# MASAMB 2025 Poster Overview

## Population Genomics

### Title to be announced

Author(s): ***Andrew Bass***

The cost of acquiring samples for genome-wide association studies (GWAS) can limit sample sizes and inhibit discovery. Additionally, a typical GWAS is analyzed in isolation, ignoring information from genetically related diseases. We introduce the surrogate functional false discovery rate (sfFDR) framework which leverages informative GWAS summary data from related diseases to improve power. The sfFDR framework provides estimates of FDR quantities such as the functional local FDR and q-value, and uses these estimates to derive a functional p-value for type I error rate control and a functional local Bayes' factor for post-GWAS analyses (e.g., fine mapping and colocalization). We show through extensive simulations that sfFDR controls the type 1 error rate and, using data from the UK Biobank, find that it substantially increases power in a study of obesity-related traits (equivalent to recruiting an additional 60% of samples). In the rare disease setting, sfFDR detected ten independent associations with eosinophilic granulomatosis with polyangiitis instead of two by a standard GWAS significance analysis. Collectively, these results highlight the utility of exploiting pleiotropy in both small and large studies, and suggest that GWAS should embrace the large body of existing datasets to increase power in future studies. We anticipate that sfFDR will have broader applications as a general framework to integrate informative data, such as those from functional annotations and SNP-level priors from AI algorithms.

### MobiVirus: simulating selective sweeps in viral epidemics

Author(s): ***Anna Efstathiou****, Katia Gkimisi, Pavlos Pavlidis, Stefanos Papadantonakis*

Selective sweep detection methods were primarily developed for eukaryotic genomes, but their performance in viral populations with high mutation rates and complex spatial transmission dynamics remains unclear, particularly for detecting recent strong positive selection during epidemics. To study this problem, we developed a forward-in-time discrete event simulation 1odelling viral genome evolution within host populations distributed across a two-dimensional space. Hosts move spatially according to SIRS epidemiological dynamics, while viral genomes are represented as binary sequences containing a beneficial genomic region where mutations create enhanced transmission rates, distinguishing normal strains from super strains. The simulator incorporates natural evolutionary processes, including mutation and recombination, with infections occurring based on the distance. This setup allows us to track strain fixation dynamics and evaluate sweep detection methods under various viral adaptation scenarios. We evaluate whether existing selective sweep detection methods can accurately identify recent strong positive selection in these spatially structured viral populations using SweeD, OmegaPlus, and classical summary statistics such as Tajima's D and

Watterson's θ under different simulation conditions. Preliminary results indicate that existing tools struggle to identify hard sweeps in these types of viral populations, with selection signals appearing to be weak or diminishing rapidly.

### Title to be announced

Author(s): ***Maria Evangelinou***

Kinship inference plays a critical role in ancient DNA (aDNA) studies, as a powerful tool to infer insights into the past and reconstruct history, social structures, and kinship relationships across diverse societies and periods of time. Ancient DNA analyses can offer a better understanding of human history than the ways we already know. However, accurately estimating degrees of relatedness from low-coverage data remains challenging which is why, several tools have been developed such as KIN, KING, and READ. Yet, their ability to identify with accuracy the kinship relationships in extreme demographic scenarios hasn't been intensely tested.

This project aims to benchmark widely across these tools for a range of simulated data in different demographic scenarios and to compare their accuracies and also, focus on their relationship specificity. Our framework includes the application of classifiers in order to assess whether identified first-degree relationships represent parent-offspring or full siblings.

By creating a standardized benchmark and introducing a targeted approach for refining first-degree relationship classification, this work offers practical tools for kinship analysis and understanding the human past.

