Vol 462|17 December 2009|doi:10.1038/nature08625

ARTICLES

Parental origin of sequence variants associated with complex diseases

Augustine Kong[1], Valgerdur Steinthorsdottir[1]*, Gisli Masson[1]*, Gudmar Thorleifsson[1]*, Patrick Sulem[1], Soren Besenbacher[1], Aslaug Jonasdottir[1], Asgeir Sigurdsson[1], Kari Th. Kristinsson[1], Adalbjorg Jonasdottir[1], Michael L. Frigge[1], Arnaldur Gylfason[1], Pall I. Olason[1], Sigurjon A. Gudjonsson[1], Sverrir Sverrisson[1], Simon N. Stacey[1], Bardur Sigurgeirsson[2], Kristrun R. Benediktsdottir[3], Helgi Sigurdsson[4], Thorvaldur Jonsson[5], Rafn Benediktsson[6], Jon H. Olafsson[2], Oskar Th. Johannsson[4], Astradur B. Hreidarsson[6], Gunnar Sigurdsson[6], the DIAGRAM Consortium†, Anne C. Ferguson-Smith[7], Daniel F. Gudbjartsson[1], Unnur Thorsteinsdottir[1,8] & Kari Stefansson[1,8]

Effects of susceptibility variants may depend on from which parent they are inherited. Although many associations between sequence variants and human traits have been discovered through genome-wide associations, the impact of parental origin has largely been ignored. Here we show that for 38,167 Icelanders genotyped using single nucleotide polymorphism (SNP) chips, the parental origin of most alleles can be determined. For this we used a combination of genealogy and long-range phasing. We then focused on SNPs that associate with diseases and are within 500 kilobases of known imprinted genes. Seven independent SNP associations were examined. Five—one with breast cancer, one with basal-cell carcinoma and three with type 2 diabetes—have parental-origin-specific associations. These variants are located in two genomic regions, 11p15 and 7q32, each harbouring a cluster of imprinted genes. Furthermore, we observed a novel association between the SNP rs2334499 at 11p15 and type 2 diabetes. Here the allele that confers risk when paternally inherited is protective when maternally transmitted. We identified a differentially methylated CTCF-binding site at 11p15 and demonstrated correlation of rs2334499 with decreased methylation of that site.

The effect of sequence variants on phenotypes may depend on parental origin. The most obvious scheme, although not the only one[1], is imprinting in which the effect is limited to the allele inherited from a parent of a specific sex. Despite this, most reports of genome-wide association studies have treated the paternal and maternal alleles as exchangeable. This is understandable, as the information required is often unavailable, but it reduces the power of such studies to discover some susceptibility variants and underestimates the effects of others, contributing to unexplained heritability. Here we describe a method that allows us to determine the parental origin of haplotypes systematically even when the parents of probands are not genotyped. We use the results to discover associations that exhibit parental-origin-specific effects.

Determining parental origin

Long-range phasing allows for accurate phasing of Icelandic samples typed with Illumina BeadChips for regions up to 10 cM in length[2]. Two advances have been made since then, stitching and parental-origin determination. Genome-wide, long-range phasing was applied to overlapping tiles, each 6 cM in length, with 3-cM overlaps between consecutive tiles. For each tile, we attempted to determine the parental origins of the two phased haplotypes regardless of whether the parents of the proband were chip-typed. Using the Icelandic genealogy database, for each of the two haplotypes of a proband a search was performed to identify, among those individuals also known to carry the same haplotype, the closest relative on each of the paternal and maternal sides (Fig. 1). Results for the two haplotypes were combined into a robust single-tile score reflecting the relative likelihood of the two possible parental-origin assignments

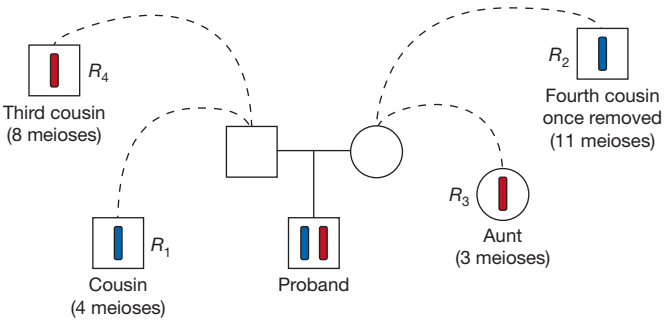

Figure 1 | An example of determination of parental origin. In blue and red are two phased haplotypes of a proband. Among other typed individuals, the closest paternal relative known also to carry the blue haplotype is R_1, a cousin; the corresponding maternal relative is R_2. For the red haplotype, a maternal aunt (R_3) carries the haplotype, and the closest known carrier on the paternal side is R_4. Because R_1 is a closer relative than R_2, and R_3 is a closer relative than R_4, the blue and red haplotypes are probably paternally and maternally inherited, respectively. The single-tile score (Methods) supporting this assignment is 0.194.

[1]deCODE genetics, Sturlugata 8, 101 Reykjavík, Iceland. [2]Department of Dermatology, [3]Department of Pathology, [4]Department of Oncology, [5]Department of Surgery, [6]Department of Endocrinology and Metabolism, Landspitali-University Hospital, 101 Reykjavik, Iceland. [7]Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge CB2 3EG, UK. [8]Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland.
*These authors contributed equally to this work.
†Lists of participants and affiliations appear at the end of the paper.

(with a score greater than zero supporting one assignment and a score less than zero supporting the other assignment; see Methods for details). We then tried to stitch the haplotypes from consecutive tiles together on the basis of sharing at the overlapping region. Stitching and parental-origin determination are complementary tasks. Specifically, if parental origin is determined with high confidence for one tile, the information can be propagated to other tiles through stitching (Supplementary Fig. 1a). Conversely, in cases in which the overlap between two adjacent tiles is homozygous for all SNPs, stitching can still be accomplished if parental origins can be determined for both tiles independently (Supplementary Fig. 1b). For haplotypes derived by stitching, a contig score for parental origin is computed by summing the individual single-tile scores.

