Supporting Online Material

Materials and methods

Selection and typing of chimpanzee SNPs

From a list of human recombination hotspots (*S1*) we extracted sequence regions containing THE1 and L2 elements with the extended motif CCTCCCTNNCCAC, and aligned them to the chimpanzee draft genome (pantro2, obtained from the UCSC website: http://genome.ucsc.edu/) in order to identify unique regions orthologously conserved in chimpanzees. We randomly selected a pool of 127 such chimpanzee elements in which the CCTCCCTNNCCAC motif was identically conserved (including the bases corresponding to the two degenerate sites), and identified SNPs from the flanking sequences using dbSNP. We used 100bp around each SNP to filter for uniqueness of the flanking sequence. We prioritized regions for genotyping using the Illumina GoldenGate design service (Illumina Inc., San Diego) to exclude SNPs with low probability of successful typing or that were unlikely to be typeable in the same panel as other included SNPs, favoring regions with the largest number of high-probability SNPs (estimated $P > 0.7$) near to, and evenly divided on, either side of the central element. We compiled a GoldenGate panel of 768 SNPs, including 10 SNPs for use as genotyping controls (from Winkler et al. (2005)) plus the 'best' THE1 and L2 motifs; these comprised 16 THE1 regions and 6 L2 regions, each with at least 15 SNPs on each side (within 40 kbp) of the motif showing $P>0.7$.

The SNP panel was used to genotype 77 chimpanzees: 36 Pan verus (western), 20 Pan troglodytes (central) and 17 Pan vellerosus chimpanzees, plus 4 that were subsequently shown to be inter-subspecific hybrids) using the Illumina GoldenGate platform (Illumina). Genotypes were called using Illumina's BeadStudio software (Illumina), and each cluster plot was checked manually. 64 SNPs were excluded because they exhibited aberrant clustering patterns, or showed departures from Hardy-Weinberg equilibrium within one or more chimpanzee populations ($p<1/700$ in any population, or $p<5/700$ in one population and $p<0.05$ in a second population), resulting in a final dataset including genotypes for 694 SNPs.

Comparing recombination between humans and chimpanzees at THE1 and L2 motif locations

To estimate recombination rates around the 16 THE1 and 6 L2 regions, we analyzed each region for each population separately using the statistical software package LDhat (*S2*), using parameter

values matching those of previous studies (*S1*,*S3*,*S4*). This software has been demonstrated to give highly similar results to alternative approaches in several different human and chimpanzee datasets (*S1*,*S4*,*S5*). LDhat produces a mean estimated recombination rate between each pair of SNPs in a dataset, using a Bayesian reversible jump MCMC approach penalizing rate changes. To produce the human and chimpanzee rates shown in Figs. 1A, 1B, and S1, S2, rates are rescaled to give units in cM/Mb based on the formula $\rho=4N_e r$, where ρ is the recombination rate per Mb estimated from the population genetic analysis via LDhat, r is the underlying rate in cM/Mb, and N_e is the effective population size. For fig. S1, we averaged recombination rates estimated from the HapMap PhaseII data for CEU individuals, at regions surrounding 216 THE1 elements and 556 L2 elements in the genome containing an exact match to the motif CCTCCCTNNCCAC, and used Ne=10,000 for the human CEU population (*S1*,*S5*). Rates shown are average estimated cM/Mb values across these regions, aligned so that position 0 is the centre of the THE1 or L2 repeat in each case. The plot shows rates for a sliding window of width 2kb, slid in 50bp increments along the region. To produce Fig. 1A, we used an identical approach, but only averaged recombination rates across the 16 THE1 surrounding regions syntenic to those for which we gathered data in chimpanzee, and separately for the 6 syntenic L2 surrounding regions. For Fig. 1B, we averaged recombination rates for western chimpanzee THE1 and L2 motif-containing regions (16 THE1 regions and 6 L2 regions) in the same way, using an estimated N_e=9,100 for western chimpanzees (*S6*). For fig. S2, we used N_e=36,000 (central chimpanzees) and N_e=18,500 (vellerosus). For these two subspecies, effective population size estimates vary or are unavailable, and so we chose our rescaling to match the total estimated cM distance across the 22 regions in the three subspecies.

We note that the informativeness of our data differs among subspecies, due to differing sample sizes and the fact that chimpanzee SNP ascertainment has been performed mainly in western chimpanzees, and in particular the majority of chimpanzee SNPs (60% of those we typed in our data) segregate in Clint, the chimpanzee sequenced by the Chimpanzee Sequencing Project (*S7*). The western chimpanzee data are clearly the most informative, on the basis of both sample size (36 chimpanzees vs. 20 and 17 for the central and vellerosus populations respectively) and the fraction of SNPs in our data segregating within each population (83% of SNPs, vs 53% and 48%).

We focus on examining power in this population, though our rate estimates for the other populations are shown in fig. S2. To evaluate our ability to identify an elevation in chimpanzee

recombination activity in motif-containing elements similar to that in humans, we used a resampling based approach. We repeatedly sampled human data matching key features of the chimpanzee data (number of regions considered, presence of a motif within a THE1 or L2 repeat, sample size, local SNP density, ascertainment, and procedure to estimate recombination rates) in order to identify whether the complete lack of elevation in recombination rate seen in chimpanzee THE1 regions could be explained by reasons of power alone. We based our human data on subsampling the CEU data from HapMap Phase II (CEU data), because the CEU have an estimated effective population size N_e of around 10,000 (*S1*), close to the recent estimated western chimpanzee N_e of 9,100 (*S6*).

We describe our procedure to examine our power to identify an elevated rate surrounding THE1 elements (an exactly corresponding procedure was used for the L2 repeats). Our chimp dataset consists of LD information for each of 16 regions, each surrounding a THE1 repeat containing the 13-mer motif. In our sampling scheme, a single "sample" means generating a random human dataset of variation information for 16 different THE1 repeat-surrounding regions, and then using this dataset and LDhat to estimate recombination rates around these 16 THE1 repeats. 10,000,000 such samples were generated, and the resulting 10^7 estimated recombination rate profiles were compared to those for the actual chimpanzee data.

For each sampled dataset, careful matching was performed so that the human data closely mirrored the actual chimpanzee data. First, we randomly chose 16 regions of the human genome across which to analyze recombination patterns. Each of the 16 regions chosen was required to contain a distinct THE1 repeat, containing an exact match to the 13-mer motif CCTCCCTNNCCAC, chosen among all 216 such possibilities (some of which do not correspond to known human hotspots). Because we had no prior knowledge, before we performed our chimpanzee experiment, whether the 16 chimpanzee regions would correspond to recombination hotspots, we do NOT use any knowledge of local human recombination rate in selecting our matching human regions. For the *ith* of the 16 human regions, region size and THE1 repeat positioning within the region was specified so as to match the *ith* of the 16 chimpanzee regions.

