Inferring adaptive gene-flow in recent African history

George Busby[1,*], **Ryan Christ**[2], **Gavin Band**[1,3], **Ellen Leffler**[1,3], **Quang Si Le**[1], **MalariaGEN**[4], **Kirk Rockett**[1,3], **Dominic Kwiatkowski**[1,3], **and Chris Spencer**[1]

[1]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK

[2]Department of Statistics, University of Oxford, 24-29 St Giles, Oxford OX1 3LB, UK

[3]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

[4]https://www.malariagen.net/network/people

[*]george@well.ox.ac.uk

Gene-flow from an ancestrally differentiated group has been shown to be a powerful source of selectively adventageous variants. To understand whether recent gene-flow may have contributed to adaptation among humans in sub-Saharan Africa we applied a novel method to identify deviations in ancestry inferred from genome-wide data from 48 populations. Among the signals of ancestry deviation that we find in the Fula ethnic group from the Gambia, who are historically pastoralist, the region that encodes the lactose persistence phenotype, *LCT/MCM6*, has the highest proportion of Eurasian ancestry in the genome. The region with the lowest proportion of non-African ancestry is across *DARC*, which encodes the Duffy null phenotype and is protective for *Plasmodium vivax* malaria. In the Jola from the Gambia and a Khoesan speaking group from Namibia we find multiple regions with inferred ancestry deviation including the Major Histocompatibility Complex. Our analysis shows the potential for adaptive gene-flow in recent human history.

Introduction

The source of adaptive genetic variation remains an important question in evolutionary biology[1]. In the canonical view of evolution, natural selection causes a change in the frequency of a mutation that was either present before the introduction of selection, so-called standing variation, or a new mutation that arose after the selective pressure began[2]. A third potential source of novel genetic variation on which natural selection can act is via gene-flow resulting from an inter or intra species admixture event. Examples of such adaptive introgression include the exchange of mimicry adaptations amongst *Heliconius* butterfly species[3], the transfer of insecticide resistance genes between sibling *Anopheles* mosquito species[4], and the spread of pesticide resistance across populations of house mice[5].

In the human lineage, recent genomic studies have shown that our demographic history is complex, involving population merges as well as splits[6–10]. These merges, also known as admixture or introgression events, open the door to the transfer of adventageous alleles. The trysts between archaic *Homo sapiens* and Neanderthals over fifty thousand years ago (kya) transfered beneficial alleles from Neanderthals into the human lineage[6–8,10–12]. While genomic regions with elevated archaic ancestry have been observed, it has also been possible to identify regions that are devoid of archaic ancestry and where it is thought that archaic alleles at these loci have been at a disadvantage[13,14].

Despite the prevalence of more recent admixture events over the past 50 kya of human history, relatively few instances of adaptation from recent merges have been found. Notable known examples include the exchange of high altitude adaptations between the ancestors of Sherpa and Tibetans[15] and the spread of the *Plasmodium vivax* malaria-proecting Duffy null mutation in Madagascar as result of gene-flow from mainland African Bantu-speakers[16].

Within Africa, multiple studies using different techniques have shown admixture to be a common theme in the recent history of the continent[17–23]. As a result, the genomes of individuals from these populations contain segments which derive from multiple ancestries. Inferring whether gene-flow has assisted adaptation requires inferring how these ancestries change locally along the genome. Although a number of strategies exist for local ancestry inference[24–27], most are designed to distinguish at best continental-level ancestry from a small number of reference populations. To identify whether haplotypes

of specific ancestry have potentially moved within, as well as from outside of, Africa, the approach we develop here involves sampling from a Hidden Markov Model to identify the likely donor haplotype from a large set of reference genomes. We use this chromosome painting approach to assign a donor ancestry label (which can be at individual, population, region, or continental level) at each locus in the genome to all recipient individuals. Iterating this process incorporates ancestry assignment uncertainty and we infer the ancestry proportions from different donor groups across the genome.

Here we use *ChromoPainter*[28] to infer ancestry across the genomes of 3,283 individuals from a set of 60 donor groups, and use this analysis to ask a simple question: across individuals in these admixed populations, are there regions of the genome where ancestry is significantly deviated away from genome-wide expectations? We construct a statistical model to test for significant deviations. We interpret deviations in local ancestry as possibly resulting from natural selection acting to either increase the frequency of the introgressed haplotype (positive selection) or prevent it from replacing an established haplotype (negative selection). We highlight specific examples where ancestry deviations align with known targets of selection, describe patterns of selection across the data, and discuss the challenges in using approaches of this kind for detecting adaptive admixture.

Results and Discussion

Inferring local ancestry

We used a published[22] dataset containing computationally phased haplotypes from 3,283 individuals, from 60 worldwide populations, typed at 328,176 high quality genome-wide SNPs [Supplementary Fig. 1]. We grouped individuals from these populations into 8 separate ancestry regions, based on genetic and ethno-linguistic similarity, as described previously[22] [Fig. 1a]. We painted chromosomes from these populations using *ChromoPainter* with donors which did not include any closely related populations, here defined as individuals from within their own ancestry region. Therefore, for each haplotype, there are seven non-local ancestries (or colours of paint) from which they can copy.

Identifying changes in local ancestry

Our approach for inferring ancestry deviations is based on the idea that ancestry tracts shared between two populations will reveal a signal of admixture[9], but will be randomly distributed amongst the genomes of individuals within those populations[29][Fig. 1b]. So, whilst individuals will have mosaic (i.e. block-like) ancestry [Fig. 1c], when summed across all individuals in a population, ancestry proportions at specific loci (SNPs) will resemble genome-wide proportions in expectation [Fig. 1d]. In this framework, we would expect ancestry proportions at a SNP to match genome-wide proportions, with some variation which we can model. A significant deviation away from expected ancestry proportions suggests that natural selection may have contributed to the change. We note that this will only ever allow for indirect inference of selection. This is nevertheless similar in spirit to other commonly used selection identification methods, such as the integrated haplotype score (iHS)[30,31] which scans the genomes of a population for longer than expected haplotypes based on their frequency, which are then (indirectly) inferred to have swept as a result of selection.

To help account for the uncertainty in ancestry inference we sampled chromosome paintings 10 times for each individual. We use a binomial likelihood to test whether the ancestry at each SNP along the genome differs from expectations. Specifically, for each recipient haplotype, we infer the expected proportion of ancestry from each non-self ancestry region from their genome-wide ancestry paintings. At each position, we fit a deviation parameter, β, that best explains the sampled ancestries at that SNP across all individuals. β can be thought of as the increase (or decrease) in expected ancestry proportions across individuals on the logit scale that best explains the proportions observed at a SNP compared to the genome-wide average. To generate a P value, we compared the likelihood ratio test statistic to a χ^2 distribution. At each SNP in the genome, this test gives a P value and estimate of β for each of the 7 non-self ancestries tested.

Modelling ancestry in this way has the advantage that it adjusts for an individual's specific ancestry proportions. The likelihood is a product across individuals which assumes that individuals are independent given the underlying ancestry proportions. Quantile-quantile (Q-Q plots) of the observed test statistics typically show high levels of inflation. There are at least three explanations for this inflation. Firstly, due to shared ancestral histories among haplotypes from the same population, the paintings are unlikely to be

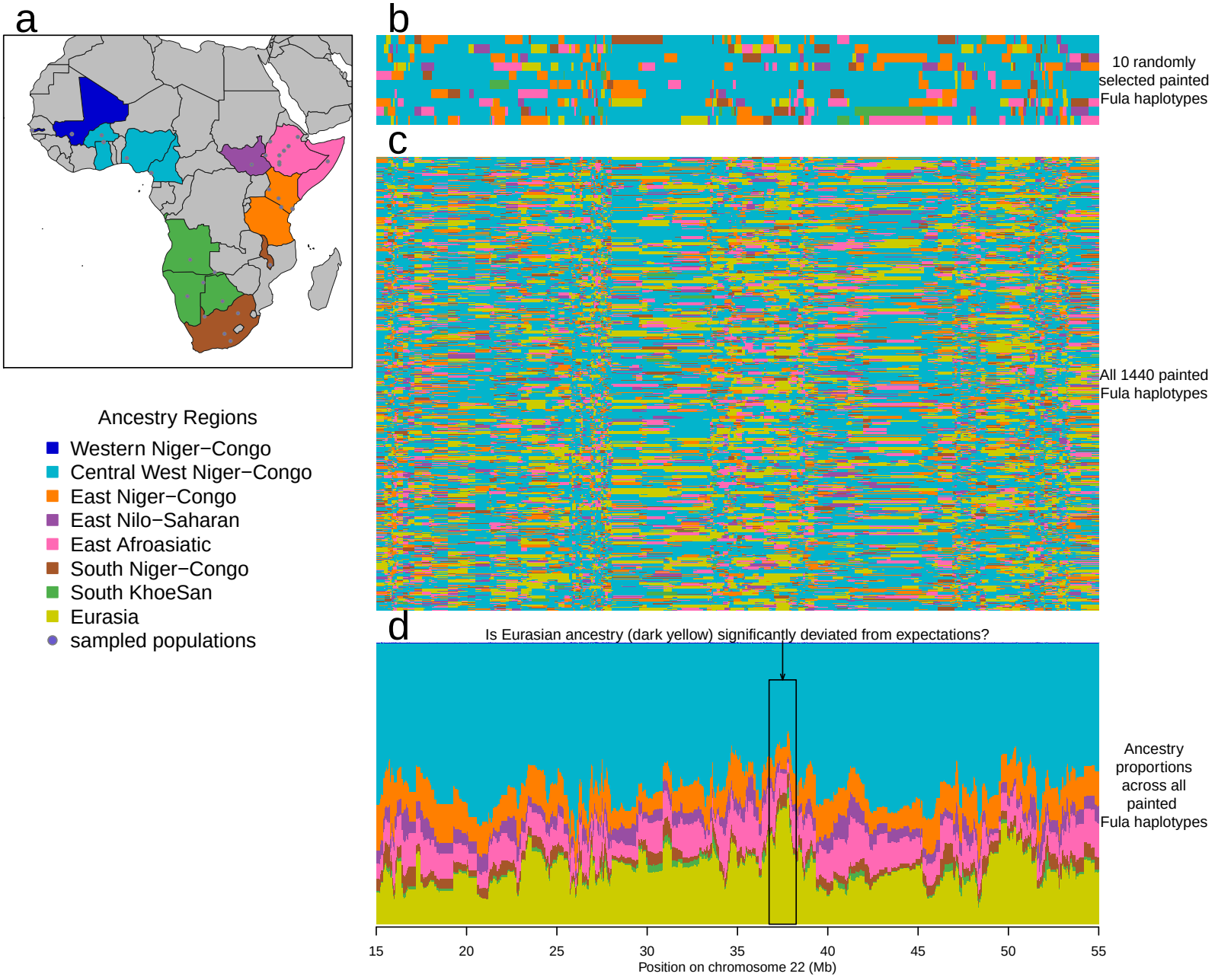


Figure 1 | Overview of local ancestry based analysis. **a** Location of the 48 populations from seven African ancestry regions used in the analysis; **b** chromosome painting samples (rows) of chromosome 22 (3392 SNPs) from ten Fula haplotypes show mosaic local ancestry; **c** chromosome painting samples from all 144 Fula haplotypes, each set of 10 consecutive rows are paintings from one recipient haplotype; **d** chromosome painting samples from (**c**) stacked to show ancestry proportions locally along chromosome 22.

independent between samples in a population, so this assumption is likely to be false. The demographic process that has given rise to the ancestry paintings is more complex than the model, leading to local variations in paintings that are not captured by the null. Secondly, the chromosome painting model explicitly uses correlations in ancestry locally in the genome (i.e. linkage disequilibrium) meaning that tests are not independent along the genome, as shown by the autocorrelation plot in Supplementary Figure 3. Finally, adaptive introgression may have a weak effect genome-wide. We discuss below alternative approaches that model the variance (as well as the expectation) in ancestry proportions across the genome. However, these limitations mean that it is difficult to estimate a genome-wide threshold for significance in the traditional sense (i.e. for defining false positive/negative rates), so we view *P* values as a ranking rather than absolute terms.

Ancestry deviations in the Fula identify known targets of selection

We illustrate our approach with an analysis of genomes from the Fula ethnic group from The Gambia. The Fula (also known as Fulani) are historically nomadic pastoralists spread across West Africa who we inferred in our previous analysis[22] to have experienced an admixture event involving a largely Eurasian source (0.19 admixture proportion) mixing with a West African source around 1,800 years ago (239CE (95% CI = 199BCE-486CE)). (Admixture events across all analysis populations inferred previously[22] are summarised in Supplementary Figure 2). This event introduced a significant amount of Eurasian ancestry into the Fula which is relatively easy to identify given that Eurasian populations split and drifted from African populations over 50kya[32]. They thus have haplotypes that are easier to differentiate from each other, making it relatively well powered to detect deviations in ancestry proportions.

In the Fula, the region of the genome with the lowest proportion of Eurasian ancestry ($-log_{10}P = 10$, β=-4 across all Fula individuals) is on chromosome 1 and contains the Duffy Antigen Receptor *DARC* gene [Fig. 2a] whilst the region with the highest level of Eurasian ancestry ($-log_{10}P = 15$, β=2) is on chromosome 2 and contains the *LCT* and *MCM6* genes [Fig. 2b]. Polymorphisms in *DARC* form the basis of the Duffy blood group system, and the Duffy null mutation, which provides resistance to *Plasmodium vivax* malaria[33] is almost fixed across Africa and absent outside[34]. Our observation of less Eurasian ancestry than expected at this locus suggests that African haplotypes have been beneficial at this locus

after the admixture event that brought the Eurasian ancestry into the Fula; that is, within the last 2,000 years. Similar effects at this locus have been observed before: based on an analysis of allele frequencies, selection following admixture between Austronesians and sub-Saharan Bantus is likely to have driven frequencies of the Duffy null mutation to higher than expectations based on admixture proportions in the Malagasy of Madagascar[16], and a recent study of Sahel populations showed an excess of African ancestry at this locus in Sudanese Arabs and Nubians[35].

