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, Middle East Technical University, Ankara, 06800, Turkey*
[4]*Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey*

[*]*Corresponding authors: arda.soylev@gidatarim.edu.tr 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 (a) most published genomes being $<1\times$ coverage, (ii) ancient DNA fragments being typically $<$80 bps. These characteristics preclude state-of-the-art CNV detection software to be effectively applied to ancient genomes. Here we present CONGA, an algorithm tailored for genotyping deletion and duplication events in genomes with low depths of coverage. Simulations show that CONGA can genotype deletions and duplications >1 Kbps with F-scores >0.77 and >0.82 at $\geq 0.5\times$, respectively. Further, down-sampling experiments using published ancient BAM files reveal that >1 Kbps deletions could be genotyped at F-score >0.75 at $\geq 1\times$ coverage. Using CONGA, we analyse deletion events at 10,018 loci in 56 ancient human genomes spanning the last 50,000 years, with coverages $0.4\times$-$26\times$. We find inter-individual genetic diversity measured using deletions and SNPs to be highly correlated, suggesting that deletion frequencies broadly reflect demographic history. We also identify signatures of purifying selection on deletions, such as an excess of singletons compared to those in SNPs. CONGA paves the way for systematic studies of drift, mutation load, and adaptation in ancient and modern-day gene pools through the lens of CNVs.

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

1 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; Saitou and Gokcumen, 2020; Stankiewicz and Lupski, 2010; Girirajan *et al.*, 2011)). 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 be highly revealing about the genetic health of lineages. For instance, a study on the last

surviving mammoth population on Wrangel Island has found an excess of deletions in this sample, which could 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; Alkan, 2020; Smith *et al.*, 2015) and read-pair information (Soylev *et al.*, 2017, 2019; Rausch *et al.*, 2012; Layer *et al.*, 2014; Chen *et al.*, 2016; Eisfeldt *et al.*, 2017), or long reads (Sedlazeck *et al.*, 2018; Chaisson *et al.*, 2015). 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. Available CNV discovery methods are therefore inapplicable to most ancient genome data sets, and so far, no specific algorithm for CNV identification in ancient genomes has been developed.

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 genomes, which estimates copy number beyond presence/absence of events. To our knowledge, CONGA is the first tool specifically developed for this purpose. There are four reasons behind the use of CNV genotyping instead of CNV discovery: (a) CNV discovery using low coverage ancient genomes is impractical if not impossible; (b) for many species studied using ancient genomics, CNV reference sets based on high quality genomes are already available; (c) ancient variation will largely overlap with modern-day variation in the vast majority of cases; (d) genotyping has much shorter running times and lower memory usage than discovery. Indeed, most human ancient genome studies to date have chosen genotyping over *de novo* SNP discovery (Prüfer, 2018; Link *et al.*, 2017). It should likewise be possible to genotype CNVs in low coverage genomes with high accuracy and in short running times using depth of coverage and split-read information. We therefore believe that CONGA's approach can efficiently open up CNV analyses to ancient genome studies, as well as low coverage whole genome sequencing datasets of extant organisms.

We first test the accuracy of CONGA using purely simulated ancient genomes as well as down-sampling simulations of real ancient genomes. We show that our algorithm exhibits reliable performance for deletions and duplications even at low depths of coverage (i.e., $<1\times$). We also discover that published ancient genome BAM files may frequently not be suitable for duplication genotyping at low coverage. We next genotype deletions in a set of real ancient human genomes using deletions ascertained in present-day African genomes (Figure 1). We find that accurate CNV genotyping is sensitive to a number of technical parameters, requiring comprehensive filtering to achieve a high quality CNV call set. Having created such a deletion CNV set, we show that CNV diversity largely represents demographic history, paralleling SNP diversity among the same ancient individuals. We further present evidence for negative selection on deletions in these ancient genomes.

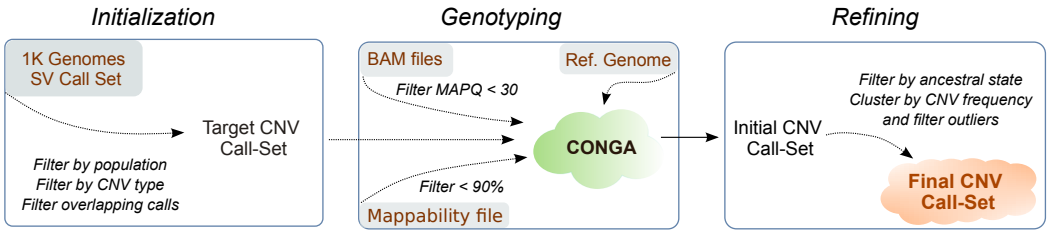

Figure 1: Overall workflow of CONGA. The first step involves initialization, where we create the input (reference) CNV file using the deletions and duplications of African populations (AFR) from Phase 3 of the 1000 Genomes Project SV call set (Sudmant *et al.*, 2015). We apply our genotyping algorithm 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.

2 Results

2.1 Overview of the algorithm

We developed CONGA to genotype given candidate CNVs in mapped read (BAM) files (Methods). 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. The genotypes can be homozygous CNV, heterozygous CNV, or no CNV. Using these likelihoods CONGA then calculates a statistic that 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).

2.2 Accuracy evaluation using simulated genomes and comparison with CNV discovery tools

To evaluate the performance of our algorithm 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. 1 for the exact numbers and length distributions of CNVs 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) as well as post-mortem damage. The data was generated at various depths: $0.05\times$, $0.1\times$, $0.5\times$, $1\times$ and $5\times$ (Methods). We then used CONGA to genotype CNVs across the simulated ancient genomes using the candidate CNV call set. In order to assess specificity in addition to 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.*, 2015). We mixed these false CNVs to the list of true CNVs with a ratio of approximately 10:1 (roughly 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 two of the widely used CNV discovery approaches: CNVnator (Abyzov *et al.*, 2011) and FREEC (Boeva *et al.*, 2012).

Table 1 shows true and false predictions by CONGA, FREEC and CNVnator, as well as their true positive rate (TPR), false discovery rate (FDR) and the F-Score for identifying deletions and duplications of small, medium and large size (as defined above). In detecting deletions and duplications, CONGA achieved a substantially lower FDR and overall higher sensitivity than the other two methods. Note that even at $0.05\times$ coverage, CONGA yields accurate results for medium and large sized CNVs, whereas the other tools clearly fail. On the other hand, the performance of each tool becomes comparable as the coverage approaches to depths of $5\times$ with large CNV sizes. For small CNVs ($<$1 Kbps), all three tools underperform, although CONGA predictions for small deletions still have higher recall and precision than the other two tools (see Supplemental Fig. 2 for precision-recall curves). We note that CONGA's genotyping approach naturally boosts its performance over FREEC and CNVnator, which are CNV discovery tools. Still, the fact that our candidate CNV call set included 10 times more false CNVs than true CNVs indicates the overall reliability of CONGA in CNV identification in ancient genomes relative to these alternatives.

