

A common sequence motif associated with recombination hot spots and genome instability in humans

Simon Myers^{1,2}, Colin Freeman², Adam Auton^{2,3}, Peter Donnelly^{2,4} & Gil McVean²

In humans, most meiotic crossover events are clustered into short regions of the genome known as recombination hot spots. We have previously identified DNA motifs that are enriched in hot spots, particularly the 7-mer CCTCCCT. Here we use the increased hot-spot resolution afforded by the Phase 2 HapMap and novel search methods to identify an extended family of motifs based around the degenerate 13-mer CCNCCNTNCCNC, which is critical in recruiting crossover events to at least 40% of all human hot spots and which operates on diverse genetic backgrounds in both sexes. Furthermore, these motifs are found in hypervariable minisatellites and are clustered in the breakpoint regions of both disease-causing nonallelic homologous recombination hot spots and common mitochondrial deletion hot spots, implicating the motif as a driver of genome instability.

Although the genomic locations of many human hot spots have been identified, an understanding of the relationship between DNA sequence and hot-spot location remains elusive. Previously, using a genome-wide map of recombination hot spots estimated from genetic variation data, we carried out an exhaustive search of short (5- to 9-mer) motifs for enrichment in hot spots and identified two, CCTCCCT and CCCCACCCC, which were strongly overrepresented in a small fraction (~10%) of human hot spots, with the first motif particularly over-represented (five- to sixfold) in THE1A/B retrotransposons within hot spots¹. Direct evidence for a role of these motifs in hot-spot activity came from studies of polymorphic hot spots, where single nucleotide variants reducing crossover activity in *cis* at hot spots DNA2 (ref. 2) and NID1 (ref. 3) disrupt motifs CCTCCCT and CCCCACCCC, respectively.

Nevertheless, the presence of either short motif is, by itself, a poor predictor of hot spot location and fails to explain most human hot spots. There are three possible solutions. First, there may be other, completely different and perhaps longer motifs that we failed to identify. Second, the identified motifs may be specific examples of

an extended degenerate family of motifs. Third, there may be no *cis*-acting sequence determinants for the majority of hot spots. To distinguish between these hypotheses, we have used an alternative approach to identifying hot spot-associated motifs that looks for sequence similarity between hot spot-associated regions both in repeats and in unique DNA. Furthermore, we have gained additional power by using recombination hot spots identified from the Phase 2 HapMap⁴ (22,699 autosomal and 608 chromosome X hot spots mapped to within 5 kb).

The rationale for the approach taken is that any generic hot spot-promoting motif should operate on diverse genetic backgrounds (such as in different repeat families). We first identified classes of repeat elements that are over-represented in hot spots (see Methods) and subsequently searched for motifs that are independently associated with enhanced hot-spot presence on multiple repeat-family backgrounds. This approach revealed the presence of a common 13-bp degenerate motif CCNCCNTNCCNC, which is related to the previously identified motif CCTCCCT (Table 1, Supplementary Tables 1–3 and Supplementary Note online). Notably, although repeats carrying the motif showed a narrow peak in average recombination rate centered at the motif, repeats with the motif-lacking consensus showed no such peak (Fig. 1). Consequently, the presence of the motif fully accounts for the enrichment of these repeat elements within hot spots.

In nonrepeat DNA, using the previously identified¹ motif CCTCCCT as a foundation, we identified individual flanking bases within 50 bp that are influential in determining hot-spot occurrence (see Methods). In agreement with our findings for repetitive DNA, this analysis revealed the presence of the 4-mer CCAC two bases downstream from the 7-mer to be the strongest additional determinant of hot-spot occurrence (Fig. 2a; $P < 10^{-30}$ for all four bases together, by FET). This implicates the core motif CCTCCCTNCCAC for both repeat (in the case of THE1A/B and L2 elements) and nonrepeat DNA. We observed an almost identical pattern of enrichment of the 13-mer motif in hot spots on the X chromosome

¹Broad Institute of Massachusetts Institute of Technology and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. ²Department of Statistics, Oxford University, 1 South Parks Road, Oxford OX1 3TG, UK. ³Department of Biological Statistics and Computational Biology, 101 Biotechnology Building, Cornell University, Ithaca, New York 14853, USA. ⁴Wellcome Trust Centre for Human Genetics, Oxford University, Roosevelt Drive, Oxford OX3 7BN, UK. Correspondence should be addressed to S.M. (myers@stats.ox.ac.uk).

