Published in final edited form as:

Psychiatry Res. 2017 June ; 252: 75–79. doi:10.1016/j.psychres.2017.02.041.

Whole-genome single nucleotide variant distribution on genomic regions and its relationship to major depression

Chenglong Yu^{a,b,*}, Bernhard T. Baune^c, Julio Licinio^{a,b}, and Ma-Li Wong^{a,b}

^aMind and Brain Theme, South Australian Health and Medical Research Institute, North Terrace, Adelaide, SA 5000, Australia

^bSchool of Medicine, Flinders University, Bedford Park, SA 5042, Australia

^cDiscipline of Psychiatry, School of Medicine, University of Adelaide, Adelaide, SA 5005, Australia

Abstract

Recent advances in DNA technologies have provided unprecedented opportunities for biological and medical research. In contrast to current popular genotyping platforms which identify specific variations, whole-genome sequencing (WGS) allows for the detection of all private mutations within an individual. Major depressive disorder (MDD) is a chronic condition with enormous medical, social and economic impacts. Genetic analysis, by identifying risk variants and thereby increasing our understanding of how MDD arises, could lead to improved prevention and the development of new and more effective treatments. Here we investigated the distributions of whole-genome single nucleotide variants (SNVs) on 12 different genomic regions for 25 human subjects using the symmetrised Kullback-Leibler divergence to measure the similarity between their SNV distributions. We performed cluster analysis for MDD patients and ethnically matched healthy controls. The results showed that Mexican-American controls grouped closer; in contrast depressed Mexican-American participants grouped away from their ethnically matched controls. This implies that whole-genome SNV distribution on the genomic regions may be related to major depression.

Keywords

major depressive disorder; whole-genome sequencing; Mexican-American; Kullback-Leibler divergence; cluster analysis

---

*Corresponding author at: Mind and Brain Theme, South Australian Health and Medical Research Institute, PO Box 11060, Adelaide, SA 5001, Australia. chenglong.yu@sahmri.com (C. Yu).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of interest

The authors declare no conflict of interest.

1. Introduction

Recent advances in high-throughput genomic technologies provide unprecedented opportunities for medical research. Patient care is moving towards genotyping/sequencing-based precision medicine. SNP (single nucleotide polymorphism) genotyping is currently the most popular technique for genome-wide association studies (GWAS) which identify significantly associated variations with a trait or disease. However, genome-wide association studies have attracted much criticism, mainly because of their assumption that common variations play a large role in elucidating the heritable variation of common disease (Visscher et al., 2012). The development of new and cheaper whole-genome sequencing (WGS) technology is allowing scientists to search for new channels to provide personalized treatment (Collins and McKusick, 2001; Hamburg and Collins, 2010). WGS can identify single nucleotide variants (SNVs), which are private genetic variants, and determine all the genetic variants within each person. As WGS costs are forecasted to drop further, researchers may have the opportunity to investigate the significance of SNVs, which involve more individual characteristics.

Major depressive disorder (MDD), as a serious chronic condition, is a main contributor to global disease burden and produces considerable morbidity and mortality (Kessler et al., 1994; Kessler et al., 2005; Lopez and Murray, 1998; Wong and Licinio, 2001; Wong and Licinio, 2004). Genetic factors play important roles in the susceptibility to major depression, as indicated by family, twin, and adoption studies (Sullivan et al., 2000; Lohoff, 2011). The heritability of MDD is estimated to range between 40%–70% (Lesch, 2004); a meta-analysis of twin studies has estimated a 37% heritability for MDD (Sullivan et al., 2000). Those data support a complex etiology that includes both genetic and environmental factors (Demirkan et al., 2011). Identifying specific genetic factors for the development of MDD remains a challenge. GWAS have identified genetic risk variants for many psychiatric illnesses; however, only recently replicated genome-wide significant loci have been identified for MDD (Amin et al., 2016; CONVERGE Consortium, 2015; Hyde et al., 2016; Peterson et al., 2017). Therefore, the underlying fundamental biology of MDD is largely unknown and much work still needs to be done to fully elucidate the genetic factors that confer susceptibility to this condition (Flint and Kendler, 2014; Sullivan et al., 2012; Yu et al., 2017).

