Genetic and linguistic borders in the Himalayan Region

Thirsa Kraaijenbrink[1], Emma J. Parkin[2], Denise R. Carvalho-Silva[3], George L. van Driem[1], Guido Barbujani[4], Chris Tyler-Smith[3], Mark A. Jobling[2], Peter de Knijff[1,5]

There are a number of competing theories about the origins of the Himalayan peoples. These theories are largely based on linguistic and/or archaeological findings, sometimes supported by the results of small-scale genetic studies. A large-scale, ethnolinguistically-informed genetic study of the greater Himalayan region might provide a definitive model for historical population events in this region, and that is why the current study was undertaken.
The geographical area of the present-day states of Nepal and Bhutan could have served as ancient corridors for human migration through the Himalayas despite their geographical position immediately south of the highest land barrier. The findings also raise the question as to whether the southern slopes of the Himalayas could have harboured a myriad of refuge areas for the ancestral Tibeto-Burman population(s) during the last glacial maximum. Alternatively, if the multitude and diversity of language communities found in these countries is a reliable indication, they could be an ancient source of genetically differentiated populations and languages. A detailed genetic study of the Himalayan region, therefore, may not only provide insights into the uniqueness and antiquity of its residents, but may also shed light on the peopling of the Himalayas and eastern Asia in general.

1. Department of Human and Clinical Genetics, Leiden University Medical Centre, Leiden, The Netherlands.

2. Department of Genetics, University of Leicester, United Kingdom.

3. The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom.

4. Dipartimento di Biologia, Università di Ferrara, Italy.

5. Corresponding author: P. de Knijff, Department of Human and Clinical Genetics, Leiden University Medical Centre, PO Box 9600, 2300 RC Leiden, The Netherlands. Phone + 31 71 526 9540, e-mail knijff@lumc.nl.

Using genetic data from 15 autosomal Short Tandem Repeat (STR) loci, we provide evidence that there is clear congruence between language and genetics. Populations speaking a language belonging to the Tibeto-Burman language family are genetically more similar to each other than to populations speaking a language belonging to the Indo-European language family. On the basis of language differences we can draw a linguistic boundary roughly running from east to west, just south of the border between India and Bhutan and through southern Nepal. A genetic boundary can be reconstructed along nearly the same route. We conclude from all these analyses that the populations of Nepal and Bhutan are likely to have originated outside their current locations, in regions where their language families are spoken, but need further work to suggest more precise origins.

Introduction

Isolation is crucial to both biological and linguistic change. Populations that are separated by barriers tend to diverge genetically because of genetic drift, and to undergo independent linguistic change, resulting in often parallel patterns of genetic and cultural differentiation. Geographical as well as cultural barriers reduce population contacts, thereby potentially increasing isolation between populations. However, both biological and linguistic change are influenced by the size of the population. It is thus also important to infer reliable information on past human demography.

The greater Himalayan region is ethnolinguistically the most complex area of the Eurasian continent. This region includes the highest land barrier on the face of this planet, and linguistic evidence unambiguously indicates that topography has shaped and channelled prehistoric population movements. The intricate ethnolinguistic mosaic of this region holds many keys to the peopling of the Eurasian continent as a whole. Whereas most language communities in the Himalayan region belong either to the Tibeto-Burman or Indo-European family, there are also Austroasiatic, Dravidian, Daic and Altaic language communities settled in the mountains, foothills and periphery of the Himalayas. Moreover, there are two language isolates, Burushaski and Kusunda, in the region. Linguistically, the Himalayas are sometimes thought to form the border between the Indo-European and Tibeto-Burman language families, though in fact the real linguistic border roughly runs parallel to the range through the hills and lowlands to the south (van Driem 2001). Some genetic studies have indicated the presence of a genetic barrier in this area, but these studies have mainly included population samples from China and India and not from populations within the Himalayan heartland: Nepal and Bhutan (Cordaux et al. 2004, Metspalu et al. 2004, Xue et al. 2006).

A very few studies include some Himalayan population samples (Cavalli-Sforza et al. 1994, Gayden et al. 2007) but were unable to sample extensively in this area.

The geographical area of the present-day states of Nepal and Bhutan could have been corridors for human migration through the Himalayas in ancient times despite their geographical position immediately south of the highest land barrier, the Himalayan mountain range: for people adapted to life at this altitude, they provide the most inviting localities. Or they could be seen as regions where human existence is difficult, and inhabited late in prehistory. Alternatively, if the multitude of language communities found in these countries is a reliable indication, they could be an ancient source of genetically differentiated populations and languages, a possible consequence of subdivision and extreme isolation over long periods. A detailed genetic study of the Himalayan region, therefore, may not only provide evidence for the uniqueness and antiquity of its residents, but may also shed light on the peopling of the Himalayas and eastern Asia in general.

In order to be able to analyse the possible correlation between the complex linguistic relationships and the genetic affinities among the many Himalayan populations and those of their neighbouring regions, we embarked upon two sampling expeditions to Nepal and Bhutan with the aim of providing answers to three major questions:

- Is there a correlation between language, genes and geography in the Himalayan region?
- Can we determine the genetic relationships (ancient ancestors) of the Nepalese and Bhutanese and deduce possible migration routes?
- Can we say something about relative ages of the various groups now living there, identifying and comparing "aboriginal" groups with the others?

In this article we will provide details of the expeditions, the samples, and the genetic systems tested. We will also describe in some detail the first autosomal DNA results. The analyses of mtDNA and Y-chromosomal data are not yet completed and will be described in a future publication.

Methods

Planning of the project

Initially, the aim of the study was to organise three expeditions, to Nepal, Bhutan and North and north-eastern India (specifically: Assam, Sikkim, and Arunachal Pradesh), to collect blood from the main ethnolinguistic groups of the greater Himalayan region. Unfortunately, it eventually turned out to be impossible

– within the time frame of our funding – to collect samples in India. Therefore, the project was restricted to the analyses of the Nepalese and Bhutanese ethnolinguistic groups, as described in detail below.

