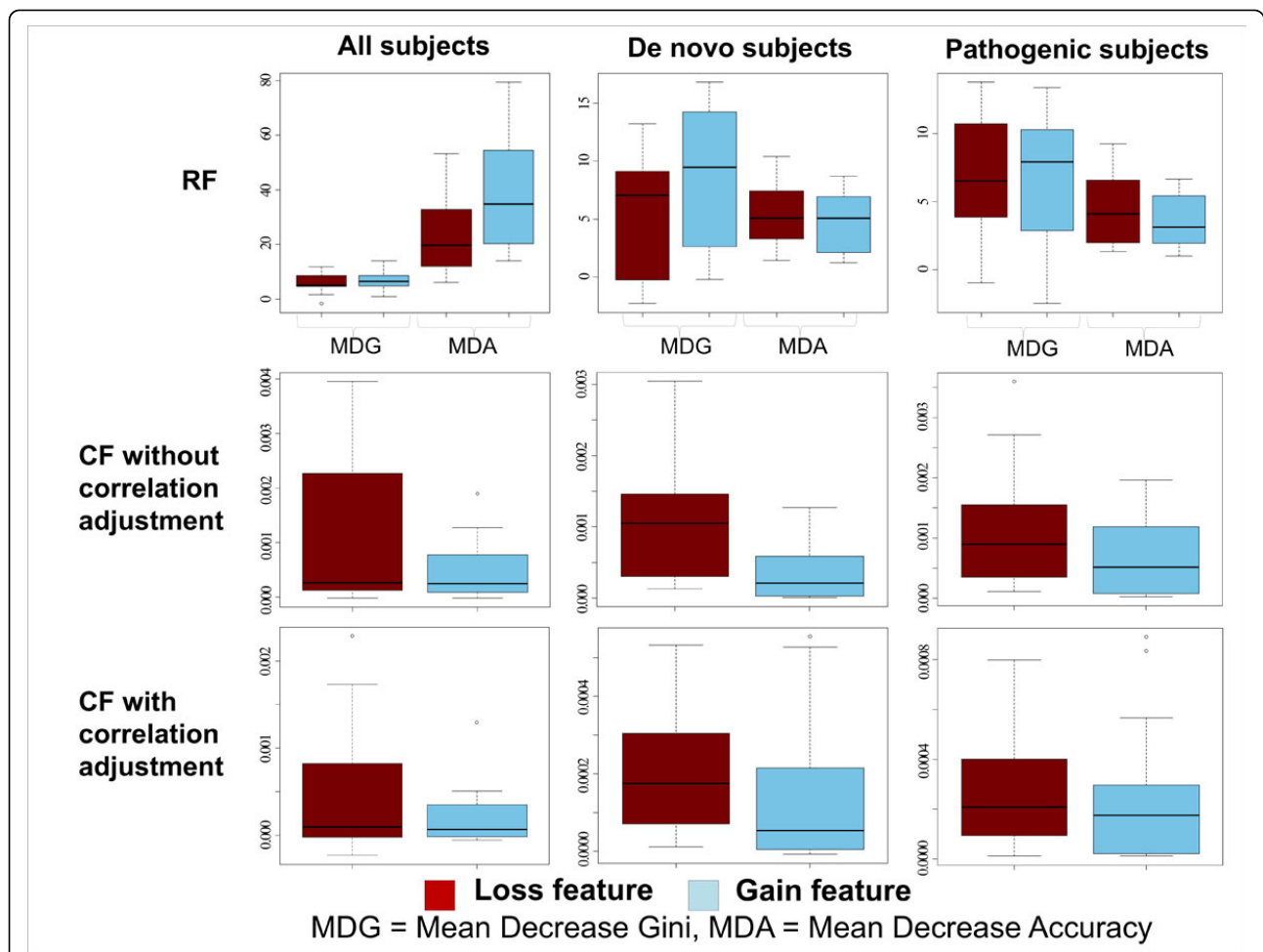
Figure 2 RF and CF feature relevance, boxplots for the 20 curated neurally-relevant features. Feature relevance boxplots for loss-based features (red) and gain-based features (blue). Mean decrease gini (MDG) and Mean decrease accuracy (MDA) were used for RF. MDA, with and without correlation adjustment, was used for CF. For all relevance metrics, higher values correspond to more relevant features.

gene count alone). When classifying all subjects, the best results for GO were achieved by MDA, either by taking the top 20 features using MDA with decorrelation or by taking the top 15% using MDA without decorrelation; after decorrelation, the top 15% features had a lower performance, suggesting that many relevant yet highly correlated features are removed by decorrelation (see Additional file 1). The best feature selection strategy achieved a slightly better performance than the manually selected GO sub-set (1 sd unit or more), but still inferior to the 20 curated neurally-relevant features (Table 3).

