René L Warren, Lauren Coombe, Johnathan Wong, Parham Kazemi, Inanc Birol

Canada's Michael Smith Genome Sciences Centre, Vancouver, Canada, V5Z 4S6

Human ancestry inference at scale, from genomic data

Using an alignment-free single nucleotide variant prediction framework that leverages integrated variant call sets from the 1000 Genomes Project, we demonstrate accurate ancestry inference predictions on over 600 human genome sequencing datasets, including complete genomes, draft assemblies, and >280 independently-generated datasets. The method presented, ntRoot, infers super-population ancestry along an input human genome in 1h15m or less on 30X sequencing data, and will be an enabling technology for cohort studies.

Our ancestry information, encoded within our DNA, is crucial for genomic studies. Individuals from different ancestries may have a different genetic makeup and associated disease risk factors, yet many association studies do not take ethnicity into consideration or simply leave it as a confounding variable, either because that information is costly to compute, is based on incomplete genotype data, or is not reliably measured. On a population scale, the HapMap Project[1] was the first to profile common patterns of variation between individuals. Subsequently, the 1000 Genomes Project (1kGP)[2] and the Simons Genome Diversity Project (SGDP)[3] catalogued human genomic variation across the whole genomes of healthy individuals from various populations, to generate a baseline for the study of population genetics.

Ancestry admixture inference can be divided into Local Ancestry Inference (LAI) and Global Ancestry Inference (GAI). LAI is the estimation of ancestry along a DNA sequence. This paradigm is based on the fact that during chromosomal crossover, proximal regions of DNA are carried together. Hence, even if an individual's ancestors were admixed generations ago, DNA fragments may be sourced from single ancestry. While GAI offers a genome-wide summary, LAI provides higher resolution ancestry inferences across genomic loci[4].

Most existing computational approaches developed for ancestry inference are based on single nucleotide variants (SNVs), leveraging the availability of population-scale genotype data from HapMap[1] and 1kGP projects[5]. Tools for GAI analysis include the widely-applied model-based STRUCTURE program[6], principal component analysis-based utilities Rye[7] and EIGENSTRAT[8], and ADMIXTURE for likelihood model-based estimation of ancestry in unrelated individuals[9]. HAPAA[10] and HAPMIX[11], both designed for LAI, adopt Hidden Markov Models (HMMs) and explicitly incorporate linkage disequilibrium (LD), the degree of correlation between allelic SNVs. RFMix[12] is an HMM-based admixture model with conditional random forest that leverages phase and LD information[13] for both GAI and LAI, and SNVstory[14] uses machine learning-based models to compute GAI from sequencing data. While these utilities have their strengths, large scale studies and routine applications may benefit from more flexible genomic input and smoother execution to better align with the demands of the current genomic era.