We organised two major expeditions, one to Nepal and one to Bhutan. The first expedition was aimed at collecting blood samples from Nepal's populations. During this expedition, held in December 2002 and January 2003, the team was assisted by several Nepali assistants from various language communities. The work in Nepal was carried out with the knowledge and cooperation of representatives of local groups and Tribhuvan University at Kirtipur. With the valuable assistance and guidance of Prof. Dr. Nirmal Man Tuladhar (Professor of Linguistics at Tribhuvan University), representatives of the ethnolinguistic groups were contacted and asked for cooperation in the project. The names of these representatives can be found in the detailed acknowledgements.

Blood donation was on a voluntary basis, often after discussing the project with local (language) communities, and a blood-sample (10 ml) was only taken if the donor had read, understood and signed the informed consent form. When a donor was unable to read or write, the consent text was read to the donor in his/her local language, after which one of the project's co-workers filled in the donor's data on the form. For some communities, detailed explanation in the local language was given and video-recorded for archival purposes.

To ascertain that a person belonged to a certain ethnolinguistic group or caste, the donor's name and place of birth were systematically checked against what is known about the names adopted by members of Nepal's diverse ethnic groups and the geographical spread of the group with which the person identified himself or herself. In addition, several team members and most of the project's co-workers speak one or more relevant Nepali languages. To consider that a person was not admixed, both parents had to belong to the same group.

The second expedition, which was aimed at collecting blood samples from Bhutanese populations, was completed in October and November 2003. The work in Bhutan was carried out with the knowledge, permission and cooperation of the Royal Government of Bhutan at Thimphu. As in Nepal, blood donation was on a voluntary basis and sampling of 10 ml of blood was only carried out after the donor had read, understood and signed the consent form. Again, when a donor was unable to read or write, the consent text was read to the donor and his or her data were recorded by one of the project's co-workers.

Donors had been pre-selected by representatives of the Royal Government of Bhutan, based on the same criteria as were used in Nepal (the only difference being that not all ethnolinguistic groups in Bhutan use group-specific names).

Blood was collected throughout Bhutan during four field trips. Members of some major Bhutanese groups were sampled in and around Thimphu (the capital

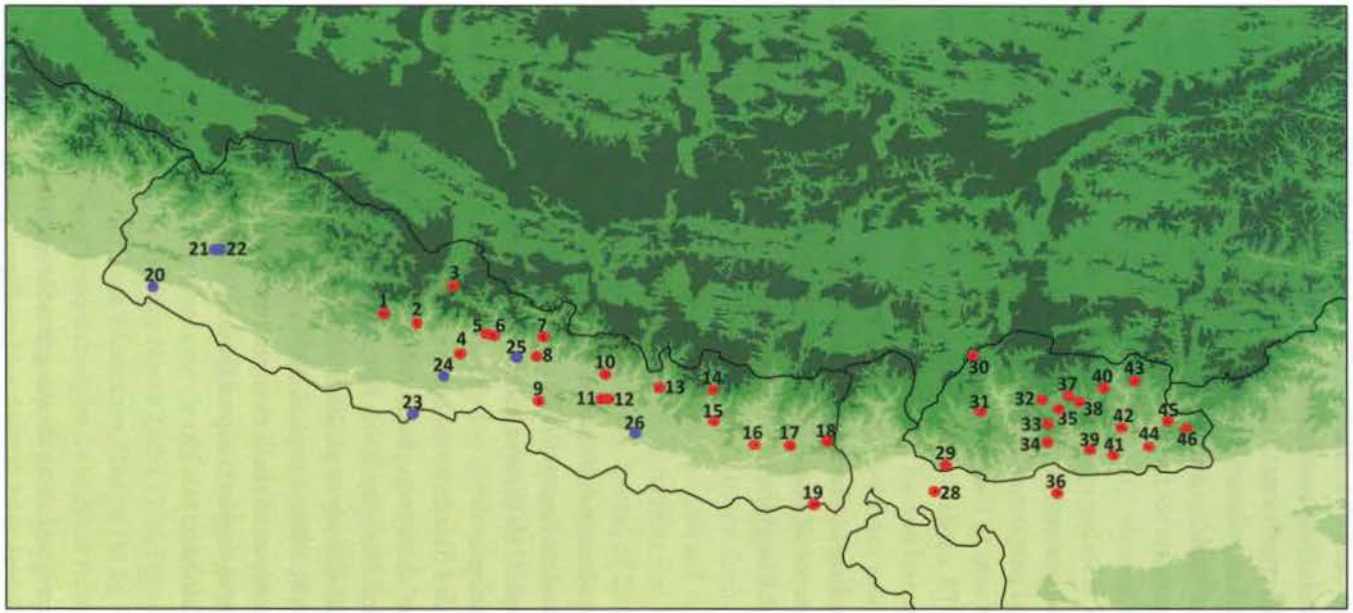

Figure 1. Distribution of ethnolinguistic groups sampled in Nepal and Bhutan

In Nepal, the blue dots reflect the Indo-European language group centres and the red dots reflect language group centres of the Tibeto-Burman speaking populations. In Bhutan, all populations speak a Tibeto-Burman language. Numbers (see also below) correspond with the numbers in Table 1 (p. 198–201). 1, Kham; 2, Chantyal; 3, Thakali; 4, Magar; 5, Gurung; 6, Dura; 7,Ghale; 8, Barâm; 9, Chepang; 10, Tamang; 11, Newar; 12, High Caste Newar; 13, Thangmi; 14, Sherpa; 15, Western Kiranti; 16, Central Kiranti; 17, Eastern Kiranti; 18, Limbu; 19, Dhimal; 20, Indo European; 21, Bahun; 22, Chetri; 23, Tharu; 24, Majhi; 25, Kumal; 26, Indo European / Tibeto Burman substrate; 28, Toto; 29, Lhokpu; 30, Layap; 31, 'Ngalop; 32, Lakha; 33, Mangde; 34, Black Mountain Mönpa; 35, Nup; 36, Bodo; 37, Bumthang; 38, Brokkat; 39, Khengpa; 40, Kurtöp; 41, Gongduk; 42, Chali; 43, Dzala; 44, Tshangla; 45, Dakpa; 46, Brokpa.

city) and when encountered during any of the four expeditions. During the various field trips in Bhutan, we were also able to collect samples from two Tibeto-Burman-speaking populations from northern India: the Bodo and the Toto. Table 1 (p. 198–201) presents descriptive statistics of the sampled individuals. Figure 1 illustrates the geographical coordinates of the centres of the various language groups sampled by us and included in our study.

