Correlation analysis between language gene polymorphism and geography/society parameter from twenty-six countries

Zongxuan Liu

Harbin Institute of Technology

Wei Xia

Harbin Institute of Technology

Bo Sun

Harbin Institute of Technology

Changlu Guo

Harbin Institute of Technology

Zhizhou Zhang (✉ zhangzzbiox@hitwh.edu.cn)

Harbin institute of Technology <https://orcid.org/0000-0002-2895-2682>

Case Report

Keywords: language gene, single nucleotide polymorphism, geography/society, correlation

Posted Date: October 8th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-960107/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Human language diversity, as a biological phenotype, shall be genetically linked with language gene polymorphism. Meanwhile, this phenotype is historically shaped by local geographical/social factors. But how many language gene polymorphisms have direct correlations with some geography/society characteristics during the long-run evolution of human languages is an interesting question and largely remains uninvestigated. This study selected a series of geography/society factors (including 13 geographical factors and 21 social factors) from 26 countries and 111 single nucleotide polymorphisms (SNPs) randomly selected from 13 language genes. Principal component analysis (PCA) was performed to explore their potential correlations. Preliminary but interesting results were obtained as follow. (1) Most geographical parameters are concentrated into one cluster in the PCA diagram. The cluster contains 12 parameters that are positively correlated with each other; (2) PCA diagrams divide social parameters into four clusters, among which exist positive and negative correlations; (3) The strongest positive correlations were observed at one of ATP2C2 gene SNPs (ATP-1: rs78371901); the strongest negative correlations were found at one of NFXL1 gene SNPs (NFX-6: rs1440228); and the least correlations with language gene SNPs were observed at four geography/society factors: aash (Annual average rainfall), fore (Forest coverage), pden (Population density of the country) and rway (Runway traffic mode).

Introduction

It is a common sense that biological phenotypes are shaped by geographical factors, and a typical phenotype is fixed and inherited by a genotype^[1-2]. Natural selection means selective reservation of phenotypes in a specific geographical environment, so human languages (there are at least 5000 different human languages worldwide) also represent marvelous results of natural selection (at least in the early stage of human language occurrence)^[3-6]. Different geographical environments yield different local languages, plus different language gene polymorphisms. But how many language gene polymorphisms have direct correlations with some geography/society characteristics during the long-run evolution of human languages is an interesting question and largely remains uninvestigated.

During the early stage of human evolution, human being and other kinds of animals shall be very similar in the context of language emergence and its usage for vocality and communication. Later on due to the uniqueness of human language genes and evolutionary advantages of human brain, human language gradually becomes a powerful characteristic of human intelligence. There are some evidence that the FOXP2 gene of Neandertals already acquired key mutations that only found in modern human being^[7], suggesting that language genes had finished their main mutational acoustic adaptations before *Homo ergaster* and other early *Homo sapiens* migrated out of Africa^[8-9]. Later on *Homo sapiens* had to get adapted to different novel geographical conditions around the world in north Europe, east Asia, south Asia, north America and south America, etc. Thus language genes had many chances to be polymorphic in order to micro-adjust the vocal system for better acoustic adaptation. Many studies^[10-12] found that design of acoustic communication systems within species appears to be directly shaped by

environmental factors, so it is with the phonological structure of human languages. Environments in which higher frequencies are less faithfully transmitted (for example, denser vegetation, higher ambient temperatures, much more precipitation, specific geomorphology, etc.) would favor greater use of sounds with lower frequencies^[13]. Besides, this environmental influence would happen in a much shorter time-span than the interval between “speciation events”^[14, 15].

Theoretically, the early stage of human being’s evolution permits an environmental situation in which vowels are not essential (full opening of mouth is not beneficial or not allowed). For example, desiccation, chilliness or many predators around, all these situations promote phonological formation of consonant prone languages. One language example is Hebrew, with 22 consonants and no vowels. The historical potential reasons may contain important societal event(s), but more probable is natural environmental situation at the early phase of Hebrew language. More and more studies indicated that low temperature and dryness significantly influence vowel usage^[10–12]. The effectiveness of different phoneme profile is determined by anatomical structure of a phonatory organ, and the latter is determined by genes. So by postulation, geographical conditions shall be also influential in the formation of language gene polymorphism during the long-term evolution time.

Of course, human languages, especially modern human language, often contain a large part of artificially designed extent in the course of standardization, and the newly designed part may have nothing to do with any geographical even historical influence. But the geographical influence on historical adaptation reflected on the local language gene polymorphism shall leave distinctive mark. Besides, natural selection happens in both natural (geographical) environment and cultural (societal) environment, though in the early stage of human civilization more selective adaptation may be implemented through geographical factors. Geographical factors bring driving force to human population, while society factors may influence individuals in a human population with adaptive pressure no less than geographical factors do^[16–17].

This study focused on correlation analysis among a total 111 single nucleotide polymorphisms (SNPs) from 13 language genes and a series of geography/society factors collected from 26 countries, and the author did find some interesting correlational parameters, including several strongest positive and negative correlations.

