Language/Cognition Gene Polymorphism Patterns Potentially Associated with Novel Teaching/Learning Technology Based on Brain-Computer Interface

Wei Xia[1,3#], Yiping Geng[2#], Shuaiyu Zhang[2#], Yongdong Xu[2*], Zhizhou Zhang[1*]

[1]BIOX Biotechnology Center, Harbin Institute of Technology, Weihai, China 264209, China
[2]School of Computer Science and Information Technology, Harbin Institute of Technology, Weihai ,264209 , China
[3]School of Languages and Literature, Harbin Institute of Technology, Weihai, China 264209, China
#equal contribution

*Corresponding author: Wei Xia (xiawei2015@hitwh.edu.cn),
Yiping Geng(1970914300@qq.com),
Shuaiyu Zhang(2316072618z@gmail.com),
Yongdong Xu (ydxu@hit.edu.cn),
Zhizhou Zhang (zhangzzbiox@hitwh.edu.cn)

Abstract. Brain-computer interfaces seem to be an inevitable direction of human evolution and will naturally be used in the field of education, including the cultivation of special talents and the prevention and treatment of specific brain diseases. Since each person's brain has individual characteristics, including differences in language and cognitive functions, the theoretical variations in language/cognitive genetic polymorphism patterns among diverse populations are essentially differences in the brain's inherent molecular hardware. This is crucial for the development of personalized brain-computer interface educational technologies. This study examined the sequence information of 239 language gene polymorphisms and 223 cognitive gene polymorphism loci in 201 whole-genome sequence samples. Through principal component analysis and two other clustering methods, we preliminarily discovered that modern humans contain at least four distinct language-cognition genetic polymorphism patterns. The first three patterns may correspond to only a minority of modern humans, while the last pattern may correspond to the vast majority. Since each pattern likely includes samples from all continents, this suggests that there may be no continent-specific language-cognition genetic polymorphism patterns.

Keywords: Brain-computer interface, Language gene, Cognition gene, Gene polymorphism, Pattern

T. Yao et al. (eds.), *Proceedings of the 2024 3rd International Conference on Engineering Management and Information Science (EMIS 2024)*, Advances in Computer Science Research 111,
https://doi.org/10.2991/978-94-6463-447-1_47

1 Introduction

As a novel learning method, brain-computer interface (BCI) is inevitably applied in new teaching/learning practice in the future. BCI is actually a computer hardware combination [1], and there are problems of whether the hardware performance is fully matched and compatible. Theoretically, the electronic signals from the computer machine system input into the human brain (biological computing device) need to be recognized and processed by language/cognitive function modules, so the corresponding hardware structural differences in language/cognitive functions for different individuals need to be basically studied. One way to find these hardware structural differences is to observe and compare the diversity of language/cognitive gene polymorphism patterns (LCGPP) among different individuals. This diversity is one of the molecular bases of the macro-performance differences of the brain's language/cognitive function modules, and will be an important reference for development of personalized BCI devices [2].

This study collected genomic sequences of 201 individuals from different populations throughout history and across the world. Using self-developed software, it conducted a diversity scan on 239 language gene polymorphism sites and 223 cognitive gene polymorphism sites. A PCA (principal component analysis) was performed on the data from 462 diversity sites of the above samples, preliminarily determining that there are roughly four types (or a continuous spectrum) of human language cognition gene polymorphism patterns. The distinctive BCI characteristics corresponding to these five main LCGPPs await further investigation in the future. Moreover, since the representativeness of the collected samples for modern humans is still far from adequate, there is a need to further improve sample information for various populations in the future. One of the benefits of including ancient samples is that it can help determine which genes or their polymorphic sites are most conservative and crucial for language/cognition functions.

