Rare coding variants as risk modifiers of the 22q11.2 deletion implicate postnatal cortical development in syndromic schizophrenia

Jhih-Rong Lin[1], Yingjie Zhao[1], M. Reza Jabalameli[1], Nha Nguyen[1], Joydeep Mitra[1], International 22q11.2DS Brain and Behavior Consortium[*], Ann Swillen[2], Jacob A. S. Vorstman[3], Eva W. C. Chow[3], Marianne van den Bree[4], Beverly S. Emanuel[5], Joris R. Vermeesch[2], Michael J. Owen[4], Nigel M. Williams[4], Anne S. Bassett[3], Donna M. McDonald-McGinn[5], Raquel E. Gur[6], Carrie E. Bearden[7], Bernice E. Morrow[1], Herbert M. Lachman[1], Zhengdong D. Zhang[1,§]

1. Department of Genetics, Albert Einstein College of Medicine, NY, USA
2. Department of Human Genetics, KU Leuven, Leuven, Belgium
3. Department of Psychiatry, University of Toronto, Toronto, ON, Canada
4. MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK
5. Division of Human Genetics and 22q and You Center, the Children's Hospital of Philadelphia, Philadelphia, PA, USA
6. Department of Psychiatry and Lifespan Brain Institute, Penn Medicine—CHOP, University of Pennsylvania, Philadelphia, PA, USA
7. Departments of Psychiatry and Biobehavioral Sciences and Psychology, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA

* A full list of members and their affiliations appears in the supplement.
§ Corresponding author (E-mail: zhengdong.zhang@einsteinmed.edu)

Keywords: 22q11.2 deletion, schizophrenia, rare variants, disease risk modification

ABSTRACT

22q11.2 deletion is one of the strongest known genetic risk factors for schizophrenia. Recent whole-genome sequencing of schizophrenia cases and controls with this deletion provided an unprecedented opportunity to identify risk modifying genetic variants and investigate their contribution to the pathogenesis of schizophrenia in 22q11.2 deletion syndrome. Here, we apply a novel analytic framework that integrates gene network and phenotype data to investigate the aggregate effects of rare coding variants and identified modifier genes in this etiologically homogenous cohort (223 schizophrenia cases and 233 controls of European descent). Our analyses revealed significant additive genetic components of rare nonsynonymous variants in 110 modifier genes (adjusted $P = 9.4E-04$) that overall accounted for 4.6% of the variance in schizophrenia status in this cohort, of which 4.0% was independent of the common polygenic risk for schizophrenia. The modifier genes affected by rare coding variants were enriched with genes involved in synaptic function and developmental disorders. Spatiotemporal transcriptomic analyses identified an enrichment of coexpression between modifier and 22q11.2 genes in cortical brain regions from late infancy to young adulthood. Corresponding gene coexpression modules are enriched with brain-specific protein-protein interactions of *SLC25A1*, *COMT*, and *PI4KA* in the 22q11.2 deletion region. Overall, our study highlights the contribution of rare coding variants to the SCZ risk. They not only complement common variants in disease genetics but also pinpoint brain regions and developmental stages critical to the etiology of syndromic schizophrenia.

INTRODUCTION

22q11.2 deletion syndrome (22q11.2DS) is a severe developmental disorder, caused predominantly by a 3-Mb genomic deletion (in ~85% of cases)[1,2]. It is one of the most common chromosomal abnormalities (~1 in 4000 live births, ~1 in 1000 fetuses)[3,4]. Neuropsychiatric dysfunction is a prominent feature among its heterogeneous phenotypic presentations: About one in four individuals with 22q11.2DS develops schizophrenia (SCZ), usually in late adolescence or early adulthood[5-7]. 22q11.2DS has been widely used as a model to study SCZ due to its highly increased disease risk and clinical features compatible with the more common, idiopathic form[7-10]. Thus, identifying genetic risk factors for SCZ in addition to 22q11.2 deletion in those with 22q11.2DS has high clinical and scientific value[9,11].

In addition to 22q11.2 deletion and common variants associated with SCZ, increasing evidence shows that many rare variants may act as the 'second hits' for SCZ in 22q11.2DS[7] and thus contribute to its variable expressivity[12,13]. Identifying rare coding variants separate from the common polygenic risk can complement risk prediction of 22q11.2DS-associated SCZ[9,11] and pinpoint genes that play a role in its pathophysiology. Analyzing such modifier genes may reveal important biological processes in the development of syndromic SCZ and their connection to disease risk genes in the 22q11.2 deletion regions.

To uncover genetic variation that contributes to the high prevalence of SCZ in 22q11.2DS, the International 22q11.2DS Brain and Behavior Consortium (IBBC) assembled and whole-genome sequenced (WGS) a cohort of 519 22q11.2DS patients with and without SCZ, providing an indispensable resource to identify modifier rare coding variants that change the disease risk of the 22q11.2 deletion in syndromic SCZ[9]. Although the cohort is by far the largest to date, the conventional rare variant association tests failed to identify significant associations with SCZ for any gene or pathway after multiple test correction, likely due to insufficient statistical power from the moderate sample size. Therefore, it is imperative to use a different approach.

We developed an integrated method – the integrated gene signal processing (IGSP) – that can substantially increase the power to identify risk genes implicated by rare coding variants in case-control sequencing studies using gene networks and mouse knockout phenotypes[14]. Importantly, IGSP follows a 'discovery-driven' strategy to score risk genes without relying on prior disease-specific knowledge to avoid scoring bias (**Supplementary note**). In this study, with a new analytic framework that incorporates IGSP, we investigated modifier rare coding variants associated with SCZ in 22q11.2DS with three interconnected aims to (1) identify the risk component of rare coding variants in each 22q11.2DS patient and its contribution to SCZ, (2) uncover modifier genes and related biological processes that play important roles in SCZ risk modification in 22q11.2DS, and (3) elucidate the functional connection between modifier genes and 22q11.2 deletion in brain development.

RESULTS

Rare coding variants contribute to the genetic risk of SCZ in 22q11.2DS

Our study cohort consisted of 223 SCZ cases and 233 controls of European ancestry in the sequenced cohort of 22q11.2DS patients (**Supplementary Table S2**). We identified 173,752 rare coding variants (170,919 SNPs and 2,833 indels) with alternative allele frequency (AAF) < 0.01 across 18,828 coding genes in the study cohort. This includes 61,900 synonymous, 101,944 missense, and 4,220 loss-of-function (stop gain and frame shift) rare variants. There was no significant difference in the exome-wide minor allele counts of either rare coding variants or rare nonsynonymous variants between SCZ cases and controls ($P = 0.857$ and $P = 0.2$, respectively) (**Supplementary Fig. S1**).

In this study, we used an integrated approach with IGSP[14] to address the issue of insufficient statistical power of the previous study[9], which found no significant rare variant association with SCZ in the same 22q11.2DS patient cohort. We developed IGSP to improve the power for prioritizing risk genes implicated by rare variants by integrating gene-level rare variant association signals with gene functional network and gene knock-out phenotypes. IGSP scores genes based on both rare variant association signals of genes and their joint functional and phenotypic relationship (**Supplementary note**). According to simulations, the IGSP score outperforms the rare variant gene association signal by 2 to 3 times[14] in prioritizing disease risk genes.

