nature genetics

Recurring exon deletions in the *HP* (haptoglobin) gene contribute to lower blood cholesterol levels

Linda M Boettger[1,2], Rany M Salem[1–4], Robert E Handsaker[1,2], Gina M Peloso[2,5], Sekar Kathiresan[2,5], Joel N Hirschhorn[1–4] & Steven A McCarroll[1,2]

One of the first protein polymorphisms identified in humans involves the abundant blood protein haptoglobin. Two exons of the *HP* gene (encoding haptoglobin) exhibit copy number variation that affects HP protein structure and multimerization. The evolutionary origins and medical relevance of this polymorphism have been uncertain. Here we show that this variation has likely arisen from many recurring deletions, more specifically, reversions of an ancient hominin-specific duplication of these exons. Although this polymorphism has been largely invisible to genome-wide genetic studies thus far, we describe a way to analyze it by imputation from SNP haplotypes and find among 22,288 individuals that these *HP* exonic deletions associate with reduced LDL and total cholesterol levels. We further show that these deletions, and a SNP that affects *HP* expression, appear to drive the strong association of cholesterol levels with SNPs near *HP*. Recurring exonic deletions in *HP* likely enhance human health by lowering cholesterol levels in the blood.

The HP protein binds free hemoglobin and facilitates its removal from the bloodstream[1,2]. A common 1.7-kb copy number variant (CNV) inside the *HP* gene determines the copy number (generally 1 or 2) of a tandem two-exon segment, including sequence that encodes a multimerization domain. This CNV is responsible for a striking protein phenotype: HP circulates as a dimer in individuals who are homozygous for the HP1 allele (encoding a single copy of the multimerization domain), but it forms multimers in individuals with the two-copy HP2 allele[3–5] (**Fig. 1**). HP2 is also a less efficient antioxidant than HP1 (ref. 6), and HP2 is required to make the tight-junction modulator protein zonulin, which is the preprocessed product of HP2 (ref. 7). Whether such functional variation contributes to human phenotypes is not well understood.

The alleles of *HP* are further divided into subtypes by nucleotide polymorphisms that cause HP to run faster or slower on a protein gel[8], hereafter called the 'F' and 'S' alleles, respectively. Both F and S alleles segregate on the HP1 background, creating the subtypes HP1F and HP1S. The most common form of HP2 contains both alleles (as paralogous sequence variants) and is called HP2FS, but a low-frequency HP2SS form also exists[9]. There are no known functional differences between the F and S alleles.

Despite the functional importance of haptoglobin—which constitutes one of the five most abundant proteins in blood[10]—and the potential functional importance of the common CNV that affects its structure, analyzing the association of this CNV with human phenotypes has proven challenging, and the CNV's relationship to genome-wide association study (GWAS) signals near *HP* has been unclear[11]. The CNV is not in strong linkage disequilibrium (LD) with any individual SNP[11], and it has not been successfully genotyped with array-based copy number analysis[12] or low-coverage sequencing[13]. Instead, the polymorphism is generally typed with protein PAGE[14], PCR[15] or quantitative PCR[16], which has in practice restricted the size of most association studies. Although the *HP* polymorphism—one of the earliest polymorphisms to be discovered in humans—has been analyzed in hundreds of studies for associations with many human phenotypes, the limited sample sizes of these studies have provided insufficient power to determine whether the common *HP* CNV, or other nearby genetic variation, contributes to genetically complex phenotypes.

Blood cholesterol levels are one of the most important known biomarkers for future health and mortality[17]. A GWAS for cholesterol levels found a definitive signal ($P = 3 \times 10^{-24}$) at markers near *HP*[18], but, as at most GWAS-implicated loci, the causal variant(s) explaining this association are not known. Although the HP protein's most familiar role is to bind hemoglobin, HP also binds cholesterol molecules[19–22]. We hypothesized that the *HP* CNV might be responsible for the genetic association of cholesterol levels with this locus. To investigate this relationship, we had to develop ways to understand a surprisingly complex form of structural variation and its relationships to SNPs and haplotypes.

RESULTS

A revised structural history of the haptoglobin gene

The alleles and mutational history of a locus provide a context for understanding whether and how the locus generates phenotypic variation. Standard genomics approaches, such as LD-based and array-based

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Division of Endocrinology, Boston Children's Hospital, Boston, Massachusetts, USA. [4]Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts, USA. [5]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. Correspondence should be addressed to S.A.M. (mccarroll@genetics.med.harvard.edu).

Received 11 August 2015; accepted 20 January 2016; published online 22 February 2016; doi:10.1038/ng.3510

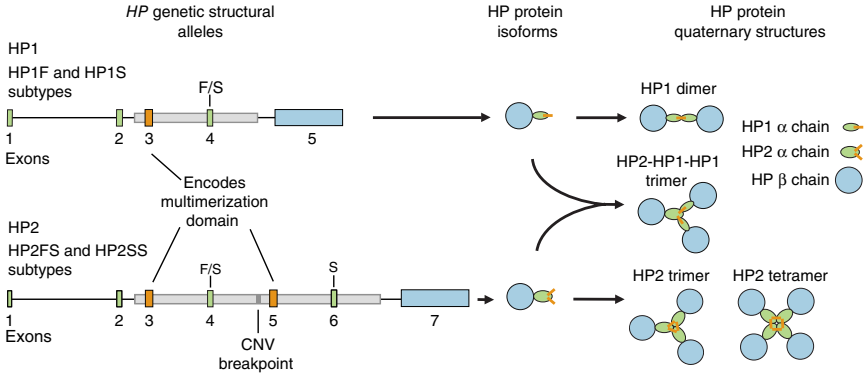

Figure 1 A common CNV in the *HP* gene is responsible for distinct molecular phenotypes. The HP2 allele contains two additional exons as compared to the HP1 allele: exons 3 and 4 are analogous to exons 5 and 6, respectively. The boundaries of the CNV are shown by the gray boxes on the gene diagrams. The HP1 allele contains one copy of sequence in exon 3 (orange), which encodes the protein multimerization domain, allowing dimers to be formed. HP2 has two copies of this multimerization domain, which results in the formation of multimers. Exons 4 and 6 (green) contain the F/S mutations responsible for the protein running faster or slower on a gel. The long final exon of HP1 and HP2 encodes the β subunit of the protein (blue), whereas the earlier exons encode the α subunit (green and orange). The α and β subunits are cleaved apart by proteolytic processing after translation but are held together by disulfide bonds[5]. The protein isoform diagrams shown here were modeled after those in an earlier publication[5].

CNV analyses, have not yet successfully captured structural variation in *HP*[11]; we sought to determine why standard methods have failed and to develop a new approach.

The long-accepted model of *HP* structural evolution[23] proposed that HP2 arose through non-homologous recombination between HP1F and HP1S to produce HP2FS. The assumption that HP2 was formed by the fusion of human HP1 alleles arose from the observation that non-human great apes lack HP2 (ref. 24) and that the left and right copies of the sequence in HP2FS are similar to sequences in HP1F and HP1S, respectively[23]. However, the low LD between the *HP* CNV and surrounding SNPs potentially suggests a more complex structural history, as has been noted previously[25]. We first sought to distinguish between the two forces that reduce LD between nearby loci: (i) recombination and (ii) recurrent mutation. If the low LD (of the CNV with flanking SNPs) were caused by frequent homologous recombination near the *HP* CNV region, then SNPs on the left and right sides of the structural variation would have low LD to one another. Conversely, if *HP* structure were affected by recurring intrachromosomal structural mutations (or by non-allelic recombination between identical sister chromatids), then low LD between SNPs and the CNV might still be accompanied by high LD between SNPs on either side of the CNV.