We now carefully thinned the HapMap Phase II CEU data for those SNPs falling within each region to match the actual chimpanzee data. Specifically, for the *ith* region we first thinned the CEU data by only considering data for a panel of 36 unrelated individuals rather than the full set of CEU HapMap samples, to match our chimpanzee sample size of 36. Next, we further thinned

the CEU data by only considering a subset of the available SNPs. HapMap SNPs were subsampled to produce the same number of SNPs on each side of the region centre, within the same distance from the centre of the THE1 repeat, as was seen in the chimpanzee data for the *ith* chimpanzee region. Furthermore, this thinning was performed to satisfy three conditions. (i) The exact same number of SNPs on each side of the region centre segregated in a single HapMap individual outside the panel of 36, and also segregated in the panel of size 36, as segregated in the single chimpanzee individual "Clint" in our data and also segregated in the 36 western chimpanzees. This matches the heavy use of "Clint", the chimpanzee sequenced for the chimpanzee reference genome, to identify SNPs in Chimpanzee: ~60% of all our typed SNPs segregate in "Clint", (ii) The exact same number of SNPs on each side of the region centre do not segregate in this single individual, but do segregate in the combined set of this single individual and another individual also not in the panel, and have frequency over 5% in the 36 individuals, as do not segregate in "Clint" in the chimpanzee data for region *i*, but do have western chimpanzee sample frequency over 5%. This matches the fact that additional informative dbSNP Chimpanzee SNPs are ascertained in other Western chimpanzees via low coverage sequencing, which we model by inclusion of a second individual for SNP ascertainment. (iii) The exact same number of SNPs on each side of the region centre do not segregate in the single individual, but do segregate, at frequency below 5% in the 36 individuals, as do not segregate in "Clint" in our data but do segregate, with frequency below 5%, in our western chimpanzee data. This matches the counts of relatively uninformative additional low frequency SNPs, not captured by (i) and (ii) above in the data for the *ith* human and *ith* chimpanzee regions. This scheme is designed so that our data for each individual thinned human region ought to closely parallel features of the data for the corresponding chimpanzee THE1 region.

Because this matching is stringent, some human regions chosen cannot be matched to the appropriate chimpanzee region – in such cases we re-chose another human region for inclusion in the set of 16. Specifically, after following the above procedure, 74 of the 216 human THE1 regions could be thinned to match at least one of the 16 chimpanzee THE1 regions. In practice, to implement the above sampling scheme efficiently, we performed both the thinning, and then subsequent recombination rate estimation using LDhat (performed exactly as for the real chimpanzee data) for these 74 regions in advance, giving each of the 74 a unique label between 1 and 16, according to which chimpanzee region they were matched to. A single "sample" corresponding to human data for 16 THE1-surrounding regions was then produced by randomly sampling a single set of rate estimates with label 1, a single set of rate estimates labeled 2, and so

on up to 16. To evaluate the signal for a rate increase at motif sites, we first normalized the rate corresponding to each single region by dividing by regional mean, then measured the average normalized rate at the position corresponding to the center of the motif-containing-repeat. This procedure aims to further reduce the effect of differences in effective population size or mean background rate between species. The values obtained are shown as a histogram in fig. S3; the empirical p-value (p=0.00012) is one-sided and based on the rank of the observed mean in the actual chimpanzee data within the 10,000,000 samples. The equivalent procedure for L2, with only 6 regions, yielded p=0.38.

Testing for biases in evolution at the hotspot motif

To test for unequal evolution caused by BGC acting at the 13-bp degenerate or less degenerate core motif we searched for asymmetry in cases where only chimpanzee, or only humans, had the motif. For details of theoretical model-based predictions see the supporting online text. The search was based on local alignments of the human (hg18) and chimpanzee (pantro2) genomes, obtained from the UCSC website: http://genome.ucsc.edu/. Because misalignment is a serious concern in analyses concentrating on differences between species (in our setting it could lead to a lack of power, or false positives if asymmetric between species) we performed additional filtering steps. We removed regions with non-unique alignment, regions containing uncalled bases, pairwise alignments of length below 100 bases, and alignments of regions on non-orthologous chromosomes (e.g. for regions mapped to chimpanzee chromosome 2a, only alignments with human chromosome 2 were considered).

After this filtering process, we identified all cases where the motif CCTCCCTNNCCAC or CCNCCNTNNCCNC was present in exactly one species. A further filtering step removed instances with a base containing a PHRED quality score <40, cases with more than 5 mismatching sites or 2 bases in indels within 40bp of the first between-species difference within the motif, and motif copies within 25bp of some other identified motif (to avoid issues caused by repeats). This resulted in a set of “human only” or “chimpanzee only” motif cases. To stratify these by background, we used the UCSC repeat annotation, only labeling a region associated with a particular repeat (or non-repeat) in cases where both the human and chimpanzee annotation were in agreement. We also repeated the above with the same filters to identify a set of shared motifs between species, to normalize the non-shared counts.

For statistical testing, we initially looked for imbalances between species (Table 1) by two approaches. The molecular clock assumption generally holds well since the human-chimpanzee split (*S8*) and so in the absence of BGC, motifs seen in exactly one species ought to be randomly distributed between humans and chimpanzees, allowing testing for an imbalance via an exact binomial test within any particular repeat class. Previous work (*S9*) has suggested the 13-bp motif acts but has a differential effect on recombination rates on 5 backgrounds: THE1, L2, Alu(Y,Sc,Sg) elements, non-repeat DNA, and other repeat DNA. We therefore performed separate testing on each background, given that *a priori* BGC is expected to be strongest on the hottest THE1 background, followed by the L2 background, and weaker in the other cases. Similarly, we tested separately the core motif CCTCCCTNNCCAC and the degenerate motif CCNCCNTNNCCNC, with the latter expected to show a weaker effect, counteracted by being much more numerous in the genome. In parallel, and to allow for context-dependent effects, on each background we also performed empirical based testing for imbalances in evolution at similar motifs. For the non-repeat group, and the general "Other repeat" and combined "All" groups, this was done by analyzing all motifs 1-bp away from the core or degenerate motif (and incorporating the 2 or 5 respective degenerate bases) – excluding motifs mismatching the more degenerate bases 3,6,8,9 and 12 within this motif – in the same way as described above, to count the number of "human only" or "chimpanzee only" cases.

To test if the actual 13-bp motif showed a stronger bias than this set, we pooled counts to produce a total human-only/chimpanzee-only count, and then tested whether this differed from the observed fraction via a 1-sided Fisher's exact test. This produced the empirical p-values shown in Table 1 (reasonably closely matching the binomial test p-values). We note this procedure might be conservative, as some motifs 1-bp away from the core motif might still possess some recombination-initiating activity and hence be subject to BGC.

Empirical testing could not be performed in the same way for the specific repeats THE1, L2 and Alu(Y,Sc,Sg) elements because these have a consensus sequence, so in general other motifs 1-bp away from the 13-bp hotspot motif occur only rarely on these backgrounds. Instead, we used previous results (*S9*) identifying specific regions within the consensus and changes to the consensus sequence in each case responsible for producing motif occurrences within these backgrounds (e.g. a mutation C to T at base 167 and of T to C at base 173 of the THE1B consensus sequence produces the core motif on the THE1B background). Based on this, we used the above procedure to test for imbalances at all motifs formed by mutating the consensus

sequence of the appropriate motif in the same way as to produce the hotspot motif (e.g. for the THE1 background, all 13-bp motifs differing from the THE1B consensus by one C to T change and one T to C change at non-degenerate bases).

We thinned the resulting motif set to eliminate motifs overlapping the 13-bp region within the consensus corresponding to the hotspot motif itself. Secondly, we took only those motifs with comparable power to the real data to detect a difference between species – for the THE1 and L2 backgrounds, we thinned to motifs with at least 50 human-only or chimpanzee-only occurrences, for the Alu background we thinned to motifs with at least 5000 human-only or chimpanzee-only occurrences. If this results in a set of N 1-sided p-values, K of which are less than or equal to that observed for the actual hotspot motif, the appropriate empirical p-value is then $(K+1)/(M+1)$. This procedure resulted in the remaining empirical p-values shown in Table 1.

