

Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager

This manuscript ([permalink](#)) was automatically generated from [apeltzer/eager2-paper@66afeff](#) on August 6, 2020.

Authors

- **James A. Fellows Yates**

 [0000-0001-5585-6277](#) ·  [jfy133](#) ·  [jafellowsyates](#)

Microbiome Sciences Group, Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany; Institut für Vor- und Frühgeschichtliche Archäologie und Provinzialrömische Archäologie, Ludwig Maximilian University, Munich, Germany · Funded by Max Planck Society

- **Thiseas C. Lamnidis**

 [0000-0003-4485-8570](#) ·  [TCLamnidis](#) ·  [TCLamnidis](#)

Population Genetics Group, Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany · Funded by Max Planck Society

- **Maxime Borry**

 [0000-0001-9140-7559](#) ·  [maxibor](#) ·  [notmaxib](#)

Microbiome Sciences Group, Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany · Funded by Max Planck Society

- **Aida Andrades Valtueña**

 [0000-0002-1737-2228](#) ·  [aidaanva](#) ·  [aidaanva](#)

Computational Pathogenomics Group, Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany · Funded by Max Planck Society

- **Zandra Fagernäs**

 [0000-0003-2667-3556](#) ·  [ZandraFagernas](#) ·  [ZandraSelina](#)

Microbiome Sciences Group, Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany · Funded by Max Planck Society

- **Stephen Clayton**

 [0000-0001-5223-9695](#) ·  [sc13-bioinf](#)

Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany · Funded by Max Planck Society

- **Maxime U. Garcia**

 [0000-0003-2827-9261](#) ·  [MaxUlysse](#) ·  [gau](#)

Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden · Funded by Barncancerfonden

- **Judith Neukamm**

 [0000-0001-8141-566X](#) ·  [JudithNeukamm](#) ·  [JudithNeukamm](#)

Palaeogenetics Group, Institute of Evolutionary Medicine, University of Zurich, Zürich, Switzerland

- **Alexander Peltzer**

 [0000-0002-6503-2180](#) ·  [apeltzer](#) ·  [alex_peltzer](#)

Quantitative Biology Center (QBiC), Eberhard-Karls-Universität, Tübingen, Germany

Abstract

The broadening utilisation of ancient DNA (aDNA) to address archaeological, palaeontological, and biological questions is resulting in a rising diversity in the size of laboratories and scale of analyses being performed. In the context of this heterogeneous landscape, we present nf-core/eager, an advanced and entirely redesigned pipeline for the analysis of ancient genomic data. nf-core/eager builds on existing ideas and concepts introduced in the original EAGER pipeline, and improves various aspects of the analysis procedure by employing computational frameworks such as Nextflow and nf-core. The pipeline aims to address three main themes: accessibility and adaptability to different research groups and their computing configurations, reproducibility to ensure robust analytical standards in the field, and updating the EAGER pipeline to the latest routine ancient genomic practises. This new version of EAGER has been developed within the nf-core initiative, to ensure high quality software development and maintenance support; contributing to a long-term lifecycle for the pipeline. nf-core/eager will assist in ensuring that ancient DNA sequencing data can be used by a diverse range of research groups and fields.

Introduction

Ancient DNA (aDNA) has become a widely accepted source of biological data, helping to provide new perspective for a range of fields including archaeology, ecology, cultural heritage, and palaeontology. The utilisation of next-generation sequencing has allowed the recovery of aDNA from a wide variety of sources, including but not limited to, the skeletal remains of animals [1,2,3,4], modern and archaic humans [5,6,7,8], bacteria [9,10,11], viruses [12,13], plants [14,15], palaeofaeces [16,17], dental calculus [18,19], sediments [20,21], medical slides [22], parchment [23], and most recently, ancient 'chewing gum' [24,25]. Improvement in laboratory protocols to increase yields of otherwise trace amounts of DNA has at the same time led to studies that can total hundreds of ancient individuals [26,27], spanning single [28] to thousands of organisms [18]. These differences of disciplines have led to a heterogeneous landscape in terms of the types of analyses undertaken, and their computational resource requirements [29,30]. Taking into consideration the unequal distribution of resources (and infrastructure such as internet connection), streamlined and efficient pipelines can help increase accessibility to high-quality analyses.

The degraded nature of aDNA poses an extra layer of complexity to standard modern genomic analysis. Through a variety of processes [31] DNA molecules fragment over time, resulting in ultra-short molecules [32]. These sequences have low nucleotide complexity making it difficult to identify with precision which part of the genome a read (sequenced DNA molecule) is derived from. When fragmentation is not a 'clean break', this can lead to uneven ends with single-stranded 'overhangs' at end of molecules, which are susceptible to chemical processes such as deamination that lead to misincorporation of bases during library construction [33]. On top of this, taphonomic processes such as heat, moisture, and microbial- and burial-environment processes lead to varying rates of degradation [34,35]. The original DNA content of a sample is therefore increasingly lost over time and supplanted by younger 'environmental' DNA. Later handling by archaeologists, museum curators, and scientists can also contribute 'modern' contamination. While these characteristics can help provide evidence towards the 'authenticity' of true aDNA sequences (e.g. aDNA C>T 'damage' profiles [36]), they also pose specific challenges such as unspecific DNA alignment and/or low coverage and miscoding lesions that can result in low-confidence genotyping. These factors often lead to prohibitive sequencing costs when retrieving enough data for modern NGS data pipelines (such as more than 1 billion reads for a 1X depth coverage *Yersinia pestis* genome [37]), and thus require aDNA-tailored methods and techniques to overcome these challenges.

Two previously published and commonly used pipelines in the field are PALEOMIX [38] and EAGER [39]. These two pipelines take a similar approach to link together standard tools used for Illumina

NGS data processing (sequencing quality control, sequencing adapter removal/and or paired-end read merging, mapping of reads to a reference genome, genotyping, etc.), but with a specific focus on tools that are designed for, or well-suited for aDNA (such as the bwa aln algorithm for ultra-short molecules [40] and mapDamage [41] for evaluation of aDNA characteristics). Yet, neither of these pipelines have had major updates to bring them in-line with current routine aDNA analyses. Metagenomic screening of off-target genomic reads for pathogens or microbiomes [18,19] has become particularly common in palaeo- and archaeogenetics, given its role in revealing widespread infectious disease and possible epidemics that had previously been undetected in the archaeological record [12,13,37,42]. Without easy access to the latest field-established analytical routines, aDNA studies risk being published without the necessary quality control checks that ensure aDNA authenticity as well as limiting the full range of possibilities from their data. Given that material from samples is limited, there are both ethical as well as economical interests to maximise analytical yield [43].

To address these shortcomings, we have completely re-implemented the latest version of the EAGER pipeline in Nextflow [44] (a domain-specific-language or ‘DSL’, specifically designed for the construction of omics analysis pipelines), introduced new features, and more flexible pipeline configuration. In addition, the newly named pipeline - nf-core/eager - has been developed in the context of the nf-core community framework [45], which enforces strict guidelines for best-practises in software development.

Results and Discussion

Scalability, Portability, and Efficiency

The re-implementation of EAGER into Nextflow offers a range of benefits over the original custom pipeline framework.

Firstly, the new framework provides immediate integration of nf-core/eager into various job schedulers in POSIX High-Performance-Cluster (HPC) environments, cloud computing resources, as well as local workstations. This portability allows users to run nf-core/eager regardless of the type of computing infrastructure or cluster size (if applicable), with minimal effort or configuration. This facilitates reproducibility and therefore maintenance of standards within the field. This is further assisted by the in-built compatibility with software environments and containers such as Conda [46], Docker [47] and Singularity [48]. These are isolated sandboxes that include all software (with exact versions) required by the pipeline, in a form that is installable and runnable by users regardless of the setup of their local software environment. Another major change with nf-core/eager is that the graphical-user-interface (GUI) set up of an EAGER run is now replaced with a command-line-interface (CLI) as the primary user interaction mode. This is more compatible and portable with most HPCs (that may not offer display of a window system), and is in line with the vast majority of bioinformatic tools. We therefore believe this will not be a hindrance to new researchers from outside computational biology. However, a GUI-based pipeline setup is now offered via the nf-core website (<https://nf-co.re/launch>), which provides a common GUI format across multiple pipelines as well as additional robusticity checks of input parameters for those less familiar with CLIs. Typically the output of the launch functionality is a JSON file that can be used with nf-core launch command as a single parameter (as with EAGER v1), however upcoming integration with tower.nf [49] will allow direct submission of pipelines without any command line usage.

Secondly, reproducibility is made easier through the use of ‘profiles’ that can define configuration parameters. These profiles can be managed at different hierarchical levels. HPC-level profiles can specify parameters for the computing environment (job schedulers, cache locations for containers, maximum memory and CPU resources etc.), which can be centrally managed to ensure all users of a group use the same settings. Pipeline-level profiles, specifying parameters for nf-core/eager itself,

allow fast access to routine pipeline-run parameters via a single flag in the nf-core/eager run command, without having to configure each new run from scratch. Compared to the original EAGER, which utilised per-FASTQ XML files with hardcoded filepaths for a specific user's server, nf-core/eager allows researchers to publish the specific profile used in their runs alongside their publications, to ensure other groups can generate the same results. Usage of profiles can also reduce mistakes caused by insufficient 'prose' based reporting of program settings that can be regularly found in the literature. The default nf-core/eager profile uses parameters evaluated in different aDNA-specific contexts (e.g. in [\[50\]](#)), and will be updated in each new release as new studies are published.

