

ORIGINAL ARTICLE

A latent genetic subtype of major depression identified by whole-exome genotyping data in a Mexican-American cohort

C Yu^{1,2}, M Arcos-Burgos^{3,4}, J Licinio^{1,2,5} and M-L Wong^{1,2}

Identifying data-driven subtypes of major depressive disorder (MDD) is an important topic of psychiatric research. Currently, MDD subtypes are based on clinically defined depression symptom patterns. Although a few data-driven attempts have been made to identify more homogenous subgroups within MDD, other studies have not focused on using human genetic data for MDD subtyping. Here we used a computational strategy to identify MDD subtypes based on single-nucleotide polymorphism genotyping data from MDD cases and controls using Hamming distance and cluster analysis. We examined a cohort of Mexican-American participants from Los Angeles, including MDD patients ($n=203$) and healthy controls ($n=196$). The results in cluster trees indicate that a significant latent subtype exists in the Mexican-American MDD group. The individuals in this hidden subtype have increased common genetic substrates related to major depression and they also have more anxiety and less middle insomnia, depersonalization and derealisation, and paranoid symptoms. Advances in this line of research to validate this strategy in other patient groups of different ethnicities will have the potential to eventually be translated to clinical practice, with the tantalising possibility that in the future it may be possible to refine MDD diagnosis based on genetic data.

Translational Psychiatry (2017) **7**, e1134; doi:10.1038/tp.2017.102; published online 16 May 2017

INTRODUCTION

The chronicity of major depressive disorder (MDD) results in tremendous medical, social, and economic impact. MDD is a major contributor to global disease burden and produces considerable morbidity and mortality.^{1–5} Despite recent advances,⁶ little is known about its underlying fundamental biology and much work still needs to be done to fully elucidate the genetic factors that confer susceptibility to MDD.^{7–10} Clinically, major depression has been classified based on various, distinct features that include course, periodicity, qualitative and quantitative types of symptoms, clinical features, age or phase of life, and cause.¹¹ Those categories are based on historical observations, and sometimes are unconvincing or controversial. For example, atypical depression is the most common form of depression in outpatients, but beyond the well-characterized constellation of symptoms (mood reactivity, leaden paralysis, hyperphagia, hypersomnia and rejection sensitivity) that define it, the biological course of this presentation remains unknown.¹² It has not yet been established if atypical depression is a stable subtype or if it is just one of several forms of MDD that an individual may express during a lifetime of recurrent depressions.^{13,14} As different subtypes of MDD may respond differentially to various medications, it is critical that we elucidate the natural course of this disorder.

Efforts to explore subtypes of depression have recently been made using sophisticated statistical models on clinical data;^{15–20} however, there has been no studies on genetic MDD subtyping. Recent advances in high-throughput genomic technologies provide considerable opportunities for medical research. Clinical care appears to be moving toward genotyping/sequencing-based

precision medicine, and single-nucleotide polymorphism (SNP) genotyping is currently the most popular technique used in genome-wide association studies, which identify variations that are significantly associated with a trait or disease.²¹ In addition to searching for SNPs or genes that are significantly associated with a disease, it is also important to understand whether genetic data could be used to identify disease subtypes.

Here we developed a computational strategy that identifies genetic subtypes using functional SNPs. A group of Mexican-American patients from Los Angeles was examined in this study. We chose this group because the Hispanic population is currently the largest ethnic minority group in the United States, representing over 37 million people, and within this group, almost 70% are Mexican-Americans.²² Although this population is growing markedly, there is little research on psychiatric diagnosis and treatment in this group.²³ The idea for this new approach arose from distance-based phylogenetic analyses of genetic sequences described by us earlier.^{24–28} In the proposed methodology, we applied Hamming distance on a SNP set to measure the genetic similarity between two individuals. Then, we reconstructed a cluster tree based on the Hamming distance matrix of all individuals; cluster relationships in the tree revealed interesting and meaningful MDD subtypes.

MATERIALS AND METHODS

The Los Angeles Mexican-American cohort

We investigated a Los Angeles Mexican-American group of 203 MDD patients (50.88%) and 196 healthy controls (49.12%) aged 19–65 years,

¹Mind and Brain Theme, South Australian Health and Medical Research Institute, Adelaide, SA, Australia; ²School of Medicine, Flinders University, Bedford Park, Adelaide, SA, Australia; ³Department of Genome Sciences, John Curtin School of Medical Research, Australian National University, Canberra, ACT, Australia; ⁴University of Rosario International Institute of Translational Medicine, Bogota, Colombia and ⁵South Ural State University Biomedical School, Chelyabinsk, Russia. Correspondence: Dr C Yu or Professor M-L Wong, Mind and Brain Theme, South Australian Health and Medical Research Institute, PO Box 11060, Adelaide, SA 5001, Australia.