Mutations in an intron in *MCM6*, a regulatory element for /emphLCT lead to the lactose persistence phenotype[36] (the ability to digest milk into adulthood), and represent one of the clearest signals of recent natural selection in the genome. Lactose persistence has evolved at least twice independently[37], in Europe and in East Africa[38]. Encouragingly, our analysis identifies this Eurasian haplotype, despite the SNP set not containing the 'European' lactase variant (13910 C>T polymorphism, rs4988235; Supplementary Fig. 4) and suggests that the European haplotype has entered the Fula as a result of gene-flow from Europe approximately 2kya. In both this case and that of *DARC* described above, our ancestry based analysis provides new insight into the potential origins of selected mutations.

The genome-wide distribution of ancestry deviations in the Fula

We note that because Eurasian and Khoesan haplotypes are likely to be more diverged from those of other sub-Saharan African groups we are better able to characterise them and so have more power to identify deviations concerning these ancestries. Indeed, over half (21) of the top 40 ancestry deviations observed in the Fula involve Eurasian ancestry [Supplementary Table 1; Supplementary Fig. 5], including increases of Eurasian ancestry across the *HLA* region on chromosome 6 ($-log_{10}P = 12.38$,$\beta = 1.29$), a group of *DSC* genes on chromosome 18 ($-log_{10}P = 13.55$, $\beta = 1.35$), and at *APOL* genes on chromosome 22 ($-log_{10}P = 13.01$, $\beta = 1.32$). Of the non-Eurasian signals, of note are a significant increase in Central West African ancestry across the *GYP* genes on chromosome 4 [Fig. 2c] ($-log_{10}P = 8.96$, $\beta = 1.13$), encompassing a region which has recently been identified as being associated with malaria[39,40], and which, based on a recent iHS analysis[41], shows signs of being under natural selection. We observe an increase in Khoesan ancestry across the *SERPIN* genes on chromosome 18 [Fig. 2d] ($-log_{10}P = 19.32$, $\beta = 2.98$), one of which, *SERPINB11*, has previously been implicated as a target of natural selection in pastoralist

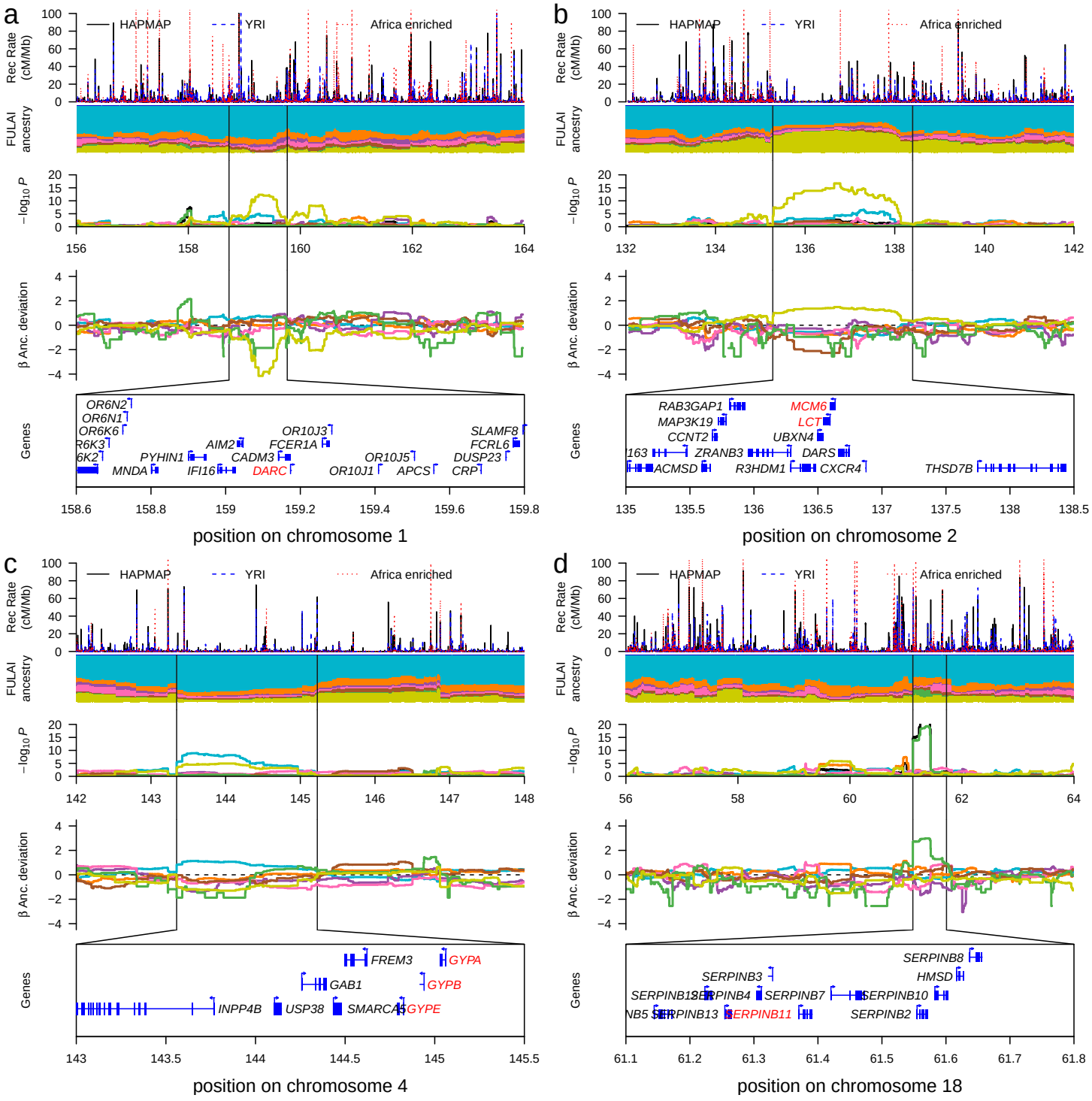


Figure 2 | Local ancestry deviations in the West African Fula. Each main panel is made up of 5 sub-panels showing, from top to bottom: the recombination rate across the region; ancestry proportions coloured by region as in Figure 1; evidence for deviation based on the binomial deviation model; *P* values for the likelihood ratio test, coloured by ancestry; the associated β coefficient fit by the LRT for each ancestry; the genes in the region. **a** lack of Eurasian ancestry around *DARC* on chromosome 1; **b** excess Eurasian ancestry around the *LCT* and *MCM6* genes on chromosome 2; **c** increased Central West African ancestry at the *GYP* genes on chromosome 4. **d** increased Khoesan copying at the *SERPIN* genes on chromosome 18

groups in Africa[42].

We next looked at the evidence for covariance in ancestry deviation at SNPs across the genome. As expected, in all cases we found negative covariance in ancestry at SNPs where we observed a significant deviation signal [Fig. 3a]. When one ancestry increases in frequency at a locus, it must be at the detriment to other ancestries. This association was more pronounced for more closely related ancestries. For example, the largest negative relationships were seen between East African groups [red boxes in Fig. 3a]. Interestingly, when we looked at the distribution of βs at significant SNPs, there was a general trend towards signals of ancestry *increases*, except in the case of Eurasian (yellow) and Central West African (light blue) ancestry [Fig. 3c].

Accounting for reciprocal local ancestry deviations

One concern with our analysis is that deviations in local ancestry might result from reciprocal shared ancestry. That is, a *non-local* donor individual might share haplotypes with a recipient population because they themselves are admixed from the recipient group. We will observe such reciprocal copying in our framework when a single, or small number of, donor individuals contribute excessively to an ancestry deviation signal. An example can be seen in Figure 4. Across all Gambian populations from West Africa (of which we show the Wollof as an example) we observed a significant increase in East African ancestry across the 32.2-33Mb region of chromosome 6, which contains several *HLA* genes [Fig. 4a and 4b]. However, this signal is driven almost entirely by Wollof individuals copying from a single Nilo-Saharan speaking individual, Anuak11. When we painted this individual with both non-local and local donors (i.e. including individuals from the Nilo-Saharan ancestry region), across this region we observe that one haplotype from this individual copies a roughly 1Mb chunk of their genome from the West African ancestry region that includes Gambian populations [dark blue in Fig. 4c] even though there were donors from the more closely related Nilo-Saharan ancestry region available. Our analysis of admixture in the Anuak showed an admixture event involving Central West African and Nilo-Saharan sources in 703CE (95% CI 427CE-1037CE) [Supplementary Fig. 2], which may have provided a vehicle for West African ancestry to enter this population.

One way to quantify this effect is to look at the number of unique donor haplotypes from an ancestry

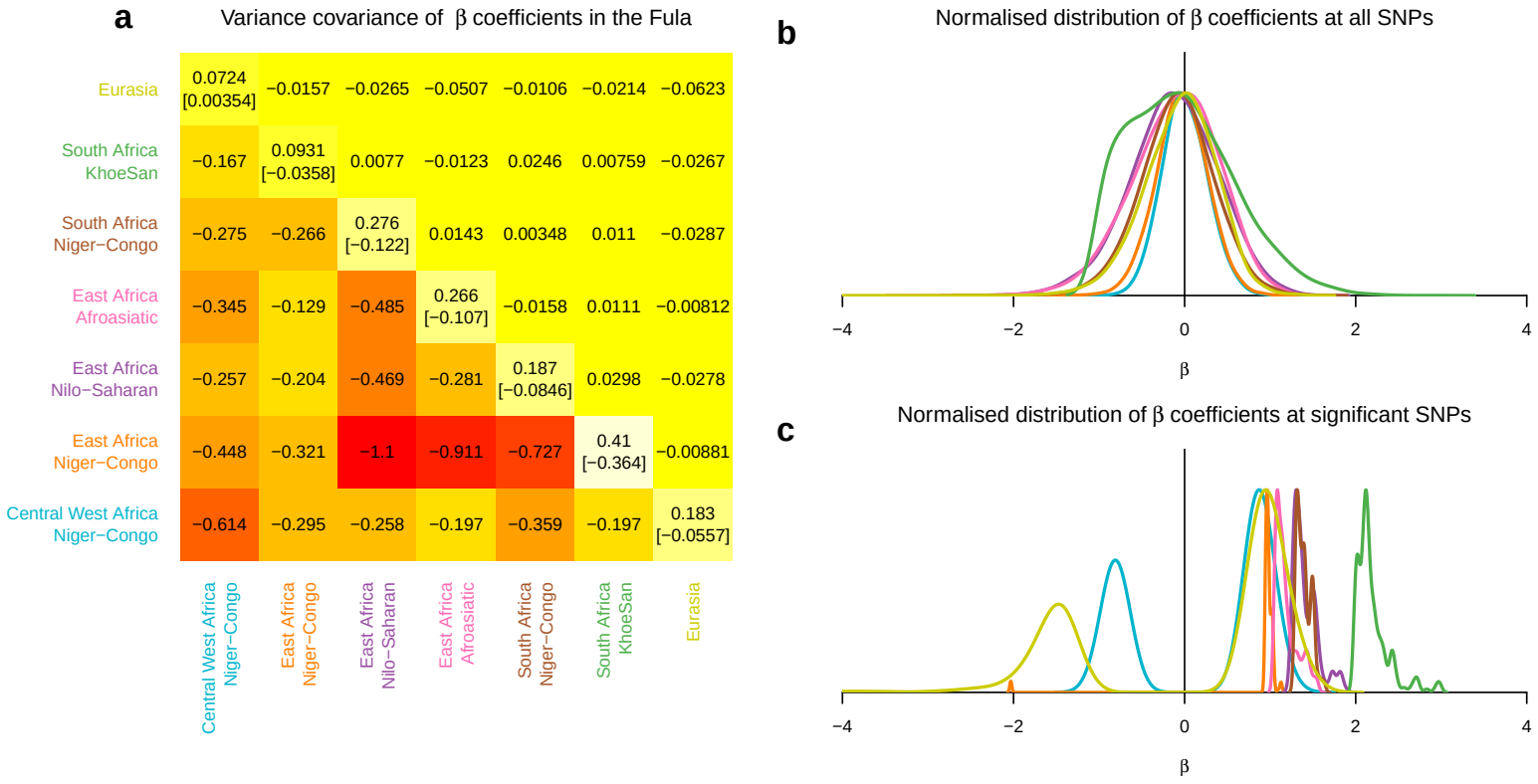


Figure 3 | Negative covariance of ancestry deviations in the Fula. **a** top right diagonal shows of heatmap shows covariance at genome-wide SNPs; bottom left are SNPs that are significant in either of the ancestry region pair; the diagonal shows the variance and mean in brackets within each ancestry region **b** the genome-wide distribution of β coefficients, split by ancestry, across all SNPs and normalised to have a maximum density of 1. **c** the distribution of β coefficients for significant SNPs, normalised to have a maximum density of 1. in panels **b** and **c** lines are coloured by ancestry as in panel **a**

region contributing to a signal as a proportion of the total number of donors. This quantity will be 1 if every recipient individual copies from a different donor haplotype and will tend towards zero as the number of unique donors copied from goes down. Whilst for Nilo-Saharan ancestry in the Wollof the median proportion of unique donors contributing to a signal genome-wide is 0.342 (95%CI = 0.20-0.55) [Fig. 4e], the average across the 32.2-33.Mb region of chromosome 6 is 0.072, which lies significantly outside of the genome-wide distribution (empirical outlying P value = 0.0007). We can see that this signal is driven by a low proportion of unique donors from the Nilo-Saharan ancestry region [purple line in Fig. 4f] and a large increase in copying from the most copied haplotype [Fig. 4g]. Because this effect may be the result of the donor individual being admixed from the recipient population, in such cases, it is not possible to infer the directionality of ancestry sharing. Therefore, to guard against reciprocal copying we performed an extra filter on the results. At each locus in the genome, we computed the number of unique donor haplotypes that contributed to the total amount of copying from that region. If the number of unique donors was ≤ 3 or the proportion of unique donors was < 0.1, we assumed the result was due to reciprocal copying.

Patterns of ancestry deviation across populations

In addition to the Fula, we observe multiple ancestry deviations in another Gambian population the Jola [Fig. 5; Supplementary Fig. 6] . In the Jola, the most deviated region of the genome involves a significant increase in Eurasian ancestry at 27.3-28.2Mb on chromosome 6 which containing several *HLA* genes ($-log_{10}P = 12.76, \beta = 1.94$), a signal that was also observed in the Wollof, also from the Gambia ($-log_{10}P = 10.41, \beta = 1.57$). We also observed a significant increases of Eurasian ancestry at *MYO3B* ($-log_{10}P = 13.53,\ \beta = 1.97$) and *SMOC2* ($-log_{10}P = 9.40,\ \beta = 1.75$), and an increase in Khoesan ancestry on chromosome 11 at *NLRP10* ($-log_{10}P = 12.87,\ \beta = 2.24$), which codes for a protein involved in inflammation and apoptosis[43,44].

