CONGA: Copy number variation genotyping in ancient genomes and low-coverage sequencing data

Arda Söylev[1*], Sevim Seda Çokoglu[2], Dilek Koptekin[3], Can Alkan[4], and Mehmet Somel[2*]

[1]*Department of Computer Engineering, Konya Food and Agriculture University, Konya, 42080, Turkey*
[2]*Department of Biology, Middle East Technical University, Ankara, 06800, Turkey*
[3]*Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, 06800, Turkey*
[4]*Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey*

[*]*Corresponding authors: asoylev@gmail.com and msomel@metu.edu.tr*

ABSTRACT

To date, ancient genome analyses have been largely confined to the study of single nucleotide polymorphisms (SNPs). Copy number variants (CNVs) are a major contributor of disease and of evolutionary adaptation, but identifying CNVs in ancient shotgun-sequenced genomes is hampered by typical low coverage ($<1\times$) and short fragments (<80 bps), precluding standard CNV detection software to be effectively applied to ancient genomes. Here we present CONGA, tailored for genotyping CNVs at low coverage. Simulations and down-sampling experiments suggest that CONGA can genotype deletions >1 kbps with F-scores >0.75 at $\geq 1\times$, and distinguish between heterozygous and homozygous states. We applied CONGA to genotype 10,002 outgroup-ascertained deletions across a heterogenous set of 71 ancient human genomes spanning the last 50,000 years, produced using variable experimental protocols. A fraction of these (21/71) display divergent deletion profiles unrelated to their population origin, but attributable to technical factors such as coverage and read length. The majority of the sample (50/71), despite originating from nine different laboratories and having coverages $0.44\times$-$26\times$ (median $4\times$) and read lengths 52-121 bp (median 69), exhibit coherent deletion frequencies. Across these 50 genomes, inter-individual genetic diversity measured using SNPs and CONGA-genotyped deletions are strongly correlated. CONGA-genotyped deletions also display purifying selection signatures, as expected. CONGA thus paves the way for systematic CNV analyses in ancient genomes, despite the technical challenges posed by low and variable genome coverage.

Keywords Genomics · ancient DNA · CNV genotyping · deletion · low coverage whole genome sequencing

Introduction

Ancient genomics, the analysis of genetic material extracted from archaeological and paleontological remains, has become a major source of information for the study of population history and evolution over the last decade Skoglund and Mathieson (2018); Frantz *et al.* (2020); Shapiro and Hofreiter (2014); Marciniak and Perry (2017). While the number of published ancient genomes is exponentially growing, their analyses have yet been nearly exclusively limited to those of single-nucleotide polymorphisms (SNPs), while structural variations (SVs) in ancient genomes remain mostly ignored. Copy number variations (CNVs) are a common type of SVs and include deletions and duplications ranging from 50 bps to several megabasepairs. Although their number, by count, is much fewer than SNPs, the fraction of the genome affected by CNVs is well past that accounted for SNPs Conrad *et al.* (2010). Likewise, CNVs are a major contributor to phenotypic variation: they are frequently discovered as the basis of diverse biological adaptations Gonzalez *et al.* (2005); Perry *et al.* (2007); Xue *et al.* (2008); Chan *et al.* (2010); McLean *et al.* (2011); Hardwick *et al.* (2011); Kothapalli *et al.* (2016); Nuttle *et al.* (2016); Hsieh *et al.* (2019) as well as genetic diseases (reviewed in Zhang *et al.* (2009); Stankiewicz and Lupski (2010); Girirajan *et al.* (2011); Saitou and Gokcumen (2020)). This renders the study of CNVs in ancient genomes two-fold attractive. First, as CNVs frequently serve as genetic material for adaptation, their study in ancient genomes can allow detailed temporal investigation of adaptive processes. Examples include evolutionary changes in salivary amylase copy numbers in humans and in dogs, thought to represent

responses to a shift to starch-rich diets Mathieson and Mathieson (2018); Bergström *et al.* (2020). Second, large deletions can be a major source of deleterious mutation load, and studying deletion frequencies in ancient genome samples from extinct species or severely bottlenecked populations can inform about the genetic health of lineages. For instance, a study on the last surviving mammoth population on Wrangel Island reported an excess of deletions in this sample, which may have compromised the population's fitness Rogers and Slatkin (2017).

Despite this appeal, the impact of CNVs on evolutionary history and ancient phenotypes remains largely unexplored Frantz *et al.* (2020). The reason lies in the significant technical challenges in CNV detection posed by ancient genomes. State-of-the-art methods for CNV discovery from shotgun genome sequencing data require at least moderate depth of coverage Abyzov *et al.* (2011); Boeva *et al.* (2012); Smith *et al.* (2015); Alkan (2020) and read-pair information Rausch *et al.* (2012); Layer *et al.* (2014); Chen *et al.* (2016); Eisfeldt *et al.* (2017); Soylev *et al.* (2017, 2019), or long reads Chaisson *et al.* (2015); Sedlazeck *et al.* (2018). However, due to the degraded and elusive nature of ancient DNA, ancient genome data is frequently produced at low coverage ($<1\times$) and the molecules retrieved are typically short, between 50-80 bps. Excess variability in genome coverage caused by taphonomic processes is another potential issue. Although CNVs have been studied in a few relatively high coverage ancient genomes using CNV discovery tools Green *et al.* (2010); Reich *et al.* (2010); Meyer *et al.* (2012); Rogers and Slatkin (2017); Bergström *et al.* (2020), these methods are inapplicable to most ancient genome data sets, and so far, no specific algorithm for CNV identification in ancient genomes has been developed and tested.

With the aim to fill this gap, here we present CONGA (**Co**py **N**umber Variation **G**enotyping in **A**ncient Genomes and Low-coverage Sequencing Data), a CNV genotyping algorithm tailored for ancient and other low coverage genomes, which estimates copy number beyond presence/absence of events. We use simulations and down-sampling experiments to assess CONGA's performance. Beyond simulations, we explore whether deletions can be reliably genotyped in heterogeneous datasets composed of ancient genomes from different laboratories, where not only low coverage, but also coverage variability caused by differences in taphonomy and experimental protocols may pose challenges. We evaluate this by studying expected patterns of genetic drift and negative selection on CONGA-genotyped deletions.

Results

Motivation and overview of the algorithm

We developed CONGA to genotype given candidate CNVs in mapped read (BAM) files (Methods). The choice of CNV genotyping over CNV discovery has obvious reasons: (a) CNV discovery using low coverage ancient genomes is impractical; (b) for many species studied using ancient genomics, CNV reference sets based on high quality genomes are already available (Supplemental Note S1); (c) variants in ancient genomes will largely overlap with present-day variants in most cases; (d) genotyping has much shorter running times and lower memory usage than discovery. Indeed, although algorithms for *de novo* SNP discovery exist Prüfer (2018); Link *et al.* (2017), most ancient genome studies to date have chosen genotyping known variants because of low coverage and DNA damage Orlando *et al.* (2021). We reasoned that it may be likewise possible to genotype CNVs in ancient genomes with high accuracy and in short running times using depth of coverage and split-read information, despite low and variable coverage.

Briefly, CONGA first calculates the number of reads mapped to each given interval in the reference genome, which we call "observed read-depth". It then calculates the "expected diploid read-depth", i.e., the GC-content normalized read-depth given the genome average. Using these values, CONGA calculates the likelihood for each genotype by modeling the read-depth distribution as Poisson, similar to common CNV callers Xie and Tammi (2009); Chiang *et al.* (2009); Yoon *et al.* (2009). The genotypes can be homozygous CNV, heterozygous CNV, or no CNV. Using these likelihoods CONGA then calculates a statistic we term the C-score, defined as the likelihood of a CNV being true (in heterozygous or homozygous state) over it being false (no CNV). For genotyping duplications, CONGA also uses an additional split-read step in order to utilize paired-end information. Briefly, it splits reads and remaps the split within the genome, treating the two segments as paired-end reads Karakoc *et al.* (2012); Soylev *et al.* (2019). Either type of signature, read-depth or paired-end, can be sufficient to call a duplication (Methods). The overall workflow is presented in Figure 1.

Accuracy evaluation using simulated genomes and comparison with published algorithms

To evaluate the performance of CONGA we first simulated genomes with CNVs of ancient-like characteristics. We employed VarSim Mu *et al.* (2015) to insert deletions and duplications into the human reference genome (GRCh37). We used three different size intervals for CNVs: small (100 bps - 1000 bps), medium (1,000 bps - 10,000 bps) and large (10,000 bps - 100,000 bps). We thus simulated three genomes, each with roughly 1,500 deletions and 1,500 duplications of a specific size range (see Supplemental Fig. S1 for the exact numbers and length distributions of CNVs

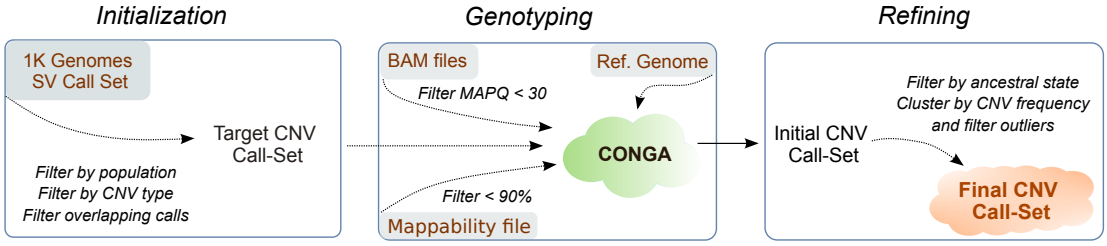

Figure 1: Overall workflow of CONGA. The first step involves initialization, where we create the input (reference) CNV file using the deletions and duplications of a high quality genome set. We apply our genotyping algorithm in the second step and create the initial CNV call set. We then perform a filtering and refining step, which is used to generate the final CNV call set.

inserted in each genome). We next used these genomes as input to the ancient read simulator Gargammel Renaud *et al.* (2017), which generates paired-end short Illumina reads with varying fragment sizes (median 66 bps) and post-mortem damage. The data was generated at various depths: 0.05×, 0.1×, 0.5×, 1× and 5× (Methods). We then used CONGA to genotype CNVs across the simulated ancient genomes, using a candidate CNV call set. In order to assess specificity and sensitivity, we also used a background (false) CNV list, prepared using published deletion and duplication calls from modern-day human long-read sequencing datasets Audano *et al.* (2019); Chaisson *et al.* (2019); Zook *et al.* (2020); Collins *et al.* (2020), as well as from African populations (AFR) from Phase 3 of the 1000 Genomes Project Sudmant *et al.* (2015b). We mixed these false CNVs to the list of true CNVs with a ratio of approximately 10:1 (∼15,000 false events vs. ∼1,500 true events) and used this mixed list as the candidate CNV call set to CONGA (Methods). To assess the performance of CONGA in identifying CNVs, we further compared it with a CNV genotyping tool, GenomeSTRiP Handsaker *et al.* (2011, 2015), and three of the widely used CNV discovery tools: CNVnator Abyzov *et al.* (2011), FREEC Boeva *et al.* (2012) and mrCaNaVaR Alkan *et al.* (2009); Kahveci and Alkan (2018); Alkan (2020).

Table 1 shows true and false predictions by CONGA, GenomeSTRiP, FREEC and CNVnator, as well as their true positive rate (TPR), false discovery rate (FDR) and the F-score (F1) for identifying deletions and duplications of small, medium and large size (as defined above). We report results with mrCaNaVaR separately in Supplemental Table S1.A as this algorithm was specifically designed to target large duplications (>10 kbps) only.

Both genotypers, CONGA and GenomeSTRiP, achieved higher performance compared to the three CNV discovery tools (Table 1; Supplemental Table S1.A). Although this may seem expected, the fact that our candidate CNV call set included 10 times more false CNVs than true CNVs is notable, and indicates that both CONGA and GenomeSTRiP achieve non-trivial performances in distinguishing true versus false CNVs in ancient genomes.

CONGA and GenomeSTRiP had comparable performances, although CONGA had lower FDR and slightly lower recall (TPR) than the latter, leading to overall higher F-scores. We note that GenomeSTRiP was performed on each genome independently here, and its performance when genotyping multiple genomes together could be higher Handsaker *et al.* (2011). However, joint genotyping may also create biases in heterogeneous datasets (see Discussion).

We observed that all tools converge in performance as the coverage approaches depths of 5×, especially with large CNVs. For small CNVs (<1 kbps), all tools under-performed, although CONGA predictions still had higher recall and precision than the other tools (see Supplemental Fig. S2 for precision-recall curves).

