Comparative genomics of *Mycobacterium africanum* Lineage 5 and Lineage 6 from Ghana suggests different ecological niches.

Authors:

Isaac Darko Otchere[1,2γ], Mireia Coscollá[3,4γ], Leonor Sánchez-Busó[5], Adwoa Asante-Poku[1], Daniela Brites[3,4,] Chloe Loiseau[3,4,] Conor Meehan[6], Stephen Osei-Wusu[1], Audrey Forson[7], Clement Laryea[8], Abdallah Iddrisu Yahayah[9], Akosua Baddoo[7], Gloria Akosua Ansa[10] Samuel Yaw Aboagye[1], Prince Asare[1], Sonia Borrell[3,4], Florian Gehre[6,11], Patrick Beckert[12,13], Thomas A Kohl[12,13], Sanoussi N'dira[14], Christian Beisel[15], Martin Antonio[11], Stefan Niemann[12,13], Bouke C de Jong[6,11], Julian Parkhill[5], Simon R Harris[5], Sebastien Gagneux[3,4]*, Dorothy Yeboah-Manu[1]*.

Author Affiliations:

[1]Noguchi Memorial Institute for Medical Research, University of Ghana, Legon, Accra, Ghana.

[2]Department of Biochemistry, Cell and Molecular Biology, University of Ghana, Legon, Accra, Ghana.

[3]Swiss Tropical and Public Health Institute, Basel, Switzerland.

[4]University of Basel, Basel, Switzerland.

[5]Wellcome Trust Sanger Institute, University of Cambridge, Hinxton, United Kingdom.

[6]Institute of Tropical Medicine, Antwerp, Belgium

[7]Chest Clinic, Korle-Bu Teaching Hospital, Accra, Ghana.

[8]37 Military Hospital, Accra, Ghana.

[9]Chest Department, Tamale Teaching Hospital, Tamale, Ghana.

[10]Public Health Department, University of Ghana Hospital, Legon, Accra, Ghana

[11]Medical Research Council, The Gambia Unit, Gambia

[12]Molecular and Experimental Mycobacteriology, Research Center Borstel, Borstel, Germany

[13]Germany German Center for Infection Research, Partner Site Hamburg-Borstel-Lübeck, Germany

[14]National Reference Laboratory for Mycobacteria, Cotonou, Benin

[15]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland.

*These authors contributed equally

[γ]These authors contributed equally

Correspondence should be addressed to: DYM (dyboah-Manu@noguchi.ug.edu.gh and SG (sebastien.gagneux@unibas.ch)

Abstract:

Mycobacterium africanum (*Maf)* causes up to half of human tuberculosis in West Africa, but little is known on this pathogen. We compared the genomes of 253 *Maf* clinical isolates from Ghana, including both L5 and L6. We found that the genomic diversity of L6 was higher than in L5, and the selection pressures differed between both groups. Regulatory proteins appeared to evolve neutrally in L5 but under purifying selection in L6. Conversely, human T cell epitopes were under purifying selection in L5, but under positive selection in L6. Although only 10% of the T cell epitopes were variable, mutations were mostly lineage-specific. Our findings indicate that *Maf* L5 and L6 are genomically distinct, possibly reflecting different ecological niches.

Introduction:

The global phylogeography of the human-adapted *Mycobacterium tuberculosis* complex (MTBC) demonstrates highest diversity in West Africa, with six out of the seven known lineages represented[1,2]. Two of these lineages, Lineage 5 (L5) and Lineage 6 (L6), together originally known as *Mycobacterium africanum* (*Maf*), are restricted to West Africa for unknown reasons. By contrast, MTBC lineages belonging to *Mycobacterium tuberculosis* sensu stricto (*Mtbss*), in particular Lineage 4 (L4), are more geographically widespread[1]. *M. africanum* has remained an important pathogen in West Africa since its first description in 1968[3], and is responsible for up to half of human tuberculosis (TB) in some regions[4].

The MTBC is thought to have originally emerged in Africa and subsequently spread to other parts of the world following waves of human migrations, trade and conquests[5–8]. Yet the reason(s) why *Maf* is limited to West Africa despite, for example, centuries of the trans-Atlantic slave trade remains unknown. Some comparative studies have identified phenotypic differences between the two *Maf* lineages[9,10], suggesting they might be fundamentally distinct and occupy different ecological niches.

Three hypotheses have been put forward to explain the restriction of *Maf* to West Africa. The first hypothesis proposes that *Maf* might have emigrated outside of Africa but was later outcompeted by *Mtbss*, which has been shown to be more virulent than *Maf* in animal models[11]. The second hypothesis states that the restriction of *Maf* to West Africa is due to its adaptation to West African human populations[9,12]. Finally, according to the third hypothesis, *Maf* might be zoonotic with an animal reservoir restricted to West Africa.

Some evidence in support of the first hypothesis is the reported association of *Maf* (L6) with HIV co-infection, attenuated ESAT-6 responses and delayed progression to active disease relative to *Mtbss*[9,13–16]. In addition, both *Maf* lineages as well as *Mtbss* L1, together described as "ancestral" MTBC lineages, have been shown to elicit a stronger early production of pro-inflammatory cytokines compared to the "modern" MTBBC L2, L3 and L4[17]. The delayed pro-inflammatory

74 immune response in the "modern" MTBC lineages might allow for more rapid disease progression
75 and transmission[17]. The second hypothesis is supported by the statistical association of L5 with the
76 native West African ethnic group known as "Ewe" reported by two independent studies in
77 Ghana[9,12]. The third hypothesis is mainly supported by the phylogenetic placement of *Maf* (L6)
78 amidst the cluster of the animal-adapted members of the MTBC in the various phylogenies of the
79 MTBC[5,7,18].

80 If the first hypothesis is true, the proportion of *Maf* associated TB in West Africa is expected to
81 decline over time. However, there are conflicting reports of the proportion of *Maf* associated TB in
82 West Africa. Even though the report of a steady decline of *Maf* associated TB in some settings
83 seems to support the first hypothesis[19–21], other studies indicate that *Maf* remains an important cause
84 of TB in West-Africa[22–24]. In Ghana for instance, a recent study showed that the proportion of TB
85 due to *Maf* remained constant over the 8 year study period[25]. Even though, the reported statistical
86 association of L5 with ethnicity in Ghana suggests a possible co-evolutionary scenario in favour of
87 the second hypothesis, genetic evidence of co-evolution/co-adaptation remains to be demonstrated.
88 In the case of the third hypothesis, the environmental or zoonotic reservoir(s) need to be identified.

89 In this study, we used whole genome sequencing of Ghanaian *Maf* clinical strains to explore
90 genomic differences between the two *Maf* lineages that might support one or more of these
91 hypotheses.
92
93

Results

Whole genome SNP distance, average nucleotide diversity and phylogeny of *Maf* in Ghana

Our data set comprised *Maf* isolates obtained from TB patients reporting to various hospitals in Ghana. After excluding genomes that did not meet the criteria for mapping (Supplementary figure S1), 253 *Maf* genomes (175 L5 and 78 L6) were used for the analysis. Patients' residential regions are provided (Supplementary figure S2). The upper right pie chart indicates those 97 patients (55 infected with L5 and 42 with L6) with no information on region of residence. We found the number of fixed SNPs (SNPs found in more than 95% of genomes) in a genome compared to the MTBC ancestor[26] to be significantly higher in L6 (1,037) compared to L5 (928) (Wilcoxon rank-sum test, $p < 0.0001$) (Fig 1A). Moreover, despite the larger number of L5 genomes (more than twice the number of L6 genomes) analyzed, the mean pairwise SNP distance between any two strains was significantly higher in L6 (360) compared to L5 (223) (Wilcoxon rank-sum test, $p < 0.0001$; Fig 1B). Finally, the whole genome average nucleotide diversity (π) for L6 (0.000110) was significantly higher compared to L5 (0.00007) (Fig 1C, non-overlapping 95% confidence interval (CI)). Taken together, these findings show that L6 in Ghana is significantly more genetically diverse than L5 irrespective of sample size. The whole genome-based phylogenetic tree of the Ghanaian *Maf* strains generated from 11,027 total polymorphic positions between the *Maf* strains and the MTBC ancestor reference excluding repetitive and mobile genetic element rooted on *M. canettii* is shown in Figure 2. The *Maf* lineages were resolved as two distinct branches of the genome-based tree with possible sub-groups (Fig 2).

Genetic diversity of L6 is significantly higher than L5 among T cell epitopes and genes of other functional categories.

We found that the higher diversity of L6 compared to L5 was reflected across all the 8 functional categories of genes analyzed (Fig. 3). Whereas pairwise nucleotide diversity (π) for L5 was below 0.0001 across all functional categories, the estimates for L6 were all above 0.0001. The most prominent difference between L6 and L5 was within 1,226 experimentally confirmed human T cell

121 epitopes of MTBC which we downloaded from the Immune Epitope Database (IEDB)[27] , for which
122 the mean π for L5 was 0.000063 compared to the 0.000149 estimated for L6, reflecting more than a
123 two-fold difference in diversity (non-overlapping 95% CI).
124 Within L5, there was no difference between the estimated π for the T cell epitopes and any of the
125 other functionally categorized genes. However, within L6, genes encoding regulatory proteins and
126 those involved with virulence, detoxification and adaptation were more diverse compared to those
127 for lipid metabolism as well as intermediate metabolism and respiration (non-overlapping 95% CI).
128 In addition, genes encoding regulatory proteins were more diverse compared to those involved with
129 lipid metabolism (non-overlapping 95% CI).
130
131 **Different selection pressures within L5 and L6 in human T cell epitopes and regulatory**
132 **proteins**
133 The average pairwise dN/dS of the concatenates of T cell epitopes as well as genes of the seven
134 other functional categories were calculated for all genomes and compared between L5 and L6.
135 Apart from sequences encoding human T cell epitopes and regulatory proteins that had median
136 average pairwise dN/dS ratios greater than 1.0 in L5 and L6, respectively (Fig 4, panel a and b), all
137 the remaining functional categories showed a dN/dS ratio of less than 1.0 in both lineages
138 (Supplementary figure S3). Human T cell epitopes of *Maf* L5 (median pairwise dN/dS = 0.64) were
139 significantly more conserved compared to L6, which exhibited higher diversity (median pairwise
140 dN/dS = 1.53) (Wilcoxon rank-sum test, W = 2265, $p < 0.0001$). Conversely, genes encoding
141 regulatory proteins were more diverse among L5 genomes (with median pairwise dN/dS = 1.03)
142 (Wilcoxon rank-sum test, W = 6303, $p = 0.0010$) compared to L6 (with median pairwise dN/dS =
143 0.85). To account for the different sample sizes; 147 L5 compared to 67 L6 genomes after
144 excluding 43 genomes differing from others with less than 10 SNPs difference (see Methods and
145 supplementary figure S1), we repeated the analysis using mean values of 10 randomly sampled sets
146 of L5 genomes with sample size 67 among human T cell epitopes (Fig 4, panel c) and regulatory
147 proteins (Fig 4 panel d) and got similar results (Wilcoxon rank-sum test, W = 1300, $p < 0.0001$, W

148 = 3466, $p < 0.0001$ for T cell epitopes and regulatory proteins, respectively).

149

150 **Lineage-specific accumulation of mutations within human T cell epitopes.**

151 When we compared the number of epitopes with amino acid mutations between lineages, we found
152 more epitopes mutated in L6 (57) compared to L5 (45), but no statistically significant difference
153 (Fig 5). In addition, we compared the number of nonsynonymous polymorphic sites between the
154 two Maf lineages within the human T cell epitopes (Supplementary Fig S4), and found that these
155 were more frequent in L6 (38) compared to L5 (28) but with no statistically significant difference
156 between L5 and L6. We compared the identity of the mutant human T cell epitopes between the two
157 *Maf* lineages (Fig 6A) and found 72 epitopes that were uniquely mutated in L5 (among 174
158 genomes) compared to 54 epitopes in L6 (among 67 genomes). Only two epitopes (IEDB IDs
159 178644 and 178609) were mutated in both lineages. However, the mutations were at different loci
160 with different amino acid substitutions (A183G and G278D in L5 compared to A177V and D277N
161 in L6). In terms of T cell antigens, there were 28 uniquely mutated in L5 compared to 19 in L6 and
162 12 mutated in both lineages involving different epitopes within the respective antigens (Fig 6B).
163 The 12 T cell antigens mutated in both lineages are summarized in Table 1. All T cell epitopes and
164 antigens with mutations among the two *Maf* lineages are listed in Supplementary table S5.

