Yu et al. *Translational Psychiatry* (2018)8:70
DOI 10.1038/s41398-018-0117-7

Translational Psychiatry

ARTICLE Open Access

Low-frequency and rare variants may contribute to elucidate the genetics of major depressive disorder

Chenglong Yu[1,2,3], Mauricio Arcos-Burgos[4], Bernhard T. Baune[5], Volker Arolt[6], Udo Dannlowski[6,7], Ma-Li Wong[2,3,8] and Julio Licinio[8]

Abstract

Major depressive disorder (MDD) is a common but serious psychiatric disorder with significant levels of morbidity and mortality. Recent genome-wide association studies (GWAS) on common variants increase our understanding of MDD; however, the underlying genetic basis remains largely unknown. Many studies have been proposed to explore the genetics of complex diseases from a viewpoint of the "missing heritability" by considering low-frequency and rare variants, copy-number variations, and other types of genetic variants. Here we developed a novel computational and statistical strategy to investigate the "missing heritability" of MDD. We applied Hamming distance on common, low-frequency, and rare single-nucleotide polymorphism (SNP) sets to measure genetic distance between two individuals, and then built the multi-dimensional scaling (MDS) pictures. Whole-exome genotyping data from a Los Angeles Mexican-American cohort (203 MDD and 196 controls) and a European-ancestry cohort (473 MDD and 497 controls) were examined using our proposed methodology. MDS plots showed very significant separations between MDD cases and healthy controls for low-frequency SNP set (P value < 2.2e−16) and rare SNP set (P value = 7.681e−12). Our results suggested that low-frequency and rare variants may play more significant roles in the genetics of MDD.

Introduction

Major depressive disorder (MDD) is a common mental illness with tremendous medical, economic, and social impact. MDD, as a principal contributor to disease load worldwide, leads to high levels of morbidity and mortality[1–5]. One significant avenue for preventing and treating depression lies in uncovering the genetics of this condition[6,7]. Despite rapid advances on genome-wide association studies (GWAS)[8–10], little is understood about its fundamental biological basis and much further research needs to be carried out to fully unravel the genetic elements that confer susceptibility to this disorder[11,12].

Many studies have been proposed to explore the genetic causes of complex diseases from a point view of the "missing heritability"[13–16]. For example, some genetic effects are not owing to the common single-nucleotide polymorphisms (SNPs) examined in the candidate-gene studies or GWAS, but due to low-frequency and rare variants, copy-number variations, and other types of genetic mutations[17]. Actually, GWAS focus on the identification of significant common (minor allele frequency (MAF) ≥ 5%) variants, thus analyses of low-frequency (0.5% ≤ MAF < 5%) and rare (MAF < 0.5%) variants would be promising to elucidate additional disease risk or trait variability[18]. Furthermore, using next-generation sequencing, family-based linkage analysis has also provided an important way to understand the role of rare variants in disease etiology[19,20]. For example, some family-

Correspondence: Chenglong Yu (chenglong.yu@flinders.edu.au) or Julio Licinio (licinioj@upstate.edu)
[1]Centre for Population Health Research, School of Health Sciences and Sansom Institute of Health Research, University of South Australia, Adelaide, SA, Australia
[2]Mind and Brain Theme, South Australian Health and Medical Research Institute, Adelaide, SA, Australia
Full list of author information is available at the end of the article

control studies applied Hamming distance to identify disease genes based on sequencing data[21,22]. However, high-priced sequencing expenses are currently a concern that restricts acquiring large datasets.

Recently, we have applied GWAS and rare-variant analysis to investigate the genetics of MDD based on a whole-exome genotyping data from a Mexican-American cohort in Los Angeles and a replication European-ancestry cohort[23]. Our results suggested that the "missing heritability" in MDD may be partly explained by rare variants, because most of the functional variations detected in the cohorts were rare. In this study, we designed a novel computational and statistical strategy to further investigate this conclusion. In our methodology, we used Hamming distance on common, low-frequency, and rare SNP sets to measure the genetic distance between two individuals. Then we built the multi-dimensional scaling pictures, in which separation between MDD cases and healthy controls revealed valuable information for hidden genetic factors of major depression. The corresponding statistical results in the pictures were reported.

Materials and methods

The two cohorts used in this study

In our previous work[23], we have investigated a cohort of MDD cases ($n = 203$) and controls ($n = 196$) of Los Angeles Mexican-Americans. They were mostly recent immigrants born in Mexico and experienced high levels of hyperactivation of the hypothalamic-pituitary-adrenal axis related to distress, challenges, and acculturation issues caused by immigration. MDD were diagnosed using the Structured Clinical Interview for DSM-IV (Diagnostic and Statistical Manual IV edition) (SCID, for abbreviation). Subjects met the diagnostic criteria for current, unipolar major depressive episode, attended a pharmacogenetic study on antidepressant treatment, and had an initial 21-Item Hamilton Depression Rating Scale (HAM-D21 for abbreviation) score of ≥18 with item number 1 (depressed mood) rated ≥2. Controls responded that they were in good health and replied questionnaires about acculturation. But they were not screened for medical illnesses and did not respond to structured psychiatric interviews. The controls were also Mexican-American and recruited from the same community in Los Angeles. The control group have similar sex ratio and age distribution (mean and standard error) to the MDD group (see Table S1). Participants submitted written informed consent, and their demographic, epidemiological, and clinical descriptions were previously described in detail[24–26]. We have registered this study in ClinicalTrials.gov (NCT00265291). The research was 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.

