Human recombination hot spots hidden within regions of strong marker association

Alec J. Jeffreys[1,*], Rita Neumann[1], Maria Panayi[1,3], Simon Myers[2] & Peter Donnelly[2]

[1] Department of Genetics,
University of Leicester,
Leicester LE1 7RH,
UK

[2] Department of Statistics,
University of Oxford,
Oxford OX1 3TG,
UK

[3] Present address: National Genetics Reference Laboratory, St Mary's Hospital, Hathersage Road, Manchester, M13 0JH, UK

* to whom correspondence should be addressed.

tel. +44 116 2523435

fax. +44 116 2523378

email ajj@le.ac.uk

The fine-scale distribution of meiotic recombination events in the human genome can be inferred from patterns of haplotype diversity in human populations[1-5] but only directly studied by high-resolution sperm typing[6-8]. Both approaches indicate that crossovers are heavily clustered into narrow recombination hot spots. However, our direct understanding of hot-spot properties and distributions is largely limited to sperm typing in the major histocompatibility complex (MHC)[7]. We now describe the analysis of an unremarkable 206 kb region on human chromosome 1, revealing localised regions of linkage disequilibrium (LD) breakdown that mark the locations of sperm crossover hot spots. The distribution, intensity and morphology of these hot spots are strikingly similar to those in the MHC. However, we also accidentally detected additional hot spots within regions of strong association. Coalescent analysis of genotype data detected most of the hot spots, but revealed significant differences between sperm crossover frequencies and "historical" recombination rates. This raises the possibility that some hot spots, in particular those in regions of strong association, may have evolved very recently and not left their full imprint on haplotype diversity. These results suggest that hot spots could be very abundant and possibly fluid features of the human genome.

Our direct understanding of fine-scale patterns of autosomal recombination in human DNA comes largely from studies of the class II region of the MHC which revealed a crisp pattern of blocks of markers in strong association (haplotype blocks) separated by narrow regions of abrupt breakdown of LD[7]. Sperm typing showed that these regions correspond to narrow (1-2 kb) crossover hot spots that tend to occur in clusters and within which recombination events appear to be initiated[7,9]. Hot-spot intensities agreed well with historical recombination rates estimated by population genetic approaches[5]. However, the MHC is highly atypical, being gene-rich and under intense selective pressure to diversify, and it is unclear whether crossover patterns seen in this region are more generally typical of the human genome, or whether haplotype blocks can instead arise by genetic drift acting on the products of more uniformly distributed crossovers[10-12].

To address these issues, we selected a 206 kb region of chromosome 1q42.3 for study. This region contains the only well-characterised autosomal crossover hot spot outside the MHC, located adjacent to the highly variable minisatellite MS32 (*D1S8*)[13]. The human genome sequence revealed that MS32 is located in a 77 kb long non-coding region between the *NID* (nidogen, entactin) and *TM7SF1* (transmembrane 7 superfamily member 1) genes, in an unremarkable chromosomal region showing typical gene density (eight genes in the megabase interval around MS32) and somewhat higher than average GC content (46%) and recombination activity (~2 cM/Mb in males, ~4 cM/Mb in females over the flanking 0.7 Mb interval[14]). The region selected includes MS32 and its associated hot spot, together with the nearest gene, *NID* (**Fig. 1a**).

To investigate patterns of LD across this region, we genotyped 200 SNPs in a panel of 80 UK semen donors of north European origin, with SNP coverage increased over regions of LD breakdown to refine the location of putative hot spots. LD analysis revealed a reasonably clear picture of haplotype blocks up to 80 kb long separated by intervals of partial or complete LD breakdown (**Fig. 1a**). Surprisingly, the MS32-associated hot spot lies within an extended region of strong marker association.

Genotypes were later analysed with three coalescent methods for detecting recombination hot spots[5,15,16] plus two methods which estimate fine-scale historical recombination rates[5,17] (**Fig. 1b**). This analysis revealed five putative hot spots, plus the known MS32 hot spot, each with support from at least two of the hot-spot detection methods. Three candidate hot spots, in and near the *NID* gene, mapped to regions of substantial LD breakdown and had high estimated recombination rates. The other two were upstream of *TM7SF1* and appeared to be weak, like the MS32 hot spot itself.

Sperm crossover assays were successfully developed for all intervals of LD breakdown identified in **Fig. 1a**, and were completed prior to coalescent analysis of historical recombination rates. Briefly, these assays involve the use of allele-specific PCR, directed to heterozygous SNP sites in haplotype blocks flanking a candidate hot spot, to selectively amplify crossover molecules from large batches of sperm DNA from men with appropriate SNP heterozygosities[6,7]. Analysis of the three candidate regions in the *NID* gene confirmed that they did indeed contain highly localised hot spots 1.1-2.0 kb wide (**Fig. 2**). Unexpectedly, the central hot spot (*NID*2) proved to be a double hot spot with centres just 1.4 kb apart. One of these hot spots (*NID*2a) maps to a region of LD breakdown, while the other (*NID*2b) lies almost entirely within a region of intense marker association and was not detected by coalescent analysis (or possibly was merged with *NID*2a).

Similar patterns of sperm crossover were seen in the region extending from minisatellite MS32 to upstream of *TM7SF1* (**Fig. 3**). One predicted hot spot, termed MSTM2, proved genuine but extremely weak in sperm, with a mean peak activity of 0.9 cM/Mb over three men tested (**Table 1**), no higher than the mean rate of recombination in the human genome of 0.9 cM/Mb in male meiosis[18]. A second hot spot (MSTM1) was again a closely-spaced doublet (MSTM1a, b) with centres separated by 2.0 kb. Hot-spot MSTM1a was located in a region of intense LD and was only detected by one of the coalescent methods.

To summarise, this survey includes the MS32 hot spot and revealed seven new sperm crossover hot spots. Five of these eight hot spots were detectable from LD breakdown and six, perhaps seven, by coalescent analyses. The properties of these hot spots (**Table 1)** are strikingly similar to those characterised in the MHC[7], in terms of clustering of hot spots, peak activity (0.9-70 cM/Mb, c.f. 0.4-140 cM/Mb for the MHC), width (1.1-2.1 kb, c.f. 1.0-1.9 kb in the MHC) and frequency (5 hot spots detectable from LD analysis over 206 kb compared with 6 hot spots over 216 kb in the MHC). There are no obvious DNA sequence similarities between different hot spots. The total mean recombination frequency r over all hot spots is 2.3 x 10^{-3}, giving an overall sperm activity of 1.1 cM/Mb over the entire 206 kb interval, very close to the genome average male rate of 0.9 cM/Mb. All hot spots showed indistinguishable rates of reciprocal crossover in all men tested and in most cases indistinguishable distributions of breakpoints in reciprocal crossover products (data not shown). Three hot spots (*NID*1, MS32 and MSTM2) showed significant rate variation in different men, with one man showing no detectable crossovers at MSTM2. The recombination rate outside hot spots appeared to be extremely low, with 16 putative exchanges seen in a total of 13.5 kb DNA within the crossover test intervals (total length 32 kb) but outside hot spots (**Figs. 2, 3**). Since some of these rare exchanges might be PCR artifacts[6], this gave an upper estimate of 0.04 cM/Mb for exchanges outside hot spots, exactly as estimated in the MHC region[7], and suggests that the great majority of crossovers occur in hot spots.

