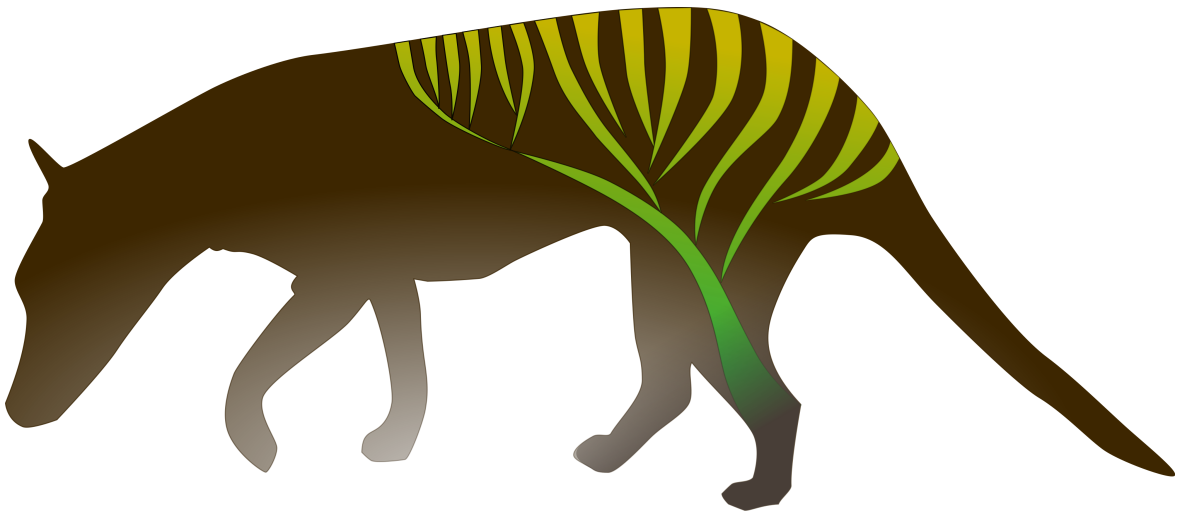
Phylomania 2018

Maths & Physics Building,
Sandy Bay Campus,
University of Tasmania, Hobart
November 21-23



Schedule

All seminars will be in Lecture Theatre 1, accessible from Level 2 of Maths & Physics Building.

Morning and afternoon tea, and lunches, will be in Room 333 on the level above.

	Wednesday	**Thursday**	**Friday**
09:00	Conference Opening	Housekeeping	Housekeeping
09:10	Rohan Mehta *Probability of Monophyly*	Kristina Wicke *Non-binary Tree-based Networks*	Yao-ban Chan *Network/Tree Reconciliation*
09:30	Jonathan Mitchell *Testing n-taxon Species Trees*	Mareike Fischer *Classes of Tree-based Networks* (combination 60 minute talk)	Cassius Manuel Pérez *Sequences with Taboos*
09:50	Venta Terauds *Are We Related?* (40 minute talk)		Ben Wilson *Embedding Trees* (40 minute talk)
10:10		Yukihiro Murakami *Reconstructing Tree-child networks*	
10:30	Morning Tea & Coffee		
11:00	Janan Sykes *Protein Structure Alignment*	Caitlin Cherryh *Treelikeness*	Olga Chernomor *Phylogenetic Terraces*
11:20	Jotun Hein *Protein Structure Evolution*	Julia Shore *Linearity and New Codon Models*	Suha Naser-Khdour *Model Violations*
11:40	Kevin Downard *Phylonumerics* (40 minute talk)	Rob Lanfear *Concordance Factors*	Matthew Hahn *Misleading Maximum Likelihood*
12:00		Jiahao Diao *Evolution of Gene Duplicates*	Aaron Darling *Surprise!*
12:20	Lunch		
14:00	David Bryant *Tree Decompositions*	Chris Burridge *Seabird Genetic Differentiation*	Michael Woodhams *Closure in Codon Models*
14:20	Michael Charleston *Fast MAD Algorithm*	Conrad Burden *2-island, 2-allele Wright Fisher* (40 minute talk)	Tim McInerney *Haplotype Blocks*
14:40	Marnus Stoltz *Backward Diffusions*		Greg Jordan *Ancestral States Matter*
15:00	Afternoon Tea & Coffee		*Posters, Prizes and Closing Remarks*
15:30	Barbara Holland *Are L-M Models more Robust?*	John Hewson *Graded Rings*	Drinks and celebrations!
15:50	Jeremy Sumner *Markov Association Schemes* (40 minute talk)	Minh Bui *Protein Evolution*	
16:10		Matilda Brown *Support Vector Machines*	
16:30	Pub O'Clock		
	World's End Brew Pub *or Preachers*	Conference Dinner from 7pm *Tom M^{c}Hugo's*	Preachers *or World's End*

"Pub O'Clock" has several options, two being the most obvious as they're our most frequent drinking holes: World's End Brew Pub on Sandy Bay Road, and Preachers in Battery Point.