The performance for pathway-based features was markedly worse, with the AUC very close to the total gene count, even when restricting to the manually-selected pathway subset and with the best feature selection strategy (top 20 features using MDA and decorrelation) (Table 3).

These results suggest that feature selection is able to manage the large number and redundancy of GO and pathway-based, although their information content for ASD classification based on rare CNV genes appears to be more limited than the curated neurally-relevant features. Pathway-based features may have particularly disappointing performance results because of the small size of the corresponding gene-sets. Based on these results, we preferred using only the 20 curated neurally-relevant features for the other analyses. It is important to consider that the 20 curated neurally-relevant features include two features derived from the gene-set union of the GO and pathways manually-selected on the basis of known ASD pathobiology.

Classification results using other classifiers and modifying CF parameters

Classification of all subjects using linear SVM or NN achieved lower or comparable AUC compared to CF

[see Additional file 1]. This suggests CF is an optimal classification method for this problem.

Additional analysis demonstrated that CF performance does not vary substantially modifying model construction settings (inferential tests and p-value threshold for feature selection operating for tree construction) [see Additional file 1].

Finally, we compared the performance of our classifier based on the 20 neurally-relevant feature to an existing classifier designed to distinguish benign from pathogenic CNV for mental retardation (GeCCO [19]). We found our classifier to have a significantly better performance [see Additional file 1].

Prioritization of subjects

We used the classification results based on the 20 curated neurally-relevant features and total gene count to prioritize subjects with potentially interesting inherited rare losses; we focused on inherited rare variation because it is more difficult to assess its significance outside known ASD loci, and we selected rare losses as these display better classification performance than rare gains.

To ensure the classification results are robust to the algorithm's stochasticity, we required a subject to be classified in at least 15 cross-validation iterations out of 20. We scanned the prediction probability cutoffs to maximize the number of correctly predicted ASD subjects without pathogenic or *de novo* CNVs with respect to incorrectly classified controls. At the ASD prediction probability cutoff of 0.52, the correctly classified ASD subjects represented 3.76% of all ASD subjects with at least one genic rare loss, which dropped to 1.25% (12 subjects) when removing subjects with pathogenic or *de novo* rare CNVs (TP*); this was still larger (about 3.5x) than 0.36% of incorrectly classified controls (FP, false positives), interpretable as a false discovery rate < 30% [see Additional file 1].

We manually inspected these 12 subjects, and found that one of them had a pathogenic loss missed by the clinical annotation, whereas the others had at least one VUS (Variant of Unknown Clinical Significance) each; we used a set of top-scoring loss-based features to identify specific genes that are more likely to contribute to ASD risk within the VUS variant [see Additional file 5].

Finally, we also noticed that prioritized subjects were significantly enriched in females compared to all autism subjects with at least one rare genic loss (Fisher's Exact Test p-value 0.0066 and OR 2.7). A similar enrichment was reported when considering pathogenic CNVs further classified as "highly penetrant" based on clinical genetic literature [32].

Conclusions

We successfully used rare CNVs and neurally-relevant gene annotations to classify ASD subjects: the best classifier

Table 3 Classification results for all subjects using 20 neurally-relevant curated features, 20 matched randomized features, Gene Ontology and pathways (mean ± sd)

Gene set (All subjects)	All CNV	Gain CNV	Loss CNV
20 curated	0.533±0.004	0.513±0.005	0.546±0.003
GO	0.512±0.005	0.506±0.002	0.519±0.002
GO (man. selected)	0.520±0.005	0.505±0.005	0.524±0.003
GO (f.s.: 20 MDA dec.)	0.524±0.003	0.510±0.003	0.529±0.005
Pathway	0.500±0.000	0.500±0.000	0.504±0.004
Pathway (man. selected)	0.500±0.000	0.500±0.001	0.510±0.004
Pathway (f.s.: 20 MDA dec.)	0.513±0.003	0.510±0.004	0.513±0.003
Random (20 curated)	0.517±0.005	0.510±0.007	0.515±0.007
Total count	0.515±0.005	0.505±0.005	0.516±0.004

achieved an AUC of 0.533, corresponding to 7.9% ASD subjects correctly classified by rare CNVs and less than 3% incorrectly classified controls; this result is reasonably close to prior expectations that about 10% of ASD subjects have rare CNV contributing to ASD risk.

