preferences is potentially a general phenomenon in the Lepidoptera. What we show here for *Utetheisa*, and attribute to genetic architecture, might in fact apply beyond lepidopterans and birds to any group that exhibits male homogamety. The protected invasion theory, therefore, might be invoked broadly to account for taxonomic variation in the degree to which sexual dimorphism is manifested in nature. □

Methods

Utetheisa rearing and mating

Larvae were reared as previously described[23] on a pinto-bean-based diet supplemented with seeds of *Crotalaria spectabilis*, a natural food plant of *Utetheisa*. For any set of progeny, two groups of 8–10 larvae each were raised under identical conditions to adulthood. As in earlier studies[24], all matings involved presenting females with a choice of two males for 24 h in humidified cylindrical containers (0.35 l). For the present purposes, the males were related neither to the female nor to each other. Males and females were of known body mass at the time of mating, and the males were all 3-day-old virgins. Events in the mating enclosures were monitored on an ongoing basis for the first hour (to ensure that both males actively courted) and then at intervals of 6 h (to check on the occurrence of mating). A record was kept of the male that mated (the males were wing-marked for recognition purposes).

The experimental protocol required that individual females be presented daily with an unfamiliar pair of males until they mated a total of six times. The two males in the pair initially offered were chosen such that, in a randomly selected half of cases, they differed by 5% in body mass, and by 10% in the remaining cases. Subsequent presentations to any one female were dependent on whether the female mated. If she did not mate, she was next presented males showing the same mass difference; if she did, her next choice was between males of the alternative mass difference.

Mating Preference Index

For experimental purposes, we raised three generations of 44 families (derived from wild *Utetheisa* females caught in Highlands County, Florida, USA). Within each family, mating preferences were assessed for six full sisters, their mother and their paternal grandmother. The mating preference index (MPI) was calculated for each female as the average of her six mating choices. Individual choices were scored as 0 if the female favoured the smaller male, and as 1 if she chose the larger male. MPI values therefore fell within the range of 0 to 1.

Statistical analyses

To estimate heritability (in the narrow sense)[14], we regressed the average MPI value of six full sisters on the MPIs of their mother and paternal grandmother. A significant positive correlation was evident in the female–paternal-grandmother regression only, indicating that the mating preference is sex-linked rather than autosomal. For calculation of heritability, therefore, the slope of the regression was multiplied by two rather than four because females share half of their sex chromosomes, and only a quarter of their autosomes, with their paternal grandparents[14]. The data were not transformed because they met the assumptions of normality (normal probability plot) and linearity (Lowess test). Furthermore, there was no need to adjust regression coefficients or standard errors for unequal variances because the variances were the same for mother and paternal-grandmother MPI values (two-tailed variance ratio test, $F = 1.728$, $P = 0.08$). Heritabilities (regression slopes) were compared using ANCOVAs[14].

Received 24 June; accepted 11 July 2002; doi:10.1038/nature01027.

1. Iyengar, V. K. & Eisner, T. Female choice increases offspring fitness in an arctiid moth (*Utetheisa ornatrix*). *Proc. Natl Acad. Sci. USA* **96,** 15013–15016 (1999).
2. Reeve, H. K. Haplodiploidy, eusociality and absence of male parental and alloparental care in Hymenoptera: a unifying genetic hypothesis distinct from kin selection theory. *Phil. Trans. R. Soc. Lond. B* **342,** 335–352 (1993).
3. Reeve, H. K. & Shellman-Reeve, J. S. The general protected invasion theory: Sex biases in parental and alloparental care. *Evol. Ecol.* **11,** 357–370 (1997).
4. Conner, W. E., Roach, B., Benedict, E., Meinwald, J. & Eisner, T. Courtship pheromone production and body size as correlates of larval diet in males of the arctiid moth, *Utetheisa ornatrix*. *J. Chem. Ecol.* **16,** 543–552 (1990).
5. Dussourd, D. E., Harvis, C., Resch, J., Meinwald, J. & Eisner, T. Pheromonal advertisement of a nuptial gift by a male moth (*Utetheisa ornatrix*). *Proc. Natl Acad. Sci. USA* **88,** 9224–9227 (1991).
6. LaMunyon, C. W. & Eisner, T. Spermatophore size as determinant of paternity in an arctiid moth (*Utetheisa ornatrix*). *Proc. Natl Acad. Sci. USA* **91,** 7081–7084 (1994).
7. LaMunyon, C. W. Increased fecundity, as a function of multiple mating, in an arctiid moth, *Utetheisa ornatrix*. *Ecol. Entomol.* **22,** 69–73 (1997).
8. González, A., Rossini, C., Eisner, M. & Eisner, T. Sexually transmitted chemical defense in a moth (*Utetheisa ornatrix*). *Proc. Natl Acad. Sci. USA* **96,** 5570–5574 (1999).
9. Iyengar, V. K. & Eisner, T. Heritability of body mass, a sexually selected trait, in an arctiid moth (*Utetheisa ornatrix*). *Proc. Natl Acad. Sci. USA* **96,** 9169–9171 (1999).
10. Opler, P. A. & Krizek, G. O. *Butterflies East of the Great Plains* (John Hopkins Univ., Baltimore, 1984).
11. LaMunyon, C. W. & Eisner, T. Postcopulatory sexual selection in an arctiid moth (*Utetheisa ornatrix*). *Proc. Natl Acad. Sci. USA* **90,** 4689–4692 (1993).
12. Lande, R. Models of speciation by sexual selection on polygenic traits. *Proc. Natl Acad. Sci. USA* **78,** 3721–3725 (1981).
13. Andersson, M. *Sexual Selection* (Princeton Univ. Press, Princeton, 1994).
14. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits*, 4th edn (Sinauer, Sunderland, 1998).
15. Houde, A. E. Sex-linked heritability of a sexually selected character in a natural population of *Poecilia reticulata* (Pisces: Poeciliidae) (guppies). *Heredity* **69,** 229–235 (1992).
16. Wilkinson, G. S., Kahler, H. & Baker, R. H. Evolution of female mating preferences in stalk-eyed flies. *Behav. Ecol.* **9,** 525–533 (1998).
17. Brooks, R. & Endler, J. A. Female guppies agree to differ: phenotypic and genetic variation in mate-choice behaviour and the consequences for sexual selection. *Evolution* **55,** 1644–1655 (2001).
18. Brooks, R. & Endler, J. A. Direct and indirect sexual selection and quantitative genetics of male traits in guppies (*Poecilia reticulata*). *Evolution* **55,** 1002–1015 (2001).
19. Wolfenbarger, L. L. & Wilkinson, G. S. Sex-linked expression of a sexually selected trait in the stalk-eyed fly, *Crytodiopsis dalmanni*. *Evolution* **55,** 103–110 (2001).
20. Sharma, V. L. Chromosome studies on two species of moths. *Bionature* **19,** 65–67 (1999).
21. Traut, W. & Frantisek, M. Sex chromosome differentiation in some species of Lepidoptera (Insecta). *Chromosome Res.* **5,** 283–291 (1997).
22. Grula, J. W. & Taylor, O. R. Jr The effect of X-chromosome inheritance on mate selection behaviour in the sulfur butterflies, *Colias eurytheme* and *C. philodice*. *Evolution* **34,** 688–695 (1980).
23. Conner, W. E., Eisner, T., Vander Meer, R. K., Guerrero, A. & Meinwald, J. Precopulatory sexual interaction in an arctiid moth (*Utetheisa ornatrix*): Role of a pheromone derived from dietary alkaloids. *Behav. Ecol. Sociobiol.* **9,** 227–235 (1981).
24. Iyengar, V. K., Rossine, C. & Eisner, T. Precopulatory assessment of male quality in an arctiid moth (*Utetheisa ornatrix*): hydroxydanaidal is the only criterion of choice. *Behav. Ecol. Sociobiol.* **49,** 283–288 (2001).