Using IGSP to leverage its improved power of prioritizing risk genes, we first defined a modifier risk score (MRS) to quantify aggregate effects of rare coding variants in risk genes at the individual level (see **Methods**). We then evaluated how much SCZ risk that MRS could account for in our study cohort with a new bootstrapping pipeline. Briefly, we randomly divided our study cohort into 500 discovery and corresponding target subcohorts. For each discovery subcohort, we ran IGSP to score the risk of genes implicated by rare variants. Next, using top-scoring genes, we calculated the MRS of each subject in the corresponding target subcohort. Finally, we averaged the MRS ranks across all target subcohorts for each subject in the study cohort (**Fig. 1**). Our simulation showed that the average MRS rank can effectively quantify individuals' relative SCZ risk from modifier rare coding variants. It approaches the optimal MRS (after standardization) when the prioritization power of risk genes increases (**Supplementary note** and **Fig. S2**).

The results of our bootstrapping analysis (**Fig. 1**) showed that SCZ cases had a significantly higher MRS than controls (**Fig. 2a**): rare nonsynonymous variants in 110 putative modifier genes prioritized by IGSP with a full integration of both gene network and phenotypes explained 4.6% of the variance in SCZ status in the study cohort (i.e., Nagelkerke's R^2 = 4.6%, 99% CI: 3.7%-5.6%, adjusted P = 9.4E-04). On average, subjects with a higher MRS (> 50%) in a target cohort have an odds ratio (OR) of ~1.3 to develop SCZ (**Fig. 2b**). We identified a smaller aggregate genetic signal from rare coding variants in 40 putative modifier genes prioritized by IGSP with only a network integration (Nagelkerke's R^2 = 3.2%, 99% CI: 2.4%-4.2%, adjusted P = 9.6E-03) (**Supplementary Fig. S3**). In contrast, we did not observe a clear aggregate effect of rare coding variants (Nagelkerke's R^2 = 1.0%, 99% CI: 0.5%-1.5%, adjusted P = 0.64) in the top 90 genes prioritized by genetic association signals alone (**Supplementary Fig. S4**). Overall, these results suggest that the integration of the gene network and mouse knockout phenotypes with genetic association signals significantly improved modifier gene prioritization.

To better understand the nature of rare coding variants as SCZ risk modifiers, we aggregated them in different ways (**Fig. 2c**). The results strongly suggested that modifier rare coding variants in the study cohort can either increase or decrease SCZ risk, as the risk contribution of the modifier rare coding variants would not have been identified without accounting for the direction derived from the burden test. Aggregation of rare variants without normalizing their effects at the gene level substantially diminished our ability to identify their contribution to disease risk, likely due to bias toward large genes or ones with a high density of rare coding variants. The results also showed that including the variant weights based on their predicted functionality (quantified as CADD scores[15]) improved the identification of the risk contribution from rare coding variants, supporting the hypothesis that functional rare variants in SCZ risk genes outside the 22q11.2 deletion region can modify the risk of developing SCZ among 22q11.2DS patients. This hypothesis was further supported by the observation that no risk contribution from rare synonymous variants was detected (**Fig. 2d**). We further examined individual types of rare nonsynonymous variants and observed an association between SCZ and the risk component from missense (P = 0.009) or loss-of-function (P = 0.028) rare variants (**Fig. 2d** and **Supplementary Fig. S5**). The agreement in risk direction between missense and loss-of-function variants suggests that most functional modifier rare variants contribute to the SCZ risk by impairing the

function of carrier genes (**Supplementary Fig. S5**). Of 173,752 rare variants in our study cohort, a very small proportion – 2,312 variants (1.3%) – have AAF > 0.01 among non-Finnish Europeans in gnomAD, a large population reference panel. After excluding them from MRS calculation, we observed a small increase in the identified risk contribution (Nagelkerke's R^2 = 4.8%, 99% CI: 4.1%-5.9%), likely due to specific ancestries and heterogeneous effects of some excluded rare variants.

Although common variants associated with idiopathic SCZ also contribute to syndromic SCZ in 22q11.2DS[9], a finding that we replicated in our analysis (Nagelkerke's R^2 = 4.9%, adjusted P = 5.0E-04, **Supplementary Fig. S6**), it is not clear how common and rare variants together contribute to the overall genetic risk for developing SCZ in 22q11.2DS. To answer this question, for all individuals in the study cohort we first calculated polygenic risk scores (PRS) for idiopathic SCZ using PRSice-2 with the P-value threshold that yielded the most robust result. Risk scores for modifier rare coding variants were obtained by combining MRS across 500 target subcohorts (i.e., the average MRS rank) as described in **Fig. 1**. We confirmed that the risk components from common and rare variants were highly orthogonal to each other: 4.0% out of 4.6% and 4.3% out of 4.9% of the variance in the SCZ status was explained exclusively by MRS and SCZ PRS, respectively (See **Methods**). With the orthogonal nature of two different types of risk scores, we showed that their combination can explain SCZ expression (Nagelkerke's R^2 = 8.9%) substantially better than MRS or SCZ PRS alone (**Fig. 3**).

Rare variants implicate genes that modify risk of SCZ in 22q11.2DS

To better understand the contribution of rare variants to the development of SCZ in 22q11.2DS, we used the full integration of IGSP to identify disease risk modifying genes (i.e., modifier genes hereafter) by scoring genes outside 22q11.2 deletion region for their connection with SCZ based on rare nonsynonymous variants in them in the full study cohort. Our statistical framework showed that rare nonsynonymous variants in the top 110 IGSP-scored genes (with a full integration) of a random subcohort (n = 406, 89% of the full study cohort) can best explain the SCZ status of the remaining subjects (n = 50) in the study cohort (**Fig. 2a**). Therefore, we selected the 110 top-scoring genes of the full study cohort (**Supplementary Table S3**) as putative modifier genes and analyzed their relevant clinical support, pathway enrichment, regulatory elements, and cross-ethnicity. Of these 110 modifier genes, 54 and 56 had a higher weighted burden of rare nonsynonymous variants among SCZ cases and controls (and thus increase or decrease SCZ risk in 22q11.2DS), respectively. These candidate modifiers have been annotated for neurodevelopmental disorders in the ClinVar database (as of July 31, 2021) (**Supplementary Table S4**).

