Online Supplemental Data

Supplemental Files

Online data supplement accompanies the paper on the American Journal of Psychiatry website (http://ajp.psychiatryonline.org/) including:

Appendix I: Catalog of Phenotypic Features.

Table S1: Mapping SNP Sets into Genomic Information. (Information obtained from HaploReg v2, dbSNP and NCBI databases)

Table S2: Mapping Genes Targeted by SNP Sets to Mental and Brain and Nervous System Disorder Categories.
Table S3: Molecular Consequences of SNP Variants.
Table S4: Molecular Pathways and Ontologies Identified in the Genotypic-Phenotypic Architecture of Schizophrenia.
Table S5: Genotypic-Phenotypic AND/OR Relationships. (Table 3 extended).
Table S6: Phenotypic Feature Grouping by PCA and Clustering the Molecular Genetics of Schizophrenia Sample (Group), and Validation with the Literature and Expert Knowledge (Symptoms)
Table S7: Reproducibility of the Genotypic-Phenotypic Relationships of the Molecular Genetics of Schizophrenia Study in the CATIE Study
Table S8: Reproducibility of the Genotypic-Phenotypic Relationships of the Molecular Genetics of Schizophrenia Study in the Portuguese Islands Study
Table S9: Summary of the Reproducibility of the Molecular Genetics of Schizophrenia Dataset in the CATIE and the Portuguese Islands Studies
Table S10: Phenotypic Features of the CATIE and the Portuguese Islands Studies
Figures S1-S7.

Genomics Dataset: Gain and NonGain Studies

We first investigated the architecture of schizophrenia (SZ) using the Gain and NonGain genome wide association studies (GWAS) as our main targets, which are coherent case-control studies performed in a single lab under similar conditions. This study contains data from 8023 subjects, 4196 patients and 3827 controls, combining data from Euro-American ancestry (EA) and African-American ancestry (AA). Genotyping was carried using the Affymetrix 6.0 array, which assays 906,600 SNPs.

This study was originally performed in part at Washington University. Study population, ascertainment, phenomics and genomic datasets, as well as other information relative to this study can be accessed in the dbGaP web page [http://www.ncbi.nlm.nih.gov/gap/] by their identifiers: *phs000021.v3.p2* and *phs000167.v1.p1* for GAIN and NonGAIN projects, respectively.

The genotype data was codified in a matrix [SNPs x subjects], where the columns and rows correspond to subjects and SNPs, respectively. In each cell of the matrix, the value for the corresponding SNP and subject is assigned as 1, 2, and 3 for the SNP allele values AA, AB, and BB, respectively. Missing values were initialized by 0.

Data Cleaning

The quality control (QC) of the genotypic data was performed following the steps detailed in references (1, 2), removing consequently all the SNPs satisfying the next criteria:

1) SNP call rate < 95% in either GAIN or NonGAIN or combined datasets.
2) Hardy-Weinberg (HWE) p - value < 10E - 6 based on control samples in either GAIN or NonGAIN or combined, (using only females for chr X SNPs).
3) Minor Allele Frequency (MAF) < 1% in combined dataset.
4) Failed plate effect test in GAIN, NonGAIN or combined dataset.
5) MENDEL errors > 2 in either GAIN or NonGAIN.
6) >1 disconcordant genotypes in either GAIN 29 duplicates or NonGAIN 32 duplicates.
7) >2 disconcordant genotypes for 93 (=3×31 trios) samples genotyped in both GAIN and NonGAIN.

A total of 209,321 SNPs were excluded due to the restrictions described above from the total 906,109 SNPs genotyped. Therefore, 696,788 SNPs passed the QC filters. Then, 2891 SNPs were pre-selected to reduce the large search space using the logistic association function included in the PLINK software suite (3), taking sex and ancestry as co-variates, and establishing a generous threshold (p-value < 0.01). This threshold was established as 0.01 because this is approximately the value used in the supplementary tables reported in (1) for AA, EA and AA-EA analyses.

Methodology: a Divide & Conquer Strategy to Dissect a GWAS into the Genotypic-Phenotypic Architecture of a Disease

Overview

To uncover the architecture of SZ we applied a "Divide & Conquer" strategy (see Figures S1 and S2) that is commonly used in computer science to solve complex problems such as those of proteomics and transcriptomics (4-7) and cancer identification (8). Here we applied this strategy to dissect a single GWAS into multiple genotypic and/or phenotypic networks (9), as an attempt to extract the maximum information even from one dataset.

The "divide" step deconstructs genotypic and phenotypic data independently, and explores multiple local patterns (i.e., SNP sets and phenotypic sets). We used non-negative matrix factorization methods that have been applied to characterize complex genomic (10-13) and social profiles (14, 15), and generalized them to approach GWA data in a purely data-driven and unbiased fashion (16). Thus, our systematic grouping strategy is not directed by previous knowledge of polygenic involvement in SZ, does not limit subjects to only one SNP set (12, 13, 17), and does not predefine the number of SNP sets, avoiding possible biases and assumptions that relationships are linear, regular, or random (9, 18-21). Unlike other approaches (17, 22), we do not constrain SNP sets to a particular genome feature or to be in linkage disequilibrium (LD), and the phenotypic status of the subjects is not considered in SNP set formation (i.e., it is unsupervised (10, 11)). After incorporating phenotypic status *a posteriori* within each set (e.g., cases and controls), we establish their statistical significance with powerful and well-founded test methods that perform the appropriate corrections for the use of SNP sets (16, 17, 22-25), as well as provide an unbiased risk surface of disease to test predictions (16, 26).

The "conquer" step consists of three stages. First, assembling the uncovered local components of the genotypic architecture into genotypic networks of SNP sets, where two SNP sets are connected if they (i) comprise different sets of subjects described by similar sets of SNPs, (ii) and/or if they have similar sets of subjects but characterized by distinct sets of SNPs, (iii) and/or if one of the two SNP sets contains a subset of subjects and SNPs of the other SNP set. Second, optimally combining (10, 11, 16, 27) the local components of the phenotypic architecture (i.e., phenotypic sets) with the genotypic sets to expose the joint genotypic-phenotypic architecture of the disease. Third, evaluating complexity in the pathway from SNP sets to phenotypic sets; some connected SNP-set networks may be candidates to converge to equifinality, whereas other disjoint networks can lead to multifinality (i.e., recognizing a collection of diseases).

Finally, we carried out independent analyses to test for possible confirmations of the heterogeneous architecture of SZ. We performed bioinformatics analysis of genes related to each uncovered relationship and their molecular consequences. Then, we computationally and clinically evaluated the genotypic-phenotypic relations to determine sub-classes of the disease based on whether the groups of SZ patients varied on a range of positive and/or negative symptoms.

Method

Given a genotype database from a GWAS represented as a matrix [SNPs x subjects], the method for dissecting the architecture of a disease is composed of 6 steps (Figure S1), where a SNP set is a sub-matrix (16) harboring subjects described by a set of SNPs sharing similar allele values (16, 28):

1) *Identify SNP sets (Implemented in our PGMRA web server* (16)*).* Use a Generalized Factorization Method (GFM) to dissect a GWAS into SNP sets (16, 28) (see below for a mathematical description of NMF). The GFM applies recurrently a basic factorization method to generate multiple matrix partitions using various initializations with different maximum numbers of sub-matrices k (e.g., $2 \leq k \leq \sqrt{\mathrm{n}}$), where n is the number of subjects, and thus, avoids any pre-assumption about the ideal number of sub-matrices (see below for a rationale about the use of unconstrained number of sub-matrices or clusters (16)). Particularly, we developed a new version of the basic bioNMF method (29) termed Fuzzy Nonnegative Matrix Factorization method (FNMF) (16), and used it as a default basic factorization method. FNMF allows overlapping among sub-matrices, and detection of outliers (16). For each run of the basic factorization method ($2 \leq k \leq \sqrt{\text{n}}$), all sub-matrices are selected to compose a family of genotypic SNP sets *G_k* ={G_*k_i*}, where $1 \leq i \leq k$. Each *G_k* family, as well as all families together *G* ={G_*k*} for all *k*, may include overlapped, partially redundant and different-size sub-matrices.

2) ***Perform a statistical analysis of SNP sets (Accessed via our PGMRA web server* (16)*)***. Use the R-project package SKAT (22, 28) to evaluate the significance of each SNP set. We used the identity-by-state (IBS) as a kernel because the analyzed variants are not rare but common, and therefore, using the "weighted IBS" kernel would not be adequate (22, 28). Since the SNP sets can overlap, we run each one separately. The sex and ancestry of the subjects were used as covariates, and the default remaining parameters were utilized.

3) ***Map a disease risk function***

3.1) *Estimate the risk of a SNP set.* Incorporate *a posteriori* the status of the subjects in a weighted average of epidemiological risks function of all subjects in a particular SNP set:

$$Risk(\text{G}\_k\_i\ ) = \frac{\sum_{i\in ST}|ST_i|Q_i}{\sum_{i\in ST}|ST_i|} \qquad (1)$$

with *ST* being the status of the instances (i.e., cases and controls) and *Q* the weights given by epidemiologic risk of SZ in each SNP set (e.g., 0 and 1 for controls and cases; 0.01, 0.1 and 1 for cases, relatives and controls, respectively) (16).

3.2) *Plot the genotype risk surface of the disease*. Encode each SNP set into a 3-tuple (X,Y,Z), where SNP sets are placed along the *x- and y-axis* using a dendrogram based on their distances in the SNP (see step 4.1, M_{SNPs}) and subject (see step 4.2, $M_{subjects}$) domains, respectively, and Z is the risk variable calculated in (eqn. 1). Interpolate and plot the surface by using the *tgp* and *latticeExtra* packages in R-project, respectively.

4) ***Discover and encode relations among SNP sets into topologically organized networks***

4.1) *Identify optimal and non-redundant relations between SNP sets based on their shared SNPs and, separately, based on their shared subjects*. Overlap of SNP sets refers to overlap of SNP loci, which, in most of our cases leads also to sharing allele values. The sharing of alleles is fully true when there is overlap of both loci and subjects.

4.1.1) Co-cluster all G_*k*_*i* SNP sets within *G* by calculating the pairwise probability of intersection among them using the Hypergeometric statistics (11, 27) (PI_{hyp}) on intersected SNPs: PI_{hyp} (G_*e*_*q*, G_*r*_*w*) (eqn. 2, see below), where *q* and *w* are SNP sets generated in runs with a maximum of *e* and *r* number of sub-matrices, respectively, and *p* in (eqn. 2) is the intersection of SNPs. Then, encode all PI_{hyp}-values, which encompass –in some extent— the distance between SNP sets, in a square [SNP set x SNP set] matrix M_{SNPs}.

4.1.2) Repeat the former procedure based on intersected subjects and determine the $M_{subjects}$ matrix.

4.1.3) Eliminate highly overlapped/redundant SNP sets, which may occur due to the repetitive application of the factorization methods, by deleting all except one SNP set where $Max(M_{SNPs}[i, j], M_{subjects}[i, j]) \leq \delta$, for all *i*, *j* indices in the matrices. Here, we used $\delta = 10E\text{-}15$.

4.2) *Organize SNP sets sharing SNPs and/or subjects into subnetworks.*

4.2.1) For each row *i* and column *j* in M_{SNPs}, $M_{SNPs}[i, j] \leq \varphi$, connect the corresponding SNP sets with a blue line, indicating that they share SNPs. In our case, we established $\varphi \leq 3E - 09$. This value results from adjusting typical p-value of 0.01 by the total number of pairwise comparisons between all possible generated SNP sets [4094 X 4094, by using the Hypergeometric-based test (eqn. 2)], likewise a Bonferroni correction (30).

4.2.2) For each row *i* and column *j* in M_{SNPs}, $M_{subjects}[i, j] \leq \varphi$, connect the corresponding SNP sets with a red line, indicating that they share subjects.

5) ***Identify genotype-phenotype latent architectures***

5.1) *Create a phenotype database.* Dissect the questionnaire based on DIGS and the Best Estimate Diagnosis into individual variables (see below, Data Reduction and Appendix I, catalog of phenotypic features). The variables can be numerical or categorical. For efficiency, in our case, each categorical variable was re-coded into different variables with binary values. The phenotype data was codified in a [phenotype features x subjects] matrix, where the columns and rows correspond to subjects and phenotypic features, respectively. In our case, because the phenotypic features from cases are different from those from the controls, we only considered the cases.

5.2) *Identify phenotype sets (Implemented in the PGMRA web server* (16)*).* Use step 1) with the phenotype database from 5.1) –instead of genotype database— to identify phenotypic sets, where a phenotypic set is a sub-matrix (16) harboring subjects described by a set of phenotypic features sharing similar values (i.e., *P_h_j*, where *j* is a phenotypic set generated in a run with a maximum of *h* number of sub-matrices).

5.3) *Identify genotypic-phenotypic relations***.** Co-cluster SNP sets with phenotype sets into relations using the Hypergeometric statistics on intersected subjects, where $R_{i,j}$ = $\mathrm{PI}_{\mathrm{hyp}}$ (G_k_i, P_h_j) (see below, eqn. 2), G_k_i and P_h_j are SNP and phenotypic sets, respectively, and p in (see below, eqn. 2) is the intersection of subjects. Relations $R_{i,j} < T$ constitute the genotypic-phenotypic architecture of a disease. The significance of the relations (T) was established by the p-value ($\mathrm{PI}_{\mathrm{hyp}}$) provided by the Hypergeometric-based test (see below, eqn. 2) (11, 27).