We used droplet digital PCR (ddPCR)[26] to genotype the *HP* CNV in 264 unrelated individuals sampled by the 1000 Genomes Project[13], phased the structural alleles onto SNP haplotypes using low-coverage sequence data[27] and clustered similar SNP haplotypes (Online Methods). We observed that, although many pairs of SNPs on opposite sides of the CNV were in high LD with each other ($r^2 > 0.95$) (**Supplementary Fig. 1**), copy number of the *HP* exons was not strongly correlated with any SNP on either side (maximum $r^2 = 0.44$ in Europeans from the 1000 Genomes Project). Three common SNP haplotypes (denoted A, B and C in **Fig. 2**) persisted through the CNV region yet segregated with both the HP1 and HP2 forms, a pattern that appears consistent with recurring structural mutations at *HP* (**Fig. 2**).

We next sought to determine whether structural mutations at *HP* involved deletions or duplications by analyzing the nucleotide variation in the CNV region. We classified 27 haplotypes as one of four conventional subtypes—HP1S, HP1F, HP2FS or HP2SS—on the basis of known sequence differences[23]. For HP2 haplotypes, we refer to the left copy of the CNV as HP2-left (which is proximal to the centromere

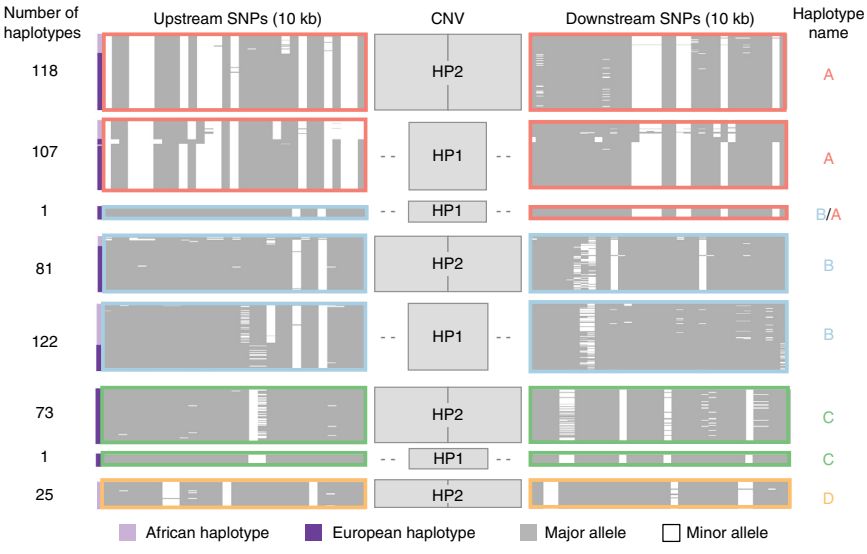

Figure 2 SNP haplotypes surrounding *HP* persist through the CNV region yet segregate with both structural forms of *HP*. This plot displays the SNP haplotypes (10 kb on each side of the *HP* CNV) segregating with HP1 and HP2 based on an analysis of 264 samples (528 haplotypes). The upstream SNPs are proximal to the centromere, whereas the downstream SNPs are distal to the centromere. Each thin horizontal line represents an individual SNP haplotype; similar or identical haplotypes are organized into clusters outlined by colored boxes. Note that the size of small clusters has been increased for visibility purposes, and the number of haplotypes contained in each cluster is indicated at the left of the plot. White represents the minor allele and gray indicates the major allele across all populations in the analysis (CEU, IBS, TSI, YRI). Haplotypes ascertained from West African (HapMap YRI) individuals are indicated with lavender bars to the left of the plot, while haplotypes ascertained in European populations (CEU, IBS, TSI) are indicated with dark purple bars to the left of the plot. Haplotypes were clustered with the *k*-means method using upstream SNP haplotypes. Similar SNP haplotypes carrying different structures are indicated with colored outlines (dark pink, light blue, green, gold) and are designated haplotypes A–D. This figure was based on analysis of 1000 Genomes Project samples and data (Online Methods).

and 5′ on the transcribed RNA) and to the right copy as HP2-right (distal to the centromere and 3′ on the transcribed RNA) (**Fig. 3a**). Some 42 nucleotide polymorphisms differed among the subtypes of *HP* (for example, between HP2FS and HP1S) but were consistent for any given subtype (**Fig. 3a** and **Supplementary Fig. 2**). To identify ancestral and derived alleles, we compared the human variants of each

