Big data reveals fewer recombination hotspots than expected in human genome

Ziqian Hao[1,3], Haipeng Li[1,2,*]

[1]CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China.

[2]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China.

[3]University of Chinese Academy of Sciences, Beijing, China.

Short title: Recombination hotspots in human genome

*Corresponding author

E-mail: lihaipeng@picb.ac.cn (HL)

Abstract

Recombination is a major force that shapes genetic diversity. The inference accuracy of recombination rate is important and can be improved by increasing sample size. However, it has never been investigated whether sample size affects the distribution of inferred recombination activity along the genome, and the inference of recombination hotspots. In this study, we applied an artificial intelligence approach to estimate recombination rates in the UK10K human genomic data set with 7,562 genomes and in the OMNI CEU data set with 170 genomes. We found that the fluctuation of local recombination rate along the UK10K genomes is much smaller than that along the CEU genomes, and recombination activity in the UK10K genomes is also much less concentrated. The same phenomena were also observed when comparing UK10K with its two subsets with 200 and 400 genomes. In all cases, analyses of a larger number of genomes result in a more precise estimation of recombination rate and a less concentrated recombination activity with fewer recombination hotpots identified. Generally, UK10K recombination hotspots are about 2.93-14.25 times fewer than that identified in previous studies. By comparing the recombination hotspots of UK10K and its subsets, we found that the false inference of population-specific recombination hotspots could be as high as 75.86% if the number of sampled genomes is not super large. The results suggest that the uncertainty of estimated recombination rate is substantial when sample size is not super large, and more attention should be paid to accurate identification of recombination hotspots, especially population-specific recombination hotspots.

Author summary

We applied FastEPRR, an artificial intelligence method to estimate recombination rates in the UK10K data set with 7,562 genomes and established the most accurate human genetic map. By comparing with other human genetic maps, we found that analyses of a larger number of genomes result in a more precise estimation of recombination rate and a less concentrated recombination activity with fewer recombination hotpots identified. The false inference of population-specific recombination hotspots could be substantial if the number of sampled genomes is not super large.

Key words: Big data; artificial intelligence; recombination hotspot; UK10K; FastEPRR

Introduction

Recombination plays a fundamental role in evolution [1]. The study of recombination helps to decipher secrets of linkage disequilibrium [2-4], nucleotide diversity [5, 6], population admixture and differentiation [7-10], natural selection [11-14] as well as diseases [15, 16]. Therefore, much attention has been paid to the inference of recombination rate and hotspots. Since pedigree studies [17, 18] and sperm-typing studies [19, 20] are practically challenging, indirect approaches, such as population genetic methods [21-25], are remarkably useful. Population genetic methods estimate recombination rate from genetic variations in DNA sequences sampled from a population. The rationale behind these indirect approaches is that recombination events can be partially recovered during the time interval when the samples have evolved since the time of the most recent common ancestor.

It has been found that recombination rate varies along genome, and recombination hotspots have much higher recombination rate than other chromosomal regions [1]. Thus many efforts have been paid to identify recombination hotspots. It was found that recombination hotspots vary between humans and great apes [14, 26, 27]. By comparing European Americans and African Americans, some common hotspots as well as many population-specific hotspots were identified [28]. Other studies also revealed that different human populations have different hotspot intensities and distributions [29, 30]. There are variabilities of hotspots even among human individuals [31]. All these findings lead to an important conclusion that recombination hotspots are under rapid evolution. It was estimated that there are at least 15,000 hotspots [25] or more than 25,000 hotspots [32] in the human genome. Recently, up to 34,381 autosomal DNA double-strand breaks hotspots per individual were identified [33]. Thus it is widely accepted that recombination hotspot is a ubiquitous feature of the human genome.

According to coalescent theory, the expected coalescent tree length increases as well as the number of recombination events in the sample, when the sample size increases. Therefore, the inference accuracy of recombination rate can be improved by using data set with larger sample size [34-36]. It is important to investigate whether a recombination hotspot inferred in a small data set can also be identified in a larger data set. Although super-large sample size could theoretically facilitate a precise estimate of recombination rate, it is an enormous challenge to estimate

recombination rate under this circumstance. In this study, to precisely estimate recombination rate, we applied FastEPRR [34], an extremely fast artificial-intelligence-based software, to the UK10K data set (*n* = 7,562 genomes) [37] and the 1000 Genomes OMNI CEU data set (Utah residents with Northern and Western European ancestry) (*n* = 170 genomes) [38]. We found that the fluctuation of local recombination rate along the UK10K genomes is smaller than that along the CEU genomes, and recombination activity is less concentrated in UK10K than that in CEU. Recombination hotspots in UK10K are also much fewer than those in CEU. When comparing the results of UK10K with that of two UK10K subsets, the same phenomena were found, indicating that these are due to the difference in sample size instead of UK10K sample-specific factors. Therefore, the uncertainty of estimated recombination rate may be underestimated when identifying recombination hotspots, especially population-specific recombination hotspots.

Results

Using FastEPRR, the average ρ $(= 4N_e r)$ per Mb of 22 autosomes estimated from the UK10K data set is 4,015.19, where N_e is the effective population size, and r the recombination rate per generation. The average r in the same chromosomal regions in the 2019 deCODE family-based genetic map is 1.2806 cM/Mb [18], thus the estimated N_e for the UK10K data set is 78,385, which is larger than the estimated N_e from the 1000 Genomes OMNI CEU data set. This implies a recent human expansion. Then the population recombination rate ρ was converted to the recombination rate r. The Pairwise Pearson correlation coefficients were calculated between the UK10K map and each of other three maps at 5-Mb and 1-Mb scale (Fig 1 and 2), which are 2019 deCODE family-based genetic map, and the maps of 1000 Genomes OMNI CEU data set established by FastEPRR and LDhat [24]. The coefficients are 0.8375, 0.9130, and 0.7191 at 5-Mb scale and 0.6915, 0.8643, and 0.5672 at 1-Mb scale, respectively. Moreover, two genetic maps were established for two randomly selected subsets (UK10K-sub200 and UK10K-sub400) of UK10K (*n* = 200 and 400 genomes). The Pearson correlation coefficients between the UK10K and UK10K-sub200 maps are 0.9479 and 0.8968 at 5- and 1-Mb scales, respectively. Those between the UK10K and UK10K-sub400 maps are 0. 9619 and 0.9292 at 5- and 1-Mb scales, respectively (Fig 1 and 2). Thus, the UK10K map is highly correlated with other human genetic

maps.

