The evolution of CHC22 clathrin influences its role in human glucose metabolism

Matteo Fumagalli [1,2,*], Yoan Diekmann [3,*], Stephane M. Camus [2,*], Alice Burke [2], Marine D. Camus [2], Paul J. Norman [4], Agnel Praveen Joseph [5], Laurent Abi-Rached [6], Andrea Benazzo [7], Rita Rasteiro [8], Iain Mathieson [9], Maya Topf [5], Peter Parham [10], Mark G. Thomas [3,11], Frances M. Brodsky [2,^]

[1] Department of Life Sciences, Silwood Park campus, Imperial College London, Ascot SL5 7PY, UK

[2] Division of Biosciences, University College London, London WC1E 6BT, UK

[3] Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK

[4] Division of Personalized Medicine and Department of Immunology, University of Colorado, Aurora, CO 80045, USA

[5] Institute of Structural and Molecular Biology, Birkbeck College and University College London, UK

[6] Equipe ATIP, URMITE UM63 CNRS 7278 IRD 198 INSERM 1095, IHU Méditerranée Infection, Aix Marseille Université, Marseille, France

[7] Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara 44121, Italy

[8] School of Biological Sciences, University of Bristol, Bristol BS8 1TQ, UK

[9] Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA 19104, USA

[10] Departments of Structural Biology and Microbiology & Immunology, Stanford University, Stanford, CA 94305, USA

[11] UCL Genetics Institute, University College London, Gower St, London WC1E 6BT, UK

* these authors contributed equally to this study

^ to whom correspondence should be addressed:

Professor Frances M. Brodsky
Division of Biosciences
University College London
Medical Sciences Building (Room 130B)
Gower Street, London WC1E 6BT
Tel: 44 (0)20 3549 5464
Fax:44 (0)20 7679 2682
f.brodsky@ucl.ac.uk

Short title: Evolution of CHC22 clathrin affects its role in glucose metabolism

List of abbreviations:

CCV: clathrin-coated vesicles
CHC: clathrin heavy chain
CLC: clathrin light chain
GLUT4: glucose transporter 4
SNP: single nucleotide polymorphism
GFP: green fluorescent protein
FRAP: fluorescence recovery after photobleaching
GSC: GLUT4 storage compartment
HA: hemagglutinin
FACS: fluorescence-activated cell sorting

ABSTRACT

CHC22 clathrin plays a key role in intracellular membrane trafficking of the insulin-responsive glucose transporter GLUT4, and so in post-prandial clearance of glucose from human blood. We performed population genetic and phylogenetic analyses of the *CLTCL1* gene, encoding CHC22, to gain insight into its functional evolution. Analysis of 58 vertebrate genomes showed independent loss of *CLTCL1* in at least two lineages after it arose from a gene duplication during the emergence of jawed vertebrates. All vertebrates studied retain the parent *CLTC* gene encoding CHC17 clathrin, which mediates endocytosis and other housekeeping pathways of membrane traffic, as performed by the single type of clathrin in non-vertebrate eukaryotes. For those species retaining *CLTCL1*, preservation of CHC22 functionality was supported by strong evidence for purifying selection over phylogenetic timescales, as seen for *CLTC*. Nonetheless, *CLTCL1* showed considerably greater allelic diversity than *CLTC* in humans and chimpanzees. In all human population samples studied, two allelic variants of *CLTCL1* segregate at high frequency, encoding CHC22 proteins with either methionine or valine at position 1316. Balancing selection of these two allotypes is inferred, with V1316 being more frequent in farming populations, when compared to hunter-gatherers, and originating an estimated 500-50 thousand years ago. Functional studies indicate that V1316-CHC22 is less effective at controlling GLUT4 membrane traffic than M1316-CHC22, leading to an attenuated insulin-regulated response, consistent with structural predictions and measurable differences in cellular dynamics of the two variants. These analyses suggest that CHC22 clathrin was subject to selection in humans with different diets, leading to allotypes that affect its role in nutrient metabolism and have potential to differentially influence the human insulin response.

INTRODUCTION

Clathrin-coated vesicles (CCVs) are key players in eukaryotic intracellular membrane trafficking (1). Their characteristic lattice-like coat is self-assembled from cytoplasmic clathrin proteins, captures membrane-embedded protein cargo and deforms the membrane into a vesicle. This process enables CCVs to mediate protein transport to and from the plasma membrane and between organelles. The triskelion-shaped clathrin molecule is formed from three identical clathrin heavy chain (CHC) subunits. Humans have two genes (*CLTC* and *CLTCL1*) that respectively encode CHC17 and CHC22 clathrins (2). CHC17 clathrin, which has three bound clathrin light chain (CLC) subunits, is expressed uniformly in all tissues and forms CCVs that control receptor-mediated endocytosis, as well as lysosome biogenesis and maturation of regulated secretory granules. These pathways are conventionally associated with clathrin function and are mediated by clathrin in all eukaryotic cells (1). In humans CHC22 clathrin is most highly expressed in muscle and adipose tissue and forms separate CCVs that are not involved in endocytosis (3). Instead, CHC22 CCVs regulate targeting of the glucose transporter 4 (GLUT4) to an intracellular compartment where it is sequestered until released to the cell surface in response to insulin (4). This insulin-responsive GLUT4 pathway is the dominant mechanism for clearing blood glucose into muscle and fat after a meal (5). In addition to their distinct tissue expression patterns and biological functions, CHC22 does not bind the CLC subunits that associate with CHC17 clathrin, even though the CHC protein sequences are 85% identical (3,6). This remarkable biochemical and functional divergence evolved since the gene duplication event that gave rise to the two different clathrins during the emergence of chordates (2). Notably, however, the *CLTCL1* gene encoding CHC22 evolved into a pseudogene in the *Mus* genus, although mice maintain an insulin-responsive GLUT4 pathway for clearing blood glucose. This observation suggests that, despite the remarkable importance of the *CLTCL1* gene product, backup pathways have evolved to compensate for the loss of CHC22 protein. To understand the evolution of the specialized function of CHC22, and the potential selective processes involved, we here

explore the phylogenetic history of the *CLTCL1* gene in vertebrates and its population genetics in humans, non-human primates and bears.

Ecological shifts create selective forces that filter variation in cellular genes. These include changes in nutritional conditions (7), as well as encounters with pathogens (8); both documented as selective forces that affect membrane traffic genes (9,10). Recent studies of the evolution of genes involved in membrane trafficking have focused on an overview of all eukaryotes with the goals of establishing the origins of membrane-traffic regulating proteins in the last common eukaryotic ancestor and defining the species distribution of various families of traffic-regulating proteins (11,12). These studies have identified common features of proteins that regulate membrane traffic (11) and revealed that extensive gene duplication has allowed lineage-specific diversification of coat proteins and other membrane traffic regulators, such as the Rab GTPases (13,14). Our earlier study of available annotated genomes in 2005 suggested that the gene duplication giving to rise to the two CHC-encoding genes occurred as a result of one of the whole genome duplications contributing to chordate evolution (2). Here we focus on the more recent evolutionary history of these genes, as well as analyze the increased number of fully annotated vertebrate genomes. We establish that the *Mus* genus is not unique in its post-chordate loss of *CLTCL1*, identifying at least one additional independent gene loss event in the clade of *Cetartiodactyla*, and possibly more. Nonetheless, there is strong evidence for CHC22 sequence conservation amongst those species that retained *CLTCL1* (2). This evolutionarily recent gene loss in some lineages and retention of the functional form in others suggested that *CLTCL1* may still be under purifying selection, so we examined *CLTCL1* variation between individuals within vertebrate populations. Comparing populations, we found *CLTCL1* to be considerably more polymorphic than *CLTC*, which encodes the clathrin found in all eukaryotes, with evidence for strong ancient purifying selection for CHC17 clathrin function and diversifying (or relaxed purifying) selection on CHC22 function. Additionally, we identified two high-frequency allelic variants of human CHC22, which have different functional properties. The derived allele

arose in ancient humans and is more frequent in farming populations when compared to hunter-gatherers. Similarly, *CLTCL1* allelic variants segregated in bears with different diets (prevalently carnivorous vs prevalently herbivorous). We previously observed that CHC22 accumulates at sites of GLUT4 retention in the muscle of insulin-resistant patients with type 2 diabetes (4) in addition to its active role in membrane traffic of GLUT4. Thus, CHC22 variation has potential to differentially affect membrane traffic pathways involved in insulin resistance, as well as alter normal glucose metabolism within human and other vertebrate populations. The analyses reported here lead us to propose that variation in the CHC22 clathrin coat may be a response to changing nutritional pressures both between and within vertebrate species.

RESULTS

Phylogenetic analyses reveal selective loss or retention of a functional *CLTCL1* gene in vertebrates

Identification of CHC-encoding genes in 62 vertebrate and non-vertebrate species (Fig. S1-S2) indicates a dynamic history of gene duplications and losses (Fig. 1). The *CLTCL1* gene was detected only in jawed vertebrates (bony vertebrates and cartilaginous fish, *Gnathostomata*). The lamprey (*Petromyzon marinus*) and hagfish (*Eptatretus burgeri*), two jawless vertebrates, have only one CHC-encoding gene, while the elephant- and whale shark (*Callorhinchus milii* and *Rhincodon typus*, respectively) genomes have both. This distribution refines the timing of the CHC-encoding gene duplication to the period after lampreys and hagfish split off the vertebrate lineage, estimated at 493.8MYA (2.5% and 97.5% HPD: 459.3, 533.8), and before the evolution of jawed vertebrates 450.8MYA (2.5% and 97.5% HPD: 432.1, 468.1) (15,16). Of the ten species of bony fish that split off the spotted gar (*Lepisosteus oculatus*) lineage (17) whose genomes are generally tetraploid, all had two versions of *CLTC*, one had two *CLTCL1* genes and one (cave fish, *Astyanax*

mexicanus) apparently lost *CLTCL1*. Excluding bony fish, *CLTC* was found to be further duplicated in zebra finch (*Taeniopygia guttata*), but in no other vertebrates. Eight additional species of vertebrates with high genome coverage and reliable annotation had no identifiable *CLCTL1* gene, implying at least five independent loss events (including the cave fish loss) that were clade- or species-specific. *CLTC* and *CLTCL1* genes are present in the *Caviomorpha* and *Scuridae* rodent suborders (guinea pig (*Cavia porcellus*), Alpine marmot (*Marmota marmot*) and ground squirrel (*Ictidomys tridecemlineatus*), and lost in the *Muroidea* suborder from the entire *Mus* genus (2) and from rat (*Rattus norvegicus*). The *Cetartiodactyla* clade has lost *CLTCL1* altogether. However, it is only represented by four genomes in our dataset: pig (*Sus scrofa*), sheep (*Ovis aries*), cow (*Bos taurus*) and Yangtze finless porpoise (*Neophocaena asiaeorientalis*). The absence in all analyzed species of both clades suggests two independent loss events before speciation. The absence of *CLTCL1* in rat clarifies why CHC22 could not be biochemically identified in rat and indicates that antibodies against CHC22 that react with rat cells must cross-react with other proteins (18). *CLTCL1* was also not detected in the genomes of the little brown bat (*Myotis lucifugus*) and the duck-billed platypus (*Ornithorhynchus anatinus*). These may represent additional independent loss events, assuming the genome annotations for these species are reliable. The intermittent loss of *CLTCL1* and the retention of *CLTC* raises the question of whether their patterns of evolution are typical for genes with related functions that duplicated in the same time frame.

CLTC and *CLTCL1* are located on paralogous regions of human chromosomes 17 and 22, respectively. For these two genes, the evolutionary rates (rate of non-synonymous mutations to rate of synonymous mutations; dN/dS) across vertebrates at each position were determined and plotted along the length of the protein sequences (Fig. 2A, B). Several adjacent paralogues have been maintained in these chromosomal regions, some of which are involved in membrane traffic, including the gene pair of *MTMR4* and *MTMR3*, encoding the myotubularin lipid phosphatases. Also, CLC subunits of CHC17 clathrin are encoded by paralogous genes (*CLTA* and *CLTB*) that arose from a local gene duplication, mapped to the

same time frame as the CHC-encoding duplication (2). Comparison of the distribution of dN/dS ratios for the three pairs revealed stronger purifying selection on the *CLTC/CLTCL1* genes than on *MTMR4/MTMR3* and *CLTA/CLTB* (Fig. 2C-E), suggesting the clathrin-encoding clade is more evolutionarily constrained. This observation is consistent with our previous identification of conserved signature residues in *CLTCL1* using DIVERGE analysis (2) and indicates conserved functions for both the *CLTC* and *CLTCL1* gene products. Furthermore, there was a striking difference in the distribution and average of evolutionary rates between *CLTC* and *CLTCL1* (Kolmogorov-Smirnov test *p*-value < 2.2e-16), with *CLTC* being significantly more constrained by purifying selection than *CLTCL1*. In contrast, there was minimal difference in the distribution and average of evolutionary rates between the two paralogs in *MTMR4/MTMR3* and *CLTA/CLTB* (Kolmogorov-Smirnov test yields *p*-values 0.003643 and 0.9959 respectively).

Human population genetic analyses indicate purifying selection with ongoing diversification for *CLTCL1*

To follow up the indication that *CLTC* and *CLTCL1* are subject to different degrees of purifying selection during evolution, we investigated their variation in human populations. We analyzed 2,504 genomes from the 1000 Genomes Project database, phase 3 (19) and identified alleles resulting from nonsynonymous substitutions for *CLTC* and *CLTCL1.* This dataset included individuals from each of five human meta-populations: European (EUR, 503), East Asian (EAS, 504), Admixed American (AMR, 347), South Asian (SAS, 489) and African (AFR, 661). The reference sequences for chimpanzee (*Pan troglodytes*) and pseudo-references for two archaic humans, Altai Neanderthal and Denisovan, were also included to relate allelic variation to the ancestral state. A median-joining network for all the inferred *CLTC* human alleles showed a very common allele (sample frequency = 0.997) with only five low-frequency variants generating a total of six alleles (Fig. 3A). In this network, the allele frequency is indicated by the size of the circle (in $\log_2$ scale) and line length between circles is proportional the number of non-synonymous changes between alleles. Each allele

encodes a variant of CHC22 (allotype), which can include one or more single-nucleotide polymorphisms (SNPs). The meta-populations in which the allele is found are indicated in color representing their percentage of the total frequency of the allele in humans.