To explore the biology involved in the disease risk modulation of modifier genes, we analyzed gene-set enrichment among them with 10 preselected gene sets relevant to SCZ etiology (see **Methods**). We identified a significant enrichment of synaptic genes (P = 8.29E-09) and developmental disorder genes (P = 9.66E-04) (**Fig. 4a** and **Supplementary Fig. S7**). Notably, although the 22q11.2 deletion rarely occurs, we also detected an enrichment of loss-of-function intolerant genes (P = 9.36E-06) and missense constrained

genes (P = 3.19E-05). Using a hypothesis-free approach, we further investigated the enrichment of modifier genes in gene sets for different biological processes and found that gene sets with the most significant enrichment were highly relevant to neurological (e.g., neurogenesis and differentiation) and developmental biological processes (e.g., neuron development and heart development) (**Fig. 4b**). While our gene-set enrichment analysis revealed biological connections between modifier genes and SCZ, further examination identified known SCZ genes among them with more specific pathological roles in synaptic function (e.g., *TNIK*[16] and *NRXN1*[17]) or calcium channel activity (e.g. *CACNA1C*[18]). In addition, some modifier genes (e.g., *BDNF* and *HIF1A*)[19-21] are associated with other mental disorders, such as bipolar disorder and major depression disorder, that share disease pathogenesis with SCZ[22].

To ascertain the role of modifier genes in modulating SCZ risk, we examined their expression (on the exon level) during brain development and its connection to the SCZ association signals of rare variants. Using BrainSpan RNA-seq data, we identified 1,318 exons in 92 (83.6%) of the 110 modifier genes with general expression in developing brains (See **Methods**). Association tests (SKAT-O) on 2,353 rare variants in those exons confirmed a significant association with SCZ (P = 1.44E-09). Moreover, 1,018 exons of the 110 modifier genes with specific expression in developing brains, and 1,477 rare variants in them showed less evidence of association with SCZ (P = 0.0015).

Given the WGS data, we were interested in not only rare but low-frequency ($1\% \leq AAF < 5\%$) noncoding variants in the regulatory elements of modifier genes. Such noncoding variants are important for two reasons. First, they may constitute additional risk components not covered by SCZ PRS (common variants) and MRS (rare coding variants). Second, their SCZ association can confirm the involvement of the identified modifier genes in developing 22q11.2DS-associated SCZ. Using SKAT-O, we tested SCZ association of rare variants aggregated in four types of regulatory regions of those 110 modifier genes (see **Methods**) and identified a significant SCZ association of transcriptional regulator binding sites in neurons (neuron-TRBS) (P = 1.87E-04). Next, we added low-frequency variants to the tests and identified a significant SCZ association of enhancers (P = 4.69E-04) (**Supplementary Table S5**).

In addition to 22q11.2DS patients of European ancestry, the WGS data generated by IBBC included patients of other minority ethnicities (**Supplementary Table S2**), among which the Hispanic subcohort was the largest (n = 31; 19 SCZ cases and 12 controls). We investigated whether the modifier genes identified among 22q11.2DS patients of European ancestry (i.e., our study cohort) were also enriched with the modifier variants among patients of other ethnicities. We first aggregated rare nonsynonymous variants in the 110 modifier genes in the Hispanic subcohort and tested the association with SCZ for this set of variants. We did not identify a significant association (SKAT-O, P = 0.094). Moreover, for the 110 modifier genes, our association test (SKAT-O) identified a significant SCZ association of rare noncoding variants in the neuron-TRBS (P = 0.03) and of rare and low-frequency noncoding variants in the brain enhancer regions (P = 0.047), respectively, in this Hispanic subcohort, despite its small sample size.

Modifier genes are potential therapeutic targets for SCZ in 22q11.2DS. To investigate whether they are involved in common biological processes with genes affected by SCZ drugs, we analyzed the expression of modifier genes and that of genes most differentially expressed in cells treated with SCZ drugs (see **Methods**). We selected 3 FDA-approved antipsychotics – i.e., Haloperidol, Clozapine, and Quetiapine – and identified a significant enrichment of coexpression between modifier genes and drug-induced top 10 differentially expressed genes (P = 8.4E-04, P = 9.4E-04, and P = 8.0E-04, respectively). The significance of the enrichment was not sensitive to different numbers of top differentially expressed genes.

Both modifier and 22q11.2 genes are involved in brain development

22q11.2 deletion is the strongest known molecular genetic risk factor for SCZ[23]. We hypothesize that the modifier genes and 22q11.2 deletion share convergent pathological mechanisms in brain development. We conducted contextualized analysis using brain transcriptomic data to assess this hypothesis. First, using data from PsychENCODE[24] we constructed gene coexpression networks corresponding to different spatiotemporal combinations of regions and stages during brain development (**Supplementary Table S6**, **S7**, and **S8**), following a previously developed approach[25]. A sliding window was used to combine three consecutive time periods into a time frame, so there was an overlap of samples between two successive time frames. First, we assessed the enrichment of coexpression among the modifier genes in the spatiotemporal combinations of brain development to test the hypothesis that modifier genes are likely involved in the same biological process in a spatiotemporal combination (see **Methods**). Our results showed that modifier genes tended to be coexpressed in the cortical region from the neonatal period to young adulthood, the limbic system from late infancy to young adulthood, and mediodorsal nucleus of the thalamus and cerebellar cortex from late fetal stage to early childhood. (**Fig. 5a** and **Supplementary Table S6** and **S7**). Next, we tested whether modifier genes tended to be coexpressed with genes in the 22q11.2 deletion region (hereafter 22q11.2 genes) in any spatiotemporal combinations of brain development (see **Methods**). We identified enriched coexpression between modifier and 22q11.2 genes with spatiotemporal combinations localized to the cortical region from late infancy to young adulthood (**Fig. 5b**); five convergent spatiotemporal combinations were implicated by both modifier genes and connection between modifier and 22q11.2 genes: P7-9/R1, P8-10/R1, P8-10/R2, P9-11/R1, and P10-11/R1. This suggests that modifier and 22q11.2 genes likely disrupt intersecting biological processes at these convergent points of postnatal cortical development.

To uncover potential intersecting biological processes disrupted by modifier genes and 22q11.2 deletion, we carried out the weighted gene co-expression network analysis (WGCNA) to identify gene coexpression modules in the aforementioned five convergent spatiotemporal combinations of brain development, all of which have at least 49 samples (**Supplementary Table S8**). Among them, P8-10/R2 did not produce a soft thresholding-based scale free topology model fit > 0.8 (**Supplementary Fig. S8**) and thus was excluded from the following module analyses. 10 ~ 18 gene coexpression modules (each with more than 100 genes) were identified at each combination (see **Methods**) (**Supplementary Fig.**