Recombination rates over different physical scales show that smaller scale has more variability and discrimination (Fig 3), which is consistent with the previous finding [1]. Therefore, to detect the recombination hotspots and compare the number of hotspots among different genetic maps, the recombination rate at 5-kb scale was used. The recombination rates for chromosome 1 were shown as examples (Fig 4). Two regions (chr1:15,425,001-15,430,000 and chr1:236,680,001-236,685,000) were found with extremely high recombination rates, in accordance with the corresponding peaks in other five genetic maps (Fig 4). These two regions harbor seven PRDM9 binding peaks [39], which is a characteristic feature of recombination hotspot [14, 40-42]. The human leukocyte antigen (HLA) region is one of the most polymorphic regions in the human genome [43]. Attention has been paid to recombination of this region [44-46]. Multiple recombination hotspots within the HLA region were observed in the six genetic maps (Fig 5). Therefore, as what previously reported [25, 32, 33], recombination hotspots are indeed a ubiquitous feature of the human genome.

However, the UK10K genetic map is less zigzag than other five human genetic maps (Fig 4), the sample sizes of which are large but much less than that of UK10K. The variances of recombination rate along the human genome were calculated from different genetic maps at 5-kb scale (4.73 cM^2/Mb^2 in UK10K; 11.11 cM^2/Mb^2 in UK10K-sub400; 15.53 cM^2/Mb^2 in UK10K-sub200; 9.14 cM^2/Mb^2 in OMNI CEU FastEPRR; 18.77 cM^2/Mb^2 in OMNI CEU LDhat; 26.04 cM^2/Mb^2 in deCODE). As expected, the variance of recombination rate along the UK10K genomes is the smallest one. Then the proportion of total recombination in various percentages of sequence was plotted and investigated (Fig 6A). Overall, recombination activity in the UK10K genomes is much less concentrated than that in other human genomes. For the latter, we found that 57.60 – 81.90% of crossover events occur in 10% of the sequence, in accordance with the previous findings [18, 32, 34]. However, only 39.30% of crossover events occur in 10% of the sequence in UK10K, which is strikingly lower than that in the previous studies [18, 32, 34]. This difference results from the precisely estimated recombination rate based on the UK10K data set with super-large sample size.

Recombination hotspots were then identified by two different methods. LDhot is a popular method to detect recombination hotspots [25, 26, 32]. However, it is impossible to apply LDhot to

the UK10K data set because of computational burden. Thus, a revised approach was adopted from LDhot. Recombination hotspot was detected by comparing the recombination rate of a 5-kb sliding window with the average recombination rate in its surrounding 200-kb region. The window was recognized as a recombination hotspot if its recombination rate was x times higher than the average, where x is the threshold. For a wide-range threshold, the number of detected recombination hotspots in OMNI CEU (LDhat), UK10K-sub200, and UK10K-sub400 is much larger than that in UK10K (Fig 7). Following one of the standards used in the previous study [25], we required estimated increase in rate by a factor of at least 5 over local background (i.e. $x = 5$). Then we identified 5,115 recombination hotspots in the UK10K genomes, which is about 2.93-6.72 times less than the previously identified [25, 32, 33]. The number of detected recombination hotspots in OMNI CEU (LDhat), UK10K-sub200 is 21,714, and 20,478, respectively, which is consistent with the previous estimates [25, 32, 33]. The number of shared hotspots is 4,099 among the four genetic maps. Second, recombination hotspot was detected when the recombination rate of the sliding window was at least 10 times larger than the genome average [17, 18]. The number of identified recombination hotspots in UK10K, UK10K-sub400, UK10K-sub200, and OMNI CEU (LDhat) are 2,413, 6,075, 7,757, and 12,121, respectively. The number of shared hotspots is 1,877 among the four genetic maps in this case. The UK10K recombination hotspots are about 6.22-14.25 times fewer than the previously identified [25, 32, 33]. In summary, analyses of a larger number of genomes result in a more precise estimation of recombination rate with fewer recombination hotpots identified.

We next wondered whether the inference of population-specific recombination hotspots could be partially affected by the uncertainty of estimated recombination rate due to the limited number of sampled genomes. The UK10K data set provides us a great opportunity to investigate this issue. False population-specific recombination hotspots can be recognized by investigating two different data sets sampled from the same population, and we suspect that fewer samples may result in more false population-specific recombination hotspots. This is well-supported by the Venn diagram showing overlap of recombination hotspots identified in three UK10K data sets (Fig 8). When pairing the two data sets with 200 and 7,562 genomes, and identifying recombination hotspots by comparing the local recombination rate with the average recombination rate in its surrounding 200-kb region (Fig 8A), 15,534 and 171 false population-specific recombination

hotspots in total are observed, respectively. The percentage of false population-specific recombination hotspots among the hotspots identified is 75.86 and 3.34, respectively. When pairing the two data sets with 400 and 7,562 genomes, 11,111 and 170 false population-specific recombination hotspots in total are observed, respectively (Fig 8A). The corresponding percentage is 69.20 and 3.32, respectively. When recombination hotspots were identified by comparing the local recombination rate with the genome-wide average recombination rate (Fig 8B), the similar results were obtained. Therefore, the false inference of population-specific recombination hotspots could be substantial if the number of sampled genomes is not super large.