Finally, nf-core/eager provides improved efficiency over the original EAGER pipeline by replacing the sample-by-sample sequential processing with Nextflow's asynchronous job parallelisation, whereby multiple pipeline steps and samples are run in parallel (in addition to natively parallelised pipeline steps). This is similar to the approach taken by paleomix, however nf-core/eager expands this by utilising Nextflow ability to customise resource settings the parameters for every job in the pipeline; reducing unnecessary resource allocation that can occur with unfamiliar users to each step of an NGS data processing pipeline. This is particularly pertinent given the increasing use of centralised HPCs or cloud computing that often use per-hour cost calculations.

Updated Workflow

nf-core/eager follows a similar structural foundation to the original version of EAGER and partially with paleomix. Given Illumina short-read FASTQ and/or BAM files and a reference FASTA file, this can be split into four main stages:

1. Pre-processing
 - Sequencing quality control: FastQC [\[51\]](#)
 - Sequencing artefact clean-up (merging, adapter clipping): AdapterRemoval2 [\[52\]](#)
 - Pre-processing statistics generation
2. Mapping and post-processing
 - Alignment against reference genome: BWA [\[40,53\]](#), CircularMapper [\[39\]](#), Bowtie2 [\[55\]](#)
 - Mapping quality filtering: SAMtools [\[56\]](#)
 - PCR duplicate removal: DeDup [\[39\]](#), Picard MarkDuplicates [\[57\]](#)
 - Mapping statistics generation: PreSeq [\[58\]](#), Qualimap2 [\[59\]](#)
3. aDNA Evaluation and Modification
 - Damage profiling: DamageProfiler [\[60\]](#)
 - aDNA reads selection: PMDtools [\[61\]](#)
 - Damage removal: Bamutils [\[62\]](#)
4. Genotyping and Consensus Sequence Generation: GATK [\[57\]](#), ANGSD [\[63\]](#) VCF2Genome [\[39\]](#)

In nf-core/eager, all tools originally used in EAGER have been updated to their latest versions, as available on Bioconda [\[64\]](#) and Conda-forge [\[65\]](#) to ensure widespread accessibility and stability of utilised tools. The MapDamage2 (for damage profile generation) [\[36\]](#) and Schmutzi (for mitochondrial contamination estimation) [\[66\]](#) methods have not been carried over to nf-core/eager, the first because a more performant successor method is now available (DamageProfiler), and the latter because a stable release of the method could not be migrated to Bioconda. We anticipate that there will be an updated version of Schmutzi in the near future that will allow us to integrate the method again in nf-core/eager, once a version is released on Bioconda. As an alternative, estimation of human *nuclear* contamination estimation is now offered through ANGSD [\[63\]](#). Support for the Bowtie2 aligner [\[55\]](#) has been updated to have default settings optimised for aDNA [\[67\]](#).

New tools to the basic workflow include fastp [\[68\]](#) for the removal of 'poly-G' sequencing artefacts that are common in 2-colour Illumina sequencing machines (such as the increasingly popular NextSeq and NovaSeq platforms [\[69\]](#)). For genotyping, we have now included FreeBayes [\[70\]](#) as an alternative

to the human-focused GATK tools, and have also added pileupCaller [71] for generation of genotyping formats commonly utilised in ancient human population analysis. We have also maintained the possibility of using the now unsupported GATK UnifiedGenotyper, as the GATK HaplotypeCaller performs *de novo* assembly around possible variants, which may not be suitable for low-coverage aDNA data.

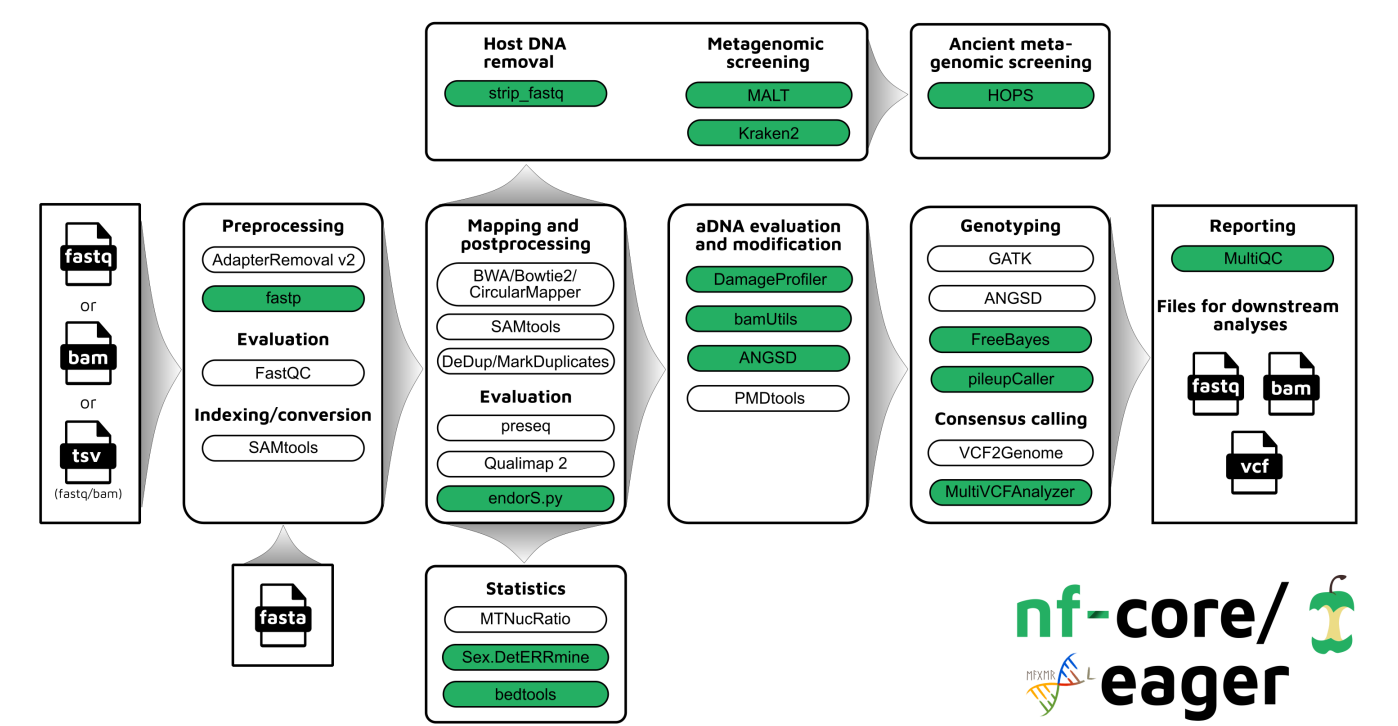


Figure 1: Simplified schematic of the nf-core/eager workflow pipeline. Green filled bubbles indicate new functionality added over the original EAGER pipeline.

We have further extended the functionality of the pipeline by adding ancient metagenomic analysis, to identify the wider taxonomic content of a sample. We have added the possibility to screen all off-target reads (not mapped to the reference genome) with two metagenomic profilers: MALT [72,73] and Kraken2 [74]. Characterisation of properties of authentic aDNA from MALT alignments is carried out with the HOPS pipeline [75]. Ancient metagenomic studies sometimes include comparative samples from living individuals [76]. To support open data, whilst respecting personal data privacy, nf-core/eager includes a ‘strip_fastq’ script which creates raw FASTQ files, but with all reads successfully mapped to the reference genome removed. This allows safe upload of sequencing data to public repositories after removal of identifiable (human) data.

Table 1: Comparison of usage and pipeline functionality of common ancient DNA processing pipelines. Tick represents full functionality, tilde represents partial functionality cross represents not implemented. .

Functionality	EAGER	paleomix	nf-core/eager
Command line setup	✗	✓	✓
GUI Setup	✓	✗	✓
Reproducible software environments offered	✓	✗	✓
HPC scheduler integration	✗	✗	✓
Cloud computing integration	✗	✗	✓
Per-process resource optimisation	✗	~	✓
Pipeline-step parallelisation	✗	✓	✓

Functionality	EAGER	paleomix	nf-core/eager
Sequencing lane merging	✓	✓	✓
Sequencing quality control	✓	✗	✓
Sequencing artefact removal	✗	✗	✓
Adapter clipping/read merging	✓	✓	✓
Post-processing sequencing QC	✗	✗	✓
Reference mapping	✓	✓	✓
Reference mapping statistics	✓	✓	✓
Multi-reference mapping	✗	✓	✗
Mapped reads filtering	✓	✓	✓
Off-target read treatment	✓	✗	✓
Off-target metagenomic profiling	✗	✗	✓
Off-target metagenomic authentication	✗	✗	✓
Library complexity estimation	✓	✗	✓
Duplicate removal	✓	✗	✓
BAM merging	✗	✓	✓
Damage read filtering	✓	✗	✓
Contamination estimation (Human)	✓	✗	✓
Biological sex determination (Human)	✗	✗	✓
Genome coverage estimation	✓	✓	✓
Damage calculation	✓	✓	✓
Damage rescaling	✗	✓	~
SNP Calling/Genotyping	✓	~	✓
Consensus sequence generation	✓	~	✓
Regions of interest statistics	~	✓	✓

Additional functionality tailored for ancient bacterial genomics includes integration of a SNP alignment generation tool, MultiVCFAnalyzer [9], which allows assessment of cross-mapping levels from different related taxa to a reference genome - a common challenge in ancient bacterial genome reconstruction [35]. The output SNP alignment FASTA file can then be used for downstream analyses such as phylogenetic tree construction. Simple coverage statistics of particular annotations (e.g. genes) of an input reference is offered by bedtools [77], which can be used in cases such as for determining functional differences between ancient bacterial strains (as in [42]). When using a human reference genome, nf-core/eager can now also give estimates of the biological sex of a given individual with Sex.DetERRmine [78]. A dedicated 'endogenous DNA' calculator (endorS.py) is also included, to provide a percentage estimate of the sequenced reads matching the reference from the total number of reads sequenced per library.

