

PUBLISHED VERSION

Chenglong Yu, Bernhard T. Baune, Julio Licinio and Ma-Li Wong

A novel strategy for clustering major depression individuals using whole-genome sequencing variant data

Scientific Reports, 2017; 7:44389-1-44389-7

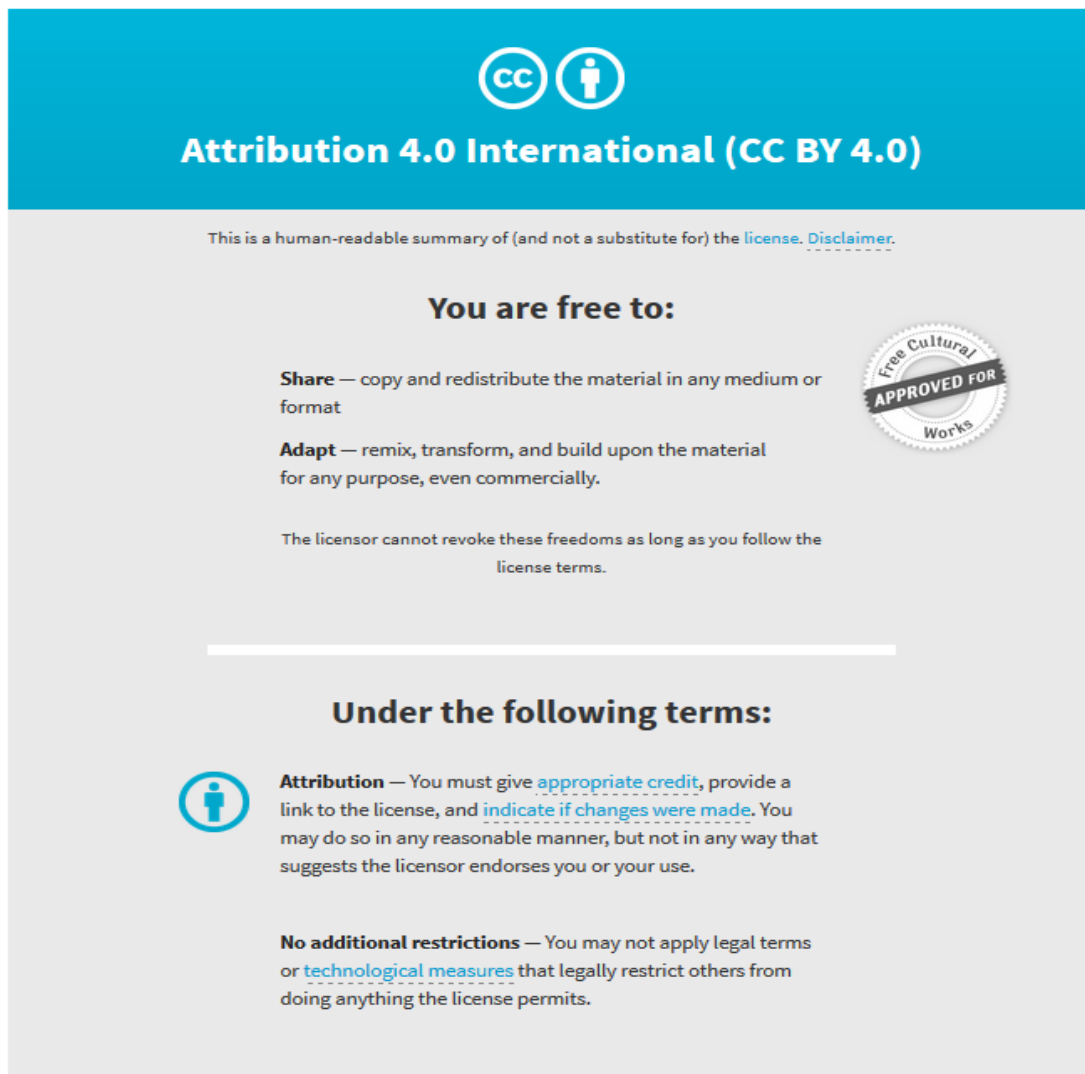
© The Author(s) 2017. This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under 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/4.0/>

Originally published at:

