Genome Biology

RESEARCH

Open Access

A refined analysis of Neanderthal-introgressed sequences in modern humans with a complete reference genome

Shen-Ao Liang[1†], Tianxin Ren[2†], Jiayu Zhang[1†], Jiahui He[3], Xuankai Wang[2], Xinrui Jiang[2], Yuan He[3], Rajiv C. McCoy[4], Qiaomei Fu[5,6], Joshua M. Akey[7], Yafei Mao[2,8*] and Lu Chen[1*]

†Shen-Ao Liang, Tianxin Ren and Jiayu Zhang contributed equally to this work.

*Correspondence: yafmao@sjtu.edu.cn; lu_chen@fudan.edu.cn

[1] State Key Laboratory of Genetic Engineering, Center for Evolutionary Biology, School of Life Science, Fudan University, Shanghai 200438, China
[2] Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education, Shanghai Jiao Tong University, Shanghai 200030, China
Full list of author information is available at the end of the article

Abstract

Background: Leveraging long-read sequencing technologies, the first complete human reference genome, T2T-CHM13, corrects assembly errors in previous references and resolves the remaining 8% of the genome. While studies on archaic admixture in modern humans have so far relied on the GRCh37 reference due to the availability of archaic genome data, the impact of T2T-CHM13 in this field remains unexplored.

Results: We remap the sequencing reads of the high-quality Altai Neanderthal and Denisovan genomes onto GRCh38 and T2T-CHM13. Compared to GRCh37, we find that T2T-CHM13 significantly improves read mapping quality in archaic samples. We then apply IBDmix to identify Neanderthal-introgressed sequences in 2504 individuals from 26 geographically diverse populations using different reference genomes. We observe that commonly used pre-phasing filtering strategies in public datasets substantially influence archaic ancestry determination, underscoring the need for careful filter selection. Our analysis identifies approximately 51 Mb of Neanderthal sequences unique to T2T-CHM13, predominantly in genomic regions where GRCh38 and T2T-CHM13 assemblies diverge. Additionally, we uncover novel instances of population-specific archaic introgression in diverse populations, spanning genes involved in metabolism, olfaction, and ion-channel function. Finally, to facilitate the exploration of archaic alleles and adaptive signals in human genomics and evolutionary research, we integrate these introgressed sequences and adaptive signals across all reference genomes into a visualization database, ASH (www.arcseqhub.com).

Conclusions: Our study enhances the detection of archaic variations in modern humans, highlights the importance of utilizing the T2T-CHM13 reference, and provides novel insights into the functional consequences of archaic hominin admixture.

Keywords: T2T-CHM13, Archaic admixture, Adaptive introgression, Database, Complete human genome sequence

Background

The availability of Neanderthal and Denisovan sequencing data has enabled studies on admixture between archaic hominins and modern humans, leading to the discovery of their genetic contributions in human genomes [1–6]. All non-Africans carry approximately 2% Neanderthal-derived DNA [1, 2, 6–9], while recent findings indicate that Africans also harbor more Neanderthal ancestry than previously thought [10]. Individuals from Oceania possess ~ 2–5% Denisovan ancestry, with smaller amounts of Denisovan sequences also present in Asian populations [7, 8, 11, 12]. To better understand the functional, phenotypic, and evolutionary impact of archaic admixture, it is crucial to identify introgressed hominin sequences in modern human genomes. For instance, substantial evidence suggests that some archaic alleles and haplotypes are adaptive and persist at high frequencies in human populations [10, 13–19]. However, most studies of archaic introgression have relied on the GRCh37 human reference genome [8, 10, 11, 13, 14, 20]. The availability of higher-quality and more complete reference genomes, such as GRCh38 and T2T-CHM13, may open new avenues for this field of research [21, 22].

Through the dedicated efforts of the Genome Reference Consortium (GRC), significant improvements were made from GRCh37 to GRCh38 [22]. Despite these advancements, GRCh38 still contains hundreds of gaps and misassembled regions [23]. However, in early 2022, the Telomere-to-Telomere (T2T) Consortium achieved a major breakthrough by assembling the first gapless human genome, T2T-CHM13 [21]. This assembly resolves previously unsequenced and misassembled regions, providing a more accurate reference for human genomics. T2T-CHM13 enhances our understanding of segmental duplications, repeats, epigenomics, genetic diversity, large-scale genomic differences, and centromeres [24–28]. Nevertheless, its impact on archaic introgression patterns in modern humans remains largely unexplored.

In light of this, the availability of the highly complete T2T-CHM13 genome presents a unique opportunity to reevaluate and advance our understanding of archaic genetic legacy in modern populations. To explore this, we remapped archaic sequencing reads onto GRCh38 and T2T-CHM13, followed by variant calling in archaic genomes. Using IBDmix, a modern population reference-free method, we identified Neanderthal-introgressed sequences in a large dataset of sequenced individuals from diverse continental populations. Furthermore, we investigated the impact of different filtering criteria on genetic variants utilized in archaic introgression analyses. Under consistent filtering conditions, we uncovered novel differences in Neanderthal sequences and adaptive introgression signals between T2T-CHM13 and other reference genomes. To facilitate access to these findings, we developed ASH (www.arcseqhub.com), a visualization database that provides a user-friendly platform for exploring archaic segments, their functional implications, and statistical analyses. ASH aims to support researchers studying archaic hominin admixture by offering an interactive and accessible resource for the scientific community.

Results

T2T-CHM13 improves read mapping quality in archaic samples

The high-coverage archaic genomes of the Altai Neanderthal and Denisovan individuals are currently available in the GRCh37 version [2, 6]. To ensure consistency in our analysis, we downloaded the raw sequencing reads from these studies [2, 6], remapped them onto three human reference genomes (GRCh37, GRCh38, and T2T-CHM13), and called variants using GATK [29] (Fig. 1, see the "Methods" section). For modern human genomes, we analyzed samples from the 1000 Genomes Project (1KGP) that were sequenced at $\geq 30\times$ coverage. Due to inconsistent pre-phasing strategies, we rephased the VCF data with both GRCh38 and T2T-CHM13 [21, 30] (see the "Methods" section). To minimize technical biases, we closely matched the analysis pipeline across all reference genomes, ensuring that any major differences were attributable to the reference genome itself rather than methodological artifacts. After processing both archaic and modern genomes, we applied IBDmix, a modern population reference-free method, to identify Neanderthal-introgressed sequences across the genome and consequently detected population-specific adaptive introgression signals in T2T-CHM13. The overall workflow is summarized in the analysis pipeline flowchart (Fig. 1).