Given the large amount of sequencing often required to yield sufficient genome coverage from aDNA data, palaeogeneticists tend to use multiple (differently treated) libraries, and/or merge data from multiple sequencing runs per library. The original EAGER pipeline could only run single libraries at a time, and in these contexts required significant manual user input in merging different FASTQ or BAM files. A major upgrade in nf-core/eager is that the new pipeline supports automated processing of

complex sequencing strategies for many samples. This is facilitated by the optional use as input of a simple table in TSV format, which includes file paths and additional metadata such as sample name, library name, sequencing lane, colour chemistry, and UDG treatment. This allows simultaneous processing and appropriate merging of heterogeneous data from multiple sequencing runs and/or library types.

The original EAGER pipeline required users to look through many independent output directories and files to make full assessment of their sequencing data. This has now been replaced with a much more extensive MultiQC [79] report. This tool aggregates the log files of every supported tool into a single interactive report, and assists users in making a fuller assessment of their sequencing and analysis runs. We have strived to ensure that there is a corresponding MultiQC module for every tool used by nf-core/eager where possible, to enable comprehensive evaluation of all stages of the pipeline.

An overview of the entire pipeline is shown in Fig. 1, and a tabular comparison of functionality between EAGER, paleomix and nf-core/eager in 1.

Accessibility

Alongside the interactive MultiQC report, we have written extensive documentation on all parts of running and interpreting the output of the pipeline. Given that a large fraction of aDNA researchers come from fields outside computational biology, and thus may have limited computational training, we have written documentation [80] that also gives guidance on how to interpret each section of the report - specifically in the context of NGS and aDNA. This includes best practices or expected output schematic images, which are published under CC-BY licenses to allow for use in other training material (an example can be seen in Fig. 2). We hope this open-access resource will make the study of aDNA more accessible to researchers new to the field, by providing practical guidelines on how to evaluate characteristics and effects of aDNA on downstream analyses.

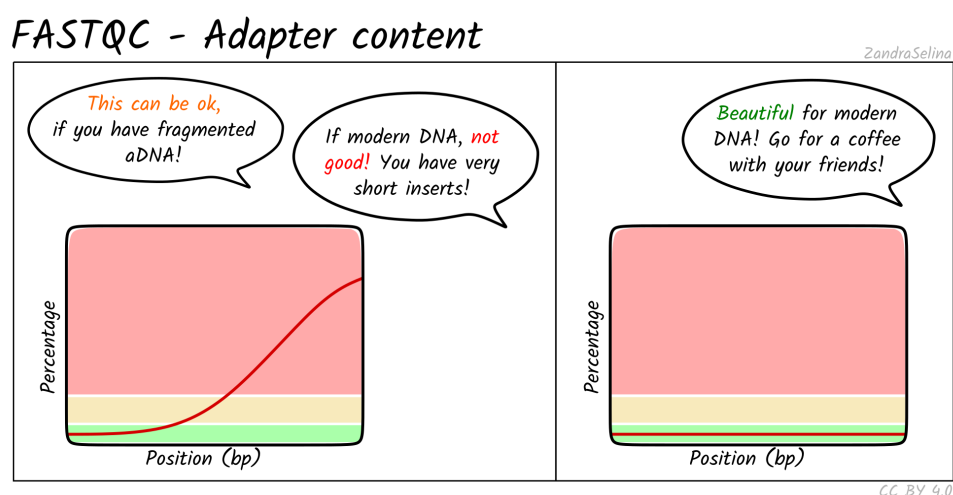


Figure 2: Example schematic images of pipeline output documentation that can assist new users in the interpretation to next-generation-sequencing aDNA processing.

The development of nf-core/eager in Nextflow and the nf-core initiative will also improve open-source development, while ensuring the high quality of community contributions to the pipeline. While Nextflow is written primarily in Groovy, the Nextflow DSL simplifies a number of concepts to an intermediate level that bioinformaticians without Java/Groovy experience can easily access (regardless of own programming language experience). Furthermore, Nextflow places ubiquitous and more widely known command-line interfaces, such as bash, in a prominent position within the code, rather than custom java code and classes. We hope this will motivate further bug fixes and feature contributions from the community, to keep the pipeline state-of-the-art and ensure a longer life-cycle.

This will also be supported by the active and welcoming nf-core community who provide general guidance and advice on developing Nextflow and nf-core pipelines.

Comparisons with other pipelines

We compared pipeline run-times of the three pipelines to show that the new implementation of nf-core/eager is equivalent or more efficient than EAGER or paleomix. While similar pipelines designed for aDNA have also been released, such as ATLAS [81] or HOPS [???], these generally have been designed with specific contexts in mind (e.g. human population genetics or pathogen screening) and do not include many of the preprocessing steps required for analysing NGS data.

We ran each pipeline on a subset of Viking-age cod (*Gadus morhua*) genomic data from [4]. This data was originally run using paleomix, and was re-run here as described but with paleomix (v1.2.14), with settings for the other two pipelines as close as possible to the original paper (EAGER with v1.92.33, and nf-core/EAGER with v2.2.0dev, commit **XXX**). Benchmarking environment and exact pipeline run settings can be seen in **METHODS** and **SUPPLEMENTARY INFORMATION**. The benchmarking environment was set using the Ubuntu 18.04 operating system on a 32 CPU and 256 GB RAM VM instance of the cloud service from the GWDG **MORE DEETZ HERE**. Resource parameters of each tool were only modified to specify the maximum available on the server and otherwise left as default. Two samples with three paired-end sequencing runs were analysed, with adapter clipping and merging (AdapterRemoval), mapping (bwa aln), duplicate removal (markduplicates) and damage profiling (paleomix: mapDamage2, EAGER and nf-core/EAGER: DamageProfiler) steps performed. We ran the commands for each tool sequentially, but repeated these batch of commands 10 times - to account for variability in the cloud service IO connection. Run times were measured using the GNU time tool (v1.7).

Table 2: Comparison of run times between three ancient DNA pipelines. paleomix has an additional run with 'optimised' parameters with fairer resources for comparison .

Tool	Version	Environment	Mean Run Time	SD Run Time
EAGER	1.92.33	singularity		
paleomix	1.2.14	conda		
paleomix (optimised)	1.2.14	conda		
nf-core/eager	2.2.0dev	conda		

The results of the benchmarking tests can be seen in Table 2. Due to 'unfair' defaults of paleomix of only running mapping with bwa on 1 thread, we also re-ran paleomix with the flag `-max-bwa-threads` set to 4 (listed in Table 2 as 'optimised'), which is equivalent to the bwa process of nf-core/eager.

Conclusion

nf-core/eager is an efficient, portable, and accessible pipeline for processing ancient genomic data. This re-implementation of EAGER into Nextflow and nf-core will improve reproducibility and inclusion of rapidly increasing aDNA datasets, for both large and small laboratories. Extensive documentation also enables newcomers to the field to get a practical understanding on how to interpret aDNA in the context of NGS data processing. Ultimately, nf-core/eager provides easier access to the latest tools and routine screening analyses commonly used in the field, and sets up the pipeline for remaining at the forefront of palaeogenetic analysis.

Methods

Installation

nf-core/eager requires a version of Java, Nextflow and either a functional Conda installation *or* Docker/Singularity container installation. A quick installation guide to follow to get started can be found in the *Quickstart* section of the nf-core/eager repository [\[82\]](#).

Running

After the installation, users can run the pipeline using standard test data by utilising some of the test profiles we provide (e.g. using Docker):

```
nextflow run nf-core/eager -r 2.1.0 -profile test,docker
```

This will download test data automatically, run the pipeline locally with all software tools containerised in a Docker image and store the output of that run in the a './results' folder of your current directory.

The default pipeline settings assumes paired end FASTQ data, and will run:

- FastQC
- AdapterRemoval2 (merging and adapter clipping)
- post-clipping FastQC (for AdapterRemoval2 performance evaluation)
- bwa mapping (with the 'aln' algorithm)
- samtools flagstat (for mapping statistics)
- endorS.py (for endogenous DNA calculation)
- DeDup (for PCR amplicon deduplication)
- PreSeq (for library complexity evaluation)
- DamageProfiler and Qualimap2 (for genome coverage statistics)
- MultiQC pipeline run report

If no additional FASTA indices are given, these will also be generated.

The pipeline is highly configurable and most modules can be turned on-and-off using different flags at the request of the user, to allow high customisation to each user's needs. For example, to include metagenomic screening of off-target reads, and sex determination based on on-target mappings of pre-clipped single-end data:

```
nextflow run nf-core/eager -r 2.1.0 \  
-profile conda \  
--input '<path>/<to>/*/R1*.fastq.gz' --single_end \  
--fasta '<path>/<to>/<reference>.fasta.gz' \  
--skip_fastqc --skip_adapterremoval \  
--run_bam_filtering --bam_discard_unmapped --bam_unmapped_type 'fastq' \  
--run_metagenomic_screening --metagenomic_tool 'malt' --database  
    '<path>/<to>/<malt_database>' \  
--run_sexdeterrmine
```

Profiles

In addition to private locally defined profiles, we utilise a central configuration repository to enable users from various institutions to use pipelines on their particular infrastructure more easily [83]. There are multiple resources listed in this repository with information on how to add your own configuration profile with help from the nf-core community.

