28th Annual New Zealand Phylogenomics Meeting

Tuesday 11 February – Friday 14 February 2025
Sudima Hotel, Kaikoura, New Zealand

Programme & Abstracts

Organisers (University of Canterbury)

- Scott Collier
- Charles Semple
- Mike Steel

Participants

Quentin Atkinson	University of Auckland
Lars Berling	University of Canterbury
Remco Bouckaert	University of Auckland
David Bryant	University of Otago
Minh Bui	Australian National University
Bruna Caveion	University of Canterbury
Alan Cooper	Charles Sturt University
Kylie Chen	University of Auckland
Alexei Drummond	University of Auckland
Andrew Francis	University of New South Wales
Nikolaos Gasparatos	University of Canterbury
Alex Gavryushkin	University of Canterbury
Russell Gray	Max Planck Institute for Evolutionary Anthropology
Jasmine Hall	Victoria University of Wellington
Simon Harris	University of Auckland
Sebastian Höhna	Ludwig Maximilians Universität München
Michael Hendriksen	University of New South Wales
Daniel Huson	University of Tübingen
Benedict King	Max Planck Institute for Evolutionary Anthropology
Egor Lappo	Stanford University
Maayan Levy	University of Canterbury
Teng Li	University of Auckland
Yu Liu	Beijing Normal University
Simone Linz	University of Auckland
Peter Lockhart	Massey University
Kerry Manson	University of Canterbury
Nicholas Matzke	University of Auckland
Trish McLenachan	Massey University

Masafumi Obara	University of Auckland
Tobia Ochsner	University of Auckland and ETH Zurich
Marcus Overwater	ETH Zurich
Andrew J. Roger	Dalhousie University
Tony Schaufelberger	University of Auckland
Charles Semple	University of Canterbury
Christine Simon	University of Connecticut
Mike Steel	University of Canterbury
Mark Stukel	University of California, Davis
Douglas Theobald	Brandeis University
Kate Truman	University of Canterbury
Geoff Whittle	Victoria University of Wellington
Sophia Witham	University of Canterbury
Yao Xiao	University of Auckland
Walter Xie	University of Auckland
Zijing Yang	University of Auckland
Lilin Zhang	University of Canterbury
Terry Zhang	University of Canterbury

Programme

All talks as well as all morning and afternoon teas are in the Pacific (whole room).

Tuesday 11 February

10.30 am	**Morning Tea**
12.00 pm	**Lunch** (Hiku Restaurant and Bar)
2.00 pm	Chris Simon *Global genomics of the man-o'-war (Physalia) reveals biodiversity across the ocean surface*
2.20 pm	Lars Berling *Statistics in the space of ranked time trees*
2.40 pm	Teng Li *Biodiversity nano-soup: Rapid biodiversity assessment using Oxford Nanopore sequencing*
3.00 pm	**Afternoon Tea**
3.30 pm	Daniel Huson *Working with phylogenetic trees and networks using PhyloSketch*
3.50 pm	Alan Cooper *Hard sweeps in feral genomes: A uniquely powerful evolutionary experiment*

Wednesday 12 February

8.45 am	**Arrival Tea and Coffee**
9.00 am	Russell Gray *Does group identity shape linguistic diversification? A quantitative test of the schismogenesis hypothesis*
9.20 am	Benedict King *Posterior predictive tests of model adequacy in phylogenetic analysis of linguistic data*
9.40 am	Kylie Chen *Simulating realistic copy number events for single-cell cancer data*
10.00 am	**Morning Tea**
10.40 am	Marcus Overwater *Conditioning on sample times in birth-death processes*
11.00 am	Michael Hendrikson *Counting caterpillar phylogenetic networks*
11.20 am	Bruna Caveion *Inferring phylogenies from scRNA-seq data*
11.40 am	Nikolaos Gasparatos *Simulating single-cell RNA sequencing data for cancer phylogenomics*
12.00 pm	**Lunch** (Hiku Restaurant and Bar)

12.00 pm	**Lunch** (Hiku Restaurant and Bar)
2.00 pm	Andrew J. Rodger *Reconstructing ancient evolutionary relationships with new deep-time phylogenetic models*
2.20 pm	Mark Stukel *Phylogenomic insights into the evolution, timing of diversification, and global biogeography of cicadas*
2.40 pm	Sophie Yang *Prior probability of clade splits under Serial-Sampled Coalescent*
3.00 pm	**Afternoon Tea**
3.30 pm	David Bryant *A new algorithm for computing the likelihood of a phylogeny*
3.50 pm	Peter Lockhart *Monday night dinners with David*
6.00 pm	**Dinner** (Pacific 2)