Acknowledgements We thank J. Ladau and W. E. Conner for the collection of field-caught individuals, and J. Schlesinger for technical assistance. This manuscript was improved through comments from J. Dale, S. M. Flaxman and E. A. Tibbetts. Research support was provided by the National Institutes of Health (T.E.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to V.K.I. (e-mail: viyengar@wooster.edu).

Detecting recent positive selection in the human genome from haplotype structure

Pardis C. Sabeti*†#, David E. Reich*, John M. Higgins* Haninah Z. P. Levine*, Daniel J. Richter*, Stephen F. Schaffner*, Stacey B. Gabriel*, Jill V. Platko*, Nick J. Patterson*, Gavin J. McDonald*, Hans C. Ackerman‡, Sarah J. Campbell‡, David Altshuler*§, Richard Cooper‖, Dominic Kwiatkowski‡, Ryk Ward† & Eric S. Lander*¶

** Whitehead Institute/MIT Center for Genome Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA*
† Institute of Biological Anthropology, University of Oxford, Oxford, OX2 6QS, UK
‡ Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK
§ Departments of Genetics and Medicine, Harvard Medical School, Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
‖ Department of Preventive Medicine and Epidemiology, Loyola University Medical School, Maywood, Illinois 60143, USA
¶ Department of Biology, MIT, Cambridge, Massachusetts 02139, USA
 # Harvard Medical School, Boston, Massachusetts 02115, USA

The ability to detect recent natural selection in the human population would have profound implications for the study of human history and for medicine. Here, we introduce a framework for detecting the genetic imprint of recent positive selection by analysing long-range haplotypes in human populations. We first identify haplotypes at a locus of interest (core haplotypes). We then assess the age of each core haplotype by the decay of its association to alleles at various distances from the locus, as measured by extended haplotype homozygosity (EHH). Core haplotypes that have unusually high EHH and a high population frequency indicate the presence of a mutation that rose to prominence in the human gene pool faster than expected under

neutral evolution. We applied this approach to investigate selection at two genes carrying common variants implicated in resistance to malaria: *G6PD*[1] and CD40 ligand[2]. At both loci, the core haplotypes carrying the proposed protective mutation stand out and show significant evidence of selection. More generally, the method could be used to scan the entire genome for evidence of recent positive selection.

The recent history of the human population is characterized by great environmental change and emergent selective agents[3]. The domestication of plants and animals at the start of the Neolithic, roughly 10,000 years ago, yielded an increase in human population density. Humans were confronted with the spread of new infectious diseases, new food sources and new cultural environments. The last 10,000 years have thus been some of the most interesting times in human biological history, and may be when many important genetic adaptations and disease resistances arose.

We sought to design a powerful approach for detecting recent selection. Our method relies on the relationship between an allele's frequency and the extent of linkage disequilibrium (LD) surrounding it. (LD often refers to association between two alleles. Here, we use it to measure the association between a single allele at one locus with multiple loci at various distances.) Under neutral evolution, new variants require a long time to reach high frequency in the population, and LD around the variants will decay substantially during this period owing to recombination[4,5]. As a result, common alleles will typically be old and will have only short-range LD. Rare alleles may be either young or old and thus may have long- or short-range LD. The key characteristic of positive selection, however, is that it causes an unusually rapid rise in allele frequency, occurring over a short enough time that recombination does not substantially break down the haplotype on which the selected mutation occurs. A signature of positive natural selection is thus an allele having unusually long-range LD given its population frequency. The decay of LD, and therefore the relative scale of 'short'- and 'long'-range LD, is dependent on local recombination rates. A general test for selection on the basis of these principles must therefore control for local variation in recombination rates.

