Schizophrenia Research xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

# Targeted resequencing of regulatory regions at schizophrenia risk loci: Role of rare functional variants at chromatin repressive states

Javier González-Peñas [a], Jorge Amigo [a,b,c], Luis Santomé [c,1], Beatriz Sobrino [a,b,c], Julio Brenlla [a,d], Santiago Agra [a,d], Eduardo Paz [a,d], Mario Páramo [a,d], Ángel Carracedo [a,b,c], Manuel Arrojo [a,d], Javier Costas [a,*]

[a] Instituto de Investigación Sanitaria (IDIS) de Santiago de Compostela, Complexo Hospitalario Universitario de Santiago de Compostela (CHUS), Servizo Galego de Saúde (SERGAS), Santiago de Compostela, Spain
[b] Grupo de Medicina Xenómica, Universidade de Santiago de Compostela (USC), Santiago de Compostela, Spain
[c] Fundación Pública Galega de Medicina Xenómica, Complexo Hospitalario Universitario de Santiago (CHUS), Servizo Galego de Saúde (SERGAS), Santiago de Compostela, Spain
[d] Servizo de Psiquiatría, Complexo Hospitalario Universitario de Santiago de Compostela, Servizo Galego de Saúde (SERGAS), Santiago de Compostela, Spain

ARTICLE INFO

*Article history:*
Received 24 February 2016
Accepted 24 March 2016
Available online xxxx

*Keywords:*
Next-generation sequencing
Polycomb
CNV
Single nucleotide variant
Rare variant
Chromatin remodeling

ABSTRACT

There is mounting evidence that regulatory variation plays an important role in genetic risk for schizophrenia. Here, we specifically search for regulatory variants at risk by sequencing promoter regions of twenty-three genes implied in schizophrenia by copy number variant or genome-wide association studies. After strict quality control, a total of 55,206 bp per sample were analyzed in 526 schizophrenia cases and 516 controls from Galicia, NW Spain, using the Applied Biosystems SOLiD System. Variants were filtered based on frequency from public databases, chromatin states from the RoadMap Epigenomics Consortium at tissues relevant for schizophrenia, such as fetal brain, mid-frontal lobe, and angular gyrus, and prediction of functionality from RegulomeDB. The proportion of rare variants at polycomb repressive chromatin state at relevant tissues was higher in cases than in controls. The proportion of rare variants with predicted regulatory role was significantly higher in cases than in controls ($P = 0.0028$, OR = 1.93, 95% C.I. = 1.23–3.04). Combination of information from both sources led to the identification of an excess of carriers of rare variants with predicted regulatory role located at polycomb repressive chromatin state at relevant tissues in cases versus controls ($P = 0.0016$, OR = 19.34, 95% C.I. = 2.45–2495.26). The variants are located at two genes affected by the 17q12 copy number variant, *LHX1* and *HNF1B*. These data strongly suggest that a specific epigenetic mechanism, chromatin remodeling by histone modification during early development, may be impaired in a subset of schizophrenia patients, in agreement with previous data.



## 1. Introduction

Recent genomic findings strongly suggest a role of regulatory variation in susceptibility to schizophrenia. For instance, only ten of the 108 genome-wide association studies (GWAS) significant loci identified by the large mega-GWAS of the Psychiatric Genomics Consortium (PGC) may be attributable to a nonsynonymous coding variant (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Variants associated with schizophrenia at GWAS are enriched in expression quantitative trait (eQTL) loci at promoter and enhancer regions (Richards et al., 2012; Roussos et al., 2014). Tag SNPs in linkage disequilibrium with 5′untranslated regions (5′UTR) are the most enriched category among most significant single nucleotide polymorphisms (SNPs) at schizophrenia GWAS in comparison with other gene categories, including exonic SNPs (Schork et al., 2013). 5′UTR is a region with an important role in transcriptional regulation (ENCODE Project Consortium et al., 2007). Copy number variants (CNVs) conferring risk to neurodevelopmental disorders also points to a role for regulation of gene expression in schizophrenia risk, as one probable mechanism for CNV effect is a change in expression levels (Henrichsen et al., 2009). This mechanism has been demonstrated for several genes at the 16p11.2 deletion and duplication, 22.q11.2 deletion, 15q11.2 duplication, or 3q29 deletion (Mehta et al., 2014; Migliavacca et al., 2015; Ye et al., 2012). Finally, whole exome sequencing failed to find uncommon coding variants of moderate effect (Need et al., 2012; Purcell et al., 2014).

