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.

Title To Be Announced

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.

Polymorphism-Aware Phylogenetic Models (PoMo) for Species Delimitation

Wenjie Zhu

Species delimitation—the identification of evolutionary distinct lineages—remains a fundamental challenge in evolutionary biology. Traditional morphology-based approaches face significant limitations: many taxonomic groups exhibit cryptic diversity with minimal phenotypic differentiation, morphological characters can be plastic or convergent, and subjective interpretation introduces inconsistency. The increasing availability of population-level genomic data, particularly single nucleotide polymorphisms (SNPs), provides unprecedented opportunities for objective, model-based species delimitation that can reveal cryptic species and test alternative delimitation hypotheses quantitatively.

Model-based approaches such as BPP and BFD* have revolutionized species delimitation by incorporating evolutionary processes and allowing statistical comparison of competing models. Polymorphism-aware phylogenetic models (PoMo) offer a complementary framework that explicitly models allele frequency evolution along phylogenetic trees, bridging population genetics and phylogenetics. Originally developed for phylogenetic inference, PoMo accounts for both evolutionary divergence and within-species genetic variation.

We present an extension of PoMo for Bayesian species delimitation using SNP data, implemented in RevBayes with Bayes factor model selection. We evaluate our approach through simulation studies and apply it to an empirical firefly (Luciola) SNP dataset, comparing performance against established delimitation methods.

Our preliminary investigations suggest that PoMo-based species delimitation shows promise, particularly in terms of computational speed and statistical robustness, potentially expanding the methodological toolkit for challenging delimitation problems in morphologically cryptic taxa.

Day 1, Session 3 – Reticulate Evolution

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.

Title To Be Announced

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.

Day 2, Session 1 – Comparative & Evolutionary Genomics

Unsupervised learning as a tool to retrieve genomes from undersampled taxa: Fast and slow evolution in myxozoans

Claudia Weber

Myxozoans are obligate endoparasites that belong to the phylum Cnidaria. Compared with their closest free-living relatives, they have evolved highly simplified body plans and reduced genomes. However, little is known about their genome architecture because of a lack of sufficiently contiguous genome assemblies. This work presents two new Kudoa genomes, one of them near-chromosomal, built entirely from low-coverage long reads from infected fish samples.

The results illustrate the potential of using unsupervised learning methods to disentangle sequences from different sources, and facilitate producing genomes from undersampled taxa. Extracting distinct components of chromatin interaction networks allows scaffolds from mixed samples to be assigned to their source genomes. Meanwhile, low-dimensional embeddings of read composition permit targeted assembly of potential parasite reads.

Despite drastic changes in genome architecture in the lineage leading to Kudoa and considerable sequence divergence between the two genomes, gene order is highly conserved. Although parasitic cnidarians show rapid protein evolution compared with their free-living relatives, there is limited evidence of less efficient selection. While deleterious substitutions may become fixed at a higher rate, large evolutionary distances between species make robustly analyzing patterns of molecular evolution challenging. These observations highlight the importance of filling in taxonomic gaps, to allow a comprehensive assessment of the impacts of parasitism on genome evolution.

Building a scalable bioinformatics strategy for sequencing historical fungal collections

Lia Obinu, Wu Huang, Niall Garvey, George Mears, Torda Varga, Maria Kamouyiaros, Ben Price, Ester Gaya

Whole genome sequencing (WGS) of a comprehensive fungal collection offers transformative potential for resolving the fungal tree of life, uncovering novel metabolic pathways, and informing biodiversity conservation. However, sequencing historical specimens, and especially old type material, featuring degraded DNA at scale introduces unique technical and analytical challenges. The variety of biological features of the specimen, inconsistent source of contaminants, the level of contamination, and DNA degradation are all key considerations. As part of the ongoing Fungarium Sequencing Project (FSP) at Kew, we are developing and optimising a dedicated bioinformatics pipeline to process over 7,000 fungal and lichen genomes, with genomic material extracted from type fungarium samples spanning decades or even hundreds of years old. Here, we present a bioinformatic roadmap towards thousands of fungarium genome assemblies.

Building upon 300 fungal genomes generated in a pilot sequencing study, we identified critical decision points in the analytical workflow — from assessing pre-sequencing quality metrics to guide lab work, read preprocessing, contamination detection and removal, to developing hybrid

assembly approaches to scale up the workflow to suit the needs of a large-scale genome assembly endeavor. The insights gained directly inform the computational strategy for the FSP, ensuring data quality, reproducibility, and efficient release to public repositories. This methodological roadmap will not only accelerate the delivery of the FSP but also establish best-practice guidelines for processing genome sequences from historical fungal collections globally.