Thursday 13 February

8.45 am	**Arrival Tea and Coffee**
9.00 am	Andrew Francis *An encoding for phylogenetic trees that is stable under the addition of new leaves*
9.20 pm	Maayan Levy *Representative sets and closure*
9.40 am	Yu Liu *Reconstructing evolutionary histories with duplication events under the framework of algorithmic information theory*
10.00 am	**Morning Tea**
10.40 am	Douglas Theobald *Cross-validation of ancestral sequence reconstruction methods by predicting extant sequences*
11.00 am	Simone Linz *Fast algorithms to embed trees in phylogenetic networks without inferring new information*
11.20 am	Terry Zhang *Asymptotic enumeration of normal and hybridization networks via tree decoration*
11.40 am	Tobia Ochsner *Estimating distributions on time trees*
12.00 pm	**Lunch** (Hiku Restaurant and Bar)

12.00 pm	**Lunch** (Hiku Restaurant and Bar)
2.00 pm	Remco Bouckaert *The BEAST2 speed-dating app*
2.20 pm	Walter Xie *Bayesian phylogenetic inference of a generative angular diffusion model for protein structure evolution*
2.40 pm	Nicholas Matzke *Model comparison supports the pitcher hypothesis for the origin of the Utricularia suction traps*
3.00 pm	**Afternoon Tea**
3.30 pm	Lilin Zhang *Machine learning for genomic prediction*
3.50 pm	Yao Xiao *Combined diploid variant calling and phylogeny inference from scDNA-seq read counts*

Friday 14 February

8.45 am	**Arrival Tea and Coffee**
9.00 am	Egor Lappo *A lattice construction for enumeration of RTCN rankings*
9.20 am	Kate Truman *Inference using the Skyline Stratigraphic Ranges Fossilized Birth-Death model*
9.40 am	Charles Semple *How well does diversity indices capture phylogenetic diversity?*
10.00 am	**Morning Tea**
10.40 am	Kerry Manson *A 'properties first' approach for measuring diversity*
11.00 am	TBA *TBA*
11.20 am	Mike Steel *Most recent common ancestors and Cayley trees*
12.00 pm	**Lunch** (Hiku Restaurant and Bar)

Abstracts

Lars Berling

(University of Canterbury, berlinglars96@gmail.com)

Statistics in the space of ranked time trees

Phylogenetic trees form high dimensional non-Eulidean spaces, rendering classical statistical approaches ineffective and adding complexity to the problem. Estimating basic statistics, such as mean and variance for samples of trees is already challenging and using tree based statistics is mostly overlooked in practice. When evaluating sets of trees, it is common to use heuristics to compute a summary tree or to project the tree to a lower dimensional Euclidean vector space to apply standard statistics. However, few tools exist to analyse trees within their respective treespace and the most prominent candidate treespaces have been shown to exhibit unwanted properties. For example the 'stickiness' property of stratified spaces such as BHV or the computational complexity of tree-rearrangement based spaces such as NNI. Moreover, it has been shown that if one is interested in rooted time-trees the underlying geometry of these treespaces is fundamentally different. Here we introduce a recently developed treespace based on local tree rearrangements that allows for computational efficient distance computation, the ranked NNI space. We highlight desirable properties of this treespace that we believe are important for developing statistics within it. Additionally, we showcase applications such as estimating means for summarizing sets of trees and using tree variances in the context of MCMC convergence assessment.

Remco Bouckaert

(University of Auckland, remco@cs.auckland.ac.nz)

The BEAST2 speed-dating app

Cube space is a convenient subspace of full phylogenetic space that contains all rooted time trees that can be drawn without crossing edges for a particular order of taxa. Last year, an algorithm for fast inference of tree distributions over cube space was introduced. However, it was restricted to using a strict clock, which makes for an unrealistic assumption for many kinds of dating analyses. Applying an uncorrelated relaxed clock is hampered by the difficulty of associating rate distributions with internal branches. Here, we introduce the averaged clock model (ARC), which associates independently distributed rates on leaf branches and has branch rates for internal nodes as function of their child branch rates. We also introduce a number of variants of the ARC model including spikes and bursts. These models are convenient to use within cube space. We show that, perhaps surprisingly, ARC produces narrower age distribution estimates than an uncorrelated relaxed clock. Simulation studies show it works well with various kind of dating information, including node calibrations, clock rate priors, and fixed as well as sampled tip dates, but so far, dating with sampled ancestors remains a challenge in this framework.

David Bryant

(University of Otago, david.bryant@otago.ac.nz)

A new algorithm for computing the likelihood of a phylogeny

Felsenstein's algorithm for calculating the likelihood of a tree is one of the foundations of phylogenetics. In this talk I will describe what is perhaps the first alternative algorithm for calculating this likelihood. Our method works in a similar way to Felsenstein's algorithm, but breaks up the tree in a different way. The new approach is especially amenable to parallel computing. We can, for example, reduce the time taken to compute the likelihood from $O(n)$ to $O(\log n)$ parallel time per site. This is joint work with Celine Scornavacca and Dave Swofford.

Bruna Caveion

(University of Canterbury, bca121@uclive.ac.nz)

Inferring phylogenies from scRNA-seq data

Single-cell RNA sequencing (scRNA-seq) is a powerful tool for understanding cellular diversity and reconstructing phylogenies. By combining real and simulated datasets, we explore how scRNA-seq data can be used to model evolutionary relationships. Through a practical case study, this work highlights the potential of computational approaches to unlock new insights into cellular evolution.

