Systems biology analysis of human genomes points to key pathways conferring spina bifida risk

Vanessa Aguiar-Pulido[a], Paul Wolujewicz[a], Alexander Martinez-Fundichely[b,c], Eran Elhaik[d], Gaurav Thareja[e], Alice Abdel Aleem[f], Nader Chalhoub[f], Tawny Cuykendall[b,c], Jamel Al-Zamer[g], Yunping Lei[h], Haitham El-Bashir[g], James M. Musser[i,j], Abdulla Al-Kaabi[k], Gary M. Shaw[l], Ekta Khurana[b,c], Karsten Suhre[e], Christopher E. Mason[a,b,c], Olivier Elemento[a,b,c,m], Richard H. Finnell[h,n,o], and M. Elizabeth Ross[a,1]

[a]Center for Neurogenetics, Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY 10021; [b]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY 10065; [c]His Royal Highness Prince Alwaleed Bin Talal Bin Abdulaziz Al-Saud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY 10065; [d]Department of Biology, Lund University SE-221 00 Lund, Sweden; [e]Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Doha, Qatar; [f]Department of Neurology, Weill Cornell Medicine-Qatar, Doha, Qatar; [g]Rehabilitation Medicine, Hamad Medical Corporation, Doha, Qatar; [h]Department of Molecular and Cellular Biology, Center for Precision Environmental Health, Baylor College of Medicine, Houston, TX 77030; [i]Department of Pathology and Genomic Medicine, Houston Methodist Research Institute, Houston, TX 77030; [j]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY 10065; [k]Sidra Medical and Research Center, Weill Cornell Medicine-Qatar, Doha, Qatar; [l]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94305; [m]Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10021; [n]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030; and [o]Department of Medicine, Baylor College of Medicine, Houston, TX 77030

Edited by Patrick Stover, Texas A&M AgriLife, College Station, TX; received April 12, 2021; accepted October 20, 2021

Spina bifida (SB) is a debilitating birth defect caused by multiple gene and environment interactions. Though SB shows non-Mendelian inheritance, genetic factors contribute to an estimated 70% of cases. Nevertheless, identifying human mutations conferring SB risk is challenging due to its relative rarity, genetic heterogeneity, incomplete penetrance, and environmental influences that hamper genome-wide association studies approaches to untargeted discovery. Thus, SB genetic studies may suffer from population substructure and/or selection bias introduced by typical candidate gene searches. We report a population based, ancestry-matched whole-genome sequence analysis of SB genetic predisposition using a systems biology strategy to interrogate 298 case-control subject genomes (149 pairs). Genes that were enriched in likely gene disrupting (LGD), rare protein-coding variants were subjected to machine learning analysis to identify genes in which LGD variants occur with a different frequency in cases versus controls and so discriminate between these groups. Those genes with high discriminatory potential for SB significantly enriched pathways pertaining to carbon metabolism, inflammation, innate immunity, cytoskeletal regulation, and essential transcriptional regulation consistent with their having impact on the pathogenesis of human SB. Additionally, an interrogation of conserved noncoding sequences identified robust variant enrichment in regulatory regions of several transcription factors critical to embryonic development. This genome-wide perspective offers an effective approach to the interrogation of coding and noncoding sequence variant contributions to rare complex genetic disorders.

neural tube defects | myelomeningocele | whole-genome sequence | rare variant enrichment | pathway analysis

The neural tube defect (NTD) spina bifida (SB), among the debilitating but survivable malformations in live births, is due to failed embryonic neural tube closure. Together, SB and the nonviable NTD anencephaly have a global prevalence ranging from one in 3,000 to one in 100 (1). Decades of clinical and animal model investigations have indicated that SB comprises a complex genetic disorder, requiring at least one (and probably several) of many genetic alterations or gene–environment interactions for neurulation to fail (2, 3). NTD-causing mutations have been reported in more than 250 mouse genes (4, 5), which has since grown to over 400 mutant genes currently listed in the Mouse Genome Informatics database, further underscoring the complex genetic origins of the disorder. Genetic heritability of human SB, or the proportion of cases that are attributable to genetic alteration, is estimated to be as much as 70% (6).

Maternal periconceptional supplementation with folic acid (vitamin B9) can reduce the occurrence of SB in offspring by as much as 70% in some populations (7–9). Despite folate supplementation campaigns and fortification of the US food supply since 1998, SB prevalence rates have only dropped 30%, suggesting that most benefits from folic acid have been achieved. Other agents such as vitamin B12, methionine, or inositol show some promise for effective prevention (10). However, the mechanisms through which these agents influence SB occurrence

Significance

Genetic investigations of most structural birth defects, including spina bifida (SB), congenital heart disease, and craniofacial anomalies, have been underpowered for genome-wide association studies because of their rarity, genetic heterogeneity, incomplete penetrance, and environmental influences. Our systems biology strategy to investigate SB predisposition controls for population stratification and avoids much of the bias inherent in candidate gene searches that are pervasive in the field. We examine both protein coding and noncoding regions of whole genomes to analyze sequence variants, collapsed by gene or regulatory region, and apply machine learning, gene enrichment, and pathway analyses to elucidate molecular pathways and genes contributing to human SB.

Author contributions: V.A.-P., J.M.M., C.E.M., R.H.F., and M.E.R. designed research; V.A.-P., P.W., E.E., N.C., T.C., and M.E.R. performed research; V.A.P., A.M.-F., E.E., G.T., A.A.A., N.C., T.C., J.A.-Z., Y.L., H.E.-B., A.A.-K., G.M.S., E.K., K.S., C.E.M., O.E., R.H.F., and M.E.R. contributed new reagents/analytic tools; V.A.-P., P.W., A.M.-F., E.E., G.T., Y.L., and O.E. analyzed data; and V.A.-P., P.W., E.E., J.M.M., G.M.S., O.E., R.H.F., and M.E.R. wrote the paper.

Competing interest statement. R.H.F. formerly held a leadership position with the now dissolved TeratOmic Consulting LLC. He also receives travel funds to attend editorial board meetings of the Journal of Reproductive and Developmental Medicine published out of the Red Hospital of Fudan University. E.E. consults for the DNA Diagnostics Center. P.S. and R.H.F. are coauthors on a 2020 paper resulting from an NIH workshop: Maruvada P et al., Knowledge gaps in understanding the metabolic and clinical effects of excess folates/folic acid: a summary, and perspectives, from an NIH workshop. Am J Clin Nutr. 2020 Nov 11;112(5):1390-1403. doi: 10.1093/ajcn/nqaa259. PMID: 33022704; PMCID: PMC7657327.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: mer2005@med.cornell.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2106844118/-/DCSupplemental.

Published December 16, 2021.