The simulation results thus suggest that CONGA can efficiently and accurately genotype deletions and duplications of length >1 kbps in ancient genomes at ≥0.5× coverage, with higher overall accuracy compared to available discovery and genotyping tools.

Copy number predictions of CNVs

Beyond the identification of deletion and duplication events, classifying individual genotypes as heterozygous or homozygous CNVs could provide valuable information for population genetic analyses of CNVs. However, predicting CNV copy numbers can be a significant challenge on low coverage genomes Kousathanas *et al.* (2017). We thus assessed the performance of CONGA to determine the copy number of a CNV based on the likelihood model described above using our simulation data. We focused on medium and large size CNVs given the weak performance of CONGA on small CNVs. We note that CONGA only evaluates the possibility of homozygous duplications (ignores copy numbers

Table 1: Summary of simulation predictions by CONGA, GenomeSTRiP, FREEC and CNVnator.

	Cov.	CONGA					GenomeSTRiP					FREEC					CNVNator				
		T	F	TPR	FDR	F1	T	F	TPR	FDR	F1	T	F	TPR	FDR	F1	T	F	TPR	FDR	F1
Dels (small) *1810 True*	.05×	1471	1887	**0.81**	**0.56**	**0.57**	829	6308	0.46	0.88	0.19	0	1221	0.00	1.00	-	3	47442	0.00	1.00	0.00
	.1×	1266	1440	**0.70**	**0.53**	**0.56**	851	4765	0.47	0.85	0.23	0	198	0.00	1.00	-	0	402	0.00	1.00	-
	.5×	1285	157	**0.71**	**0.11**	**0.79**	853	1549	0.47	0.64	0.41	0	6761	0.00	1.00	-	0	806	0.00	1.00	-
	1×	1410	46	**0.78**	**0.03**	**0.86**	888	719	0.49	0.45	0.52	0	1916	0.00	1.00	-	0	263	0.00	1.00	-
	5×	1593	8	**0.88**	**0.00**	**0.93**	917	89	0.51	0.09	0.65	20	392	0.01	0.95	0.02	341	493	0.19	0.59	0.26
Dups (small) *1751 True*	.05×	601	548	0.34	**0.48**	**0.41**	829	3834	**0.47**	0.82	0.26	0	44	0.00	1.00	-	7	47700	0.00	1.00	0.00
	.1×	719	404	0.41	**0.36**	**0.50**	1048	2691	**0.60**	0.72	0.38	0	7	0.00	1.00	-	0	28699	0.00	1.00	-
	.5×	856	64	0.49	**0.07**	**0.64**	1077	686	**0.62**	0.39	0.61	0	3	0.00	1.00	-	0	9	0.00	1.00	-
	1×	1155	14	**0.66**	**0.01**	**0.79**	1127	311	0.64	0.22	0.71	0	555	0.00	1.00	-	0	884	0.00	1.00	-
	5×	1448	1	**0.83**	0.00	**0.91**	888	1270	0.73	0.05	0.82	35	77	0.02	0.69	0.04	2	0	0.00	**0.00**	0.00
Dels (med.) *1680 True*	.05×	1136	1704	0.68	**0.60**	**0.50**	1430	5670	**0.85**	0.80	0.33	0	83	0.00	1.00	-	0	68	0.00	1.00	-
	.1×	1273	1308	0.76	**0.51**	**0.60**	1452	4383	**0.86**	0.75	0.39	0	237	0.00	1.00	-	1	216	0.00	1.00	0.00
	.5×	1423	171	0.85	**0.11**	**0.87**	1495	1467	**0.89**	0.50	0.64	239	6433	0.14	0.96	0.06	187	257	0.11	0.58	0.18
	1×	1506	53	**0.90**	**0.03**	**0.93**	1501	699	0.89	0.32	0.77	421	2135	0.25	0.84	0.20	330	257	0.20	0.44	0.29
	5×	1569	9	**0.93**	**0.01**	**0.96**	1510	102	0.90	0.06	0.92	929	485	0.55	0.34	0.60	949	423	0.56	0.31	0.62
Dups (med.) *1684 True*	.05×	792	551	0.47	**0.41**	**0.52**	1104	3813	**0.66**	0.78	0.33	0	3	0.00	1.00	-	0	114	0.00	1.00	-
	.1×	950	422	0.56	**0.31**	**0.62**	1160	2701	**0.69**	0.70	0.42	0	3	0.00	1.00	-	0	102	0.00	1.00	-
	.5×	1340	60	**0.80**	**0.04**	**0.87**	1322	685	0.79	0.34	0.72	271	15	0.16	0.05	0.28	2	4	0.00	0.67	0.00
	1×	1451	11	**0.86**	**0.01**	**0.92**	1389	333	0.82	0.19	0.82	582	937	0.35	0.62	0.36	16	2	0.01	0.11	0.02
	5×	1553	1	**0.92**	**0.00**	**0.96**	1473	95	0.87	0.06	0.91	1000	329	0.59	0.25	0.66	105	2	0.06	0.02	0.12
Dels (large) *1385 True*	.05×	1208	1812	0.87	**0.60**	**0.55**	1330	5891	**0.96**	0.82	0.31	0	87	0.00	1.00	-	84	131	0.06	0.61	0.11
	.1×	1251	1309	0.90	0.51	**0.63**	1337	4371	**0.97**	0.77	0.38	0	754	0.00	1.00	-	560	246	0.40	**0.31**	0.51
	.5×	1293	157	0.93	**0.11**	**0.91**	1335	1496	**0.96**	0.53	0.63	664	3136	0.48	0.83	0.26	1049	293	0.76	0.22	0.77
	1×	1299	53	0.94	**0.04**	**0.95**	1338	759	**0.97**	0.36	0.77	1239	156	0.89	0.11	0.89	1204	309	0.87	0.20	0.83
	5×	1299	4	0.94	**0.00**	**0.97**	1336	230	**0.96**	0.15	0.91	1260	154	0.91	0.11	0.90	1265	453	0.91	0.26	0.82
Dups (large) *1532 True*	.05×	1263	563	0.82	**0.31**	**0.75**	1271	3900	**0.83**	0.75	0.38	0	6	0.00	1.00	-	4	354	0.00	0.99	0.00
	.1×	1327	366	**0.87**	**0.22**	**0.82**	1327	2855	**0.87**	0.68	0.46	0	0	-	-	-	455	315	0.30	0.41	0.40
	.5×	1420	58	0.93	**0.04**	**0.94**	1424	964	**0.93**	0.40	0.73	589	97	0.38	0.14	0.53	1039	77	0.68	0.07	0.78
	1×	1426	20	0.93	**0.01**	**0.96**	1445	623	**0.94**	0.30	0.80	1305	266	0.85	0.17	0.84	1216	94	0.79	0.07	0.86
	5×	1428	9	0.93	**0.01**	**0.96**	1447	454	**0.94**	0.24	0.84	1304	294	0.85	0.18	0.83	1350	165	0.88	0.11	0.89

The table shows CNV prediction performances of CONGA, GenomeSTRiP, FREEC and CNVnator on simulated genomes with depths 0.05×, 0.1×, 0.5×, 1× and 5×, for deletions (Dels) and duplications (Dups) of multiple CNV size intervals including 100 bps - 1 kbps (small), 1 kbps - 10 kbps (medium) and 10 kbps - 100 kbps (large). Here, **T** (True) and **F** (False) refer to correct and incorrect predictions respectively, **TPR** is true positive rate (or recall) and **FDR** is false discovery rate $(1 - Precision)$ of each algorithm. **F1** (F-score), is calculated as $(2 \times Precision \times Recall)/(Precision + Recall)$. Bold values in each row represent the highest TPR, lowest FDR, or highest F1 across the tools. See Supplemental Table S1.A for details and mrCaNaVaR predictions for large variations. Commands that we used to run each tool are also given in Supplemental Material. The results here were generated using C-Score <0.5 for CONGA, while no read-pair or mappability filters were applied.

≥3). Figure 2 shows CONGA's copy number prediction performance for deletions and duplications using F-scores for each coverage tested. We found that F-scores were above 0.7 at coverages ≥0.5×. Encouragingly, CONGA had comparable power in identifying heterozygous and homozygous events of size >1 kbps (Supplemental Table S1.B).

Down-sampling experiments with real ancient genomes

We next studied the performance of CONGA in identifying CNVs at various depths of coverage using real ancient genome data. As no ground truth CNV call-set is available, we used the following approach: (i) we chose three published ancient genomes of relatively high coverage (≥9×), (ii) we genotyped CNVs using the full genome data with CONGA and using a modern-day human CNV call set as input, (iii) we down-sampled the ancient genome data to lower coverages, (iv) we assessed CONGA's performance in genotyping the same CNVs at low coverage (Methods).

Specifically, we selected a (∼ 23.3×) ancient Eurasian genome (Yamnaya) de Barros Damgaard *et al.* (2018b), a 13.1× ancient genome from Greenland (Saqqaq) Rasmussen *et al.* (2010), and a 9.6× ancient genome from Ethiopia (Mota) Llorente *et al.* (2015). The Yamnaya genome was only available as a BAM file, while the latter two were

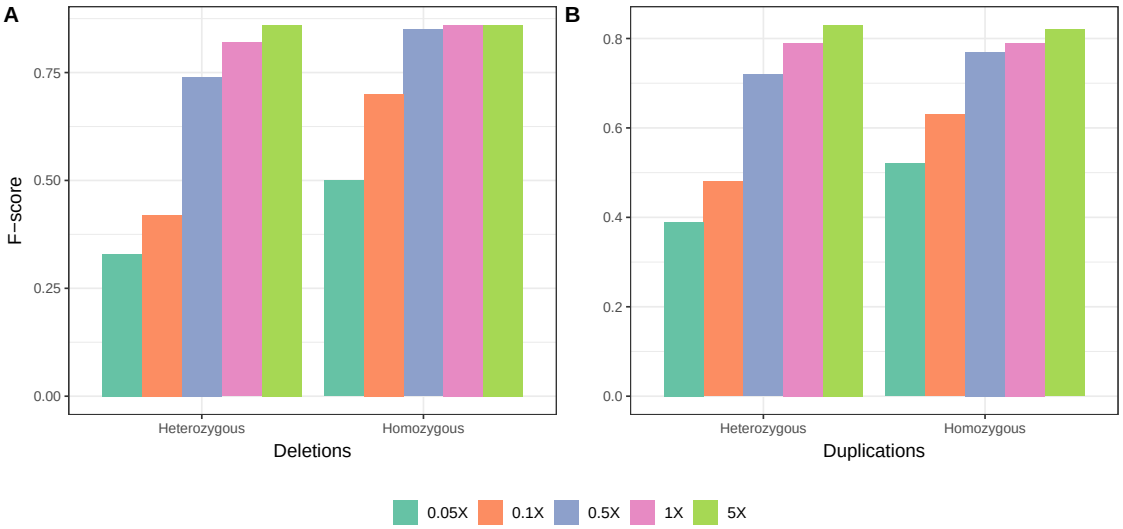

Figure 2: Performance (F-scores) of CONGA in predicting copy-numbers of (A) deletions and (B) duplications using merged sets of medium and large CNVs, at various coverage values.

available as FASTQ files, which we processed into BAM files (Methods). We used a list of modern-day human CNVs as candidate CNV set ($n = 17,392$ deletions and $n = 14,888$ duplications) (Methods) as input to CONGA. We thus genotyped between 688-1,581 deletions and 638-4,097 duplications across these three genomes using the full data. We then down-sampled all three BAM files to various depths, and repeated the genotyping for each genome. We estimated CONGA's TPR and FDR on down-sampled genomes by treating the CNVs genotyped using the full data as ground truth (Methods).

CONGA displayed satisfactory performance in identifying deletions in all three genomes even at coverages around $0.5\times$, with TPR of $>70\%$ and FDR of $<45\%$ (Figure 3, Supplemental Table S1.C). For duplications, however, CONGA showed poor performance: at around $1\times$ coverage, duplication TPR was $>40\%$ in the Saqqaq and Mota genomes, and only 22% in the Yamnaya genome. A detailed analysis of these results suggested that pre-publication quality filtering of BAM files may have obliterated read-depth-based duplication signals in the data (Supplemental Note S2).