Received 15 February; accepted 18 July; published online 24 August 2008; doi:10.1038/ng.213

Table 1 Repeat elements enriched in hot spots and their hot spot–associated motifs

Repeat element	Hot spots containing element	Cold spots containing element	Corrected <i>P</i> value	Hot spot–motif region	Motif <i>P</i> value	Chromosome X <i>P</i> value
THE1B	1,196	606	$<2 \times 10^{-16}$	TGTGAGGCCTCCCTAGCCAC [—] CGTGGAACT	$<2 \times 10^{-16}$	5.5×10^{-4}
THE1A	234	89	1.8×10^{-13}	GAGGCCTCCCTAGCCAC [—] GT	$<2 \times 10^{-16}$	4.6×10^{-2}
GA-rich/ CT-rich	976	662	4.5×10^{-12}	CCTCCCTT	4.0×10^{-4}	6.1×10^{-3}
	1,005	737	7.5×10^{-8}	CCTCCTT	4.2×10^{-2}	
L2	10,113	9,271	7.8×10^{-7}	GAGGCCTCCCTGACCACC	$<2 \times 10^{-16}$	2.9×10^{-1}
AluY	3,634	3,284	1.4×10^{-2}	GATCCGCC [—] NCCTTGGCCTCCCA	$<2 \times 10^{-16}$	1.2×10^{-3}

Total number of narrow hot spots and matched cold spots on autosomes is 22,673. In hot spot–motif regions, mutations relative to the repeat family consensus are underlined. Sequences matching the consensus CCNCCNTNCCNC are shown in bold. For AluY, the N indicates that all four bases were observed in hot spot–enriched motifs at this position. The motif *P* value reported is the smallest Bonferroni-corrected *P* value.

(Table 1), indicating that the 13-bp motif operates on multiple backgrounds in both males and females.

Further testing of motifs occurring outside repeats and mismatching a single base of the 13-bp ‘core’ revealed additional degeneracy within the motif at positions 3, 6 and 12, with mismatches at these three bases still consistent with some hot-spot activity (Fig. 2b and Supplementary Methods online) and a consensus of CCNCCNTNCCNC. These degenerate positions correspond exactly to the mismatching sites within repeat motifs (Table 1) and motifs mismatching at these locations still showed hot-spot activity, albeit reduced, across the other repeats as well (Fig. 1a). Notably, the polymorphism at the DNA2 hot spot corresponds to the first apparently degenerate position within the motif², suggesting that site-wise degeneracy may fail to represent more complex dependencies between nucleotides.

In order to estimate what proportion of the 22,699 narrowly defined autosomal hot spots require the presence of the 13-base motif, we applied a maximum likelihood approach (Supplementary Methods). Our approach attempted to account for two facts: first, that motif

occurrences only stimulate a hot spot with some probability; and second, that each hot spot (defined to 3–5 kb) can contain several motifs, on different backgrounds, each of which may contribute (we assume independently) to hot-spot activity. We estimate that the location of $41.1\% \pm 1.4\%$ of all human hot spots is determined by the presence of the motif (95% confidence interval estimated by bootstrapping, see Supplementary Methods). Mechanistically, this means that recruitment of crossover events to a particular hot spot requires the presence of one or more copies of the motif in 41% of hot spots.

It is also important to know how well the presence of the motif predicts hot spots (that is, the extent to which the presence of the motif is sufficient). The penetrance of the motif varied between genetic backgrounds (Supplementary Table 4 online). For example, the presence of CCTCCCTNCCAC in a THE1A background resulted in a detected hot spot 73% of the time, whereas in unique DNA it led to a detected hot spot 10% of the time. The high predictive power on the THE1A background is, in part, because the context of the repeat element leads to the presence of the most recombinogenic nucleotides