6) Annotate genes, and symptoms/classes of disease.

6.1) *Map latent architectures to the genome.* For each SNP set, we analyze all genes being affected by each of the SNPs in a SNP set. This analysis includes the SNP location with respect to a gene, the type and number of genes being affected by one SNP (e.g., protein coding, ncRNA genes, and pseudogenes), the possible transcripts being affected and the position where they are affected (e.g. coding region, distance to stop codon, splicing site, intron, UTR, ect.), and finally promoter and intergenic regions' features are inspected for annotation if the SNP does not overlap with a gene then regulatory. Moreover the possible molecular consequences of each SNP over function is provided, as well as, the corresponding allele values. Annotation information was obtained from the Haploreg DB (31) and from the Ensembl (www.ensembl.org) and NCBI (www.ncbi.nlm.nih.gov) web services (see below).

Once we obtain the information described above, we generate a list of relevant genes that it is used to query the Nextbio web site (32) in order to find diseases related to each gene. NextBio uses proprietary algorithms to calculate and rank the diseases and drugs most significantly correlated with a queried gene, where rank values are established relative to the top-scored result (score set to 100). Therefore, although a low-scoring result might have less statistical significance compared to the top-ranked result, it could still have real biological relevance. In our case, out of all possible diseases, only the categories "Mental Disorders" and "Brain and Nervous System Disorders" were considered from the "Disease Atlas".

6.2) *Map latent architectures to disease symptoms or classes of disease.*

6.2.1) Characterize each phenotypic feature by the type of symptoms that they represent. First, explore the distribution of the phenotypic dataset by calculating the principal components (PCA, Statistic Toolbox, Matlab R2011a) of the Phenotypic sample, where the columns are subjects and the rows are the phenotypic variables. Here we used as many PCs as needed to account for the 75% of the sample (5 PCs). In the sample with the phenotypic features as rows and the PCs as columns, cluster the rows by using Hierarchical Clustering (Correlation and Maximum as inter and intra-clustering measurements, Statistic Toolbox, Matlab R2011a). This clustering process generates natural groups of features constitution natural partition hypotheses about the phenotypic features (Figure S5). Second, evaluate each phenotypic feature included in the phenotype

database using curated information from experts[1] and the literature (33, 34) and individually classify each item based on the symptoms as purely positive (1), purely negative (4), primarily positive (2) or primarily negative symptoms (3).
6.2.2) For each phenotypic set *P_h_j* related to a SNP set *G_k_i* in $R_{i,j}$
re-code each phenotypic feature by their positive and/or negative symptoms in a [$R_{i,j}$ X phenotypic feature] matrix $M_{symptoms}$.
6.2.3) Cluster the encoded features by factorizing $M_{symptoms}$ into sub-matrices using a basic factorization method with a maximum number of sub-matrices defined by the Cophenetic index (35).
6.2.4) Label the latent classes of the diseases. (Our current results provided 8 classes, see Figure 3B.)

Mathematical description of NMF

We consider a GWA data set consisting of a collection of N_M subject samples (e.g., cases and controls), which we use to characterize a domain of genotypic (SNPs) states of interest. The data are represented as an n_M x N_M matrix *M*, whose rows contain the allele values of the n_M SNPs in the N_M subject samples. Using the FNMF, we find a manageable number of SNP sets *k*, positive local and linear combinations of the N_M subjects and the n_M SNPs, which can be used to distinguish the genetic profiles of the subtypes contained in the data set. Mathematically, this corresponds to finding an approximate factoring, $M \sim W_M$ x H_M, where both factors have only positive entries and hence are biologically meaningful (12-14, 16). W_M is an n_M x *k* matrix that defines the SNP set decomposition model whose columns specify how much each of the subjects contributes to each of the *k* SNP set. H_M is a *k* x N_M matrix whose entries

[1] Thought disorder is the most ambiguous to classify. Some thought disorder clearly belongs to positive symptoms (e.g., delusions of persecution or reference) while some, such as thought withdrawal, echoing, or broadcasting (labeled as "bizarre" in DSM), can appear in multiple SZ presentations, including negative and disorganized SZ. Based on our clinical experience, we classified these "bizarre" thought symptoms as observable in either positive or negative SZ, although we found them to be more prevalent in negative and disorganized SZ. In our opinion, these bizarre symptoms are most likely to reflect a disorder of integrative thought processes in self-aware consciousness that results in a blurring of the distinction between internal thoughts and external reality. Consequently we allowed for the possibility that such bizarre phenomena can observed in both positive and negative forms of SZ, even though we expected it would be rare or absent in some paranoid subtypes.

represent the SNP allele values of the *k* SNP sets for each of the N_M subject samples. In our implementation either a subject or SNP can belong to more than one SNP set (10, 11, 16).

Rationale for the Use of Unconstrained Number of Clusters

Although there are many indices that estimate the appropriate number of clusters for a given partition, we previously demonstrated that they are often constrained by the type of cluster, and metrics utilized (11, 36). Therefore, it is hard to obtain a consensus from all of them, and they very often provide contradictory results. Moreover, given that the target of the method is to obtain good relations among clusters from different domains of knowledge, it is not known which cluster in one domain will match another cluster in a different domain, and thus, the more varied the clusters, the better the chance of identifying posterior inter-domain relations (11, 36, 37). To do so, we repeatedly applied a basic clustering method in one domain of knowledge to generate multiple clustering results using various numbers of clusters initializations (from 2 to $\sqrt{n}$, where n is the number of observations/subjects).

Coincident Test Index: Co-clustering and Establishing Relations Between Sets

The degree of overlapping between two SNP or phenotypic sets was assessed by calculating the pairwise probability of intersection among them based on the Hypergeometric distribution (11, 27) (PI_{hyp}, Figure S2):

$$PI_{hyp}(P_i, G_j) = 1 - \sum_{q=0}^{p-1} \binom{h}{q}\binom{g-h}{n-q} \Big/ \binom{g}{h} \qquad \begin{array}{c} h = |P_i| \\ n = |G_j| \\ p = P_i \cap G_j \end{array} \tag{2}$$

where *p* observations belong to a set P_i of size *h*, and also belong to a set G_j of size *n*; and *g* is the total number of observations. Therefore, the lower the $PI_{hyp,}$ the higher the overlapping (17). The (p-value of) hypergeometric "test" is used here as a measure of association strength. The real test (p-value) of genotypic-phenotypic relationship was provided through the permutation procedure.

Permutation Test for Genotypic-Phenotypic Relations

Statistical significance reported values were obtained by 4000 independent permutations due to the comparisons between all possible generated SNP sets (i.e., 4094, from 2 to $\sqrt{n}$), and possible overlapped SNP sets here identified were generated as following (16, 38): a) assign random subjects to a phenotypic cluster of random size; b) assign random subjects to a genotype cluster (set) of random size; c) calculate the Hypergeometric statistic (PI_{hyp}, (27) eqn 2) between the two clusters and accumulate the value. These values form an empirical null distribution of PI_{hyp} used to calculate the empirical p-value of an identified relation. All optimal relations had empirical p-value ≤ value < 4.7E-03.

Resampling Statistics of the NMF Sets

To guarantee the submatrices converge to the same solution and, given the non-deterministic nature of NMF and its dependence on the initialization of the *W* and *H* vectors, we run it 40 times for any *k* maximum number of allowed submatrices with different random initializations of the vectors to select those that that best approximates the input matrix (12). Besides, to estimate the precision of sample statistics of the SNP sets (variance of the *W* and *H* vectors) we use a leave-one-out technique (jackknifing) 1000 times on the SNP domain and obtained a 94% support for all identified sets with an average variance of c.a. ±5% of their corresponding *W* and *H* vectors (29). Finally, we already modified this sampling technique to ensure the occurrence of the remaining sets after a leave-one-set-out (7) and applied to our current sample with >90% of support.

Data Reduction

Data reduction was not applied because many Principal Components (PCs) were required in this study (data not shown), consistent with the demonstration that clustering with the PCs instead of the original variables does not necessarily improve, and often degrades, cluster quality and interpretability (39). Moreover, likewise in phenomics (40), partially correlated variables reinforce the association and clarify the symptom identification process. Therefore, we used initially 93 phenotypic features listed in Appendix I, catalog of phenotypic features.

Briefly, phenotypic features used in the search process included all available data from the interviews. That is, replies to DIGS as well as to the Best Estimate Diagnosis code sheet submitted by GAIN/NONGAIN to dbGaP. Unbiased compilation of all of the data resulted in an initial set of 93 features. To capture items specific for positive and negative schizophrenia and avoid symptoms with affective elements, symptoms reported by acutely psychotic patients, and redundant items the original set of was pruned based on authors clinical experience, and computational feature validation (Table S6 and above in Method, step 6.2.1).

Bioinformatics Analysis: Genotypic Organization of the SZ Architecture Accounts for Multiple Genetic Sources of the Disease

Given that genotypic SZ architecture is composed of multiple networks, we matched each SNP set composing these networks with the corresponding genomic location of their SNPs, and in turn, with the mapped genes (Figure 3A, Table S1) to investigate what these SNP sets represent in terms of genomic information. We uncovered a list of genes with many different functions and distinct roles in different molecular networks (Table S1-S3).

***A single SNP Set Can Map Different Classes of Genes, Located in Different Chromosomes, and Distinct Types of Genetic Variants*.**

The uncovered SNP sets contain SNPs that map gene, promoter and intergenic regions (IGRs) located anywhere in the genome, without being constrained by genomic features such as a specific gene or haplotype (28). For example, SNP set 81_13 contains SNPs in chromosomes 8 and 16, whereas SNP set 42_37 has SNPs located in chromosomes 2 and 11 (Figure 3A, Table S1). SNP set 75_67 has SNPs in chromosomes 4, 8, 15, and 16, among others, and maps >30 genes, as expected by its generality (Figure 3A, Table S1). The latter SNP set is in the same network as SNP sets 56_30, 76_74 and 81_13, and thus shares some genes with them. Despite being in the same network, the last three SNP sets map to particular genes specific to each of them (Figure 3A, Table S1).

In addition to mapping genes in different locations, SNP variants within the SNP sets affect distinct classes of genes including protein-coding, non-coding (ncRNA) genes, and pseudogenes, with different molecular consequences depending on the altered region (coding, UTRs, introns, Table S3). For example, only 25% of SNPs in SNP set 75_67 affect protein-coding genes, which are the targets most often considered in genetic studies of diseases, whereas another 25% of SNPs affect ncRNAs (lincRNAs, antisense RNAs, miRNAs). One of these lincRNAs is SOX2-OT, which is associated with > 15 possible transcripts (Table S3); it is contained inside the SOX2 transcription factor that is predominantly expressed in the human brain where SOX2-OT is also highly enriched (41, 42).

Likewise, SNPs from SNP set 22_11 are located within a large intergenic region corresponding to two overlapping and newly characterized long ncRNAs AC068490.2 and AC096570.2 (Table S3). Moreover, two SNP variants of SNP set G19_2 affect miRNA AL354928.1 and small nuclear RNA U4, as well as protein-coding GOLGA1 gene (Figure S6A, Table S3). Finally, the SNP sets can map to large genomic regions. That is the case with all SNPs in SNP set 22_11 (with risk of 73 %), and a few in SNP set 81_13 (with risk of 95%), which correspond to two different structural CNVs already annotated (www.ensembl.org). These results point to accumulation of possible regulatory alterations of gene expression pattern in these groups (Table S3), which suggests an underlying complex and dynamic architecture of molecular processes that influence vulnerability to distinct forms of SZ.

Bioinformatics Analysis of the SNP Set-Related Genes Reveals Disparate Molecular Consequences.

A detailed analysis of SNPs and mapped genes revealed at least three complex scenarios affecting multiple genes in different fashions (activation, repression, antisense modulation) and producing different molecular consequences (Table S3). First, we determined that even *a single SNP within a SNP set could produce different consequences in affected transcripts* (Table S3). For example, one SNP from SNP set 81_13 was located in a protein-coding region of the SNTG1 gene, which can produce either a change in an intron or in a transcript affecting nonsense-mediated protein decay that would be eliminated by a surveillance pathway containing a premature stop codon (Table S3). Second, we found that *multiple SNPs within a SNP set can affect multiple genes in different ways*. This heterogeneity is exemplified by SNPs from SNP set 19_2 intersecting with both ncRNAs and the GOLGA1 gene (Figure 4a). Third, we uncovered that *multiple SNPs within different SNP sets can distinctively affect single genes.* For example, SNP sets 71_55 and 14_6 are located in different networks since they have neither SNPs nor subjects in common (Figure 3 and Figure S5). Yet, all SNPs within both SNP sets are located in the same NTRK3 gene, which influences hippocampal function, but at different locations (Figure 6B), which thereby may modify risk for SZ differentially (43). Consequently it is not surprising that each SNP set is observed in different individuals with distinct phenotypic consequences. Overall, since a single SNP can affect multiple gene transcripts, or multiple SNP sets may influence a single gene transcript, we must consider the specific transcription pathway in order to understand antecedent mechanisms that result in equifinality and multifinality.

Genes Mapped by SNP Sets at Risk Correlate with Different Aspects of Neurodevelopment.

