Brain & Language 174 (2017) 1–8

Contents lists available at ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

Short communication

Axon guidance pathways served as common targets for human speech/language evolution and related disorders

Huimeng Lei [a,*,1], Zhangming Yan [b,1], Xiaohong Sun [a], Yue Zhang [c], Jianhong Wang [c], Caihong Ma [d], Qunyuan Xu [a], Rui Wang [e], Erich D. Jarvis [f,g], Zhirong Sun [b,*]

[a] Department of Neurobiology, Beijing Institute for Brain Disorders, Beijing Center of Neural Regeneration and Repair, Key Laboratory for Neurodegenerative Diseases of the Ministry of Education, Capital Medical University, Beijing 100069, China
[b] MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China
[c] Department of Children Healthcare, Capital Institute of Pediatrics, Beijing, 100020, China
[d] Reproductive Medicine Center of Peking University Third Hospital, Beijing, 100191, China
[e] Hengkuan Telegenomics Co., Ltd., 36/F, 5 Meiyuan Rd., Tianjin 300384, China
[f] Department of Neurobiology, Duke University Medical Center, Durham, NC 27710, USA
[g] Howard Hughes Medical Institute, Chevy Chase, MD, 20815-6789, USA

ARTICLE INFO

Article history:
Received 22 August 2016
Revised 17 May 2017
Accepted 29 June 2017

Keywords:
Genomic convergence
PRAME
ROBO1
Axon guidance
Human speech/language

ABSTRACT

Human and several nonhuman species share the rare ability of modifying acoustic and/or syntactic features of sounds produced, i.e. vocal learning, which is the important neurobiological and behavioral substrate of human speech/language. This convergent trait was suggested to be associated with significant genomic convergence and best manifested at the *ROBO-SLIT* axon guidance pathway. Here we verified the significance of such genomic convergence and assessed its functional relevance to human speech/language using human genetic variation data. In normal human populations, we found the affected amino acid sites were well fixed and accompanied with significantly more associated protein-coding SNPs in the same genes than the rest genes. Diseased individuals with speech/language disorders have significant more low frequency protein coding SNPs but they preferentially occurred outside the affected genes. Such patients' SNPs were enriched in several functional categories including two axon guidance pathways (mediated by netrin and semaphorin) that interact with ROBO-SLITs. Four of the six patients have homozygous missense SNPs on *PRAME* gene family, one youngest gene family in human lineage, which possibly acts upon retinoic acid receptor signaling, similarly as *FOXP2*, to modulate axon guidance. Taken together, we suggest the axon guidance pathways (e.g. *ROBO-SLIT*, *PRAME* gene family) served as common targets for human speech/language evolution and related disorders.

1. Introduction

While language is generally considered a unique component of humanity, its behavioral substrate, vocal learning, has been found in several distantly related nonhuman species (Doupe & Kuhl, 1999). Anatomical studies revealed that human and these nonhuman vocal learners share specialized corticostriatal loops and direct connections from motor cortical areas to brainstem vocal motor neurons, which are absent in vocal non-learners including non-human primates (Jarvis, 2004; Petkov & Jarvis, 2012). Comparative studies suggested that the above neural convergence in vocal learners is accompanied with certain genomic and transcriptomic convergence (Pfenning et al., 2014; Wang, 2011; Wang et al., 2015). Some revealed genes were shown to possibly participate in the development of vocal learning or even human speech and language. For example, *C4ORF21* as the only gene reported so far to exhibit accelerated evolution in human and other two mammalian vocal learners (elephants and microbats) but not their vocal non-learning relatives was found to be a candidate gene for childhood apraxia of speech (Peter et al., 2016; Wang, 2011).

Here we focused on the list of convergent amino acid changes shared by humans and other two mammalian vocal learners, i.e. the so-called *single non-random amino acid pattern* (SNAAP) changes in 73 genes (**in bold** hereafter) (Wang, 2011). Compared to those genetic changes associated with various aspects of human speech/language as reported by human patient studies, this list relates to genetic changes associated with a more definitive and

* Corresponding authors.
E-mail addresses: leihm@ccmu.edu.cn (H. Lei), sunzhr@mail.tsinghua.edu.cn (Z. Sun).
[1] These authors contributed equally to the paper as first authors.

http://dx.doi.org/10.1016/j.bandl.2017.06.007
0093-934X/

concrete neurobiological and behavioral feature of speech/language that differentiates human from other primates. Interestingly, genes with such amino acid convergence show strong ties to *ROBO-SLIT* axon guidance pathway (***ROBO1***, ***NEO1***, ***PITPNA***, ***CKAP5***, ***PCM1***, ***CEP192***, ***PTPRB*** and ***EFNB1***), which correlates with neural connectivity differences between vocal learners and non-learners (Pfenning et al., 2014; Wang, 2011; Wang et al., 2015) and the convergent expression patterns of ***ROBO1*** and its ligand *SLIT1* across humans and three orders of avian vocal learners (Pfenning et al., 2014; Wang, 2011; Wang et al., 2015) as well as the fact that human dyslexia and speech sound disorders have susceptible mutations in and near the ***ROBO1*** locus (Bates et al., 2011; Hannula-Jouppi et al., 2005). The differential expression of ***ROBO1*** and *SLIT1* is developmentally regulated in the song learning production brain region in juvenile zebra finches, a songbird, during its critical period for song learning (Wang, 2011; Wang et al., 2015). Notably, *SLIT1* serves a direct downstream target for the speech-language related gene *FOXP2*, the first gene whose mutation was shown to cause a Mendelian form of speech/language disorder (Konopka et al., 2009; Vernes et al., 2007). The roles of other genes with SNAAPs enriched in vocal learning mammals (***PARP1***, ***FRAS1***, ***GDAP1***, ***HAL***, ***E2F3***, ***CASP8AP2*** and ***USHBP1***) in speech-language, hearing and vocal production were reviewed in (Wang, 2011). ***WDFY3*** is recently suggested a causative disease gene for autism whose core symptoms include language and communication deficits (Orosco et al., 2014).

The above findings make it tempting to assess how such convergent amino acid changes were maintained during the population evolution of humans and relate to intra-species trait variations of speech/language. Here we were allowed to do so by the availability of human genetic variation data and by the ability to sequence exomes of individuals with different speech/language disorders. We found evidence for these convergent sites being highly fixed in human populations and having accumulated more associated protein coding SNPs in other parts of the same genes. These associated variations were also fixed and spread throughout human populations. Diseased individuals recruited significant more uncommon and presumably deleterious protein coding SNPs. These diseased patient SNPs do not preferentially occur in SNAAP genes; however, their affected genes are significantly overrepresented in functional categories that interact closely with *ROBO-SLIT* genes, e.g. the netrin axon guidance pathway and the semaphorin axon guidance pathways.

