Received: 15 May 2021 | Revised: 15 July 2021 | Accepted: 11 August 2021

DOI: 10.1111/ahg.12444

ORIGINAL ARTICLE

Annals of human genetics WILEY

The hazards of genotype imputation in chromosomal regions under selection: A case study using the Lactase gene region

Aminah T. Ali | **Anke Liebert** | **Winston Lau** | **Nikolas Maniatis** | **Dallas M. Swallow**

University College London Research Department of Genetics Evolution and Environment, London, UK

Correspondence
Dallas M. Swallow, University College London Research Department of Genetics Evolution and Environment, London, UK
Email: d.swallow@ucl.ac.uk

Present address: Aminah T. Ali, Department of Medical and Molecular Genetics, School of Basic and Medical Biosciences, King's College London, London, UK.

Present address: Anke Liebert, The Francis Crick Institute, 1 Midland Road, London, UK.

Abstract

Although imputation of missing SNP results has been widely used in genetic studies, claims about the quality and usefulness of imputation have outnumbered the few studies that have questioned its limitations. But it is becoming clear that these limitations are real—for example, disease association signals can be missed in regions of LD breakdown. Here, as a case study, using the chromosomal region of the well-known lactase gene, *LCT*, we address the issue of imputation in the context of variants that have become frequent in a limited number of modern population groups only recently, due to selection. We study SNPs in a 500 bp region covering the enhancer of *LCT*, and compare imputed genotypes with directly genotyped data. We examine the haplotype pairs of all individuals with discrepant and missing genotypes. We highlight the nonrandom nature of the allelic errors and show that most incorrect imputations and missing data result from long haplotypes that are evolutionarily closely related to those carrying the derived alleles, while some relate to rare and recombinant haplotypes. We conclude that bias of incorrectly imputed and missing genotypes can decrease the accuracy of imputed results substantially.

KEYWORDS
Derived alleles, haplotypes, imputation, Lactase, selection

1 | INTRODUCTION

Imputation of missing Single Nucleotide Polymorphism (SNP) results (Marchini & Howie, 2010), originally made necessary by the inclusion of different SNPs in commercial genome-wide genotyping platforms, as a way of combining data from different sources, continues to be widely used in disease association studies (see e.g., Bycroft et al., 2018). Many claims have been made about its high level of accuracy (Howie et al., 2012; Howie et al., 2011). Imputation has also been used in the context of population genetics, (Hellenthal et al., 2014; Ilardo et al., 2018) and for ancient DNA where coverage is low (Allentoft et al., 2015; Martiniano et al., 2017). It has been suggested that incorrect imputation can occur for SNPs adjacent to or within regions of the breakdown of Linkage Disequilibrium (LD) (Weng et al., 2014). Here we address another issue, namely that of variants that have become frequent only recently, in particular modern populations, due to selection. We investigate this in relation to variants present in an enhancer upstream of the lactase gene, *LCT*.

Functional regulatory variants in this *LCT* enhancer have emerged at a detectable frequency only in the last 5000 years or so, as assessed from ancient DNA samples (Mathieson et al., 2018), haplotype approaches (Bersaglieri et al., 2004; Coelho et al., 2005) and also using modelling approaches (Itan et al., 2009). The recent history of these variants, with little time for recombination, combined with the selection that has increased their frequency, means that they each occur on very extended haplotypes (Bersaglieri et al., 2004; Liebert et al., 2017; Poulter et al., 2003; Ranciaro et al., 2014). So far, five variants have been convincingly confirmed as functional ((Liebert et al., 2016) and references therein), of which one, rs4988235 (*−13910*T*), has been studied in most detail and is present predominantly in Europeans (reviewed in Ingram et al., 2009; Segurel & Bon, 2017). Imputation has been used to help assess its presence in ancient DNA samples with missing data (Allentoft et al., 2015; Ilardo et al., 2018). The apparent detection of lactase persistence (LP) in samples of bronze age pastoralists from the western Steppe region contributed to the belief held by some that LP arose in that population, even though the evidence for this is not strong, since the early occurrences (~4000 BP) were found in western Europe as well as Ukraine (Mathieson et al., 2015, 2018).

Here we test the accuracy of imputation of 10 LP enhancer region SNPs by comparing with true genotyping results, using modern DNA samples. While 95% of the genotypes are correct the 5% that are wrongly imputed or fail to impute are non-random with respect to gain or loss of an allele or genotype. We examine the haplotype pairs of these individuals in an attempt to understand the basis of this bias.

2 | METHODS

2.1 | Strategy

The global 1000 Genomes Phase III data (1000 Genomes Project Consortium et al., 2012, 2015) were used as a reference panel for all the imputations we conducted, with all groups combined as recommended. For the study panels, we used two different datasets (Population studies 1 and 2). Each study provided a "validation" set of directly genotyped (true) genotypes that were then compared with the genotypes we obtained after imputation using the SHAPEIT2 +IMPUTE 2 approach (Marchini & Howie, 2010). The overall experimental strategy is shown in Figure S1A, B.