Quantifying structural variants in chromosomes using landmark-based disparity

Jeff Streicher

Chromosomal architecture has played a key role in the evolution of biodiversity. Detecting structural variants (SVs) on chromosomes has informed the study of speciation, sex determination, adaptation, and some of the earliest divergences in the tree of life. Here we present a computationally non-intensive approach, based on geometric morphometrics, that uses conserved DNA sequences as landmarks to quantify structural disparities of focal chromosomes across multiple species, individuals, or cell types. Based on two approaches, we show that this 'geno-metric' method can be applied at micro- and macroevolutionary scales to discover and diagnose SVs. Using human X-linked genes and ultraconserved elements as landmarks, we provide empirical demonstrations with amniote sex chromosomes, the Drosophila virilis group, and placental mammal genomes. Landmark-based structural disparity analysis effectively identifies chromosomal rearrangements and has parallels with traditional morphometrics regarding chromosome size, landmark orientation and landmark availability. Using simulations, we show that structural disparity inferred from ultraconserved elements is correlated with overall levels of chromosome evolution; an attribute which is consistent with observed disparity between and within mammalian orders. We also found that the disparity patterns of SVs have significant phylogenetic signal, giving them broad importance for studying evolutionary biology. Structural disparity analyses are a valuable addition to the comparative genomic toolkit in that they offer an intuitive, rapid mechanism for detecting SVs associated with single copy genetic landmarks and the potential to reveal broader patterns of genome evolution.

Title To Be Announced

Artemis Kotoula

Octocorallia are a group of invertebrates crucial for the sustainability of marine environments. Their ecological importance is evident through their formulation and preservation of reef structure, which aids in supplying habitats and maintaining biodiversity for a wide range of marine species. In spite of their significance, their genomes stay understudied when compared to other marine animals. This project takes advantage of coral reference genome assemblies to study their biodiversity and adaptations using methods like whole genome alignment, evolutionary rate analysis, synteny, orthology inference and gene family evolution. Comparing and combining the results of these methodologies across species allows for the observation of conserved and divergent coding and non-coding genomic regions, structural variations, gene duplication and loss. This way, insights can be obtained about the evolution and the genomic biodiversity of

Octocorallia, and light can be shed on adaptations that these species might have developed due to a wide range of environmental parameters. Studying this diverse group of species can help not only fill in an existing knowledge gap in the field of invertebrate genomics by understanding their biodiversity, ecological importance, and evolutionary history, but can also help with revealing potential coral vulnerabilities and adaptive pathways, offering insights into the sustainability of varying marine ecosystems.

Investigating contamination events in SARS-CoV-2 genome data

Olivier Anoufa, Nicola De Maio and Nick Goldman

Contamination can occur during genome sequencing when foreign DNA mixes with the sample's genome. This can cause the sample's consensus sequence to incorrectly incorporate parts of the contaminant's genome. Such errors can negatively affect the inference of the evolutionary histories and recombination.

Here, we investigate contamination and related biases in SARS-CoV-2 genome data. We propose new computationally efficient approaches to identify contaminated samples and mixed infections and prevent associated errors in consensus genomes. Two methods are presented. The first consists in masking genome positions with unexpectedly low sequencing depth. The second is a Hidden Markov Model describing contamination and its impact on sequence read data and consensus genomes.

We apply these methods to 4,471,579 SARS-CoV-2 sequences to both identify contaminated samples and to correct their consensus sequences. Using the first method, we were able to detect hundreds of putatively contaminated samples.

Advancing the classification of variants of uncertain significance: analysis frameworks for precision genome editing at the single cell level

Magdalena Strauss

Determining whether DNA mutations labelled as variants of uncertain significance are harmful or benign remains a major obstacle to clinical application. To address this, we developed experimental and statistical tools to map mutational impact by integrating single-cell RNA and DNA sequencing in the gene editing context. This talk presents our mathematical and statistical tools, which reveal how individual mutations disrupt cell function and alter gene regulation.

We studied cellular responses to interferon gamma (IFNγ)—a molecule that signals immune activity—across different mutations to the JAK1 gene, demonstrating the accuracy of our tools by linking genotype with transcriptional phenotype and achieving low error rates for known genotype-phenotype relationships. Our approach identified transcriptional heterogeneity of IFNγ response across groups of JAK1 missense mutations, highlighting its potential for systematic classification of variants of uncertain significance.