We developed an experimental design to detect positive selection at a locus using the breakdown of LD as a clock for estimating the ages of alleles. We began by genotyping a collection of single nucleotide polymorphisms (SNPs) in a small 'core region' to identify the 'core haplotypes'. We selected SNPs of sufficient density, so that recombination between them would be extremely rare and the core haplotypes could be explained in terms of a single gene genealogy (Supplementary Fig. 1). Zones of very low historical recombination were identified by looking for clusters of SNPs where Hudson's R_M was 0 and $|D'|$ was one[6,7] (see Supplementary Fig. 1).

We then added increasingly distant SNPs to study the decay of LD from each core haplotype. To visualize this process, we generated haplotype bifurcation diagrams that branch to reflect the creation of new, extended haplotypes by historical recombination proximal and distal to the core region. We measured LD at a distance x from the core region by calculating the extended haplotype homozygosity (EHH). EHH is defined as the probability that two randomly chosen chromosomes carrying the core haplotype of interest are identical by descent (as assayed by homozygosity at all SNPs[8]) for the entire interval from the core region to the point x. EHH thus detects the transmission of an extended haplotype without recombination. Our test for positive selection involves finding a core haplotype with a combination of high frequency and high EHH, as compared with

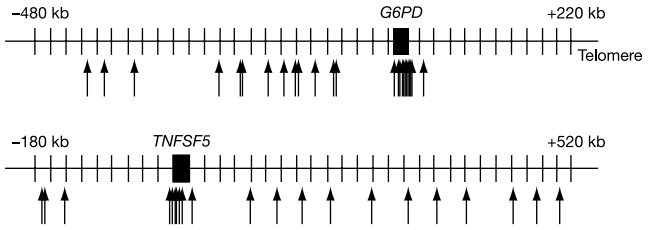

Figure 1 Experimental design of core and long-range SNPs for *G6PD* and *TNFSF5*. The core region is highlighted by a cluster of densely spaced SNPs (arrows) at the gene. Additional, widely separated flanking SNPs, used to examine the decay of LD from each core haplotype, are also shown. Markers distal to *G6PD* were within repetitive subtelomeric sequence and could not be genotyped.

Table 1 **Core haplotype frequencies in six populations**

(**a**) *G6PD*

Core haplotype	Core SNP alleles (kb)											Core haplotype frequencies in six populations						
	−10	−2	−2	−1	0	1	1	3	3	4	4	Total	Beni	Yoruba	Shona	African American	European American	Asian
1	C	G	C	G	G	A	C	C	G	C	C	0.13 (54)	0.15 (9)	0.21 (18)	0.13 (11)	0.13 (12)	0.03 (2)	0.07 (2)
2	–	–	–	–	–	–	–	–	–	–	T	0.16 (67)	0 (0)	0 (0)	0 (0)	0.07 (6)	0.57 (37)	0.80 (24)
3	–	–	–	–	–	–	–	–	–	T	–	0.25 (102)	0.23 (14)	0.24 (21)	0.45 (37)	0.19 (17)	0.17 (11)	0.07 (2)
4	–	–	–	–	–	–	–	T	–	–	–	0.01 (5)	0.02 (1)	0 (0)	0.04 (3)	0.01 (1)	0 (0)	0 (0)
5	–	–	–	–	–	–	–	T	A	–	–	0.10 (41)	0.22 (13)	0.11 (10)	0.06 (5)	0.14 (13)	0 (0)	0 (0)
6	–	–	–	–	–	G	–	–	–	–	–	0.12 (48)	0.17 (10)	0.10 (9)	0.11 (9)	0.22 (20)	0 (0)	0 (0)
7	–	–	T	A	–	G	G	–	–	–	–	0.04 (17)	0.05 (3)	0.07 (6)	0.06 (5)	0.03 (3)	0 (0)	0 (0)
8	–	A*	T	A	A	G	G	–	–	–	–	0.13 (53)	0.17 (10)	0.23 (20)	0.13 (11)	0.13 (12)	0 (0)	0 (0)
9	T	–	–	–	–	–	–	–	–	–	T	0.07 (28)	0 (0)	0.03 (3)	0.02 (2)	0.07 (6)	0.23 (15)	0.07 (2)
											N	415	60	87	83	90	65	30

(**b**) *TNFSF5*

Core haplotype	Core SNP alleles (kb)					Core haplotype frequencies in six populations						
	−6	0	1	3	4	Total	Beni	Yoruba	Shona	African American	European American	Asian
1	T	T	T	T	G	0.03 (12)	0.06 (4)	0.03 (3)	0.02 (2)	0.04 (3)	0 (0)	0 (0)
2	C	–	–	–	–	0.50 (200)	0.32 (20)	0.38 (33)	0.46 (38)	0.40 (32)	0.78 (49)	1.00 (28)
3	C	–	–	C	–	0.08 (32)	0.10 (6)	0.12 (10)	0.06 (5)	0.14 (11)	0 (0)	0 (0)
4	–	C*	–	–	–	0.25 (100)	0.38 (24)	0.38 (33)	0.26 (21)	0.27 (22)	0 (0)	0 (0)
5	–	–	C	–	–	0.12 (47)	0.14 (9)	0.07 (6)	0.18 (15)	0.14 (11)	0.10 (6)	0 (0)
6	–	–	C	–	A	0.02 (10)	0 (0)	0 (0)	0.01 (1)	0.01 (1)	0.13 (8)	0 (0)
7	–	C*	C	–	A	0 (1)	0 (0)	0.01 (1)	0 (0)	0 (0)	0 (0)	0 (0)
8	C	–	C	–	–	0 (1)	0 (0)	0 (0)	0 (0)	0.01 (1)	0 (0)	0 (0)
					N	403	63	86	82	81	63	28

Observed core haplotypes at *G6PD* and *TNFSF5* in six populations of African, European and Asian descent. Relative distances of core SNP alleles from the putative malaria resistance mutations are given in kb. Frequencies for haplotypes (and numbers of observations) are given for all populations. There are no apparent recombinants among the *G6PD* core haplotypes, and R_M is 0. There are 2 recombinant haplotypes among the 403 *TNFSF5* chromosomes, and R_M would also be 0 if the 2 haplotypes appearing only once were removed from the analysis[6].
*Both proposed mutations associated with malaria resistance (*G6PD*-202A and *TNFSF5*-726C) are observed only in Africans and occur on *G6PD*-CH8 and on *TNFSF5*-CH4 and *TNFSF5*-CH7, respectively.

other core haplotypes at the locus. An attractive aspect of this approach is that the various core haplotypes at a locus serve as internal controls for one another, adjusting for any unevenness in the local recombination rate.