In spite of these hints, there is a lack of studies on regulatory variation in schizophrenia due to the difficulty in interpretation of the consequences of variants. Variation at regulatory regions is more probably associated with changes at specific tissues and developmental states. Therefore, consideration of the spatio-temporal regulation of gene expression may be of relevance to interpret regulatory variation.

* Corresponding author at: Instituto de Investigación Sanitaria (IDIS) de Santiago de Compostela, Hospital Clínico Universitario, edificio Consultas, andar-2, Grupo de xenética psiquiátrica, despacho 15, E-15706, Santiago de Compostela, Spain.
*E-mail address:* javier.costas.costas@sergas.es (J. Costas).
[1] Current address: Health in Code, A Coruña, Spain.

http://dx.doi.org/10.1016/j.schres.2016.03.029
0920-9964/

In this context, several large scale projects arose in recent years, such as the ENCODE Project, whose goal is to build a catalog of functional variants along the genome, the RoadMap Epigenomics Project, focused on changes in chromatin states during development based on histone modification patterns, or BrainSpan, studying transcriptional changes during human brain development (ENCODE Project Consortium, 2012; Kang et al., 2011; Roadmap Epigenomics Consortium et al., 2015). In addition, several tools help with the interpretation of this large amount of data such as RegulomeDB (Boyle et al., 2012) or HaploReg (Ward and Kellis, 2016).

The usefulness of these big datasets has recently become apparent. For instance, using RoadMap Epigenomics data, the PGC detected an enrichment of schizophrenia-associated variants at enhancers active in mid-frontal lobe and angular gyrus (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Roussos et al. (2014) showed that the enrichment of eQTL in schizophrenia-associated variants was greater after using functional annotation for human cerebral cortex, taken from ENCODE and RoadMap Epigenomics Projects. Duan et al. (2014) used ENCODE data from neuronal cell lines and fetal brain to select regulatory elements around the microRNA *MIR137* for schizophrenia-control sequencing and to classify these regions as promoter, enhancer or insulators. Analysis by burden tests detected association of rare variants with schizophrenia, restricted to promoters and enhancers.

In the present work, we sequenced regulatory regions around the transcriptional start sites (TSS) of different genes associated with schizophrenia by GWAS as well as candidate genes at CNVs involved in schizophrenia susceptibility to find rare putative functional variants involved in risk, under the hypothesis that rare variants are not captured by common GWAS arrays. Interpretation of these findings was done using relevant data from big projects such as ENCODE and RoadMap Epigenomics, and predictions of functionality from RegulomeDB.

## 2. Materials and methods

### 2.1. Subjects

A total of 516 schizophrenic patients from the Santiago de Compostela area at the Galician Healthcare System (NW Spain) were included in the present work after written informed consent. Patients were classified as suffering schizophrenia according to DSM-IV criteria after consensus of at least two experienced psychiatrists. Around half of the patients were inpatients under many years of treatment. A total of 516 Galician controls from blood donors at the Galician Transfusion Center (Santiago de Compostela) were also included. Previous work on these subjects allowed removal of cryptic relationships (Carrera et al., 2012). There is no stratification problems as determined on that work. The work was approved by the appropriated Ethical Committee (Comité Ético de Investigación Clínica de Galicia).

