European Journal of Human Genetics (2007) 15, 1071–1078

www.nature.com/ejhg

ARTICLE

Enrichment of HapMap recombination hotspot predictions around human nervous system genes: evidence for positive selection ?

Jan Freudenberg*[,1], Ying-Hui Fu[1] and Louis J Ptáček[1,2]

[1]*Department of Neurology, Institute of Human Genetics, University of California San Francisco, San Francisco, CA, USA;* [2]*Howard Hughes Medical Institute, University of California San Francisco, San Francisco, CA, USA*

Channels and developmental genes belong to the molecular key players in the human central nervous system (CNS). Mutations in these genes often cause monogenic neurological disease and interspecies comparisons had shown reduced divergence. On the other hand, accelerated evolution of genes with roles in neurotransmission and development had indicated widespread positive selection in hominids. In the present study, we hypothesized that recombination hotspots could be enriched at genes with particularly important role in the CNS, because at those loci beneficial mutations may occur on a highly constrained background and consequently increased recombination could promote their fixation. To test this hypothesis, we retrieved CNS genes based on keyword search, expression data and expert knowledge. Consistent with our hypothesis, we find an enrichment of hotspot predictions around genes that are retrieved by all three strategies. Moreover, when comparing human genes based on their Gene Ontology annotations, we find hotspot predictions preferentially located around channels and neurodevelopmental genes. Taken together with the distinct sequence evolution that was reported by comparative genomic studies, this finding indicates continued positive selection at many CNS gene loci. In support of this interpretation, we also find an enrichment of recombination hotspot predictions around conserved noncoding regions that were reported to display a signature of accelerated evolution in the human lineage. Widespread positive selection acting on CNS gene loci could relate to the high prevalence of human nervous system disorders with genetically complex inheritance, potentially under an ancestral susceptibility allele model.
European Journal of Human Genetics (2007) **15,** 1071–1078; doi:10.1038/sj.ejhg.5201876; published online 13 June 2007

Keywords: CNS gene function; recombination hotspot; sequence evolution; complex CNS disorder

Introduction

Disorders of the central nervous system (CNS) constitute a major global health burden. Accordingly, not only the role of genetic variation at candidate gene loci in CNS disorders, but also molecular function and evolution of CNS genes has been of interest in many studies. One important result from the comparative genomic analysis of CNS gene loci is their higher level of interspecies conservation.[1–3] This indicates a higher potential for deleterious mutations and is consistent with gene function in an organ system as complex and indispensable as the CNS. On the other hand, coding and noncoding regions of certain types of CNS genes, in particular channel and developmental genes, had shown evidence for accelerated evolution in hominids as compared to murids.[3–6] This may be attributed to positively selected adaptations in genes with important role in the human CNS. Because the

*Correspondence: Dr J Freudenberg, Department of Neurology, Institute of Human Genetics, University of California San Francisco, 1550 4th Street, San Francisco, CA 94143-2922, USA.
Tel: +1 415 514 9310; Fax: +1 415 502 5641;
E-mail: jan.freudenberg@ucsf.edu
Received 13 November 2006; revised 13 March 2007; accepted 15 May 2007; published online 13 June 2007

realized strength of positive selection on a constrained background is increased when recombination is present,[7] more recombination near genes with important role in CNS function and development could thus be advantageous. If more recombination would be detectable at CNS genes, it would thus provide evidence for ongoing selection in the human lineage, based on the assumption that local recombination rates evolve under the influence of natural selection. Ongoing positive selection at CNS gene loci might be related to the high prevalence of genetically complex CNS disorders, potentially under an ancestral susceptibility allele model.[8]

Consistent with a functionally biased location of meiotic recombination events, recombination rates and linkage disequilibrium (LD) are known to vary greatly across the human genome.[9–12] The recently published HapMap dataset allowed for the first time the identification of functional categories of genes that are preferentially located within regions of weak LD (immune response and sensory perception) and strong LD (DNA and RNA metabolism and cell cycle).[13,14] In addition, several functional categories related to neurotransmission showed reduced LD.[13] These LD patterns result from the joint influence of drift and different modes of selection on multiple sites under a complex population history. Nevertheless, the biased LD patterns gave rise to the hypothesis that it could be advantageous for certain types of genes to be located in regions of increased recombination.[13] This makes sense theoretically, because LD can occur between multiple variable sites under selection. By allowing mutations to evolve independently from their haplotype background, the increase of recombination rates between such sites can cause selection to act more efficiently.[15] On the basis of recent genome-wide estimates of fine-scale recombination rates,[16,17] it is now possible to take a closer look at the relationship between recombination rate and gene function. These recombination rate estimates are based on the polymorphism data that specifically reflect human evolutionary history over the past 200 000 years.[14] One result of the analysis of fine-scale recombination rates is the widespread existence of recombination hotspots.[16,17] Recombination hotspots appear to evolve in relatively short periods, because little conservation of hotspots between human and chimpanzee was observed.[18,19]

In the present study, we hypothesized that recombination hotspot predictions might be enriched at CNS gene loci. By allowing beneficial mutations to evolve independently from a highly constrained haplotype background, recombination hotspots could be particularly advantageous at CNS gene loci.