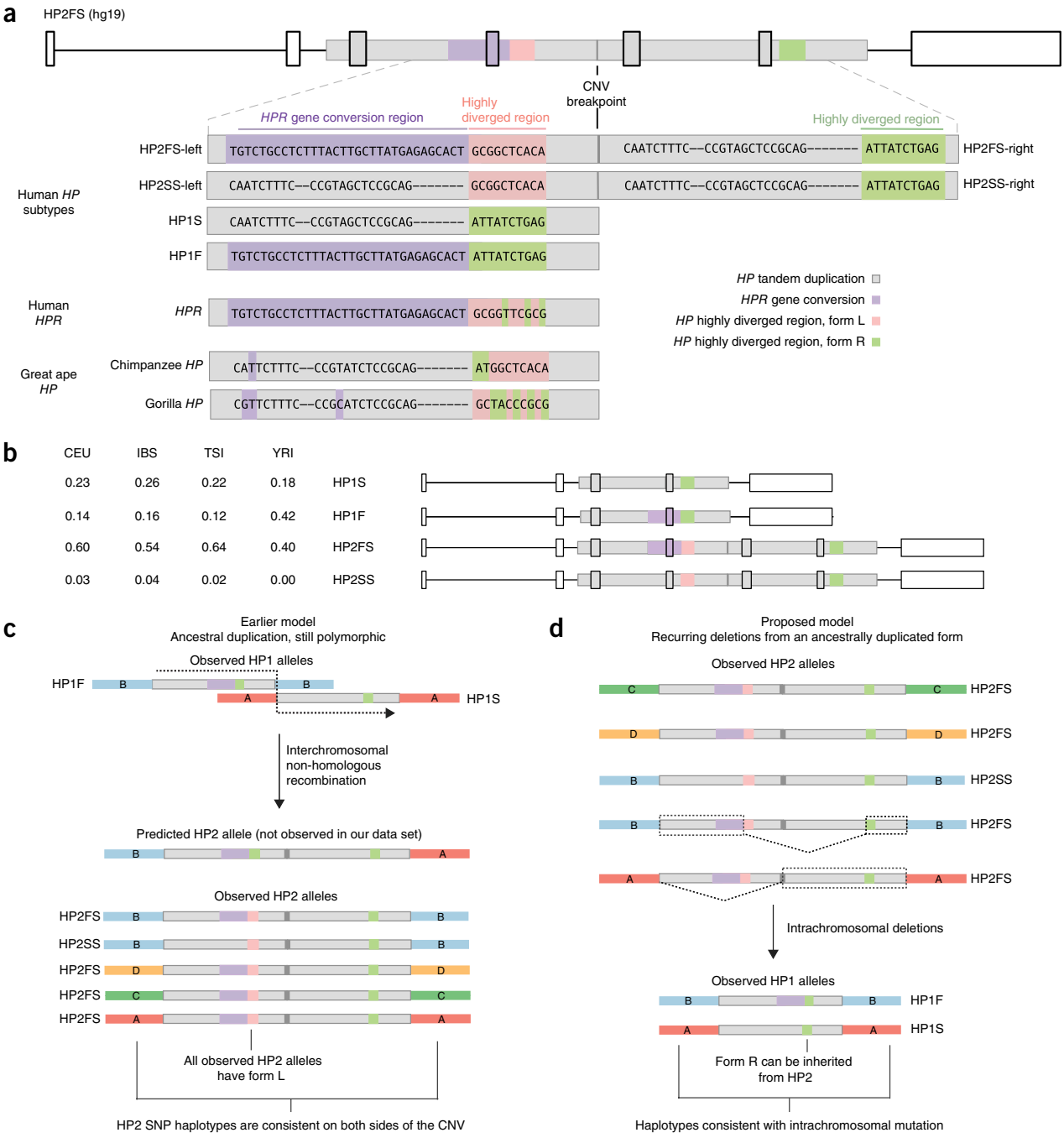

Figure 3 SNP haplotypes and sequence differences between *HP* subtypes inform structural history. (**a**) This alignment shows base-pair differences between the *HP* structural forms analyzed from 27 haplotypes. Only polymorphic bases are depicted. The HP2FS haplotype contains a 300-bp segment with a derived paralogous gene conversion from *HPR* (lavender) and a 250-bp region that is highly diverged between subtypes (green/pink). Each form of the highly diverged region contains a mixture of ancestral and derived alleles. The dashes reflect a 2-bp and a 7-bp indel; the other sites shown are individual SNPs. The sequence data used to create this alignment are available online (GenBank, KT923758–KT923784). (**b**) The frequency of each *HP* haplotype in four populations. (**c**) The earlier model of *HP* structural evolution (interchromosomal non-homologous recombination) would predict the HP1F SNP haplotype background (haplotype B) upstream of HP2 and the HP1S SNP haplotype background (haplotype A) downstream of HP2. Additionally, it would predict form R of the highly diverged region in HP2-left. However, neither of these predictions was observed in any of the HP2 alleles in this study. (**d**) Both HP1F and HP1S can be created through simple deletions in HP2FS. The dashed lines indicate deleted sequence, and the dashed boxes indicate the sequence required to create each HP1 haplotype. The deletion model is also consistent with the observed SNP haplotype backgrounds surrounding the CNV.

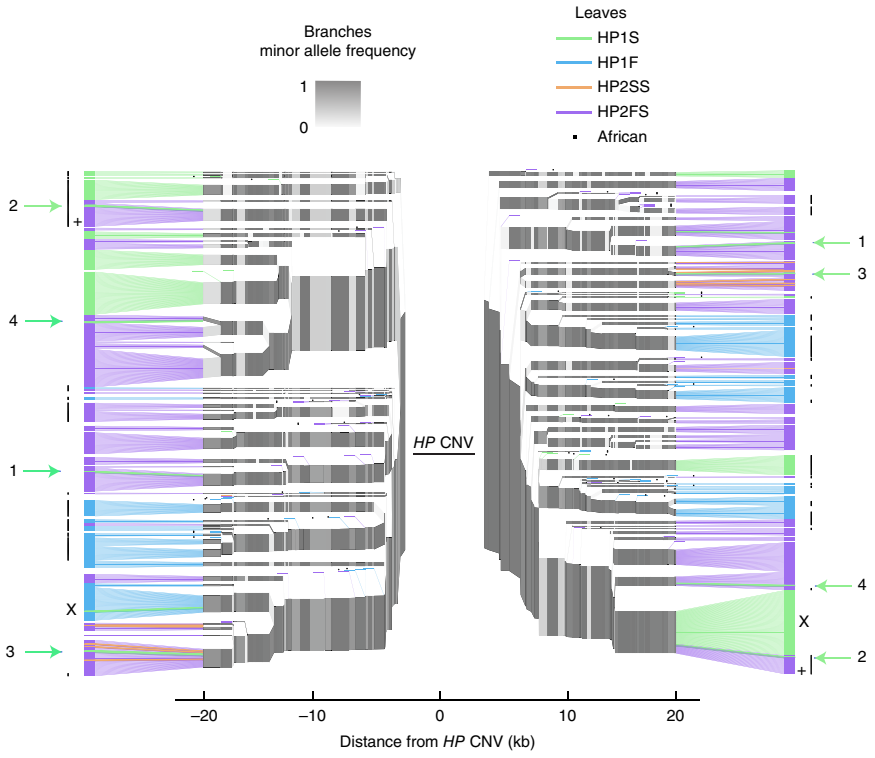

Figure 4 Lone HP1S structural alleles segregate on common HP2FS SNP haplotypes. SNP haplotype data are shown for three European populations (CEU, IBS, TSI) and one African population (YRI) totaling 528 haplotypes. SNPs on the left half of the plot exist upstream of the *HP* duplication (proximal to the centromere), whereas SNPs on the right half of the plot physically reside downstream of the duplication (distal to the centromere). Branch points represent markers at which the depicted haplotypes diverge owing to mutation and/or recombination with other haplotypes. The structures are represented on the leaves to clarify their relationships to SNP haplotypes, but the CNV and the paralogous gene conversion physically reside within the gap at the center of the plot. The African individuals are identified with a dot after the leaf. Arrows with numbers indicate HP1 alleles segregating with the standard HP2 SNP haplotypes for at least 20 kb on both sides of the CNV. A plus sign identifies the SNP haplotype branch that carries HP2FS in almost all sampled Africans but carries HP1S in all sampled Europeans. This SNP haplotype is identical downstream of the CNV and differs by a single nucleotide upstream of the CNV. The X indicates the single haplotype observed in this study with apparent recombination in the CNV region (B/A in **Fig. 2**). This recombination event appears to be recent because it is identical to standard haplotypes for at least 20 kb on either side of the CNV.

polymorphism to great ape versions of the *HP* gene, great ape paralogs of *HP* and the human *HPR* gene (encoding haptoglobin-related protein), which lies 2.2 kb downstream of *HP* and has 90% sequence identity with it (**Fig. 3a**, Online Methods and **Supplementary Fig. 3**).

This analysis showed that HP1F and HP2FS-left (the left copy of the CNV segment on HP2FS) share a 300-bp segment containing 30 derived variants that is nearly identical to a portion of the human *HPR* gene. This segment is likely the result of paralogous gene conversion, through which a segment of *HPR* sequence was transferred into the *HP* gene (**Fig. 3a** and **Supplementary Figs. 2** and **3**). This gene conversion is responsible for the F mutations in HP1F and HP2FS. We believe that this gene conversion event has complicated detection of the CNV in genomic studies, as the copy number–variable sequence can appear to arise partly from *HP* and partly from *HPR*, and likely explains why CNV data resources[12,13] have lacked genotypes for this CNV. Our analysis also identified a highly diverged 250-bp region that has ten fixed differences (between subtypes), including a mixture of derived and ancestral alleles in each segment (**Fig. 3a** and **Supplementary Figs. 2** and **3**). We refer to this sequence as the 'highly diverged region' and call the allele present in HP1S, HP1F and HP2-right 'form R' and the allele present in HP2-left 'form L'. We confirmed that these sequence differences are consistent at the population level by genotyping the boundaries of each variable region using ddPCR in DNA from 590 individuals sampled by HapMap[28] and the 1000 Genomes Project[27] (**Fig. 3b**, Online Methods and **Supplementary Fig. 4**).

The sequence differences between the *HP* subtypes shown in **Figure 3** indicate that neither modern HP2 subtype (HP2FS nor HP2SS) could be created through the fusion of known HP1F and HP1S subtypes in the way that the earlier model[23] proposed (**Fig. 3c**); for the earlier model to be true, HP2 would need to have arisen from a fusion of HP1S with a hypothetical diverged HP1 allele (containing form L of the highly diverged region) that no longer segregates at an appreciable frequency in human populations. Alternatively, we propose that HP2 could be much older than previously thought, allowing these (non-allelic) sequences the time to diverge strongly from each other as paralogous sequence variants on an HP2 structure. HP2 does have all the sequences required to form HP1 alleles by simple non-allelic homologous recombination (NAHR) between the two tandem copies of the two-exon segment on HP2FS (**Fig. 3d**).

