

Linkage disequilibrium extends across putative selected sites in FOXP2

Submission for a letter

Susan Ptak^{1&#}, Wolfgang Enard^{1&}, Victor Wiebe¹, Ines Hellmann², Johannes Krause¹,
Michael Lachmann¹ and Svante Pääbo¹.

¹ Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany

² Department of Integrative Biology, University of California, Berkeley, CA, 94720 USA.

Currently at: Max F. Perutz Laboratories, Dr. Bohr Gasse 5, 1030 Vienna, Austria

& These authors contributed equally

To whom correspondence should be addressed.

Email : ptak@eva.mpg.de

Phone : +49-341-3550-550

Fax : +49-341-3550-555

Key words : evolution, selection, linkage disequilibrium, FOXP2, human

Running head : Molecular evolution of human FOXP2

Title length: 70 characters (including spaces)

Abstract length: 140 words

Main text length: 11,658 characters (including spaces)

Page requirement for figure 1: 0.5-1 page

Number of references : 29

Abstract

Polymorphism data in humans suggest that the gene encoding the transcription factor *FOXP2*, which influences speech and language development, has been subject to a selective sweep within the last 260,000 years. It has been proposed that one or both of two substitutions that occurred on the human evolutionary lineage and changed amino acids were the targets for selection. In apparent contradiction to this is the observation that these substitutions are present in Neandertals who diverged from humans maybe 300,000-400,000 years ago. We have collected polymorphism data upstream and downstream of the substitutions. Contrary to what is expected following a selective sweep, we find that the haplotypes extend across the two sites. We discuss possible explanations for these observations. One of them is that the selective sweep reflected in *FOXP2* polymorphism data was not associated with the two amino acid substitutions.

Main text

Humans carrying only one functional copy of the gene encoding the transcription factor *FOXP2* are impaired in speech and language acquisition (Vargha-Khadem et al. 2005), showing deficits in a wide range of expressive and receptive language skills, particularly in the timing and sequencing of orofacial movements (Alcock KJ 2000; Watkins 2002). Two lines of evidence suggest that positive selection has affected *FOXP2* and that it has acquired novel properties on the human evolutionary lineage (Enard et al. 2002; Zhang, Webb, and Podlaha 2002). First, two amino acids encoded in exon 7 of the *FOXP2* gene changed on the human lineage after separation from the common ancestor with chimpanzees (Enard et al. 2002; Zhang, Webb, and Podlaha 2002) and are now fixed among humans (Supl. Data). Since *FOXP2* is extremely conserved among mammals, this represents a significant accumulation of amino acid substitutions (Enard et al. 2002; Zhang, Webb, and Podlaha 2002; Clark et al.

2003; Nickel et al. 2008). Furthermore, when these substitutions are introduced into the endogenous mouse *Foxp2*, they affect the function of cortico-striatal circuits and alter vocalization (Enard 2009). Second, when a ~14,000 bp segment immediately upstream of exon 7 was sequenced in 20 individuals from all major continents, an excess of rare alleles, including rare ancestral alleles, was found (Enard et al. 2002). It is unlikely that a neutral scenario with demographic models appropriate for humans explains such a signal (Enard et al. 2002). Since the signature of a selective sweep decays quickly (especially the signature of an excess of rare ancestral alleles) (Przeworski 2002), the fixation of the sweep seen in *FOXP2* is likely to have occurred within the last 260,000 years (Enard et al. 2002; Supl. Data). A second dating approach even suggests that the sweep began as recently as 42,000 years ago (Coop et al. 2008). The two amino acid substitutions have been considered as the best candidates for the cause of this sweep (e.g. Enard et al. 2002; Zhang, Webb, and Podlaha 2002; Krause et al. 2007; Coop et al. 2008). There is, however, no direct evidence linking either of the amino acids with the sweep. Furthermore, two 43,000-year-old Neandertals have been found to carry the amino acid substitutions (Krause et al. 2007), suggesting that the amino acid substitutions predate the separation of human and Neandertal ancestral populations 300,000 to 400,000 years ago (Stringer and Hublin 1999; Noonan et al. 2006). Here we further explore the hypothesis that one (or both) of the amino acids is the cause of the selective sweep signal seen in polymorphism data.

