

.janno file details

Contents

1	Background	1
2	Identifiers	1
3	Relations among samples/individuals	2
4	Spatial position	3
5	Temporal position	3
5.1	General structure	3
5.2	The columns in detail	4
6	Genetic summary data	5
6.1	Individual properties	5
6.2	Library properties	5
6.3	Data yield	6
6.4	Data quality	6
6.4.1	Contamination	7
7	Context information	7

1 Background

The `.janno` file columns are specified in the Poseidon package specification [here](#). The following documentation includes additional background information for many of the variables. This should make it more easy to compile the necessary information for both published and unpublished data. The `.pdf` version of the latest version of this document is available [here](#).

2 Identifiers

The `Poseidon_ID` column represents each sample with an ideally world-wide unique identifier string often equal to the identifier used in the respective accompanying publication. There is no central authority to issue these identifiers, so it remains in the hand of the authors to avoid duplication. We're aware of this inconsistency and hope the aDNA community will eventually come together to establish a mechanism to ensure uniqueness of identifiers. If there are multiple samples from one individual, then they have to be clearly distinguished with

relevant suffixes added to the `Poseidon_ID`. `Poseidon_ID`s are also employed in the genetic data files in a Poseidon package and therefore have to adhere to certain constraints.

The column `Alternative_IDs` provides a way to list other IDs used for the respective individual. These might for example be names used in different publications or popular names like “Iceman”, “Ötzi”, “Girl of the Uchter Moor”, “Tollund Man”, etc.. The `Relation_*` columns described below allow to more precisely express the relationship type “identical” among samples in a Poseidon package.

The `Collection_ID` column stores an additional, secondary identifier as it is often provided by collaboration partners (archaeologists, museums, collections) that provide specimen for archaeogenetic research. These identifiers might have a very heterogenous structure and may not be unique across different projects or institutions. The `Collection_ID` column is therefore a free form text field.

The `Group_Name` column contains one or multiple group or population names for each individual, separated by `;`. The first entry must be identical to the one used in the genotype data for the respective sample in a Poseidon package. Assigning group and population names is a hard problem in archeogenetics [1], so the `.janno` file allows for more than one identifier.

3 Relations among samples/individuals

To systematically document biological relationships uncovered among samples/individuals in one or multiple Poseidon datasets (e.g. with software like READ [2] or BREADR [3]), the `.janno` file can be fit with a set of columns featuring the `Relation_*` prefix. They together should be capable to encode all kinds of pairwise, biological relationships an individual might have.

`Relation_To` is a string list column (so: multiple values are possible if separated by `;`) that stores the `Poseidon_ID`s of other samples/individuals to which the current individual has some relationship.

`Relation_Degree` stores a formal description of the closeness of this relationship as measured purely from aDNA data. It is therefore also a list column that can hold the following values for each relationship:

- `identical`: The two samples are from the same individual or from identical twins
- `first`: The two individuals are closely related – a first degree relationship (e.g. siblings, parent-offspring)
- `second`: A second degree relationship (e.g. cousins, grandparent to grandchild)
- `thirdToFifth`: A third to fifth degree relationship (e.g. great-grandparent to great-grandchild)
- `sixthToTenth`: A sixth to tenth degree relationship
- `unrelated`: Unrelated – this is the default state among all individuals, which does not have to be expressed explicitly. This category will therefore probably never be used
- `other`: Any other kind of relationship not covered by the aforementioned categories

For each entry in `Relation_To` there must be a corresponding entry in `Relation_Degree`.

`Relation_Type` allows to add more verbose details about the relationship type, if it was possible to reconstruct that from the archaeological or historical context. Because there are too many possible permutations, there is no pre-defined set of values for what can and cannot be entered here. It is advisable, though, to stick to a general scheme like the following, which describes a given relationship from the point of view of the current individual:

- `father_of`: This individual is likely the father of the partner individual
- `grandchild_of`: This individual is likely the grandchild of the partner individual
- `mother_or_daughter_of`: This individual is likely either the mother or daughter of the partner individual (which might be unclear, in case of imprecise archaeological dating)

69 • ...

