Research paper

Investigation of copy number variation in subjects with major depression based on whole-genome sequencing data

Chenglong Yu[a,b,*], Bernhard T. Baune[c], Ma-Li Wong[a,b], Julio Licinio[a,b,d,*]

[a] Mind and Brain Theme, South Australian Health and Medical Research Institute, North Terrace, Adelaide, SA 5000, Australia
[b] School of Medicine, Flinders University, Bedford Park, SA 5042, Australia
[c] Discipline of Psychiatry, School of Medicine, University of Adelaide, Adelaide, SA 5005, Australia
[d] South Ural State University Biomedical School, Chelyabinsk, Russian Federation

ARTICLE INFO

ABSTRACT

Background: Despite recent intensive research using genome-wide association studies, the underlying biological basis of major depressive disorder (MDD) still remains unknown. In contrast to genotyping platforms which identify specific variations, whole-genome sequencing (WGS) allows us to detect all private genetic variations within an individual. So far there have been no studies investigating copy number variations (CNVs) in subjects with MDD using WGS data.

Methods: We obtained complete WGS paired-end reads data of 15 MDD patients and 10 ethnically matched healthy controls. We performed alignments for the sequencing reads and used GASV package to call CNVs including deletion, inversion, translocation and divergence for those subjects.

Results: Our results show that, in the Mexican-American sample, deletion CNVs were significantly richer in MDD cases than healthy controls on each of 23 chromosomes. However, other types of CNVs failed to reach any significance. In the Australian sample, there was no statistically significant difference of CNVs between MDD cases and controls. Furthermore, we found that the Australian group had significantly more deletion CNVs than the Mexican-American group.

Limitations: High quality WGS costs limited obtaining larger datasets. The GASV package does not currently support duplication or insertion CNVs.

Conclusions: To our knowledge this is the first time that CNVs detected by WGS data are used to study major depression. The conclusion that deletion CNVs are significantly richer in MDD cases than healthy controls is consistent with the previous finding about recurrent depressive disorder by genome-wide association analysis of CNVs on a large genotyping microarray data.

1. Introduction

Major depressive disorder (MDD) is a common but serious psychiatric and behavioural condition associated with high medical, social, and economic burdens. MDD produces significant morbidity and mortality, and leads to high rates of suicide (Kessler et al., 2005; Lopez and Murray, 1998; Wong and Licinio, 2001, 2004). Genetic factors play significant roles in the development of MDD. Despite recent progress, little is known about the underlying fundamental biology of MDD and much work still needs to be done to fully elucidate the genetic factors that confer susceptibility to this condition (Amin et al., 2016; CONVERGE Consortium, 2015; Hyde et al., 2016; Sullivan et al., 2012; Wong et al., 2016; Yu et al., 2017d).

Recent advances in high-throughput sequencing technologies provide unprecedented opportunities for biological and medical research (Soon et al., 2013; Yu et al., 2010, 2014). Specifically, the growth of newer and cheaper whole-genome sequencing (WGS) technology has allowed the development of new methods towards personalized treatment (Collins and McKusick, 2001; Hamburg and Collins, 2010; Yu et al., 2017b, 2017c). WGS can identify single nucleotide variants (SNVs), small insertions and deletions (INDELs), and copy number variations (CNVs), which are private genetic variations, and determine all the genetic variations within each person. As WGS costs are dropping further, scientists may have the opportunity to investigate the significance of these variations, which involve more individual characteristics (Belkadi et al., 2015).

CNVs refer to a type of structural variation with abnormal copy number changes involving long DNA fragments and resulting in gains (duplication or insertion), losses (deletion), or other rearrangements (inversion or translocation) of the genome (Pirooznia et al., 2015). The size of CNVs is typically defined as larger than 50 bp, whereas smaller elements are known as INDELs (Zarrei et al., 2015). There is increasingly robust recent evidence that CNVs play important roles in the genetics of MDD (Glessner et al., 2010; O'Dushlaine et al., 2014; Perlis et al., 2012; Rucker et al., 2013). All these studies used genomic microarrays as their diagnostic tool for detecting CNVs; however, to our knowledge, there have been no studies that have investigated CNVs in subjects with MDD using WGS data thus far. The aim of the current study was to identify CNVs in MDD participants and determine if these differ from ethnically matched healthy controls using WGS data.

2. Materials and methods

2.1. The Mexican-American sample

We have recently studied whole-exome genotyping of a Los Angeles Mexican-American cohort of 203 MDD patients and 196 healthy controls (Wong et al., 2016; Yu et al., 2017a). Participants in this cohort had three or more grandparents born in Mexico. They provided written informed consent, and detailed demographic, epidemiological and clinical descriptions have been previously described (Dong et al., 2009; Wong et al., 2012, 2014). Participants with MDD met DSM-IV (Diagnostic and Statistical Manual IV edition) criteria for current, unipolar major depressive episode using the Structured Clinical Interview for Diagnostic (SCID) and had a HAM-D21 (21-Item Hamilton Depression Rating Scale) score of 18 or greater with item number 1 (depressed mood) rated 2 or greater. Controls were in general good health but were not screened for medical or psychiatric illnesses; they were age- and gender-matched Mexican-American individuals recruited from the same community in Los Angeles. The study was registered in ClinicalTrials.gov (NCT00265291), and approved by the institutional review boards of the University of California Los Angeles and University of Miami, USA, and by the human research ethics committees of the Australian National University and Bellbery Ltd, Australia.