Overall, both our simulations and down-sampling experiments with real genomes suggest that CONGA can efficiently genotype >1 kbps deletion events at depths of coverage of $0.5\times$, and even at $0.1\times$. CONGA could thus be applied on a large fraction of ancient shotgun sequenced genomes available for deletion genotyping. In contrast, CONGA's low performance in duplication genotyping in the down-sampled Yamnaya BAM data implies that identifying duplications in published low coverage ancient genomes may not be feasible, as the data are mainly submitted in BAM format in public repositories (see Discussion). We therefore limited downstream analyses on real ancient genomes to deletions >1 kbps.

Analysis of 71 real ancient genomes and technical influences on deletion genotyping

Although CONGA's above performance in deletion genotyping was promising, heterogeneous sets of real ancient genomes may pose additional challenges, as they are obtained from DNA samples of complex taphonomic history and are produced via different experimental protocols. Hence, whether consistent biological signals may still be extracted from low coverage genome sets remains unclear. To explore this, we genotyped deletions with CONGA across a diverse sample of real ancient human genomes. We then studied their diversity with expectation that deletions, like SNPs, should display genome-wide similarity patterns that reflect population origin, i.e., shared genetic drift, among individuals Conrad and Hurles (2007); Levy-Sakin *et al.* (2019); Almarri *et al.* (2020).

We thus collected BAM files for 71 ancient human genomes belonging to a time range between c.2,800-45,000 years Before Present (BP) (Supplemental Table S2) Rasmussen *et al.* (2014); Günther *et al.* (2015); Hofmanová *et al.* (2016); Jones *et al.* (2015); Kılınç *et al.* (2016); de Barros Damgaard *et al.* (2018b); Gamba *et al.* (2014); González-Fortes *et al.* (2017); de Barros Damgaard *et al.* (2018a); Keller *et al.* (2012); Sikora *et al.* (2019); Olalde *et al.* (2014); Lazaridis *et al.* (2014); Antonio *et al.* (2019); Allentoft *et al.* (2015); Haber *et al.* (2019); Fu *et al.* (2014); Broushaki *et al.* (2016); Seguin-Orlando *et al.* (2014); Jones *et al.* (2017); Haber *et al.* (2017); Raghavan *et al.* (2014); Martiniano *et al.* (2017);

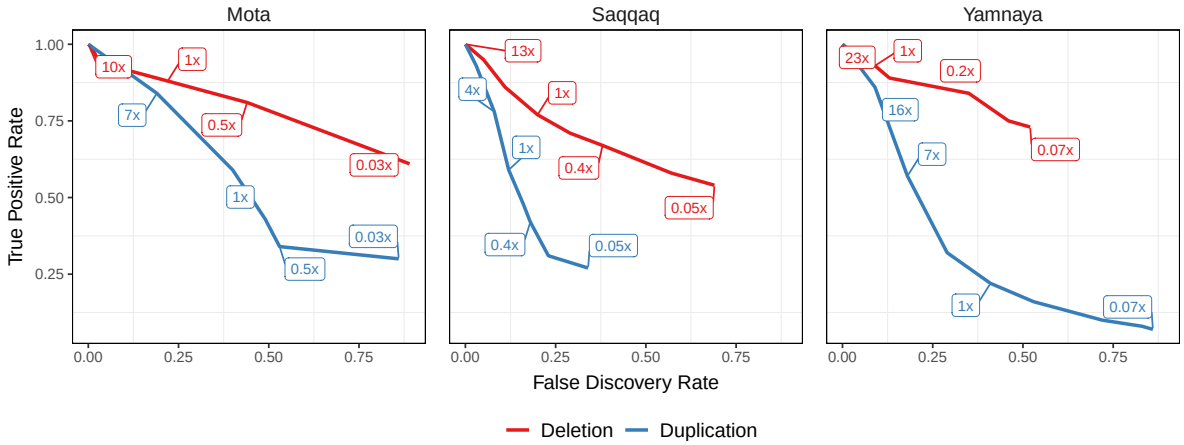

Figure 3: TPR vs FDR curves for deletion and duplication predictions of CONGA using Mota, Saqqaq and Yamnaya genomes down-sampled to various depths from their original coverages of 9.6×, 13.1× and 23.3×, respectively. The numbers inside boxes show the down-sampled coverage values. We calculated TPR and FDR for down-sampled genomes assuming that our CONGA-based predictions with the original genomes (full data) reflect the ground truth. These predictions, in turn, were made using modern-day CNVs as candidate CNV list. The purpose of the experiment was to evaluate accuracy at lower coverage relative to the full data (Methods).

Krzewińska *et al.* (2018); Yaka *et al.* (2021). These were chosen to bear diverse characteristics, including a wide range in mean coverage (0.04×-26×, median = 3.45×), population origin (West and East Eurasia and North America), the laboratory of origin (10 different laboratories), the use of shotgun vs. whole-genome capture protocols, or the use of uracil–DNA–glycosylase (UDG) treatment Rohland *et al.* (2015). For genotyping, we used a candidate CNV dataset of 11,390 autosomal deletions (>1 kbps with mean 10,735 bps) identified among African populations (AFR) from Phase 3 of the 1000 Genomes Project Sudmant *et al.* (2015b) (Methods). Our motivation for using an African sample here was to avoid ascertainment bias Clark *et al.* (2005) in studying deletion frequencies, as all of the 71 ancient individuals were non-African, and thus African populations represent an outgroup to our sample set. We further filtered these for high mappability (mean >0.9) and to be derived in the human lineage using chimpanzee and bonobo genomes to represent the ancestral state, leaving us with 10,002 deletion events (Methods).

Genotyping the 10,002 loci across 71 BAM files, we found 8,780 (88%) genotyped in at least one genome (as deletion or reference). Further, 5,467 (55%) loci genotyped as a deletion (in heterozygous or homozygous state) at least once. Across the 71 genomes, we detected a median number of 490 deletion events [396-2,648] again in either heterozygous or homozygous state.

We studied deletion copy number (frequency) variation across these 71 ancient genomes using a battery of heatmaps, hierarchical clustering, multidimensional scaling plots (MDS) and principal components analysis (PCA) (Supplemental Fig. S3; Supplemental Fig. S4). This revealed a minority of genomes exhibiting highly divergent frequencies, without obvious association with their population of origin. Given the close evolutionary relationship among Eurasian human populations, we reasoned that these divergent signals most likely originate from experimental artifacts, data processing artifacts, or variability of DNA preservation among samples. Supporting this, mean deletion frequencies across the 71 genomes could be explained by laboratory-of-origin (Kruskall-Wallis test , $p = 0.08$). We identified a subset of 21 divergent, or outlier genomes, and removing these also removed the laboratory-of-origin effect (Kruskall-Wallis test, $p = 0.22$; Supplemental Note S3). We could further recognize a number of attributes that could explain these divergent deletion profiles. First, the 21 divergent genomes had on average shorter read length compared to the rest (median = 57 vs. 69; Wilcoxon rank sum test $p < 0.001$; Supplemental Fig. S5A). One of these was the Iceman, with unusually short (50 bps) reads. Second, the coverage of the 21 divergent genomes was lower compared to the remaining 50 (median = 3.31 vs. 3.98; Wilcoxon rank sum test, $p = 0.014$; Supplemental Fig. S5B). For instance, all three genomes with <0.1× coverage in our dataset (ne4, ko2, and DA379) were among the outliers. The number of non-genotyped loci was likewise higher in the divergent group (median = 1509 vs. 1886; Wilcoxon rank sum test, $p = 5.39 \times 10^{-5}$; Supplemental Fig. S5C). UDG-treatment did not appear to be related to outlier behaviour (binomial test $p = 2.633 \times 10^{-9}$; Supplemental Fig. S5D). Meanwhile, Bon002, the only sample produced using whole-genome capture, was among the most extreme outliers, suggesting that the capture procedure distorts coverage. Consequently we removed these 21 genomes from further analyses.

A comparison of deletion and SNP diversity across 50 ancient genomes

The above filtering steps resulted in a dataset of 8,780 derived deletions genotyped in at least one of the 50 ancient Eurasian genomes, with 396-748 deletions (median = 467.5) detected in heterozygous or homozygous state per genome, and 29% detected in at least one genome.

We used this dataset to test three hypotheses: (i) that CONGA-called deletion diversity patterns should parallel SNP diversity patterns, reflecting shared demographic history (genetic drift and admixture) among genomes, (ii) that CONGA-called deletions should be evolving under some degree of negative selection (caused by gene expression alterations, exon loss, or frame-shifts), and (iii) that variation in deletion load among genomes may be correlated with variation in deleterious SNP load. We note that the first two patterns (hypotheses i and ii) have been previously described using large modern-day CNV datasets (see Conrad and Hurles (2007); Levy-Sakin *et al.* (2019); Almarri *et al.* (2020) for drift, and Conrad *et al.* (2010); Cooper *et al.* (2011); Sudmant *et al.* (2015a) for selection), and our goal here was mainly to perform a sanity check and assess CONGA's effectiveness in producing reliable biological signals.

To test the first hypothesis, we compared pairwise genetic distances among the 50 individuals (Figure 4A) calculated using either SNPs or deletion genotypes. For this, we collected 38,945,054 autosomal SNPs ascertained in African individuals in the 1000 Genomes Dataset and genotyped our 50 ancient genomes at these loci (Methods). We then calculated pairwise outgroup-f_3 statistics, a measure of shared genetic drift between a pair of genomes relative to an outgroup population Patterson *et al.* (2012). Using the Yoruba as outgroup, we calculated genetic distances for all pairs of ancient genomes as (1 - f_3), using either SNPs or deletions. We observed strong positive correlation between the two resulting distance matrices (Spearman $r = 0.671$, Mantel test $p = 0.001$) (Figure 4B). Summarizing SNP- and deletion-based distances using multidimensional scaling (MDS) also revealed highly similar patterns, with clear clustering among west and east Eurasian genomes observed with either type of variation (Figure 4C, D). This result was encouraging in showing that diversity patterns based on deletion genotyping with CONGA in a heterogeneous sample of low coverage ancient genomes reveals expected signals of shared demographic history.

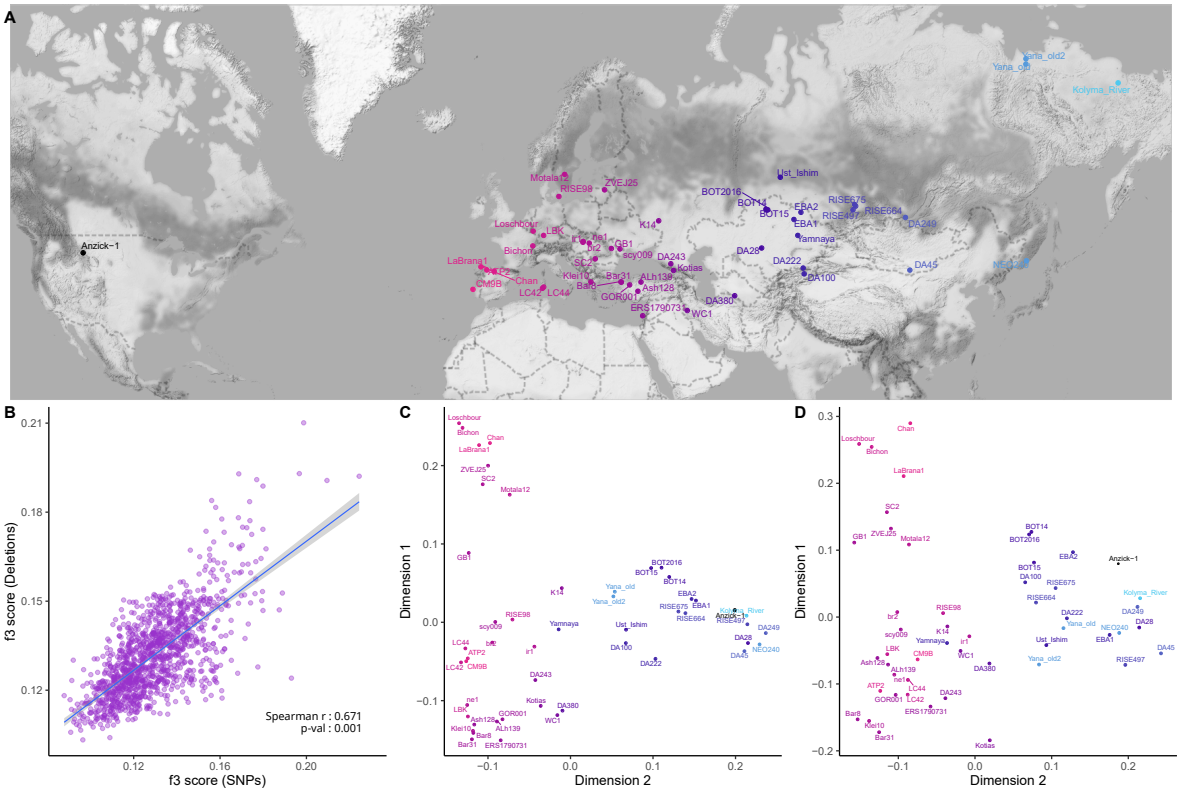

Figure 4: (A) Geographic locations of the 50 ancient individuals. (B) Comparison of genetic distances calculated using SNPs and deletions. We calculated the Spearman correlation coefficient between two matrices and then calculated Mantel test p-value using the "mantel" function in R package "vegan" (v2.5-7). (C) and (D) represent multidimensional scaling plots that summarize outgroup-f_3 statistics calculated across all pairs among the 56 ancient individuals using SNPs and deletions, respectively.