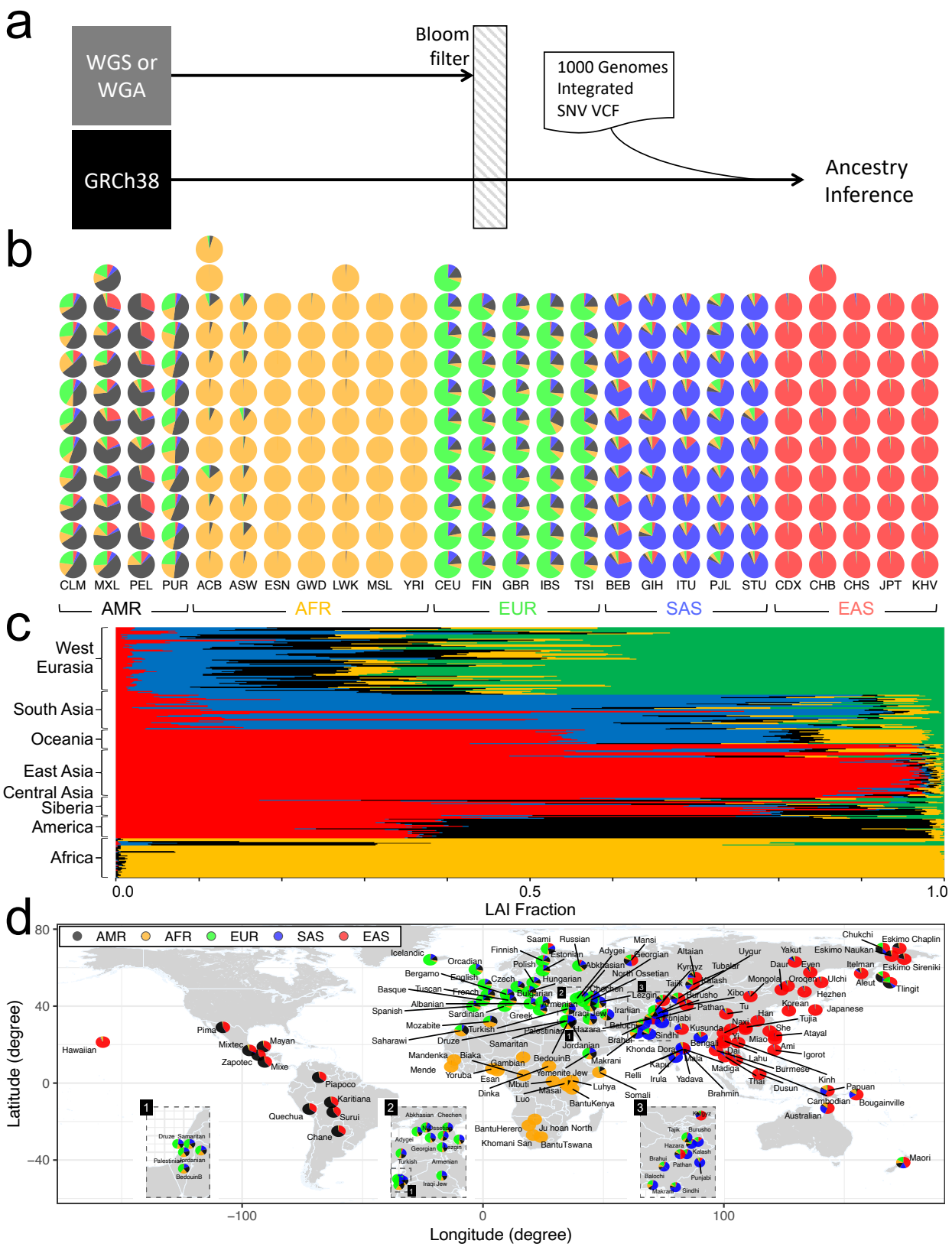

Fig. 1: Ancestry inference from genomic data. **a**, Schematic of the ntRoot ancestry inference workflow on whole genome sequencing (WGS) or assembly (WGA). Each genomic base of the human genome reference GRCh38 queries a Bloom filter built from an individual WGS or WGA data; Identified SNVs are cross-referenced with the 1kGP integrated variant call set[2]. Human

ancestry composition estimates derived from LAI on **b**, 266 WGS 1kGP datasets grouped by 5 super-populations (AMR:American; AFR:African; EUR:European; SAS:South Asian; EAS:East Asian) and 26 populations (CLM:Colombian; MXL:Mexican-American; PEL:Peruvian; PUR:Puerto Rican; ACB:African-Caribbean; ASW:African-American South West; ESN:Esan; GWD:Gambian; LWK:Luhya; MSL:Mende; YRI:Yoruba; CEU:CEPH (Centre d'Étude du Polymorphisme Humain – Utah); FIN:Finnish; GBR:British; IBS:Spanish; TSI:Tuscan; BEB:Bengali; GIH:Gujarati; ITU:Indian; PJL:Punjabi; STU:Sri Lankan; CDX:Dai Chinese; CHB:Han Chinese; CHS:Southern Han Chinese; JPT:Japanese; KHV:Kinh Vietnamese), and **c**, 279 SGDP WGS datasets from 129 distinct populations **d**, shown in the global context (labels), with inset regions zoomed-in

Here, we present ntRoot, a method for super-population level LAI and GAI analysis from whole genome assemblies or raw genome sequencing datasets. It uses a sequence alignment-free genome variant detection framework[15] that leverages the integrated variant call sets from 1kGP. We adapted the base-editing utility ntEdit[15] for SNV prediction and ancestry inference, where each genomic base of the human genome reference (GRCh38) is interrogated in the genomic data from the individual of interest, using a reduced sequence representation (Fig. S1) to detect putative SNVs (Fig. 1a). Cross-referencing the putative SNVs identified with 1kGP integrated variant call sets, we calculate the average allele frequency (AF) within each of the five super-populations. We keep an independent tally of non-zero AF SNVs at both local (per tile) and global (whole genome) levels. We rank the super-populations separately for each tile and the genome using those combined metrics, as an indicator of the most likely ancestry.

