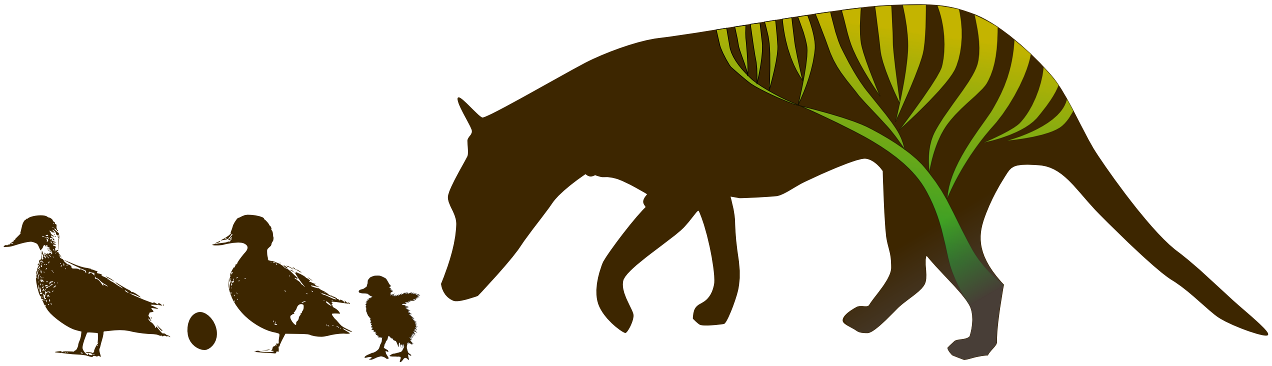
Phylomania 2022
Ducks in a Row

Hobart
November 23-25
2022



with generous support from

ARC CENTRE OF EXCELLENCE FOR
PLANT SUCCESS
IN NATURE AND AGRICULTURE

Important Information

Venue

For those of you who are attending in person, the conference room is **Harvard Lecture Theatre 2**, which is part of the **Centenary Building** at the low end of Sandy Bay Campus.

Links

The video links for recorded talks will be on the Phylomania web page at
http://www.maths.utas.edu.au/phylomania/2022/index.html.

Zoom details (for all three days):

- **Meeting URL:** https://utas.zoom.us/j/88462896982
- **Meeting ID:** 884 6289 6982

Phylomania 2022 Conference Dinner

Location: **Room For A Pony**, 338 Elizabeth St North Hobart.

Time: Thurs 24/11, 6pm (approx). Fine to arrive whenever is convenient. Menu: http://www.roomforapony.com.au Ordering: Please order and pay for food/drinks at the bar. See the menu above for dietary options.

Transport: The Metro 501 bus travels directly from the Sandy Bay Campus to North Hobart (up Elizabeth st). Otherwise, catching any bus to the Hobart CBD and then another up Elizabeth St will work (google maps is your friend!) You can pay the driver for a bus ticket using cash ($3.50). It takes approx 25mins to walk from the campus to the CBD and another 20mins CBD to Room For A Pony. A uber/taxi fare Sandy Bay to North Hobart will cost $20-$30.

Pre/After-party: I highly recommend a visit to Boodle Beasley (diagonally across the road) if you can fit it in!

Covid Safety

We are doing out best to protect everyone during this ongoing Covid-19 pandemic: here are some of the measures we have in place.

Please do your best to look after yourself and others!

- Mask wearing is encouraged!
- Please space out as much as possible; you may even be able to to have an empty chair between you and others during the sessions.
- We will provide masks and a good spread of hand sanitizer at entry to, or in the lecture theatre.
- The caterers will be asked to place the individual food portions at a few locations in the common space, to avoid bottlenecking. Do use the hand sanitizers, which will be spread out across the common space.
- If weather permits, we encourage you to move into the open-air centenary courtyard for breaks. If not, we encourage you to spread out across the large common indoor space outside the theatre, rather than cluster.
- Rapid Antigen Test (RAT) kits will be available if you feel any desire to test yourself.
- If you feel ill, please stay away!

Keynote Speaker 2022

Senior Lecturer/Consultant in Statistics
School of Biological Sciences
Faculty of Science
The University of Queensland

Simone Blomberg started out as a lizard ecologist (PhD with R. Shine (USyd)). After a successful postdoc on phylogenetic comparative methods in the USA with T. Garland and A. Ives (U. Wisconsin, Madison and U. California, Riverside), Simone returned to Australia without a job. After trying (and failing) to obtain a position as a lizard ecologist, Simone re-trained, gaining a Master's degree in Applied Statistics from the ANU. In 2007 Simone was appointed to a lectureship in the School of Biological Sciences at the University of Queensland, with a special focus on providing a statistical consulting service to academic staff and postgraduate students. She has been there ever since, researching the interface between statistics and evolutionary theory, supervising postgraduate students, and teaching courses in evolutionary theory, systematics, and statistics. She works to further increase the quality of science produced by the UQ School of Biological Sciences.

The Conference Organizers gratefully thank the
ARC Center of Excellence for Plant Success in Nature and Agriculture
for financial support, allowing us to bring Dr Blomberg to Phylomania 2022.

Session Guide

When	Who	What
		WEDNESDAY
8:00-8:40am	*Everyone*	Registration: Just arrive and write your name on a sticker!
8:40-9:00	Committee	Welcome and Housekeeping
	Simon Ellingsen	Opening Remarks and Conference Official Opening
	Duncan Robinson	Acknowledgement of Country
Mathematical delights		
9:00-9:40	Simone Blomberg	**KEYNOTE**: Stochastic diffusion models for quantitative characters
9:40-10:00	Mike Steel	Modelling feature diversity and phylogenetic loss due to rapid extinction at the present
10:00-10:20	Michael Baake	Aspects of the Markov embedding problem
10:20-10:40	Conrad Burden	Feller Diffusions, Coalescence and Sampling Distributions
10:40-11:10	*Morning Tea*	
Computational Methods		
11:10-11:30	Gavin Huttley	Meta-science, meta-genomics and meta-programming with cogent3
11:30-11:50	Patrick Gemmell	A phylogenetic method linking nucleotide substitution rates to continuous phenotypic change
11:50-12:10pm	Ben Silke	An alignment-free approach to identifying highly divergent sequences
12:10-12:30	Joshua Hoyle	Evolution of SARS CoV-2 Coronavirus Variants of Concern with Mass-based Phylogenetics
12:30-1:30	*Lunch*	
Networks		
1:30-1:50	Samuel Barton	Hypergraph Models with Application
1:50-2:10	Jonathan Mitchell	Detecting recurrent evolution in Sorghum with convergence-divergence models
2:10-2:30	Luke Cooper	Can gene duplication explain the observed topology of gene regulatory networks?
2:30-2:50	Diane Donovan	Topological Measures on Hypergraphs: with Application to Gene Expression Data
2:50-3:20	*Afternoon Tea*	
Evolutionary Models		
3:20-3:40	Thomas Wong	MAST: Phylogenetic Inference with Mixtures Across Sites and Trees
3:40-4:00	Huaiyan Ren	The estimation and application of mixture models in phylogenetics
4:00-4:20	Qian Feng	An improved JHMM for detecting recombination from *Plasmodium falciparum* malaria parasite *var* genes
4:20-4:40	Xia Hua	Protracted speciation and extinction process

