ChASM: A Statistically Rigorous Method for the Detection of Chromosomal Aneuploidies in Ancient DNA Studies

A Preprint

Adam B. Rohrlach
Department of Archaeogenetics
Max Planck Institute for Evolutionary Anthropology
Leipzig, Germany
adam_ben_rohrlachch@eva.mpg.de

Jonathan Tuke
School of Computer and Mathematical Sciences
University of Adelaide
Adelaide, Australia
simon.tuke@adelaide.edu.au

Kay Prüfer
Department of Archaeogenetics
Max Planck Institute for Evolutionary Anthropology
Leipzig, Germany
pruefer@eva.mpg.de

Wolfgang Haak
Department of Archaeogenetics
Max Planck Institute for Evolutionary Anthropology
Leipzig, Germany
wolfgang_haak@eva.mpg.de

Abstract

Chromosomal aneuploidies are the condition defined by the presence of an abnormal number of copies of the chromosomes that are present in the nuclei of the cells. Chromosomal aneuploidies represent the most common large-scale chromosomal abnormalities in human populations, and can affect autosomal chromosomes (e.g. Down syndrome) as well as the sex chromosomes (e.g. Klinefelter syndrome). The physical manifestations of different karyotypes can vary from critical, and resulting in miscarriage, such as in Edwards syndrome, to so mild that carriers are unaware that they carry an aneuploidy (e.g. Jacobs syndrome). Remains of individuals who carried aneuploidies from archaeological contexts present skeletal pathologies consistent with many other conditions, and accurate or confirmatory cases must rely on genetic diagnoses. Here we present *ChASM* (Chromosomal Aneuploidy Screening Methodology), a statistically rigorous Bayesian method for the detection of autosomal and sex chromosomal aneuploidies that leverages chromosome-wise read counts and takes into account differences in sequencing methodology, genetic coverage and condition rarity to produce posterior probability estimates for the screening of small and large databases of sequence data. To facilitate the ease of use, ChASM has been implemented in R as the package `RChASM`.

Keywords Chromosomal Aneuploidy · Ancient DNA · Trisomy

1 Introduction

Through the study of ancient DNA (aDNA), researchers have uncovered large- and small-scale events that have shaped our genomes[Haak et al., 2015, Mathieson et al., 2018]. An important driver of human evolution is disease and disorder, with studies investigating the history of pathogens, such as Yersinia pestis or Treponema pallidum, revealing our co-evolution[Barquera et al., 2025, Harbeck et al., 2013]. Recent studies have begun to explore genetic conditions in the human record, and how these more rare and personal diseases may have been viewed in past societies[Rohrlach et al., 2024, Maixner et al., 2021, Anastasiadou et al., 2024, Fuchs et al., 2021, Dorado-Fernández et al., 2023].

Chromosomal aneuploidies, the presence of an abnormal number of chromosomes in a cell, are one of the most common forms of genetic abnormality in human beings, often resulting in miscarriage[Crowe et al., 1997, Orr et al., 2015]. Among newborns, trisomy 21 (Down syndrome) is the most common aneuploidy, and trisomy 16 is the most common

cause of pregnancy loss[Griffin, 1996]. Aneuploidies can be manifest in one of three ways. Full aneuploidies, where missing or additional copies of entire chromosome(s) are present in all cells, partial aneuploidies, where only a portion of the extra chromosome(s) is present in all of the cells, and mosaic aneuploidies, where either the entire, or a portion of the extra chromosomes is present in *some* of the cells[Orr et al., 2015].

Excluding exceedingly rare cases of trisomy 22, the only aneuploidies that are not always lethal are trisomy 13 (Patau syndrome), trisomy 18 (Edwards syndrome) and trisomy 21 (Down syndrome)[Crowe et al., 1997]. Symptoms frequently observed in individuals with trisomies 13 and 18, *e.g.* facial clefts, cardiac and urinary complications and limb and nervous system defects, can be severe[Springett et al., 2015], and hence the modern 5-year survival rates for trisomy 13 (9.7%) and trisomy 18 (12.3%) would have likely been much lower before the advent of modern medicine[Costello et al., 2015]. Conversely, individuals today with Trisomy 21 have a rate of survival to adulthood of 95%[Arumugam et al., 2016]. However, trisomy 21 can lead to a wide range of developmental abnormalities, such as intellectual disabilities, cardiac defects, autoimmune disorders and recurrent infections[Carfì et al., 2014]. While no individual skeletal pathology can be used to diagnose any of these trisomies, genetic diagnoses may help to explain osteological observationsRohrlach et al. [2024].

Sex chromosomal aneuploidies often lead to less life-threatening symptoms, although developmental issues are common, and some disease risk factors can be increased. For example, individuals carrying karyotype XXX are at an increased risk of malformations of the reproductive system (5-16%)[Tartaglia et al., 2010], seizures (11-15%)[Tartaglia et al., 2010] and have been observed to suffer from poor dentition (23.9%)[Wigby et al., 2020]. Individuals carrying karyotype XYY have an increased risk of infertility (11.27-fold)[Berglund et al., 2020], Valgus (2527-fold)[Berglund et al., 2020] and a high observed frequency of dental issues (22%)[Bardsley et al., 2013]. The most common sex chromosomal aneuploidy, XXY or Klinefelter syndrome, has been shown to lead to high rates of infertility (91-99%)[Groth et al., 2013] and an increased risk of Mediastinal cancer (500-fold)[Groth et al., 2013], as well as learning and mental health difficulties[Turriff et al., 2017].

Due to the difficulty in diagnosing these conditions osteologically, very little is known about them in the archaeological record[Garcia-Heras, 2024]. Two recent studies from Anastasiadou *et al.*and Rohrlach *et al.* have uncovered cases of trisomies 18 and 21, as well as Klinefelter, Turner and Jacobs syndromes[Anastasiadou et al., 2024, Rohrlach et al., 2024]. Both studies, using related but different methods, suggest using read counts as the primary method for identifying such cases. Where Anastasiadou *et al.* suggest pre-defined cut offs, Rohrlach *et al.* instead opted for modeling read counts via a Bayesian methodology, leveraging a Dirichlet-multinomial distribution to account for differences in data generation and sample quality. Both methods have been shown to be able to identify cases of trisomies, and attempt to identify cases when contamination may be obscuring true signal, but differ in the way in which they report karypotype classifications. Specifically, where Anastasiadou *et al.* return only karyotype classifications, Rohrlach *et al.* method instead returns posterior probabilities for all karyotypes, and hence the highest-posterior classification, we also generate diagnostic plots and tests for contamination and abnormal sequencing.

Here we present a software package implementing a generalised approach to the Bayesian method we presented in Rohrlach *et al.* 2024. We generalise the method to include sex chromosomal aneuploidies, and derive statistics to classify contamination and sequence data that diverges too far from expectation. We use simulations to show the power of the method to reliably identify cases of aneuploidies, and to calculate quality control thresholds for best use recommendations. Finally, we compare the performance of our method to that of Anastasiadou *et al.* to benchmark our method against the only currently available method for aneuploidy detection designed for aDNA.

2 Algorithm

2.1 A distribution for read counts per autosomal chromosome

Consider the problem of describing the proportion of N_j total reads, mapping to a set of chromosomes A, where $|A| = n$, for individual j. For ease of notation, let the sex chromosomes X and Y be represented as chromosomes 23 and 24, respectively.