Users can customise this infrastructure profile by themselves, with the nf-core community, or with their local system administrator to make sure that the pipeline runs successfully, and can then rely on the Nextflow and nf-core framework to ensure compatibility upon further infrastructure changes. For example, in order to run the nf-core/eager pipeline at the Max Planck Institute for the Science of Human History (MPI-SHH) in Jena, users only have to run:

```
nextflow run nf-core/eager -r 2.1.0 -profile shh_cdag,test
```

This runs the testing profile of the nf-core/eager pipeline with parameters specifically adapted to the HPC system at the MPI-SHH. In some cases, similar institutional configs for other institutions may already exist (originally utilised for different nf-core pipelines), so users need not write their own.

Inputs

The pipeline can be started using (raw) FASTQ files from sequencing or pre-mapped BAM files. Additionally, the pipeline requires a FASTA reference genome. If BAM input is provided, an optional conversion to FASTQ is offered, otherwise BAM files processing will start from the post-mapping stage.

If users have complex set-ups, e.g. multiple sequencing lanes that require merging of files, the pipeline can be supplied with a tab separated value (TSV) file to enable such complex data handling. Both FASTQs and BAMs can be provided in this set up. FASTQs with the same library name and sequencing chemistry but sequenced across multiple lanes will be concatenated after adapter removal and prior mapping. Libraries with the sample name and with the same UDG treatment, will be merged after deduplication. If libraries with the sample name have different UDG treatment, these will be merged after the aDNA modification stage (i.e. BAM trimming or PMDtools, if turned on), prior to genotyping, as shown in Figure 3.

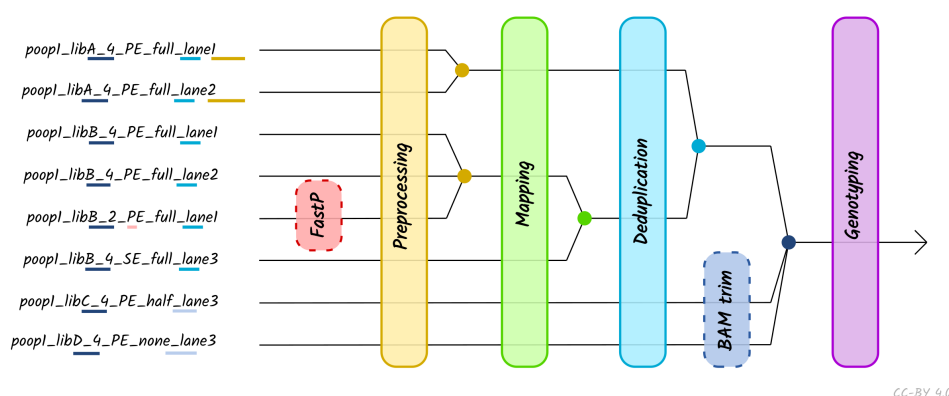


Figure 3: Schematic of different processing and merging points based on the nature of different libraries, as specified in metadata of a TSV file. Dashed boxes represent optional library-specific processes

As Nextflow will automatically download files from URLs, profiles and/or TSV files can include links to publicly available data (e.g. the ENA FTP server). This assists in reproducibility, because if profiles or TSV files are uploaded with a publication, a researcher wishing to re-analyse the data in the same way can use the exact settings and merging procedures in the original publication, without having to reconstruct this from prose.

Monitoring

Users can either monitor their pipeline execution with the messages Nextflow prints to the console while running, or utilise projects such as Nextflow Tower [49] to monitor their analysis pipeline during runtime.

Output

The pipeline produces a multitude of output files in various file formats, with a more detailed listing available in the user documentation. These include metrics, statistical analysis data, and standardised output files (BAM, VCF) for close inspection and further downstream analysis, as well as a MultiQC report. If an emailing daemon is set up on the server, the latter can be emailed to users automatically, when starting the pipeline with a dedicated option (--email you@yourdomain.org).

Data and software availability

All code is available on github at <https://github.com/nf-core/eager> and archived with Zenodo under the DOI [10.5281/zenodo.1465061](https://doi.org/10.5281/zenodo.1465061). The version of nf-core/eager that this preprint is based on is the current 'dev' branch of the GitHub repository (2.2.0dev), and on publication will be released as 2.2.0.

This paper was written with Manubot [84].

Competing Interests

No competing interests are declared.

Acknowledgements

We thank the nf-core community for general support and suggestions during the writing of the pipeline. We also thank Arielle Munters, Hester van Schalkwyk, Irina Velsko, Katherine Eaton, Luc Venturini, Marcel Keller, Pierre Lindenbaum, Pontus Skoglund, Raphael Eisenhofer, Torsten Günter, and Kevin Lord for bug reports and feature suggestions. We are grateful to the members of the Department of Archaeogenetics at the Max Planck Institute for the Science of Human History who performed beta testing of the pipeline. We thank the aDNA twitter community for responding to polls regarding design decisions during development.

The GWDG kindly provided computational infrastructure for benchmarking. We also want to thank Selina Carlhoff, Alexander Herbig and Wolfgang Haak for providing comments and suggestions on this manuscript, and acknowledge Christina Warinner, Stephan Schiffels and the Max Planck Society who provided funds for travel to nf-core events. This project was also supported by the ERC Starting Grant project (FoodTransforms) ERC-2015-StG 678901 funded by the European Research Council awarded to Philipp W. Stockhammer (Ludwig Maximilian University, Munich).

References

- 1. Complete Genomes Reveal Signatures of Demographic and Genetic Declines in the Woolly Mammoth**
Eleftheria Palkopoulou, Swapan Mallick, Pontus Skoglund, Jacob Enk, Nadin Rohland, Heng Li, Ayça Omrak, Sergey Vartanyan, Hendrik Poinar, Anders Götherström, ... Love Dalén
Current Biology (2015-05) <https://doi.org/34d>
DOI: [10.1016/j.cub.2015.04.007](https://doi.org/10.1016/j.cub.2015.04.007) · PMID: [25913407](https://pubmed.ncbi.nlm.nih.gov/25913407/) · PMCID: [PMC4439331](https://pubmed.ncbi.nlm.nih.gov/PMC4439331/)
- 2. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse**
Ludovic Orlando, Aurélien Ginolhac, Guojie Zhang, Duane Froese, Anders Albrechtsen, Mathias Stiller, Mikkel Schubert, Enrico Cappellini, Bent Petersen, Ida Moltke, ... Eske Willerslev
Nature (2013-06-26) <https://doi.org/q7n>
DOI: [10.1038/nature12323](https://doi.org/10.1038/nature12323) · PMID: [23803765](https://pubmed.ncbi.nlm.nih.gov/23803765/)
- 3. Ancient pigs reveal a near-complete genomic turnover following their introduction to Europe**
Laurent A. F. Frantz, James Haile, Audrey T. Lin, Amelie Scheu, Christina Geörg, Norbert Benecke, Michelle Alexander, Anna Linderholm, Victoria E. Mullin, Kevin G. Daly, ... Greger Larson
Proceedings of the National Academy of Sciences (2019-08-27) <https://doi.org/gf9hnf>
DOI: [10.1073/pnas.1901169116](https://doi.org/10.1073/pnas.1901169116) · PMID: [31405970](https://pubmed.ncbi.nlm.nih.gov/31405970/) · PMCID: [PMC6717267](https://pubmed.ncbi.nlm.nih.gov/PMC6717267/)
- 4. Ancient DNA reveals the Arctic origin of Viking Age cod from Haithabu, Germany**
Bastiaan Star, Sanne Boessenkool, Agata T. Gondek, Elena A. Nikulina, Anne Karin Hufthammer, Christophe Pampoulie, Halvor Knutsen, Carl André, Heidi M. Nistelberger, Jan Dierking, ... James H. Barrett
Proceedings of the National Academy of Sciences (2017-08-22) <https://doi.org/gbt8b2>
DOI: [10.1073/pnas.1710186114](https://doi.org/10.1073/pnas.1710186114) · PMID: [28784790](https://pubmed.ncbi.nlm.nih.gov/28784790/) · PMCID: [PMC5576834](https://pubmed.ncbi.nlm.nih.gov/PMC5576834/)
- 5. 137 ancient human genomes from across the Eurasian steppes**
Peter de Barros Damgaard, Nina Marchi, Simon Rasmussen, Michaël Peyrot, Gabriel Renaud, Thorfinn Korneliussen, J. Víctor Moreno-Mayar, Mikkel Winther Pedersen, Amy Goldberg, Emma Usmanova, ... Eske Willerslev
Nature (2018-05-09) <https://doi.org/gd8hs5>
DOI: [10.1038/s41586-018-0094-2](https://doi.org/10.1038/s41586-018-0094-2) · PMID: [29743675](https://pubmed.ncbi.nlm.nih.gov/29743675/)
- 6. A Draft Sequence of the Neandertal Genome**
R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Y. Fritz, ... S. Paabo
Science (2010-05-06) <https://doi.org/c2x>
DOI: [10.1126/science.1188021](https://doi.org/10.1126/science.1188021) · PMID: [20448178](https://pubmed.ncbi.nlm.nih.gov/20448178/) · PMCID: [PMC5100745](https://pubmed.ncbi.nlm.nih.gov/PMC5100745/)
- 7. A High-Coverage Genome Sequence from an Archaic Denisovan Individual**
M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prufer, C. de Filippo, ... S. Paabo
Science (2012-08-30) <https://doi.org/q8p>
DOI: [10.1126/science.1224344](https://doi.org/10.1126/science.1224344) · PMID: [22936568](https://pubmed.ncbi.nlm.nih.gov/22936568/) · PMCID: [PMC3617501](https://pubmed.ncbi.nlm.nih.gov/PMC3617501/)
- 8. The genome of the offspring of a Neanderthal mother and a Denisovan father**
Viviane Slon, Fabrizio Mafessoni, Benjamin Vernot, Cesare de Filippo, Steffi Grote, Bence Viola, Mateja Hajdinjak, Stéphane Peyrégne, Sarah Nagel, Samantha Brown, ... Svante Pääbo
Nature (2018-08-22) <https://doi.org/cs64>
DOI: [10.1038/s41586-018-0455-x](https://doi.org/10.1038/s41586-018-0455-x) · PMID: [30135579](https://pubmed.ncbi.nlm.nih.gov/30135579/) · PMCID: [PMC6130845](https://pubmed.ncbi.nlm.nih.gov/PMC6130845/)