In contrast to *CLTC*, we identified 46 non-synonymous SNPs in *CLTCL1*, present in 52 distinct haplotypes (referred to here as alleles, following the definition given above, see Tables S1-S2). A median-joining network for the most common *CLTCL1* alleles showed that they are widely distributed within the human macro-populations (Fig. 3B). Each macro-population tends to have private, less frequent alleles. Nevertheless, all the macro-populations comprised two main allelic clades, together constituting a sample frequency of 77%. Note that the two main alleles were downsampled by a third of their frequency, in the plot only, for visualization purpose. These alleles differ by a single M1316V substitution (SNP ID rs1061325 with genomic location chr22:19184095 on hg19 assembly). The valine at position 1316 is predicted to have a functional effect on the protein since it was categorized as “probably damaging” with a probability of 0.975 by PolyPhen (20) and as “damaging” by SIFT (21). The ancestral sequences from chimpanzee and archaic humans have the M1316 allotype, suggesting that M1316 is likely to represent the ancestral state. To further investigate this, we inspected raw sequencing data from both Altai Neanderthal and Denisovan (Table S3). We inferred the most likely genotype to be homozygous for the M1316 amino acid (minimum depths equal to 40 and 28, respectively). We further extracted sequencing data for additional 13 archaic and ancient humans (Table S3). We found that the V1316 amino acid is present in Pleistocene hunter-gatherers and Neolithic farmers but not in other Neanderthals or Holocene hunter-gatherers in this limited data set. The equivalent residue in human CHC17 (encoded by *CLTC*) is also methionine, suggesting that methionine at this position likely pre-dated the initial duplication generating *CLTCL1*. For the non-human species analyzed (Fig. 1), all CHC-encoding genes present have M1316, further indicating a conserved role in CHC structure or function.

To quantify the levels of nucleotide and allelic diversity for non-synonymous sites within human populations, several summary statistics were calculated. For populations within each meta-population, we separately calculated the number of segregating sites (S), the number of segregating singletons (SS), Tajima's D (TD), Fu and Li's D* (FLDs) and F* (FLFs), the sum of square allele frequencies including the most common allele (H1) and excluding it (H2), and the ratio between H2 and H1 (H2H1) (Table S4). All populations display negative TD values with many populations exhibiting negative FLDs and FLFs values. These values are consistent with low diversity within common alleles and an excess of low-frequency variants (namely due to singletons).

To assess whether such summary statistics are expected or not under neutral evolution in each population, a set of control genes with the same coding length as *CLTCL1* were similarly processed (Table S5). Since we analyzed only non-synonymous substitutions, which are likely to behave differently than silent mutations, standard coalescent simulations are not suitable for deriving the expected distribution of genetic diversity. By analyzing more than 500 control genes, we computed the empirical percentile rank for each statistic in each population. High or low values (e.g. >0.99 or <0.01) are indicative of statistics for *CLTCL1* being outlier in the empirical null-distribution with values unlikely to occur under mutation and neutral genetic drift alone. Summary statistics and populations were then clustered according to their empirical ranks and plotted on a heat map (Fig. 4). All populations exhibited high empirical ranks for S, with *CLTCL1* being an outlier in the distribution for many populations (percentile rank > 0.95). Similarly, most of them, apart from CHS (East Asians) and STU (South East Asian), showed a high ranking for SS (the number of singletons), although this metric is very sensitive to sample size. Nevertheless, all populations *CLTCL1* showed high values for H2 and H2H1 and low values for H1, all suggesting an unusually high frequency for the second most common allele, a situation likely to occur under balancing selection (22) or because of a soft sweep (23). That *CLTCL1* was low ranking in all populations for H1, a statistic representing the frequency of the most common allele, also

supported diversifying selection rather than hard sweeps. Similarly, TD, FLDs, and FLFs values for *CLTCL1* were generally lower than observed for the >500 control genes, although this was not consistent across all populations. Inconsistency may be partly explained by these statistics not having the power to capture complex selective scenarios. Nevertheless, the distribution of empirical percentile ranks for all analytical parameters of *CLTCL1* variation (Fig. 4) was bimodal. Under neutrality, uniform distribution would be expected, so this deviation is suggestive of diversifying selection acting on *CLTCL1*. We also calculated a measure of genetic differentiation (fixation index F_{ST}) between pairs of canonical reference populations, namely Yoruba from Nigeria (YRI), North Americans with Caucasian European ancestry (CEU), and Han Chinese from Beijing (CHB). We did not find any suggestion that F_{ST} values for *CLTCL1* are outliers in the empirical distribution of control genes, with percentiles ranks between 0.15 and 0.37. This is compatible with selection acting on distinct populations to different extents. As we considered nonsynonymous changes only, the limited number of SNPs considered per gene may create a larger variance in these empirical distributions, especially for allele-based statistics. We therefore sought further evidence for the high frequency of the second most common allele by investigating the whole genomic variation, including silent SNPs. We observed a local increase of H2 statistics in *CLTCL1* for European populations, which already showed a large value based on nonsynonymous changes (Fig. S3). This analysis also indicates that selection signatures are restricted to a local genomic region encompassing *CLTCL1*.

The high frequency of two major alleles of *CLTCL1* that appear in all modern human populations (Fig. 3B, Table S1) is a potential indicator of balancing selection (22). Such a distribution was confirmed using a different data set of more than 50 sampled human populations (Fig. S4). In several populations, we also observed an apparent excess of heterozygosity at SNP rs1061325 (Table S6), suggestive of heterozygote advantage (overdominance) for the two encoded allotypes differing at residue 1316. Specifically, all European populations show a ratio of observed vs. expected (assuming Hardy-Weinberg

equilibrium) heterozygosity greater than 1, with the highest value of 1.24 for Iberic Spanish (IBS) (Fig. 5). The excess of heterozygosity for some populations was confirmed when only the two main alleles (as in Fig. 3B) were considered (Table S7). A selective pressure that might be acting on *CLTCL1*, irrespective of population distribution, could be changes in dietary habits that have been occurring globally since the advent of agriculture. Farming, followed by food manufacture has gradually, then dramatically, increased carbohydrate availability and consumption by humans. As CHC22 clathrin, the gene product of *CLTCL1*, is required for formation of the intracellular pathway critical for an insulin response, its genetic history could potentially be influenced by these changes. To address the hypothesis that nutritional habits conferred selective pressure on *CLTCL1*, we compared the frequency of SNP rs1061325 in farming vs hunter-gatherer population samples from ancient and modern humans. Although the appearance of SNP rs1061325 predates the advent of farming (Table S3), the observed frequencies of this allele, which encodes the CHC22-V1316 allotype, are consistent with a tendency to increase once farming became common practice for a population (Fig. 6), although limited sample size for modern humans may have reduced the power to reach statistical significance. The highest difference in allele frequency was observed between early farmers and hunter-gatherers from west Eurasia. However, we caution that these two populations are highly diverged, and it remains possible that this significant difference in allele frequency is due to genetic drift shaped by population history, rather than natural selection.

Genetic variation in non-human vertebrate populations supports functional diversification of *CLTCL1*

We analyzed allelic variation for *CLTC* and *CLTCL1* in the genomes of 79 individuals representing six species of great ape, two species each for chimpanzee, gorilla and orangutan (*Pan troglodytes*, *Pan paniscus*, *Gorilla beringei*, *Gorilla gorilla*, *Pongo abellii*, *Pongo pygmaeus*). After data filtering and haplotype phasing, we found no non-synonymous SNPs for *CLTC* and 64 putative non-synonymous SNPs for *CLTCL1* (Table S8). In gorillas,

orangutans and chimpanzees one of the non-synonymous changes in *CLTCL1* leads to a premature stop-codon at amino position 41, with an overall frequency of 36%. However, sequences containing the stop-codon exhibited only a marginal increase of nucleotide diversity (+4.7% as measured by Watterson's index (24)) compared to the full-length sequences, suggesting that these are relatively new variants. Notably, for all the non-human primates analyzed, *CLTCL1* does not have the V1316 allotype, which appears private to humans. However, in gorillas, orangutans and chimpanzees we observe a common but different substitution at the same amino acid position to T1316.

To further investigate variation in non-human primates, we increased the sample size per species by analyzing *CLTCL1* variation in 70 genomes of chimpanzees and bonobos. A median-joining network for the inferred 8 *CLTCL1* alleles (Table S9) showed a major allele common to different species with less frequent alleles primarily restricted to individual ones (Fig. 7). We observed considerable diversity in non-human primates, with a potential tendency towards multiple variants. However, amino acid 1316 was not covered in this data set, possibly due to poor sequencing mapping quality associated with the high nucleotide diversity observed. In another data set of 20 individuals (25), we confirmed a frequency of 10% of T1316 amino acid polymorphism in chimpanzee, but not in bonobos.

We further investigated *CLTCL1* variation in polar bears (*Ursus maritimus*) and their closest related species, brown bears (*Ursus arctos*). These two species, which diverged 479-343 KYA (10), experience very different diets (10,26) and are phylogenetically closer to each other than chimpanzees and humans. Polar bears subsist on a high fat, low carbohydrate diet and brown bears consume a more varied diet of carbohydrate, protein and fat. Analysis of 22 bear genomes (6 polar bears and 16 brown bears) (27), revealed three amino acid substitutions at positions 1267, 1389, and 1522 which are fixed in polar bears and either absent or polymorphic in brown bears. Genetic differentiation between polar and brown bears, as measured by F_{ST}, is markedly higher for *CLTCL1* (0.56) than for *CLTC* (0.26) (10).

This sample of bear populations supports the emergence of multiple *CLTCL1* variants within a species and potential diet-related selection.

Modelling CHC22 variation based on *CLTCL1* polymorphism suggests an effect on clathrin lattice contacts

One explanation for selection of the human-specific *CLTCL1* allele encoding the CHC22-V1316 allotype is that this amino acid change might confer a functional change in the clathrin lattice. This was predicted by the PolyPhen and SIFT analyses, highlighting the change as potentially structure-altering. Many humans are heterozygous for the M1316/V1316 allotypes (44% based on all individuals from 1000 Genomes project), and there may potentially be special properties for mixed lattices formed from the two gene products. To address the possibility that this polymorphism affects protein function, we used MODELLER (28) to produce a homology model of the two CHC22 allotypes based on the crystal structure of CHC17 clathrin (PDB 1B89), taking advantage of the 85% protein sequence identity between human CHC17 and CHC22 (Fig. 8). Modeling using UCSF Chimera (29) showed that residue 1316 is found at a key interface between triskelion legs in assembled clathrin (Fig. 8A and top of panel B). If M1316 is substituted by V1316, the smaller side chain creates a void that would be energetically unfavorable (Fig. 8, bottom of panel B), such that the triskelion leg might twist slightly to close the void. In the clathrin lattice, the legs have a torque that rotates the assembly interface along the protein sequence (30), so a further twist could slightly adjust the interface, altering assembly interactions. Changes in the assembly interface could affect integrity of the lattice and potentially influence kinetics of assembly and disassembly. Mixed lattices of the two CHC22 allotypes would therefore have different properties from CHC22 coats formed in homozygotes for the two major *CLTCL1* alleles. CHC22 is needed for the traffic of GLUT4 to its intracellular storage compartment, where GLUT4 awaits release to the plasma membrane in response to insulin. However, CHC22 also accumulates at the GLUT4 storage compartment when it expands due to impaired GLUT4 release in cases of insulin-resistant type 2 diabetes (4). Thus, genetic variation of

CHC22 could alter rates of retention and release of GLUT4 in both healthy and disease states.

CHC22 variants display functional differences

To test whether the evolutionary change from M1316 to V1316 in CHC22 clathrin alters its properties, three aspects of CHC22 biochemistry and function were compared for the two allotypes. HeLa cells were transfected with constructs encoding each CHC22 variant or CHC17 tagged with green fluorescent protein (GFP). Atypically for their epithelial cell origin but not for transformed cells, HeLa cells express CHC22 clathrin (they are homozygous for the M1316 allotype) (31,32). We observed that the transfected fluorescently tagged CHC22 allotypes were both concentrated in the perinuclear region of the cell, similar to endogenous CHC22-M1316 detected by antibody, and did not overlap with endogenous CHC17 (Fig. 9A). Conversely, transfected GFP-CHC17 did not overlap with endogenous CHC22, so expression of the transfected CHCs reflected their natural distribution (3). Using these constructs, the dynamics of membrane association for the two allotypes of CHC22 and for CHC17 were assessed by Fluorescence Recovery After Photobleaching (FRAP). To assess clathrin turnover, as an indicator of clathrin coat stability, cells expressing fluorescent proteins were photobleached in the perinuclear area (Fig. 9B) and the rate of recovery of fluorescence was measured. Recovery of CHC17 fluorescence was the most rapid, consistent with its more soluble properties compared to CHC22 (3). CHC22-M1316 showed the slowest recovery and CHC22-V1316 was intermediate (Fig. 9C-E), suggesting that it is more exchangeable in the CHC22 coat than the M1316 allotype.