2.2 | Population Study 1

Our genotype and sequencing data were available for 52 SNPs (Table S1A, B, which shows the full nomenclature of key SNPs) covering a 1.77Mb region (from rs1446525 at chr2:135637847 to rs6711718 at chr2:137407012; build GRCh37/hg 19), from multiple populations (Liebert et al., 2017). Figure 1 shows the positions of the genotyped SNPs and genes, in relation to a Linkage Disequilibrium Unit (LDU) map of the 1.77 Mb and 500 bp genomic intervals on chromosome 2. Forty-two of the 52 SNPs in the study panel were used as scaffold SNPs (blue diamonds) for imputing genotypes for the 10 markers (red diamonds) within the 500 bp segment covering the enhancer region. Scaffold SNPs were selected so that they were distributed to take into account LD, rather than physical distance (Liebert et al., 2017). Plotting the high-resolution genetic maps with distances expressed in LDU on the physical map shows the non-linear relationship between the two and the underlying structure of LD, which is typically a "Block-Step" structure. LDU blocks represent areas of conserved LD and low haplotype diversity, while Steps (increasing LDU distances) define LD breakdown, primarily caused by recombination since crossover profiles agree precisely with the corresponding LDU steps (Maniatis et al., 2007).

The 802 individuals included in this study were of "European", "Middle Eastern" and "African" origin (Liebert et al., 2017), and further details about these data collections are given in the Supporting Information of that paper (see Liebert et al., 2017, Table S2b for samples). Diplotypes of the samples tested are also shown in the Supporting Information of the current paper. For imputation, the three continental groups were phased and imputed independently.

2.3 | Population Study 2

As an independent analysis, public datasets from HapMap were used. SNP genotype data for the same 1.77 Mb segment of chromosome 2 were obtained for individuals from the HapMap populations ASW, CEU, GIH, LWK, MXL, TSI, YRI, CHB and JPT, for which there was rs4988235 SNP data https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html. These data were used [rather than whole genome sequence (WGS) data] to mimic a more realistic imputation experiment where the study panel is of lower resolution than the WGS reference panel. Quality control (QC) cut-offs were applied to each population, MAF 0.01, geno (missingness/SNP) 0.1, mind (missingness per individual) 0.1, HWE 0.001 and applied to all scaffold SNPs. These filtered populations were then merged to produce one population (n = 475) in which the 490 overlapping SNPs were retained and used for

