North and East African mitochondrial genetic variation needs further characterization towards precision medicine

Anke Fähnrich[1,*], Isabel Stephan[1,*], Misa Hirose[2], Mosab Ali Awadelkareem[3,4], Saleh Ibrahim[2,5], Hauke Busch[1,6,+], Inken Wohlers[1,+§]

[1]Medical Systems Biology Division, Lübeck Institute of Experimental Dermatology and Institute for Cardiogenetics, University of Lübeck, Ratzeburger Allee 160, 23562, Lübeck, Germany

[2]Genetics Division, Lübeck Institute of Experimental Dermatology, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany

[3]Faculty of Medical Laboratory Sciences, Al-Neelain University, Khartoum, Sudan

[4]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford OX3 9DU, United Kingdom

[5]Sharjah Institute for Medical Research, University of Sharjah, Sharjah, United Arab Emirates

[6]University Cancer Center Schleswig-Holstein, University Hospital of Schleswig-Holstein, Campus Lübeck, 23538 Lübeck, Germany

[*]Equal contributing first authors

[+] Equal contributing last authors

[§] Correspondence to Inken.Wohlers@uni-luebeck.de

Abstract

Mitochondria are maternally inherited cell organelles with their own genome, and perform various functions in eukaryotic cells such as energy production and cellular homeostasis. Due to their inheritance and manifold biological roles in health and disease, mitochondrial genetics serves a dual purpose of tracing the history as well as disease susceptibility of human populations across the globe. This requires a comprehensive catalogue of commonly observed genetic variations in the mitochondria for all regions throughout the world. So far, however, certain regions, such as North and East Africa have been understudied. Towards this, we have created the most comprehensive quality-controlled North and East African mitochondrial dataset to date by compiling 11 published cohorts with novel data of mitochondrial genomes from 159 Sudanese individuals. We combined these 641 mitochondrial sequences with sequences from the 1000 Genomes (n=2,504) and the Human Genome Diversity Project

(n=828) and used the tool haplocheck for extensive quality control and detection of in-sample contamination. Using a subset of high-coverage mitochondrial sequences we predict 15 potentially novel haplogroups in North and East African subjects and observe likely phylogenetic deviations from the established PhyloTree reference for haplogroups L0a1 and L2a1. This demonstrates common hitherto unexplored variants in mitochondrial genomes of the North and East African region that lead to novel phylogenetic relationships, calling for further in-depth population genetic studies in that region.

Author summary

Does currently available human mitochondrial genetic data sufficiently cover North and East African diversity to investigate precision medicine for mitochondria-related diseases? To address this question, we have integrated and compared novel with existing sequence data from that region and world-wide. We find that variant profiles, i.e. haplogroups, carried by North and East Africans are shared with or related to those observed in Europe, Africa and Asia alike. Nonetheless, for a few haplogroups, world-wide data is inaccurate and incomplete for North and East Africans. Thus, we call for additional, future studies to more comprehensively represent human mitochondrial genetic diversity towards implementing precision medicine.

Introduction

Mitochondria are energy-producing, double-membrane-bound organelles that contain their own genome and that are central to many processes in eukaryotic cells (1). The mitochondrial (MT) genome is circular and typically 16,569 bases long in humans (2). Mitochondrial dysfunction is often linked to inefficient oxidative phosphorylation (OXPHOS) capacity with compromised ATP and increased mitochondrial reactive oxygen species (ROS) production that change cellular function and signaling (3). Dysfunction accompanies various pathologic conditions, such as metabolic diseases, neurodegenerative disorders (4), and autoimmune condition as well as cancer (5) and aging (6). Its root cause might stem from variations of the MT genome, which has thus moved into the center of research as determining factor for complex diseases and precision medicine.

Since mitochondria are exclusively maternally inherited, the mitochondrial phylogeny represents a maternal genealogy, which has been summarized in an established and widely applied reference phylogenetic tree called PhyloTree (7,8). Its current version 17 is based on more than 24,000 mitochondrial sequences and constitutes more than 5,400 haplogroups, i.e., recurrent genetic variant profiles common in human mitochondria.

Recent studies assessed mitochondrial haplogroups in world-wide human diversity reference data sets from the 1000 genomes project (1000G; n=2,504) (9,10), the Human Genome Diversity Project (HGDP; n=623 in (11); n=102 in (12)) and very recently gnomAD (13). Such mitochondrial phylogenetics studies have traced modern human settlement of the world (14) and support the hypothesis of Africa as origin of modern humans. The "Out-of-Africa" theory proposes the so-called "mitochondrial Eve", which represents the mitochondrial Reconstructed Sapiens Reference Sequence (RSRS) as the phylogenetic root and modern human origin (15).

In this work, we address the state of mitochondrial reference data from the regions of North Africa and East Africa. As of now, there is little genetic data available from this world region, as it has been forgone by large genome projects such as 1000G (16), gnomAD (17) or Topmed (18). An exception is Egypt, for which an initial reference dataset has been compiled (19) and efforts are ongoing to construct a much larger Egyptian genome reference towards implementing genomics-based precision medicine.

To address this reference data limitation, we sequenced *de novo* mitochondrial genomes of a Sudanese (n=159) and an Egyptian cohort (n=217) and integrated the results with an Egyptian cohort having whole genome sequencing data (n=110). Both Egyptian cohorts have previously been investigated for haplogroup distribution (19,20) and are here combined with mitochondrial sequencing data of ancient Egyptian mummies (21) as well as mitochondrial sequences from Genbank with origins of North or East African countries.

In addition to providing additional mitochondrial genomes from this understudied region, we reanalyzed all available sequencing data using a novel tool, called haplocheck (22), to identify potential sample contamination and phantom variants (23). Our combined and new quality-controlled data results in 146 exceptionally high-quality and 641 high quality mitochondrial genome sequences that we used to phylogenetically re-characterize genetic MT variation in North and East Africa with respect to world-wide MT diversity. We find that the haplogroups of the region cover nearly all world-wide clades and identified 15 potential novel

96 subhaplogroups, among them haplogroups L0a1 and L2a1, for which sequences from North
97 and East Africa suggest a different phylogenetic relationship as described by the reference
98 PhyloTree.
99
100
101 **Results & Discussion**
102
103 North and East African mitochondrial data collection and filtering
104 To comprehensively cover the phylogenetic tree of the North and East African population we
105 sequenced *de novo* a cohort of n=159 Sudanese individuals which we combined with
106 mitochondrial sequencing data of n=217 Egyptians. Additional n=110 MT genomes were
107 extracted from whole genome sequencing data of n=100 and n=10 Egyptians from Pagani et
108 al. (20) and Wohlers et al. (19), respectively. We further included mapped mitochondrial
109 sequencing data of n=90 ancient Egyptian mummies from Schuenemann et al. (21); for three
110 ancient Egyptian mummies WGS data were available (cf. Supplementary Table 1 for details).
111 As controls, we used two world-wide whole-genome references, 1000G with n=2,504 samples
112 (16) and HGDP with n=828 samples (24) from which we extracted the mitochondrial
113 sequences. Further, we obtained quality-controlled and aligned Genbank sequences from
114 countries in North and East Africa from the Supplement of McInerny et al. (25), which we
115 manually traced back to their eight respective publications provided in Suppl. Table 1.
116 Sampling location or ethnicity as well as number of sequences are provided in Figure 1.
117 Notably, five of eight studies relating to Genbank entries include only specific haplogroups.
118 These are two studies on haplogroup U6 (Maca-Meyer et al. (26), n=11; Olivieri et al. (27),
119 n=28) addressing Berbers in Morocco, two studies on haplogroup L3 in Morocco (Harich et al.
120 (28), n=8) and Ethiopia (Soares et al. (29), n=70) and one study on both M1 and U6 addressing
121 Morocco (Pennarun et al. (30), n=56). Notably, we also included n=18 Genbank MT sequences
122 of ancient Moroccans (Fregel et al. (31)). Genbank-related studies that do not focus on a
123 specific haplogroup are addressing Egypt (Kujanová et al. (32), n=35) and Tunisia (Costa et al.
124 (33), n=42).
125 To exclude study-specific batch effects and sequencing errors, we applied a two-step quality
126 control (QC) filter on all sequencing-based North and East African data sets. This was
127 important to correctly assess haplogroup distributions and to identify putatively new

haplogroups. For the filter step, we kept all samples from Genbank as well as those (i) having a haplocheck-reported major haplogroup equal to minor haplogroup, and (ii) having contamination assessed and not detected (i.e. “NO”). An overall of 641 samples from four sequencing-based North and East African data sets and from eight Genbank studies pass this filter (cf. Suppl. Table 11 and in Suppl. Table 17 for aligned sequences). For the second QC filtering step, we excluded Genbank sequences, as contamination checks were unfeasible, and additionally required a minimum sequencing coverage greater than 1000x, which is considerably higher than the minimum 600x coverage required by haplocheck to detect a 1% or greater contamination (22). The number of total and filtered samples in each QC step are provided in Figure 1 for every cohort included. For mitochondrial sequences obtained from Genbank, the overall number is provided, only.

