Published in final edited form as:

Hum Genet. 2018 July ; 137(6-7): 553–567. doi:10.1007/s00439-018-1910-3.

The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease

Jiaqi Liu^{#1,2,3}, Yangzhong Zhou^{#2,4}, Sen Liu^{#1,2,5}, Xiaofei Song⁶, Xinzhuang Yang⁷, Yanhui Fan⁸, Weisheng Chen^{1,2,5}, Zeynep Coban Akdemir⁶, Zihui Yan^{1,2,5}, Yuzhi Zuo^{1,2,5}, Renqian Du⁶, Zhenlei Liu^{2,9}, Bo Yuan⁶, Sen Zhao^{1,2,5}, Gang Liu^{1,2,5}, Yixin Chen^{1,2,5}, Yanxue Zhao^{1,2,5}, Mao Lin^{1,2,5}, Qiankun Zhu^{1,2,5}, Yuchen Niu^{2,5,7}, Pengfei Liu⁶, Shiro Ikegawa¹⁰, You-Qiang Song⁸, Jennifer E. Posey⁶, Guixing Qiu^{1,2,5}, DISCO (Deciphering disorders Involving Scoliosis and COmorbidities) study, Feng Zhang^{11,12}, Zhihong Wu^{2,5,7}, James R. Lupski^{6,13,14}, and Nan Wu^{1,2,5,*}

¹Department of Orthopedic Surgery, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, 100730, China

²Beijing Key Laboratory for Genetic Research of Skeletal Deformity, Beijing, 100730, China

³Department of Breast Surgical Oncology, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China

⁴Department of Internal Medicine, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, 100730, China

⁵Medical Research Center of Orthopedics, Chinese Academy of Medical Sciences, Beijing, 100730, China

⁶Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, 77030, United States of America

⁷Department of Central Laboratory, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, 100730, China

⁸School of Biomedical Sciences, The University of Hong Kong, Hong Kong, China

⁹Department of Neurosurgery, Xuanwu Hospital, Capital Medical University, Beijing, 100053, China

*Corresponding author: Nan Wu, Department of Orthopedic Surgery, Beijing Key Laboratory for Genetic Research of Skeletal Deformity, Medical Research Center of Orthopedics, all at Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, No.1 Shuaifuyuan, Beijing 100730, China, China. Tel: +861069152809, dr.wunan@pumch.cn.

Conflict of Interest

J.R.L. has stock ownership in 23andMe and Lasergen, is a paid consultant for Regeneron Pharmaceuticals, and is a coinventor on multiple the United States and European patents related to molecular diagnostics for inherited neuropathies, eye diseases and bacterial genomic fingerprinting. The Department of Molecular and Human Genetics at Baylor College of Medicine derives revenue from the chromosomal microarray analysis and clinical exome sequencing offered in the Baylor Genetics Laboratory (<http://bmgl.com>).

¹⁰Laboratory of Bone and Joint Diseases, Center for Integrative Medical Sciences, RIKEN, Tokyo, 108-8639, Japan

¹¹Obstetrics and Gynecology Hospital, Institute of Reproduction and Development, Fudan University, Shanghai, 200433, China

¹²Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, 200433, China

¹³Departments of Pediatrics, Baylor College of Medicine, Houston, Texas, 77030, United States of America

¹⁴Texas Children's Hospital, Houston, Texas, 77030, United States of America

These authors contributed equally to this work.

Abstract

With the recent advance in genome-wide association studies (GWAS), disease-associated single nucleotide polymorphisms (SNPs) and copy number variants (CNVs) have been extensively reported. Accordingly, the issue of incorrect identification of recombination events that can induce the misannotation of multi-allelic or hemizygous variants has received more attention. However, the potential distorted calculation bias or significance of a detected association in a GWAS that may result due to the coexistence of CNVs and SNPs in the same genomic region may remain under-recognized. Here we performed the association study within a congenital scoliosis (CS) cohort whose genetic etiology was recently elucidated as a compound inheritance model including mostly one rare variant deletion CNV allele and one common variant noncoding hypomorphic haplotype of the *TBX6* gene. We demonstrate that the existence of a deletion in *TBX6* led to an overestimation of the contribution of the SNPs on the hypomorphic allele. Furthermore, we generalized a model to explain the calculation bias, or distorted significance calculation for an association study that can be 'induced' by CNVs at a locus. Meanwhile, overlapping between the disease-associated SNPs from published GWAS and common CNVs (overlap 10%) and pathogenic/likely pathogenic CNVs (overlap 99.69%) was significantly higher than the random distribution ($p < 1 \times 10^{-6}$ and $p = 0.034$, respectively), indicating that such locus co-existence of CNV and SNV alleles might generally influence data interpretation and potential outcomes of a GWAS. We also verified and assessed the influence of colocalizing CNVs to the detection sensitivity of disease-associated SNP variant alleles in another adolescent idiopathic scoliosis (AIS) genome-wide association study. We propose that detecting co-existent CNVs when evaluating the association signals between SNPs and disease traits may improve genetic model analyses and better integrate GWAS with robust Mendelian principles

Keywords

Copy Number Variants (CNV); Single-nucleotide Polymorphisms; Single Nucleotide Variants (SNV); Association Analysis

Introduction

Recent efforts in high-throughput sequencing are generating large catalogs of genetic variants in general populations and patients with various diseases (Genomes Project et al. 2010; Mills et al. 2011; Tennesen et al. 2012). However, false assignments of disease-gene association and pathogenicity can result in suboptimal utility of previous resources in both basic and clinical research. It has been demonstrated that inaccurate data interpretation and identification of recombination events due to duplication CNV could result in mis-mapping the true disease-associated locus in a linkage study (Lupski et al. 1991; Matise et al. 1994). Though guidelines for identifying disease-causing genes or variants have been proposed, the risk of false-positive reports still cannot be ignored (MacArthur et al. 2014; Yang et al. 2013b). Notably, recent genomic studies and statistical analyses of variant data from general populations have cast serious doubts on the previously published disease causality of numerous severe loss-of-function (LoF) mutations (Lek et al. 2016; MacArthur et al. 2014; Tarailo-Graovac et al. 2017), suggesting the possibility of false-positive findings in some previous studies.

Single-nucleotide polymorphisms (SNPs) are present at a frequency of more than 1% in the human genome, either in the coding regions or non-coding regions affecting exon splicing or transcription (Genomes Project et al. 2012). In general populations, SNPs have been recognized as predictive markers in complex traits such as height (Visscher et al. 2010), and associated with multiple common diseases, including type II diabetes mellitus (Flannick and Florez 2016), Crohn disease (Franke et al. 2010), schizophrenia (International Schizophrenia et al. 2009; Lee et al. 2012), and breast cancer (Han et al. 2016). Further validation of the association significance of these SNPs includes functional validation in model organisms such as mice (Flint and Eskin 2012), zebrafish (Gonzaga-Jauregui et al. 2015), *Drosophila* (Yoon et al. 2017), plant (Crossa et al. 2010), or other experimental organisms, and informatic validation supported by tools like PLINK (Purcell et al. 2007) and GCTA (Yang et al. 2011). However, it remains a challenge to generally interpret the functional significance and gene/variant allele involved for novel disease-associated SNPs identified from either genome-wide association studies (GWAS) or next-generation sequencing (NGS) data. Practically, distinct transcriptional and post-transcriptional models may underlie the observed effects of genetic variants (Chick et al. 2016). Additional modes of inheritance such as parental imprinting or compound heterozygosity also add to the complexity (Albers et al. 2012; Wu et al. 2015).