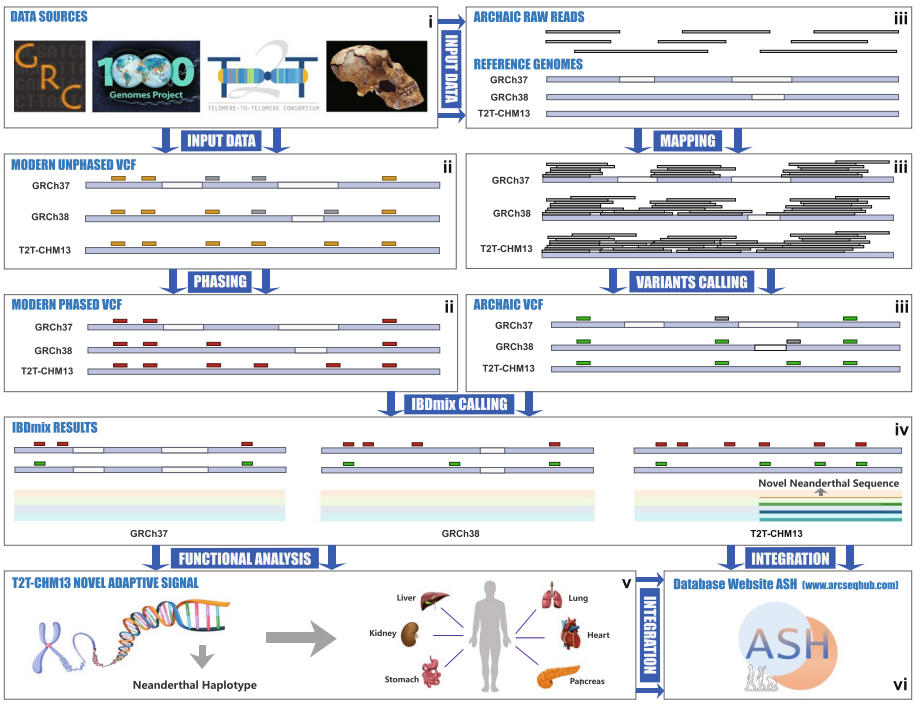

Fig. 1 Workflow overview. The core flow of this study includes six steps. (i) Data collection. (ii) Rephasing genotype data of 2504 samples from 1KGP. In genotype VCFs, rectangles represent high-quality (orange) and low-quality (gray) unphased genotypes. In phased VCFs, red rectangles indicate high-quality phased genotypes. (iii) Remapping archaic sequencing reads on GRCh37, GRCh38, and T2T-CHM13 individually. Rectangles in archaic VCFs represent high-quality archaic variants (green) and low-quality variants (gray). (iv) Detection of archaic introgressed sequence in 1KGP samples based on three references using IBDmix. (v) Functional analysis on adaptive introgression signals in T2T-CHM13. (vi) Integration of the introgression information on three references into a visualization database website, called ASH (www.arcseqhub.com)

A previous study demonstrated that T2T-CHM13 improves paired-end read alignment in modern human samples across populations [21]. To assess whether T2T-CHM13 also enhances short-read alignment in archaic genomes, we examined mapping and variant-calling metrics for Neanderthal and Denisovan sequencing data. Compared to GRCh38, an additional 1.9×10^7 (1.03%) sequencing reads were mapped to T2T-CHM13 (Additional file 2: Table S1). Notably, relative to the original GRCh37 reference, T2T-CHM13 significantly improved mapping rates across all chromosomes (Fig. 2a). This effect was particularly pronounced for acrocentric chromosomes, where mapping rates increased from ~ 80% up to > 95%. Furthermore, using a Wilcoxon rank-sum test, we observed a significant reduction in the standard deviation of read depth, a proxy for mapping quality in complex genomic region [31]. GRCh38 showed a significant improvement over GRCh37 (p-value < 1e − 323), while T2T-CHM13 further outperformed GRCh38 (p-value = 2.2e − 104), indicating greater coverage uniformity (Fig. 2b). The analysis was replicated using Denisovan genome data, yielding consistent results (Additional file 1: Fig. S1a, Fig. S1b). Collectively, these findings demonstrate that T2T-CHM13 significantly improves read mapping and variant detection in archaic genomes, enhancing the accuracy of introgression analyses.

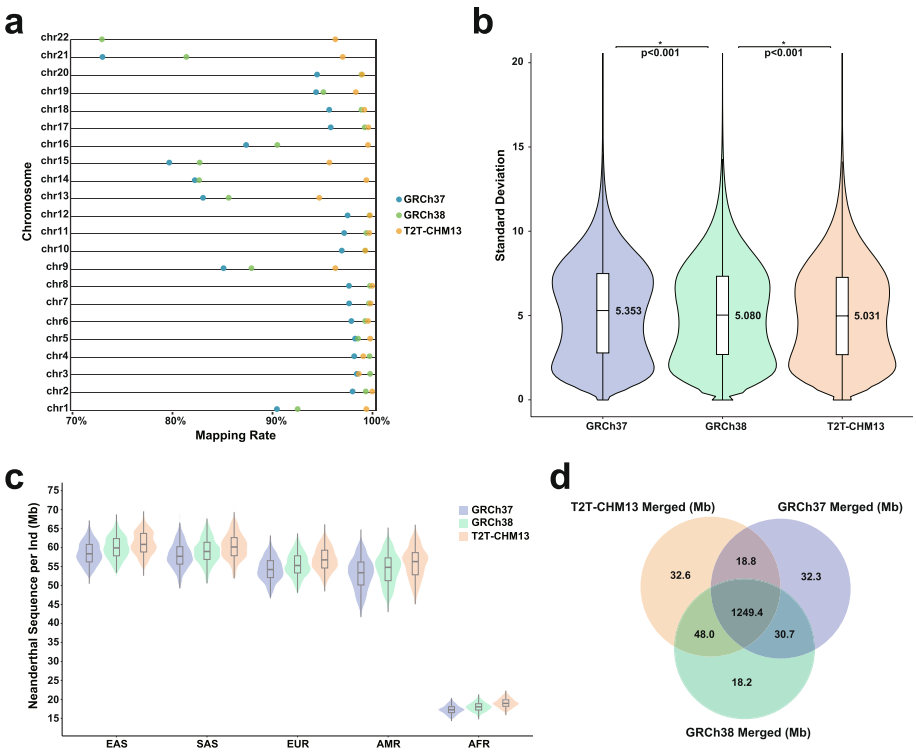

Fig. 2 Comparison of Neanderthal ancestry across three reference genomes. **a** Read mapping rate of the Neanderthal sample in GRCh37, GRCh38, and T2T-CHM13. **b** The standard deviation (s.d.) of read depth in GRCh37, GRCh38, and T2T-CHM13. **c** Violin plot showing Neanderthal sequence identified in five populations from 1KGP data. **d** Venn diagram of the amount of Neanderthal sequence covered in the genome across GRCh37, GRCh38, and T2T-CHM13 references

Pre-processing pipelines in public 1KGP phased data introduce bias in Neanderthal ancestry estimates

While preparing modern human variation data from the 1000 Genomes Project (1KGP) for archaic introgression detection, we observed substantial discrepancies between the pre-phasing filtering strategies used to generate phased Variant Call Format (VCF) files for GRCh38 (Strategy 1) [30] and T2T-CHM13 (Strategy 2) (see the "Methods" section) [32]. To assess the impact of these distinct filtering strategies on archaic introgression analysis, and to establish a universal criterion that minimizes differences attributable to technical artifacts rather than reference genome variations, we replicated both strategies on unphased VCFs relative to GRCh38 and T2T-CHM13 [30, 33]. We then performed phasing using Shapeit [34] and identified Neanderthal sequences using IBDmix [10] on the rephased datasets.

We observed the pattern of excluded biallelic variants was consistent between GRCh38 and T2T-CHM13 under identical pre-phasing strategies (Additional file 1: Fig. S2a, Fig. S2b). However, the primary differences between the strategies stemmed from the Minor Allele Count (MAC) cutoff and Variant Quality Score Log Odds (VQSLOD) cutoff (Additional file 1: Fig. S2a, Fig. S2b, Additional file 2: Table S2). Notably, singletons are excluded to mitigate potential genotype-calling artifacts in the IBDmix approach used for Neanderthal sequence detection. Importantly, the stringent VQSLOD threshold in Strategy 2 introduced a systematic bias, leading to a higher proportion of variants supporting evidence against identity-by-descent (IBD)—specifically, an increase in "2–0" genotype pattern variants via the IBDmix algorithm (see the "Methods" section). This effect was primarily driven by the large number of variants absent in modern human data (Additional file 1: Fig. S3a, Fig. S3b).

As expected, this bias extended to Neanderthal introgression callsets. In GRCh38 callsets, the mean number of Neanderthal-introgressed sequences per individual was 15–20% higher when using Strategy 1 compared to Strategy 2 (Additional file 1: Fig. S4a), a result that remained consistent across populations. Significantly, the discrepancy was even more pronounced in T2T-CHM13-based callsets, where Strategy 1 identified up to 40% more Neanderthal sequences than Strategy 2 (Additional file 1: Fig. S4b). This finding aligns with our observation that Strategy 2 yielded a higher proportion of "2–0" genotype pattern alleles in T2T-CHM13 compared to GRCh38 (~70% vs. ~40%, Additional file 1: Fig. S3a).

