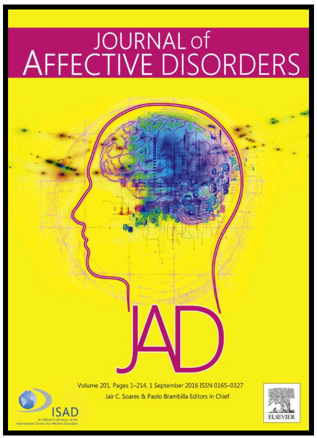

JOURNAL of
AFFECTIVE DISORDERS
JAD
Volume 201, Pages 1–214, 1 September 2016 ISSN 0165-0327
Jair C. Soares & Paolo Brambilla Editors in Chief
ISAD
ELSEVIER

www.elsevier.com/locate/jad

Genetic clustering of depressed patients and normal controls based on single-nucleotide variant proportion

Chenglong Yu, Bernhard T. Baune, Ke-Ang Fu, Ma-Li Wong, Julio Licinio

PII: S0165-0327(17)31162-X
DOI: https://doi.org/10.1016/j.jad.2017.11.023
Reference: JAD9337

To appear in: *Journal of Affective Disorders*

Received date: 7 June 2017
Revised date: 27 September 2017
Accepted date: 6 November 2017

Cite this article as: Chenglong Yu, Bernhard T. Baune, Ke-Ang Fu, Ma-Li Wong and Julio Licinio, Genetic clustering of depressed patients and normal controls based on single-nucleotide variant proportion, *Journal of Affective Disorders,* https://doi.org/10.1016/j.jad.2017.11.023

Genetic clustering of depressed patients and normal controls based on single-nucleotide variant proportion

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

[a]Robinson Research Institute, Adelaide Medical School, University of Adelaide, Adelaide, SA, Australia.

[b]Mind and Brain Theme, South Australian Health and Medical Research Institute, Adelaide, SA, Australia.

[c]College of Medicine and Public Health, Flinders University, Bedford Park, SA, Australia.

[d]Discipline of Psychiatry, Adelaide Medical School, University of Adelaide, Adelaide, SA, Australia.

[e]School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, Zhejiang, China.

[f]College of Medicine, Departments of Psychiatry, Pharmacology and Medicine, State University of New York, Upstate Medical University, Syracuse, NY, USA.

*Correspondence to: Level 8, Adelaide Health & Medical Sciences Building, North Terrace, Adelaide SA 5000, Australia. Phone: +61 8 83136343. chenglong.yu@adelaide.edu.au (C. Yu).

Abstract

Background

Genetic components play important roles in the susceptibility to major depressive disorder (MDD). The rapid development of sequencing technologies is allowing scientists to contribute new ideas for personalized medicine; thus, it is essential to design non-invasive genetic tests on sequencing data, which can help physicians diagnose and differentiate depressed patients and healthy individuals.

Methods

We have recently proposed a genetic concept involving single-nucleotide variant proportion (SNVP) in genes to study MDD. Using this approach, we investigated combinations of distance metrics and hierarchical clustering criteria for genetic clustering of depressed patients and ethnically matched controls.

Results

We analysed clustering results of 25 human subjects based on their SNVPs in 46 newly discovered candidate genes.

Conclusions

According to our findings, we recommend Canberra metric with Ward's method to be used in hierarchical clustering of depressed and normal individuals. Futures studies are needed to advance this line of research validating our approach in larger datasets, those may also be allow the investigation of MDD subtypes.

Limitations

High quality sequencing costs limited our ability to obtain larger datasets.

Abbreviations

MDD, major depressive disorder; SNV, single-nucleotide variant; SNVP, single-nucleotide variant proportion; WGS, whole-genome sequencing; UPGMA, Unweighted Pair Group Method with Arithmetic mean; WPGMA, Weighted Pair Group Method with Arithmetic mean; UPGMC, Unweighted Pair Group Method with Centroid; WPGMC, Weighted Pair Group Method with Centroid.