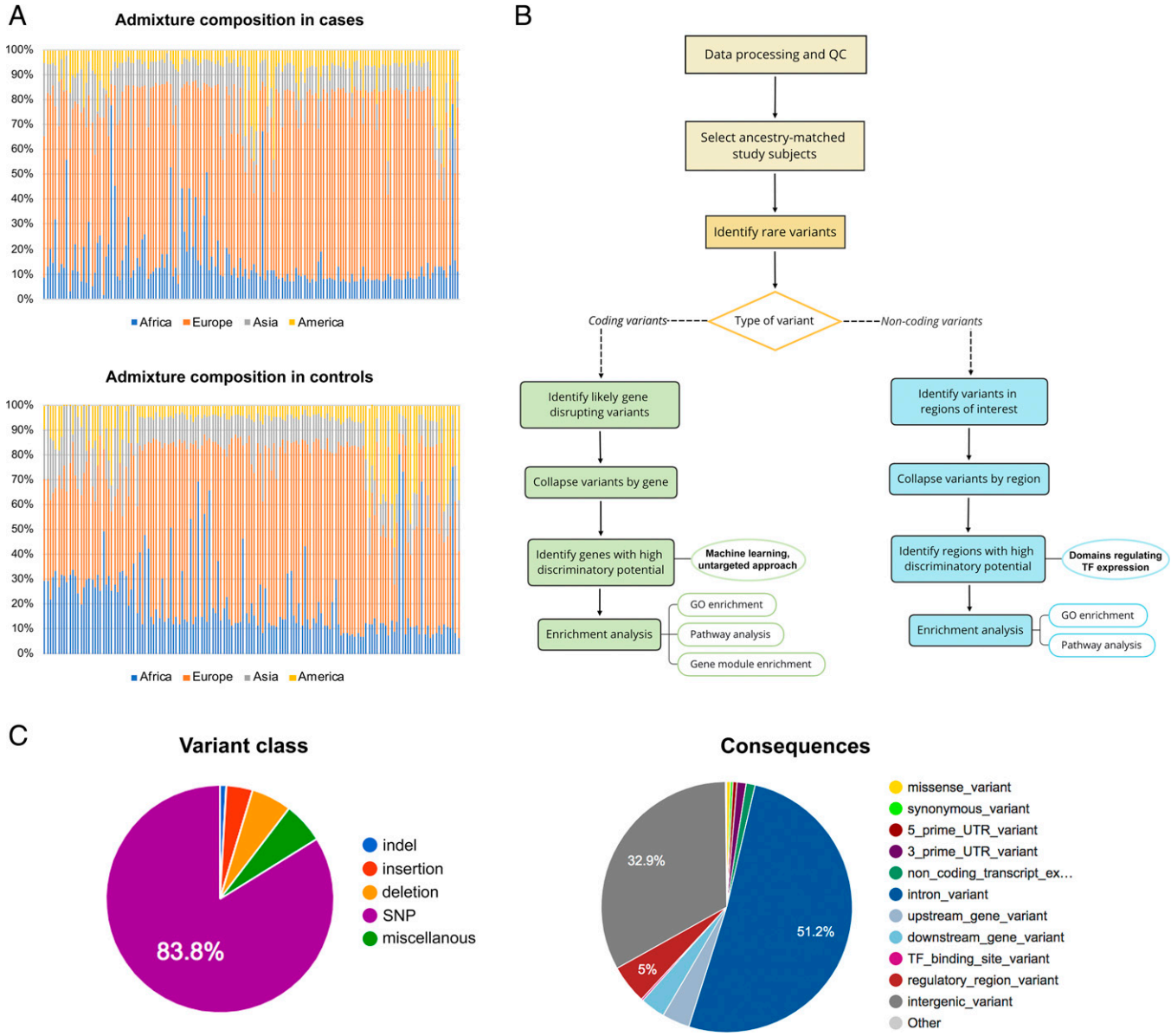

Fig. 1. WGS analysis overview. (*A*) The admixture composition of the ethnically diverse cohort of 149 SB cases and 149 ancestry-matched controls used in the analysis. For brevity, the nine gene pools were collapsed by continent. (*B*) The strategy used to interrogate WGS data. (*C*) The proportion of variants found in the cohort by type.

remains elusive, and based on mouse models, responses to supplements like folic acid vary with the genetic context (3, 11–14). Although powerful, the mouse is an imperfect surrogate for humans on several counts, among them intergenic regions that differ significantly from the human genome, with less species conservation than protein-coding regions. At present, it is not possible to identify maternal–fetal genotypes that indicate vulnerability to a teratogenic drug or toxin or to predict which preventive therapy will best ensure healthy pregnancy outcomes for individual couples. There is a pressing need to identify patterns of human SB genetic predisposition that could lead to better understanding individual prognosis, improved care of SB-afflicted children, and enhanced capabilities for birth defect prevention.

Next-generation sequencing offers increasing insight into risk factors for common complex genetic disorders including type II diabetes (1 in 10 in the United States) (15), schizophrenia (1 in 100) (16, 17), and autism spectrum disorders (1 in 59) (18, 19). However, less prevalent complex genetic disorders are particularly challenging, as they affect relatively small and globally diverse populations [e.g., in the United States, 1 in 3,000 for NTDs, 1 in 700 for orofacial clefts (20), and 1 in 140 for congenital heart disease (CHD) (21)], which requires pooling genetically diverse cohorts that may confound downstream analyses. Genetic studies (including genome-wide association studies, GWAS) of the more prevalent structural birth defects such as CHD have indicated that, while sequence variants that are common in human populations probably contribute to birth defects, they account for only a small proportion of genetic risk (21) so that GWAS will require thousands of subjects to identify common variants that increase risk for structural birth defects. Nevertheless, NTD cases display significant enrichment in rare variants (22, 23), suggesting that genes bearing rare variants will have stronger associations (greater effect sizes) and may be identifiable in smaller cohorts. Taking an approach distinct from a GWAS, this is a multicenter SB case-control study that

mounts a comprehensive, ancestry-matched whole-genome sequence (WGS) analysis from a systems biology perspective. This study seeks to identify pathways and biological functions that are disrupted in SB as reflected in their enrichment with genes or regulatory regions harboring rare, likely damaging mutations.

Results

We obtained WGS data of 310 individuals encompassing 157 SB cases and 153 controls. After quality control screening, which included the extent of genomic regions sequenced and average depth of coverage (30× for over 80% of the genome), the remaining samples were analyzed to identify the population substructure and optimize the case-control pairing to minimize stratification. For that, genetic ancestry was identified using a mixed admixture model, and only ancestry paired-matched cases and controls with paired-matched ancestry were included for further analysis (see *Materials and Methods* for details). The admixture proportions of the selected individual pairs are shown in Fig. 1*A*. Downstream computations were carried out on 298 individuals, including 149 cases and 149 nonmalformed controls. The mean sequencing depth of the samples was >30× regardless of their origin (i.e., venipuncture or newborn blood spots) (*SI Appendix*, Fig. S1). Variants were called using a standard pipeline (see *Materials and Methods*).

Decades of clinical and animal model investigations underlie the consensus that rare variants likely lead to SB (show higher penetrance) more often than common ones, and only a few examples are cited here (10, 24–27). Among the reasons, if common alleles were the main driver of SB occurrence, the condition should be more highly prevalent in the population. However, common variants do not necessarily equate with common phenotypes, and relatively common alleles may contribute to SB genetic risk. We therefore used a somewhat relaxed definition of rare, including for further analysis those variants with an allele frequency (AF) < 0.01 in any population from 1,000 Genomes (28), National Heart, Lung, and Blood Institute (NHLBI) and the Exome Sequencing Project (ESP) (29), and the Genome Aggregation Database (gnomAD) (30). Hence, from a total of 41,005,720 variants at the cohort level, 22,502,019 variants were retained for subsequent analyses. No statistically significant difference was found in the variant distributions between cases and controls (two-tailed Student's t test P value = 0.7197, *SI Appendix*, Fig. S2). Fig. 1*B* outlines the workflow for the genome-wide analyses carried out on the ancestry-matched samples reported in this manuscript. Fig. 1*C* provides a breakdown of the different types of variants found in the study cohort.

Coding Variant Analysis Supports Existing Literature and Identifies Pathways Involving Inflammation, Innate Immunity, and Cytoskeletal Regulation. We previously reported an increased burden of nonrecurrent, private, loss-of-function variants in genomes of NTD patients compared to controls, which was consistent across three cohorts with different ancestry (22). Herein, we extended the analysis of coding regions to include all rare likely gene disrupting (LGD) protein-coding variants (i.e., frameshift, nonsense, splice donor/acceptor, stop gain/loss, and missense predicted deleterious). Thus, of the 22.5 million rare variants in our cases and controls, 56,210 met criteria as LGD single-nucleotide variants (SNVs) and insertions-deletions (InDels) and were included in the current analysis. No statistically significant difference was found in the rare LGD variant distributions between cases and controls (two-tailed Student's t test P value = 0.6547, *SI Appendix*, Fig. S2). Traditional GWAS approaches typically involve finding an association between a variant and the disorder. However, with a limited sample size, this study is underpowered to carry out even a rare variant association analysis, which involves aggregating the effect of rare variants within a gene. Assuming minor allele frequencies (MAF) of 5%, over 3,000 cases will be required to reach statistical significance at the gene level at 80% power (see *Materials and Methods*). Unsurprisingly, rare variant aggregate association tests such as SNV-set (Sequence) Kernel Association Test (31) did not render any statistically significant results at the single-gene level after multiple testing correction.

Considering the complex genetic nature of SB and its relatively low prevalence, systems biology approaches are more appropriate to find statistically significant results after correcting for multiple hypothesis testing. Since the rare (or even private), potentially deleterious variants found in cases are likely to affect different genes within several common pathways or biological processes and functions, we surmised that a machine learning approach can help reduce the genomic search space. A total of 13,526 genes harbored the 56,210 LGD variants identified in our cohort, and these genes were further analyzed to determine which ones allowed distinguishing cases from controls (i.e., had high discriminatory potential). Embedded feature selection was employed to perform this process; thus, we selected a machine learning algorithm that pinpointed relevant genes during the learning process. Random Forest (RF) (32)—a machine learning technique which uses numerous decision trees—was employed to build a predictive model of SB utilizing the 13,526 genes that harbored LGD variants as input. The best, most optimized model was selected by comparing cross-validation results, and additionally, a completely separate subset of the data (hold-out dataset) was utilized to estimate the generalization error. Method performance was assessed by calculating the area under the receiver operating characteristic curve (AUROC) (33) (see *Materials and Methods*). The selected RF model, which encompassed 100 trees, was able to achieve an AUROC of 0.78 on the hold-out dataset, indicating that the model performs well on new, unseen data (*SI Appendix*, Fig. S3). A list of 439 genes were identified as relevant to distinguish cases from controls by this technique (Dataset S1). This gene list was then used for enrichment analyses in order to identify pathways, biological processes, molecular functions, and cellular components that were overrepresented.

