nature genetics

Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies

Shin Lin, Aravinda Chakravarti & David J Cutler

Genome-wide disease-association mapping has been heralded as the study design of the next generation, but the lack of analytical methods to use genotype data fully is a large stumbling block. Here we describe an algorithm and statistical method that efficiently and exhaustively exploits haplotype information by subjecting alleles (a marker or contiguous sets of markers) from sliding windows of all sizes to transmission disequilibrium tests. By applying our method to simulated data and to Hirschsprung disease, we show that it can detect both common and rare disease variants of small effect. These results show that the theoretical benefits of genome-wide association studies are at last realizable.

Genome-wide linkage studies have been instrumental in elucidating the etiology of numerous single-gene diseases. For complex diseases, such as schizophrenia, autism and diabetes, these methods have proven less successful. In 1996, Risch and Merikangas[1] established, on theoretic grounds, the increased power of association studies over linkage methods.

Risch and Merikangas argued that in an idealized future, disease-association studies could be carried out with substantial power by typing roughly 1 million functional single-nucleotide polymorphisms (SNPs; or perfect proxies) in the genome, the 'direct' method[2]. This future has still not arrived. The ability to assess the probable functional importance of genomic regions is improving[3] but far from perfect. The alternative, to type all variants, is equally inviable, as there are estimated to be at least 15 million SNPs at frequency 1% or greater in the human genome[4,5]. Given that all SNP databases[6–9] together contain fewer than 10 million unique SNPs, a substantial fraction of this variation has yet to be discovered and will not be characterized or have working experimental assays in the foreseeable future.

The HapMap project[9], charged with characterizing human variation, has opted, for practical reasons, to focus on variants of frequency 5% or greater. Although by design some fraction of variants responsible for disease will be missed, genetic mapping studies incorporating the HapMap SNPs can still be useful because of linkage disequilibrium (LD) between a disease-causing mutation and a nearby typed site[1]. Testing individual SNPs for disease association may not, however, make full use of the genotype data set. Seeking association of combinations of SNPs that are inherited together with the disease-causing mutation (*i.e.*, using haplotypes; the 'indirect' method[2]) may be more powerful.

The most useful way of thinking about LD in the disease-mapping setting may not be as discrete blocks. An individual site may show greater LD over longer ranges than with surrounding sites[10]. This implies that a 'signal' for a disease-causing mutation may be optimally detected by testing the full length of a disease-associated haplotype for association (**Fig. 1**). In general, the beginning and end sites of a disease-associated haplotype are unknown, and they may be positioned irrespective of the boundaries of high LD blocks. Therefore, we tested allele windows of all positions and lengths.

The challenge to our approach is one of multiple tests. Sliding windows of haplotypes are correlated. Many haplotypes are extremely rare and have little power to detect association *a priori*[11]. Bonferroni correction by the total number of tests results in vastly diminished power[12]. Moreover, carrying out statistical tests on all sliding windows is computationally intensive. There is no generally accepted methodology capable of handling genome-scale data to test hypotheses using a genome-wide significance level.

We report a new algorithm and associated computer implementation that exhaustively searches all alleles (here taken to mean individual SNPs as well as continuous haplotypes of all lengths) of input sequence data to find the set yielding the lowest transmission disequilibrium test (TDT) *P* values. These *P* values are then adjusted to multiple test–corrected genome-wide significance by permutation tests[13]. We call this method and implementation the exhaustive allelic TDT (EATDT). The computer implementation of EATDT is efficient, allowing it to achieve the high level of performance required for the permutation approach to multiple testing adjustment.

RESULTS

Application to simulated sequences

Today, multiplex assays exist to genotype hundreds of thousands of SNPs in thousands of individuals[14]. To ascertain the extent to which exhaustive exploitation of observed haplotypes could compensate for genotyping at currently feasible densities, which are unlikely to include the disease-causing mutation, we ran a series of computer simulations. We generated 5-Mb sequences under the infinite sites neutral

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Broadway Research Building, Suite 475, 733 N. Broadway, Baltimore, Maryland 21205, USA. Correspondence should be addressed to D.J.C. (dcutler@jhmi.edu).

Published online 24 October 2004; doi:10.1038/ng1457

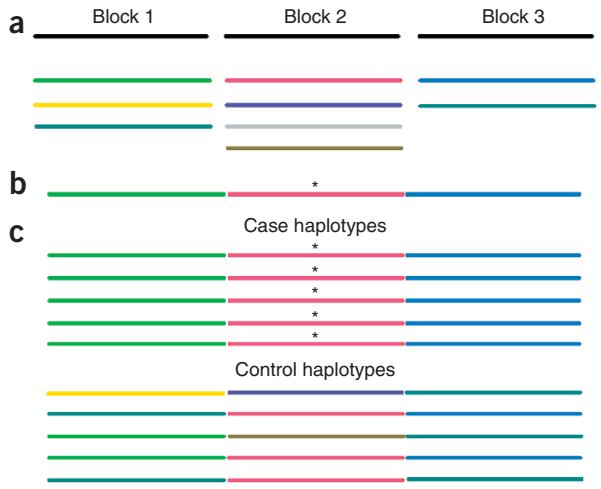

Figure 1 LD blocks, rare mutations and haplotypes. (**a**) A region of the genome is characterized by three blocks of high LD, which have three, four and two haplotypes, respectively. The haplotypes are composed of common variants. (**b**) A disease-causing mutation, represented by an asterisk, occurs on an allele containing a particular haplotype from each block of high LD. Together, the haplotypes form a single disease-associated haplotype. (**c**) A depiction of case and control alleles assuming that the mutation is rare, the disease is recessive and the blocks are interrupted by recombinational hotspots. Comparing case and control haplotypes across the entire region (five versus zero) yields a greater signal than doing so within block 2 (five versus three). Although this depiction is simplified, it is the essential feature that allows common variants to detect rare mutations.

coalescent model[4] with uniform recombination[15]. We formed diplotypes of individuals in families with these sequences and assigned disease status of family members using the mixed model of inheritance[16]. We selected trios in which the child was affected. We kept only the common sites whose frequencies were biased to reflect the distribution of variants found in public databases. We selected the final marker set at random to reach a density of 1 SNP per 10 kb. Diplotypes were phased by a computer algorithm and subjected to both an individual SNP association analysis and EATDT. Adjusted *P* values were considered significant at a genome-wide level of 0.05 using the Sidák method. We calculated power values for linkage assuming an affected sibling-pair design comprising 500 markers with a fully informative one closely linked (recombination fraction = 0) to the disease-associated locus, as in Risch and Merikangas[1].

