nature genetics

Refining the impact of *TCF7L2* gene variants on type 2 diabetes and adaptive evolution

Agnar Helgason[1], Snæbjörn Pálsson[1,2], Gudmar Thorleifsson[1], Struan F A Grant[1,13], Valur Emilsson[1], Steinunn Gunnarsdottir[1], Adebowale Adeyemo[3], Yuanxiu Chen[3], Guanjie Chen[3], Inga Reynisdottir[1], Rafn Benediktsson[4,5], Anke Hinney[6], Torben Hansen[7], Gitte Andersen[7], Knut Borch-Johnsen[7,8], Torben Jorgensen[9], Helmut Schäfer[10], Mezbah Faruque[3], Ayo Doumatey[3], Jie Zhou[3], Robert L Wilensky[11], Muredach P Reilly[11], Daniel J Rader[11], Yu Bagger[12], Claus Christiansen[12], Gunnar Sigurdsson[4,5], Johannes Hebebrand[6], Oluf Pedersen[7,8], Unnur Thorsteinsdottir[1], Jeffrey R Gulcher[1], Augustine Kong[1], Charles Rotimi[3] & Kári Stefánsson[1]

We recently described an association between risk of type 2 diabetes and variants in the transcription factor 7-like 2 gene (*TCF7L2*; formerly *TCF4*), with a population attributable risk (PAR) of 17%–28% in three populations of European ancestry[1]. Here, we refine the definition of the *TCF7L2* type 2 diabetes risk variant, $HapB_{T2D}$, to the ancestral T allele of a SNP, rs7903146, through replication in West African and Danish type 2 diabetes case-control studies and an expanded Icelandic study. We also identify another variant of the same gene, HapA, that shows evidence of positive selection in East Asian, European and West African populations. Notably, HapA shows a suggestive association with body mass index and altered concentrations of the hunger-satiety hormones ghrelin and leptin in males, indicating that the selective advantage of HapA may have been mediated through effects on energy metabolism.

We recently reported an association between three markers (the composite allele X of microsatellite DG10S478 and the T alleles of SNPs rs12255372 and rs7903146) and type 2 diabetes in Icelanders and replicated the finding in two additional case-control series of European ancestry from the USA and Denmark (Denmark A)[1]. These markers are located within a 64-kb block of strong linkage disequilibrium (LD) containing exon 4 and parts of two large flanking introns of the 217-kb *TCF7L2* gene on chromosome 10 (**Supplementary Fig. 1** online), and are highly correlated in populations of European ancestry. To replicate and refine this association to type 2 diabetes, we genotyped these markers in 1,149 affected individuals and 2,400 controls in another Danish sample (Denmark B) and in a more genetically diverse West African group[2] consisting of 621 affected individuals and 448 controls. We also increased the number of genotyped Icelandic controls from 931 to 9,950. In Denmark B, all three variants were strongly associated with disease risk. However, the association of rs7903146 T (relative risk (RR) = 1.49 (95% confidence interval (c.i.): 1.34–1.66), $P = 6.4 \times 10^{-13}$, PAR = 21%) was noticeably stronger than those of the other two variants (**Table 1** and **Supplementary Table 1** online). Haplotypes carrying DG10S478 X or rs12255372 T but not rs7903146 T did not confer any risk. In contrast, haplotypes carrying rs7903146 T conferred similar risk regardless of whether they carried DG10S478 X or rs12255372 T. Comparable results were obtained with the expanded Icelandic case-control study (**Supplementary Fig. 2** online). In the West African study group, the association of rs7903146 T to type 2 diabetes was replicated (RR = 1.45 (1.19–1.77), $P = 0.00021$, PAR = 20%), after adjusting for relatedness and ancestry. However, the association was not significant for DG10S478 X and was weaker for rs12255372 T. Overall, these results rule out DG10S478 X and rs12255372 T as causal variants but identify rs7903146 T as either the risk variant itself or its closest known correlate.

This is consistent with results from recent replication studies, which have reported slightly higher risk at greater statistical significance for rs7903146 T (albeit with no formal attempt to evaluate the difference in effect)[3,4]. Equally important, these results support the notion that relatively diverse populations, such as those of West Africa, provide the

[1]deCODE genetics, 101 Reykjavik, Iceland. [2]University of Iceland, 101 Reykjavik, Iceland. [3]National Human Genome Center, Department of Community and Family Medicine, Howard University, Washington DC 20060, USA. [4]Icelandic Heart Association, 201 Kopavogur, Iceland. [5]National University Hospital, 101 Reykjavik, Iceland. [6]Department of Child and Adolescent Psychiatry, Rheinische Kliniken Essen, University of Duisburg-Essen, 45147 Essen, Germany. [7]Steno Diabetes Center, 2820 Gentofte, Denmark. [8]University of Aarhus, 8000 Aarhus, Denmark. [9]Research Centre for Prevention and Health, University Hospital Glostrup, 2600 Glostrup, Denmark. [10]Institute of Medical Biometry and Epidemiology, Philipps-University of Marburg, 35037 Marburg, Germany. [11]University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104 USA, [12]Center for Clinical and Basic Research A/S, 2750 Ballerup, Denmark. [13]Current address: Center for Applied Genomics, Abramson Research Center, The Children's Hospital of Philadelphia, 3516 Civic Center Blvd, Philadelphia, Pennsylvania 19104, USA. Correspondence should be addressed to A.H. (agnar@decode.is) or K.S. (kstefans@decode.is).