To benchmark the performance of ntRoot, we derived super-population-level ancestry for a random subset of the 1kGP whole genome sequencing (WGS) datasets[16] organized into 26 human populations (Fig. S2, Table S1, n=266, 10-12 per population) and observed correct ancestry inferences for all 266 individuals compared to true labels when derived globally (Fig. S3; Tables S1-S3). Ancestry composition was inferred locally using 5Mbp or 2Mbp tiles (Table S4), and the estimated super-population contributions were reported for the entire genome of each 266 and 279 WGS 1kGP and SGDP datasets, respectively (Fig. 1b-d, Fig. S4-S5, Table S5). We note that, based on LAI, the ancestry composition of Americans (AMR), Europeans (EUR) and South Asians (SAS) is more diverse compared with that of Africans (AFR) and East Asians (EAS), consistent with historical accounts of population movement[17]. Inferring ancestries locally is information-rich (Fig. 2) but we caution about deriving a GAI label from LAI (i.e., largest pie chart slice) as, out of the 266 1kGP datasets, one (CLM:Columbian, 0.4%) and three labels (1 CLM, 2 PUR:Puerto Rican, 1.1%) are inconsistent with the highest super-population fraction, when using 5Mbp vs. 2Mbp tiles, respectively (Table S4). Interestingly, all four incongruent samples represent individuals from the Americas, where population admixture has occurred[17] and hence, ancestry is less homogeneous.