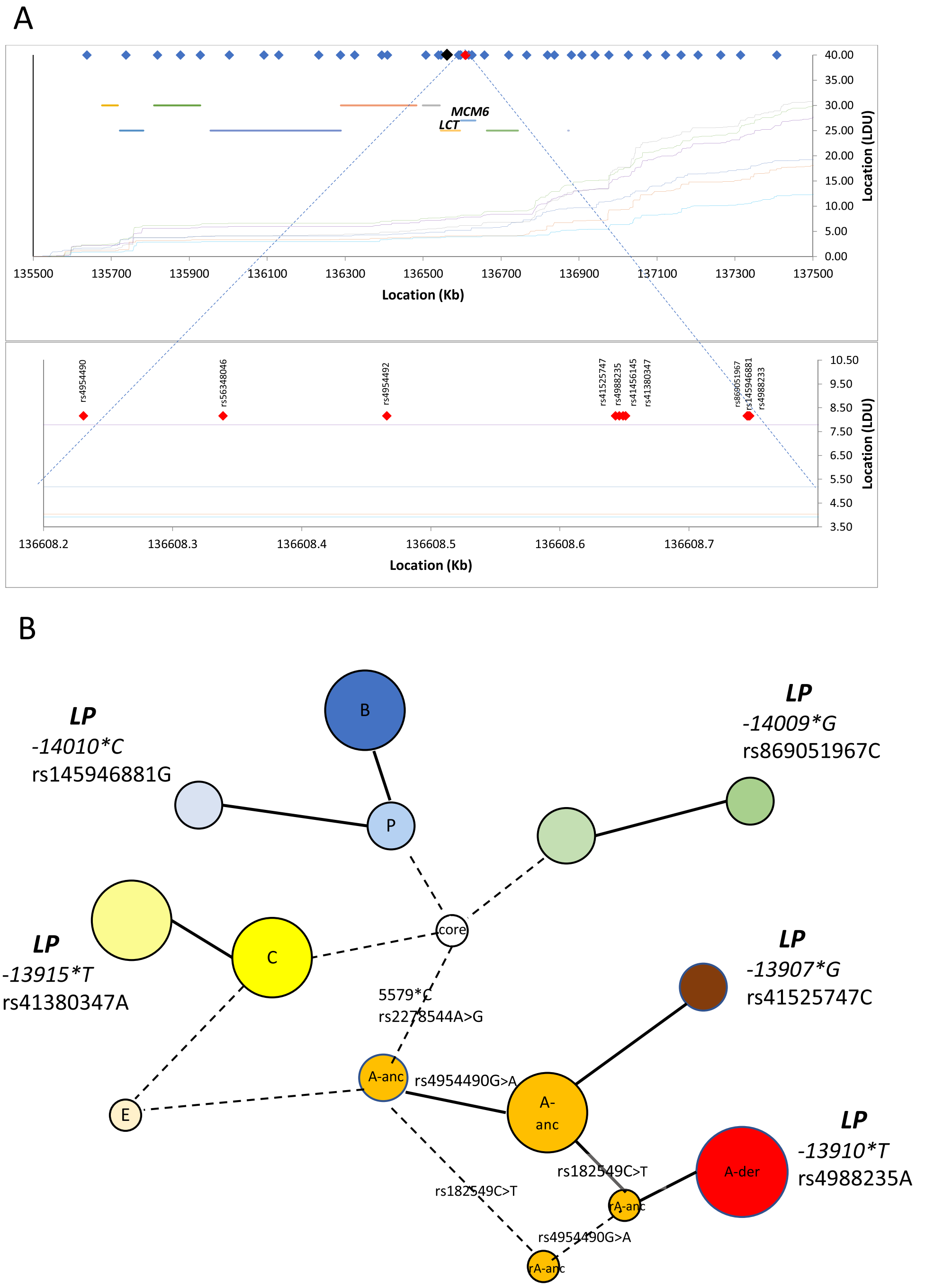

FIGURE 1 **(A) Diagram showing the positions of the genotyped SNPs and genes, in relation to the LDU map of the 1.77 Mb and 500 bp genomic intervals on chromosome 2 (build 37; hg19).** Physical location on chromosome 2 is shown on the horizontal axis and genetic location in linkage disequilibrium units (LDU) on the vertical axis. The maps were constructed using HapMap data for the six populations YRI (grey), LWK (green) ASW (purple), MKK (blue), TSI (orange), CEU (turquoise). *Upper plot*. 52 SNPs used in Population study 1 are represented by blue and red diamonds. Blue SNPs were used as a scaffold for imputation and red SNPs were masked. The larger black diamond shows rs182549 *Lower plot*. Shows the position of the 10 analysis (masked) SNPs with rs IDs, in the 500 bp interval (maps ASW, MKK, TSI, CEU) within an intron of MCM6. Sequence position from Build 37. Note that there is no evidence of recombination (no change in LDU location) in this small region in any of the groups. **(B) Core haplotype network showing the LP variants (with literature nomenclature) and haplotypes mentioned in this paper.** Continuous lines represent a single nucleotide change; dotted lines represent several changes and /or recombinations. Core haplotype names follow those used previously (Liebert et al., 2017) and colour code follows colours in Table S7

imputing rs4988235. *Note that the reference panel includes most of these same HapMap samples which should provide a better opportunity for the software to impute the missing data correctly*.

2.4 | Imputation

The scaffold SNPs for the Population studies 1 and 2 were phased with SHAPEIT2 (Delaneau et al., 2013) and all the untyped and masked SNPs in the full 1.77 Mb region were then imputed with IMPUTE2 (Howie et al., 2011) using the 1000 Genomes data as the reference panel (1000 Genomes Project Consortium et al., 2012, 2015; Delaneau, Marchini, and 1000 Human Genomes Project Consortium, 2014). Imputed genotypes were called with GTOOL (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html) using a calling threshold of 0.9 (so-called "hard" calls), and later for comparison using a less stringent threshold of 0.8.

2.5 | Haplotype Pairs

To obtain individual haplotype pairs for Population study 1, 52 SNPs (Table S1A,B) were also phased using PHASE (Stephens & Donnelly, 2003; Stephens et al., 2004) and assigned to the core haplotype groups of Hollox and colleagues (2001) as reported in Liebert et al. (2017) using the key markers listed in Table S1A. The three continental groups were combined as done previously (Liebert et al., 2017). The diplotypes can be found in the Supporting Information.

For Population study 2 the 40 SNPs spanning 653 kb (Chr2:136324225-136976936, GRCh37/hg19) in the HapMap dataset were also phased as one group, using PHASE, and haplotypes assigned to those of Hollox (Hollox et al., 2001), Table S1A, using informative overlapping SNPs (rs2304370 and rs3739022, as well as rs4988235). As an independent test of phasing, and for direct comparison with dataset 1, we extracted from the 1000 Genomes sequence data (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502) all the populations in which there were erroneous or missing imputations (CEU, ASW, GIH, MXL, TSI) as well as three others not included in the SNP dataset, GBR, FIN and IBS that segregate rs4988235 at good frequency. We used 44 of the same 52 SNPs as in Population study 1 (i.e., excluding the rare enhancer SNPs) from the full 1.77 Mb region (see Table S1A). The aim was to generate the longer phased 1.77 Mb haplotypes as done for our own data, using the same method, namely PHASE. Thirty-one additional SNPs in the *LCT/MCM6* region were added to improve accuracy (total of 75 SNPs, see Table S1A), while keeping computational time manageable.

3 | RESULTS

For Population Study 1 pairwise *D*' values across the core *LCT/MCM6* haplotype region show the previously reported similarity and very low level of recombination across this chromosomal region, not only across continental groups, but when subdivided by linguistic group (Fig. S2). Figure 1 likewise shows the similarity of the LD profile in different Hapmap population groups.

3.1 | Imputation of the Masked SNPs of Population Study 1

Initially, all 42 phased "scaffold" study SNPs were used for imputation. The first important observation was that two of the 10 SNPs that were to be imputed (rs41525747 and rs869051967), of which there were 66 and 32 derived allele occurrences, respectively, in the directly genotyped data, somewhat surprisingly were not present in the reference 1000G data. Therefore they could not be imputed within the test sample (Table S2), leaving eight SNPs that could be imputed.

For these eight SNPs there was a total of 98 wrongly imputed genotypes in the 802 samples. While the overall discordance rate is quite low (1%–3% per SNP; Table 1, Table S2), this is not a complete reflection of the true situation, as summarised in Table 1. In most cases, the errors were asymmetric, so that in the case of rs4988235 most were due to incorrect imputation of the presence of the derived allele where it is not present in the directly genotyped data (14/16), while other SNPs were in the other direction (i.e., failure to impute the derived allele), for example, rs56348046 (11 out of 15) and rs41380347 (six out of seven).

