

Investigating the Evolutionary Importance of Denisovan Introgressions in Papua New Guineans and Australians

Ya Hu,^{1,2,4} Qiliang Ding,^{1,2,3,4} Yi Wang,¹ Shuhua Xu,²
Yungang He,² Minxian Wang,² Jiucun Wang,¹ Li Jin^{1,2}

¹ State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200438, China.

10 ² CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological
11 Sciences, Chinese Academy of Sciences, Shanghai 200031, China.

12 ³ Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA.

13 ⁴ YH and QD contributed equally to this work.

Corresponding author: Li Jin, lijin@fudan.edu.cn

For Supplemental Files 1 and 2, and the software package of the method developed in this paper, please contact qd29@cornell.edu

Running title: Denisovan Introgression in Papuans

21 Keywords: Denisovan, Archaic Introgression, Papua New Guinean, Australian, Local Adaptation.

23 **Abstract**

24 Previous research reported that Papua New Guineans (PNG) and Australians contain
25 introgressions from Denisovans. Here we present a genome-wide analysis of Denisovan
26 introgressions in PNG and Australians. We firstly developed a two-phase method to detect
27 Denisovan introgressions from whole-genome sequencing data. This method has relatively
28 high detection power (79.74%) and low false positive rate (2.44%) based on simulations.
29 Using this method, we identified 1.34 Gb of Denisovan introgressions from sixteen PNG
30 and four Australian genomes, in which we identified 38,877 Denisovan introgressive alleles
31 (DIAs). We found that 78 Denisovan introgressions were under positive selection. Genes
32 located in the 78 introgressions are related to evolutionarily important functions, such as
33 spermatogenesis, fertilization, cold acclimation, circadian rhythm, development of brain,
34 neural tube, face, and olfactory pit, immunity, etc. We also found that 121 DIAs are missense.
35 Genes harboring the 121 missense DIAs are also related to evolutionarily important functions,
36 such as female pregnancy, development of face, lung, heart, skin, nervous system, and
37 male gonad, visual and smell perception, response to heat, pain, hypoxia, and UV, lipid
38 transport, metabolism, blood coagulation, wound healing, aging, etc. Taken together, this
39 study suggests that Denisovan introgressions in PNG and Australians are evolutionarily
40 important, and may help PNG and Australians in local adaptation. In this study, we also
41 proposed a method that could efficiently identify archaic hominin introgressions in modern
42 non-African genomes.

43 **Introduction**

44 Genomic introgressions from two extinct hominin species that used to reside on the
45 Eurasian continent, namely Neanderthal and Denisovan, were detected in modern non-
46 Africans (Green et al. 2010, Reich et al. 2010). Neanderthals were found to have admixed
47 with the ancestors of all non-Africans. Denisovan introgressions, in contrast, were mostly
48 found in PNG and Australians (Reich et al. 2011).

49 Despite the relatively low proportion, archaic hominin introgressions in non-Africans
50 were found of great evolutionary importance. Neanderthal and Denisovan introgressions
51 were found at immunity genes *STAT2*, *OAS* gene clusters, and *HLA* Class I genes (Abi-
52 Rached et al. 2011, Mendez et al. 2012a, 2012b). Neanderthal introgression encompassing
53 *HYAL2*, a gene related to the response to UV-B irradiation, was found at high frequency
54 and under positive selection in East Asians (Ding et al. 2014a). Neanderthal introgression
55 at the skin color gene *MC1R* was found to have introduced a functional missense allele
56 V92M into the gene pool of anatomically modern human (AMH, Ding et al. 2014b). These
57 studies suggest that archaic hominin introgressions might help modern non-Africans in
58 local adaptation.

59 In addition, two previous studies have proposed methods in identifying Neanderthal
60 introgressions in Eurasians at genome-wide scale (Sankararaman et al. 2014, Vernot and
61 Akey 2014). These methods are based on the high divergence between Neanderthal
62 introgressions and Africans, as well as the recent introgression time. The method proposed
63 by Plagnol and Wall (2006), further extended by Vernot and Akey (2014), is based on
64 high LD between variants absent from Africans. The method proposed by Sankararaman
65 et al. (2014) is based on derived alleles that are consistent with Neanderthal but absent
66 from Africans. They (Sankararaman et al. 2014, Vernot and Akey 2014) subsequently
67 analyzed evolutionary importance of Neanderthal introgressions. To date, however, no
68 study has focused on analyzing evolutionary importance of Denisovan introgressions in
69 PNG and Australians.

70 One major issue in studying archaic hominin introgressions in AMH is to confidently
71 demonstrate archaic origin of the identified introgressions. Improperly handling this issue
72 could introduce false positives (Ding et al. 2014c). To address this issue, a previous study

73 (Ding et al. 2014b) proposed a three-step approach. Putative introgressions should satisfy
74 the following three criteria to be considered as from archaic introgression: (1) close
75 phylogenetic relationship with Neanderthal or Denisovan; (2) the divergence time with
76 Neanderthal or Denisovan postdates the African – archaic hominin population split time
77 (i.e., 550 thousand years ago [KYA], Prüfer et al. 2014); and (3) reject the alternative model
78 (i.e., incomplete lineage sorting model).

79 Here we present the analysis of Denisovan introgressions in PNG and Australians.
80 We proposed a two-phase method to confidently identify Denisovan introgression. This
81 method has relatively high detection power and low false positive rate. We then applied
82 this method to twenty PNG and Australians genomes, and identified 1.34 Gb of Denisovan
83 introgressive haplotypes and 38,877 DIAs. We found that 78 Denisovan introgressions were
84 under positive selection, and 121 DIAs are missense. Genes in the 78 positively selected
85 introgressions and harboring the 121 missense DIAs are related to a variety of evolutionarily
86 important functions. In conclusion, Denisovan introgressions may play an important role
87 in the evolution and local adaptation of PNG and Australians.

88

89 **Results**

90 *A Two-phase Method in Identifying Denisovan Introgressions*

91 Here we propose a two-phase method in identifying Denisovan introgressions in
92 modern non-African genomes. The first phase is the *identification phase*. This method in
93 identifying archaic hominin introgressions (denoted as *introgressive haplotypes* hereafter)
94 is based on the long divergence time between archaic hominins and AMH. The second
95 phase is the *filtering phase*, in which we employed an established three-step approach
96 (Ding et al. 2014b) to reduce false positives.

97 Since the divergence time between African and archaic hominin (550 KYA, Prüfer et
98 al. 2014) is much longer than that between African and non-African (<100 KYA, Jin and
99 Su 2000), more mutations are expected to be accumulated on the archaic hominin lineage
100 than on the non-African lineage after their divergence with Africans, and most of such
101 mutations are absent from Africans. For two alleles on a SNP, if an allele is absent from
102 all the African genomes, while could be observed on the non-African genomes, this allele
103 is defined as an *E-allele* (Fig. 1). Such alleles have also been studied previously (Plagnol
104 and Wall 2006, Wall et al. 2009, Sankararaman et al. 2014, Vernot and Akey 2014). There
105 are three possible sources for E-alleles. The first source is mutations in archaic hominins
106 that occurred after their divergence with Africans. The second source is mutations in non-
107 Africans that occurred after their divergence with Africans. The third source is alleles that
108 existed in the ancestors of Africans and non-Africans, but were not observed in African
109 samples, due to either genetic drift or a limited African sample size. Based on simulation
110 (Supplemental Text, Supplemental Table 1), we showed that introgressive haplotypes
111 have higher density of E-alleles (denoted as *E-allele rate* hereafter) than the rest of the
112 non-African genomes. We then implemented a hidden Markov model (HMM)-based
113 approach with two hidden states (i.e., *lower* and *higher* E-allele rate) to partition each non-
114 African genome into segments with different E-allele rates (see Material and Methods).
115 The *first* hidden state includes segments with *lower* E-allele rates, and the *second* hidden
116 state includes segments with *higher* E-allele rates. To reduce noises, segments with length
117 less than 10 Kb will be discarded. We consider segments with higher E-allele rates (i.e.,
118 labeled as the second state) are more likely of archaic origin, and these segments will

119 be subjected to filtering by the second phase to minimize false positives.