70 Unlike `Relation_Degree`, `Relation_Type` can be left empty even if there are entries in `Relation_To`. But
71 if it is filled, then the number of values must be equal to the number of entries in both `Relation_To` and
72 `Relation_Degree`.

73 The `Relation_Note` column allows to add free-text information about the relationships of this individual. This
74 might also include information about the method used to infer the degree and type.

75 4 Spatial position

76 The `.janno` file contains five columns to describe the spatial origin of an individual sample: `Country`,
77 `Location`, `Site` and finally `Latitude` and `Longitude`.

78 The `Country` column should contain a present-day political country name following the `English short name`
79 in [ISO 3166](#).

80 The `Country_ISO` column should contain the present-day political country of origin of the sample, expressed in
81 codes using the standard ISO 3166-1 alpha-2, like “AR” for Argentina or “NO” for Norway.

82 The `Location` column allows for free form text entry and can contain further, unspecified location information.
83 This might be the name of an administrative or geographic region, or an arbitrary unit of reference like a
84 mountain, lake or city close to the point of discory of the respective sample.

85 The `Site` column should contain a site name, ideally in the latin alphabet and ideally the name that is
86 commonly used in publications.

87 The `Latitude` and `Longitude` column should contain geographic coordinates (WGS84) in decimal degrees
88 (DD) with a precision of not more than five places after the decimal point. This yields a precision of about
89 [1.1132m at the equator](#) which is sufficient to describe the position of an archaeological site. Coordinates in other
90 formats like for example Degrees Minutes Seconds (DMS) or in completely different coordinate reference systems
91 should be transformed. There exist many open source software solutions to do that, most based on the [PROJ](#)
92 [library](#) e.g. the [The World Coordinate Converter](#).

93 5 Temporal position

94 The temporal position of a sample is encoded with seven different columns in the `.janno` file: `Date_C14_Labnr`,
95 `Date_C14_Uncal_BP`, `Date_C14_Uncal_BP_Err`, `Date_BC_AD_Median`, `Date_BC_AD_Start`, `Date_BC_AD_Stop`,
96 `Date_Type`

97 5.1 General structure

98 The `Date_Type` column handles the general distinction between the most common forms of age information:

- 99 • `modern` : Applies to present day reference samples, so not ancient DNA.
- 100 • `C14` : Applies if there is a set of radiocarbon dates explicitly listed in the columns `Date_C14_Labnr`,
101 `Date_C14_Uncal_BP` and `Date_C14_Uncal_BP_Err` whose post-calibration probability distribution is a
102 meaningful prior for the individual’s year of death. The dates do not always have to be directly from the
103 individual’s tissue, but they should be immediately relevant for their year of death (e.g. a date from a
104 grain kernel recovered from the individual’s grave).

- **contextual**: Applies in all other cases if the columns **Date_BC_AD_Median**, **Date_BC_AD_Start**, **Date_BC_AD_Stop** can be filled. This includes age attribution based on the archaeologically determined stratigraphy or typological information. **contextual** should also be chosen if the sample is dated very indirectly with radiocarbon dating (e.g. radiocarbon dates from other, unrelated features of the same site) or dated with other physical or chemical dating methods (e.g. dendrochronology or optically stimulated luminescence).

So **Date_C14_Labnr**, **Date_C14_Uncal_BP** and **Date_C14_Uncal_BP_Err** only go along with **Date_Type = C14**, whereas **Date_BC_AD_Median**, **Date_BC_AD_Start**, **Date_BC_AD_Stop** complement both **Date_Type = C14** and **Date_Type = contextual**. Radiocarbon dates that only serve as secondary evidence for a contextual dating should not be reported in **Date_C14_Labnr**, **Date_C14_Uncal_BP** and **Date_C14_Uncal_BP_Err**.

5.2 The columns in detail

Each radiocarbon date has a unique identifier: the “lab number”. It consists of a lab code issued by the journal [Radiocarbon](#) for each laboratory and a serial number. This lab number makes the date well identifiable and should be reported in **Date_C14_Labnr** with the lab code separated from the serial number with a minus symbol.

