A common sex-dependent mutation in a *RET* enhancer underlies Hirschsprung disease risk

Eileen Sproat Emison[1]*, Andrew S. McCallion[1]*, Carl S. Kashuk[1], Richard T. Bush[1], Elizabeth Grice[1], Shin Lin[1], Matthew E. Portnoy[2], David J. Cutler[1], Eric D. Green[2,3] & Aravinda Chakravarti[1]

[1]*McKusick – Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA*
[2]*Genome Technology Branch and* [3]*NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA*

* These authors contributed equally to this work

The identification of common variants that contribute to the genesis of human inherited disorders remains a significant challenge. Hirschsprung disease (HSCR) is a multifactorial, non-mendelian disorder in which rare high-penetrance coding sequence mutations in the receptor tyrosine kinase RET contribute to risk in combination with mutations at other genes. We have used family-based association studies to identify a disease interval, and integrated this with comparative and functional genomic analysis to prioritize conserved and functional elements within which mutations can be sought. We now show that a common non-coding *RET* variant within a conserved enhancer-like sequence in intron 1 is significantly associated with HSCR susceptibility and makes a 20-fold greater contribution to risk than rare alleles do. This mutation reduces *in vitro* enhancer activity markedly, has low penetrance, has different genetic effects in males and females, and explains several features of the complex inheritance pattern of HSCR. Thus, common low-penetrance variants, identified by association studies, can underlie both common and rare diseases.

HSCR, or congenital aganglionosis with megacolon, occurs in one in every 5,000 live births. Heritability of HSCR is nearly 100% with clear multigenic inheritance. Although *RET* is the main HSCR gene implicated[1,2], mutations also occur in seven other genes involved in enteric development, specifically *ECE1*, *EDN3*, *EDNRB*, *GDNF*, *NRTN*, *SOX10* and *ZFHX1B*[3]. However, less than 30% of patients have mutations in these eight genes; additional HSCR-causing mutations in *RET* and/or other genes must therefore exist.

We recently identified a high-frequency HSCR-associated *RET* haplotype in an inbred Mennonite population[4]. However, this haplotype lacks a frank coding-sequence mutation. Association between HSCR and specific *RET* marker alleles has also been shown in other populations[5–7]. Importantly, even in families with established linkage to the *RET* locus on chromosome 10q11.21, *RET* coding mutations are frequently absent[1,2]. Additionally, we demonstrated that the *RET* Mennonite haplotype is also present in the general population and is overtransmitted to affected offspring[8], indicating that the same mutation might underlie HSCR in Mennonites and in the general outbred population. We advance two competing, but not mutually exclusive, hypotheses to explain these findings: first, one or more non-coding mutations at *RET*, and second, a mutation(s) in a second gene that is tightly linked to *RET*. The shared conjecture of these hypotheses is that a common HSCR-causing mutation lies in a functional sequence within or flanking the *RET* locus. Resolving these hypotheses can be accomplished by a congruence of association studies and a comprehensive catalogue of all functional elements within the greater *RET* locus.

Family-based association studies

Genome sequence data (http://www.genome.ucsc.edu/cgi-bin/hgGateway; build 35) identifies two additional genes in the 350-kilobase (kb) region surrounding *RET*. *GALNACT-2*, a chondroitin *N*-acetylgalactosaminyltransferase[9,10], contains eight exons spanning 46.8 kb and begins 9 kb from the last *RET* exon. Thirteen exons encode *RASGEF1A*, a predicted guanyl-nucleotide exchange factor that spans 72 kb and begins 65 kb 3′ of *RET*. To refine the association within this locus genetically, we initially genotyped 28 single-nucleotide polymorphisms (SNPs) spanning 175 kb in 126 HSCR-affected individuals and their parents, ascertained from the general outbred population (Table 1). The genomic interval encompasses *RET*, *GALNACT-2* and *RASGEF1A*.

Transmission disequilibrium tests (TDTs)[11] on each SNP showed statistically significant disease associations spanning a region immediately 5′ of *RET* through to *RASGEF1A* (Fig. 1a, Table 1). Specifically, 13 of 17 *RET* SNPs, 3 of 7 *GALNACT-2* SNPs and 2 of 4 *RASGEF1A* SNPs tested are significantly associated with HSCR (Table 1), reflecting the high background linkage disequilibrium in this region (data not shown). However, the greatest statistical significance, and, more importantly, the largest transmission distortions ($\tau \geq 0.7$), occurred among eight SNPs in a 27.6-kb segment from 4.2 kb 5′ of *RET* through to *RET* exon 2 (Fig. 1a). Within this region the highest association was within *RET* intron 1.

We performed and analysed three resequencing experiments to identify additional variants, with particular emphasis being given to multi-species conserved sequences (MCSs; see below) within the 27.6-kb region of highest association. Specifically, we identified the SNP RET+3 (marked with an asterisk in Fig. 1a) within MCS+9.7 by resequencing HSCR patients from families with demonstrated *RET*-linkage but no identified coding sequence mutations. TDTs of RET+3 in all 126 trios showed the largest transmission distortion ($\tau = 0.8$) and the highest statistical significance ($P = 10^{-11}$). When association tests are factored by offspring gender, a known risk factor in HSCR, RET+3 and the adjacent marker 1.1SfcI (3.3 kb away) are the only two SNPs demonstrating association in females. Two additional variants (rs2506005 and rs2506004) lie within MCS+9.7 and are located 76 nucleotides (nt) 5′ and 217 nt 3′ of RET+3, respectively; both are in complete linkage disequilibrium with RET+3 and each other. The HSCR-associated allele at each of

these additional SNPs is the ancestral allele. The RET+3:C allele is highly conserved in all nine mammalian species examined (Supplementary Fig. 1), and it is the derived polymorphic allele (RET+3:T) that is overtransmitted. We postulate that RET+3 is the most likely site of the disease variation.

We queried whether susceptibility to HSCR within this locus can be explained by *RET* alone or whether additional common variants might be present at *GALNACT-2* or *RASGEF1A*. We used the exhaustive allelic TDT (EATDT), a method for iteratively and successively testing all possible haplotypes of all possible sizes for association with HSCR[12,13]. Seventeen haplotypes are significantly associated with HSCR, but they have two critical properties (Fig. 1b): first, no associated haplotype is limited to markers across *GALNACT-2* or *RASGEF1A*, and second, all haplotypes involve *RET* SNPs alone, particularly those in intron 1. These results strongly indicate a role for a single, common variant within *RET*. Because all but one haplotype involves RET+3, we conclude that the HSCR association can be attributed to one of the following: RET+3 is in tight linkage disequilibrium with a yet unknown disease-susceptibility variant, RET+3 is the disease-causing mutation alone, or RET+3 is a disease-causing variant that acts synergistically with additional disease variants on the associated haplotype.

