Title

The recombination landscape of the Khoe-San likely represents the upper limits of recombination divergence in humans

Journal

Genome Biology, 23(1)

ISSN

1474-760X

Authors

van Eeden, Gerald
Uren, Caitlin
Pless, Evlyn
et al.

Publication Date

2022

DOI

10.1186/s13059-022-02744-5

Peer reviewed

Genome Biology

RESEARCH

Open Access

The recombination landscape of the Khoe-San likely represents the upper limits of recombination divergence in humans

Gerald van Eeden[1], Caitlin Uren[1,2], Evlyn Pless[3], Mira Mastoras[3], Gian D. van der Spuy[1,2,4], Gerard Tromp[1,2,4], Brenna M. Henn[3] and Marlo Möller[1,2*]

*Correspondence: marlom@sun.ac.za

[1] DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[2] Centre for Bioinformatics and Computational Biology, Stellenbosch University, Stellenbosch 7602, South Africa
[3] Department of Anthropology, Center for Population Biology and the Genome Center, University of California (UC) Davis, Davis, CA, USA
[4] SAMRC-SHIP South African Tuberculosis Bioinformatics Initiative (SATBBI), Center for Bioinformatics and Computational Biology, Cape Town, South Africa

Abstract

Background: Recombination maps are important resources for epidemiological and evolutionary analyses; however, there are currently no recombination maps representing any African population outside of those with West African ancestry. We infer the demographic history for the Nama, an indigenous Khoe-San population of southern Africa, and derive a novel, population-specific recombination map from the whole genome sequencing of 54 Nama individuals. We hypothesise that there are no publicly available recombination maps representative of the Nama, considering the deep population divergence and subsequent isolation of the Khoe-San from other African groups.

Results: We show that the recombination landscape of the Nama does not cluster with any continental groups with publicly available representative recombination maps. Finally, we use selection scans as an example of how fine-scale differences between the Nama recombination map and the combined Phase II HapMap recombination map can impact the outcome of selection scans.

Conclusions: Fine-scale differences in recombination can meaningfully alter the results of a selection scan. The recombination map we infer likely represents an upper bound on the extent of divergence we expect to see for a recombination map in humans and would be of interest to any researcher that wants to test the sensitivity of population genetic or GWAS analysis to recombination map input.

Keywords: Recombination rate, Recombination map, Genetic map, Khoe-San, Selection scan

Background

Recombination enables the evolution of complex traits by shuffling novel genetic variants, brought about by mutation, into new combinations with existing alleles from varying genomic origins [1]. Due to the evolutionary significance of recombination, many implementations of software packages that infer the recombination rate have been

developed. Some of these packages rely on inferring past recombination events by analysing pedigrees [2], by detecting changes in ancestry [3, 4] or by the boundaries of blocks of identity by descent (IBD) [5]. These methods require large sample numbers (> 2000) to accurately infer the recombination rate at fine-scales. Other packages use linkage disequilibrium (LD) [6] or derivatives thereof, e.g. summary statistics [7], to infer recombination into the very distant past and require fewer individuals to infer recombination at fine-scales. LD-based recombination maps, however, are strongly influenced by past demographic events, e.g. population bottlenecks [8]. Recombination inference software that is aware of changes in the effective population size (N_e) of a population, such as pyrho [9], can be used to mitigate this effect.

The rate of recombination varies between species [10], between populations within species [9, 11] and even among individuals [12]. The recombination rate across the genome is generally expressed as a ratio of genetic distance and physical distance, known as a recombination map. It has been shown that at low resolutions (> 1 Mb), population specific recombination maps are fairly similar [13] and at high resolutions they correlate according to continental levels of population differentiation [11]. For instance, the pedigree-based deCODE [14] map, based on the Icelandic population, correlates better at fine scales to the linkage-disequilibrium-based (LD-based) HapMap II [15] map of the CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) than it does to the HapMap II map of the YRI (Yoruba in Ibadan, Nigeria) [7, 14]. Many population-specific recombination maps have been inferred to date, but none have been inferred for any southern African populations [16] and researchers studying these populations have had to use available maps that might not suit their analysis.

In this manuscript, we present a novel recombination map for the Nama—an indigenous population of southern Africa [17] that forms part of a larger group of geographically close and culturally related individuals known collectively as the "Khoe-San". The Khoe-San are reported to have the most divergent lineages of any other living population [18–22], and it is believed that they have largely remained isolated until ~2000 years ago [17, 18, 23]. Therefore, a recombination map for this population may be very different at fine scales compared to recombination maps that have been inferred for other populations. The Khoe-San also contribute a significant ancestral component (15–75%) to admixed southern African groups, like the South African Coloured (SAC) population and southern Bantu-speaking populations [24, 25], and a recombination map for diverse Khoe-San populations could benefit studies involving these groups. The demographic history of the Nama is multi-layered, with 5–25% gene flow from Eastern African caprid and cattle pastoralists ~2000 years ago [26] and genetic exchange with the Damara—a hunter-gatherer population of West-Central African ancestry who became economic clients of the Nama. These events were finally followed by recent admixture with European colonists and to a lesser degree ~250 years ago.

We used whole genome sequencing (WGS) data of 54 unrelated Nama individuals [27] to infer a LD-based recombination map that is adjusted according to past changes in N_e. Demographic history was inferred using SMC++ [28] for distant changes in N_e and AS-IBDNe [29] for recent N_e changes. The N_e size changes were then combined, and the demography-aware LD-based method pyrho [9] was used for recombination rate inference. The resultant population-specific recombination map was then compared to other

publicly available recombination maps using the Spearman rank correlation coefficient. Finally, we assessed the fine scale differences between the inferred Nama recombination map and the combined Phase II HapMap recombination map in a region of chromosome 1 and demonstrated how the use of different recombination maps can affect the results from a selection scan.