After filtering based on various quality and yield criteria, 289,658 autosomal markers and 8,411 markers on chromosome X were used. Excluding those with no parent listed in the genealogy database or with a genotyping yield of less than 98%, 38,167 individuals, the majority typed with Illumina HumanHap300 or CNV370 BeadChips (Supplementary Information), were processed. For these individuals, 97.8% of the heterozygous genotypes were long-range phased, and in 99.8% of these the parental origin was determined. Overall, 3,841,331,873 heterozygous genotypes, or 97.7% of all heterozygous genotypes, had parental origin assigned. The data includes 2,879 typed trios. To evaluate the accuracy of our method empirically, a run was performed with the data for parents in these trios removed when determining parental origin. For 231,585,437 heterozygous genotypes in the probands/offspring, parental origin was determined both by our method and using the trio data directly, with 500,330 discrepancies, an error rate of 0.22%. Because the trios tested passed heritability checks in preprocessing, the error rate for individuals with fewer than two parents genotyped is probably higher. Nevertheless, the overall error rate is probably less than 0.4% (Supplementary Information).

Imprinting and disease association

Although many mechanisms can lead to parental-origin-specific association with a phenotype, sequence variants located close to imprinted genes are more likely to exhibit such behaviour a priori. Through two sources, ref. 3 and the Imprinted Gene Catalogue[4,5], we found forty-eight genes known to be imprinted in humans (Supplementary Table 1). Selecting regions that fall within 500 kilobases (kb) of any of these genes (NCBI build 36 of the human genome assembly) amounts to approximately 1% of the genome. The 500-kb threshold was chosen because imprinted genes often occur in clusters and the imprinting status of genes close to known imprinted genes is often undetermined. It is also known that a sequence variant can directly affect the function of a gene located some distance away. Among the 298,069 SNPs we processed, 3,840 fall within these selected regions.

By consulting the US National Institutes of Health Office of Population Studies catalogue of published genome-wide association studies[6] (accessed 25 April 2009), we intersected reported SNP–disease associations with $P < 5 \times 10^{-8}$ with the selected regions (Supplementary Table 2). After further restriction to diseases for which genome scans have been published based on Icelandic data, four associations remained. Three other SNP associations we were aware of that fall within the imprinted regions, one recently published for basal-cell carcinoma[7] and two new type 2 diabetes (T2D) variants discovered in the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) sample set (unpublished data, DIAGRAM Consortium; Supplementary Information), were also examined.

Association analysis

For each disease–SNP association, five tests were performed (Table 1). We performed a standard case-control test without taking parental origin into account to provide a baseline. Then we performed a

Table 1 | Parental-origin-specific analyses of disease-susceptibility variants

Disease, SNP [alleles]*		Standard case-control test		Tests of association with parental origins						
NCBI build 36 position, *N*	*M*, F_{con}	OR	P‡	Paternal allele§		Maternal allele§		2-d.f. test‖	Paternal vs maternal (case only)	
				OR	*P*	OR	*P*	*P*	n12:n21¶	*P*
Breast cancer, rs3817198† [C/T]										
C11 1,865,582,	34,909,									
1,803	0.303	1.04	0.36	1.17	0.038	0.91	0.11	0.0040	437:339	6.2×10^{-4}
Basal-cell carcinoma, rs157935 [T/G]										
C7 130,236,093,	37,041,									
1,118	0.676	1.23	1.8×10^{-5}	1.40	1.5×10^{-6}	1.09	0.19	3.8×10^{-6}	237:182	0.010
T2D, rs2237892 [C/T]										
C11 2,796,327,	34,706,									
1,468 (discovery)	0.925	1.19	0.044	1.14	0.24	1.24	0.071	0.095	81:90	0.51
783 (replication)		1.08	0.43	0.87	0.30	1.43	0.024	0.050	35:59	0.014
2,251 (combined)		1.15	0.043	1.03	0.71	1.30	0.0084	0.027	116:149	0.054
T2D, rs231362† [C/T]										
C11 2,648,047,	33,377,									
1,423 (discovery)	0.551	1.09	0.051	0.97	0.67	1.23	0.0010	0.0037	329:401	0.014
750 (replication)		1.10	0.073	1.00	0.99	1.22	0.011	0.037	158:191	0.098
2,173 (combined)		1.10	0.013	0.98	0.73	1.23	6.2×10^{-5}	2.6×10^{-4}	487:592	0.0032
T2D, rs4731702 [C/T]										
C7 130,083,924,	34,706,									
1,468 (discovery)	0.439	1.15	0.0018	1.07	0.24	1.23	6.4×10^{-4}	0.0013	335:374	0.17
783 (replication)		0.95	0.38	0.84	0.024	1.08	0.31	0.048	163:204	0.037
2,251 (combined)		1.08	0.039	0.99	0.79	1.17	0.0010	0.0041	498:578	0.022
T2D, rs2334499 [T/C]										
C11 1,653,425,	34,706,									
1,468 (discovery)	0.412	1.11	0.017	1.41	4.3×10^{-9}	0.87	0.020	3.5×10^{-9}	437:276	7.0×10^{-9}
783 (replication)		1.02	0.71	1.23	0.0055	0.84	0.023	0.0018	222:157	8.0×10^{-4}
2,251 (combined)		1.08	0.034	1.35	4.7×10^{-10}	0.86	0.0020	5.7×10^{-11}	659:433	4.1×10^{-11}

NCBI build 36 position is shown in terms of chromosome and base number. *N*, case sample size; *M*, control set size; F_{con}, control frequency (frequency of the risk allele in controls).
* The first allele is the risk allele on the basis of analyses that do not take into account parent of origin.
† Imputed allele probabilities were used.
‡ Genomic control was applied (true for all *P* values shown).
§ The effect of the paternally inherited allele was tested by comparing the corresponding alleles in cases with those in controls. The effect of the maternally inherited allele was tested similarly.
‖ The test assumes a multiplicative effect for the paternally and maternally inherited alleles, but allows the effects to be different under the alternative hypothesis when the null hypothesis of no effect is tested.
¶ To test directly whether the paternally and maternally inherited alleles have different effects, their allele frequencies were compared within the cases. Information for this test was mainly captured by the counts of the two types of heterozygote: n12 denotes the number of cases who have inherited allele 1 from the father and allele 2 from the mother, and n21 denotes the number of cases who have inherited allele 2 from the father and allele 1 from the mother.

case-control analysis separately for the paternally and maternally inherited alleles. A 2-d.f. test was applied to evaluate the joint effect. A multiplicative model was assumed for the two alleles, but the magnitude and direction of the effect were allowed to differ. Finally, the difference between the effects of the paternally and maternally inherited alleles was directly tested by comparing their allele frequencies within cases. The information for this test came mainly from the counts of the two types of heterozygote within cases (Supplementary Information).