## Cancer

### Modelling Random Mutations in Extra-Chromosomal DNA Dynamics

Author(s): ***Jordan Cobley****, Weini Huang*

Extrachromosomal DNA (ecDNA) is a genetic error found in more than 30% of tumour samples across various cancer types and is correlated to worse clinical outcomes. Unlike chromosomal DNA where genetic materials are on average equally divided to daughter cells, controlled by centromeres during mitosis, the segregation of ecDNA copies is due to random partition and leads to a fast accumulation of cell-to-cell heterogeneity in copy numbers. This heterogeneity in copy numbers as well as in mutations arising in those copies are critical for tumour progress and treatment resistance. While previous studies have been focused on the ecDNA copy number dynamics, we are interested in understanding mutation accumulation processes in ecDNA elements, and whether and how they differ from the corresponding dynamics in the chromosomal DNA. We use stochastic processes to model the random segregation of ecDNA copies as well as random mutations in each ecDNA copy during cell divisions.

We will present preliminary results of ecDNA mutation patterns and their comparison with well-studied mutation patterns in the chromosomal DNA.

### Title to be announced

Author(s): ***Poulami Somanya Ganguly***

Oncogene amplification on circular extrachromosomal DNA (ecDNA) has been linked to poor prognosis and higher treatment resistance in multiple types of human cancer. ecDNA are mobile genetic elements lacking centromeres that segregate randomly into daughter cells during mitotic cell division. While random segregation of ecDNA has been shown to drive intra-tumour gene-copy-number heterogeneity in cancer cell populations, how ecDNA contribute to phenotypic heterogeneity, and therefore to cell fitness, is less well understood. Cancer cell populations with ecDNA show remarkable levels of heterogeneity in mRNA and protein concentrations, which is not predicted by the heterogeneity in gene copy number alone. I will present ongoing work on phenotypic heterogeneity in cancer cells, where heterogeneity arises as a consequence of stochastic processes during gene expression. Using a mathematical model of stochastic gene expression, which predicts heterogeneity even in isogenic populations, I will demonstrate how we quantify phenotypic heterogeneity beyond copy-number heterogeneity and discuss the implications of this for cancer cell fitness.

### Reconstructing clonal dynamics using persistent DNA damage

Author(s): ***James Hayes****, Martin S Taylor, Michael D Nicholson, Craig J Anderson, Sarah J Aitken, Duncan T Odom, Liver Cancer Evolution Consortium*

Despite recent advances, the dynamics of the earliest stages of tumorigenesis remain unclear. In our work, we leverage the mutational structure associated with DNA damage that persists, unrepaired, for multiple rounds of cell division, and use this structure to quantify key aspects of early tumour evolution. Specifically, we (i) measure tumour transformation times, that is, the number of cell divisions between the occurrence of DNA damage and a tumour's most recent common ancestor (MRCA); (ii) devised a novel test for the occurrence of subclonal selection after a tumour's MRCA. The validity of our methodology was evaluated extensively by simulation. We applied our method to a multi-strain mouse model of DNA damage induced tumorigenesis. This initial work demonstrates the utility and potential of persistent DNA damage as a powerful tool with which to investigate clonal dynamics.

### BDDTree: Birth-Death-Dormant Tree for reconstructing somatic evolution using methylation clock data

Author(s): ***Bingxin Lu*** *(University of Surrey), Bingjie Chen (Guangzhou Medical University)*

DNA methylation patterns at CpG sites represent somatic epigenetic changes that accumulate with age and can serve as molecular clocks. These patterns provide a complementary dimension to genetic data for reconstructing somatic evolutionary histories. While methylation signatures, particularly those from circulating cell-free DNA, have shown promise in early cancer detection and prognosis, methods for inferring phylogenetic trees directly from DNA methylation data remain limited.

To address this gap, we are developing BDDtree, a novel Bayesian framework for reconstructing cancer phylogenies from methylation data generated by a wide range of technologies using either bulk or single-cell samples, including methylation arrays, whole-genome bisulfite sequencing, and reduced representation bisulfite sequencing. The compatibility with diverse data types makes BDDtree broadly applicable, increasing its potential for widespread use in epigenetic and somatic evolutionary studies.