Flanking SNP haplotypes also suggest that HP2 did not arise from recombination between HP1F and HP1S. All HP1F alleles segregate with SNP haplotype B, and almost all HP1S alleles segregate with SNP haplotype A (**Supplementary Fig. 5**). If HP2 had been created via non-allelic recombination between these two alleles, SNP haplotype B would be upstream of HP2 and SNP haplotype A would be downstream (**Fig. 3c**); however, characteristic HP2 SNP haplotypes persist across the CNV region and do not appear to involve such recombinant haplotypes (**Fig. 3d**; see **Supplementary Fig. 6** for our complete model of *HP* structural evolution).

An alternative model would be that HP2 is in fact the ancestral allele in humans and that HP1 alleles arose (and may continue to arise) by simple exonic deletions (due to NAHR) on an HP2 background. HP2-to-HP1 deletions have been observed at low frequency in the somatic and sperm cells of homozygous-HP2 individuals[29], demonstrating that the *HP* gene is prone to this type of structural mutation.

We sought to use information from long SNP haplotypes to further evaluate the alternative hypothesis that HP2-to-HP1 deletions gave rise to the structural variation at *HP*. If HP2-to-HP1 deletions occur intrachromosomally (or between sister chromatids) and are transmitted to offspring, then rare HP1 subtypes might segregate on SNP haplotypes that are usually associated with common HP2 subtypes. Whereas the short (10-kb) haplotypes immediately around *HP* cluster into a small number of groups (**Fig. 2**), we found that longer SNP haplotypes have much more information and cluster into a larger number of smaller groups (**Fig. 4**). A dendrogram analysis of these longer haplotypes showed that several common HP2FS-flanking SNP haplotypes also contain rare (singleton) HP1S alleles (**Fig. 4** and Online Methods).

Table 1 Imputation of *HP* structural features from surrounding SNPs

	Europeans (CEU, TSI, IBS)	
Subtype	Imputation (r^2)	r^2 (tag SNP)
HP1S	0.94	0.86 (rs217181)
HP1F	0.98	0.83 (rs9302635)
HP2FS	0.94	0.40 (rs217181)
HP2SS	0.75	0.58 (rs34914030)
HP1 versus HP2	0.94	0.44 (rs217181)

This table shows the correlation (r^2) between *HP* structural alleles (as identified by direct molecular analysis) and predictions from imputation from SNP haplotypes, using SNPs on the Illumina Omni2.5 array. The correlation between each structural feature and the most strongly correlated individual SNP is also displayed. The CEU, IBS and TSI populations were merged into a single European population for this analysis.

Four of these rare HP1S structures segregate with SNP haplotypes that are identical to common HP2FS SNP haplotypes for at least 20 kb on either side of the CNV. The HP2FS and HP1S alleles from the same SNP haplotype branch also share derived mutations within the CNV region, consistent with shared ancestry (**Supplementary Fig. 7**). These observations indicate that these four HP1S alleles likely result from recent exonic deletions that occurred on an earlier HP2 allele. We also identified a SNP haplotype that carried the HP2FS allele in 15 of 16 sampled Africans but had the HP1S allele in 16 of 16 sampled Europeans, consistent with a deletion event in an African ancestor whose descendants migrated to Europe (**Fig. 4**). We conclude that *HP* structural variation reflects a combination of ancient and recent deletions that continue to create HP1 alleles from HP2 alleles.

For common HP1 alleles to be derived deletions, HP2 would have to be ancient. The HP2 allele has not been observed in non-human primates, prompting the earlier model[23] that it was a derived, recent[30] allele. However, high-coverage genome sequences from ancient hominins are now available. We found that the *Homo neanderthalensis*[31] and *Homo denisova*[32] genomes both have many sequence reads containing the breakpoint sequence that is present on HP2 but not on HP1 and that they also contain all other sequences that define the HP2FS subtype (**Supplementary Table 1**). The presence of HP2FS in Neanderthals, Denisovans, and both modern and ancient[33] African humans (**Fig. 3b** and **Supplementary Table 1**) indicates that HP2 arose before the divergence of these hominins 400 to 600 thousand years ago[34]. (An earlier study, which assumed that HP2 was the derived allele, estimated the age of HP2 at less than 100 thousand years[30], but this is contradicted by the ancient hominin genome sequences[31,32].) SNP haplotypes further support the idea that HP2 is an ancient structural form: unlike HP1, HP2 segregates on all four common human SNP haplotypes identified at this locus (A–D in **Fig. 2**).

HP structural alleles can be imputed from SNP haplotypes

It is important to understand how complex, recurring variation contributes to human phenotypes. The gene conversion history and limited LD between the *HP* CNV and surrounding SNPs have made it challenging to study this structural variation. We sought to develop a way to integrate *HP* structural variation into large-scale genetic studies whose large sample sizes enable robust analysis of relationships with phenotypes. We hypothesized that, although *HP* structural mutations have occurred many times among human ancestors, the subset of these mutations that are old and common today might segregate on characteristic SNP haplotypes in many different individuals. Indeed, the above analysis of highly specific SNP haplotypes showed that such haplotypes usually segregate with a characteristic *HP* structural allele (**Fig. 4**).