It is tempting to assume that there exists some probability vector $\mathbf{p}^{\boldsymbol{c}}$ that describes the multinomial distribution $\mathbf{N}_j \sim MN\left(N_j, \mathbf{p}^{\boldsymbol{c}}\right)$, where

$$\mathbf{N}_j = (N_{1j}, \cdots, N_{nj})$$

is the number of reads mapping to each chromosome, and $\boldsymbol{c}$ is an n-dimensional vector describing the aneuploidy. We define the elements of $\boldsymbol{c} = [c_i]$ to be the expected fold-change in reads mapping to the chromosome (or set) represented by element i compared to an "average non-carrier". For example, if we consider the set of all autosomal chromosomes ($n = 22$) to be $c_i = 1$, then a similar $\boldsymbol{c}$, but with $c_{21} = 1.5$, would represent a karyotype with an additional copy of chromosome 21, and hence of trisomy 21 (Down syndrome).

Unfortunately, sequencing protocols and sample quality are rarely so consistent, and hence the true distribution of $\boldsymbol{N}_j$ is over-dispersed. To model this we also assume that $\mathbf{p}^{\boldsymbol{c}}$ has a Dirichlet prior distribution of the form $\mathbf{p}^{\boldsymbol{c}} \sim Dirichlet\left(\boldsymbol{\alpha}\right)$. Since the Dirichlet distribution is a conjugate prior for the multinomial distribution, the posterior distribution for $\boldsymbol{N}_j$ is a Dirichlet-multinomial (DM) distribution of the form

$$\boldsymbol{N}_j \sim DM\left(N_j, \boldsymbol{\alpha}\right).$$

We now have, for

$$\alpha_0 = \sum_{i=1}^{n} \alpha_i,$$

that

$$E\left[N_{ij} \middle| \boldsymbol{\alpha}\right] = N_j \frac{\alpha_i}{\alpha_0}$$

and

$$\text{Var}\left(N_{ij} \middle| \boldsymbol{\alpha}\right) = N_j \frac{\alpha_i}{\alpha_0}\left(1 - \frac{\alpha_i}{\alpha_0}\right)\left(\frac{N + \alpha_0}{1 + \alpha_0}\right).$$

We note that we commonly use three chromosome sets:

1. The "Autosomal set", $\boldsymbol{A}_a = \{1, \cdots, 22\}$. Here we consider chromosomes one through to twenty-two, and do not consider the sex chromosomes. We use this set to look at three autosomal trisomies: trisomy 13 (Patau syndrome), trisomy 18 (Edwards syndrome) and trisomy 21 (Down syndrome), but to ignore genetic sex.
2. The "Sex Chromosomal set", $\boldsymbol{A}_s = \{A_a, 23, 24\}$. Here we consider the X and Y chromosomes, and merge all autosomal chromosomes into one set.
3. The "Autosomal Z-score set", $\boldsymbol{A}_z = A_a \setminus \{13, 18, 21\}$. We use this set to calculate Z-scores for assigning samples to a classification of "too unlike the reference data". Here we again consider the autosomal chromosomes, but this time ignore chromosomes 13, 18 and 21 in order to not classify carriers of Patau, Edwards or Down syndrome as otherwise dissimilar, and again ignore genetic sex.

2.2 Estimating the common reference distribution

sec:estimating) We now estimate a "common" karyotype (i.e., $c_i = 1$ for $i = 1, \cdots, n$) from a quality-filtered subset of the reference data (see Supplementary Section S3). Of importance, we filter for (a) a lower- and upper-bound cut off of mapped reads and (b) statistical outliers when forming clusters. We then use maximum likelihood to estimate the Dirichlet parameters using the `dirichlet.mle()` function from the *sirt* R-package[Robitzsch and Robitzsch, 2017].

From these concepts we estimate the Dirichlet distribution for the common autosomal reference, and the common references for genetically female (XX) and male (XY) individuals, denoted $\boldsymbol{\alpha}^{\boldsymbol{c}_a^1}$, $\boldsymbol{\alpha}^{\boldsymbol{c}^{XX}}$ and $\boldsymbol{\alpha}^{\boldsymbol{c}^{XY}}$, respectively. We use $\boldsymbol{\alpha}^{\boldsymbol{c}^{XY}}$ as the basis for modified sex chromosomal aneuploidies, but retain the parameter for reads mapping to the Y chromosome (denoted ϵ^+) to account for errors in read mapping caused by, among other factors, sequencing error.

2.3 Adjusting $\boldsymbol{\alpha}$ for new karyotypes

Unfortunately, especially in the cases of rare karyotypes, we likely cannot sample enough individuals with "uncommon" karyotypes in order to estimate the associated $\boldsymbol{\alpha}$. However, we can adjust the $\boldsymbol{\alpha}$ for the more common karyotypes such that the expectation matches reality, and such that the variance is relatively unchanged.

Consider a karyotype where chromosome t undergoes a fold-change of k, with all others remaining the same. We then set a karyotype $\boldsymbol{c}^*$ where $c_i^* = 1$ for $i \neq t$, but $c_t^* = k$, and the element-wise multiplication yields that

$$\alpha_i^{\boldsymbol{c}^*} = \begin{cases} \alpha_i, & \text{if } i \neq t, \\ k\alpha_t, & \text{if } i = t. \end{cases}$$

For

$$\alpha_0^{\boldsymbol{c}^*} = \sum_{i=1}^{n} \alpha_i^{\boldsymbol{c}^*},$$

we now have that

$$
\begin{aligned}
\mathrm{E}\left[N_i \middle| \boldsymbol{\alpha}^{\boldsymbol{c}^*}\right] &= N_i \frac{\alpha_i^{\boldsymbol{c}^*}}{\alpha_0^{\boldsymbol{c}^*}} \\
&= \begin{cases} N_i \dfrac{\alpha_i}{\alpha_0^{\boldsymbol{c}^*}}, & \text{if } i \neq t, \\ N_t \dfrac{k\alpha_t}{\alpha_0^{\boldsymbol{c}^*}}, & \text{if } i = t, \end{cases} \\
&= \begin{cases} N_j \dfrac{\alpha_i}{\alpha_0} \dfrac{\alpha_0}{\alpha_0^{\boldsymbol{c}^*}}, & \text{if } i \neq t, \\ kN_t \dfrac{\alpha_t}{\alpha_0} \dfrac{\alpha_0}{\alpha_0^{\boldsymbol{c}^*}}, & \text{if } i = t, \end{cases} \\
&\approx \begin{cases} E\left[N_{ij} \middle| \boldsymbol{\alpha}\right], & \text{if } i \neq t, \\ kE\left[N_{ij} \middle| \boldsymbol{\alpha}\right], & \text{if } i = t, \end{cases}
\end{aligned}
$$

assuming that $\frac{\alpha_0}{\alpha_0^{\boldsymbol{c}^*}} \approx 1$. This slight adjustment in expected values reflects that more (or less) reads will map to chromosome t as there are no more (or less) copies.

Additionally, if $\frac{\alpha_0}{\alpha_0^{\boldsymbol{c}^*}} \approx 1$, then

$$
\mathrm{Var}\left(N_{ij} \middle| \boldsymbol{\alpha}^{\boldsymbol{c}^*}\right) \approx \mathrm{Var}\left(N_{ij} \middle| \boldsymbol{\alpha}\right).
$$

Note that for sex chromosomal aneuploidies, we adjust the parameter vector for genetically male individuals, $\boldsymbol{\alpha}^{\boldsymbol{c}^{XY}}$, but retain the information on the proportion of reads erroneously mapping to the Y chromosome for genetically female individuals.