The impact of CHC22 variation on GLUT4 retention was then assessed. Because HeLa cells express CHC22, they can form a GLUT4 storage compartment (GSC), when transfected to express GLUT4. These cells can sequester GLUT4 intracellularly, and then release it to the plasma membrane in response to insulin, behaving like muscle and adipocytes, though with more modest insulin response (33-35). To detect GLUT4 release to the cell surface, we used

a construct expressing GLUT4 tagged with mCherry and a hemagglutinin (HA) epitope embedded in an exofacial loop of the transporter (HA-GLUT4-mCherry). Appearance of surface GLUT4 in response to insulin can be detected by fluorescence-activated cell sorting (FACS) using an antibody to the HA epitope (Fig. 10A). Transfection of HeLa cells with siRNA depleting CHC22 ablates this insulin-responsive pathway (33) (Fig. 10A). We then assessed whether siRNA inhibition of insulin-responsive GLUT4 release can be rescued by expression of CHC22-M1316-GFP or CHC22-V1316-GFP. These are the same constructs characterized in Fig. 9A and they are siRNA-resistant, as well as being GFP-tagged. We observed that, when endogenous CHC22 was depleted, CHC22-M1316 was able to restore the insulin response but CHC22-V1316 was not, when the rescue constructs were expressed at the same levels in cells (measured by intensity of GFP fluorescence) (Fig. 10A). CHC17 expression also did not rescue insulin-induced GLUT4 expression, as shown previously (4). However, CHC22-V1316 is functional for trapping GLUT4 intracellularly because CHC22-transgenic mice that express CHC22-V1316 in their muscle from the natural human promoter show excessive GLUT4 sequestration in their muscle compared to wild-type mice without CHC22, leading to higher blood glucose in the transgenic animals (4). To analyze GLUT4 sequestration in another way, cells depleted for CHC22 and then transfected with mCherry-GLUT4 plus either CHC22 allotype or CHC17 were each divided into three populations expressing equivalently low, medium and high levels of the transfected CHC-GFP. Then, the total GLUT4 content of the cells was measure by mCherry fluorescence. We observed higher levels of GLUT4 in cells expressing CHC22-M1316-GFP, compared to cells expressing either CHC22-V1316-GFP or CHC17-GFP at both medium and high levels of CHC expression (Fig. 10B). This suggests that GLUT4 is sequestered more effectively from degradative membrane traffic pathways when trafficked by CHC22-M1316 than by CHC22-V1316, indicating that the M1316 variant is more efficient at GLUT4 targeting to the GSC. As indicated by their weak insulin response compared to muscle or fat cells, HeLa cells are only just able to form a functional GSC from which GLUT4 can be released. For these cells, the less effective CHC22-V1316 is inadequate to restore GSC

formation when their endogenous CHC22-M1316 is depleted. Using this HeLa model is necessitated by the lack of natural models for the CHC22-dependent GLUT4 pathway in myoblasts and adipocytes, as well as a lack of antibodies that detect surface GLUT4. Nonetheless, these experiments demonstrate a functional difference between CHC22-M1316 and CHC22-V1316 and suggest that CHC22-V1316 is less efficient at GLUT4 sequestration.

DISCUSSION

We analyzed the phylogenetics and population genetics of CHC22 clathrin to understand the functional variation of this protein in relation to its evolutionary history. CHC22 clathrin is a key player in post-prandial clearance of glucose from human blood through its role in intracellular packaging of the GLUT4 glucose transporter in muscle and fat, the tissues in which CHC22 and GLUT4 are expressed (4). The CHC22 pathway positions GLUT4 for cell surface release in response to insulin and consequent uptake of glucose into these tissues (36). The gene encoding CHC22 resulted from gene duplication that we have now dated to 494-451 MYA, early in vertebrate evolution when jawed vertebrates emerged. We had previously shown that the gene encoding CHC22 (*CLTCL1*) is a pseudogene in mice (2). Expanding analysis to 58 vertebrate genomes (>5X coverage) we could not detect *CLTCL1* in nine vertebrate genomes. Six of these absences can be ascribed to two independent gene loss events in branches of the *Rodentia* and the *Cetartidactylae*. The three others may represent additional gene losses or incomplete genome annotation. All vertebrate and non-vertebrate eukaryotes considered here have retained the parent *CLTC* gene encoding CHC17 clathrin, which mediates endocytosis and other housekeeping membrane traffic pathways (12). Our new analysis establishes that *CLTC* is under strong purifying selection. Notably, *CLTCL1* also shows evidence of purifying selection in the species in which it has been retained, supporting the functionality of CHC22 in those species. Compared to *CLTC*,

extensive allelic diversity was observed for *CLTCL1* in all species where populations were analyzed, including humans, chimpanzees and bears. Variant alleles were species-specific in most cases and unique variants were found in species with extreme dietary habits, such as polar bears. In all human populations, two allelic variants of *CLTCL1* are present in high frequency, differing only at one nucleotide, resulting in CHC22 protein with either methionine or valine at position 1316. The V1316 allotype occurs only in humans, but some non-human primates have different variation at the same position. Analysis of ancient humans dated the appearance of the V1316 variant to 500-50 KYA and indicated that M1316, which is present also in CHC17 clathrin, is the ancestral state. Analyses of human population genetic data provided support for balancing selection of the two allotypes of CHC22. Together these analyses suggest that CHC22 clathrin is undergoing selection in humans and we hypothesize that selective pressure results from its role in nutrient metabolism. Consistent with this hypothesis, we observed functional differences between the two CHC22 allotypes in their ability to control GLUT4 membrane traffic, as predicted by structural modeling and differences in cellular dynamics of the two allotypes.

Our updated phylogenetic analysis reveals that after the ancestral gene encoding CHC duplicated in vertebrates, *CLTC* was retained in all vertebrate species, consistent with the encoded CHC17 mediating the cellular housekeeping clathrin functions shared by all eukaryotes. Species lacking the paralogue *CLTCL1* still have the insulin-responsive GLUT4 pathway in which the encoded CHC17 functions, but the presence of CHC22 improves the efficiency of this pathway. Data presented here (Fig. 10A) and our recent mapping of the intracellular pathway mediated by CHC22 (33) indicate that CHC22 clathrin enhances an existing transport route from the secretory pathway to the insulin-responsive GSC. This CHC22 pathway complements the endocytic pathway for GLUT4 targeting to the GSC, so species without CHC22 can rely primarily on endocytosis for GLUT4 trafficking to the GSC, while species with CHC22 use both pathways. Thus, we hypothesize that species with functional CHC22 clathrin are better at intracellular GLUT4 sequestration, resulting in lower

surface GLUT4 in the absence of insulin, and tighter regulation of GLUT4 release in response to insulin. The trade-off is that these species have an inherent increased tendency to insulin resistance as their GLUT4 is sequestered more effectively. The presence of CHC22 in humans would have been favorable for the development of our large brain, preserving blood glucose levels for brain function during starvation (37). The two main vertebrate branches that have lost CHC22 include *Muridae* (mice and rats) who are incessant herbivores and *Cetartiodactyla* (sheep, cattle, porpoise and pigs) which include the lineage of ruminants (sheep and cattle) whose muscle uptake of glucose is critical for muscle function but not a main source of glucose clearance (38). Both groups of species require greater availability of GLUT4 on their cell surfaces, so more efficient GLUT4 sequestration by CHC22 would not be favorable to their nutritional needs. The fact that CHC22 alters the balance of membrane traffic to the GSC means that species losing *CLTCL1* could evolve compensatory pathways more compatible with their diets. Thus, transgenic mice expressing CHC22 over-sequester GLUT4 in their muscle and develop high blood glucose with aging (4). The blind cave fish, which also apparently lacks *CLTCL1,* has independently evolved mutations in the insulin receptor, another pathway to insulin resistance, such that the presence of CHC22 on top of this mechanism might be detrimental (39). The loss of *CLTCL1* is consistent with the insulin responsive GLUT4 pathway being a target for natural selection driven by diet and might explain the loss from other vertebrate species during vertebrate evolution.

The allelic variation reported here for *CLTCL1* in human and bear populations further supports the hypothesis that *CLTCL1* has undergone continued selection during vertebrate evolution in relation to diet. While purifying selection appears to be operating on *CLTCL1* in the species that have retained it, *CLTCL1* is still far more variable than *CLTC* in the same species. In humans, we find support for balancing selection maintaining two major and functionally distinct alleles at remarkably similar frequencies in all population samples studied. The SNP distinguishing these alleles is human-specific and likely arose 550-50 KYA

(post-Neanderthal, pre-Neolithic). Statistical analysis comparing modern farming populations with modern hunter-gatherers shows an apparent increase of the V1316 variant, suggesting a correlation with regular consumption of digestible carbohydrate. The fact that the two alleles are functionally distinct is consistent with diversifying selection operating on *CLTCL1*. The allelic diversity of *CLTCL1* in other primate species could have the potential effect of diluting its function. Whilst chimpanzees are omnivores and gorillas herbivores, both rely on extensive foraging for carbohydrate for nutrition. It is also notable that polar bears, which have very low carbohydrate diets compared to their brown bear relatives, have distinct allelic variants of CHC22 with unknown functionality, again consistent with *CLTCL1* undergoing selection driven by nutritional ecology.

Clathrins are self-assembling proteins and function as a latticed network in the protein coat that they form on transport vesicles. Our structural modelling predicts that the single amino acid difference between the two main human CHC22 allotypes could influence the strength of molecular interactions in the CHC22 clathrin lattice, as position 1316 occurs at a lattice assembly interface (Fig. 8). When expressed in cells, both CHC22 variants had the same overall intracellular distribution, but CHC22-V1316 showed more rapid turnover from membranes than CHC22-M1316 and was less effective at GLUT4 sequestration (Fig. 8). These properties are consistent with the Met-to-Val change attenuating GLUT4 retention. This interpretation is supported by a GLUT4 translocation assay, which indicated that the V1316 variant is not as effective in GSC formation as the ancestral M1316 form of CHC22. Thus, mixed lattices that would occur in heterozygous individuals might reduce GLUT4 sequestration compared to M1316 homozygotes, which would have the effect of improving glucose clearance. It could be argued that human consumption of digestible carbohydrate on a regular basis (37), requiring increased glucose clearance, might be the selective force driving this genetic adaptation. This is consistent with the increased frequency of the V1316 variant in farmers. It is also possible that some forms of polar bear CHC22 are super-active

at GLUT4 sequestration, providing a route to maintain high blood glucose, as occurs through other mutations in the cave fish (39).

Regulators of fundamental membrane traffic pathways have diversified through gene duplication in many species over the timespan of eukaryotic evolution. Retention and loss can, in some cases, be correlated with special requirements resulting from species differentiation, such as the extensive elaboration of genes in the secretory pathway of *Tetrahymena* (12,40)**.** The evolutionary history of *CLTCL1*, following vertebrate-specific gene duplication, suggests that differentiation of nutritional habits has shaped selection for the presence and absence of *CLTCL1* in some vertebrate species, and its diversification in humans and potentially other species. Though its highest expression is in muscle and adipose tissue, transient expression of CHC22 during human brain development has also been documented (41). This was noted in a study of a very rare null mutant of *CLTCL1* that caused loss of pain sensing in homozygotes and no symptoms for heterozygotes (41). Attenuated CHC22 function of the V1316 variant might lead to a spectrum of pain-sensing in humans but this is unlikely to be a strong selective force affecting reproductive success, whereas glucose homeostasis, as suggested by our analysis is more likely. The presence of the *CLTCL1* gene encoding CHC22 clathrin confers tight regulation on glucose uptake into muscle and fat tissue, which was likely beneficial to nutrition of the large human brain, and perhaps reproduction by influencing insulin resistance during pregnancy. However, in recent human history, readily available carbohydrate has increased our need to clear glucose from the blood, such that selection continues to act on *CLTCL1*. Our cell biology studies have demonstrated that CHC22 increases GLUT4 retention. While we would not expect the major *CLTCL1* polymorphism to influence the development of Type 2 diabetes (T2D), CHC22 accumulates on expanded GSC that forms in cases of insulin-resistant T2D (4), so its variation could potentially exacerbate insulin resistance to different degrees. The genetic diversity that we report here may reflect evolution towards reversing a human tendency to insulin resistance.

MATERIAL AND METHODS

Phylogenetics

Vertebrate genomes as well as genomes of *Drosophila melanogaster*, *Caenorhabditis elegans*, *Ciona intestinalis* and *Ciona savignyi* were downloaded from Ensembl (42), all accessed on the 23/04/2016 except for pig (14/12/17), marmoset and hagfish (both 09/08/18), excluding vertebrate species sequenced below five-fold genome coverage, *i.e.* with less than five reads per site on average. In addition, we downloaded the genomes of the elephant-shark (43), whale-shark (44), marmot (The Alpine Marmot Genome, BioProject PRJEB8272 on NCBI) and porpoise (45). All potential homologs for the human isoforms of *CLTC/CLTCL1*, *MTMR4/MTMR3*, and *CLTA/CLTB*, in the above genomes were retrieved via BLAST (46). An *e*-value threshold of 0.001 with additional constraints applied by default in InParanoid version 7 (47) were used (at least 50% of the sequences are covered by the alignment, and at least 25% of the residues are aligned). For *CLTC* and *CLTCL1* the polar bear (*Ursus maritimus*) (10), brown bear (*Ursus arctos*) (27) and black bear (*Ursus americanus*) CHC17 and CHC22 protein sequences were manually added. For *CLTCL1* only the elephant and horse sequences (XP_023397213.1 and XP_023502410.1 respectively) were manually added.

The sequences corresponding to the longest transcripts were aligned with MAFFT (48) and phylogenetic trees generated with PhyML (49). The last two steps were repeated after manually removing outlier sequences lying on long branches (*CLTC/CLTCL1*: ENSTNIP00000007811.1, ENSTGUP00000014952.1, XP_023397213.1, XP_023502410.1; *CLTA/CLTB*: ENSPSIP00000012669.1) and in the case of genomes not retrieved from Ensembl (therefore lacking the gene-to-transcript mapping) sequences most likely corresponding to alternative transcripts (XP_015350877.1, XP_007899998.1, XP_007899997.1, XP_007904368.1, XP_007904367.1, XP_020375861.1,

XP_020375865.1, XP_020375862.1, XP_020392037.1, XP_020375864.1, XP_020375859.1). Trees were manually reconciled based on the Ensembl species tree extended by elephant-, whale shark, brown-, black bear, porpoise and marmot with TreeGraph (50). Branch lengths were estimated based on the multiple sequence alignment (MSA) with PhyML fixing the manually reconciled topology, with options '-u' and '--constraint_file'. With this approach no support values for splits are calculated. The resulting trees were used as input to generate a new phylogeny-aware MSAs with PRANK (51). Branch lengths of the reconciled topologies were then re-estimated based on the MSA generated by PRANK.

To compute evolutionary rates, the sequences and subtrees corresponding to *CLTC* and *CLTCL1* clades after duplication (i.e. excluding non-vertebrates) were extracted and sequences from species without either *CLTC* or *CLTCL1* were removed. The same procedure was performed for *MTMR4/MTMR3* and *CLTA/CLTB*. A phylogeny-aware MSA was computed with PRANK on the remaining sequences, and the amino acid alignment was converted to a codon alignment with PAL2NAL (52). Finally, dN/dS ratios (i.e. the ratio of the rate of nonsynonymous substitutions to the rate of synonymous substitutions) were inferred based on the codon alignments with PAML (53) for the six proteins independently using the site model M7. Model M7 fits a Beta-distribution to the site rates by estimating the two Beta parameters shape and scale. Rates are estimated per site over the entire phylogeny, and therefore represent time averages.

