Evolution of the rapidly-mutating human salivary agglutinin gene (*DMBT1*) and population subsistence strategy

Shamik Polley (1), Sandra Louzada (2), Diego Forni (3), Manuela Sironi (3), Theodosius Balaskas (1), David Hains (4), Fengtang Yang (2), Edward J Hollox (1)*

1. Department of Genetics, University of Leicester, Leicester LE2 1TE, UK
2. Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK
3. Scientific Institute IRCCS E. Medea, Bioinformatic Lab, 23852 Bosisio Parini, Italy
4. Division of Pediatric Nephrology, University of Tennessee Health Science Center, Le Bonheur Children's Hospital, 50 N. Dunlap Ave, Rm 652R Children's Foundation Research Tower, Memphis, TN 38103

* Corresponding author

Dr Ed Hollox

Department of Genetics, Adrian Building,

University of Leicester

University Rd,

Leicester LE1 7RH

ejh33@le.ac.uk

+44 (0) 116 252 3407

BIOLOGICAL SCIENCES: Genetics

Keywords: copy number variation, saliva, agriculture, agglutinin, mutation, *DMBT1*, structural variation,
Abstract

The dietary change resulting from the domestication of plant and animal species and development of agriculture at different locations across the world was one of the most significant changes in human evolution. An increase in dietary carbohydrates caused an increase in dental caries following the development of agriculture, mediated by the cariogenic oral bacterium *Streptococcus mutans*. Salivary agglutinin (SAG, encoded by the *DMBT1* gene) is an innate immune receptor glycoprotein that binds a variety of bacteria and viruses, and mediates attachment of *S. mutans* to hydroxyapatite on the surface of the tooth. In this study we show that multiallelic copy number variation (CNV) within *DMBT1* is extensive across all populations and is predicted to result in between 7 to 20 scavenger-receptor cysteine-rich (SRCR) domains within each SAG molecule. Direct observation of de novo mutation in multi-generation families suggests these CNVs have a very high mutation rate for a protein-coding locus, with a mutation rate of up to 5% per gamete. Given that the SRCR domains bind *S. mutans* and hydroxyapatite in the tooth, we investigated the association of sequence diversity at the SAG-binding gene of *S. mutans*, and *DMBT1* CNV. Furthermore, we show that *DMBT1* CNV is also associated with a history of agriculture across global populations, suggesting that environmental change as a result of agriculture has shaped the pattern of CNV at *DMBT1*, and that the *DMBT1*-SAG interaction is a promising model of host-pathogen-culture co-evolution in humans.

Significance statement

Humans have undergone an evolutionary very recent change in environment of their own making. The development of agriculture profoundly altered diet and exposure to pathogens, and yet the evolutionary response to this is still poorly understood. Here, we characterise extensive copy number variation (CNV) of the gene encoding salivary agglutinin (*DMBT1*). Salivary agglutinin comprises 10% of salivary protein and binds bacteria, including mediating the attachment of the causative agent of dental caries, *Streptococcus mutans*, to teeth. We show that *DMBT1* is a very fast-mutating protein-coding locus, and *DMBT1* CNV correlates with a population history of agriculture. Furthermore, we examine the relationship between variation of the *Streptococcus mutans* region that binds salivary agglutinin and CNV of the *DMBT1* gene.
Introduction

The effect of the agricultural transition on human genome variation has been extensive (1). In addition to the indirect effect of an exponential increase in population size, direct effects on particular genes have occurred, most notably the evolution, at multiple locations through multiple alleles, of lactase persistence at the LCT gene, enabling adults to drink milk generated from domesticated mammals (2). The agricultural transition is also thought to have had an impact on the oral commensal microbiota, in particular Streptococcus mutans, the causative agent of dental caries which is the most common chronic infectious disease in humans. Analysis of ancient skeletal remains (3) and modern genomic diversity (4) has suggested that S. mutans became a major oral pathogen only after the development of agriculture and the concomitant increase in availability of sugars consumed directly or derived from starchy foods. The increased level of caries in individuals from agricultural societies is observed in both modern and prehistoric populations (5-7). This was likely to have profound consequences to the health of the individuals concerned prior to the development of modern dental treatment (8). Caries left untreated leads to tooth loss, potential severe infections, and a decrease in masticatory efficiency potentially leading to a reduction in access of enzymes to the food bolus (9). It is unclear whether human genetic variation has responded to this change via natural selection.

We analysed the variation of the DMBT1 gene encoding a major salivary glycoprotein salivary agglutinin, also known as gp-340, hensin or muclin, and hereafter referred to DMBT1$^{SAG}$ (10). This protein comprises ~10% of total salivary protein in children and 5% in adults (11), and is also present at other mucosal surfaces (12). DMBT1$^{SAG}$ is a component of innate immunity, acting as a pattern recognition receptor interacting with bacteria such as S. mutans and Helicobacter pylori and viruses such as HIV-1 and influenza (12). Variation between host saliva also affects the adhesion of S. mutans (13), and protein variants of DMBT1$^{SAG}$ have been suggested to affect caries susceptibility in children (14).

Copy number variation (CNV) describes a difference in DNA dosage between different individuals, and includes simple deletion and duplications as well as more complex multicopy and multiallelic variation (15). CNV can affect gene expression by altering the total number of copies of individual genes and therefore gene dosage, by changing tissue-specific enhancers or by varying the number of exons within a gene, potentially altering the number of protein-coding subunits, for example (16). CNV can also show a germline mutation rate at least an order of magnitude higher than single
nucleotide substitutions, because of the distinct mutational processes that underly copy number
change (16). Genomewide, CNVs are enriched for genes that encode proteins that interact with the
environment, particularly those in host defence (17), and a high mutation rate of these loci may
contribute to immunological individuality of the host. Whether selection or a relaxation of functional
constraint is responsible for this bias in genomewide distribution remains unresolved, although
there are strong arguments for the role of gene duplication in evolution (18). There is convincing
evidence that CNV in humans can affect the host’s susceptibility to infectious diseases, including the
well-established effect of α-globin deletion on malaria susceptibility (19). Furthermore, it has been
suggested that the frequency of high copy number alleles of the salivary amylase gene AMY1 has
increased by natural selection in populations that eat a carbohydrate-rich diet (20).

DMBT15AG mostly consists of an array of scavenger receptor cysteine-rich (SRCR) domains which bind
bacteria, including S. mutans (21) and promote their adherence to hydroxyapatite of the tooth (22,
23), which is critical for the cariogenic activity of the bacteria. The canonical DMBT1 gene annotated
in the hg19 human genome assembly has 13 repeats each containing a SRCR domain (Figure 1a). The
repeats containing the SRCR domain, hereafter known as SRCR repeats, within the DMBT1 gene are
distinct at the DNA level but share ~80% identity at the protein level. Within the SRCR domain,
smaller regions that bind to S. mutans and hydroxyapatite have been identified (Figure 1d), although
bacterial binding is inhibited by sialidases, showing glycosyl groups are also important in bacterial
binding (24). Two polymorphic deletions within the DMBT1 gene, involving variable numbers of SRCR
repeats, has been partially described previously but the nature and extent of CNV within this gene
remained incompletely characterised. For example, a polymorphic deletion involving SRCR3- SRCR6
was associated with Crohn’s disease. Furthermore, a polymorphic deletion involving at least repeats
SRCR9-SRCR11 has been described (25, 26). Genomewide arrayCGH (aCGH) analysis has identified
two CNVs consistent with the known polymorphic deletions, but showing extensive complex loss and
gain of copy number (17) (Figure 1b).

We aimed to fully characterise the CNV involving DMBT1, and investigate its mutation rate, question
whether it has adapted to different environments across human populations and dissect the
variation involving its interaction with S. mutans.
Results

Characterisation of copy number variation at DMBT1

Genomewide analysis using high-resolution tiling array CGH (aCGH) identified two CNVs whose location was consistent with the known polymorphic deletions described in the literature, but showing a much more extensive and complex loss and gain of copy number (17) (Figure 1b). We used these two CNV regions as a starting point for our analysis, by interrogating the two CNVs (CNV1 involving SRCR3-SRCR6 and CNV2 involving SRCR9-SRCR11, Figure 1b) separately. To this end, we designed several paralog ratio tests (PRTs), a form of quantitative PCR that is particularly robust in accurately calling CNVs (27). These PRTs were used to estimate copy number at each CNV in 270 individuals from HapMap phase 1. To verify our novel PRTs, we used concordance between PRT assays, clustering of PRT copy number estimates into distinct groups reflecting integer copy number and comparison with aCGH probe intensities (Figure 1c, Figure S1). By typing samples in duplicate, we estimated an upper 95% confidence limit for the error rate in determining CNV1 and CNV2 copy number to be 0.37% (366 samples, no discrepancies) and 0.33% (407 samples, no discrepancies) respectively. In addition, we further validated a subset of samples using long-PCR and fiber-FISH (Figure S2, S3). We show that CNV1 is a multiallelic CNV with a copy number varying between 0 and 5 per diploid genome, and the copy number variable unit includes 4 SRCR domains. For CNV1, zero, one and more than one copies reflect homozygous deletion, heterozygous deletion and normal genotype of the deletion described previously (26). Sanger sequencing of homozygous deleted samples suggests that non-allelic homologous recombination (NAHR) between the 98% identical SRCR repeats carrying SRCR2 and SRCR6 is responsible for CNV1 (Figure 1d, Supplementary methods). It is also clear that CNV2 is considerably more complex than the small deletion described previously, being a multiallelic CNV ranging between 1 and 11 copies per diploid genome with each repeat unit carrying a single SRCR domain.