Coalescent methods, which estimate sex-averaged recombination rates over long periods of time, gave some marked differences from current male frequencies assayed in sperm (**Table 1**). For example, of the *NID* hot spots, the one predicted from coalescent analysis to be the weakest (hot-spot *NID*3; **Fig. 1b)** was in fact the most intense in sperm **(Fig. 2c)**. Some hot spots such as *NID*1, *NID*2a and MSTM2 appear to have higher historical rates than those seen in sperm. In contrast, three hot spots (*NID*2b, MS32, MSTM1a) have left little imprint on LD yet are as active in sperm as MHC hot spots that have caused complete breakdown of LD[7]. A similar discrepancy is seen for the most intense hot spot, *NID*3, which nevertheless shows significant association between markers upstream and downstream of the hot spot (**Table 1**). Furthermore, 12% of crossovers at this hot spot map within an upstream region of intense LD (**Fig. 2**) (D' = 0.98, likelihood ratio (LR) in favour of significant LD = 10^{21}:1) despite substantial crossover activity in this region ($r = 1 \times 10^{-4}$ per sperm).

We carried out simulation studies to better understand the joint effects of hot-spot intensity and genetic drift on simple measures of association, and whether the striking differences between historical and sperm rates might simply be due to uncertainties in the coalescent estimation method. We first simulated two markers with recombination frequency *r* between them, using an effective population size appropriate to north Europeans[19] (**Fig 4**). Starting with absolute association (only two haplotypes for a pair of markers), LD as measured by D' and the associated likelihood ratio (LR) decay, but on occasion genetic drift can subsequently re-establish significant LD. LR identifies these periods of strong pairwise association more reliably than the relatively chaotic D' measure (**Fig. 4a**). As expected, these periods increase in frequency and duration as *r* decreases. Next, we used coalescent simulations to determine the likelihood of a hot spot being within such a period (**Fig. 4b)**. For hot spots with $r > 10^{-3}$, it is unlikely, under the parameters used for simulation, that drift will generate significant associations. In contrast, weak hot spots ($r < 10^{-5}$) are unlikely to be in a period of low LD. For intermediate rates, such as at hot-spots *NID*2b and MSTM1a ($r = 0.3 \times 10^{-4}$, 0.9×10^{-4} respectively) located in regions of strong marker association, whether or not a hot spot corresponds to LD breakdown will be a matter of chance population history.

Finally, we generated coalescent simulations across the entire region that matched the real data for SNP locations, SNP allele frequencies, and missing data. These simulations used the positions and intensities of crossover hot spots detected in sperm, and assumed a constant male:female ratio of crossover rates. We simulated constant size populations plus two population bottleneck models proposed for European populations[20-22], and estimated historical recombination rates from each set of simulated polymorphism data. These rates, estimated from the simulated genotypes, cluster around the rates used in the simulations and are little affected even by quite extreme bottlenecks (**Fig. 5a**). Irrespective of demographic scenario, the relative historical intensities of hot spots estimated from the real genotype data were more discordant than any of the relative historical intensities from these simulations. We can therefore reject the null hypothesis that relative historical rates match relative sperm rates at $P = 0.001$. Analysis of individual hot spots showed that *NID*2a, MS32 and MSTM2 have significantly different historical and sperm rates (**Fig. 5b**). We conclude that the differences between the sex- and time-averaged rates estimated from genotype data and recombination frequencies in sperm are likely to be real, rather than an artefact of the estimation method.

Hot spots such as MSTM2, which have higher historical rates than those seen in sperm, may simply be more active in female meiosis[6] or in a subset of men yet to be tested. On the other hand, the observed associations across the stronger hot-spots *NID*3 and MS32 (**Table 1**) are difficult to explain by genetic drift under population demographic histories plausible for European populations, even if the hot spots are not functional in female meiosis. We can never formally exclude the existence of an extreme population history that would largely eliminate evidence of historical recombination, but it seems unlikely that it would do so at only some hot spots, whilst increasing evidence for recombination at others. The simplest alternative explanation for active hot spots such as *NID*3 and MS32, located in regions of strong association, is that they are young and have evolved so recently that they have failed to leave their full mark on haplotype diversity in Europeans. This would be consistent with evidence from diversity surveys that hot spots are ephemeral during primate evolution[23,24,25] and predicts that some such hot spots might be polymorphic in human populations, with ancestral inactive haplotypes still in existence; this prediction is testable. Conversely, where historical rates are larger than sperm rates, it might be that the hot spots are older and en route to extinction. Our observation that the hot-spot MSTM2 shows polymorphism in sperm crossover frequency is consistent with this hypothesis.

The present study establishes that the pattern of recombination hot spots seen in the MHC, which has been widely used as a model for interpreting human haplotype structures (see refs. 26,27), is likely to be more widely applicable in the human genome. However, it has also shown that regions of strong marker association are not necessarily recombinationally suppressed and can harbour even quite active hot spots. Historical recombination rates play a key role in shaping patterns of haplotype diversity within populations. These rates appear to differ from contemporary crossover frequencies in sperm, plausibly because of rapid evolution of the hot spots themselves. It is historical recombination rates that are relevant for association mapping of disease genes and for most population genetic analyses. In contrast, contemporary sperm rates are needed to unravel the determinants of hot-spot activity and the origin and evolutionary dynamics of recombination hot spots in the human genome. In particular, systematic surveys of sperm crossovers within haplotype blocks are needed to establish the true incidence of "hidden", and potentially evolutionarily very young, hot spots.

Methods

Genotyping. We collected, with approval from the Leicestershire Health Authority Research Ethics Committee, semen and blood samples with informed consent from 200 UK men of north European descent including volunteers and men attending fertility clinics. We selected 80 men showing good sperm DNA yields for analysis. Sperm DNAs were whole-genome amplified by MDA[28] and genotyped by subsequent PCR amplification of appropriate 3-8 kb targets followed by allele-specific oligonucleotide (ASO) hybridisation to dot blots of PCR products[6]. Of the 200 SNPs used in this study, we recovered 121 by BLAST searches of the NCBI SNP database[29] and 79 by resequencing selected DNA regions in 6-8 north Europeans. Details of SNPs, ASOs and genotypes can be found at http://www.le.ac.uk/ge/ajj/MS32/. None of the markers showed significant deviation from Hardy-Weinberg equilibrium. We performed LD analysis on unphased genotype data as described previously[7].

Historical recombination rate estimation. We used coalescent analyses to estimate recombination rates from genotype data established for all 200 SNPs in the panel of 80 semen donors, using LDhat (ref 5) with a smoothing penalty of 5 and 10^8 iterations of which the first 10% were discarded as burn in. PHASE (http://www.stat.washington.edu/stephens/software.html) was run with the setting X10, as recommended in the documentation.