In the current study we acquired complete WGS data for a group of 15 participants selected from the cohort, 10 MDD patients and 5 controls. WGS was performed using Illumina HiSeq. 2000 (BGI-Shenzhen, Shenzhen, Guangdong, China). We have confirmed that in the cohort there was no family or population structure among all those participants (Wong et al., 2016). We confirmed that there were no blood relatives between the 15 Mexican-American subjects. All those subjects were female, and the MDD case group had an average age of 38.8 years with standard deviation 8.15 and the control group had an average age of 39.6 years with standard deviation 7.36. Thus case and control groups had basically the same age distribution (see Tables S1 and S2).

2.2. The Australian of European-Ancestry sample

We also included WGS data from a group of 10 Australian participants of European-ancestry. Participants included 5 MDD cases and 5 healthy controls who gave written informed consent and were recruited under the cognitive function and mood study (CoFaM-study) protocol conducted by the Discipline of Psychiatry, University of Adelaide, South Australia, Australia (Baune and Air, 2016). The SCID was used to ascertain that healthy controls were free from lifetime history of psychiatric disorders, and the main diagnostic and mood assessment instruments used were DSM-IV criteria for MDD and HAM-D21, respectively. The study was approved by human research ethics committees at the University of Adelaide and Flinders University, South Australia, Australia. WGS was performed using HiSeq X (Garvan Institute, Sydney, New South Wales, Australia).

In Tables S1 and S2, we presented descriptive statistics of gender, age and HAM-D scores for the 15 Mexican-American subjects and the 10 Australian subjects. The Australian sample of European-ancestry included both acutely depressed and remitted depressed patients who had by definition a lower depression severity compared to acutely depressed patients. Thus the HAM-D scores for the Australian sample may be lower on average than the Mexican-American sample.

2.3. WGS analysis and CNV calling

We aligned the high quality paired-end sequencing reads of each participant to the human reference genome (hg19, Genome Reference Consortium GRCh37) using Burrows-Wheeler Aligner (Li and Durbin, 2009) to get SAM (sequence alignment/map) format files. SAM files were converted, sorted and merged into one BAM (binary version of a SAM file) format file using SAMtools (Li et al., 2009). In our previous work (Wong et al., 2016), we used the mpileup command in SAMtools/BCFtools (Li, 2011) to produce VCF (variant call format) files with SNV and INDEL information. The results of SNVs and INDELs in those 25 subjects have been reported (Wong et al., 2016). In the current research, we used alignmented BAM files to call CNVs employing the GASV package (Sindi et al., 2009), which implements a geometric method for classification and comparison of structural variants. GASV was designed for next generation paired-end sequencing data and can identify CNVs including deletion, inversion, translocation and divergence (divergent variation indicates an inter-chromosomal rearrangement that is not a deletion, inversion, or insertion) (Sindi et al., 2012).

WGS analysis and CNV calling were performed using eResearch South Australia's high-performance computers (www.ersa.edu.au).

2.4. Statistical analysis

Difference between two group means was tested using independent two-sample t-test. For multiple testing, P-values were corrected using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995) and significance was set at ≤ 0.05. All statistical calculations and graphs were performed using the R software (www.r-project.org).

3. Results

We investigated CNVs of deletion, inversion, translocation and divergence on 25 subjects (10 Mexican-American MDD cases; 5 Mexican-American controls; 5 Australian MDD cases; 5 Australian controls) based on complete WGS data. Fig. 1 illustrates the CNV comparisons between MDD case and healthy control groups on two populations (Mexican-American and Australian). In the Mexican-American sample, significant difference was found between MDD cases and controls for deletion CNV (p-value $= 0.009$, t-statistic $= -3.066$); the MDD group had significantly more deletion CNVs than the control group. However, there was no statistically significant difference between MDD cases and controls for inversion, translocation and divergence on both populations.

Interestingly, the Australian group had significantly more deletion CNVs than the Mexican-American group (see Fig. 1). This is consistent with our previous finding showing that Mexican-American individuals have significantly more SNVs when compared to Australian individuals of European-Ancestry (Wong et al., 2016), as more deletions directly lead to more loss of nucleotides in DNA sequences which may contain SNVs. Furthermore, our Mexican-American cohort and the International haplotype map project (HapMap) cohort were recruited from the same community in Los Angeles; thus, they have 45% Indigenous American, 49% European, and 5% African ancestries (Johnson et al., 2011). Results from the International HapMap 3 Consortium and the 1000 Genomes Project Consortium showed that individuals with African ancestry have increased number of variants, and Spanish populations have excess of rare variants (Genomes Project