Analysis of further samples from the CEPH-Human Genome Diversity Project (HGDP) panel of 971 individuals from 52 populations worldwide (28) showed rare individuals with a CNV2 copy number of zero. Sanger sequencing of PCR products from these individuals showed that all the 0 copy CNV2 alleles had a breakpoint within 33bp of sequence identical between SRCR8 and SRCR11 (Figure S4), just upstream of the exon encoding the SRCR domain, suggesting that this allele was generated by NAHR between these repeats (Supplementary methods). This finding suggests that other larger CNV2 alleles have also been generated by NAHR between any of the repeats carrying SRCR domains 8-11.
DMBT1 copy number variation has a high mutation rate

The extensive allelic diversity and repetitive genomic structure of DMBT1, together with the knowledge that NAHR is likely to have mediated generation of new alleles, led us to consider whether CNV1 and CNV2 had a high mutation rate. To study this directly, we used our validated PRT assays to call copy number of DMBT1 CNV1 and CNV2 on 522 samples from 40 large multigenerational families from the Centre d’Etude de Polymorphisme Humain (CEPH) collection. We robustly identified de novo copy number mutations at both loci (Figure S5). The mutation rate at CNV1 is estimated to be 1.4% per gamete (9 out of 632 meioses, 95%CI 0.7-2.7%) and the mutation rate at CNV2 is 3.3% per gamete (21 out of 632 meioses, 95%CI 2.1-5.0%). These mutation rate estimates place both loci amongst the most highly mutating loci known, with comparable rates seen only for non-coding minisatellites (29) or for the mitochondrial D-loop (30). Error rates for CNV1 and CNV2 of 0.37% and 0.33% respectively are below the lower 95%CI bound of both mutation rates, showing that these high mutation rates are not due to errors in copy number calling. Furthermore, examination of the copy number calls of the individuals showing de novo mutations indicates a high posterior probability of that copy number call (Figure S5). All mutations were of a loss or gain of one CNV repeat unit, with no evidence of a bias towards loss or gain. Analysis of our data showed that two CNV1 mutational events and one CNV2 mutational event were associated with a crossover involving flanking marker exchange at the correct position in that individual. This suggests that although NAHR events involving homologous chromosomes do contribute to CNV mutation rate, most events are likely to be inter- or intra-chromatidal.

Global diversity of DMBT1 and agriculture

To examine global diversity of DMBT1, we determined diploid copy number of CNV1 and CNV2 on the CEPH-HGDP panel. We observed a similar range for CNV1 (0-5 copies per diploid genome) and for CNV2 (0-11 copies per diploid genome) as in the HapMap samples (Figure S5). Although there was no linkage disequilibrium between copy number alleles at CNV1 and CNV2 ($r^2=0$ for CEU parents, $r^2=0.01$ for YRI parents, $r^2=0.01$ for all HGDP, Figure 2a), there was a clear negative relationship between average copy number at CNV1 and CNV2 at the population level ($r^2=0.11$, Figure 2b) and at the continental level ($r^2=0.43$). Because, across populations, the increase in CNV2 is mirrored in part by a decrease in CNV1, the total predicted number of the SRCR domains for the two copies of DMBT1 on homologous chromosomes does not mirror this trend. Nevertheless the total number of SRCR domains per diploid DMBT1 is highly variable, and the number of SRCR domains in a given DMBT1SAG molecule is predicted to range between 7 and 20, at least (Figure S5g).
Given the fact that the SRCR domain is known to bind \textit{S. mutans}, we considered that the development of agriculture and consequent increase carbohydrate-rich foods, oral \textit{S. mutans} and dental caries might be a selective pressure influencing the frequency distributions of CNV1 and CNV2. To test this, we correlated CNV1 and CNV2 mean copy number for each HGDP population with an index of extent of agricultural practice for each population, as published previously (31). We used two statistical approaches. First, a regression analysis corrected for population effects by using distance from East Africa as a covariate. Seondly, a partial Mantel analysis corrected using a population pairwise geographical distance matrix, a population pairwise genetic distance matrix, or distance from East Africa as covariates (32). We found a negative relationship between agricultural populations and the mean CNV1 copy number, and a positive relationship between agricultural populations and CNV2 (Table 1). By resampling from an empirical distribution of partial Mantel r correlations with agriculture, we estimated the genomewide significance of this observation to be \( p = 0.0467 \), using the geographic distance matrix as a covariate, and \( p=0.0410 \) using the genetic distance matrix as a covariate. We also tested the association between mean copy number of a population and a subsistence strategy based on carbohydrate-rich foods, such as cereals, roots and tubers, as defined previously (Figure 2c, (33)). We found an association for CNV1 and a weak association for CNV2, both in the expected directions (Table 1). This finding suggests that the subsistence history of a population has affected the frequency distribution of both CNVs within \textit{DMBT1}.

To investigate this association further, we called \textit{DMBT1} CNV1 and CNV2 copy number genotype using sequence read depth analysis on three published ancient DNA samples ((34-36), Table S1, Figures S5e, f). Both Denisova and Neanderthal hominins show a high CNV1 copy number of 3 and a low CNV2 copy number of 3, within the range of hunter-gatherer human populations. Analysis of an 8000-year old hunter-gatherer from Loschbour in Luxembourg provides a more recent directly ancestral calibration point. He had a CNV1 genotype of 1, which is common in modern Europeans, and a CNV2 genotype of 4, which is less common but still present in modern Europeans. This tentatively suggests that a reduction in CNV1 copy number had occurred, or was occurring, but an increase in CNV2 was yet to occur, but further samples are required before any firm conclusions can be drawn, as the observed genotype is consistent with a copy number distribution that is unchanged from modern Europeans.

To provide further support for natural selection in shaping frequencies of CNV1 and CNV2 copy number, we used forward simulations to model CNV1 and CNV2 in situations of population
expansion but no selection. Using a mutation rate derived from our pedigree analysis as well as initial allele frequency distributions based on our largest hunter-gatherer population sample (Biaka) we simulated 1000 populations at CNV1 and CNV2 using a stepwise mutation model and a realistic demographic model (Supplementary methods). The resulting distribution of mean copy number for both CNVs provides an empirical test of the departure of our observed populations from this neutral stepwise mutation model. We interpreted a departure from the model as evidence of natural selection occurring on the locus on ancestors of individuals in that population. Using an estimate of mutation rate of the lower 95% confidence interval for both CNV1 and CNV2, 99.5% of simulated populations had a CNV1 mean copy number above 2.4 and 99.5% of populations had a CNV2 mean copy number above 7.4 (Figure 2b). This shows that CNV2 copy numbers are lower than expected given a neutral stepwise mutation model for all populations, probably reflecting selective constraints on DMBT1^SAG protein length. For CNV1, however, six populations show mean copy number consistent with a neutral stepwise mutation model. Four of those populations are from Africa, one from East Asia and one from South America, yet all six have been classified as non-agricultural (33).

The increase in CNV2 copy number and decrease in CNV1 copy number might be due to selection for a particular phenotype more favoured following the transition to agriculture. Our favoured hypothesis is that a particular SRCR domain that binds S. mutans or hydroxyapatite of the tooth more weakly, thereby reducing the likelihood of caries. Analysis of the S. mutans binding domain sequence shows that there is no difference between SRCR domains in CNV1 and CNV2 (Figure 1d). However, when manually inspecting the human GRCh37/UCSC hg19 reference sequence but also 10 HGDP samples sequenced to high-depth ((34), Table S1), the SRCR domains in CNV2 share a serine to tyrosine change which disrupts the hydroxyapatite-binding domain (37) and abolishes a strong potential mucin-type O-linked glycosylation site (Figure 1d). Replacement of CNV1-type SRCR domains with CNV2-type SRCR domains has therefore allowed this tyrosine substitution to propagate rapidly through the DMBT1^SAG molecule. This suggests that the transition to agriculture has been accompanied by a partial replacement of canonical SRCR domains with SRCR domains that either bind the tooth less strongly or, because glycosylation is important for binding, bind S. mutans less strongly.

Evolutionary relationship between DMBT1 and its ligand in Streptococcus mutans

If the interaction between DMBT1^SAG and S. mutans is co-evolving, we might expect to see a relationship between variation of S. mutans and CNV1 and CNV2 of DMBT1 across different individuals. Adaptation of S. mutans to the DMBT1^SAG phenotype of different mouths reflects a more recent evolutionary timescale than adaptation of DMBT1 in humans. However, given that S. mutans
colonises the mouth in early childhood (38) and the doubling time of biofilm-attached S. mutans is in the order of hours (39), it is likely that it can adapt genetically to the oral environment. We genotyped 125 adult individuals resident in Leicester UK, 92 of European origin, for CNV1 and CNV2, and sequenced part of the S. mutans gene spaP from DNA isolated from matched saliva. spaP encodes Agl/II which is the ligand for human DMBT1$_{SAG}$ (40). We focused on a 1kb region of the spaP gene encoding 336 amino acids from the C-terminal region known to contain two binding domains for human DMBT1$_{SAG}$, namely Ad1 and Ad2 (41). For 98 of our cohort (78%), only one S. mutans strain (as defined by homozygosity of the sequenced region) was found. This observation suggests very low within-mouth diversity of S. mutans, and that most people are colonised by only one strain. However, alignment of sequences showed 136 single nucleotide polymorphisms, 82 of which altered amino acid sequence, reflecting very high levels of diversity between individuals (Figure S6).

A difference between the allele frequency spectra (AFS) of non-synonymous and synonymous polymorphisms can indicate selection, if synonymous polymorphisms are assumed to be neutral. Comparison of the AFS of the S. mutans spaP gene shows a difference in both the total and European-only cohort (p=0.015 and p=0.018 respectively), with an enrichment of polymorphisms with rare alleles (MAF<1%, p=0.003 for total cohort, p=0.022 for European-only cohort, Figure S6e,f). This result indicates that weak negative selection is acting on spaP, consistent with previous genomewide approaches (4).

If a particular spaP allele was adapted to a particular DMBT1$_{SAG}$ phenotype, we might expect the spaP allele and the DMBT1 genotype to be associated across a number of individuals. Across our cohort, we identified 13 polymorphisms that changed amino acid sequence in spaP Ad1 or Ad2. Given the low derived allele frequencies of these polymorphisms, we had limited power to detect an association with CNV1 or CNV2 copy number. However, two of these polymorphisms affect an amino acid highly conserved across oral Streptococci (Figure S6g,h), and in one the derived allele (A1090D, PDB P11657) was associated with lower copy number at CNV1 (nominal p=0.0416) and CNV2 (nominal p=0.0169) in the European-origin cohort, which leads to an overall association with low DMBT1 copy number (p=0.046, corrected for multiple comparisons). Only CNV2 remains associated with low copy number (nominal p=0.0196) in the full cohort, suggesting an effect of ethnicity on this interaction.