The data in **Table 1** show that indirect, genome-wide association studies, carried out today with randomly selected common markers from public databases, will be able to uncover mutations with effects that are undetectable by linkage. The differences are most pronounced

Table 1 Power to detect disease-associated alleles

				Linkage				Single-SNP TDT[a]				EATDT[b]			
γ	q	λ_{as}	h_a^2	250 ASPs	500 ASPs	750 ASPs	1,000 ASPs	500 trios	1,000 trios	1,500 trios	2,000 trios	500 trios	1,000 trios	1,500 trios	2,000 trios
4	0.01	1.0857	0.0312	0.00	0.01	0.02	0.03	0.00	0.00	0.00	0.02	0.02	0.56	0.78	0.94
	0.03	1.2326	0.1405	0.07	0.28	0.55	0.77	0.02	0.26	0.64	0.88	0.72	1.00	1.00	1.00
	0.05	1.3494	0.2464	0.27	0.75	0.95	0.99	0.46	0.74	0.98	0.94	0.94	1.00	1.00	1.00
	0.1	1.5367	0.3878	0.73	0.99	1.00	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.2	1.6416	0.5239	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.5	1.3924	0.3819	0.38	0.87	0.99	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	0.01	1.0097	0.0035	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.08	0.34	0.62
	0.03	1.0435	0.0164	0.00	0.02	0.03	0.06	0.00	0.02	0.08	0.22	0.30	0.80	0.98	1.00
	0.05	1.0758	0.0302	0.02	0.08	0.18	0.32	0.06	0.20	0.62	0.84	0.58	0.98	0.98	1.00
	0.1	1.1142	0.0502	0.11	0.42	0.72	0.90	0.32	0.90	0.96	0.98	0.84	1.00	1.00	1.00
	0.2	1.1142	0.0649	0.28	0.77	0.96	1.00	0.76	0.98	1.00	0.98	0.98	1.00	1.00	1.00
	0.5	1.0500	0.0354	0.11	0.42	0.72	0.90	0.92	1.00	1.00	1.00	0.96	1.00	1.00	1.00
2	0.01	1.0025	0.0009	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	0.03	1.0113	0.0042	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.20	0.44
	0.05	1.0205	0.0077	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.22	0.42	0.70
	0.1	1.0333	0.0133	0.00	0.00	0.00	0.02	0.02	0.22	0.38	0.62	0.12	0.68	0.88	0.98
	0.2	1.0404	0.0184	0.01	0.01	0.02	0.09	0.28	0.76	0.90	0.98	0.42	0.90	0.98	0.98
	0.5	1.0205	0.0104	0.01	0.01	0.02	0.09	0.58	0.94	0.98	0.98	0.56	0.96	1.00	0.98
1.5	0.01	1.0025	0.0009	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.03	1.0113	0.0042	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.05	1.0205	0.0077	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02
	0.1	1.0333	0.0133	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.04	0.00	0.04	0.14	0.22
	0.2	1.0404	0.0184	0.00	0.00	0.00	0.00	0.00	0.18	0.30	0.42	0.00	0.20	0.32	0.50
	0.5	1.0205	0.0104	0.00	0.00	0.00	0.00	0.08	0.48	0.68	0.74	0.04	0.42	0.60	0.66

[a]TDT with SNPs and permutation test. [b]TDT with all alleles and permutation test.
Each value for the single-SNP TDT and EATDT methods is based on 50 simulations. In total, the computer runtime was ~4 d on a 70-node cluster for the 96 sets of 50 simulations represented (the single-SNP TDT and EATDT results are paired results). The genotype density is 300,000 SNPs per genome. γ, GRR; q, frequency of disease-associated allele; λ_{as}, sibling relative risk attributable to the main locus; h_a^2, proportion of variance explainable by the main locus (heritability if a single-locus disorder).