Discussion

The fewer recombination hotspots in the human genome were revealed by the UK10K data set, which can be explained by the super-large sample size of UK10K. The number of sampled genomes is 170 and 7,562 in the OMNI CEU and UK10K data sets, respectively. Under the constant population size model, the expected tree length of the UK10K genomes is 1.67 times larger than that of the CEU genomes. Therefore, the number of recombination events occurred in UK10K is also 1.67 times more than that in CEU. Based on coalescent simulations (see Materials and Methods), if an exponential growth scenario is considered, the number of recombination events in UK10K is 3.32 times more than that in CEU. For a bottleneck scenario, the number of recombination events in UK10K is 2.06 times more than that in CEU. More recombination events in the UK10K genomes result in a more precise UK10K estimation of recombination rate with fewer recombination hotpots identified. Moreover, coalescent simulations demonstrate that the relative standard deviations of tree length of independent loci of UK10K in the three models are 1.66, 2.80, and 2.05 times lower than those of CEU, respectively. It indicates again that the recombination rate estimated with fewer samples is more variable along the genome than that with more samples, which corresponds with our results.

This study is the first attempt to identify recombination hotspots using such a super-large population genetic data set. By processing genomes with sample size in four orders of magnitude, we also demonstrate that artificial intelligence approaches [35, 47-49] benefit population genetic analysis. The results suggest that the uncertainty of estimated recombination rate could be

underestimated when identifying recombination hotspots in previous studies. Since this uncertainty is also largely affected by demography [50], we suggest that demography should be considered when detecting recombination hotspots. This could be done by providing demographic parameters to FastEPRR [34], or revising the simulation code of LDhot [25, 26] and LDpop [50]. If demography is unknown, this effect could be partially alleviated by analyzing more samples.

Materials and Methods

How to build genetic map and to fine-tune FastEPRR

Recently, we developed an extremely fast software (FastEPRR) which utilized an artificial intelligence method based on the finite-site model to estimate recombination rate from single nucleotide polymorphism data [34]. FastEPRR first calculates compact folded site frequency spectrum (SFS) [51] and four recombination-related statistics for sliding windows. Next, training set is generated by simulating data using a modified version of *ms* simulator [52], under the constraints of compact folded SFS and a value set of ρ. The gamboost package [53] is used to fit the training set and build the regression model. Recombination rate can be then inferred given the summary statistics observed in the real data.

In this study we fine-tuned the FastEPRR software. For optimization of speed, for-loops were replaced by "apply" family of functions because R language is good at vectorized operations. Input and output operations were also minimalized to speed up. Moreover, users can set the values of ρ when generating the training samples, choose the number of independent training samples to generate given a specific value of ρ, and decide whether to estimate confidence interval according to their needs. All these improvements and updates are assembled in FastEPRR 2.0. Users can also provide an arbitrary demographic model, so the effect of demography can be easily counted.

In total, the genomic variants of 3,781 unrelated individuals (n = 7,562 genomes) from the UK10K data set were taken into consideration. To build genetic maps, the window size was set as 10 kb, and the step length was 5 kb. Indels were removed. Windows were excluded if these overlapped with the known sequencing gaps in human reference genome (hg19) or the number of non-folded-singleton SNPs was less than 10. Default value set of ρ(0, 0.5, 1, 2, 5, 10, 20, 40, 70, 110, 170, 180, 190, 200, 220, 250, 300 and 350) was used to simulate the training sets. There were

fewer than 0.5% windows, the estimated value ρ of which was large than 350. When the estimated value is beyond the training set boundary, linear extrapolation is employed, which might make the results imprecise. Thus default values of ρ were adjusted and FastEPRR was re-ran for these windows until the estimated values fell into the range of the given set of ρ. The fore-described method in FastEPRR was used to evaluate the effects for variable recombination rates within windows.

Sub-samples (UK10K-sub200 and UK10K-sub400) were randomly chosen from UK10K data set. The number of genomes in the two sub-samples was 200 and 400, respectively. Seven individuals (14 genomes) were shared in both sub-samples. The lists of chosen individual IDs are available on request.

Coalescent simulations

The expected tree length under the constant size model is well known [54-56]. Simulations under the exponential growth and bottleneck models were performed according to the procedures described previously [57, 58]. For exponential growth model, we assumed $N_0/N_1 = 100$, where N_0 and N_1 are the current and ancestral effective population size, respectively. The expansion duration is 0.1, where time is rescaled as one unit for $2N_0$ generations as usual. For bottleneck model, we assumed $N_0/N_1 = 10$, where N_0 and N_1 are the current effective population size, and the effective population size during the bottleneck, respectively. The ancestral effective population is equal to N_0. The time back to the bottleneck is 0.1, and the duration of bottleneck 0.1. For each demographic scenario, 10^6 simulations were performed to estimate the length of coalescent tree.

Data and software availability

The UK10K data (3,781 unrelated individuals; or *n* = 7,562 genomes) are kindly shared by the UK10K Project Consortium. FastEPRR 2.0 is integrated in the eGPS cloud at www.egps-software.org (Yu, et al. 2019). The desktop version and the genetic maps established in this study are freely available on the institutional website www.picb.ac.cn/evolgen/softwares/.

Acknowledgments

We thank Dr. Yi-Hsuan Pan for the comments, Peng-Yuan Du for preparing figures, and the