Additionally, we show how our computational methods improve scDNA-seq and scRNA-seq analysis in gene editing, enhancing detection of biological signal amid technical noise in an application to DNA repair. In another application, we analysed transcriptional profiles of drug-resistant colon cancer cells after exposure to dabrafenib and cetuximab, characterising distinct types of drug resistance. Finally, we present ongoing work assessing the pathogenicity and cellular impact of missense mutations in a prevalent lung cancer oncogene, and integrating data using hierarchical models.

VI-guided NSGA-II: a novel Evolutionary Multi-Objective Optimisation Algorithm for Feature Selection for Single-cell Classification

Rowbottom, M., Strauss, M. and Aishwaryaprajna

Single-cell RNA-sequencing (scRNA-seq) has made it possible to understand cellular heterogeneity at a detailed level, and to classify cells into known functional categories. However, the identification of the most informative features driving this heterogeneity poses an ongoing challenge. To address this, we formulated the problem in terms of multi-objective optimisation, with the dual conflicting aims of maximising classifier accuracy while minimising the number of selected features. Non-dominated Sorting Genetic Algorithm II (NSGA2) is a highly effective multi-objective evolutionary algorithm with applications across a broad spectrum of feature selection problems. We propose a modified NSGA2 algorithm for exploring the trade-off between feature list size and the per-class classification accuracy, and apply it to colorectal cancer cells with a broad range of DNA mutations that each represent one of three distinct categories. The developed algorithm is modified to leverage existing knowledge on the interactions between the considered gene features, directly guiding the algorithm's learning process. Evaluation is performed through comparison with a generic NSGA2 algorithm as well as a LASSO regression approach. Further experimentation is then performed on the broader impact of both warm-start population initialisation and the application of alternative model evaluation metrics in optimisation. Such comparisons facilitate an exploration into the potential for evolutionary algorithms to streamline the feature selection process in scRNA-seq analysis. Furthermore, in understanding what features best distinguish the different functional groups, candidates for future study into the inhibition of tumour cell growth can be identified.

Day 2, Session 2 – Phylogenomics

Phylogenetic inference with not-so-rare mutations and wee tiny organisms

Rui Borges

A common assumption in molecular evolution is that mutations are rare. This assumption is widely used in the development of evolutionary theory, an example being the infinite sites model, which posits that the genome has infinitely many sites, such that each new mutation necessarily occurs at a unique site. This is clearly a simplifying assumption, and it is well established that recurrent mutations - particularly in highly diverse organisms such as bacteria and viruses - occur

frequently. In this study, we challenge the assumption of rare mutation in phylogenetic inference. In a recent study, we demonstrated that ignoring recurrent mutations leads to inflated estimates of effective population size. We therefore expect that unaccounted recurrent mutations influence the rate of molecular evolution and, consequently, distort phylogenetic inference. To investigate this, we developed two models of evolution: one that incorporates recurrent mutations, and another that allows only boundary mutations (akin to the infinite sites model). We generated multiple sequence alignments using the former and inferred phylogenies using the latter. While the inferred tree topologies were consistent across models, the estimated branch lengths and mutation rates were substantially biased. We also show that standard nucleotide substitution models fail to accurately estimate phylogenies - although it appears that the correct topology may still be recovered given sufficiently large data. Finally, we discuss the implications of these findings for phylogenetic inference, drawing on viral case studies of epidemiological and medical relevance.

Modelling compositional and exchange rate changes over time

Peter Foster

The process of evolution can change over time, including compositional tree heterogeneity (CTH) and exchange rate tree heterogeneity (ERTH). Models that can accommodate CTH and ERTH in molecular evolution are described. Fit of these models was compared using a likelihood ratio test in maximum likelihood, and in Bayesian analysis using the conditional predictive ordinate (CPO)-based log pseudomarginal likelihood (LPML), also leave-one-out cross-validation (LOO-CV). CTH and ERTH can be flexibly modelled in a Bayesian framework with tree-heterogeneous models that tune themselves to the amount of heterogeneity in the data being analysed. Since phylogenetic analysis is usually done using tree-homogeneous models, effects of CTH and ERTH on subsequent phylogenetic analysis using such models were described. Compositional effects due to CTH were seen as expected, for example where unrelated taxa with similar compositions would group together in homogeneous analysis. Similar effects were also demonstrated due to ERTH. Detection of CTH and ERTH by modelling is compared to detection using matched pairs tests (MPTs) that have been used to test molecular sequences for stationarity, reversibility, and homogeneity (SRH). Comparisons between modelling and MPTs on data simulated on very simple trees showed that the two approaches were equivalent, but simulations on larger trees showed that the two approaches differed greatly. Modelling showed greater power, especially in detection of ERTH, and some ERTH was completely invisible to MPTs but was decisively detected by modelling. Detection and modelling of CTH and ERTH is shown in two empirical examples.