Materials and methods

Recombination hotspot predictions were retrieved from the UCSC Genome Database (version hg17),[20] based on HapMap phase I genotype data.[14] These hotspots are predicted by the LDhot method,[17] that infers population recombination rates from patterns of LD. The method is sufficiently fast to be applicable to genome-wide data sets. In an evaluation on experimentally verified hotspots, it predicted correctly four out of eight hotspots with no false positives.[21] Autosomal gene model annotations were retrieved from the Ensembl Core database (version 36).[22]

To test the enrichment of hotspots around genes from a certain category, we calculated the odd that a gene from this category displays a recombination hotspot in its flanking regions *versus* the odd that a gene not belonging to this category displays a recombination hotspot in its flanking regions. This odds ratio (OR) was separately calculated for upstream and downstream flanking regions (30 kb upstream or downstream, respectively) and then averaged to obtain a summary score for the respective category. We excluded genes that do not have at least five polymorphic HapMap Phase I SNPs in their 30 kp up- and downstream regions. To measure the significance of the association between a category and recombination hotspots, we employed a permutation-based strategy. We randomly permuted gene identifiers across gene model annotations in the human genome, while leaving the functional annotations that are assigned to gene identifiers unchanged and also leaving the position of hotspots unchanged. *P*-values for all categories were initially determined by 1000 random permutations. For categories where maximally 1% of random permutations showed a more extreme hotspot association than the actually observed score (ie higher average OR), the number of permutations was increased by a factor of 10. For the most significant categories, 100 000 random permutations were performed. Around genes from the most significant categories ($P \leq 10^{-5}$), the actually observed enrichment of hotspots was stronger than seen in any of the 100 000 random permutations. This permutation strategy accounts for the location of genes and hotspots in the human genome. However, in combination with unknown patterns of chromosomal clustering of functionally related genes, it could be biased by the overproportional sampling of regions that flank two neighboring genes from a same category. Thus, categories that show a higher degree of chromosomal clustering may attain anticonservative *P*-values, if they have overproportionally many hotspots in chromosomal clusters, whereas such categories may attain conservative *P*-values, if they have a relative depletion of hotspots in chromosomal clusters.

To be robust against the influence of unknown patterns of chromosomal clustering on the reported results, we additionally employed a shifting window strategy to test the association between hotspots and gene functions. This shifting window strategy moves a nonoverlapping 50 kb window across the genome. Genomic windows were required to contain at least five polymorphic SNPs and

overlap at least one gene to be included. For each category, we retrieved all windows that overlap with at least one gene from the respective category. Then we calculated for each category, the relative frequency of windows that overlap at least one hotspot prediction as test statistic. Again we used random permutation of gene identifiers across gene model annotations to determine the significance of the test statistic. Because hotspot predictions are known to be rare within genes,[16] this shifting window strategy may have a bias against categories that are enriched for large-sized genes since they have more intragenic windows.

CNS genes termed 'neuronal genes' were defined as human genes that were returned by the EntrezGene server for the query '(neuronal* or glial* or neural* or neurite or axon) and not olfactory' (date of query: 3 November 2006). EntrezGene provides manually curated free text annotation of genes. Gene expression data were retrieved from GNF2 dataset,[23] as represented in the UCSC Genome Annotation database[20] (version hg 17). Cross-references to Ensembl genes were obtained from the UCSC database. We defined CNS expression as the expression in the tissues olfactory bulb, temporal cortex, parietal cortex, prefrontal cortex, occipital cortex, cingulate cortex, cerebellum, cerebellar pedunculus, amygdala, hypothalamus, thalamus, subthalamic nuclei, caudate nucleus, globus pallidus, pons, medulla and spinal cord. We excluded fetal and neoplastic tissues from the analysis. Multiple measurements of a gene in a tissue were averaged and genes were taken as expressed in a given tissue, if the signal exceeded 200 arbitrary units, as suggested in the original publication.[23]

Ensembl genes[22] were cross-referenced via the International Protein Index (version April 2006)[24] to the Gene Ontology (GO) Annotation database (version April 2006).[25] Obeying the 'true path rule', GO annotations were transitively assigned for all parent categories. GO categories[26] were analyzed, if annotated to 20 or more human Ensembl genes. To account for multiple testing, *Q*-values were estimated based on the permutation-based *P*-value distribution over the 1266 GO categories[27] showing that the most significant *P*-values of $\leq 10^{-5}$ correspond to a *Q*-value of $\leq 10^{-4}$, whereas a false discovery rate of 5% is attained when calling GO categories with $P \leq 0.002$ significant. To be included in the analysis, genes were required to have at least one GO annotation.[26]