We hypothesized that if two individuals have similar SNV distributions on the genomic regions, then those two individuals would have similar phenotypes, such as traits or diseases. In this study, we investigated the whole-genome SNV distributions on 12 different genomic regions for 25 human subjects including MDD cases and ethnically-matched controls. To accurately compare two SNV distributions, an effective similarity measure may be applied. In our previous work (Yu et al., 2011), we used the Kullback-Leibler divergence (also called relative entropy) to obtain a symmetrised measure and successfully studied the phylogeny of various DNA sequences based on their probability distributions. The Kullback-Leibler divergence, which is a powerful tool to measure the difference between two probability distributions, has been widely used in many fields, such as bioinformatics (Kaitchenko, 2004; Liu et al., 2014; McClendon et al., 2012; Hoang et al., 2015), genomics (Huang et al., 2014; Lin, 2015; Siegel et al., 2015), and machine learning (Lee and Lee, 2006; Rubinstein

and Kroese, 2013). By means of this measure, we performed cluster analysis for those subjects including MDD cases and controls. Our results suggest that whole-genome SNV distribution on the genomic regions may have some relationship with MDD.

2. Methods

2.1 The Mexican-American Sample

In our recent work (Wong et al., 2016) we have investigated whole-exome genotyping of a Los Angeles Mexican-American cohort of 203 MDD patients and 196 healthy controls, aged 19–65 years. Participants in this cohort had three or more grandparents born in Mexico. Participants with MDD were diagnosed using the Structured Clinical Interview for Diagnostic (SCID) and Statistical Manual of mental disorders, and the DSM-IV (Diagnostic and Statistical Manual IV edition) diagnostic criteria for current, unipolar major depressive episode with a HAM-D21 (21-Item Hamilton Depression Rating Scale) score of 18 or greater with item number 1 (depressed mood) rated 2 or greater; they participated in pharmacogenetic studies of antidepressant treatment. MDD was defined as 5 out of 9 criteria in the SCID. A Hamilton score >18 was used to characterize entry criteria for the pharmacogenetic study but not to diagnose MDD. Raters were experienced bilingual clinical personnel (nurses, social workers and physicians) using Spanish or English versions of questionnaires and rating scales, and diagnosis was confirmed by a research psychiatrist. 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. Participants provided written informed consent, and detailed demographic, epidemiological, and clinical descriptions were previously described (Dong et al., 2009; Wong et al., 2012; Wong et al., 2014). 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 Bellberry Ltd, Australia. In this work we obtained 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 individuals and no blood relationship among the 15 selected participants.

2.2 The European-Ancestry Sample

We also included WGS data from a group of 10 Australians of European-Ancestry as a comparison group. Those 10 participants, including 5 MDD cases and 5 healthy controls, gave written informed consent and were recruited under the protocol of Characteristics of the Cognitive Function and Mood Study (CoFaM-Study) conducted by the Discipline of Psychiatry, University of Adelaide, South Australia, Australia (Baune and Air, 2016). The MDD diagnoses were made using Mini International Neuropsychiatric Interview (MINI) for both the cases and the healthy controls. We also assessed HAM-D21 scores for those participants. This study was under 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).

Table 1 summarizes descriptive statistics of gender, age and HAM-D scores for the 15 Mexican-American subjects and the 10 Australian subjects. The Australians of European-Ancestry sample includes both acutely depressed and remitted depressed patients who have 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 Data Analysis

We performed the variant calling analysis of high quality paired-end sequencing reads using a previously described pipeline (Wong et al., 2016). Briefly, we aligned the 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 to the BAM (binary version of a SAM file) format files using SAMtools (Li et al., 2009). BAM files were then merged into one BAM file, and the mpileup command in SAMtools was applied to calculate the likelihood of data given each possible genotype, and store the likelihoods in a binary file. The output was supplied to SAMtools/BCFtools (Li, 2011) which created the SNV and INDEL (small insertions and deletions) calling to produce VCF (variant call format) files. Finally, the popular ANNOVAR software (Wang, et al., 2010) was applied to annotate SNV information through 12 different genomic regions. WGS data analysis were performed using highperformance computers in eResearch South Australia (www.ersa.edu.au).

2.4 SNV Count Distribution on Genomic Regions

After variant annotation, we can categorize the SNV count distribution on 12 genomic regions into: exonic synonyms, exonic nonsynonyms, intronic, intergenic, splicing, UTR3, UTR5, downstream, upstream, ncRNA exonic, ncRNA intronic, and ncRNA splicing. According to ANNOVAR results, all the SNVs are located on those 12 genomic regions. There is no other genomic region to be considered. Thus each subject corresponds to a SNV distribution on the 12 genomic regions. In order to obtain an accurate comparison between two SNV distributions, a proper similarity measure should be carefully employed to test our hypothesis in this study that if two individuals have similar SNV distributions, then those two individuals would have similar phenotypes, such as traits or diseases.