Copy number variants (CNVs) are defined as sequence variants ranging from 50 bps to several megabases (Mbs) in size, including deletions, duplications, triplications, insertions, complex genomic rearrangements (CGR) and other CNVs (Carvalho and Lupski 2016). Compared with single-nucleotide variants (SNVs), CNVs are responsible for more than ten times the heritable sequence differences in general populations (Pang et al. 2010), and their genome-wide map has been comprehensively studied (Zarrei et al. 2015). Likewise, they are involved in the pathogenesis of both sporadic Mendelian disorders and complex multifactorial disease (Andrews et al. 2015; Stankiewicz and Lupski 2010; Weischenfeldt et al. 2013). It is noteworthy that common CNVs and SNPs in relatively nearby regions have been associated with an increased likelihood of causing the same phenotype, such as Crohn

disease (McCarroll et al. 2008a), rheumatoid arthritis and type 1 diabetes etc. (Wellcome Trust Case Control et al. 2010). Tandem and segmental duplications or even triplications would result in multi-allelic variants involving more than one locus and of further importance (Bailey et al. 2001; Campbell et al. 2016; Fredman et al. 2004), e.g. the 24-kb-long Charcot-Marie-Tooth disease type 1A-repeats (CMT1A-REPs) that sponsor deletions and duplications at 17p12 (Lindsay et al. 2006; Lupski 2003; Lupski et al. 1991). This situation becomes even more complex when examining structurally and evolutionarily unstable loci in the genome, as exemplified by studies of 2q13, 17q21.31 and 16p12.1 regions in human genomes (Antonacci et al. 2010; Yuan et al. 2015; Zody et al. 2008). Multi-allelic SNVs account for ~2.3% of all autosomal SNVs in the 1000 Genomes Project Phase 3 (Genomes Project et al. 2012), ~6.4% in ExAC (Exome Aggregation Consortium, <http://exac.broadinstitute.org>) database), and ~2.2% from a previous study (Campbell et al. 2016). The combination of CNV and SNV polymorphisms underscores the limitation of the current haploid human genome reference which does not annotate structural variant polymorphisms (Carvalho and Lupski 2008; Yuan et al. 2015) at a locus and may affect genomic data interpretation especially when dealing with complex regions of the genome. Although limited, some insights have been gained in understanding the impact of the coexistence of CNVs and SNPs in the same region, especially those associated with human diseases. And misannotation of multi-allelic variants in clinical genome sequencing could lead to missing diagnoses (Campbell et al. 2016; Trivellin et al. 2014).

Furthermore, studies of common deletion polymorphisms also demonstrated essentially the same distribution of linkage disequilibrium with surrounding SNPs, indicating that these mutations are likely to be ancestral and share evolutionary history (Hinds et al. 2006). Deletions would lead to a hemizygous allele at corresponding genomic regions and contribute to the allelic architecture of both carrier and recessive disease-causing mutations (Boone et al. 2013; Flipsen-ten Berg et al. 2007). The recent report of a compound inheritance model in congenital scoliosis (CS) provides a model to simulate this particular scenario (Wu et al. 2015). CS is a relatively rare and severe congenital disease which is defined as a lateral curvature of the spine exceeding 10 degrees due to a congenital vertebral malformation during somitogenesis. CS affects 0.5~1 in 1,000 live births (Giampietro et al. 2003). We previously conducted an association analysis using a candidate gene approach in the Han Chinese population and reported that two SNPs (rs2289292 and rs3809624) in the *TBX6* gene were significantly associated with CS (Fei et al. 2010). Following the identification of this association, we demonstrated that up to 11% of sporadic CS cases had a rare null allele of the *TBX6* gene, due to either rare variant CNVs (16p11.2 deletions) or SNVs (nonsense or frameshift variants), in combination with one common variant hypomorphic allele in the same region ('T-C-A', defined by the coexistence of three common SNPs, rs2289292-rs3809624-rs3809627). Additional functional and multi-racial evidence were subsequently provided to establish the *TBX6* compound inheritance model in sporadic CS patients (Wu et al. 2015). This model was further repeatedly observed in a Japanese and a European CS cohort independently (Lefebvre et al. 2016; Takeda et al. 2017).

After the introduction of this compound inheritance model of a rare CNV with an *in trans* common variant hypomorphic haplotype, we hypothesized that the associated haplotype

(rs2289292 and rs3809624) identified previously (Fei et al. 2010) could be misinterpreted as homozygous SNPs and the calculated significance in the GWAS distorted given that a hemizygous allele was taken as homozygous in patients with deletions. This will result in an overestimation of the *TBX6* haplotype prevalence in the patient population (Fig 1), which is a realistic model to investigate the effect of coexisting CNVs on the interpretation of disease-associated SNPs. In this study, we evaluated the effects of deletion CNVs and SNP which occur in the same genomic region on the disease-association analysis using data from the previously described CS cohort (Wu et al. 2015). We further document a generalized model that was introduced to reveal the mechanism by which CNVs cause distorted calculations of the association power. To investigate the universality of the distorted calculations, we measured the overlap of significant SNVs identified by GWAS with common CNVs (frequency > 1%) and pathogenic/likely pathogenic CNVs, and also validate the influence in another genome-wide association study of adolescent idiopathic scoliosis (AIS). Collectively, we proposed a conceptual framework delineating how to avoid these potential genotyping misinterpretations and distorted calculations in SNVs association analysis when analyzing and interpreting the results of an association study.

Materials and Methods

Subjects

A case-control design was adopted in this study, with all participants enrolled from the Peking Union Medical College Hospital (PUMCH) in China between October 2010 and June 2014. A total of 161 unrelated sporadic CS patients and 166 unrelated healthy controls were recruited (Wu et al. 2015). Furthermore, a GWAS of 196 AIS patients and 303 subjects without AIS was carried out from the Centre of Genomics Sciences in the University of Hong Kong for further validation. The diagnoses of CS and AIS were confirmed by clinical experts based on the patients' radiological findings. The 469 controls without CS or AIS were diagnosed by MRI scan of the spine. Written informed consents were obtained from all the participants (those who were no less than 18 years of age at the time of enrollment) or their guardians (for participants who were less than 18 years of age). The study protocols were approved by the Ethics Committee of PUMCH and the University of Hong Kong (IRB approval number: UW 08–158).