Methods

Language genes and their SNPs

Language is an emergent complicated function of human being, though many other animals also have their own ‘languages’. If a gene mutation is statistically or experimentally associated with a certain language function loss, it would be called language gene. Language gene SNP data were all randomly selected for each gene in the dbSNP database: <https://www.ncbi.nlm.nih.gov/snp/>; Table 1 listed 13

language genes, and a total 111 SNPs from these 13 genes were randomly selected for this study (Table 2, Supplementary file-1). Quantification of SNPs was described in Supplementary file-2.

Table 1
Thirteen language genes

	Gene name	Encoded protein	Compromised ability when mutated (example)
1	FOXP1	Forkhead box protein P1;	Expressive language ^[18]
2	FOXP2	Forkhead box protein P2;	Speech ^[19]
3	CNTNAP2	Contactin-associated protein-like 2	Early language development ^[20–22]
4	TPK1	Thiamin pyrophosphokinase 1	Syntactic and lexical ability ^[23–24]
5	DCDC2	Doublecortin domain-containing protein 2	Reading, dyslexia ^[25–27]
6	KIAA0319	Dyslexia-associated protein KIAA0319	Reading, dyslexia ^[28, 20, 25, 29]
7	TM4SF20	Transmembrane 4 L6 family member 20	Language delay; communication disorder ^[30]
8	FLNC	Filamin-C; Muscle-specific filamin	Reading, language ^[31]
9	ATP2C2	Calcium-transporting ATPase type 2C member 2	Memory ^[32]
10	ROBO1	Roundabout homolog 1;	Phonological buffer ^[33–34]
11	ROBO2	Roundabout homolog 2;	Expressive vocabulary ^[35]
12	CMIP	C-Maf-inducing protein	Reading, memory ^[20, 25, 32]
13	NFXL1	NF-X1-type zinc finger protein NFXL1;	Speech ^[36]

Table 2
111 SNPs of thirteen language genes

	Variables	Abbreviation		Variables	Abbreviation
1	ROBO1 rs34841026	ROBO1	57	DCDC2 rs33943110	DCD-4
2	ROBO2 rs11127602	ROBO-1	58	DCDC2 rs33914824	DCD-5
3	ROBO2 rs10865561	ROBO-2	59	DCDC2 rs9467075	DCD-6
4	ROBO2 rs5788280	ROBO-3	60	DCDC2 rs9460973	DCD-7
5	ROBO2 rs3923745	ROBO-4	61	DCDC2 rs3846827	DCD-8
6	ROBO2 rs3923744	ROBO-5	62	DCDC2 rs3789219	DCD-9
7	ROBO2 rs1163750	ROBO-6	63	CNTNAP2 rs1637842	CNTN-1
8	ROBO2 rs1163749	ROBO-7	64	CNTNAP2 rs1637841	CNTN-2
9	ROBO2 rs1163748	ROBO-8	65	CNTNAP2 rs1479837	CNTN-3
10	ROBO2 rs1031377	ROBO-9	66	CNTNAP2 rs1468370	CNTN-4
11	TM4SF20 rs6724955	TM1	67	CNTNAP2 rs1062072	CNTN-5
12	TM4SF20 rs44675173	TM2	68	CNTNAP2 rs1062071	CNTN-6
13	TM4SF20 rs4675172	TM3	69	CNTNAP2 rs987456	CNTN-7
14	TM4SF20 rs4673192	TM4	70	CNTNAP2 rs700309	CNTN-8
15	TM4SF20 rs4438464	TM5	71	CNTNAP2 rs700308	CNTN-9
16	TM4SF20 rs4428010	TM6	72	CNTNAP2 rs3194	CNTN-10
17	TM4SF20 rs4408717	TM7	73	CMIP rs201316817	CMI-1
18	TPK1 rs113536847	TPK-1	74	CMIP rs183876152	CMI-2
19	TPK1 rs79464600	TPK-2	75	CMIP rs183075361	CMI-3
20	TPK1 rs77358162	TPK-3	76	CMIP rs114894868	CMI-4
21	TPK1 rs28380423	TPK-4	77	CMIP rs79979027	CMI-5
22	TPK1 rs17170295	TPK-5	78	CMIP rs74031247	CMI-6
23	TPK1 rs12333969	TPK-6	79	CMIP rs60152409	CMI-7
24	TPK1 rs6953807	TPK-7	80	CMIP rs57603843	CMI-8
25	NFXL1 rs1964425	NFX-1	81	CMIP rs35429777	CMI-9
26	NFXL1 rs1822030	NFX-2	82	CMIP rs34119643	CMI-10