Overall, our results reveal that variant pre-processing pipelines in public 1KGP phased panels can introduce significant discrepancies in Neanderthal ancestry detection using IBDmix. To ensure a balance between variant quantity and quality, and to minimize pipeline-induced biases, we conducted all subsequent analyses using Strategy 1, which applies a high-quality yet less stringent filtering approach, for both GRCh38 and T2T-CHM13 datasets.

IBDmix identifies more Neanderthal sequences in T2T-CHM13

Using a concordant pre-phasing strategy, we analyzed datasets from GRCh38 and T2T-CHM13 to identify Neanderthal sequences segregating in modern human populations. All analyses were performed on autosomes and queryable genomic regions only (Additional file 2: Table S3, see the "Methods" section). The genomic regions included for

archaic ancestry detection were similar in size across the reference genomes (Additional file 2: Table S3). Compared to previously reported GRCh37 callsets [10], we identified more Neanderthal sequences per individual in both GRCh38 and T2T-CHM13, with this enrichment being consistent across populations (Fig. 2C). However, the enrichment was modest, averaging ~ 1 Mb more Neanderthal sequence per individual when comparing GRCh37 to GRCh38 or GRCh38 to T2T-CHM13 (Additional file 1: Fig. S5, Additional file 2: Table S4).

Notably, our analysis revealed approximately 51.3 Mb of Neanderthal sequences uniquely identified in T2T-CHM13 compared to GRCh38 (Fig. 2d). Of these, ~ 1.68 Mb fall within the "newly-resolved" regions of T2T-CHM13 [21]. Despite these unique discoveries, we observed substantial overlap among the three reference genomes, with ~ 94% of the Neanderthal sequences identified in T2T-CHM13 being shared with those previously reported in GRCh37 (Fig. 2d) [10]. Due to pipeline discrepancies making GRCh37 less comparable to GRCh38 and T2T-CHM13, subsequent analyses focused on the comparison between GRCh38 and T2T-CHM13.

Small-scale variations impact the identification of Neanderthal sequences

Compared to GRCh38, we identified 2087 novel Neanderthal-introgressed segments in the T2T-CHM13 callset, spanning approximately 51.3 Mb of the genome. Of these, 242 segments (~ 15.92 Mb) do not overlap with introgressed sequences identified in GRCh38 and are designated as "independent sequences." The remaining 1,845 segments (~ 35.35 Mb) extend introgressed sequences already present in GRCh38 and are referred to as "extending sequences" (Fig. 3a). Notably, the lengths of these segments vary significantly, ranging from a few base pairs to several hundred kilobases. The length distribution revealed two distinct peaks: the first representing the "extending sequences" and the second corresponding to the "independent sequences" (Fig. 3b).

Small genetic differences between the T2T-CHM13 and GRCh38 assemblies can affect read mapping quality and genotyping accuracy, potentially leading to discrepancies in the detection of archaic introgressed signals via IBD inference. To investigate whether the novel Neanderthal-introgressed signals observed in T2T-CHM13 stems from local genetic differences between the reference genomes, we utilized PAV [35, 36] to identify genetic variants across the genome. Variants exceeding 10 bp in length, including insertions, deletions, and inversions, were systematically screened and intersected with the T2T-CHM13 novel introgressed segments. This analysis revealed that 1564 segments (74.94% of the total) spanning ~ 40.31 Mb (78.57% of the genomic coverage of the total segments) overlapped with 4,196 variants ranging in size from 10 bp to 1.16 Mb (Fig. 3c, Additional file 2: Table S5).

Furthermore, among these T2T-CHM13-specific Neanderthal-introgressed segments associated with local variants, there are intriguing cases particularly noteworthy. For instance, a ~ 1 kb insertion on chromosome 9 in T2T-CHM13 introduces novel signals of Neanderthal introgression, predominantly found at high frequencies (> 5%) in all non-African populations. These Neanderthal introgression signals span the gene *MUSK*, which is implicated in congenital myasthenic syndrome (Fig. 3d) [37]. These findings underscore how local genetic differences between reference genomes influence the detection of archaic introgression. Importantly, the complete and more accurate

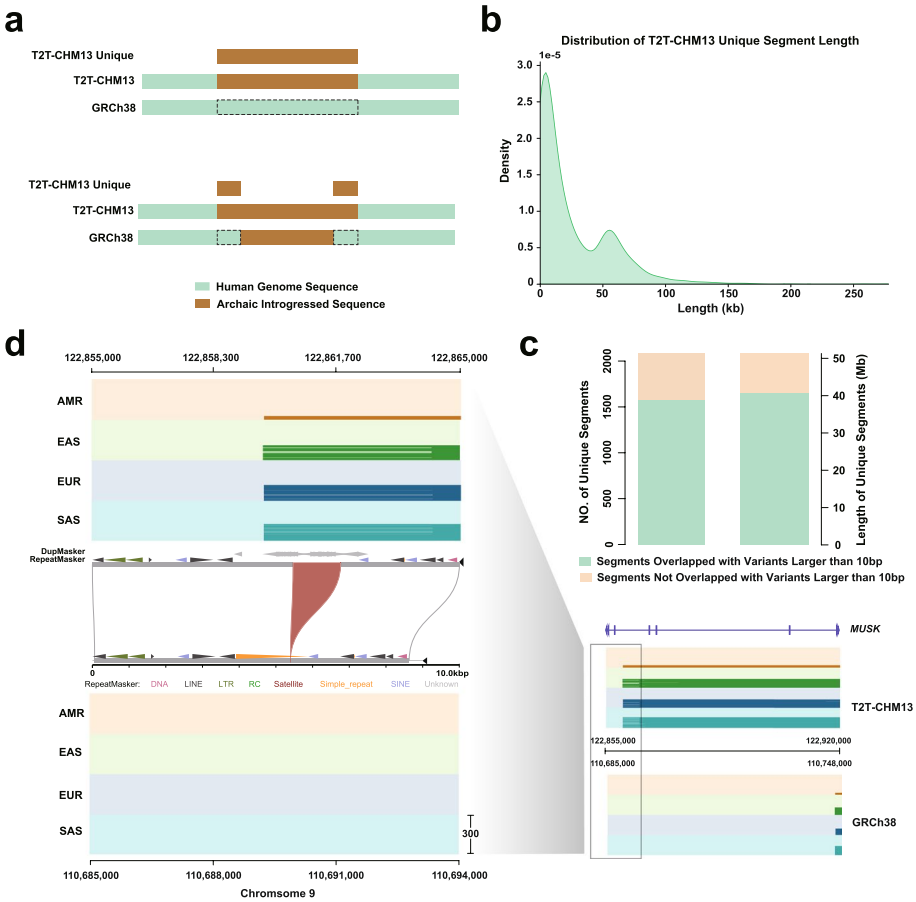

Fig. 3 Enrichment of T2T-CHM13 unique introgressed sequences overlapping genetic variants between T2T-CHM13 and GRCh38. **a** Schematic representation of T2T-CHM13 unique Neanderthal sequences in the genome. The brown box indicates the Neanderthal-introgressed sequence, while the green box represents the modern human sequence. Two types of T2T-CHM13 unique introgressed sequence are shown: completely independent sequences (top) and extensions of the GRCh38 sequences (bottom). **b** Length distribution of T2T-CHM13 unique sequences. **c** Barplot shows the count (left) and length (right) of T2T-CHM13 unique sequences overlap with variants > 10 bp. **d** An example of T2T-CHM13 unique introgressed sequences covering the *MUSK* gene on chromosome 9 due to a ~ 1 kb insertion. Neanderthal segments are in dark brown (AMR), green (EAS), blue (EUR) and cyan (SAS)

T2T-CHM13 reference genome provides new insights into archaic introgressed signals that were previously undetectable.