CNV detection and SNPs genotyping

To screen the 16p11.2 micro-deletion, i.e. the deletion CNV, in all the participants in the CS cohort, we performed quantitative polymerase chain reaction (qPCR) analyses. Two test loci (named PA and PB) in the 16p11.2 deletion region and one reference locus (named P1) outside of the deletion region were used. (Primer Sequence: P1-F, GGGGAAGGAAGTTACATGAC; P1-R, TCGTGTTTCCCTGTTGTACC; PA-F, GGTCTAAGCCACACACTAAC; PA-R, TGAGTTTAGGGACCAATCTA; PB-F, GCTGCCAGTATGTGACCGAGA; PB-R, GGGTGGAGGAGAGGATAGGG) The qPCR primers are provided in the table below. The experiments were conducted using the SYBR Green Real-time PCR Master Mix (TOYOBO, Japan) and ABI Prism™ 7900HT Sequence Detection System. Three replicate experiments were conducted for each assay. The average Ct values, Ct (PA-P1) and Ct (PB-P1) were calculated for each sample. Any 16p11.2

deletion candidates suggested by qPCR were confirmed by CGH microarray. Finally, detection of the 16p11.2 micro-deletion was confirmed by array-based comparative genomic hybridization microarray (aCGH) in all the candidates from the qPCR assay. In all the participants, full-length *TBX6* plus 1 kb of the upstream region were amplified by long-range PCR using LA Taq polymerase (Takara, Japan). The detailed methods have been described in our previous study (Wu et al. 2015).

To further verify our model in a genome-wide association study, 196 cases with AIS were genotyped with Affymetrix Genome-Wide Human SNP Array 6.0, and 303 subjects without AIS were genotyped with Affymetrix Human Mapping 500K Array. SNPs and CNVs were analyzed by Genotyping Console Software (Affymetrix, USA).

Correction of SNPs genotyping

To depict the true allelic distributions of the three previously identified SNPs (rs2289292, rs3809624, and rs3809627) in CS cohort, we removed the miscounted alleles induced by false interpretation of homozygosity resulting from the 16p11.2 deletions in our cohort. The genotypes of patients with 16p11.2 deletion were corrected to reflect a hemizygous state from miscounted homozygous alleles.

Modeling the parameters affecting distorted calculations

The distorted calculations of SNPs significance due to the coexistence of CNVs was simplified and simulated based on the simple chi-square equation (two-sided). The curve-fitting was conducted by the coordinate of the crossover significance ($p=0.05$) of the difference between the allele frequencies (AFs) in cases and controls with varying CNV frequency, sample size and the association between the SNPs and CNVs. Additionally, we took the three SNPs (rs2289292, rs3809624, and rs3809627) into this model to verify the simulated distorted calculations.

Genomic data collection and views

The dataset of significant SNPs related to clinical conditions or phenotypes was obtained from data deposited in the National Human Genome Research Institute (NHGRI) Catalog of Published GWAS (<http://www.ebi.ac.uk/gwas>, Table S1) (Welter et al. 2014), and the dataset of CNVs was acquired from the Database of Genomic Variants (DGV) (<http://dgv.tcag.ca/dgv/app/downloads>) (MacDonald et al. 2014). After filtering duplicate variants, common CNVs (frequency > 1%) were respectively compared with the loci of significant SNPs and the human hg19 genome to calculate each overlap area. The pathogenic/likely pathogenic CNVs were acquired from UCSC database (<https://genome.ucsc.edu>), including tracks of ClinGenCNV (Kaminsky et al. 2011; Miller et al. 2010), cnvDevDelayCase (Coe et al. 2014; Cooper et al. 2011), clinvarCnv (Landrum et al. 2016), and coriellDelDup (NCBI:dbGaP). We also incorporated a list of expert-curated CNV-associated syndromes from DECIPHER (<https://decipher.sanger.ac.uk/>). After filtering out benign/likely benign variants since the frequencies of them demonstrate no difference between case and control groups, and the remaining variants were further grouped into loss or gain for those with the type annotation. The aneuploidies (e.g. trisomy 13, 18, 21, and X) were not included in the group of pathogenic CNVs in this study. The frequency of each interval was calculated by

counting the variants that cover this interval (Table S2). Circos plots were provided to visualize the significant SNPs from GWAS and common/pathogenic CNVs against the human hg19 genome using the circos-0.67–2 package of Perl.

Statistical analysis

To test if the detected SNPs in our dataset fit Hardy-Weinberg equilibrium (HWE), we performed a goodness-of-fit Chi-Square test. Association analysis was performed for allelic, genotypic association utilizing the SPSS software v15.0 (SPSS, USA). Haplotyping results were assessed by version 4.2 of the Haploview program (Barrett et al. 2005). The difference of AFs between case and control groups was compared using Pearson χ^2 test (two-sided). Odds ratio (OR) with 95% confidence interval (CI) was used to measure the influence of *TBX6* polymorphisms on the occurrence of CS by using the unconditional logistic regression built in the SNPstats software (<http://bioinfo.iconcologia.net/SNPstats>). $p < 0.05$ (two-sided) was considered as statistically significant.

Results

Detection of CNV and SNVs in the *TBX6* locus

The 16p11.2 deletion was identified in 12/161 unrelated CS patients by qPCR analyses and confirmed by aCGH. Sequencing of *TBX6* revealed one nonsense and four frameshift variants in five CS patients without the 16p11.2 deletion. Thus, Lof variant alleles, presumably representing null mutations of *TBX6*, were found in a total of 17 CS patients. We went on to identify the common hypomorphic allele consisting of three SNPs of rs2289292, rs3809624, and rs3809627 as the other contributing factor, i.e. biallelic variants at the *TBX6* locus, in the compound inheritance model (Fig 1A, Table 1) (Wu et al. 2015). Distributions of each of these three SNPs did not deviate from HWE ($P > 0.05$; Table 1).

Allelic, genotypic and haplotypic association analyses

The AFs in *TBX6* were significantly different between the cases and the controls in rs2289292 and rs3809624 (T and C, respectively), but not in rs3809627 (A, Table 1). The two non-reference alleles of rs2289292 and rs3809624 were the risk alleles for CS (OR=1.38, 95% CI 1.02-1.88; OR=1.44, 95% CI 1.06-1.96; Table 1). However, genotypic association showed no significant difference between the case and control groups (data not shown).

In the 12 CS patients with the 16p11.2 deletion, the hemizygous allele in *TBX6* was miscounted as two identical (homozygous) alleles during annotating the variant data (Fig 1B). To depict the true AFs of those three SNPs, we removed the over-counted alleles due to the coexistence of the deletion. After correction, there was no significant difference in the allelic distributions between the cases and controls for all the three SNPs (Table 1).

The *TBX6* risk haplotype was shown to be a significant risk haplotype in CS patients (OR=1.42, 95% CI 1.02-1.97, $P = 0.038$; Table 2). After correcting the miscounted alleles induced by the 16p11.2 deletion, the significance shown above disappeared (OR=1.32, 95% CI 0.95-1.84, $P = 0.110$; Table 2).

