bioRxiv preprint doi: https://doi.org/10.1101/105346; this version posted February 2, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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

OXFORD

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. Currently, PMD can only be addressed by adjusting sequencing quality scores or by removing potentially damaged data. Here we present ATLAS, a suite of methods to analyze ancient samples that properly account for PMD. It works directly from raw BAM files and contains all necessary methods to infer patterns of PMD, recalibrate base quality scores and accurately genotype ancient DNA, along with many other useful tools. ATLAS enables the building of complete and customized pipelines for the analysis of ancient and low-depth samples in a very user-friendly way. Using 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

Thanks to technological advances, ancient genomes are currently being generated a rapid pace. These provide unique insights into past populations and can substantially enhance the inference of demographic and selective events that shaped modern genetic diversity (Slatkin and Racimo, 2016). 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 methylated or unmethylated cytosin (C) to uracil (U) or thymine (T), respectively (Sassa *et al.*, 2016). In both cases, amplification and sequencing lead to a C $\rightarrow$ T transition on the affected and a G $\rightarrow$ A on the opposite strand (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 thus more exposed to damage, and then decays roughly exponentially towards the center of the fragment (Jónsson *et al.*, 2013). As these transitions are artifacts not reflective of the sample's original DNA sequence they must not be considered as variants in downstream analyses.

One strategy to reduce the presence of PMD in the data is to treat the extracted DNA with Uracil-DNA-Glycosylase (UGD) and endonuclease VIII, which cleave fragments at unmethylated C's affected by PMD. However, this technique is restricted to one class of PMD (Briggs *et al.*, 2010) and leads to a loss of often precious material. Another option is to reduce the calling of false variants bioinformatically, for instance by trimming reads of their first few nucleotides, as these show the highest rates of PMD (e.g. Gamba *et al.*, 2014). This may lead 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 by rescaling the quality scores of a base according to its probability of being damaged. However, like the other approaches, mapDamage 2.0 accounts for PMD only indirectly by reducing the influence of possibly affected bases, and hence also leads to loss of information.

© 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 February 2, 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.

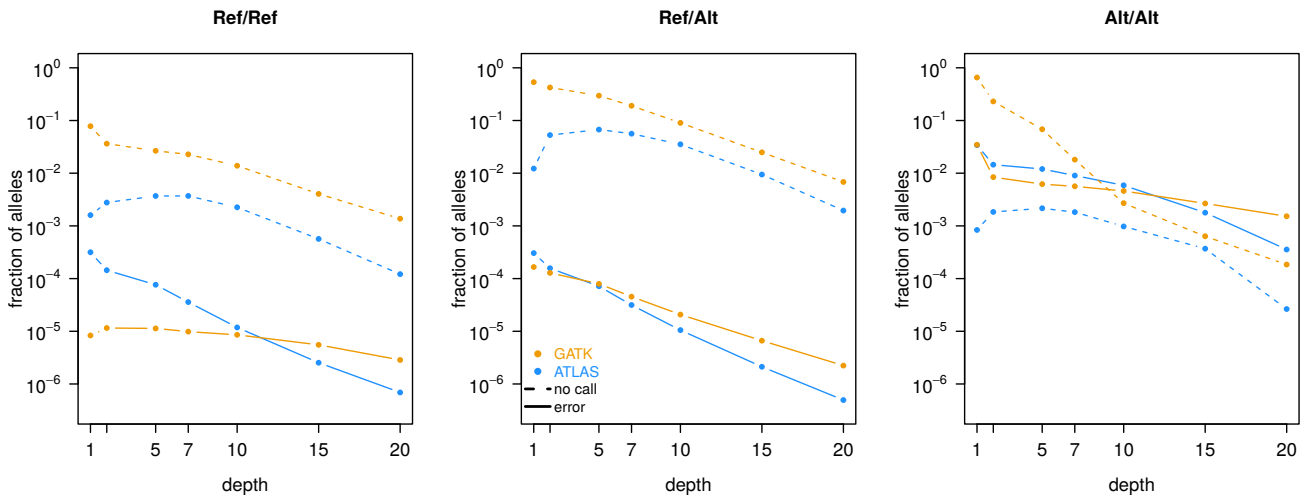

Fig. 1. Fraction of alleles not called or wrongly called at sites with sequencing depth > 0 in a simulated sample of ten ancient chromosomes. The calls are classified according to the underlying true genotype (Ref=reference allele, Alt=alternative allele).

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.*, 2015; Kousathanas *et al.*, 2016). ATLAS works directly from raw BAM files and contains all necessary methods to infer patterns of PMD, recalibrate base quality scores and accurately genotype ancient samples, as well as to infer population genetic estimates of genetic diversity directly from genotype likelihoods and produce gVCF files to call population samples with GATK (DePristo *et al.*, 2011). ATLAS further includes many auxiliary tools to build complete and customized pipelines to work with aDNA or other low-depth samples.

2 Methods

While ATLAS offers a large number of useful features, we illustrate here the most important ones by outlining the pipeline we recommend for single-end sequencing of ancient samples. We will then briefly discuss how to use ATLAS on paired-end sequencing data and describe additional tools likely to be useful to many users. An exhaustive list of all tools available and details on their usage are available on the project page.