Table 2 Validity measures of disease-associated allele

		Validity of haplotype identified															
		500 trios				1,000 trios				1,500 trios				2,000 trios			
γ	q	Se[a]	Sp[b]	PPV[c]	NPV[d]	Se[a]	Sp[b]	PPV[c]	NPV[d]	Se[a]	Sp[b]	PPV[c]	NPV[d]	Se[a]	Sp[b]	PPV[c]	NPV[d]
4	0.01	0.97	1.00	0.94	1.00	0.87	1.00	0.93	1.00	0.81	1.00	0.95	1.00	0.79	1.00	0.93	0.99
	0.03	0.82	1.00	0.95	0.99	0.81	0.99	0.91	0.99	0.82	0.99	0.92	0.99	0.81	1.00	0.95	0.99
	0.05	0.81	0.98	0.86	0.98	0.79	0.99	0.87	0.98	0.76	0.99	0.91	0.97	0.79	0.99	0.93	0.97
	0.1	0.81	0.95	0.82	0.95	0.76	0.96	0.82	0.94	0.86	0.97	0.86	0.97	0.84	0.96	0.85	0.96
	0.2	0.80	0.91	0.83	0.90	0.76	0.92	0.83	0.88	0.84	0.93	0.86	0.92	0.8	0.88	0.77	0.90
	0.5	0.50	0.62	0.70	0.40	0.59	0.65	0.75	0.47	0.65	0.70	0.80	0.53	0.64	0.71	0.80	0.53
3	0.01	NA	NA	NA	NA	0.90	0.99	0.71	1.00	0.83	1.00	0.90	1.00	0.81	1.00	0.85	1.00
	0.03	0.89	0.99	0.82	0.99	0.79	1.00	0.91	0.99	0.80	0.99	0.88	0.99	0.82	0.99	0.90	0.99
	0.05	0.86	0.99	0.90	0.99	0.82	0.99	0.85	0.98	0.78	0.99	0.87	0.98	0.78	0.99	0.92	0.98
	0.1	0.72	0.98	0.89	0.95	0.81	0.97	0.85	0.96	0.78	0.98	0.88	0.96	0.75	0.98	0.87	0.95
	0.2	0.78	0.94	0.85	0.90	0.83	0.92	0.83	0.93	0.82	0.95	0.87	0.92	0.76	0.91	0.80	0.89
	0.5	0.52	0.59	0.67	0.43	0.61	0.67	0.74	0.52	0.54	0.63	0.71	0.46	0.63	0.71	0.77	0.55
2	0.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	0.03	NA	NA	NA	NA	0.75	0.99	0.83	0.99	0.78	1.00	0.92	0.99	0.75	0.99	0.88	0.99
	0.05	NA	NA	NA	NA	0.76	1.00	0.96	0.98	0.80	0.99	0.91	0.98	0.79	0.99	0.88	0.98
	0.1	0.87	0.96	0.78	0.98	0.80	0.99	0.91	0.97	0.74	0.98	0.84	0.96	0.74	0.97	0.81	0.96
	0.2	0.86	0.94	0.85	0.95	0.85	0.95	0.86	0.94	0.81	0.93	0.80	0.93	0.76	0.91	0.75	0.91
	0.5	0.68	0.82	0.83	0.66	0.59	0.65	0.71	0.53	0.61	0.66	0.71	0.55	0.66	0.73	0.77	0.61
1.5	0.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	0.03	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	0.05	NA	NA	NA	NA	0.95	1.00	1.00	1.00	NA	NA	NA	NA	0.83	1.00	0.95	0.99
	0.1	NA	NA	NA	NA	0.88	0.99	0.96	0.98	0.88	0.97	0.82	0.98	0.8	0.99	0.94	0.97
	0.2	NA	NA	NA	NA	0.87	0.98	0.94	0.96	0.82	0.96	0.87	0.94	0.79	0.89	0.70	0.93
	0.5	0.97	1.00	1.00	0.96	0.68	0.62	0.68	0.63	0.69	0.71	0.74	0.66	0.63	0.67	0.69	0.61

[a]Sensitivity = P(A | B). [b]Specificity = P(¬ A | ¬ B). [c]Positive predictive value = P(B | A). [d]Negative predictive value = P(¬ B | ¬ A). Let A be the event that a chromosome is identified by the exhaustive allelic method as carrying a disease-associated chromosome. Let B be the event that a chromosome carries the disease-causing mutation. Let the symbol ¬ denote the complement. Each value is based on the simulations in which a haplotype was determined significant from **Table 1**. Calculations were done on simulations for which an allele was declared by EATDT to be of genome-wide significance. The allele giving the lowest *P* value was used. The number of replicates on which each four-value set is based can be found by multiplying 50 by the corresponding EATDT value in **Table 1**.

for mutations with moderate to small effects (genotype risk ratio $\gamma = 3$ or 2) and will be greater in real studies because a fully informative marker will rarely be tightly linked to the disease-causing mutation in a linkage design. Neither method has high power for detecting mutations with miniscule effects ($\gamma = 1.5$). Nor does the direct method in Risch and Merikangas[1]. **Table 1** also shows that EATDT achieves a substantial increase in power over individual SNP analysis for many of the disease models considered. That this observation is especially true for rare disease-causing mutations is consistent with the notion that rare alleles are generally more recent and have had less time for recombination events and mutations to degrade surrounding patterns of LD[17,18].

Because we used the permutation approach to multiple tests, the gains from considering haplotypes in our exhaustive allelic method are not overshadowed by the penalty paid for doing far more tests, as some researchers have worried[19]. There is, of course, a substantial computational load, but not an insurmountable one. A genome-wide data set of 2,000 trios genotyped at a density of 1 SNP per 10 kb takes ~20 d to analyze on a single-processor personal computer. With a modestly sized cluster of computers, a whole-genome association analysis takes no longer than a weekend to complete.

We found that alleles with genome-wide significance for disease association were fairly predictive of chromosomes carrying disease-causing mutations (**Table 2**) and the general locations of these causal variants (**Table 3**). Accurate identification of chromosomal regions carrying mutations allows genetic dissection of loci at which multiple, distinct disease-causing alleles exist, a situation shown to be plausible by simulations[10]. (Of course, the alleles must each be of sufficient effect.)

We found that in a substantial fraction of simulations, the disease-associated haplotype yielding the lowest *P* value did not precisely localize the disease-causing mutation (**Table 3**). This finding is not unexpected because the most informative alleles are not necessarily the ones closest to the disease-causing mutation[20]. Other methodologies exist specifically to identify mutation location in the context of fine mapping[21–24]. Often, these algorithms assume that there is a single disease-causing mutation at a given locus or are unsuitable for genome-wide analysis because of their computational burden. In such cases, our method may be used to isolate distinct sets of chromosomes and specific regions containing disease-associated alleles, which may then be amenable to analysis by the location algorithms.

Application to Hirschsprung disease samples

We ran the power calculations for different SNP densities ranging from 1 SNP per 6 kb to 1 SNP per 300 kb. The power decreased gradually and was able to detect mutations of certain frequencies and