Conditional Inference Forest (CF) typically performed as well as or better than other classifiers, and was also found to have a stable performance when using parameter settings different than defaults. Losses alone displayed a markedly stronger classification power than gains; in addition, features based on brain expression, synaptic component and neuro-phenotypes had a superior performance to the full collection of GO and pathways, even after the latter were pre-processed by feature selection and de-correlation methods. This classification approach can be used for other case-control rare CNV data-sets; the features we found to be optimal for ASD are likely to perform well for other neurodevelopmental and neuropsychiatric disorders (e.g. developmental delay, schizophrenia), which also display a specific burden of rare variation for neuronal and neurodevelopmental genes [54-57].

The CF classifier based on neurally-relevant features was also successfully used to extract subjects with inherited losses potentially contributing to ASD risk, but not classified as pathogenic CNVs according to clinical annotation. Nonetheless, since classification performance is particularly high for pathogenic and *de novo* CNV carriers, this type of analysis can also be used to prioritize subjects in the absence of clinical annotations.

Within this classification framework, improved performance could perhaps be achieved by (i) using additional indexes of genic intolerance to variation [58] (ii) modeling the different level of constraint for different gene components (coding exons, introns, UTRs), (iii) expanding the gene annotation feature set, for instance using gene interaction network distances from known ASD disease genes [59], (iv) modeling non-coding RNAs, regulatory sequences and other inter-genic elements [60]), (v) weighting differently X-linked variants based on the subjects' genders. Modeling gender more accurately is of particular interest, considering the higher prevalence of autism in males compared to females; it would be ideal to train and assess the performance of the classifier for male and female subjects separately, which was not possible for our data-set owing to the small number of female subjects.

Availability of supporting data
The complete set of stage-1 CNV calls is currently available in dbGAP as phs000267.v3.p2; stage-2 will be soon available in dbGAP as phs000267.v4.p2. Rare variants for ASD subjects and controls are provided as additional files [see Additional file 3-4].

List of abbreviations used (if any)
ASD: Autism Spectrum Disorders; AUC: Area Under the ROC Curve; CF: Conditional Inference Forest; CNV: Copy Number Variant; GO: Gene Ontology; HPO: Human Phenotype Ontology; MDA: Mean Decrease Accuracy; MGI: Mouse Genome Informatics; MPO: Mammalian Phenotype Ontology; MRMR: Minimum Redundancy Maximum Relevance Feature Selection; NN: Neural Network; OOB: Out Of Bag; RF: Random Forest; SD: Standard Deviation; SVM: Support Vector Machine(s); VUS: Variant of Unknown Clinical Significance

Additional material

> **Additional file 1: Supplementary methods and results.**
>
> **Additional file 2: Twenty curated neurally-relevant feature genes** Gene-set names and genes (entrez gene ID) used to construct the twenty curated neurally-relevant features.
>
> **Additional file 3: Rare CNV data for ASD subjects** Rare CNV data for ASD subjects.
>
> **Additional file 4: Rare CNV data for controls** Rare CNV data for controls.
>
> **Additional file 5: Prioritized ASD subjects carrying rare inherited losses** Subject IDs, CNV coordinates, CNV clinical annotation, CNV gene mapping and gene annotations for the 12 prioritized ASD subjects carrying rare inherited losses potentially contributing to ASD risk.

Authors' details
[1]Data and Knowledge Engineering Laboratory, School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok

10140, Thailand. [2]Neurotechnology and Plasticity Lab, School of Computational Science and Engineering, McMaster University, Hamilton, Ontario L8S 4L8, Canada. [3]Program in Genetics and Genome Biology, The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario M5G 0A4, Canada. [4]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada. [5]McLaughlin Centre, University of Toronto, Toronto, Ontario M5G 0A4, Canada. [6]Centre of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, P.O. Box: 80216, Jeddah 21589, KSA.

