Archaic Hominin Introgression in Africa Contributes to Functional Salivary *MUC7* Genetic Variation

Duo Xu,[1] Pavlos Pavlidis,[2] Recep Ozgur Taskent,[1] Nikolaos Alachiotis,[3] Colin Flanagan,[1] Michael DeGiorgio,[4] Ran Blekhman,[5] Stefan Ruhl,*[,6] and Omer Gokcumen*[,1]

[1]Department of Biological Sciences, University at Buffalo, The State University of New York, Buffalo, NY

[2]Institute of Molecular Biology and Biotechnology (IMBB), Foundation for Research and Technology – Hellas, Heraklion, Crete, Greece

[3]Institute of Computer Science (ICS), Foundation for Research and Technology – Hellas, Heraklion, Crete, Greece

[4]Department of Biology and the Institute for CyberScience, Pennsylvania State University, University Park, PA

[5]Department of Genetics, Cell Biology, and Development, University of Minnesota, Twin Cities, MN

[6]Department of Oral Biology, School of Dental Medicine, University at Buffalo, The State University of New York, Buffalo, NY

*__Corresponding authors:__ E-mails: shruhl@buffalo.edu; omergokc@buffalo.edu.

Associate editor: Rasmus Nielsen

Abstract

One of the most abundant proteins in human saliva, mucin-7, is encoded by the *MUC7* gene, which harbors copy number variable subexonic repeats (PTS-repeats) that affect the size and glycosylation potential of this protein. We recently documented the adaptive evolution of *MUC7* subexonic copy number variation among primates. Yet, the evolution of *MUC7* genetic variation in humans remained unexplored. Here, we found that PTS-repeat copy number variation has evolved recurrently in the human lineage, thereby generating multiple haplotypic backgrounds carrying five or six PTS-repeat copy number alleles. Contrary to previous studies, we found no associations between the copy number of PTS-repeats and protection against asthma. Instead, we revealed a significant association of *MUC7* haplotypic variation with the composition of the oral microbiome. Furthermore, based on in-depth simulations, we conclude that a divergent *MUC7* haplotype likely originated in an unknown African hominin population and introgressed into ancestors of modern Africans.

Key words: **CNV, structural variation, human evolution, recurrent mutation, ABC simulation, saliva, mucin, microbiome.**

Introduction

MUC7 encodes one of the most abundant proteins in human saliva (Biesbrock et al. 1997). Mucin-7 is a small soluble mucin protein which, like other mucins, harbors subexonic repeat sequences rich in the amino acids proline, threonine, and serine (PTS repeats; Naganagowda et al. 1999). PTS repeats are primary targets for O-glycosylation of the protein (Bennett et al. 2012). Unlike the larger soluble gel-forming mucins, where dense O-glycosylation aids in lubricating mucosal surfaces and protecting them from physical, chemical, or microbial insult (Hollingsworth and Swanson 2004; Frenkel and Ribbeck 2015), the O-glycans of MUC7 are primary targets for a great variety of commensal and pathogenic microorganisms (Smith and Bobek 2001; Takamatsu et al. 2006; Heo et al. 2013; Thamadilok et al. 2016). *MUC7* also bears no genetic homology to any of the other mucins (Dekker et al. 2002). It originated in the ancestor of placental mammals, making it one of the youngest members of the mucin functional family of proteins (Xu et al. 2016).

The numbers of PTS repeats in human *MUC7* vary from 5 to 6 haploid copies (Biesbrock et al. 1997) (fig. 1), but origins and evolution of the haplotypes harboring these different copy number variants are unknown. Previous locus-specific studies in three independent cohorts suggested that the less-common five-copy allele is associated with protection against asthma (Kirkbride et al. 2001; Rousseau et al. 2006; Watson et al. 2009). However, the association of this locus with asthma was not confirmed in genome wide association studies (GWAS) in larger cohorts (Torgerson et al. 2011). Traditional GWAS assess copy number variation associations only through interrogating "tag" single nucleotide variants that are in high linkage disequilibrium with haplotypes carrying particular copy number variations. As such, it is plausible that the genotyping platforms used in those GWAS studies were unable to accurately "tag" the copy number variation in *MUC7* alleles (Mills et al. 2011; Sudmant et al. 2015). Thus, it remains unresolved whether *MUC7* genetic variation is indeed associated with protection against asthma.

Here, to better address the question of disease association of *MUC7* PTS-repeats, we resolved the haplotype architecture of the locus and investigated genetic variation of *MUC7* in humans from a more fundamental, evolutionary perspective. We already reported that the copy number of PTS-repeats evolved recurrently in different primate lineages, leading to an unusually high variation of MUC7 proteins across primate species (Xu et al. 2016). Despite this variation in PTS repeat

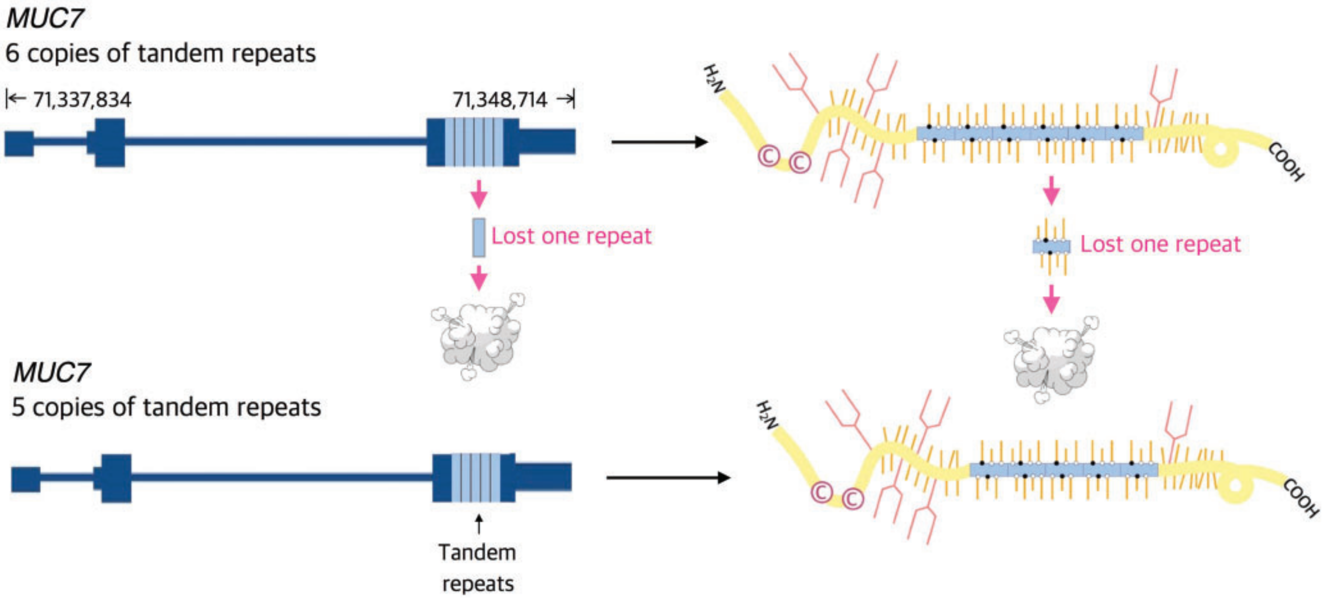

Fig. 1. The genetic and protein structure of MUC7. Schematic representation of the gene structure (left) and the coded protein (right) of two common *MUC7* haplotypes with five and six PTS-repeats on the top and bottom panel, respectively. On the left panel, the thin bars show introns, thicker bars show the untranslated regions (utr), and the thickest bars indicate coding exons. On the third exon (second coding exon), individual PTS-repeats are shown in light blue bars. The GRCh37/hg19 locations for the genes are indicated on the top left panel, which shows the reference allele. On the right panel, the predicted protein structure is shown based on previous studies (Gururaja et al. 1998). The yellow line indicates the non-repeated parts of the protein and the blue line indicates where the PTS-repeats are. Attached N-glycans are symbolized by orange, fork-like extensions. O-glycans are symbolized by shorter yellow sticks.

number, the O-glycosylation potential for individual PTS-repeats remained conserved. We further found that lineage-specific adaptive forces shaped the copy number variation of PTS repeats in primate species, including humans. On the basis of these results, we hypothesized that adaptive pressures likely involved the interaction of MUC7 with the oral microbiome or systemic pathogens occasionally traversing the mouth-saliva environment in primates. We now focus our investigation on how *MUC7* variation has evolved within the human lineage and how it may affect disease susceptibility.