In silico prediction of binding consensus sequences and degeneracy for C2H2 zinc-finger proteins

We obtained the full set of classical human C2H2 zinc-finger containing protein peptide sequences from the C2H2 ZNF database (http://kzfgd.pzr.uni-rostock.de:8080/KZGD2007/index.jsp). Sequences for *PRDM9* and other genes in different mammalian species were obtained as described below. We identified zinc-fingers within each sequence by searching for matches to the PROSITE C2H2 pattern (PS00028) C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H. Here e.g. x(8) corresponds to any 8 amino acids, while x(2,4) corresponds to any 2-4 amino acid stretch.

To explore predicted DNA binding sequences for a protein with a given sequence of zinc fingers, we assumed tandem binding of successive fingers, and employed an implementation of a recently developed scoring method (*S10*) which offers improvements in in silico prediction of binding specificity relative to alternative methods (*S10*) to score candidate DNA target motifs for each protein. The approach uses a SVM framework to predict binding assuming tandem binding of zinc fingers, and based on the observed protein sequence at the canonical DNA-contacting amino acids. The x(8) section of the C2H2 pattern above defines residues -2 to 6 within the alpha helix component of the zinc finger (residues numbered -2,-1,1,2,..,6), and the canonical DNA-contacting amino acids are residues -1,2,3,6 (*S11,S12*), corresponding to sites 2,4,5 and 8 respectively within the x(8) segment. Each finger is assumed to contact a 4-bp region overlapping

1bp between fingers, and interactions between fingers can be allowed for in the model by using a non-linear kernel.

To allow a high throughput analysis, we downloaded the model (with a quadratic polynomial kernel) developed in (*S10*) and used it to make binding predictions in conjunction with the software package SVM-light (*S13*). For a given candidate DNA target motif of the appropriate length, the method gives a positive or negative score for binding of the protein. To obtain a predicted consensus DNA binding sequence for a protein, we maximized this score by hill climbing as follows. We began by finding the motif maximizing the score calculated assuming no interactions *between* fingers (via simply adding scores for different fingers, similar to the linear scoring SVM (*S10*)), while still using the polynomial kernel model to allow for interactions between amino acids *within* a zinc finger. This was done analytically via a dynamic programming algorithm moving along the zinc finger array, considering 4^4=256 possible 4-bp motif extensions after incorporating each successive zinc finger.

Although readily obtainable, the approximate motif solution this produces is not necessarily optimal for the scoring calculated as in (*S10*), as it fails to account for potential interactions between zinc fingers. Starting with the approximate motif, we therefore switched to the full scoring model of (*S10*), and then successively made that single base change within the motif giving the biggest score increase, until no score increase was possible, and took this maximum to be the final predicted binding motif for each protein. In practice, the approximate and final motifs were typically very similar for most proteins (in particular, for PRDM9 both contain the degenerate motif CCNCCNTNNCCNC), and only a small number of hill-climbing steps were typically required.

This approach was applied to the complete list of human zinc-finger protein obtained as described above. We searched through the list of predicted binding motifs for occurrences of the degenerate motif CCNCCNTNNCCNC or its complement. We also applied the method to predict consensus binding sequences for PRDM9 in chimpanzee and mouse fig. S5).

To explore predicted binding motif degeneracy for a given protein (Fig. 2, fig. S4-5), we began by calculating the scores for the predicted consensus motif, and all 1-bp alterations of this motif (a total of $3M+1$ scores for a motif of length M bases). These scores are real numbers, which must be transformed into estimates of relative binding probability in order to quantify degeneracy

fully and produce standard logo plots revealing this degeneracy. We used the makelogo (*S14*) software to produce plots, with relative heights for a given base at a certain position proportional to the estimated probability of motif binding given that base, and overall letter height given by a measure of the entropy at the base. More specifically, for the letter at position i within a motif and for base L in $\{A,C,G,T\}$ we estimated the probability of motif binding

$$P_i(L) = P(\text{Binding} \mid \text{Letter } L \text{ at position } i)$$

and took the height for letter *i* proportional to $P_i(L)$, for consistency with (*S9*). The overall stack height H_i for position i is given by the entropy relative to the case where all bases at position i give identical binding probabilities:

$$H_i = \sum_{L\in\{A,C,G,T\}} P_i(L)\log\left(\frac{P_i(L)}{\sum_{L\in\{A,C,G,T\}} P_i(L)}\right)$$

It remains to calculate $P_i(L)$.To do this, we ignored dependencies among loci and assumed that the scores produced by the SVM for different motifs can be thought of as affine transforms of scores constructed using a position sensitive scoring matrix (PSSM). Under this model, given a score S for a sequence $L_1, L_2, \ldots, L_M$ of letters in a sequence of length M, S is a transformed sum of log-odds ratios, of the form

$$S = a + b\sum_{i=1}^{M}\log\left(\frac{P(L_i \text{ at position } i \mid \text{Binding})}{P(L_i \text{ at position } i \mid \text{No binding})}\right) \quad (1)$$

where a and b are arbitrary constants to be determined.

To determine appropriate values for use in producing Fig. 2 and fig. S4, in the absence of other information we matched previously estimated overall probabilities of hotspot activity for the motif. Using Bayes formula in equation (1) we obtain

$$S = c + b\sum_{i=1}^{M}\log\left(\frac{P_i(L)}{1-P_i(L)}\right) \quad (2)$$

where c is a new constant. Finally, given two possible letters L_i^1 and L_i^2 at a given position i within the motif and identical letters elsewhere in the motif, the difference in scores S_1 and

S_2 between these two letters relates to the difference in binding probability between L_i^1 and L_i^2 and depends only on the parameter b:

$$S_1 - S_2 = b\left[\log\left(\frac{P_i(L_i^1)}{1-P_i(L_i^1)}\right) - \log\left(\frac{P_i(L_i^2)}{1-P_i(L_i^2)}\right)\right] \quad (3)$$

We used equation (3) to separately estimate a reasonable value for b for each potential letter within the motif-binding protein. First, we assume that for each position i, the choice $L_i^1 = C_i$, the optimal base at position i within the consensus sequence, gives a binding probability of 10%, corresponding to the previously estimated probability of hotspot activity for an exact match to the 13-bp core motif on a random (non-repeat) DNA background. This specifies $P_i(L_i^1) = 0.1$ and hence using the optimal score S_1 for the consensus motif itself, and choosing the appropriate value of S_2 from the set of $3M$ scores for all the single base mismatches from the binding consensus, it is possible to obtain the probability of binding for every alternative letter $P_i(L_i^2)$ at position i in terms of the unknown parameter b. Next, we choose b so that the probability of binding is reduced to, on average, 2.5% for any alternative letter within the 13-bp stretch corresponding to the motif. This is similar to the previously estimated average probability of hotspot activity for a 1-bp mismatch to the 13-bp core motif on a random (non-repeat) DNA background. Note that although this scheme results in probability estimates influenced by observations in real human hotspots, it is applied independently and identically for each potential binding protein in Fig. 2 and fig. S4. The result of this choice is that the absolute maximum binding probabilities, and average degeneracy levels, within the 13-bp motif for the candidate proteins roughly match each another and those of the real 13-bp motif. In contrast, the relative degeneracy of different bases is not affected, because this is determined by the amount the unscaled score decrease for different changes relative to the motif consensus – as a result, we restrict our discussion of the motifs produced to this relative degeneracy.

