Molecular Psychiatry (2010) 15, 216–225

www.nature.com/mp

ORIGINAL ARTICLE

The complex global pattern of genetic variation and linkage disequilibrium at catechol-*O*-methyltransferase

N Mukherjee[1], KK Kidd[1], AJ Pakstis[1], WC Speed[1], H Li[1], Z Tarnok[2], C Barta[3], SLB Kajuna[4] and JR Kidd[1*]

[1]*Department of Genetics, Yale University School of Medicine, New Haven, CT, USA;* [2]*Vadaskert Child and Adolescent Psychiatric Clinic, Budapest, Hungary;* [3]*Institute of Medical Chemistry, Molecular Biology and Pathobiochemistry, Semmelweis University, Budapest, Hungary and* [4]*Department of Biochemistry and Molecular Biology, Hubert Kairuki Memorial University, Dar es Salaam, Tanzania*

Genetic variation at the catechol-*O*-methyltransferase (*COMT*) gene has been significantly associated with risk for various neuropsychiatric conditions such as schizophrenia, panic disorder, bipolar disorders, anorexia nervosa and others. It has also been associated with nicotine dependence, sensitivity to pain and cognitive dysfunctions especially in schizophrenia. The non-synonymous single nucleotide polymorphism (SNP) in exon 4—Val108/158Met—is the most studied SNP at *COMT* and is the basis for most associations. It is not, however, the only variation in the gene; several haplotypes exist across the gene. Some studies indicate that the haplotypic combinations of alleles at the Val108/158Met SNP with those in the promoter region and in the 3′-untranslated region are responsible for the associations with disorders and not the non-synonymous SNP by itself. We have now studied DNA samples from 45 populations for 63 SNPs in a region of 172 kb across the region of 22q11.2 encompassing the *COMT* gene. We focused on 28 SNPs spanning the *COMT*-coding region and immediately flanking DNA, and found that the haplotypes are from diverse evolutionary lineages that could harbor as yet undetected variants with functional consequences. Future association studies should be based on SNPs that define the common haplotypes in the population(s) being studied.
Molecular Psychiatry (2010) **15,** 216–225; doi:10.1038/mp.2008.64; published online 24 June 2008

Keywords: SNP; haplotype; population; association; lineage; evolution

Introduction

The catechol-*O*-methyltransferase (*COMT*) gene (MIM 116790) warrants study for several reasons: (1) the enzyme is very important in degradation of catecholamines (for example, dopamine) and hence in neural functioning, (2) polymorphisms within the gene and in the upstream and downstream regions of the gene have been associated with a variety of neuropsychiatric phenotypes[1–5] and (3) some of those polymorphisms produce functionally different forms of the enzyme.[6,7] COMT exists in two different isoforms. The soluble isoform (S-COMT) has 221 amino-acid residues; its translation begins at exon 3 and is regulated by the P1 promoter located in intron 2 (reference Tenhunen *et al.*[8]). The membrane-bound isoform (MB-COMT) has an additional 50 amino-acid residues encoding a transmembrane domain at the N-terminal end of the protein. The MB-COMT is regulated by the P2 promoter located at the 5′ end of the gene.[8] It was shown about 30 years ago that RBC COMT activity varies among siblings and other first-degree relatives showing autosomal recessive inheritance.[9] This variation was later identified as due to a single nucleotide change that results in an amino-acid change from valine to methionine at codon 158 for MB-COMT and codon 108 for S-COMT (Val108/158Met).[10] The Val108/158Met difference affects the thermostability and the activity of both forms of the enzyme[11,12] and has been the most frequently studied polymorphism in *COMT*.

Palmatier *et al.*[13] detailed the inconsistency among results of various association studies with the Val108/158Met single nucleotide polymorphism (SNP) in multiple populations. Association of this SNP to different neuropsychiatric disorders continues to be debatable.[1,4,5,14–19] We have shown in our earlier studies that the non-synonymous SNP and other SNPs in the gene vary in frequencies among populations.[13,20,21] In addition to the catalogued neuropsychiatric disorders, genetic variation in *COMT* has also been studied with respect to pain sensitivity and breast cancer.[22,23]

To establish a more detailed molecular and evolutionary profile of the *COMT* gene we have extended

*Correspondence: Dr JR Kidd, Department of Genetics, Yale University School of Medicine, 333 Cedar Street, PO Box 208005, New Haven, CT 06520-8005, USA.
E-mail: Judith.Kidd@yale.edu
Received 7 January 2008; revised 13 May 2008; accepted 16 May 2008; published online 24 June 2008

our previous studies to a total of 63 SNPs across a 172 kb region in 45 populations ($N = 2335$) that represent most of the geographic regions of the world. In this paper we focus on the 51.7 kb region extending from ~12 kb upstream of the P2 promoter to ~10 kb downstream of *COMT* covered by 28 SNPs with an average spacing of less than 2 kb and a maximum interval of 5.4 kb. Our studies resolve the complexity of the haplotypes across the region into subregions that have simpler patterns of haplotypes wherein each haplotype represents a largely distinct evolutionary lineage throughout the recent dispersal and diversification of modern humans. Diverse combinations of those distinct lineages occur in various populations around the world.

Materials and methods

Populations and DNA samples

We have studied the same 38 population samples that were used by Palmatier *et al.*[20] plus samples from seven additional populations. The additional populations are African Masai and Sandawe of Tanzania, European Archangel'sk Russians and Hungarians, the East Asian Laotians and Koreans, and the South American Quechua from Peru. The 45 populations are from nine major geographical regions—Africa, Southwest (SW) Asia, Europe, Northwest (NW) Asia, Eastern Siberia, East Asia, Pacific Islands, North America and South America. Sample sizes ranged from 44 chromosomes (Masai) to 238 chromosomes (Laotians) with an average of just over 108 chromosomes per population. All populations and sample sizes are given in Supplementary Table S1 by geographic region. DNA was purified from lymphoblastoid cell lines. Samples were collected with informed consent of all the participants.

Markers typed

The 63 SNPs and their relative sequence positions from genome build 36.2 are given in Supplementary Table S2. The SNPs were typed by various methods that include PCR-restriction fragment length polymorphism assay, TaqMan assay and the Illumina platform. For more detailed analyses we have focused on the region immediately encompassing *COMT* because of its possible relevance in various disorders based on association studies. The 28 SNPs on which the current analysis is focused are given in Table 1. For simplicity we will refer to SNPs by number 1–28 in column 1 in this table.