		THURSDAY
Europe		
8:00-8:20	Mareike Fischer	Edge-based phylogenetic networks and their proximity measures
8:20-8:40	Ana Serra Silva	An Update on Clumps, Or: Yes, We Can Recover Informative Clusters of Trees with Non-Identical Leaf Sets
8:40-9:00	Luke Kelly	Lagged couplings diagnose MCMC phylogenetic inference
9:00-9:20	Liam J. Maher	Network-based polyploid phylogenetics
9:20-9:40	George Aliatimis	Tropical logistic regression on phylogenetic trees
9:40-10:10	Morning Tea	
North America		
10:10-10:30	Hector Baños	The Tree of Blobs of a Species Network: Identifiability under the Coalescent
10:30-10:50	Jonathan N Odumegwu	Heterogeneity of Gene Tree Topologies
10:50-11:10	Marina Garrote-López	Using semi-algebraic constraints to design new weights for quartet-based methods
11:10-11:30	David Barnhill	Hit and Run Sampling from the Space of Phylogenetic Trees
11:30-1:00pm	*Lunch*	**and Poster Session!**
Tree Space		
1:00-1:20	Lena Collienne	Subtree Prune and Regraft on Ranked Trees
1:20-1:40	Tom Nye	Phylogenetic information geometry and its implications
1:40-2:00	Stephan Huckemann	On the wald space for phylogenetic trees
2:00-2:30(ish)	Matthew Macaulay	Hyperbolic Tree Embeddings: a Continuous Representation of Discrete Trees
2:30-3:10	*Afternoon Tea*	
Computational Methods		
3:10-3:30	Joshua Stevenson	Subflattenings: What are they good for?
3:30-3:50	Frederick Jaya	Finding tree topologies from an alignment using site-rate variation
3:50-4:10	Robert McArthur	Is time-oriented phylogenetic reconstruction of thousands of sequences possible?
4:10-4:30	Andrew Francis	Networks and Covers
	Conference Dinner	
6:00pm	*Anyone*	Room for a Pony, North Hobart

When	Who	What
FRIDAY		
Matrix Analytic Methods		
9:00am-9:20	Małgorzata O’Reilly	Matrix-analytic methods for the evolution of species trees, gene trees, and their reconciliation (Part I)
9:20-9:40	Albert C. Soewongsono	Matrix-analytic methods for the evolution of species trees, gene trees, and their reconciliation (Part II)
9:40-10:00	Jiahao Diao	Matrix-analytic methods for the evolution of species trees, gene trees, and their reconciliation (Part III)
10:00-10:20	Mingqi He	Approximate Bayesian computation for Markovian binary trees in phylogenetics
10:20-10:50	*Morning Tea*	
Statistical Methods		
10:50-11:10	Katherine Caley	Is the Human genome in mutation equilibrium?
11:10-11:30	Lars Berling	Statistics in the space of ranked time trees
11:30-11:50	Luke Yates	Model validation and selection in phylogenetic comparative analyses using posterior predictive methods
11:50-12:10	Qin Liu	Unpacking phylogenetic inference: residual diagnostics and goodness-of-fit tests
12:10-12:30	Yao-ban Chan	The large-sample asymptotic behaviour of quartet-based summary methods for species tree inference
12:30-1:30pm	*Lunch*	
Applications / Tenuously Connected To Phylogenetics		
1:30-1:50	Hanh Minh Vo	Which clade(s) of eudicot *NCED* genes are triggered by dehydration?
1:50-2:10	Bennet McComish	Using genomic signatures of natural selection to elucidate MS genetics
2:10-2:30	Greg Jordan	Evolutionary implications of genome size and cell size
2:30-2:50	Keaghan J Yaxley	Global variation in the relationship between avian phylogenetic diversity and functional dispersion is driven by environmental constraints
2:50-3:10	Nick Fountain-Jones	TBC
3:10-3:30	Raima Carol Appaw	Generating networks for epidemiological dynamics of infectious disease
3:30-4:00	*Afternoon Tea*	
Awards and Closing Remarks		

Abstracts

Stochastic diffusion models for quantitative characters

Simone Blomberg, The University of Queensland
s.blomberg1@uq.edu.au

Mathematical Delights

Stochastic processes are commonly used to model various aspects of the evolution of organisms, both at the level of the population (e.g., gene frequencies), and at the level of the species (e.g., traits). This is because stochastic processes allow us to model both the deterministic aspects of evolution, such as selection, as well as the stochastic aspects of evolution, such as genetic drift and other sources of "randomness" in the same modelling framework. Indeed, the development of modern phylogenetic comparative methods has focused attention on the need for explicit models of evolution in order to test hypotheses about the phenotypic correlations among traits, and correlations between traits and environments. Further, explicit models of evolution are necessary in order to understand the dynamics of trait change and disparity through "deep" time.

I will discuss the mathematical structure of some commonly used stochastic process models of evolution (diffusions). I will discuss similarities and differences among models and introduce two new evolutionary models that may be of use particularly when traits are non-Gaussian. Pitfalls and issues associated with testing hypotheses about model parameters will be addressed. I will also talk about my recent work in applying diffusions to multivariate evolutionary problems, such as modelling the G-matrix over ”deep” time.

Modelling feature diversity and phylogenetic loss due to rapid extinction at the present

Mike Steel, University of Canterbury
mathmomike@gmail.com

Mathematical Delights

(Joint work with Marcus Overwater)

Previous work has mathematically modelled the relative loss of phylogenetic diversity (PD) when species at the present become extinct. PD is often taken in conservation biology as a proxy for feature diversity (FD); but modelling FD using an evolutionary process on a tree (either fixed or random) leads to somewhat different behaviour than for PD. This talk presents some recent mathematical results for FD loss in a phylogenetic context.

Aspects of the Markov embedding problem

Michael Baake, Faculty of Mathematics, Bielefeld University
mbaake@math.uni-bielefeld.de

Mathematical Delights

The problem whether a given Markov matrix can occur in a continuous-time Markov chain has many facets, some of them astonishingly subtle and difficult. Here, based on joint work with Ellen Baake and Jeremy Sumner, some recent progress will be discussed, with focus on recombination models in genetics and its consequences for some popular discrete time models.

Feller Diffusions, Coalescence and Sampling Distributions

Conrad Burden, Mathematical Sciences Institute, Australian National University
`Conrad.burden@anu.edu.au`

Mathematical Delights

(Joint work with Prof. Robert Griffiths, School of Mathematics, Monash University)

In 1951 William Feller published the solution to the diffusion equation

$$u_t(t,x) = \frac{1}{2}\{xu(t,x)\}_{xx} - \alpha\{xu(t,x)\}_x$$

for an initial condition $u(0,x) = \delta(x - x_0)$. The equation models the growth of a population of independently reproducing individuals, and manifests as the infinite population, small time step limit of a Bienaymé-Galton-Watson branching process, or as the limit of a continuous-time birth-death process in which the birth and death rates become infinite while their difference remains equal to α.
We have calculated the distribution of the random variable $A_n(s,t)$, defined as the number of ancestors at a time s in the past of a sample of size n taken from the infinite population of a Feller diffusion at a time t since since its initiation. We illustrate two applications: The first is the stationary sampling distribution of a subcritical diffusion of a population of multiple types undergoing neutral mutations between types. The second is the construction of a coalescent tree for a supercritical diffusion assuming a uniform prior on the time since initiation, and calculation of sampling distributions for a multi-type population undergoing neutral mutations from this tree.

Meta-science, meta-genomics and meta-programming with cogent3

Gavin Huttley, Australian National University
`Gavin.Huttley@anu.edu.au`

Computational Methods I

Open source software is a driver for scientific innovation. It serves as the layer connecting the vision and needs of experimentalists with the quantitative tools produced by methods developers. Software libraries are toolboxes that significantly reduce the effort to develop custom applications. cogent3 is one such Python library. It originated from research aimed at understanding mutagenic processes using maximum-likelihood-based phylogenetic methods. It has capabilities well beyond this scope that have led to its incorporation in applications tackling various problems, such as meta-genomics. I will describe key features of cogent3, including data sampling, sequence alignment, and a "grammar" for succinctly specifying context-dependent substitution models. I will also outline plans for its ongoing evolution, including integration with other widely used tools.

A phylogenetic method linking nucleotide substitution rates to continuous phenotypic change

Patrick Gemmell, Statistics & Biology, Harvard University
`pgemmell@fas.harvard.edu`

Computational Methods I

(Joint work with Timothy B. Sackton, Scott V. Edwards, and Jun S. Liu)

Animal genomes contain highly conserved non-coding sequences that are important to biological function e.g., gene regulation. But what elements are related to what traits? This talk presents a phylogenetic method that answers this question by associating nucleotide substitution rates with changes in a continuous trait of interest. The method takes as input a multiple sequence alignment of conserved elements, continuous trait data observed in extant species, and a background phylogeny and substitution process. Gibbs sampling is used to assign rate categories (background, conserved, accelerated) to lineages and explore whether the assigned rate categories are associated with increases or decreases in the rate of trait evolution. The approach will be illustrated using publicly available data e.g., mammalian elements and lifespan data. The software has been implemented using R and C++ and is sufficiently fast that it can be run on hundreds of thousands of elements from many species. We hope that phylogenetic methods linking genotype to phenotype will ultimately increase our understanding of natural history. We also hope they will allow us to use data from diverse species to shine a spotlight on parts of our own genome that are of biomedical interest.