E-mail: chenglong.yu@sahmri.com or mali.wong@sahmri.com

Received 9 November 2016; revised 4 April 2017; accepted 10 April 2017

which was a convenience sample, as we had previously obtained and performed classical genetic analysis on functional SNP data in this cohort.²⁹ Participants provided written informed consent, and detailed demographic, epidemiological and clinical descriptions were previously described.^{30–32} 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, Australia.

Individuals in this group had three or more grandparents born in Mexico. MDD was diagnosed using the SCID (Structured Clinical Interview for DSM-IV (Diagnostic and Statistical Manual IV edition)). Patients met diagnostic criteria for current, unipolar major depressive episode, participated in a pharmacogenetic study of antidepressant treatment and had an initial HAM-D 21 (21-Item Hamilton Depression Rating Scale) score of 18 or greater with item number 1 (depressed mood) rated 2 or greater. MDD was defined as five out of nine criteria in the SCID. The structured clinical interview for the DSM-IV Axis I Disorders had a mean kappa score for sensitivity and specificity among raters of 0.84–0.85. 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.^{30–32} Control subjects responded that they were in good health and answered to acculturation questionnaires. However, they were not screened for medical illnesses and did not respond to structured psychiatric interviews. They were age- and gender-matched Mexican-American individuals recruited from the same community in Los Angeles.

SNP genotyping data analyses

The cohort was genotyped by the Australian Genome Research Facility (North Melbourne, VIC, Australia; www.agrf.org.au) using the Illumina HumanExome BeadChip-12v1_A (San Diego, CA, USA), which exonic content consists of >250 000 markers representing diverse populations and a range of common conditions. All samples passed the Illumina expected SNP call rate (>99%). Detailed genotyping data analyses have been reported in our recent work and briefly described here.²⁹ We analyzed 83 898 common and rare SNP variants that remained after raw whole-exome SNP data (247 909 variants) from 399 Mexican-American subjects, were filtered by a pipeline that considered call rate, number of alleles and Hardy–Weinberg equilibrium deviation. The identity by descent matrix between all pairs of individuals was estimated after linkage disequilibrium pruning and used for quality control and for the mixed linear models analyses. Then, the association between MDD and those SNPs was analyzed using single- and multi-locus linear mixed-effect models³³ with up to 10 steps in the backward/forward optimization algorithm. Models included fixed (SNPs, gender and age) and random (family or population structure) effects and were both implemented in SVS 8.3.0 (Golden Helix, Bozeman, MT, USA). A total of 19 common SNPs (rs41310573, rs201935337, rs140395831, rs56293203, rs78562453, rs115054458, rs143696449, rs748441912, rs62001028, rs150952348, rs782472239, rs112610420, rs142029931, rs201483250, rs200897153, rs3744550, rs115668237, rs56344012 and rs200520741) in 18 genes were significantly associated with MDD at the genome-wide false discovery rate < 0.05. It is worth mentioning that principal component analysis of random effects clearly showed the absence of family or population stratification in this cohort.²⁹ In the approaches described below, we tested all 83 898 variants and the 19 significant variants separately.

The Hamming distance between two individuals

The traditional genetic distance, such as in Nei *et al.*³⁴ and Goldstein *et al.*,³⁵ is designed as a measure of the genetic divergence between populations within a species; and thus it is not appropriate to use this approach to explore the genetic variations associated with a complex disease within a human population, namely, Mexican-American. Here we introduce the Hamming distance,³⁶ which is a natural distance without the assumption of any model of mutation/substitution rate, to investigate the genetic similarity between two individuals based on a set of SNPs.

Let S be a SNP set that contains n SNPs. We use SNP_k to represent the SNP indexed k ($k = 1, \dots, n$). Thus, $S = \{SNP_1, SNP_2, \dots, SNP_n\}$. If P and Q are two individuals, their genotypes in SNP sets are respectively named S_P and S_Q . Let S_P be $\{SNP_1^P, SNP_2^P, \dots, SNP_n^P\}$ and S_Q be $\{SNP_1^Q, SNP_2^Q, \dots, SNP_n^Q\}$. Then, the Hamming distance between the two individuals

Table 1. Hamming distances of three subjects in a six-SNP set

Genotype	SNP1 (C/T)	SNP2 (A/T)	SNP3 (G/T)	SNP4 (A/G)	SNP5 (C/G)	SNP6 (C/T)
Subject P	CC	AT	GG	AG	CG	CC
Subject Q	CT	AA	GG	GG	CC	TT
Subject R	CT	AA	GG	AA	CC	CC

Abbreviation: SNP, single-nucleotide polymorphism.