Results

Briefly, 84 Nama individuals were sequenced to 4x-8x depth via Illumina short read sequencing, variant-called and phased in combination with additional African low coverage genomes as well as 1000 Genomes Phase 3 as part of the African Genome Resource [30]. Genomes were variant-called with GATK3.4 following best practices and phased with SHAPEIT2. Further details regarding the production of this dataset are described in Ragsdale et al. [27]. Global ancestry estimates for the Nama, as compared to other Africans from the African Genome Resource along with representative Europeans (CEU), were inferred using ADMIXTURE. Ancestry estimates indicate that the bulk of the Nama's ancestry is Khoe-San, which is rare elsewhere in the African continent with the exception of the southern Bantu-speaking Sotho and Zulu (Fig. S1). There is a sharp cline in European ancestry across individuals, ranging from ~0 to 50% as may occur with a recent pulse of admixture which has not yet reached equilibrium in a few generations. A subset of individuals carry ancestry frequent in Bantu-speaking and eastern African populations, likely reflecting recent Damara or Herero marriage as indicated in demographic interviews with participants. Ancestry proportion among the full set of related individuals was similar to the subset of unrelated individuals (Fig. S2).

The inferred demographic history of the Nama

We inferred the N_e for the Nama using SMC++ (Fig. 1 A right) and AS-IBDNe (Fig. 1 A left). The results from SMC++ represent the N_e change from 50,000 to 260 generations into the past. The AS-IBDNe results represent the N_e change from 50 to 4 generations into the past and the N_e was inferred using IBD segments of Khoe-San ancestry exclusively. An N_e of ~30,000 approximately 10,000 generations ago with a reduction in N_e to ~21,000 approximately 5000 generations ago is consistent with previously published inferred N_e for the Nama [31]. Inconsistent with previous results, there is a further reduction in N_e to ~10,000 approximately 1000 generations ago. The inferred N_e by SMC++ then stops at 260 generations, because SMC++ can infer N_e approximately 6—120 thousand years ago (kya) with low error [28] and by default SMC++ uses an heuristic to calculate these timepoints automatically given the data.

We therefore used AS-IBDNe to estimate population fluctuations over the past thousand years [32]. We deconvoluted 84 Nama genomes (SNP array) into local ancestry tracts with three possible ancestry states: Khoe-San ancestry, European ancestry and Western-Central African ancestry as represented by Nama, GBR (British in England and Scotland) and LWK (Luhya in Webuye, Kenya) population samples. We tested the accuracy of RFMix via simulation in the Nama as well as testing both SNP array and low coverage genome data in order to determine the best dataset for local ancestry inference. We simulated continuous gene flow from 3 ancestral groups: European admixture starting 8 generations ago with 1% contribution per generation, Bantu admixture starting 14

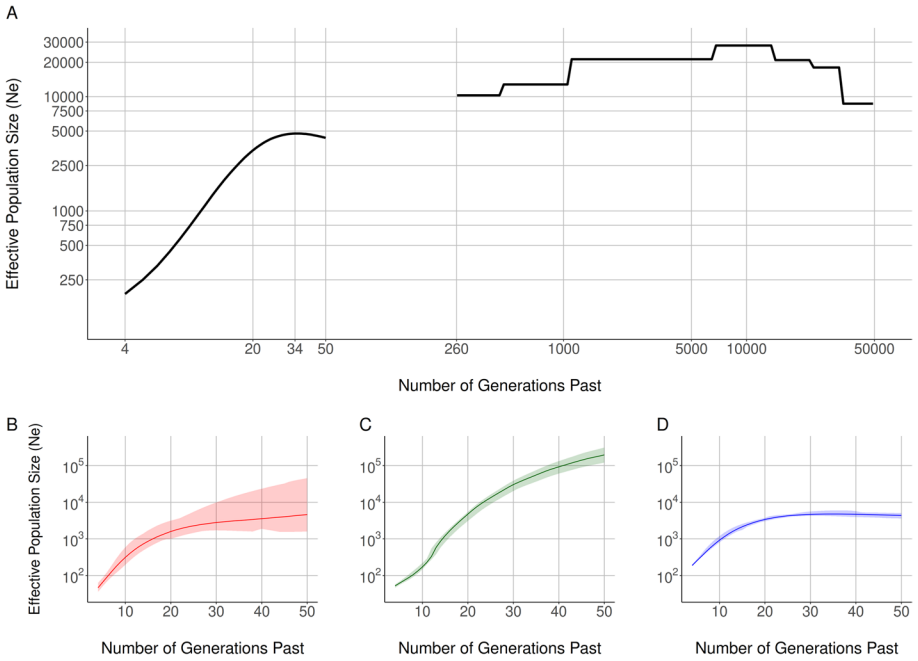

Fig. 1 **A** The inferred effective population size history for the Nama plotted on a log10 scale with SMC++ results on the right and AS-IBDNe results for the Nama on the left. **B**–**D** The AS-IBDNe results for the LWK (**B**), GBR (**C**) and Nama (**D**) ancestral components in the Nama

generations ago with 2% contribution per generation, and the remaining contribution for each generation coming from the Khoe-San. The population randomly mates to create each subsequent generation, also taking into account the recombination landscape to accurately copy haplotype blocks. Eleven individuals were used for each ancestral population: French individuals as the European reference, Bantu-speaking as the West African reference and Nama individuals with >90% Khoe-San ancestry (that were not later used in RFmix runs) for the Khoe-San ancestry component. The average global LAI accuracy, allowing the reference individuals to themselves be admixed, was ~92% on average for the simulated individuals. European ancestry-specific accuracy was 97.6% with individuals being 12.8% European on average, Khoe-San ancestry-specific accuracy was 92.4% with individuals being 76.8% Khoe-San, and Bantu accuracy was 81.5%, with individuals having 10.4% Bantu ancestry overall.