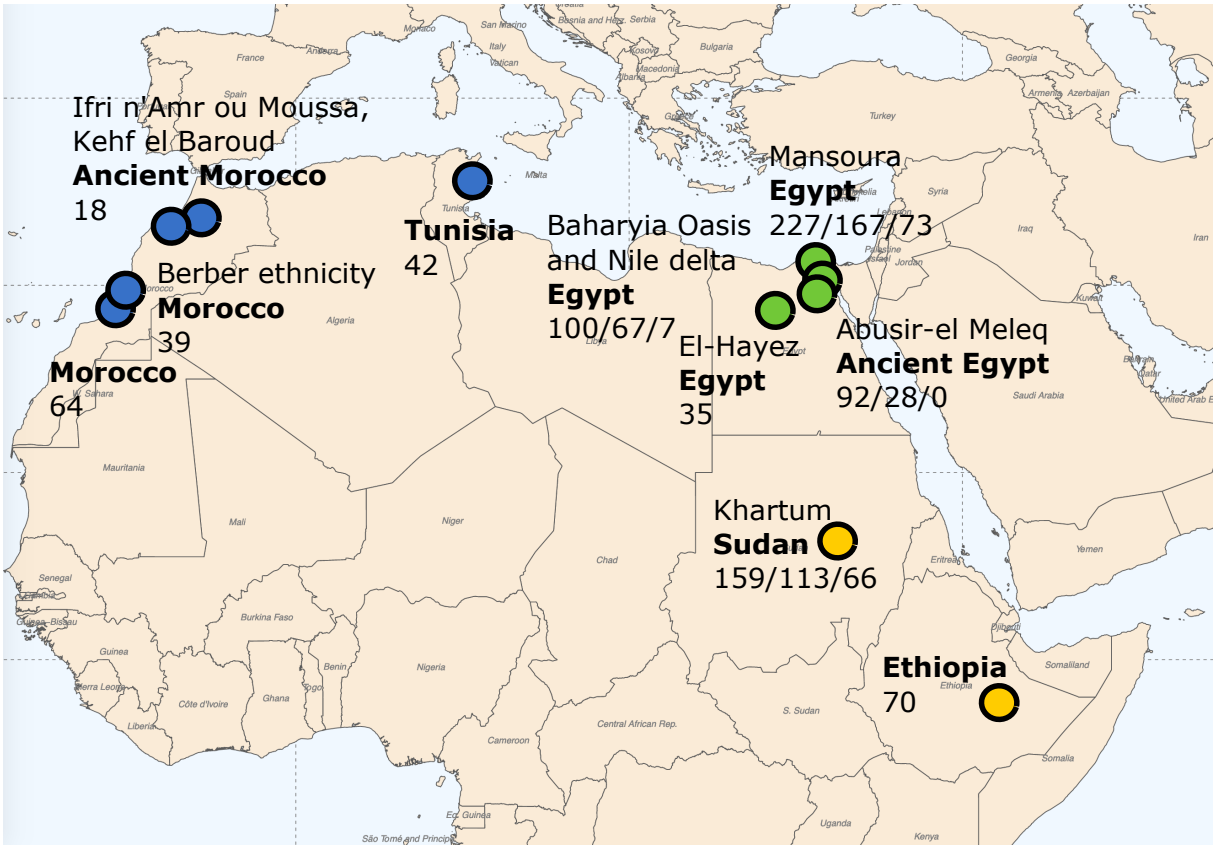
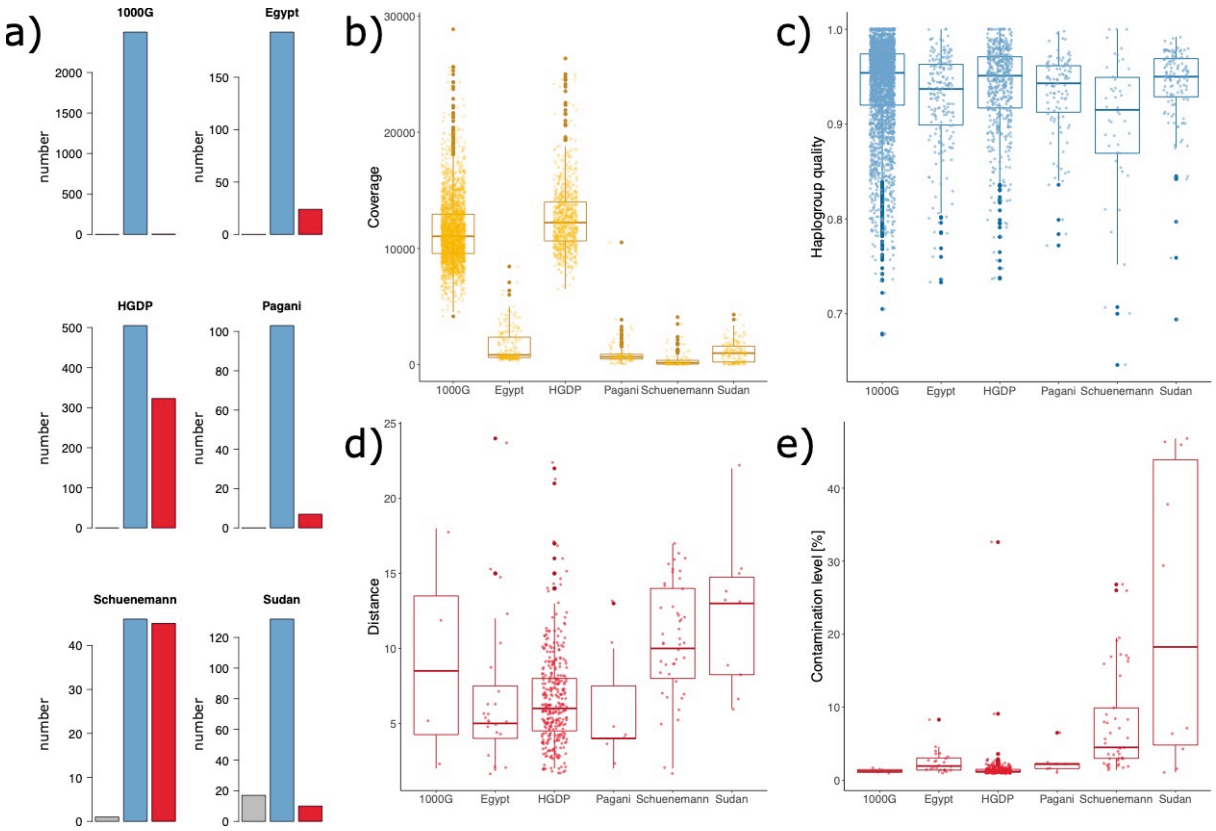


Figure 1: Mitochondrial data compiled from North and East Africa. Blue color denotes cohorts referred to as North African, green color cohorts referred to as Northeast African and yellow color cohorts referred to as East African. Cohorts analyzed from raw sequencing data are denoted by three numbers: The overall number of samples, the number of samples remaining after lenient QC filtering and the number of samples remaining after strict QC filtering. Locations with only one number refer to Genbank MT sequences obtained via McInerny et al.

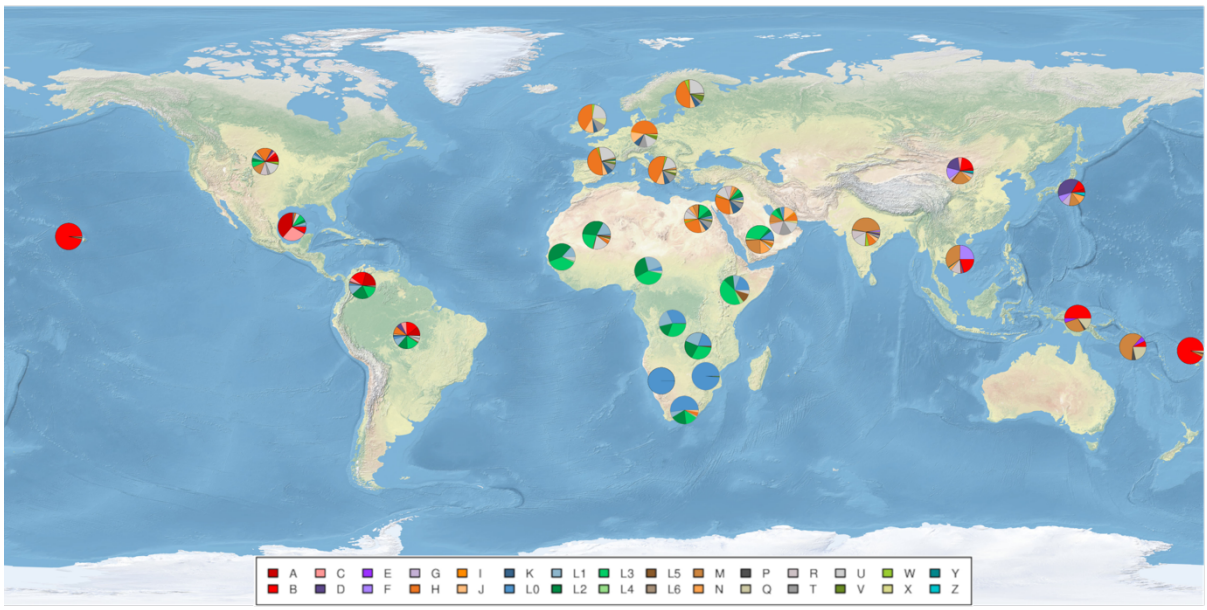
147 (25). Whenever we do not explicitly refer to Northeast Africa, it is included in North and/or
148 East Africa.
149
150 Haplogroup-based contamination assessment for diverse sequencing settings
151 We assessed potential contamination and related sequencing and variant calling metrics for
152 all six sequencing-based datasets, including the world-wide references 1000G and HGDP (cf.
153 Supplementary Tables 2-9 and 13-16 for details). Coverage of the two latter is high with a
154 median coverage greater than 10,000X, while the former four cohorts have a significantly
155 lower coverage of 840, 670, 163 and 992 for the Egyptian MT sequencing, the Egyptian WGS,
156 the ancient Egyptian mummies and the Sudanese MT sequencing cohorts, respectively, see
157 Figure 2 b.
158



