Journal of Human Genetics (2017) 62, 577–580
© 2017 The Japan Society of Human Genetics All rights reserved 1434-5161/17
www.nature.com/jhg

SHORT COMMUNICATION

Single-nucleotide variant proportion in genes: a new concept to explore major depression based on DNA sequencing data

Chenglong Yu[1,2], Bernhard T Baune[3], Julio Licinio[1,2] and Ma-Li Wong[1,2]

Major depressive disorder (MDD) is a common psychiatric illness with significant medical and socioeconomic impact. Genetic factors are likely to play important roles in the development of this condition. DNA sequencing technology has the ability to identify all private genetic mutations and provides new channels for studying the biology of MDD. In this proof-of-concept study we proposed a novel concept, single-nucleotide variant proportion (SNVP), to investigate MDD based on whole-genome sequencing (WGS) data. Our SNVP-based approach can be used to test newly found candidate genes as a complement to genome-wide genotyping analysis. Furthermore, we performed cluster analysis for MDD patients and ethnically matched healthy controls, and found that clusters based on SNVP may predict MDD diagnosis. Our results suggest that SNVP may be used as a potential biomarker associated with major depression. Our methodology could be a valuable predictive/diagnostic tool as one can test whether a new subject falls within or close to an existing MDD cluster. Advances in this study design have the potential to personalized treatments and could include the ability to diagnose patients based on their full or part DNA sequencing data.

Journal of Human Genetics (2017) **62,** 577–580; doi:10.1038/jhg.2017.2; published online 2 February 2017

The development of new and cheaper sequencing technology is allowing scientists to search for new approaches to perform personalized treatment.[1] Whole-genome sequencing (WGS) can determine single-nucleotide variants (SNVs) which are private genetic variants and identify all genetic variants within each person.[2] A channel for studying major depressive disorder (MDD) is to identify its association with gene expression which suggests a measurable impact of current MDD status on gene expression.[3] This approach inspired us to investigate whether single-base variation could also be translated to quantitative measurements. The variations occurring within a DNA sequence influence gene structure and its protein function. Furthermore, SNVs in genes correlate with differences in the way individuals respond to a drug treatment or in their susceptibility to a complex disease such as major depression.

Here we propose a novel concept: single-nucleotide variant proportion (SNVP) in genes, to explore MDD based on DNA sequencing data. Investigating SNVP in MDD-associated genes may help identify pathways involved in MDD, and SNVP-MDD associations may reveal hidden genetic structures in this complex disorder. We have recently identified common and rare variations in a total of 46 genes that may confer susceptibility to MDD in a Mexican–American cohort.[4] We obtained complete WGS data for a group of 15 participants selected from a Mexican–American cohort,[5–7] 10 MDD patients and 5 controls. We have confirmed that in the cohort there was no family or population structure among all those individuals[4] and no blood relationship among the 15 selected participants. We also included WGS data from a group of 10 Australians of European-Ancestry including 5 MDD cases and 5 controls as a comparison group. We performed SNV-calling analysis of high quality WGS paired-end reads using a previously described pipeline.[4]

SNVP in a gene is defined as the ratio of the number of SNVs to the number of all nucleotides in this gene sequence. For example, the gene G has L nucleotides in its DNA sequence, and there are n SNVs in this gene in a given person. The SNVP in gene G for this person is $\frac{n}{L}\times 100\%$. We calculated the SNVP of 46 genes in 25 human subjects. Difference between two group means was tested using independent two-sample t-test. As we studied 46 genes, P-values were corrected using the false discovery rate method[8] and significance was set at $\leqslant 0.05$. We calculated the Euclidean distance between two subjects based on the SNVP for 46 genes. After obtaining the distance matrix across individuals, the multi-dimensional scaling (MDS) method[9] was used to detect distance relationships between individuals in a two-dimensional picture. We then used the neighbor-joining method[10] on the distance matrix to reconstruct the cluster tree drawn using MEGA software.[11] For further details see Supplementary Materials and Methods.

