MASAMB 2025 Seminar Overview

Day 1, Session 1 – Population Genomics

Detecting balancing selection using deep learning

Matteo Fumagalli, Sandipan Paul Arnab, Cindy Santander, Michael DeGiorgio

Balancing selection is a mode of natural selection that maintains genetic diversity through various mechanisms, including overdominance and negative frequency-dependent selection. However, distinguishing the genomic signature among balancing selection modes, and other forms of selection, is a significant challenge as these processes leave signals that are largely overlapping. In this talk, I will present a journey on how we first approached this problem from a statistical perspective. I will then move towards our efforts to develop deep learning solutions to this aim and discuss how the use of full genomic data, in contrast to summary statistics, helped improve our predictions. Afterwards, I will present some recent and ongoing work on integrating temporal data and transfer learning in our inferential framework. In this work, we demonstrate how resource-efficient deep transfer learning, combined with novel data preprocessing and the modeling of genomic autocovariation, can effectively characterise modes of balancing selection using either phased or unphased genotypes, and with or without temporal data from ancient DNA. Finally, I will offer practical recommendations for both empiricists and method developers on advancing the detection of transient balancing selection in genomic data.

Merging evolutionary timescales to quantify adaptation

Ioanna Kotari, Carolin Kosiol, Rui Borges

Positive diversifying selection promotes recurrent changes in amino acid sequences, driving adaptive evolution. In phylogenetics, it is commonly inferred using the ratio of non-synonymous to synonymous substitution rates (dN/dS). Under the assumption that synonymous substitutions are neutral, dN/dS can inform us about the direction and strength of natural selection. However, standard codon models typically rely on a single representative genome per species and neglect population-level processes such as selection on synonymous codons and GC-biased gene conversion (gBGC), both of which can bias dN/dS estimates and generate false positives, especially in the presence of within-species variation. The growing availability of genome sequences in the genomics era makes it possible to study adaptive evolution with greater resolution by combining interspecific and intraspecific variation. In this study, we introduce a Polymorphism-aware Phylogenetic Model for Codon Evolution (PoMo-cod), which models adaptive evolution at a population genetics scale, while leveraging both fixed and polymorphic differences between species. PoMo-cod advances codon modelling by integrating mutation biases, gBGC, synonymous codon and diversifying selection. We compare our framework with standard codon models to assess how population-level forces bias signals of positive selection. We observed that while the dN/dS-based inference remains a valid and useful tool, it can overestimate adaptive signals when population-level processes influence sequence evolution beyond idealised assumptions. This study highlights current limitations in detecting diversifying

selection and how population-aware approaches can improve the accuracy of evolutionary inference.

ESPClust: A Method for Unsupervised Identification of Effect Modifiers in Omics Studies

Francisco Pérez-Reche, Nathan Cheetham, Ruth Bowyer, Ellen Thompson, Francesca Tettamanzi, Cristina Menni, Claire Steves

High-throughput omics technologies have transformed our ability to link individual traits with biological characteristics. However, current analytical approaches often overlook the crucial role of covariates as effect modifiers, leading to a "one effect-size fits all" paradigm. This neglect risks skewed effect size estimates and disregards vital heterogeneity in human populations, which is fundamental for personalized medicine.

We introduce ESPClust, a novel unsupervised method to identify covariates modifying effect sizes in omics association studies. ESPClust extends effect modification by analyzing the 'effect size profile' (ESP)—a collection of effect sizes linking multiple omics variables to an outcome. The method divides the covariate space into regions with approximately homogeneous ESPs, thereby uncovering subpopulations with distinct associations. This approach allows for the identification of effect size modifiers even in datasets typically considered too small for traditional univariate stratified analyses.

We demonstrate ESPClust's versatility and ability to uncover nuanced effect size modifications through applications to synthetic data and real-world studies. These include blood metabolomics and insulin resistance, and pre-pandemic metabolomics influencing COVID-19 symptom manifestation. ESPClust provides a robust statistical framework for understanding complex omics data, holding significant promise for advancing personalized medicine by revealing how biological associations vary across individuals based on specific covariates.

Lost Founders, Found Haplotypes: Reconstructing Parental Genomes from Offspring Sequencing Data

Hadi Khan, Richard Durbin

In experimental genetics, the loss of founding parents of crosses before sequencing is a common and costly mistake. Without these genomes the precise origin of variation in subsequent generations is unknown which can stall further research. My work presents a solution: a program that computationally reconstructs founder genomes using only their descendants.