Received 27 June 2006; accepted 8 December 2006; published online 7 January 2007; doi:10.1038/ng1960

Table 1 Association of type 2 diabetes risk with the markers DG10S478, rs12255372 and rs7903146

Study group (number of affected individuals/controls)	Marker (allele)	Frequency in individuals with type 2 diabetes	Frequency in controls	RR (95%c.i.)	*P* value[a]
Denmark B (1,149/2,400)	DG10S478 (X)	0.316	0.252	1.37 (1.23–1.53)	2.11×10^{-8}
	rs12255372 (T)	0.320	0.256	1.36 (1.22–1.52)	3.37×10^{-8}
	rs7903146 (T)	0.345	0.261	1.49 (1.34–1.66)	6.46×10^{-13}
Iceland (1,185/9,950)	DG10S478 (X)	0.364	0.288	1.41 (1.28–1.56)	2.35×10^{-12}
	rs12255372 (T)	0.363	0.289	1.41 (1.28–1.55)	8.98×10^{-12}
	rs7903146 (T)	0.391	0.304	1.47 (1.33–1.62)	5.19×10^{-15}
West Africa: combined (621/448)[b]	DG10S478 (X)	-	-	1.20 (0.91–1.59)	0.19
	rs12255372 (T)	-	-	1.31 (1.01–1.69)	0.044
	rs7903146 (T)	-	-	1.45 (1.19–1.77)	0.00021

[a]*P* values for the Icelandic and the West African studies were adjusted for relatedness of the individuals using simulations. [b]The Mantel-Haenszel model was used to combine the results obtained from the four West African subgroups (results for each subgroup are shown in **Supplementary Table 1**). As the estimates of RR in the combined West African group are not directly derived from the allele frequencies in this group, we do not show these frequencies here.

means to refine association signals detected in relatively homogeneous populations (such as those of European ancestry) characterized by larger regions of strong LD. Sequencing all exons of *TCF7L2* and ~90% of a 120-kb region (114.38–114.5 Mb, NCBI build 34) spanning the LD block containing rs7903146 did not uncover any additional variants with a stronger association to type 2 diabetes than rs7903146 T. Given the pattern of LD surrounding rs7903146 in the HapMap groups, we conclude that if the T allele is not itself the risk variant, then the unidentified functional variant it tags is unlikely to lie outside the aforementioned sequenced region.

A phylogenetic reconstruction of the evolutionary relationships between haplotypes within the *TCF7L2* exon 4 LD block uncovered two major lineages (**Fig. 1**). The more diverse lineage, HapB, contains all but one of the haplotypes carrying rs7903146 T, a subset we refer to as $HapB_{T2D}$. The second major lineage, HapA, consists of a relatively homogeneous cluster of haplotypes, noticeably divergent from HapB, that show a 'star-like' pattern of diversity, with a single basal haplotype accounting for 81% of its chromosomes. There is an unusual degree of divergence between the HapMap groups in this genomic region, with the frequency of HapA ranging from 10% in Nigerian Yoruba (YRI) to 58% in individuals of European ancestry from the Utah pedigree of the Centre d'Etude du Polymorphisme Humain (CEU) and 95% in East Asians (CHB+JPT) (**Fig. 1**). Thus, unlike in populations of European and African ancestry, rs7903146 T ($HapB_{T2D}$) cannot account for a large fraction of type 2 diabetes in East Asian populations, because its frequency is only ~2% in the CHB+JPT HapMap group. Notably, the pattern of diversity shown in **Figure 1** is suggestive of a positive selective sweep that may have rapidly driven HapA almost to fixation in East Asians and to lower frequencies in Europeans and Africans.

To test this hypothesis, we applied two different methods, an F_{ST}-based test and the long-range haplotype (LRH) test[5], to search for signals of positive selection in the HapMap groups. A history of positive selection may be indicated for loci where F_{ST}, the observed proportion of the overall genetic variation due to differences between groups, is in the upper extreme of the range expected under neutral evolution. Using haplotypes constructed from 42 consecutive SNPs within the exon 4 LD block, we obtained $F_{ST} = 0.306$ for the three HapMap groups. Based on 10,000 coalescent simulations (using demographic settings from ref. 6), it emerged that such a high F_{ST} value is very unlikely under neutral evolution ($P = 0.0018$). Examination of F_{ST} values between HapMap group pairs showed that the deviation from neutral expectations is due to the difference between the East Asian group and the other two ($P < 0.002$ in both cases, but $P = 0.567$ for the difference between the CEU and YRI groups). Using only rs7924080 T to tag HapA, we obtained $F_{ST} = 0.7559$ for the three HapMap groups, with only 0.36% of 2,643,043 HapMap SNPs (those typed in all three groups, polymorphic in at least one and not on the X chromosome) showing an equal or greater value of F_{ST}. These results are consistent with a positive selective sweep of HapA in East Asians and no selection (or less intense selection) in the other two HapMap groups.

When a variant is driven rapidly to high frequency by positive selection, insufficient time passes for recombination and mutation to produce the magnitude of genetic diversity expected on the background of a neutral variant of the same frequency[5,7]. The LRH test statistic, relative extended haplotype homozygosity (rEHH), measures the background diversity of a putative positively selected core haplotype relative to the background diversity of other core haplotypes constructed from alleles of the same loci. We defined core haplotypes using three SNPs (rs7924080, rs11196199 and rs12255372) spanning a 20-kb region within the exon 4 LD block (rs7924080 efficiently tags HapA in all HapMap groups). For the three HapMap groups (**Fig. 2**), we calculated rEHH at increasing distances (within a 400-kb region) from the HapA core haplotype and the distribution of rEHH values expected for equally frequent haplotypes under neutral evolution (obtained from coalescent simulations based on demographic settings from ref. 6). Overall, these tests indicate that strong positive selection drove HapA to near fixation in East Asian populations and that weaker or shorter selective episodes may also have increased the frequency of HapA in Europe and Africa. We obtained similar results when we performed a genome-wide comparison of integrated rEHH (irEHH) values for HapA with those of equally frequent alleles for each of the HapMap groups (see Methods). Thus, only 0.7% of 15,497 alleles (East Asians), 7.9% of 17,010 alleles (CEU) and 2.7% of 42,354 alleles (YRI) had irEHH values greater than or equal to those observed for HapA, indicating that the signature of positive selection observed for HapA is unusual in all three HapMap groups when compared with the rest of the genome, especially in the East Asians. We obtained rough age estimates for HapA based on its recombination history[8]: 11,933, 8,401 and 4,051 years for the CEU, East Asian and YRI HapMap groups, respectively. Although tentative, these ages coincide broadly with the onset of agriculture in the three geographic regions represented by the HapMap groups[9].