Results

Copy Number Variation of *MUC7* PTS Repeats Evolved Recurrently in Humans

Discovering and genotyping the copy number of short repeat segments using short read sequences is challenging due to alignment and assembly errors (Mills et al. 2011; Zhao et al. 2016). Indeed, the 1,000 Genomes Project data did not capture the copy number variation affecting *MUC7* (Sudmant et al. 2015). To genotype the full range of PTS-repeat copy numbers, we directly genotyped 251 randomly selected samples from European, African, and Asian populations already sequenced by the 1,000 Genomes Project (supplementary table S1, Supplementary Material online). Then, to understand the haplotypic variation of *MUC7* subexonic copy number alleles, we constructed a maximum likelihood tree using the nonrepeated sections of *MUC7* from 5,008 phased haplotypes out of the 1,000 Genomes Project Phase 3 (fig. 2*a*, supplementary fig. S1, Supplementary Material online). The haplotypes fell into eight divergent haplogroup clusters (MUC7*A-H; fig. 2*b*). When we superimposed the number of PTS-repeat copies and the corresponding haplotypes constructed by single nucleotide variants, we observed that two haplogroups (groups G and E) harbor 5 PTS-repeat alleles, whereas the other groups harbor 6 PTS-repeat alleles (supplementary table S1, Supplementary Material online). This suggests that *MUC7* PTS-repeat copy number variation evolved independently in at least two different branches of the human phylogenetic tree, indicating recurrent evolution of these variants.

To test whether recurrence indeed shaped *MUC7*'s genetic variation, we directly sequenced five-copy haplotypes from haplogroups G and E, as well as six-copy haplotypes. To do this, we amplified the PTS-repeat region in heterozygote individuals, separated five- and six-repeat haplotypes from each other by DNA gel electrophoresis and excised bands containing the five- and six-copy haplotypes from the gel for sequencing. We discovered that the repeat sequences of haplogroup G were distinct from haplogroup E, even though they both contain five repeat copies (fig. 2*c*). Specifically, compared with the human reference genomes, which all have the six PTS-repeat allele haplotype, each repeat has distinctive single nucleotide differences from each other. As such, it was possible to distinguish individual repeats. We found that haplogroup G is missing the fourth PTS-repeat, whereas haplogroup E is missing the fifth PTS-repeat, counted from the 5′ end. Assuming that the five-copy PTS-repeat was ancestral as we suggested in our earlier work (Xu et al. 2016), our results suggest at least three independent mutational events: 1) the duplication event creating the six-copy PTS-repeat must have occurred before haplogroup E diverged from the other human *MUC7* haplotypes; 2) haplogroup G has lost the

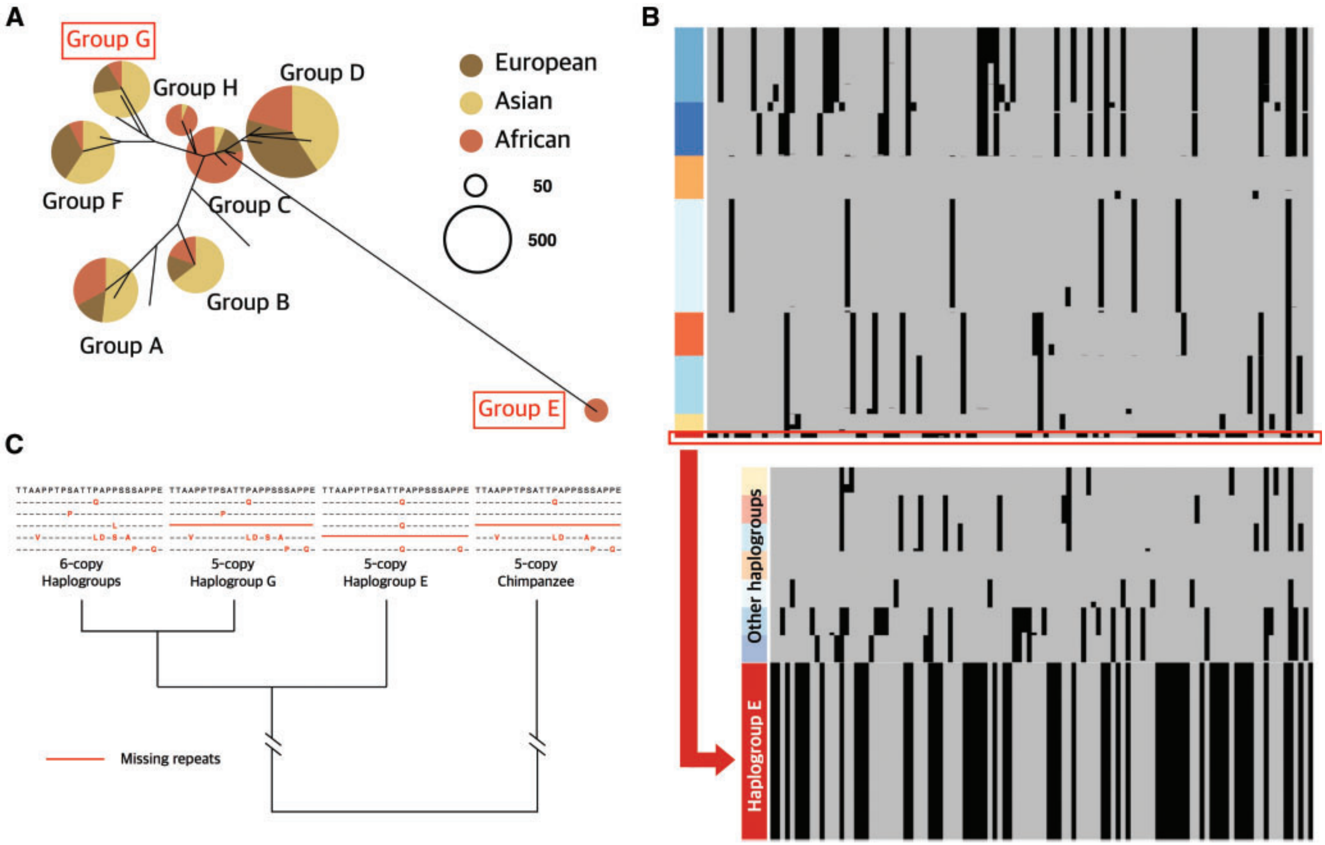

Fig. 2. *MUC7* genetic variation in humans. (*a*) Simplified single nucleotide variation-based phylogenetic tree showing the major haplotype groups (haplogroups). The size of each bubble is correlated with the respective allele frequency in each haplogroup. The pie-charts within each haplogroup indicate the frequency of its occurrence within continental populations according to data from the 1,000 Genomes Project Phase 3 (sample names can be found in supplementary table S1, Supplementary Material online). "Admixed" populations are not shown in this tree. Haplogroups carrying five copy PTS-repeats are highlighted in red boxes; (*b*) haplogroup clustering. The upper panel shows the k-means clustering of haplotypes of the *MUC7* gene, using single nucleotide variants with frequency $\geq$ 1% from the 1,000 Genomes Project (Phase 3). Each column indicates a single nucleotide variant in *MUC7* ordered by their location along the chromosome 4. Each row indicates a haplotype from the 1,000 Genomes Project. Black color indicates single nucleotide variants that differ from the reference genome, whereas gray indicates variants that have the reference genome allele. The small cluster on the bottom, circled in red, shows haplogroup E. In the lower, expanded panel, we show the clustering of haplogroup E cluster and 10 randomly chosen haplotypes from each of the other clusters for clarity; (*c*) *MUC7* PTS-repeat sequences in diverse human haplotypes as determined by Sanger sequencing and by the reference sequence for chimpanzees. Missing repeats are indicated by a red line. The protein sequence of the first repeat which remains the same in all haplotypic backgrounds is shown on top. The other PTS repeats are listed below (second to sixth repeats from the top). Single amino acid differences to the first repeat sequence are highlighted with red letters.

fourth copy; and 3) haplogroup E lost the fifth copy. Independent of the exact scenario, it is clear that two different five-copy PTS-repeat alleles are present in two independent haplotypic backgrounds, which strongly suggests recurrent formation of these alleles and a high mutation rate in this locus. Overall, we conclude that the contemporary variation of *MUC7* PTS-repeat copy number has evolved through multiple independent mutational events happening since the divergence of humans and chimpanzees.

Imputation of *MUC7* PTS-Repeat Haplotypes Allows Reevaluation of the Prior Association with Asthma Susceptibility

Accurate imputation of copy number variants is essential to incorporate these variants to genome-wide association studies conducted by single nucleotide genotyping platforms. As previously noted, the *MUC7* PTS-repeat copy number was associated with asthma susceptibility in three independent locus-specific association studies (Kirkbride et al. 2001; Rousseau et al. 2006; Watson et al. 2009). However, due to the recurrent evolution of *MUC7* PTS-repeat copy number variation, the individual single nucleotide variants do not accurately predict the copy number state of PTS-repeats. For example, the haplogroup E, which we showed to carry five MUC7 PTS-repeats, was not tagged by GWAS. As such, GWAS do not have the power to accurately tag *MUC7* PTS-repeats. To address this, we identified multiple single nucleotide variants that tag the two independently evolved five PTS-repeat alleles with very high accuracy in all populations tested (supplementary table S1, Supplementary Material online). Combining the allele frequencies of these tag single nucleotide variants would theoretically allow evaluation of PTS-repeat copy number in existing single nucleotide variant based genotyping studies. However, when we attempted to use these single nucleotide variants to impute the PTS-repeat copy numbers in multiple GWAS studies for asthma (Anon

1999; Boushey et al. 2005; Sorkness et al. 2007; Moore et al. 2010), we realized that some of the tag variants were not included in the genotyping platform that was used in these studies. To address this shortcoming, we used only single nucleotide variants that were used in the GWAS study to predict the PTS-repeat copy number states of 1,000 Genomes data. This approach allowed us to find combinations of single nucleotide variants genotyped in the GWAS platform to accurately predict the copy number status of *MUC7* PTS-repeats (supplementary fig. S2, Supplementary Material online). A similar approach was recently used for imputing structural variants within the haptoglobin locus (Boettger et al. 2016). This analysis showed that it is possible to tag the PTS-repeat copy number state (and the haplogroups) of *MUC7* using only the data available in GWAS studies with high accuracy (99.8%, supplementary table S2, Supplementary Material online). Using this methodology, we were unable to replicate the association between *MUC7* PTS-repeat copy number and asthma prevalence (nominal, locus-specific *P*-value = 0.2402 for African Americans, *P*-value = 0.1507 for European descendants, Cochran–Armitage test for trend test; supplementary table S2, Supplementary Material online). Also, the directionality of five-copy frequency in cases and controls are opposite in Africans and European cohorts, where African controls having a higher frequency of five-copy alleles as compared with the cases (supplementary fig. S2, Supplementary Material online). It is plausible that there is an indirect, weak correlation that depends on the external factors (e.g., the abundance of allergens in the environment). A more parsimonious explanation however is that the previous, locus-specific studies found spurious associations between *MUC7* PTS-repeat copy number and asthma prevalence.