We applied our approach to two genes that have been implicated in resistance to the malaria parasite *Plasmodium falciparum*. Glucose-6-phosphate dehydrogenase (*G6PD*) is a classical example of a gene where variants can confer malaria resistance[9]. Evidence over the past 40 years has shown that the common variant *G6PD*-202A confers partial protection against malaria, with a case-control study estimating a reduction in disease risk of about 50% (ref. 1). The CD40 ligand gene (*TNFSF5*) encodes a protein with a critical role in immune response to infectious agents. One case-control study suggested that a common variant in the promoter region,

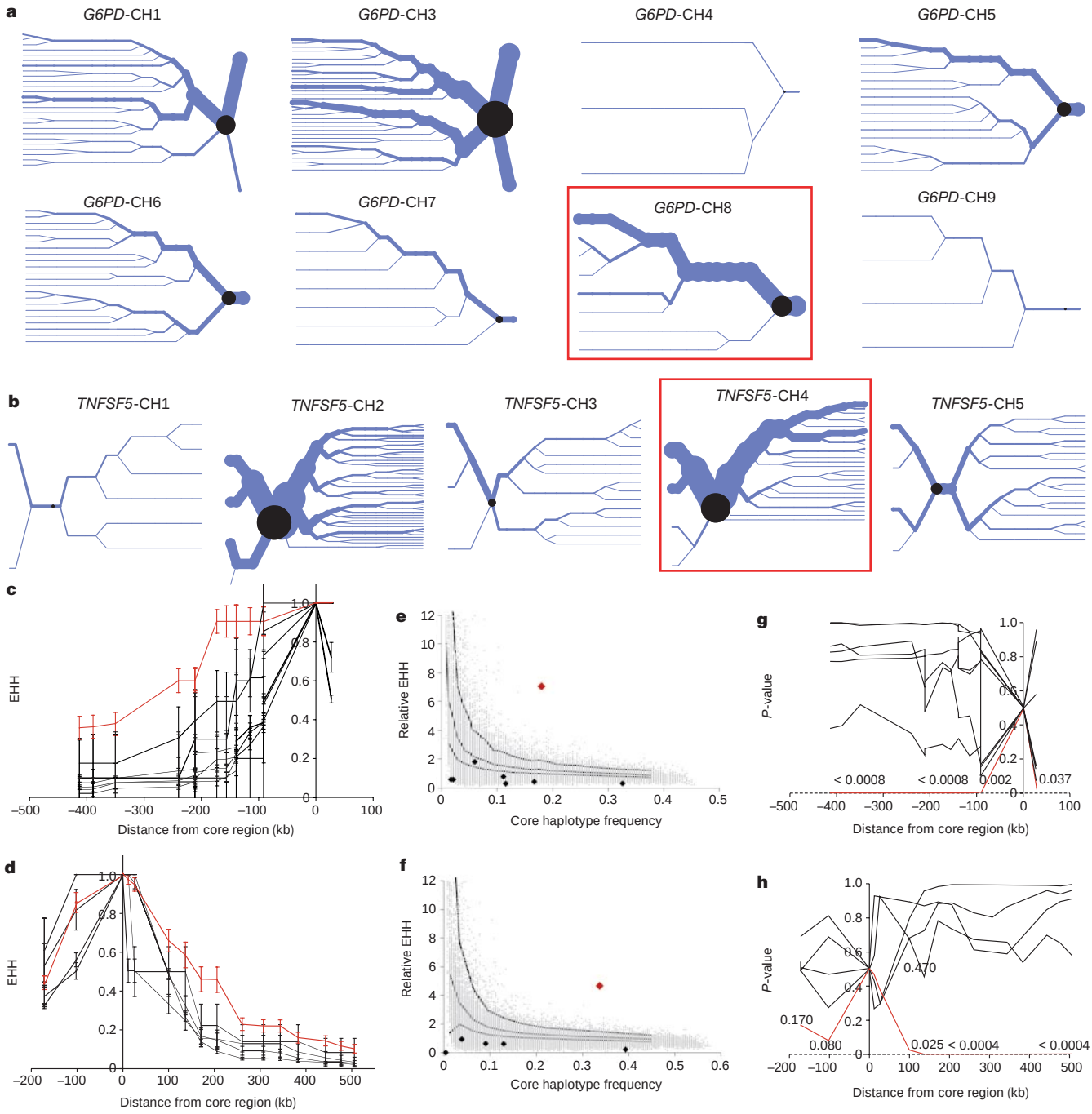

Figure 2 Core haplotype frequency and relative EHH of *G6PD* and *TNFSF5*. **a**, **b**, Haplotype bifurcation diagrams (see Methods) for each core haplotype at *G6PD* (**a**) and *TNFSF5* (**b**) in pooled African populations demonstrate that *G6PD*-CH8 and *TNFSF5*-CH4 (boxed or labelled in red) have long-range homozygosity that is unusual given their frequency. **c**, **d**, The EHH at varying distances from the core region on each core haplotype at *G6PD* (**c**) and *TNFSF5* (**d**) demonstrates that *G6PD*-CH8 and *TNFSF5*-CH4 have persistent, high EHH values. **e**, **f**, At the most distant SNP from *G6PD* (**e**) and *TNFSF5* (**f**) core regions, the relative EHH plotted against the core haplotype frequency is presented and compared with the distribution of simulated core haplotypes (on the basis of simulation of 5,000 data sets; represented by grey dots and given with 95th, 75th and 50th percentiles). The observed non-selected core haplotypes in our data are represented by black diamonds. **g**, **h**, We calculated the statistical significance of the departure of the observed data from the simulated distribution at each distance from the core. *G6PD*-CH8 (**g**) and *TNFSF5*-CH4 (**h**) demonstrate increasing deviation from a model of neutral drift at further distances from the core region in both directions.