Sample processing

After blood sampling, a small amount of blood was spotted in duplicate onto FTA filter-paper cards for archival purposes (4 spots of ~1 cm diameter per FTA card).

The blood samples were sent to the Netherlands via DHL as soon as possible after sampling (tubes and FTA cards were sent separately). One set of FTA cards is currently stored in Leiden and one set in Leicester for future reference.

The blood in the tubes was used for DNA isolation, using the Autopure LS* from Gentra Systems, according to the manufacturer's specifications. All blood samples yielded a sufficient amount of good quality DNA. Aliquots of all samples were shipped to the United Kingdom for Y-chromosomal and mitochondrial-DNA analyses.

Genotyping

After DNA isolation, all Nepalese and Bhutanese samples were genotyped for 21 forensic autosomal Short Tandem Repeat (STR) loci, contained in three commercially available kits: Powerplex 16 (Promega), AMPFlSTR Identifiler (Applied Biosystems) and FFFL (Promega).

To our own data, we added data from many reference populations from India and China (Kraaijenbrink et al. in prep., Rajkumar and Kashyap 2003, Gaikwad and Kashyap 2003, Neeta and Kashyap 2004, Kasyap et al. 2004, Hima Bindu et al. 2005, Krithika et al. 2006, Kasyap et al. 2006, Hima Bindu et al. 2007, Xue et al. 2006, Quintana-Murci et al. 2001, Lee et al. 2004). From many of these reference populations, only genotypes of the loci contained in the Powerplex 16 kit were available. Therefore it was decided to limit the analyses reported here to the 15 autosomal STR loci contained in this kit.

Statistical analyses

Population structure was examined using the program Structure 2.2 (Pritchard et al. 2000, Falush et al. 2003, Falush et al. 2007) based on the admixture model with correlation between allele frequencies across clusters. For each number of clusters

K, five independent Structure runs were performed, all using a burn-in of 20,000 iterations, followed by 10,000 iterations of MCMC for estimates of clustering.

Pairwise FST for all population pairs was calculated using the Excel add-in Genalex 6.1 (Peakall and Smouse, 2006). In order to compare with the results obtained using Structure, the pairwise FST values were used in multi-dimensional scaling (MDS) analyses performed with the program NCSS. The first two dimensions resulting from the NCSS analyses were used for creating an MDS plot in Excel.

From the Structure analyses with K=2, we took the estimated values of attribution to the two clusters of each of the population samples as input for a spatial distribution map using the Kriging procedure with the Surfer 8 software (Golden Software, http://www.goldensoftware.com).

Results and discussion

In total, we collected DNA samples from 947 unrelated Nepalese volunteers (764 males and 183 females) and 1029 unrelated Bhutanese volunteers (839 males and 190 females), belonging to 40 major ethnolinguistic groups from the Tibeto-Burman family, and 11 ethnolinguistic groups from the Indo-European family (Table 1 p. 198–201). All samples were analysed for the 21 forensic autosomal STRs comprised in the Powerplex16, Identifiler and FFFL kits. Due to small sample size for some population samples, a total of 23 individuals from Nepal (see Table 1 p. 198–201) were not included in this study.

The general allele frequency distributions of these autosomal STR loci for Nepal and Bhutan have been published previously (Kraaijenbrink et al. 2007a, Kraaijenbrink et al. 2007b). When analysing the data in Structure, the two-cluster model (K=2) was the best fit of our total dataset. All runs for K=2 produced the distribution shown in Figure 2A. Most populations belonging to the Indo-European family are grouped together in one cluster which is predominantly blue in Figure 2A, and most populations belonging to the Tibeto-Burman family are grouped together in the other cluster which is predominantly yellow in Figure 2A.

When we increased the number of clusters, the clear distinction between the Tibeto-Burman and the Indo-European language groups was lost. Instead, with K=6, three populations became clearly clustered individually (Figure 2B). These populations, the Toto from North India (T in Figure 2B), and the Lhokpu (L in Figure 2B) and the Black Mountain Mönpa (M in Figure 2B) from Bhutan, are known to have been almost completely isolated from their neighbouring populations until relatively recently, due to both geographical and cultural barriers, which provides a possible explanation for this separate clustering.