References

1. Lai MC, Lombardo MV, Baron-Cohen S: Autism. *Lancet* 2014, **383**(9920):896-910.
2. Elsabbagh M, Divan G, Koh YJ, Kim YS, Kauchali S, Marcin C, Montiel-Nava C, Patel V, Paula CS, Wang C, *et al*: Global prevalence of autism and other pervasive developmental disorders. *Autism research: official journal of the International Society for Autism Research* 2012, **5**(3):160-179.
3. Constantino JN, Todorov A, Hilton C, Law P, Zhang Y, Molloy E, Fitzgerald R, Geschwind D: Autism recurrence in half siblings: strong support for genetic mechanisms of transmission in ASD. *Molecular psychiatry* 2013, **18**(2):137-138.
4. Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, Kendall J, Marks S, Lakshmi B, Pai D, Ye K, *et al*: Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* 2011, **70**(5):886-897.
5. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, *et al*: Structural variation of chromosomes in autism spectrum disorder. *American journal of human genetics* 2008, **82**(2):477-488.
6. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, Chu SH, Moreau MP, Gupta AR, Thomson SA, *et al*: Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 2011, **70**(5):863-885.
7. Autism Genome Project C, Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu XQ, Vincent JB, Skaug JL, Thompson AP, *et al*: Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature genetics* 2007, **39**(3):319-328.
8. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, *et al*: Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 2010, **466**(7304):368-372.
9. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A, *et al*: De novo gene disruptions in children on the autistic spectrum. *Neuron* 2012, **74**(2):285-299.
10. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, *et al*: Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 2012, **485**(7397):242-245.
11. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, *et al*: De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 2012, **485**(7397):237-241.
12. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, *et al*: Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 2012, **485**(7397):246-250.
13. Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, Ataman B, Schmitz-Abe K, Harmin DA, Adli M, Malik AN, *et al*: Using whole exome sequencing to identify inherited cause of autism. *Neuron* 2013, **77**:259-273.
14. Jiang YH, Yuen RK, Jin X, Wang M, Chen N, Wu X, Ju J, Mei J, Shi Y, He M: Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *American journal of human genetics* 2013, **93**(2):249-63.
15. Devlin B, Scherer SW: Genetic architecture in autism spectrum disorder. *Current opinion in genetics & development* 2012, **22**(3):229-237.
16. Betancur C: Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain research* 2011, **1380**:42-77.
17. Buxbaum JD, Daly MJ, Devlin B, Lehner T, Roeder K, State MW, Autism Sequencing Consortium: The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron* 2013, **76**:1052-1056.
18. Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST, Working Group of the American College of Medical Genetics Laboratory Quality Assurance C: American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genetics in Medicine : official journal of the American College of Medical Genetics* 2011, **13**(7):680-685.
19. Hehir-Kwa JY, Wieskamp N, Webber C, Pfundt R, Brunner HG, Gilissen C, de Vries BB, Ponting CP, Veltman JA: Accurate distinction of pathogenic from benign CNVs in mental retardation. *PLoS computational biology* 2010, **6**(4): e1000752.
20. Breiman L: Random Forest. *Machine Learning Journal* 2001, **45**:5-32.
21. Breiman L, Friedman JH, Olshen RA, Stone CJ: Classification and regression trees. Belmont, California: Wadsworth; 1984.
22. Strobl C, Boulesteix AL, Zeileis A, Hothorn T: Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics* 2007, **8**:25.
23. Hothorn T, Hornik K, Zeileisa A: Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 2006, **15**:651-671.
24. Burges CJ: A Tutorial on Support vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 1998, **2**(2):121-167.
25. Haykin S: Neural Networks and Learning Machines. Prentice Hall;, 3 2008.
26. Lord C, Rutter M, Goode S, Heemsbergen J, Jordan H, Mawhood L, Schopler E: Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. *Journal of autism and developmental disorders* 1989, **19**(2):185-212.
27. Lord C, Rutter M, Le Couteur A: Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders* 1994, **24**(5):659-685.
28. Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, Fisher S, Fox L, Howells W, Bertelsen S, *et al*: A genome-wide association study of alcohol dependence. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(11):5082-5087.
29. Figueiredo JC, Lewinger JP, Song C, Campbell PT, Conti DV, Edlund CK, Duggan DJ, Rangrej J, Lemire M, Hudson T, *et al*: Genotype-environment interactions in microsatellite stable/microsatellite instability-low colorectal cancer: results from a genome-wide association study. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2011, **20**(5):758-766.
30. Newcomb PA, Baron J, Cotterchio M, Gallinger S, Grove J, Haile R, Hall D, Hopper JL, Jass J, Le Marchand L, *et al*: Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2007, **16**(11):2331-2343.
31. Fox CS, Liu Y, White CC, Feitosa M, Smith AV, Heard-Costa N, Lohman K, Consortium G, Consortium M, Consortium G, *et al*: Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS genetics* 2012, **8**(5):e1002695.
32. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, Thiruvahindrapuram B, Xu X, Ziman R, Wang Z: Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *American journal of human genetics* 2014, **94**(5):677-694.
33. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, *et al*: Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature biotechnology* 2011, **29**(6):512-520.
34. Huang N, Lee I, Marcotte EM, Hurles ME: Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics* 2010, **6**(10): e1001154.
35. BrainSpan: Atlas of the Developing Human Brain. [http://developinghumanbrain.org].
36. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al*: A gene atlas of the mouse and human

protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(16):6062-6067.