Table 3 Location measures of the identified disease-associated allele

		500 trios			1,000 trios			1,500 trios			2,000 trios		
γ	q	Size[a]	Con[b]	Dist[c]	Size[a]	Con[b]	Dist[c]	Size[a]	Con[b]	Dist[c]	Size[a]	Con[b]	Dist[c]
4	0.01	73.2	0	529	212	0.54	109	223	0.51	175	211	0.47	173
	0.03	117	0.61	55.2	123	0.64	68	117	0.60	56.1	116	0.78	41.2
	0.05	92.2	0.68	34.1	96.6	0.66	39.4	90.9	0.62	40.1	84	0.70	36.7
	0.1	57.8	0.64	26.3	59.3	0.68	19.7	58	0.68	21.4	54	0.70	22.4
	0.2	34.8	0.54	15.8	34.2	0.50	17.5	28.4	0.54	14.9	29.9	0.52	17
	0.5	17.3	0.32	12.0	10.9	0.30	11.5	17.3	0.36	12.6	11.6	0.40	7.77
3	0.01	NA	NA	NA	101	0.25	163	163	0.41	209	197	0.48	176
	0.03	97.6	0.8	34.4	126	0.62	50.7	106	0.55	75.0	96.5	0.70	47.1
	0.05	68.1	0.69	34.3	85.4	0.71	34.0	97.8	0.80	36.1	76.9	0.68	30.6
	0.1	74.0	0.69	24.7	55.5	0.60	28.1	70.4	0.72	23.5	65.6	0.74	22.3
	0.2	36.7	0.47	15.7	30.9	0.42	19.3	27.1	0.46	15.1	31.4	0.52	16.4
	0.5	11.3	0.27	9.19	11.1	0.36	8.82	16.6	0.36	11.2	12.5	0.30	8.94
2	0.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	0.03	NA	NA	NA	171	0.8	108	154	0.50	66.2	132	0.55	76.7
	0.05	NA	NA	NA	132	0.82	40.2	111	0.67	42.4	89.7	0.63	40.6
	0.1	39.2	0.5	15.2	59.4	0.68	19.2	73	0.75	26	74.5	0.73	39.6
	0.2	31.1	0.57	9.93	31.7	0.51	14.5	33	0.61	13.9	28.1	0.55	15.4
	0.5	14.6	0.36	8.24	13.1	0.33	9.52	10.7	0.22	12.2	10.9	0.29	14.6
1.5	0.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	0.03	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	0.05	NA	NA	NA	99.7	1	2.58	NA	NA	NA	72.9	0	47.1
	0.1	NA	NA	NA	18	0	14.8	54.1	0.6	23.4	63.9	0.55	30.8
	0.2	NA	NA	NA	25.8	0.50	11.5	26.7	0.56	16.9	17.2	0.52	6.84
	0.5	4.75	1	1.59	5.69	0.29	3.89	8.2	0.27	7.59	13.9	0.27	23.6

[a]Average size of allele identified (kb). [b]Fraction of haplotypes containing disease-associated SNP. [c]Physical distance from middle of allele to disease-associated SNP (kb).
Calculations were done on simulations for which an allele was declared by EATDT to be of genome-wide significance in **Table 1**.

effects, even at the lowest SNP density (data not shown). Encouraged by these findings and to see how well our theoretical results would hold for real data, we applied our methods to a data set comprising 35 trios with Hirschsprung disease (HSCR) from the Old Order Mennonite community[25]. The genome-wide scan consisted of 4,244 SNPs typed using the WGSA-EcoRI-p502 array (Affymetrix).

Analysis by EATDT identified three loci with genome-wide significance (**Table 4**). Two markers ~1 Mb away from the disease-causing mutation in *EDNRB* resulting in the amino acid substitution W276C (ref. 26) yielded the lowest P values, which were equivalent whether the markers were taken singly or together (note that the disease-causing mutation was not in the genome scan). Two other markers in the genotype set were closer to the known disease-causing mutation, but one of the significant SNPs was 90 bp away from the marker closest to the disease-causing mutation. The genotyped markers in this data set were not evenly spaced. No haplotype obtained a P value any lower than the individual P values of the two significant SNPs.

Table 4 Genome-wide association analysis of HSCR with EATDT

	Exhaustive allelic method	Individual SNP analysis
Locus	Adjusted P value	Adjusted P value
EDNRB (13q22)	$<10^{-6}$	$<10^{-6}$
21q21	0.037	>0.05
RET? (10q21)	0.042	0.014

For the significant findings, multiple alleles with different corresponding P values were associated to the three loci. The allele giving the lowest P value is reported.

The second locus of genome-wide significance corresponded to an eight-marker haplotype spanning 4 Mb on chromosome 21q21. This region contains a single annotated gene, *NCAM2* (encoding neural cell adhesion molecule 2); this is notable because HSCR is a neurocristopathy. The location of the region on chromosome 21 is also notable, given the association between Down syndrome and HSCR[27]. The single-SNP analysis missed this locus entirely; the individual SNPs of that region yielded adjusted P values of >0.80.

The third locus of genome-wide significance was a four-marker haplotype covering 1.6 Mb on chromosome 10q21. This region is located 12 Mb telomeric of the gene *RET* on 10q11 previously reported to be mutated in HSCR disease[28,29]. Whether the positive signal corresponds to *RET* or another gene nearby warrants further scientific investigation. Like *EDNRB*, the locus could have been found without using haplotypes, given this specific data set. There were no markers in the genome scan within *RET*; the closest was 1 Mb away. Application of our computational method using markers specific to the region yielded findings that were significant on a single–candidate gene level but not on a genome-wide level (data not shown). EATDT should work just as well for SNPs typed for fine mapping and candidate gene approaches as for genome-wide studies.

DISCUSSION

The power calculations from the simulations are not perfectly realistic but are probably underestimates. First, the Sidák method used to

Table 5 Values for TDT

		Untransmitted	
		Specific allele	Other alleles
Transmitted	Specific allele	*a*	*b*
	Other alleles	*c*	*d*

determine genome-wide significance was conservative because it treated the genome as 560 independent 5-Mb sequences. Of course, adjacent sequences are correlated and ideally should not be penalized as separate tests in *P* value adjustments. The underestimation of power, though, is probably not substantial because treating the genome as 2,800 independent 1-Mb sequences yielded values only slightly lower than those shown in **Table 1** (data not shown). Second, the genotype sets used did not involve choosing an optimally informative set of markers, called tag SNPs[9]. This additional step would yield results equal to or better than those from the simple, random SNP selection scheme that we used. Perhaps most importantly, the underlying allele frequency spectrum was assumed to be well approximated by a neutral model with uniform recombination. We know the recombination assumption is not correct in detail[30], and the neutral model's applicability to disease-associated loci, which are presumably under some degree of selection, is unproven[12]. Nevertheless, the coalescent simulation assuming constant population leads to conservative power estimates. Had a population expansion, as is presumed to have occurred in the ancestral history of humans, been modeled, LD would have extended over greater stretches. In such a scenario, on the other hand, the rough location capabilities of our method as demonstrated in the data shown in **Table 3** would be less precise.

