Published in final edited form as:
Genet Med. 2021 July ; 23(7): 1211–1218. doi:10.1038/s41436-021-01126-9.

Genome-wide investigation identifies a rare copy number variant burden associated with human spina bifida

Paul Wolujewicz[1], **Vanessa Aguiar-Pulido**[1], **Alice AbdelAleem**[2], **Vidya Nair**[2], **Gaurav Thareja**[3], **Karsten Suhre**[3], **Gary M. Shaw**[4], **Richard H. Finnell**[5], **Olivier Elemento**[1,6,7,8], **M. Elizabeth Ross**[1]

[1]Center for Neurogenetics, Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY

[2]Neurogenetics Research, Weill Cornell Medicine Qatar, Doha Qatar

[3]Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Doha, Qatar

[4]Pediatrics, Stanford University School of Medicine, Stanford CA

[5]Center for Precision Environmental Health, Departments of Molecular and Cellular Biology, Molecular and Human Genetics and Medicine, Baylor College of Medicine, Houston, TX, USA

[6]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY

Correspondence to: **M. Elizabeth Ross, MD, PhD**, Head, Laboratory of Neurogenetics and Development, Director, Center for Neurogenetics, **Weill Cornell Medicine**, 413 East 69th, Box 240, New York, NY 10021, Phone 646-962-6144, mer2005@med.cornell.edu.

3. Author Information
In adherence to the CRediT -Contributor Roles Taxonomy and publish each author's contribution to the study in the manuscript is as follows.
1. Conceptualization: P.W.; V.A-P.; O.E.; M.E.R.
2. Data curation (including subject enrollment): A.A.A.; A.K.; M.E.R.; R.H.F.; G.M.S.; V.N.; G.T.; K.S.
3. Formal Analysis: V.A-P.; P.W.; O.E.; R.H.F.; M.E.R.
4. Funding acquisition: M.E.R.; R.H.F.; P.W.; A.A.A.
5. Investigation: All co-authors
6. Methodology: P.W.; V.A-P.; O.E.; M.E.R.
7. Project administration: M.E.R.
8. Resources: M.E.R.; O.E.; K.S.; A.K.; A.A.A.; R.H.F.
9. Software: P.W.; V.A-P.
10. Supervision: V.A-P.; O.E.; M.E.R.; K.S.,
11. Validation: P.W.; V.N.; A.A.A.; G.M.S.; R.H.F.
12. Visualization: P.W.; V.A-P.; M.E.R.; O.E.
13. Writing – original draft: P.W.; V.A-P.; M.E.R.
14. Writing – review & editing: P.W.; V.A-P.; M.E.R.; R.H.F.; O.E.; and all co-authors

4. Ethics Declaration
Recruitment protocols were approved by Institutional Review Boards (IRBs) in the US (state of California and Stanford University, the University of Texas at Austin, Weill Cornell Medical College-NY, IRB# 1106011749, specific to this study), and the Middle Eastern population receiving their healthcare in Qatar (Hamad Medical Corporation and Weill Cornell Medical College-Qatar IRB # 15-00065, specific to this study), including informed consent documentation provided in both English and Arabic.

COMPETING FINANCIAL INTERESTS STATEMENT
Dr. Finnell 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.
The other co-authors have nothing to report.

[7]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA

[8]Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA

Abstract

Purpose: Next generation sequencing has implicated some risk variants for human spina bifida (SB), but the genome-wide contribution of structural variation to this complex genetic disorder remains largely unknown. We examined copy number variant (CNV) participation in the genetic architecture underlying SB risk.

Methods: A high-confidence ensemble approach to whole genome sequences (WGS) was benchmarked and employed for systematic detection of common and rare CNVs in two separate ancestry-matched SB case-control cohorts.

Results: SB cases were enriched with exon disruptive rare CNVs, 44% of which were under 10 KB, in both ancestral populations (P=6.75×10-7; P=7.59×10-4). Genes containing these disruptive CNVs fall into molecular pathways, supporting a role for these genes in SB. Our results expand the catalog of variants and genes with potential contribution to genetic and gene-environment interactions that interfere with neurulation, useful for further functional characterization.

Conclusion: This study underscores the need for genome-wide investigation and extends our previous threshold model of exonic, single nucleotide variation toward human SB risk to include structural variation. Since WGS data affords detection of CNVs with greater resolution than microarray methods, our results have important implications toward a more comprehensive understanding of the genetic risk and mechanisms underlying neural tube defect pathogenesis.

INTRODUCTION

Neural tube defects (NTDs) are anomalies of the central nervous system (CNS) present at birth that manifest with varying subtypes and severity and are among the most common structural birth defects. In more severe NTD subtypes the rostral neural tube fails to close, exposing brain (anencephaly) or brain and cervical-thoracic spine (craniorachischisis), resulting in intrauterine or neonatal death. In contrast, spina bifida aperta (SB, myelomeningocele) is a neural tube closure defect most often confined to the caudal spine below the level of T10. With advances in surgical repair and management, the majority of spina bifida patients will live into adulthood, but will experience lifelong physical challenges including paralysis, associated hydrocephalus requiring CSF shunting, autonomic dysfunction, orthopedic issues and more. With heritability estimates as high as 70%[1], NTDs are thought to arise through an interplay of multiple gene-gene interactions determining genetic predisposition and environmental factors that tip the balance toward failed neurulation[2].

The successes of folic acid supplementation for prevention has led to an emphasis in genetic and epidemiological NTD research on the disease association with candidate genes involved in folate metabolic pathways. Moreover, it has prompted numerous studies in animals of genes involved in one-carbon metabolism and their link with structural birth defects. In addition, genetically engineered animal models of NTD have revealed the importance to neurulation of signaling pathways such as Wnt / planar cell polarity (PCP)[3], sonic hedgehog (Shh)[4], and protein kinase A (PKA)[5]. Mouse models of NTDs have established more than 250 genes whose variants predispose to NTD in the mouse, often showing incomplete penetrance that would suggest additional factors in the genetic background or fetal environment are required for NTD to be manifested[6]. These insights have spurred a number of candidate gene searches among affected patients and sometimes parents in an attempt to identify genetic variants that confer risk for developing an NTD. Nevertheless, there remains a translational challenge to reconcile mouse and human NTD data in order to pinpoint genes, pathways and eventually discern patterns of genetic variation that predispose risk in humans.