In just under two thirds (31) of the other 46 sub-Saharan African populations we observed no significant genome-wide deviations in ancestry using a conservative significance threshold of $-log_{10}P \geq 8$ [Fig. 5; Supplementary Table 1], and observed multiple (> 5) significant signals in just 3 populations: the Jola and Fula from the Gambia, and the Ju/hoansi from Namibia. Around a quarter (39/159; 24.5%) of the loci

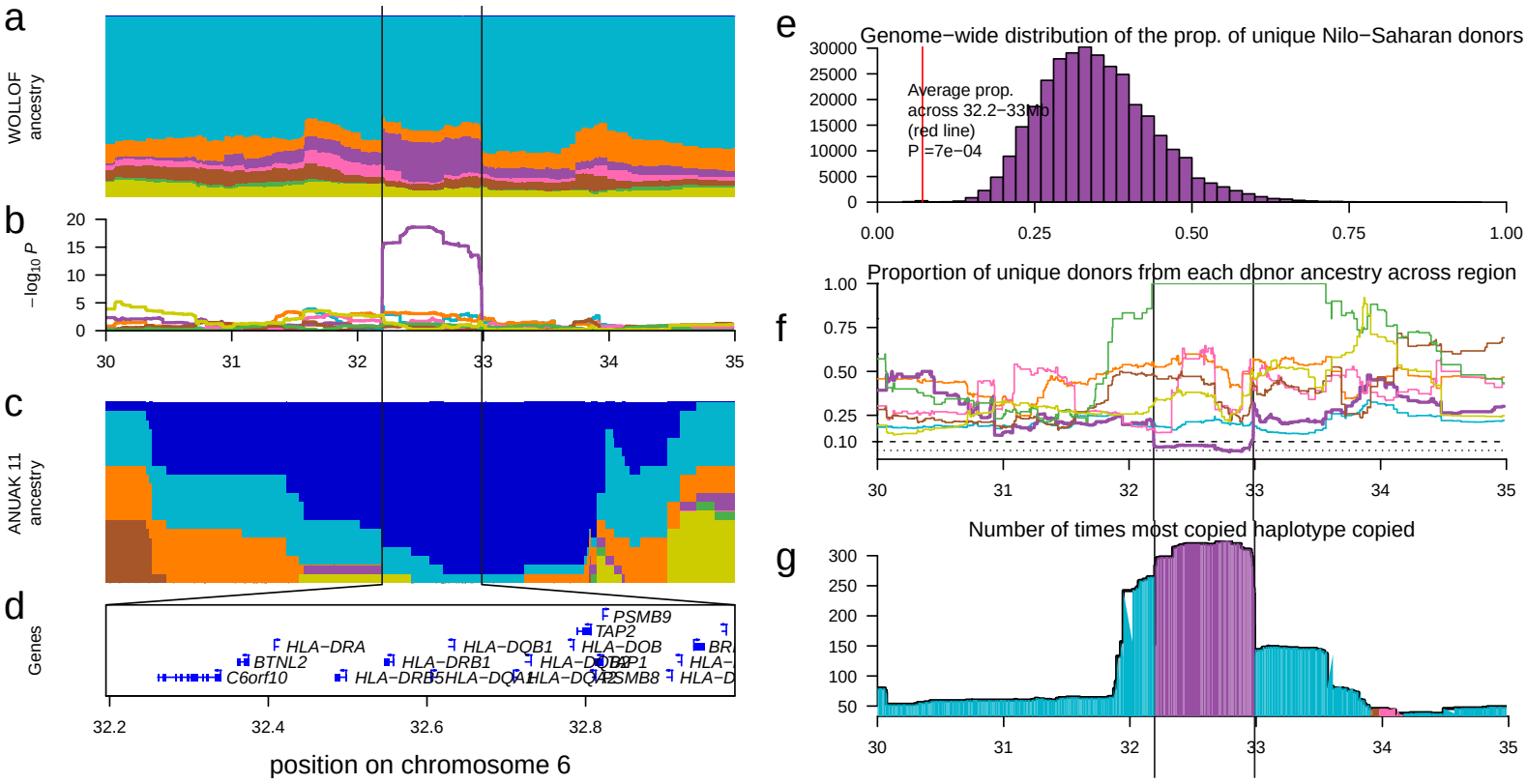


Figure 4 | Investigating local ancestry at the *HLA* in the Wollof. **a** local ancestry across 2160 Wollof haplotypes at the 30-35Mb region of chromosome 6. **b** results of the binomial ancestry deviation test shows a significant increase in Nilo-Saharan (purple) ancestry between 32.2 and 33Mb. **c** ancestry proportions from a local painting analysis in a single Anuak individual showing Western African (dark blue) ancestry. **d** UCSC genes across this region. **e** genome-wide distribution of unique Nilo-Saharan donors in the Wollof. **f** the proportion of unique donors when each of 7 non-self ancestries are copied across the 30-35Mb region of chromosome 6. **g** the number of times the most common haplotype is copied; the colour under the line represents the ancestry of the most copied individual.

identified across the 48 populations, of which 32 were in the Jola, Fula and Ju/hoansi, were discounted from further analysis because of potential reciprocal copying, resulting in a final list of 114 loci, 79% of which were in the Fula, Jola, and Ju/hoansi.

In the Ju/hoansi, the top region of differentiation in the genome involves an increase in East Africa Niger-Congo ancestry at 26.1-26.6Mb on chromosome 10 ($-log_{10}P = 17.14$, $\beta = 2.07$) which contains *MYO3A*, a gene involved in hearing[45], and *GAD2*, which has been associated with decreased body mass index in Europeans[46]. We also observe a significant increase ($-log_{10}P = 11.48$, $\beta = 1.86$) of East Africa Afroasiatic ancestry at 99.08-99.20Mb on chromosome 13, a region containing *FARP1* and *STK24*. *FARP1* is thought to be a member of the band 4.1 protein superfamily[47], a major structural component of red blood cell membranes[48]. We also see an increase in West African Niger-Congo ancestry at 47.65-47.76Mb on chromosome 6 a region containing *CD2AP*, *OPN5*, *GPR111*, and *GPR115* genes, and a large increase in Eurasian ancestry at *DISP1* on chromosome 1 ($-log_{10}P = 10.70$, $\beta = 2.88$).

Eight populations [Fig. 5] showed evidence of significant ancestry deviation at a short haplotype in the middle of the *LDB2* gene on chromosome 4. All 8 of these populations come from East Africa and are either Nilo-Saharan or Afroasiatic speakers [Supplementary Table 1; Supplementary Fig. 8]. We sought to characterise the nature of this signal further by combining data across all 235 individuals from these 8 populations (Sudanese, Somali, Anuak, Gumuz, Ari, Tigray, Amhara and Oromo) and the two other populations from the region with nominal evidence of deviation (Afar and Wolayta). In all cases, we observed a signal of increased Khoesan ancestry [Supplementary Fig. 8a] and localised this to increased copying specifically from the Ju/hoansi, Xun, and Amaxhosa of southern Africa [Supplementary Fig. 8d]. This differs considerably from a randomly selected SNP 1.5Mb downstream, with a more usual copying profile and where the Luhya and several Eurasian groups were typically copied more. *LDB2* codes for a LIM-domain binding protein, LDB2, which is a transcriptional regulator and has been linked to carotid atheroschlerosis[49].

Considerations of our approach to infer adaptive gene-flow

Whilst admixture appears to be an almost universal characteristic of the populations that we use in this analysis, we only find evidence of substantial adaptive gene-flow in three populations: the Fula and Jola

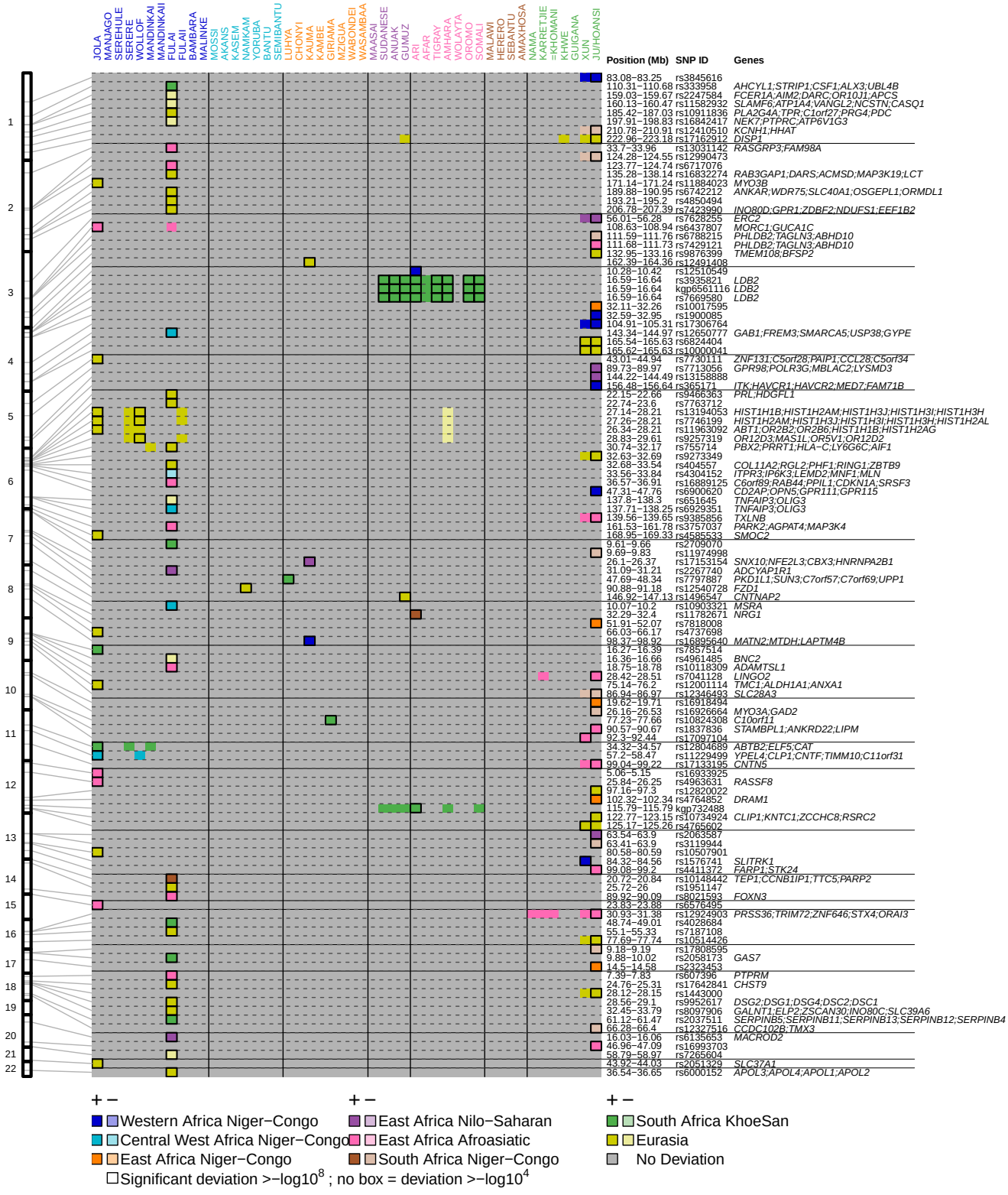


Figure 5 | Genome-wide ancestry deviations across sub-Saharan Africa. For each of the 114 loci (rows) where we observe ancestry deviations with a $-log_{10}P \geq 8$, for each population (columns) we show the ancestry that has changed (colour) and the direction of change (lighter colours show decrease, darker increase) and the strength of significance (black box $-log_{10}P \geq 8$; no box $8 \leq -log_{10}P \geq 4$; grey = $-log_{10}P < 4$. The positions of the significant loci are shown on the right and genes of interest, together with the detailed genomic region, are shown on the left.

from the Gambia and the Ju/hoansi from Namibia. Importantly, we infer significant recent admixture involving diverged source populations in these three groups, from Eurasian-like sources in the Gambians, and East African Asiatic-like source in the Ju/hoansi [Supplementary Fig. 2][22]. Our method is likely to have greater power to detect deviations in these scenarios. As such, it is perhaps not surprising that across populations from the Central West African ancestry region, where we previously inferred admixture events involving mostly closely related African sources[22] we infer a single example of adaptive gene-flow in the Namkam from Ghana at *FZD1* on chromosome 8 [Fig. 5]. It is important to note that even though all Gambian populations in our analysis had similar sample sizes and were painted using exactly the same procedure as the Jola and Fula, we do not see similar numbers of local ancestry changes in them, suggesting that such deviations are not linked to our chromosome painting procedure.

In addition to working with populations that have substantial recent admixture, there are further considerations to the results that come from our approach. The local ancestry inference procedure that we have used[28] is based on a fully probabilistic framework. To access uncertainty in local ancestry with the paintings, we generated 10 realisations (sampled paintings) of the painting algorithm and inferred ancestry proportions from these paintings. Because of the stochastic nature of such an approach, we are unlikely to capture the full uncertainty in local ancestry, but chose this approach based on computational requirements and the use of a similar approach in previous admixture inference work[9]. Nevertheless, future work to optimise the computation issues of storing the $n * n$ posterior copying probability matrix at each locus in the genome will likely provide a clearer assessment of local ancestry uncertainty. We used a binomial likelihood model, which takes each ancestry separately and models genome-wide proportions against local deviations. Using this model, we found no evidence that more than one ancestry increases or decreases in concert at a locus, at least in the Fula [Fig. 3]. Modelling changes across all ancestries at the same time, for example by using a multi-variate normal model, might provide further insights into part of the genome where we observe co-deviations.