An alignment-free approach to identifying highly divergent sequences (Student presentation)

Ben Silke, Australian National University, Research School of Biology
`u6675274@anu.edu.au`

Computational Methods I

(Joint work with Ben Silke, Yu Lin, and Gavin Huttley)

Clustering of sequences based on similarity is a common problem in bioinformatics. Phylogeny-based clustering methods which utilise the phylogenetic tree demonstrate promise but require the input of phylogenetic trees or sequence alignments. The complexity of constructing these inputs is often very high such that the methods may be impractical for large-scale sequence data. We examined the properties of an alignment-free estimation of the most divergent members of a sequence collection. Specifically, we identify the subset of sequences that maximise average Jensen-Shannon divergence (JSD) computed from k-mer frequencies. We implement the approach in 'divergent', a python command line application. We evaluated the performance of 'divergent' using both artificial and natural data. We show that lineages identified using 'divergent' exhibit close to the maximum genetic distance. 'divergent' has respectable compute performance – identifying eight lineages from a collection of 1015 whole microbial genomes in $\sim$2 minutes on an old laptop.

Evolution of SARS CoV-2 Coronavirus Variants of Concern with Mass-based Phylogenetics (Student presentation)

Joshua Hoyle, Infectious Disease Responses Laboratory, Prince of Wales Clinical Research Sciences,
`joshua.hoyle@scientia.org.au`

Computational Methods I

(Joint work with Christian Mann, Kevin M. Downard)

The global spread of SARS-CoV2 coronavirus and the emergence of a growing number of highly infectious variants underscores the need for both rapid and sensitive detection and a means with which to explore the virus' evolutionary trajectory. For over a decade, this laboratory has championed the use of high resolution mass spectrometry as a means to both detect and characterise respiratory viruses into different types, subtypes and lineages. Using protein mass maps generated for individual components, or the whole virus, we have shown that the detection of specific signature peptides can be used for this purpose without the need for sequencing of the strains in question. Subsequently, we have used the same maps to assemble phylogenetic-like mass trees using numerical datasets that avoid the need for gene or protein sequences or any sequence alignment. Here we employ the combined strategy to identify and rapidly distinguish SARS-CoV2 coronavirus strains across five major variants of concern. Deletions or mutations within the surface spike protein across these variants, that originated in the UK, South Africa, Brazil and India (known as the alpha, beta, gamma and delta variants respectively), lead to associated mass differences in the mass maps that can be used to identify and distinguish the variants. The same mass map profiles were utilised to construct reliable evolutionary histories even where complete protein coverage is not achieved, while greater coverage enables the direct identification and display of variant-associated mutations on the mass tree.

Hypergraph Models with Application (Student presentation)

Samuel Barton, The University of Queensland, School of Mathematics and Physics
`s.barton@uqconnect.edu.au`

Networks

(Joint work with Diane Donovan and James Lefevre, University of Queensland)

Networks, also known as graphs, are a common and useful tool used to model interactions among objects. However, graphs only capture pairwise interactions and relationships. This property often limits modelling and results as many real-world systems encompass multi-way interactions. As such, modelling systems with hypergraphs, being the generalisation of a graph, provides a more informative framework. In this talk, we will introduce hypergraphs and discuss some properties and measures which can be used to study the hypergraph structure. With this, we will then model an example data set and discuss the results.

Detecting recurrent evolution in Sorghum with convergence-divergence models

Jonathan Mitchell, University of Tasmania
Jonathan.Mitchell@utas.edu.au Networks

(Joint work with Barbara Holland (University of Tasmania), Yongfu Tao, David Jordan, Emma Mace (The University of Queensland), Jeremy Sumner (University of Tasmania))

Convergence-divergence models (CDMs) are alternatives to graph-theoretic phylogenetic networks that allow some populations to continuously become more similar over time. A Markov model describes convergence of some populations, which become more similar (e.g., in their nucleotide sequences), and divergence of others, which become less similar. CDMs can model many biological processes, including recurrent evolution and introgressive hybridization, alongside divergent evolution of other populations.
Recurrent evolution is the process where distinct populations independently evolve similarly, often due to similar selection pressures. It is suspected to be influencing patterns of presence/absence of gene families among Sorghum populations. Studying the relationships between the presence/absence patterns and phenotype to understand recurrent evolution could guide better crop development. We infer a CDM on the Sorghum populations. Future work is to detect which gene families could have undergone recurrent evolution.

Can gene duplication explain the observed topology of gene regulatory networks? (Student presentation)

Luke Cooper, University of Tasmania
luke.cooper@utas.edu.au Networks

Gene regulatory networks (GRNs) describe the regulatory relationships between genes in a particular tissue type, or an entire organism. They are typically represented by directed graphs, where nodes represent genes, and an edge from node A to node B, means that A regulates B.
GRNs reconstructed from real data exhibit some common topological properties, such as hierarchical structure and the presence of a few highly connected genes, and many genes that are connected to only 1 or 2 genes.
We propose a simple model for the evolution of gene regulatory networks that allows us to construct evolutionary paths between networks with discrete time steps. These discrete time steps correspond to gene duplication events, followed by some ‘divergence’ whereby edges of the recently duplicated pair of genes can be modified.
We discuss the history of these types of models, and the biological motivations behind our proposed model. We also discuss limitations, a handful of interesting results, and future directions.

Topological Measures on Hypergraphs: with Application to Gene Expression Data

Diane Donovan, School of Mathematics and Physics The University of Queensland
dmd@maths.uq.edu.au Networks

(Joint work with James Lefevre, Samual Barton, Daniel Ortiz-Barrientos, and Zoe Broad)

Hypergraphs generalise graphs in that edges consist of subsets of vertices of varying size as distinct from subsets of size 1 or 2. In a network modelling setting, generalising the definition facilitates the study of multi-way interactions as opposed to pairwise interactions. To gain advantage from this generalisation we need to adapt the standard definitions of adjacency, paths and distance and use these to study measures such as diameter and betweenness centrality. I will address these issues in this talk and also discuss how the results can be used to interrogate gene expression data.

MAST: Phylogenetic Inference with Mixtures Across Sites and Trees

Thomas Wong, Research School of Biology, Australian National University
Thomas.Wong@anu.edu.au

Evolutionary Models

(Joint work with Caitlin Cherryh, ANU; Allen G Rodrigo, University of Auckland, NZ; Matthew W Hahn, Indiana University, USA; Bui Quang Minh, ANU; and Robert Lanfear, ANU)

Hundreds or thousands of loci are now routinely used in modern phylogenomic studies. Concatenation approaches to tree inference assume that there is a single topology for the entire dataset, but different loci may have different evolutionary histories due to incomplete lineage sorting, introgression, and/or horizontal gene transfer; even single loci may not be treelike due to recombination. To overcome this shortcoming, we introduce the mixture across sites and trees (MAST) model, which uses a mixture of bifurcating trees to represent multiple histories in a single concatenated alignment. The MAST model allows each tree to have its own topology, branch lengths, substitution model, nucleotide or amino acid frequencies, and model of rate heterogeneity across sites. We implemented the MAST model in a maximum-likelihood framework in the popular phylogenetic software, IQ-TREE. Simulations show that we can accurately recover the true model parameters, including branch lengths and tree weights (i.e. frequencies) for a given set of tree topologies. We also show that we can use standard statistical inference approaches to reject a single-tree model when data are simulated under multiple trees (and vice versa). We applied the MAST model to multiple primate datasets and found that it can recover the signal of incomplete lineage sorting in the Great Apes, as well as the asymmetry in minor trees caused by introgression among several macaque species. When applied to a dataset of four Platyrrhine species for which standard concatenated maximum likelihood and gene tree approaches disagree, we find that MAST gives the highest weight to the tree favoured by gene tree approaches. These results suggest that the MAST model is able to analyse a concatenated alignment using maximum likelihood while avoiding some of the biases that come with assuming there is only a single tree. The MAST model can offer unique biological insights when applied to datasets with multiple evolutionary histories.