Ancestral alleles

We determined ancestral states of 62 SNPs (one could not be determined) either directly from genotype data on our nonhuman primate samples or by comparing human sequence with the consensus sequences available at GenBank for chimpanzee or other primates. For most of the SNPs, we typed three individuals from each of five species: bonobo, chimpanzee, gorilla, orangutan and gibbon. The ancestral alleles of the SNPs based on our data are given in Table 1 and Supplementary Table S2. We inferred the ancestral states of all those SNPs that could not be typed by our typing methods from the chimpanzee consensus sequence. In cases with gaps in the chimpanzee sequence at the SNP position, the sequence was compared with both rhesus and orangutan sequences.

Statistical methods

Allele frequencies of the SNPs were calculated by gene counting assuming codominant inheritance. All the sites were also tested for Hardy–Weinberg (H-W) ratios by χ^2-test using the program FENGEN[24] or by a permutation-based exact test using HWSIM[25] when small observed numbers make the χ^2 approximation inappropriate. In general 1000 permutations were carried out, but when the *P*-value deviated significantly at the 1% level, 10 000 permutations were done. Expected heterozygosities were estimated as $1-\sum_{P_i}{}^2$. The expectation maximization (EM) algorithm implemented in HAPLO[26] was used to obtain maximum likelihood estimates of haplotype frequencies. The EM algorithm and various alternative methods of haplotype inference based on coalescent and Bayesian theories typically give essentially identical frequency estimates for the haplotypes occurring at common frequencies in a population but can differ for rare haplotypes.[27] The small differences arise because the methods implemented in PHASE[28,29] and similar programs assign genotypes to individuals and thus haplotype frequencies are restricted to units of 1/(2N). Ambiguous situations are 'resolved' by PHASE to a specific genotype in those instances whereas the EM algorithm maintains the ambiguity as a probability altering the haplotype frequency estimate. We confirmed identity of haplotype frequency estimates for our data except for rare haplotypes (<5%) using HAPLO[26] and fastPHASE,[29] and report only the maximum likelihood estimates from HAPLO.

Pairwise linkage disequilibrium (LD) was estimated as r^2 (reference Devlin and Risch[30]) and D′ (reference Lewontin[31]) with significance levels determined by a permutation test.[32] The significance level for each pairwise measure of LD in each population was determined by comparing the maximum likelihood value from HAPLO for haplotype frequencies based on randomly permuting the genotypes at the two markers independently among the individuals in the population 1000 times. Observed likelihoods greater than any of the 1000 permutation values are assigned a significance value of <0.001.

Comparative plots of regions of high LD were done using HAPLOT.[33] The default values for the agglomerative algorithm in HAPLOT were used. Markers are grouped starting with adjacent pairs with r^2 values >0.4 and adjacent SNPs were added if the average pairwise r^2 values of that SNP with others already in the group exceeded 0.3. The resulting groups are represented in the figures as double-headed arrows.

Table 1 Details of 28 SNPs across the *COMT* gene

SNP number	*Site as in ALFRED*	*rs# dbSNP build 127*	*Ancestral allele*	*SNP (strand typed) forward/ reverse*	*Position (genome build 36.2)*	*Distance to next SNP*	*Ancestral SNP frequency range*	*Mean frequency (45 pops)*	*Average Het (45 pops)*	*Fst (45 pops)*
1	rs5993875	rs5993875	G	G/A (F)	18 295 326	4079	0.378–0.952	0.687	0.396	0.080
2	rs4485648	rs4485648	G	G/A (R)	18 299 405	954	0.00–0.600[b]	0.213	0.243	0.209
3	C_2539253	rs9606186	G	G/C (F)	18 300 359	3047	0.396–0.934	0.643	0.433	0.056
4	rs5746848	rs5746848	G	G/A (R)	18 303 406	3740	0.391–0.882[b]	0.588	0.455	0.060
5	rs5748489	rs5748489	C	C/A (F)	18 307 146	946	0.391–0.936	0.695	0.389	0.084
6	Promoter	rs2075507	T	T/C (R)	18 308 092	792	0.410–0.951	0.653	0.427	0.057
7	C_11731880	rs2020917	C	C/T (F)	18 308 884	1237	0.474–1.00	0.814	0.263	0.131
8	rs737865	rs737865	T	T/C (F)	18 310 121	1286	0.475–1.00	0.803	0.282	0.119
9	C_2539273	rs933271	C	C/T (F)	18 311 407	2644	0.088–0.942[b]	0.434	0.433	0.118
10	rs174675	rs174675	G	G/A (R)	18 314 051	3482	0.064–0.878	0.571	0.430	0.122
11	C_3274705	rs5993882	T	T/G (F)	18 317 533	5464	0.525–0.975	0.804	0.284	0.099
12	rs5746849	rs5746849	G	G/A (F)	18 322 997	2180	0.04–0.675[b]	0.409	0.424	0.124
13	C_11804654	rs740603	G	G/A (F)	18 325 177	3160	0.039–0.669[b]	0.397	0.424	0.115
14	C_11804650	rs4646312	T	T/C (F)	18 328 337	676	0.400–1.00	0.757	0.315	0.142
15	C_2539306	rs165722	C	C/T (F)	18 329 013	939	0.390–0.962	0.611	0.432	0.091
16	C_2538746	rs6269	G	G/A (F)	18 329 952	283	0.00–0.583[b]	0.303	0.375	0.102
17	C_2538747	rs4633	C	C/T (F)	18 330 235	33	0.368–0.981	0.641	0.419	0.091
18	C_2255322	rs740602	G	G/A (F)	18 330 268	939	0.711–1.00	0.948	0.085	0.146
19	C_2538750	rs4818	G	G/C (F)	18 331 207	64	0.00–0.593[b]	0.262	0.332	0.123
20	Exon 4 *Nla*III	rs4680	G	G/A (F)	18 331 271	533	0.383–0.991	0.66	0.406	0.092
21	C_7543740	rs769224	G[a]	G/A (F)	18 331 804	2654	0.630–1.0[a]	0.943	0.073	0.088
22	C_2255331	rs174699	T	T/C (F)	18 334 458	1804	0.190–1.00	0.746	0.263	0.305
23	Exon_6BglI	rs362204	DEL	DEL/G (F)	18 336 262	519	0.113–0.909[b]	0.466	0.437	0.122
24	C_2255335	rs165599	G	G/A (F)	18 336 781	242	0.260–0.973	0.566	0.419	0.147
25	C_2255336	rs165728	T	T/C (F)	18 337 023	2450	0.190–1.00	0.739	0.276	0.285
26	rs165815	rs165815	G	G/A (R)	18 339 473	2482	0.118–0.973[b]	0.446	0.384	0.223
27	rs887199	rs887199	G	G/A (F)	18 341 955	5113	0.027–0.882[b]	0.55	0.385	0.223
28	rs887204	rs887204	G	G/A (F)	18 347 068	—	0.237–0.973[b]	0.553	0.424	0.143

Abbreviations: SNP, single nucleotide polymorphism; Het, heterozygosity; #, number.
[a]See text discussion under *Ancestral alleles* section of Results.
[b]Derived allele frequencies high in several populations.