What was the phenotypic effect of HapA that increased the reproductive success of its carriers in the past? We previously reported a suggestive nonsignificant negative association between $HapB_{T2D}$ and body mass index (BMI) in type 2 diabetes patients[1] (see also ref. 3). To

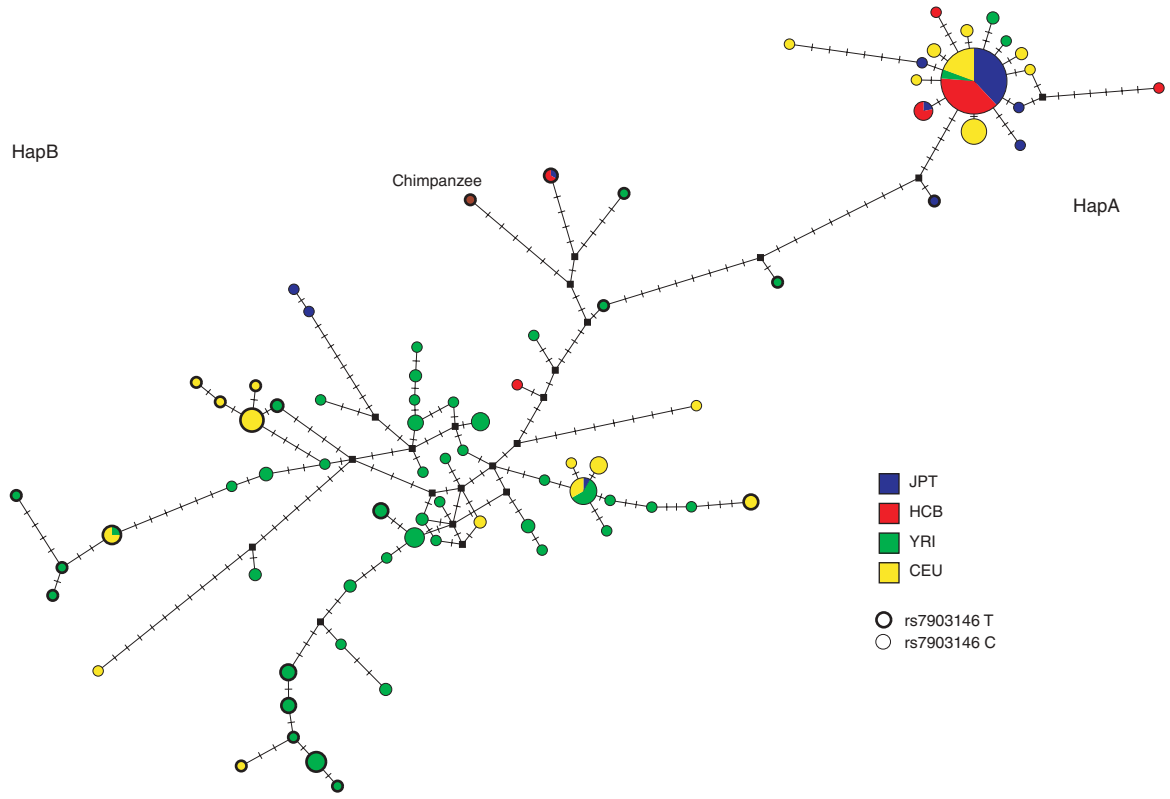

Figure 1 Median-joining (MJ) network[30] describing the evolutionary relationships between 78 distinct haplotypes, inferred from the genotypes of 63 SNPs and one microsatellite (DG10S478) from the *TCF7L2* exon 4 LD block. Each haplotype is represented by a circle whose area reflects the overall number of copies observed and whose color-coding indicates the frequency of the haplotype in the HapMap groups. Lines between circles represent mutational evolutionary pathways between haplotypes reconstructed by the MJ algorithm; line length is proportional to the number of differences between haplotypes. Filled squares represent non-sampled haplotypes that were reconstructed by the MJ algorithm as evolutionary intermediaries between observed haplotypes. The network shows two basic clusters of haplotypes: first, the relatively diverse cluster HapB, to which about 90% of the YRI haplotypes belong. As all the type 2 diabetes risk haplotypes, now defined by rs7903146 T, also belong to HapB, we refer to this set of haplotypes as $HapB_{T2D}$ (circles with thick outlines). $HapB_{T2D}$ is not a monophyletic lineage. Indeed, the relatively widespread distribution of rs7903146 T in the network is consistent with it being the ancestral allele (that is, identical to the chimpanzee reference sequence), with ample time to recombine with a variety of haplotype backgrounds, particularly in the relatively diverse West African population. The second cluster, HapA, contains a more homogeneous set of haplotypes that show a star-like pattern of diversity suggestive of a positive selective sweep on a monophyletic lineage. HapA accounts for 10%, 58% and 95% of the YRI, CEU and CHB+JPT HapMap samples, respectively. The long branch that separates HapA from HapB represents a set of 20 strongly correlated SNPs, any one of which can be used as a surrogate for HapA.

follow up on these observations, we tested for association of HapA and $HapB_{T2D}$ to BMI in five diabetes patient groups and six control groups (**Table 2**). Here, $HapB_{T2D}$ represents haplotypes that carry rs7903146 T, whereas HapA represents haplotypes with rs10885406 A and rs7903146 C. This definition of HapA is almost perfectly correlated ($r^2 = 0.96$) with that used in the preceding selection analyses in the combined HapMap data. Combining results from five type 2 diabetes patient groups using the Mantel-Haenszel model, HapA is associated with increased BMI (1.4% per copy, $P = 0.0014$), whereas $HapB_{T2D}$ is associated with decreased BMI (–1.3% per copy, $P = 0.0016$). The results for HapA and $HapB_{T2D}$ are not independent, as about 80% of the haplotypes in populations of European ancestry are either HapA or $HapB_{T2D}$. The estimated effect for haplotypes that are neither HapA nor $HapB_{T2D}$ is intermediate, but there is insufficient power to distinguish it clearly from the effects of HapA or $HapB_{T2D}$. The negative association between $HapB_{T2D}$ and BMI in affected individuals does not necessarily imply that $HapB_{T2D}$ leads to reduced BMI, as it could simply reflect the joint independent risk of BMI and $HapB_{T2D}$ for type 2 diabetes. When we combined results from the six control groups, we observed a weaker association in the same direction as in the affected individuals: an increase of 0.39% per copy of HapA ($P = 0.075$) and a decrease of 0.44% per copy of $HapB_{T2D}$ ($P = 0.092$).