TNFSF5-726C, is associated with a similar degree of protection against malaria[2].

We first studied *G6PD* (Fig. 1). We defined a core region of 15 kilobases (kb) at *G6PD* and genotyped 11 SNPs in 3 African and 2 non-African populations. The SNPs defined 9 core haplotypes (Table 1a) (denoted *G6PD*-CH1 to 9, for core haplotypes 1 to 9). The *G6PD*-202A allele, which has been associated with protection from malaria, was carried on only one core haplotype, *G6PD*-CH8. Notably, *G6PD*-CH8 is common in Africa (18%), where malaria is endemic, but is absent outside of Africa. For carrying out our test for selection, we focused on the three African populations, which did not differ significantly with respect to core haplotype frequencies (by Fisher's exact test[10]) and hence were pooled for the main analysis. (Analyses were also performed separately for each population, yielding qualitatively similar results; see below.)

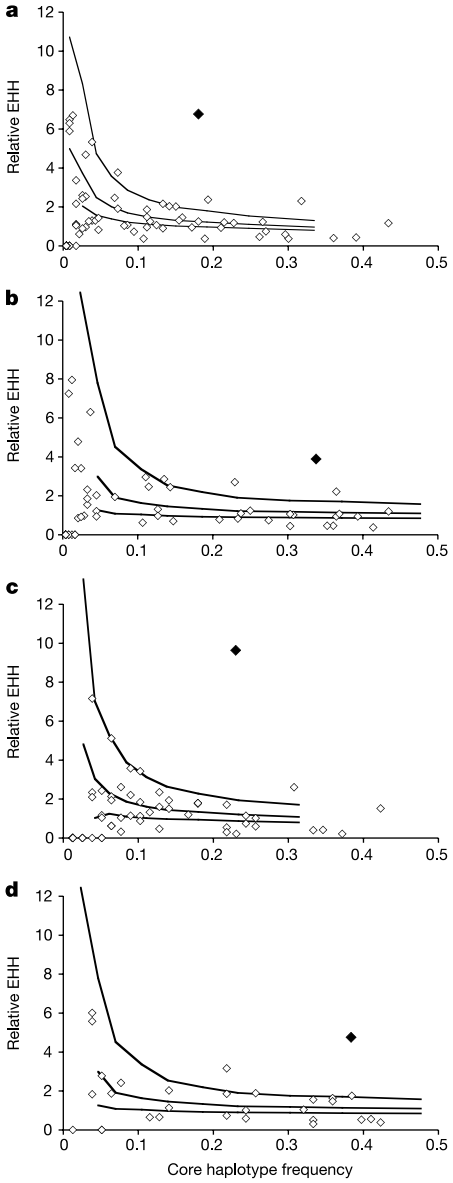

Figure 3 Control regions: core haplotype frequency against relative EHH. To provide an empirical, non-simulation-based evaluation of the signal of selection, we compared the frequency and relative EHHs for *G6PD* (**a**) and *TNFSF5* (**b**) with patterns observed in randomly chosen genes in the genome (see Methods). **c**, **d**, We performed the entire analysis again on *G6PD* (**c**) and *TNFSF5* (**d**) for a subset of 78 Yoruban haplotypes (using family trios where phase could be determined). We were able to match 30 to 87 core haplotypes (indicated by outlined diamonds) from the control regions to our data. The 95th, 75th and 50th percentiles for simulated data are also shown. *G6PD*-CH8 and *TNFSF5*-CH4 (indicated by black diamonds) clearly stand out from the pattern seen at other loci in the genome, suggesting a true signal of selection.

G6PD-CH8 demonstrates clear long-range LD (as seen by the predominance of one thick branch in the haplotype bifurcation diagram (Fig. 2a)) and has correspondingly high EHH. The EHH is 0.38 at the largest distance tested; that is, 413 kb (Fig. 2c). For each core haplotype we calculated the relative EHH; specifically, the factor by which EHH decays on the tested core haplotype compared with the decay of EHH on all other core haplotypes combined (Methods).

To test formally for selection, for each core haplotype, we compared the allele frequency to the relative EHH at various distances (Fig. 2e shows the comparison at 413 kb proximal to *G6PD*). *G6PD*-CH8 has a much higher relative EHH than other haplotypes of comparable frequency, but is this statistically significant? To obtain a sense of how unusual our observation is, we simulated haplotypes using a coalescent process (see Methods and Fig. 2e)[11]. The deviation from the simulation results is highly significant and becomes progressively more marked with increasing distance (Fig. 2g) (*P*-values at 413 kb proximal are: constant-sized population, $P < 0.0008$; expansion, $P < 0.0006$; bottleneck, $P < 0.0008$; population structure, $P < 0.0008$; see Methods and Supplementary Table 2 for details of the demographic models we considered). The frequency and LD properties of *G6PD*-CH8 are incompatible with what is expected under a model of neutral evolution for a wide range of demographies. Furthermore, when the three African populations comprising the pooled sample were considered separately, a signal of selection was identified independently in each population (Yoruba, $P < 0.0012$; Beni, $P < 0.0440$; and Shona, $P < 0.0030$, based on simulation of a constant-sized population), demonstrating that the signal of selection is not an artefact of pooling the three population samples.