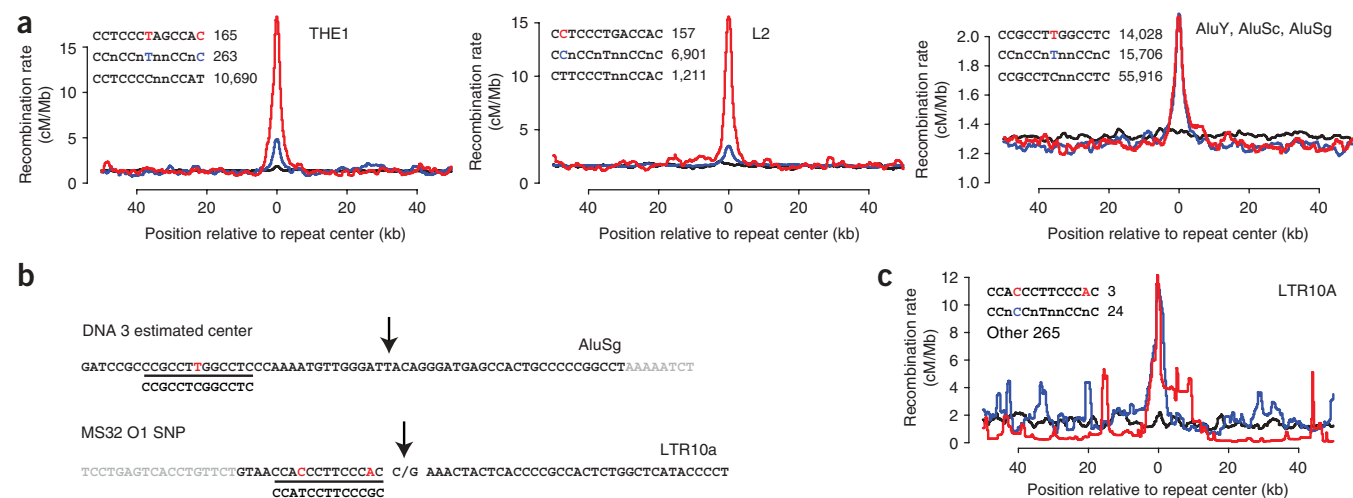


Figure 1 A common hot-spot motif acts across different repeat families. **(a)** Average recombination rates around repeat elements THE1, L2 and combined AluY, AluSc, AluSg in the genome (see Supplementary Fig. 2 online for separate plots for each Alu family and the Supplementary Note for a discussion of other Alu families). For each plot, the three lines relate to repeats carrying the identified repeat-specific hot-spot motif (red; the motif indicated by the top sequence), repeats with other sequences that match the degenerate hot-spot motif (blue; the consensus is shown in the middle line) and repeats carrying the consensus for that family (black line; sequence shown at the bottom; numbers indicate sample size). Deviations from the repeat-family consensus in sequence are indicated by colored letters. Note that we allow degeneracy in the repeat-family consensus at positions 8 and 9 of the hot-spot motif consensus. **(b)** Known hot spots carrying newly identified instances of the hot-spot motif. The center of the DNA3 hot spot⁷ occurs in an AluSg sequence containing the Alu-specific hot-spot motif, which differs from the Alu consensus at position 7 (repeat-family consensus shown below). The O1 SNP in the MS32 hot spot⁸, which influences hot-spot activity, occurs in an LTR10A repeat element one base pair 3' from a match to the hot-spot consensus motif. This motif differs from the LTR10A repeat-family consensus at two positions (indicated in red). **(c)** Estimated recombination rates around LTR10A elements carrying the same 13-base sequence as at MS32 (red; excluding MS32), other matches to the hot-spot consensus motif (blue) and other sequences (black).

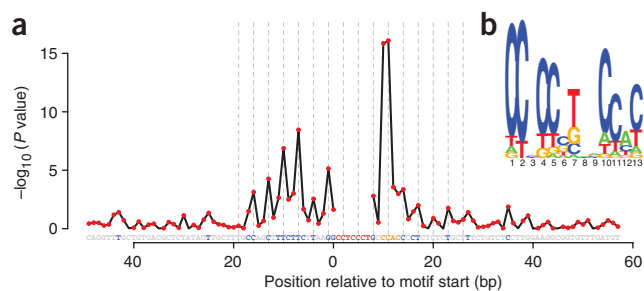


Figure 2 The role of flanking sequence and motif degeneracy in determining hot-spot activity. **(a)** The evidence ($-\log_{10} P$ value using χ^2 test) for a difference in base composition at nucleotides surrounding the motif CCTCCCT between narrowly defined hot spots and matched cold spots. The sequence shows the most over-represented base at each position; positions showing significant differences ($P < 0.01$) are blue, the fixed CCTCCCT motif is shown in red and the other specified bases within the repeat-based 13-bp motif are shown in orange. Vertical dotted lines spaced at 3-bp intervals highlight periodic occurrence of strongly signaled bases. **(b)** Degeneracy within the core 13-base motif estimated by comparing counts of each motif mismatching exactly 1 bp of the 13-bp core CCTCCCTNNCCAC in hot spots and matched cold spots. The combined height of the stacked letters at each position is proportional to the $-\log_{10} P$ value and the relative height of each letter is proportional to its over-enrichment in hot spot-associated motifs (**Supplementary Methods**).

outside the motif, as defined in **Figure 2a**. Overall, we found that only a fraction of hot spots are driven by highly penetrant motif-background combinations (for example, the location of 3.5% of hot spots is determined by motif-background combinations with a relative risk of ten or more; **Supplementary Table 4**). For the majority of hot spots, other factors (including motif density and the additional context features shown in **Figure 2a**) must interact with motif

presence. Reports of distant or even *trans* effects on crossover activity^{5,6} also indicate that the presence of short sequence motifs is not fully sufficient to determine hot-spot activity.