Variables			Abbreviation			Variables			Abbreviation		
27	NFXL1	rs1822029	NFX-3	83		ATP2C2	rs78371901	ATP-1			
28	NFXL1	rs1812964	NFX-4	84		ATP2C2	rs74038217	ATP-2			
29	NFXL1	rs1545200	NFX-5	85		ATP2C2	rs62640935	ATP-3			
30	NFXL1	rs1440228	NFX-6	86		ATP2C2	rs62640932	ATP-4			
31	NFXL1	rs1371730	NFX-7	87		ATP2C2	rs62640931	ATP-5			
32	NFXL1	rs1036681	NFX-8	88		ATP2C2	rs62050917	ATP-6			
33	NFXL1	rs978094	NFX-9	89		ATP2C2	rs16973859	ATP-7			
34	NFXL1	rs920462	NFX-10	90		ATP2C2	rs13334642	ATP-8			
35	FXP1	rs7638391	FXP1	91		ATP2C2	rs4782970	ATP-9			
36	FOXP1	rs76145927	FOXP1-1	92		ATP2C2	rs4782948	ATP-10			
37	FOXP1	rs75214049	FOXP1-2	93		KIAA0319	rs138160539	KIA-1			
38	FOXP1	rs17008544	FOXP1-3	94		KIAA0319	rs117692893	KIA-2			
39	FOXP1	rs17008063	FOXP1-4	95		KIAA0319	rs114195393	KIA-3			
40	FOXP1	rs11914627	FOXP1-5	96		KIAA0319	rs699461	KIA-4			
41	FOXP1	rs7639736	FOXP1-6	97		KIAA0319	rs699462	KIA-5			
42	FOXP1	rs1499893	FOXP1-7	98		KIAA0319	rs699463	KIA-6			
43	FOXP1	rs1053797	FOXP1-8	99		KIAA0319	rs730860	KIA-7			
44	FLNC	rs2291569	FLN-1	100		KIAA0319	rs10946705	KIA-8			
45	FLNC	rs2291568	FLN-2	101		KIAA0319	rs75674723	KIA-9			
46	FLNC	rs2291566	FLN-3	102		KIAA0319	rs75720688	KIA-10			
47	FLNC	rs2291565	FLN-4	103		FOXP2	rs10227893	FOXP2-1			
48	FLNC	rs2291563	FLN-5	104		FOXP2	rs10244649	FOXP2-2			
49	FLNC	rs2291562	FLN-6	105		FOXP2	rs12705977	FOXP2-3			
50	FLNC	rs2291561	FLN-7	106		FOXP2	rs61732741	FOXP2-4			
51	FLNC	rs2291560	FLN-8	107		FOXP2	rs61758964	FOXP2-5			
52	FLNC	rs2291558	FLN-9	108		FOXP2	rs62640396	FOXP2-6			
53	FLNC	rs2249128	FLN-10	109		FOXP2	rs73210755	FOXP2-7			
54	DCDC2	rs35029429	DCD-1	110		FOXP2	rs1058335	FOXP2-8			

Variables			Abbreviation		
55	DCDC2 rs2274305	DCD-2	111	FOXP2 rs61753357	FOXP2-9
56	DCDC2 rs34584835	DCD-3			

Collection of geography/society parameters

Geography/society parameters were collected from three websites, including Baidu (<https://baike.baidu.com>), Bing (<https://cn.bing.com>), and United Nations databases (UND) (<http://data.un.org/Default.aspx>). (Supplementary file-3) Detailed data are not provided in this manuscript because of page limitation, but can be requested from the corresponding author.

Table 3 Geography/society parameters

	Variables (in number)	Abbreviation	Data source
1	Active duty army(10^4)	army	Baidu, Bing
2	Agriculture,forstry, husbandry and fishery	agrc	Baidu, Bing
3	Airport	airp	Baidu, Bing
4	Annual average rainfall (mm)	aash	Baidu, Bing
5	Area of the country (km^2)	area	Baidu, Bing
6	Average GDP per person	agdp	Baidu, Bing
7	City number with population over 100000	cnpl	UND
8	Country and region for exportation	ceex	Baidu, Bing
9	Country and region for importation	crim	Baidu, Bing
10	Country/regions that speak the same language	regi	Baidu, Bing
11	Cultivated land area (10^4 hectare)	cula	Baidu, Bing
12	Fertility rate, total (births per woman)	fert	UND
13	Forest coverage (%)	fore	Baidu, Bing
14	Geographical neighbor country	gnei	Baidu, Bing
15	Geographical resource (minerals)	gres	Baidu, Bing
16	Geomorphology feature type	geft	Baidu, Bing
17	Human development index	hdi	Baidu, Bing
18	Industry type	indu	Baidu, Bing
19	Main exported product	mepr	Baidu, Bing
20	Main imported product	mipr	Baidu, Bing
21	Main river	mrvr	Baidu, Bing
22	Mortality rate, infant (per 1,000 live births)	mrta	UND
23	Population aged 65 years or older (thousands)	pold	UND
24	Population density of the country	pden	Baidu, Bing
25	Population in the sample country	popu	Baidu, Bing
26	Port	port	Baidu, Bing
27	Race in the country	race	Baidu, Bing