We next applied our approach to the CD40 ligand gene (Fig. 1). We defined a core region of 10 kb and genotyped 5 SNPs. The SNPs defined seven core haplotypes (Table 1b). The *TNFSF5*-726C allele, which has been associated with protection from malaria, was present on *TNFSF5*-CH4, which is common in Africa (34%), but is absent outside of Africa. *TNFSF5*-CH4 demonstrates high LD as seen in the haplotype bifurcation diagrams (Fig. 2b) and has high EHH at long distances (Fig. 2d). *TNFSF5*-CH4 is a clear outlier when compared with other haplotypes (Fig. 2f), and its frequency and LD properties are incompatible with neutral evolution under multiple demographic models (*P*-values at 506 kb distal are: constant-sized population, $P < 0.0012$; expansion, $P < 0.0008$; bottleneck, $P < 0.0012$; population structure, $P < 0.0008$; see Methods and Supplementary Table 2 for details). Again, the *P*-value is increasingly significant at further distances from the core region both proximally and distally (Fig. 2h). When each African population was analysed separately, a signal of selection was significant (Yoruba, $P < 0.0008$; Beni, $P < 0.0023$; Shona, $P < 0.0242$). These results thus provide independent evidence supporting the proposed role of CD40 ligand in malaria resistance[2].

We tested our conclusion of positive selection by performing a similar analysis on 17 randomly chosen control regions across the human genome in the same African populations. We only used data from each control if it was closely matched to our data in terms of the number of chromosomes studied and the homozygosity at the core haplotype and at long distances from the core (Fig. 3). *G6PD*-CH8 and *TNFSF5*-CH4 clearly stand out from the other loci, showing that the *P*-values determined by simulation are also supported by direct, empirical comparison. In measuring *P*-values for the controls where there is no prior hypothesis of selection, a Bonferonni correction for multiple-hypothesis testing was applied. Notably, one core haplotype, from the monocyte chemotactic protein 1 region, shows frequency and LD properties similar to

G6PD and *TNFSF5*, although this nominally significant result may be simply a false-positive owing to the large number of hypotheses examined.

We used a linkage-disequilibrium-based technique[12] to estimate dates of origin of the two resistance variants. The estimates were about 2,500 years for *G6PD* and about 6,500 years for *TNFSF5* (see Supplementary Information for details). The date for G6PD is consistent with a recent independent age estimate for *G6PD*-202A based primarily on microsatellite data[13].

Finally, we explored whether positive selection could have been detected with traditional tests (Supplementary Table 3)[14]. We performed Tajima's *D*-test[15], Fu and Li's *D*-test[16], Fay and Wu's *H*-test[17], the Ka/Ks test[18], the McDonald and Kreitman test[19], and the Hudson–Kreitman–Aguadè (HKA) test[20]. None showed significant deviation from neutral evolution for either *G6PD* or *TNFSF5*, consistent with their low power to detect recent selection.

Our approach, which we refer to as the long-range haplotype (LRH) test, provides a way to detect recent positive selection by analysing haplotype structure in random individuals from a population. How far back in human history can one detect positive selection? Selective events occurring less than 400 generations ago (10,000 years assuming 25 years per generation) should leave a clear imprint at distances of over 0.25 centiMorgans (cM). The signal of such long-range LD should be distinguishable from the background extent of LD for common haplotypes in the genome[21], which are typically tens of thousands of generations old[4] and hence extend 0.02 cM or less. Over many tens of thousands of years, the signal of selection will become lost as recombination whittles the long-range haplotypes to the typical size of haplotype blocks in the human genome[21].

The LRH test can be used to search for evidence of positive selection by testing each common haplotype in a gene, without prior knowledge of a specific variant or selective advantage. Once the signature of selection is found, one must then decipher its cause. The LRH test could be applied to scan the entire human genome for evidence of recent positive selection simply by applying it to each haplotype block in reference data sets from human populations, as will be collected by the Human Haplotype Map project[21]. In this fashion, it should be possible to shed light on how the human genome was shaped by recent changes in culture and environment. The LRH test should also be useful for studying selection in other organisms, including domestic animals and parasites such as the malaria parasite *P. falciparum*[22]. □

Methods

Human subjects

DNA samples from 252 males from Africa were used in the study: 92 Yoruba and 73 Beni from Nigeria, and 87 Shona from Zimbabwe. Additional DNA samples from 29 Yoruban trios (father–mother–child clusters) were genotyped at the 17 control regions. The Yoruba and Shona males were healthy individuals obtained as part of the International Collaborative Study of Hypertension in Blacks. The Beni samples were from civil servants in Benin City. Samples from four non-African populations and four primates were also used (Supplementary Information).

SNP genotyping

We genotyped 49 SNPs distributed around *G6PD* and 37 SNPs distributed around *TNFSF5* using mass spectrometry (Sequenom)[23]. The SNPs were identified by our own re-sequencing and through previous discovery efforts[2,24,25]. For *G6PD*, we focused on genotyping SNPs proximal to the gene, as SNPs in the repetitive subtelomeric sequence distal to *G6PD* could not be genotyped. A total of 25 SNPs around *G6PD* and 21 SNPs around *TNFSF5* were successfully genotyped and used in analysis (see Supplementary Information for details).

Haplotype bifurcation diagrams

To visualize the breakdown of LD on core haplotypes, we created bifurcation diagrams using MATLAB. The root of each diagram is a core haplotype, identified by a black circle. The diagram is bi-directional, portraying both proximal and distal LD. Moving in one direction, each marker is an opportunity for a node; the diagram either divides or not based on whether both or only one allele is present. Thus the breakdown of LD on the core haplotype background is portrayed at progressively longer distances. The thickness of the lines corresponds to the number of samples with the indicated long-distance haplotype.

Extended haplotype homozygosity and relative EHH

EHH at a distance x from the core region is defined as the probability that two randomly chosen chromosomes carrying a tested core haplotype are homozygous at all SNPs[8] for the entire interval from the core region to the distance x. EHH is on a scale of 0 (no homozygosity, all extended haplotypes are different) to 1 (complete homozygosity, all extended haplotypes are the same). Relative EHH is the ratio of the EHH on the tested core haplotype compared with the EHH of the grouped set of core haplotypes at the region not including the core haplotype tested. Relative EHH is therefore on a scale of 0 to infinity.