In figure 2B we also show the predicted contacts between the consensus binding sequence and the PRDM9 zinc fingers themselves. Note that contacts actually occur on the complementary strand to that shown (due to the fact that we chose to present the binding sequence for the strand corresponding to the orientation of the 13-mer hotspot motif).

To compare PRDM9 predicted binding targets across species in fig. S5, we generated logos as described above, using the scaling value $b = 0.24$ generated from the human analysis of PRDM9 (in the absence of appropriate recombination information for other species) and again assuming a sequence background such that the optimal base at a given position gives a 10% chance of binding, in equation (3).

Comparing C2H2 zinc-finger proteins between human and chimpanzee

We sought to evaluate the level of sharing of zinc fingers between human and chimpanzee across the range of zinc-finger genes. To evaluate sharing, we concentrated on comparing the canonical DNA-contacting amino acids within these fingers. Predicted human-chimpanzee zinc-finger containing orthologue pairs were identified by first taking the intersection of the human and chimpanzee gene sets downloaded from the Ensembl genome browser (Ensembl 53; http://www.ensembl.org/index.html), and then thinning to take only those genes whose collection of predicted protein sequences contains at least one instance of the PROSITE C2H2 pattern (PS00028) in each species. This process mapped 609 human predicted C2H2 zinc finger genes to a corresponding chimpanzee gene (75% of the total of 811 human predicted C2H2 zinc finger genes), including three of the five candidate 13-bp motif binding genes: *PRDM9*, *RP1-153G14.3* and *ZNF808*. The fourth candidate *ZNF697* was present on the list but had a protein sequence missing six zinc fingers relative to the C2H2 ZNF database annotation, and *ZNF124* was not present on the list.

We identified each zinc finger in each gene, and recorded the four amino acids at positions -1,2,3 and 6 within the zinc finger. A human or chimpanzee zinc finger was regarded as "conserved" between the species if the other species contained a zinc finger in the corresponding protein which was identical at all four of these amino acids, and not conserved otherwise. Sharing between species at a given gene was then quantified by the total conserved zinc fingers coded for by the human and chimpanzee exonic sequences (without regard to the order of these fingers within the sequence). In the Ensembl annotation, *RP1-153G14.3*, *ZNF808* and *ZNF697* show complete sharing of all encoded zinc fingers (22 unique fingers for *ZNF808*, 9 unique fingers for *RP1-153G14.3* and 5 unique fingers for *ZNF697*). *PRDM9* however shows sharing between species of only one of 15 unique zinc finger types present in either human or chimpanzee (as in Fig. 2).

By additional manual comparison of the human and chimpanzee genomes, we further determined that there are no predicted amino acid differences at the DNA-contacting bases between the species at any of the zinc fingers coded by the remaining gene, *ZNF124* (though gaps in the chimpanzee assembly overlap two of the nine zinc fingers in this gene, preventing our checking these cases), or in the additional six zinc fingers coded within *ZNF697*. Thus only *PRDM9* shows differences between the species by our measure. To test whether the level of sharing between human and chimpanzee in *PRDM9* of only one zinc finger type is unusual, we used the fraction of unique zinc fingers present in either human or chimpanzee that were in fact conserved in both species as a test statistic, and calculated this statistic for each identified gene pair (the observed statistic for *PRDM9* was 1/15=0.0667).

We considered all 544 pairs of genes encoding at least two C2H2 zinc fingers in each species and calculated an empirical p-value for the observed *PRDM9* statistic. 0.0667 was the lowest value observed in the set, giving p=1/544=0.00184. Finally, we performed manual checking between human and chimpanzee at *PRDM9* based on the genome sequence, to verify that the lack of sharing observed was not a result of details of the Ensembl protein prediction. We identified potential zinc fingers directly based on the reference sequence. The identified fingers incorporated exactly those identified in the human and chimpanzee PRDM9 protein sequences, as well as five putative additional zinc fingers in the chimpanzee case (for which there is no evidence of actual incorporation into any chimpanzee *PRDM9* gene product), none of which were conserved with human and perhaps indicating additional divergence in the zinc finger array between the species.

Given the uncertainty regarding the chimpanzee PRDM9 zinc finger array, we present predicted binding motifs and degeneracy in chimpanzee for both the Ensembl prediction and the larger set of all 18 possible chimpanzee zinc fingers (fig. S5).

Examination of PRDM9 zinc fingers in mammalian species

Predicted amino acid sequences for PRDM9 orthologues in chimpanzee (Refseq: XP_517829), orangutan (Ensembl peptide ID: ENSPPYG00000015366), macaque (XP_001082663), mouse (Q96EQ9), rat (P0C6Y7) and elephant (ENSLAFG00000015425) were downloaded from the Ensembl and NCBI protein databases, choosing the annotation containing the largest number of identified zinc fingers in each species. Zinc fingers were identified as described above, and the amino acids at bases -1,2,3 and 6 within each zinc finger are shown in Fig. 2. Manual checking

was also performed as described above, revealing single additional, degenerate, potential zinc fingers immediately 5’ of the tandem zinc finger array in several species including human, chimpanzee and mouse. We chose not to include these in Fig. 2, as the functional ability of these fingers is uncertain.

Supporting online text

We used theoretical models based on those from (*S15*), together with plausible estimated human parameters, to obtain several predictions regarding BGC and reported in the main text. These models support three specific points. First, we estimated and reported in the main text the fraction of motifs expected to be lost on the THE1 and L2 backgrounds in the case where the 13-bp hotspot motif has been active on the human lineage since the human-chimpanzee split. Second, we investigated patterns of loss and gain expected on the human and chimpanzee lineages in the case where the 13mer has only been active on the human lineage for some fraction of the time since the human-chimpanzee split, in order to conclude that the observed data patterns are consistent with 1-2MY of motif activity, occurring only on the human lineage. Third, we verified that theoretical modeling predicts two features seen in the real data: motif losses were more imbalanced (and these imbalances were statistically significant with smaller total counts) between human and chimpanzee for backgrounds and versions of the motif that were more recombinationally active, while motif gains never exhibited significant imbalances between species. For this third verification, under our theoretical model we calculated expected imbalances in motif losses and gains, and overall event counts required to give good power to see an imbalance, as a function of motif recombination activity. This calculation was repeated under a range of possible models for when the 13-bp motif became active on the human lineage (Table S1).

The framework required to perform these calculations is a simple extension of that in (*S15*) but is presented here for completeness. All calculations are based on the estimated probabilities of motif gain and loss, which can be shown (e.g. (*S16*)) to be a function of

1. The product $\mu_L T_A$ of the mutation rate μ_L per site per meiosis and the time T_A since the human-chimpanzee split, measuring species divergence.
2. The scaled "drive" parameter $4 N_e g = 2\alpha(r_A - r_B)$, measuring the strength of BGC. Here N_e is the effective population size, and r_A, r_B are the recombination probabilities at a motif site before and after mutation respectively, in units of

mcM (Morgans x 10^{-5}). Assuming N_e=10000, it was previously estimated that α=0.55-1.09 (*S15*).