**Discussion**

We have shown that DMBT1 shows extensive CNV, with two distinct regions (termed CNV1 and CNV2) showing extensive copy number polymorphism across a wide range of populations. This
polymorphism is predicted to underly the variable number of tandemly-repeated SRCR domains of the DMBT1\textsuperscript{SAG} protein observed in different individuals. These SRCR domains have both \textit{S. mutans}- and hydroxyapatite-binding activities (13, 22, 26, 42). Direct observation of de novo mutations in pedigrees show that both CNV1 and CNV2 have exceptionally high mutation rates, 1.4\% and 3.3\% per gamete per generation respectively; to our knowledge the fastest mutation rates affecting coding sequence yet described in humans. Analysis of breakpoints suggests that, at least for CNV2, NAHR drives the mutation process and that most, but not all, NAHR events are inter- or intrachromatidal, rather than between homologous chromosomes. Such a bias has also been observed at the tandemly-repeated DEFA1A3 locus (43) and at alpha-satellite DNA (44), suggesting a shared mechanism.

Populations which practice agriculture generally show a low copy number of CNV1 and high copy number of CNV2, distinct from hunter-gatherer populations and ancient hominins. This pattern of CNV increases the number of copies of a particular type of SRCR domain containing an amino acid change predicted to disrupt binding to hydroxyapatite in the tooth and to abolish a mucin-like O-glycosylation site. Given that cariogenic \textit{S. mutans} became prevalent after the development of agriculture, our data suggest that DMBT1\textsuperscript{SAG} has evolved to modulate its binding to the tooth surface or \textit{S. mutans} (or both) by rapidly-mutating SRCR domain units carrying the appropriate binding motifs. This presupposes that caries was an agent of natural selection prior to the development of modern dentistry (8, 45). We think that this is not unlikely, given the known acute consequences of caries, such as the increased risk of abscess (46) and chronic consequences increase in difficulty eating, particularly in children, and reduced weight/height gain (47, 48). Nevertheless, hypotheses about the agents of evolutionary change in humans are very difficult to prove, and we note that, given DMBT1\textsuperscript{SAG} protein is expressed on other mucosal epithelia and interacts with other microbes, other evolutionary scenarios are possible, such as adaptation to an altered microbiome of the gut.

Indeed, CNV1 of \textit{DMBT1} corresponds to a previously described deletion (26), with 0 copies reflecting a homozygous deletion and 1 copy reflecting a heterozygous deletion. This deletion has previously been associated, in a small case-control study, with increased susceptibility to Crohn’s disease (25), an intestinal inflammatory disease. If this association is confirmed, this would represent an interesting case of pathogen/culture-driven selection increasing the allele frequency of an autoimmune susceptibility allele.

We also investigated variation of \textit{S. mutans} in the context of DMBT1\textsuperscript{SAG} CNV. The overall pattern of variation of the DMBT1\textsuperscript{SAG}-binding region of AgI/II is that of weak negative selection across the population, where new amino acids changes are selected against when transferred from host to
host. Our data also supports the lack of geographical structure of S. mutans, as our sampling of a
restricted population effectively captures the global diversity of sequences analysed elsewhere, at
least for this particular region, and again argues for weak negative selection and background
selection being the dominant force shaping diversity in S. mutans. There is weak evidence that there
is a relationship between sequence variation at AgI/II and CNV at DMBT1. As minor alleles at AgI/II
are generally rare, this needs a much larger cohort and functional analysis to tease apart the natural
variation modulating this interaction.

Our results now provide a framework for understanding the full nature and functional effect of
sequence variation at this locus, which will have an unclear linkage disequilibrium relationship with
neighbouring SNPs. One study has highlighted DMBT1 to be a strong candidate for ancient balancing
selection in the genome (49). Another recent study has identified the region upstream of DMBT1 to
show unusually negative Tajima’s D (in the 5th percentile genomewide) in Europeans, supporting our
model of selection (50). However, the effect on SNP diversity of a rapidly mutating CNV undergoing
fluctuating geographically structured selection remains unclear. The relationship between the CNV
we describe here and diseases should also be studied further, in particular those diseases with an
infectious or immune component to their etiology. Taken further we hope that the rapidly-mutating
DMBT1 gene will become a paradigm of host-pathogen evolutionary study leading to important
insights in understanding the process of caries formation, and other host-microbe interactions, in
humans.
Methods

Full details on the methods used, and details of the samples, are described in the supplementary materials. CNV was characterised and typed using multiple paralog ratio tests (PRT), a form of quantitative PCR that uses the same primer sequences for test and reference loci to minimise amplification bias. PRTs were validated both by examining assay concordance and by testing a subset of samples using long PCR, array CGH, and fiber-FISH.

*S. mutans* sequences were derived by Sanger sequencing following PCR amplification of a region of the *spaP* gene directly from DNA insulated from human mouthwash samples. Population simulations were conducted using simuPOP. Correlation of population subsistence with mean population copy number was performed as previously described (31, 33).

Acknowledgements

This work was funded by a Government of India Ministry of Social Justice and Empowerment PhD studentship to SP and EJH. EJH was supported in part by an MRC New Investigator Grant (GO801123). SL and FY were supported by the Wellcome Trust (WT098051). DSH was supported by a NIH grant RC4DK090937-01. We would like to thank Gurdeep Lall, Jenny Bowdrey, and Sejjal Patel for help and support, and Rita Neumann, Alec Jeffreys, Mark Jobling and Jan Mollenhauer for DNA samples. This research used the ALICE and SPECTRE High Performance Computing Facilities at the University of Leicester.

References


---

**Figure Legends**

**Figure 1 Analysis of DMBT1 structure and CNV**

a) Dotplot of DMBT1 gene (exons and introns) against itself. Lines indicate high similarity and emphasise the repeated nature of the structure of the gene. Individual SRCR domains are indicated and numbered. Note that the canonical DMBT1 gene sequence has one fewer SRCR domain that that predicted by the genome assembly, and the extra SRCR domain is labelled 9'.

b) Exon-intron structure of the DMBT1 gene with CNV signals. Three DMBT1 gene annotations derived from different transcripts are shown. CNV signals from the Database of Genome Variants (http://dgv.tcag.ca/dgv/app/home) are shown below, with red indicating loss of signal compared to a reference genome, blue gain of signal and brown both loss and gain of signal observed in different samples. Note that these annotations are often larger than the actual CNV, since they can represent large insert clones that detect a CNV, but with the CNV boundaries unknown. CNV1 and CNV2 are annotated, with the reference genomic sequence showing one copy of the CNV1 region and four copies of the CNV2 region. Figure is based on UCSC Genome Browser screenshot hg19 assembly.

c) Comparison of copy number calling methods for CNV1 (left) and CNV2 (right). Each point on the scatterplot represents an individual sample, with different symbols reflecting the final copy number call. The x-axes individuate the copy number value estimated from paralog ratio tests (PRTs), and the y-axes indicate the first principal component of probe intensity data for probes spanning the CNV in array comparative genomic hybridisation (data from (17)).

d) Sequence relationship of SRCR repeats. A maximum-likelihood tree shows the relationship of the SRCR repeats (between 3 and 4kb) carrying the SRCR-coding-domain exon. Scale bar indicates 0.1 substitutions per site. Amino acid sequences of the SRCR domains corresponding to the *S. mutans*
and hydroxyapatite-binding regions are arranged alongside the tree, ordered according to the order of SRCR domains on the tree. Note that the divergent SRCR14 domain does not bind bacteria (54), is not coded by repeated DNA region (see this figure, part a), and is located at the C-terminal end of the salivary agglutinin protein.

**Figure 2 Distribution of DMBT1 copy number values in the CEPH-HGDP panel**

a) Across individuals. Each point represents the mean unrounded PRT copy number of an individual, with the histogram on each axis representing the distribution of CNV1 copy numbers (x axis) and CNV2 copy numbers (y axis).

b) Across populations. The means of CNV1 and CNV2 in each population are plotted, coloured according to continent of origin. The red dashed lines represent the value above which 99.5% of mean copy numbers of simulated populations fall.

c) Average CNV1 and CNV2 copy number in agricultural and non-agricultural populations. Populations are colored according to region (legend in part b), thick line indicates median value and thin lines are 25th and 75th centiles, and p values from logistic regression, with distance from Africa as a covariate (table 1). Agricultural population definition is according to (33).
<table>
<thead>
<tr>
<th>Locus</th>
<th>Cereals, roots, tubers populations as ref. 33</th>
<th>Percentage of time spent on agriculture as ref.31</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds ratio (95% CI) *</td>
<td>Beta value from regression (95% CI) *</td>
</tr>
<tr>
<td>CNV1</td>
<td>0.14 (0.02-0.84) p=0.041</td>
<td>-0.010 (-0.017 to -0.003) p=0.009</td>
</tr>
<tr>
<td>CNV2</td>
<td>2.00 (1.01-4.33) p=0.057</td>
<td>0.027 (0.010 to 0.045) p=0.003</td>
</tr>
</tbody>
</table>

*Distance from Africa as covariate

** pairwise geographical distance matrix as covariate

*** pairwise genetic distance matrix as covariate
Methods

DNA samples and extraction

HapMap DNA samples for the three Phase I populations were purchased from Coriell Cell Repositories and diluted to 10ng/µl in water. Four different populations were analysed, Yoruba from Ibadan in Nigeria (YRI), Japanese from Tokyo in Japan (JPT), the Han Chinese from Beijing in China (CHB) and Americans of European ancestry from Salt Lake City, Utah (CEU).

Samples from the Centre d’etude de polymorphisme human Human Genome Diversity Project (CEPH-HGDP) samples were purchased from CEPH and diluted to 5ng/µl in water. The sample set comprises has been described previously (1), but comprises DNA from 1050 individuals from 52 populations, of which a subset of 971 individuals were analysed in this study (2).