2. Results

To verify the significance of SNAAP sites, we downloaded the pre-calculated Genomic Evolutionary Rate Profiling (GERP) scores from UCSC database (http://hgdownload.cse.ucsc.edu/gbdb/hg19/bbi/All_hg19_RS.bw). GERP scores are site specific and measure evolutionary constraint using maximum likelihood evolutionary rate estimation, i.e. a greater score means greater evolutionary constraint inferred to be acting on that site (Davydov et al., 2010). These scores were calculated based on UCSC alignments of 35 mammals to human genome (hg19). We compared the mean GERP scores of the SNAAP sites and their surrounding 10 bp and 100 bp protein coding regions, respectively (Fig. 1A and B). Furthermore, we selected 100 random sets of protein coding sites from the reference 3256 orthologous gene sets and compared their average profile of mean GERP sores and those of surrounding 10 bp and 100 bp regions, respectively (Fig. 1C and D). We found the SNAAP sites have much lower GERP scores than surrounding regions, a phenomenon not seen in random sets of sites. The SNAAP sites are in regions that have stronger selective constraints (most GERP scores greater than 2) than the random sets of sites (most GERP scores smaller than 2). These results are specific to SNAAP sites and cannot be explained by their differences in overall GERP scores at SNAAP gene levels (Fig. 1E and F). These results are in line with the prior findings that the SNAAP sites experienced significant non-neutral selection and occurred in regions with unusually high negative selection (Wang, 2011).

2.1. Vocal learning SNAAPs are well conserved and have significantly more associated protein coding SNPs fixed throughout major human populations

SNP analyses of the 1000 Genomes Project (1KG) data on the 73 genes with SNAAP sites present in vocal learners revealed that only five of these sites had evolved SNPs in human populations (Supplementary Table 1). Three of the five SNPs (in genes ***C1ORF87***, ***SELP*** and ***TAF1C***) result in minor allelic amino acid types the same as found in monkey (Rhesus macaque), but are of extremely low frequencies in humans (MAF < 0.001). A fourth SNP (in ***PLEKHH1***) results in an allelic amino acid site the same as in chimpanzee, with a MAF = 0.479 in humans, supporting the conclusion in prior study that this site in PLEKHH1 is a false positive convergent site; therefore, we excluded this site for subsequent analysis. A fifth SNP in ***TMEM87B*** results in a deletion not present in either monkey or chimpanzee.

We also found seven human SNPs that are in the same codon of, but not at the SNAAP sites (Supplementary Table 1). All seven SNPs are rare in humans: one with MAF < 0.01 and six with MAF < 0.001. Furthermore, none of these SNPs results in minor allelic amino acid types to the same as in monkey and chimpanzee. Three of the seven SNPs are missense SNPs: in ***ROBO1*** A→V, where chimpanzee and monkey has T; ***C1ORF116*** K→Q, where monkey has R; and ***MROH2B*** I→V, where monkey has M. Notably, ***ROBO1***'s SNAAP site was in the 1st codon position, while the 2nd and 3rd positions of the same codon have evolved two SNPs in human populations yet with very low MAFs (0.0002 and 0.0004). These results suggest that the SNAAP site amino acid differences between human and non-human primates are fixed by strong negative selection in human populations so that the human amino acid types hardly mutate into the forms of their non-human primate relatives.

We compared the number of human SNPs across the protein coding sequence (cSNPs) of genes with SNAAP sites relative to random gene sets of equal sizes drawn from a reference set of 3256 genes where the SNAAP genes were identified (Wang, 2011). The SNAAP genes and random gene sets have comparable protein coding sequence lengths (Fig. 2A). We found the SNAAP genes had significantly more cSNPs (157.53 ± 18.57, mean ± std) than the random 100 gene sets (99.86 ± 17.52, mean ± std; one-tailed *t*-test, p-value < 1E−15; Fig. 2B). This excess of cSNPs is largely attributed to common cSNPs, i.e. cSNPs with MAF > 0.01 (Fig. 2C; one-tailed *t*-test, p-value < 1E−15; 147.79 ± 14.70 versus 93.78 ± 15.74, mean ± std), but not rare cSNPs, i.e. cSNPs with MAF < 0.01 (Fig. 2D; one-tailed *t*-test, p-value = 1; 6.08 ± 2.69 versus 5.42 ± 3.24, mean ± std). Taken together, these results suggest that the SNAAP genes in vocal learners have a stronger level of fixation at the SNAAP sites, but conversely have gained significantly more cSNPs in other parts of their protein-coding regions; these over-represented cSNPs have been well fixed throughout the human population.

2.2. Low frequency protein coding SNPs are enriched in patients with impaired speech/language abilities but do not preferentially occur in SNAAP genes

To test whether these SNAAP sites and other sites in the same genes could be associated with variations in the vocal learning trait, the exomes of a total of six patients with impaired speech