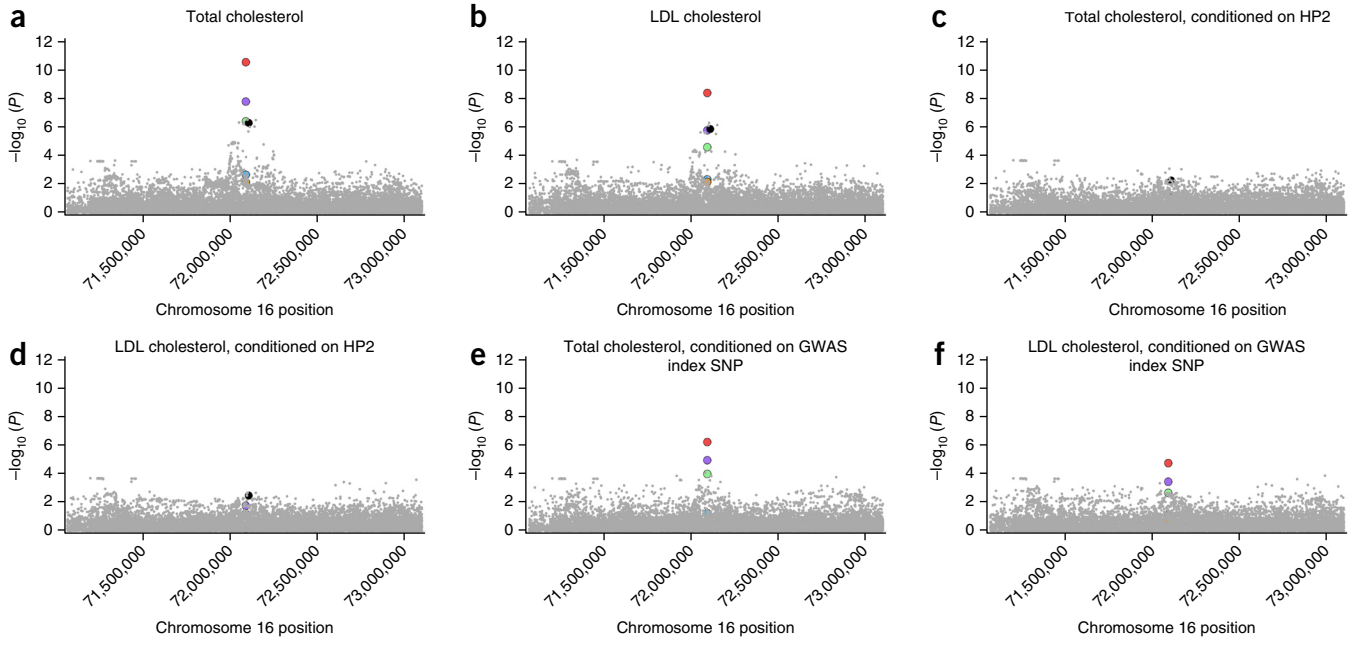

Figure 5 The HP2 allele associates with increased total cholesterol levels and increased LDL cholesterol levels. The imputed structural variants and all regional SNPs imputed from 1000 Genomes Project data are shown for this analysis of 22,288 individuals. (**a**,**b**) The HP2 variant is the strongest regional association for both total cholesterol levels ($P = 2.79 \times 10^{-11}$) (**a**) and LDL cholesterol levels ($P = 4.3 \times 10^{-9}$) (**b**). (**c**,**d**) Conditioning on the HP2 variant causes most of the association of the GWAS index SNP with total cholesterol levels ($P = 0.006$) (**c**) and LDL cholesterol levels ($P = 0.004$) (**d**) to disappear. (**e**,**f**) Conditioning on the GWAS index SNP only has a moderate effect on the association of HP2 with total cholesterol levels ($P = 5.95 \times 10^{-7}$) (**e**) and LDL cholesterol levels ($P = 2.02 \times 10^{-5}$) (**f**).

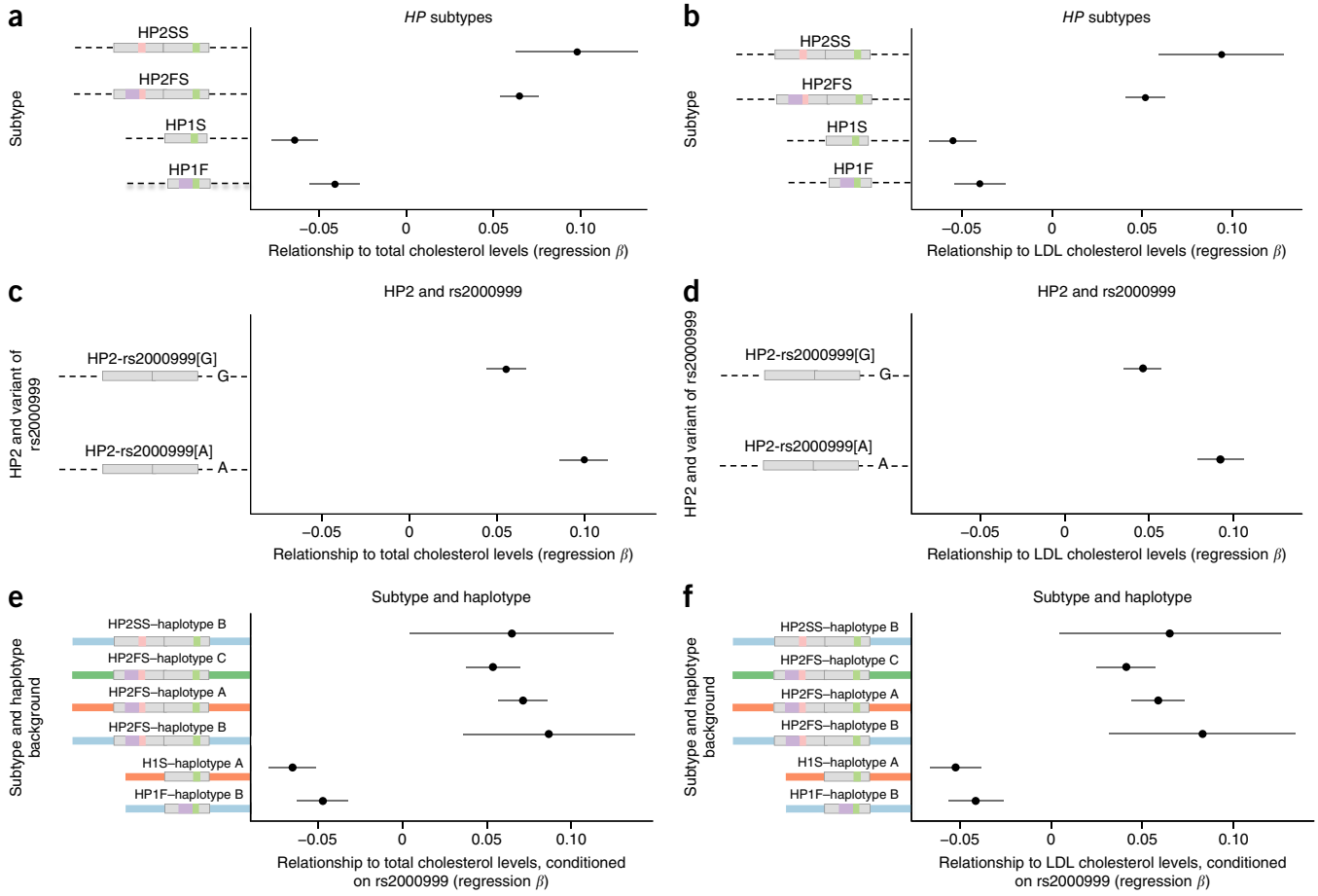

Figure 6 The rs2000999[A] allele on the HP2 background is associated with a greater increase in total cholesterol and LDL cholesterol levels than the rs2000999[G] allele. (**a**,**b**) The regression β values of the HP1 and HP2 alleles with total cholesterol (**a**) and LDL cholesterol (**b**) levels are shown with the standard error for this analysis of 22,288 individuals. (**c**,**d**) The regression β value of each allele of rs2000999 with total cholesterol (**c**) and LDL cholesterol (**d**) levels is shown with the standard error. (**e**,**f**) The regression β value of each *HP* subtype and total cholesterol (**e**) and LDL cholesterol (**f**) levels separated by SNP haplotype background is plotted with the standard error. The β value for each HP1 allele was calculated by a comparison with HP2 alleles only, and the β value of HP2 alleles was calculated through a comparison with HP1 alleles only (Online Methods and **Supplementary Note**).

To test this hypothesis, we phased *HP* structural alleles with SNP haplotypes to create reference chromosomes for imputation[35–37] (**Supplementary Data Set**). To measure the efficacy of imputation (using Beagle[35]), we implemented a series of leave-one-out trials, in each of which we removed an individual's *HP* gene structure from the reference panel and attempted to infer what structure was present on the basis of the surrounding SNP haplotype and the rest of the reference panel (Online Methods and **Supplementary Note**). Although no individual SNP 'tagged' *HP* CNV status (HP2 versus HP1) with high accuracy (maximum $r^2 = 0.44$), we were able to impute *HP* CNV status from multi-SNP haplotypes in both African and European population samples with high accuracy ($r^2 = 0.94$ in a European (CEU, IBS, TSI) population sample and $r^2 = 0.92$ in a Yoruba (YRI) sample), using only SNPs present on common SNP genotyping arrays (**Table 1**, **Supplementary Fig. 8** and **Supplementary Tables 2–5**). We believe this result reflects the fact that, despite recurring mutation at *HP*, most HP1 alleles trace back to a few ancient mutations in common ancestors (more recent deletion events likely reduce the efficacy of imputation but are more rare) (**Table 1**). Our imputation approach allows *HP* structural variation to be incorporated into large genetic studies using existing SNP data.

Haptoglobin and blood cholesterol levels

Both total cholesterol and low-density lipoprotein (LDL) cholesterol levels associate strongly ($P = 3 \times 10^{-24}$ and 2×10^{-22}, respectively, in a cohort of >100,000 individuals[18]) with the SNP rs2000999, which is within 15 kb of *HP*. Given that the HP protein binds to multiple types of cholesterol molecules[19–22] and that the HP1/HP2 difference has at least a modest correlation with this SNP ($r^2 = 0.14$), we hypothesized that the recurring structural variation that causes the HP1/HP2 difference could be responsible for the association of cholesterol levels with variation in this region.