UK10K Project Consortium for sharing the data. This work was supported by grants from the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040800), the National Natural Science Foundation of China (no. 91731304). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Coop G, Przeworski M. An evolutionary view of human recombination. Nat Rev Genet. 2007;8(1):23-34. doi: 10.1038/nrg1947. PMID: 17146469.
2. Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor Appl Genet. 1968;38(6):226-31. doi: 10.1007/bf01245622. PMID: 24442307.
3. Ohta T, Kimura M. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics. 1971;68(4):571-80. PMID: 5120656.
4. Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nat Rev Genet. 2002;3(4):299-309. doi: 10.1038/nrg777. PMID: 11967554.
5. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-wide patterns and properties of de novo mutations in humans. Nat Genet. 2015;47(7):822-6. doi: 10.1038/ng.3292. PMID: 25985141.
6. Kong A, Thorleifsson G, Frigge ML, Masson G, Gudbjartsson DF, Villemoes R, et al. Common and low-frequency variants associated with genome-wide recombination rate. Nat Genet. 2014;46(1):11-6. doi: 10.1038/ng.2833. PMID: 24270358.
7. Payseur BA, Rieseberg LH. A genomic perspective on hybridization and speciation. Mol Ecol. 2016;25(11):2337-60. PMID: 26836441.
8. Keinan A, Reich D. Human population differentiation is strongly correlated with local recombination rate. PLoS Genet. 2010;6(3):e1000886. doi: 10.1371/journal.pgen.1000886. PMID: 20361044.
9. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 2009;5(6):e1000519. doi: 10.1371/journal.pgen.1000519. PMID: 19543370.
10. Wegmann D, Kessner DE, Veeramah KR, Mathias RA, Nicolae DL, Yanek LR, et al. Recombination rates in admixed individuals identified by ancestry-based inference. Nat Genet. 2011;43(9):847-53. doi: 10.1038/ng.894. PMID: 21775992.
11. Schumer M, Xu CL, Powell DL, Durvasula A, Skov L, Holland C, et al. Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science. 2018;360(6389):656-9. doi: 10.1126/science.aar3684. PMID: 29674434.
12. Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. Recombination: the good, the bad and the variable. Philos Trans R Soc B-Biol Sci. 2017;372(1736):20170279. doi: 10.1098/rstb.2017.0279. PMID: 29109232.
13. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. Science. 2011;331(6019):920-4. doi: 10.1126/science.1198878. PMID: 21330547.

14. Stevison LS, Woerner AE, Kidd JM, Kelley JL, Veeramah KR, McManus KF, et al. The time scale of recombination rate evolution in Great Apes. Mol Biol Evol. 2016;33(4):928-45. doi: 10.1093/molbev/msv331. PMID: 26671457.
15. Hussin JG, Hodgkinson A, Idaghdour Y, Grenier JC, Goulet JP, Gbeha E, et al. Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nat Genet. 2015;47(4):400-4. doi: 10.1038/ng.3216. PMID: 25685891.
16. Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. Trends Genet. 2002;18(1):19-24. doi: Doi 10.1016/S0168-9525(01)02550-1. PMID: 11750696.
17. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010;467(7319):1099-103. doi: 10.1038/nature09525. PMID: 20981099.
18. Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science. 2019;363(6425):eaau1043. doi: 10.1126/science.aau1043. PMID: 30679340.
19. Webb AJ, Berg IL, Jeffreys A. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. Proc Natl Acad Sci USA. 2008;105(30):10471-6. doi: 10.1073/pnas.0804933105. PMID: 18650392.
20. Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proc Natl Acad Sci USA. 2015;112(7):2109-14. doi: 10.1073/pnas.1416622112. PMID: 25646453.
21. Hudson RR. Two-locus sampling distributions and their application. Genetics. 2001;159(4):1805-17. PMID: 11779816.
22. Fearnhead P, Donnelly P. Estimating recombination rates from population genetic data. Genetics. 2001;159(3):1299-318. PMID: 11729171.
23. McVean G, Awadalla P, Fearnhead P. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics. 2002;160(3):1231-41. PMID: 11901136.
24. Auton A, McVean G. Recombination rate estimation in the presence of hotspots. Genome Res. 2007;17(8):1219-27. doi: 10.1101/gr.6386707. PMID: 17623807.
25. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004;304(5670):581-4. doi: DOI 10.1126/science.1092500. PMID: 15105499.
26. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, et al. Comparison of fine-scale recombination rates in humans and chimpanzees. Science. 2005;308(5718):107-11. doi: 10.1126/science.1105322. PMID: 15705809.
27. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, et al. Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet. 2005;37(4):429-34. doi: 10.1038/ng1529. PMID: 15723063.
28. Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, et al. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet. 2004;36(7):700-6. doi: 10.1038/ng1376. PMID: 15184900.
29. Evans DM, Cardon LR. A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. Am J Hum Genet. 2005;76(4):681-7. doi: Doi 10.1086/429274. PMID: 15719321.
30. Conrad DF, Jakobsson M, Coop G, Wen XQ, Wall JD, Rosenberg NA, et al. A worldwide survey

of haplotype variation and linkage disequilibrium in the human genome. Nat Genet. 2006;38(11):1251-60. doi: 10.1038/ng1911. PMID: 17057719.

31. Neumann R, Jeffreys AJ. Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation. Hum Mol Genet. 2006;15(9):1401-11. doi: 10.1093/hmg/ddl063. PMID: 16543360.
32. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science. 2005;310(5746):321-4. doi: 10.1126/science.1117196. PMID: 16224025.
33. Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. Recombination initiation maps of individual human genomes. Science. 2014;346(6211):1256442. doi: 10.1126/science.1256442. PMID: 25395542.
34. Gao F, Ming C, Hu WJ, Li HP. New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era. G3-Genes Genomes Genet. 2016;6(6):1563-71. doi: 10.1534/g3.116.028233. PMID: 27172192.
35. Lin K, Futschik A, Li HP. A fast estimate for the population recombination rate based on regression. Genetics. 2013;194(2):473-84. doi: 10.1534/genetics.113.150201. PMID: 23589457.
36. Sall T, Nilsson NO. The robustness of recombination frequency estimates in intercrosses with dominant markers. Genetics. 1994;137(2):589-96. PMID: 8070668.
37. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526(7571):82-9. PMID: 26367797.
38. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65. doi: 10.1038/nature11632. PMID: 23128226.
39. Altemose N, Noor N, Bitoun E, Tumian A, Imbeault M, Chapman JR, et al. A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. elife. 2017;6:e28383. doi: 10.7554/e.Life.28383. PMID: 29072575.
40. Parvanov ED, Petkov PM, Paigen K. Prdm9 controls activation of mammalian recombination hotspots. Science. 2010;327(5967):835-. doi: 10.1126/science.1181495. PMID: 20044538.
41. Hochwagen A, Marais GAB. Meiosis: a PRDM9 guide to the hotspots of recombination. Curr Biol. 2010;20(6):R271-R4. doi: 10.1016/j.cub.2010.01.048. PMID: 20334833.
42. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, et al. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science. 2010;327(5967):836-40. doi: 10.1126/science.1183439. PMID: 20044539.
43. Wu RG, Li HX, Peng D, Li R, Zhang YM, Hao B, et al. Revisiting the potential power of human leukocyte antigen (HLA) genes on relationship testing by massively parallel sequencing-based HLA typing in an extended family. J Hum Genet. 2019;64(1):29-38. doi: 10.1038/s10038-018-0521-0. PMID: 30348993.
44. Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet. 2001;29(2):217-22. doi: DOI 10.1038/ng1001-217. PMID: 11586303.
45. Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. Am J Hum Genet. 2002;71(4):759-76. doi: Doi 10.1086/342973. PMID: 12297984.
46. Miretti MM, Walsh EC, Ke XY, Delgado M, Griffiths M, Hunt S, et al. A high-resolution

linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. Am J Hum Genet. 2005;76(4):634-46. doi: Doi 10.1086/429393. PMID: 15747258.