The simulation results thus suggest that CONGA can efficiently and accurately genotype deletions and duplications of length $>$1 Kbps in ancient genomes at $\geq 0.5\times$ coverage. We then turned to studying the performance of CONGA in identifying copy numbers for medium and large size CNVs.

2.2.1 Copy number prediction 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 ancient 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. For simplicity, CONGA only evaluates the possibility of homozygous duplications (ignoring copy numbers $\geq$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 $\geq$0.5 at coverages $\geq 0.5\times$. Importantly, CONGA had comparable power in identifying heterozygous and homozygous events of size $>$1 Kbps (Supplemental Table S1.B).

2.3 Down-sampling experiments with real ancient genomes

Beyond pure simulations, we further 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 (BAM or FASTQ files) of relatively high coverage ($\geq 9\times$), (ii) we genotyped CNVs using the full genome data with CONGA and with a modern-day human CNV call set, (iii) we

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

	Cov.	Total	CONGA					FREEC					CNVNator				
			True	False	TPR	FDR	F-Score	True	False	TPR	FDR	F-Score	True	False	TPR	FDR	F-Score
Dels (small)	0.05×	1810	1474	3724	**0.81**	**0.72**	**0.42**	0	1221	0.00	1.00	-	3	47442	0.00	1.00	0.00
	0.1×	1810	1268	2415	**0.70**	**0.66**	**0.46**	0	198	0.00	1.00	-	0	402	0.00	1.00	-
	0.5×	1810	767	456	**0.42**	**0.37**	**0.51**	0	6761	0.00	1.00	-	0	806	0.00	1.00	-
	1×	1810	834	189	**0.46**	**0.18**	**0.59**	0	1916	0.00	1.00	-	0	263	0.00	1.00	-
	5×	1810	1236	89	**0.68**	**0.07**	**0.79**	20	392	0.01	0.95	0.02	341	493	0.19	0.59	0.26
Dups (small)	0.05×	1751	230	1878	**0.13**	**0.89**	**0.12**	0	44	0.00	1.00	-	7	47700	0.00	1.00	0.00
	0.1×	1751	128	1116	**0.07**	**0.90**	**0.09**	0	7	0.00	1.00	-	0	28699	0.00	1.00	-
	0.5×	1751	304	28	**0.17**	**0.08**	**0.29**	0	3	0.00	1.00	-	0	9	0.00	1.00	-
	1×	1751	618	22	**0.35**	**0.03**	**0.52**	0	555	0.00	1.00	-	0	884	0.00	1.00	-
	5×	1751	1183	30	**0.68**	0.02	**0.80**	35	77	0.02	0.69	0.04	2	0	**0.00**	0.00	0.00
Dels (medium)	0.05×	1680	868	791	**0.52**	**0.48**	**0.52**	0	83	0.00	1.00	-	0	68	0.00	1.00	-
	0.1×	1680	821	328	**0.49**	**0.29**	**0.58**	0	237	0.00	1.00	-	1	216	0.00	1.00	0.00
	0.5×	1680	1066	68	**0.63**	**0.06**	**0.76**	239	6433	0.14	0.96	0.06	187	257	0.11	0.58	0.18
	1×	1680	1268	56	**0.75**	**0.04**	**0.84**	421	2135	0.25	0.84	0.20	330	257	0.20	0.44	0.29
	5×	1680	1380	53	**0.82**	**0.04**	**0.89**	929	485	0.55	0.34	0.60	949	423	0.56	0.31	0.62
Dups (medium)	0.05×	1684	193	29	**0.11**	**0.13**	**0.20**	0	3	0.00	1.00	-	0	114	0.00	1.00	-
	0.1×	1684	384	20	**0.23**	**0.05**	**0.37**	0	3	0.00	1.00	-	0	102	0.00	1.00	-
	0.5×	1684	990	10	**0.59**	**0.01**	**0.74**	271	15	0.16	0.05	0.28	2	4	0.00	0.67	0.00
	1×	1684	1233	15	**0.73**	**0.01**	**0.84**	582	937	0.35	0.62	0.36	16	2	0.01	0.11	0.02
	5×	1684	1507	13	**0.89**	**0.01**	**0.94**	1000	329	0.59	0.25	0.66	105	2	0.06	0.02	0.12
Dels (large)	0.05×	1385	931	1017	**0.67**	**0.52**	**0.56**	0	87	0.00	1.00	-	84	131	0.06	0.61	0.11
	0.1×	1385	1025	549	**0.74**	**0.35**	**0.69**	0	754	0.00	1.00	-	560	246	0.40	0.31	0.51
	0.5×	1385	1093	334	**0.79**	0.23	**0.78**	664	3136	0.48	0.83	0.26	1049	293	0.76	**0.22**	0.77
	1×	1385	1083	373	0.78	0.26	0.76	1239	156	**0.89**	**0.11**	**0.89**	1204	309	0.87	0.20	0.83
	5×	1385	1072	383	0.77	0.26	0.75	1260	154	0.91	**0.11**	**0.90**	1265	453	**0.91**	**0.26**	0.82
Dups (large)	0.05×	1532	966	116	**0.63**	**0.11**	**0.74**	0	6	0.00	1.00	-	4	354	0.00	0.99	0.00
	0.1×	1532	1157	138	**0.76**	**0.11**	**0.82**	0	0	-	-	-	455	315	0.30	0.41	0.40
	0.5×	1532	1388	205	**0.91**	0.13	**0.89**	589	97	0.38	0.14	0.53	1039	77	0.68	**0.07**	0.78
	1×	1532	1402	223	**0.92**	0.14	**0.89**	1305	266	0.85	0.17	0.84	1216	94	0.79	**0.07**	0.86
	5×	1532	1410	243	**0.92**	0.15	0.89	1304	294	0.85	0.18	0.83	1350	165	0.88	**0.11**	**0.89**

Table shows the CNV prediction performance of CONGA, FREEC and CNVnator on simulated genomes with depth 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, **True** and **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. The **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 F-Score across the tools. CONGA consistently shows the best performance in identifying small and medium sized deletions and duplications, as well as the best performance in identifying large CNVs at low coverage. See Supplemental Material for the commands used to run each tool and Supplemental Table S1.A for details.