Any substitution initially occurs on one particular chromosomal haplotype. If that substitution is beneficial and increases in frequency, all derived and ancestral alleles on this haplotype also increase in frequency. However, recombination during the sweep can bring the beneficial substitution to other haplotypes that carry different alleles than the haplotype upon which the substitution initially occurred. Thus recombination can rescue alleles that otherwise would be lost during a selective sweep. Upon fixation, alleles rescued by recombination will

tend to be at low frequency, while alleles on the haplotype on which the selected substitution arose will tend to be at high frequency (fig. 1A). Some of these low-frequency alleles will be derived and some will be ancestral. In contrast, mutations during and after the sweep create low-frequency derived alleles (barring back-mutations). Thus, recombination leads to an increased occurrence of low-frequency ancestral (or high-frequency derived) alleles in the vicinity of the selected site (Fay and Wu 2000). Indeed, 16 rare alleles, 7 of which are ancestral, were seen in one Nigerian individual among the 20 individuals initially sequenced, supporting a selective sweep (Enard et al. 2002). If one or both of the amino acid substitutions caused the sweep, the Nigerian haplotype upstream of exon 7 is expected to have been rescued by a crossing-over event between it and exon 7 during the sweep (fig. 1A). However, since polymorphic sites on either side of a selected site are expected not to be in linkage disequilibrium (LD) (Kim and Nielsen 2004; Stephan, Song, and Langley 2006; Jensen et al. 2007; McVean 2007), the Nigerian individual is not expected to carry any rare haplotype downstream of exon 7.

To test if LD breaks down across exon 7, we sequenced 7,606 bp downstream of exon 7 in the Nigerian individual that carries the 7 low-frequency ancestral alleles as well as in three other individuals (Supl. Data). We found five sites at which the Nigerian individual was heterozygous and carried ancestral alleles that were absent in the human reference genome and the other individuals sequenced. This suggests that the LD seen upstream of exon 7 extends across exon 7. To test this, we sequenced a 3,575 bp region containing the polymorphic sites downstream and two of the 7 sites upstream of exon 7 in 14 additional Nigerian individuals. We find two main haplotypes: one that carries the derived state at the sites upstream and at the first four sites downstream, and a second haplotype that carries the ancestral state at all 6 sites (fig. 1B, Table S1). Thus, contrary to what one would expect following a sweep where the selected site was within exon 7, we find significant LD

($\omega=r^2=D'=1$; $p<0.001$) between sites upstream and downstream of exon 7 (fig. 1B). Simulations suggest that such a pattern of complete LD is unlikely both for a standard sweep model and a sweep model with demography appropriate for the Nigerian population ($p\leq 0.001$; Supl. Data).

These observations may be explained by five different scenarios. The first scenario is that the signals for positive selection are spurious and that no sweep affected *FOXP2*. Given the evidence (Enard et al. 2002; Zhang, Webb, and Podlaha 2002), this seems unlikely, but remains a possibility.

The second scenario is that the sweep is old and began prior to the split between humans and Neandertals. This was deemed most likely in (Krause et al. 2007), based on the dating method used in (Enard et al. 2002), which make numerous assumptions about human demography and the selective sweep. However, the phylogenetic method used by (Coop et al. 2008) is based only on the accumulation of substitutions in the chromosomal region and thus makes no such assumptions. Hence we do not find this scenario likely.

The third scenario is that at least one of the amino acid substitutions arose in humans after their divergence from Neandertals and then swept to fixation. The amino acid substitutions seen in Neandertals are either due to gene flow between humans and Neandertals or to experimental contamination from modern humans. A number of lines of evidence suggest that this is unlikely. No evidence for gene flow from Neandertals to humans have hitherto been found (Serre et al. 2004; Noonan et al. 2006; Krause et al. 2007), although low levels of admixture cannot presently be rejected. Krause et al. consistently saw the derived state at both amino acid positions in the Neandertal extracts in several experiments in two Neandertal individuals from Spain, and for one of these individuals the results were independently reproduced in two other laboratories (but see Coop et al. 2008). The derived

state of both amino acid positions has now been observed in high-throughput sequencing of other Neandertal individuals from Croatia (Burbano and Pääbo, unpublished data).