MUC7 Genetic Variation in Humans Is Associated with Oral Microbiome Composition

MUC7 is known to be recognized and bound by bacteria residing in the oral cavity as well as by systemic pathogens that occasionally traverse the mouth environment (Walz et al. 2009; Heo et al. 2013; Thamadilok et al. 2016). Thus, we hypothesized that genetic variation in *MUC7* could affect bacterial colonization in the oral cavity. To test this hypothesis, we conducted a locus-specific association study, by analyzing genetic variation within 50 kb upstream and 50 kb downstream of *MUC7* from 93 humans and correlated it with the microbial composition in fifteen body sites of the same individuals (Human Microbiome Project Consortium 2012; Blekhman et al. 2015). From all these 15 different body sites ranging from stool to skin, we found significant correlations only in the oral cavity (fig. 3*a*, supplementary table S3, Supplementary Material online; *P*-value $\leq 9.66 \times 10^{-5}$ in each of the oral cavity sites). More specifically, we found 20 significant associations between genetic variants and microbial taxon abundance in the anterior nares, palatine tonsils, saliva, supragingival plaque, and tongue dorsum. For example, we highlighted the strongest signal for the association between single nucleotide variant rs12498483 and the abundance of the genus *Eubacterium* in the palatine tonsil (fig. 3*b*). Other associations include bacterial genera, such as *Lautropia* and *Neisseria*, which are known colonizers of mucosal surfaces and the oral cavity (Rossmann et al. 1998; Mager et al. 2003).

To understand haplotype-level variation that defines these genetic-microbial associations around the *MUC7* locus, we investigated linkage disequilibrium patterns among the variants significantly associated with microbiome composition. We found one haplotype, defined by rs12498483, that fully includes *MUC7* genetic variation (supplementary fig. S3, Supplementary Material online), and is made up of variants that broadly indicate samples that do not belong to haplogroups A, B, and C (fig. 2*a*). The haplotype is very common (>30% in most populations), reaching close to 50% in African populations. It is linked with *Eubacterium*, which has been reported to be a prominent and highly diverse genus in the oral cavity (Downes et al. 2001). Even though the host genotype data that we use in our microbiome association analysis comprise a limited number of variants, we were able to further identify dozens of single nucleotide variants that belong to this haplotype using 1,000 Genomes data set (supplementary fig. S3, Supplementary Material online). These include putatively functional variants, including seven that affect coding sequence (supplementary table S3, Supplementary Material online).

Our microbiome analysis primarily involved individuals with Eurasian ancestry. In addition, the bacterial associations are at the genus and occasionally class level, limiting the routes to direct functional testing. As such, our results are by no means definitive, and an adaptive impact of *MUC7*-microbiome interactions has yet to be resolved. However, our results suggest that the genetic variation within *MUC7* influences the oral microbiome. These results, together with the fact that microbiome composition varies across human populations (Yatsunenko et al. 2012; Morton et al. 2015; Gomez et al. 2016), motivate the need to design more focused studies to tackle specific gene–microbe interactions in diverse human groups and generate novel hypotheses with regard to the adaptive role of *MUC7* genetic variation.

Haplogroup E Is Unusually Divergent

During our analyses, we noticed an apparently divergent haplotype group, which is also one of the haplotypes carrying five PTS-repeat allele (haplogroup E, fig. 2*a* and *b*). We asked whether this haplotype is unusually divergent as compared with genome-wide expectations. To investigate this, we devised a statistical test inspired by Wall's B (Wall 1999), which measures the maximum number of SNPs in a locus that are in perfect LD ($r^2 = 1.0$) between themselves. Similar to Wall's B, our statistical method obtains high values if nonoverlapping subgroups of the sample are sufficiently distinct genetically. Wall's B measures the number of single nucleotide variants that are in perfect LD with each other and are subsequent (i.e., no other variant in between). Our statistical approach is different from Wall's B in that it measures the maximum number of single nucleotide variants that are in perfect LD with each other within a given locus independent of their location within that locus. We also normalize this measure by the total

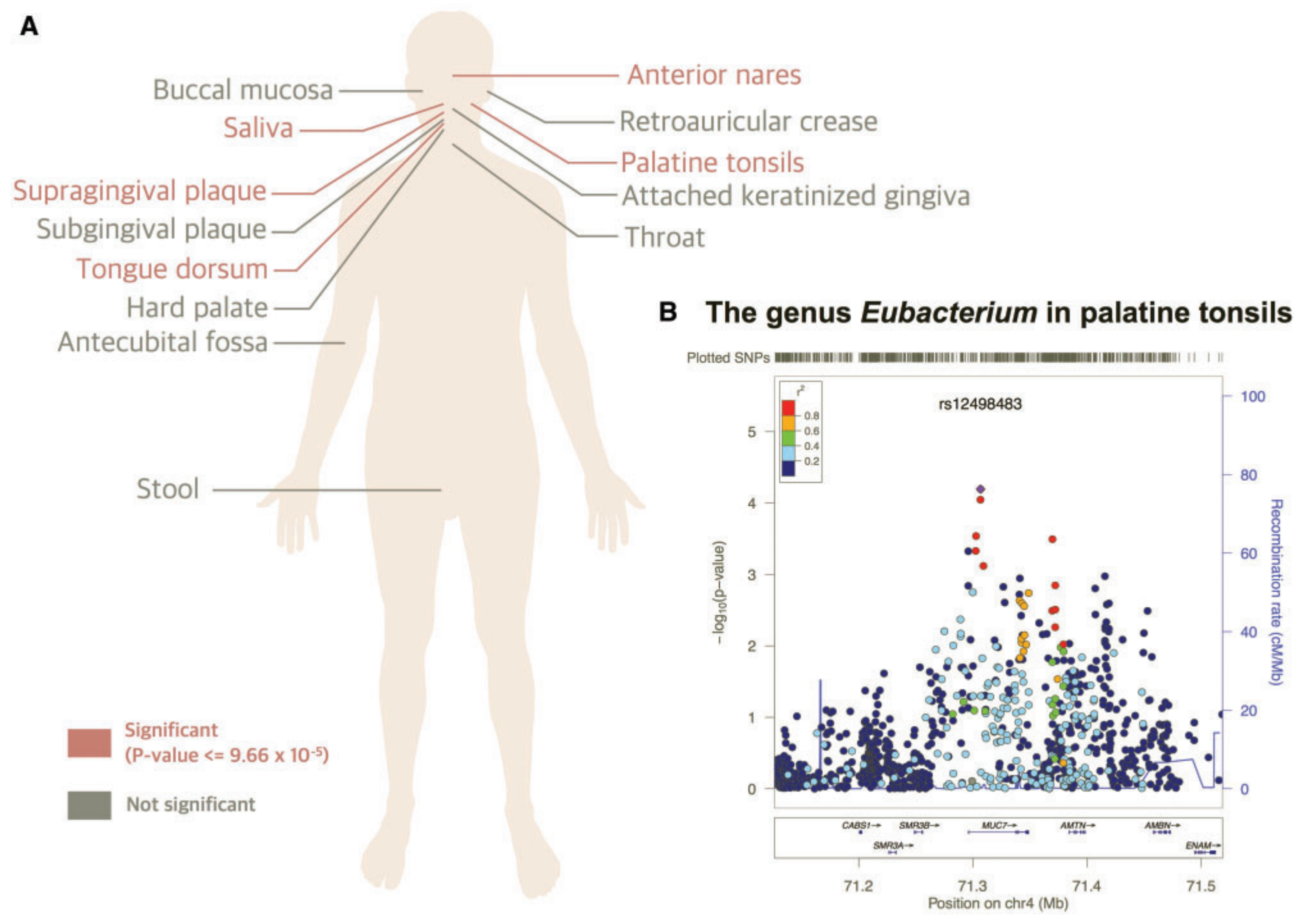

FIG. 3. The association of human genetic variation around *MUC7* with the microbiome. (*a*) Body sites where we found significant correlations between microbiome composition and *MUC7* genetic variants (indicated in red; *P*-value $\leq 9.66 \times 10^{-5}$). The body sites where we did not find any significant correlations are indicated with gray text. (*b*) The correlation between SNPs in the *MUC7* locus and the abundance of *Eubacterium* in the palatine tonsils. The *x*-axis shows the location on chromosome 4. The top panel (labeled as Plotted SNPs) indicates the locations and density of tested host (i.e., human) single nucleotide variants. At the bottom of the graph, the locations of individual genes are indicated. The main graph indicates the correlation (-$\log_{10}$ (*P*-value) on the *y*-axis) between host SNPs and the abundance of *Eubacterium* in the tonsils. The colors represent the R^2 value showing the correlation between each SNP and rs12498483, which has the strongest association with Eubacterium (shown in purple at the top and center of the plot). The *y*-axis on the right shows the recombination rate, which is indicated in the plot by the blue line. As it can be observed, there is no indication of recombination affecting the *MUC7* locus.