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 ($\sim 23.3\times$) ancient Eurasian genome (Yamnaya) (de Barros Damgaard *et al.*, 2018b), a $13.1\times$ ancient genome from Greenland (Saqqaq) (Rasmussen *et al.*, 2010), and a $9.6\times$ ancient genome from Ethiopia (Mota) (Llorente *et al.*, 2015). The Yamnaya genome was only available as a BAM file, while the latter two were 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 688-1581 deletions and 638-4097 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 the CONGA's TPR (true positive rate) and FDR (false discovery rate) 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, with TPR of $>70\%$ and FDR of $<45\%$ at coverages around $0.5\times$ (Figure 3, Supplemental Table S1.C). For duplications, however, CONGA's

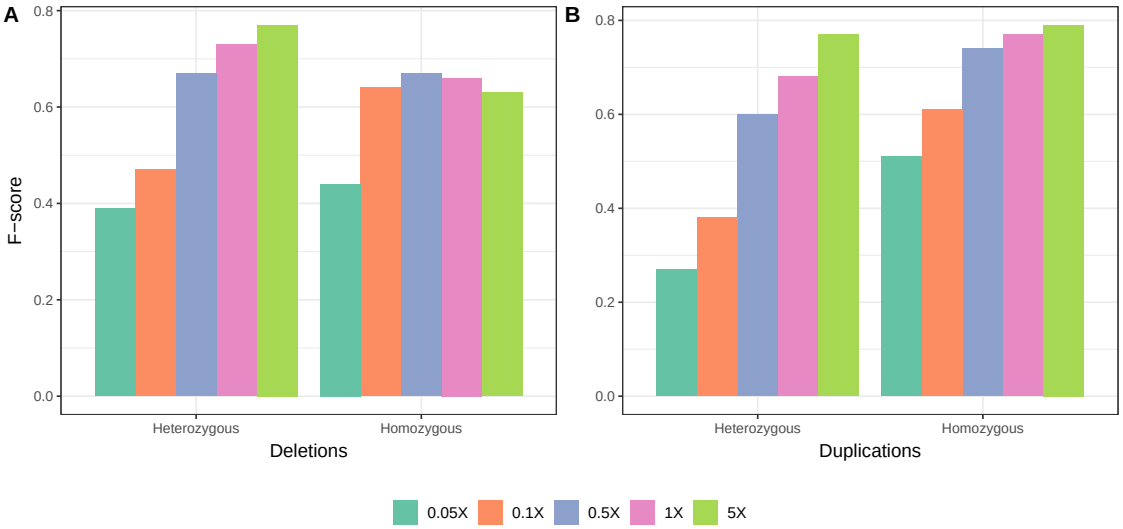

Figure 2: Copy-number prediction performance of CONGA using F-score for (A) deletions and (B) duplications using medium and large CNVs merged for various coverage values.

performance was lower. At around $1\times$ coverage, duplication TPR was $>40\%$ in the Saqqaq and Mota genomes, and only 22% in the Yamnaya genome.

To understand the reasons behind CONGA's striking under-performance at low coverages, especially on the Yamnaya genome, we studied the results in further detail. We noticed that only 47 (2.8%) of the 1,661 originally called duplication events in the full Yamnaya genome were genotyped using read-depth information (C-score), while the rest were genotyped using paired-end read information. This can explain CONGA's under-performance at low coverage in the Yamnaya genome: as coverage decreases, the number of paired-end reads supporting a duplication falls rapidly, compromising recall. The lack of usable read-depth information in the full Yamnaya genome, in turn, could be standard quality filters applied to the BAM file before data publication. Such filtering could have erased the read-depth signature, leaving only paired-end information available, which is not helpful at low coverages. This scenario is supported by the fact CONGA duplication calls are clearly more successful in the Saqqaq and Mota genomes at low coverage. These two were retrieved as original FASTQ files instead of processed BAM files. Moreover, 35% (1498/4027) and 27% (172/638) of duplication events were genotyped using read-depth information on the full Saqqaq and Mota genomes, respectively, in contrast to only 2.8% in the Yamnaya genome (see Discussion).

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. In contrast, although CONGA's accuracy in duplication identification was comparable to those of deletions in simulations, it was dramatically lower in correctly finding duplications in the down-sampled Yamnaya genome even at $1\times$ coverage. We suspect this low performance is caused by pre-publication quality filtering of BAM files. Consequently, duplication genotyping with CONGA directly on published ancient genomes appears infeasible, as ancient genomes are mainly submitted in BAM format in public repositories (see Discussion). We therefore limited the following analyses on real ancient genomes to deletions >1 Kbps.

2.4 Deletion genotyping across ancient genomes reveals strong influence by technical variables

We genotyped deletions with CONGA across a diverse sample of real ancient human genomes. Our naive hypothesis was that CNVs, 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 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

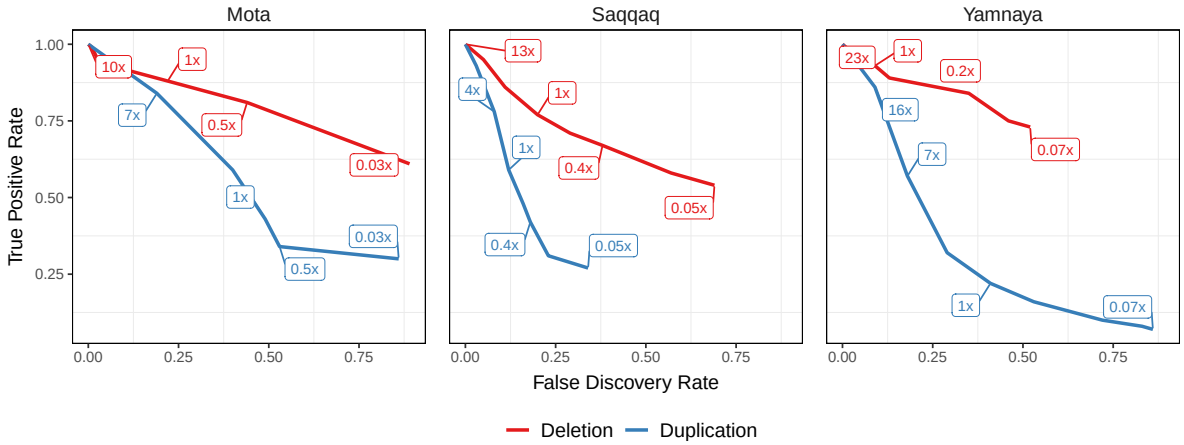

Figure 3: TPR vs FDR curves for deletion and duplication predictions of CONGA using Mota, Saqqaq and Yamnaya genomes down-sampled to various depths from their original coverages of 9.6×, 13.1× and 23.3×, respectively. The numbers inside boxes show the down-sampled coverage values. We calculated TPR and 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).

et al., 2017; Haber *et al.*, 2017; Raghavan *et al.*, 2014; Martiniano *et al.*, 2017; 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.*, 2015) (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. Genotyping these 11,390 deletions across the 71 BAM files resulted in a median number of 905 deletions [380-3,649] detected per genome, in either heterozygous or homozygous state. Of the 11,390 deletions, 8,106 (71%) were detected in at least one genome.