Comparative genomics to define functional elements

The finding of association across an intron indicated a need to identify functional elements within the *RET* locus. Systematic comparisons of orthologous sequences can uncover coding and non-coding functional elements on the assumption that such regions evolve more slowly than non-functional (neutral) sequences[14–17]. We obtained and compared the genomic sequence of a ~350-kb segment encompassing human *RET* with the orthologous intervals in 12 non-human vertebrates. MCSs were identified as the intersection of elements that satisfied the criteria of Bray[18] and Margulies[19]. Synteny is preserved across this interval in all vertebrates examined, although the fraction of sequence that can be aligned with the human sequence decreases with increasing evolutionary distance (Fig. 2a).

A total of 84 MCSs were identified (Supplementary Table 1), with 44% (37 of 84) of the identified MCSs corresponding to exons of *RET*, *GALNACT-2* and *RASGEF1A*. The remaining 47 MCSs are likely to be non-coding because no matching complementary DNA sequence or open reading frame greater than 20 amino acids in length was found. We identified five such elements within the most highly associated 27.6 kb around *RET* intron 1 (MCS−5.2, MCS−1.3, MCS+2.8, MCS+5.1 and MCS+9.7, identified by their distance in kilobases from the *RET* start site as in Fig. 3a).

Although *GALNACT-2* and *RASGEF1A* are unlikely to harbour common HSCR variants, they might carry rare mutations and be important in HSCR, just as some of the 126 patients we studied also have rare *RET* mutations. To test their involvement in enteric development and HSCR, we characterized their temporal and spatial expression in humans and mice. Transcription of *RASGEF1A* is limited to brain and several tissues (bone marrow, testis, colon and placenta) with high replicative capacity (Fig. 2b–d). *RET* and *GALNACT-2* share overlapping, nearly ubiquitous postnatal expression patterns. Importantly, *GALNACT-2* and *RASGEF1A* are both highly expressed at 13.5 days *post coitum*, coincident with peak *RET* expression and colonization of the gut by neural-crest-derived neuronal precursors (Fig. 2c), a feature disrupted in HSCR. Consequently, *GALNACT-2* and *RASGEF1A* expression patterns are consistent with a potential role in enteric neural crest migration. However, our analysis of morpholino-based gene

Table 1 **Analysis of disease associations**

Gene	Marker	dbSNP ID	A1	A2	All affected individuals			Male offspring			Female offspring		
					T	U	τ	T	U	τ	T	U	τ
5′ *RET*	RET−6	rs3097565	G	T	45	43	0.51	27	32	0.46	18	11	0.62
	RET−5	rs2742250	G	C	56	44	0.56	41	27	0.60	15	17	0.47
	RET−4	rs3026707	A	G	45	33	0.58	32	19	0.63	13	14	0.48
	RET−3	rs3026720	T	C	38	29	0.57	27	17	0.61	11	12	0.48
	RET−2†	rs741763	G	C	69	26	0.73**	47	14	0.77**	22	12	0.65
	RET−1†	rs2505997	C	T	57	19	0.75**	43	10	0.81**	14	9	0.61
RET int1	RET+1†	rs2435365	T	C	76	29	0.72**	53	17	0.76**	23	12	0.66
	RET+2†	rs2435364	A	G	73	27	0.73**	50	15	0.77**	23	12	0.66
	1.1Sfcl†	rs2435362	A	C	100	27	0.79***	72	14	0.84***	28	13	0.68*
	RET+3†‡	rs2435357	T	C	101	25	0.80***	73	12	0.86***	28	13	0.68*
	RET+4†	rs752975	A	G	74	29	0.72**	51	17	0.75**	23	12	0.66
	INT1.4b†	rs2505535	G	A	92	28	0.77***	68	14	0.83***	24	14	0.63
RET protein-coding region	X2EagI†	rs1800858	A	G	96	28	0.77***	72	14	0.84***	24	14	0.63
	INT8	rs3026750	G	A	60	40	0.60*	42	22	0.66*	18	18	0.50
	X13TaqI	rs1800861	G	T	59	38	0.61*	41	21	0.66*	18	17	0.51
	I18BbvI	rs2742237	C	G	59	28	0.68*	41	14	0.75**	18	14	0.56
	I18StyI	rs2742239	A	G	52	27	0.66*	34	12	0.74**	18	15	0.55
	I19BsgI	rs2075912	T	C	55	25	0.69*	37	12	0.76**	18	13	0.58
GALNACT-2	GN−1	rs3026787	G	A	17	14	0.55	15	10	0.60	2	4	0.33
	GN+1	rs4948705	C	T	59	29	0.67*	40	13	0.75**	19	16	0.54
	GN+2	rs1864393	A	G	35	15	0.70*	27	9	0.75**	8	6	0.57
	GN+3	rs2435337	G	C	57	29	0.66*	39	14	0.74**	18	15	0.55
	GN+4	rs2505556	C	T	63	59	0.52	42	39	0.52	21	20	0.51
	GN+5	rs2435384	G	T	57	39	0.59	39	21	0.65*	18	18	0.50
	GN+6	rs2435381	T	C	55	41	0.57	37	28	0.57	18	13	0.58
RASGEF1A	RAS+2	rs1254958	T	C	56	27	0.67*	38	13	0.75**	18	14	0.56
	RAS+1	rs1254965	T	C	56	27	0.67*	38	12	0.76**	18	15	0.55
	RAS−1	rs1272142	G	T	55	41	0.57	38	22	0.63*	17	19	0.47
	RAS−2	rs1955356	A	T	51	39	0.57	33	24	0.58	18	15	0.55

A1 and A2 are associated and unassociated alleles, respectively; T and U are transmitted and untransmitted allele counts, respectively; τ is transmission frequency of A1.
† SNPs highlighted by the black box on Fig. 1a and displayed in Fig. 3a.
‡ rs2506005 and rs2506004 lie 76 nt upstream and 217 nt downstream of RET+3, respectively. These three SNPs were found to be in complete linkage disequilibrium; the TDT results are therefore identical with those for RET+3.
Significant *P* values are represented as follows: asterisk, $0.001 < P < 0.05$; two asterisks, $10^{-6} < P < 10^{-3}$; three asterisks, $10^{-11} < P < 10^{-9}$.

knockdowns in zebrafish has uncovered only mid-gastrulation defects in convergence and extension for *Galnact-2* and neuronal cell death in the central nervous system by 24 h after fertilization for *Rasgef1a* (data not shown). In contrast, a similar disruption of *RET* results in incomplete colonization of the digestive tube by enteric neurons[20,21]. These functional analyses cannot exclude either *GALNACT-2* or *RASGEF1A* as HSCR candidate genes, because the observed embryonic lethality occurred before the onset of migration of neural crest cells into the digestive tube. However, genetic association tests have excluded the occurrence of a common mutation at *GALNACT-2* or *RASGEF1A* as a contributory factor in HSCR.