SNPs with very low heterozygosity in any population are skipped, and if within a group, are represented as an open (uncolored) segment of an arrow.

We used the long-range haplotype test (LRH)[34] to test all 63 SNPs for signs of selection using the P2 promoter SNP (SNP 6) and the non-synonymous SNP in exon 4 (SNP 20) in the core haplotypes for all the 45 populations. The unpublished utility P-select (Sheng Gu) was used to implement the LRH analyses.

Results

The data in this study extend our previous studies[20,21] with the addition of 21 SNPs and 7 populations for a total of 28 SNPs in 45 populations extending from ~12 kb upstream of the P2 promoter to ~10 kb downstream of the gene (details of 35 additional SNPs in all 45 populations are reported (Supplementary Table S2) here but not further analyzed). Of the 28 SNPs typed, two involved the promoter regions: SNP 6 (rs2075507 previously rs2097603, the *Hin*dIII polymorphism) in the P2 promoter, and SNP 16 (rs6269) in the P1 promoter. Five coding SNPs were typed including the most studied exon 4 non-synonymous SNP.

Allele frequencies

The allele frequencies for 63 SNPs across the 172 kb region for 45 populations are available in ALFRED (http://alfred.med.yale.edu) and are retrievable using the rs numbers in Supplementary Table S2. As expected, allele frequencies varied among the populations. Of 2835 H-W tests (63 SNPs times 45 populations) performed, 147 tests deviated significantly at the nominal 5% level compared to 142 expected by chance for independent tests whereas 28 tests deviated significantly at the nominal 1% level, exactly what is expected for 2835 independent tests. However, the 63 markers are not independent of each other. We note 18 regions within the 172 kb studied that frequently have high LD and so consider a *P*-value of 0.00006 (0.05/810; due to $18 \times 45 = 810$ 'fully independent' tests) to be a conservative Bonferroni correction. By this criterion, none of the tests significantly deviates from H-W. This is consistent with the pattern of the 28 nominally significant tests scattered among the populations and SNPs. Thus, we conclude that none of the population samples is significantly stratified and that none of the SNP assays gives a significant systematic typing error.

Ancestral alleles

Of the 28 SNPs, 16 SNPs gave results for at least some of the nonhuman apes. The genotypes for the apes are given in Supplementary Table S3. For SNP 21, rs769224 (A/G), frequency of the 'G' allele is high in all the populations. Both chimpanzee and bonobo samples were homozygous for the 'A' allele, whereas the gorilla, gibbon and orangutan samples were homozygous for the 'G' allele. This fixed difference between gorilla and the two chimpanzee species presents a more complex inference problem. As we cannot be certain of the order of mutation for the polymorphism, we have examined LD with the immediately flanking SNPs. Both have high frequencies for the ancestral allele in Africa and are nearly completely associated with the 'G' allele at SNP 21. It therefore seems probable that 'G' is the ancestral allele in humans. We note that a fraction of the human genome is more similar to gorilla than to chimpanzee.[35] For one SNP (rs885987; SNP 12 in Supplementary Table S2), our nonhuman primate samples did not give any result. The nucleotide at the same position is a 'C' in the nonhuman primate consensus sequences instead of the 'A' or 'G' polymorphism in humans. No chimpanzee sequence data exist for the immediate SNP 22 (rs174699) region and the ancestral state was based on consensus sequences for rhesus and orangutan. We were able to genotype this SNP on all our nonhuman primate samples including the chimpanzee samples giving the same allelic state as in rhesus and orangutan consensus suggesting that absence of chimpanzee sequence is a gap in the consensus sequence, not a deletion in the species. For the remainder of the 28 SNPs, the ancestral allele is based on the chimpanzee consensus sequence (Table 1).

Heterozygosities

Average heterozygosities and other population characteristics of the 28 SNPs are given in Table 1. Except for two synonymous coding SNPs (SNP 18, rs740602 and SNP 21, rs769224), all the other SNPs showed an average heterozygosity above 0.24, 13 of these were above 0.4. However, not reflected in heterozygosity is a wide range of variation in frequency of the ancestral allele, both within and among sites across the populations. The largest range for ancestral allele frequency across the populations is 0.027–0.887 and the smallest range is 0.711–1.0. F_{st} values for nine SNPs are above 0.14, the average F_{st} value estimated for 369 SNPs.[36] SNP 22 has the highest F_{st} value, 0.31, and three SNPs (SNP 25, 26 and 27) at the 3′ end also have high F_{st} values (0.22–0.28).

Linkage disequilibrium patterns

The HAPLOT[33] patterns of LD for all the 63 SNPs are complex, varying across the region and among populations (data not shown). The patterns of LD with 28 SNPs are shown in Figure 1. As seen at many other loci,[37,38] the patterns of LD at *COMT* varied among populations studied (Figure 1; somewhat different patterns exist using D' as shown in Supplementary Figure S1). In most populations high levels of LD exist in both the 5′ and the 3′ regions, but not in the coding region. An interesting feature is that a region of high LD ($r^2 > 0.5$) exists from SNP 14 in intron 2 through the non-synonymous SNP 20 in exon 4 in most populations of SW Asian, European and NW Asian ancestries and in populations of Native American ancestry. The other populations do not show such consistently high LD in this region. Except for SNP