<http://doi.org/10.1038/srep44389>

PERMISSIONS

<http://creativecommons.org/licenses/by/4.0/>



The image shows a Creative Commons Attribution 4.0 International License graphic. It features a blue header with the CC and BY icons and the text "Attribution 4.0 International (CC BY 4.0)". Below the header, there is a disclaimer: "This is a human-readable summary of (and not a substitute for) the license. Disclaimer." The main content is divided into two sections: "You are free to:" and "Under the following terms:". Under "You are free to:", there are three bullet points: "Share" (copy and redistribute), "Adapt" (remix, transform, and build upon), and a note that the licensor cannot revoke these freedoms as long as the license terms are followed. A circular seal on the right says "Free Cultural APPROVED FOR Works". Under "Under the following terms:", there are two bullet points: "Attribution" (you must give appropriate credit, provide a link to the license, and indicate if changes were made) and "No additional restrictions" (you may not apply legal terms or technological measures that legally restrict others from doing anything the license permits).

28 June 2017

<http://hdl.handle.net/2440/105224>

SCIENTIFIC REPORTS



OPEN

A novel strategy for clustering major depression individuals using whole-genome sequencing variant data

Chenglong Yu^{1,2}, Bernhard T. Baune³, Julio Licinio^{1,2} & Ma-Li Wong^{1,2}

Major depressive disorder (MDD) is highly prevalent, resulting in an exceedingly high disease burden. The identification of generic risk factors could lead to advance prevention and therapeutics. Current approaches examine genotyping data to identify specific variations between cases and controls. Compared to genotyping, whole-genome sequencing (WGS) allows for the detection of private mutations. In this proof-of-concept study, we establish a conceptually novel computational approach that clusters subjects based on the entirety of their WGS. Those clusters predicted MDD diagnosis. This strategy yielded encouraging results, showing that depressed Mexican-American participants were grouped closer; in contrast ethnically-matched controls grouped away from MDD patients. This implies that within the same ancestry, the WGS data of an individual can be used to check whether this individual is within or closer to MDD subjects or to controls. We propose a novel strategy to apply WGS data to clinical medicine by facilitating diagnosis through genetic clustering. Further studies utilising our method should examine larger WGS datasets on other ethnical groups.

With the development of new and cheaper whole genome sequencing (WGS) technology, patient care may move towards precision medicine. Ever since the first human genome was fully sequenced, scientists have been searching for approaches to provide personalized care¹. WGS allows us to identify single nucleotide variants (SNVs), which are private genetic variants, and determine all the genetic variants within each person. Single nucleotide polymorphism (SNP) genotyping is currently the gold-standard technique for genome-wide association studies (GWAS), as WGS costs remain relatively high; however, as WGS costs are projected to drop further, researchers may have the opportunity to examine the significance of SNVs, which involve more individual characteristics.

Major depressive disorder (MDD) is a chronic condition with great medical, social, and economic impacts. MDD is a main contributor to global disease burden and produces significant morbidity and mortality²⁻⁶. Despite recent advances⁷⁻⁹, little is known about its underlying fundamental biology. The existing psychiatric genetic studies have not found common consistently replicated gene variants of large effect in the pathogenesis of MDD¹⁰⁻¹², and thus much work still needs to be done to fully elucidate the genetic factors that confer susceptibility to this condition. For our current research, we tested whether the combined effect of all SNVs at the whole-genome sequence level could confer genetic liability to the MDD risk.

In this study, we focus on a sample of Los Angeles Mexican-American participants who had three or more grandparents born in Mexico. MDD participants were diagnosed using the Structured Clinical Interview (SCID) for Diagnostic and Statistical Manual of Mental Disorders (DSM), and the DSM-IV 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 a pharmacogenetic study of antidepressant treatment. 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 Angeles¹³⁻¹⁶. Here, we establish a new computational approach to cluster subjects based on all of their WGS variants. We believe that clustering of patients based on their SNV profiles may provide

¹Mind and Brain Theme, South Australian Health and Medical Research Institute, North Terrace, Adelaide, SA 5000, Australia. ²School of Medicine, Flinders University, Bedford Park, SA 5042, Australia. ³Discipline of Psychiatry, School of Medicine, University of Adelaide, Adelaide, SA 5005, Australia. Correspondence and requests for materials should be addressed to C.Y. (email: chenglong.yu@sahmri.com)