We were able to obtain genome-wide SNP data from 22,288 individuals of European descent with cholesterol measurements (Online Methods, **Supplementary Table 6** and **Supplementary Note**). In this sample, we found that the GWAS index SNP (rs2000999) was associated as expected with total cholesterol levels ($P = 5.15 \times 10^{-8}$) and LDL cholesterol levels ($P = 1.43 \times 10^{-7}$) (**Fig. 5a**,**b**). We used our approach to impute the most likely *HP* subtypes in each individual's genome. The imputed HP2 state was associated with cholesterol phenotypes much more strongly ($P = 2.8 \times 10^{-11}$ for total cholesterol levels and $P = 4.3 \times 10^{-9}$ for LDL cholesterol levels) than any SNP in the *HP* region did (**Fig. 5a**,**b**). Furthermore, in analyses controlling for the HP1/HP2

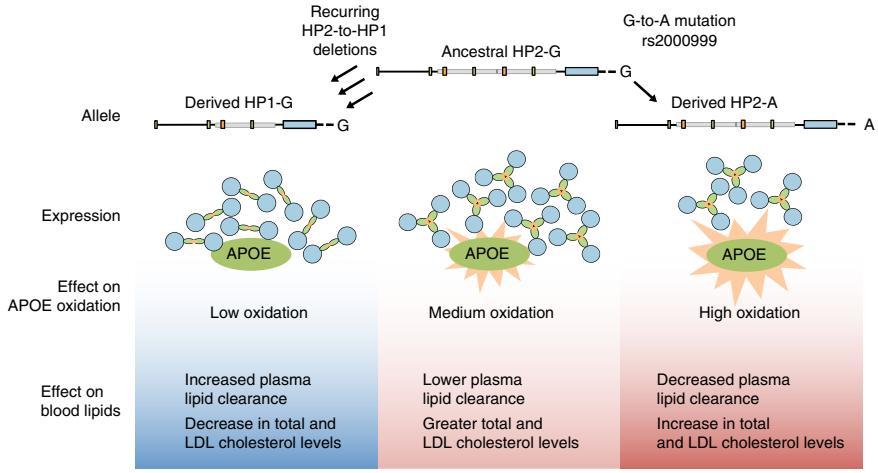

Figure 7 A model for the influence of *HP* genetic polymorphisms on total and LDL cholesterol levels. Because HP serves as an antioxidant for bound APOE[20,22] and HP1 has greater antioxidant activity than HP2 (ref. 6), we propose that HP1 alleles (arising from HP2-to-HP1 deletions) lessen the oxidative burden on APOE, allowing it to more effectively clear plasma lipids. Conversely, the rs2000999[A] allele decreases *HP* expression[40,41] and thus reduces antioxidant protection for APOE, contributing to elevated cholesterol levels.

difference, the association at the index SNP was reduced to $P = 0.006$ for total cholesterol levels and $P = 0.004$ for LDL cholesterol levels (**Fig. 5c**,**d**) (notably, these are still nominally positive associations, which we further explore below), whereas the HP1/HP2 variant continued to associate more strongly with cholesterol levels ($P = 5.95 \times 10^{-7}$ for total cholesterol levels and $P = 2.02 \times 10^{-5}$ for LDL cholesterol levels) in analyses controlling for the GWAS index SNP (**Fig. 5e**,**f**).

Both HP2 subtypes (HP2FS and HP2SS) were associated with increased cholesterol levels (**Fig. 6a**,**b**). Although the HP1F and HP1S subtypes segregate on very different SNP haplotype backgrounds (**Supplementary Fig. 5**), they were associated with similar levels of protection from elevated cholesterol levels (**Fig. 6a**,**b**), further supporting the idea that *HP* structural variation (rather than nearby sequence variation) is the primary driver of the association with these structural alleles.

The GWAS index SNP, rs2000999, is located in a strong enhancer sequence for hepatocytes[38,39] (the primary source of HP), and the derived allele of this variant associates with reduced *HP* expression[40,41]. Although the above analysis more strongly implicated the *HP* structural variation than this SNP in cholesterol levels, we hypothesized that both the CNV and the rs2000999 variant might affect cholesterol levels through their respective effects on haptoglobin structure and abundance. The derived rs2000999[A] allele is present almost exclusively on HP2 haplotypes ($D' = 0.96$), so we examined the effect of each rs2000999 allele on the HP2 background. (The LD between rs2000999 and the *HP* subtypes is shown in **Supplementary Table 7**). We found that, whereas all HP2 alleles were associated with an increase in total and LDL cholesterol levels when compared to HP1 alleles (**Fig. 6a**,**b**), the effect was modestly enhanced for HP2 alleles with the derived rs2000999[A] allele (**Fig. 6c**,**d**). When we corrected for the effect of rs2000999, HP2 alleles on all European SNP haplotype backgrounds (A–C as shown in **Fig. 2**) were associated with similarly elevated cholesterol levels (**Fig. 6e**,**f**). We believe that the impact of rs2000999 on *HP* expression explains the residual nominal association that is present at this SNP in analyses conditioning on HP1/HP2. The imputation efficacy of the HP1/HP2 difference is similar for each SNP haplotype background (haplotype A, $r^2 = 0.93$; haplotype B, $r^2 = 0.95$; haplotype C, $r^2 = 0.95$; using SNPs on the Illumina Omni2.5 array), indicating that imperfect imputation is unlikely to have strongly biased the association toward rs2000999 or any other SNP.

This analysis indicates that the association of cholesterol phenotypes with SNPs near *HP* reflects a complex allelic architecture arising from multiple variants and historical mutations (structural and single nucleotide) at the locus. The status of rs2000999 as the lead (index) SNP at this locus likely reflects a combination of (i) a true genetic effect of this SNP, arising from an effect on *HP* expression levels and explaining an increase of ~1.49 mg/dl in total cholesterol concentration, and (ii) partial LD ($r^2 = 0.14$) with a larger effect (2.11 mg/dl increase in total cholesterol concentration) arising from *HP* structural variation that changes the encoded protein (**Supplementary Table 8**).

DISCUSSION

We have presented multiple lines of evidence that recurrent deletions in HP2 have created new HP1 alleles, a phenomenon that likely explains the low LD between individual SNPs and HP1/HP2. We also found that *HP* is polymorphic for paralogous gene conversion from *HPR*, which has obscured the CNV from analysis by earlier sequencing and array-based CNV studies. Although recurring deletions and paralogous gene conversion have historically made studying this structural variation challenging, we demonstrated that *HP* subtypes can be imputed from SNP haplotypes with high accuracy, an approach that should make it possible to resolve longstanding uncertainty about how genetic variation at *HP* relates to many human phenotypes. We used this imputation strategy to study *HP* variation in 22,288 individuals and showed that a complex allelic architecture, shaped most strongly by the *HP* CNV and also by a *cis*-acting expression effect, is likely responsible for the strong association of cholesterol levels with 16q22.2 in GWAS[18].

Haptoglobin interacts with the APOE protein, which is critical to maintaining low total cholesterol and LDL cholesterol levels[42]. Oxidation of APOE impairs its ability to clear plasma lipids[43]. The HP protein directly binds APOE[20,22] and serves as an APOE antioxidant[20]. The HP2 form of the protein is a less efficient antioxidant than the HP1 form[6], providing a potential mechanism for the association we observe. Decreased HP levels due to rs2000999[A] may have a similar phenotypic effect but by reducing the level (rather than changing the protein structure) of HP. HP2 and rs2000999[A] could contribute to increased total and LDL cholesterol levels by providing insufficient antioxidant activity for APOE (**Fig. 7**).