Genes were classified into broad annotation categories as an overview of the biological processes that were affected (shown in *SI Appendix*, Fig. S4). Interestingly, out of the 439 genes with high discriminatory potential between cases and controls (i.e., SB discriminative genes), nine were differentially expressed in a previous transcriptome analysis of fetuses with NTDs (34). That small study performed genome microarray–based transcription profiling

Table 1. Genes found in this machine learning strategy to have high discriminatory potential between SB cases and controls that were previously found to be differentially expressed in human fetal NTD versus healthy control amniocytes

Gene	Expression up/down	Fold change (log2)	Adjusted P value
*CGAS**	+	2.82	0.02
GRIN2D†	+	3.15	0.00
*MYH11**	+	2.92	0.01
ODF3B	+	2.69	0.04
IVL	–	2.96	0.01
*LAMC2**	–	2.37	0.02
SLITRK6†	–	2.69	0.02
*USP2**	–	2.81	0.00
ZNF750	–	2.23	0.04

Amniocyte data (log2 fold changes and adjusted P values) reported by Nagy et al., 2006 (34).
*These differentially expressed genes are also found in significantly overrepresented pathways obtained in our analysis.
†These differentially expressed genes have been associated with neuronal synapse assembly and axon pathfinding.

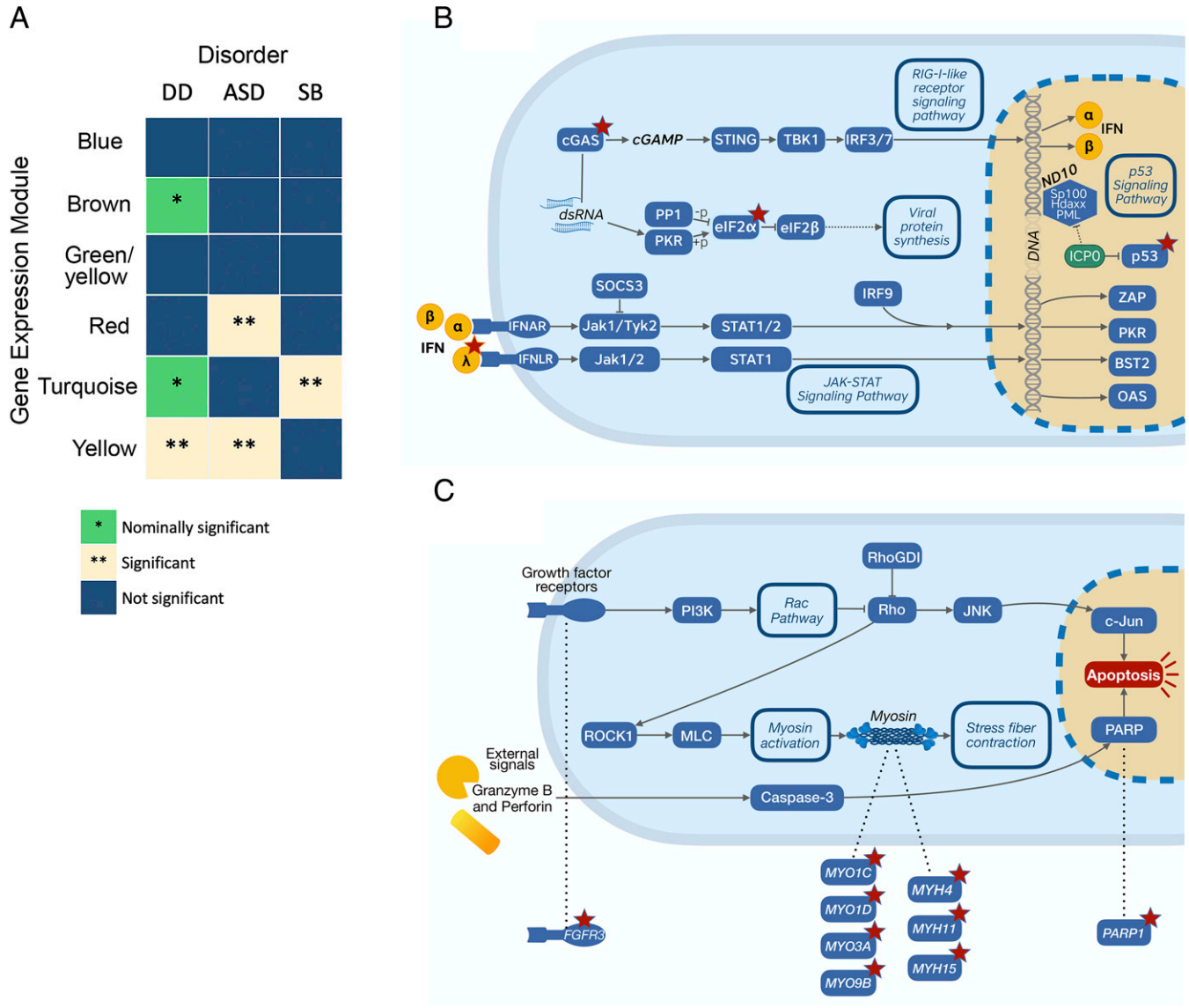

Fig. 2. The genes with the greatest potential to discriminate between SB cases and non-SB controls and their relationships in pathways. (*A*) The genes with high discriminatory potential to distinguish SB cases and controls significantly enrich an early progenitor class, gene coexpression module identified in a transcriptome WGCNA study of midgestation human cortex (35). The modules most highly enriched in rare variants found in individuals with developmental delay (DD, neuronal regulation module) or autism spectrum disorder (ASD, neuronal regulation and neurobehavior modules) (Walker et al.) are distinct from SB (this study, early neural progenitor proliferation module). (*B*) The pathways related to immunity are enriched with genes that contain LGD mutations in SB cases and impact the interferon arm of the HSV-1 pathway. (*C*) SB risk genes affect cytoskeletal regulation. The genes enriched with LGD variants in SB cases disrupt RhoGDI signaling and actin-myosin components of the cytoskeleton. Red stars in *B* and *C* indicate LGD-enriched genes.

of human fetal amniocyte-derived messenger RNA (mRNA) from pregnant women at 17 to 19 wk gestation, comprising seven NTD-affected pregnancies compared to five healthy controls (34) (Table 1). Next, we sought to determine whether any of the fetal cortical clusters of genes ("gene modules") identified in a previously published analysis based on human midgestational (weeks 14 to 21) RNA sequence data (35) were enriched in SB discriminative genes (see *Materials and Methods*). We found that the human fetal gene expression cluster referred to as the "turquoise module" by Walker et al. was the only one of six modules that was significantly enriched in genes with high discriminatory potential in our SB cases (adjusted $P < 0.02$) (Fig. 2*A*). This turquoise module was described (35) as enriched for specific brain cell types or brain-relevant Gene Ontology (GO) terms involving mitotic progenitors and cell division and therefore represents an early progenitor class.

A total of 20 relevant pathways were overrepresented in genes with the potential to discriminate between SB cases and controls (Table 2). Importantly, the pathways with greatest significance were those related to central metabolism (Carbon Metabolism and Cobalamin Transport and Metabolism, adjusted P value < 0.001). The variant-enriched genes within these pathways suggest lipid (fatty acid) and glucose metabolism as the aspects most affected in our cohort. This is particularly interesting in that the epidemiological data accumulating post-introduction of folic acid fortification into the US food supply indicates that the persistent risks for NTD may be largely attributable to the rise in obesity and diabetes (36–38). This finding provides strong evidence that the proposed approach is pinpointing relevant pathways.

Additional pathways associated with human SB encompass genes linked to innate immunity and inflammatory response

Table 2. Pathways enriched with genes of high discriminatory potential between SB cases and healthy controls

Pathway	Adjusted *P* value	Genes
Carbon metabolism	0.00081	*ADPGK, EHHADH, ACAT2, ASNS, ENO4, MDH2, H6PD, MMUT, PGD, TKTL2*
Cobalamin (Cbl, vitamin B12) transport and metabolism	0.00099	*ABCD4, CTRB2, MMUT, TCN1*
Glyoxylate and dicarboxylate metabolism	0.00358	*ACAT2, HYI, MDH2, MMUT*
Propanoate metabolism	0.00449	*EHHADH, ACAT2, LDHAL6B, MMUT*
Herpes simplex virus 1 infection	0.00685	*HLA-A, EIF2AK3, EIF2AK4, ZNF439, CGAS, ZNF283, ZNF160, ZNF616, ZNF8, ZNF790, ZNF233, ZNF850, TP53, ZNF708, ZNF273, ZNF682, ZNF814, ZNF562, ZNF736*
DNA damage	0.00885	*CIP2A, CAD, CDT1, BRCA1, PKMYT1, CCP110, POLR2A, MUTYH, CENPF, USP2, TTK, TDP1, TP53*
ECM–receptor interaction	0.01062	*FREM2, COL18A1, LAMC2, SV2C, ITGA8, TNR, RELN*
RhoGDI pathway	0.01321	*FGFR3, MYO3A, MYH11, MYH15, MYO1C, MYH4, MYO9B, MYO1D, PARP1*
Codeine and morphine metabolism	0.01358	*ABCC2, CYP2D6, SLCO1B3*
Sertoli-Sertoli cell junction dynamics	0.01605	*EPN3, MYO3A, MYH11, MYH15, MYO1C, MYH4, NPR1, SPTBN1, STX5, MYO9B, ITGA8, MYO1D, CGN, SPTB, CLDN6*
Homologous DNA pairing and strand exchange	0.01793	*EXO1, BRCA1, POLD3, RAD51D, RAD9B*
NAD metabolism, sirtuins, and aging	0.02171	*FOXO1, PARP1*
Pentose phosphate pathway	0.02427	*H6PD, PGD, TKTL2*
PERK regulates gene expression	0.02427	*EXOSC5, ASNS, EIF2AK3*
Interaction between L1 and ankyrins	0.02427	*SPTBN1, SCN5A, SPTB*
Cell cycle checkpoints	0.02468	*EXO1, BRCA1, PKMYT1, PSMA8, MCM10, RAD9B, MCM6, HIST3H3, TP53*
Valine, leucine, and isoleucine degradation	0.02575	*ACSF3, EHHADH, ACAT2, MMUT*
Amino acid transport across the plasma membrane	0.02638	*SLC7A7, SLC6A6, SLC6A12*
Aurora A signaling	0.02638	*DLGAP5, BRCA1, TP53*
Oncogene-induced senescence	0.02860	*ETS2, TNRC6B, TP53*