Subjects	Gender	Age	Total SNVs	Total INDELs	dbSNP
MA-Depression-1	Female	35	8,348,095	522,994	4,031,869
MA-Depression-2	Female	30	7,921,961	513,462	3,993,392
MA-Depression-3	Female	41	8,037,674	514,135	3,986,882
MA-Depression-4	Female	32	8,021,058	511,756	3,903,495
MA-Depression-5	Female	45	7,839,942	511,053	4,001,897
MA-Depression-6	Female	38	7,834,986	516,002	4,021,724
MA-Depression-7	Female	36	7,935,708	512,681	3,911,549
MA-Depression-8	Female	59	7,694,178	514,095	3,949,370
MA-Depression-9	Female	41	7,778,564	520,337	3,987,191
MA-Depression-10	Female	31	8,073,958	526,792	4,045,542
MA-Control-1	Female	50	7,879,192	519,009	4,042,162
MA-Control-2	Female	45	6,974,138	517,756	4,021,858
MA-Control-3	Female	39	6,911,665	526,897	3,999,059
MA-Control-4	Female	29	7,197,066	518,675	4,011,644
MA-Control-5	Female	35	7,487,135	517,667	4,031,216
AU-Depression-1	Male	44	3,883,255	555,785	3,888,831
AU-Depression-2	Female	19	3,938,868	541,109	3,956,682
AU-Depression-3	Female	19	3,925,906	560,127	3,928,997
AU-Depression-4	Female	25	3,933,654	557,712	3,935,804
AU-Depression-5	Female	18	3,905,386	555,542	3,920,378
AU-Control-1	Female	20	3,898,847	569,129	3,923,876
AU-Control-2	Male	18	3,920,681	558,496	3,903,217
AU-Control-3	Male	30	3,861,132	552,110	3,861,584
AU-Control-4	Female	18	3,922,531	568,501	3,911,346
AU-Control-5	Male	20	3,820,520	449,055	3,773,974

Table 1. Whole-genome sequencing variation analysis of 25 human subjects. MA, Mexican-American; AU, Australian; SNVs, single nucleotide variants; INDELs, small insertions and deletions; dbSNP (the number of SNVs and INDELs that are found in the dbSNP database in NCBI).

valuable clues for prognostics, diagnostics, and therapeutics, as it takes into account all of their genetic data. The idea for this approach arose from distance-based phylogenetic analysis of DNA/protein sequences proposed by us earlier^{17–21}. In our proposed methodology we used a well-defined metric in mathematics, the Jaccard distance, to measure the similarity/dissimilarity between subjects using all the SNV information from each individual and from that we construct cluster trees based on the Jaccard distance matrices. Clustering relationships in the trees showed that Mexican-American MDD patients grouped together, and were clustered far from ethnically matched healthy controls. This discovery may be translated to clinical practice since we may be able to predict the MDD status of a given Mexican-American subject based on his/her WGS data.

Materials and Methods

The Mexican-American Sample. In our recent work¹⁶, we have investigated the whole-exome genotyping data of a Los Angeles Mexican-American cohort aged 19–65 years of 203 MDD patients and 196 healthy controls. Participants provided written informed consent, and detailed demographic, epidemiological, and clinical descriptions were previously described^{13–15}. 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^{13–15}. In this study, we obtained complete WGS data for a group of 15 participants selected from the cohort, 10 MDD patients and 5 controls. In Table 1, we present the gender (all are female) and age information of the 15 Mexican-American subjects. We have confirmed that in the cohort there was no family or population structure among all those individuals¹⁶ and no any blood relationship among the 15 selected participants.

The European-Ancestry Sample. For comparison as an outgroup sample, we also include WGS data from a group of 10 Australians of European-Ancestry. Those 10 participants gave written informed consent and were recruited under the Cognitive function and mood disorders study (conducted by the Discipline of Psychiatry, University of Adelaide, South Australia, Australia). This sample was studied under approved Human Research Ethics Committees protocols at the University of Adelaide and Flinders University, South Australia, Australia. In Table 1, we present the gender and age information of these 10 subjects.