2.4 Calculating posterior probabilities

Once $\boldsymbol{\alpha}$ has been defined, the posterior probability for an observed number of read counts per chromosome, can be calculated as

$$
P\left(\boldsymbol{N}_j \middle| \boldsymbol{\alpha}, k\right) = \frac{\Gamma\left(\alpha_0\right)\Gamma\left(N_j + 1\right)}{\Gamma\left(N_j + \alpha_0\right)} \prod_{k \in K} \frac{\Gamma\left(N_{jk} + \alpha_k\right)}{\Gamma\left(\alpha_k\right)\Gamma\left(N_{jk} + 1\right)}.
$$

Given the prior probability of each possible karyotype, the posterior probability for karyotype k is then

$$
P\left(k \middle| \boldsymbol{N}_j, \boldsymbol{\alpha}\right) = \frac{P\left(\boldsymbol{N}_j \middle| \boldsymbol{\alpha}, k\right) P(k)}{\sum_{s \in K} P\left(\boldsymbol{N}_j \middle| \boldsymbol{\alpha}, ks\right) P(s)}. \tag{1}
$$

We calculate the probabilities defined in equation 1 in the log-space using the `ddirmnom()` function from the *extraDistr* R-package[Wolodzko, 2020], and calculate the denominator we use the `logSumExp()` function from the *matrixStats* R-package[Bengtsson, 2025].

2.5 Identifying departure from the reference distribution

For individual j, assume that we have read counts for each of n chromosomes of the form

$$
\boldsymbol{N}_j = \left(N_{1j}, \cdots, N_{nj}\right),
$$

and let the total number of reads attributed to individual j be

$$
N_j = \sum_{i=1}^{n} N_{ij}.
$$

Assume also that $\boldsymbol{N}_j$ has a Dirichlet-Multinomial distribution $DM(N_j, \boldsymbol{\alpha})$, where

$$
\boldsymbol{\alpha} = \left(\alpha_1, \cdots, \alpha_n\right),
$$

with normalising constant

$$\alpha_0 = \sum_{i=1}^{n} \alpha_i,$$

that has been estimated from a reference data set (such as in **??**).

It is known that the expected value and variance of the N_{ij} are, respectively,

$$\begin{aligned} \mathrm{E}\left[N_{ij}\right] &= N_j \frac{\alpha_i}{\alpha_0} \\ &= \mu_{ij} \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}\left(N_{ij}\right) &= N_j \frac{\alpha_i}{\alpha_0}\left(1 - \frac{\alpha_i}{\alpha_0}\right)\left(\frac{N_j + \alpha_0}{1 + \alpha_0}\right) \\ &= \sigma_{ij}^2. \end{aligned}$$

Defining the per chromosome Z-score to be

$$Z_{ij} = \frac{N_{ij} - \mu_{ij}}{\sigma_{ij}},$$

clearly the expected value and variance of Z_{ij} are zero and one, respectively, and assuming that $N_{ij} >> 0$, then $Z_{ij} \dot{\sim} N(0, 1)$.

Next, define

$$\lambda_j = \sum_{i=1}^{n} Z_{ij}^2, \tag{2}$$

which result in

$$\lambda_j \dot{\sim} \chi_n^2,$$

and the number of chromosome-wise Z-scores that are considered "significant", denoted

$$F_j = \sum_{i=1}^{n} \delta_{|Z_{ij}|>2}.$$

An individual is considered an outlier that does not represent the baseline if the observed χ^2 statistic is significant, and if F_j is sufficiently large (we suggest $F_j > 2$).

Finally, we calculate the full χ_{22}^2 on all autosomal chromosomes, without chromosomes 13, 18 and 21, when calculating the χ_{19}^2 . We do this as, if an individual carries a (possible) trisomy on chromosomes 13 (Patau syndrome), chromosome 18 (Edwards syndrome) or chromosome 21 (Down syndrome), then we *expect* an abnormal read count on the associated chromosome.

2.6 Accounting for potential contamination

In some cases, when there is contamination between XX and XY individuals, a mixture of the two read counts can be observed. With low numbers of reads, the variation can overlap with the distribution of reads of a true XXY carrier. For methods that only look at the ratio of reads mapping to the X and Y chromosome, this can appear like the outcome of an XXY karyotype. Since we consider the proportion of reads mapping to the autosomes, we aim to distinguish between a true case of XXY and contamination (for a sufficient total number of reads N_j).

Consider the associated Dirichlet-multinomial distributions for the XX and the XY karyotype, denoted $\boldsymbol{\alpha}^{\boldsymbol{c}^{XX}}$ and $\boldsymbol{\alpha}^{\boldsymbol{c}^{XY}}$, respectively. We assume that there is some proportion of XX-associated data, denoted $\gamma \in [0, 1]$, and hence the remaining proportion of $1 - \gamma$ comes from XY-associated sequence data. We do not assume that there is a combination of XX or XY any other karyotype (*i.e.* XXY or X0), and consider the probability of a random mix of two rare sex chromosomal aneuploidies negligible.

It must be then that the number of observed reads follows a distribution of the form

$$\mathbf{N}_j \sim DM\left(N_j, \boldsymbol{\alpha}^{\boldsymbol{c}^{XX}\gamma\boldsymbol{c}^{XY}}\right),$$

where

$$\boldsymbol{\alpha}^{\boldsymbol{c}^{XX\gamma XY}} = \gamma\boldsymbol{\alpha}^{\boldsymbol{c}^{XX}} + (1-\gamma)\boldsymbol{\alpha}^{\boldsymbol{c}^{XY}} \tag{3}$$

We estimate γ by consider the number of reads mapping to the X and Y chromosomes, denoted $n_{j,x}$ and $n_{j,y}$, relative to the autosomes, and comparing this to the expected values,

$$E\left[\boldsymbol{N}_j|\boldsymbol{\alpha}^{\boldsymbol{c}^{XX}\gamma\boldsymbol{c}^{XY}}\right] = \gamma E\left[\boldsymbol{N}_j|\boldsymbol{\alpha}^{\boldsymbol{c}^{XX}\gamma}\right] + (1-\gamma)E\left[\boldsymbol{N}_j|\boldsymbol{\alpha}^{\boldsymbol{c}^{XY}\gamma}\right].$$

If we assume that or the reads mapping to the X and Y chromosomes are as expected, this yields that for $k \in \{X, Y\}$,

$$\begin{aligned}
&E\left[N_{jk}|\boldsymbol{\alpha}^{\boldsymbol{c}^{XX}\gamma\boldsymbol{c}^{XY}}\right] = n_{js} \\
\implies &\hat{\gamma}N_j\frac{\alpha_k^{\boldsymbol{c}^{XX}}}{\alpha_0^{\boldsymbol{c}^{XX}}} + (1-\hat{\gamma})N_j\frac{\alpha_k^{\boldsymbol{c}^{XY}}}{\alpha_0^{\boldsymbol{c}^{XY}}} = n_{js} \\
\implies &\hat{\gamma}N_j\left(\frac{\alpha_k^{\boldsymbol{c}^{XX}}}{\alpha_0^{\boldsymbol{c}^{XX}}} + \frac{\alpha_k^{\boldsymbol{c}^{XY}}}{\alpha_0^{\boldsymbol{c}^{XY}}}\right) = n_{js} - \frac{\alpha_k^{\boldsymbol{c}^{XY}}}{\alpha_0^{\boldsymbol{c}^{XY}}} \\
\implies &\hat{\gamma} = \frac{n_{js} - \frac{\alpha_k^{\boldsymbol{c}^{XY}}}{\alpha_0^{\boldsymbol{c}^{XY}}}}{N_j\left(\frac{\alpha_k^{\boldsymbol{c}^{XX}}}{\alpha_0^{\boldsymbol{c}^{XX}}} + \frac{\alpha_k^{\boldsymbol{c}^{XY}}}{\alpha_0^{\boldsymbol{c}^{XY}}}\right)}
\end{aligned}$$

Using this estimate of the mixing parameter, we can now calculate $\boldsymbol{\alpha}^{\boldsymbol{c}^{XX\gamma XY}}$ per equation 3.