MCS+9.7 functions as an enhancer *in vitro*

Although MCS+9.7 is probably a functional element, the specific function of this sequence and the mechanism by which it exhibits a deleterious effect is not known. MCS+9.7 demonstrates a minimum identity of 72.5% with all mammalian species examined. No predicted structural or regulatory RNAs were identified in MCS+9.7 with the QRNA algorithm[22]. The MCS+9.7 sequence includes a gamut of predicted transcription-factor-binding sites (Supplementary Table 2), including two retinoic acid response elements (RARE) within 4 nt on each side of the RET+3 site. However, no predicted binding sites are disrupted directly by the mutant RET+3:T allele or the alleles at the rs2506004 and rs2506005 sites. Retinoic acid has already been documented as a negative and a positive regulator of *RET* expression in cardiac and renal development, respectively[23,24]. Furthermore, exogenous retinoic acid delays the colonization of the hindgut by *RET*-positive enteric neuroblasts and results in ectopic *RET* expression during embryogenesis[25]. Although the mutation or mutations do not introduce or destroy a predicted RARE, it might introduce a new site that permits competition with, or reduces access to, the neighbouring predicted RAREs. Thus, the ultimate proof of disease causation will require the synthesis of the trait, from one or all three of the MCS+9.7 variants, in an appropriate model organism.

On the basis of its location, we predicted that the MCS+9.7 element functions as a transcriptional enhancer or suppressor. Using transient transfection assays, we tested the function of two *RET* intron 1 constructs in the mouse neuroblastoma cell line Neuro-2a. Amplicons containing MCS+9.7 and MCS+5.1/+9.7 show enhancer activity in this cell line (Fig. 3b), although this activity in HeLa cells is negligible (data not shown), indicating that the activity of MCS+9.7 might be dependent on cell type. Importantly, amplicons harbouring the mutant allele have significantly lower enhancer activity (sixfold to eightfold decrease) than those containing the wild-type allele (*t*-test, $P \leq 0.001$). These data indicate that the mutation might lie within, and might compromise

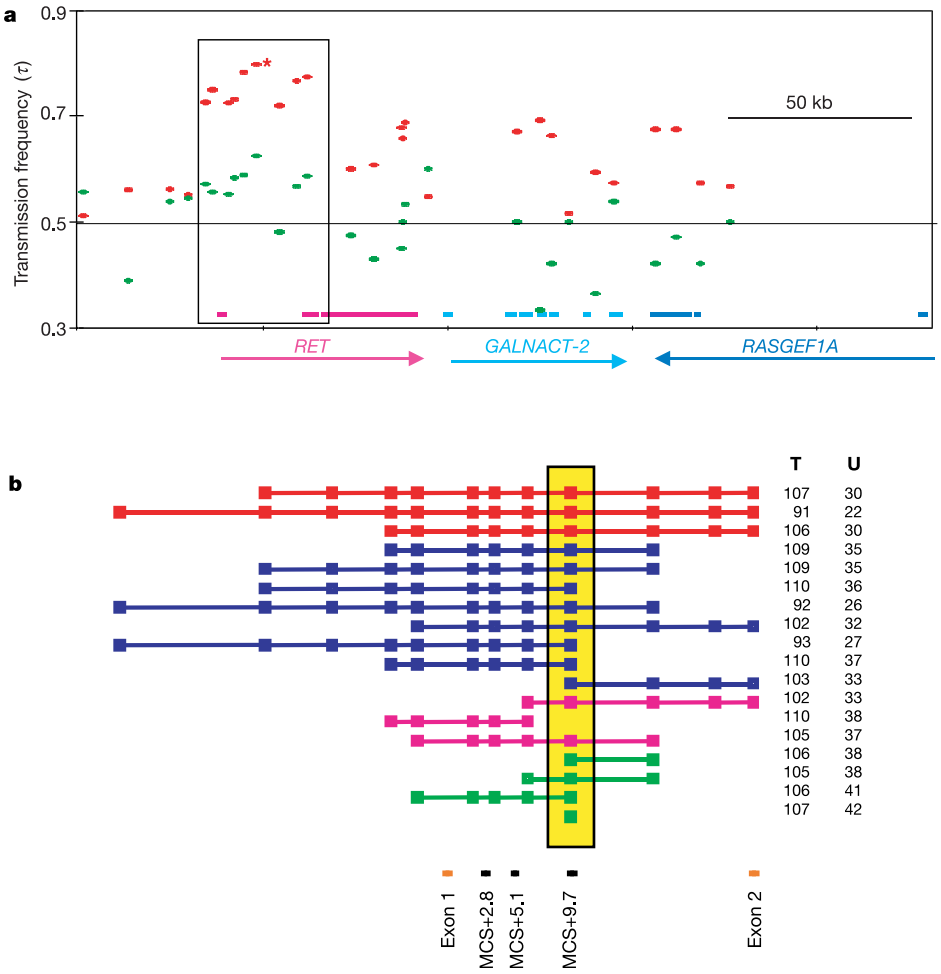

Figure 1 Transmission disequilibrium tests. **a**, TDT tests of individual SNPs. The region of 10q11.21 including *RET*, *GALNACT-2* and *RASGEF1A*. The horizontal line at 50% transmission indicates expectation under the null hypothesis. The asterisk identifies RET+3. Exons are marked as coloured boxes. The black rectangle represents the 27-kb area in Fig. 3a. Red symbols, affected offspring; green symbols, unaffected offspring. **b**, EATDTs. The 5′-most SNP shown is RET−5, the 3′-most SNP is X2EagI. Counts of transmitted (T) and untransmitted (U) chromosomes are given in columns at the right. All haplotypes with permutation-based *P* values less than or equal to the single most significantly associated SNP (RET+3) are shown. Red, $P < 10^{-8}$; blue, $P=10^{-8}$; purple, $P=10^{-7}$; green, $P=10^{-6}$.