The software takes unphased VCF genotype data from multi-generational offspring (F1, F2, F3+) and reassembles the set of haplotypes which made up the founders. It works by finding patches of homozygosity in the offspring, converting these into blocks of haplotypes and subtracting the found haplotype from other samples to find the full set of haplotypes around each loci. It then uses inheritance and recombination patterns through the pedigree to infer the most probable ancestral sequences that link these block level haplotypes. This approach effectively resurrects

lost genetic data, enabling critical analyses like QTL mapping for crosses that were previously considered unusable.

The utility of this method extends beyond lab experiments. It provides a new framework for investigating the deep history of natural populations. For instance, one potential further development is to apply this logic to estimate the founding haplotypes of the initial human population that expanded out of Africa, offering a novel lens through which to view our own origins. My tool turns the incomplete genotypes of descendants into a complete picture of their ancestors.

Day 1, Session 2 – Reticulate Evolution

Inferring gene flow from phylogenies with too many genomes

Diogo Ribeiro, Rui Borges

Gene flow is now widely recognized as a central force in shaping species evolution, but despite its importance, it remains challenging to quantify across lineages. This task becomes even more complicated with confounding processes, such as ancient gene flow and incomplete lineage sorting, making it harder to reconstruct species history. Nonetheless, recent advances in sequencing technologies provide extensive genomic data needed to revisit these questions at scale. By integrating both within-species and between-species sequence variation, we aim to disentangle the contribution of gene flow and other short and long-scale evolutionary processes.

We introduce a new model of species evolution with gene flow (FlowT) that incorporates multiple interacting evolutionary forces, such as mutation bias and genetic drift, within a phylogenetic framework, facilitating analysis at evolutionary timescales where gene flow is understudied. FlowT leverages the observed allele frequencies, making it well-suited for large-scale genomic datasets. Additionally, because it automatically integrates all the possible genealogies to estimate the species tree, it avoids the computational burden of traversing the whole genealogical space.

To assess the performance of our model, we simulated multiple sequence alignments across a range of evolutionary scenarios. Using Bayesian inference, we accurately recover the mutation, gene flow rates, and divergence times. As such, these results demonstrate the model's ability to infer gene flow across phylogenetic timescales using large-scale genomic data.

Inferring gene flow between cryptic sibling species of *Anopheles* mosquitoes

Yuttapong Thawornwattana

Gene flow between species is an important evolutionary process that can facilitate adaptation and lead to species diversification. Despite a recent surge in studies of gene flow from genomic data, our understanding of the prevalence of gene flow across the tree of life remains

incomplete. This is partly because only a small fraction of organisms has genome data available and methods commonly used to study gene flow tend to be heuristic or approximate, with limited ability to detect gene flow. Here, we show that it is possible to study gene flow in groups of organisms for which no genomic resources exist and there is no prior evidence for gene flow. We generated new genome data for four groups of sibling species of Anopheles mosquitoes in North America. By performing full-data likelihood inference under the multispecies coalescent models of gene flow and applying Bayesian tests, we find strong evidence of gene flow in all four groups. Using a commonly used heuristic method failed to detect gene flow or detected gene flow that was not supported by genealogical heterogeneity. The resulting species phylogenies with gene flow not only clarify taxonomic status of these species but also lay the groundwork for studying the evolution of key traits in this group such as the ability to transmit human malaria. Overall, this work illustrates the feasibility of studying gene flow in new groups of organisms, paving the way towards better understanding the prevalence and role of gene flow across the tree of life.

Concatenation in the anomaly zone

Menno J. de Jong, Axel Janke

Inferring species trees from concatenated loci is often criticised for failing to account for gene tree discordance – particularly when using character-based methods. However, this criticism does not apply to distance-based concatenation trees, which can be shown to be statistically consistent even in anomaly zones. Building on this insight, we introduce DIST (Distance-based Inference of Species Trees), an intuitive and scalable method that infers species trees from population-level distance matrices containing multi-locus estimates of Dxy, FST or coalescence units (τ). DIST derives these values from between-individual sequence dissimilarity estimates, E(p), using basic equations from coalescence theory. Under certain conditions, DIST can also quantify gene tree discordance and distinguish whether it arises from gene flow or incomplete lineage sorting alone. While conceptually related to more sophisticated summary methods, DIST differs in that it does not seek the species tree which best explains a set of gene trees. Instead, it searches for the species tree which best explains an average gene tree, of which all branch lengths reflect mean coalescence time, E(t). Although this average gene tree is rarely observed empirically, it is approximated by an individual-level distance-based tree, traditionally referred to as a 'tree of individuals'. The DIST algorithm is implemented in the R package SambaR, which now accepts input in the form of pairwise E(p) estimates.