number of single nucleotide variants in that locus (see Materials and Methods for details). Thus, compared with Wall's B, our approach is more robust to noise generated by rare variants or sequencing errors. Our results showed a significantly higher than expected number of variants that are in linkage disequilibrium with each other in the *MUC7* locus compared with 1,000 randomly selected regions of the genome for African (YRI) populations (*P*-value = 0.050, Wilcoxon rank-sum test), but not for Eurasian populations (*P*-values > 0.1, Wilcoxon rank-sum test, supplementary table S4, Supplementary Material online, fig. 4*a*). We also confirmed this observation using other statistics such as theta (θ) and Wall's B (supplementary fig. S4, Supplementary Material online). These results support that haplogroup E indeed constitutes a distinct clade separated by an unexpectedly long branch in the gene tree.

The unusually high divergence of haplogroup E from other haplogroups suggests an increased coalescence time to the most recent common ancestor in the *MUC7* locus. To investigate the coalescent depth of the haplogroups, we used the BEAST software (Drummond and Rambaut 2007) to construct a phylogeny of two haplotypes from each haplogroup described above, including the chimpanzee and rhesus macaque reference haplotypes as outgroups (fig. 4*b*). This allowed us to estimate the most likely coalescence time for haplogroup E and the other human *MUC7* haplotypes to ~4.5 million years before present (95% confidence interval ranges from 3.2 to 6.3 million years before present). This indicates that for *MUC7* both tree depth among humans and divergence between humans and chimpanzees date back further than most other parts of the genome (Schiffels and Durbin 2014). Our analysis also confirms that the depth of the tree is primarily due to the divergent haplotype group E.

On the basis of the coalescent tree, we noticed that haplogroup E originated prior to the assumed Human–Neanderthal population divergence, previously dated to between 260 and 765 thousand years before present (Prüfer et al. 2014). To quantify this, we first calculated the

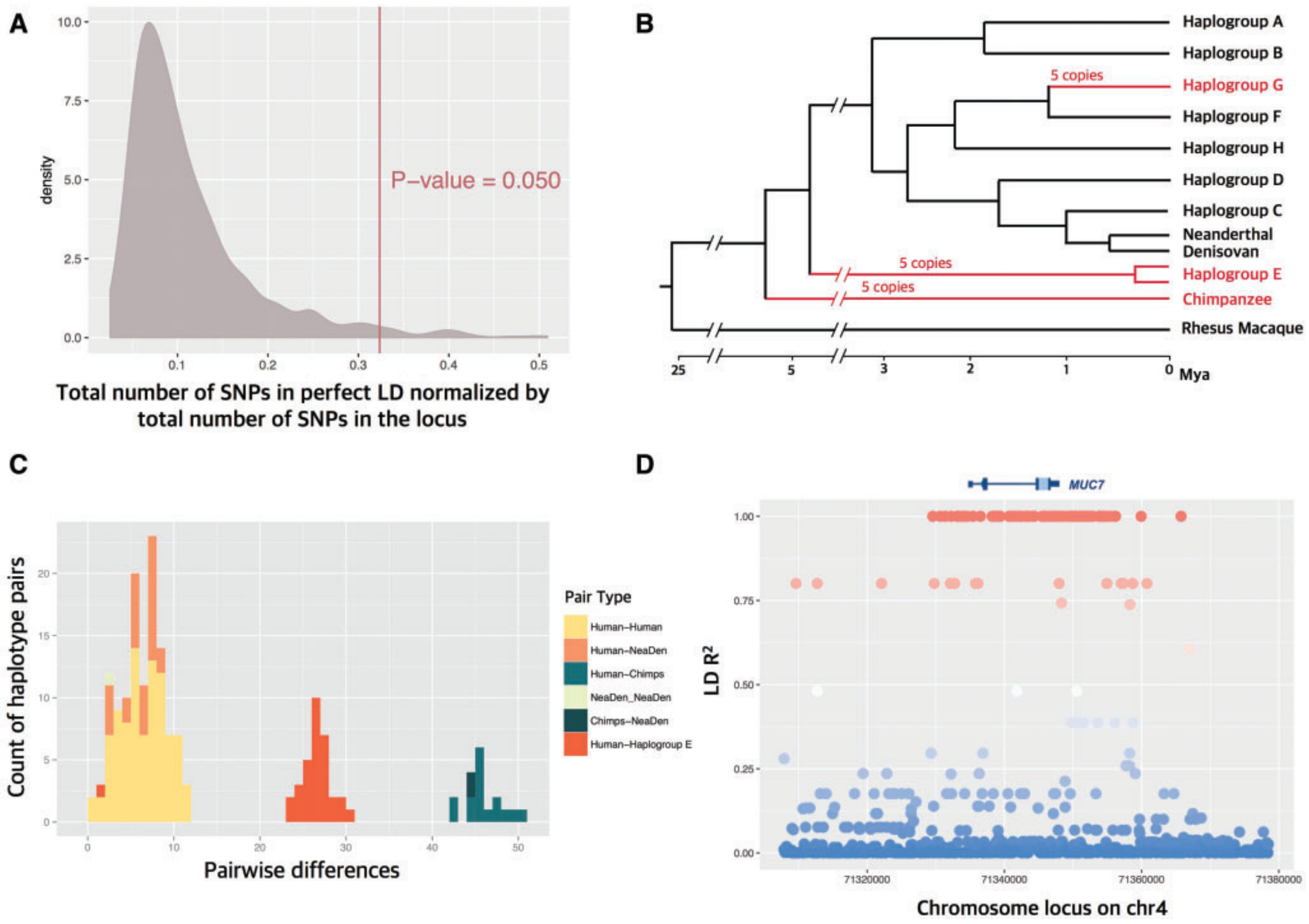

Fig. 4. Unusually derived haplogroup E sequences. (*a*) Total number of SNPs in perfect LD normalized by total number of SNPs in *MUC7* as compared with 1,000 random human genomic regions calculated for 1,000 Genomes Phase 3 data set and compared with *MUC7* (red line). The *P*-values for testing whether *MUC7* genetic variation is within this distribution were calculated by Wilcoxon rank-sum test, are shown on the plots. (*b*) Phylogenetic tree constructed by BEAST software with coalescence times shown at the bottom. Note that the tree deviates from scale in two instances indicated by line breaks. We also represent two haplotypes from haplogroup E to show that there is little variation within this haplogroup, as compared with the others. This indicates that haplogroup E was likely introduced to the population more recently, or that it underwent a recent bottleneck. (*c*) Counts of haplotype pairs (*y*-axis) of *MUC7* haplotypes. Different colors were used to indicate groups of haplotype comparisons (e.g., Neanderthal and Denisovan (NeaDen) vs. other human haplotypes or Chimpanzee [Chimps]). (*d*) The haplotype block of haplogroup E in the YRI population. The *x*-axis of the plot shows the chromosomal location on chromosome 4, whereas the *y*-axis shows the linkage disequilibrium (R^2) between all single nucleotide variation around *MUC7* and rs7684907, which tags haplogroup E. SNPs with high LD with rs7684907 (R^2 0.75) are indicated in red, whereas those with $R^2 < 0.5$ are marked in blue.

number of nucleotide differences between each human haplotype in a pairwise fashion. Then we plotted these differences as a histogram (fig. 4*c*). On the basis of this plot, nucleotide differences between haplotypes generate "clusters" of similar haplotypes, whereas more distant haplotypes are further apart. We found that human and chimpanzee haplotypes clustered separately. We also found that human–Neanderthal and human–Denisovan haplotypes, as well as most human–human haplotypes clustered together. Strikingly, haplogroup E haplotypes when compared with other haplotypes form a distinct cluster, suggesting that haplogroup E is unusually divergent from other human haplotypes. On the basis of linkage disequilibrium analysis, we further showed that the haplogroup E covers the entire *MUC7* gene and corresponds to an unbroken haplotype block of ~20 kb (fig. 4*d*).