To what extent is the expanded, degenerate motif responsible for determining the location of any of the 17 hot spots identified and studied using direct analysis of human sperm? In addition to the DNA2 hot spot (where the presence of the 7-mer motif was described previously¹), both the DNA3 hot spot in the HLA class II region⁷ and the MS32 hot spot⁸ contained the degenerate 13-mer motif within a few base pairs of the estimated center (16 bp and 1 bp, respectively). In addition, an exact match to the 13-bp core was found within 300 bp of the estimated center of the HLA hot spot DMB2 (ref. 7). The probability of such proximity between hot-spot center and the degenerate motif occurring by chance alone is 0.0046 for DNA3, 0.0016 for MS32 and 0.059 for DMB2, although note that DMB2 contains the core motif, which occurs only once every 450 kb on average (these P values are not Bonferroni corrected, see the **Supplementary Note** for additional discussion and supporting evidence). In the MS32 hot spot, sperm typing has identified a single-base mutation within an LTR10A repeat element that associates with hot-spot activity⁸. This C/G polymorphism is located 1 bp downstream of the hot-spot motif (**Fig. 1b**). On the basis of our examination of sequence flanking nonrepeat motifs (**Fig. 2a**), we would predict that mutation to a G allele at this site would reduce crossover activity, as observed. Comparison of other LTR10A elements confirmed that the presence of the motif is specifically associated with a local increase in recombination rate (**Fig. 1c**).

Our results implicate the 13-mer motif in allelic crossover activity during meiosis. A natural question is whether the motif might also have a role in other forms of recombination or recombination-associated genome rearrangement, including nonallelic homologous recombination (NAHR), minisatellite mutation and repeat-associated deletion and rearrangement. To date, breakpoints of NAHR-generated

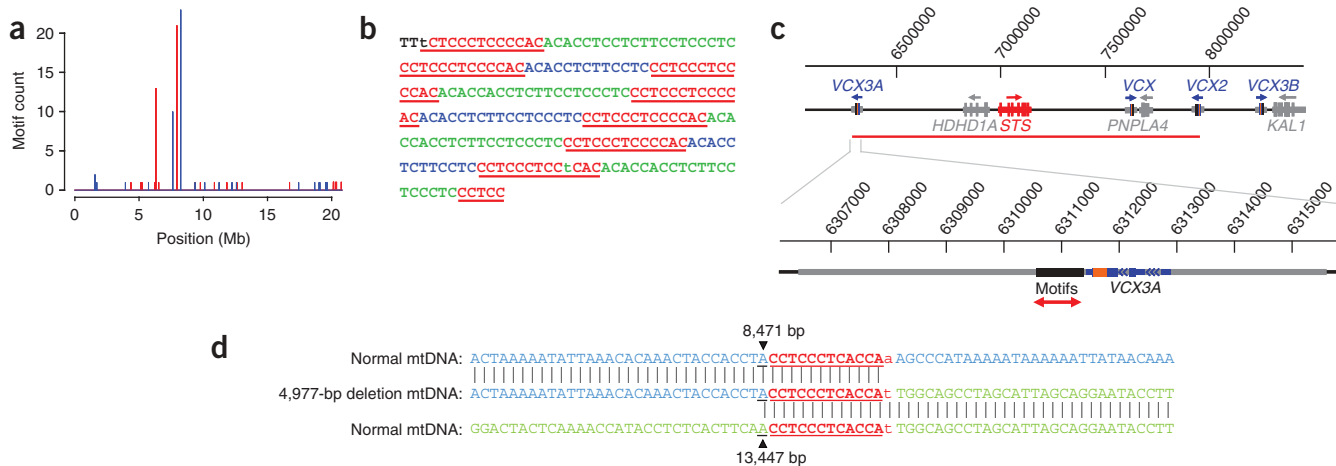


Figure 3 Hot-spot sequence motifs at STS deletion hot spot and common mitochondrial deletion endpoints. **(a)** The density of exact matches to the motif CCTCCCTNNCCAC in 5-kb windows along the first 20 Mb of chromosome X shows four grouped clusters of motif occurrences (motifs on the plus (red) and minus (blue) strands are shown separately). **(b)** Each motif cluster corresponds to a tandem repeat region downstream of a *VCX* gene family member; alternating repeats are colored in blue and green. Shown is the genome reference sequence for the first seven repeats downstream of the *VCX3A* gene. Matches to the hot-spot motif occur once per repeat and are shown red, underlined (mismatching bases, lower case). **(c)** Nonallelic homologous recombination (NAHR) between two of four homologous repeats, each containing *VCX* genes, removes the *STS* gene (red) and causes X-linked ichthyosis. The four repeats are marked as *VCXx* and arrows show their orientation. NAHR between directly oriented *VCX3A* and *VCX* can delete *STS*. The lower plot shows the architecture of the *VCX3A*-containing homologous repeat (thick gray line), including the gene itself and the downstream motif-containing tandem repeat within which deletion breakpoints have been shown to cluster¹². **(d)** The mtDNA 'common deletion'. The top line shows normal 5' mtDNA surrounding base 8,471, the bottom line 3' mtDNA surrounding base 13,447 and the middle sequence deletion carrying mtDNA (base matches shown by lines). The deletion occurs within a 13-bp direct repeat (underlined) of which 12 bp overlap almost exact matches to the hot-spot motif (red).