2 Methods

2.1 Language/Cognition Genes and Their Snps

Language/Cognition abilities are closedly ssociated with several dozens of genes, and those genes can be called language gene or cognition gene after the gene’s fucntion is confirmed especially experimentally. For both language gene and cognition gene, SNP sites in the dbSNP database were selected in a way that the each whole gene region was relatively equally spanned by the selected sites, plus those already with known clinical effects (seen in the Genecards database).This study employed 36 language/cognition genes, and a total 239 SNPs from 18 language genes were selected, while 223 SNPs from 18 cognition genes were selected (Information for these genes and their SNP sites seen in ref.[3-8, 9-15]).

2.2 Sample Genome Sequences

All genome sequences were downloaded from ENA database (https://www.ebi.ac.uk/ena/browser/). Total 201 whole genomes (including 68 ancient genomes) from 5 continents (Africa, Asia, Europe, North America, and South America) were collected, among which, there are 32 from EastAsia, 21 from Africa, 28 from Europe, 12 from SouthAm, 4 from NorthAm, 6 forbirds, 9 for fish, 33 for Primates, 24 for OtherA group (including rodents, reptiles, Laurasistheria) and 18 for OtherB animal group (Table 1). More information can be requested from the authors.

2.3 SNP Information Abstraction and PCA Analysis

The authors used python-based hash07plus03 software to extract all 462 SNP (single nucleotide polymorphism)sequences from each genome. In all 201 genomes, the sizes mainly range from 10G to 200G. Genomes less than 10G were neglected or only used as a reference. Principal Component Analysis（PCA）was performed using R packages FactoMineR, factoextra and ggplot2.The main codes are listed as below.

```
> library(FactoMineR)
> library(factoextra)
> library(ggplot2)
>region<- read.delim('C:/RBook/20220516fastqSNPdata.txt', row.names = 1, sep = '\t')
>region<- t(region)
>region.pca <- PCA(region, ncp = 2, scale.unit = TRUE, graph = FALSE)
> plot(region.pca)
> pca_sample <- data.frame(region.pca$ind$coord[ ,1:2])
> head(pca_sample)
> pca_eig1 <- round(region.pca$eig[1,2], 2)
> pca_eig2 <- round(region.pca$eig[2,2],2 )
> pca_eig1
> pca_eig2
> group <- read.delim('C:/RBook/group3.txt', row.names = 1, sep = '\t', check.names = FALSE)
> group <- group[rownames(pca_sample), ]
> pca_sample <- cbind(pca_sample, group)
> pca_sample$samples <- rownames(pca_sample)
> head(pca_sample)
> library(ggrepel)
> ggplot(data = pca_sample, aes(x = Dim.1, y = Dim.2)) +geom_point (aes(color = group), size = 3) + scale_color_manual (values = c ('purple', 'red', 'green','blue','brown', 'pink','yellow','orange','grey')) + theme(panel.grid = element_blank(), panel.background = element_rect (color = 'black', fill = 'transparent'), legend.key = element_rect (fill = 'transparent')) + labs(x = paste('PCA1:', pca_eig1, '%'), y = paste('PCA2:', pca_eig2, '%'), color = '') + geom_text_repel (aes (label = samples), size = 3, show.legend = FALSE, box.padding = unit(0.25, 'lines'))
```