Two of the seven associations examined, one with prostate cancer and another with coronary artery disease, did not exhibit parental-origin-specific effects (Supplementary Information and Supplementary Table 3). The five associations that did are presented here.

Breast cancer. Allele C of rs3817198 in the 11p15 region (Fig. 2) was reported[8] to be associated with breast cancer with an allelic odds ratio of OR = 1.07 ($P = 3 \times 10^{-9}$). This study included about 21,860 cases and 22,578 controls, allowing this modest effect to achieve genome-wide significance. A study[9] of 9,770 cases and 10,799 controls in the Cancer Genetic Markers of Susceptibility project reported odds ratios of 1.02 and 1.12 for heterozygous and homozygous carriers of the same variant, respectively. Using information in their supplementary material, we deduced a *P* value of 0.06. Marker rs3817198 is not on the Illumina chips used to type the majority of the Icelandic samples, but is included on the Illumina 1M BeadChips for which we have data on 124 trios. We used a single-track assay to type another 90 trios, giving a total of 214 trios with genotypes for rs3817198, which translates to a training set of 856 haplotypes. Adapting the statistical model used by IMPUTE[10], allele probabilities of rs3817198 were calculated for individuals with phased and parental-origin-determined haplotypes for this region (Supplementary Information). With the imputation results for 1,803 cases and 34,909 controls (Table 1), the standard case-control test gave a non-significant odds ratio of 1.04 ($P = 0.36$). However, when parental origin was taken into account, the paternally inherited allele showed a significant association (OR = 1.17, $P = 0.0038$). The direct test of parental-origin-specific effects that used only the case data was even more significant ($P = 6.2 \times 10^{-4}$). This is because the estimated effect of allele C when maternally inherited, although not significant ($P = 0.11$), is protective (OR = 0.91).

Basal-cell carcinoma. We recently identified association of allele T of rs157935, located at 7q32 (Fig. 3), with basal-cell carcinoma (OR = 1.23, $P = 5.7 \times 10^{-10}$)[7]. Limiting the analysis to samples for which parental origin could be determined, the paternally inherited allele was significantly associated with the disease (OR = 1.40, $P = 1.5 \times 10^{-6}$), but the effect of the maternally inherited allele, although it was in the same direction, was not significant (OR = 1.09, $P = 0.19$; Table 1). Tested directly, the effects of the paternally and maternally inherited alleles were significantly different ($P = 0.01$).

Type 2 diabetes. Allele C of rs2237892 in the maternally expressed gene *KCNQ1* was first observed to be associated with T2D in Asian populations[11,12]. The power to detect association in populations of European ancestry is low owing to the high frequency of the variant there (~93% compared with ~61% in Asians), but the association has nonetheless been replicated[11,12]. In the T2D samples we have previously used in genome scans (Table 1) including 1,468 cases, none of the tests involving parental origin were significant for rs2237892. However, with the addition of another 783 patients, giving a total of 2,251 cases (Supplementary Information), allele C was significantly associated with the disease (OR = 1.30, $P = 0.0084$) when maternally transmitted, whereas the results for the paternally inherited allele were flat (OR = 1.03, $P = 0.71$).

Through a meta-analysis of eight T2D genome-wide scans of DIAGRAM sample sets with additional follow-up (Supplementary Information), allele C of rs231362 was shown to associate with the disease (OR = 1.08, $P = 3 \times 10^{-13}$). Marker rs231362 is also located in *KCNQ1* (Fig. 2), but it is not substantially correlated with rs2237892 (correlation coefficient, $r^2 = 0.002$). Also, it is not on any of the Illumina chips used. A training set of 912 haplotypes, created through single-track-assay genotyping of 228 trios, was used for imputation of rs231362 into the Icelandic samples. Using the imputed results, the standard case-control test gave an odds ratio

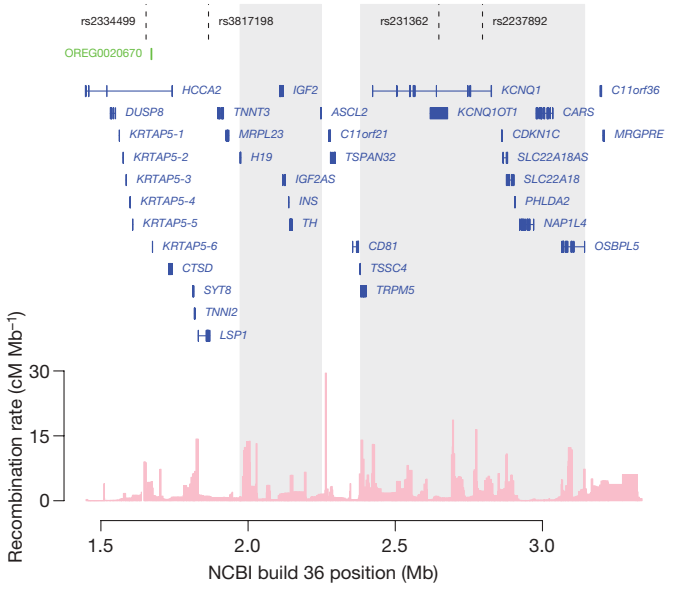

Figure 2 | Chromosome 11p15 locus. Markers associated with T2D (rs2334499, rs231362 and rs2237892) and breast cancer (rs3817198) are indicated. The two regions containing clusters of imprinted genes are shaded. The location of the CTCF-binding region studied (OREG0020670) and gene annotations were taken from the University of California, Santa Cruz, genome browser (http://genome.ucsc.edu/). Estimated recombination rates, from the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/), are plotted to reflect the linkage disequilibrium structure in the region. Mb, megabase.