Hot-spot detection by coalescent analysis. We applied three methods, each using different approximations to the coalescent model, and so in effect differing aspects of the data. Each method compares, via the likelihood ratio, the fit of a model that explicitly includes a hot spot in a particular small window with the fit of a model with constant recombination rate, to look for evidence of a hot spot within the window, and then slides the window across the region of interest. There is reasonable agreement between the putative hot spots detected by each method, but some differences, as expected since the different methods rely on different “signals” in the data and so will differ in power in different parts of the region depending on SNP density, allele frequencies and genotype configurations. We applied the method LDhot from ref 5 directly to the genotype data and used a simple model for SNP ascertainment, namely that a Poisson number of chromosomes with mean three was chosen from a panel of 12 chromosomes and a SNP detected if it was polymorphic in the chosen set. We used haplotypes estimated from the PHASE run described above as inputs to the other two hot-spot detection methods, Hotspotter[15] and Fearnhead's method[16]. The latter only specifies hot-spot location within a window of 6 SNPs, so we used LDhat rate estimates to further refine the location. For Hotspotter, we tested successive SNP intervals for the presence of a hot spot, establishing significance for each interval by performing a one-sided test that assumed an equal mixture of a chi-squared distribution on one degree of freedom and a point mass at zero for the likelihood ratio statistic. Only one of the hot spots analysed in sperm could not be detected by any of the methods and it seems unlikely that population-based approaches would be able to separate overlapping hot spots such as *NID*2a and *NID*2b. Fearnhead's method appears to be the most powerful of these coalescent approaches for detecting hot spots. Our implementation of Hotspotter may be susceptible to false positives. These conclusions about the performance of the methods may be quite sensitive to the dense SNP spacing around the hot spots in this study, and may not necessarily generalize to other settings. P. Fearnhead, C. Freeman, J. Marchini and C. Spencer contributed to the coalescent analyses described above.

Sperm crossover analysis. For each target region, we selected SNPs in the 5' and 3' haplotype blocks (usually three per block) and designed allele-specific primers (ASPs) for each SNP. We assayed these ASPs by PCR on genomic DNA from individuals homozygous for the correct or incorrect allele, to identify primers that showed excellent efficiency and allele specificity and to determine optimal annealing temperatures. We identifed men who were heterozygous for appropriate 5' and 3' SNPs and with additional intervening SNP heterozygosities needed for subsequent crossover mapping. We established the linkage phase between 5' and 3' SNPs in these men by PCR on genomic DNA using these ASPs. We prepared sperm DNA for crossover analysis as described previously under conditions designed to minimise the risk of contamination[6]. Multiple batches of sperm DNA each containing, depending on crossover rate, 700-22,000 amplifiable molecules of each progenitor haplotype (8.4-264 ng DNA containing 0.4-2 crossover molecules) were amplified in 96-well plates by long PCR using ASPs in repulsion phase, selected for compatible optimal annealing temperatures, to selectively amplify crossover molecules. We digested these primary PCR products with S1 nuclease to remove single-stranded DNA and PCR artifacts and re-amplified them using nested internal ASPs in repulsion phase. We analysed these secondary PCRs by agarose gel electrophoresis and visualised DNA by staining with ethidium bromide to identify crossover-positive reactions. We re-amplified secondary PCRs using nested non-allele-specific primers and mapped crossover exchange points by dot-blot hybridisation of these tertiary PCR products with ^{32}P-labelled ASOs. All crossover assays included multiple aliquots of blood DNA; no examples of crossovers were detected in these negative controls. Full details of long PCR, S1 digestion, crossover mapping and Poisson correction for more than one crossover per PCR are given elsewhere[6,7]. We provide details of ASPs and crossover assay conditions for each target region at http://www.le.ac.uk/ge/ajj/MS32/.

Association simulations. To investigate the effects of recombination frequency on LD decay, we seeded multiple populations of 10,000 diploid individuals (an effective population size appropriate for north Europeans[19]) with equal numbers of haplotypes AB and ab and allowed them to evolve with random sampling of gametes from the parental generation and with crossovers between the two loci occuring at a given recombination frequency. Every 100 generations we sampled 80 individuals to match the sample size used in LD surveys (**Fig. 1**). Simulations were terminated when either minor allele frequency fell below 0.15, corresponding to the threshold used in **Figs. 1-3**. We combined the haplotypes in the 80 sampled individuals into unphased genotypes, and used these to estimate the D' measure of LD between the two markers and the associated likelihood ratio LR in favour of LD, exactly as in **Fig. 1**. LR is defined as the ratio of the probability of observing the genotypes at the maximum likelihood haplotype frequencies compared to the probability of observing the genotypes at the haplotype frequencies expected if markers are in free association. Strengths of association between a pair of markers were similarly investigated by coalescent simulations carried out using the ancestral recombination graph[30] to simulate samples of 160 two-locus haplotypes at stationarity for different specified recombination rates between the loci. We again set the effective population size to 10,000. Importance sampling was used to correct for the conditioning on each locus being segregating, haplotypes were combined randomly to give 80 two-locus genotypes, and samples with a minor allele frequency <0.15 at either locus were discarded. LR analysis of each remaining sample (1000 samples for a given recombination rate) was carried out as above.

To assess the precision of the coalescent estimates of historical rates, we ran coalescent-based simulations to produce data matching the actual genotype data for numbers of SNPs, allele frequencies and missing data. The simulations had recombination hot spots at the positions observed in sperm, with recombination rates calculated from the sperm crossover frequencies in the hot spots and from estimated background rates outside hot spots, and assuming as from the genetic map that females had twice the crossover rate of males across the region. For each simulated data set we estimated recombination rates using LDhat. For each demographic scenario, the set of 1000 simulations was used to estimate an effective population size by comparing the estimated LDhat rate across the entire region with the known rate used in the simulation. The respective N_e estimates were then used, as for the real genotype data, to convert LDhat estimates into crossover rates. The collection of estimated rates for each hot spot (total rate across the hot-spot region as defined in **Table 1**) for each simulation was then compared with the rate estimated from the actual genotype data.

Three different demographic scenarios were used in these simulations: constant population size (N_e = 10,000); constant population size (N_e = 10,000) followed by an instantaneous bottleneck (F=0.18) 800 generations ago; and constant population size (N_e = 10,000) followed by an instantaneous bottleneck (F=0.18) 1600 generations ago (generation time 25 years in each case), the latter two scenarios following those suggested elsewhere[21,22]. Some such bottleneck is thought likely to have occurred in European ancestry[20]. While the effect of a bottleneck is to reduce the evidence of historical recombination, the sampling distribution of LDhat rate estimates across the hot spots was in fact little affected by even these quite severe bottlenecks (**Fig 5a**).

For each simulated dataset, we used Spearman's rank correlation coefficient to measure the discrepancy between the estimated relative hot spot intensities and the true relative intensities. The value for the actual data (0.286) was lower (i.e. more discrepant from the truth) than for any simulated dataset, for each demographic scenario. Note that this comparison of hot-spot intensities *relative* to each other provides robustness to some departures from the assumptions underpinning our simulations.