120 The second phase is a filtering process designed to reduce false positive rate. We
121 adopted a previously established three-step approach (Ding et al. 2014b) in the second
122 phase. The segments identified by the first phase will need to satisfy the following three
123 criteria to be considered as from Denisovan introgression: (1) the segment is phylogenetically
124 closest to Denisovan, among Denisovan, Neanderthal, African, and chimpanzee sequences;
125 (2) the divergence time between the segment and Denisovan is < 550 KYA (Prüfer et al.
126 2014), and the divergence time between the segment and Denisovan is less than that
127 between the segment and Neanderthal; and (3) incomplete lineage sorting model is rejected
128 for the segment by length of the segment (i.e., > 0.01089 cM, details see Material and
129 Methods, Ding et al. 2014a). This filtering phase could reduce false positives introduced
130 by ancient population structure, background selection, and GC-biased gene conversion
131 (see Discussion).

132 In summary, we proposed a two-phase method in identifying Denisovan introgression
133 in this section. We evaluated the detection power and false positive rate of this method
134 in the following section.

135

136 *Detection Power and False Positive Rate of the Two-phase Method*

137 We evaluated the performance (i.e., detection power and false positive rate) of the two-
138 phase method by simulation. Demographic parameters were from Vernet and Akey (2014,
139 for Eurasians, Africans and archaic hominins) and Prüfer et al. (2014, for PNG). The archaic
140 hominin gene flow rate was from Vernet and Akey (2014), and the archaic hominin – African
141 divergence time (550 KYA) was from Prüfer et al. (2014). Parameters for recombination
142 hotspot and recombination rate were from Hellenthal et al. (2008).

143 Using msHOT (Hellenthal and Stephens 2007), we simulated non-African genomes
144 with archaic introgressions. Sequence length for each simulation was set to 1 Mb. We define
145 true archaic introgressions on the simulated non-African sequences as segments that (1)
146 coalesced first with archaic hominin (based on the trees generated by msHOT); (2) diverged
147 with archaic hominin < 550 KYA (based on the divergence times provided by msHOT);

148 and (3) length of the segments are consistent with an introgression time after 550 KYA
149 (i.e., genetic length > 0.01089 cM).

150 We then employed our method to the simulated non-African sequences, and evaluated
151 the performance of our method. Detection power is defined as the proportion of the simulated
152 archaic introgressions that were detected by the two-phase method, and false positive
153 rate is defined as the proportion of the segments identified by the two-phase method that
154 are not simulated archaic introgressions. We repeated the simulation for 1,000 times to
155 evaluate detection power and false positive rate.

156 In the first phase of our method, there is an adjustable parameter named penalty value
157 (λ). To obtain an optimized λ , we calculated detection power and false positive rate of
158 our method using different λ values. Detection power reached maximum of 79.74% at
159 $\lambda=14$, with false positive rate of 2.44% (Table 1, Supplemental Table 2). Therefore, in the
160 following analyses on PNG and Australians, we set penalty value $\lambda=14$.

161 As aforementioned, we assume segments with *higher* E-allele rate identified by the first
162 phase are more likely of archaic origin, and these segments will be subjected to filtering
163 by the second phase. To provide support for this assumption, we hereby allow *all* segments
164 identified by the first phase to enter the second phase, and re-calculated detection power
165 and false positive rate (Supplemental Table 2). It was observed that the false positive rate
166 rises, while maintaining similar detection power, which is in support of our assumption.

167

168 *Detecting Denisovan Introgressions in PNG and Australian Genomes*

169 We applied the two-phase method to the sixteen PNG genomes and the four Australian
170 genomes in the Simons Genome Diversity Project dataset with $\lambda=14$ where the detection
171 power reached maximum in simulation. We identified 1.34 Gb of Denisovan introgressive
172 haplotypes. We estimated that PNG and Australians carry $1.42\pm0.04\%$ more Denisovan
173 ancestry than Chinese Han in Beijing (CHB). This estimation is consistent with Meyer et
174 al. (2012), which estimated that Papuan carry $2.0\pm0.9\%$ more Denisovan ancestry than
175 Han Chinese. We also replicated the findings in Mendez et al. (2012b), which reported
176 Denisovan introgression encompassing the OAS gene cluster in Papuans. Furthermore,

177 the identified Denisovan introgressive haplotypes diverged with Denisovan, Neanderthal,
178 and Africans at 353.24 KYA, 641.97 KYA, and 957.77 KYA, respectively. This is consistent
179 with reports in Mendez et al. (2012b) and Prüfer et al. (2014). The identified Denisovan
180 introgressive haplotypes are included in Supplemental File 1, and is the data source for
181 the subsequent analyses.

182

183 ***Denisovan Introgressive Alleles***

184 To investigate evolutionary importance of Denisovan introgressions, we identified alleles
185 that were introduced by Denisovan introgression in PNG and Australians (denoted as *DIA*
186 [*Denisovan Introgressive Alleles*] hereafter). Alleles that absent in Africans but carried by
187 all Denisovan introgressive haplotypes were considered as *candidate DIAs*. To minimize
188 effect of possible sequencing error, candidate DIAs that are singletons, doubletons, or
189 tripletons were discarded.

190 We then implemented three stringent criteria identical to the three-step approach used
191 in the second phase of the two-phase method to filter the candidate DIAs to minimize false
192 positives. Specifically, a given candidate DIA could pass the stringent filter only if *all* of
193 the haplotypes carrying the candidate DIA satisfy the three criteria: (1) phylogenetically
194 closest to the Denisovan among PNG and Australian haplotypes, chimpanzee, African,
195 Neanderthal, and Denisovan; (2) diverged with Denisovan < 550 KYA; and (3) genetic
196 length is > 0.01089 cM to reject the incomplete lineage sorting model.

197 For example, for the candidate DIA rs117911431-T (chr2: 98206189, GRCh37), there
198 are 80.00% (32/40) of the PNG and Australian haplotypes carrying this allele. To validate
199 Denisovan origin for rs117911431-T, we first reconstructed a phylogenetic tree for PNG
200 and Australian haplotypes, along with Denisovan, Neanderthal, African, and chimpanzee
201 haplotypes (Fig. 2A). It was observed that all haplotypes carrying the rs117911431-T
202 coalesced with Denisovan first, before joining the rest of haplotypes that do not carry this
203 allele. Second, the haplotypes carrying the rs117911431-T diverged with Denisovan at
204 153.49 KYA, which postdates the divergence time between Denisovan and AMH (550
205 KYA). Third, the haplotypes carrying the rs117911431-T are 0.13 cM long, and therefore
206 the incomplete lineage sorting model could be rejected ($p < 2.98 \times 10^{-16}$). In summary, all

207 haplotypes that carry the rs117911431-T satisfy the three stringent criteria, and thus the
208 rs117911431-T is considered as a DIA.

