1. Lead CI Name: Jaime Gongora

2. Project Title: Genome-wide SNP and population genetics of platypuses and oryxes

3. Project Description:

This project will analyse the diversity of adaptive genes and their expression levels in two iconic species. The platypus, arguably one of the most irreplaceable mammals existing today, and the Arabian oryx, the first species to be rescued from extinction in the wild.

We will analyse whole genome sequencing data from three different datasets. The first dataset consists of 26 platypus samples that were collected in the unregulated Tenterfield Creek (n=11) and below the dam (n=8), and above the dam (n=7) in the regulated Severn River (Mijangos, Bino, et al., 2022). The second dataset consists of 57 platypus samples from across the whole species range in eastern mainland Australia and Tasmania (Martin et al., 2018). The third dataset consists of 71 Arabian oryx samples collected in Oman.

In this project, we will use new and updated reference genomes (Zhou et al., 2021) along with novel programs for read mapping (NextGenMap; Sedlazeck et al., 2013), genotype calling (Octopus; Cooke et al., 2021) and data analysis (dartR; Mijangos, Gruber, et al., 2022).

We will identify genes under selection using a variety of methods such as loadings of a PCA analysis (Price et al., 2006), identification of FST outliers (Whitlock & Lotterhos, 2015), genotype-environment association studies (Brauer et al., 2018) and comparative genomics using the program Genespace (Lovell et al., 2022).

We will analyse diversity in various gene regions and perform expression analysis to evaluate whether adaptations to local changes are already present in populations. In the platypus dataset, we will compare the genetic variation across the different populations. We anticipate changes in diversity across populations that are considered fragmented based on a prior study in the field and early findings already observed (Mijangos, Bino, et al., 2022). Also, we anticipate changes in diversity and expression levels in platypus populations residing in the regulated river (Severn River) compared to platypus populations in the unregulated river (Tenterfield Creek). In the Arabian oryx, we will investigate genes that might contribute to adaptation to desert life (*e.g.* heat loss, water conservation, and metabolism), as well as the genomic cost of a near extinction event.

The information gathered by this project will allow us to understand the adaptive potential of these populations and inform their conservation.

4. Investigators:

Dr Jaime Gongora is the Associate Dean for Indigenous Strategy at the Faculty of Science at The University of Sydney, where he is also an Associate Professor teaching conservation biology and animal genetics. Dr Gongora has dedicated his career to studying the genetic variety of wild animals, as well as their population genetics, phylogenetics, and phylogeography.

Relative to opportunity, Dr Mijangos has a stellar record, as shown by his high-quality publications and the creation of a very important and capacity-building software for genomics-based research

(dartR). Dr Mijangos's combination of advanced bioinformatic and programming abilities with a firm grasp of genetic principles makes him uniquely qualified to apply genomic methods to wildlife preservation. Dr Mijangos has ample experience using computer clusters such as GADI and Artemis at the University of Sydney.

5. Technical Plan:

As the backbone of our bioinformatics pipeline, we will use scripts developed, optimised to scale, and benchmarked in the distributed computing environment of NCI Raijin by bioinformaticians Tracy Chew, Cali Willet, and Rosemarie Sadsad from the Sydney Informatics Hub (https://github.com/Sydney-Informatics-Hub/Bioinformatics), a core facility at the University of Sydney, and HPC specialists Andrey Bliznyuk and Roger Edberg from NCI.

We will map reads using the mapping program NextGenMap (Sedlazeck et al., 2013), which is more than twice as fast as BWA while achieving a mapping sensitivity similar to Stampy or Bowtie2. We will call genotypes using the novel program Octopus (Cooke et al., 2021). This variant caller uses a polymorphic Bayesian genotyping model capable of modelling sequencing data from a range of experimental designs within a unified haplotype-aware framework. We will use a new whole genome assembly for the Arabian oryx samples that our research group just finalised.

Our bioinformatic pipeline will be divided into two parts. The first part takes raw sequencing data as input (FASTQ files), performs data pre-processing and produces BAM files (Fig. 1). The second part of our pipeline takes as input the BAM files and calls short germline variants across the samples (Fig. 2).