	Variables (in number)	Abbreviation	Data source
28	Religion in the country	relg	Baidu, Bing
29	Road (10 ⁴ km)	road	Baidu, Bing
30	Runway traffic mode	rway	Baidu, Bing
31	Rural population (% of total population)	rupo	UND
32	Tourist/visitor arrivals(2017)(Thousands)	tova	UND
34	Weather feature type	weft	Baidu, Bing

Selected countries with decent representation

This study is a part of our larger scale one that has to collect much more parameters (language gene single nucleotide polymorphism or SNPs, geographical factors, societal factors, etc.) except for education/culture parameters. Besides 34 geography/society parameters (Table 3), we are collecting 111 (SNPs) + 61 (Education/ Culture/ Geography etc.) = 172 parameters in total for around 150 countries, and only 26 countries have been collected for all 172 parameters by the time this manuscript is written. The names of the countries are Japan, China, Mongolia, Vietnam, Nepal, Myanmar, Bhutan, Sri Lanka, Bangladesh, Iran, Slovenia, Greece, Bosnia and Herzegovina, Benin, Cameroon, Gambia, Kenya, Senegal, Tunisia, Tanzania, Barbados, Dominican, Peru, France, UK, and Germany. Though only 26 in number, the name list contains descent representation, including one eastern (Asia) developed country, three western developed countries, two developing countries in eastern Asia, three ordinary European countries, seven African countries, six ordinary countries in south Asia, and several typical ones in west Asia (one) and Middle America (three). So the results in this study would be preliminary but valuable.

Correlation analysis

The software Origin was employed to analyze the potential correlation relationship between all parameters. By using Origin-2019, principal components analysis (PCA) was performed to extract potential correlation patterns among multiple numerical variables. PCA is a feasible technique to emphasize variation and visualize strong patterns in a large dataset. In a typical PCA diagram, correlated variables are drawn as short or long arrows in which long arrow represents strong correlation and short arrow represents weak correlation; plus, two arrows can form a right, obtuse or acute angle, representing no-correlation, negative correlation or positive correlation, respectively. Quantitative correlation between any two factors was undertaken as follow. The basic variables to measure in a PCA result plot are arrow length and the angle between two arrows. The correlation score = angle score × arrow score; the angle score ranges from -5 to +5. The angles (0-15], (15-30], (30-45], (45-60], (60-75], (75-105], (105-120], (120-135], (135-150], (150-165], (165-180] are scored 5, 4, 3, 2, 1, 0, -1, -2, -3, -4, -5, respectively. The arrow score =

the length of arrow-1× the length of arrow-2. The angle value and the arrow length value can be conveniently obtained with ImageJ software (Supplementary file-4, Supplementary file-5).

Results And Discussion

General correlations among all selected geography parameters

Figure 1 showed the general correlations among all selected 34 geography/society parameters, in which geographical factors themselves have specific intra-correlation characteristics. First, most geographical parameters are concentrated into one cluster. This cluster contains 12 parameters that correlate positively with each other: aash, cula, gres, mrvr, area, road, gnei, airp, port, geft, rway, and agrc. Only one parameter, fore, is not in the cluster. On the contrary, the fore parameter has a negative correlation with all the above 12 parameters. Second, in the above mentioned cluster, the port parameter has least correlation with other parameters, which is somehow unexpected. But it is relatively easy to understand why the fore parameter negatively correlated with all other geographical parameters. The more forest, the less human activity, the less values of those human activity-related geographical factors.

General correlations among all selected society parameters

When we look at correlation profiles among the 21 society parameters, we found that they are basically divided into four clusters: cluster 1 (popu, army, pold, cnpl, indu, mepr, relg, mipr, weft, tova, aash), cluster 2 (hdi, agdp, pden), cluster 3 (mrta, fert, regi, rupo) and cluster 4 (crim, ceex, race). Cluster 1 has 11 parameters that positively correlate with each other; cluster 1 has little correlations with cluster 2 and cluster 4, but it has a strong negative correlation with cluster 3. Cluster 2 and cluster 3 has a more negative correlation. Cluster 2 contains hdi (Human development index) and agdp (Average GDP per person). This means the more level of economic development, the less values of those in cluster 3. Economy can decrease the diversity of human society.

Correlations between geography and society parameters

The hdi and agdp parameters in the cluster 2 have 8-9 negatively correlated factors: fore (Forest coverage), rupo (Rural population), regi (Country/regions that speak the same language), mrta (Mortality rate), fert (Fertility rate), crim (Country and region for importation), ceex (Country and region for exportation), race (Race in the country) and aash (Annual average rainfall), in which fore, regi and aash belong to geographical factors. So hdi and agdp also have negative correlations with some geographical parameters (though the correlation level is not high), which means economic progress also diminishes the diversity level of geographical elements.

There are two parameters, port and fore, specially positioned in all PCA diagrams (supplementary file 4). These two parameters are strongly negative with each other, and neither have strong positive correlations with most other geography/society parameters. There are nine parameters negatively correlating with port: pden, fore, rupo, mrta, fert, regi, crim, ceex and race. All other parameters either positively correlate, or

have little correlations, with port. So the parameter port influences the geography/society environment, though not strongly, but broadly.

