Science. Author manuscript; available in PMC 2013 November 15.

Published in final edited form as:
Science. 2010 February 12; 327(5967): . doi:10.1126/science.1182363.

Drive Against Hotspot Motifs in Primates Implicates the *PRDM9* gene in Meiotic Recombination:

We provide evidence that a rapidly evolving gene is involved in determining recombination hotspot locations in humans

Simon Myers[1,2,*], **Rory Bowden**[1,2,*], **Afidalina Tumian**[1], **Ronald E. Bontrop**[3], **Colin Freeman**[2], **Tammie S. MacFie**[4], **Gil McVean**[1,2,#], and **Peter Donnelly**[2,1,#]

[1]Department of Statistics, Oxford University, 1 South Parks Road, Oxford OX1 3TG, UK [2]Wellcome Trust Centre for Human Genetics, Oxford University, Roosevelt Drive, Oxford OX3 7BN, UK [3]Department of Comparative Genetics & Refinement, Biomedical Primate Research Center, Rijswijk, Netherlands [4]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK [Present address: Institute of Cell & Molecular Science, Barts and The London School of Medicine and Dentistry, 4 Newark Street, London E1 2AT]

Abstract

Although present in both humans and chimpanzees, recombination hotspots, at which meiotic cross-over events cluster, differ markedly in their genomic location between the species. We report that a 13-bp sequence motif previously associated with the activity of 40% of human hotspots does not function in chimpanzee, and is being removed by self-destructive drive in the human lineage. Multiple lines of evidence suggest that the rapidly evolving zinc-finger protein, PRDM9 binds to this motif and that sequence changes in the protein may be responsible for hotspot differences between species. The involvement of PRDM9, which causes Histone H3 Lysine 4 trimethylation, implicates a common mechanism for recombination hotspots in eukaryotes but raises questions about what forces have driven such rapid change.

In humans and most other eukaryotes, meiotic cross-over events typically cluster within narrow regions termed hotspots (1-5). Previously (6), we identified a degenerate 13-bp motif, CCNCCNTNNCCNC, overrepresented in human hotspots. Both LD-based analysis (6) and sperm typing at currently active hotspots (7) implicated this motif in the activity of 40% of hotspots.

Remarkably, despite nearly 99% identity at aligned bases, humans and chimpanzees show little, if any, sharing of hotspot locations (4, 5), although it has remained undetermined whether the recently-identified hotspot motif is also active in the chimpanzee. To resolve this question, we collected chimpanzee genetic variation data at 22 loci where there is both an inferred hotspot at the orthologous location in humans, and human-chimpanzee sequence conservation of the 13-mer: 16 motifs within THE1 elements and 6 within L2 elements, chosen for their high activity of a particular “core” version of the motif in humans (Fig. S1). We used the statistical software LDhat to estimate recombination rates separately in each region in different populations of both species (8). For humans, we used the HapMap Phase

Correspondence should be addressed to: myers@stats.ox.ac.uk.
*Both first authors contibuted equally.
#Both last authors contributed equally.

Europe PMC Funders Author Manuscripts

II data. For chimpanzees, we genotyped 36 Western, 20 Central and 17 Vellorosus chimpanzees at a total of 694 chimpanzee SNPs, an average of 31.5 per region.

Because these regions are inferred human hotspots, the average estimated recombination rate surrounding the motif in humans showed a strong peak for both L2 and THE1 elements (Fig. 1A). In contrast, chimpanzees showed no evidence of increased recombination rates for either background. In Western chimpanzees, the THE1 estimated recombination rate around the motif was similar to the regional average, while a weak peak in mean rate for the L2 elements was produced solely by a single potential hotspot in one of the six regions (Fig. 1B). Results for the other chimpanzee subspecies were less informative (8, Fig. S2) but did not reveal a different pattern. To ensure that unknown haplotypic phase, smaller sample size, less dense data, and SNP ascertainment in chimpanzee had not compromised power to detect hotspots, we repeatedly sampled from the CEU HapMap to produce human datasets comparable to those from chimpanzee in terms of these features (8). Importantly, we conditioned only on the presence of the 13-mer in THE1 and L2 elements and not the presence of a hotspot. This bootstrap technique revealed that the differences between humans and chimpanzee rates cannot be explained by differences in power ($p=0.00052$), though the signal was only significant for THE1 elements when analyzed separately ($p=0.00012$; Fig. S3). These results provide evidence that the 13-mer motif does not recruit hotspots in chimpanzees, implying changes in recombination machinery between humans and chimpanzees. The existence of factors capable of such changes in recombination genome-wide has been demonstrated in C. elegans (9) and notably by the mapping in mice of a trans-acting factor responsible for differences in hotspot location among inbred strain crosses (10, 11).

