ARTICLE

A Chromosomal Rearrangement Hotspot Can Be Identified from Population Genetic Variation and Is Coincident with a Hotspot for Allelic Recombination

Sarah J. Lindsay, Mehrdad Khajavi, James R. Lupski, and Matthew E. Hurles

Insights into the origins of structural variation and the mutational mechanisms underlying genomic disorders would be greatly improved by a genomewide map of hotspots of nonallelic homologous recombination (NAHR). Moreover, our understanding of sequence variation within the duplicated sequences that are substrates for NAHR lags far behind that of sequence variation within the single-copy portion of the genome. Perhaps the best-characterized NAHR hotspot lies within the 24-kb-long Charcot-Marie-Tooth disease type 1A (CMT1A)–repeats (REPs) that sponsor deletions and duplications that cause peripheral neuropathies. We investigated structural and sequence diversity within the CMT1A-REPs, both within and between species. We discovered a high frequency of retroelement insertions, accelerated sequence evolution after duplication, extensive paralogous gene conversion, and a greater than twofold enrichment of SNPs in humans relative to the genome average. We identified an allelic recombination hotspot underlying the known NAHR hotspot, which suggests that the two processes are intimately related. Finally, we used our data to develop a novel method for inferring the location of an NAHR hotspot from sequence variation within segmental duplications and applied it to identify a putative NAHR hotspot within the LCR22 repeats that sponsor velocardiofacial syndrome deletions. We propose that a large-scale project to map sequence variation within segmental duplications would reveal a wealth of novel chromosomal-rearrangement hotspots.

The sequencing of the human genome revealed that at least 5% of the genome consists of long, highly similar duplicated sequences known as "low-copy repeats" (LCRs), or segmental duplications.[1,2] These segmental duplications can have high sequence similarity (>90%), can be several hundreds of kilobases in length, and are enriched in ape genomes relative to genomes of other species.[2–4] Segmental duplications have been shown to have unusual patterns of sequence evolution relative to single-copy sequences, both in terms of orthologous sequence divergence and of reticulate evolution processes between duplications within the same genome,[5,6] and they may well play a central role in the evolution of novel gene function after gene duplication. Moreover, duplicated sequences appear to harbor unusual patterns of sequence variation within humans[7–10] that may result from gene conversion (the nonreciprocal transfer of sequence information between two homologous stretches of DNA) occurring between the duplicated copies.

Recent studies have revealed extensive structural variation within the human genome, with a marked enrichment of deletions, duplications, and inversions in and around segmental duplications.[11–15] The study of the functional importance of these structural variants is in its infancy, but the number of genetic diseases in which the structural dynamism conferred by segmental duplications plays a major role has been growing rapidly. The dominant role that genomic architecture plays in diseases—such as Charcot-Marie-Tooth disease type 1A (CMT1A [MIM 118220]) and hereditary neuropathy with pressure palsies (HNPP [MIM 162500]), due to reciprocal duplication and deletion on chromosome 17p12,[16] Smith-Magenis syndrome (SMS [MIM 182290]) and dup(17)(p11.2p11.2) due to reciprocal deletions and duplications on chromosome 17p11.2,[17] and deletions causing neurofibromatosis type I (NF1 [MIM 162200]) on chromosome 17q11.2[18] and Sotos syndrome (MIM 117550) on 5q35[19,20]—has led to them being classified as "genomic disorders."[21,22] Despite this work, there remains a substantial proportion of segmental duplications whose propensity to undergo nonallelic homologous recombination (NAHR) and to generate potentially disease-causing rearrangements is unknown. Although the association between structural variation and segmental duplications has been observed, the experimental demonstration of NAHR as a mechanism for such changes remains to be fully documented.

The germline mutational process underlying these observations of structural dynamism at segmental duplications is known as "NAHR" and can result in duplications, deletions, and inversions of genomic segments between copies of a duplicated sequence.[23] NAHR shares a number of important features with allelic meiotic recombination, which has led to suggestions that the two processes operate by similar mechanisms.[24] One striking similarity between allelic homologous recombination (AHR) and NAHR is the existence of hotspots of recombinatorial activity in

From the Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom (S.J.L.; J.R.L.; M.E.H.); and Departments of Molecular and Human Genetics (M.K.; J.R.L.) and Pediatrics (J.R.L.), Baylor College of Medicine, Houston

Received May 1, 2006; accepted for publication August 22, 2006; electronically published September 26, 2006.

Address for correspondence and reprints: Dr. Matthew Hurles, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom. E-mail: meh@sanger.ac.uk

Am. J. Hum. Genet. 2006;79:890–902. © 2006 by The American Society of Human Genetics. All rights reserved. 0002-9297/2006/7905-0011$15.00

which both crossovers and gene-conversion events cluster. In all genomic disorders in which the precise breakpoints of numerous independent rearrangements have been mapped, it has been found, by DNA sequence analysis of the products of recombination, that the breakpoints cluster within small intervals of greatly enhanced recombinatorial activity.[25] The likelihood of a breakpoint falling within one of these NAHR hotspots can be >2 orders of magnitude greater than in the surrounding sequence. These NAHR hotspots have size and morphology similar to experimentally determined AHR hotspots.[25] The study of AHR hotspots has been revolutionized by the genomewide inference of local recombination rates from patterns of sequence variation within populations. Whereas there are only ~10–20 experimentally determined AHR hotspots, the locations of ~50,000 AHR hotspots have been inferred throughout the genome from population genetic data.[26]

No such revolution has yet accelerated the discovery of NAHR hotspots. A genomewide map of NAHR hotspots would facilitate the identification of loci at which rearrangements result in embryonic lethality, would catalyze the discovery of other genomic disorders, and would inform our understanding of the origins of structural variation. Patterns of sequence variation and linkage disequilibrium (LD) within segmental duplications remain largely uncharacterized; segmental duplications are deemed outside the portion of the genome amenable to genomewide haplotype mapping.[27] Three types of variant sites are apparent within sequence alignments of duplicated sequences: sites that differ between allelic copies (i.e., SNPs), fixed sites that differ between paralogous copies (i.e., paralogous sequence variants [PSVs]), and a special class of SNPs that are polymorphic across paralogous sequences (i.e., multisite variants [MSVs]).

Given the role that NAHR hotspots potentially play in disease-causing rearrangements, it is of great interest to be able to characterize sequence variation within segmental duplications and to identify signatures of NAHR hotspot activity from these data. To develop a method that will enable identification of hotspots for NAHR solely from sequence variation, it is necessary to improve our understanding of the evolutionary processes occurring within segmental duplications. Elsewhere, we have demonstrated that, at two known Y-chromosomal NAHR hotspots, the presence of an NAHR hotspot could be inferred from comparisons between human and great ape sequences of the duplicated sequence containing the hotspot.[5] The extension of this method to autosomal NAHR hotspots has been thrown into doubt by the demonstration of the short-lived evolutionary nature of AHR hotspots.[28,29]

The 24-kb-long CMT1A-repeat (REP) segmental duplications[30] that sponsor pathogenic HNPP deletions and reciprocal CMT1A duplications are ideal loci for exploring the consequences of duplication on sequence evolution and for developing methods to identify NAHR hotspots. The CMT1A-REPs were duplicated recently on 17p11.2-12 in the common ancestor of humans and chimpanzees, with the distal copy ancestral, and the human copies share 98.7% sequence similarity.[30,31] These repeats contain a well-characterized ~600-bp-long NAHR hotspot that has an ~50-fold elevated rate of crossover compared with the surrounding sequence and is shared among populations.[32,33]

In this study, we characterized structural and sequence variation at the CMT1A-REPs in humans and hominoid species, by using a combination of Southern hybridization and resequencing by shotgun haplotyping.[34] We demonstrate that post-duplication gene conversion has altered the pattern and rate of sequence evolution in the CMT1A-REPs, and we develop a robust, novel method for identifying NAHR hotspots from patterns of sequence diversity within humans.