The Conference Outing on Saturday

- The traditional Saturday Outing will happen again. Stay tuned for details (and we will have to be flexible to weather and people's changing desires) but the rough plan is to optionally go to the Salamanca Market on Saturday morning (quite touristy but definitely worth a visit if you haven't already), and then take a mid-level walk from about noon in the foothills of Mt Wellington.
- Sturdy-ish shoes are required (sneakers *may* be ok…).
- If you do intend to take the walk, you must bring a sunhat and sunscreen. Tasmanian sun is brutal and you will get burned if you don't take adequate precautions.
- There are lots of other interesting things to do in Hobart! There's MONA, Mawson's Hut (a replica), the museum and art gallery (TMAG), and a fair number of pubs. Ask us for more ideas.

Abstracts

Matilda Brown, University of Tasmania

Using two-class support vector machines to identify ecological overlap

Support vector machines (SVMs) are a class of machine learning classifiers that can be exploited to solve a set of problems faced by ecologists. Many ecologists considering past climates, invasive species or predicting species' responses to climate change are interested in questions about the ecological similarity of populations or species. These questions are mostly studied using correlative modelling techniques which use the climatic data from occurrence records to construct probabilistic models of habitat suitability, which are then compared to evaluate ecological overlap.
I will argue that applying SVMs to pairs of species in n-dimensional environmental space is a better method to answer these questions because it does not require the boundaries of each hypervolume to be delineated. Instead, the occupation of environmental space by each taxon can be directly compared. This means that the results obtained using this method are less likely to be affected by sampling bias, environmental availability and low sample size (due to geographical or climatic restriction). The theoretical advantages of using two-class SVMs are supported by the results from real-world data on conifers, showing that this method provides improved estimates of ecological similarity.

David Bryant, University of Otago

Tree decompositions make things fast
(Joint work with Celine Scornavacca)

In 2004, Brodal, Faberberg and Pedersen came up with an extremely clever $O(n \log n)$ time algorithm for computing the number of quartets that two trees have in common. Key to the algorithm was a particular hierarchical decomposition of the set of edges in a phylogeny. Celine Scornavacca and I modified and simplified the decomposition as part of an $O(n \log n)$ time algorithm for computing the path-length distance between trees. I think this new decomposition could have wider application. As an example, I'll give an $O(\log n)$ (parallel) time algorithm for computing the likelihood of a tree.

Minh Bui, Australian National University

A new model of protein evolution

Protein sequence analyses are based on empirical models of amino-acid replacements. However, the estimation of the most widely used models (LG and WAG) had several limitations. I will present a much improved estimation procedure implemented in IQ-TREE and a new model (IQ-REV) that outperforms LG and WAG. I will also introduce another version (IQ-NONREV) relaxing the time-reversibility assumption and its application to infer the root of Archaea.

Conrad Burden, Australian National University

Stationary distribution of a 2-island 2-allele Wright-Fisher diffusion model
(Joint work with Robert Griffiths, University of Oxford)

Population genetics models of migration between partially isolated subpopulations, known as 'island models' have existed since the pioneering work of Wright in 1943 (*Isolation by distance*, Genetics **28**, 114). Here we consider a 2-island, 2-allele Wright-Fisher model with small mutation rates between the two alleles and small migration rates between the islands. In the diffusion limit of infinite population sizes, the allele distribution becomes a density $f(x, y)$ defined on the unit square $\Omega = [0, 1] \times [0, 1]$, where x and y are the relative proportions of type-1 alleles on the first and second islands respectively. The stationary distribution is obtained by solving the forward-Kolmogorov equation to leading order in mutation and migration rates as a set of line densities on the edges of Ω, corresponding to states for which one island is bi-allelic and the other island is non-segregating, and a set of point masses at the corners of the sample space, corresponding to states for which both islands are simultaneously non-segregating. Analytic results for the corner probabilities and line densities are verified independently using the backward generator and for the corner probabilities using the coalescent.