S9 and **Table S9-12**). We then searched for modules that may be affected by both modifier and 22q11.2 genes. For the deletion, we assembled a gene set (i.e., 22q-SB-PPI genes) including both 22q11.2 genes and their direct interaction partners in a spatiotemporal brain protein-protein interaction network (SB-PPI) (see **Methods**) since 22q11.2 genes have been shown to form a brain-developmental PPI network that may affect SCZ-associated modules[26]. As we assessed whether 22q11.2 deletion affects a module by testing its enrichment of 22q-SB-PPI genes, we identified significant concurrent enrichment of both modifier and 22q-SB-PPI genes in modules of three out of four spatiotemporal combinations: $M3_{P7\text{-}9/R1}$, $M4_{P9\text{-}11/R1}$, and $M1_{P10\text{-}11/R1}$ (**Fig. 5c-f**). $M3_{P7\text{-}9/R1}$ corresponded to the period from late infancy to late childhood, while $M4_{P9\text{-}11/R1}$ and $M1_{P10\text{-}11/R1}$ corresponded to the period from middle childhood to young adulthood and the period from adolescence to young adulthood, respectively. The three modules were from the same brain region (R1) and highly overlapped, sharing 638 genes among them (**Supplementary Fig. S10**). They were associated with nervous system development, cell migration and angiogenesis (**Supplementary Fig. S11**) and were enriched with cell-marker genes for neurons and astrocytes. To investigate specific 22q11.2 genes that contributed to the observed enrichment, we tested the enrichment of SB-PPI using individual 22q11.2 genes as the seed gene in these three modules. We identified significant enrichment of SB-PPI for *SLC25A1* in all of them, *COMT* in $M4_{P9\text{-}11/R1}$ and $M1_{P10\text{-}11/R1}$, *SEPT5* and *CRKL* in $M3_{P7\text{-}9/R1}$, and *PI4KA* and *CLDN5* in $M1_{P10\text{-}11/R1}$ (**Fig. 5f**).

DISCUSSION

Using an integrated approach to gene prioritization and a risk-scoring framework based on bootstrapping, we analyzed the WGS-based genotype data of a 22q11.2DS cohort to identify rare variants that modify syndromic SCZ risk. Our modifier risk scoring uncovered potential rare coding variant-based genetic risk to develop SCZ in 22q11.2DS at the individual patient level that were largely independent of common polygenic risk for idiopathic SCZ. The implicated modifier genes were enriched with genes involved in neurodevelopment and synaptic functions. Recent genetic studies of rare and common variants for idiopathic SCZ suggest the convergent disease origin especially in synaptic biology[27-29] and highlight the importance of using different approaches. Our results shed new light on genetic modifiers of SCZ in 22q11.2DS and show that synapse genes harbor genetic variants modifying SCZ risk caused by 22q11.2 deletion, suggesting a convergent disease origin between idiopathic and syndromic SCZ. In addition, our results provide not only insights to improve risk prediction for 22q11.2DS-associated SCZ but genetic support of drug targets for treatment in this cohort. In addition to rare coding variants, we identified SCZ associations of rare and low-frequency variants in enhancer regions of modifier genes active in the prefrontal cortex (PFC), suggesting that their gene regulatory activities in the PFC are involved in the development of SCZ. Finally, our gene expression analysis revealed a concurrent enrichment of modifier and 22q11.2-connected genes in gene coexpression modules localized to the cortical region from late infancy to young adulthood. This result suggested that modifiers are involved in biological pathways for the postnatal cortical development perturbed by 22q11.2 deletion.

Earlier studies of 22q11.2 genes have provided important biological insights into the potential impact of 22q11.2 deletion on brain development. For example, 22q11.2 genes were found to be enriched with spatiotemporal PPIs during childhood that may be driven by the pathologies of associated brain disorders[30]. Also, SCZ-associated neurodevelopmental modules implicated by common risk variants in idiopathic SCZ were found to contain many genes in the brain developmental PPI network of 22q11.2 genes, especially *SEPT5*, *PI4KA*, and *SNAP29*[26]. In our study, we used a conceptionally different, bottom-up approach to first find gene coexpression modules associated with SCZ in 22q11.2DS in brain development implicated by modifier genes and then investigated the connection between those modules and 22q11.2 genes. Notably, we uncovered such gene coexpression modules in childhood that were enriched with 22q11.2-connected genes. In addition to *SEPT5* and *PI4KA*, our analysis suggested that the enrichment was driven by other genes especially *COMT* and *SLC25A1*. *COMT* is one of the most widely studied 22q11.2 genes for SCZ due to its role in degradation of dopamine[31,32]. *SLC25A1* is a mitochondrial protein whose interactome was recently found to participate in synaptic function and was altered in SCZ patients with 22q11.2DS[33,34]. Overall, our study confirms results from previous ones and provides additional insights specific to SCZ in 22q11.2DS.

Previous studies of rare coding variants in idiopathic SCZ were limited to extremely rare and highly deleterious variants due to high genetic heterogeneity and selective pressure in this disorder[28,35]. We were able to demonstrate aggregate effects of rare coding variants on SCZ in this 22q11.2DS cohort, despite the relatively moderate sample size, likely for three reasons. First, our analytic framework incorporates a gene network and phenotypes (through IGSP) to improve prioritization of risk genes. Second, the cohort shares the same disease-causing copy number variant – i.e., 22q11.2 deletion – and thus likely also shares certain modifier risk variants. Third, SCZ risk variants among carriers of 22q11.2 deletion tend to have larger effect sizes according to the liability model. Modifier genes by nature are candidate drug targets to treat 22q11.2DS-associated SCZ. Of note, modifier genes that we identified include targets of antipsychotic drugs (e.g., *DRD1*, *DRD2*, and *CACNA1C*) and, especially, ones that have not been well-studied (e.g., *C3* and *OPRK1*). This highlights the potential of studying modifier rare coding variants in 22q11.2DS to identify therapeutic targets for SCZ in the general population.

The recent SCHEMA study implicated ultra-rare coding variants (URVs) in 10 genes as conferring substantial risk for SCZ (odds ratios of 3–50, $P < 2.14E-06$) and 32 genes at a false discovery rate of <5%[28]. Among the latter, *NR3C2* also appeared as a modifier risk gene in our study. Two reasons may account for the lack of more substantial overlap. First, we tested SCZ association based on the weighted burden of all nonsynonymous rare variants in a gene, not limited to ultra-rare coding variants. Second, many ultra-rare coding variants with large effect sizes may be too rare to be observed in a study cohort with hundreds of subjects. Nevertheless, those genes implicated by ultra-rare coding variants are also highly relevant to synapse functions, suggesting a convergence of disease mechanisms at the pathway level regardless whether the disease is part of a syndrome.

Having shown that modifier genes are enriched in modules associated with nervous system development and cell migration, we considered whether the mouse phenotype data

integration could potentially add scoring bias. As per the design of the IGSP algorithm, genes with certain phenotypes in human tend to have higher phenotype scores only if stronger gene association signals are enriched among genes whose mouse orthologs have annotations of relevant phenotypes. Therefore, although certain types of mouse knockout phenotypes (e.g., of the nervous system) are better studied and thus probably more complete than others, such a research variation does not introduce scoring bias in IGSP since the distribution of gene association signals is independent of the degree of completeness of mouse phenotype annotation. On the other hand, the power of our integrative statistical framework will be reduced when the disease and related phenotypes are less studied in mouse models. While the moderate sample size is no doubt a limiting factor of this study, this problem is being actively addressed by the Consortium through continuously recruiting more 22q11.2DS patients.