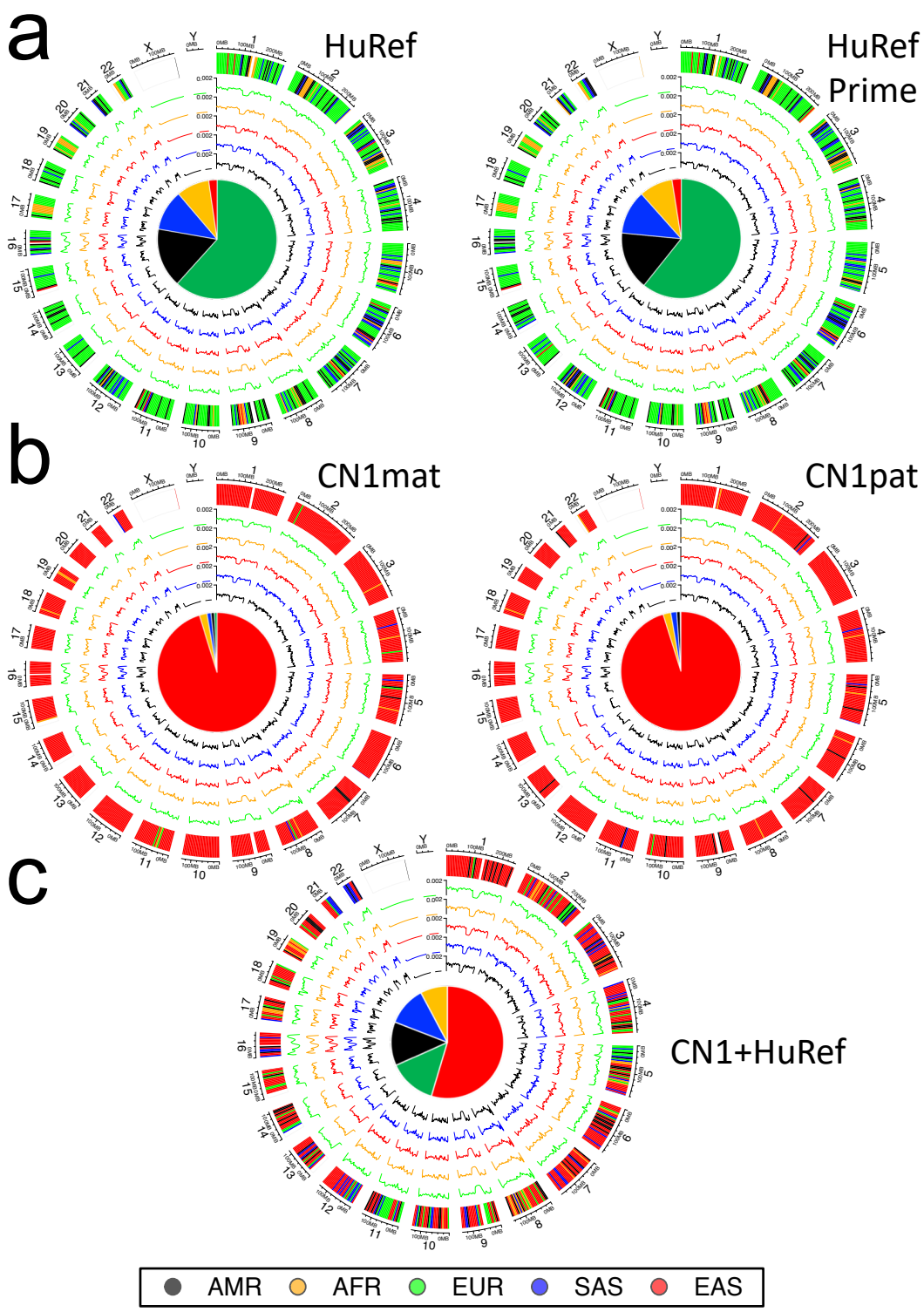

Fig. 2: Circular genome representations of ancestry inference using WGA for **a,** European HuRef[18]and HuRefPrime haplotypes, **b,** Han Chinese CN1[19] (mat:maternal, pat:paternal) haplotypes and **c,** simulated HuRef/CN1 diploid rearrangement. The outer track shows ntRoot local ancestry assignments (LAI) along each chromosome using 5Mbp tiles (AMR:American; AFR:African; EUR:European; SAS:South Asian; EAS:East Asian), with each subsequent inner track showing the corresponding SNV density for each super-population. Inner pie charts report the predicted global ancestry composition (GAI).

ntRoot LAI inferences using the chromosome-scale diploid HuRef[18] (Fig. 2a), CN1[19] (Fig. 2b) and haploid KOREF[20] (Fig. S6) genome assemblies of three individuals places the prevalent ancestry based on highest composition (EUR:54%, EAS:87%, EAS:87%, respectively), consistent with both highest GAI scores, and the published European[18], Chinese[19] and Korean[20] ancestry for each (Tables S6-S7). Further, ntRoot correctly infers EAS and EUR as the two most prevalent super-population ancestries from a synthetic diploid genome admix of HuRef and CN1 (Fig. 2c, Table S7).

Using orthogonal datasets distinct from the 1kGP data, WGS-based SGDP results (Table S5) and WGA-based GAI/LAI results (Tables S6 and S7, respectively) independently validate our approach. Unlike existing tools, ntRoot can infer ancestries directly from a genome assembly, even when it is in a draft stage (Fig. S7). Processing haploid WGA with ntRoot averaged 29m and consumed at most 13 GB RAM, while it ran in under 1h36m requiring at most 68 GB RAM on average, on 279 SGDP WGS datasets (22—87-fold genome coverage, average 44-fold). Despite varying WGS sequencing coverage, ntRoot produced robust predictions, using as low as 22X coverage (Tables S5 and S8). Compared to SNVstory[14], a recently developed GAI-based utility, ntRoot demonstrates the same accuracy while running ~4X faster (1h09m vs. 4h25m) and with less memory when inferring the ancestry of 100 1kGP individuals from WGS data (Tables S8 and S9).

With its streamlined workflow, ntRoot generates GAI and LAI-based admixture profiles that offer detailed insights into population genetics. We anticipate ntRoot will broadly facilitate ancestry inference and association studies providing reliable objective demographics information.