Chris Burridge, University of Tasmania

Predictors of genetic differentiation among seabird populations…or lack thereof.
(Joint work with Anicée Lombal)

Understanding the historical and contemporary factors leading to genetic differentiation of populations is highly desirable to help identify conservation priorities and maintain viability of species. In this study, we evaluated a candidate set of generalized linear models (GLMs) to identify contemporary contributors to population differentiation in mitochondrial DNA (mtDNA) for 73 seabird species. The lack of mutation-drift equilibrium observed in 19% of seabird species coincided with lower estimates of genetic structure. Historical fragmentation was the best predictor of genetic differentiation within Tropical and Southern Temperate species and was supported by variation in phenotypic traits, whereas differences in post-breeding movement patterns among colonies and International Union for Conservation of Nature (IUCN) status did not appear as significant predictors of population genetic structure in any of our GLMs. Hierarchical comparisons of genetic partitioning showed that all Southern Temperate species sampled on the Falkland Islands exhibited the highest change in allele frequency around that zone. This suggests that isolation of those colonies during the Pleistocene may have been reinforced by contemporary factors such as high prey availability in the Patagonian shelf and limited foraging movements. These results show that signatures of historical events still dominate as contributors to genetic structuring among seabird colonies but are reinforced by contemporary factors such as ocean productivity.

Yao-ban Chan, The University of Melbourne

Reconciliation of a gene network and species tree

(Joint work with Charles Robin, The University of Melbourne)

The phylogenetic trees of genes and the species which they belong to are similar, but distinct due to various evolutionary processes which affect genes but do not create new species. Reconciliations map the gene tree into the species tree, explaining the discrepancies by events including gene duplications and losses. However, when duplicate genes undergo recombination (a phenomenon known as non-allelic homologous recombination or paralog exchange), the phylogeny of the genes becomes a network, not a tree. In this talk, we explore how to reconcile a gene network to a species tree with duplications and losses. We show that an extension of the lowest common ancestor (LCA) mapping solves the problem for a restricted class of networks, give a polynomial-time algorithm which bounds the position of each gene, and show that the general problem is fixed-parameter tractable in the level of the network.

Michael Charleston, University of Tasmania

A fast MAD algorithm to root phylogenetic trees

(Joint work with David Bryant, University of Otago)

The Minimal Ancestral Deviation (MAD) method is a recently introduced procedure for estimating the root of a phylogenetic tree, based only on the shape and branch lengths of the tree. The method is loosely derived from the midpoint rooting method, but, unlike its predecessor, makes use of all pairs of OTUs when positioning the root. In this note we establish properties of this method and then describe a fast and memory efficient algorithm. As a proof of principle, we use our algorithm to determine the MAD roots for simulated phylogenies with up to 100,000 OTUs. The calculations take a few minutes on a standard laptop.

Olga Chernomor, CIBIV, University of Vienna

Characteristics of phylogenetic terraces and their influence on tree space exploration

(Joint work with Lukasz Reszczynski (CIBIV, Medical University of Vienna), Arndt von Haeseler (MFPL, CIBIV, University of Vienna, Medical University of Vienna))

In phylogenomics, the analysis of concatenated gene alignments, the so-called supermatrix, is commonly accompanied by the assumption of partition models. Under such models, each gene, or more generally partition, is allowed to evolve under its own evolutionary model. Though partition models provide a more comprehensive analysis of supermatrices, missing data may hamper the tree search algorithms due to the existence of phylogenetic terraces - collections of different species-trees with identical score (maximum likelihood or parsimony score).
For sparse supermatrices with a lot of missing data, the number of terraces and the number of trees on the terraces can be very large. If terraces are not taken into account, a lot of computation time might be unnecessarily spent to evaluate many trees that in fact have identical score. Thus, exploration of tree-space is inefficient and due to limitations of numerical accuracies, the trees on a terrace will have slightly different scores, another unwanted effect.
In our previous work, we provided an efficient way of saving computational time by identifying consecutive species-trees that lie on the same terrace, and generalized the concept to partial terraces, which occur more frequently than “full” terraces providing additional timesaving possibilities.
Here, we continue with the characterization of phylogenetic terraces and provide insights into combinatorial properties of phylogenetic tree search in the presence of missing data. Efficient sampling of trees from the terrace allows studying terraces for any type of coverage patterns. Among others, we discuss the probability to leave a terrace under random nearest neighbor interchange and whether it is useful to consider more than one tree from the terrace during the search. Such insights are valuable for understanding the tree space exploration for contemporary phylogenomic alignments.
(i) O. Chernomor, B.Q. Minh, and A. von Haeseler (2015) Consequences of Common Topological Rearrangements for Partition Trees in Phylogenomic Inference. *Journal of Computational Biology*, Dec; **22**(12):1129-42. (DOI:10.1089/cmb.2015.0146);
(ii) O. Chernomor, A. von Haeseler, and B.Q. Minh. (2016) Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology* (DOI:10.1093/sysbio/syw037)