A separate process, predicted to cause rapid evolution of individual hotspots, is the self-destructive drive inherent in double-strand break (DSB) formation, known as biased gene conversion (BGC) (12). Mutations reducing DSB formation in cis at recombination hotspots are preferentially transmitted as a consequence of repair of DSBs initiated on the other, more recombinogenic, strand in heterozygotes and are thus favored in a manner mimicking natural selection (13). This phenomenon could lead to rapid hotspot loss (14, 15). Direct evidence from sperm typing (16) has shown BGC at one polymorphic point mutation disrupting an occurrence of the 13-bp motif. More generally, BGC is predicted to eliminate copies of any recombination promoting motif from the genome. The species-specific recombination activity of the 13-bp human hotspot motif suggests that losses of this motif should have occurred preferentially on the human lineage, rather than that leading to chimpanzees.

To examine the evidence for BGC driven motif loss, we therefore characterized rates and patterns of molecular evolution for the degenerate 13-mer and the “core” version of the motif, on specific backgrounds: THE1 elements, L2 elements, AluY/Sc/Sg elements (degenerate motif only), other repeats, and unique non-repeat DNA (Table 1). We found a consistent substitution pattern imbalance, with chimpanzees having more copies of the motif than humans (empirical $p=0.003$ for the most active form, with three of four independent backgrounds showing $p<0.05$; $p=0.002$ for the degenerate 13-mer motif, with $p<0.05$ for three of five individual backgrounds (8)). As predicted by theoretical considerations of BGC (14, 15, supporting online text, Table S1) the magnitude of the imbalance was strongest for cases where the motif has greatest activity. To assess whether motifs have been gained in chimpanzee or lost in humans we used the published Macaque (17), and draft Orangutan (18) genome sequences to infer ancestral sequence. For THE1 elements, L2 elements, and non-repeat DNA, we observe an excess of human losses of the most active motif relative to chimpanzee ($p<0.05$ in each case, Fig. 1C, Table S2) and similar results for the degenerate 13-mer motif (Table S3). The effect strength again correlates with hotspot activity. In contrast, there are no significant differences between species in motif gains ($p>0.3$). Alu

elements were not analyzed because of a high rate of uncertainty in inferring the ancestral base.

To ask whether motif activity has been lost on the chimpanzee lineage or gained on the human lineage we compared our observations to a population-genetics model (14, 15, supporting online text). On the human lineage, approximately 16% of motifs on the THE1 and 8% on the L2 background have been lost in humans since human-chimpanzee divergence (Fig. 1C). If the motif had been active since the time of speciation, we predict that 46-56% and 31-38% of motifs in THE1 and L2 elements respectively should have been lost. The observed patterns of motif evolution in humans are instead consistent with a recent (1-2 million years ago) activation of the 13-bp motif on the human lineage, rather than inactivation on the chimpanzee lineage.