The estimation and application of mixture models in phylogenetics (Student presentation)

Huaiyan Ren, Australian National University
u7151703@anu.edu.au

Evolutionary Models

Estimating accurate phylogenies requires good substitution models. Mixture models and partition models are two approaches to estimating substitution models, and both have been shown to improve phylogenetic accuracy. However, although, mixture models have many advantages over partition models, they are still rarely used. In this talk I will present research which shows first that one of the central assumptions of partition models (that a single model is adequate for a single partition) is unlikely to be true. I will then present an algorithm which allows users to easily estimate optimal phylogenetic mixture models for any dataset, and show using simulated data that the algorithm works well. Finally, I show the results of applying this algorithm on some empirical data.

An improved JHMM for detecting recombination from *Plasmodium falciparum* malaria parasite *var* genes

(Student presentation)
Qian Feng, Melbourne Integrative Genomics, University of Melbourne
fengq2@student.unimelb.edu.au

Evolutionary Models

(Joint work with Mun Hua Tan, Kathryn E. Tiedje, Karen P. Day. These three are all from School of BioSciences, The University of Melbourne, Bio21 Institute, Melbourne, Australia; Heejung Shim, Yao-ban Chan. Both are from Melbourne Integrative Genomics / School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia.)

The antigen PfEMP1 (*Plasmodium falciparum* erythrocyte membrane protein 1) plays a key role in the pathogenicity and immune evasion of *Plasmodium falciparum*, the deadliest malaria parasite. This antigen is encoded by the highly diverse *var* gene family. Recombination is one of the primary mechanisms for maintaining this diversity, and identifying recombination is thus of major interest to biologists. However, the lack of a reliable alignment for the *var* genes does not allow most standard methods to be used. The dominant model, developed by Zilversmit *et al.*, utilizes a jumping hidden Markov model (JHMM) that allows recombination between any two points in a pair of sequences, which is both biologically inaccurate and inefficient.

We propose an improved JHMM that constrains recombination to only act between nearby positions in an unaligned set of sequences, and use it to find recombination events between two time-separated datasets. We demonstrate the efficiency and accuracy of our method through simulations. In addition, we apply this algorithm to a large longitudinal dataset of *var* genes from Ghana. Our results are similar to previously published findings from a cross-sectional study, and we are additionally able to study the distribution of recombination breakpoints through time. This method provides a general framework for identifying recombinations in unaligned datasets.

Protracted speciation and extinction process

Xia Hua, Mathematical Sciences Institute, Australian National University
xia.hua@anu.edu.au

Evolutionary Models

How long does speciation take? The answer to this important question in evolutionary biology lies in the genetic difference not only among species, but also among lineages within each species. With the advance of genome sequencing in non-model organisms and the statistical tools to improve accuracy in inferring evolutionary histories among recently diverged lineages, we now have the lineage-level trees to answer these questions. However, we do not yet have an analytical tool for inferring speciation processes from these trees. What is needed is a model of speciation processes that generates both the trees and species identities of extant lineages. The model should allow calculation of the probability that certain lineages belong to certain species and have an evolutionary history consistent with the tree. Here we propose such a model and test the model performance on both simulated data and real data. By accounting for different rates and types of speciation processes across lineages in a species group, our model allows us to delimit species in a probabilistic way and to test speciation theories across all lineages in a species group. By explicitly linking evolutionary processes on lineage level to species level, the model provides a new phylogenetic approach to study not just when speciation happened, but how speciation happened.

Edge-based phylogenetic networks and their proximity measures

Mareike Fischer, University of Greifswald, Germany
`mareike.fischer@uni-greifswald.de` Europe

(Joint work with Tom N. Hamann and Kristina Wicke)

Phylogenetic networks which are, as opposed to trees, suitable to describe processes like hybridization and horizontal gene transfer, play a substantial role in evolutionary research. However, while non-treelike events need to be taken into account, they are relatively rare, which implies that biologically relevant networks are often assumed to be similar to trees in the sense that they can be obtained by taking a tree and adding some additional edges. This observation led to the concept of so-called tree-based networks, which recently gained substantial interest in the literature. Unfortunately, though, identifying such networks in the unrooted case is an NP-complete problem. Therefore, classes of networks for which tree-basedness can be guaranteed are of the utmost interest.
The most prominent such class is formed by so-called edge-based networks, which have a close relationship to generalized series parallel graphs known from graph theory. They can be identified in linear time and are in some regards biologically more plausible than general tree-based networks. While concerning the latter proximity measures for general networks have already been introduced, such measures are not yet available for edge-basedness. This means that for an arbitrary unrooted network, the "distance" to the nearest edge-based network could so far not be determined. In my talk, I will fill this gap by introducing two classes of proximity measures for edge-basedness.

An Update on Clumps, Or: Yes, We Can Recover Informative Clusters of Trees with Non-Identical Leaf Sets

Ana Serra Silva, The Natural History Museum, London
`a.da-silva@nhm.ac.uk` Europe

(Joint work with Mark Wilkinson, The Natural History Museum, London, UK)

Post-processing of trees often focuses on (multi)sets of phylogenetic trees with identical leaf sets, which can be effectively summarised using a multitude of consensus and clustering methods. However, with the increased popularity of phylogenomic analyses, the amount of tree sets with non-identical leaf sets that require specialised clustering approaches also increases. This need stems from the necessity to generalise tree-to-tree distances to pairs of trees with non-identical leaf sets, including how to deal with pairs of trees with no taxonomic overlap, and extreme size disparity between trees. In an attempt to tackle the former, we define a tree-to-supertree distance-based subsetting approach, "clumps of trees". Using examples from amphibian phylogenetics we will illustrate the expected outputs, the effects of changing analytical parameters and some of the potential pitfalls of the clumping approach. Time permitting, we will briefly show that clumping can also be applied to (multi)sets of trees with identical taxonomic sampling.

Lagged couplings diagnose MCMC phylogenetic inference

Luke Kelly, University College Cork
`lkelly@ucc.ie` Europe

(Joint work with Robin Ryder, Université Paris-Dauphine and Grégoire Clarté, University of Helsinki)

Phylogenetic inference attempts to reconstruct the ancestry of a set of observed taxa and is an intractable statistical problem on a complex, high-dimensional space. The likelihood function is an integral over unobserved evolutionary events on a tree and is often multimodal. Markov chain Monte Carlo (MCMC) methods are the primary tool for Bayesian phylogenetic inference but constructing sampling schemes to efficiently explore the associated posterior distributions or assess their performance is difficult.
Couplings have recently been used to construct unbiased MCMC estimators and compute bounds on the convergence of chains to their stationary distribution. We describe a procedure to couple a pair of Markov chains targeting a posterior distribution over a space of phylogenetic tree topologies, branch lengths, scalar parameters and latent variables such that the chains meet exactly at a random, finite time. We use the meeting times to diagnose convergence in total variation distance jointly across all components of the model on trees with up to 200 leaves.

Network-based polyploid phylogenetics (Student presentation)

Liam J. Maher, University of East Anglia, School of Computing Sciences
liamjmaher96@gmail.com Europe

(Joint work with K. T. Huber and T. Wu., University of East Anglia, School of Computing Sciences, UK.)

Cells with multiple (3+) complete sets of chromosomes is a main biological feature of certain species, for example bread wheat. This phenomenon is thought to have arisen via *polyploidisation*. Such events are rare and by calling the number of complete sets of chromosomes the *ploidy level*, the goal of my talk is to help shed light on the evolutionary history of polyploids based solely on the ploidy levels of a dataset of species. We address this question within a *phylogenetic network* framework (Semple, C. and Steel, M. *Phylogenetics.* Oxford University Press, 2003) in this talk. We begin with a brief introduction and explanation of relevant concepts before presenting some of our key findings.