9. **Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis**
Kirsten I. Bos, Kelly M. Harkins, Alexander Herbig, Mireia Coscolla, Nico Weber, Iñaki Comas, Stephen A. Forrest, Josephine M. Bryant, Simon R. Harris, Verena J. Schuenemann, ... Johannes Krause
Nature (2014-08-20) <https://doi.org/f6nk4g>
DOI: [10.1038/nature13591](https://doi.org/10.1038/nature13591) · PMID: [25141181](https://pubmed.ncbi.nlm.nih.gov/25141181/) · PMCID: [PMC4550673](https://pubmed.ncbi.nlm.nih.gov/PMC4550673/)
10. **Integrative approach using *Yersinia pestis* genomes to revisit the historical landscape of plague during the Medieval Period**
Amine Namouchi, Meriam Guellil, Oliver Kersten, Stephanie Hänsch, Claudio Ottoni, Boris V. Schmid, Elsa Pacciani, Luisa Quaglia, Marco Vermunt, Egil L. Bauer, ... Barbara Bramanti
Proceedings of the National Academy of Sciences (2018-12-11) <https://doi.org/ggfn3h>
DOI: [10.1073/pnas.1812865115](https://doi.org/10.1073/pnas.1812865115) · PMID: [30478041](https://pubmed.ncbi.nlm.nih.gov/30478041/) · PMCID: [PMC6294933](https://pubmed.ncbi.nlm.nih.gov/PMC6294933/)
11. **Ancient genomes reveal a high diversity of *Mycobacterium leprae* in medieval Europe**
Verena J. Schuenemann, Charlotte Avanzi, Ben Krause-Kyora, Alexander Seitz, Alexander Herbig, Sarah Inskip, Marion Bonazzi, Ella Reiter, Christian Urban, Dorthe Dangvard Pedersen, ... Johannes Krause
PLOS Pathogens (2018-05-10) <https://doi.org/gdrj4v>
DOI: [10.1371/journal.ppat.1006997](https://doi.org/10.1371/journal.ppat.1006997) · PMID: [29746563](https://pubmed.ncbi.nlm.nih.gov/29746563/) · PMCID: [PMC5944922](https://pubmed.ncbi.nlm.nih.gov/PMC5944922/)
12. **Ancient hepatitis B viruses from the Bronze Age to the Medieval period**
Barbara Mühlemann, Terry C. Jones, Peter de Barros Damgaard, Morten E. Allentoft, Irina Shevnina, Andrey Logvin, Emma Usmanova, Irina P. Panyushkina, Bazartseren Boldgiv, Tsevel Bazartseren, ... Eske Willerslev
Nature (2018-05-09) <https://doi.org/gddxvj>
DOI: [10.1038/s41586-018-0097-z](https://doi.org/10.1038/s41586-018-0097-z) · PMID: [29743673](https://pubmed.ncbi.nlm.nih.gov/29743673/)
13. **Neolithic and medieval virus genomes reveal complex evolution of hepatitis B**
Ben Krause-Kyora, Julian Susat, Felix M Key, Denise Kühnert, Esther Bosse, Alexander Immel, Christoph Rinne, Sabin-Christin Kornell, Diego Yepes, Sören Franzenburg, ... Johannes Krause
eLife (2018-05-10) <https://doi.org/gdhck2>
DOI: [10.7554/elife.36666](https://doi.org/10.7554/elife.36666) · PMID: [29745896](https://pubmed.ncbi.nlm.nih.gov/29745896/) · PMCID: [PMC6008052](https://pubmed.ncbi.nlm.nih.gov/PMC6008052/)
14. **Ancient reveals the timing and persistence of organellar genetic bottlenecks over 3,000 years of sunflower domestication and improvement**
Nathan Wales, Melis Akman, Ray H. B. Watson, Fátima Sánchez Barreiro, Bruce D. Smith, Kristen J. Gremillion, M. Thomas P. Gilbert, Benjamin K. Blackman
Evolutionary Applications (2018-02-13) <https://doi.org/gf568v>
DOI: [10.1111/eva.12594](https://doi.org/10.1111/eva.12594) · PMID: [30622634](https://pubmed.ncbi.nlm.nih.gov/30622634/) · PMCID: [PMC6304678](https://pubmed.ncbi.nlm.nih.gov/PMC6304678/)
15. **The origins and adaptation of European potatoes reconstructed from historical genomes**
Rafal M. Gutaker, Clemens L. Weiß, David Ellis, Noelle L. Anglin, Sandra Knapp, José Luis Fernández-Alonso, Salomé Prat, Hernán A. Burbano
Nature Ecology & Evolution (2019-06-24) <https://doi.org/ggxkk8>
DOI: [10.1038/s41559-019-0921-3](https://doi.org/10.1038/s41559-019-0921-3) · PMID: [31235927](https://pubmed.ncbi.nlm.nih.gov/31235927/)
16. **The *Prevotella copri* Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations**
Adrian Tett, Kun D. Huang, Francesco Asnicar, Hannah Fehlner-Peach, Edoardo Pasolli, Nicolai Karcher, Federica Armanini, Paolo Manghi, Kevin Bonham, Moreno Zolfo, ... Nicola Segata

Cell Host & Microbe (2019-11) <https://doi.org/ggc9dc>
DOI: [10.1016/j.chom.2019.08.018](https://doi.org/10.1016/j.chom.2019.08.018) · PMID: [31607556](https://pubmed.ncbi.nlm.nih.gov/31607556/) · PMCID: [PMC6854460](https://pubmed.ncbi.nlm.nih.gov/PMC6854460/)

17. CoproID predicts the source of coprolites and paleofeces using microbiome composition and host DNA content

Maxime Borry, Bryan Cordova, Angela Perri, Marsha Wibowo, Tanvi Prasad Honap, Jada Ko, Jie Yu, Kate Britton, Linus Girdland-Flink, Robert C. Power, ... Christina Warinner
PeerJ (2020-04-17) <https://doi.org/dr8x>
DOI: [10.7717/peerj.9001](https://doi.org/10.7717/peerj.9001) · PMID: [32337106](https://pubmed.ncbi.nlm.nih.gov/32337106/) · PMCID: [PMC7169968](https://pubmed.ncbi.nlm.nih.gov/PMC7169968/)

18. Pathogens and host immunity in the ancient human oral cavity

Christina Warinner, João F Matias Rodrigues, Rounak Vyas, Christian Trachsel, Natallia Shved, Jonas Grossmann, Anita Radini, Y Hancock, Raul Y Tito, Sarah Fiddyment, ... Enrico Cappellini
Nature Genetics (2014-02-23) <https://doi.org/r4n>
DOI: [10.1038/ng.2906](https://doi.org/10.1038/ng.2906) · PMID: [24562188](https://pubmed.ncbi.nlm.nih.gov/24562188/) · PMCID: [PMC3969750](https://pubmed.ncbi.nlm.nih.gov/PMC3969750/)

19. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus

Laura S. Weyrich, Sebastian Duchene, Julien Soubrier, Luis Arriola, Bastien Llamas, James Breen, Alan G. Morris, Kurt W. Alt, David Caramelli, Veit Dresely, ... Alan Cooper
Nature (2017-03-08) <https://doi.org/f9szrm>
DOI: [10.1038/nature21674](https://doi.org/10.1038/nature21674) · PMID: [28273061](https://pubmed.ncbi.nlm.nih.gov/28273061/)

20. Fifty thousand years of Arctic vegetation and megafaunal diet

Eske Willerslev, John Davison, Mari Moora, Martin Zobel, Eric Coissac, Mary E. Edwards, Eline D. Lorenzen, Mette Vestergård, Galina Gussarova, James Haile, ... Pierre Taberlet
Nature (2014-02-05) <https://doi.org/f2zr4s>
DOI: [10.1038/nature12921](https://doi.org/10.1038/nature12921) · PMID: [24499916](https://pubmed.ncbi.nlm.nih.gov/24499916/)