We next investigated the function of the 13-mer motif. Previously, we suggested that the human hotspot motif was likely bound by a zinc finger protein with at least 12 zinc fingers, on the basis of an extended 30-40bp region of weaker sequence specificity containing the motif, and a 3bp periodicity of influential bases (6). We therefore set out to identify candidates for such a protein using a computational algorithm that predicts DNA binding specificity for C2H2 zinc-finger proteins (19). Among the 691 identified human C2H2 zinc-finger proteins the 13-mer motif was present within the predicted binding sequence of five (Fig. S4). Binding specificity was then further explored in silico by comparing predicted motif degeneracy for each candidate (inferred by calculating the relative binding score for every 1-bp mutation relative to the consensus) to empirical degeneracy patterns in the 13-bp motif (Fig. 2C). Predictions for one of the candidates, PRDM9, exactly matched the observed degeneracy at positions 3,6,8,9 and 12 within the 13-bp motif (Fig. 2B) and lack of degeneracy at the other 8 positions. Predictions for the other four candidates showed features inconsistent with the observed degeneracy (Fig. S4). Intriguingly, the predicted binding sequence for PRDM9 also contains an exact match on the opposite strand for an 8-bp region of the extended motif, upstream of the 13-bp degenerate motif, perhaps suggesting that PRDM9 zinc fingers might contact both DNA strands. Finally, the number (thirteen) of zinc fingers in this protein, the positioning of the match to the 13-bp motif within the longer predicted binding sequence, and strong influence of this 13-bp region on specificity, all match our previous predictions (6).

The lack of activity of the 13-bp motif in chimpanzees demonstrated above suggests that in addition to having the predicted binding specificity, any motif-binding protein candidate should also show differences between humans and chimpanzees. For four of the five candidates the predicted DNA contacting amino acids within the zinc fingers are identical between human and chimpanzee. Chimpanzee PRDM9, however, has a dramatically different predicted binding sequence (Fig. S5). Although PRDM9 has multiple zinc fingers in both species (12 and 13 respectively), the DNA-contacting residues −1, 2, 3 and 6 are only shared between species in the first finger (Fig. 2C). Such rapid evolution is exceptional. Comparing these residues among all 544 C2H2-containing zinc-finger protein human-chimpanzee orthologue pairs, PRDM9 is the most diverged (p=0.0018). The PRDM9 sequences in five additional mammals (Elephant, Mouse, Rat, Macaque and Orangutan) exhibit rapid evolution, variation in zinc-finger number (between 8 and 12) and patterns of substitution suggestive of complex repeat shuffling (Fig 2C; 20).

Multiple lines of evidence point to a role for the orthologous mouse gene, *Prdm9*, in recombination. *Prdm9* lies within a 5.1Mb region containing a locus that influences genome-wide hotspot locations (10, 11), and is exclusively expressed during meiotic prophase, with *Prdm9* knockouts showing infertility and failure to properly repair DSBs (21). Mouse PRDM9 trimethylates lysine 4 of histone H3 (H3K4me3) (21), an epigenetic mark

specifically enriched on mouse chromatids carrying recombination initiation sites within the mouse hotspot *Psmb9* (22). In yeast, mutation of the sole gene, *Set1*, encoding H3K4me3 reduces cross-over activity at 84% of hotspots (23). The lack of well-defined target sequence specificity of Set1 (which is not a zinc finger protein) may indicate why no dominant hotspot motif has been identified in yeast. Intriguingly, *Prdm9* is also the only species-incompatibility gene yet identified in mouse (24), with differences among 9 PRDM9 zinc fingers between mouse strains potentially playing a causal role in male sterility.