Population genetics

Phased genotypes were obtained by querying Variant Call Format (VCF) files (54) from the 1000 Genomes Project database Phase 3 (19) for all available 2,504 samples. Only high-quality variants were retained using *vcflib* (https://github.com/vcflib/vcflib) with options "VT = SNP & QUAL > 1 & AC > 1 & DP > 5000". Missing genotypes were assigned to homozygotes for the reference alleles. Finally, only sites with a recorded annotated function

of being missense, nonsense, stop-loss or frame-shift for tested genes according to the UCSC Table Browser were retained (55) (tables snp150 and snp150CodingDbSnp). For each retained position, the reference sequence for chimpanzee from the UCSC Table Browser (55) (table snp150OrthoPt5Pa2Rm8) was initially used to infer the putative ancestral state. For ambiguous or multiallelic states in the chimpanzee sequence, the human reference base was used as an initial proxy for the ancestral state. The predicted functional impact of amino acid replacements was obtained by using Polyphen (20) and SIFT (21). Additional frequency information for a single mutation of interest in more than 50 human populations was retrieved from the HGDP CEPH Panel (56) from http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/. Genotype data for farmers and hunter-gatherer individuals were collected from the Simons Genome Diversity Project Dataset (57). Populations were merged based on their assigned geographical region with the following classification for hunter-gatherers: Africa (Biaka, Ju hoan North, Khomani San, Mbuti), Central Asia and Siberia (Aluet, Chukchi, Eskimo Chaplin, Eskimo Naukan, Eskimo Sireniki, Even, Itelman, Tlingit, Tubalar, Ulchi, Atayal), East and South Asia (Atayal, Kusunda). Farmer and hunter-gatherer allele frequencies were compared following a previously described approach (58). Briefly, we analytically computed the probability that the V allele is more frequent in farmers than in hunter gatherers while fully accounting for the uncertainty in the individual frequency estimates. V allele frequencies were inferred from allele counts of M and V in a Bayesian framework with a conjugate Beta uniform prior. We recorded maximum a posteriori estimates with 95% highest posterior density credible intervals computed with the *Smisc* R library, version 0.3.9. We collected further published ancient DNA data from western Eurasia and classified into three genetic grouping: hunter-gatherer (HG), early farmer (EF) and steppe (SP), using supervised ADMIXTURE (59) as previously described (60). These are genetic groups and not directly based on differences in material culture or subsistence, but importantly in the case of HG and EF, these genetic classifications correspond closely to hunter-gatherer and agricultural subsistence strategies (61-63). We then restricted to samples dated between 10,000 and 5,000 years before present (BP) that

were classified as either HG or EF, leading to a dataset of 119 HG and 316 EF of which 85 and 188 respectively had coverage at rs1061325. Frequencies for South-East Asians and ancient Eurasians were down sampled to ensure numerical stability.

High-coverage VCF files for 79 individuals from six species of great apes were retrieved (64). Data was filtered using *vcflib* on the combined data set with the options “QUAL > 32 & DP > 50 & DP < 7000 & FS < 27 & MQ > 25 & AC > 1”, similarly to the original manuscript describing this data set (64). To retrieve nonsynonymous changes, only variants where the translated proteins for each allele differ were retained. We finally phased the data and assigned individual haplotypes using *shapeit* v2.r837 with the options “-burn 50 -prune 20 -main 100 -window 0.5 -effective-size 20000”. Additional 110 genomes of chimpanzees and bonobos were analyzed (25,65). Data filtering, functional annotation and haplotype phasing were performed as described above.

Full genome VCF files for two high-coverage archaic humans, namely one Neandertal Altai (66) and one Denisova were retrieved (67). Low-quality sites were filtered out using *vcflib* with the options “QUAL > 1 & DP > 10”. A pseudo-reference sequence for each archaic human was constructed by replacing the heterozygous sites with the previously inferred human ancestral state. Sequencing data information for additional ancient human samples were obtained from previously published high-quality whole genome sequences (62,68-74). Genotype likelihoods were calculated using the standard GATK model (75). Median-joining network plots were generated in R using *pegas* package (76).

Several summary statistics were calculated on the inferred haplotypes to describe their levels of nucleotide diversity. Specifically, for each population separately, the number of polymorphisms (S), the number of segregating singletons (SS), Tajima’s D (TD) (77), Fu and Li’s D* (FLDs) and F* (FLFs) (78), the sum of square allele frequencies including the most common allele (H1) and excluding it (H2) and their ratio (79) were calculated. We also

computed genetic differentiation (F_{ST}) (80) between pairs of canonical reference populations, namely Yoruban (YRI), Europeans (CEU), and Han Chinese (CHB).

To assess whether the observed summary statistics are expected under neutral evolution, genes with a coding length approximately equal (+/- 5%) to the one observed for the tested gene, *CLTCL1*, were selected. For this analysis, the longest isoform for each gene, and its annotation was considered according to refGene table from the UCSC Genome Browser. We discarded genes on chromosome 6 and on sex chromosomes, as well as *CLTA*, *CLTB* and *CLTC*. This set was further reduced to ensure that the coding length of *CLTCL1* (5,343 bp) was the median value of the control distribution. A total of 631 control genes were retrieved. As summary statistics can be calculated only in case of genetic variability, genes showing no SNPs (S=0) within each population were disregarded. Therefore, the number of analyzed control genes varied for each population, with a median value of 475. For each summary statistic, the empirical percentile rank for the value observed in *CLTCL1* compared to the whole distribution of control genes was calculated. Very low (e.g. < 0.01) or high (e.g. > 0.99) values are suggestive of *CLTCL1* being an outlier in the empirical distribution. For plotting purposes, summary statistics and populations were clustered according to a dendrogram inferred from their respective distances based on the calculated matrix of empirical percentile ranks. That is, populations clustering together exhibit similar patterns of percentile ranks, and thus of summary statistics. The underlying dendrograms are not reported. The heatmap plot was generated using the function heatmap.2 in R (81) with the package *gplots*. Cells with an empirical percentile rank lower than 0.1, 0.05 or 0.01 were denoted with a single, two or three asterisks, respectively.

Structure prediction by modelling

MODELLER v9.13 (28) was used to model the structure of the proximal leg segment of CHC22, using the crystal structure of bovine CHC17 (PDB 1B89) (82) as a template. The model of the M1316V mutant was derived in a similar way using a mutated sequence.

Structure visualization and analysis of residue interactions at the mutation site M1316 were performed using UCSF Chimera (29). The wild type and mutant homology models were positioned in the cryo-electron microscopy map of the bovine clathrin lattice (EMD: 5119) (83) by structural superposition on the atomic model originally fitted in the map (PDB:1XI4).

Functional experiments

Antibodies, plasmids and reagents

Mouse monoclonal anti-CHC17 antibodies X22 (84), TD.1 (85) and affinity-purified rabbit polyclonal antibody specific for CHC22 and not CHC17 (4) were produced in the Brodsky laboratory. Commercial sources of antibodies were as follows: mouse monoclonal anti-β-actin (clone AC-15, Sigma), mouse monoclonal anti-HA (clone 16B12, Covance), rabbit polyclonal anti-CHC22 (Proteintech). Secondary antibodies coupled to HRP were from ThermoFisher, the secondary antibody coupled to Brilliant Violet 421 was from BioLegend. The HA-GLUT4-mCherry was generated by replacing the GFP from the HA-GLUT4-GFP construct (gift from Dr Tim McGraw (86)) with mCherry using KpnI and EcoRI. The generation of the CHC22 variant expressing a valine at position 1316 (CHC22V) was previously described (87). The CHC22 variant expressing a methionine at position 1316 (CHC22M) was generated from CHC22V by quick-change mutagenesis (New England Biotechnologies, USA) following manufacturer's instructions.

Small RNA interference

Targeting siRNA was produced to interact with DNA sequences AAGCAATGAGCTGTTTGAAGA for CHC17 (87) (Qiagen), TCGGGCAAATGTGCCAAGCAA and AACTGGGAGGATCTAGTTAAA for CHC22 (1:1 mixture of siRNAs were used) (4) (Dharmacon). Non-targeting control siRNA was the Allstars Negative Control siRNA (Qiagen).

Cell culture

HeLa cells were grown in Dulbecco's Modified Eagle Medium high glucose (Gibco) supplemented with 10% FBS (Gibco), 50 U/mL penicillin, 50 µg/mL streptomycin (Gibco), 10 mM Hepes (Gibco) and maintained at 37°C in a 5% CO_2 atmosphere. HeLa cells were free of mycoplasma infection.

siRNA and DNA transfection

Cells were transfected for 72 hours with 20 nM of siRNA. Silencing was assessed by immunoblotting. Transient DNA transfections for rescue experiments were performed during the third day of silencing. For FACS experiments, cells (per well of 6-well plate, 70% confluent) were transiently transfected with 1 µg DNA for CHC22M-GFP and CHC22V-GFP, 1.5 µg DNA for CHC17-GFP and HA-GLUT4-mCherry. For FRAP experiments, cells (per glass bottom dish, 60% confluent) were transfected with 0.75 µg DNA for CHC22-GFP (M or V) or 1.5 µg DNA for CHC17-GFP. FACS and FRAP experiments were carried out 24 hours later. All transfections were performed using JetPrime transfection reagent (PolyPlus) following manufacturer's instructions.

GLUT4 translocation assay using flow cytometry

HeLa cells were grown in 6-well plates and transiently transfected with either HA-GLUT4-mCherry alone or in combination with GFP-tagged CHC22 (M or V) or CHC17-GFP the day before the experiment. The next day, cells were serum-starved (2 hours) before insulin stimulation (170 nM or vehicle (water) for 15 minutes, 37°C). Cells were then placed on ice and rapidly washed (2X, PBS, 4°C) and fixed (PFA 2%, 15 min). After fixation, cells were washed (2X, PBS, RT) then blocked for 1 hour (PBS 2% BSA, RT) before incubation with monoclonal anti-HA antibody (45 min, RT) to detect surface GLUT4. After incubation, cells were washed (3X, PBS, RT) and incubated with anti-mouse secondary Ig coupled to Brilliant Violet 421 (45 min, RT). Cells were then washed (5X, PBS, RT), gently lifted using a cell scraper (Corning), pelleted (800xg, 8 min) and re-suspended (PBS, 2% BSA, 4°C). Data was acquired with Diva acquisition software by LSRII flow cytometer (Becton Dickinson).

Typically, 10,000 to 30,000 events were recorded and Mean Fluorescence Intensity (MFI) values for surface GLUT4 (Brilliant Violet 421) and total GLUT4 (mCherry) were recorded using 450/50 and 530/30 filters, respectively. The ratio of surface to total MFI was calculated to quantify the extent of GLUT4 translocation. Mean Fluorescence Intensity values for total GLUT4 (mCherry) were plotted for GLUT4 stability assays. Histograms and post-acquisition analysis were performed using FlowJo software (Treestar). Total GLUT4 and surface GLUT4 values are reported separately for cells expressing CHC22 variants with equalized GFP signals at the top third (high), middle third (medium) and bottom third (low) levels of expression.

Fluorescence Recovery After Photobleaching

The imaging of transiently transfected HeLa cells grown on Cellview glass bottom culture dishes (Greiner, Germany) was performed at 37°C in a 5% CO_2 atmosphere, using low 488 nm laser power to minimize photobleaching using a 63x (1.4 NA) lens on a Leica SP8 confocal microscope. A 2.0 μm^2 circular region of interest was positioned in the perinuclear region of the transfected cells, a region where both CHC22 and CHC17 naturally occupy. A 100% laser power (488 nm) coupled to 11 iterations was performed to achieve GFP photobleaching. Recovery of fluorescence was recorded from 4 to 10 independent cells per dish. The experiment was repeated at least three times.

Immunoblotting

HeLa cells protein extracts were quantified by BCA (Pierce), separated by SDS-PAGE (10% acrylamide), transferred to nitrocellulose membrane (0.2 µm, Biorad), labelled with primary antibodies (1-5 µg/mL), washed and labelled with species-specific horseradish peroxidase-conjugated secondary antibodies (ThermoFisher). Peroxidase activity was detected using Western Lightning Chemiluminescence Reagent (GE Healthcare). The molecular migration position of transferred proteins was compared to the PageRuler Prestain Protein Ladder 10

to 170 kDa (Thermo Fisher Scientific). Signals were detected using the Chemidoc XRS+ imaging system (Biorad) and quantifications were performed using Image J software (NIH).

Statistical analyses

Graphs and statistical analyses were performed using Prism software (Graphpad). Detailed statistical information including statistical test used, number of independent experiments, *p* values, definition of error bars is listed in individual figure legends. All experiments were performed at least three times.

Acknowledgments: This work was supported by grants from the National Institutes of Health (USA) DK095663 (FMB and MF) and AI090905 (PP). We are grateful to Libby Guethlein for helpful discussions on genomic data retrieval and processing.

References

(1) Brodsky FM. Diversity of clathrin function: new tricks for an old protein. Annu Rev Cell Dev Biol 2012;28.

(2) Wakeham DE, Abi-Rached L, Towler MC, Wilbur JD, Parham P, Brodsky FM. Clathrin heavy and light chain isoforms originated by independent mechanisms of gene duplication during chordate evolution. Proc Natl Acad Sci U S A 2005 May 17;102(20):7209-7214.

(3) Dannhauser PN, Camus SM, Sakamoto K, Sadacca LA, Torres JA, Camus MD, et al. CHC22 and CHC17 clathrins have distinct biochemical properties and display differential regulation and function. J Biol Chem 2017 Dec 22;292(51):20834-20844.

(4) Vassilopoulos S, Esk C, Hoshino S, Funke BH, Chen CY, Plocik AM, et al. A role for the CHC22 clathrin heavy-chain isoform in human glucose metabolism. Science 2009 May 29;324(5931):1192-1196.

(5) Shepherd PR, Kahn BB. Glucose transporters and insulin action—implications for insulin resistance and diabetes mellitus. N Engl J Med 1999;341(4):248-257.