We then studied deletion copy numbers, or frequencies, across the 71 ancient genomes. Unexpectedly, both hierarchical clustering (displayed as a heatmap in Supplemental Fig. 3A) and principal components analysis (PCA) (Figure 4A) revealed that genomes grouped into two in terms of deletion frequencies (Supplemental Fig. 3A). We noticed that this grouping mainly represented laboratories-of-origin, irrespective of the population-of-origin. One set, which we call Group 1, contained 30 individuals from 8 laboratories, while the other, Group 2, included 41 genomes from 4 laboratories (Figure 4A). Only 2 of the 10 laboratories had data assigned to both groups. The laboratory-of-origin was a highly significant factor explaining deletion frequencies in PC1 (Kruskal-Wallis test $p < 10^{-8}$). Further, 735 (9%) of 8106 deletions identified in at least one genome showed differences in frequency among laboratories (Benjamini-Hochberg corrected Kruskal-Wallis test $p < 0.05$). Also notably, Group 2 genomes had substantially higher frequencies of heterozygous deletions: among 735 deletions identified as significantly different among the 71 ancient genomes, the mean within Group 1 was ∼4, while that for Group 2 was ∼28.

We considered it unlikely that the observed dual grouping could have a biological reason and sought for technical explanations. To investigate this, we first sorted the 11,390 deletions into $k = 8$, as well as $k = 16$ clusters using k-means clustering, based on frequency similarities across the 71 genomes (Supplemental Fig. 4, 5, 6A). We found that in some, albeit not in all deletion clusters, the Group 1 and 2 genomes visibly differ in their mean frequencies (Supplemental Fig. 4, 5). In total, 4 out of $k = 8$ clusters (comprising 14% of the total number of deletions) showed clear differences between Group 1 and Group 2 (effect size, Cohen's d >0.8). We obtained similar results using for $k = 16$ (5 out of 16 clusters, comprising 11% deletions). We hypothesized that deletion loci in these 4 outlier clusters may have low mappability, i.e., the unique mapping potential of reads at a locus (Supplemental Fig. 7; Methods). Indeed, we found that the same clusters showing the strongest Group 1 vs. Group 2 differences also show the lowest average mappability (on average 16% and 24% lower mappability than rest of the clusters for $k = 8$ and for $k = 16$, respectively). We further reasoned that differential filtering of BAM data before publishing might explain different

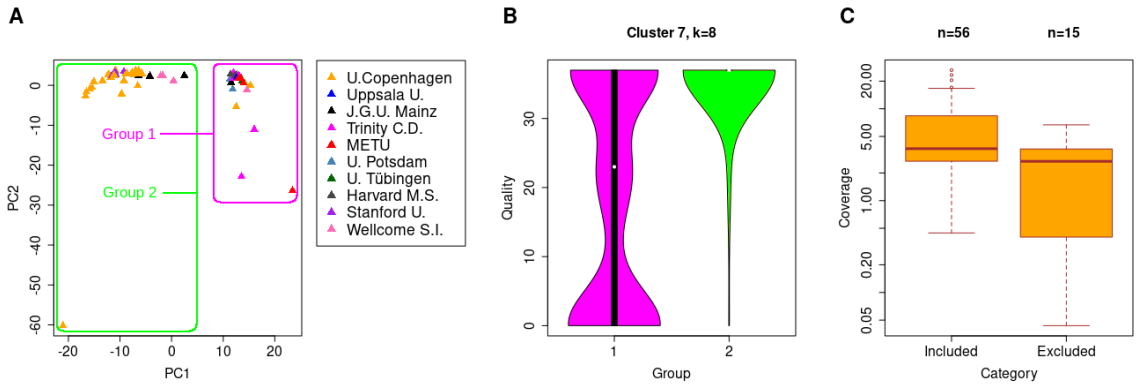

Figure 4: (A) A principal components analysis of deletion frequencies across the 71 genomes. There is significant distinction between the two groups observed in PC1 that explains 18% of the total variation (Wilcoxon rank sum test on PC1 values $p < 10^{-16}$). (B) Violin plot of MAPQ scores of 514 deletions belonging to outlier cluster 7 ($k = 8$), for Group 1 ($n = 30$) and Group 2 ($n = 41$) genomes. The two distributions are significantly different (Wilcoxon rank sum test, $p < 10^{-15}$). The same pattern is observed for the 2 other outlier clusters, which also happen to have low mappability, but not the remaining 5 clusters. (C) Boxplots representing the coverages of the included and excluded genomes in the refined dataset. The excluded genomes have significantly lower coverages on average (Wilcoxon rank sum test, $p = 0.01$).

deletion estimates at these loci between Group 1 vs. Group 2 genomes. Studying aligned read mapping quality (MAPQ) scores for reads mapping to each deletion in all 71 genomes, we discovered that Group 1 genomes had lower MAPQ scores than Group 2 genomes for deletions in the same 4 outlier clusters (Figure 4B; Supplemental Fig. 8, 9). This suggests that the observed laboratory-of-origin effect on deletion frequencies derives from differences among laboratories in BAM filtering before data submission: Group 1 data were not filtered for MAPQ, and consequently were genotyped accurately at low mappability deletions; hence their lower MAPQ scores at deletions in outlier clusters. In contrast, Group 2 BAM files were filtered for MAPQ, and at these loci, which lacked reads, CONGA genotyped artificial deletions.

If this assessment is true, the laboratory-of-origin effect could be removed by increasing the mappability score threshold per locus and filtering all BAM files for higher MAPQ. We thus increased the average mappability score threshold to ≥ 0.9 per deletion, thus discarding 12% of 11,390 deletions. We also filtered reads for MAPQ ≥ 30. Using the 10,018 deletions that remained after this double filtering, we repeated the principle components analysis (Supplemental Fig. 6B) and replotted the heat map (Supplemental Fig. 3B). We now observed no laboratory-based grouping on the new PC1 (Kruskal-Wallis test $p = 0.14$), and no difference in PC1 values between the previously identified Group 1 vs Group 2 genomes (Wilcoxon rank sum test $p = 0.61$), suggesting that mappability and MAPQ filters effectively removed the laboratory-of-origin effect. We next tested each of the 10,018 deletions for a laboratory-of-origin effect on their frequencies; this identified only 130 (1.3%) showing a significant effect (Benjamini-Hochberg corrected Kruskal-Wallis test $p < 0.05$). We note that some of these cases could represent authentic population differences in deletion frequency, because data from different laboratories vary in their population origins.