The uncalibrated radiocarbon measurement can be described by a Gaussian distribution with mean and standard deviation. So the column **Date_C14_Uncal_BP** holds the mean of that distribution in years before present (BP) as usually reported by radiocarbon laboratories. The age is always a positive integer value starting from a zero that corresponds to 1950 AD. The column **Date_C14_Uncal_BP_Err** holds the respective standard deviation for each date in years. This should be the 1-sigma distance, so that the probability that the actual uncalibrated age of the measured sample is within the $\text{Date_C14_Uncal_BP} \pm \text{Date_C14_Uncal_BP_Err}$ range is about 68%.

Date_C14_Labnr, **Date_C14_Uncal_BP** and **Date_C14_Uncal_BP_Err** each can hold multiple values separated by **;** to allow for multiple radiocarbon dates for each aDNA sample. With multiple values the number and order of values in the columns must be equal.

In the columns **Date_BC_AD_Median**, **Date_BC_AD_Start**, **Date_BC_AD_Stop** ages are reported in years BC and AD, so in relation to the zero point of the Gregorian calendar. BC dates are represented with negative, AD with positive integer values.

- If radiocarbon dates are available (**Date_Type = C14**): **Date_BC_AD_Median** should report the median age after calibration. With multiple dates this can be determined either with sum calibration or more complex (e.g. bayesian) age modelling. **Date_BC_AD_Start** and **Date_BC_AD_Stop** should report the starting/ending age of a 95% probability window around the age median.
- If only contextual (e.g. from archaeological typology) age information is available (**Date_Type = contextual**): **Date_BC_AD_Start** and **Date_BC_AD_Stop** should simply report the approximate starting and end date determined by the respective source of scientific authority (e.g. an archaeologist knowledgeable about the relevant typological sequences). In this case **Date_BC_AD_Median** should be calculated as the mean of **Date_BC_AD_Start** and **Date_BC_AD_Stop** rounded to an integer value.
- If the sample is a modern reference sample (**Date_Type = modern**): **Date_BC_AD_Median**, **Date_BC_AD_Start**, **Date_BC_AD_Stop** should all be set to the value 2000, for 2000AD.

The column **Date_Note** allows to add arbitrary free-text information about the dating of a sample.

144 6 Genetic summary data

145 6.1 Individual properties

146 The **Genetic_Sex** column should encode the biological sex as determined from the DNA read distribution on
147 the X and Y chromosome. It only allows for the entries

- 148 • **F** : female
- 149 • **M** : male
- 150 • **U** : unknown

151 This limitation stems from the genotype data formats by Plink and the Eigensoft software package. Edge cases
152 (e.g. XXY, XYY, X0, ...) can not be expressed with this format and should be reported as **U** with an additional
153 comment in the free text **Note** field. Genetic sex determination for ancient DNA can be performed for example
154 with Sex.DetERRmine [4].

155 The **MT_Haplogroup** column is meant to store the human mitochondrial DNA haplogroup for the respective
156 individual in a simple string. The entry can be arbitrarily precise. A software tool to determine the MT haplogroup
157 is for example Haplogrep [5].

158 The **Y_Haplogroup** column holds the respective human Y-chromosome DNA haplogroup in a simple string.
159 The notation should follow a syntax with the main branch + the most terminal derived Y-SNP separated with a
160 minus symbol (e.g. R1b-P312).

161 6.2 Library properties

162 The **Source_Tissue** column documents the skeletal, soft tissue or other elements from which source material for
163 DNA library preparation have been extracted. If multiple libraries have been taken from different elements, these
164 can be listed separated by **;**. Specific bone names should be reported with an underscore (e.g. bone_phalanx,
165 tooth_molar).

166 The **Nr_Libraries** column holds a simple integer value of the number of libraries that have been prepared for
167 an individual.

168 The **Capture_Type** column specifies the general pre-sequencing preparation methods that have been applied to
169 the library. See [6] for a review of the different techniques (not including younger developments). This field can
170 hold one of multiple different values, but also multiple of these separated by **;** if different methods have been
171 applied for different libraries.