Keywords: major depressive disorder; sequencing; distance metric; hierarchical clustering; candidate gene; Canberra distance; Ward's method

1. Introduction

Major depressive disorder (MDD) has a lifetime prevalence of 10%-20% in the general population; it produces significant morbidity and mortality, and leads to high suicide rates (Wong and Licinio, 2001). Genetic factors have been proven to play important roles in the development of MDD (Flint and Kendler, 2014). The growth of newer and cheaper high-throughput sequencing technologies has allowed the progress of novel methods towards personalized treatment (Soon et al., 2013). Specifically, whole-genome sequencing can identify most or even all private genetic variations such as single-nucleotide variants (SNVs), small insertions and deletions, and copy number variations (Belkadi et al., 2015). Thus it is desirable to develop a non-invasive genetic test using sequencing data, which can help physicians diagnose and differentiate depressed patients from normal healthy people. Genetic clustering has provided a promising conduit to explore this topic (Yu et al., 2017a).

Recently, we proposed a new genetic concept, single-nucleotide variant proportion (SNVP) in genes, to study MDD based on DNA sequencing data (Yu et al., 2017b)., Multivariate cluster analysis such as hierarchical cluster tree method can be designed to identify depressed individuals and normal controls using SNVPs in a range of candidate genes (Yu et al., 2017b). Since hierarchical clustering is sensitive to both the choice of distance metric technique (e.g., Euclidean distance, Manhattans distance, etc.) and the criterion for determining the order of clusters (e.g., complete linkage, average linkage, etc.), various combinations of those approaches may yield different results. Thus, the distance metric technique and the clustering criterion should be carefully selected.

In this report, we focused on methodological refinement by investigating different combinations of distance metrics and hierarchical clustering criteria for genetic clustering of depressed patients and normal controls. Based on SNVPs in 46 candidate genes associated with major depression, different clustering trees were compared to evaluate the robustness of combination results.

2. Materials and methods

2.1 Subjects

We have obtained complete whole-genome sequencing data of two samples of human participants (Yu et al., 2017b, 2017c). Two samples were respectively recruited from two populations: Mexican-American in Los Angeles, California, USA (Wong et al., 2016) and Australian of European-ancestry in Adelaide, South Australia, Australia (Baune and Air, 2016). The Mexican-American sample included 15 subjects (10 MDD patients and 5 healthy controls, sequenced by Illumina HiSeq 2000 at BGI-Shenzhen, Shenzhen, Guangdong, China), and the Australian sample consisted of 10 subjects (5 MDD patients and 5 healthy controls, sequenced by Illumina HiSeq X at Garvan Institute, Sydney, New South Wales, Australia). All the participants provided written informed consent, and we confirmed that there were no blood relatives among them. The details about MDD and healthy control diagnostic criteria can be found in our previous work (Wong et al., 2016). The study was registered in ClinicalTrials.gov (NCT00265291) and approved by the Institutional Review Boards of the University of California, Los Angeles and the University of Miami in USA, and the Human Research Ethics Committees of the Australian National University, the University of Adelaide, the Flinders University and Bellberry Limited in Australia.

2.2 SNVP in 46 candidate genes

SNVP in a gene was defined as the ratio of the number of SNVs to the number of all nucleotides in the gene sequence (Yu et al., 2017b). In recent work, using genome-wide association study for common mutations (Minor Allele Frequency, MAF $\geq$ 0.01) and rare-variant analysis for rare mutations (MAF < 0.01) on exome genotyping data from a large Mexican-American cohort, we identified 46 genes which may confer susceptibility to MDD, namely, *ALDH3B1, ANKMY2, ANO8, ARHGAP8, BCAR3, C10orf27, C19orf39, C2orf54,* *CACNA1G, CIZ1, CNTD1, CNTNAP1, CRAMP1L, EMR2, FAM69B, FASN, FSCB, GNA15, GRK4, HOMER3, KRBA1, LILRA1, LRRC24, LRWD1, MUC5B, MUC6, MYH13, OR1L4, OR2T12, OR52I1, OR6C4, ORAI1, PHF21B, PLEKHG6, PRR5, RABAC1, SLC2A8, SLX4, TBC1D2B, TMEM150B, TMEM151A, TRIO, TRPM2, TRPV4, UNC13D*, and *VENTX* (Wong et al., 2017). We have performed SNV callings on whole-genome sequencing data analysis and calculated SNVPs in those 46 genes for 25 human subjects (Yu et al., 2017b) using high-performance computers in eResearch South Australia (www.ersa.edu.au).