To assess whether historical rate estimates differed from sperm rates at particular hot spots, we used a two-sided test of whether the actual historical rate estimate was an outlier amongst rate estimates from the simulated data. This test showed that three hot spots (*NID*2a, MS32, and MSTM2) have significant differences between historical and sperm crossover rates for all three demographic scenarios ($P< 0.008$, $P<0.018$, $P<0.002$ respectively). False discovery rate (FDR) analysis suggested that these *P* values were significant in all three cases, based on an overall FDR of 0.05. For two other hot spots (*NID*1, *NID*3) the *P* values are small for all three scenarios, suggesting the discrepancies may also be real.

Acknowledgements

We thank J. Blower and volunteers for providing semen and blood samples, S. Mistry for oligonucleotide synthesis, K. Holloway for DNA samples, C. May for web site design, A. Webb for bioinformatics advice, P. Fearnhead, C. Freeman, J. Marchini and C. Spencer for coalescent analyses, and colleagues for helpful discussions. This work was supported by grants to AJJ from the Medical Research Council, the Royal Society *and the Louis-Jeantet Foundation, and to PD from the NIH, the Nuffield Trust and the Wolfson Foundation.*

References

1. Chakravarti, A. *et al.* Nonuniform recombination within the human beta-globin gene cluster. *Am. J. Hum. Genet.* **36**, 1239-1258 (1984).
2. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225-2229 (2002).
3. Reich, D.E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet*. **32**, 135-142 (2002).
4. Crawford, D.C. *et al.* Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.* **36**, 700-706 (2004).
5. McVean, G.A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581-584 (2004).
6. Jeffreys, A.J., Ritchie, A. & Neumann, R. High-resolution analysis of haplotype diversity and meiotic crossover in the human *TAP2* recombination hotspot. *Hum. Mol. Genet*. **9**, 725-733 (2000).
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. May, C.A., Shone, A.C., Kalaydjieva, L., Sajantila, A. & Jeffreys, A.J. Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene *SHOX*. *Nat. Genet.* **31**, 272-275 (2002).
9. Jeffreys, A.J. & May, C.A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat. Genet.* **36**, 151-156 (2004).
10. Wang, N., Akey, J.M., Zhang, K., Chakraborty, R. & Jin, L. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* **71**, 1227-

1234 (2002).

11. Phillips, M.S. *et al.* Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.* **33**, 382-387 (2003).

12. Stumpf, M.P. Haplotype diversity and SNP frequency dependence in the description of genetic variation. *Eur. J. Hum. Genet.* **12**, 469-477 (2004).

13. 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).

14. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241-247 (2002).

15. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213-2233 (2003).

16. Fearnhead, P., Harding, R.M., Schneider, J.A., Myers, S. & Donnelly, P. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics* **167,** 2067-2081 (2004).

17. Stephens, M. & Donnelly, P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**, 1162-1169 (2003).

18. Gyapay, G. *et al.* The 1993-94 Généthon human genetic linkage map. *Nat. Genet.* **7**, 246-339 (1994).

19. Morton, N.E. *Outline of genetic epidemiology* (Karger, Basel, 1982).

20. Reich, D.E. *et al*. Linkage disequilibrium in the human genome. *Nature* **411**, 199-204 (2001).

21. Sabeti, P.C. *et al*. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-837 (2002).

22. Akey, J.M. *et al*. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol*. **2**, e286 (2004).

23. Wall, J.D., Frisse, L.A., Hudson, R.R. & Di Rienzo, A. Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. *Am. J. Hum. Genet.* **73**, 1330-1340 (2003).

24. Ptak, S.E. *et al.* Absence of the TAP2 human recombination hotspot in chimpanzees. *PLoS Biol.* **2**, 849-855 (2004).

25. Winckler, W. *et al*. Comparison of fine-scale recombinaion rates in humans and chimpanzees. *Science* 10 Feb Epub (2005).

26. Stumpf, M.P. & McVean, G.A. Estimating recombination rates from population-genetic data. *Nat. Rev. Genet.* **4**, 959-968 (2003).

27. Kauppi, L., Jeffreys, A.J. & Keeney, S. Where the crossovers are: recombination distributions in mammals. *Nat. Rev. Genet.* **5**, 413-424 (2004).

28. Dean, F.B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. USA* **99**, 5261-5266 (2002).

29. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308-311 (2001).

30. Griffiths, R.C. & Marjoram, P. An ancestral recombination graph. In *IMA Volume on Mathematical Population Genetics*, ed. P. Donnelly and S. Tavaré (Springer-Verlag, Berlin/Heidelberg/New York) pp. 257-270 (1996).

Figure legends

Figure 1 Patterns of linkage disequilibrium (LD) and historical recombination in a 206 kb interval around minisatellite MS32. (**a**) LD profile across MS32 and the neighbouring *NID* gene established from 200 SNPs genotyped in a panel of 80 UK semen donors of north European origin. Maximum likelihood haplotype frequencies for each pair of SNPs were used to estimate |D'| levels of LD (lower right), plus the associated likelihood ratio (LR) versus free association (upper left), and are colour-coded as indicated[7]. SNPs with minor allele frequencies <0.15 (25 in total) were excluded from analysis. The locations of the remaining 175 SNPs are shown below and to the right of the plot, with positions centred on the middle of MS32 at co-ordinate 0. LD blocks were identified visually as regions where most marker pairs are in strong (|D'| > 0.8) and highly significant (LR > 10^4) association. Regions of LD breakdown targeted for sperm crossover analysis are shown; analysis of the region around MS32 has been reported previously[13]. (**b**) Historical recombination rates and positions of putative recombination hot spots (marked above plot) estimated from coalescent analyses of genotype data. Population recombination rates ρ, defined as $\rho = 4N_e r$ where N_e is the effective population size and r is per-generation recombination rate, were estimated across the region using LDhat (red) (ref 5) and PHASE (blue) (ref 17). These in turn were converted to r assuming that $N_e = 10{,}000$ (ref. 19) and used to estimate the local sex-averaged recombination activity in cM/Mb. Coloured triangles show putative recombination hot spots significant at $P < 0.01$ for three different hot-spot detection methods: LDhot (ref 5) (red), Hotspotter (ref 15) (blue) and Fearnhead's method[16] (green). See Methods for further details. All coalescent analyses were undertaken after sperm typing and without knowledge of the sperm typing results.

Figure 2 LD and sperm crossover profiles across the *NID* gene. (**a**) Organisation of the 3' end of the *NID* gene, with exons shown in black. (**b**) Expanded LD profiles and the location of LD block boundaries across the three regions of LD breakdown, plotted as in **Fig. 1**. (**c**) Sperm crossover profiles in each region. Data for each region were derived by typing sperm from a single man for reciprocal (A-type, B-type) crossover molecules (Cd and cD crossovers in a man heterozygous for haplotypes CD and cd). Exchange points in all crossovers were mapped and used to determine the number of exchanges, shown in italics, in each interval between heterozygous markers and thence the recombination activity in cM/Mb. The location of each marker is shown as a tick above the plot; discordancies in markers between the LD and recombination plots arise from high frequency markers that are homozygous in the semen donor and from low frequency markers excluded from LD analysis but informative in the donor. Red curves show the underlying crossover distributions assuming that crossovers are normally distributed across each hot spot, established by least-squares best-fit analysis of cumulative crossover frequency distributions[7]. The fits are generally good (hot-spot *NID*3, $\chi^2_{[2\text{ d.f.}]} = 10.0$, $P = 0.02$; *NID*2a plus 2b, $\chi^2_{[5\text{ d.f.}]} = 0.1$, $P = 1$; *NID*1, $\chi^2_{[2\text{ d.f.}]} = 0.5$, $P = 0.8$). The crossover distribution across *NID*2a and 2b does not fit a single normal distribution ($\chi^2_{[5\text{ d.f.}]} = 63$, $P < 0.001$), and the double hot spot was further verified in a second man who showed the same bimodal crossover profile (not shown). The following numbers of crossovers and progenitor molecules were analysed: *NID*3, 221 A-type and 249 B-type crossovers, each recovered from 230,000 amplifiable molecules of each progenitor haplotype ; *NID*2a plus 2b, 146 A and 138 B crossovers, each recovered from 1,250,000 amplifiable molecules; *NID*1, 51 A and 67 B crossovers, each recovered from 180,000 molecules.