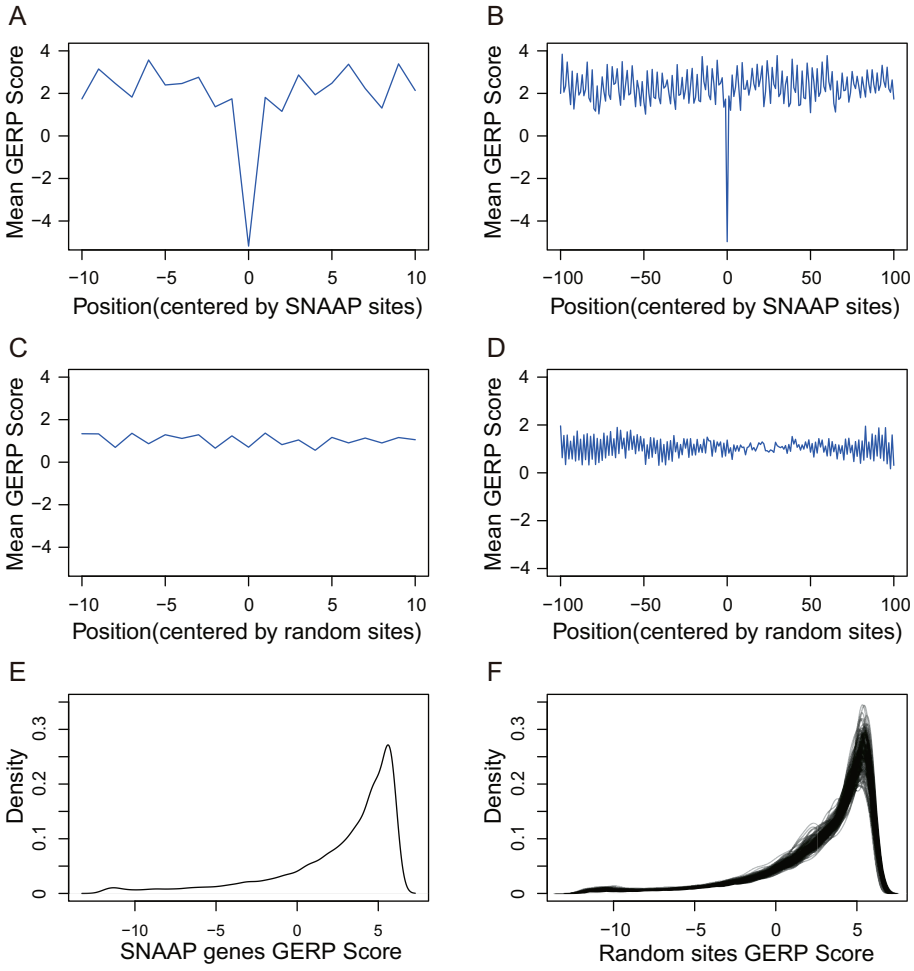

Fig. 1. Comparison of mean GERP scores of SNAAP sites and surrounding 10 bp (A) and 100 bp (B) regions, with those of randomly selected sites and surrounding 10 bp (C) and 100 bp (D) regions. Density plots of overall GERP scores of SNAAP genes (E) and 100 random sets of genes (F) exhibit similar distributions.

and language abilities (Han Chinese) were sequenced and studied (Table 1). Three of them were diagnosed to have delayed speech and language development (#1, #2, #3). The #1 and #2 patients' IQ scores were 77 and 70, respectively. The #3 patient has a lower IQ score of 66 and his hearing ability was found to be normal. The fourth patient (#4) was discovered to have a hearing deficit in left ear whose IQ was unable to be assessed. These four children are male and 2–3 years old. Two elder children (8–9 years old, male) were diagnosed to have stuttering/dysarthria (#5; IQ score: 103) and dyslexia (#6; IQ score: 72), respectively.

Here we focused on low frequency cSNPs (including cSNPs of MAF < 0.1 and *de novo* SNPs; denoted as lfcSNPs) for two reasons: (1) they are more representative for specific populations (Moore et al., 2013); (2) and deleterious variants are more likely to occur at low frequency and thus be associated with a role in the etiology of diseases (Diogo et al., 2013). Scanning over 19,868 genes, we found the six patients had significantly more lfcSNPs than non-African populations (Fig. 3A; one-tailed *t*-test, p-value = 4.6E−11, excluding African population). The African population had significantly more lfcSNPs than patients or non-African populations (one-tailed *t*-test, p-value < 1E−15), possibly due to the higher genetic diversity of African populations (Lohmueller et al., 2008; Tishkoff & Williams, 2002) or/and an artificially high level of ascertainment bias in current SNP databases derived from arrays based on predominantly non-African populations (Teo, Small, & Kwiatkowski, 2010). The six patients had significantly higher numbers of lfcSNPs than their East Asian cohorts with no overlap in values between the two groups (Fig. 3A), indicating the large difference seen in the patients is not due to ancestry differences with the controls. Furthermore, the numbers of lfcSNPs in patients (mean ± std: 5821 ± 195) were 1.8 times as many as that of the non-African populations (mean ± std: 3172 ± 334), and fell in tails of distributions of control groups representing five major ancestry lineages (Fig. 3A). Interestingly, though patients accumulated significant more lfcSNPs than general populations, we found patients have comparable numbers of lfcSNPs as general populations in the SNAAP genes (Fig. 3B). In other words, the significant excess of lfcSNPs in patients occurred outside the SNAAP genes.

A substantial portion of these lfcSNPs in patients were not documented by the 1K Genome Project in general populations (*de novo* cSNPs or dncSNPs; Supplementary Table 2). In the six patients, we found the percentages of dncSNPs in the lfcSNPs that occur in the SNAAP genes (18.1% ± 4.8%) were significantly lower (one-tailed *t*-test, p-value = 2.6E−6) than those in all genes (55.5% ± 0.5%). Combining with the above results, this result indicates that patients not only accumulated more lfcSNPs but also more dncSNPs outside SNAAP genes. This may suggest either that SNAAP genes are resistant against *de novo* variations, or alternatively, that variations on SNAAP genes have been easier to spread and get tolerated throughout the populations so that they do not appear *de novo*. Given the higher number of cSNPs in SNAAP genes in normal populations as suggested in Section 2.1, the second scenario seems more likely.

PANTHER gene ontology (GO) analysis of 2966 genes with dncSNPs in patients revealed three biological processes (containing

