Human Genetics and Languages

David Comas, *Universitat Pompeu Fabra, Barcelona, Spain*
Elena Bosch, *Universitat Pompeu Fabra, Barcelona, Spain*
Francesc Calafell, *Universitat Pompeu Fabra, Barcelona, Spain*

Advanced article

Article Contents

- Correlation between Genes and Languages
- Genetic and Linguistic Landscapes
- Exceptions in the Correlation between Genes and Languages

Online posting date: 15th July 2008

The similarities between the mode of inheritance and evolution of genes and languages have fostered interest in the joint analysis of both disciplines. The correlation between genes and languages was firstly demonstrated by Cavalli-Sforza and collaborators comparing a tree build from 'classical' genetic markers to a linguistic tree of languages. Several criticisms to this correlation have been raised and a large number of exceptions have been described. It has been shown that the most plausible factor that influences in the correlation of genes and languages is geography. However, the information provided by genetics and linguistics, as well as the one provided by other disciplines, will allow us to reconstruct the history of humankind.

The link between the evolution of the human species and languages was already pointed by Charles Darwin (1859) in *Origin of Species*, where he suggested that the reconstruction of the human evolutionary tree would shed light into the classification of human languages. The data provided by palaeoanthropology and the development of molecular anthropology have allowed us to reconstruct the evolutionary tree of extant humans in a detailed manner. However, the reconstruction of the language tree, particularly in its deepest branches, remains sketchy and controversial, due to the much faster evolution rate of languages, and, as we will discuss later, to different processes that make language evolution less tree-like than that of genes.

Undoubtedly, speech is one of the features that defines humankind. It is beyond the scope of the present article to discuss the anatomical, physiological, neurological and, ultimately, genetic features that make speech possible, since we will focus not on what is shared by almost all humans (namely, the ability to produce speech), but on what is diverse among humans (that is, the language that is spoken). However, we will summarize some of the recent, crucial findings in the evolution of speech. The *FOXP2* gene has been found to be involved in language capability (Lai *et al*., 2001) and molecular analyses have pointed to a recent evolution of this gene in the human lineage (Enard *et al*., 2002). These analyses have shown that this gene has been the target of recent adaptation in humans compared to other primates, suggesting that the expansion of modern humans, approximately 200 000 years ago, could have been fostered by the acquisition of a sophisticated speech. However, the recent deoxyribonucleic acid (DNA) data provided by the analysis of two Neanderthal individuals shows that the derived alleles found in modern humans for the *FOXP2* gene are shared with Neanderthals (Krause *et al*., 2007). Therefore, the appearance of these changes in the *FOXP2* gene predates the common ancestor of Neanderthals and modern humans.

Human speech is a universal characteristic: all human groups have developed a sophisticated language, although some of them have not developed a writing system until recently. Almost all humans have the same biological aptitude to acquire a given language. However, there is no genetic determination of the language spoken, and each individual learns the vocabulary and grammar of a specific language depending on the cultural background he/she is born into. This is especially patent in child adoptions: the adopted children easily learn their new parents' language. Languages are part of human culture, and therefore, they are transmitted from parents to offspring, from one generation to the next generation, in the same way as other cultural traits, such as religion, technology or ethical rules, are transmitted. This fact has profound implications in the correlation of genes and languages. Owing to their cultural mode of inheritance and transmission, languages can be easily influenced, modified or replaced by other languages, and these changes can happen as fast as less than a generation.