21. Neandertal and Denisovan DNA from Pleistocene sediments

Viviane Slon, Charlotte Hopfe, Clemens L. Weiß, Fabrizio Mafessoni, Marco de la Rasilla, Carles Lalueza-Fox, Antonio Rosas, Marie Soressi, Monika V. Knul, Rebecca Miller, ... Matthias Meyer
Science (2017-05-12) <https://doi.org/b6jd>
DOI: [10.1126/science.aam9695](https://doi.org/10.1126/science.aam9695) · PMID: [28450384](https://pubmed.ncbi.nlm.nih.gov/28450384/)

22. Plasmodium vivax Malaria Viewed through the Lens of an Eradicated European Strain

Lucy van Dorp, Pere Gelabert, Adrien Rieux, Marc de Manuel, Toni de-Dios, Shyam Gopalakrishnan, Christian Carøe, Marcela Sandoval-Velasco, Rosa Fregel, Iñigo Olalde, ... Carles Lalueza-Fox
Molecular Biology and Evolution (2020-03) <https://doi.org/ggqzq2>
DOI: [10.1093/molbev/msz264](https://doi.org/10.1093/molbev/msz264) · PMID: [31697387](https://pubmed.ncbi.nlm.nih.gov/31697387/) · PMCID: [PMC7038659](https://pubmed.ncbi.nlm.nih.gov/PMC7038659/)

23. Paging through history: parchment as a reservoir of ancient DNA for next generation sequencing

M. D. Teasdale, N. L. van Doorn, S. Fiddyment, C. C. Webb, T. O'Connor, M. Hofreiter, M. J. Collins, D. G. Bradley
Philosophical Transactions of the Royal Society B: Biological Sciences (2015-01-19)
<https://doi.org/ggqzq3>
DOI: [10.1098/rstb.2013.0379](https://doi.org/10.1098/rstb.2013.0379) · PMID: [25487331](https://pubmed.ncbi.nlm.nih.gov/25487331/) · PMCID: [PMC4275887](https://pubmed.ncbi.nlm.nih.gov/PMC4275887/)

24. A 5700 year-old human genome and oral microbiome from chewed birch pitch

Theis Z. T. Jensen, Jonas Niemann, Katrine Højholt Iversen, Anna K. Fotakis, Shyam Gopalakrishnan, Åshild J. Vågene, Mikkel Winther Pedersen, Mikkel-Holger S. Sinding, Martin R. Ellegaard, Morten E. Allentoft, ... Hannes Schroeder

Nature Communications (2019-12-17) <https://doi.org/ggfm6x>
DOI: [10.1038/s41467-019-13549-9](https://doi.org/10.1038/s41467-019-13549-9) · PMID: [31848342](https://pubmed.ncbi.nlm.nih.gov/31848342/) · PMCID: [PMC6917805](https://pubmed.ncbi.nlm.nih.gov/PMC6917805/)

25. Ancient DNA from mastics solidifies connection between material culture and genetics of mesolithic hunter-gatherers in Scandinavia

Natalija Kashuba, Emrah Kirdök, Hege Damlien, Mikael A. Manninen, Bengt Nordqvist, Per Persson, Anders Götherström

Communications Biology (2019-05-15) <https://doi.org/gggzqz>
DOI: [10.1038/s42003-019-0399-1](https://doi.org/10.1038/s42003-019-0399-1) · PMID: [31123709](https://pubmed.ncbi.nlm.nih.gov/31123709/) · PMCID: [PMC6520363](https://pubmed.ncbi.nlm.nih.gov/PMC6520363/)

26. The Beaker phenomenon and the genomic transformation of northwest Europe

Iñigo Olalde, Selina Brace, Morten E. Allentoft, Ian Armit, Kristian Kristiansen, Thomas Booth, Nadin Rohland, Swapan Mallick, Anna Szécsényi-Nagy, Alissa Mittnik, ... David Reich

Nature (2018-02-21) <https://doi.org/gcx74m>
DOI: [10.1038/nature25738](https://doi.org/10.1038/nature25738) · PMID: [29466337](https://pubmed.ncbi.nlm.nih.gov/29466337/) · PMCID: [PMC5973796](https://pubmed.ncbi.nlm.nih.gov/PMC5973796/)

27. The genomic history of southeastern Europe

Iain Mathieson, Songül Alpaslan-Roodenberg, Cosimo Posth, Anna Szécsényi-Nagy, Nadin Rohland, Swapan Mallick, Iñigo Olalde, Nasreen Broomandkhoshbacht, Francesca Candilio, Olivia Cheronet, ... David Reich

Nature (2018-02-21) <https://doi.org/gc2n9h>
DOI: [10.1038/nature25778](https://doi.org/10.1038/nature25778) · PMID: [29466330](https://pubmed.ncbi.nlm.nih.gov/29466330/) · PMCID: [PMC6091220](https://pubmed.ncbi.nlm.nih.gov/PMC6091220/)

28. A draft genome of *Yersinia pestis* from victims of the Black Death

Kirsten I. Bos, Verena J. Schuenemann, G. Brian Golding, Hernán A. Burbano, Nicholas Waglechner, Brian K. Coombes, Joseph B. McPhee, Sharon N. DeWitte, Matthias Meyer, Sarah Schmedes, ... Johannes Krause

Nature (2011-10-12) <https://doi.org/fk87wk>
DOI: [10.1038/nature10549](https://doi.org/10.1038/nature10549) · PMID: [21993626](https://pubmed.ncbi.nlm.nih.gov/21993626/) · PMCID: [PMC3690193](https://pubmed.ncbi.nlm.nih.gov/PMC3690193/)

29. Bioinformatics Education–Perspectives and Challenges out of Africa

O. Tastan Bishop, E. F. Adebisi, A. M. Alzohairy, D. Everett, K. Ghedira, A. Ghouila, J. Kumuthini, N. J. Mulder, S. Panji, H.-G. Patterson, (for the H3ABioNet Consortium, as members of The H3Africa Consortium)

Briefings in Bioinformatics (2014-07-02) <https://doi.org/f67hix>
DOI: [10.1093/bib/bbu022](https://doi.org/10.1093/bib/bbu022) · PMID: [24990350](https://pubmed.ncbi.nlm.nih.gov/24990350/) · PMCID: [PMC4364068](https://pubmed.ncbi.nlm.nih.gov/PMC4364068/)

30. Highlights on the Application of Genomics and Bioinformatics in the Fight Against Infectious Diseases: Challenges and Opportunities in Africa

Saikou Y. Bah, Collins Misita Morang'a, Jonas A. Kengne-Ouafo, Lucas Amenga-Etego, Gordon A. Awandare

Frontiers in Genetics (2018-11-27) <https://doi.org/gfrxbz>
DOI: [10.3389/fgene.2018.00575](https://doi.org/10.3389/fgene.2018.00575) · PMID: [30538723](https://pubmed.ncbi.nlm.nih.gov/30538723/) · PMCID: [PMC6277583](https://pubmed.ncbi.nlm.nih.gov/PMC6277583/)

31. Instability and decay of the primary structure of DNA

Tomas Lindahl

Nature (1993-04) <https://doi.org/d9c9vq>
DOI: [10.1038/362709a0](https://doi.org/10.1038/362709a0) · PMID: [8469282](https://pubmed.ncbi.nlm.nih.gov/8469282/)

32. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins

Matthias Meyer, Juan-Luis Arsuaga, Cesare de Filippo, Sarah Nagel, Ayinuer Aximu-Petri, Birgit Nickel, Ignacio Martínez, Ana Gracia, José María Bermúdez de Castro, Eudald Carbonell, ... Svante Pääbo

Nature (2016-03-14) <https://doi.org/bdcn>
DOI: [10.1038/nature17405](https://doi.org/10.1038/nature17405) · PMID: [26976447](https://pubmed.ncbi.nlm.nih.gov/26976447/)

33. Patterns of damage in genomic DNA sequences from a Neandertal

A. W. Briggs, U. Stenzel, P. L. F. Johnson, R. E. Green, J. Kelso, K. Prufer, M. Meyer, J. Krause, M. T. Ronan, M. Lachmann, S. Paabo

Proceedings of the National Academy of Sciences (2007-08-21) <https://doi.org/bs4w7h>
DOI: [10.1073/pnas.0704665104](https://doi.org/10.1073/pnas.0704665104) · PMID: [17715061](https://pubmed.ncbi.nlm.nih.gov/17715061/) · PMCID: [PMC1976210](https://pubmed.ncbi.nlm.nih.gov/PMC1976210/)

34. A new model for ancient DNA decay based on paleogenomic meta-analysis

Logan Kistler, Roselyn Ware, Oliver Smith, Matthew Collins, Robin G. Allaby

Nucleic Acids Research (2017-06-20) <https://doi.org/gf58ts>
DOI: [10.1093/nar/gkx361](https://doi.org/10.1093/nar/gkx361) · PMID: [28486705](https://pubmed.ncbi.nlm.nih.gov/28486705/) · PMCID: [PMC5499742](https://pubmed.ncbi.nlm.nih.gov/PMC5499742/)

35. A Robust Framework for Microbial Archaeology

Christina Warinner, Alexander Herbig, Allison Mann, James A. Fellows Yates, Clemens L. Weiß, Hernán A. Burbano, Ludovic Orlando, Johannes Krause

Annual Review of Genomics and Human Genetics (2017-08-31) <https://doi.org/gf5wqv>
DOI: [10.1146/annurev-genom-091416-035526](https://doi.org/10.1146/annurev-genom-091416-035526) · PMID: [28460196](https://pubmed.ncbi.nlm.nih.gov/28460196/) · PMCID: [PMC5581243](https://pubmed.ncbi.nlm.nih.gov/PMC5581243/)

36. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters

Hákon Jónsson, Aurélien Ginolhac, Mikkel Schubert, Philip L. F. Johnson, Ludovic Orlando

Bioinformatics (2013-07) <https://doi.org/gb5g2t>
DOI: [10.1093/bioinformatics/btt193](https://doi.org/10.1093/bioinformatics/btt193) · PMID: [23613487](https://pubmed.ncbi.nlm.nih.gov/23613487/) · PMCID: [PMC3694634](https://pubmed.ncbi.nlm.nih.gov/PMC3694634/)

37. Early Divergent Strains of *Yersinia pestis* in Eurasia 5,000 Years Ago

Simon Rasmussen, Morten Erik Allentoft, Kasper Nielsen, Ludovic Orlando, Martin Sikora, Karl-Göran Sjögren, Anders Gorm Pedersen, Mikkel Schubert, Alex Van Dam, Christian Mollin Outzen Kapel, ... Eske Willerslev

Cell (2015-10) <https://doi.org/f3mxqd>
DOI: [10.1016/j.cell.2015.10.009](https://doi.org/10.1016/j.cell.2015.10.009) · PMID: [26496604](https://pubmed.ncbi.nlm.nih.gov/26496604/) · PMCID: [PMC4644222](https://pubmed.ncbi.nlm.nih.gov/PMC4644222/)

38. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX

Mikkel Schubert, Luca Ermini, Clio Der Sarkissian, Hákon Jónsson, Aurélien Ginolhac, Robert Schaefer, Michael D Martin, Ruth Fernández, Martin Kircher, Molly McCue, ... Ludovic Orlando

Nature Protocols (2014-04-10) <https://doi.org/f5x3qm>
DOI: [10.1038/nprot.2014.063](https://doi.org/10.1038/nprot.2014.063) · PMID: [24722405](https://pubmed.ncbi.nlm.nih.gov/24722405/)

39. EAGER: efficient ancient genome reconstruction

Alexander Peltzer, Günter Jäger, Alexander Herbig, Alexander Seitz, Christian Kniep, Johannes Krause, Kay Nieselt

Genome Biology (2016-03-31) <https://doi.org/ggqzpk>
DOI: [10.1186/s13059-016-0918-z](https://doi.org/10.1186/s13059-016-0918-z) · PMID: [27036623](https://pubmed.ncbi.nlm.nih.gov/27036623/) · PMCID: [PMC4815194](https://pubmed.ncbi.nlm.nih.gov/PMC4815194/)

40. Fast and accurate short read alignment with Burrows-Wheeler transform

H. Li, R. Durbin

Bioinformatics (2009-05-18) <https://doi.org/dqt59j>
DOI: [10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324) · PMID: [19451168](https://pubmed.ncbi.nlm.nih.gov/19451168/) · PMCID: [PMC2705234](https://pubmed.ncbi.nlm.nih.gov/PMC2705234/)

41. mapDamage: testing for damage patterns in ancient DNA sequences

Aurelien Ginolhac, Morten Rasmussen, M. Thomas P. Gilbert, Eske Willerslev, Ludovic Orlando

Bioinformatics (2011-08-01) <https://doi.org/cn45v7>
DOI: [10.1093/bioinformatics/btr347](https://doi.org/10.1093/bioinformatics/btr347) · PMID: [21659319](https://pubmed.ncbi.nlm.nih.gov/21659319/)

42. The Stone Age Plague and Its Persistence in Eurasia

Aida Andrades Valtueña, Alissa Mittnik, Felix M. Key, Wolfgang Haak, Raili Allmäe, Andrej Belinskij, Mantas Daubaras, Michal Feldman, Rimantas Jankauskas, Ivor Janković, ... Johannes Krause
Current Biology (2017-12) <https://doi.org/cgmv>
DOI: [10.1016/j.cub.2017.10.025](https://doi.org/10.1016/j.cub.2017.10.025) · PMID: [29174893](https://pubmed.ncbi.nlm.nih.gov/29174893/)

43. Novel Substrates as Sources of Ancient DNA: Prospects and Hurdles

Eleanor Green, Camilla Speller
Genes (2017-07-13) <https://doi.org/gf57tz>
DOI: [10.3390/genes8070180](https://doi.org/10.3390/genes8070180) · PMID: [28703741](https://pubmed.ncbi.nlm.nih.gov/28703741/) · PMCID: [PMC5541313](https://pubmed.ncbi.nlm.nih.gov/PMC5541313/)

44. Nextflow enables reproducible computational workflows

Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, Cedric Notredame
Nature Biotechnology (2017-04-11) <https://doi.org/gfj52z>
DOI: [10.1038/nbt.3820](https://doi.org/10.1038/nbt.3820) · PMID: [28398311](https://pubmed.ncbi.nlm.nih.gov/28398311/)

45. The nf-core framework for community-curated bioinformatics pipelines

Philip A. Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, Sven Nahnsen
Nature Biotechnology (2020-02-13) <https://doi.org/ggk3qh>
DOI: [10.1038/s41587-020-0439-x](https://doi.org/10.1038/s41587-020-0439-x) · PMID: [32055031](https://pubmed.ncbi.nlm.nih.gov/32055031/)

46. Conda — Conda documentation <https://docs.conda.io/en/latest/>

47. Empowering App Development for Developers | Docker <https://www.docker.com/>

48. Home

Sylabs.io
<https://sylabs.io/>

49. Nextflow Tower <https://tower.nf/>

50. Improving ancient DNA read mapping against modern reference genomes

Mikkel Schubert, Aurelien Ginolhac, Stinus Lindgreen, John F Thompson, Khaled AS AL-Rasheid, Eske Willerslev, Anders Krogh, Ludovic Orlando
BMC Genomics (2012) <https://doi.org/gb3ff7>
DOI: [10.1186/1471-2164-13-178](https://doi.org/10.1186/1471-2164-13-178) · PMID: [22574660](https://pubmed.ncbi.nlm.nih.gov/22574660/) · PMCID: [PMC3468387](https://pubmed.ncbi.nlm.nih.gov/PMC3468387/)

51. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

52. AdapterRemoval v2: rapid adapter trimming, identification, and read merging

Mikkel Schubert, Stinus Lindgreen, Ludovic Orlando
BMC Research Notes (2016-02-12) <https://doi.org/gfzqhb>
DOI: [10.1186/s13104-016-1900-2](https://doi.org/10.1186/s13104-016-1900-2) · PMID: [26868221](https://pubmed.ncbi.nlm.nih.gov/26868221/) · PMCID: [PMC4751634](https://pubmed.ncbi.nlm.nih.gov/PMC4751634/)

53. Fast and accurate long-read alignment with Burrows–Wheeler transform

Heng Li, Richard Durbin

Bioinformatics (2010-03-01) <https://doi.org/cm27kg>
DOI: [10.1093/bioinformatics/btp698](https://doi.org/10.1093/bioinformatics/btp698) · PMID: [20080505](https://pubmed.ncbi.nlm.nih.gov/20080505/) · PMCID: [PMC2828108](https://pubmed.ncbi.nlm.nih.gov/PMC2828108/)

54. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

Heng Li

arXiv (2013-05-28) <https://arxiv.org/abs/1303.3997>

55. Fast gapped-read alignment with Bowtie 2

Ben Langmead, Steven L Salzberg

Nature Methods (2012-03-04) <https://doi.org/gd2xzn>

DOI: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923) · PMID: [22388286](https://pubmed.ncbi.nlm.nih.gov/22388286/) · PMCID: [PMC3322381](https://pubmed.ncbi.nlm.nih.gov/PMC3322381/)

56. The Sequence Alignment/Map format and SAMtools

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin,
1000 Genome Project Data Processing Subgroup

Bioinformatics (2009-06-08) <https://doi.org/ff6426>

DOI: [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352) · PMID: [19505943](https://pubmed.ncbi.nlm.nih.gov/19505943/) · PMCID: [PMC2723002](https://pubmed.ncbi.nlm.nih.gov/PMC2723002/)

57. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo

Genome Research (2010-07-19) <https://doi.org/bnzbn6>

DOI: [10.1101/gr.107524.110](https://doi.org/10.1101/gr.107524.110) · PMID: [20644199](https://pubmed.ncbi.nlm.nih.gov/20644199/) · PMCID: [PMC2928508](https://pubmed.ncbi.nlm.nih.gov/PMC2928508/)

58. Predicting the molecular complexity of sequencing libraries

Timothy Daley, Andrew D Smith

Nature Methods (2013-02-24) <https://doi.org/gfx6f5>

DOI: [10.1038/nmeth.2375](https://doi.org/10.1038/nmeth.2375) · PMID: [23435259](https://pubmed.ncbi.nlm.nih.gov/23435259/) · PMCID: [PMC3612374](https://pubmed.ncbi.nlm.nih.gov/PMC3612374/)

59. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data

Konstantin Okonechnikov, Ana Conesa, Fernando García-Alcalde

Bioinformatics (2015-10-01) <https://doi.org/ggxrmx>

DOI: [10.1093/bioinformatics/btv566](https://doi.org/10.1093/bioinformatics/btv566) · PMID: [26428292](https://pubmed.ncbi.nlm.nih.gov/26428292/) · PMCID: [PMC4708105](https://pubmed.ncbi.nlm.nih.gov/PMC4708105/)

60. Integrative-Transcriptomics/DamageProfiler: DamageProfiler v0.4.9

Judith Neukamm, Alexander Peltzer

Zenodo (2019-11-29) <https://doi.org/ggxrmz>

DOI: [10.5281/zenodo.3557708](https://doi.org/10.5281/zenodo.3557708)

61. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal

Pontus Skoglund, Bernd H. Northoff, Michael V. Shunkov, Anatoli P. Derevianko, Svante Pääbo, Johannes Krause, Mattias Jakobsson

Proceedings of the National Academy of Sciences (2014-02-11) <https://doi.org/f2z5sw>

DOI: [10.1073/pnas.1318934111](https://doi.org/10.1073/pnas.1318934111) · PMID: [24469802](https://pubmed.ncbi.nlm.nih.gov/24469802/) · PMCID: [PMC3926038](https://pubmed.ncbi.nlm.nih.gov/PMC3926038/)

62. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data

Goo Jun, Mary Kate Wing, Gonçalo R. Abecasis, Hyun Min Kang

Genome Research (2015-06) <https://doi.org/f7dz2d>

DOI: [10.1101/gr.176552.114](https://doi.org/10.1101/gr.176552.114) · PMID: [25883319](https://pubmed.ncbi.nlm.nih.gov/25883319/) · PMCID: [PMC4448687](https://pubmed.ncbi.nlm.nih.gov/PMC4448687/)

63. **ANGSD: Analysis of Next Generation Sequencing Data**
Thorfinn Sand Korneliussen, Anders Albrechtsen, Rasmus Nielsen
BMC Bioinformatics (2014-11-25) <https://doi.org/gb8wpz>
DOI: [10.1186/s12859-014-0356-4](https://doi.org/10.1186/s12859-014-0356-4) · PMID: [25420514](https://pubmed.ncbi.nlm.nih.gov/25420514/) · PMCID: [PMC4248462](https://pubmed.ncbi.nlm.nih.gov/PMC4248462/)
64. **Bioconda: sustainable and comprehensive software distribution for the life sciences**
Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, Johannes Köster, The Bioconda Team
Nature Methods (2018-07-02) <https://doi.org/gd2xzp>
DOI: [10.1038/s41592-018-0046-7](https://doi.org/10.1038/s41592-018-0046-7) · PMID: [29967506](https://pubmed.ncbi.nlm.nih.gov/29967506/)
65. **conda-forge | community driven packaging for conda**<https://conda-forge.org/>
66. **Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA**
Gabriel Renaud, Viviane Slon, Ana T. Duggan, Janet Kelso
Genome Biology (2015-10-12) <https://doi.org/f72mvg>
DOI: [10.1186/s13059-015-0776-0](https://doi.org/10.1186/s13059-015-0776-0) · PMID: [26458810](https://pubmed.ncbi.nlm.nih.gov/26458810/) · PMCID: [PMC4601135](https://pubmed.ncbi.nlm.nih.gov/PMC4601135/)
67. **Assessing DNA Sequence Alignment Methods for Characterizing Ancient Genomes and Methylomes**
Marine Pouillet, Ludovic Orlando
Frontiers in Ecology and Evolution (2020-05-06) <https://doi.org/ggzwqr>
DOI: [10.3389/fevo.2020.00105](https://doi.org/10.3389/fevo.2020.00105)
68. **fastp: an ultra-fast all-in-one FASTQ preprocessor**
Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu
Bioinformatics (2018-09-01) <https://doi.org/gd9mrb>
DOI: [10.1093/bioinformatics/bty560](https://doi.org/10.1093/bioinformatics/bty560) · PMID: [30423086](https://pubmed.ncbi.nlm.nih.gov/30423086/) · PMCID: [PMC6129281](https://pubmed.ncbi.nlm.nih.gov/PMC6129281/)
69. **QC Fail Sequencing » Illumina 2 colour chemistry can overcall high confidence G bases**<https://sequencing.qcfail.com/articles/illumina-2-colour-chemistry-can-overcall-high-confidence-g-bases/>
70. **Haplotype-based variant detection from short-read sequencing**
Erik Garrison, Gabor Marth
arXiv (2012-07-24) <https://arxiv.org/abs/1207.3907>
71. **stschiff/sequenceTools**
Stephan Schiffels
(2020-07-19) <https://github.com/stschiff/sequenceTools>
72. **MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman**
Alexander Herbig, Frank Maixner, Kirsten I. Bos, Albert Zink, Johannes Krause, Daniel H. Huson
bioRxiv (2016-04-27) <https://doi.org/ggxkk9>
DOI: [10.1101/050559](https://doi.org/10.1101/050559)
73. **Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico**
Åshild J. Vågene, Alexander Herbig, Michael G. Campana, Nelly M. Robles García, Christina Warinner, Susanna Sabin, Maria A. Spyrou, Aida Andrades Valtueña, Daniel Huson, Noreen Tuross, ... Johannes Krause

Nature Ecology & Evolution (2018-01-15) <https://doi.org/ggxkk7>
DOI: [10.1038/s41559-017-0446-6](https://doi.org/10.1038/s41559-017-0446-6) · PMID: [29335577](https://pubmed.ncbi.nlm.nih.gov/29335577/)

74. Improved metagenomic analysis with Kraken 2

Derrick E. Wood, Jennifer Lu, Ben Langmead

Genome Biology (2019-11-28) <https://doi.org/ggfk55>

DOI: [10.1186/s13059-019-1891-0](https://doi.org/10.1186/s13059-019-1891-0) · PMID: [31779668](https://pubmed.ncbi.nlm.nih.gov/31779668/) · PMCID: [PMC6883579](https://pubmed.ncbi.nlm.nih.gov/PMC6883579/)

75. HOPS: automated detection and authentication of pathogen DNA in archaeological remains

Ron Hübler, Felix M. Key, Christina Warinner, Kirsten I. Bos, Johannes Krause, Alexander Herbig

Genome Biology (2019-12-16) <https://doi.org/ggxkmb>

DOI: [10.1186/s13059-019-1903-0](https://doi.org/10.1186/s13059-019-1903-0) · PMID: [31842945](https://pubmed.ncbi.nlm.nih.gov/31842945/) · PMCID: [PMC6913047](https://pubmed.ncbi.nlm.nih.gov/PMC6913047/)

76. Microbial differences between dental plaque and historic dental calculus are related to oral biofilm maturation stage

Irina M. Velsko, James A. Fellows Yates, Franziska Aron, Richard W. Hagan, Laurent A. F. Frantz, Louise Loe, Juan Bautista Rodriguez Martinez, Eros Chaves, Chris Gosden, Greger Larson, Christina Warinner

Microbiome (2019-07-06) <https://doi.org/ggxkmc>

DOI: [10.1186/s40168-019-0717-3](https://doi.org/10.1186/s40168-019-0717-3) · PMID: [31279340](https://pubmed.ncbi.nlm.nih.gov/31279340/) · PMCID: [PMC6612086](https://pubmed.ncbi.nlm.nih.gov/PMC6612086/)

77. BEDTools: a flexible suite of utilities for comparing genomic features

Aaron R. Quinlan, Ira M. Hall

Bioinformatics (2010-03-15) <https://doi.org/cmrms3>

DOI: [10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033) · PMID: [20110278](https://pubmed.ncbi.nlm.nih.gov/20110278/) · PMCID: [PMC2832824](https://pubmed.ncbi.nlm.nih.gov/PMC2832824/)

78. Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe

Thiseas C. Lamnidis, Kerttu Majander, Choongwon Jeong, Elina Salmela, Anna Wessman, Vyacheslav Moiseyev, Valery Khartanovich, Oleg Balanovsky, Matthias Ongyerth, Antje Weihmann, ... Stephan Schiffels

Nature Communications (2018-11-27) <https://doi.org/ggxkk6>

DOI: [10.1038/s41467-018-07483-5](https://doi.org/10.1038/s41467-018-07483-5) · PMID: [30479341](https://pubmed.ncbi.nlm.nih.gov/30479341/) · PMCID: [PMC6258758](https://pubmed.ncbi.nlm.nih.gov/PMC6258758/)

79. MultiQC: summarize analysis results for multiple tools and samples in a single report

Philip Ewels, Måns Magnusson, Sverker Lundin, Max Käller

Bioinformatics (2016-10-01) <https://doi.org/f3s996>

DOI: [10.1093/bioinformatics/btw354](https://doi.org/10.1093/bioinformatics/btw354) · PMID: [27312411](https://pubmed.ncbi.nlm.nih.gov/27312411/) · PMCID: [PMC5039924](https://pubmed.ncbi.nlm.nih.gov/PMC5039924/)

80. <https://nf-co.re/eager/docs>

81. ATLAS: Analysis Tools for Low-depth and Ancient Samples

Vivian Link, Athanasios Kousathanas, Krishna Veeramah, Christian Sell, Amelie Scheu, Daniel Wegmann

bioRxiv (2017-03-24) <https://doi.org/gg668z>

DOI: [10.1101/105346](https://doi.org/10.1101/105346)

82. nf-core/eager

nf-core

(2020-07-27) <https://github.com/nf-core/eager>

83. nf-core/configs

nf-core

(2020-08-03) <https://github.com/nf-core/configs>

84. **Open collaborative writing with Manubot**

Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter

PLOS Computational Biology (2019-06-24) <https://doi.org/c7np>

DOI: [10.1371/journal.pcbi.1007128](https://doi.org/10.1371/journal.pcbi.1007128) · PMID: [31233491](https://pubmed.ncbi.nlm.nih.gov/31233491/) · PMCID: [PMC6611653](https://pubmed.ncbi.nlm.nih.gov/PMC6611653/)