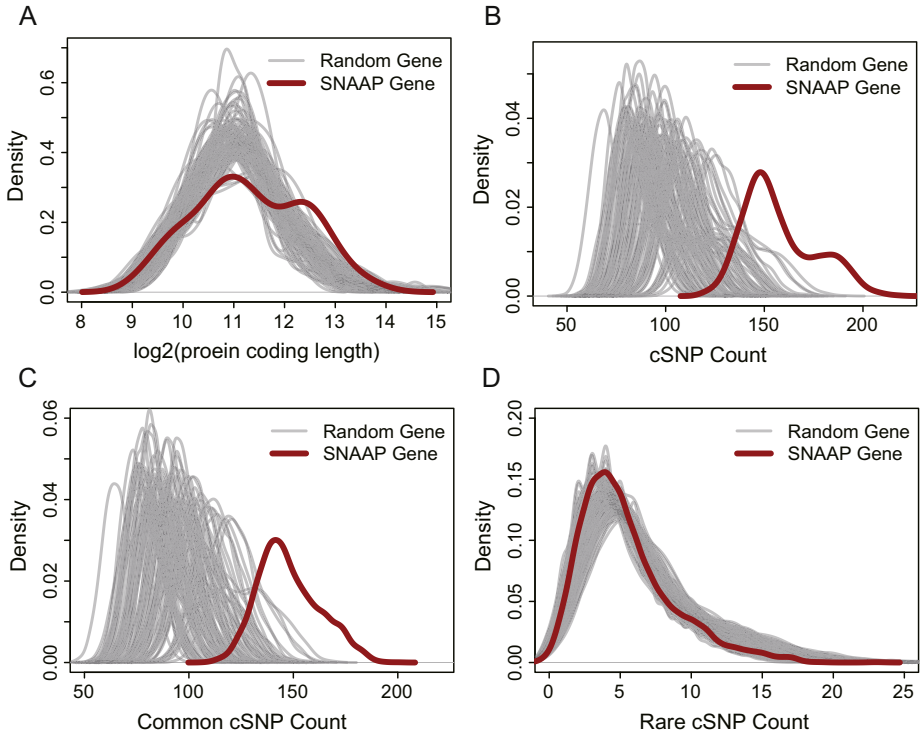

Fig. 2. Comparison of cSNPs in SNAAP genes and random gene sets. (A) Density plots of the number of cSNPs in SNAAP genes (red line) and 100 random gene sets (grey lines). (B) Density plots of the number of common cSNPs in SNAAP genes and 100 random gene sets. (C). Density plots of the number of rare cSNPs in SNAAP genes and 100 random gene sets. (D) Density plots of the coding region length of SNAAP genes and 100 random gene sets.

at least 20 genes) with >2-fold enrichment and p value < 0.05 with Bonferroni correction for multiple testing (Thomas et al., 2003): fertilization, sensory perception of sounds and cell matrix adhesion (Supplementary Table 3A, grey shading). PANTHER pathway analysis revealed three pathways (containing at least 20 genes) with >2-fold enrichment (p value < 0.05 without Bonferroni correction for multiple testing; no pathways have p value < 0.05 with Bonferroni correction): two glutamate receptor pathways (ionotropic and metabotropic) and insulin/IGF pathway (Supplementary Table 3A, no shading). Further, we tested the subset of patients' dncSNPs that are not documented in any public database and performed the same analysis as above. In this case, only two biological processes were identified: sensory perception of sounds and cell matrix adhesion, while the above three pathways remained significant.

For non-dncSNPs of the lfcSNPs, we counted the numbers for each gene in patients and the East Asian population and used the Fisher's exact test to assess if any genes have significant enrichment of such lfcSNPs in patients relative to the East Asian population. Of the 19,848 genes analyzed, 1819 genes were identified at p value < 0.05 (Supplementary Table 2). We performed PANTHER analyses on these genes as above. No significant biological processes were identified. In contrast, PANTHER pathway analysis revealed five significantly overrepresented pathways (Supplementary Table 3B). The top two pathways are mediated by Netrin and Semaphorins, which are known axon guidance molecules that interact closely with the ROBO-SLIT axon guidance pathway (Pasterkamp, 2012; Stein & Tessier-Lavigne, 2001).

2.3. Multiple patients share rare homozygous cSNPs in the PRAME gene family

Due to the wide spectrum of symptoms in speech/language disorders, it is usually difficult to identify any commonly shared variations even for a large cohort of patients with the same diagnosis. Not surprisingly, we examined rare non-synonymous cSNPs (MAF < 0.05, i.e. rncSNP) that are predicted to have "probably damaging" effects by PolyPhen (by default parameters; i.e. at false positive rate thresholds of 5% for HumDiv model and 10% for HumVar model) in the SNAAP genes of the patients and found the resultant 13 rncSNPs (Table 1; in grey shading) were unique to each patient. Two rncSNPs were from the same gene (***NEIL1***) and belong to two different patients. None of these rncSNPs were at the SNAAP sites. Specifically, patient #1 had one rncSNP in SNAAP gene ***CEP192*** that functions critically in centrosome mediated neuronal migration regulated by the ROBO-SLIT pathway (Gomez-Ferreria et al., 2007; Higginbotham & Gleeson, 2007). Patient #2 did not have any identifiable rncSNP in SNAAP genes. Patient #3 had 3 rncSNPs in SNAAP genes (***GPA33***, ***NUP188*** and ***URB1***). No prior knowledge that we could find has supported a direct role of these genes in speech/language disorders. Patient #4 had 2 rncSNPs in the SNAAP genes ***TAF1C*** and ***NEIL1***. SNPs in ***TAF1C*** were shown to be significantly associated with autism spectrum disorders (Anney et al., 2010, 2012). ***NEIL1*** was recently revealed as a disease gene for human autosomal recessive steroid-resistant nephrotic syndrome (SRNS) and children with SRNS are at higher risk of sensorineural hearing impairment (Sanna-Cherchi et al., 2011). But it remains unclear if this is caused by drug over-dosage or genetics. Patient #5 also had a rncSNP in ***NEIL1*** but at a different position from that in patient #4. Patient #6 had one rncSNP in SNAAP gene ***LAMB4*** and two rncSNPs in SNAAP gene ***LCT***. The two rncSNPs in ***LCT*** resulted in two consecutive Prolines, i.e. Pro-Pro, to replace the original residues: Thr-Ser. This supposedly would induce a drastic change on protein structure from a linear shape into a curved shape. But it is unclear how mutations on ***LCT*** may affect speech/language.

We also examined the non-synonymous dncSNPs that are homozygous in each patient and are predicted to be "probably damaging" by PolyPhen. There was a total of 13 such dncSNPs in

Table 1
The "probably damaging" rare non-synonymous SNPs in patients. In grey shading: rncSNPs in SNAAP genes; no shading: homozygous dncSNPs.