Most dramatically, the samples that evaded imputation (and returned as "missing") were also skewed with respect to genotype. In addition to the 2 SNPs that could not be imputed at all, a total of 218 further genotypes were not imputed, that is, 0%–10%, (Table 1), and were thus missing from the imputed dataset (at a calling threshold 0.9), and in most cases this was not evenly distributed between carriers and non-carriers of the derived alleles. This effect was particularly prominent for rs41380347 (–13915*T*>*G*), where 79 out of 80 samples that failed imputation were either heterozygous ($n = 53$), or homozygous ($n = 26$) for the derived allele. The summary of this is shown in Table 1

TABLE 1 **Incorrect allele calls and missing data in the first imputation study, Population study 1.** Data expressed as gains and losses of derived allele, and % error or missingness shaded in grey; total counts for each group in bold. * See Supporting Information for details of total counts for each locus; genome positions are shown. Details of the correct genome nomenclature are shown in Table S1A,B. For example −13910 C>T is rs4988235, or genomic reference NC_000002.11:g.136608646G>A or ClinVar ref. NM_005915.6:c.1917+326C>T (MCM6:c.1917+326C>T)

SNP literature name		**−13495 C>T**		**−13603 C>T**		**−13730 T>G**		**−13910 C>T**		**−13913 T>C**		**−13915 T>G**		**−14010 G>C**		**−14011 C>T**	
SNP rs ref		**rs4954490**		**rs56348046**		**rs4954492**		**rs4988235**		**rs41456145**		**rs41380347**		**rs145946881**		**rs4988233**	
Overall Minor Allele Frequency		**0.240**		**0.083**		**0.027**		**0.063**		**0.004**		**0.128**		**0.017**		**0.002**	
								Incorrect allele calls and percentages									
Population subdivision	**N**	**gain|loss**	**%**	**gain|loss**	**%**	**gain|loss**	**%**	**gain|loss**	**%**	**gain|loss**	**%**	**gain|loss**	**%**	**gain|loss**	**%**	**gain|loss**	**%**
South Europe	15	0|1	7	-	-	-	-	2|0	13	-	-	-	-	-	-	0|1	7
Northwest/Central Europe	38	1|0	3	-	-	-	-	2|0	5	-	-	-	-	-	-	-	-
East/Southeast Europe	39	6|0	15	1|0	3	-	-	6|0	15	-	-	-	-	-	-	-	-
Total Europe	**92**	**8|0**	9	**1|0**	1	-	-	**10|0**	11	-	-	-	-	-	-	**0|1**	1
Total Middle Eastern	**327**	**4|5**	3	**1|6**	2	**4|0**	1	**4|1**	2	**0|2**	1	**1|4**	2	**1|0**	0	**0|2**	1
North Africa	102	1|1	2	1|3	4	1|1	2	0|1	1	0|1	1	0|2	2	1|0	1	2|0	2
East Africa	241	2|3	2	1|1	1	3|2	2	-	-	0|3	1	-	-	4|1	2	5|0	2
West Africa	40	1|0	3	0|1	3	1|0	3	-	-	-	-	-	-	-	-	-	-
Total Africa	**383**	**4|4**	2	**2|5**	2	**5|3**	2	**0|1**	0	**0|4**	1	**0|2**	1	**5|1**	2	**7|0**	2
Total incorrect allele calls	**98**	**16|9**	3	**4|11**	2	**9|3**	1	**14|2**	2	**0|6**	1	**1|6**	1	**6|1**	1	**7|3**	1
Grand Total	**802***	**25**		**15**		**12**		**16**		**6**		**7**		**7**		**10**	
								Missing genotype counts and percentages									
Total missing genotypes	**218**	**27**	4	**26**	3	**15**	2	**2**	0.2	**0**	0	**80**	10	**17**	2	**51**	6
Missing; carrier|non-carrier		**13|14**		**7|19**		**15|0**		**2|0**		**0|0**		**79|1**		**6|11**		**1|50**	

TABLE 2 LCT core haplotype pairs of the 16 discordant individuals in Population study 1, showing sequenced and imputed genotypes for rs4988235, and PHASE derived haplotypes. More detail is shown in Table S5. The capital letters denote the Hollox core haplotype defined by the SNPs shown in Table S1. **A**-der = + derived allele rs4988235 +derived allele rs182549; r**A**-anc = ancestral allele rs4988235 +derived allele rs182549; **C**-der = + derived allele rs41380347

Person	Continent	Sequenced genotype	Imputed genotype	Haplotype 1	Haplotype 2
1	South Europe	GG	AG	r**A**-anc	**P**-like
2	South Europe	GG	AG	**B**	r**A**-anc
3	Northwest/Central Europe	AG	AA	r**A**-anc	**A**-der
4	Northwest/Central Europe	AG	AA	**A**-der	r**A**-anc
5	East/Southeast Europe	GG	AG	r**A**-anc	**D**
6	East/Southeast Europe	GG	AG	r**A**-anc	**C**
7	East/Southeast Europe	GG	AG	**B**	r**A**-anc
8	East/Southeast Europe	GG	AG	r**A**-anc	**A**-anc
9	East/Southeast Europe	GG	AG	r**A**-anc	**A**-anc
10	East/Southeast Europe	GG	AG	rare	r**A**-anc
11	Middle East	GG	AG	r**A**-anc	**P**
12	Middle East	GG	AG	**P**	r**A**-anc
13	Middle East	GG	AG	**C**	r**A**-anc
14	Middle East	GG	AG	**B**	r**A**-anc
15	Middle East	AG	GG	**C**-der	**A**-der
16	North Africa	AG	GG	**A**-der	**P**

and details in Table S2. It should be noted that this anomaly occurred notwithstanding having an "info score" of more than 0.9 in many cases (Table S3, see particularly rs41380347 for Middle Eastern and African samples). Table S3 also shows allele frequencies in the different continental groups compared with the 1000 Genomes samples, where the MAF for rs41380347 is 0.0006 in the reference panel.