The estimated effects of the haplotypes on BMI are consistently stronger for male affected individuals (1.9% per copy of HapA; $P = 0.00041$) and male controls (0.63% per copy; $P = 0.041$). Corresponding results for $HapB_{T2D}$ are –1.8% ($P = 0.00049$) and –0.45% ($P = 0.18$), respectively (**Table 2**). Overall, the association between the *TCF7L2* variants and BMI seems real, albeit modest in the controls. Indeed, all reported *P* values are two-sided and thus slightly conservative for replication groups, where it may be appropriate to test only for increased BMI with HapA and decreased BMI with $HapB_{T2D}$. We are presently unable to determine whether the effect of *TCF7L2* haplotypes on BMI is driven by HapA or $HapB_{T2D}$ or whether they exert opposing effects on BMI. Also, there may be differences in the magnitude of effect among the groups (notwithstanding that some estimates have large standard errors).

As BMI is a major risk factor for type 2 diabetes, it is intriguing, and highlights the complexity of the disease[10], that $HapB_{T2D}$ is associated with reduced BMI. This suggests that individuals who acquire type 2

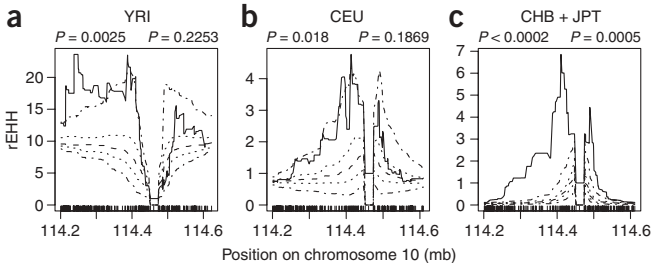

Figure 2 The LRH selection test at increasing physical distance from the core haplotype region. Shown are (**a**) YRI, (**b**) CEU and (**c**) East Asian HapMap groups. The solid line denotes observed rEHH values for HapA. The broken lines denote the 0.025, 0.25, 0.5, 0.75 and 0.975 quantiles from the distribution of rEHH values expected under neutral evolution, based on at least 5,000 coalescent simulations for each HapMap group. Simulations were conditioned on SNP minor allele frequencies, estimated recombination rates[28] and best-fit demographic settings[6] for the HapMap groups. Only simulated data sets that yielded a core haplotype at the same frequency as HapA in the three HapMap groups (0.10, 0.58 and 0.95, respectively) were considered. *P* values indicate the rank percentile of the observed aggregate rEHH value among the expected aggregate values. All three groups show significant *P* values in the region to the left of the core haplotype region, although the deviation from neutral expectations is marginal for the CEU sample. Only the East Asian group yielded a significant *P* value to the right of the core haplotype region. Similar results were obtained using neutral simulations based on other standard demographic models (**Supplementary Fig. 3** online).

diabetes through $HapB_{T2D}$ may have a disease different from that of those who acquire it through obesity. Conversely, since HapA is associated with increased BMI, it is legitimate to ask whether HapA and $HapB_{T2D}$ might be pulling individuals toward opposite ends of a distribution of physiological function with relation to their impact on BMI, providing distinct routes toward a common end, type 2 diabetes. Pooling results from the five diabetes case-control sets, HapA was estimated to have a slightly, but not significantly, higher risk than haplotypes that are neither $HapB_{T2D}$ nor HapA (RR = 1.06, $P = 0.22$). Thus, $HapB_{T2D}$ remains the only clear-cut risk factor for type 2 diabetes at this locus.

Given the association of HapA and $HapB_{T2D}$ to BMI in males, and noting a recent report of lower insulin secretion in $HapB_{T2D}$ carriers with type 2 diabetes[3], we examined their relationship to 14 metabolic traits in a subset of 918 Icelandic controls. Two of these traits, the fasting plasma concentrations of the hormones ghrelin and leptin, showed nominally significant differences (a decrease in the level of ghrelin and an increase in the level of leptin) by HapA copy number in males ($P = 0.0058$ and $P = 0.025$, respectively; **Table 3**). A suggestive increase in the fasting plasma concentration of insulin per HapA copy was also observed in males ($P = 0.09$). No such associations were observed for $HapB_{T2D}$ (**Supplementary Table 2** online). While these results are not significant at the 0.05 level after strict correction for the number of tests performed, they are suggestive. In particular, given the stronger association of HapA to BMI in males, it is noteworthy that the effects are observed only in males.

Ghrelin and leptin are of interest in the context of type 2 diabetes and the past selective forces that may have acted on variants of *TCF7L2* because of their role in the short-term neuroendocrine regulation of appetite and the long-term regulation of fat storage and energy metabolism[11]. In the long-term, fasting concentrations of ghrelin and leptin are strongly influenced by BMI such that weight gain increases the basal level of leptin and reduces the basal level of ghrelin[12]. Nonetheless, the strongest relationships we observed, HapA with ghrelin and the ghrelin-to-leptin ratio (GLR) in males, remain nominally significant even when adjusted for the impact of BMI ($P = 0.015$ and $P = 0.016$, respectively), whereas the weaker relationship between HapA and BMI disappears when adjusted for the impact of these traits.

Further studies are needed to confirm the impact of HapA on energy metabolism and to determine whether this phenotype contributed to the enhanced fitness of HapA in the past. Indeed, *TCF7L2* encodes a transcription factor and might thus influence multiple

Table 2 Association of HapA and $HapB_{T2D}$ with BMI[a] in individuals with type 2 diabetes and controls