We also included the European-ancestry cohort of MDD cases ($n = 473$) and controls ($n = 497$), which was used for replication in our previous study[23]. In this cohort, the MDD group also have similar sex ratio and age distribution (mean and standard error) to the control group (see Table S1). Those participants provided written informed consent and were recruited under two protocols: (1) Münster mood disorder studies (consisted of the neuroimaging and the mood-in-flame studies), which have been conducted by the Department of Psychiatry and Psychotherapy, University of Münster, Münster, Germany, and (2) the Characteristics of the Cognitive Function and Mood Study (CoFaM-Study) conducted by the Discipline of Psychiatry, University of Adelaide, South Australia, Australia[27]. The SCID/MINI (Mini International Neuropsychiatric Interview) was used to ascertain that healthy controls were free from lifetime history of psychiatric disorders; for this cohort, we also used DSM-IV criteria and HAM-D21 for the main diagnostic of MDD and mood assessment. The study on this cohort was approved by Human Research Ethics Committee protocols at the University of Münster, Germany, and University of Adelaide and Flinders University, South Australia, Australia.

Our previous power analysis on the same cohorts[23] suggested that, when 100,000 variants are tested for association studies, 200 cases and 200 controls are sufficient to detect 80% true positives and a medium size of effect defined by the Cohen's h parameter. This value of 100,000 overcomes the numbers of SNPs in common, low-frequency and rare-variant groups studied here. Actually, based on the effect sizes for the 19 MDD GWAS significant variants in the Mexican-American cohort[23], the post hoc statistical power ranges between >60% (SNP-exm2249659, Cohen's $h = 0.335$) and >99% (SNP-exm1508600, Cohen's $h = 0.643$). Cohen's h suggests 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes, respectively; thus, it indicates that our study had enough power to detect medium to large effect sizes for current association tests.

Whole-exome SNP genotyping

The two cohorts were genotyped by the Australian Genome Research Facility (North Melbourne, VIC, Australia; www.agrf.org.au) using the Illumina HumanExome BeadChip-12v1_A, in which exonic content consists of >250,000 markers representing diverse populations and a range of common conditions. All the human samples passed the Illumina expected SNP calling rate (>99%). Then we filtered the raw whole-exome SNPs by a pipeline considering variant call rate, allele numbers, and Hardy–Weinberg equilibrium deviations. For this follow-

up study, we analyzed 83,898 SNPs for the Mexican-American cohort and 121,174 SNPs for the European-ancestry cohort, which remained after quality control (QC) and filtering out criteria. Detailed QC and filtering analyses have been well reported in our previous work[23].

SNP classification and population stratification

Considering MAF, we divided the 83,898 SNPs in the Mexican-American population into 27,575 common variants, 17,838 low-frequency variants, and 38,485 rare variants, and divided the 121,174 SNPs in the European-ancestry population into 12,530 common variants, 12,902 low-frequency variants, and 95,742 rare variants. As expected, we found that most SNPs are rare (MAF < 0.5%) in the HumanExome BeadChip, because this chip has been designed to concentrate on rare variants rather than common ones[28], which contrast to conventional genome-wide genotyping arrays that do not tag low-frequency and rare variants[29].

We then used the four classes of SNPs (all, common, low-frequency and rare) to check population stratifications of the two cohorts. PLINK software[30], which provides a powerful tool for population stratification based on pairwise identity-by-state (IBS) distance and multi-dimensional scaling (MDS) plots, was used here.

Table 1 Hamming distances of three subjects in an 8-SNP set

Genotype	SNP1 (A/T)	SNP2 (G/T)	SNP3 (C/G)	SNP4 (C/T)	SNP5 (C/T)	SNP6 (A/G)	SNP7 (A/C)	SNP8 (C/T)
Subject *X*	AT	GG	CG	CC	CC	AG	AC	TT
Subject *Y*	AA	GG	CC	TT	CT	GG	CC	CT
Subject *Z*	AA	GG	CC	CC	CT	AA	CC	TT

SNP single-nucleotide polymorphism

Hamming distance between two individuals

In this study, we use Hamming distance[31], a natural distance without assuming any model mutation/substitution rate, to investigate the genetic distance between two individuals based on a set of SNPs.