We found that imputation could be used to extend the analysis of a complex CNV locus to very large samples ($n = 22,288$) for which SNP data were available. The large sample was critical for resolving multiple effects that are in partial LD and made it possible to appreciate (at high levels of significance) effects that were not apparent in an earlier, smaller study[44]. There are currently controversies about the role of the *HP* CNV in heart disease, cancer, malaria, Crohn's disease and numerous other human phenotypes. Our approach to imputing complex structural alleles and the imputation resource we make available here (**Supplementary Data Set**) should make it possible to resolve these questions in a definitive way using large existing SNP

data sets. A similar approach might be useful at the hundreds of other loci affected by complex and multiallelic CNVs.

GWAS has identified thousands of genetic variants that associate with genetically complex traits. At almost all of these loci, the responsible, functional variants have yet to be found and the underlying allelic architectures are unknown. A particularly intriguing question involves the extent to which the underlying allelic architectures will turn out to be simple (for example, a single responsible functional variant) or complex. Haptoglobin appears to offer an early example of a locus at which an association signal arises from the combined effects of many different common functional alleles with different kinds of effects—a set of many alleles that affect protein structure and an additional allele that affects expression level. It will be interesting and important to understand how widespread such allelic complexity is in human biology.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data are available on GenBank under accessions KT923758–KT923784.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank C. Usher for comments on the manuscript and work on the figures. This work was supported by a grant from the National Human Genome Research Institute (R01HG006855 to S.A.M.). The Yerkes Center (grant P51OD011132) provided primate DNA samples. R.M.S. was supported by a US National Institutes of Health/National Heart, Lung, and Blood Institute K99 award (1K99HL122515-01A1) and an advanced postdoctoral fellowship award from the Juvenile Diabetes Research Foundation (JDRF 3-APF-2014-111-A-N). G.M.P. was supported by the National Heart, Lung, and Blood Institute of the US National Institutes of Health under award K01HL125751.

AUTHOR CONTRIBUTIONS

L.M.B., S.A.M. and R.E.H. designed the experiments for understanding *HP* structural evolution. R.M.S., L.M.B. and G.M.P. performed imputation and association analyses of cholesterol cohorts. L.M.B. performed computational analyses of HapMap and 1000 Genomes Project data, constructed the imputation reference panels and performed all laboratory experiments. L.M.B. and S.A.M. wrote the manuscript. J.N.H. and S.K. provided advice on data analysis. All authors contributed to interpretations of data and to revisions of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Allison, A.C. & Rees, W.A. The binding of haemoglobin by plasma proteins (haptoglobins); its bearing on the renal threshold for haemoglobin and the aetiology of haemoglobinuria. *BMJ* **2**, 1137–1143 (1957).
2. Langlois, M.R. & Delanghe, J.R. Biological and clinical significance of haptoglobin polymorphism in humans. *Clin. Chem.* **42**, 1589–1600 (1996).
3. Smithies, O. & Walker, N.F. Genetic control of some serum proteins in normal humans. *Nature* **176**, 1265–1266 (1955).
4. Wejman, J.C., Hovsepian, D., Wall, J.S., Hainfeld, J.F. & Greer, J. Structure and assembly of haptoglobin polymers by electron microscopy. *J. Mol. Biol.* **174**, 343–368 (1984).
5. Nielsen, M.J. & Moestrup, S.K. Receptor targeting of hemoglobin mediated by the haptoglobins: roles beyond heme scavenging. *Blood* **114**, 764–771 (2009).
6. Melamed-Frank, M. *et al.* Structure-function analysis of the antioxidant properties of haptoglobin. *Blood* **98**, 3693–3698 (2001).
7. Tripathi, A. *et al.* Identification of human zonulin, a physiological modulator of tight junctions, as prehaptoglobin-2. *Proc. Natl. Acad. Sci. USA* **106**, 16799–16804 (2009).
8. Smithies, O., Connell, G.E. & Dixon, G.H. Inheritance of haptoglobin subtypes. *Am. J. Hum. Genet.* **14**, 14–21 (1962).
9. Shindo, S. Haptoglobin subtyping with anti-haptoglobin α chain antibodies. *Electrophoresis* **11**, 483–488 (1990).
10. Martosella, J. & Zolotarjova, N. Multi-component immunoaffinity subtraction and reversed-phase chromatography of human serum. *Methods Mol. Biol.* **425**, 27–39 (2008).
11. Cahill, L.E. *et al.* Currently available versions of genome-wide association studies cannot be used to query the common haptoglobin copy number variant. *J. Am. Coll. Cardiol.* **62**, 860–861 (2013).
12. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
13. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
14. Levy, A.P. *et al.* Haptoglobin phenotype and prevalent coronary heart disease in the Framingham offspring cohort. *Atherosclerosis* **172**, 361–365 (2004).
15. Koch, W. *et al.* Genotyping of the common haptoglobin Hp 1/2 polymorphism based on PCR. *Clin. Chem.* **48**, 1377–1382 (2002).
16. Soejima, M. & Koda, Y. TaqMan-based real-time PCR for genotyping common polymorphisms of haptoglobin (HP1 and HP2). *Clin. Chem.* **54**, 1908–1913 (2008).
17. Zethelius, B. *et al.* Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. *N. Engl. J. Med.* **358**, 2107–2116 (2008).
18. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
19. Salvatore, A. *et al.* Haptoglobin binding to apolipoprotein A-I prevents damage from hydroxyl radicals on its stimulatory activity of the enzyme lecithin-cholesterol acyltransferase. *Biochemistry* **46**, 11158–11168 (2007).
20. Salvatore, A., Cigliano, L., Carlucci, A., Bucci, E.M. & Abrescia, P. Haptoglobin binds apolipoprotein E and influences cholesterol esterification in the cerebrospinal fluid. *J. Neurochem.* **110**, 255–263 (2009).
21. Spagnuolo, M.S. *et al.* Analysis of the haptoglobin binding region on the apolipoprotein A-I–derived P2a peptide. *J. Pept. Sci.* **19**, 220–226 (2013).
22. Cigliano, L., Pugliese, C.R., Spagnuolo, M.S., Palumbo, R. & Abrescia, P. Haptoglobin binds the antiatherogenic protein apolipoprotein E—impairment of apolipoprotein E stimulation of both lecithin:cholesterol acyltransferase activity and cholesterol uptake by hepatocytes. *FEBS J.* **276**, 6158–6171 (2009).
23. Maeda, N., Yang, F., Barnett, D.R., Bowman, B.H. & Smithies, O. Duplication within the haptoglobin Hp2 gene. *Nature* **309**, 131–135 (1984).
24. McEvoy, S.M. & Maeda, N. Complex events in the evolution of the haptoglobin gene cluster in primates. *J. Biol. Chem.* **263**, 15740–15747 (1988).
25. Hardwick, R.J. *et al.* Haptoglobin (*HP*) and haptoglobin-related protein (*HPR*) copy number variation, natural selection, and trypanosomiasis. *Hum. Genet.* **133**, 69–83 (2014).
26. Hindson, B.J. *et al.* High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.* **83**, 8604–8610 (2011).
27. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
28. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
29. Asakawa, J., Kodaira, M., Nakamura, N., Satoh, C. & Fujita, M. Chimerism in humans after intragenic recombination at the haptoglobin locus during early embryogenesis. *Proc. Natl. Acad. Sci. USA* **96**, 10314–10319 (1999).
30. Rodriguez, S. *et al.* Molecular and population analysis of natural selection on the human haptoglobin duplication. *Ann. Hum. Genet.* **76**, 352–362 (2012).
31. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
32. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
33. Gallego Llorente, M. *et al.* Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science* **350**, 820–822 (2015).
34. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).
35. Browning, S.R. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.* **124**, 439–450 (2008).
36. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
37. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
38. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).
39. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
40. Froguel, P. *et al.* A genome-wide association study identifies rs2000999 as a strong genetic determinant of circulating haptoglobin levels. *PLoS One* **7**, e32327 (2012).
41. Soejima, M. *et al.* Genetic factors associated with serum haptoglobin level in a Japanese population. *Clin. Chim. Acta* **433**, 54–57 (2014).
42. Ishibashi, S., Herz, J., Maeda, N., Goldstein, J.L. & Brown, M.S. The two-receptor model of lipoprotein clearance: tests of the hypothesis in "knockout" mice lacking the low density lipoprotein receptor, apolipoprotein E, or both proteins. *Proc. Natl. Acad. Sci. USA* **91**, 4431–4435 (1994).
43. Yang, Y., Cao, Z., Tian, L., Garvey, W.T. & Cheng, G. VPO1 mediates ApoE oxidation and impairs the clearance of plasma lipids. *PLoS One* **8**, e57571 (2013).
44. Guthrie, P.A.I. *et al.* Complexity of a complex trait locus: *HP*, *HPR*, haemoglobin and cholesterol. *Gene* **499**, 8–13 (2012).