Kylie Chen

(University of Auckland, kche309@aucklanduni.ac.nz)

Simulating realistic copy number events for single-cell cancer data

Many cancers have significant departures from the two copies of each gene typically found in a healthy human genome. Variations in the number of copies in a genome region are known as copy number variants. For example, in colorectal cancer, short tandem repeats — a few base pairs long — are duplicated many times due to errors in DNA repair during cell replication. Previous studies show that duplications and deletions occur across a range of cancer types, and these changes are thought to contribute to the emergence of the cancer phenotype. In our work, we aim to simulate copy duplications, deletions, and mutations guided by evidence from empirical cancer copy signature studies. These biologically realistic simulations will serve as a first step in exploring how copy number variants can be leveraged alongside single nucleotide variants in phylogenetic analysis.

Alan Cooper

(Charles Sturt University, alacooper@csu.edu.au)

Hard sweeps in feral genomes: A uniquely powerful evolutionary experiment

Feral animals represent a microcosm of evolution, as they adapt rapidly to harsh environments (e.g. house mouse in Australian deserts) using an extremely limited genetic diversity. The resulting genetic signatures (e.g. hard sweeps) can identify critical genetic pathways and interacting networks, revealing evolutionary systems that are not obvious from conventional genetic studies. Humans are a feral species outside of Africa, and our genomes record many marks of strong genetic selection (e.g. > 57 hard sweeps) as we adapted to the non-African environment, during a prolonged standstill on the Arabian Peninsula. We are extending this approach to study genomic evolution in feral species across New Zealand and Australia, and have detected functional patterns that could represent the genetic blueprint of how species adapt to new environments.

Andrew Francis

(University of New South Wales, a.francis@unsw.edu.au)

An encoding for phylogenetic trees that is stable under the addition of new leaves

Any storage and manipulation of a tree requires an encoding of the tree, whether that's as a set of edges and vertices, or in a format like Newick. When new leaves need to be added to a tree—through new sequences being obtained, such as in the rapidly changing context of an epidemic—those leaves need to be placed on the tree, and the encoding updated. There are algorithms to place a new leaf optimally ('online phylogenetics'), but current methods require recomputing the encoding of the tree, which can change in complex ways. In this talk I will describe a new encoding for trees that is stable under the addition of new leaves. That is, each leaf is given an 'address', which together forms a 'folio' that encodes the tree. New leaves that are added are assigned new addresses, but this does not change the addresses of the existing leaves: it simply adds a new row to the folio for each new leaf. Features of this encoding are that one can readily identify hierarchical structures in the tree (assisting in classification), easily extract subtrees, and compute distances between leaves, directly from the encoding. Joint work with Mark Tanaka and Ruiting Lan.

Nikolaos Gasparatos

(University of Canterbury, nikogasp@yahoo.de)

Simulating single-cell RNA sequencing data for cancer phylogenomics

As single-cell RNA sequencing (scRNA-seq) technologies advance, they offer valuable insights into the genetic and transcriptional diversity of individual cells. This is particularly relevant in cancer research, where tumor cells exhibit complex evolutionary processes and heterogeneity. By combining information from single nucleotide variants (SNVs) and gene expression profiles, it is possible to reconstruct phylogenies of cancer cells and better understand how tumors develop and respond to treatment. However, the sparsity and technical limitations of scRNA-seq data pose challenges for existing phylogenetic methods. In my talk I am going to introduce a simulation framework designed to generate synthetic scRNA-seq data that incorporates both evolutionary dynamics and sequencing constraints. This tool aims to support the development and evaluation of phylogenetic approaches tailored to scRNA-seq data.

Russell Gray

(Max Planck Institute for Evolutionary Anthropology, rd.gray@auckland.ac.nz)

Does group identity shape linguistic diversification? A quantitative test of the schismogenesis hypothesis

The diversity of human languages poses a paradox. If language is a tool for efficient communication, why are there so many languages? One often proffered answer is that language is not just about communication but also about signalling group identity. If that claim were true, we might expect to see specific patterns in language evolution. First, lineages with more splitting events should show more divergence than those with fewer. Second, this "schismogenesis effect" should be stronger in the aspects of language that speakers were more aware of. Here, we test those claims using recently developed phylogenetic methods that enable the impact of schismogenesis to be quantified and four different datasets—grammatical features, basic vocabulary, phonological change and raw sound files. We focus on linguistic divergence in the island nation of Vanuatu, which is famous for having more languages per capita than any other part of the world. Our preliminary results find strong support for both the presence of a "schismogenesis effect" and the prediction that this effect should be more pronounced in lexical and phonological divergences. This talk is joint work with Jordan Douglas, Mary Walworth, and Simon Greenhill.