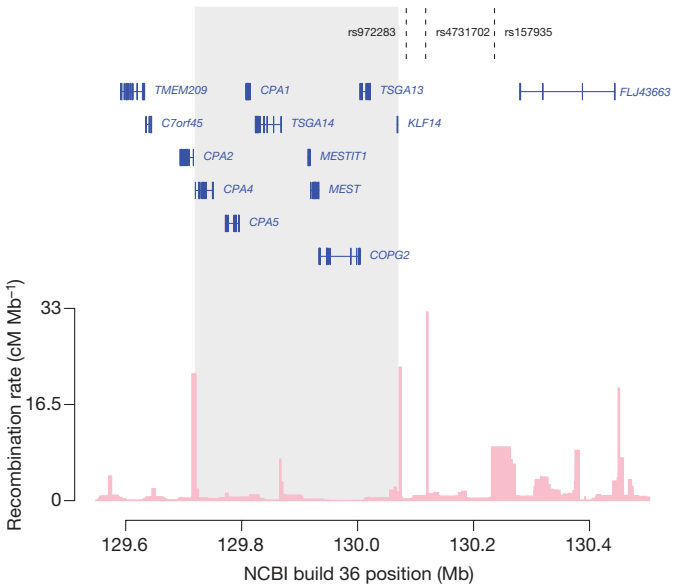

Figure 3 | Chromosome 7q32 locus. Markers associated with T2D (rs4731702 and rs972283 ($r^2 = 1$; HapMap CEU)) and basal-cell carcinoma (rs157935) are indicated. The region containing the known imprinted genes is shaded. Gene annotations were taken from the University of California, Santa Cruz, genome browser. Estimated recombination rates (from HapMap) are plotted to reflect the linkage disequilibrium structure in the region.

of 1.10 ($P = 0.013$). The effect, however, appears to be limited to the maternally inherited allele (OR = 1.23, $P = 6.2 \times 10^{-5}$).

Another association with T2D in DIAGRAM samples involves allele C of rs4731702 at 7q32 (OR = 1.07, $P = 2 \times 10^{-10}$; Fig. 3). In our combined Icelandic samples, the association was again restricted to the maternally inherited allele (OR = 1.17, $P = 0.0010$; OR = 0.99, $P = 0.79$ for the paternally inherited allele).

Evaluating the seven known susceptibility variants jointly (the five highlighted above plus the two variants for prostate cancer and coronary artery disease), the test for no parental-specific effect for all gave a P value of $<5 \times 10^{-6}$. Also, an analysis of false-discovery rate[13] indicates that it is likely that at least four of the five highlighted variants have true parental-origin-specific effects (Supplementary Information).

A new diabetes susceptibility variant. Properly evaluating the statistical significance of the susceptibility variants described above requires adjusting for relatedness of the participants using the method of genomic control[14]. This required us to perform genome scans for these diseases (Supplementary Table 4 gives parental-origin test results for established susceptibility variants located outside the selected regions). The T2D scan performed with the initial sample set (Supplementary Information and Supplementary Fig. 2) gave a striking result (Table 1). Allele T of rs2334499, at 11p15 (Fig. 2), showed such a weak association (OR = 1.11, $P = 0.017$) in the standard case-control test that it does not stand out in a genome-wide scan. However, taking into account parental origin, both the paternally inherited allele (OR = 1.41, $P = 4.3 \times 10^{-9}$) and the 2-d.f. test ($P = 3.5 \times 10^{-9}$) were genome-wide significant. Most notably, the maternally inherited allele also showed nominally significant association, but the effect of allele T was protective (OR = 0.87, $P = 0.020$). Tested directly, the difference between the effects of the paternally and maternally inherited alleles was also genome-wide significant ($P = 7.0 \times 10^{-9}$). This SNP falls within 350 kb of a large cluster of imprinted genes, making the results even more compelling. However, the observation that allele T is protective when maternally inherited required replication. For this, we used an additional set of 783 chip-typed T2D cases. All tests involving parental origin were significantly replicated. For the combined analysis of the two sample sets (Supplementary Information and Supplementary Fig. 3), the paternally inherited allele had an odds ratio of 1.35 ($P = 4.7 \times 10^{-10}$) and the maternally inherited allele had an odds ratio of 0.86 ($P = 0.0020$). The 2-d.f. test and the paternal-versus-maternal test gave P values of 5.7×10^{-11} and 4.1×10^{-11}, respectively.

As there are known examples in an imprinted setting where the paternal and maternal alleles interact[15], we tested rs2334499 for an interactive effect. This test was not significant ($P > 0.4$; Supplementary Information) indicating that the multiplicative model provides an adequate fit. Specifically, in comparison with CT (first allele paternal, second allele maternal), CC, TT and TC have relative risks of 1.17, 1.35 and 1.57, respectively.

The transmitted maternal allele has an effect in all four T2D variants in Table 1. Because prenatal maternal conditions may be a factor in conferring risk on the offspring, we examined the role of the non-transmitted maternal allele. No significant effect was observed (Supplementary Information).

Imprinted regions at 11p15 and 7q32

Imprinted genes at 11p15.5 fall into two clusters, *H19*/*IGF2* and *KCNQ1* (Fig. 2), regulated through separate imprinting control regions, each of which controls expression of a number of genes within the cluster[16]. The *H19*/*IGF2* imprinting control region is regulated through a differentially methylated region that is normally methylated only on the paternal chromosome. Binding of the insulator protein CTCF in the imprinting control region is permitted only on the unmethylated maternal chromosome, resulting in expression of *IGF2* only from the paternal methylated chromosome and expression of *H19* from the maternal chromosome[17]. The breast cancer paternally associated marker rs3817198 resides within *LSP1*, 100 kb downstream of *H19* and within the same linkage disequilibrium block. The effect of this marker on breast cancer could thus be through the *H19*/*IGF2* imprinted locus. Loss of imprinting at the *H19*/*IGF2* locus, resulting in activation of *IGF2* expression, has been reported in a number of different tumour types[18]. Furthermore, loss of imprinting at the *H19*/*IGF2* locus in normal tissue has also been shown to indicate a predisposition to colorectal cancer[18].

The *KCNQ1* cluster is regulated through an imprinting control region located in the promoter region of *KCNQ1OT1*, a paternally expressed non-coding antisense RNA. Hypermethylation of the maternal allele results in monoallelic activity of the neighbouring maternally expressed protein-coding genes. The two T2D-associated markers at this locus, rs231362 and rs2237892, are both located within the maternally expressed *KCNQ1*, consistent with the risk associations, rs231362 also residing within the *KCNQ1OT1* antisense transcript (Fig. 2).

Although both the T2D marker rs2334499 and the breast cancer marker rs3817198 fall within 350 kb of imprinted genes, the region harbouring them has not been reported to be imprinted[19] (Fig. 2). Marker rs2334499 resides within the first intron of *HCCA2*, a gene which spans 300 kb containing several other genes (Fig. 2) including *KRTAP5-1* to *KRTAP5-6*, *DUSP8* and *CTSD*[20]. To determine whether genes in this region showed signs of imprinting, we performed allele-specific expression analysis of *HCCA2*, *CTSD* and *DUSP8* (Fig. 2), as well as three genes known to be imprinted in the 11p15.5 region (*IGF2*, *KCNQ1* and *KCNQ1OT1*), in RNA isolated from peripheral blood and adipose. Whereas allele-specific expression of *IGF2*, *KCNQ1* and *KCNQ1OT1* was confirmed in this data set, clear biallelic expression was seen for *HCCA2* and *DUSP8*. However, excess paternal expression could not be ruled out for *CTSD* (Supplementary Information and Supplementary Table 6).