Modeling the distorted calculations of the SNP allele frequency due to a coexisting CNV

The effect of distorted calculations was further simulated based on the simple chi-square equation. The significance of SNPs of interest was simultaneously affected by the sample size, frequencies of SNPs and CNVs, and the coexistence of the SNPs and CNVs (Fig 2). First, with an increasing sample size, more subtle differences between the cases and controls can be shown with significance (Fig 2A). When the CNVs are not associated with any disease-causing allele, they simply affect the number of total alleles and distorted calculations of the significance, which is similar to the effect of changing sample size. While deletion is statistically similar to a decrease of sample size, duplication is similar to an increase of sample size (Fig 2B). When a particular CNV is associated with a particular allele in the case group, the calculated frequency of this SNP increases (in case of deletion) or decreases (in case of duplication) according to the frequency of the CNV. In the case simulated in Fig 2, the sample size is fixed and the CNV is entirely associated with the SNP (coexistence frequency=10%), the areas of significance deviate in opposite directions depending on the existence of deletions or duplications (Fig 2C). In this study, the 16p11.2 deletion was associated with the 'T-C-A' haplotype in affected cases; therefore the calculated frequency of this haplotype became higher than the real frequency of the SNP alleles. The validated set of data from the CS cohort (N=160) was also applied in this model. If the coexistence of 16p11.2 deletion is not considered in the association analysis, rs2289292 and rs3809624 are found in the area of significance. When the frequency of 16p11.2 deletion is considered in the correction of the AFs of the SNPs, none of the three SNPs is in the area of significance. Thus, the previous distorted calculations of association between the two SNPs and CS is corrected (Fig 2D).

Disease-associated SNPs significantly enriched with CNVs

A typical design of GWAS requires the comparison of a population with certain phenotypes (cases) to the general populations (control). If the CNV is associated with a certain disease phenotype, there is supposed to be a skewed distribution signal the case group compared with the control population. However, the overlapping of common CNVs with SNPs could substantially influence the SNP calculation. To address this question, the distribution of 20,726 significant SNPs collected from the NHGRI Catalog of Published GWAS (Table S1) (Welter et al. 2014) was matched with the locations of common CNVs (population frequency > 1%) acquired from the DGV (MacDonald et al. 2014) and pathogenic/likely pathogenic CNVs acquired from the UCSC database (<https://genome.ucsc.edu>) (Fig 3).

The common CNVs in the general population included 15,264 unique losses and 2,308 unique gains. The significant SNPs and common CNVs were visualized in the circos plot against the human hg19 genome (Fig 3A). Notably, about 10% of the SNPs (2,042/20,726) across the human genome were found to map to the same genomic regions of common CNVs which were observed in only 4.2% of the human genome. The biased overlap between the significant SNPs identified in GWAS and common CNVs indicates that common CNVs have the potential to affect the detection of significant SNPs ($p < 1 \times 10^{-6}$, OR=2.41, OR±95% CI=2.37–2.60, Pearson χ^2 test, Table 3).

We also investigated the overlap of the significant SNPs identified by GWAS (Welter et al. 2014) and the pathogenic/likely pathogenic CNVs acquired from the UCSC genome database (Table S2). As shown in Fig 3B, pathogenic CNVs were observed in 97.63% of the human genome, and most of the significant SNPs (99.69%) were located in the same regions with the pathogenic CNVs. The biased overlap between these SNPs and the pathogenic CNVs also suggested that potential distorted calculations could occur in the GWAS ($p=0.034$, $OR=1.02$, $OR\pm 95\%CI=1.00-1.04$, Pearson χ^2 test, Table 3). Moreover, the detection of significant SNPs might be more likely affected by pathogenic/likely pathogenic CNVs smaller than 10Mb ($p=1.8\times 10^{-8}$, $OR=1.06$, $\pm 95\%CI 1.04-1.08$, Pearson χ^2 test, Table 3).

To verify our findings, we further performed a genome-wide association study of 196 AIS patients and 303 subjects without AIS. In the total of 27,096 SNPs with significant difference ($p<0.05$, Table S3) between cases and controls, 585(2.16%) SNPs were identified overlapping with CNVs in the same region either in the patients or controls (Table S4). To avoid the bias of CNV to disease-associated SNP analysis between the cases and controls, we assessed the influence of the CNVs on the significance of their overlapping SNPs. Finally, four SNPs (rs1999435, rs2847443, rs5996945, and rs389625) were found overlapped with CNVs in both groups, and their significance in associating SNPs to AIS changed when the distorted calculations caused by CNVs were corrected (rs1999435, $p=0.013$ to 0.0017 ; rs2847443, $p=0.047$ to 0.0020 ; rs5996945, $p=4.8\times 10^{-9}$ to 6.2×10^{-9} ; rs389625, $p=0.069$ to 0.0049 ; Table S5).

Discussion

CNV magnifies the association signals of SNPs in the CS patients

In this study, we quantitatively evaluated the effects of CNV on the calculated association signals of SNPs from the same region. Using the association analysis of CS as an example, we provide evidence that CNVs, particularly deletions, could potentially exaggerate the significance of the corresponding SNPs in the association study. The *TBX6* compound inheritance model of one null mutation and one common hypomorphic allele was established in up to 11% of CS patients (Fig 1B) (Wu et al. 2015). Additionally, the common hypomorphic alleles include two common SNPs (rs2289292 and rs3809624) which were previously reported as being associated with CS in the Han Chinese population (Fei et al. 2010). In the allelic association analyses, we also identified that AFs in *TBX6* were significantly different between the cases and the controls in these two SNPs. However, after correcting for the distorted calculations of the prevalence of these SNPs (Fig 1B), there was no significant association signal between the SNPs and the occurrence of CS, consistent with our hypothesis that CNV causes overestimation of the disease association of the coexisting SNPs. As the current sample size was limited, the association between the CS phenotype and these SNPs warrants further evaluation in a larger cohort. The bias in the analysis of SNPs to test for association with complex disease, which is caused by the 'miscounting' of the number of alleles has been shown in the previous study (Marenne et al. 2013). In our study, we further illustrate the theory by providing a realistic case in which the SNP was linked with the CNV.

CNVs might extensively disturb the association power of SNP association analysis

This distorted calculations of SNP association with disease, which we demonstrated in the CS cohort, can also be identified in a real genome-wide association study and readily generalized to other genomic studies. In the GWAS replication, 2.16% of SNPs with significant difference between AIS patients and controls were identified overlapping with CNVs in the same region. Furthermore, four SNPs overlapping with CNVs in both groups showed changed significance in association to AIS when the distorted calculations caused by CNVs were reinterpreted. To reveal the effect of the distorted calculations of association power and accuracy caused by CNVs, we further simplified the effect of distorted calculations based on the chi-square equation. According to this simplified model, the significance of SNPs of interest was simultaneously affected by sample size, frequencies of SNPs and CNVs, and the coexistence of the SNPs and CNVs (Fig 2). In this case, the size of the error in the AF analysis will not be reduced with even larger sample sizes, and the common SNPs in the other allele can always be detected with significance in the association analysis, thus causing false positive associations.

To illustrate the importance of recognizing these distorted calculations, we visualized and matched the significant SNPs with common and pathogenic CNVs according to published databases. After comparing the known significant SNPs with the common CNVs (MacDonald et al. 2014; Welter et al. 2014), ~10% of the SNPs identified by previous GWAS overlap with common CNVs, indicating that the overlap of SNPs and CNVs is very common and cannot be ignored (Table 3). Moreover, the significant SNPs were more likely located in the same regions as the pathogenic CNVs, which also suggested that such a distorted calculation could occur in published GWAS studies (Table 3). The bias in the association analysis caused by the coexistence of CNVs and SNPs depends on their corresponding allelic frequencies. Though the pathogenic CNVs in our illustration were not common in the general population, special attention should still be paid when studying patients with specific diseases, as shown in our analysis of the CS patients.

