The Journal of Open Source Software

PileupCaller: A command-line tool to sample genotypes from low-coverage sequencing data of ancient DNA

Stephan Schiffels [1]

1 Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: AHM Mahfuzur Rahman

Reviewers:

- @dkoptekin
- @josecediel25

Submitted: 09 May 2025
Published: unpublished

Summary

Next generation sequencing data is ubiquitous in medical and biological sciences. It has also become the primary tool in archaeogenetics, where ancient DNA is extracted from archaeological organic (often human skeletal) material, processed into DNA sequencing libraries and then sequenced (Orlando et al., 2021). As a testimony to the rapid and accellerating growth of the field, we today have close to ten thousand published ancient human genomes available in the public record (Mallick et al., 2024; Schmid et al., 2024), and many smaller datasets of other organisms. A key step in processing raw sequencing data is the estimation of genotypes at specific variable positions along the genome. Such positions are often pre-selected because they are informative about ancestry or of particular biological relevance (Haak et al., 2015; Mathieson et al., 2015; Rohland et al., 2022). While established tools exist for this task for high-quality modern sequencing data (Danecek et al., 2021; Van der Auwera & O'Connor, 2020), these are often not appropriate for ancient DNA, which has often too low sequencing-coverage and a higher error rate due to post-mortem DNA damage. PileupCaller is a command-line tool written in Haskell, which randomly samples genotypes from raw alignment data at predefined bi-allelic positions. Several modes can be selected, geared towards specific input data features and research questions.

Statement of need

Present-day DNA, for example from medical studies results in raw sequencing data with relatively low per-base error rates and sequencing-coverages of at least several multiples of 1 (for example (1000 Genomes Project Consortium et al., 2015)) but in fact up to 20-30x coverage. Dedicated tools to process such data include samtools/bcftools (Danecek et al., 2021) and GATK (Van der Auwera & O'Connor, 2020) among many other tools. Ancient DNA seuqencing data often comes with substantially lower coverage and substantially higher error rates. In terms of coverage, most ancient genomes have genome-wide coverage often below 1x and in fact very often even below 0.1x. Such low coverage means that any given genomic site is more likely not covered by a sequencing read than covered. At the same time, the low fraction of sites that is actually covered has higher error rates than modern DNA, due to ancient-DNA damage. These two factors violate the assumptions behind statistical genotype callers like `bcftools call` or `HaplotypeCaller` from GATK.

As is widely used practice in the field, very low-coverage ancient DNA data is often "called", simply by randomly selecting reads at a given position of interest. PileupCaller is a command-line tool that does exactly that, by reading in a list of SNP positions and a stream of sequencing data, some optional filtering options, and then performs random samples at every position of interest for multiple individuals. Even before this paper, `pileupCaller` has been widely used since its creation in 2017, mostly because of its simple use and low-memory footprint thanks to streaming.

Usage and key functionality

pileupCaller relies on the pileup format defined in (Danecek et al., 2021). This format lists for each site the nucleotides from all reads covering that site, including base-qualities. An example pileup-file can be found in the repository. These files are rarely saved, but used as an intermediate format for streaming, specifically from samtools mpileup(). A typical usage command line for pileupCaller could be:

```
samtools mpileup -R -B -q30 -Q30 -f <reference_genome.fasta> \
    Sample1.bam Sample2.bam Sample3.bam | \
pileupCaller --randomHaploid --sampleNames Sample1,Sample2,Sample3 \
    --samplePopName MyPop -f <Eigenstrat.snp> \
    -e <My_output_prefix>
```

Among the options passed to samtools mpileup, we highlight the -B flag, which switches off base-alignment quality recalibration and is critical for avoiding reference bias with low-coverage ancient DNA. A key input ingredient is the SNP list, which needs to be given in Eigenstrat-format (Price et al., 2006), defined in (Mah et al., 2021). Example files for Pileup and Eigenstrat-formatted data can be found under test/testDat in the software repository. The SNP file not only lists the positions of the variants that should be called, but also the two possible alleles at each site. This is important, as pileupCaller will ensure that only those two alleles are called. Sites at which a third allele is called will be output as missing.

In terms of output formats, pileupCaller currently supports Eigenstrat, Plink (https://www.cog-genomics.org/plink/2.0/) and VCF (Danecek et al., 2011), with an option to additionally compress the output in gzip-format. Command line options are documented inline via pileupCaller --help.

PileupCaller is part of the "sequenceTools" package, which contains multiple other minor scripts and command-line tools, with pileupCaller being the central and most popular tool. The sequenceTools package makes key use of the "sequence-formats" Haskell library (Schiffels, 2025), which contains parsers for the Pileup-, the Plink-, the Eigenstrat and the VCF-Format.

References

1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, Coriell Institute for Medical Research, Max Planck Institute for Molecular Genetics, US National Institutes of Health, Analysis group, Affymetrix, Albert Einstein College of Medicine, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Massachusetts General Hospital, McGill University, New York Genome Center, Ontario Institute for Cancer Research, Pennsylvania State University, … Peruvian in Lima, P. (pel). (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. http://www.nature.com/doifinder/10.1038/nature15393

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R. G., Hallgren, F., Khartanovich, V., … Reich, D. E. (2015). Massive migration from the steppe was a source for indo-european languages in europe. *Nature*, *522*(7555), 207–211. http://www.nature.com/doifinder/10.1038/nature14317

Mah, M., Patterson, N., & Price, A. (2021). Eigensoft. In *GitHub repository*. GitHub. https://github.com/DReichLab/EIG

Mallick, S., Micco, A., Mah, M., Ringbauer, H., Lazaridis, I., Olalde, I., Patterson, N., & Reich, D. (2024). The allen ancient DNA resource (AADR) a curated compendium of ancient human genomes. *Scientific Data*, *11*(1), 1–10. https://doi.org/10.1038/s41597-024-03031-7

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E. R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J. L., Castro, J. M. B. de, Carbonell, E., … Reich, D. E. (2015). Genome-wide patterns of selection in 230 ancient eurasians. *Nature*. http://www.nature.com/doifinder/10.1038/nature16152

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., & Warinner, C. (2021). Ancient DNA analysis. *Nature Reviews. Methods Primers*, *1*(1). https://doi.org/10.1038/s43586-020-00011-0

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. E. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. http://www.nature.com/doifinder/10.1038/ng1847

Rohland, N., Mallick, S., Mah, M., Maier, R., Patterson, N., & Reich, D. (2022). Three assays for in-solution enrichment of ancient human DNA at more than a million SNPs. *Genome Research*, *32*(11-12), 2068–2078. https://doi.org/10.1101/gr.276728.122

Schiffels, S. (2025). Sequence-formats. In *GitHub repository*. GitHub. https://github.com/stschiff/sequence-formats

Schmid, C., Ghalichi, A., Lamnidis, T. C., Mudiyanselage, D. B. A., Haak, W., & Schiffels, S. (2024). *Poseidon – a framework for archaeogenetic human genotype data management*. https://doi.org/10.7554/elife.98317

Van der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the cloud: Using docker, GATK, and WDL in terra*. O'Reilly Media.