We confirm all methods and experiments in this study were performed in accordance with relevant guidelines and regulations.

Whole-Genome Sequencing (WGS) and Analysis. Samples from fifteen Mexican-American participants (10 MDD patients and 5 controls) were whole-genome sequenced using Illumina HiSeq 2000 (BGI-Shenzhen, Shenzhen, Guangdong, China) and samples from ten European-Ancestry Australian participants

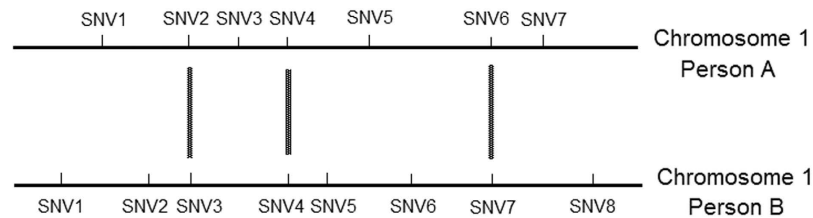


Figure 1. A hypothetical illustration of the distribution of SNVs on chromosome (chr) 1 of two individuals. SNV2, SNV4 and SNV6 in person A and SNV3, SNV4 and SNV7 in person B occupy the same respective positions in chr 1.

(5 MDD patients and 5 controls) were whole-genome sequenced using HiSeq X (Garvan Institute, Sydney, New South Wales, Australia). After obtaining paired-end sequencing reads of those 25 participants, we did the variant calling analysis using the following pipeline. The reads of each participant were aligned to the human reference genome (hg19, Genome Reference Consortium GRCh37) using Burrows-Wheeler Aligner (BWA)²² to get SAM (sequence alignment/map) format files. SAM files were converted to the BAM (binary version of a SAM file) format files using SAMtools²³. BAM files were then merged into one BAM file, and the mpileup command in SAMtools was used to calculate the likelihood of data given each possible genotype, and store the likelihoods in a binary file. The output was supplied to SAMtools/BCftools²⁴ which created the SNV/INDEL (small insertions and deletions) calling to generate VCF (variant call format) files. Then, ANNOVAR²⁵ was used to annotate SNV/INDEL information and their classifications. For WGS and analysis details, please see Supplementary Information. Only the SNV information was used in the following methodology.

Clustering Subjects on SNV Sets. To take in consideration all the SNV information of those subjects, it was important to define a distance between two subjects when running the cluster analysis. We clustered the subjects at the chromosome level; consequently, we defined a distance between two people based on SNV information obtained from a given chromosome, e.g., chromosome 1. First, we considered the SNVs distribution on that chromosome. In Fig. 1, we give a hypothetical SNVs distribution on chromosome 1 for two individuals. In a real case scenario, in a given chromosome two individuals may have many same position SNVs, e.g., SNV6 in person A and SNV7 in person B in Fig. 1. Our hypothesis was that if two individuals shared more same position SNVs, then those two individuals would have more similar phenotypes, such as traits or diseases. Therefore, a proper distance (dis-similarity) between two SNV sets could be employed.

Let S_1 and S_2 be SNV sets in a given chromosome from subject 1 and subject 2. We use $|S|$ to denote the cardinality of set S . The Jaccard metric²⁶, a statistics tool for measuring the similarity and diversity of sample sets, is introduced here. The Jaccard metric between S_1 and S_2 is defined as

$$J(S_1, S_2) = 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} = 1 - \frac{|S_1 \cap S_2|}{|S_1| + |S_2| - |S_1 \cap S_2|} \quad (1)$$

If S_1 and S_2 are exactly the same, then $S_1 \cap S_2 = S_1 \cup S_2$, $J(S_1, S_2) = 0$. If S_1 and S_2 have no common elements, then $S_1 \cap S_2 = \phi$, $J(S_1, S_2) = 1$. Unlike many distances used in comparative genomics (e.g., some distances based on alignment methods), the Jaccard metric satisfies the triangle inequality, i.e., $J(S_1, S_3) \leq J(S_1, S_2) + J(S_2, S_3)$, which guarantees that the Jaccard distance is a well-defined metric in mathematics²⁷. Considering Fig. 1 as an example, $S_A = 7$, $S_B = 8$, $S_A \cap S_B = 3$, so

$$J(S_A, S_B) = 1 - \frac{|S_A \cap S_B|}{|S_A| + |S_B| - |S_A \cap S_B|} = 1 - \frac{3}{7 + 8 - 3} = 1 - 0.25 = 0.75 \quad (2)$$

In this proof-of-principle study we use the Jaccard metric to calculate the distance matrices of those 25 participants in each chromosome. Then cluster trees based on each chromosome can be reconstructed. Clustering relationships shown in the trees may reveal significant medical information that may be translated into clinical practice.