Methods

Ancestry inference workflow

On whole genome sequencing (WGS) read datasets, the ntRoot workflow (driven by the python script, ntroot --reads) launches ntEdit, followed by the ancestry inference step. The ntEdit pipeline first executes ntHits (v1.0.2, user-defined parameter k=55, https://github.com/bcgsc/ntHits), which outputs a Bloom filter[21] (BF) of robust (non-error) read k-mers. ntEdit (v2.0.0, parameters -s 1 -Y 0.55 -j 3 -l 1000GP_integrated_snv_v2a_27022019.GRCh38.phased_gt1.vcf.gz, https://github.com/bcgsc/ntEdit) loads the generated BF and, using the human genome (hg) reference GRCh38, interrogates every k-mer it comprises in turn, querying the BF for presence/absence. For each k-mer of the hg reference, ntEdit permutates the 3'-end base and evaluates its presence in the BF, as well as k/j k-mer variants (where -j is a user-defined jump

parameter to skip k-mers, thus limiting the search space) overlapping the base under scrutiny (Fig. S1). When the base under consideration is supported by a sufficient number of k-mers ($k_{ct}/j \geq Y$), it is reported as a single nucleotide variant (SNV) and output by ntEdit in a variant call format (vcf) file (_variants.vcf). ntRoot loads a succinct version of the 1kGP integrated variant call sets where reported SNVs have an allele frequency (AF) of 1% or more in at least one human super-population (EAS, AFR, EUR, SAS or AMR) and the SNVs reported by ntEdit are cross-referenced against the 1kGP call set input vcf.

On genome sequence assemblies (WGA) and other 1X representation of an individual's genome, ntRoot builds a BF using the btllib common code library[22] (ntroot --genome). The remaining steps in WGA follow the same process as WGS.

Once the ntEdit execution has completed, the ntEdit output vcf is parsed, summing and averaging the variant AF for each of the five super-populations while tracking the number and rate of non-zero (nz) AF SNVs over all cross-referenced SNVs. Separate tallies are kept for the whole genome and for each 5 Mbp tile (user-defined parameter --tile, default 5 Mbp). Global (GAI) and local (LAI) ancestry inference scores are calculated as:

$$Average\ AF\ x\ nz\ SNV\ rate$$

for the whole genome and locally, for each tile. For a super-population sp, the nz SNV rate is defined as:

$$nz\ SNV\ rate_{sp} = \frac{nz\ AF\ SNV\ count_{sp}}{total\ nz\ AF\ SNV\ count}$$

The GAI and LAI scores are bounded within the range [0, 1], with 0 and 1 representing lowest and highest confidence, respectively. The highest-scoring ancestry is reported for the former (GAI), globally, and the associated super-population label assigned to each tile for the latter (LAI), locally. When each tile has been assigned, the LAI super-population fraction for super-population sp is reported as:

$$LAI_fraction_{sp} = \frac{\sum_{i=1}^{num_tiles} length(tile_i)\ if\ tile_{i_{super-population}} == sp}{\sum_{i=1}^{num_tiles} length(tile_i)}$$

The GAI and LAI scores are output in a tab-separated variable (tsv) output file (_ancestry-predictions_tile<tile size>.tsv). When the --lai option is specified, ntRoot will additionally output the ancestry predictions and scores per tile (_ancestry-predictions-tile-resolution_tile<tile size>.tsv).

Ancestry inference of 266 1kGP individuals (validation set), using whole genome sequencing datasets

The 30X WGS data was downloaded from https://www.internationalgenome.org/data-portal/data-collection/30x-grch38 (Supplementary Table S10). We randomly selected 266 individuals, with 10-12 representatives per each 26 human populations (defined in[2,16]) and downloaded the associated WGS data with enaBrowserTools (https://github.com/enasequence/enaBrowserTools). We executed ntRoot while tracking the wall-clock run time and memory. An example ntRoot command on a 1kGP WGS dataset:

```
ntroot -k 55 --draft GRCh38.fa --reads ERR3242308_ -t 48 -Y 0.55 -l
1000GP_integrated_snv_v2a_27022019.GRCh38.phased_gt1.vcf.gz
```

where GRCh38.fa contains the GRCh38 human genome reference sequence with chromosome numbers as FASTA header, ERR3242308_ is the fastq file prefix for the sequencing reads and 1000GP_integrated_snv_v2a_27022019.GRCh38.phased_gt1.vcf.gz the integrated variant call set from 1kGP. See Supplementary Table S10 for the file source.