Comparing RFmix runs for the SNP array and genome data to previously obtained ADMIXTURE ancestry percentage estimates, we found that the Khoe-San ancestry was systematically under-called in the genomes compared to the global ancestry estimates from ADMIXTURE, with European and Bantu ancestry consistently higher. This trend is however improved when using the SNP-array data for LAI. Therefore in the AS-IBDNe analysis, admixture deconvolution was performed on MEGA SNP array [33] data in order to facilitate larger numbers of haplotypes in the reference populations.

Beginning 50 generations ago, we infer an N_e of 4360 for the Khoe-San component (Fig. 1D). The N_e starts to decline 34 generations ago and continues to decline until an N_e of 190 inferred 4 generations ago; estimation stops 4 generation ago to avoid coalescent events based on genealogical relationships. The rapid population decline substantially

predates the arrival of European settlers in the Richtersveld in 1760 (or ~7 generations ago), an arid region just south of the Orange River [34]. The N_e results inferred for each set of ancestry specific IBD segments (Fig. 1B–D) have very narrow 95% confidence intervals.

The correlation between the inferred Nama recombination map and other publicly available maps

Fine-scale recombination rate differences between pairs of populations are correlated according to continental levels of population differentiation [9, 11]. Considering the long period that the Nama were isolated and their complex demographic history, we hypothesise that there is no available recombination map that is representative of the Nama. Therefore, we compared the inferred recombination map for the Nama with 26 other publicly available recombination maps derived from [9] using the Spearman rank correlation at a 2-kilobase resolution (Fig. 2). These maps were inferred for populations from the 1000 Genomes [35] dataset, and the populations are classified into various super-populations representing major ancestry differences. We find that pairwise correlations between all 27 maps cluster according to continental levels of population differentiation

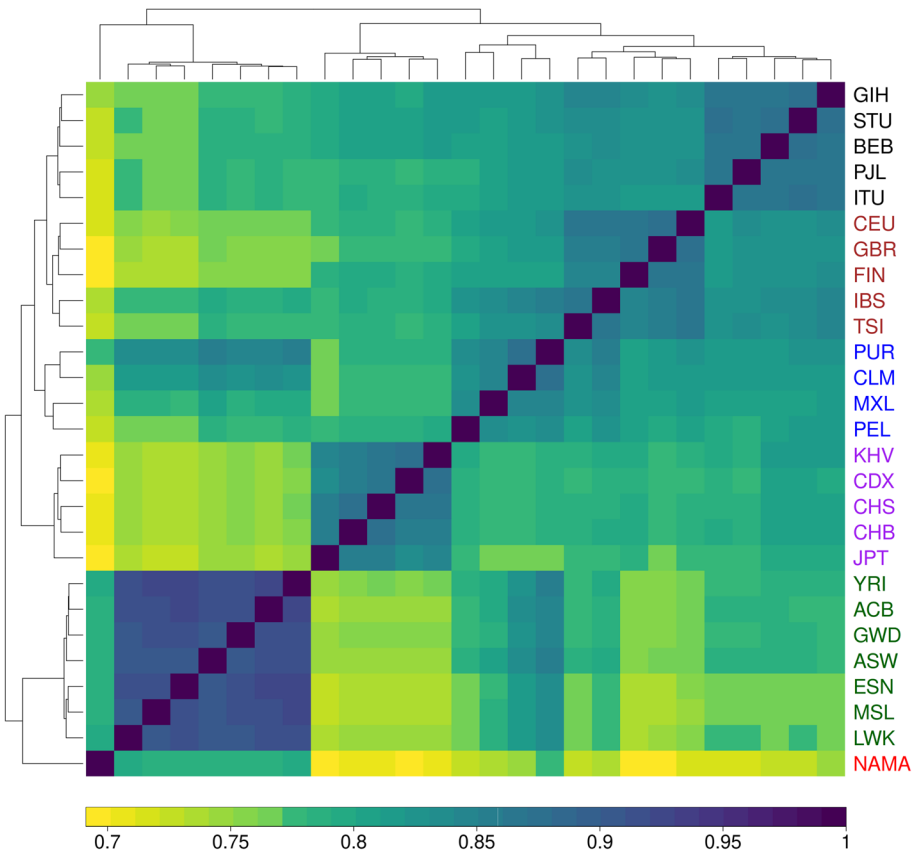

Fig. 2 Heatmap indicating the Spearman rank correlation between the genetic maps of 27 populations, including the Nama, at a 2-kilobase resolution. The colour of the population labels represent distinct super-population groups, with the Nama highlighted in red. There is clear clustering according to super-population groups and the Nama recombination map correlates the best with other African populations

(super-populations). Furthermore, we find that the Nama are more closely related to other African populations than to other continental groups (< 0.75); however, the pairwise correlations between the Nama and the other African populations are much weaker (~0.79) than the pairwise correlations between the African populations (> 0.90). These values represent correlations between inferred maps and, therefore, include any noise potentially introduced during inference. The true maps are likely to be more similar than these values suggest.

The Spearman rank correlation mitigates potential differences in map length that would influence the Pearson correlation coefficient. Therefore, we neglect the magnitude of the recombination rate in favour of qualitative aspects of the maps. Inspecting the qualitative aspects of recombination maps is especially relevant when LD-based recombination maps are compared, since LD-based methods produce population recombination rates that need to be scaled using N_e and therefore assume an accurate estimate for N_e.