Figure 1 profiles SNVP in those 46 genes for 25 human subjects in a heat map. We can find that Mexican–American individuals have significantly different SNVPs on many genes when compared

[1]Mind and Brain Theme, South Australian Health and Medical Research Institute, North Terrace, Adelaide, SA, Australia; [2]School of Medicine, Flinders University, Bedford Park, SA, Australia and [3]Discipline of Psychiatry, School of Medicine, University of Adelaide, Adelaide, SA, Australia
Correspondence: Dr C Yu, Mind and Brain Theme, South Australian Health and Medical Research Institute, PO Box 11060, Adelaide, SA 5001, Australia.
E-mail: chenglong.yu@sahmri.com

Received 5 December 2016; revised 21 December 2016; accepted 26 December 2016; published online 2 February 2017

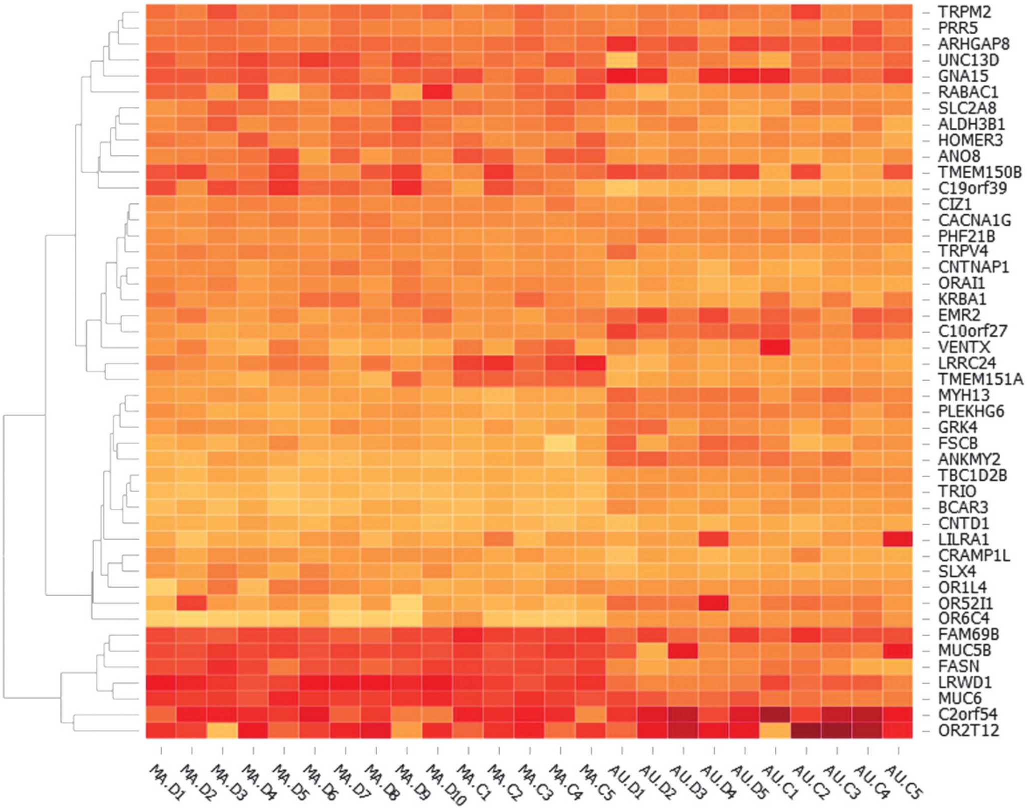

Figure 1 Heat map showing SNVP in 46 genes for 25 human subjects. AU, Australian; C, control; D, depression; MA, Mexican–American; SNVP, single-nucleotide variant proportion. The deeper the color, the higher the SNVP abundance.

to Australian individuals of European-Ancestry (see Supplementary Table S1). Statistical test results for SNVP across those 46 genes are summarized in Table 1. In the Mexican–American sample significant differences were found between MDD case and control in 9 genes, namely the *CNTD1*, *GRK4*, *LRRC24*, *MYH13*, *PHF21B*, *SLX4*, *TBC1D2B*, *TMEM151A* and *TRPV4* genes. For the Australian group, significant results were found for the *MUC6* and *TBC1D2B* genes between MDD cases and controls. Specifically, SNVP in gene *TBC1D2B* which is located in Chr 15q24.3–q25.1 shows significant results in both Mexican–American and Australian groups.