Conclusion

Gene-flow between closely related groups is increasingly being implicated as a key driver of adaptation. Here we have sought to address whether this process has resulted from admixture events which we

previously inferred in African populations within the last 4,000 years. Our approach uncovered regions of the genome previously known to have undergone selection such as the *LCT/MCM6* gene complex in relation to lactose persistence, and *DARC* in relation to *P. vivax* malaria in the Fula of West Africa. In both cases, combining these observations with our understanding of the admixture history of the Fula demonstrates that changes in local ancestry can help pinpoint regions of the genome where natural selection may have been working. In addition, we identified other regions of the genome where local ancestry is deviates from expectations [Fig. 5], raising the possibility that adaptive gene-flow may be a more general, albeit rare, phenomenon. However, to evaluate how widespread this phenomenon is further work is needed to develop approaches that are able to detect more subtle signatures of ancestry deviations, and to account for the problems of non-independence highlighted above. Nevertheless, whilst adaptive introgression (i.e. adaptation resulting from interspecific gene-flow) has a deep literature, both theoretically[50] and empirically (for example in animals reviewed here[1]), examples of such an effect within species on a recent timescale are less common. Nonetheless, as large genome-scale datasets from humans and other species become available, the framework and methodology described here can be used to understand how widespread the phenomenon of adaptive gene-flow is.

References

1. Hedrick, P. W. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology* **22**, 4606–4618 (2013).

2. Hermisson, J. & Pennings, P. S. Soft Sweeps Molecular Population Genetics of Adaptation From Standing Genetic Variation. *Genetics* **169**, 2335–2352 (2005).

3. Consortium, T. H. G. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).

4. Clarkson, C. S. *et al.* Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation. *Nature Communications* **5**, 4248 (2014).

5. Staubach, F. *et al.* Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (Mus musculus). *PLOS Genetics* **8**, e1002891 (2012).

6. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328**, 710 –722 (2010).

7. Patterson, N. *et al.* Ancient Admixture in Human History. *Genetics* **192**, 1065–1093 (2012).

8. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).

9. Hellenthal, G. *et al.* A Genetic Atlas of Human Admixture History. *Science* **343**, 747–751 (2014).

10. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).

11. Abi-Rached, L. *et al.* The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans. *Science* **334**, 89–94 (2011).

12. Gittelman, R. M. *et al.* Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. *Current Biology* **26**, 3375–3382 (2016).

13. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **advance on** (2014).

14. Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics* **16**, 359–371 (2015).

15. Jeong, C. *et al.* Admixture facilitates genetic adaptations to high altitude in Tibet. *Nature Communications* **5** (2014).

16. Hodgson, J. A. *et al.* Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. *Proceedings of the Royal Society B: Biological Sciences* **281**, 20140930 (2014).

17. Pagani, L. *et al.* Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool. *The American Journal of Human Genetics* **91**, 83–96 (2012).

18. Schlebusch, C. M. *et al.* Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science* **338**, 374–379 (2012).

19. Pickrell, J. K. *et al.* The genetic prehistory of southern Africa. *Nature Communications* **3**, 1143 (2012).

20. Pickrell, J. K. *et al.* Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 2632–2637 (2014).

21. Pickrell, J. K. & Reich, D. Toward a new history and geography of human genes informed by ancient DNA. *Trends in Genetics* **30**, 377–389 (2014).

22. Busby, G. B. *et al.* Admixture into and within sub-Saharan Africa. *eLife* **5**, e15266 (2016).

23. Patin, E. *et al.* Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* **356**, 543–546 (2017).

24. Price, A. L. *et al.* Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genet* **5**, e1000519 (2009).

25. Baran, Y. *et al.* Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics (Oxford, England)* **28**, 1359–1367 (2012).

26. Brisbin, A. *et al.* PCAdmix: Principal Components-Based Assignment of Ancestry along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Human biology* **84**, 343–364 (2012).

27. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics* **93**, 278–288 (2013).

28. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics* **8**, e1002453 (2012).

29. Baird, S. J. E. Fisher's markers of admixture. *Heredity* **97**, 81–83 (2006).

30. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).

31. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A Map of Recent Positive Selection in the Human Genome. *PLoS Biol* **4**, e72 (2006).

32. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

33. Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The Resistance Factor to Plasmodium vivax in Blacks. *New England Journal of Medicine* **295**, 302–304 (1976).

34. Howes, R. E. *et al.* The global distribution of the Duffy blood group. *Nature Communications* **2**, 266 (2011).

35. Triska, P. *et al.* Extensive Admixture and Selective Pressure Across the Sahel Belt. *Genome Biology and Evolution* **7**, 3484–3495 (2015).

36. Ingram, C. J. E., Mulcare, C. A., Itan, Y., Thomas, M. G. & Swallow, D. M. Lactose digestion and the evolutionary genetics of lactase persistence. *Human Genetics* **124**, 579–591 (2008).

37. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics* **74**, 1111–1120 (2004).

38. Ranciaro, A. *et al.* Genetic Origins of Lactase Persistence and the Spread of Pastoralism in Africa. *The American Journal of Human Genetics* **94**, 496–510 (2014).

39. Malaria Genomic Epidemiology Network. A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature* **526**, 253–257 (2015).

40. Leffler, E. M. *et al.* A structural variant encoding hybrid glycophorins is associated with resistance to severe malaria. *bioRxiv* (2016).

41. Johnson, K. E. & Voight, B. F. Patterns of shared signatures of recent positive selection across human populations. *bioRxiv* (2017).

42. Seixas, S., Ivanova, N., Ferreira, Z., Rocha, J. & Victor, B. L. Loss and Gain of Function in SERPINB11: An Example of a Gene under Selection on Standing Variation, with Implications for Host-Pathogen Interactions. *PLoS ONE* **7**, e32518 (2012).

43. Inohara, N. & Nuñez, G. NODs: intracellular proteins involved in inflammation and apoptosis. *Nature Reviews. Immunology* **3**, 371–382 (2003).

44. Tschopp, J., Martinon, F. & Burns, K. NALPs: a novel protein family involved in inflammation. *Nature Reviews. Molecular Cell Biology* **4**, 95–104 (2003).

45. Walsh, T. *et al.* From flies' eyes to our ears: mutations in a human class III myosin cause progressive nonsyndromic hearing loss DFNB30. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 7518–7523 (2002).

46. Boesgaard, T. W. *et al.* A -243A–¿G polymorphism upstream of the gene encoding GAD65 associates with lower levels of body mass index and glycaemia in a population-based sample of 5857 middle-aged White subjects. *Diabetic Medicine: A Journal of the British Diabetic Association* **24**, 702–706 (2007).

47. Koyano, Y. *et al.* Molecular cloning and characterization of CDEP, a novel human protein containing the ezrin-like domain of the band 4.1 superfamily and the Dbl homology domain of Rho guanine nucleotide exchange factors. *Biochemical and Biophysical Research Communications* **241**, 369–375 (1997).

48. Conboy, J., Kan, Y. W., Shohet, S. B. & Mohandas, N. Molecular cloning of protein 4.1, a major structural element of the human erythrocyte membrane skeleton. *Proceedings of the National Academy of Sciences* **83**, 9512–9516 (1986).

49. Mm, S. *et al.* Lim domain binding 2: a key driver of transendothelial migration of leukocytes and atherosclerosis. *Arteriosclerosis, thrombosis, and vascular biology* **34**, 2068–2077 (2014).

50. Lewontin, R. C. & Birch, L. C. Hybridization as a Source of Variation for Adaptation to New Environments. *Evolution* **20**, 315–336 (1966).

51. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179–181 (2012).

Acknowledgements (not compulsory)

We thank ...

Author contributions statement

All authors reviewed the manuscript.

Additional information

To include, in this order: **Accession codes** (where applicable); **Competing financial interests** (mandatory statement).

Materials and Methods

Dataset

African population genetics dataset

We used the dataset of 3,283 individuals from 48 African and 12 Eurasian populations typed at 328,000 SNPs previously published here[22], and available at https://www.malariagen.net/resource/18. This dataset comprises 14 Eurasian and 46 African groups, mostly from sub-Saharan Africa [Fig. 1]. Half (23) of the African groups represent subsets of samples collected as part to the MalariaGEN consortium. Details of partners involved in the consortium and recruitment of samples in relation to studying malaria genetics are published elsewhere[39]. The remaining groups are from publicly available datasets and the 1000 Genomes Project, with Eurasian groups from the latter included to help understand the genetic contribution from outside of the continent. To study patterns of diversity, we created an integrated dataset at a common set of high-quality SNPs and used SHAPEITv2[51] to haplotypically phase these genotype data, generating a final set of 6566 haplotypes.

Defining Ancestry Regions

We used a combination of geographic, ethnic, and genetic information to group our sample into populations and eight ancestry regions [Fig. 1]. We analysed each population in turn, and inferred local ancestry using only those populations from outside of their ancestry region, which we term *non-local* paintings.

Chromosome painting procedure

Inferring population specific prior copying probabilities and painting parameters

We utilised *ChromoPainter*[28] to infer local ancestry. In order to do so we first needed to generate population-specific prior copying probabilities. The algorithm initially assumes that each donor haplotype is equally likely to be copied from, which in reality is not the case. Moreover, when there is less contextual information available for painting, for example at the beginning and ends of chromosomes, in its default setting, the algorithm will assign ancestry based on the on the initial prior copying probabilities, which will lead to noisy ancestry inference. We therefore performed initial runs of *ChromoPainter* where for each recipient population we used the Expectation-Maximisation (EM) algorithm to infer the prior

copying probabilities from each recipient group (using the $-ip$ flag). We ran ten EM iterations across 4 chromosomes (1,4,10,15) using 5 recipient individuals per population, and averaging across individuals and chromosomes to infer population specific prior copying probabilities.

We next estimated the recombination scaling constant, using the $-in$ flag, and the mutation (HMM emission) probabilities, using the $-im$ flag, separately for the same 5 individuals and 4 chromosomes noted above for each recipient population, and again averaged across chromosomes and individuals. We then painted all individuals from a given recipient population using these population-specific prior copying probabilities and model parameters.

Inferring local ancestry with ChromoPainter

To estimate local ancestry uncertainty, we take multiple sample paths through the HMM to generate 10 ancestry paintings per recipient haplotype. We note that in an ideal scenario, we would use *ChromoPainter* to infer the true uncertainty in the ancestry at a SNP by computing the full set of posterior marginal copying probabilities for each recipient haplotype at every position of the genome, giving us the probability that a recipient haplotype copies from all donor haplotypes at a given position. However, given N recipient haplotypes and M donor individuals, this would give an $N*M$ matrix at each of L loci for each recipient population, which quickly leads to computational storage issues. To overcome this, we instead store 10 painting samples, leading to $N*10$ matrix being stored at each locus and use these painting samples to estimate uncertainty in the local ancestry inference.

Identifying local deviations in local ancestry

Binomial logistic model

The approach we take for identifying deviations in local ancestry is based on the idea that, in the absence of any other population genetic process, an individual's ancestry proportions at a test SNP should be within the distribution of their genome-wide ancestry proportions. In theory, we can generate this null distribution of ancestry expectations based on the paintings at all genome-wide SNPs, or a subset thereof. In practice, we use a leave-one-out approach, and compute an individual's genome-wide ancestry proportions from all SNPs that are not on the same chromosome as our test SNP that we want to test.

We model the proportion from each donor ancestry region separately via a binomial model with a

mean estimated genome-wide and a linear increase on the logistic scale determined by β. We used a likelihood ratio test to determine if this parameter was significantly different from zero at its maximum.

$$H_0 : y_i \sim \text{Binom}(\theta, S) \ , \ \log\left[\frac{\theta}{1-\theta}\right] = \mu_i$$

$$H_1 : y_i \sim \text{Binom}(\theta, S) \ , \ \log\left[\frac{\theta}{1-\theta}\right] = \mu_i + \beta$$

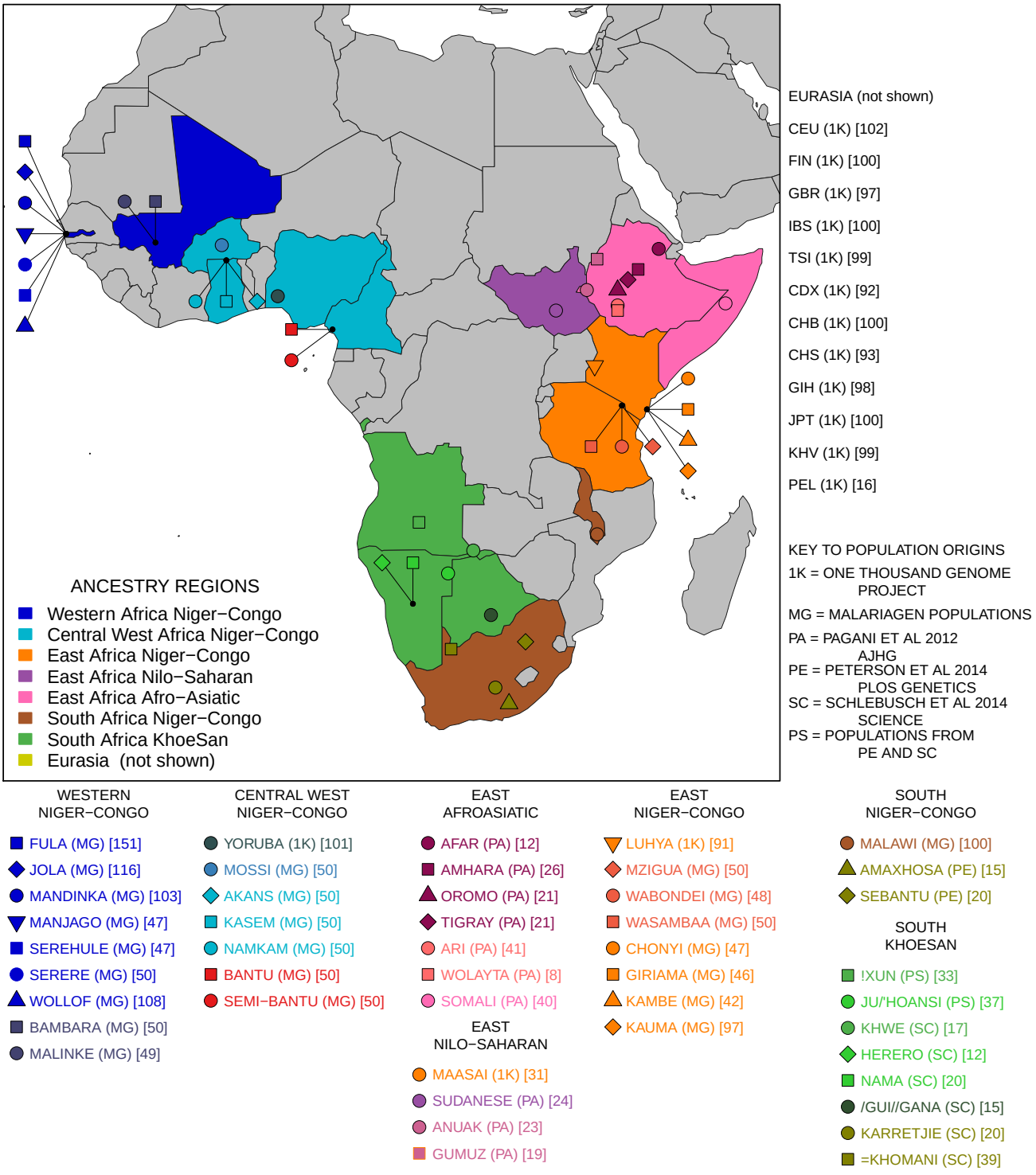
We calculate the MLE of beta by finding $\hat{\beta}$ that solves

$$\frac{1}{S}\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \frac{e^{\mu_i + \hat{\beta}}}{1 + e^{\mu_i + \hat{\beta}}}$$

