Data Management Plan.

The genomics community has generally been a leader in open data sharing, recognizing both the importance of reproducibility and that the ultimate impact of our work is maximized if our data can be used by other researchers to perform expanded analyses or test different hypotheses. This principle is especially relevant to the bioinformatics educational goals of this proposal, as this skillset is so valuable precisely because of the wide availability of (under-analyzed) genomic-scale data.

The genomics field and its journals have certain minimum standards for data accessibility, e.g., requiring raw data generated as part of a study to be deposited in public, permanent repositories. This standard will be met for all generated project data (as described below). However, I have found that data usability and interpretability are increased markedly if appropriate intermediate and final outputs are also made available. For example, in addition to raw sequence read data from a project, the availability of intermediate files with read mapping summaries and the final individual genotype estimates can be of equal or higher value. Accordingly, for all of papers published from my lab, we have deposited relevant processed data files to appropriate public databases. Demonstrating the value of this effort to the broader community, the data from four papers to date that we have deposited in the Dryad Digital Repository have already been downloaded a total of 132 times (as of 15 July 2015), despite three of these papers being published only within the past year (and note that this does not include downloads for the associated raw data that are stored on other repositories). I am likewise committed to making our bioinformatics code available through similar repositories, to facilitate future applications. All of my lab's current bioinformatics projects are explicitly planning towards making our analytical tools available to a developing anthropological bioinformatics community. Finally, I have worked with the Galaxy (a free-use bioinformatics platform hosted by Penn State) team to develop and implement tools for evolution- and conservation-minded population genomic analyses (including for non-model species, for example for all of the analyses conducted in my 2013 aye-aye population genomics paper published in *PNAS*). This will be one of the outcomes of the integrated educational component of the proposed project, as our bioinformatics workshop 'hackathon' goals will include the development and implementation of a standard-interface, user-friendly toolkit on Galaxy for the identification and genotyping of STRs from shotgun genome sequencing and targeted DNA capture sequencing read data, and demographic history reconstructions with the resulting genotypes. An additional benefit is that through this process the workshop participants will develop intimate familiarity with Galaxy and its potential value.

Illumina BeadChip human and cattle SNP genotype data. The human Malagasy 1M SNP genotype data will be deposited in the database of Genotypes and Phenotypes (dbGAP) maintained by the National Center for Biotechnology Information (NCBI). Approval for deposition was provided by the participants as part of the informed consent process; this was also approved by the Penn State IRB and Madagascar Ministry of Health. The cattle 770k SNP genotype data will be deposited in NCBI's Gene Expression Omnibus repository.

Raw sequence data. All genomic sequence read data (Illumina HiSeq) generated through this study (for both the shotgun sequencing and DNA capture steps) will be deposited to the NCBI Short Read Archive (SRA) prior to publication. This is the standard repository for sequence read data and is required prior to publication by most journals.

Intermediate and final data processing files. All potentially valuable output from our bioinformatics processing steps will be deposited in the Dryad Digital Repository, including (as applicable for the various species) ADMIXTURE population ancestry results and HAPMIX estimates of SNP haplotype ancestries for each individual, STR and flanking region sequences and their mapped genomic locations for each species, DNA capture probe set sequences for each species, SNP and STR genotype results including pileup summary data for each individual, and the simulated STR and SNP datasets. These data will be published on Dryad prior to or upon the publication of each associated manuscript arising from this project.

Bioinformatics code. All code developed as part of this project will be made freely available through GitHub. Additionally, as mentioned above, certain modules will also be developed specifically for integration with the Galaxy bioinformatics platform as part of the integrated educational component of the project.