Is the deuterostome clade real?

Serra Silva A., Natsidis P., Piovani L., Kapli P., and Telford M.J.

Much of our understanding of early animal evolution rests on the existence of two bilaterian clades, Deuterostomia and Protostomia. An exhaustive comparison of multiple independent phylogenomic datasets revealed disparate levels of support for these clades; however, while

strong support for protostome monophyly is widespread across datasets, support for deuterostomes is equivocal and linked to conditions known to lead to systematic errors in tree inference (inadequate substitution models, presence of long-branches and short-internal branches). Using a new large metazoan dataset, we systematically explored sources of error suspected to underpin support for Deuterostomia. To parse the effects of long-branch artefacts and inadequate substitution models, we compared the support for Deuterostomia under site-homogeneous and site-heterogeneous substitution models on two sets of taxon jackknifed alignments – including or excluding long-branched taxa. This systematic approach confirmed that, when sources of error are mitigated, it is nearly impossible to distinguish between monophyletic Deuterostomia and its paraphyletic alternatives, and that long-branch artefacts have a higher impact on support for monophyletic deuterostomes than model inadequacy. Our results suggest that even if Deuterostomia is monophyletic, many of its purported synapomorphies were probably present in the last common ancestor of Bilateria. These results have implications for our understanding of bilaterian relationships and evolution.

Scalable phylogenetic inference with long indels

Mattes Mrzik, Julija Pecerska, Maria Anisimova, Manuel Gil

The rapid expansion of genomic data creates a demand for scalable phylogenetics. A central challenge in phylogenetic modeling is the treatment of indels, which are often ignored—either by treating gaps as missing data or by removing or trimming gap-rich alignment regions. Yet, indel patterns can be highly informative for tree inference, especially at deeper divergences. Recent work from our group included modeling indels as single site events, which provides a simple solution to reduce computational complexity, although at the expense of making unrealistic assumptions. Although Bayesian methods can accommodate complexity, such as multiple-site and overlapping indels, they scale poorly, making it impossible to analyze larger datasets. Consequently, the substantial inferential power offered by larger datasets cannot be fully harnessed.

We present a scalable frequentist approach for tree inference based on the long-indel model TKF92. Since the calculation of the marginal probability of the alignment given the tree is exponential under this model, we include indel histories as a dependent variable and thereby achieve an algorithmic complexity that is linear in the number of sequences and the length of the alignment. Our method iteratively co-optimizes tree topology and indel histories: tree improvements are made via likelihood-increasing moves and edge length adjustments, followed by reestimation of indel histories for affected edges.

This framework leverages the phylogenetic signal encoded in indel patterns while maintaining computational efficiency, enabling robust inference on large and diverse genomic datasets. Additionally, this iterative frequentist optimization can be extended to include a realignment step, enabling joint inference of alignment and tree.

Inferring niche shifts from phylogenies and species distributions

Nathan Clark, Josh Tyler, Will Pearse

One of the perennial questions in evolutionary ecology is how niche space is opened and partitioned by distinct evolutionary lineages. The modern scale of genetic and ecological datasets allow us to test this by identifying when particular clades are undergoing rapid ecological change. We propose an extension to Bayesian phylogenetic regression that adds learnable scalings of branch length, giving us a new method to explore and test for shifts in niche-evolution rate across the tree.

We empirically test this on a plant community dataset from Utah (USA). We show that niche shifts are more common in clades that are closely related to invasive species, suggesting that the ability to coexist is conserved in evolutionary time. We outline how this method can be used to tease apart the ecological and genetic factors determining species distributions without requiring phenotypic information.