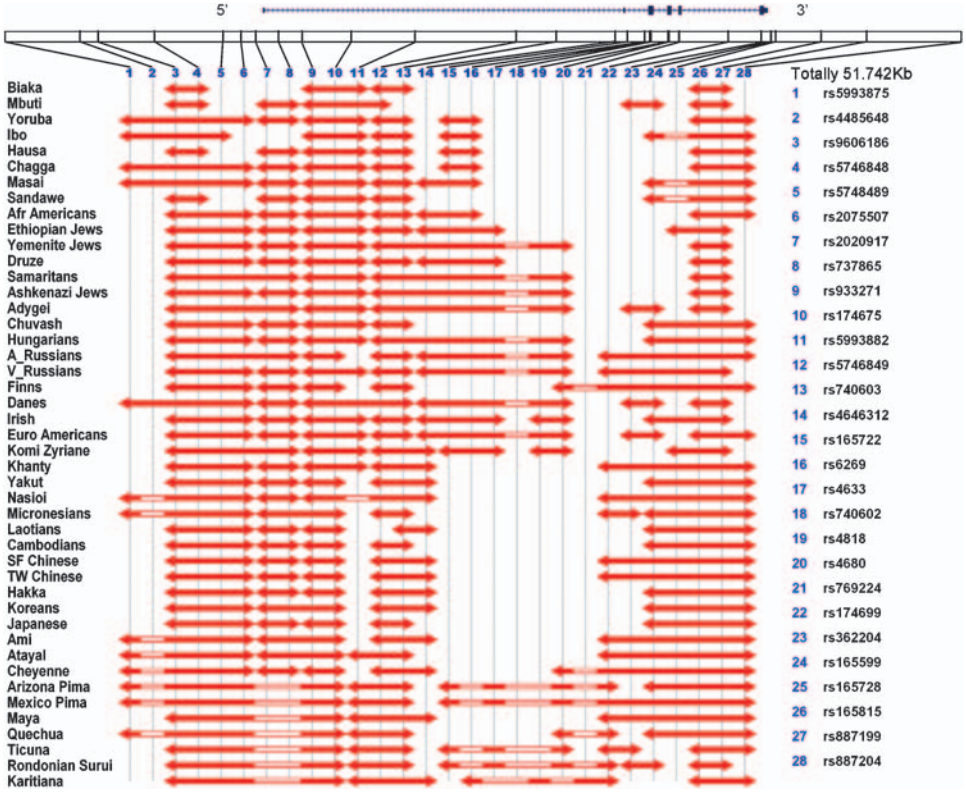

Figure 1 HAPLOT diagram showing regions of high linkage disequilibrium (LD) based on r^2 values with 28 single nucleotide polymorphisms (SNPs) at catechol-*O*-methyltransferase (*COMT*). The pairwise r^2 values and default parameters were used (see text). The red arrows indicate regions of high LD; open parts of arrows indicate noninformative sites.

18, all the other SNPs have relatively high heterozygosities (0.21–0.45; except in Atayal) and therefore, low heterozygosity is not the reason for the observed low LD. In contrast, almost all populations including African populations show strong LD across the much longer region extending from SNP 3 upstream of the gene through SNP 14 near the beginning of exon 2. Another region of high LD in most populations extends from the 3′ UTR through the 3′ end (SNP 24–28). In all of these subregions the intensity of LD varies among specific pairs of SNPs and among populations; the values also vary depending on the statistic used to measure LD. Defining regions of high LD ('LD blocks') is a simplification of the actual pairwise LD values. Supplementary Table S4 illustrates the complexity of pairwise r^2 and D' values for the 14 SNPs across the upstream 'half' of *COMT* for nine specific populations, one from each geographic region.

Haplotype frequencies and heterozygosities

Estimating the haplotypes across the gene using all 28 SNPs shows an extremely complex pattern with many haplotypes, none of which is very frequent (data not shown). Therefore, we have used HAPLOT to identify subregions of adjacent SNPs to analyze and represent graphically. Using both r^2 (Figure 1) and D' (Supplementary Figure S1) the regions that are most consistent among populations for high LD and reduced haplotype diversity are SNPs 3–6 (7.73 kb), 7–11 (8.65 kb), 12–16 (6.95 kb), 17–20 (1.03 kb) and 24–28 (10.28 kb). Haplotype heterozygosities were compared between haplotypes with all 28 SNPs and haplotypes for these shorter regions. A graphical representation of the haplotype heterozygosities is shown in Figure 2.

Haplotype frequencies for the subsets of SNPs are in ALFRED. Figure 3 is a summary of haplotype frequencies for each subregion representing the geographical averages for populations of four major geographical regions (Africa, Europe, East Asia and America).

Haplotypes for the first of these regions (SNPs 3–6, 7.73 kb) include the P2 promoter SNP and three upstream SNPs and show relatively low levels of haplotype diversity (2–5 haplotypes; Figure 3; Supplementary Figure S2a). Two major haplotypes (one completely ancestral, GGCT and the other totally derived, CAAC) constitute more than 80% of the chromosomes in all populations of non-African ancestry. In the populations of African ancestry, two additional common (>5%) haplotypes are required to account for at least 80% of the chromosomes. No simple scheme of haplotype evolution can explain the common extant haplotypes. Ancestral recombination must be involved (recurrent mutations would be far

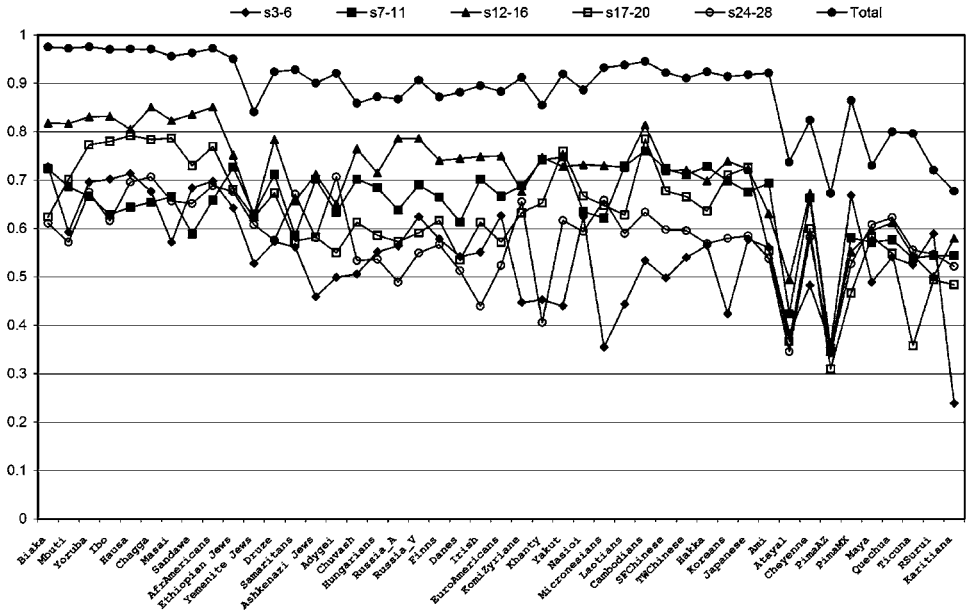

Figure 2 Summary of haplotype heterozygosities for the selected regions of generally simple evolutionary patterns (see text). Haplotype heterozygosities with all 28 single nucleotide polymorphisms (SNPs) (total) have been compared to haplotypes with fewer SNPs (s3–6; s7–11; s12–16; s17–20; s24–28). The numbers indicate the SNPs used to estimate haplotypes based on the order in Table 1.