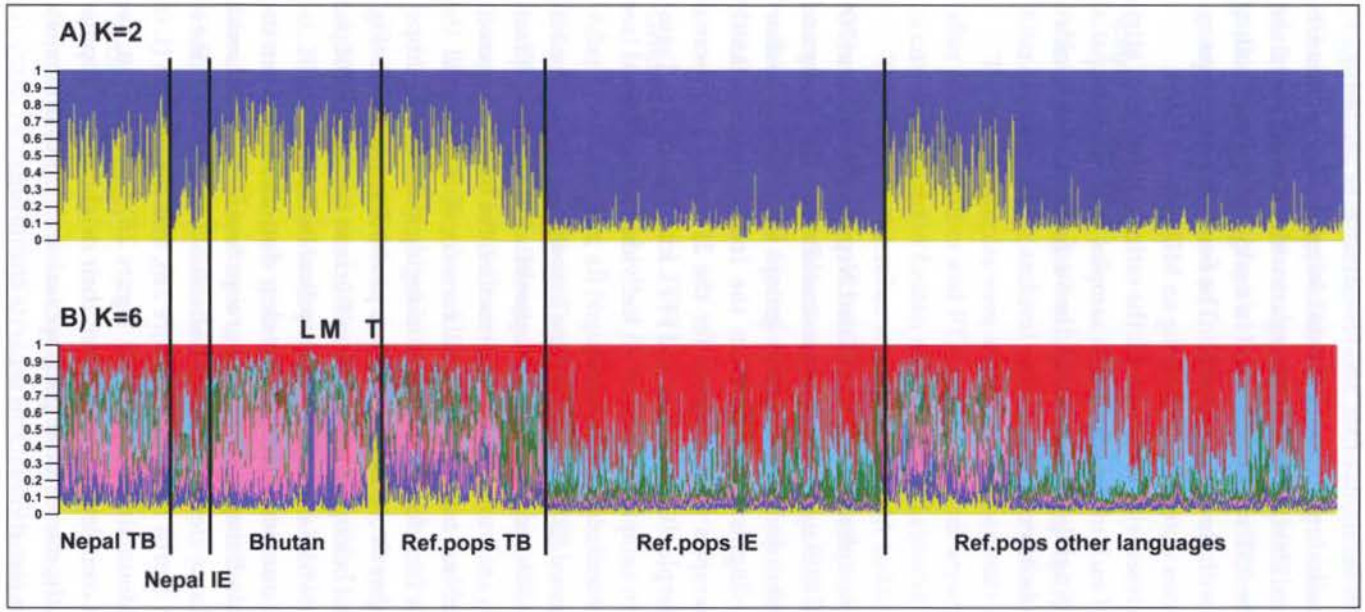

Figure 2. Results of unsupervised Structure analyses

The colours represent the proportion of inferred ancestry from K ancestral populations.
For K=2 (2A), the inferred ancestry largely corresponds to the linguistic family to which the populations belong with Tibeto-Burman speaking populations mostly being assigned to the yellow cluster and Indo-European speaking populations to the blue cluster.
For K=6 (2B) the majority of the "linguistic clustering" is lost in favour of the separation of the Lhokpu (L), Black Mountain Mönpa (M) and Toto (T), although some differences can still be observed between the Tibeto-Burman and Indo-European populations.

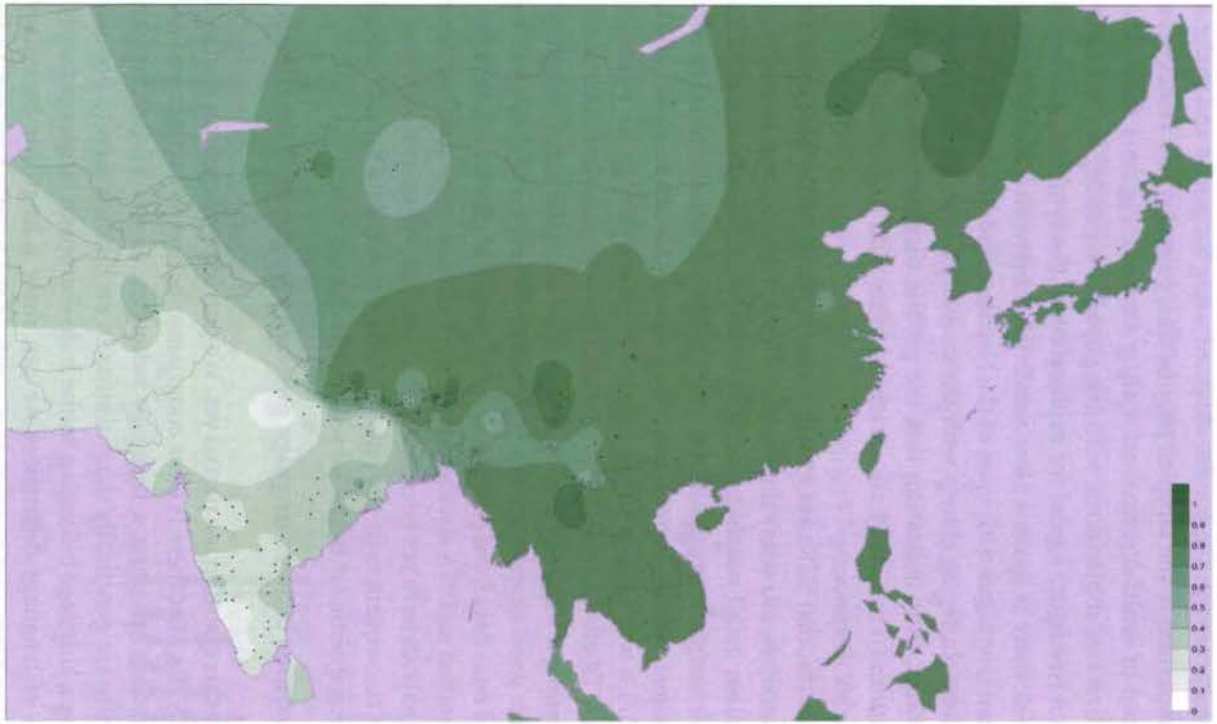

Figure 3. Geographical mapping of the unsupervised Structure K=2 results

The spatial mapping of the Structure K=2 results illustrates an approximate north-east vs. south-west clinal gradient with the steepest gradient located in the southern Himalayan foothills. The green colourscale indicates the percentage of "Tibeto-Burman" genetic contribution which is the highest in Eastern China, and the lowest in Southern India.

The spatial mapping of the Structure K=2 results (Figure 3) illustrates an approximate north-east vs. south-west clinal gradient with the steepest gradient located in the southern Himalayan foothills.