Despite years of clinical studies and investigations in animal models, the patterns of human genetic variation that predispose to NTD remain elusive, limiting our efforts to clearly define the genetic architecture underlying the etiology of NTDs. This may in part be attributed to the relatively small effect sizes of individual genes and the narrow focus of investigations on variation within protein coding regions of the genome. Whole genome sequencing (WGS) data coupled with recent advances in algorithmic detection of genomic variation offer opportunities to interrogate under-explored forms of potential NTD risk such as structural variants (SVs), whose effects on NTD risk are not well understood. SVs have been shown to alter the structure and dosage of many genes and rare SVs may exert stronger effects on gene expression compared to rare single nucleotide polymorphisms (SNPs). Rare SVs may also ablate exons and create gene fusions, affecting downstream functionality independent from gene expression. Efforts thus far to interrogate SVs and CNVs in NTD cases have largely relied on array-based platforms or whole exome sequencing. Using those tools, deletions in several genes involving cilia and proteoglycans have been implicated in NTD risk[7]. Therefore, we hypothesize that novel NTD risk genes and mechanistic insight will be gained from a comprehensive genome-wide evaluation of CNVs from WGS data.

We report our first in a series of structural variant analyses of WGS data from two separate ancestry-matched human SB case-control cohorts, focusing on the landscape of high-confidence, likely gene disrupting CNVs. For complex diseases including NTDs, it is important to account for population genetic differences to avoid confounding effects of ancestral variation; thus, we sought to maintain a representative balance of population admixtures in each of our cohorts. In this work, we benchmarked an ensemble of computational tools for CNV detection that was then employed to characterize both common and rare CNVs in well-defined human SB cohorts with the objective of further defining genomic risk contribution in these CNS anomalies.

MATERIALS AND METHODS

Study cohorts and whole genome sequencing

SB subjects who displayed non-syndromic myelomeningocele were selected as cases for this study. Of the 140 SB cases, 67 were collected in the US and 73 from Qatar. The US cohort included an additional 46 unrelated controls with another 30 control subjects obtained from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium[8]. These additional 30 germline samples were analyzed and confirmed to maintain similar admixtures as the US cases. The Qatar cohort included an additional 107 unrelated individuals from the same geographic region and who, again, displayed a population admixture similar to cases. In all, the US cohort comprised 143 individuals and the Qatar cohort included 180 individuals. Altogether, 323 anonymized subjects were included in our study encompassing the two populations. The presence of Y-chromosome markers in individuals was used to determine sex ratios in our case-control cohorts. The male:female ratio within in our case group is 0.69 and is 0.68 within our control group, indicating no sex bias in our study.

Genomic DNA was extracted from de-identified infant blood spot cards obtained from the California Genetic Diseases Screening Program as well as from venipuncture samples collected from subjects participating in the national Spina Bifida Clinic at Hamad Medical Corporation in Qatar. DNA extraction was done using the Puregene DNA Extraction Kit (Qiagen, Valencia, California) and all DNA samples were submitted for whole genome sequencing using an Illumina HiSeq2500 platform to yield short insert paired end 2x100bp reads. Ancestry-aware study cohorts were obtained by extracting relevant loci via PLINK (v.1.9)[9] to calculate specific admixture components, and each cohort constituted a representative admixture balance for both cases and controls.

Alignment, SNV Calling and pre-processing

FASTQ reads were aligned to reference genome hg38 using BWA[10]. After reads were sorted and duplicates were removed, SNV and InDel calling was performed with GATK4 and joint genotyping was carried out on the whole cohort according to GATK Best Practices recommendations[11]. Only variants with a “PASS” in the filter column were retained. SAMtools[12] was used on individual bam files to run quality control measures and to assess read depth uniformity. Read depth statistics were also employed in SAMtools or GATK. The median insert size for samples included in the analysis was 413bp.

CNV and MEI detection

To maximize high-quality CNV detection from short read WGS data, we deployed numerous callers known to perform well individually and integrated their results into a consensus framework that we optimized and packaged as an ensemble approach (Supplementary Figure 1). Manta (v.1.4)[13], Delly (v.0.7.7)[14], Lumpy (v.0.2.13)[15], CNVnator (v.0.3.3)[16] and ERDS (v.1.1)[17] were each employed on individual sample bam files. ERDS also used SNV calls as input for each individual to refine CNV breakpoints. The CNVs detected using the read depth tools (CNVnator and ERDS) were only kept if there was an agreement between calls from each tool, which we defined as those CNVs that are of the same type and that have breakpoints located within close proximity of each other (<2 kb for CNVS < 100kb; <5kb for CNVS > 100kb). Only calls ≥ 1kb were retained given the limited ability of these read depth methods to capture smaller CNVs. We added to this set of CNVs the consensus calls made from the tools that also utilize split-read and read pair signatures (Manta, Lumpy and Delly). A consensus call entailed two out of these three tools to agree on the CNV type and that individual breakpoints are located within 1 kb of each other. Only calls ≥ 300bp and ≤ 100kb were retained, since CNVs outside of this range are prone to yield false positives. MELT (v. 2.0.5)[18] was used for detection of mobile element insertions, namely Alu, SVA and L1 elements.

To ensure against software performance dependencies across study cohorts, our ensemble CNV detection approach was implemented as a Docker image and run on individual subjects, generating a single consensus VCF for each sample. Sample VCFs were merged across all samples via SURVIVOR[19] resulting in a non-redundant set of high-quality candidate CNVs.

SV annotation and filtering repetitive and low complexity regions

Each CNV with any predicted overlap with any coding sequencing of the canonical transcript of 20,246 protein-coding genes was annotated as coding. Deletions were considered loss-of-function if they overlapped any coding sequence and duplications were considered loss-of-function if they affected an exon without extending outside the transcript boundaries. Duplications were considered to be copy gain if they spanned the entirety of a transcript. AnnotSV (v.2.0)[20] was used for annotation of VCF using reference genome hg38. We filtered those CNVs from our call sets that had > 70% reciprocal overlap with repetitive and low-complexity regions, which may confound genomic variant detection. As a comprehensive set of repetitive and low-complexity regions, we combined four datasets: (1) the set of assembly gaps defined by UCSC, including centromeres, telomeres, constitutive heterochromatin domains, gaps between or within clones and contigs, and the repeat-dominated short arms of chromosomes 13, 14, 15, 21, and 22; (2) the UCSC list of segmental duplications; (3) the pseudo-autosomal regions of the sex chromosomes; and (4) repeat regions as defined by RepeatMasker[21].

Benchmarking and CNV simulation analyses

In order to select the optimal combination of software and parameters included in our ensemble CNV approach, we conducted a number of benchmarking analyses deploying numerous detection algorithms on both real and simulated genomes. We utilized the well characterized HG002 genome for benchmarking deletion calls obtained from the Genome in a Bottle (GIAB) consortium[30], which provides Tier 1 benchmark regions of high-quality deletions that we utilized as our ground truth data set. For our simulation data, we used RSVsim to simulate deletions and duplications of a range of sizes at various genomic coordinates. Wgsim in the SAMtools package was used to create comparable whole genome sequencing reads similar to our study cohorts for further benchmarking and analysis. For sensitivity and precision evaluations, we used Truvari for HG002 deletion benchmarking and in-house scripts for accuracy metrics on simulated data.