- 172 • **Shotgun** : Sequencing without any enrichment (whole genome sequencing, screening etc.)
- 173 • **1240k** : Target enrichment with hybridization capture optimised for sequences covering the 1240k SNP
174 array [7], [8], [9]
- 175 • **ArborComplete**, **ArborPrimePlus**, **ArborAncestralPlus** : Target enrichment with hybridization cap-
176 ture as provided by Arbor Biosciences in three different kits branded [myBaits Expert Human Affinities](#)
- 177 • **TwistAncientDNA** : Target enrichment with hybridization capture as provided by Twist Bioscience [10]
- 178 • **OtherCapture** : Target enrichment with hybridization capture for any other set of sequences
- 179 • **ReferenceGenome** : Modern reference genomes where aDNA fragmentation is not an issue and other
180 sample preparation techniques apply

181 The **UDG** column documents if the libraries for the respective individual went through UDG (USER enzyme)
182 treatment. This wet lab protocol step removes molecular damage in the form of deaminated cytosines characteristic

183 of ancient DNA.

- 184 • **minus** : A protocol without UDG treatment (e.g. [11])
- 185 • **half** : A protocol with UDG-half treatment (e.g. [12])
- 186 • **plus** : A protocol with UDG-full treatment (e.g. [13])
- 187 • **mixed** : Multiple later merged libraries went through different UDG treatment approaches

188 The **Library_Names** column should contain the names for the library as used in the publication.

189 The **Library_Built** column describes the library preparation method regarding single- or double-stranded
190 protocols. See e.g. [14] for more information.

- 191 • **ds** : Double-stranded library preparation
- 192 • **ss** : Single-stranded library preparation
- 193 • **mixed** : If multiple libraries with different strandedness were used. See also the Sequencing Source File in
194 the Poseidon package as a way to provide details.

195 The **Genotype_Ploidy** column stores a characteristic of the aDNA data treatment. Humans have two complete
196 sets of chromosomes in their cells and hence are diploid organisms. For many computational aDNA applications
197 it is more practical, though, to work with pseudo-haploid data, so data where only one read per position is selected
198 by a random sampling process.

- 199 • **diploid** : No random read selection
- 200 • **haploid** : Random read selection to produce pseudo-haploid data

201 The column **Data_Preparation_Pipeline_URL** should finally store an URL that links to a complete and
202 human-readable description of the computational pipeline (for example a specific configuration for nf-core/eager
203 [15] by which the sample data was processed.

204 6.3 Data yield

205 The **Endogenous** column holds the percentage of mapped reads over the total amount of reads that went into
206 the mapping pipeline. That boils down to the DNA percentage of the library that matches the (human) reference.
207 It should be determined from Shotgun libraries (so before any hybridization capture), not on target and without
208 any quality filtering. In case of multiple libraries only the highest value should be reported. The % endogenous
209 DNA can be calculated for example with the [endorS.py](#) script.

210 The **Nr_SNPs** column gives the number of SNPs reported in the genotype data files for this individual.

211 The **Coverage_on_Target_SNPs** column reports the mean SNP coverage on the target SNP array (e.g. 1240K)
212 for the merged libraries of this sample. To calculate the coverage it is necessary to determine which SNPs are
213 covered how many times by the mapped reads. Individual SNPs might be covered multiple times, whereas others
214 may not be covered at all by the highly deteriorated ancient DNA. The coverage for each SNP is therefore a
215 number between 0 and n. The statistic can be determined for example with the QualiMap [16] software package.
216 In case of multiple libraries, the coverage can be given as a mean across all of them.

217 6.4 Data quality

218 The **Damage** column contains the % damage on the first position of the 5' end for the main Shotgun library used
219 for sequencing or capture. This is an important statistic to verify the age of ancient DNA. In case of multiple
220 libraries you should report a value from the merged read alignment.

221 6.4.1 Contamination

222 Contamination of ancient DNA with foreign reads is a major challenge for archaeogenetics. There exist multiple
223 competing ideas, algorithms and software tools to estimate the degree of contamination for individual samples
224 (e.g. ANGSD [17], contamLD [18] or hapCon [19]), with some methods only applicable under certain circumstances
225 (e.g. popular X-chromosome based approaches only work on male individuals). Also the results of different
226 methods tend to differ both in the degree of contamination they estimate and in the way the output is usually
227 encoded. To cover the multitude of methods in this domain, and to make the results representable in the `.janno`
228 file, we offer the `Contamination_*` column family.