The imprinted region at 7q32 consists of maternally expressed genes (*CPA4* and *KLF14*) flanking paternally expressed genes (*MEST* and *MESTIT1*) (Fig. 3). The T2D-associated marker rs4731702 is located 14 kb from the maternally expressed *KLF14* transcription factor[21] and only increases risk of T2D when carried on the maternal chromosome. The basal-cell carcinoma variant rs157935, conferring risk through the paternal allele, is located 170 kb telomeric to the imprinted region.

We previously[22] correlated SNP genotypes from the Illumina 300K chip with gene expression using RNA samples from adipose tissue ($N = 603$) and peripheral blood ($N = 745$). Here, taking parental origin into account, we re-evaluated the correlation between the six variants in Table 1 and expression of genes at the 7q32 and 11p15.5 loci. The T2D risk allele of rs4731702 at 7q32 correlated with lower expression of *KLF14* in adipose tissue ($P = 3 \times 10^{-21}$) when inherited maternally, but there was no effect when it was inherited paternally (Supplementary Table 7). Similar correlation was not seen in blood. Conversely, no strong correlation with parental-origin-specific gene expression was seen for the other disease-associated variants at 7q32 or 11p15.5 (Supplementary Table 7).

Methylation of a novel CTCF-binding site

Recent studies have mapped regions of CTCF-binding genome-wide for identification of insulator elements[23,24]. One of the sites identified (OREG0020670) is a 2-kb region located 17 kb centromeric to the T2D marker rs2334499 (Fig. 2 and Supplementary Fig. 4). We assessed the methylation status of this CTCF-binding region in DNA samples derived from peripheral blood, using bisulphite sequencing. We identified a differentially methylated region of 180 base pairs including seven CpG dinucleotides (Supplementary Fig. 4) where the ratio of 5-methyl cytosine (Cp) varied from around 0.1 to 0.6. Methylation at five of the seven CpG dinucleotides (CpG-1 to CpG-5; Supplementary Fig. 4) was highly correlated (Supplementary Table 9). The estimated Cp ratio was tested for correlation with SNPs in a two-megabase surrounding region. The most significant correlation was observed between methylation status at CpG-4 and

Table 2 | Correlation between methylation of a CTCF-binding region and the T2D risk-variant rs2334499

CpG dinucleotide	Percentage methylation (mean, s.e.)*	Effect†	P‡
CpG-1	22.9 (0.9)	−5.7	6.5×10^{-7}
CpG-2	15.3 (0.7)	−3.1	0.00055
CpG-3	13.3 (0.8)	−2.5	0.017
CpG-4	56.7 (0.9)	−8.4	2.6×10^{-13}
CpG-5	34.9 (0.9)	−6.7	6.8×10^{-8}
CpG-6	22.2 (0.6)	−1.0	0.24
CpG-7	52.9 (1.1)	−1.5	0.30

* Mean and standard error of the C/T allele ratio estimated by bisulphite sequencing of 168 individuals.
† Change in percentage methylation per allele T of rs2334499 carried.
‡ Significance of the correlation between methylation and rs2334499.

rs2334499, for which $P = 2.6 \times 10^{-13}$ (Table 2). Furthermore, correlation between rs2334499 and methylation of CpG-1 to CpG-5 was significant. For these five CpG dinucleotides, the T2D risk allele correlated with decreased methylation and this effect was observed regardless of whether the allele was inherited from the father or the mother. By contrast, neither the breast cancer variant nor the two other T2D markers at 11p15.5 showed any correlation with the methylation status of this CTCF-binding site.

Discussion

Being able to determine parental origin of alleles and haplotypes in large samples opens new avenues to study associations between sequence variants and human traits. Standard association analysis provides suboptimal power to discover disease susceptibility variants that exhibit parental-origin-specific effects. Even when association can be established, the true effect is underestimated. Marker rs2334499 did not gain serious attention even after the large collaborative effort of the DIAGRAM Consortium. However, its contribution to T2D, measured by the recurrent risks of siblings generated, is second only to that of the *TCF7L2* variant among the known susceptibility variants (Supplementary Information and Supplementary Fig. 2). Sequence variants, such as rs2334499, that can confer both risk and protection depending on parental origin can lead to balanced selection and as a result promote diversity.

Functional imprinting is extremely tissue and stage specific, and although some genes retain their imprinted status throughout life, the main role of imprinting is believed to be during prenatal growth and development. However, the associations of rs4731702 C with T2D and *KLF14* expression in adult adipose tissue, in both cases only when maternally inherited, strongly implicates this transcription factor as the disease gene.

We searched for evidence of epigenetic marks around the T2D risk variant rs2334499, as it is located some distance away from the established 11p15.5 imprinted genes. A CTCF-binding site in the region was found to be differentially methylated and the rs2334499 risk allele was shown to be correlated with decreased methylation. Given the well-established role of CTCF in imprinting, this new site could differentially influence the dosage of imprinted genes on the two parental chromosomes.

Despite their successes, genome-wide association studies have so far yielded sequence variants that explain only a small fraction of the estimated heritability of most of the human traits studied. Obvious contributors to the unexplained heritability, or 'dark matter', include rare variants not well tagged by common SNPs and common variants that have very small effects individually. Results presented here demonstrate that a portion of the heritability of some common/complex traits is hidden in more complex relations between sequence variants and the risks of these variants.

METHODS SUMMARY

Subjects. We used 38,167 Icelandic individuals who were genotyped using an Illumina SNP chip and processed for long-range phasing. See Supplementary Information for details of disease and control groups.

Genotyping. We performed genome-wide genotyping using various Illumina BeadChips. Individual genotyping of two SNPs was done using Centaurus assays.

Determination of parental origin. We used the Icelandic genealogy database to identify the closest relatives who shared a haplotype with the proband. Parental origin was then assigned to the two haplotypes of a proband on the basis of a computed score (Methods).

Data imputation. On the basis of a training set of trios, by adapting the statistical model used by IMPUTE[10] to our setting, we computed allele probabilities of the paternal and maternal chromosomes separately for samples for which an SNP was not genotyped.