P and Q is defined as $H(P, Q) = \sum_{i=1}^n \delta(SNP_i^P, SNP_i^Q)$, where $\delta(x, y) = \begin{cases} 0 & \text{if } x \text{ and } y \text{ are the same} \\ 1 & \text{otherwise} \end{cases}$, that is, the number of positions at

which the corresponding SNPs are different in the SNP set S . And, the normalized Hamming distance is defined as: $NH(P, Q) = \sum_{i=1}^n \delta(SNP_i^P, SNP_i^Q) / n$. In Table 1, individuals P , Q and R show their genotypes in a six-SNP set. Thus, the Hamming distance between P and Q is 5, the Hamming distance between P and R is 4 and the Hamming distance between Q and R is 2. Our hypothesis was that if two individuals have a closer Hamming distance, then those two individuals would have more similar phenotypes, such as diseases or traits. In the above example, we assume that Q and R possess more similar phenotypes.

The population stratification must be corrected before the SNP set can be used in this method. Principal component analysis was used to confirm there was no family or population structure among individuals in our Mexican-American cohort.²⁹ In a given group of individuals, we can calculate their Hamming distance matrix based on a specific SNP set. After obtaining the distance matrix across the individuals, two methods can be used to map the distance matrix into a two-dimensional picture: (1) the multi-dimensional scaling (MDS) method and (2) the clustering tree method.

MDS and the Hamming distance matrix

The classical MDS method proposed by Torgerson³⁷ is aimed at representing high-dimensional data or a distance matrix into a low-dimensional space with preservation of similarities between data points, which can visually disclose some structures hidden in the data. We used the classical MDS method to map the Hamming distance matrix into a two-dimensional Euclidean plane, and in this plane, each individual is represented by one point in the scatter plot. The MDS method gives a data visualization of all individuals and aims at preserving the between-individual Hamming distances in a two-dimensional space as accurately as possible.

The clustering tree method and the Hamming distance matrix

We drew cluster trees employing hierarchical cluster analysis of Hamming distance matrix data. We used the popular distance-based neighbor-joining method,³⁸ which is a bottom-up agglomerative strategy for reconstructing trees. Each external node in the tree represents one individual, and the edge length in the tree indicates exactly the Hamming distances among individuals. Cluster trees were drawn using the MEGA 6 software³⁹ (www.megasoftware.net).

Statistical analysis

The difference between two group means on each item was tested using an independent two-sample Student's t -test. Multiple testing was addressed by correcting P -values using the false discovery rate method, and the significance level was set at ≤ 0.05 .

Code availability

All data were analyzed using the R software (www.r-project.org), and the code can be accessed from the authors.

RESULTS

MDS visualization on two SNP sets

In Supplementary Table S1, we summarize descriptive statistics of gender, age, Hamilton depression rating scale (HAM-D) scores and educational levels for all Mexican-American subjects. Following the proposed method, we used the normalized Hamming distances (NHD) to obtain the distance matrices cross the 399