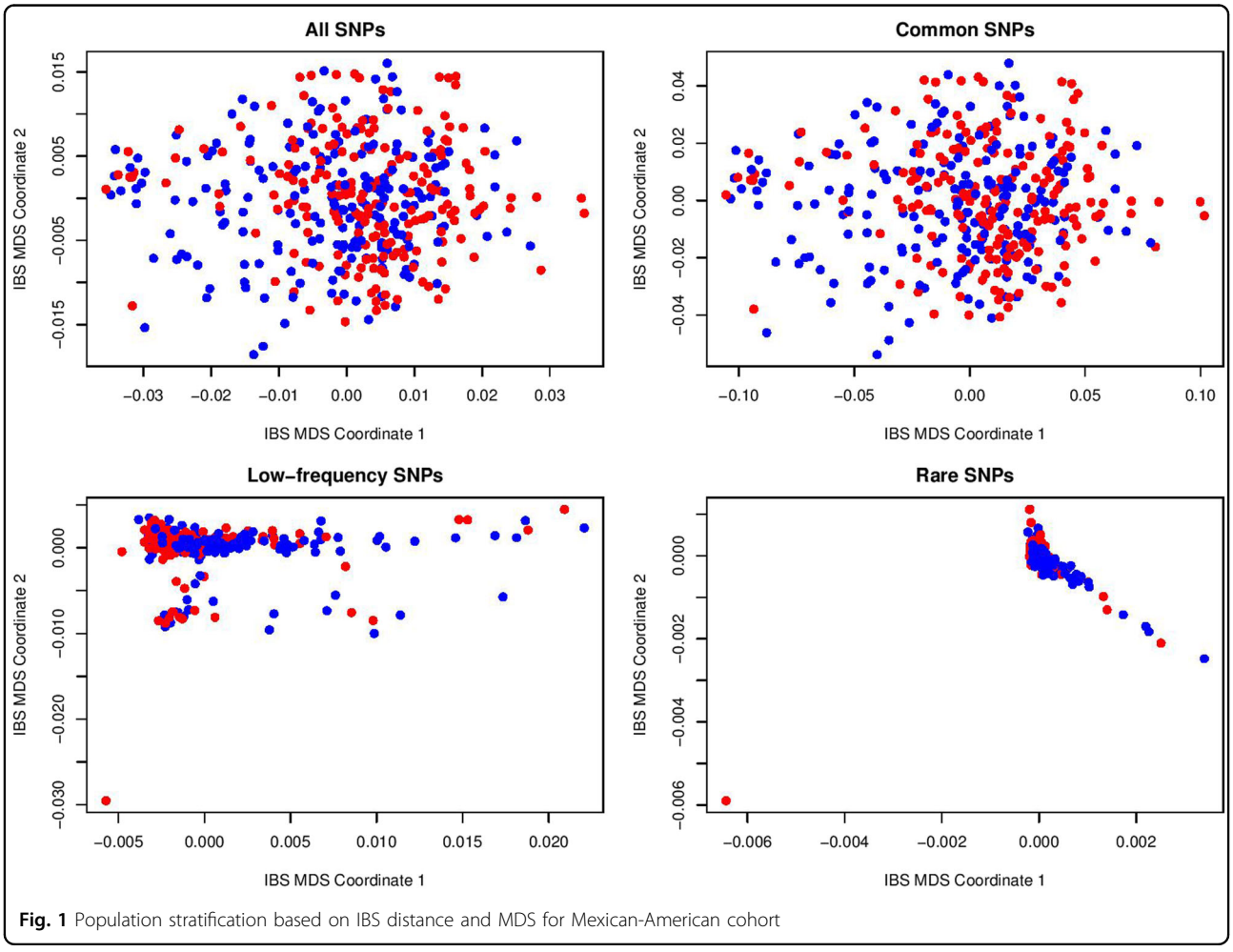

Fig. 1 Population stratification based on IBS distance and MDS for Mexican-American cohort