To further ensure that the data is not influenced of major technical artifacts, we repeated the clustering analysis with the 10,018 deletions that remained after filtering. This revealed that some singular ancient genomes behave as outliers in their deletion frequencies (Supplemental Fig. 4, 5). We could visually identify 15 such genomes out of the 71 total (Supplemental Table S2). A number of attributes could explain outlier behaviour. First, the coverage of the 15 outlier genomes was lower compared to the remaining 56 (Wilcoxon rank sum test, $p = 0.01$; Figure 6C). For instance, all three genomes with $<0.1\times$ coverage in our dataset (ne4, ko2, and DA379) were among the outliers. Second, the 15 outlier genomes had on average shorter read length compared to the rest (median = 52.7 vs 65.8, Wilcoxon rank sum test, $p < 0.001$). One of these was the Iceman, with unusually short (50 bps) reads. Other characteristics could also play a role. Among the most extreme outliers was Bon002, which was the only sample produced using whole genome hybridization capture with MyBaits probes (Kılınç *et al.*, 2016), suggesting that the capture procedure may distort coverage. Consequently we decided to remove these 15 genomes from further analysis.

2.5 A comparison of deletion and SNP diversity across 56 ancient genomes

The above genotyping and filtering steps resulted in a refined set of 10,018 autosomal deletions ascertained in present-day Africa and genotyped in 56 ancient Eurasian genomes, with 328-649 deletions (median = 411) detected in heterozygous or homozygous state per genome, and 28% detected in at least one genome. We used this dataset to test two hypotheses on deletion frequencies: (i) that deletion diversity patterns will broadly parallel SNP diversity patterns as previously reported for modern-day CNV datasets (Conrad and Hurles, 2007; Levy-Sakin *et al.*, 2019; Almarri *et al.*, 2020), and (ii) that deletion frequencies will reflect some degree of negative selection, which may be expected as deletions can bear harmful due to various effects including expression level changes, exon loss, or frame-shifting.

To investigate the first hypothesis, we compared pairwise genetic distances among the 56 individuals (Figure 5A) calculated using either SNPs or deletion genotypes. For this, we collected 5,991,735 autosomal SNPs ascertained in African Yoruba individuals in the 1000 Genomes Dataset and genotyped our 56 ancient genomes at these loci (Methods). We then calculated pairwise outgroup-f3 statistics, a measure of shared genetic drift between a pair of genomes relative to an outgroup population (Patterson *et al.*, 2012). Outgroup-f3 values were calculated for all pairs of ancient genomes using either SNPs or deletions, and these were then used these to calculate SNP -or deletion- based genetic distances between each pair (1-f3). We observed strong positive correlation between the two resulting distance matrices (Spearman $r = 0.71$, Mantel test $p = 0.001$) (Figure 5B). We further investigated relationships among ancient individuals by summarizing these distances using multidimensional scaling (MDS). This revealed highly similar patterns in the SNP -and deletion- based MDS plots, with clear clustering among west and east Eurasian genomes (Figure 5C, D). These results support the notion that deletion diversity patterns are at least partly shaped by demographic history and shared genetic drift.

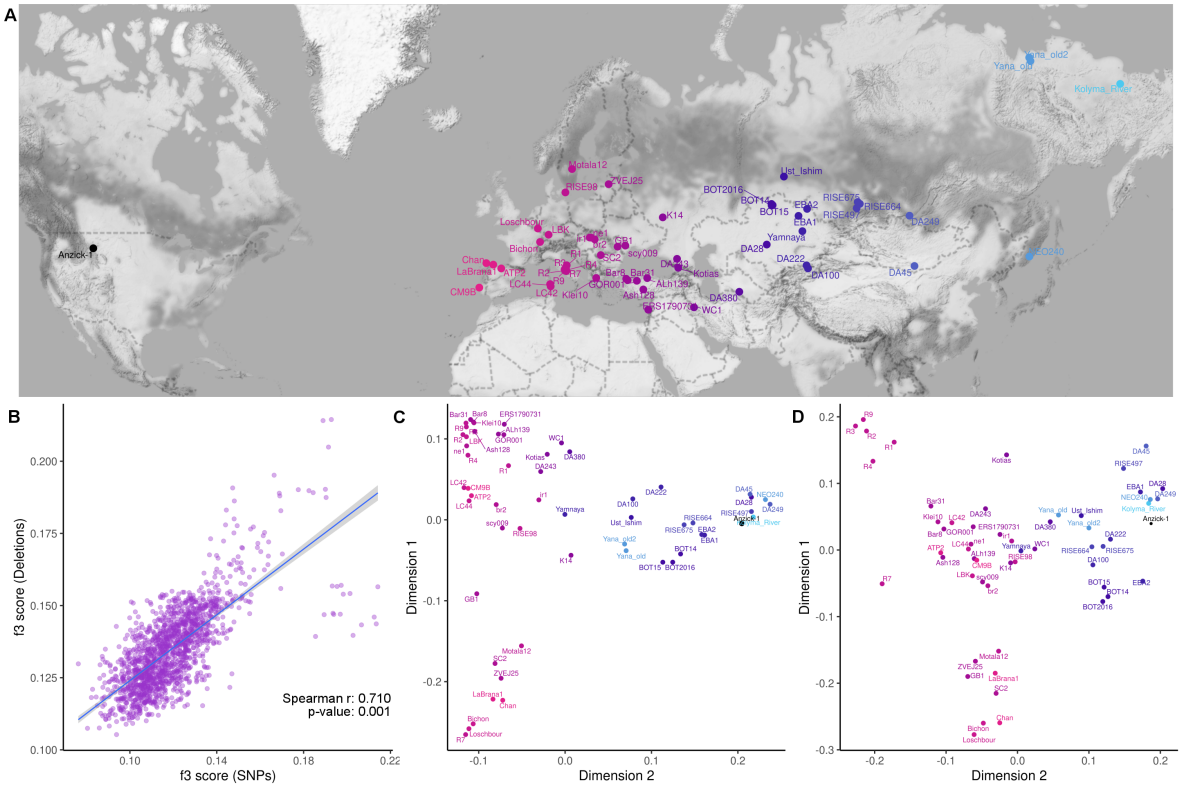

Figure 5: (A) Geographic locations of the 56 ancient individuals. (B) Comparison of genetic distances calculated using SNPs and deletions. We calculated the Spearman correlation coefficient between two matrices and then calculated Mantel test p-value using the "mantel" function in R package "vegan" (v2.5-7). (C) and (D) represent multidimensional scaling plots that summarize outgroup-f3 statistics calculated across all pairs among the 56 ancient individuals using SNPs and deletions, respectively.

2.6 Negative selection on deletion variants

We next studied the site-frequency-spectrum (SFS) of autosomal deletions across these 56 genomes in comparison with the SNP SFS. For this, we first polarised deletions as well as SNPs using chimpanzee and bonobo genomes to represent the ancestral state, thus determining 9,579 deletions and 4,907,535 SNPs derived in the human lineage (Methods). Second, 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 a stark excess of singletons among deletions, which is consistent with stronger negative selection on the latter (Figure 6A).