ONLINE METHODS

Genotyping *HP* structural variants. To determine the copy number of the *HP* CNV and the other structural polymorphisms, we used a droplet-based digital PCR method[26] to measure copy number at four locations (boundaries A, B, D and E in **Supplementary Fig. 4b**). We designed a pair of PCR primers and a dual-labeled fluorescence-FRET oligonucleotide probe to the sequence of each *HP* boundary and to a two-copy control locus. Intermediate copy number calls were repeated with triplicate measurements. We used a PCR assay for boundary C to verify the consistency of this boundary in HP2SS haplotypes. Only individuals predicted to carry the HP2SS haplotype on the basis of ddPCR measurements produced an amplicon. A sufficient number of assays were designed such that no single incorrect copy number measurement would mistakenly identify a diploid subtype pair as another subtype pair (**Supplementary Table 9**). Allelic copy numbers were determined on the basis of a biallelic copy number model for each sequence boundary (**Supplementary Table 9**). Hardy-Weinberg equilibrium of *HP* subtypes (**Supplementary Table 10**) and faithful transmission of *HP* subtypes in family trios (**Supplementary Table 11**) were verified. One likely three-copy allele was found as well as two rare mutations that interfere with assay amplification (**Supplementary Table 12**). Phasing confirmation of recent deletion alleles was performed with Drop-Phase[45], which is further discussed below (**Supplementary Table 13**). Primer sequences are provided in **Supplementary Table 14**.

SNP haplotype analysis for *HP* CNV and subtypes. *HP* structural variants were phased with SNP haplotypes from the 1000 Genomes Project Phase I data and were used to show short haplotypes in four common clusters (**Fig. 2**) and longer closely related haplotypes in a dendrogram (**Fig. 4**). Haplotypes in **Figure 2** were clustered with the *k*-means method using upstream SNP haplotypes only. All recent deletion alleles (HP1S subtype) as shown in **Figure 4** were from individuals who have two standard HP2FS SNP haplotypes. Each deletion allele was phased onto the correct SNP haplotype using the Drop-Phase technique[45], a new method for phasing based on the idea that physically linked sequences are more frequently partitioned into the same droplets (**Supplementary Table 13**). *HP* structural variants were also phased with SNPs from common SNP genotyping arrays to evaluate the potential for these variants to be imputed from existing GWAS data. Phasing and encoding for the structural alleles are further discussed in the **Supplementary Note**.

Population sequencing in the CNV region. We performed Sanger sequencing of the CNV region of 27 human haplotypes segregating on diverse SNP haplotype backgrounds and the analogous region of four great apes: chimpanzee, bonobo, gorilla and orangutan. Individual human haplotypes were sequenced by targeting a single structural allele from HP1/HP2 heterozygotes. The primers for the human HP2 allele target the HP2 breakpoint. HP1 haplotypes were obtained with size selection through gel extraction. HP1 sequencing primers were designed to be compatible with the chimpanzee, gorilla and orangutan reference genomes and were also used to sequence the corresponding region in each great ape. All primer pairs are specific to *HP* and do not amplify haptoglobin-related protein (*HPR*) or primate haptoglobin (*HPP*). The hg19 reference genome supplied the human *HPR* sequence. The chimpanzee and gorilla *HP* genes were sequenced in samples NS03489 and PR00107, respectively (DNA provided by Coriell Cell Repositories). The sequence for the chimpanzee *HPR* and *HPP* genes was supplied by previously sequenced clones (GenBank, M84462.1 and M84463.1) (see **Supplementary Figs. 2** and **3** for additional great apes and results by sample and see **Supplementary Table 14** for primer sequences).

Creating and testing imputation reference panels. We evaluated the efficacy of imputation for the *HP* CNV as well as *HP* subtypes (HP1F, HP1S, HP2FS and HP2SS) using reference panels composed of experimentally determined *HP* structural alleles and SNPs ascertained from the 1000 Genomes Project[13] and HapMap[28]. Separate reference panels were created and tested from each of the following SNP data sets: Illumina Omni2.5, HapMap 3 (Illumina 1M and Affymetrix 6.0) and Illumina 1M and Affymetrix 6.0 individually (**Table 1** and **Supplementary Tables 2–5**). Separate reference panels were created and tested for European and African populations because of differences in SNP haplotype backgrounds for *HP* subtypes (**Fig. 4**). We performed a series of leave-one-out trials to evaluate the efficacy of imputation for *HP* structural variants. See the **Supplementary Note** for more information and **Supplementary Table 15** for population identifiers.

Imputation of *HP* structural variation into cohorts for cholesterol association study. A reference panel composed of encoded *HP* structural alleles and SNPs surrounding the CNV region from the Illumina Omni, Illumina 1M and Affymetrix 6.0 arrays was developed and used to impute *HP* structural variation into cohorts with cholesterol information using Beagle (v2.3.1) imputation software. See the **Supplementary Note** for further detail.

Association analysis. The analysis of association between imputed *HP* structural variants and the four lipid traits (total cholesterol, LDL cholesterol, high-density lipoprotein (HDL) cholesterol and triglycerides) was performed in six studies of 22,288 individuals of European ancestry. Each lipid trait was regressed on age and sex and was inverse-normal transformed before analysis. Linear regression was performed to test the association between imputed structural variants or SNPs in the locus and the lipid trait, assuming an additive genetic model, using PLINK[46] (v1.07). The imputed *HP* structural variants and genotypes were analyzed as dosages to account for imputation uncertainty, and poorly imputed variants were discarded (imputation info <0.4). All analyses were adjusted for ten study-specific principal components. Study-specific results were combined via the inverse-variance fixed-effects meta-analysis method implemented in Metal[47]. Sensitivity and specificity phenotype analyses were performed to assess the influence of type 2 diabetes (condition or removal of samples) and cholesterol-lowering/statin medication use (recalculating values, condition or removal of samples). All analyses were performed using baseline lipid measurements for cohorts with longitudinal follow-up. See **Supplementary Figure 9** for HDL cholesterol and triglyceride association results and the **Supplementary Note** for more information.

Code availability. The following packages were used to analyze data and are publicly available online: Beagle[35] (v2.3.1), PLINK[46] (v1.07), SHAPEIT2 (ref. 48; version 2.644), IMPUTE2 (ref. 49; version 2.3), Metal[47] and SMARTPCA[50]. The following custom scripts are available upon request: R scripts used to format data, to perform linear regression analyses and to cluster the haplotypes in **Figure 2**, a Python script used to cluster the haplotypes in **Figure 4** and Perl scripts used to format data.

45. Regan, J.F. *et al.* A rapid molecular approach for chromosomal phasing. *PLoS One* **10**, e0118270 (2015).
46. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
47. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
48. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
49. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
50. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).