Most genes mapped by the SNP sets are involved in neurodevelopment (Table S2). For example, the SNP set 81_13 (Figure 3A) maps to SNTG1, PXDNL, and GP2 genes (Table S1). SNTG1 is a syntrophin that mediates dystrophin binding in brain specifically. It is down-regulated in neurodevelopmental disorders, sleep disorders, and dementia (Table S2). PXDNL encodes a peroxidasin-like protein, which affects risk of SZ and dementia (Table S2). GP2 encodes glycoprotein 2 (zymogen granule membrane) and is down-regulated in neuropathy and basal ganglia disorders, but up-regulated in Alzheimer's disease

(Table S2). Cumulatively, characterization of all genes in terms of related diseases supports the biological impact of these SNP sets.

Pathways

We identified distinct pathways (see Tables S1 and S4, and Figure S7) including genes that have already been reported as associated with SZ by GWAS, as well as genes known to be abnormally expressed in the brain of SZ patients. Overall, the products of genes uncovered by the SNP sets are included in several well-known, relevant and interconnected signaling pathways. Annotation information was manually curated and obtained from the Haploreg DB (31) and from the Ensembl (www.ensembl.org) and NCBI (www.ncbi.nlm.nih.gov) web services.

PI3K / Akt Signaling. Akt is a Serine/threonine Kinase, it is activated by tyrosine kinase receptors, integrins, T and B cell receptors, cytokine receptors, G-proteins-coupled receptors and other stimuli that involves the production of PIP3 triphosphate (phosphatidylinositol triphosphate) by PI3K (phosphoinositide 3 kinase).

PI3K can be activated by different ways:

•FOXR2 (forkhead box R2) is a proto-oncogene when it is mutated, maintained cell growth and proliferation through activation of RAS (GTPase) increase aberrant signaling through pathways PI3K/AKT/mTOR and RAS/MAP/ERK, inhibiting apoptosis (44-46).

•SOD3 (superoxide dismutase 3) causes increased of phosphorylation of ERK / Ras and PIP3 because PI3K, SOD3 may be Phosphorilated by Erk1/2 (47-49).

•Sema3A inhibits the proliferation and cell growth in neurons and prevents axonal growth by inhibiting the PI3K/Akt via inhibition of Ras. Neuropilin and SEMA1 bound active apoptosis via PI3K/Akt (50-53).

•RAS (GTPase) can be activated by FOXR2 mutated by SOD3 and inhibited by Sema3A. Ras and PI3K can activate mTORC1 by cRaf/MEK/ERK (44, 49).

•SNX19 inhibits Akt phosphorylation resulting in apoptosis (54, 55).

•STYK1 oncogene that binds to Akt to activate the cascade signaling downstream and leading to increased tumor cells and increasing the risk of metastasis (56).

•CHST9 catalyzes the sulfates transfer to N -acetylgalactosamine residues, inhibits Cd19/p85/PI3K-p110 complex (57, 58).

•RRAGB is part of RAG proteins that interact with mTORC1 family and are required for activation of amino acids via mTORC1(59).

Signaling Pathways Activating MAPK/p38/p53. p38 MAPKs (α, β, γ, and δ) are members of the MAPK family that are activated by a variety of environmental stresses and inflammatory cytokines. As with other MAPK cascades, the membrane-proximal component is a MAPKKK, typically a MEKK or a mixed lineage kinase (MLK). The MAPKKK phosphorylates and activates MKK3/6, the p38 MAPK kinases. MKK3/6 can also be activated directly by ASK1, which is stimulated by apoptotic stimuli. p38 MAPK is involved in regulation of HSP27, MAPKAPK-2 (MK2), MAPKAPK-3 (MK3), and several transcription factors including ATF-2, Stat1, the Max/ Myc complex, MEF-2, Elk-1, and indirectly CREB via activation of MSK1.

This pathway may be activated by activation of PI3K way Rac/MEK/ERK

•DUSP4 is a MKP able of inhibiting p38MAPK 12 and 14a, is regulated by TNF-a expression. Decreases ERK 1/2 and reducing the cellular viability by alteration of the NF-kB / MAPK pathways (60, 61).

•MAGEH1 expression causes apoptosis of melanoma cells through the interaction with the inner region to the membrane of the p75 neurotrophin receptor (p75NTR) one TNF receptor type, and possibly also through competition with the TNF receptor associated factor - 6 (TRAF6) and catalytic neurotrophin receptor (TRK) for the same site of interaction with p75 (62-64).

Nucleus

•TRPS1 The gene encodes for an atypical member of the GATA family. It can activate Snail 1 to produce inhibition of cadherines inside of nucleus (65, 66).

•ST18 is a promoter of hypermethylation, ST18 loss of expression in tumor cells suggests that this epigenetic mechanism responsible for the specific down - regulation of tumor (67, 68).

•SPATA7 may be involved in the preparation of chromatin in early meiotic prophase in the nuclei for the initiation of meiotic recombination (69, 70).

•ZC3H14 a protein with zinc finger Cys3His evolutionarily conserved that specifically binds to RNA and polyadenosine therefore postulated to modulate post-transcriptional gene expression (71).

•U4, is part of snRNP small nucleolar ribonucleic particles (RNA –protein), each one bind specifically to individual RNA . The function of the human U4 3'SL micro RNA is unclear. It exists to enable the formation of nucleoplasm in Cajal bodies (72).

•PPP1R1C (Protein phosphatase 1, regulatory subunit 1C) is a protein-coding gene and inhibitor of PP1, and is itself regulated by phosphorylation. It promotes cell growth and may protect against cell death, particularly when induced by pathological stress (73, 74).

•PRPF31 main function is thought to recruit and strap for U4/U6 U5 tri – snRNP (75, 76).
•EVI5 works in G1/S phases, prevents phosphorylation of Emi 1 by Plk1and therefore inactive APC/C and accumulates cyclin A. In prometaphase, Plk1 phosphorylates to EVI5, producing its inactivation and subsequent activation of APC/C and downstream signaling pathways to complete the mitotic cycle (77-79).
•SNORA42: The main functions of snoRNAs has long been thought to modify, mature and stabilize rRNAs . These posttranslational modifications - transcriptional are important for production of accurate and efficient ribosome. Moreover, some snoRNAs are processed to produce small RNAs (80, 81).
•SNORD112. SnoRNAs act as small nucleolar ribonucleoproteins (snoRNPs), each of which consists of a C / D box or box H / ACA RNA guide, and four C / D and H / ACA snoRNP associated proteins. In both cases, snoRNAs specifically hybridize to the complementary sequence in the RNA, and protein complexes associated then perform the appropriate modification to the nucleotide that is identified by the snoRNAs (80, 81).
•SMARCAD1 contributes as part of a large complex with HDAC1, HDAC2 , and KAP1 G9A to integrate with nucleosome spacing and histone deacetylation. H3K9 methylation is required for heterochromatin restore apparently facilitates histone deacetylation and H3K9me3 . How chromatin remodeling is done by deacetylation is unknown, but it seems to coordinate spacing between nucleosomes with H3K9 acetylation and monomethylation (82).

Mitochondria
•SLC25A14 uncoupling protein that facilitates the transfer of anions from the inside of the mitochondria to the outer mitochondrial membrane and the return transfer of protons from the outside to the inner mitochondrial membrane. SLC25A14 functional role in cellular energy supply and the production of superoxide after it overexpressed in neuronal cells. In untreated culture conditions, overexpression of MMP and SLC25A14 significantly decreased content of intracellular ATP (83-85).
•TMEM135, some studies have demonstrated TMEM135 association with mitochondrial's fat metabolism, and a possible role for TMEM135 recently identified in improving fat storage (86).
•VDAC3 selective Anions voltage-dependent channels (VDACs) are proteins that form pores allowing permeability of the mitochondrial outer membrane. A growing body of evidence indicates that VDAC plays a major role in metabolite flow in and out of mitochondria, resulting in regulation of mitochondrial functions (87).

Membrane

•SLC20A2 the proteins of this group transport stream comprises an initial joining of a Na + ion, followed by a random interaction between Pi (inorganic phosphorus) monovalent and second ion Na + . Reorientation loaded carrier, then leads to the release substrate in the cytosol (88-90).

•NALCN encoding a voltage-independent, cationic, non-selective, non-inactivating, permeable to sodium, potassium and calcium channel when expressed exogenously in HEK293 cells. Sodium is important for neuronal excitability in vivo, the NALCN channel seems to be the main source of sodium leak in hippocampal neurons and because these two processes are strongly altered in schizophrenia is the hypothesis had to
NALCN could show a genetic association with schizophrenia (91).

•HACE1 is a tumor suppressor, catalyses poly - Rac1 ubiquitylation at lysine 147 upon activation by HGF, resulting in its proteasomal degradation. HACE1 controls NADPH oxidase. HACE1 promotes increased binding to Rac1 regulating the NADPH oxidase, decrease the production of oxygen free radicals , and inhibit the expression of cyclin D1 and decrease susceptibility to damage DNA. HACE1 loss leads to overactive NADPH oxidase, increased ROS generation, also the expression of cyclin D1 and DNA damage induced by ROS (92-95).

•NCAM1 is a constitutive molecule expressed on the surface of various cells, promotes neurite outgrowth, nerve branching, fasciculation and cell migration (96, 97).

•OPN5 apparent gabaergic interaction in Synaptic space (98, 99).

•NETO2 is an auxiliary subunit determines the functional propiedadde KARS proteins (kainate, a subfamily of ionotropic glutamate receptors –iGluRs-) that mediate excitatory synaptic transmission, regulate the release of neurotransmitters and in selective distribution in brain (100, 101).

•VANGL1 This gene encodes a member of the family tretraspanin. Mutations in this gene are associated with neural tube defects . Alternative splicing results in multiple transcript variants (102, 103).

•DKK4 is a DKK to block the expression of LRP and thus union with the complex Frizzled and Wnt / SFRP / WIF blocking the release of b –catenin (104).

•NTRK3 is a member of the family of neurotrophin receptors and is critical for the development of the nervous system. Published studies suggested that NTRK3 is a dependence receptor , which signals both the ligand -bound state ("on") and the free ligand ("off") state (see chart) . When present the ligand neurotrophin -3 (NT-3), NTRK3 trigger signals within the cell via a tyrosine kinase domain in promoting cell proliferation and survival. In the absence of NT-

3, NTRK3 signals for cell death by triggering apoptosis. Therefore, NTRK3 have the potential to be an oncogene or tumor suppressor gene function of the presence of NT-3 (43, 105-108).

Reticular Endoplasmic Reticulum

- PSMC1 is involved in the destruction of the protein in bulk at a fast or slow rate in a wide variety of biological processes such as cell cycle progression, apoptosis, regulation of metabolism, signal transduction, and antigen processing (109).
- PTBP2 Ptbp1 and Ptbp2 regulate the alternative splicing of various RNA target assemblies, suggesting that the roles of Ptbp1 / 2 proteins are different in different cellular contexts. Ptbp2 functions in the brain are not clear (110, 111).
- RyR3s is a type of ion channel that intracellular free Ca2 + when opened from the endoplasmic reticulum (ER). It is very similar to the inositol triphosphate receptor (inositol - 1, 4,5- triphosphate) IP3R. The main signal to trigger the opening of RyRs are Ca2 + has usually entered through voltage-dependent channels of cell membrane. RyR3 is expressed in several cell types including the brain in small quantities, RyR3 deficient mice have impaired hippocampal synaptic plasticity and impaired learning. ATP also stimulates the activity of the channels RyR3. The therapeutic targets focus on molecules that induce release control, internalization and calcium mobilization (112, 113).
- RPL35 is a protein binding to the signal recognition particle (SPR) and its receptor (SR). They mediate targeting complexes nascent chain - ribosome to the endoplasmic reticulum (114).
- RPL5 is an MDM2 binding protein (MDM2 oncogene, protein E3 ubiquitin ligase) and SRSF1 (serine / rich splicing factor arginine 1) to stabilize p53 oncogene and to induce cell senescence. RPL can join RPL11 and other ribosomal proteins to silence Hdm2 and p53 (114, 115).
- FAM69A calico dependent kinase, extracellular and intracellular, localized in the endoplasmic reticulum (78, 116)

Other Organelles

- GOLGA1 is part transport proteins of the Golgi apparatus, which participates in glycosylation and transport of proteins and lipids in the secretory pathway (117, 118).
- EML5 blocks EMAP via MAP or stabilization of microtubules (74).
- ARPC5L component can function as Arp2 / 3 complex which is involved in the regulation of actin polymerization and together with the activation of factor inducing nucleation (NPF) mediates the formation of branched networks of actin. It belongs to the family Arpc5 (119).

•CSMD1 in the TGF-B pathway, CSMD1 permits the TGF-B receptor I junction, allowing it to phosphorylate Smad3 and thus allow complex formation: phosphorylated Smad3 / phosphorylated Smad2 / Smad4; the complex is internalized into the cellular nucleus and bound to a transforming factor leads to apoptosis. In addition, the TGF-B receptor II binds the phosphorylated complex, allowing for subsequent binding Smad1/5/8 with Smad4, and nuclear internalizing inducing apoptosis mediated by binding to a transforming factor (111, 120).