Ancestry inference of 100 1kGP individuals (benchmarking set), using whole genome sequencing datasets

The 30X WGS data was downloaded from https://www.internationalgenome.org/data-portal/data-collection/30x-grch38. We randomly selected 100 individuals in total, 20 for each of the five super-populations (defined in[2,16]), ensuring no overlap with the 1kGP WGS validation set, and downloaded the associated WGS data with enaBrowserTools (https://github.com/enasequence/enaBrowserTools). We executed ntRoot (v1.0.0; ntroot --reads -t 48) on each while tracking the wall-clock run time and memory.

To perform ancestry inference using SNVstory[14], we followed the approach described in the SNVstory manuscript. For each of the 100 1kGP individuals, we aligned the corresponding short reads to the human reference GRCh38 using bwa mem[23] (v 0.7.17-r1188; -t 48 -R "@RG\tID:1\tLB:lib1\tPL:ILLUMINA\tSM:<sample name>\tPU:unit1"). Next, we generated a vcf file from the short-read alignments using the GATK[24,25] suite of tools. The necessary files for the GRCh38 reference genome were generated using samtools[26] faidx (v1.17) and gatk CreateSequenceDictionary (v4.5.0.0). Next, the variants were called using gatk HaplotypeCaller (v4.5.0.0), with separate runs launched for each chromosome (option -L), and at most 12 parallel jobs running at a given time. For each individual, all chromosome-specific vcf files generated using gatk HaplotypeCaller were merged using bcftools[26] concat (v 1.17; --threads 48). Finally, the ancestry for the individual was predicted using SNVstory[14] (v3.0.2; --genome-ver 38 --mode WGS), providing the generated vcf file as input. The GAI inference for SNVstory was taken as the 1kGP continental prediction with the highest probability.

Ancestry inference on reference, 1kGP-independent WGS datasets

In separate experiments, we downloaded the whole-genome sequencing datasets (Supplementary Table S10) of an individual of European ancestry, HuRef (Illumina HiSeq X Five) and of Korean ancestry, KOREF (Illumina HiSeq X Ten / PacBio_SMRT Sequel II) and executed ntRoot (v1.0.0; ntroot --reads -t 48) on those datasets while tracking the wall-clock run time and memory. This experiment served as an independent validation of our 1kGP-based approach, since HuRef and KOREF are not part of the 1kGP cohort.

Ancestry inference on 279 WGS datasets (1kGP-independent, discovery set) from the Simons Genome Diversity Project

We downloaded all available 279 genomes sequencing datasets (WGS) from the diverse populations of the Simons Genome Diversity project[3] using enaBrowserTools and executed ntRoot (v1.0.0; ntroot --reads -t 48) on each read set separately while tracking the wall-clock run time and memory. We tabulated the fractions of ntRoot ancestry inference from every tsv output (in _ancestry-predictions_tile<tile size>.tsv) within the cohort and detailed in

Supplementary Table S5. Subsequently, we represented these fractions for a randomly-selected representative individual, for each of the distinct 129 populations represented in Figure 1d. This large data served as a large-scale and independent validation of our approach.

Ancestry inferences using complete and draft whole genome sequences

In separate experiments, we downloaded the complete diploid genome assemblies (Supplementary Table S10) of an individual of European ancestry (Super-population EUR: European), HuRef/HuRefPrime (complete genome, principal/alternate haplotypes), of Han-Chinese (EAS: East Asian) ancestry, CN1mat/pat (complete genome, maternal/paternal haplotypes), Korean (EAS: East Asian) ancestry, KOREF (complete genome, pseudohaploid) and African (AFR: African) ancestry, HG02055 (complete genome, maternal and paternal haplotypes) and executed ntRoot (v1.0.0; ntroot --genome -t 48) while tracking the wall-clock run time and memory. We also ran ntRoot on a synthetic HuRef-CN1 admix diploid WGA and three pseudohaploid genome assemblies[27] of HG02055 to test ntRoot on admixed and pseudohaploid assemblies vs. separated, maternal and paternal haplotypes. For the synthetic HuRef-CN1 admix diploid experiment, using a custom script (available on zenodo), we first randomly selected complete paternal or maternal chromosome sequences from HuRef and HuRefPrime (random_chromosome_mix.py GCA_000002125.1_HuRef_genomic.chr-renamed.fa GCA_000212995.1_HuRefPrime_genomic.chr-renamed.fa HuRef-hap.fa), and then separately from CN1mat and CN1pat (random_chromosome_mix.py GWHCBHM00000000.genome.fasta GWHCBHQ00000000.genome.fasta CH1hap). We then executed ntRoot (v1.0.0; ntroot --genome -t 48) on the combined HuRef-hap.fa and CN1-hap.fa FASTA files.