165

166 **Conservation of human T cell epitopes of L5 is not affected by patient ethnicity.**

167 We previously reported an association between L5 and Ewe patient ethnicity[9,12]. Hence to test if
168 conservation of T cell epitopes and/or the diversity of regulatory proteins in L5 was influenced by
169 patient ethnicity, we estimated pairwise dN/dS for sequences encoding T cell epitopes and
170 regulatory proteins of L5 genomes stratified by patient ethnicity (Fig 7). The median dN/dS of T
171 cell epitopes were all below 1.0 irrespective of patient ethnicity (Fig 7A). However, the median
172 dN/dS of regulatory proteins were marginally above 1.0 among L5 from Ewe TB patients and
173 below 1.0 among L5 from non-Ewe TB patients (Fig 7B). There was no statistically significant

174 difference between the estimated dN/dS of either the sequences encoding T cell epitopes (Fig 7A)
175 or regulatory proteins (Fig 7B) between L5 from TB patients of Ewe and non-Ewe ethnicities. In
176 addition, there was no difference in either the number of T cell epitopes with amino acid
177 substitutions (Supplementary figure S6A) or the number of non-redundant SNPs (Supplementary
178 figure S6B) between L5 strains from patients of the Ewe ethnicity and those of other ethnicities.
179
180

Discussion

In this study, we compared the largest collection of *Maf* genomes including both L5 and L6 reported so far. We found that, 1) at the whole genome level, L6 had significantly more pairwise nucleotide diversity, higher number of fixed SNPs as well as lower average pairwise SNPs relative to L5, 2) L5 had overall more conserved human T cell epitopes compared to L6, 3) conservation of T cell epitopes in L5 was not influenced by patient ethnicity, and 4) genes encoding regulatory proteins of L5 had lower pairwise nucleotide diversity but a higher ratio of non-synonymous to synonymous substitution rate than L6.

Our finding that *Maf* L6 has a higher number of fixed SNPs and higher average pairwise SNPs relative to L5 suggests that L6 has diversified more compared to L5 since the emergence of the two lineages[5,28,29]. This observation is corroborated by the higher genome-wide nucleotide diversity of L6 compared to L5. The higher diversity of L6 might be linked to an earlier emergence. However, recent whole genome-based phylogenies rooted on *Mycobacterium canettii* show that following the branch leading to *Maf* and all animal-adapted members of the MTBC defined by the characteristic deletion in RD9[30,31], L5 branches off much earlier than L6[32]. Hence, L5 is ancestral to L6, a notion which is also supported by genomic deletion analyses which show that in addition to RD9, L6 and all the animal-adapted members of the MTBC harbor the deletions of RD7, RD8 and RD10[30]. Hence other factors are likely to account for the higher diversity of L6 compared to L5 in Ghana.

Maf is highly restricted to West Africa, and thus could be seen as an ecological specialist compared to the other MTBC lineages. Specialists are expected to harbour less diversity across strains compared to generalists[8]. The observed lower genome-wide nucleotide diversity of L5 hence supports the hypothesis that L5 might be a specialist maintained in West Africa by adaptation to specific human genotypes. In contrast, the higher genome-wide diversity of L6 indicates a generalist pathogen, and hence would have been expected to be globally distributed instead of

displaying restriction to West Africa[4,8]. The observed diversity of L6 therefore may indicate a pathogen with a wider host range, supporting the hypothesis of maintenance in West Africa by possible environmental or zoonotic reservoir(s). Alternatively, the higher diversity of L6 coupled with the higher number of fixed SNPs could mean that it has a higher intrinsic mutation rate compared to L5.

Even though 90% of T cell epitopes were highly conserved in both L5 and L6 (Fig 5), which is in line with previous reports for the whole MTBC[26,8,33], we found T cell epitopes in L5 to exhibit less nucleotide diversity and to be under purifying selection compared to L6. The purifying selection of mutations within L5 is comparable to that reported for the specialist sub-lineages of L4[8]. Interestingly, dN/dS within essential genes for survival in macrophages did not differ between L5 and L6 (Supplementary Figure S3) supporting the notion that the genes in this category perform key functions in both L5 and L6. Since T cell response might partially drive the pathogenesis of TB[34], the relative conservation of T cell epitopes in L5 indicate that it might elicit a more efficient T cell response compared to L6 in its particular host population. This therefore suggests L5 may be a more human-specific pathogen and L6, with significantly more diverse T cell epitopes, a potential opportunistic environmental or zoonotic pathogen. Even though the conserved T cell epitopes of L5 could account for geographical restriction to West Africa and the association with the Ewe ethnicity[1,4,9,12], we found no difference between the diversity of L5 isolated from TB patients of Ewe and those of non-Ewe ethnic backgrounds. The limited number of L5 genomes from Ewe TB patients could possibly account for the lack of observed difference in diversity of T cell epitopes of L5 from TB patients of Ewe and non-Ewe ethnicities, and hence larger sample sizes are required to explore this further. L5 isolated from TB patients of the Ewe ethnicity were shown to be distributed all across the L5 clade of the *Maf* phylogeny instead of clustering in a particular sub-clade (Supplementary figure S7). This suggests that, if L5 is indeed maintained in West Africa by its co-evolution/adaptation with the Ewe ethnic group of West Africa (Cote d'Ivoire, Ghana, Nigeria,

Togo and Benin)[9,12], there is no specific sub-group of L5 that is responsible for this association but rather the whole of L5.

Members of the MTBC survive in the host mostly by modulation of the host immune response via the action of secretory proteins which form part of regulons controlled by specific regulatory proteins[35,36]. In addition, some regulatory proteins are involved in the regulation of transcription and translation of these secretory effectors as well as gene expression of other proteins involved with diverse functions[36,37]. Regulatory proteins in the MTBC hence play an important role in the survival and propagation of the bacteria. Therefore, our finding that regulatory proteins in L5 are under neutral selection (pairwise dN/dS = 1.03) compared to L6 in which they appear under purifying selection indicates that the mutations within regulatory proteins might be lineage-specific. This result is comparable to an earlier report comparing mutations within regulatory proteins between *Mtbss* and *M. bovis*, which found most of the *M. bovis* to harbor majority of the mutations[36]. As mutations within some regulatory proteins have been associated with attenuated virulence[38–40], our observation could account for the reported attenuated virulence of *Maf* relative to *Mtbss*[14,15,17,41]. This calls for further comparative studies of regulatory proteins between L5, L6 and other MTBC lineages to ascertain the role of regulatory proteins.

Our data is limited by the fact that, the number of L5 genomes was almost 3 times the number of L6 genomes; however, we used 1,000x bootstrap sampling with replacement of both L5 and L6 of equal sample size to limit any possible bias when comparing both lineages due to differences in sample size. In addition, a number of the L5 genomes did not have data on ethnicity and hence affected the number of L5 isolated from patients of the Ewe ethnicity for which we used average estimates of 10 random samples of L5 isolated from patients of non-Ewe origin in comparisons to account for the different sample sizes.

In conclusion, our findings indicate that the two *Maf* lineages L5 and L6 are distinct in terms of genomic diversity, and selection pressure on T cell epitopes and regulatory proteins, possibly reflecting different ecological niches. Whereas L5 may be maintained in West Africa by its co-evolution or adaptation with native West Africans, L6 may be maintained by an environmental reservoir, possibly a zoonotic source. This genomic analysis of *Maf* from Ghana gives a glimpse of the often neglected diversity within *Maf* and the MTBC overall. More studies are needed from representative genomes of *Maf* from across West Africa to understand the full diversity of these members of the MTBC. Improved knowledge of *Maf* will have implications for our understanding of human TB and the development of better control tools.

Tables

Table 1: Functions of the 12 T cell antigens mutated in both L5 and L6

T cell antigen	Function
Rv0288	encodes low molecular weight antigen 7 **EsxH** involved with cell wall and cell processes
Rv0934	encodes periplasmic phosphate-binding lipoprotein **PstS1** involved with cell wall and cell processes
Rv2029c	encodes 6-phosphofructokinase **PfkB** involved with intermediate metabolism and respiration
Rv2627c	encoding a conserved hypothetical protein
Rv3003c	Encodes the large subunit of acetolactate synthase involved with valine and isoleucine biosynthesis
Rv3024c	encodes a probable tRNA involved with information pathways
Rv3763	encodes a 19 kDa lipoprotein antigen precursor **LpqH** involved with cell wall and cell processes
Rv3804c	encodes the secreted antigen 85-a **FbpA** involved with lipid metabolism
Rv3823c	encodes conserved integral membrane transport protein **MmpL8** involved with cell wall and cell processes
Rv3825c	encodes polyketide synthase **Pks2** involved with lipid metabolism
Rv3879c	encodes ESX-1 secretion-associated protein **EspK** involved with cell wall and cell processes
Rv3883c	encodes membrane-anchored myosin **MycP1** involved with intermediate metabolism and respiration

274 **Figure legends**

275 Figure 1: Number of SNPs per *Maf* Lineage (175 L5 and 78 L6 genomes). *a*: Number of SNPs
276 between *Maf* genomes and the hypothetical MTBC ancestor (the median fixed SNPs of L5 (934) is
277 lower (W = 417, p-value < 2.2e-16) compared to L6 (1,039). *b*: Pairwise SNPs between genomes
278 within each lineage (the median of the pairwise SNPs is lower (W = 234, p-value < 2.2e-16) in L5
279 (212) compared to L6 (334). *c*: Whole genome average nucleotide diversity (π) between L5 and L6
280 (the mean diversity of L5 (0.000076) is significantly (non-overlapping 95% confidence intervals)
281 lower than L6 (0.000110). Error bars indicate 95% confidence intervals.

282

283 Figure 2: Phylogeny of Ghanaian *Maf* strains. (The maximum likelihood phylogenetic tree of 253
284 Ghanaian *Maf* isolates is based on 11,027 variable positions. The tree was rooted on *M. canettii* and
285 the confidence of nodes was assessed by bootstrapping 1000 pseudo replicates. Each lineage clade
286 is colored according to the conventional MTBC lineage color codes[1].

287

288 Figure 3: Averaged nucleotide diversity (π) of *Maf* within genes of eight functional categories. *epit*
289 – genes encoding human T cell epitopes, *esmac* – genes essential for growth in macrophages,
290 *intmedres* – genes involved with intermediate metabolism and respiration, *lipmet* – genes involved
291 with lipid metabolism, *virdetad* – genes involved with virulence, detoxification and adaptation,
292 *cwallproc* - genes involed with cell wall and cell processes, *regprot* – genes encoding regulatory
293 proteins and *infopath* – genes involved with information pathways. Error bars are indications of
294 95% confidence intervals.

295

296 Figure 4: Pairwise dN/dS of genes encoding human T cell epitopes and regulatory proteins in L5
297 and L6. Estimation of pairwise dN/dS of epitopes (*a*) and regulatory proteins (*b*) using the entire
298 147 L5 against the 67 L6 genomes. Estimation of pairwise dN/dS of epitopes (*c*) and regulatory

299 proteins (*d*) using the mean dN/dS values of 10 random samples (size =67, with replacement) of L5
300 against the 67 L6 genomes.
301
302 Figure 5: Number of human T cell epitopes with nonsynonymous SNPs (nsSNPs) stratified by *Maf*
303 lineage. No significant difference (X-squared = 1.487, df = 1, p-value = 0.22) between the number
304 of epitopes with nsSNPs among the 67 L6 genomes and L5 (mean values of 10 random samples of
305 size=67 with replacement).
306
307 Figure 6: Number of human T cell epitopes (*a*) and human T cell antigens (*b*) with amino acid
308 substitutions stratified by *Maf* lineage. Green represents L6-specific mutant antigens or epitopes.
309 Brown represents L5-specific mutant antigens or epitopes. Yellow represents antigens or epitopes
310 mutated in both L5 and L6 but at different loci with different amino acid substitutions.
311
312 Figure 7: Pairwise dN/dS of sequences encoding human T cell epitopes (*a*) and genes encoding
313 regulatory proteins (*b*) of L5 by patient ethnicity. L5 genomes from strains isolated from patients of
314 the Ewe ethnicity (15 genomes) against, average values of 10 random samples of size 15 of L5
315 genomes of isolates from Non-Ewe patients.
316
317

References

1. Gagneux, S. *et al.* Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 2869–2873 (2006).
2. Gagneux, S. & Small, P. M. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. *Lancet Infect. Dis.* **7,** 328–37 (2007).
3. Castets, M., Boisvert, H., Grumbach, F., Brunel, M. & Rist, N. Tuberculosis bacilli of the African type: preliminary note. *Rev. Tuberc. Pneumol. (Paris).* **32,** 179–84 (1968).
4. de Jong, B. C., Antonio, M. & Gagneux, S. Mycobacterium africanum-review of an important cause of human tuberculosis in West Africa. *PLoS Negl. Trop. Dis.* **4,** e744 (2010).
5. Comas, I. *et al.* Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat. Genet.* **45,** 1176–82 (2013).
6. Comas, I. *et al.* Population Genomics of Mycobacterium tuberculosis in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa. *Curr. Biol.* 3260–3266 (2015). doi:10.1016/j.cub.2015.10.061
7. Hershberg, R. *et al.* High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. *PLoS Biol.* **6,** e311 (2008).
8. Stucki, D. *et al.* Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat. Genet.* **48,** 1535–1543 (2016).
9. Asante-Poku, A. *et al.* Molecular epidemiology of Mycobacterium africanum in Ghana. *BMC Infect. Dis.* **16,** 385 (2016).
10. Yeboah-manu, D. *et al.* Genotypic diversity and drug susceptibility patterns among M. tuberculosis complex isolates from South-Western Ghana. *PLoS One* **6,** e21906 (2011).
11. Bold, T. D. *et al.* Impaired fitness of Mycobacterium africanum despite secretion of ESAT-6. *J. Infect. Dis.* **205,** 984–90 (2012).
12. Asante-Poku, A. *et al.* Mycobacterium africanum Is Associated with Patient Ethnicity in Ghana. *PLoS Negl. Trop. Dis.* **9,** e3370 (2015).