37. Bayes A, van de Lagemaat LN, Collins MO, Croning MD, Whittle IR, Choudhary JS, Grant SG: **Characterization of the proteome, diseases and evolution of the human postsynaptic density.** *Nature neuroscience* 2011, **14**(1):19-21.

38. Darnell JC, Van Driesche SJ, Zhang C, Hung KY, Mele A, Fraser CE, Stone EF, Chen C, Fak JJ, Chi SW, *et al*: **FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism.** *Cell* 2011, **146**(2):247-261.

39. Ascano M Jr., Mukherjee N, Bandaru P, Miller JB, Nusbaum JD, Corcoran DL, Langlois C, Munschauer M, Dewell S, Hafner M, *et al*: **FMRP targets distinct mRNA sequence elements to regulate protein expression.** *Nature* 2012, **492**(7429):382-386.

40. Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J, *et al*: **The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data.** *Nucleic acids research* 2014, **42**(Database issue):D966-974.

41. Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE, Mouse Genome Database G: **The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse.** *Nucleic acids research* 2014, **42**(Database issue):D810-817.

42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature genetics* 2000, **25**(1):25-29.

43. Refaeilzadeh P, Tang L, Liu H: **Cross-validation.** *Encyclopedia of Database Systems* 2009, 532-538.

44. Liaw A, Wiener M: **Classification and Regression by randomForest.** *R News* 2002, **2**(3):18-22.

45. Hothorn T, Hornik K, Zeileis A: **party: A Laboratory for Recursive Partytioning.** *CRAN* 2013.

46. Chang CC, Lin CJ: **LIBSVM: a library for support vector machines.** *ACM Transactions on Intelligent Systems and Technology* 2011, **2**:27:21-27:27.

47. Peng H, Long F, Ding C: **Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy.** *IEEE transactions on pattern analysis and machine intelligence* 2005, **27**(8):1226-1238.

48. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, *et al*: **Reactome: a database of reactions, pathways and biological processes.** *Nucleic acids research* 2011, **39**(Database issue):D691-697.

49. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic acids research* 2000, **28**(1):27-30.

50. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data, information, knowledge and principle: back to metabolism in KEGG.** *Nucleic acids research* 2014, **42**(Database issue):D199-205.

51. Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, D'Eustachio P, Stein L: **Annotating cancer variants and anti-cancer therapeutics in reactome.** *Cancers* 2012, **4**(4):1180-1211.

52. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database.** *Nucleic acids research* 2009, **37**(Database issue):D674-679.

53. Nishimura D: **BioCarta.** *Biotech Software & Internet Report* 2001, **2**:117-120.

54. Webber C, Hehir-Kwa JY, Nguyen DQ, de Vries BB, Veltman JA, Ponting CP: **Forging links between human mental retardation-associated CNVs and mouse gene knockout models.** *PLoS genetics* 2009, **5**(6):e1000531.

55. Shaikh TH, Haldeman-Englert C, Geiger EA, Ponting CP, Webber C: **Genes and biological processes commonly disrupted in rare and heterogeneous developmental delay syndromes.** *Human molecular genetics* 2011, **20**(5):880-93.

56. Kirov G, Pocklington AJ, Holmans P, Ivanov D, Ikeda M, Ruderfer D, Moran J, Chambert K, Toncheva D, Georgieva L, *et al*: **De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia.** *Molecular psychiatry* 2012, **17**(2):142-53.

57. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, O'Dushlaine C, Chambert K, Bergen SE, Kähler A: **A polygenic burden of rare disruptive mutations in schizophrenia.** *Nature* 2014, **506**(7487):185-90.

58. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB: **Genic intolerance to functional variation and the interpretation of personal genomes.** *PLoS genetics* 2013, **9**(8):e1003709.

59. Oti M, Snel B, Huynen MA, Brunner HG: **Predicting disease genes using protein-protein interactions.** *Journal of medical genetics* 2006, **43**(8):691-698.

60. Walker S, Scherer SW: **Identification of candidate intergenic risk loci in autism spectrum disorder.** *BMC genomics* 2013, **14**:499.