MASAMB 2025 Seminar Overview

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.

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 – Medical Genomics

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.