The Divergent *MUC7* Haplotype Likely Introgressed from an Archaic African Hominin Population

We focused on three possible scenarios to explain the unusual divergence of haplogroup E. First, our null hypothesis was that *ancient structure* within the human lineage may be responsible for this divergent haplotype in the *MUC7* locus. Ancient structure refers to the remnants of genetic structure of populations ancestral to modern humans, which have been maintained across the human genome. It has been reported previously that such variation is especially detectable in regions with low recombination and high mutation rate (μ; Plagnol and Wall 2006; Lin et al. 2015). Second, we considered long-term balancing selection as a possibility to explain the maintenance of highly divergent lineages within the *MUC7* region. Indeed, our previous work as well as that of others have shown evidence of such loci where ancient lineages have been maintained within human populations (Gokcumen, Zhu, et al. 2013; Leffler et al. 2013; DeGiorgio et al. 2014; Key et al. 2014; Pajic et al. 2016). However, based on the phylogenetic tree (fig. 4*b*) and distribution of pairwise differences (fig. 4*c*), both ancestral structure and balancing selection scenarios are unlikely. To be more specific, the depth of the branches within haplogroup E is relatively shallow despite the fact that they are highly divergent from other haplogroups. This indicates that haplogroup E's origin within the human population is relatively recent as compared with other haplogroups. Corroborating this observation is that the haplotype block that defines haplogroup E remained intact (i.e., did not undergo recombination or gene conversion events, fig. 4*d*), also indicating that the haplogroup did not have the time to recombine within the human lineage. These observations fit neither to ancestral structure nor to balancing selection scenarios very well (Charlesworth 2006; Green et al. 2010; Yang et al. 2012).

Third, we considered introgression from an archaic hominin as a source of this haplotype, which will explain the divergence of haplogroup E from other haplogroups, as well as the apparently short coalescence of this haplogroup within the human lineage. This assumption was based on recent observations of similarly divergent haplotypes that were due to archaic introgression (Huerta-Sánchez et al. 2014; Dannemann et al. 2016). However, introgression of haplogroup E from Neanderthals or Denisovans is unlikely, because haplogroup E is exclusively found in sub-Saharan African populations, whereas introgression events between Neanderthals or Denisovans with modern humans happened after modern humans migrated out of Africa (Vernot and Akey 2014; fig. 5*a*, supplementary fig. S5, Supplementary Material online). As such, we hypothesize that an archaic hominin, at that time still roaming in Africa, contributed haplogroup E to the ancestors of extant Africans. The possibility of such an introgression has been discussed in recent studies (Hammer et al. 2011; Hsieh et al. 2016).

To distinguish between these scenarios and test our hypothesis in a more rigorous and quantitative way, we simulated the effects of the hypothesized ancient introgression to present day African genetic variation. It is important to note that such an introgression will have a genome-wide and not locus-specific effect. As such, before testing whether *MUC7* haplogroup E was indeed the result of introgression, we first asked whether such introgression has happened and, if it did, we wanted to simulate its signatures in present day African genomes. To do this, we first assumed the demographic scenario described in Gravel et al. (2011), which is characterized by a population expansion that took place 148,000 years ago to its final present-day effective population size of 14,474 (supplementary table S5, Supplementary Material online). On the basis of these parameters, we considered two competing scenarios (fig. 5*b*). First, we considered a scenario with no introgression and where *all* the variation in observed summary statistics can be explained by drift, and variation of mutation and recombination rates in the genome. Second, we considered a scenario where some of the observed variation can be explained by invoking introgression from a "ghost" species that shares a common ancestry with the modern human lineage. To distinguish between these two competing scenarios, we used an Approximate Bayesian Computation (ABC) approach (Beaumont et al. 2002; Csilléry et al. 2012) to find the posterior probability distribution of multiple parameters (supplementary table S5, fig. S6*A*–*F*, Supplementary Material online), including the rate and duration of the putative introgression, that would explain the genetic variation observed in each of the 50 random 10 kb fragments along chromosome 4. Using this approach, we generated 300,000 simulations for each of the 50 fragments. Our results showed evidence for introgression from an unsampled archaic population to the genomes of present-day Africans (supplementary fig. S7, Supplementary Material online, Bayes Factor > 19), and as such, are concordant with recent studies describing such an introgression in Africa (Hsieh et al. 2016).

For each simulated data set we recorded the parameter values, as well as the number of introgression events that occurred during the genealogical history of the sample. Using these simulated data sets, we were able to analyze the impact of such an introgression event to the modern day genomic variation (supplementary fig. S7, Supplementary Material online). For example, Wall's B and Watterson's θ Estimator are both clearly higher for simulated fragments that were affected by introgression ("introgressed") than for those that were not ("nonintrogressed"; fig. 5*c*). This set-up allowed us to specifically test whether *MUC7*'s genetic variation is most likely explained under a scenario which involves introgression or under one that does not. Indeed, in all statistics that were influenced by introgression events, the respective values for *MUC7* are best explained by invoking introgression (fig. 5*c*, supplementary fig. S7, Supplementary Material online).

To further interrogate whether introgression best explains the data, we used S^* statistics (Plagnol and Wall 2006) as applied recently to detect introgression(s) from Neanderthal and Denisovan genomes into the modern gene pool (Vernot and Akey 2014; Vernot et al. 2016). We first calculated S^* for the 50 kb region that encompasses *MUC7* in five replicates (each with 20 randomly chosen YRI genomes) using Northwestern European genomes as outliers (see supplementary methods, Supplementary Material online). High S^* scores indicate haplotypes that harbor higher numbers of derived variants and are shown to indicate introgressed regions in the genome (Vernot and Akey 2014; Hsieh et al. 2016). To empirically contextualize our results, we first compared the S^* values we got from our replicates to S^* values from other 50 kb regions across chromosome 4 (fig. 5*d*). We found that the *MUC7* region is a clear and significant outlier (*P*-value < 0.0001). To verify this, we simulated the S^* distribution where we use the number of segregating sites and recombination rate matching what is observed for *MUC7* (0.17 Cm/Mb, 495 segregating sites). In these simulations, we assume no introgression (see Materials and Methods for details). Note that the number of segregation sites is very high and the recombination rate is very low. This set of parameters is the most conservative allowing derived