47. Lin K, Li HP, Schlotterer C, Futschik A. Distinguishing positive selection from neutral evolution: Boosting the performance of summary statistics. Genetics. 2011;187(1):229-44. doi: 10.1534/genetics.110.122614. PMID: 21041556.
48. Flagel L, Brandvain Y, Schrider DR. The unreasonable effectiveness of convolutional neural networks in population genetic inference. Mol Biol Evol. 2019;36(2):220-38. doi: 10.1093/molbev/msy224. PMID: 30517664.
49. Pavlidis P, Jensen JD, Stephan W. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics. 2010;185(3):907-22. doi: 10.1534/genetics.110.116459. PMID: 20407129.
50. Kamm JA, Spence JP, Chan J, Song YS. Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. Genetics. 2016;203(3):1381-99. doi: 10.1534/genetics.115.184820. PMID: 27182948.
51. Li HP, Stephan W. Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. Genetics. 2005;171(1):377-84. doi: 10.1534/genetics.105.041368. PMID: 15972464.
52. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002;18(2):337-8. PMID: 11847089.
53. Hothorn T, Buehlmann P, Kneib T, Schmid M, Hofner B. mboost: Model-Based Boosting, R package version 2.9-1, https://CRAN.R-project.org/package=mboost. 2018.
54. Kingman JFC. The coalescent. Stoch Process Their Appl. 1982;13(3):235-48.
55. Kingman JFC. On the genealogy of large populations. J Appl Probab. 1982;19:27-43. PMID: 11338422.
56. Tajima F. Evolutionary relationship of DNA-sequences in finite populations. Genetics. 1983;105(2):437-60. PMID: 6628982.
57. Li HP, Stephan W. Inferring the demographic history and rate of adaptive substitution in Drosophila. PLoS Genet. 2006;2(10):1580-9. doi: 10.1371/journal.pgen.0020166. PMID: 17040129.
58. Griffiths RC, Tavare S. Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc London, Ser B. 1994;344(1310):403-10. doi: DOI 10.1098/rstb.1994.0079. PMID: 7800710.

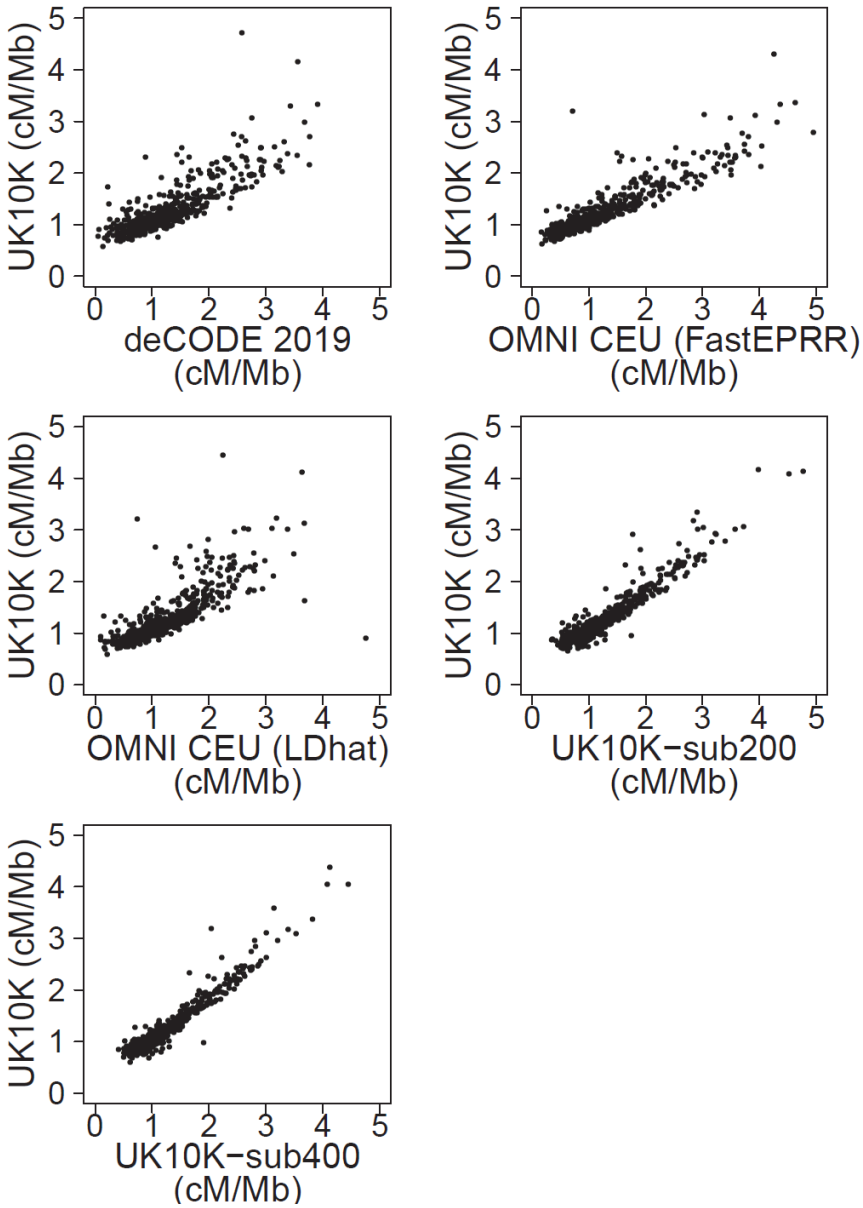

Fig 1. Correlation between the UK10K genetic map and five other maps at 5-Mb scale.