209 In total, 38,877 alleles passed the aforementioned filtering process, and were thus
210 considered as DIAs. To identify regions containing Denisovan introgression, we computed
211 pairwise linkage disequilibrium (LD) r^2 score for the identified DIAs. We define Denisovan
212 introgressive regions (DIRs) as regions containing a set of DIAs in complete LD ($r^2=1.00$),
213 and a DIA randomly chosen from each DIR is used as a *tag DIA* to represent the DIR. In
214 total, we identified 2,802 DIRs. The DIAs and DIRs are listed in Supplemental File 2 and
215 Supplemental Table 3. It was observed that some tag DIAs are in LD (but not in complete
216 LD), these tag DIAs could represent a same introgression event.

217 Among the 2,802 DIRs, frequency of Denisovan introgression in 2.64% (74/2,802) and
218 31.12% (872/2,802) of the DIRs reached 50% (20/40) and 25% (10/40), respectively.

219

220 *Denisovan Introgressions in 78 DIRs were Under Positive Selection*

221 We investigated whether the DIRs were under positive selection using the integrated
222 haplotype score (iHS) test (Voight et al. 2006). Since the iHS is most effective in detecting
223 positive selections with the frequency of selected sites $\geq 20\%$, we excluded DIRs with the
224 frequency of the tag DIAs $< 20\%$ from the analysis for positive selection. Further, false
225 positive rate of the iHS is not likely be elevated by archaic hominin introgression (unpublished
226 results).

227 We first computed the standardized iHS for SNPs with minor allele frequency $\geq 5\%$ in
228 the twenty PNG and Australian genomes using the Whole-Genome Homozygosity Analysis
229 and Mapping Machina (Voight et al. 2006). For each DIR, the Denisovan introgression in
230 the DIR will be considered as under positive selection if more than one (not including one)
231 of the SNPs in strong LD ($r^2 \geq 0.80$) with the tag DIA has standardized $|iHS| \geq 2.00$, since
232 extreme iHS values (i.e., $|iHS| \geq 2.00$) caused by positive selection are usually in strong LD
233 (Voight et al. 2006). We intend to reduce false positive by not including those regions with
234 only one extreme iHS value. We further applied the Extended Haplotype Homozygosity
235 (EHH) test (Sabeti et al. 2002) and drew haplotype bifurcation graph (Gautier and Vitalis,

236 2012) for the SNP with highest |iHS| to validate that the Denisovan introgression is the
237 target of positive selection.

238 For example, for the DIR represented by rs373805722-T (chr20: 50305814, GRCh37,
239 carried by 42.50% [17/40] of the PNG and Australian haplotypes), it was observed that
240 three SNPs in complete LD with the rs373805722, namely rs372900695, rs373840702,
241 and rs371966430, have standardized |iHS| of 2.515, 2.515, and 2.458, respectively. This
242 observation suggests that the Denisovan introgression tagged by rs373805722-T could
243 be under positive selection. This conclusion was further supported by EHH (Fig. 2B) and
244 haplotype bifurcation graphs (Fig. 2C), since haplotypes carrying the DIAs have higher
245 haplotype homozygosity. We thus concluded that the Denisovan introgression tagged by
246 rs373805722-T was under positive selection in PNG and Australians.

247 Totally, we found that Denisovan introgressions in 78 DIRs were under positive selection
248 (Table 2, Supplemental Table 4). Further, genes located in the 78 positively selected DIRs
249 are related to a variety of evolutionarily important functions, such as reproduction (*SPEF2*),
250 response to environmental conditions (*SAXO1* [temperature], *STK39* [stress], *UBE2E2*
251 [DNA damage], and *KCNH7* [circadian rhythm]), development (*SALL4* [limb and neural
252 tube], *SPEF2* [brain], *ALDH1A3* [face and olfactory pit], *MEGF11* [retina], and *ADAMTS9*
253 [melanocyte]), and immunity (*STK39* and *ZNF639*). Positive selection on the DIRs harboring
254 these genes may suggest that Denisovan introgressions at these genes are adaptive in
255 PNG and Australians.

256

257 *One hundred and twenty-one DIAs are Missense Alleles*

258 Besides positive selection, we are also interested in missense DIAs, since missense
259 alleles are more likely to be functional. We obtained annotation for all 38,877 DIAs from
260 the Variant Effect Predictor of the Ensembl Genome Browser (Flicek et al. 2014). Among
261 all 38,877 DIAs, we found 121 missense DIAs (Table 3, Supplemental Table 5).

262 Similar to the above discovery that the genes located in the positively selected DIRs
263 are evolutionarily important, genes harboring the 121 missense DIAs are also of evolutionary
264 importance. Functions of the genes harboring the missense DIAs include reproduction

265 (*PVRL3*, *PSG1*, *PSG3*, *PSG7*, *SPESP1*, *OR1D2*, and *RXFP2*), response to environmental
266 conditions (*NUP210* [heat], *ATR* [heat, UV, and drug], *PAWR* [vitamin E and pain], *OR1D2*
267 [smell], *HMCN1* [vision], *BBS9* [vision], *VTN* [blood coagulation and wound healing], *AKAP1*
268 [blood coagulation], *PDE9A* [blood coagulation], *TTN* [blood coagulation], *KLK8* [response
269 to wounding], and *PINK1* [toxic substance and hypoxia]), development (*ERBB2* [heart and
270 peripheral nervous system], *RCAN1* [central nervous system], *AHI1* [central nervous system],
271 *CRISPLD2* [face and lung], *KLK8* [skin], and *FRAS1* [skin]), metabolism (*NUP210*, *LDHD*,
272 *ACAD9*, *TNXB*, *SARM1*, *OSBPL3* and *GBP6*), aging (*DDC8*), and immunity (*NUP210*,
273 *ERBB2*, *HPR*, *HP*, *CTBP1*, *PSG3*, *CHMP4C*, *VTN*, *SARM1*, and *GBP6*).

274 In summary, we reported that Denisovan introgressions introduced 121 missense alleles
275 into PNG and Australians, and further suggest that the 121 missense DIAs are located in
276 evolutionarily important genes.

277

278 **Discussion**

279 In this study, we proposed a two-phase method in identifying Denisovan introgression in
280 PNG and Australians. We then evaluated the performance of this method by simulations.
281 This two-phase method could effectively detect Denisovan introgressions (detection power
282 = 79.74%), while maintaining a low false positive rate (2.44%). We applied this two-phase
283 method to twenty PNG and Australian genomes, and identified ~ 1.34 Gb of Denisovan
284 introgressive haplotypes. We further found that Denisovan introgressions in 78 DIRs were
285 under positive selection, and 121 DIAs are missense alleles. In addition, genes located in
286 the 78 positively selected DIRs and harboring the 121 missense DIAs are of evolutionary
287 importance. Taken together, the above findings suggest that Denisovan introgressions are
288 evolutionarily important, and may play an important role in the evolution and local adaptation
289 of PNG and Australians.

290

291 *Control for False Positives*

292 In this study, we proposed a new method to identify Denisovan introgressions in non-
293 African genomes. In this method, we carefully controlled for false positives by implementing
294 a three-step filtering process (Ding et al. 2014b) after the identification process.

295 In the first (i.e., identification) phase of the two-phase method in identifying Denisovan
296 introgressions, we identified candidates of Denisovan introgressive haplotypes based on
297 their high divergence with Africans. However, some non-African haplotypes that are not
298 of Denisovan origin may also have high divergence with Africans, such as haplotypes from
299 incomplete lineage sorting (ancient population structure), background selection, and GC-
300 biased gene conversion.