Statistical analysis. For SNPs directly typed, we used likelihood-based procedures to study disease associations taking parental origin into account. For imputed SNPs, we used logistic regressions and *t*-tests. Genomic control was used to control for relatedness among subjects.

Methylation analysis. Bisulphite sequencing was used to estimate the level of methylation. For each CpG dinucleotide, we determined the methylation status by calculating the C/T allele ratio at that site.

Gene expression. We tested associated SNPs for correlation with expression of genes located in a one-megabase window centred on the variant, in a data set of expression measurements in whole blood and adipose tissue. The same expression library was used to determine parental origin of expression by using allele-specific probes where the parental origin of each allele was known.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 August; accepted 29 October 2009.

1. Rampersaud, E., Mitchell, B. D., Naj, A. C. & Pollin, T. I. Investigating parent of origin effects in studies of type 2 diabetes and obesity. *Curr. Diabetes Rev.* **4,** 329–339 (2008).
2. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genet.* **40,** 1068–1075 (2008).
3. Luedi, P. P. *et al.* Computational and experimental identification of novel human imprinted genes. *Genome Res.* **17,** 1723–1730 (2007).
4. Morison, I. M., Paton, C. J. & Cleverley, S. D. The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res.* **29,** 275–276 (2001).
5. Morison, I. M., Ramsay, J. P. & Spencer, H. G. A census of mammalian imprinting. *Trends Genet.* **21,** 457–465 (2005).
6. Hindorff, L. A., Junkins, H. A., Mehta, J. P. & Manolio, T. A. A Catalog of Published Genome-Wide Association Studies. *OPG: Catalog Published Genome-Wide Assoc. Studies* ⟨http://www.genome.gov/gwastudies⟩ (2009).
7. Stacey, S. N. *et al.* New common variants affecting susceptibility to basal cell carcinoma. *Nature Genet.* **41,** 909–914 (2009).
8. Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447,** 1087–1093 (2007).
9. Thomas, G. *et al.* A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (*RAD51L1*). *Nature Genet.* **41,** 579–584 (2009).
10. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genet.* **39,** 906–913 (2007).
11. Yasuda, K. *et al.* Variants in *KCNQ1* are associated with susceptibility to type 2 diabetes mellitus. *Nature Genet.* **40,** 1092–1097 (2008).
12. Unoki, H. *et al.* SNPs in *KCNQ1* are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nature Genet.* **40,** 1098–1102 (2008).
13. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100,** 9440–9445 (2003).
14. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
15. Georges, M., Charlier, C. & Cockett, N. The callipyge locus: evidence for the *trans* interaction of reciprocally imprinted genes. *Trends Genet.* **19,** 248–252 (2003).
16. Ideraabdullah, F. Y., Vigneau, S. & Bartolomei, M. S. Genomic imprinting mechanisms in mammals. *Mutat. Res.* **647,** 77–85 (2008).
17. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405,** 482–485 (2000).
18. Feinberg, A. P. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447,** 433–440 (2007).
19. Goldberg, M., Wei, M., Yuan, L., Murty, V. V. & Tycko, B. Biallelic expression of HRAS and MUCDHL in human and mouse. *Hum. Genet.* **112,** 334–342 (2003).
20. Authier, F., Metioui, M., Fabrega, S., Kouach, M. & Briand, G. Endosomal proteolysis of internalized insulin at the C-terminal region of the B chain by cathepsin D. *J. Biol. Chem.* **277,** 9437–9446 (2002).
21. Parker-Katiraee, L. *et al.* Identification of the imprinted *KLF14* transcription factor undergoing human-specific accelerated evolution. *PLoS Genet.* **3,** e65 (2007).
22. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).
23. Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128,** 1231–1245 (2007).

24. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* **19,** 24–32 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This project was funded in part by FP7-MC-IAPP Grant agreement no. 218071 (CancerGene) to deCODE genetics.

Author Contributions A.K. and K.S. planned and directed the research. A.K. wrote the first draft of the paper and, together with K.S., V.S., G.M., G.T. and U.T., wrote most of the final version. A.K. and G.M. designed the method to determine parental origin. G.M., with assistance from P.I.O., implemented the algorithm. D.F.G. wrote the code for association analysis taking parental origin into account and performed some initial analyses. P.S., S.B. and S.S. tabulated the established disease-associated variants and the regions known to harbour imprinted genes. V.S. and G.T. contributed to the analysis of the diabetes data and, together with A.K. and U.T., planned the follow-up association and functional studies. A.G., A.K. and M.L.F. imputed the untyped SNPs. S.N.S. and P.S. were responsible for the breast cancer and basal-cell carcinoma data. A.B.H., G.S. and R.B. provided clinical data for T2D, O.Th.J., T.J. and H.S. provided clinical data for breast cancer, and J.H.O., B.S. and K.R.B. provided clinical data for basal-cell carcinoma. The DIAGRAM Consortium provided the novel T2D-associated variants that are close to imprinted genes. Aslaug J., A.S., Adalbjorg J., K.Th.K. and S.A.G. performed the methylation and expression studies. A.C.F.-S. assisted in the interpretation of the results from the association and functional studies.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.K. (kong@decode.is) or K.S. (kstefans@decode.is).