2.3 Distance metrics

When clustering subjects, we put subjects with similar features into the same cluster and dissimilar subjects into different clusters. The distance metric showing similarity and dissimilarity is significant as it determines how different two subjects are. Mathematically, a distance function $D(X, Y)$ between two N-dimensional numerical vectors $X = (x_1, x_2, ..., x_N)$ and $Y = (y_1, y_2, ..., y_N)$ is said to be metric if it satisfies the following properties:

(i) Non-negativity: $D(X, Y) \geq 0$;

(ii) Identity of indiscernibles: $D(X, Y) = 0$ if and only if $X = Y$;

(iii) Symmetry: $D(X, Y) = D(Y, X)$;

(iv) Triangle inequality: $D(X, Z) \leq D(X, Y) + D(Y, Z)$ for any X, Y and Z.

Here we compared the following distance metrics for clustering. Since SNVP is a ratio number between 0 and 1, we consider $0 < x_i,\ y_i < 1$ for $i = 1, 2, \ldots, N$.

(1) Euclidean distance: $D(X,Y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$.

(2) Manhattan distance: $D(X,Y) = \sum_{i=1}^{N} | x_i - y_i |$.

(3) Canberra distance: $D(X,Y) = \sum_{i=1}^{N} \frac{| x_i - y_i |}{| x_i | + | y_i |}$.

(4) Chebyshev distance: $D(X,Y) = \max_i |x_i - y_i|$.

(5) Minkowski distance: $D(X,Y) = \left(\sum_{i=1}^{N} |x_i - y_i|^P\right)^{1/P}$, for $P \geq 1$. Actually, for $P = 1$, it is Manhattan distance and for $P = 2$, it is Euclidean distance.

(6) NTV distance (Nieto et al., 2003): $D(X,Y) = \dfrac{\sum_{i=1}^{N} |x_i - y_i|}{\sum_{i=1}^{N} \max(|x_i|,\ |y_i|)}$. The triangle inequality for this metric has also been verified by a simple proof (Dress and Lokot, 2003).

In this study, we use $D(X, Y)$ to represent the genetic distance between two human subjects X and Y. Here each subject is represented by an N-dimensional numerical vector, i.e., a set of SNVP values for N candidate genes ($N = 46$ in work presented here).

2.4 Hierarchical clustering criteria

The hierarchical clustering results are commonly displayed in a dendrogram, and there are many different linkage criteria for this aim. In this study we compared the following agglomerative algorithms: single linkage (nearest neighbour method), complete linkage (farthest neighbour method), average linkage (Unweighted Pair Group Method with Arithmetic mean - UPGMA), McQuitty's linkage (Weighted Pair Group Method with Arithmetic mean - WPGMA), centroid linkage (Unweighted Pair Group Method with Centroid - UPGMC), median linkage (Weighted Pair Group Method with Centroid - WPGMC) and Ward's method (Arabie et al., 1996). We also used bootstrap resampling techniques to assess hierarchical clustering uncertainty. By using Pvclust package (Suzuki and Shimodaira, 2006) two types of probability values, approximately unbiased (au) probability value and bootstrap probability (bp) value, were calculated and shown in the dendrogram. All statistical calculations and dendrogram graphs in this study were performed using the R software (www.r-project.org).

We tested different combinations of eight distance metric techniques and seven hierarchical cluster analysis criteria to cluster those 25 subjects based on SNVPs in the 46 candidate genes. For the Minkowski metric, we examined P = 1.25, 1.5, and 1.75. The expected clustering result was supposed to firstly separate the two population groups (Mexican-American and Australian) and then separate two subgroups (depressed and normal control) within each population sample.

3. Results

In Table S1, we present descriptive statistics of gender, age and HAM-D (Hamilton depression) scores for the 15 Mexican-American and the 10 Australian subjects whose whole genome data were analysed in this study. The Australian group included both acutely depressed and remitted depressed patients who had by definition a lower depression severity compared to acutely depressed patients; thus, the HAM-D scores for Australian sample may be lower on average than the Mexican-American sample.