Similarly, multi-allelic sites, for which the genomic position may vary to more than one alternate nucleotide, have been reported in genomic analyses (Lindsay et al. 2006). Multi-allelic sites could be even more common in regions of gene duplication. It was estimated that multi-allelic sites account for more than 6.4% (600,072 of 9,462,741) of all autosomal positions with SNVs in the ExAC database, and the importance of identifying multi-allelic variants increases in concert with increasing sample size (Campbell et al. 2016). In this case, failure to map the DNA sequence reads and identify multi-allelic sites could contribute to the erroneous inference of variant sites. In association analysis, this causes practical issues because the patterns and significance of association power could be changed if the multi-allelic sites are not recognized (Fig 2). However, the scenario is further complicated by the recognition of PSVs (paralogous sequence variants) caused by mis-mapping of variation in different copies of a segmental duplication (Bailey et al. 2002; Lindsay et al. 2006; Lupski 2003). It is challenging to distinguish the PSVs from multisite variants without proper data collections and toolsets. And the importance of multi-allelic sites is far more fully recognized in public databases at present.

Investigation of coexistence of CNVs and SNPs is recommended in analyzing genomic data of complex diseases

Pathogenesis of complex disease is heterogeneous considering the multiple genetic and environmental contributing factors. Therefore, the inference of genetic determinants for such diseases is particularly challenging. So far, numerous SNPs have been reported to be associated with complex diseases, largely supported by linkage disequilibrium in large case-control paired cohorts of up to hundreds of thousands of individuals (Manolio 2009). However, several SNPs and their overlapping CNVs were reported to be both associated with the same disease or trait, such as schizophrenia (rs1009153 and rs4778334 in 15q11.2 deletion (Zhao et al. 2013), rs165774 in 22q11.2 deletion and 22q11.2 duplication (Higashiyama et al. 2016; Rees et al. 2016)), obesity/body mass index (rs12446632 in 16p12.3 deletion (Locke et al. 2015; Yang et al. 2013a)), and breast cancer (rs17370615 and rs6001376 in deletions of *APOBEC3* gene (Long et al. 2013; Marouf et al. 2016)) (Table S6). This might cause the overestimation or misannotation of association signals of the SNPs in the same regions.

Assisted by a variety of algorithms, data from microarray-based GWAS can also be used for CNV detection. Compared with other techniques, this is a cost-effective approach particularly if GWAS efforts have already been performed or even published. Technically, both proprietary software from Illumina or Affymetrix and academically developed packages such as QuantiSNP or PennCNV have been made available for CNV calling purposes (Colella et al. 2007; Wang et al. 2007). The use of a second algorithm on the same dataset to produce more informative and reliable results is always recommended (Pinto et al. 2011). However, if only a limited number of selected SNPs are included in the analysis, it is difficult to detect CNV in the same region, unless a targeted CNV assay is included (Itsara et al. 2009). It has been reported that 82% of the genotyped sites can be tagged by the HapMap SNPs near the CNV; however, each high-density genome-wide SNP platform effectively tagged only about ~50% of the common deletions (Conrad et al. 2010; Cooper et al. 2008). Therefore, limited sensitivity of PennCNV and QuantiSNP impedes the application of the statistical models for joint analysis of CNV and SNP (Dellinger et al. 2010; Marenne et al. 2013; McCarroll et al. 2008b). Additional CNV detection through aCGH (Lai et al. 2005), MLPA (multiplex ligation-dependent probe amplification) (Shen and Wu 2009) or qPCR (Weaver et al. 2010) is needed under these circumstances. In addition, droplet digital PCR (ddPCR) can further distinguish amplification and copy number gains beyond duplication [e.g. triplication, quadruplication (Gu et al. 2016)]. Moreover, based on the advances in SNP calling from next-generation sequencing data (Nielsen et al. 2011), CNVs can easily be detected by particular algorithms with an adequate read depth (Goodwin et al. 2016; Hormozdiari et al. 2009; Iossifov et al. 2014).

Thus, we suggest the careful validation and detection of CNVs when evaluating the association between SNPs and diseases (Fig 4). For population studies, the existence of CNVs in the same loci as identified SNPs needs to be investigated because even rare CNV can result in misannotation of the pathogenicity of the SNPs.

This study has several limitations. Firstly, only two realistic examples in which the SNPs was overlapping with the CNVs were assessed in our study, which may reduce the

generalizability of the pipeline we suggested. Secondly, we investigated the universality of the potential distorted calculations by measuring the overlap of significant SNVs identified by GWAS with common CNVs and pathogenic/likely pathogenic CNVs in public database, in which the ethnic information is not complete. However, the result might be biased without pairing the SNPs and CNVs in same ethnic populations. Therefore, we suggest each association study of SNPs was detected after SNP genotyping and fine mapping in the discovery cohort. If any CNV exists in the same region of association in the same population, the potential overestimation should be corrected. Thus, more accurate association and the real pattern of variation in specific loci can be revealed through such fine resolution and experimental approach.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank all the individuals involved in the study for their participation.

Funding

This research was funded in part by the National Natural Science Foundation of China (81501852 to N.W., 81472046 and 81772299 to Z.W., 81472045 and 81772301 to G.Q.), Beijing Natural Science Foundation (7172175 to N.W.), Beijing Nova Program (Z161100004916123 to N.W.), Beijing Nova Program Interdisciplinary Collaborative Project (xxjc201717 to N.W.), 2016 Milstein Medical Asian American Partnership Foundation Fellowship Award in Translational Medicine (to N.W.), The Central Level Public Interest Program for Scientific Research Institute (2016ZX310177 to N.W.), PUMC Youth Fund & the Fundamental Research Funds for the Central Universities (3332016006 to N.W.), CAMS Initiative Fund for Medical Sciences (2016-I2M-3-003 to G.Q. and N.W., 2016-I2M-2-006 and 2017-I2M-2-001 to Z.W.), the Distinguished Youth Foundation of Peking Union Medical College Hospital (JQ201506 to N.W.), the 2016 PUMCH Science Fund for Junior Faculty (PUMCH-2016-1.1 to N.W.), the US National Institutes of Health, National Institute of Neurological Disorders and Stroke (NINDS R01NS058529 and R35NS105078 to J.R.L.), National Human Genome Research Institute/National Heart, Lung, and Blood Institute (NHGRI/NHLBI UM1 HG006542 to J.R.L.), the National Human Genome Research Institute (NHGRI K08 HG008986 to J.E.P.).