In additional to 22q11.2 deletion and SCZ-associated common variants, modifier rare variants also contribute to the disease risk of syndromic SCZ. Although their aggregate effects may not be directly significant to idiopathic SCZ, given the genetic heterogeneity of the disease in general, the identification and analysis of modifier rare variants is complementary to that of highly pathogenic ultra-rare variants in SCZ of the general population. For example, our study shows how rare variants across genes important in brain development collectively contribute to substantial risk in syndromic SCZ, which shares similar clinal features with idiopathic SCZ. This sheds light on important biological processes in the etiology of the disease and highlights the value of studying SCZ using 22q11.2DS as a model.

METHODS

22q11.2DS cohort and rare variants

The original WGS data consisted of 519 unrelated 22q11.2DS patients recruited across 22 international sites and underwent stringent measures of quality control[9] (**Supplementary note**). The study complies with all relevant ethical regulations and was approved by local institutional research ethics boards and have informed consent from all studied subjects. Among 519 22q11.2DS patients, there were 259 SCZ cases who have been diagnosed at any age by a stringent case consensus procedure and 260 controls who had no history of any psychotic illness when assessed at age ≥ 25 years. To maximize the sample size with a homogeneous genetic background, our study cohort was the subcohort of European ancestry that included 223 SCZ cases and 233 controls (**Supplementary note**). The largest subcohort of minorities, consisting of 31 subjects of Hispanic ancestry, was used to investigate whether modifier genes identified in the study cohort were enriched with modifier rare variants. WGS was performed with the Illumina pipeline. Sequence alignment was carried out with PEMapper to map WGS reads to the human genome build hg38. Variants were called with PECaller[9] (**Supplementary note**). Rare variants were defined as variants with AAF < 1% in the study cohort. Rare variants with a missing genotype rate > 0.01 in the study cohort were excluded from the analyses. This study

focused on autosomal rare variants outside the 22q11.2 deletion region; variants in the 22q11.2 deletion region were not analyzed.

Variant annotation method

We identified coding variants using CADD annotation (v.1.6, 'CodingTranscript'). In our study, different types of coding variants were defined by the Ensembl Variant Effect Predictor (VEP) as part of CADD: nonsynonymous variants are coding variants not annotated as 'synonymous' variants, while loss-of-function variants are either 'stop_gained' or 'frameshift' variants.

Modifier rare coding variants

The statistical framework that we used to identify and analyze modifier rare coding variants included sample bootstrapping, risk gene prioritization based on rare coding variants (the IGSP method), modifier risk calculation and evaluation to examine the aggregate effects of rare coding variants on SCZ risk (**Fig. 1**).

First, we bootstrapped the study cohort to randomly generate 500 target subcohorts and their paring discovery subcohorts. In each iteration, we randomly selected $n_t/2$ cases and $n_t/2$ controls from the study cohort to create the target subcohort; the remaining $223-n_t/2$ cases and $233-n_t/2$ controls in the study cohort constituted its corresponding discovery subcohort. n_t determines not only the sample size of target subcohorts and thus the rank resolution (see details below) but also the sample size of the corresponding discovery subcohorts and thus the power for prioritizing risk genes. In this study, we set $n_t = 50$ and showed that out result was not sensitive to selection of n_t within a reasonable range (**Supplementary Fig. 12**). Our sample bootstrapping procedure ensured that every individual in the study cohort was included in at least one target subcohort. Using each discovery subcohort, we applied IGSP to prioritize genes for risk modification based on the SCZ association of rare coding variants weighted by orthogonal information contained in the gene network and mouse phenotypes. Specifically, we collected gene association signals of each protein-coding gene outside the 22q11.2 deletion region by applying the weighted burden test (using the R package SKAT) to rare nonsynonymous variants with their CADD scores (v.1.6). The test used the SCZ status as the phenotype and included as covariates sex and the top 10 principal components to account for the subpopulation structure, which were obtained using PLINK (v.1.9) based on common variants (MAF > 0.05). The output association *P*-values were then used as input to IGSP to score and rank genes.

Next, we defined a modifier risk score for each individual in the target subcohorts based on rare nonsynonymous variants in the top n putative modifier genes:

$$\sum_{i=1}^{n}\left(\frac{d_i}{v_i}\sum_{j=1}^{v_i} c_{ij}\cdot a_{ij}\right) \quad (1)$$

in which d_i is the risk direction of gene i, taking on the value of 1 or −1 if the direction of variant burden is on SCZ cases or controls, respectively, v_i the number of rare nonsynonymous variants in gene i, c_{ij} and a_{ij} the CADD score and the number of alternative alleles of rare nonsynonymous variant j in gene i, respectively. For each putative modifier gene (one of the top n genes scored by IGSP), the equation calculates the weighted sum of alternative alleles for rare nonsynonymous variants in the gene weighted by the corresponding CADD scores. To avoid scoring bias to large genes and genes harboring a high density of rare coding variants, we performed a normalization process in which the weighted sum was divided by the number of rare nonsynonymous variants for the same gene observed in the corresponding discovery subcohort.

Finally, we evaluated whether the MRS explained the variance in SCZ status in the target subcohorts. Since MRS across different target subcohorts may involve different top scoring genes and thus are not directly comparable, we ranked subjects in each target subcohort in ascending order of their MRS and summarized each subject's modifier risk by taking the average of his MRS ranks over target subcohorts. The relationship between the summarized MRS and SCZ status was evaluated using logistic regression including as covariates sex and the top 10 principal components for correcting population substructure. Subjects may appear in different target subcohorts, which could introduce uncertainty to average MRS rank. We used another bootstrapping procedure to evaluate rare variant SCZ risk characterized by the average MRS rank. Briefly, in each bootstrap replicate, we calculated the variance of SCZ status explained by the average MRS rank from 20 random observations of each subject (i.e., 20 random target subcohorts including the subject) (**Supplementary Fig. S13**). The estimated variance of SCZ status explained by the average MRS rank and 99% confidence interval are derived from 201 replicates (median and the range after removing the first and last estimates after sorting, respectively).

We started the process with the MRS based on the top 10 genes and repeated the process by including the next 10 genes until the top 500 genes were examined. Overall, 50 tests are carried out to test relationship between MRS and SCZ status; however, the tests are correlated because the tested genes are highly overlapped. Therefore, we used the minimal P-value test[36] to calculate the corrected P-value of the most significant association in 50 tests. First, we calculated e, the effective number of independent tests as :

$$e = M - \sum_{i=1}^{M} [I(\lambda_i > 1)(\lambda_i - 1)] , \quad (2)$$

in which M equals the number of tests (50), λs are the eigenvalues of the $M \times M$ correlation matrix of the P-values of M tests, and I is an indicator function. The P-value correlation matrix can be calculated based on the Pearson correlation coefficient between the vectors of P-values from the 201 bootstrap replicates. Next, given e and the lowest P-value among M tests ($P_{\min}$), the corrected P-value can be calculated as:

$$P = 1 - (1 - P_{\min})^e . \quad (3)$$

Common polygenic risk for idiopathic SCZ

We calculated common polygenic risk of idiopathic SCZ for individuals with 22q11.2DS based on the PRS analysis using PRSice-2[37,38]. We first downloaded the summary statistics of the idiopathic SCZ GWAS[27] from the Psychiatric Genomics Consortium (PGC). We selected common SNPs in our 22q11.2DS cohort (internal MAF > 5%) and carried out LD clumping if they were within 250 kbps and $R^2 > 0.1$. Next, we used 19 *P*-value thresholds – 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.01, 1E-3, 1E-4, 1E-5, 1E-6, 1E-7, 1E-8, 1E-9, and 1E-10 – to select SNPs for PRS scoring after clumping. Sex and the top 10 principal components for population substructure correction were used as covariates in PRSice-2.