The signals captured by our exhaustive allelic approach in the trios with HSCR were found to be a superset of those detected by individual SNP analysis, consistent with our conclusion from simulations that our method exploiting LD is as powerful as, or more powerful than, single-SNP methods. The three loci identified by our method are largely concordant with earlier work by Puffenberger and colleagues[26,31] using identity-by-descent analyses and LD mapping on microsatellite markers from a superset of the trios used in this report. The biggest difference is that evidence for association involving loci on 21q and 10q in this report was statistically significant after genome-wide adjustment for multiple tests. (*EDNRB* was judged significant in both studies.)

In our analysis of HSCR trios, genotyping error and missing data were handled during phase reconstruction with hap2 (ref. 32). The program checks for mendelian inconsistencies, which are recoded as missing data, and all missing data are then inferred. This inference process is generally accurate, but not perfect. The exact impact of extremely high error rates or large amounts of missing data in the context of EATDT is unknown and will require further research.

Our methodology solves several problems confronting the analysis of association studies. When an association study genotypes multiple SNPs, multiple test correction of *P* values is often missing or obscure.

Table 6 Related parameter values of simulations

γ	q	p	λ_{as}	h_a^2	Z	t	d	Prevalence	C/E	λ_{as}'	λ_s	h^2
4	0.01	0.99	1.0857	0.0312	2	1.6522	0.4027	0.0241	1	1.051	3.274	0.4786
	0.03	0.97	1.2326	0.0889	2	1.6522	0.4027	0.0270	1	1.130	3.277	0.4990
	0.05	0.95	1.3494	0.1405	2	1.6522	0.4027	0.0301	1	1.195	3.302	0.5230
	0.1	0.9	1.5367	0.2464	2	1.6522	0.4027	0.0384	1	1.324	3.168	0.5537
	0.2	0.8	1.6416	0.3878	2	1.6522	0.4027	0.0582	1	1.431	2.842	0.5932
	0.5	0.5	1.3924	0.5239	2	1.6522	0.4027	0.1422	1	1.325	1.988	0.6140
3	0.01	0.99	1.0384	0.0139	2	1.1752	0.4349	0.0237	1	1.020	3.254	0.4727
	0.03	0.97	1.1063	0.0402	2	1.1752	0.4349	0.0256	1	1.063	3.237	0.4834
	0.05	0.95	1.1632	0.0642	2	1.1752	0.4349	0.0275	1	1.101	3.218	0.4937
	0.1	0.9	1.2656	0.115	2	1.1752	0.4349	0.0328	1	1.165	3.112	0.5103
	0.2	0.8	1.3532	0.1841	2	1.1752	0.4349	0.0446	1	1.239	2.877	0.5330
	0.5	0.5	1.2656	0.2319	2	1.1752	0.4349	0.0910	1	1.209	2.215	0.5450
2	0.01	0.99	1.0097	0.0035	2	0.6654	0.4657	0.0232	1	1.009	3.236	0.4670
	0.03	0.97	1.0276	0.0102	2	0.6654	0.4657	0.0241	1	1.020	3.213	0.4703
	0.05	0.95	1.0435	0.0164	2	0.6654	0.4657	0.0251	1	1.033	3.177	0.4718
	0.1	0.9	1.0758	0.0302	2	0.6654	0.4657	0.0275	1	1.047	3.103	0.4765
	0.2	0.8	1.1142	0.0502	2	0.6654	0.4657	0.0328	1	1.080	2.969	0.4864
	0.5	0.5	1.1142	0.0649	2	0.6654	0.4657	0.0512	1	1.086	2.554	0.4944
1.5	0.01	0.99	1.0025	0.0009	2	0.3666	0.4819	0.0230	1	1.004	3.254	0.4678
	0.03	0.97	1.0071	0.0026	2	0.3666	0.4819	0.0234	1	1.010	3.232	0.4678
	0.05	0.95	1.0113	0.0042	2	0.3666	0.4819	0.0239	1	1.007	3.207	0.4676
	0.1	0.9	1.0205	0.0077	2	0.3666	0.4819	0.0251	1	1.016	3.157	0.4686
	0.2	0.8	1.0333	0.0133	2	0.3666	0.4819	0.0275	1	1.031	3.058	0.4697
	0.5	0.5	1.0404	0.0184	2	0.3666	0.4819	0.0355	1	1.031	2.820	0.4747

γ, GRR; q, frequency of disease-associated allele; p, $1 - q$; λ_{as}, sibling relative risk attributable to the main locus; h_a^2, proportion of variance explainable by the main locus (heritability if multilocus inheritance $C = 0$); C/E, C is the variance due to multilocus inheritance and E is the variance due to random effects; Z, threshold on liability scale; t, distance in liability between homozygotes of the disease-associated and normal alleles; d, degree of dominance ($d = 0$, 1/2 and 1 for dominant, codominant and recessive gene actions, respectively); λ_{as}', sibling relative risk of major locus under multiplicative model; λ_s, sibling relative risk; h^2, heritability.

When the study simultaneously tests an unspecified number of haplotypes with unknown correlation, this problem becomes much worse. Here we show that a study can test every SNP and every haplotype in every sliding window of every size. In doing so, the study will incur a multiple test penalty. By assessing that penalty in a permutation framework, however, we show that the additional information gained by testing all alleles overcomes the penalty paid for the additional tests. These concepts and results are consistent with previous work[33], which tested for interactions between all alleles, albeit on data sets with no more than eight SNPs. Although we tested only contiguous sets of SNPs, our method is more powerful than current genome-wide approaches and provides a computer implementation that is efficient and runs easily on large genome-wide data sets.