159
160 **Figure 2:** Characterization of the six sequencing-based data sets using numbers reported by
161 haplocheck. a) Histogram of number of samples with haplocheck contamination status not
162 available ("NA") (gray), not contaminated ("NO") (blue) and predicted contaminated ("YES")
163 (red). b) Box plots for sequencing coverage. c) Box plots for haplogroup quality of those
164 samples not predicted to be contaminated. d) Box plots for distance between major and minor
165 haplogroup within the phylogenetic tree for those samples predicted to be contaminated. e)

166 Box plots of the contamination levels for samples predicted as contaminated. Box plots display
167 median and lower/upper quartiles; whiskers denote the most extreme data point no more
168 than 1.5 times the interquartile range; outliers are data points extending beyond whiskers.
169
170 Contamination is identified using the computational tool haplocheck, which identifies
171 polymorphic sites within samples. It attempts to decompose the variants into a so-called
172 major and another, minor haplogroup. If such decomposition is possible and coherent, a
173 contamination of a sample of the major haplogroup with another sample of the minor
174 haplogroup is predicted and a respective contamination level reported. Interestingly, we
175 found a large heterogeneity in contamination calls among the studies. Few of the WGS-based
176 1000G samples are predicted to be contaminated, yet more than 30% of the WGS-based HGDP
177 samples (Figure 2a), albeit at a low contamination level of 2% (median 1.2%, Figure 2e).
178 Comparable, low level high distance contamination is predicted for WGS samples of Wohlers
179 et al. and Pagani et al. Possibly, these contamination calls are caused by nuclear mitochondrial
180 DNA sequences (NUMTs) causing mapping artefacts, although a core genome coverage of 30x,
181 a typically large haplogroup quality (Figure 2c) and a large distance between major and minor
182 haplogroup (Figure 2d) indicate that this is rather not the case. High contamination levels that
183 indicate a mix-up during library preparation are restricted to a few samples from the Sudan
184 cohort. For the ancient Egyptian mummy data set, haplocheck predicts contaminations in
185 haplogroups related to H2a2a, which is very likely caused by a lack of sequencing coverage,
186 which causes the revised Reference Sequence (rCRS) bases to be identified that refer to
187 haplogroup H2a2a1. This is supported by the very low coverage for this cohort.
188
189 Distribution of Western Asian and Northeast African mitochondrial haplogroup diversity
190 The sequential accumulation of polymorphisms in the MT sequence over time across
191 generations has led to the emergence of geographically isolable variation profiles and MT
192 haplogroups. As a result, geographic regions have characteristic distributions of mitochondrial
193 haplogroups. The analysis of n=9,264 world-wide MT sequences (see Methods) yielded
194 mitochondrial haplogroup distributions for distinct geographical regions as seen in Figure 3a.
195

a)



b)

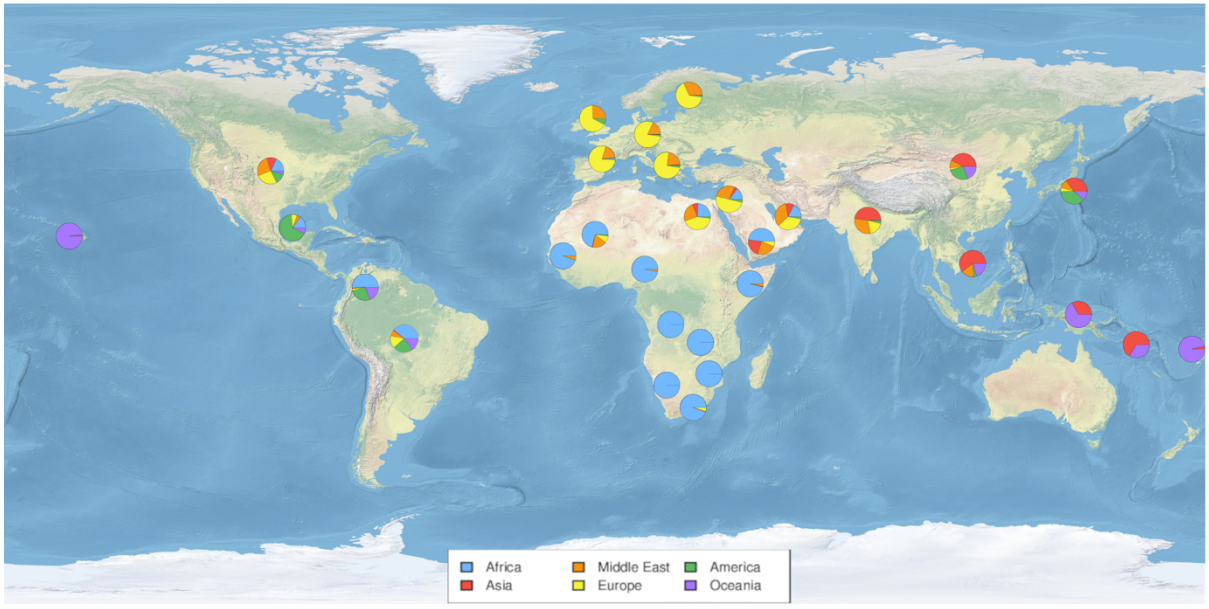
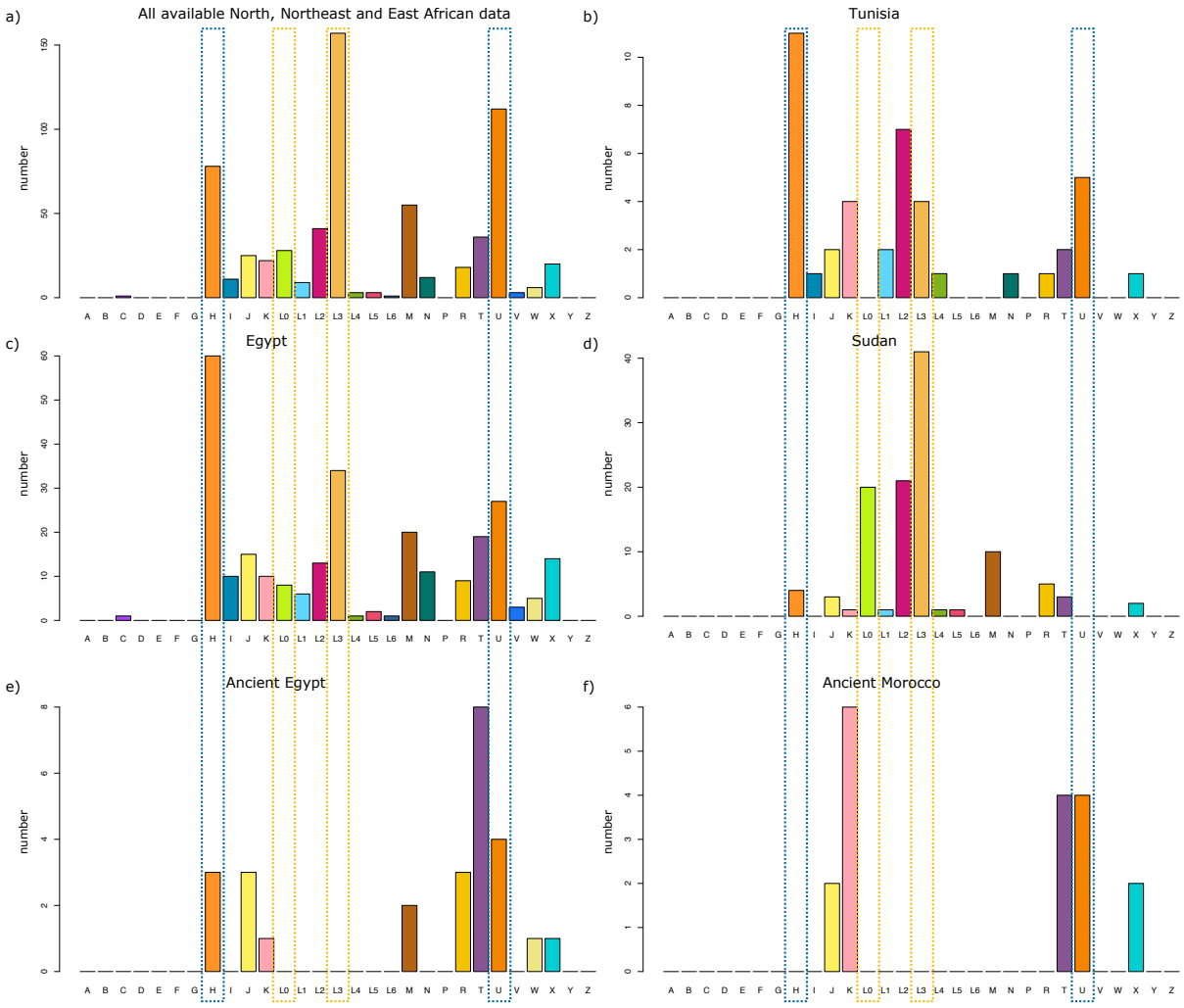


Figure 3: World-wide haplogroup distributions by geographic region. a) Macrohaplogroup level b) Macrohaplogroups have been assigned to hypothetical continents or regions of origin. We decided on the following assignment: Africa: L haplogroups; Asia: E, F, G, M, Y, Z; Middle East: N, R, U, I, W; Europe: H, J, K, T, V; America: A, C, D, X; Oceania: B, P, Q.