Table 2 Hypervariable human minisatellites and hot-spot motifs

Minisatellite (reference)	Consensus repeat unit	Repeat unit contains motif	Mutation rate and array size correlation	Mutational polarity	Pedigree mutation rate (%)
MS1 (ref. 16)	<u>CCACCC</u> TATCCACCCTAT ^a	Yes	–	No ^b	5.20
CEB1 (refs. 17,18)	AGCCAGGGACCTCCGAGGCCACCC <u>TCCCTCCCC</u> CTC	Yes	Yes	Weak	6.70
B6.7 (ref. 21)	CTCATGTCCTATAGAGACCC <u>CCTCACTG</u> TCCACC	Yes	Yes	No	5.50
D7S22 (ref. 15)	TCCTGCCT <u>TCCCTGCC</u> CACACCGCCCGCTTTCTAT ^c	Yes	Yes	No	1.30
MS32 (ref. 8)	GAGCAGGTGGCCAGGGGTGACTCAGAATG	No	No	Yes	0.40
MS20 (ref. 20)	GGCGGGTGAAGTAGACGGCGTGCCTAGGGGCCGAGCGGGGTGTGACGG	No	No	Yes	0.40
MS31 (ref. 19)	CCACCTCCACAGACTGC	No	–	Yes	0.75
Insulin (INS-VNTR) (ref. 22)	ACAGGGGTGGGG	No	No	Weak	~0.10

^aThe MS1 consensus is only 9 bp long, and the motif is formed by tandem copies of the consensus; therefore, two tandem copies of the repeat unit are shown (the first copy is underlined).

^bFor MS1, inter-allelic events show weak, but not statistically significant, clustering towards the 3' end of the array. ^cA subset of variant repeats at D7S22 perfectly match the hot-spot core sequence.

rearrangements have been mapped at the sequence level for only a few diseases. In several cases, this has revealed strong breakpoint clustering into hot spots within particular genomic repeats^{9–12}, and in three diseases (NF1 microdeletion¹³, Charcot-Marie-Tooth disease type 1A (CMT1A) and hereditary neuropathy with pressure palsies¹⁴), these coincide with hot spots for allelic crossover. To assess whether the identified motif could be responsible for causing NAHR events, we examined the sequence surrounding NAHR hot spots for the six diseases that satisfy the following conditions: (i) rearrangements occur within an autosome or on the X chromosome; (ii) independent, *de novo* events with breakpoints mapping inside homologous genomic repeats occur in different individuals with the disease; and (iii) fine mapping of breakpoints has been performed in multiple cases and reveals clustering within the hot-spot region (see Methods). In all cases, the degenerate motif CCNCCNTNCCNC occurred within the hot-spot region of the appropriate low complexity repeat (**Supplementary Fig. 1** online, $P = 0.00055$; **Supplementary Note**). Examination of secondary, weaker hot spots for Smith-Magenis syndrome and NF1 did not reveal motif presence. The case of X-linked ichthyosis is particularly notable. This recessive skin disorder (incidence of 1 in 5,000 live births) is caused by deletions of the *STS* gene resulting from NAHR between two of four genomic repeats on chromosome X, each of which carries a copy of the *VCX* gene and a tandem repeat of the core 13-mer motif (making this the most dense concentration of the 13-mer in the genome at the megabase scale, **Fig. 3**). Fine mapping of four breakpoints¹² has revealed that all occurred precisely within the motif-rich tandem repeats.

Mutational processes at hypervariable human minisatellites have been examined in depth for eight minisatellites^{15–22}. These loci broadly fall into two classes. In one class, mutations patterns suggest initiation outside the repeat array because there is no correlation between array length and mutation rate and there is (except for the insulin minisatellite) a strong 'polarity' whereby mutation events cluster at one end of the minisatellite^{19,20,22}. In contrast, in the other class most mutations seem to be initiated within the array itself, with a strong correlation between array length and mutation rate and no apparent polarization (**Table 2**). Notably, we found that for every locus in the second class the minisatellite repeat unit contained a region almost perfectly matching the core hot-spot motif (**Table 2**). The presence of part of the motif, CCTCCCT, within CEB1 was previously noted²³. To ask whether this was likely to have occurred by chance, we calculated the motif occurrence in other human tandem repeat sequences of the same repeat length and GC content. We found that a motif match at all four loci is extremely unlikely to occur by chance ($P = 10^{-7}$ via permutation test). In contrast, none of the

minisatellites in the first class showed a match to the motif, consistent with event initiation at flanking hot spots outside the array. One such locus is MS32, where the motif explains the flanking hot spot (**Fig. 1b**). For both minisatellite classes we found a strong local association with elevated recombination rate (**Fig. 4**).