		Individuals with type 2 diabetes				Controls			
		$HapB_{T2D}$		HapA		$HapB_{T2D}$		HapA	
Cohort	Subgroup (affected individuals/controls)	Effect (s.e.m.)	*P* value	Effect (s.e.m.)	*P* value	Effect (s.e.m.)	*P* value	Effect (s.e.m.)	*P* value
Iceland	Both sexes (1,146/9,222)	–0.015 (0.007)	0.036	0.021 (0.007)	0.0024	–0.0049 (0.0039)	0.21	0.0058 (0.0035)	0.094
	Males (676/3,992)	–0.014 (0.008)	0.080	0.018 (0.008)	0.026	–0.0037 (0.0049)	0.45	0.0069 (0.0044)	0.12
Philadelphia	Both sexes (326/497)	–0.002 (0.013)	0.88	0.007 (0.014)	0.63	–0.0145 (0.0138)	0.29	0.0163 (0.0123)	0.19
	Males (230/310)	0.001 (0.014)	0.94	0.005 (0.015)	0.76	–0.0123 (0.0167)	0.46	0.0091 (0.0145)	0.53
Denmark A	Females (227/507)	0.014 (0.014)	0.31	–0.005 (0.013)	0.71	–0.0054 (0.0113)	0.63	0.0106 (0.0105)	0.31
Denmark B	Both sexes (1,129/2,400)	–0.029 (0.008)	0.00017	0.018 (0.007)	0.016	–0.0016 (0.0048)	0.73	0.0006 (0.0043)	0.88
	Males (684/1,138)	–0.035 (0.009)	0.00011	0.028 (0.009)	0.0013	–0.0007 (0.0062)	0.91	0.0012 (0.0055)	0.83
West Africa	Both sexes (621/439)	–0.002 (0.011)	0.86	–0.013 (0.018)	0.49	–0.0296 (0.0148)	0.045	0.0220 (0.0224)	0.32
	Males (364/262)	–0.006 (0.014)	0.66	–0.006 (0.024)	0.79	–0.0332 (0.0201)	0.10	0.0570 (0.0294)	0.055
Germany[b]	Both sexes (0/1942)	–	–	–	–	0.0001 (0.0071)	0.99	–0.0027 (0.0063)	0.67
	Males (0/950)	–	–	–	–	–0.0074 (0.0091)	0.41	0.0106 (0.0080)	0.19
Combined	Both sexes (3,449/15,007)	–0.0133 (0.0042)	0.0016	0.0139 (0.0044)	0.0014	–0.0044 (0.0026)	0.092	0.0039 (0.0022)	0.075
	Males (1,954/6,652)	–0.0180 (0.0052)	0.00049	0.0190 (0.0054)	0.00041	–0.0045 (0.0034)	0.18	0.0063 (0.0031)	0.041

[a]Multiple regression was performed on log-transformed BMI values. [b]A test of transmission disequilibrium was performed for HapA and $HapB_{T2D}$ in 985 German family trios with morbidly obese children (of which 446 involved only obese boys), uncovering a slight excess in the transmission of HapA to the obese children ($P = 0.44$ for both sexes and $P = 0.1$ for just boys). Conversely, there was a slight deficit in the transmission of $HapB_{T2D}$ to the obese children ($P = 0.3$ for both sexes, $P = 0.045$ for boys only). *P* values are two-sided in each instance.

Table 3 Association of HapA with 14 phenotypic traits linked to energy metabolism in Icelandic controls

	Both sexes ($N = 918$)		Males ($N = 419$)		Females ($N = 499$)	
Trait	Effect (s.e.m.)	*P* value[b]	Effect (s.e.m.)	*P* value[b]	Effect (s.e.m.)	*P* value[b]
Percentage body fat	0.195 (0.474)	0.68	1.498 (0.796)	0.06	–1.006 (0.574)	0.08
Waist-hip ratio	0.002 (0.004)	0.64	–0.003 (0.006)	0.62	0.006 (0.006)	0.28
Systolic blood pressure[a]	0.004 (0.007)	0.58	0.001 (0.009)	0.89	0.006 (0.009)	0.51
Diastolic blood pressure	0.668 (0.493)	0.18	0.236 (0.691)	0.73	1.074 (0.687)	0.12
Total cholesterol[a]	–0.002 (0.010)	0.81	0.013 (0.014)	0.33	–0.017 (0.014)	0.22
High-density lipoprotein[a]	–0.002 (0.013)	0.86	–0.016 (0.016)	0.30	0.011 (0.019)	0.58
Low-density lipoprotein	–0.009 (0.050)	0.86	0.056 (0.070)	0.43	–0.068 (0.068)	0.32
Triglyceride[a]	0.012 (0.027)	0.65	0.055 (0.039)	0.16	–0.027 (0.035)	0.45
Glucose	–0.003 (0.008)	0.69	–0.014 (0.011)	0.18	0.007 (0.010)	0.48
Creatinine	0.457 (0.613)	0.46	0.844 (0.931)	0.36	0.106 (0.779)	0.89
Insulin[a]	0.065 (0.037)	0.076	0.093 (0.055)	0.090	0.041 (0.048)	0.40
Leptin[a]	2.206 (2.723)	0.42	6.977 (3.111)	0.025	–2.082 (4.057)	0.61
Ghrelin[a]	–0.056 (0.033)	0.086	–0.116 (0.042)	0.0058	–0.003 (0.046)	0.95
Ghrelin-leptin ratio (GLR)[a]	–0.115 (0.069)	0.10	–0.281 (0.108)	0.0092	0.032 (0.081)	0.69

[a]Values for these traits were log-transformed for the association analysis. [b]*P* value adjusted for relatedness of individuals by simulations through population-wide genealogies.

phenotypic traits. However, variation in energy metabolism must have been important for survival throughout human evolution, and it may be more than coincidence that the age of HapA in each of the HapMap groups coincides approximately with the transition to agriculture—a move that may have brought serious nutritional challenges along with the rewards. Whether the selective advantage of HapA acted through effects on energy metabolism, we note that our findings contradict a key prediction of the thrifty-genotype hypothesis[13], insofar as $HapB_{T2D}$, a major genetic risk factor for type 2 diabetes, is negatively associated with BMI and is not the variant that contributed to adaptive evolution in the recent past.

METHODS

Genotype data. *TCF7L2* spans a 217-kb region on chromosome 10 (114374798–114592005, NCBI build 34), and its product is a high-mobility box–containing transcription factor that has a role in the Wnt signaling pathway and in type 2 diabetes. Our analyses are based on 348 SNPs from release 19 of the HapMap data that spanned the region 114200000 to 114620000 (420 kb) and were polymorphic in at least one of the four HapMap groups. To these data, we added 13 SNPs (which we had identified through sequencing and genotyped) and one microsatellite, DG10S478. As in our previous publication[1], allele X of DG10S478 is defined as all alleles larger than 0. All of these loci are contained within region 114397469–114489563 on chromosome 10, NCBI build 34. Out of the total of 361 SNPs, 236 were polymorphic in the combined East Asian group, 298 were polymorphic in the YRI group and 276 were polymorphic in the CEU group. **Supplementary Figure 1** shows the LD structure in the CEU, YRI and East Asian HapMap groups across a 545-kb region of chromosome 10 that contains the *TCF7L2* gene, and it shows the distribution of SNPs across the region and the relative positions of introns and exons.