13. de Jong, B. C. *et al.* Progression to active tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia. *J. Infect. Dis.* **198,** 1037–43 (2008).

14. Homolka, S., Niemann, S., Russell, D. G. & Rohde, K. H. Functional genetic diversity among Mycobacterium tuberculosis complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. *PLoS Pathog.* **6,** e1000988 (2010).

15. Tientcheu, L. D. *et al.* Differences in T-cell responses between Mycobacterium tuberculosis and Mycobacterium africanum -infected patients. *Eur. J. Immunol.* **44,** 1387–1398 (2014).

16. Jong, B. C. De *et al.* Mycobacterium africanum elicits an attenuated T cell response to early secreted antigenic target, 6 kDa, in patients with tuberculosis and their household contacts. *J. Infect. Dis.* **193,** 1279–86 (2006).

17. Portevin, D. *et al.* Human macrophage responses to clinical isolates from the Mycobacterium tuberculosis complex discriminate between ancient and modern lineages. *PLoS Pathog.* **7,** e1001307 (2011).

18. Niemann, S., Merker, M., Kohl, T. & Supply, P. Impact of Genetic Diversity on the Biology of Mycobacterium tuberculosis Complex Strains. *Microbiol. Spectr.* **4,** (2016).

19. Källenius, G. *et al.* Evolution and clonal traits of Mycobacterium tuberculosis complex in Guinea-Bissau. *J. Clin. Microbiol.* **37,** 3872–3878 (1999).

20. Niobe-Eyangoh, S. N. *et al.* Genetic biodiversity of Mycobacterium tuberculosis complex strains from patients with pulmonary tuberculosis in Cameroon. *J. Clin. Microbiol.* **41,** 2547–53 (2003).

21. Dosso, M. *et al.* Primary resistance to antituberculosis drugs: a national survey conducted in Côte d'Ivoire in 1995–1996* for the Ivoirian Study Group on Tuberculosis Resistance ** Projet Santé Abidjan (Coopération S U M M A R Y. *INT J TUBERC LUNG DIS* **3,** 805–809 (1999).

22. Koro, F. K. *et al.* Population dynamics of tuberculous bacilli in cameroon as assessed by

370 spoligotyping. *J. Clin. Microbiol.* **51,** 299–302 (2013).

371 23. Gehre, F. *et al.* The first phylogeographic population structure and analysis of transmission
372 dynamics of M. africanum West African 1--combining molecular data from Benin, Nigeria
373 and Sierra Leone. *PLoS One* **8,** e77000 (2013).

374 24. Lawson, L. *et al.* A molecular epidemiological and genetic diversity study of tuberculosis in
375 Ibadan, Nnewi and Abuja, Nigeria. *PLoS One* **7,** e38409 (2012).

376 25. Yeboah-Manu, D. *et al.* Spatio-Temporal Distribution of Mycobacterium tuberculosis
377 Complex Strains in Ghana. *PLoS One* **11,** e0161892 (2016).

378 26. Comas, I. *et al.* Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily
379 hyperconserved. *Nat. Genet.* **42,** 498–503 (2010).

380 27. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43,** (2015).

381 28. Bos, K. I. *et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New
382 World human tuberculosis. *Nature* **514,** 494–497 (2014).

383 29. Kay, G. L. *et al.* Eighteenth-century genomes show that mixed infections were common at
384 time of peak tuberculosis in Europe. *Nat. Commun.* **6,** 6717 (2015).

385 30. Brosch, R. *et al.* A new evolutionary scenario for the Mycobacterium. *Proc. Natl. Acad. Sci.*
386 **99,** 3684–3689 (2002).

387 31. Mostowy, S., Cousins, D., Brinkman, J., Aranaz, A. & Behr, M. a. Genomic deletions
388 suggest a phylogeny for the Mycobacterium tuberculosis complex. *J. Infect. Dis.* **186,** 74–80
389 (2002).

390 32. Coscolla, M. & Gagneux, S. Consequences of genomic diversity in mycobacterium
391 tuberculosis. *Semin. Immunol.* **26,** 431–444 (2014).

392 33. Coscolla, M. *et al.* M. tuberculosis T Cell Epitope Analysis Reveals Paucity of Antigenic
393 Variation and Identifies Rare Variable TB Antigens Reveals Paucity of Antigenic Variation
394 and Identifies Rare Variable TB Antigens. *Cell Host Microbe* **18,** 538–548 (2015).

395 34. Pai, M. *et al.* Tuberculosis. *Nat. Rev. Dis. Prim.* **2,** 16076 (2016).

35. Brodin, P., Rosenkrands, I., Andersen, P., Cole, S. T. & Brosch, R. ESAT-6 proteins: Protective antigens and virulence factors? *Trends Microbiol.* **12,** 500–508 (2004).

36. Bigi, M. M. *et al.* Polymorphisms of 20 regulatory proteins between Mycobacterium tuberculosis and Mycobacterium bovis. *Microbiol. Immunol.* (2016). doi:10.1111/1348-0421.12402

37. Raman, S., Hazra, R., Dascher, C. C. & Husson, R. N. Transcription regulation by the Mycobacterium tuberculosis alternative sigma factor SigD and its role in virulence. *J. Bacteriol.* (2004). doi:10.1128/JB.186.19.6605-6616.2004

38. Gonzalo-asensio, J., Malaga, W., Pawlik, A. & Astarie-dequeker, C. Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. **111,** (2014).

39. Peirs, P., Parmentier, B., De Wit, L. & Content, J. The Mycobacterium bovis homologous protein of the Mycobacterium tuberculosis serine/threonine protein kinase Mbk (PknD) is truncated. *FEMS Microbiol. Lett*. **188,** 135–9 (2000).

40. Saïd-Salim, B., Mostowy, S., Kristof, A. S. & Behr, M. A. Mutations in Mycobacterium tuberculosis Rv0444c, the gene encoding anti-SigK, explain high level expression of MPB70 and MPB83 in Mycobacterium bovis. *Mol. Microbiol.* **62,** 1251–1263 (2006).

41. de Jong, B. C. *et al.* Differences between tuberculosis cases infected with Mycobacterium africanum, West African type 2, relative to Euro-American Mycobacterium tuberculosis: an update. *FEMS Immunol. Med. Microbiol.* **58,** 102–5 (2010).

42. Otchere, I. D. *et al.* Detection and characterization of drug-resistant conferring genes in Mycobacterium tuberculosis complex strains: A prospective study in two distant regions of. *Tuberculosis* **99,** 147–154 (2016).

43. Warren, R. M. *et al.* Molecular evolution of Mycobacterium tuberculosis: phylogenetic reconstruction of clonal expansion. *Tuberculosis* **81,** 291–302 (2001).

44. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J. Clin. Microbiol.* **35,** 907–14 (1997).

422 45. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina
423 sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
424 46. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
425 *Bioinformatics* **25,** 1754–1760 (2009).
426 47. Koboldt, D. C. *et al.* VarScan 2 : Somatic mutation and copy number alteration discovery in
427 cancer by exome sequencing VarScan 2 : Somatic mutation and copy number alteration
428 discovery in cancer by exome sequencing. *Genome Res.* **22,** 568–576 (2012).
429 48. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide
430 polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-
431 2; iso-3. *Fly (Austin).* **6,** 80–92 (2012).
432 49. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
433 phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
434 50. Lew, J. M., Kapopoulou, A., Jones, L. M. & Cole, S. T. TubercuList - 10 years after.
435 *Tuberculosis* **91,** 1–7 (2011).
436 51. Popescu, A.-A., Huber, K. T. & Paradis, E. ape 3.0: New tools for distance-based
437 phylogenetics and evolutionary analysis in R. *Bioinformatics* **28,** 1536–1537 (2012).
438 52. R Core Team. R: A language and environment for statistical computing version 3.2.3. *R*
439 *Foundation for Statistical Computing, Vienna, Austria* (2015). Available at:
440 https://www.coursehero.com/file/pe12cv/Code-R-version-323-2015-12-10-Wooden-
441 Christmas-Tree-Copyright-C-2015-The-R/. (Accessed: 26th July 2017)
442 53. Gardner, M. J. & Altman, D. G. Statistics in Medicine Confidence intervals rather than P
443 values: estimation rather than hypothesis testing. *Br. Med. J. (Clin. Res. Ed).* **292,** 746–750
444 (1986).
445 54. Nakagawa, S. & Cuthill, I. C. Effect size, confidence interval and statistical significance: A
446 practical guide for biologists. *Biol. Rev.* **82,** 591–605 (2007).
447 55. Charif D, L. J. *SeqinR 1.0-2: a contributed package to the R-project for statistical computing*

devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto MERH, Vendruscolo M, editors. Structural approaches to sequence evolution: Molecules, networks, populations. Springer Verlag (2007).

56. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30,** 2725–9 (2013).

454 **Acknowledgements**

455 Bacterial Isolation and DNA preparations were done in the Biosafety level 3 facility at the Noguchi
456 Memorial Institute for Medical Research, University of Ghana. Bioinformatics analyses were
457 performed using the scientific computing core (sciCORE) at the University of Basel and the
458 computing facility of the Wellcome Trust Sanger Institute, Genome Campus, Cambridge
459 University. This work was supported by the Wellcome Trust Intermediate Fellowship awarded to
460 DYM (Grant Number 097134/Z/11/Z) and by the Swiss National Science Foundation (grants
461 310030_166687, IZRJZ3_164171 and IZLSZ3_170834), the European Research Council (309540-
462 EVODRTB) and SystemsX.ch.

463

464

465 **Author contributions**

466 Conceived the idea: DYM, SG

467 Designed experiments: DYM, SG, IDO, MC, SRH, JP

468 Contributed reagents and performed experiments: IDO, MC, AAP, LSB, MC, SOW, AF, CL, GAA,
469 AIY, AB, SYA, PA, CL, DB, SB, FG, PB, TK, SN, MA, SN, CB, BCDJ, JP and SRH

470 Analysed Data: IDO, MDC, SRH, LSB, SG, DYM

471 Wrote manuscript: IDO, MC, SG and DYM

472 All authors critically reviewed the manuscript

473

474 **Competing financial interests**

475 None declared.

476

477 **Data availability.**

478 All the analyzed and/or generated data in this study are included in this article and its

supplementary information files. Whole genome sequence reads have been submitted to the EMBL-EBI European Nucleotide Archive (ENA) Sequence Read Archive (SRA) with accession numbers provided in the supplementary document attached (Supplementary data S8).

Supplementary Information

Supplementary figure S1: Flow chart of Ghanaian *Maf* genomes used for the study. Sites of DNA sequencing as well as number of genomes used for each specific analysis are indicated.

Supplementary figure S2: Map of Ghana showing the regional distribution of *Maf* isolates. The sizes of pie charts correspond to number of isolates from the respective regions.

Supplementary figure S3: Pairwise dN/dS of genes of six functional categories using all the 147 L5 against the 67 L6 genomes. *esmac* – genes essential for growth in macrophages, *intmedres* – genes involved with intermediate metabolism and respiration, *lipmet* – genes involved with lipid metabolism, *virdetad* – genes involved with virulence, detoxification and adaptation, *cwallproc* - genes involved with cell wall and cell processes and *infopath* – genes involved with information pathways

Supplementary figure S4: Number of non-redundant SNPs within human T cell epitopes stratified by *Maf* lineage. No difference (X-squared = 0.0055391, df = 1, p-value = 0.9407) between the proportion of nsSNPs between L6 (67 genomes) and L5 (mean values of 10 random samples of size = 67 with replacement)

Supplementary table S5: Mutated Human T cell antigens and epitopes of *Maf* with mutations stratified by lineage

505 Supplementary figure S6: Number of epitopes with and without nsSNPs stratified by Lineage 5
506 with patient ethnicity (*a*). Number of non-redundant SNPs in epitopes stratified by ethnicity of L5
507 infected patients (*b*). (The values for the non-Ewe associated L5 are mean estimates of 10 random
508 samples of size 15 with replacement)
509
510 Supplementary figure S7: Uniform distribution of L5 isolated from ethnic Ewe TB patients among
511 the L5 clade.
512
513 Supplementary data S8: Genomes and accession numbers of *Maf* from Ghana.
514

Methods

Ethical Statement and Participant Enrollment

The study and its protocols were reviewed by the Scientific and Technical Committee and approved by the Institutional Review Board (IRB) of the Noguchi Memorial Institute for Medical Research, Legon-Ghana with Federal Wide Assurance number FWA00001824.

Mycobacterium africanum Strains

Isolates used for this study were cultivated from July 2007 to November 2014 in Ghana[9,12], West Africa, involving consecutive sputum smear positive pulmonary TB cases recruited from two different studies with isolates spanning four regions of Ghana (three in the south and one from the North) (supplementary figure S2).

