Archival Report

Biological Psychiatry

Genome-wide Analysis of the Role of Copy Number Variation in Schizophrenia Risk in Chinese

Zhiqiang Li, Jianhua Chen, Yifeng Xu, Qizhong Yi, Weidong Ji, Peng Wang, Jiawei Shen, Zhijian Song, Meng Wang, Ping Yang, Qingzhong Wang, Guoyin Feng, Benxiu Liu, Wensheng Sun, Qi Xu, Baojie Li, Lin He, Guang He, Wenjin Li, Zujia Wen, Ke Liu, Fang Huang, Juan Zhou, Jue Ji, Xingwang Li, and Yongyong Shi

ABSTRACT

BACKGROUND: Compelling evidence suggested the role of copy number variations (CNVs) in schizophrenia susceptibility. Most of the evidence was from studies in populations with European ancestry. We tried to validate the associated CNV loci in a Han Chinese population and identify novel loci conferring risk of schizophrenia.

METHODS: We performed a genome-wide CNV analysis on 6588 patients with schizophrenia and 11,904 control subjects of Han Chinese ancestry.

RESULTS: Our data confirmed increased genome-wide CNV (>500 kb and <1%) burden in schizophrenia, and the increasing trend was more significant when only >1 Mb CNVs were considered. We also replicated several associated loci that were previously identified in European populations, including duplications at 16p11.2, 15q11.2-13.1, 7q11.23, and *VIPR2* and deletions at 22q11.2, 1q21.1-q21.2, and *NRXN1*. In addition, we discovered three additional new potential loci (odds ratio >6, $p < .05$): duplications at 1p36.32, 10p12.1, and 13q13.3, involving many neurodevelopmental and synaptic related genes.

CONCLUSIONS: Our findings provide further support for the role of CNVs in the etiology of schizophrenia.

Keywords: Copy number variation, Han Chinese, Schizophrenia, 1p36.32, 10p12.1, 13q13.3

http://dx.doi.org/10.1016/j.biopsych.2015.11.012

Schizophrenia (SZ) is a major mental illness that results in significant social or occupational dysfunction. It affects about 1% of the population, with heritability estimates of up to 80% (1). Most geneticists view SZ as a complex genetic disease; the etiology of SZ is only partially resolved, and our knowledge is still nascent (2). Dozens of genome-wide association studies (GWASs) of common (single nucleotide polymorphisms [SNPs]) and rare (copy number variations [CNVs]) genetic risk variants were used to find candidate genes or genome regions that contribute to SZ (3–16). Both SNPs and CNVs and the complex interplay between them were involved in the genetic susceptibility to SZ. For SZ genetic risk variants, the common SNPs usually have relatively small effects (odds ratio [OR] <1.5), whereas the rare CNVs have much larger effects (OR = 2–30) (2,17). Because of their rarity, the CNVs make a much smaller contribution to the total risk (2).

Numerous studies have been conducted to identify CNVs underlying susceptibility to SZ (6–15,18–24). The 22q11.2 deletion was the first CNV implicated in SZ and is one of the major known genetic risk factors for SZ (8,10). The prevalence of 22q11.2 deletion in patients with SZ is about .3% (10,20). A recent study showed evidence that duplications of 22q11.2 protect against SZ (6). Deletions on 1q21.1 (10,12), 15q13.3 (7,10,12), 15q11.2 (11,12), 2p16.3 (7,23), 17p12 (11), 3q29 (7,22), 17q12 (21), and distal 16p11.2 (9) and duplications on 15q13.1 (23), 16p11.2 (13,19,22), 16p13.1 (11,18), 1p36.33 (13), *VIPR2* (7,8), and *CGNL1* (13) were found to be associated with SZ. Because of the complexity of SZ and the rarity of CNVs, not all these associations showed very strong statistical evidence. Also, to our knowledge, most of the previous studies are mainly carried out in Caucasian populations. Cross-ethnic validation of the susceptibility CNVs and identifying novel ones in Han Chinese populations is important for understanding the etiology and genetic risk factors for SZ.

In this study, we tried to discover novel SZ susceptibility CNVs and validate the reported CNVs identified in Caucasian populations (14,25) using large SZ samples of Han Chinese origin. We examined rare CNVs in three independent data sets. The first data set contained 2992 patients with SZ and 5176 control subjects, the second data set contained 3596 patients with SZ and 2636 control subjects, and the third data set contained only 4092 control subjects. The total samples analyzed comprised 6588 patients with SZ and 11,904 control subjects.

ISSN: 0006-3223

METHODS AND MATERIALS

Samples

All SZ cases of Chinese ancestry were inpatients or outpatients from various mental health centers in China, as described in our previous report (5). All patients were interviewed by two independent psychiatrists, and SZ was diagnosed according to DSM-IV criteria, with at least a 2-year history of the disease. The control subjects were randomly selected from volunteers (from hospitals and community/school surveys). Written requests were issued to volunteers for evaluation of their medical histories. Subjects with major mental illnesses or positive family history for major mental illness were excluded. All participants provided written informed consent. The study received approval from the local Ethics Committee of Human Genetic Resources.