Population genomic databases and BAM confirmation

We utilized the following population reference databases to extract population allele frequencies for our detected CNVs: 1000 Genomes Project[22], Database of Genomic Variants[23], gnomAD-SV[24]. If the coordinates for a given population reference were relative to reference assembly GRCh37/h19, then they were converted using the University of California Santa Cruz (UCSC) Batch Coordinate Conversion (LiftOver) tool.

To manually assess and validate rare coding CNVs, Integrative Genomics Viewer[25] and Samplot[26] was used to visually compare the read depth of the CNV with that of the surrounding regions. This manual curation entailed examining deviations in read depth corresponding to the predicted change in copy number; that is a 50% reduction for a heterozygous deletion or a 50% increase for a heterozygous duplication. Predicted CNVs were required to have unambiguous start and end breakpoints, which was refined using split-read and/or read pair information.

Pathway and Statistical Analyses

Ingenuity Pathway Analysis (IPA), Webgestalt[27] and GeneAnalytics[28] were used to investigate the genes affected by rare coding CNVs and both cases and controls. We utilized the IPA software to identify the top canonical pathways associated with our dataset, which consisted of the genes impacted by rare coding CNVs in our SB cases. We also conducted a pathway overrepresentation analysis with the KEGG functional pathway database, considering only protein-coding genes perturbed by rare CNVs since our aim was to assess the impact of coding CNVs. For the burden analyses, we applied two-sided Wilcoxon rank-sum tests to analyze the distributions of common and rare CNVs as well as MEIs in our SB cohorts. Significance of differences between cases and controls in mean values for the number of rare coding CNVs per genome were assessed in each population cohort using two-tailed Student *t*-tests.

Real-time quantitative PCR

Select rare coding CNVs identified in our cohorts were validated using real-time quantitative PCR in samples for which DNA remained available after whole genome sequencing. DNA from four separate individuals that harbored a rare coding CNV was amplified using primers designed to hybridize in the region containing the putative CNV. Fold changes of expression were calculated and compared to the average value of three control samples, which contained two copies of the gene-specific region in which primers were designed. All reactions were performed in technical triplicates and ß-actin was used as an internal control. Fold changes of expression were calculated using the $2^{-\Delta\Delta CT}$ method and the gene-specific primers used for four CNVs are listed in Supplementary Table 2.

RESULTS

Cohort Characteristics and CNV Workflow

Two separate ancestry-matched cohorts (US and Qatar) were subjected to whole genome sequencing and were analyzed in this study comprising a total of 323 subjects. The US cohort included 67 cases and 76 controls and the Qatar cohort included 73 cases and 107 controls. Cases in each cohort had a clinical diagnosis of non-syndromic SB and controls were balanced with similar ancestry admixture components as the cases in each study cohort. Principal components analysis (PCA) of the population admixtures for each cohort show that the cases and controls are comprised of similar ancestral backgrounds (Figure 1a).

We devised an optimized methodology for CNV detection featuring a consensus strategy that utilizes five CNV / SV callers as well as input from SNV detection for call refinement.

This ensemble approach leveraged multiple genomic signatures into a joint consensus integrative framework (Supplementary Figure 1). The approach employed a combination of read depth callers that performed among the best with regard to accuracy metrics in benchmarking studies and our own analyses[29]. Our approach also incorporated split read and discordant read pair evidence from several tools in order to yield high quality deletions and duplications >300 base pairs (bp). As part of initial analyses, our method was benchmarked against the well-characterized HG002 genome from the GIAB consortium[30] yielding an F1-measure=82.53 in calling deletions. Since this benchmarking data set does not include tandem duplications or copy number gains, results were further refined and tested on simulated genomes with 30X coverage harboring a number of duplications of various sizes. These initial benchmarking efforts led to our goal of utilizing an ensemble approach for CNV detection that performs at an optimized balance of sensitivity and precision compared to other combinations of callers and tested parameters (Supplementary Figure 2).

Detection of Common and Rare Coding CNVs

Our ensemble CNV approach detected a mean of 2389 deletions and 692 duplications per genome with only slightly more CNVs in the Qatar cohort, consistent with similar observations in population substructures including African ancestry (Figure 2a). Although CNV sizes spanned from several hundred bp to several megabases (Mb) in size, no statistically significant difference in the CNV size distribution was seen between cases and controls (P=0.548) (Figure 2d). In addition to our CNV analyses, we deployed a computational pipeline in a subset of our NTD cases and controls to analyze and compare the distributions of mobile element insertions (MEIs). Aside from encompassing over 50% of human genomes, genomic variation caused by Alu, SVA and LINE-1 (L1) elements are associated with risks for multiple human diseases[31]. We investigated these abundant forms of genomic variation in order to ascertain whether they contribute to NTD cases disproportionately more than in controls. Finding no significant difference in the distribution of the number of MEIs between our SB cases and controls (P=0.491), our data did not support a role of mobile elements in SB (Supplementary Figure 3).

We sought to identify CNVs with greatest potential to disrupt gene function and so contribute to NTD pathophysiology. Therefore, our analyses focused on coding CNVs, which we defined as those variants that overlap a coding exonic region by at least one nucleotide. We analyzed common and rare CNVs separately under the assumption that while common CNVs may modify genetic risk, rare CNVs are under more selective pressure and are thus inherently more deleterious. Among common coding CNVs (>1% minor allele frequency (MAF) in population genomic databases), we observed no significant difference between cases and controls in our two cohorts (Figure 3a, 3c). In contrast, among the rare coding CNVs, which we defined as those <0.1% MAF, we observed a statistically significant enrichment in cases compared to controls in both the US cohort and the Qatar cohort (P=7.59×10-4 and P=6.75×10-7, respectively) (Figure 3b, 3d). As with the common coding CNVs, variant size was not a significant factor between cases and controls for the rare coding CNVs (P=0.917), including when we stratified by each cohort (Qatar: P=0.914; US: P=0.652). Moreover, the distributions for the coding CNVs that were within the 0.1-1% MAF range did not reach significance in either cohort (Qatar: P=0.182; US: P=0.234). More

CNVs in the Qatar SB case-control cohort were classified as rare compared to the US SB case-control cohort, presumably due to less Middle Eastern representation in existing population databases. This, however, does not alter the significance of the burden analysis as each case-control comparison was ancestry matched.