The reconstruction of the ancestry of two individuals, groups, populations or species can be achieved using evolutionary genetic tools. The principle in its simplest form is the following: the more genetically similar two entities are the more recent is their common ancestor. This basic principle implies that genetic differences accumulate

ELS subject area: Evolution and Diversity of Life

How to cite:
Comas, David; Bosch, Elena; and, Calafell, Francesc (July 2008) Human Genetics and Languages. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0020810

through time and, therefore, the longer the time passed since the split of two entities, the larger the genetic differences between them. In a parallel way, the linguistic similarity that we observe nowadays in spoken languages points to their common origin from an ancestral language, also called a proto-language. Besides the parallelisms between linguistics and genetics, linguistic replacement might be more rapid than genetic change, since genetic transmission is vertical, i.e. from individuals to their offspring, whereas linguistic transmission might be vertical and horizontal, i.e. from individuals to any other subject (related or not) in the population. Moreover, some processes, both in genetics and linguistics, can mask the extant differences between populations, making distantly related populations appear much closer to each other as they really are. In genetics, convergent evolution, homoplasy, adaptation or admixture may diminish or erase the genetic differences accumulated through time since the split of two populations. For instance, populations such as central Africans and Australian aborigines, who split more than 50 000 years ago, may present similar genetic adaptations to solar radiation in genes related to skin colour, despite their remote common ancestors. In the same way, processes such as linguistic borrowing or language replacement might show linguistic similarities due to recent events not related to the split of languages.

Both disciplines, genetics and linguistics, use similar methods to reconstruct the histories of genes and languages, respectively. Genetic and linguistic distances can be calculated in different ways in both disciplines and represented them afterwards as trees. Genetic distances can be calculated from allele frequencies or DNA differences. In a similar way, linguistics can use list of common words, such as the Swadesh list (**Table 1**), to quantify differences between languages. Although trees provide a nice graphical way to represent data, they impose a bifurcating model onto a distance matrix that may not have such a structure, in particular in closely related entities. In addition, trees cannot represent processes such as linguistic borrowings or population admixture events, thus limiting and biasing some of the data representations. Nonetheless, if the purpose of the analysis is to determine whether a correlation between genes and languages exists, trees are not actually needed: a correlation between (genetic and linguistic) distances can be calculated directly from the raw distance matrices, by means of the Mantel test. Moreover, such an analysis can

Table 1 A 25-word subset of Swadesh's list in English, French, German, Italian, Catalan, Dutch, Swedish and Latin

No	English	French	German	Italian	Catalan	Dutch	Swedish	Latin
1	I	je	ich	io	jo	ik	jag	ego
2	you (sing.) thou	tu, vous (formal)	du, Sie (formal)	tu, lei (formal)	tu, vostè, vós (formal)	jij, je, U (formal)	du	tu
3	He	il	er	lui, egli	ell	hij	han	is, ea
4	we	nous	wir	noi	nosaltres	wij,we	vi	nos
5	you (pl.)	vous	ihr, Sie (formal)	voi	vosaltres, vostès (formal)	jullie	ni	vos
6	they	ils, elles	sie	loro, essi	ells, elles	zij, ze	de	ii, eae
7	this	ceci	dieses	questo	aquest	deze, dit	dethär	hic, is
8	that	cela	jenes	quello	aquell	die, dat	detdär	ille
9	here	ici	hier	qui, qua	aquí	hier	här	hic
10	there	là	dort	là	allà	daar	där	ibi
11	who	qui	wer	chi	qui	wie	vem	quis
12	what	quoi	was	che	què	wat	vad	quid
13	where	où	wo	dove	on	waar	var	ubi
14	when	quand	wann	quando	quan	wanneer	när	quando
15	how	comment	wie	come	com	hoe	hur	quam, quomodo
16	not	ne…pas	nicht	non	no	niet	inte, ej	non
17	all	tout	alle	tutto	tot	al, alle	alla	omnis
18	many	plusieurs	viele	molti	molts	veel	många	multi
19	some	quelques	einige	alcuni	alguns, uns	enkele, sommige	några, vissa	aliqui, aliquot
20	few	peu	wenige	pochi	poc	weinig	få	pauci
21	other	autre	andere	altro	altres	ander	annan	alter, alius
22	one	un	eins	uno	un	een	ett	unus
23	two	deux	zwei	due	dos	twee	två	duo
24	three	trois	drei	tre	tres	drie	tre	tres
25	four	quatre	vier	quattro	quatre	vier	fyra	quattuor

allow for discounting confounding factors such as geographical distance; see examples of this approach in Rosser *et al*. (2000), Rubicz *et al*. (2002), Karafet *et al*. (2002), Ayub *et al*. (2003) and Wood *et al*. (2005), among others.