BDDtree employs Markov chain Monte Carlo (MCMC) techniques to infer the posterior distributions of both tree topologies and key model parameters, which include methylation and demethylation rates, cell birth and death rates, sampling probability, and a dormant rate representing the proportion of cells that are non-dividing and non-dying. The inclusion of the dormant rate allows us to capture spatial constraints and periods of evolutionary stasis during tumour progression.

The underlying birth-death-dormant model accounts for stochastic changes in population size and the high likelihood of sampling dormant cells from multi-regional tumour biopsies. BDDtree thus provides a powerful new tool for studying somatic evolution from an epigenetic perspective, enriching our understanding of tumour heterogeneity and dynamics beyond what genetic data alone can reveal.

## Comparative & Evolutionary Genomics

### Modeling varying demography along phylogenies from large-scale genomic data

Author(s): ***Yari Cerruti,*** *Sebastian Höhna, Rui Borges*

Several methods exist today for modeling variation in population size over time. These approaches rely on different types of data and theoretical frameworks: the simplest use summary statistics, while more complex methods rely on coalescent-based inference from site frequency spectra or ancestral recombination graphs. However, these methods share common drawbacks: they can become computationally intensive when applied to large genomic datasets, and often fail to capture more ancient demographic events due to loss of statistical power at longer time scales.

Here, we propose an extension of polymorphism-aware phylogenetic models to account for fluctuating demography along the phylogeny in single or multiple populations or species. This method has two main objectives: (1) to accommodate reversible mutations in the inference of demography, and (2) to account for the divergence history of several

populations or species. Together, these extensions will improve our ability to detect population size changes at longer evolutionary timescales.

## Tracing Dopamine Receptor Evolution: Phylogenetic Insights and a New Synteny Metric

Author(s): ***Olympia-Dialekti Vouzina****, Constantina Theofanopoulou and Tereza Manousaki*

Dopamine receptors (DRs) are key components of vertebrate neurobiology, associated with cognition, behavior, and various neuropsychiatric disorders. They belong to the protein family of G protein-coupled receptors, and are traditionally divided into two classes, D1-like and D2-like, based on functional and structural characteristics. This project investigates the evolutionary history of DRs across vertebrates using phylogenetic and comparative genomic approaches.

Protein sequences from representative species were analyzed through homology inference, multiple sequence alignment, and maximum likelihood phylogenetics. The results support the classical D1/D2 classification, reveal conserved subfamilies within each class, and highlight lineage-specific duplication and loss events, suggesting potential adaptive diversification.

To complement the phylogenetic analysis, a tool is being developed to visualize the microsynteny of DR gene neighborhoods across vertebrate genomes. Using information from the prior homology analysis, this tool calculates microsynteny scores based on gene presence/absence, order, and strand orientation, and generates comparative maps that highlight conserved and rearranged regions.

Together, these analyses—supported by the inclusion of several recently published vertebrate genomes—provide new insights into the evolution of dopamine receptors and the organization of their surrounding genomic regions across vertebrates.

## Tracing Positive Selection: A Multi-Dimensional Analysis of Mammalian Genomes

Author(s): ***Yuqing Peng****, Richard Nichols, Mario Dos Reis*

Unraveling the molecular mechanisms of adaptive evolution in mammals necessitates the identification of genes under positive selection. Here, we conducted a genome-wide screen for positive selection across 190 mammalian genomes, employing codon substitution models implemented in CODEML (PAML package). Our analysis of over 15,000 one-to-one orthologous genes from OrthoMaM, utilizing site models (M1a vs. M2a, M7 vs. M8) and likelihood ratio tests, revealed numbers of genes exhibiting evidence of adaptive evolution.

Functional enrichment analyses using Gene Ontology and PANTHER classification highlighted immune response, sensory perception, and metabolic processes as key

functional categories enriched for positively selected genes. Furthermore, structural investigations indicated a tendency for positively selected amino acid residues to cluster within specific protein domains, suggesting a potential link between these regions and adaptive function. While acknowledging the sensitivity of codon models to initial parameters and alignment quality in certain cases, subsequent re-analyses reinforced the robustness of our findings. This study provides a comprehensive genome-wide perspective on molecular adaptation in the mammalian lineage, pinpointing candidate genes likely driven by selective pressures associated with immunity, environmental interactions, and metabolic adaptation.