Cross-analysis of SCZ risk from common and rare coding variants

We analyzed SCZ status using four logistic regression models with (1) SCZ PRS and average MRS rank as predictor variables, and sex and top 10 principal components for population substructures as covariates, (2) SCZ PRS and covariates, (3) average MRS rank and covariates, and (4) covariates only. The independent risk component exclusively from common and rare coding variants, respectively, can be evaluated across these models. The combined genetic risk scores that characterize the overall genetic risk from both common and rare coding variants for SCZ in 22q11.2DS can be derived by applying coefficients from the logistic regression model (1).

Gene-set enrichment analysis

We systematically investigated gene sets enriched with 110 modifier genes using both hypothesis-driven and hypothesis-free approaches. In the hypothesis-driven approach, we investigated the enrichment of the modifier genes in 10 gene sets that were commonly used in SCZ studies: synaptic genes, FMRP target, postsynaptic density from human neocortex (hPSD), neurotransmitter system, presynaptic genes, calcium channel activity, developmental disorder, loss-of-function intolerant genes, and missense constrained genes (**Supplementary Table S13**). The first eight gene sets are relevant to SCZ biology and the last two were enriched with rare variants in idiopathic SCZ[35]. In the hypothesis-free approach, we investigated the enrichment of modifier genes in 7,481 gene sets annotated with gene ontology (GO) terms for different biological processes from the Molecular Signatures Database (v.7.4)[39,40]. For each gene set, we considered only the subset of genes that was scored by IGSP. We used Fisher's exact test to assess the significance of enrichment of modifier genes for a gene set, using 8,028 genes scored by IGSP (the full integration requires the phenotype annotation of mouse gene knockouts, which cover only half of human gene homologs) as the background. Bonferroni multiple test correction was used to correct *P*-values for 10 and 7,481 tests performed in the two approaches, respectively.

To investigate whether a gene coexpression module was enriched with modifier genes, we used Fisher's exact test and considered only genes in modules scored by IGSP (8,028 genes as the background). To investigate whether a gene coexpression module was enriched with

22q-SB-PPI genes or cell-marker genes[41], we used Fisher's exact test with all 21,196 coding genes as the background.

Exon expression of modifier genes

We used RNA-seq gene expression data at the exon level from BrainSpan (https://www.brainspan.org/static/download.html)[42] to identify regions (exons) of modifier genes with general and specific expression during brain development, defined as exons with median RPKM ≥ 1 and < 1, respectively, across all samples cataloged in BrainSpan regardless of brain regions or age.

Regulatory regions of modifier genes

We investigated noncoding rare variants in four different types of regulatory regions of modifier genes: promoters, brain-active enhancers, transcriptional regulator binding sites in neurons (neuron-TRBS), and DNase I hypersensitive sites in neural tissues (neural-DHS). We defined a gene's promoter as the genomic region within 500 bp of its representative transcription start site based on FANTOM CAGE data[43] and obtained its brain-active enhancer(s) (in the PFC) from the gene regulatory network generated by PsychENCODE[44]. We collected neuron-TRBS and neural-DHS as potential regulatory elements for a gene within its 50-kb upstream and 50-kb downstream regions from ReMap database (Homo sapiens; nonredundant peaks)[45] and an online repository (https://zenodo.org/record/3838751#.Y7yD3i-B2-x)[46], respectively.

SCZ association for a set of variants

We used the weighted burden test to derive rare variant associations at the gene level as input to IGSP because the direction of rare variant burden was required to calculate MRS. For other analyses that aimed to simply test SCZ association for a set of variants, we used weighted SKAT-O[47] (with CADD scores as the variant weights). When testing associations in the study cohort ($n = 456$) and discovery subcohorts ($n = 406$), we used sex and the top 10 principal components for the subpopulation structure correction as covariates. When testing associations in the Hispanic subcohort ($n = 31$), we used sex and the top 4 principal components as covariates.

Modifier genes and drug-induced differentially expressed genes

We tested whether modifier genes tend to be coexpressed with differentially expressed genes induced by SCZ drugs. Top differentially expressed genes were identified using drug-induced gene expression signatures from the Expanded CMap LINCS Resource 2020 (https://clue.io/data/CMap2020#LINCS2020)[48]. Briefly, for an antipsychotic drug, we identified in CMap n genes that most frequently appear in the list of top n most overexpressed and underexpressed genes, respectively, across different treatments of the drug. For this coexpression analysis, we focused on brain tissues between middle childhood and young adulthood. We considered two gene to be coexpressed if there is significant coexpression between them in any one of the 12 spatiotemporal combinations of four brain

regions and three stages of brain development (**Supplementary Tables S6**, **S7**, and **S8**) (see the next subsection). We assessed the degree of coexpression using a permutation test with 100,000 iterations of randomization, in each of which the degree of coexpression between 110 random IGSP-scored genes and drug-induced top differentially expressed genes was calculated to construct the null distribution.

Transcriptional analysis in brain development

Gene expression data were downloaded from PsychENCODE (http://development.psychencode.org/). Four brain regions and 11 overlapping stages of brain development defined in a previous study[25] were used to construct 44 spatiotemporal combinations (**Supplementary Tables S6, S7 and S8**). For each spatiotemporal combination, we measured high-confidence coexpression (1 or 0) based on a stringent threshold of Pearson correlation coefficient (the absolute value ≥ 0.85). We measured the degree of coexpression among modifier genes based on the sum of high-confidence coexpression between each pair of modifier genes. To assess the degree of coexpression among modifier genes in a spatiotemporal combination, we performed a permutation test with 100,000 iterations of randomization, in each of which the degree of coexpression among 110 random IGSP-scored genes was calculated to construct the null distribution. We measured the degree of coexpression with 22q11.2 genes for each IGSP-scored gene based on the number of high-confidence coexpression between the gene and each of 46 22q11.2 genes[49]. To investigate whether modifier genes tended to be coexpressed with 22q11.2 genes, we performed logistic regression to regress the status of modifier genes (i.e., 110 modifier genes and the remaining 7,918 IGSP-scored genes as non-modifier genes) on the degree of coexpression with 22q11.2 genes. Bonferroni correction was applied to 44 spatiotemporal combinations. To identify gene coexpression modules in a spatiotemporal combination, we performed the weighted gene co-expression network analysis (WGCNA), which uses a soft-thresholding method to better detect gene modules. First, to determine a soft thresholding power, we run "pickSoftThreshold" function to obtain the first soft thresholding power (starting from 2 to 15) of which the corresponding scale-free topology model fit $R^2 > 0.8$ without considering sign of coexpression (networktype = 'unsigned'). Next, we transformed the adjacency matrix of a coexpression network to topological overlap matrix (TOM; unsigned 'networkType' and 'TOMType') using the selected soft thresholding power and clustered genes based on the corresponding dissimilarity as the distance measure. The gene clustering was based on a hierarchical clustering function 'flashClust' (the 'average' method). Last, we determined gene modules by running 'cutreeDynamic' function (the 'tree' method) with a minimum module size = 100 genes. Genes in a coexpression network not classified into any modules (i.e., grey genes) were excluded from our module analyses.