Rare Coding CNVs and Potential Functional Significance

Chromosomal locations of rare CNVs seen in SB cases as well as size breakdowns of rare coding CNVs in our cohort are shown in Figure 4. Slightly more than half of the rare coding CNVs detected were gene disrupting deletions compared to duplications (52.75% vs 47.75%), which was true for both cohorts (Supplementary Figure 4). The mean number of genes per genome that were disrupted by rare coding CNVs was significantly higher in SB cases relative to controls in both cohorts ($P < 0.01$) (Figure 4c). The genes affected by rare CNVs in our human SB cases were subjected to pathway analysis using IPA software and, in both cohorts, defined several canonical signaling pathways associated with NTDs, including retinoic acid and protein kinase A (PKA) signaling, and, in the US cohort, WNT/PCP pathways (Figure 5a). However, these pathways did not reach statistical significance after correction for multiple hypothesis testing. Taking another approach, over-representation of genes perturbed by rare coding CNVs in KEGG pathways did suggest potential disruption for cAMP signaling (P=1.18×10-3), though at a relaxed false discovery rate (FDR) of 0.176 (Supplementary Figure 5).

Within metabolic pathways, we found rare coding CNVs in genes serving several cellular and mitochondrial processes, including choline transport and catabolism, which are closely associated with NTD formation (Figure 5b). Choline's oxidation to betaine within the mitochondria provides a link to folate-dependent, one-carbon metabolism, suggesting that the observed gene disrupting variants are likely to contribute to SB risk. For example, *SLC44A2* and *SLC44A3* belong to the SLC44 family of transporters that exhibit choline transmembrane transporter activity. *BHMT2*, which has been associated with NTDs[32], functions as a methyl transferase to catalyze the transfer of a methyl group from betaine to homocysteine. Homocysteine is an intermediate of methionine metabolism and has also been linked with NTD risk in a number of studies[33]. *DMGDH* is directly involved in the catabolism of choline to form sarcosine and is an essential enzyme in the glycine cleavage system, an important mitochondrial process known to harbor genetic risk variants for SB[34]. The visual representations of the CNVs we highlight in the metabolic pathway in Figure 5b are also included as Supplementary Figure 6 to aid in clearer inspection of the reads near the breakpoints of the CNVs. These visualizations display consistent genomic read signatures for each of these CNVs, and was part of the manual curation and validation process. Additional potentially SB-relevant rare CNVs that were detected only in cases are listed in Supplementary Table 1. This includes partial duplications we identified in SB cases impacting *PARD3*, an established NTD risk gene[35] that directs polarized cell growth and affects asymmetric division, as well as *VAV2*, which has several roles in actin dynamics and cytoskeletal remodeling. These cell polarity and cytoskeletal processes are increasingly associated with NTD risk and have led to the discovery of additional novel candidate genes[36]. For the samples harboring relevant CNVs in which DNA remained available, qPCR was performed to experimentally validate our *in silico* findings and we include examples as

Supplementary Figure 7 (primer designs provided in Supplementary Table 2). Finally, Supplementary Table 3 shows the positions of qualifying rare exon disruptive CNVs that fell within IPA-defined signaling pathways in both Qatar and US cohorts and overlap those CNV regions previously identified[41] as having an impact on human neurodevelopment. Five of the CNVs so identified in our study overlapped CNVs associated with autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), obsessive compulsive disorder (OCD), or schizophrenia and were considered by those investigators to be clinically relevant according to American College of Medical Genetics (ACMG) guidelines.

DISCUSSION

Here we report a systematic interrogation of the CNV landscape in human SB and find an increased burden of rare CNVs directly affecting coding nucleotides. Integrative CNV pipelines from WGS data provide better resolution for variant detection over other conventional methods including array comparative genomic hybridization (aCGH) and whole exome sequencing (WES). In particular, 44% of the rare coding CNVs in our cohorts were less than 10kb in size, and many would have gone undetected using array-based or WES assays. The ensemble approach used here for CNV discovery in SB cases detected rare CNVs disrupting genes not previously associated with SB, but that participate in pathways of biological significance for neurulation. Moreover, the observation of rare CNVs in some known NTD risk genes and pathways, including one-carbon metabolism, reinforces the validity of this strategy.

Digenic variants have been observed in a number of mouse models of NTDs[37] as well as in human studies, suggesting synergistic deleterious effects of variants in genes involved in folate metabolism[38] or in PCP component genes[39]. While larger cohorts will be needed to reach statistical power necessary to pinpoint specific gene combinations indicative of individual risk, an oligogenic or polygenic model of SB risk is gaining traction and should be considered when evaluating genomic contribution. Indeed, we previously reported evidence using predicted deleterious exonic SNPs genome-wide to propose an omnigenic model of NTD risk. This threshold model of NTD risk was based on accumulation of singleton loss-of-function variants (SLoFVs), regardless of the genes harboring these variants[40]. In the current study, demonstration of the enrichment of rare gene disrupting CNVs in cases supports extending this threshold burden model of SB risk to include these SVs. That the burden of rare coding CNVs is present in both SB cohorts interrogated in this study supports the notion that our results are not due to effects of population stratification that could confound the interpretation of rare CNVs.

The overlap of the rare gene disrupting CNVs identified here contributes to a resource that may one day enhance clinical utility as it may soon be possible to examine the WGS of an infant with SB for prognostic indicators. For example, SB individuals with rare CNVs disrupting genes previously associated with neurodevelopmental disorders may alert to the need for early and vigorous intervention to optimize cognitive development and communication skills in addition to physical therapy. There is much more to be explored, as our high-confidence detection approach to identify rare CNVs using short-read WGS data almost certainly under-estimates the contribution of structural genome variation to SB risk.

In the future, multi-platform approaches and integration of long-read sequencing technology promise to enable detection of more SVs per genome. Clearly, SVs, including CNVs, are an understudied form of genomic variation in SB that warrants further investigation.

Our analyses of the CNV landscape in our SB cohorts underscore that candidate gene approaches limited to exons do not capture the full scope of genomic variation contributing to risk. Functional experiments will ultimately be critical for vetting genomic variants as they relate to NTD predisposition. Nevertheless, interrogating CNVs genome-wide expands the repertoire in SB research of variants and genes with potential to contribute to genetic and gene-environment interactions that interfere with neurulation. Evidence is accumulating to support the view that threshold burden models of SB pathogenesis and WGS analyses will achieve a more thorough characterization of the genetic architecture of NTDs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the Qatar national Spina Bifida Clinic, Hamad Medical Corporation, and Ms. Amira Assad, Project Coordinator at WCM-Q, for invaluable efforts toward patient enrollment. We thank the California Department of Public Health Maternal Child and Adolescent Health Division for providing data. Findings and conclusions herein are those of the authors and do not necessarily represent the official position of the California Dept. of Public Health. This project was supported by the National Institutes of Health (P01HD067244, R01NS076465, R01HD081216, T32HD060600; T32 ES027801); CDC (6U01DD000489); Qatar Foundation (NPRP4-149-3-049; the Biomedical Research Program at Weill Cornell Medicine Qatar).