Tropical logistic regression on phylogenetic trees (Student presentation)

George Aliatimis, Lancaster University
g.aliatimis@lancaster.ac.uk Europe

(Joint work with Ruriko Yoshida)

A phylogenetic tree is a tree representations of evolutionary history between species.
Tree leaves are labelled with current species and internal nodes representing ancestral species are unlabelled.
Recent technological advancements in genetics and genomics led to cheap and fast genome data generation.
This, in turn, paved the way for the development of *phylogenomics*, a new field that applies tools in phylogenetics to conduct comparative analyses on genome data.
Our goal is to develop statistical methods over the space of phylogenetic trees, which coincides with the ultrametric space and conforms to spaces from *tropical geometry*.
We developed the tropical logistic regression model, which projects phylogenetic trees onto a tropical hyperplane and classifies the corresponding genes to certain species.

The Tree of Blobs of a Species Network: Identifiability under the Coalescent

Hector Baños, Dalhousie University
hbassnos@gmail.com North America

(Joint work with John A. Rhodes and Elizabeth S. Allman, University of Alaska Fairbanks and Jonathan Mitchell, UTAS)

The network multispecies coalescent model (NMSC) is a standard probabilistic model describing the formation of gene trees in the presence of hybridization and incomplete lineage sorting. Inference methods under the NMSC are severely limited by heavy computational demands along with the uncertainty of how complicated a network can be to be consistently inferred. We present a step toward inferring a general species network by showing the identifiability of its tree of blobs, in which non-cut edges are contracted to nodes, so only tree-like relationships between the taxa are shown. This result depends upon an analysis of gene quartet concordance factors under the model, together with a new combinatorial inference rule.

Heterogeneity of Gene Tree Topologies

Jonathan N Odumegwu, Department of Biostatistics, School of Global Public Health, New York University
jonathan.odumegwu@nyu.edu North America

(Joint work with James H. Degnan of the Department of Mathematics and Statistics, University of New Mexico)

Multilocus phylogenetic studies often show a high degree of gene tree heterogeneity — gene trees that have different topologies from each other as well as from the species tree topology. In some cases, this can lead to studies with hundreds of loci having a distinct gene tree topologies. The degree of heterogeneity is expected to increase when there is a high degree of incomplete lineage sorting due to short branches (as measured in coalescent units) in the species tree. Other potential sources of heterogeneity include other biological processes such as introgression, recombination within genes, ancestral population structure, gene duplication and loss, and horizontal gene transfer, as well as gene tree estimation error due to short DNA sequences or inadequate substitution models. Here we examine the relationships between speciation and extinction rates and gene tree heterogeneity with both gene tree estimation error and no gene tree estimation error. In particular, higher speciation rates lead to shorter branches in the species tree and therefore higher levels of incomplete lineage sorting. In many cases, it might not be surprising that every gene tree has a unique topology, even for data sets with 1000 gene trees. We also propose using the average pairwise Robinson-Foulds (RF) distance between gene trees as a measure of heterogeneity as opposed to using the average RF distance between gene trees and the true species tree.

Using semi-algebraic constraints to design new weights for quartet-based methods

Marina Garrote-López, Department of Mathematics, University of British Columbia
mgarrote@math.ubc.ca North America

(Joint work with Marta Casanellas, Universitat Politècnica de Catalunya; Jesús Fernández-Sánchez, Universitat Politècnica de Catalunya)

In this talk we present new phylogenetic reconstruction methods based on algebraic and semi-algebraic tools. Algebraic tools have been incorporated in phylogenetic reconstruction methods during the last decade, however, in order to improve these methods, it is important to consider the semi-algebraic constraints imposed by the stochasticity of the parameters of the evolutionary models considered. Based on these constraints, we propose two new weighting systems for quartets evolving under a general Markov model. Furthermore, we will present the performance of different quartet-based methods in combination with these weights when reconstructing phylogenetic trees with more than four leaves. Finally, we will provide results on simulated and real data to illustrate the success of our methods.

Hit and Run Sampling from the Space of Phylogenetic Trees (Student presentation)

David Barnhill, Naval Postgraduate School Department of Operations Research
david.barnhill@nps.edu North America

(Joint work with Professor Ruriko Yoshida, Ph.D, Naval Postgraduate School and Professor Keiji Miura, Ph.D, Kwansei Gakuin University)

In this presentation we introduce a Markov Chain Monte Carlo (MCMC) Hit and Run (HAR) uniform sampler over a tropically convex space of ultrametrics. This is particularly important because by sampling from the space of ultrametrics, we are sampling from the space of phylogenetic trees, or tree space. This has wide ranging implications to statistical inference relating to this tree space. Specifically, we show how this HAR sampler can be employed to sample over the space of ultrametrics in order to non-parametrically estimate the phylogenetic tree distribution using what we call tropical density estimator (TDE) with the tropical metric. We compare the results of the TDE using the tropical metric against often used density estimation methods using the Billera-Holmes-Vogtman metric to show that TDE is more accurate and computationally less expensive.

[Poster] **Stochastic Safety Radius on UPGMA** (Student presentation)
Lillian Paul, Operations Research Department, Naval PostGraduate School
lillian.paul17@gmail.com Posters

(Joint work with Dr. Ruriko Yoshida, Naval Postgraduate School and LTC Peter Nesbitt, Naval PostGraduate School)

Unweighted Pair Group Method with Arithmetic Mean (UPGMA) is one of the most popular distance-based methods to reconstruct a phylogenetic tree from a distance matrix computed from an alignment of sequences. Stochastic safety radius introduced by Steel and Gascuel provides a lower bound for the probability that a phylogenetic tree reconstruction method returns the true tree topology from a given distance matrix. In this article, we compute stochastic safety radius of UPGMA for a phylogenetic tree with n leaves.

[Poster] **McCoy: A complete workflow for close-to-real-time phylodynamic analyses**
Wytamma Wirth, The University of Melbourne
wytamma.wirth@unimelb.edu.au Posters

(Joint work with Simon Mutch, Robert Turnbull, and Sebastian Duchene)

Here we present McCoy, a sequences to results pipeline for close-to-real-time phylodynamic analyses. This pipeline is built in Snakemake and enables online phylodynamic inference through a simple command-line interface. The pipeline includes fasta intake, alignment, quality control, phylodynamic and reporting modules. McCoy is an opinionated pipeline that streamlines routine phylodynamic analyses. Install McCoy with 'pip install mccoy'. Extensive documentation can be found at https://mccoy-devs.github.io/mccoy/index.html.

[Poster] **Further explorations of split space**
Michael Charleston, University of Tasmania
michael.charleston@utas.edu.au Posters

TBC

Subtree Prune and Regraft on Ranked Trees
Lena Collienne, University of Canterbury, School of Mathematics and Statistics
lena@lenacoll.de Tree Space

A number of methods for reconstructing phylogenetic trees from sequence data rely on tree sampling algorithms. For example, Bayesian inference methods like MrBayes, RevBayes, and BEAST use MCMC algorithms where in every step a tree similar to the current tree is proposed and accepted if it fulfils certain conditions. Often used for proposals are tree rearrangement operations like NNI (Nearest Neighbour Interchange) and SPR (Subtree Prune and Regraft), which perform a local change to a given tree to produce a similar but not identical tree. These operations can then be used to define a distance between two trees as the minimum number of rearrangements needed to transform one tree into another. In contrast to other distance measures like the Robinson-Foulds distance, these distance measures are biologically motivated and furthermore define tree spaces, which allow statistical analyses of distribution over trees. Decades of research on tree rearrangements NNI and SPR revealed that computing distances is NP-hard on unranked trees. Only recently it has been found that a modification of the Nearest Neighbour Interchange operation to ranked trees, where internal nodes are ordered according to times of the corresponding evolutionary events, allows efficient computation of distances. Though research on the corresponding RNNI (ranked NNI) tree space is still in its early stages, having a tree space with efficiently computable distances facilitates analysing distributions of trees, for example posterior distributions derived by software packages like BEAST.