Table 1. Genome samples employed in this study

Sample	Group	Region	Age*	Sample	Group	Region	Age*	Sample	Group	Region	Age*
et1	Africa	Ethiopia	4500	pa4	SouthAsia	Pakistan	0	x3	others	unkown	0
ga1	Africa	Gambia	0	pa5	SouthAsia	Pakistan	0	x4	others	unkown	0
ga2	Africa	Gambia	0	pa6	SouthAsia	Pakistan	0	x5	others	unkown	0
ga3	Africa	Gambia	0	sr1	SouthAsia	SriLanka	0	x6	others	unkown	0
ga4	Africa	Gambia	0	sr2	SouthAsia	SriLanka	0	p1	Primates	unkown	0
ga5	Africa	Gambia	0	sr3	SouthAsia	SriLanka	0	p10	Primates	unkown	0
ga6	Africa	Gambia	0	kz1	Euro	Poland	4500	p11	Primates	unkown	0
ke1	Africa	Kenya	0	cz1	Euro	Czech	45000	p12	Primates	unkown	0
ke2	Africa	Kenya	0	de2	Euro	Russia	100000	p13	Primates	unkown	0
ke3	Africa	Kenya	0	de3	Euro	Russia	78000	p14	Primates	unkown	0
le1	Africa	unknown	0	de4	Euro	Russia	100000	p15	Primates	unkown	0
le2	Africa	unknown	0	de5	Euro	Russia	78000	p16	Primates	unkown	0
le3	Africa	unknown	0	dep	Euro	Russia	78000	p17	Primates	unkown	0
mo1l	Africa	Morocco	15000	fi1	Euro	Finnish	0	p18	Primates	unkown	0
mo1s	Africa	Morocco	15000	fi2	Euro	Finnish	0	p19	Primates	unkown	0
sa1	Africa	Southern Africa	0	fi3	Euro	Finnish	0	p2	Primates	unkown	0
sa2	Africa	Southern Africa	0	ge1	Euro	Georgia	9700	p20	Primates	unkown	0
sa3	Africa	Southern Africa	0	la1	Euro	Latvia	5900	p21	Primates	unkown	0
ss1	Africa	sub-Sahara	4500	nd1	Euro	Russia	50000	p22	Primates	unkown	0
ss2	Africa	sub-Sahara	7900	nd10	Euro	Spain	430000	p23	Primates	unkown	0
ss3	Africa	sub-Sahara	3160	nd2	Euro	Spain	90000	p24	Primates	unkown	0
b1	Birds	unknown	0	nd3	Euro	Spain	90000	p25	Primates	unkown	0
b2	Birds	unknown	0	nd4n	Euro	Russia	50300	p26	Primates	unkown	0
b3	Birds	unknown	0	nd5n	Euro	Russia	60000	p27	Primates	unkown	0
b4	Birds	unknown	0	nd6	Euro	Belgium	120000	p28	Primates	unkown	0
b5	Birds	unknown	0	nd7	Euro	Germany	120000	p29	Primates	unkown	0
b6	Birds	unknown	0	nd8	Euro	Russia	60000	p3	Primates	unkown	0
c4	EastAsia	China	0	nd9	Euro	Russia	96700	p30	Primates	unkown	0
c5	EastAsia	China	0	sp1	Euro	Spain	0	p31	Primates	unkown	0
c6	EastAsia	China	0	sp2	Euro	Spain	0	p32	Primates	unkown	0
c7	EastAsia	China	7000	sp3	Euro	Spain	0	p33	Primates	unkown	0
c8	EastAsia	China	7000	sp4	Euro	Spain	0	p4	Primates	unkown	0
c9	EastAsia	China	6000	sp5	Euro	Spain	0	p5	Primates	unkown	0
c11	EastAsia	China	4000	sp6	Euro	Spain	0	p6	Primates	unkown	0
c12	EastAsia	China	4000	F1	Fish	unkown	0	p7	Primates	unkown	0
c13	EastAsia	China	2100	F2	Fish	unkown	0	p8	Primates	unkown	0