2. Funding Sources:

National Institutes of Health (P01HD067244, R01NS076465, R01HD081216, T32HD060600;T32 ES027801); CDC (6U01DD000489); Qatar Foundation (NPRP4-149-3-049, and the Biomedical Research Program at Weill Cornell Medicine Qatar).

1. Data availability

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). Deidentified data will be made available upon request.

REFERENCES

1. Leck I Causation of neural tube defects: clues from epidemiology. Br Med Bull. 1974;30(2):158–163. [PubMed: 4619574]
2. Jansen AC, Keymolen K. Fetal and neonatal neurogenetics. Handb Clin Neurol. 2019;162:105–132. doi:10.1016/B978-0-444-64029-1.00005-9. [PubMed: 31324307]
3. Chen Z, Lei Y, Cao X, et al. Genetic analysis of Wnt/PCP genes in neural tube defects. BMC Med Genomics. 2018;11(1):38. [PubMed: 29618362]
4. Renard E, Chéry C, Oussalah A, et al. Exome sequencing of cases with neural tube defects identifies candidate genes involved in one-carbon/vitamin B12 metabolisms and Sonic Hedgehog pathway. Hum Genet. 2019;138(7):703–713. [PubMed: 31139930]
5. Wu J, Lu X, Wang Z, et al. Association between PKA gene polymorphism and NTDs in high risk Chinese population in Shanxi. Int J Clin Exp Pathol. 2013;6(12):2968–2974. [PubMed: 24294386]

6. Harris MJ, Juriloff DM. 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. 2010;88(8):653–669. [PubMed: 20740593]
7. Chen X, Shen Y, Gao Y, et al. Detection of copy number variants reveals association of cilia genes with neural tube defects. PLoS One. 2013;8(1):e54492. [PubMed: 23349908]
8. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82–93. [PubMed: 32025007]
9. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. [PubMed: 25722852]
10. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–1760. [PubMed: 19451168]
11. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1110):11.10.1–11.10.33. [PubMed: 25431634]
12. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–2079. [PubMed: 19505943]
13. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32(8):1220–1222. [PubMed: 26647377]
14. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):i333–i339. [PubMed: 22962449]
15. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6):R84. [PubMed: 24970577]
16. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974–984. [PubMed: 21324876]
17. Zhu M, Need AC, Han Y, et al. Using ERDS to infer copy-number variants in high-coverage genomes. Am J Hum Genet. 2012;91(3):408–421. [PubMed: 22939633]
18. Gardner EJ, Lam VK, Harris DN, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 2017;27(11):1916–1929. [PubMed: 28855259]
19. Jeffares DC, Jolly C, Hoti M, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. 2017;8:14061. [PubMed: 28117401]
20. Geoffroy V, Herenger Y, Kress A, et al. AnnotSV: an integrated tool for structural variations annotation. Bioinformatics. 2018;34(20):3572–3574. [PubMed: 29669011]
21. Smit AFA, Hubley R & Green P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>.
22. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. [PubMed: 26432245]
23. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42(Database issue):D986–D992. [PubMed: 24174537]
24. Collins RL, Brand H, Karczewski KJ, et al. A structural variation reference for medical and population genetics. Nature. 2020;581(7809):444–451. [PubMed: 32461652]
25. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–26. [PubMed: 21221095]
26. Belyeu JR, Nicholas TJ, Pedersen BS, et al. SV-plaudit: A cloud-based framework for manually curating thousands of structural variants. Gigascience. 2018;7(7):giy064.
27. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. 2019;47(W1):W199–W205. [PubMed: 31114916]

28. Ben-Ari Fuchs S, Lieder I, Stelzer G, et al. GeneAnalytics: An Integrative Gene Set Analysis Tool for Next Generation Sequencing, RNAseq and Microarray Data. OMICS. 2016;20(3):139–151. [PubMed: 26983021]
29. Trost B, Walker S, Wang Z, et al. A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data. Am J Hum Genet. 2018;102(1):142–155. [PubMed: 29304372]
30. Zook JM, Hansen NF, Olson ND, et al. A robust benchmark for detection of germline large deletions and insertions [published correction appears in Nat Biotechnol. 2020 7 22;:]. Nat Biotechnol. 2020;38(11):1347–1355. [PubMed: 32541955]
31. Saleh A, Macia A, Muotri AR. Transposable Elements, Inflammation, and Neurological Disease. Front Neurol. 2019;10:894. Published 2019 8 20. [PubMed: 31481926]
32. Zinck JW, de Groh M, MacFarlane AJ. Genetic modifiers of folate, vitamin B-12, and homocysteine status in a cross-sectional study of the Canadian population. Am J Clin Nutr. 2015;101(6):1295–1304. [PubMed: 25948668]
33. Yang M, Li W, Wan Z, Du Y. Elevated homocysteine levels in mothers with neural tube defects: a systematic review and meta-analysis. J Matern Fetal Neonatal Med. 2017;30(17):2051–2057. [PubMed: 27633659]
34. Shah RH, Northrup H, Hixson JE, Morrison AC, Au KS. Genetic association of the glycine cleavage system genes and myelomeningocele. Birth Defects Res A Clin Mol Teratol. 2016;106(10):847–853. [PubMed: 27620832]
35. Gao Y, Wang J, Shangguan S, et al. Quantitative Measurement of PARD3 Copy Number Variations in Human Neural Tube Defects. Cell Mol Neurobiol. 2018;38(3):605–614. [PubMed: 28623428]
36. Beaumont M, Akloul L, Carré W, et al. Targeted panel sequencing establishes the implication of planar cell polarity pathway and involves new candidate genes in neural tube defect disorders. Hum Genet. 2019;138(4):363–374. [PubMed: 30838450]
37. Murdoch JN, Damrau C, Paudyal A, et al. Genetic interactions between planar cell polarity genes cause diverse neural tube defects in mice. Dis Model Mech. 2014;7(10):1153–1163. [PubMed: 25128525]
38. Dutta HK, Borbora D, Baruah M, Narain K. Evidence of gene-gene interactions between MTHFD1 and MTHFR in relation to anterior encephalocele susceptibility in Northeast India. Birth Defects Res. 2017;109(6):432–444. [PubMed: 28398708]
39. Wang L, Xiao Y, Tian T, et al. Digenic variants of planar cell polarity genes in human neural tube defect patients. Mol Genet Metab. 2018;124(1):94–100. [PubMed: 29573971]
40. Chen Z, Lei Y, Zheng Y, et al. Threshold for neural tube defect risk by accumulated singleton loss-of-function variants [published correction appears in Cell Res. 2019 9;29(9):776]. Cell Res. 2018;28(10):1039–1041. [PubMed: 29976953]
41. Zarrei M, Burton CL Engchuan W, et al. A large data resource of genomic copy number variation across neurodevelopmental disorders. NPJ Genom Med 2019; 4:26. [PubMed: 31602316]