As described previously, multiple haplogroups are predominant in different continental region. Here, we decided on the following assignment: Africa: L haplogroups; Asia: E, F, G, M, Y, Z; Middle East: N, R, U, I, W; Europe: H, J, K, T, V; Americas: A, C, D, X; Oceania: B, P, Q. Of note, we are aware that haplogroups overlap between continents, e.g., especially in Europe and the Middle East. The grouping above therefore does not entail exclusivity with respect to

207 the listed haplogroups. Based on our assignment, we determined the relative proportion of
208 haplogroups associated with different continents for each clustered group (see Figure 3b).
209 Regions in Asia, North America, Northeast Africa, and the Middle East are particularly diverse.
210 Historically, this can be attributed to events such as colonization and slave trade in the
211 Americas as well as early human migration in the Middle East (34). In addition to its unique
212 transcontinental location between Africa and Asia, Northeast Africa is adjacent to the center
213 of modern human settlement in the world (14) and has been affected by complex events of
214 prehistoric back migrations (35).
215 The relative frequency of mitochondrial macrohaplogroups of the 641 QC filtered North and
216 East African samples are depicted in Figure 4a and the individual haplogroup assignment per
217 sample is given in in Supplementary Table 11.
218



219
220 **Figure 4:** Histograms of macrohaplogroup frequencies. Blue dashed line boxes indicate
221 predominantly North African and yellow dashed line boxes predominantly East African
222 macrohaplogroups within this region. a) For all n=641 North, Northeast and East African

223 samples passing the lenient QC filter and including Genbank sequences. This includes also
224 Genbank sequences of studies addressing particular haplogroups (e.g. L3, M1 and U6) and
225 thus is not a geographically representative distribution. b) Tunisian haplogroup numbers out
226 of overall n=42. c) Egyptian haplogroup numbers out of overall n=297. d) Sudanese
227 haplogroup numbers out of overall n=113. e) Haplogroup numbers in ancient Egyptian
228 mummies. f) Haplogroup numbers in ancient Moroccan individuals.
229
230 Country-specific haplogroup frequencies for Tunisia, Egypt, Sudan, ancient Morocco and
231 ancient Egypt are shown in Figures 4b-f. Geographic neighborhoods are clearly reflected in
232 the haplogroup distributions and indicate a geographic gradient from North Africa via
233 Northeast Africa towards East Africa. This can be noted particularly for haplogroups H and U,
234 that are most prevalent in Tunisia, less abundant in Egypt, and rare in Sudan. Vice versa, L3
235 and L0 are prevalent in Sudan, and to a lesser extend in Egypt but rarely observed in Tunisia.
236 Interestingly, L2 is commonly observed in both Tunisia and Sudan, but rarer in Egypt. Overall,
237 most macrohaplogroups are observed in the Egyptian cohort (compiled from three data sets).
238 This finding might be explained by the larger sample sizes and compilation of various data sets
239 available for Egypt, compared to only one data set for Tunisia and Sudan, respectively. Yet,
240 more macrohaplogroups reach high fractions compared to Tunisia and Sudan, supporting the
241 hypothesis that Egypt is the region of highest mitochondrial genetic diversity within North-
242 and East Africa.
243
244 Northeast African mitochondrial lineages cover nearly all major world-wide clades
245 In order to elucidate the relative haplogroup abundance in North and East Africa relative to
246 the rest of the world, we computed a phylogenetic tree from the 641 sequences in
247 combination with the 1000G and HGDP reference data. The resulting tree in Figures 5a-b
248 demonstrates that maternal ancestry is shared with nearly all major, world-wide
249 mitochondrial clades and is generally not restricted to specific subclades, apart from the
250 haplogroup M, which is mainly represented by M1. This is a haplogroup known to be restricted
251 to North and East Africa that is absent from the 1000G and present in HGDP data only once
252 (see also Supplementary Figures 1 and 2 with added sample names and corresponding
253 haplogroups).

The world-wide macrohaplogroups not observed in North and East Africa are A, B, D, E, F, G, P, Y and Z, which are predominantly North Asian, Northeast Asian, East Asian, Southeast Asian, Native American or Oceanian.

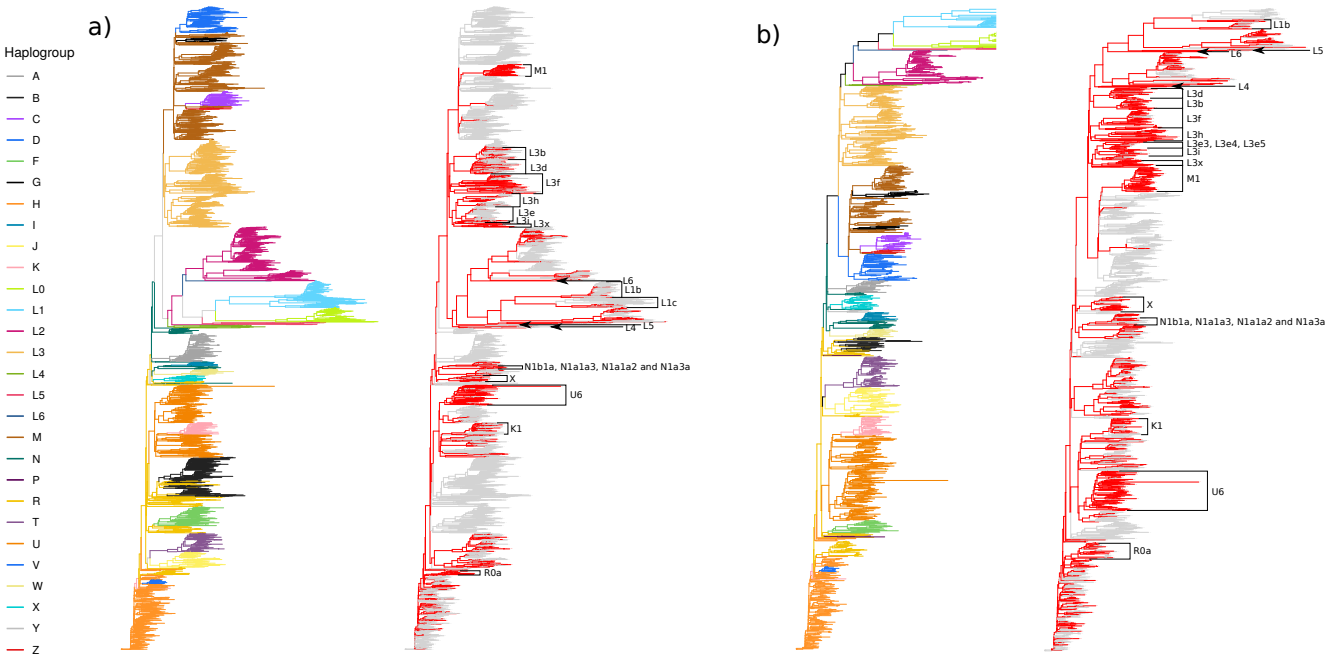


Figure 5: Phylogenetic tree of 641 North and East African mitochondrial sequences together with sequences from two worldwide reference data sets. The tree is colored according to major macrohaplogroup assigned by haplocheck (left) and colored according to data source, with North, Northeast and East African mitochondrial sequences compiled in this study in red and reference data in gray (right). a) Reference data is 1000G b) Reference data is HGDP.

The predominantly European haplogroup H is more common in Tunisia and Egypt and less common in Sudan. The same applies to the haplogroups I, J, K, N and U.

Haplogroup I represents a with respect to RSRS not very deep-rooting sister clade of N1a1b, as described earlier (36). Haplogroup J is also observed in both ancient data sets. North African sequences belong to J1 as well as J2 and are with respect to RSRS more deep-rooting than 1000G and HGDP sequences.

Haplogroup K represents a subgroup of U8, as expected (37). In the North and East African data, only K1 is observed, whereas 1000G and HGDP also cover K2.