Day 2, Session 3 – Cancer Evolution & Genomics, and Others

The dynamic fitness landscape of ageing haematopoiesis through clonal competition

Nathaniel V. Mon Père, Francesco Terenzi, Benjamin Werner

For most organisms, the soma participates little (if at all) in its species' evolution. Yet, despite the brevity of a single lifetime, it is not exempt from the forces of evolution, which drive it to age and ultimately develop cancer. Recent advances in sequencing have emphasised the ubiquity of fitness-driven evolution in human tissues. Blood presents an excellent example of this, where large clonal expansions are observed in healthy elderly individuals – a condition termed clonal haematopoiesis. Yet how precisely this evolution unfolds remains to be understood.

We investigated the evolutionary dynamics of the haematopoietic stem cell pool by analysing two large scale datasets of different type: one with clone trajectories from multi-timepoint sequencing, and the other single stem cell-derived colony expansions.

By applying population genetics models to predict clone trajectories and sample site frequency spectra, we showed that clonal competition plays a significant role in shaping distinctive fitness landscapes. Applying Bayesian statistics, we quantified the distribution of fitness effects and the occurrence rate of fit mutants in blood. These calibrations we then used to estimate the fitness effects of 2000 haematopoietic clone trajectories across three distinct datasets. This revealed a multi-stage mode of clone evolution, where the fittest clones only occur through accumulation of multiple beneficial mutations.

Modelling the evolutionary dynamics of multiple extrachromosomal DNA types in cancer

Elisa Scanu, Benjamin Werner, Weini Huang

Extrachromosomal DNA (ecDNA) is an important driver of genomic heterogeneity and rapid adaptation in cancer. By enabling oncogene amplification outside chromosomal constraints, ecDNA promotes accelerated tumour evolution and therapeutic resistance. While single-species ecDNA dynamics have been studied, tumour cells often harbour multiple ecDNA types, whose interactions and evolutionary trajectories remain poorly understood.

We present a mathematical model for the evolution of multiple ecDNA types within a proliferating cancer cell population. The model integrates stochastic processes of ecDNA replication, segregation, phenotypic and genotypic alterations and interspecies interactions, enabling us to explore how different ecDNA types co-evolve and influence overall system dynamics. We derive analytical and computational results characterising ecDNA copy number distributions, the emergence of dominant configurations, and the stability of ecDNA heterogeneity over time.

Our results reveal that inter-ecDNA interactions can produce non-trivial population structures and suggest conditions under which ecDNA diversity is maintained or collapses. This work provides a quantitative framework for understanding the evolutionary logic of ecDNA-driven genomic architecture in cancer and raises new questions about the role of ecDNA cooperation and competition in tumour adaptation.

Our approach contributes to the growing interface between mathematical oncology and computational modelling, offering a novel perspective on cancer evolution grounded in population dynamics. It lays the groundwork for future integration with genomic data and provides a flexible framework to explore how ecDNA dynamics shape tumour heterogeneity.

Dual clustering approaches for pathway discovery and risk stratification using longitudinal high-dimensional biomarker data

Julie Fendler, Paul D.W. Kirk

We develop and apply statistical methods for characterising high-risk populations using longitudinal biomarker data across multiple time points, with applications to complex clinical outcomes. Our framework employs two complementary clustering approaches to identify both underlying putative biological mechanisms and clinically relevant patient subgroups.

We first group proteins that show similar expression patterns into co-expression modules with shared biological functions at specific time points, and investigate whether these modules are expressed differently in patients who develop adverse outcomes compared to controls. This helps to shed light on which biological pathways might be involved in disease progression or adverse events.

We subsequently group patients based on their overall biomarker profiles and risk levels, and identify the protein markers that characterise the discovered high-risk subgroups.

We describe the computational methods for both analyses, including variable selection procedures and clustering algorithms optimised for high-dimensional biomarker data leveraging temporal structure. Our dual framework—examining both biomarker co-expression networks

and patient risk stratification—provides complementary insights into putative disease mechanisms while identifying clinically actionable risk signatures.

Potential applications include maternal health outcomes, cancer prognosis, and cardiovascular risk assessment. The integration of these methods offers a comprehensive framework for precision medicine, potentially improving risk prediction across diverse clinical contexts and enabling personalised treatment strategies.

Jointly Modelling RNA-Seq and Legacy Microarray Data for Improved Power in Biomarker Discovery

Thomas Chen

While RNA-Sequencing (RNA-Seq) is now the standard method for measuring gene expression, numerous collections of valuable legacy microarray datasets exist. Although statistically challenging, the successful integration of these heterogeneous data types would increase sample size and statistical power, enabling more robust scientific findings in areas such as gene regulation, target identification, and patient stratification.