References

- Albers CA et al. (2012) Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit *RBM8A* causes TAR syndrome. *Nat Genet* 44:435–439, S431–432. doi:10.1038/ng.1083 [PubMed: 22366785]
- Andrews T et al. (2015) The clustering of functionally related genes contributes to CNV-mediated disease. *Genome Res* 25:802–813. doi:10.1101/gr.184325.114 [PubMed: 25887030]
- Antonacci F et al. (2010) A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat Genet* 42:745–750. doi:10.1038/ng.643 [PubMed: 20729854]
- Bailey JA et al. (2002) Recent segmental duplications in the human genome. *Science* 297:1003–1007. doi:10.1126/science.1072047 [PubMed: 12169732]
- Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* 11:1005–1017. doi: 10.1101/gr.187101 [PubMed: 11381028]
- Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265. doi:10.1093/bioinformatics/bth457 [PubMed: 15297300]

- Boone PM et al. (2013) Deletions of recessive disease genes: CNV contribution to carrier states and disease-causing alleles. *Genome Res* 23:1383–1394. doi:10.1101/gr.156075.113 [PubMed: 23685542]
- Campbell IM et al. (2016) Multiallelic positions in the human genome: challenges for genetic analyses. *Hum Mutat* 37:231–234. doi:10.1002/humu.22944 [PubMed: 26670213]
- Carvalho CM, Lupski JR (2008) Copy number variation at the breakpoint region of isochromosome 17q. *Genome Res* 18:1724–1732. doi:10.1101/gr.080697.108 [PubMed: 18714090]
- Carvalho CM, Lupski JR (2016) Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* 17:224–238. doi:10.1038/nrg.2015.25 [PubMed: 26924765]
- Chick JM et al. (2016) Defining the consequences of genetic variation on a proteome-wide scale. *Nature* 534:500–505. doi:10.1038/nature18270 [PubMed: 27309819]
- Coe BP et al. (2014) Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* 46:1063–1071. doi:10.1038/ng.3092 [PubMed: 25217958]
- Colella S et al. (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 35:2013–2025. doi:10.1093/nar/gkm076 [PubMed: 17341461]
- Conrad DF et al. (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464:704–712. doi:10.1038/nature08516 [PubMed: 19812545]
- Cooper GM et al. (2011) A copy number variation morbidity map of developmental delay. *Nat Genet* 43:838–846. doi:10.1038/ng.909 [PubMed: 21841781]
- Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet* 40:1199–1203. doi:10.1038/ng.236 [PubMed: 18776910]
- Crossa J et al. (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186:713–724. doi:10.1534/genetics.110.118521 [PubMed: 20813882]
- Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ (2010) Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res* 38:e105. doi:10.1093/nar/gkq040 [PubMed: 20142258]
- Fei Q et al. (2010) The association analysis of *TBX6* polymorphism with susceptibility to congenital scoliosis in a Chinese Han population. *Spine (Phila Pa 1976)* 35:983–988. doi:10.1097/BRS.0b013e3181bc963c [PubMed: 20228709]
- Flannick J, Florez JC (2016) Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat Rev Genet* 17:535–549. doi:10.1038/nrg.2016.56 [PubMed: 27402621]
- Flint J, Eskin E (2012) Genome-wide association studies in mice. *Nat Rev Genet* 13:807–817. doi:10.1038/nrg3335 [PubMed: 23044826]
- Flipsen-ten Berg K et al. (2007) Unmasking of a hemizygous *WFS1* gene mutation by a chromosome 4p deletion of 8.3 Mb in a patient with Wolf-Hirschhorn syndrome. *Eur J Hum Genet* 15:1132–1138. doi:10.1038/sj.ejhg.5201899 [PubMed: 17637805]
- Franke A et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 42:1118–1125. doi:10.1038/ng.717 [PubMed: 21102463]
- Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ (2004) Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet* 36:861–866. doi:10.1038/ng1401 [PubMed: 15247918]
- Genomes Project C et al. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073. doi:10.1038/nature09534 [PubMed: 20981092]
- Genomes Project C et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65. doi:10.1038/nature11632 [PubMed: 23128226]
- Giampietro PF et al. (2003) Congenital and idiopathic scoliosis: clinical and genetic aspects. *Clin Med Res* 1:125–136. doi:10.3121/cmr.1.2.125 [PubMed: 15931299]
- Gonzaga-Jauregui C et al. (2015) Exome sequence analysis suggests that genetic burden contributes to phenotypic variability and complex neuropathy. *Cell Reports* 12:1169–1183. doi:10.1016/j.celrep.2015.07.023 [PubMed: 26257172]

- Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17:333–351. doi:10.1038/nrg.2016.49 [PubMed: 27184599]
- Gu S et al. (2016) Mechanisms for the generation of two quadruplications associated with split-hand malformation.. *Hum Mutat* 37:160–164. doi:10.1002/humu.22929 [PubMed: 26549411]
- Han MR et al. (2016) Genome-wide association study in East Asians identifies two novel breast cancer susceptibility loci. *Hum Mol Genet* 25:3361–3371. doi:10.1093/hmg/ddw164 [PubMed: 27354352]
- Higashiyama R et al. (2016) Association of copy number polymorphisms at the promoter and translated region of *COMT* with Japanese patients with schizophrenia. *Am J Med Genet B Neuropsychiatr Genet* 171B:447–457. doi:10.1002/ajmg.b.32426 [PubMed: 26852906]
- Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* 38:82–85. doi:10.1038/ng1695 [PubMed: 16327809]
- Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* 19:1270–1278. doi:10.1101/gr.088633.108 [PubMed: 19447966]
- International Schizophrenia C et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460:748–752. doi:10.1038/nature08185 [PubMed: 19571811]
- Iossifov I et al. (2014) The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515:216–221. doi:10.1038/nature13908 [PubMed: 25363768]
- Itsara A et al. (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* 84:148–161. doi:10.1016/j.ajhg.2008.12.014 [PubMed: 19166990]
- Kaminsky EB et al. (2011) An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genet Med* 13:777–784. doi:10.1097/GIM.0b013e31822c79f9 [PubMed: 21844811]
- Lai WR, Johnson MD, Kucherlapati R, Park PJ (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* 21:3763–3770. doi:10.1093/bioinformatics/bti611 [PubMed: 16081473]
- Landrum MJ et al. (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 44:D862–868. doi:10.1093/nar/gkv1222 [PubMed: 26582918]
- Lee SH et al. (2012) Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* 44:247–250. doi:10.1038/ng.1108 [PubMed: 22344220]
- Lefebvre M et al. (2016) Autosomal recessive variations of *TBX6*, from congenital scoliosis to spondylocostal dysostosis. *Clin Genet* 91: 908–912. doi:10.1111/cge.12918
- Lek M et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–291. doi:10.1038/nature19057 [PubMed: 27535533]
- Lindsay SJ, Khajavi M, Lupski JR, Hurles ME (2006) A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. *Am J Hum Genet* 79:890–902. doi:10.1086/508709 [PubMed: 17033965]
- Locke AE et al. (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518:197–206. doi:10.1038/nature14177 [PubMed: 25673413]
- Long J et al. (2013) A common deletion in the *APOBEC3* genes and breast cancer risk. *J Natl Cancer Inst* 105:573–579. doi:10.1093/jnci/djt018 [PubMed: 23411593]
- Lupski JR (2003) 2002 Curt Stern Award Address. Genomic disorders recombination-based disease resulting from genomic architecture. *Am J Hum Genet* 72:246–252. doi:10.1086/346217 [PubMed: 12596790]
- Lupski JR et al. (1991) DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* 66:219–232. doi:10.1016/0092-8674(91)90613-4 [PubMed: 1677316]
- MacArthur DG et al. (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature* 508:469–476. doi:10.1038/nature13127 [PubMed: 24759409]

- MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 42:D986–992. doi:10.1093/nar/gkt958 [PubMed: 24174537]
- Manolio TA (2009) Cohort studies and the genetics of complex disease. *Nat Genet* 41:5–6. doi: 10.1038/ng0109-5 [PubMed: 19112455]
- Marenne G, Chanock SJ, Malats N, Genin E (2013) Advantage of using allele-specific copy numbers when testing for association in regions with common copy number variants. *PLoS One* 8:e75350. doi:10.1371/journal.pone.0075350 [PubMed: 24040408]
- Marouf C, Gohler S, Filho MI, Hajji O, Hemminki K, Nadifi S, Forsti A (2016) Analysis of functional germline variants in *APOBEC3* and driver genes on breast cancer risk in Moroccan study population. *BMC Cancer* 16:165. doi:10.1186/s12885-016-2210-8 [PubMed: 26920143]
- Matisse TC et al. (1994) Detection of tandem duplications and implications for linkage analysis. *Am J Hum Genet* 54:1110–1121. [PubMed: 8198134]
- McCarroll SA et al. (2008a) Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nat Genet* 40:1107–1112. doi:10.1038/ng.215 [PubMed: 19165925]
- McCarroll SA et al. (2008b) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 40:1166–1174. doi:10.1038/ng.238 [PubMed: 18776908]
- Miller DT et al. (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* 86:749–764. doi:10.1016/j.ajhg.2010.04.006 [PubMed: 20466091]
- Mills RE et al. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470:59–65. doi:10.1038/nature09708 [PubMed: 21293372]
- NCBI:dbGaP Genotyping NIGMS Chromosomal Aberration and Inherited Disorder Samples. (Study Accession ID:phs000269.v1.p1) https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000269.v1.p1
- Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 12:443–451. doi:10.1038/nrg2986 [PubMed: 21587300]
- Pang AW et al. (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* 11:R52. doi:10.1186/gb-2010-11-5-r52 [PubMed: 20482838]
- Pinto D et al. (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* 29:512–520. doi:10.1038/nbt.1852 [PubMed: 21552272]
- Purcell S et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575. doi:10.1086/519795 [PubMed: 17701901]
- Rees E et al. (2016) Analysis of intellectual disability copy number variants for association with schizophrenia. *JAMA Psychiatry* 73:963–969. doi:10.1001/jamapsychiatry.2016.1831 [PubMed: 27602560]
- Shen Y, Wu BL (2009) Designing a simple multiplex ligation-dependent probe amplification (MLPA) assay for rapid detection of copy number variants in the genome. *J Genet Genomics* 36:257–265. doi:10.1016/S1673-8527(08)60113-7 [PubMed: 19376486]
- Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. *Annu Rev Med* 61:437–455. doi:10.1146/annurev-med-100708-204735 [PubMed: 20059347]
- Takeda K et al. (2017) Compound heterozygosity for null mutations and a common hypomorphic risk haplotype in *Tbx6* causes congenital scoliosis. *Human Mutation* 38:317–323. doi:10.1002/humu.23168 [PubMed: 28054739]
- Tarailo-Graovac M, Zhu JYA, Matthews A, van Karnebeek CDM, Wasserman WW (2017) Assessment of the ExAC data set for the presence of individuals with pathogenic genotypes implicated in severe Mendelian pediatric disorders. *Genet Med* 19:1300–1308. doi:10.1038/gim.2017.50 [PubMed: 28471432]
- Tennessen JA et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337:64–69. doi:10.1126/science.1219240 [PubMed: 22604720]

- Trivellin G et al. (2014) Gigantism and acromegaly due to Xq26 microduplications and GPR101 mutation. *N Engl J Med* 371:2363–2374. doi:10.1056/NEJMoa1408028 [PubMed: 25470569]
- Visscher PM, Yang J, Goddard ME (2010) A commentary on ‘common SNPs explain a large proportion of the heritability for human height’ by Yang et al. (2010). *Twin Res Hum Genet* 13:517–524. doi:10.1375/twin.13.6.517 [PubMed: 21142928]
- Wang K et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17:1665–1674. doi:10.1101/gr.6861907 [PubMed: 17921354]
- Weaver S et al. (2010) Taking qPCR to a higher level: Analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. *Methods* 50:271–276. doi:10.1016/j.ymeth.2010.01.003 [PubMed: 20079846]
- Weischenfeldt J, Symmons O, Spitz F, Korbel JO (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* 14:125–138. doi:10.1038/nrg3373 [PubMed: 23329113]
- Wellcome Trust Case Control C et al. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 464:713–720. doi:10.1038/nature08979 [PubMed: 20360734]
- Welter D et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42:D1001–1006. doi:10.1093/nar/gkt1229 [PubMed: 24316577]
- Wu N et al. (2015) *TBX6* null variants and a common hypomorphic allele in congenital scoliosis. *N Engl J Med* 372:341–350. doi:10.1056/NEJMoa1406829 [PubMed: 25564734]
- Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88:76–82. doi:10.1016/j.ajhg.2010.11.011 [PubMed: 21167468]
- Yang TL, Guo Y, Li SM, Li SK, Tian Q, Liu YJ, Deng HW (2013a) Ethnic differentiation of copy number variation on chromosome 16p12.3 for association with obesity phenotypes in European and Chinese populations *Int J Obes (Lond)* 37:188–190. doi:10.1038/ijo.2012.31 [PubMed: 22391884]
- Yang Y et al. (2013b) Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N Engl J Med* 369:1502–1511. doi:10.1056/NEJMoa1306555 [PubMed: 24088041]
- Yoon WH et al. (2017) Loss of nardilysin, a mitochondrial co-chaperone for alpha-ketoglutarate dehydrogenase, promotes mTORC1 activation and neurodegeneration. *Neuron* 93:115–131. doi:10.1016/j.neuron.2016.11.038 [PubMed: 28017472]
- Yuan B et al. (2015) Comparative Genomic Analyses of the Human *NPH1* Locus Reveal Complex Genomic Architecture and Its Regional Evolution in Primates. *PLoS Genet* 11:e1005686. doi:10.1371/journal.pgen.1005686 [PubMed: 26641089]
- Zarrei M, MacDonald JR, Merico D, Scherer SW (2015) A copy number variation map of the human genome. *Nat Rev Genet* 16:172–183. doi:10.1038/nrg3871 [PubMed: 25645873]
- Zhao Q et al. (2013) Rare CNVs and tag SNPs at 15q11.2 are associated with schizophrenia in the Han Chinese population. *Schizophr Bull* 39:712–719. doi:10.1093/schbul/sbr197 [PubMed: 22317777]
- Zody MC et al. (2008) Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet* 40:1076–1083. doi:10.1038/ng.193 [PubMed: 19165922]

Fig 1. Distorted calculations of the allele frequency of the *TBX6* haplotype

A. Three common single-nucleotide polymorphisms (SNPs) in *TBX6*. rs3809627 (C/A) and rs3809624 (T/C) are located in the 5' non-coding region, while rs2289292 (C/T) is a synonymous SNP in the last exon. B. The coexistence of a deletion and a T-C-A haplotype at 16p11.2 could cause a misannotation and lead to a distorted calculation of the T-C-A allele frequency.