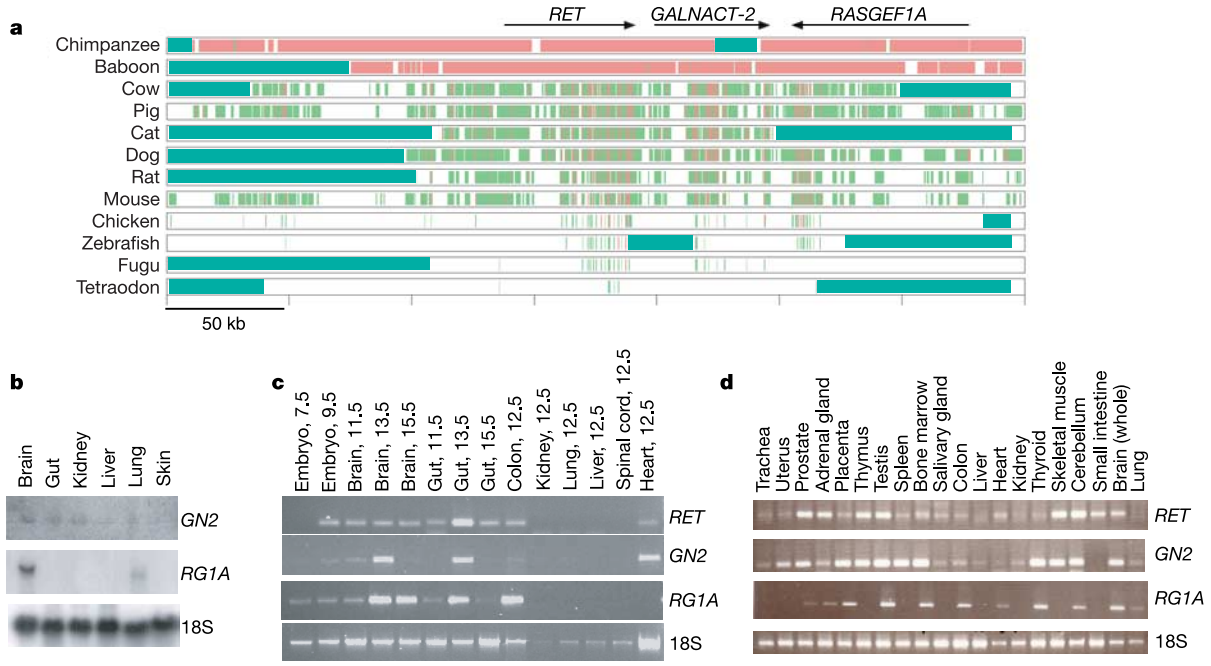

Figure 2 Identification and characterization of conserved sequence elements within 350 kb encompassing *RET*. **a**, Multi-PIP (per cent identity plot) alignment of genomic sequence from 12 vertebrates compared with the human sequence. Red, more than 75% sequence identity over 100 nt; green, more than 50% sequence identity over 100 nt; blue, gaps in contig of 500 nt or more; white, unalignable. **b**, Northern blots showing expression of *GALNACT-2* (GN2) and *RASGEF1A* (RG1A) in adult mouse tissues. **c**, **d**, Expression of *RET*, *GALNACT-2* (GN2) and *RASGEF1A* (RG1A) by RT–PCR in embryonic mouse (**c**) and adult human tissues (**d**). Numbers in **c** are days *post coitum*.

the activity of, an enhancer-like sequence in *RET* intron 1. *RET* coding sequence mutations in HSCR are always loss-of-function alleles. Thus, our finding that the RET+3 mutation decreases transcription is consistent with HSCR biology. We can localize the enhancer function, and the genetic change that diminishes that function, to the 900-nt fragment tested in the MCS+9.7 construct. Within this region there exist three segregating sites (rs2506005, RET+3 and rs2506004) in complete linkage disequilibrium. In principle, any one of these three sites, or their combination, can be the disease susceptibility factor.

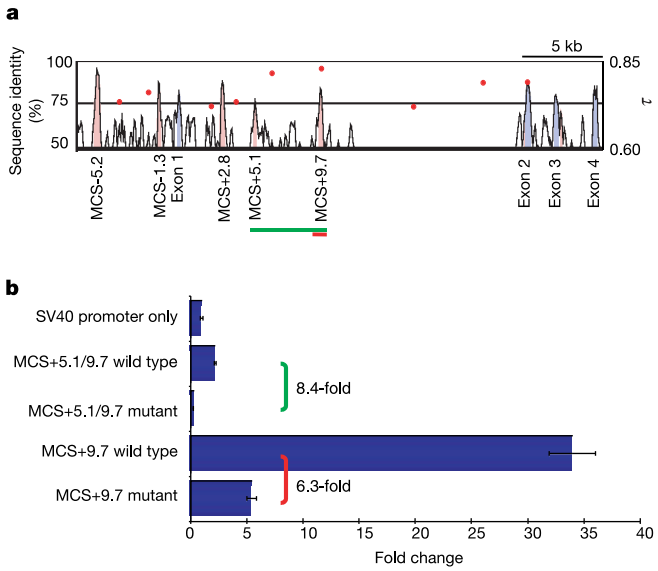

Figure 3 Identification and functional characterization of *RET* MCS+9.7. **a**, VISTA plot displaying percentage identity between mouse and human in the 5′ region of *RET*. Estimated transmission frequencies to affected offspring are shown by red circles. **b**, Reporter gene expression in Neuro-2a cells with the use of amplicons MCS+9.7 and MCS+5.1/9.7 (mutant and wild type correspond to nucleotides T and C, respectively). The smaller of the tested constructs (MCS+9.7 only) is bracketed in red; the MCS+5.1/9.7 amplicon encompassing both MCS+9.7 and the adjacent MCS+5.1 is bracketed in green. All assays were conducted in triplicate and were repeated three times (nine data points total). Error bars represent standard error.