Circular genome representations

For all aforementioned experiments and for each human super-population, we tracked tile assignments and SNV density in .tsv files for all non-zero AF SNVs reported by ntRoot and cross-referenced with the 1kGP integrated variant call set. That information was then plotted in R using circlize[28]. All scripts developed to support our study are available on zenodo (10.5281/zenodo.10869033).

Benchmarking

ntRoot was mainly (WGA and 1kGP validation set) benchmarked on a server with 144 Intel(R) Xeon(R) Gold 6150 2.70GHz CPUs with 2.1TB RAM, using 48 threads. Additionally, we used a server-class system with 144 Intel(R) Xeon(R) Gold 6254 3.1 GhZ CPUs with 2.9 TB RAM for processing the large SGDP WGS data volumes and benchmarking SNVstory. We extracted the wall clock run time and peak memory from the output of the UNIX time command preceding all ntRoot and SNVstory processes.

Data availability

The data used in this study is summarized in supplementary material, Supplementary Table S10.

Code availability

ntRoot is freely available on GitHub (https://github.com/bcgsc/ntroot).

Acknowledgements

This study is supported by the Canadian Institutes of Health Research (CIHR) [PJT-183608, I.B.]. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funding organizations. The funding organizations did not have a role in the design of the study, the collection, analysis and interpretation of the data, or in writing the manuscript.

References

1. Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426**, 789–796 (2003).
2. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
3. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

4. Tang, H. *et al*. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol. Off. Publ. Int. Genet. Epidemiol. Soc.* **28**, 289–301 (2005).
5. McVean, G. A. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
6. Pritchard, J. K. *et al*. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
7. Conley, A. B. *et al.* Rye: genetic ancestry inference at biobank scale. *Nucleic Acids Res.* **51**, e44 (2023).
8. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
9. Alexander, D. H. *et al*. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
10. Sundquist, A. *et al*. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Res.* **18**, 676–682 (2008).
11. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5**, e1000519 (2009).
12. Maples, B. K. *et al*. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
13. Uren, C. *et al*. Putting RFMix and ADMIXTURE to the test in a complex admixed population. *BMC Genet.* **21**, 1–8 (2020).
14. Bollas, A. E. *et al.* SNVstory: inferring genetic ancestry from genome sequencing data. *BMC Bioinformatics* **25**, 76 (2024).

15. Warren, R. L. *et al.* ntEdit: scalable genome sequence polishing. *Bioinforma. Oxf. Engl.* **35**, 4430–4432 (2019).

16. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).

17. Willerslev, E. & Meltzer, D. J. Peopling of the Americas as inferred from ancient genomics. *Nature* **594**, 356–364 (2021).

18. Levy, S. *et al.* The Diploid Genome Sequence of an Individual Human. *PLoS Biol.* **5**, e254 (2007).

19. Yang, C. *et al.* The complete and fully-phased diploid genome of a male Han Chinese. *Cell Res.* **33**, 745–761 (2023).

20. Kim, H.-S. *et al.* KOREF_S1: phased, parental trio-binned Korean reference genome using long reads and Hi-C sequencing methods. *GigaScience* **11**, giac022 (2022).

21. Bloom, B. H. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM* **13**, 422–426 (1970).

22. Nikolić, V. *et al.* btllib: A C++ library with Python interface forefficient genomic sequence processing. *J. Open Source Softw.* **7**, 4720 (2022).

23. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013) doi:10.48550/ARXIV.1303.3997.

24. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

25. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. (O'Reilly Media, 2020).

26. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, (2021).

27. Wong, J. *et al.* Linear time complexity de novo long read genome assembly with GoldRush. *Nat. Commun.* **14**, 2906 (2023).

28. Gu, Z. *et al*. circlize Implements and enhances circular visualization in R. *Bioinforma. Oxf. Engl.* **30**, 2811–2812 (2014).