Healthy volunteers were also recruited from local staff and students with informed consent (Leicester cohort), resident in Leicester or the surrounding county, following ethical approval from the University of Leicester ethics committee. Ethnic origin was given by self-identification, with 110 of native European origin and 42 of non-European origin (27 South Asian, 6 Middle Eastern, 4 East Asian, and 5 sub-Saharan African). Buccal cells were collected using standard mouthwash (Sainsbury’s economy), pelleted by centrifugation and genomic DNA extracted following an in-house phenol-chloroform-based method. DNA was quantified by spectrophotometry and diluted to 10ng/µl in water.

Copy number determination and validation at CNV1

Long PCR.

Initial analysis used the primers described previously to type the deletion variant at CNV1, L1595 5’CTGCTGAGCATTGCCTGTGTTCTA-3’ and L1591 5’GTCATATCAGCTCTGAATAGAAAAGTGC-3’ (3, 4), used to co-amplify both long allele and deleted allele of DMBT1. Long PCRs were performed in a total volume of 25µl reactions on a Veriti thermal cycler (Life Technologies) with 45mM Tris-HCl(pH8.8), 11mM (NH₄)₂SO₄, 4.5mM MgCl₂, 6.7mM 2-mercaptoethanol, 4.4mM EDTA, 1mM of each dNTP (sodium salt), 113µg/ml bovine serum albumin, 5µM of each primer (forward and reverse primer), 0.6U of Taq DNA Polymerase (Kapa Biosaystems) and 0.03U Pfu DNA polymerase (Stratagene) and 20-50ng of good quality unsheared DNA. Long PCR reactions were performed with thermo cycler conditions of an initial denaturation of 94°C for 1 min, a first stage consisting of 20 cycles each of 94°C for 15s and 68°C for 10 min, and a second stage consisting of 12 cycles each of
94°C for 15s and 68°C for 10 min (plus 15s/cycle); these were followed by a single chase phase of 72°C for 10 min. We used two control genomic DNAs U87-MG and U343-MG (gift from Dr. Jan Mollenhauer, University of Southern Denmark, Denmark) to standardize our long PCR. U87-MG is the representative for the intact configuration (largest allelic variant) and U343-MG has the internal deletion (shortest allelic variant). The long allele was predicted to generate a 16.9kb amplicon, and the deleted allele a 4.2kb amplicon. Sanger sequencing of the 4.2kb amplicon confirmed that the deletion is generated by NAHR between repeats carrying SRCR3 and SRCR7.

**Allele specific PCR.**

Amplification of the long allele using the primers above was critically dependent on DNA quality, and for some highly-sheared samples an intact 16.9kb amplicon could not be generated. To limit false negatives – i.e. the presence of a long allele not detected by long PCR, we designed a PCR that would amplify a short fragment only in the presence of the long allele. The PCR reactions were done with 10ng of DNA with 1X low dNTPs buffer (50mM Tris-HCl (pH8.8), 12.5mM (NH₄)₂SO₄, 1.4mM MgCl₂, 7.5mM 2-mercaptoethanol, 200µM of each dNTP (sodium salt), 125µg/ml bovine serum albumin, and were supplemented with 5 µM of each primer (forward 5’GGTTGACACAAAACCAACCC-3’ and reverse 5’TGAGTGCTTGACTGCAATTC-3’ primer), 0.5U Taq DNA Polymerase (Kapa Biosystems) in a total volume of 10µl. The PCRs were performed with initial denaturation of 94°C for 4 minutes, followed by 35 cycles of 94°C for 30 seconds, 63°C for 30 seconds and 72°C for 30 seconds, followed by a final extension 72°C for 10 minutes.

**Block-specific PCR**

We also designed a PCR that produced different size amplicons from different SRCR-repeats (“blocks”, Figure S1) for further validation. Homozygous deletion of CNV1 (0 copy sample) resulted in successful amplification of a 1240bp and 1119bp amplicon but absence of a 1549bp amplicon, while other genotypes produced all three amplicons. Primers were 5’GGATGATGTGCGCTGCTCAGGACA-3’ and 5’CTGGGGACTCACCTGGCGCT-3’, and PCR conditions were as described above, for long PCR.

**ArrayCGH data**

Quantile-normalised arrayCGH data for HapMap samples (5) were made available by Dr Don Conrad (Washington University). The intensity values for the probes mapping to CNV1 (12 probes, hg18 chr10:124328301-124341214) and CNV2 (18 probes, hg18 chr10:124343497-124359814) were taken and separately analysed using principal component analysis using the R software CNVtools. The value of each sample for PC1 was then taken to represent a value reflecting the copy number for CNV1 or CNV2.
**Paralog Ratio Tests**

All PRT assays were designed using PRTprimer (6), and checked for predicted amplification sites using in-silico PCR.

For analysis of CNV1 copy number, two PRT assays were designed (Figure S1). All 13 SRCR repeats were aligned and small misaligned regions were targeted for PRT1 assay. The copy number variable (test) and non-variable (reference) SRCR repeats were amplified in PRT1 assays using one pair of primers within DMBT1 gene. The test region was amplified PCR product from SRCR3 repeat and reference region was amplified PCR product from SRCR7 repeat. 1.0 µl of 10X KAPA Taq Buffer A (15mM Mg²⁺, KAPA biosystems), 1 µl of 10X PRT supplementary buffer (10x: 0.5M TrisHCl pH8.8, 125mM (NH₄)₂SO₄, 14mM MgCl₂, 75mM 2-mercaptoethanol, 2mM of each dNTP, 1.25mg/ml deacetylated bovine serum albumin), 3 µM of each primer (labeled 5’FAM/HEX CTTGAGCCTTCATAAACC-3’ and reverse 5’ CTAAGGAATGTTCCACACT-3’ primer), 0.5U Taq DNA Polymerase (Kapa biosystems), and 5-10 ng of DNA as template in every 10 µl PCR reaction. The following thermo cycler conditions were used for PRT1 PCR amplification: initial denaturation of 94°C for 4 minutes, followed by 25 cycles of 94°C for 30 seconds, 62°C for 30 seconds and 72°C for 30 seconds, followed by a final extension 70°C for 10 minutes. 1.0 µl of PCR product was mixed with 0.1 µl MapMarker® 1000 size standard (BioVentures, Inc., USA) and 10 µl HiDi formamide (Life Technologies, UK) before being size-separated by capillary electrophoresis on a 3130xl genetic analyser (Life Technologies, UK) according to the manufacturer’s instructions. The PCR peak area data was analysed using GeneMapper® software v.3.7 (Life Technologies UK).

For PRT2, the test PCR product was amplified from 5th SRCR/SID block and reference PCR product was from 1st SRCR/SID block. PCR components were as PRT1, except 3 µM of each primers (labeled 5’FAM/HEX-TCCACTGGGGTCACAGG-3’ forward and reverse 5’-CTACAGGGGAACACAAGAAC-3’ primer) were used in a final volume of 10 µl. The following thermo cycler conditions were used for PRT2 PCR amplification: initial denaturation of 94°C for 4 minutes, followed by 24 cycles of 94°C for 30 seconds, 63°C for 30 seconds and 72°C for 30 seconds, followed by a final extension 70°C for 10 minutes. Electrophoresis was performed as for PRT1.

The area ratio of peaks at 143bp/153 bp (PRT1) and 234bp/229bp (PRT2) were measured for PRT1 and PRT2 assays for CNV1 of DMBT1, and calibrated against eight HapMap DNA samples (NA18956 – 0 copies, NA18555 – 1 copy, NA12752 – 1 copy, NA18517 – 2 copies, NA10855 – 2 copies, NA12044 – 2 copies, NA07056 - 2 copies, NA18507 – 3 copies) that were run in every experiment. Integer CNV1 copy number was called by combining information from all two assays (PRT1 and PRT2) using a Gaussian mixture model, implemented in the statistical language R (package CNVtools version...
1.42.3) (7). Clustering quality values (Q) are shown in table S2, representing the averaged value of signal-to-noise ratios between adjacent copy numbers with weights reflecting copy number frequency (7). The number of components of the Gaussian mixture model was determined from histogram of average normalized PRT ratios for CNV1 assays, and mean PRT ratios were transformed to have a standard deviation of 1 to facilitate clustering.

**Copy number determination and validation at CNV2**

**Long PCR.**

We aligned different repeats carrying the SRCR domains and primers were designed to amplify sequence across CNV2. The forward primer was selected from 5' region of CNV2 region and the reverse primer was from 3' region of CNV2 region. The PCR amplified different sizes product depending on copy number of CNV2 of *DMBT1*. Long PCRs were performed in a total volume of 25μl reactions on a Veriti thermal cycler with with 45mM Tris-HCl(pH8.8), 11mM (NH₄)₂SO₄, 4.5mM MgCl₂, 6.7mM 2-mercaptoethanol, 4.4mM EDTA, 1mM of each dNTP (sodium salt), 113μg/ml deacetylated bovine serum albumin, 5μM of each primer (forward 5'-AACACCAGTTGTGTAACTGAGACCC-3' and reverse primer 5'-GCCTTCCCAGGTACTTTGGCAGTC-3'), 0.6U of Taq DNA Polymerase (Kapa Biosystems) and 0.03U Pfu DNA polymerase (Stratagene) and 25-50 ng of good quality unsheared DNA. Long PCR reactions were performed with an initial denaturation of 94°C for 1 min, a first stage consisting of 20 cycles each of 94°C for 15s and 68°C for 10 min, and a second stage consisting of 12 cycles each of 94°C for 15s and 68°C for 10 min (plus 15s/cycle); these were followed by a single chase phase of 72°C for 10 min. The products were then analysed following standard agarose gel electrophoresis. Mapping the primers on to the reference genome assembly showed that we would expect a 3244bp product for a 0 copy allele (i.e. containing one SRCR domain repeat only), 7303bp product for a 1 copy allele, 11362bp for a 2 copy allele and 15421bp for a 3 copy allele. In practice, products from the 0 and 1 alleles could be reliably generated, from the 2 copy allele generated reliably from unsheared DNA and the 3 copy allele generated only intermittently. Sanger sequencing of the zero copy allele PCR product from 25 individuals from the HGDP panel confirmed that, in all cases, the SRCR domain amplified was a fusion generated by NAHR between SRCR8 and SRCR11, with a breakpoint localised to 32 bp immediately 5' of the exon coding for the SRCR domain.