cascades. For example, the herpes simplex virus 1 (HSV-1) infection pathway (Table 2, adjusted $P = 0.00685$ and Fig. 2*B*) includes Cyclic GMP–AMP Synthase (*CGAS*), whose expression was increased in fetal cells from human NTD cases (Table 1). Interestingly, this gene is interferon inducible and part of innate immunity (39). The rare LGD variants of the analysis cohort also impact genes associated with three critical cascades: RIG-I–like receptor signaling, JAK–STAT signaling, and p53 signaling (Fig. 2*B*). Another relevant pathway implicated in human SB is associated with the response to DNA damage (adjusted $P = 0.00885$) and includes Ubiquitin Specific Peptidase 2 (USP2), which is required for Tumor Necrosis Factor alpha (TNF-alpha)–induced Nuclear Factor kappa B (NF-κB) signaling. *USP2* was also differentially expressed in the fetal cells from NTD cases shown in Table 1. Together, these pathways are consistent with previous data implicating immune responses in SB pathogenesis (40) and suggest fetal contributions beyond maternal factors in utero produce SB.

The extracellular matrix (ECM)–receptor interaction (Table 2, adjusted $P = 0.01062$), cytoskeletal regulation (Rho GDP-dissociation inhibitor [RhoGDI] pathway, Table 2, adjusted $P = 0.01321$ and Fig. 2*C*), and cell–cell signaling (Sertoli–Sertoli Cell Junction Dynamics, Table 2, adjusted $P = 0.01605$) pathways were also significantly impacted by rare, LGD variant–enriched genes in SB cases. Among these cascades, genes such as Laminin Subunit Gamma 2 (*LAMC2*) and Myosin Heavy Chain 11 (*MYH11*) were also found to be differentially expressed in NTD fetal cells (Table 1).

GO enrichment analysis identified biological processes to be statistically significant (Dataset S2). Among these are processes related to cellular/molecular transport (Intraciliary Anterograde Transport, adjusted $P = 0.00008$; Amino Acid Transmembrane Transport, adjusted $P = 0.00612$), cell migration and morphogenesis (Lateral Motor Column Neuron Migration, Positive Regulation of Trophoblast Cell Migration and Auditory Receptor Cell Morphogenesis, adjusted $P < 0.01$), the response to stress (EiF2alpha Phosphorylation in Response to Endoplasmic Reticulum Stress, adjusted $P = 0.00081$; Negative Regulation of Translational Initiation in Response to Stress, adjusted $P = 0.00315$), mitochondrial and nuclear DNA (Mitochondrial DNA Repair and Mitochondrial DNA Metabolic Process,

Table 3. TF genes whose regulatory regions are significantly enriched with rare variants

TF	Full name	Adjusted *P* value	Coordinates
ZNF274	Zinc finger protein 274	1.64E-11	GeneHancer
RFX5	Regulatory factor X5	5.25E-05, 7.56E-08	hESC CTCF loops (naïve and primed)
MAX	MYC associated factor X	6.37E-05, 0.018	CTCF loop conserved across tissues and hESC (naïve)
JUND	JunD proto-oncogene, AP-1 TF	0.026	hESC CTCF loops (primed)

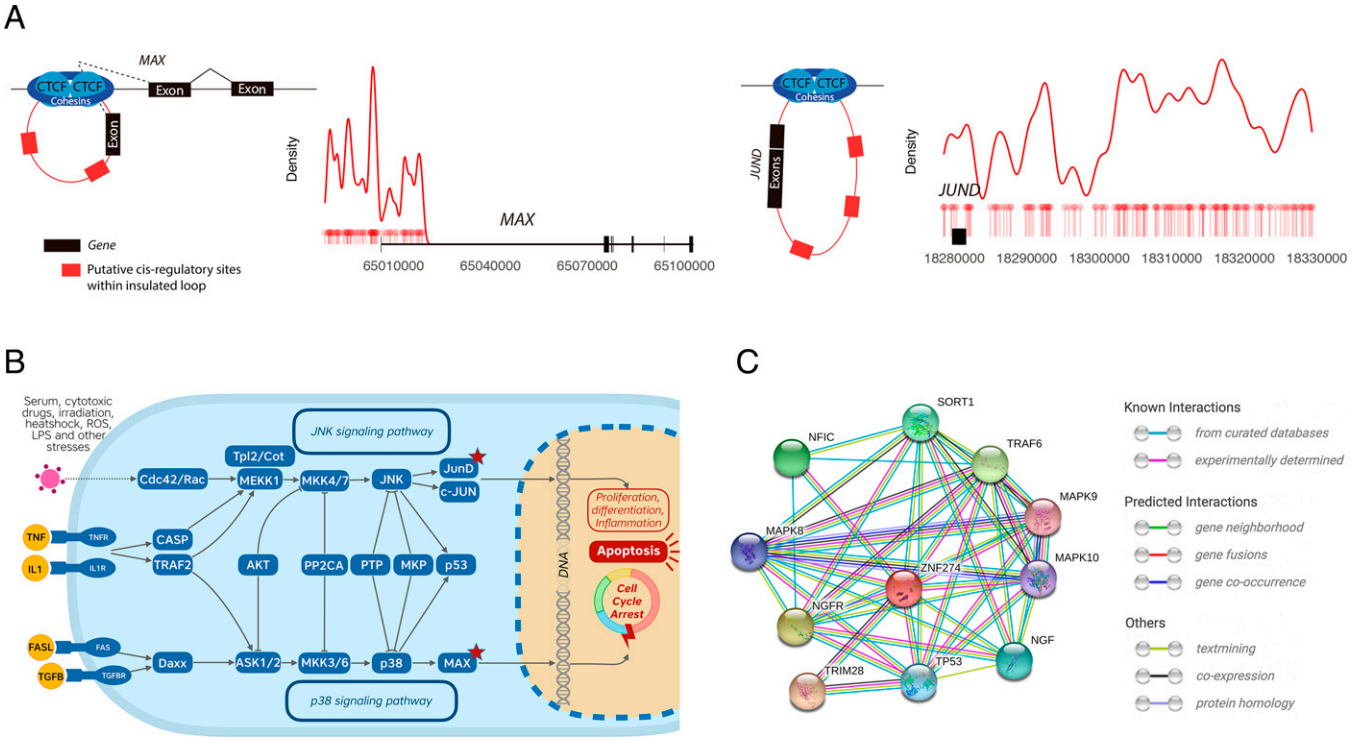

Fig. 3. TF genes whose regulatory regions are enriched with rare noncoding SNVs and their interactions. (*A*) The location of rare noncoding variants within the CTCF loops spanning MAX and JUND in cases. (*B*) The pathways regulating cell processes are impacted by rare noncoding variants. The regulatory regions of MAX and JUND are enriched in rare SNVs, impacting the JNK and p38 signaling pathways. Red stars indicate rare variant enrichment of regulatory regions in SB cases. (*C*) The interaction partners of ZNF274 based on data from the Search Tool for the Retrieval of Interacting GenesProteins (STRING).

DNA Replication, and Regulation of DNA-dependent DNA Replication Initiation, adjusted $P < 0.005$), metabolism (Cobalamin Metabolic Process, adjusted $P = 0.00099$) and one-carbon metabolism (Response to Methotrexate [a folate analog], adjusted $P = 0.00486$). Additional results can be found for GO enrichment analysis of cellular components and molecular functions (Datasets S3 and S4). LGD variant–enriched genes related to the ciliary base were overrepresented (adjusted $P = 0.03580$, *SI Appendix*, Table S4), consistent with our previous identification of SB-associated variants in the primary ciliary G Protein-Coupled Receptor 161 (*GPR161*) that caused a mislocalization of the receptor and disrupted downstream signaling (41).

