Visual Research Summary of Previous Projects

Project undertaken at University of Cambridge

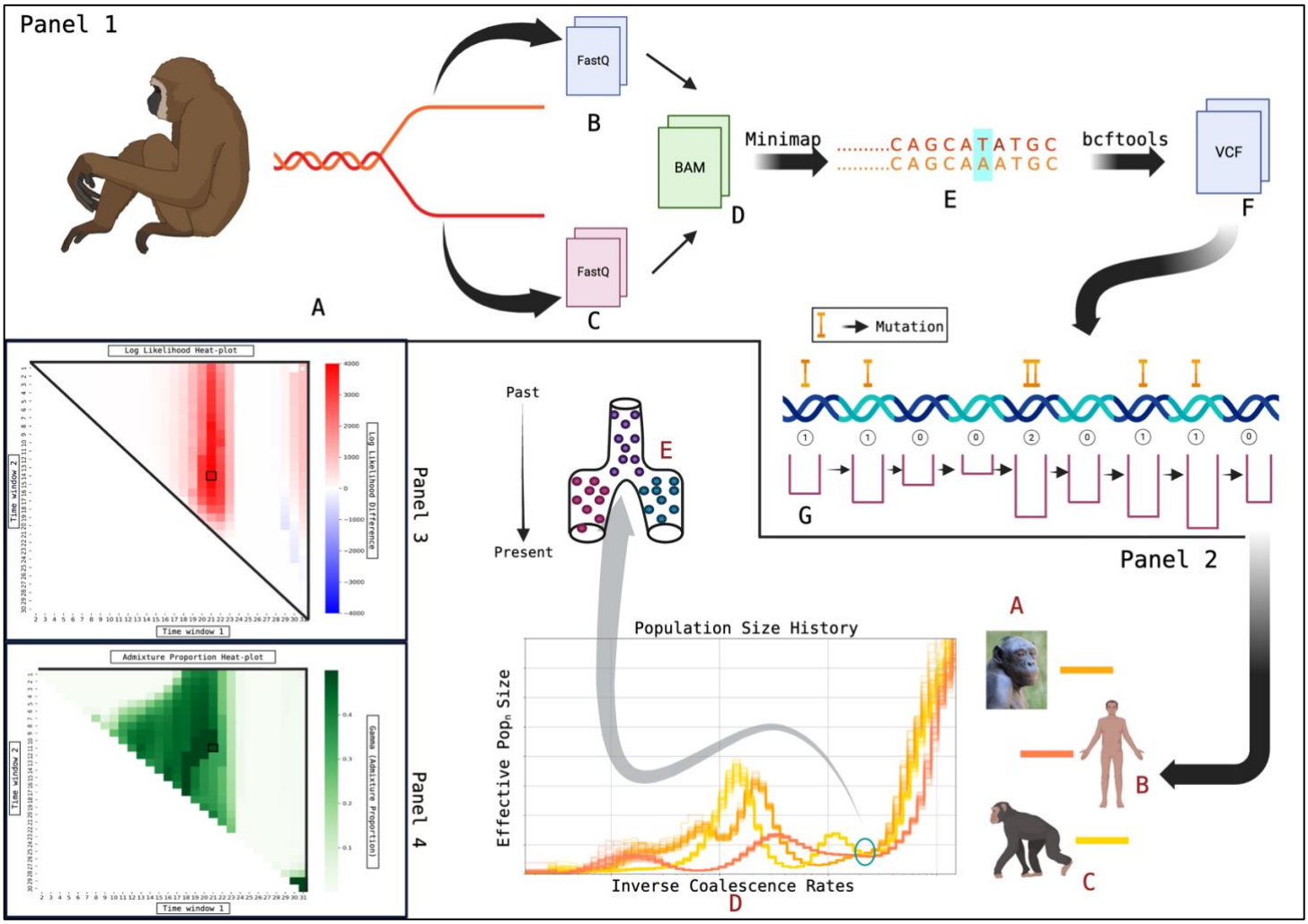

Figure 1. *Computational workflow and multi-method inference of primate demographic history using T2T genomes.*

***Panel 1*:** *Data generation and variant calling*. Telomere-to-telomere (T2T) assemblies from primate species (**A**) were split into pseudo-diploid FASTQ pairs (**B**, **C**) and aligned using minimap2 to generate BAM files (**D**). Variants were called using bcftools, producing VCFs (**F**) that encode genome-wide heterozygosity patterns (**E**).

Panel 2: *PSMC estimation of a historically effective population size.* Genome-wide variants were parsed into mutation-indexed segments (**G**). Regions with fewer mutations correspond to recent coalescent events, whereas dense mutation clusters represent deeper ancestral branches. The modified PSMC (**D**) used these intervals to estimate population-size trajectories (**A** (Bonobo), **B** (Human), **C** (Chimpanzee)), representing long-term fluctuations in N_e over time. The point (**E**) from where the trajectories diverge reflects the time of divergence into separate species from the ancestral population.