Mycobacterial Sub-Culturing and Chromosomal DNA Extraction

Mycobacterium africanum strains were revived by sub-culturing on Lowenstein Jensen (LJ) slants; one supplemented with 0.4% sodium pyruvate the other with glycerol to enhance the growth of Lineage 5 and Lineage 6 strains of respectively. The cultures were incubated at 37 °C and monitored regularly until growth was observed. When confluence was achieved, five loops full of colonies were fetched into 2 mL cryo-vials containing 1 mL of sterile nuclease-free water, heat-inactivated at 98 °C for 60 minutes for DNA extraction using a hybrid DNA extraction protocol[42]. The isolates were confirmed MTBC by PCR amplification of IS*6110*, genotyped as *Maf* by large sequence polymorphism (LSPs) detecting region of difference (RD) 9 and 12[43]. Lineage identification was achieved by spoligotyping as previously described[44]. Strains confirmed as belonging either L5 or L6 were sequenced by the illumina platform at the Wellcome Trust Sanger Institute, United Kingdom.

DNA Sequencing, Mapping of Sequence Reads, Variance Calling and Generation of Whole

Genome Fasta files

Samples were sequenced as multiplexed libraries on the Illumina HiSeq platform to produce paired end reads of 125 nt in length. Genomes provided by the Research Center Borstel was obtained by sequencing DNA libraries prepared with the Nextera XT kit and run on Illumina MiSeq (250 and 300 bp, paired end) and NextSeq (150 bp, paired end) according to the manufacturer's instruction (Illumina, San Diego, USA). The FastQ files containing the raw paired-end reads were processed using a python pipeline developed in house as follows. The reads were first adapter- and quality-trimmed with Trimmomatic v0.33[45]. Reads lower than 20 bp were not kept for the downstream analysis. Overlapping paired-end reads were then merged with SeqPrep (https://github.com/jstjohn/SeqPrep). The resulting filtered reads were mapped to a hypothetical reconstructed MTBC ancestor[26] with BWA v0.7.12[46]. Duplicated reads were marked by the MarkDuplicates module of Picard v 2.1.1 (https://github.com/broadinstitute/picard). The RealignerTargetCreator and IndelRealigner modules of GATK v.3.4.0 (https://software.broadinstitute.org/gatk/download/archive) were used to perform local realignment of reads around indels. SNPs were called with Samtools v1.2 (https://sourceforge.net/projects/samtools/files/samtools/1.2/) and VarScan v2.4.1[47] using the following thresholds: minimum mapping quality of 20, minimum base quality at a position of 20 and minimum read depth at a position of 7X. SNPs were considered fixed at a frequency of $\geq$ 90% and alleles were considered ancestral when the SNP frequency was $\leq$ 10%. Furthermore, SNPs were called only if the alternative basecall was supported by at least five reads and without strand bias. All variants were annotated using snpEff v4.11[48], in accordance with the *M. tuberculosis* H37Rv reference annotation (AL123456.3). SNPs falling in regions with at least 50 bp identity to other regions in the genome were excluded from the analysis.

Generation of Variable Positions and Phylogenetic Analysis

The variable SNPs alignment was obtained by concatenating the SNP calls present in the variant

567 calling file of each genome, using the IUPAC nucleotide ambiguity codes for heterozygous calls. A
568 position was considered variable if at least one genome had a SNP at that position. Called deletions
569 and positions not called according to the minimum threshold of 7 were encoded as gaps. Positions
570 for which the proportion of gaps exceeded 50% were excluded from the alignment. Maximum
571 likelihood phylogeny of the variable positions with 1000 bootstraps was then generated using
572 RAxML version 8.2.3[49] with GTR substitution matrix and other default settings with the final tree
573 evaluated and optimized under GAMMA with accuracy of 0.1 Log likelihood units. The best tree
574 was then, rooted on *M. canettii* and annotated using figtree
575 (http://www.webcitation.org/getfile?fileid=27177ee8dd2f34cfd254b9c5e6c6fdf4b65329f6).
576
577 **Comparative genomics analysis of isolates using genes encoding proteins of 8 functional**
578 **categories.**
579 Experimentally confirmed human MTBC T cell epitope (1,226 epitopes) sequences (spanning 304
580 antigens with some overlapping sequences) retrieved from the Immune Epitope Database (IEDB),
581 tested in human T cell assays, with no major histocompatibility complex (MHC) restrictions and
582 have genomic coordinates in the H37Rv reference strain[32,8] were *in silico* extracted from the fasta
583 whole genome files and concatenated excluding sequence redundancy using customized bash
584 algorithms. Complementary sequences of epitopes encoded by the reversed strand were first
585 transcribed before the concatenation to have all the sequences in the same direction. In addition,
586 MTBC genes of other seven functional categories namely those encoding regulatory proteins
587 (regprot; 196), genes involved with lipid metabolism (limpet; 267), genes involved with
588 intermediate metabolism (intmedres; 917), genes involved with virulence, detoxification and
589 adaptation (virdetad; 216), genes involved with information pathways (infopath; 234), genes
590 involved with cell wall and cell processes (cwallproc; 768) and genes essential for growth in
591 macrophages (esmac; 125) according to the tuberculist database[50] were also retrieved and
592 concatenated as described above excluding genes involved with drug resistance.
593

Estimation of Pairwise Nucleotide Diversity

Pairwise SNP distances of the whole genome excluding sites associated with drug resistance, concatenates of T cell epitopes and the genes of other seven functional categories were calculated with the *dna.dist* function of *ape* package[51] of R version 3.2.3[52] as previously described[8]. Average pairwise nucleotide diversity per site (π) and confidence intervals for the π was calculated as previously described[8] and plotted with *ggplot2* package implemented in R. The upper and lower levels of confidence were attained by estimating the 97.5th and 2.5th quantiles of the π distribution obtained by bootstrapping (1000 replicates) as previously described[8]. Non-overlapping confidence intervals of π were taken as evidence of statistically significant differences[53,54]. Details of the algorithm for this analysis are available upon request.

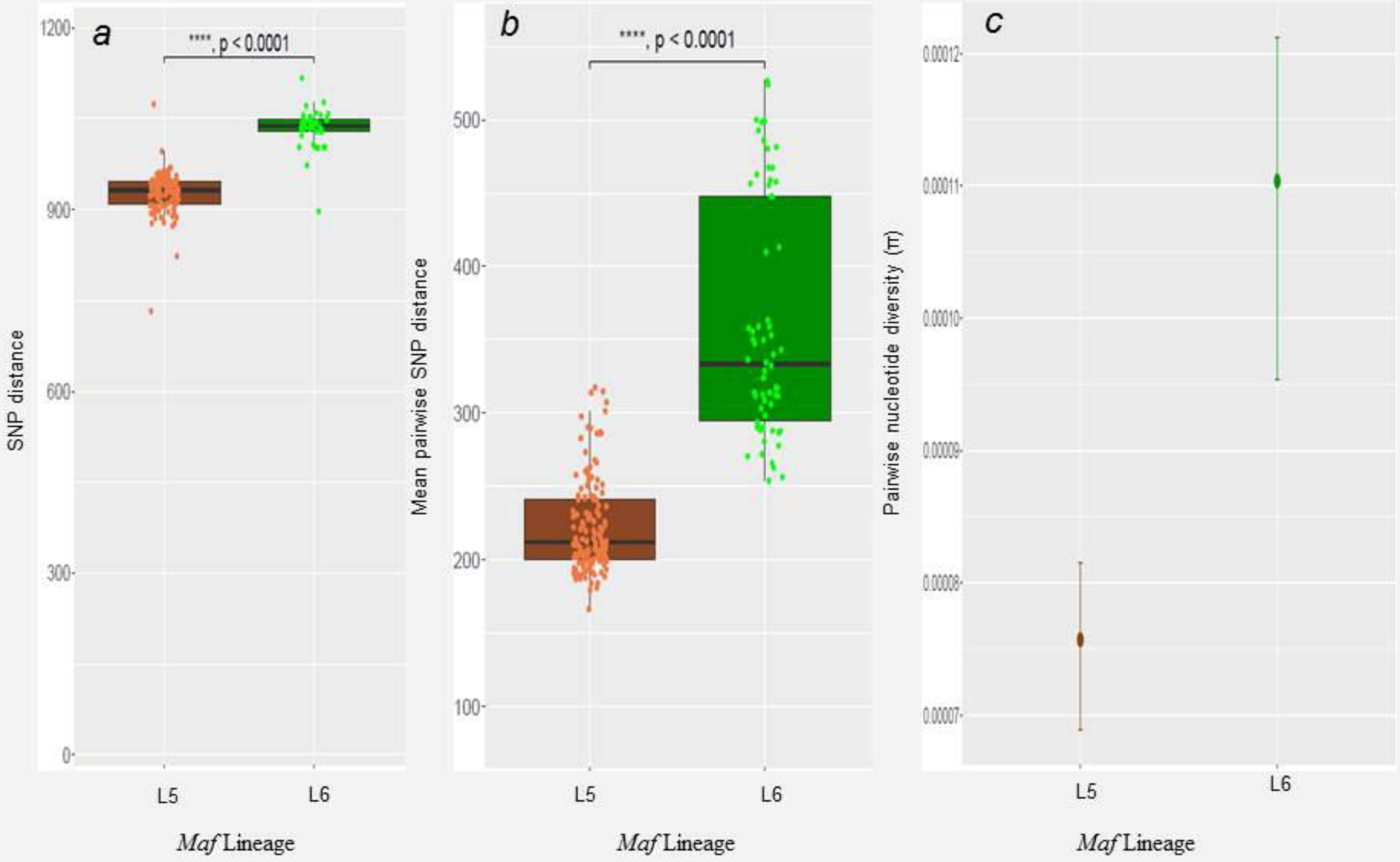
Estimation of Pairwise dN/dS

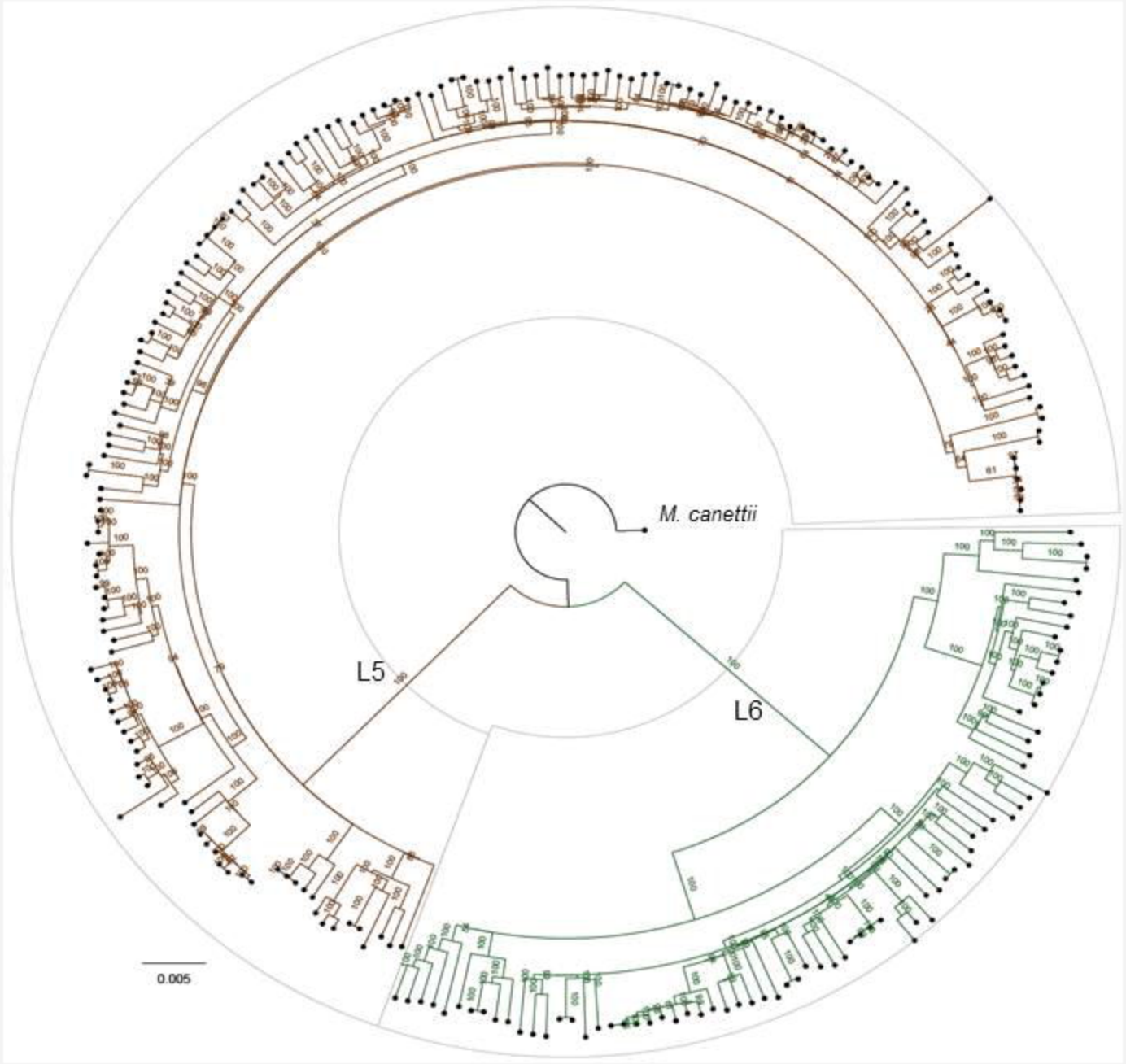
The concatenates of the human T cell epitopes and the other genes of seven functional categories were also used for estimation of dN/dS ratios stratified by lineage. As a follow up, dN/dS of T cell epitopes and regulatory proteins were also estimated for 15 L5 genomes from Ewe TB patients and 77 from non-Ewe TB patients. The dN/dS estimates were calculated with all polymorphic sites within each lineage using the *kaks* function of the *seqinr* packgage[55] as previously described[8] and box plotted using *ggplot2* package in in R version 3.2.3. Statistical difference of the estimates between the *Maf* lineages was accessed using the non-parametric Wilcoxon rank-sum tests with continuity correction in R version 3.4.0.