In order to compare with the results obtained using Structure and make a more detailed comparison of the sampled populations, pairwise FST values were calculated and used in multi-dimensional scaling analyses. Figure 4 shows the MDS plot of the first two dimensions, with the populations coded according to language affiliation (see figure legend for explanation of the symbols). As can be seen from Figure 4, there is again a clear subdivision between Tibeto-Burman and Indo-European languages with most of the Nepalese and Bhutanese Tibeto-Burman populations clustering closely with the majority of the Tibeto-Burman reference populations, thus indicating that the genetic distances observed between the populations in this study are correlated more with linguistic distance than with geographic distance.

Even though autosomal STRs are usually thought not to be the best tools for a refined genetic analysis, our study shows that, at least in the Greater Himalayan Region, even a rather small number (n=15) of highly-variable autosomal STRs can give a valuable insight into population (pre-) history. Based on initial results from Y-chromosomal and/or mitochondrial analyses (Metspalu et al. 2004, Gayden et al. 2007) it was already suggested that there is evidence for a genetic difference between Tibeto-Burman and Indo-European populations. Our autosomal analyses among a large number of populations from the actual language border area provide good support for this hypothesis. We expect that this will be further confirmed once data from potentially more powerful genetic markers (autosomal SNPs, detailed mtDNA data and detailed Y-chromosome data) become available. These analyses have not been completed yet, but will be available soon.

What are the consequences of the results of the present data for our initial three main research questions? Below we will briefly discuss this:

- Is there a correlation between language, genes, and geography in the Himalayan region?
  The answer to this question is a partial yes. We provide evidence, on the basis of autosomal STRs, that there is clear congruence between language and genetics. Populations speaking a language belonging to the Tibeto-Burman language family are genetically more similar to each other than to populations speaking a language belonging to the Indo-European language family. On the basis of language differences we can draw a linguistic boundary roughly running from east to west, just south of the border between India and Bhutan, and running through Nepal. A genetic boundary can be reconstructed along nearly the same route.

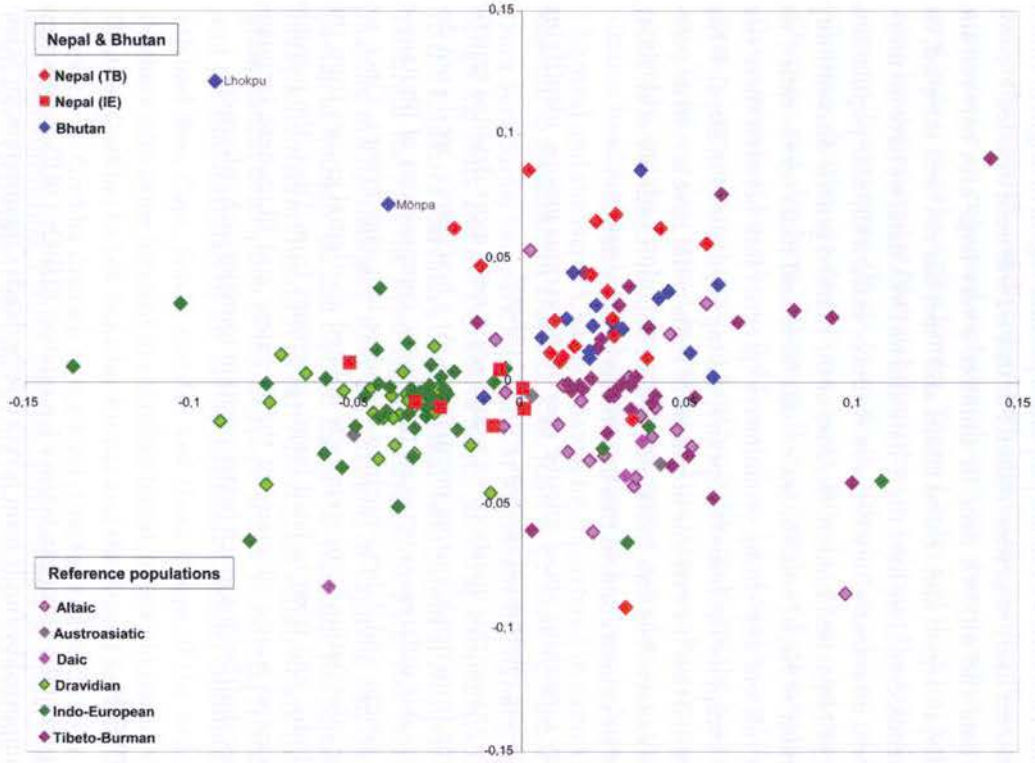

Figure 4. MDS plot

MDS plot, based on pairwise FST values between all populations. The symbol-coding legend included in the figure reveals the language affiliation of the reference populations and the populations from Nepal and Bhutan. Nearly all Indo-European and Dravidian speaking populations (including those in Nepal) are very clearly separated by the first dimension (x-axis) from populations speaking Tibeto-Burman or other languages.

- Can we determine the genetic relationships (ancient ancestors) of the Nepalese and Bhutanese and deduce possible migration routes?
  This question is very difficult to answer on the basis of these autosomal data. Generally speaking we detect a very close genetic relationship between various, mostly eastern Chinese, populations and the Nepalese and Bhutanese populations. However, we can not pinpoint a possible region of origin in China and reconstruct likely migration routes. For this we will need detailed Y-chromosomal data and mtDNA data. In addition, a very large area between the sub-Himalayan region and central China and Tibet has not been sampled. As a consequence, even if we have the Y-data and mtDNA data, we have no information about an essential link between the more eastern Chinese populations and those of Nepal and Bhutan. Because of the marked genetic discontinuity between most of Nepal and Bhutan on one hand and India on the other, we can safely rule out any strong evolutionary (genetic) link between these two countries, except for the Indo-European speaking populations in Nepal. It has been suggested that the extreme northeast of India could have served as a corridor from eastern Asia into India and perhaps Bhutan (Cordeaux et al. 2004). At present we cannot confirm this hypothesis in more detail.
- Can we say something about relative ages of the various groups, identifying and comparing "aboriginal" groups with the others?
  In order to answer this question, we would again need more detailed genetic information from neighbouring populations. At a first glance, we do not detect any notable differences between the genetic compositions of Bhutanese language groups and Tibeto-Burman Nepalese language groups. Once we have more information from seemingly isolated aboriginal groups from the north of India, and from isolated language groups from Tibet (the possible direct ancestral source of many of the Nepalese and Bhutanese language groups), we shall be in a much better position to make such inferences.