Results

To test the association of recombination hotspots predictions[14,17] with genes from a certain category, we employed a permutation based analysis strategy (see Materials and Methods for details). We initially focused the analysis on hotspots that are predicted in the 30 kb up- and downstream flanking regions of genes, because it was shown that hotspots most often exist in these genomic windows.[16] To this end, we first calculated as test statistic the ratio between the odd of a hotspot prediction in up- or downstream regions a gene annotated by the respective category and the odd of a hotspot prediction around a gene not annotated by the respective category. In addition, we employed a shifting window strategy by moving a 50 kb window across the genome. For each category, we retrieved all windows that overlap at least one gene from the respective category and calculated as test statistic the relative frequency of windows that overlap at least one hotspot prediction.

Recombination hotspots are enriched around CNS genes

Notably, we found hotspots enriched around genes with particularly important role in the CNS when retrieving such genes by three different strategies (Table 1). As one strategy, we followed the suggestion to identify CNS genes by keyword search in text-based gene annotations.[5] We retrieved 'neuronal genes' as those that are returned from

Table 1 Enrichment of recombination hotspots around CNS genes that are retrieved by different definitions

CNS gene definition	*# with 5′ hs*	*# w/o 5′ hs*	*5′ OR*	*# 3′ hs*	*# w/o 3′ hs*	*3′ OR*	*3′–5′ OR*	*3′–5′ P-value*	*#50 kb windows*	*% with hs*	*Window P-value*
Neuronal genes	286	742	1.33	289	739	1.22	1.28	0.00001	3166	40.1	0.00001
Max CNS expressed	397	1096	1.26	409	1084	1.19	1.22	0.00001	4680	39.6	0.00001
Neuronal ⋂ Max CNS	70	146	1.61	63	153	1.26	1.44	0.00001	1162	43.0	0.00001
AD candidate genes	59	176	1.16	68	167	1.29	1.21	0.029	758	41.2	0.0014

Neuronal genes are defined by keyword search of text-based EntrezGene annotations and compared to other GO-annotated genes. Max CNS expressed genes are defined by their maximal expression in a CNS tissue based on the GNF2 dataset and compared to other genes with expression annotations. 'AD candidate' genes are defined by their independently suggested status as candidate for affective disorder. The columns of the table report the number of genes the category is annotated to and that display at least one hotspot in the 30 kb upstream (# 5 'hs) and downstream (# 3 'hs) region or that do not display any hotspot in the upstream (# w/o 5 'hs) and downstream (# w/o 3′ hs) region. The upstream odds ratio (5′ OR), the downstream odds ratio (3′ OR) and the average OR measure the enrichment of hotspots around genes from the respective category. In addition, the number of shifting windows (#50kb windows) is shown that overlap at least one gene from the respective category and the percentage of such window that overlap at least on HapMap hotspot prediction (% with hs). *P*-Values were calculated by permutation analysis.

the EntrezGene server by searching for the keywords 'neuronal', 'glial', 'neural', 'neurit' or 'axon'. On the basis of this definition, we identified 1026 neuronal genes that conform to our inclusion criteria, that is having at least one GO annotation and at least five HapMap SNPs in flanking regions. Consistent with the hypothesis of an enrichment of hotspots around CNS genes, we found flanking regions of neuronal genes enriched for recombination hotspots (upstream OR: 1.33, downstream OR: 1.22, average OR: 1.28; $P<$10E-05). We further observed a somewhat larger effect size for the subset of 145 neuronal genes that are additionally GO annotated by 'cell adhesion' (upstream OR: 1.56, downstream OR: 1.25, average OR: 1.4; $P=$ 0.0057). This may be interesting, because noncoding accelerated hominid evolution was found associated with 'cell adhesion' genes,[5] what had mainly relied on a subset of 'cell adhesion' genes that belonged to a larger group of 'neuronal' genes as defined by keyword search.

We next used the human GNF2 dataset[23] to identify CNS genes. Here we defined CNS genes as those with maximal expression in a CNS tissue. In total, we retrieved 14 600 human autosomal Ensembl genes with available expression and SNP annotation for their flanking regions and found 1493 of them having their maximal expression in a CNS tissue. On the basis of their maximal expression in a CNS tissue, one may predict that these genes play important roles in nervous system function. This definition of CNS genes also showed an enrichment of hotspot predictions (OR: 1.22, $P\leq10^{-5}$). A subset of 216 of the 1493 maximal CNS expressed genes was also contained in the above set of 1026 genes that were retrieved by keyword search. This intersected set of 216 genes produced an even stronger association with recombination hotspots (OR: 1.44, $P\leq10^{-5}$).