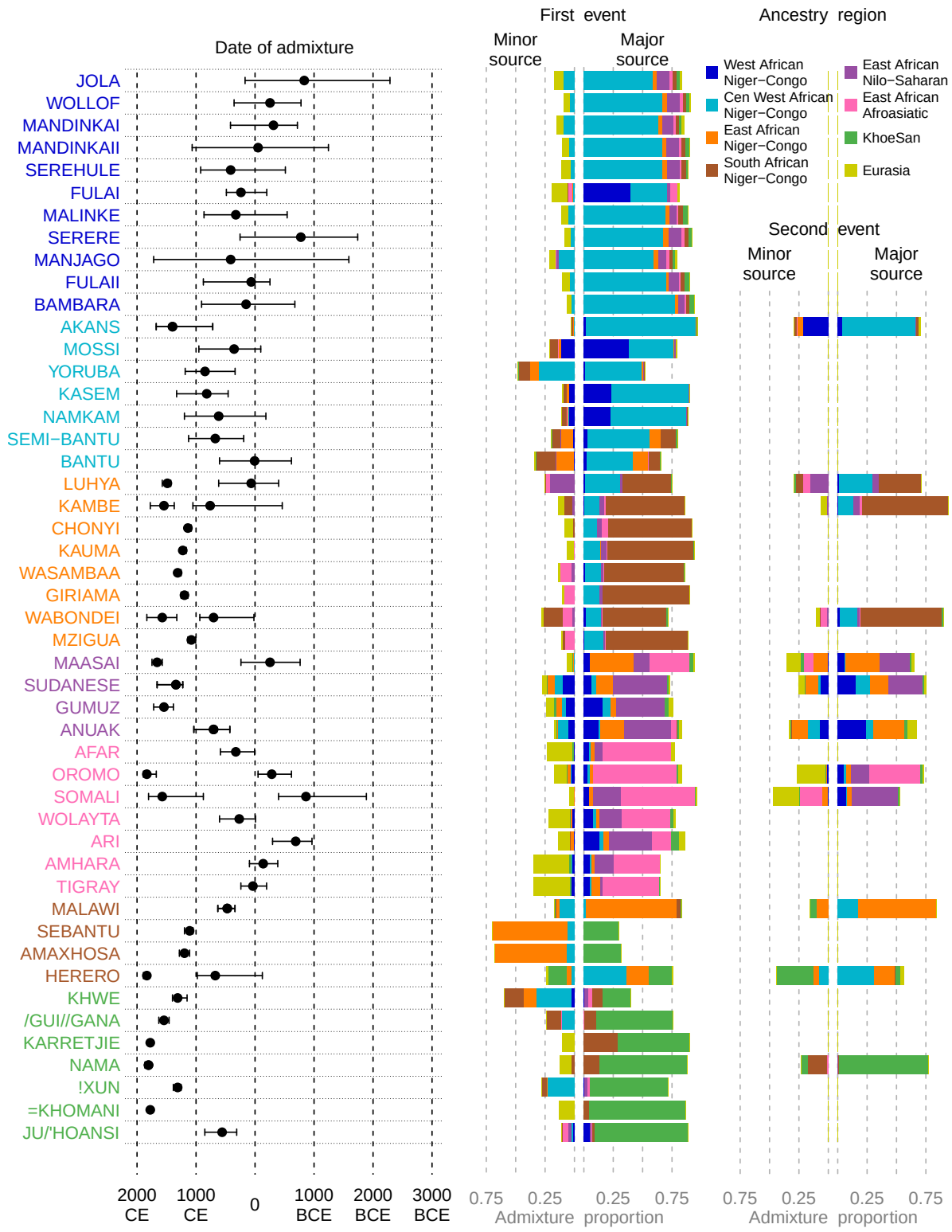
Then we use that MLE to calculate our likelihood ratio test statistic Λ and compare it to a χ_1^2 distribution to get our pvalue.

$$\Lambda = 2\left(\hat{\beta}\sum_{i=1}^{N} y_i + S\sum_{i=1}^{N} \log\left[\frac{1 + e^{\mu_i}}{1 + e^{\mu_i + \hat{\beta}}}\right]\right)$$

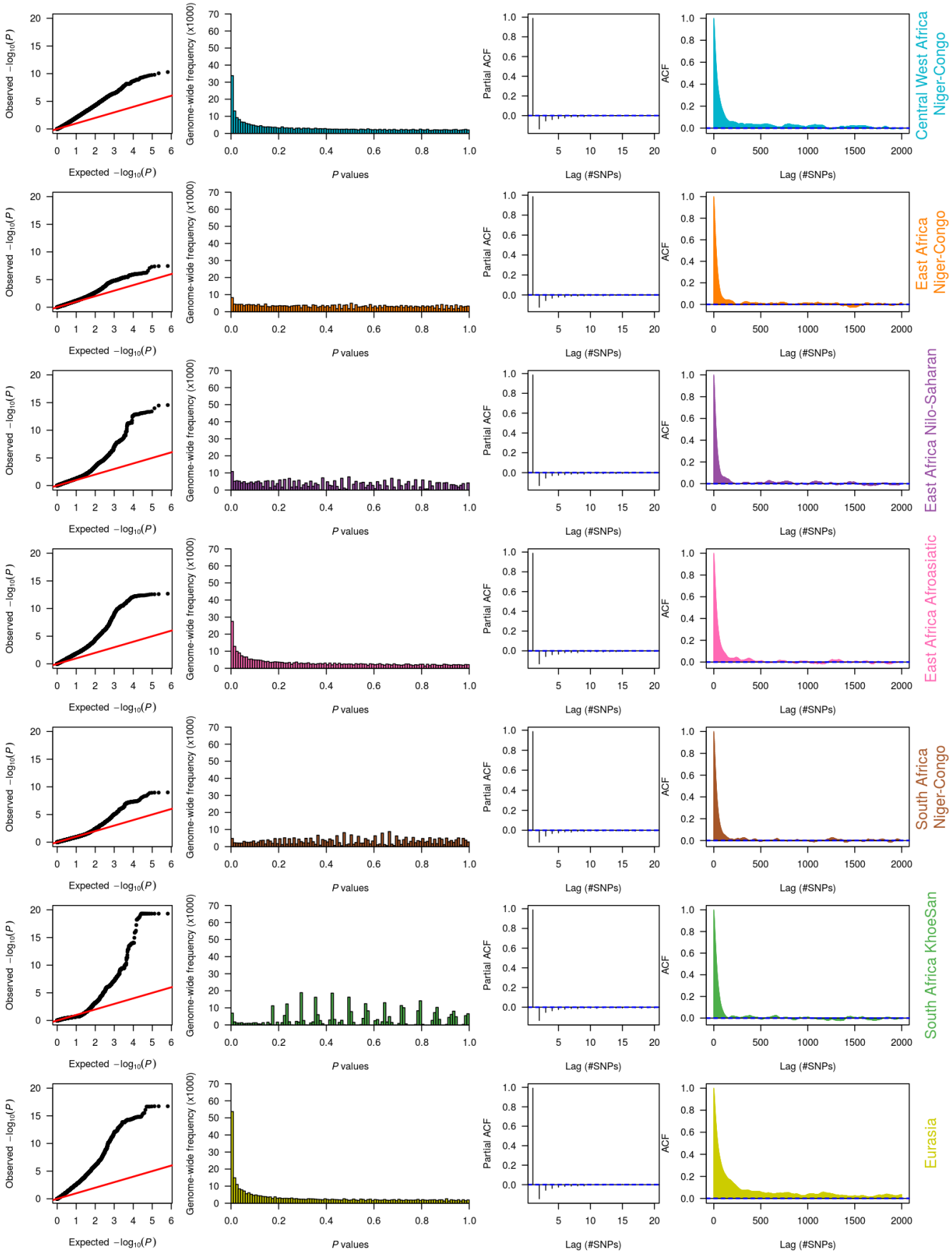
We then use chi-squared distribution to test if this statistic is in the tail.



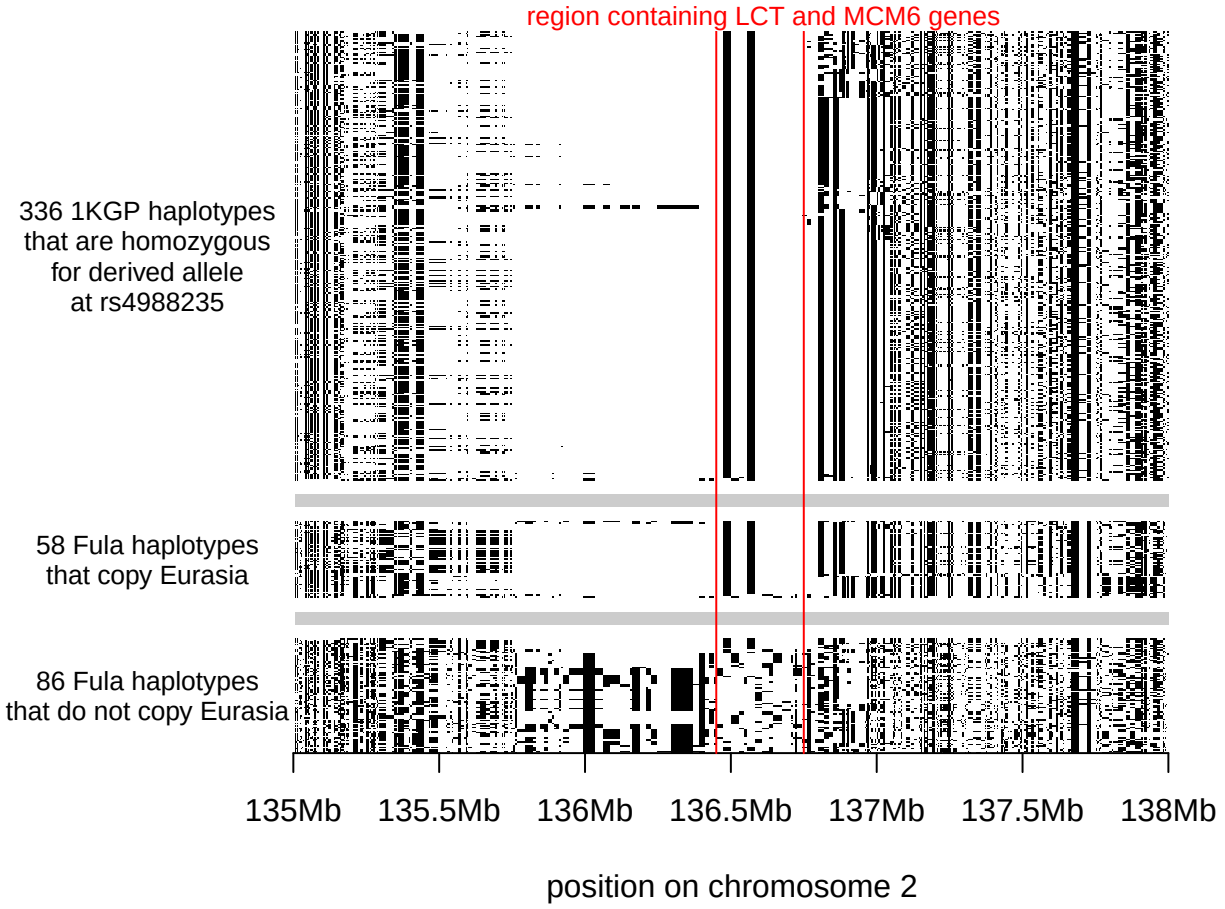
Supplementary Figure 1 | Description of populations used. Africa is divided into 7 ancestry regions based on ethnography, geography, and genetics.