DIAGRAM Consortium Benjamin F. Voight[1,2,3], Laura J. Scott[4], Valgerdur Steinthorsdottir[5], Christian Dina[6,7], Eleftheria Zeggini[8,9], Cornelia Huth[10,11], Yurii S. Aulchenko[12], Ryan P. Welch[4], Gudmar Thorleifsson[5], Laura J. McCulloch[13], Teresa Ferreira[9], Harald Grallert[10,11], Najaf Amin[12], Guanming Wu[14], Cristen J. Willer[4], Soumya Raychaudhuri[1,2,15], Shaun Purcell[1,2,16], Steve A. McCarroll[1,17], Claudia Langenberg[18], Oliver M. Hoffmann[19], Josée Dupuis[20], Lu Qi[21,22], Ayellet V. Segrè[1,17], Mandy van Hoek[23], Pau Navarro[24], Kristin Ardlie[1], Beverley Balkau[25,26], Rafn Benediktsson[27,28], Amanda J. Bennett[13], Roza Blagieva[29], Eric Boerwinkle[30], Lori L. Bonnycastle[31], Kristina Bengtsson Boström[33], Bert Bravenboer[34], Suzannah Bumpstead[8], Noël P. Burtt[1], Guillaume Charpentier[35], Peter S. Chines[31], Marilyn Cornelis[22], David J. Couper[36], Gabe Crawford[1], Alex S. F. Doney[37,38], Katherine S. Elliott[9], Amanda L. Elliott[1,17], Michael R. Erdos[31], Caroline S. Fox[39,40], Christopher S. Franklin[41], Martha Ganser[4], Christian Gieger[10], Niels Grarup[42], Todd Green[1,2], Simon Griffin[18], Christopher J. Groves[13], Candace Guiducci[1], Samy Hadjadj[43], Neelam Hassanali[13], Christian Herder[44], Bo Isomaa[45,46], Anne U. Jackson[4], Paul R. V. Johnson[47], Torben Jørgensen[48], Wen H. L. Kao[49], Norman Klopp[10], Augustine Kong[5], Peter Kraft[21], Johanna Kuusisto[50], Torsten Lauritzen[51], Man Li[52], Alouisius Lieverse[53], Cecilia M. Lindgren[9], Valeriya Lyssenko[54], Michel Marre[55,56], Thomas Meitinger[10], Kristian Midthjell[57], Mario A Morken[31], Narisu Narisu[31], Peter Nilsson[54], Katharine R. Owen[13], Felicity Payne[8], John R. B. Perry[58,59], Ann-Kristin Petersen[10], Carl Platou[57], Christine Proença[6], Inga Prokopenko[9,13], Wolfgang Rathmann[60], N. William Rayner[9,13], Neil R. Robertson[9,13], Ghislain Rocheleau[61,62,63], Michael Roden[44,64], Michael J. Sampson[65], Richa Saxena[1,2,66], Beverley M. Shields[58,59], Peter Shrader[67,68], Gunnar Sigurdsson[27,28], Nicholas Smith[8], Thomas Sparsø[42], Klaus Strassburger[60], Heather M. Stringham[4], Qi Sun[21], Amy J. Swift[31], Barbara Thorand[10], Jean Tichet[69], Tiinamaija Tuomi[45,70], Rob van Dam[22], Thijs van Herpt[23,53], G. Bragi Walters[5], Michael N. Weedon[58,59], Jacqueline Witteman[12], Richard N. Bergman[71], Stephane Cauchi[6], Francis S. Collins[72], Anna L. Gloyn[13], Ulf Gyllensten[73], Torben Hansen[42,74], Winston A. Hide[19], Graham A. Hitman[75], Albert Hofman[12], David Hunter[21], Kristian Hveem[57,76], Markku Laakso[50], Karen L. Mohlke[77], Andrew D. Morris[37,38], Colin N. A. Palmer[37,38], Peter P. Pramstaller[78], Igor Rudan[41,79,80], Eric Sijbrands[23], Lincoln D. Stein[14], Jaakko Tuomilehto[81], Andre Uitterlinden[23], Mark Walker[82], Nicholas J. Wareham[18], Richard M. Watanabe[83], Goncalo R. Abecasis[4], Inês Barroso[8], Bernhard O. Boehm[29], Harry Campbell[41], Mark J. Daly[1,2], Jose C. Florez[1,2,3], Timothy M. Frayling[58,59], Leif Groop[54,70], Andrew T. Hattersley[58,59], Frank B. Hu[21,22], James B. Meigs[3,67], Andrew P. Morris[9], James S. Pankow[84], Oluf Pedersen[42,85,86], Rob Sladek[61,62,63], Unnur Thorsteinsdottir[5,87], H.-Erich Wichmann[10,11], James F. Wilson[41], Thomas Illig[10], Philippe Froguel[6,88], Cornelia M. van Duijn[12], Kari Stefansson[5,87], David Altshuler[1,2,3,17,66,89], Michael Boehnke[4], Mark I. McCarthy[9,13,90].