We then calculate the posterior probability of contamination, by calculating

$$P\left(C|\boldsymbol{N}_j\right) = \frac{P\left(\boldsymbol{N}_j|C\right)P\left(C\right)}{\sum_{k^*}^{K^*}P\left(\boldsymbol{N}_j|k^*\right)},$$

where $k^* = \{XX, XY, XXY, C\}$. We restrict K^* to only include these karyotypes as these are the only possible karyotypes that could be confused as contamination.

3 Results

3.1 Assessing the performance of ChASM via simulation

To assess the performance of ChASM, we simulated 5×10^5 realisations for each autosomal and sex chromosomal karyotype, based on empirical data, resulting in 1.1×10^6 total realisations (see Supplementary Section 4). We used these simulations to test the performance of ChASM method to correctly identify chromosomal aneuploidies, for varying levels of coverage (total number of reads)

3.1.1 Simulating Sex Chromosomal Aneuploidies

ChASM achieves 97.35% overall accuracy when assigning autosomal aneuploidies. We observe that XYY is by far the least accurately assigned karyotype (98.59%, Figure 1)), owing to the fact that the Y chromosome is significantly smaller than the X chromosomes, and hence the overlap between the distribution of karyotypes XY and XYY for low read count totals is relatively large (Supplementary Figure 2). For this reason, we observe that the minimum number of total reads required to achieve 95% accuracy for all sex chromosomal aneuploidies is 60k (Figure 1), and we apply this as the recommended minimum number of reads for analyses using ChASM.

After applying this threshold, we see the misclassification rate for karyotype XYY drop to 1.41%, and overall ChASM achieves 99.76% accuracy, 99.71% specificity, 99.8% AUC_{roc} and a Cohen's κ of 0.996. Clearly, this threshold could be considered conservative for karyotypes other than XYY, and this decision can be made by researchers in context.

3.1.2 Simulating Autosomal Aneuploidies

ChASM achieves 95.16% accuracy when assigning autosomal aneuploidies. Critically, we find that no realisations generated without an aneuploidy were erroneously classified as any of the possible trisomies. However, we find that all

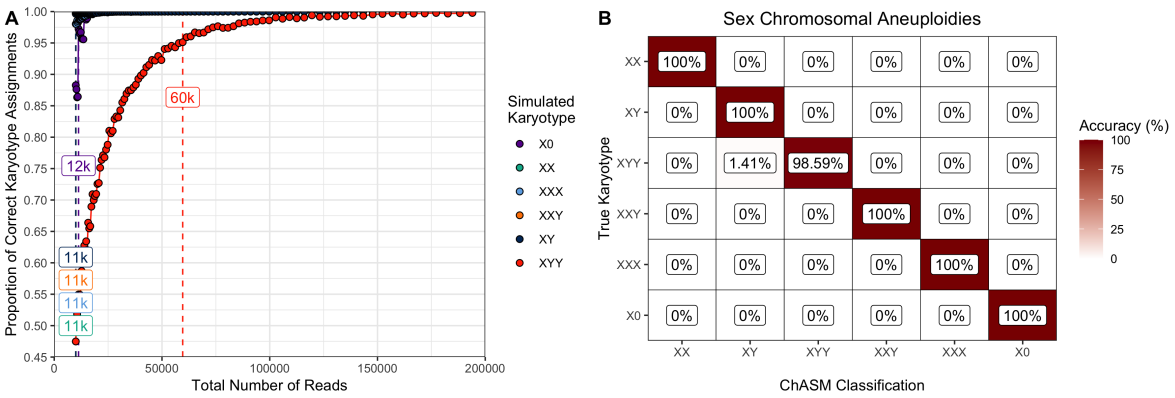

Figure 1: Results of the simulation study for sex chromosomal aneuploidies. (A) observed proportions of reads that are correctly classified for each karyotype (colour) with the minimum number of total reads for which at least 95% accuracy was achieved (minimum possible 11,000 reads), (B) a confusion matrix for classification accuracy for simulations with at least 60k reads.

three trisomies can be misclassified at rates of between 3.51% to 9.12%, for trisomy 18 and 13, respectively (Figure 2 A). However, when we apply the minimum cut off of at least 60,000 reads, then trisomies 13, 18 and 21 are correctly classified in 98.95%, 99.98% and 99.92% of simulations, respectively (Figure 2 B). Overall, when applying the filter, ChASM has 99.77% accuracy, 99.77% specificity, 99.8% AUC_{roc} and a Cohen's κ of 0.997.

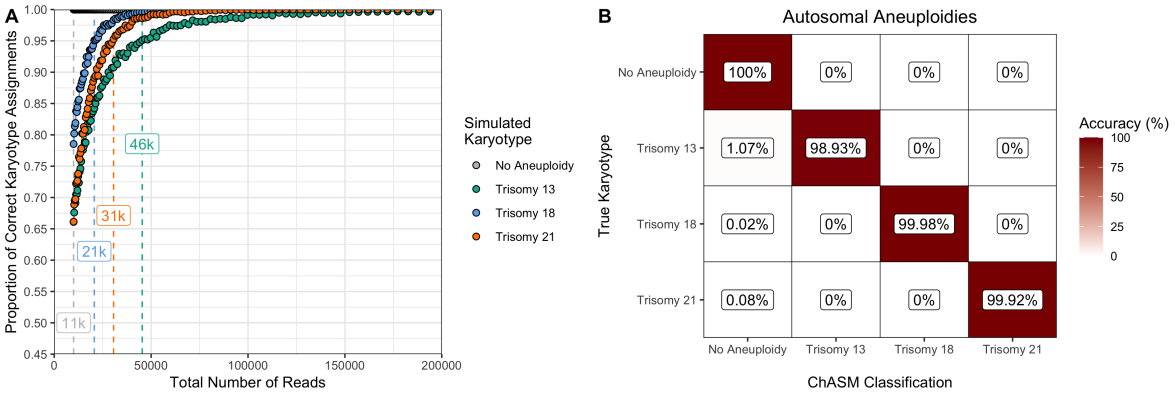

Figure 2: Results of the simulation study for autosomal aneuploidies. (A) observed proportions of reads that are correctly classified for each karyotype (colour) with the minimum number of total reads for which at least 95% accuracy was achieved (minimum possible 11,000 reads), (B) a confusion matrix for classification accuracy for simulations with at least 60,000 reads.

