Neuroscience Research 79 (2014) 22–33

Contents lists available at ScienceDirect

Neuroscience Research

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

Comprehensive survey of CNVs influencing gene expression in the human brain and its implications for pathophysiology☆

Divya Mehta[a,b,1], Kazuya Iwamoto[c], Junko Ueda[b], Miki Bundo[c], Naoki Adati[d,2], Toshio Kojima[d,3], Tadafumi Kato[b,*]

[a] Max Planck Institute of Psychiatry, Munich 80804, Germany
[b] Laboratory for Molecular Dynamics of Mental Disorders, RIKEN Brain Science Institute, Saitama 351-0198, Japan
[c] Department of Molecular Psychiatry, Graduate School of Medicine, University of Tokyo, Tokyo 113-8654, Japan
[d] Comparative Systems Biology Team, RIKEN Genomic Sciences Center, Yokohama 230-0045, Japan

ARTICLE INFO

Article history:
Received 8 July 2013
Received in revised form 22 October 2013
Accepted 29 October 2013
Available online 6 November 2013

Keywords:
Copy number variation (CNV)
Human post-mortem brains
Schizophrenia
Bipolar disorder
Transcriptome
Genomics

ABSTRACT

Copy number variations (CNVs) contribute to neuropsychiatric diseases, which may be partly mediated by their effects on gene expression. However, few studies have assessed the influence of CNVs on gene expression in the brain. The objective was to perform an unbiased comprehensive survey of influence of CNVs on gene expression in human brain tissues. CNV regions (CNVRs) were identified in 72 individuals (23 schizophrenia, 23 bipolar disorder and 26 controls). Significant associations between the CNVRs and gene expression levels were observed for 583 CNVR-expression probe pairs (293 unique eCNVRs and 429 unique transcripts), after corrections for multiple testing and controlling the effect of the number of subjects with CNVRs by label swapping permutations. These CNVRs affecting gene expression (eCNVRs) were significantly enriched for rare/low frequency ($p = 1.087 \times 10^{-10}$) and gene-harboring CNVRs ($p = 1.4 \times 10^{-6}$). Transcripts overlapping CNVRs were significantly enriched for glutathione metabolism and oxidative stress only for cases but not for controls. Moreover, 72 (24.6%) of eCNVRs were located within the chromosomal aberration regions implicated in psychiatric-disorders: 16p11.2, 1q21.1, 22q11.2, 3q29, 15q11.2, 17q12 and 16p13.1. These results shed light on the mechanism of how CNVs confer a risk for psychiatric disorders.

1. Introduction

A key objective in genetic research is to link genomic variation to phenotype differences to uncover normal as well as pathological variation. The influence of single nucleotide polymorphisms on phenotypic variation has been extensively studied; however, it is only recently that other DNA alterations such as copy number variations are being investigated. Copy number variations (CNVs) are DNA segments present at variable copy numbers and owing to their large size, contribute to a substantial proportion of the variation in the human genome (Ionita-Laza et al., 2009; Redon et al., 2006). Among the CNVs, rare CNVs are of more interest because they are presumably enriched in de novo events. Under the rare-variant common disease hypothesis, multiple rare variants with high effect sizes in aggregation, contribute substantially to the illness, hence these rare variants are of great interest since they have not been subject to selection as yet (Zhang et al., 2009). Rare copy number variations have been reported in individuals with neurological and psychiatric disorders such as schizophrenia (International Schizophrenia Consortium, 2008; Levinson et al., 2011; McCarthy et al., 2009), autism (Hedges et al., 2012), bipolar disorder (Zhang et al., 2009) and mental retardation (Guilmatre et al., 2009).

Although a large number of CNVs have been identified in a variety of different species and range of diseases, the functional impact of CNVs at the molecular level remains largely unexplored. One way to assess the functional impact of copy number variations is via its

Abbreviations: CNVs, copy number variations; LCLs, lymphoblastoid cell lines; CNVRs, copy number variable regions; eCNVRs, expression influencing copy number variable regions; kb, kilo base pairs; Mb, mega base pairs; fdr, false discovery rate.

* Corresponding author at: Laboratory for Molecular Dynamics of Mental Disorders, RIKEN Brain Science Institute, Hirosawa 2-1, Wako, Saitama 351-0198, Japan. Tel.: +81 48 467 6949; fax: +81 48 467 6947.
E-mail address: kato@brain.riken.jp (T. Kato).
[1] Current address: Complex Trait Genomics Group, Queensland Brain Institute, Upland Road St Lucia, Brisbane 4067, Australia.
[2] Current address: Research Equipment Center, Hamamatsu University School of Medicine, Hamamatsu 431-3192, Japan.
[3] Current address: Research Center for Physical Fitness, Sports and Health, Toyohashi University of Technology, Toyohashi, 441-8580, Japan.

0168-0102/$ – see front matter © 2013 The Authors. Published by Elsevier Ireland Ltd and the Japan Neuroscience Society. All rights reserved.
http://dx.doi.org/10.1016/j.neures.2013.10.009

effect on different cellular processes such as gene expression levels. The initial study exploring the transcriptome-wide impact of CNVs on gene expression profiles in lymphoblastoid cell lines (LCLs) identified that approximately 20% of variation in gene expression could be attributed to copy number variations in the genome (Stranger et al., 2007). While there is a plethora of studies assessing the influence of single nucleotide polymorphisms on gene expression profiles, to the best of our knowledge, there are only four studies interrogating the influence of CNVs on gene expression in humans. Moreover, due to the limited availability of human tissues such as brain samples, three of the four studies assessing the influence of CNVs on gene expression in normal tissues till date have been performed on LCLs (Luo et al., 2012; Schlattl et al., 2011; Stranger et al., 2007) while only one recent study (Ye et al., 2012) has assessed gene expression in the human brain.