301 To reduce false positives, we implemented a filtering process after the identification
302 phase. The three criteria used in the filtering process are from an earlier study (Ding et al.
303 2014b), and are summarized from several studies (Mendez et al. 2012a, 2012b, Ding et al.
304 2014a). The three criteria could effectively distinguish true Denisovan introgressions from
305 false positives. In specific, the first criterion (close phylogenetic relationship with Denisovan)
306 can filter out false positives from background selection. The second criterion (divergence

307 time with Denisovan should postdate Denisovan – African divergence time) can filter out
308 false positives from incomplete lineage sorting, background selection, and biased gene
309 conversion. To minimize the impact of local mutation rate variation on time estimations
310 (such as accelerated mutation rate caused by GC-biased gene conversion, Galtier and
311 Duret 2007), we used local mutation rates calibrated from human – chimpanzee pairwise
312 alignment in time estimations. The third criterion (rejection of alternative model by genetic
313 length) can filter out false positives from incomplete lineage sorting. In summary, the filtering
314 phase ensures that most false positives picked up in the identification phase were excluded
315 from the final result. This is consistent with the simulation results (false positive rate 2.44%
316 for PNG, and 6.44% for Eurasians, see next section). This performance, based on our
317 simulations, is comparable to, if not better than, previously proposed approaches.

318 In this study, we also proposed a method to identify DIAs. The aforementioned three-
319 step filtering process was also used in this method to filter out false positives. Specifically,
320 for a given candidate DIA, it will be considered as a DIA only if *all* haplotypes carrying the
321 allele satisfied the aforementioned three criteria (i.e., of Denisovan origin). False positives
322 are not likely to pass this filter, since haplotypes carrying the false positives are not all of
323 Denisovan origin. This stringent filtering process ensures that false positives are minimized
324 in the final result.

325 In summary, we used a previously proposed three-step approach to control for false
326 positives in this study. Since false positives are well controlled, we are confident that the
327 conclusions of this study are not likely be influenced by false positives.

328

329 *Comparison with Other Methods*

330 Before this study, two methods were proposed to identify Neanderthal introgressions
331 in Eurasian genomes (Plagnol and Wall 2006, Sankararaman et al. 2014, Vernot and
332 Akey 2014). We compared our two-phase method in identifying Denisovan introgressions
333 in non-Africans proposed in this paper with the two other methods.

334 The method proposed by Vernot and Akey (2014) aims to detect the pattern of high
335 LD between variants absent from Africans (also see Plagnol and Wall 2006, Wall et al. 2009).

336 Despite both our method and the method of Vernot and Akey (2014) are based on the
337 high divergence between introgressive haplotypes and Africans, we do not investigate
338 LD of neighboring E-alleles (i.e., variants absent from Africans). Instead, we use high E-
339 allele rate as criterion to identify introgressive haplotypes in the first phase of our method.
340 Since the method proposed here is not based on LD, the identification of an introgressive
341 haplotype in a given non-African genome is independent from the allele patterns in other
342 non-African genomes.

343 The method proposed by Sankararaman et al. (2014) focus on alleles that are derived, not
344 observed in Africans and consistent with Neanderthal, which is an effective way to reduce
345 false positives. In the first phase of our method, however, we do not require E-alleles to
346 be derived or consistent with Denisovan. Instead, we implemented a separate filtering
347 process (i.e., the second phase) to reduce false positives. While both methods performed
348 well in false positive control (details see below), the method proposed here may have
349 the potential to detect introgressive haplotypes from unknown archaic hominins in future
350 studies, since the first phase of this method is independent from the genomes of known
351 archaic hominins.

352 In addition to above comparisons, we also compared the performance of our method
353 with the two other methods. Since the evaluation of performance of our method in Results
354 is based on the demography of PNG (detection power = 79.74%, false positive rate =
355 2.44%), we re-evaluated the performance of our method using the Eurasian demography
356 here.

357 Demographic parameters of Eurasians and archaic hominin gene flow rate we used
358 here are same as Vernot and Akey (2014). There are only a few small differences between
359 our simulation and the simulation in Vernot and Akey (2014). Length of simulated Eurasian
360 sequences is 1 Mb in our simulation, and is 50 Kb in Vernot and Akey (2014). Furthermore,
361 our definition of archaic introgressions on the simulated sequences is: (1) coalesced first
362 with archaic hominin (based on trees generated by msHOT); (2) diverged with archaic
363 hominin < 550 KYA (Prüfer et al. 2014); and (3) length of segments are consistent with
364 an introgression time after 550 KYA. In Vernot and Akey (2014), archaic introgressions

365 were identified in trees (generated by ms) in which archaic sequence joins the tree before
366 the simulated join time (< 700 or 400 KYA).

367 For simulated Eurasian genomes, the detection power of our method reached maximum
368 of 85.31% at $\lambda=10$, with false positive rate of 6.44% (Table 1, Supplemental Table 2). The
369 detection power and false positive rate reported in Vernot and Akey (2014) are ~ 30% and
370 20%, respectively. The detection power and false positive rate reported in Sankararaman
371 et al. (2014) are 38.4% to 55.2% and < 10%, respectively. It could therefore be suggested
372 that the performance of our method is comparable to, if not better than, the two other methods.
373 However, this comparison is preliminary, since simulation parameters and definition of true
374 archaic introgressions on simulated sequences in the three studies are not exactly identical.

375

376 *Prospects*

377 In this paper, we proposed a two-phase method in identifying Denisovan introgressions
378 in non-Africans. This method may be applicable to other populations and other types of
379 archaic introgressions. In combination with the method in identifying archaic introgressive
380 alleles (i.e., DIA in this study) proposed in this study, it would be interesting to explore the
381 functional and evolutionary importance of Neanderthal introgressions in Eurasians. Further,
382 our two-phase method may also be applicable in identifying introgressions from unknown
383 archaic hominins, since the first phase of our method is independent from Neanderthal
384 and Denisovan genomes. Without the existence of known archaic hominin genome, however,
385 extra caution should be excised to distinguish true archaic introgressions (from unknown
386 archaic hominins) from false positives.

387 **Material and Methods**

388 *The Two-phase Method in Identifying Denisovan Introgressions*

389 In this paper, we proposed a two-phase method to identify Denisovan introgressions
390 in non-Africans. In the first (i.e., identification) phase, we identified candidate Denisovan
391 introgressions based on their high E-allele rate. In the second (i.e., filtering) phase, we
392 employed a three-step filtering process to reduce false positives. Here we describe the
393 details of this two-phase method.

394 In the first phase, we define *E-allele* as follows: for two alleles on a bi-allelic SNP, if
395 an allele is absent from all African samples (507 Africans in the 1000 Genomes Project
396 Phase 3 [1000 Genomes Project Consortium 2012], and 13 Africans from Prüfer et al. 2014,
397 for details see below), but could be observed in PNG or Australian genome, this allele is
398 defined as an E-allele. E-allele rate of a given segment is calculated as the number of E-
399 alleles on the segment divided by the number of polymorphic sites in the genomic region
400 of the segment. Simulations suggest that introgressive haplotypes have higher E-allele
401 rate than the rest of the non-African genomes (Supplemental Text). We thus developed
402 a hidden Markov model (HMM)-based approach with two hidden states (i.e., lower and
403 higher E-allele rate) to partition each PNG and Australian genome into segments with
404 different E-allele rates. Segments with *lower* E-allele rate were labeled as the *first* state,
405 and segments with *higher* E-allele rate were labeled as the *second* state. To reduce noises,
406 segments with length < 10 Kb were discarded. Segments with higher E-allele rates (i.e.,
407 labeled as the second state) are more likely of Denisovan origin, and will be subjected
408 to filtering by the second phase. Below we describe the HMM-based approach.