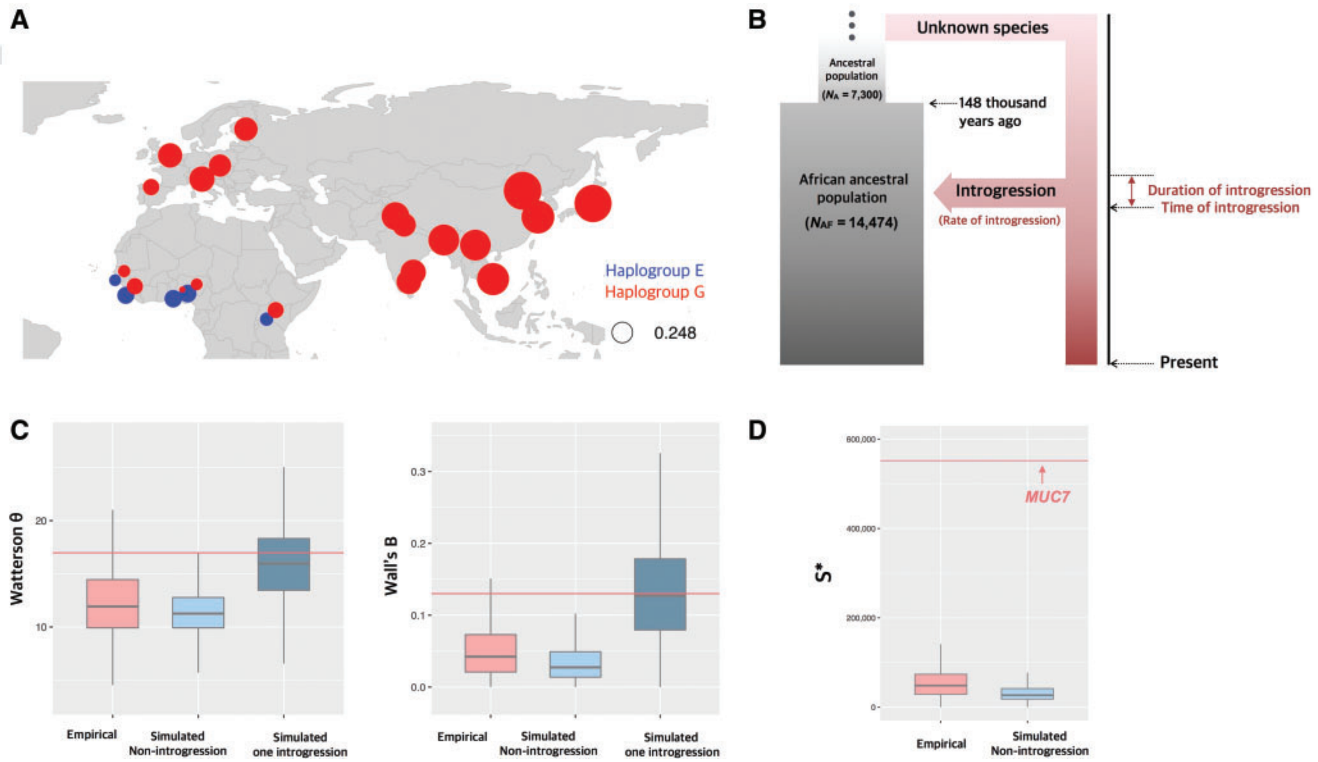

Fig. 5. (*a*) A global map showing the allele frequencies of five-copy variants. The size of the bubbles is proportional with the allele frequency of the five-copy alleles. Note that this map only shows five-copy haplogroups. The allele frequencies of six-copy variants are not shown. The red and blue colors indicate haplogroups G and E, respectively (see fig. 2*a* for phylogenetic relationships). (*b*) The model for simulations that we presented in this study. Our model is based on our hypothesis that an archaic hominin species in Africa contributed genetic material to ancestors of modern Africans. Our model is built on demographic parameters summarized by Gravel et al. (2011). We used a Bayesian framework to estimate other parameters including mutation and recombination rates as well as, time, duration, and rate of the putative introgression (see supplementary methods, Supplementary Material online). (*c*) Comparison of multiple population statistics of simulated data with observed values for *MUC7*. The *y*-axes of the plots show Watterson estimator (θ), and Wall's B values, respectively. The red boxes show the empirical values observed across chromosome 4, the light and dark blue boxes show the statistics calculated for the simulated fragments that had no introgression and introgression, respectively. Here, we show the results from the simulated sequences where introgression happened once, even though we also considered multiple introgression scenarios (data shown in supplementary fig. S7, Supplementary Material online). The horizontal red line shows the value calculated for the *MUC7* loci. Note that for both statistics, *MUC7* values cluster with introgression scenario. (*d*) The last plot shows the comparison of empirical and simulated S^* values with observed S^* values for *MUC7*, respectively. The empirical distribution is calculated for 50 kb windows across chromosome 4. The simulated distribution is calculated assuming the demographic model in Gravel et al. (2011), and recombination and mutation rates matching those observed in *MUC7*. In both plots the horizontal red line indicates the S^*-score measured for the *MUC7* locus.

haplotypes without invoking introgression. Even then, our results show that *MUC7* is a clear outlier and show unusually high S^* scores (P-value < 0.0001), indicating the effect of introgression in this locus (fig. 5*d*).

It is plausible that long-term balancing selection may have maintained old lineages, mimicking the signatures of introgression. To test this possibility, we used SLiM (Messer 2013), which allows simulation of complex evolutionary scenarios, to model long-term heterozygote advantage using the posterior distribution of the models that we described above alongside with the Yoruba demographic history described in Gravel et al. (2011) (see supplementary methods, Supplementary Material online). Two new parameters included here that are different from previous simulations are the selection coefficient (s) and dominance parameter (h). The former is the strength of the selection (here we considered $s = 0.001$, $s = 0.01$, and $s = 0.1$) and the latter is the equilibrium frequency at which the allele is balanced (here we considered equilibrium frequencies ranging from 0.5 [$h = 100$] to 0.9 [$h = 1.125$]). Overall, we performed 1,000 forward simulations for each of the different balancing selection scenarios with different combinations of s and h. None of these scenarios explains the observed variation in *MUC7* fully, where multiple summary statistics do not fit the simulated results invoking balancing selection (supplementary fig. S8, Supplementary Material online). We conclude based on these results that balancing selection in the form of heterozygote advantage with different s and h values cannot explain the observed variation as well as the introgression model.

Last, to test which of the above-described scenarios explain *MUC7* genetic variation in a collective and comprehensive manner, we conducted a principal component analysis of all summary statistics for all simulation results. Briefly, we constructed a matrix of summary statistic distributions for

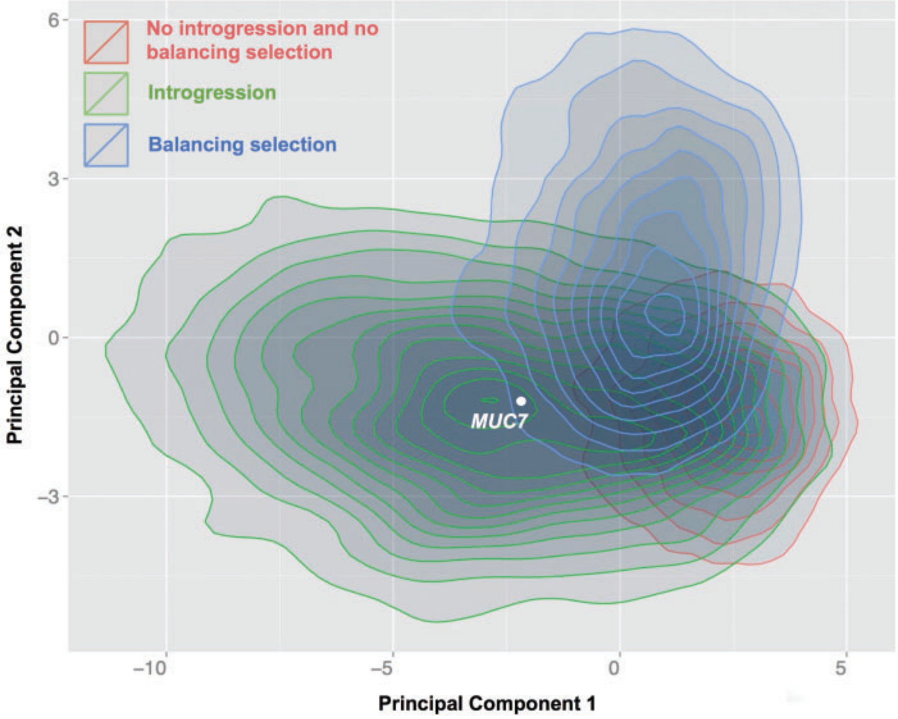

Fig. 6. Principal component analyses of all summary statistics for different scenarios. For this diagram, we used all the summary statistics presented individually in supplementary fig. S7, Supplementary Material online to create a 2-D density plot of principal component 1 (*x*-axis) and principal component 2 (*y*-axis), which explains 49.1% and 26.5% of the variation, respectively. We have categorized the data points into three scenarios, where we considered no-introgression neutral (red), introgression neutral (green), and balancing selection (blue). It is apparent that *MUC7* is placed in the middle of the introgression distribution, and distant from other scenarios. We further used ABC-based categorization to show that indeed introgression is the most likely scenario explaining *MUC7* summary statistics (Bayes Factor $\geq$ 12, supplementary methods, Supplementary Material online).

simulated sequences that are generated under no introgression, introgression and balancing selection scenarios. On the basis of this matrix, we were able to calculate principal components of the multivariate data where principal components 1 and 2, explain 49.1% and 26.5% of the data (fig. 6). We showed that *MUC7* is placed within the introgressed sequences and away from simulated sequences belonging to other scenarios. To quantify this observation, we have conducted an ABC-based categorization of *MUC7*, where we showed that indeed *MUC7* genetic variation is best explained by the introgression scenario with robust statistical support (Bayes Factor $>$ 36 vs. no introgression; Bayes Factor $>$ 12 vs. balancing selection, see supplementary methods, Supplementary Material online). The simulation results show with high confidence that introgression is a better model to explain the polymorphism patterns of the *MUC7* region than the alternative models which do not account for such an introgression.

Taken together, our analyses collectively support the hypothesis that ancient introgression has influenced modern African genomes and that genetic variation in *MUC7* locus is influenced by this introgression event.

Discussion

MUC7 is an abundant member of only a few principal proteins native to human saliva (i.e., those that are secreted exclusively by salivary glands; Ruhl 2012). Like other mucins, the key structural feature of the MUC7 protein is its densely glycosylated PTS-repeat domain (Dekker et al. 2002). The genetic variation affecting these repeats likely has multiple implications on the protein's function, including its impact on the rheological properties of saliva (Inoue et al. 2008), as well as its interactions with commensal and pathogenic microorganisms (Murray et al. 1992; Takamatsu et al. 2006; Walz et al. 2009; Heo et al. 2013; Thamadilok et al. 2016). Earlier studies have claimed that the copy number variation of *MUC7* PTS-repeats in humans is related to upper-respiratory infections (Kirkbride et al. 2001) as it was shown for the larger, functionally but not evolutionarily related mucin in saliva, *MUC5B* (Roy et al. 2014). Specifically, five-copy PTS-repeat alleles of *MUC7* have been associated with protection against asthma in independent cohorts of Europeans (Kirkbride et al. 2001) and Africans (Watson et al. 2009). It is relevant to note here that MUC7 was reported to be involved in maintaining salivary and mucus viscous properties, quantified by *Spinnbarkeit* (Inoue et al. 2008). It is plausible, therefore, that the six PTS-repeat alleles, while providing more protection against certain pathogens, increased the viscosity of saliva, thereby indirectly influencing the severity and pathology of asthma. However, we were not able to replicate the locus-specific associations of *MUC7*'s genetic variation with protection against asthma (Torgerson et al. 2011).

We conclude that the previous studies likely suffer from inadequate sampling and hence reported spurious associations. It is also possible that the interaction of *MUC7*'s genetic variation with commensal and pathogenic microbiome composition may indirectly influence asthma susceptibility. Indeed, our results showed that genetic variation around *MUC7* is associated with microbiome composition in an European ancestry cohort. As such, further studies are needed to clarify the impact of *MUC7*'s genetic variation on human colonization by specific commensal and pathogenic bacteria, and its broader biomedical relevance.