The African haplogroups L0, L1, L2 and L3 that are commonly found in Africa show a high abundance in Egypt, Sudan and Tunisia with the exceptions of L0 in Tunisia and L1 in Sudan being absent and lowly abundant respectively. L3 is the founder of haplogroup N, which has

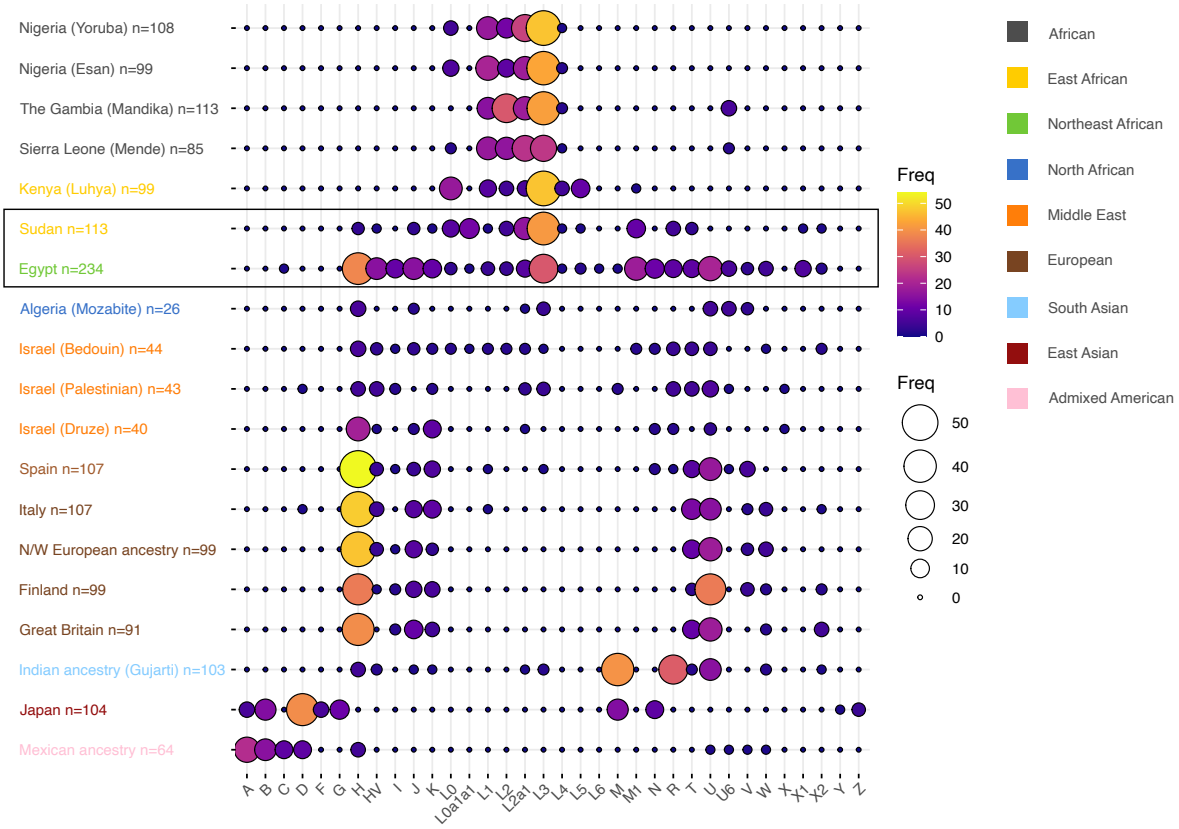
276 given rise to all world-wide haplogroups outside of Africa, and as such it has been studied
277 extensively and exclusively. Haplogroup L3 is observed in all data sets, but predominantly in
278 Sudan with a share of 40%. This is in line with previous studies that suggest East Africa as the
279 origin of L3 (29).
280 Besides African-wide common haplogroups we observe macrohaplogroups L4, L5 and L6. All
281 these macrohaplogroups have previously been observed in East Africa, mainly. Due to low
282 prevalence and due to being restricted to this region, they have not been studied explicitly.
283 All but one sample relating to North and East African M haplogroups belong to the M1 clade,
284 which is known to be restricted to this region and is the only non-Asian M subclade. The data
285 sets compiled here cover a large diversity of M1. Haplogroup M1 has neither been observed
286 in the comparably small Tunisian nor in the ancient Moroccan data set, but was present in the
287 ancient Egyptian mummy data.
288 Sequences belonging to the N sub-haplogroups build subclades that are located at different
289 positions in the phylogenetic tree generated here. All of them are in the same phylogenetic
290 branch, the same holds true for the haplogroup R. We only observe sub-haplogroups of R0a
291 (n=17) that has been described and interpreted previously (38), and encounter haplogroup R
292 among the filtered sequences of an ancient Egyptian mummy once.
293 Nearly all T sub-haplogroups are detected in North or East Africa, and their share is elevated
294 in both Moroccan and Egyptian ancient data sets. The presence of haplogroup T, as well as its
295 sister clade J, has previously been observed and interpreted as Near Eastern genetic impact
296 on the region (21,32).
297 North and East African sequences of macrohaplogroup U cover all sub-haplogroups (U1 to U8)
298 except the rare U9 subgroup. Haplogroup U6 has been studied in a targeted way (26,27,30)
299 because it is, together with M1, the only haplogroup of a world-wide macrohaplogroup that
300 is autochthonous in Africa, i.e., predominantly occurring there. Corresponding U6 sequences
301 from (26) and (27) that we integrated in our data compilation as well as Egyptian and world-
302 wide reference data support this notion.
303 Finally, haplogroup X is present in data from all North and East African regions, including data
304 from three ancient individuals. The coalescence of X in the phylogenetic tree with 1000G and
305 HGDP as reference, respectively, is not identical, involving N9b and R11 as sister clades, both
306 of which are haplogroups that are poorly represented in the 1000G and HGDP reference data,
307 resulting in a limited resolution of the phylogenetic relationship of these clades.

308

309 Clades occurring predominantly in North and East Africa are often poorly characterized

310 We next investigated North and East African haplogroup profiles within the geographic
311 context using selected 1000G and HGDP populations from neighboring regions, see Figure 6.
312 A corresponding histogram is shown in Supplementary Figure 3.

313



314

315 **Figure 6:** Haplogroup frequencies of selected 1000G and HGDP populations together with the
316 Sudanese and Egyptian haplogroups of this work (highlighted in box). The populations are
317 colored according to geographic region.

318

319 This shows that both Egyptian and Sudanese mitochondrial profiles are mediating between
320 the profiles of populations assessed by these public references, constituting a genetic
321 continuum between them. For example, 1000G covers African, European, South Asian, East
322 Asian and Admixed American populations, none of which represents mitochondrial
323 haplogroup occurrences we observe in Sudan or Egypt according to data generated and
324 compiled in this work (cf. Figure 6). Middle Eastern and North African populations of the HGDP
325 reference show a combination of predominantly European, African and Asian haplogroups
326 comparable to Sudan and Egypt. However, HGDP population sample sizes are low and thus

327 corresponding haplogroup distributions may not be representative. Further, some
328 haplogroups from the North and East African region are so far absent from both major global
329 references 1000G and HGDP. This applies especially to specific haplogroups that exclusively
330 or predominantly occur in North and/or East Africa, e.g. subclades of L0a1a1, L3, L4, L6, M1,
331 U6 and X1. Some of those, with a particular interest to world-wide human prehistoric
332 migration routes, have explicitly been addressed by studies and selectively been sequenced
333 before, e.g., L3, M1 and U6 and are included in our study. However, local clades possibly not
334 related to archaeogenetic hypothesis are more poorly characterized, e.g., L0a1a1 and X1.
335 Finally, there is little information on rare haplogroups such as L6. They differ a lot from all
336 other, well characterized haplogroups and as such may convey different mitochondrial
337 function and disease susceptibility. Thus, also and especially such rare haplogroups will need
338 better characterization in the context of mitochondrial genetics-based precision medicine.
339
340 Variation suggesting novel haplogroups
341 To obtain a dataset of exceptional high-quality we next selected (n=146) samples with high
342 sequencing coverage and without predicted contamination (provided in Suppl. Table 12). The
343 phylogenetic tree of these sequences with corresponding haplogroups is depicted in Figure 7.

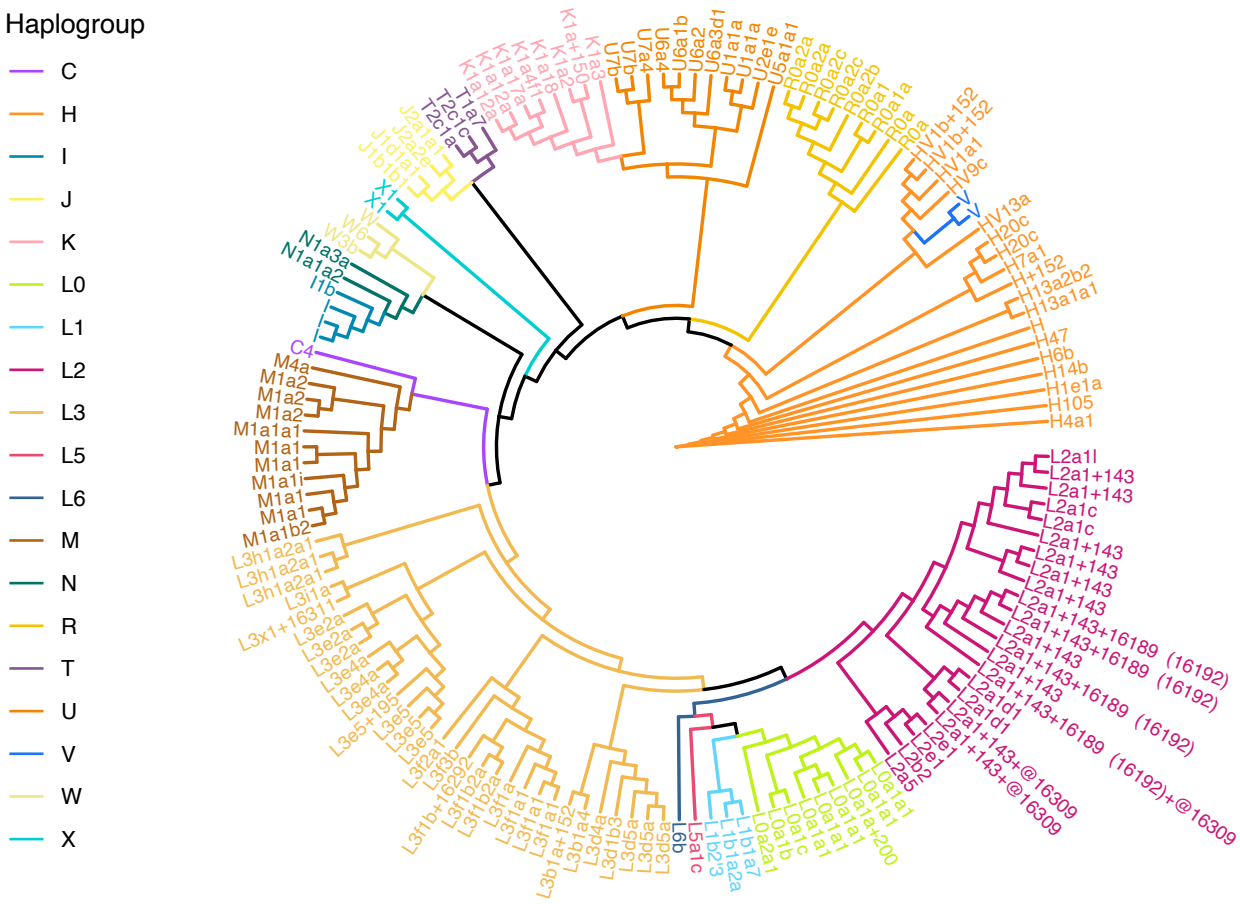


Figure 7: Phylogenetic tree of the n=146 exceptional high-quality sequences used for detection of novel haplogroups. The circular tree depicts the topological information, while branch lengths are not preserved. Two alternative visualizations preserving branch lengths and a histogram of the macrohaplogroup distribution are shown in Supplementary Figure 4.