Noncoding Variant Analysis Points to Perturbed Core Signaling Pathways because of the Dysregulation of Transcription Factor Genes. When assessing the impact of rare variants in intergenic, nonprotein-coding regions, it is critical to identify those variants likely to have a deleterious impact on gene regulation as well as to determine which gene(s) may exhibit altered expression. Enhancers are *cis*-regulatory elements well known to modulate gene expression by binding transcription factors (TFs) to facilitate or suppress transcription. Within these enhancer regions, transcription factor–binding sites (TFBSs)—short motifs demonstrated to bind TFs—can be affected by single-nucleotide changes (42). Nevertheless, when bound by TFs, enhancers can loop long distances to contact and regulate specific genes; therefore, it cannot be assumed that a rare SNV in a specific enhancer will impact the closest gene. Recent studies elucidate the constraints that restrict each of the ∼one million documented enhancers in the human genome to specific target-gene interactions (43, 44). Several studies have observed that chromatin loops mediated by CTC-Factor (CTCF) and cohesin bound on both anchors at the loop ends isolate genes from active enhancers, thus leading to a dysregulation of gene transcription units partially or fully within the loop when disrupted (44–48).

Based on the potential of SNVs to abrogate binding of cognate factors, the noncoding portion of the genome was interrogated, which included 17,548,500 rare noncoding SNVs at the cohort level. Regulatory regions of 106 TF genes previously identified as relatively conserved throughout evolution (49) were analyzed for enrichment in rare noncoding SNVs. Two complementary approaches were used to determine the regulatory region coordinates for these: a more traditional one based on curated enhancer regions [obtained from GeneHancer data available through GeneCards (50)] and a state-of-the-art approach based on CTCF loops [conserved across tissues (51) and mapped during early and later differentiation stages in human embryonic stem cells (hESC) (44)] (see *Materials and Methods*). Our use of CTCF loop maps to infer the gene being regulated by a rare variant enriched transcription enhancer region is a strategy that leverages critical insights into the three-dimensional configuration of human ES genomes. This is highly relevant to SB, as the neurulation defect arises in or proximal to germinal epithelium within the first 30 to 40 d of gestation within the typical staging of the maps to which we refer.

SI Appendix, Figs. S5–S8, *A* show the distribution of rare variants in noncoding regions for cases normalized to controls for the coordinates corresponding to each set of coordinates. In this analysis, we identified four TF genes that were enriched for rare SNVs in their regulatory regions in cases compared to controls (Table 3). Quantile–quantile (Q–Q) plots, cumulative distribution function (CDF) plots, and probability–probability (P–P) plots are also included in *SI Appendix* (*SI Appendix*, Figs. S5–S8). Fig. 3*A* shows a schematic of the CTCF loops and the distribution within them of those rare noncoding SNVs present only in cases for two TF genes: *MAX* (Myc-associated Factor

X) and *JUND* (JunD Proto-Oncogene, AP-1 Transcription Factor Subunit). The CTCF loop associated with MAX contains the *cis*-regulatory regions in the 3′ end, including the 3′ untranslated region (UTR). The disruption of the 3′ UTR could affect localization, stability, export, and translation efficiency of mRNA. JUND's CTCF loop isolates both the TF and its regulatory regions; hence, disrupting this loop could affect its transcription interactions. In both cases, the disruption of regulatory loci within *MAX*- and *JUND*-specific CTCF loops positions the variants to affect the expression of these TF genes, which can impact the genes those TFs regulate and, ultimately, the pathways in which they are involved.

Finally, pathway enrichment analysis was carried out using the TF genes identified in the previous analysis (Table 3) as input. The results encompassed several overrepresented signaling pathways pertaining to immunity and the regulation of essential cellular processes, such as cell growth, differentiation, and proliferation (Dataset S5). As expected, GO enrichment analysis involved biological processes predominantly related to transcription (Dataset S6). Nonetheless, it is worth highlighting that terms pertaining to the central nervous system were overrepresented (Response to Axon Injury, adjusted $P = 0.00615$; Neuron Apoptotic Process adjusted $P = 0.00807$). Additional results can be found for GO enrichment analysis of cellular components and molecular functions (Datasets S7 and S8).

Discussion

This study comprises a multicenter, population-based, ancestry-matched genome-wide analysis of SB WGS data. Because of the multifactorial nature of SB, our genomic interrogations were stringent, seeking rare changes that produce potentially damaging mutations in protein-encoding sequences or noncoding TF regulatory regions. The strategy pursued here has yielded significant results that stand up to multiple testing correction. Furthermore, the validity of our analyses was supported by two sources of transcriptomic data from human neurodevelopment. First, nine of the genes with discriminatory potential and found in enriched pathways in our study were also differentially expressed in a small survey of differentially expressed mRNAs from midgestation fetal amniocytes of NTD-affected pregnancies. Second, using our data in a previously described gene module enrichment method (weighted gene coexpression network analysis, WGCNA), we found that variant enriched genes from our SB data overlapped with a gene coexpression module from a study (35) of midgestation human cortex, a module that was classified as representing an early progenitor network. This result is appropriate for a structural birth defect that involves germinal epithelium over the first gestational month and supports the relevance of the SB associations detected here.

Among the pathways, defined by PathCards and GO, that were most highly enriched with genes that were discriminative for SB were "Carbon Metabolism" and "Cobalamin (Cbl, Vit B12) Transport and Metabolism." It is interesting that the carbon metabolism–related genes found to discriminate SB cases from controls in our cohort are not core players in folate, one-carbon metabolism but instead relate to lipid and glucose metabolism. This is particularly relevant in that postfolate fortification epidemiological data have suggested that persistent risks for NTD may be attributable to concomitant population increases in obesity and diabetes (36–38). Obesity, metabolic syndrome, and diabetes are rising public health concerns both in the US and Qatar populations (38, 52–54). This systems biology approach may be particularly suited for the detection of physiologically relevant pathways contributing to SB and may be less subject to ascertainment bias than candidate gene approaches more common in the field.

Another intriguing pathway emerging in this study highlights processes of innate immunity (HSV-1 infection and DNA damage). Among their SB-discriminatory genes, *CGAS* is known to encode a sensor of viral double strand RNA (dsRNA) and DNA-damaged double strand DNA (dsDNA), participates in the RIGI-like signaling pathway, and was differentially expressed in NTD-affected human fetal amniocytes (34). Pathways involving ECM and cytoskeletal regulation mechanisms may illuminate folate resistant mechanisms at work in human NTD, as *Frem2* mutant mice are not protected by folic acid supplementation (12). The RhoGDI pathway is enriched in several unconventional myosin family members known as regulators of actin-based molecular motors. In particular, MYO1D is necessary for the asymmetric localization of planar cell polarity (PCP) protein VANGL1 (55). MYO1C serves as actin transport to the leading edge of motile cells, while MYO9B is a RhoGTPase activator. Along with the myosin heavy-chain MYH gene products, these molecules regulate cell junction dynamics and cytoskeletal contractile elements modulating cell morphologies and so are positioned to facilitate morphogenetic changes in neural tube cells.

This SB study reaches beyond protein coding sequences to examine nucleotide variation in intergenic functional domains of SB patients. The approach presented here identified four TF genes whose regulatory regions are enriched in variants and are likely contributors to SB risk. Among these, *MAX*, *JUND*, and *ZNF274* (zinc finger protein 274) stand out (Table 3). ZNF274 is a transcriptional repressor involved in epigenetically modified chromatin complexes with SETB1-TRIM28 (56) [Fig. 3*C* (57)] and has been associated with p75 neurotrophin-mediated signaling (Dataset S5), which participates in key events in spinal cord neuron survival and plasticity (58, 59). MAX is a bHLH protein, a transcriptional repressor acting via the recruitment of a chromatin remodeling complex with histone methylase activity. Among the overrepresented pathways encompassing multiple TF genes, the MAPK signaling pathway (Dataset S5) is of particular interest in view of its critical role in brain function (60) and immunity (61). Mutations in either MAX or JUND, through changes in the p38 and JNK signaling pathways respectively, could disrupt inflammation processes as well as cell proliferation and differentiation through the cell cycle and induction of apoptosis (Dataset S5 and Fig. 3*B*).

This genome-wide search has identified significant SB-associated pathways and regulatory regions that are hypothesized to be key drivers of SB. The strength of this approach is that it avoids cherry picking among genes and pathways already implicated through mouse genetic studies. For example, pathways such as central metabolism—strongly significant here—have been overshadowed in human genetic studies by candidate gene searches for PCP or one-carbon metabolism genes because of undeniably important insights from animal model investigations. The genes highlighted here in lipid metabolism and glycolytic pathways can be tested in genetic replication studies and biological models. Similarly, regulatory regions for transcription factors *MAX*, *JUND*, and *ZNF274* are now proposed for further scrutiny.