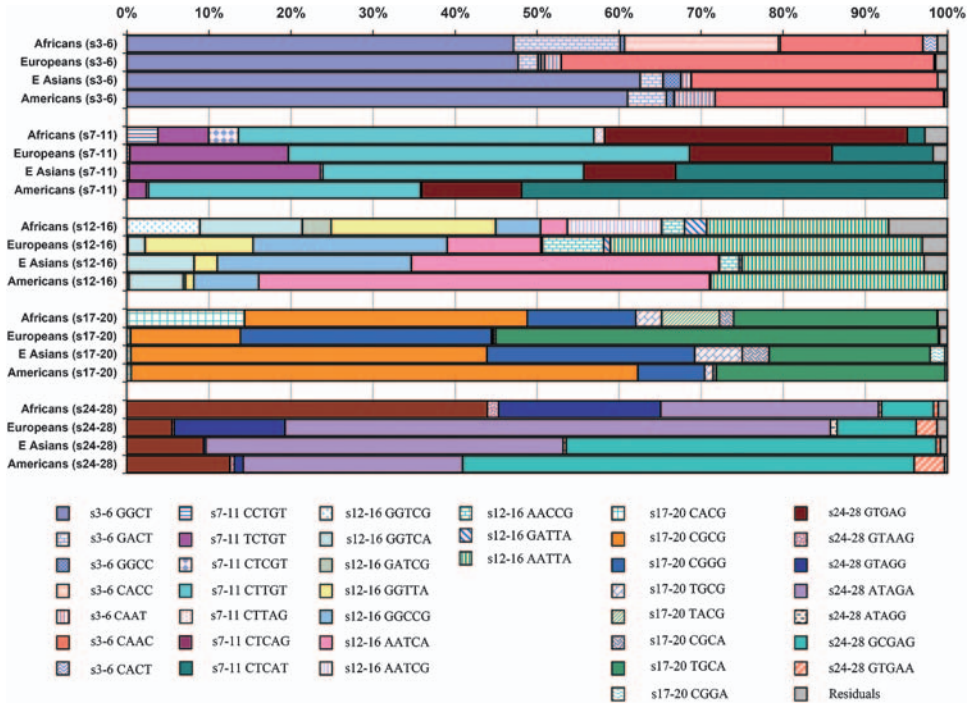

Figure 3 Average haplotype frequencies for four major geographic regions (Africa, Europe, East Asia and America) for each subregion of the catechol-*O*-methyltransferase (*COMT*) gene. The length of each colored bar represents the frequency of the corresponding haplotype. The populations averaged are Africans (Biaka, Mbuti, Yoruba, Ibo, Hausa, Chagga, Masai, Sandawe, African Americans and Ethiopian Jews), Europeans (Ashkenazi Jews, Adygei, Chuvash, Hungarians, Russians—Archangel'sk, Russians—Vologda, Finns, Danes, Irish and European Americans), East Asians (Laotians, Cambodians, San Francisco Chinese, Taiwanese Chinese, Hakka, Japanese, Ami and Atayal) and Americans (Cheyenne, Arizona Pima, Mexican Pima, Maya, Quechua, Ticuna, Rondonian Surui and Karitiana).

rarer events); however, the order of mutation and recombination is ambiguous as most intermediate forms are not seen. The difficulty in determining the evolutionary history of the haplotypes is illustrated by the never common but widely present haplotype GGCC. It has only one nucleotide difference from the

ancestral haplotype but other 'simple' evolutionary intermediates are not seen. Thus, GGCC may be frequently regenerated by recombination between the two most common haplotypes. In any case, although the four common haplotypes in African populations can be explained by accumulation of four mutations (with one intermediate not seen), that evolutionary pathway does not involve GGCC. A separate pathway involving GGCC does not involve the common (in Africa) haplotype GACT. Clearly, existing data do not allow a simple resolution of the evolutionary sequence.

Haplotypes for the next five SNPs (SNPs 7–11, 8.65 kb extending from upstream of the gene through intron 1) also show low levels of haplotype diversity (Figure 3; Supplementary Figure S2b). The ancestral haplotype is found at low frequencies in eight of ten populations of African ancestry and occasionally elsewhere. Four major haplotypes constitute more than 90% of chromosomes in 38 of the 45 populations. Globally seven common haplotypes occur at > 5% in at least one population. In contrast to the immediately upstream region, these seven common haplotypes seem to result from single-step mutation events; no common obviously recombinant haplotype exists. Although LD (as measured by r^2 and D') is low (Figure 1; Supplementary Figure S1) between SNPs 8 and 9, there is no evidence of recombination.

Haplotypes across SNPs 12–16 (6.95 kb, extending from intron 1 through the P1 promoter) show relatively higher levels of haplotype diversity in all the populations of African, SW Asian and European ancestries than in populations of East Asian, Pacific and Native American ancestries. The ancestral haplotype is found in nine of the ten African populations and exists in only two other populations in very low frequencies. Two haplotypes (AATTA and AATCA), both frequent globally, are most likely formed by recombination events between SNPs 13 and 14 (Figure 3; Supplementary Figure S2c). These two haplotypes constitute more than 50% of the chromosomes among the populations of East Asian, Pacific and American ancestries. Figure 1 shows the breakdown of LD between SNPs 13 and 14. A complex repetitive region with length variation is present between these two SNPs,[20] which might be a likely region for frequent recombination.