Michael Hendriksen

(University of New South Wales, m.hendriksen@unsw.edu.au)

Counting caterpillar phylogenetic networks

Expanding covers provide a way to encode a large class of phylogenetic networks, called labellable networks (Francis & Steel, 2023). This class includes many familiar types of networks, including orchard, normal, tree-child and tree-sibling networks. As expanding covers are a combinatorial structure, it is possible that they can be used as a tool for counting such classes for a fixed number of leaves and reticulations, for which, in many cases, a formula has not yet been found. More recently, a new class of networks was introduced, called spinal networks, which are analogous to caterpillar trees for phylogenetic trees and can be fully described using covers (Francis, Marchei, & Steel, 2024). In this talk, we describe a method for counting the intersection of spinal networks with some familiar classes, with the hope that these form a base case from which to attack the more general classes.

Daniel Huson

(University of Tuebingen, daniel.huson@uni-tuebingen.de)

Working with phylogenetic trees and networks using PhyloSketch

Researchers studying mathematical properties of phylogenetic trees and networks desire easy access to examples. The PhyloSketch app allows users to quickly sketch trees or networks and these are captured and provided in Newick format. Constructing a phylogenetic tree or network by hand from a image can be a tedious process, and so the app provides a simple image processing feature that aims at capturing a tree or network in a semi-automated fashion. We present some of the algorithm issues and will explore some applications.

Benedict King

(Max Planck Institute for Evolutionary Anthropology, benedict_king@eva.mpg.de)

Posterior predictive tests of model adequacy in phylogenetic analysis of linguistic data

In recent years Bayesian phylogenetic analysis of language data has converged on a set of standards or best practices. In a typical analysis, the covarion substitution model is applied to a dataset of cognate-coded basic vocabulary data and combined with a Birth-Death Skyline (BDSKY) tree prior and a relaxed clock model. However, the appropriateness of these models for language data have not yet been fully investigated. I performed posterior predictive simulation to test the ability of the BDSKY model to accurately reproduce empirical trees, using the Oceanic subgroup of the Austronesian language family as a test case. The posterior predictive tree sample was generated by the

R package TreeSim taking combinations of BDSKY parameter values from the beast2 log files of the empirical analysis. The posterior and posterior predictive tree samples were then compared using a number of metrics calculated in the TreeStat2 BEAST2 package. The BDSKY performed poorly on tree balance metrics and the ratio of external to internal branch lengths. Based on the results of the posterior predictive simulations, I performed an additional analysis with a finer division of the most recent BDSKY time intervals. This led to a far better model fit as measured by Bayes factors, and a younger and more realistic root age. Surprisingly however, this second model performed worse under posterior predictive tests, performing poorly on tree length and the gamma statistic in addition to external/internal branch length ratio and tree balance. These results highlight the complexities of modelling linguistic phylogenies. Although refining the BDSKY model by including finer time divisions appeared to increase model fit, it led to poorer performance in posterior predictive tests. This suggests that increasing model complexity can reduce the predictive power of a model even when the model fit to the data is substantially better. The poor performance of the BDSKY model on tree balance metrics suggests that the widely-adopted BDSKY model fails to accurately describe the structure of language trees. These findings reveal that the current set of best practices for phylogenetic analysis of linguistic data are not always sufficient for complex datasets.

Egor Lappo

(Stanford University, elappo@stanford.edu)

A lattice construction for enumeration of RTCN rankings

A ranking of a phylogenetic network is a temporal ordering of its internal vertices, corresponding to the sequence of events in the evolutionary history of the lineages represented by the network. Enumeration of the possible rankings of a network can aid in evaluating the computational complexity of phylogenetic computations and inference algorithms. We prove an equivalence between rankings for a phylogenetic tree T and a certain path-counting problem on a certain lattice associated with the partial order described by T. For any tree-based network N with its support tree T, lattice paths for T could be equipped with "roadblocks" specified by the network structure of N. We show that the enumeration of rankings for a network N corresponds to the enumeration of non-roadblocked paths on a lattice associated with T. Finally, we extend our construction to ranked tree-child networks (RTCNs), addressing the problem of enumerating rankings for a given RTCN topology. Our construction introduces a novel algebraic structure into mathematical phylogenetics and provides a conceptual framework for analysis of networks through their displayed trees.

Maayan Levy

(University of Canterbury, maayanlevy321@gmail.com)

Representative sets and closure

Supertree algorithms often break down phylogenetic trees into units called rooted triples. A set of rooted triples might represent a specific tree or a pool of trees. A natural question that arises is the following: given a set of triples, which subsets minimally encode the same information? This talk covers a representation of these minimal sets as the spanning trees of a graph. We build on a result that the minimal sets form the bases of a matroid; however, matroid theory will only be touched on lightly for those interested.

Teng Li

(University of Auckland, teng.li@auckland.ac.nz)

Biodiversity nano-soup: Rapid biodiversity assessment using Oxford Nanopore sequencing

Rapid biodiversity assessment is pivotal for understanding ecological dynamics and conserving biodiversity. Traditional methods, reliant on morphological identification and PCR-based DNA barcoding, are time-consuming and often limited by the need for prior knowledge of the species present. Given the high cellular abundance of mitochondrial genomes in animals and their compact size of typically 15-20 kb, we developed a novel approach leveraging advances in Nanopore sequencing for direct, PCR-free, and barcoding-free sequencing of mixed animal DNA samples to generate complete mitochondrial genomes without assembly. We also introduced MashEM, a method based on the Expectation-Maximization (EM) algorithm that integrates Minimap2 mapping results and Mash distance to generate species-level taxonomic classifications and abundance profiles from Nanopore raw reads. This project aims to revolutionize biodiversity assessments by providing a faster, cost-effective, and accurate method for large-scale biodiversity monitoring.