409 For a non-African chromosome, we assumed it contains M alleles on M SNPs. F_m is the E-
410 allele status indicator for the m^{th} allele. $F_m=1$ indicates that this allele is an E-allele, while
411 $F_m=0$ indicates that the allele is not an E-allele. We therefore transformed M alleles into
412 a string of 0 and 1s. The goal of the HMM-based approach is to partition the string into
413 N segments and label each segment with a hidden state. We solved the partitioning and
414 labeling problem with a classic Viterbi algorithm. The steps of the HMM-based approach
415 are as follows: (1) set a penalty for state transitions (λ) and an initial E-allele rate for each
416 hidden state; (2) use the Viterbi algorithm to partition the non-African chromosome and

417 label each segment with a hidden state; (3) re-estimate the E-allele rate for both hidden
418 states; (4) repeat steps 2 and 3 until convergence. We set the initial E-allele rate of the first
419 state as the observed genome-wide E-allele rate of the non-African genomes, and the
420 initial E-allele rate of the second state as the observed genome-wide E-allele rate of the
421 Denisovan genome. In our search for Denisovan introgressions in PNG and Australians,
422 the initial E-allele rates for the first and the second states were set as 0.0058 and 0.0276,
423 respectively.

424 Above we described the details of the first phase of the two-phase method. Here we
425 describe the details of the second (i.e., filtering) phase. We intended to reduce false positive
426 rate by implementing this filtering process. Segments identified in the first phase that have
427 higher E-allele rate (i.e., labeled as the second state) are considered as candidates of
428 introgressive haplotypes, and are subjected to this filtering. Candidates will be considered
429 as true Denisovan introgressive haplotypes if they satisfy all three following criteria: (1)
430 phylogenetically closest to Denisovan, among Denisovan, Neanderthal, African, and
431 chimpanzee sequences; (2) diverged with Denisovan < 550 KYA, and the divergence
432 time with Denisovan is less than the divergence time with Neanderthal; and (3) genetic
433 length > 0.01089 cM.

434 For the first criterion, the phylogenetic trees are reconstructed using the parsimony
435 method implemented in PHYLIP (Felsenstein 1989). For the second criterion, we estimated
436 the divergence time between the candidates and Denisovan, Neanderthal and African (a
437 San individual, HGDP01029) using a maximum likelihood-based method (Mendez et al.
438 2012a). The mutation rates used in the time estimations are calibrated using the human
439 (GRCh37) – chimpanzee (PanTro3) pairwise alignment, assuming a human – chimpanzee
440 divergence time of 6,500 KYA. For the third criterion, the purpose is to reject the incomplete
441 lineage sorting model. Based on the incomplete lineage sorting model, the candidates
442 existed in the human gene pool before the Denisovan – African population divergence
443 (i.e., 550 KYA). Probability of persistence of a haplotype over T_p years could be expressed
444 as $p=e^{-(\theta \times T_p)/(100 \times T_g)}$, where T_g is time per generation (20 years), and θ is the genetic length
445 of the haplotype. When $p \leq 0.05$, $T_p = 550,000$ (based on the incomplete lineage sorting
446 model), and $T_g=20$, the θ could be computed as ≥ 0.01089 cM. Thus, the third criterion
447 requires the candidates to be longer than 0.01089 cM to reject the incomplete lineage

448 sorting model. Genetic map used in this study is from Hinch et al. (2011).

449

450 *Evaluation of Performance of the Two-phase Method by Simulations*

451 In Results, we evaluated the detection power and false positive rate of our two-phase
452 method by simulation, based on the demographic parameters of PNG. In Discussion, in
453 order to compare with two other published methods, we evaluated the detection power
454 and false positive rate of our method using the demographic parameters of Eurasians. Here
455 we describe the details of the simulations.

456 Using msHOT (Hellenthal and Stephens 2007), we simulated non-African genomes
457 with archaic introgressions. Demographic parameters for Africans, Eurasians, and archaic
458 hominins are from Vernot and Akey (2014), Schaffner et al. (2005), Tennessen et al. (2012),
459 and Gravel et al. (2011). Demographic parameters for PNG are from Prüfer et al. (2014).
460 In specific, the demographic parameters we used are as follows. Generation time was set
461 to 20 years, and mutation rate was set to 2×10^{-8} per base per generation (Fu et al. 2014).
462 Population size (N_e) of archaic hominin after their divergence with Africans is 1,500. Before
463 148 KYA, N_e of Africans was 7,300, and after 148 KYA, N_e of Africans was 14,474. There
464 was exponential growth in Africans from 5,115 years ago, and present N_e of Africans is
465 424,000. Ancestors of non-Africans diverged with Africans at 60 KYA. Divergence time
466 between Europeans and East Asians was randomly set between 36 KYA and 50 KYA.
467 Before the European – East Asian divergence, N_e of Eurasians was set to the same as
468 N_e of Europeans. Before 23 KYA, total N_e of East Asians and Europeans was randomly
469 set between 4,000 and 18,353, with ratio of N_e of Europeans to N_e of East Asians randomly
470 set between 1 and 2.5. We require N_e of Europeans less than 9,475 and N_e of East Asians
471 less than 8,879 before 23 KYA. There was exponential growth in Europeans and East
472 Asians from 23 KYA to 5,115 years ago, with N_e of Europeans and East Asians increased
473 to 9,475 and 8,879, respectively. Since 5,115 years ago, there was rapid exponential
474 growth of N_e in Europeans and East Asians, and the present N_e of Europeans and East
475 Asians are 512,000 and 1,370,990, respectively. For Papuans, they diverged with Eurasians
476 at 36 KYA, and we assume that their demographic history is same as East Asians before
477 5,115 years ago. Since 5,115 years ago, there was exponential growth in Papuans, and

478 the present N_e of Papuans is 424,000 (close to Africans, Prüfer et al. 2014). Before the
479 Eurasian divergence, migration rate between Africans and Eurasians was 1.498975×10^{-4} ,
480 i.e., 1.498975×10^{-4} of Africans were new migrants from Eurasians each generation. After
481 the Eurasian divergence, the African – European, African – East Asian, and European –
482 East Asian migration rates were 2.498291×10^{-5} , 7.794668×10^{-6} , and 3.107874×10^{-5} ,
483 respectively. Time of archaic hominin introgression into the ancestors of Eurasians was
484 randomly set between 55 KYA and the Eurasian divergence time, with migration rate 0.0015.
485 This introgression continued for 500 years. Time of archaic hominin introgression into East
486 Asians was 500 years after the Eurasian divergence, with migration rate randomly set
487 between 0.00015 and 0.0005. This introgression continued for 500 years. Time of archaic
488 introgression into Papuans was 500 years after the Papuan – East Asian divergence, with
489 migration rate of 0.0015.

490 In our simulation, sequence length of simulated non-African genome was set to 1 Mb.
491 Parameters for recombination hotspot and recombination rate were from Hellenthal et al.
492 (2008). Background recombination rate was 2.325×10^{-9} between two bases per generation.
493 Number of hotspots for each sequence was randomly drawn from a Poisson distribution
494 with mean value 25 (i.e., one hotspot per 40 Kb on average). Width of each hotspot was
495 randomly set from 1 to 2 Kb. For intensity of each hotspot (i), we set $\log_{10}(i)$ as random
496 value between 1 and 2.5.

497 We performed simulations using the aforementioned parameters. In each simulation,
498 we simulated 1,040 Africans, one archaic hominin, one East Asian, and one PNG (or
499 European, for the simulation using Eurasian demography). On the simulated non-African
500 sequences, we define simulated archaic introgressions as follows, as per the definition of
501 true archaic introgressions: (1) they coalesced first with archaic hominin from trees generated
502 by msHOT in simulation; (2) they diverged with archaic hominin later than 550 KYA, based
503 on divergence times provided by msHOT; and (3) length of the segments is consistent
504 with an introgression time after 550 KYA (i.e., longer than 0.01089 cM, see above). We
505 repeated the simulation for 1,000 times using the PNG demography and the Eurasian
506 demography, respectively.