229 `Contamination` is a list column to represent the different contamination values estimated for a sample with
230 one or multiple software tools. As usual multiple values are separated by `;`.

231 `Contamination_Err` is another list column to store the respective (standard) error term for the values in
232 `Contamination`.

233 Some tools for contamination estimation do not return a mean plus a standard error. ContamMix, for example,
234 yields a 95% confidence interval instead, to better represent assymetric output distributions. `Contamination`
235 and `Contamination_Err` can not represent this. We suggest to derive a mean and a standard error from these
236 alternative outputs. The latter can be calculated as the largest distance from the mean to the limits of the
237 confidence interval.

238 `Contamination_Meas` finally is the third necessary list column, which contextualizes the values in
239 `Contamination` and `Contamination_Err`. Each measure in these columns has to be accompanied by the
240 software and software version used to calculate it. The individual entries might e.g. look like this:

- 241 • `ANGSD v0.935`
- 242 • `hapCon v0.4a1`
- 243 • `custom script`

244 This setup has the consequence that the columns `Contamination`, `Contamination_Err`, `Contamination_Meas`
245 always have to have the same number of `;`-separated values.

246 The `Contamination_Note` column is a free text field to add additional information about the contamination
247 estimates, e.g. which parameters were used with the respective software tools.

248 7 Context information

249 The `Genetic_Source_Accession_IDs` column was introduced to link the derived genotype data in Poseidon
250 with the raw sequencing data typically uploaded to archives like the ENA [20] or SRA [21]. There projects or
251 even individual samples are given clear identifiers: Accession IDs. This janno column is supposed to store one
252 or multiple of these Accessions IDs for each individual/sample in Poseidon. If multiple are entered, then they
253 should be arranged by descending specificity from left to right (e.g. project id > sample id > sequencing run id).

254 The `Primary_Contact` column is a free form text field that stores the name of the main or the corresponding
255 author of the respective paper for published data.

256 The `Publication` column holds either the value `unpublished` for (yet) unpublished samples or – for published
257 data – one or multiple citation-keys of the form `AuthorJournalYear` without any spaces or special characters.
258 These keys have to be identical to the BibTeX citation-keys identifying the respective entries in the `.bib` file of
259 the package. BibTeX is a file format to store bibliographic information, where each entry (article, book, website,

260 ...) is defined by a series of parameters (authors, year of publication, journal, ...). Here's an example `.bib` file
261 with two entries for [22] and [23]:

```
262 @article{Cassidy2015,  
263     doi = {10.1073/pnas.1518445113},  
264     url = {https://doi.org/10.1073%2Fpnas.1518445113},  
265     year = 2015,  
266     month = {dec},  
267     publisher = {Proceedings of the National Academy of Sciences},  
268     volume = {113},  
269     number = {2},  
270     pages = {368--373},  
271     author = {Lara M. Cassidy and Rui Martiniano and Eileen M. Murphy and  
272             Matthew D. Teasdale and James Mallory and Barrie Hartwell  
273             and Daniel G. Bradley},  
274     title = {Neolithic and Bronze Age migration to Ireland and establishment  
275             of the insular Atlantic genome},  
276     journal = {Proceedings of the National Academy of Sciences}  
277 }  
278  
279 @article{Feldman2019,  
280     doi = {10.1126/sciadv.aax0061},  
281     url = {https://doi.org/10.1126%2Fsciadv.aax0061},  
282     year = 2019,  
283     month = {jul},  
284     publisher = {American Association for the Advancement of Science ({AAAS})},  
285     volume = {5},  
286     number = {7},  
287     pages = {eaax0061},  
288     author = {Michal Feldman and Daniel M. Master and Raffaella A. Bianco and  
289             Marta Burri and Philipp W. Stockhammer and Alissa Mittnik and  
290             Adam J. Aja and Choongwon Jeong and Johannes Krause},  
291     title = {Ancient {DNA} sheds light on the genetic origins of early Iron Age  
292             Philistines},  
293     journal = {Science Advances}  
294 }
```

295 The string `CassidyPNAS2015` is the citation-key of the first entry. To cite both publications in the `Publication`
296 column, one would enter `CassidyPNAS2015;FeldmanScienceAdvances2019` .