Simone Linz

(University of Auckland, s.linz@auckland.ac.nz)

Fast algorithms to embed trees in phylogenetic networks without inferring new information

The reconstruction of a phylogenetic network that captures not only speciations but also non-treelike reticulation events frequently results in an inferred network that represents information that only has little support from the data. For instance, this is the case for phylogenetic networks that are reconstructed from a set of phylogenetic trees. Typically, the resulting network embeds additional trees that are not in the input. In this talk, we discuss polynomial-time algorithms that embed a set of phylogenetic trees in a level-1 or normal network without inferring any new trees if such a network exists. Joint work with Magnus Bordewich, Janosch Döcker, and Charles Semple.

Yu Liu

(Beijing Normal University, yu.ernest.liu@hotmail.com)

Reconstructing evolutionary histories with duplication events under the framework of algorithmic information theory

Algorithmic information theory (AIT) provides a fundamental concept for describing the complexity of an object X: by examining all programs that generate X, the length of the shortest program can define X's complexity, known as Kolmogorov complexity. This concept has been proven useful in defining a metric termed information distance, which has applications in classification tasks, the construction of phylogenetic trees, and more. Recently, the significance of Kolmogorov complexity has been revisited, particularly with the rise of large language models. However, Kolmogorov complexity is uncomputable in principle. In practical use, it can be approximated through methods such as traditional compression algorithms (e.g., gzip). In recent years, we proposed an approach called Ladderpath, which iteratively discovers repeated substructures within an object to reconstruct a hierarchical relationship of reuse and nesting among these substructures. Conceptually, this can be understood as a compression algorithm rooted in AIT. On the one hand, Ladderpath can be used to define information distance (and subsequently construct phylogenetic trees). On the other hand, perhaps more importantly, it reveals hierarchical nesting relationships among repetitive substructures. We have applied this hierarchical relationship to various domains, including analyzing protein sequences of different species to study evolutionary relationships (Physical Review Research, 2024), exploring the structure-function relationships in neural networks (npj Complexity, 2024), and designing new peptide drugs (Journal of Chemical Information and Modeling, 2024). Ladderpath is a general-purpose method based on AIT. In this talk, we will focus on our recent (unpublished) studies related to phylogenetics, including how reuse relationships among substructures identified by Ladderpath can be leveraged to define information distances. Additionally, we will explore how this method can be applied to reconstruct evolutionary histories involving sequence duplication events or horizontal transfer events (which are typically not well-suited for description using binary phylogenetic trees). Through this, we aim to highlight its potential applications.

Peter Lockhart

(Massey University, p.j.lockhart@massey.ac.nz)

Monday night dinners with David

Kerry Manson

(University of Canterbury, kerrymanson320@gmail.com)

A 'properties first' approach for measuring diversity

The evolutionary isolation of individual species can be quantified using certain measures called phylogenetic diversity indices. These indices share out the phylogenetic diversity (PD) value of a tree among the species present at the leaves, depending on their position relative to others. There are many possible ways this sharing out can be achieved. Current approaches begin by trying to imagine a fair method of allocating PD and hope that the result has nice mathematical properties. In this talk we advocate taking the opposite approach. That is, determining the properties our desired methods should possess before systematically searching across all possible diversity indices for an optimal one. We discuss examples of how this approach can be used in relation to a key property of diversity indices: the difference between summed index scores and PD for subsets of leaves. The diversity index found by our properties first approach is better than current approaches at being able to balance this noted tradeoff between local and global perspectives of diversity. Joint work with Martin Frohn.

Nicholas Matzke

(University of Auckland, n.matzke@auckland.ac.nz)

Model comparison supports the pitcher hypothesis for the origin of the Utricularia suction traps

The origin of carnivorous plant traps has been a research subject ever since Darwin's 1875 book on the subject. However, the origin of the more complex trap, the tiny suction traps of genus Utricularia (bladderworts) mystified Darwin and subsequent researchers. In this research, we build a phylogenetic model for the evolution of carnivorous plant traps to test the "pitcher hypothesis" for the origin of the Utricularia trap, proposing a gradual evolutionary transition from simple adhesive traps to pitcher traps, and ultimately to Utricularia unique suction traps. We assembled phylogenetic trees for carnivorous plant species to test our hypothesis with statistical model comparison: the fit of a model where Utricularia bladder traps are essentially miniaturised pitcher traps is compared to less constrained null models where any trap type can evolve into any other. The results suggested that among the 18 phylogenetic models evaluated, the res7abprCTE model, aligning with the pitcher hypothesis, emerged as the best-fitting model, with an AIC weight of 60%, and two other similar pitcher-hypothesis models garner the remaining 40%. We propose that by statistically comparing models representing detailed, stepwise, pathways for the evolution of complex adaptations, we should be able to convert exercises in “adaptive storytelling”, where verbal scenarios are subjectively judged on plausibility, into the modern phylogenetic framework of statistical model comparison.