Novel signals of adaptive Neanderthal introgression in the T2T-CHM13 reference

Previous studies have reported the instances of adaptive archaic introgression, which are the regions that harbor high-frequency archaic haplotypes in specific populations [10, 16–19]. To extend these findings using alternative reference genomes, we applied established methods to identify population-level adaptive variants [10]. These variants were located within introgressed segments identified by IBDmix, matched the Neanderthal allele, and displayed significant differences in derived allele frequency (DAF) between populations (e.g., Europeans vs. East Asians, Africans vs. Europeans, and Africans vs. East Asians) (Fig. 4a, see the "Methods" section). Across the genomes analyzed, we

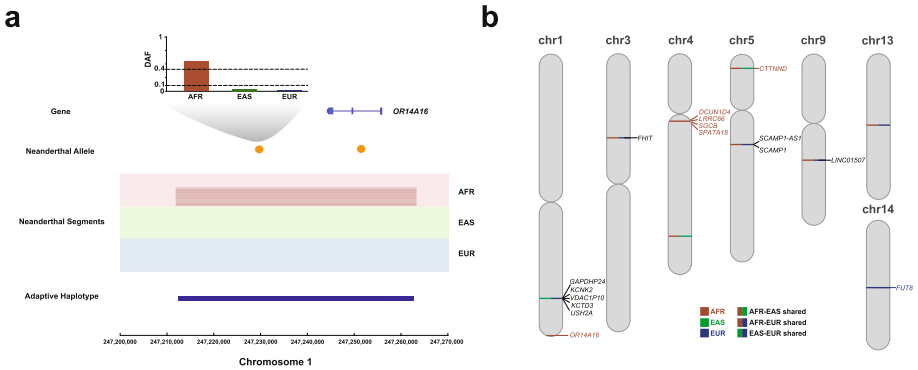

Fig. 4 Novel population-specific high frequency introgressed segments in T2T-CHM13. **a** An illustrated example of T2T-CHM13 novel Neanderthal adaptive haplotype in AFR covering the gene *OR14A16* on chromosome 1. AFR-specific high-frequency-derived alleles (DAF > 40%) that match the Altai Neanderthal genome are shown as orange circles. Neanderthal segments identified by IBDmix are plotted in dark red for AFR, in contrast to the absence of introgression observed in EAS and EUR. The AFR adaptive haplotype merged by Neanderthal segments is shown as a dark blue line. **b** Novel population-specific adaptive haplotypes and related genes identified in T2T-CHM13 across the entire genome are shown in AFR (red), EAS (green), and EUR (blue), mixed colors indicate population-shared high-frequency adaptive haplotypes

identified 87, 87, and 94 population-specific high-frequency Neanderthal haplotypes in GRCh37, GRCh38, and T2T-CHM13, respectively.

We compared high-frequency haplotypes in T2T-CHM13 with GRCh37 and GRCh38. ~ 90% (84/94) of T2T-CHM13 haplotypes were consistent with those found in the other references (Additional file 2: Table S6). Among the ten novel population-specific adaptive haplotypes detected exclusively in T2T-CHM13, two were African-specific and two were non-African-specific (Fig. 4b, Additional file 2: Table S7). These regions included previously recognized targets of adaptive introgression, such as *SGCB* and *SPATA18* [13, 14, 17]. Additionally, these four haplotypes spanned genomic regions newly found to be associated with Neanderthal adaptive introgression, with enrichment for genes involved in metabolism, ion channel function, and olfactory processes, including *FUT8*, *OR14A16*, *KCNK2*, and *KCTD3* (Fig. 4b, Additional file 2: Table S7).

Moreover, we also identified four novel high-frequency haplotypes shared by Africans and Europeans, and two shared by Africans and East Asians in T2T-CHM13, encompassing genes enriched in cancer metabolism, such as *CTNND2*, *FHIT*, and *LINC01507* (Fig. 4b, Additional file 2: Table S7). These findings demonstrate the value of using T2T-CHM13 to uncover new evidence of adaptive introgression, enhancing our understanding of evolutionary history, genomic variations under selection, and their ongoing significance in the human genome.

Database of Neanderthal-introgressed sequences—ASH

To promote broader and more accessible application of archaic hominin admixture studies, we developed ArcSeqHub (ASH), a user-friendly web interface. This platform integrates 985,148, 1,006,918, and 1,037,491 Neanderthal sequences across 2504 samples from geographically diverse populations (1KGP), along with corresponding adaptive Neanderthal introgression signals identified in GRCh37, GRCh38, and T2T-CHM13, respectively. Built using Hypertext Markup Language (HTML) and supplementary

scripts, ASH offers two intuitive search options: the Gene Query and the Locus Query. These tools enable users to visualize Neanderthal-introgressed sequences and associated functional genes, as well as access key statistics, such as introgression ratios across super-populations or individual samples, based on the three reference genomes (Fig. 5). All datasets, statistics, and materials available on ASH are freely downloadable, facilitating ease of use and data accessibility for the research community.

Discussion

In this study, we reprocessed and realigned two archaic genomes onto the GRCh38 and T2T-CHM13 references, leveraging the original raw sequencing reads [2, 6]. This allowed us to analyze Neanderthal ancestry in modern populations using the updated genome references. We then discovered a modest enrichment of introgressed sequences detected by IBDmix when using T2T-CHM13, compared to other references. We also demonstrate that improved variant calling in the fully assembled T2T-CHM13 reference contributes to the identification of unique Neanderthal sequences. Furthermore, we refined population-specific signals of adaptive introgression and developed a database visualizing introgressed segments across diverse modern populations. Overall, our findings provide a valuable resource and novel insights for future studies of archaic admixture using T2T-CHM13.

The differences in detected archaic sequences between T2T-CHM13 and other references primarily arise from improved read mapping and variant calling. T2T-CHM13 enables corrections to variant positions and genotypes, which are crucial for accurate introgression signal detection. Consistent with previous studies showing enhanced accuracy for modern human genome analysis with T2T-CHM13 compared to GRCh38 and GRCh37 [24], we observed similar improvements in archaic genome analyses (Fig. 2a, b, Additional file 1: Fig. S1a, Fig. S1b). Furthermore, collapsed segmental duplications in

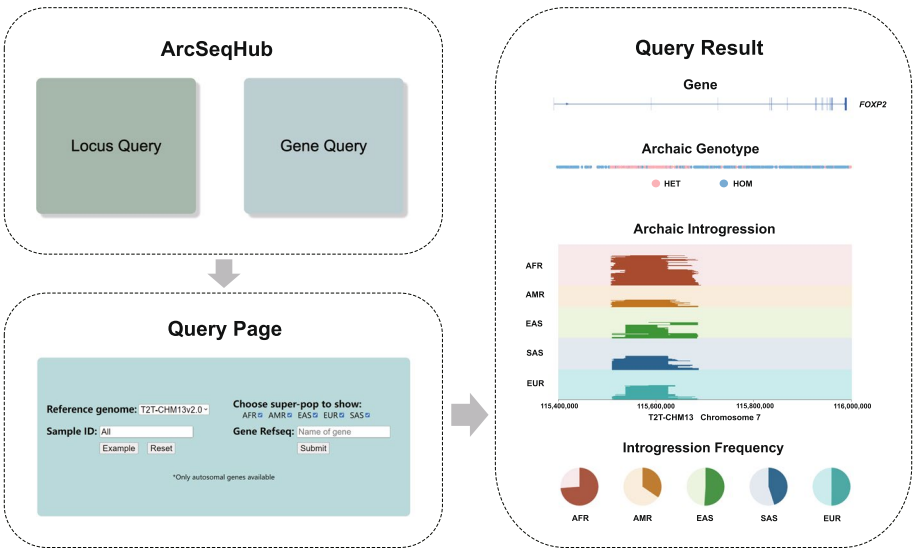

Fig. 5 ArcSeqHub (ASH) database website. The information of Neanderthal-introgressed sequence and the adaptive signals based on three reference genomes are integrated into a visualization database website (www.arcseqhub.com)

GRCh38 and GRCh37, such as the *GPRIN2A/B* region [28, 38], often result in erroneous variant calls. T2T-CHM13's comprehensive segmental duplication map allowed us to exclude such biased regions, improving the quality of variant calls for both archaic and modern human genomes. These improvements are crucial for more reliable detection of introgressed segments.