Caitlin Cherryh, ANU

A new test for treelikeness in phylogenetic data

(Joint work with Bui Quang Minh (ANU), Robert Lanfear (ANU))

Most phylogenetic analyses assume that the evolutionary history of a locus can be described by a single bifurcating tree. We call this the treelikeness assumption. Treelikeness may be violated by alignment error, incomplete lineage sorting, introgression, or recombination. If the treelikeness assumption is violated, this may affect downstream inferences, such as those which seek to estimate species trees from sets of gene trees. Here, we develop an approach to measure the treelikeness of a sequence alignment, and to test whether a single bifurcating tree is an adequate representation of that alignment. We evaluate the test with a comprehensive set of simulations and compare the test to some related methods. The approach is freely available with open source code, and we hope that it will enable biologists to choose appropriate loci when conducting phylogenetics analyses.

Jiahao Diao, University of Tasmania

Model for evolution of the family of gene duplicates

(Joint work with Tristan L. Stark and David A. Liberles, Temple University; Małgorzata M. O'Reilly and Barbara R. Holland, University of Tasmania.)

Consider a Markov model for the evolution of the family of genes proposed in Teufel *et al.* 2014 (Section 10), in which the state (n, m) records the number copies $n = 0, 1, 2, \ldots$, and the number of *redundant* copies $m = 0, 1, , \ldots, n$ of a gene in the family. By redundant copies we mean copies whose loss will not result in the loss of the functions of the gene when these are preseved by some other genes in the family.

The transition rates between the state of the Markov model in Teufel *et al* depend on the following key processes: duplication of a gene; loss of one copy of a gene; one copy acquiring a new function (neofunctionalisation); and a loss of some regulatory region in one of the genes which leads to a number of genes required to fulfill some function (subfunctionalisation).

The aim of this work is to develop suitable expressions for the transition rates function. As an initial inspiration of our analysis, we consider a Markov model with a more detailed state, represented as a matrix, where rows corresponds to the various genes in the family, and columns to their functions. That is, state is a binary matrix $\mathbf{A} = [A_{ij}]_{i=1,\ldots,M;j=1,\ldots,Z}$, where M is the number of genes in the family, and Z the number of functions, such that $A_{ij} = 1$ when function j is performed by some regulatory region of gene i, and 0 otherwise.

Although we suspect that the model with the detailed representation will not be tractable for developing analytic solutions (except in very small cases) we propose to use it as a simulation model to give insight into suitable transition rates in the reduced state space representation.

(i) A. Teufel, J. Zhao, M. O'Reilly, L. Liu, and D.A. Liberles. On Mechanistic Modeling of Gene Content Evolution: Birth-Death Models and Mechanisms of Gene Birth and Gene Retention. *Computation* 2014, **2**(3), 112-130; https://doi.org/10.3390/computation2030112

Kevin Downard, University of New South Wales, Sydney

Phylonumerics: A New Mass-Based Phylogenetics Approach to Study Mechanisms of Antiviral Resistance in the Influenza Virus

A new and novel mass-based "phylonumerics" approach and algorithm has been developed to study the evolution of any organism from the protein perspective using datasets commonly employed in proteomics. Sets of peptide masses (known as mass maps), rather than gene or protein sequence data, can be employed to construct phylogenetic trees and these "mass trees" used to trace and study the evolutionary history of organisms from which the proteins are derived. Furthermore, mass differences associated with single amino acid mutations can be charted and interrogated across the tree in terms of their frequency, position and lineage. Understanding the mechanisms by which antiviral drug resistance mutations manifest and are compensated for remains elusive despite its importance to improving responses to the influenza virus. The phylonumerics approach is shown to be able to investigate the emergence of antiviral resistance mutations in influenza neuraminidase. Frequent ancestral and descendant mutations to antiviral resistance mutations are identified in N2 neuraminidase. The majority occur in the head region around the active site and drive hydrophilicity changes, primarily through the incorporation or loss of hydroxyl groups. The mass tree approach allows the evolution of influenza to be viewed from a global protein perspective and putative epistatic and compensatory mutations, remote in their sequence and structure, to be proposed.