Material and Methods

Samples

Complete CMT1A-REP sequences were generated from genomic DNA from cell lines of (i) 10 unrelated males from the European Collection of Cell Cultures (ECACC) ethnic diversity panel (2 Australian Aborigine, 2 from the United Kingdom, 1 Italian, 1 Japanese, 2 Zulu, and 2 Native American) and (ii) chimpanzee, gorilla, orangutan, and gibbon, from the ECACC primate panel.

Southern hybridization and limited resequencing was performed on 72 samples from the CEPH Human Genome Diversity Panel and on samples from the Baylor College of Medicine control panel (93 African American, 98 Hispanic, 95 European American, and 72 Asian American).

Southern Hybridization Screening for Structural Variation of CMT1A-REPs

Restriction-enzyme digests were performed according to the manufacturer's instructions. We used a dosage-analysis approach, using a CMT1A-REP probe derived from a purified restriction fragment from a cosmid described elsewhere.[16] Probes labeled with ^{32}P-α-deoxycytidine triphosphate with the Rediprime II labeling kit (Amersham Pharmacia Biotech) identified two *Eco*RI restriction fragments on chromosome 17p11.2-12 (a 7.9-kb *Eco*RI fragment localized to the proximal CMT1A duplication monomer region and a 6.1-kb *Eco*RI fragment mapping to the distal CMT1A region). The 7.9-kb proximal and 6.1-kb distal *Eco*RI fragments are contained entirely within the CMT1A-REP sequence.

Long PCR of CMT1A-REPs

Each CMT1A-REP was amplified in two portions, with the use of a non–repeat-specific internal primer and an external primer located in flanking single-copy sequence. The distal repeat was amplified in two portions with use of the oligo pairs (1) CMT1AD2 CCACATTACTGCTTCCTCATGTGT and CMT1AINT5 GTTCATGGTTCATGCTGAGGGTTG and (2) CMT1AD1 GGGGGTAGAAAAGGGGTCTCATTTTCC and CMT1AINT3 ATTACAGCTACTGTTGCAGCAGTG, which amplified products of 12,777 and 11,327 bp, respectively. The proximal repeat was amplified in two portions, with use of the oligo pairs (3) CMT1AP2 CTTAGCCATTGCCCATTGATGGAC and CMT1AINT5 GTTCATGGTTCATGCTGAGGGTTG and (4) CMT1AP1 CCATTAGAGAGCTTTCCTCATTGC and

CMT1AINT3 ATTACAGCTACTGTTGCAGCAGTG, which amplified products of 12,600 and 11,344 bp, respectively.

These PCR fragments do not overlap in the center of the repeat, so additional primers were designed to obtain the genotypic sequence for the middle portion of the repeats. The gap between PCR products runs from 10,230 to 11,304 bp in our alignment, so the SNP information in this region between haplotypic sequences comes from genotypic sequences unless specified. To obtain gap sequence, long PCR was performed as described above, but with the use of primers CMT1AP1 CCATTAGAGAGCTTTCCTCATTGC and CMT1A_Join1 GCAGTGATGCTCAGTAGAAAG, at an annealing temperature of 60°C and an extension time of 13 min, and with CMT1AD1 GGGGGTAGAAAAGGGGTCTCATTTTCC and CMT1A_Join3 GGGCTGATGTTTAGTAAACAA, at an annealing temperature of 57°C and an extension time of 13 min.

All PCR reactions were performed in a 50-μl volume with the use of the Expand 20Kb Plus PCR kit (Roche Applied Science) and 200 ng of genomic DNA as template. Unless otherwise stated, the reactions were performed following the manufacturer's protocol, with an extension time of 11 min and an annealing temperature of 57°C. All oligos were synthesized by Sigma Genosys.

Resequencing of CMT1A-REPs

The long-PCR products were fragmented, cloned, and shotgun sequenced to a high depth (>20× coverage) with the use of the shotgun-haplotyping method,[34] which recovers haplotypic sequence across the length of the PCR product by assembling read pairs from the two alleles into separate assemblies. To obtain sequence data from the middle of the repeat, we direct sequenced additional PCR products, using the PCR primer CMT1A_Join3 and an additional internal sequencing primer, CMT1A_Join2 CATAGAAATGTGTGGACCAAT.

The sequence data were then assembled using the Gap4 assembly software; SNPs were automatically called, and then the haplotypic sequences were exported. In two individuals, one Native American male (AMA) and one U.K. male (C07220), the alleles from one of the long PCRs (3 and 4, respectively) were monomorphic. Targeted resequencing of individuals with unusual Southern banding patterns was performed using shotgun haplotyping or with staged primers (see table 1) after amplification with PCR primers listed above. The GenBank accession numbers for the sequences generated in this study are DQ480370–DQ480419.

Statistical and Evolutionary Analysis

For genotypic analyses of the entire CMT1A-REP, the component haplotypic sequences were arbitrarily spliced together to form 24-kb allelic sequences for each individual, with each half containing true haplotypic sequence. Sequences were aligned with the CMT1A-REP GenBank reference sequences, with the use of BioEdit and Se-Al. All analyses not dependent on having true haplotypes were derived from these alignments of spliced haplotypes. Repetitive elements were detected by RepeatMasker. Jukes-Cantor distances and nucleotide diversity (π) were calculated from full 24-kb primate and human reference sequences for each repeat, with the use of PHYLIP and DNASP,[35] respectively.

Phylogenetic networks and trees were constructed, using SplitsTree,[36] from full 24-kb sequences from the human and primate panels for each repeat. Neighbor-joining trees were constructed using 24-kb sequences from the human and primate panels, with the use of PHYLIP under Felsenstein 84 (F84) and Jukes-Cantor models of evolution with 1,000 bootstrap replicates. TREE-PUZZLE[37] was used to construct maximum-likelihood trees, to compute branch lengths, and to perform the likelihood ratio to test the molecular clock hypothesis. Single alleles from each CMT1A-REP from all the primate species and from a single human were used and were modeled using F84 distances, with the gibbon sequence specified as an outgroup.

Table 1. Primers Used in Targeted Resequencing

The table is available in its entirety in the online edition of *The American Journal of Human Genetics*.

The sliding window plots (with window size 700 bp) of the two indices for identifying NAHR hotspots—concerted index and hotspot index—were generated in Excel, from lists of variant site (i.e., SNPs, PSVs, and MSVs) output from alignments, by code written in Interactive Data Language 6.0 (Research Systems). A permutation test written in Interactive Data Language 6.0 was performed to test the significance of the hotspot index; 10,000 replicates were performed, in which the positions of the observed numbers of MSVs, SNPs, and PSVs were randomized along a 24-kb stretch of DNA.

The haplotype-reconstruction program PHASE 2.1[38,39] was used to predict AHR hotspots in the proximal and distal CMT1A-REPs. All human CMT1A-REP sequences were entered in genotypic form for the analysis, to avoid any generation of false recombinants by erroneous splicing of haplotypes. The PHASE algorithm was iterated 10 times, to increase accuracy.

Results

Structural Variation in CMT1A-REPs Is Driven by Retroelement Insertion

To investigate structural diversity within the CMT1A-REPs in humans and chimpanzees, we performed Southern analysis with panels of diverse, phenotypically normal individuals. This Southern assay uses a probe across a region of the repeats that contains an *Eco*RI site specific to the distal repeat, and, therefore, different-sized bands for proximal and distal CMT1A-REPs are obtained.[16] Accordingly, in individuals who have two copies of each repeat, two bands of the same intensity will be seen. In patients with CMT1A and HNPP, copy-number changes manifest themselves as an altered relative intensity of the two bands.[40] Among 165 globally diverse humans, three unusual additional patterns were observed (figs. 1 and 2). Among chimpanzees ($n = 20$), four different banding patterns were observed.