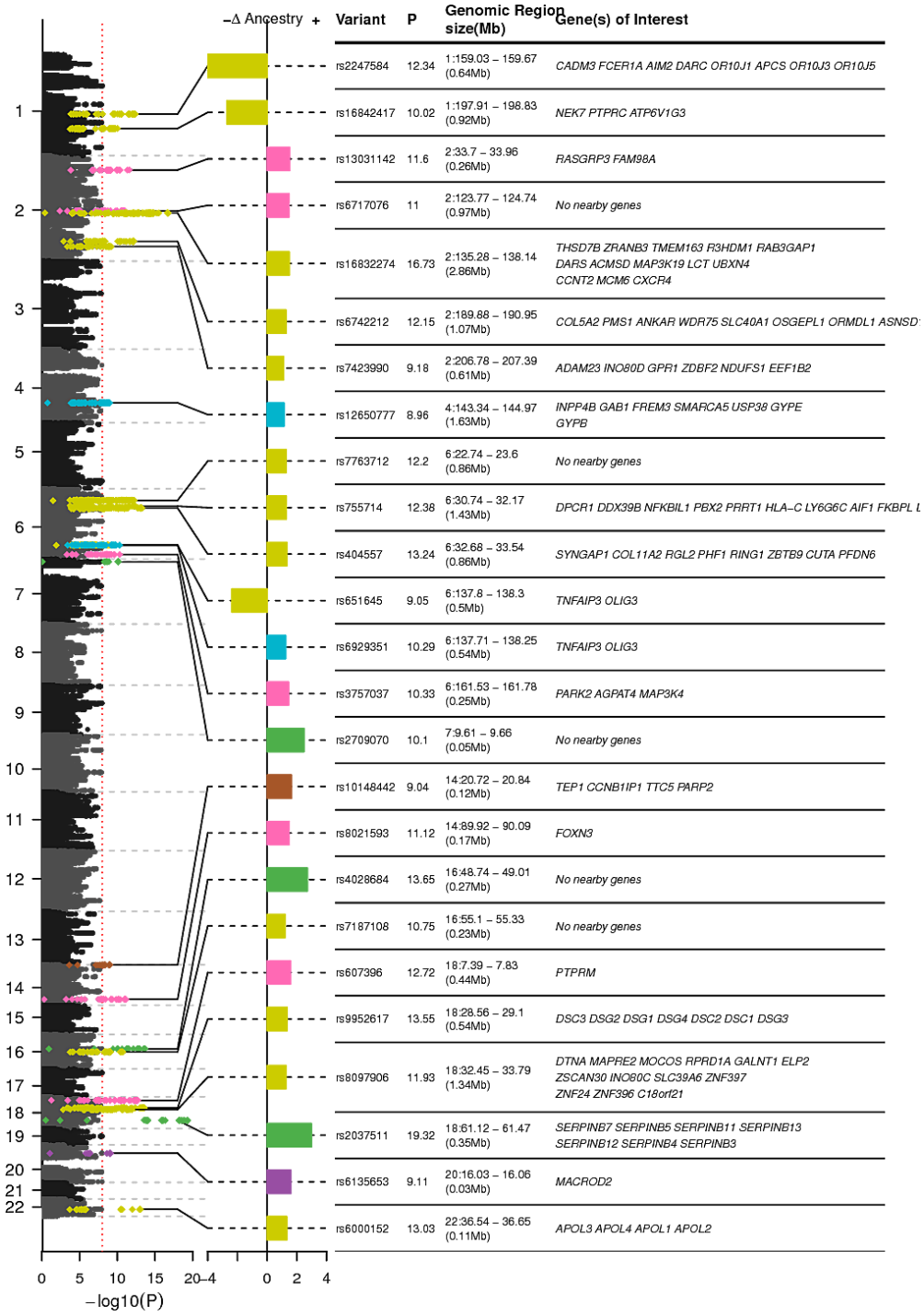
Supplementary Figure 2 | Overview of admixture events inferred in Sub-Saharan Africa. For reference, we show the main admixture events inferred previously by Busby et al, 2016[22]. For each of 48 African populations, we characterise the main admixture event inferred from chromosome painting using GLOBETROTTER[9] including the date(s) of admixture and the ancestry composition of the two admixture sources contributing to an admixture event. In 12 populations we infer two admixture events which happened at the same time if there is a single admixture date, or, in the case of two events, reflect the source composition of the earlier event. In all cases the sum of the minor and major admixture sources bars adds up to 1



Supplementary Figure 3 | Quantile-Quantile plots of the model statistics in the Fula show inflation. We show QQ plots for *P* values generated by the binomial likelihood model for each ancestry separately.

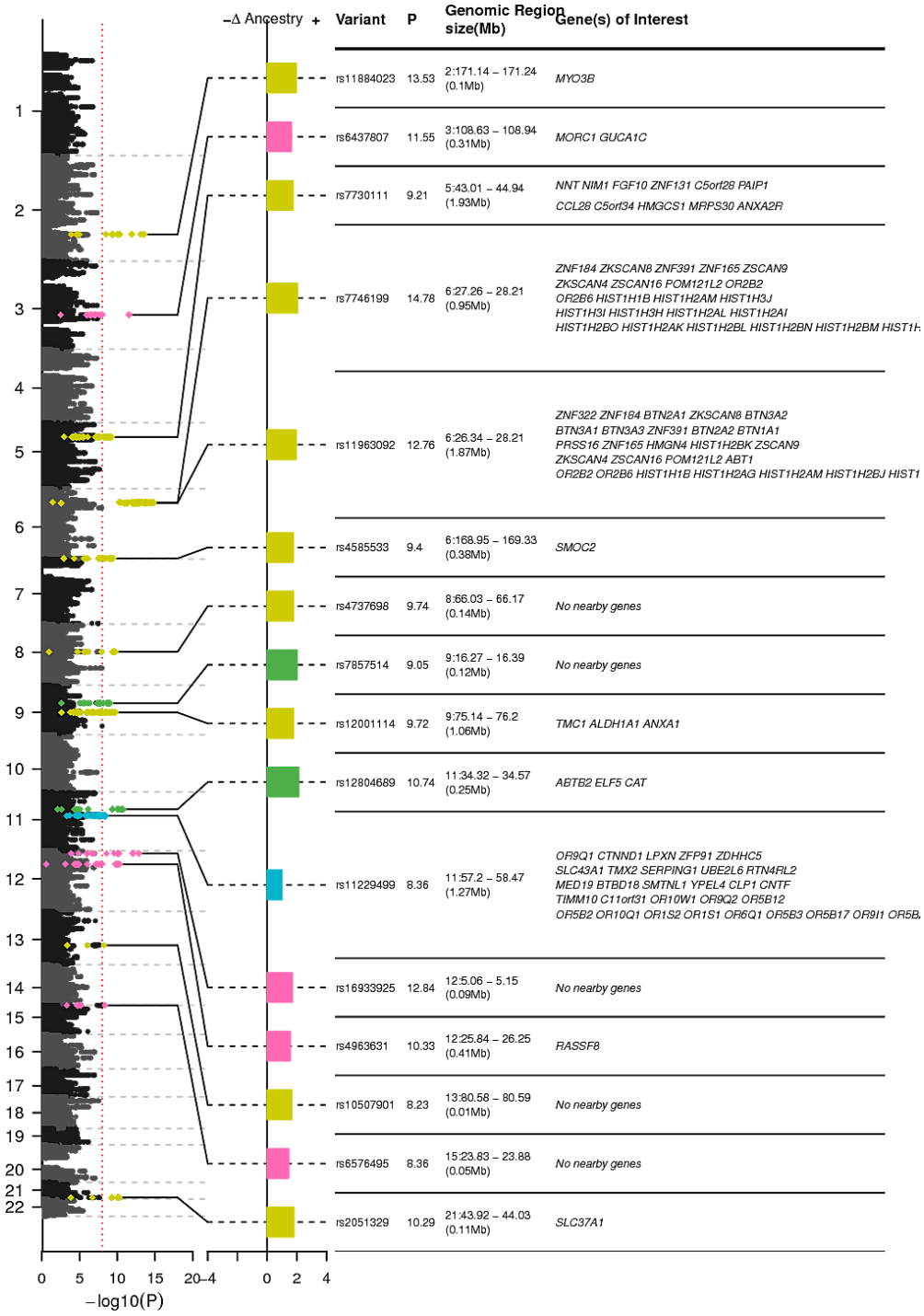


Supplementary Figure 4 | Comparison of Fula and 1KGP haplotypes at the *LCT/MCM6* region. rs4988235 is one of the two main SNPs present in the lactase persistence haplotype. The SNP is not present in our dataset. To ascertain whether the signal we identified in this region with our method is indeed due to the presence of the derived allelle at this SNP, we show the phased haplotypes of all 1KGP Eurasian individuals who are homozygous for the derived variant at rs4988235 as well as the haplotypes for those Fula who copy Eurasia donors across the *LCT* region, and those that do not. The same haplotype is shared across the known 1KGP lactase persistence allele carriers and the Fula individuals who copy Eurasia, suggesting that the signal is indeed related to the presence of the lactase persistence haplotype in the Fula.

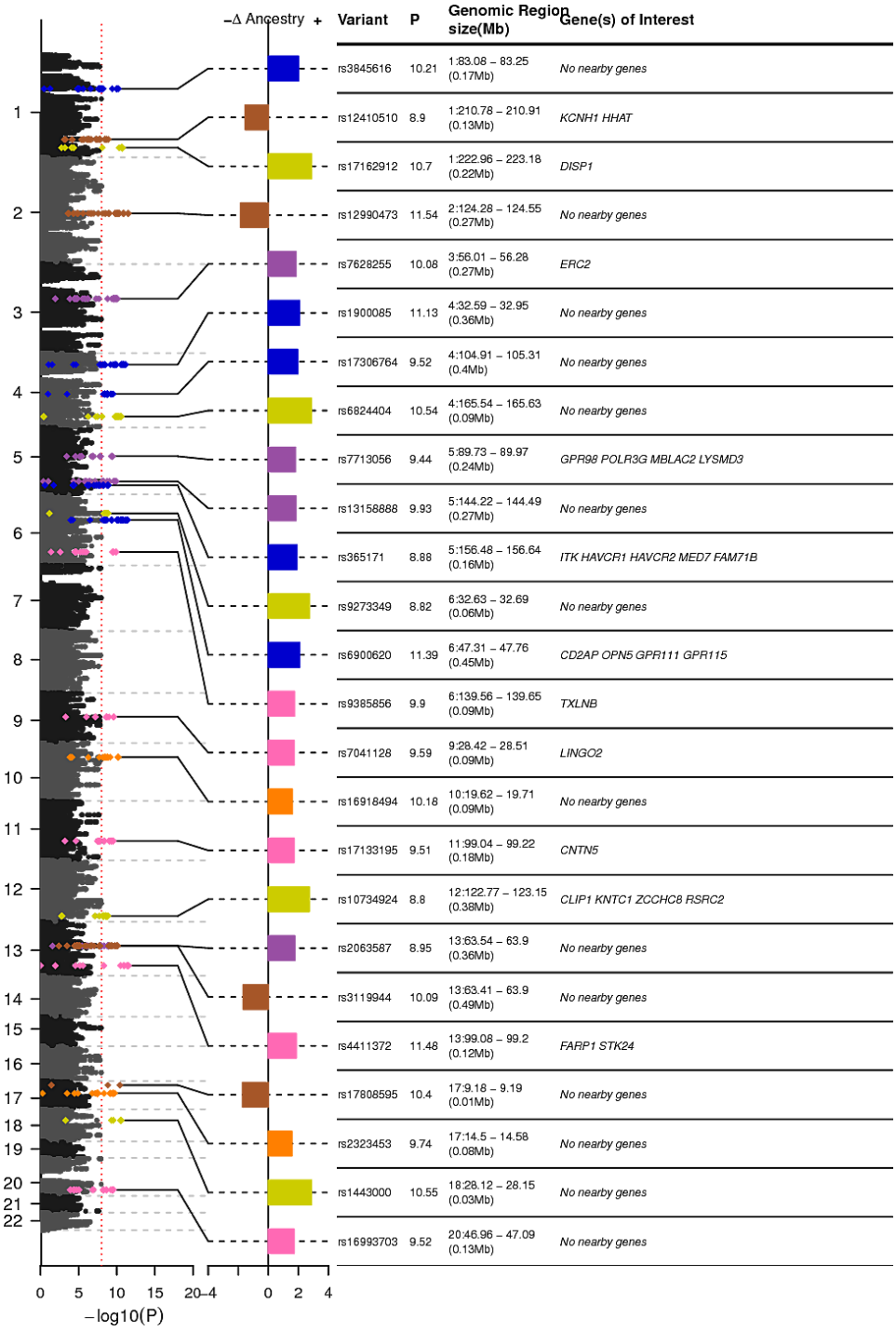


Variant	P	Genomic Region size(Mb)	Gene(s) of Interest
rs2247584	12.34	1:159.03 – 159.67 (0.64Mb)	*CADM3 FCER1A AIM2 DARC OR10J1 APCS OR10J3 OR10J5*
rs16842417	10.02	1:197.91 – 198.83 (0.92Mb)	*NEK7 PTPRC ATP6V1G3*
rs13031142	11.6	2:33.7 – 33.96 (0.26Mb)	*RASGRP3 FAM98A*
rs6717076	11	2:123.77 – 124.74 (0.97Mb)	*No nearby genes*
rs16832274	16.73	2:135.28 – 138.14 (2.86Mb)	*THSD7B ZRANB3 TMEM163 R3HDM1 RAB3GAP1 DARS ACMSD MAP3K19 LCT UBXN4 CCNT2 MCM6 CXCR4*
rs6742212	12.15	2:189.88 – 190.95 (1.07Mb)	*COL5A2 PMS1 ANKAR WDR75 SLC40A1 OSGEPL1 ORMDL1 ASNSD*
rs7423990	9.18	2:206.78 – 207.39 (0.61Mb)	*ADAM23 INO80D GPR1 ZDBF2 NDUFS1 EEF1B2*
rs12650777	8.96	4:143.34 – 144.97 (1.63Mb)	*INPP4B GAB1 FREM3 SMARCA5 USP38 GYPE GYPB*
rs7763712	12.2	6:22.74 – 23.6 (0.86Mb)	*No nearby genes*
rs755714	12.38	6:30.74 – 32.17 (1.43Mb)	*DPCR1 DDX39B NFKBIL1 PBX2 PRRT1 HLA–C LY6G6C AIF1 FKBPL*
rs404557	13.24	6:32.68 – 33.54 (0.86Mb)	*SYNGAP1 COL11A2 RGL2 PHF1 RING1 ZBTB9 CUTA PFDN6*
rs651645	9.05	6:137.8 – 138.3 (0.5Mb)	*TNFAIP3 OLIG3*
rs6929351	10.29	6:137.71 – 138.25 (0.54Mb)	*TNFAIP3 OLIG3*
rs3757037	10.33	6:161.53 – 161.78 (0.25Mb)	*PARK2 AGPAT4 MAP3K4*
rs2709070	10.1	7:9.61 – 9.66 (0.05Mb)	*No nearby genes*
rs10148442	9.04	14:20.72 – 20.84 (0.12Mb)	*TEP1 CCNB1IP1 TTC5 PARP2*
rs8021593	11.12	14:89.92 – 90.09 (0.17Mb)	*FOXN3*
rs4028684	13.65	16:48.74 – 49.01 (0.27Mb)	*No nearby genes*
rs7187108	10.75	16:55.1 – 55.33 (0.23Mb)	*No nearby genes*
rs607396	12.72	18:7.39 – 7.83 (0.44Mb)	*PTPRM*
rs9952617	13.55	18:28.56 – 29.1 (0.54Mb)	*DSC3 DSG2 DSG1 DSG4 DSC2 DSC1 DSG3*
rs8097906	11.93	18:32.45 – 33.79 (1.34Mb)	*DTNA MAPRE2 MOCOS RPRD1A GALNT1 ELP2 ZSCAN30 INO80C SLC39A6 ZNF397 ZNF24 ZNF396 C18orf21*
rs2037511	19.32	18:61.12 – 61.47 (0.35Mb)	*SERPINB7 SERPINB5 SERPINB11 SERPINB13 SERPINB12 SERPINB4 SERPINB3*
rs6135653	9.11	20:16.03 – 16.06 (0.03Mb)	*MACROD2*
rs6000152	13.03	22:36.54 – 36.65 (0.11Mb)	*APOL3 APOL4 APOL1 APOL2*

Supplementary Figure 5 | Manhattan plot showing ancestry deviations in the Fula. The left panel shows $-log_{10}\,P$ values based on the binomial likelihood test for each of the 7 non-self ancestries tested in the Fula. SNPs with $-log_{10}P$ values greater than 8 are highlighted with the colour of the ancestry that the deviation is seen. The central panel shows the change (Δ) in ancestry observed at significant SNPs.The right panel gives information about the top 'hit' SNP and the genes within the genomic region around the hit.