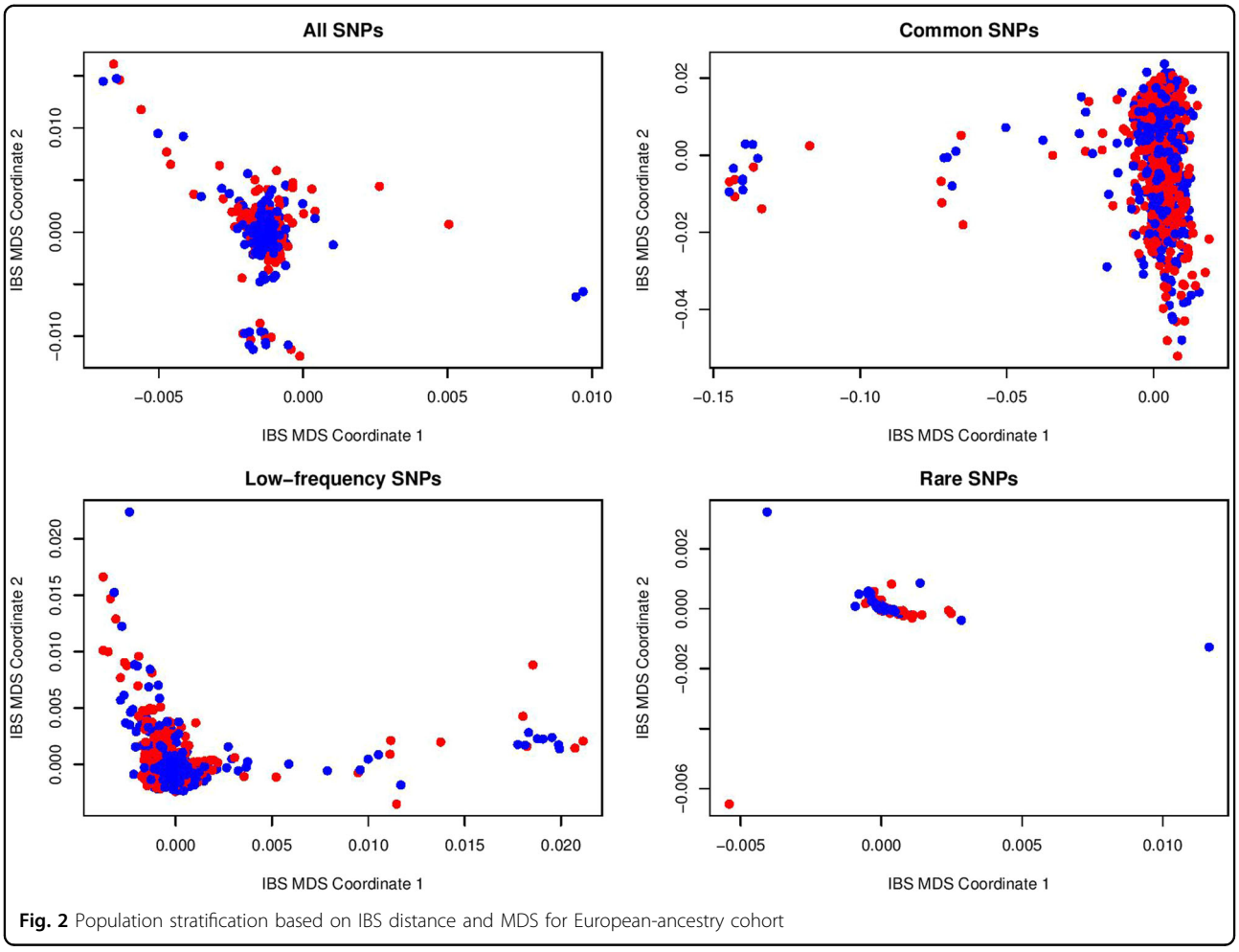

Fig. 2 Population stratification based on IBS distance and MDS for European-ancestry cohort

Let S be an SNP set which contains n SNPs. We use SNP_k to represent the SNP indexed k ($k = 1, \ldots, n$). Thus, $S = \{SNP_1, SNP_2, \ldots, SNP_n\}$. Suppose that X and Y are two individuals who have their own genotypes on this SNP set S, namely and respectively, S_X and S_Y. Let S_X be $\{SNP_1^X, SNP_2^X, \ldots, SNP_n^X\}$ and S_Y be $\{SNP_1^Y, SNP_2^Y, \ldots, SNP_n^Y\}$. Then the Hamming distance between the two individuals X and Y is defined as

$$H(X,Y) = \sum_{i=1}^{n} \delta(SNP_i^X, SNP_i^Y), \quad (1)$$

where $\delta(a,b) = \begin{cases} 0 & \text{if } a \text{ and } b \text{ are the same} \\ 1 & \text{otherwise} \end{cases}$, that is, the number of positions at which the corresponding SNPs are different on the SNP set S. Considering the size of the SNP set, we can also get the normalized Hamming distance as

$$NH(X,Y) = \frac{\sum_{i=1}^{n} \delta(SNP_i^X, SNP_i^Y)}{n}. \quad (2)$$

Take Table 1 as a simple example, individuals X, Y, and Z show their genotypes on an SNP set of eight SNPs. The Hamming distance between X and Y is 7 (SNP1, SNP3, SNP4, SNP5, SNP6, SNP7, and SNP8 are different). Similarly, the Hamming distance between Y and Z is 3, and the Hamming distance between X and Z is 5. Our hypothesis was that if two individuals have closer Hamming distance in this way, then those two individuals would have closer genetic distance and more similar phenotypes such as diseases or traits. We assume that Y and Z have more similar phenotypes in the above example.