Phased haplotypes were generated for the 60 CEU parents, 60 YRI parents and 90 East Asian individuals from the HapMap project. The phase of alleles in haplotypes was estimated using the EM algorithm[14], in combination with the family trio information for the CEU and YRI groups (where the genotypes from the 30 children in each of the groups were used to help infer the allelic phase of the haplotypes). To obtain haplotypes across the entire 420-kb region, phase was estimated in informative multiple overlapping segments, using an approach analogous to that applied in the phasing of entire chromosomes by the HapMap project[15].

New case-control groups used for type 2 diabetes association refinement studies. Three type 2 diabetes case-control data sets were used to replicate and refine the association initially detected in ref. 1. First, the Africa America Diabetes Mellitus study, which was originally designed as an affected sibling pair study with enrollment of available spouses as controls. It has since been expanded to include other family members of the affected pairs and population controls. Recruitment strategies and eligibility criteria for the families enrolled in this report have been described previously[1]. This West African case-control series consisted of individuals from the Yoruba (156 affected individuals, 250 controls) and Igbo (186 affected individuals, 113 controls) groups from Nigeria and the Akan (196 affected individuals, 54 controls) and Gaa-Adangbe (83 affected individuals, 31 controls) groups from Ghana, which have previously shown little evidence of significant population substructure[16].

The second set of individuals with type 2 diabetes were from the Steno Diabetes Center in Copenhagen ($N = 1{,}018$) and from the Inter99 population-based sample of 30- to 60-year-old individuals living in the greater Copenhagen area and sampled at Research Centre for Prevention and Health[17] ($N = 359$). This data set is referred to in the text as Denmark B. Diabetes and pre-diabetes categories were diagnosed according to the 1999 World Health Organization (WHO) criteria. A total of 1,149 of these individuals were genotyped. An effectively random subset ($N = 2{,}400$) of Danish controls with BMI measurements were obtained from the Inter99 collection. Informed written consent was obtained from all subjects before participation. The study was approved by the Ethical Committee of Copenhagen County and was in accordance with the principles of the Helsinki Declaration.

The third set of affected individuals with type 2 diabetes from Iceland was described previously[1] and was analyzed together with an expanded set of controls described in the next section.

Cohorts used to assess association to BMI. Descriptive statistics in this section are presented as mean ± s.d., with age measured in years and BMI measured in kg/m^2. The Icelandic BMI cohort ($N = 9{,}222$; age = 61.2 ± 16.6; BMI = 27.4 ± 5.2) was composed of individuals who participated in studies of the genetic etiology of cardiovascular and metabolic diseases. Most subjects were recruited as unaffected relatives of probands or as controls and did not have any history of cardiovascular diseases (CVD). Detailed information was gathered for all individuals who participated in these studies, including biometric obesity and fasting concentration of blood lipids and glucose. A familial-enriched subset of the CVD cohort ($N = 918$), selected for use in gene expression studies, was extensively measured for clinical markers, including fasting concentrations of ghrelin and leptin. The German control cohort was composed of unrelated underweight, normal-weight and obese students ($N = 558$; age = 24.97 ± 3.55; BMI = 20.93 ± 4.44) and unrelated parents of obese children ($N = 1384$; age = 42.75 ± 5.89; BMI = 30.40 ± 6.16), as detailed in ref. 18. The parents belong to 708 independent German families containing at least one obese child (BMI = 31.03 ± 5.95; age 14.02 ± 3.71). The association between *TCF7L2* variants

and BMI was also assessed in type 2 diabetes case-control groups from Denmark and Pennsylvania, both of which were described and analyzed in our previous publication[1].

Active ghrelin was measured in a set of 918 Icelandic population controls using ELISA (sensitivity was 2.5 fmol/ml) from Linco Research. Leptin was measured using ELISA (sensitivity at least 7.8 pg/ml) from R&D systems. All samples, standards, two controls (high and low) and unknown treated plasma samples were measured in duplicate. Blood lipids, glucose and other biomarkers were measured as previously described[19].

Analysis of association to diabetes. We used a likelihood ratio test to calculate two-sided *P*-values for single-marker association to type 2 diabetes. For the Denmark B case-control study, we attempted to genotype all individuals reported in **Table 1** for DG10S478, rs12255372 and rs7903146. The average yield was about 96%. As the markers were in strong LD, when the genotype of one marker was missing for an individual, the genotypes of the other two markers were used to provide partial information through a likelihood approach we have described and used previously[20]. This ensured that results for all three markers presented in **Table 1** were always based on the same individuals, allowing meaningful comparisons. The average yield of the three markers was about 89% for the West African study, and we used the same method to handle missing genotypes. Allele frequencies rather than carrier frequencies are presented for the markers. Allele-specific RR was calculated assuming a multiplicative model[21]. The results for the West African study were obtained by combining data and results from four ethnic groups using a Mantel-Haenszel model[22] in which each group was allowed to have different population frequencies for alleles, but alleles were assumed to have identical relative risks. Standard errors and confidence intervals were computed assuming that the log of the estimate of RR has a normal distribution.

Correction for relatedness. We tested the association of an allele to diabetes using the signed ('+' for excess and '–' for deficit in affected individuals) square root of a standard likelihood ratio statistic that has an asymptotic standard normal distribution under the null hypothesis when subjects are unrelated. However, some of the individuals in the Icelandic and West African case-control groups were related to each other (see above). The genotypes of closely related individuals are not independent, causing the s.d. of the aforementioned association test statistic to be >1, which, if not corrected for, would lead to *P* values that are anticonservative. An adjustment for relatedness was performed using a previously described procedure[1,20]. In short, we performed 100,000 simulations that took into account the known relationships of the individuals to obtain the actual s.d. of the test statistic under the null hypothesis of no association.