Panel 3: *Structured-coalescent log-likelihood surface from cobraa.* "cobraa" evaluates a grid of demographic models characterized by different timings of admixture. The resulting log-likelihood heatmap visualizes support for historical splits and merging events. Red regions indicate high likelihood difference, revealing signatures of ancient structure that are not captured by PSMC alone.

Panel 4: *Temporal admixture proportions and evidence for ghost introgression.* This heatmap quantifies the estimated admixture proportions across successive time windows. Elevated values indicate periods where structured-coalescent models infer gene flow from unsampled or "ghost" populations.

Master's Thesis Undertaken at LMU (1)

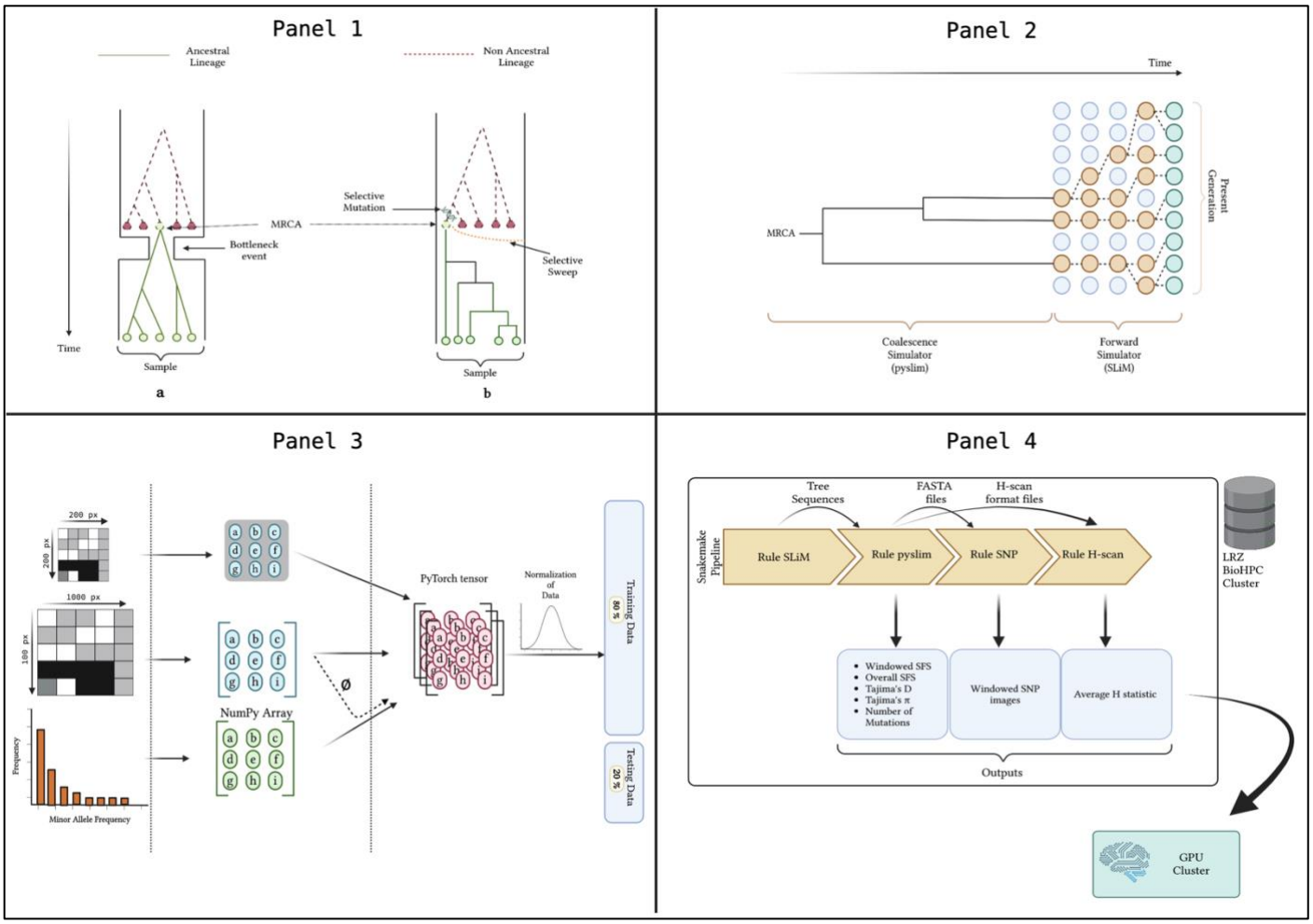

Figure 2. *Workflow for simulating and classifying bottlenecks versus selective sweeps.*

Panel 1 illustrates the genealogical differences between a demographic bottleneck (a) and a selective sweep (b), highlighting how each event shapes coalescent histories and genetic diversity.

Panel 2 shows the combined coalescent (pyslim) and forward-time (SLiM) simulation framework used to generate realistic genomic data under both scenarios.