In addition, for both the second and third scenario we still need to explain the finding of LD across the selected site. A possibility is gene conversion during the sweep (Jones and Wakeley 2008). Yet, simulations of a selective sweep in the presence of gene conversion, for this region, suggest that such strong LD occurs only for high rates of gene conversion that are restricted to the area around exon 7 (Supl. Data). Although we can not rule out gene conversion, gene-conversion alone is not sufficient to explain the findings at *FOXP2*, given the presence of the amino acid substitutions in Neandertals and the estimated timing of the sweep.

The fourth scenario is that both amino acid substitutions occurred prior to the split of modern humans and Neandertals, but at least one of them was not fixed and persisted in both populations after the split. At some point, this amino acid spread to a second haplotype. The two haplotypes carrying the amino acids then became selectively favored and swept to fixation in humans. However, the signals of an excess of rare alleles and rare ancestral alleles are less likely under such a scenario than under a model of selection on new variation (Innan and Kim 2004; Hermisson and Pennings 2005; Przeworski, Coop, and Wall 2005; Teshima, Coop, and Przeworski 2006) and become increasingly less likely as the frequency of the selected allele increases (Hermisson and Pennings 2005; Przeworski, Coop, and Wall 2005). Furthermore, the paucity of variation present in this region (Coop et al. 2008) is unlikely if the selected site persisted long enough prior to selection to spread to two haplotypes.

The fifth scenario is what the LD and Neandertal data are telling us *prima facie*, namely that the two amino acids are not associated with the selective sweep. This suggests that sometime after the human-chimpanzee split but prior to the modern human-Neandertal split, the two amino acid substitutions arose and became fixed, possibly due to positive

selection. A selective sweep then affected *FOXP2* in humans after their separation from Neandertals. The variant that drove this sweep remains unknown, but is likely to be either upstream or downstream of the region studied to date. It is possible that the selected variant has not reached fixation in humans, but since genome scans have failed to find evidence for ongoing selection at *FOXP2* (Voight et al. 2006; Wang et al. 2006) and since the polymorphism patterns seen at *FOXP2* are more compatible with a finished or nearly finished sweep, this seems less likely. That two selective events affected a single gene during human evolution is obviously less likely than that a single event did so. However, multiple selective events have been suggested for other genes in humans (Clark et al. 2003; Bustamante et al. 2005; Nielsen et al. 2005) and as much as 10% of the human genome may carry signatures of recent selection (Williamson et al. 2007). In any event, that two selective events may have affected a human gene involved in aspects of speech and language is intriguing. It will therefore be important to analyze any functional elements located upstream or downstream of exon 7 that carry differences between current humans and Neandertals.

Acknowledgements

We are grateful to Barbara Höffner and Sebastian Werner for technical assistance, to Graham Coop, Jeff Jensen, Gil McVean, Rasmus Nielsen, Pavlos Pavlidis, Pleuni Pennings, Molly Przeworski, David Reich and Wolfgang Stephan for helpful discussions, and to Jeff Jensen, Phillip Johnson, Rasmus Nielsen, Pleuni Pennings, Noah Rosenberg, Weiwei Zhai and anonymous reviewers for comments on earlier versions of this manuscript. This work was supported by the Max Planck Society and the European Commission's Sixth Framework Programme for New and Emerging Science and Technology (PKB140404). IH is supported by the Human Frontier Science Program, Grant LT00794.