Replicability

Replicability of the Phenotypic Features: Psychiatric Assessment in International Schizophrenia Consortium (ISC), and the Psychiatric Genetics consortium for schizophrenia (PGC-SCZ)

Unfortunately, there are no large studies except MGS that used DIGS for the assessment of SZ and the other large study (ISC) used a wide variety of instruments. There has been little attention to the need for consistency in obtaining a detailed and rich phenotypic description of SZ, with meta-analysis based on reducing assessment to the dichotomy of SZ or not based on DSM or ICD diagnoses (1, 24, 121-124). The supplementary material in (122) is the main description of the ISC assessment data. The ISC is composed of the following samples using different diagnostic procedures: Aberdeen (723 cases, clinical interview is SCID, diagnosis by OPCRIT, which is a checklist completed by clinician), Cardiff (cases from Bulgaria diagnosed by "about 50" psychiatrists in 5 different hospitals using a translation of SCAN – schedule for clinical assessment in neuropsychiatry for diagnosis of SZ by DSM-IV), Dublin (no structured interview, diagnosis by Operational Criteria Checklist OPCRIT by clinicians, using DSM-IV criteria), Edinburgh (assessed by SADS-L supplemented by case notes and collateral information, diagnosis based on DSM-IV criteria), London (diagnosis by ICD10 diagnosis of schizophrenia in medical case history and confirmed by SADS-L diagnosis, diagnosed as at least "probably for schizophrenia based on Research Diagnostic Criteria RDC), Portuguese Islands Collection (used DIGS, SANS, and SAPS, and OPCRIT), Sweden (diagnoses were made by treating clinician based on diagnoses of ICD 8,9, or 10, with a DSM-IV checklist review of medical record in a small subsample of 111 to confirm that 95% of cases could be confirmed to meet DSM-IV criteria. In summary, the ISC phenotypic analysis were based on the diagnosis of SZ according to a variety of criteria (ICD 8, 9, or 10; DSM-IV, and RDC) by a variety of structured interviews

(DIGS, SCAN, SADS-L, and RDC) filtered through a variety of checklists (DSM-IV, OPCRIT) or no structured interview at all (based on medical charts by attending psychiatrists) to produce a dichotomous phenotype (SZ or not) for meta-analysis, as detailed in the supplementary material attached to (122).

The PSC-SCZ attempted to evaluate the quality of the phenotypic data that it collected by pooling ISC, MGS, and some other studies (CATIE, Cardiff, 5 SGENE sites and Zucker Hillside hospital) that used variable assessment and diagnostic methods. This is described in the supplementary material (part A. Recruitment and assessment of subjects) of (123). Possible contributors were rated according to an 18-item questionnaire covering the assessment protocol and quality control procedures. Nine of these 18 items were agreed by consensus as key for acceptance into the study: the use of any structured interview, systematic training of interviewers in use of the instrument, systematic quality control of diagnostic accuracy, reliability trials, review of medical record information, best estimate procedure employed, specific inclusion and exclusion criteria developed and utilized, MDs or PhDs as making the final diagnostic determination, and special additional training for the final diagnostics. Each study was scored on this 9-point scale. Most but not all satisfied 7 of the 9 criteria and were judged of high quality, 3 others were accepted anyway, and 1 was excluded due to inadequate quality control of the diagnostic process.

Replicability of Results: The Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) and the Portuguese Islands samples (PIS) from the PGC-SCZ

The CATIE sample was collected as part of the Clinical Antipsychotic Trials of Intervention Effectiveness project, and ascertainment was previously described (24, 125-128). The cases are comprised of 738 (544 males and 194 females) of the 1,460 CATIE participants, who donated a DNA sample from multiple sites in the United States of America (US) of which 402 and 336 had European and African ancestries, respectively. The control sample used for the CATIE GWAS was collected by MGS (123). In the CATIE GWAS, the utilized MGS controls totaled 733 (493 males and 240 females) including European and African ancestries. The array platform was Affymetrix 500K and Perlegen 164K. We considered 44 phenotypic variables from the CATIE study (Table S10), which used variable assessment and diagnostic methods including the Positive and Negative Syndrome Scale, the Quality of Life Questionnaire, and the Structured Clinical Interview for DSM-IV (24, 123, 125-127).

The cases in the PIS lived in Portugal, the Azorean and Madeiran islands, or were the direct (first or second generation) Portuguese immigrant population in the US (24, 128, 129). 346 cases (213 males and 133 females) were used in this human subjects protocol approved by State University of New York – Uptown Medical Center, Syracuse, New York. The controls were not related to cases, with the exception of 3 controls that married into families but were not biologically related to cases. The control sample used in this analysis was comprised of 215 controls (80 males and 135 females). Like the cases, they also lived in Portugal, the Azorean and Madeiran islands, or were the direct (first or second generation) Portuguese immigrant population in the USA. The array platform was Affymetrix 5.0. Likewise the MGS samples, the Portuguese sample used DIGS for the assessment of SZ (128, 129). Here, we utilized 35 of these features (Table S10).

References

1. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature. 2009;460I:753-7.

2. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010;5I:1564-73.

3. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MaR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81I:559-75.

4. Babcock DF. Development. Smelling the roses? Science. 2003;299I:1993-4.

5. Spehr M, Gisselmann G, Poplawski A, Riffell JA, Wetzel CH, Zimmer RK, et al. Identification of a testicular odorant receptor mediating human sperm chemotaxis. Science. 2003;299I:2054-8.

6. Ozkan SB, Wu GA, Chodera JD, Dill KA. Protein folding by zipping and assembly. Proc Natl Acad Sci U S A. 2007;104I:11987-92.

7. Harari O, Park SY, Huang H, Groisman EA, Zwir I. Defining the plasticity of transcription factor binding sites by Deconstructing DNA consensus sequences: the PhoP-binding sites among gamma/enterobacteria. PLoS Comput Biol. 2010;6I:e1000862.

8. Peppercorn J, Perou CM, Carey LA. Molecular subtypes in breast cancer evaluation and management: divide and conquer. Cancer Invest. 2008;26I:1-10.

9. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. Nature. 2000;407I:651-4.

10. Zwir I, Shin D, Kato A, Nishino K, Latifi T, Solomon F, et al. Dissecting the PhoP regulatory network of Escherichia coli and Salmonella enterica. Proc Natl Acad Sci U S A. 2005;102I:2862-7.

11. Zwir I, Huang H, Groisman EA. Analysis of Differentially-Regulated Genes within a Regulatory Network by GPS Genome Navigation. Bioinformatics. 2005;21I:4073-83.

12. Pascual-Montano A, Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Marqui RD. bioNMF: a versatile tool for non-negative matrix factorization in biology. BMC Bioinformatics. 2006;7I:366.

13. Tamayo P, Scanfeld D, Ebert BL, Gillette MA, Roberts CW, Mesirov JP. Metagene projection for cross-platform, cross-species characterization of global transcriptional states. Proc Natl Acad Sci U S A. 2007;104I:5959-64.

14. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401I:788-91.

15. Zhou T, Kuscsik Z, Liu JG, Medo M, Wakeling JR, Zhang YC. Solving the apparent diversity-accuracy dilemma of recommender systems. Proc Natl Acad Sci U S A. 2010;107I:4511-5.

16. Arnedo J, del Val C, de Erausquin GA, Romero-Zaliz R, Svrakic D, Cloninger CR, et al. PGMRA : a web server for (phenotype x genotype) many-to-many relation analysis in GWAS. Nucleic Acid Research. 2013;75.

17. Beyene J, Tritchler D. Multivariate analysis of complex gene expression and clinical phenotypes with genetic marker data. Genetic Epidemiology. 2007;31I:S103-S9.

18. Ciliberti S, Martin OC, Wagner A. Innovation and robustness in complex regulatory gene networks. Proc Natl Acad Sci U S A. 2007;104I:13591-6.

19. Wagner A. Robustness against mutations in genetic networks of yeast. Nat Genet. 2000;24I:355-61.

20. Wuchty S. Scale-free behavior in protein domain networks. Mol Biol Evol. 2001;18I:1694-702.

21. Albert R. Scale-free networks in cell biology. J Cell Sci. 2005;118I:4947-57.

22. Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of …. 2011I:82-93.

23. Lin X, Cai T, Wu MC, Zhou Q, Liu G, Christiani DC, et al. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. Genet Epidemiol. 2011;35I:620-31.

24. Lee SH, DeCandia TR, Ripke S, Yang J, Sullivan PF, Goddard ME, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nature Genetics. 2012;44I:247-U35.

25. Wu MC, Maity A, Lee S, Simmons EM, Harmon QE, Lin X, et al. Kernel Machine SNP-Set Testing Under Multiple Candidate Kernels. Genet Epidemiol. 2013;37I:267-75.

26. Bezdek JC. Pattern Analysis. In: Pedrycz W, Bonissone PP, Ruspini EH, editors. Handbook of Fuzzy Computation. Bristol: Institute of Physics; 1998. p. F6.1.-F6..20.

27. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999;22I:281-5.

28. Wu M, Kraft P, Epstein M. Powerful SNP-set analysis for case-control genome-wide association studies. The American Journal of …. 2010I:929-42.

29. Schachtner R, Lutter D, Knollmuller P, Tome AM, Theis FJ, Schmitz G, et al. Knowledge-based gene expression classification via matrix factorization. Bioinformatics. 2008;24I:1688-97.

30. Dunn OJ. Multiple Comparisons Among Means. Journal of the American Statistical Association. 1961;56I:52-64.

31. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40I:D930-4.

32. Kupershmidt I, Su QJ, Grewal A, Sundaresh S, Halperin I, Flynn J, et al. Ontology-based meta-analysis of global collections of high-throughput public data. PLoS One. 2010;5I:epublish.

33. American_Psychiatric_Association. Diagnostic and Statistical Manual of Mental Disorders. Washington, DC: American Psychiatric Publishing Inc.; 1994.

34. Wray NR, Visscher PM. Narrowing the Boundaries of the Genetic Architecture of Schizophrenia. Schizophrenia Bulletin. 2010;36I:14-23.

35. Sokal R, Rohlf J. The Comparison of Dendrograms by Objective Methods. Taxon. 1962;11.

36. Ruspini EH, Zwir I. Automated generation of qualitative representations of complex objects by hybrid soft-computing methods. In: Pal SK, Pal A, editors. Pattern recognition : from classical to modern approaches. New Jersey.: World Scientific; 2002. p. 454-74.

37. Romero-Zaliz R, Del Val C, Cobb JP, Zwir I. Onto-CC: a web server for identifying Gene Ontology conceptual clusters. Nucleic Acids Res. 2008;36I:W352-7.

38. Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell. 2004;117I:185-98.

39. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. Bioinformatics. 2001;17I:763-74.

40. Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. Nat Rev Genet. 2011;11I:855-66.

41. Amaral PP, Neyt C, Wilkins SJ, Askarian-Amiri ME, Sunkin SM, Perkins AC, et al. Complex architecture and regulated expression of the Sox2ot locus during vertebrate development. RNA. 2009;15I:2013-27.

42. Brennand KJ, Simone A, Jou J, Gelboin-Burkhart C, Tran N, Sangar S, et al. Modelling schizophrenia using human induced pluripotent stem cells. Nature. 2011;473I:221-5.
43. Otnaess MK, Djurovic S, Rimol LM, Kulle B, Kahler AK, Jonsson EG, et al. Evidence for a possible association of neurotrophin receptor (NTRK-3) gene polymorphisms with hippocampal function and schizophrenia. Neurobiol Dis. 2009;34I:518-24.
44. Katoh M. Human FOX gene family (Review). Int J Oncol. 2004;25I:1495-500.
45. Katoh M. Identification and characterization of human FOXK1 gene in silico. Int J Mol Med. 2004;14I:127-32.
46. Katoh M, Igarashi M, Fukuda H, Nakagama H. Cancer genetics and genomics of human FOX family genes. Cancer Lett. 2013;328I:198-206.
47. Laurila JP, Castellone MD, Curcio A, Laatikainen LE, Haaparanta-Solin M, Gronroos TJ, et al. Extracellular superoxide dismutase is a growth regulatory mediator of tissue injury recovery. Mol Ther. 2009;17I:448-54.
48. Miao L, St Clair DK. Regulation of superoxide dismutase genes: implications in disease. Free Radic Biol Med. 2009;47I:344-56.
49. Fattman CL, Schaefer LM, Oury TD. Extracellular superoxide dismutase in biology and medicine. Free Radic Biol Med. 2003;35I:236-56.
50. Eastwood SL, Law AJ, Everall IP, Harrison PJ. The axonal chemorepellant semaphorin 3A is increased in the cerebellum in schizophrenia and may contribute to its synaptic pathology. Mol Psychiatry. 2003;8I:148-55.
51. Tamagnone L, Giordano S. Semaphorin pathways orchestrate osteogenesis. Nat Cell Biol. 2006;8I:545-7.
52. Chacon MR, Fernandez G, Rico B. Focal adhesion kinase functions downstream of Sema3A signaling during axonal remodeling. Mol Cell Neurosci. 2010;44I:30-42.
53. Neufeld G, Kessler O. The semaphorins: versatile regulators of tumour progression and tumour angiogenesis. Nat Rev Cancer. 2008;8I:632-45.
54. Kan A, Ikeda T, Saito T, Yano F, Fukai A, Hojo H, et al. Screening of chondrogenic factors with a real-time fluorescence-monitoring cell line ATDC5-C2ER: identification of sorting nexin 19 as a novel factor. Arthritis Rheum. 2009;60I:3314-23.
55. Harashima SI, Harashima C, Nishimura T, Hu Y, Notkins AL. Overexpression of the autoantigen IA-2 puts beta cells into a pre-apoptotic state: autoantigen-induced, but non-autoimmune-mediated, tissue destruction. Clin Exp Immunol. 2007;150I:49-60.