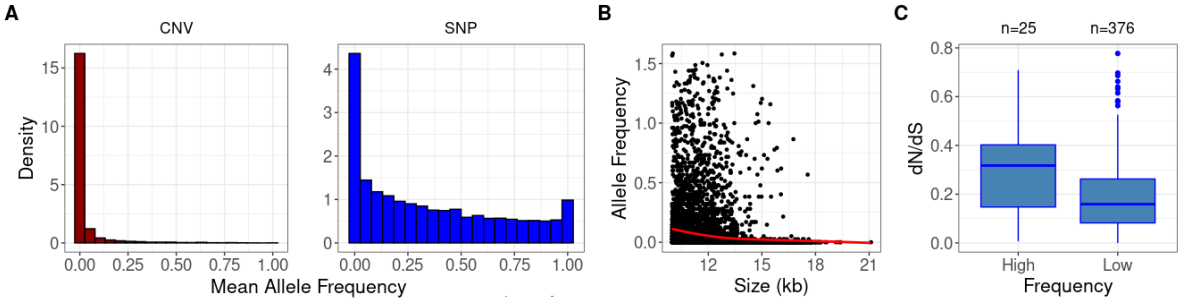

Figure 6: (A) The site-frequency-spectra of derived deletion alleles (on the left) and derived SNP alleles (on the right). The x-axes show mean allele frequency for each locus calculated using only those genomes where a locus has been observed (e.g. an allele observed in 10 out of 40 genomes will be 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 $r = -0.35$, $p < 10^{-16}$). Both axes were log_2-scaled. (C) Boxplots representing the dN/dS values of high ($n = 25$) vs. low ($n = 376$) frequency deletions.

If deletions are indeed under strong negative selection, as indicated by their SFS, we may expect longer deletions, as well as deletions containing evolutionary conserved genes, to be segregating at lower frequencies. To test the first idea, we compared deletion allele frequencies with their length. As hypothesized, deletion size and its frequency were negatively correlated across the 56 genomes (Spearman correlation $r = -0.35$, $p < 10^{-16}$) (Figure 6B).

To test the second hypothesis, we first determined annotated genes overlapping with our deletion dataset and found 461 Ensembl (v75) human genes. Overall, 5% of the 9579 derived deletions overlapped with at least one gene, and among those deletions, each overlapped with an average of 1.33 genes. We then collected mouse-human dN/dS ratios (Methods) for these genes (n = 401, 0-0.77, median = 0.16, mean = 0.20), which is an inverse measure of protein sequence conservation. We tested for a difference between lower frequency (below average) and higher frequency (above average) deletions in terms of dN/dS values of the genes they overlapped with. We found that deletions with lower allele frequency also had lower dN/dS values (Mann-Whitney U test, two-sided $p = 0.002$; Figure 6C), with an effect size of Cohen's $d = 0.68$. This suggests that deletions disrupting highly conserved genes tend to be rarer, again consistent with the action of negative selective.

2.7 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 ∼12 hours in total with as low as 2.2 GB of peak-memory consumption. This is ∼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× coverage (Sudmant *et al.*, 2015). The analysis took just slightly longer, ∼15 minutes average per genome, with similar memory usage.

We also compare the time and memory requirements of CONGA, FREEC and CNVnator in Table 2. In order to benchmark these tools, we used a 5× simulated genome (the same genome with medium sized CNVs used in the simulation experiments described above) with the same computing resources[1]. CONGA was 2× slower than the other two software in time performance when run in default mode, i.e. when genotyping duplications and deletions together. The reason for this is CONGA's use of its own small-scale read mapper for split-reads, which creates the bottleneck for time and memory usage. However, when only genotyping deletions, split-read information is not utilized; in this case CONGA's running time reduces to $\sim 1/3$ and memory usage to $\sim 1/7$ of the other two algorithms.

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

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× Yamnaya genome (with coverages between 23× and 0.07×).

Table 2: Time and Memory Consumption

Tools	Time (h:m)	Peak Memory Usage (GB)
CNVnator	0:32	12
FREEC	0:39	7.1
CONGA (deletions only)	0:12	1.2
CONGA (deletions and duplications)	1:19	10

Time and memory consumption of each algorithm for a simulated genome of 5× depth of coverage. "Time" refers to wall clock time and "Peak memory usage" is the maximum resident set size.

3 Discussion

Modern human genome sequencing experiments today typically reach coverages >20× 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×) among ancient genome datasets.

3.1 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× 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 true positive as well as true negative rates, especially at coverages <1×. This is unsurprising, as these latter tools were developed for the purpose of discovering novel CNVs in relatively high coverage genome data. In contrast, CONGA has been specifically developed for analyzing low coverage genomes by genotyping CNVs at previously ascertained loci.

CONGA further achieved acceptable accuracy (∼50% TPR and <30% FDR) in deletion copy number estimates in genomes of 0.5× 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 weaker, as we discuss below.

Nevertheless, the relatively high accuracy at 0.5× coverage suggests that the method could be used to genotype deletions across a considerable fraction of published shotgun sequenced ancient genomes. Beyond aDNA, CONGA will also be 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)).

3.2 Caveats in duplication genotyping

In simulated genome experiments, CONGA's performance in genotyping duplications was similar to that in deletions. Here, read-depth information provided comparable power as in genotyping deletions; furthermore, paired-end information from split reads also yielded results with acceptable accuracy (c.50%), although only at 5× coverage (Supplemental Table S1.E). 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× ancient BAM file, Yamnaya. Our results suggest the following reasons behind this discrepancy: (i) The Yamnaya genome was available as a BAM file, and we presume it was subject to pre-publication processing (at the alignment or quality filtering steps) that removed excess reads at duplicated loci. (ii) Consequently, 97% of duplications CONGA genotyped in the original (23×) BAM file were called using paired-end information, instead of read-depth information. (iii) Because paired-end information is more severely affected than read-depth information with decreasing coverage (as the former uses information from only reads that overlap breakpoints, while the latter uses information from across the length of the CNV), genotyping duplications in this BAM files became infeasible at <5× coverage.

The majority of shotgun ancient genomes in public databases are only published as BAM files. The motivation is to avoid publishing environmental DNA reads, which constitute a large fraction of shotgun sequencing experiments.

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 (see below).

3.3 Caveats in deletion genotyping

Applying CONGA to genotype deletions on 71 real ancient shotgun genomes from North Eurasia and the Americas using a dataset of modern-day African deletions yielded further critical observations. First, we found that genotyping deletions on published BAM files can also suffer from artifacts caused by pre-filtering of published reads based on mapping quality, a practice which apparently varies among research groups. Pre-filtered BAM files contain a paucity of reads originating from loci with relatively low mappability, creating artificial deletion signals. Nevertheless, we find that this issue can be overcome by increasing the average mappability threshold for deletion loci in the reference set and by applying higher mapping quality filters to all BAM files - a solution not available for genotyping duplications.