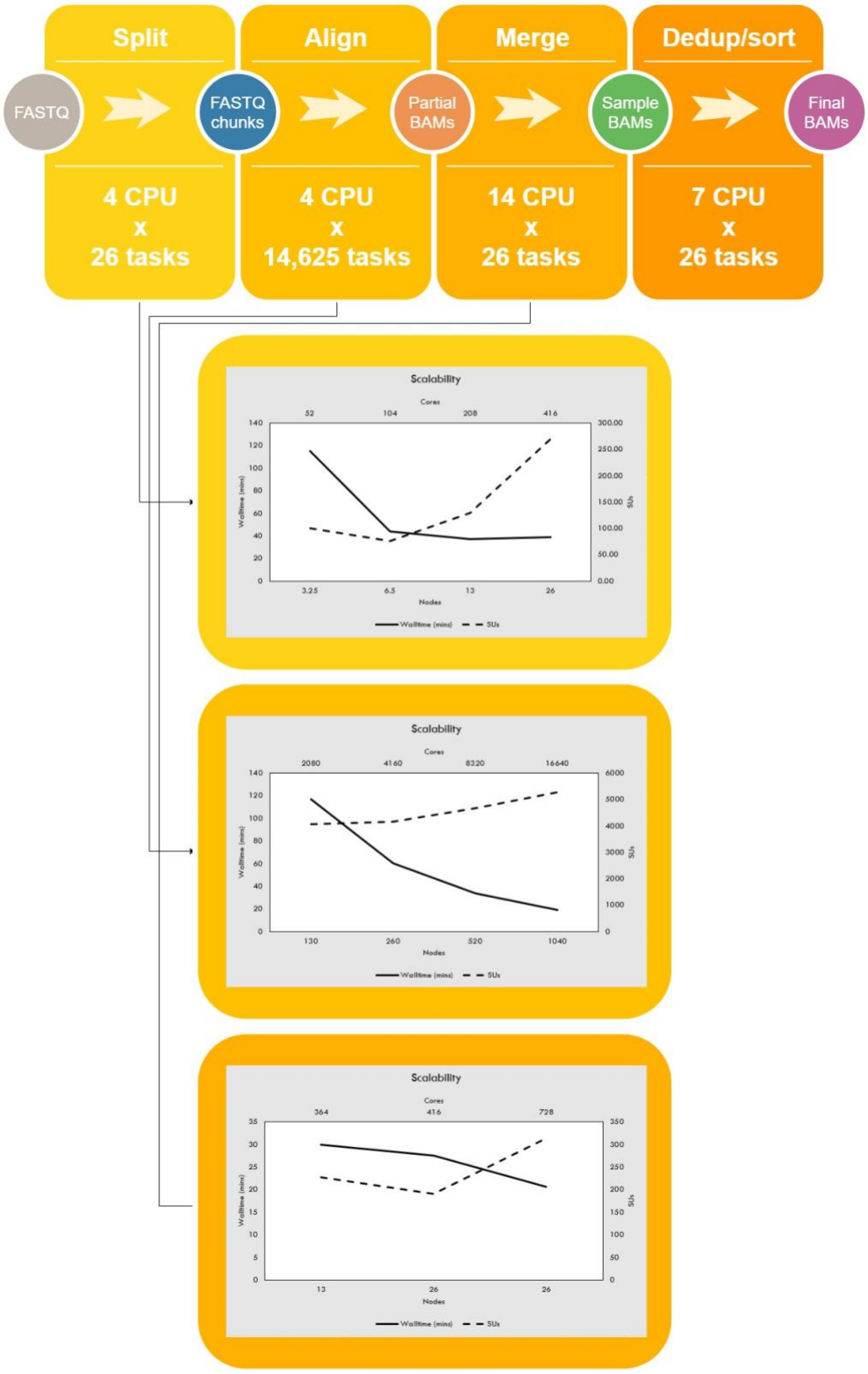

Figure 1 Example of the workflow of the first part of the pipeline. This part consists of 4 scalable jobs that are run sequentially. Each job contains independent tasks that can be run in parallel.