Worldwide distribution of MCS+9.7 variants

The global distribution of the RET+3:T allele was determined by genotyping individuals from 51 unselected populations. The mutant T allele is virtually absent within Africa (a frequency of less than 0.01), has intermediate frequency in Europe (0.25) but reaches high frequency (0.45) in Asia (Fig. 4). Additionally, we generated haplotypes for seven SNPs from 180 individuals, 60 each from Africa, Europe and Asia, derived from the above worldwide set, and compared them against haplotypes from HSCR patients (Supplementary Table 3). Haplotypes bearing the RET+3:T allele probably have a single origin, some time after modern humans emerged from Africa. The high frequency of the RET+3:T allele, and the susceptibility haplotype, in east Asia correlates with an increased incidence of short-segment HSCR among Asian newborns (3.1 versus 1.5 out of every 10,000 births in Asian American versus European American births in California between 1983 and 1997; C. Torfs, personal communication). This same haplotype has a 66% frequency in Chinese sporadic HSCR patients[6]; consequently, a doubling of the mutant allele frequency translates into an approximate doubling of disease incidence. We suspect that RET+3:T is a marker for short-segment HSCR because the low frequency of the RET+3:T allele in Africa correlates with a lower frequency of short-segment HSCR among African Americans[3].

These data strongly suggest that, of the three SNPs within MCS+9.7, only the RET+3 variant is the susceptibility mutation. The associated alleles at rs2506005, RET+3 and rs2506004 are the ancestral, derived and ancestral alleles, respectively. Given our knowledge of human evolution and that the susceptibility haplotype has 1% frequency in Africa, the ancestral haplotype (with ancestral alleles at each SNP) was virtually extinct within Africa

until it increased in frequency with the occurrence of the RET+3:T mutation.

This finding of a common allele that rapidly increased in frequency but is associated with a disease predisposition can be explained in one of three ways: first, recurrent mutations from the wild type to the same deleterious mutant; second, chance increase by genetic drift; and third, a selective advantage of the mutation in heterozygotes. Our finding of a common haplotype indicates that the first explanation may be unlikely. To distinguish between the two remaining alternatives, we performed two analyses: we estimated an F_{ST} value of 0.027, and we compared our worldwide mutant allele distribution (summarized as an allele frequency of less than 5% in Africa, more than 25% in Europe, and more than 40% in China and Japan) to that of 8,247 SNPs from the ENCODE loci[26]. Only 38 sites (0.46%) show the observed or a more extreme pattern, strongly indicating a selective advantage to the mutation.

We do not know the nature of the selective agent or whether such selection was only in the past or exists today. Common mutations and polymorphisms have survived for long periods of human evolution and have been exposed to numerous biological and demographic factors that have affected their survival and current frequency. If polymorphisms make substantial contributions to common disorders, a significant fraction of them must have been exposed to selection. It is not therefore surprising that most common disease associations involve alleles that provided (α-globin, β-globin[27], glucose-6-phosphate dehydrogenase G6PD[28–30], HLA[31], Fy[32] and other variants in malaria), or are suspected of providing (CCRΔ32 in HIV infection[33–35]), a survival advantage to humans. Thus, many common variants in currently common disorders perhaps stem from alleles that were, or are, protective for another phenotype, providing mechanistic support to the common variant, common disease model of genetic disease[36,37].

Before the advent of corrective surgical methodologies in the 1950s, HSCR was a uniformly fatal disorder, necessitating positively acting selective forces to maintain this deleterious allele at high frequency. Our demonstration that the RET+3:T allele is a derived allele that is virtually absent in Africa but rose to a frequency of 0.25 in Europe and 0.45 in Asia in 100,000 years or less is indicative of such a selective force. *RET* is a tyrosine kinase receptor on the surface of neuroblasts and many other cell types, and it is not inconceivable that it might be a target of pathogen entry, similar to the chemokine receptors involved in HIV and malaria.

Genetic properties of the RET+3 susceptibility allele

A pervasive feature of HSCR is the marked gender difference in expression and incidence, with males being four times more likely to be affected than females. These sex differences could arise from mutations on the X chromosome, but genome-wide mapping studies[1,2] have consistently failed to identify an X-linked gene. Consequently, we tested whether the RET+3 variant at MCS+9.7 shows sex-specific effects. As shown in Table 1, transmission frequency of the associated allele in the *RET* region is always smaller to affected daughters than to affected sons, with rare exceptions at non-significant SNPs. Indeed, given the lower female penetrance, there were fewer affected daughters than sons in our sample, and among them only the mutant SNP (boys, $\tau = 0.86$, $P = 3.7 \times 10^{-11}$; girls, $\tau = 0.68$, $P = 0.02$) and the SNP at 1.1SfcI, 3.3 kb away, are statistically significantly different from 0.50. Nevertheless, a trend test for a difference in transmission frequency in male and female offspring is highly significant and estimates the male-to-female transmission ratio to be approximately 2 ($P = 0.0007$). Thus, the genetic effect at MCS+9.7 is significantly greater in sons than in daughters.

Two other features of the RET+3 mutation display sex differences consistent with the greater incidence in males than females. First, as shown in Table 2, the transmission frequency to affected sons and daughters leads to a 5.7-fold and 2.1-fold increase in susceptibility in males and in females, respectively, assuming a multiplicative model for penetrance. Second, genotype frequencies of affected individuals can be used to estimate the penetrance, which varies between 6.2×10^{-5} and 1.8×10^{-3} (Table 2) and is considerably smaller than that for long-segment HSCR. Our finding of gender differences in penetrance is consistent with the greater incidence of HSCR in males. For all traits with gender-specific differences in incidence, affected individuals from the less frequently affected sex (females for HSCR) have a higher mean liability. Therefore, when we consider all susceptibility loci, we expect females with HSCR to carry more susceptibility alleles than their male counterparts[38]. It