In order to check the effect of a probability threshold for imputation, we lowered the stringency of the cut-off to 80% for calling imputed genotypes. The results were varied. For example, in the case of rs41380347, the missing calls for almost all derived allele carriers were replaced by correct calls. However, this was not the case for most other SNPs, where the number of incorrect calls increased. In the case of rs4988235 the results were identical. This is summarised in Table S4.

3.2 | Examination of haplotype background of rs4988235 (*−13910C>T*): Population Study 1

Haplotypes determined by PHASE showed that all 104 chromosomes carrying the derived allele for rs4988235 share the same core "**A**" haplotype (Hollox et al., 2001; Poulter et al., 2003), which is extended over more than 200 kb in most cases (Liebert et al., 2017). This haplotype also carries the derived allele of rs4954490, an SNP that is within the analysis region, and in 103/104 cases also the derived allele of rs182549 C>T (also known as *−22kb G>A* in the lactase literature), an SNP outside the analysis region (Fig. 1).

3.3 | rs4954490, rs182549 and LCT core haplotypes in relation to errors in imputation of rs4988235

The haplotype pairs determined by PHASE and assigned to the core haplotypes of Hollox (Hollox et al., 2001; Poulter et al., 2003) for each of the 16 discordant individuals are shown in Table 2.

The most noteworthy observation was that at rs4988235, all 10 of the European individuals and four Middle Eastern individuals in which alleles were wrongly imputed, carry the rare ancestral **A** haplotype with a derived T at rs182549, but *without* the derived allele at rs4988235, of which there were only 16 in the dataset (Fig. 1B). While most of these carry an A at rs4954490, five also lack the derived allele at this locus as well. These haplotypes are not frequent and are closely ancestral to the haplotype which carries rs4988235A, the European functional allele also known as *−13910*T* (Poulter et al., 2003) and for simplicity we call them both "r (rare) **A**-anc" in Fig. 1B and Table 2, while we call *−13910*T/* rs4988235A "**A**-der".

We therefore examined these two additional SNPs in detail, by removing (rs4954490) or adding them (rs182549) to the masked region, resulting in a 43 and 41 SNP scaffold, respectively. In neither case did this result in more accurate imputation; these results are shown in detail in Table S5. Minor changes were seen for the other loci, with little change to the asymmetry, apart from the case of rs4988233 which curiously was reversed (see Table S6). There were also similar levels of missingness (Table S6).

In just two of the discrepant individuals in the experiment with 42 scaffold SNPs (samples 15 and 16 in Table 2 and Table S5), the rs4988235 A allele (*–13910*T*) was missed, even though this allele was present on the usual **A** haplotype background. In both these cases, the second chromosome of these individuals was a less common haplotype. In one case this was a **C** haplotype carrying the derived *LP* allele at *–13915T>G* (rs41380347), which the software had also failed to impute, and this individual was also wrongly imputed when using the 43 SNP scaffold (Table S5). In the other case, the second chromosome carried the rare **P** haplotype, and this sample was correctly imputed for rs4988235 when using the 43 SNP scaffold (Table S5).

In summary for rs4988235, 14/16 of the wrong imputations carried directly ancestral **A** haplotype chromosomes, three of these also carrying uncommon haplotypes on the other chromosome. The remaining two of the 16 also carried uncommon haplotypes. The two missing imputations are less easy to understand. One person was homozygous for the derived allele haplotype and the other carried it in combination with the common **B** haplotype.

3.4 | Imputation using HapMap data (Population study 2)

Using the publicly available HapMap SNP data in which rs4988235 had been typed we were able to examine the accuracy of the imputation of this same chromosomal region with a more than tenfold higher scaffold of study SNP density (490) over the 1.77 Mb region (Fig. S1B). This resulted in seven out of 475 genotype discordances, at probability 0.9, (Table S7; see also Table S3 for Info scores), and 37 failures to impute rs4988235 at probability 0.9. Note that this was the case even though the HapMap SNP dataset is actually a subset of the 1000 Genomes. One of the seven discordances was for an individual (NA20810) who was heterozygous at rs182549, but homozygous for the ancestral rs4988235G (not GA as imputed), that is, the "difficult" genotype combination found in Population study 1. This was not the case for the other two rs4988235A gains. However, of the 37 with failure to impute at 0.9 probability, another sample (NA19676) carried this allelic combination (Table S7).

For the HapMap data, reducing the imputation confidence to 0.8, fourteen more calls were made for rs4988235 and eight of these were correct. However, six were erroneous. The details of this are also given in Table S7.