Correlation between a specific gene SNP and Geography/society parameters

PCA analysis was undertaken for 13 language genes one by one with 34 geography/society parameters. Each language gene gives around 10 different SNPs (see SNP number in Table 2). All PCA results were quantified and demonstrated in Figure 2.

For the 1-10 SNPs of each gene, most appear in all four quadrants in the four-quadrant PCA diagram at the same time, though the SNPs of TM1-7 only appear in the 2nd and 4th quadrants (Supplementary file-4). No all SNPs from a single gene stay aggregated in a corner of the PCA map (only in one quadrant) as expected; because all genes passed through million years of mutual adaptation; if all SNPs from a single gene stay aggregated, that would mean that many factors or SNPs from other genes counteract with them. Such a single gene would likely be lost during evolution.

The strongest positive correlations were seen at (ATP-1~army, ATP-1~gres, ATP-1~pold, ATP-1~popu, and ATP-1~road). ATP-1 (rs78371901) is one of the SNPs of language gene ATP2C2; this gene encodes the ATPase secretory pathway Ca²⁺ transporting-2 protein. Diseases associated with ATP2C2 include specific language impairment and some oral communication disorders. The army (Active duty army), gres (Geographical resource), pold (Population aged 65 years or older), popu (Population in the sample country) and road (Road) contain only two geographical factors: gres and road. The gres mainly represents natural resources such as different types of mineral resources.

The strongest negative correlations were seen at (NFX-6~area, NFX-6~army, NFX-6~gres, NFX-6~mrwr, and NFX-6~road). NFX-6 (rs1440228) is one of SNPs of NFXL1 gene. NFXL1 encodes a Nuclear Transcription Factor (X-Box Binding-Like 1). Gene Ontology annotations related to this gene include DNA-binding transcription factor activity and proximal promoter DNA-binding transcription repressor activity, plus RNA polymerase II-specific activity. It is associated with a disease of Specific Language Impairment. The area (Area of the country), army (Active duty army), gres (Geographical resource), mrwr (Main river) and road contain four geographical factors. Interestingly, gres and road and involved both the strongest positive and strongest negative correlations.

Four geography/society parameters demonstrated least correlations with language gene SNPs (Figure 2C), and they are aash (Annual average rainfall), fore (Forest coverage), pden (Population density of the country) and rway (Runway traffic mode); Another several parameters demonstrated second least correlations with language gene SNPs, and they are ceex (Country and region for exportation), crim (Country and region for importation), agrc (Agriculture, forestry, husbandry and fishery) and relg (Religion in the country).

In Figure 2C, there is another interesting point. For each geography/society parameter, the number of SNPs with positive correlations with it is almost the same as the number of SNPs with negative

correlations with it. That suggests that each parameter is coincidentally balanced by similar numbers of language gene SNPs with opposite correlations with it.

Discussion

In this study, the basic data include 13 language genes and their randomly selected 111 single nucleotide polymorphisms (SNPs), SNP profiles in 26 countries, and 34 geography/society parameters in 26 countries. In order to undertake principal component analysis (PCA), SNP genotypes and all geography/society parameters have to be quantitatively represented into numerical values.

Why these parameters in Table 3? Actually this manuscript only contains very general geography/society parameters, and is apparently lacking in detailed description of geography/society features for selected countries. More geographical parameters needed in the future, such as temperature, altitude, rugosity, some googlemap factors, etc. The number of lowest altitude, number of highest altitude, river direction—mountain orientation—precipitation amount, precipitation type, humidity, etc. Also, parameter combinations shall be described in the future because synergistic effects have been rarely considered.

One of the limitations in this study is the ‘data age’. Only those factors representing geography/society features in the early stage of human evolution can leave significant marks in language gene polymorphisms. Though modern geography/society factors still work for language gene polymorphism’s evolution, modern time duration is too short for the whole history of human being. Influence of modern time on language gene SNPs may be only steadily seen after a long time in the future. Data in Table 3 are mostly modern parameters; only a few parameters can last long time and have few changes, such as Geomorphology feature type. But even Geomorphology feature may change a lot after several thousand years.

A limitation also exists in the data of language gene polymorphisms. At present, human genome sequencing samples are not balanced for different countries, and many developing countries contribute much less to genome sequence data than developed countries. So data from poor or most developing countries have been not representative enough.

In PCA analysis, the strongest positive correlations were seen as (ATP-1~army, ATP-1~gres, ATP-1~pold, ATP-1~popu, and ATP-1~road); the strongest negative correlations were seen as (NFX-6~area, NFX-6~army, NFX-6~gres, NFX-6~mrvr, and NFX-6~road); the least correlations with language gene SNPs were found at aash (Annual average rainfall), fore (Forest coverage), pden (Population density of the country) and rway (Runway traffic mode). At this moment, we are actually not sure what the above correlations mean. What messages they convey needs further future studies with more sophisticated data.