Negative selection on deletion variants

We next studied the impact of negative (purifying) selection on deletions by comparing the site-frequency-spectrum (SFS) of autosomal deletions with those of SNPs. We used the 8,780 human-derived deletions and 32,304,437 human-derived SNP alleles across the 50 ancient genomes (Methods). To allow comparison with the pseudo-haploidized SNP genotype data, we randomly chose one allele per genome (i.e., deletion or no event) in the deletion dataset. Set side by side with the SNP SFS, we observed an excess of singletons and a lack of fixed derived variants among deletions, consistent with stronger negative selection on the latter (Figure 5A). The excess of undetected and singleton deletions does not appear to be related to low recall, as both high and low coverage genomes show the same trend (Supplemental Fig. S6).

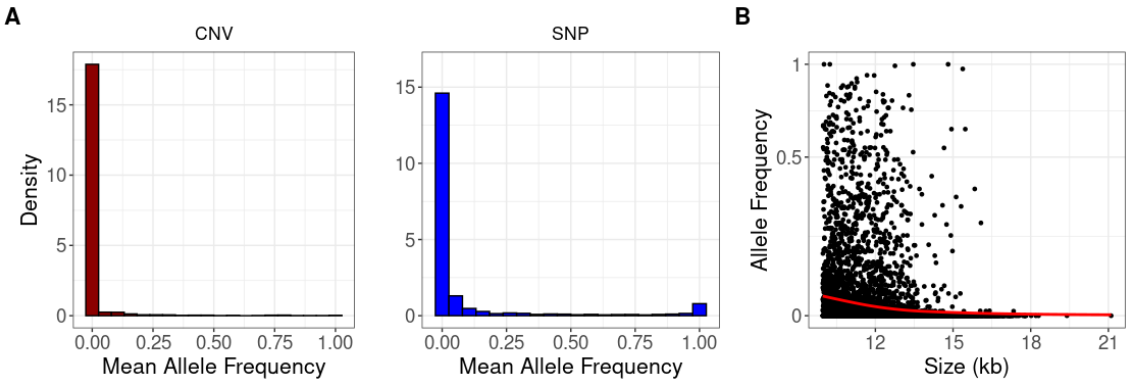

Figure 5: (A) The site-frequency-spectra of derived deletion alleles (on the left, n=3,472) and derived SNP alleles (on the right, n=57,307). The x-axes show mean allele frequency for each locus calculated using only those genomes where a locus has been observed (e.g. an allele observed in 10 out of 40 genomes will be represented as 25%). The two distributions are significantly different from each other (Kolmogorov-Smirnov test $p < 10^{-15}$). (B) The size distribution (in kbps) of the deletions versus mean allele frequency. The red line shows the fitting of smoothing spline and indicates a negative correlation (Spearman correlation $r = -0.33$, $p < 10^{-16}$). Both axes were log_2-scaled.

If deletions are under negative selection we may also expect longer deletions, or deletions containing evolutionary conserved genes, to be segregating at lower frequencies. Indeed, we found that deletion allele frequencies were negatively correlated with deletion size across the 50 genomes (Spearman correlation $r = -0.33$, $p < 10^{-16}$) (Figure 5B). To test the second idea, we determined deletions overlapping Ensembl human genes. Overall, 26% of the 8,780 derived deletions overlapped minimum one gene. We then collected mouse-human dN/dS ratios, an inverse measure of protein sequence conservation (Methods). We found that deletions with lower (below-mean) allele frequency had slightly lower dN/dS values compared to deletions with higher (above-mean) allele frequency (median = 0.086 vs. 0.097; Mann-Whitney U test, one-sided $p = 0.055$). These observations, along with the SFS comparison, follow the notion that deletions are evolving under negative selection.

We further asked whether inter-individual variation in the total deletion mutation burden that we measure in our dataset may be correlated with variation in the burden of functional deleterious SNPs based on their impact on protein sequence. Demographic bottlenecks can theoretically cause variable levels of mutation burden –as deletions and/or as SNPs– among ancient genomes, and these burden levels could be correlated especially if their phenotypic impacts are comparable (see Discussion). To test this we collected (a) total deletion length and (b) the number of genes affected by deletions, for each of the 50 ancient genomes (Supplemental Fig. S7). We further collected SIFT scores (an estimate of how protein sequence would be affected by a SNP Ng and Henikoff (2003)) for n=22,996 SNPs in our dataset, predicted to be "deleterious" or "tolerated", and used these to calculate a deleterious/tolerated ratio per genome (Methods; Supplemental Fig. S8A). We then compared the deleterious/tolerated ratio-based burden levels with deletion-based mutation burden levels (total length and number of genes), but found no significant correlation (Spearman $r = 0.09$ and $r = -0.05$, respectively, $p > 0.5$; Supplemental Fig. S8B). This could be explained by high noise and lack of statistical power, as well as differences in phenotypic impacts between deletions and SNPs (see Discussion). We also did not observe any correlation between historical age and deletion frequencies in this sample (Supplemental Fig. S9).

Time and memory consumption

Finally, we examined time and memory requirements of CONGA. We first tested our performance of deletions with BAM files of the 71 ancient genomes presented above. This finished in $\sim$12 hours in total with as low as 2.2 GB of peak-memory consumption. This is $\sim$10 minutes per genome. In order to evaluate CONGA's performance with a higher coverage genome sample, we ran 30 genomes (randomly selected 10 CEU, 10 YRI, 10 TSI) from the 1000 Genomes Project Phase 3, which had mean 7.4$\times$ coverage Sudmant *et al.* (2015b). The analysis took just slightly longer, $\sim$15 minutes average per genome, with similar memory usage.

We also compared the time and memory requirements of CONGA, GenomeSTRiP, FREEC and CNVnator in Table 2. In order to benchmark these tools, we used a 5$\times$ simulated genome (the same genome with medium sized CNVs used in the simulation experiments described above) with the same computing resources[1]. CONGA has the lowest runtime and memory footprint among the other tools.

In Supplemental Table S1.D we report the effects of parameter choices when using with CONGA on runtime and memory usage. We note that using split-reads for duplication genotyping (intended for higher coverage genomes) increases runtime and memory consumption significantly because here CONGA uses its own small-scale read mapper, which creates a bottleneck.

We further provide a comparison of CONGA's performance on genomes of various depths of coverage in Supplemental Table S1.D, calculated using the down-sampled 23$\times$ Yamnaya genome (with coverages between 23$\times$ and 0.07$\times$).

Table 2: Time and Memory Consumption

Tools	Time (h:m)	Peak Memory Usage (GB)
CONGA	0:09	1.2
GenomeSTRiP	1:22	2.2
FREEC	0:39	7.1
CNVnator	0:32	14.1

Time and memory consumption of each algorithm for a simulated genome of 5$\times$ depth of coverage with 1680 deletions and 1684 duplications. "Time" refers to wall clock time and "Peak memory usage" is the maximum resident set size. Note that GenomeSTRiP has two steps in its pipeline: preprocessing and genotyping. Here, time was calculated by summing the running times of each step, and memory by taking the maximum. For CONGA, we used default parameters used in the simulation experiments.

Discussion

Modern human genome sequencing experiments today typically reach coverages $>$20$\times$ and increasingly use long read technology, and such experiments can employ diverse read signatures to reliably identify CNVs Alkan *et al.* (2011). CONGA's approach that mainly relies on the read-depth signature is naive in comparison; however, using read-depth appears as the main practical solution given the short fragment size and the predominance of low coverage (around or $<$1$\times$) among ancient genome datasets.

CONGA's overall performance and utility

Despite these challenges, our experiments using simulated genomes and down-sampled real ancient FASTQ data showed that CONGA can relatively efficiently genotype deletions and duplications of size $>$1 kbps at 1$\times$ coverage, or even lower. CONGA outperformed two "modern DNA" CNV discovery algorithms, FREEC and CNVnator, two methods previously employed in ancient genome analyses Smith *et al.* (2017); Bhattacharya *et al.* (2018). CONGA exceeded both tools in TPR and true negative rates, especially at coverages $<$1$\times$. This is unsurprising, as these tools were developed for discovering novel CNVs in relatively high coverage genome data. Meanwhile, compared to GenomeSTRiP, a CNV genotyper that also uses both different sources of information within a Bayesian framework Handsaker *et al.* (2011, 2015), CONGA performed better in achieving lower FDR rates at all coverages, while GenomeSTRiP had higher recall at coverages 0.5$\times$ or below. In time and memory use, CONGA surpassed all three tools.

In terms of deletion copy number estimates, CONGA again achieved acceptable accuracy ($\sim$75% TPR and $<$30% FDR) in genomes of 0.5$\times$ coverage. At lower depths of coverage and also when genotyping deletions $<$1 kbps, recall and/or precision were weaker. CONGA's performance on duplications was also poor, as we discuss below.

[1]Intel(R) Xeon(R) CPU E5-2640 v2 @ 2.00GHz: 2CPUs * 8 cores each=16 cores total and 216GB RAM

Overall, the relatively high accuracy at $\geq 0.5\times$ coverage suggests that CONGA could be used to genotype deletions across a considerable fraction of published shotgun sequenced ancient genomes. CONGA and GenomeSTRiP could also be used in parallel, as they appear to complement each other in recall and specificity. Further, GenomeSTRiP can be used in population samples for jointly genotyping low coverage genomes, which could potentially increase performance. We caution, however, that joint genotyping can create ascertainment biases if coverage and ancestry co-vary among jointly analysed genomes.

Beyond aDNA, CONGA is suitable for CNV analyses for any low depth whole-genome sequencing (WGS) experiment. Such studies are increasing in number due to the trade-off between budget limitations and the wealth of genome-wide information that can be used in population and conservation genetics (e.g. Vieira *et al.* (2016)).

Caveats in duplication genotyping

In simulated genome experiments, CONGA's performance in genotyping duplications was similar to that in deletions. Beyond read-depth information (also used in genotyping deletions), duplication genotyping could also effectively benefit from paired-end information from split reads. Using paired-end information alone yielded >0.65 F-Score for duplication genotyping, though only at $5\times$ coverage and with variants >10 kbps (Supplemental Table S1.E; Supplemental Fig. S10). In down-sampling experiments, CONGA showed slightly lower performance in duplication genotyping than in deletion genotyping when using two ancient genomes available as FASTQ files. However, CONGA's performance was dramatically low on the $23\times$ ancient BAM file, Yamnaya. This can be explained as follows (see Supplemental Note S2): (i) The available Yamnaya data was processed in such a way that excess reads at duplicated loci, i.e. read-depth information, was lost. (ii) Consequently, nearly all (97%) duplications CONGA genotyped in the original ($23\times$) BAM file were called only using paired-end information. (iii) Because paired-end information is rapidly lost with decreasing coverage (as it requires reads overlapping breakpoints), and read-depth information was lacking, genotyping duplications in this BAM files became infeasible at $<5\times$ coverage.

The majority of shotgun ancient genomes in public databases are only published as BAM files. The majority of published files are also at $<5\times$ coverage. Hence, most published ancient shotgun genomes are not amenable to duplication genotyping with CONGA. This is highly unfortunate, as gene duplications are a major source of evolutionary adaptation that would be valuable to study also in ancient populations.

Caveats in deletion genotyping

Applying CONGA to genotype deletions on a heterogeneous set of real ancient shotgun genomes revealed conspicuous technical influences on deletion genotyping, with a significan fraction of the 71 analysed genomes displaying outlier behaviour in their deletion frequencies. We could notice technical particularities for the 21 genomes identified as outliers, such as lower coverage, shorter read lengths, or the application of whole-genome hybridization capture. Our results suggest that $0.4\times$ coverage may be close to the lower threshold for deletion genotyping of >1 kbps events, slightly higher than the threshold in our simulation results. We also find that whole-genome hybridization capture and extra short reads (roughly ≤ 55 bp) compromise deletion genotyping, while UDG-treatment does not show a significant effect. That said, we lack clear explanations for outlier deletion frequency patterns for some of these 21 genomes. For instance, the genome SI-45 has coverage $>3\times$ and an average read length of 60 bps, but nevertheless displays unusual deletion patterns. We suspect that such unexpected patterns might reflect technical peculiarities in library preparation, sequencing or data filtering. Unique taphonomic processes influencing DNA preservation and variability in coverage may also be at play.