To investigate the mutational processes underlying these different banding patterns, we resequenced portions of the CMT1A-REPs in individuals with some of the above unusual banding patterns. We also sequenced the entire proximal and distal CMT1A-REPs in a chimpanzee and the entire ancestral CMT1A-REP in one gorilla, orangutan, and gibbon (see below), and we identified additional structural variants. We chose to sequence these regions, using the method of shotgun haplotyping from long-PCR products anchored in single-copy sequence. Because of the inher-

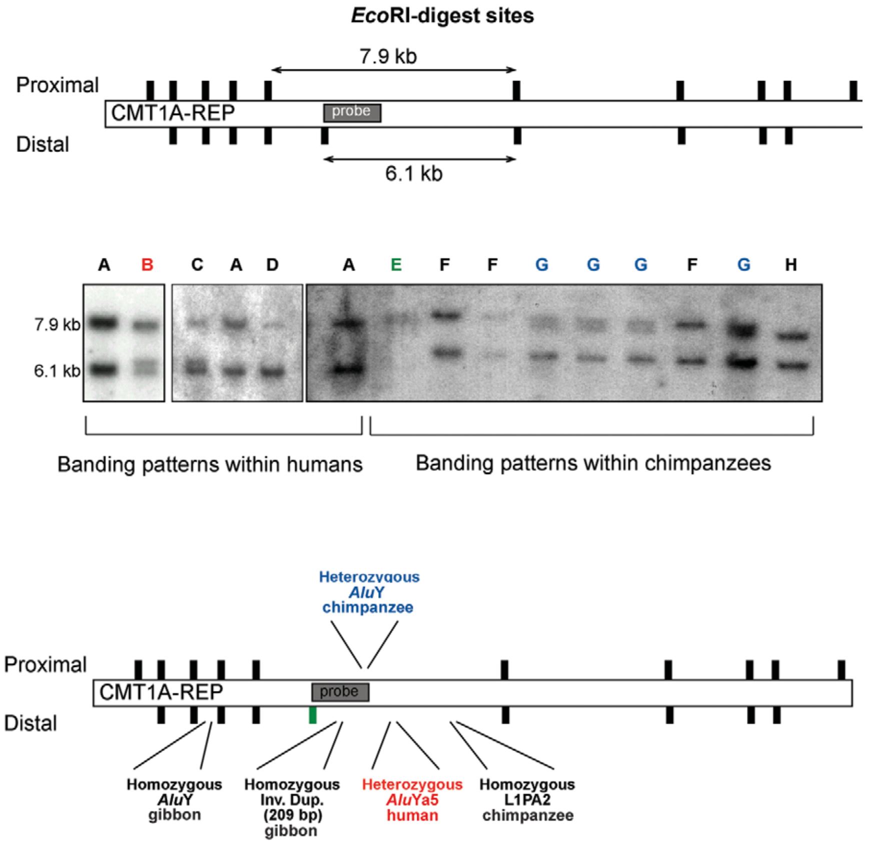
EcoRI-digest sites
7.9 kb
Proximal
CMT1A-REP
probe
Distal
6.1 kb
A B C A D A E F F G G G F G H
7.9 kb
6.1 kb
Banding patterns within humans
Banding patterns within chimpanzees
Heterozygous AluY chimpanzee
Proximal
CMT1A-REP
probe
Distal
Homozygous AluY gibbon
Homozygous Inv. Dup. (209 bp) gibbon
Heterozygous AluYa5 human
Homozygous L1PA2 chimpanzee

Figure 1. Structural variation within CMT1A-REPs. Top panel shows the Southern probe binding site (*horizontal shaded rectangle*) and the positions of *Eco*RI sites in the CMT1A-REPs (*black vertical bars*). Middle panel shows the different Southern banding patterns identified in humans and chimpanzees, with each distinct pattern given a one-letter code. Bottom panel shows the sites of insertion events identified in the sequences described in the text. These insertion events are color coded, to identify which banding pattern they correspond with. A variable *Eco*RI restriction site is shown as a green bar, color coded to reflect the associated banding pattern. The absence of this restriction site, in combination with the presence of the partial (2,123-bp) L1PA2 insertion, accounts for the larger distal REP fragment apparent in chimpanzee Southern banding patterns F, G, and H, as compared to humans.

ently high (>20 ×) sequence coverage required for shotgun haplotyping, we can have much higher confidence in identifying variant sites than we would with standard genotypic sequencing.[34] Moreover, the ratio of alleles at a variant site depends on the number of sequences that were coamplified, with a deviation from 1:1 ratios indicating the presence of additional copies. In none of these data did we see evidence of copy numbers of the CMT1A-REPs different from expectation (two in human and chimpanzee and one in gorilla, orangutan, and gibbon). Whereas losses and a gain of an *Eco*RI restriction site were identified in our sequencing and could explain some of the unusual banding patterns, more striking was the observation of four independent retroelement insertions events (three *Alu*Y and an L1PA2) within a 7.5-kb interval that encompasses the known NAHR hotspot (fig. 1). This includes a polymorphic *Alu*Ya5 insertion within the distal CMT1A-REP, which is found in an African American population with an allele frequency of 3% and is not in the dbRIP database of retroelement insertion polymorphisms.[41] These data show a high frequency of independent insertions, which are not mediated by gene conversion or duplication from *Alu* insertions already present in the repeats, since there are no homologues for them inside the CMT1A-REPs and each insertion site displays the characteristic target side duplication of a retroelement transposition.

Rate of Sequence Evolution in CMT1A-REPs Increases after Duplication

It has been suggested from simulations, theoretical considerations, and limited comparative resequencing that the rate of sequence evolution may increase after a duplication event.[5,42] Since the duplication event generating CMT1A-REPs has already been placed on the primate phylogeny,[30,31] we decided to investigate sequence divergence in the CMT1A-REPs, to test this hypothesis. We sequenced full CMT1A-REP sequences from chimpanzee, gorilla, orangutan, and gibbon. We compared the sequence divergence in CMT1A-REPs between humans and other

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Figure 2. Global Southern analysis of CMT1A-REPs. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics.*

hominoid species with the genome averages for intergenic nonrepetitive sequences for these species[43] and could not identify significantly discrepant sequence divergences. This kind of analysis is, however, not robust to regional variation in the rate of sequence evolution. Analysis of the chimp draft genome shows substantial regional variation around the average value of 1.23%.[44] Acceleration in the rate of sequence evolution at the CMT1A-REPs would be masked if these sequences were evolving more slowly than the genome-average rate, as a result of selective constraint on functional sequences within the CMT1A-REPs. There is only a single short exon of the *COX10* gene in one of the CMT1A-REPs.[30] However, conserved (PhastCons) sequence elements comprise twice as much of the CMT1A-REPs (~8%)[45] than the genome average (~4.3%), so it is plausible that sequence constraint does depress the overall rate of sequence evolution at these loci.

A more powerful method of detecting unequal rates of sequence evolution is to compare rates of sequence evolution at the same locus in different lineages. We compared the likelihood of obtaining the observed sequences of the CMT1A-REPs in the hominoids mentioned above, under two models in which (i) rates of sequence evolution were constrained to be the same on all lineages (constant rates) and (ii) rates of sequence evolution were allowed to vary across lineages (free rates). A likelihood-ratio test rejected the null hypothesis of a constant rate of evolution at the 5% level; allowing the evolutionary rates to vary between lineages is a better fit to the data (see table 2). These results suggest that the rates of sequence evolution have not been equal in the CMT1A-REPs in different lineages. If we compare branch lengths in the two phylogenies, we can identify which lineages appear to be evolving more rapidly than expected. Three of the four post-duplication lineages (both human lineages and one of the chimpanzee lineages) have longer branch lengths in the free-rates phylogeny than in the constant-rates phylogeny, despite the observation that sequence evolution is generally slower in humans and chimpanzees than in other hominoid species,[46] which suggests that our likelihood-ratio test is actually conservative. This conclusion is also robust to CpG biases, since the GC content in the distal CMT1A-REP is not significantly differentiated among hominoid sequences (range 40.30%–40.45%).