The coding region common to both the MB—and S—forms of COMT shows low levels of LD except in the Eurasian populations. The four coding region SNPs (SNP17–20, spanning only ~1 kb), form five haplotypes but only three of those are common in the populations of non-African ancestry (Supplementary Figure S2d). The ancestral haplotype (CGGG) is not seen in a few populations but common (20–50%) in most African and Eurasian populations. Two historical recombination events appear to have occurred in this region, one between the SNPs 17 and 18 and the other between SNPs 19 and 20. Haplotypes CACG, TACG, TGCA and CGCA indicate recombination between SNPs 17 and 18, whereas haplotypes TGCG, TGCA, CGCG and CGCA indicate recombination between SNPs 19 and 20. The derived allele 'A' of SNP 18 is always associated with the 'CG' haplotype of SNPs 19 and 20.

Finally, haplotypes with five SNPs in the 3′ region (SNP 24–28, 10.28 kb) of the gene show higher heterozygosities for the populations of African, SW Asian and European ancestries and lower heterozygosities for the populations of East Asian, Pacific and American ancestries. The ancestral haplotype (GTGAG) is not observed in any population (Figure 3; Supplementary Figure S2e). One apparently recombinant haplotype (ATAGA) is globally frequent, possibly reflecting a single historical recombination event in the human lineage between SNPs 27 and 28.

Patterns of obligate recombination

Haplotypes were estimated for adjacent SNP pairs by HAPLO, using SNPs 5–28, omitting individuals when a typing result was missing for one or both sites. Most pairs showed presence of all four haplotypes, indicating obligate ancestral crossovers. However, for several SNP pairs (7 and 8, 8 and 9, 14 and 15, 15 and 16, 18 and 19, 21 and 22, 24 and 25, 25 and 26, 26 and 27) the fourth haplotype is seen in very few populations and at low frequencies (< 5%). The fourth haplotype with sites 5 and 6 is only seen in a few populations and only outside Africa, suggesting that a recombination event might have occurred after humans migrated out of Africa. However, the fourth haplotype may exist in Africa but was not observed in our African samples.

Test for selection (EHH and REHH)

As seen in Supplementary Figures S2a and d, the European populations have particular haplotypes in high frequency both at the P2 promoter and at the coding region. Also, LD levels are higher at both these regions among the Europeans than the other populations. Therefore, it seemed appropriate to test the European populations for presence of extended haplotypes for one or both of these cores.

We used the LRH test[34] for selection with all 63 SNPs in all the populations. The basis of this test is that when any SNP is under positive selection, it increases in frequency more rapidly than expected by random genetic drift. Other SNPs adjacent to the selected SNP also increase in frequency as a hitchhiking effect that would not be expected were the increase to have occurred over many generations. Thus, geographic-specific haplotypes can exist at frequencies higher than elsewhere when selection is geographically specific. On the basis of evidence that the P2 promoter SNP and the exon 4 non-synonymous SNP are probably functional,[12] we used one core haplotype centered on the P2 promoter SNP and another core haplotype centered on the exon 4 non-synonymous SNP.

The core haplotype with the exon 4 non-synonymous SNP did not show an extended haplotype. The core haplotype using the P2 promoter SNP and three

SNPs upstream to it showed extended haplotypes (EHH) up to 40 kb (beyond the coding region of the gene) in several populations but more consistently in the European populations. However, the relative EHH (REHH)[34] value was found to be more than 4 only in three of the European populations (Hungarians, Irish and European Americans). Figures 4a and b show the EHH and REHH profiles for the European populations. Some populations from the other parts of the world showed extended haplotypes but REHH values were low (data not shown).

Discussion

This is the first study of *COMT* with a large number of SNPs on multiple populations from around the world. We have previously shown, using fewer markers and populations, that the frequencies of haplotypes at *COMT* vary widely among populations.[13,20,21] The much greater number of markers now allows a more refined picture of this variation. We report several interesting features across the gene with respect to frequencies of alleles, LD pattern and haplotype heterozygosity.

The derived allele frequencies for 12 out of 28 SNPs are high in at least half of the population samples. We have confirmed that for one SNP (SNP 21), rs769224, the allelic states differ in chimpanzee and gorilla. We infer from haplotype evidence that the gorilla allele is the ancestral allele for humans consistent with the theory that a fraction of the human genome being more similar to gorilla than to chimpanzee.[35]

Considerable complexity in haplotypes and LD patterns exists across the *COMT* gene and those complex patterns vary among populations around the world. However, it is possible to subdivide the region into segments that individually show much simpler patterns involving only a few haplotypes that vary in frequency among populations. Most regions of the gene show the quite standard geographic pattern of decreasing heterozygosity with increasing distance from Africa. Most individual regions show no or little evidence of frequent recombination though some common haplotypes are most parsimoniously explained as having arisen by an early historical crossover subsequently transmitted as a single evolutionary lineage maintaining identity by descent for virtually all extant copies.

There are clear implications for this pattern of evolutionary lineage for subregions but not the gene as a whole: testing markers in one region of the gene will convey little information on evolutionary lineages in other regions of the gene. The upstream and downstream regions of the gene show high LD for all the populations. However, the coding region of the gene shows high levels of LD for the Europeans and Native Americans only but no LD for the East Asians (despite the fact that almost all the SNPs are highly heterozygous) and low LD for the Africans. Although there are two SNPs with documented/probable functional consequences, each separate lineage that has been distinct for most of modern human evolution (and certainly in the 'out-of-Africa' populations) may harbor cryptic functional variation not yet identified.

Because the P2 promoter region shows a reasonable amount of LD in all the populations and this and the exon 4 non-synonymous SNP are probably functional,[12] we have tested for positive selection on this region. Our test for positive selection with the exon 4 non-synonymous SNP did not show any evidence of selection. However, the core haplotype with the P2 promoter SNP showed an extended haplotype in European populations. In spite of a similar pattern of EHH in European populations, the REHH pattern varied considerably among these populations. Most other populations did not show evidence of an extended haplotype. EHH extending to about 100 kb or more is considered a definite signature of positive selection acting on the locus,[34,39] but in this study the EHH that is seen does not extend that far.