Toward genetic validation of these drivers, if limited to SNVs and InDels, our results indicate WGS on some 3,300 SB cases will be needed to establish significance for individual genes and begin to address common variant contributions to SB. Greater power may be gained from combining gene enrichment by deleterious variants along with damaging structural variation [e.g., CNVs (23)], demanding new computational approaches to accomplish this task. It is unlikely that a single variant or gene would greatly impact nonsyndromic SB risk. However, it is entirely possible that a single pathway could be predisposing if it contained variants affecting multiple genes in the pathway. Furthermore, only a few LGD-containing genes may be

necessary to result in SB in an individual, as there are examples in the mouse of NTDs caused by digenic mutations within a pathway. For example, NTDs have been seen in mice heterozygous for mutations in pairs of PCP pathway genes *Vangl2*/*Ptk7* (62, 63), *Cobl*/*Vangl2*, *Vangl2*/*Scrb*, and *Vangl2*/*Celsr1* (64), digenic mutations in cytoskeletal regulators *Enah*/*Vasp* (65), or cell adhesion genes *Itga1*/*Itga6* (66). A case-control study limitation is that it can only illuminate components of SB risk for the affected individual. Determining the recurrence risk for a couple will require WGS investigations of case–parent trios, and these efforts are ongoing. Trio analyses will enable the identification of inherited versus de novo variants. Further computational approaches are needed to find potential genetic interactions in individual cases. Also, genomics will only provide one piece of the complex puzzle that undoubtedly includes epigenetic modifications of genomic DNA and chromatin, often in response to maternal nutrition and/or environmental exposures. As we build population-based genome investigations, it will be important to gather gene-expression data from the same subjects—for example, from amniocytes of SB-affected and control pregnancies.

Toward functional testing of SB genetic contributions, the hypotheses generated using systems biology–based computational strategies will require biological validation, likely utilizing genome editing of individual genes and regulatory sequences, singly and in combination, in the mouse in vivo and human stem cell systems (reviewed in ref. 67). These animal and human models offer additional opportunities to test environmental stressors that mimic toxic exposures and intrauterine conditions that undoubtedly interact with the genome and impact the epigenome to tip the epistatic load toward SB (20). The approach demonstrated here represents an important step toward an integrated systems biology view of genetic factors underpinning human neural tube defects. The genomic efforts will have to be combined with epigenetic, multiomic, and environmental investigations to obtain a full picture required for precision medicine (68). Importantly, the identification of recurrence risk toward new avenues for prevention is only one use of precision medicine. Knowledge of the genetic risk of an individual SB infant—even which pathways are most likely impacted in that individual—could inform prognosis and allow for devising novel early interventions toward optimizing the developmental potential of the child. Systems biology approaches will enable inclusion of relatively rare, complex genetic disorders such as SB in this future of 21st-century disease prevention and improved individualized healthcare.

Materials and Methods

Study Subjects. For this case/control study, case subjects with nonsyndromic SB who displayed myelomeningocele were selected (69, 70). The initial cohort comprised 310 subjects from two different countries. Of the 157 SB-affected individuals, 85 were collected in the United States and 72 in Qatar. Among the 153 remaining controls are 45 unrelated subjects from the United States and 108 unrelated individuals living in Qatar (71).

The human subject research study protocol was approved by Institutional Review Boards in the US (Weill Cornell Medical College-NY) and the Middle Eastern population receiving their healthcare in Qatar (Hamad Medical Corporation and Weill Cornell Medical College-Qatar). Consent documentation was provided in both English and Arabic; all participants provided informed consent.

We have no information regarding prenatal folate status on any of the individuals in this study. The US food supply has been fortified with folic acid since 1998, and Qatar began a fortification program in 2009 with a nationwide folate supplementation program prior to that. Case-control collections from both countries spanned pre- and postfortification periods, with reported NTD prevalence in both countries relatively stable over the encompassed 25 y span at 0.5 to 1/1,000. We surmise that there is a mixture of folate statuses (at least with respect to timing of samples collected relative to fortification programs) across both US and Qatar cohorts, making it unlikely that the results would be skewed in a particular direction based on folate intake.

WGS. The subject material included genomic DNA extracted specifically for this project from deidentified infant blood spot cards collected from the California Genetic Diseases Screening Program and referred from the California Birth Defects Monitoring Program (72). Genomic DNA was also derived from venipuncture samples collected from subjects participating in the national Spina Bifida Clinic at the Hamad Medical Corporation. Genomic DNA was extracted, whether from bloodspots or venipuncture samples, using the Puregene DNA Extraction Kit (Qiagen). Input amounts of DNA from infant blood spots were 200 to 500 ng, and inputs from venipuncture samples were 2 to 3 μg. All DNA samples were whole-genome sequenced using Illumina chemistries (v3) on HiSeq 2500 instruments to yield short insert paired-end reads of 2×100 base pairs (bp).

Population Structure Analyses and Case-Control Matches. Stratification bias occurs when the variants that distinguished the compared cohorts (cases and controls) are recognized because of mismatching ancestries rather than variation with respect to the phenotype, and this bias is a major shortcoming of genomic trials (73). Addressing the problem of stratification bias requires an unbiased estimator of genomic ancestry (74) that would provide a gene pool breakdown of the studied samples irrespective of the study cohort and a procedure that optimizes the case-control matches by their genomic ancestry. For that, we first extracted a set of 130,000 ancestry informative markers (AIMs) reported by Elhaik et al. (75) from the genotype data. Employing this AIMs set was reported to improve the genomic ancestral inferences (76). Next, we calculated the ancestry of each individual in relation to nine gene pools representing distinct geographic regions around the world (e.g., South Africa) (75) using *supervised ADMIXTURE* (77). The output was the admixture proportions of each individual corresponding to those global gene pools. Subsequently, we applied the Pair Matcher (*PaM*) tool that matches the cases with the controls by their genomic distances (78). Briefly, *PaM* calculates the genetic distances between every two individuals as the sum of differences between their nine admixture proportions. The pairing assignments are then optimized to maximize the numbers of ancestry-matched pairs and ensure that a genomically ancestry-balanced cohort is used in the analysis.

The final study cohort employed for further analysis included 298 human subjects. Of the 149 SB-affected individuals, 77 were from the United States and 72 from Qatar. Among the 149 ancestry-matched controls were 43 unrelated subjects from the United States and 72 unrelated individuals living in Qatar. The remaining 34 controls matching the ancestry of US subjects were selected from the Pan-Cancer Analysis of Whole Genomes study (79), all of which were germline samples obtained from Caucasian subjects.

Read Mapping, Variant Calling, and Annotation. The sequence data were processed using standard pipelines as described in the Broad Institute's Genome Analysis Tool Kit (GATK) Best Practices (80). Reads were aligned to the hg38 reference provided as part of the GATK Bundle using the Burrows-Wheeler Aligner (BWA) (81). Variant calling was performed with GATK4 (82), and joint genotyping was carried out on the whole cohort followed by Variant Quality Score Recalibration. Quality control (following standard practices such as obtaining sequencing metrics, per sample missing rate, and level of heterozygosity) was done to check for DNA contamination and identify outliers, removing those samples with poor quality. Per-variant quality was also assessed, and only variants with a "PASS" in the filter column were retained and annotated utilizing Ensembl Variant Effect Predictor (VEP) v.95 (83).

Rare Coding Variant Analysis. Variants in coding regions were filtered to retain only those that are globally rare [MAX_AF < 0.01 as provided by the max_af flag in the VEP annotation, which reports the highest AF observed in any population from the 1,000 Genomes Project (28), NHLBI - ESP (29), and gnomAD (84)]. Next, LGD variants were identified as SNVs and InDels, including 1) loss-of-function variants (i.e., nonsense, frameshift, splicing, stop gain, or stop lost) and 2) missense variants predicted deleterious [by SIFT (85) and/or PolyPhen (86)]. Variants meeting the previous criteria (from now on, "qualifying variants") were collapsed by gene, that is, a matrix with the number of qualifying variants per gene per subject was obtained. A power calculation for individual gene association was performed using the Genetic Association Study Power Calculator (87). This tool was used to determine the minimum number of subjects required to reach statistical significance at the gene level (*P* value = 0.0000025). Therefore, assuming a power of 80% and an MAF of 5%, at least 3,300 cases are necessary.

Since the sample size necessary to achieve statistically significant single-gene association using rare variant association analysis was well above the available number of cases, an alternative approach based on machine learning

(88) was proposed. The matrix of qualifying variants was used as input to a machine learning classifier for embedded feature selection. Hence, genes were selected as part of the learning algorithm using as class label the group to which each individual belongs (i.e., case or control). The input data were divided into two parts pseudorandomly to ensure proportions were maintained: one part encompassing 80% of the data, which was used for training and cross-validation to select the best, most optimized model, and a second part including the remaining 20% (hold-out dataset), which was used for further, independent assessment of the method's performance on new, unseen data. RF (32)—a machine learning technique which uses numerous decision trees—was employed to build a predictive model of SB using Python's scikit-learn library (89). Hyperparameter tuning was performed utilizing the *random search* and *grid search* functions within scikit-learn. The model performance was assessed by calculating the AUROC (33) utilizing the scikit-learn library. Threefold cross-validation was employed to select the best model as provided by the same library. To estimate the generalization error, the selected model was further tested on the hold-out dataset. As an additional quality control check, we created RF models on 10 sets that were generated by randomly shuffling the group (case/control) labels. This analysis sought to ensure that the model was not learning the noise existing in the data and, as a consequence, would not generalize well.