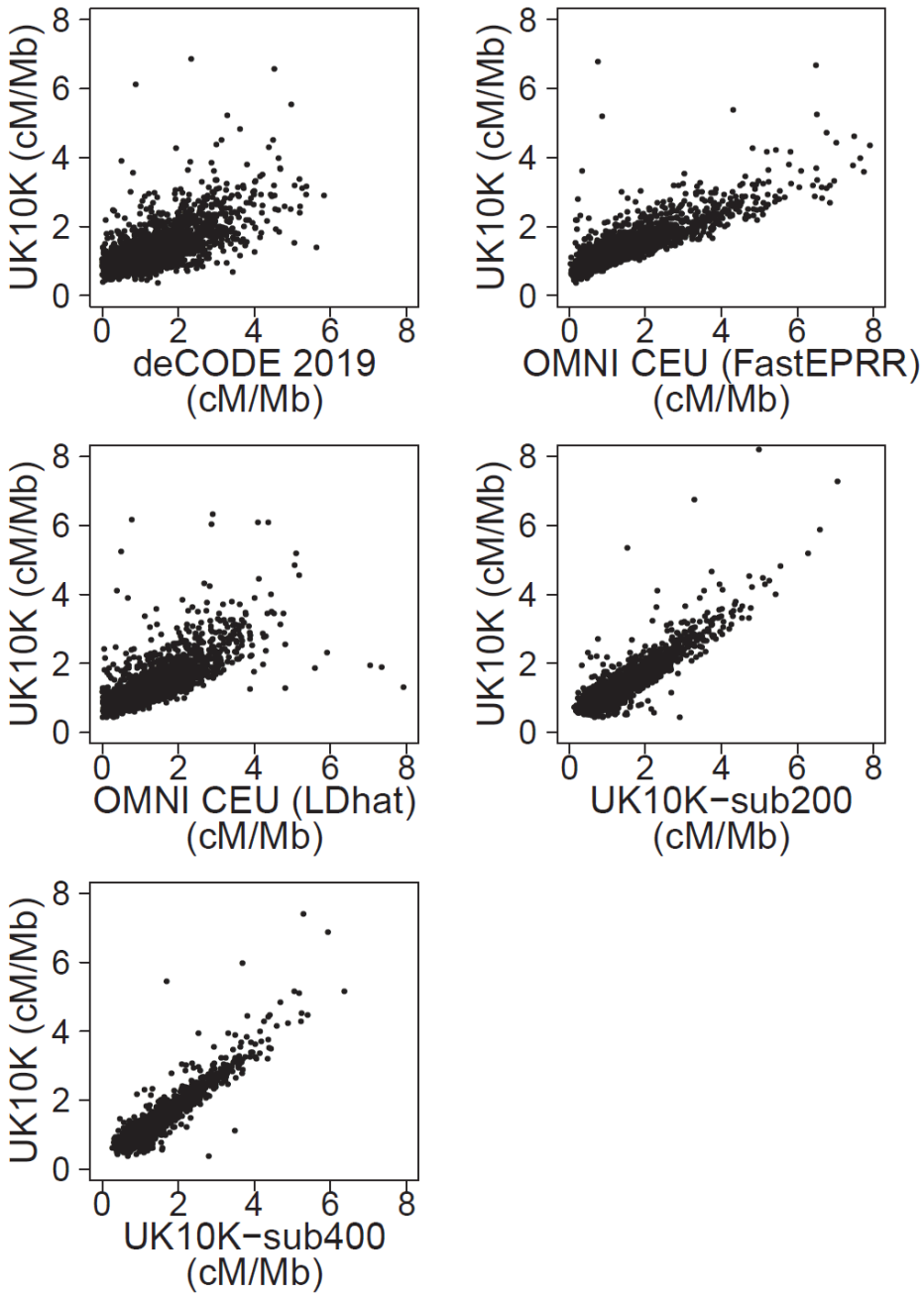

Fig 2. Correlation between the UK10K genetic map and five other maps at 1-Mb scale.

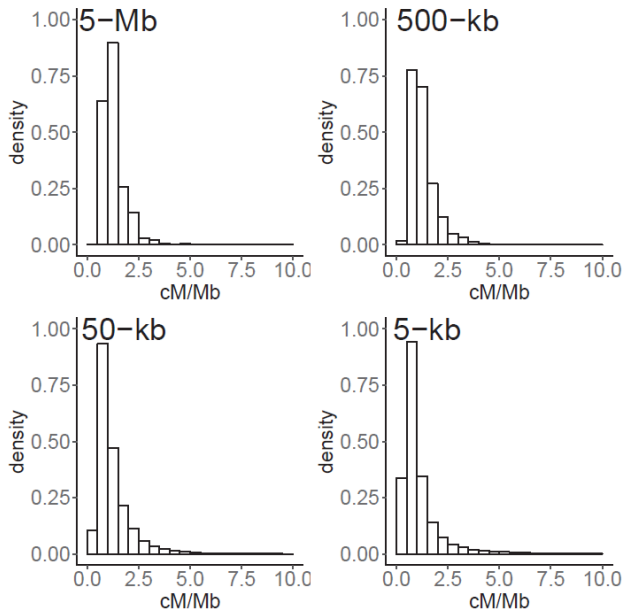

Fig 3. Recombination rates of the human autosomal genome over different physical scales.