Fine-scale recombination as applied in selection scans

The combined Phase II HapMap recombination map is derived from 270 individuals who represent four geographically diverse populations, including the Yoruba from Western Africa. It is sometimes used as a proxy [36] for southern African populations, since all other available recombination maps derived from African populations are of western African ancestry, a globally diverse map is thought to be the best substitute. Even though population-specific recombination maps are similar at low resolutions, certain analyses, such as selection scans, might benefit from a high-resolution population-specific recombination map that accurately captures fine-scale differences. Figure 3 illustrates

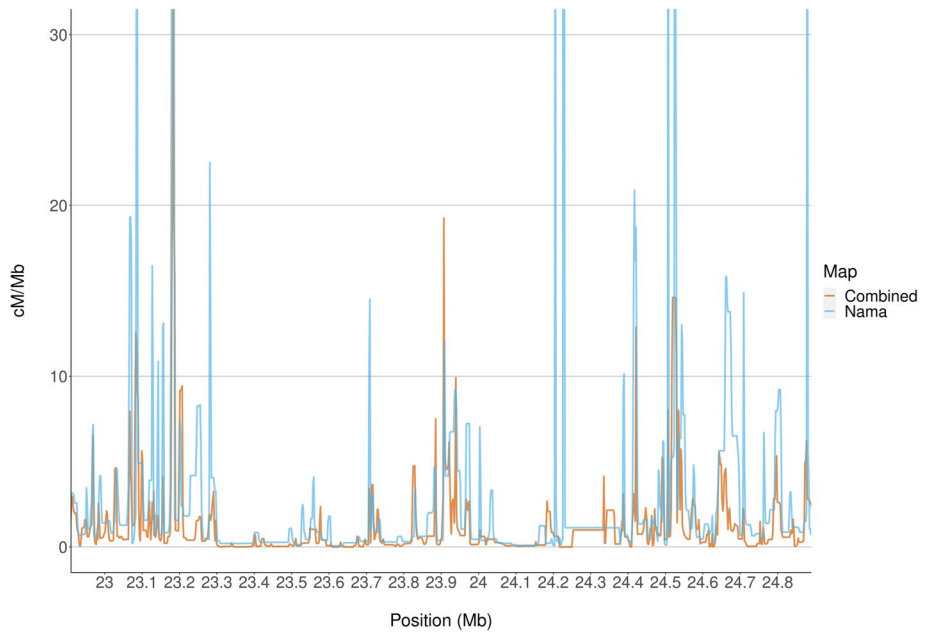

Fig. 3 The recombination rate of the combined Phase II HapMap recombination map and the inferred recombination map for the Nama plotted over a segment of chromosome 1. There is a high degree of overlap between the maps across this region, but there are positions with recombination hotspots indicated by the Nama map that are not indicated by the combined Phase II HapMap map, e.g. at 24.2 Mb

the recombination rate (cM/Mb) plotted over part of chromosome 1 for the combined Phase II HapMap recombination map (orange) and the inferred recombination map for the Nama (blue). The positions of regions of high recombination (hotspots) are largely concordant between the two maps and mainly differ in magnitude. However, in the region at 24.2 Mb, there are hotspots present in the Nama recombination map that are absent from the combined Phase II HapMap recombination map. To further investigate the effects that these differences could have, we performed genome-wide selection scans on Nama SNP array data using the combined Phase II HapMap map and the Nama map. We focused on the integrated haplotype scores (iHS), a selection statistic which detects recent positive selection, by evaluating haplotype homozygosity for the ancestral and derived haplotypes extending from a locus of interest [37]. iHS is most effective at detecting alleles that have been swept to intermediate frequencies, and it is among the most common statistics cited in other comparable selection scans in the Khoe-San.

After taking the absolute value of the integrated Haplotype Scores (iHS) and filtering for the highest 1.0% of the scores, we found an overlap of 1504 candidate genes (50%) between the two maps. However, the run using the combined Phase II HapMap map and the run using the Nama map identified 808 and 713 unique candidate genes respectively (Fig. 4). The difference in the number of top 1.0% of hits is due to the change in the relative length of the maps. The Pearson correlation (r) between the iHS scores found using the combined Phase II HapMap and Nama maps is 0.93. We compiled a list of 131 candidate genes [18, 20, 36], previously identified using iHS, that are under selection in the Khoe-San and compared this list to our results. We found an overlap of three genes (*CTNNAL1*, *ALDH1A2* and *SYT14*) between the previously identified genes and the run using the combined Phase II HapMap map but only an overlap of one gene (*TRIM39*)

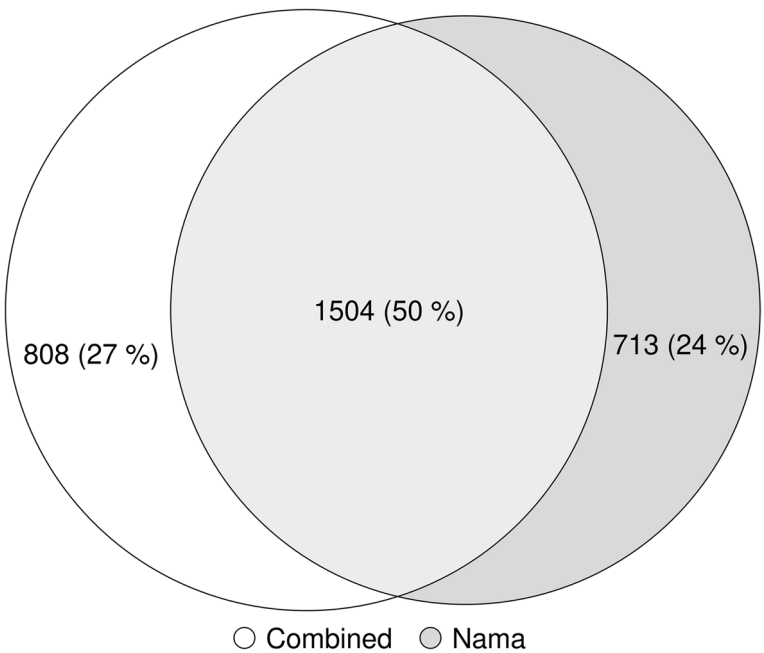

Fig. 4 Venn diagram of the candidate genes found using the 1.0% highest selection scan results (absolute value iHS) for the selection scan using the combined Phase II HapMap map (white) and the selection scan using the Nama map (grey)

between the previously identified genes and the run using the Nama map. *TRIM39* encodes for a ring finger protein associated with diseases including Behcet's syndrome; it regulates p21 and plays an important role in determining cell fate [38]. Previous research has demonstrated that selection statistics such as iHS are sensitive to phasing, sample size and ascertainment bias [39]. Our results indicate that a population-specific recombination map should also be considered in attempts to fine-map adaptive haplotypes.