Most importantly, our method largely defuses the debate over the genetic basis of complex traits embodied by the common disease–common variant[2,34,35] and common disease–rare variant viewpoints[10,36]. Many researchers in the field previously believed that adoption of one philosophy or the other had large implications for study design and analysis. Specifically, there was doubt that a rare disease variant could be detected by typing common SNPs. We show that such concerns may be unwarranted. Because rare mutations are, on average, younger than the surrounding SNPs, they often exist on long haplotypes spanning multiple blocks of high LD. We can therefore use long haplotypes to find rare mutations and use short haplotypes, or even individual SNPs, to find common mutations. Because EATDT tests both, we simultaneously investigate both potential genetic architectures.

The study design and analysis adopted in our simulations are not optimally efficient. They do not use *a priori* knowledge of the LD structure of the genome: SNPs are picked at random. They make no *a priori* assumptions about the potential functional importance of genomic regions. Incorporating such information could only improve power. But the 'prior-free' approach of the indirect method ensures its ability to make new, unexpected discoveries. The rationale behind the design of the HapMap project centers on the premise that only after the genetic basis of most complex diseases is understood can we truly know which SNPs are functional.

The genetics community has so far been reluctant to generate data at the magnitude we simulated. The HSCR data set that we analyzed is 1,000 times smaller than any of the simulations. But we found that even with current capabilities for genotyping and without further development of choosing tag SNPs, association studies using EATDT are very powerful. These findings should encourage investigators to go forward with large-scale genome-wide association studies, using the results of the HapMap project, to definitively elucidate complex diseases.

METHODS

EATDT Algorithm. For any given window, there were several distinct alleles. For each allele of a window, we created a bit word of $2n$ positions. If individual i ($i = 1..n$) was heterozygous with respect to the allele at hand and that allele was transmitted to an affected offspring, then the $2i$–1th position was set to one; if that allele was not transmitted, then the $2i$th position was set to one. All other digits of the bit word were left zero.

We determined the bit words of all window positions and lengths of the given set of alleles. We call the set of such words Y. A particular element $y \in Y$ may have more than one corresponding allele (either SNP or haplotype). The sum of the bits of y gives $b + c$ in a TDT table (**Table 5**) for the bit word's corresponding allele(s).

To obtain b, we constructed a $2n$ bitword, called ts, in which the odd positions were one and the even positions were zero. The odd bits of ts represent transmitted alleles. To find b, we applied a bitwise AND operation between y and ts; the sum of the bits of this operation is b. With b (and thereby c, as $b + c$ is known), the P value from McNemar's test may be computed (exact calculations are stored in a lookup table). We calculated the P value for each element of Y. In this manner, we carried out TDTs on all possible alleles (SNPs and haplotypes of every size) in a given set of sequences and identified the allele(s) with the lowest TDT P value(s).

Suppose p is the lowest P value yielded by the algorithm, which was derived from a complicated probability density, which is dependent on the various steps outlined above and likely to be without closed form. To find the true statistical significance of p, we carried out a permutation test. For NN iterations, we constructed the kth permutation of the original data by creating a random bitword ts_k. We set the odd number bits, denoted $ts_{k,i}$ ($i = 1,3,5\ldots$), equal to one with probability 0.5 and zero otherwise. We set the even number bits, $i + 1$, denoted $ts_{k,i+1}$, equal to $\sim ts_{k,i}$, where $\sim$ is the logical not. In this way, the transmitted and untransmitted status for each pair of alleles is randomized. For any given permutation, for all $y \in Y$, we carried out a bitwise AND operation between y and ts_k, thereby obtaining the b and c values for this haplotype and this permutation. Let p_k be the lowest P value observed in the kth permutation. The adjusted P value of our procedure is the number p_k ($k = 1..NN$) smaller than p, divided by NN. An example and more detailed explanation of our algorithm is included in **Supplementary Methods** online.

Simulations. We simulated nucleotide sequences from an infinite sites[37] model with recombination[38]. Because we were working with extremely large regions, we reimplemented the classical Hudson algorithm with several computational improvements, including rewriting naturally recursive routines in a nonrecursive fashion, efficient searches for regions experiencing most recent common ancestor events and special procedures for careful management of memory storage. These enhancements allowed us to simulate the 'whole' effective human population of 40,000 alleles for sequences parameterized by $\theta = 4{,}000$ (nucleotide diversity) and $4Nr = 2{,}000$ (N is the effective population size; r, the recombination rate per individual per generation). Assuming that θ per site is 8×10^{-4} (ref. 39), N is the traditional 10,000 and $r = 30$ exchanges per 3×10^9 bp per meiosis, the values of θ and $4Nr$ correspond to sequence variation and recombination found in 5 Mb of sequence. We formed an individual's diplotype by choosing two sequences with replacement from the 40,000 haplotypes (thereby maintaining Hardy-Weinberg equilibrium). We randomly paired individuals to form couples, to which offspring were assigned in accordance to the Poisson distribution (mean = 2) under independent assortment of haplotypes.

Our version of the mixed model of inheritance follows[16,40]. For a dichotomous trait, the model assumes that the trait is due to an underlying, but unobservable, liability scale (y) to which mendelian inheritance of a single gene (l), multifactorial transmission (*i.e.*, other genes; c) and random environmental effects (e) contribute additively and independently: $y = l + c + e$. Affectation status is defined by a threshold Z on the liability scale such that all individuals with a liability value above Z are considered affected. For parents, the variables c and e are random variables chosen from normal distributions with means zero and variances C and E, respectively, such that $C + E = 1$. The major locus has two alleles, G and g with the disease-associated allele g having frequency q. (In our simulations, the actual disease-causing mutation was chosen to be the most central site whose minor allele frequency was between $q \pm 0.2q$.) The difference, in units of s.d., between the means of the liability distributions of the two homozygous classes is t; the degree of dominance is d. Therefore, l equals 0, td or t depending on whether the parent has genotype GG, Gg or gg at the major locus. For children, y is determined in the same way except c is obtained by summing the midparental (average of the two parental values) c value and a random number chosen from a normal distribution with mean zero and variance $C/2$.

In simulations of the mixed model of inheritance assuming $C = E$, a user must input q, t, d and Z or q, genotype risk ratio (GRR) and Z. (Note in the latter case, Z has no bearing on the impact of the disease-causing mutation. It merely influences the number of individuals who need to be simulated to acquire a given number of TDT trios.) The formal relationship between the

mixed model parameters and GRR, as with sibling relative risk (λ_s) and heritability (h^2) is developed below.