3. The fraction *f* of the time since the human-chimpanzee split that the motif has been active on the human lineage. Setting *f*=1 corresponds to activity for the whole time since the human-chimpanzee split.

We make simplifying assumptions based on the idea that primate mutation rates are low: first, that all observed motif gains in a species are the result of a unique event since the human-chimpanzee ancestor, with no other mutations having occurred other than this event. Second, we assume that any mutation within some set of *k* sites inside a motif (non-degenerate bases) results in motif loss, and neglect the possibility that a loss-causing mutation can be reversed by a subsequent gain mutation. We set *k*=8 throughout, to ignore the 5 degenerate bases within the 13-bp motif. To remain comparable with (*S15*), we also conservatively assume every motif gain-causing or loss-causing mutation alters the recombination rate by a common factor of 2. Real rates may be higher – this would result in a proportional increase in the drive parameter for each hotspot. We assume that mutation rates are identical in humans and chimpanzees and identical across sites. Given this assumption, we use a value for $\mu_L T_A$ matching the observed level of human-chimpanzee divergence of 1.2%. Finally, we assume motif loss and gain events are immediately either lost or fix in the population.

Newly arising loss mutations occur in a motif of length k at rate μ_L per site per meiosis, and if BGC operates for time T the probability of no such mutation fixing is given by (*S15*):

$$\exp\left[-k\mu_L T\left(\frac{4N_e g}{1-\exp(-4N_e g)}\right)\right]$$

Thus the overall probability of loss of a particular motif on the human lineage is:

$$\mathrm{P(Motif\ loss)} = 1-\exp\left[-k\mu_L T_A\left(1-f+\frac{4N_e g}{1-\exp(-4N_e g)}f\right)\right]$$

The corresponding probability for the chimpanzee lineage is given by letting $4\,N_e\,g$ equal zero in the above expression (no BGC).

Gains give a more complex expression, since once motif activity begins, motif gains are also targeted by drive to remove them. We consider a sequence 1-bp away from an active hotspot motif, and add the assumption that 1/3 of all possible mutations at the single non-matching base result in mutation to the hotspot motif form, so that motif generating mutations occur at rate $\mu_L/3$ per site per meiosis. In a similar way to the above, it is straightforward to show that the probability such a sequence gains a motif on the human lineage is given by

$$\mathrm{P(Motif\ gain)} =$$

$$\exp\left[-\mu_L T_A\left(\left[(k-1)+\frac{1}{3}\right]\left[1-f\right]+\frac{4N_e g}{1-\exp(-4N_e g)}kf\right)\right]\frac{\frac{4N_e g/3}{\exp(4N_e g)-1}}{\frac{4N_e gk}{1-\exp(-4N_e g)}-\frac{4N_e g/3}{\exp(4N_e g)-1}-(k-1)}\times$$

$$\left(\exp\left[\mu_L T_A f\left(\frac{4N_e gk}{1-\exp(-4N_e g)}-\frac{4N_e g/3}{\exp(4N_e g)-1}-(k-1)\right)\right]-1\right)+$$

$$\exp\left[-\mu_L T_A\left(k\left[1-f\right]+\frac{4N_e g}{1-\exp(-4N_e g)}kf\right)\right]\times\frac{1}{2}\times\left(\exp\left[\frac{2}{3}\mu_L T_A\left(1-f\right)\right]-1\right).$$

Again, the appropriate chimpanzee gain probability is obtained by setting $4\,N_e\,g$ equal to zero (in the limit).

The required loss and gain probabilities can now readily be calculated for a set of hotspots of known mcM intensity. To investigate the expected probability of loss of the core motif within a THE1 element on the human lineage in the case of motif activity throughout the time since the human-chimpanzee ancestor, we used HapMap Phase II rate estimates for the 216 such motifs present in the human genome today. The recombination distance contributed by each hotspot motif was estimated by taking the recombination distance for the 10kb centered at the motif itself and subtracting from it an expected background distance, calculated using the average recombination rate for the flanking 16kb on either side of this 10kb region. Estimated distances below zero were set

to zero. This gave a value for r_H for each hotspot, which we then substituted into P(Motif loss) above, setting f=1 and using α=0.55-1.09 to give a likely range of loss probabilities. This resulted in a loss probability for each hotspot; the mean probability is an estimate of the fraction of hotspots expected to be lost on the human lineage, and gave 46% (α=0.55) and 56% (α=1.09). Note this may be a conservative estimate, since we assume only mutations at non-degenerate bases experience drive, and that such mutations only halve hotspot intensity. For the expected probability of loss of L2 motifs, we repeated the above analysis for the 556 L2 motif-containing elements present in humans today to give a range of 31-38%. These values become 15-18% (THE1) and 11-13% (L2) for f=0.2 (corresponding to ~1 million years of 13-bp motif activity), reasonably close to the observed values of 16% and 8%. For ~2 million years (f=0.4) of motif activity the expected loss fractions are 24-30% (THE1), and 16-20% (L2), higher than observed values, suggesting 1-2 million years as a conservative upper bound for the onset of 13-bp motif activity within humans.

Table S1 investigates the relative strength of BGC in influencing hotspot losses and gains under the first model class, and for varying hotspot intensity r_H and different possibilities for f. The ratio of human:chimpanzee loss and gain events are immediately calculable from the above formulae, and reveal that imbalances grow with recombination rate (consistent with Table 1, Fig. 1C, Tables S2, S3), while losses consistently show much larger imbalances than gains. For each table entry, power was explored by first fixing the expected number of chimpanzee loss (or gain) events, a quantity uninfluenced by BGC on the other (human) lineage. This fixes the overall number M of 13-bp motifs (or sequences 1-bp from the motif) in the human-chimpanzee ancestor. We then simulated loss and gain events on the two lineages, before performing binomial testing of human-only versus chimp-only events, exactly as for the real data. For each entry we explore a range of possible values for M. We then record the minimum expected number of chimpanzee (either loss or gain) events required to give 80% power to observe an imbalance significant, at $p<0.05$, in this binomial testing. In fact, we were able to perform exact power calculations based on the appropriate binomial distribution. In general, a much greater number of gain events than loss events on the chimpanzee lineage are

required for 80% power, particularly when motif activity began recently (e.g. f=0.25) (Table S1), so our observation of losses driving between-species imbalances (Fig. 1C) is expected. This is a result of BGC causing a rapid accumulation of motif losses, while a drop in gains is less obvious unless the sample size is very large.