Our results suggested that the copy number variation of highly O-glycosylated *MUC7* PTS-repeats has recurrently evolved in the human lineage. This observation supports our previous work, where we showed that copy number variation of *MUC7* PTS-repeats has rapidly evolved under adaptive pressures likely shaped by ever-changing pathogenic pressures among primates (Xu et al. 2016). As such, geography-specific complex forces likely shaped the distribution of the five- and six-copy PTS-repeat alleles in human populations as well. Such geography-specific, complex adaptive forces have been recently shown to be more prevalent than previously thought (Eaaswarkhanth et al. 2014; Quintana-Murci 2016). Testing whether such complex adaptive forces helped maintain *MUC7* genetic variation in humans is complicated due to the recurrent nature of genetic variation in this locus, as well as the presence of introgressed haplotypes. It remains an intriguing question for future studies.

The most important contribution of our study to the understanding of human evolution is the serendipitous discovery that introgression from an enigmatic African population likely contributed to the noteworthy genetic variation of the *MUC7* gene at both single nucleotide and copy number variation levels. Our finding agrees with recent reports of such an introgression in sub Saharan African populations (Hammer et al. 2011; Hsieh et al. 2016), as well as the unexpectedly old human remains (Hublin et al. 2017) and lineages (Schlebusch et al. 2017). The role of *MUC7* in host–microbe interaction adds an important novel aspect to the ongoing discussion about the impact of archaic introgression in the human genomes. Indeed, recent studies have shown that haplotypes introgressed from Neanderthals have shaped the variation in the immune system of modern humans (Quach et al. 2016). For example, even though all other five-copy alleles have the haplogroup G background in Eurasia, the introgressed haplogroup E alone constitutes the majority of five PTS-repeat alleles in Africa. Therefore, it is plausible that adaptive forces, that favored five-copy alleles, may have maintained the introgressed haplogroup E in African populations. Within contemporary data sets, we found no obvious trends in pathogenic pressures or geophysical factors (e.g., humidity or temperature) that explains the distribution of *MUC7* PTS-repeats copy number variation in contemporary human populations (data not shown). However, additional analyses, especially among indigenous or ancient populations may reveal such patterns. Lachance et al. suggested that another locus with known copy number variation involving the *CSMD1* gene shows evidence of archaic admixture among Sub-Saharan African populations (Lachance et al. 2012). Thus, our results contribute to the emerging notion that functional variation introgressed from an African hominin population into modern humans and has been maintained among contemporary African populations.

Collectively, our study exemplifies how combined locus-specific and genome-wide approaches can shed light onto the evolutionary history and functional implications of complex, recurrent structural variation. This complements the recent work on complex structural variants within disease-susceptibility loci (Boettger et al. 2016; Eaaswarkhanth et al. 2016; Pajic et al. 2016; Sekar et al. 2016). Given that sub-exonic PTS repeat variation in other mucins, similar to that occurring in *MUC7*, may have critical roles for disease susceptibility and evolutionary innovations (Barreiro et al. 2005; Madsen et al. 2008; Gokcumen, Tischler, et al. 2013; Kirby et al. 2013; Schaper et al. 2014), we anticipate that our work will serve as a model to study other yet-to-be-discovered recurrent variants in similarly complex regions of the genome.

Materials and Methods

The samples that were genotyped for PTS-repeat copy number and sequenced for individual repeat haplotypes were purchased from Coriell Institute. These samples are included in the 1,000 Genomes Project, but the PTS-repeat copy number variation in these samples was not genotyped. The specific sample names for these human samples can be found in supplementary table S1, Supplementary Material online. There are multiple transcripts of *MUC7*, all with the same coding sequence. For this analysis, we used the most common transcript (GRCh37/hg19, chr4:71337834-71348714). We used PCR to amplify the region that contains the entire PTS-repeat region of *MUC7* in humans (GRCh37/hg19, chr4:71346912-71347433). For the population genetic analyses, we analyzed the entire gene but omitted the tandem repeat regions (GRCh37/hg19: chr4:71337834-71346953 and 71347368-71348714). The microbiome analysis was conducted based on 16S rRNA sequencing performed as part of the Human Microbiome Project and as described previously (Human Microbiome Project Consortium 2012). The details for both experimental, bioinformatic and population genetic analysis can be found in supplementary methods, Supplementary Material online. We provide the codes for statistical analyses and simulations used in this study on our website (gokcumenlab.org), as well as in GitHub (https://github.com/duoduoo/Xu_2017).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors thank Derek Taylor, Paul Cullen, and James Berry for critical readings of previous versions of this manuscript. This study is primarily funded by OG's start-up funds, as well as IMPACT grant both from University at Buffalo Research

Foundation. P.P. was funded by the grant FP7 REGPOT-InnovCrete (No. 316223) grant, as well as by the FP7-PEOPLE-2013-IEF EVOGREN (625057). M.D. is supported by funds from the Alfred P. Sloan Foundation. S.R. is funded by NIH grants R01DE019807 and R21DE025826. We thank Andrew Clark, Ruth Ley, and Dirk Gevers for their help with obtaining and analyzing the HMP microbiome and host genetic data, and commenting on the manuscript. The authors declare that there is no conflict of interest with organization regarding to the materials discussed in this article.

References

Anon. 1999. The Childhood Asthma Management Program (CAMP): design, rationale, and methods. Childhood Asthma Management Program Research Group. *Control. Clin. Trials*. 20:91–120.

Barreiro LB, Patin E, Neyrolles O, Cann HM, Gicquel B, Quintana-Murci L. 2005. The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region. *Am. J. Hum. Genet*. 77:869–886.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

Bennett EP, Mandel U, Clausen H, Gerken TA, Fritz TA, Tabak LA. 2012. Control of mucin-type O-glycosylation: a classification of the polypeptide GalNAc-transferase gene family. *Glycobiology* 22:736–756.

Biesbrock AR, Bobek LA, Levine MJ. 1997. MUC7 gene expression and genetic polymorphism. *Glycoconj. J*. 14:415–422.

Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, Spector TD, Keinan A, Ley RE, Gevers D, et al. 2015. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol*. 16:191.

Boettger LM, Salem RM, Handsaker RE, Peloso GM, Kathiresan S, Hirschhorn JN, McCarroll SA. 2016. Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. *Nat. Genet*. 48:359–366.

Boushey HA, Sorkness CA, King TS, Sullivan SD, Fahy JV, Lazarus SC, Chinchilli VM, Craig TJ, Dimango EA, Deykin A, et al. 2005. Daily versus as-needed corticosteroids for mild persistent asthma. *N. Engl. J. Med*. 352:1519–1528.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet*. 2:e64.

Csilléry K, François O, Blum MGB. 2012. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol*. 3:475–479.

Dannemann M, Andrés AM, Kelso J. 2016. Introgression of Neanderthal- and Denisovan-like haplotypes contributes to adaptive variation in human toll-like receptors. *Am. J. Hum. Genet*. 98:22–33.

DeGiorgio M, Lohmueller KE, Nielsen R. 2014. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet*. 10:e1004561.

Dekker J, Rossen JW, Büller HA, Einerhand AW. 2002. The MUC family: an obituary. *Trends Biochem. Sci*. 27:126–131.

Downes J, Munson MA, Spratt DA, Kononen E, Tarkka E, Jousimies-Somer H, Wade WG. 2001. Characterisation of Eubacterium-like strains isolated from oral infections. *J. Med. Microbiol*. 50:947–951.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol*. 7:214.

Eaaswarkhanth M, Pavlidis P, Gokcumen O. 2014. Geographic distribution and adaptive significance of genomic structural variants: an anthropological genetics perspective. *Hum. Biol*. 86:260–275.

Eaaswarkhanth M, Xu D, Flanagan C, Rzhetskaya M, Hayes MG, Blekhman R, Jablonski N, Gokcumen O. 2016. Atopic dermatitis susceptibility variants in Filaggrin Hitchhike Hornerin selective sweep. *Genome Biol. Evol*. 8(10):3240–3255.

Frenkel ES, Ribbeck K. 2015. Salivary mucins protect surfaces from colonization by cariogenic bacteria. *Appl. Environ. Microbiol*. 81:332–338.

Gokcumen O, Tischler V, Tica J, Zhu Q, Iskow RC, Lee E, Fritz MH-Y, Langdon A, Stütz AM, Pavlidis P, et al. 2013. Primate genome architecture influences structural variation mechanisms and functional consequences. *Proc. Natl. Acad. Sci. U. S. A*. 110:15764–15769.

Gokcumen O, Zhu Q, Mulder LCF, Iskow RC, Austermann C, Scharer CD, Raj T, Boss JM, Sunyaev S, Price A, et al. 2013. Balancing selection on a regulatory region exhibiting ancient variation that predates human–Neanderthal divergence. *PLoS Genet*. 9:e1003404.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, 1000 Genomes Project, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U. S. A*. 108:11983–11988.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A draft sequence of the Neanderthal genome. *Science* 328:710–722.

Gururaja TL, Ramasubbu N, Venugopalan P, Reddy MS, Ramalingam K, Levine MJ. 1998. Structural features of the human salivary mucin, MUC7. *Glycoconj. J*. 15:457–467.