Overall, we find that ChASM reliably correctly identifies autosomal aneuploidies. Critically, the false positive rate is zero, meaning that researchers can be confident that non-common karyotype assignments are not due to model miss-specification.

3.2 Comparison with Karyo R_xR_y

The only other published method for assigning chromosomal aneuploidies to aDNA is Karyo R_xR_y from Anastasiadou *et al.*[Anastasiadou et al., 2024]. This method calculates the coverage on the X and Y chromosomes, normalised by the coverage on the autosomes, and then using coverage thresholds, identifies cases of aneuploidies. The method considers Klinefelter, Jacobs, XXX and Turner syndromes. This method also performs a similar calculation for chromosome 21 to identify trisomy 21. The thresholds for identifying aneuploidies are calculated specifically for shotgun data, but an option to change the thresholds for 1240k capture data is also included, although the method is not calibrated for any other existing capture methods. It is due to this pre-calibrated approach that Karyo R_xR_y can be run on a single individual/sample, where ChASM requires training data. Hence, we cannot compare to the samples from Anastasiadou

et al. as we have access to only the six positive cases, and no other other similarly produced samples from the same laboratory[Anastasiadou et al., 2024].

Another critical difference is that ChASM uses a distributional approach to modeling read mapping. This means that ChASM can (a) return posterior probabilities of chromosomal aneuploidies, (b) take into account the prior probabilities of aneuploidies via modern estimates of the rates of prevalence, and (c) take into account coverage using the posterior Dirichlet-multinomial distribution. Hence, ChASM returns not just the karyotype assignments, but also uncertainty around about these classifications, with informative diagnostic visualisations.

To compare the performance of ChASM and Karyo R_xR_y we performed two comparative analyses. First, on the data for which Karyo R_xR_y was calibrated to work best (shotgun and 1240k), and then on a capture assay for which Karyo R_xR_y was not calibrated (immuno-capture). To compare and quantify the performance of the two methods on shotgun and 1240k sequence data, we use published empirical data from the Bronze Age Iberian site of La Almoloya (ALM)[Villalba-Mouco et al., 2021], the Neolithic French site of Gurgy (GRG)[Rivollat et al., 2023], the Iron Age Thai site of Yappa Nhae (YPN)[Carlhoff et al., 2023], the Bronze Age Bulgarian site of Yunatsite (YUN)[Penske et al., 2023] and a collection of sites from the Irish Neolithic period[Cassidy et al., 2020]. To compare and quantify the performance of the two methods on immuno-capture data, we use the published empirical data from the Neolithic site of Gurgy[Rivollat et al., 2023]. We chose ALM, YPN and YUN as they have reported cases of sex chromosomal and autosomal aneuploidies. Unfortunately, we know of no such cases where aneuploidies have been identified for anything other than shotgun or 1240k capture data. However, using GRG we can still call the common genetic sexes XX and XY using both methods, and compare this to the genetic sexes assigned by Rivollat *et al.* in their study. The authors of these archaeogenetic studies report two cases of XXX syndrome (ALM062 and YPN020), one case of Klinefelter syndrome (CLL011) and two cases of trisomy 21 (YUN039 and PN07).

3.2.1 Shotgun and 1240k capture data

ChASM and Karyo R_xR_y correctly identify the case of trisomy 21, the case of XXY and both cases of XXX syndrome for both shotgun and 1240k capture data. Further, we see complete agreement between the calls of XX (n=81) and XY (n=78) between the two methods. From the diagnostic plots (Figure 3), we can see that the positions of ALM062, CLL011 and YPN020 are consistent with the reported sex chromosomal aneuploidies (Figure **??**) and that YUN039 and PN07 yield significantly more reads to chromosome 21, consistent with the diagnoses of trisomy 21 (Figure 3 B and C). Hence, we see that ChASM and Karyo R_xR_y both agree completely and correctly with the assignments of sex chromosomal and autosomal aneuploidies. Hence, we show that ChASM works equally well as Karyo R_xR_y on these calibrated sequence data types.

The agreement between the methods is expected as the informative statistics used by these methods are highly correlated, with correlation coefficients of 0.999 for both R_x and p_x and R_y and p_y in our empirical examples ($p \leq 2.2 \times 10^{-16}$) Hence, we expect broad agreement between the two methods for shotgun or 1240k sequence data for which careful calibration has been considered.

3.2.2 Immuno-capture data

The assignments of sex chromosomal karyotypes for the immuno-capture data for GRG via ChASM completely agree with the genetic sex assignments from the authors[Rivollat et al., 2023], who used the approach given by Mittnik *et al.*[Mittnik et al., 2016]. Conversely, due to the lack of calibration to immuno-capture data, Karyo R_xR_y only achieves an overall accuracy of 40.86%. This sharp decrease in accuracy is due to the fact that, while Karyo R_xR_y correctly identifies every XX individual, only one of the 56 XY individuals is identified as XY, and the remaining XY individuals are instead assigned to the "Contamination" class.

While the R_x values are quite similar for all three sequencing data types, the R_y values differ significantly, likely due to the desire to target informative sites on the Y chromosome in order to call Y haplogroups, and hence the increased number of sites targeted on the Y chromosome for 1240k, relative to the length of the chromosome (see Figure 4).

It should be noted that this problem was present whether the "–capture" flag was used or not used for Karyo R_XR_y. For new capture assays, such as the 1.4M SNP capture assay using the Twist technology, which more than doubles the number of sites targeted on the Y chromosome from 32,670 to 81,925[Rohland et al., 2022], this departure from calibration may present similar issues.

Finally, both ChASM and Karyo R_xR_y identify no cases of autosomal aneuploidies, in agreement with the findings from screening the shotgun and 1240k sequence data for the same individuals.