Figure 3 LD and sperm crossover profiles near minisatellite MS32. (**a**) The region analysed. (**b**) Expanded LD profiles across MS32 and the two regions of LD breakdown. (**c**) Sperm crossover profiles across MS32 and the two additional regions, plotted as in **Fig. 2**. Data for MS32 were taken from ref. 13 and were averaged over three men tested. The following numbers of crossovers and progenitor molecules were analysed: hot-spot MS32, 250 crossovers in 640,000 molecules; MSTM1a plus 1b, 169 A crossovers in 600,000 molecules and 119 B crossovers in 400,000 molecules from one man; MSTM2, 17 A and 21 B crossovers, each in 1,100,000 molecules from one man. The goodness-of-fit of the normal distributions shown are: MS32, $\chi^2_{[8\ \mathrm{d.f.}]} = 8.0$, $P = 0.5$; MSTM1a plus 1b, $\chi^2_{[4\ \mathrm{d.f.}]} = 13.4$, $P = 0.02$; MSTM2, $\chi^2_{[2\ \mathrm{d.f.}]} = 2.1$, $P = 0.4$. The distributions for hot-spots MSTM1a plus 1b are approximate since not all of MSTM1a is included in the test interval and MSTM1b lacks markers near the centre of the hot spot.

Figure 4 Impact of crossover frequencies on strengths of pairwise marker association. (**a**) Influence on D' measures of LD and on the strength of marker association. Constant-sized random-mating populations of 10,000 individuals were simulated for two markers separated by the indicated recombination frequency *r*, with the population initially containing just two equifrequent haplotypes. A sample of 80 individuals, corresponding to the sample size used in genotyping, was taken every 100 generations and used to estimate D' and the likelihood ratio LR in favour of linkage disequilibrium, as in **Fig. 1**. The simulations shown maintained minor allele frequencies above 0.15, the threshold frequency used in **Figs. 1-3**. Simulations initiated with more than two haplotypes, and with various levels of initial LD, produced very similar profiles after

the initial phase of LD decay (data not shown). (**b**) Probability from coalescent simulations that a marker pair with minor allele frequencies above 0.15 will show a strength of association above the indicated LR thresholds (dashed lines in (**a**)). Even with no recombination, some simulations give low LRs due to relatively uninformative genotype configurations. Observed mean sperm crossover frequencies at the hot spots near MS32 are indicated below.

Figure 5 Testing whether sperm crossover frequencies are compatible with coalescent estimates of historical recombination rates. We simulated 1000 datasets under each of three different demographic scenarios: model 1, solid line, constant population size of 10,000; model 2, dashed line, bottleneck 800 generations ago with inbreeding coefficient 0.18; model 3, dotted line, bottleneck 1600 generations ago with inbreeding coefficient 0.18. Simulations used background recombination rates, hot-spot positions and intensities as seen in sperm; see Methods for details. For each simulation, we estimated historical recombination rates from simulated genotypes using LDhat as for the real genotype data, calculating the recombination rate across each hot spot. **(a)** The smoothed empirical distributions of rates for each hot spot and demography, with the vertical green line giving the recombination rate under which the data were simulated, and the red vertical line showings the LDhat estimate from the real genotype data. A red line located in the tail of the empirical distribution indicates a historical rate inconsistent with the sperm rate under the model. **(b)** Empirical *p*-values for each hot spot estimated from these simulations, using the rank of the observed value in the empirical distribution to gain a (two-sided) *p*-value for a test of whether the observed LDhat rate estimate is compatible with the sperm-based rate estimate. The minimal *p*-value possible from 1000 simulations is 0.002.