One mechanism by which duplicated sequences may have a higher rate of sequence evolution than single-copy sequences is gene conversion between paralogous sequences, which can act to introduce additional variation into a duplicated sequence. The presence of gene conversion can perturb the normal phylogenetic relationships between orthologous and paralogous sequences.[5] The phylogenetic relationships among the hominoid CMT1A-REPs sequences mentioned above are already known: the duplication event occurred in the common ancestor of humans and chimpanzees, and, thus, the proximal CMT1A-REP in humans should share a common ancestor with the proximal repeat in CMT1A-REPs in chimpanzees, to the exclusion of all other sequences. A neighbor-joining phylogeny reconstructed from the hominoid CMT1A-REP sequences does not show the expected phylogenetic relationships: the human proximal CMT1A-REP seems most closely related to the human distal CMT1A-REP (see fig. 3). Low bootstrap values around the nodes of phylogeny relating chimpanzee and human sequences suggest that this phylogeny is not well supported by the data. These results might be explained either by complex speciation processes resulting in the gene tree not reflecting the species tree[47] or by gene conversion shuffling variation between CMT1A-REPs. To further explore whether gene conversion may be the cause of this phylogenetic discrepancy, we constructed a SplitsTree of the same sequences (fig. 3). A SplitsTree is a network-based phylogeny that displays conflicting signals within the data as cycles (also known as "reticulations"). Cycles can be generated by any evolutionary process that introduces conflicting evolutionary signals into sequence alignments—for example, frequent recurrent mutation, reversion mutation, or any process leading to concerted evolution, such as either gene conversion or unequal crossing over. If gene conversion is acting on the duplicated CMT1A-REPs, then cycles should appear in the SplitsTree but only among the post-duplication lineages. By contrast, complex speciation[47] might give a phylogeny that is different from the expected phylogeny (given the species tree), but this phylogeny should be constant along the length of the CMT1A-REPs, and so cycles should not be observed. We observed clear cycles in the SplitsTree, with these cycles being confined to the relationships among post-duplication lineages.

Table 2. Results of Molecular Clock Likelihood-Ratio Test

	Tree Branch Length (substitutions per base)		
Lineage	Free Rate[a]	Constant Rate	Fold Difference
Human distal	**.00643**	.00561	1.15
Human proximal	**.00592**	.00561	1.06
Chimp proximal	**.00666**	.00609	1.09
Chimp distal	.00466	.00609	.77
Gorilla	.00703	.00743	.95
Orangutan	.01565	.01582	.99
Gibbon	.01899	.01898	1.00

[a] Longer tree branch lengths in the ingroup species are shown in bold.

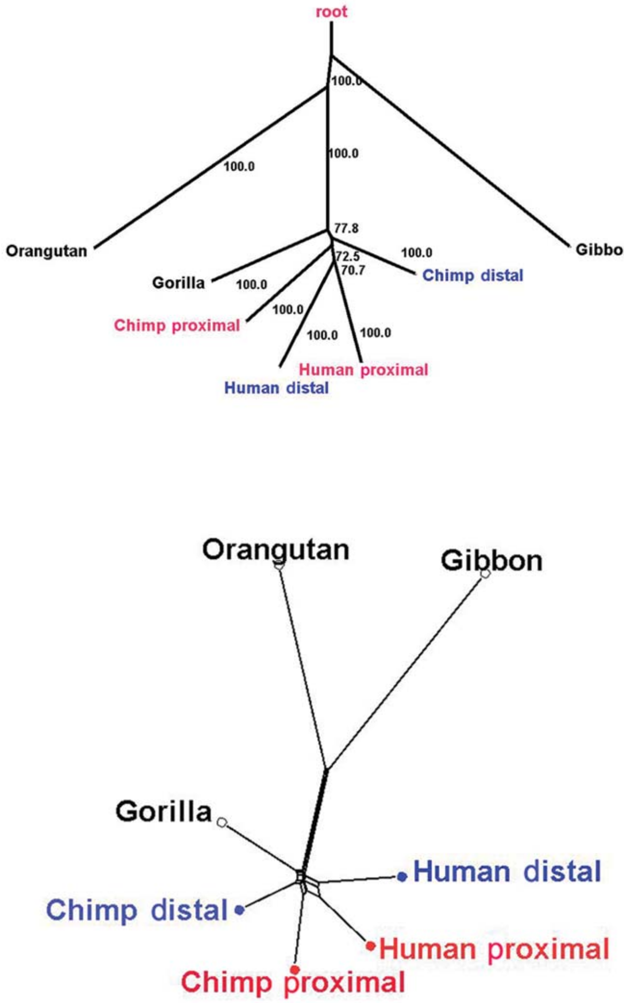

Figure 3. Phylogenies of CMT1A-REPs. Upper panel displays a rooted neighbor-joining phylogeny of CMT1A-REPs in hominoid species, labeled with bootstrap percentages. Proximal sequences are shown in red, and distal sequences in blue. Lower panel displays a SplitsTree of CMT1A-REPs in hominoid species, constructed using uncorrected *P* distances. Proximal sequences are shown in red, and distal sequences in blue.

Gene Conversion Elevates Sequence Diversity in the CMT1A-REPs

There is an apparent enrichment of SNPs in segmental duplications (including the CMT1A-REPs) in dbSNP.[48] It has been argued that this increased SNP density in segmental duplications could be due to the mismapping of variation between copies of a segmental duplication, which could lead to PSVs (positions conserved in each copy but different between them) (see fig. 4) being mistaken for SNPs.[2] An alternative explanation is that this increased diversity is genuine and results from gene conversion acting to increase sequence diversity by introducing a PSV from one repeat into the other as a SNP.[42]

We resequenced the CMT1A-REPs in 10 globally diverse humans, blind to any possible structural variation. By using the method of shotgun haplotyping from long-PCR products anchored in single-copy sequence, we can be sure that we are not confusing PSVs with SNPs and that we have much higher confidence in identifying SNPs than we would by using standard genotypic sequencing.[34] These sequences showed greatly elevated values of π in both proximal (0.00177) and distal (0.00165) CMT1A-REPs, which are more than twofold greater than the genome average of 0.000751.[49] There is no evidence of a more regional (as opposed to CMT1A-REP–specific) elevation of sequence diversity in the levels of heterozygosity or of the frequency of dbSNP entries in the single-copy sequence flanking the distal CMT1A-REP (HapMap).

To explore whether gene conversion is the cause of this elevated π, we constructed a SplitsTree comprising only human CMT1A-REP sequences (fig. 5). Although the proximal and distal REPs are clearly differentiated, there are extensive cycles in both major clusters within the SplitsTree, which suggests the influence of gene conversion.

Three types of variant sites are apparent within alignments of proximal and distal CMT1A-REP sequences (fig. 4). In addition to SNPs (sites that differ between allelic copies) and PSVs (fixed sites that differ between paralogous copies), we also identify SNPs that are found at the same location and with the same alleles in both proximal and distal REPs. Here, this third class of variants are known as "MSVs" and correspond to the variant class MSV_2, as defined by Fredman et al.[10]; elsewhere, these have also been dubbed "shared polymorphic sites."[9]

If gene conversion is operating between proximal and distal CMT1A-REPs, we should expect (i) that the number of SNPs shared between repeats (MSVs) is much greater than we would expect if the two repeats were evolving independently and (ii) that repeat-specific variants (PSVs) are frequently gene converted into the other repeat and appear as new SNPs. Given that we identified 180 SNPs in the proximal repeat and 143 SNPs in the distal repeat, if the locations of these SNPs in proximal and distal repeats were independent, we would expect, on average, to observe one or two MSVs (this number is simply the product of the frequencies of variant sites at either locus multiplied by the length of the CMT1A-REP). However, we observed 24 MSVs (SNPs shared between the repeats), which is a highly statistically significant enrichment (χ^2 test $P = 2 \times 10^{-60}$). Removing MSVs that might conceivably represent recurrent mutation at CpG dinucleotides still leaves 11 MSVs, which remains a highly significant enrichment (χ^2 test $P = 1 \times 10^{-11}$). It is more difficult to confirm the second expectation; PSVs defined from aligning a single proximal and a single distal CMT1A-REP from humans (typically taken from the human reference sequence) are not independent of SNP locations in humans. Some apparent PSVs defined in this manner would be derived SNP alleles, and this confounds any attempts to test whether SNPs are more frequent at PSVs than we would expect. Ideally, we would know what the ancestral PSVs