Elsewhere in this issue, Baudat and colleagues demonstrate that variation in *PRDM9* among humans correlates with variability in genome-wide hotspot usage, and that PRDM9 binds the 13-bp motif in a sequence specific manner in vitro. The findings of both studies imply that PRDM9 determines human hotspot locations, with PRDM9 evolution explaining lack of hotspot conservation in other species. Exactly how PRDM9 functions, for example through altering transcription of DSB repair genes or directly recruiting DSB repair proteins, remains unknown. These findings also raise the question of why such an important gene is evolving so rapidly. The DNA sequence of the zinc-finger array of PRDM9 constitutes a coding minisatellite, suggesting a high intrinsic mutation rate resulting from repeat instability. However, patterns of evolution within the zinc finger array, notably the clustering and coordination of changes at sites that interact with DNA bases, strongly suggest positive selection on binding specificity (20). Selection could possibly arise from the gradual degradation of hotspots through BGC leading to a loss in fitness either through promotion of deleterious alleles within hotspots (15) or through having insufficient cross-over events to support proper disjunction(14, 15). Alternatively, the rapid evolution of *PRDM9* could be indicative of genetic conflict, such as meiotic drive or conflict involving mobile elements (25, 26). While there is no direct evidence for this, it is intriguing to note that mouse *Prdm9* lies within one of the inversions characterizing the meiotic-drive *t*-complex (27).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank N. Mundy for advice and provision of chimpanzee samples, C. Mitchell and E. Nerrienet for assisting in chimpanzee sample collection. We thank D. Falush and G. Coop for helpful conversations. Part of the work was completed while SM was a fellow at the Broad Institute. We would like to acknowledge funding from the Leverhulme Trust (GM), the Royal Society (PD), and the Wellcome Trust (SM, CF, GM, PD). The chimpanzee genotype data we generated is freely downloadable from www.stats.ox.ac.uk/~myers/chimpanzeehotspotstudy.html.

References and Notes

1. Gerton JL, et al. Proc Natl Acad Sci U S A. 2000; 97:11383. [PubMed: 11027339]
2. McVean GA, et al. Science. 2004; 304:581. [PubMed: 15105499]
3. Paigen K, et al. PLoS Genet. 2008; 4:e1000119. [PubMed: 18617997]
4. Ptak SE, et al. Nat Genet. 2005; 37:429. [PubMed: 15723063]
5. Winckler W, et al. Science. 2005; 308:107. [PubMed: 15705809]
6. Myers S, Freeman C, Auton A, Donnelly P, McVean G. Nat Genet. 2008; 40:1124. [PubMed: 19165926]
7. Webb AJ, Berg IL, Jeffreys A. Proc Natl Acad Sci U S A. 2008; 105:10471. [PubMed: 18650392]
8. Materials and methods are available in the supporting material and at Science Online.
9. Mets DG, Meyer BJ. Cell. 2009; 139:73. [PubMed: 19781752]
10. Parvanov ED, Ng SH, Petkov PM, Paigen K. PLoS Biol. 2009; 7:e36. [PubMed: 19226189]
11. Grey C, Baudat F, de Massy B. PLoS Biol. 2009; 7:e35. [PubMed: 19226188]

12. Nicolas A, Treco D, Schultes NP, Szostak JW. Nature. 1989; 338:35. [PubMed: 2537472]
13. Nagylaki T. Proc Natl Acad Sci U S A. 1983; 80:6278. [PubMed: 6578508]
14. Coop G, Myers SR. PLoS Genet. 2007; 3:e35. [PubMed: 17352536]
15. Boulton A, Myers RS, Redfield RJ. Proc Natl Acad Sci U S A. 1997; 94:8058. [PubMed: 9223314]
16. Jeffreys AJ, Neumann R. Nat Genet. 2002; 31:267. [PubMed: 12089523]
17. Gibbs RA, et al. Science. 2007; 316:222. [PubMed: 17431167]
18. 2007. http://genome.ucsc.edu/cgi-bin/hgGateway?clade=mammal&org=Orangutan&db=0
19. Persikov AV, Osada R, Singh M. Bioinformatics. 2009; 25:22. [PubMed: 19008249]
20. Oliver PL, et al. PLoS Genet. 2009; 5:e1000753. [PubMed: 19997497]
21. Hayashi K, Yoshida K, Matsui Y. Nature. 2005; 438:374. [PubMed: 16292313]
22. Buard J, Barthes P, Grey C, de Massy B. Embo J. 2009; 28:2616. [PubMed: 19644444]
23. Borde V, et al. EMBO J. 2009; 28:99. [PubMed: 19078966]
24. Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. Science. 2009; 323:373. [PubMed: 19074312]
25. Presgraves DC. Bioessays. 2007; 29:386. [PubMed: 17373698]
26. Wyckoff GJ, Wang W, Wu CI. Nature. 2000; 403:304. [PubMed: 10659848]
27. Schimenti J. Trends Genet. 2000; 16:240. [PubMed: 10827448]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