Supporting figures

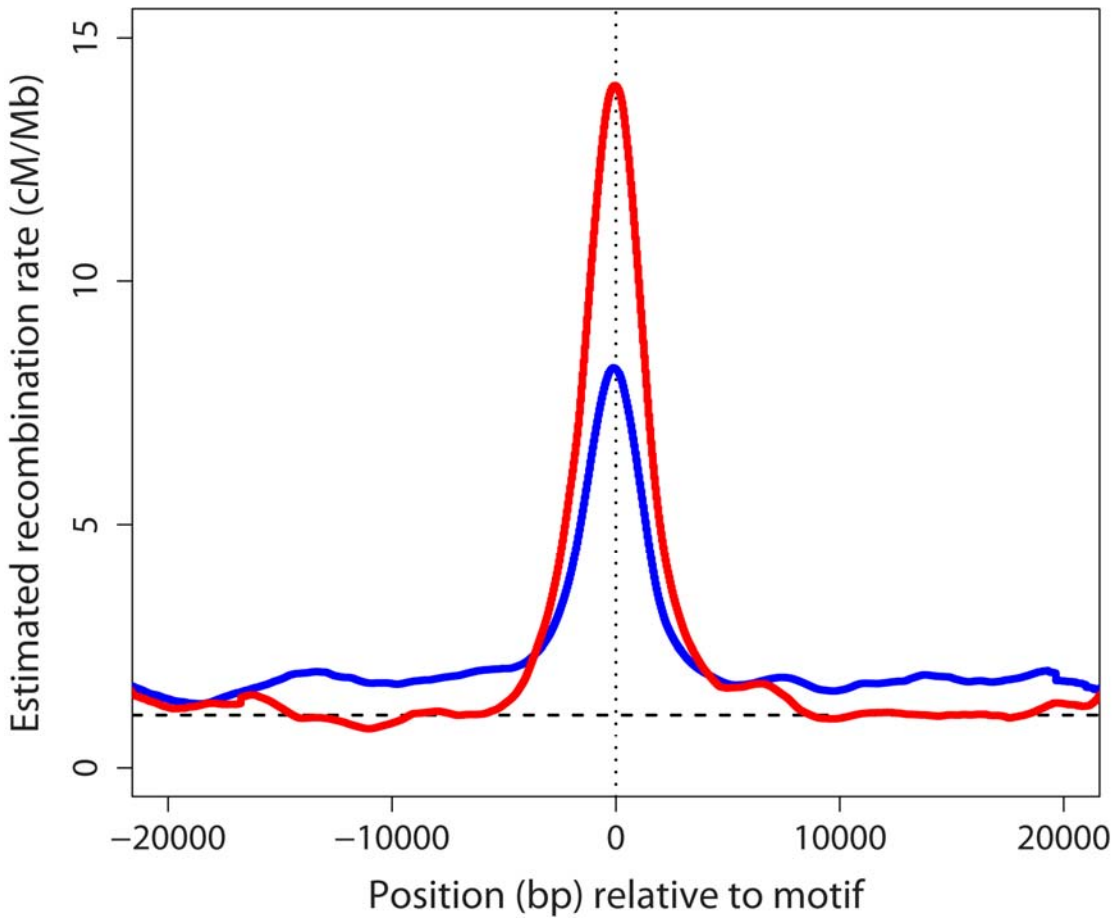

Figure S1. Estimated HapMap Phase II recombination rate across the 40kb surrounding all human THE1 elements (red line) and L2 elements (blue line) containing an exact match to the 13-bp core motif. Rates are smoothed using a 2kb sliding window slid in 50bp increments, averaged across elements. Horizontal dashed line: the human average recombination rate of 1.1cM/Mb. Vertical dotted line: the centre of the repeat.

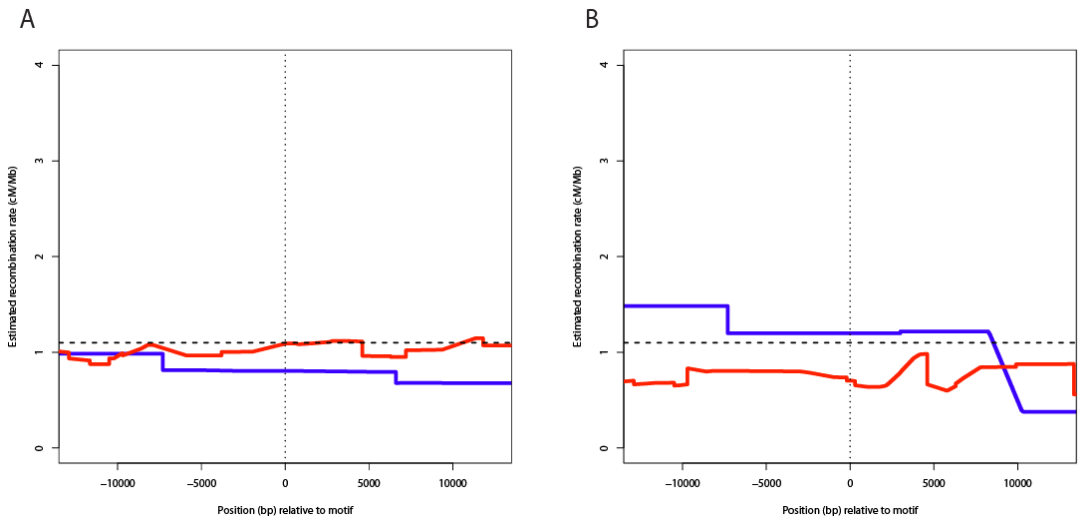

Figure S2. Estimated recombination rates around chimpanzee motif sites in central and vellerosus chimpanzees. For additional details, see Materials and Methods. **A** Average LDhat-estimated recombination rate for our central chimpanzee data, across the 25kb surrounds of 16 chimpanzee THE1 elements (red line) and 6 L2 elements (blue line) containing an 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 LDhat estimated recombination L2 and THE1 rates for our vellerosus chimpanzee data. Other details identical to A.