Acknowledgements

This work, as part of the European Science Foundation EUROCORES Programme OMLL, was supported by funds from NWO (Netherlands Organisation for Scientific Research, grant no. 231-70-001), the Arts and Humanities Research Board of Great Britain (grant no. AN/9585/APN15314) and the EC Sixth Framework Programme (Contract no. ERAS-CT-2003-980409). CTS is supported by The Wellcome Trust. MAJ is a Wellcome Trust Senior Fellow in Basic Biomedical Science (grant number 057559).

The research in Nepal was conducted in association with the Centre for Nepal and Asian Studies (CNAS) at Kirtipur under the Bilateral Agreement for Academic Cooperation between Tribhuvan University (TU) and Leiden University. The research team received the organisational support and enthusiastic assistance of many grass-roots community service organisations and from the many informed volunteers of indigenous language communities who took an active interest in the research and discussed with interest the ramifications of genetic investigations for an enhanced understanding of our shared prehistorical past. Special gratitude is due the Arjun Limbu, Yograj Limbu and the Kirat Yakthung Chumlung headquartered at Mahalakshmisthan in Lalitpur; the Praja Capacity Development Programme (Praja Samudayik Vikash Karyakram) at Shaktikhor in Chitwan district; to Ajay Praja, Santa Bahadur Praja and Dambar Bahadur Chepang and the Nepal Chepang (Praja) Sangh at Kathmandu; to our many Tharu friends at Sauraha; to Bal Gopal Shrestha of Sankhu and the Newar community organisation Friends of Sankhu; to Bharat Rai of the National Youth Service Trust, our dear friend Ashok Rai and the Danuwar Rai community; to Dilendra Subba and the Limbu Literary Development Association; to the Kirat Yakkha Chuma headquartered at Mahalakshmisthan in Lalitpur; to Shree Mani Chand Chantyal and the Nepal Chantyal Samaj headquartered in Samakhushi in Kathmandu; to Buddhiman Dura, Lt. Col. John P. Cross, Ritu Kumar Dura and the Dura Seva Samaj Sampark Karyalaya at Ram Bajar in Pokhara; to Kishor Dura, Singh Raj Dura and the Dura Seva Samaj Samiti at Vasundhara in Kathmandu; to the Kshetriya Karyalaya of the Thakali Seva Samiti at Nadipur in Pokhara; to Dashrath Rai and the Kendriya Karyalaya of the Kirat Rai Yayokkha at Kathmandu; to Lile Thangmi, Kavi Raj Thangmi of Lapilang and the Nepal Thami Samaj; to Avinath Rai and Ganesh Rai of the Wambule Rai Samaj Nepal; to our many kind Baram friends of Gorkha district, for whom we are still completing a grammatical description of the language; to our wonderful friend Temba Bhote and the Buddhist half of the Lhomi (Shingsaba, Bhote) community; to Lt. Col. Michael Roe, Capt. Simon Garside and those troops of the British Gurkhas at Pokhara who came forward to volunteer blood; to our old friend Vishva 'Bishow' Bhatt of Gorkha; to Tek Bahadur Kumal and the Kumal Service Organisation at Chevetar in Gorkha district; to the kind director and staff of the Youth Awareness Environmental Forum and the Environmental Library at Badegaon at Godawari. Gratitude is also due to the Nepal Janjati Mahasangh at Anamnagar for their scholarly interest and kind advice. Furthermore, we owe gratitude to Cas F. de Stoppelaar, Honorary Consul-General of Nepal in the Netherlands and to Kari Cuelenare of the Dutch Consulate in Nepal. Last but not least, we thank our old and dear friends Narayan Prasad 'Yangsarumba' Pangyangu Subba and Gram

Bahadur 'Sarumba' Pangyangu Subba of Tamphula in Tehrathum district and our dear friend Surendra Raj Dhakal of Gorkha.

The research in the Kingdom of Bhutan was conducted as part of the long-standing bilateral cooperation between the Royal Government of Bhutan and Leiden University. The field campaign was carried out in accordance with the Memorandum of Understanding concluded between the Dzongkha Development Authority (DDA) and Leiden University. Much gratitude is due to the Chairmen of the Dzongkha Development Authority, Their Excellencies 'Lönpo Sangay Ngedup and 'Lönpo Thinley Gyamtsho, as well as to the Home Minister, His Excellency 'Lönpo Jigmi Yoezer Thinley, for encouraging and facilitating the research in every way. Logistics and preparations at district and village level were coordinated in a timely and thorough fashion by Dr'âsho Sangye Dorji, the Honourable Secretary of Dzongkha Development Authority, and by Tshewang Dorji, the Chief Research Officer.

Much gratitude is due to all the village headmen and district officers who assisted the research team on site. Particular gratitude is due to our friends 'Adap Dôji and Seta of the Lhokpu language community at Loto Kucu, Karma Chen and Dhani Ram Toto at Phüntsho'ling, Tandri and friends at Riti in the Black Mountains, 'Ap Drakpa of Phajong Pam and friends in the Gongduk language community. Furthermore, we thank Cecila Keijzer, Peter Newsum and Yanchen Doma of SNV (Netherlands Development Organisation), Bhutan. During the two expeditions we were assisted in the field by Ivo van Asperen, Janine van Nes, and Dr. Mariëtte Hoffer. Some data, included in this manuscript, and to be published in detail elsewhere (Kraaijenbrink et al. in prep.) contains genetic data provided to us by V.K. Kashyap, B. Su, H. Shi, C.J. Xiao, W.R. Tang, and Y. Xue. Here we would like to acknowledge them for this. Special thanks are due to our old friend Karma Tshering of Gaselô, who really made the work happen. Karma involved himself and made himself indispensable at every level, anticipated and prevented all possible difficulties, and facilitated every aspect of the research programme with unparalleled perspicacity. Finally we would like to thank Mr. H.N.W. van Gent of the Netherlands Ministry of Foreign Affairs in The Hague for his assistance in sending the materials to Nepal and Bhutan.