Integration of gene expression and CNV data will allow the prioritization of CNV-harboring candidate regions where the CNVs significantly alter gene expression levels of transcripts thereby providing evidence of a downstream functional consequence. The aim of this study was to perform a comprehensive and unbiased genome-wide search for functional CNVs in the human brain and to interrogate the nature of these CNVs.

2. Materials and methods

2.1. Samples

Gene expression and copy number variations (CNVs) were obtained from prefrontal cortices of postmortem brains of 105 individuals (35 controls, 35 bipolar disorder [one of which was later excluded due to the alteration of diagnosis] and 35 schizophrenia patients) from the Stanley Medical Research Institute. As described in our previous report (Iwamoto et al., 2011), to reduce confounding factors due to previously identified effects of sample pH, we interrogated 72 individuals (26 controls, 23 bipolar disorder and 23 schizophrenia patients) which were preselected for high pH levels ($pH \geq 6.4$).

2.2. Gene expression

Gene expression levels were assessed using the Affymetrix HU133A microarray which comprised of 22,283 expression probes, details of which are described elsewhere (Iwamoto et al., 2011). Briefly, the raw gene expression data was preprocessed using MAS5 (*Affymetrix*) and filtered for probes which were called present in more than 50% of the samples, allowing a total of 11,920 probes for subsequent analysis. Microarray data had been deposited to the GEO database and is available on the GEO server (GES12649) and on the Stanley Medical Research Institute database (https://www.stanleygenomics.org/).

2.3. Copy number variation

Copy number variation was measured on the Agilent 450k early access CGH array (Agilent Technologies, Inc., Santa Clara, CA, USA), which is designed based on the database of known CNVs. Sample and reference DNA (3.0 μg each) was labeled with Cy5 or Cy3 using the DNA labeling kit from Agilent. Hybridization and washes were performed following the manufacturer's recommendation. The arrays were scanned with a MicroArray Scanner G2505A (Agilent). The obtained TIFF image data were processed with Agilent Feature Extraction software (version 9.5.3.1) using the CGH-v4_95_Feb07 protocol (Agilent).

DNA from one female (NA15510, Coriell Cell Repository, Camden, NJ, USA) was used as a reference to allow detection of copy number changes. This was in accordance with previous reports which have shown that usage of a single reference increases the sensitivity to detect more CNVs and produces more consistent and reproducible data as compared to using a pooled reference (Haraksingh et al., 2011). The raw data were imported into Agilent DNA Analytics 70 software and analyzed using the Aberration Detection Method 2 (ADM-2) algorithm (Lipson et al., 2006) which uses log2 ratios weighted by log2 ratios error as calculated by Feature Extraction software to identify genomic intervals with copy number differences between the samples and the reference. The Agilent Feature Extraction software was used to compute Quality Control metrics. The Agilent protocol recommended thresholds including average signal intensity at each probe, background signal (noise) (<5) using non-hybridizing control probes and signal-to-noise ratios (>30) were used to assess the quality of DNA and the experimental workflow. The derivative log ratio spread (dLRsd) was used to calculate the robust standard deviation (spread) of the log ratio differences between consecutive probes across all chromosomes. Three samples that did not satisfy the QC metrics thresholds and had dLRsd of >0.30 were excluded from further analysis.

The following parameters were used in this analysis: threshold of ADM-2: 6.0 with a bin size of 10; fuzzy zero: on; GC correction: on, aberration filters: on (maxAberrations = 100,000 AND percentPenetrance = 0); feature level filters: on (gIsSaturated = true OR rIsSaturated = true OR gIsFeatNonUnifOL = true OR rIsFeatNonUnifOL = true). A minimum three contiguous suprathreshold probes were mandatory to define a copy number change. Data were centralized and calls with average log2 ratios of ≤0.25 were excluded from the analysis. Data were normalized using the GC correction algorithm that corrects for wavy artifacts associated with the GC content of genomic regions and fuzzy zero correction that allows correction of extended aberrant segments with low absolute mean ratios that might represent noise. In the current study we assessed only autosomal CNVs since analysis of X and Y chromosomal CNVs are difficult to interpret. After filtering, a total of 34,453 CNV probes corresponding to 6836 copy number variable regions (CNVRs) were used for the analysis.

2.4. Statistical analysis

Physical positions and annotations of the gene expression and CNV array probes were updated to the Genome build GRCh37 (hg19) using the UCSC genome browser Liftover tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Separate analysis was performed for the continuous log2 CNV ratio and the simplified CNV state (1 = loss, 2 = normal state and 3 = gain) for each CNVR. A log2 ratio of 0 was considered the normal state, a log2 ratio of <−0.25 was considered a loss and log2 ratio of >0.25 was considered a gain. To identify the influence of CNVRs on gene expression, for each CNVR we probed a cis-window of ±1 Mb from the CNVR coordinates. For all gene expression probes located fully or partially (at least 1 bp overlap) within this window, we calculated the association between the CNVR state/CNVR log-ratio and the gene expression levels using general linear models (glm) in R, whilst covarying for age, gender, ethnicity and post-mortem interval (PMI) hours and the results were corrected for multiple testing using 5% false discovery rate (fdr). Results of the association between the CNVR state/CNVR log-ratio and the gene expression levels were very similar, therefore only the results for the CNVR state are presented.

To correct for the different numbers of gene expression probes tested for each CNVR and to account for possible inflation in association *p*-values which might result due to outliers especially for the singletons, we repeated the association analysis using label-swapping adaptive permutations with a maximum of 100,000 permutations in PLINK (Purcell et al., 2007). This method is widely accepted as the most appropriate method for multiple testing