Figure 1

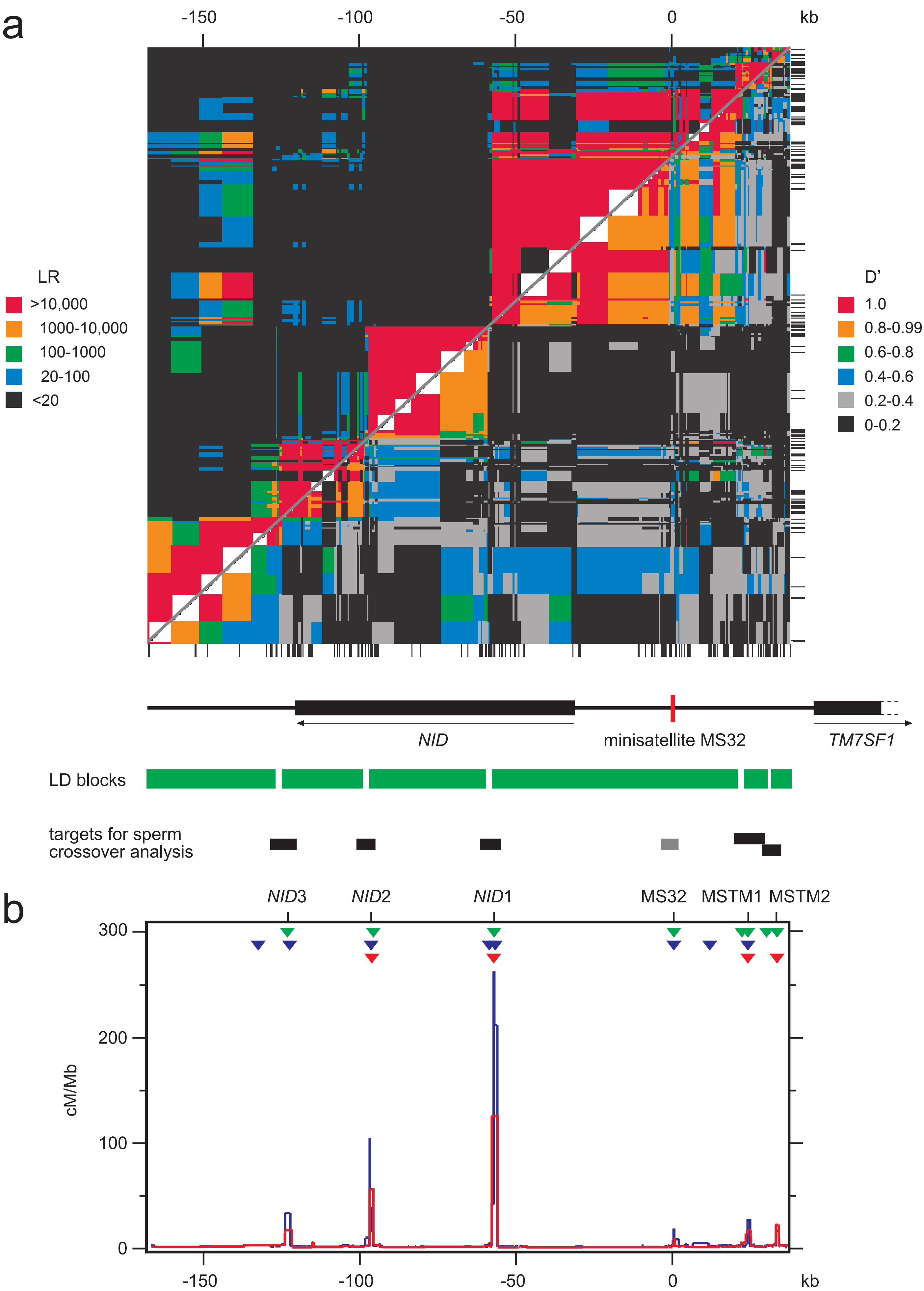

a
-150
-100
-50
0
kb
LR
>10,000
1000-10,000
100-1000
20-100
<20
D'
1.0
0.8-0.99
0.6-0.8
0.4-0.6
0.2-0.4
0-0.2
NID
minisatellite MS32
TM7SF1
LD blocks
targets for sperm
crossover analysis
b
NID3
NID2
NID1
MS32
MSTM1
MSTM2
300
200
100
0
cM/Mb
-150
-100
-50
0
kb

Figure 2

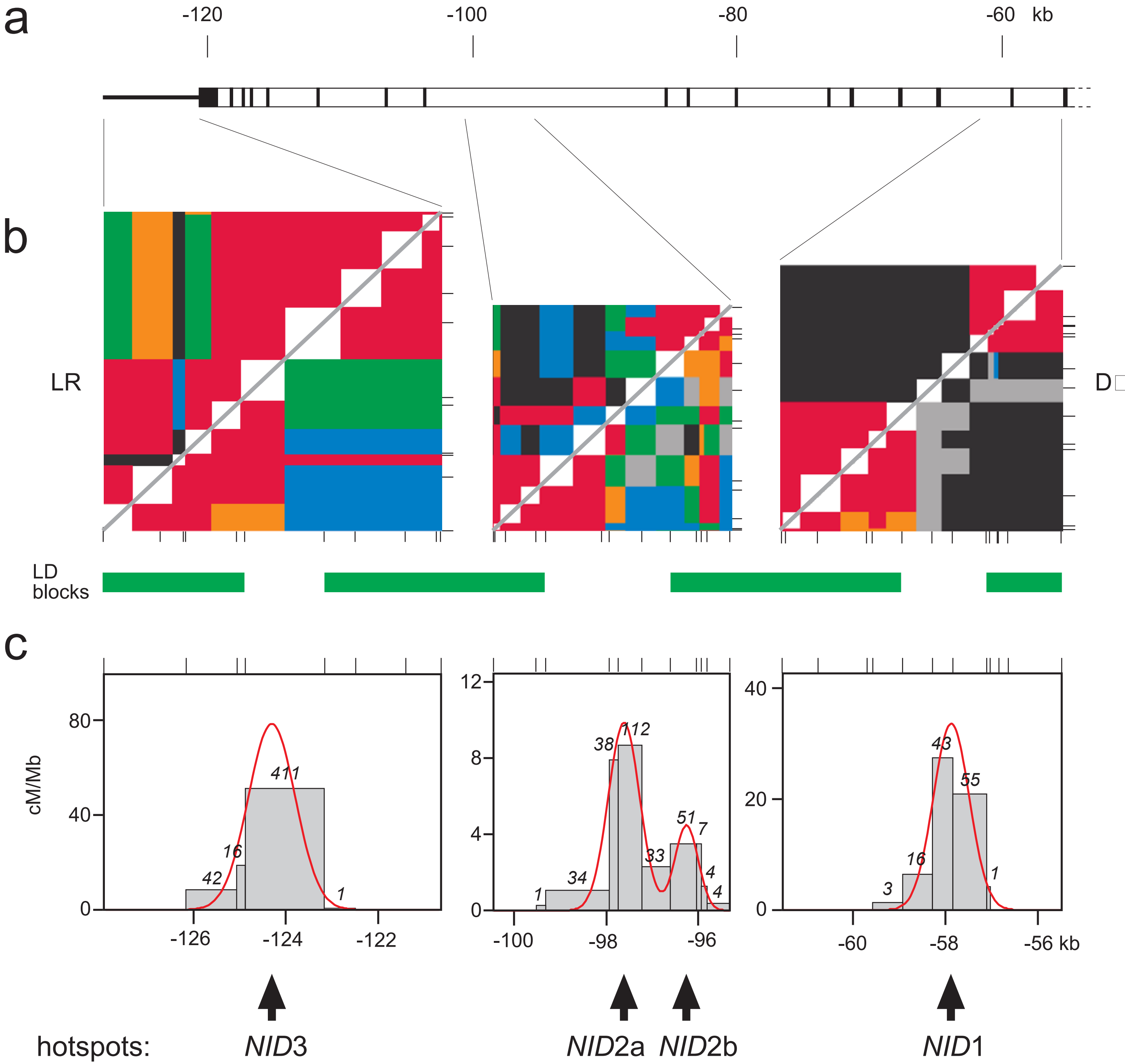

a
-120
-100
-80
-60
kb
b
LR
D
LD blocks
c
cM/Mb
80
40
0
-126
-124
-122
42
16
411
1
12
8
4
0
-100
-98
-96
1
34
38
112
33
51
7
4
4
40
20
0
-60
-58
-56 kb
3
16
43
55
1
hotspots:
NID3
NID2a
NID2b
NID1