After checking the dendrogram results of all combinations, we found that the Ward's method as a hierarchical clustering criterion generated the best outcomes for all the distance metrics except the Chebyshev metric (see Figure 1). Clustering relationships in the trees showed that the two populations were well separated. Furthermore, within the Mexican-American group, MDD patients clustered together and ethnically matched controls clustered away and separately. However, although the Australian individuals stably stand as a separated group, within that group the depressed and control subjects could not be well distinguished for all the combinations of distance metrics and hierarchical cluster analysis criteria. The clustering results on other hierarchical clustering criteria were presented in Figures S1-S6.

We also examined all the distance metrics and found that the Canberra metric produced the best results for all the hierarchical clustering criteria except for the single linkage algorithm (see Figure 1 and Figures S1-S6). Actually, single linkage method could not perform well for all the distance metrics. To assess the uncertainty in hierarchical clustering by the Canberra metric and the Ward's method, we used the Pvclust package to perform 5,000 iterations bootstrapping in order to construct the cluster dendrogram with au/bp values (%) as shown in Figure 2. The au values for two population clusters are 100 and 100, and for the two subgroups within Mexican-American cluster are 82 and 80. The au values computed by multiscale bootstrap resampling better approximate to unbiased probability values. Au value of a cluster indicates how strongly this cluster is supported by data. Our result implies very strong clusters for the two populations (100% and 100%), and the two subgroups (82% and 80%) within the Mexican-American sample. Therefore, based on the SNVP data, we recommend that the Canberra metric with Ward's hierarchical clustering method be used in genetic cluster analysis of depressed individuals and normal controls.

4. Discussion

In this study, we tested different combinations of distance metrics and hierarchical clustering criteria based on SNVP. The Canberra metric with Ward's criterion clearly surpassed the other combinations. Actually, the Canberra metric has attracted attention from computational geneticists and has been used in checking ranked lists of molecular biomarkers (Jurman et al., 2008). The Ward's criterion, which focuses on minimizing the total within-cluster variance, may also reveal information on genetic variation between/within MDD and control clusters. Further investigation on this combination should examine larger genetic data sets and additional replication samples from other ethnic groups.

Our approach could cluster the 15 Mexican-American subjects into two groups in the hierarchical cluster tree: major depression and normal control. We could potentially determine how close new Mexican-American participants were to the existing depressed or control group using genetic clustering. Subjects within or close to the depression group in the cluster tree could be predicted to be individuals with depression.

Cluster analysis based on SNVP not only could derive a predictive/diagnostic tool, as one can test whether a new subject falls within or close to an existing diseased cluster, but may also provide an alternative way for determining MDD subtypes. Major depression as a clinically heterogeneous illness has been classified based on distinct clinical features that include course, periodicity, qualitative and quantitative types of symptoms, age or phase of life, and cause (van Loo et al., 2012). Since different subtypes of MDD may respond differentially to various medications, there has been considerable interest in studying classification systems and subgroupings of depressed patients (Hybels et al., 2013; Ulbricht et al., 2015). Clinical data-driven subtypes of MDD remain largely controversial due to the heterogeneity of this disorder (Harald and Gordon, 2012; Bosaipo et al., 2016); thus, our methodology of genetic clustering on sequencing data would bring a new direction to this field. However, high quality deep sequencing costs are currently still a concern that limits obtaining larger datasets. Furthermore, the fact that Australian subjects fail cluster into case and control subgroups may imply that our current computational strategy may be restricted to specific populations, with a higher degree of genetic diversity, such as Mexican-Americans (International HapMap 3 Consortium, 2010). Thus future studies on larger genetic data of other ethnical groups will be needed to test the robustness of our method.

Author Disclosure Statement

Contributors

Chenglong Yu, Mind and Brain Theme, South Australian Health and Medical Research Institute, North Terrace, Adelaide, SA 5000, Australia and School of Medicine, Flinders University, Bedford Park, SA 5042, Australia.

Bernhard T. Baune, Discipline of Psychiatry, School of Medicine, University of Adelaide, Adelaide, SA 5005, Australia.

Ke-Ang Fu, School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, Zhejiang 310018, China.

Ma-Li Wong, Mind and Brain Theme, South Australian Health and Medical Research Institute, North Terrace, Adelaide, SA 5000, Australia and School of Medicine, Flinders University, Bedford Park, SA 5042, Australia.

Julio Licinio, Mind and Brain Theme, South Australian Health and Medical Research Institute, North Terrace, Adelaide, SA 5000, Australia and School of Medicine, Flinders University, Bedford Park, SA 5042, Australia.