We also found that discrepancies in introgressed regions between T2T-CHM13 and GRCh38 often stem from small genetic differences between the assemblies (Fig. 3c, d). These differences, caused by genomic polymorphisms or inaccuracies in GRCh38, can disrupt the initiation and extension of introgressed segment detection by IBDmix. Therefore, in addition to use a complete genome reference, accounting for such reference-specific genetic differences is essential in future archaic admixture studies [38]. Incorporating a pangenome graph approach into archaic genome research in the future may provide a promising solution [39].

In the newly resolved 8% of the genome in T2T-CHM13, we identified only ~ 1.6 Mb of novel Neanderthal sequences. This limited discovery is largely attributable to the complexity and repetitive nature of these regions, which include all centromeric regions and the entire short arms of five human chromosomes [21]. Due to the short length of next-generation sequencing reads in archaic genome data, many of these regions were masked during analysis to ensure accurate alignment. Similarly, the modern human genome data used in this study, also generated with next-generation sequencing, face challenges in accurately mapping to these complex regions. To address these limitations, the development of specialized alignment tools tailored for repetitive genomic regions, along with the integration of long-read sequencing data for modern human genomes will be essential. These advancements could provide a more comprehensive understanding of archaic introgression patterns within the newly assembled 8% of the genome in future studies.

One more insight afforded by our analyses is the influence of variant filtering pipelines on the detection of Neanderthal ancestry. Discordant homozygotes (i.e., variants where the genotype of the archaic individual is homozygous alternative and the modern human genotypes are homozygous reference, referred to as "2–0" genotype pattern variants in "Results") serve as strong evidence against introgression in IBDmix [10]. Specifically, we observed that applying a more stringent VQSLOD criterion (Strategy 2) excluded many variants, which would be further misclassified as homozygous reference alleles in modern samples by IBDmix when the archaic genotype is homozygous for the alternative allele. This would introduce a bias against introgression and remarkably reduce the detected amount of Neanderthal ancestry. These findings highlight the importance of carefully selecting variant filtering criteria based on research goals and algorithmic requirements as the advancements of the genomics era emerge. Researchers relying on default or public-released settings risk introducing biases that may compromise conclusions [40]. In our analysis, adjusted VQSLOD filters applied in higher-quality reference panels manually "impute" more sites against IBD by downstream introgression calling method, albeit with producing smaller but more accurate variant sets. The choice of filters involves trade-offs between variant accuracy and the amount of informative data and also depends on distinct characteristics of methods on detecting archaic ancestry.

In conclusion, we demonstrate the difference in archaic introgressed sequence detection by IBDmix between T2T-CHM13 and GRCh38, primarily driven by improvements

in genome assembly, variant calling, and filtering strategies. While T2T-CHM13 identified more Neanderthal sequences using IBDmix, further validation with alternative methods, such as Hidden Markov Models (HMMs) or S* statistics [8, 14, 20], is necessary to confirm these observations. Despite pipeline complexities and method-specific characteristics, our findings underscore the advantages of using T2T-CHM13 as a reference for studying archaic admixture.

Conclusions

Leveraging the complete, more accurate, and representative T2T-CHM13 reference genome, we demonstrated its critical role in advancing studies of archaic admixture. The use of T2T-CHM13 significantly improved the mapping of archaic reads, enabling more accurate detection of introgressed sequences. Additionally, we found that applying different pre-phasing filtering strategies, commonly used in public datasets, resulted in substantial variation in the amount of archaic sequences detected by IBDmix. This underscores the importance of carefully selecting pre-processing pipelines for identifying archaic introgression. Compared to GRCh38, we discovered ~ 51 Mb Neanderthal sequences unique to T2T-CHM13, with approximately 80% of these sequences overlapping ~ 4200 genetic variants that distinguish T2T-CHM13 from GRCh38. Additionally, T2T-CHM13 enabled the discovery of novel population-specific Neanderthal adaptive haplotypes associated with genes involved in metabolism, olfactory function, and ion-channel activity, such as *FUT8*, *OR14A16*, and *KCNK2*. To support further research, we integrated Neanderthal sequences from all references into a publicly accessible database, ASH (www.arcseqhub.com), to facilitate the exploration of archaic alleles and adaptive signals in human genomics and evolutionary studies. Together, our findings highlight the advantages of utilizing T2T-CHM13 in archaic admixture studies, providing new insights into the variation, functionality, and evolutionary significance of archaic ancestry in humans.

Methods

Data in this study

We downloaded Altai Neanderthal and Altai Denisovan sequencing reads from https://www.ebi.ac.uk/ena/browser/view/PRJEB1265 [2, 41], and https://www.ebi.ac.uk/ena/browser/view/PRJEB3092 [6, 42]. We downloaded 1000 Genomes project phased VCFs phase 3 version 5b from https://www.internationalgenome.org/data-portal/data-collection/phase-3 [43–45]. We downloaded 1000 Genomes 30 × genotype VCFs on GRCh38 from https://www.internationalgenome.org/data-portal/data-collection/30x-grch38 [30, 44, 46]. We downloaded 1000 Genomes project genotype VCFs recalled on T2T-CHM13v2.0 from https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/assemblies/variants/1000_Genomes_Project/chm13v2.0/ [21, 47]. We downloaded reference genome data from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/human_g1k_v37.fasta.gz (GRCh37) [43], ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz (GRCh38) [22] and https://github.com/marbl/CHM13 (T2T-CHM13) [21].

Mapping archaic sequencing reads

The mapping approaches employed in this study are consistent with those used in previous research [2, 6]. Initially, the first and last two bases were trimmed from the reads to reduce the effects of remaining ancient DNA damage. Then we mapped the reads of Altai Neanderthal and Altai Denisovan to GRCh37, GRCh38, and T2T-CHM13 reference genomes separately using BWA version 0.7.17 [48] with parameters "–n 0.01 –o 2 –l 65,536". Using BWA's samse command [48], alignments were converted to the SAM format, and then we sorted the SAMs and converted to BAMs with SAMtools version 1.3.1 [49] (coordinate-sorted). Afterward, we excluded the unmapped single reads, QC-failed single reads, and reads shorter than 35 bp. Subsequently, we annotated the NM and MD tag in BAM files using SAMtools calmd [49], and then removed reads with an edit distance of more than 20% of the sequence length. Finally, we utilized a Python script [50] to exclude duplications for each library.

After the removal of duplicates, we combined the BAM files for all libraries using SAMtools merge [49]. Then we used GATK IndelRealigner [29] to realign sequences in the identified genomic regions. After local realignment, the NM/MD fields in BAMs were recalculated using SAMtools calmd [49], and reads with an edit distance of more than 20% of the sequence length were removed.

To obtain mapping statistics for reads, we employed SAMtools stats [49]. To assess the coverage of reads across each base in the genome, we utilized BEDtools version 2.30.0 genomecov [51]. Additionally, to further assess read mapping quality, we calculated the standard deviation (s.d.) of read depth of mapped reads within the unmasked regions. Initially, contiguous bases in the included regions were treated as individual windows. Then average read depth of each window was estimated, resulting in a set of average read depths for all windows. Subsequently, we obtained the mean read depth across all windows, followed by the calculation of the s.d. of read depth of mapped regions at the whole-genomic level. This approach allows for a comprehensive examination of the distribution of reads across the entire genome. The utilization of the s.d. of read depth in archaic read mapping can be treated as an index to assess the consistency and quality of the mapping process. A lower s.d. of read depth indicates higher quality of the reference genome.