Panel 3 summarizes the preprocessing pipeline, in which simulated SNP matrices and summary statistics are converted into NumPy arrays and normalized PyTorch tensors for deep learning.

Panel 4 presents the complete computational workflow—SLiM → pyslim → SNP processing → H-scan—used to generate windowed SFS, Tajima's D/π, SNP images, and haplotype statistics on HPC and GPU clusters for model training.

Master's Thesis Undertaken at LMU (2)

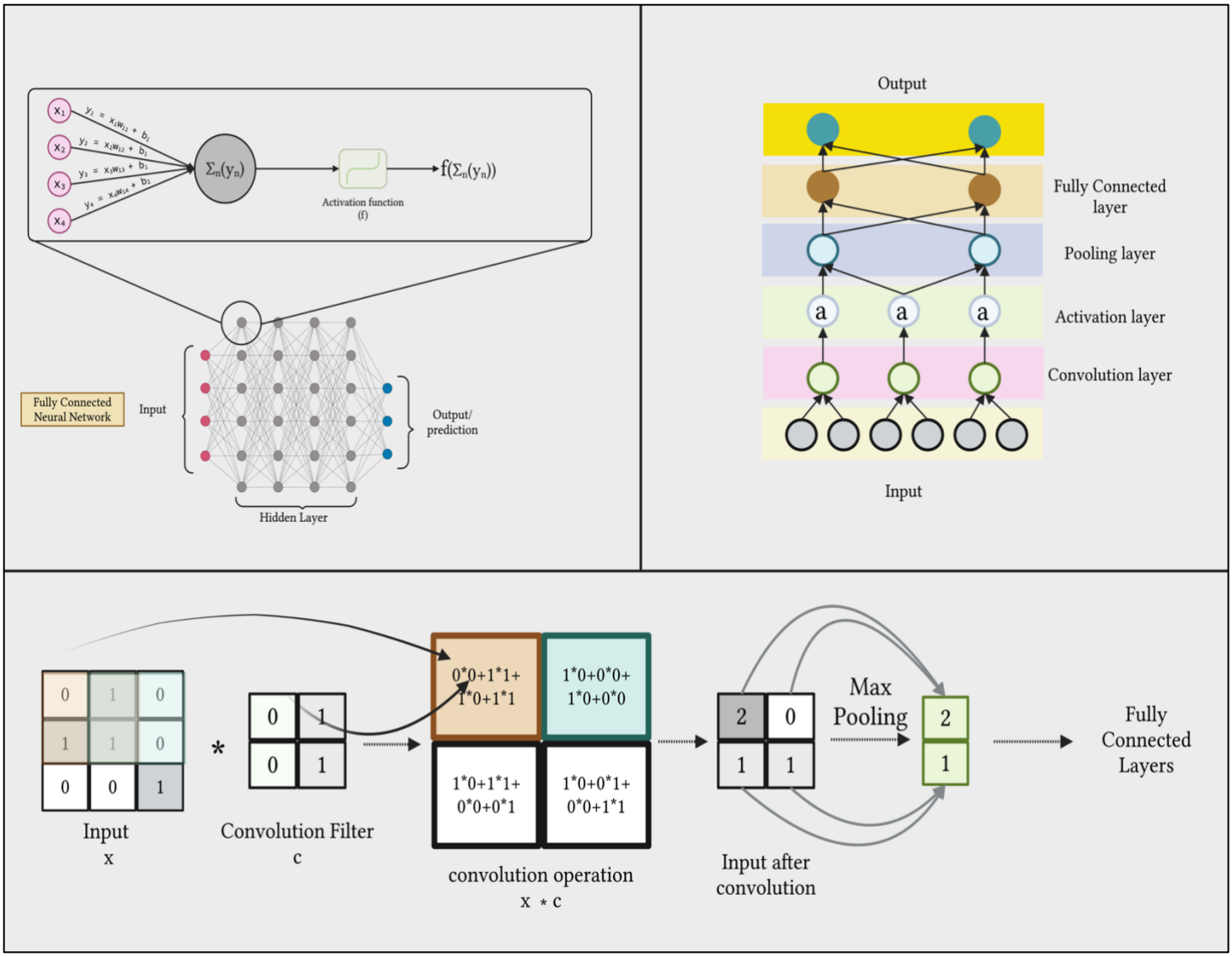

Figure 3. *Overview of neural network architecture and convolutional operations used in model training.*

The top-left panel illustrates the structure of a fully connected neural network (FCNN), where input features are linearly combined and passed through an activation function to generate nonlinear representations across hidden layers.

The top-right panel shows the hierarchical architecture of a convolutional neural network (CNN), including convolution, activation, pooling, and fully connected layers leading to the final output.

The bottom panel provides a step-by-step depiction of the convolution operation: an input matrix is scanned with a convolution filter to produce feature maps, which are then downsampled through max-pooling before being passed to fully connected layers for classification.