Spatiotemporal brain-PPI network of 22q11.2 genes

We defined a spatiotemporal brain protein-protein interaction network of 22q11.2 (22q-SB-PPI) genes for a spatiotemporal combination as genes that have spatiotemporal brain protein-protein interactions with at least one 22q11.2 gene. We determined a spatiotemporal brain protein-protein interaction between two genes if they satisfied two

conditions: first, their protein products physically interacted according to MIST (v.5.0, Homo sapiens)[50]; second, both genes are transcribed in a specific spatiotemporal combination and have evidence of coexpression at transcriptional level (i.e., the absolute value of Pearson correlation coefficient ≥ 0.7).

DATA AVAILABILITY

This study is a secondary data analysis of the whole genome sequencing data of SCZ in 22q11.2DS generated by the International 22q11.2DS Brain and Behavior Consortium (IBBC). All summary statistics for the SCZ association of rare coding variants in the IBBC 22q11.2DS cohort are available at http://zdzlab.einsteinmed.org/1/sz-22q.html. All predicted modifier genes are available in the supplementary material. Due to privacy concerns for our research participants, individual-level genetic data from the IBBC study of SCZ in 22q11.2DS are not publicly available; however, access to anonymized data can be requested from a qualified academic investigator to the IBBC Executive Committee, providing the data transfer is approved by the Institutional Review Board and regulated by a material transfer agreement.

CODE AVAILABILITY

IGSP is publicly available at https://zenodo.org/record/1034362#.X-JWQNgzY2w. All other software used in our analyses is open source and described in Methods.

ACKNOWLEDGEMENTS

This work was supported by funding to ZDZ from NIH Grants U01 MH101720, R01 AG057909, U19 AG056278, R01 AG061521, R01 AG057706, RF1 AG057341, R01 AG061155, P01 AG047200, P01 AG017242, P30 AG038072, and a Career Scientist Award from the Irma T. Hirschl Trust.

COMPETING INTERESTS

All authors declare no competing interests.

REFERENCES

1. Edelmann, L., Pandita, R.K. & Morrow, B.E. Low-copy repeats mediate the common 3-Mb deletion in patients with velo-cardio-facial syndrome. *Am J Hum Genet* **64**, 1076-86 (1999).
2. Shaikh, T.H. *et al.* Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum Mol Genet* **9**, 489-501 (2000).
3. Tezenas Du Montcel, S., Mendizabai, H., Ayme, S., Levy, A. & Philip, N. Prevalence of 22q11 microdeletion. *J Med Genet* **33**, 719 (1996).
4. Botto, L.D. *et al.* A population-based study of the 22q11.2 deletion: phenotype, incidence, and contribution to major birth defects in the population. *Pediatrics* **112**, 101-7 (2003).
5. Murphy, K.C., Jones, L.A. & Owen, M.J. High rates of schizophrenia in adults with velo-cardio-facial syndrome. *Arch Gen Psychiatry* **56**, 940-5 (1999).
6. Fung, W.L. *et al.* Elevated prevalence of generalized anxiety disorder in adults with 22q11.2 deletion syndrome. *Am J Psychiatry* **167**, 998 (2010).
7. Bassett, A.S. *et al.* Rare Genome-Wide Copy Number Variation and Expression of Schizophrenia in 22q11.2 Deletion Syndrome. *Am J Psychiatry* **174**, 1054-1063 (2017).
8. Gur, R.E. *et al.* A neurogenetic model for the study of schizophrenia spectrum disorders: the International 22q11.2 Deletion Syndrome Brain Behavior Consortium. *Mol Psychiatry* **22**, 1664-1672 (2017).
9. Cleynen, I. *et al.* Genetic contributors to risk of schizophrenia in the presence of a 22q11.2 deletion. *Mol Psychiatry* (2020).
10. Sumitomo, A. *et al.* A mouse model of 22q11.2 deletions: Molecular and behavioral signatures of Parkinson's disease and schizophrenia. *Sci Adv* **4**, eaar6637 (2018).
11. Davies, R.W. *et al.* Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. *Nat Med* **26**, 1912-1918 (2020).
12. Merico, D. *et al.* Whole-Genome Sequencing Suggests Schizophrenia Risk Mechanisms in Humans with 22q11.2 Deletion Syndrome. *G3 (Bethesda)* **5**, 2453-61 (2015).
13. Michaelovsky, E. *et al.* Risk gene-set and pathways in 22q11.2 deletion-related schizophrenia: a genealogical molecular approach. *Transl Psychiatry* **9**, 15 (2019).
14. Lin, J.R., Zhang, Q., Cai, Y., Morrow, B.E. & Zhang, Z.D. Integrated rare variant-based risk gene prioritization in disease case-control sequencing studies. *PLoS Genet* **13**, e1007142 (2017).
15. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894 (2019).
16. Wang, Q. *et al.* The psychiatric disease risk factors DISC1 and TNIK interact to regulate synapse composition and function. *Mol Psychiatry* **16**, 1006-23 (2011).
17. Kirov, G. *et al.* Neurexin 1 (NRXN1) deletions in schizophrenia. *Schizophr Bull* **35**, 851-4 (2009).