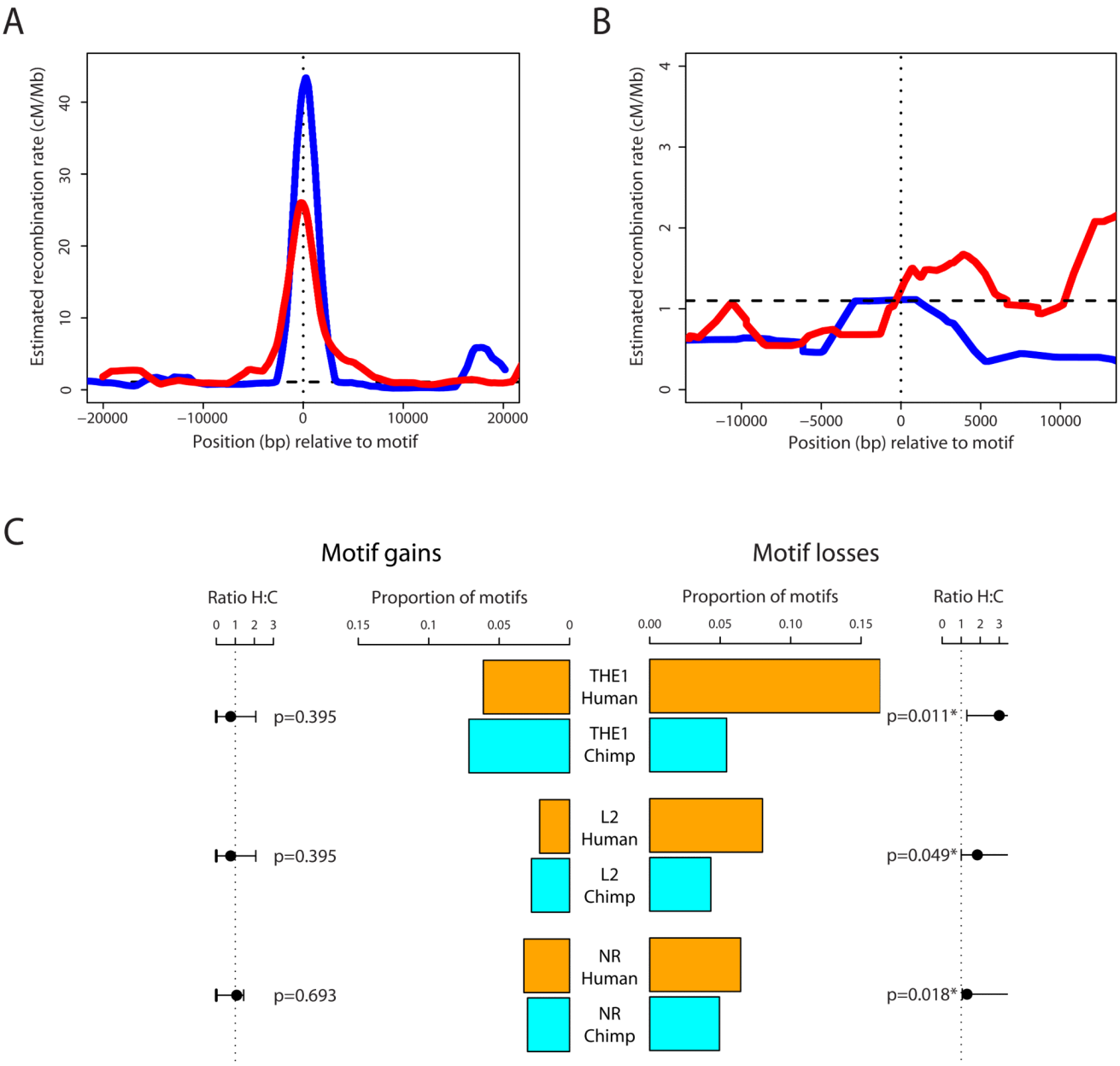

Figure 1.