Correlation between Genes and Languages

Human demographic expansions, fuelled by technological or cultural innovations, might produce the spread of people and, therefore, the spread of genes and languages to a new territory. The colonization of a new territory as a result of a demographic expansion implies at first a genetic and linguistic similarity among the colonizers that could subsequently diverge. Therefore, we can infer that vast populations sharing a language or a group of closely related languages might be the result of population expansions occurring so recently that there has been no time for language divergence. In the same way, populations occupying a territory and speaking diverse languages of the same linguistic family could be inferred to have an ancient origin (**Figure 1**). Several well-known examples of the spread of language families are the cases of the Indo-European, Bantu or Austronesian. These linguistic expansions correlate almost perfectly with genetic expansions, in the way that populations speaking nowadays these groups of languages are also genetically closely related. Most of these expansions might have been triggered by the spread of food producers, and therefore the spread of the ancestral language spoken by these groups, into the territories occupied by hunter-gatherer populations. As a result of this process, the extant distribution of these language families (and genes) is wide. The distribution of the Indo-European, the Sino-Tibetan, the Bantu and the Austronesian family languages have been associated with the agriculture expansion from the Middle East, north of China, the border between Nigeria and Cameroon and the south of China, respectively. After these expansions, the languages (and genes) might have started to differentiate and subsequently, what we observe today is a group of contiguous distributions of related languages and genes. Nonetheless, the rationale of joint spread of languages and genes linked to a cultural innovation such as food production is as valid as one of its major premises, namely, the demic diffusion of the agriculture, where the cultural innovation is carried by people, and therefore jointly carried by genes. However, other processes such as acculturation, which implies the diffusion of cultural innovations without population replacement, might lead to cultural replacement (that might include language) without genetic replacement.

The Neolithic expansions are one category of population movements that could have spread simultaneously genes and languages. Renfrew (1994) produced a general framework in which the current geographical location of the major linguistic families is linked to one or more of a few (pre)historical population movements. In summary: (i) some linguistic families (Khoisan, Nilo-Saharan, Caucasian, Austric, Australian, Indo-Pacific and Amerindian; see **Figure 1**) owe their geographic locations solely to the initial spread of anatomically modern humans in the Upper Palaeolithic; (ii) Neolithic dispersals would have spread the Nigero-Kordofanian, Afro-Asiatic, Indo-European, Elamo-Dravidian, Sino-Tibetan and Austronesian families; (iii) a milder climate in the higher boreal latitudes may have lead to the colonization of northern regions and to the