56. Li J, Wu F, Sheng F, Li YJ, Jin D, Ding X, et al. NOK/STYK1 interacts with GSK-3beta and mediates Ser9 phosphorylation through activated Akt. FEBS Lett. 2012;586I:3787-92.

57. Shimizu K, Okamoto N, Miyake N, Taira K, Sato Y, Matsuda K, et al. Delineation of dermatan 4-O-sulfotransferase 1 deficient Ehlers-Danlos syndrome: observation of two additional patients and comprehensive review of 20 reported patients. Am J Med Genet A. 2011;155AI:1949-58.

58. Zhao X, Wu Q, Fu X, Yu B, Shao Y, Yang H, et al. Examination of copy number variations of CHST9 in multiple types of hematologic malignancies. Cancer Genet Cytogenet. 2010;203I:176-9.

59. Sekiguchi T, Hirose E, Nakashima N, Ii M, Nishimoto T. Novel G proteins, Rag C and Rag D, interact with GTP-binding proteins, Rag A and Rag B. J Biol Chem. 2001;276I:7246-57.

60. Balko JM, Schwarz LJ, Bhola NE, Kurupi R, Owens P, Miller TW, et al. Activation of MAPK pathways due to DUSP4 loss promotes cancer stem cell-like phenotypes in basal-like breast cancer. Cancer Res. 2013;73I:6346-58.

61. Cagnol S, Rivard N. Oncogenic KRAS and BRAF activation of the MEK/ERK signaling pathway promotes expression of dual-specificity phosphatase 4 (DUSP4/MKP2) resulting in nuclear ERK1/2 inhibition. Oncogene. 2013;32I:564-76.

62. Tcherpakov M, Bronfman FC, Conticello SG, Vaskovsky A, Levy Z, Niinobe M, et al. The p75 neurotrophin receptor interacts with multiple MAGE proteins. J Biol Chem. 2002;277I:49101-4.

63. Ojima H, Yoshikawa D, Ino Y, Shimizu H, Miyamoto M, Kokubu A, et al. Establishment of six new human biliary tract carcinoma cell lines and identification of MAGEH1 as a candidate biomarker for predicting the efficacy of gemcitabine treatment. Cancer Sci. 2010;101I:882-8.

64. Selimovic D, Sprenger A, Hannig M, Haikel Y, Hassan M. Apoptosis related protein-1 triggers melanoma cell death via interaction with the juxtamembrane region of p75 neurotrophin receptor. J Cell Mol Med. 2012;16I:349-61.

65. Kobayashi H, Hino M, Inoue T, Nii E, Ikeda K, Son C, et al. GC79/TRPS1 and tumorigenesis in humans. Am J Med Genet A. 2005;134I:341-3.

66. Bonnard C, Strobl AC, Shboul M, Lee H, Merriman B, Nelson SF, et al. Mutations in IRX5 impair craniofacial development and germ cell migration via SDF1. Nat Genet. 2012;44I:709-13.

67. Yang J, Siqueira MF, Behl Y, Alikhani M, Graves DT. The transcription factor ST18 regulates proapoptotic and proinflammatory gene expression in fibroblasts. FASEB J. 2008;22I:3956-67.

68. Shukla R, Upton KR, Munoz-Lopez M, Gerhardt DJ, Fisher ME, Nguyen T, et al. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. Cell. 2013;153I:101-11.

69. Wang H, den Hollander AI, Moayedi Y, Abulimiti A, Li Y, Collin RW, et al. Mutations in SPATA7 cause Leber congenital amaurosis and juvenile retinitis pigmentosa. Am J Hum Genet. 2009;84I:380-7.

70. Perrault I, Hanein S, Gerard X, Delphin N, Fares-Taie L, Gerber S, et al. Spectrum of SPATA7 mutations in Leber congenital amaurosis and delineation of the associated phenotype. Hum Mutat. 2010;31I:E1241-50.

71. Leung SW, Apponi LH, Cornejo OE, Kitchen CM, Valentini SR, Pavlath GK, et al. Splice variants of the human ZC3H14 gene generate multiple isoforms of a zinc finger polyadenosine RNA binding protein. Gene. 2009;439I:71-8.

72. Novotny I, Blazikova M, Stanek D, Herman P, Malinsky J. In vivo kinetics of U4/U6.U5 tri-snRNP formation in Cajal bodies. Mol Biol Cell. 2011;22I:513-23.

73. Chen G, Zhou X, Florea S, Qian J, Cai W, Zhang Z, et al. Expression of active protein phosphatase 1 inhibitor-1 attenuates chronic beta-agonist-induced cardiac apoptosis. Basic Res Cardiol. 2010;105I:573-81.

74. Chen J, Lee G, Fanous AH, Zhao Z, Jia P, O'Neill A, et al. Two non-synonymous markers in PTPN21, identified by genome-wide association study data-mining and replication, are associated with schizophrenia. Schizophr Res. 2011;131I:43-51.

75. Deery EC, Vithana EN, Newbold RJ, Gallon VA, Bhattacharya SS, Warren MJ, et al. Disease mechanism for retinitis pigmentosa (RP11) caused by mutations in the splicing factor gene PRPF31. Hum Mol Genet. 2002;11I:3209-19.

76. Wilkie SE, Vaclavik V, Wu H, Bujakowska K, Chakarova CF, Bhattacharya SS, et al. Disease mechanism for retinitis pigmentosa (RP11) caused by missense mutations in the splicing factor gene PRPF31. Mol Vis. 2008;14I:683-90.

77. Guardavaccaro D, Pagano M. Stabilizers and destabilizers controlling cell cycle oscillators. Mol Cell. 2006;22I:1-4.

78. Alcina A, Fernandez O, Gonzalez JR, Catala-Rabasa A, Fedetz M, Ndagire D, et al. Tag-SNP analysis of the GFI1-EVI5-RPL5-FAM69 risk locus for multiple sclerosis. Eur J Hum Genet. 2010;18I:827-31.

79. Yu H. Cdc20: a WD40 activator for a cell cycle degradation machine. Mol Cell. 2007;27I:3-16.

80. Di Leva G, Briskin D, Croce CM. MicroRNA in cancer: new hopes for antineoplastic chemotherapy. Ups J Med Sci. 2012;117I:202-16.

81. Mannoor K, Liao J, Jiang F. Small nucleolar RNAs in cancer. Biochim Biophys Acta. 2012;1826I:121-8.

82. Alabert C, Groth A. Chromatin replication and epigenome maintenance. Nat Rev Mol Cell Biol. 2012;13I:153-67.

83. Jezek P. Possible physiological roles of mitochondrial uncoupling proteins--UCPn. Int J Biochem Cell Biol. 2002;34I:1190-206.

84. Echtay KS. Mitochondrial uncoupling proteins--what is their physiological role? Free Radic Biol Med. 2007;43I:1351-71.

85. Kwok KH, Ho PW, Chu AC, Ho JW, Liu HF, Yiu DC, et al. Mitochondrial UCP5 is neuroprotective by preserving mitochondrial membrane potential, ATP levels, and reducing oxidative stress in MPP+ and dopamine toxicity. Free Radic Biol Med. 2010;49I:1023-35.

86. Scheideler M, Elabd C, Zaragosi LE, Chiellini C, Hackl H, Sanchez-Cabo F, et al. Comparative transcriptomics of human multipotent stem cells during adipogenesis and osteoblastogenesis. BMC Genomics. 2008;9I:340.

87. Reina S, Palermo V, Guarnera A, Guarino F, Messina A, Mazzoni C, et al. Swapping of the N-terminus of VDAC1 with VDAC3 restores full activity of the channel and confers anti-aging features to the cell. FEBS Lett. 2010;584I:2837-44.

88. Biber J, Hernando N, Forster I. Phosphate transporters and their function. Annu Rev Physiol. 2013;75I:535-50.

89. Forster IC, Hernando N, Biber J, Murer H. Phosphate transporters of the SLC20 and SLC34 families. Mol Aspects Med. 2013;34I:386-95.

90. Inden M, Iriyama M, Takagi M, Kaneko M, Hozumi I. Localization of type-III sodium-dependent phosphate transporter 2 in the mouse brain. Brain Res. 2013;1531I:75-83.

91. Souza RP, Rosa DV, Romano-Silva MA, Zhen M, Meltzer HY, Lieberman JA, et al. Lack of association of NALCN genetic variants with schizophrenia. Psychiatry Res. 2011;185I:450-2.

92. Daugaard M, Nitsch R, Razaghi B, McDonald L, Jarrar A, Torrino S, et al. Hace1 controls ROS generation of vertebrate Rac1-dependent NADPH oxidase complexes. Nat Commun. 2013;4I:2180.

93. Castillo-Lluva S, Tan CT, Daugaard M, Sorensen PH, Malliri A. The tumour suppressor HACE1 controls cell migration by regulating Rac1 degradation. Oncogene. 2013;32I:1735-42.

94. Mettouchi A, Lemichez E. Ubiquitylation of active Rac1 by the E3 ubiquitin-ligase HACE1. Small GTPases. 2012;3I:102-6.

95. Torrino S, Visvikis O, Doye A, Boyer L, Stefani C, Munro P, et al. The E3 ubiquitin-ligase HACE1 catalyzes the ubiquitylation of active Rac1. Dev Cell. 2011;21I:959-65.

96. Xu Z, He Z, Huang K, Tang W, Li Z, Tang R, et al. No genetic association between NCAM1 gene polymorphisms and schizophrenia in the Chinese population. Prog Neuropsychopharmacol Biol Psychiatry. 2008;32I:1633-6.

97. Sullivan PF, Keefe RS, Lange LA, Lange EM, Stroup TS, Lieberman J, et al. NCAM1 and neurocognition in schizophrenia. Biol Psychiatry. 2007;61I:902-10.

98. Tamura H, Kawata M, Hamaguchi S, Ishikawa Y, Shiosaka S. Processing of neuregulin-1 by neuropsin regulates GABAergic neuron to control neural plasticity of the mouse hippocampus. J Neurosci. 2012;32I:12657-72.
99. Tarttelin EE, Bellingham J, Hankins MW, Foster RG, Lucas RJ. Neuropsin (Opn5): a novel opsin identified in mammalian neural tissue. FEBS Lett. 2003;554I:410-6.
100. Fisher JL, Mott DD. The auxiliary subunits Neto1 and Neto2 reduce voltage-dependent inhibition of recombinant kainate receptors. J Neurosci. 2012;32I:12928-33.
101. Ivakine EA, Acton BA, Mahadevan V, Ormond J, Tang M, Pressey JC, et al. Neto2 is a KCC2 interacting protein required for neuronal Cl- regulation in hippocampal neurons. Proc Natl Acad Sci U S A. 2013;110I:3561-6.
102. Katoh M. WNT/PCP signaling pathway and human cancer (review). Oncol Rep. 2005;14I:1583-8.
103. Katoh Y, Katoh M. Comparative genomics on Vangl1 and Vangl2 genes. Int J Oncol. 2005;26I:1435-40.
104. Saini S, Majid S, Dahiya R. The complex roles of Wnt antagonists in RCC. Nat Rev Urol. 2011;8I:690-9.
105. Braskie MN, Kohannim O, Jahanshad N, Chiang MC, Barysheva M, Toga AW, et al. Relation between variants in the neurotrophin receptor gene, NTRK3, and white matter integrity in healthy young adults. Neuroimage. 2013;82I:146-53.
106. Durany N, Michel T, Zochling R, Boissl KW, Cruz-Sanchez FF, Riederer P, et al. Brain-derived neurotrophic factor and neurotrophin 3 in schizophrenic psychoses. Schizophr Res. 2001;52I:79-86.
107. Durany N, Thome J. Neurotrophic factors and the pathophysiology of schizophrenic psychoses. Eur Psychiatry. 2004;19I:326-37.
108. Pan Y, Zhang J, Liu W, Shu P, Yin B, Yuan J, et al. Dok5 is involved in the signaling pathway of neurotrophin-3 against TrkC-induced apoptosis. Neurosci Lett. 2013;553I:46-51.
109. Tanahashi N, Suzuki M, Fujiwara T, Takahashi E, Shimbara N, Chung CH, et al. Chromosomal localization and immunological analysis of a family of human 26S proteasomal ATPases. Biochem Biophys Res Commun. 1998;243I:229-32.
110. Keppetipola N, Sharma S, Li Q, Black DL. Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. Crit Rev Biochem Mol Biol. 2012;47I:360-78.
111. Aberg KA, Liu Y, Bukszar J, McClay JL, Khachane AN, Andreassen OA, et al. A comprehensive family-based replication study of schizophrenia genes. JAMA Psychiatry. 2013;70I:573-81.