Tobia Ochsner

(University of Auckland and ETH Zurich, tobia.ochsner@inf.ethz.ch)

Estimating distributions on time trees

In Bayesian phylogenetics, Markov Chain Monte Carlo (MCMC) methods generate samples of time trees to capture the posterior distribution of the phylogenetic tree. We explore novel approaches to fit distributions on these time trees samples. In particular, we extend the family of Conditional Clade Distributions (CCD) to model the tree topology and branch lengths simultaneously. Our distributions are designed to capture both central tendencies and dependencies within this high-dimensional and non-Euclidean space. They facilitate a straightforward reconstruction of a summary tree and manage to capture the tails of the time tree distribution. This property extends their utility beyond the robust reconstruction of a single summary tree, enabling more sophisticated downstream analyses like exploring credible sets or evaluating alternative evolutionary hypotheses.

Marcus Overwater

(ETH Zurich, marcus.overwater@bsse.ethz.ch)

Conditioning on sample times in birth-death processes

Birth-death models are widely used in phylodynamics to infer population parameters from time-scaled phylogenetic trees. In these models tip dates are generated by a sampling process where each individual in the population is sampled at a constant rate. These sampling assumptions are often violated by real data, and have been shown to bias inferences when the sampling process is misspecified. One potential solution is to use the sequentially sampled coalescent where the sampling times are fixed parameters of the model. The other solution, which I present here, is to condition on the sampling times of the birth-death process. I explore the mathematical properties of this model and describe the probability density of a phylogenetic tree. This density can easily be applied as a tree prior in Bayesian inference. I present the results of a well-calibrated simulation study, which validates an implementation of the model in BEAST2. Finally, I show that in a scaling limit this model is equivalent to the sequentially sampled coalescent with a stochastic effective population size given by a diffusion process, providing a connection between birth-death and coalescent phylodynamic models.

Andrew J. Roger

(Dalhousie University, andrew.roger@dal.ca)

Reconstructing ancient evolutionary relationships with new deep-time phylogenetic models

The origin of eukaryotic cells from prokaryotic ancestors remains one of the most enigmatic major transitions in the evolution of life on Earth. It is widely accepted that the nucleocytoplasmic lineage of eukaryotes is related to asgard Archaea and mitochondria evolved from endosymbionts related to Alphaproteobacteria. However, the precise phylogenetic positions of these two ancestral lineages, the nature of additional genetic contributors, the position of the root of the eukaryotic tree and the relative timing of events that occurred in eukaryogenesis remain unclear. This lack of clarity stems, in part, from artefacts induced by the inadequacy of standard phylogenetic models of amino acid sequence evolution (e.g. LG+Gamma) to capture the dynamics of sequence change on the billion-year timescale. Here we introduce several new models for deep-time phylogenetic estimation that account for: i) heterogeneity in the amino acid substitution process over sites, ii) functional shifts in molecules over splits, and iii) simultaneous across-site and branch-heterogeneity in amino acid frequencies. We show how these methods were applied to robustly determine the root of the eukaryote tree and other deep-time phylogenetic problems. This is joint work with Kelsey Williamson, Laura Eme, Hector Baños, Charley McCarthy, Edward Susko, Ryoma Kamikawa, Sergio Muñoz-Gómez, Bui Quang Minh, and Alastair Simpson

Charles Semple

(University of Canterbury, charles.semple@canterbury.ac.nz)

How well does diversity indices capture phylogenetic diversity?

Phylogenetic diversity (PD) is a popular measure for quantifying the biodiversity of a collection Y of taxa, while (phylogenetic) diversity indices provide a way to apportion PD to individual taxa. In conservation prioritisation, diversity indices have been proposed and used as an alternative to PD. However, for some specific diversity index, the PD of Y is typically not equal to the sum of the diversity indices apportioned to the taxa in Y. In this talk, we investigate the extent of this difference for the commonly-used indices Fair Proportion and Equal Splits. This is joint work with Magnus Bordewich.

Chris Simon

(University of Connecticut, chris.simon@uconn.edu)

Global genomics of the man-o'-war (Physalia) reveals biodiversity across the ocean surface