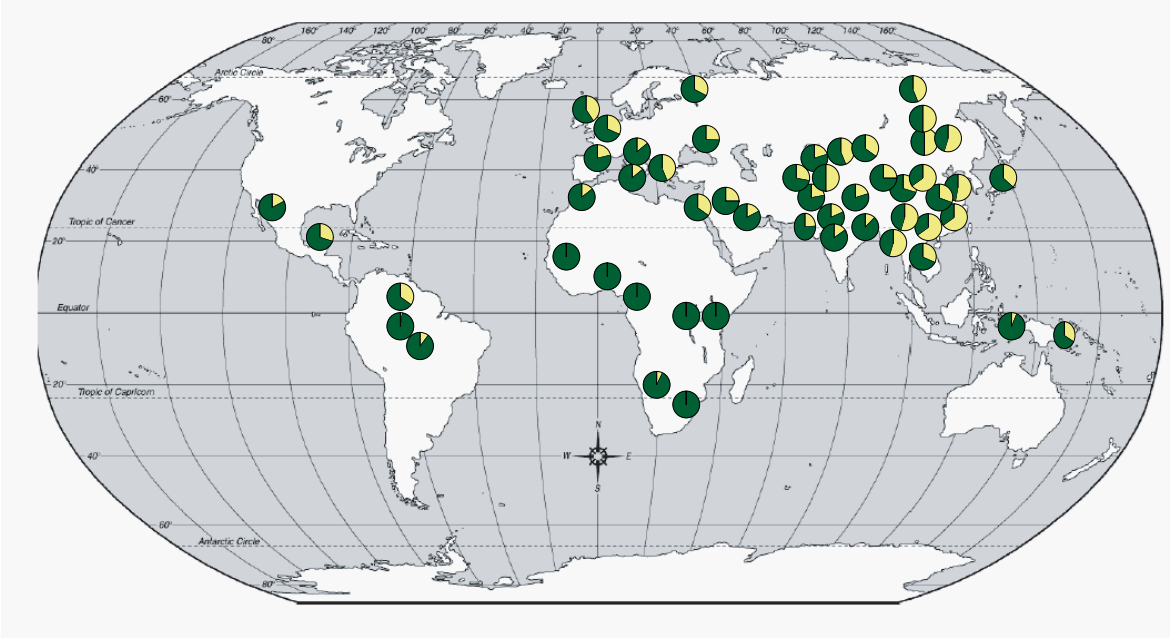

Figure 4 Worldwide allele frequencies of RET+3. Frequencies of the putative wild-type (green, C) and mutant (yellow, T) alleles are given for 51 populations comprising 1,064 individuals from the CEPH Human Genome Diversity Panel.

follows that the penetrance of any specific mutation must be lower for the less-affected sex, as observed here.

To assess the genetic impact of this common mutation we estimated the proportion of the total variance in susceptibility that the RET+3 mutation explains. Only 2.63% and 1.14% of the variation are explained by the action of this mutation in males and females, respectively (Table 2). This is in contrast with the meagre 0.1% of the total variance in susceptibility explained by all known coding mutations at *RET*[3]. Consequently, the MCS+9.7 enhancer mutation explains a 10–20-fold greater susceptibility variation than all other known *RET* mutations. However, our findings also show that a considerable number of additional loci may remain to be identified.

A final gender difference is that the mutant allele arises from mothers and fathers in 35 and 18 of the 53 informative families, respectively. This is significantly different from expectation ($P = 0.02$) and similar to the effect that we observed previously in linkage analysis of *RET* in a different series of families[2]. The cause of this bias is unknown because *RET* is not known to be imprinted; however, whether *RET* shows specific imprinting in neuroblasts is unknown.

Discussion

A combination of human genetic, comparative genomic, functional, and population genetic analyses cumulatively demonstrates a common non-coding mutation at *RET* in HSCR. This common polymorphism is the explanation of our previous failure to find coding sequence mutations in the majority of HSCR cases, even in patients from families showing linkage to *RET*. This polymorphism means that many parents in multiplex families are homozygous for the variant allele and consequently fail to show segregation of *RET*. We conclude that *RET* mutations, coding and/or non-coding, are probably a necessary feature in all cases of HSCR. However, *RET* mutations are not sufficient for HSCR because the disease incidence also requires mutations at additional loci[1,2,4].

Our identification of the RET+3 mutation was aided by comparative sequence analysis and emphasized by its likely selection. This is not to suggest that additional *RET* variants cannot exist, but if they do not lie in identifiable functional sites then they cannot be confirmed. This finding has several implications for the genetic analyses of both mendelian and complex disease. First, mutation searches in human disease should include both the coding sequences of genes and neighbouring non-coding elements. This is critical not only to common complex disorders, in which non-coding mutations may conspire with mutations at additional genes for disease to occur, but also in rare mendelian phenotypes in which 10–15% of patients can have no recognized mutations despite incontrovertible evidence for a single known gene. Second, not all mutations for rare diseases are required to be rare or have 100% penetrance. Thus, the criterion of identifying mutations as sequence changes that are absent from controls may not be appropriate for a significant fraction of alterations and may exclude legitimate mutations. Third, the inheritance patterns of single gene traits due to common variants are somewhat different from those that we have come to expect from rare mendelizing mutations, particularly when penetrance is not complete. Thus, apparent genetic heterogeneity in linkage or bilineal inheritance does not imply that mutations do not exist at a single locus.

One of the major challenges emphasized by our research is the critical need to define functional non-coding elements. A variety of non-coding elements are involved in transcription, translation, recombination, replication and repair, but we remain ignorant of the full nature and function of these sequences. Comparative genomics provides the current best avenue for recognizing such elements in a generic way, but this depends on the assumption that functional sites evolve recognizably more slowly than non-functional sites. These analyses have shown that only 1.5% of the human genome is devoted to coding exons, and as much as 3% to conserved non-coding sequences[39], implying that the latter might be particularly important as sites of mutation. Nevertheless, comparative methods are inefficient at recognizing function when it is determined by only a few nucleotides, such as at transcription-factor-binding sites. The recent focused efforts at experimental detection of non-coding regulatory elements are therefore very important to disease genetics[26]. In the future, with a more complete knowledge of genomic function, mutation detection in human disease should encompass all functional elements, not only exons.

Our studies provide a molecular view to a multifactorial disorder: the most common mutation is non-coding, it has low (marginal) penetrance, the mutation has sex-dependent effects and explains only a small fraction of the total susceptibility to HSCR. Nevertheless, it has three features that are relevant to the analysis of common complex disorders. First, although the known protein-coding HSCR mutations have higher (51–72%) penetrance, their rarity in the population implies they explain only a minute fraction (0.1%) of the disorder. Additional genes or environmental factors are therefore necessary to explain disease incidence. Second, about 11% of our HSCR patients carry known *RET* coding mutations in addition to the RET+3:T variant. It is not unlikely that coding and non-coding mutations might act synergistically to affect disease penetrance; in other words, there might be more than one mutation per gene. Third, an enhancer mutation allows us to speculate that additional factors (proteins) interact with this element and can mitigate or attenuate its genetic effect on *RET* transcription. Thus, for common mutations, we expect that mutation penetrance will depend on other alleles and genes (genetic background), epigenetic effects (such as those associated with sex-linked gene dosage) or even the environment. □

Table 2 **Genetic characteristics of the *RET* enhancer mutation**

Genotype	Observed genotype counts†		Expected frequency‡	Penetrance ($\times 10^5$)§	
	Males	Females		Males	Females
CC	40	15	0.58	16.1 ± 2.2	6.2 ± 0.9
CT	50	17	0.37	34.5 ± 3.8	6.4 ± 1.3
TT	37	26	0.06	175.0 ± 22.9	35.9 ± 8.0
Risk ratio (γ)#				5.7	2.1
Variation (%)				2.63	1.14

The results assume a population incidence of 1 in 5,000, with S-HSCR and L-HSCR representing 80% and 20% of cases, respectively, and a sex ratio of 2:1 (M:F).
†Based on genotyping 185 affected individuals.
‡Estimated on the basis of allele frequencies of 0.76 (C) and 0.24 (T) on untransmitted chromosomes.
§Penetrance estimates are given with one standard deviation.
#Estimated by assuming a multiplicative model under which $\tau = \gamma/(1+\gamma)$.