18. Bhat, S. *et al.* CACNA1C (Cav1.2) in the pathophysiology of psychiatric disease. *Prog Neurobiol* **99**, 1-14 (2012).
19. Fernandes, B.S. *et al.* Peripheral brain-derived neurotrophic factor (BDNF) as a biomarker in bipolar disorder: a meta-analysis of 52 studies. *BMC Med* **13**, 289 (2015).
20. Kishi, T., Yoshimura, R., Ikuta, T. & Iwata, N. Brain-Derived Neurotrophic Factor and Major Depressive Disorder: Evidence from Meta-Analyses. *Front Psychiatry* **8**, 308 (2017).
21. Shibata, T. *et al.* The alteration of hypoxia inducible factor-1 (HIF-1) and its target genes in mood disorder patients. *Prog Neuropsychopharmacol Biol Psychiatry* **43**, 222-9 (2013).
22. Brosch, K. *et al.* Reduced hippocampal gray matter volume is a common feature of patients with major depression, bipolar disorder, and schizophrenia spectrum disorders. *Mol Psychiatry* **27**, 4234-4243 (2022).
23. Bassett, A.S. *et al.* Clinically detectable copy number variations in a Canadian catchment population of schizophrenia. *J Psychiatr Res* **44**, 1005-9 (2010).
24. Li, M. *et al.* Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* **362**(2018).
25. Willsey, A.J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007 (2013).
26. Forsyth, J.K. *et al.* Synaptic and Gene Regulatory Mechanisms in Schizophrenia, Autism, and 22q11.2 Copy Number Variant-Mediated Risk for Neuropsychiatric Disorders. *Biol Psychiatry* **87**, 150-163 (2020).
27. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502-508 (2022).
28. Singh, T. *et al.* Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* **604**, 509-516 (2022).
29. Iyegbe, C.O. & O'Reilly, P.F. Genetic origins of schizophrenia find common ground. *Nature* **604**, 433-435 (2022).
30. Lin, G.N. *et al.* Spatiotemporal 16p11.2 protein network implicates cortical late mid-fetal brain development and KCTD13-Cul3-RhoA pathway in psychiatric diseases. *Neuron* **85**, 742-54 (2015).
31. Egan, M.F. *et al.* Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proc Natl Acad Sci U S A* **98**, 6917-22 (2001).
32. Williams, H.J., Owen, M.J. & O'Donovan, M.C. Is COMT a susceptibility gene for schizophrenia? *Schizophr Bull* **33**, 635-41 (2007).
33. Gokhale, A. *et al.* Systems Analysis of the 22q11.2 Microdeletion Syndrome Converges on a Mitochondrial Interactome Necessary for Synapse Function and Behavior. *J Neurosci* **39**, 3561-3581 (2019).
34. Li, J. *et al.* Mitochondrial deficits in human iPSC-derived neurons from patients with 22q11.2 deletion syndrome and schizophrenia. *Transl Psychiatry* **9**, 302 (2019).
35. Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat Neurosci* **19**, 1433-1441 (2016).
36. Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71-76 (2019).

37. Euesden, J., Lewis, C.M. & O'Reilly, P.F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466-8 (2015).
38. Choi, S.W. & O'Reilly, P.F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**(2019).
39. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
40. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739-40 (2011).
41. Zhang, X. *et al.* CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res* **47**, D721-D728 (2019).
42. BrainSpan Atlas of the Developing Human Brain [Internet]. Available at http://brainspan.org. (2011).
43. Consortium, F. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462-70 (2014).
44. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**(2018).
45. Hammal, F., de Langen, P., Bergon, A., Lopez, F. & Ballester, B. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res* **50**, D316-D325 (2022).
46. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244-251 (2020).
47. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224-37 (2012).
48. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452 e17 (2017).
49. Zinkstok, J.R. *et al.* Neurobiological perspective of 22q11.2 deletion syndrome. *Lancet Psychiatry* **6**, 951-960 (2019).
50. Hu, Y. *et al.* Molecular Interaction Search Tool (MIST): an integrated resource for mining gene and protein interaction data. *Nucleic Acids Res* **46**, D567-D574 (2018).

FIGURE LEGEND

Figure 1. Statistical framework to examine the aggregate effects of modifier rare coding variants on SCZ. We first randomly split the study cohort of SCZ in 22q11.2DS into a target subcohort with 25 SCZ cases and 25 controls and a discovery subcohort with the remaining subjects. We then calculated the relative SCZ risk of subjects in a target subcohort based on their rare coding variants in genes prioritized in the corresponding discovery subcohort (MRS, Equation 1). We repeated the calculation 500 times and evaluated MRS for SCZ across subcohorts based on the average MRS rank.

Figure 2. SCZ risk from rare coding variants in 22q11.2DS. a. SCZ status and MRS. The plot shows the fraction of variance in SCZ status in target subcohorts explained by the MRS (*y*-axis) based on rare nonsynonymous variants in the top *n* genes prioritized by IGSP scoring with a full integration (using gene network and phenotypes) in discovery subcohorts (*x*-axis). D(+/−) denotes that a higher MRS corresponds to a higher/lower risk of SCZ. The vertical bars denote a 99% confidence interval. **b.** Odds ratio. To directly evaluate whether MRS quantifies the SCZ risk, for each of 500 target subcohorts, we calculated odds ratio of SCZ patients using top 50% MRS (based on the top 110 genes in the discovery subcohort) as the exposure. The *x*- and *y*-axis show the number of observed rare nonsynonymous variants and the number of affected genes, respectively, in a target subcohort. **c.** SCZ status vs. MRS calculated in four different ways. S1 is for using Equation 1, while S2, S3, and S4 denote using Equation 1 without c_{ij} for considering the predicted functionality of rare coding variants (i.e., CADD-score weighting), without v_i for normalizing variant effects at the gene level, and without d_i for considering risk direction, respectively. **d.** SCZ status vs. MRS calculated for different types of rare coding variants.

Figure 3. Composition of genetic risk of SCZ in the study cohort. The horizontal bars show min-max normalized PRS, average MRS rank, and combined risk scores for 456 22q11.2DS patients with and without SCZ (i.e., cases and controls). The combined risk scores were calculated by applying coefficients of PRS and average MRS rank obtained from a logistic regression model (with the SCZ status as outcome and sex and top 10 principal components for population substructure correction as covariates). The same SCZ cases are connected by pink lines. Three SCZ cases with PRS lower than the median but with high MRS are highlighted by red connecting lines.

Figure 4. Gene-set enrichment analysis of modifier genes. a. Ten gene sets known to be highly relevant to SCZ. **b.** Gene sets for GO terms of biological processes. Of the 7,481 such gene sets, only 10 with the most significant enrichment are shown. Nominal enrichment *P*-values calculated by Fisher's exact tests are indicated. The error bars represent a 95% confidence interval. The ones in red denote significant enrichment (adjusted $P < 0.05$) after Bonferroni correction.

Figure 5. Transcriptomic analyses of modifier and 22q11.2 genes in brain development. a. Enrichment of coexpression among modifier genes in brain development. In each spatiotemporal combination, we tested whether modifier genes tended to be coexpressed using permutation tests. **b.** Enrichment of coexpression between modifier and

22q11.2 genes in brain development. In each spatiotemporal combination, we tested whether modifier genes tended to be coexpressed with 22q11.2 genes using logistic regression. **c-f.** Concurrent enrichment of modifier and 22q-SB-PPI genes in modules of P7-9/R1, P8-10/R1, P9-11/R1, and P10-11/R1. The heatmap shows the enrichment of cell markers among module genes for different cell types (only relevant cell types with at least 50 marker genes cataloged in CellMarker[41] were considered). N: Neuron. A: Astrocyte. M: Macrophage. O: Oligodendrocyte. **g.** Contribution of specific 22q11.2 genes to the enrichment of 22q-SB-PPI genes in $M3_{P7\text{-}9/R1}$, $M4_{P9\text{-}11/R1}$, and $M1_{P10\text{-}11/R1}$. One, two, and three asterisks denote Bonferroni adjusted $P < 0.05$, < 0.01 and < 0.001 for 138 tests (46 22q11.2 genes and 3 modules), respectively.