As a third definition of CNS genes, we applied independent expert knowledge by using a manually defined list of candidate genes for affective disorder,[28] 235 of them fulfilling our inclusion criteria. These CNS candidate genes displayed a significant association with hotspots with similar effect size (OR: 1.21, $P=0.03$). Thus, hotspots appear enriched around genes with important roles in the human CNS when such genes are defined by text-based annotations, expression data or independent expert knowledge.

Recombination hotspots are enriched around genes with CNS-related GO Annotations

To investigate the relationship between hotspot location and gene function more systematically, we tested all GO categories for the enrichment of recombination hotspots. Altogether, we compared 1266 GO categories based on 15 354 autosomal genes that have at least one GO annotation. The existence of functional categories that are enriched for hotspots and categories that are depleted of hotspots is indicated by the distribution of *P*-values over all GO categories, showing an excess of *P*-values close to 0 or close to 1 (Figure 1).

Consistent with our hypothesis, GO categories most significantly enriched for recombination hotspots comprise several categories annotated to channel activity genes (Table 2 and Supplementary Table 1). Among a total of 359 genes annotated with 'alpha-type channel activity', we found 104 with at least one hotspot in their upstream region and 119 with at least one hotspot in their downstream region (OR: 1.49, $P\leq10^{-5}$). A similarly significant enrichment of hotspots ($P\leq10^{-5}$) was found around genes annotated with '(voltage-gated) ion channel activity', '(metal) ion transport' and '(channel or pore class) transporter activity'. By maintaining delicate ion balances at cellular membranes, the genes annotated by these categories are directly responsible for the ability of nervous system cells to generate signals. Interestingly, a significant association with hotspots ($P\leq10^{-5}$) also exists for the more general GO categories 'membrane' and (integral/intrinsic to) plasma membrane. Thus, the enrichment of hotspots around channels appears to be the more extreme reflection of a general pattern that applies to genes encoding membrane proteins.

Because more general GO categories are annotated to a higher number of genes, they have more power to attain a significant association. Nevertheless, owing to the above association of 'ion channel activity' with recombination hotspots, the respective subcategories deserve attention too (Supplementary Table 1). Among these, the most

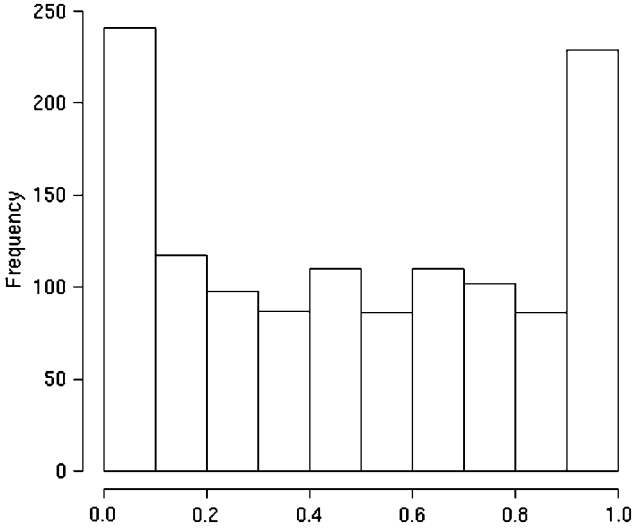

Figure 1 Frequency histogram of OR-based *P*-values that are estimated for the 1266 tested GO categories. The excess of *P*-values that are close to 0 and *P*-values that are close to 1 indicates the existence of GO categories that are enriched for hotspots and the existence of GO categories that are depleted of hotspots. We observe 30 GO categories (Table 1) with *P*-values $\leq$0.0005 as compared to one GO category that would be expected by chance alone under the expectation of a uniform distribution.

Table 2 List of 30 Gene Ontology (GO) categories that are most significantly associated with predictions of recombination hotspots, see legend of Table 1 for further description of columns