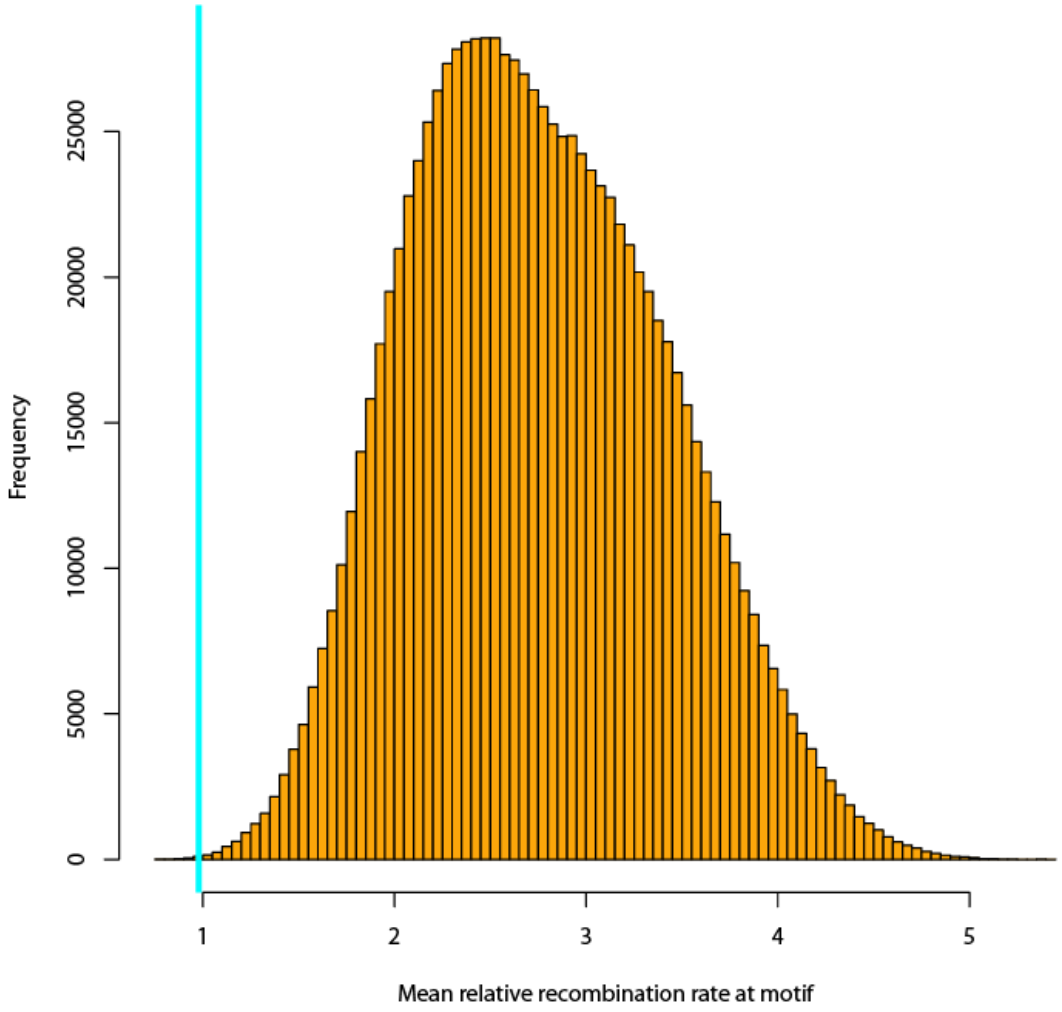

Figure S3. Histogram showing distribution of mean estimated relative recombination rate at the THE1 motif, across 10^7 random samples of 16 human motif-containing regions matched to those in chimpanzee in containing the 13-bp motif within a THE1 LTR, sample size, SNP density and ascertainment, and analysed in the same way as the chimp data. Cyan vertical line: value observed for actual chimpanzee data. For additional details, see Materials and Methods.

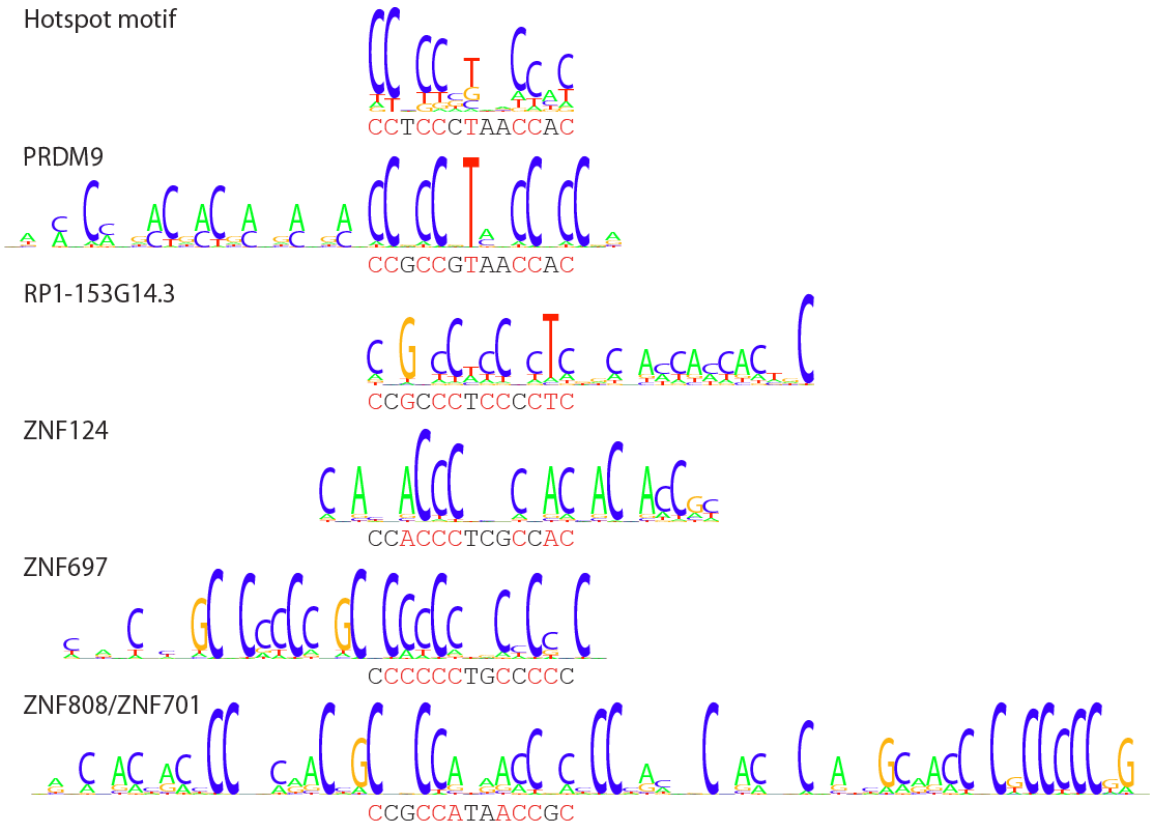

Figure S4. Degeneracy examination in candidate human proteins potentially binding the 13bp hotspot motif. The plot shows aligned pairs of logo plots and text below each plot. The PRDM9 logo plot is reproduced from Fig. 2 and shows degeneracy within the core hotspot motif while the text shows the consensus 13-bp hotspot motif (influential bases highlighted in red, degenerate bases black). Each following pair corresponds to a named candidate protein whose predicted binding sequence contains the degenerate motif CCNCCNTNNCCNC (Materials and Methods), aligned at matches to this degenerate motif. Each logo plot summarizes the predicted binding sequence and degeneracy for one candidate – relative letter height corresponds to estimated probability of protein binding for each possible base at each position within the sequence, with lower letter stack height corresponding to more degenerate positions. The text below each logo shows the predicted ideal sequence within the motif region, with bases as or more degenerate than base 12 in the "hotspot motif" plot regarded as degenerate and shown in black, other bases shown in red. Note only PRDM9 matches the required degeneracy at all sites – e.g. all other candidates mismatch by being both non-degenerate at base 6 and degenerate at base 7.