2.1 Recommended pipeline for single-end data

Step 1: Split reads by length. The probability of PMD depends on a nucleotide's distances from the fragment ends. These distances are known for reads spanning the entire fragment, but not for those shorter than their fragment. ATLAS thus implements a functionality to classify the reads accordingly in order to infer PMD patterns independently for both groups.

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.*, 2016). While the former is more accurate at the end of the reads, imposing the generalized model reduces estimation noise in the rest of the read where PMD observations are rare.

Step 3: Recalibrating base quality scores. Base quality scores given by sequencing machines such as Illumina are typically not reflective of the true error probability and must be recalibrated. ATLAS offers two recalibration methods recently developed by us: 1) a direct extension of Base Quality Score Recalibration (BQSR, DePristo *et al.*, 2011) to aDNA (Hofmanová *et al.*, 2015) applicable to populations for which detailed knowledge on known polymorphisms is available, and 2) a reference-free method that exploits haploid or ultra-conserved regions of the genome (Kousathanas *et al.*, 2016).

Step 4: Variant Calling. ATLAS implements three different variant callers: 1) An MLE genotype caller similar to the one by Li (2011). 2) A Bayesian genotype caller that puts a prior on the genotypes based on nucleotide frequencies and the heterozygosity of the genomic region. 3) A Bayesian haploid-level caller particularly suited for very low sequencing depth that determines the allele with the most evidence to be present using the same prior. While all these callers are applied to single individuals, ATLAS also allows calling of population-level samples by producing individual-specific gVCF files that can be analyzed jointly by GATK.

2.2 Additional functionalities

ATLAS can be readily applied to paired-end data, but requires a different pipeline, which is detailed on the project page. For instance, paired-reads from aDNA fragments are usually merged prior to genotyping as they frequently overlap due to their reduced length. Aside from genotype calling, ATLAS also estimates local heterozygosity in a genomic region accounting for the uncertainty of the local genotypes and PMD (Kousathanas *et al.*, 2016), offers tools to reduce modern contamination (similar to PMDS Skoglund *et al.*, 2014), generates input files for PSMC (Li and Durbin, 2011), ANGSD (Nielsen *et al.*, 2011) or BEAGLE (Ayres *et al.*, 2012) while accounting for PMD, and adjusts the quality scores in BAM files to reflect the recalibration and base-specific PMD probabilities, just to name a few.

2.3 Implementation

ATLAS is written in C++ and uses the BamTools library (Barnett *et al.*, 2011) to parse and write BAM files.

3 Results

We simulated a random sample of ten diploid chromosomes each of ten Mbp in length with associated genetic variation generated according to the expected site frequency spectrum for $\theta = 0.001$. We then simulated ancient single-end sequencing data with errors and PMD at rates commonly observed following Kousathanas *et al.* (2016). We used both ATLAS and GATK with comparable parameters (supplementary section 1) to recalibrate base quality scores with BQSR and to produce two sets of

bioRxiv preprint doi: https://doi.org/10.1101/105346; this version posted February 2, 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.

gVCF files. The GATK set was corrected for PMD with mapDamage 2.0. Variants were then called with GATK's GenotypeGVCFs jointly for all chromosomes within a set.

As revealed by comparing them to the latter, a much higher proportion of sites could be called when analyzing the data with ATLAS than with GATK regardless of the underlying true genotype (Fig. 1). We note that this resulted in a slightly higher error rate of ATLAS at low sequencing depths, but not at depths above five for Ref/Alt and ten for the rest of the calls. At these depths, ATLAS makes increasingly less errors than GATK, articularly at sites affected by PMD.

As previously suggested (Hwang *et al.*, 2015), GATK has an inherent reference bias, as it made more mistakes at heterozygous than at homozygous reference sites, while ATLAS showed no such difference. Further, the difference in error rates is highest for the homozygous reference genotypes, whereas for the heterozygous and homozygous alternative sites GATK makes almost as many errors as ATLAS. We tried to remove this bias by testing alternative configurations of GATK (supplementary section 1), but to no avail.

4 Funding

This study was supported by the Swiss National Foundation grant 31003A_149920 to DW.

References

Ayres, D. L. *et al.* (2012). BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Systematic biology*, **61**(1), 170–3.

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

Briggs, A. and Stenzel, U. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, **104**, 14616–14621.

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

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

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

Hofmanová, Z. *et al.* (2015). Early farmers from across europe directly descended from neolithic aegeans. *bioRxiv*.

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

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

Kousathanas, A. *et al.* (2016). Inferring heterozygosity from ancient and low coverage genomes. Technical report.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, **27**, 2987–93.

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

Nielsen, R. *et al.* (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Sassa, A. *et al.* (2016). Mutagenic consequences of cytosine alterations site-specifically embedded in the human genome. *Genes and environment : the official journal of the Japanese Environmental Mutagen Society*, **38**, 17.

Skoglund, P. *et al.* (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, **111**, 2229–2234.

Slatkin, M. and Racimo, F. (2016). Ancient DNA and human history. *Proceedings of the National Academy of Sciences*, **113**, 6380–6387.