(6) Liu SH, Towler MC, Chen E, Chen CY, Song W, Apodaca G, et al. A novel clathrin homolog that co-distributes with cytoskeletal components functions in the trans-Golgi network. EMBO J 2001 Jan 15;20(1-2):272-284.

(7) Babbitt CC, Warner LR, Fedrigo O, Wall CE, Wray GA. Genomic signatures of diet-related shifts during human origins. Proc Biol Sci 2011 Apr 7;278(1708):961-969.

(8) Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS genetics 2011;7(11):e1002355.

(9) Elde NC, Malik HS. The evolutionary conundrum of pathogen mimicry. Nature Reviews Microbiology 2009;7(11):787-797.

(10) Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, Xiong Z, et al. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell 2014;157(4):785-794.

(11) Rout MP, Field MC. The Evolution of Organellar Coat Complexes and Organization of the Eukaryotic Cell. Annu Rev Biochem 2017(0).

(12) Dacks JB, Robinson MS. Outerwear through the ages: evolutionary cell biology of vesicle coats. Curr Opin Cell Biol 2017;47:108-116.

(13) Diekmann Y, Seixas E, Gouw M, Tavares-Cadete F, Seabra MC, Pereira-Leal JB. Thousands of rab GTPases for the cell biologist. PLoS computational biology 2011;7(10):e1002217.

(14) Guerrier S, Plattner H, Richardson E, Dacks JB, Turkewitz AP. An evolutionary balance: conservation vs innovation in ciliate membrane trafficking. Traffic 2017;18(1):18-28.

(15) Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. Tree of life reveals clock-like speciation and diversification. Mol Biol Evol 2015;32(4):835-845.

(16) dos Reis M, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PC, Yang Z. Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. Current Biology 2015;25(22):2939-2950.

(17) Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics 2011 Aug;188(4):799-808.

(18) Towler MC, Kaufman SJ, Brodsky FM. Membrane traffic in skeletal muscle. Traffic 2004;5(3):129-139.

(19) 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015;526(7571):68-74.

(20) Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods 2010;7(4):248-249.

(21) Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 2009;4(7):1073.

(22) Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. PLoS genetics 2006;2(4):e64.

(23) Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends in ecology & evolution 2013;28(11):659-669.

(24) Watterson G. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 1975;7(2):256-276.

(25) Teixeira JC, de Filippo C, Weihmann A, Meneu JR, Racimo F, Dannemann M, et al. Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees, and bonobos. Mol Biol Evol 2015;32(5):1186-1196.

(26) Bojarska K, Selva N. Spatial patterns in brown bear Ursus arctos diet: the role of geographical and environmental factors. Mamm Rev 2012;42(2):120-143.

(27) Benazzo A, Trucchi E, Cahill JA, Maisano Delser P, Mona S, Fumagalli M, et al. Survival and divergence in a small group: The extraordinary genomic history of the endangered Apennine brown bear stragglers. Proc Natl Acad Sci U S A 2017 Nov 7;114(45):E9589-E9597.

(28) Benjamin W, Sali A. Comparative protein structure modeling using. Modeller.Curr.Protoc.bioinfo. 2014;47:5-6.

(29) Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. Journal of computational chemistry 2004;25(13):1605-1612.

(30) Wilbur JD, Hwang PK, Brodsky FM. New faces of the familiar clathrin lattice. Traffic 2005;6(4):346-350.

(31) Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK, et al. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature 2013;500(7461):207.

(32) Landry JJ, Pyl PT, Rausch T, Zichner T, Tekkedil MM, Stutz AM, et al. The genomic and transcriptomic landscape of a HeLa cell line. G3 (Bethesda) 2013 Aug 7;3(8):1213-1224.

(33) Camus SM, Camus MD, Sadacca LA, Esk C, Gould GW, Kioumourtzoglou D, et al. CHC22 Clathrin Diverts GLUT4 from the ER-to-Golgi Intermediate Compartment for Intracellular Sequestration. bioRxiv 2018:242941.

(34) Trefely S, Khoo PS, Krycer JR, Chaudhuri R, Fazakerley DJ, Parker BL, et al. Kinome Screen Identifies PFKFB3 and Glucose Metabolism as Important Regulators of the Insulin/Insulin-like Growth Factor (IGF)-1 Signaling Pathway. J Biol Chem 2015 Oct 23;290(43):25834-25846.

(35) Haga Y, Ishii K, Suzuki T. N-glycosylation is critical for the stability and intracellular trafficking of glucose transporter GLUT4. J Biol Chem 2011 Sep 9;286(36):31320-31327.

(36) Bryant NJ, Govers R, James DE. Regulated transport of the glucose transporter GLUT4. Nature reviews Molecular cell biology 2002;3(4):267-277.

(37) Hardy K, Brand-Miller J, Brown KD, Thomas MG, Copeland L. The importance of dietary carbohydrate in human evolution. Q Rev Biol 2015;90(3):251-268.

(38) Hocquette JF, Bornes F, Balage M, Ferre P, Grizard J, Vermorel M. Glucose-transporter (GLUT4) protein content in oxidative and glycolytic skeletal muscles from calf and goat. Biochem J 1995 Jan 15;305 (Pt 2)(Pt 2):465-470.

(39) Riddle MR, Aspiras AC, Gaudenz K, Peuß R, Sung JY, Martineau B, et al. Insulin resistance in cavefish as an adaptation to a nutrient-limited environment. Nature 2018.

(40) Bright LJ, Kambesis N, Nelson SB, Jeong B, Turkewitz AP. Comprehensive analysis reveals dynamic and evolutionary plasticity of Rab GTPases and membrane traffic in Tetrahymena thermophila. PLoS genetics 2010;6(10):e1001155.

(41) Nahorski MS, Al-Gazali L, Hertecant J, Owen DJ, Borner GH, Chen Y, et al. A novel disorder reveals clathrin heavy chain-22 is essential for human pain and touch development. Brain 2015;138(8):2147-2160.

(42) Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res 2016 Jan 4;44(D1):D710-6.

(43) Venkatesh B, Lee AP, Ravi V, Maurya AK, Lian MM, Swann JB, et al. Elephant shark genome provides unique insights into gnathostome evolution. Nature 2014;505(7482):174.

(44) Read TD, Petit RA, Joseph SJ, Alam MT, Weil MR, Ahmad M, et al. Draft sequencing and assembly of the genome of the world's largest fish, the whale shark: Rhincodon typus Smith 1828. BMC Genomics 2017;18(1):532.

(45) Yuan Y, Zhang P, Wang K, Liu M, Li J, Zheng J, et al. Genome Sequence of the Freshwater Yangtze Finless Porpoise. Genes 2018;9(4):213.

(46) Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, et al. BLAST: a more efficient report with usability improvements. Nucleic Acids Res 2013;41(W1):W29-W33.

(47) Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 2009;38(suppl_1):D196-D203.

(48) Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30(4):772-780.

(49) Guindon S, Dufayard J, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010;59(3):307-321.

(50) Stöver BC, Müller KF. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 2010;11(1):7.

(51) Loytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A 2005 Jul 26;102(30):10557-10562.

(52) Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 2006;34(suppl_2):W609-W612.

(53) Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;24(8):1586-1591.

(54) Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011;27(15):2156-2158.

(55) Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, et al. The UCSC genome browser database: 2016 update. Nucleic Acids Res 2015;44(D1):D717-D725.

(56) Cann HM, De Toma C, Cazes L, Legrand M, Morel V, Piouffre L, et al. A human genome diversity cell line panel. Science 2002;296(5566):261-262.

(57) Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 2016;538(7624):201-206.

(58) Raineri E, Dabad M, Heath S. A note on exact differences between beta distributions in genomic (Methylation) studies. PLoS One 2014;9(5):e97349.

(59) Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009 Sep;19(9):1655-1664.

(60) Mathieson S, Mathieson I. FADS1 and the timing of human adaptation to agriculture. bioRxiv 2018:337998.

(61) Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 2015;522(7555):207.

(62) Skoglund P, Malmstrom H, Omrak A, Raghavan M, Valdiosera C, Gunther T, et al. Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. Science 2014 May 16;344(6185):747-750.

(63) Skoglund P, Malmstrom H, Raghavan M, Stora J, Hall P, Willerslev E, et al. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 2012 Apr 27;336(6080):466-469.

(64) Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature 2013;499(7459):471-475.

(65) De Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science 2016;354(6311):477-481.

(66) Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 2014;505(7481):43-49.

(67) Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science 2012 Oct 12;338(6104):222-226.

(68) Broushaki F, Thomas MG, Link V, Lopez S, van Dorp L, Kirsanow K, et al. Early Neolithic genomes from the eastern Fertile Crescent. Science 2016 Jul 29;353(6298):499-503.

(69) Hofmanova Z, Kreutzer S, Hellenthal G, Sell C, Diekmann Y, Diez-Del-Molino D, et al. Early farmers from across Europe directly descended from Neolithic Aegeans. Proc Natl Acad Sci U S A 2016 Jun 21;113(25):6886-6891.

(70) Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 2014;513(7518):409-413.

(71) Olalde I, Allentoft ME, Sánchez-Quinto F, Santpere G, Chiang CW, DeGiorgio M, et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. Nature 2014;507(7491):225-228.

(72) Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 2014;505(7481):87-91.

(73) Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas AS, Manica A, Moltke I, et al. Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. Science 2014 Nov 28;346(6213):1113-1118.

(74) Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 2014;514(7523):445-449.

(75) McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010 Sep;20(9):1297-1303.

(76) Paradis E. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics 2010;26(3):419-420.

(77) Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989 Nov;123(3):585-595.

(78) Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics 1993 Mar;133(3):693-709.

(79) Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS genetics 2015;11(2):e1005004.

(80) Reynolds J, Weir BS, Cockerham CC. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 1983 Nov;105(3):767-779.

(81) Team RC. R: A language and environment for statistical computing.Vienna, Austria: R Foundation for Statistical Computing; 2014 2014.

(82) Ybe JA, Brodsky FM, Hofmann K, Lin K, Liu S, Chen L, et al. Clathrin self-assembly is mediated by a tandemly repeated superhelix. Nature 1999;399(6734):371-375.

(83) Fotin A, Cheng Y, Sliz P, Grigorieff N, Harrison SC, Kirchhausen T, et al. Molecular model for a complete clathrin lattice from electron cryomicroscopy. Nature 2004;432(7017):573-579.

(84) Brodsky FM. Clathrin structure characterized with monoclonal antibodies. I. Analysis of multiple antigenic sites. J Cell Biol 1985 Dec;101(6):2047-2054.

(85) Näthke IS, Heuser J, Lupas A, Stock J, Turck CW, Brodsky FM. Folding and trimerization of clathrin subunits at the triskelion hub. Cell 1992;68(5):899-910.

(86) Lampson MA, Racz A, Cushman SW, McGraw TE. Demonstration of insulin-responsive trafficking of GLUT4 and vpTR in fibroblasts. J Cell Sci 2000 Nov;113 (Pt 22)(Pt 22):4065-4076.

(87) Esk C, Chen CY, Johannes L, Brodsky FM. The clathrin heavy chain isoform CHC22 functions in a novel endosomal sorting step. J Cell Biol 2010 Jan 11;188(1):131-144.

(88) Lippincott-Schwartz J, Snapp E, Kenworthy A. Studying protein dynamics in living cells. Nature reviews Molecular cell biology 2001;2(6):444.

LIST OF ITEMS

Figure legends

Figure 1. Phylogenetic profile *of CLTC/CLTCL1* homologs, with gene presence in the corresponding genome indicated by a filled black circle. All sequences used have less than 5% unspecified residues (‘X’s). Divergent gene sequences with low sequence similarity but that still fall within the *CLTC* clade are shown as empty circles (see Materials and Methods

for similarity threshold and sequence IDs). Based on the profile and species tree the most parsimonious phylogenetic tree edge for loss and duplication events is inferred and shown as red stars and blue squares, respectively.

Figure 2. Evolutionary rates expressed as dN/dS ratios inferred for each position in *CLTC* (A) and *CLTCL1* (B). Rates are averages over an entire phylogenetic tree and therefore not specific to the human proteins. However, to assist interpretation, only rates for residues present in the human proteins are shown. Kernel density estimates of the distributions of dN/dS ratios per paralogous pair of proteins (C-E). Mean dN/dS ratios averaged over all sites are shown as hatched marks.

Figure 3. Median joining network of human alleles for *CLTC* (A) and *CLTCL1* (B). Each circle represents a unique allele whose global frequency is proportional to its circle's size. For *CLTC*, sizes were transformed in 10 times $\log_2$ scale for ease of visualization. For *CLTCL1*, only alleles with a frequency greater than 20 were plotted and the two main alleles were reduced in size by a factor of three for visualization purpose. The frequency within each macro-population is color-coded. Localization of the M1316V variant is depicted with a hashed line, with alleles carrying the M variant on the right-hand side, and alleles carrying the V variant on the left-hand side.

Figure 4. Matrix of summary statistics for *CLTLC1*. For each population (on the rows) we calculated several summary statistics (on the columns) and reported their percentile rank against their corresponding empirical distribution based on >500 control genes. The resulting matrix was then sorted on both axes as a dendrogram (not reported) based on the pairwise distances between each pair of populations. As depicted in the color legend, red and yellow denote low and high percentile ranks, respectively. Asterisks indicate levels of significance (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$). The overall distribution of rank percentiles is shown inside the color legend.

Figure 5. Worldwide distribution of heterozygosity of M1316V. A red triangle indicates an excess of heterozygosity (third tertile of the ratio between observed and expected heterozygosity), a blue circle a deficiency of heterozygosity (first tertile), and a grey square otherwise. The size of each symbol is proportional to the frequency of heterozygous genotypes.

Figure 6. Distribution of V1316 frequency in populations of hunter-gatherers and farmers. Maximum a posteriori estimates, and 95% highest posterior density credible intervals, of the frequency of V1316 are compared between hunter-gatherer and farmer populations in three continents. Probability of V allele being at a higher frequency in farmers, labelled as p(f>hg), is also reported.

Figure 7. Median joining network of chimpanzee and bonobo alleles for *CLTCL1*. Each circle represents a unique allele whose global frequency is proportional to its circle's size. Rare alleles were magnified by a factor of three for ease of visualization. The frequency within each species is color-coded.