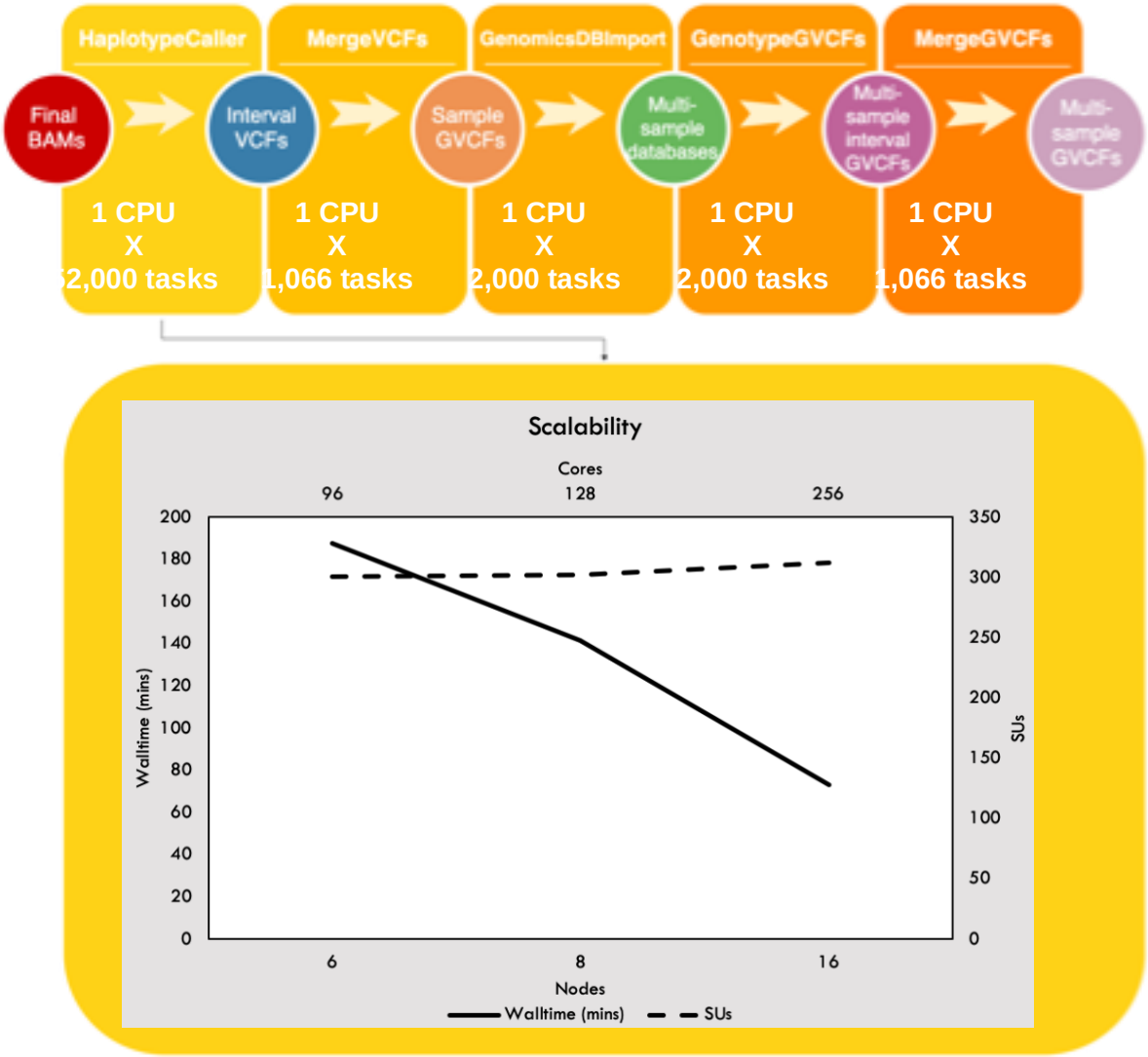

Figure 2 Example of the workflow of the second part of the pipeline. This part consists of 5 scalable jobs that are run sequentially. Each job contains independent tasks that can be run in parallel.

For data analyses, we will use the R package dartR (Mijangos, Gruber, et al., 2022).

6. Data Plan:

After the Adapter allocation period is completed, the data will be transferred and stored in the Research Data Store (RDS) service of the University of Sydney.

7. References:

Brauer, C. J., Unmack, P. J., Smith, S., Bernatchez, L., & Beheregaray, L. B. (2018). On the roles of landscape heterogeneity and environmental variation in determining population genomic structure in a dendritic system. *Molecular Ecology, 27*(17), 3484-3497.

Cooke, D. P., Wedge, D. C., & Lunter, G. (2021). A unified haplotype-based method for accurate and comprehensive variant calling. *Nature Biotechnology, 39*(7), 885-892.

Lovell, J. T., Sreedasyam, A., Schranz, M. E., Wilson, M. A., Carlson, J. W., Harkess, A., Emms, D., Goodstein, D. M., & Schmutz, J. (2022). GENESPACE: syntenic pan-genome annotations for eukaryotes. *Biorxiv*.

Martin, H. C., Batty, E. M., Hussin, J., Westall, P., Daish, T., Kolomyjec, S., Piazza, P., Bowden, R., Hawkins, M., & Grant, T. (2018). Insights into platypus population structure and history from whole-genome sequencing. *Molecular Biology and Evolution, 35*(5), 1238-1252.

Mijangos, J., Bino, G., Hawke, T., Kolomyjec, S., Kingsford, R., Sidhu, H., Grant, T., Day, J., Dias, K., & Gongora, J. (2022). Fragmentation by major dams and implications for the future viability of platypus populations.

Mijangos, J. L., Gruber, B., Berry, O., Pacioni, C., & Georges, A. (2022). dartR v2: an accessible genetic analysis platform for conservation, ecology, and agriculture. *Methods in Ecology and Evolution*.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics, 38*(8), 904-909.

Sedlazeck, F. J., Rescheneder, P., & Von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics, 29*(21), 2790-2791.

Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of FST. *The American Naturalist, 186*(S1), S24-S36.

Zhou, Y., Shearwin-Whyatt, L., Li, J., Song, Z., Hayakawa, T., Stevens, D., Fenelon, J. C., Peel, E., Cheng, Y., & Pajpach, F. (2021). Platypus and echidna genomes reveal mammalian biology and evolution. *Nature, 592*(7856), 756-762.