The open ocean is a vast, highly connected environment, and the organisms found there have been hypothesized to represent massive, well-mixed populations. Of these, the Portuguese man-o'-war (Physalia) is uniquely suited to dispersal, sailing the ocean surface with a muscular crest. We tested the hypothesis of a single, panmictic Physalia population by sequencing 151 genomes, and found five distinct lineages, with multiple lines of evidence showing strong reproductive isolation despite range overlap. We then scored thousands of citizen-science photos and identified four recognizable morphologies corresponding to these lineages. Three lineages occur in New Zealand. Of these, one lineage was previously unrecognized and represents a new species restricted to New Zealand and Tasmania. Within lineages, we detected individual long-distance dispersal events and regionally endemic subpopulations, and use physical modeling to show that these are connected by winds and currents. We find that, even in these sailing species, genetic variation is highly partitioned across the open ocean. This is joint work with Samuel H. Church and Casey W. Dunn along with R.B. Abedon, N. Ahuja, C.J. Anthony, D. Destanović, D.A. Ramirez, L.M. Rojas, A.E. Albinsson, I. Álvarez Trasobares, R.E. Bergemann, O. Bogdanovic, D.R. Burdick, T.J. Cunha, A. Damian-Serrano, G. D'Elía, K.B. Dion, T.K. Doyle, J.M. Conçalves, A.G. Rajal, S.H.D. Haddock, R.L. Helm, D. Le Gouvello, Z.R. Lewis, B.I.M.M. Magalhães, M.J. Mańko, C.G. Mayorga-Adame, A. de Mendoza, C.C. Moura, C. Munro, R. Nel, K. Oguchi, J.N. Perelman, L. Prieto, K.A. Pitt, M. Roughan, A. Schaeffer, A.L. Schmidt, J. Sellanes, N.G. Wilson, G. Yamamoto, E.A. Lazo-Wasem, M.B. Decker, and J.M. Coughlan.

Mike Steel

(University of Canterbury, mathmomike@gmail.com)

Most recent common ancestors and Cayley trees

In Part 1, we consider the question of how many leaves need to be randomly sampled from a tree so that the root of the induced subtree is close to (or exactly equal to) the root of the entire tree. We describe how an early result of Michael J Sanderson can be reformulated to reveal more precisely the impact of tree shape on this question. In Part 2, we explore the distribution of a Robinson-Foulds type of distance on Cayley trees (rather than on the usual phylogenetic trees). This question was motivated by a 2024 paper (Khayatian, Valiente and Zhang) that introduced this metric on trees arising in modelling tumour cell evolution. Applying asymptotic enumeration techniques and some probabilistic tools, precise results and exact expressions for various questions arising in Parts 1 and 2 can be derived.

Mark Stukel

(University of California, Davis, mark.stukel@uconn.edu)

Phylogenomic insights into the evolution, timing of diversification, and global biogeography of cicadas

The Cretaceous-Paleogene (K-Pg) mass extinction resulted in a massive turnover in earth's biodiversity. Numerous lineages diversified after the K-Pg boundary, including birds, placental mammals, frogs, and snakes. With over 3,400 described species, the globally-distributed charismatic insect family Cicadidae (singing cicadas) appears to have also diversified after the K-Pg boundary, since the oldest fossils unambiguously assigned to the family date to the Paleocene. We assembled a global phylogenomic dataset of 490 nuclear AHE loci as well as mitochondrial genomes for Cicadidae, sampling all five subfamilies, 85% of tribes, and 25% of genera. We resolved the phylogenetic relationships among Cicadidae subfamilies using concatenated maximum-likelihood and multi-species coalescent approaches. We combined Bayesian multispecies coalescent and fossilized birth-death models to estimate divergence times of Cicadidae lineages using 44 fossil taxa. We also conducted ancestral area reconstruction for Cicadidae, representing the first biogeographic analysis for the entire family. We find that singing cicadas originated in the Cretaceous, with four of the five subfamilies diversifying shortly after the K-Pg extinction event. We also find a Gondwanan origin for Cicadidae, with many lineages dispersing across the globe. Our results add to the body of literature about the effects of the K-Pg mass extinction in shaping present-day diversity. This is joint work with Chris Simon and Nick Matzke.

Douglas Theobald

(Brandeis University, dtheobald@brandeis.edu)

Cross-validation of ancestral sequence reconstruction methods by predicting extant sequences

Ancestral sequence reconstruction (ASR) is a phylogenetic method used to study ancient biomolecules and molecular evolution. However, its accuracy remains unknown, as true ancestral proteins cannot be directly compared. To address this, we introduce "extant sequence reconstruction" (ESR), a cross-validation method that reconstructs present-day sequences using ASR techniques, allowing comparison with known true sequences. Our findings show that the commonly used average probability metric is a reliable indicator of reconstruction accuracy when the evolutionary model is accurate or over parameterized. However, it is unreliable for comparing different models, as more accurate models can yield reconstructions with lower probabilities. While better models may result in lower sequence identity, they produce reconstructions that are biophysically closer to true ancestors. In addition, ESR analysis shows that the choice of sequence alignment methodology has a much greater impact on reconstruction accuracy than evolutionary model choice, and that on average the highest log-likelihood alignment results in fewer total sequence reconstruction mistakes. These findings highlight the importance of model selection in ASR and the value of sampling reconstructions to analyze

ancestral protein properties. ESR provides a practical approach for validating evolutionary models and can be applied to any phylogenetic analysis of biological sequences.