Following the proposed method, we calculated the Euclidean distance matrix of 25 human subjects using their SNVPs of all 46 genes. Then we applied the classical MDS method to map the distance matrix in a two-dimensional plane. Each point in the plane represents one individual. In Figure 2a there is an apparent separation between the Mexican–American group and the Australian group, as expected. However, there is no clear distinction between MDD cases and controls. As the MDS transformation from distance matrix to two-dimensional space may lose some high-dimensional distance information, we constructed the neighbor-joining tree to represent the clustering results. Figure 2b shows that all 10 Mexican–American MDD patients are grouped together in a cluster, and within that cluster the Mexican–American control group were separated far away from the MDD group. The Australians of European-ancestry group, as a different population, assembled as an obvious outgroup away from Mexican–Americans. These interesting results imply that the SNVP in some genes may be associated with MDD.

Our method may be a valuable predictive/diagnostic tool. A subject can be represented by a multi-dimensional feature vector that indicates the SNVP of every candidate gene. In the corresponding multi-dimensional feature space, some binary classifiers such as logistic regression[12] and support vector machine[13] can be used to distinguish cases and controls, and then prediction accuracy, sensitivity and specificity can be obtained. As an example, the support vector machine classifier separated MDD cases (all $\text{SNVP} < 0.00126$) and controls (all $\text{SNVP} > 0.00126$) with a 100% prediction accuracy in our Australian sample with a threshold of 0.00126 using the one-dimensional feature vector with the SNVP of the *TBC1D2B* gene (this SNVP was found to be significantly different between MDD and controls in both the Mexican–American and Australian groups). Our approach could work in a large number of subjects by using a combination of several candidate genes.

To the best of our knowledge this is the first study that formulates the concept SNVP. It may bring a new methodology of quantitative sequencing analysis at the gene level. Our proposed SNVP may also be a quantitative variable similarly to gene expression level,[14] as differences in SNVP may reflect differences in medication response or disease risk. Our approach may be complementary to genome-wide genotyping data analysis by testing newly found candidate genes. In this study, we used WGS data to calculate the SNVP. However, it would be sufficient to obtain targeted sequencing data of specific regions of interest, this would allow for a more efficient use of funding and bioinformatics resources and time.