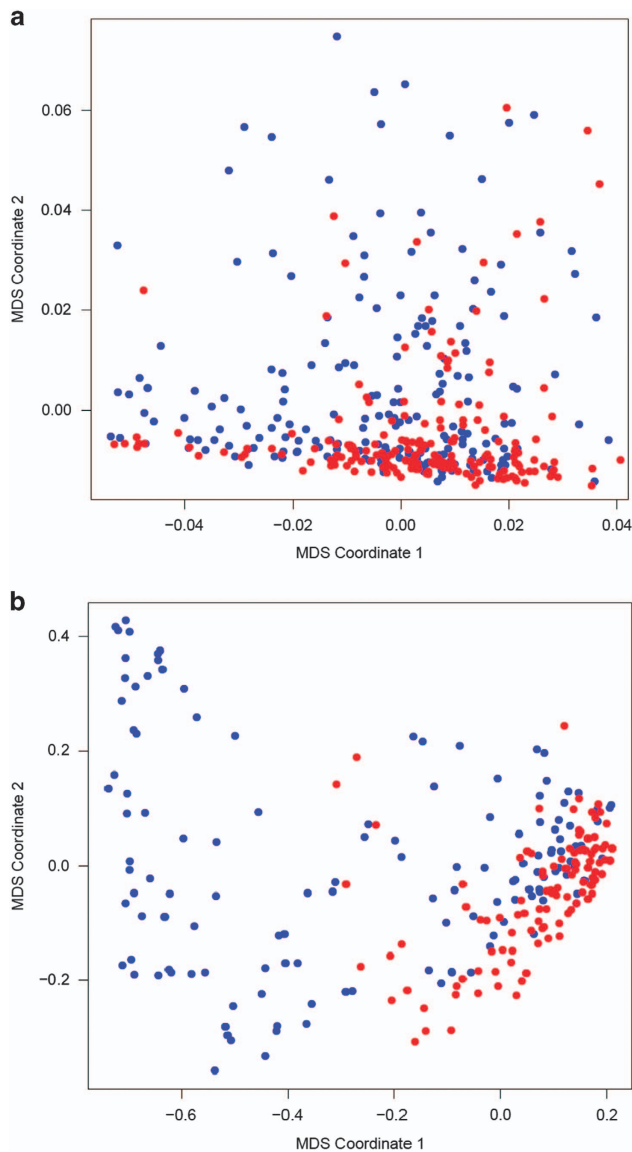


Figure 1. MDS two-dimensional visualization of 399 Mexican-American subjects (MDD cases are represented by blue dots and controls are represented by red dots) in (a) the 83 898 SNP set and in (b) the 19 significant SNP set. MDD, major depressive disorder; MDS, multi-dimensional scaling; SNP, single-nucleotide polymorphism.

participants for the SNP set of 83 898 variants and the SNP set of 19 variants. We checked the pairwise NHD in the (i) MDD group, (ii) control group and (iii) MDD cross-control group. In the MDD cross-control group, each pairwise distance was calculated between one individual in the MDD group and one individual in the control group. For the SNP set of 83 898 variants, the NHD mean \pm s.d. was 0.185 ± 0.016 , 0.184 ± 0.015 and 0.186 ± 0.008 for the MDD, control and MDD cross-control groups, respectively. For the SNP set of 19 variants, the NHD mean \pm s.d. was 0.516 ± 0.301 , 0.222 ± 0.163 and 0.426 ± 0.308 for the MDD, control and MDD cross-control groups, respectively. Then, we applied the classical MDS to map the two distance matrices in a two-dimensional Euclidean plane, in which each point in the plane represents one individual. Figure 1a, which is based on 83 898 SNPs, shows no significant difference between cases (blue points) and controls (red points). However, in Figure 1b, which is based on 19 significant variants, there are clearly some cases (blue points) that scatter far away from other individuals, especially from controls

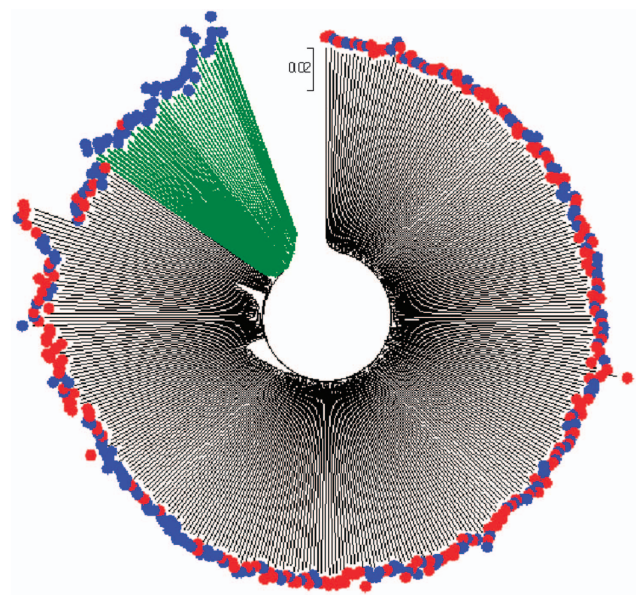


Figure 2. Cluster tree for 399 Mexican-American subjects (MDD cases are represented by blue external nodes and controls are represented by red external nodes) in the 83 898 SNP set. MDD, major depressive disorder; SNP, single-nucleotide polymorphism.