In the search for novel and previously undefined subclades, we investigated this set using the phantom mutation module of the online haplogrep2 tool. In this module, sequences are checked for recurrent variants that are not haplogroup-defining and that have a 'Soares score' of less than or equal to 2, i.e., having been detected no more than twice in a large collection of mitochondrial sequences (according to Table S3 of Soares et al. (39)). We checked for such variants occurring in samples of the same haplogroup to ensure that the variation is likely specific to haplogroup carriers and defines a novel, previously undefined subclade. A summary is provided in Table 1.

Haplogroup	Subhaplogroup variant	# HQ	# LQ	Soares score
HV1b+152	8640T;4590G	2	1	0;0
I	567C	3	0	0
I	567C;4772C	2	3	0;2
L0a1a1	8869G	2	1	1
L0a1a1	8017T	3	9	0
L2a1c	5186G;9116C	2	0	2;2
L2a1+143	7897A;398C;9389G	2	0	1;2;2
L2e1	13356C	2	0	0
L3d5a	15457A;14410A	2	1	0;1
L3e4a	470G; 9455G	2	2	0;1
L3e5	711C	2	2	0
L3f1a1	12441C;736T;328G; 15151G;13967T;14554G	3	3	0;0;0;1;2;2
M1a2	16182G;2858G; 15401G;1282A	2	0	1;1;1;1
R0a2c	5981C	2	1	0
V	9126C	2	1	0

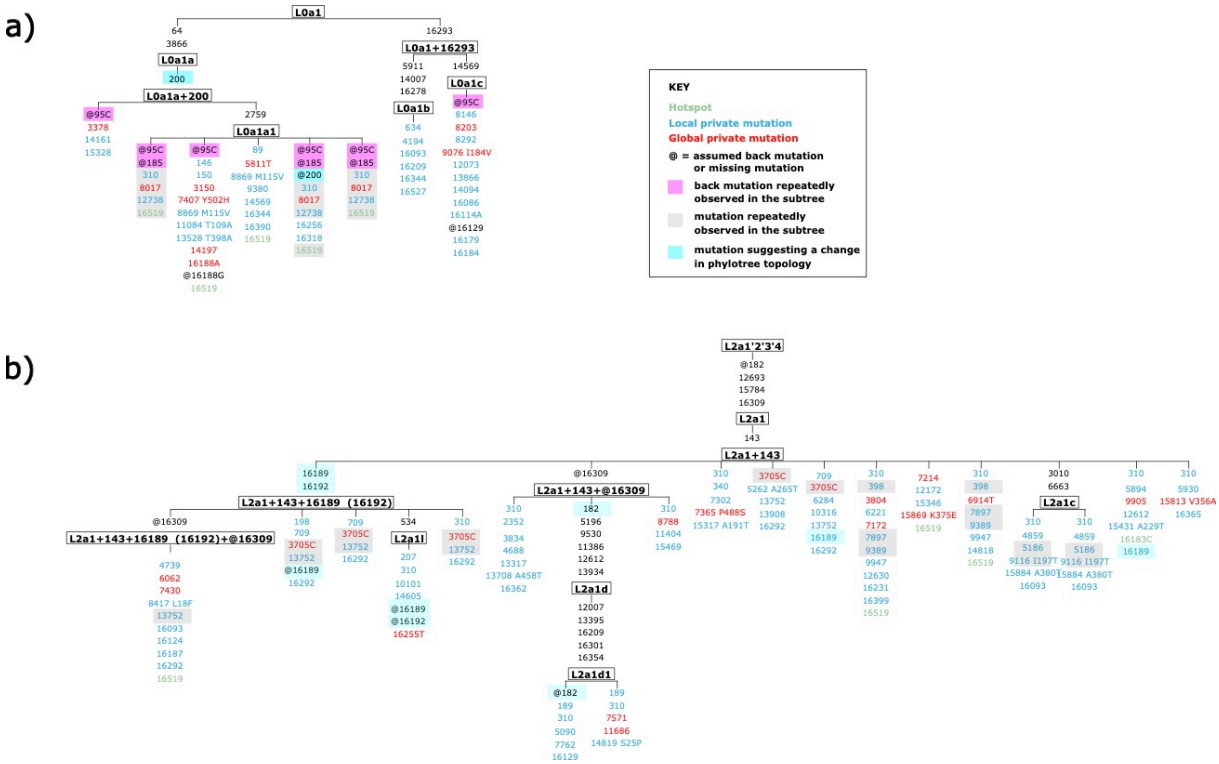
Table 1: Haplogroups having potentially novel and previously undescribed sub-haplogroups. The table columns denote, in this order, the haplogroup, the recurrent novel variants, the number of samples after strict filtering (# HQ; out of n=146) and after lenient filtering (#LQ; out of n=641-146=495) and the Soares score (39), which represents the previous occurrence of the variant in world-wide data.

For ten putatively novel haplogroups, we find further support among samples with lower sequencing coverage. We find one, two and 3-6 novel recurrent variants for respectively seven, five and three haplogroups each. Thereby, 14 of 30 potential novel haplogroup variants have a Soares score of 0 (47%), 9 have a score of 1 (30%) and the remaining 7 have a score of 2 (23%). The haplogroups of which potential novel sub-haplogroups have been identified are either haplogroups that are largely restricted to North and/or East Africa, e.g., L2a1, M1, and different L3 sub-haplogrups (d, e and f), or are potential local clades diverging from basic macrohaplogroups, e.g., I and V.

375

376 Variation suggesting revision of phylogenetic relationships

377 For the 146 exceptional high-quality sequences we examined the phylogenetic tree according
378 to the PhyloTree reference for potential new haplogroups using a function from the web-
379 based haplogrep2 tool (see Supplementary Figure 5). For this, we considered all detected and
380 haplogroup-defining expected, but missing, variants in the samples. We find two clades which
381 are better represented, when revising or extending the phylogenetic tree provided by
382 PhyloTree, L0a1a1 and L2a1, see Figure 8



383

384 **Figure 8:** Phylogenetic tree according to the PhyloTree reference for those samples that were
385 assigned to subclades a) L0a1 and b) L2a1. We highlight with magenta, gray and turquoise
386 boxes those variants that indicate that a different phylogenetic tree may better explain the
387 samples from North and East Africa.

388

389 There are eight samples assigned to L0a1 sub-haplogroups, see Figure 8a. In most of them,
390 haplogroup-defining variant 95C is not observed, which is already noted within the PhyloTree
391 reference. Further, three samples show a back mutation 185. Such a back mutation is
392 expected for haplogroup L0a2 according to PhyloTree, so the samples we sequenced may link
393 clades L0a1 and L0a2. In one sample, an expected 200 mutation was not observed, showing

394 the possibility that 95C and 185 together with 310, 8017, 12738 and possibly 16519 are a
395 hitherto undescribed subclade of L01a1.
396 The second branch of interest is L2a1. Out of 19 samples to which sub-haplogroups of L2a1
397 have been assigned, most show extensive, recurrent deviations from the PhyloTree
398 phylogenetic tree, see Figure 8b. For example, resolving the haplogroup variants 16189 and
399 16192 within the PhyloTree reference seems not representative for North and East African
400 samples. Further, variants 3705C, 13752, 7897, 9389, 5186 and 9116, which are recurrent in
401 the L2a1 clade, are not part of PhyloTree, yet. However, if included, they would implicate
402 changes in phylogeny.
403
404
405 **Conclusion**
406
407 We investigated whether current global mitochondrial genetic data adequately represents the
408 genetic mitochondrial variation of the North and East African region. Genetic data from this
409 region is sparse until now, despite an apparent interest due to autochthonous ancestry
410 components being identified (19,40) and despite the importance of the region as origin or
411 transit of modern humans outside of Africa. Beside in-depth analysis of mitochondrial
412 sequencing data underlying Egyptian haplogroup distributions that we have reported
413 previously (19), we here generated novel sequencing data of mitochondrial genomes from
414 Sudan, another region with little genetic data available. We characterized North and East
415 African mitochondrial variation by combining this data with publicly available sequencing data
416 as well as with mitochondrial sequences from Genbank relating to 10 further studies. This
417 results, to the best of our knowledge, in the largest mitochondrial data collection from this
418 region to date.
419 With a stringent quality control via the novel tool haplocheck we related North and East
420 African phylogenetics to the world-wide context by restricting the analysis to mitochondrial
421 sequences unaffected by potential contamination or other shortcomings such as insufficient
422 sequencing coverage. An additional filter for a minimal 1000x sequencing coverage results in
423 exceptionally high-quality mitochondrial sequences, which makes us very confident that novel
424 mitochondrial variations, novel haplogroups and observed PhyloTree inconsistencies are
425 indeed caused by a hitherto insufficient genetic representation of this geographic region.