Another observation was that 15 of the 71 genomes analysed displayed unexpectedly high deletion frequencies at loci that are rarely or never genotyped in the rest of the 71 individuals. Given the close evolutionary relationship among Eurasian human populations, the majority of these signals are unlikely to be authentic, but rather originate from experimental artifacts and/or variability of DNA preservation among samples. Indeed, we could notice particularities in experimental protocols for some of these genomes, such as lower coverage, shorter read lengths, or the application of whole-genome hybridization capture.

These observations on real genomes indicate that $0.1\times$ coverage may be the lower threshold for deletion genotyping of >1 Kbps events, which is also consistent with the simulation results. Whole-genome hybridization capture may also be incompatible with reliable CNV detection. That said, we still lack clear explanations for outlier deletion frequency patterns for some of these 15 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 DNA extraction, library preparation, sequencing or data filtering.

3.4 Community recommendations for improving CNV analyses in ancient genomes

We mark the urgent need for changes in practice in producing and publishing ancient genomes to allow reliable study of CNVs in addition to 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 data. We also note that whole-genome capture data may be incompatible with CNV discovery.
- Researchers should strive to publish their data either as FASTQ files, or as BAM files aligned with the most relaxed parameters and without applying any filtering. Our results show that this is absolutely necessary to allow duplication genotyping at low coverage and to avoid biases in deletion genotyping on BAM files.
- 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.

We further recommend the following for comparative CNV analyses in ancient genomes:

- Studying and –if necessary– normalizing mapping quality distributions across published BAM files is crucial to avoid biases.
- Given the unavoidable heterogeneity of ancient genome data sources and the sensitivity of CNV detection and genotyping tools to such heterogeneity, researchers should pay utmost care to identify and filter out abnormal patterns of CNV frequencies in joint datasets.

3.5 Studying drift and selection in past populations

Beyond these technical aspects, our analysis of >1 Kbps deletions genotyped in 56 ancient genomes revealed how this CNV species has been evolving under genetic drift and negative selection, simultaneously. First, we found that genetic distances based on deletion frequencies were overall strongly correlated with those of SNP frequencies, and that they reflected spatial distances among populations. This is consistent with the observation that CNV frequencies in modern-day human populations at least partly reflect demographic processes, such as geographical isolation, drift and admixture (Conrad and Hurles, 2007; Levy-Sakin *et al.*, 2019; Almarri *et al.*, 2020).

Second, we observed that the deletion frequency spectrum in ancient Eurasians was considerably steeper than that of SNPs. While a notable proportion of derived SNPs polymorphic in Africans were fixed in our ancient Eurasian sample, we found no such fixed deletion event. We also found that large deletions, as well as deletions disrupting highly conserved genes, segregated at significantly lower frequencies compared to the rest of the deletion set. These results indicate that deletions are on average less evolutionarily neutral than SNPs and evolve under negative selection pressure. Intriguingly, these deletions have apparently continued to segregate across continents for possibly more than hundred thousand years, perhaps aided by bottleneck events. CONGA thus opens up a new avenue of research into deleterious CNV mutation loads in natural populations, which can now be studied using the rapidly growing number of low coverage WGS experiments.

4 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. 10, 11). The first reason is that the majority of shotgun ancient genomes are produced at low coverage (typically $<1\times$), which limits 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.

4.1 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.

Alternative approaches could also be considered within the same framework. One is using the negative binomial to model the read-depth (Miller *et al.*, 2011). However, the Poisson is simple and a most commonly used tool, and our simulations (described in the following) empirically demonstrate its effectiveness. Another approach could be calculating the likelihood of >2 copies of duplication events, such as multicopy genes (Sudmant *et al.*, 2010). These alternatives may be considered in future work.

4.2 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 7. We therefore designed CONGA to include both paired-end and single-end reads as input and evaluate the 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 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.

4.2.1 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).

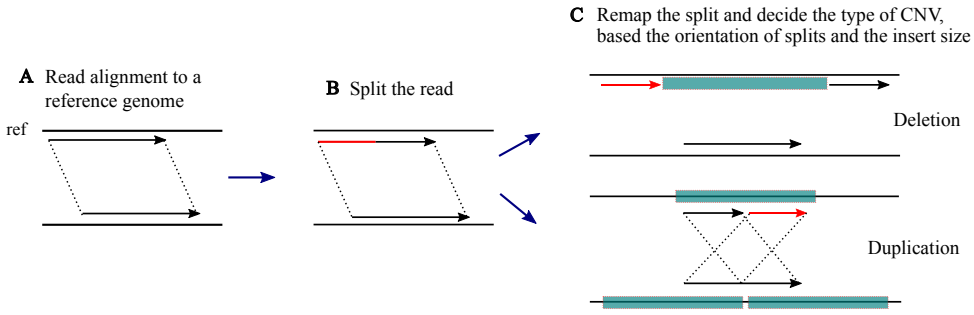

Figure 7: The figure shows our split read approach to emulate paired-end using single-end reads. (A) We are given short-read Illumina mappings in a BAM file as input. The top horizontal line represents the reference genome sequence ("ref"), while the lower line represents the donor genome. (B) We split the read from the middle, keeping the initial mapping as one element and the other subsequence (split segment) as the second element of a pair. (C) 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.

Based on the distance between the reads (insert-size) and orientation, we then evaluate the type of putative CNV. As Figure 7C 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).

4.3 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), available at http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/. Using this data, for each CNV event (S_i), we calculated the average mappability value within its breakpoints. We used either of two different minimum average mappability thresholds for the CNV events we analyzed: 0.5 (relaxed) or 0.9 (strict). Our deletion frequency analysis results show that the strict filter is more reliable especially when analyzing data sets of heterogeneous origin (see Results).

4.4 Simulation and down-sampling experiments

4.4.1 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. 1.

To evaluate specificity in addition to 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.*, 2015). 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, 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.

4.4.2 Simulating ancient genome read data

We then 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. As Gargammel is able to generate aDNA fragments with size distribution given as input, we used a subset of (Fu *et al.*, 2014), which is its 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 (calculated using 1× coverage data, but others 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).

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

We ran CNVnator (Abyzov *et al.*, 2011) and FREEC (Boeva *et al.*, 2012) on the simulated genomes with parameters described in the Supplemental Material 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.

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). Note that the number of True CNVs are: 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.