3.5 | Haplotype Distribution in Relation to Errors in Imputation in Population Dataset 2

We next used PHASE to obtain haplotypes for the 40 SNPs in the central 635 kb region using the HapMap SNP data, and as an independent test also the 75 SNP 1.77 Mb PHASE haplotypes from the 1000 Genomes sequence data (Fig. S1B). Using both sets of data we determined the *LCT* gene region core haplotypes (extending from the enhancer region to the end of the coding region of *LCT*) and found them in good agreement in both analyses, despite fewer key markers in the HapMap SNP dataset (Table 7B). As for Population study 1, all of the derived (A alleles) for rs4988235 were located on an extended **A** haplotype background carrying both the derived alleles of rs4954490 and rs182549.

The haplotypes in all the samples in which imputation was erroneous or failed are summarised in Table S7b. For the 13 erroneously imputed samples at 80% probability (in which there were fewer missing calls), the eight that gained an A allele each carried an almost identical extended ancestral **A** haplotype (haplotype h82 and h83 in Table 7B), of which there were only 72 in the entire dataset of 950 chromosomes; one was also heterozygous for the previously described recombinant **E** haplotype which results from intragenic recombination of **A** and **C** (29). Of the five that showed allelic "loss", two carried rather similar recombined **A** haplotypes; one was heterozygous for a derived **A** haplotype together with a rare **P** haplotype.

For the samples that failed to impute rs4988235 (missing results) at either genotype calling probability ($n = 21$) 10 individuals (11 chromosomes) carried ancestral **A** haplotypes, of which 10 were haplotypes h82 or h83, with the eleventh being the very similar haplotype h84, which occurs once in the dataset, and there were again two E haplotype chromosomes.

In summary, of the 13 wrongly imputed and 21 samples missing imputation at 80% probability, 19 (8 and 11, respectively) carried extended **A** haplotypes closely ancestral to **A**-der, and three of these carried the rare **E** recombinant haplotype on the other chromosome, while 3 had one rare recombinant chromosome in combination with a normal extended **A**-der chromosome (Table S7).

4 | DISCUSSION

In this study we show that despite the fact that the imputation algorithm has a high level of apparent accuracy (97%–99%) there are particular issues in imputing genotypes of SNPs that have undergone a recent dramatic change in allele frequency.

First, because of the heterogeneity of allele frequency spectra across population groups, there is a reasonable chance that the relevant SNP is not found at all in the reference panel used for imputation, as is the case here for two of the SNPs that are more frequent in East Africans, and cannot, therefore, be imputed using the 1000G; secondly and perhaps more seriously, incorrect imputation and failure to impute is not random in direction, nor across genotype. While in the case of rs41380347 this might partly be due to rarity in the reference panel, the same effect was found for rs4988235, in samples from Europeans, the Middle East and also in Africans (Population study 1, Table S2) and in the HapMap data, (Population study 2), CEU, TSI, ASW, MEX, ASW and GIH (Table S7A).

The results with rs4988235 using our data show that the algorithm is unable to properly distinguish the haplotype most recently ancestral to the derived haplotype (which we call r**A**-anc and carries the derived allele for rs182549 but not rs4988235). The HapMap (Population study 2) dataset included only two individuals carrying this r**A**-anc haplotype but the analysis that we have conducted shows that both these two chromosomes also cause difficulties in this dataset. In addition, other closely related extended **A**-anc haplotypes are major confounders, as perhaps unsurprisingly are a few rare recombinant haplotypes. The more surprising issue is that in a few cases of apparent homozygosity of the common extended derived **A** haplotype there was a failure to impute. Since there is apparently no recombination in at least two of these individuals over more than 600 kb (NA12043 and NA19777) one can only surmise that other recent derived alleles present in the sequence data are acting as confounders in the same way as rs182549. For the HapMap analysis it should be emphasised that the "experimental" data used for imputation comes from a group of individuals, most of whom are included within the 1000G reference panel.

This study focusses on one small genetic region, a region known to have been under-selection and uses just one imputation method under standard conditions. The results nevertheless indicate that great caution should be taken over the use of imputation. Clearly, this is a case study, and it will be interesting to consider the extent to which this is a problem in other parts of the genome with a different evolutionary history. It is currently not practicable to conduct this kind of study, that uses visual inspection of haplotype pairs, genome-wide. However other loci studied in our group, involving non-selected gene regions that are disease association signals, show similar examples of allelic imbalance (Maniatis and colleagues, unpublished). Although we only used one imputation method, it has been reported that the accuracy of IMPUTE2 is similar to, or slightly greater than other approaches (Shi et al., 2018). But since the principle behind the genotype imputation is the same across these, it seems likely that the result from other methods would be similar.

For association studies, the most efficient approach, in our view, is to use directly genotyped data with multi-marker methods for association that do not require the use of the same SNPs in different SNP-arrays for meta-analysis. In particular, the multi-marker method that utilises population-specific LDU maps, has the power to estimate causal locations without the need for imputation (Andrew et al., 2008; Direk et al. 2014; Elding et al. 2011, 2013; Lau et al., 2017). Our own work on Crohn's disease, for example, allowed us to detect association of novel loci using only directly genotyped SNP data (Elding et al., 2011). For example, CYLD was first found this way (Elding et al., 2013) and this has been followed up with strong functional evidence, in that Cyld knockout mice show severe colonic inflammation and damage to the intestinal lining (Mukherjee et al., 2020). For evolutionary genetics and ancient DNA where high coverage sequencing is not possible, the best approach may be targeted capture of specific genome regions (Cruz-Davalos et al., 2018) carrying population-specific derived alleles of particular interest to the study of recent adaptation that have been identified from modern DNA.