[1]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [2]Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [3]Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, USA. [5]deCODE genetics, Sturlugata 8, 101 Reykjavik, Iceland. [6]CNRS-UMR-8090, Institute of Biology and Lille 2 University, Pasteur Institute, F-59019 Lille, France. [7]INSERM, UMR915, CNRS, ERL3147, 44007 Nantes, France. [8]Wellcome Trust Sanger Institute, Hinxton CB10 1HH, UK. [9]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [10]Institute of Epidemiology, Helmholtz Zentrum Muenchen, 85764 Neuherberg, Germany. [11]Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität, 81377 Munich, Germany. [12]Department of Epidemiology, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands. [13]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford OX3 7LJ, UK. [14]Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto, Ontario M5G 0A3, Canada. [15]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [16]Stanley Center for Psychiatric Research, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [17]Department of Molecular Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. [18]MRC Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [19]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [20]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts 02118, USA. [21]Departments of Nutrition and Epidemiology, Harvard School of Public Health, 665 Huntington Avenue, Boston, Massachusetts 02115, USA. [22]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Avenue, Boston, Massachusetts 02115, USA. [23]Department of Internal Medicine, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands. [24]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh EH4 2XU, UK. [25]INSERM U780, F-94807 Villejuif, France. [26]University Paris-Sud, F-91405 Orsay, France. [27]Landspitali University Hospital, 101 Reykjavik, Iceland. [28]Icelandic Heart Association, 201 Kopavogur, Iceland. [29]Division of Endocrinology, Diabetes and Metabolism, Ulm University, 89081 Ulm, Germany. [30]The Human Genetics Center and Institute of Molecular Medicine, University of Texas Health Science Center, Houston, Texas 77030, USA. [31]National Human Genome Research Institute, National Institute of Health, Bethesda, Maryland 20892, USA. [33]R&D Centre, Skaraborg Institute, 541 30 Skövde, Sweden. [34]Department of Internal Medicine, Catharina Hospital, PO Box 1350, 5602 ZA Eindhoven, The Netherlands. [35]Endocrinology-Diabetology Unit, Corbeil-Essonnes Hospital, F-91100 Corbeil-Essonnes, France. [36]Department of Biostatistics and Collaborative Studies Coordinating Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [37]Diabetes Research Centre, [38]Pharmacogenomics Centre, Biomedical Research Institute, University of Dundee, Ninewells Hospital, Dundee DD1 9SY, UK. [39]National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts 01702, USA. [40]Division of Endocrinology, Diabetes, and Hypertension, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [41]Centre for Population Health Sciences, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, UK. [42]Hagedorn Research Institute, DK-2820 Gentofte, Denmark. [43]CHU de Poitiers, Endocrinologie Diabetologie, CIC INSERM 0801, INSERM U927, Université de Poitiers, UFR, Médecine Pharmacie, 86021 Poitiers Cedex, France. [44]Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. [45]Folkhälsan Research Center, FIN-00014 Helsinki, Finland. [46]Malmska Municipal Health Center and Hospital, 68601 Jakobstad, Finland. [47]DRWF Human Islet Isolation Facility and Oxford Islet Transplant Programme, University of Oxford, Old Road, Headington, Oxford OX3 7LJ, UK. [48]Research Centre for Prevention and Health, Glostrup University Hospital, DK-2600 Glostrup, Denmark. [49]Department of Epidemiology, Department of Medicine, and Welch Center for Prevention, Epidemiology, and Clinical Research, Johns Hopkins University, Baltimore, Maryland 21287, USA. [50]Department of Medicine, University of Kuopio and Kuopio University Hospital, FIN-70211 Kuopio, Finland. [51]Department of General Medical Practice, University of Aarhus, DK-8000 Aarhus, Denmark. [52]Department of Epidemiology, Johns Hopkins University, Baltimore, Maryland 21287, USA. [53]Department of Internal Medicine, Maxima MC, PO-Box 90052, 5600 PD Eindhoven, The Netherlands. [54]Department of Clinical Sciences, Diabetes and Endocrinology Research Unit, University Hospital Malmö, Lund University, 205 02 Malmö, Sweden. [55]Department of Endocrinology, Diabetology, Nutrition, Bichat-Claude Bernard University Hospital, Assistance Publique des Hôpitaux de Paris, 75877 Paris Cedex 18, France. [56]INSERM U695, Université Paris 7, 75870 Paris Cedex 18, France. [57]HUNT Research Center, Department of Community Medicine and General Practice, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway. [58]Genetics of Complex Traits, Institute of Biomedical and Clinical Science, Peninsula Medical School, University of Exeter, Magdalen Road, Exeter EX1 2LU, UK. [59]Diabetes Genetics, Institute of Biomedical and Clinical Science, Peninsula Medical School, University of Exeter, Barrack Road, Exeter EX2 5DW, UK. [60]Institute of Biometrics and Epidemiology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. [61]Department of Human Genetics, McGill University, Montreal H3H 1P3, Canada. [62]Department of Medicine, Faculty of Medicine, McGill University, Montreal H3A 1A4, Canada. [63]McGill University and Genome Quebec Innovation Centre, Montreal H3A 1A4, Canada. [64]Department of Medicine/Metabolic Diseases, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. [65]Department of Endocrinology and Diabetes, Norfolk and Norwich University Hospital NHS Trust , Norwich NR1 7UY, UK. [66]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [67]General Medicine Division, Massachusetts General Hospital, Boston, Massachusetts, USA. [68]Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. [69]Institut Interrégional pour la Santé, F-37521 La Riche, France. [70]Department of Medicine, Helsinki University Hospital, University of Helsinki, FIN-00290 Helsinki, Finland. [71]Department of Physiology and Biophysics, University of Southern California School of Medicine, Los Angeles, California

90033, USA. [72]National Institutes of Health, Bethesda, Maryland 20892, USA. [73]Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, S-751 85 Uppsala, Sweden. [74]University of Southern Denmark, DK-5230 Odense, Denmark. [75]Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London E1 1BB, UK. [76]Department of Medicine, The Hospital of Levanger, N-7600 Levanger, Norway. [77]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599-7264, USA. [78]Institute of Genetic Medicine, European Academy Bozen/Bolzano, Viale Druso 1, 39100 Bolzano, Italy. [79]Croatian Centre for Global Health, Faculty of Medicine, University of Split, Soltanska 2, 21000 Split, Croatia. [80]Institute for Clinical Medical Research, University Hospital ''Sestre Milosrdnice'', Vinogradska 29, 10000 Zagreb, Croatia. [81]Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki FIN-00300, Finland. [82]Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK. [83]Department of Preventive Medicine, Keck Medical School, University of Southern California, Los Angeles, California 90089-9001, USA. [84]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota 55454, USA. [85]Department of Biomedical Science, Panum, Faculty of Health Science, University of Copenhagen, 2200 Copenhagen, Denmark. [86]Faculty of Health Science, University of Aarhus, DK-8000 Aarhus, Denmark. [87]Faculty of Medicine, University of Iceland, 101 Reykjavík, Iceland. [88]Genomic Medicine, Imperial College London, Hammersmith Hospital, London W12 0NN, UK. [89]Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts 02144, USA. [90]Oxford National Institute for Health Research Biomedical Research Centre, Churchill Hospital, Old Road, Headington, Oxford OX3 7LJ, UK.

METHODS

Assignment of parental origin. Let H be a haplotype for a tile T. For a particular proband, $f(T, H)$ and $m(T, H)$ were calculated as the meiotic distances to the closest relatives on the paternal side and, respectively, the maternal side known to carry H. Descendants of the parents of the proband, for example siblings of the proband, were excluded from this calculation. Also, a value of 10,000 was assigned when no relatives carrying the haplotype was found. Let A and B be the two phased haplotypes of the proband. The single-tile score for parental origin was calculated as

$$\begin{aligned}\text{score}(T,A,B) &= \text{score}(T,A) - \text{score}(T,B)\\ &= [\log(1-2^{-m(T,A)}) - \log(1-2^{-f(T,A)})]\\ &\quad - [\log(1-2^{-m(T,B)}) - \log(1-2^{-f(T,B)})]\end{aligned}$$

A score that is greater than zero supports the assignment of A as the paternally inherited haplotype and B as the maternally inherited haplotype, whereas a score that is less than zero supports the reverse. Although it is not meant to be optimal in any formal sense, this system of scoring was chosen to have two properties. First, for the same absolute difference between $m(T, H)$ and $f(T, H)$, the absolute value of score(T, H) is higher when the lesser of $m(T, H)$ and $f(T, H)$ is smaller, thus giving more weight to situations in which a close relative who shared a haplotype is found. Second, the scoring was designed such that the result from one haplotype in one tile could not completely dominate the contributions from other haplotypes and adjacent tiles when results were combined (see below).

When haplotypes for n consecutive tiles, $T_1, \ldots, T_n$, could be stitched together to form $A = (A_1, \ldots, A_n)$ and $B = (B_1, \ldots, B_n)$, the contig score for parental origin assignment was calculated as

$$\text{contig\_score}(T_1, \ldots, T_n) = \sum_{i=1}^{n} \text{score}(T_i)$$

Parental origins were assigned on the basis of whether the contig score was greater than or less than zero. The accuracy of this procedure was evaluated using the trio test.