297 When creating a new Poseidon package the `.bib` file should be filled together with the `Publication` column.
298 One of the most simple ways to obtain the BibTeX entries may be to request them with the doi from the [doi2bib](#)
299 web app. It could be necessary to adjust the result manually, though. The citation-key, for example, has to be
300 replaced by the one used in the `Publication` column.

301 The `Note` column is a free form text field that can contain small amounts of additional information that is not
302 yet expressed in a more systematic form in the the other `.janno` file columns.

The **Keywords** column was introduced to allow for tagging individuals with arbitrary keywords. This should simplify sorting and filtering in personal Poseidon package repositories. Each keyword is a string and multiple keywords can be separated with `;`.

- [1] S. Eisenmann *et al.*, “Reconciling material cultures in archaeology with genetic data: The nomenclature of clusters emerging from archaeogenomic analysis,” *Scientific Reports*, vol. 8, no. 1, Aug. 2018, doi: [10.1038/s41598-018-31123-z](https://doi.org/10.1038/s41598-018-31123-z).
- [2] J. M. Monroy Kuhn, M. Jakobsson, and T. Günther, “Estimating genetic kin relationships in prehistoric populations,” *PLOS ONE*, vol. 13, no. 4, p. e0195491, Apr. 2018, doi: [10.1371/journal.pone.0195491](https://doi.org/10.1371/journal.pone.0195491).
- [3] A. B. Rohrlach, J. Tuke, D. Popli, and W. Haak, “BREADR: An r package for the bayesian estimation of genetic relatedness from low-coverage genotype data,” Apr. 2023, doi: [10.1101/2023.04.17.537144](https://doi.org/10.1101/2023.04.17.537144).
- [4] T. C. Lamnidis *et al.*, “Ancient fennoscandian genomes reveal origin and spread of siberian ancestry in europe,” *Nature Communications*, vol. 9, no. 1, Nov. 2018, doi: [10.1038/s41467-018-07483-5](https://doi.org/10.1038/s41467-018-07483-5).
- [5] S. Schönherr, H. Weissensteiner, F. Kronenberg, and L. Forer, “Haplogrep 3 - an interactive haplogroup classification and analysis platform,” *Nucleic Acids Research*, vol. 51, no. W1, pp. W263–W268, Apr. 2023, doi: [10.1093/nar/gkad284](https://doi.org/10.1093/nar/gkad284).
- [6] M. Knapp and M. Hofreiter, “Next generation sequencing of ancient DNA: Requirements, strategies and perspectives,” *Genes*, vol. 1, no. 2, pp. 227–243, Jul. 2010, doi: [10.3390/genes1020227](https://doi.org/10.3390/genes1020227).
- [7] Q. Fu *et al.*, “An early modern human from romania with a recent neanderthal ancestor,” *Nature*, vol. 524, no. 7564, pp. 216–219, Jun. 2015, doi: [10.1038/nature14558](https://doi.org/10.1038/nature14558).
- [8] W. Haak *et al.*, “Massive migration from the steppe was a source for indo-european languages in europe,” *Nature*, vol. 522, no. 7555, pp. 207–211, Mar. 2015, doi: [10.1038/nature14317](https://doi.org/10.1038/nature14317).
- [9] I. Mathieson *et al.*, “Genome-wide patterns of selection in 230 ancient eurasians,” *Nature*, vol. 528, no. 7583, pp. 499–503, Nov. 2015, doi: [10.1038/nature16152](https://doi.org/10.1038/nature16152).
- [10] N. Rohland, S. Mallick, M. Mah, R. Maier, N. Patterson, and D. Reich, “Three assays for in-solution enrichment of ancient human DNA at more than a million SNPs,” *Genome Research*, vol. 32, no. 11–12, pp. 2068–2078, Nov. 2022, doi: [10.1101/gr.276728.122](https://doi.org/10.1101/gr.276728.122).
- [11] F. Aron, G. U Neumann, and G. Brandt, “Non-UDG treated double-stranded ancient DNA library preparation for illumina sequencing v1,” Dec. 2019, doi: [10.17504/protocols.io.bakricv6](https://doi.org/10.17504/protocols.io.bakricv6).
- [12] F. Aron, G. U Neumann, and G. Brandt, “Half-UDG treated double-stranded ancient DNA library preparation for illumina sequencing v1,” Sep. 2020, doi: [10.17504/protocols.io.bmh6k39e](https://doi.org/10.17504/protocols.io.bmh6k39e).
- [13] F. Aron, G. U Neumann, and G. Brandt, “Full-UDG treated double-stranded ancient DNA library preparation for illumina sequencing v1,” Dec. 2020, doi: [10.17504/protocols.io.bqbpmsmn](https://doi.org/10.17504/protocols.io.bqbpmsmn).
- [14] M.-T. Gansauge and M. Meyer, “Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA,” *Nature Protocols*, vol. 8, no. 4, pp. 737–748, Mar. 2013, doi: [10.1038/nprot.2013.038](https://doi.org/10.1038/nprot.2013.038).
- [15] J. A. Fellows Yates *et al.*, “Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager,” *PeerJ*, vol. 9, p. e10947, Mar. 2021, doi: [10.7717/peerj.10947](https://doi.org/10.7717/peerj.10947).
- [16] K. Okonechnikov, A. Conesa, and F. García-Alcalde, “Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data,” *Bioinformatics*, vol. 32, no. 2, pp. 292–294, Oct. 2015, doi: [10.1093/bioinformatics/btv566](https://doi.org/10.1093/bioinformatics/btv566).
- [17] T. S. Korneliussen, A. Albrechtsen, and R. Nielsen, “ANGSD: Analysis of next generation sequencing data,” *BMC Bioinformatics*, vol. 15, no. 1, Nov. 2014, doi: [10.1186/s12859-014-0356-4](https://doi.org/10.1186/s12859-014-0356-4).