Kate Truman

(University of Canterbury, kate.truman@pg.canterbury.ac.nz)

Inference using the Skyline Stratigraphic Ranges Fossilized Birth-Death model

The Fossilized Birth-Death (FBD) models are a popular class of models which allow fossil data to be included without assuming that lineages go extinct at fossilization. The Stratigraphic Ranges Fossilized Birth-Death (SRFBD) model, introduced by Stadler et al. (2018), includes additional information by explicitly including stratigraphic ranges, that is, multiple fossils of the same species with different ages. The model can also be applied to epidemiological contexts, where multiple samples may be taken over time from the same infected patient. The SRFBD model, which has recently been implemented in BEAST, would be expected to produce less biased results than a general FBD model with the same diversification rates, but which does not allow multiple samples of the same species or patient. However, the SRFBD model only allows for constant diversification rates (birth, extinction and sampling), which is a major simplification of real world processes. We thus focus on an extension, the skyline SRFBD model, in which diversification rates take piecewise-constant forms. We discuss considerations for its implementation in BEAST, so that it can be used to infer evolutionary histories.

Walter Xie

(University of Auckland, walter@cs.auckland.ac.nz)

Bayesian phylogenetic inference of a generative angular diffusion model for protein structure evolution

Protein structure evolution plays a critical role in understanding molecular function and adaptation, but traditional modelling approaches often face limitations in capturing structural complexity and correlations. We implemented an innovative framework (Garcıa-Portugue's et al. 2017, Golden et al. 2017) into LPhy and BEAST2, that utilises dihedral angle sequences (ϕ and ψ) of C-alpha atoms to represent protein structures. This approach eliminates the need for structural alignment and also reduces dimensionality. The model employs the Wrapped Normal (WN) diffusion, modelling the evolutionary trajectory of a pair of dihedral angles as a bivariate diffusion process on the torus. Compared to Cartesian coordinate-based models, this model requires fewer parameters while preserving structural dependencies, leading to more accurate and computationally efficient phylogenetic inference. In this talk, we will demonstrate a simple simulation for this model and share some preliminary results from Bayesian phylogenetic analyses. This is joint work with Clementine Yan and Alexei Drummond.

Yao Xiao

(University of Auckland, yxia415@aucklanduni.ac.nz)

Combined diploid variant calling and phylogeny inference from scDNA-seq read counts

With the development of single-cell sequencing technology, phylogenetics can be applied to cell and developmental biology. For example, during development, cellular variations can be viewed as markers of evolution. Consequently, evolutionary models offer new perspectives for understanding somatic development and cancer evolution. However, most existing methods handle data filtering, variant calling, and phylogenetic inference separately. This workflow does not fully leverage the information in the data, potentially leading to biased results. To address these issues, we propose a Bayesian inference–based model that simultaneously performs variant calling and phylogenetic tree inference directly from single-cell DNA sequencing read counts. Our model has the following features: it considers all 16 possible diploid genotypes and accounts for both sequencing errors and allele dropout that may occur during sequencing. Our research aims to provide a powerful new tool for studying cancer evolution and developmental biology.

Sophie Yang

(University of Auckland, sophiezijing@gmail.com)

Prior probability of clade splits under Serial-Sampled Coalescent

Conditional Clade Distributions (CCDs) provide an effective way to represent posterior tree distributions in Bayesian phylogenetic analysis. The prior distribution of clade splits is an important element of CCDs, which is useful for combining multiple datasets in large-scale phylogenetic studies. While the prior probability of clade splits can be calculated easily for contemporaneously sampled data, the case of serial-sampled data is more complex. Here we introduce a dynamic programming algorithm to calculate the prior probability of clade splits under the coalescent process with serial-sampled data. To overcome computational challenges, we develop a heuristic method that improves running time while maintaining high accuracy. Additionally, we propose a simulation-based approach to integrate over various population sizes. Our results highlight the influence of serial sampling intervals on clade split probabilities, enhancing our understanding of prior clade distributions. This work provides a foundation for further optimization and scaling in complex phylogenetic analyses.

Lilin Zhang

(University of Canterbury, lzh207@uclive.ac.nz)

Machine learning for genomic prediction

With the advancement of sequencing technologies, millions of genetic markers (e.g. SNPs) have been discovered in the genome of many species. With the availability of cost-effective SNP chips, genomic predictions that use many genotyped SNPs across the genome have been widely applied to animals and plants to improve genetic progress and to humans to predict polygenic risks of diseases. This talk focuses on sharing my experiences from experimenting with several machine learning models on genomic prediction datasets, each containing sequence lengths of millions of SNPs.

Terry Zhang

(University of Canterbury, terry.zhang@pg.canterbury.ac.nz)

Asymptotic enumeration of normal and hybridization networks via tree decoration

Adding randomly k arcs between edges of a rooted binary phylogenetic tree with n leaves may or may not result in a normal network. In this talk, we show that the probability of obtaining a normal network through such approach tends to 1 as n grows if k is fixed or even grows with n at the rate of $o(n^{\frac{1}{3}})$. We also investigate the exact enumeration of hybridization network with 2 reticulation vertices and conclude that the same asymptotic results apply to hybridization network. This talk is based on joint work with Mike Steel and Michael Fuchs.