Figure 8. Modelling of the structural changes in clathrin caused by the Met-Val dimorphism at residue 1316. Model of the CHC17 clathrin lattice (A) is reproduced with permission (83) with the region comprising residue 1316 boxed. Panel B (top part) is the magnified boxed region in A with a CHC22-M1316 model (residues 1210 to 1516) docked into one of the four clathrin heavy chains (CHC-1) forming the edge of the lattice. The black arrows show the location of the amino acid residue 1316 with the side chain highlighted in CHC-1. The density of the other three CHCs is indicated. Computational models of human CHC22 (residues 1210 to 1516) with either Met or Val at position 1316 (B, lower parts). The yellow circle encloses space opened by reducing the side chain size, which would require a shift in CHC torque to regain structurally favorable side chain contacts.

Figure 9. Turnover differences between the CHC22-M1316 and CHC22-V1316 allotypes measured by fluorescence recovery after photobleaching (FRAP).

HeLa cells were transfected with CHC22-M1316-GFP (CHC22M) or CHC22-V1316-GFP (CHC22V) or CHC17-GFP and the expressed constructs were localized relative to endogenous CHC22 and CHC17, which were also compared to each other (A). Endogenous CHC22, CHC17 and the transfected proteins were visualized by immunofluorescence with anti-CHC22 rabbit polyclonal antibody (pAb, red), anti-CHC17 mouse monoclonal antibody (mAb, green) and anti-GFP chicken polyclonal antibody (green for CHC17-GFP or red for CHC22-GFP), respectively. Bars represent 3 μm (untransfected and CHC22M-GFP) and 5 μm (CHC22V-GFP and CHC17-GFP). Transfectants were photobleached in the circular region indicated (B) and recovery of fluorescence (FRAP) was visualized over time (bars, 10 μm) and quantified within the bleached regions (C). For the data in (C), area under the curves (D) and mobile fractions Mf (E) were calculated (88). We performed a one-way analysis of variance (ANOVA) with Tukey's multiple comparison post-hoc test: *$p<0.05$, **$p<0.01$.

Figure 10. Differences in intracellular GLUT4 sequestration and stability in cells expressing CHC22-M1316 or CHC22-V1316 allotypes. HeLa cells were treated with siRNA to deplete endogenous CHC22 or with control siRNA, then transfected to co-express HA-GLUT4-mCherry along with CHC17-GFP (CHC17), CHC22-M1316-GFP (CHC22M) or CHC22-V1316-GFP (CHC22V) (A and B). Total levels of expressed GLUT4 and CHC22 were measured by FACS (mean fluorescence intensity (MFI) for mCherry or GFP, respectively). Surface levels of GLUT4 were measured with anti-HA antibody at basal conditions (-) or after 30 minutes of exposure to insulin (+) and surface/total GLUT4 is reported as measure of GLUT4 translocation to the cell surface (A) in cells expressing equivalent total levels of CHC-GFP. Transfected cells treated with siRNA to deplete endogenous CHC22, but not treated with insulin, were gated into thirds expressing equivalently low (L), medium (M) and

high (H) levels of CHC-GFP for each type of CHC, then total levels of HA-GLUT4-mCherry in each population were plotted (B). We performed a one-way analysis of variance (ANOVA) with Tukey's multiple comparison post-hoc test: *p<0.05.

SUPPLEMENTARY INFORMATION

Figure S1. Unreconciled phylogenetic tree for *CLTC* and *CLTCL1* across all investigated species.

Figure S2. Reconciled phylogenetic tree (therefore missing support values) for *CLTC* and *CLTCL1* across all investigated species.

Figure S3. Variation of H2 statistics along a genomic region surrounding *CLTCL1* in four European populations. Each window has a size of 20 kbp and step is 5 kbp.

Figure S4. Geographical distribution of M1316V alleles across human populations in the HGDP-CEPH panel data set. The ancestral allele T refers to amino acid M, while the derived nucleotide C refers to V.

Table S1. Human alleles for the coding region of *CLTLC1* extracted from the 1000 Genomes project data set. For each unique allele (hap_ID on the first column), the frequency in each macro-population (EUR, EAS, AMR, SAS, AFR), archaic humans and modern chimpanzees is reported, from column 2 to 9. From column 10 the nucleotide sequence for each unique allele is reported for all 46 retrieved SNPs, indexed on the remaining columns.

Table S2. Functional annotation for each coding polymorphism of *CLTCL1* reported in Table S1. Columns represent: chromosome, genomic position, SNP ID, reference allele, alternate allele, functional annotation for each isoform.

Table S3. Archaic and ancient human M1316V genotypes retrieved. Genotype likelihoods and sources for each sample are reported.

Table S4. Summary statistics of nucleotide diversity for *CLTLC1* calculated for all analyzed human populations. For each population on the rows, several summary statistics are reported on the columns. The notation follows the one reported in the main text and NSAM is the number of sampled chromosomes.

Table S5. List of control genes used to assess deviation from neutrality in the summary statistics of *CLTCL1*.

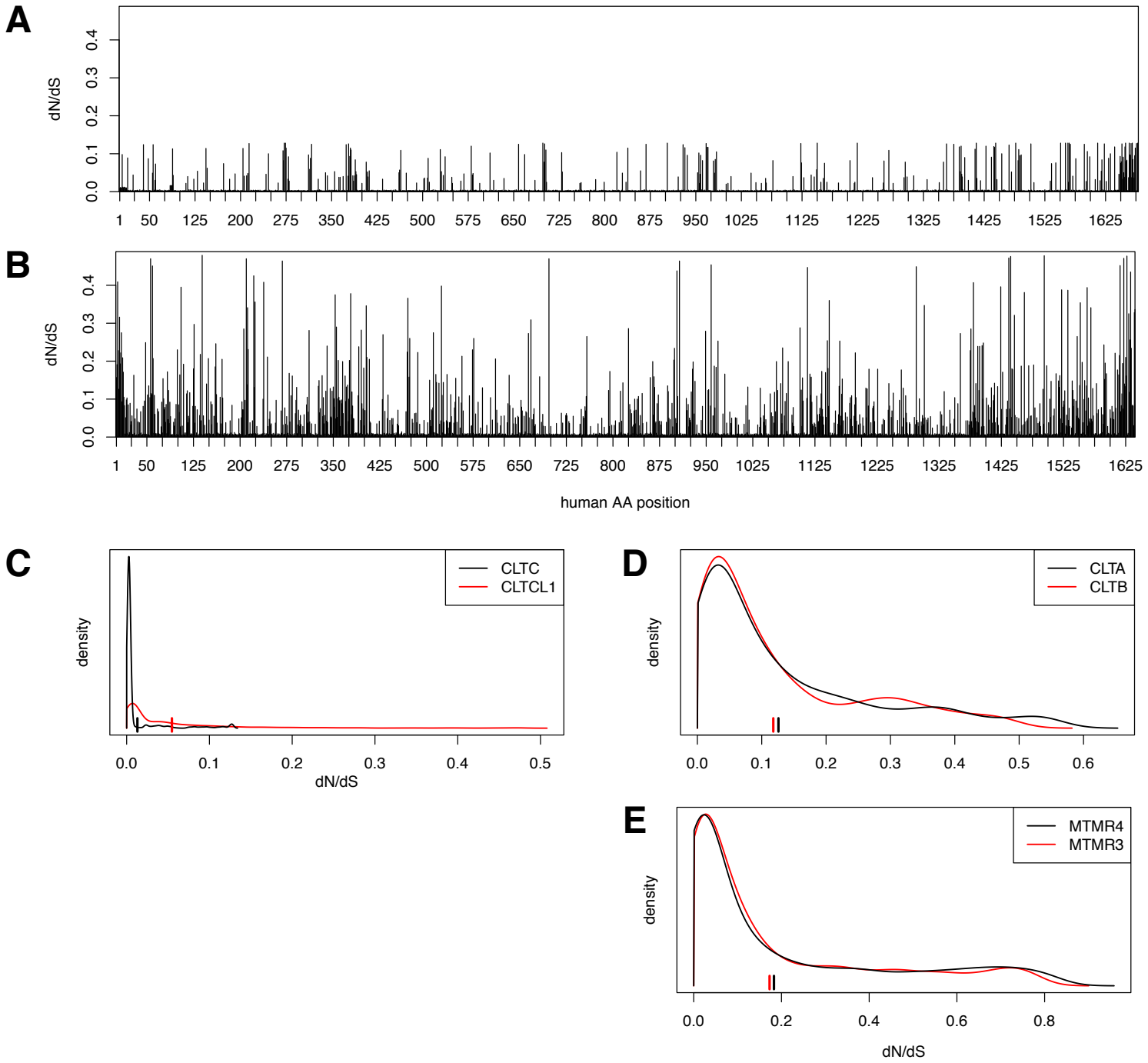
Table S6. Expected and observed heterozygosity for M1316V for all analyzed human populations. For each population (on rows), several metrics are reported included the frequency of observed homozygous (Obs.Homo1 and Obs.Homo2) and heterozygous genotypes (Obs.Hetero). Likewise, the corresponding expected values under Hardy-Weinberg equilibrium are reported (Exp.Homo1, Exp.Homo2, Exp.Hetero). Finally, the ratio between observed and expected heterozygosity is reported.

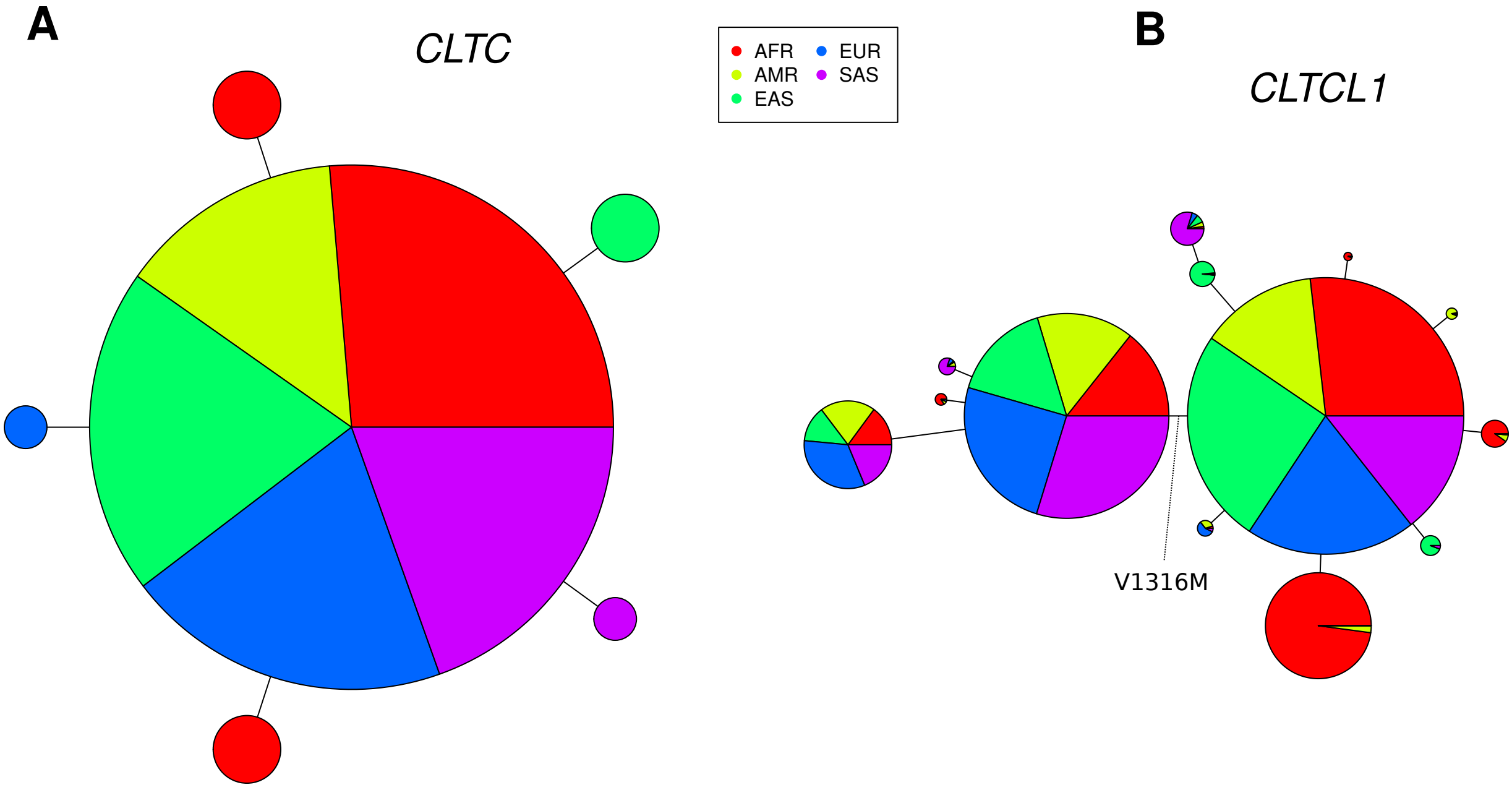
Table S7. Expected and observed heterozygosity for the two main alleles in *CLTCL1* (see Fig. 3B and Table S1) for all analyzed human populations. See legend for Table S6 for a description of its content.