Patient	Gene	Chr	Pos	dbSnp ID	Ref	Alt	AA Change	MAF[a]	EAS MAF	Gene function	Phenotype
#1	CEP192	18	13008577	rs114794290	C	G	Ser138Cys	0.0088	0.0179	neuronal migration	Delayed speech language development
	FAM115C	7	143416945	rs62486260	C	T	Arg101Trp	-	-	cold sensation	
	CYP2D6	22	42523558	rs202102799	T	C	Tyr355Cys	-	-	Neurorecognit-ion, eating disorder	
	PRAMEF5	1	13368549	rs201258581	C	T	Pro416Leu	-	-	unknown	
#2	PRAMEF7	1	12980040	-	A	G	Tyr411Cys	-	-	unknown	
#3	GPA33	1	167032960	rs188793256	G	C	Pro144Ala	0.0006	0.003	cell recognition	
	NUP188	9	131765214	rs17433024	C	T	Ala1419Val	0.0016	-	nucleoporin	
	URB1	21	33706638	rs148134142	A	G	Leu1564Pro	0.0086	0.0417	unknown	
	ACE2	X	15582334	-	G	A	Arg708Trp	-	-	amino acid homeostasis	
	IRS4	X	107979111	-	G	A	Ser155Phe	-	-	glucose homeostasis	
	ZNF717	3	75787269	-	G	A	Thr502Ile	-	-	unknown	
	POTEI	2	131220502	-	C	T	Val1039Met	-	-	unknown	
#4	NEIL1	15	75646094	rs140982397	G	A	Gly245Arg	0.0006	0.003	DNA repair	Hearing deficit
	TAF1C	16	84212765	rs750949459	A	T	Leu798Met	-	-	part of the transcription complex	
	PRAMEF5	1	13368549	rs201258581	C	T	Pro416Leu	-	-	unknown	
	ANKRD36	2	97854842	-	C	G	Pro721Arg	-	-	unknown	
	TIGD5	8	144681518	-	G	C	Arg482Pro	-	-	unknown	
	GAGE13	X	49192710	rs201491405	G	C	Asp75His	-	-	unknown	
	AC008132.13	22	18835784	-	G	A	Arg447Gln	-	-	unknown	
#5	NEIL1	15	75641449	rs187873972	C	A	Pro68His	0.0042	0.0198	DNA repair	Stuttering
	ZNF717	3	75786516	rs141393015	G	T	Thr753Asn	-	-	unknown	
	ZNF717	3	75788130	-	C	T	Gly165Glu	-	-	unknown	
	PRAMEF5	1	13368549	rs201258581	C	T	Pro416Leu	-	-	unknown	
#6	LCT	2	136570172	rs76854071	A	G	Ser688Pro	-	-	encoding lactase	Reading disability
	LCT	2	136570175	-	T	G	Thr687Pro	-	-	encoding lactase	
	LAMB4	7	107664494	-	C	T	Cys1759Tyr	-	-	cell attachment and migration	
	AGAP4	10	46321527	rs200049550	G	A	Arg610Cys	-	-	unknown	

[a] MAF is minor allele frequency from 1000 Genome Project.

13 genes (Table 1, no shading) found in all six patients, and over a half of these genes are of unknown functions. Surprisingly, some SNPs were shared in multiple patients (rs201258581 in gene *PRAMEF5* of patient #1, #4 and #5), or in different locations of the same gene in different patients (ZNF717 of patient #3 and #5). One *ZNF717* mutation has been isolated among the 16 rare homozygous variants from at least 2 families that have patients with Joubert Syndrome, a disorder characterized by autistic behav-

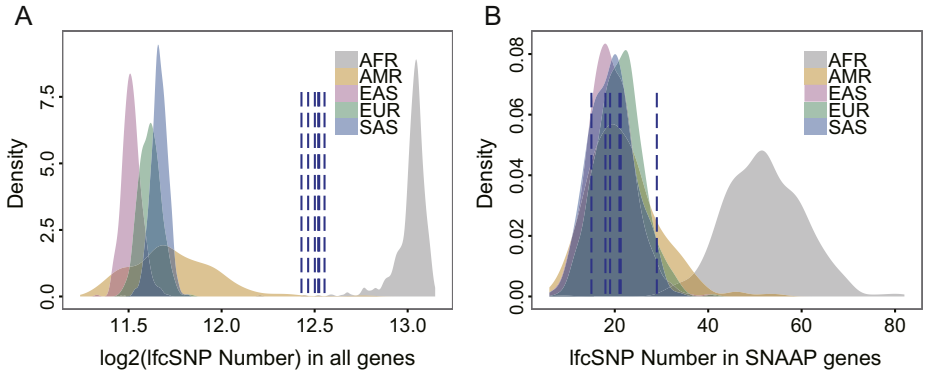

Fig. 3. The number of lfcSNPs in patients and 1 KG populations. A. Density plot in a log2 scale of the total number of lfcSNPs in 19,848 protein coding genes of the 1 KG populations. The five super human ethnic populations are plotted separately different colors. The corresponding numbers of lfcSNPs in 6 patients are shown as six blue vertical lines. B. Density plot in a log2 scale of the total number of lfcSNPs in SNAAP genes of 1 KG populations. The corresponding numbers of lfcSNPs in 6 patients are shown as six blue vertical lines.

ior and mental retardation (Huang et al., 2011). Furthermore, patient #2 also had a SNP in *PRAMEF7* gene paralogous to *PRAMEF5*. In other words, there was a total of four patients (#1, #2, #4, #5) with such homozygous missense mutations on the *PRAME* gene family, which is among the youngest human gene families that arose in the last 3 million years, i.e. well after human-chimpanzee divergence, and were positively selected in the human population (Birtle, Goodstadt, & Ponting, 2005). Interestingly, the most studied *PRAME* family member gene *PRAME* is reported as a dominant repressor of retinoic acid receptor signaling (Epping et al., 2005). Since retinoic acid acts as a chemoattractant to guide growth cone turning (Dmetrichuk, Carlone, & Spencer, 2006) and its receptor signaling interacts with the language gene *FOXP2* to drive neuronal differentiation (Devanna, Middelbeek, & Vernes, 2014), we suggest that the axon guidance pathways involving the recently evolved *PRAME* gene family have been critical for both speech/language evolution and related disorders.

3. Discussion

We caution that it will be completely biased if one tries to understand our results by the standards of population genetics. Here it is more useful to include fewer patients and less homogenous phenotypes in order to bridge results from the two research paradigms: comparative genomics and population genetics. The genetic markers we obtained from the comparative genomic studies are based on the analyses of species with or without the vocal learning trait, a primitive but distinguishing substrate for human speech language. There is no priori to correlate findings from vocal learning studies with any specific or well defined speech language disorders, because such findings may relate to a broad spectrum of functionalities necessary for human speech language, e.g. motor, sensory, neurodevelopment, emotion or their combinations. In this regard, by including the few not-so-well-defined patients, we in fact increased the signal-to-noise ratio, which is in sharp contrast with conventional design of human population genetic studies which requires well-defined large cohorts.

The recent discovery of more-than-expected genome-wide convergence on protein sequences of echolocation mammals has been disputed (Parker et al., 2013). It raises questions about whether certain phenotypic convergence would be associated with any genomic convergence that can be appropriately detected by known models of protein evolution (Thomas & Hahn, 2015; Zou & Zhang, 2015). Here our analyses on SNAAP sites and genes that exhibit amino acid convergence in vocal learners supported their significant relevance with human speech/language evolution and related disorders. Our results imply that SNAAP genes gained further functional specialization in human evolution and such specialization has been widely fixed as common SNPs relevant to human speech/language functions. The strong negative selection over SNAAP sites that differentiates human from nonhuman primates throughout human evolution could have served as a prerequisite condition for such specialization to occur.