426 We find that North and East African mitochondrial genomes are phylogenetically diverse and
427 cover nearly all major clades worldwide. Likewise, specific subclades are prevalent in the
428 region, supporting that North and East Africa, and particularly Northeast Africa harbors
429 ancestry components typically attributed to Africa, Europe and Asia (19) , all of which are likely
430 present in the region since prehistoric times (41). Further, several North or East African clade
431 representatives highlighted by our work and not studied in-depth so far may still add further
432 information to worldwide maternal human genetic history, e.g., the haplogroups L4, L6, N, I
433 and X.
434 In conclusion, the initial North and East African mitochondrial reference dataset compiled as
435 part of our work calls for further extension in order to support genetics-based precision
436 medicine.
437
438
439 **Material and Methods**
440
441 Haplogroup data compilation for world-wide mitochondrial haplogroup distribution
442 Based on a comprehensive literature research, n=7,414 publicly available mtDNA sequences
443 with available haplogroup and geographic information in the form of the country of origin
444 were collected. Additionally, data from two reference datasets, the 1000 Genomes dataset
445 (16) (n=2,504) and the Human Genome Diversity Project (24) (n=828) were added, resulting
446 in a total number of n=10,746 entries. As quality assurance, for all database entries
447 haplogroups were reassigned using for mitochondrial sequences haplogrep2 (9) and for next
448 generation sequencing-based cohorts haplocheck (22). Subsequently, entries were filtered
449 such that only those remained that were representative for the population of a region.
450 Towards this, studies analyzing only specific haplogroups as well as data from archaeological
451 studies was removed. Eventually, 9,264 entries were retained for further analysis.
452
453 Objective geographic grouping for world-wide mitochondrial haplogroup distributions
454 For determining the relative amount of haplogroups per geographic region, we grouped the
455 database entries into clusters of minimal geographic distance and having a size of at least 100
456 entries. At the beginning, only countries with less than 100 entries were considered, i.e., those
457 that could not form a group on their own. Countries with an initial number less than 100 were

458 assigned to group 1, all other countries to group 2. In order to classify a country from group
459 1, an attempt was made to combine it in a first step with other countries or already connected
460 groups of countries in group 1. A continent-dependent maximum distance was set as a
461 termination criterion. For Europe, this was set to 1,000 km, for Africa to 1,750 km, for the
462 Middle East to 1,600 km, for Asia to 1,900 km, for the Americas to 2,600 km and for Oceania
463 to 1,600 km. This distance was determined as a function of the average area of a country on
464 the continent and the existing sample density. If after this first step, the sample number of a
465 country or a group was still less than 100, the countries with an initial sample number greater
466 than or equal to 100 were included as possible connection partners. In this second step, the
467 country or group to be classified was in any case merged with the country from group 1 or
468 group 2 with minimal distance.
469
470 Mitochondrial DNA sequencing
471 Genomic DNA samples were processed for library preparation, using the Human mtDNA
472 Genome protocol for Illumina Sequencing Platform
473 (https://emea.support.illumina.com/content/dam/illumina-
474 support/documents/documentation/chemistry_documentation/samplepreps_legacy/huma
475 n-mtdna-genome-guide-15037958-01.pdf) as described previously (19,42) for Egyptian
476 samples and 63 Sudanese samples. In brief, a primer set [MTL-F1
477 (AAAGCACATACCAAGGCCAC) and MTL-R1 (TTGGCTCTCCTTGCAAAGTT); MTL-F2
478 (TATCCGCCATCCCATACATT) and MTL-R2 (AATGTTGAGCCGTAGATGCC)] was used to amplify
479 the mtDNA by long-range PCR (LR-PCR). Library preparation of the LR-PCR products was
480 performed using a Nextera XT DNA Library Preparation Kit (Illumina Inc., CA, USA). For 96
481 Sudanese samples, we applied a slightly modified protocol to improve the evenness of the
482 coverage. More specifically, the LR-PCR was conducted using another primer set [hMT-1_F
483 (AAATCTTACCCCGCCTGTTT) and hMT-1_R (AATTAGGCTGTGGGTGGTTG); hMT-2_F
484 (GCCATACTAGTCTTTGCCGC) and hMT-2_R (GGCAGGTCAATTTCACTGGT)] and the library
485 preparation of the LR-PCR products of hMT1 and hMT2 primers was conducted using DNA
486 Prep Tagmentation (Illumina Inc.). This approach allows the fragment size of the tagmented
487 LR-PCR products to be evenly 300 bp, as the transposome reaction occurs every 300 bp on the
488 beads, in contrast to the Nextera XT DNA library preparation Kit, by which transposome
489 reaction occurs randomly, thus, results in various size of the tagmented LR-PCR products (i.e.,

490 larger fragments were removed by beads purification steps). The libraries prepared by both
491 methods were further purified, concentration quantified using Qubit Fluorometer (Thermo
492 Fisher GmbH, Dreieich, Germany) and the library size was determined by 2100 Bioanalyzer
493 Instrument (Agilent, Santa Clara, CA, USA). The final library was sequenced on the Illumina
494 MiSeq sequencing platform, using v2 chemistry (2 × 150 bp paired-end reads) (Illumina Inc.).
495
496 Sudanese and Egyptian mitochondrial sequencing characteristics and QC
497 Running FastQC on all Sudanese and Egyptian mitochondrial sequencing FASTQ files resulted
498 in no systematic, data quality compromising warnings or errors. Sudanese sequencing data
499 has an average length of 140 bases, Egyptian sequencing comprises samples sequenced with
500 150 and samples sequenced with 300 bp target read length, resulting in an average length of
501 205 bases across the cohort. Average bases mapped per sample are about 18 Mb for the
502 Sudanese cohort and about 30 Mb for the Egyptian cohort, amounting to an approximate
503 average coverage of more than 1000X. In the Sudanese cohort, five samples have less than
504 1000 reads and are considered dropouts. The Egyptian cohort does not contain samples with
505 less than 1000 reads.
506
507 Variant calling, haplogroup assignment and contamination detection
508 Raw reads from FASTQ files of Sudanese and Egyptian sequencing data were mapped to
509 mitochondrial sequence MT of GRCh38, which resembles the historic revised Cambridge
510 Reference Sequence (rCRS), a sequence of haplogroup H2a2a1. For samples with available
511 BAM files, we used these directly. Our analysis is largely based on haplocheck (version 1.1.2)
512 (22) and its internal accompanying tools haplogrep2 (9) and mutserv (version 1.3.4) as
513 implemented via the mtDNA-server (43). Haplocheck is a recent tool which allows to detect
514 sample contaminations from sequencing data by assessing if heteroplasmic variant calls can
515 be broken down into a mixture of two haplogroups, one denoted as “major”, the other as
516 “minor”. Variant caller mutserv was run with default settings, which amounts to reporting
517 heteroplasmic variants at a level larger than 1% (--level 0.01), calling deletions (--deletions),
518 base alignment quality turned off (--noBaq), considering a mapping quality of 20 (--mapQ 20)
519 and a base quality of 20 (--basQ).
520
521 Phylogenetic analyses

522 In order to obtain a FASTA format mitochondrial sequence that relates to the haplogrep2
523 output reported by haplocheck, we perform variant calling with mutserv using the default
524 parameters also used within haplocheck. Then we select from the VCF those variants for which
525 the major base is the non-reference base. Using bcftools consensus, we then convert the VCF
526 file into FASTA. Note that mutserv does not call insertions, only deletions which are denoted
527 with a gap symbol (within FASTA, it is "-"), thus ensuring that the sequence length is always
528 exactly the length of rCRS, i.e. 16,569 bases, and that sequences align correctly without re-
529 alignment. The same holds for the Genbank files processed by McInerny et al. (25). Thus, we
530 can simply combine FASTA files of cohorts by pasting them into one combined FASTA file. The
531 combined FASTA file was then loaded into Jalview (44) where a Neighbour Joining-based tree
532 was computed and exported into Newick format. This tree was then visualized with ggtree
533 (45) using various graphical settings.
534
535 Ethics statement
536 This study was approved by the Central Institution Review Board, Al-Neelain University,
537 Sudan, IRB Serial Number NU-IRB-17-7-7-91. All subjects gave written informed consent in
538 accordance with the Declaration of Helsinki.
539
540
541 **Data availability**
542 The public data compiled as well as the corresponding publications and repository IDs are
543 provided in Supplementary Table 1. Sudanese raw sequencing data generated as part of this
544 study is available from the European Genome Phenome Archive (EGA) under study ID
545 EGAD00001008315. The analysis workflow used to run haplocheck, perform variant calling
546 and generate and compile mitochondrial sequences is available at
547 https://github.com/iwohlers/2020_mt_analyses.
548
549
550 **Acknowledgements**
551 The authors thank all volunteers who donated blood for the study as well as everyone involved
552 in the collection of blood samples. We also thank Miriam Freitag for excellent technical
553 assistance. IW, HB and SI acknowledge funding by the Deutsche Forschungsgemeinschaft

(DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 22167-390884018. AF, IS, HB and IW acknowledge support through the high-performance computer cluster (OMICS-Cluster) of the University of Lübeck.