c14	EastAsia	China	2200	F3	Fish	unkown	0	p9	Primates	unkown	0
c15	EastAsia	China	3100	F4	Fish	unkown	0	bz1	South Am	Brazil	8000
c16	EastAsia	China	3900	F5	Fish	unkown	0	ch1	South Am	Chile	4700
c17	EastAsia	China	4100	F6	Fish	unkown	0	me1	South Am	Mexica	0
c18	EastAsia	China	4100	F79	Fish	unkown	0	ur1	South Am	Uruguay	668
c19	EastAsia	China	5200	lc1	Fish	Tanzania	0	ur2	South Am	Uruguay	1400
c20	EastAsia	China	40000	st1	Fish	unkown	0	pe1	South Am	Peru	0
c21	EastAsia	China	4000	km1	NorthAm	US	9000	pe2	South Am	Peru	0
c22	EastAsia	China	4000	sc1	NorthAm	US	10000	pe3	South Am	Peru	0
c23	EastAsia	China	5300	us1	NorthAm	US	2000	so1	South Am	Brazil	10000
c24	EastAsia	China	3700	us2	NorthAm	US	12500	so2	South Am	Argentina	500
c25	EastAsia	China	5500	d1	rodents	unkown	0	so7	South Am	Chile	4500
c26	EastAsia	China	4000	d2	rodents	unkown	0	so8	South Am	Chile	6000
c27	EastAsia	China	2300	d3	rodents	unkown	0	ap1	Another	unkown	0
dc1	EastAsia	China	0	d4	rodents	unkown	0	bc1	Another	unkown	0
dc2	EastAsia	China	0	d5	rodents	unkown	0	bc2	Another	unkown	0
dc3	EastAsia	China	0	d6	rodents	unkown	0	cm1	Another	unkown	0
dg1	EastAsia	Russia/China	8000	L1	Laurasiatheria	unkown	0	cm2	Another	unkown	0
dg2	EastAsia	Russia/China	8000	L2	Laurasiatheria	unkown	0	dp1	Another	unkown	0
in2	South Asia	India	0	L3	Laurasiatheria	unkown	0	dp2	Another	unkown	0
in4	South Asia	India	0	L4	Laurasiatheria	unkown	0	dp3n	Another	unkown	0
in5	South Asia	India	0	L5	Laurasiatheria	unkown	0	gb1	Another	unkown	0
mg1	EastAsia	Mongolia	34000	L6	Laurasiatheria	unkown	0	ha1	Another	unkown	0
ne10m	South Asia	Nepal	2000	R1	reptiles	unkown	0	hu1	Another	unkown	0
ne2	South Asia	Nepal	2000	R2	reptiles	unkown	0	pr1	Another	unkown	0
ne3	South Asia	Nepal	2000	R3	reptiles	unkown	0	rr1	Another	unkown	0
ne5	South Asia	Nepal	2000	R4	reptiles	unkown	0	rr2	Another	unkown	0
ne9m	South Asia	Nepal	2000	R5	reptiles	unkown	0	rt1	Another	unkown	0
ja2	EastAsia	Japan	3500	R6	reptiles	unkown	0	rt2	Another	unkown	0
jm1	EastAsia	Japan	3000	x1	others	unkown	0	rt3	Another	unkown	0
jm2	EastAsia	Japan	3000	x2	others	unkown	0	su1	Another	unkown	0

* Age: 0 means the present year; 4500 means 4500 before present

3 Results and Discussion

Figure 1 illustrates the SNP polymorphism patterns of modern and ancient human population samples, containing four circles. The leftmost circle encompasses one modern sample (c5), the second circle also contains one modern sample (c4), the third circle includes two modern samples (pa4, c6), and the rightmost circle contains at least fifteen modern samples. This essentially suggests that modern humans have at least four different language/cognition gene polymorphism patterns. The first three circles on the left imply that some modern humans still possess genetic polymorphism patterns from ancient times, whereas the rightmost circle indicates that some ancient genetic polymorphism patterns have continued into the contemporary populations.

The first circle on the left roughly includes six East Asians, three Europeans, two Africans, and one American; the second circle comprises three Asians, two Europeans, and one American; the third circle consists of five Asians and two Africans. The fourth circle, the rightmost one, contains at least seventeen Asians, nine Africans, ten Europeans, and three Americans. Although the samples in this study were not evenly drawn from each continent (for instance, based on population sizes), the aforementioned four circles all generally encompass samples from several continents. This suggests that throughout human evolutionary history, interactions among populations across various continents have always been taking place, leading to the possibility that there are likely no continent-specific language /cognition gene polymorphism patterns within modern populations; this is positive news for the development of brain-computer interface technologies, although there is still a need to develop specific technological products for small population groups.