2.5 Symmetrised Kullback-Leibler Divergence

As discussed in the introduction, the Kullback-Leibler divergence (KLD) provides a similarity measure between two probability distributions (Kullback and Leibler, 1951). Let $\lambda_1 = (p_1, p_2, \dots, p_n)$, $p_i > 0$ for all n , and $\lambda_2 = (q_1, q_2, \dots, q_n)$, $q_i > 0$ for all n , be two discrete probability distributions on a finite n -dimensional set X . Then the KLD, denoted as $H(\lambda_1, \lambda_2)$ of λ_1 with respect to λ_2 is defined by

$$H(\lambda_1, \lambda_2) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i}. \quad (1)$$

$H(\lambda_1, \lambda_2) = 0$ if and only if $\lambda_1 = \lambda_2$. $H(\lambda_1, \lambda_2)$ is often called a distance, but it is not a true metric because $H(\lambda_1, \lambda_2) \neq H(\lambda_2, \lambda_1)$ in general. Furthermore, it does not satisfy the triangle inequality. We have defined a simple symmetrised measure of KLD (Yu et al., 2011), denoted by $d(\lambda_1, \lambda_2)$, as

$$d(\lambda_1, \lambda_2) = \frac{H(\lambda_1, \lambda_2) + H(\lambda_2, \lambda_1)}{2}. \quad (2)$$

Clearly, $d(\lambda_1, \lambda_2) = d(\lambda_2, \lambda_1)$. We will use this symmetrised KLD to compare the SNV distributions among those subjects. In this study, we consider the 12 genomic regions through the whole genome, so we take n as 12.

In this proof-of-principle work we use this symmetrised KLD to calculate the distance matrix of those 25 participants. Then cluster tree based the distance matrix can be constructed. Clustering relationships shown in the tree may reveal significant medical information that may be translated into clinical practice.

3. Results

3.1 SNV Distribution Analysis

In Figure 1 and Table S1, we provided the whole-genome SNV count distributions through the 12 genomic regions for four groups of 25 human subjects (10 Mexican-American MDD cases, 5 Mexican-American controls, 5 Australian MDD cases, and 5 Australian controls). Figure 1 shows that the Mexican-American MDD and control groups have close SNV distribution curves, but there also exist some discrepancies on some genomic regions, such as intronic and intergenic. Furthermore, the Mexican-American and Australian groups have very different SNV distributions on the genomic regions studied. Mexican-American individuals have significantly more SNVs when compared to Australian individuals of European-Ancestry. Our Mexican-American cohort and the Human Haplotype Matching Project (HapMap) cohort were recruited from the same community in Los Angeles; thus, they have 49% European, 45% Indigenous American, and 5% African ancestries (Johnson et al., 2011). Mexican-Americans have more polymorphic SNPs in Mexican-Americans than northern Europeans (International HapMap 3 Consortium, 2010), which is compatible with their African and Spanish ancestries. 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 (1000 Genomes Project Consortium, 2012).

For statistical analysis, the independent two-sample t-test is used to determine whether group means are significantly different from each other. As we studied 12 genomic regions, the significance level was set at < 0.0042 by Bonferroni correction. For Mexican-American, there is no significant difference between SNVs of MDD and control groups on genomic regions of exonic synonyms, exonic nonsynonyms, splicing, UTR3, UTR5, downstream, upstream, ncRNA exonic, ncRNA intronic, and ncRNA splicing. The intronic SNVs in MDD group (mean of 3027113.7 and standard deviation of 72837.85) is significantly

different from those in control group (mean of 2729536.2 and standard deviation of 143102.09) with twotailed p-value of 0.0003. The intergenic SNVs in MDD group (mean of 4123066.8 and standard deviation of 84412.48) is significantly different from those in control group (mean of 3801548.4 and standard deviation of 165288.21) with two-tailed p-value of 0.0005. However, for Australian of European-Ancestry, there is no significant difference between SNVs of MDD and control groups on all the 12 genomic regions.

Although Figure 1 shows some visualized distinguishing information among those subjects, for acquiring precise cluster analysis we used the symmetrised KLD to numerically compare those 25 SNV distributions.