507 We then employed our two-phase method to the simulated non-African genomes, with

508 different λ values. The initial E-allele rates for the first and second states are the averaged
509 E-allele rates for the simulated non-African (i.e., PNG or Eurasian) and the simulated
510 archaic hominin sequences in the 1,000 simulations, respectively. For the simulated PNG
511 genomes, initial E-allele rates for the first and the second states were 0.0058 and 0.0654,
512 respectively. For the simulated Eurasian genomes, initial E-allele rates for the first and
513 second states were 0.0044 and 0.0547, respectively. Detection power is defined as the
514 proportion of the simulated archaic introgressions that were detected by the two-phase
515 method, and false positive rate is defined as the proportion of the segments identified by
516 the two-phase method that are not simulated archaic introgressions. We computed detection
517 power and false positive rate for each integer λ value from 1 to 25 for the simulated PNG
518 dataset, and from 1 to 15 for the simulated Eurasian dataset.

519

520 *Data Sources*

521 After we demonstrated that our two-phase method could effectively detect Denisovan
522 introgressions in PNG by simulation and obtained an optimized λ value, we applied this
523 method to the sixteen PNG genomes and the four Australian genomes in the Simons
524 Genome Diversity Project dataset.

525 When identifying E-alleles, 507 African genomes in the five African populations in the
526 1000 Genomes Project Phase 3 dataset (Esan in Nigeria, Gambian in Western Divisions in
527 the Gambia, Mende in Sierra Leone, Luhya in Webuya, Kenya, and Yoruba in Ibadan,
528 Nigeria, 1000 Genomes Project Consortium 2012) and 13 high-coverage African genomes
529 from Prüfer et al. (2014, 5 Yoruba, 2 Mbuti, 2 Dinka, 2 San, 2 Mandenka) were used as
530 Africans. In addition, we obtained high-coverage Neanderthal genome from Prüfer et al.
531 (2014), and high-coverage Denisovan genome from Meyer et al. (2012). We phased the
532 13 high-coverage African genomes, the Neanderthal genome, and the Denisovan genome
533 using the SHAPEIT software (Delaneau et al. 2013). In all of our analyses, we focused
534 on bi-allelic SNPs, and discarded tri-allelic SNPs and insertions/deletions.

535

536

537 *Identifying DIAs*

538 To explore the evolutionary importance of the identified Denisovan introgressions, we
539 identified alleles that were introduced by Denisovan introgressions, hereby denoted as
540 *Denisovan introgressive alleles (DIAs)*. To identify DIAs, alleles that absent in Africans but
541 carried by all Denisovan introgressive haplotypes were considered as candidate DIAs. We
542 then applied a stringent filter identical to the three-step approach used in the second phase
543 of the two-phase method to control for false positives. For a given candidate DIA, it will be
544 considered as a DIA if *all* haplotypes carrying the allele are of Denisovan origin, i.e.,
545 satisfy the following three criteria: (1) phylogenetically closest to Denisovan, among PNG
546 and Australian haplotypes, Neanderthal, Denisovan, chimpanzee, and African (HGDP01029,
547 a San individual); (2) diverged with Denisovan no earlier than 550 KYA; and (3) genetic
548 length < 0.01089 cM. For the first criterion, the phylogenies are reconstructed using the
549 neighbor-joining method implemented in PHYLIP and MEGA (Tamura et al. 2013). For
550 the second criterion, the divergence times are estimated using the method in Mendez et
551 al. (2012a). For the third criterion, the purpose is to reject the incomplete lineage sorting
552 model.

553

554 *Scan for Positive Selection, Annotation of DIAs and Genes*

555 We computed the standardized integrated haplotype score (iHS) for SNPs with minor
556 allele frequency more than 5% in the twenty PNG and Australian genomes using the
557 Whole-Genome Homozygosity Analysis and Mapping Machina (Voight et al. 2006). The
558 genetic distance is from Hinch et al. (2011). For each DIR, the Denisovan introgression
559 in the DIR will be considered as under positive selection if more than one (not including
560 one) of the SNPs in strong LD ($r^2 \geq 0.80$) with the tag DIA of the DIR has standardized
561 $|iHS| \geq 2.00$, since extreme iHS values (i.e., $|iHS| \geq 2.00$) caused by positive selection are
562 usually in strong LD (Voight et al. 2006). We further applied the EHH test (Sabeti et al.
563 2002) and drew haplotype bifurcation graph (Gautier and Vitalis, 2012) for the SNP with
564 highest $|iHS|$ to validate that the Denisovan introgression is the target of positive selection.

565 We obtained annotation for the 38,877 DIAs from the Variant Effect Predictor of the
566 Ensembl Genome Browser (Flicek et al. 2014).

567 **Data Access**

568 Denisovan introgressive haplotypes in PNG and Australians identified in this study are
569 included as Supplemental File 1. DIAs identified in this study are included as Supplemental
570 File 2.

571

572 **Acknowledgements**

573 This research was supported by the National Science Foundation of China (31271338
574 and 31330038) and the Chinese National Basic Research Program (2012CB944600).

575

576 **Disclosure Declaration**

577 The authors declare that there is no conflict of interest.

578

579

580 **Figure and Figure Legends**

581 Figure 1. Definition of E-allele.

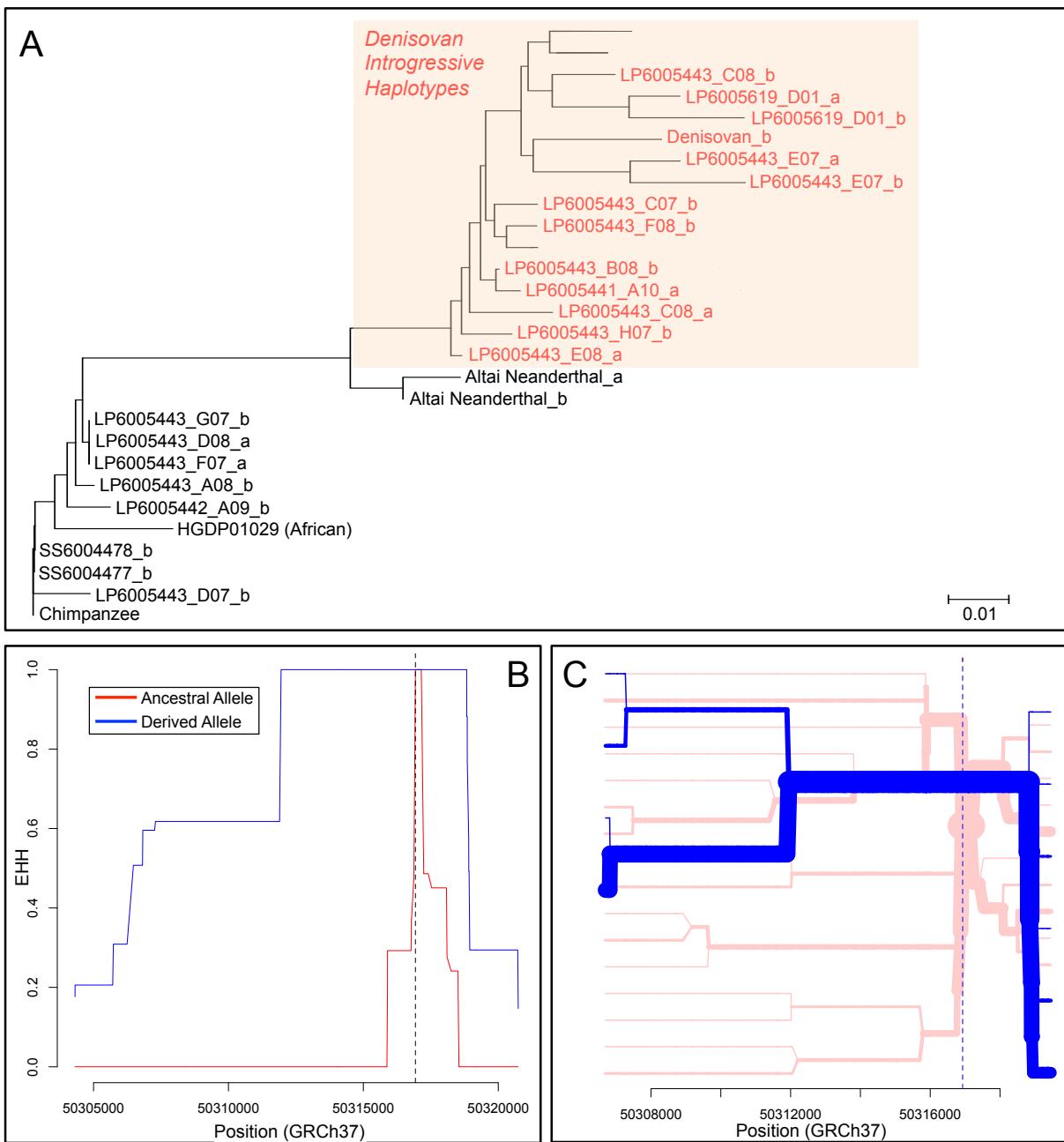
582 For two alleles on a SNP, if an allele is absent from all the African genomes, while could
583 be observed on the non-African genomes, this allele is defined as an E-allele. In the Fig.
584 1, alleles *a* and *b* are E-alleles, since both alleles are absent from all African genomes,
585 while was observed in non-African genomes.