Archaic VCF calling

We used the HaplotypeCaller from GATK version 4.3.0.0 [29] to produce genotype calls for single nucleotide variants (SNVs) and insertions and deletions (INDELs) over all sites separately for Altai Neanderthal and Altai Denisovan. To identify high-quality variants, we excluded heterozygous variants with aligned reads that carry reference_allele_count/alternative_allele_count less than 1/3 or more than 3. Finally, we generated per-chromosome Variant Call Format (VCF) files in block-gzip compressed form with a tabix (http://samtools.sourceforge.net/tabix.shtml) index file based on three reference genomes respectively.

Whole genome data processing

Before conducting archaic introgression calling, it is necessary to perform masking operations on the whole-genome data (1000 Genomes, Altai Neanderthal, and Altai Denisovan genomes). Our procedures closely align with those outlined in the previous study [10], and also with some differences in certain details:

- CpGs mask

CpGs were masked as in [2]. In the case of GRCh38 and T2T-CHM13, we performed "liftover" of the variant data from 15 African hunter-gatherer populations to the GRCh38 and T2T-CHM13 coordinates individually. Different from previous studies, when employing the CpG mask from closely related species, for conservative considerations, we included one-to-many alignment scenarios in the generation of sequence alignments for closely related species. Consequently, this allowed for the masking of nearly all CpG sites.

- Mappability mask

Mappable regions were determined by examining all 35 base long reads that overlap each site. A site is mappable if the majority of overlapping reads are mapped uniquely or without 1-mismatch hit to GRCh37, GRCh38, and T2T-CHM13.

Segmental duplication (SD) mask

- Segmental duplication (SD) mask

SD for three reference genomes were removed and downloaded from: https://genome.ucsc.edu/cgi-bin/hgTables [52].

- Indel mask

Sites within 5 bp of INDELs in VCFs were removed.

- Accessibility mask

The 1000 Genomes accessibility mask was applied to three reference genomes.

GRCh37 accessibility mask data were obtained from: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20141020.strict_mask.whole_genome.bed.

T2T-CHM13 accessibility mask data were obtained from: https://github.com/arangrhie/T2T-HG002Y/tree/main/accessibility_masks.

In the case of GRCh38, given the high-coverage nature of the 1000 Genomes data used, the accessibility mask data published was not utilized. Instead, we employed a methodology analogous to that for T2T-CHM13 [24] to generate the accessibility mask for GRCh38.

For archaic samples, we applied three filters as in [2].

- Tandem repeat filter (TRF)

We downloaded the Tandem Repeat Finder annotation for GRCh37, GRCh38, and T2T-CHM13 from: https://genome.ucsc.edu/cgi-bin/hgTables [52].

- Mapping quality filter (MQ30)

We computed the root-mean-square mapping quality from BAMs using a custom Python script, as opposed to extracting the MQ (Mapping Quality) field from VCFs [2]. This choice was necessitated by modifications in the GATK tools.

- Genome alignability filter

We produced the map35_50% track, which requires that at least 50% of all possible 35 mers overlapping a position do not find a match to any other position in the genome allowing for up to one mismatch [2].

- T2T-CHM13 cenSat mask

We merged consecutive centromere/αSat Higher Order Repeat (HOR) array regions until ct (centromeric transition region) appeared in the cenSat annotation track from UCSC, and masked these cenSat regions in T2T-CHM13 archaic introgression analysis.

Neanderthal introgression calling

Following the completion of all masking procedures, we employed a typical archaic introgression detection method without any modern reference panel, IBDmix [10] to identify the Neanderthal sequence in all populations including Africans. Utilizing phased VCFs of 2504 modern samples, archaic VCFs, and the masked region BEDs, we executed IBDmix with parameters (1) –LOD-threshold 4.0, (2) –minor-allele-count-threshold 1, (3) –archaic-error 0.01, (4) –modern-error-max 0.002, (5) –modern-error-proportion 2 and (6) the cutoff for putative introgressed sequence length at 50 kb. Finally, we obtained five distinct sets of IBDmix results corresponding to three different reference genomes (one for GRCh37, and two datasets each for GRCh38 and T2T-CHM13, owing to distinct pre-phasing filtering strategies).

Phasing modern VCFs

We needed modern phased 1KGP panels combined with archaic genotype panels to call archaic introgression. However, the pre-phasing filtering strategies applied to the 1KGP phased VCF datasets for GRCh38 and T2T-CHM13 differ [30, 32]. The pre-phasing filtering strategy for public GRCh38 phased data (Strategy 1): (1) FILTER (column in the VCF) = PASS, (2) GT missingness rate < 5%, (3) HWE exact test p-value > 1e − 10 in at least > one super-population, (3) mendelian error rate (MER) ≤ 5%, (4) minor allele count (MAC) ≥ 2. The pre-phasing filtering strategy for public T2T-CHM13 phased data (Strategy 2): (1) exclude FILTER (column in the VCF) = PASS, (2) exclude variants with an alt allele of "*" after multiallelic splitting, (3) exclude GT missingness rate < 5%, (4) exclude Hardy–Weinberg p-value < 1e − 10 in any 1000G subpopulation, (5) exclude sites where Mendelian Error Rate (Mendelian errors/num alleles) ≥ 0.05, (6) exclude homoallelic sites (MAC = 0), (7) exclude variants with a high chance of being errors as predicted by computational modeling.

Therefore, to ensure comparability of results, we reapplied the two distinct pre-phasing filtering strategies to the unphased VCF datasets of GRCh38 and T2T-CHM13 separately. Initially, we annotated the unphased VCF files with p values from the Hardy–Weinberg equilibrium (HWE) exact test [53], stratified by super-population, employing the BCFtools version 1.9 fill-tags plugin [54]. Subsequently, multiallelic sites were

segregated into distinct rows, and INDELs underwent left-normalized representation using the BCFtools norm tool [48]. Following this, we computed the Mendelian Error Rate (MER) for each variant row using the Mendelian plugin in BCFtools [54]. Finally, we employed the BCFtools toolkit [54] to filter the variants based on predefined two pre-phasing filtering criteria.

After completing the filtering steps, we employed Shapeit version 5.1.1 [34] for variant phasing. We utilized the phase_common module for variants with a minor allele frequency (MAF) greater than 0.1%, generating a haplotype scaffold. Subsequently, phase_rare module was applied for phasing variants with MAF less than 0.1%. Following this, multiallelic sites were merged into separate rows, and the left-normalized representation of INDELs was performed using the BCFtools norm tool [54]. Subsequent to these preprocessing steps, we utilized the BCFtools toolkit [54] to extract 2,504 unrelated individuals, obtaining four phased VCF datasets (two pre-phasing filtering strategies for two reference genomes) containing haplotypes for these 2,504 individuals.

"2–0" genotype pattern variants

There is strong evidence against the IBD mode (i.e., the allele is less likely to be inherited from the archaic hominins) in IBDmix, which derives from discordant homozygotes (i.e., variants where the genotype of the archaic individual is homozygous alternative and the modern human genotypes are all homozygous reference, referred to as "2–0" genotype pattern variants). These variants may result from imputation as homozygous reference for missing sites in modern humans, as there would be no variation at these sites in the modern human panel. Therefore, a large number of variants are excluded due to the overly stringent VQSLOD cutoff in Strategy 2, which may lead to an unexpectedly larger number of "2–0" genotype pattern variants and significant bias in detecting archaic introgressed signals when combining modern human phased panels with archaic genotype panels.

T2T-CHM13-unique introgression analysis

Due to the limitations in coverage of modern human data for the GRCh37 reference genome and inherent disparities during data processing between GRCh37 and the other two reference genomes, our analysis focused exclusively on the T2T-CHM13 and GRCh38 datasets to investigate the implications of reference genome updates on the emergence of novel archaic introgression signals and mitigate potential biases stemming from data and technical variations.