Coalescent simulations

We used a computer program by Hudson that simulates gene history with recombination[11]. The program was modified to generate data such as we collected. We simulated a long region of DNA (1.3 cM), with one end defined as the 'core'. We progressively added SNPs at the core until they matched our data (for the *G6PD* or *TNFSF5* core) to within ±12.5% in terms of the homozygosity. To mimic the SNP selection strategy used by The SNP consortium[25], which was the source of most of the SNPs in our study, we only included simulated SNPs in our analysis if different alleles were observed at two randomly chosen chromosomes from the sample. At longer distances, we added additional SNPs, only choosing SNPs for analysis that matched our data in terms of frequency (within a ±12.5% window) and also broke down EHH to the same extent as was observed in our data (within ±12.5%).

We repeated the simulations for 5,000 data sets (each producing typically 6–8 core haplotypes) to generate many data points with which to compare our data. *P*-values were obtained by first binning the simulated data by core haplotype frequency into 30 bins of equal size, each containing about 1,000 data points. We then ranked an observed core haplotype's relative EHH compared with that of all simulated data points within the bin containing haplotypes of the same frequency—the rank determines the *P*-value. For simulations of additional demographic histories, we considered two models of expansion, an extreme bottleneck, and a highly structured population. For expansions, we simulated a population that was constant at size 10,000 until 200 or 5,000 generations ago, when it expanded suddenly by a factor of 1,000. For an extreme bottleneck, we simulated a population that was constant at size 10,000 except for a brief bottleneck (inbreeding coefficient 0.18) that occurred 800 generations ago. (An inbreeding coefficient[26] of 0.18 is generated by dropping the population size to 800 chromosomes for 160 generations.) For a structured population, we simulated two equal-sized populations of size $N/2$ that exchanged migrants throughout history with a probability of $1/8N$ per generation per chromosome.

Results of the simulations remained qualitatively similar when we explored additional demographies and when we varied the stringency of matching to SNP allele frequencies and to haplotype homozygosities. A comprehensive, simulation-based exploration of the LRH test will be presented elsewhere, along with explorations of the statistical power of the test and computer code for implementing the LRH test on other data sets, including those with missing or unphased data (D.E.R., manuscript in preparation).

Control regions

To obtain control data for comparison to the *G6PD* and *TNFSF5* haplotypes, we genotyped the same population samples in 17 randomly chosen autosomal genes (*ACVR2B*, *TGFB1*, *DDR1*, *GTF2H4*, *COL11A2*, *LAMB1*, *WASL*, *SLC6A12*, *KCNA1*, *ARGHDIB*, *PCI*, *PRKCB1*, *NF1*, *SCYA2*, *PAI2*, *IL17R* and *HCF2*) selected previously as part of a genome-wide survey of linkage disequilibrium[26]. For our analyses we randomly picked chromosomes to match the numbers sampled for *G6PD* and *TNFSF5*, and we only included those control regions that we could match to our data in terms of homozygosity at the core and homozygosity at long distance (±25% stringency of matching). After the filtering process, we evaluated *G6PD* at 240.3 kb proximal to the gene and *TNFSF5* at 343.9 kb distal to the gene, because we could not make enough comparisons to control regions at the further distances. Seven genes matched to *G6PD* and seven genes matched to *TNFSF5*. We repeated the analysis using 78 Yoruban chromosomes for which phase information was known experimentally (because of genotyping in trios) and for which phase for the most part did not have to be inferred computationally. Six genes matched to *G6PD* and six genes matched to *TNFSF5* (see Supplementary Information for details).

Received 7 June; accepted 19 September 2002; doi:10.1038/nature01140.
Published online 9 October 2002.

1. Ruwende, C. & Hill, A. Glucose-6-phosphate dehydrogenase deficiency and malaria. *J. Mol. Med.* **76,** 581–588 (1998).
2. Sabeti, P. *et al.* CD40L association with protection from severe malaria. *Genes Immun.* **3,** 286–291 (2002).
3. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, 1994).
4. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge/New York, 1983).
5. Stephens, J. C. *et al.* Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *Am. J. Hum. Genet.* **62,** 1507–1515 (1998).
6. Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111,** 147–164 (1985).
7. Lewontin, R. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49,** 49–67 (1964).
8. Nei, M. *Molecular Evolutionary Genetics* Eqn. 8.4 (Columbia Univ. Press, New York, 1987).
9. Luzatto, L., Mehta, A. & Vulliamy, T. *The Metabolic & Molecular Bases of Inherited Disease* 4517–4553 (McGraw-Hill, New York, 2001).
10. Raymond, M. & Rousset, F. An exact test for population differentiation. *Evolution* **49,** 1280–1283 (1995).
11. Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23,** 183–201 (1983).

12. Reich, D. E. & Goldstein, D. B. *Microsatellites: Evolution and Applications* 128–138 (Oxford Univ. Press, Oxford/New York, 1999).
13. Tishkoff, S. A. *et al.* Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293,** 455–462 (2001).
14. Rozas, J. & Rozas, R. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15,** 174–175 (1999).
15. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).
16. Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133,** 693–709 (1993).
17. Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155,** 1405–1413 (2000).
18. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335,** 167–170 (1988).
19. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351,** 652–654 (1991).
20. Hudson, R. R., Kreitman, M. & Aguade, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153–159 (1987).
21. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **23,** 2225–2229 (2002).
22. Wootton, J. C. *et al.* Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* **418,** 320–323 (2002).
23. Tang, K. *et al.* Chip-based genotyping by mass spectrometry. *Proc. Natl Acad. Sci. USA* **96,** 10016–10020 (1999).
24. Vulliamy, T. J. *et al.* Linkage disequilibrium of polymorphic sites in the G6PD gene in African populations and the origin of G6PD A. *Gene Geogr.* **5,** 13–21 (1991).
25. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
26. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411,** 199–204 (2001).