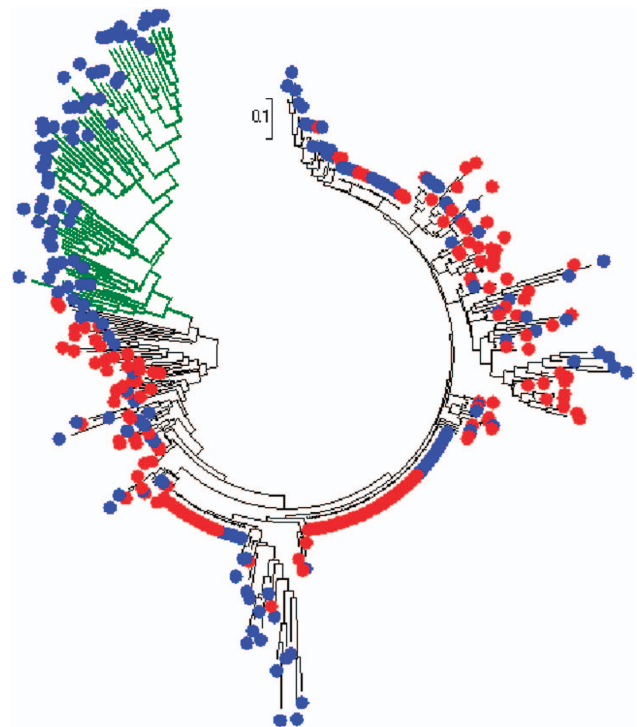


Figure 3. Cluster tree for 399 Mexican-American subjects (MDD cases are represented by blue external nodes and controls are represented by red external nodes) in the 19 significant SNP set. MDD, major depressive disorder; SNP, single-nucleotide polymorphism.

(red points). This interesting finding implies that a latent subgroup of MDD cases may exist in this cohort.

Subtype identification using cluster tree

We applied the neighbor-joining method to reconstruct the cluster trees in Figures 2 and 3 to the distance matrices data

Table 2. The statistical analysis on HAM-D scores for the group of MDD patients in the identified subtype and the group of MDD patients not in the identified subtype

HAM-D 21 items	Score range	Mean score for MDDs in the subtype	Mean score for MDDs not in the subtype	t-value	P-value	FDR
1. Depressed mood	0–4	1.88	1.73	0.5851	0.2799	0.4024
2. Feelings of guilt	0–4	1.36	1.42	–0.2836	0.3887	0.4686
3. Suicide	0–4	0.70	0.52	1.0234	0.1544	0.3222
4. Insomnia early	0–2	1.06	1.11	–0.2761	0.3915	0.4686
5. Insomnia middle	0–2	0.76	1.03	–1.7261	0.0438*	0.2519
6. Insomnia late	0–2	0.97	1.14	–0.9664	0.1681	0.3222
7. Work and activities	0–4	1.85	1.81	0.1653	0.4345	0.4759
8. Retardation	0–4	0.73	0.53	1.1102	0.1349	0.3169
9. Agitation	0–4	0.61	0.52	0.6073	0.2725	0.4024
10. Anxiety	0–4	1.55	1.14	1.9417	0.0276*	0.2377
11. Anxiety somatic	0–4	1.48	1.50	–0.0874	0.4653	0.4865
12. Somatic symptoms gastro-intestinal	0–2	0.39	0.34	0.4015	0.3445	0.4661
13. Somatic symptoms general	0–2	1.27	1.13	1.0965	0.1378	0.3169
14. Genital symptoms	0–2	1.18	1.28	–0.6405	0.2617	0.4024
15. Hypochondriasis	0–4	0.28	0.48	–1.3961	0.0830	0.2795
16a. Loss of weight (rating by history)	0–2	0.45	0.31	0.7707	0.2217	0.3922
16b. Loss of weight (measured weight change)	0–2	0.36	0.31	0.2368	0.4075	0.4686
17. Insight	0–2	0.24	0.38	–1.3139	0.0960	0.2795
18a. Diurnal variation: AM worse	0–2	0.82	0.82	0.0082	0.4967	0.4967
18b. Diurnal variation: PM worse	0–2	0.56	0.86	–1.5833	0.0590	0.2714
19. Depersonalization and derealisation	0–4	0.15	0.47	–2.3350	0.0108*	0.2377
20. Paranoid symptoms	0–3	0.03	0.20	–1.8881	0.0310*	0.2377
21. Obsessional and compulsive symptoms	0–2	0.06	0.17	–1.3071	0.0972	0.2795
Total score	0–67	18.12	18.42	–0.1815	0.4282	NA

Abbreviations: FDR, false discovery rate; HAM-D 21, 21-Item Hamilton Depression Rating Scale; MDD, major depressive disorder; NA, not available. We use * to show the significant test results.