We postulate a twofold role of possible specializations brought by SNAAPs. Firstly, the specialized changes may have influenced brain development through axon guidance pathways that allows formation of new projections, thus enabling further development of specialized circuits. Such changes could be mediated by *ROBO/SLIT* and centrosome related genes for neuronal migration. For speech language pathology, this might provide a plausible explanation to the clinical finding that many individuals with primary microcephaly, a disease usually caused by defected centrosomal proteins, show no obvious motor deficits but suffer from speech language delay (Faheem et al., 2015). Patients in our study have significant enrichment of probably damaging SNPs in axon guidance pathways and this might have induced abnormal cross-talks with *ROBO/SLIT* or centrosome mediated neuronal migration to affect their speech/language abilities. Secondly, the specialized changes may have affected the ciliated cells, e.g. those on the lungs, respiratory tract and inner ear that are critical organs for speech language, through centrosome related genes and others. Pathologically, centrosomal defects are known to cause ciliopathies. Further, Usher syndrome is a typical ciliopathy with impaired speech language symptoms and molecularly linked with ***USHBP1*** (Piatti, De Santi, Brogi, Castorina, & Ambrosetti, 2014). The SNAAP genes ***E2F3*** and ***RB1CC1*** play critical roles in maintaining inner ear ciliated hair cells (Mantela et al., 2005). Interestingly, the dyslexia candidate gene *DYX1C1*, regulated by ***PARP1***, had another role recently recognized as a dynein axonemal assembly factor, linking itself to ***ODF1*** for normal ciliary motor functions (Tarkar et al., 2013). In this regard, the enriched defects of our patients in fertility functional category may possibly reflect some malfunction not just related with sperm but general ciliary development, which is in line with their enriched defects in sound perception and underlies their impaired speech/language abilities.

Recent analyses on the known 10 genes associated with speech language disorders revealed three axon guidance molecules (***ROBO1***, *ROBO2*, *CNTNAP2*) and their regulator *FOXP2* to have SNPs under significant positive selection during human population evolution (Mozzi et al., 2016). It strengthens our argument that the axon guidance pathways were important targets in human speech/language evolution. In contrast, few pervasive or episodic positive selection events were detected for these genes to be associated with human speech language or vocal communication traits.

This could be attributed to the more complicated evolutionary scenarios of genetic substrate for speech language at species level. It was suggested that many SNAAP sites might have experienced a loss of ancestral types at early evolutionary stages of placental mammals but then recapitalized independently in the vocal learners through reversal mutations, a situation where natural selection first favors a mutation, then favors its removal, and later still favors its ultimate restoration (Wang, 2011). Further, the common ancestor of primates might have gained certain genetic substrate for speech/language, which human further evolved but other primates lost (Wang, 2011).

We noted quite a few functionally interconnected SNAAP genes for axon guidance. Among them, genes ***ROBO1***, ***NEO1***, ***PITPNA*** and ***PTPRB*** modulate responsiveness of growing axons to guidance cues, such as Netrin-1, Slits and Ephrins (e.g. ***EFNB1***), stabilizes migrating neurons and facilitates axonal navigation (Holland, Peles, Pawson, & Schlessinger, 1998; Kim et al., 2015; Leyva-Díaz et al., 2014; Stein & Tessier-Lavigne, 2001; Sun, Bahri, Schmid, Chia, & Zinn, 2000; Xie et al., 2005). Once a migrating neuron encounters cues like Slits, it retracts the leading process and extends a new process on the opposite side by re-orienting the centrosome into the newly formed process where three SNAAP genes (***CKAP5***, ***PCM1*** and ***CEP192***) are involved (Barr & Gergely, 2008; Dammermann & Merdes, 2002; Gomez-Ferreria et al., 2007). SLIT/ROBO pathway transcriptionally regulates *HES1* and in turn affects craniofacial structure development (Akimoto et al., 2010; Borrell et al., 2012), whereas mutated *EFNB1* caused craniofrontonasal abnormalities (Wieland et al., 2004). In addition, ***ROBO1***'s ligand *SLIT1* and ***EFNB1*** are both transcriptionally upregulated by *FOXP2* (Konopka et al., 2009; Vernes et al., 2011).

Compared to *FOXP2* and other susceptible genes known so far, our analyses on SNAAP genes and human patients provide a more systematic framework to better understand a patient's pathology. The speech/language disorders of the six patients did not seem to involve mutations that are likely causative from SNAAP genes. Rather, patients have accumulated more low frequency SNPs, potentially deleterious, in genes other than SNAAP genes. As they were fixed at low levels in human populations rather than *de novo*, we suspect these mutations are genetically transmitted from their parents and have served as risk factors. The rest are *de novo* mutations and many preferentially occur in neural functional related genes. We argue these *de novo* mutations are likely to contain causative factors, e.g. homozygous mutations in *PRAME* gene family, which also potentially links to the axon guidance pathways. The above findings and postulations are summarized in Supplementary Fig. 1.

4. Material and methods

The public datasets of human single nucleotide polymorphism (SNP) used in this study were retrieved from the latest release of the 1000 Genomes Project (1 KG) data (Phase 3, May 2013 release). It includes 2504 individuals representing 5 super populations worldwide: African (AFR), Ad Mixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS). Six patients (age 3–10) of Han Chinese ancestry were recruited from unrelated families from Department of Children Healthcare of Capital Institute of Pediatrics, Beijing, China with approval by the ethical review board of the hospital and informed consent obtained from the participants. Cotton swab derived buccal cells were scraped from the inner side of cheeks of patients, and genomic DNA was extracted and purified by QIAamp DNA mini Kit (Cat#: 51306, QIAGEN). The patients' samples yielded about 1μg DNA and were sent for exome sequencing. The intelligence quotient of patients was assessed based on the Developmental Diagnostic Scale of Children Aged 0–6 Years and Wechsler Intelligence Scale for Children (v3) for older patients, respectively. The diagnosis for impaired speech/language abilities were determined by the "test of speech/language" in the above tests. A hearing test was performed by auditory brainstem response (ABR). Whole exome enrichment libraries were made with NimbleGen SeqCap EZ Exome + UTR enrichment kit (Roche NimbleGen Inc.) and sent for pair-end whole exome sequencing at Annoroad Gene Technology Co., Ltd., Beijing. The resultant reads were at least 30× coverage when mapped to a reference human genome (hg19, obtained from the University of California, Santa Cruz (UCSC) Genome Browser) using BWA tools. The mapped BAM files were processed for variant calling and filtering with the best practice pipeline of GATK. All variant calling results were saved in VCF format. SnpEff was used to annotate the VCF files of 1K Genome populations and patients. Coding SNPs (cSNPs) were extracted from annotated VCF files based on the SnpEff gene-based annotation.