To identify the novel archaic introgressed sequences in T2T-CHM13, we "liftovered" GRCh38 merged callset to T2T-CHM13 coordinates, after which the sequences were excluded from T2T-CHM13 merged callset. To further scrutinize whether the signals of novel introgression were in those localized different regions between the two reference genomes, we used T2T-CHM13 as the reference genome and GRCh38 as the query to run PAV version 2.0.0 [35, 36], obtaining all variants in GRCh38 relative to T2T-CHM13. Subsequently, we extracted all the variants larger than 10 bp that we thought may influence read mapping quality.

To identify the variants overlapped with the T2T-CHM13 novel archaic introgressed sequences, we processed sequences shorter than 50 kb (Fig. 3a) in a special manner.

Considering the IBDmix cutoff for a minimum sequence length of 50 kb, any alteration in signals within dynamic programming algorithms might hinder the extension of local introgression sequences. Therefore, we extended novel archaic introgressed sequences in T2T-CHM13 shorter than 50 kb and closely connected downstream to GRCh38 introgression sequences, extending them to 50 kb. Similarly, we extended sequences less than 50 kb and closely connected upstream to GRCh38 introgressed sequences, extending them to 50 kb. Additionally, for sequences shorter than 50 kb and tightly connected both upstream and downstream to GRCh38 sequences, we extended them in both directions until they reached 50 kb.

Subsequently, we identified all T2T-CHM13 introgression sequences covering variants larger than 10 bp utilizing BEDtools [51]. To identify significant and larger updates in archaic introgressed signals specific to the T2T-CHM13 reference genome, we further required that the novel introgressed sequences must cover structural variations (variants > 50 bp), and these sequences had to be present in at least 5% of the total population across a minimum of two modern human populations.

Identifying novel population-specific adaptive introgressed signals in T2T-CHM13

Initially, we identified population-specific high-frequency introgressed alleles and merged haplotypes for EUR, EAS, and AFR based on GRCh37, GRCh38, and T2T-CHM13 reference genomes separately, primarily drawing from the methodology outlined in [10]. As for the infer for the derived and ancestral state, in the case of GRCh37, we used derived allele frequencies calculated and ancestral states tagged by 1000 Genomes Project. For GRCh38, the derived and ancestral states were inferred based on the chimpanzee state in the Ensembl v110 EPO 10 primate alignment [55]. As for T2T-CHM13, we "liftovered" the GRCh38 data to T2T-CHM13 coordinates, as there is currently no published EPO version available for T2T-CHM13 coordinates.

To identify novel population-specific adaptive signals in T2T-CHM13, we "liftovered" population-specific high-frequency introgressed alleles and merged haplotypes previously identified in GRCh37 and GRCh38 to T2T-CHM13 coordinates. Subsequently, we excluded GRCh37 and GRCh38 data with T2T-CHM13 coordinates from the T2T-CHM13 callset to obtain the novel introgressed haplotypes set and the novel introgressed Neanderthal alleles set separately employing BEDtools [51]. Finally, we intersected the two datasets utilizing BEDtools [51] to identify novel population-specific merged introgressed haplotypes covered by at least one Neanderthal-derived allele.

The archaic sequence hub website development

The Archaic Sequence Hub (ArcSeqHub) provides a user-friendly interface for efficiently exploring archaic introgression in modern humans. The query page of ArcSeqHub features two approaches: one based on gene name and the other utilizing segment coordinates. Each query allows users to select a subset of super-populations of interest or specify particular samples. The website's interface was crafted and supported with Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), JavaScript (JS), and Django version 4.1 [56]. In addition, a custom R script was developed in-house to show archaic introgression for human samples with the R package transPlotR [57]. The pie charts were generated with ECharts and enhanced with

Bootstrap version 3.4.1. The website operation is supported by the Aliyun server, in conjunction with nginx version 1.20.1. Functional testing has been conducted on various popular web browsers, here, we recommend using Google Chrome for a better experience. We welcome user contributions, suggestions for improvements, and bug reports through www.arcseqhub.com/contact/.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03502-z.

Additional file 1. Figure S1. Comparison of Denisovan statistics across three reference genomes. Figure S2. Count of biallelic variants excluded by each filtering criterion in pre-phasing filtering strategies. Figure S3. '2-0' genotype pattern variants under two pre-phasing strategies. Figure S4. Comparison of Neanderthal introgression detected under different pre-phasing filtering strategies. Figure S5. Difference of Neanderthal introgression per individual between T2T-CHM13 and GRCh38.

Additional file 2. Table S1. The reads mapping statistics of Altai Neanderthal and Denisovan in three reference genomes. Table S2. Count of biallelic variants excluded by each filtering criterion in pre-phasing filtering Strategy 1 & Strategy 2 in T2T-CHM13 and GRCh38. Table S3. Masked region statistics for Altai Neanderthal and Denisovan in three reference genomes. Table S4. Mean value of Neanderthal introgressed sequence (Mb/ind.) in populations in three reference genomes. Table S5. 1,564 T2T-CHM13-unique Neanderthal sequences are overlapped by 4,196 variants larger than 10 bp. Table S6. Population-specific haplotypes identified in GRCh37, GRCh38 and T2T-CHM13. Table S7. T2T-CHM13-unique population-specific haplotypes and covered genes in AFR, EAS and EUR compared with GRCh37.

Acknowledgements
We thank members of the Chen laboratory for their suggestions and discussions. We thank Dr. Aaron B. Wolf for assistance with adaptive introgression analysis.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
L.C. and Y.M. designed the study. J.Z., X.W., X.J., Y.H., J.H. and S.L. performed the data analysis. T.R. built up the database website. R.M., Q.F., J.M.A., T.R., Y.M., S.L. and L.C. wrote the manuscript. All authors read, edited, and approved the manuscript.

Authors' information
Shen-Ao Liang, Tianxin Ren, and Jiayu Zhang contributed equally to this work as first authors.

Funding
This work was supported by grants from the National Natural Science Foundation of China (32270668) and 111 project (B25056) to L.C., and by grants from the Natural Science Foundation of Chongqing, China (CSTB2024NSCQ- JQX0004) and Shanghai Jiao Tong University 2030 Initiative (WH510363003/016) to Y.M.

Data availability
The modern VCFs based on GRCh37 and GRCh38 from 1000 Genomes datasets are publicly available from the International Genome Sample Resource (IGSR) [44] at: https://www.internationalgenome.org/data-portal/data-collection/phase-3 [43, 45], and https://www.internationalgenome.org/data-portal/data-collection/30x-grch38 [30, 46]. The modern VCFs based on T2T-CHM13 is publicly available from Telomere-to-Telomere Consortium CHM13 project at https://github.com/marbl/CHM13 [21, 47]. Sequencing reads data for two previously published high-coverage Altai Neanderthal and Altai Denisovan [2, 6] are publicly available at the European Nucleotide Archive (ENA: https://www.ebi.ac.uk/ena) under the accession numbers: ERP002097 and ERP001519, respectively [41, 42]. The database website ASH in this study can be accessed at www.arcseqhub.com [58]. All data generated or analyzed during this study are included in this published article, its supplementary information files and the database website ASH. Archaic VCFs and introgression calls can be accessed via Zenodo: https://doi.org/10.5281/zenodo.14552025 [59], and ASH: https://www.arcseqhub.com/download/ [58].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Yafei Mao is currently an Editorial Board Member of Genome Biology, but was not when the manuscript was submitted, and had no input into the editorial handling of the manuscript. The other authors have no competing interests to declare.

Author details
[1]State Key Laboratory of Genetic Engineering, Center for Evolutionary Biology, School of Life Science, Fudan University, Shanghai 200438, China. [2]Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education, Shanghai Jiao Tong University, Shanghai 200030, China. [3]Ministry of Education Key Laboratory of Contemporary Anthropology, Center for Evolutionary Biology, School of Life Science, Fudan University, Shanghai 200438, China. [4]Department of Biology, Johns Hopkins University, Baltimore, MD 21212, USA. [5]Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China. [6]University of the Chinese Academy of Sciences, Beijing 100049, China. [7]The Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08540, USA. [8]Center for Genomic Research, International Institutes of Medicine, The Fourth Affiliated Hospital, Zhejiang University, Yiwu 322000, China.