Mareike Fischer, Greifswald University

Classes of treebased networks

(Joint work with Kristine Wicke)

Recently, so-called treebased phylogenetic networks have gained considerable interest in the literature, where a treebased network is a network that can be constructed from a phylogenetic tree, called the base tree, by adding additional edges. In my talk, I will provide some sufficient criteria for treebasedness by reducing phylogenetic networks to related graph structures. While it is generally known that deciding whether a network is treebased is NP-complete, one of these criteria, namely edgebasedness, can be verified in polynomial time. Besides these edgebased networks, I will introduce some more classes of treebased networks and analyze their relationships.

Matthew Hahn, Indiana University

Cases in which maximum likelihood will be positively misleading

Genome-scale sequencing has been of great benefit in recovering species trees, but has not provided final answers. Despite the rapid accumulation of molecular sequences, resolving short and deep branches of the tree of life has remained a challenge, and has prompted the development of new strategies that can make the best use of available data. One such strategy — the concatenation of gene alignments — can be successful when coupled with many tree estimation methods, but has also been shown to fail when there are high levels of incomplete lineage sorting. Here, we focus on the failure of likelihood-based methods in retrieving a rooted, asymmetric four-taxon species tree from concatenated data when the species tree is in or near the anomaly zone — a region of parameter space where the most common gene tree does not match the species tree because of incomplete lineage sorting. First, we use coalescent theory to prove that most informative sites will support the species tree in the anomaly zone, and that as a consequence maximum-parsimony succeeds in recovering the species tree from concatenated data. We further show that maximum-likelihood tree estimation from concatenated data fails both inside and outside the anomaly zone, and that this failure cannot be easily predicted from the topology of the most common gene tree. We show that likelihood-based methods often fail in a region partially overlapping the anomaly zone, likely because of the lower relative cost of substitutions on discordant gene tree branches that are absent from the species tree. Our results confirm and extend previous reports on the performance of these methods applied to concatenated data from a rooted, asymmetric four-taxon species tree, and highlight avenues for future work improving the performance of methods aimed at recovering species tree.

Jotun Hein, University of Oxford

Protein Structure Evolution

(Joint work with Michael Golden, Thomas Hamelryck, Kanti Mardia, Eduardo Portugues, Michael Sørensen.)

Recently described stochastic models of protein evolution have demonstrated that the inclusion of structural information in addition to amino acid sequence leads to a more reliable estimation of evolutionary parameters. We present a generative, evolutionary model of protein structure and sequence that is valid on a local length scale. The model concerns the local dependencies between sequence and structure evolution in a pair of homologous proteins. The evolutionary trajectory between the two structures in the protein pair is treated as a random walk in dihedral angle space, which is modelled using a novel angular diffusion process on the two-dimensional torus. Coupling sequence and structure evolution in our model allows for modelling both "smooth" conformational changes and "catastrophic" conformational jumps, conditioned on the amino acid changes. The model has interpretable parameters and is comparatively more realistic than previous stochastic models, providing new insights into the relationship between sequence and structure evolution. For example, using the trained model we were able to identify an apparent sequence-structure evolutionary motif present in a large number of homologous protein pairs. The generative nature of our model enables us to evaluate its validity and its ability to simulate aspects of protein evolution conditioned on an amino acid sequence, a related amino acid sequence, a related structure or any combination thereof.

John Hewson, UTAS

Graded rings of Markov invariants

(Joint work with Jeremy Sumner, University of Tasmania)

By considering binary Markov models on phylogenetic trees and their associated probability distributions, a group action on a tensor product space is naturally identified together with the associated graded ring of invariant functions. In the case of three taxa and below, these invariants can be completely accounted for; however at four taxa and above the situation becomes much more complicated. This talk will review the methodology we have applied to categorise these functions on higher number of taxa and give results obtained thus far.

Barbara Holland, University of Tasmania

Are Lie-Markov models more robust to taxon-sampling than models without the closure property?

(Joint work with Ned Goodman, Jeremy Sumner, and Michael Charleston, UTAS)

Lie-Markov (LM) models were introduced in Sumner et al (2012). They have the property of being closed under matrix multiplication. This means that if you take two rate matrices from the same model class and let one act for some time t_1 and then the other for some time t_2 then the overall process can also be described by a substitution matrix from the same class. This property seems particularly desirable when we consider that evolution is most likely a heterogeneous process that does not act the same way in all the different branches of a phylogenetic tree. Although the Lie Markov property is not sufficient to ensure that there is a single rate matrix that can describe an "average" process over the whole tree, it does at least ensure that over particular branches of the tree an average process can be found.