The features (i.e., the 439 genes with high discriminatory potential) were ranked according to importance as returned by the RF classifier based on the Gini impurity metric. Gini impurity provides a measurement of the likelihood of incorrect classification of a new instance of a random variable (if that new instance was randomly classified according to the distribution of class labels from the dataset), and those with an importance value > 0 were selected for subsequent steps. Genes were next broadly categorized based on GO Slim using WebGestalt (90), thus obtaining a high-level summary of biological categories based on GO terms. The same genes with high discriminatory potential were used as input to GeneAnalytics (91) for pathway and GO enrichment analyses. Within GeneAnalytics, *P* values are calculated assuming an underlying binomial distribution and corrected for multiple comparison using false discovery rate (FDR) (92). Finally, gene module enrichment was carried out as described by Walker et al. (35). Briefly, clusters (gene modules) were obtained by these authors as a result of applying WGCNA (93) to bulk RNA sequencing data from the midgestational (weeks 14 to 21) human cortex. In the present work, gene module enrichment was calculated employing the same logistic regression model described by Walker and colleagues: is.disease $\sim$ is.module + gene covariates. *P* values were adjusted to correct for multiple testing, applying a Bonferroni correction (as described in the same publication).

Rare Noncoding Variant Analysis. Variants in noncoding regions were filtered to retain only those SNVs that are rare (MAX_AF < 0.01 as provided by the VEP annotation using the max_af flag) in any given population part of 1,000 Genomes, ESP, and gnomAD. Regions regulating 106 TF genes previously identified as relatively relevant during development (49) were obtained. These regulatory regions were defined using data pertaining to curated enhancer GeneHancer data (50) and within CTCF loops spanning each TF gene of interest. Three different sets of coordinates—or catalogs—were used to determine the region coordinates for the CTCF loops, including a dataset of conserved loops across multiple tissues (51) and loops mapped in hESCs at an earlier (naïve) and later (primed) developmental stage (44). CTCF maps from these sources are highly appropriate for this purpose, as many CTCF loops are conserved across tissues and developmental stages, and we specifically interrogated those that are known to be conserved. Furthermore, SB arises early in development—before 35 d gestation—and the neural tube is a germinal epithelium, so SB is closely related to hESCs at early (naïve) and more differentiated (primed) progenitor stages.

The subsequent steps were performed for each catalog. First, BEDTools (94) was employed to identify those rare noncoding SNVs that fell within regulatory regions. Similar to the analysis of variants in protein coding exons, variants in noncoding sequences were collapsed by regulatory regions to determine the frequency of SNVs occurring within TF gene regulatory regions. To identify regions with high discriminatory potential for SB, regulatory regions associated to a TF gene were tested for enrichment in cases versus controls. For this purpose, the proportion of SNVs in cases divided by controls was calculated, and the *fitdist* function within the fitdistrplus R package (95) was used to determine which regulatory regions were significantly enriched. *P* values were FDR adjusted to correct for multiple comparisons.

Finally, the list of TF genes whose regulatory regions were significantly enriched with SNVs (FDR < 0.05) in at least one of the catalogs was used as input to pathway and GO enrichment analysis. Similar to the coding variant analysis, this was carried out employing GeneAnalytics.

Data Availability. Genes found to have high discriminatory potential and those that enriched pathways in this study are provided in Datasets S1–S8 in Supplementary Information. Data pertaining to specific variants generated during the downstream analyses, which support the findings of this study, are available upon request to the corresponding author (MER). The whole genome sequencing data cannot be shared in a public database, due to IRB restrictions.

ACKNOWLEDGMENTS. We thank Ms. Amira Assad, Project Coordinator at Weill Cornell Medicine-Qatar, for invaluable efforts toward patient enrollment. We thank the California Department of Public Health Maternal Child and Adolescent Health Division for providing data. The findings and conclusions herein are ours and do not necessarily represent the official position of the California Department of Public Health. This project was supported by the NIH (Grants P01HD067244, R01NS076465, R01HD081216, and T32HD060600) and the Qatar Foundation (Grant NPRP4-149-3-049; the Biomedical Research Program at Weill Cornell Medicine-Qatar).

1. M. E. Ross, C. E. Mason, R. H. Finnell, Genomic approaches to the assessment of human spina bifida risk. *Birth Defects Res.* **109**, 120–128 (2017).
2. M. E. Ross, Gene-environment interactions, folate metabolism, and the embryonic nervous system. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2**, 471–480 (2010).
3. J. B. Wallingford, L. A. Niswander, G. M. Shaw, R. H. Finnell, The continuing challenge of understanding, preventing, and treating neural tube defects. *Science* **339**, 1222002 (2013).
4. J. T. Eppig *et al.*, Mouse Genome Database Group, The Mouse Genome Database (MGD): From genes to mice–A community resource for mouse biology. *Nucleic Acids Res.* **33**, D471–D475 (2005).
5. M. J. Harris, D. M. Juriloff, An update to the list of mouse mutants with neural tube closure defects and advances toward a complete genetic perspective of neural tube closure. *Birth Defects Res. A Clin. Mol. Teratol.* **88**, 653–669 (2010).
6. L. B. Jorde, R. M. Fineman, R. A. Martin, Epidemiology and genetics of neural tube defects: An application of the Utah Genealogical Data Base. *Am. J. Phys. Anthropol.* **62**, 23–31 (1983).
7. A. E. Czeizel, I. Dudás, Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation. *N. Engl. J. Med.* **327**, 1832–1835 (1992).
8. MRC Vitamin Study Research Group, Prevention of neural tube defects: Results of the Medical Research Council Vitamin Study. *Lancet* **338**, 131–137 (1991).
9. R. J. Berry *et al.*, Collaborative Project for Neural Tube Defect Prevention, Prevention of neural-tube defects with folic acid in China. *N. Engl. J. Med.* **341**, 1485–1490 (1999).
10. P. Wolujewicz, M. E. Ross, The search for genetic determinants of human neural tube defects. *Curr. Opin. Pediatr.* **31**, 739–746 (2019).
11. J. D. Gray *et al.*, Functional interactions between the LRP6 WNT co-receptor and folate supplementation. *Hum. Mol. Genet.* **19**, 4560–4572 (2010).
12. A. Marean, A. Graf, Y. Zhang, L. Niswander, Folic acid supplementation can adversely affect murine neural tube closure and embryonic survival. *Hum. Mol. Genet.* **20**, 3678–3683 (2011).
13. K. E. Christensen *et al.*, Moderate folic acid supplementation and MTHFD1-synthetase deficiency in mice, a model for the R653Q variant, result in embryonic defects and abnormal placental development. *Am. J. Clin. Nutr.* **104**, 1459–1469 (2016).
14. L. G. Mikael, L. Deng, L. Paul, J. Selhub, R. Rozen, Moderately high intake of folic acid has a negative impact on mouse embryonic development. *Birth Defects Res. A Clin. Mol. Teratol.* **97**, 47–52 (2013).
15. C. Fuchsberger *et al.*, The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
16. D. Avramopoulos, Recent advances in the genetics of schizophrenia. *Mol. Neuropsychiatry* **4**, 35–51 (2018).
17. M. Schreiber, M. Dorschner, D. Tsuang, Next-generation sequencing in schizophrenia and other neuropsychiatric disorders. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **162B**, 671–678 (2013).
18. E. F. Sener, H. Canatan, Y. Ozkul, Recent advances in autism spectrum disorders: Applications of whole exome sequencing technology. *Psychiatry Investig.* **13**, 255–264 (2016).
19. J. Zhou *et al.*, Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.* **51**, 973–980 (2019).
20. T. G. Beames, R. J. Lipinski, Gene-environment interactions: Aligning birth defects research with complex etiology. *Development* **147**, dev191064 (2020).
21. S. N. Nees, W. K. Chung, The genetics of isolated congenital heart disease. *Am. J. Med. Genet. C. Semin. Med. Genet.* **184**, 97–106 (2020).
22. Z. Chen *et al.*, Threshold for neural tube defect risk by accumulated singleton loss-of-function variants. *Cell Res.* **28**, 1039–1041 (2018).
23. P. Wolujewicz *et al.*, Genome-wide investigation identifies a rare copy-number variant burden associated with human spina bifida. *Genet. Med.* **23**, 1211–1218 (2021).
24. M. Ishida *et al.*, GOSgene, A targeted sequencing panel identifies rare damaging variants in multiple genes in the cranial neural tube defect, anencephaly. *Clin. Genet.* **93**, 870–879 (2018).