Such effects could be investigated by future studies compiling larger datasets with detailed experimental descriptions. Meanwhile, our results point to the necessity of rigorous quality control and outlier filtering when calling deletions in heterogeneous datasets, similar to practices traditionally adopted in transcriptome analyses. This is particularly essential when combining genomes produced using different experimental protocols and sequencing platforms.

Community recommendations for improving CNV analyses in ancient genomes

The above observations mark the urgent need for new practices in producing and publishing ancient genomes to allow reliable study of both deletions and duplications, beyond SNPs.

- Most published ancient genome data to date is SNP capture data, which is largely worthless for CNV analyses. Our results underscore the long-term value of shotgun sequencing data over SNP capture, as well as whole-genome capture.
- Publishing data as raw FASTQ files should be priority. The main motivation behind publishing BAM files instead of raw data is to avoid publishing environmental DNA reads, which constitute a large fraction of

reads from shotgun sequenced aDNA experiments. Saving microbial (e.g. pathogenic) aDNA fractions for investigations is another motivation. Nevertheless, our results show that raw FASTQ data is absolutely necessary for duplication genotyping at low coverage and also helpful against biases in deletion genotyping. In the long term, publishing raw data will be for the whole community's benefit.

- Sharing all details on DNA extraction, library construction, as well as the alignment and preprocessing steps used in creating the exact version of datasets submitted to public databases is crucial for healthy reuse of the data.

Purifying selection and mutation loads in past populations

Our analysis of >1 kbps deletions genotyped in 50 ancient genomes revealed how variation in deletion frequencies reflect (a) demographic history, as reflected in strong correlation with SNP variation and spatial clustering, and (b) negative selection, as reflected in a steeper SFS than of SNPs, lower frequencies of large deletions, and lower frequencies of deletions overlapping conserved genes. These results show that CONGA can identify reliable biological signals in technically heterogeneous and noisy datasets, which is a non-trivial outcome.

Beyond expected patterns, we also studied possible correlation between deletion loads and deleterious SNP loads per genome across the 50 ancient individuals. High deleterious mutation loads could arise by relaxation of negative selection due to strong bottlenecks, as suggested for Wrangel Island mammoths Rogers and Slatkin (2017) or for dogs Marsden *et al.* (2016). Conversely, bottlenecks can cause high inbreeding levels, and this may lead the purging of recessive deleterious variants, as recently described for a founder population of killer whales Foote *et al.* (2021). In our dataset we found no significant relationship between deletion-related loads and deleterious SNP loads. This could be due to lack of strong variability among Eurasian genomes in deleterious mutation burdens or due to low statistical power, as we only use deletions segregating in Africa. The result could also reflect differences in dominance effects or fitness effects between SNPs and deletions.

A full analysis of this question could be possible with the creation of a geographically comprehensive genomic time-series, especially genomes of non-Eurasian populations with variable demographic histories. It would further require CNV discovery in carefully processed high-coverage ancient genomes and subsequent genotyping on low coverage data using CONGA. We hope that our study opens the way for such work, bringing deeper insight into the impacts of selection and drift in humans and other species.

Methods

Among various approaches developed for CNV discovery using high throughput sequencing data, almost all use the fact that read-depth, i.e., the density of reads mapped to the reference genome, will be on average lower in deleted regions and higher in duplicated regions Alkan *et al.* (2011); Ho *et al.* (2020). The distance between paired-end reads, their orientation, and split-read information (start and end of reads mapping to different locations) are further sources of information used in determining CNVs. Although available CNV discovery algorithms generally perform well in modern-day human genome sequencing data with high coverage, this is not necessarily the case for ancient genomes, as well as other low coverage sequencing experiments (Supplemental Fig. S11, S12). The first reason is that the majority of shotgun ancient genomes are produced at low coverage (typically $<1\times$), which constrains the use of read-depth information. Second, ancient DNA fragments are short and of variable size (typically between 50-100 bps) Shapiro and Hofreiter (2014). Thus, paired-end information is absent, and available split-read information is also limited. Variability in ancient DNA preservation and genome coverage Pedersen *et al.* (2014) is yet another noise source that is expected to limit efficient CNV discovery. CONGA overcomes these limitations using genotyping instead of *de novo* discovery. It estimates whether a candidate CNV, the location of which is provided as input, is present in a genome in BAM format. It also estimates the genotype, i.e., the heterozygous or homozygous state. CONGA makes use of read-depth information for deletions, and both read-depth and split-read information for duplications.

Likelihood-based read-depth calculation for deletion and duplication genotyping

The input to the algorithm is (1) a list of candidate CNV locations and CNV type, i.e., deletion or duplication, and (2) a data set of reads aligned to the linear reference genome, e.g., using BWA Li and Durbin (2009), which should be in BAM format.

In order to calculate the likelihood of a CNV at a given locus based on read-depth information, CONGA uses an approach akin to Soylev *et al.* (2019). Let (S_i) be the i^{th} CNV in our CNV input list, defined by the breakpoint interval (B_l, B_r) and the type of CNV: a deletion or duplication. At this locus, CONGA calculates the likelihood of the three possible genotype states, k, given the read alignment data and CNV type. The genotype states are: no event ($k = 0$), a

heterozygous state ($k = 1$), or a homozygous state ($k = 2$). The likelihood, in turn, is calculated by comparing the observed (O_i) read-depth versus the expected (E_{ik}) read-depth within (B_l, B_r), given the three different genotypes. We detail the steps below.

1. We count the total number of mapped reads within that locus (falling fully within the interval (B_l, B_r)). This is the observed read-depth, (O_{RD}).
2. We calculate expected read-depth under a "no event" scenario, i.e., representing the diploid state. Here we account for the GC bias in high-throughput sequencing data Smith *et al.* (2008), by using LOESS smoothing to normalize read-depth for GC content. Specifically, for each chromosome, we calculate the read-depth values per GC percentile for sliding windows of size 1,000 bps (step size = 1 bp). We then calculate the average read-depth per GC percentile. Then, using the chromosome-wide average GC value for the interval (B_l, B_r), we calculate the expected diploid read-depth, $E_{i_{k=0}}$.
3. We model the read-depth distribution as Poisson, using the expected read-depth values for $k = 0$, $k = 1$, $k = 2$. We calculate the probability $P(RD_{S_i}|state = k)$ as:

$$P(RD_{S_i}|state = k) = \frac{E_{ik}^{O_i} \times e^{-E_{ik}}}{O_i!},$$

   where E_{i_k} is the expected read-depth given $state = k$, and O_i is the observed read-depth at that specific locus. A typical autosomal human locus is diploid (has copy number = 2); therefore when there is no CNV event $(k = 0)$, the expected value of O_i should be $E_{i_{k=0}}$.
   If a genome is homozygous for a deletion, we expect no reads mapping to the region, thus $O_i \sim E_{i_{k=2}} = 0$. For heterozygous deletions, the expected number of mapped reads in that interval will be half of the expected diploid read-depth: $O_i \sim E_{i_{k=1}} = E_{i_{k=0}}/2$. For homozygous duplications, we expect $O_i \sim E_{i_{k=2}} = E_{i_{k=0}} \times 2$. For heterozygous duplications, we expect $O_i \sim E_{i_{k=1}} = E_{i_{k=0}} \times 1.5$.
4. We calculate a likelihood-based score, which we term the C-score, to estimate how likely locus S_i carries a non-reference variant in a genome, in either one copy or two copies. For this we use the calculated likelihoods for the three states. We define the C-score as the maximum of the likelihoods of (S_i) being present in heterozygous state $(k = 1)$ or in homozygous state $(k = 2)$ in that genome, over the likelihood of no event $(k = 0)$. We use the log function to avoid numerical errors.
   $C - score(S_i) =$

$$\frac{max(log(P(RD_{S_i}|k{=}1)), log(P(RD_{S_i}|k{=}2)))}{log(P(RD_{S_i}|k{=}0))},$$

   The C-score is distributed between 0 and +∞, with lower scores indicating higher likelihood of a true CNV event.

Results from our simulations and down-sampling experiments suggest that the relatively simple Poisson distribution can be effectively used to model copy number states, especially in the face of potentially non-independent errors due to ambiguous mapping of short and damaged reads or GC content heterogeneity. We note that alternative models have also been used for analysing CNVs in short read sequencing data, such as the negative binomial distribution Miller *et al.* (2011) or Gaussian mixed models Handsaker *et al.* (2011). We also note CONGA's our approach could be expanded in the future by including the evaluation of duplication events involving >2 copies, as in multicopy genes Sudmant *et al.* (2010).

Split-read and paired-end signatures for duplication genotyping

Beyond read-depth, information of paired-end reads or read fragments that do not linearly map to the genome can be used to identify CNVs. Ancient genomes are sometimes single-end and sometimes paired-end sequenced, but in the latter case, short overlapping reads are typically merged into a single read before alignment. Ancient genome data is thus practically single-read. However, the split-read method can be applied on single-read ancient genome data, which emulates paired-end information for genotyping duplications. This approach is visualized in Figure 6. We therefore designed CONGA to include both paired-end and single-end reads as input and to evaluate paired-end signature information.

First, assume a read of length L mapped to position pos_x in the reference genome, where pos_x is assumed to be one of the breakpoints of a putative CNV. There always exists a subsequence $\geq L/2$ that will have at least one mapping in the reference genome with some error threshold. Thus, we can split a read into two subsequences, assigning the

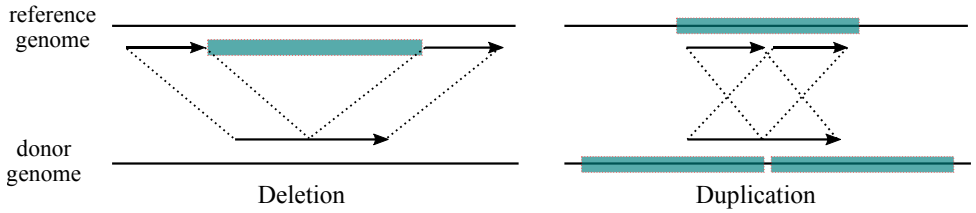

Figure 6: The figure shows our split-read approach to emulate paired-end using single-end reads. We use short-read Illumina mappings in a BAM file as input. We split each discordant read (whose mapping quality is larger than the given threshold and does not overlap with a known satellite) from the middle, keeping the initial mapping as one element and the other subsequence (split segment) as the second element of a pair. We remap the split segment to the reference genome, and evaluate the position and the orientation of both reads to identify the presence of putative CNVs.

actual mapping to one of the pairs and remapping the other subsequence ("split segment") as a second pair. There are two possible split strategies: an even decomposition, where both subsequences are of equal lengths, or an uneven decomposition, where the subsequences are of unequal lengths. Given the infeasibility of testing each split position and the fact that ancient reads are typically already short, we follow Karakoc *et al.* (2012) and split the read from the middle to obtain two reads with equal lengths $L/2$. If a read overlaps a duplication breakpoint, and assuming that the expected position of the breakpoint will be uniformly distributed within the read, the split segment will map to the reference genome with insert size—the distance between the split-read pairs—greater than zero.

With this simple observation, the need to observe all possible breakpoints can be eliminated. Thus, given a single-end read Rse_i, we define $Rpe_i = (l(Rpe_i[pos_x : pos_x + RL/2])$ and $r(Rpe_i[pos_y : pos_y + RL/2]))$, where pos_x is the initial mapping position of the single-end read, pos_y is the remapping position of the split read, RL is the length of the single-end read observed before the split, $l(Rpe_i[pos_x : pos_x + RL/2])$ is the left pair within pos_x and $pos_x + RL/2$ and $r(Rpe_i[pos_y : pos_y + RL/2])$ is the right pair within pos_y and $pos_y + RL/2$ of the paired-end reads. We use this information as described in the following section.