Conflict of interest

The authors declare no competing interests.

Authors' contributions

H.L. and R.W. conceived the concept of the study. H.L., R.W., Z.Y., Z.S., Q.X. and E.J. designed study plan and analyzed data. X.S., Y.Z., J. W., C.M., and H.L. recruited patients and performed the diagnostic tests. X.S. and H.L. prepared the patients' sample DNA and supervised exome sequencing. H.L., R.W. and Z.Y. wrote the paper. H.L., R.W., Z.S., Q.X. and E.J. revised the paper.

Funding

This work was supported by grants from the National Natural Science Foundation of China [31171051 to H.L., 31371108 to H.L., 31171274 to Z.S.], 973 Projects of China [2012CB725203 to Z.S.], Natural Science Foundation of Beijing [5112008, 5132007 to H. L.], the General Program of Science and Technology Development Project of Beijing Municipal Education Commission of China [KM201110025001 to H.L.], Beijing Municipal Technology Foundation for Selected Overseas Chinese Scholar to H.L., and the Capital Health Research and Development of Special [2014-1-4091 to C. M.], Howard Hughes Medical Institute [to E.D.J.].

Statement of significance to the neuroscience of language

- Confirmed the genetics of vocal learning to be further specialized in humans.
- Identified possible genetic risk factors and causative mutations for speech language disorders linking to axon guidance.
- Proposed the role of axon guidance pathways on speech language evolution and related disorders.

Acknowledgements

The authors thank colleagues from Annoroad Inc., and Hengkuan Genomics Co., Ltd. in facilitating exome sequencing and data collection.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bandl.2017.06.007.

References

Akimoto, M., Kameda, Y., Arai, Y., Miura, M., Nishimaki, T., Takeda, A., et al. (2010). Hes1 is required for the development of craniofacial structures derived from ectomesenchymal neural crest cells. *Journal of Craniofacial Surgery, 21*(5), 1443–1449.

Anney, R., Klei, L., Pinto, D., Almeida, J., Bacchelli, E., Baird, G., et al. (2012). Individual common variants exert weak effects on the risk for autism spectrum disorders. *Human Molecular Genetics, 21*(21), 4781–4792.

Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T. R., et al. (2010). A genome-wide scan for common alleles affecting risk for autism. *Human Molecular Genetics, 19*(20), 4072–4082.

Barr, A. R., & Gergely, F. (2008). MCAK-independent functions of ch-Tog/XMAP215 in microtubule plus-end dynamics. *Molecular and Cellular Biology, 28*(23), 7199–7211.

Bates, T., Luciano, M., Medland, S., Montgomery, G., Wright, M., & Martin, N. (2011). Genetic variance in a component of the language acquisition device: ROBO1 polymorphisms associated with phonological buffer deficits. *Behavior Genetics, 41*(1), 50–57.

Birtle, Z., Goodstadt, L., & Ponting, C. (2005). Duplication and positive selection among hominin-specific PRAME genes. *BMC Genomics, 6*(1), 1–19.

Borrell, V., Cardenas, A., Ciceri, G., Galceran, J., Flames, N., Pla, R., et al. (2012). Slit/Robo signaling modulates the proliferation of central nervous system progenitors. *Neuron, 76*(2), 338–352.

Dammermann, A., & Merdes, A. (2002). Assembly of centrosomal proteins and microtubule organization depends on PCM-1. *The Journal of Cell Biology, 159*(2), 255–266.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLOS Computational Biology, 6*(12), e1001025.

Devanna, P., Middelbeek, J., & Vernes, S. C. (2014). FOXP2 drives neuronal differentiation by interacting with retinoic acid signaling pathways. *Frontiers in Cellular Neuroscience, 8*, 305.

Diogo, D., Kurreeman, F., Stahl, E. A., Liao, K. P., Gupta, N., Greenberg, J. D., et al. (2013). Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. *The American Journal of Human Genetics, 92*(1), 15–27.

Dmetrichuk, J. M., Carlone, R. L., & Spencer, G. E. (2006). Retinoic acid induces neurite outgrowth and growth cone turning in invertebrate neurons. *Developmental Biology, 294*(1), 39–49.

Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience, 22*(1), 567–631.

Epping, M. T., Wang, L., Edel, M., Carlee, L., Hernandez, M. J. M., & Bernards, R. (2005). The human tumor antigen PRAME is a dominant repressor of retinoic acid receptor signaling. *Cell, 122*(6), 835–847.

Faheem, M., Naseer, M. I., Rasool, M., Chaudhary, A., Kumosani, T., Ilyas, A. M., et al. (2015). Molecular genetics of human primary microcephaly: An overview. *BMC Medical Genomics, 8*(1), 1–11.

Gomez-Ferreria, M. A., Rath, U., Buster, D. W., Chanda, S. K., Caldwell, J. S., Rines, D. R., et al. (2007). Human Cep192 is required for mitotic centrosome and spindle assembly. *Current Biology, 17*(22), 1960–1966.

Hannula-Jouppi, K., Kaminen-Ahola, N., Taipale, M., Eklund, R., Nopola-Hemmi, J., Kaariainen, H., et al. (2005). The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia. *PLoS Genetics, 1*(4), 467–474.

Higginbotham, H. R., & Gleeson, J. G. (2007). The centrosome in neuronal development. *Trends in Neurosciences, 30*(6), 276–283.

Holland, S. J., Peles, E., Pawson, T., & Schlessinger, J. (1998). Cell-contact-dependent signalling in axon growth and guidance: Eph receptor tyrosine kinases and receptor protein tyrosine phosphatase [beta]. *Current Opinion in Neurobiology, 8*(1), 117–127.

Huang, L., Szymanska, K., Jensen, Victor L., Janecke, Andreas R., Innes, A. M., Davis, Erica E., et al. (2011). TMEM237 is mutated in individuals with a Joubert Syndrome related disorder and expands the role of the TMEM family at the ciliary transition zone. *The American Journal of Human Genetics, 89*(6), 713–730.

Jarvis, E. D. (2004). Learned birdsong and the neurobiology of human language. *Annals of the New York Academy of Sciences, 1016*(1), 749–777.

Kim, M., Fontelonga, T., Roesener, A. P., Lee, H., Gurung, S., Mendonca, P. R. F., et al. (2015). Motor neuron cell bodies are actively positioned by Slit/Robo repulsion and Netrin/DCC attraction. *Developmental Biology, 399*(1), 68–79.