Figure 3

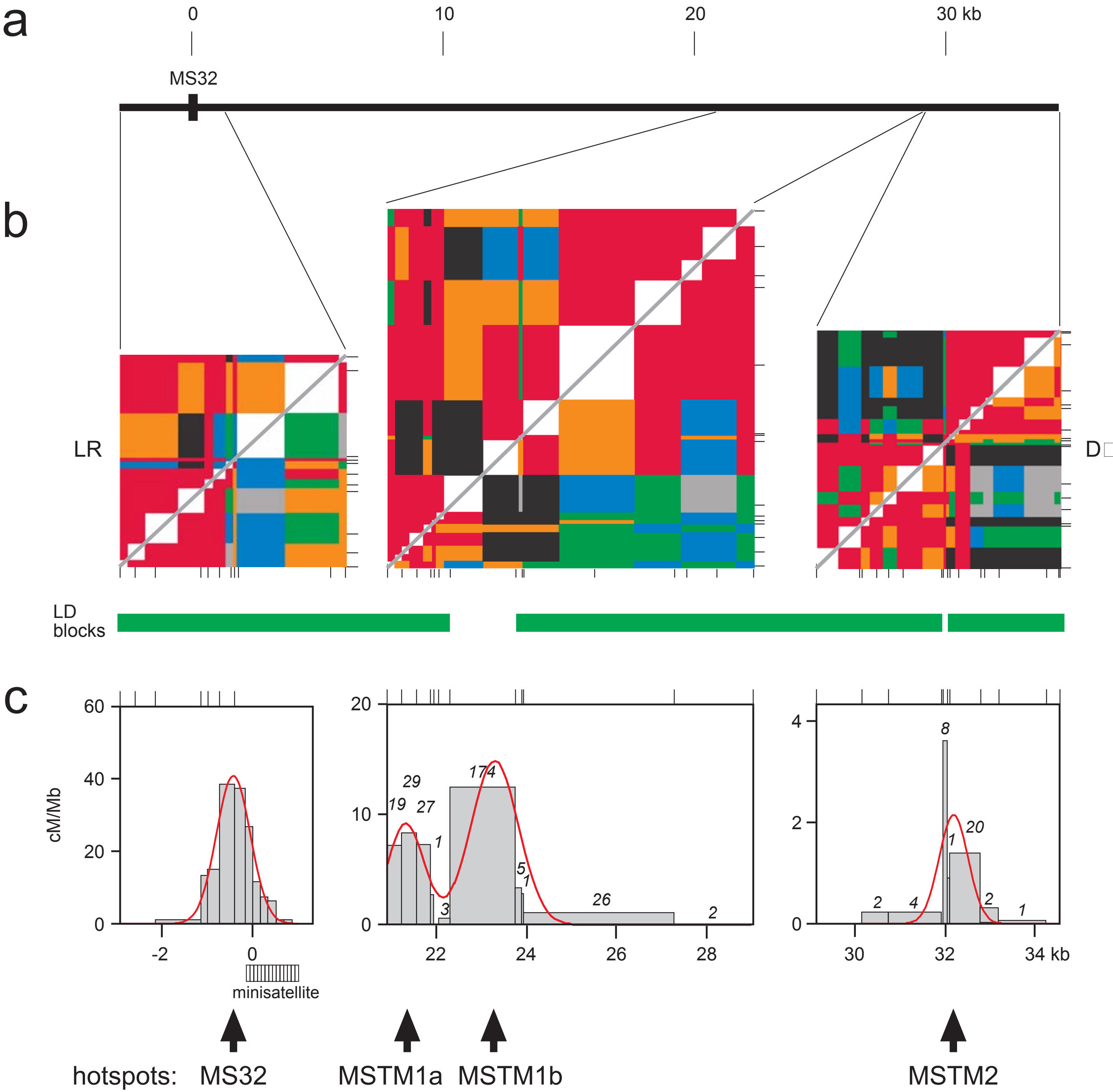

a
0
10
20
30 kb
MS32
b
LR
D□
LD
blocks
c
cM/Mb
60
40
20
0
-2
0
minisatellite
20
10
0
19
29
27
1
3
174
5
1
26
2
22
24
26
28
4
2
0
8
1
20
2
4
2
1
30
32
34 kb
hotspots: MS32
MSTM1a
MSTM1b
MSTM2

Figure 4

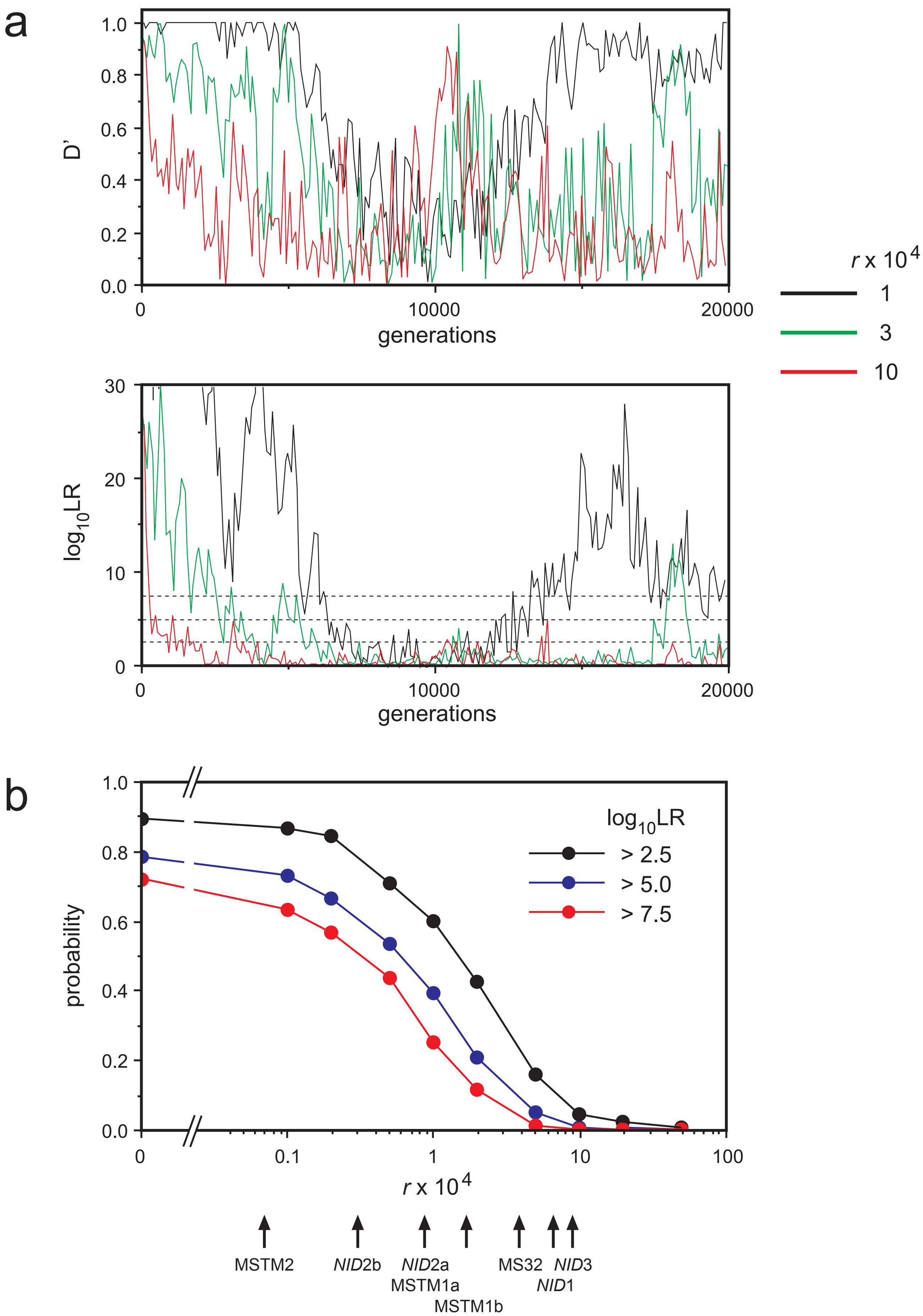
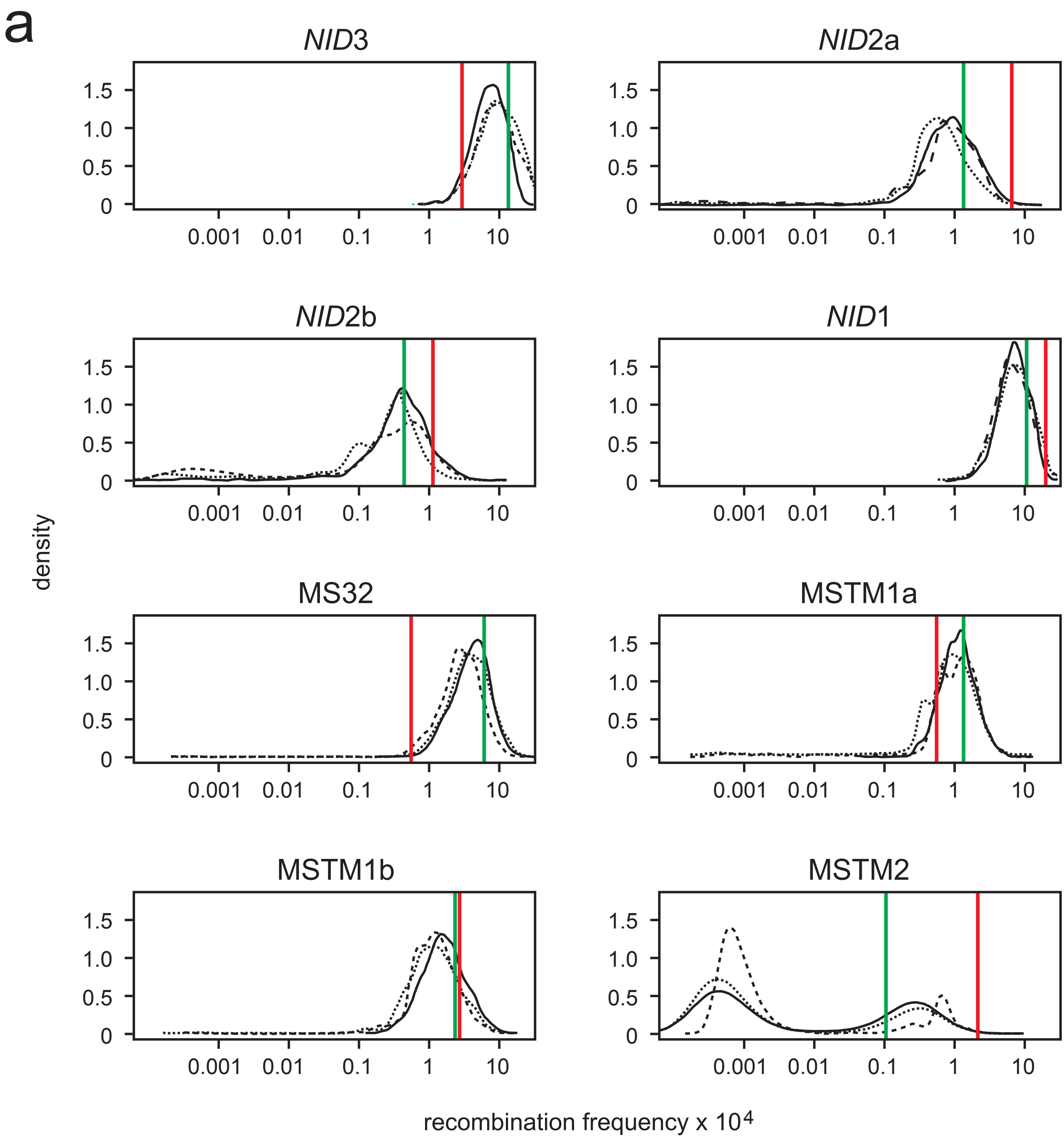