across the 399 subjects for the SNP set of 83 898 variants and the SNP set of 19 variants, respectively. In Figure 2, which is based on 83 898 SNPs, there is a significant subgroup of MDD cases (marked in green color to display its branches) in the cohort, although the normalized Hamming distances of this subgroup are not significantly away from the other subjects. In Figure 3, which is based on 19 significant variants, the MDD subgroup (also marked in green color) in the tree looks very obvious. This also implies that the 19 significant SNPs detected by genome-wide association studies can indeed capture most of the information from the 83 898 common SNPs. The two newly found subgroups in Figures 2 and 3 contain the same 41 MDD subjects (see Supplementary Figures S1 and S2 and Supplementary Table S1 for details), which we consider as a latent subtype in the Mexican-American MDD group.

Our new approach shows that a potential subtype exists in our Mexican-American MDD sample. We have confirmed that there were no blood relatives between those Mexican-American individuals.²⁹ Therefore, the identification of subgroups in the cluster trees was not due to genetic relatedness. Supplementary Table S1 also summarizes descriptive statistics of gender, age, HAM-D scores and educational levels for the identified subgroup of 41 MDD cases. There are 150 males and 249 females (sex ratio of 60.2%) in our Mexican-American cohort with an average age of 39.2 years with s.d. 11.5. In the latent MDD subgroup, there are 41 subjects (15 males and 26 females) with the sex ratio of 57.7% and an average age of 38.9 years with s.d. 10.3. Therefore, this latent MDD subgroup was not associated with gender or age.

Table 2 contains the statistical results of HAM-D 21 items for the MDD latent subtype group and the remainder group of MDD patients. Although false discovery rate did not identify significant results, the original *t*-tests showed potential significant symptom differences between two groups—insomnia middle (decreased), anxiety (increased), depersonalization and derealisation (decreased), and paranoid symptoms (decreased).

DISCUSSION

In this study, we developed a computational strategy to identify MDD subtypes based on SNP genotyping data using Hamming distance and cluster analysis. The results in cluster trees indicate that a significant latent subtype exists in the Mexican-American MDD group. The individuals in the hidden subtype have increased common genetic substrates related to MDD and they may also have more anxiety, and less middle insomnia, depersonalization, derealisation and paranoid symptoms.

We used the Hamming distance on SNP data to generate the distance matrix for subjects. To show the close/distant relationships of subjects in a two-dimensional space, we used the MDS and neighbor-joining tree methods. Both methods can project individuals in a two-dimensional plane. The MDS worked well with the set of 19 significant SNPs (Figure 1b) but not with the set of 83 898 SNPs (Figure 1a). However, the neighbor-joining tree worked well for the sets of 19 significant SNPs and 83 898 SNPs; therefore, both Figures 2 and 3 identify the same MDD subtype. It is known that the MDS method may lose distance information in the conversion process from the distance matrix to a two-dimensional projection. Therefore, the Euclidean distance (visual representation) between two points in the MDS two-dimensional plane may differ from the original Hamming distance. In contrast, the neighbor-joining tree method preserves the original distance between points. The additive distance between two leaf nodes in the tree is identical to the one in the distance matrix. Furthermore, among distance-based tree construction methods, the neighbor-joining algorithm does not assume a constant rate of evolution, as opposed to the molecular clock hypothesis that has always been controversial.^{40,41} Because of its low computational complexity, the neighbor-joining algorithm can be performed very fast and is widely used to generate phylogenetic trees of a large number of biological species or other entities.⁴² Thus, to obtain an accurate subtyping identification, we recommend the phylogenetic cluster

approach to build the hierarchical tree for those subjects. Actually, the Hamming distance is a powerful mathematical tool that can be used for many genetic applications, such as checking population structures⁴³ or family-control analysis.⁴⁴

Our Mexican-American cohort and the International Haplotype Map Project (HapMap) cohort were recruited from the same community in Los Angeles by the same team; thus, they both have an admixture of 49% European, 45% Indigenous American and 5% African ancestries.⁴⁵ According to the International HapMap 3 Consortium⁴⁶ and the 1000 Genomes Project Consortium,⁴⁷ it would be expected that individuals with African ancestry, such as Mexican-Americans, have an increased number of variants compared with other populations, such as Northern European. Future work is needed to extend our new methodology to other populations. Some distance formulae such as the Hamming distance may need adjustments for different populations.