112. Blayney LM, Zissimopoulos S, Ralph E, Abbot E, Matthews L, Lai FA. Ryanodine receptor oligomeric interaction: identification of a putative binding region. J Biol Chem. 2004;279I:14639-48.
113. Zissimopoulos S, Seifan S, Maxwell C, Williams AJ, Lai FA. Disparities in the association of the ryanodine receptor and the FK506-binding proteins in mammalian heart. J Cell Sci. 2012;125I:1759-69.
114. Shimamura A, Alter BP. Pathophysiology and management of inherited bone marrow failure syndromes. Blood Rev. 2010;24I:101-22.
115. Fregoso OI, Das S, Akerman M, Krainer AR. Splicing-factor oncoprotein SRSF1 stabilizes p53 via RPL5 and induces cellular senescence. Mol Cell. 2013;50I:56-66.
116. Dudkiewicz M, Lenart A, Pawlowski K. A novel predicted calcium-regulated kinase family implicated in neurological disorders. PLoS One. 2013;8I:e66427.
117. Mudge J, Miller NA, Khrebtukova I, Lindquist IE, May GD, Huntley JJ, et al. Genomic convergence analysis of schizophrenia: mRNA sequencing reveals altered synaptic vesicular transport in post-mortem cerebellum. PLoS One. 2008;3I:e3625.
118. Alzhanova D, Hruby DE. A host cell membrane protein, golgin-97, is essential for poxvirus morphogenesis. Virology. 2007;362I:421-7.
119. Millard TH, Behrendt B, Launay S, Futterer K, Machesky LM. Identification and characterisation of a novel human isoform of Arp2/3 complex subunit p16-ARC/ARPC5. Cell Motil Cytoskeleton. 2003;54I:81-90.
120. Tang MR, Wang YX, Guo S, Han SY, Wang D. CSMD1 exhibits antitumor activity in A375 melanoma cells through activation of the Smad pathway. Apoptosis. 2012;17I:927-37.
121. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460I:748-52.
122. Consortium I. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature. 2008;455I:237-41.
123. Consortium SPG-WASG. Genome-wide association study identifies five new schizophrenia loci. Nat Genet. 2011;43I:969-76.
124. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, Akterin S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013.
125. Stroup TS, McEvoy JP, Swartz MS, Byerly MJ, Glick ID, Canive JM, et al. The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development. Schizophr Bull. 2003;29I:15-31.

126. Lieberman JA. Comparative effectiveness of antipsychotic drugs. A commentary on: Cost Utility Of The Latest Antipsychotic Drugs In Schizophrenia Study (CUtLASS 1) and Clinical Antipsychotic Trials Of Intervention Effectiveness (CATIE). Arch Gen Psychiatry. 2006;63I:1069-72.
127. Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, Stroup TS, et al. Genomewide association for schizophrenia in the CATIE study: results of stage 1. Mol Psychiatry. 2008;13I:570-84.
128. Consortium TSPG. Genome-wide association study identifies five new schizophrenia loci. Nat Genet. 2011;43I:969-76.
129. Sklar P, Pato MT, Kirby A, Petryshen TL, Medeiros H, Carvalho C, et al. Genome-wide scan in Portuguese Island families identifies 5q31-5q35 as a susceptibility locus for schizophrenia and psychosis. Mol Psychiatry. 2004;9I:213-8.
130. Wang X, Liu B, Li N, Li H, Qiu J, Zhang Y, et al. IPP5, a novel protein inhibitor of protein phosphatase 1, promotes G1/S progression in a Thr-40-dependent manner. J Biol Chem. 2008;283I:12076-84.

Supplemental *Figure Legends*

FIGURE S1. Methodology Workflow of the Divide & Conquer Strategy. Processes involving SNP and phenotypic sets are indicated in blue and red, respectively, whereas procedures concerning phenotypic-genotypic relations are shown in violet. The web server application PGMRA for identifying genotype-phenotype relations is publically available at http://phop.ugr.es/fenogeno (16). Statistical analysis was performed by the SNP-set Kernel Association Test (SKAT) (22, 28), which is also accessible via the web server cited above.

FIGURE S2. The Hypergeometric Statistics (see eqn. 2). (**A**) Relations are evaluated by the probability of intersection (PI) measures the degree of overlapping between two sets/clusters, assigning the lowest p-value to the higher overlapping (red: low; green: high). Illustration of the relations based on two dimensions, where the x-y axes correspond to SNP sets (G_i) and phenotypic sets (P_j), respectively. The similarity of intersection (SI) measures the explanatory quality of a set and it is represented by the size of the circles (big: high; small: low). (**B**) Sagittal diagram illustrating the intersection operation used to build relations and the corresponding evaluation as is reported in (**A**). Colors correspond to the PI and line widths to the SI.

FIGURE S3. **SNP Sets Represented as Submatrices Composed of SNPs (y-axis) Shared by Distinct Subsets of Subjects (x-axis)**. Allele values are indicated as AA (light blue), AB (intermediate blue), BB (dark blue), and missing (black). SNP and subject names/codes are not shown. Subject status was superimposed after SNP set identification: cases (red) and controls (green). SNP sets are labeled by a pair of numbers representing the maximum number of sub-matrices and the order in which they were selected by the method, as described in Figure 2. Row and column dendograms were superimposed *a posteriori* into each sub-matrix for visualization purposes.

FIGURE S4. **The Distribution of Risk for the Uncovered SNP Sets.** Histogram representing the distribution of SNP sets as measured by their density (y-axis, frequency of SNP sets divided by the length of the interval) at different risk values (x-axis). The histogram distribution can be approximated by a normal distribution, where SNP sets at >70% of risk were selected for our analysis, with only 42 non-redundant SNP sets (Table 1).

FIGURE S5. Genotypic Dissection and Identification of the SZ Architecture. Nodes indicate SNP sets linked by shared SNPs (blue line) and/or subjects (red line) without any pre-assumption about the subject status (i.e., case or control).

FIGURE S6. Bioinformatics Analysis of SNPs Derived from SNP Sets Targeting Genomic Regions[2]. (**A**) Multiple SNPs within a SNP set can affect a single gene in many ways. 5 SNPs from the SNP set 19_2 (100% of risk) can affect GOLGA1: SNPs rs10986471 and rs640052 may produce downstream variations; SNP rs634710 can generate missense variations; SNP rs7031479 may introduce intron variants; and SNP rs687434 may create non-coding exon variants (www.ensembl.org, Tables S1 and S3). Two SNP variants of the SNP set 19_2 affect the regulatory region of ncRNAs genes: miRNA AL354928.1 and small nuclear RNA (U4 snRNA) (Table S1). The rs640052 SNP lies between regulatory regions downstream and upstream of U4 and the GOLGA1 gene,

[2] The protein coding genes include the 5′ and 3′ untranslated region (3′ UTR, 5′UTR), exons that code for the coding sequence (CDS) and introns. The ncRNA genes are defined only in terms of exons and introns. The promoter upstream and downstream region for both types of genes have been defined as the segment of 5000bp before the beginning of the 5′ UTR, and 5000bp after the 3′UTR end. The remaining space between the upstream and downstream region of a gene is here defined as the intergenic region.

which may be functionally related. The U4 snRNAs conform the splicesome, which is involved in the splicing process that generates diverse mRNA species from a single pre-mRNA. Consistently, the GOLGA1 gene has substantial variation in alternative splice isoform expression and alternative polyadenylation in cerebellar cortex between normal individuals and SZ patients (117, 130). (**B**) All SNPs from SNP set 71_55 are located in the intergenic region upstream of the NTRK3 gene, in a location of a predicted enhancer (Table S1). Nevertheless, those SNPs of the 14_6 SNP set are located within NTRK3, principally in intronic regions and within the upstream region of pseudogene RP11-356B18.1 (Table S1). The latter pseudogene is harbored in an intron of NTRK3 that is processed in the NTRK-005 transcript variant, which does not code neurotrophin receptor-3 protein. This suggests that a mutation in the first SNP set may inhibit the transcription of the corresponding gene, whereas mutations in the second SNP set may block or decrease production of the corresponding protein (Table S3).

FIGURE S7. Pathway Analysis. Distinct pathways identified by the SNP sets are well known, relevant and interconnected signaling pathways for neural development, neurotrophin function, neurotransmission, and neurodegenerative disorders (see Tables S1 and S4). Other genes uncovered are also overwhelmingly expressed in the brain, and participate in regulation of intracellular signaling, oxidative stress, apoptosis, neuroimmune regulation, protein synthesis, and epigenetic gene expression.

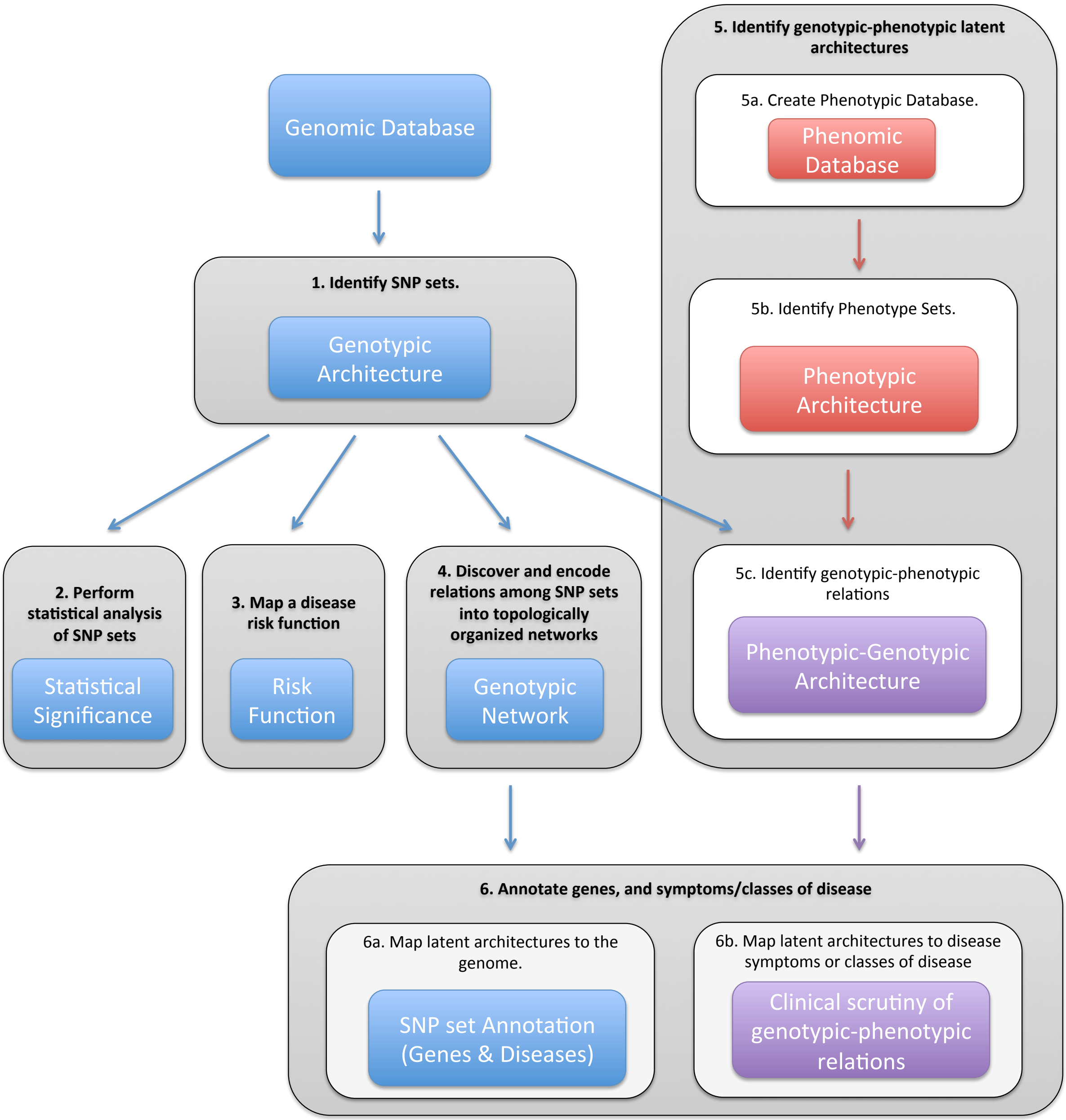

Genomic Database
1. Identify SNP sets.
Genotypic Architecture
2. Perform statistical analysis of SNP sets
Statistical Significance
3. Map a disease risk function
Risk Function
4. Discover and encode relations among SNP sets into topologically organized networks
Genotypic Network
5. Identify genotypic-phenotypic latent architectures
5a. Create Phenotypic Database.
Phenomic Database
5b. Identify Phenotype Sets.
Phenotypic Architecture
5c. Identify genotypic-phenotypic relations
Phenotypic-Genotypic Architecture
6. Annotate genes, and symptoms/classes of disease
6a. Map latent architectures to the genome.
SNP set Annotation (Genes & Diseases)
6b. Map latent architectures to disease symptoms or classes of disease
Clinical scrutiny of genotypic-phenotypic relations