3.2 Cluster Analysis of 25 Participants

Following the proposed method, we used the symmetrised KLD and SNV distributions to obtain the distance matrix among those 25 subjects. Here the symmetrised KLD calculation was done using the “seewave” package (Sueur et al., 2016) in R software (www.r-project.org). We then applied the popular neighbor-joining method (Saitou and Nei, 1987) on the distance matrix to reconstruct the cluster tree, which was drawn using software MEGA 6 (Tamura et al., 2013). Figure 2 shows the cluster tree for the 25 participants. We found that all the five Mexican-American controls grouped together in a cluster, and ten Mexican-American MDD patients were separated from that group. The Australian individuals of European-Ancestry, as a different population, assembled as an obvious outgroup from the Mexican-American subjects.

4. Discussion

To the best of our knowledge, this is the first time that whole-genome SNV distribution on the genomic regions are considered to study MDD. The SNV distribution, as a potential genetic marker, could be a hotspot in bioinformatics and medical research in the foreseeable future. The cluster analysis results obtained by our approach support our assumption that two individuals who have similar SNV distributions on genomic regions would have similar phenotypes, such as traits or diseases. Clustering relationships in the tree show that the Mexican-American controls group together, and ethnically matched MDD patients grouped away and separately. It implies that SNV distribution on the genomic regions may relate to this condition. The fact that Australian subjects fail to be clustered into case and control groups may suggest that our method may be restricted to specific populations, with a higher degree of genetic diversity, such as Mexican-Americans. Our results showed that the symmetrised KLD, as a numerical measure to compare the SNV distributions, is appropriate to cluster Mexican-American MDD subjects in this work. Further studies will be needed to determine whether this measure can replicate our findings in larger datasets. Costs and computational resources needed for WGS and data analysis continue to be high, limiting the availability of larger datasets. The aim of this paper is not to confirm or refute previous results, but rather to bring a completely new methodology to investigate the relationship between SNV and disease. For example, the distributions of intronic and intergenic SNVs in Mexican-American MDD and control groups are significantly different. It implies that

functional variants in these two genomic regions may be more numerous in MDD cases than controls.

In this study we focused only on single base variations on 12 genomic regions: exonic synonyms, exonic nonsynonyms, intronic, intergenic, splicing, UTR3, UTR5, downstream, upstream, ncRNA exonic, ncRNA intronic, and ncRNA splicing. INDELs were not included in the current analysis because their distributions on those regions are ambiguous. For example, since an INDEL involve in insertion or deletion of a fragmental sequence, it may occur at the boundary of regions and cross the two regions. In this scenario, we could not determine its position in one genomic region. Furthermore, exonic synonyms and nonsynonyms are always considered for point mutations, thus INDELs are not applicable for these regions.

Cluster analysis that we used in this paper could be used as a predictive/diagnostic model. For example, our approach can cluster the 15 Mexican-Americans into two groups in the cluster tree: MDD and control groups. For new participants, we could determine how close those persons are to the existing MDD or control groups using cluster analysis. Subjects within or close to the MDD group in the cluster tree could be predicted to be potential MDD cases. Furthermore, future studies using cluster analysis on a larger sample may cluster specific symptoms and clinical characteristics, and explore potential MDD subtyping biomarkers. Advances in this line of research could provide a more integrated mechanistic view of whether subjects with similar SNV distributions on the genomic regions may also be phenotypically similar.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors have been supported by grants NHMRC APP1051931 (MLW), NHMRC APP1060524 (BTB), and NIH grant GM61394 (JL and MLW), and institutional funds from the South Australian Health and Medical Research Institute.

References

- 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012; 491:56–65. [PubMed: 23128226]
- Amin N, Jovanova O, Adams HH, Dehghan A, Kavousi M, Vernooij MW, Peeters RP, de Vrij FM, van der Lee SJ, van Rooij JG, van Leeuwen EM, Chaker L, Demirkan A, Hofman A, Brouwer RW, Kraaij R, Willems van Dijk K, Hankemeier T, van Ijcken WF, Uitterlinden AG, Niessen WJ, Franco OH, Kushner SA, Ikram MA, Tiemeier H, van Duijn CM. Exome-sequencing in a large population-based study reveals a rare Asn396Ser variant in the LIPG gene associated with depressive symptoms. *Mol Psychiatry*. 2016; doi: 10.1038/mp.2016.101
- Baune BT, Air T. Clinical, Functional, and Biological Correlates of Cognitive Dimensions in Major Depressive Disorder-Rationale, Design, and Characteristics of the Cognitive Function and Mood Study (CoFaM-Study). *Front Psychiatry*. 2016; 7:150.doi: 10.3389/fpsy.2016.00150 [PubMed: 27616997]
- Collins FS, McKusick VA. Implications of the Human Genome Project for medical science. *JAMA*. 2001; 285:540–544. [PubMed: 11176855]