Supplementary Information accompanies the paper on *Nature*'s website (**http://www.nature.com/nature**).

Acknowledgements We thank B. Blumenstiel, M. DeFelice, A. Lochner, J. Moore, H. Nguyen and J. Roy for assistance in genotyping the 17 control regions. We also thank L. Gaffney, S. Radhakrishna, T. DiCesare and T. Lavery for graphics and technical support, B. Ferrell for the Beni samples, and A. Adeyemo and C. Rotimi for helping to collect the Yoruba and Shona samples. Finally, we thank M. Daly, E. Cosman, B. Gray, V. Koduri, T. Herrington and L. Peterson for comments on the manuscript. P.C.S. was supported by grants from the Rhodes Trust, the Harvard Office of Enrichment, and by a Soros Fellowship. This work was supported by grants from the National Institute of Health.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.S.L. (e-mail: lander@genome.wi.mit.edu).

Voltage-sensing mechanism is conserved among ion channels gated by opposite voltages

Roope Männikkö*, Fredrik Elinder* & H. Peter Larsson†

** Department of Neuroscience, The Nobel Institute for Neurophysiology, Karolinska Institutet, SE-171 77 Stockholm, Sweden*
† Neurological Sciences Institute, Oregon Health & Science University, 505 NW 185th Avenue, Beaverton, Oregon 97006

Hyperpolarization-activated cyclic-nucleotide-gated (HCN) ion channels are found in rhythmically firing cells in the brain and in the heart[1], where the cation current through HCN channels (called I_h or I_f) causes these cells to fire repeatedly[2]. These channels are also found in non-pacing cells, where they control resting membrane properties, modulate synaptic transmission, mediate long-term potentiation, and limit extreme hyperpolarizations[3–7]. HCN channels share sequence motifs with depolarization-activated potassium (Kv) channels, such as the fourth transmembrane segment S4[8,9]. S4 is the main voltage sensor of Kv channels, in which transmembrane movement of S4 charges triggers the opening of the activation gate[10–17]. Here, using cysteine accessibility methods[10–12], we investigate whether S4 moves in an HCN channel. We show that S4 movement is conserved between Kv and HCN channels, which indicates that S4 is also the voltage sensor in HCN channels. Our results suggest that a conserved voltage-sensing mechanism operates in the oppositely voltage-gated Kv and HCN channels, but that there are different coupling mechanisms between the voltage sensor and activation gate in the two different channels.

If S4 is the voltage sensor in HCN channels, some of its charges must move relative to the electric field across the membrane during activation. To test this hypothesis, we measured the solvent accessibility of several residues in S4 of spHCN channels[9] (see Methods) in both open and closed channels, to look for state-dependent changes in accessibility. We introduced cysteines at eight different positions in S4 of spHCN channels (Fig. 1a, b) and assayed the accessibility of the cysteines by applying the membrane-impermeable thiol reagent MTSET ([2-(trimethylammonium)ethyl] methanethiosulphonate, bromide)[10–12,17].

In Fig. 1c we show a typical family of wild-type spHCN channel currents. All mutations but R341C expressed well. Figure 1d shows typical normalized conductance versus voltage curves ($G(V)$) for wild-type channels and all seven mutations. The channels with charge-neutralizing cysteine substitutions activated at more hyperpolarized potentials than did the wild-type spHCN channels. The voltage shifts caused by the cysteine substitutions were similar to the hyperpolarizing shifts caused by neutralizations of S4 charges in other HCN channels[18,19].

Perfusion of external and internal MTSET (100 μM for 2 min) had very small effects on wild-type spHCN channels, mainly a <10% change in the current amplitude. However, a similar application of MTSET on the cysteine-substituted channels resulted in clear, irreversible changes in the amplitude of the current, in the activation rate of the current, or in the voltage dependence of the channels. (K329C was not clearly affected by MTSET at any potential and was, therefore, not included in the subsequent study.) The modification rate of the cysteines was measured by plotting the amplitude of the MTSET-induced change as a function

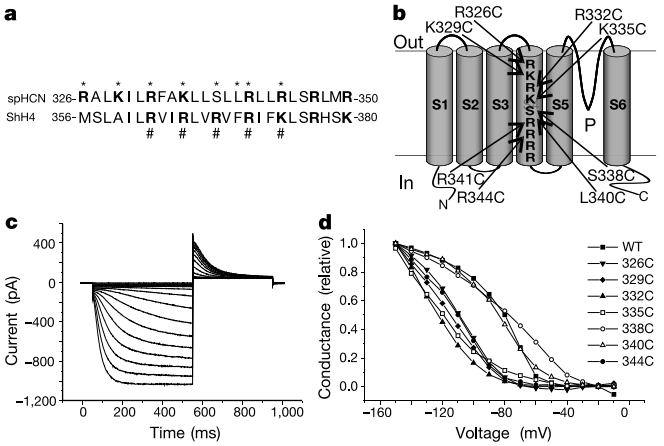

Figure 1 Basic features of spHCN channels. **a**, Sequence alignment of spHCN and Shaker channels in the S4 region. Bold indicates positive charges, asterisk indicates residues mutated in spHCN, and hash indicates the positive charges contributing to voltage sensing in Shaker[11,13,14]. **b**, Putative transmembrane topology showing the positions of the introduced cysteines. **c**, Patch-clamp recording of currents through wild-type (WT) spHCN channels in response to voltage pulses from −10 mV to −150 mV, from a holding potential of −10 mV. Tail potential, +50 mV. **d**, Conductance versus voltage $G(V)$ curves for WT and cysteine-substituted spHCN channels measured in excised patches. $G(V)$ curves measured in whole oocytes were shifted to the right by approximately 20–30 mV, as previously reported[9]. The conductance was calculated from tail currents, normalized as $G(V) = (I(V) - I_{min})/(I_{max} - I_{min})$.