587 Figure 2. Phylogenetic tree, EHH analysis, and haplotype bifurcation graph.

588 (A) Phylogenetic tree for PNG and Australian haplotypes, Denisovan, Neanderthal, and
589 chimpanzee sequences, reconstructed using the neighbor-joining method in MEGA 6.
590 Polymorphic sites within chr2: 98197888–98216073 (GRCh37, shared boundary of all
591 rs117911431-T-carrying-haplotypes) were used to reconstruct the tree. Chimpanzee
592 sequence was used as root. Trivial monophyletic clusters were collapsed to simplify tree
593 presentation. Haplotypes carrying the rs117911431-T were colored in red. It was observed
594 that these haplotypes coalesced with Denisovan haplotype first, before joining the rest of
595 haplotypes that do not carry this allele, suggesting that the rs117911431-T is a DIA. (B)
596 EHH graph of the SNP with highest $|iHS|$ (rs372900695) in the DIR tagged by rs373805722-
597 T. The derived allele of the SNP (rs372900695-A) is also a DIA, and is in complete LD
598 with the tag DIA (rs373805722-T) of the DIR. The haplotype homozygosity for the derived
599 allele (i.e., DIA) decays much slower than that of the ancestral allele, suggesting that the
600 Denisovan introgression tagged by rs373805722-T is the target of positive selection. (C)
601 Haplotype bifurcation graph of the SNP with highest $|iHS|$ (rs372900695) in the DIR tagged
602 by rs373805722-T. The haplotype homozygosity for the derived allele (i.e., DIA) decays
603 much slower than that of the ancestral allele, which is consistent with Fig. 2B, and also
604 suggests that the Denisovan introgression tagged by rs373805722-T is the target of positive
605 selection.

606 (See next page for Figure 2)



607

608

609 **Tables**

610 Table 1. Detection power and false positive rate of our two-phase method (λ from 6 to 15).

611 See Supplemental Table 2 for full table (with λ from 1 to 25). PNG: Estimated using the
612 demographic parameters of PNG. Eurasian: Estimated using the demographic parameters
613 of Eurasian. Numbers in red are the maximum detection power and its corresponding false
614 positive rate.

λ	PNG		Eurasian	
	Detection Power	False Positive Rate	Detection Power	False Positive Rate
6	0.7116	0.0201	0.8405	0.0604
7	0.7425	0.0210	0.8508	0.0606
8	0.7655	0.0207	0.8516	0.0624
9	0.7770	0.0213	0.8512	0.0631
10	0.7837	0.0219	0.8531	0.0644
11	0.7900	0.0224	0.8506	0.0643
12	0.7916	0.0233	0.8473	0.0651
13	0.7957	0.0241	0.8456	0.0672
14	0.7974	0.0244	0.8426	0.0675
15	0.7972	0.0250	0.8397	0.0677

615

616 Table 2. A list of DIRs under positive selection and harbor evolutionarily important genes.

617 For a full list of DIRs under positive selection, see Supplemental Table 4. Chr.: chromosome. Tag: position of the tag DIA of the
618 DIR. Frequency: frequency of the Denisovan introgression. Top Position: position of the SNP with highest |iHS|. r^2 : LD r^2 score
619 between the tag DIA and the allele with highest |iHS|. All positions are in GRCh37. Genes in red are evolutionarily important.

Chr.	Tag	Allele	Frequency	Highest iHS	Top Position	r^2	Gene	Function
20	50305814	T	0.425	2.515	50316930	1.000	ATP9A <i>SALL4</i>	Embryonic limb morphogenesis, neural tube closure
5	35760424	G	0.400	2.015	35776369	1.000	<i>SPEF2</i>	Spermatogenesis, fertilization, brain morphogenesis
15	101432654	G	0.325	3.463	101449530	1.000	<i>LOC100507452 ALDH1A3</i>	Face development, olfactory pit development
2	168951543	T	0.275	2.552	168953002	1.000	<i>STK39</i>	Response to stress, regulation of inflammation
3	23612125	C	0.275	2.260	23630927	1.000	<i>UBE2E2</i>	Cellular response to DNA damage stimulus
15	66534679	T	0.225	3.648	66551937	1.000	<i>MEGF11</i>	Retina layer formation
3	64584252	G	0.225	2.294	64612943	0.861	<i>ADAMTS9</i>	Positive regulation of melanocyte differentiation
2	163726790	A	0.225	3.399	163665773	0.861	<i>KCNH7</i>	Circadian rhythm
3	179024584	G	0.200	2.758	179061537	1.000	<i>ZNF639 MFN1 GNB4</i>	Negative regulation by host of viral transcription
9	18942544	T	0.200	2.410	18969095	0.848	<i>SAXO1</i>	Cold acclimation