Existing methods do not adequately deal with platform-specific differences, most importantly the background effect in microarray data caused by inefficient hybridisation, and do not perform well at identifying differentially expressed (DE) genes. Therefore, we propose a novel Bayesian hierarchical model for joint analysis of microarray and RNA-seq data. This model focuses on summary data for computational efficiency, estimates the size of the background interference in microarray, and includes a jointly estimated correction factor, derived from first principles, dependent on the estimated background effect, the gene-specific effect size, and the measured gene expression.

Simulations demonstrate strict control of the FDR and Type I error rates under the model assumptions and realistic perturbations of them. In contrast, the inverse-variance weighting method suffers from inflated error rates (FDR >58%, Type I error >28%). We applied the model to four datasets (one RNA-Seq, three microarray) on early- vs late-stage breast cancer. The model identified 83 DE-genes, whereas no significant hits were found using Storey's method on the RNA-Seq dataset in isolation.

Our model provides a robust framework for incorporating legacy microarray datasets leading to increased statistical power.

Dynamic Adaptive Sampling for Human Trio Sequencing

Isabel Poetzsch, Nicola De Maio, Nick Goldman

Rare genetic diseases are commonly diagnosed using whole genome trio sequencing. This sequencing mode is among the most expensive and time-consuming because it requires sequencing three entire diploid human genomes at high coverage, i.e. the patient plus their two – usually healthy – parents. The resulting sequences are compared and analysed to identify a set

of candidate variants in the patient. Compared to the size of the three genomes, the candidate set will be drastically smaller such that a more targeted sequencing approach, like dynamic adaptive sampling, could lead to greater efficiency.

Dynamic adaptive sampling as implemented in BOSS-RUNS (Weilguny et al. 2023) builds on the ability of nanopore sequencing to reject fragments before their sequencing is completed. Whether a fragment is rejected is encoded by the decision strategy, which is dynamically recalculated as the sequencing experiment progresses based on what has already been sequenced, as well as the time cost of rejection and resampling. This allows redistributing the sequencing capacity to sites of particular interest as identified by the genome patterns gathered earlier in the experiment. We believe such an approach could yield a reduced sequencing load, reducing cost as well as sequencing time.

In this talk, we present our initial progress on showing feasibility and potential gains of applying dynamic adaptive sampling to trio sequencing. We report the quantities of data needed to identify single nucleotide variants in trio data according to our Bayesian model and show our novel test setup to evaluate the benefit of different sequencing strategies.

Flexible and efficient count-distribution and mixed-model methods for eQTL mapping with quasar

Jeffrey M Pullin, Chris Wallace

Identifying genetic variants that affect gene expression — expression quantitative trait loci (eQTLs) — is a major focus of modern genomics. Today, various tools exist for eQTL mapping, each using different statistical and methodological approaches. However, it is unclear which approaches lead to better performance, and challenges, particularly scalability as datasets continue to increase in size, remain. Here, we introduce quasar, a flexible and efficient C++ software program for cis and trans eQTL mapping. Compared to existing eQTL mapping methods, quasar implements a wider variety of statistical models, including the linear model, Poisson and negative binomial generalised linear models, linear mixed model and Poisson and negative binomial generalised linear mixed models. Methodologically, we introduce and implement a faster, analytic approximation to the score test variance in mixed models. Furthermore, we highlight that difficulties with accurately estimating the negative binomial dispersion parameter, previously identified in the context of RNA-seq differential expression analysis, also apply to eQTL mapping. Therefore, quasar implements the Cox-Reid adjusted profile likelihood (APL) which enables unbiased estimation of the negative binomial dispersion parameter. We assess quasar's performance and compare it to three existing eQTL mapping methods – apex, jaxQTL and tensorQTL – on the OneK1K dataset. We demonstrate that quasar's output agrees with existing methods where their methodology aligns but that quasar is 6-264x faster. One benefit of having multiple approaches within the same software package is the ability to compare statistical models for eQTL mapping without confounding by implementation. We find that: count-based models have higher power, that mixed models do not show better performance in a dataset without substantial relatedness, and that the adjusted profile likelihood improves Type 1 error control when using the negative binomial likelihood. Overall, quasar provides a performant and

versatile program for eQTL mapping and we nominate the negative binomial GLM model, incorporating APL dispersion estimation, as the statistical model with the best performance.