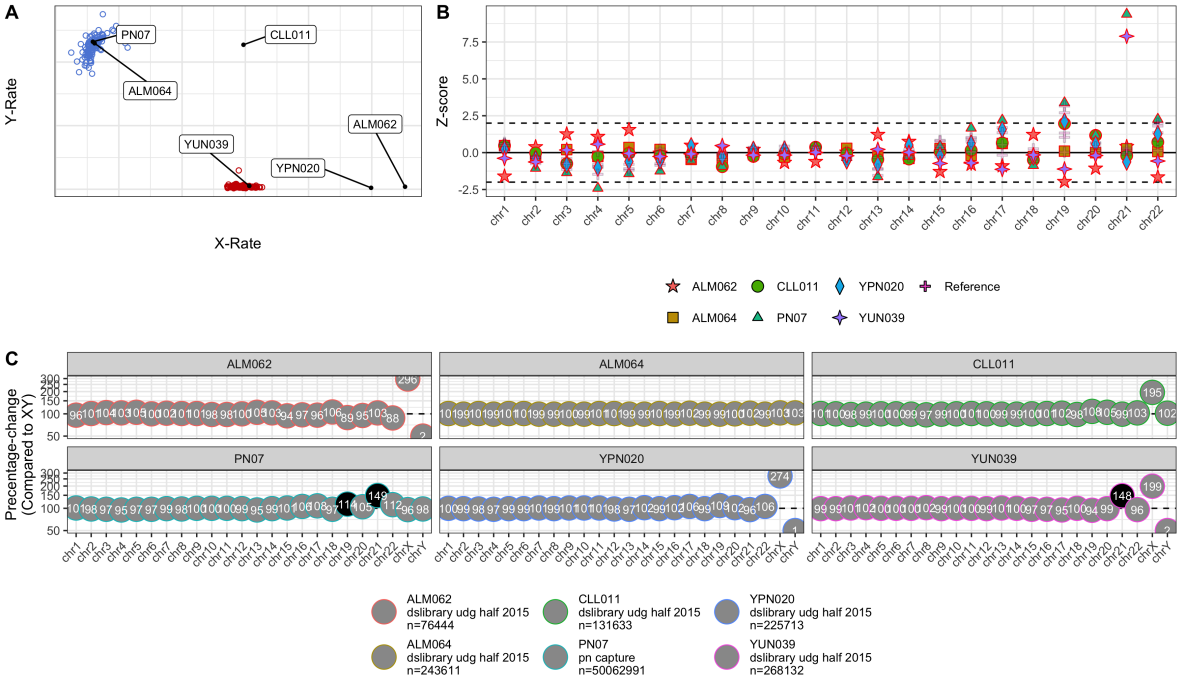

Figure 3: The diagnostic plot for the empirical analysis of the shotgun data for the cases of: XXX syndrome (ALM062 and YPN020), Klinefelter syndrome (CLL011), trisomy 21 (YUN039 and PN07) and an individual with no aneuploidies (ALM064). (A) a scatter plot of the proportions of reads mapping to the X and Y chromosomes, (B) Z-scores per autosomal chromosome, and (C) the percentage-increase of mapped reads per chromosomes compared to expectation (from the Dirichlet-multinomial distribution for an XY individual). Grey and black filled circles indicate $|Z| \geq 2$ for associated Z-scores.

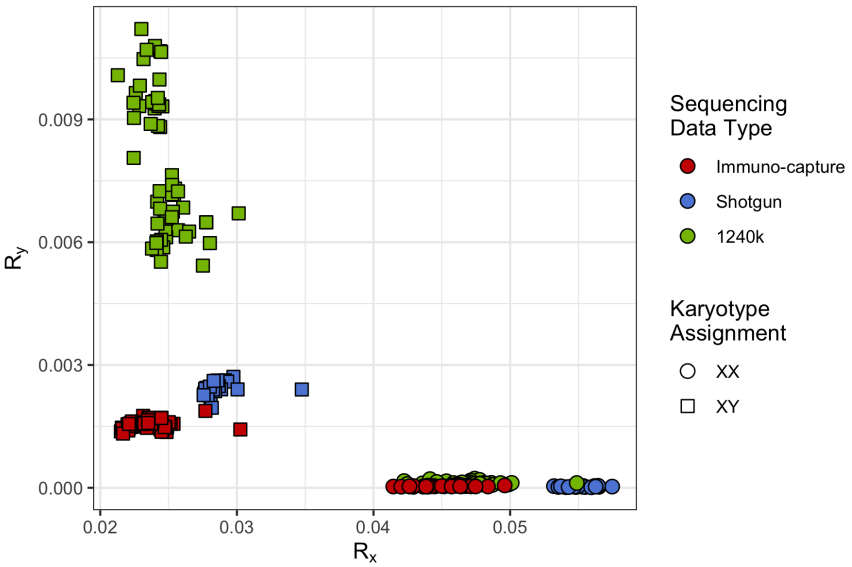

Figure 4: A scatter plot of the X-rate (R_x) and Y-rate (R_y) for samples from Gurgy (GRG)[Rivollat et al., 2023] as calculated by Karyo R_xR_y. Points are coloured by the sequencing data type, and shapes indicate the study from which the samples are sourced.

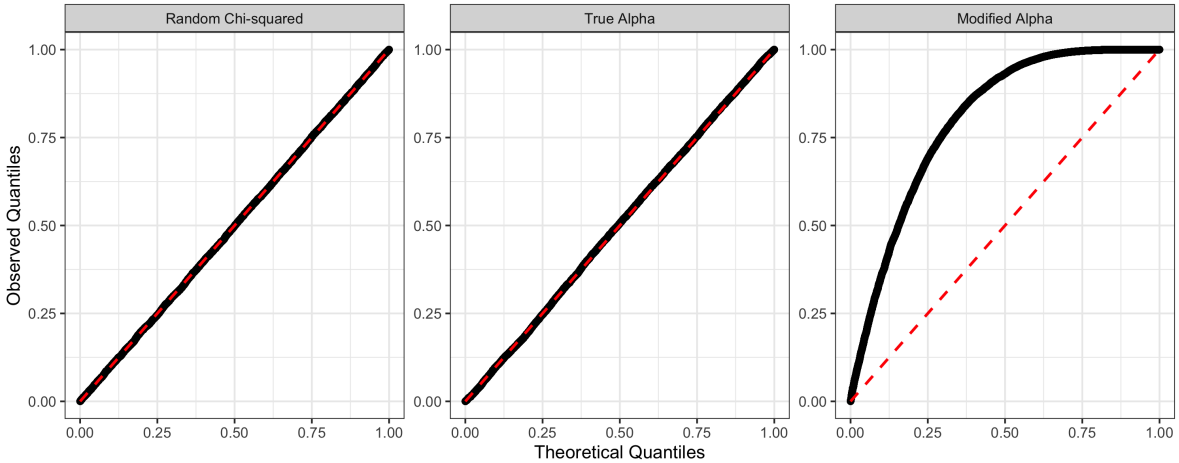

Figure 5: QQ-plots for the random sample from a χ^2_{22} distribution, from the simulations with the true $\boldsymbol{\alpha}$, and with a modified $\boldsymbol{\alpha}$ vs the theoretical quantiles from a χ^2_{22} distribution.

3.2.3 Departure from the reference distribution ("unusual samples")

It is possible that samples do not resemble a realisation from any of the possible $\boldsymbol{\alpha}$ associated with a karyotype due to abnormal DNA degradation or abnormal data production conditions. Under these conditions, one of the karyotypes will fit "best", but we must also assess that they fit sufficiently well. To identify departures from the reference distribution, we calculate λ, a χ^2-statistic (Section 2.5). To assess the performance of λ, and based on empirical data, we simulated 10^5 realisations of read counts, across varying levels of total read counts, half of which are made to depart from the reference distribution (Section 5). We also varied how much we modified $\boldsymbol{\alpha}$ (defined as "error rate").

We see that the test statistics produced under the modified simulations do not fall on the red dashed line, indicating a poor fit to the χ^2_{22} reference distribution. However, the test statistics calculated from the true simulations fit the theoretical quantiles, as well as the random sample from the true reference distribution (Figure 5. Hence, our reference distribution for λ_j appears to be well calibrated.

We were then interested in the performance of the method in predicting whether a realisation was produced from a true or a modified simulation (see Figure 1). We observe that the λ_j are significantly ($p < 2.2 \times 10^{-16}$) greater for modified simulations (Figure 1A). We also compared the associated p-values for the tests, and note that while the p-values from the true simulations appear to form follow the expected uniform distribution, the p-values from the modified simulations are highly left-skewed (Supplementary Figure 1B).

To test the effect of simulation parameters on λ, we modeled the test outcomes using logistic regression (Supplementary Section 6). We find that for simulations where $\boldsymbol{\alpha}$ is unmodified, total read count and error rate have no significant effect on the performance of the test statistic λ ($p = 0.069$ and $p = 0.177$, respectively), and yields an expected false positive rate of 5%. For the simulations where $\boldsymbol{\alpha}$ was modified, both read count and error rate were significant (2×10^{-16}), and the ability of λ to correctly identify unusual samples increase as the total read count and the error rate increase, with the effect size of the error rate was 3.41 times larger than the total read count. Finally, we find that reads proportions need only change by around 3% for λ to have an expected accuracy of 95% or greater (Supplementary Figure 1).