An intriguing association between the hot-spot motif and a recurrent genome rearrangement is the 'common deletion' in mitochondria. This deletion is the most common mitochondrial rearrangement, can result in Pearson's syndrome, CPEO and Kearns Sayre syndrome and also accumulates within cells during normal human aging²⁴. The deletion event is mediated by two 13-bp direct repeats separated by 4,977 bp, which have only one mismatch to the canonical 13-mer motif (**Fig. 3d**). Although mitochondria do not undergo meiosis, the motif might cause deletions by stimulating the formation of double-stranded breaks during mitochondrial replication, as has been shown experimentally in mice²⁵.

Our results provide the first evidence that a substantial fraction of human recombination hot spots share a common mechanism. Does the nature of the motif offer any clues as to the molecular basis for recombination hot spots? A notable feature of the degeneracy both within and beyond the 13-mer core is a threefold periodicity (**Fig. 2a**).

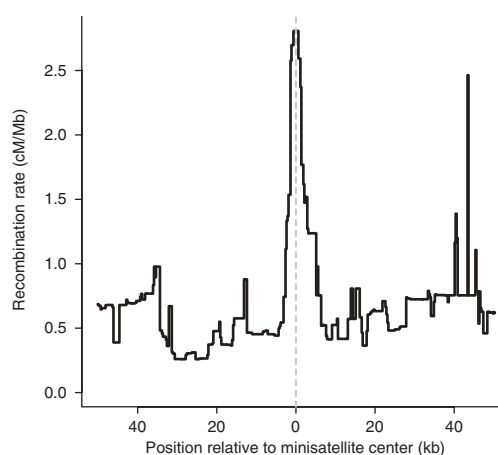


Figure 4 Recombination and hypermutable minisatellites. The plot shows median recombination rate (calculated in non-overlapping windows of 250 bp) around 9 of 10 previously identified hypermutable human minisatellites²⁸, excluding minisatellite CEB1 owing to a lack of typed flanking SNPs in HapMap. The highest median rate corresponds precisely to the minisatellite location. Only minisatellite CEB36 has a recombination rate estimate below the genome-wide average; all others have a marked local increase in recombination rate focused at or near the minisatellite.

This pattern is unlikely to reflect coding sequence because hot spots actively avoid coding regions^{1,26}. Rather, the pattern might reflect direct interaction with flanking DNA of a motif-binding protein. The periodicity might reflect cooperative binding of a protein interacting with 3-bp DNA units, as occurs for RAD51, which promotes DNA strand exchange in humans²⁷. However, the spacing is also reminiscent of the 3-bp binding unit of individual fingers within zinc-finger binding proteins, which can possess long consensus sequences. The identification of factors that interact with the hot-spot motif should provide further insight into the process of human recombination and its evolution.

METHODS

Generating hot spots and matched cold spots. We estimated hot spots from the HapMap Phase II data as previously described^{1,4}. For each of the 34,142 resultant hot spots, we identified an identically sized cold-spot region on the same chromosome where there was no evidence ($P = 1.0$) of excess recombination activity relative to the surrounding DNA. Each cold spot was required to match the paired hot spot in terms of local GC content (to within 10%) and SNP density (to within 10%), and we chose the closest possible cold spot conditional on these constraints. We matched 99.8% of the hot spots in this manner (the remaining 0.2% were excluded from comparisons).

Testing for over- and under-representation of repeat elements in hot spots. Using the locations of all identified repeats (from the RepeatMasker track of the UCSC browser, May 2004 assembly, hg17), we tested each repeat class and family and individual repeats for differences in the extent of overlap with narrowly localized (5 kb or less in size) hot spots and cold spots. For each repeat type, we recorded the number of hot spots (and cold spots) overlapping at least one repeat of the specified type, and compared the two totals via a binomial test. P values obtained were Bonferroni corrected for multiple testing using the number of repeats represented at least 25 times in either hot spots or cold spots (Supplementary Tables 1–3).

Motif testing in repeat backgrounds. We tested for hot spot-associated motifs separately within the following repeat backgrounds with at least 25 copies in hot spots or cold spots and showing $P < 0.01$: THE1B, THE1A, GA-rich and CT-rich (combined), L2, AluY, THE1A-int, MIRb, C-rich and G-rich combined, LTR49, MIR, AluSg, Tigger2a, MER61A, LTR5B, MLT1D, polypurine, LTR1, AluSg1 and (CCA)_n and (TGG)_n combined. We also tested several backgrounds similar to over-represented backgrounds: THE1C, THE1D, Alu, AluJ/FLAM, AluJb, AluJo, AluS, AluSc, AluSg/x, AluSp, AluSp/q, AluSq, AluSq/x, AluSx, AluYa8, AluYb9 and MLT1C, yielding a total of 36 backgrounds to test. For each background, we created a 'hot' set of all occurrences of the repeat overlapping a narrowly defined hot spot. We also created a comparable 'cold' set containing occurrences of the repeat not overlapping any hot spot, with the maximal number of 'cold' repeats possible added from each chromosome in turn so as to match the size distribution of the 'hot' repeats (fraction in successive 10% size bins). On each background, we tested for differences between hot spots and cold spots for every possible nondegenerate DNA motif of length 5–9 bp via Fisher's exact test. Within each motif size, we then applied Bonferroni correction for the number of motifs, and recorded motifs showing $P < 0.05$. Over-represented motifs of length 7 or more were mapped against the consensus sequence for each element. This typically identified a series of overlapping segments within each element, the union of which is shown in Table 1. We tested the X chromosome separately from the autosomes.