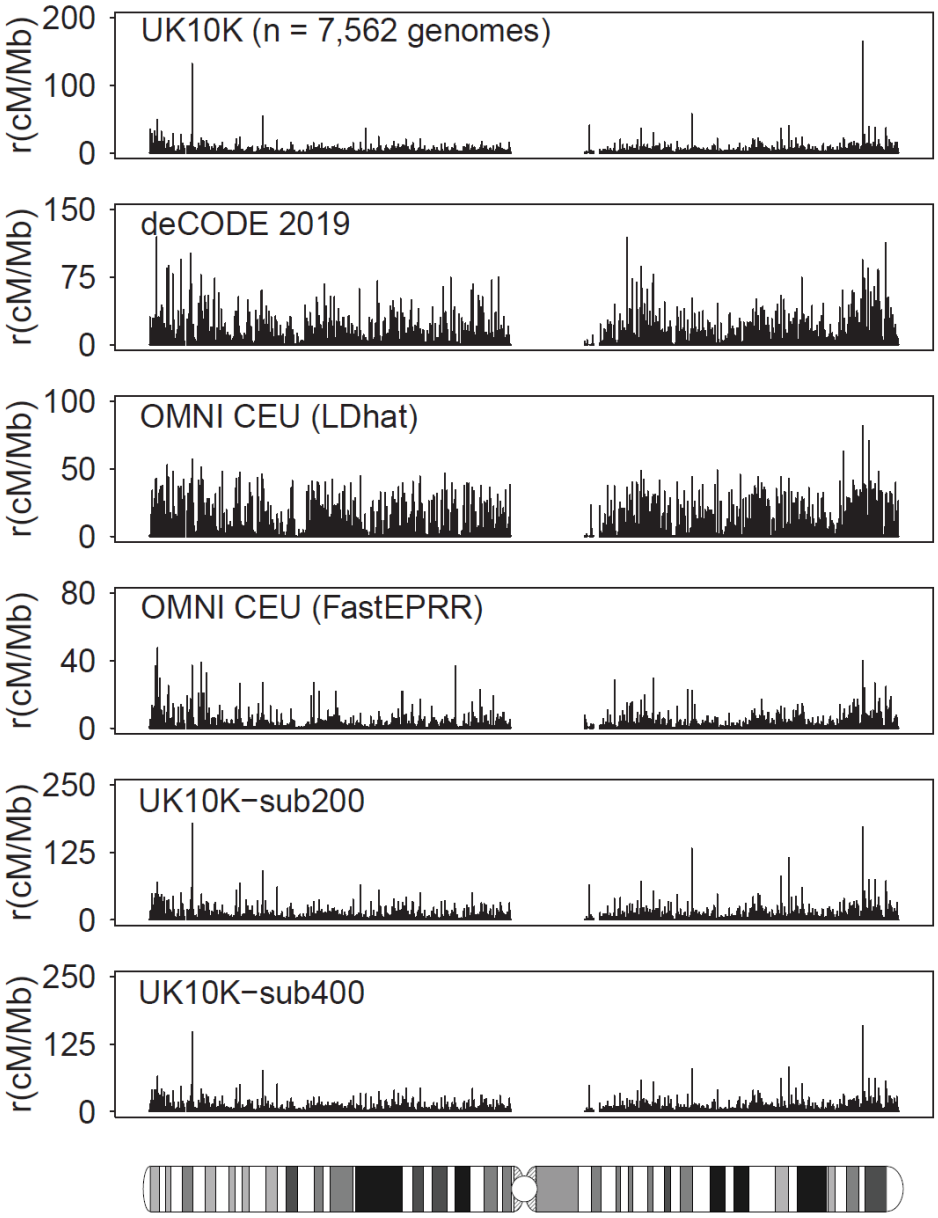

Fig 4. Recombination rates of human chromosome 1 for six genetic maps at 5-kb scale. The cartoon at the bottom is a visualization of the chromosome.