Results

Whole-Genome Sequencing (WGS) Data Analysis. Table 1 provides the results of WGS variation in 25 human subjects and shows that Mexican-American individuals have significantly more SNVs when compared to Australian individuals of European-Ancestry. For total SNVs, Australian's mean value is 3901078 ($n = 10$), Mexican-American's mean value is 7729021.3 ($n = 15$), the t test p -value for the two groups is $2.09e-15$. This is consistent with data from the Human Haplotype Matching Project (HapMap). We contributed the Mexican-American sample to HapMap, from the same community as subjects in this study. That study showed that Mexican-Americans have more polymorphic SNPs in Mexican-Americans than in northern Europeans²⁸. Mexican-Americans from that Los Angeles community have median ancestry proportions that are 45% Indigenous American, 49% European and 5% African²⁹. According to results from the International HapMap 3 Consortium and the 1000 Genomes Project Consortium, it would be expected that individuals with African

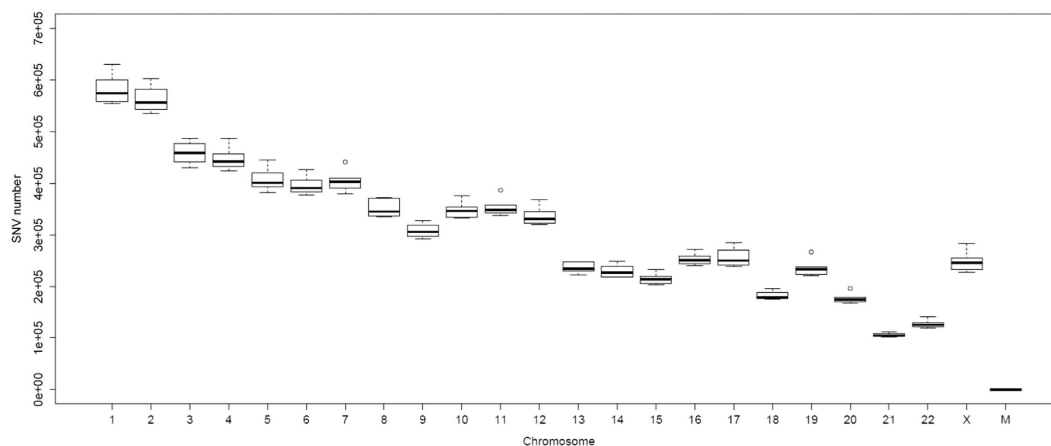
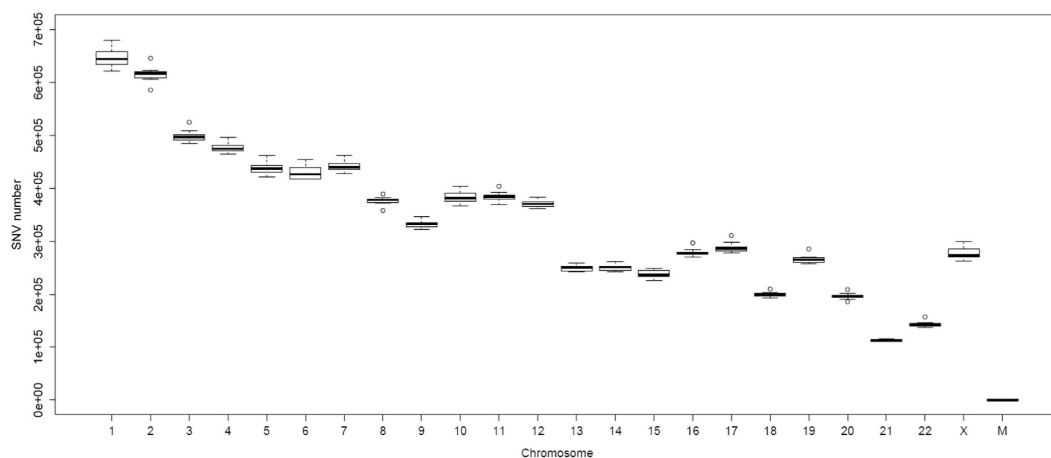
a**b**