- 324 [18] N. Nakatsuka, É. Harney, S. Mallick, M. Mah, N. Patterson, and D. Reich, “ContamLD: Estimation of ancient nuclear DNA contamination using breakdown of linkage disequilibrium,” *Genome Biology*, vol. 21, no. 1, Aug. 2020, doi: [10.1186/s13059-020-02111-2](https://doi.org/10.1186/s13059-020-02111-2).
- 325 [19] Y. Huang and H. Ringbauer, “hapCon: Estimating contamination of ancient genomes by copying from reference haplotypes,” *Bioinformatics*, vol. 38, no. 15, pp. 3768–3777, Jun. 2022, doi: [10.1093/bioinformatics/btac390](https://doi.org/10.1093/bioinformatics/btac390).
- 326 [20] J. Burgin *et al.*, “The european nucleotide archive in 2022,” *Nucleic Acids Research*, vol. 51, no. D1, pp. D121–D125, Nov. 2022, doi: [10.1093/nar/gkac1051](https://doi.org/10.1093/nar/gkac1051).
- 327 [21] K. Katz, O. Shutov, R. Lapoint, M. Kimelman, J. R. Brister, and C. O’Sullivan, “The sequence read archive: A decade more of explosive growth,” *Nucleic Acids Research*, vol. 50, no. D1, pp. D387–D390, Nov. 2021, doi: [10.1093/nar/gkab1053](https://doi.org/10.1093/nar/gkab1053).
- 328 [22] L. M. Cassidy *et al.*, “Neolithic and bronze age migration to ireland and establishment of the insular atlantic genome,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 2, pp. 368–373, Dec. 2015, doi: [10.1073/pnas.1518445113](https://doi.org/10.1073/pnas.1518445113).
- 329 [23] M. Feldman *et al.*, “Ancient DNA sheds light on the genetic origins of early iron age philistines,” *Science Advances*, vol. 5, no. 7, p. eaax0061, Jul. 2019, doi: [10.1126/sciadv.aax0061](https://doi.org/10.1126/sciadv.aax0061).