In this talk we consider a modification of the Subtree Prune and Regraft (SPR) rearrangement to ranked trees and introduce two different tree spaces based on it. We discuss some properties of these tree spaces, focusing on those relating to the complexity of computing distances under ranked SPR operations as well as similarities and differences to other tree rearrangement based tree spaces. We furthermore show how adding leaves to trees can change the distance under ranked SPR operation significantly, which may have relevance to practical understanding of the results of tree sampling algorithms given uncertain "wandering taxa"".

Phylogenetic information geometry and its implications

Tom Nye, School of Mathematics, Statistics and Physicsl; Newcastle University; UK
tom.nye@ncl.ac.uk

Tree Space

(Joint work with Jonas Lueg, University of Goettingen, Germany; Maryam Garba, Newcastle University, UK; and Stephan Huckemann, University of Goettingen, Germany)

Most existing metrics between trees directly measure differences in topology and edge weights, and are unrelated to the models of evolution used to infer trees. We describe metrics which instead are based on distances between the probability models of discrete or continuous characters induced by trees. These behave very differently from existing metrics and we illustrate this using some simple examples. It is also desirable to construct shortest paths between trees, or geodesics, using these metrics.

The Billera-Holmes-Vogtmann (BHV) tree space is an example of a space with geodesics, and existence of geodesics has enabled development of a variety of statistical methods for analysing data sets of trees in BHV space. We briefly describe how construction of information-based geodesics leads to the recently proposed wald space of phylogenetic trees, and look at prospects for using wald space for statistical analysis of trees.

On the wald space for phylogenetic trees

Stephan Huckemann, Georg-August-University Göttingen
huckeman@math.uni-goettingen.de

Tree Space

(Joint work with Tom Nye, Jonas Lueg, and Maryam Garba)

We further explore the wald space for phylogenetic trees introduced in Tom Nye's talk. As a point set, it sits between the BHV space (Billera, Holmes and Vogtmann, 2001) and the edge-product space (Moulton and Steel 2004). It has a natural embedding in the space of positive definite matrices, equipped with the information geometry. Thus, singularities such as overlapping leaves are infinitely far away, proper forests, however, comprising the "BHV-boundary at infinity", are part of the wald space, adding boundary correspondences to groves (corresponding to orthants in the BHV space). In fact the wald space contracts to the complete disconnected forest. Further, it is a geodesic space, exhibiting the structure of a Whitney stratified space of type (A) where strata carry compatible Riemannian metrics. We explore some more geometric properties, but the full picture remains open. We conclude by identifying interesting and pressing open problems.

Hyperbolic Tree Embeddings: a Continuous Representation of Discrete Trees

Matthew Macaulay, The University of Technology Sydney
matthew.macaulay@sydney.edu.au

Tree Space

Navigating the high dimensional space of discrete trees for phylogenetics is a challenging problem for tree optimisation. Hyperbolic embeddings of trees offer a fruitful way to encode trees but require a differentiable tree decoder. This talk introduces soft-NJ, a differentiable version of neighbour-joining that enables gradient-based optimisation over the space of trees.

I'll illustrate the potential of soft-NJ with three examples in frequentist and Bayesian phylogenetics: maximum likelihood, maximum a posteriori and variational inference. Optimising trees (or tree distributions) in hyperbolic space can produce results close to state-of-art methods. However, geometric frustrations of the embedding locations produce local optima that pose a challenge for optimisation.

Subflattenings: What are they good for? (Student presentation)

Joshua Stevenson, University of Tasmania
joshua.stevenson@utas.edu.au

Computational Methods II

Since the dawn of time, phylogeneticists have asked the question: "Are these two groups of taxa separated by an edge in the tree?". Flattenings and subflattenings offer a direct method for answering this very question, in theory. But what about in practice? In this talk I will discuss our forthcoming paper that begins to explore this question, as well as possible avenues for future work.

Finding tree topologies from an alignment using site-rate variation (Student presentation)

Frederick Jaya, Research School of Biology, The Australian National University
frederick.jaya@anu.edu.au

Computational Methods II

(Joint work with Robert Lanfear, ANU)

Many phylogenetic analyses assume that all sites of a sequence alignment have evolved along a single tree. However, reticulate evolution and bioinformatic errors can often violate this single-tree assumption. In this talk I present a new method to identify a set of putative tree topologies from an alignment, and share preliminary applications to empirical data. The method leverages the observation that when sites are modeled on an incorrect tree, they will appear as fast-evolving sites. It uses the FreeRate model to identify classes of fast-evolving sites, and then determines whether these sites are likely to have evolved along a separate tree topology from the remainder of the alignment. I will share results from applying this method to two empirical datasets, as well as plans to use simulations to identify the strengths and weaknesses of the approach.

Is time-oriented phylogenetic reconstruction of thousands of sequences possible? (Student presentation)

Robert McArthur, School of Computing, Australian National University
robert.mcarthur@anu.edu.au

Computational Methods II

(Joint work with Yu Lin and Gavin Huttley)

Applying phylogenetic reconstruction techniques to data with thousands of homologous sequences is difficult. Due to the larger number of possible solutions, it becomes even more difficult when the desired outcome is a time-oriented tree. This difficulty arises for both the alignment and the tree reconstruction steps. We employ a divide-and-conquer strategy to improve tractability, examining the suitability of two published algorithms for this purpose: (1) the Disk Coverage Method (Roshan et al 2004), which splits a set into over-lapping subsets; (2) the Semple and Steele (2000) algorithm, for merging rooted trees. The latter algorithm is the principal bottleneck due to its reliance on many iterations of min-cut. We estimated the scalability of these algorithms in terms of both memory and compute time. We show that even given algorithm (2) limitations, tackling data sets with 1000 sequences can be feasible on a laptop computer.

Networks and Covers

Andrew Francis, Western Sydney University
a.francis@westernsydney.edu.au

Computational Methods II

The set of forests of phylogenetic trees is in bijection with the set of partitions of finite sets [F & Jarvis, 2022]. This bijection generalises correspondences for phylogenetic trees, and in particular for binary trees, which correspond with perfect matchings [Diaconis & Holmes, 1998]. But is there a set-theoretic correspondence for phylogenetic networks? In this talk I will discuss some of the challenges with this generalisation and some steps towards a resolution involving covers of finite sets.

Matrix-analytic methods for the evolution of species trees, gene trees, and their reconciliation

Małgorzata O'Reilly, University of Tasmania
malgorzata.oreilly@utas.edu.au

Matrix Analytic Methods

(Joint work with Albert C. Soewongsono[1]*, Jiahao Diao*[1]*, Tristan Stark*[1]*, Amanda E. Wilson*[2]*, David A. Liberles*[2]*, and Barbara R. Holland*[1]. [1]*: University of Tasmania;* [2] *Temple University)*

These are three talks, Parts I-III, by Małgorzata, Albert, and Jiahao, respectively.
In the reconciliation problem the task is to map a given gene tree into the known species tree. Our goal is to maximize the likelihood of such fitting, given the data is incomplete. That is, it is not known which nodes in the gene tree are duplications and which are speciations, and furthermore, the extinctions are not observed.
Part I: Małgorzata M. O'Reilly
We describe a Markovian Binary Tree (MBT) model for the species tree and a Quasi-Birth-and-Death process (QBD) model for the gene tree. We derive new results using the theory of matrix-analytic methods (MAMs) and describe efficient algorithms for the computation of a range of useful metrics. We present methodology for the analysis of the reconciliation problem.
Part II: Albert Soewongsono
We construct examples of MBT models for the evolution of species. We illustrate the application of our theoretical results in the analysis of key metrics of a species tree through numerical examples. We provide the physical interpretations of these quantities and how they can be applied to data.
Part III: Jiahao Diao
We construct examples of QBD models for the evolution of gene families. We describe how they can be applied in the analysis of gene tree balance. We present an application of our methodology for the reconciliation problem through numerical examples.