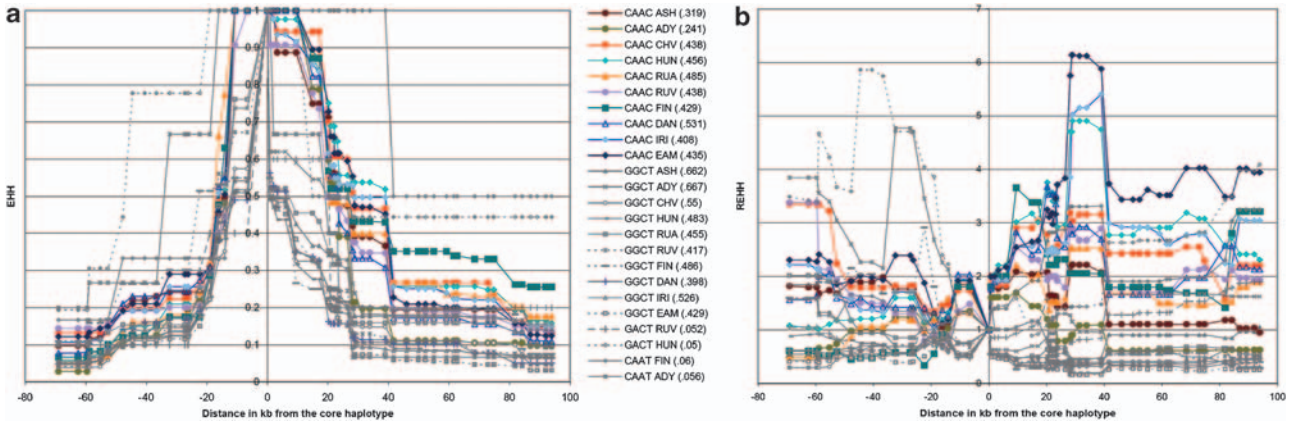

Figure 4 Diagram showing extended haplotype homozygosities (EHH, **a**) and relative extended haplotype homozygosities (REHH, **b**). Core haplotype includes the P2 *Hin*dIII (rs2075507) single nucleotide polymorphism (SNP) and three upstream SNPs. At the fourth site (*Hin*dIII), T, site absent; C, site present. The gray lines either do not show EHH or exist in very low frequencies. EAM, European Americans; ASH, Ashkenazi Jews; CHV, Chuvash; DAN, Danes; FIN, Finns; IRI, Irish; RUV, Vologda Russians; RUA, Archangel'sk Russians; ADY, Adygei; HUN, Hungarians.

Thus, whether the observed EHH in the European populations is the effect of random genetic drift or represents some amount of positive selection acting on this region is not clear. REHH values are neither high nor consistent among populations even in the same geographic region. Parsimony argues that one cannot conclude the evidence is significant for positive selection on the P2 promoter region in Europe or anywhere else.

Haplotypes with different combinations of SNPs across the gene, usually involving the exon 4 non-synonymous SNP Val108/158Met, have been associated with various neuropsychiatric disorders.[1,3,4,5,14–17] It has become clear that both alleles at Val108/158Met exist on chromosomes with each of the two alleles in the P2 promoter. However, there has been no comprehensive analysis of the LD pattern of this region except for initial reports from our laboratory.[20,21] Our data argue that the relationship could be even more complex. As seen clearly in Supplementary Figure S2, one cannot even assume that all European populations are the same. These data now provide a common reference data set of genetic variation and haplotypes across the gene and around the world. These data can be used to determine the degree to which different studies using fewer markers are studying the same or different haplotypic lineages. Such understanding will be essential to resolving the inconsistencies in studies based on different combinations of SNPs in different populations.

Acknowledgments

This work was funded in part by National Institute of Health grant GM057672 (to JRK) and in part by NIH grant AA009379 (to KKK). We acknowledge Dr Sheng Gu for his help with the various computer programs such as HAPLOT and P-Select. We thank Eva Straka and Daniel Votava for their excellent technical help. We also acknowledge and thank the following people who helped assemble the samples from the diverse populations: FL Black, LL Cavalli-Sforza, K Dumars, J Friedlaender, K Kendler, W Knowler, F Oronsaye, J Parnas, L Peltonen, L O Schulz, D Upson, EL Grigorenko, NJ Karoma, JJ Kim, R-B Lu, A Odunsi, F Okonofua, OV Zhukova 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, and the African American samples were obtained from the Coriell Institute for Medical Research. Special thanks are due to the many hundreds of individuals who volunteered to give blood samples. Without such participation of individuals from diverse parts of the world, we would be unable to obtain a true picture of the genetic variation in our species.

Websites used:
ALFRED: http://alfred.med.yale.edu/alfred/index.asp
UCSC: genome browser: http://genome.ucsc.edu/
BLAT: http://genome.ucsc.edu/cgi-bin/hgBlat
dbSNP homepage: http://www.ncbi.nlm.nih.gov/SNP/
BLAST: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi.

