The effect of read-to-reference evolutionary relatedness on bioinformatic mapping efficiency of genomic sequencing data

Dante Celani, Randy Klabacka
Department of Biological Sciences
Utah Tech University, St. George, Utah

INTRODUCTION

A major obstacle in genomic research with non-model organisms (99.99% of all organisms on earth) is the selection of an appropriate reference genome to use for bioinformatic processing of genetic sequencing data. Raw genomic data is ideally mapped to a reference genome from the same organism, but limited research of non-model organisms typically means the absence of a reference genome. One solution is to use the genome from a closely-related organism. However, mutations that accumulate between species scale with time, meaning that genetic dissimilarity in more distantly related taxa will result in decreased mapping efficiency. Among amniote vertebrates, squamate reptiles are a great system for testing mapping efficiency across clades since this group is the most species-rich of all amniotes.

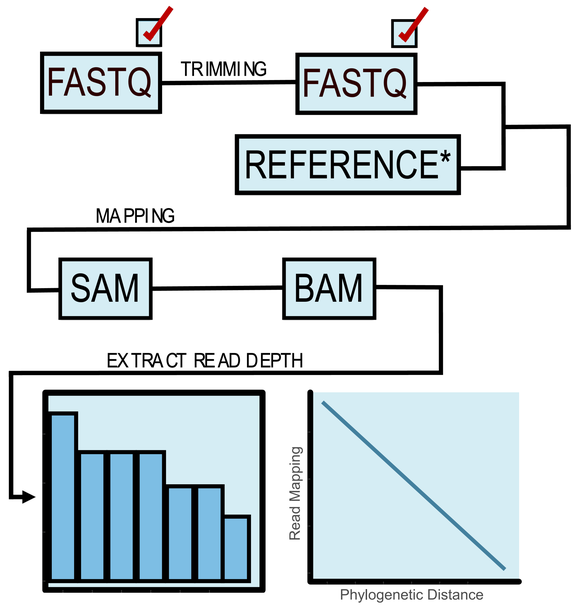
PREDICTION

We predict that mapping success (as measured by read depth) will be positively associated with genetic relatedness, based on increased genetic dissimilarity between species that are more distantly related.

METHODS

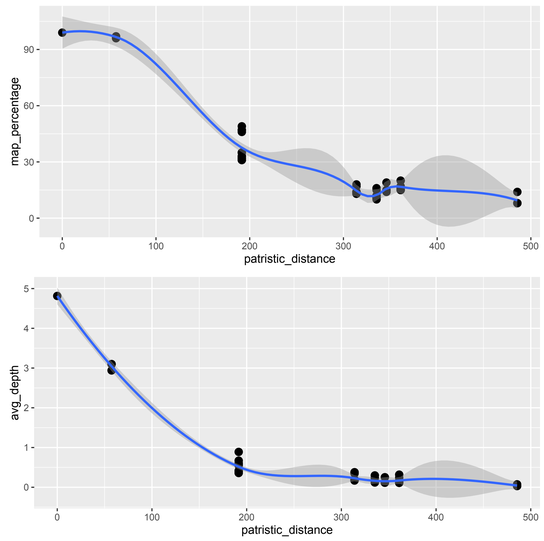
We test the mapping efficiency to reference genomes of varying relatedness using whole-genome sequence data from four species of Aspidoscelis (whiptail lizards), which is a non-model study system with a wealth of information on sexual/asexual reproduction. To determine the relationships, we run the raw reads through the pipeline. The raw reads must be cleaned before mapping them to each of the 28 vertebrate genomes. The resulting BAM files are merged for simplification. From the condensed bam files the % of reads mapped to each genome are determined along with average depth and number of reads.

BIOINFORMATICS PIPELINE



Schematic of our bioinformatics pipeline (including quality check points) and read mapping predictions based on evolutionary distance from reads to reference genome.