ACKNOWLEDGEMENTS

We thank the other co-authors of Liebert et al. (2017) for their contributions towards the collection of data reused in this paper. WL and NM acknowledge the Wellcome Trust, United Kingdom, for financial support 547 (Seed Award 209106/Z/17/Z); DS and AL funding by an EU Marie Curie ITN FP7 Framework Programme grant, LeCHE, grant ref 215362-2.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare

AUTHORS CONTRIBUTIONS

AA conducted all the Imputation analyses; AL provided SNP and sequence data for study 1; WL accessed and formatted the public data and prepared the LDU maps; AA, AL, DS and WL conducted the haplotype analyses: NM and DS conceived and designed the project; all authors contributed to writing the paper.

DATA ACCESSIBILITY STATEMENT

Genotype and sequencing data for Population Study 1 (frequencies) were previously published (Liebert et al., 2017). The diplotypes used for the current study are shown in the Supplementary data. The SNP data used for Population Study 2 are openly available at https://www.sanger.ac.uk/resources/downloads/human/hapmap3.htm

ORCID

Aminah T. Ali https://orcid.org/0000-0003-1089-9278
Anke Liebert https://orcid.org/0000-0002-5849-6147
Winston Lau https://orcid.org/0000-0001-5501-4619
Nikolas Maniatis https://orcid.org/0000-0002-6567-5936
Dallas M. Swallow https://orcid.org/0000-0001-7310-2735

REFERENCES

Allentoft, M. E., Sikora, M., Sjogren, K. G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P. B., Schroeder, H., Ahlström, T., Vinner, L., Malaspinas, A. S., Margaryan, A., Higham, T., Chivall, D., Lynnerup, N., Harvig, L., Baron, J., Della Casa, P., Dąbrowski, P., … Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, *522*, 167–172. https://doi.org/10.1038/nature14507

Andrew, T., Maniatis, N., Carbonaro, F., Liew, S. H., Lau, W., Spector, T. D., & Hammond, C. J. (2008). Identification and replication of three novel myopia common susceptibility gene loci on chromosome 3q26 using linkage and linkage disequilibrium mapping. *Plos Genetics*, *4*, e1000220. https://doi.org/10.1371/journal.pgen.1000220

1000 Human Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., … McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

1000 Human Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., & Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics*, *74*, 1111–1120. https://doi.org/10.1086/421051

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Coelho, M., Luiselli, D., Bertorelle, G., Lopes, A. I., Seixas, S., Destro-Bisol, G., & Rocha, J. (2005). Microsatellite variation and evolution of human lactase persistence. *Human Genetics*, *117*(4), 329–339. https://doi.org/10.1007/s00439-005-1322-z

Cruz-Dávalos, D. I., Nieves-Colón, M. A., Sockell, A., Poznik, G. D., Schroeder, H., Stone, A. C., Bustamante, C. D., Malaspinas, A.-S., & Ávila-Arcos, M. C. (2018). In-solution Y-chromosome capture-enrichment on ancient DNA libraries. *Bmc Genomics [Electronic Resource]*, *19*(1), 608. https://doi.org/10.1186/s12864-018-4945-x

Delaneau, O., Marchini, J. & 1000 Genomes Project Consortium. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature communications*, *5*, 3934. https://doi.org/10.1038/ncomms4934

Delaneau, O., Zagury, J. F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5–6. https://doi.org/10.1038/nmeth.2307

Direk, K., Lau, W., Small, K. S., Maniatis, N., & Andrew, T. (2014). ABCC5 transporter is a novel type 2 diabetes susceptibility gene in European and African American populations. *Annals of Human Genetics*, *78*(5), 333–344. https://doi.org/10.1111/ahg.12072

Elding, H., Lau, W., Swallow, D. M., & Maniatis, N. (2011). Dissecting the genetics of complex inheritance: linkage disequilibrium mapping provides insight into Crohn disease. *American Journal of Human Genetics*, *89*(6), 798–805. https://doi.org/10.1016/j.ajhg.2011.11.006

Elding, H., Lau, W., Swallow, D. M., & Maniatis, N. (2013). Refinement in localization and identification of gene regions associated with Crohn disease. *American Journal of Human Genetics*, *92*(1), 107–113. https://doi.org/10.1016/j.ajhg.2012.11.004

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*(6172), 747–751. https://doi.org/10.1126/science.1243518

Hollox, E. J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, A. I., & Swallow, D. M. (2001). Lactase haplotype diversity in the Old World. *American Journal of Human Genetics*, *68*(1), 160–172. https://doi.org/10.1086/316924

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959. https://doi.org/10.1038/ng.2354

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 (Bethesda)*, *1*(6), 457–470. https://doi.org/10.1534/g3.111.001198

Ilardo, M. A., Moltke, I., Korneliussen, T. S., Cheng, J., Stern, A. J., Racimo, F., de Barros Damgaard, P., Sikora, M., Seguin-Orlando, A., Rasmussen, S., van den Munckhof, I. C. L., Ter Horst, R., Joosten, L. A. B., Netea, M. G., Salingkat, S., Nielsen, R., & Willerslev, E. (2018). Physiological and Genetic Adaptations to Diving in Sea Nomads. *Cell*, *173*(3), 569–580 e515. https://doi.org/10.1016/j.cell.2018.03.054