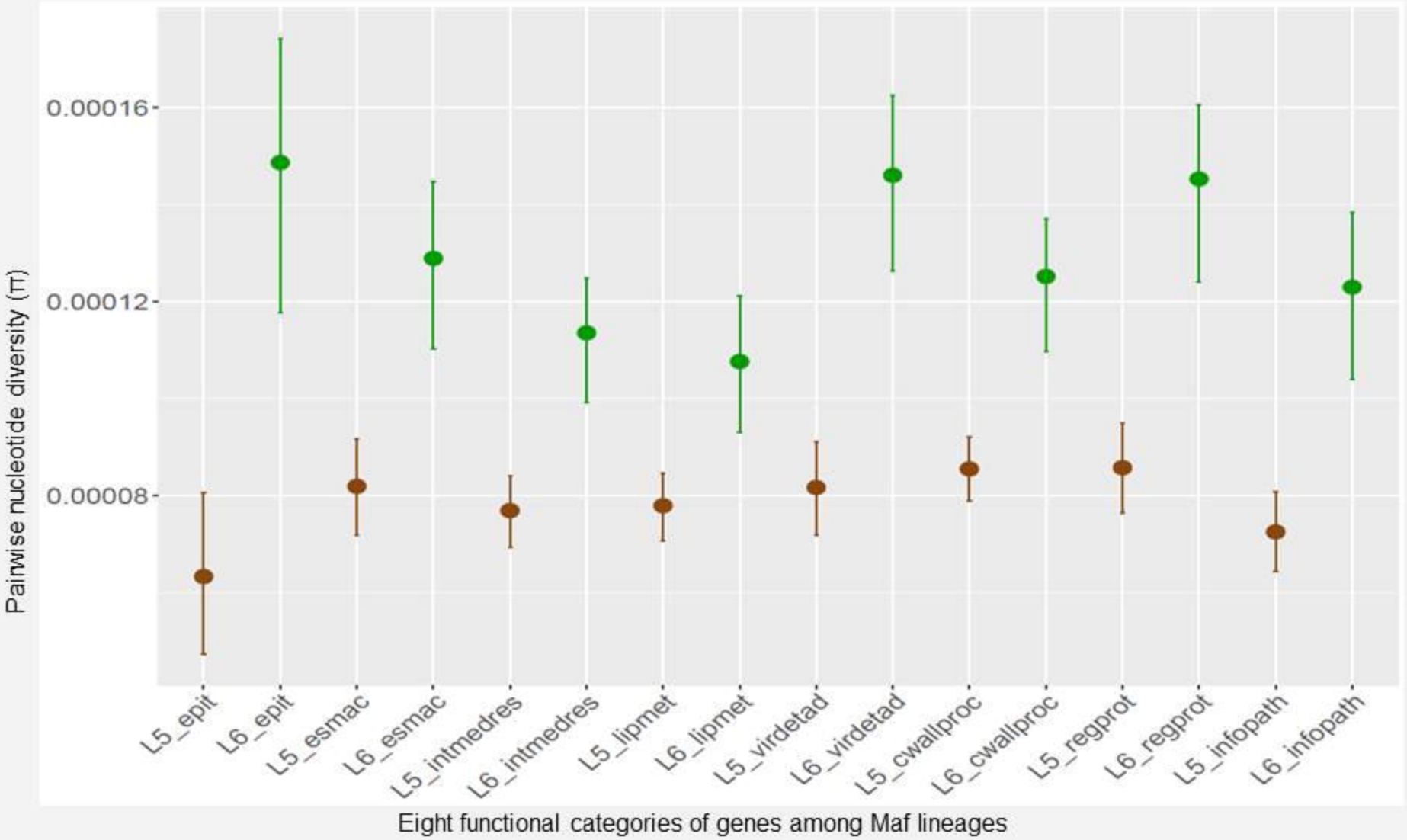
Human T cell Epitopes with Non-Synonymous SNPs and Count of Non-Redundant SNPs

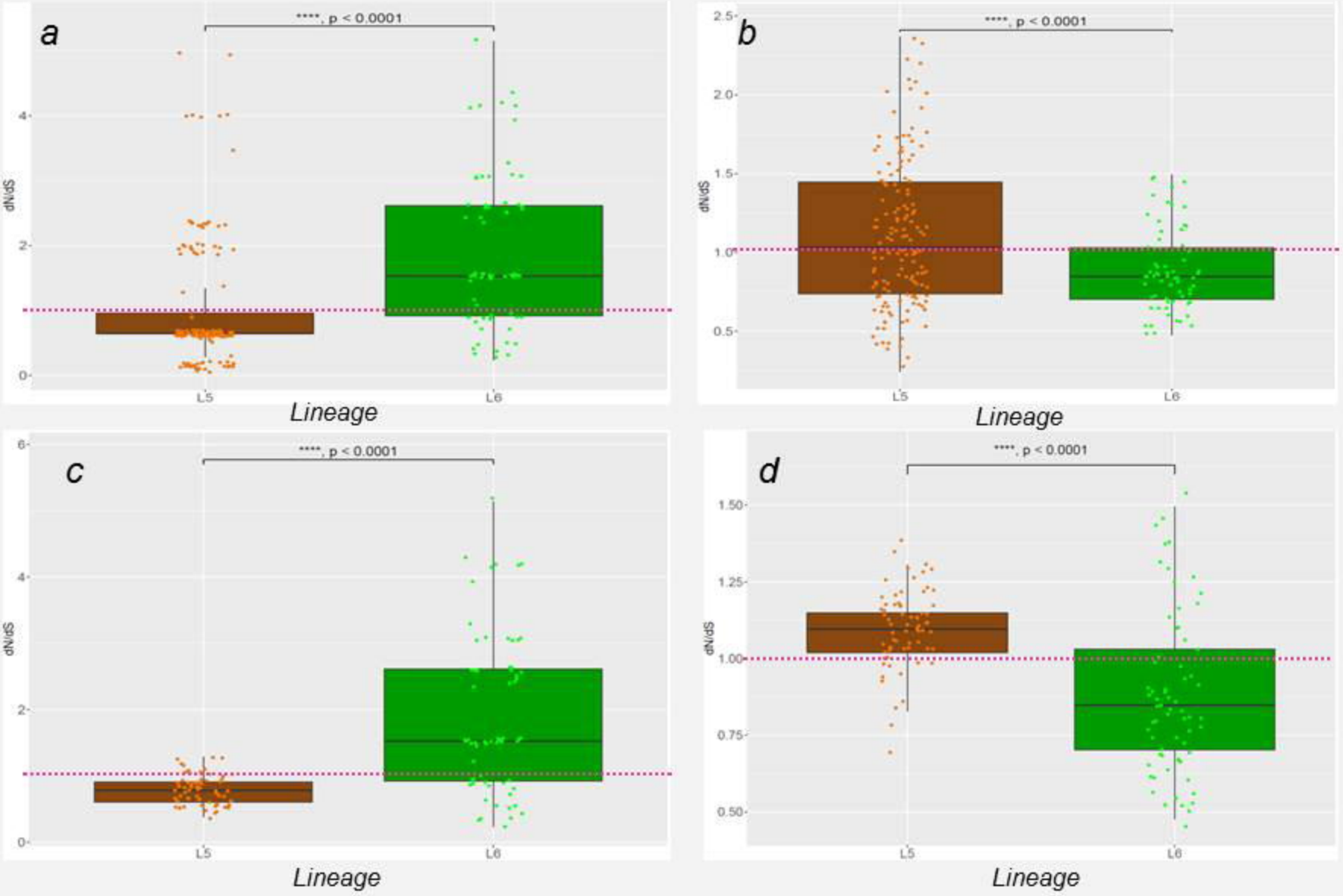
Synonymous and non-synonymous mutations within the coordinates of each epitope were extracted from the variant calling file (VCF) obtained for each genome. The specific human T cell epitopes with non-synonymous SNPs were compared between the *Maf* lineages for lineage-specific mutated epitopes and *Maf*-specific mutated epitopes.

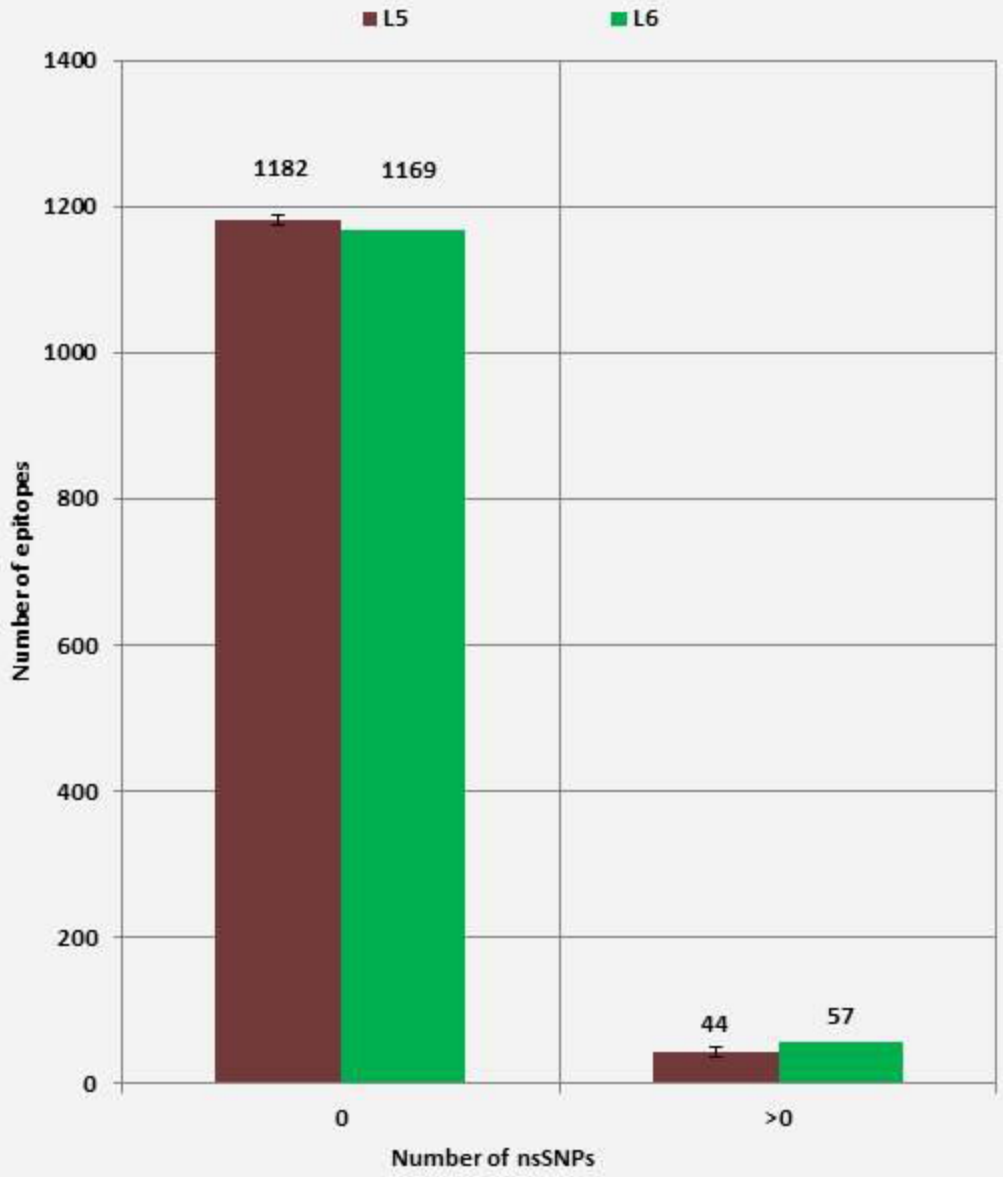
620 Furthermore, the number of pairwise non-redundant SNPs was estimated for the *Maf* lineages (67
621 L6 genomes and the 10 random samples of L5 of equal size as L6) as well as L5 genomes stratified
622 by patient ethnicity (15 L5 from Ewe patients and 10 random samples of L5 from non-Ewe TB
623 patients of size 15) using Mega6[56]. The number of SNPs per each group was plotted and compared
624 between the groups using the fisher's exact test for statistical significance in R version 3.2.3.
625