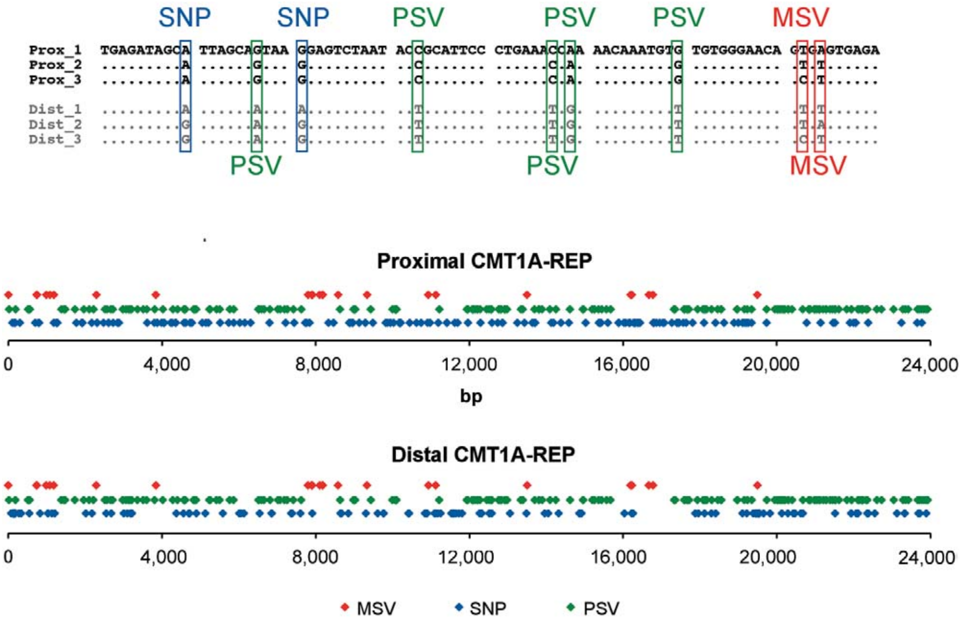

Figure 4. Definition and distribution of SNPs, PSVs, and MSVs. Upper panel shows examples of SNPs, PSVs, and MSVs within a portion of an alignment of three distal (*gray*) and three proximal (*black*) CMT1A-REP sequences. Lower panels display the positions of these three classes of variants within proximal and distal CMT1A-REPs in 20 human chromosomes.

were between the most recent common ancestors of the two repeats. Although we don't have this information, we do have full sequences of both CMT1A-REPs in chimpanzee, and defining PSVs from these sequences is independent of SNPs in humans. We find that 19 of the 266 PSVs identified in chimpanzees are SNPs in humans, with the same alleles as the alternate states of the chimpanzee PSVs. This is a highly significant enrichment of human SNPs at sites of chimpanzee PSV (χ^2 test $P = 1.6 \times 10^{-8}$). Thus, gene conversion explains why such a high percentage (~38%) of the apparent PSVs ($n = 260$) identified from aligning the proximal and distal CMT1A-REPs in the human reference sequence are variable between alleles in our data.

Known NAHR Hotspot Can Be Identified from Patterns of Sequence Variation within Species but Not between Species

It has been shown elsewhere that a sliding window statistic known as the "concerted index" can be used to identify localized gene conversion resulting from a known NAHR hotspot in alignments of AZFa-HERVs in great ape species.[5] This statistic contrasts the sequence similarity between paralogous and orthologous sequences, to identify regions where the paralogous sequences have become homogenized and the orthologous sequences have diverged. If there is appreciable divergence between orthologous sequences but paralogous sequences are very similar, it is likely that the ancestral sequences were not very similar, and, hence, any similarity between paralogues is due to active homogenization rather than to ancestral similarity. We applied the concerted index to an alignment of complete CMT1A-REP proximal and distal sequences from a single human and chimpanzee (fig. 6). In contrast to results from the study of AZFa-HERVs, there is no evidence of a peak of the concerted index around the known NAHR hotspot or elsewhere in the alignment of CMT1A-REPs. In addition, we also applied phylogenetic profiling,[50] to search for evidence of location-specific NAHR within alignments of CMT1A-REP sequences from both repeats in hominoid species, with no success (data not shown). Phylogenetic profiling is best suited to the identification of recombination events in sequence alignments; identifying localized gene-conversion hotspots requires that narrow window sizes be used, and this increases the variance, which confounds visual analysis.

To devise a statistic to seek a signal of the known NAHR hotspot within alignments of human CMT1A-REP sequences, we considered the likely impact of gene conversion on the frequency of the three types of variants that exist in such alignments: SNPs, PSVs, and MSVs (fig. 4). Gene conversion of a PSV will generate a new SNP in the locus receiving the gene-conversion tract. Thus, gene conversion can cause the number of PSVs to decrease and the number of SNPs to increase. Gene conversion of the derived allele of a SNP present in one CMT1A-REP will generate a SNP in the same location in the paralogous CMT1A-REP. This shared SNP is reclassified as an MSV. Thus, gene conversion has caused the number of MSVs to increase and the number of SNPs to decrease. In summary, frequent gene conversion should convert both SNPs and PSVs into MSVs. Thus, in regions of elevated NAHR characterized by

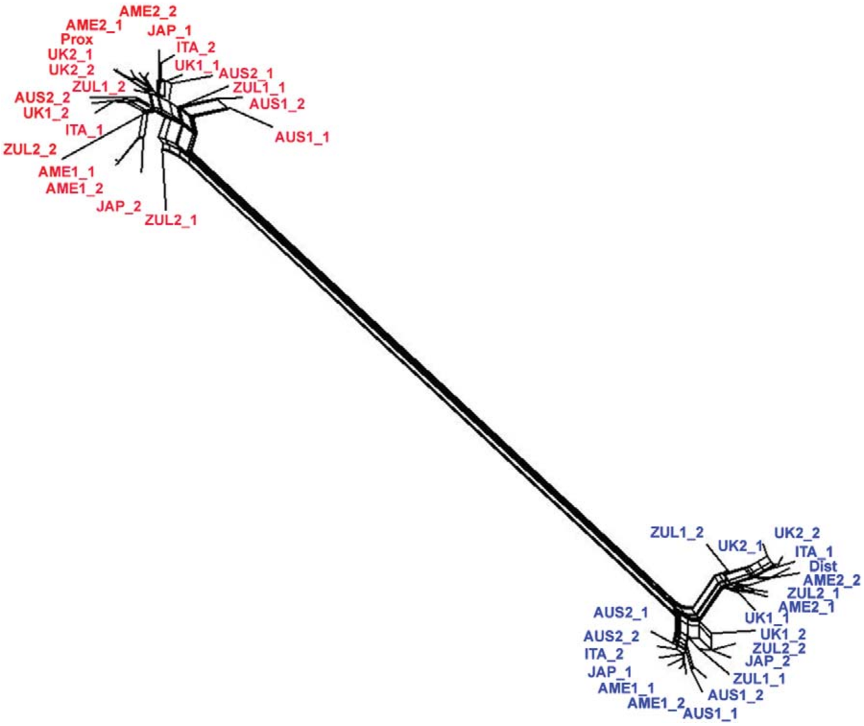

Figure 5. SplitsTree of all proximal (*red*) and distal (*blue*) human CMT1A-REPs

high rates of paralogous gene conversion, the frequency of MSVs should be higher, and the frequency of SNPs and PSVs lower, than elsewhere within the alignment. We devised the hotspot index (MSVs/(SNPs+PSVs)) on the basis of the frequency of different types of variant sites within a window of a sequence alignment, to reflect these expectations. The hotspot index should give high values in localized regions with elevated paralogous gene conversion and low values where gene conversion is absent. The hotspot index can be calculated in sliding windows along an alignment, and a simple permutation test is used to identify regions of significantly elevated values.

If we apply our hotspot index to the alignment of human CMT1A-REP sequences, we see a very strong and significant signal of a paralogous gene-conversion hotspot ~8 kb into the 24-kb alignment (fig. 6). This hotspot precisely overlaps the known, experimentally determined NAHR hotspot characterized from patients with HNPP deletions or CMT1A duplications. An additional, much weaker signal of paralogous gene conversion is seen at a location ~16 kb into the alignment, which corresponds with a much weaker NAHR hotspot.[51] It can be seen that these two signals of NAHR hotspot activity clearly coincide with a cluster of MSVs (fig. 4). Both of these peaks are highly significant, compared with the 99th percentile of permutations. These observations are robust to the removal of all MSVs that could potentially represent recurrent mutations at CpG dinucleotides. Whereas the absolute intensity of the hotspot index is reduced by removing these MSVs, the strongest peak in signal is still over the NAHR hotspot and remains highly significant, compared with the 99th percentile (fig. 7).