Figure 2. Descriptive statistical distributions of SNVs on all the chromosomes. (a) The Mexican-American control group. **(b)** The Mexican-American MDD group.

ancestry, such as Mexican-Americans, have increased number of variants, and, moreover, the Spanish population have excess of rare variants^{28,30}.

For the Mexican-American sample, both depression and control subjects have approximately 7,000,000 to 8,000,000 SNVs; 5,100,000 to 5,200,000 INDELS, and 3,900,000 to 4,000,000 SNVs in dbSNP (the SNP database). We calculated the SNV distributions on each chromosome for the Mexican-American and Australian samples. In Fig. 2a, we used boxplot to show the descriptive statistical distributions of SNVs in each chromosome for the Mexican-American control group. Descriptive statistical distributions of SNVs of each chromosome for the Mexican-American depression group are provided in Fig. 2b. Since only female Mexican-American samples were used for this study, we include chromosome X in the results. We found that the depression and control groups have basically the same SNV distributions for all chromosomes. Table S1 provides detailed information of SNV distributions for all the chromosomes in the 25 subjects.

Clustering Subjects using Cluster Trees. Following the proposed method, we use the Jaccard metric and SNV sets to obtain the distance matrices between those 25 participants for each chromosome. Jaccard distance calculation was done using R programming language. We used the popular neighbor-joining method³¹ on the distance matrices to construct cluster trees, which were drawn using software MEGA 6³². Figure 3a shows the cluster tree for 25 subjects in chromosome 1. We found that all the 10 Mexican-American MDD patients grouped together in a cluster, and 5 Mexican-American controls 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. This fact is also consistent with the genetic distance between different populations³³.

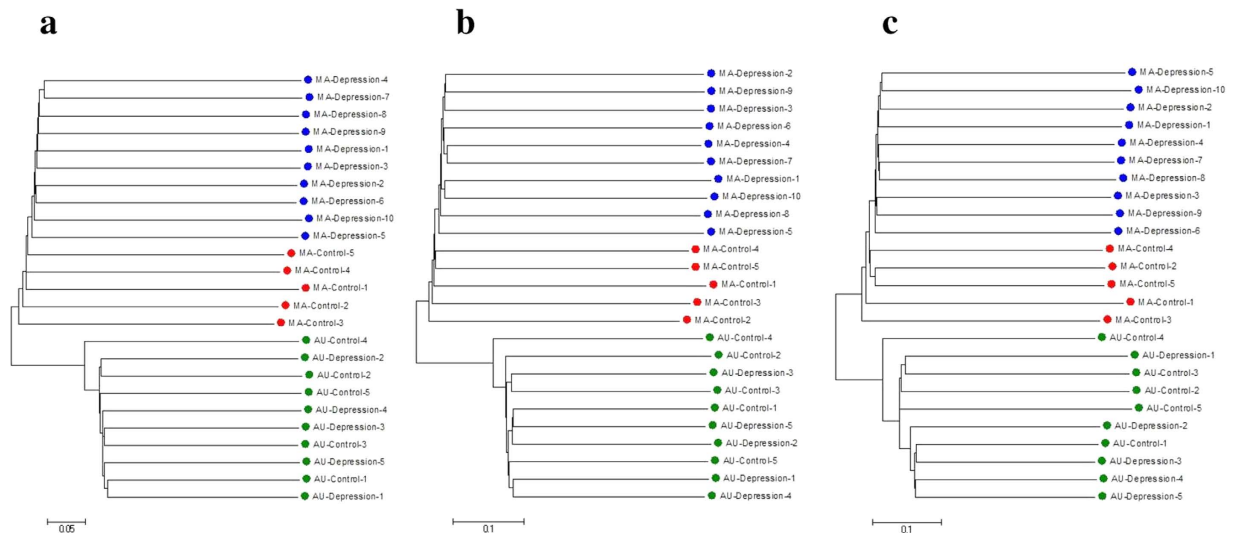


Figure 3. Cluster tree of 25 subjects for different chromosomes. (a) On chromosome 1. (b) On chromosome 22. (c) On chromosome X.