Matrix-analytic methods for the evolution of species trees, gene trees, and their reconciliation (Part II)

(Student presentation)

Albert Christian Soewongsono, School of Natural Sciences, University of Tasmania
albert.soewongsono@utas.edu.au

Matrix Analytic Methods

(Joint work with Jiahao Diao, Tristan Stark, Amanda E. Wilson, David A. Liberles, Barbara R. Holland, Małgorzata M. O'Reilly)

Please refer to the abstract submitted by Małgorzata M. O'Reilly for her talk.

Matrix-analytic methods for the evolution of species trees, gene trees, and their reconciliation (Part III) Species-gene tree reconciliation

Jiahao Diao, University of Melbourne
jiahao.diao@unimelb.edu.au

Matrix Analytic Methods

We consider the reconciliation problem, in which the task is to find a suitable gene tree that fits the given species tree, so as to maximize the likelihood of such fitting, given the available data. We describe a model for the species tree, a model for the gene tree, and construct methodology for the analysis of reconciliation. We derive our results using the theory of matrix-analytic methods and describe efficient algorithms for the computation of a range of useful metrics. We illustrate the theory with examples and provide the physical interpretations of the discussed quantities, with the focus on the practical applications of the theory to incomplete data. (Part III of Matrix-analytic methods for the evolution of species trees, gene trees, and their reconciliation.)

Approximate Bayesian computation for Markovian binary trees in phylogenetics (Student presentation)

Mingqi He, School of Mathematics and Statistics, University of Melbourne
mingqih1@student.unimelb.edu.au

Matrix Analytic Methods

(Joint work with Yao-ban Chan, Sophie Hautphenne; School of Mathematics and Statistics, University of Melbourne)

Phylogenetic trees describe the relationships between species in the evolutionary process and provide information about the rates of diversification. To understand the mechanisms underneath macroevolution, we consider a class of multitype branching processes called the Markovian Binary trees (MBTs). MBTs allow for variation in the diversification rates, and provide a flexible and realistic model for the growth of phylogenetic trees. However, due to its complex structure, the analytical likelihood function for a given tree is intractable. Approximate Bayesian Computation (ABC) is a likelihood-free inference method that provides a solution to this problem.

In this research, we use ABC to infer the rates in MBTs by exploiting the shape of phylogenetic trees. We evaluate the accuracy of this inference method using simulation studies. Results suggest that ABC methods can detect the variation in the diversification rates. This finding demonstrates the potential of ABC methods in the analysis of diversification rates in phylogenetics, which takes a step forward in the implementation of complex diversification models in phylogenetic trees.

Is the Human genome in mutation equilibrium?

Katherine Caley, Research School of Biology, The Australian National University
katherine.caley@anu.edu.au

Statistical Methods

(Joint work with Ben Kaehler, School of Science, University of New South Wales; Von Bing Yap, Department of Statistics and Applied Probability, National University of Singapore; Gavin Huttley, Research School of Biology, The Australian National University)

Most models of sequence divergence assume the composition of nucleotides does not change through time. This assumption implies a state of mutation equilibrium which is almost impossible if the processes affecting mutagenesis change through time. Considerable empirical evidence strongly suggests that this may be incorrect. In my Honours, I have addressed this possibility through developing the following statistical measures: a test for the existence of mutation disequilibrium, a test of its equivalence, and a measurement of the magnitude of mutation disequilibrium. I used carefully constructed edge cases with simulated data to establish the consistency of the statistics with theoretical expectations. I applied the statistics to empirical data from cases with striking prior evidence for recent perturbations affecting: an entire genome (loss of DNA methylation in *Drosophila melanogaster*); or, a small genomic segment (FXY in *Mus musculus*). Using paired experimental designs, I show the predicted vast excess of small probabilities from the statistical tests. I further show the statistical measure of magnitude is also elevated in these case. Applying the methods to the Human evolution, I conservatively estimate $> 50\%$ of our genome is in mutation disequilibrium.

Statistics in the space of ranked time trees (Student presentation)
Lars Berling, School of Mathematics and Statistics, University of Canterbury
berlinglars96@gmail.com Statistical Methods

It is common to encounter large collections of phylogenetic trees with many commonalities: Bayesian methods are based on sampling many trees, multiple different trees can be equally likely or have the same parsimony score or assessing uncertainty of a dataset via rerunning either Maximum Likelihood or Parsimony methods can result in slightly different trees each time. For such collections of phylogenetic trees it has become standard practice to use consensus or summary tree methods to compute a single tree as a representative. However, this is often a challenge as uncertainties in sequence data can lead to conflicting trees and result in polytomies. These are helpful for displaying the uncertainty in data but many summary tree heuristics generally result in partially unresolved trees even though a fully resolved tree could be constructed. Another common phenomenon are tree islands which are distinct sets of different trees that represent the given data equally well and hence lead to multiple possible results in repeated analyses. Applying consensus methods to datasets with multiple such islands will result in a summary tree with low support by the data and possibly many polytomies. Although these problems are well known, no solution addressing these problems is currently available. Most notably, the Billera-Holmes-Vogtmann (BHV) tree space has been developed and much progress on the development of statistical tools within it has been made. However, it has since been shown that problems such as stickiness of mean trees, which is a general tendency of summary trees to be partially unresolved, make this space unsuitable.
Here we present statistical analysis within the recently developed RNNI tree space which is a modification of the popular nearest neighbour interchange move for ranked trees. We present an algorithm that approximates mean trees, which are inherently non-sticky, and show that it outperforms popularly used tree summaries on several tree based measures, including the likelihood values. Moreover, we are able to showcase further properties of the RNNI tree space that are desirable in the context of statistical analysis of time trees.

Model validation and selection in phylogenetic comparative analyses using posterior predictive methods
Luke Yates, School of Natural Sciences, University of Tasmania
luke.yates@utas.edu.au Statistical Methods
(Joint work with Ben Halliwell, Barbara Holland)

To be supplied.

Unpacking phylogenetic inference: residual diagnostics and goodness-of-fit tests (Student presentation)
Qin Liu, University of Tasmania
qin.liu@utas.edu.au Statistical Methods

Model selection is an essential process in the pipeline of phylogenetic inference since choosing the incorrect model can lead to biased inferences. However, the current procedure of model selection typically involves relative model selection. That is, the best-fit model is selected from among a set of candidate models by using some selection criteria (e.g., AIC or BIC), and the chosen model could be the “least-bad” model among all the candidate models. In other words, the “best” model may still perform poorly. For this reason, it is crucial to assess the absolute adequacy of the best-fit model to the data and, most importantly, to determine what factors influence the model violation and how badly the lack of fit affects the inference.
Goodness-of-fit tests and residual diagnostics show whether and how the model fits poorly. Residual diagnostics are more useful and powerful than goodness-of-fit tests for two reasons. Firstly, the goodness-of-fit tests, which measure the discrepancy between the observed values and the fitted model’s expected, only determine whether the model is adequate but do not explain how the inadequate fits arise. Secondly, the goodness-of-fit tests may always fail the model if we have large data. This is because, compared to a smaller data set, a larger data set provides more opportunities to show discrepancies between the observed and fitted values and increases the chances of rejecting the model. This does not mean that the model offers insufficient overall adequacy to the data.
Developing some effective residual diagnostic tools is the main focus of our current research. We are also reviewing the existing methods for assessing goodness-of-fit. We aim to build an R package including useful residual diagnostics and goodness-of-fit tests for model validation.

The large-sample asymptotic behaviour of quartet-based summary methods for species tree inference

Yao-ban Chan, School of Mathematics and Statistics / Melbourne Integrative Genomics, The University of Melbourne
`yaoban@unimelb.edu.au`

Statistical Methods

(Joint work with Qiuyi Li, School of Mathematics and Statistics / Melbourne Integrative Genomics, The University of Melbourne, and Celine Scornavacca, Institut des Sciences de l'Evolution, Université Montpellier)

Summary methods seek to infer a species tree from a set of gene trees. A desirable property of such methods is that of statistical consistency; that is, the probability of inferring the wrong species tree (the error probability) tends to 0 as the number of input gene trees becomes large. A popular paradigm is to infer a species tree that agrees with the maximum number of quartets from the input set of gene trees; this has been proved to be statistically consistent under several models of gene evolution. In this talk, we show that the asymptotic behaviour of the error probability of such methods in this limit decays exponentially, as well as the Robinson-Foulds distance between the true and inferred species trees. For a 4-taxon species tree, we derive a closed form for the asymptotic behaviour in terms of the probability that the gene evolution process produces the correct topology. We also derive bounds for the sample complexity (the number of gene trees required to infer the true species tree with a given probability), which outperform existing bounds. We then extend our results to bounds for the asymptotic behaviour of the error probability for any species tree, and compare these to the true error probability for some model species trees using simulations.