Sequence variation remains uncharacterized in the other segmental duplications harboring known NAHR hotspots; however, one recent study[9] has used overlapping finished BAC sequences to characterize sequence variation at the LCR22 repeats known to sponsor the deletion causing velocardiofacial syndrome (VCFS [MIM 192430]). Although no NAHR hotspot has yet been experimentally identified at these segmental duplications, we applied the hotspot index to these sparser LCR22 data and found a number of potential NAHR hotspots that were highly significant (fig. 8). Using related considerations, the authors of this recent study of sequence variation at LCR22[9] highlighted a similar set of regions within these segmental duplications as being candidate NAHR hotspots, and it remains to be seen whether the breakpoints of patients with VCFS cluster at this purported hotspot.

Relationship between NAHR Hotspots and AHR Hotspots

It has been suggested that NAHR hotspots may originate from the presence of AHR hotspots within duplicated sequences. The locations of AHR hotspots have been inferred from large genomewide SNP genotyping data. However, inference of accurate recombination-rate estimates requires high genotyping accuracy and high SNP density. High-quality SNP genotyping data are at a much lower

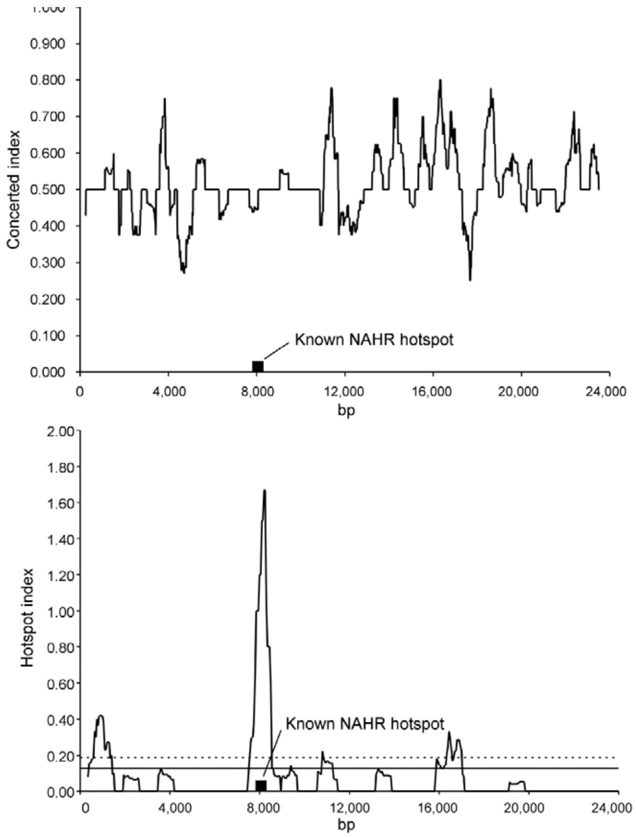

Figure 6. Identifying NAHR hotspots from patterns of sequence variation. Upper panel shows how the concerted index varies along the length of a 24-kb alignment of human and chimpanzee CMT1A-REPs. Lower panel shows how an alternative sliding window statistic, the hotspot index, varies along an alignment of human proximal and distal CMT1A-REPs. Solid and dashed horizontal lines indicate the 95th and 99th percentiles, respectively, of the hotspot index, as determined by 10,000 random permutations of the positions of SNPs, PSVs, and MSVs.

density within segmental duplications in the Phase I HapMap data,[27] and, as a result, segmental duplications appear very much depleted for AHR hotspots, as compared with single-copy regions of the genome. There are no AHR hotspots within the CMT1A-REPs inferred from either HapMap Phase I or Perlegen data.[26,27]

Rather than use data from these genomewide surveys of LD, we can use our resequencing data to seek signals of an AHR hotspot. We estimated recombination rates across both proximal and distal CMT1A-REPs from our resequencing data, using the software PHASE (fig. 9). In both proximal and distal CMT1A-REPs, there is an apparent recombination hotspot at the site of the known NAHR hotspot, with the hotspot in the distal CMT1A-REP very much stronger than the hotspot in the proximal CMT1A-REP (~400 vs. ~6 times greater than background, respectively). It is unclear how NAHR will confound estimates of allelic recombination rates; however, we have shown above that gene conversion (NAHR) between CMT1A-REPs influences patterns of sequence variation in a highly localized fashion, whereas an allelic recombination hotspot should influence LD over longer distances. Therefore, we removed all the SNPs within known NAHR hotspots from the analysis and reestimated recombination rates in both REPs. The apparent recombination hotspot in the proximal CMT1A-REP disappears entirely, whereas an attenuated recombination hotspot remains in the distal CMT1A-REP. This attenuated recombination hotspot appears to lie adjacent to the known NAHR hotspot and retains a recombination rate intensity ~60-fold greater than the surrounding background recombination rate (90% symmetric CI 9 to 889 times).

We searched within the known NAHR hotspot, for sequence motifs that have recently been identified as being enriched within allelic recombination hotspots,[26] and identified the 7-mer sequence CTCCTCC as being present in both proximal and distal CMT1A-REPs. This motif is closely related to the top-scoring hotspot-enriched motif CCTCCCT[26] but is not associated with the apparently recombinogenic THE1B repeat.

Discussion

We have shown that duplication of the CMT1A-REPs in the common ancestor of humans and chimpanzees has had a profound effect on subsequent processes of sequence evolution. The pattern of this sequence evolution is strikingly reticulate, both between humans and chimpanzees and among humans. This is consistent with recent studies showing that reticulate evolution is a common feature of duplicated sequences in the human genome, due to processes of concerted evolution that may include both gene conversion and unequal crossing over.[6] We have also demonstrated accelerated sequence evolution in the CMT1A-REPs after duplication. This finding is in agreement both with simulations of gene conversion and with limited empirical evidence from other loci.[5] The significant enrichment of SNPs shared between CMT1A-REPs in humans (i.e., MSVs) provides strong additional support that gene conversion is the predominant mechanism driving the unusual patterns of variation in the CMT1A-REPs.

One unexpected finding of our comparative sequencing and resequencing was the prevalence of retroelement insertions within a 7-kb interval containing the known NAHR hotspot. Whereas it might be tempting to infer a relationship between the NAHR hotspot and these insertion events, it is difficult to determine the significance of

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Figure 7. The hotspot index with potential CpG sites removed. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics*.

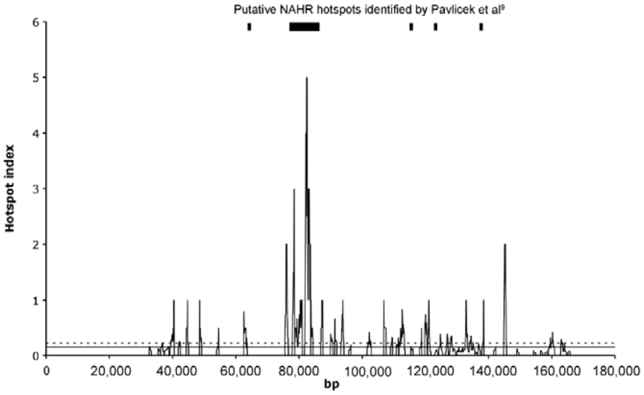

Figure 8. Putative NAHR hotspots in LCR22-2 and LCR22-4. This figure shows how the hotspot index varies along an ~180-kb alignment of LCR22-2 and LCR22-4. Solid and dashed horizontal lines indicate the 95th and 99th percentiles, respectively, of the hotspot index, as determined by 10,000 random permutations of the positions of SNPs, PSVs, and MSVs.