The Impact of Sequencing and Genotyping Errors on Bayesian Analysis of Genomic Data under the Multispecies Coalescent Model

Jiayi Ji, Paschalia Kapli, Tomáš Flouri, Ziheng Yang

The multispecies coalescent (MSC) model accounts for genealogical fluctuations across the genome and provides a framework for analyzing genomic data from closely related species to estimate species phylogenies and divergence times, test interspecific gene flow, and delineate

species boundaries. As the MSC model assumes correct sequences, sequencing errors at low coverage may be a serious concern. We used computer simulation to assess the impact of genotyping errors in phylogenomic data on Bayesian inference of the species tree and population parameters such as species split times, population sizes, and the rate of gene flow. The base-calling error rate is found to be extremely influential. At the low rate of e = 0.001 (phred score of 30), estimation of species trees and population parameters are little affected by genotyping errors even at low coverage of ~3x. At high error rates (e = 0.005 or 0.01) and low coverage (less than 10x), genotyping errors can reduce the power of species tree estimation, and introduce biases in estimates of population sizes and the rate of gene flow. Treating heterozygotes in the sequences as missing data (ambiguities) may reduce the impact of genotyping errors. We found it advantageous in terms of inference precision and accuracy to sequence a few samples at high coverage than many samples at low coverage.

Day 1, Session 3 – Machine Learning Applications

Characterization of selective pressures acting on protein sites with Deep Learning

Estelle Bergiron, Luca Nesterenko, Julien Barnier, Philippe Veber, Bastien Boussau

It is often useful to identify the selective pressures acting on a particular site of a protein to better understand its function. This is typically done with likelihood-based approaches applied to codon sequences in a phylogenetic context. However, these approaches are computationally costly. Here we use the phyloformer neural network architecture, which has been shown to be able to reconstruct accurate phylogenies from sequence alignments, to identify selective pressures acting on individual amino acid sites. We design different versions of the architecture and train and test them on simulations. We compare the results of one of our best models to the state-of-the-art approach codeml and find that it outperforms it when it is applied to data that resemble its training data, but that it performs less well when applied to data that does not resemble the training data. In all cases, our approach operates at a fraction of codeml's computational cost. These results suggest that a phyloformer-based architecture, trained on realistic simulations, could compare favorably to state-of-the-art approaches to characterize selection pressures acting on coding sequences.

Title To Be Announced

Luc Blassel

Phylogenetic inference, the task of reconstructing how related sequences evolved from common ancestors, is a central task in evolutionary genomics.

The current state-of-the-art methods exploit probabilistic models of sequence evolution along phylogenetic trees, by searching for the tree maximizing the likelihood of observed sequences, or by estimating the posterior of the tree given the sequences in a Bayesian framework.

Both approaches typically require computing likelihoods, which is only feasible under simplifying assumptions such as independence of the evolution at the different positions of the sequence, and even then remains a costly operation.

Here we present Phyloformer 2, a likelihood-free inference method for posterior distributions over phylogenies, trained end-to-end from sequences to trees.

Phyloformer 2 exploits a novel encoding for pairs of sequences that makes it more scalable than previous approaches, and a parameterized probability distribution factorized over a succession of subtree merges.

The resulting network outperforms both state-of-the-art maximum likelihood methods and a previous likelihood-free method for point estimation in topological accuracy. Phyloformer also runs orders of magnitude faster than even distance methods, leveraging the parallel processing power of GPUs.

It opens the way to fast and accurate phylogenetic inference under realistic models of sequence evolution.

Re-analysis of RNASeq Data suggests alternative explanations to long distance transport of mRNA: ∀ grafting RNASeq experiment ∃ plenty of false positives

Pirita Paajanen

Short-read RNA-seq studies of grafted plants have led to the proposal that thousands of messenger RNAs (mRNAs) move over long distances between plant tissues, potentially acting as signals and promising ample biotechnology applications. To curate a well-founded dataset for machine learning applications, I downloaded all the existing data from sequencing archives and performed a meta-analysis of existing mobile mRNA datasets and examined the associated bioinformatic pipelines. Taking technological noise, biological variation, potential contamination and incomplete genome assemblies into account, we find that a high percentage of currently annotated graft-mobile transcripts are left without statistical support from available RNA-seq data. This meta-analysis challenges the findings of previous studies and current views on mRNA communication and shows the power of mathematics and statistics in molecular biology studies.

