bioRxiv preprint doi: https://doi.org/10.1101/105346; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Genetics and Population Analysis

Genetic and population analysis

ATLAS: Analysis Tools for Low-depth and Ancient Samples

Vivian Link [1,2], Athanasios Kousathanas [1,2], Krishna Veeramah [3], Christian Sell [4], Amelie Scheu [4] and Daniel Wegmann [1,2]*

[1]Department of Biology, University of Fribourg, Fribourg, 1700, Switzerland,
[2]Swiss Institute of Bioinformatics, Fribourg, 1700, Switzerland,
[3]Department of Ecology and Evolution, Stony Brook University, 11794, Stony Brook, USA and
[4]Institute of Anthropology, Johannes Gutenberg-University, Mainz, 55128, Germany.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Post-mortem damage (PMD) obstructs the proper analysis of ancient DNA samples and can currently only be addressed by removing or down-weighting potentially damaged data. Here we present ATLAS, a suite of methods to accurately genotype and estimate genetic diversity from ancient samples, while accounting for PMD. It works directly from raw BAM files and enables the building of complete and customized pipelines for the analysis of ancient and other low-depth samples in a very user-friendly way. Based on simulations we show that, in the presence of PMD, a dedicated pipeline of ATLAS calls genotypes more accurately than the state-of-the-art pipeline of GATK combined with mapDamage 2.0.
Availability: ATLAS is an open-source C++ program freely available at https://bitbucket.org/phaentu/atlas.
Contact: Daniel.Wegmann@unifr.ch
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Ancient genomes provide unique insights into past populations and substantially enhance the inference of demographic and selective events that shaped modern genetic diversity. However, there are two main challenges when genotyping ancient DNA (aDNA): first, only low numbers of unique aDNA fragments remain for sequencing and second, these fragments can be subject to post-mortem DNA damage (PMD). The most common form of PMD is the deamination of cytosin (C), which leads to a C $\rightarrow$ T transition on the affected and a G $\rightarrow$ A on the opposite strand after amplification and sequencing (Briggs and Stenzel, 2007). The probability of a C deamination is highest at the ends of the DNA fragments, as these are often single-stranded, and then decays roughly exponentially towards their center (Jónsson *et al.*, 2013). As these transitions are artifacts not reflective of the original genome, they must not be considered as variants.

The presence of PMD in the data can be reduced by an enzymatic treatment that cleaves fragments at C's affected by PMD. However, this technique is restricted to PMD at unmethylated C's (Briggs *et al.*, 2010) and removes precious material. Alternatively, the false calling of variants can be reduced bioinformatically, for instance by trimming reads of their first few nucleotides (e.g. Gamba *et al.*, 2014). This leads to a problematically high loss of data, however, if done conservatively. Data loss is much smaller when using mapDamage 2.0 (Jónsson *et al.*, 2013), which incorporates the effect of PMD into genotyping pipelines such as Paleomix (Schubert *et al.*, 2014) by rescaling base quality scores to additionally reflect the probability of being damaged. But this also leads to a loss of information since PMD is only accounted for indirectly.

We here present the Analysis Tools for Low-depth and Ancient Samples (ATLAS), a collection of statistical methods built upon a dedicated genotyping model that comprehensively accounts for PMD (Hofmanová *et al.*, 2016; Kousathanas *et al.*, 2017). ATLAS works directly from raw BAM files (Li *et al.*, 2009) and contains all necessary methods to accurately genotype and estimate genetic diversity from ancient samples. ATLAS further includes many auxiliary tools to build complete and customized pipelines to work with aDNA or other low-depth samples.

2 Methods

2.1 Recommended pipeline for single-end data

ATLAS is written in C++, uses the BamTools library (Barnett *et al.*, 2011) to parse and write BAM files, and implements many tools, listed in supplementary section 1. Here we present some highlights by outlining the recommended pipeline for single-end sequencing data of ancient samples.

Step 1: Classify reads by length. Since PMD rates depend on the distance from fragment ends, ATLAS distinguishes reads spanning the entire fragment, for which these distances are known, from those shorter than their fragment.

© The Author 2017. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

bioRxiv preprint doi: https://doi.org/10.1101/105346; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