**Paralog Ratio Tests**

A total of three PRT assays were designed to estimate copy number for CNV2 region of *DMBT1*: PRT3, PRT4 and PRT5. For PRT3, the test PCR product was amplified from the repeats carrying SCCR9, 9’, 10 and 11, and the reference region was PCR product amplified from the repeat carrying
SRCR8 close to reference region of PRT1. The PCR reactions were performed with 5-10 ng of DNA as template in a total volume of 10µl with 1µl of 10X KAPA Taq Buffer A (15mM Mg²⁺, KAPA biosystems), 1 µl of 10X PRT supplementary buffer (10x: 0.5M TrisHCl pH8.8, 125mM (NH₄)₂SO₄, 14mM MgCl₂, 75mM 2-mercaptoethanol, 2mM of each dNTP, 1.25mg/ml deacetylated bovine serum albumin), 3 µM of each primer (labeled 5’FAM/HEX-TGTGGTCACTTAGGACAGGG-3’ and reverse 5’-CCTCACAGTGAGAGGATCCC-3’ primer), 0.5U Taq DNA Polymerase (kapa biosystems), in every PCR reaction. The PCR was performed with following thermo cycler conditions were: initial denaturation of 94°C for 4 minutes, followed by 24 cycles of 94°C for 30 seconds, 62°C for 30 seconds and 72°C for 30 seconds, followed by a final extension 72°C for 10 minutes. Electrophoresis and analysis was performed as for PRT1.

For PRT4, the test PCR region was similar to PRT3 (SRCR9, 9’ 10 and 11) but a total four reference regions from different blocks were co-amplified (repeats carrying SRCR2, 8, 12, and 13) using forward primer 5’FAM/HEX-GGTGTCATCTGCTCAGGT-3’ and reverse primer 5’-TCTTCTGCCCTCTCCGC-3’. PCR conditions and electrophoresis conditions were as PRT3.

For PRT5 the test PCR region was similar to PRT3 and PRT4 (SRCR9 and SRCR10) but a reference region from a different chromosome was chosen (chromosome 9). Primers used were forward 5’-GATCACATCCTTCACAAGGT-3’ and reverse 5’-CCAGTTGTCTCTGCTCTCAAAC-3’. PCR conditions and electrophoresis conditions were as PRT3.

The area ratio of peaks of the test and reference amplicons were measured for PRT3, PRT4, and PRT5 assays for CNV2 of DMBT1, and calibrated against eight HapMap DNA samples (NA18956 – 2 copies, NA18555 – 2 copies, NA12752 – 3 copies, NA18517 – 2 copies, NA10855 – 4 copies, NA12044 – 5 copies, NA07056 - 6 copies, NA18507 – 3 copies). Scatterplots showed that although PRT5 raw copy number estimates agreed well with PRT3 and PRT4, PRT5 data did not show strong evidence of clustering about integer copy numbers and were not used for calculating final integer copy number, which was done using CNVtools as described for CNV1 above.

**Analysis of CNV2 breakpoints**

We selected 24 individuals across 15 populations that had a copy number of one for CNV2 (Table S3). Long PCR followed by gel electrophoresis (described above) confirmed heterozygosity for a 0 and 1 alleles, and allowed excision, purification and Sanger sequencing of a single copy of the 0 allele. Alignment of the sequences with the genomic sequence showed that the 5’ end of the 0 allele is from the SRCR8 repeat and the 3’ end from the SRCR11 repeat, with the breakpoint in all 24 alleles within a 32bp region of 100% identity between the two repeats. The high identity of the breakpoint
and the rapid mutation rate of the locus strongly suggest that the CNV is generated by recurrent non-allelic homologous recombination (NAHR). The sequence motif associated with recombination hotspots (8) was not found within or around this region, but a sequence motif previously associated with deletions is 26 bp 3’ of the CNV breakpoint region ([9] Figure S4).

**Fiber-FISH**

Two pairs of PCR primers were designed to amplify two ~2kb parts of the *DMBT1* SRCR repeat (5’ CTGAGGCTGGTGAATGGA 3’ forward and 5’-TATCCCTYTCCCTGCCCRAGCA-3’ reverse., 5’ TCAGCAATGGYRTCWGATGT forward and CTACAGGGGAACACAAGA-3’ reverse). Degenerate bases were used within the primers to allow amplification from different SRCR repeats.

The probes used consisted of PCR products from the SRCR repetitive regions of the *DMBT1* gene, a fosmid clone (G248P8718G1) and a BAC clone (RP11-144H6), selected from the UCSC Genome Browser (GRCh37/hg19 assembly). Plasmid DNA was purified using a PhasePrep BAC DNA kit (Sigma-Aldrich) following manufacturer’s protocol and amplified using a GenomePlex® Whole Genome Amplification (WGA) kit (Sigma-Aldrich) following manufacturer’s protocol. SRCR products were labelled with digoxigenin-11-dUTP (Roche) and DNP-11-dUTP (Perkin Elmer), G248P8718G1 clone was labelled with digoxigenin-11-dUTP (Roche) and RP11-144H6 clone was labelled with biotin-16dUTP (Roche), by using a modified WGA re-amplification kit (Sigma-Aldrich) as described before (10).

Single DNA molecule fibers were prepared using the molecular combing method (11) according with the manufacturer’s instructions (Genomic Vision). Briefly, the cells were washed in 1×PBS (Invitrogen) and embedded in a 1.2% low melting point agarose (Lonzar) plugs (~1 million cells/plug), followed by overnight incubation (16-18 hours) at 50°C with digestion solution [8:1:1, 0.5M EDTA: 10% sarkosyl: proteinase K (Ambion)]. The next day plugs were washed with 1×TE (10mM Tris, 1mM EDTA, pH8.0), stained with YOYO-1 (Molecular Probes, Life Technologies), transferred to 0.5M MES buffer (pH5.5) solution, melted at 70°C for 25 minutes and incubated overnight (16-18 hours) at 42°C with β-agarase enzyme (BioLabs). The following day coated coverslips (Genomic Vision) were soaked in the DNA solution and pulled out of the solution at a constant vertical speed by using a molecular combing system (Genomic Vision), which allowed the production of single-molecule DNA fibers with a constant stretching factor (2Kb=1μm). The cover slips were then baked for 4 hours at 68°C.

The probe mix was denatured at 65°C for 10 minutes before being applied onto the coverslips and the hybridisation was carried out in a 37°C incubator overnight. Post-hybridisation washes consisted of two rounds of 50% formamide/2×SSC (v/v) washes followed by two additional washes in 2×SSC. All washes were done at 25°C, 5 minutes for each time. Digoxigenin-labeled probes were detected
using a 1:100 dilution of monoclonal mouse anti-DIG antibody (Sigma-Aldrich) and a 1:100 of Texas Red-X-conjugated goat anti-mouse IgG (Invitrogen); DNP-labeled probes were detected using with a 1:100 dilution of Alexa 488-conjugated rabbit anti-DNP IgG and 1:100 Alexa 488-conjugated donkey anti-rabbit IgG (Invitrogen); biotin-labeled probes were detected with one layer of 1:100 of Cy3-avidin (Sigma-Aldrich). After detection, slides were mounted with SlowFade Gold® mounting solution containing 4’, 6-diamidino-2-phenylindole (Invitrogen). Images were visualised on a Zeiss AxioImager D1 fluorescent microscope equipped with narrow band-pass filters for DAPI, FITC, Cy3 and Texas Red fluorescence and an ORCA-EA CCD camera (Hamamatsu). Digital image were captured and processed using the SmartCapture software (Digital Scientific UK).

10-12 DNA fibers were captured for every sample. The lengths of the reference and test regions were measured independently using ImageJ software (12). The well stretched fibers were used for further analysis and broken fibers were excluded. Independent measurements of labelled fiber length from two individuals’ measurements (EJH and SP) were used to avoid measurement errors, using ImageJ. The numbers of red and green spots that represented the first and second part of each SRCR domains were also counted by eye. The length of the labelled DMBT1 regions (test) was calibrated based on size of reference fosmid clone, labelled in red, and 40kb in size.

**Determination of error rate**

For CNV1 and CNV2, 366 and 407 samples respectively were completely typed twice. Samples were taken from the HapMap cohort, Leicester cohort, and HGDP cohort. After calling using CNVtools as described above, no discrepancies were found, leading to a sample-wise error rate of 0-0.37% for CNV1 and 0-0.33% for CNV2, with 95%CI calculated using a Poisson approximation.