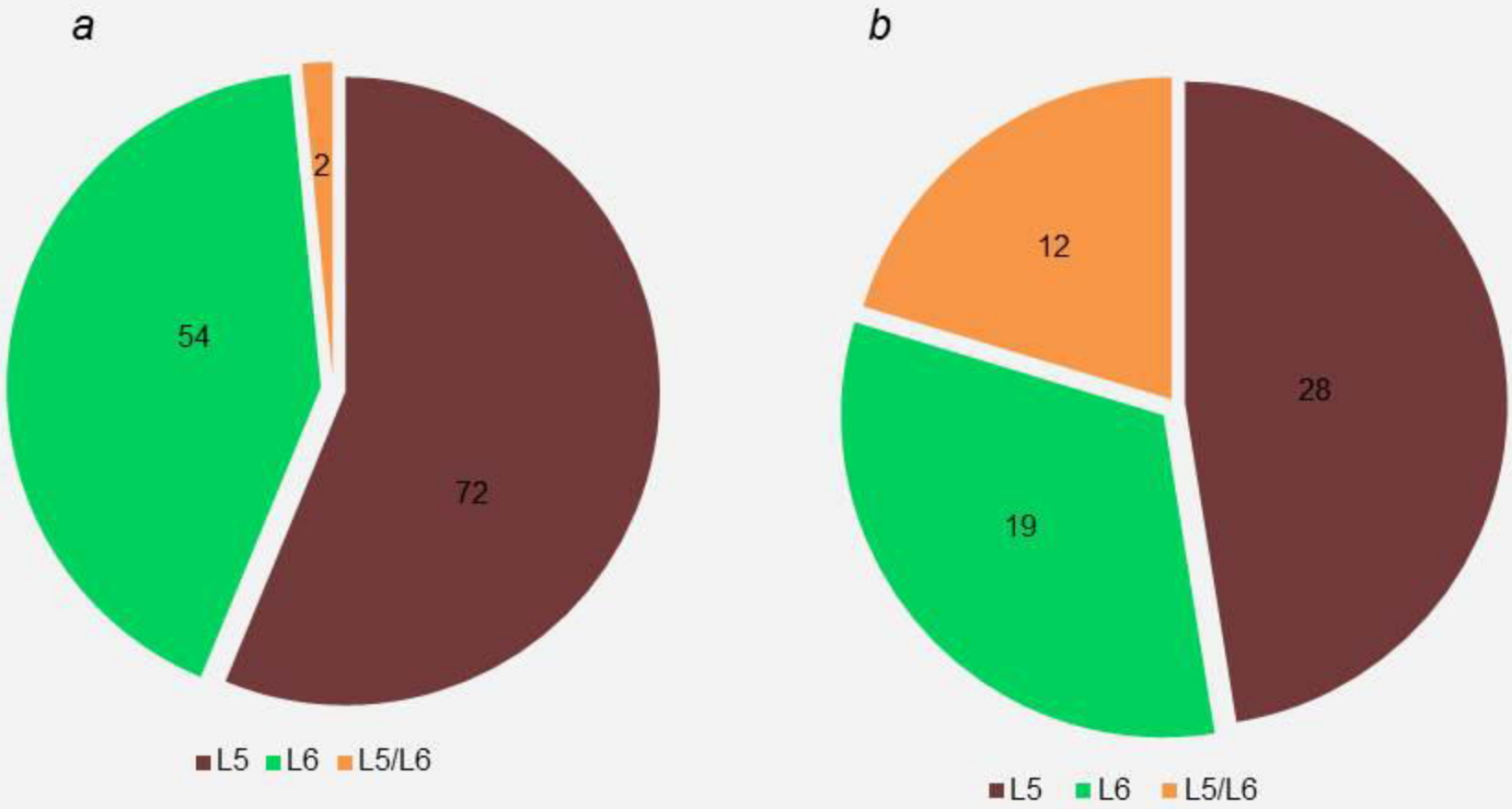

a
****, p < 0.0001
SNP distance
1200
900
600
300
0
L5
L6
Maf Lineage
b
****, p < 0.0001
Mean pairwise SNP distance
500
400
300
200
100
L5
L6
Maf Lineage
c
Pairwise nucleotide diversity (π)
0.00012
0.00011
0.00010
0.00009
0.00008
0.00007
L5
L6
Maf Lineage


M. canettii
L5
L6
0.005

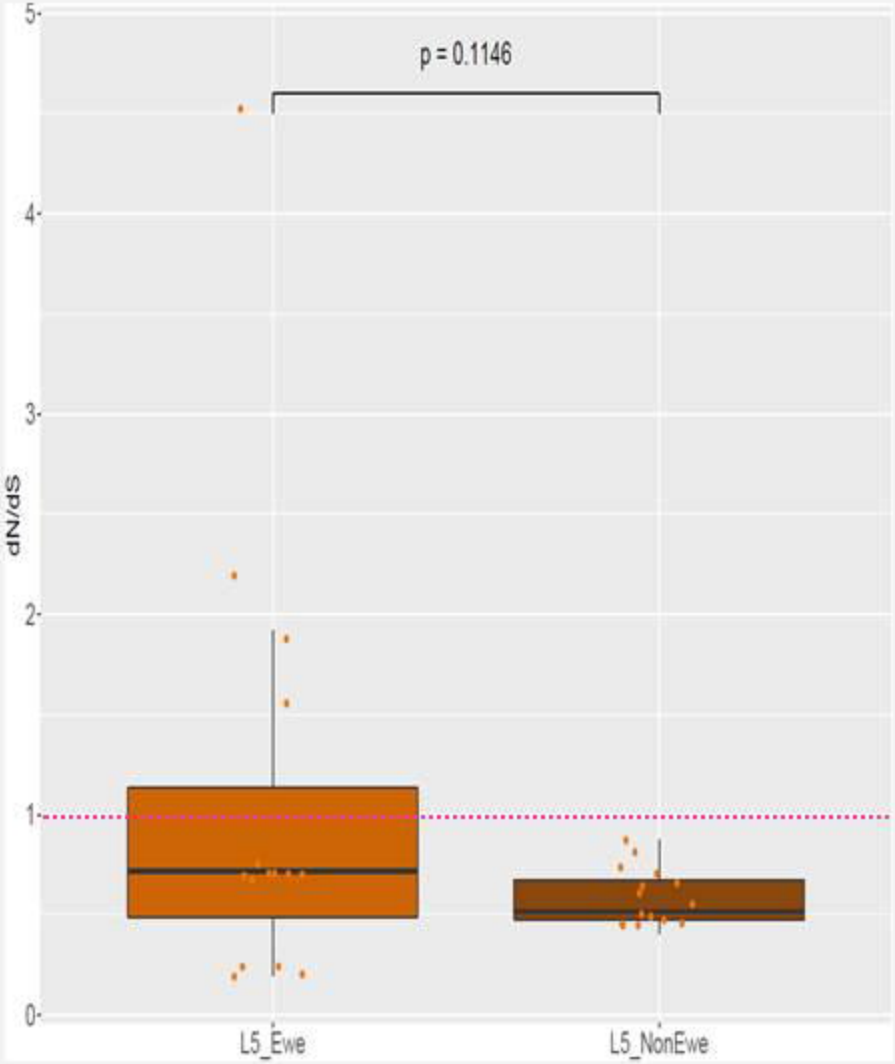

Pairwise nucleotide diversity (π)
0.00016
0.00012
0.00008
L5_epit
L6_epit
L5_esmac
L6_esmac
L5_intmedres
L6_intmedres
L5_lipmet
L6_lipmet
L5_virdetad
L6_virdetad
L5_cwallproc
L6_cwallproc
L5_regprot
L6_regprot
L5_infopath
L6_infopath
Eight functional categories of genes among Maf lineages


a
****, p < 0.0001
dN/dS
0
2
4
L5
L6
Lineage
b
****, p < 0.0001
dN/dS
0.5
1.0
1.5
2.0
2.5
L5
L6
Lineage
c
****, p < 0.0001
dN/dS
0
2
4
6
L5
L6
Lineage
d
****, p < 0.0001
dN/dS
0.50
0.75
1.00
1.25
1.50
L5
L6
Lineage


L5
L6
1400
1200
1000
800
600
400
200
0
Number of epitopes
1182
1169
44
57
0
>0
Number of nsSNPs