Discussion and conclusions

Recombination maps are important resources for epidemiological and evolutionary analyses; however, there are currently no recombination maps that represent southern African populations [16]. The Nama, a southern African indigenous population, would likely produce a distinct recombination landscape from publicly available recombination maps, because of their complex demographic history. The recent rapid population decline (shown in Fig. 1) partially illustrates this complex history. Despite gene flow from Eastern African pastoralists ~2000 years ago and recent admixture with Europeans, the Nama do not cluster with any of the continental groups that we have representative recombination maps for (Fig. 2). Therefore, their recombination landscape is indeed unique and epidemiological studies that involve the Nama or any other related populations, like other Khoe-San populations or southern African Bantu-speaking groups, would benefit from our inferred map. This recombination map also represents a likely upper bound on the extent of divergence we expect to see for a recombination map in humans and would be of interest to any researcher that wants to test the sensitivity of population genetic or GWAS analysis to recombination map input. Fine-scale differences in recombination can meaningfully alter the results of a selection scan (demonstrated in Figs. 3 and 4). However, it should be noted that recent studies found that population-specific recombination maps have little effect on phasing [40], imputation [40] and local ancestry inference [41]. Therefore, the combined Phase II HapMap recombination map's proxy status with regards to the Nama is dependent on the analysis that the map is used for.

There are many available techniques [42] to infer the recombination rate and some have contrasting limitations which means that not all techniques would allow accurate, fine-scale estimates for a given dataset. Assuming limitless resources, we would have preferred pedigree-based methods, because these allow sex-specific recombination rate inference and rely on inferring individual recombination events between successive generations based largely on observed meioses. However, pedigree-based methods require many thousands of individuals to produce fine-scale maps [2]. Other options are IBD-based and LAI-based methods, but they too require in the order of a couple thousand individuals for fine-scale estimates [5]. Our small sample size (54 unrelated individuals) made LD-based methods the obvious choice for fine-scale estimates. However, there are many assumptions that accompany LD-based methods that make them less than ideal, for instance the assumption of a constant N_e and the potential bias from gene flow when inferring recombination in admixed populations [43]. Therefore, the complex demographic history of the Nama made demography-aware methods, like pyrho, the ideal compromise between data availability and accuracy. Even so, the population-specific

recombination map presented here is likely an accurate representation of the recombination landscape of the Nama and future epidemiological and evolutionary research will benefit from this resource.

Methods

Inferring demographic history

It has been shown that demographic history, especially recent bottlenecks, can greatly impact LD-based recombination inference. We, therefore, inferred the demographic history of the Nama to improve our recombination rate estimates. Two methods, SMC++ (v1.15.2) [28] and IBDNe (v23Apr20) [32], were used and the results combined. See Fig. 5 for an overview of the methods.

SMC++ uses LD information to infer demographic histories and can infer divergence times between 6 and 120 kya with low error [28]. A whole genome sequencing (WGS) dataset (EGAD00001006198) of 84 Nama individuals (54 unrelated) was used. The input for SMC++ was created separately for each chromosome, from the unrelated individuals in the WGS dataset, by using the *vcf2smc* program with 10 randomly selected "distinguished" (see Terhorst et al. [28] for more information on this) individuals. The result is 10 separate datasets for each chromosome. This creates a composite likelihood which, according to the authors, may lead to improved estimates. A per-generation mutation rate of 1.25e−8 was assumed and all of the input files were then included in an estimate of the N_e through time using the *estimate* program. Since SMC++ regards uncalled

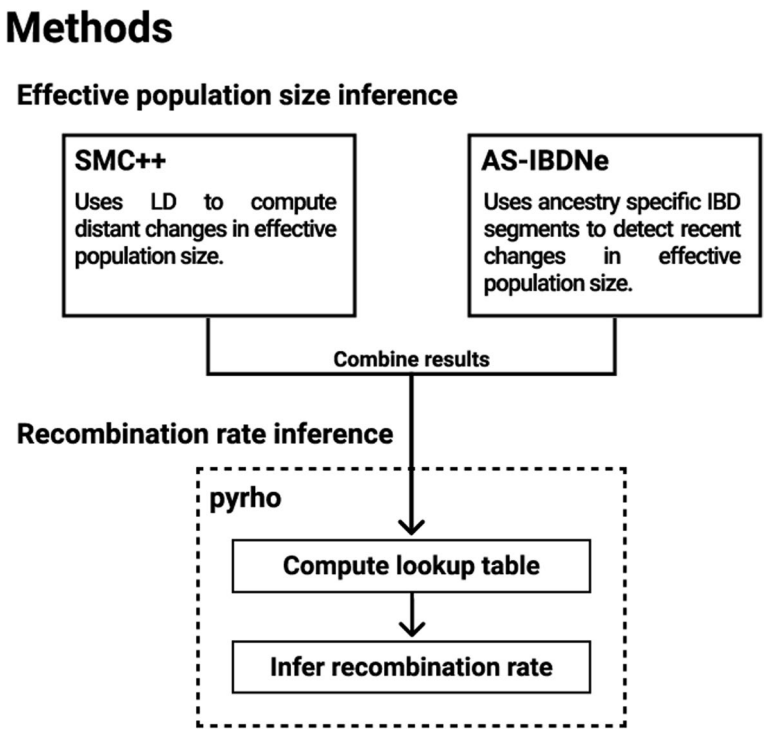

Fig. 5 A brief overview of the methods used in effective population size inference and the subsequent recombination rate inference

regions as long runs of homozygosity, Stephen Schiffels' mappability mask (created for human genome build GRCh37 using SNPable [44]) was used to mask regions of low mappability. All other default parameters were used.