Analysis of association to BMI and metabolic traits. The association between BMI and HapA or $HapB_{T2D}$ was analyzed using multiple regression. Log(BMI) was taken as the response variable. The estimated number of copies of HapA (or $HapB_{T2D}$) carried by an individual was entered as an explanatory variable. Adjustment for sex and age was accomplished by including sex ($1 + age + age^2$) as explanatory variables in analyses that combined data from both sexes. Analyses for males included age and age^2 as explanatory variables. For the West African cohort, indicator variables for the different ethnic groups were also included as explanatory variables. Our haplotype analysis program NEMO[23] (which implements methods based on maximum likelihood) was used to estimate of number of copies of HapA or $HapB_{T2D}$ carried by individuals. Because of the strong LD between rs10885406 and rs7903146, the two SNPs used to define HapA, there is little uncertainty in phase for most individuals. Results from different cohorts were combined using a Mantel-Haenszel-like model[22]. Specifically, for testing, a weighted average of the *t* statistics was calculated using weights proportional to the inverse of the standard errors of the estimated effects. The combined estimated effect was computed as a weighted average of the individual estimated effects using weights proportional to the inverse of the standard errors squared. Missing genotypes and correction for relatedness of individuals were performed using methods like those we described above for the diabetes association analysis. Similar association analyses were performed for other metabolic traits.

Alignment of the human and chimpanzee reference sequences. An alignment of the reference human and chimpanzee sequences, between positions 114274597 and 114674597 (NCBI human build 34) was obtained from the Ensembl genome browser. This alignment was used to determine the ancestral allelic states of the SNPs from the HapMap project, the SNPs typed by us and the microsatellite DG10S478.

Statistical tests of positive selection. Evidence for the impact of natural selection on the *TCF7L2* exon 4 LD block was examined by applying two different methods to the HapMap project groups. In these analyses, the Japanese (JPT) and Chinese (CHB) HapMap groups were combined to form a single East Asian group (CHB+JPT). First, we examined whether the degree of population divergence in haplotype frequencies among the HapMap groups exceeded expectations based on neutral evolution for the exon 4 LD block. Under neutrality, allele and haplotype frequency differences between populations are shaped by the counteracting forces of genetic drift, gene flow and mutation. The range of expected outcomes is constrained by the demographic history experienced by the populations. Thus, when the observed divergence between populations is in the upper extreme of the expected range (or outside it), the neutral model can be rejected in favor of one in which population differences in the intensity selective forces have caused an unusual degree of divergence in the frequency of particular variants[24]. In cases where positive selection has driven a variant to high frequency in a subset of the populations tested, it is expected that unusually great divergence between populations will be accompanied by unusually small variation within the populations that have experienced the most selection. We used the analysis of molecular variance (AMOVA)[25] approach to calculate the distribution of genetic diversity within and among populations in terms of haplotype frequencies. The expected range of divergence between populations can be determined either by means of a comparison with a large set of putatively neutral genomic regions from the same set of samples or by means of a comparison with multiple simulated data sets. The latter approach is more effective when dealing with haplotype data from LD blocks. On the one hand, it would be hard to match comparative data from other parts of the genome with relation to recombination rates. On the other hand, there is reason to believe that genomic regions within LD blocks are more gene-rich and therefore less likely to have experienced neutral evolution than regions that have weaker LD[15]. Coalescent simulations were performed to evaluate the statistical significance of the observed F_{ST} values, when controlling for demography, recombination rates and minor allele frequencies of SNPs. The SNPs used for the F_{ST} selection test are listed in **Supplementary Table 3** online.

The second method used to detect signals of positive selection in *TCF7L2* is based on examining the pattern of diversity within populations. Under neutrality, there is an expected positive relationship between the frequency of an allele or haplotype, its age, the variability at linked sites and the extent to which linkage disequilibrium (LD) with other loci decays at increasing physical distance. Common alleles or haplotypes with unusually low diversity at linked sites and/or slow decay of LD with increasing physical distance represent likely targets of recent positive selection[5,7]. According to this scenario, a variant is driven so rapidly to high frequency by positive selection that insufficient time has passed for the accumulation of background mutational and recombinational diversity expected for a neutral variant with the same frequency. We used the long-range haplotype (LRH) test[5] to examine the decay of LD at increasing distance from HapA, defined with the boundaries of the exon 4 LD block (the core haplotype area). This test is based on the estimation of the relative extended haplotype homozygosity (rEHH), which measures the fragmentation of a putative selected haplotype, caused by recombination and mutation, relative to other haplotypes made up of alleles from the same loci, as the LD to increasingly distant loci is assessed.

An advantage to the rEHH statistic is that it evaluates the extended LD around a putative selected haplotype in relation to the most informative comparative data: namely, other haplotypes formed by alleles from the same loci in the same samples[5]. In order to determine whether the observed pattern of genetic diversity at *TCF7L2* could have been generated by neutral evolution, we performed coalescent simulations as described below. In order to determine whether the observed rEHH values could have been produced by the demographic scenarios tested, we summed rEHH values from the observed data

across the genomic region examined and compared this aggregate rEHH value with a null distribution obtained from multiple coalescent simulations. The proportion of simulations that yield aggregate rEHH values greater than or equal to the observed value can be used as an empirical *P* value with which to evaluate the null hypothesis.

We also sought to determine the rank of rEHH values obtained for variants in *TCF7L2* in a particular HapMap group relative to an empirical distribution of rEHH values obtained for all other SNP alleles of identical frequency in the same HapMap group. These calculations were performed using release 19 of the HapMap data (http://www.hapmap.org). To simplify comparisons between different genomic regions, we calculated a single integrated rEHH value for each allele (using an approach similar to ref. 8). The resulting integrated rEHH (irEHH) value represents the area beneath the line defined by the rEHH point estimates that are obtained as haplotypes are extended in both directions from the allele being tested (until the EHH value in both directions has fallen below 0.05). In order to make comparisons of irEHH values meaningful between regions with different rates of recombination, the positions of SNPs were defined in cM for these calculations (using recombination rate maps for phase I of the HapMap that were interpolated for the phase II HapMap data). Although a genome-wide distribution of irEHH values cannot be considered an appropriate null distribution to test for signals of positive selection[8,26], the rank of a particular irEHH value in the overall distribution does provide a valuable indication of how unusual the underlying pattern of diversity is in the genome.

A rough estimate of the age of a selected haplotype was obtained using the approach described in ref. 8, based on the formula $EHH = e^{-2rg}$, where EHH represents the probability that two randomly picked chromosomes are homozygous at a recombination distance r from a core haplotype region, given a common ancestor g generations ago. We assume a generation interval of 25 years. The age of the putative selected haplotype, HapA, was estimated only for the upstream region that showed the strongest signal of selection and was based on the average number of generations obtained for EHH values between 0.05 and 0.1.