## Machine Learning Applications

### Harnessing Simulations to Identify Balancing Selection in Evolve-and-Resequence Experiments

Author(s): ***Baron Koylass***, *Max Reuter, Michael DiGiorgio, Matteo Fumagalli*

Natural selection is one of the key processes that governs how genetic variants become fixed, lost, or remain segregating within populations and species over time. Identifying alleles targeted by balancing selection is crucial to provide insights into mechanisms underlying the maintenance of common variants. With the emergence of large sequencing datasets and the need for ever complex models, deep learning algorithms, trained by simulations, have shown great potential to detect cryptic selection signals. Despite this, at the moment, only few studies employed deep learning methods to detect and characterise the genomic footprint of balancing selection, largely due to the complexity of producing high-fidelity simulations.

To address this gap, we sought to replicate a previously performed evolve-and-resequence experiment on sexual antagonism in Drosophila melanogaster. We implemented a novel, efficient, and flexible pipeline to simulate various nuances of the experiment including variations in population sizes, distinct replicates, linked selection, temporal sampling, and uncertainty of sequencing data. Input data derived from the Drosophila Genetic Reference Panel were seamlessly incorporated into the protocol.
Using this framework, we demonstrated that generated synthetic data closely replicate allele frequency dynamics observed in the empirical experiment. Notably, we were able to highlight a region of interest with a pattern compatible with balancing selection. We further postulate and propose how such deep learning synthetic training data allows analysis in a model-agnostic way, a task that is still considered challenging under most scenarios. We conclude by highlighting further technological advancement in devising deep learning algorithms for time-series genomic data.

## Phylogenomics & Evolutionary Networks

### Title to be announced

Author(s): ***Nok Ting Lam***, *Vladislav Ivanov, László Péter Biró, Zsolt Bálint, Krisztián Kertész, Gábor Piszter, Leonardo Dapporto, Roger Vila, Marko Mutanen, Vlad Dincă*

Species complexes with allopatric populations scattered across a wide geographic area represent some of the most challenging topics for taxonomy, but also for the prioritization of limited conservation resources. A prime example is represented by cold-adapted butterflies of the *Polyommatus eros* species group, with a fragmented distribution across Eurasia. In Europe, this group was often regarded as comprising two species (*Polyommatus eros* (Ochsenheimer) and *Polyommatus eroides* (Frivaldszky)), although different taxonomic views recognized from one to four species. Based on limited material and genetic data, a single species (*P. eros*) is recognized by the most recent checklists of European butterflies.

Here we use genome-wide data (double digest RAD sequencing) coupled with information on wing spectral characteristics to investigate evolutionary relationships among European butterflies of the *P. eros* species group. Taken together, results of phylogenetic analyses, genetic structure, gene flow, species delimitation and wing reflectance revealed a complex pattern likely shaped by the Pleistocene glaciations. While differentiation between some populations is relatively low, two lineages (one detected in mountains of northern Albania and Montenegro, the other, assignable to taxon *menelaos*, endemic to southern Greece), appear to be undergoing accelerated differentiation, suggesting a process of paraphyletic speciation. The endangered taxon *menelaos* is particularly differentiated in terms of genetics and wing spectral properties, the latter suggesting reinforcement with respect to Balkan *eroides*. Although assessed at a continental scale, a more complete understanding of the intricate patterns displayed by this species group will only be possible with the inclusion of extra-European populations.