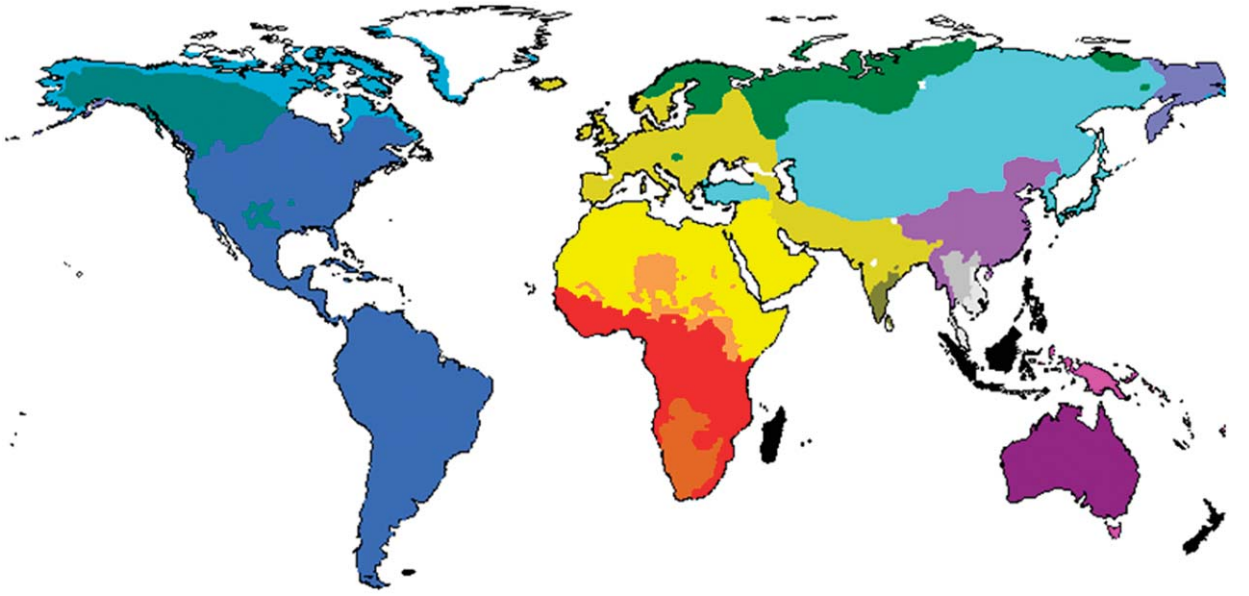

Figure 1 Map of main human linguistic families. Adapted from Ruhlen (1991).

expansion of the Uralic, Chukchi-Kamchatkan, Na-Dene and Eskimo-Aleut families and (iv) later movements under an elite-dominance model, that is, in which a well-structured minority can rule over a territory and impose a culture and a language; that was the case for the expansion of the Altaic family from a reduced homeland in Central Asia, or of the Indo-European family to South Asia, or the introduction of the same Indo-European family to vast tracts of the world in the post-Columbian European colonization.

The joint spread of agriculture and people (or, for that matter, any of the (i)–(iii) processes mentioned above) might create a correlation between genetic and linguistic evolutionary distances or trees. Cavalli-Sforza (1997) and Cavalli-Sforza *et al.* (1998) were the first to demonstrate this correlation between genes and languages in human populations. Data of around 120 'classical' (i.e. those detected in gene products rather than in genes themselves, such as blood groups and other protein polymorphisms) genetic polymorphisms from 42 worldwide populations was collected and genetic distances represented in an evolutionary tree. Subsequently, this genetic tree was compared to a linguistic tree of languages (Ruhlen, 1991) and a high correlation between both trees was shown (**Figure 2**). A few exceptions to this global correlation were found, such as populations belonging to the same linguistic family being very genetically different, or, on the contrary, genetically close populations belonging to different language