**Inferring CNV1 and CNV2 copy number from high-depth genomic sequence**

We selected 13 samples for analysis: 1 Denisovan (13), 1 Neanderthal (Denisova cave, (14)) and 1 modern human (Loschbaur, hunter-gatherer, (15)), together with 10 HGDP samples sequenced by the same laboratory using same technology to a similar depth as the ancient DNA samples (13). Sequence alignment (.bam) files were downloaded either from the European Nucelotide Archive or directly from http://cdna.eva.mpg.de/denisova/BAM/human/. Bam files were indexed locally using samtools version 1.1 (16), and number of mapped reads obtained using the samtools view -c -F 4 command. Three regions were selected: DMBT1 CNV1 region (hg19) chr10:124338311-124351224, DMBT1 CNV2 region chr10:124353507-124369824 and a reference region proximal to DMBT1 which shows no segmental duplications nor any CNV, according to the Database of Genome Variants (dgv.tgac.ca). Following calculation of a ratio of CNV sequence reads:reference sequence reads for both CNV1 and CNV2, the HGDP samples were used as a calibration curve to enable CNV1 and CNV2 copy number to be called from ancient DNA samples.
Genotyping FUT2 for secretor status phenotype

Human SAG has been shown to be glycosylated by the enzyme alpha-(1, 2) fucosyltransferase, encoded by the FUT2 gene and responsible for the secretor status of ABO and Lewis blood group antigens (17-19). Because binding to host SAG is thought to be mediated in part by glycosylated SAG residues, we genotyped the common secretor polymorphism in our Leicester volunteer cohort. This is the FUT2 W143X polymorphism (rs601338), resulting in an inactive FUT2 (20). We used the fact that the secretor allele generates an AvaII restriction digest site to design a simple PCR-RFLP assay using the primers 5′-GAGTACGTCCGCTTACC-3′ and 5′-CTTCCACACTTTTGCGCATGA-3′. The PCR reactions were performed with 5-10 ng of DNA as template in a total volume of 10µl with 5 µM of each primer, 0.5 µl of 2.5mM dNTPs mix (Promega), 0.5U Taq DNA Polymerase (KAPA biosystems), 1µl of 10X KAPA Taq Buffer A (15mM Mg²⁺, KAPA biosystems) in every PCR reaction. The PCR was performed with following thermo cycler conditions were: initial denaturation of 94°C for 4 minutes, followed by 35 cycles of 94°C for 30 seconds, 58°C for 30 seconds and 72°C for 30 seconds, followed by a final extension 72°C for 10 minutes. After PCR, 5µl of 1X CutSmart™ Buffer was added to each sample containing 2 units of AvaII enzyme (New England Biolabs) and incubated at 37°C for 1-2 hours. An invariant AvaII site was included in the PCR product to act as a control for complete digestion. Following agarose gel electrophoresis, ethidium bromide staining and scoring of the genotypes, the secretor status of each sample was deduced by calling rs601338 genotypes GG and GA as secretors and AA genotype as non-secretors.

S. mutans diversity analysis

The part of the spaP gene encoding the SAG-binding C-terminal region of the Agl/II protein was amplified from 10-25ng genomic DNA derived from mouthwash samples in a total volume of 25µl with 5 µl of 5X KAPA HiFi Fidelity Buffers (containing 2 mM MgCl₂ at 1X), 5 µM of each primer (5′ACTGTTCATTTCCATTACTTTAAACTAGC-3′ and 5′ GTTAATCTTAGGAACATTATTGATAACG-3′), 0.5 U KAPA HiFi (HotStart) DNA Polymerase, cycling using an initial denaturation of 95°C for 4 minutes, followed by 35 cycles of 98°C for 20 seconds, 58°C for 30 seconds and 72°C for 30 seconds, followed by a final extension 72°C for 5 minutes. Following confirmation of amplification of a 1008bp product by agarose gel electrophoresis, the band was excised and extracted using a QIAquick Gel Extraction Kit (Qiagen). Two internal primers (5′CGCTTCTTCTGGATAATCATCTACATAG-3′ and 5′ ATCCATGTGGTGGGTGCACAAC-3′) were used to Sanger sequence 20-30ng of the purified PCR product using standard BigDye Terminator 3.1 chemistry. Four samples were also sequenced using the PCR primers to generate sequence from the alternative strand, with fully concordant results. Sequences were trimmed and aligned using ClustalX (21), and variable sites confirmed by visual inspection of chromatograms.
Following alignment of 1008bp across the 149 sequences showed that 7 sequences showed heterogeneity at one position and 22 at more than one position indicating more than one strain had been amplified and sequenced. Those 29 sequences were divided into two sequences so total 178 nucleotide sequences were obtained. Using BLAST, a further 76 homologous sequences from sequenced genomes from oral streptococci strains were identified, including 4 from *Streptococcus intermedius*, 1 from *S. downei*, 2 from *S. gordonii*, 1 from *S. oralis*, 1 from *S. sanguinis*, 3 from *S. sobrinus*, and aligned with our sequences.

Phylogenetic analysis was performed using the software package MEGA5 (22). A maximum-likelihood tree showed a clear clade where 26 sequences amplified from volunteer mouthwash DNA were grouped together with reference sequences from *S. intermedius*. These 26 sequences were assumed to be derived from oral *S. intermedius* and were removed from further analysis.

Allele frequency spectra were compared using the k-sample Anderson-Darling test, implemented in the kSamples package (version 1.0) in the statistical software R. This non-parametric, distribution-free test is insensitive to ties and has more power to differences at the tails of distributions compared to the more-commonly used Kolmogorov Smirnov test. Variants were arbitrarily called as “rare” if they had a minor allele frequency less than 1%, and the binning of polymorphic sites into those with rare and frequent alleles allowed a standard Fisher’s exact test to assess the enrichment of rare non-synonymous polymorphisms.

We tested the relationship between *DMBT1* CNV1 and CNV2 and the non-synonymous substitutions within Ad1 and Ad2 using logistic regression. The allele of each non-synonymous substitution within was the binomial (1 or 0) outcome variable and tested in turn in a logistic regression model with CNV1 copy number, CNV2 copy number and secretor status as predictor variables, using the statistical software R. A combined p-value reflecting significance of association with *DMBT1* copy number was generated using Fisher’s approach, then adjusted for multiple comparisons using Bonferroni’s method.

**Estimation of mutation rate at CNV1 and CNV2**
We determined copy number of CNV1 and CNV2 for 522 samples from 40 families from the CEPH collection using the methods described above. We also genotyped two short tandem repeats (STRs) distal to the CNVs to confirm Mendelian inheritance of chromosomes. These were DMBT-m1 (chr:10:124393773-124364053-hg19), amplified using a fluorescently-labelled forward primer 5’CTCCAGAGGGTGATCTGCTCTG-3’ and reverse primer 5’GTGACAGAGCGAAGTCCATGTC-3’, and DMBT1-m2 (chr:10:124403465-124403739-hg19), amplified using a fluorescently-labelled forward primer 5’GGGCACAAGCTATGTCAC-3’ and reverse primer 5’CATTCATTCCCTGCCTCCATGC-3’.
3’. 5-10 ng of template DNA was amplified in 50mM Tris-HCl pH8.8, 12.5mM (NH₄)₂SO₄, 1.4mM MgCl₂, 7.5mM 2-mercaptoethanol, 200µM of each dNTP, 125µg/ml BSA, 5 µM of each primer (labeled forward and reverse primer), 0.3U Taq DNA Polymerase (Kapa biosystems) in a total volume of 10µl. PCR reactions were performed with thermal cycler conditions of an initial denaturation of 94°C for 4 minutes, followed by 25 cycles of 94°C for 30 seconds, 63°C for 30 seconds and 72°C for 30 seconds, followed by a final extension of 72°C for 10 minutes.

We constructed pedigrees and identified copy number changes in the children that were consistent with de novo mutational events. To confirm that this were likely to be real events we analysed the posterior probability scores of the individuals showing the discrepant copy number calls. For all except two individuals, copy number posterior probability was greater than 0.99, and the remaining two calls, both for CNV2, the posterior probability greater than 0.85.

We took a conservative approach and called as de novo copy number changes only those events that could not be explained by crossover in a non CNV region in one of the parental haplotypes. This is a conservative assumption, as it is possible that some of these events are genuine crossovers in the CNV region causing copy number changes. After removing 4 events for CNV2 only, 9 CNV1 events and 21 CNV2 events in 632 meioses remained. We calculated mutation rate based on this result, and estimated binomial 95% confidence intervals using binom package (version 1.1) in R using the Klopper-Pearson method.

To determine whether the de novo copy number changes were accompanied by flanking marker exchange, we also genotyped two STRs proximal to the CNVs. These were DMBT-m3 (hg19-chr10:124335292-12435438), amplified using a fluorescently-labelled forward primer 5’ CCCATGCTCATCCAGTCAATC -3’ and reverse primer 5’ TCCAGCCTAGACAACAGAGC-3’, and DMBT1-m5 (hg19-chr10:124227879-124228113 ), amplified using a fluorescently-labelled forward primer 5’ GCATTGCTAGAECTCCAGGCA -3’ and reverse primer 5’ ATGGCAAATTCGGAGATGC -3’. Amplification and analysis conditions for these two short tandem repeats were as for the first two distal STRs.

Human population genetic analysis

Analysis of the nature of subsistence.

The nature of subsistence for the 52 HGDP populations was taken from two different sources. The first source was a previously published analysis (23) that used data from Murdock (24) or the Encyclopedia of World Cultures (25) to define each population trait according to the main dietary component. We considered those who had cereals, roots and tubers as their main dietary bcomponent to have a carbohydrate-rich diet, and those relying on fats, meat, milk or other foods to
have a carbohydrate-poor diet. This grouping was used as a predictor variable in a logistic regression, with distance from Africa as a covariate.

The second data source was from Fumigalli et al. (26) who also used data from Murdock but reported the data as a percentage of time spent on agricultural activities. These data were used in a linear regression model as a predictor variables, with distance from Addis Ababa used as a covariate to account for the general loss of diversity observed as part of the out-of-Africa range expansion, using with mean CNV1 or mean CNV2 of the population as the dependent variable. For the Fumigalli dataset, because it is a continuous variable and the difference between a population that spends 10% of time spent agriculture (relative to other subsistence strategies) and a population that spends 90% could be regarded as a “cultural distance matrix”, we also used a partial Mantel test to test for the association of agricultural subsistence and mean CNV1 or mean CNV2, with three different covariate matrices: geographical distance from Addis Ababa, geographical distance between populations, and genetic distance between populations, as measured by the Reynolds’ distance (27) calculated from genomewide SNPs genotyped on the HGDP populations.

We estimated genomewide significance of our observation by constructing an empirical distribution of partial Mantel r correlations with agriculture as above, for the average repeat length for each population of 783 autosomal microsatellites (28). We resampled a random pair of microsatellites 10000 times from this distribution, and recorded the number of pairs that showed an r value equal to or greater than the r values shown by DMBT1 CNV1 and CNV2. All calculations were performed using the statistical language R.

Population simulation

We used a forward simulation approach to simulate the behaviour of both CNV1 and CNV2 in an expanding population assuming no selection. We used simuPOP version 1.1.1, a forward population simulator based in Python (29), and a modified version of the simudCV script (30) to simulate copy number genotypes of individuals in a population with a given mutation rate, population size for a given number of generations, assuming no selection and random mating. The average diploid copy number of the population at the end of the simulation was calculated and used for subsequent analyses. 1000 replicate populations were simulated for each simulation condition chosen.