Remapping paired-reads and utilizing paired-read information

According to our remapping strategy, we use a seed-and-extend approach similar to that implemented in mrFAST Alkan *et al.* (2009), where a read is allowed to be mapped to multiple positions. Our main concern here is that the split segment, due to its short length, can be mapped to unrealistically high numbers of positions across the genome. To overcome this problem we use the approach developed in TARDIS Soylev *et al.* (2017), allowing the split segment to be mapped only up to 10 positions within close proximity (15 kbps by default) of the original mapping position and applying a Hamming distance threshold for mismatches (5% of the read length by default).

Based on the distance between the reads (insert-size) and orientation, we then evaluate the type of putative CNV. As Figure 6C shows, if the split segment maps behind the initially mapped segment of the same pair to generate a reverse-forward mapping orientation, this would be an indication of a duplication.

In order to utilize this paired-read information, for each CNV locus used as input to our algorithm, we count the number of read-pair (i.e. split segments) that map around +/- 5 kbps of the breakpoints. Each such read-pair is treated as one observation. We use these counts in combination with the C-score (read-depth information) to genotype duplications (see below). We do not use this read-pair information for genotyping deletions due to its low effectiveness in our initial trials (Supplemental Table S1.E).

Mappability filtering

The probability of unique alignment of a read of certain size varies across the genome, mainly due to repetitive sequences. Various algorithms estimate this probability, termed mappability, across the genome for k-mers of specific length Koehler *et al.* (2011); Derrien *et al.* (2012); Karimzadeh *et al.* (2018); Pockrandt *et al.* (2020). This is calculated by extracting k-mers of given length through the genome, remapping them to the reference genome, and measuring mappability as the proportion of unique mappings Karimzadeh *et al.* (2018). Because low mappability regions can be confounded with real deletions, we use mappability information to filter out CNV loci that could represent false positives.

CONGA accepts any mappability file in BED format, where values are distributed between 0 and 1. These can then be used to filter out CNVs for minimum mappability.

In our experiments, we used the 100-mer mappability data from the ENCODE Project ENCODE Project Consortium (2012). Using this data, for each CNV event (S_i), we calculated the average mappability value within its breakpoints. We used a minimum average mappability threshold of 0.9 for the CNV events we analyzed.

Our deletion frequency analysis results suggest that the strict filter should be used especially when analyzing data sets of heterogeneous origin. This is because published BAM files frequently differ in mapping quality filters applied before publishing (and these filters are usually not indicated), and such filtered BAM files will produce artificial deletion signals at low mappability regions, while unfiltered BAM files will not.

Simulation and down-sampling experiments

Simulating ancient genomes with implanted deletions and duplications

Our goal here was to study the performance of CONGA on different sized deletions or duplications using simulated genomes containing implanted CNVs and to determine thresholds for reliably calling these variants. We first employed VarSim Mu *et al.* (2015) to simulate and insert deletions and duplications into the human reference genome GRCh37. We repeated this three times, for small (100 bps - 1000 bps), medium (1000 bps - 10,000 bps), and large (10,000 bps - 100,000 bps) CNVs. As a result we generated three CNV-implanted genomes, with around 1500 deletions and 1500 duplications each (between 1385 and 1810). The CNVs were produced so that they were non-overlapping, and their length distribution and exact counts are provided in Supplemental Fig. S1.

To evaluate specificity and sensitivity, we also included a background (false) CNV set in the experiment, which would not be implanted but would be queried as part of the candidate list. This background set was prepared using recently published deletion and duplication calls from human genome sequencing experiments Audano *et al.* (2019); Chaisson *et al.* (2019); Zook *et al.* (2020); Collins *et al.* (2020) and also sequencing data from African populations (AFR) from Phase 3 of the 1000 Genomes Project Sudmant *et al.* (2015b). We compiled a list of 17,392 deletions and 14,888 duplications that were non-overlapping and of size $>\sim$1000 bps using BEDTools mergeBed Quinlan and Hall (2010). When evaluating genomes with small CNVs (100 bps - 1,000 bps), we additionally included small CNVs from Chaisson *et al.* (2019). Specifically we added 4,623 deletions and 3,750 duplications of size 100 bps - 1,000 bps to the above background list.

In order to assess CONGA's performance, we added the true CNVs generated using VarSim to this background set (and removed overlapping CNVs from the candidate genotype set), such that only $\sim$10% of the input candidate CNV list were true events. Finally, we determined how many of these true events could be correctly called by CONGA and other software.

Simulating ancient genome read data

We used the above-described simulated genomes as input to Gargammel Renaud *et al.* (2017), which generates ancient-like Illumina reads, i.e., short reads of variable size bearing postmortem damage (i.e., C-to-T transitions at read ends) and including adapters. Gargammel can generate aDNA fragments following a size distribution given as input, and we used a subset of Fu *et al.* (2014), which is default for this software. We used Gargammel to produce reads at various depths of coverage: 0.05×, 0.1×, 0.5×, 1× and 5×. We then removed adapters and merged overlapping reads Schubert *et al.* (2016) to generate single-end Illumina reads. These reads had sizes ranging between 34 bps and 139 bps, with average 69 bps and median 66 bps (these statistics were calculated using 1× coverage data, but other data also had similar distributions). We mapped the Gargammel-output reads back to the human reference genome (hg19, or GRCh37) using BWA-aln Li and Durbin (2009) with parameters "-l 16500 -n 0.01 -o 2" (Supplemental Material). Note that BWA-aln has been shown to be more accurate for short ancient reads than BWA-mem Oliva *et al.* (2021).

Evaluation of CONGA, GenomeSTRiP, CNVnator and FREEC with simulated ancient genome data

We ran CNVnator Abyzov *et al.* (2011), FREEC Boeva *et al.* (2012) and GenomeSTRiP Handsaker *et al.* (2011) on the simulated genomes with parameters described in the Supplementary Information and CONGA with two values for the C-score ($<$0.3 and $<$0.5). We used the above-described list of CNVs as the input candidate set for CONGA and GenomeSTRiP.

To determine true calls, we used $>$50% reciprocal overlap for the two CNV events (the event in the input event set and the called event) to be considered the same. This calculation was done using BEDTools Quinlan and Hall (2010). The number of true CNVs were: 1810 deletions and 1751 duplications for 100 bps - 1000 bps; 1680 deletions and 1684 duplications for 1000 bps - 10,000 bps; and 1385 deletions and 1532 duplications for 10,000 bps - 100,000 bps.

Down-sampling experiment with real ancient genomes

We used three relatively high coverage ($\sim$23.3$\times$, $\sim$13.1$\times$ and $\sim$9.6$\times$ respectively) genomes of a Yamnaya culture-related individual from early Bronze Age Karagash (hereafter Yamnaya), Kazakhstan de Barros Damgaard *et al.* (2018b), a Saqqaq culture-related individual from Bronze Age Greenland (hereafter Saqqaq) Rasmussen *et al.* (2010), and a 4500-year old East African hunter-gatherer individual from Mota Cave in Ethiopia (hereafter Mota) Llorente *et al.* (2015). Using this data, and the above-described 17,392 deletions and 14,888 duplications of size $>$1 kbps (see above) as input, we genotyped 2639 deletions and 1972 duplications in Yamnaya (deletion sizes: 1 kbps to 4 Mbps, median = 4 kbps, mean = 23 kbps; duplication sizes: 1 kbps to 28 Mbps, median = 14 kbps, mean = 80 kbps); 1581 deletions and 4097 duplications in Saqqaq (deletion sizes: 1 kbps to 5 Mbps, median = 5 kbps, mean = 17 kbps; duplication sizes: 1 kbps to 28 Mbps, median = 16 kbps, mean = 70 kbps); and 688 deletions and 638 duplications in Mota (deletion sizes: 1 kbps to 130 kbps, median = 4 kbps, mean = 7 kbps; duplication sizes: 1 kbps to 28 Mbps, median = 6 kbps, mean = 82 kbps).

We then randomly down-sampled the BAM files to various depths using Picard Tools Pic (2019): between 16-0.07$\times$ for Yamnaya; 9-0.05$\times$ for Saqqaq; 7-0.03$\times$ for Mota. We note that this down-sampling procedure does not produce the exact targeted depths, which is the reason why we obtain variable coverages in Fig. 3.

For calling deletions we used C-score$<$0.5. For calling duplications, we called events that fulfilled either of the following conditions (a) C-score$<$0.5, or (b) C-score$<$10 and read-pair support $>$10. Finally, treating the results of the original data as the correct call-set, we calculated TPR (true positive rate) and FDR (false discovery rate) for the down-sampled genomes. We considered CNVs with $\geq 50\%$ reciprocal overlap as representing the same event, calculated using BEDTools Quinlan and Hall (2010).

C-score and read-pair cutoffs and minimum CNV size

We ran CONGA with a range of parameter values for the C-score [0.1-5] and for minimum read-pair support (from 0 support to $>$30), and using the above-described true event sets as the input candidate set involving medium and large CNVs (1680 deletions and 1684 duplications for 1000 bps - 10,000 bps, and 1385 deletions and 1532 duplications for 10,000 bps - 100,000 bps).

We used simulation results (Supplemental Table S1.E) to choose an effective cutoff for calling CNVs. For both deletions and duplications, we decided to use C-score $<$0.5, which appears to yield a good trade-off between recall and precision. Specifically, in simulations, this cutoff ensured an F-score of $>$0.5 at 0.1$\times$ for $>$1 kbps deletions, and superior F-scores at higher coverages (Supplemental Fig. S13).

In addition, we observed that read-pair support $>$10 could be useful for identifying duplications in the absence of read-depth support, but only when coverages were $\geq$1$\times$ (Supplemental Table S1.E; Supplemental Fig. S10). Moreover, read-pair support was not valid for detecting deletions.

We note that CONGA outputs the C-scores and read-pair counts for all input CNVs. Users can choose alternative cutoffs to increase recall (higher C-scores) or precision (lower C-scores).

The simulation experiments showed that CONGA was not efficient in identifying events $<$1 kbps. CONGA therefore ignores events $<$1 kbps under default parameters. This can be modified by the user if needed.

Analysis of real ancient genomes

Ancient genome selection and preprocessing

We selected 71 ancient shotgun or whole-genome captured genomes from individuals excavated in West and East Eurasia and in North America (Supplemental Table S2). Our sample set belongs to a time range between c.2,800-45,000 years Before Present (BP). Samples from 10 different laboratories were selected in order to study the effects of different data production protocols on deletion genotyping. We also chose genomes with a range of coverage levels (0.04$\times$-26$\times$, median = 3.45$\times$) and that included both UDG-treated and non-UDG-treated libraries. The only capture-produced data was Bon002 Kılınç *et al.* (2016), produced using whole-genome hybridization with myBaits (Arbor Biosciences, USA) probes.

Selected ancient genomes were mapped to the human reference genome (hg19, or GRCh37) using BWA aln/samse (0.7.15) Li and Durbin (2009) with parameters "-n 0.01, -o 2". PCR duplicates were removed using FilterUniqueSAM-Cons.py Kircher (2012).

We also removed reads with $>10\%$ mismatches to the reference genome, those of size $<$35 bps, and with $<$30 mapping quality (MAPQ).

Candidate CNV call set for real ancient genomes

Here our goal was to study properties of deletion variants in ancient genomes and to compare these with SNP variation in terms of demographic history and purifying selection. Polymorphism data sets can suffer from ascertainment bias in downstream evolutionary analyses Clark *et al.* (2005). A common practice to avoid this bias is to use SNPs ascertained in a population that is an outgroup to the focal populations. We therefore used variants ascertained in modern-day African populations for both calling SNP and deletion variants in our ancient genomes.

In order to create a candidate deletion call set to be used as input to CONGA, we downloaded deletions of size >1000 bps identified among 661 African population (AFR) genomes of the 1000 Genomes Project Phase 3 Sudmant *et al.* (2015b). When a deletion was located inside the breakpoints of another deletion, we removed the internal one. In addition, for pairs of deletions that had $>50\%$ overlap, we filtered out the smaller one. Finally, we filtered out deletion loci with $<50\%$ average mappability (see above). This resulted in $11,390$ autosomal >1000 bps deletions from 661 AFR genomes.

We filtered these deletions for high mappability (≥ 0.9 average mappability) and being derived in the human lineage (see section "Ancestral state determination" below). This left us with 10,002 deletion loci.

Deletion genotyping in ancient genomes

We genotyped all the chosen 71 ancient genomes using the 11,390 AFR autosomal deletion data set (>1 kbps with mean 10,735 bps). We used C-score <0.5 as cutoff for calling deletions, and >2 for calling the reference homozygous genotype (0/0). To limit false negatives, C-scores between 0.5 and 2 were coded as missing (NA). Note that these cutoffs can be modified by the user.