Fig 2. The coexistence of various CNVs causes allele frequency distorted calculations of SNPs in the same genomic regions

To simulate the effect of distorted calculations, allele frequencies (AFs) of one particular SNP of interest were compared between cases and controls of various sample sizes using Pearson χ^2 test (two-sided). $p < 0.05$ is recognized as statistically significant, illustrated as the dark-blue areas. $y=x$ was demonstrated as the black dotted line. B. When the sample size is fixed (N=1000), the effect of a coexisting CNV which is not associated with any particular SNP is similar to the effect of sample size variation. While deletion is statistically similar to a decrease of sample

size, duplication is similar to an increase of sample size. C. When association exists between the CNV and the SNP, the significance is overestimated or underestimated depending on the particular type of the CNV and the coexistence frequency between the CNV and the SNP. In the case simulated above, the sample size is fixed ($N=1000$) and the CNV is entirely associated with the SNP (coexistence frequency=10%), the areas of significance deviate to opposite directions depending on the existence of deletions or duplications. D. The validated set of data from the congenital scoliosis cohort ($N=161$) is applied in the model above. If the coexistence of 16p11.2 deletion is not considered in the association analysis, rs2289292 (black) and rs3809624 (red) are found in the area of significance (left). When the frequency of 16p11.2 is considered in the correction of the AFs of the SNPs, none of the three SNPs is in the area of significance (right). Thus, the previous misannotation of the association between the two SNPs and congenital scoliosis is corrected.

Fig 3. The genome-wide atlas of CNVs and significant SNPs identified by GWAS

Circos plots are used to show the coexistence of the SNPs which were significantly associated with certain clinical conditions or phenotypes and common CNVs (frequency > 1%) or pathogenic/likely pathogenic CNVs. The human hg19 genome was used as the reference genome in this case. The significant SNPs were obtained from The National Human Genome Research Institute Catalog of Published GWAS Catalog, the common CNVs were acquired from the online Database of Genomic Variants and the pathogenic CNVs were acquired from the UCSC database (<https://genome.ucsc.edu>). The biased overlap between these SNPs and the common CNVs and pathogenic/likely pathogenic CNVs suggested that the distorted calculations could occur in the GWAS studies. A. ~10% of the significant SNPs (2,042/20,726) identified by GWAS were in the same regions with the common CNVs which were observed in only 4.2% of the human genome. The height of each radiation represented the frequency of each SNP or CNV. B. Most of the significant SNPs (99.69%) identified by GWAS were in the same regions with the pathogenic/likely pathogenic CNVs, while the pathogenic CNVs were observed in 97.63% of the human genome. The height of each radiation of CNV represented the number of times that each CNV had been reported in the database. SNP, single-nucleotide polymorphisms; CNV, copy number variants

Fig 4. A suggested framework to avoid misannotation when evaluating the association between SNPs and diseases.

First, the association of SNPs was detected after SNP genotyping and fine mapping in the discovery cohort. If any CNV exists in the same region of association, the potential overestimation should be corrected. Additionally, the association analyses should be re-evaluated after the correction. Furthermore, the biological and informatics validations were still needed in larger cohorts to investigate the function of SNPs in conjunction with their compound CNVs to reveal the underlying inheritance model.

Table 1.
The association analysis of three SNPs in *TBX6* investigated in 161 CS cases and 166 controls

SNP	Allele Frequency			HWE			Corrected Allele Frequency		
	Alleles	χ^2	p	OR (95% CI)	p		Alleles	χ^2	p
1(rs2289292)	T C						T C		
Case	184 138				0.151		172 138		
Control	163 169	4.249	0.042	1.382 (1.016-1.882)	0.877		163 169	2.621	0.114
									1.292 (0.947-1.763)
2(rs3809624)	C T						C T		
Case	188 134				0.517		176 134		
Control	164 168	5.313	0.023	1.437 (1.055-1.957)	0.756		164 168	3.502	0.069
									1.346 (0.986-1.835)
3(rs3809627)	A C						A C		
Case	205 117				0.232		193 117		
Control	188 144	3.376	0.067	1.342 (0.980-1.837)	0.754		188 144	2.107	0.149
									1.264 (0.921-1.733)

The difference of allele frequencies (AFs) between case and control groups were compared using Pearson χ^2 test (two-sided). Wild-type/mutation: C/T in rs2289292, T/C in rs3809624 and C/A in rs3809627. The Hardy-Weinberg equilibrium (HWE) of the AF of the detected SNPs were validated in our data using the goodness-of-fit Chi-Square test. $p < 0.05$ (two-sided) was considered as statistically significant.

Abbreviation: SNP, single nucleotide polymorphisms; CS, congenital scoliosis; OR, odds ratio; CI, confidence interval; HWE, Hardy-Weinberg equilibrium.

Table 2. The haplotypes association analysis of three SNPs in *TBX6* investigated in 161 CS cases and 166 controls

Haplotype	Before corrected			Corrected		
	Frequency in case	Frequency in control	OR with 95% CI	p	Frequency in case	Frequency in control
T-C-A	0.550	0.482	1.418 (1.021-1.969)	0.038	0.512	0.482
T-C-C	0.003	0.009	0.427 (0.044-4.164)	0.635		
T-T-A	0.006	0.000	2.282 (1.984-2.625)	0.195		
T-T-C	0.012	0.000	2.282 (1.984-2.625)	0.039		
C-C-A	0.025	0.003	10.255 (1.264-83.219)	0.013		
C-C-C	0.006	0.000	2.282 (1.984-2.625)	0.195		
C-T-A	0.056	0.081	0.855 (0.448-1.631)	0.744		
C-T-C	0.342	0.425	1	--	0.340	0.424
						1
						--

The haplotypes of case and control groups were compared using Pearson χ^2 test (two-sided). Wild-type/mutation: C/T in rs2289292, T/C in rs3809624 and C/A in rs3809627. Abbreviation: SNP, single nucleotide polymorphisms; CS, congenital scoliosis; OR, odds ratio; CI, confidence interval.

Table 3.

The relation of the significant SNPs and CNVs based on published database.

Total	Significant SNPs (bp)	Whole genome (bp)
	20,726	3,095,677,412
Within the common CNVs	2,042	130,618,321
Outside the common CNVs	18,684	2,965,059,091
OR with 95% CI	2.41 (2.37-2.60)	
<i>p</i>	$<1 \times 10^{-6}$	
Within the pathogenic CNVs	20,658	3,022,313,748
Outside the pathogenic CNVs	68	73,363,664
OR with 95% CI	1.02 (1.00-1.04)	
<i>p</i>	0.03	
Within the pathogenic CNVs (<10 Mb)	20,116	2,842,153,180
Outside the pathogenic CNVs (<10 Mb)	610	253,524,232
OR with 95% CI	1.06 (1.04-1.08)	
<i>p</i>	1.8×10^{-8}	

The distribution of SNPs in regions with or outside CNVs was compared using Pearson χ^2 test (two-sided). The significant SNPs collected from the National Human Genome Research Institute (NHGRI) Catalog of Published GWAS was matched with the locations of common CNVs (population frequency > 1%) acquired from the Database of Genomic Variants (DGV) and pathogenic CNVs acquired from the UCSC database.

Abbreviation: SNP, single nucleotide polymorphisms; CNV, copy number variant; OR, odds ratio; CI, confidence interval.