IBDNe can infer the N_e size 4-50 generations into the past by using identity by descent (IBD) information. By separating IBD segments by ancestry before inferring the N_e, one can obtain an estimate of N_e localised to each population ancestry. We developed a Snakemake pipeline (Fig. 6), called AS-IBDNe (https://github.com/hennlab/AS-IBDNe), to estimate ancestry specific N_e from a given SNP array dataset. The pipeline was adapted from the procedure used in Browning et al. [29]. We ran it on 84 Nama individuals genotyped on the Multi-Ethnic Global Array (MEGA) [33]. The pipeline takes in SNP-array data in plink binary file format, uses plink v1.9 [45] to break the data by chromosome, and shapeIT v2 [46] to phase the chromosomes. The dataset is then converted to VCF format using SHAPEIT2 and split into one file containing the reference individuals and one file containing the admixed individuals using BCFtools [47]. Next, RFMix v2.0 [48] is run on these two vcf files to estimate the ancestry of arbitrarily sized segments across the genome. Simultaneously, RefinedIBD and merge-ibd-segments.17Jan20.102.jar [49]

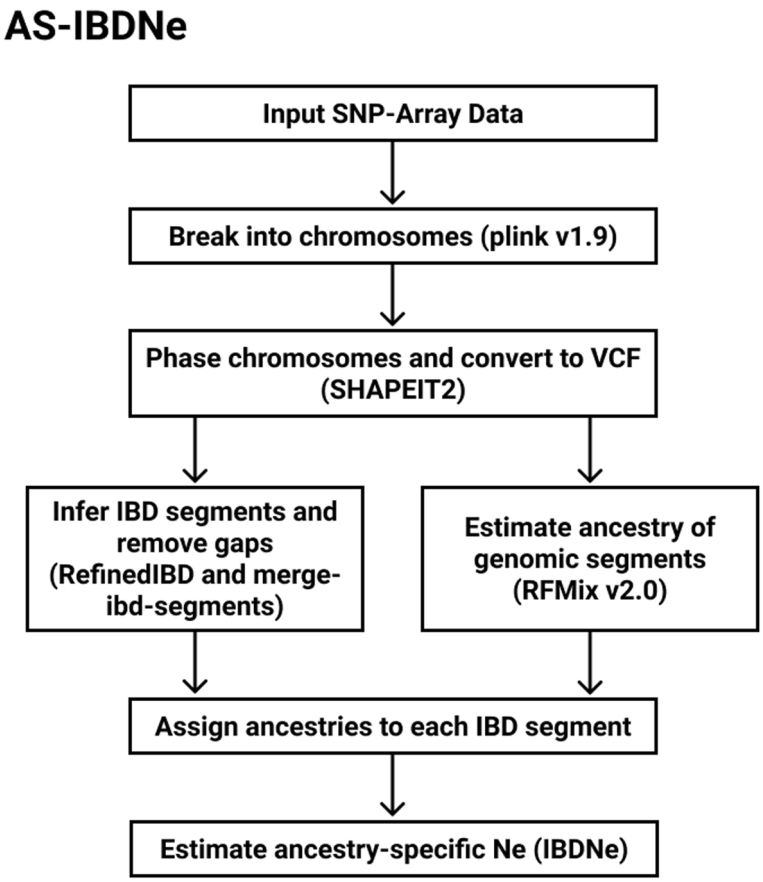

Fig. 6 An overview of the AS-IBDNe pipeline. Input SNP array data in plink binary format is split by chromosomes using plink v1.9. Each chromosome is then phased and converted to vcf format by SHAPEIT2. IBD segments are next inferred using RefinedIBD, and merge-ibd-segments is used to remove gaps between them. Meanwhile, RFMix2.0 is run to estimate the ancestry of differently sized genomic segments. Finally, RFMix-produced ancestries are assigned to each IBD segment, and IBDNe is run to produce ancestry-specific effective population size estimates

is run on the phased data to infer ibd segments and remove any gaps between them. The ancestries produced from RFMix v2.0 are then assigned to each IBD segment using a custom python script. This information is provided to the program IBDNe [32], which produces estimates of historical population size for each ancestry. The RFMix results were also used to create the ternary diagrams in Fig. S2 using the ggtern package in R. All the default parameters of RFMix, RefinedIBD and IBDNe were used except RFMix, which was run with 3 expectation maximisation iterations and the *reanalyse-reference* flag, and IBDNe, which was run with the *mincM* flag set to 3. The combined Phase II HapMap recombination map was used whenever a recombination map was required during the inference. The output of SMC++ can be converted to a csv where the time-scale and the N_e estimates are linear. The output from AS-IBDNe can then be added to the linear output from SMC++, and this file can then be used during recombination rate inference in pyrho [9].