the rate and location of retroelement insertions because of the manner in which some of them were ascertained. By screening for structural variation with a Southern assay with a probe near the known NAHR hotspot, we selectively enriched for retroelement insertions in this region. However, two of the retroelement insertions that we discovered—the L1 insertion in a chimpanzee and an *Alu*Y gibbon insertion—were ascertained through sequencing with no prior knowledge of structural variation; both of these insertions occur within the 7-kb interval containing the known NAHR hotspot. This suggests that there may be some element of mutational fragility to this region of the CMT1A-REP and that this fragility has persisted over the span of ~20 million years. This suggestion is supported by the existence of other sites of chromosomal fragility that are also segmentally duplicated on 17p and have been demonstrated to have been involved in structural variation within and between species.[52]

Previous studies based on genotyping have demonstrated that there are odd patterns of sequence variation within segmental duplications that do not conform to our expectations of single-copy SNPs.[10] We have demonstrated not only that these patterns of sequence variation are different but also that the density is different. There is more than twofold greater SNP diversity within the CMT1A-REPs than in the genome average. The prospect that segmental duplications might harbor greater sequence diversity than single-copy sequences was first suggested by the observation that segmental duplications are enriched for dbSNP entries.[48] There are two possible explanations for this observation: first, that PSVs have been assigned to the wrong REP[2] and so generated a false SNP, and, second, that SNP diversity really is greater in segmental duplications.[42] These explanations are not necessarily mutually exclusive. We have shown directly that elevated SNP diversity is a real feature of sequence variation at autosomal segmental duplications. This is, however, not to say that dbSNP is not contaminated with mismapped PSVs masquerading as SNPs. If we compare the dbSNP entries in proximal and distal REPs with our own resequencing data based on 20 chromosomes, we find discrepant patterns of variation within dbSNP that suggest that there may be PSV contamination. A far greater proportion of SNPs in dbSNP are shared between both CMT1A-REPs (34%) than in our data (7%), and yet the total number of SNPs in dbSNP (n = 478) is not correspondingly greater than in our data (n = 323). Moreover, in our data, 38% of the 290 PSVs apparent when the reference sequences of the CMT1A-REPs are aligned appear to be SNPs, whereas, in dbSNP, 84% of these PSVs appear as SNPs. This suggests that, although segmental duplications may well harbor elevated levels of sequence diversity, dbSNP may not be a reliable guide to this diversity.

The CMT1A-REPs harbor a known hotspot for NAHR.

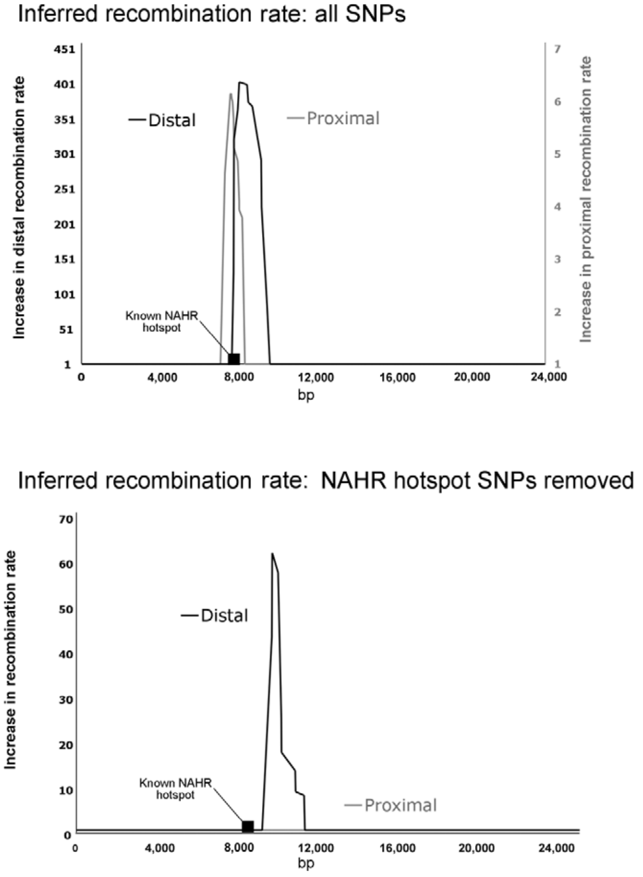

Figure 9. Allelic recombination rates within CMT1A-REPs. Upper panel shows the estimated local recombination rates along proximal (*gray*) and distal (*black*) CMT1A-REPs with the use of all SNPs. Lower panel shows the estimated local recombination rates in proximal and distal CMT1A-REPs once SNPs in the known ~700-bp hotspot (7,800–8,500 bp) have been removed. The location of the known NAHR hotspot is indicated.

We considered that NAHR hotspots may share with AHR hotspots the feature that a hotspot for crossing over is also a hotspot for gene conversion.[8,53,54] Using a priori expectations of the influence of gene conversion on the distribution of three types of variants (SNPs, PSVs, and MSVs) apparent within alignments of CMT1A-REP sequences, we found that, by far, the strongest signal of localized gene conversion was tightly localized around the known NAHR hotspot.

We have shown elsewhere that an NAHR hotspot on the human Y chromosome can be identified from patterns of sequence variation between humans and chimpanzees.[5] The sliding window statistic used in that study failed to identify the known NAHR hotspot in the present study. The success of a within-species measure of NAHR—and the failure of a between-species measure—suggest that the known NAHR hotspot in the CMT1A-REPs is recent in origin. This is supported by the sequence divergence between CMT1A-REPs not being suppressed in the known NAHR hotspot, as would be expected had the hotspot been undergoing homogenization by gene conversion over a longer evolutionary time frame. This recent origin of an autosomal NAHR hotspot is reminiscent of the shallow time depth of AHR hotspots, which also appear to not be shared between humans and chimpanzees.[28,29] This observation raises the question of whether NAHR hotspots are simply AHR hotspots within segmental duplications, as has been suggested recently.[24] When we sought signals of an allelic recombination hotspot in our resequencing data and removed the confounding influence of SNPs within the NAHR hotspot, we identified an AHR hotspot flanking the NAHR hotspot in the distal CMT1A-REP, with a recombination rate ~60-fold greater than the surrounding sequence. This provides strong evidence that the NAHR hotspot in the distal CMT1A-REP is closely correlated with an AHR hotspot, which was not apparent from analysis of the genotyping data produced by the HapMap project.[27] More detailed studies of local allelic recombination rates in other NAHR hotspots are required to further support the relationship between AHR and NAHR hotspots.

The number of known NAHR hotspots could be greatly increased, with a concomitant improvement in our understanding of the underlying mutational mechanism, if many more NAHR hotspots could be identified from patterns of sequence variation rather than from the laborious mapping of rearrangement breakpoints in patients. It should also be possible to identify NAHR hotspots that sponsor rearrangements that are never seen in the population because of embryonic lethality. To achieve this revolution in understanding, surveys of sequence variation within segmental duplications are required. It is notoriously difficult to characterize sequence variation within segmental duplications, and the methods used in this study will not easily scale to genomewide analyses. However, sufficiently informative data might be collected by other means that are more scalable. We could consider that PSVs identified from alignments of segmental duplications in the human reference sequence and in existing dbSNP entries might give us a sufficiently dense set of variants, so that a genotyping-based approach, rather than a sequencing-based approach, might be feasible. To test this hypothesis, we thinned our sequencing data by removing any variants that were not apparent as PSVs in the alignment of the human reference CMT1A-REPs or in dbSNP. Encouragingly, we found that the known NAHR hotspot still represents the strongest signal of the hotspot index and is still highly significant, as judged by the permutation test (fig. 10). We are embarking on experiments to try these more scalable methods on other known NAHR hotspots.

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Figure 10. Hotspot index on thinned genotypic data. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics.*

The success of our method for identifying further NAHR hotspots will depend not only on the age of the hotspot but also on SNP density and on the frequency of gene conversion and crossovers occurring between the repeats. Once we have further characterized sequence variation in segmental duplications, it seems likely that we will soon be able to construct a genomewide map of the locations of NAHR hotspots as a more disease-orientated counterpart to the recently published map[26] of allelic recombination hotspots in the single-copy portion of the genome. This map would catalyze our identification of haplolethal loci (where deletions cause embryonic lethality), would aid the discovery of novel genomic disorders, and would increase our understanding of the mechanisms generating structural variation in the human genome.