Population modelling requires assumptions, and we took care to use the most accurate assumptions we could. For the demographic model, we used a demographic model based on estimates from whole genome sequences (31) and genomewide nucleotide mutation rates from analysis of pedigree data (32). We used 80 generations (~2000 years) at a population size of 2x10³ followed by exponential growth over 2,400 generations (~60,000 years) to a final population size of 10⁵. We also
used a population model for the same number of generations but with a fixed population size of 10,000, equivalent to the long-term effective population size of humans, to explore the effect of the exponential growth of the population.

Initial allele frequencies of a pre-agricultural pre-expansion African population were based on the Biaka population, which is a hunter gatherer group nearest to Ethiopia in our analyses. It is important to note that we do not regard the Biaka as representative of an ancestral group in any other respect apart from it being non-agricultural and not having undergone an expansion out of Africa. Copy number allele frequencies were inferred from diploid copy numbers assuming random assortment of alleles and Hardy-Weinberg equilibrium using ConVem (33).

We used a stepwise mutation model in our simulations. Based on our observation in families, where there were no examples of gain and loss of more than one copy, we similarly restricted our stepwise mutation model to loss or gain of one copy.

Supplementary figure legends

**Figure S1 Copy number typing of DMBT1 CNV1 and CNV2**

a) The *DMBT1* gene intron-exon structure is shown in blue, with the CNV1 region highlighted in red. Above the gene, the primer binding sites and expected length products of the CNV1-calling PCR are shown, with annotated examples of agarose gels. Below the gene are the regions amplified by PRT1 and PRT2 primers, corresponding to the reference (green) and test (red) loci.

b) A pairwise scatterplot for raw values from the PRT1 and PRT2 assays on the 270 HapMap samples.

c) The *DMBT1* gene intron-exon structure is shown in blue, with the CNV2 region highlighted in green. Above the gene, the primer binding sites and expected length products of the CNV2-calling PCR are shown, with an annotated example agarose gel on the left. Also shown above the gene are the regions amplified by PRT3 primers, corresponding to the reference (orange) and test (blue) loci. Below the gene the primer positions for the long PCR for calling CNV2 are shown, together with an annotated example gel on the right. Below these primer positions, the primer positions for PRT4 (red, test locus; green, reference locus) and PRT5 (blue, test locus; green, reference locus) are shown.
d) A pairwise scatterplot for raw values from the PRT3 and PRT4 assays on the 270 HapMap samples.
e) and f) pairwise scatterplot for raw values for PRT5 versus PRT3 and PRT4 assays on the 270 HapMap samples. This illustrates the clustering quality of PRT3 and PRT4 and, while PRT5 clearly measures copy number at this locus, there is no strong evidence of clustering.

Figure S2 Validation of DMBT1 CNV1 structure using molecular combing fiber-FISH

a) Schematic representation of fiber-FISH strategy. The 171kb BAC was labelled blue, with a reference fosmid, mapping to a non-CNV region, labelled red, and two PCR products designed to map to the 5'(red) and 3'(green) regions of the repeats carrying the SRCR domains within DMBT1.

b) Example fiber-FISH images on DNA from cell lines derived from a YRI HapMap trio (Y045). This trio was selected because it was invariant at CNV2 and showed variation at CNV1. Two different length alleles can be distinguished in the heterozygous daughter GM19202.

c) Individual measurements of DMBT1 probe fiber length from the father (GM19200, red squares), mother (GM19201, green triangles) and daughter (GM19202, blue circles). Each point represents two independent measurements of a single fiber, with measurements taken by the authors EJH (y-axis) and SP (x-axis). For each individual several fibers were not included for measurement due to evidence of a broken or bent fiber. The mother clearly shows longer DMBT1 alleles than the father, with the daughter having both long and short alleles. Importantly, the difference in length between the 2 copy allele and the 3 copy allele, and therefore the repeat unit size, is consistent with the 12.7kb (4 repeats of around 3kb each, carrying SRCR domains), as predicted from PCR and analysis of the gene structure.

Figure S3 Validation of CNV2 structure using molecular combing fiber-FISH

a) Schematic representation of fiber-FISH strategy. The 171kb BAC was labelled blue, with a reference fosmid, mapping to a non-CNV region, labelled red, and two PCR products designed to map to the 5'(red) and 3'(green) regions of the repeats carrying the SRCR domains within DMBT1.

b) Example fiber-FISH images on DNA from cell lines derived from a CEU HapMap trio (family 1447). This trio was selected because it was invariant at CNV1 and showed variation at CNV2. Two different length alleles can be distinguished by eye in the father GM12762.
c) Individual measurements of DMBT1 probe fiber length from the father (GM12762, red squares), mother (GM12763, green triangles) and daughter (GM12753, blue circles). Each point represents two independent measurements of a single fiber, with measurements taken by the authors EJH (y-axis) and SP (x-axis). For each individual several fibers were not included for measurement due to evidence of a broken or bent fiber. The mother clearly shows longer DMBT1 alleles than the father, with the daughter having both long and short alleles. Importantly, the difference in length between the 2 copy allele and the 3 copy allele, and therefore the repeat unit size, is consistent with 16kb (4 repeats of about 4kb each, carrying SRCR domains), as predicted from PCR and analysis of the gene structure.

**Figure S4 Sequence alignment of CNV2 breakpoint region**

Alignment of an example 0 copy allele with paralogous sequence from SRCR8 and SRCR11 from human reference genome. The 0 copy allele switches from SRCR8 sequence (blue) to SRCR11 sequence (red) within the region indicated by arrows. The sequence motif enriched in deletions (9) is highlighted within a box. Sequence alignment used Clustal Omega (34).

**Figure S5 Copy number typing and distribution of CNV1 and CNV2 in CEPH families and HGDP samples**

Scatterplots for a) CNV1 and b) CNV2 showing mean unrounded copy number values from PRT (x axis) and Bayesian posterior probabilities of copy number calls (y axis). Points represent a single sample, with red crosses representing a copy number call from a child having a de novo copy number mutation. Summary histograms are shown above and to the side of both axes, with numbers indicating the copy number call made by CNVtools for that cluster. It should be noted that the x-axes represent raw PRT values of test peak area:reference peak area before normalisation, so that, for example in CNV1, a value of 0.5 represents a diploid copy number of one for that repeat unit.

c) and d) Histogram of transformed raw copy number values for c) CNV1 and d) CNV2 for 971 HGDP samples. The x axis represents the copy number signal used to fit the Gaussian mixture model for calling integer copy number, and is equal to the average raw PRT value divided by the standard deviation of all samples, so that the entire dataset has a standard deviation of 1 to optimise model fitting. Lines indicate the Gaussian curves fitted to the data, and represent the posterior probabilities of integer copy number calls.
e) and f) Read depth analysis of 10 HGDP samples previously sequenced at high depth for CNV1 and CNV2. Black datapoints are HGDP samples plotted according to observed sequence read depth ratio (x axis) and PRT copy number (y axis). The red line shows the best-fit linear regression line for the HGDP samples. Three ancient DNA samples (Loschbour, Denisovan and Neanderthal), also previous sequenced to high depth, are plotted according to their observed read depth ratio, with the dashed line indicating inferred diploid copy number.

g) Bar chart showing the calculated diploid tandemly-repeated SRCR domain count of DMBT1 in the HGDP samples, sorted by continent of origin. The canonical DMBT1 sequence has 13 tandemly-repeated SRCR domains, so that an individual homozygous for this canonical sequence will have 26 diploid SRCR domains.

Figure S6 Analysis of salivary agglutinin binding region of S. mutans spaP gene

a) Phylogeny of spaP gene in oral Streptococci from European-only cohort. Grey nodes represent S. mutans, red nodes represent S.intermedius and blue nodes represent other oral streptococci. Nodes annotated with accession numbers represent those identified using from sequence database searches. 203 aligned DNA sequences of 1008nt in length were analysed, and the tree was inferred by using the maximum likelihood method based on the Tamura-Nei model, and the tree with the highest log likelihood is shown. The scale bar indicates the number of substitutions per site.

b) Subtree of a, with non-mutans and non-intermedius Streptococci sequences removed.

c) Phylogeny of spaP gene in oral Streptococci from full cohort. Grey nodes represent S.mutans, red nodes represent S.intermedius and blue nodes represent other oral streptococci. Extra sequences not shown in the tree in figure a are annotated as triangles at the node.

d) Subtree of a, with non-mutans and non-intermedius Streptococci sequences removed.

e) Allele frequency spectrum of S.mutans spaP sequences of European-only cohort. The synonymous allele frequency spectrum is in red, and the non-synonymous frequency spectrum in blue.

f) Allele frequency spectrum of S. mutans spaP sequences of full cohort.

g) Alignment of Ad1 and Ad2 DMBT1-SAG-binding regions of spaP gene in S. mutans and its orthologs in other oral streptococci. Sites polymorphic in S. mutans are shown with the alternative allele in orange. Colours correspond to the regions highlighted on the crystal structure as shown in part e.
h) Crystal structure of the C-terminal region of *S. mutans* AgI/Il protein encoded by *spaP*. Amino acids corresponding to the Ad1 and Ad2 DMBT1<sup>Ag</sup>-binding regions are highlighted in green and blue respectively. Known polymorphic amino acids within Ad1 and Ad2 are highlighted in orange.

References


a) [Diagram showing chromosomal regions and copy numbers]

b) Father GM19200
   - CNV1: 2 copies CNV2: 2 copies
   - CNV1: 1 copy allele
   - CNV1: 2 copy allele

Mother GM19201
   - CNV1: 4 copies CNV2: 2 copies
   - CNV1: 2 copy allele

Daughter GM19202
   - CNV1: 3 copies CNV2: 2 copies
   - CNV1: 2 copy allele
   - CNV1: 1 copy allele

c) Predicted allele size difference: 12.7 kb
a) Diagram showing a region of 171 Kb with genomic markers.

b) Results for different individuals:
- Father GM12762: CNV1 2 copies, CNV2 8 copies.
- Mother GM12763: CNV1 2 copies, CNV2 3 copies.
- Daughter GM12753: CNV1 2 copies, CNV2 5 copies.