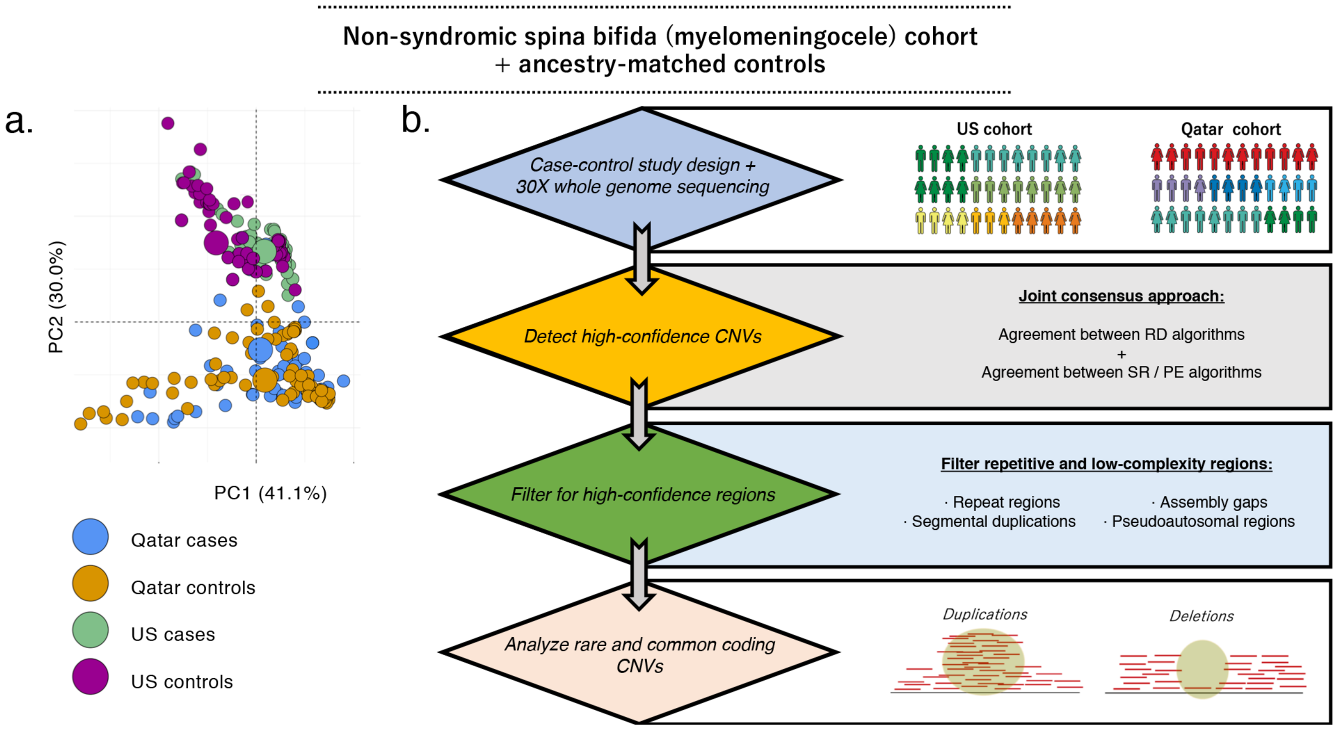
Non-syndromic spina bifida (myelomeningocele) cohort
+ ancestry-matched controls
a.
PC2 (30.0%)
PC1 (41.1%)
Qatar cases
Qatar controls
US cases
US controls
b.
Case-control study design +
30X whole genome sequencing
US cohort
Qatar cohort
Detect high-confidence CNVs
Joint consensus approach:
Agreement between RD algorithms
+
Agreement between SR / PE algorithms
Filter for high-confidence regions
Filter repetitive and low-complexity regions:
· Repeat regions
· Segmental duplications
· Assembly gaps
· Pseudoautosomal regions
Analyze rare and common coding CNVs
Duplications
Deletions

Figure 1. Spina bifida cohorts and analytical workflow.

(a) Two first principal components (PCs) from population admixture data of the individuals in the study colored by cohort and case status. (b) Study design and approach for high-confidence CNV detection in reliable genomic regions using whole genome sequencing data. RD=read depth; SR=split-read; PE=paired end.