Given a group of individuals, we can compute their Hamming distance matrix based on a specific SNP set such as common, low-frequency, or rare-variant set. After obtaining the distance matrix between all pairs of individuals, MDS approach[32] can be used to observe the distance relationships among those individuals in a two-dimensional graph. The display of scatters representing individuals which shows separating variability between MDD cases and healthy controls can reveal interesting genetic information hidden in the SNP sets. Then for

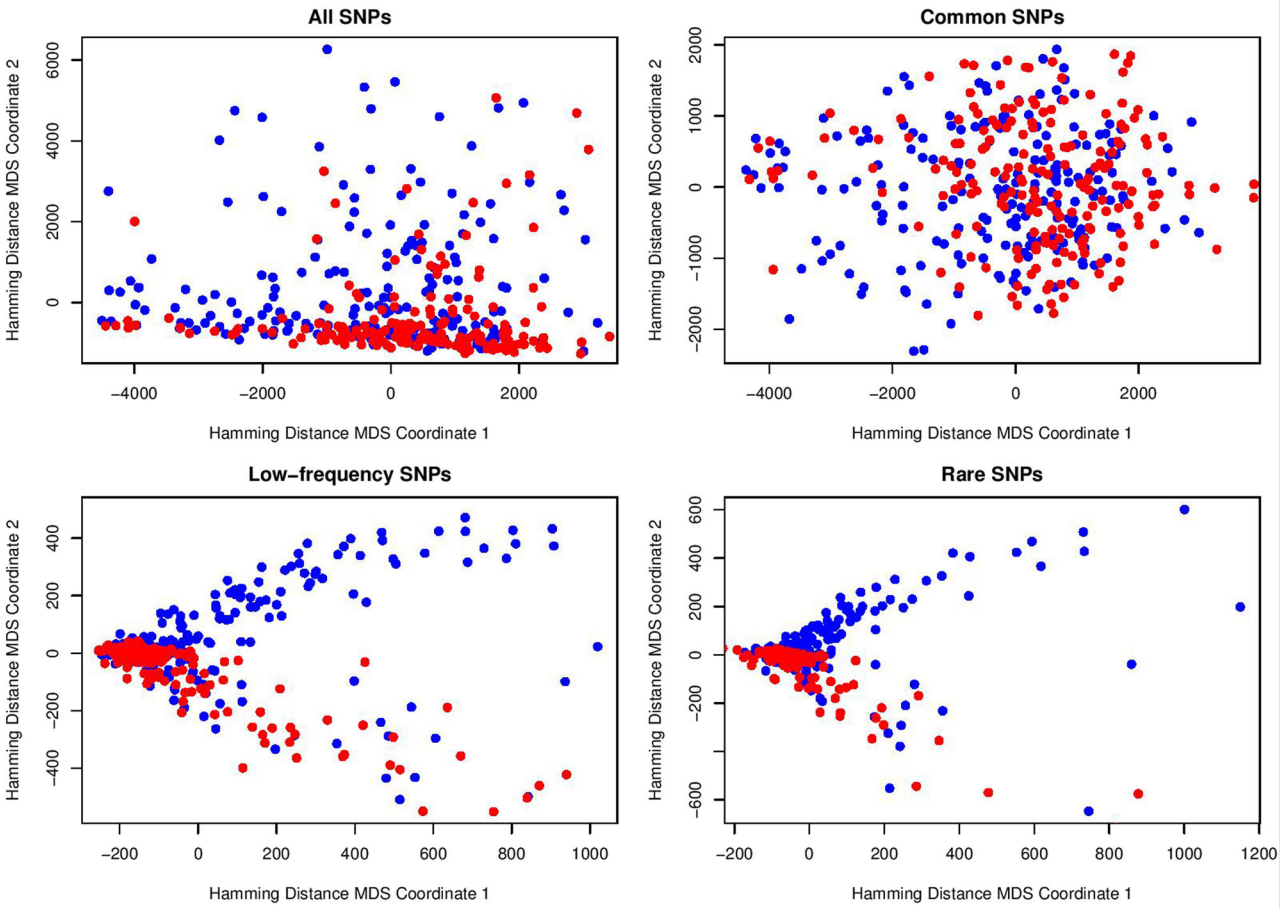

Fig. 3 MDS visualization based on Hamming distance for Mexican-American cohort

statistical analysis, Hotelling's T^2 test for two independent samples is used to examine whether the means of the two groups (case and control) are equal.

Code availability

The codes (by R software; www.r-project.org) of data analysis for this study can be accessed from the authors.

Results

Population stratification

In Figs. 1 and 2, we presented the population stratification results based on all, common, low-frequency, and rare SNP sets for the Mexican-American cohort and the European-ancestry cohort. Although several far outliers are found in Mexican-American cohort for low-frequency and rare SNPs and in European-ancestry cohort for rare SNPs, there is no significant separation between depressed cases (blue points) and controls (red points) in the IBS-MDS plots.