Potential interaction between ATP2C2 and NFXL1 genes may be worth tackling since both hold strongest correlations with geography/society parameters^[37]. By now there is no any report on their potential interaction, though bioinformatics search can generate some hints (Figure 3). Figure 2 illustrates known

interactions among proteins encoded by 13 language genes, in which most interactions have not been experimentally confirmed. ROBO1 has interactions with other eight genes; FOXP1, CMIP, CNTNAP2 and ATP2C2 have interactions with seven genes, but TPK1, FLNC and TM4SF20 have not been found to interact with other language genes. As for the interaction between ATP2C2 and NFXL1, known information is little because only one interaction between the two genes exists as text-mining nature (Figure 3). But both are co-expressed with DCDC2 gene, which is associated with reading disability and modulates neuronal development in the brain [38]. By postulation, if a protein interacts with more other proteins, its polymorphisms will have more influential biological consequences, thus have more chances to have correlations with more other proteins. To note, 13 language genes almost all are involved in human learning capacity development, thus participate the determination of one's ability to interact with geography/society factors; so those SNPs that have most or strongest correlations with geography/society factors may also hold stronger correlations with other language genes' SNPs. This point will be investigated using more language gene SNPs in the near future.

This study is based on one assumption: geography/society factors directly influence acoustic adaptation of *Homo sapiens*, and the latter is reflected on gradual or sudden optimization of anatomical structure of a phonatory organ in a local geographical/societal environment, while continuous language gene mutational evolution directly facilitates the above optimization process. So in the time scale of human being evolution, there should be meaningful correlations between some geography/society factors and human language gene polymorphisms. Such a 'meaningful' correlation will be more manifested only in a condition that more sophisticated data and better pattern recognition algorithms are both ready for use. On the other hand, this study is also highly dependent on one question: what is the relationship between anatomical structure of a phonatory organ and language gene itself? This relationship could be direct or indirect. If it is direct, then functional optimization of a phonatory organ will be easily reflected on the change in language gene SNP profiles, and such a correlation will be relatively easy to pinpoint between geographical/societal factors and language gene SNPs; if it is indirect, the expected correlation will be much harder to find. By now there have been already a few studies on the proteomic profiles of (human) vocal tissue/organs, but proteins in Table 1 were not found in these studies only except for FLNC [41–44]. Biological studies on this matter in the future 5-10 years are greatly expected for providing a help.

Conclusions

This study tackles the question whether there is some correlation between human language gene polymorphisms and geography-society parameters collected from twenty-six countries. By the method of principal component analysis, the main correlations within geography parameters, within society parameters, between geography and society parameters, and between language gene SNPs and geography-society parameters, were analyzed. This study obtained its preliminary results, the main points of which contain: (1) most geographical parameters are concentrated into one cluster. This cluster contains 12 parameters that correlate positively with each other; (2) society parameters are divided into four clusters in the PCA diagrams and among them exist positive and negative correlations; (3) the

strongest positive correlations were seen as (ATP-1~army, ATP-1~gres, ATP-1~pold, ATP-1~popu, and ATP-1~road); the strongest negative correlations were seen as (NFX-6~area, NFX-6~army, NFX-6~gres, NFX-6~mrwr, and NFX-6~road); the least correlations with language gene SNPs were observed at aash, fore, pden and rway. But the meanings of such correlations need time to decipher in the future.

The authors have to admit that this study still has several points to improve as follow. In this study, whole countries are not always an appropriate unit of analysis for language gene polymorphisms and their correlations, especially when one considers very large and arguably heterogeneous ones such as "France" or "China". All twenty-six countries have language sampling points as shown in the WALS database^[45], but it is not clear whether the SNP data fit those geographical points. Besides, each language gene has hundreds of DNA polymorphism variants and most of them have not been functionally characterized. Later on we may have to focus on a few SNPs that have been linked with language disorders^[46, 47]. Importantly, GALTON'S problem raises questions about the nature of explanation in cross-national research^[48], and this problem likely exists in this version manuscript. For more than a decade it has been standard in such studies to include controls at least for Galton's problem and for country contact. Careful steps have to take in order to solve the above problems later on.

Declarations

ACKNOWLEDGMENTS

This study was supported by State Language Commission Research Grant (YB135-117) and National Research Center for Foreign Language Education Grant (ZGWYJYJJ10A042). The authors thank Nie Qi, Pan Yusheng for their help in collecting sample data.

Conflicts of interest

The authors declare no potential conflicts of interest for this study.

References

1. Nothnagel M, Lu TT, Kayser M, Krawczak M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Hum Mol Genet*,2010,19(15):2927–2935.
2. Shen X. Study on genetic and geographical factors of non-syndromic cleft lip and palate disease. Lanzhou University,2011.
3. Munroe, R. L.,M. Silander. Climate and the consonant-vowel (CV) syllable: Replication within language families. *Cross-Cultural Research*,1999,33: 43–62.
4. Urban M, Moran S. Altitude and the distributional typology of language structure: Ejectives and beyond. *PLoS One*, 2021, 16(2):e0245522.