Recombination rate inference

Previous published guidelines [42] aided the choice of recombination rate inference method and pyrho, a demography-aware LD-based method, was selected. We assumed the same per-generation mutation rate of 1.25e−8 for all the inference steps. The most computationally laborious task when using pyrho is the generation of a lookup table which enables subsequent processes to be computationally faster. The combined SMC++/AS-IBDNe demographic history was used to generate a lookup table for the unrelated subset of 54 Nama WGS individuals using pyrho *make_table*. A convenient feature of this lookup table is that it is compatible with other recombination rate inference software, e.g LDhat [6], which make use of exact two-locus sampling probabilities with the added benefit of already taking the specified demographic history into account. This lookup table and the combined demographic history were employed to find optimal hyperparameters to be used for recombination rate inference with pyrho *hyperparam*. The parameters that yielded the highest overall accuracy were a smoothness penalty of 15 and a window size of 30. These parameters and the lookup table were then used to infer the recombination rate with pyrho *optimise*. The output provides the per base pair per generation recombination rate for a given interval.

Selection scans

For the selection scans, we used data from 104 Nama individuals who were genotyped on the Illumina Omni2.5 array as part of the African Genome Diversity Project. Close relatives were identified from demographic interviews and verified via allele-based kinship coefficients in *plink*. Individuals with more than 50% European, and Damara or Herero admixture were excluded. Ancestry estimates were obtained using ADMIXTURE with $k=6$ possible ancestral clusters: Nama, Northern San, Near Eastern, East African Nilotic, West African, and European (see also Fig. S1) [50]. After QC, kinship and ancestry exclusions, we analysed $n=55$ individuals. We calculated iHS using selscan 1.3.0 [51] and default parameters. For the --map flag, we used recombination rates from the custom Nama map in one run and from the combined Phase II HapMap in a second

run. We filtered for the most extreme iHS scores (absolute value) by taking the highest 1.0% of the scores.

We annotated these positions using the gene range list provided by Plink (https://www.cog-genomics.org/plink/1.9/resources). We compared the candidate genes found in each run of selscan to create a Venn Diagram. We also calculated the Pearson correlation between iHS scores for each SNP as calculated by each run of selscan.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02744-5.

Additional file 1: Figure S1. Ancestry estimates for the Nama obtained using ADMIXTURE with k=4 to k=7 possible ancestral clusters. Figure S2. (A) Ternary diagram showing the African, European and Khoe-San ancestry contributions, as represented by the LWK, GBR and Nama respectively, for the whole WGS dataset of 84 individuals. (B) Ternary diagram showing the African, European and Khoe-San ancestry contributions, as represented by the LWK, GBR and Nama respectively, for the unrelated subset of the WGS dataset.

Additional file 2.

Acknowledgements
The authors would like to thank Prof. Carina Schlebusch and Dr. Torsten Günther for providing previously published data on the demographic history of the Nama. We thank Aaron Ragsdale, Jeffrey Spence and one anonymous reviewer for their thoughtful comments and careful examination of the manuscript. We thank Elizabeth Atkinson for comparative ancestry analysis. We express our gratitude to the Nama community for their generous contribution of DNA, family interviews and ethics consultation without which this research would not be possible.

Review history
The review history is available as Additional file 2.

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
GvE performed effective population size inference, created recombination maps, compared recombination maps and wrote the main body of the article. EP calculated and filtered integrated haplotype scores and wrote the sections relating to selection scans. M Mastoras formalised the AS-IBDNe pipeline and provided a figure and description for the pipeline. CU, GvdS, GCT, BMH and M Möller conceptualised and reviewed. All authors read and approved the final manuscript.

Funding
This research was funded (partially or fully) by the South African government through the South African Medical Research Council and the National Research Foundation. GvE was supported by the DSI-NRF Innovation Doctoral Scholarship. This research was supported by NIH grant R35GM133531 (to BMH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials
Sequence data has been deposited at the ega (EGA), which is hosted by the EBI and the CRG, under accession number EGAD00001006198 [52]. These data are described more fully in Ragsdale et al. [27]. The recombination map inferred for the Nama can be found at https://github.com/TBHostGen/nama-recombination-map [53]. Our pipeline for inferring AS-IBDNe is available at https://github.com/hennlab/AS-IBDNe.

Declarations

Ethics approval and consent to participate
Approved by the Health Research Ethics Committee 2 of Stellenbosch University under ethics reference number S20/02/034. The Health Research Ethics Committee (HREC) complies with the SA National Health Act No. 61 of 2003 as it pertains to health research. The HREC abides by the ethical norms and principles for research, established by the World Medical Association (2013). Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects; the South African Department of Health (2006). Guidelines for Good Practice in the Conduct of Clinical Trials with Human Participants in South Africa (2nd edition), as well as the Department of Health (2015). Ethics in Health Research: Principles, Processes and Structures (2nd edition). The Health Research Ethics Committee reviews research involving human subjects conducted or supported by the Department of Health and Human Services, or other federal departments or agencies that apply the Federal Policy for the Protection of Human Subjects to such research (United States Code of Federal Regulations Title 45 Part 46), and/or clinical investigations regulated by the Food and Drug Administration (FDA) of the Department of Health and Human Services.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 19 December 2021 Accepted: 1 August 2022
Published online: 09 August 2022