Table 1 The statistical test results for SNVP in four groups cross the 46 genes

	MA t-*test*			*AU* t-*test*		
Genes	t-*statistic*	P-*value*	*FDR*	t-*statistic*	P-*value*	*FDR*
ALDH3B1	−0.3646	0.0208	0.0653	0.3691	0.7216	0.9454
ANKMY2	3.2069	0.2158	0.3546	1.7515	0.1180	0.7688
ANO8	1.9672	0.2336	0.3553	1.3393	0.2173	0.7688
ARHGAP8	1.3924	0.0686	0.1577	−0.0606	0.9532	0.9981
BCAR3	−0.3267	0.0132	0.0567	0.2537	0.8061	0.9549
C10orf27	−1.4068	0.8107	0.9323	0.9345	0.3774	0.8680
C19orf39	0.8754	0.0169	0.0620	NA	NA	NA
C2orf54	1.4466	0.8444	0.9373	−1.0602	0.3200	0.8660
CACNA1G	−0.0286	0.0488	0.1181	0.6038	0.5627	0.9370
CIZ1	1.4626	0.1565	0.2940	−1.4303	0.1905	0.7688
CNTD1	0.8927	0.0063	0.0365[a]	0.1754	0.8652	0.9707
CNTNAP1	−1.3588	0.0437	0.1122	0.4829	0.6421	0.9454
CRAMP1L	0.5842	0.0175	0.0620	−0.5917	0.5704	0.9370
EMR2	0.0179	0.3607	0.4880	0.6096	0.5591	0.9370
FAM69B	1.5004	0.1137	0.2378	−1.4686	0.1801	0.7688
FASN	1.0466	0.8762	0.9373	0.1781	0.8631	0.9707
FSCB	0.1166	0.0317	0.0910	1.8190	0.1064	0.7688
GNA15	1.9089	0.3736	0.4910	0.3695	0.7214	0.9454
GRK4	1.3689	0.0016	0.0148[a]	1.5466	0.1605	0.7688
HOMER3	0.1958	0.3109	0.4333	−0.0586	0.9547	0.9981
KRBA1	−1.2718	0.2869	0.4124	−2.1679	0.0620	0.7130
LILRA1	−1.5110	0.4740	0.6057	−0.3236	0.7546	0.9454
LRRC24	0.6382	0.0002	0.0076[a]	−0.8018	0.4458	0.9322
LRWD1	−0.9616	0.0909	0.1992	−2.6737	0.0282	0.4324
MUC5B	3.1247	0.9660	0.9660	−0.1095	0.9155	0.9981
MUC6	0.4790	0.2061	0.3546	5.4194	0.0006	0.0145[a]
MYH13	2.3582	0.0003	0.0076[a]	1.1670	0.2768	0.8011
OR1L4	−0.5262	0.5589	0.6765	NA	NA	NA
OR2T12	−0.2592	0.8581	0.9373	−0.2490	0.8096	0.9549
OR52I1	1.4056	0.2100	0.3546	0.8485	0.4208	0.9218
OR6C4	−1.2060	0.5874	0.6929	−1.0000	0.3466	0.8680
ORAI1	0.5694	0.1295	0.2590	0.5369	0.6060	0.9454
PHF21B	−0.0660	0.0016	0.0148[a]	0.3155	0.7605	0.9454
PLEKHG6	−0.1960	0.1598	0.2940	1.1622	0.2787	0.8011
PRR5	3.2448	0.2395	0.3553	−1.3799	0.2049	0.7688
RABAC1	−0.2985	0.9306	0.9512	−1.2060	0.2623	0.8011
SLC2A8	0.6859	0.9278	0.9512	−0.6299	0.5464	0.9370
SLX4	0.2535	0.0032	0.0211[a]	0.4508	0.6641	0.9454
TBC1D2B	0.8479	0.0030	0.0211[a]	−8.9077	<0.0000	0.0009[a]
TMEM150B	1.5578	0.5022	0.6244	2.0246	0.0775	0.7130
TMEM151A	−0.6468	0.0006	0.0096[a]	−0.7303	0.4860	0.9370
TRIO	0.2082	0.0439	0.1122	−0.4933	0.6351	0.9454
TRPM2	1.3355	0.0213	0.0653	−0.9506	0.3697	0.8680
TRPV4	3.3540	0.0071	0.0365[a]	1.5155	0.1681	0.7688
UNC13D	0.2611	0.0136	0.0567	−0.3809	0.7132	0.9454
VENTX	1.5307	0.2294	0.3553	–0.7220	0.4909	0.9370

Abbreviations: AU t-test, the t-test for two groups AUD and AUC; AUC, Australian control; AUD, Australian MDD case; FDR, false discovery rate; MA t-test, the t-test for two groups MAD and MAC; MAC, Mexican–American control; MAD, Mexican–American MDD case; MDD, major depressive disorder; NA, not applicable; SNVP, single-nucleotide variant proportion.

[a] We use to show the significant test results after multiple testing correction (FDR <0.05).