- CONVERGE consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature*. 2015; 523:588–591. [PubMed: 26176920]
- Demirkan A, Penninx BW, Hek K, Wray NR, Amin N, Aulchenko YS, Van Dyck R, de Geus EJ, Hofman A, Uitterlinden AG, Hottenga JJ. Genetic risk profiles for depression and anxiety in adult and elderly cohorts. *Mol Psychiatry*. 2011; 16(7):773–783. [PubMed: 20567237]
- Dong C, Wong ML, Licinio J. Sequence variations of ABCB1, SLC6A2, SLC6A3, SLC6A4, CREB1, CRHR1 and NTRK2: association with major depression and antidepressant response in Mexican-Americans. *Mol Psychiatry*. 2009; 14:1105–1118. [PubMed: 19844206]
- Flint J, Kendler KS. The genetics of major depression. *Neuron*. 2014; 81(3):484–503. [PubMed: 24507187]
- Hamburg MA, Collins FS. The path to personalized medicine. *N Engl J Med*. 2010; 363:301–304. [PubMed: 20551152]
- Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR, Tung JY, Hinds DA, Perlis RH, Winslow AR. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nature Genet*. 2016; 48:1031–1036. [PubMed: 27479909]
- Hoang T, Yin C, Zheng H, Yu C, He RL, Yau SST. A new method to cluster DNA sequences using Fourier power spectrum. *J Theor Biol*. 2015; 372:135–145. [PubMed: 25747773]
- Huang HH, Yu C, Zheng H, Hernandez T, Yau SC, He RL, Yang J, Yau SST. Global comparison of multiple-segmented viruses in 12-dimensional genome space. *Mol Phylogenet Evol*. 2014; 81:29–36. [PubMed: 25172357]
- International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010; 467:52–58. [PubMed: 20811451]
- Johnson NA, Coram MA, Shriver MD, Romieu I, Barsh GS, London SJ, Tang H. Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genet*. 2011; 7:e1002410. [PubMed: 22194699]
- Kaichenko, A. Algorithms for estimating information distance with application to bioinformatics and linguistics; Electrical and Computer Engineering, Canadian Conference on IEEE; 2004. p. 2255-2258.
- Kessler RC, Chiu WT, Demler O, Merikangas KR, Walters EE. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry*. 2005; 62:617–627. [PubMed: 15939839]
- Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, Wittchen HU, Kendler KS. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey Arch Gen Psychiatry. 1994; 51:8–19. [PubMed: 8279933]
- Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat*. 1951; 22:79–86.
- Lee C, Lee GG. Information gain and divergence-based feature selection for machine learning-based text categorization. *Inf Process Manag*. 2006; 42:155–165.
- Lesch KP. Gene-environment interaction and the genetics of depression. *J Psychiatry Neurosci*. 2004; 29(3):174–184. [PubMed: 15173894]
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25:1754–1760. [PubMed: 19451168]
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009; 25:2078–2079. [PubMed: 19505943]
- Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011; 27:2987–2993. [PubMed: 21903627]
- Liu R, Yu X, Liu X, Xu D, Aihara K, Chen L. Identifying critical transitions of complex diseases based on a single sample. *Bioinformatics*. 2014; 30:1579–1586. [PubMed: 24519381]
- Lin S. Kullback-Leibler divergence for detection of rare haplotype common disease association. *Eur J Hum Genet*. 2015; 23:1558–1565. [PubMed: 25735482]
- Lohoff FW. Overview of the genetics of major depressive disorder. *Curr Psychiatry Rep*. 2010; 12(6): 539–546. [PubMed: 20848240]