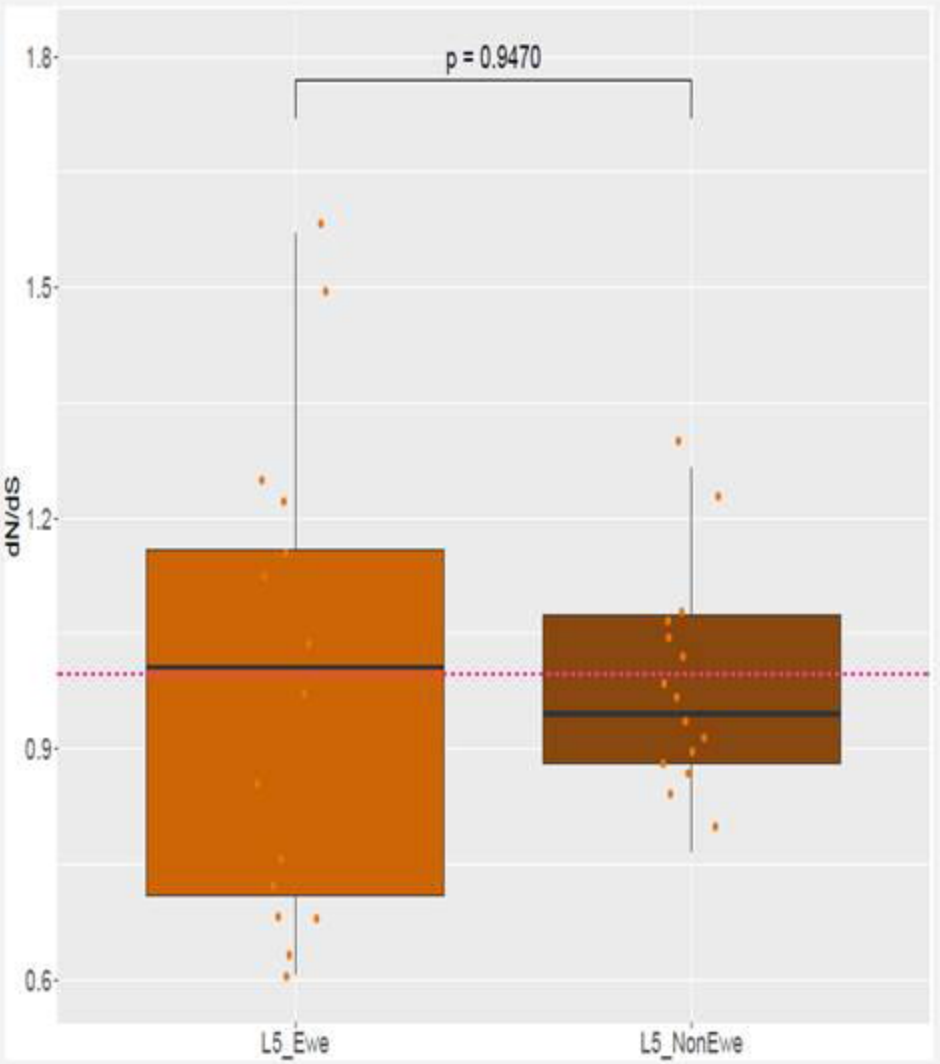

a
2
54
72
L5
L6
L5/L6
b
12
28
19
L5
L6
L5/L6

a

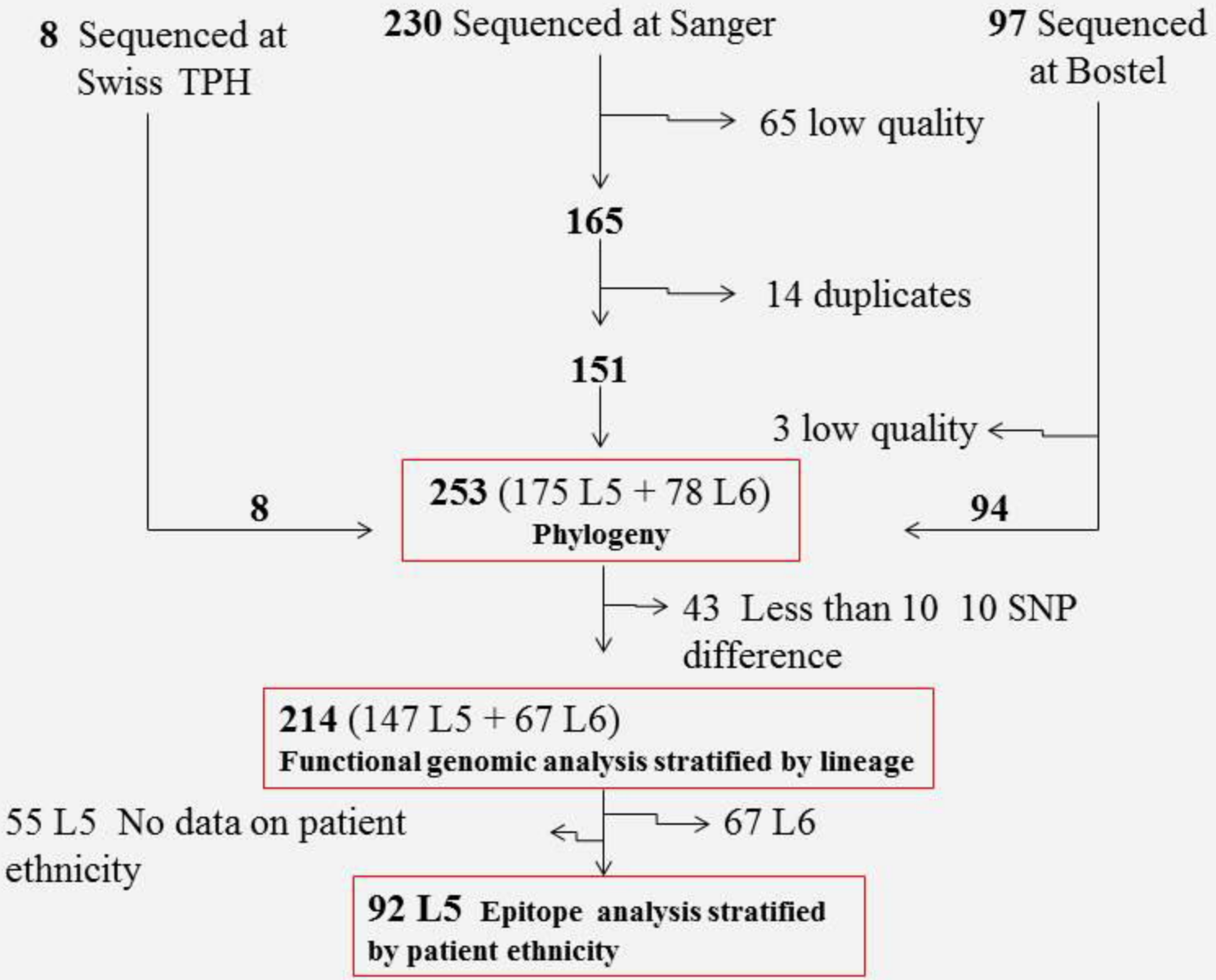


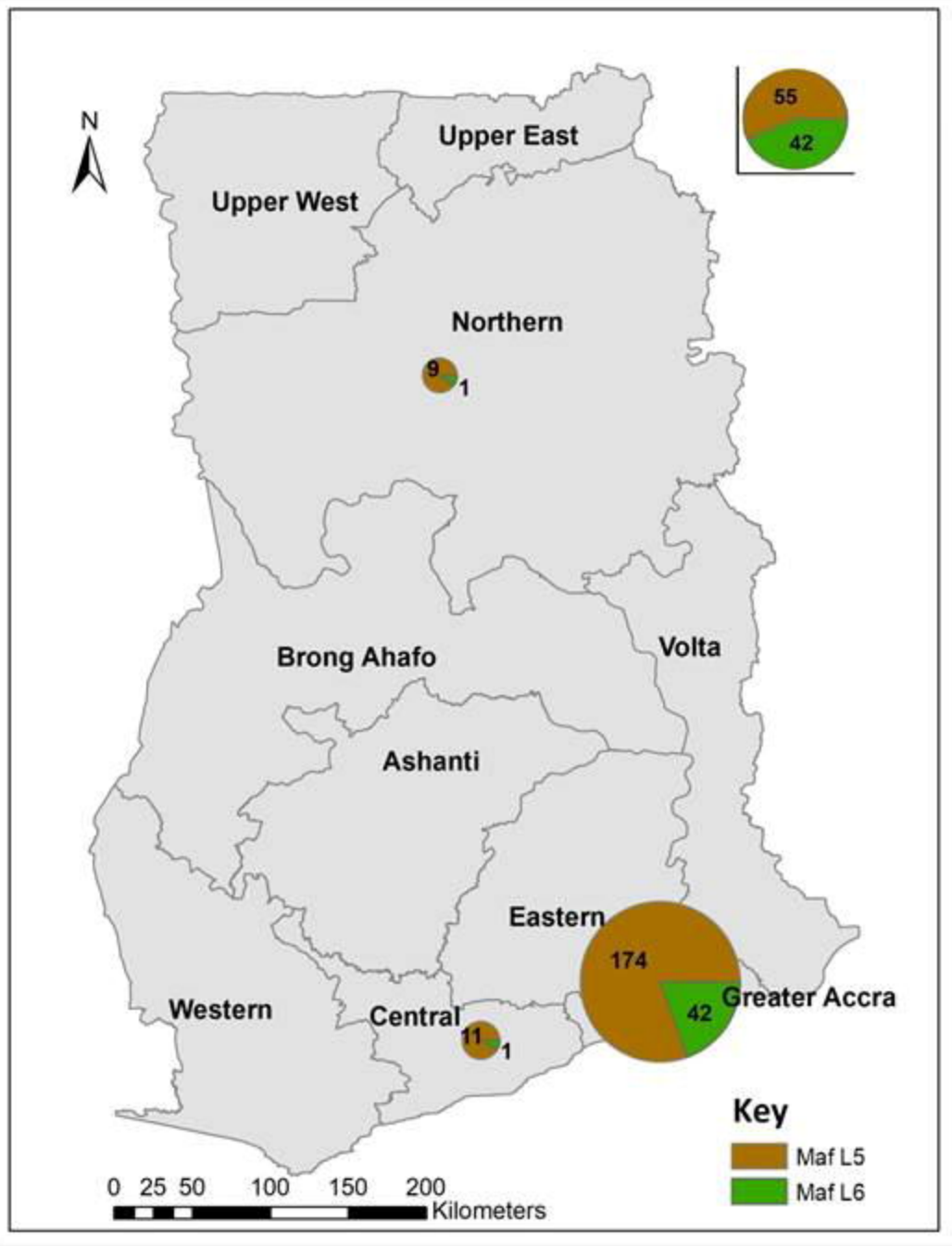
L5 with Ethnicity

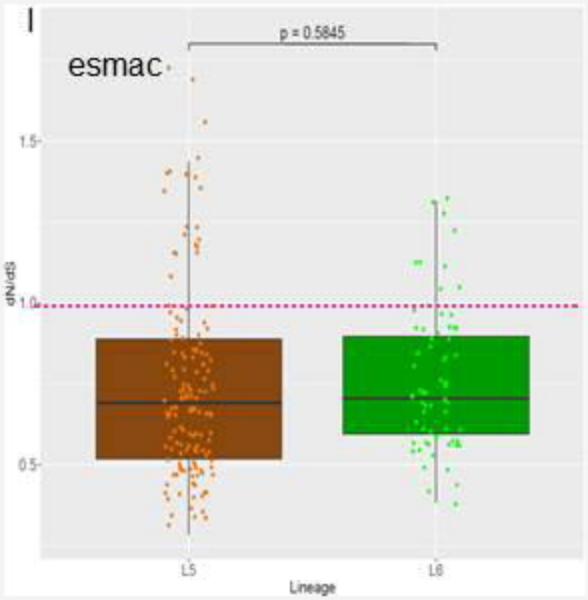
b

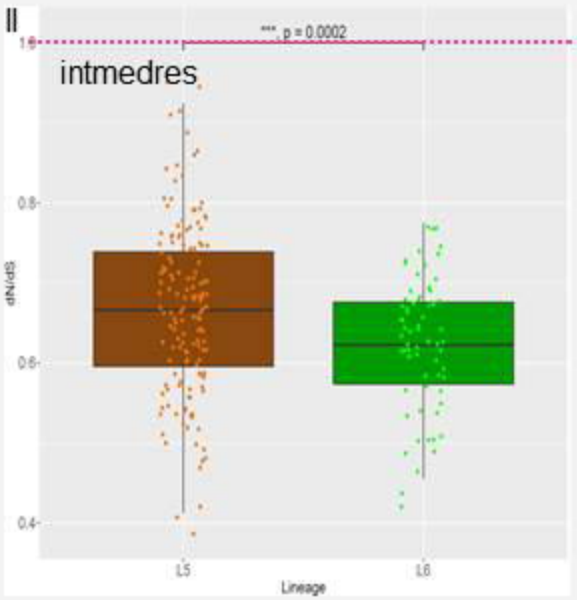


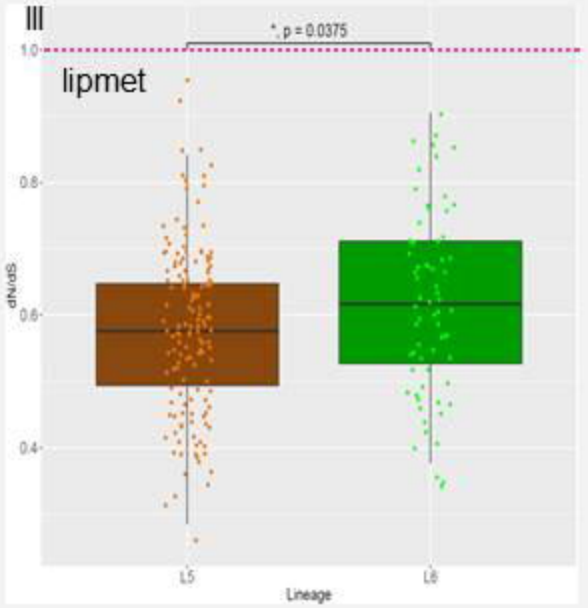
L5 with Ethnicity

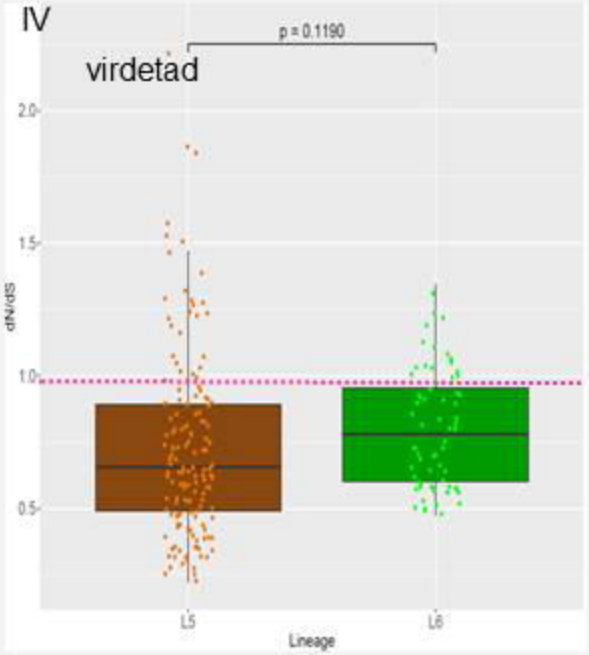

8 Sequenced at Swiss TPH
230 Sequenced at Sanger
97 Sequenced at Bostel
65 low quality
165
14 duplicates
151
3 low quality
8
253 (175 L5 + 78 L6)
Phylogeny
94
43 Less than 10 10 SNP difference
214 (147 L5 + 67 L6)
Functional genomic analysis stratified by lineage
55 L5 No data on patient ethnicity
67 L6
92 L5 Epitope analysis stratified by patient ethnicity

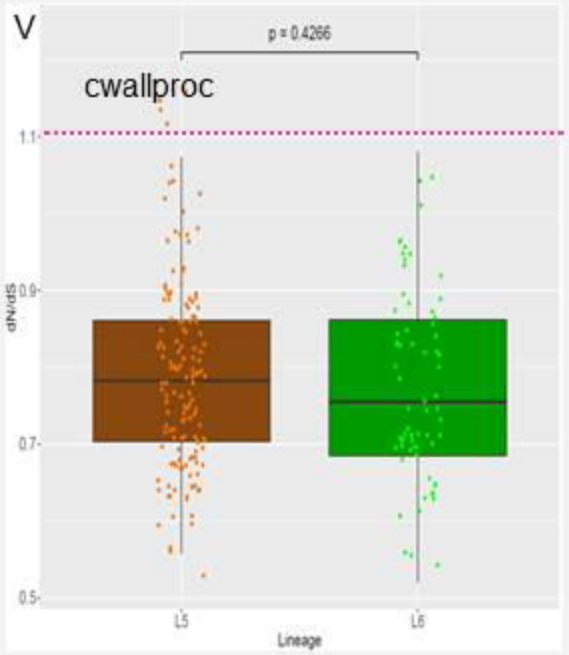

N
Upper East
Upper West
Northern
Brong Ahafo
Volta
Ashanti
Eastern
Western
Central
Greater Accra
55
42
9
1
174
42
11
1
Key
Maf L5
Maf L6
0 25 50 100 150 200
Kilometers