620

621 Table 3. Missense DIAs harbored by evolutionarily important genes.

622 Chr.: chromosome. See Supplemental Table 5 for a full list of missense DIAs.

Chr.	Position	Allele	Frequency	Gene	Function
16	84907552	C	0.550	<i>CRISPLD2</i>	Face morphogenesis, lung development
17	37873628	A	0.425	<i>ERBB2</i>	Heart development, peripheral nervous system development, innate immune response
3	13363815	C	0.400	<i>NUP210</i>	Cellular response to heat, carbohydrate metabolic process, viral life cycle
6	135768231	T	0.375	<i>AHI1</i>	Central nervous system development, regulation of behavior
16	75148067	T	0.375	<i>LDHD</i>	D-lactate dehydrogenase (cytochrome) activity
16	72110404	T	0.350	<i>HPR</i>	Positive regulation of defense response to virus by host
16	72094680	G	0.350	<i>HP</i>	Defense response to bacterium, immune system process
3	128598793	A	0.350	<i>ACAD9</i>	Mitochondrial respiratory chain complex I assembly
3	110866284	G	0.350	<i>PVRL3</i>	Fertilization
19	43237142	A	0.325	<i>PSG3</i>	Female pregnancy, defense response
4	1231802	T	0.250	<i>CTBP1</i>	White fat cell differentiation, viral genome replication
19	43372914	T	0.250	<i>PSG1</i>	Female pregnancy
19	43433744	A	0.250	<i>PSG7</i>	Female pregnancy
7	24874134	T	0.250	<i>OSBPL3</i>	Lipid transport
3	142177832	T	0.200	<i>ATR</i>	Cellular response to UV, cellular response to heat, response to drug
8	82665330	T	0.200	<i>CHMP4C</i>	Regulation of viral process
1	185897772	T	0.200	<i>HMCN1</i>	Visual perception, response to stimulus
15	69238594	G	0.175	<i>SPESP1</i>	Fusion of sperm to egg plasma membrane
17	26691654	T	0.150	<i>VTN</i>	Regulation of blood coagulation, regulation of wound healing, immune response
17	26715448	A	0.150	<i>SARM1</i>	Response to glucose, innate immune response
6	32017161	T	0.150	<i>TNXB</i>	Triglyceride metabolic process, fatty acid metabolic process, collagen metabolic process
17	76887402	G	0.150	<i>DDC8</i>	Aging, central nervous system development, response to hormone
1	20971008	G	0.150	<i>PINK1</i>	Cellular response to toxic substance, cellular response to hypoxia
17	55183263	C	0.150	<i>AKAP1</i>	Blood coagulation
21	35893716	T	0.150	<i>RCAN1</i>	Blood circulation, central nervous system development
7	33313567	T	0.125	<i>BBS9</i>	Response to stimulus, visual perception, fat cell differentiation
12	80083615	G	0.125	<i>PAWR</i>	Cellular response to estradiol stimulus, cellular response to vitamin E, sensory perception of pain
21	44117550	T	0.125	<i>PDE9A</i>	Blood coagulation
2	179612635	T	0.125	<i>TTN</i>	Blood coagulation, muscle contraction
1	89848223	A	0.125	<i>GBP6</i>	Metabolic process, immune response, defense response to bacterium
17	2996178	A	0.125	<i>OR1D2</i>	Sensory perception of smell, single fertilization
4	79396653	C	0.125	<i>FRAS1</i>	Skin development
19	51501095	A	0.100	<i>KLK8</i>	Memory, response to wounding, keratinocyte proliferation
13	32365978	T	0.100	<i>RXFP2</i>	Oocyte maturation, male gonad development

625 **References**

626 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from
627 1,092 human genomes. *Nature* 491: 56-65.

628 Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F,
629 Gharizadeh B, Luo M, Plummer FA, et al. 2011. The shaping of modern human immune
630 systems by multiregional admixture with archaic humans. *Science* 334: 89-94.

631 Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. 2013. Haplotype estimation using
632 sequencing reads. *Am J Hum Genet* 93: 687-696.

633 Ding Q, Hu Y, Xu S, Wang J, Jin L. 2014a. Neanderthal introgression at chromosome
634 3p21.31 was under positive natural selection in East Asians. *Mol Biol Evol* 31: 683-695.

635 Ding Q, Hu Y, Xu S, Wang CC, Li H, Zhang R, Yan S, Wang J, Jin L. 2014b.
636 Neanderthal origin of the haplotypes carrying the functional variant Val92Met in the
637 MC1R in modern humans. *Mol Biol Evol* 31: 1994-2003.

638 Ding Q, Hu Y, Jin L. 2014c. Non-Neanderthal origin of the HLA-DPB1*0401. *J Biol
639 Chem* 289: 10252-10252.

640 Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5:
641 164-166.

642 Flícek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P,
643 Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res* 42: D749-D755.

644 Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PL, Aximu-Petri
645 A, Prüfer K, de Filippo C, et al. 2014. Genome sequence of a 45,000-year-old modern
646 human from western Siberia. *Nature* 514: 445-449.

647 Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null
648 hypothesis of molecular evolution. *Trends Genet* 23: 273-277.

649 Gautier M, Vitalis R. 2012. rehh: an R package to detect footprints of selection in
650 genome-wide SNP data from haplotype structure. *Bioinformatics* 28: 1176-1177.

651 Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA,
652 1000 Genomes Project, Bustamante CD. 2011. Demographic history and rare allele
653 sharing among human populations. *Proc Natl Acad Sci U S A* 108: 11983-11988.

654 Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H,
655 Zhai W, Fritz MH, et al. 2010. A draft sequence of the Neandertal genome. *Science*
656 328: 710-722.

657 Hellenthal G, Stephens M. 2007. msHOT: modifying Hudson's ms simulator to
658 incorporate crossover and gene conversion hotspots. *Bioinformatics* 23: 520-521.

659 Hellenthal G, Auton A, Falush D. 2008. Inferring human colonization history using a
660 copying model. *PLoS Genet* 4: e1000078. DOI: 10.1371/journal.pgen.1000078.

661 Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K,
662 Buxbaum SG, Akylbekova EL, et al. 2011. The landscape of recombination in African
663 Americans. *Nature* 476: 170-175.

664 Jin L, Su B. 2000. Natives or immigrants: modern human origin in east Asia. *Nat Rev
665 Genet* 1: 126-133.

666 Mendez FL, Watkins JC, Hammer MF. 2012a. A haplotype at STAT2 Introgressed from
667 Neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am
668 J Hum Genet* 91: 265-274.

669 Mendez FL, Watkins JC, Hammer MF. 2012b. Global genetic variation at OAS1
670 provides evidence of archaic admixture in Melanesian populations. *Mol Biol Evol* 29:
671 1513-1520.

672 Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F,
673 Prüfer K, de Filippo C, et al. 2012. A high-coverage genome sequence from an archaic
674 Denisovan individual. *Science* 338: 222-226.

675 Plagnol V, Wall JD. 2006. Possible ancestral structure in human populations. *PLoS
676 Genet* 2: e105. DOI: 10.1371/journal.pgen.0020105.

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

680 Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW,
681 Stenzel U, Johnson PL, et al. 2010. Genetic history of an archaic hominin group from
682 Denisova Cave in Siberia. *Nature* 468: 1053-1060.

683 Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AM, Ko YC,
684 Jinam TA, Phipps ME, et al. 2011. Denisova admixture and the first modern human
685 dispersals into Southeast Asia and Oceania. *Am J Hum Genet* 89: 516-528.

686 Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB,
687 Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection
688 in the human genome from haplotype structure. *Nature* 419: 832-837.

689 Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, Patterson N,
690 Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day
691 humans. *Nature* 507: 354-357.

692 Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a
693 coalescent simulation of human genome sequence variation. *Genome Res* 15: 1576-
694 1583.

695 Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular
696 Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 30: 2725-2729.

697 Tennesen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R,
698 Liu X, Jun G, et al. 2012. Evolution and functional impact of rare coding variation from
699 deep sequencing of human exomes. *Science* 337: 64-69.

700 Vernot B, Akey JM. 2014. Resurrecting surviving Neandertal lineages from modern
701 human genomes. *Science* 343: 1017-1021.

702 Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection
703 in the human genome. *PLoS Biol* 4: e72. DOI: 10.1371/journal.pbio.0040072.

704 Wall JD, Lohmueller KE, Plagnol V. 2009. Detecting ancient admixture and estimating
705 demographic parameters in multiple human populations. *Mol Biol Evol* 26: 1823-1827.