Funding

The authors have been supported by grants NHMRC APP1051931 (Ma-Li Wong), NHMRC APP1060524 (Bernhard T. Baune), and NIH grant GM61394 (Julio Licinio and Ma-Li Wong), and institutional funds from the South Australian Health and Medical Research Institute.

Role of the funding source

The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Conflict of interest

The authors declare no conflict of interest.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version of this journal.

References

Arabie, P., Hubert, L.J., De Soete, G., 1996. Clustering and classification. World Scientific Publishing, Singapore.

Baune, B.T., Air, T., 2016. 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.

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.L., Abel, L., 2015. Whole-genome sequencing is more powerful than whole- exome sequencing for detecting exome variants. Proc. Natl. Acad. Sci. U.S.A. 112(17), 5473-5478.

Bosaipo, N.B., Foss, M.P., Young, A.H., Juruena, M.F. 2016. Neuropsychological changes in melancholic and atypical depression: a systematic review. Neurosci. Biobehav. Rev. 2016; 73: 309-325.

Dress, A., Lokot, T., 2003. A simple proof of the triangle inequality for the NTV metric. Appl. Math. Lett. 16(6), 809-813.

Flint, J., Kendler, K.S. 2014. The genetics of major depression. Neuron 81(3), 484-503.

Harald, B., Gordon, P., 2012. Meta-review of depressive subtyping models. J. Affect. Disord. 139(2), 126-140.

Hybels, C.F., Landerman, L.R., Blazer, D.G., 2013. Latent subtypes of depression in a community sample of older adults: can depression clusters predict future depression trajectories? J. Psychiatr. Res. 47(10), 1288-1297.

International HapMap 3 Consortium, 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467(7311), 52-58.

Jurman, G., Merler, S., Barla, A., Paoli, S., Galea, A., Furlanello, C., 2008. Algebraic stability indicators for ranked lists in molecular profiling. Bioinformatics 24(2), 258-264.

Nieto, J.J., Torres, A., Vázquez-Trasande, M.M., 2003. A metric space to study differences between polynucleotides. Appl. Math. Lett. 16(8), 1289-1294.

Soon, W.W., Hariharan, M., Snyder, M.P., High-throughput sequencing for biology and medicine. Mol. Syst. Biol. 9(1), 640.

Suzuki, R., Shimodaira, H., 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22(12), 1540-1542.

Ulbricht, C.M., Rothschild, A.J., Lapane, K.L., 2015. The association between latent depression subtypes and remission after treatment with citalopram: a latent class analysis with distal outcome. J. Affect. Disord. 188, 270-277.

van Loo, H.M., De Jonge, P., Romeijn, J.W., Kessler, R.C., Schoevers, R.A., 2012. Data-driven subtypes of major depressive disorder: a systematic review. BMC Med., 10(1), 156.

Wong, M.L., Arcos-Burgos, M., Liu, S., Velez, J.I., Yu, C., Baune, B.T., Jawahar, M.C., Arolt, V., Dannlowski, U., Chuah, A., Huttley, G.A., Fogarty, R., Lewis, M.D., Bornstein, S.R., Licinio, J., 2016. The PHF21B gene is associated with major depression and modulates the stress response. Mol. Psychiatry, 22(7), 1015-1025.

Wong, M.L., Licinio, J., 2001. Research and treatment approaches to depression. Nat. Rev. Neurosci. 2(5), 343-351.

Yu, C., Arcos-Burgos, M., Licinio, J., Wong, M.L., 2017a. A latent genetic subtype of major depression identified by whole-exome genotyping data in a Mexican-American cohort. Transl. Psychiatry 7(5), e1134.