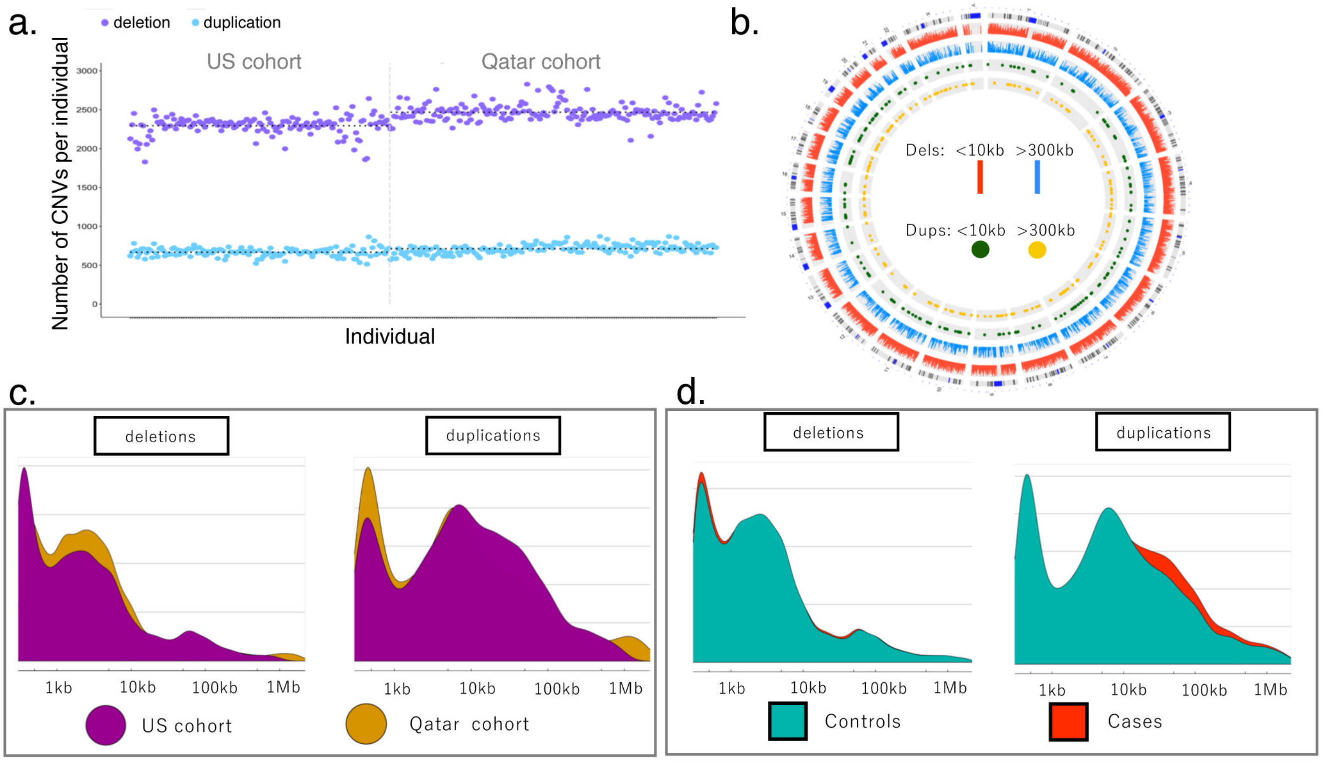

Figure 2. Copy Number Variation Call Set.

(a) Total ascertained CNVs per individual in the respective cohorts (purple=deletion; blue=duplication). All samples included for common and rare variant analyses exhibited comparable read depth and insert size profiles. (b) Circos plot representing small (<10kb) and large (>300kb) CNVs observed in our cohorts (red=deletions under 10kb; blue= duplications under 10kb; green=deletions over 300kb; yellow=duplications over 300kb). (c-d) CNV size distributions of all deletions or duplications is nearly identical between cases and controls in (c) the Qatar cohort or (d) the US cohort.

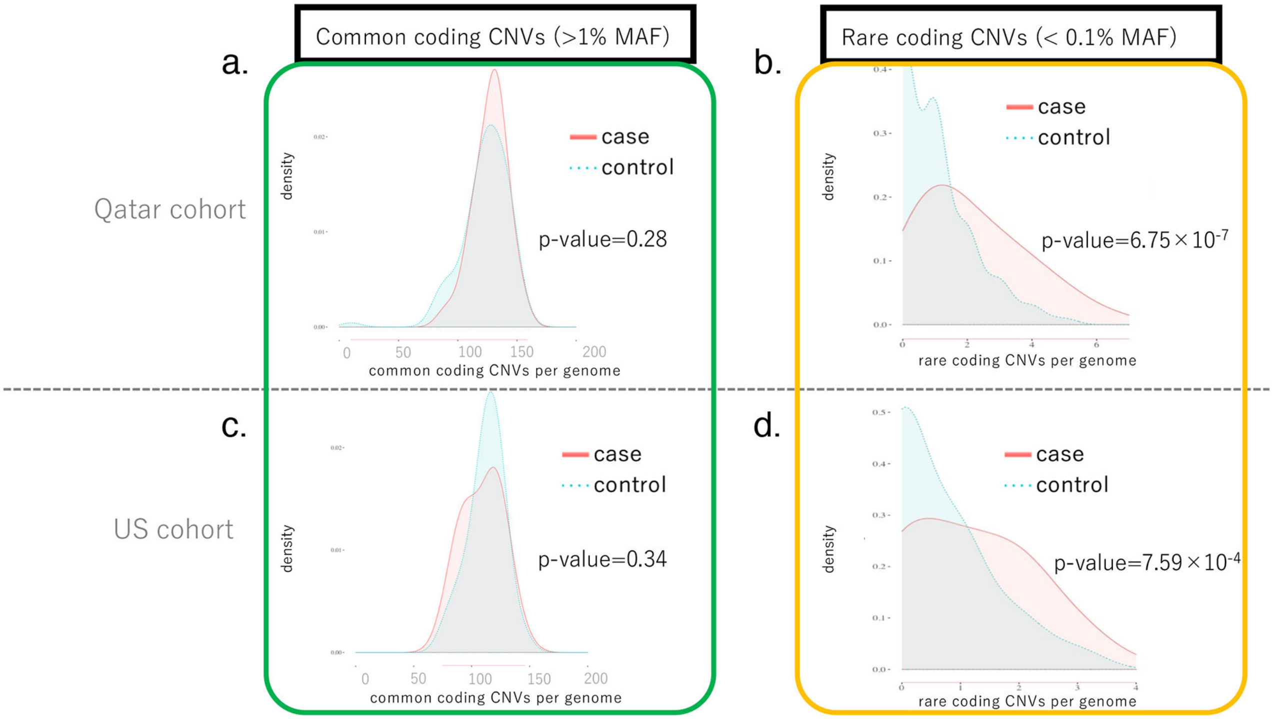

Figure 3. Burden of coding CNVs in SB cases vs controls.

(a,b) Relative distributions of common and rare coding CNVs in the Qatar case-control cohort. (a) The frequency distribution of common coding CNVs does not significantly differ between cases and controls (P=0.28), (b) in contrast, a significant enrichment of rare coding CNVs is observed in cases compared to controls (P=6.75×10-7). (c-d) Common and rare coding CNV case-control comparison in the US cohort. (c) Common CNV per genome distributions do not significantly differ in cases vs. controls (P=0.34), (d) However, a significant enrichment of rare coding CNVs is found in cases compared to controls in this group (P=7.59×10-4).