GWAS Genotyping and Quality Control

The genome-wide genotyping experiments were performed using different platforms for the three data sets. The Affymetrix Genome-Wide Human SNP 6.0 arrays (Affymetrix, Santa Clara, California) were used in data set 1, the Affymetrix Axiom myDesign Human Genotyping arrays (Exome 319 plus customized design) were used in data set 2, and the Illumina HumanOmni 1M-Quad or Human660W-Quad chips (Illumina, Inc., San Diego, California) were used in data set 3. Systematic quality control (QC) for SNP-based GWAS analyses was conducted as described elsewhere (5). Briefly, samples were excluded if they 1) had low QC assessments (contrast QC $<$ .4 for Affymetrix SNP 6.0, dish QC $<$ .82 for Affymetrix Axiom myDesign), suggesting poor quality; 2) had unknown established sex or displayed inconsistent sex; 3) had overall genotype call rates of $<$95%; 4) were a member of a pair of relatedness (PI_HAT values $>$.25, on the basis of identity-by-descent analysis) with the lower call rate; or 5) were the population outliers based on the principal components analysis. Samples after QC were kept for further analysis (Supplemental Table S1).

CNV Calling and QC

For the Affymetrix arrays, the genotyping calls and signal intensity data were generated using Affymetrix Power Tools and then converted into log R ratio (LRR) and B allele frequency (BAF) values using PennCNV-Affy protocol (26). For the Illumina arrays, the genotyping and signal intensity data (GType, LRR, and BAF) were directly exported from the GenomeStudio project files. The subsequent CNV calling analyses were performed using PennCNV according to the manual (26). The clean_cnv.pl program was adopted to merge adjacent calls of the same individual using the fraction threshold (.5, calculated as base pair length). Samples that met the following QC measures were thought to be of good quality for further analysis: 1) within the SD of the normalized intensity (LRR $<$.35); 2) GC base pair wave factor $<$.05; 3) with CNV call count $<$70. Finally, 6558 patients with SZ and 11,904 control subjects were retained for the CNV association analysis (Supplemental Table S1). The CNVs were then excluded if they were covered by $<$10 probes or had a probe density $<$1 probe per 20 kb.

Statistical Analysis of CNVs

The genome-wide CNV burden (number of CNV segments and number of genes overlapped) analyses in patients with SZ versus control subjects were performed using PLINK (27). Permutation procedures (10,000$\times$) were used to assess statistical significance (one side test). Only CNVs with frequencies $<$1% (filtered for frequency in each respective data set) were used, and two CNV sizes ($>$500 kb and $>$1 Mb) were tested in the burden analyses. Gene annotation based on UCSC Genome Browser (University of California, Santa Cruz; http://genome.ucsc.edu/) Human Genome version 19 (hg19)/National Center for Biotechnology Information Build 37 was used throughout the study. To control for differences in the CNV rate between the different arrays, the analyses were conducted separately for data sets 1 and 2. To identify specific genic CNV regions associated with SZ, the CNV frequency between patients with SZ and control subjects was evaluated at each gene using Fisher's exact test. We considered only loci that were nominally significant between patients with SZ and control subjects ($p < .05$) for the novel ones. Also, we excluded loci that met any of the following criteria: 1) captured with poor probes coverage; 2) residing at centromere or telomere proximal regions; 3) arising from a "peninsula" of common CNV; 4) with extremes of GC content; or 5) samples contributing to multiple CNV regions. Only the loci further validated using an independent method were considered. Any locus presented in the Results was manually inspected in the raw CNV calls from each data set to check that CNVs were not filtered from this region.

CNV Validation by Affymetrix CytoScan HD Arrays

The Affymetrix CytoScan HD Array is a high-density microarray designed to detect whole-genome CNV, interrogating the entire genome using $>$2.6 million markers. The experiment was performed following the manufacturer's instructions. The CNV calls were generated using Chromosome Analysis Suite software. The CNVs with a ratio ($>$70%) of the intersection were considered to be true DNA features.

RESULTS

Global CNV Burden Analysis

To investigate the global impact of CNVs on disease susceptibility, global CNV burden analysis was conducted for data sets 1 and 2 for rare CNVs ($<$1% in the sample for each data set) of different sizes ($>$500 kb and $>$1 Mb). As shown in Table 1, in both data sets, the rate of large ($>$500 kb) rare CNVs (deletions and duplications) were higher in patients with SZ (data set 1, case/control ratio = 1.07, $p = .182$; data set 2, ratio = 1.29, $p = .011$), and these CNVs intersected with more genes ("gene count"; date set 1, case/control ratio = 1.23, $p = .066$; data set 2, ratio = 1.30, $p = .031$). When deletions ($>$500 kb) or duplications ($>$500 kb) only were considered, greater genome-wide burdens were also observed for both rate and "gene count" in our SZ sample except the rate of deletions only in the data set 2. However, although the rate of deletions is lower in the SZ sample of data set 2 (ratio = .96), the "gene count" showed a 1.13-fold increase in cases. Thus, a common