AliFilter: a Machine Learning Approach to Alignment Filtering

Giorgio Bianchini, Rui Zhu, Francesco Cicconardi, Edmund RR Moody

Many applications in bioinformatics and computational biology employ multiple sequence alignments, which are used to identify homologous residues in nucleotide and protein sequences. However, highly divergent sequence alignments often contain a significant proportion of noise. Reducing this noise is normally achieved through filtering the alignment by trimming columns

that are poorly aligned or offer minimal useful information; either automatically using software, or manually by visualising the alignment and identifying regions to remove. Manual approaches are labour-intensive and less reproducible, but can utilise the researcher's specialist knowledge, rather than relying on filtering criteria that might not be adequate for each alignment.

AliFilter is a new tool that uses machine-learning to automate manual alignment filtering. AliFilter creates a model from a small number of manually annotated alignments, then uses this model to accurately reproduce the manual annotation (98% accuracy), while being resilient to mistakes in the training data. Users can use the program with a default model (included) or create customised models for individual datasets or filtering criteria. AliFilter reduces the execution time of tree inference by 35% for a phylogenomics-level dataset, whilst retaining results that were almost identical to the full alignment, unlike other alignment filtering tools we tested. AliFilter is a free and open-source software written in the C# programming language; it is distributed under a GPLv3 licence from https://github.com/arklumpus/AliFilter, from where both the source code and standalone executables for Windows, macOS and Linux can be downloaded.

Deep learning: Balancing, linkage and effects of selection (DEEPBLUES)

Antonio Pacheco

The phenomenon of Balancing Selection (BS) observed across a multitude of organisms refers to the processes that sustain genetic variation over large numbers of generations. The detection of BS continues to be a challenge in the field of evolutionary genomics, being often entangled with Linkage Disequilibrium (LD). The subtle nature of BS, and its complex interactions with factors such as LD have significant effects even over timescales longer than speciation events. The implementation of polymorphism-aware phylogenetic models (PoMos) allows the inference of phylogenies and directional selection and, more importantly, BS with the recently developed PoMoBalance. However, the influence of LD on detecting BS within this framework needs to be investigated further.

We combine the PoMoBalance approach with the training of convolutional neural network CNNs. The CNNs are applied to data simulated with SLiM in different scenarios combining LD and BS with corresponding ancestral recombination graphs (ARGs). Training and testing data are simulated from a three population out of Africa demographic scenario for humans, incorporating the effect of dominance at intermediate allele frequencies that represent BS.

We present a comprehensive comparison of the performance of our PoMoBalance with CNN approach using (i) standard population genetic summary statistics (SS) such as site frequency spectra and Tajima's D, and (ii) SS that tree based including features of the ARGs calculated from branch lengths, tree topology and lineage-through-time plots. In particular, we have tested the performance for classification of balancing selection and directional selection under the influence of complex recombination and demographic scenarios.

Hierarchical Patterns of Soil Biodiversity in Extreme Environments: Insights Across Biological Scales

Laura Villegas, Laura Pettrich, Esteban Acevedo-Trejos, Lucy Jiménez, Arunee Suwanngam, Nadim Wassey, Miguel L Allende, Alexandra Stoll, Oleksandr Holovachov, Ann-Marie Waldvogel, Philipp H. Schiffer

Information about geographical patterns of biota, species diversity and distribution, is scarce for soils, despite their pivotal role as ecosystems. The Atacama is the driest non-polar desert on earth and it is believed that only specialized taxa can survive there. Above ground invertebrates have been reported in the Atacana Desert but its soils have not been comprehensively analyzed. By studying different areas across the Atacama, we aimed to better understand resilience of soil organisms in times of global aridification. Facing, two major methodological challenges we investigated diversity of soil nematodes at the genomic, genetic, taxonomic, community and life-cycle levels: firstly the vastness of the area, which makes comprehensive sampling all but impossible and secondly the large number of new, undescribed species, which cannot be easily assigned to taxa. We thus implemented an approach using (statistical) classifiers to model the distribution of species in space. In a second approach we used machine learning methods to also assign nematodes to the genera and species level based on the presence and absence of conserved genetic elements. From this, we predict that distribution of asexual taxa is more likely to occur at higher altitudes, and that the distribution of genera richness in the Atacama follows a latitudinal diversity gradient and is influenced by (rare) precipitation. We also show that using classifiers it will be possible to assing species level identity at the OTU level to novel taxa. Our work shows that even under extreme environmental conditions stable, healthy soil communities can persist.