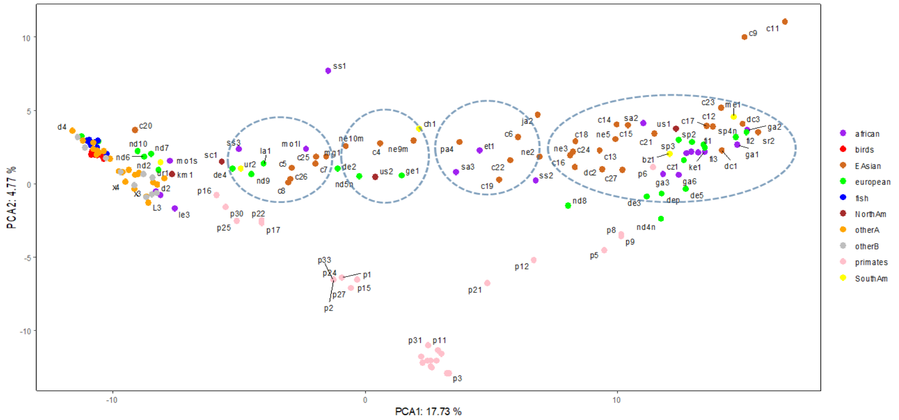

Fig. 1. PCA results using SNP data from 201 genome samples. Note: some samples were not marked in the figure due to crowdedness. The author also used two other clustering methods, K-Mean clustering and Hierarchical Clustering, and obtained similar results (data not shown).

In figure 1, p5, p6, p8 and p9 represent Gorilla gorilla, Homo sapiens, Pan paniscus and Pan troglodytes, respectively. The positions of the above four samplessupport that p8 (Pan paniscus) and p9 (Pan troglodytes) possess most similar language/cognition gene polymorphism patterns as modern human, so these two types of model animals shall be suitable to test some functions of BCI devices.

This study preliminarily discovered at least 4 different language-cognition genetic polymorphism patterns in the population. The characteristic genes and sequence polymorphism sites corresponding to these patterns certainly require further exploration. In this process, much more sequence polymorphism information is needed, such as inverted sequences, sequence deletions, repeated sequences, coding region sequences, non-coding sequences, remote regulatory sequences, remote spatially adjacent sequences, etc., rather than just simple SNP sites. Using more forms of polymorphism, even more language/cognition genes, and more samples will provide the basic supporting information needed for BCI product development, thereby determining which factors are used to develop universal products and which factors are used to develop personalized products. Furthermore, based on the genomic sequence polymorphisms corresponding to hundreds of human diseases [16-17], there is an inexhaustible resource treasure trove for the future development of diverse BCI products.

As for the specific implications or functional implications of these patterns, it is still too early to describe. The authors checked out the pattrn-specific conservative SNP sites, and found that the left two patterns in figure 1 basically had few conservative SNP contents within each group(data not shown). However, from left to right circles (patterns), the number of conservative SNP sites increased quickly, which is worth investigating in future. Differential contents of conservative SNP contents in each pattern definitely affect language and cognition characteristics, thus directly influencing BCI features to which each pattern can get adapted.

4 Conclusions

This study explores the molecular basis for the future development of individualized brain-computer interface (BCI) technologies in the field of education from the perspective of language and cognitive genetic polymorphism patterns. Using software developed by our research team, we examined the sequence information of 239 language gene polymorphisms and 223 cognitive gene polymorphism loci in 201 whole-genome sequence samples from ancient and modern times, as well as from different parts of the world. Through principal component analysis and two other clustering methods, we preliminarily discovered that modern humans contain at least four distinct language-cognition genetic polymorphism patterns. The first three patterns may correspond to only a minority of modern humans, while the last pattern may correspond to the vast majority. Since each pattern includes samples from all continents, this suggests that there may be no continent-specific language-cognition genetic polymorphism patterns.

Acknowledgments

This study was supported by State Language Commission Research Grant (YB135-117), Association of Chinese Graduate Education Grant (B-2017Y0505-079) and National Research Center for Foreign Language Education Grant (ZGWYJYJJ10A042).