a
1.0
0.8
0.6
0.4
0.2
0.0
D'
0
10000
20000
generations
r x 10^4
1
3
10
30
20
10
0
log10LR
0
10000
20000
generations
b
1.0
0.8
0.6
0.4
0.2
0.0
probability
log10LR
> 2.5
> 5.0
> 7.5
0
0.1
1
10
100
r x 10^4
MSTM2
NID2b
NID2a
MSTM1a
MSTM1b
MS32
NID3
NID1

Figure 5



b

	*NID*3	*NID*2a	*NID*2b	*NID*1	MS32	MSTM1a	MSTM1b	MSTM2
model 1	0.094	**0.008**	0.176	**0.020**	**0.004**	0.222	0.504	**0.002**
model 2	0.076	**0.006**	0.110	0.056	**0.018**	0.290	0.238	**0.002**
model 3	0.088	**0.004**	0.072	0.092	**0.016**	0.474	0.280	**0.002**

Table 1 Properties of recombination hot spots in the *NID*-MS32 region.

hot spot:	***NID*3**	***NID*2a**	***NID*2b**	***NID*1**	**MS32**	**MSTM1a**	**MSTM1b**	**MSTM2**
no. men tested	3	2	2	7	3	2	2	3
no. crossovers typed	1094	302	107	1345	250	179	374	46
mean recombination frequency *r* x 10,000	9.0	0.85	0.30	6.7	3.9	0.88[a]	1.5	0.070
recombination frequency *r* range x 10,000	7.7, 9.2, 10.2	0.81, 0.90	0.29, 0.30	2.4-16.2	1.9, 2.0, 8.9	0.75[a], 1.0[a]	1.2, 1.8	<0.03[b], 0.04, 0.17
centre	-124,250[c]	-97,630	-96,260	-57,900	-400	21,320[a]	23,300[c]	32,230
centre location	Alu Jo element downstream of *NID*	single copy DNA in *NID* intron 12	single copy DNA in *NID* intron 12	Alu Yc5 element in *NID* intron 4	intergenic, in RTLV-LTR	intergenic, in single copy DNA	intergenic, in single copy DNA	intergenic, in single copy DNA
95% width, kb[d]	2.0[c]	1.4	1.1	1.5	1.5	1.6[a]	2.1[c]	1.3
mean peak cM/Mb	70[c]	10	4	70	40	9[a]	16[c]	0.9
historical peak cM/Mb	16	55[e]	-[e]	120	8.1	4.5	16	23
median D' across hot spot[f]	0.46	0.49	0.97	0.07	0.88	0.95	0.49	0.22
median $\log_{10}$LR across hot spot[g]	2.8	0.7	25.0	0.04	7.7	5.6	1.2	0.4

a, approximate values since part of the hot spot lies outside the test interval for sperm DNA analysis.

b, no MSTM2 crossovers were seen in this man, and the upper 95% C.I. is therefore given.

c, approximate values due to lack of markers near hot-spot centre.

d, width of hot spot within which 95% of crossovers occur, estimated from best-fit normal distributions.

e, hot-spots *NID*2a and *NID*2b were not resolved by coalescent analysis (**Fig. 1b**) and a single value is therefore given for historical recombination intensity.

f, median D' values for all pairwise comparisons of all markers in the haplotype block upstream of the hot spot with all markers in the downstream haplotype block, excluding markers with a minor allele frequency <0.15.

g, median $\log_{10}$ of the likelihood ratios in favour of linkage disequilibrium, for all marker comparisons between haplotype blocks flanking the hot spot.