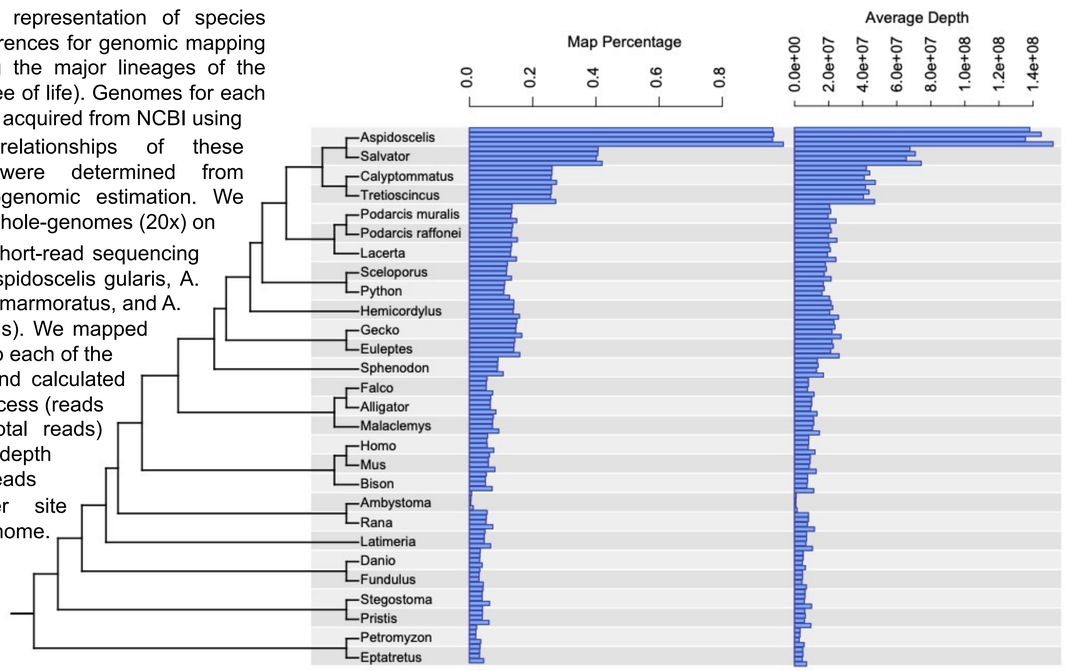
READ-REF PHYLO DISTANCE



Schematic of our bioinformatics pipeline (including quality check points) and read mapping predictions based on evolutionary distance from reads to reference genome. Line estimate is from geom_smooth r package using span of 0.5.

MAP PERCENTAGE AND AVERAGE DEPTH ACROSS VERTEBRATES

Phylogenetic representation of species used as references for genomic mapping (representing the major lineages of the vertebrate tree of life). Genomes for each lineage were acquired from NCBI using ftp. The relationships of these organisms were determined from recent phylogenomic estimation. We sequenced whole-genomes (20x) on an Illumina short-read sequencing approach (Aspidoscelis gularis, A. inornatus, A. marmoratus, and A. septemvittatus). We mapped these data to each of the references and calculated mapping success (reads mapped / total reads) and average depth (number of reads mapped per site along the genome.



CONCLUSIONS

- For the most part the mapping percentage and average depth were negatively correlated to phylogenetic distance as expected.
- The first of the phylogenetic distance graphs follow an inverse sigmoidal curve suggesting a potential inverse logarithmic relationship between map percentage and phylogenetic distance.
- In the second of the phylogenetic distance graphs there is a rapid decrease in average depth, indicating a potential negative exponential relationship between average depth and phylogenetic distance.
- These results show that mapping to an organism outside of your focal genus results in rapid decrease in mapping efficiency, which will result in less downstream data (e.g., variants)

FUTURE DIRECTIONS

- Examination of our raw whole-genome sequence data shows that the sequencing reads were of high quality.
- This is relevant information for researchers who work on non-model organisms to help inform their choice of reference. While there appears to be a relationship between read mapping and evolutionary relatedness, it is possible this relationship is non-linear and/or affected by genome quality.

REFERENCES AND ACKNOWLEDGEMENTS

1- Andrade P. et al. Regulatory changes underlie color polymorphisms in the wall lizard. PNAS. 2019
2- Rhie A, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021
3- Castoe T.A. Sequencing the genome of the Burmese python as a model for studying extreme adaptations in snakes.
4- Srikulnath K. et al. Highly conserved synteny and gene order in the Toxicofera clade. Chrom Res. 2013
5- Leitão H.G. Chromosome-Level Genome Assembly of the Cape Cliff Lizard (Hemicordylus capensis). GBL. 2023.
6- Westfall, A.K. et al. A chromosome-level genome assembly for the eastern fence lizard. GigaScience. 2021
7- Andrews, S. FastQC: A quality control tool for high throughput sequence data. Babraham Bioinformatics.
8- Chen, S. et al. fastp: an ultraf-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018
9- Li H. et al. Fast and accurate short read alignment with Burrows-Wheeler Transform Bioinformatics. 2009.

We thank the University of Utah Center for High Performance Computing (CHPC) for providing the cluster for our genomic data processing.