Conflict of interest

The authors have declared no conflicting interests.

References

1. Wallace DC, Fan W, Procaccio V. Mitochondrial energetics and therapeutics. Annu Rev Pathol. 2010;5:297–348.
2. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981 Apr 9;290(5806):457–65.
3. Chandel NS. Evolution of Mitochondria as Signaling Organelles. Cell Metab. 2015 Aug 4;22(2):204–6.
4. Wang W, Zhao F, Ma X, Perry G, Zhu X. Mitochondria dysfunction in the pathogenesis of Alzheimer's disease: recent advances. Mol Neurodegener. 2020 May 29;15(1):30.
5. Hsu C-C, Tseng L-M, Lee H-C. Role of mitochondrial dysfunction in cancer progression. Exp Biol Med (Maywood). 2016 Jun;241(12):1281–95.
6. Wallace DC. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. Annu Rev Genet. 2005;39:359–407.
7. van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009 Feb;30(2):E386-394.
8. van Oven M. PhyloTree Build 17: Growing the human mitochondrial DNA tree. Forensic Science International: Genetics Supplement Series. 2015 Dec 1;5:e392–4.
9. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt H-J, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016 08;44(W1):W58-63.
10. Rishishwar L, Jordan IK. Implications of human evolution and admixture for mitochondrial replacement therapy. BMC Genomics. 2017 Feb 8;18(1):140.

586 11. Lippold S, Xu H, Ko A, Li M, Renaud G, Butthof A, et al. Human paternal and maternal
587 demographic histories: insights from high-resolution Y chromosome and mtDNA
588 sequences. Investig Genet. 2014;5:13.
589 12. Rieux A, Eriksson A, Li M, Sobkowiak B, Weinert LA, Warmuth V, et al. Improved
590 calibration of the human mitochondrial clock using ancient genomes. Mol Biol Evol.
591 2014 Oct;31(10):2780–92.
592 13. Laricchia KM, Lake NJ, Watts NA, Shand M, Haessly A, Gauthier L, et al. Mitochondrial
593 DNA variation across 56,434 individuals in gnomAD. bioRxiv. 2021 Jan
594 1;2021.07.23.453510.
595 14. Bergström A, Stringer C, Hajdinjak M, Scerri EML, Skoglund P. Origins of modern human
596 ancestry. Nature. 2021 Feb;590(7845):229–37.
597 15. Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli E-L, Silva NM, et al. A
598 "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J
599 Hum Genet. 2012 Apr 6;90(4):675–84.
600 16. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang
601 HM, et al. A global reference for human genetic variation. Nature. 2015 Oct
602 1;526(7571):68–74.
603 17. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational
604 constraint spectrum quantified from variation in 141,456 humans. Nature.
605 2020;581(7809):434–43.
606 18. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of
607 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021
608 Feb;590(7845):290–9.
609 19. Wohlers I, Künstner A, Munz M, Olbrich M, Fähnrich A, Calonga-Solís V, et al. An
610 integrated personal and population-based Egyptian genome reference. Nat Commun.
611 2020 Sep 18;11(1):4719.
612 20. Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, et al. Tracing the route of
613 modern humans out of Africa by using 225 human genome sequences from Ethiopians
614 and Egyptians. Am J Hum Genet. 2015 Jun 4;96(6):986–91.
615 21. Schuenemann VJ, Peltzer A, Welte B, van Pelt WP, Molak M, Wang C-C, et al. Ancient
616 Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in
617 post-Roman periods. Nat Commun. 2017 May 30;8:15694.

618 22. Weissensteiner H, Forer L, Fendt L, Kheirkhah A, Salas A, Kronenberg F, et al.
619 Contamination detection in sequencing studies using the mitochondrial phylogeny.
620 Genome Res. 2021 Jan 15;
621 23. Brandstätter A, Sänger T, Lutz-Bonengel S, Parson W, Béraud-Colomb E, Wen B, et al.
622 Phantom mutation hotspots in human mitochondrial DNA. Electrophoresis. 2005
623 Sep;26(18):3414–29.
624 24. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into
625 human genetic variation and population history from 929 diverse genomes. Science.
626 2020 20;367(6484).
627 25. McInerney TW, Fulton-Howard B, Patterson C, Paliwal D, Jermiin LS, Patel HR, et al. A
628 globally diverse reference alignment and panel for imputation of mitochondrial DNA
629 variants. BMC Bioinformatics. 2021 Sep 1;22(1):417.
630 26. Maca-Meyer N, González AM, Pestano J, Flores C, Larruga JM, Cabrera VM.
631 Mitochondrial DNA transit between West Asia and North Africa inferred from U6
632 phylogeography. BMC Genet. 2003 Oct 16;4:15.
633 27. Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, et al. The mtDNA legacy
634 of the Levantine early Upper Palaeolithic in Africa. Science. 2006 Dec
635 15;314(5806):1767–70.
636 28. Harich N, Costa MD, Fernandes V, Kandil M, Pereira JB, Silva NM, et al. The trans-
637 Saharan slave trade - clues from interpolation analyses and high-resolution
638 characterization of mitochondrial DNA lineages. BMC Evol Biol. 2010 May 10;10:138.
639 29. Soares P, Alshamali F, Pereira JB, Fernandes V, Silva NM, Afonso C, et al. The Expansion
640 of mtDNA Haplogroup L3 within and out of Africa. Mol Biol Evol. 2012 Mar;29(3):915–
641 27.
642 30. Pennarun E, Kivisild T, Metspalu E, Metspalu M, Reisberg T, Moisan J-P, et al. Divorcing
643 the Late Upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in
644 Africa. BMC Evol Biol. 2012 Dec 3;12:234.
645 31. Fregel R, Méndez FL, Bokbot Y, Martín-Socas D, Camalich-Massieu MD, Santana J, et al.
646 Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb
647 from both the Levant and Europe. Proc Natl Acad Sci U S A. 2018 Jun 26;115(26):6774–
648 9.

649 32. Kujanová M, Pereira L, Fernandes V, Pereira JB, Cerný V. Near eastern neolithic genetic
650 input in a small oasis of the Egyptian Western Desert. Am J Phys Anthropol. 2009
651 Oct;140(2):336–46.
652 33. Costa MD, Cherni L, Fernandes V, Freitas F, Ammar El Gaaied AB, Pereira L. Data from
653 complete mtDNA sequencing of Tunisian centenarians: testing haplogroup association
654 and the "golden mean" to longevity. Mech Ageing Dev. 2009 Apr;130(4):222–6.
655 34. Rodriguez-Flores JL, Fakhro K, Agosto-Perez F, Ramstetter MD, Arbiza L, Vincent TL, et al.
656 Indigenous Arabs are descendants of the earliest split from ancient Eurasian
657 populations. Genome Res. 2016 Feb;26(2):151–62.
658 35. Cabrera VM, Marrero P, Abu-Amero KK, Larruga JM. Carriers of mitochondrial DNA
659 macrohaplogroup L3 basal lineages migrated back to Africa from Asia around 70,000
660 years ago. BMC Evol Biol. 2018 Jun 19;18(1):98.
661 36. Olivieri A, Pala M, Gandini F, Hooshiar Kashani B, Perego UA, Woodward SR, et al.
662 Mitogenomes from two uncommon haplogroups mark late glacial/postglacial
663 expansions from the near east and neolithic dispersals within Europe. PLoS One.
664 2013;8(7):e70492.
665 37. González AM, García O, Larruga JM, Cabrera VM. The mitochondrial lineage U8a reveals
666 a Paleolithic settlement in the Basque country. BMC Genomics. 2006 May 23;7:124.
667 38. Gandini F, Achilli A, Pala M, Bodner M, Brandini S, Huber G, et al. Mapping human
668 dispersals into the Horn of Africa from Arabian Ice Age refugia using mitogenomes. Sci
669 Rep. 2016 May 5;6:25472.
670 39. Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, et al. Correcting for purifying
671 selection: an improved human mitochondrial molecular clock. Am J Hum Genet. 2009
672 Jun;84(6):740–59.
673 40. Hodgson JA, Mulligan CJ, Al-Meeri A, Raaum RL. Early back-to-Africa migration into the
674 Horn of Africa. PLoS Genet. 2014 Jun;10(6):e1004393.
675 41. Serra-Vidal G, Lucas-Sanchez M, Fadhlaoui-Zid K, Bekada A, Zalloua P, Comas D.
676 Heterogeneity in Palaeolithic Population Continuity and Neolithic Expansion in North
677 Africa. Curr Biol. 2019 Nov 18;29(22):3953-3959.e4.
678 42. Russlies J, Fähnrich A, Witte M, Yin J, Benoit S, Gläser R, et al. Polymorphisms in the
679 Mitochondrial Genome Are Associated With Bullous Pemphigoid in Germans. Front
680 Immunol. 2019;10:2200.

681 43. Weissensteiner H, Forer L, Fuchsberger C, Schöpf B, Kloss-Brandstätter A, Specht G, et al.
682 mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA
683 in the cloud. Nucleic Acids Res. 2016 Jul 8;44(W1):W64-69.
684 44. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a
685 multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009 May
686 1;25(9):1189–91.
687 45. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. Curr Protoc Bioinformatics.
688 2020 Mar;69(1):e96.

689