References

1. Kawala-Sterniuk, A. et al. (2021) Summary of over Fifty Years with Brain-Computer Interfaces—A Review. Brain Sci., 11(1): 43. DOI: 10.3390/brainsci11010043.
2. Ma Y, Gong A, Nan W, Ding P, Wang F, Fu Y.(2023) Personalized Brain–Computer Interface and Its Applications. *Journal of Personalized Medicine*, 13(1): 46.https://doi.org/10.3390/jpm13010046.
3. Zhang, Z., Zhang, S., Zhou, H., & Xu, Y. (2024) A general evolution landscape of language and cognition genes. *BioRxiv*.https://doi.org/10.1101/2024.03.11.584338.
4. Liu, Z., et al. (2021) Correlation analysis between language gene polymorphism and geography/society parameter from twenty-six countries. Research Square. https://doi.org/10.21203/rs.3.rs-960107/v1.
5. Xia, W., Zhang, Z. Z., & Guo, C. L. (2019)Correlation analysis between English-Chinese translation-based writing error types and language gene polymorphisms for Chinese graduate students. International Journal of Learning and Teaching, 5(4):333-338. DOI: 10.18178/ijlt.5.4.333-338.
6. Xia, W., Zhang, Z., & Guo, C. (2019) Novel education technology may derive from personal genome data: a language gene polymorphism site potentially associated with translation-writing errors in a bilingual classroom of Chinese students. In: 2019 International Conference on Modern Educational Technology.Nanjing.pp. 45-48.https://doi.org/10.1145/3341042.3341067.
7. Xia, W., Qin, H., Li, X., & Zhang, Z. (2018) Language gene basis for precision personalized education and liberal education (I). In: ICSET 2018. Taipei. pp. 112–116. https://doi.org/10.1145/3268808.3268859.
8. Xia, W., & Zhang, Z. (2017) Language gene network patterns may facilitate relationship setting-up between language genotypes and students' class-performance. International Journal of Learning and Teaching, 3(4): 259-263.DOI: 10.18178/ijlt.3.4.259-263.
9. Li M, Zhang W, Zhou X. (2020) Identification of genes involved in the evolution of human intelligence through combination of inter-species and intra-species genetic variations. PeerJ, 8:e8912. http://doi.org/10.7717/peerj.8912.
10. Goriounova, N. A., & Mansvelder, H. D. (2019) Genes, Cells and Brain Areas of Intelligence. Front Hum Neurosci,13: 390595. DOI: 10.3389/fnhum.2019.00044.
11. Savage, J.E., Jansen, P.R., Stringer, S. et al. (2018) Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence.*Nature genetics*, 50: 912–919. https://doi.org/10.1038/s41588-018-0152-6.
12. Sniekers, S. et al. (2017) Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nature genetics*, 49: 1107–1112. https://doi.org/10.1038/ng.3869.
13. Xia, W., & Zhang, Z. (2022)Language gene polymorphism pattern survey provided important information for education context in human evolution. *Biorxiv*. https://doi.org/10.1101/2022.10.31.514632.
14. Shi, L., Hu, E., Wang, Z.et al. (2017) Regional selection of the brain size regulatinggene CASC5 provides new insight into human brain evolution. *Human genetics*, 136(2):193-204.DOI: 10.1007/s00439-016-1748-5.
15. Tattersall, I. (2023) Endocranial volumes and human evolution. F1000Research, 12:565. https://doi.org/10.12688/f1000research.131636.1.
16. Ramadan, R.A., Altamimi, A.B. (2024) Unraveling the potential of brain-computer interface technology in medical diagnostics and rehabilitation: A comprehensive literature review. Health Technol, 14: 263–276. https://doi.org/10.1007/s12553-024-00822-1.

17. Lorenz, E. A., Su, X., & Skjæret-Maroni, N. (2024) A review of combined functional neuroimaging and motion capture for motor rehabilitation. Journal of NeuroEngineering and Rehabilitation, 21(1): 3. https://doi.org/10.1186/s12984-023-01294-6.