## Arboreal Networks Through the Lens of Augmented Trees

Author(s): ***Darren Overman***, *Katharina T. Huber*

The development of powerful strategies to help protect the planets biodiversity will undoubtedly be informed by the understanding of the evolutionary history of plants and animals, amongst others. Usually expressed in the form of a phylogenetic tree, such a structure might not always be appropriate to capture the complex evolutionary picture of organisms, in that their past could have been influenced by non-tree like evolutionary events such as introgression which is known to effect butterfly evolution [1]. In such cases more general structures called (phylogenetic) networks have proven useful [1,2]. Originally introduced as single-rooted directed acyclic graphs these have recently been studied in multi-rooted form.

A major challenge in the context of networks in general is the development of methodologies and software tools for constructing them from biological data. In this poster, we present a recent encoding result for augmented trees (i.e. phylogenetic trees

with an edge weighting in terms of non-negative integers) [3], which are closely related to a type of multi-rooted network called an arboreal network.

References
[1] R. W. Wallbank et al., "Evolutionary novelty in a butterfly wing pattern through enhancer shuffling," PLoS Biology, vol. 14, no. 1, p. e1002353, 2016.
[2] K. T. Huber, V. Moulton, and G. E. Scholz, "Forest-based networks," Bulletin of Mathematical Biology, vol. 84, no. 10, p. 119, 2022.
[3] K. T. Huber and D. Overman, "Arboreal networks and their underlying trees," arXiv preprint arXiv:2503.22419, 2025.

## Genetic structure, spectral characteristics and evolutionary history of the European *Polyommatus eros* species group (Lepidoptera, Lycaenidae)

## Exploration of bacterial rDNA evolution using long-read datasets

Author(s): ***Laura Tingley***, *Katharina T. Huber, Jo Dicks*

Conservation of rDNA has underpinned bacterial species identification for decades. Housed within rDNA operons, bacteria contain 1-15 copies of the 16S rRNA gene, with unknown levels of sequence variation between them. So, how much intra- and inter-genome 16S variation exists? Can this alter identification outcomes? And how does this impact our understanding of rDNA evolution?

We utilised UKHSA's NCTC3000 dataset, comprising PacBio long reads from ~3000 diverse bacterial strains, to investigate. Long reads can span both the 16S gene and beyond the rDNA operon, achieving copy specificity. We assessed variation between distinct 16S sequence copies in 42 complete *Escherichia coli* genomes by developing a computational pipeline to extract operon copy-specific reads and create consensus sequences for the 7 distinct rDNA operons. Within- and between-strain analysis was conducted alongside *Shigella* strains to provide insight into the evolution of rDNA within these closely related taxa.

NeighborNet analysis of copy-specific sequences enabled us to develop a model of rDNA evolution in *E. coli*. Furthermore, 16S variation at the single nucleotide level is being evaluated to understand its impact on phylogenetic relationships within and between species and how it affects bacterial identification. In summary, our new pipeline is shedding light on bacterial rDNA evolution and enabling us to discriminate between phylogenetically similar but distinct bacterial species including *E. coli* and *Shigella*.

## Other

### Learning Disease Dynamics with Flow Models

Author(s): ***Kyriakos Flouris***

We explore a method for learning the relationship between epidemiological observations and underlying disease parameters using a flow matching model. By combining historical outbreak data with simulated trajectories, we train the model to approximate the posterior distribution over key parameters conditioned on observed data. We then propose a way to utilize this posterior to support predictions of future disease progression. This approach aims to provide a flexible and data-efficient framework for integrating mechanistic insight with modern deep learning tools in epidemiological forecasting.