References

Cavalli-Sforza, Luigi L., Menozzi, Paolo and Piazza, Alberto. 1994. *The History and Geography of Human Genes.* Princeton, New Jersey: Princeton University Press.

Cordaux, Richard, Weiss, Gunter, Saha, Nilmani and Stoneking, Mark. 2004. "The northeast Indian passageway: a barrier or corridor for human migrations?" *Molecular Biology and Evolution* 21 (8): 1525–1533.

van Driem, George L. 2001. *Languages of the Himalayas: An Ethnolinguistic Handbook of the Greater Himalayan Region, containing an Introduction to the Symbiotic Theory of Language* (2 vols.). Leiden: Brill.

Falush, Daniel, Stephens, Matthew and Pritchard, Jonathan K. 2003. "Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies." *Genetics* 164: 1567–1587.

Falush, Daniel, Stephens, Matthew and Pritchard, Jonathan K. 2007. "Inference of population structure using multilocus genotype data: dominant markers and null alleles." *Molecular Ecology Notes* 7: 574–578.

Gaikwad, Sonali and Kashyap, V.K. 2003. "Genetic diversity in four tribal groups of western India: a survey of polymorphism in 15 STR loci and their application in human identification." *Forensic Science International* 134 (2–3): 225–231.

Gayden, Tenzin, Cadenas, Alicia M., Regueiro, Maria, Singh, Nanda B., Zhivotovsky, Lev A., Underhill, Peter A., Cavalli-Sforza, Luigi L., and Herrera, Rene J. 2007. "The Himalayas as a directional barrier to gene flow." *American Journal of Human Genetics* 80: 884–894.

Hima Bindu, G., Trivedi, R. and Kashyap, V.K. 2005. "Genotypic polymorphisms at seventeen autosomal short tandem repeat loci in four tribal populations of Andhra Pradesh, India." *Journal of Forensic Sciences* 50 (4): 978–983.

Hima Bindu, G., Trivedi, R. and Kashyap, V.K. 2007. "Allele frequency distribution based on 17 STR markers in three major Dravidian linguistic populations of Andhra Pradesh, India." *Forensic Science International* 170 (1): 76–85.

Kashyap, V.K., Ashma, Richa, Gaikwad, Sonali, Sarkar, B.N. and Trivedi, R. 2004. "Deciphering diversity in populations of various linguistic and ethnic affiliations of different geographical regions in India: analysis based on 15 microsatellite markers." *Journal of Genetics* 83 (1): 49–63.

Kashyap, V.K., Guha, Saurav, Sitalaximi, T., Hima Bindu, G., Hasnain, Seyed E. and Trivedi, R. 2006. "Genetic structure of Indian populations based on fifteen autosomal microsatellite loci." *BMC Genetics* 7: 28.

Kraaijenbrink, Thirsa, van Driem, George L., Opgenort, Jean-Robert M.L., Tuladhar, Nirmal M. and de Knijff, Peter. 2007. "Allele frequency distribution for 21 autosomal STR loci in Nepal." *Forensic Science International* 168 (2–3): 227–231.

Kraaijenbrink, Thirsa, van Driem, George L., Karma Tshering of Gaselô and de Knijff, Peter. 2007. "Allele frequency distribution for 21 autosomal STR loci in Bhutan." *Forensic Science International* 170 (1): 68–72.

Krithika, S., Trivedi, R., Kashyap, V.K. and Vasulu, T.S. 2006. "Antiquity, geographic contiguity and genetic affinity among Tibeto-Burman populations of India: a microsatellite study." *Annals of Human Biology* 33 (1): 26–42.

Lee, Andrew C., Kamalam, Angamuthu, Adams, Susan M. and Jobling, Mark A. 2004. "Molecular evidence for absence of Y-linkage of the Hairy Ears trait." *European Journal of Human Genetics* 12: 1077–1079.

Metspalu, Mait, Kivisild, Toomas, Metspalu, Ene, Parik, Jüri, Hudjashov, Georgi, Kaldma, Katrin, Serk, Piia, Karmin, Monika, Behar, Doron M., Gilbert, M. Thomas P., Endicott, Phillip, Mastana, Sarabjit, Papiha, Surinder S., Skorecki, Karl, Torroni, Antonio and Villems, Richard. 2004. "Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans." *BMC Genetics* 5: 26.

Neeta, Sarkar and Kashyap, V.K. 2004. "Allelic variation at 15 microsatellite loci in one important Australoid and two Indocaucasoid groups of Chhattisgarh-India." *Journal of Forensic Sciences* 49 (1): 184–188.

Peakall, Rod and Smouse, Peter E. 2006. "GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research." *Molecular Ecology Notes* 6: 288–295.

Pritchard, Jonathan K, Stephens, Matthew and Donnelly, Peter. 2000. "Inference of population structure from multilocus genotype data." *Genetics* 155: 945–959.

Quintana-Murci, Lluís, Krausz, Csilla, Zerjal, Tatiana, Sayar, S. Hamid, Hammer, Michael F., Mehdi, S. Qasim, Ayub, Qasim, Qamar, Raheel, Mohyuddin, Aisha, Radhakrishna, Uppala, Jobling, Mark A., Tyler-Smith, Chris and McElreavey, Ken. 2001. "Y-Chromosome lineages trace diffusion of people and languages in Southwestern Asia." *American Journal of Human Genetics* 68: 537–542.