MDS on Hamming distance

In Fig. 3 we presented MDS results on Hamming distance for all, common, low-frequency, and rare SNP sets of the Mexican-American cohort. For common SNPs, there is no significant separation between MDD cases (blue points) and controls (red points). However, for low-frequency and rare SNPs, we found that all the healthy controls were scattered in the lower half-plane, and all points in the upper half-plane were blue points representing depressed cases. We use Hotelling's T^2 test to statistically examine these visual separations between case and control points in the MDS plane. For common variants, the result is P value = 5.891e−05 and T^2 statistic = 9.983. For low-frequency variants, the Hotelling's T^2 test result is Pvalue <2.2e−16 and T^2 statistic = 42.958. For rare variants, the Hotelling's T^2 test result is Pvalue = 7.681e−12 and T^2 statistic = 27.32. Therefore, the separation of cases and controls shown in the MDS plane is much more significant for low-frequency and rare SNPs than for common SNPs.

In Fig. 4, we showed the results of MDS on Hamming distance for all, common, low-frequency, and rare SNP

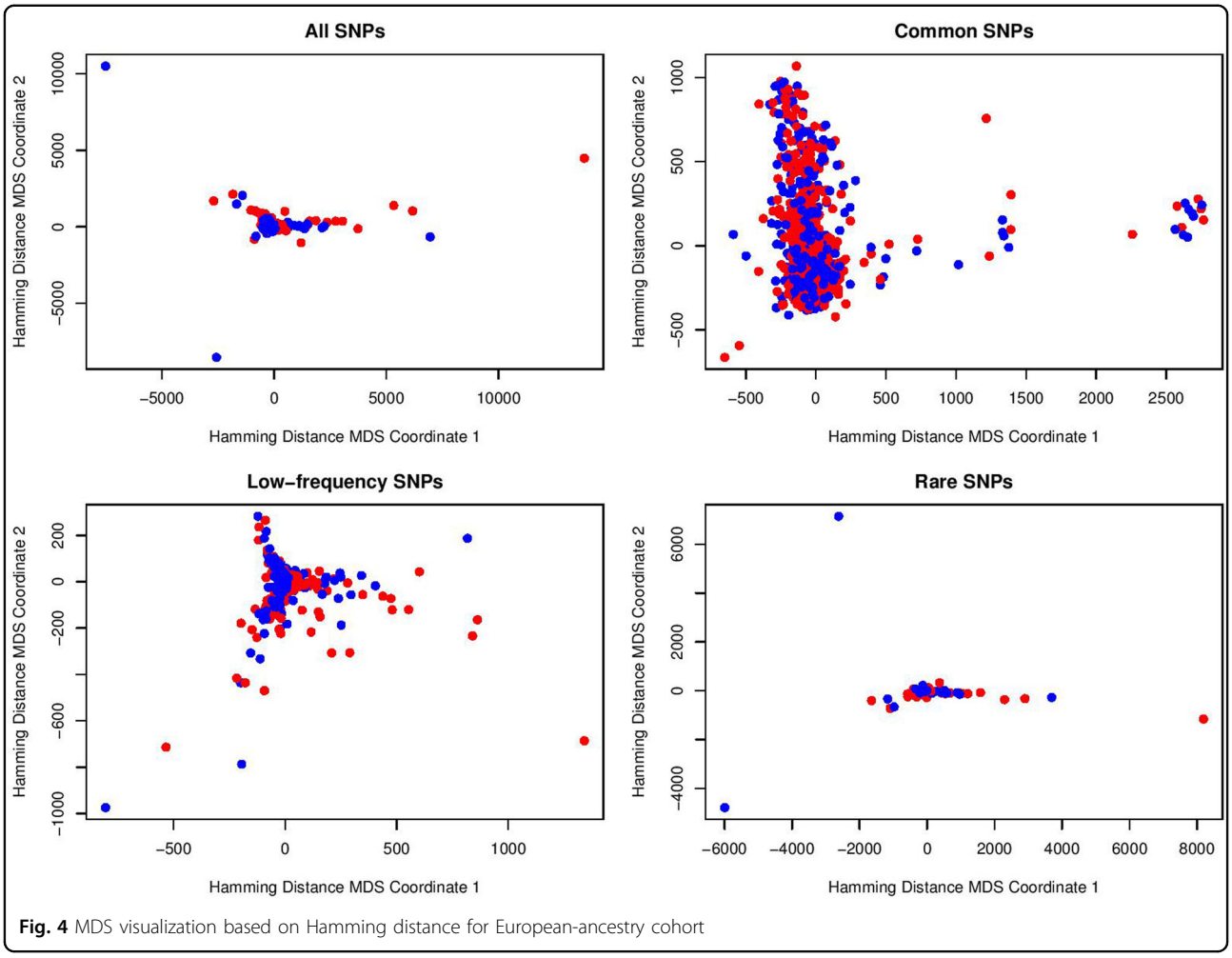

Fig. 4 MDS visualization based on Hamming distance for European-ancestry cohort

sets of the European-ancestry cohort. For all the SNP sets, we see that there is no significant visual separation between MDD cases (blue points) and controls (red points) in the MDS planes. The same results can also be found in Figure S1 by excluding some far outliers.