At a previous Phylomania meeting Gavin Huttley suggested that a good test of whether or not LM models provide benefits for accurate phylogenetic inference would be to test if they are more robust to differences in taxon sampling. We set up an experiment where for a variety of real data sets we created new alignments by deleting subsets of taxa at random. We then computed the average RF and SPR distance between all pairs of trees (restricted to the subset of taxa that overlap). Our hypothesis was that LM models would be more robust to taxon-sampling than non LM models. An important confounding variable is the overall quality of the model – somewhat reassuringly we found that models with good AIC scores were more robust than models with poor AIC scores.

Preliminary analysis of the data revealed no clear advantage for LM models over non-LM models but I hope to torture the data a bit more before November 21!

Greg Jordan, University of Tasmania

Why ancestral states matter (well to me anyway)

It's been a busy year and I've left this far too long, but this is a rough idea of what I will talk about. I work at the interface between evolutionary biology and palaeoenvironmental science. I'll explain how ancestral state reconstruction underpins almost everything I have to deal with in some form or another. I'll consider the concepts of implicit and explicit reconstructions, intrinsic versus extrinsic traits, and something about uncertainty (and the uncertainty in the uncertainty).

Rob Lanfear, Australian National University

New methods to calculate concordance factors for phylogenomic datasets

(Joint work with Bui Quang Minh and Matthew Hahn)

Measures of underlying variation are often useful in interpreting inferences made from a dataset. In phylogenetics, we are typically interested in inferences about particular branches on a phylogeny. Concordance factors provide useful measures of underlying variation, but there are some unresolved issues with calculating and interpreting them. We implement a simple approach to calculate gene concordance factors for phylogenomic datasets with lots of missing data. We also introduce a new 'site concordance factor', which provides an estimate of the proportion of informative sites in a dataset that support a particular branch in the tree.

Tim McInerney, National Centre for Indigenous Genomics, The Australian National University

Phylogenetically congruent haplotype blocks in the human genome identified by character-compatibility matrices.

Background: Haplotype blocks (haploblocks) are contiguous stretches of DNA whose shared pattern of inheritance is undisrupted by recombination. Each haploblock is an independent realisation of the stochastic evolutionary process. Evolutionary relationships of samples at sites within a haploblock can be explained by a single phylogenetic tree, but additional phylogenies are required when sites are in different haploblocks. Evolutionary relationships inferred through phylogenies allow researchers to to detect cryptic relatedness in 'unrelated' individuals, and utilise that information to detect disease associations with greater power than is possible by assuming independence of sites and samples [1]. Researchers can also investigate the evolutionary history of discrete regions of the genome or analyse the whole genome average [2] and identify outliers as putative regions for natural selection [3]. Identifying haploblocks is problematic. Linkage disequilibrium (LD) approaches lack specificity as boundaries may extend across recombination points, leading to haploblocks with multiple evolutionary histories. Four-gametes test lacks sensitivity towards homoplasy-induced noise, resulting in premature termination of haploblock boundaries. We applied a clique-based clustering algorithm [4; 5] to character-compatibility matrices [6] with the specificity to identify haploblocks with a single, congruent evolutionary history and sensitivity tolerant of homoplasy in a 1,000 Genomes Project population.
Results and Conclusions: Haploblock structure was largely concordant between character-compatibility and LD; 4% of haploblocks were identical, 82% of character-compatibility haploblocks were contained within a larger LD haploblock, and 7% vice-versa. On average, character-compatibility haploblocks were more phylogenetically congruent (80%) compared to LD haploblocks (55%). Our approach appears to reliably identify haploblock structures that are more phylogenetically congruent than LD-based approaches, consistent with our theory that evolutionary relationships within LD haploblocks are explained by multiple phylogenies. We posit that our approach to haploblock identification is superior when downstream analyses assume, or benefit from, phylogenetic congruence of sites in the region(s) under investigation.
References
1. Thompson and Fardo (2016) Comparing performance of non-tree-based and tree-based association mapping methods. BMC Proceedings 10(Supplement 7): 405-410.
2. Heled and Drummond (2008) Bayesian inference of population size history from multiple loci. BMC Evolutionary Biology 8(1): 1-15.
3. Voight, et al. (2006) A Map of Recent Positive Selection in the Human Genome. PLoS Biology 4(3): e72.
4. Kim, et al. (2017) A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. Bioinformatics 34(3): 388-397.
5. Yoo, et al. (2015) Clique-Based Clustering of Correlated SNPs in a Gene Can Improve Performance of Gene-Based Multi-Bin Linear Combination Test. BioMed Research International 2015: 11.
6. Jakobsen and Easteal (1996) A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Bioinformatics 12(4): 291-295.