References

1 Li T, Ball D, Zhao J, Murray RM, Liu X, Sham PC *et al.* Family-based linkage disequilibrium mapping using SNP marker haplotypes: application to a potential locus for schizophrenia at chromosome 22q11. *Mol Psychiatry* 2000; **5**: 77–84. Erratum in: Mol Psychiatry 2000; 5: 452.
2 Egan MF, Goldberg TE, Kolachana BS, Callicott JH, Mazzanti CM, Straub RE *et al.* Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proc Natl Acad Sci USA* 2001; **9**: 6917–6922.
3 Shifman S, Bronstein M, Sternfeld M, Pisante-Shalom A, Lev-Lehman E, Weizman A *et al.* A highly significant association between a COMT haplotype and schizophrenia. *Am J Hum Genet* 2002; **71**: 1296–1302.
4 Shifman S, Bronstein M, Sternfeld M, Pisanté A, Weizman A, Reznik I *et al.* COMT: a common susceptibility gene in bipolar disorder and schizophrenia. *Am J Med Genet B Neuropsychiatr Genet* 2004; **128**: 61–64.
5 Pooley EC, Fineberg N, Harrison PJ. The met(158) allele of catechol-*O*-methyltransferase (COMT) is associated with obsessive-compulsive disorder in men: case–control study and meta-analysis. *Mol Psychiatry* 2007; **12**: 556–561.
6 Tunbridge EM, Weinberger DR, Harrison PJ. A novel protein isoform of catechol *O*-methyltransferase (COMT): brain expression analysis in schizophrenia and bipolar disorder and effect of Val158Met genotype. *Mol Psychiatry* 2006; **11**: 116–117.
7 Tunbridge EM, Lane TA, Harrison PJ. Expression of multiple catechol-*o*-methyltransferase (COMT) mRNA variants in human brain. *Am J Med Genet B Neuropsychiatr Genet* 2007; **144**: 834–839.
8 Tenhunen J, Salminen M, Lundström K, Kiviluoto T, Savolainen R, Ulmanen I. Genomic organization of the human catechol *O*-methyltransferase gene and its expression from two distinct promoters. *Eur J Biochem* 1994; **223**: 1049–1059.
9 Weinshilboum RM, Raymond FA. Inheritance of low erythrocyte catechol-*o*-methyltransferase activity in man. *Am J Hum Genet* 1977; **29**: 125–135.
10 Lachman HM, Papolos DF, Saito T, Yu YM, Szumlanski CL, Weinshilboum RM. Human catechol-*O*-methyltransferase pharmacogenetics: description of a functional polymorphism and its potential application to neuropsychiatric disorders. *Pharmacogenetics* 1996; **6**: 243–250.
11 Lotta T, Vidgren J, Tilgmann C, Ulmanen I, Melén K, Julkunen I *et al.* Kinetics of human soluble and membrane-bound catechol *O*-methyltransferase: a revised mechanism and description of the thermolabile variant of the enzyme. *Biochemistry* 1995; **34**: 4202–4210.
12 Chen J, Lipska BK, Halim N, Ma QD, Matsumoto M, Melhem S *et al.* Functional analysis of genetic variation in catechol-*O*-methyltransferase (COMT): effects on mRNA, protein, and enzyme activity in postmortem human brain. *Am J Hum Genet* 2004; **75**: 807–821. Erratum in: Am J Hum Genet. 2005; 76: 1089.
13 Palmatier MA, Kang AM, Kidd KK. Global variation in the frequencies of functionally different catechol-*O*-methyltransferase alleles. *Biol Psychiatry* 1999; **46**: 557–567.
14 Lang UE, Bajbouj M, Sander T, Gallinat J. Gender-dependent association of the functional catechol-*O*-methyltransferase Val158-Met genotype with sensation seeking personality trait. *Neuropsychopharmacology* 2007; **32**: 1950–1955.
15 Domschke K, Freitag CM, Kuhlenbaumer G, Schirmacher A, Sand P, Nyhuis P *et al.* Association of the functional V158M catechol-*O*-methyl-transferase polymorphism with panic disorder in women. *Int J Neuropsychopharmacol* 2004; **7**: 183–188.
16 Beuten J, Payne TJ, Ma JZ, Li MD. Significant association of catechol-*O*-methyltransferase (COMT) haplotypes with nicotine dependence in male and female smokers of two ethnic populations. *Neuropsychopharmacology* 2006; **31**: 675–684.

17 Funke B, Malhotra AK, Finn CT, Plocik AM, Lake SL, Lencz T *et al.* COMT genetic variation confers risk for psychotic and affective disorders: a case control study. *Behav Brain Funct* 2005; **1**: 19.
18 Williams HJ, Glaser B, Williams NM, Norton N, Zammit S, Macgregor S *et al.* No association between schizophrenia and polymorphisms in COMT in two large samples. *Am J Psychiatry* 2005; **162**: 1736–1738.
19 Fan JB, Zhang CS, Gu NF, Li XW, Sun WW, Wang HY *et al.* Catechol-*O*-methyltransferase gene val/met functional polymorphism and risk of schizophrenia: a large-scale association study plus meta-analysis. *Biol Psychiatry* 2005; **57**: 139–144.
20 Palmatier MA, Pakstis AJ, Speed W, Paschou P, Goldman D, Odunsi A *et al.* COMT haplotypes suggest P2 promoter region relevance for schizophrenia. *Mol Psychiatry* 2004; **9**: 859–870.
21 DeMille MM, Kidd JR, Ruggeri V, Palmatier MA, Goldman D, Odunsi A *et al.* Population variation in linkage disequilibrium across the *COMT* gene considering promoter region and coding region variation. *Hum Genet* 2002; **111**: 521–537.
22 Diatchenko L, Slade GD, Nackley AG, Bhalang K, Sigurdsson A, Belfer I *et al.* Genetic basis for individual variations in pain perception and the development of a chronic pain condition. *Hum Mol Genet* 2005; **14**: 135–143.
23 Hu Z, Song CG, Lu JS, Luo JM, Shen ZZ, Huang W *et al.* A multigenic study on breast cancer risk associated with genetic polymorphisms of ER Alpha, COMT and CYP19 gene in BRCA1/BRCA2 negative Shanghai women with early onset breast cancer or affected relatives. *J Cancer Res Clin Oncol* 2007; **133**: 969–978.
24 Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A *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.
25 Cubells JF, Kobayashi K, Nagatsu T, Kidd KK, Kidd JR, Calafell F *et al.* Population genetics of a functional variant of the dopamine beta-hydroxylase gene (*DBH*). *Am J Med Genet* 1997; **74**: 374–379.
26 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.
27 Zhang S, Pakstis AJ, Kidd KK, Zhao H. Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. *Am J Hum Genet* 2001; **69**: 906–912.
28 Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**: 978–989.
29 Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* 2005; **76**: 449–462.
30 Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 1995; **29**: 311–322.
31 Lewontin RC. The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics* 1964; **49**: 49–67.
32 Zhao H, Pakstis AJ, Kidd JR, Kidd KK. Assessing linkage disequilibrium in a complex genetic system. I. Overall deviation from random association. *Ann Hum Genet* 1999; **63**: 167–179.
33 Gu S, Pakstis AJ, Kidd KK. HAPLOT: a graphical comparison of haplotype blocks, tagSNP sets and SNP variation for multiple populations. *Bioinformatics* 2005; **21**: 3938–3939.
34 Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002; **419**: 832–837.
35 Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 2006; **441**: 1103–1108.
36 Kidd KK, Pakstis AJ, Speed WC, Kidd JR. Understanding human DNA sequence variation. *J Hered* 2004; **95**: 406–420.
37 Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, Brookes AJ *et al.* Linkage disequilibrium patterns vary substantially among populations. *Eur J Hum Genet* 2005; **13**: 677–686.
38 Gu S, Pakstis AJ, Li H, Speed WC, Kidd JR, Kidd KK. Significant variation in haplotype block structure but conservation in tagSNP patterns among global populations. *Eur J Hum Genet* 2007; **15**: 302–312. Erratum in: Eur J Hum Genet. 2007; 15: 818.
39 Han Y, Gu S, Oota H, Osier MV, Pakstis AJ, Speed WC *et al.* Evidence of positive selection on a class I ADH locus. *Am J Hum Genet* 2007; **80**: 441–456.

Supplementary Information accompanies the paper on the Molecular Psychiatry website (http://www.nature.com/mp)