Acknowledgments

S.J.L. and M.E.H. are supported by the Wellcome Trust. M.K. and J.R.L. are supported by the National Institute of Neurological Disorders and Stroke, National Institutes of Health, and by the Charcot-Marie-Tooth Association. The authors are grateful to David Willey, Chris Clee, Lucy Matthews, and Bob Plumb, for help with shotgun haplotyping; to Chris Tyler-Smith, for guidance with Southern analysis; to Matthew Stephens, for advice on using PHASE; to Bernice Morrow and Adam Pavlicek, for sharing sequence alignments; and to Daniel Turner, for comments on the manuscript. We are also grateful to HGDP-CEPH for providing DNA samples. A detailed description of the panel has been published elsewhere.[55]

Web Resources

Accession numbers and URLs for data presented herein are as follows:

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for CMT1A-REP sequences [accession numbers DQ480370–DQ480419])

HapMap, http://www.hapmap.org/

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for CMT1A, HNPP, SMS, NF1, Sotos syndrome, and VCFS)

PHYLIP, http://evolution.genetics.washington.edu/phylip.html

RepeatMasker, http://www.repeatmasker.org/

Se-Al, http://evolve.zoo.ox.ac.uk/software.html?id = seal

References

1. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE (2001) Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 11: 1005–1017
2. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. Science 297: 1003–1007
3. Samonte RV, Eichler EE (2002) Segmental duplications and the evolution of the primate genome. Nat Rev Genet 3:65–72
4. Stankiewicz P, Lupski JR (2002) Molecular-evolutionary mechanisms for genomic disorders. Curr Opin Genet Dev 12:312–319
5. Hurles ME, Willey D, Matthews L, Hussain SS (2004) Origins of chromosomal rearrangement hotspots in the human genome: evidence from the AZFa deletion hotspots. Genome Biol 5:R55
6. Jackson MS, Oliver K, Loveland J, Humphray S, Dunham I, Rocchi M, Viggiano L, Park JP, Hurles ME, Santibanez-Koref M (2005) Evidence for widespread reticulate evolution within human duplicons. Am J Hum Genet 77:824–840
7. Hallast P, Nagirnaja L, Margus T, Laan M (2005) Segmental duplications and gene conversion: human luteinizing hormone/chorionic gonadotropin beta gene cluster. Genome Res 15:1535–1546
8. Bosch E, Hurles ME, Navarro A, Jobling MA (2004) Dynamics of a human inter-paralog gene conversion hotspot. Genome Res 14:835–844
9. Pavlicek A, House R, Gentles AJ, Jurka J, Morrow BE (2005) Traffic of genetic information between segmental duplications flanking the typical 22q11.2 deletion in velo-cardio-facial syndrome/DiGeorge syndrome. Genome Res 15:1487–1495
10. Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ (2004) Complex SNP-related sequence variation in segmental genome duplications. Nat Genet 36:861–866
11. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38:75–81
12. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951
13. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528
14. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78–88
15. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM (2006) Common deletion polymorphisms in the human genome. Nat Genet 38:86–92
16. Pentao L, Wise CA, Chinault AC, Patel PI, Lupski JR (1992) Charcot-Marie-Tooth type-1a duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. Nat Genet 2:292–300
17. Chen KS, Manian P, Koeuth T, Potocki L, Zhao Q, Chinault AC, Lee CC, Lupski JR (1997) Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. Nat Genet 17:154–163
18. Lopez Correa C, Brems H, Lazaro C, Marynen P, Legius E (2000) Unequal meiotic crossover: a frequent cause of *NF1* microdeletions. Am J Hum Genet 66:1969–1974
19. Kurotaki N, Stankiewicz P, Wakui K, Niikawa N, Lupski JR (2005) Sotos syndrome common deletion is mediated by directly oriented subunits within inverted Sos-REP low-copy repeats. Hum Mol Genet 14:535–542
20. Visser R, Shimokawa O, Harada N, Kinoshita A, Ohta T, Niikawa N, Matsumoto N (2005) Identification of a 3.0-kb major recombination hotspot in patients with Sotos syndrome who carry a common 1.9-Mb microdeletion. Am J Hum Genet 76: 52–67
21. Lupski JR, Stankiewicz P (2005) Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet 1:e49
22. Lupski JR, Stankiewicz P (2006) Genomic disorders: the genomic basis of disease. Humana Press, Totawa, New Jersey
23. Stankiewicz P, Lupski JR (2002) Genome architecture, rearrangements and genomic disorders. Trends Genet 18:74–82
24. Lupski JR (2004) Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. Genome Biol 5:242
25. Hurles ME, Lupski JR (2006) Recombination hotspots in nonallelic homologous recombination. In: Lupski JR, Stankiewicz P (eds) Genomic disorders: the genomic basis of disease. Humana Press, Totawa, New Jersey
26. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324
27. IHMC (2005) A haplotype map of the human genome. Nature 437:1299–1320
28. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P, Altshuler D (2005) Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308:107–111
29. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S (2005) Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet 37:429–434
30. Reiter LT, Murakami T, Koeuth T, Gibbs RA, Lupski JR (1997) The human *COX10* gene is disrupted during homologous recombination between the 24 kb proximal and distal CMT1A-REPs. Hum Mol Genet 6:1595–1603
31. Kiyosawa H, Chance PF (1996) Primate origin of the CMT1A-REP repeat and analysis of a putative transposon-associated recombinational hotspot. Hum Mol Genet 5:745–753
32. Reiter LT, Murakami T, Koeuth T, Pentao L, Muzny DM, Gibbs RA, Lupski JR (1996) A recombination hotspot responsible for

two inherited peripheral neuropathies is located near a mariner transposon-like element. Nat Genet 12:288–297
33. Lopes J, Ravise N, Vandenberghe A, Palau F, Ionasescu V, Mayer M, Levy N, Wood N, Tachi N, Bouche P, Latour P, Ruberg M, Brice A, LeGuern E (1998) Fine mapping of de novo CMT1A and HNPP rearrangements within CMT1A-REPs evidences two distinct sex-dependent mechanisms and candidate sequences involved in recombination. Hum Mol Genet 7:141–148
34. Lindsay SJ, Bonfield JK, Hurles ME (2005) Shotgun haplotyping: a novel method for surveying allelic sequence variation. Nucleic Acids Res 33:e152
35. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497
36. Huson DH (1998) SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14:68–73
37. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504
38. Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens M (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet 36:700–706
39. Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233
40. Chance PF, Abbas N, Lensch MW, Pentao L, Roa BB, Patel PI, Lupski JR (1994) Two autosomal-dominant neuropathies result from reciprocal DNA duplication/deletion of a region on chromosome-17. Hum Mol Genet 3:223–228
41. Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P (2006) dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat 27:323–329
42. Hurles ME (2002) Are 100,000 "SNPs" useless? Science 298: 1509
43. Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68:444–456
44. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87
45. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–1050
46. Elango N, Thomas JW, Yi SV (2006) Variable molecular clocks in hominoids. Proc Natl Acad Sci USA 103:1370–1375
47. Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103–1108
48. Estivill X, Cheung J, Pujana MA, Nakabayashi K, Scherer SW, Tsui LC (2002) Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. Hum Mol Genet 11:1987–1995
49. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933
50. Weiller GF (1998) Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol Biol Evol 15:326–335
51. Reiter LT, Hastings PJ, Nelis E, de Jonghe P, van Broeckhoven C, Lupski JR (1998) Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. Am J Hum Genet 62:1023–1033
52. Stankiewicz P, Shaw CJ, Withers M, Inoue K, Lupski JR (2004) Serial segmental duplications during primate evolution result in complex human genome architecture. Genome Res 14: 2209–2220
53. Jeffreys AJ, May CA (2004) Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat Genet 36:151–156
54. Bi W, Park SS, Shaw CJ, Withers MA, Patel PI, Lupski JR (2003) Reciprocal crossovers and a positional preference for strand exchange in recombination events resulting in deletion or duplication of chromosome 17p11.2. Am J Hum Genet 73: 1302–1315
55. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, et al (2002) A human genome diversity cell line panel. Science 296:261–262