In total, 1,222 deletion loci (12%) out of 10,002 were missing across all the 71 genomes. Of the remaining, 5,467 were genotyped as a deletion in heterozygous or homozygous state in at least one genome. Genotyping rates (non-missing values) in the full dataset was overall 80.0%.

Analyzing the ancient deletion dataset

We generated a heatmap summarizing deletion copy numbers using the R "gplots" package "heatmap.2" function Warnes *et al.* (2020). Further, we performed a principal components analyses (PCA) on the deletion copy number data set (removing missing values) with 71, 60 (first outlier filter) and 50 (refined data set) ancient genomes (Supplemental Fig. S4). PC1 and PC2 values were computed using the R "stats" package "prcomp" function using the default parameters R Core Team (2020). On the same 3 genome sets, we likewise created multidimensional scaling plots (MDS) calculated with parameter "k=2" with the R "cmdscale" function on a Euclidean distance matrix of deletion frequencies (without removing NAs), and hierarchical clustering trees summarizing Manhattan distance matrices, calculated with the R "dist" and "hclust" functions. This analysis revealed visible outliers in deletion frequency among samples, which we defined as the "divergent" genome set (Supplemental Fig. S3A; Supplemental Note S3).

Based on this observation, we compared the total number of missing values, average read length, and coverage between the divergent genome set (n=21) and the rest, which we refer to as the "coherent" set (n=50), using the Mann-Whitney U test with the R "wilcox.test" function, and visualized the data with R utility function "boxplot" R Core Team (2020) (Supplemental Fig. S5A, B, C, D). We likewise compared average deletion frequencies between UDG-treated and untreated genomes using the Mann-Whitney U test.

Creating and analyzing the refined deletion data set and the SNP data set

SNP genotyping in ancient genomes

Following the same reasoning as above regarding ascertainment bias, we used an African population to create a SNP genotyping set for calling SNPs in the ancient genomes. Specifically, we used the 1000 Genomes Yoruba data set, which included a total of 38,945,054 autosomal bi-allelic SNPs (minor allele frequency > 0) in 661 African genomes of the 1000 Genomes Project Phase 3 The 1000 Genomes Project Consortium (2015). First, all reads in all BAM files were clipped (trimmed) using the trimBam algorithm implemented in BamUtil Jun *et al.* (2015). Following standard practice Mittnik *et al.* (2018), we trimmed (a) the end 2 bases of each read for samples prepared with the Uracil-DNA-glycosylase (UDG) protocol, and (b) the end 10 bases of each read for non-UDG samples.

Using these BAM files of the 50 ancient individuals and the above-described SNP list, we generated pseudo-haploid SNP calls at these target SNP positions by randomly selecting one read and recording the allele carried on that read

as the genotype. This was performed using the pileupCaller software (https://github.com/stschiff/sequenceTools) on samtools mpileup output (base quality>30 and MAPQ>30) Li *et al.* (2009).

Ancestral state determination

To polarize deletion and SNP alleles for being ancestral or derived in the human lineage, we mapped loci from hg19 (GRCh37) to panTro6 (chimpanzee) and to panPan2 (bonobo) using the UCSC Genome Browser tool "liftOver" with default parameters Kent *et al.* (2002). For deletions, we filtered out deletions that did not fully map to either chimpanzee or bonobo reference genomes, as these could represent derived insertions in the human lineage. The remaining deletions could thus be inferred to be alleles that were derived in humans. For SNPs, we removed the positions not represent in either chimpanzee or bonobo reference genomes and assigned the ancestral state as the Pan allele, only if both chimpanzee and bonobo carried same allele. This left us with 32,344,446 SNP positions with derived allele information.

Creating the refined deletion data set

We removed 21 genomes identified as outliers in both heatmap, PCA and MDS analyses. Next, we genotyped the 8,780 AFR deletions in the remaining 50 genomes. We call this the "refined data set". After refining our data set, we also checked its general properties. We plotted size distribution in logarithmic scale, deletion allele frequency distribution and relative frequency distribution among observed heterozygous deletions over homozygous deletions using R's "graphics" package hist function (Supplemental Fig. S14) R Core Team (2020). We also plotted relative deletion (homozygous or heterozygous) frequencies of 8,780 deletions for each individual in our refined data set using R's "graphics" package matplot function R Core Team (2020).

Genetic distance and selection analyses using deletions and SNPs

Here our goal was to calculate pairwise genetic distances among the 50 ancient genomes using deletion allele frequencies and using SNPs, and further to compare the distances. We calculated distances using the commonly used outgroup-f_3 statistics, which measures shared genetic drift between two samples relative to an outgroup, and is implemented as qp3pop in Admixtools v.7.0 Patterson *et al.* (2012). The outgroup-f_3 values were calculated for each pair of 50 individuals (a) in the deletion and (b) in the SNP data sets, using the African Yoruba as outgroup in both cases. To convert the deletion data set to eigenstrat format, which Admixtools requires, we encoded the first nucleotide of each deletion as the reference allele, and the alternative allele was randomly assigned among the remaining 3 nucleotides using custom Python script. We thus calculated a pairwise similarity matrix for both data sets. Genetic distances were calculated as 1-f_3. Distances were then summarized using multidimensional scaling (MDS) with the "cmdscale" function of R R Core Team (2020) (Figure 4C, D; Supplemental Fig. S4).

We further performed the Mantel test to compare the f_3-based similarity matrices calculated using SNPs and deletions. We used the "mantel" function in the R-package "vegan" with parameter "method=spearman" Oksanen *et al.* (2013).

Site frequency spectrum calculation for deletions and SNPs

Here our goal was to compare the SFS across deletions and SNPs called in ancient genomes. Because the ancient SNP genotypes are pseudo-haploidized, we performed the same pseudo-haploidization process on the deletion data set. For this, for any heterozygous call in the deletion data set, we randomly assigned either of the homozygous states, using the R "sample" function (i.e., we converted 1's to 0's or 2's with 50% probability). We then counted derived alleles at each locus, for deletions and for SNPs, and divided by the total number of genomes where an allele was observed at that locus (i.e., removing the missing data). We plotted the site-frequency spectrum analysis on both deletions and SNPs using R's "ggplot2" package geom_histogram function Wickham (2016). We also calculated the Spearman correlation between the deletion size in logarithmic scale and the frequency using R's "stats" package "cor.test" function R Core Team (2020). Further, we plotted the site-frequency spectrum analysis on deletions in high and low coverage genomes using R's "ggplot2" package geom_histogram function Wickham (2016) (Supplemental Fig. S6). The threshold is considered to be the median coverage (3.98×).

Evolutionary conservation

To measure evolutionary conservation for genes that overlapped deletions, we retrieved non-synonymous (dN) and synonymous (dS) substitution rate estimates between human (GRCh37) and the mouse genome (GRCm38) per gene from Ensembl (v75) via the R package "biomaRt" Durinck *et al.* (2005). We queried 18,112 genes with dN, dS values and calculated the dN/dS ratio (or Ka/Ks) per gene. The ratio for genes with more than one dN or dS values were calculated as the mean dN or dS per gene. We then intersected our deletions with the genes with dN/dS values using

BEDTools Quinlan and Hall (2010) and found 2,221 Ensembl (v75) human genes. Overall, 34% of the 10,002 derived deletions overlapped with at least one gene. We then collected mouse-human dN/dS ratios (Methods) for these genes (n = 2,221, 0-1.18, median = 0.09, mean = 0.13). For deletions overlapping with multiple genes, we calculated the mean dN/dS per deletion. We then divided the deletions in our data set into two groups by the deletion allele frequency: high versus low relative to the median. We plotted the dN/dS ratios of the deletion groups defined above using the R package "ggplot2" and the "geom_boxplot" function Wickham (2016).

Comparison with SIFT predictions and temporal change

Here our goal was to study deleterious mutation loads per genome in the form of SIFT-predicted harmful SNPs and CONGA-predicted deletions, across the 50 ancient genomes. We used SIFT predictions available in Ensembl (v75) collected via the R package "biomaRt" Durinck *et al.* (2005, 2009). We retrieved SIFT predictions of "tolerated" and "deleterious" impact and SIFT scores for all 1000 Genomes human SNPs from Ensembl, and subsetted the African SNP set used for genotyping the ancient genomes. This resulted in 22,996 SNPs with SIFT predictions. Further, we calculated a ratio representing the total number of SIFT-predicted "deleterious" SNPs over the number of "tolerated" SNPs, for each of the 50 individuals. In addition, we calculated the total CONGA-predicted deletion length and the total number of genes overlapping CONGA-predicted deletions per individual, ignoring homozygous or heterozygous state. We plotted these three mutation load scores, i.e. SIFT-predicted deleterious/tolerated ratios per individual, the number of affected genes, and the total deletion length, using R base function "plot" (Supplemental Fig. S7) R Core Team (2020). We further estimated pairwise correlations between the three scores, fitting the values into a linear model using the R "lm" function and calculating Spearman's rank correlation. We plotted the linear models using the R base function "pairs" (Supplemental Fig. S8B) R Core Team (2020).

We finally tested whether the mean deletion allele frequency changed over time by fitting the values in a linear model using the R "lm" function (Supplemental Fig. S9).

Software Availability

CONGA is implemented in C programming language and its source code is available under BSD 3-clause license at https://github.com/asylvz/CONGA, as well as Supplemental Code. Simulated datasets and predictions of each tool can be accessed through Zenodo (10.5281/zenodo.5555990). Mappability data was downloaded from `http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/`

Competing interest statement

The authors declare no competing interests.

Acknowledgements

The authors would like to thank Gözde Zeliha Turan for her suggestions with mappability data and Kıvılcım Başak Vural for technical support. This work was supported by the ERC Consolidator grant "NEOGENE" (Project No.: 772390).

Author Contributions

AS developed and implemented the algorithm, performed simulations and down-sampling experiments. SSÇ and DK conducted technical and evolutionary analyses on real data. CA contributed to algorithm design. MS led the project and coordinated the activities. All authors contributed to editing the manuscript and participated in weekly discussions.

References

(2019). Picard toolkit. `https://broadinstitute.github.io/picard/`.

Abyzov, A. *et al.* (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*, **21**(6), 974–984.

Alkan, C. (2020). Automatic characterization of copy number polymorphism using high throughput sequencing. *Turkish Journal of Electrical Engineering & Computer Sciences*, **28**(1), 253–261.

Alkan, C. *et al.* (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet*, **41**(10), 1061–1067.

Alkan, C. *et al.* (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet*, **12**(5), 363–376.

Allentoft, M. E. *et al.* (2015). Population genomics of bronze age eurasia. *Nature*, **522**(7555), 167.

Almarri, M. A. *et al.* (2020). Population structure, stratification, and introgression of human structural variation. *Cell*, **182**(1), 189–199.

Antonio, M. L. *et al.* (2019). Ancient rome: A genetic crossroads of europe and the mediterranean. *Science*, **366**(6466), 708–714.

Audano, P. A. *et al.* (2019). Characterizing the major structural variant alleles of the human genome. *Cell*, **176**(3), 663–675.

Bergström, A. *et al.* (2020). Origins and genetic legacy of prehistoric dogs. *Science*, **370**(6516), 557–564.

Bhattacharya, S. *et al.* (2018). Whole-genome sequencing of atacama skeleton shows novel mutations linked with dysplasia. *Genome research*, **28**(4), 423–431.

Boeva, V. *et al.* (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics (Oxford, England)*, **28**, 423–425.

Broushaki, F. *et al.* (2016). Early neolithic genomes from the eastern fertile crescent. *Science*, **353**(6298), 499–503.

Chaisson, M. J. P. *et al.* (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, **517**, 608–611.

Chaisson, M. J. P. *et al.* (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, **10**, 1784.

Chan, Y. F. *et al.* (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitx1 enhancer. *science*, **327**(5963), 302–305.

Chen, X. *et al.* (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, **32**, 1220–1222.

Chiang, D. Y. *et al.* (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods*, **6**(1), 99–103.

Clark, A. G. *et al.* (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome research*, **15**(11), 1496–1502.

Collins, R. L. *et al.* (2020). A structural variation reference for medical and population genetics. *Nature*, **581**(7809), 444–451.

Conrad, D. F. and Hurles, M. E. (2007). The population genetics of structural variation. *Nature genetics*, **39**(7), S30–S36.

Conrad, D. F. *et al.* (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, **464**(7289), 704–712.

Cooper, G. M. *et al.* (2011). A copy number variation morbidity map of developmental delay. *Nature genetics*, **43**(9), 838–846.

de Barros Damgaard, P. *et al.* (2018a). 137 ancient human genomes from across the eurasian steppes. *Nature*, **557**(7705), 369–374.

de Barros Damgaard, P. *et al.* (2018b). The first horse herders and the impact of early bronze age steppe expansions into asia. *Science*, **360**(6396), eaar7711.