We constructed cluster trees for all chromosomes. Except for the mitochondrial genome, all cluster trees clustered the Mexican-American MDD patients as group distinct from the controls. Although the Australian subjects stably stand as an outgroup, within that group the MDD and control individuals could not be well distinguished as in the Mexican-American group. Figure 3b and c show the cluster trees in chromosome 22 and chromosome X, respectively. In Figure S1, we provided the cluster trees for all the other 20 chromosomes and mitochondrial genome.

WGS data analysis and Jaccard distance calculations were performed using high-performance computers in eResearch South Australia (<https://www.ersa.edu.au/>).

Discussion

The results obtained by our new approach support the assumption that two individuals who share more of the same position SNVs would have more similar phenotypes, such as traits or diseases. Clustering relationships in the trees show that the Mexican-American MDD patients group together, and ethnically matched controls grouped away and separately. The fact that Australian subjects fail to be clustered into case and control groups may imply that this computational method may be restricted to specific populations, with a higher degree of genetic diversity, such as Mexican-Americans. It should be noted that the choice of Jaccard metric was not random. When measuring similarity between two SNV sets, the intersection of two sets denotes the shared same position SNVs of two people, and the union of two sets is used to normalize the similarity to a value between 0 and 1. All the SNV information for two sets is fully utilized in this metric. Furthermore, the Jaccard metric is a rigorous mathematical distance. Our results showed that it is appropriate to cluster Mexican-American MDD subjects in this study. Among distance-based tree construction methods, the neighbor-joining technique does not assume a constant rate of evolution, as opposed to the molecular clock hypothesis. Due to its low computational complexity it can be performed quickly and is widely used to generate cluster trees of individuals^{34,35}.

We have confirmed that there were no blood relatives between those Mexican-American subjects, thus the clustering relationships in the trees were not due to genetic relatedness. For the Mexican-American sample, all the subjects were female, and the MDD case group had an average age of 38.8 years with standard deviation 8.15 and the control group had an average age of 39.6 years with standard deviation 7.36. The two groups have basically the same age distribution. Thus the clustering results were not associated with gender and age. For our approach, confounding phenotypes with complex genetic architecture may be reflected in the measured distance and this could alter the observed clustering. Therefore, before performing our method, it is necessary to control confounding factors such as ethnicity, MDD diagnostic and control selection criteria, genetic relatedness, gender and age.

Our aim in this paper was not to confirm or refute previous genetic research of depression such as candidate gene studies or GWAS³⁶ but rather to bring a novel direction using comparative genomic analysis at the whole-genome sequence level. In our methodology, the combined effect of all SNVs in the complete genome, including all genomic regions such as coding and non-coding, was considered as a genetic factor to the depression risk. Our computational approach allowed us to perform a global comparison of whole-genome information in the subjects, which no other existing method can achieve. Once a Jaccard distance matrix has been constructed, the results in the clustering tree can be displayed and viewed graphically; this is user-friendly and allows even non-expert to understand the relationships among the subjects. Furthermore, most existing genome-wide analysis methods involve many statistical models. The different choices of these models can lead to inconsistent results. Our method does not involve any statistical model and it depends only on the genetic distance between two individuals by considering their whole-genome SNV information. Therefore, our approach is stable and produces a unique analysis result.

High quality full genome sequencing costs are currently still a concern that limits obtaining larger datasets; another limitation is the high level of computational resources needed for sequencing data analysis. Future studies utilising our method should examine additional replication data on other ethnical groups.

We have developed a novel methodology to cluster subjects based on their WGS data. To the best of our knowledge, this is the first time that SNV and cluster analysis are used to study major depressive disorder. Our approach could be a useful predictive/diagnostic tool; i.e., one could test whether WGS data from a new subject could contribute to determine whether that subject would be within or close to an existing MDD or control cluster. Advances in this line of research have the potential to be rapidly translated to clinical practice and could include the ability to diagnose patients based on WGS data.