Rohan Mehta, Stanford University

The probability of monophyly of a sample of gene lineages on a species tree.
(Joint work with David Bryant, University of Otago, Dunedin, NZ; Noah Rosenberg, Stanford University, USA.)

Monophyletic groups — groups that consist of all the descendants of a common ancestor — are important objects of study in fields that concern genealogy or population history, including phylogeography, species delimitation, and phylogenetics. Recent work has investigated mathematical aspects of monophyletic groups under coalescent models, generating predictions about the properties of monophyly in relation to the genealogy of a set of populations or species. We derive the probability that a group of individuals within a larger genealogy is monophyletic conditional on a species tree of any size and shape. We also extend two-species computations to compute the probability of reciprocal monophyly for samples from three or four species. We analyze the effects of species tree height, branch lengths, and sample size on monophyly probabilities. We also use an example dataset from the study of maize domestication to demonstrate that the theoretical probabilities we obtain are comparable to those found in computations from data. Finally, we present software for computing monophyly probabilities under the assumptions of coalescent models.

Jonathan Mitchell, University of Alaska Fairbanks

Testing n-Taxon Species Trees with the Multispecies Coalescent Model
(Joint work with Elizabeth Allman, John Rhodes, both University of Alaska Fairbanks)

Incomplete lineage sorting, where gene tree topologies can differ from species tree topologies, can be modeled by the multispecies coalescent model. Here we describe a test for an n-taxon species tree, with gene trees expected to arise in specific frequencies under the multispecies coalescent model. A substantial departure from these frequencies can be interpreted as evidence to reject the species tree and/or the multispecies coalescent model. A species tree may be rejected in favour of a network that models more complex biological processes such as hybridisation.

Yukihiro Murakami, Delft University of Technology

Reconstructing Tree-Child networks from Reticulate Edge-Deleted Subnetworks
(Joint work with Leo van Iersel, Remie Janssen, Mark Jones (TU Delft); Vincent Moulton (University of East Anglia))

Network reconstruction lies at the heart of phylogenetic research. Two well-studied classes of phylogenetic networks include tree-child networks and level k networks. In a tree-child network, every non-leaf node has a child that is a tree node or a leaf. In a level k network, the maximum number of reticulations contained in a biconnected component is k. Here, we show that level k tree-child networks are encoded by their reticulate edge-deleted subnetworks, which are subnetworks obtained by deleting a single reticulation edge, if $k \geq 2$.
Following this, we provide a polynomial-time algorithm for uniquely reconstructing such networks from their reticulate-edge-deleted subnetworks. Moreover, we show that this can even be done when considering subnetworks obtained by deleting one reticulation edge from each biconnected component with k reticulations.

Suha Naser-Khdour, Australian National University

The Prevalence and Impact of Model Violations in Phylogenetics

Most phylogenetic methods rely on mathematical substitution models that approximate evolutionary process. A common assumption in these models is that the sequences have evolved under stationary, reversible and homogeneous (SRH) conditions. Although such assumptions are often criticized, the extent of SRH violations and their effects on phylogenetic inference remains largely unknown. In order to address this gap, we used the matched-pairs test of symmetry to assess the scale and impact of SRH model violations in empirical data. In particular, I will discuss our results from 3,572 partitions and 35 published phylogenomic datasets.

Cassius Manuel Pérez, Medical University of Vienna / CIBIV

Sequences of any alphabet and with any set of taboos

(Joint work with Arndt von Haeseler and Michael Charleston)

It is regularly assumed that each of the sites in a DNA sequence evolves independently. However, it is known that some dependencies exist, for example the ones induced by restriction enzymes in bacteria. These enzymes recognize a specific sequence of nucleotides and produce a double-stranded cut in the DNA. Therefore, the recognition sequence is, let us say, a taboo: if a random mutation made this sequence appear at some site of the host DNA sequence, it would be recognized as foreign DNA and eliminated (assuming that methylation is not the case).
We give some properties of the set of sequences not containing some taboos, describe how this can affect the error of the estimated evolution time and explain how to carry out a Monte-Carlo-simulation.

Julia Shore, University of Tasmania

Good old reliable linearity: a new way to build codon models.