Derrien, T. *et al.* (2012). Fast computation and applications of genome mappability. *PloS one*, **7**(1), e30377.

Durinck, S. *et al.* (2005). Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**(16), 3439–3440.

Durinck, S. *et al.* (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, **4**(8), 1184–1191.

Eisfeldt, J. *et al.* (2017). TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. *F1000Research*, **6**, 664.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Foote, A. D. *et al.* (2021). Evidence of long-term purging of mutation load in killer whale genomes. *Biorxiv*.

Frantz, L. A. *et al.* (2020). Animal domestication in the era of ancient genomics. *Nature Reviews Genetics*, pages 1–12.

Fu, Q. *et al.* (2014). Genome sequence of a 45,000-year-old modern human from western siberia. *Nature*, **514**(7523), 445–449.

Gamba, C. *et al.* (2014). Genome flux and stasis in a five millennium transect of european prehistory. *Nature communications*, **5**(1), 1–9.

Girirajan, S. *et al.* (2011). Human copy number variation and complex genetic disease. *Annual review of genetics*, **45**, 203–226.

Gonzalez, E. *et al.* (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, **307**(5714), 1434–1440.

González-Fortes, G. *et al.* (2017). Paleogenomic evidence for multi-generational mixing between neolithic farmers and mesolithic hunter-gatherers in the lower danube basin. *Current Biology*, **27**(12), 1801–1810.

Green, R. E. *et al.* (2010). A draft sequence of the Neandertal genome. *Science*, **328**(5979), 710–722.

Günther, T. *et al.* (2015). Ancient genomes link early farmers from atapuerca in spain to modern-day basques. *Proceedings of the National Academy of Sciences*, **112**(38), 11917–11922.

Haber, M. *et al.* (2017). Continuity and admixture in the last five millennia of levantine history from ancient canaanite and present-day lebanese genome sequences. *The American Journal of Human Genetics*, **101**(2), 274–282.

Haber, M. *et al.* (2019). A transient pulse of genetic admixture from the crusaders in the near east identified from ancient genome sequences. *The American Journal of Human Genetics*, **104**(5), 977–984.

Handsaker, R. E. *et al.* (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet*, **43**(3), 269–276.

Handsaker, R. E. *et al.* (2015). Large multiallelic copy number variations in humans. *Nat Genet*, **47**(3), 296–303.

Hardwick, R. J. *et al.* (2011). A worldwide analysis of beta-defensin copy number variation suggests recent selection of a high-expressing defb103 gene copy in east asia. *Human mutation*, **32**(7), 743–750.

Ho, S. S. *et al.* (2020). Structural variation in the sequencing era. *Nature Reviews Genetics*, **21**(3), 171–189.

Hofmanová, Z. *et al.* (2016). Early farmers from across europe directly descended from neolithic aegeans. *Proceedings of the National Academy of Sciences*, **113**(25), 6886–6891.

Hsieh, P. *et al.* (2019). Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science*, **366**(6463).

Jones, E. R. *et al.* (2015). Upper palaeolithic genomes reveal deep roots of modern eurasians. *Nature communications*, **6**(1), 1–8.

Jones, E. R. *et al.* (2017). The neolithic transition in the baltic was not driven by admixture with early european farmers. *Current Biology*, **27**(4), 576–582.

Jun, G. *et al.* (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale dna sequence data. *Genome research*, **25**(6), 918–925.

Kahveci, F. and Alkan, C. (2018). Whole-genome shotgun sequence CNV detection using read depth. *Methods in molecular biology*, **1833**, 61–72.

Karakoc, E. *et al.* (2012). Detection of structural variants and indels within exome data. *Nat Methods*, **9**(2), 176–178.

Karimzadeh, M. *et al.* (2018). Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Research*, **46**(20), e120–e120.

Keller, A. *et al.* (2012). New insights into the tyrolean iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature communications*, **3**(1), 1–9.

Kent, W. J. *et al.* (2002). The human genome browser at ucsc. *Genome research*, **12**(6), 996–1006.

Kılınç, G. M. *et al.* (2016). The demographic development of the first farmers in anatolia. *Current Biology*, **26**(19), 2659–2666.

Kircher, M. (2012). Analysis of high-throughput ancient dna sequencing data. In *Ancient DNA*, pages 197–228. Springer.

Koehler, R. *et al.* (2011). The uniqueome: a mappability resource for short-tag sequencing. *Bioinformatics*, **27**(2), 272–274.

Kothapalli, K. S. *et al.* (2016). Positive selection on a regulatory insertion–deletion polymorphism in fads2 influences apparent endogenous synthesis of arachidonic acid. *Molecular biology and evolution*, **33**(7), 1726–1739.

Kousathanas, A. *et al.* (2017). Inferring heterozygosity from ancient and low coverage genomes. *Genetics*, **205**(1), 317–332.

Krzewińska, M. *et al.* (2018). Ancient genomes suggest the eastern pontic-caspian steppe as the source of western iron age nomads. *Science advances*, **4**(10), eaat4457.

Layer, R. M. *et al.* (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*, **15**(6), R84.

Lazaridis, I. *et al.* (2014). Ancient human genomes suggest three ancestral populations for present-day europeans. *Nature*, **513**(7518), 409–413.

Levy-Sakin, M. *et al.* (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature communications*, **10**, 1025.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**(14), 1754–1760.

Li, H. *et al.* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Link, V. *et al.* (2017). Atlas: analysis tools for low-depth and ancient samples. *BioRxiv*, page 105346.

Llorente, M. G. *et al.* (2015). Ancient ethiopian genome reveals extensive eurasian admixture in eastern africa. *Science*, **350**(6262), 820–822.

Marciniak, S. and Perry, G. H. (2017). Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics*, **18**(11), 659–674.

Marsden, C. D. *et al.* (2016). Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences*, **113**(1), 152–157.

Martiniano, R. *et al.* (2017). The population genomics of archaeological transition in west iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS genetics*, **13**(7), e1006852.

Mathieson, S. and Mathieson, I. (2018). Fads1 and the timing of human adaptation to agriculture. *Molecular biology and evolution*, **35**(12), 2957–2970.

McLean, C. Y. *et al.* (2011). Human-specific loss of regulatory dna and the evolution of human-specific traits. *Nature*, **471**(7337), 216–219.

Meyer, M. *et al.* (2012). A high-coverage genome sequence from an archaic denisovan individual. *Science*, **338**(6104), 222–226.

Miller, C. A. *et al.* (2011). ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PloS one*, **6**, e16327.

Mittnik, A. *et al.* (2018). The genetic prehistory of the baltic sea region. *Nature communications*, **9**(1), 1–11.

Mu, J. C. *et al.* (2015). VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics*, **31**(9), 1469–1471.

Ng, P. C. and Henikoff, S. (2003). Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, **31**(13), 3812–3814.

Nuttle, X. *et al.* (2016). Emergence of a homo sapiens-specific gene family and chromosome 16p11. 2 cnv susceptibility. *Nature*, **536**(7615), 205–209.

Oksanen, J. *et al.* (2013). Package 'vegan'. *Community ecology package, version*, **2**(9), 1–295.

Olalde, I. *et al.* (2014). Derived immune and ancestral pigmentation alleles in a 7,000-year-old mesolithic european. *Nature*, **507**(7491), 225–228.

Oliva, A. *et al.* (2021). Bwa-mem is not the best aligner for ancient dna short reads. *bioRxiv*.

Orlando, L. *et al.* (2021). Ancient dna analysis. *Nature Reviews Methods Primers*, **1**(1), 1–26.

Patterson, N. *et al.* (2012). Ancient admixture in human history. *Genetics*, **192**(3), 1065–1093.

Pedersen, J. S. *et al.* (2014). Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome research*, **24**(3), 454–466.

Perry, G. H. *et al.* (2007). Diet and the evolution of human amylase gene copy number variation. *Nature genetics*, **39**(10), 1256–1260.

Pockrandt, C. *et al.* (2020). Genmap: ultra-fast computation of genome mappability. *Bioinformatics*.

Prüfer, K. (2018). snpad: An ancient dna genotype caller. *Bioinformatics*, **34**(24), 4165–4171.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raghavan, M. *et al.* (2014). Upper palaeolithic siberian genome reveals dual ancestry of native americans. *Nature*, **505**(7481), 87–91.

Rasmussen, M. *et al.* (2010). Ancient human genome sequence of an extinct palaeo-eskimo. *Nature*, **463**(7282), 757–762.

Rasmussen, M. *et al.* (2014). The genome of a late pleistocene human from a clovis burial site in western montana. *Nature*, **506**(7487), 225–229.

Rausch, T. *et al.* (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**(18), i333–i339.

Reich, D. *et al.* (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468**(7327), 1053–1060.

Renaud, G. *et al.* (2017). gargammel: a sequence simulator for ancient dna. *Bioinformatics*, **33**(4), 577–579.

Rogers, R. L. and Slatkin, M. (2017). Excess of genomic defects in a woolly mammoth on wrangel island. *PLoS genetics*, **13**(3), e1006601.

Rohland, N. *et al.* (2015). Partial uracil–dna–glycosylase treatment for screening of ancient dna. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **370**(1660), 20130624.

Saitou, M. and Gokcumen, O. (2020). An evolutionary perspective on the impact of genomic copy number variation on human health. *Journal of molecular evolution*, **88**(1), 104–119.

Schubert, M. *et al.* (2016). Adapterremoval v2: rapid adapter trimming, identification, and read merging. *BMC research notes*, **9**(1), 1–7.

Sedlazeck, F. J. *et al.* (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature methods*, **15**, 461–468.

Seguin-Orlando, A. *et al.* (2014). Genomic structure in europeans dating back at least 36,200 years. *Science*, **346**(6213), 1113–1118.

Shapiro, á. and Hofreiter, M. (2014). A paleogenomic perspective on evolution and gene function: new insights from ancient dna. *Science*, **343**(6169), 1236573.

Sikora, M. *et al.* (2019). The population history of northeastern siberia since the pleistocene. *Nature*, **570**(7760), 182–188.

Skoglund, P. and Mathieson, I. (2018). Ancient genomics of modern humans: the first decade. *Annual review of genomics and human genetics*, **19**, 381–404.

Smith, D. R. *et al.* (2008). Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res*, **18**(10), 1638–1642.

Smith, S. D. *et al.* (2015). Grom-rd: resolving genomic biases to improve read depth detection of copy number variants. *PeerJ*, **3**, e836.

Smith, S. D. *et al.* (2017). Evolutionary adaptation revealed by comparative genome analysis of woolly mammoths and elephants. *DNA Research*, **24**(4), 359–369.

Soylev, A. *et al.* (2017). Toolkit for automated and rapid discovery of structural variants. *Methods*, **129**, 3–7.

Soylev, A. *et al.* (2019). Discovery of tandem and interspersed segmental duplications using high-throughput sequencing. *Bioinformatics*, **35**, 3923–3930.

Stankiewicz, P. and Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annu Rev Med*, **61**, 437–455.

Sudmant, P. H. *et al.* (2010). Diversity of human copy number variation and multicopy genes. *Science*, **330**(6004), 641–646.

Sudmant, P. H. *et al.* (2015a). Global diversity, population stratification, and selection of human copy-number variation. *Science*, **349**(6253).

Sudmant, P. H. *et al.* (2015b). An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**(7571), 75–81.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74.

Vieira, F. G. *et al.* (2016). Estimating ibd tracts from low coverage ngs data. *Bioinformatics*, **32**(14), 2096–2102.

Warnes, G. R. *et al.* (2020). *gplots: Various R Programming Tools for Plotting Data*. R package version 3.1.1.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Xie, C. and Tammi, M. T. (2009). Cnv-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC bioinformatics*, **10**(1), 1–9.

Xue, Y. *et al.* (2008). Adaptive evolution of ugt2b17 copy-number variation. *The American Journal of Human Genetics*, **83**(3), 337–346.

Yaka, R. *et al.* (2021). Variable kinship patterns in neolithic anatolia revealed by ancient genomes. *Current Biology*, **31**(11), 2455–2468.

Yoon, S. *et al.* (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res*, **19**(9), 1586–1592.

Zhang, F. *et al.* (2009). Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics*, **10**, 451–481.

Zook, J. M. *et al.* (2020). A robust benchmark for detection of germline large deletions and insertions. *Nature Biotechnology*, pages 1–9.