Coalescent simulations. The Simcoal simulation software[27] was used to generate samples under the neutral coalescent in order to evaluate the statistical significance of values obtained from selection tests for the observed data. A major complication in the use of coalescent simulations to detect signals of natural selection in genetic data is that a population's demographic history has a major impact on the expected patterns of diversity within its gene pool. In particular, it is difficult to distinguish between the impact of a positive selective sweep at a locus and the impact of a rapid population expansion. In most previous studies that have used a coalescent-based approach, the solution to this problem has been to simulate genetic data under a relatively wide range of demographic scenarios (typically, constant size, population expansion and bottleneck with expansion) that could be thought to apply to the population from which the observed data was obtained. We follow this procedure using three simple models: constant size at $N_e = 10{,}000$, a recent instantaneous expansion from 10,000 to 10,000,000 (200 generations ago) and a dramatic bottleneck taking N_e from 10,000 to 800 (800 generations ago), a condition that lasts for 160 generations, after which there is a re-expansion of N_e to 10,000. We also applied demographic settings obtained from an important recent study[6], wherein a wide range of demographic parameter values for populations of African, European and East Asian ancestry was evaluated using the genome-wide collection of SNP genotypes from the HapMap project with the aim of identifying the best-fitting demographic settings. The latter are referred to throughout the text as the 'best fit' demographic model and are listed in **Supplementary Table 4** online.

Recombination rates between adjacent loci are also key parameters in coalescent simulations involving multiple linked loci. We set recombination rates between adjacent SNPs within *TCF7L2*, according to values available from HapMap project website. These recombination rates were inferred on the basis of a population genetics approach[28] using release 16c of the HapMap data, which contained roughly 25% of the SNPs available in release 19. We note that the direct estimates of recombination rates between microsatellites in the deCODE genetics map[29] are very similar to the inferred rates from the HapMap data for this region. Additional SNPs from release 19 of the HapMap and other SNPs genotyped by us were incorporated into the existing recombination map by means of interpolation using the physical map positions of the SNPs. The resulting recombination map is shown in **Supplementary Table 5** online.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank all the individuals that provided DNA samples and other information that made this study possible. A.H., H.S. and J.H. were supported by the German National Genome Network. O.P., T.H., G.A., K.B.-J. and T.J. were supported by the Danish Medical Research Council, the Danish Diabetes Association and the European Economic Community (EUGENE 2 LSHM-CT-2004-512013). Support for the Africa America Diabetes Mellitus (AADM) study is provided by NIH grant 3T37TW00041-03S2 from the Office of Research on Minority Health. This project is also supported in part by the National Center for Research Resources (NCRR), the National Human Genome Research Institute (NHGRI) and by the National Institute for Diabetes and Digestive and Kidney Diseases (grant DK-54001). Requests for materials should be addressed to A.H. (agnar@decode.is) or K.S. (kstefans@decode.is).

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Grant, S.F.A. *et al.* Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323 (2006).
2. Rotimi, C.N. *et al.* In search of susceptibility genes for type 2 diabetes in West Africa: the design and results of the first phase of the AADM study. *Ann. Epidemiol.* **11**, 51–58 (2001).
3. Florez, J.C. *et al. TCF7L2* polymorphisms and progression to diabetes in the Diabetes Prevention Program. *N. Engl. J. Med.* **355**, 241–250 (2006).
4. Groves, C.J. *et al.* Association analysis of 6,736 U.K. subjects provides replication and confirms *TCF7L2* as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes* **55**, 2640–2644 (2006).
5. Sabeti, P.C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
6. Schaffner, S.F. *et al.* Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576–1583 (2005).
7. Slatkin, M. & Bertorelle, G. The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics* **158**, 865–874 (2001).
8. Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
9. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, New Jersey, 1994).
10. O'Rahilly, S., Barroso, I. & Wareham, N.J. Genetic factors in type 2 diabetes: the end of the beginning? *Science* **307**, 370–373 (2005).
11. Arora, S. & Anubhuti. Role of neuropeptides in appetite regulation and obesity: a review. *Neuropeptides* **40**, 375–401 (2006).
12. Erdmann, J., Lippl, F., Wagenpfeil, S. & Schusdziarra, V. Differential association of basal and postprandial plasma ghrelin with leptin, insulin, and type 2 diabetes. *Diabetes* **54**, 1371–1378 (2005).
13. Neel, J.V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *Am. J. Hum. Genet.* **14**, 353–362 (1962).
14. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **39**, 1–38 (1977).
15. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
16. Adeyemo, A.A., Chen, G., Chen, Y. & Rotimi, C. Genetic structure in four West African population groups. *BMC Genet.* **6**, 38 (2005).
17. Andersen, G. *et al.* Studies of the association of the GNB3 825C>T polymorphism with components of the metabolic syndrome in white Danes. *Diabetologia* **49**, 75–82 (2006).
18. Hinney, A. *et al.* Melanocortin-4 receptor gene: case-control study and transmission disequilibrium test confirm that functionally relevant mutations are compatible with a major gene effect for extreme obesity. *J. Clin. Endocrinol. Metab.* **88**, 4258–4267 (2003).
19. Jonsdottir, L.S., Sigfusson, N., Gudnason, V., Sigvaldason, H. & Thorgeirsson, G. Do lipids, blood pressure, diabetes, and smoking confer equal risk of myocardial infarction in women as in men? The Reykjavik Study. *J. Cardiovasc. Risk* **9**, 67–76 (2002).
20. Amundadottir, L.T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38**, 652–658 (2006).
21. Falk, C.T. & Rubinstein, P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* **51**, 227–233 (1987).
22. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22**, 719–748 (1959).

23. Gretarsdottir, S. *et al.* The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat. Genet.* **35**, 131–138 (2003).
24. Beaumont, M.A. & Nichols, R.A. Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. B* **263**, 1619–1626 (1996).
25. Excoffier, L., Smouse, P.E. & Quattro, J.M. Analysis of molecular variance inferred from metric distances among DNA haplotypes - Application to human mitochondrial-DNA restriction data. *Genetics* **131**, 479–491 (1992).
26. Teshima, K.M., Coop, G. & Przeworski, M. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* **16**, 702–712 (2006).
27. Laval, G. & Excoffier, L. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* **20**, 2485–2487 (2004).
28. McVean, G., Awadalla, P. & Fearnhead, P. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–1241 (2002).
29. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
30. Bandelt, H.J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48 (1999).