Rajkumar, Ravathi and Kashyap V.K. 2003. "Evaluation of 15 biparental STR loci in human identification and genetic study of the Kannada-speaking groups of India." *American Journal of Forensic Medicine and Pathology* 24 (2): 187–192.

Xue, Yali, Zerjal, Tatiana, Bao, Weidong, Zhu, Suling, Shu, Qunfang, Xu, Jiujin, Du, Ruofy, Fu, Songbin, Li, Pu, Hurles, Matthew E., Yang, Huanming and Tyler-Smith, Chris. 2006. "Male demography in East Asia: a North-South contrast in human population expansion times." *Genetics* 172: 2431–2439.

Table 1. Descriptive statistics of the population samples from Nepal and Bhutan

Population / Pool	Code	Language family	country
Barâm	BAR	TB	Nepal
Chantyal	CHN	TB	Nepal
Chepang (Praja)	CHP	TB	Nepal
Central Kiranti*	CKI	TB	Nepal
Dhimal	DHI	TB	Nepal
Dura	DUR	TB	Nepal
Eastern Kiranti**	EKI	TB	Nepal
Ghale	GHL	TB	Nepal
Gurung	GUR	TB	Nepal
High Caste Newar	HCN	TB	Nepal
Kham (Magar)	KHM	TB	Nepal
Limbu†	LIM	TB	Nepal
Magar	MGR	TB	Nepal
Newar	NWR	TB	Nepal
Sherpa (Solu-Khumbu)	SHE	TB	Nepal
Thangmi	THG	TB	Nepal
Thakali	THK	TB	Nepal
Tamang	TMG	TB	Nepal
Western Kiranti‡	WKI	TB	Nepal
Bahun (Brahmin)	BHU	IE	Nepal
Chetri (Kshetriya)	CHE	IE	Nepal
Indo European / Tibeto Burman substrate§	IET	IE	Nepal
Indo European¶	IEU	IE	Nepal
Kumal	KUM	IE	Nepal
Majhi (Bote)	MAJ	IE	Nepal
Tharu	THR	IE	Nepal
Brokpa (Bj'op)	BRP	TB	Bhutan
Bumthang	BUM	TB	Bhutan
Chali	CHL	TB	Bhutan
Dakpa (Dwagspo)	DAK	TB	Bhutan
Dzala	DZA	TB	Bhutan
Gongduk	GNG	TB	Bhutan
Brokkat	KAT	TB	Bhutan
Khengpa	KHG	TB	Bhutan
Kurtöp	KUR	TB	Bhutan
Lakha	LAK	TB	Bhutan
Layap	LAY	TB	Bhutan
Lhokpu (Lhop, Doya)	LHP	TB	Bhutan

n#males	n#females	Number on map	Lat (dec)	Long (dec)
32	6	8	28,07	84,67
21	2	2	28,40	83,37
20	7	9	27,58	84,70
42	6	16	27,13	87,05
20	2	19	26,50	87,70
27	8	6	28,28	84,20
12	7	17	27,14	87,43
17	8	7	28,28	84,73
40	6	5	28,30	84,12
24	6	12	27,62	85,43
13	1	1	28,50	83,00
56	7	18	27,19	87,83
40	6	4	28,08	83,83
44	10	11	27,62	85,40
20	5	14	27,73	86,58
16	2	13	27,75	86,00
20	9	3	28,82	83,75
41	9	10	27,88	85,42
51	14	15	27,38	86,60
25	8	21	29,17	81,17
37	10	22	29,17	81,20
33	6	26	27,25	85,75
26	14	20	28,75	80,50
21	5	25	28,05	84,45
21	6	24	27,83	83,67
28	7	23	27,42	83,33
40	10	46	27,40	91,72
50	10	37	27,67	90,55
50	11	42	27,38	91,02
49	10	45	27,47	91,52
51	11	43	27,90	91,15
46	10	41	27,08	90,93
24	5	38	27,73	90,43
52	10	39	27,132	90,68
51	13	40	27,82	90,82
50	10	32	27,68	90,15
25	5	30	28,15	89,40
39	8	29	26,95	89,12

Population / Pool	Code	Language family	country
Mangde ('Nyenkha, Henke)	MNG	TB	Bhutan
Black Mountain Mönpa	MON	TB	Bhutan
'Ngalop (Dzongkha)	NGA	TB	Bhutan
Nup	NUP	TB	Bhutan
Tshangla (Shâchop)	TSH	TB	Bhutan
Bodo	BOD	TB	India
Toto	TOT	TB	India
Darai	DAR	IE	Nepal
Giri	GIR	IE	Nepal
Jirel	JIR	TB	Nepal
Lhomi (Shingsaba)	LHM	TB	Nepal
Rana	RAN	IE	Nepal
Shah	SHH	IE	Nepal
Tibetan	TIB	TB	Nepal

* Pool containing Bantawa, Chintang, Chamling, Dungmali, Kulung, Nachiring, Puma and Sampang population samples.

** Pool containing Athpahariya, Lohorung and Yakkha population samples.

† Pool containing Chathare, Pañcthare, Phedappe, Tamarkhole and general Limbu population samples.

‡ Pool containing Bahing, Dumi, Jero, Khaling, Sunwar, Thulung and Wambule population samples.

§ Pool containing Danuwar and Kachariya Danuwar populations samples.

¶ Pool containing Damai, Sarkí, Sonar and Visvakarma population samples.

n#males	n#females	Number on map	Lat (dec)	Long (dec)
54	10	33	27,42	90,22
40	18	34	27,22	90,22
50	10	31	27,53	89,48
27	10	35	27,58	90,33
50	11	44	27,18	91,32
37	2	36	26,67	90,33
54	16	28	26,67	89,00
0	2	small sample size, not included in this study		
3	1	small sample size, not included in this study		
2	0	small sample size, not included in this study		
10	2	small sample size, not included in this study		
1	0	small sample size, not included in this study		
0	1	small sample size, not included in this study		
1	0	small sample size, not included in this study		