Received: 10 July 2024 Accepted: 11 February 2025
Published online: 17 February 2025

References

1. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, et al. A draft sequence of the Neandertal genome. Science. 2010;328:710–22.
2. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505:43–9.
3. Prufer K, de Filippo C, Grote S, Mafessoni F, Korlevic P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyregne S, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017;358:655–8.
4. Mafessoni F, Grote S, de Filippo C, Slon V, Kolobova KA, Viola B, Markin SV, Chintalapati M, Peyregne S, Skov L, et al. A high-coverage Neandertal genome from Chagyrskaya Cave. Proc Natl Acad Sci U S A. 2020;117:15132–6.
5. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468:1053–60.
6. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338:222–6.
7. Sankararaman S, Mallick S, Patterson N, Reich D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Curr Biol. 2016;26:1241–7.
8. Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, Dannemann M, Grote S, McCoy RC, Norton H, et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science. 2016;352:235–9.
9. Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, Gignoux C, Woerner A, Hammer MF, Slatkin M. Higher Levels of Neanderthal Ancestry in East Asians than in Europeans. Genetics. 2013;194:199-+.
10. Chen L, Wolf AB, Fu W, Li L, Akey JM. Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. Cell. 2020;180(677–687): e616.
11. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell. 2018;173(53–61): e59.
12. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538:201–6.
13. Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, Paabo S, Patterson N, Reich D. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014;07:354–7.
14. Vernot B, Akey JM. Resurrecting surviving Neandertal lineages from modern human genomes. Science. 2014;43:1017–21.
15. Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014:512**:**194–197.
16. Dannemann M, Kelso J. The Contribution of Neanderthals to Phenotypic Variation in Modern Humans. Am J Hum Genet. 2017;101:578–89.
17. Gittelman RM, Schraiber JG, Vernot B, Mikacenic C, Wurfel MM, Akey JM. Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. Curr Biol. 2016;26:3375–82.
18. Racimo F, Gokhman D, Fumagalli M, Ko A, Hansen T, Moltke I, Albrechtsen A, Carmel L, Huerta-Sanchez E, Nielsen R. Archaic Adaptive Introgression in TBX15/WARS2. Mol Biol Evol. 2017;34:509–24.
19. Simonti CN, Vernot B, Bastarache L, Bottinger E, Carrell DS, Chisholm RL, Crosslin DR, Hebbring SJ, Jarvik GP, Kullo IJ, et al. The phenotypic legacy of admixture between modern humans and Neandertals. Science. 2016;351:737–41.
20. Skov L, Hui R, Shchur V, Hobolth A, Scally A, Schierup MH, Durbin R. Detecting archaic introgression using an unadmixed outgroup. PLoS Genet. 2018;14: e1007641.
21. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. The complete sequence of a human genome. Science. 2022;376:44–53.
22. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27:849–64.
23. Church DM. A next-generation human genome sequence. Science. 2022;376:34–5.
24. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, et al. A complete reference genome improves analysis of human genetic variation. Science. 2022:376**:**eabl3533.

25. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ, et al. Complete genomic and epigenetic maps of human centromeres**.** Science. 2022:376:eabl4178.
26. Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, Jain M, Shumate A, Razaghi R, Koren S, et al. Epigenetic patterns in a complete human genome. Science. 2022:376:eabj5089.
27. Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, de Lima LG, Limouse C, Halabian R, Wojenski L, Rodriguez M, et al. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science. 2022:376:eabk3112.
28. Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, Diekhans M, Sulovari A, Munson KM, Lewis AP, et al. Segmental duplications and their variation in a complete human genome. Science. 2022:376**:**eabj6965.
29. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
30. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell. 2022;185(3426–3440): e3419.
31. Yang X, Wang X, Zou Y, Zhang S, Xia M, Fu L, Vollger MR, Chen NC, Taylor DJ, Harvey WT, et al. Characterization of large-scale genomic differences in the first complete human genome. Genome Biol. 2023;24:157.
32. Lalli, J., Bortvin, A., McCoy R., Werling D.W. Phased T2T 1KGP panel (1.1). Datasets. Zenodo. https://doi.org/10.5281/zenodo.7612953 (2024).
33. Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, Hook PW, Koren S, Rautiainen M, Alexandrov IA, et al. The complete sequence of a human Y chromosome. Nature. 2023;621:344–54.
34. Hofmeister RJ, Ribeiro DM, Rubinacci S, Delaneau O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. Nat Genet. 2023;55:1243–9.
35. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021:372**:**eabf7117.
36. Porubsky D, Vollger MR, Harvey WT, Rozanski AN, Ebert P, Hickey G, Hasenfeld P, Sanders AD, Stober C, Human Pangenome Reference C, et al. Gaps and complex structurally variant loci in phased genome assemblies. Genome Res. 2023:33**:**496–510.
37. Chevessier F, Faraut B, Ravel-Chapuis A, Richard P, Gaudon K, Bauche S, Prioleau C, Herbst R, Goillot E, Ioos C, et al. MUSK, a new target for mutations causing congenital myasthenic syndrome. Hum Mol Genet. 2004;13:3229–40.
38. Mao Y, Zhang G. A complete, telomere-to-telomere human genome sequence presents new opportunities for evolutionary genomics. Nat Methods. 2022;19:635–8.
39. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. A draft human pangenome reference. Nature. 2023;617:312–24.
40. Hemstrom W, Grummer JA, Luikart G, Christie MR. Next-generation data filtering in the genomics era. Nat Rev Genet. 2024;25:750–67.
41. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. Raw sequencing reads of the Altai Neanderthal. Datasets. European Nucleotide Archive. https://www.ebi.ac.uk/ena/browser/view/PRJEB1265 (2016).
42. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, et al. Raw sequencing reads of the Altai Denisovan. Datasets. European Nucleotide Archive. https://www.ebi.ac.uk/ena/browser/view/PRJEB3092 (2016).
43. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015:526**:**68–74.
44. Clarke L, Fairley S, Zheng-Bradley X, Streeter I, Perry E, Lowy E, et al. The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. Nucleic Acids Res. 2017;45:D85–9.
45. The 1000 Genomes Project Consortium. VCF files from the 1000 Genomes Project phase 3 release. Datasets. The International Genome Sample Resource. https://www.internationalgenome.org/data-portal/data-collection/phase-3 (2015).
46. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. VCF files of 30× sequencing depth data from the 1000 Genomes Project on GRCh38. Datasets. The International Genome Sample Resource. https://www.internationalgenome.org/data-portal/data-collection/30x-grch38 (2022).
47. Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, Hook PW, Koren S, Rautiainen M, Alexandrov IA, et al. VCF files from the 1000 Genomes Project recalled on T2T-CHM13v2.0**.** Datasets. Github. https://github.com/marbl/CHM13 (2023).
48. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009:25**:**2078–2079.
50. Kircher M. Analysis of High-Throughput Ancient DNA Sequencing Data. In: Shapiro B, Hofreiter M, editors. Ancient DNA. Methods in Molecular Biology, vol 840. Totowa: Humana Press; 2012. p. 197–228. https://doi.org/10.1007/978-1-61779-516-9_23.
51. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
52. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014;42:D764-770.
53. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet. 2005;76:887–93.

54. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.
55. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 2008;18:1814–28.
56. Django Software Foundation. Django. https://djangoproject.com
57. Jun Z. transPlotR: An elegant package to visualize gene structures. https://github.com/junjunlab/transPlotR
58. Liang S, Ren T, Zhang J, et al. Archaic introgressed sequence database-Arcseqhub. http://www.arcseqhub.com
59. Liang S, Ren T, Zhang J, et al. Archaic data based on GRCh38 and T2T-CHM13. 2024. Datasets Zenodo. https://doi.org/10.5281/zenodo.14552025.

Publisher's Note