25. X. Chen *et al.*, Rare deleterious PARD3 variants in the aPKC-binding region are implicated in the pathogenesis of human cranial neural tube defects via disrupting apical tight junction formation. *Hum. Mutat.* **38**, 378–389 (2017).
26. P. Lemay *et al.*, Rare deleterious variants in GRHL3 are associated with human spina bifida. *Hum. Mutat.* **38**, 716–724 (2017).
27. J. Zou *et al.*, Association between rare variants in specific functional pathways and human neural tube defects multiple subphenotypes. *Neural Dev.* **15**, 8 (2020).
28. G. R. Abecasis *et al.*, 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
29. J. A. Tennessen *et al.*, Broad GO; Seattle GO; NHLBI Exome Sequencing Project, Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
30. K. J. Karczewski *et al.*, Genome Aggregation Database Consortium, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
31. S. Lee *et al.*, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
32. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
33. J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
34. G. R. Nagy *et al.*, Use of routinely collected amniotic fluid for whole-genome expression analysis of polygenic disorders. *Clin. Chem.* **52**, 2013–2020 (2006).
35. R. L. Walker *et al.*, Genetic control of expression and splicing in developing human brain informs disease mechanisms. *Cell* **179**, 750–771.e22 (2019).
36. H. Y. Huang, H. L. Chen, L. P. Feng, Maternal obesity and the risk of neural tube defects in offspring: A meta-analysis. *Obes. Res. Clin. Pract.* **11**, 188–197 (2017).
37. M. R. Loeken, Mechanisms of congenital malformations in pregnancies with pre-existing diabetes. *Curr. Diab. Rep.* **20**, 54 (2020).
38. A. H. Mokdad *et al.*, The continuing epidemics of obesity and diabetes in the United States. *JAMA* **286**, 1195–1200 (2001).
39. Q. Chen, L. Sun, Z. J. Chen, Regulation and function of the cGAS-STING pathway of cytosolic DNA sensing. *Nat. Immunol.* **17**, 1142–1149 (2016).
40. K. J. Denny *et al.*, Neural tube defects, folate, and immune modulation. *Birth Defects Res. A Clin. Mol. Teratol.* **97**, 602–609 (2013).
41. S. E. Kim *et al.*, Dominant negative GPR161 rare variants are risk factors of human spina bifida. *Hum. Mol. Genet.* **28**, 200–208 (2019).
42. M. V. Rockman, G. A. Wray, Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.* **19**, 1991–2004 (2002).
43. K. C. Akdemir *et al.*, PCAWG Structural Variation Working Group; PCAWG Consortium, Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.* **52**, 294–305 (2020).
44. X. Ji *et al.*, 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* **18**, 262–275 (2016).
45. Z. Tang *et al.*, CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015).
46. J. M. Dowen *et al.*, Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374–387 (2014).
47. D. Hnisz *et al.*, Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
48. M. Ganji *et al.*, Real-time imaging of DNA loop extrusion by condensin. *Science* **360**, 102–105 (2018).
49. L. Arbiza *et al.*, Genome-wide inference of natural selection on human transcription factor binding sites. *Nat. Genet.* **45**, 723–729 (2013).
50. S. Fishilevich *et al.*, GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**, bax028 (2017).
51. E. M. Liu *et al.*, Identification of cancer drivers at CTCF insulators in 1,962 whole genomes. *Cell Syst.* **8**, 446–455.e8 (2019).
52. W. R. Rowley, C. Bezold, Y. Arikan, E. Byrne, S. Krohe, Diabetes 2030: Insights from yesterday, today, and future trends. *Popul. Health Manag.* **20**, 6–12 (2017).
53. F. M. Ali, Z. Nikoloski, H. Reka, O. Gjebrea, E. Mossialos, The diabetes-obesity-hypertension nexus in Qatar: Evidence from the World Health Survey. *Popul. Health Metr.* **12**, 18 (2014).
54. E. Ullah *et al.*, Harnessing Qatar Biobank to understand type 2 diabetes and obesity in adult Qataris from the First Qatar Biobank Project. *J. Transl. Med.* **16**, 99 (2018).
55. P. S. Hegan, E. Ostertag, A. M. Geurts, M. S. Mooseker, Myosin Id is required for planar cell polarity in ciliated tracheal and ependymal epithelial cells. *Cytoskeleton (Hoboken)* **72**, 503–516 (2015).
56. S. Frietze, H. O'Geen, K. R. Blahnik, V. X. Jin, P. J. Farnham, ZNF274 recruits the histone methyltransferase SETDB1 to the 3′ ends of ZNF genes. *PLoS One* **5**, e15082 (2010).
57. D. Szklarczyk *et al.*, STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
58. C. A. Bentley, K. F. Lee, p75 is important for axon growth and Schwann cell migration during development. *J. Neurosci.* **20**, 7706–7715 (2000).
59. G. K. Chu, W. Yu, M. G. Fehlings, The p75 neurotrophin receptor is essential for neuronal cell survival and improvement of functional recovery after spinal cord injury. *Neuroscience* **148**, 668–682 (2007).
60. H. H. Ryu, Y. S. Lee, Cell type-specific roles of RAS-MAPK signaling in learning and memory: Implications in neurodevelopmental disorders. *Neurobiol. Learn. Mem.* **135**, 13–21 (2016).
61. G. Huang, L. Z. Shi, H. Chi, Regulation of JNK and p38 MAPK in the immune system: Signal integration, propagation and termination. *Cytokine* **48**, 161–169 (2009).
62. X. Lu *et al.*, PTK7/CCK-4 is a novel regulator of planar cell polarity in vertebrates. *Nature* **430**, 93–98 (2004).
63. L. Wang *et al.*, Digenic variants of planar cell polarity genes in human neural tube defect patients. *Mol. Genet. Metab.* **124**, 94–100 (2018).
64. J. N. Murdoch *et al.*, Genetic interactions between planar cell polarity genes cause diverse neural tube defects in mice. *Dis. Model. Mech.* **7**, 1153–1163 (2014).
65. A. S. Menzies *et al.*, Mena and vasodilator-stimulated phosphoprotein are required for multiple actin-dependent processes that shape the vertebrate nervous system. *J. Neurosci.* **24**, 8029–8038 (2004).
66. A. De Arcangelis, M. Mark, J. Kreidberg, L. Sorokin, E. Georges-Labouesse, Synergistic activities of alpha3 and alpha6 integrins are required during apical ectodermal ridge formation and organogenesis in the mouse. *Development* **126**, 3957–3968 (1999).
67. P. Wolujewicz, J. W. Steele, J. A. Kaltschmidt, R. H. Finnell, M. E. Ross, Unraveling the complex genetics of neural tube defects: From biological models to human genomics and back. *Genesis.* 10.1002/dvg.23459 (2021).
68. J. C. Denny, F. S. Collins, Precision medicine in 2030-seven ways to transform healthcare. *Cell* **184**, 1415–1419 (2021).
69. Y. Lei *et al.*, Mutations in planar cell polarity gene SCRIB are associated with spina bifida. *PLoS One* **8**, e69262 (2013).
70. Y. Lei *et al.*, Identification of novel CELSR1 mutations in spina bifida. *PLoS One* **9**, e92207 (2014).
71. P. Kumar *et al.*, Evaluation of SNP calling using single and multiple-sample calling algorithms by validation against array base genotyping and Mendelian inheritance. *BMC Res. Notes* **7**, 747 (2014).
72. L. A. Croen, G. M. Shaw, N. G. Jensvold, J. A. Harris, Birth defects monitoring in California: A resource for epidemiological research. *Paediatr. Perinat. Epidemiol.* **5**, 423–427 (1991).
73. S. Yusuf, J. Wittes, Interpreting geographic variations in results of randomized, controlled trials. *N. Engl. J. Med.* **375**, 2263–2271 (2016).
74. H. Carress, D. J. Lawson, E. Elhaik, Population genetic considerations for using biobanks as international resources in the pandemic era and beyond. *BMC Genomics* **22**, 351 (2021).
75. E. Elhaik *et al.*, Genographic Consortium, Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat. Commun.* **5**, 3513 (2014).
76. E. Elhaik, Why most principal component analyses (PCA) in population genetic studies are wrong. *bioRxiv* [Preprint] (2021) https://doi.org/10.1101/2021.04.11.439381. Accessed 10 October 2021.
77. D. H. Alexander, K. Lange, Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).
78. E. Elhaik, D. M. Ryan, Pair Matcher (PaM): Fast model-based optimization of treatment/case-control matches. *Bioinformatics* **35**, 2243–2250 (2019).
79. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
80. G. A. Van der Auwera *et al.*, From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 111011–111033 (2013).
81. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
82. R. Poplin *et al.*, Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* [Preprint] (2018) https://doi.org/10.1101/201178. Accessed 30 July 2021.
83. W. McLaren *et al.*, The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
84. K. J. Karczewski *et al.*, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
85. P. C. Ng, S. Henikoff, SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
86. I. A. Adzhubei *et al.*, A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
87. J. L. Johnson, G. R. Abecasis, GAS Power Calculator: Web-based power calculator for genetic association studies. *bioRxiv* [Preprint] (2017) https://doi.org/10.1101/164343. Accessed 15 January 2021.
88. M. W. Libbrecht, W. S. Noble, Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).
89. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
90. Y. Liao, J. Wang, E. J. Jaehnig, Z. Shi, B. Zhang, WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019).
91. S. Ben-Ari Fuchs *et al.*, GeneAnalytics: An integrative gene set analysis tool for next generation sequencing, RNAseq and microarray data. *OMICS* **20**, 139–151 (2016).
92. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate–A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B. Stat. Methodol.* **57**, 289–300 (1995).
93. P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
94. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
95. M. L. Delignette-Muller, C. Dutang, fitdistrplus: An R package for fitting distributions. *J. Stat. Softw.* **64**, 1–34 (2015).