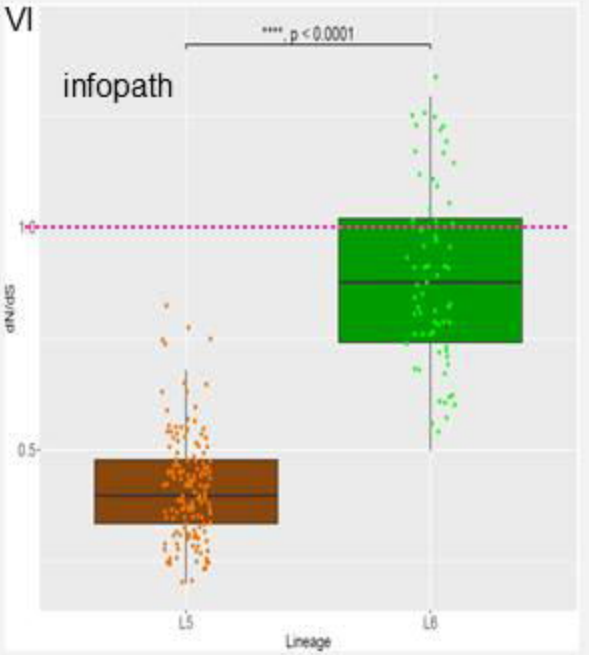

I
esmac
p = 0.5845
dN/dS
1.5
1.0
0.5
L5
L6
Lineage

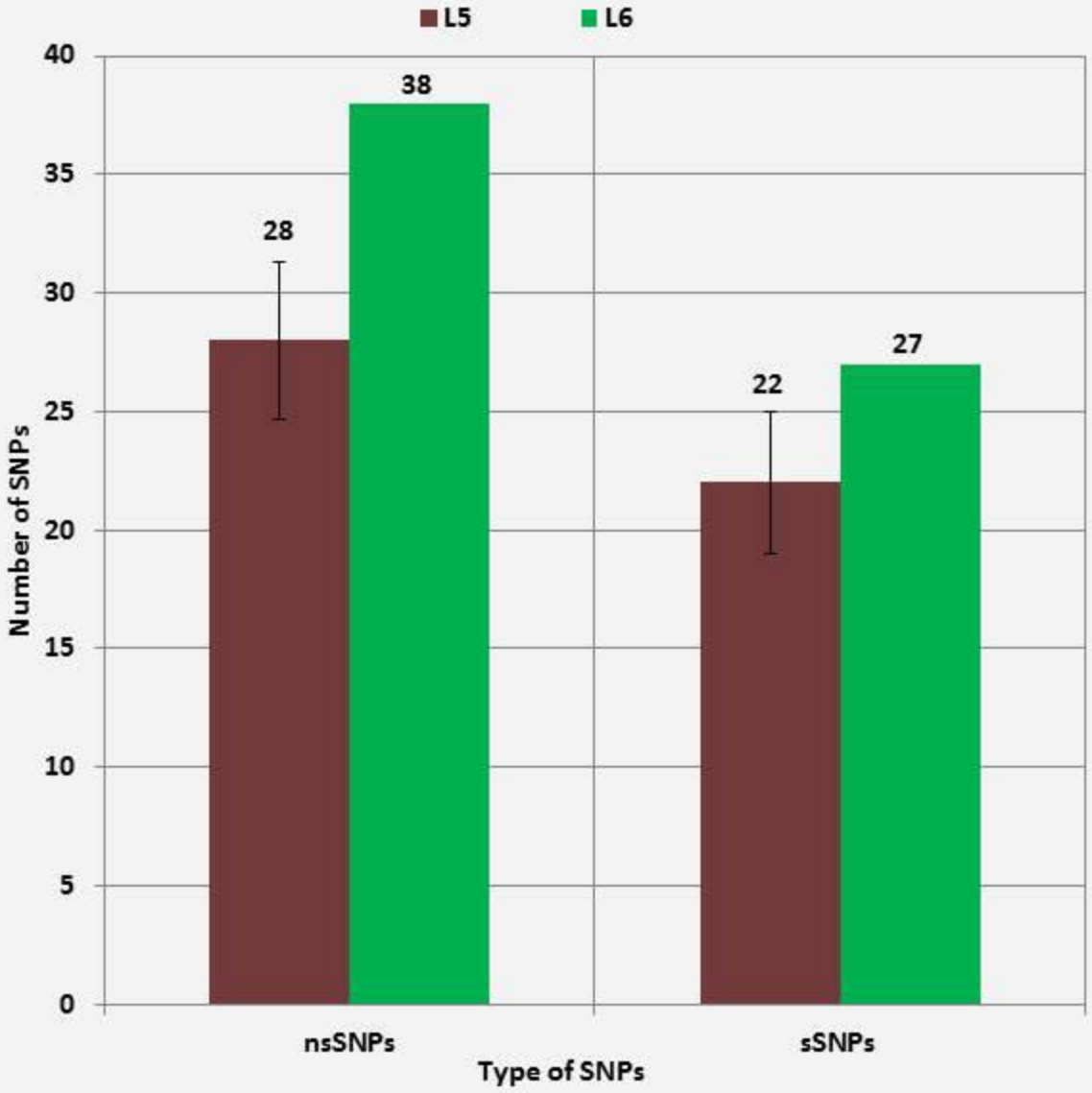

II
intmedres
***, p = 0.0002
dN/dS
1.0
0.8
0.6
0.4
L5
L6
Lineage

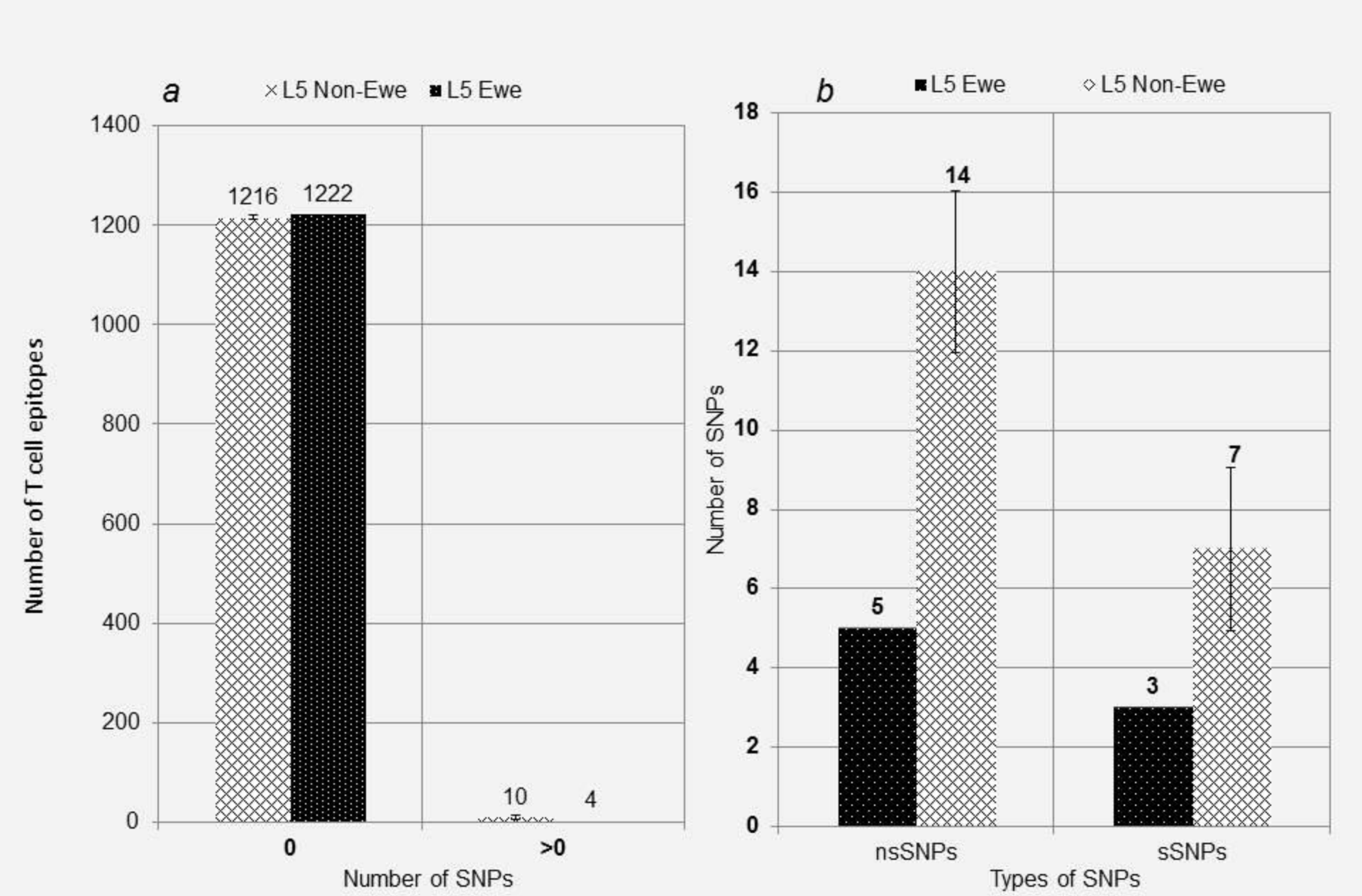

III
lipmet
*, p = 0.0375
dN/dS
1.0
0.8
0.6
0.4
L5
L6
Lineage

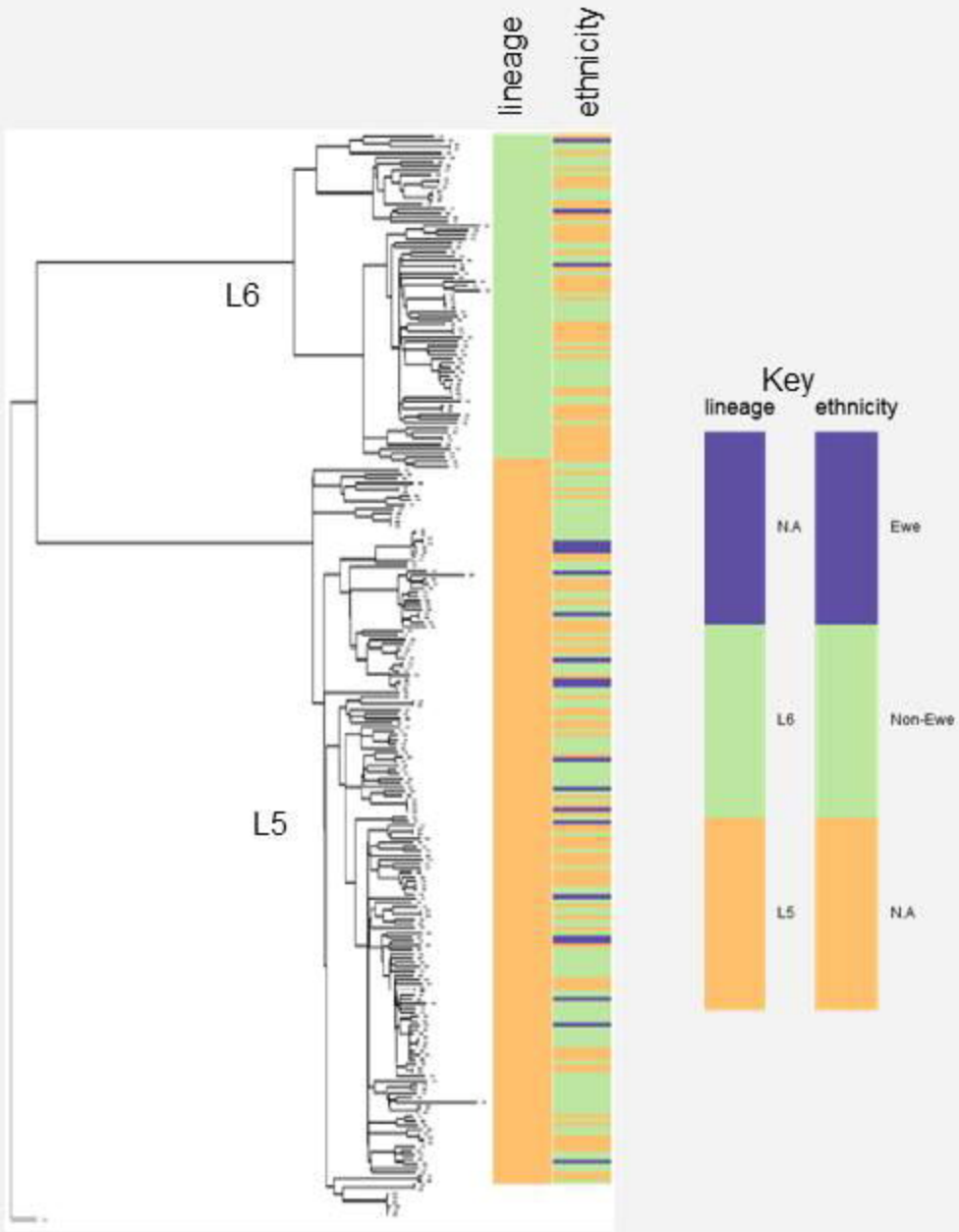

IV
virdetad
p = 0.1190
dN/dS
2.0
1.5
1.0
0.5
L5
L6
Lineage


V
cwallproc
p = 0.4266
dN/dS
1.1
0.9
0.7
0.5
L5
L6
Lineage


VI
infopath
****, p < 0.0001
dN/dS
1.0
0.5
L5
L6
Lineage


L5
L6
40
35
30
25
20
15
10
5
0
Number of SNPs
28
38
22
27
nsSNPs
sSNPs
Type of SNPs


a
L5 Non-Ewe
L5 Ewe
Number of T cell epitopes
1400
1200
1000
800
600
400
200
0
1216
1222
10
4
0
>0
Number of SNPs
b
L5 Ewe
L5 Non-Ewe
Number of SNPs
18
16
14
12
10
8
6
4
2
0
14
7
5
3
nsSNPs
sSNPs
Types of SNPs


lineage
ethnicity
L6
L5
Key
lineage
ethnicity
N.A
L6
L5
Ewe
Non-Ewe
N.A