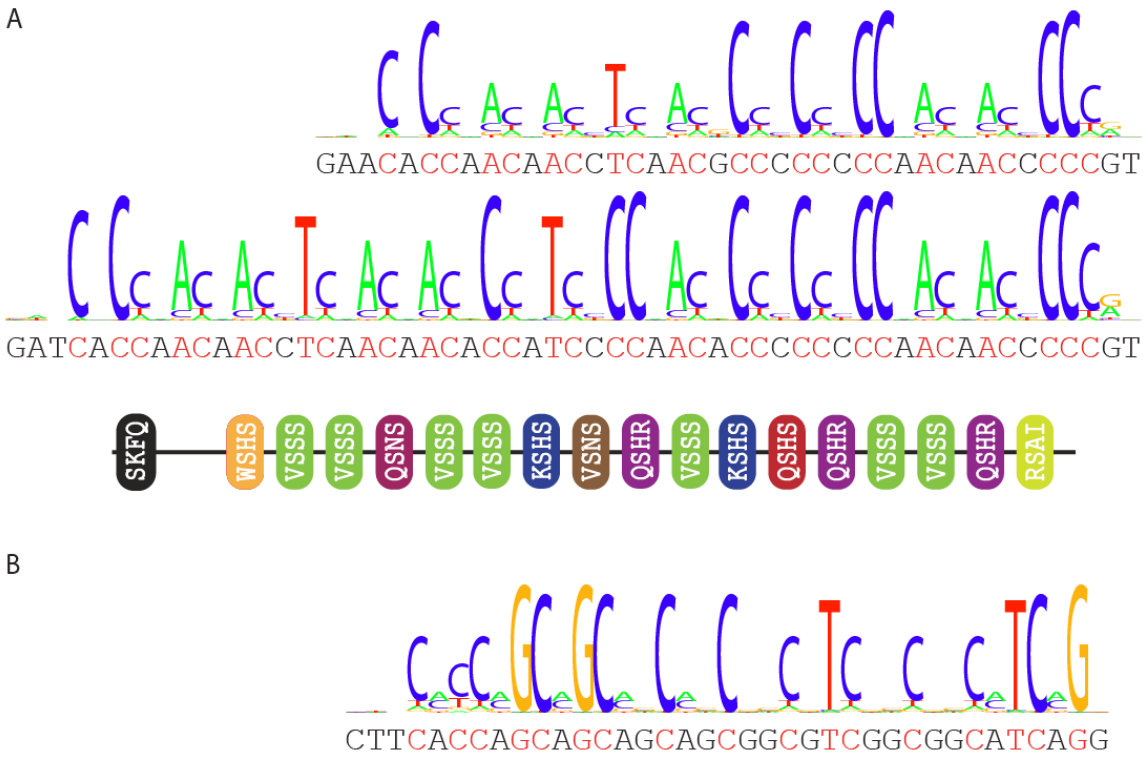

Figure S5. Predicted chimpanzee (**A**) and mouse (**B**) PRDM9 binding sequence and degeneracy. Each part shows a logo plot and aligned text below giving the predicted consensus sequence and degeneracy based on the predicted DNA contacting amino acids within the chimpanzee PRDM9 zinc finger protein sequence (Fig. 2C), produced in exactly the same way as the plots shown in fig. S4 and Fig. 2B. The above and below chimpanzee logo plots correspond to the Ensembl PRDM9 chimpanzee protein sequence prediction shown in Fig. 2C, and a potential alternative chimpanzee protein incorporating five additional zinc fingers, respectively. All predicted consensus sequences differ strongly from the predicted human PRDM9 consensus, and in particular do not contain matches to the 13-bp degenerate sequence.

Supporting tables

Table S1. Predictions of the effect of BGC on relative motif gains and losses between human and chimpanzee under a model where motif activity is restricted to the human lineage. Each cell corresponds to a given combination of the BGC parameter $4N_e g$ (between 1 and 50) and the fraction of the time (5%- 95%) since the human-chimpanzee ancestor that the 13-bp motif has been active. For each cell there are two lines corresponding to species-specific gains and losses respectively. First number on each line: predicted odds ratio for human-specific losses relative to chimpanzee-specific losses, and chimpanzee-specific gains relative to human-specific gains (both expected to be >1 in the presence of drive) respectively. Second (bracketed) number on each line: minimum expected number of events unique to the chimpanzee lineage, rounded up to the next integer, required for 80% power to reject at $p<0.05$ the hypothesis that the odds ratio is 1 (no BGC model; this number not determined if >450). For more details see supporting online text. Bold lines (always in fact highlighting losses) indicate whether losses or gains require a smaller chimpanzee event expectation for 80% power.

Fraction of time motif active		Drive parameter				
		1	5	10	20	50
5%	Gains	1.02 (>450)	1.06 (>450)	1.08 (>450)	1.10 (>450)	1.19 (>450)
	Losses	1.03 (>450)	**1.21 (368)**	**1.47 (76)**	**2.00 (22)**	**3.67 (6)**
10%	Gains	1.05 (>450)	1.13 (>450)	1.16 (>450)	1.22 (354)	1.42 (138)
	Losses	1.06 (>450)	**1.42 (92)**	**1.94 (23)**	**3.04 (8)**	**6.68 (3)**
25%	Gains	1.12 (>450)	1.39 (158)	1.49 (118)	1.69 (79)	2.44 (28)
	Losses	1.15 (>450)	**2.06 (20)**	**3.44 (7)**	**6.49 (3)**	**18.26 (1)**
50%	Gains	1.28 (256)	2.13 (36)	2.50 (28)	3.20 (20)	6.72 (12)
	Losses	**1.30 (184)**	**3.17 (8)**	**6.17 (3)**	**13.43 (2)**	**49.93 (1)**
95%	Gains	1.68 (79)	14.13 (10)	30.21 (8)	48.88 (8)	200.37 (8)
	Losses	**1.57 (53)**	**5.33 (3)**	**11.91 (2)**	**30.89 (1)**	**187.72 (1)**

Table S2. Less degenerate "core" motif loss and gain counts since the human-chimpanzee ancestor

Sequence background	CCTCCCTNNCCAC losses				CCTCCCTNNCCAC gains			
	Human lineage	Chimp lineage	Ratio (H/C)	P-value	Human lineage	Chimp lineage	Ratio (H/C)	P-value
THE1	18	6	3.0	0.011*	6	8	0.75	0.395
L2	24	13	1.85	0.049*	6	8	0.75	0.395
Non-repeats	150	115	1.30	0.018*	73	68	1.09	0.693

Table S3. Degenerate motif loss and gain counts since the human-chimpanzee ancestor

Sequence background	CCNCCNTNNCCNC losses				CCNCCNTNNCCNC gains			
	Human lineage	Chimp lineage	Ratio	Binomial p-value	Human lineage	Chimp lineage	Ratio	Binomial p-value
THE1	24	7	3.43	0.0017**	20	21	0.95	0.500
L2	200	180	1.11	0.1649	113	115	0.98	0.474
Non-repeats	2215	2024	1.09	0.0018**	1320	1249	1.06	0.922

Supporting references

S1. K. A. Frazer *et al.*, *Nature* **449**, 851 (2007).
S2. G. A. McVean *et al.*, *Science* **304**, 581 (2004).
S3. S. Myers, L. Bottolo, C. Freeman, G. McVean, P. Donnelly, *Science* **310**, 321 (2005).
S4. W. Winckler *et al.*, *Science* **308**, 107 (2005).
S5. A. J. Jeffreys, R. Neumann, M. Panayi, S. Myers, P. Donnelly, *Nat Genet* **37**, 601 (2005).
S6. J. L. Caswell *et al.*, *PLoS Genet* **4**, e1000057 (2008).
S7. The Chimpanzee Sequencing and Analysis Consortium, *Nature* **437**, 69 (2005).
S8. N. Patterson, D. J. Richter, S. Gnerre, E. S. Lander, D. Reich, *Nature* **441**, 1103 (2006).
S9. S. Myers, C. Freeman, A. Auton, P. Donnelly, G. McVean, *Nat Genet* **40**, 1124 (2008).
S10. A. V. Persikov, R. Osada, M. Singh, *Bioinformatics* **25**, 22 (2009).
S11. M. Elrod-Erickson, M. A. Rould, L. Nekludova, C. O. Pabo, *Structure* **4**, 1171 (1996).
S12. N. P. Pavletich, C. O. Pabo, *Science* **252**, 809 (1991).
S13. T. Joachims, in *Advances in Kernel Methods-Support Vector Learning*, B. Schölkopf, C. Burges, A. Smola, Eds. (MIT-Press, 1999), pp. 169-184.
S14. T. D. Schneider, R. M. Stephens, *Nucleic Acids Res* **18**, 6097 (1990).
S15. G. Coop, S. R. Myers, *PLoS Genet* **3**, e35 (2007).

S16. T. Nagylaki, *Proc Natl Acad Sci U S A* **80**, 6278 (1983).