Discussion

Genetic factors play important roles in the susceptibility to major depression, as indicated by family, twin, and adoption studies[33]. The heritability of MDD is estimated to range between 40 and 70%[34]. In this article, we designed a novel methodology to explore the "missing heritability" of MDD. The significant separations between cases and controls for low-frequency and rare variants in the planes of MDS on Hamming distance supported our previous conclusion[23] that most of the functional variants detected in the Mexican-American cohort were rare. The corresponding statistical results also show that the separation for low-frequency and rare SNPs is much more significant than for common SNPs. Thus, our findings further suggest that low-frequency and rare variants may play more significant roles in the development of MDD. Low-frequency and rare variants are currently not tagged by conventional genome-wide genotyping arrays, thus they may represent an important but understudied component of MDD genetics[29]. There are many different types of technical designs that identify low-frequency and rare variants. In the current study, we applied whole-exome-wide genotyping array data, which are relatively cost-effective. With the rapid development of next-generation sequencing technologies (whole-genome sequencing, whole-exome sequencing, and targeted sequencing of candidate genes), it is now possible to collect most or even all low-frequency and rare genetic variants in large samples and test their roles in human disease risk[35–37]. Thus, future work would be needed to further investigate our methodology on much larger SNP sets.

The traditional genetic distance[38] considers mutation rates and is designed as a measure of genetic divergence

between populations within a species. Therefore, it is not appropriate to examine the genetic variation associated with a complex disorder within a human population, namely Mexican-American. In this work, Hamming distance was used to investigate the genetic distance between two individuals based on their SNP sets. As sequencing costs are currently dropping further, we may examine single-nucleotide variants (SNVs) which involve much more individual genetic information. For example, we recently proposed a new concept of SNV proportion in genes and employed it to develop a predictive approach for major depression[39]. Using similar classification and cluster analysis methods, a potential tool for MDD diagnosis could also be constructed based on Hamming distance and low-frequency/rare SNPs.

Using our methodology with a case–control study, we could examine the effect of a group of variants within a specific range of MAF. The design and methodology we have developed can also be extended to other complex disorders. Our approach based on exome genotyping array or sequencing data may reveal the "missing heritability" resulted from allele frequency for many complex diseases.

Our European-ancestry cohort failed to show significant separations of cases and controls in the MDS planes on Hamming distance. The reasons could be as follows. First, MDD is clearly a gene–environment interaction disorder[34]. Our Mexican-American cohort is comprised of first-generation individuals (60%) who have experienced significant levels of stress and hyperactivation of the hypothalamic-pituitary-adrenal axis related to acculturation issues[40,41]. In contrast to the European-ancestry cohort, the significant stressful life events for Mexican-Americans could cause much higher levels of depression. Therefore, the depression effect differences between case and control in two cohorts may be large. Further studies using our methodology could be tested on a larger size of European-ancestry sample. Second, the European-ancestry cohort that we studied have much lower levels of genetic variants. In our previous work[23], whole-genome sequencing analyses of a subset of the two cohorts revealed that European-ancestry subjects have a significantly reduced (around 50%) number of SNVs compared with Mexican-American subjects. For this reason, the roles of low-frequency and rare variants may vary across populations.

Acknowledgements
We have been supported by grants APP1051931 (M.-L.W. and M.A.-B.), APP1070935 (M.-L.W.), and APP1060524 (B.T.B.) from NHMRC of Australia, the German Research Foundation Grant FOR 2107, DA1151/5-1 (UD), NIH Grant GM61394 (M.-L.W.), and institutional funds from the South Australian Health and Medical Research Institute, Flinders University. and the Australian National University.

Author details
[1]Centre for Population Health Research, School of Health Sciences and Sansom Institute of Health Research, University of South Australia, Adelaide, SA, Australia. [2]Mind and Brain Theme, South Australian Health and Medical Research Institute, Adelaide, SA, Australia. [3]College of Medicine and Public Health, Flinders University, Bedford Park, SA, Australia. [4]GENIUROS group, Center for Research in Genetics and Genomics, Institute of Translational Medicine, School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia. [5]Discipline of Psychiatry, Adelaide Medical School, University of Adelaide, Adelaide, SA, Australia. [6]Department of Psychiatry and Psychotherapy, University of Münster, Münster, Germany. [7]Department of Psychiatry and Psychotherapy, University of Marburg, Marburg, Germany. [8]Departments of Psychiatry, Pharmacology and Medicine, College of Medicine, State University of New York, Upstate Medical University, Syracuse, NY, USA

Conflict of interest
The authors declare that they have no conflict of interest.

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41398-018-0117-7.