References

- Collins, F. S. & McKusick, V. A. Implications of the Human Genome Project for medical science. *JAMA* **285**, 540–544 (2001).
- 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).
- Lopez, A. D. & Murray, C. C. The global burden of disease, 1990–2020. *Nat. Med.* **4**, 1241–1243 (1998).
- Wong, M. L. & Licinio J. Research and treatment approaches to depression. *Nat. Rev. Neurosci.* **2**, 343–351 (2001).
- 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).
- 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).
- Hyde, C. L. *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Genet.* **48**(9), 1031–1036. (2016).
- 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*, doi: 10.1038/mp.2016.101 (2016).
- CONVERGE consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**(7562), 588–591 (2015).
- 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).
- Lopez-Leon, S. *et al.* Meta-analyses of genetic studies on major depressive disorder. *Mol. Psychiatry* **13**(8), 772–785 (2008).
- Lohoff, F. W. Overview of the genetics of major depressive disorder. *Curr. Psychiatry Rep.* **12**(6), 539–546 (2010).
- 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).
- 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).
- 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).
- Wong, M. L. *et al.* The PHF21B gene is associated with major depression and modulates the stress response. *Mol. Psychiatry* doi: 10.1038/mp.2016.174 (2016).
- Yu, C., Liang, Q., Yin, C., He, R. L. & Yau, S. S. T. A novel construction of genome space with biological geometry. *DNA Res.* **17**, 155–168 (2010).
- Deng, M., Yu, C., Liang, Q., He, R. L. & Yau, S. S. T. A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS One* **6**(3), e17293 (2011).
- Yu, C. *et al.* Protein space: a natural method for realizing the nature of protein universe. *J. Theor. Biol.* **318**, 197–204 (2013).
- Yu, C., He, R. L. & Yau, S. S. T. Protein sequence comparison based on *K*-string dictionary. *Gene* **529**(2), 250–256 (2013).
- Hoang, T. *et al.* A new method to cluster DNA sequences using Fourier power spectrum. *J. Theor. Biol.* **372**, 135–145 (2015).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16), 2078–2079 (2009).
- Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
- Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
- Levandowsky, M. & Winter, D. Distance between sets. *Nature* **234**(5323), 34–35 (1971).
- Lipkus, A. H. A proof of the triangle inequality for the Tanimoto distance. *J. Math. Chem.* **26**(1–3), 263–265 (1999).
- International HapMap 3 Consortium. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**(7311), 52–58 (2010).
- Johnson, N. A. *et al.* Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genet.* **7**(12), e1002410 (2011).
- 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422), 56–65 (2012).
- Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**(4), 406–425 (1987).
- Tamura, K., Stecher, G., Peterson, D., Filipowski, A. & Kumar, S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**(12), 2725–2729 (2013).
- Nei, M. Genetic distance between populations. *Am. Nat.* **106**, 283–292 (1972).
- Mihaescu, R., Levy, D. & Pachter, L. Why neighbor-joining works. *Algorithmica* **54**(1), 1–24 (2009).
- Gascuel, O. & Steel, M. Neighbor-joining revealed. *Mol. Biol. Evol.* **23**(11), 1997–2000 (2006).
- Levinson, D. F. *et al.* Genetic studies of major depressive disorder: Why are there no GWAS findings, and what can we do about it? *Biol. Psychiatry* **76**(7), 510 (2014).

Acknowledgements

The authors have been supported by grants APP1051931 and APP1070935 (M.L.W.), and APP1060524 (B.T.B.) from the National Health and Medical Research Council (Australia), NIH grant GM61394 (J.L. and M.L.W.), and institutional funds from the South Australian Health and Medical Research Institute.

Author Contributions

C.Y. conceived of the study and performed all data analyses. C.Y., B.T.B. and M.L.W. analyzed and interpreted the results. C.Y., M.L.W. and J.L. wrote the paper. The final version was done by all authors.

Additional Information

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing Interests: The authors declare no competing financial interests.

How to cite this article: Yu, C. *et al.* A novel strategy for clustering major depression individuals using whole-genome sequencing variant data. *Sci. Rep.* 7, 44389; doi: 10.1038/srep44389 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 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/4.0/>

© The Author(s) 2017