3.2.4 Contamination

In order to test the usefulness of our method for detecting contamination (of karyotypes XY into XX), we simulated 250,000 realisations without contamination, and 250,000 realisations with contamination rates between 5% and 95%. We find that our method has an overall accuracy of only 92.06%. While the sensitivity is high (99.82%), the specificity is lower (86.41%), indicating that the false positive rate would be too high for ChASM to be used as a generalised method for contamination.

However, we only aim to use the contamination estimates from ChASM to estimate if contamination may drive the classification of a sample as karyotype XXY. Hence, for the 52,218 simulations of contamination that resulted in a classification of XXY, the simulated contamination rate was between 14.45% and 71.77% (Figure 3). In these cases, only 5 simulations (9.58×10^{-3}%) failed to identify that contamination was the true reason for the misclassification. Hence, if a sample results in a call of XXY, then the rate of contamination is likely to be quite high, and the

contamination warning generated by ChASM will be reliable. Nevertheless, we suggest using established methods to detect contamination[Renaud et al., 2015, Olalde et al., 2014, Peyrégne and Peter, 2020] that should be implemented before any downstream aDNA analysis, and that this method not be used for general contamination estimates.

4 Discussion

While cases of chromosomal aneuploidies are not frequent, they still represent the most common class of genetic abnormalities leading to genetic conditions. As the number of available ancient genomes continues to rapidly increase, screening for aneuploidies can offer potential explanations for observed skeletal abnormalities or pathologies. Autosomal aneuploidies can lead to significant health challenges in individuals[cite], for which no osteological markers are diagnostic. Sex chromosomal aneuploidies, while often undiagnosed in modern individuals, have been shown to lead to symptoms which may have caused health issues, mental developmental issues or elevated rates of gender dysphoria[cite]. Hence, detecting and diagnosing cases of aneuploidies leads to a more complete understanding of an individual, and the community in which they lived, as well as osteological markers that may otherwise have many other possible explanations.

ChASM can be used in single studies where the statistically required minimum of $n = 23$ individuals is met. However, the more samples of a similar data production type that can be utilised, the better the Dirichlet-multinomial distribution will be calibrated. Hence, we encourage researchers to build large legacy databases of read counts for regular screening of their data, which could form part of an analysis pipeline. Since ChASM requires only 60,000 reads mapping to the human genome, large databases of read counts can be generated from shotgun screening data.

ChASM is a powerful and statistically rigorous tool for detecting chromosomal aneuploidies. We show that the method works for samples with a low as 60,000 mapped reads, equating to approximately 0.0006X and 0.00144X coverage. Additionally, we show that the diagnostic statistic used by ChASM to detect departures from the Dirichlet-multinomial distribution performs well. We suggest that calculating the proportion of samples from individual sequencing runs may perform well as a quality control tool to detect abnormal sequencing runs. Finally, we show that ChASM can reliably detect levels of contamination where they may cause spurious classifications of XXY. However, we warn that this method of contamination estimation lacks power to detect low amounts of contamination, and hence ChASM should not be used as the sole method for detecting contamination in studies.

5 Acknowledgments

The authors wish to thank Associate Professor Elina Salmela, Professor Nigel Bean and Dr Vincent Braunack-Mayer for enlightening and instructive conversations.

6 Data availability

We implemented ChASM in the R-package `RChASM`, available on The Comprehensive R Archive Network (CRAN). A comprehensive vignette for each step of a standard analysis is available at `https://jonotuke.github.io/RChASM/articles/example_analysis.html`. All scripts for the R analyses presented in the study can be found at `https://github.com/BenRohrlach/ChASM_RAnalyses`.

7 Funding

This research was funded by the European Research Council under the European Union's Horizon 2020 research and innovation programme under the grant agreement number 101141408-ROAMANCE.

8 Author Information

A.B.R. and W.H. conceived the study, A.B.R and J.T. performed data analysis and developed statistical tools. W.H. and K.P. provided access to resources and methodology, A.B.R wrote the paper, A.B.R., W.H., K.P. and J.T. edited the paper.

9 Ethics declarations

The authors declare no competing interests.

References

Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stewardson, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207–211, 2015.

Iain Mathieson, Songül Alpaslan-Roodenberg, Cosimo Posth, Anna Szécsényi-Nagy, Nadin Rohland, Swapan Mallick, Iñigo Olalde, Nasreen Broomandkhoshbacht, Francesca Candilio, Olivia Cheronet, et al. The genomic history of southeastern Europe. *Nature*, 555(7695):197–203, 2018.

Rodrigo Barquera, T Lesley Sitter, Casey L Kirkpatrick, Darío A Ramirez, Arthur Kocher, Maria A Spyrou, Lourdes R Couoh, Jorge A Talavera-González, Mario Castro, Tanya von Hunnius, et al. Ancient genomes reveal a deep history of Treponema pallidum in the Americas. *Nature*, 640(8057):186–193, 2025.

Michaela Harbeck, Lisa Seifert, Stephanie Hänsch, David M Wagner, Dawn Birdsell, Katy L Parise, Ingrid Wiechmann, Gisela Grupe, Astrid Thomas, Paul Keim, et al. Yersinia pestis DNA from skeletal remains from the 6th century AD reveals insights into Justinianic Plague. *PLoS pathogens*, 9(5):e1003349, 2013.

Adam B. Rohrlach, Maïté Rivollat, Patxuka de Miguel-Ibáñez, Ulla Nordfors, Anne-Mari Liira, João C Teixeira, Xavier Roca-Rada, Javier Armendáriz-Martija, Kamen Boyadzhiev, Yavor Boyadzhiev, et al. Cases of trisomy 21 and trisomy 18 among historic and prehistoric individuals discovered from ancient DNA. *Nature communications*, 15(1): 1294, 2024.

Frank Maixner, Julia Gresky, and Albert Zink. Ancient DNA analysis of rare genetic bone disorders. *International Journal of Paleopathology*, 33:182–187, 2021.

Kyriaki Anastasiadou, Marina Silva, Thomas Booth, Leo Speidel, Tony Audsley, Christopher Barrington, Jo Buckberry, Diana Fernandes, Ben Ford, Mark Gibson, et al. Detection of chromosomal aneuploidy in ancient genomes. *Communications Biology*, 7(1):14, 2024.

Katharina Fuchs, Biaslan Ch Atabiev, Florian Witzmann, and Julia Gresky. Towards a definition of Ancient Rare Diseases (ard): presenting a complex case of probable Legg-calvé-Perthes disease from the north Caucasian Bronze Age (2200-1650 cal BCE). *International Journal of Paleopathology*, 32:61–73, 2021.

Enrique Dorado-Fernández, Jesús Herrerín-López, Ildefonso Ramírez-González, Loreto Parro-González, and Albert Isidro-Llorens. Survival in Mudejar Spain in the Middle Ages (thirteenth–fourteenth centuries): Ancient rare diseases—an uncommon diagnosis in archaeological human remains. *International Orthopaedics*, 47(11):2869–2875, 2023.