Identifying hot-spot motifs in nonrepeat DNA. We identified all autosomal occurrences of CCTCCCT in nonrepeat DNA surrounded by at least 50 bp of nonrepeat DNA on each side and thinned occurrences to give a minimum separation of 100 bp. We compared base composition at given positions relative to the motif with a χ^2 test (3 degrees of freedom), to produce the P values shown in Figure 2a. This approach showed an enrichment of CCAC 2 bp downstream from the CCTCCCT motif, as observed in repeat elements. Testing this 4-mer motif at the same location revealed an even stronger signal for

enrichment of CCAC in hot-spot cases via a χ^2 test (OR = 2.2, $P < 10^{-30}$). This enrichment was the strongest for any 4-bp motif at any site within the 50 bp surrounds of the CCTCCCT motif (results not shown).

Motif degeneracy and estimating the proportion of hot spots explained by the motif. Details of these analyses are available in the Supplementary Methods.

Analysis of hypermutable minisatellites. We considered recombination rate estimates surrounding ten of the most mutable human minisatellites known (identified in ref. 28). One of these, CEB1, was excluded because of low SNP density in HapMap. Eight human minisatellites have previously been examined using minisatellite variant repeat mapping by PCR^{15–22} (MVR-PCR): B6.7, CEB1, MS1, MS32, MS205 and the insulin minisatellite, where male germ-line mutations have been studied in sperm, and MS31 and D7S22, where pedigree mutants have been studied. We observed an occurrence of the hot-spot motif (mismatching one base in each case) within each of the four minisatellites where previous minisatellite variant repeat mapping suggested initiation of events within the repeat array. To test whether this was evidence of enrichment of the hot-spot motif, we resampled 10^7 sets of four repeat sequences from the collection of autosomal tandem repeats (with at least 3 repeats and at least 88% homology to the repeat consensus, as seen in the set studied here—all eight MVR-PCR studied minisatellites were included in this collection). The tandem repeats were downloaded from the Simple Repeats track of the UCSC genome browser. Each resampled repeat was chosen to match the GC content and repeat unit size (both within 5%) of the corresponding MVR-PCR-studied minisatellite. We counted how frequently all four resampled repeats matched the motif so closely, yielding $P = 1 \times 10^{-7}$ for the observed data. Testing via permutation of the consensus repeat unit within each minisatellite gave a similar result ($P = 2 \times 10^{-7}$).

Examination of sequence surrounds at five NAHR hot spots. We used previous literature^{9–14,29,30} to identify six major disease-related nonallelic homologous recombination hot spots (CMT1A, NF1, Sotos syndrome, Smith-Magenis syndrome (SMS), Williams-Beuren syndrome and X-linked ichthyosis). Each hot spot occurs within some homologous pair of low-copy repeats. The sequences were obtained for each respective low-copy repeat, as defined using the Segmental Duplications track of the UCSC genome browser. For each region we identified all copies of the degenerate hot-spot motif CCNCNTNNCCNC within the first low-copy repeat (an arbitrary choice) of the pair involved in NAHR events. Finally, we curated motif occurrences, defining occurrences within Alu elements but outside of AluY, AluSc or AluSg elements as likely to be inactive in promoting recombination. This annotation was used to produce Supplementary Figure 1. To test whether our observation of degenerate motifs within all five hot spots was expected by chance, separately within each LCR we resampled 10,000 hot-spot positions (chosen so that all of the hot spot lay within the LCR), and calculated the proportion of cases the hot spot contained a putatively active copy of the motif. Multiplying these P values together yielded $P = 0.00055$. For further discussion of the motifs found within each hot spot, see the Supplementary Note.

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

ACKNOWLEDGMENTS

We thank J. Lupski for help with compiling information on NAHR hot spots and G. Coop for discussion. We thank the Engineering and Physical Sciences Research Council (EPSRC), the Wolfson Foundation, the EU and the Wellcome Trust for financial support.

AUTHOR CONTRIBUTIONS

S.M., P.D. and G.M. designed the study; S.M., A.A. and C.F. performed the analyses; S.M. and G.M. wrote the paper.

Published online at <http://www.nature.com/naturegenetics/>

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>

- Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324 (2005).