GRR is the increased chance that an individual with a particular genotype will develop a given disease. Suppose the risk for individuals of genotype *Gg* is γ times greater than the risk for individuals with genotype *GG*: GRR = γ. Under a multiplicative assumption for two *g* alleles, the GRR for genotype *gg* is γ^2. To clarify the relationship of these GRR parameters to the mixed model of inheritance, we first write out the disease prevalence using the latter's formalism:

$$K = \Phi(-Z)p^2 + \Phi(-Z+dt)2pq + \Phi(-Z+t)q^2 \quad (1)$$

where K is the prevalence, p is $1 - q$ and Φ is the cumulative distribution function of a standard normal distribution. Dividing equation 1 by the probability of disease given genotype *gg*, Φ $(-Z + t)$, we obtain the following equation:

$$\begin{aligned} K' &= K/\Phi(-Z+t) \\ &= (\Phi(-Z)/\Phi(-Z+t))p^2 + (\Phi(-Z+dt))/\Phi(-Z+t))2pq + q^2 \end{aligned} \quad (2)$$

Simply, $(\Phi(-Z)/\Phi(-Z + t))$ and $(\Phi(-Z + dt))/\Phi(-Z + t))$ are γ^2 and γ, respectively. Given γ, λ_s easily follows. If this single locus were the only genetic contribution to disease ($C = 0$), λ_s would be the increased risk of developing a disease for a sibling of an affected individual over the population prevalence. If multiple loci contribute, the parameter (call it λ_{as} for the $C > 0$ case) is the increased risk attributable to this one locus. From Risch and Merikangas[1],

$$\lambda_{as} = \left(1 + \frac{pq(\gamma - 1)^2}{2(p + \gamma q)^2}\right)^2$$

Risch and Merikangas actually use K' as prevalence throughout their 1996 paper, because the probability of disease given genotype *gg* cancels out of the final equations used to calculate power.

Finally, we may calculate h^2, which, in the absence of other loci ($C = 0$) and for an additive locus ($d = 1/2$), is the narrow sense heritability of the trait, or the extent to which liability is determined by the additive effects of genes transmitted from parents to offspring. In the presence of other genetic loci, the parameter (call it h_a^2) can be interpreted as the proportion of phenotypic variance explained by the additive effects of this locus. From Falconer & Mackay (ref. 41), we have the equation

$$h_a^2 = 2(m_S - m)/i \quad (3)$$

where m is the mean deviation of the liability of the general population, m_S is the mean deviation for siblings of affected individuals and i is the mean deviation of affected individuals from m. Substituting the parameters already derived into equation 3, we obtain the following equation:

$$h_a^2 = 2K(-\Phi^{-1}(K) + \Phi^{-1}(K\lambda_s))/f(-\Phi^{-1}(K)) \quad (4)$$

where f is the probability density function of a standard normal distribution[40].

Table 6 shows input and tabulated parameters related to the simulations underlying **Tables 1–4**. To derive the values in **Table 6**, for a given set of γ, q and Z and assuming $C = E = 1/2$, the above equations may be used to calculate analytically λ_{as}, t, d and K. We calculated h_a^2 with $C = 0$ and $E = 1/2$. We derived λ_{as}, λ_s and h^2 by tabulating the affected individuals among the simulated individuals and making appropriate calculations with information regarding family structures. The former value assumes a multiplicative model for the penetrance of unlinked loci[42].

Because association studies in the future will use SNPs characterized in public databases, we simulated our marker sets to reflect the bias in allele frequency spectrum of SNPs from those databases. To do so, we assumed that an allele with derived allele frequency x in the general population was contained in the human SNP database dbSNP with probability $p(x)$. To derive $p(x)$, we noted that dbSNP was largely created by aligning random small reads (generally 500 bases or less) against the human genome. Assuming that the number of reads that align at any given SNP is Poisson-distributed (a reasonable assumption given the enormous size of the genome relative to the size of each individual read) with mean η, we calculated the probability that any given SNP with derived allele frequency x is contained in dbSNP as follows:

$$\begin{aligned} p(x) &= \sum_{i=0}^{\infty} P(\text{SNP is in database} \mid i \text{ reads } P(i \text{ reads}) \\ &= \sum_{i=0}^{\infty} \frac{e^{-\eta}\eta^i}{i!}[1 - x^{i+1} - (1-x)^{i+1}] \\ &= \sum_{i=0}^{\infty} \frac{e^{-\eta}\eta^i}{i!} - xe^{-\eta(1-x)} \sum_{i=0}^{\infty} \frac{e^{-\eta}(\eta x)^i}{i!} \\ &\quad - (1-x)e^{-\eta x} \sum_{i=0}^{\infty} \frac{e^{-\eta(1-x)}[\eta(1-x)]^i}{i!} \\ &= 1 - xe^{-\eta(1-x)} - (1-x)e^{-\eta x} \end{aligned}$$

We chose sites consistent with the above frequency distribution and thinned the marker set by first dropping sites whose minor allele frequency was less than 0.2 and then dropping sites at random until the density was 1 SNP per 10 kb.

We selected TDT trios at random based on the affectation status of offspring up to a specified sample size. We derived multiple trios from multiplex sibships. We phased diplotypes using hap2 (ref. 32) at 10,000 iterations, with the first 5,000 discarded as 'burn-in' and the remainder thinned by storing every 20th iteration. We then subjected the phased trios to our exhaustive allelic algorithm and adjusted P values of the alleles for multiple tests by permutation tests comprised of 11,000 iterations. For a given allele, we increased the numerator and denominator of the corresponding adjusted P value by one (call the result p') to account for Monte Carlo error[43]. Given the fact that hap2 may be modified to output a posterior distribution of haplotypes, strictly speaking, the P values should have been computed over these distributions rather than with a single realization of the phasing program. Given the high accuracy of haplotyping trios[32] and the computational burden required by the former method, however, the latter procedure was adopted instead.

Because we simulated ~5 Mb of contiguous sequence, we resorted to an analytical correction method to account for genome-scale multiple testing. As of human genome build 34, the total size of the sequenced human genome is ~2,800 Mb. To account for the multiple tests from 2,800/5 other 5-Mb regions, we applied the Sidák correction[44] to the p' value to determine the final adjusted P value (call it p^*): $p^* = 1 - (1 - p')^{2,800/5}$.

We considered p^* to be statistically significant at the nominal 0.05 value. The Sidák correction is conservative but slightly less so than the more popular Bonferroni method. As with the latter, the former treats each 5-Mb region as independent, which is not true for adjacent regions of the genome. We used analogous procedures for single-SNP analysis corrected by permutation tests.

Evidence that the permutation procedure of EATDT produces the proper type I error rate is given in **Supplementary Figure 1** online.

Old Order Mennonite data. We mapped the 4,244 SNPs of each member of the 35 trios onto build 34 of the human genome. We created bit words separately for each of the 22 autosomes, because creating bit words on the whole data set would have generated meaningless interchromosomal alleles. We aggregated the bit words and analyzed them as we analyzed the simulations. No post-permutation test adjustment (Sidák correction) was necessary. To break ties among alleles with the same P value, we reported any individual SNPs that existed. Otherwise, we chose the longest allele.

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

ACKNOWLEDGMENTS

We thank K. Broman and D. Valle for their insightful comments and discussions regarding this work and J. Kloss and C. Kashuk for their technical assistance. S.L. is supported by the Medical Scientist Training Program and is a student of the Predoctoral Training Program in Human Genetics at Johns Hopkins School of Medicine. A.C. and D.J.C. are supported by project grants from the US National Institutes of Health.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 11 August; accepted 28 September 2004
Published online at http://www.nature.com/naturegenetics/

1. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
2. Collins, F.S., Guyer, M.S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
3. Thomas, J.W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
4. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
5. Kruglyak, L. & Nickerson, D.A. Variation is the spice of life. *Nat. Genet.* **27**, 234–236 (2001).
6. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
7. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
8. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
9. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
10. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
11. Chapman, J.M., Cooper, J.D., Todd, J.A. & Clayton, D.G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* **56**, 18–31 (2003).
12. Long, A.D. & Langley, C.H. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **9**, 720–731 (1999).
13. Churchill, G.A. & Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971 (1994).
14. Kennedy, G.C. *et al.* Large-scale genotyping of complex DNA. *Nat. Biotechnol.* **21**, 1233–1237 (2003).
15. Hudson, R.R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
16. Morton, N.E. & MacLean, C.J. Analysis of family resemblance. 3. Complex segregation of quantitative traits. *Am. J. Hum. Genet.* **26**, 489–503 (1974).
17. Hudson, R.R. The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**, 611–631 (1985).
18. de la Chapelle, A. & Wright, F.A. Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proc. Natl. Acad. Sci. USA* **95**, 12416–12423 (1998).
19. Terwilliger, J.D. & Weiss, K.M. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotechnol.* **9**, 578–594 (1998).
20. Liang, K.Y., Hsu, F.C., Beaty, T.H. & Barnes, K.C. Multipoint linkage-disequilibrium-mapping approach based on the case-parent trio design. *Am. J. Hum. Genet.* **68**, 937–950 (2001).
21. McPeek, M.S. & Strahs, A. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.* **65**, 858–875 (1999).
22. Service, S.K., Lang, D.W., Freimer, N.B. & Sandkuijl, L.A. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am. J. Hum. Genet.* **64**, 1728–1738 (1999).
23. Morris, A.P., Whittaker, J.C. & Balding, D.J. Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Am. J. Hum. Genet.* **67**, 155–169 (2000).
24. Liu, J.S., Sabatti, C., Teng, J., Keats, B.J. & Risch, N. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.* **11**, 1716–1724 (2001).
25. McCallion, A.S. *et al.* Genomic variation in multigenic traits: Hirschsprung disease. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 373–381 (2003).
26. Puffenberger, E.G. *et al.* Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. *Hum. Mol. Genet.* **3**, 1217–1225 (1994).
27. Chakravarti, A. & Lyonnet, S. Hirschsprung disease. in *The Metabolic and Molecular Bases of Inherited Disease* (ed. Charles R.S.) (McGraw-Hill, New York, 2001).
28. Edery, P. *et al.* Mutations of the RET proto-oncogene in Hirschsprung's disease. *Nature* **367**, 378–380 (1994).
29. Romeo, G. *et al.* Point mutations affecting the tyrosine kinase domain of the RET proto-oncogene in Hirschsprung's disease. *Nature* **367**, 377–378 (1994).
30. Jeffreys, A.J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29**, 217–222 (2001).
31. Puffenberger, E.G. *et al.* A missense mutation of the endothelin-B receptor gene in multigenic Hirschsprung's disease. *Cell* **79**, 1257–1266 (1994).
32. Lin, S., Chakravarti, A. & Cutler, D.J. Haplotype and missing data inference in nuclear families. *Genome Res.* **14**, 1624–1632 (2004).
33. Jannot, A.S., Essioux, L., Reese, M.G. & Clerget-Darpoux, F. Improved use of SNP information to detect the role of genes. *Genet. Epidemiol.* **25**, 158–167 (2003).
34. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
35. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
36. Weiss, K.M. & Terwilliger, J.D. How many diseases does it take to map a gene with SNPs? *Nat. Genet.* **26**, 151–157 (2000).
37. Watterson, G.A. On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
38. Hudson, R.R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
39. Halushka, M.K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22**, 239–247 (1999).
40. Falconer, D.S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29**, 51–76 (1965).
41. Falconer, D.S. & Mackay, T.F.C. *Introduction to Quantitative Genetics* (Longman, Essex, 1996).
42. Risch, N. Assessing the role of HLA-linked and unlinked determinants of disease. *Am. J. Hum. Genet.* **40**, 1–14 (1987).
43. Davison, A.C. & Hinkley, D.B. *Bootstrap Methods and Their Application* (Cambridge University Press, Cambridge, 1997).
44. Sidak, Z. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *J. Am. Stat. Assoc.* **62**, 626–633 (1967).