Carol A Crowe, Stuart Schwartz, Cynthia J Black, and Vikram Jaswaney. Mosaic trisomy 22: a case presentation and literature review of trisomy 22 phenotypes. *American journal of medical genetics*, 71(4):406–413, 1997.

Bernardo Orr, Kristina M Godek, and Duane Compton. Aneuploidy. *Current Biology*, 25(13):R538–R542, 2015.

Darren K Griffin. The incidence, origin, and etiology of aneuploidy. *International review of cytology*, 167:263–296, 1996.

Anna Springett, Diana Wellesley, Ruth Greenlees, Maria Loane, Marie-Claude Addor, Larraitz Arriola, Jorieke Bergman, Clara Cavero-Carbonell, Melinda Csaky-Szunyogh, Elizabeth S Draper, et al. Congenital anomalies associated with trisomy 18 or trisomy 13: A registry-based study in 16 European countries, 2000–2011. *American journal of medical genetics Part A*, 167(12):3062–3069, 2015.

John P Costello, Allison Weiderhold, Clauden Louis, Conner Shaughnessy, Syed M Peer, David Zurakowski, Richard A Jonas, and Dilip S Nath. A contemporary, single-institutional experience of surgical versus expectant management of congenital heart disease in trisomy 13 and 18 patients. *Pediatric cardiology*, 36(5):987–992, 2015.

Ashokan Arumugam, Kavitha Raja, Mahalakshmi Venugopalan, Baskaran Chandrasekaran, Kesava Kovanur Sampath, Hariraja Muthusamy, and Nagarani Shanmugam. Down syndrome—a narrative review with a focus on anatomical features. *Clinical anatomy*, 29(5):568–577, 2016.

Angelo Carfì, Manuela Antocicco, Vincenzo Brandi, Camilla Cipriani, Francesca Fiore, Donatella Mascia, Silvana Settanni, Davide L Vetrano, Roberto Bernabei, and Graziano Onder. Characteristics of adults with down syndrome: prevalence of age-related conditions. *Frontiers in medicine*, 1:51, 2014.

Nicole R Tartaglia, Susan Howell, Ashley Sutherland, Rebecca Wilson, and Lennie Wilson. A review of trisomy X (47, XXX). *Orphanet journal of rare diseases*, 5(1):8, 2010.

Kristen Wigby, Lisa Cordeiro, Rebecca Wilson, Kathleen Angkustsiri, Tony J Simon, and Nicole Tartaglia. Adaptive functioning in children and adolescents with Trisomy X: an exploratory analysis. In *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, volume 184, pages 456–468. Wiley Online Library, 2020.

Agnethe Berglund, Kirstine Stochholm, and Claus Højbjerg Gravholt. Morbidity in 47, XYY syndrome: a nationwide epidemiological study of hospital diagnoses and medication use. *Genetics in Medicine*, 22(9):1542–1551, 2020.

Martha Zeger Bardsley, Karen Kowal, Carly Levy, Ania Gosek, Natalie Ayari, Nicole Tartaglia, Najiba Lahlou, Breanna Winder, Shannon Grimes, and Judith L Ross. 47, XYY syndrome: clinical phenotype and timing of ascertainment. *The Journal of pediatrics*, 163(4):1085–1094, 2013.

Kristian A Groth, Anne Skakkebæk, Christian Høst, Claus Højbjerg Gravholt, and Anders Bojesen. Klinefelter syndrome—a clinical update. *The Journal of Clinical Endocrinology & Metabolism*, 98(1):20–30, 2013.

Amy Turriff, Ellen Macnamara, Howard P Levy, and Barbara Biesecker. The impact of living with Klinefelter syndrome: a qualitative exploration of adolescents and adults. *Journal of Genetic Counseling*, 26(4):728–737, 2017.

Jaime Garcia-Heras. The discovery of common chromosome aneuploidies with medical implications through innovative analysis of ancient DNA (aDNA). *Journal of the Association of Genetic Technologists*, 50(3), 2024.

Alexander Robitzsch and Maintainer Alexander Robitzsch. Package 'sirt'. *Computer software]. https://www. maths. bris. ac. uk/R/web/packages/sirt/sirt. pdf*, 2017.

Tymoteusz Wolodzko. extradistr: Additional univariate and multivariate distributions. *R package version*, 1(1), 2020.

Henrik Bengtsson. *matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors)*, 2025. URL `https://CRAN.R-project.org/package=matrixStats`. R package version 1.5.0.

Vanessa Villalba-Mouco, Camila Oliart, Cristina Rihuete-Herrada, Ainash Childebayeva, Adam B Rohrlach, María Inés Fregeiro, Eva Celdrán Beltrán, Carlos Velasco-Felipe, Franziska Aron, Marie Himmel, et al. Genomic transformation and social organization during the Copper Age-Bronze Age transition in southern Iberia. *Science advances*, 7(47): eabi7038, 2021.

Maïté Rivollat, Adam Benjamin Rohrlach, Harald Ringbauer, Ainash Childebayeva, Fanny Mendisco, Rodrigo Barquera, András Szolek, Mélie Le Roy, Heidi Colleran, Jonathan Tuke, et al. Extensive pedigrees reveal the social organization of a Neolithic community. *Nature*, 620(7974):600–606, 2023.

Selina Carlhoff, Wibhu Kutanan, Adam B Rohrlach, Cosimo Posth, Mark Stoneking, Kathrin Nägele, Rasmi Shoocongdej, and Johannes Krause. Genomic portrait and relatedness patterns of the Iron Age Log Coffin culture in northwestern Thailand. *Nature Communications*, 14(1):8527, 2023.

Sandra Penske, Adam B Rohrlach, Ainash Childebayeva, Guido Gnecchi-Ruscone, Clemens Schmid, Maria A Spyrou, Gunnar U Neumann, Nadezhda Atanassova, Katrin Beutler, Kamen Boyadzhiev, et al. Early contact between late farming and pastoralist societies in southeastern Europe. *Nature*, 620(7973):358–365, 2023.

Lara M Cassidy, Ros Ó Maoldúin, Thomas Kador, Ann Lynch, Carleton Jones, Peter C Woodman, Eileen Murphy, Greer Ramsey, Marion Dowd, Alice Noonan, et al. A dynastic elite in monumental Neolithic society. *Nature*, 582 (7812):384–388, 2020.

Alissa Mittnik, Chuan-Chao Wang, Jiří Svoboda, and Johannes Krause. A molecular approach to the sexing of the triple burial at the upper paleolithic site of Dolní Věstonice. *PloS one*, 11(10):e0163019, 2016.

Nadin Rohland, Swapan Mallick, Matthew Mah, Robert Maier, Nick Patterson, and David Reich. Three assays for in-solution enrichment of ancient human DNA at more than a million SNPs. *Genome research*, 32(11-12):2068–2078, 2022.

Gabriel Renaud, Viviane Slon, Ana T Duggan, and Janet Kelso. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome biology*, 16(1):224, 2015.

Iñigo Olalde, Morten E Allentoft, Federico Sánchez-Quinto, Gabriel Santpere, Charleston WK Chiang, Michael DeGiorgio, Javier Prado-Martinez, Juan Antonio Rodríguez, Simon Rasmussen, Javier Quilez, et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature*, 507(7491):225–228, 2014.

Stéphane Peyrégne and Benjamin M Peter. AuthentiCT: a model of ancient DNA damage to estimate the proportion of present-day DNA contamination. *Genome biology*, 21(1):246, 2020.