- CNV2 alleles:
  - Father: alleles 3 and 5
  - Mother: alleles 1 and 2
  - Daughter: alleles 3 and 2

c) Graph showing the predicted size difference of one copy from genome assembly is 4.2 kb.
**CNV2 breakpoints**

**Deletion motif**
<table>
<thead>
<tr>
<th>POP</th>
<th>CVN1</th>
<th>CVN2</th>
<th>POP7</th>
<th>Cereals_roots_tubers</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adygei</td>
<td>2</td>
<td>5.12</td>
<td>Europe</td>
<td>1</td>
<td>6066.984</td>
</tr>
<tr>
<td>Balochi</td>
<td>1.75</td>
<td>4.21</td>
<td>SouthAsia</td>
<td>0</td>
<td>5842.057</td>
</tr>
<tr>
<td>Bantu_N</td>
<td>2.36</td>
<td>3.36</td>
<td>Africa</td>
<td>1</td>
<td>5469.913</td>
</tr>
<tr>
<td>Bantu_S</td>
<td>2.38</td>
<td>3.63</td>
<td>Africa</td>
<td>1</td>
<td>4306.19</td>
</tr>
<tr>
<td>Bedouin</td>
<td>1.91</td>
<td>4.66</td>
<td>MiddleEast</td>
<td>0</td>
<td>2887.254</td>
</tr>
<tr>
<td>Biak</td>
<td>2.63</td>
<td>2.44</td>
<td>Africa</td>
<td>0</td>
<td>1335.496</td>
</tr>
<tr>
<td>Brahui</td>
<td>1.84</td>
<td>4.44</td>
<td>SouthAsia</td>
<td>0</td>
<td>6475.605</td>
</tr>
<tr>
<td>Burushi</td>
<td>1.72</td>
<td>4.04</td>
<td>SouthAsia</td>
<td>1</td>
<td>6132.567</td>
</tr>
<tr>
<td>Cambodiari</td>
<td>2</td>
<td>2.5</td>
<td>EastAsia</td>
<td>1</td>
<td>11762.11</td>
</tr>
<tr>
<td>Colombi</td>
<td>2.71</td>
<td>4.57</td>
<td>America</td>
<td>0</td>
<td>24177.34</td>
</tr>
<tr>
<td>Dai</td>
<td>2.2</td>
<td>6.3</td>
<td>EastAsia</td>
<td>1</td>
<td>10213.13</td>
</tr>
<tr>
<td>Daur</td>
<td>2.3</td>
<td>4.7</td>
<td>EastAsia</td>
<td>1</td>
<td>10896.21</td>
</tr>
<tr>
<td>Druze</td>
<td>1.95</td>
<td>4.82</td>
<td>MiddleEast</td>
<td>1</td>
<td>2887.254</td>
</tr>
<tr>
<td>French</td>
<td>1.89</td>
<td>5.11</td>
<td>Europe</td>
<td>1</td>
<td>5249.04</td>
</tr>
<tr>
<td>French_Bas</td>
<td>1.46</td>
<td>4.88</td>
<td>Europe</td>
<td>1</td>
<td>5857.478</td>
</tr>
<tr>
<td>Han</td>
<td>1.98</td>
<td>4.18</td>
<td>EastAsia</td>
<td>1</td>
<td>9854.752</td>
</tr>
<tr>
<td>Hazara</td>
<td>2.3</td>
<td>4.7</td>
<td>SouthAsia</td>
<td>0</td>
<td>6253.618</td>
</tr>
<tr>
<td>Hezhen</td>
<td>1.89</td>
<td>3.56</td>
<td>EastAsia</td>
<td>0</td>
<td>9299.626</td>
</tr>
<tr>
<td>Japanese</td>
<td>1.72</td>
<td>4.28</td>
<td>EastAsia</td>
<td>1</td>
<td>9919.105</td>
</tr>
<tr>
<td>Kalash</td>
<td>1.67</td>
<td>4.96</td>
<td>SouthAsia</td>
<td>1</td>
<td>5704.994</td>
</tr>
<tr>
<td>Karitiana</td>
<td>2.5</td>
<td>4.29</td>
<td>America</td>
<td>1</td>
<td>24330.14</td>
</tr>
<tr>
<td>Lahu</td>
<td>1.75</td>
<td>3.13</td>
<td>EastAsia</td>
<td>1</td>
<td>9875.318</td>
</tr>
<tr>
<td>Makrani</td>
<td>1.92</td>
<td>4.76</td>
<td>SouthAsia</td>
<td>0</td>
<td>6178.761</td>
</tr>
<tr>
<td>Mandenka</td>
<td>2.29</td>
<td>3.21</td>
<td>Africa</td>
<td>1</td>
<td>3629.649</td>
</tr>
<tr>
<td>Maya</td>
<td>2.45</td>
<td>4.68</td>
<td>America</td>
<td>1</td>
<td>19825.71</td>
</tr>
<tr>
<td>Mbuti</td>
<td>2.62</td>
<td>2.54</td>
<td>Africa</td>
<td>0</td>
<td>3872.417</td>
</tr>
<tr>
<td>Miaozu</td>
<td>2.3</td>
<td>4.6</td>
<td>EastAsia</td>
<td>1</td>
<td>10290.53</td>
</tr>
<tr>
<td>Mongola</td>
<td>1.8</td>
<td>5</td>
<td>EastAsia</td>
<td>0</td>
<td>9131.367</td>
</tr>
<tr>
<td>Mozabite</td>
<td>2.07</td>
<td>4.59</td>
<td>MiddleEast</td>
<td>1</td>
<td>2844.946</td>
</tr>
<tr>
<td>NAN_Mela</td>
<td>1.15</td>
<td>4.54</td>
<td>Oceania</td>
<td>1</td>
<td>14843.12</td>
</tr>
<tr>
<td>Naxi</td>
<td>1.33</td>
<td>3.89</td>
<td>EastAsia</td>
<td>1</td>
<td>7071.967</td>
</tr>
<tr>
<td>North_Itali</td>
<td>1.69</td>
<td>5.77</td>
<td>Europe</td>
<td>1</td>
<td>5305.81</td>
</tr>
<tr>
<td>Orcadian</td>
<td>2.13</td>
<td>5.67</td>
<td>Europe</td>
<td>1</td>
<td>4659.217</td>
</tr>
<tr>
<td>Oroqen</td>
<td>2.67</td>
<td>3.78</td>
<td>EastAsia</td>
<td>0</td>
<td>10817.81</td>
</tr>
<tr>
<td>Palestinian</td>
<td>1.98</td>
<td>4.76</td>
<td>MiddleEast</td>
<td>1</td>
<td>3946.438</td>
</tr>
<tr>
<td>Papuan</td>
<td>1.88</td>
<td>5.59</td>
<td>Oceania</td>
<td>1</td>
<td>22662.78</td>
</tr>
<tr>
<td>Pathan</td>
<td>1.71</td>
<td>3.67</td>
<td>SouthAsia</td>
<td>1</td>
<td>6201.7</td>
</tr>
<tr>
<td>Pima</td>
<td>2.5</td>
<td>3.79</td>
<td>America</td>
<td>1</td>
<td>18015.79</td>
</tr>
<tr>
<td>Russian</td>
<td>1.8</td>
<td>5.2</td>
<td>Europe</td>
<td>1</td>
<td>6012.261</td>
</tr>
<tr>
<td>San</td>
<td>2.83</td>
<td>2.83</td>
<td>Africa</td>
<td>0</td>
<td>5842.057</td>
</tr>
<tr>
<td>Sardinian</td>
<td>1.79</td>
<td>5.46</td>
<td>Europe</td>
<td>1</td>
<td>5118.374</td>
</tr>
<tr>
<td>She</td>
<td>2.1</td>
<td>3.9</td>
<td>EastAsia</td>
<td>1</td>
<td>9832.496</td>
</tr>
<tr>
<td>Sindhi</td>
<td>1.88</td>
<td>4.08</td>
<td>SouthAsia</td>
<td>1</td>
<td>10123.19</td>
</tr>
<tr>
<td>Surui</td>
<td>2</td>
<td>6</td>
<td>America</td>
<td>1</td>
<td>24330.14</td>
</tr>
<tr>
<td>Tu</td>
<td>2</td>
<td>3.3</td>
<td>EastAsia</td>
<td>0</td>
<td>7110.29</td>
</tr>
<tr>
<td>Tujia</td>
<td>1.7</td>
<td>3.5</td>
<td>EastAsia</td>
<td>1</td>
<td>8868.139</td>
</tr>
<tr>
<td>Tuscan</td>
<td>2</td>
<td>5.25</td>
<td>Europe</td>
<td>1</td>
<td>5116.516</td>
</tr>
<tr>
<td>Uygur</td>
<td>2.3</td>
<td>2.5</td>
<td>SouthAsia</td>
<td>0</td>
<td>10260.55</td>
</tr>
<tr>
<td>Xibo</td>
<td>1.89</td>
<td>3.89</td>
<td>EastAsia</td>
<td>1</td>
<td>9328.79</td>
</tr>
<tr>
<td>Language</td>
<td>Value 1</td>
<td>Value 2</td>
<td>Region</td>
<td>Count</td>
<td>Frequency</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>---------</td>
<td>-------------</td>
<td>-------</td>
<td>-----------</td>
</tr>
<tr>
<td>Yakut</td>
<td>1.92</td>
<td>4.4</td>
<td>EastAsia</td>
<td>0</td>
<td>16168.51</td>
</tr>
<tr>
<td>Yizu</td>
<td>2</td>
<td>2.9</td>
<td>EastAsia</td>
<td>1</td>
<td>9869.853</td>
</tr>
<tr>
<td>Yoruba</td>
<td>2.73</td>
<td>3.41</td>
<td>Africa</td>
<td>1</td>
<td>2384.859</td>
</tr>
</tbody>
</table>