Methods

Patient samples

We genotyped trios with 126 probands, all their parents (of whom 3 were affected) plus 24 unaffected siblings; for the penetrance studies we also genotyped additional probands for a total of 450 samples. All forms of HSCR (short segment, long segment, and total colonic aganglionosis) were represented in the patient sample. Of the ascertained cases, 11% presented with additional anomalies including defined neurocristopathies, chromosomal abnormalities (for example, trisomy 21) and other defects. Ascertainment was conducted under informed consent approved by the Institutional Review Board of the Johns Hopkins University School of Medicine. In addition to the HSCR patients and their families, we also genotyped 1,064 samples representing individuals from six continents from the CEPH Human Genome Diversity Panel (http://www.cephb.fr/HGDP-CEPH-Panel/; ref. 37).

SNP genotyping

We selected SNPs with a minimum minor allele frequency of 10%, with physical map locations covering the three genes *RET*, *GALNACT-2* and *RASGEF1A* and emphasizing the associated region within *RET*[8]. From dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), we selected SNPs with known heterozygosity and/or SNPs with both alleles observed twice ('double hit' SNPs); we used markers for which robust genotyping assays could be developed. All SNPs are referred to by their rs numbers. Genotypes were generated with the fluorogenic 5′ nuclease assay (Taqman; Applied Biosystems). A TECAN Genesis workstation was used for all liquid handling, thermal cycling was

completed on MJ Research Tetrads, and end-point measurements were made on an ABI 7900 (Applied Biosystems). Genotypes were determined with SDS 2.1 (Applied Biosystems) and verified by the instrument operator. Of the samples, 10% ($n = 45$) were genotyped in duplicate for all 30 markers; no discrepancies were observed between the 1,350 paired replicate genotypes.

Transmission disequilibrium test

The TDT χ^2-test statistic was used to identify significant deviation from the expected 1:1 mendelian transmission[11]. The transmission frequency (τ) from heterozygous parents to offspring was estimated from all family genotype data at each SNP by maximum likelihood. We examined the following three hypotheses: single τ; τ different by parent gender (τ_m, τ_f); or different transmission rates to male (b) and female (g) children (τ_b, τ_g). χ^2 tests with one degree of freedom based on the appropriate likelihood ratio were used to test whether $\tau = 1/2$, $\tau_m = \tau_f$ or $\tau_b = \tau_g$.

Haplotype reconstruction and EATDT

For family-based samples, haplotypes were inferred by using hap2, a method that combines traditional family-based reconstructions with population-based linkage disequilibrium information to achieve extremely accurate reconstruction within nuclear families[12]. Haplotypes for control HGDP individuals were reconstructed with PHASE[41]. EATDTs were performed, after haplotype reconstruction, for all sliding windows of all numbers of SNPs at all positions[13]. Within each window of any size, all observed haplotypes were tested for association by the TDT. To assess overall significance while accounting for multiple tests, 10^8 permutations were performed to estimate *P*.

Resequencing

Three resequencing experiments were performed and analysed to identify novel SNPs: first, DNA chip-based resequencing[42] of the non-repeat sequence in a 90-kb interval containing *RET* in 32 Mennonites (15 HSCR cases and 17 controls); second, resequencing MCSs within *RET* intron 1 in 22 HSCR patients from families with *RET*-linkage but no identified coding sequence mutations; and third, resequencing 9 kb around RET+3 in four and eight individuals each homozygous for the RET+3:T and the RET+3:C allele, respectively. These analyses identified numerous rare and novel SNPs, additional low-frequency SNPs existing in dbSNP, and a high-frequency SNP within intron 1 enriched in patients, namely RET+3. In addition to RET+3, we identified variants within three additional intron 1 conserved elements (see below) by resequencing in HSCR patients.

Allele distribution at ENCODE loci

The ENCODE project[26] has identified all segregating sites at five loci on human chromosomes 2p16.3, 2q37.1, 4q26, 7q21.13 and 7q31.33, each about 500 kb in length. All SNPs were genotyped in the HapMap samples from four populations, namely Utah CEPH, Yoruba from Ibadan, Nigeria, Han Chinese from Beijing, and Japanese from Tokyo, Japan (www.HapMap.org). We estimated allele frequencies at 8,247 SNPs in the three continental regions (Europe, Africa and Asia; 60 independent individuals each) and compared them with those of the RET+3:T allele. We estimated the probability of observing the allele frequency as less than 5% in Yoruba, more than 25% in Europe and more than 40% in China/Japan in all 8,247 SNPs as 0.0046. To reduce effects of linkage disequilibrium, we sampled every second (4,121 SNPs), fourth (2,059 SNPs), eighth (1,028 SNPs) and sixteenth (512 SNPs) SNP to obtain probabilities of 0.0036, 0.0049, 0.0068 and 0.0059, respectively. An identical analysis with the F_{ST} statistics gave a *P* value of 0.027 (0.023–0.029).

Additional methods are provided in Supplementary Information.

Received 14 December 2004; accepted 15 February 2005; doi:10.1038/nature03467.