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD. 2011. Genetic evidence for archaic admixture in Africa. *Proc. Natl. Acad. Sci. U. S. A*. 108:15123–15128.

Heo S-M, Choi K-S, Kazim LA, Reddy MS, Haase EM, Scannapieco FA, Ruhl S. 2013. Host defense proteins derived from human saliva bind to *Staphylococcus aureus*. *Infect. Immun*. 81:1364–1373.

Hollingsworth MA, Swanson BJ. 2004. Mucins in cancer: protection and control of the cell surface. *Nat. Rev. Cancer*. 4:45–60.

Hsieh P, Woerner AE, Wall JD, Lachance J, Tishkoff SA, Gutenkunst RN, Hammer MF. 2016. Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. *Genome Res*. 26:291–300.

Hublin J-J, Ben-Ncer A, Bailey SE, Freidline SE, Neubauer S, Skinner MM, Bergmann I, Le Cabec A, Benazzi S, Harvati K, et al. 2017. New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature* 546:289–292.

Huerta-Sánchez E, Jin X, Asan Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512:194–197.

Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–214.

Inoue H, Ono K, Masuda W, Inagaki T, Yokota M, Inenaga K. 2008. Rheological properties of human saliva and salivary mucins. *J. Oral Biosci*. 50:134–141.

Key FM, Teixeira JC, de Filippo C, Andrés AM. 2014. Advantageous diversity maintained by balancing selection in humans. *Curr. Opin. Genet. Dev*. 29C:45–51.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol*. 16:111–120.

Kirby A, Gnirke A, Jaffe DB, Barešová V, Pochet N, Blumenstiel B, Ye C, Aird D, Stevens C, Robinson JT, et al. 2013. Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. *Nat. Genet*. 45:299–303.

Kirkbride HJ, Bolscher JG, Nazmi K, Vinall LE, Nash MW, Moss FM, Mitchell DM, Swallow DM. 2001. Genetic polymorphism of MUC7: allele frequencies and association with asthma. *Eur. J. Hum. Genet*. 9:347–354.

Lachance J, Vernot B, Elbers CC, Ferwerda B, Froment A, Bodo J-M, Lema G, Fu W, Nyambo TB, Rebbeck TR, et al. 2012. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* 150:457–469.

Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* 339:1578–1582.

Lin Y-L, Pavlidis P, Karakoc E, Ajay J, Gokcumen O. 2015. The evolution and functional impact of human deletion variants shared with archaic hominin genomes. *Mol. Biol. Evol*. 32(4):1008–1019.

Madsen BE, Villesen P, Wiuf C. 2008. Short tandem repeats in human exons: a target for disease mutations. *BMC Genomics*. 9:410.

Mager DL, Ximenez-Fyvie LA, Haffajee AD, Socransky SS. 2003. Distribution of selected bacterial species on intraoral surfaces. *J. Clin. Periodontol*. 30:644–654.

Messer PW. 2013. SLiM: simulating evolution with selection and linkage. *Genetics* 194:1037–1039.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* 470:59–65.

Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, D'Agostino R Jr, Castro M, Curran-Everett D, Fitzpatrick AM, et al. 2010. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am. J. Respir. Crit. Care Med*. 181:315–323.

Murray PA, Prakobphol A, Lee T, Hoover CI, Fisher SJ. 1992. Adherence of oral streptococci to salivary glycoproteins. *Infect. Immun*. 60:31–38.

Naganagowda GA, Gururaja TL, Satyanarayana J, Levine MJ. 1999. NMR analysis of human salivary mucin (MUC7) derived O-linked model glycopeptides: comparison of structural features and carbohydrate–peptide interactions. *J. Pept. Res*. 54:290–310.

Pajic P, Lin Y-L, Xu D, Gokcumen O. 2016. The psoriasis-associated deletion of late cornified envelope genes LCE3B and LCE3C has been maintained under balancing selection since Human Denisovan divergence. *BMC Evol. Biol*. 16:265.

Papadantonakis S, Poirazi P, Pavlidis P. 2016. CoMuS: simulating coalescent histories and polymorphic data from multiple species. *Mol. Ecol. Resour.* 16(6):1435–1448.

Plagnol V, Wall JD. 2006. Possible ancestral structure in human populations. *PLoS Genet*. 2:e105.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505:43–49.

Quach H, Rotival M, Pothlichet J, Loh Y-HE, Dannemann M, Zidane N, Laval G, Patin E, Harmant C, Lopez M, et al. 2016. Genetic adaptation and Neanderthal admixture shaped the immune system of human populations. *Cell* 167:643–656.e17.

Quintana-Murci L. 2016. Understanding rare and common diseases in the context of human evolution. *Genome Biol*. 17:225.

Rossmann SN, Wilson PH, Hicks J, Carter B, Cron SG, Simon C, Flaitz CM, Demmler GJ, Shearer WT, Kline MW. 1998. Isolation of Lautropia mirabilis from oral cavities of human immunodeficiency virus-infected children. *J. Clin. Microbiol*. 36:1756–1760.

Rousseau K, Vinall LE, Butterworth SL, Hardy RJ, Holloway J, Wadsworth MEJ, Swallow DM. 2006. MUC7 haplotype analysis: results from a longitudinal birth cohort support protective effect of the MUC7*5 allele on respiratory function. *Ann. Hum. Genet*. 70:417–427.

Roy MG, Livraghi-Butrico A, Fletcher AA, McElwee MM, Evans SE, Boerner RM, Alexander SN, Bellinghausen LK, Song AS, Petrova YM, et al. 2014. Muc5b is required for airway defence. *Nature* 505:412–416.

Ruhl S. 2012. The scientific exploration of saliva in the post-proteomic era: from database back to basic function. *Expert Rev. Proteomics*. 9:85–96.

Schaper E, Gascuel O, Anisimova M. 2014. Deep conservation of human protein tandem repeats within the eukaryotes. *Mol. Biol. Evol* 31(5):1132–1148.

Schiffels S, Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet*. 46:919–925.

Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, Munters AR, Steyn M, Soodyall H, Lombard M, et al. 2017. Ancient genomes from southern Africa pushes modern human divergence beyond 260,000 years ago. *bioRxiv [Internet]*. 145409. Available from: http://biorxiv.org/content/early/2017/06/05/145409. abstract

Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, Tooley K, Presumey J, Baum M, Van Doren V, et al. 2016. Schizophrenia risk from complex variation of complement component 4. *Nature* 530:177–183.

Smith CJ, Bobek LA. 2001. Bactericidal and fungicidal activity of salivary mucin (MUC7) peptide fragments. *J. Dent. Res*. 80:601.

Sorkness CA, Lemanske RF Jr, Mauger DT, Boehmer SJ, Chinchilli VM, Martinez FD, Strunk RC, Szefler SJ, Zeiger RS, Bacharier LB, et al. 2007. Long-term comparison of 3 controller regimens for mild-moderate persistent childhood asthma: the Pediatric Asthma Controller Trial. *J. Allergy Clin. Immunol*. 119:64–72.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526:75–81.

Takamatsu D, Bensing BA, Prakobphol A, Fisher SJ, Sullam PM. 2006. Binding of the streptococcal surface glycoproteins GspB and Hsa to human salivary proteins. *Infect. Immun*. 74:1933–1940.

Thamadilok S, Roche-Håkansson H, Håkansson AP, Ruhl S. 2016. Absence of capsule reveals glycan-mediated binding and recognition of salivary mucin MUC7 by *Streptococcus pneumoniae*. *Mol. Oral Microbiol*. 31:175–188.

Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, Himes BE, Levin AM, Mathias RA, Hancock DB, et al. 2011. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat. Genet*. 43:887–892.

Vernot B, Akey JM. 2014. Resurrecting surviving Neanderthal lineages from modern human genomes. *Science* 343:1017–1021.

Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, Dannemann M, Grote S, McCoy RC, Norton H, et al. 2016. Excavating Neanderthal and Denisovan DNA from the genomes of Melanesian individuals. *Science* 352:235–239.

Wall JD. 1999. Recombination and the power of statistical tests of neutrality. *Genet. Res*. 74:65–79.

Walz A, Odenbreit S, Stühler K, Wattenberg A, Meyer HE, Mahdavi J, Borén T, Ruhl S. 2009. Identification of glycoprotein receptors within the human salivary proteome for the lectin-like BabA and SabA adhesins of *Helicobacter pylori* by fluorescence-based 2-D bacterial overlay. *Proteomics* 9:1582–1592.

Watson AM, Ngor W-M, Gordish-Dressman H, Freishtat RJ, Rose MC. 2009. MUC7 polymorphisms are associated with a decreased risk of a diagnosis of asthma in an African American population. *J. Investig. Med*. 57:882–886.

Xu D, Pavlidis P, Thamadilok S, Redwood E, Fox S, Blekhman R, Ruhl S, Gokcumen O. 2016. Recent evolution of the salivary mucin MUC7. *Sci. Rep*. 6:31791.

Yang MA, Malaspinas A-S, Durand EY, Slatkin M. 2012. Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. *Mol. Biol. Evol*. 29:2987–2995.

Zhao X, Emery SB, Myers B, Kidd JM, Mills RE. 2016. Resolving complex structural genomic rearrangements using a randomized approach. *Genome Biol*. 17:126.