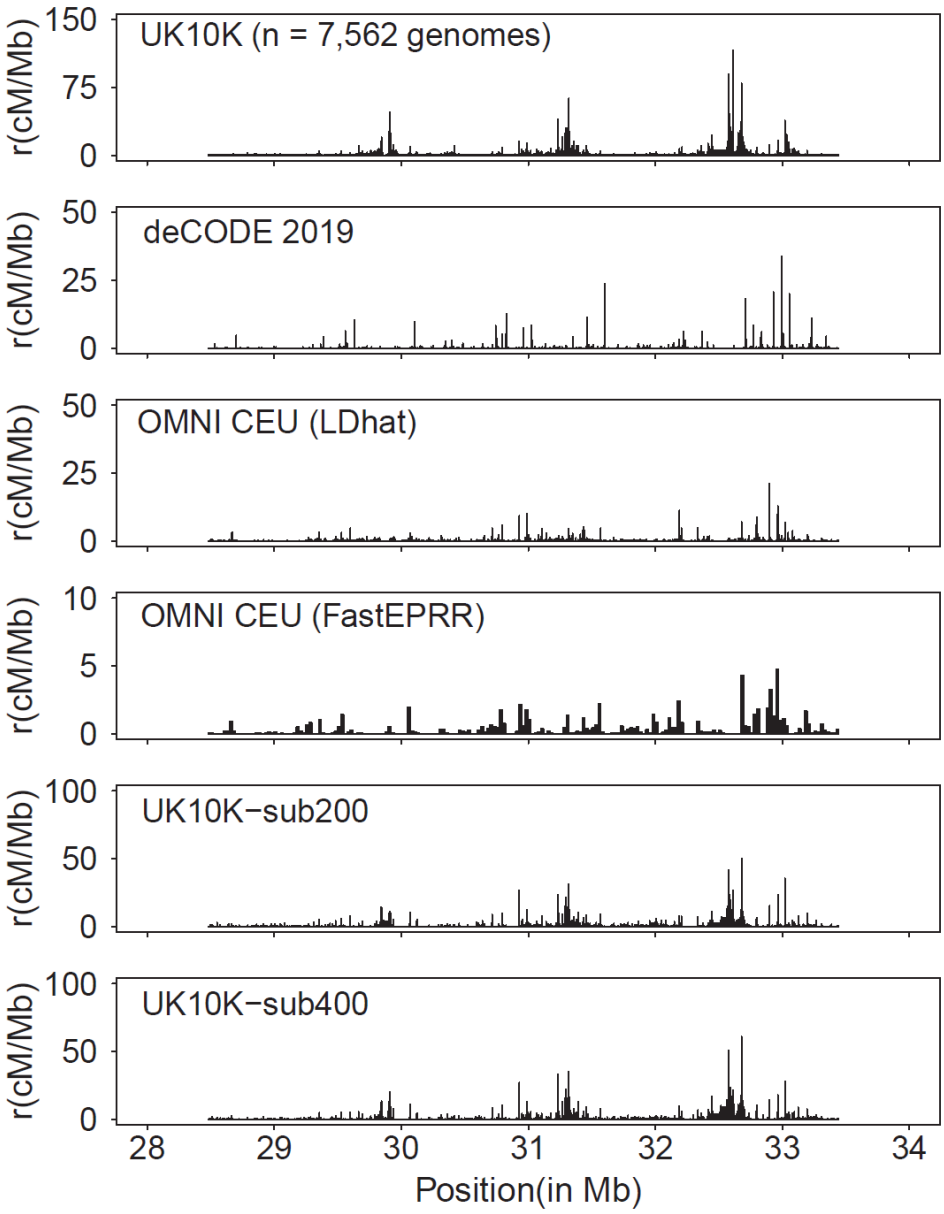

Fig 5. Recombination rates of human HLA region for 6 genetic maps at 5-kb scale.

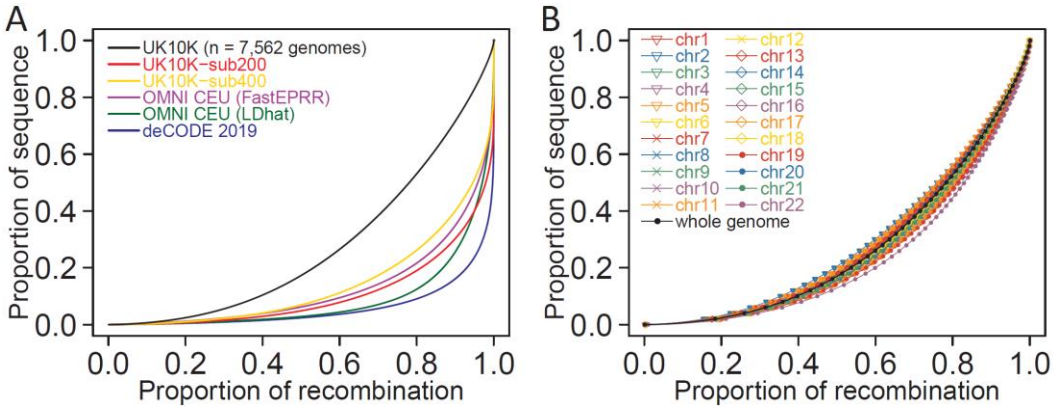

Fig 6. Proportion of recombination.

(A) Proportion of recombination of different genetic maps.

(B) Proportion of recombination of different chromosomes from the UK10K data set. Each colored line represents one chromosome, while the black line denotes the autosomal genome.

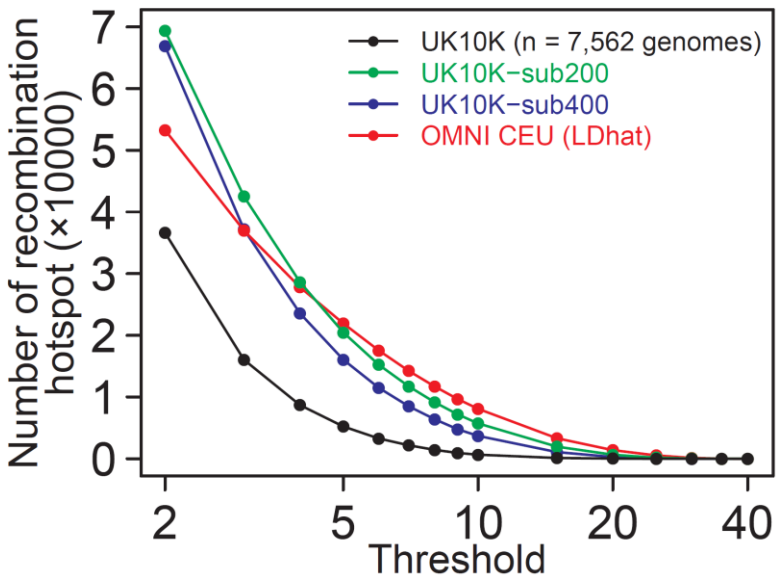

Fig 7. Number of recombination hotspots revealed at different threshold.

Recombination hotspots were identified by comparing the local recombination rate with the average recombination rate in its surrounding 200-kb region.

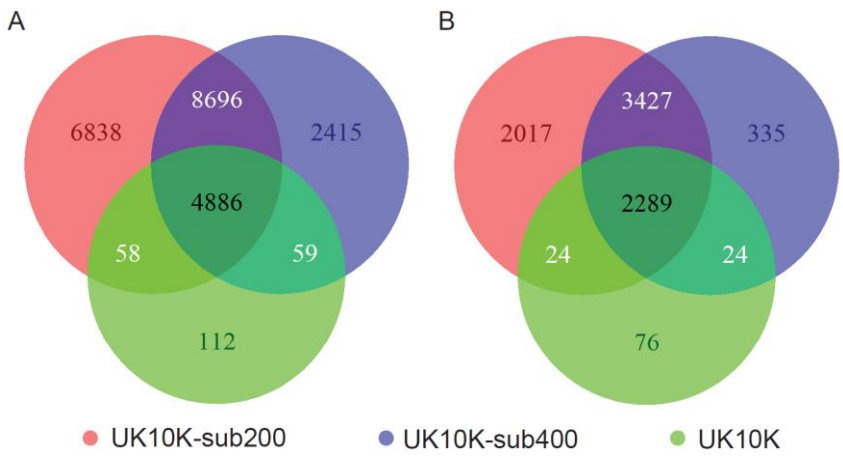

Fig 8. Venn diagram showing overlap of recombination hotspots identified in three UK10K data sets.

(A) Recombination hotspots were identified by comparing the local recombination rate (a threshold of 5) with the average recombination rate in its surrounding 200-kb region.

(B) Recombination hotspots were identified by comparing the local recombination rate (a threshold of 10) with the genome-wide average recombination rate.