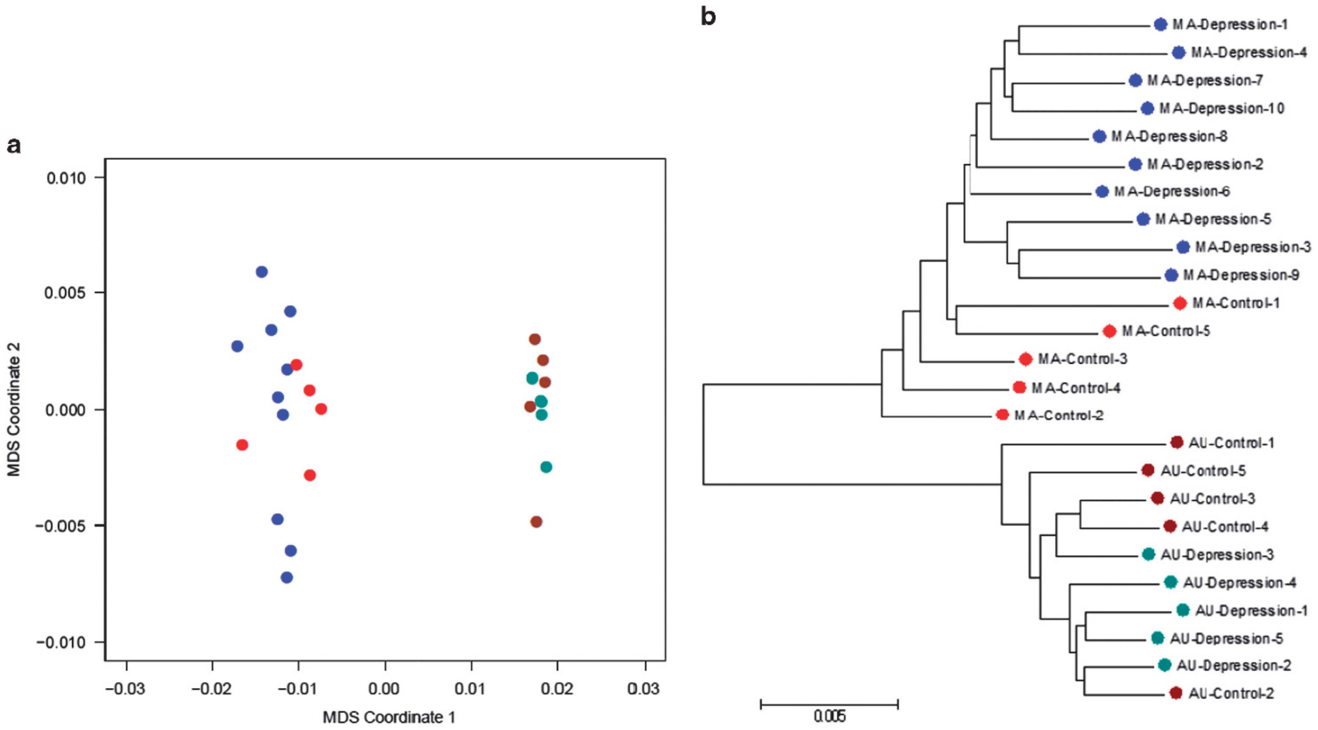

Figure 2 (**a**) MDS two-dimensional visualization of 25 human subjects (Mexican–American MDD case, blue point; Mexican–American control, red point; Australian MDD case, green point; Australian control, brown point). (**b**) Cluster tree for 25 human subjects based on their SNVPs of all the 46 genes. MDD, major depressive disorder; MDS, multi-dimensional scaling; SNVP, single-nucleotide variant proportion.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

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

1 Hamburg, M. A. & Collins, F. S. The path to personalized medicine. *N. Engl. J. Med.* **363**, 301–304 (2010).

2 Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl Acad. Sci. USA* **112**, 5473–5478 (2015).

3 Jansen, R., Penninx, B. W., Madar, V., Xia, K., Milaneschi, Y., Hottenga, J. J. *et al.* Gene expression in major depressive disorder. *Mol. Psychiatry* **21**, 339–347 (2016).

4 Wong, M. L., Arcos-Burgos, M., Liu, S., Velez, J. I., Yu, C., Baune, B. T. *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).

5 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).

6 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).

7 Wong, M. L., Dong, C., Flores, D. L., 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* **171**, 1297–1309 (2014).

8 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* **57**, 289–300 (1995).

9 Torgerson, W. S. Multidimensional scaling: I. theory and method. *Psychometrika* **17**, 401–419 (1952).

10 Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

11 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).

12 Zhu, J. & Hastie, T. Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5**, 427–443 (2004).

13 Yu, C., Deng, M., Zheng, L., He, R. L., Yang, J. & Yau, S. S. DFA7, a new method to distinguish between intron-containing and intronless genes. *PloS ONE* **9**, e101363 (2014).

14 Mehta, D., Menke, A. & Binder, E. B. Gene expression studies in major depression. *Curr. Psychiatry Rep.* **12**, 135–144 (2010).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)