1. Bolk, S. *et al.* A human model for multigenic inheritance: phenotypic expression in Hirschsprung disease requires both the RET gene and a new 9q31 locus. *Proc. Natl Acad. Sci. USA* **97,** 268–273 (2000).
2. Gabriel, S. B. *et al.* Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nature Genet.* **31,** 89–93 (2002).
3. Chakravarti, A. & Lyonnet, S. in *The Metabolic and Molecular Bases of Inherited Disease* 8th edn (eds Scriver, C. R., Beaudet, A. R., Sly, W. & Valle, D.) Ch. 251, 6231–6255 (McGraw-Hill, New York, 2001).
4. Carrasquillo, M. M. *et al.* Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nature Genet.* **32,** 237–244 (2002).
5. Borrego, S. *et al.* RET genotypes comprising specific haplotypes of polymorphic variants predispose to isolated Hirschsprung disease. *J. Med. Genet.* **37,** 572–578 (2000).
6. Garcia-Barcelo, M. M. *et al.* Chinese patients with sporadic Hirschsprung's disease are predominantly represented by a single RET haplotype. *J. Med. Genet.* **40,** e122 (2003).
7. Sancandi, M. *et al.* Single nucleotide polymorphic alleles in the 5′ region of the RET proto-oncogene define a risk haplotype in Hirschsprung's disease. *J. Med. Genet.* **40,** 714–718 (2003).
8. McCallion, A. S. *et al.* Genomic variation in multigenic traits: Hirschsprung disease. *Cold Spring Harb. Symp. Quant. Biol.* **68,** 373–381 (2003).
9. Uyama, T. *et al.* Molecular cloning and expression of a second chondroitin *N*-acetylgalactosaminyltransferase involved in the initiation and elongation of chondroitin/dermatan sulfate. *J. Biol. Chem.* **278,** 3072–3078 (2003).
10. Sato, T. *et al.* Molecular cloning and characterization of a novel human β1,4-*N*-acetylgalactosaminyltransferase, β4GalNAc-T3, responsible for the synthesis of *N*,*N*′-diacetyllactosediamine, galNAc β1-4GlcNAc. *J. Biol. Chem.* **278,** 47534–47544 (2003).
11. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52,** 506–516 (1993).
12. Lin, S., Chakravarti, A. & Cutler, D. J. Haplotype and missing data inference in nuclear families. *Genome Res.* **14,** 1624–1632 (2004).
13. Lin, S., Chakravarti, A. & Cutler, D. J. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nature Genet.* **36,** 1181–1188 (2004).
14. Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288,** 136–140 (2000).
15. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424,** 788–793 (2003).
16. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 241–254 (2003).
17. Nobrega, M. & Pennacchio, L. A. Comparative genomic analysis as a tool for biological discovery. *J. Physiol.* **554,** 31–39 (2003).
18. Bray, N., Dubchak, I. & Pachter, L. AVID: A global alignment program. *Genome Res.* **13,** 97–102 (2003).
19. Margulies, E. H., Blanchette, M., Haussler, D. & Green, E. D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13,** 2507–2518 (2003).
20. Shepherd, I. T., Pietsch, J., Elworthy, S., Kelsh, R. N. & Raible, D. W. Roles for GFRalpha1 receptors in zebrafish enteric nervous system development. *Development* **131,** 241–249 (2004).
21. Shepherd, I. T., Beattie, C. E. & Raible, D. W. Functional analysis of zebrafish GDNF. *Dev. Biol.* **231,** 420–435 (2001).
22. Rivas, E. & Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2,** 8 (2001).
23. Shoba, T., Dheen, S. T. & Tay, S. S. Retinoic acid influences the expression of the neuronal regulatory genes Mash-1 and c-ret in the developing rat heart. *Neurosci. Lett.* **318,** 129–132 (2002).
24. Batourina, E. *et al.* Vitamin A controls epithelial/mesenchymal interactions through Ret expression. *Nature Genet.* **27,** 74–78 (2001).
25. Pitera, J. E., Smith, V. V., Woolf, A. S. & Milla, P. J. Embryonic gut anomalies in a mouse model of retinoic Acid-induced caudal regression syndrome: delayed gut looping, rudimentary cecum, and anorectal anomalies. *Am. J. Pathol.* **159,** 2321–2329 (2001).
26. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).
27. Haldane, J. B. S. The rate of mutation of human genes. *Hereditas* **35** (suppl.), 267–273 (1948).
28. Allison, A. C. G-6-PD deficiency in red blood cells of East Africans. *Nature* **186,** 531–532 (1960).
29. Allison, A. C. & Clyde, D. F. Malaria in African children with deficient erythrocyte glucose-6-phosphate dehydrogenase. *Br. Med. J.* **5236,** 1346–1349 (1961).
30. Motulsky, A. Metabolic polymorphisms and the role of infectious disease in human evolution. *Hum. Biol.* **32,** 28–62 (1960).
31. Hill, A. V. *et al.* Common west African HLA antigens are associated with protection from severe malaria. *Nature* **352,** 595–600 (1991).
32. Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The resistance factor to *Plasmodium vivax* in blacks. The Duffy-blood-group genotype, FyFy. *N. Engl. J. Med.* **295,** 302–304 (1976).
33. Samson, M. *et al.* Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* **382,** 722–725 (1996).
34. Dean, M. *et al.* Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science* **273,** 1856–1862 (1996).
35. Huang, Y. *et al.* The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. *Nature Med.* **2,** 1240–1243 (1996).
36. Collins, F. S. *et al.* New goals for the U.S. Human Genome Project: 1998–2003. *Science* **282,** 682–689 (1998).
37. Lander, E. S. The new genomics: global views of biology. *Science* **274,** 536–539 (1996).
38. Falconer, D. S. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Ann. Hum. Genet.* **31,** 1–20 (1967).
39. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
40. Cann, H. M. *et al.* A human genome diversity cell line panel. *Science* **296,** 261–262 (2002).
41. Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68,** 978–989 (2001).
42. Cutler, D. J. *et al.* High-throughput variation detection and genotyping using microarrays. *Genome Res.* **11,** 1913–1925 (2001).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank members of the Chakravarti laboratory for their discussions on this manuscript, M. Kenton for assistance with family recruitment, and E. Margulies and M. Blanchette for help with multi-species sequence analysis. We thank the NISC Comparative Sequencing Program for generating the multi-species sequence data. We also acknowledge the many participants of the NISC Comparative Sequencing Program, especially the leadership provided by G. Bouffard and B. Blakesley. We also acknowledge the many participants of the NISC Comparative Sequencing Program, especially the leadership provided by G. Bouffard and B. Blakesley. This work was supported by grants from the US National Institute of Child Health and Development.

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

Correspondence and requests for materials should be addressed to A.C. (aravinda@jhmi.edu).

GenBank accession numbers for genomic sequences reported in this paper are as follows: AC125509 and AC125512 (baboon), AC124166 (cat), AC138567 (chicken), RP43-171H18 (chimpanzee), AC124163 and AC124164 (cow), AC123973 (dog), AC124911 and AC125500 (fugu), AC122156 and AC124165 (pig), AC114881 (rat), AC135546 (tetra) and AC124155 (zebrafish).