GO key	*GO name*	*# with 5′ hs*	*# w/o 5′ hs*	*5′ OR*	*# with 3′ hs*	*# w/o 3′hs*	*3′ OR*	*3′–5′ OR*	*3′–5′ P-value*	*# 50 kb window*	*% with hs*	*Window P-value*
GO:0005244	Voltage-gated ion channel activity	51	114	1.54	61	104	1.86	1.70	0.00001	480	0.456	0.00001
GO:0005261	Cation channel activity	75	171	1.52	85	161	1.67	1.59	0.00001	735	0.457	0.00001
GO:0005216	Ion channel activity	95	234	1.40	112	217	1.64	1.52	0.00001	985	0.429	0.00001
GO:0015268	Alpha-Type channel activity	104	255	1.41	119	240	1.58	1.49	0.00001	1016	0.432	0.00001
GO:0030154	Cell differentiation	129	311	1.44	132	308	1.36	1.40	0.00001	1097	0.435	0.00001
GO:0007275	Development	443	1135	1.39	458	1120	1.33	1.36	0.00001	3655	0.411	0.00001
GO:0005576	Extracellular region	283	707	1.41	269	721	1.19	1.29	0.00001	1833	0.453	0.00001
GO:0006811	Ion transport	177	505	1.21	202	480	1.34	1.28	0.00001	1769	0.414	0.00001
GO:0005887	Integral to plasma membrane	294	742	1.39	274	762	1.14	1.27	0.00001	2476	0.438	0.00001
GO:0031226	Intrinsic to plasma membrane	295	751	1.38	277	769	1.14	1.26	0.00001	2511	0.437	0.00001
GO:0005886	Plasma membrane	397	1086	1.29	400	1083	1.18	1.23	0.00001	3598	0.416	0.00001
GO:0005215	Transporter activity	312	881	1.24	330	863	1.22	1.23	0.00001	2689	0.401	0.00001
GO:0015267	Channel or pore class transporter activity	105	268	1.36	121	252	1.53	1.44	0.00002	1039	0.428	0.00001
GO:0048513	Organ development	136	317	1.49	130	323	1.27	1.38	0.00002	1026	0.448	0.00001
GO:0030001	Metal ion transport	97	256	1.31	111	242	1.46	1.38	0.00008	987	0.447	0.00001
GO:0007155	Cell adhesion	158	379	1.45	143	394	1.15	1.30	0.00011	1734	0.434	0.00001
GO:0048731	System development	118	270	1.52	106	282	1.19	1.36	0.00014	1195	0.423	0.00001
GO:0016020	Membrane	1132	3598	1.12	1209	3521	1.12	1.12	0.00017	9815	0.388	0.00001
GO:0051179	Localization	637	1969	1.13	680	1926	1.13	1.13	0.00022	5698	0.371	0.00001
GO:0007399	Nervous system development	116	270	1.49	106	280	1.20	1.34	0.00024	1193	0.422	0.00001
GO:0005267	Potassium channel activity	40	92	1.50	46	86	1.69	1.59	0.00030	405	0.447	0.00001
GO:0005615	Extracellular space	121	278	1.51	106	293	1.14	1.33	0.00031	631	0.458	0.00001
GO:0006813	Potassium ion transport	48	109	1.52	52	105	1.56	1.54	0.00035	486	0.449	0.00001
GO:0015075	Ion transporter activity	162	469	1.19	185	446	1.32	1.26	0.00036	1584	0.399	0.00001
GO:0043565	Sequence-specific DNA binding	129	305	1.47	129	305	1.34	1.40	0.00001	799	0.406	0.00037
GO:0006812	Cation transport	115	354	1.12	148	321	1.47	1.293	0.00038	1237	0.418	0.00001
GO:0051234	Establishment of localization	632	1961	1.13	677	1916	1.14	1.132	0.00027	5669	0.372	0.00002
GO:0016021	Integral to membrane	883	2766	1.13	928	2721	1.10	1.111	0.00057	7309	0.398	0.00001
GO:0031224	Intrinsic to membrane	884	2775	1.12	931	2728	1.10	1.110	0.00060	7334	0.398	0.00001
GO:0008076	Voltage-gated potassium channel complex	27	55	1.69	33	49	2.13	1.91	0.00012	213	0.479	0.00008

The complete results for all tested 1266 GO categories are given in Supplementary Table 1.

significant GO categories comprise 'voltage-gated potassium channel complex/activity'. Owing to the assembly of functional potassium channels from multiple independent monomers, these categories are still annotated to a relatively high number of genes. An even stronger effect size but weaker significance can be observed for the less densely populated GO categories 'calcium channel activity', 'sodium channel activity' and 'chloride channel activity'. Thus, a contribution to the association of channels with recombination hotspots comes from channels with selective permeability for all major ions that determine the cellular membrane potential.

The second major group of GO categories that shows association with recombination hotspots is constituted by annotations to developmental and particularly neurodevelopmental genes (Table 2 and Supplementary Table 1). Among the 1578 genes that are annotated with a role in 'development', 443 show at least one hotspot in their upstream region and 458 show at least one hotspot in their downstream region (OR: 1.36, $P \leq 10^{-5}$). Moreover, significant associations ($P \leq 10^{-4}$) were found for the GO categories 'cell differentiation' and 'organ/system development'. Thus, in addition to channels which are important in neuronal signal processing, developmental genes are associated with recombination hotspots. Among more specific developmental categories, the strongest signals were found for 'nervous system development' and 'muscle development'. Thus, hotspots are particularly enriched around genes that play a role in the development of excitable tissues. Further support for an enrichment of hotspots around neurodevelopmental genes is provided by the results obtained for the GO category 'cell adhesion' (OR: 1.30 $P = 1.1 \times 10^{-4}$). This GO category is annotated to many genes with a role in neurodevelopment and communication between neurons such as neurologins, contactins and cadherins.