Figure S1

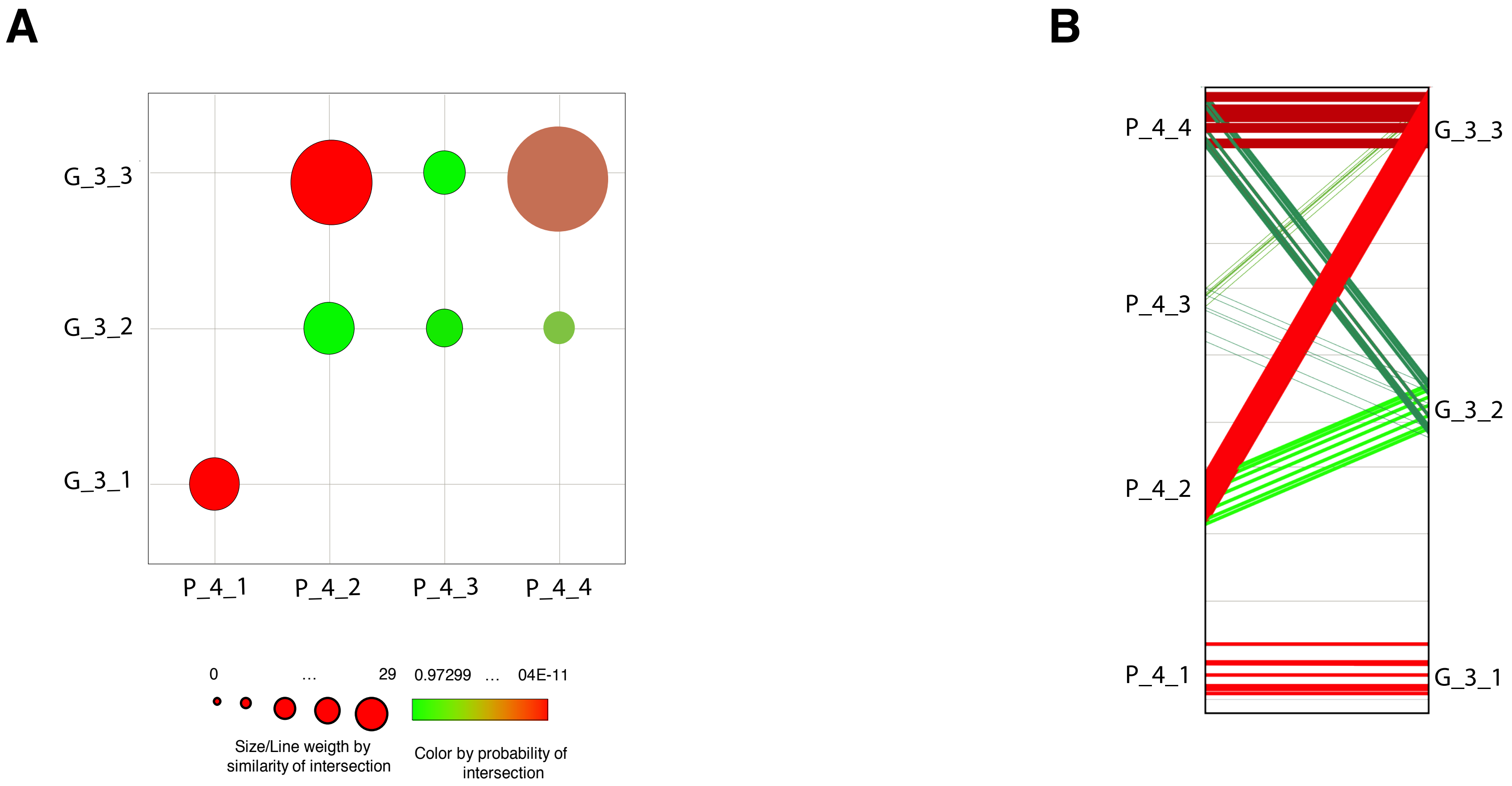

A
G_3_3
G_3_2
G_3_1
P_4_1
P_4_2
P_4_3
P_4_4
0
...
29
0.97299
...
04E-11
Size/Line weigth by similarity of intersection
Color by probability of intersection
B
P_4_4
P_4_3
P_4_2
P_4_1
G_3_3
G_3_2
G_3_1

Figure S2

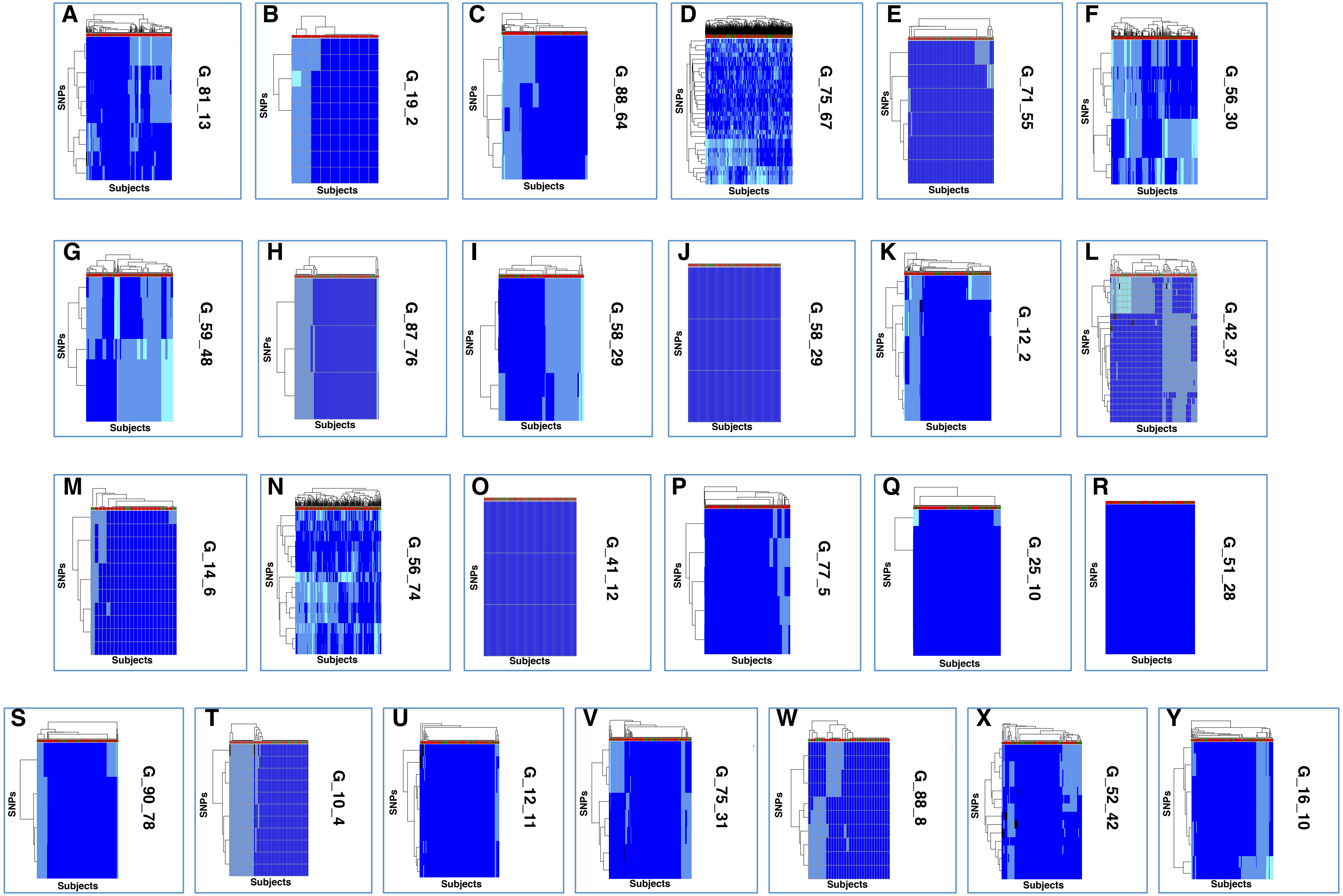

A
SNPs
Subjects
G_81_13
B
SNPs
Subjects
G_19_2
C
SNPs
Subjects
G_88_64
D
SNPs
Subjects
G_75_67
E
SNPs
Subjects
G_71_55
F
SNPs
Subjects
G_56_30
G
SNPs
Subjects
G_59_48
H
SNPs
Subjects
G_87_76
I
SNPs
Subjects
G_58_29
J
SNPs
Subjects
G_58_29
K
SNPs
Subjects
G_12_2
L
SNPs
Subjects
G_42_37
M
SNPs
Subjects
G_14_6
N
SNPs
Subjects
G_56_74
O
SNPs
Subjects
G_41_12
P
SNPs
Subjects
G_77_5
Q
SNPs
Subjects
G_25_10
R
SNPs
Subjects
G_51_28
S
SNPs
Subjects
G_90_78
T
SNPs
Subjects
G_10_4
U
SNPs
Subjects
G_12_11
V
SNPs
Subjects
G_75_31
W
SNPs
Subjects
G_88_8
X
SNPs
Subjects
G_52_42
Y
SNPs
Subjects
G_16_10

Figure S3

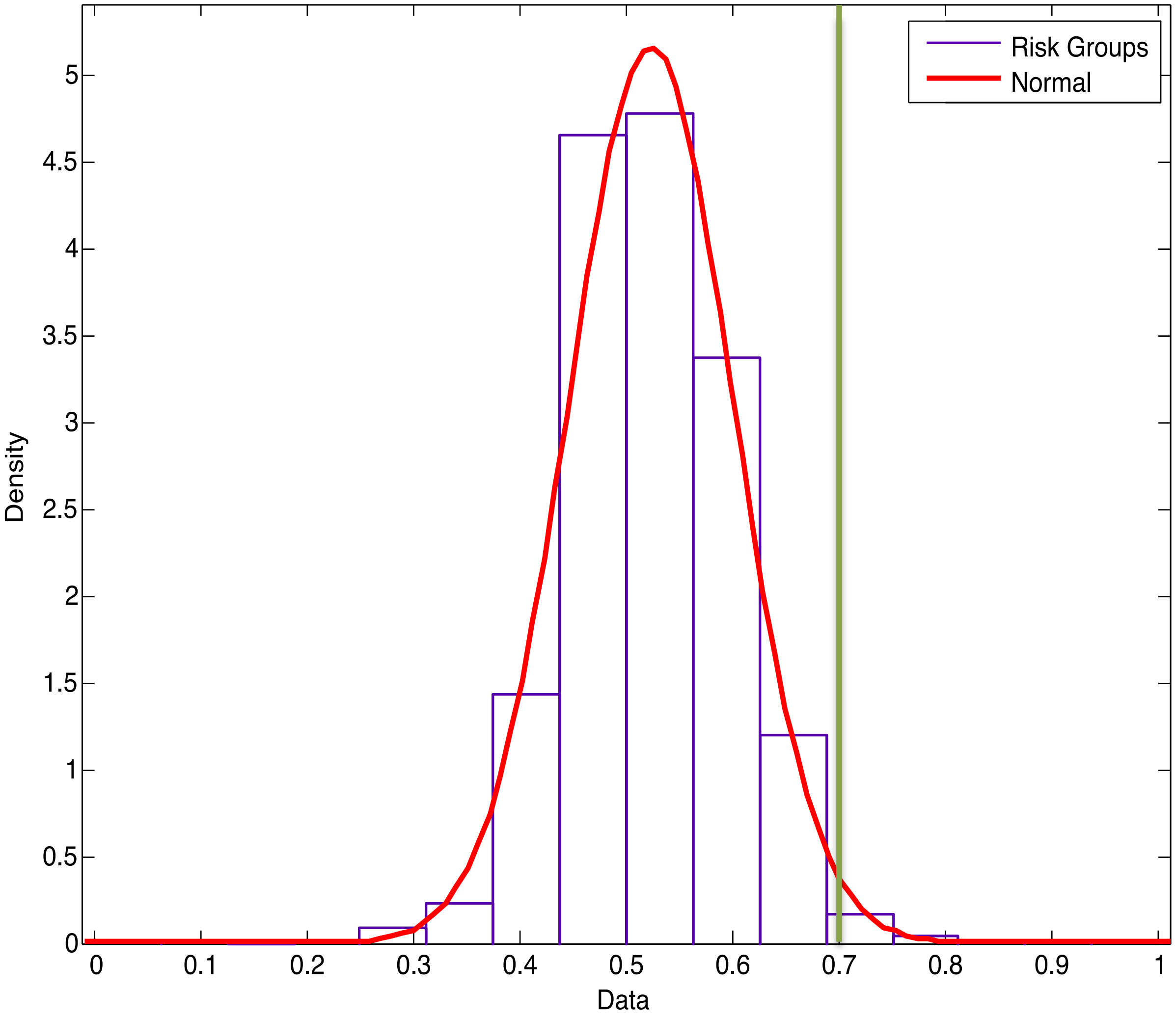

Risk Groups
Normal
Density
Data
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure S4