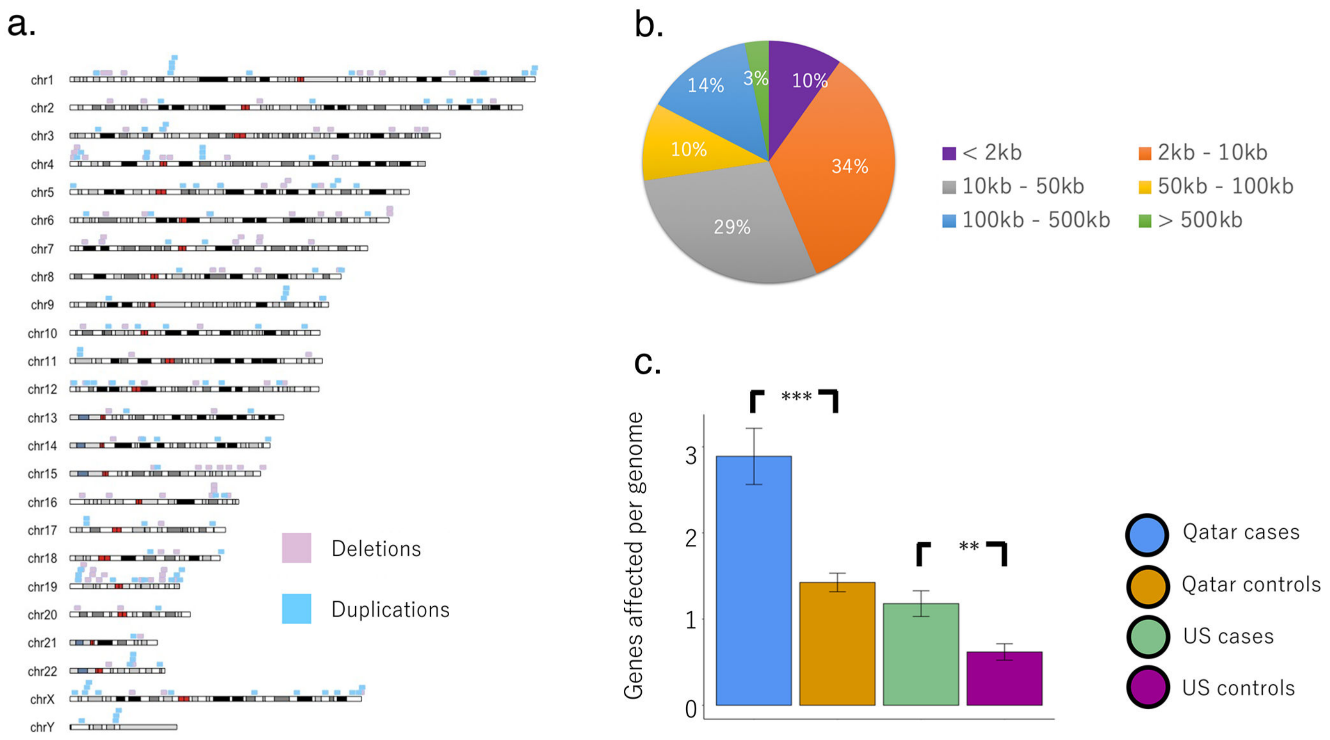

Figure 4. Characterization of detected rare coding CNVs by locus, size and genic content.

(a) Chromosomal location of rare coding CNVs found in SB cases shows broad distribution across the genome. (b) Size breakdown of the rare coding CNV call set in SB cases. Observed rare coding CNVs from both cohorts are categorized into six bins corresponding to the detected CNV size. (c) Comparison within cohorts of the mean value of genes affected by rare coding CNVs per genome in cases compared to controls **p<0.01; ****p<0.0001.

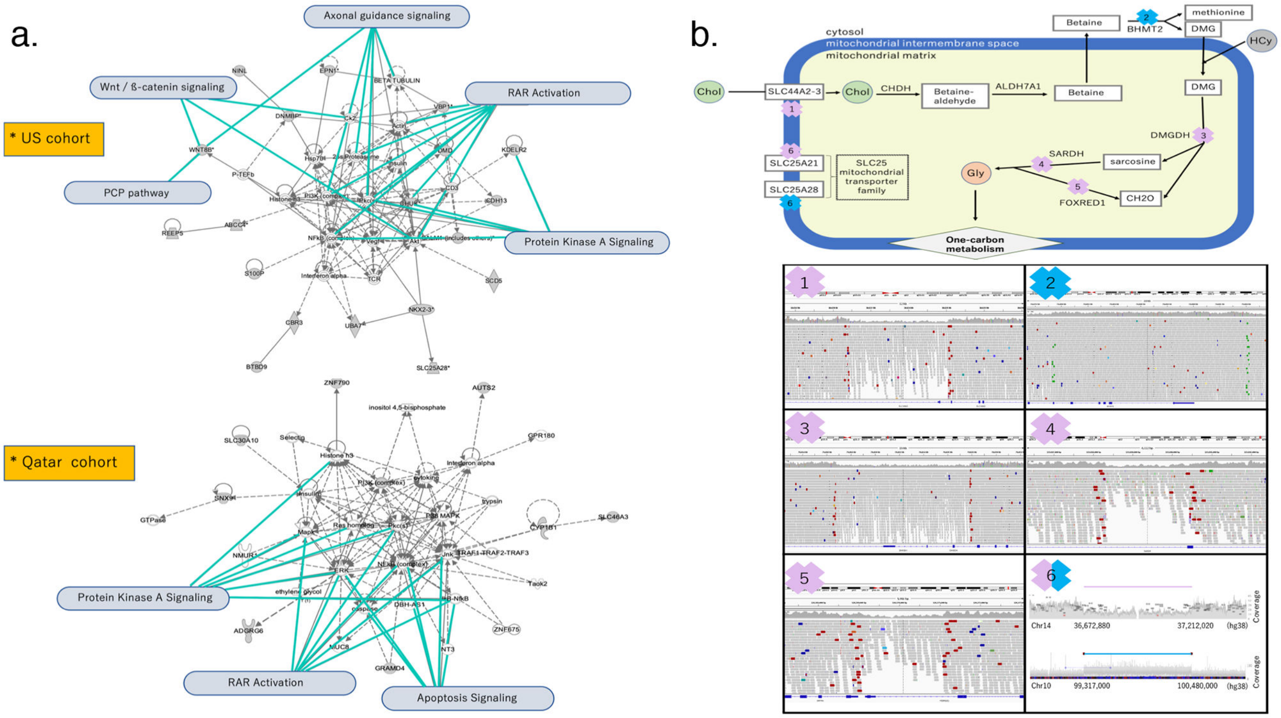

Figure 5. Rare CNVs in SB cases participate in SB relevant pathways and affect various aspects of one-carbon metabolism.

(a) Signaling pathways enriched in cases by gene disrupting, rare coding CNVs detected in the US and Qatar cohorts. Shown are Ingenuity Pathway Analysis (IPA) plots. Several of these pathways in both cohorts emerged as enriched in SB cases. PCP=planar cell polarity; RAR=retinoic acid receptor. (b) Metabolic processes of choline transport and one-carbon metabolism that are disrupted by rare coding CNVs found in SB cases are labeled 1-6. Purple=deletions; blue=duplications. Their corresponding CNVs are represented in the lower panel in (b) using the Integrative Genomics Viewer (IGV) (for CNV 1-5) and samplot (for CNV 6). Chol=choline; Hcy=homocysteine; Gly=glycine.