Which clade(s) of eudicot *NCED* genes are triggered by dehydration? (Student presentation)

Hanh Minh Vo, School of Natural Sciences, UTAS
`hanhminh.vo@utas.edu.au`

Tenuously Connected to Phylogenetics

(Joint work with Michael Charleston, Timothy Brodribb, and Frances C. Sussmilch)

In response to dehydration stress, some plants can rapidly biosynthesise the stress hormone abscisic acid (ABA). ABA limits water loss by inducing closure of pores called stomata on leaves and promotes desiccation tolerance. Within the ABA biosynthesis pathway, the genes that encode the rate-limiting nine-*cis*-epoxycarotenoid dioxygenase (NCED) enzymes are the only genes induced within the time frame of stomatal closure. Most plant species have multiple *NCED* genes. In this study, we tested whether there is a phylogenetic pattern to which genes are associated with dehydration stress responses in the largest group of flowering plants: the eudicots.

We first performed phylogenetic analysis using available sequence information for diverse land plant species to understand the evolutionary history of the *NCED* gene family. We found evidence of an ancient duplication event in the angiosperm (flowering plant) lineage yielding two major NCED clades (NCEDI and NCEDII). Our results showed a second major duplication event in the NCEDII clade in eudicots, giving rise to two large, eudicot-specific clades (NCEDIIA and NCEDIIB) with subsequent diversification. We then tested which *NCED* genes were transcriptionally induced by dehydration under controlled conditions in three diverse eudicot species - *Arabidopsis thaliana*, pea (*Pisum sativum*) and tomato (*Solanum lycopersicum*). Our results show that minor dehydration stress can trigger expression of *NCED* genes from the NCEDIIB clade. These results suggest that genes in this clade may share an evolutionarily conserved role in rapid responses to dehydration stress in eudicots. We can use this information for improved, phylogenetically informed prediction about genes of interest for dehydration responses within this important multigene family in other eudicot species.

Using genomic signatures of natural selection to elucidate MS genetics

Bennet McComish, Menzies Institute for Medical Research, University of Tasmania
bennet.mccomish@utas.edu.au

Tenuously Connected to Phylogenetics

Multiple sclerosis prevalence shows a heterogeneous geographical pattern, with higher prevalence in populations of European ancestry, as well as increasing with distance from the equator within those populations. This pattern has likely been shaped by both natural selection and neutral genetic drift. Identifying genes that have undergone selection at MS risk loci will improve our understanding of the causative mechanisms behind the disease. Population genomics can be used to identify functional variation that has been subject to natural selection at loci associated with MS risk.

We carried out genome-wide scans for natural selection using cross-population extended haplotype homozygosity in population genomic data. MS-related selection was localised by targeting genes prioritised in the large genome-wide association study carried out by the International Multiple Sclerosis Genetics Consortium. Strong signatures of natural selection in European and Asian populations were identified in several MS risk genes. Further analysis of these genes is underway to identify likely causes of selection and mechanisms by which they may contribute to MS risk. This approach allows us to narrow down candidate genes and pinpoint a small number of top causal candidates and mechanisms. This may enable more informed targeting of the molecular mechanisms behind the disease.

Evolutionary implications of genome size and cell size

Greg Jordan, University of Tasmania
greg.jordan@utas.edu.au

Tenuously Connected to Phylogenetics

It is now clear that cell size and total genome size are somehow linked, both within species (e.g., through ploidy changes) and phylogenetically. Furthermore, there is good evidence that cell sizes of different organs are correlated, and that both genome size and cell size have important implications for plant function and the habitats that species occupy.

I will discuss the links between these different factors, with an eye to extracting some general patterns and also to understanding the challenges of interpreting the various correlations involved. One central idea is that genome size is one of a set of factors that have global impacts on phenotype. Such generic impacts have benefits in terms of assisting in coordinating function among tissues and costs that the coordination may create imbalances in function. I will even provide some evidence! However, I will also discuss the exceptions to these rules, and what they mean for understanding of evolutionary processes.

Global variation in the relationship between avian phylogenetic diversity and functional dispersion is driven by environmental constraints (Student presentation)

Keaghan J Yaxley, Department of Archaeology and Anthropology, The University of Cambridge
kjy25@cam.ac.uk

Tenuously Connected to Phylogenetics

(Joint work with Alexander Skeels and Robert A Foley)

If evolutionary distance is akin to evolutionary opportunity, then it follows that species assemblages that are distantly related will also be more disparate in terms of their traits, features and the niches they occupy. Yet, studies have found that the total phylogenetic distance of assemblages, known as phylogenetic diversity, is an unreliable surrogate for functional diversity. Here we characterise the global relationship between Faith's Phylogenetic Diversity and two measures of functional diversity (Functional Dispersion and Mean Pairwise Functional Distance) in birds using morphological traits for over 9,000 species across more than 17,000 assemblages. Globally, after correcting for species richness differences, assemblages that were phylogenetically diverse tended be less functionally dispersed than expected, however this relationship showed considerable variation across latitude and between regions. In the Northern Hemisphere, the negative relationship between phylogenetic diversity and functional dispersion gradually weakens with distance from the equator and at the highest latitudes becomes positive. In the Southern Hemisphere we observe a consistent but weak relationship between the two variables across all latitudes. Interestingly, the effect also varies longitudinally. While in the Americas and Asia and Oceania, the relationship between richness-corrected phylogenetic diversity and functional dispersion varies with latitude, whereas in Europe and Africa it appears to have no impact. We suggest that latitudinal effects are in part driven by the disproportionate representation of morphologically constrained and phylogenetically clustered migratory species at high latitudes in North America and Eurasia. We also show that the Tropical Andes and the Hengduan Mountains are regions with exceptional avian functional dispersion but relatively low phylogenetic diversity, and that low-lying tropical regions such as the Amazon and the Indonesian Archipelago are characterised by functionally constrained species assemblages. Our study highlights the importance of evolutionary and environmental context when considering the surrogacy of phylogenetic diversity for measures of functional diversity, and the value in characterising the variation of that surrogacy across environmental gradients.

TBC

Nick Fountain-Jones, University of Tasmania
nick.fountainjones@utas.edu.au

Tenuously Connected to Phylogenetics

(Joint work with TBC)

TBC

Generating networks for epidemiological dynamics of infectious disease (Student presentation)

Raima Carol Appaw, UTAS (Mathematics, School of Natural Sciences)
raima.appaw@utas.edu.au

Tenuously Connected to Phylogenetics

Contact networks, often represented by mathematical graphs, have long been studied in terms of their structural features as predictors of epidemiological dynamics of infectious disease. We consider a range of standard network types: so-called random graphs (where every individual has the same probability of connections), small-world, lattice, scale-free, a "spatial" network, and networks generated by the exponential random graph models (ERGMs). Our investigation focuses on how the number of individuals (network's size) is impacted by network structural features (e.g., mean degree) in the prediction of disease spread across the standard networks. We compare ERGMs to the standard models using various network similarity measures such as isomorphism, vertex entropy, Kolmogorov-Smirnov distance, single value decomposition, the correlation coefficient of graph spectrum, and machine learning classification algorithms to determine which model best replicates network properties (especially Fiedler and Spectral radius) of a given empirical network. We find that ERGMs reproduce most network properties better than the standard models. In addition, epidemic characteristics of a given empirical network – such as peak size, time to peak, and outbreak duration – were better captured by ERGMs compared with the standard networks. We conclude that for comparing artificial with empirical networks and for reliable prediction of epidemics on networks (especially those of small size), it is crucial to capture the appropriate network generative model that can reproduce both the empirical network characteristics and epidemic dynamics.

Everyone