Received: 24 September 2017 Revised: 1 December 2017 Accepted: 30 December 2017
Published online: 27 March 2018

References

1. Kessler, R. C. et al. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Arch. Gen. Psychiatry* **51**, 8–19 (1994).
2. Lopez, A. D. & Murray, C. C. The global burden of disease, 1990–2020. *Nat. Med.* **4**, 1241–1243 (1998).
3. Wong, M. L. & Licinio, J. Research and treatment approaches to depression. *Nat. Rev. Neurosci.* **2**, 343–351 (2001).
4. Wong, M. L. & Licinio, J. From monoamines to genomic targets: a paradigm shift for drug discovery in depression. *Nat. Rev. Drug. Discov.* **3**, 136–151 (2004).
5. Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R. & Walters, E. E. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* **62**, 617–627 (2005).
6. Lohoff, F. W. Overview of the genetics of major depressive disorder. *Curr. Psychiatry Rep.* **12**, 539–546 (2010).
7. Flint, J. & Kendler, K. S. The genetics of major depression. *Neuron* **81**, 484–503 (2014).
8. CONVERGE Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).
9. Amin, N. et al. Exome-sequencing in a large population-based study reveals a rare Asn396Ser variant in the LIPG gene associated with depressive symptoms. *Mol. Psychiatry* **22**, 537–543 (2017).
10. Hyde, C. L. et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Genet.* **48**, 1031–1036 (2016).
11. Sullivan, P. F., Daly, M. J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet.* **13**, 537–551 (2012).
12. Peterson, R. E. et al. The genetic architecture of major depressive disorder in Han Chinese women. *JAMA Psychiatry* **74**, 162–168 (2017).
13. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
14. Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).
15. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).

16. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. USA* **109**, 1193–1198 (2012).
17. Wray, N. R. & Maier, R. Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability. *Curr. Epidemiol. Rep.* **1**, 220–227 (2014).
18. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
19. Ott, J., Wang, J. & Leal, S. M. Genetic linkage analysis in the age of whole-genome sequencing. *Nat. Rev. Genet.* **16**, 275–284 (2015).
20. Knowles, E. E. et al. Genome-wide linkage on chromosome 10q26 for a dimensional scale of major depression. *J. Affect. Disord.* **191**, 123–131 (2016).
21. Imai, A. et al. Beyond homozygosity mapping: family-control analysis based on Hamming distance for prioritizing variants in exome sequencing. *Sci. Rep.* **5**, 12028 (2015).
22. Imai, A. et al. HDR: a statistical two-step approach successfully identifies disease genes in autosomal recessive families. *J. Hum. Genet.* **61**, 959–963 (2016).
23. Wong, M. L. et al. The PHF21B gene is associated with major depression, and modulates stress response. *Mol. Psychiatry* **22**, 1015–1025 (2017).
24. Dong, C., Wong, M. L. & 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* **14**, 1105–1118 (2009).
25. Wong, M. L., 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* **17**, 624–633 (2012).
26. Wong, M. L. et al. Clinical outcomes and genome-wide association for a brain methylation site in an antidepressant pharmacogenetics study in Mexican Americans. *Am. J. Psychiatry* **171**, 1297–1309 (2014).
27. Baune, B. T. & 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* **7**, 150 (2016).
28. Guo, Y. et al. Illumina human exome genotyping array clustering and quality control. *Nat. Protoc.* **9**, 2643–2662 (2014).
29. Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* **7**, 16 (2015).
30. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
31. Hamming, R. W. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**, 147–160 (1950).
32. Torgerson, W. S. Multidimensional scaling: I. theory and method. *Psychometrika* **17**, 401–419 (1952).
33. Sullivan, P. F., Neale, M. C. & Kendler, K. S. Genetic epidemiology of major depression: review and meta-analysis. *Am. J. Psychiatry* **157**, 1552–1562 (2000).
34. Lesch, K. P. Gene–environment interaction and the genetics of depression. *J. Psychiatry Neurosci.* **29**, 174–184 (2004).
35. Dunn, E. C. et al. Genetic determinants of depression: recent findings and future directions. *Harv. Rev. Psychiatry* **23**, 1–18 (2015).
36. Yu, C., Baune, B. T., Licinio, J. & Wong, M. L. A novel strategy for clustering major depression individuals using whole-genome sequencing variant data. *Sci. Rep.* **7**, 44389 (2017).
37. Yu, C., Baune, B. T., Licinio, J. & Wong, M. L. Whole-genome single nucleotide variant distribution on genomic regions and its relationship to major depression. *Psychiatry Res.* **252**, 75–79 (2017).
38. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, New York, 2000).
39. Yu, C., Baune, B. T., Licinio, J. & Wong, M. L. Single-nucleotide variant proportion in genes: a new concept to explore major depression based on DNA sequencing data. *J. Hum. Genet.* **62**, 577–580 (2017).
40. Caplan, S. et al. Cultural influences on causal beliefs about depression among Latino immigrants. *J. Transcult. Nurs.* **24**, 68–77 (2013).
41. Korenblum, W. et al. Elevated cortisol levels and increased rates of diabetes and mood symptoms in Soviet Union-born Jewish immigrants to Germany. *Mol. Psychiatry* **10**, 974–975 (2005).