Literature Cited

- Bustamante CD, Fledel-Alon A, Williamson S, et al. (14 co-authors). 2005. Natural selection on protein-coding genes in the human genome. *Nature*. 437:1153-1157.
- Clark AG, Glanowski S, Nielsen R, et al. (17 co-authors). 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*. 302:1960-1963.
- Coop G, Bullaughey K, Luca F, Przeworski M. 2008. The timing of selection at the human FOXP2 gene. *Mol Biol Evol*. 25:1257-1259.
- Enard W, Hammerschmidt K, Brückner M, et al. (56 authors). 2009. A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. *Cell*. 137:961-971.
- Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Pääbo S. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*. 418:869-872.
- Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics*. 155:1405-1413.
- Hermisson J, Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*. 169:2335-2352.
- Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. *Proc Natl Acad Sci USA*. 101:10667-10672.
- Jensen JD, Thornton KR, Bustamante CD, Aquadro CF. 2007. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics*. 176:2371-2379.
- Jones DA, Wakeley J. 2008. The influence of gene conversion on linkage disequilibrium around a selective sweep. *Genetics*. 180:1251-1259.
- Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics*. 167:1513-1524.

- Krause J, Lalueza-Fox C, Orlando L, et al. 2007. (13 co-authors). The derived FOXP2 variant of modern humans was shared with Neandertals. *Curr Biol.* 17:1908-1912.
- McVean G. 2007. The structure of linkage disequilibrium around a selective sweep. *Genetics.* 175:1395-1406.
- Nickel GC, Tefft DL, Goglin K, Adams MD. 2008. An empirical test for branch-specific positive selection. *Genetics.* 179:2183-2193.
- Nielsen R, Bustamante C, Clark AG, et al. (13 co-authors). 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 3:e170.
- Noonan JP, Coop G, Kudaravalli S, et al. (11 co-authors). 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science.* 314:1113-1118.
- Przeworski M. 2002. The signature of positive selection at randomly chosen loci. *Genetics.* 160:1179-1189.
- Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution.* 59:2312-2323.
- Serre D, Langaney A, Chech M, Teschler-Nicola M, Paunovic M, Mennecier P, Hofreiter M, Possnert G, Pääbo S. 2004. No evidence of Neanderthal mtDNA contribution to early modern humans. *PLoS Biology.* 2:313-317.
- Stephan W, Song YS, Langley CH. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics.* 172:2647-2663.
- Stringer CB, Hublin J. 1999. New age estimates for the Swanscombe hominid, and their significance for human evolution. *J Hum Evol.* 37:873-877.
- Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* 16:702-712.
- Vargha-Khadem F, Gadian DG, Copp A, Mishkin M. 2005. FOXP2 and the neuroanatomy of speech and language. *Nat Rev Neurosci.* 6:131-138.

- Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.
- Wang ET, Kodama G, Baldi P, Moyzis RK. 2006. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc Natl Acad Sci USA.* 103:135-140.
- Watkins K, Dronkers NF, and Vargha-Khadem F. 2002. Behavioural analysis of an inherited speech and language disorder: comparison with acquired aphasia. *Brain.* 125:452-464.
- Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet.* 3:e90.
- Zhang J, Webb DM, Podlaha O. 2002. Accelerated protein evolution and origins of human-specific features: *Foxp2* as an example. *Genetics.* 162:1825-1835.

Figure legend

Fig. 1. LD structure around FOXP2. (A) Schematic figure illustrating how crossing-over events cause alleles to occur at low-frequency at sites around the beneficial allele (star). Since such crossing-over events occur independently on the two sides of the beneficial allele, they eliminate linkage disequilibrium (LD) across the selected site. (B) Schematic figure of the part of *FOXP2* surrounding exon 7, which encodes the two amino acid substitutions that occurred during human evolution. Positions of polymorphic sites are indicated by bars (see also Table S1) and the number of inferred haplotypes carrying derived (D) and ancestral (A) alleles in 30 chromosomes are shown. The observed haplotype configuration is unlikely under this simple model of crossing-over. Below, strong LD (R^2 and $D' = 1$, $p < 0.001$) is indicated by red and medium LD ($R^2 = 0.36$, $D' = 0.74$; $p = 0.006$) by pink.

Figure 1