### ZINBGT: Exploratory Data Analysis of Single-Cell Transcriptomic Expression Using Mixture Models

Author(s): ***Toby Kettlewell***, *Yiyi Cheng, Thomas D. Otto, Vincent Macaulay and Mayetri Gupta*

Single-cell RNA sequencing (scRNA-seq) provides data on the signals associated with protein production within individual cells. This allows for the discovery of novel cell types, inference of cell trajectories, and fine-grained comparisons of different tissues. The analysis of scRNA-seq data uses a pipeline of methods, but benchmarking is currently unable to establish which methods are best for a given dataset. Because the conclusions drawn from an analysis depend on the choice of method, novel forms of exploratory data analysis are needed to investigate how datasets differ and the circumstances in which a given method is likely to perform best. Any such method needs to run quickly and provide easily interpretable visualisations. A family of mixture distributions on count data will be introduced, which capture the salient aspects of gene expression with the associated parameters acting as summary statistics of each gene. A novel variant of a 2d histogram will be proposed, allowing efficient exploration and comparison of large, high-dimensional datasets, while problematic genes are highlighted using a combination of distance between model and data with bootstrapping. Human immune cells will be explored in terms of gene expression, and comparison with simulations will reveal differences that could compromise benchmarking.

### Clustering Inflammation, Inferring Causality: Endotypes in Traumatic Brain Injury

Author(s): ***Romit Samanta***, *Chiollaz, A. C., Needham, E., Yue, J. K., Helmy, A., Zanier, E. R., Wang, K. K. W., Kobeissy, F., Posti, J. P., Summers, C., Manley, G. T., Maas, A. I., Tenovuo, O., Sanchez, J. C., Menon, D. K., TRACK-TBI investigators and participants, CENTER-TBI investigators and participants*

The inflammatory response in patients with traumatic brain injury (TBI) offers opportunities for stratification and intervention. Our aim was to describe immunologically based sub-phenotypes that leant themselves to future precision medicine approaches.

A panel of 30 inflammatory mediators were measured in serum and plasma samples from two international, multi-centre observational studies of TBI (CENTER-TBI, Europe; TRACK-TBI, USA). Semi-supervised hierarchical clustering was used to identify sub-groups of patients with similar inflammatory signatures. Network methods were used to identify biological pathways associated with the mediators associated with poor outcomes. Penalised, elastic net regression was used to identify the best performing mediators for predicting outcomes. Causal inference methods (BRMS) were used to delineate the role of inflammatory response on mediating neurological outcomes whilst accounting for key confounders which included brain injury biomarkers, age and non-cranial injuries.

Two clusters were identified, termed early- and pauci-inflammatory, in both cohorts. A two-cytokine signature (IL-15 and MCP-1) was the best performing predictor of cluster assignment. Clusters associated with greater age but not extra-cranial injury. Inflammation demonstrated a significant mediated path for the association between initial injury and outcome when age was accounted for.

Plausible inflammatory subphenotypes can be demonstrated in patients with TBI that offer the opportunity for patient stratification.

## Towards a Stable and Scalable Taxonomy: Consolidating and Integrating Ruminant-Associated Metataxonomic Data for a Global Survey

Author(s): ***John-Paul Wilkins***

Amplicon-based metataxonomics has become a widely used tool for microbial profiling in ruminant studies. Heterogeneity in variable-region primers and classification schemes limits the potential for comprehensive meta-analysis. We have designed a computational framework to consolidate and integrate all existing ruminant-associated 16S rRNA gene sequence metataxonomic data, based on nucleotide sequence similarity alone. This de novo classification avoids inherited methodological biases and inconsistencies in reference databases. By interlinking databases through the full-length 16S sequence a unified classification is produced combining information from multiple variable regions. Our database is based on >300k 16S sequences from a wide range of microbial environments, supplemented by >20k 16S sequences extracted from ruminant-associated MAGs. These established classifications remain intact while underpinning the rapid, large-scale incorporation of additional metataxonomic data. Preliminary analysis shows that combining broad reference and niche-specific databases enhances taxonomic resolution of ruminant-associated microbes. By unifying disparate ruminant microbiome data under a single, scalable taxonomy, we particularly demonstrate the robust recovery of rare clades typically discarded as noise. The

integrated database enables the multidirectional flow of diverse information, ranging from study metadata to sequence-derived functional data, supporting downstream functional and evolutionary analyses. Future work will develop advanced statistical methods to address biological questions arising from the survey, including host–microbiome interactions and evolutionary dynamics.