References

1. Peñalba JV, Wolf JBW. From molecules to populations: appreciating and estimating recombination rate variation. Nat Rev Genet. 2020;21:476–92.
2. 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.
3. 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;43:847–53.
4. Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, et al. The landscape of recombination in African Americans. Nature. 2011;476:170–5.
5. Zhou Y, Browning BL, Browning SR. Population-specific recombination maps from segments of identity by descent. Am J Hum Genet. 2020;107:137–48.
6. Auton A, McVean G. Recombination rate estimation in the presence of hotspots. Genome Res. 2007;17:1219–27.
7. Gao F, Ming C, Hu W, Li H. New software for the fast estimation of population recombination rates (fastepr) in the genomic era. G3 (Bethesda). 2016;6:1563–71.
8. Dapper AL, Payseur BA. Effects of demographic history on the detection of recombination hotspots from linkage disequilibrium. Mol Biol Evol. 2018;35:335–53.
9. Spence JP, Song YS. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. Sci Adv. 2019;5:eaaw9206.
10. Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T, et al. A fine-scale chimpanzee genetic map from population sequencing. Science. 2012;336:193–8.
11. Graffelman J, Balding DJ, Gonzalez-Neira A, Bertranpetit J. Variation in estimated recombination rates across human populations. Hum Genet. 2007;122:301–10.
12. Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. DNA recombination. Recombination initiation maps of individual human genomes. Science. 2014;346:1256442.
13. Serre D, Nadon R, Hudson TJ. Large-scale recombination rate patterns are conserved among human populations. Genome Res. 2005;15:1547–52.
14. 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;467:1099–103.
15. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449:851–61.
16. Swart Y, van Eeden G, Sparks A, Uren C, Möller M. Prospective avenues for human population genomics and disease mapping in southern Africa. Mol Genet Genom. 2020;295:1079–89.
17. Uren C, Kim M, Martin AR, Bobo D, Gignoux CR, van Helden PD, et al. Fine-Scale human population structure in southern Africa reflects ecogeographic boundaries. Genetics. 2016;204:303–14.
18. Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, et al. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci USA. 2011;108:5154–62.
19. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet. 2011;43:1031–4.
20. Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 2012;338:374–9.
21. Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, et al. The genetic prehistory of southern Africa. Nat Commun. 2012;3:1143.
22. Barbieri C, Hübner A, Macholdt E, Ni S, Lippold S, Schröder R, et al. Refining the Y chromosome phylogeny with southern African sequences. Hum Genet. 2016;135:541–53.
23. Barbieri C, Vicente M, Rocha J, Mpoloka SW, Stoneking M, Pakendorf B. Ancient substructure in early mtDNA lineages of southern Africa. Am J Hum Genet. 2013;92:285–92.
24. Uren C, Möller M, van Helden PD, Henn BM, Hoal EG. Population structure and infectious disease risk in southern Africa. Mol Genet Genomics. 2017;292:499–509.
25. Sengupta D, Choudhury A, Fortes-Lima C, Aron S, Whitelaw G, Bostoen K, et al. Genetic substructure and complex demographic history of South African Bantu speakers. Nat Commun. 2021;12:2080.
26. Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, et al. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. Proc Natl Acad Sci USA. 2008;105:10693–8.
27. Ragsdale AP, Weaver TD, Atkinson EG, Hoal E, Möller M, Henn BM, et al. A weakly structured stem for human origins in Africa. BioRxiv. 2022.
28. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 2017;49:303–9.
29. Browning SR, Browning BL, Daviglus ML, Durazo-Arvizu RA, Schneiderman N, Kaplan RC, et al. Ancestry-specific recent effective population size in the Americas. PLoS Genet. 2018;14:e1007385.
30. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015;517:327–32.
31. Schlebusch CM, Sjödin P, Breton G, Günther T, Naidoo T, Hollfelder N, et al. Khoe-San genomes reveal unique variation and confirm the deepest population divergence in Homo sapiens. Mol Biol Evol. 2020;37:2944–54.

32. Browning SR, Browning BL. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. Am J Hum Genet. 2015;97:404–18.
33. Martin AR, Lin M, Granka JM, Myrick JW, Liu X, Sockell A, et al. An unexpectedly complex architecture for skin pigmentation in Africans. Cell. 2017;171:1340–1353.e14.
34. Smith AB. Einiqualand: studies of the Orange river Frontier. Rondebosch: Uct Press; 1995.
35. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
36. Vicente M, Jakobsson M, Ebbesen P, Schlebusch CM. Genetic affinities among southern Africa hunter-gatherers and the impact of admixing farmer and herder populations. Mol Biol Evol. 2019;36:1849–61.
37. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4:e72.
38. Zhang L, Mei Y, Fu N, Guan L, Xie W, Liu H, et al. TRIM39 regulates cell cycle progression and DNA damage responses via stabilizing p21. Proc Natl Acad Sci USA. 2012;109:20937–42.
39. Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW. Limited evidence for classic selective sweeps in African populations. Genetics. 2012;192:1049–64.
40. Hassan S, Surakka I, Taskinen M-R, Salomaa V, Palotie A, Wessman M, et al. High-resolution population-specific recombination rates and their effect on phasing and genotype imputation. Eur J Hum Genet. 2020.
41. van Eeden G, Uren C, van der Spuy G, Tromp G, Möller M. Local ancestry inference in heterogeneous populations-are recent recombination events more relevant? Brief. Bioinformatics. 2021.
42. van Eeden G, Uren C, Möller M, Henn BM. Inferring recombination patterns in African populations. Hum Mol Genet. 2021;30:R11–6.
43. Samuk K, Noor MAF. Gene flow biases population genetic inference of recombination rate. BioRxiv. 2021.
44. Li H. SNPable. 2009. http://lh3lh3.users.sourceforge.net/snpable.shtml. Accessed 1 May 2021.
45. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
46. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 2014;10:e1004234.
47. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10.
48. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013;93:278–88.
49. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 2013;194:459–71.
50. Lin M, Siford RL, Martin AR, Nakagome S, Möller M, Hoal EG, et al. Rapid evolution of a skin-lightening allele in southern African Khoe-San. Proc Natl Acad Sci USA. 2018;115:13324–9.
51. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 2014;31:2824–7.
52. Collection of Genotypic and Ethnographic Information from Individuals of South African Ethnic Groups. European Genome-Phenome Archive. EGAD00001006198. https://ega-archive.org/datasets/EGAD00001006198
53. van Eeden G, Uren C, Pless E, Mastoras M, van der Spuy G, Tromp G, et al. Nama recombination map. 2021. https://github.com/TBHostGen/nama-recombination-map. Accessed 26 Jun 2022.

Publisher's Note