5. Blasi DE, Moran S, Moisik SR, Widmer P, Dediu D, Bickel B. Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science*, 2019, 363(6432):eaav3218.
6. Maddieson I. Language Adapts to Environment: Sonority and Temperature. *Front*, 2018, 3:28.
7. Krause J, Lalueza-Fox C, Orlando L, et al. The derived FOXP2 variant of modern humans was shared with Neandertals. *Curr Biol*, 2007, 17(21):1908–1912.
8. Dennell R, Roebroeks W. An Asian perspective on early human dispersal from Africa. *Nature*, 2005, 438(7071):1099–1104.
9. Dennell R. Dispersal and colonisation, long and short chronologies: how continuous is the Early Pleistocene record for hominids outside East Africa? *J Hum Evol*, 2003, 45(6):421–440.
10. Maddieson I, C Coupé. Human spoken language diversity and the acoustic adaptation hypothesis. *Acoust*, 2015, 138:1838.
11. Wilkins MR, Seddon N, Safran RJ. Evolutionary divergence in acoustic signals: causes and consequences. *Trends Ecol Evol*, 2013, 28(3):156–166.
12. Everett C. Languages in Drier Climates Use Fewer Vowels. *Front Psychol*, 2017, 8:1285.
13. Maddieson I, C Coupé. Human language diversity and the acoustic adaptation hypothesis. *Proceedings of Meetings on Acoustics* [Acoustical Society of America 5th Pacific Rim Underwater Acoustics Conference - Vladivostok, Russia (23-26 September 2015)], 2016, 25:060005.
14. Yoder AD, Olson LE, Hanley C, et al. A multidimensional approach for detecting species patterns in Malagasy vertebrates. *Proc Natl Acad Sci U S A*, 2005, 102(SI):6587-6594.
15. Manjil Hazarika. HOMO ERECTUS/ERGASTER AND OUT OF AFRICA: RECENT DEVELOPMENTS IN PALEOANTHROPOLOGY AND PREHISTORIC ARCHAEOLOGY. *EAA Summer School eBook*, 2007, 1: 35–41.
16. Stein MB, Chen CY, Jain S, et al. Genetic risk variants for social anxiety. *Am J Med Genet B Neuropsychiatr Genet*, 2017, 174(4):470–482.
17. Hollister BM, Farber-Eger E, Aldrich MC, Crawford DC. A Social Determinant of Health May Modify Genetic Associations for Blood Pressure: Evidence From a SNP by Education Interaction in an African American Population. *Front Genet*, 2019, 10:428.
18. Bacon C, Rappold GA. The distinct and overlapping phenotypic spectra of FOXP1 and FOXP2 in cognitive disorders. *Hum Genet*, 2012, 131(11):1687–1698.
19. Lai CS, Fisher SE, Hurst JA, et al. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*, 2001, 413(6855):519–523.
20. Newbury DF, Paracchini S, Scerri TS, et al. Investigation of dyslexia and SLI risk variants in reading- and language-impaired subjects. *Behav Genet*, 2011, 41(1):90–104.
21. Vernes SC, Newbury DF, Abrahams BS, et al. A functional genetic link between distinct developmental language disorders. *N Engl J Med*, 2008, 359(22):2337–2345.
22. Whitehouse AJ, Bishop DV, Ang QW, et al. CNTNAP2 variants affect early language development in the general population. *Genes Brain Behav*, 2011, 10(4):451–456.