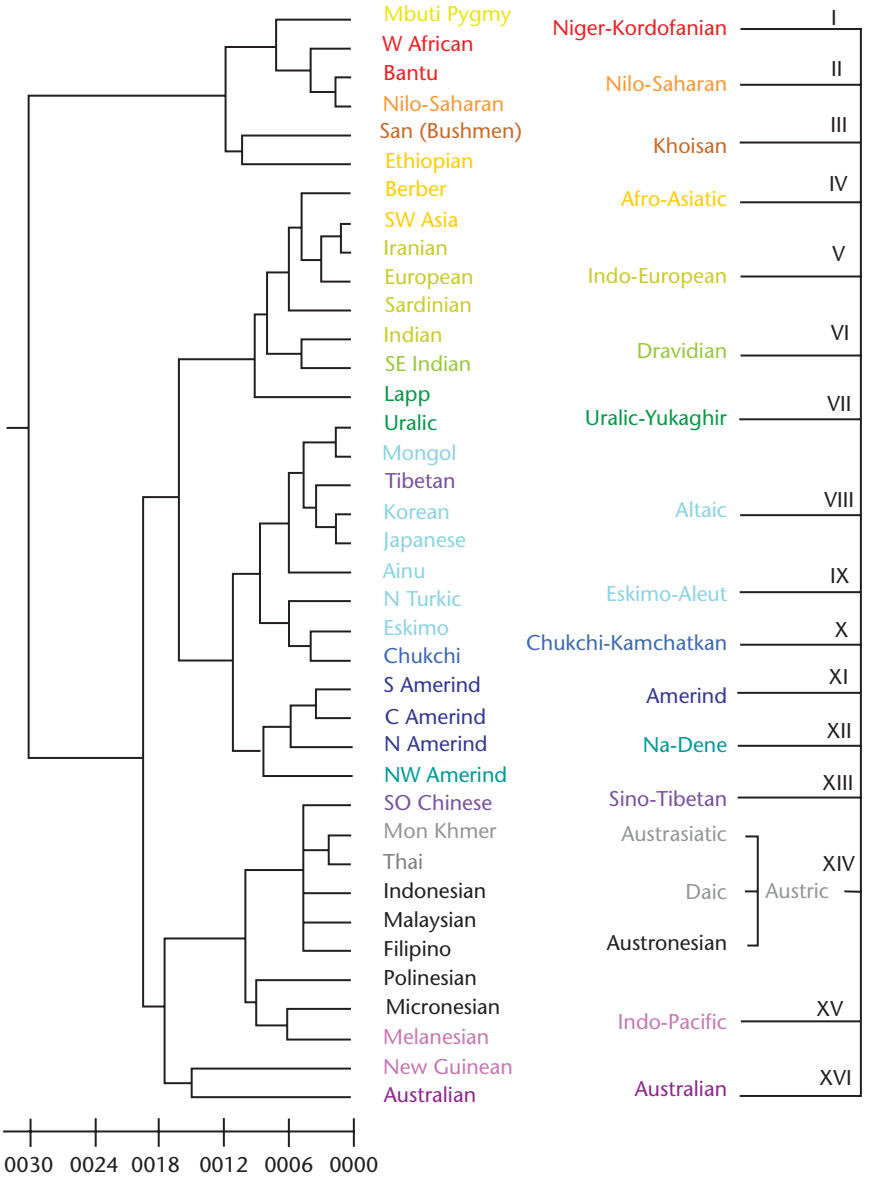

Figure 2 Trees relating the genetic (left) and linguistic (right) human diversity. Continental areas in white represent either noninhabited areas or linguistic isolates. Linguistic families are coloured as in **Figure 1**.

families. However, several criticisms have been raised concerning the data and the methodology used by Cavalli-Sforza and collaborators (Bateman *et al*., 1990). The arbitrarily defined populations and languages used, the low sample sizes and the tree reconstruction are some of the points criticized.

Despite the numerous exceptions reported, the correlation between languages and genes remains good. However, this correlation can be attributed to factors other than the direct link between languages and genes, such as the correlation of both entities to a third variable. The most plausible factor that influences in the correlation of genes and languages is geography: if genes are correlated to geography because human movements, expansions and migrations are channelled by geographical features, and languages are diffused with the limitations of the very same geographical barriers; therefore, genes and languages could be correlated as a result of geography. Some genetic data has shown high correlation with geography but not with languages. For instance, Rosser *et al*. (2000) showed that the extant genetic diversity on the Y-chromosome in Europe displays a clinal pattern highly correlated with geography but not with languages, suggesting a population expansion from the Middle East and linguistic heterogeneities.

A differential correlation between genes and languages depending on the genetic marker analysed has been found and has potent implications. It has been shown that the genetic diversity found in the Y-chromosome correlates better with the linguistic classification than mitochondrial DNA does (Poloni *et al*., 1997; Perez-Lezaun *et al*., 1999; Bosch *et al*., 2006). This fact could be explained by the different migration rates between males and females. It has been postulated (Seielstad *et al*., 1998) that the female migration rate is higher than male rate, which would explain why male lineages are more geographically structured, that is, more heterogeneous in space, than female lineages. Although the native language of speakers is usually called their 'mother' tongue, the better correlation of linguistic diversity with male-transmitted genetic diversity suggests that 'father' tongue would be a more appropriate phrase.