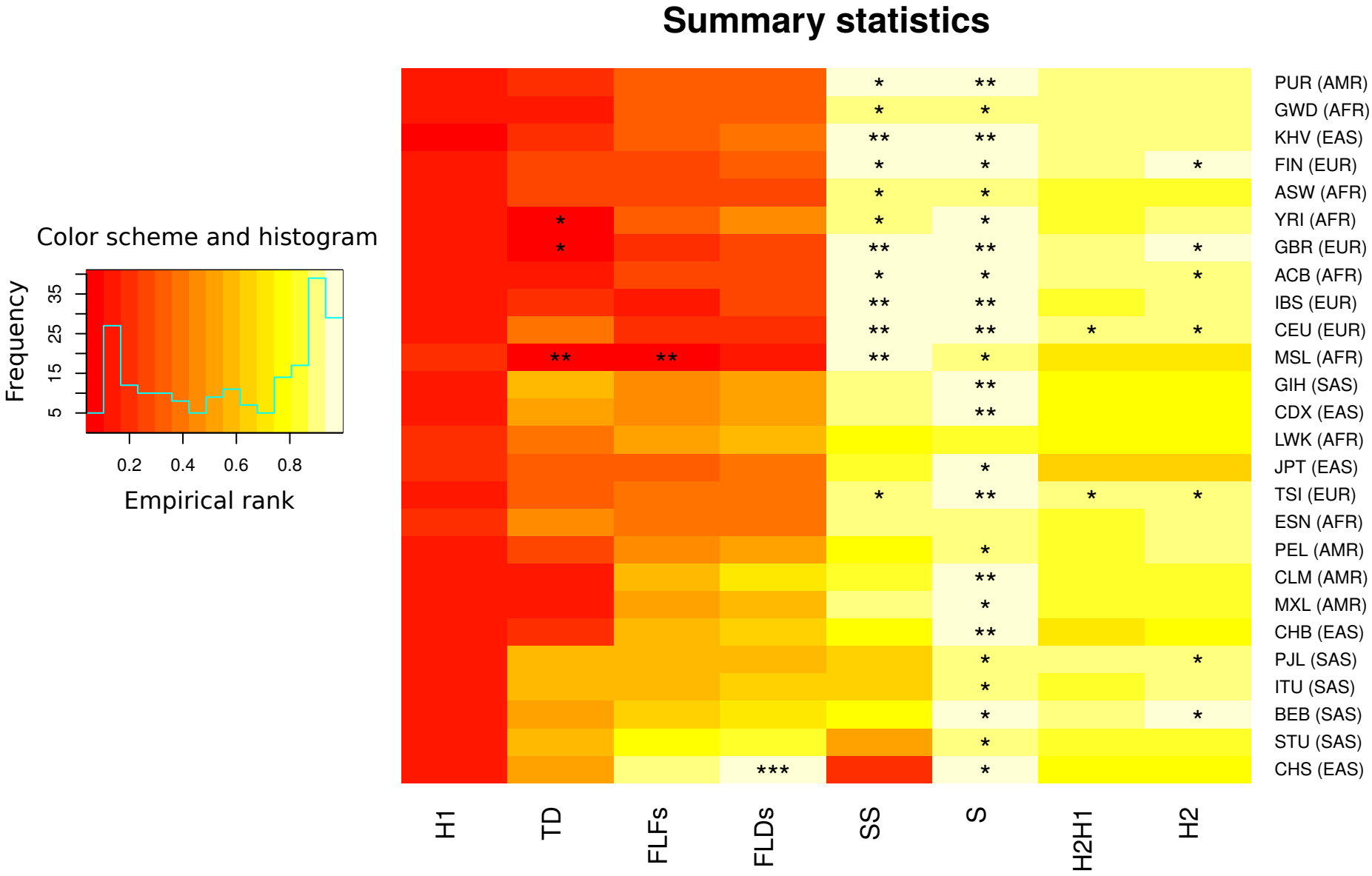
Table S8. Inferred *CLTCL1* alleles for great apes. For each unique allele (hap_ID on the first column), the frequency in each species is reported. From column 13 the amino acid sequence for each unique allele is reported for the whole protein.

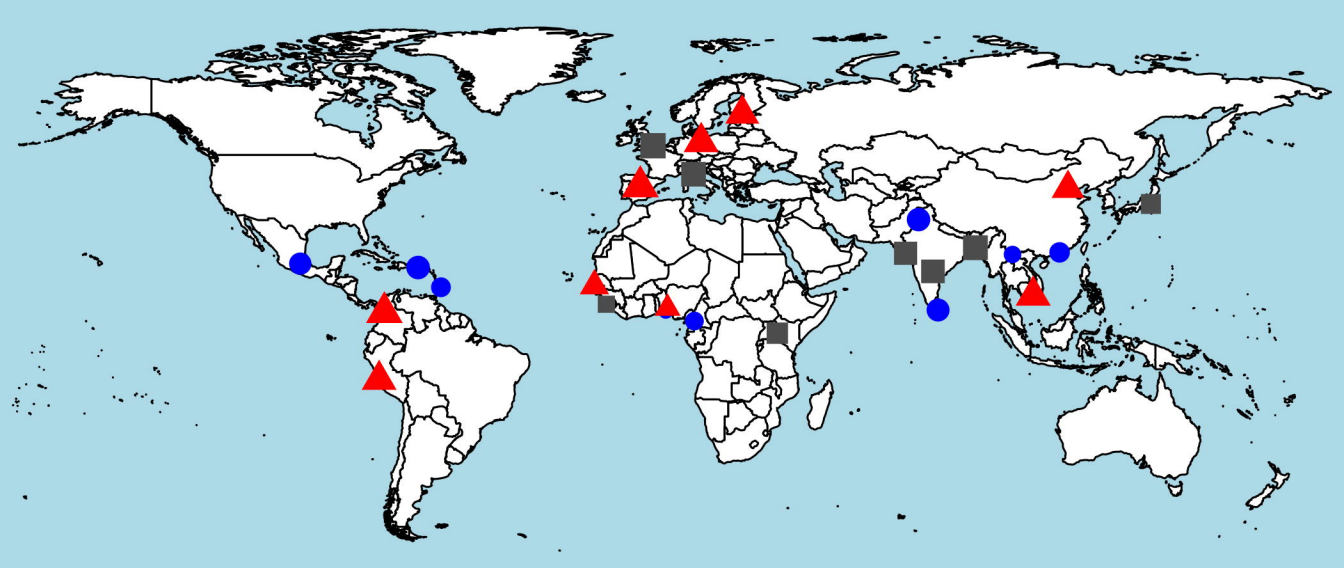
Table S9. Inferred *CLTCL1* alleles for chimpanzees and bonobos. For each unique allele (hap_ID on the first column), the frequency in each species is reported. From column 10 the amino acid sequence for each unique allele is reported for the whole protein.

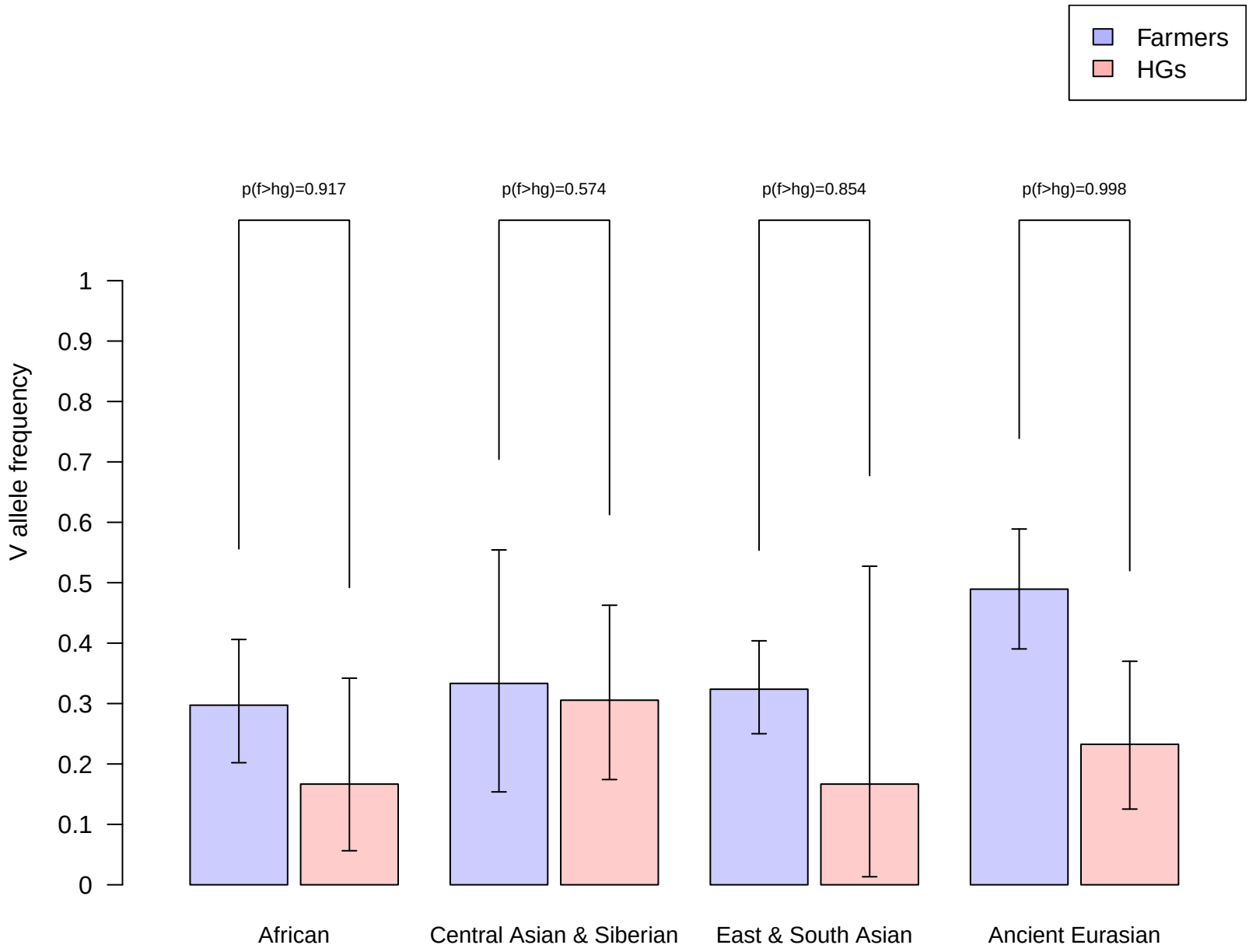
sequence
divergent sequence
loss event
duplication
Drosophila melanogaster
Caenorhabditis elegans
Ciona intestinalis
Ciona savignyi
Petromyzon marinus
Eptatretus burgeri
Callorhinchus milii
Rhincodon typus
Lepisosteus oculatus
Danio rerio
Astyanax mexicanus
Gadus morhua
Tetraodon nigroviridis
Takifugu rubripes
Oreochromis niloticus
Gasterosteus aculeatus
Oryzias latipes
Poecilia formosa
Xiphophorus maculatus
Latimeria chalumnae
Xenopus tropicalis
Anolis carolinensis
Pelodiscus sinensis
Taeniopygia guttata
Ficedula albicollis
Anas platyrhynchos
Meleagris gallopavo
Gallus gallus
Ornithorhynchus anatinus
Sarcophilus harrisii
Monodelphis domestica
Loxodonta africana
Dasypus novemcinctus
Sus scrofa
Neophocaena asiaeorientalis
Bos taurus
Ovis aries
Myotis lucifugus
Equus caballus
Felis catus
Canis familiaris
Mustela putorius furo
Ailuropoda melanoleuca
Ursus americanus
Ursus maritimus
Ursus arctos
Oryctolagus cuniculus
Cavia porcellus
Mus musculus
Rattus norvegicus
Ictidomys tridecemlineatus
Marmota marmota
Otolemur garnettii
Callithrix jacchus
Chlorocebus sabaeus
Papio anubis
Macaca mulatta
Nomascus leucogenys
Pongo abelii
Gorilla gorilla
Pan troglodytes
Homo sapiens
CLTC
CLTCL1

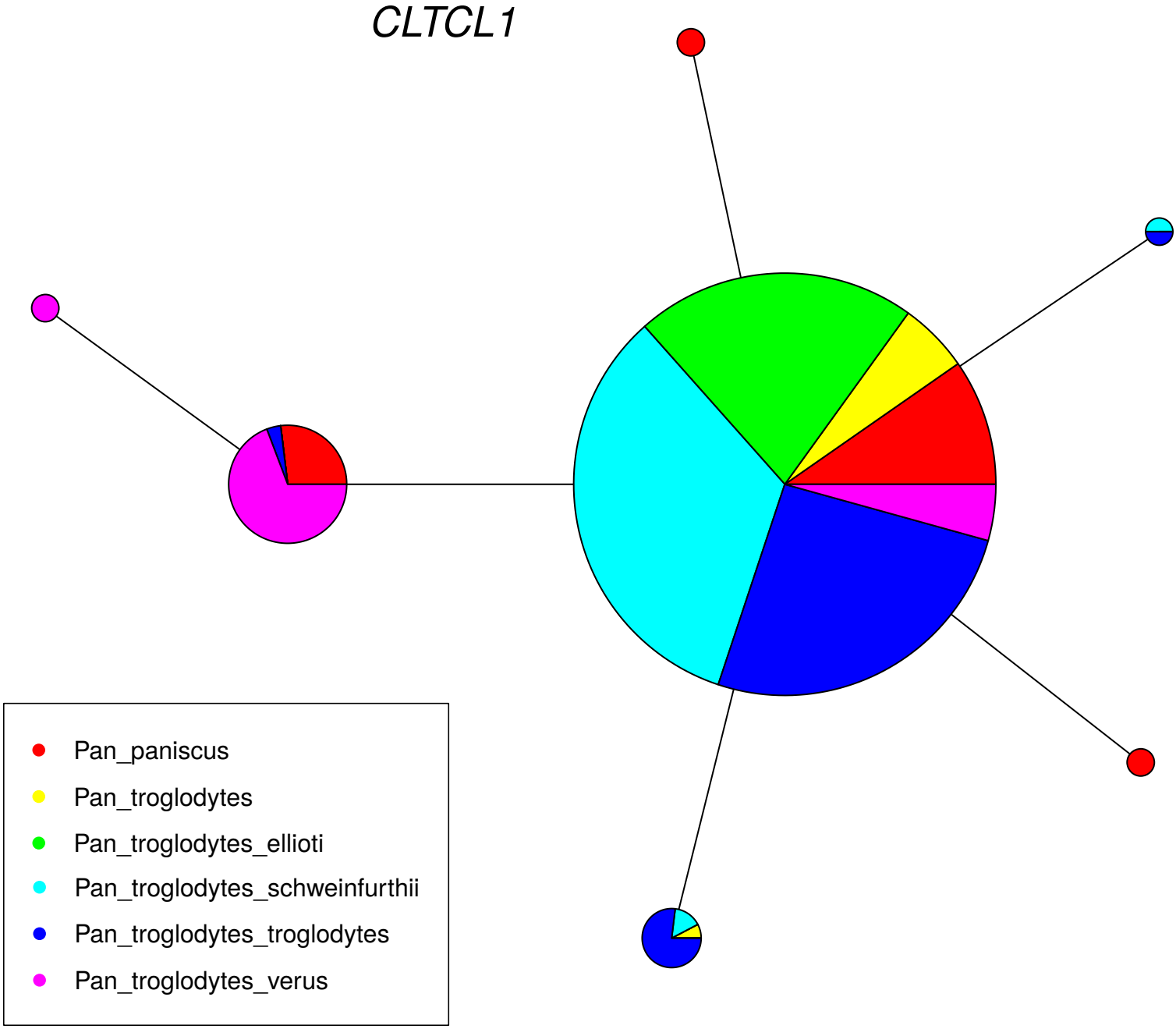

A
dN/dS
0.0
0.1
0.2
0.3
0.4
1 50 125 200 275 350 425 500 575 650 725 800 875 950 1025 1125 1225 1325 1425 1525 1625
B
dN/dS
0.0
0.1
0.2
0.3
0.4
1 50 125 200 275 350 425 500 575 650 725 800 875 950 1025 1125 1225 1325 1425 1525 1625
human AA position
C
CLTC
CLTCL1
density
0.0 0.1 0.2 0.3 0.4 0.5
dN/dS
D
CLTA
CLTB
density
0.0 0.1 0.2 0.3 0.4 0.5 0.6
E
MTMR4
MTMR3
density
0.0 0.2 0.4 0.6 0.8
dN/dS

