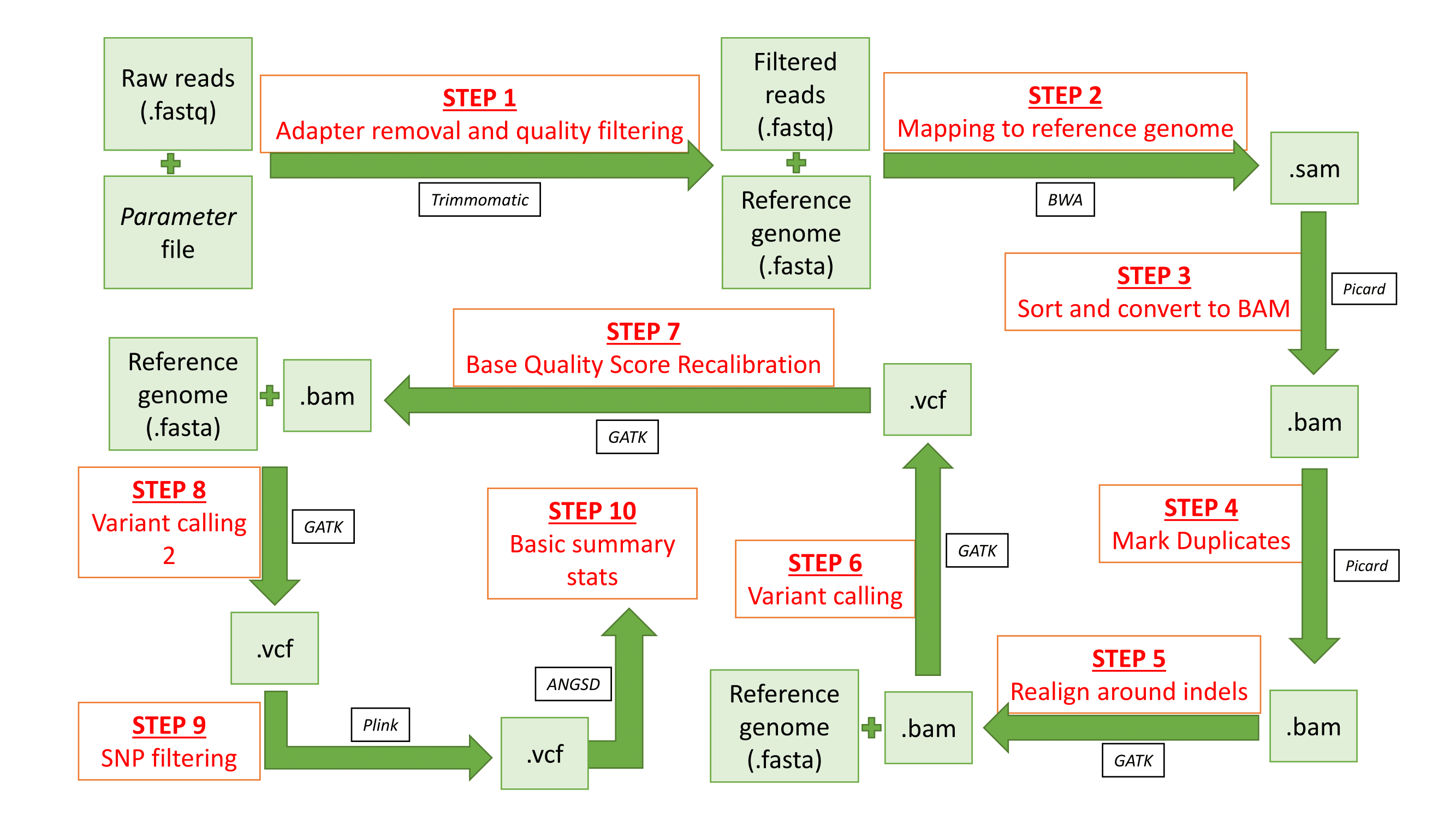
Programming and Data Science for Biology

March 2018

Lucas Rocha Moreira

**Project Proposal**

Low-coverage whole genome sequences have become a new trend in population genetics and phylogenetics due to the rise of new sequencing platforms that yield massive amounts of genetic data for reasonably low prices1,2. The analysis of these data, however, is very computationally demanding and requires several consecutive steps that use different bioinformatic algorithms. Despite the recent popularity of this approach, very few bioinformatic pipelines have been developed to automatize this process and make it more user-friendly. I propose to develop a program, called *iGenome*, that takes as input a set of individual fastq files and a “parameters” file and returns a multi-sample vcf file with nicely filtered and curated SNPs. The program is a wrapper inspired by ipyrad3 and will be written in python with the purpose of being primarily accessed through a command-line interface (CLI). The workflow involves ten sequential steps (shown in the diagram below), which call different programs and use parallel processing. Each step will be run separately and can be branched out so that different parameters can be tested. The program will be tested with a subset of low-coverage whole genome data from *Picoides villosus*, recently obtained. *iGenome* will be freely available through GitHub and will hopefully be of use to other interested users.



References

1. Oyler-McCance, S. J., Oh, K. P., Langin, K. M. & Aldridge, C. L. A field ornithologist’s guide to genomics: practical considerations for ecology and conservation. *Auk* **133,** 626–648 (2016).

2. Lemmon, E. M. & Lemmon, A. R. High-throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* **44,** 99–121 (2013).

3. Eaton, D. A. R. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* **30,** 1844–1849 (2014).