Genetic and Linguistic Landscapes

The genetic composition of populations, that is, their allele or haplotype frequencies at relevant loci, changes in space. Obviously, a major determinant of the rate of change is the geographical obstacles to population movement (and, thus, to gene flow) such as oceans or mountain ranges. However, linguistic differences can also play a role in the creation of the genetic landscape, by difficulting the mating of people speaking different languages and reinforcing endogamy within a linguistic group. Barbujani and Sokal (1990) analysed the rate of change in space of allele frequencies in Europe, determined where the genetic change was steepest, and found that such genetic borders coincided with linguistic boundaries more often than expected, independently of geographical barriers, which implied that languages had indeed restricted gene flow across populations in Europe. In particular, a clear genetic boundary was found around the Basques (Calafell and Bertranpetit, 1994), a linguistic and cultural isolate that is not separated from the rest of Iberia by any insurmountable physical barrier. A similar approach was also applied to populations in Italy (Zei *et al*., 1993; in this case, genetic differentiation was estimated from surnames, which can be modelled as a single locus with many alleles at the Y-chromosome), Britain (Falsetti and Sokal, 1993) and Japan (Sokal and Thomson, 1998).

Exceptions in the Correlation between Genes and Languages

The correlation between genes and languages has become to be expected and can be regarded, paraphrasing statistics, as a null hypothesis; thus, exceptions to the rule become more interesting and provide insights into many local population histories. The exceptions in the seminal studies by Cavalli-Sforza and coworkers were attributed to phenomena of language or population replacement. Some well-known examples of language replacement are the Pygmies, who speak languages of the Niger-Congo or the Nilo-Saharan family, while they remain clearly genetically differentiated from non-Pygmy speakers of the same language groups. Examples of language replacement abound in former European colonies, where many native groups managed to retain their genetic distinctiveness, while their languages were replaced with those of the colonial powers. These are examples of the elite-dominance model in Renfrew's general framework. In this case, a limited group of individuals take the political, religious or social rule of a general population imposing a new language. As a consequence, a linguistic replacement might take place without the genetic replacement of the population. This process is only possible in very structured populations with a social hierarchy, where the replacement of a small amount of leading individuals might affect the general population. A further example of this elite language replacement model is the Turkic language spoken nowadays in Turkey. In the eleventh century AD, Turkic tribes coming from Central Asia imposed their Turkic language, replacing Greek and other Indo-European languages and leading some of these to extinction. Nowadays, the geographical distribution of the Indo-European family of languages presents a gap in Turkey due to this language replacement, but the genetic composition of Turks is closer to other Middle Eastern and European samples rather than to Central Asian populations (Calafell *et al*., 1996; Comas *et al*., 1996; Cinnioglu *et al*., 2004). Similarly, the cultural descendants of the Romans in the Balkans, that is, Romance-speaking Romanians and Aromuns, are more closely genetically related to non-Romance Balkan populations than to Italians (Bosch *et al*., 2006).

The joint analysis of genes and languages is in the crossroad of the knowledge provided by several distant disciplines beyond genetics and linguistics, such as archaeology, cognitive sciences, development and palaeontology, among others. The final consequence of unravelling the relationship between genes and languages will lead us to the reconstruction of not only the migrations and cultural exchanges between human populations, but to the unique evolutionary history of our species.

References

Ayub Q, Mansoor A, Ismail M *et al*. (2003) Reconstruction of human evolutionary tree using polymorphic autosomal microsatellites. *American Journal of Physical Anthropology* **122**(3): 259–268.

Barbujani G and Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. *Proceedings of the National Academy of Sciences of the USA* **87**(5): 1816–1819.

Bateman R, Goddard I, O'Grady R *et al*. (1990) Speaking of forked tongues: the feasibility of reconciling human phylogeny and the history of language. *Current Anthropology* **31**(1): 1–13.