### 2.2. Selection of targeted genes

Genes were selected for sequencing based on two different sources, those significant at GWAS or similar studies involving large sample sizes and those located at CNVs associated with schizophrenia. Within each CNV affecting more than one gene, we selected the candidate gene or genes with higher probability to be involved in central nervous system phenotypes according to Gene Ontology biological function as well as literature, mainly Girirajan et al. (2012). A final criteria for selection was based on quality of PCR designs (see below). The final set of genes was *ZNF804A*, *TCF4*, *SLC39A8*, *AKT3*, *DPYD*, and *MIR137* from GWAS (Carrera et al., 2012; O'Donovan et al., 2008; Steinberg et al., 2011; Ripke et al., 2013), and *BCL9* at 1q21.1, *CYFIP1* at 15q11, *MAGEL2* and *CHRNA7* at 15q11.2–13.3, *KCTD13*, *MVP*, and *TAOK2* at 16p11.2, *PMP22* at 17p12, *AATF*, *HNF1B*, and *LHX1* at 17q12, *DGCR8* and *COMT* at 22q11.2, *DLG1* and *PAK2* at 3q29, *CNTNAP2* at 7q36, and *NRXN1* at 2p16 from CNV studies (Girirajan et al., 2012; Malhotra and Sebat, 2012).

### 2.3. Target enrichment by long range PCR

In order to unambiguously assign a regulatory region to a specific gene, we considered only regulatory regions around TSS of selected genes. Regulatory sequences surrounding the TSS are symmetrically distributed, with no bias towards upstream regions (ENCODE Project Consortium et al., 2007). The intensity signals of chromatin immunoprecipitation assays for transcription factors-revealed enrichment extending at least $\pm$1000 bp. Therefore, we tried to design PCR assays to amplify regions about 2000 bp centered on the TSS. Design of primers at TSS regions is more difficult than at coding regions due to particular characteristics of sequence such as repetitive motifs or high GC content. Because of that, long range PCRs were attempted at many regions, including several different TSS at each one of the selected genes. Each assay was tested in a few samples using at least one of three different high fidelity polymerases, Q5 High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA), Phusion High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA) or *PfuUltra* II fusion HS DNA polymerase (Agilent technologies, Santa Clara, CA), as some regions performed well only using one of them. After these preliminary PCRs, we finally selected those regions with consistent high intensity and specificity of the amplification signal (Supplementary Table 1). The total length of the targeted sequences was 73,997 bp. Prior to PCR, samples were pooled in equimolar quantities in groups of six by diagnosis, 86 pools of cases and 86 pools of controls, after three independent measurements of concentration using the Qubit dsDNA BR assay and a Qubit 2.0 fluorimeter (Invitrogen, Carlsbad, CA).

### 2.4. Resequencing by applied biosystems SOLiD system

Resequencing was carried out by the 5500xl SOLiD System (Applied Biosystems, Carlsbad, CA) using 75 + 25 bp paired-end libraries, following manufacturer's instructions. The 172 DNA pools were split in four sequencing pools of equal number of cases and controls. Each DNA pool was barcoded during library construction to allow identification. Mapping of reads to the hg19 version of the human genome was initially done with LifeScope (Applied Biosystems, Carlsbad, CA). Duplicates were removed with Picard tools v 1.106, (http://picard.sourceforge.net) following by local realignment around indels and base quality score recalibration using GATK (DePristo et al., 2011).

### 2.5. Single nucleotide variants (SNVs) detection

Only those positions with >95% of DNA pools presenting a depth of coverage higher than 180× were subject to further analyses. Prior to SNV calling, BAM alignment files were converted to PILEUP files with SAMtools (Li, 2011). SNV calling was done with VarScan 2, a program specially designed for detection of SNVs in pools (Koboldt et al., 2012). The following parameters were defined: minimum read depth at a position to make a call, 180; minimum supporting reads at a position to call variants, 4; minimum base quality at a position to count a read, 0; default *P*-value threshold for calling variants, 0.05; minimum variant allele frequency threshold, 0.01. Inspection of results revealed several positions with many pools presenting SNVs at frequency near 0.01, considerably lower than the expected frequency for alleles present once in a pool (0.083), suggesting that there were many false positive results. Therefore, two additional filters were applied using a custom R script, minimum variant allele frequency threshold, 0.025; and minimum number of reads presenting each allele at each strand, 3. Finally, to eliminate variants prone to false positives, we compared SNV calls before and after application of these two filters and removed those positions that showed any difference in the number of pools presenting

Please cite this article as: González-Peñas, J., et al., Targeted resequencing of regulatory regions at schizophrenia risk loci: Role of rare functional variants at chromatin repressive s..., Schizophr. Res. (2015), http://dx.doi.org/10.1016/j.schres.2016.03.029