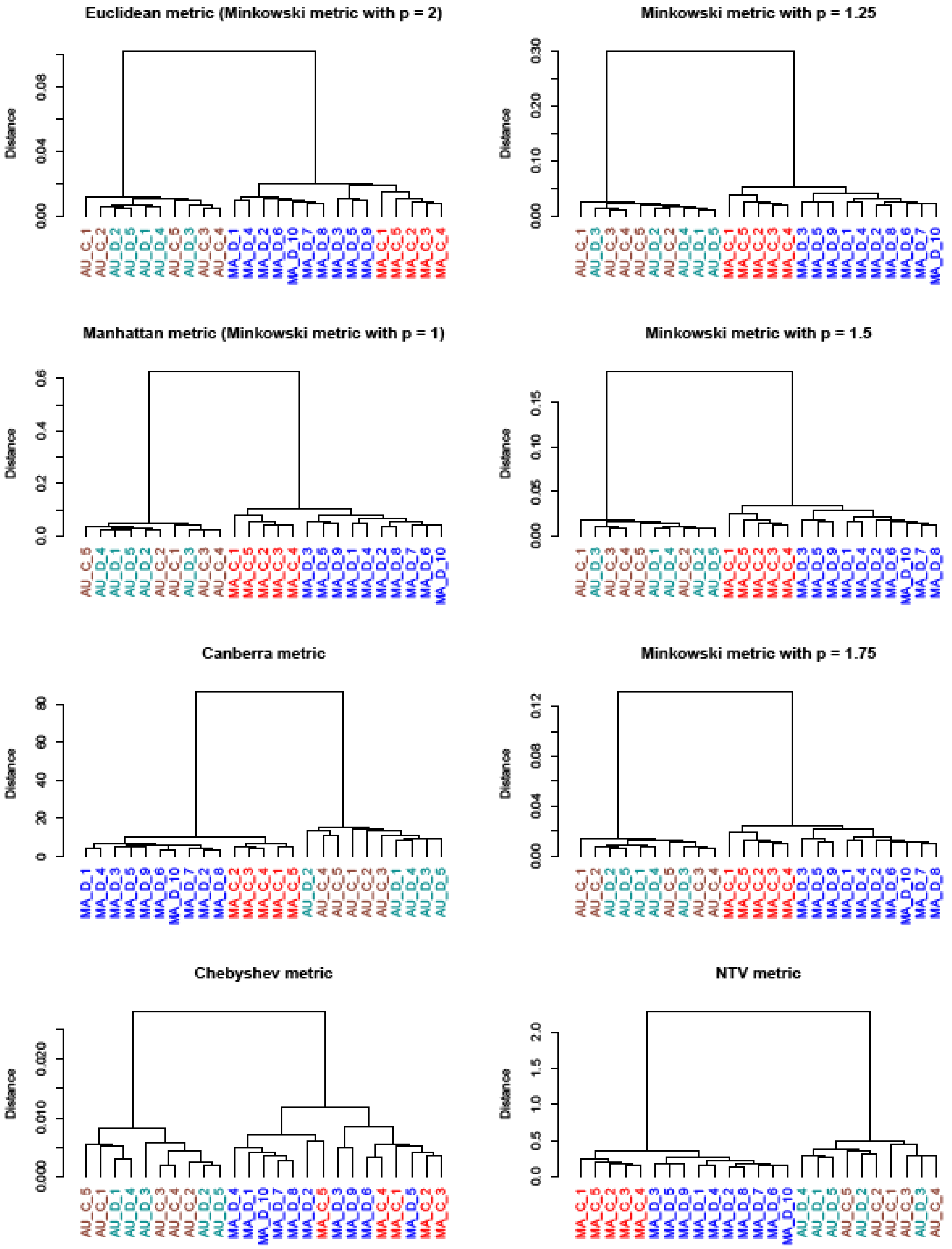
Yu, C., Baune, B.T., Licinio, J., Wong, M.L., 2017b. Single-nucleotide variant proportion in genes: a new concept to explore major depression based on DNA sequencing data. J. Hum. Genet. 62(5), 577-580.

Yu, C., Baune, B.T., Licinio, J., Wong, M.L., 2017c. Whole-genome single nucleotide variant distribution on genomic regions and its relationship to major depression. Psychiatry Res. 252, 75-79.

Figure legends

Figure 1: The dendrograms constructed by Ward’s method using eight different distance metrics. MA_D, Mexican-American MDD case; MA_C, Mexican-American control; AU_D, Australian MDD case; AU_C, Australian control.

Figure 2: Hierarchical clustering of 25 human subject with au/bp values (%) on Canberra metric and Ward’s method.