Recombination rates and patterns of motif gain and loss in human and chimpanzee. For additional details, see (8). **A** Estimated HapMap Phase II recombination rate across the 40kb surrounding 16 human THE1 elements (red line) and 6 L2 elements (blue line) orthologous to the 22 regions analyzed in chimpanzee, and each containing a conserved exact match to the 13-bp core motif. Rates are smoothed using a 2kb sliding window slid in 50bp increments, averaged across elements. Horizontal dashed line: the human average recombination rate of 1.1cM/Mb. Vertical dotted line: the centre of the repeat. **B** Average estimated recombination rate for the western chimpanzee data across around the 16 THE1 elements (red line) and six L2 elements (blue line) containing the 13-bp core motif. Other details as for A. **C** Numbers of core motif gains (left hand bars) versus losses (right bars), inferred using macaque and orang-utan outgroup information (8), in humans (orange bars)

and chimpanzees (cyan bars) on three backgrounds; THE1, L2 and non-repeat (NR). For each background, gains are shown as a fraction of motifs currently present in each species, losses as a fraction of motifs inferred in the human-chimpanzee ancestor. The intervals flanking the plot on each side show exact 1-sided 95% confidence intervals and associated p-values for testing equality of gain/loss rate between the species (8).

Europe PMC Funders Author Manuscripts

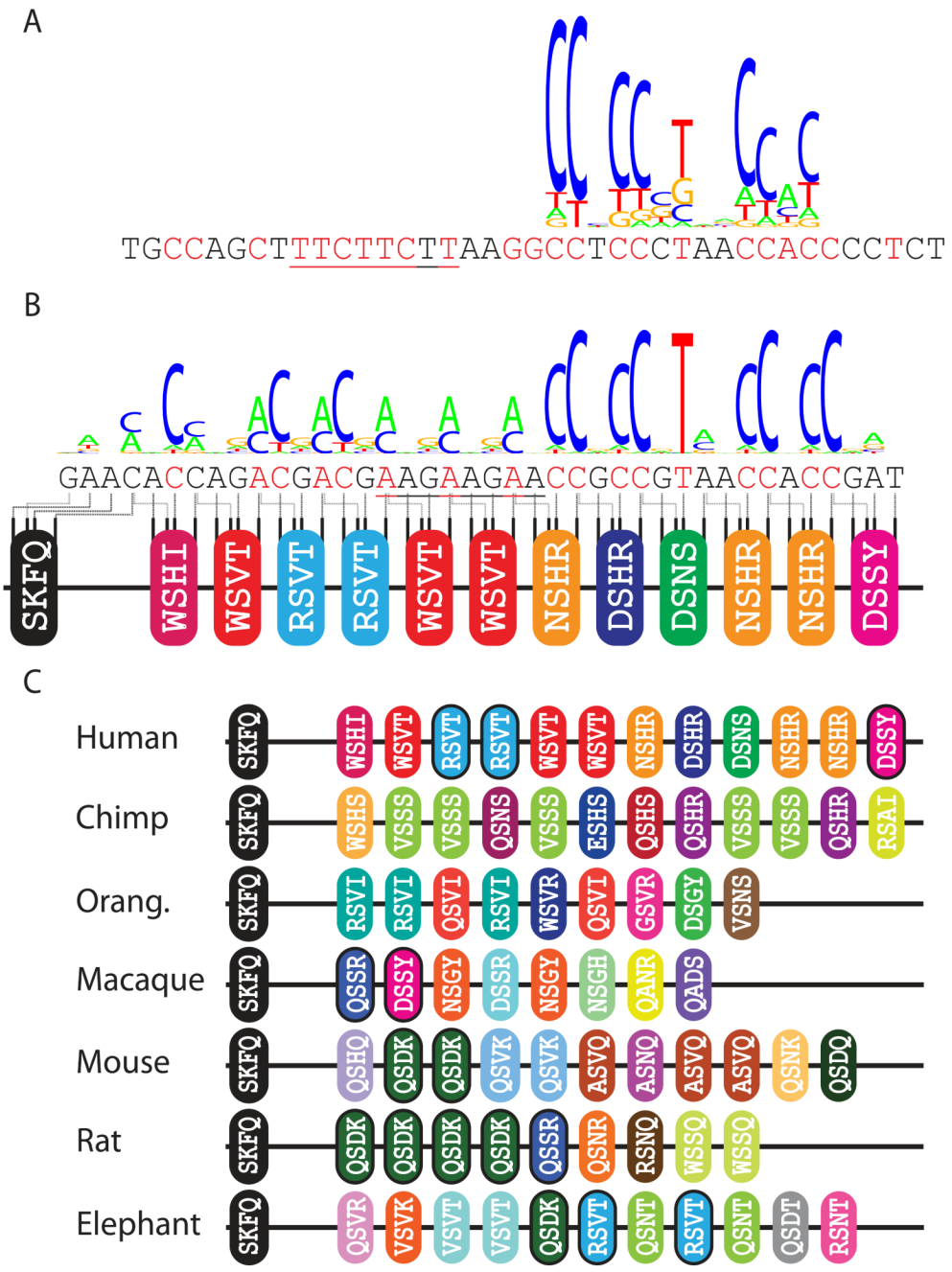

Figure 2.

A Previously estimated degeneracy of the 13-bp hotspot motif (6) (logo plot; relative letter height proportional to estimated probability of hotspot activity, total letter height determined by degree of base specificity) as well as an extended ~39-bp motif (text below logo, with influential positions ($p<0.01$) shown in red). **B** In silico prediction of the binding consensus for PRDM9, aligned with the 13-mer, with more influential positions shown in red. Underlined in both