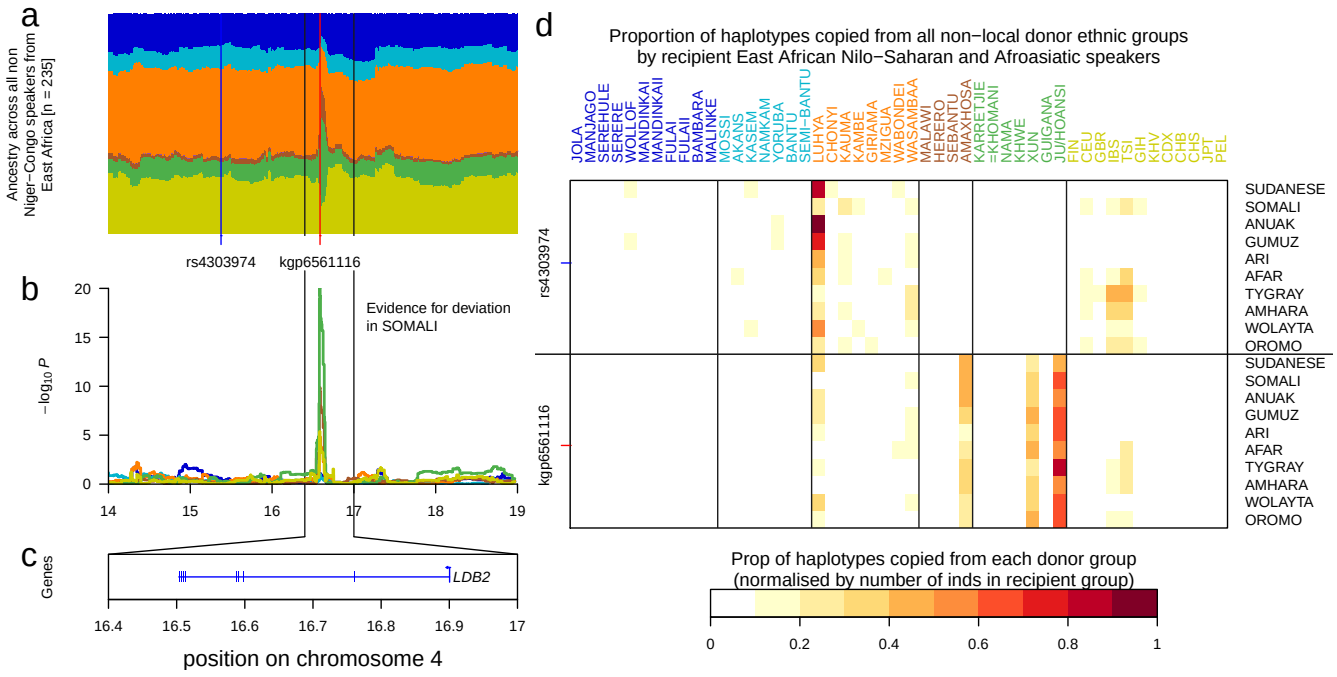


Supplementary Figure 6 | Manhattan plot showing ancestry deviations in the Jola. The left panel shows $-log_{10}\,P$ values based on the binomial likelihood test for each of the 7 non-self ancestries tested in the Fula. SNPs with P values greater than 8 are highlighted with the colour of the ancestry that the deviation is seen. The central panel shows the change (Δ) in ancestry observed at significant SNPs.The right panel gives information about the top 'hit' SNP and the genes within the genomic region around the hit.



Variant	P	Genomic Region size(Mb)	Gene(s) of Interest
rs3845616	10.21	1:83.08 – 83.25 (0.17Mb)	*No nearby genes*
rs12410510	8.9	1:210.78 – 210.91 (0.13Mb)	*KCNH1 HHAT*
rs17162912	10.7	1:222.96 – 223.18 (0.22Mb)	*DISP1*
rs12990473	11.54	2:124.28 – 124.55 (0.27Mb)	*No nearby genes*
rs7628255	10.08	3:56.01 – 56.28 (0.27Mb)	*ERC2*
rs1900085	11.13	4:32.59 – 32.95 (0.36Mb)	*No nearby genes*
rs17306764	9.52	4:104.91 – 105.31 (0.4Mb)	*No nearby genes*
rs6824404	10.54	4:165.54 – 165.63 (0.09Mb)	*No nearby genes*
rs7713056	9.44	5:89.73 – 89.97 (0.24Mb)	*GPR98 POLR3G MBLAC2 LYSMD3*
rs13158888	9.93	5:144.22 – 144.49 (0.27Mb)	*No nearby genes*
rs365171	8.88	5:156.48 – 156.64 (0.16Mb)	*ITK HAVCR1 HAVCR2 MED7 FAM71B*
rs9273349	8.82	6:32.63 – 32.69 (0.06Mb)	*No nearby genes*
rs6900620	11.39	6:47.31 – 47.76 (0.45Mb)	*CD2AP OPN5 GPR111 GPR115*
rs9385856	9.9	6:139.56 – 139.65 (0.09Mb)	*TXLNB*
rs7041128	9.59	9:28.42 – 28.51 (0.09Mb)	*LINGO2*
rs16918494	10.18	10:19.62 – 19.71 (0.09Mb)	*No nearby genes*
rs17133195	9.51	11:99.04 – 99.22 (0.18Mb)	*CNTN5*
rs10734924	8.8	12:122.77 – 123.15 (0.38Mb)	*CLIP1 KNTC1 ZCCHC8 RSRC2*
rs2063587	8.95	13:63.54 – 63.9 (0.36Mb)	*No nearby genes*
rs3119944	10.09	13:63.41 – 63.9 (0.49Mb)	*No nearby genes*
rs4411372	11.48	13:99.08 – 99.2 (0.12Mb)	*FARP1 STK24*
rs17808595	10.4	17:9.18 – 9.19 (0.01Mb)	*No nearby genes*
rs2323453	9.74	17:14.5 – 14.58 (0.08Mb)	*No nearby genes*
rs1443000	10.55	18:28.12 – 28.15 (0.03Mb)	*No nearby genes*
rs16993703	9.52	20:46.96 – 47.09 (0.13Mb)	*No nearby genes*

Supplementary Figure 7 | Manhattan plot showing ancestry deviations in the Ju/hoansi. The left panel shows $-log_{10}\,P$ values based on the binomial likelihood test for each of the 7 non-self ancestries tested in the Fula. SNPs with P values greater than 8 are highlighted with the colour of the ancestry that the deviation is seen. The central panel shows the change (Δ) in ancestry observed at significant SNPs.The right panel gives information about the top 'hit' SNP and the genes within the genomic region around the hit.



Supplementary Figure 8 | Increased Khoesan ancestry at *LDB2*. **a** local ancestry across 4700 painted chromosomes from East African Nilo-Saharan and Afroasiatic speakers at the 14-19Mb region of chromosome 4. **b** results of the binomial ancestry deviation test shows a significant increase in Khoesan (green) ancestry between 16.2 and 16.8Mb. **c** UCSC genes across this region. **d** comparison of donor copying at two SNPs on chromosome 4, rs43003974 is 1.5Mb upstream of the main signal and most copying at this SNP comes from the Luhya; conversely at kgp656116 most copying comes from three Southern African groups: The Ju/hoansi, Xun, and Amaxhosa.

Supplementary Table 1 | Evidence for ancestry deviations in African populations. Details of the ancestry deviation signals identified using the ancestry deviation model. We show the chromosome and position of the region where we observe a significant deviation in ancestry together with the rsID of the SNP with the largest deviation, the size and direction of the deviation, β and associated P value, the donor ancestry region of the deviation and the genes in the region of the signal for all significant signals of ancestry deviation.

Ethnic Gp	Chrom: position	rsID	β	P	Ancestry Region	Genes
JOLA	2: 171.14-171.24Mb	rs11884023	1.969	13.527	Eurasia	MYO3B
JOLA	3: 108.63-108.94Mb	rs6437807	1.623	11.553	East Africa Afroasiatic	MORC1; GUCA1C
JOLA	5: 43.01-44.94Mb	rs7730111	1.735	9.206	Eurasia	NNT; NIM1; FGF10; ZNF131; C5orf28; PAIP1; CCL28; C5orf34; HMGCS1; MRPS30; ANXA2R
JOLA	6: 27.26-28.21Mb	rs7746199	2.037	14.783	Eurasia	ZNF184; ZKSCAN8; ZNF391; ZNF165; ZSCAN9; ZKSCAN4; ZSCAN16; POM121L2; OR2B2; OR2B6; HIST1H1B; HIST1H2AM; HIST1H3J; HIST1H3I; HIST1H3H; HIST1H2AL; HIST1H2AI; HIST1H2BO; HIST1H2AK; HIST1H2BL; HIST1H2BN; HIST1H2BM; HIST1H2AJ; HIST1H4L; HIST1H4J; HIST1H4K

Continued on next page

Supplementary Table 1 Continued from previous page

Ethnic Gp	Chrom: position	rsID	β	P	Ancestry Region	Genes
JOLA	6: 26.34-28.21Mb	rs11963092	1.941	12.763	Eurasia	ZNF322; ZNF184; BTN2A1; ZKSCAN8; BTN3A2; BTN3A1; BTN3A3; ZNF391; BTN2A2; BTN1A1; PRSS16; ZNF165; HMGN4; HIST1H2BK; ZSCAN9; ZKSCAN4; ZSCAN16; POM121L2; ABT1; OR2B2; OR2B6; HIST1H1B; HIST1H2AG; HIST1H2AM; HIST1H2BJ; HIST1H2AH; HIST1H3J; HIST1H3I; HIST1H3H; HIST1H2AL; HIST1H2AI; HIST1H2BO; HIST1H2AK; HIST1H2BL; HIST1H2BN; HIST1H2BM; HIST1H2AJ; HIST1H4I; HIST1H4L; HIST1H4J; HIST1H4K
JOLA	6: 168.95-169.33Mb	rs4585533	1.750	9.404	Eurasia	SMOC2
JOLA	8: 66.03-66.17Mb	rs4737698	1.764	9.738	Eurasia	No nearby genes
JOLA	9: 16.27-16.39Mb	rs7857514	2.005	9.048	South Africa KhoeSan	No nearby genes
JOLA	9: 75.14-76.2Mb	rs12001114	1.762	9.724	Eurasia	TMC1; ALDH1A1; ANXA1
JOLA	11: 34.32-34.57Mb	rs12804689	2.121	10.740	South Africa KhoeSan	ABTB2; ELF5; CAT
JOLA	11: 57.2-58.47Mb	rs11229499	0.992	8.362	Central West Africa Niger-Congo	OR9Q1; CTNND1; LPXN; ZFP91; ZDHHC5; SLC43A1; TMX2; SERPING1; UBE2L6; RTN4RL2; MED19; BTBD18; SMTNL1; YPEL4; CLP1; CNTF; TIMM10; C11orf31; OR10W1; OR9Q2; OR5B12; OR5B2; OR10Q1; OR1S2; OR1S1; OR6Q1; OR5B3; OR5B17; OR9I1; OR5B21
JOLA	12: 5.06-5.15Mb	rs16933925	1.685	12.839	East Africa Afroasiatic	No nearby genes

Continued on next page

Supplementary Table 1 Continued from previous page

Ethnic Gp	Chrom: position	rsID	β	P	Ancestry Region	Genes
JOLA	12: 25.84-26.25Mb	rs4963631	1.556	10.326	East Africa Afroasiatic	RASSF8
JOLA	13: 80.58-80.59Mb	rs10507901	1.664	8.231	Eurasia	No nearby genes
JOLA	15: 23.83-23.88Mb	rs6576495	1.439	8.359	East Africa Afroasiatic	No nearby genes
JOLA	21: 43.92-44.03Mb	rs2051329	1.797	10.291	Eurasia	SLC37A1
WOLLOF	6: 27.14-28.21Mb	rs13194053	1.568	10.414	Eurasia	ZNF184; ZKSCAN8; ZNF391; PRSS16; ZNF165; ZSCAN9; ZKSCAN4; ZSCAN16; POM121L2; OR2B2; OR2B6; HIST1H1B; HIST1H2AM; HIST1H3J; HIST1H3I; HIST1H3H; HIST1H2AL; HIST1H2AI; HIST1H2BO; HIST1H2AK; HIST1H2BL; HIST1H2BN; HIST1H2BM; HIST1H2AJ; HIST1H4L; HIST1H4J; HIST1H4K
WOLLOF	6: 28.83-29.61Mb	rs9257319	1.500	9.255	Eurasia	OR12D3; MAS1L; OR5V1; OR12D2
FULAI	1: 110.31-110.68Mb	rs333958	2.322	8.081	South Africa KhoeSan	AHCYL1; STRIP1; CSF1; ALX3; UBL4B
FULAI	1: 159.03-159.67Mb	rs2247584	-4.152	12.336	Eurasia	CADM3; FCER1A; AIM2; DARC; OR10J1; APCS; OR10J3; OR10J5
FULAI	1: 160.13-160.47Mb	rs11582932	-2.100	8.225	Eurasia	COPA; DCAF8; SLAMF6; ATP1A4; VANGL2; NCSTN; CASQ1; PEA15; NHLH1
FULAI	1: 185.42-187.03Mb	rs10911836	1.070	8.440	Eurasia	HMCN1; PLA2G4A; TPR; C1orf27; PRG4; PDC; PTGS2; OCLM
FULAI	1: 197.91-198.83Mb	rs16842417	-2.706	10.022	Eurasia	NEK7; PTPRC; ATP6V1G3
FULAI	2: 33.7-33.96Mb	rs13031142	1.509	11.597	East Africa Afroasiatic	RASGRP3; FAM98A
FULAI	2: 123.77-124.74Mb	rs6717076	1.477	11.001	East Africa Afroasiatic	No nearby genes