Bosch E, Calafell F, Gonzalez-Neira A *et al*. (2006) Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. *Annals of Human Genetics* **70**: 459–487.

Calafell F and Bertranpetit J (1994) Principal component analysis of gene frequencies and the origin of the Basques. *American Journal of Physical Anthropology* **93**(2): 201–215.

Calafell F, Underhill P, Tolun A, Angelicheva D and Kalaydjieva L (1996) From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *Annals of Human Genetics* **60**: 35–49.

Cavalli-Sforza LL (1997) Genes, peoples, and languages. *Proceedings of the National Academy of Sciences of the USA* **94**: 7719–7724.

Cavalli-Sforza LL, Piazza A, Menozzi P and Mountain J (1998) Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences of the USA* **85**: 6002–6006.

Cinnioglu C, King R, Kivisild T *et al*. (2004) Excavating Y-chromosome haplotype strata in Anatolia. *Human Genetics* **114**: 127–148.

Comas D, Calafell F, Mateu E, Pérez-Lezaun A and Bertranpetit J (1996) Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations. *Molecular Biology and Evolution* **13**(8): 1067–1077.

Darwin C (1859) *On the Origin of Species*. London: John Murray.

Enard W, Przeworski M, Fisher SE *et al*. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**: 869–872.

Falsetti AB and Sokal RR (1993) Genetic structure of human populations in the British Isles. *Annals of Human Biology* **20**(3): 215–229.

Karafet TM, Osipova LP, Gubina MA *et al*. (2002) High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Human Biology* **74**(6): 761–789.

Krause J, Lalueza-Fox C, Orlando L *et al*. (2007) The derived FOXP2 variant of modern humans was shared with Neandertals. *Current Biology* **17**(21): 1908–1912.

Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F and Monaco AP (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**: 519–523.

Perez-Lezaun A, Calafell F, Comas D *et al*. (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *American Journal of Human Genetics* **65**(1): 208–219.

Poloni ES, Semino O, Passarino G *et al*. (1997) Human genetic affinities for Y-chromosome P49a,f/Taql haplotypes show strong correspondence with linguistics. *American Journal of Human Genetics* **61**: 1015–1035.

Renfrew C (1994) World linguistic diversity. *Scientific American* **270**: 116–123.

Rosser ZH, Zerjal T, Hurles ME *et al*. (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *American Journal of Human Genetics* **67**: 1526–1543.

Rubicz R, Melvin KL and Crawford MH (2002) Genetic evidence for the phylogenetic relationship between Na-Dene and Yeniseian speakers. *Human Biology* **74**(6): 743–760.

Ruhlen M (1991) *A Guide to the World's Languages*, 2nd edn. Stanford: Stanford University Press.

Seielstad MT, Minch E and Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. *Nature Genetics* **20**(3): 278–280.

Sokal RR and Thomson BA (1998) Spatial genetic structure of human populations in Japan. *Human Biology* **70**(1): 1–22.

Wood ET, Stover DA, Ehret C *et al*. (2005) Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *European Journal of Human Genetics* **13**(7): 867–876.

Zei G, Barbujani G, Lisa A *et al*. (1993) Barriers to gene flow estimated by surname distribution in Italy. *Annals of Human Genetics* **57**: 123–140.

Further Reading

Belle EM and Barbujani G (2007) Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity. *American Journal of Physical Anthropology* **133**: 1137–1146.

Diamond J and Bellwood P (2003) Farmers and their languages: the first expansions. *Science* **300**: 597–603.

Gray RD and Atkinson QD (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**: 435–439.

Saha A, Udhayasuriyan PT, Bhat KV and Bamezai R (2003) Analysis of Indian population based on Y-STRs reveals existence of male gene flow across different language groups. *DNA and Cell Biology* **22**: 707–719.

Sahoo S and Kashyap VK (2005) Influence of language and ancestry on genetic structure of contiguous populations. *BMC Genetics* **6**(1): 4.