4.4.4 Down-sampling experiment with real ancient genomes

We used three relatively high coverage ($\sim$23.3×, $\sim$13.1× and $\sim$9.6× 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× for Yamnaya; 9-0.05× for Saqqaq; 7-0.03× for Mota. We note that this down-sampling procedure does not produce the exact targeted depths, which is the reason why we have 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).

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

We ran CONGA with a range of parameter values for the C-score [0.1-10] 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 an 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× for >1 Kbps deletions, and much higher F-scores at higher coverages (Supplemental Fig. 12).

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 ≥1× (Supplemental Table S1.E; Supplemental Fig. 13). 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. We therefore designed CONGA not to evaluate events <1 Kbps under default parameters. This can be modified by the user if needed.

4.5 Analysis of real ancient genomes

4.5.1 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×-26×, median = 3.45×) and that included both UDG-treated and non-UDG-treated libraries.

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 and those of size <35 bps. For preparing the "refined data set", reads with <30 mapping quality (MAPQ) were additionally removed.

4.5.2 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.*, 2015). 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 large deletions from 661 AFR genomes.

4.5.3 Deletion genotyping in ancient genomes: the raw data set

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. We call this the "raw data set". In total 8,106 deletions were identified in at least one genome.

4.6 Analyzing the raw data set of ancient deletions

4.6.1 Cluster analyses of ancient genomes

We performed a principal components analysis (PCA) on the deletion copy number data set where we clustered the 71 ancient genomes (Figure 4A). PC1 and PC2 values were computed using the R "stats" package "prcomp" function

using the default parameters (R Core Team, 2020). We also generated a heatmap summarizing deletion copy numbers using the R "gplots" package "heatmap.2" function (Warnes *et al.*, 2020) (Supplemental Fig. 3A). These analyses revealed visible differences in deletion frequency among samples reflecting the laboratory-of-origin, and the grouping of genomes into two main groups, that we called Group 1 and Group 2. We compared PC1 values among the 10 laboratories, and between Group 1 vs. Group 2, using the Wilcoxon rank-sum test (R "stats" package "wilcox.test" function) and the Kruskal-Wallis test (R "stats" package "kruskal.test" function), respectively (R Core Team, 2020). We further tested each of the 8106 deletions identified in at least one genome for differences in frequency among the 10 laboratories using the Kruskal-Wallis test, and corrected for multiple testing with the Benjamini-Hochberg correction (R "stats" package "p.adjust" function with the "BH" parameter) (R Core Team, 2020).

4.6.2 Cluster analyses of deletion events

We conducted k-means clustering analysis of deletions in the "raw data set" by using the "kmeans" function offered by the R "stats" package (R Core Team, 2020), and clustering deletions based on deletion copy numbers across the 71 genomes. We set the seed 123 for reproducibility. Analysis of within-group sum of squares did not reveal an optimal k (Supplemental Fig. 6A), and therefore used both $k = 8$ and $k = 16$. After clustering, we plotted mean copy numbers of the individuals in our sample set for each cluster using the R "ggplot2" package "ggplot" function (Wickham, 2016) (Supplemental Fig. 4, 5).

We observed that Group 1 and Group 2 were different in terms of their mean copy numbers as expected. We conducted the Wilcoxon rank-sum test on the mean deletion copy numbers of the two groups for each cluster (R "stats" package "wilcox.test" function) (R Core Team, 2020). We also calculated effect size in mean deletion copy numbers between Group 1 vs. Group 2 using a web-based effect size calculator (Becker, 2000).

4.6.3 Mappability and mapping quality analysis of deletions clusters

For deletion events in each k-means cluster, we compiled (i) average mappability values (see above) per deletion, and (ii) mapping quality (MAPQ) values of reads mapping to those locations, separately for Group 1 and Group 2 genomes. We compared these using violin plots of average mappability per deletion for each cluster ("geom_violin" function in the R "ggplot2" package)(Supplemental Fig. 7, 8, 13) (Wickham, 2016). We plotted the mean difference in MAPQ values between Group 1 and 2 versus the total difference in copy number between Group 1 and 2 using the "ggplot" function (R "ggplot2" package) (Supplemental Fig. 14) (Wickham, 2016).

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

4.7.1 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 5,987,516 autosomal bi-allelic SNPs with a minor allele frequency of $\geq 10\%$ in 108 African Yoruba genomes in Phase 3 of the 1000 Genomes Project (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) [REF], 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 56 ancient individuals and the above-described list of 5,987,516 SNPs, 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).

4.7.2 Ancestral state determination

To estimate the site frequency spectrum (SFS) for derived alleles, we polarized both deletion and SNP alleles for being ancestral or derived in the human lineage. For this, 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 4,907,535 SNP positions with derived allele information.

4.7.3 Creating the refined deletion data set

We filtered deletions in the raw data set for high mappability (≥ 0.9 average mappability), which left us with 10,018 deletions. We further removed 15 genomes identified as outliers in the k-means clustering analyses. We then selected reads with $>$30 mapping quality (MAPQ) in all BAM files, and we genotyped the 10,018 AFR deletions in the remaining 56 genomes. We call this the "refined data set". In addition, we created a subset of this data set than only includes the 9,579 deletions derived in the human lineage (see above). 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. 15) (R Core Team, 2020). We also plotted relative deletion (homozygous or heterozygous) frequencies of 9579 deletions for each individual in our refined data set using R's "graphics" package matplot function (R Core Team, 2020).

4.7.4 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% chance). 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 correlation between the deletion size in logarithmic scale and the frequency using R's "stats" package cor.test with method parameter set to "s" indicating Spearman's correlation (R Core Team, 2020).

4.7.5 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) using the BiomaRt tool. 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), which yielded $n = 401$ genes. We divided the deletions in our data set into two groups by the deletion allele frequency per gene: high versus low frequency. We calculated the mean frequency among the 401 genes and set this value as a threshold for high and low frequency. The frequencies above the threshold were considered high and vice versa. We plotted the dN/dS ratios of the groups defined above using the R package "ggplot2" and the "geom_boxplot" function (Wickham, 2016).

4.8 Genetic distance analyses using deletions and SNPs

Here our goal was to calculate overall genetic distances among the 56 ancient genomes using deletion allele frequencies and using SNPs, and further to compare the distances. We calculated distances using the commonly used outgroup-f3 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-f3 values were calculated for each pair of 56 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-f3. Distances were then summarized using multidimensional scaling (MDS) with the "cmdscale" function of R (R Core Team, 2020).

We further performed the Mantel test to compare the f3-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).

Data access

CONGA is implemented in C programming language and its source code is available under BSD 3-clause license at https://github.com/asylvz/CONGA. Simulated datasets and predictions of each tool can be accessed through Zenodo (10.5281/zenodo.5555990)

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.

Becker, L. A. (2000). Effect size (es).

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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*.

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.

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.* (2015). 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.

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.

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.