Continued on next page

Supplementary Table 1 Continued from previous page

Ethnic Gp	Chrom: position	rsID	β	P	Ancestry Region	Genes
FULAI	2: 135.28-138.14Mb	rs16832274	1.489	16.730	Eurasia	THSD7B; ZRANB3; TMEM163; R3HDM1; RAB3GAP1; DARS; ACMSD; MAP3K19; LCT; UBXN4; CCNT2; MCM6; CXCR4
FULAI	2: 189.88-190.95Mb	rs6742212	1.278	12.152	Eurasia	COL5A2; PMS1; ANKAR; WDR75; SLC40A1; OSGEPL1; ORMDL1; ASNSD1; MSTN
FULAI	2: 193.21-195.2Mb	rs4850494	1.087	8.686	Eurasia	No nearby genes
FULAI	2: 206.78-207.39Mb	rs7423990	1.117	9.177	Eurasia	ADAM23; INO80D; GPR1; ZDBF2; NDUFS1; EEF1B2
FULAI	4: 143.34-144.97Mb	rs12650777	1.134	8.960	Central West Africa Niger-Congo	INPP4B; GAB1; FREM3; SMARCA5; USP38; GYPE; GYPB
FULAI	6: 22.15-22.66Mb	rs9466363	1.084	8.636	Eurasia	PRL; HDGFL1
FULAI	6: 22.74-23.6Mb	rs7763712	1.280	12.205	Eurasia	No nearby genes
FULAI	6: 30.74-32.17Mb	rs755714	1.288	12.375	Eurasia	DPCR1; DDX39B; NFKBIL1; PBX2; PRRT1; HLA-C; LY6G6C; AIF1; FKBPL; LY6G5B
FULAI	6: 32.68-33.54Mb	rs404557	1.331	13.243	Eurasia	SYNGAP1; COL11A2; RGL2; PHF1; RING1; ZBTB9; CUTA; PFDN6
FULAI	6: 33.56-33.84Mb	rs4304152	-1.004	8.198	Central West Africa Niger-Congo	ITPR3; IP6K3; LEMD2; MNF1; MLN
FULAI	6: 36.57-36.91Mb	rs16889125	1.321	8.421	East Africa Afroasiatic	CPNE5; C6orf89; RAB44; PPIL1; CDKN1A; SRSF3
FULAI	6: 137.8-138.3Mb	rs651645	-2.370	9.046	Eurasia	TNFAIP3; OLIG3
FULAI	6: 137.71-138.25Mb	rs6929351	1.243	10.286	Central West Africa Niger-Congo	TNFAIP3; OLIG3
FULAI	6: 161.53-161.78Mb	rs3757037	1.438	10.325	East Africa Afroasiatic	PARK2; AGPAT4; MAP3K4
FULAI	7: 9.61-9.66Mb	rs2709070	2.481	10.097	South Africa KhoeSan	No nearby genes
FULAI	7: 31.09-31.21Mb	rs2267740	1.529	8.141	East Africa Nilo-Saharan	ADCYAP1R1

Continued on next page

Supplementary Table 1 Continued from previous page

Ethnic Gp	Chrom: position	rsID	β	P	Ancestry Region	Genes
FULAI	8: 10.07-10.2Mb	rs10903321	1.126	8.859	Central West Africa Niger-Congo	MSRA
FULAI	9: 16.36-16.66Mb	rs4961485	-2.048	8.042	Eurasia	BNC2
FULAI	9: 18.75-18.78Mb	rs10118309	1.295	8.044	East Africa Afroasiatic	ADAMTSL1
FULAI	14: 20.72-20.84Mb	rs10148442	1.630	9.035	South Africa Niger-Congo	TEP1; CCNB1IP1; TTC5; PARP2
FULAI	14: 25.72-26Mb	rs1951147	1.097	8.870	Eurasia	No nearby genes
FULAI	14: 89.92-90.09Mb	rs8021593	1.482	11.122	East Africa Afroasiatic	FOXN3
FULAI	16: 48.74-49.01Mb	rs4028684	2.705	13.649	South Africa KhoeSan	No nearby genes
FULAI	16: 55.1-55.33Mb	rs7187108	1.204	10.747	Eurasia	No nearby genes
FULAI	17: 9.88-10.02Mb	rs2058173	2.335	8.202	South Africa KhoeSan	GAS7
FULAI	18: 7.39-7.83Mb	rs607396	1.566	12.717	East Africa Afroasiatic	PTPRM
FULAI	18: 24.76-25.31Mb	rs17642841	1.052	8.120	Eurasia	CHST9
FULAI	18: 28.56-29.1Mb	rs9952617	1.345	13.549	Eurasia	DSC3; DSG2; DSG1; DSG4; DSC2; DSC1; DSG3
FULAI	18: 32.45-33.79Mb	rs8097906	1.266	11.933	Eurasia	DTNA; MAPRE2; MOCOS; RPRD1A; GALNT1; ELP2; ZSCAN30; INO80C; SLC39A6; ZNF397; ZNF24; ZNF396; C18orf21
FULAI	18: 61.12-61.47Mb	rs2037511	2.980	19.316	South Africa KhoeSan	SERPINB7; SERPINB5; SERPINB11; SERPINB13; SERPINB12; SERPINB4; SERPINB3
FULAI	20: 16.03-16.06Mb	rs6135653	1.596	9.108	East Africa Nilo-Saharan	MACROD2
FULAI	20: 58.79-58.97Mb	rs7265604	-2.209	8.561	Eurasia	No nearby genes
FULAI	22: 36.54-36.65Mb	rs6000152	1.320	13.028	Eurasia	APOL3; APOL4; APOL1; APOL2
NAMKAM	7: 90.88-91.18Mb	rs12540728	3.570	8.667	Eurasia	FZD1

Continued on next page

Supplementary Table 1 Continued from previous page

Ethnic Gp	Chrom: position	rsID	β	P	Ancestry Region	Genes
LUHYA	7: 47.69-48.34Mb	rs7797887	1.340	8.314	South Africa KhoeSan	ABCA13; PKD1L1; SUN3; C7orf57; C7orf69; UPP1; HUS1; C7orf65
KAUMA	3: 162.39-164.36Mb	rs12491408	1.788	10.000	Eurasia	No nearby genes
KAUMA	7: 26.1-26.37Mb	rs17153154	1.233	8.104	East Africa Nilo-Saharan	SNX10; NFE2L3; CBX3; HNRNPA2B1
KAUMA	8: 98.37-98.92Mb	rs16895640	1.352	9.061	Western Africa Niger-Congo	MATN2; MTDH; LAPTM4B
GIRIAMA	10: 77.23-77.66Mb	rs10824308	1.680	8.630	South Africa KhoeSan	C10orf11
SUDANESE	4: 16.59-16.64Mb	kgp6561116	2.496	10.238	South Africa KhoeSan	LDB2
GUMUZ	4: 16.59-16.64Mb	rs3935821	2.405	10.562	South Africa KhoeSan	LDB2
GUMUZ	7: 146.92-147.13Mb	rs1496547	2.191	8.826	Eurasia	CNTNAP2
ARI	4: 10.28-10.42Mb	rs12510549	1.418	8.568	Western Africa Niger-Congo	No nearby genes
ARI	8: 32.29-32.4Mb	rs11782671	2.228	9.677	South Africa Niger-Congo	NRG1
ARI	12: 115.79-115.79Mb	kgp732488	1.408	8.227	South Africa KhoeSan	No nearby genes
AMHARA	4: 16.59-16.64Mb	rs7669580	2.341	11.803	South Africa KhoeSan	LDB2
XUN	4: 165.62-165.63Mb	rs10000041	2.919	9.294	Eurasia	No nearby genes
XUN	10: 92.3-92.44Mb	rs17097104	1.770	8.135	East Africa Afroasiatic	No nearby genes
XUN	12: 125.17-125.26Mb	rs4765602	3.276	14.085	Eurasia	No nearby genes
XUN	13: 84.32-84.56Mb	rs1576741	1.995	9.312	Western Africa Niger-Congo	SLITRK1
JU/HOANSI	1: 83.08-83.25Mb	rs3845616	2.008	10.212	Western Africa Niger-Congo	No nearby genes
JU/HOANSI	1: 210.78-210.91Mb	rs12410510	-1.531	8.905	South Africa Niger-Congo	KCNH1; HHAT
JU/HOANSI	1: 222.96-223.18Mb	rs17162912	2.885	10.698	Eurasia	DISP1
JU/HOANSI	2: 124.28-124.55Mb	rs12990473	-1.840	11.537	South Africa Niger-Congo	No nearby genes
JU/HOANSI	3: 56.01-56.28Mb	rs7628255	1.849	10.076	East Africa Nilo-Saharan	ERC2

Continued on next page

Supplementary Table 1 Continued from previous page

Ethnic Gp	Chrom: position	rsID	β	P	Ancestry Region	Genes
JU/HOANSI	3: 111.59-111.76Mb	rs6788215	-1.495	8.602	South Africa Niger-Congo	PHLDB2; TAGLN3; ABHD10
JU/HOANSI	3: 111.68-111.73Mb	rs7429121	1.596	8.061	East Africa Afroasiatic	PHLDB2; TAGLN3; ABHD10
JU/HOANSI	3: 132.95-133.16Mb	rs9876399	2.653	8.075	Eurasia	TMEM108; BFSP2
JU/HOANSI	4: 32.11-32.26Mb	rs10017595	1.471	8.569	East Africa Niger-Congo	No nearby genes
JU/HOANSI	4: 32.59-32.95Mb	rs1900085	2.076	11.133	Western Africa Niger-Congo	No nearby genes
JU/HOANSI	4: 104.91-105.31Mb	rs17306764	1.956	9.522	Western Africa Niger-Congo	No nearby genes
JU/HOANSI	4: 165.54-165.63Mb	rs6824404	2.869	10.543	Eurasia	No nearby genes
JU/HOANSI	5: 89.73-89.97Mb	rs7713056	1.798	9.444	East Africa Nilo-Saharan	GPR98; POLR3G; MBLAC2; LYSMD3
JU/HOANSI	5: 144.22-144.49Mb	rs13158888	1.835	9.928	East Africa Nilo-Saharan	No nearby genes
JU/HOANSI	5: 156.48-156.64Mb	rs365171	1.908	8.879	Western Africa Niger-Congo	ITK; HAVCR1; HAVCR2; MED7; FAM71B
JU/HOANSI	6: 32.63-32.69Mb	rs9273349	2.726	8.816	Eurasia	No nearby genes
JU/HOANSI	6: 47.31-47.76Mb	rs6900620	2.093	11.393	Western Africa Niger-Congo	CD2AP; OPN5; GPR111; GPR115
JU/HOANSI	6: 139.56-139.65Mb	rs9385856	1.742	9.903	East Africa Afroasiatic	TXLNB
JU/HOANSI	7: 9.69-9.83Mb	rs11974998	-1.455	8.226	South Africa Niger-Congo	No nearby genes
JU/HOANSI	8: 51.91-52.07Mb	rs7818008	1.460	8.428	East Africa Niger-Congo	No nearby genes
JU/HOANSI	9: 28.42-28.51Mb	rs7041128	1.718	9.588	East Africa Afroasiatic	LINGO2
JU/HOANSI	9: 86.94-86.97Mb	rs12346493	-1.498	8.612	South Africa Niger-Congo	SLC28A3
JU/HOANSI	10: 19.62-19.71Mb	rs16918494	1.599	10.180	East Africa Niger-Congo	No nearby genes
JU/HOANSI	10: 26.16-26.53Mb	rs16926664	-1.499	8.623	South Africa Niger-Congo	MYO3A; GAD2
JU/HOANSI	10: 90.57-90.67Mb	rs1837836	1.627	8.456	East Africa Afroasiatic	STAMBPL1; ANKRD22; LIPM
JU/HOANSI	11: 99.04-99.22Mb	rs17133195	1.713	9.515	East Africa Afroasiatic	CNTN5
JU/HOANSI	12: 97.16-97.3Mb	rs12820022	2.666	8.198	Eurasia	No nearby genes

Continued on next page

Supplementary Table 1 Continued from previous page

Ethnic Gp	Chrom: position	rsID	β	P	Ancestry Region	Genes
JU/HOANSI	12: 102.32-102.34Mb	rs4764852	1.473	8.591	East Africa Niger-Congo	DRAM1
JU/HOANSI	12: 122.77-123.15Mb	rs10734924	2.722	8.799	Eurasia	CLIP1; KNTC1; ZCCHC8; RSRC2
JU/HOANSI	13: 63.54-63.9Mb	rs2063587	1.758	8.951	East Africa Nilo-Saharan	No nearby genes
JU/HOANSI	13: 63.41-63.9Mb	rs3119944	-1.669	10.090	South Africa Niger-Congo	No nearby genes
JU/HOANSI	13: 99.08-99.2Mb	rs4411372	1.855	11.483	East Africa Afroasiatic	FARP1; STK24
JU/HOANSI	16: 30.93-31.38Mb	rs12924903	1.641	8.618	East Africa Afroasiatic	ITGAM; ITGAX; SETD1A; STX1B; FUS; KAT8; ZNF668; PRSS36; TRIM72; ZNF646; STX4; ORAI3; PRSS53; BCKDK; PRSS8; VKORC1; HSD3B7; PYCARD; PYDC1
JU/HOANSI	16: 77.69-77.74Mb	rs10514426	2.676	8.311	Eurasia	No nearby genes
JU/HOANSI	17: 9.18-9.19Mb	rs17808595	-1.704	10.399	South Africa Niger-Congo	No nearby genes
JU/HOANSI	17: 14.5-14.58Mb	rs2323453	1.565	9.738	East Africa Niger-Congo	No nearby genes
JU/HOANSI	18: 28.12-28.15Mb	rs1443000	2.870	10.548	Eurasia	No nearby genes
JU/HOANSI	18: 66.28-66.4Mb	rs12327516	-1.448	8.187	South Africa Niger-Congo	CCDC102B; TMX3
JU/HOANSI	20: 46.96-47.09Mb	rs16993703	1.713	9.517	East Africa Afroasiatic	No nearby genes