2. Jeffreys, A.J. & Neumann, R. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat. Genet.* **31**, 267–271 (2002).
3. Jeffreys, A.J. & Neumann, R. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum. Mol. Genet.* **14**, 2277–2287 (2005).
4. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
5. Baudat, F. & de Massy, B. Cis- and trans-acting elements regulate the mouse Psmb9 meiotic recombination hotspot. *PLoS Genet.* **3**, e100 (2007).
6. Neumann, R. & Jeffreys, A.J. Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation. *Hum. Mol. Genet.* **15**, 1401–1411 (2006).
7. Jeffreys, A.J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29**, 217–222 (2001).
8. Jeffreys, A.J., Murray, J. & Neumann, R. High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol. Cell* **2**, 267–273 (1998).
9. Bi, W. *et al.* 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 (2003).
10. Reiter, L.T. *et al.* A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. *Nat. Genet.* **12**, 288–297 (1996).
11. Lopez-Correa, C. *et al.* Recombination hotspot in NF1 microdeletion patients. *Hum. Mol. Genet.* **10**, 1387–1392 (2001).
12. Van Esch, H. *et al.* Deletion of VCX-A due to NAHR plays a major role in the occurrence of mental retardation in patients with X-linked ichthyosis. *Hum. Mol. Genet.* **14**, 1795–1803 (2005).
13. Raedt, T.D. *et al.* Conservation of hotspots for recombination in low-copy repeats associated with the NF1 microdeletion. *Nat. Genet.* **38**, 1419–1423 (2006).
14. Lindsay, S.J., Khajavi, M., Lupski, J.R. & Hurles, M.E. A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. *Am. J. Hum. Genet.* **79**, 890–902 (2006).
15. Andreassen, R. & Olaisen, B. De novo mutations and allelic diversity at minisatellite locus D7S22 investigated by allele-specific four-state MVR-PCR analysis. *Hum. Mol. Genet.* **7**, 2113–2120 (1998).
16. Berg, I., Neumann, R., Cederberg, H., Rannug, U. & Jeffreys, A.J. Two modes of germline instability at human minisatellite MS1 (locus D1S7): complex rearrangements and paradoxical hyperdeletion. *Am. J. Hum. Genet.* **72**, 1436–1447 (2003).
17. Buard, J., Bourdet, A., Yardley, J., Dubrova, Y. & Jeffreys, A.J. Influences of array size and homogeneity on minisatellite mutation. *EMBO J.* **17**, 3495–3502 (1998).
18. Buard, J., Shone, A.C. & Jeffreys, A.J. Meiotic recombination and flanking marker exchange at the highly unstable human minisatellite CEB1 (D2S90). *Am. J. Hum. Genet.* **67**, 333–344 (2000).
19. Jeffreys, A.J., Neil, D.L. & Neumann, R. Repeat instability at human minisatellites arising from meiotic recombination. *EMBO J.* **17**, 4147–4157 (1998).
20. May, C.A., Jeffreys, A.J. & Armour, J.A. Mutation rate heterogeneity and the generation of allele diversity at the human minisatellite MS205 (D16S309). *Hum. Mol. Genet.* **5**, 1823–1833 (1996).
21. Tamaki, K., May, C.A., Dubrova, Y.E. & Jeffreys, A.J. Extremely complex repeat shuffling during germline mutation at human minisatellite B6.7. *Hum. Mol. Genet.* **8**, 879–888 (1999).
22. Stead, J.D. & Jeffreys, A.J. Allele diversity and germline mutation at the insulin minisatellite. *Hum. Mol. Genet.* **9**, 713–723 (2000).
23. Lopes, J., Ribeyre, C. & Nicolas, A. Complex minisatellite rearrangements generated in the total or partial absence of Rad27/hFEN1 activity occur in a single generation and are Rad51 and Rad52 dependent. *Mol. Cell. Biol.* **26**, 6675–6689 (2006).
24. Linnane, A.W. *et al.* Mitochondrial gene mutation: the ageing process and degenerative diseases. *Biochem. Int.* **22**, 1067–1076 (1990).
25. Srivastava, S. & Moraes, C.T. Double-strand breaks of mouse muscle mtDNA promote large deletions similar to multiple mtDNA deletions in humans. *Hum. Mol. Genet.* **14**, 893–902 (2005).
26. Coop, G., Wen, X., Ober, C., Pritchard, J.K. & Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* **319**, 1395–1398 (2008).
27. Baumann, P., Benson, F.E. & West, S.C. Human Rad51 protein promotes ATP-dependent homologous pairing and strand transfer reactions in vitro. *Cell* **87**, 757–766 (1996).
28. Vergnaud, G. & Denoeud, F. Minisatellites: mutability and genome architecture. *Genome Res.* **10**, 899–907 (2000).
29. Bayes, M., Magano, L.F., Rivera, N., Flores, R. & Perez Jurado, L.A. Mutational mechanisms of Williams-Beuren syndrome deletions. *Am. J. Hum. Genet.* **73**, 131–151 (2003).
30. Visser, R. *et al.* 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 (2005).