We used our existing whole-exome SNP genotyping data in the work described here. The fact that those subjects fail to be clustered as expected into two big groups (cases and controls) may be due to the SNP set selection. Without doubt, a larger SNP set would reveal more interesting and comprehensive findings. As whole-genome sequencing costs are projected to decrease further, we may have the opportunity to examine single-nucleotide variant set, which involves much more individual genetic data. Consequently, further studies utilizing our method should examine larger genotyping or sequencing data. In this work, we tested the clinical relevance of our new genetic subtype using HAM-D 21 items. Our statistical results did not show significant differences between patients in the new subtype and the remaining patients for most HAM-D items. This could be explained by the high complexity of MDD clinical symptoms. Therefore, to investigate more external clinical variables, future research needs to be performed using longitudinal data for depressed patients with detailed information on course of depression, antidepressant responses and suicidality.

To the best of our knowledge, this is the first study on genetic subtyping of MDD. Genome-wide association studies examine genome-wide genetic variants in case and control samples to identify variants that are associated with a trait or disease. Genome-wide association study association findings do not directly contribute to disease subtyping. Therefore, the knowledge that our 19 SNPs were significantly associated with MDD did not translated directly into clear MDD subtypes, which could not be identified without the introduction of the Hamming distance and neighbor-joining tree analyses. Results displayed graphically in cluster trees are user-friendly and allow non-experts to easily visualize the close/distant relationships between subjects. Our approach may result in a useful future clinical predictive/diagnostic tool. One could evaluate whether genotyping data from a new subject could be used to determine whether that subject would be within or close to an existing MDD genetic subtype. However, further studies are needed using cohort of different ethnicities to determine whether such a strategy may be successfully translated into clinical practice. This method, in concert with clinical symptom data, has the potential to eventually be translated to clinical practice and could refine the ability to diagnose and classify depressed patients. Better understanding of MDD subtypes may help optimize treatment approaches.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

This study was supported by grants APP1051931 (MLW and MAB) and APP1070935 (MLW and JL) from NHMRC of Australia; grant GM61394 (JL and MLW) from NIH of USA; grant 02.A03.21.0011 (JL) from the 5-100 Project, under the Act 211 of the

Government of the Russian Federation; and institutional funds from the Australian National University, Flinders University and the South Australian Health and Medical Research Institute. We are grateful for the contributions of Israel Alvarado, Deborah L Flores, Rita Jepson, Lorraine Garcia-Teague, Patricia Reyes, Isabel Rodriguez, Gabriela Marquez and the University of California, Los Angeles General Clinical Research Center (UCLA GCRC) staff in recruiting and caring of the Mexican-American participants.

REFERENCES

- Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S *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* 1994; **51**: 8–19.
- Lopez AD, Murray CC. The global burden of disease, 1990–2020. *Nat Med* 1998; **4**: 1241–1243.
- Wong ML, Licinio J. Research and treatment approaches to depression. *Nat Rev Neurosci* 2001; **2**: 343–351.
- 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.
- 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.
- Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR *et al*. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat Genet* 2016; **48**: 1031–1036.
- Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet* 2012; **13**: 537–551.
- CONVERGE Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* 2015; **523**: 588–591.
- Amin N, Jovanova O, Adams HH, Dehghan A, Kavousi M, Vernooij MW *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* 2017; **22**: 537–543.
- Rao S, Yao Y, Ryan J, Li T, Wang D, Zheng C *et al*. Common variants in FKBP5 gene and major depressive disorder (MDD) susceptibility: a comprehensive meta-analysis. *Sci Rep* 2016; **6**: 32687.
- American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fifth edition (DSM-5)*. American Psychiatric Publishing: Arlington, Virginia, USA, 2013.
- Nierenberg AA, Alpert JE, Pava J, Rosenbaum JF, Fava M. Course and treatment of atypical depression. *J Clin Psychiatry* 1997; **59**: 5–9.
- Angst J, Gamma A, Benazzi F, Ajdacic V, Rössler W. Melancholia and atypical depression in the Zurich study: epidemiology, clinical characteristics, course, comorbidity and personality. *Acta Psychiatr Scand* 2007; **115**: 72–84.
- Oquendo MA, Barrera A, Ellis SP, Li S, Burke AK, Grunebaum M *et al*. Instability of symptoms in recurrent major depression: a prospective study. *Am J Psychiatry* 2004; **161**: 255–261.
- Lamers F, de Jonge P, Nolen WA, Smit JH, Zitman FG, Beekman AT *et al*. Identifying depressive subtypes in a large cohort study: results from the Netherlands Study of Depression and Anxiety (NESDA). *J Clin Psychiatry* 2010; **71**: 1582–1589.
- van Loo HM, de Jonge P, Romeijn JW, Kessler RC, Schoevers RA. Data-driven subtypes of major depressive disorder: a systematic review. *BMC Med* 2012; **10**: 156.
- Harald B, Gordon P. Meta-review of depressive subtyping models. *J Affect Disord* 2012; **139**: 126–140.
- Hybels CF, Landerman LR, Blazer DG. Latent subtypes of depression in a community sample of older adults: can depression clusters predict future depression trajectories? *J Psychiatr Res* 2013; **47**: 1288–1297.
- Huang HH, Xu T, Yang J. Comparing logistic regression, support vector machines, and permanent classification methods in predicting hypertension. *BMC Proc* 2014; **8**(Suppl 1): S96.
- Ulbricht CM, Rothschild AJ, Lapane KL. The association between latent depression subtypes and remission after treatment with citalopram: A latent class analysis with distal outcome. *J Affect Disord* 2015; **188**: 270–277.
- Lohoff FW. Overview of the genetics of major depressive disorder. *Curr Psychiatry Rep* 2010; **12**: 539–546.
- US Census Bureau. *The Hispanic Population in the United States*. Census Bureau: Washington, DC, USA, 2001.
- Rogler LH, Malgady RG, Rodriguez O. *Hispanics and Mental Health: a Framework for Research*. Robert E Krieger: Malabar, Fla, 1989.
- Yu C, Liang Q, Yin C, He RL, Yau SST. A novel construction of genome space with biological geometry. *DNA Res* 2010; **17**: 155–168.