Hotspot predictions are enriched around genes with distinct coding sequence evolution

On the basis of detailed comparative genomic analysis of genes from different GO categories that has been published,[3] we now looked for evidence of distinct sequence

evolution of the top 30 GO categories that are associated with recombination hotspot predictions. All the 14 GO categories that are annotated to channels displayed accelerated protein sequence evolution in hominids as compared to murids.[3] The same observation holds true for the three GO categories 'development' (GO:0007275) 'nervous system development' (GO:0007399) and 'cell adhesion' (GO:0007155). The more general and heterogeneous GO categories '(integral to) plasma membrane' and 'extracellular region/space' were not reported to display accelerated protein sequence evolution in hominids.[3] However, in the human–chimp lineage, the former two GO categories '(integral to) plasma membrane' showed evidence for slower coding sequence evolution, whereas the latter two GO categories 'extracellular region/space' showed evidence for faster coding sequence evolution.[3] Thus the comparative genomic analysis of all top hotspot-associated GO categories (Table 2) had either provided evidence for increased positive selection, increased negative selection or both.

One may also apply less stringent significance thresholds to look for a relation between coding sequence evolution and hotspot enrichment among functional gene categories. In total, 477 GO categories were analyzed both in our analysis and the published comparative analysis of human, chimp, rat and mouse protein sequence evolution.[3] Among these, 226 categories displayed slower coding sequence evolution in the human–chimp lineage (defined as p_low smaller than 0.05 in Supplementary Table S26 of International_Chimp_Genome_Consortium[3]), 46 of them were enriched for hotspot prediction (defined as 3′–5′ *P*-value smaller than 0.05 in Supplementary Table 1). This falls close to the random expectation of 47 categories, based on the simplifying assumption of independence. On the other hand, 95 categories showed faster coding sequence evolution in the human–chimp lineage (defined as p_high smaller than 0.05 in Supplementary Table S26 of International_Chimp_Genome_Consortium[3]) and we see 26 of them enriched for hotspot predictions. This number is somewhat larger than the 17 categories that would be randomly expected under the assumption of independence. Of particular interest are now those categories with accelerated protein sequence evolution in hominids as compared to murids. Among the 98 categories that were reported to display accelerated hominid protein sequence evolution as compared to murids (defined as P_hominid_Ka > murid_Ka smaller than 0.05 in Supplementary Table S30 of International_Chimp_Genome_Consortium[3]), we find 34 enriched for hotspot predictions. This is nearly twice the number of 18 categories that would be expected by chance under independence.

To test the relationship between hotspots and prior evidence for accelerated noncoding sequence evolution more directly, we went back to look at the above genes that are contained in the sets of neuronal genes, maximal CNS expressed genes and affective disorder candidates. Among the 2383 genes altogether, we found 165 to be located closest to a human accelerated noncoding region.[5] These 165 genes show significant enrichment of hotspots as compared to the remaining 2218 CNS genes (upstream: OR: 1.46, $P=0.012$; downstream: OR:1.53, $P=0.0053$). Finally, we directly analyzed the published list of regions with evidence for accelerated evolution.[5] For each region with accelerated evolution in either human or mouse, we calculated the interval that is centered on the respective region and extends for 25 kb into both directions. We then excluded intervals around mouse-accelerated regions that overlap intervals around human-accelerated regions. In further support of our hypothesis, we found a significant enrichment of hotspots within intervals around human-accelerated regions (OR: 1.3, $P=0.00015$), comparing the 992 intervals centered on human-accelerated region with the 4207 intervals centered mouse-accelerated regions.

Discussion

We report the preferential location of recombination hotspot predictions around human genes with important role in the CNS, including channels, developmental and neuronal cell adhesion genes. Strikingly, such CNS genes were also found to display the combination of strong sequence conservation[1–3,29,30] and accelerated sequence evolution in hominids.[3–6] Accelerated sequence evolution in hominids was interpreted as more prevalent positive selection or alternatively relaxed selective constraint in hominids.[3] Our results support more prevalent positive selection as a cause of accelerated evolution, based on the assumption that recombination hotspots can evolve to increase the efficacy of positive selection on a constrained background.[7] Thus, the enrichment of hotspots at CNS gene loci might indicate widespread continued positive selection in the human lineage.

Both the influence of negative and positive selection intensity on local recombination modifiers is consistent with theoretical models.[31,32] However, an exclusive influence of negative selection would not explain the strongest enrichment of hotspots at loci with accelerated hominid evolution. Acknowledging the large body of previous work on recombination promoting selection,[33,34] we therefore explain our observation by an influence of both positive *and* negative selection on local recombination rate. In this context it is most important that earlier results had indicated that the realized strength of selection is increased when recombination is present.[7] Therefore, hotspots may often have evolved at CNS genes to ease positive selection of beneficial mutations that require particularly strong sequence conservation in their neighboring sites. In that case, our results indirectly support the claim of widespread positive selection on genes with important function in the human nervous system.[4,5] If reasoning into the opposite

direction, our results support an influence of selection intensity on the location of recombination modifiers. An important part of the respective selective forces could act on noncoding, presumably regulatory, sequences.[35] Accordingly, similar types of CNS genes that are associated with hotspot predictions also had shown genomic characteristics of genes with more tightly regulated expression.[30]