23. Villanueva P, Newbury DF, Jara L, et al. Genome-wide analysis of genetic susceptibility to language impairment in an isolated Chilean population. *Eur J Hum Genet*,2011,19(6):687–695.
24. Fattal I, Friedmann N, Fattal-Valevski A. The crucial role of thiamine in the development of syntax and lexical retrieval: a study of infantile thiamine deficiency. *Brain*,2011,134(6):1720–1739.
25. Scerri TS, Morris AP, Buckingham LL, et al. DCDC2, KIAA0319 and CMIP are associated with reading-related traits. *Biol Psychiatry*,2011,70(3):237–245.
26. Deffenbacher KE, Kenyon JB, Hoover DM, et al. Refinement of the 6p21.3 quantitative trait locus influencing dyslexia: linkage and association analyses. *Hum Genet*,2004,115(2):128–138.
27. Schumacher J, Anthoni H, Dahdouh F, et al. Strong genetic evidence of DCDC2 as a susceptibility gene for dyslexia. *Am J Hum Genet*,2006,78(1):52–62.
28. Paracchini S, Thomas A, Castro S, et al. The chromosome 6p22 haplotype associated with dyslexia reduces the expression of KIAA0319, a novel gene involved in neuronal migration. *Hum Mol Genet*,2006,15(10):1659–1666.
29. Francks C, Paracchini S, Smith SD, et al. A 77-kilobase region of chromosome 6p22.2 is associated with dyslexia in families from the United Kingdom and from the United States. *Am J Hum Genet*,2004,75(6):1046–1058.
30. Wiszniewski W, Hunter JV, Hanchard NA, et al. TM4SF20 ancestral deletion and susceptibility to a pediatric disorder of early language delay and cerebral white matter hyperintensities. *Am J Hum Genet*,2013,93(2):197–210.
31. Gialluisi A, Newbury DF, Wilcutt EG, et al. Genome-wide screening for DNA variants associated with reading and language traits. *Genes Brain and Behavior*,2014,13(7): 686–701.
32. Newbury FD, Winchester L, Addis Paracchini S, et al. CMIP and ATP2C2 modulate phonological short-term memory in language impairment. *American Journal of Human Genetics*,2009,85(2): 264–272.
33. Hannula-Jouppi K, Kaminen-Ahola N, Taipale M, et al. The Axon Guidance Receptor Gene ROBO1 Is a Candidate Gene for Developmental Dyslexia. *Plos Genetics*,2005,1(4): e50.
34. Bates TC, Luciano M, Medland SE, et al. Genetic variance in a component of the language acquisition device: ROBO1 polymorphisms associated with phonological buffer deficits. *Behavior Genetics*,2011,41(1): 50–57.
35. Pourcain BS, Cents R, Whitehouse A, et al. Common variation near ROBO2 is associated with expressive vocabulary in infancy. *Nature Communications*,2014,5: 4831.
36. Villanueva P, Nudel R, Hoischen A, et al. Exome Sequencing in an Admixed Isolated Population Indicates NFXL1 Variants Confer a Risk for Specific Language Impairment. *PLoS Genetics*,2015,11(3): e1004925.
37. Chen XS, Reader RH, Hoischen A, et al. Next-generation DNA sequencing identifies novel gene variants and pathways involved in specific language impairment. *Sci Rep*. 2017;7:46105.
38. Meng H, Smith SD, Hager K, et al. DCDC2 is associated with reading disability and modulates neuronal development in the brain. *Proc Natl Acad Sci USA*. 2005;102(47):17053–17058.

39. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015;43(Database issue):D447-452.
40. Chen SJ, Liao DL, Chen CH, et al. Construction and Analysis of Protein-Protein Interaction Network of Heroin Use Disorder. *Sci Rep.* 2019;9(1):4980.
41. Welham NV, Chang Z, Smith LM, Frey BL. Proteomic analysis of a decellularized human vocal fold mucosa scaffold using 2D electrophoresis and high-resolution mass spectrometry. *Biomaterials.* 2013;34(3):669–676.
42. Gugatschka M, Darnhofer B, Grossmann T, et al. Proteomic Analysis of Vocal Fold Fibroblasts Exposed to Cigarette Smoke Extract: Exploring the Pathophysiology of Reinke's Edema. *Mol Cell Proteomics.* 2019;18(8):1511–1525.
43. Welham NV, Yamashita M, Choi SH, Ling C. Cross-sample validation provides enhanced proteome coverage in rat vocal fold mucosa. *PLoS One.* 2011;6(3):e17754.
44. Karbiener M, Darnhofer B, Frisch MT, et al. Comparative proteomics of paired vocal fold and oral mucosa fibroblasts. *J Proteomics.* 2017;155:11–21.
45. Dryer, M. S., and Haspelmath, M. (2016). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
46. Mohammad Sarowar Uddin et al. CNTNAP2 gene polymorphisms in autism spectrum disorder and language impairment among Bangladeshi children: a case-control study combined with a meta-analysis. *Hum Cell.* 2021 Sep;34(5):1410–1423.
47. Siyun Zhang et al. The Association Between Genetic Variation in FOXP2 and Sensorimotor Control of Speech Production. *Front Neurosci.* 2018 Sep 20;12:666.
48. GALTON'S PROBLEM IN CROSS-NATIONAL RESEARCH. By MARC HOWARD ROSS and ELIZABETH HOMER

Figures

Figure 1

PCA for language gene SNPs and geography/society parameters. (A) FOXP2 SNPs and 34 geography/society parameters; (B) FOXP1 SNPs and geography/society parameters; (C) ATP2C2 SNPs and geography/society parameters; (D) NFXL1 SNPs and geography/society parameters.

(A)

(B)

(C)

Figure 2

Correlation among language gene SNPs and 34 geography/society parameters. (A) Five strongest positive correlations were shown; (B) Five strongest negative correlations were shown; (C) Smallest correlations were demonstrated clearly at four geography/society parameters (aash, fore, pden and rway).

Figure 3

Interactions among 13 language genes depicted with STRING [39-40]. Detailed figure legends can be seen in the STRING website (<https://string-db.org/cgi/>). The line edge colors represent gene co-occurrence (blue), from curated databases (yellow blue), text-mining (yellow), co-expression (black), experimentally determined (pink), gene fusions (red), gene neighborhood (green) and protein homology (shallow blue), respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementaryfilesforLZXpaper.zip](#)