European Journal of Human Genetics (2003) 11, 760–769

www.nature.com/ejhg

ARTICLE

Global survey of haplotype frequencies and linkage disequilibrium at the RET locus

Pratima Chattopadhyay[1], Andrew J Pakstis[1], Namita Mukherjee[1], Sudha Iyengar[2], Adekunle Odunsi[3], Friday Okonofua[4], Batsheva Bonne-Tamir[5], William Speed[1], Judith R Kidd[1] and Kenneth K Kidd*[,1]

[1]*Department of Genetics, Yale University School of Medicine, New Haven, CT 06520-8005, USA;* [2]*Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA;* [3]*Department of Gynecological Oncology, Roswell Park Cancer Institute, Buffalo, NY, USA;* [4]*Department of Obstetrics and Gynecology, Faculty of Medicine, University of Benin, Benin City, Nigeria;* [5]*Department of Genetics, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel*

We have constructed haplotypes based on normal variation at six polymorphic sites–five single nucleotide polymorphisms (SNPs) and one short tandem repeat polymorphism (STRP)–at the *RET* locus for samples of normal individuals from 32 populations distributed across the major continental regions of the world. The haplotyped system spans 41.6 kilobases and encompasses most of the coding region of the gene. All of the markers are polymorphic in all regions of the world and in most individual populations. Expected heterozygosities for the six-site haplotypes range from 82 to 94% for all populations studied except for two Amerindian groups from the Amazon basin at 61 and 76%. Individual populations had from four to eight haplotypes with frequencies exceeding 5%. In general, African, southwest Asian and European groups have the highest numbers of total and of commonly occurring haplotypes; the lowest numbers are observed in Amerindian populations. Overall linkage disequilibrium (LD) for the five SNP sites was very significant ($P \leqslant 0.001$) for all the non-African populations, but significant at that level for only one of the seven African populations. In general, the permutation-based ξ coefficient that quantifies overall LD tends to increase the farther the population is from Africa, but variability of this measure of LD is often large within geographic regions. Pairwise LD measures among the SNPs also show considerable variation among populations. Association of STRP alleles with the SNP-defined background haplotypes is generally higher outside of Africa than in Africa, but is highly variable.
European Journal of Human Genetics (2003) **11**, 760–769. doi:10.1038/sj.ejhg.5201036

Keywords: haplotypes; linkage disequilibrium; population genetics; polymorphisms

Introduction

RET is the gene symbol for the ret proto-oncogene located near the centromere of chromosome 10 in band q11.2. The protein encoded by *RET* is the common component of the receptors for the GDNF family of neurotrophic factors.[1–3] Gain of function mutations in *RET* result in dominant oncogenic conversion causing MEN2 (types A and B).[4,5] Yet, mutations in *RET* resulting in loss of biological activity are associated with Hirschsprung disease.[6,7] The early linkage studies also showed that *RET* was within a region of markedly reduced recombination.[8] The entire *RET* genomic sequence has been cloned in a contig of cosmids encompassing 150 kilobases (kb), from the STRP sTCL-2 to the region upstream of the *RET* promoter.[9] The entire genomic sequence of the region is now available (see GenBank accessions AJ243297, AC010864, and AL591116).

*Correspondence: Dr KK Kidd; Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06520-8005, USA. Tel.: +1 203 785 2654; fax: +1 203 785 6568;
E-mail: Kidd@biomed.med.yale.edu
Received 2 December 2002; revised 9 April 2003; accepted 11 April 2003

Genetic studies of complex disorders have been shown to be more powerful if linkage disequilibrium (LD) exists between susceptibility alleles and normal genetic markers.[10] However, little is known about the magnitude and extent of LD in humans except that it varies among loci studied and among populations (eg see Tishkoff *et al*,[11] Kidd KK *et al*,[12] Kidd JR *et al*,[13] Stephens *et al*,[14] Reich *et al*,[15] and Osier *et al*[16]). LD can be evaluated by a variety of statistics[17,18] that are functions of the haplotype frequencies in the population. Thus, the starting point for studies of LD is determining haplotype frequencies. The expectation–maximization (EM) algorithm gives accurate estimates of the frequencies of common haplotypes[19–21] especially when there is significant disequilibrium. Those haplotype frequencies can also provide information on evolutionary histories, beyond what can be learned from individual markers.[16,22] We are studying *RET* both in order to understand the evolutionary histories of normal allelic variation at this locus and as an example of a centromeric locus in a region of known reduced recombination. In this study, we examine the haplotype frequencies and LD relations of six *RET* polymorphisms (Figure 1) in 32 populations distributed around the world. These data reinforce the growing consensus that African populations have significantly less LD than non-African populations.[12,14,22,23]

Materials and methods

Populations sampled

The 32 populations we have typed for six markers at *RET* are listed by geographic region in Table 1. Sample sizes range from 23 individuals (Nasioi) to 118 individuals (Irish) with a mean sample size of about 53 individuals. Descriptive information and literature citations for these population samples can be found in the allele frequency database ALFRED (http://alfred.med.yale.edu/) under the UIDs in Table 1.

All samples were collected with informed consent from the participants and approval from the appropriate institutional review boards. The DNA in this study was purified by means of standard phenol–chloroform extraction and ethanol precipitation[24] from Epstein–Barr virus–transformed lymphoblastoid cell lines[25] for all samples.

Polymorphic sites and typing protocols

The six *RET* markers typed for this study include five biallelic single nucleotide polymorphisms (SNPs) – intron 1 G/C SNP (this study), exon 2 *Hae*III,[26] exon 13 *Taq*I,[27] exon 15 *Rsa*I,[28] and intron 19 *Taq*I (this study) – and a 13-allele $(CA)_n$ short tandem repeat polymorphism (STRP)[29] within intron 5, altogether spanning 41.6 kb as shown in Figure 1. All typings were PCR-based, using the primers and protocols described in ALFRED and at the Kidd Lab Web Site.

The three coding region SNPs are synonymous changes described previously. The intron 1 G/C SNP at nucleotide position 8019 of intron 1 was identified by resequencing in our laboratory and was typed on all samples by fluorescence polarization.[30] The SNP within intron 19 is a G/A SNP at position 765 of the intron and alters a *Taq*I restriction site. This SNP was noted in *in silico* analysis of GenBank sequences and validated in African-American and European-American samples. For the four restriction site markers, the PCR product was digested with the appropriate enzyme, according to the manufacturers' protocols, and the fragments were electrophoresed on agarose gels and stained with ethidium bromide. For the intron 5 $(CA)_n$ STRP, the amplification products were run on a 5% polyacrylamide gel on an ABI 377 DNA sequencer. Fragment sizes were determined by the Genescan and Genotyper software. All typing results were entered, as individual phenotypes, into PhenoDB2, our client–server database system for genetic marker data.[31]

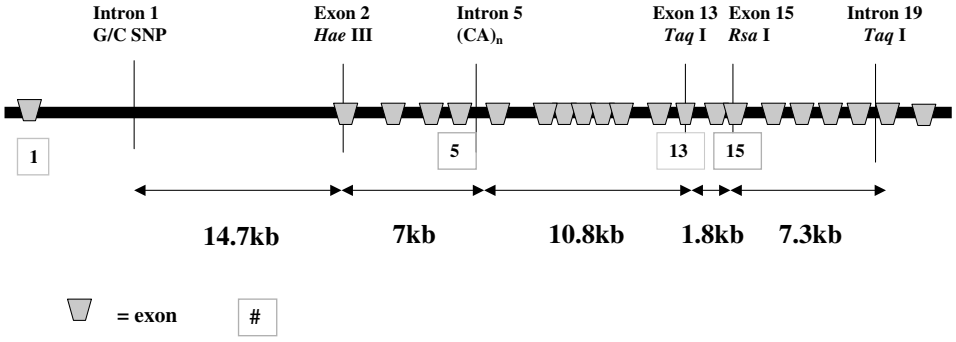

Figure 1 Physical map of the *RET* locus and positions of the polymorphisms studied. The ALFRED UIDs for these sites and haplotypes are as follows: intron 1 G/C SNP: SI000753P; exon 2 *Hae*III: SI000194O; intron 5 $(CA)_n$: SI000767U; exon13 *Taq*I: SI000160H; exon 15 *Rsa*I: SI000161I; intron 19 *Taq*I SI000695U; five-SNP haplotype: SI000861P; six-SNP haplotype: SI000860O.

Table 1 Frequencies for the nine most frequent haplotypes

Geographic region	*Population*	*2n*	*Het.*	*C2212*	*G2212*	*G1111*	*C2222*	*G2112*	*C2112*	*G1212*	*G2111*	*G2222*	*Residual*	*ALFRED Sample UID*
Africa	Biaka pygmies	138	0.73	0.315	0.391	0	0.029	0.027	0.093	0.071	0	0	0.074	SA000005F
	Mbuti pygmies	78	0.76	0.267	0.370	0	0	0.066	0.139	0	0.025	0.038	0.095	SA000006G
	Yoruba	156	0.79	0.138	0.251	0	0	0.305	0.139	0	0	0.110	0.056	SA000036J
	Ibo	94	0.72	0.229	0.452	0	0.025	0.105	0.076	0	0.021	0.080	0.012	SA000099S
	Hausa	78	0.78	0.341	0.249	0	0	0.149	0.052	0.026	0	0.141	0.043	SA000100B
	Ethiopian Jews	62	0.83	0.236	0.173	0.055	0	0.227	0.041	0.113	0	0.113	0.042	SA000015G
	African Americans	180	0.78	0.246	0.354	0	0.102	0.181	0.057	0.011	0.017	0.009	0.023	SA000101C
SW Asia	Yemenite Jews	86	0.86	0.218	0.104	0.225	0.127	0.087	0	0.063	0.070	0.020	0.086	SA000016H
	Druze	148	0.84	0.181	0.173	0.163	0.241	0.044	0.023	0.079	0.020	0	0.076	SA000008I
Europe	Adygei	108	0.86	0.148	0.265	0.133	0.158	0.057	0.024	0.022	0.012	0.010	0.171	SA000017I
	Russians	96	0.83	0.158	0.298	0.203	0.103	0.038	0.015	0.018	0.014	0	0.152	SA000019K
	Finns	72	0.81	0.125	0.148	0.162	0.181	0.034	0	0.296	0.040	0	0.014	SA000018J
	Danes	102	0.83	0.243	0.204	0.155	0.139	0.042	0	0.138	0	0.030	0.050	SA000007H
	Irish	226	0.86	0.194	0.251	0.099	0.059	0.070	0.049	0.097	0.021	0.073	0.088	SA000057M
	Euro Americans	184	0.83	0.300	0.228	0.109	0.065	0.083	0.025	0.079	0.025	0.045	0.039	SA000020C
East Asia	SF Chinese	118	0.74	0.231	0.194	0.404	0.055	0	0.017	0.018	0.043	0.019	0.019	SA000009J
	TW Chinese	98	0.77	0.333	0.079	0.332	0	0	0.021	0.043	0.031	0.022	0.139	SA000001B
	Hakka	82	0.74	0.304	0.132	0.375	0	0.013	0	0.014	0.076	0.023	0.064	SA000003D
	Japanese	102	0.79	0.102	0.165	0.383	0.102	0.020	0.009	0.063	0.103	0.010	0.043	SA000010B
	Ami	80	0.66	0.126	0.136	0.537	0.111	0	0	0	0.050	0.014	0.026	SA000002C
	Atayal	84	0.72	0.142	0.117	0.456	0	0	0	0	0.177	0	0.108	SA000021D
	Cambodians	50	0.71	0.319	0.160	0.399	0	0	0.021	0	0	0	0.101	SA000022E
Pacific	Nasioi	46	0.73	0.454	0.103	0.190	0	0	0	0.108	0.027	0	0.116	SA000012D
	Micronesians	74	0.86	0.176	0	0.118	0	0.126	0.216	0	0.178	0	0.187	SA000063J
Siberia	Yakut	102	0.80	0.202	0.111	0.348	0.147	0.040	0.026	0	0.081	0	0.045	SA000011C
North America	Cheyenne	112	0.82	0.211	0.081	0.297	0.135	0	0	0	0.110	0.042	0.123	SA000023F
	Pima Arizona	102	0.82	0.278	0.213	0.075	0.212	0.025	0	0	0.057	0.059	0.082	SA000025H
	Pima Mexico	106	0.73	0.178	0.187	0.045	0.441	0	0.020	0	0.037	0	0.092	SA000026I
	Maya	102	0.84	0.158	0.131	0.181	0.271	0.010	0	0.015	0.050	0.012	0.172	SA000013E
South America	Ticuna	130	0.69	0.292	0.015	0.446	0.169	0	0	0.039	0.008	0	0.031	SA000027J
	R. Surui	94	0.42	0.085	0	0.127	0.744	0	0	0.011	0	0	0.034	SA000014F
	Karitiana	110	0.56	0.631	0	0	0.132	0	0.159	0	0	0	0.077	SA000028K

The haplotype labels consist of the alleles present at each of the individual sites, listed in chromosomal order as in Figure 1. The intron 1 G/C SNP is represented by the nucleotide (see text). The alleles at the four SNPs altering enzyme sites are labeled as '1' indicating the restriction site is absent and '2' indicating the restriction site is present. The residual column gives the total frequency of all the remaining haplotypes in the population. Het.=heterozygosity.

Determining ancestral alleles

The primate ancestral alleles for the three coding SNPs were determined by comparing the homologous sequences obtained from samples of other apes using the logic in Iyengar *et al*.[32] The PCR primers used for typing the human polymorphisms were used to generate template for sequencing from genomic DNA from one chimpanzee, one gorilla, and one orangutan for each site. For the fourth site, intron 19 *Taq*I site, PCR products of two gorillas and two orangutans were tested by digestion with *Taq*I. Two chimpanzee samples did not amplify with the same primers. The ancestral allele for the intron 1 G/C SNP has not yet been determined.

Statistical methods

Allele frequencies at the individual sites were calculated by gene counting. The assumption of Hardy–Weinberg ratios was tested for the separate sites by means of an auxiliary program, FENGEN.[13] Variation in allele and haplotype frequencies across populations was measured as F_{st}

estimated as $\sigma_p^2/(\bar{p}\bar{q})$ for each biallelic site and as the weighted average of the standardized variance for each allele for the STRP.[33] For each site and the haplotype, expected heterozygosities were computed as $1-\Sigma p_i^2$ where the p_i are the individual allele frequencies. The multi-site haplotype frequency estimates were calculated with HAPLO,[34] which implements the EM algorithm. The HAPLO program also calculates two kinds of standard error estimates, jack-knife and binomial. Using the haplotype frequency estimates, pairwise LD coefficients were computed both as the conventional pairwise D' values[35] and as Δ^2.[17] The HAPLO/P[18] and PERMSTAT programs were used to compute the overall LD values in the form of ξ coefficients for the five-SNP and the six-site haplotypes. The permutation-based calculations also provide a test of whether overall disequilibrium is statistically significant.[18] The HAPLO/P program was also used for the group test of whether the STRP showed significant LD against the 32 background haplotypes defined by the five SNPs, as discussed elsewhere.[18,36,37]

Results

Primate ancestral alleles

Based on identical sequences for the three non-human primates at the three coding SNP sites, the site-absent or 'noncutting' alleles, coded as '1', are ancestral at the exon 15 *Rsa*I (CTAC→GTAC) and the exon 13 *Taq*I (GCGA→TCGA) sites, while the site-present or 'cutting' allele, coded as '2', is ancestral for the exon 2 *Hae*III (GGCC→AGCC) site. The intron 19 *Taq*I site also has a site-present or cutting allele ancestral based on the presence of the restriction site in gorillas and orangutans. The primate sequences for the three coding SNPs have been submitted to GenBank (AF520976-78, AF520980-82, and AF520984-86).

Allele frequencies at individual sites

Marker typings for the six sites have been collected on a total of 1704 individuals. Typing was more than 98% complete across all markers and populations. None of the data deviated significantly from the Hardy–Weinberg expectation. Frequencies and standard errors of all six polymorphisms are given in ALFRED. Allele frequencies for the five biallelic sites are graphed in Figure 2. While frequencies of alleles at the SNPs generally do not differ much among populations within the same geographic region, there are occasional exceptions obvious in Figure 2. Allele frequency variation globally is highly significant at all five sites. With the following few exceptions, all five SNPs are polymorphic in all 32 populations. In the samples of Yoruba and Ibo, only the G allele (site present) occurs at the exon 2 *Hae*III site. In the Atayal and Cambodian samples, only the C allele (site absent) is found at the exon 15 *Rsa*I site.

At the intron 5 STRP, 13 different alleles have been seen globally with sizes ranging from 97 to 121 bp in a perfect 2 bp ladder. The frequencies are all in ALFRED. Sequencing of selected homozygotes indicates that this range corresponds to 13–25 tandem repeats of the CA dinucleotide. The numbers and size ranges of alleles are greater in African populations than anywhere else. Only five of the 13 different alleles are globally common: 107, 109, 111, 113, and 115. These account for from 70% of all chromosomes (in the Mexican Pima) up to 95% of all chromosomes (in the Karitiana) with an average around the world of more than 80%. Two of these are always 'common': 109 with a

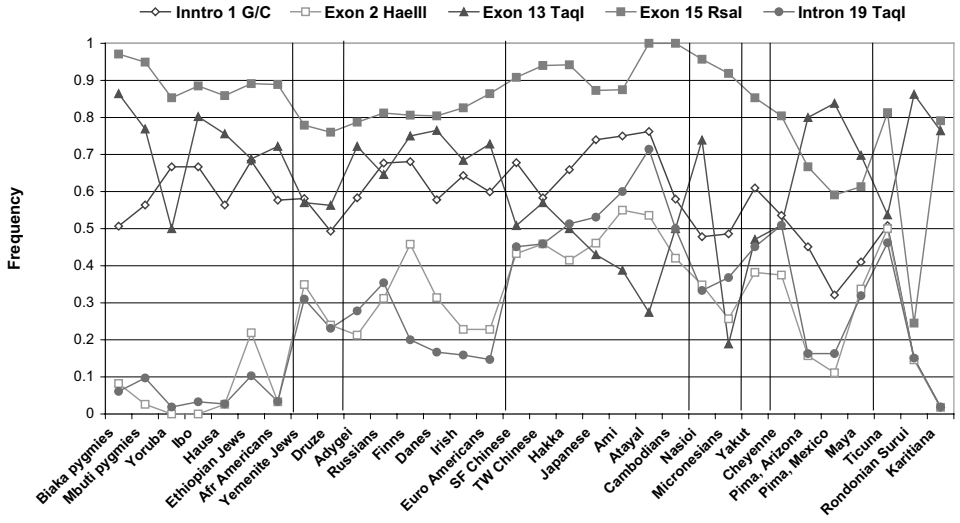

Figure 2 Allele frequencies for *RET* SNPs. Allele frequencies are plotted for the G allele of intron 1 G/C SNP and allele 1 (site absent) of the four other SNPs in 32 populations. Populations are ordered and grouped as in Table 1.

mean frequency of 0.313 and a range of 0.138–0.745, and 111 with a mean frequency of 0.298 and a range of 0.123–0.585. The other three are very rare to absent in at least one population.

The F_{st} values for these polymorphisms, based on these 32 populations, are 0.1 for intron 1 G/C SNP, 0.14 for exon 2 *Hae*III, 0.13 for exon 13 *Taq*I, 0.16 for exon 15 *Rsa*I, 0.18 for intron 19 *Taq*I, and 0.06 for intron 5 $(CA)_n$ STRP. The biallelic F_{st} values are close to the mean value of about 0.14 that we have obtained for the distribution of F_{st} values of more than 100 SNPs that we have studied on the same populations.[38] The F_{st} for the STRP site is close to the mean value of 0.07 reported by Calafell *et al*[39] in a series of 45 STRPs typed on 10 populations that are a subset of the 32 population samples reported here and similar to the mean in the larger study by Rosenberg *et al*.[40]

Haplotype frequency distributions

The maximum-likelihood estimates of the frequencies of the 32 possible five-SNP haplotypes are given in ALFRED for the 32 populations studied. Of the 32 possible haplotypes, nine occur at a frequency above 10% in at least one population (Table 1). Only one haplotype is present in every population sampled: C2212. Haplotype G2212 is common in populations in all regions except the South Americans, while haplotype G1111 is common in populations in all regions other than African. Haplotype heterozygosity is greater than 0.5 for all populations except one South American group and generally greater than 0.7 (Table 1).

Regional averages of the five-SNP haplotype frequencies are graphed in Figure 3. Although there is frequency variation among populations within each region, especially in Africa, these averages make the regional trends more obvious than the data in Table 1. Including the STRP increases the number of possible haplotypes to 416 of which 175 have a nonzero estimate in at least one population. In most of the world, the most common STRP allele, 109, occurs on the two most common SNP haplotypes, C2212 and G2212. Heterozygosity of the six-site haplotype is generally significantly greater for each population than for the five-SNP haplotype, reflecting multiple STRP alleles on many of the SNP-defined haplotypes.

Linkage disequilibrium

The regional pattern of LD is clear for the five-SNP haplotypes: overall nonrandomness, as quantified by ξ, is relatively low and generally nonsignificant in African populations while relatively higher and generally highly significant ($P \leqslant 0.001$) in populations from the rest of the world (Figure 4). However, there are individual populations that do not clearly fit the pattern. Two of the African population samples, Hausa and Ethiopian Jews, show borderline significance ($0.05 > P > 0.01$). Among those population samples significant at $P < 0.001$, the ξ-value quantifies the nonrandomness and shows considerable variation.

The overall LD pattern for the six-site haplotypes is similar to the five-SNP LD pattern but is at a considerably higher level and the statistical significance levels are stronger (data not shown). Figure 4 also shows the association of the STRP with the five-site backbone haplotypes. Except in sub-Saharan African populations, the association is generally significant at $P < 0.01$. Four populations showed marginally significant associations at $0.1 > P > 0.01$: Russians, Japanese, Ami, and R. Surui.

Pairwise LD values, as Δ^2, are given in Table 2. While there is variability within each geographical region for the

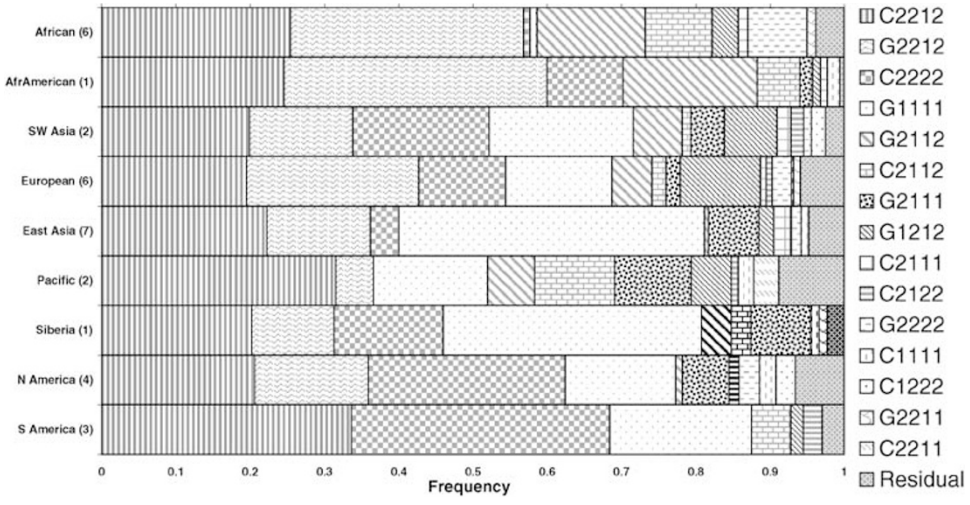

Figure 3 Regional averages of frequencies of common five-SNP haplotypes. Values within parentheses represent number of populations averaged in the geographical region. Data from Table 1.

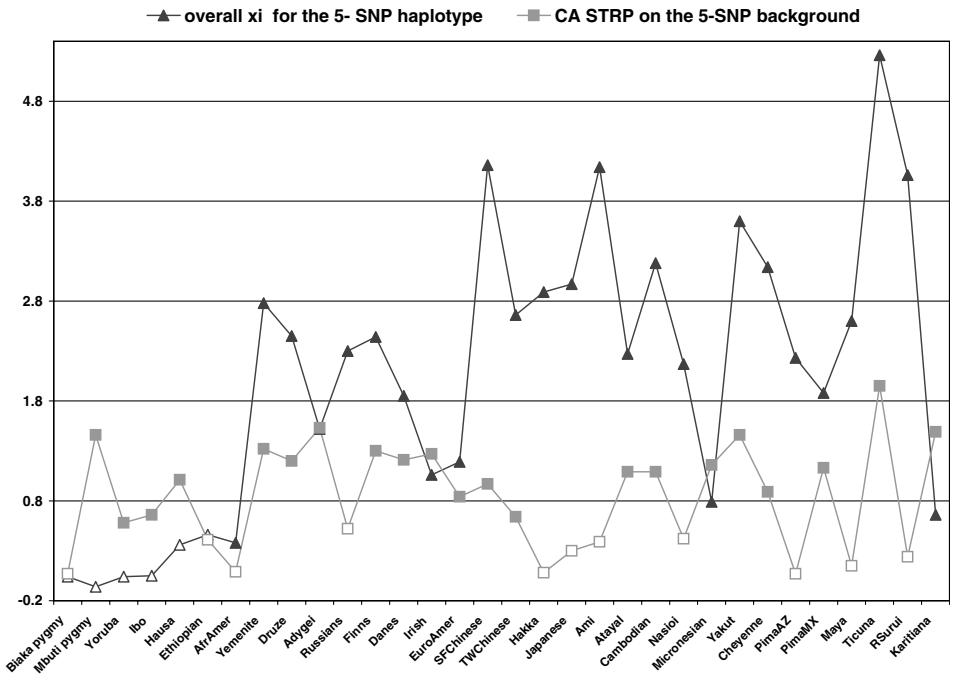

Figure 4 LD measures. ξ-values are plotted for overall nonrandomness for the five-SNP *RET* haplotypes (triangles) and for the association of STRP alleles with the five-site haplotypes (squares). Solid triangles indicate $P \leqslant 0.001$; open triangles indicate $P > 0.02$ except Hausa ($P = 0.002$) and Ibo ($P = 0.004$). Solid squares indicate $P < 0.01$; open squares indicate $P > 0.1$ except Russian ($P = 0.016$), Japanese ($P = 0.06$), Ami ($P = 0.03$), and Rondonian Surui ($P = 0.04$).

pairwise LD results, the overall trend across the 32 populations is remarkably similar for all four independent chromosomal segments (in boldface) defined by the five SNPs although the intervals differ in size (14.7, 17.8, 1.8, and 7.3 kb). These pairwise disequilibrium values are on average smallest in Africans, somewhat larger in Europeans and Eastern Asians, and largest for Native Americans. Interestingly, however, the values for non-African populations are consistently higher for the two longest of the four regions except for the Native American populations. The two smaller regions show the greatest similarity in LD values across populations (Table 3). With two exceptions, the Δ^2-values in Table 2 that are >0.21 have permutation-based *P*-values <0.010 while all 48 LD values $\geqslant 0.5$ have *P*-values <0.001. The two exceptions involve the Nasioi where LD values of 0.3–0.4 were not significant at the 1% level. For the remaining 87 Δ^2-values in the range between 0.21 and 0.50, all have *P*-values <0.010 while 64% have *P*-values <0.001. For the Δ^2-values $\leqslant 0.21$, 30% are significant at the 1% level and a third of these have *P*-values <0.001.

Discussion

Frequency estimates and variation

Extensive genetic variation is shown at all of the six polymorphisms that we have thus far studied at the *RET* locus. The average heterozygosities for the five SNPs are highest in the European samples. This is not surprising since three of the five SNPs were originally identified as RFLPs in samples of individuals of European ancestry. This ascertainment bias is undoubtedly the explanation for that aspect of the global patterns. The heterozygosities are generally lower in the African populations with considerable variation among the populations and SNPs. Outside of Africa, the heterozygosities of the exon 2 *Hae*III and intron 19 *Taq*I sites are highly correlated and generally high (data not shown) as apparent graphically in Figure 2. The correlations are themselves indicators of LD between those two sites. In contrast, the exon 15 *Rsa*I site shows a different pattern with lower heterozygosities in eastern Asian populations and higher heterozygosities in North American Indian populations. Although it is not a striking difference, the heterozygosity for the STRP is higher in Africans than elsewhere. Again, this is not surprising and reflects the well-known higher levels of variation in African populations.[41,42] As a result, the heterozygosities of the six-site haplotypes are, on average, higher in African than European populations. This pervasive genetic diversity makes it feasible to compare LD for the different intervals and populations. The 32 populations studied here are sufficient in number with diverse enough geographic origins to allow many general conclusions about the *RET* locus.

Table 2 Pairwise LD as Δ^2 for all pairs of SNPs

Geographic region	*Population*	***Intron 1–exon 2*** *14.7 kb*	*Intron 1–exon 13* *32.5 kb*	*Intron 1–exon 15* *34.3 kb*	*Intron 1–intron 19* *41.6 kb*	***Exon 2–exon13*** *17.8 kb*	*Exon 2–exon 15* *19.6 kb*	*Exon 2–intron19* *26.9 kb*	***Exon13–exon 15*** *1.8 kb*	*Exon 13–intron 19* *9.1 kb*	***Exon 15–intron 19*** *7.3 kb*
Africa	Biaka pygmies	0.044	0.061	0.031	0.012	0.014	0.003	0.006	0.005	0.011	0.002
	Mbuti pygmies	0.001	0.035	0.041	0.034	0.008	0.100	0.006	0.016	0.000	0.000
	Yoruba	NC	0.015	0.005	0.039	NC	NC	NC	0.043	0.019	0.003
	Ibo	NC	0.002	0.008	0.017	NC	NC	NC	0.034	0.040	0.020
	Hausa	0.021	0.003	0.127	0.058	0.009	0.004	0.001	0.053	0.139	0.007
	Ethiopian Jews	0.017	0.081	0.060	0.006	0.005	0.034	0.404	0.061	0.029	0.014
	African Americans	0.009	0.031	0.123	0.001	0.028	0.004	0.233	0.048	0.091	0.004
SW Asia	Yemenite Jews	0.169	0.348	0.280	0.263	0.174	0.019	0.286	0.083	0.587	0.085
	Druze	0.354	0.101	0.348	0.145	0.169	0.123	0.397	0.086	0.568	0.111
Europe	Adygei	0.194	0.124	0.320	0.017	0.371	0.073	0.261	0.104	0.166	0.006
	Russians	0.157	0.091	0.170	0.136	0.380	0.000	0.579	0.000	0.590	0.000
	Finns	0.398	0.107	0.515	0.119	0.049	0.205	0.121	0.041	0.759	0.061
	Danes	0.280	0.113	0.183	0.100	0.213	0.076	0.351	0.030	0.649	0.049
	Irish	0.090	0.005	0.015	0.017	0.094	0.063	0.297	0.010	0.333	0.024
	Euro Americans	0.169	0.115	0.011	0.091	0.142	0.002	0.353	0.038	0.429	0.004
East Asia	SF Chinese	0.308	0.262	0.105	0.344	0.691	0.041	0.770	0.032	0.905	0.048
	TW Chinese	0.431	0.269	0.018	0.359	0.401	0.000	0.427	0.019	0.658	0.000
	Hakka	0.368	0.362	0.001	0.316	0.491	0.010	0.532	0.002	0.847	0.003
	Japanese	0.224	0.176	0.338	0.180	0.323	0.124	0.384	0.114	0.890	0.099
	Ami	0.407	0.442	0.310	0.417	0.772	0.175	0.721	0.226	0.947	0.214
	Atayal	0.079	0.230	NC	0.210	0.355	NC	0.317	NC	0.942	NC
	Cambodians	0.524	0.195	NC	0.278	0.587	NC	0.724	NC	0.839	NC
Pacific	Nasioi	0.582	0.385	0.049	0.207	0.467	0.085	0.302	0.128	0.704	0.090
	Micronesians	0.015	0.160	0.007	0.300	0.041	0.012	0.199	0.021	0.086	0.015
Siberia	Yakut	0.333	0.473	0.274	0.453	0.550	0.107	0.596	0.194	0.658	0.142
North America	Cheyenne	0.165	0.346	0.098	0.302	0.492	0.110	0.454	0.236	0.931	0.210
	Pima Arizona	0.021	0.202	0.087	0.222	0.349	0.003	0.308	0.055	0.733	0.030
	Pima Mexico	0.197	0.063	0.438	0.287	0.128	0.086	0.552	0.039	0.281	0.114
	Maya	0.093	0.323	0.403	0.194	0.316	0.045	0.449	0.204	0.595	0.131
South America	Ticuna	0.850	0.748	0.245	0.774	0.715	0.201	0.797	0.146	0.910	0.169
	R. Surui	1.000	0.751	0.540	0.759	0.751	0.540	0.759	0.493	0.834	0.481
	Karitiana	1.000	0.059	0.069	1.000	0.059	0.069	1.000	0.026	0.059	0.069

The distance between sites is given for each pair. Independent pairs are in bold. NC indicates noncomputable values due to fixation of allele at one or both site polymorphisms.

Table 3 Correlation coefficients of the pairwise Δ^2-values across populations

	In1–Ex2	***Ex2–Ex13***	***Ex13–Ex15***	*Ex13–In19*
In1–Ex2				
Ex2–Ex13	0.53			
Ex13–Ex15	0.47	0.62		0.47
Ex15–In19	0.60	0.61	0.94	0.53

The interval data are those in the corresponding columns of Table 2. Independent pairs are in bold. In=intron; Ex=exon.

The haplotype frequency estimates in Table 1 that are $\geqslant 10\%$ should be accurate estimates of the common haplotypes in these populations and reflect the global variation in haplotype frequencies. Haplotypes with true frequencies of $\sim 1/2N$ might not have been observed in our samples. Also, haplotypes estimated to be absent in a population might actually have been present in the sample but not unambiguously so. Conversely, haplotypes with estimated frequencies on the order of $1/N$ or smaller may not actually be present.[19] In all, 19 of these populations had more than 40% of the individuals with unambiguous marker phenotypes (homozygous for four or five sites) for the five-SNP haplotypes. A total of 10 populations had 31–39% of the individuals with unambiguous marker phenotypes. The smallest percentages were in the Finns with only 19% unambiguous marker phenotypes. Thus, the haplotype frequency estimates are usually based on considerable phase-known data.

The haplotype frequency data (Table 1) and the regional averages (Figure 3) show that the haplotype frequencies do reflect the geographical clustering of our samples. Therefore, the *RET* polymorphisms will be useful markers for studies of population relationships, especially on a global level. As can be seen in Figure 2, individual sites can be selected as more informative for comparisons among populations from specific regions. For example, the intron 19 *Taq*I site is not informative among sub-Saharan Africans, but the exon 13 *Taq*I site and the exon 15 *Rsa*I site should be very informative among such populations.

Linkage disequilibrium

Although only six sites across 41.6 kb are involved, the *RET* data illustrate many of the complexities of studying LD in human populations. The classic approach to LD is use of a coefficient such as D and D'[35] or other measure[17,43] to quantify the pairwise associations of alleles on chromosomes in a population. Recently, researchers have begun to consider defining the segments of DNA within which only a few haplotypes account for the majority of the chromosomes in a population.[23,42] The value of this 'hap-map' approach is that it may be possible to identify the very small subset of the SNPs in a block that will serve to identify and discriminate among these common haplotypes. Although a short segment, *RET* is interesting to consider from this second perspective. In all non-African populations, there is highly significant overall LD for the five-SNP haplotypes. However, in European and Southwest Asian populations, this does not translate into a subset of these markers being sufficient to identify all common haplotypes. The most common haplotype is never $>30\%$ and except for the Adygei and Russians who require more, six different haplotypes are required to reach a cumulative frequency of 90%. All five SNPs are required to define these six haplotypes. Some other regions of the world require fewer haplotypes to account for the majority of chromosomes in a population because of reduced heterozygosity. This reduced heterozygosity can be explained as ascertainment bias for high heterozygosity of each site in Europeans. However, even where only three haplotypes account for $>80\%$ (eg Native Americans), three of the four SNPs are still required to discriminate among them.

The test of LD of the STRP against the background haplotypes provides evidence on the mutation rate at the STRP and on recent human evolution. The absence of significant LD in African populations but the presence of significant LD in most non-African populations argues that the African populations have existed for a longer time than the non-African populations. The finding of several non-African populations with nonsignificant LD indicates that the mutation rate at the STRP is at least moderate relative to the time since founding of the non-African populations. This is also indicated by the occurrence of more than one STRP allele on evolutionarily derived haplotypes.

In contrast to an expectation of a negative correlation between pairwise LD and interval length (higher LD for shorter intervals, lower LD for longer intervals), we find that the two longer of the four independent intervals have generally higher LD than the two shorter intervals (Table 2). In the African populations, all intervals generally have nonsignificant (or noncalculable) LD values, but the longer internal intervals, intron 1 G/C SNP to exon 2 *Hae*III and exon 2 *Hae*III to exon 13 *Taq*I (17.8 kb), have markedly increased LD for non-African populations. This pattern suggests that the founder effect associated with the expansion of modern humans out of Africa established this pattern of relatively higher LD for the longer regions and lower LD for the shorter regions. Subsequent random genetic drift of different magnitudes for different populations and another founder effect associated with migration into the Americas would then modify this general pattern. The situation is more complex, however, because the pairwise LD spanning the 9.1 kb between exon 13 *Taq*I and intron 19 *Taq*I is generally much higher in all non-African populations than the LD across either of the two internal intervals exon 13 *Taq*I to exon 15 *Rsa*I and exon 15 *Rsa*I to intron 19 *Taq*I. In fact, for many populations this is the largest Δ^2-value. The correlation coefficients in Table 3 show that these two smaller segments have a very similar pattern of LD among populations ($r=0.94$), while neither is as highly correlated with the LD for the region encompassing both ($r=0.47$ and 0.53). Since the same few haplotypes are involved in all cases, the unusual pattern does not relate to recombination but to frequency differences among the haplotypes. Such complex patterns of LD emphasize that LD is a statistical phenomenon, not an inherent property of a segment of DNA and that random genetic drift, more than 'hot spots' of recombination, may be the major factor determining patterns of LD.

In a study of Hirschsprung disease in the genetically isolated Old Order Mennonite community, Carrasquillo *et al*.[44] found a 'block' of LD at the 5′ end of the gene that was strong in chromosomes transmitted to Hirschsprung patients and less pronounced in the untransmitted chromosomes from those families. A more diffuse 'block' was also present in the 3′ part of the gene in both sets of chromosomes. The untransmitted chromosomes are likely to represent random chromosomes from the Old Order Mennonites. The strong LD patterns they found at the 5′ and the 3′ ends of RET may well be due to a founder effect in their sample of Mennonites and differ from what we find with our more limited number of markers in our samples of normal variation in most populations including the European populations. Three of the markers, intron 5 $(CA)_n$, exon 13 *Taq*I, and exon 15 *Rsa*I, were the same as we have typed. Our intron 1 and exon 2 markers fall within

the 5′ block of Carrasquillo *et al* and our exon 13, exon 15, and intron 19 markers fall within their 3′ block. In just over half of the populations we studied, the LD between the intron 1 and exon 2 markers ('intrablock') is less than the LD between the exon 2 and exon 13 marker ('interblock'), two intervals of roughly equal length. As noted above, the LD values for the exon 13 to exon 15 and the exon 15 to intron 19 intervals, which are both within the '3′ block', are generally much smaller even though the interval lengths are shorter.

Recently, a founder haplotype, possibly accounting for many cases of Hirschsprung disease in Spaniards, was described using several SNPs at *RET*.[45] The exon 2 *Hae*III, exon 13 *Taq*I, and exon 15 *Rsa*I in the present paper were included in that study along with others. By extrapolation of the association of these markers with Hirschsprung disease, a possible susceptibility variant close to exon 1 of the gene was postulated. The distance of the extrapolation was 20 kb. It is generally impossible to predict association across such distances because, as shown here for these and other *RET* SNPs, the patterns of LD are complex and do not follow simple regressions with molecular distance. However, the new intron 1 G/C SNP will now allow that hypothesis to be more robustly tested. The LD among markers in normal individuals bears no necessary relation to the association expected for any one marker and a disease susceptibility allele. Therefore, our study does not predict what might be found for association of this intron 1 G/C SNP with Hirschpsrung disease.

Haplotype variation at *RET* contrasts with the pattern seen at DRD2,[12] PAH,[13] and COMT[37] for many of the same population samples. At those other loci, there were more haplotypes and higher heterozygosities in the African populations. However, the data are consistent among all these loci in showing less LD in the African samples than elsewhere. Thus, *RET* is another locus that shows low levels of LD in multiple African populations strengthening the conclusion that low levels of LD is a characteristic of African populations in general. While the SNP-defined haplotype diversity at *RET* does not support an Out-of-Africa model of recent human evolution, the STRP diversity and its disequilibrium with the background, SNP-defined haplotypes (Figure 4) do support an Out-of-Africa model with a significant founder effect in the ancestry of the non-African populations.

Electronic databases cited

ALFRED (ALlele FREquency Database): http://alfred.med.yale.edu/alfred/

Kidd Lab Web Site: http://info.med.yale.edu/genetics/kkidd

Genbank: http://www.ncbi.nlm.nih.gov/Genbank/index.html

Acknowledgements

This work was supported in part by NIH Grants GM57672 (KKK and JRK) and MH62495 (KKK). We thank Valeria Ruggeri and Mellissa MC DeMille for their help in running the sequencing gels and determining the fragment sizes of STRP alleles with the Genescan software. We are indebted to the following people who helped assemble the diverse population collection used in this study: FL Black, LL Cavalli-Sforza, K Dumars, J Friedlaender, D Goldman, E Grigorenko, K Kendler, W Knowler, F Oronsaye, J Parnas, L Peltonen, LO Schulz, and K Weiss. In addition, some of the cell lines were obtained from the National Laboratory for the Genetics of Israeli Populations at Tel Aviv University, Israel, and the African-American samples were obtained from the Coriell Institute for Medical Research, Camden, NJ, USA. Special thanks are due to the many hundreds of individuals who volunteered to give blood samples for studies such as this. Without such participation of individuals from diverse parts of the world, we would have been unable to obtain a true picture of the genetic variation in our species.

References

1 Durbec P, Marcos-Gutierrez CV, Kilkenny C *et al*: GDNF signaling through the RET receptor tyrosine kinase. *Nature* 1996; **381**: 789–793.
2 Treanor JJ, Goodman L, de Sauvage F *et al*: Characterization of a multicomponent receptor for GDNF. *Nature* 1996; **382**: 80–83.
3 Trupp M, Arenas E, Fainzilber M *et al*: Functional receptor for GDNF encoded by the c-ret proto-oncogene. *Nature* 1996; **381**: 785–789.
4 Santoro M, Carlomagno F, Romano A *et al*: Activation of RET as a dominant transforming gene by germline mutations of MEN2A and MEN2B. *Science*. 1995; **267**: 381–383.
5 Eng C: Multiple endocrine neoplasia type 2 and the practice of molecular medicine. *Rev Endocr Metab Disord* 2000; **1**: 283–290.
6 Pasini B, Borrello MG, Greco A *et al*: Loss of function effect of RET mutations causing Hirschsprung disease. *Nat Genet* 1995; **10**: 35–40.
7 Parisi MA, Kapur RP: Genetics of Hirschsprung disease. *Curr Opin Pediatr* 2000; **12**: 601–607.
8 Simpson NE, Kidd KK: The mapping of the locus for multiple endocrine neoplasia type 2A by linkage with chromosome 10 markers. *J Horm Metab Res* 1989; **21**: 5–9.
9 Pasini B, Ceccherini I, Romeo G: RET mutations in human disease. *TIG* 1996; **12**: 138–144.
10 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
11 Tishkoff SA, Goldman A, Calafell F *et al*: A global haplotype analysis of the DM locus: implications for the evolution of modern humans and the origin of myotonic dystrophy mutations. *Am J Hum Genet* 1998; **62**: 1389–1402.
12 Kidd KK, Morar B, Castiglione CM *et al*: A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Hum Genet* 1998; **103**: 211–227.
13 Kidd JR, Pakstis AJ, Zhao H *et al*: Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus (PAH) in a global representation of populations. *Am J Hum Genet* 2000; **66**: 1882–1899.
14 Stephens JC, Schneider JA, Tanguay DA *et al*: Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 2001; **293**: 489–493.
15 Reich DE, Cargill M, Bolk S *et al*: Linkage disequilibrium in the human genome. *Nature* 2001; **411**: 199–204.
16 Osier MV, Pakstis AJ, Soodyall H *et al*: A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. *Am J Hum Genet* 2002; **71**: 84–99.

17 Devlin B, Risch N: A comparison of linkage disequilibrium measures for fine scale mapping. *Genomics* 1995; **29**: 311–322.
18 Zhao H, Pakstis AJ, Kidd JR, Kidd KK: Assessing linkage disequilibrium in a complex genetic system I. Overall deviation from overall association. *Ann Hum Genet* 1999; **63**: 167–179.
19 Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK: The accuracy of statistical methods for estimating haplotype frequencies: an example from the CD4 locus. *Am J Hum Genet* 2000; **67**: 518–522.
20 Fallin D, Schork NJ: Accuracy of haplotype frequency estimation for biallelic loci, via the expectation–maximization algorithm for unphased diploid genotype data. *Am J Hum Genet* 2000; **67**: 947–959.
21 Zhang S, Pakstis AJ, Kidd KK, Zhao H: Comparisons of two methods for haplotype reconstruction and haplotype frequency estimates from population data. *Am J Hum Genet* 2001; **69**: 906–912.
22 Tishkoff SA, Dietzsch E, Speed W *et al*: Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 1996; **271**: 1380–1387.
23 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.
24 Sambrook J: *Molecular cloning: a laboratory manual*. Cold Spring Harbor, NY: Cold Spring Harbor Press, 1989.
25 Anderson MA, Gusella JF: Use of cyclosporin A in establishing Epstein–Barr virus-transformed human lymphoblastoid cell lines. *In Vitro* 1984; **20**: 856–858.
26 Edery P, Attie T, Mulligan LM *et al*: A novel polymorphism in the coding sequence of the human RET proto-oncogene. *Hum Genet* 1994; **94**: 579–580.
27 Tahira T, Shiraishi M, Ishizaka Y *et al*: A Taq I RFLP in the human ret proto-oncogene. *Nucleic Acids Res* 1990; **18**: 7472.
28 Ceccherini I, Hofstra RM, Luo Y *et al*: DNA polymorphisms and conditions for SSCP analysis of the 20 exons of the ret proto-oncogene. *Oncogene* 1994; **9**: 3025–3029.
29 Pasini B, Hofstra RMW, Yin L *et al*: The physical map of the human RET proto-oncogene. *Oncogene* 1995; **11**: 1737–1743.
30 Chen X, Levine L, Kwok PY: Fluorescence polarization in homogeneous nucleic acid analysis. *Genome Res* 1999; **9**: 492–498.
31 Cheung KH, Nadkarni P, Silverstein S *et al*: PhenoDB: an integrated client/server database for linkage and population genetics. *Comput Biomed Res* 1996; **29**: 327–337.
32 Iyengar S, Seaman M, Deinard AS *et al*: Analyses of cross species polymerase chain reaction products to infer the ancestral state of human polymorphisms. *DNA Sequence* 1998; **8**: 317–327.
33 Wright S: *The theory of gene frequencies*. Chicago: University of Chicago Press, 1969, vol II.
34 Hawley ME, Kidd KK: HAPLO: a program using the EM algorithm to estimate the frequencies of multi–site haplotypes. *J Hered* 1995; **86**: 409–411.
35 Lewontin RC: The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics* 1964; **49**: 49–67.
36 Zhao H, Pakstis AJ, Kidd KK, Kidd JR: Overall and segmental levels of linkage disequilibrium. *Am J Hum Genet* 1997; **61** (Suppl): A17.
37 DeMille MMC, Kidd JR, Ruggeri V *et al*: Population variation in linkage disequilibrium across the *COMT* gene considering promoter region and coding region variation. *Hum Genet* 2002; **111**: 521–537.
38 Pakstis AJ, Kidd JR, Kidd KK: A reference distribution of Fst values for biallelic DNA markers. *Am J Hum Genet* 2002; **71** (Suppl): 371.
39 Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK: Short tandem repeat polymorphism evolution in humans. *Eur J Hum Genet* 1998; **6**: 38–49.
40 Rosenberg NA, Pritchard JK, Weber JL *et al*: Genetic structure of human populations. *Science* 2002; **298**: 2381–2385.
41 Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL: High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 1994; **368**: 455–457.
42 Patil N, Berno A, Hinds DA *et al*: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001; **294**: 1719–1723.
43 Ardlie KG, Kruglyak L, Seielstad M: Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 2002; **3**: 299–309.
44 Carrasquillo MM, McCallion AS, Puffenberger EG *et al*: Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nat Genet* 2002; **32**: 237–244.
45 Borrego S, Wright FA, Fernandez RM *et al*: A founding locus within the RET proto-oncogene may account for a large proportion of apparently sporadic Hirschsprung disease and a subset of cases of sporadic medullary thyroid carcinoma. *Am J Hum Genet* 2003; **72**: 88–100.