If selection intensity would exert an important influence on the distribution of meiotic recombination sites across the human genome, the maintenance of large-scale recombination rates[36] could be the reflection of persistent negative and recurrent positive selection that acts on regulatory sequences located in a wider genomic range around their targets genes. An influence of selection intensity on hotspot activity appears reasonable, because recombination hotspots seem to evolve as sequence-related traits, as shown by their relationship to the CCTCCCT motif[16] and the earlier observation of recombigenic alleles within hotspots.[37] In any case, the enrichment of recombination hotspots around CNS genes indicates that LD mapping studies of complex CNS disorders are particularly challenging. It also could make it more difficult to find a signature of positive selection at genes with important roles in the human CNS. This may particularly apply with multiple hotspots emerging and vanishing at those types of genes,[38,39] what may contribute to the limits of the outlier approach.[40] In this context, it might be worth mentioning that our hypothesis does not require that single recombination promoting alleles underlie the population genetic signatures that lead to hotspot predictions.

Recombination rates are further known to correlate with GC-content,[41] but GC-content explains only a small fraction of fine scale recombination rate variation.[17] Therefore, we consider it unlikely that GC-content exerts an important confounding influence on the functional bias in hotspot location. Moreover, further analysis shows that GC-content is rather below average than above average around important CNS genes[29] (and own unpublished data). Instead, the hypothesis that more intense selection leads to increased recombination (at CNS genes and elsewhere) raises the question, whether the correlation between GC-content and recombination could be confounded by the parallel correlation between GC-content and expression levels (and consequently selection intensity[42]). It is also possible to speculate about many additional confounding factors, such as for instance, LTR repeats,[16] which could interfere with the functional bias in hotspot location. However, even this speculative case does not necessarily provide a neutral explanation of the reported functional bias, because a biased location of recombination modifying LTR repeat could equally be the consequence of increased selection intensity at CNS gene loci. Moreover, any such speculative confounding factor would have to explain that the same types of CNS genes not only show an association with recombination hotspot predictions, but also display accelerated evolution in comparative genomic data.

An explanation for widespread ongoing selection at CNS gene loci might be given by the high prevalence of genetically complex CNS disorders with relatively early onset, for instance, epilepsy, migraine, schizophrenia and affective disorders. Because these CNS disorders are probably evolutionarily disadvantageous traits, selection may act against the underlying susceptibility alleles and in favor of protective alleles. Many respective susceptibility alleles could be a consequence of genetic changes that lead to human-specific CNS traits. Thus, successive genetic adaptations that produced human-specific traits might have changed the genetic background in a way that turned certain ancestral alleles into susceptibility alleles. That genetic background can determine disease allele status is well known from model organism genetics and monogenic disease. Therefore, one may hypothesize under an ancestral susceptibility allele model[8] that positive selection might act in favor of derived protective alleles. Ancestral alleles might be preferentially turned into susceptibility alleles, if encoding the 'conserved core' of the archetypic vertebrate brain.[30,43,44] Such CNS susceptibility genes may differ from other genes in the strength of constraint on sites that neighbor a positively selected allele, as indicated by their stronger sequence conservation.[1–3,29,30] Therefore, more recombination could be advantageous at these loci. If this explanation is correct, widespread positive selection would be the other side of a high locus heterogeneity of complex CNS disorders that is based on susceptibility alleles that are ancestral and common. The soon availability of genome-wide association data for various complex CNS disorders will allow for a systematical testing of this hypothesis.

In conclusion, we provide strong evidence for the preferential location of recombination hotspots at genes with important molecular role in the human CNS. We propose that the combined influence of intense negative and positive selection could have promoted the enrichment of hotspots at these loci. This would imply widespread positive selection on human CNS genes, which could be related to the high prevalence of nervous system disorders.

Acknowledgements

We thank Neil Risch, Jeff Wall, Yun Freudenberg-Hua and the anonymous review for comments. This work was supported by the Howard Hughes Medical Institute and a Sandler grant for Neurogenetics. LJP is an investigator of the Howard Hughes Medical Institute.