- Lopez AD, Murray CC. The global burden of disease, 1990–2020. *Nature Med.* 1998; 4:1241–1243. [PubMed: 9809543]
- McClendon CL, Hua L, Barreiro G, Jacobson MP. Comparing conformational ensembles using the Kullback-Leibler divergence expansion. *J Chem Theory Comput.* 2012; 8:2115–2126. [PubMed: 23316121]
- Peterson RE, Cai N, Bigdeli TB, Li Y, Reimers M, Nikulova A, Webb BT, Bacanu SA, Riley BP, Flint J, Kendler KS. The genetic architecture of major depressive disorder in Han Chinese women. *JAMA Psychiatry.* 2017; 74(2):162–168. [PubMed: 28002544]
- Rubinstein RY, Kroese DP. The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning. Springer Science & Business Media. 2013
- Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 1987; 4:406–425. [PubMed: 3447015]
- Siegel K, Altenburger K, Hon YS, Lin J, Yu C. PuzzleCluster: A novel unsupervised clustering algorithm for binning DNA fragments in metagenomics. *Curr Bioinform.* 2015; 10:231–252.
- Sueur, J., Aubin, T., Simonis, C., Lellouch, L., Brown, EC., Depraetere, M., Desjonqueres, C., Francois, F., Gasc, A., LaZerte, S., Lees, J., Marchal, J., Pavoine, S., Stotz, A., Villanueva-Rivera, LJ., Ross, Z., Witthoft, CG., Zhivomirov, H. Package ‘seewave’. 2016. <https://cran.r-project.org/web/packages/seewave>
- Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Rev Genet.* 2012; 13:537–551. [PubMed: 22777127]
- Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: review and meta-analysis. *Am J Psychiatry.* 2000; 157(10):1552–1562. [PubMed: 11007705]
- Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol.* 2013; 30:2725–2729. [PubMed: 24132122]
- Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet.* 2012; 90:7–24. [PubMed: 22243964]
- Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010; 38:e164. [PubMed: 20601685]
- Wong ML, Arcos-Burgos M, Liu S, Velez JI, Yu C, Baune BT, Jawahar MC, Arolt V, Dannlowski U, Chuah A, Huttley GA, Forgy R, Lewis M, Bornstein SR, Licinio J. The PHF21B gene is associated with major depression, and modulates stress response. *Mol Psychiatry.* 2016; doi: 10.1038/mp.2016.174
- Wong ML, Dong C, Andreev V, Arcos-Burgos M, Licinio J. Prediction of susceptibility to major depression by a model of interactions of multiple functional genetic variants and environmental factors. *Mol Psychiatry.* 2012; 17:624–633. [PubMed: 22449891]
- Wong ML, Dong C, Flores DL, Ehrhart-Bornstein M, Bornstein S, Arcos-Burgos M, Licinio J. Clinical outcomes and genome-wide association for a brain methylation site in an antidepressant pharmacogenetics study in Mexican Americans. *Am J Psychiatry.* 2014; 171:1297–1309. [PubMed: 25220861]
- Wong ML, Licinio J. Research and treatment approaches to depression. *Nature Rev Neurosci.* 2001; 2:343–351. [PubMed: 11331918]
- Wong ML, Licinio J. From monoamines to genomic targets: a paradigm shift for drug discovery in depression. *Nat Rev Drug Discov.* 2004; 3:136–151. [PubMed: 15040578]
- Yu C, Baune BT, Licinio J, Wong ML. Single nucleotide variant proportion in genes: a new concept to explore major depression based on DNA sequencing data. *J Hum Genet.* 2017; doi: 10.1038/jhg.2017.2
- Yu C, Deng M, Yau SST. DNA sequence comparison by a novel probabilistic method, Inform. Sciences. 2011; 181:1484–1492.

Highlights

- The whole-genome SNV distributions on different genomic regions are investigated.
- The symmetrised Kullback-Leibler divergence is used to measure SNV distributions.
- The cluster analysis result shows that the healthy controls grouped closer.
- The whole-genome SNV distribution may be related to major depression.

Fig 1. SNV distribution curves of four groups in 12 genomic regions. MA, Mexican American; AU, Australian. Error bars represent standard deviation.

Fig 2.
Cluster tree for 25 human subjects based on their SNV distributions.

The descriptive statistics for MDD cases and controls. μ : average value; SD: standard deviation; HAM-D: Hamilton depression rating scale; F: female; M: male.

Table 1

Group	Status	Number	Sex	Age μ	Age SD	HAM-D μ	HAM-D SD
Mexican-American	MDD	10	10 F	38.8	8.15	23	4.08
	Control	5	5 F	39.6	7.36	N/A	N/A
Australasian of European-Ancestry	MDD	5	1 M, 4 F	25	9.82	8.6	3.26
	Control	5	3 M, 2 F	21.2	4.49	2	1.26