Konopka, G., Bomar, J. M., Winden, K., Coppola, G., Jonsson, Z. O., Gao, F., et al. (2009). Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature, 462*(7270), 213–217.

Leyva-Díaz, E., del Toro, D., Menal, Maria. J., Cambray, S., Susín, R., Tessier-Lavigne, M., et al. (2014). FLRT3 is a Robo1-interacting protein that determines Netrin-1 attraction in developing axons. *Current Biology, 24*(5), 494–508.

Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., et al. (2008). Proportionally more deleterious genetic variation in European than in African populations. *Nature, 451*(7181), 994–997.

Mantela, J., Jiang, Z., Ylikoski, J., Fritzsch, B., Zacksenhaus, E., & Pirvola, U. (2005). The retinoblastoma gene pathway regulates the postmitotic state of hair cells of the mouse inner ear. *Development, 132*(10), 2377–2388.

Moore, C. B., Wallace, J. R., Wolfe, D. J., Frase, A. T., Pendergrass, S. A., Weiss, K. M., et al. (2013). Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. *PLoS Genetics, 9*(12), e1003959.

Mozzi, A., Forni, D., Clerici, M., Pozzoli, U., Mascheretti, S., Guerini, F. R., et al. (2016). The evolutionary history of genes involved in spoken and written language: Beyond FOXP2. *Scientific Reports, 6*, 22157.

Orosco, L. A., Ross, A. P., Cates, S. L., Scott, S. E., Wu, D., Sohn, J., et al. (2014). Loss of Wdfy3 in mice alters cerebral cortical neurogenesis reflecting aspects of the autism pathology. *Nature Communications, 5*. http://dx.doi.org/10.1038/ncomms5692.

Parker, J., Tsagkogeorga, G., Cotton, J. A., Liu, Y., Provero, P., Stupka, E., et al. (2013). Genome-wide signatures of convergent evolution in echolocating mammals. *Nature, 502*(7470), 228–231.

Pasterkamp, R. J. (2012). Getting neural circuits into shape with semaphorins. *Nature Reviews Neuroscience, 13*(9), 605–618.

Peter, B., Wijsman, E. M., Nato, A. Q., Matsushita, M., Chapman, K. L., Stanaway, I. B., et al. (2016). Genetic candidate variants in two multigenerational families with childhood apraxia of speech. *PLoS ONE, 11*(4), e0153864.

Petkov, C. I., & Jarvis, E. D. (2012). Birds, primates, and spoken language origins: Behavioral phenotypes and neurobiological substrates. *Frontiers in Evolutionary Neuroscience*. http://dx.doi.org/10.3389/fnevo.2012.00012.

Pfenning, A. R., Hara, E., Whitney, O., Rivas, M. V., Wang, R., Roulhac, P. L., et al. (2014). Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science, 346*(6215). http://dx.doi.org/10.1126/science.1256846.

Piatti, G., De Santi, M. M., Brogi, M., Castorina, P., & Ambrosetti, U. (2014). Emerging ciliopathies: Are respiratory cilia compromised in Usher syndrome? *American Journal of Otolaryngology, 35*(3), 340–346.

Sanna-Cherchi, S., Burgess, K. E., Nees, S. N., Caridi, G., Weng, P. L., Dagnino, M., et al. (2011). Exome sequencing identified MYO1E and NEIL1 as candidate genes for human autosomal recessive steroid-resistant nephrotic syndrome. *Kidney International, 80*(4), 389–396.

Stein, E., & Tessier-Lavigne, M. (2001). Hierarchical organization of guidance receptors: Silencing of netrin attraction by slit through a Robo/DCC receptor complex. *Science, 291*(5510), 1928–1938.

Sun, Q., Bahri, S., Schmid, A., Chia, W., & Zinn, K. (2000). Receptor tyrosine phosphatases regulate axon guidance across the midline of the Drosophila embryo. *Development, 127*(4), 801–812.

Tarkar, A., Loges, N. T., Slagle, C. E., Francis, R., Dougherty, G. W., Tamayo, J. V., et al. (2013). DYX1C1 is required for axonemal dynein assembly and ciliary motility. *Nature Genetics, 45*(9), 995–1003.

Teo, Y.-Y., Small, K. S., & Kwiatkowski, D. P. (2010). Methodological challenges of genome-wide association analysis in Africa. *Nature Reviews Genetics, 11*(2), 149–160.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., et al. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research, 13*(9), 2129–2141.

Thomas, G. W. C., & Hahn, M. W. (2015). Determining the null model for detecting adaptive convergence from genomic data: A case study using echolocating mammals. *Molecular Biology and Evolution, 32*(5), 1232–1236.

Tishkoff, S. A., & Williams, S. M. (2002). Genetic analysis of African populations: Human evolution and complex disease. *Nature Reviews Genetics, 3*(8), 611–621.

Vernes, S. C., Oliver, P. L., Spiteri, E., Lockstone, H. E., Puliyadi, R., Taylor, J. M., et al. (2011). Foxp2 regulates gene networks implicated in neurite outgrowth in the developing brain. *PLoS Genetics, 7*(7), e1002145.

Vernes, S. C., Spiteri, E., Nicod, J., Groszer, M., Taylor, J. M., Davies, K. E., et al. (2007). High-throughput analysis of promoter occupancy reveals direct neural targets of FOXP2, a gene mutated in speech and language disorders. *The American Journal of Human Genetics, 81*(6), 1232–1250.

Wang, R. (2011). *Dissecting the genetic basis of convergent complex traits based on molecular*. Duke University. <http://dukespace.lib.duke.edu/dspace/handle/10161/5630>.

Wang, R., Chen, C.-C., Hara, E., Rivas, M. V., Roulhac, P. L., Howard, J. T., et al. (2015). Convergent differential regulation of SLIT-ROBO axon guidance genes in the brains of vocal learners. *Journal of Comparative Neurology, 523*(6), 892–906.

Wieland, I., Jakubiczka, S., Muschke, P., Cohen, M., Thiele, H., Gerlach, K. L., et al. (2004). Mutations of the ephrin-B1 gene cause craniofrontonasal syndrome. *The American Journal of Human Genetics, 74*(6), 1209–1215.

Xie, Y., Ding, Y. Q., Hong, Y., Feng, Z., Navarre, S., Xi, C. X., et al. (2005). Phosphatidylinositol transfer protein-alpha in netrin-1-induced PLC signalling and neurite outgrowth. *Nature Cell Biology, 7*(11), 1124–1132.

Zou, Z., & Zhang, J. (2015). Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Molecular Biology and Evolution, 21*(8), 2085–2096.