- 25 Deng M, Yu C, Liang Q, He RL, Yau SST. A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS ONE* 2011; **6**: e17293.
- 26 Yu C, Deng M, Cheng SY, Yau SC, He RL, Yau SST. Protein space: a natural method for realizing the nature of protein universe. *J Theor Biol* 2013; **318**: 197–204.
- 27 Yu C, He RL, Yau SST. Protein sequence comparison based on *K*-string dictionary. *Gene* 2013; **529**: 250–256.
- 28 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.
- 29 Wong ML, Arcos-Burgos M, Liu S, Velez JI, Yu C, Baune BT *et al*. The PHF21B gene is associated with major depression, and modulates stress response. *Mol Psychiatry*; e-pub ahead of print 25 October 2016; doi: 10.1038/mp.2016.174.
- 30 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.
- 31 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.
- 32 Wong ML, Dong C, Flores DL, Ehrhart-Bornstein M, Bornstein S, Arcos-Burgos M *et al*. 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.
- 33 Segura V, Vilhjalmsón BJ, Platt A, Korte A, Seren U, Long Q *et al*. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* 2012; **44**: 825–830.
- 34 Nei M, Tajima F, Tateno Y. Accuracy of estimated phylogenetic trees from molecular data. *J Mol Evol* 1983; **19**: 153–170.
- 35 Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW. An evaluation of genetic distances for use with microsatellite loci. *Genetics* 1995; **139**: 463–471.
- 36 Hamming RW. Error detecting and error correcting codes. *Bell Syst Tech J* 1950; **29**: 147–160.
- 37 Torgerson WS. Multidimensional scaling: I. theory and method. *Psychometrika* 1952; **17**: 401–419.
- 38 Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987; **4**: 406–425.
- 39 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.
- 40 Mihaescu R, Levy D, Pachter L. Why neighbor-joining works. *Algorithmica* 2009; **54**: 1–24.
- 41 dos Reis M, Donoghue PC, Yang Z. Bayesian molecular clock dating of species divergences in the genomics era. *Nat Rev Genet* 2016; **17**: 71–80.
- 42 Gascuel O, Steel M. Neighbor-joining revealed. *Mol Biol Evol* 2006; **23**: 1997–2000.
- 43 Wang C, Kao W-H, Hsiao CK. Using Hamming distance as information for SNP-sets clustering and testing in disease association studies. *PLoS ONE* 2015; **10**: e0135918.
- 44 Imai A, Nakaya A, Fahiminiya S, Tetreault M, Majewski J, Sakata Y *et al*. Beyond homozygosity mapping: family-control analysis based on Hamming distance for prioritizing variants in exome sequencing. *Sci Rep* 2015; **5**: 12028.
- 45 Johnson NA, Coram MA, Shriver MD, Romieu I, Barsh GS, London SJ *et al*. Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genet* 2011; **7**: e1002410.
- 46 International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–58.
- 47 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56–65.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>

© The Author(s) 2017

Supplementary Information accompanies the paper on the Translational Psychiatry website (<http://www.nature.com/tp>)