References

1 International_Rat_Genome_Consortium: Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 2004; **428**: 493–521.

2 Sironi M, Menozzi G, Comi GP, Cagliani R, Bresolin N, Pozzoli U: Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. *Hum Mol Genet* 2005; **14**: 2533–2546.
3 International_Chimp_Genome_Consortium: Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 2005; **437**: 69–87.
4 Dorus S, Vallender EJ, Evans PD *et al*: Accelerated evolution of nervous system genes in the origin of homo sapiens. *Cell* 2004; **119**: 1027–1040.
5 Prabhakar S, Noonan JP, Paabo S, Rubin EM: Accelerated evolution of conserved noncoding sequences in humans. *Science* 2006; **314**: 786.
6 Pollard KS, Salama SR, Lambert N *et al*: An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 2006; **443**: 167–172.
7 Rice WR, Chippindale AK: Sexual recombination and the power of natural selection. *Science* 2001; **294**: 555–559.
8 Di Rienzo A: Population genetics models of common diseases. *Curr Opin Genet Dev* 2006; **16**: 630–636.
9 Kong A, Gudbjartsson DF, Sainz J *et al*: A high-resolution recombination map of the human genome. *Nat Genet* 2002; **31**: 241–247.
10 Reich DE, Cargill M, Bolk S *et al*: Linkage disequilibrium in the human genome. *Nature* 2001; **411**: 199–204.
11 Abecasis GR, Noguchi E, Heinzmann A *et al*: Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 2001; **68**: 191–197.
12 Pritchard JK, Przeworski M: Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001; **69**: 1–14.
13 Smith AV, Thomas DJ, Munro HM, Abecasis GR: Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res* 2005; **15**: 1519–1534.
14 International_HapMap_Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
15 Hill WG, Robertson A: The effect of linkage on limits to artificial selection. *Genet Res* 1966; **8**: 269–294.
16 Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: A fine-scale map of recombination rates and hotspots across the human genome. *Science* 2005; **310**: 321–324.
17 McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: The fine-scale structure of recombination rate variation in the human genome. *Science* 2004; **304**: 581–584.
18 Ptak SE, Hinds DA, Koehler K *et al*: Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet* 2005; **37**: 429–434.
19 Winckler W, Myers SR, Richter DJ *et al*: Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* 2005; **308**: 107–111.
20 Karolchik D, Hinrichs AS, Furey TS *et al*: The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 2004; **32**: D493–D496.
21 Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P: Human recombination hot spots hidden in regions of strong marker association. *Nat Genet* 2005; **37**: 601–606.
22 Hubbard T, Andrews D, Caccamo M *et al*: Ensembl 2005. *Nucleic Acids Res* 2005; **33**: D447–D453.
23 Su AI, Wiltshire T, Batalov S *et al*: A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 2004; **101**: 6062–6067.
24 O'Donovan C, Apweiler R, Bairoch A: The human proteomics initiative (HPI). *Trends Biotechnol* 2001; **19**: 178–181.
25 Camon E, Magrane M, Barrell D *et al*: The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* 2004; **32**: D262–D266.
26 Ashburner M, Ball CA, Blake JA *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**: 25–29.
27 Storey JD, Tibshirani R: Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003; **100**: 9440–9445.
28 Hattori E, Liu C, Zhu H, Gershon ES: Genetic tests of biologic systems in affective disorders. *Mol Psychiatry* 2005; **10**: 719–740.
29 Siepel A, Bejerano G, Pedersen JS *et al*: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005; **15**: 1034–1050.
30 Freudenberg J, Fu YH, Ptacek LJ: Bioinformatic analysis of human CNS expressed ion channels as candidates for episodic nervous system disorders. *Neurogenetics* 2007. Advance Online Publication.
31 Keightley PD, Otto SP: Interference among deleterious mutations favours sex and recombination in finite populations. *Nature* 2006; **443**: 89–92.
32 Roze D, Barton NH: The Hill–Robertson effect and the evolution of recombination. *Genetics* 2006; **173**: 1793–1811.
33 Barton NH, Charlesworth B: Why sex and recombination ? *Science* 1998; **281**: 1986–1990.
34 Otto SP, Lenormand T: Resolving the paradox of sex and recombination. *Nat Rev Genet* 2002; **3**: 252–261.
35 King MC, Wilson AC: Evolution at two levels in humans and chimpanzees. *Science* 1975; **188**: 107–116.
36 Tiemann-Boege I, Calabrese P, Cochran DM, Sokol R, Arnheim N: High-resolution recombination patterns in a region of human chromosome 21 measured by sperm typing. *PLoS Genet* 2006; **2**: e70.
37 Jeffreys AJ, Neumann R: Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet* 2002; **31**: 267–271.
38 Spencer CC, Deloukas P, Hunt S *et al*: The influence of recombination on human genetic diversity. *PLoS Genet* 2006; **2**.
39 Coop G, Myers SR: Live hot, die young: transmission distortion in recombination hotspots. *PLoS Genet* 2007.
40 Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM: Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res* 2006; **16**: 980–989.
41 Meunier J, Duret L: Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol* 2004; **21**: 984–990.
42 Subramanian S, Kumar S: Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* 2004; **168**: 373–381.
43 Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y: The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol* 2006; **7**: R43.
44 Striedter GF: Precis of principles of brain evolution. *Behav Brain Sci* 2006; **29**: 1–12. Discussion 12–36.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)