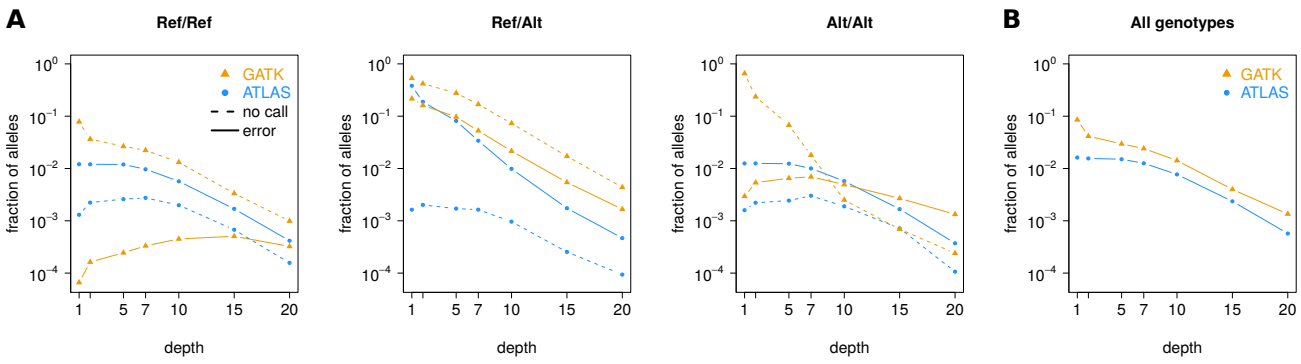

Fig. 1. A) Fraction of alleles not (dashed) or wrongly (solid) called at sites with sequencing depth >0 in a simulated chromosome of 10Mbp. The calls are classified according to the underlying true genotype (Ref=reference allele, Alt=alternative allele). B) Fraction of alleles that were not correctly called (either no or a wrong call) with a sequencing depth >0.

Step 2: Inferring PMD patterns. ATLAS either infers position-specific PMD patterns, as in MapDamage 2.0 (Jónsson *et al.*, 2013), or fits a generalized model of exponential decay (Kousathanas *et al.*, 2017), which reduces estimation noise at positions with few PMD observations.

Step 3: Recalibrating base quality scores. Base quality scores given by sequencing machines are typically distorted and must be recalibrated. ATLAS offers two recalibration methods we recently developed: a direct extension of Base Quality Score Recalibration (BQSR, DePristo *et al.*, 2011) to aDNA (Hofmanová *et al.*, 2016) applicable to populations with well characterized polymorphisms, and a reference-free method exploiting haploid or ultra-conserved genomic regions (Kousathanas *et al.*, 2017).

Step 4: Variant Calling. ATLAS offers three variant callers: a Maximum-Likelihood (MLE) caller similar to GATK (DePristo *et al.*, 2011), a Bayesian caller that forms a prior from local diversity and nucleotide composition, and a Bayesian haploid-level caller for very low sequencing depth that identifies the allele with the most evidence to be present using the same prior. For population-level samples, ATLAS further produces individual-specific gVCF files to be analyzed jointly by GATK.

2.2 Additional functionalities

For paired-end data ATLAS requires a slightly different pipeline (Supplementary Section 2). Aside from genotype calling, ATLAS also estimates region-specific heterozygosity while accounting for genotype uncertainty (Kousathanas *et al.*, 2017), offers tools to reduce modern contamination (similar to PMDS Skoglund *et al.*, 2014), and generates input files for PSMC (Li and Durbin, 2011) and BEAGLE (Ayres *et al.*, 2012), all of which while accounting for PMD.

3 Results

Following Kousathanas *et al.* (2017), we simulated a random diploid chromosome of ten Mbp with heterozygosity $\theta = 0.01$ and corresponding ancient single-end sequencing data with errors and PMD at rates commonly observed. We used ATLAS and GATK with comparable parameters (supplementary section 3) to recalibrate base quality scores with BQSR and produce MLE genotype calls. Our GATK pipeline additionally included PMD correction with mapDamage 2.0.

As previously reported (Hwang *et al.*, 2015), GATK has an inherent reference bias, resulting in about ten times less genotyping errors at true homozygous reference (Ref/Ref) than homozygous alternative (Alt/Alt) sites, depending on sequencing depth (Fig. 1A). This bias is further accentuated since GATK also less frequently emitted calls at Alt/Alt sites, particularly a low depth, and could not be removed with alternative parameterizations (supplementary section 4). In contrast, ATLAS does not show any such bias (Fig. 1A), and will thus result in higher power to detect true differences between ancient and modern samples.

The two methods also implement different strategies to deal with uncertain genotypes. While ATLAS is designed to emit calls for all sites with data, albeit with low quality scores reflecting that uncertainty, GATK ignores many sites where data is limited. This resulted in slightly higher error rates of ATLAS at lower depth (Fig. 1A), but in a larger fraction of the total genome being called correctly (Fig. 1B). At higher depths, and by accounting for PMD, ATLAS made increasingly less errors than GATK.

4 Acknowledgements

We thank Pontus Skoglund for valuable comments on an earlier version of this work, which was supported by the Swiss National Foundation grant 31003A_149920 to DW.

References

Ayres, D. *et al.* (2012). BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.*, **61**, 170–3.

Barnett, D. W. *et al.* (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, **27**, 1691–2.

Briggs, A. and Stenzel, U. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. USA*, **104**, 14616–14621.

Briggs, A. W. *et al.* (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.*, **38**, 1–12.

DePristo, M. a. *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–8.

Gamba, C. *et al.* (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.*, **5**, 5257.

Hofmanová, Z. *et al.* (2016). Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl Acad. Sci. USA*, **113**, 6886–6891.

Hwang, S. *et al.* (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.*, **5**, 17875.

Jónsson, H. *et al.* (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, **29**, 1682–4.

Kousathanas, A. *et al.* (2017). Inferring Heterozygosity from Ancient and Low Coverage Genomes. *Genetics*, **205**(1), 317–332.

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–496.

Li, H. *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Schubert, M. *et al.* (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.*, **9**(5), 1056–1082.

Skoglund, P. *et al.* (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl Acad. Sci. USA*, **111**, 2229–2234.