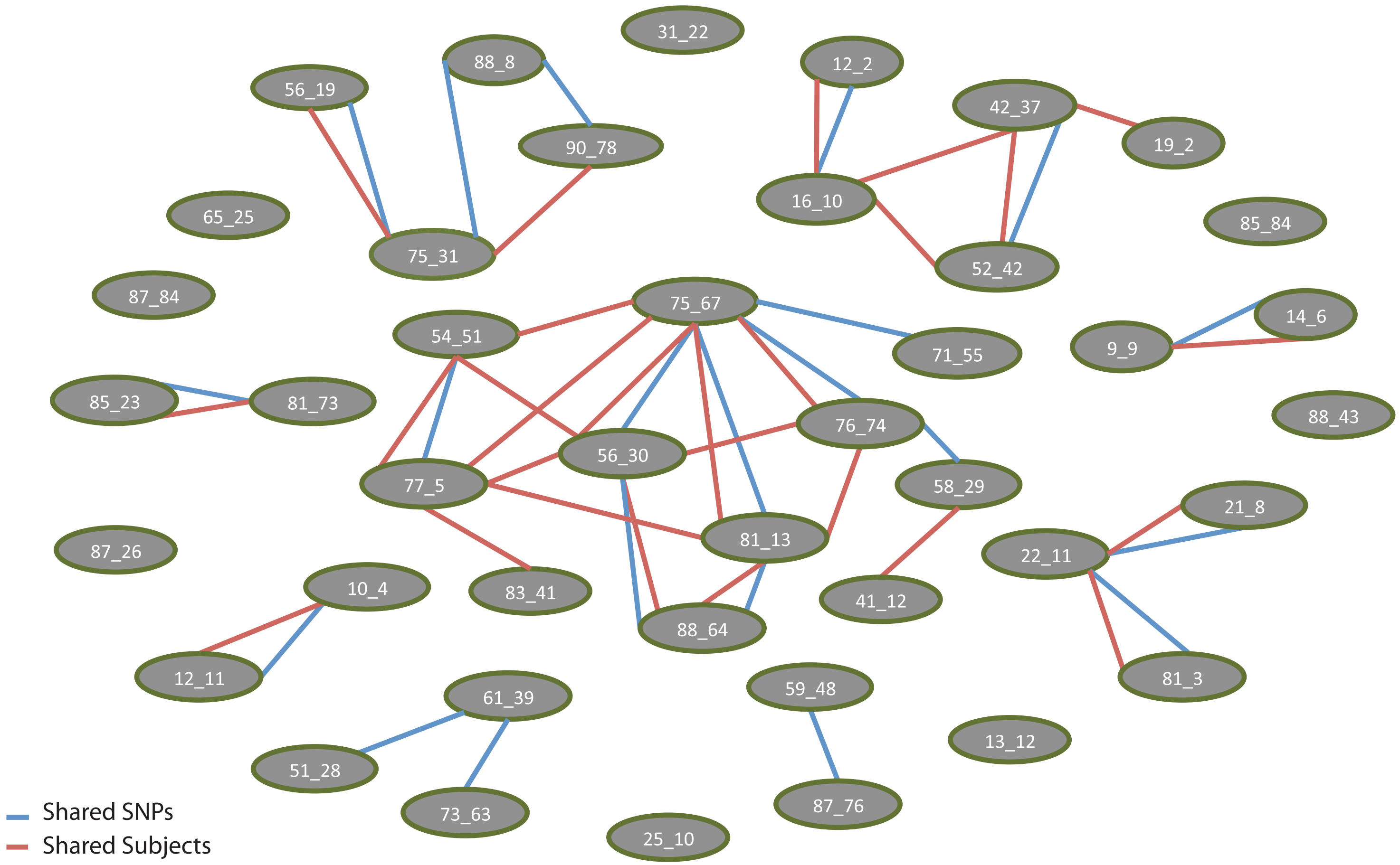
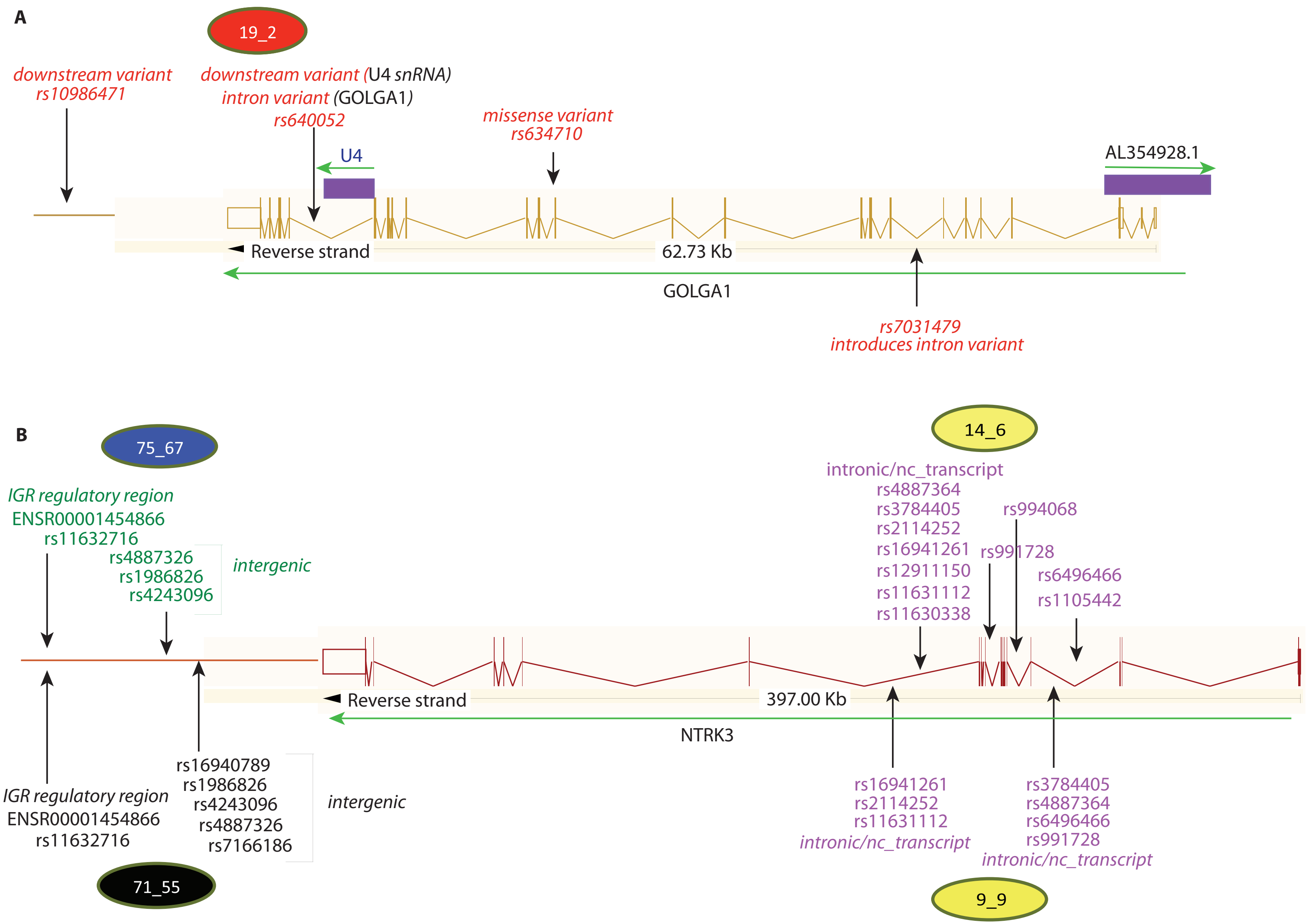

31_22
88_8
12_2
56_19
42_37
90_78
19_2
16_10
65_25
85_84
75_31
52_42
87_84
75_67
14_6
54_51
71_55
9_9
85_23
81_73
88_43
76_74
56_30
77_5
58_29
21_8
81_13
87_26
22_11
10_4
83_41
41_12
88_64
12_11
59_48
81_3
61_39
13_12
51_28
87_76
Shared SNPs
73_63
Shared Subjects
25_10

Figure S5



A
19_2
downstream variant
rs10986471
downstream variant (U4 snRNA)
intron variant (GOLGA1)
rs640052
missense variant
rs634710
U4
AL354928.1
Reverse strand
62.73 Kb
GOLGA1
rs7031479
introduces intron variant
B
75_67
IGR regulatory region
ENSR00001454866
rs11632716
rs4887326
rs1986826
rs4243096
intergenic
14_6
intronic/nc_transcript
rs4887364
rs3784405
rs2114252
rs16941261
rs12911150
rs11631112
rs11630338
rs994068
rs991728
rs6496466
rs1105442
Reverse strand
397.00 Kb
NTRK3
IGR regulatory region
ENSR00001454866
rs11632716
rs16940789
rs1986826
rs4243096
rs4887326
rs7166186
intergenic
71_55
rs16941261
rs2114252
rs11631112
intronic/nc_transcript
rs3784405
rs4887364
rs6496466
rs991728
intronic/nc_transcript
9_9

Figure S6

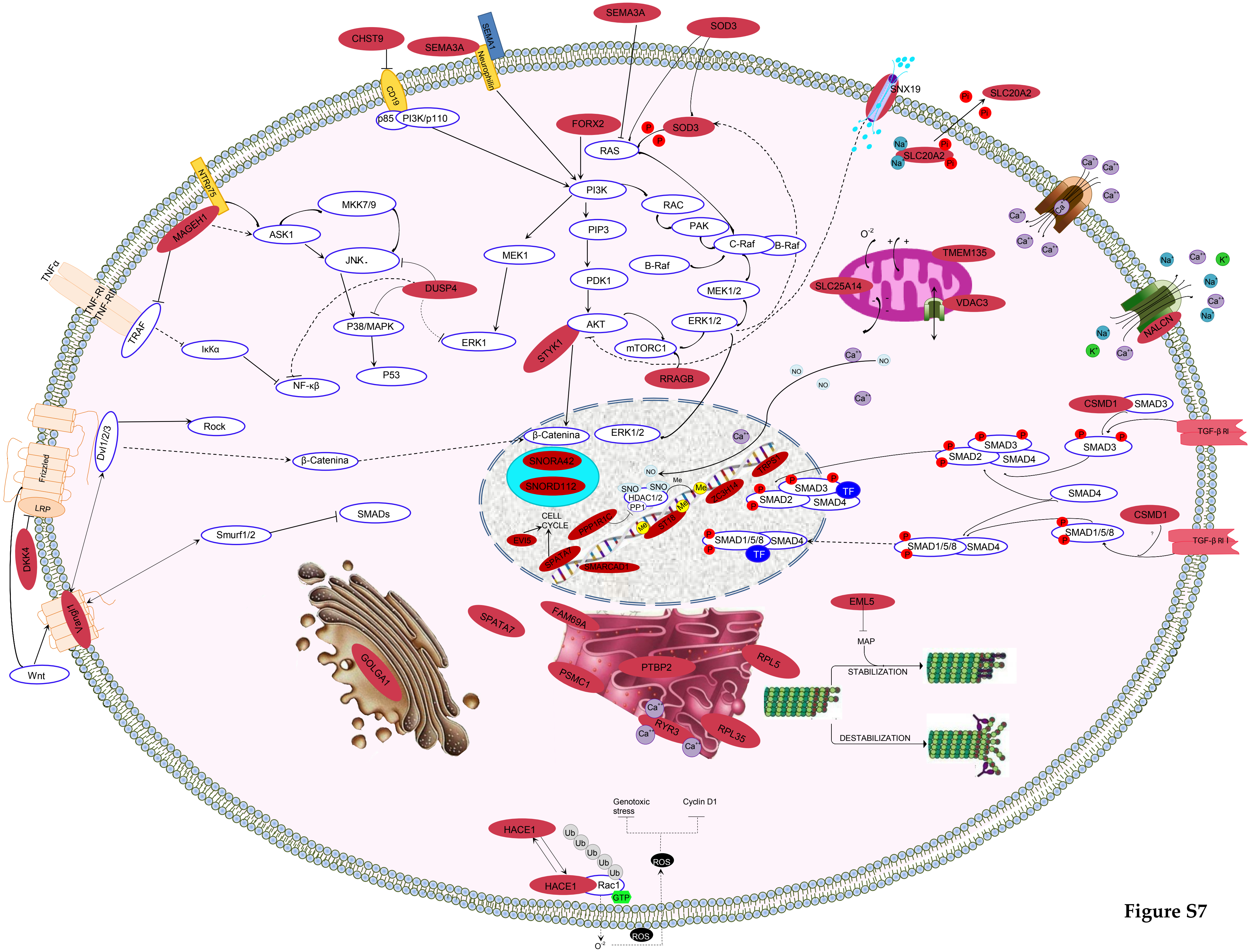

CHST9
SEMA3A
SEMA1
Neurophilin
SEMA3A
SOD3
CD19
p85
PI3K/p110
FORX2
P
P
SOD3
RAS
SNX19
Pi
SLC20A2
Pi
Na+
Pi
SLC20A2
Na+
Pi
NTRp75
MAGEH1
MKK7/9
ASK1
JNK
PI3K
PIP3
RAC
PAK
C-Raf
B-Raf
B-Raf
MEK1
PDK1
MEK1/2
TNFα
TNF-RI
TNF-RII
DUSP4
TRAF
P38/MAPK
ERK1
AKT
ERK1/2
mTORC1
IκKα
NF-κβ
P53
STYK1
RRAGB
O-2
TMEM135
SLC25A14
VDAC3
Ca++
Ca
Na+
K+
NALCN
NO
Dvl1/2/3
Rock
Frizzled
LRP
β-Catenina
β-Catenina
ERK1/2
SNORA42
SNORD112
CSMD1
SMAD3
TGF-β RI
SMAD3
SMAD2
SMAD4
SMAD3
TRPS1
SNO
SNO
HDAC1/2
PP1
Me
ZC3H14
SMAD3
TF
SMAD2
SMAD4
SMAD4
CELL CYCLE
PPP1R1C
ST18
EVI5
SPATA7
SMARCAD1
SMAD1/5/8
SMAD4
TF
SMAD1/5/8
SMAD4
SMAD1/5/8
CSMD1
TGF-β RI I
Smurf1/2
SMADs
DKK4
Vangl1
Wnt
GOLGA1
SPATA7
FAM69A
PTBP2
RPL5
PSMC1
RYR3
RPL35
EML5
MAP
STABILIZATION
DESTABILIZATION
Genotoxic stress
Cyclin D1
ROS
HACE1
Ub
HACE1
Rac1
GTP
ROS
O-2

Figure S7