A and B is an additional 8-bp matching sequence. The logo shows predicted degeneracy within this consensus (8). Below the text is the sequence of four predicted DNA-contacting amino acids for the 13 successive human PRDM9 zinc fingers (1 oval per finger, differing colors for differing fingers, separated finger is gapped N-terminal

from others), and their predicted base contacts within the motif. **C** Sequence of four predicted DNA-contacting amino acids for the PRDM9 zinc fingers in 7 mammalian species, presented as in B. Distinct fingers given different colors; fingers present in at least two species have black border.

Europe PMC Funders Author Manuscripts

Table 1

Motif imbalance between human and chimpanzee. For the core motif and the degenerate motif, we analyzed cases where the motif occurs in exactly one of human and chimpanzee. Results are shown for the full set of non-shared motifs and stratified into five backgrounds which differ in average human recombination activity. Significance levels are calculated in two ways: p-values for ratios are based on a 1-sided exact binomial test of fewer human-only cases, as the motif is known to be active in humans. Empirical p-values are 1-sided and obtained by comparisons of counts for the core or degenerate motif to counts observed for motifs of the same length and GC content on the same backgrounds (8).

Sequence background	Core motif CCTCCCTNNCCAC				Degenerate motif CCNCCNTNNCCNC			
	Human only	Chimp only	Ratio (p-value)	Empirical p-value	Human only	Chimp only	Ratio (p-value)	Empirical p-value
All	425	515	1.21 (0.0018**)	0.0033**	19448	20245	1.04 (3.2e-05**)	0.0020**
THE1	20	39	1.95 (0.0092**)	0.0050**	50	76	1.52 (0.0128*)	0.0093**
L2	30	47	1.57 (0.0338*)	0.0307*	432	496	1.15 (0.0193*)	0.0219*
AluY,Sc,Sg	-	-	-	-	3642	3924	1.08 (0.0006**)	0.1119
Other repeats	99	131	1.32 (0.0204*)	0.0346*	10126	10254	1.01 (0.1868)	0.4373
Non-repeats	276	298	1.08 (0.2135)	0.2206	5198	5495	1.06 (0.0021**)	0.0215*