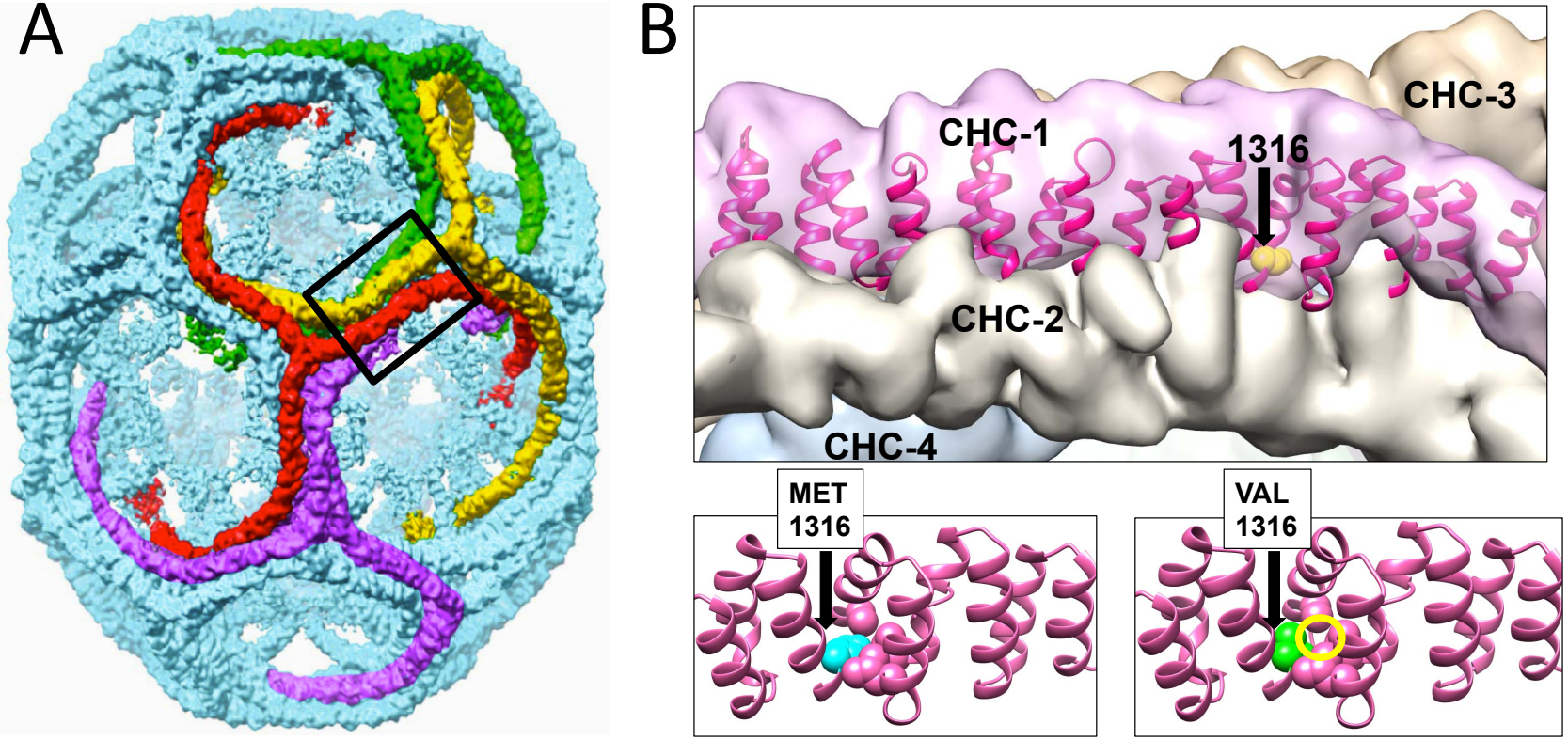

A
CLTC
AFR
EUR
AMR
SAS
EAS
B
CLTCL1
V1316M


Summary statistics
Color scheme and histogram
Frequency
5
15
25
35
0.2
0.4
0.6
0.8
Empirical rank
PUR (AMR)
GWD (AFR)
KHV (EAS)
FIN (EUR)
ASW (AFR)
YRI (AFR)
GBR (EUR)
ACB (AFR)
IBS (EUR)
CEU (EUR)
MSL (AFR)
GIH (SAS)
CDX (EAS)
LWK (AFR)
JPT (EAS)
TSI (EUR)
ESN (AFR)
PEL (AMR)
CLM (AMR)
MXL (AMR)
CHB (EAS)
PJL (SAS)
ITU (SAS)
BEB (SAS)
STU (SAS)
CHS (EAS)
H1
TD
FLFs
FLDs
SS
S
H2H1
H2

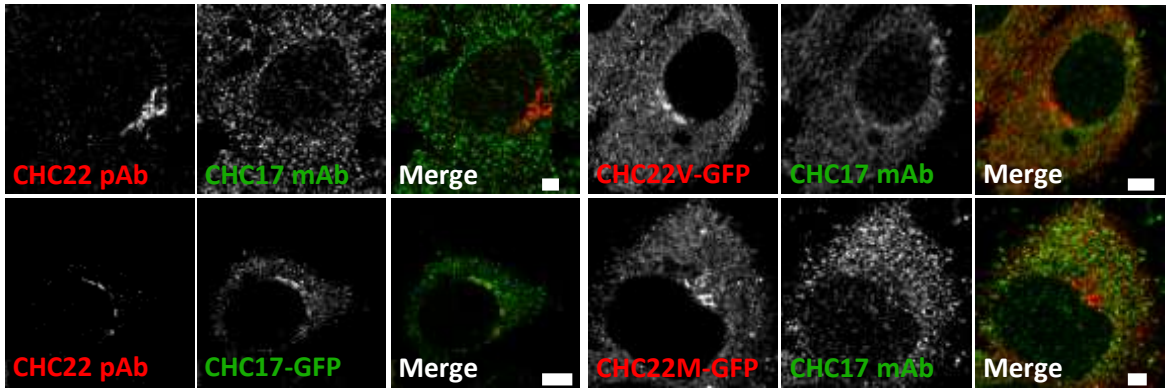



Farmers
HGs
p(f>hg)=0.917
p(f>hg)=0.574
p(f>hg)=0.854
p(f>hg)=0.998
V allele frequency
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
African
Central Asian & Siberian
East & South Asian
Ancient Eurasian

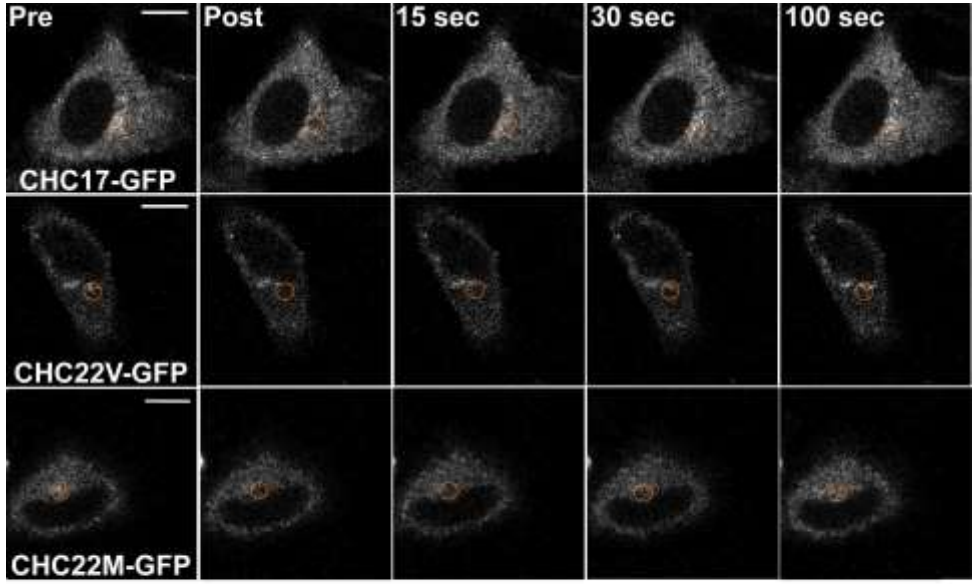

CLTCL1
Pan_paniscus
Pan_troglodytes
Pan_troglodytes_ellioti
Pan_troglodytes_schweinfurthii
Pan_troglodytes_troglodytes
Pan_troglodytes_verus


A
B
CHC-1
CHC-2
CHC-3
CHC-4
1316
MET
1316
VAL
1316

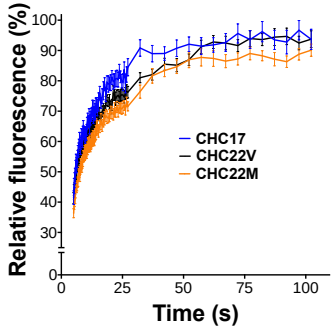
A


CHC22 pAb
CHC17 mAb
Merge
CHC22V-GFP
CHC17 mAb
Merge
CHC22 pAb
CHC17-GFP
Merge
CHC22M-GFP
CHC17 mAb
Merge

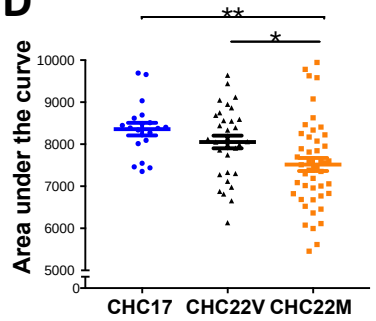
B


Pre
Post
15 sec
30 sec
100 sec
CHC17-GFP
CHC22V-GFP
CHC22M-GFP

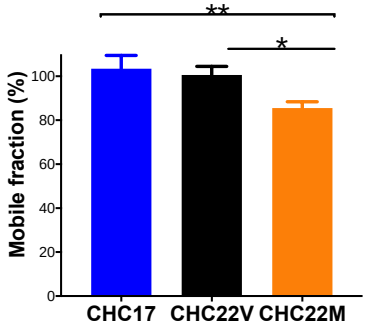
C

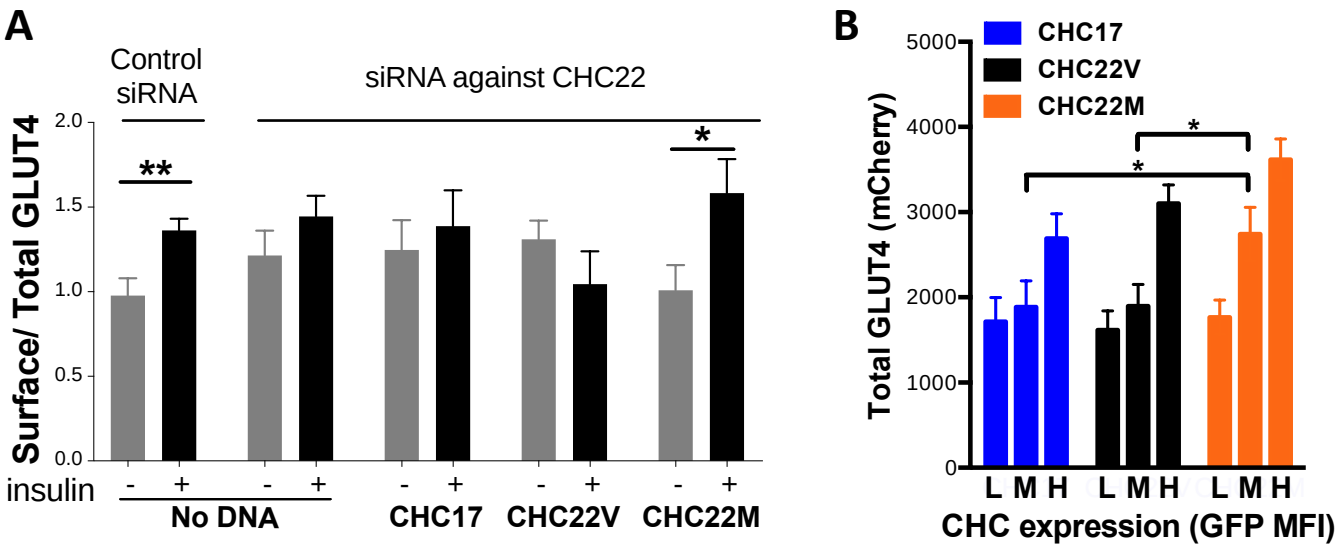

Relative fluorescence (%)
100
90
80
70
60
50
40
30
0
0
25
50
75
100
Time (s)
CHC17
CHC22V
CHC22M

D


**
*
Area under the curve
10000
9000
8000
7000
6000
5000
0
CHC17
CHC22V
CHC22M

E


**
*
Mobile fraction (%)
100
80
60
40
20
0
CHC17
CHC22V
CHC22M


A
Control siRNA
siRNA against CHC22
Surface/ Total GLUT4
2.0
1.5
1.0
0.5
0.0
**
*
insulin
- + - + - + - + - +
No DNA
CHC17
CHC22V
CHC22M
B
CHC17
CHC22V
CHC22M
Total GLUT4 (mCherry)
5000
4000
3000
2000
1000
0
*
*
L M H L M H L M H
CHC expression (GFP MFI)