Euclidean metric (Minkowski metric with p = 2)

Minkowski metric with p = 1.25

Manhattan metric (Minkowski metric with p = 1)

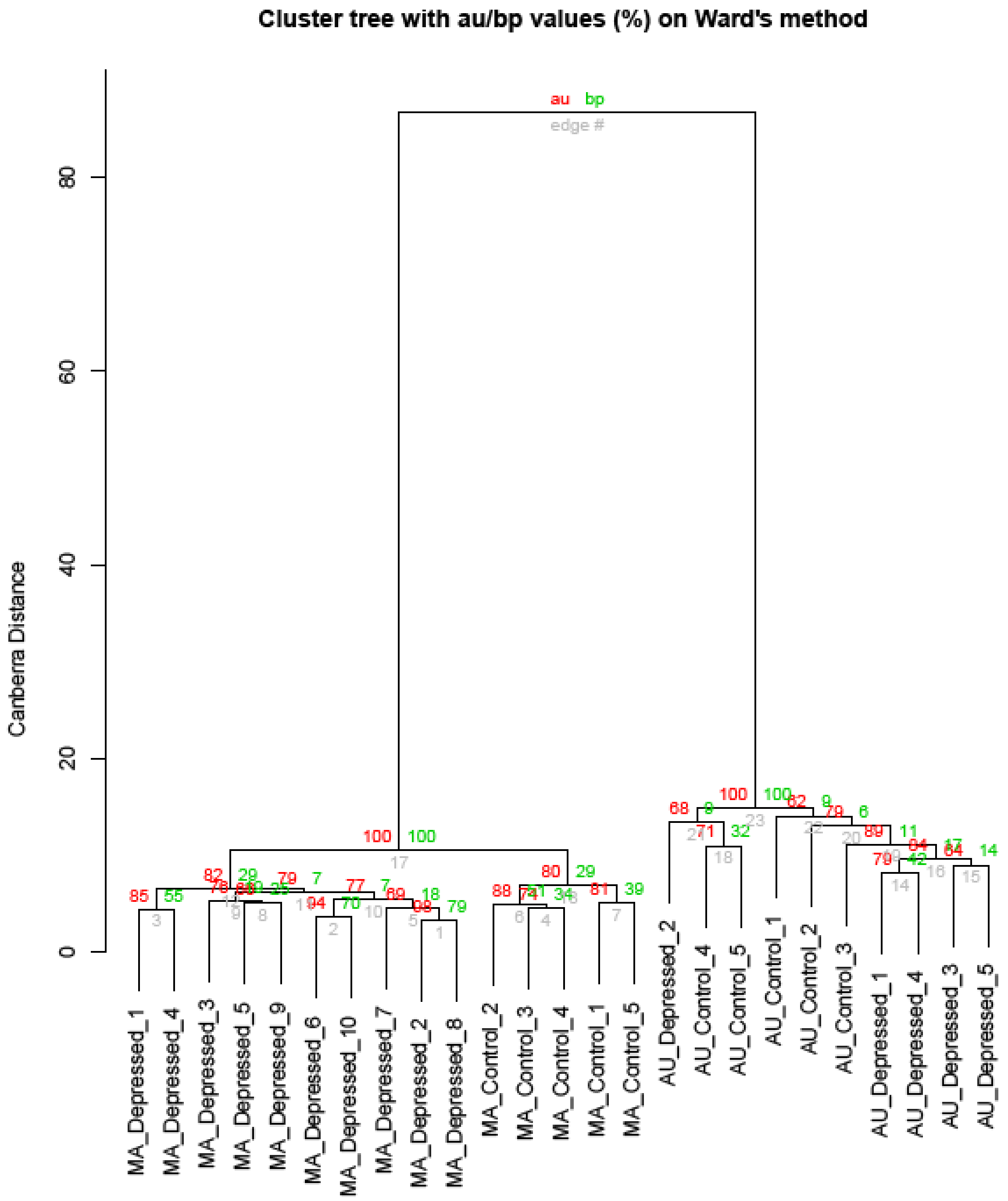
Minkowski metric with p = 1.5

Canberra metric

Minkowski metric with p = 1.75

Chebyshev metric

NTV metric



Highlights

- Sequencing allows us to detect all single-nucleotide variants within an individual.
- It is desirable to develop non-invasive genetic tests by using sequencing data.
- Multivariate cluster analysis can differentiate depressed cases and controls.
- We investigated combinations of distance metrics and clustering criteria.

- Canberra metric and Ward's method are recommended for genetic clustering.