Ingram, C. J. E., Mulcare, C. A., Itan, Y., Thomas, M. G., & Swallow, D. M. (2009). Lactose digestion and the evolutionary genetics of lactase persistence. *Human Genetics*, *124*(6), 579–591. https://doi.org/10.1007/s00439-008-0593-6

Itan, Y., Powell, A., Beaumont, M. A., Burger, J., & Thomas, M. G. (2009). The origins of lactase persistence in Europe. *Plos Computational Biology*, *5*(8), e1000491. https://doi.org/10.1371/journal.pcbi.1000491

Lau, W., Andrew, T., & Maniatis, N. (2017). High-Resolution Genetic Maps Identify Multiple Type 2 Diabetes Loci at Regulatory Hotspots in African Americans and Europeans. *American Journal of Human Genetics*, *100*(5), 803–816. https://doi.org/10.1016/j.ajhg.2017.04.007

Liebert, A., Jones, B. L., Danielsen, E. T., Olsen, A. K., Swallow, D. M., & Troelsen, J. T. (2016). In Vitro Functional Analyses of Infrequent Nucleotide Variants in the Lactase Enhancer Reveal Different Molecular Routes to Increased Lactase Promoter Activity and Lactase Persistence. *Annals of Human Genetics*, *80*(6), 307–318. https://doi.org/10.1111/ahg.12167

Liebert, A., López, S., Jones, B. L., Montalva, N., Gerbault, P., Lau, W., Thomas, M. G., Bradman, N., Maniatis, N., & Swallow, D. M. (2017). World-wide distributions of lactase persistence alleles and the complex effects of recombination and selection. *Human Genetics*, *136*(11-12), 1445–1453. https://doi.org/10.1007/s00439-017-1847-y

Maniatis, N., Collins, A., & Morton, N. E. (2007). Effects of single SNPs, haplotypes, and whole-genome LD maps on accuracy of association mapping. *Genetic Epidemiology*, *31*(3), 179–188. https://doi.org/10.1002/gepi.20199

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511. https://doi.org/10.1038/nrg2796

Martiniano, R., Cassidy, L. M., O'Maolduin, R., McLaughlin, R., Silva, N. M., Manco, L., Fidalgo, D., Pereira, T., Coelho, M. J., Serra, M., Burger, J., Parreira, R., Moran, E., Valera, A. C., Porfirio, E., Boaventura, R., Silva, A. M., & Bradley, D. G. (2017). The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *Plos Genetics*, *13*(7), e1006852. https://doi.org/10.1371/journal.pgen.1006852

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N., Mallick, S., Olalde, I., Broomandkhoshbacht, N., Candilio, F., Cheronet, O., Fernandes, D., Ferry, M., Gamarra, B., Fortes, G. G., Haak, W., Harney, E., Jones, E., Keating, D., Krause-Kyora, B., … Reich, D. (2018). The genomic history of southeastern Europe. *Nature*, *555*(7695), 197–203. https://doi.org/10.1038/nature25778

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E. R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J. L., De Castro, J. M. B., Carbonell, E., … Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, *528*(7583), 499–503. https://doi.org/10.1038/nature16152

Mukherjee, S., Kumar, R., Tsakem Lenou, E., Basrur, V., Kontoyiannis, D. L., Ioakeimidis, F., Mosialos, G., Theiss, A. L., Flavell, R. A., & Venuprasad, K. (2020). Deubiquitination of NLRP6 inflammasome by Cyld critically regulates intestinal inflammation. *Nature Immunology*, *21*(6), 626–635. https://doi.org/10.1038/s41590-020-0681-x

Poulter, M., Hollox, E., Harvey, C. B., Mulcare, C., Peuhkuri, K., Kajander, K., Sarner, M., Korpela, R., & Swallow, D. M. (2003). The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. *Annals of Human Genetics*, *67*(Pt 4), 298–311.

Ranciaro, A., Campbell, M. C., Hirbo, J. B., Ko, W.-Y., Froment, A., Anagnostou, P., Kotze, M. J., Ibrahim, M., Nyambo, T., Omar, S. A., & Tishkoff, S. A. (2014). Genetic origins of lactase persistence and the spread of pastoralism in Africa. *American Journal of Human Genetics*, *94*(4), 496–510. https://doi.org/10.1016/j.ajhg.2014.02.009

Segurel, L., & Bon, C. (2017). On the Evolution of Lactase Persistence in Humans. *Annual Review of Genomics and Human Genetics*, *18*, 297–319. https://doi.org/10.1146/annurev-genom-091416-035340

Shi, S., Yuan, N., Yang, M., Du, Z., Wang, J., Sheng, X., Wu, J., & Xiao, J. (2018). Comprehensive Assessment of Genotype Imputation Performance. *Human Heredity*, *83*(3), 107–116. https://doi.org/10.1159/000489758

Stephens, M., & Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, *73*(5), 1162–1169. https://doi.org/10.1086/379378

Stephens, M., Smith, N. J., & Donnelly, P. (2004). Documentation for PHASE, version 2.1.

Weng, Z. Q., Saatchi, M., Schnabel, R. D., Taylor, J. F., & Garrick, D. J. (2014). Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Genetics, Selection, Evolution.*, *46*, 34. https://doi.org/10.1186/1297-9686-46-34

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Ali AT, Liebert A, Lau W, et al. The hazards of genotype imputation in chromosomal regions under selection: A case study using the Lactase gene region. *Ann Hum Genet*. 2022;*86*:24–33. https://doi.org/10.1111/ahg.12444