(Joint work with Jeremy Sumner, Barbara Holland, Peter Wills, Kay Nieselt)

Since it was found that there is no practical solution for a multiplicatively closed codon model (Shore *et al.* 2018) which would have eradicated the misestimation of the synonymous/non-synonymous rate ratio (ω) caused by lack of multiplicative closure, it is now of interest to study linear codon models as it was observed by Kaine (2011) that when a Markov model is linear, the errors in parameter estimation are smaller. We explore two ways of making an existing set of matrices linear: one is of finding the smallest linear space which contains a matrix set and the other is changing each non-linear constraint of a matrix set to a linear one. In the context of codon models in particular, this process brings about the interesting question of how the ω parameter should be interpreted in a linear model and what values it now makes sense for it to take.

(i) B. T. Kaine. The effect of closure in phylogenetics. Honours Thesis, University of Tasmania 2011.
(ii) Julia A. Shore, Jeremy G. Sumner, and Barbara R. Holland. Closed codon models: just a hopeless dream? arXiv:1804.11249, 2018.

Marnus Stoltz, University of Otago

Inferring phylogenies along trees using backward diffusions

(Joint work with David Bryant)

In this talk we describe an algorithm for efficiently computing the likelihood of a species tree from unlinked binary markers or allele frequency data. The model assumptions are similar to those implemented in SNAPP however, unlike SNAPP, the method can handle hundreds or even thousands of individuals. Our approach is based on a diffusion approximation of gene dynamics. However we work directly with backwards processes to compute the probability of every marker individually, bypassing the need to compute the entire joint allele frequency across all species. We point out some of the challenges encountered along the way such as boundary conditions for the backwards diffusion to ensure uniqueness and existence, computational bottlenecks and parameter mappings between models.

Jeremy Sumner, University of Tasmania

Markov association schemes

(Joint work with Julia Shore)

I will discuss what (we hope!) is a compelling example of the mathematics of phylogenetics leading to a new (and interesting!) algebraic structure.
The motivation for this work comes from a simple model of aminoacyl-tRNA synthetase (aaRS) evolution devised by Julia Shore (UTAS) and Peter Wills (U Auckland).
Starting with a proposed rooted tree describing the specialization of aaRS through evolution of the genetic code, their model produces a space of Markov rate matrices that form (by a minor miracle!) a commutative algebra under matrix multiplication. Reasonably, we like to refer to each of these as a ‘tree-algebra’.
A natural mathematical question then arises: is the existence of these tree algebras purely a happy accident or can it be understood in more abstract terms as a special case of an already existing algebraic structure? At first blush, it appeared the resolution would follow by identifying the tree-algebras as an instance of what algebraists refer to as ‘association schemes’. However this turned out not to be the case, and further study has revealed an intriguing resolution: both the tree-algebras and association schemes occur as a special case of a more general (and novel!) algebraic structure which we have, reasonably, coined ‘Markov association schemes’.
I will recall the sequence of ideas that led to these investigations, confer their resolution, and describe our first steps in attempting to characterize the class of all Markov association schemes — an inquiry that goes well beyond our initial humble motivations coming from evolution of the genetic code!

Janan Sykes, University of Tasmania

Comparing Protein Structure Alignment Methodologies
(Joint work with Barbara Holland and Michael Charleston, both UTAS)

Most attempts to understand the protein universe involve sorting proteins by structural similarity. However, there are inconsistencies between attempts and manual curation plays a large role. Finding an extremely effective and reliable method of protein structure alignment and comparison may solve this problem. Dozens of different methodologies for this task have been put forward, using techniques such as 'contact map comparison', 'information compression analysis', 'direct superimposition and distance calculation', and 'protein structure alphabet analysis'. Eighteen of these techniques were compared in their ability to distinguish pairs of proteins known to share differing levels of structural similarity and to cluster proteins from several different folds into their appropriate groups. SP-AlignNS was found to be extremely effective in both cases.
Methods (where possible) were split into the alignment method and the score used to assess similarity. It was intended that this would allow independent assessment of algorithmic performance. Surprisingly, we found that some hybrids of mismatched scores and alignment methods performed better than the originals.

Venta Terauds, University of Tasmania

Are we related? Detecting an evolutionary signal between pairs of circular genomes
(Joint work with Jeremy Sumner, University of Tasmania)

The calculation of evolutionary distance via models of genome rearrangement has an inherent factorial complexity. Various algorithms and estimators have been used to address this, however many of these set quite specific conditions for the underlying model.
In recent work, we showed that a technique for calculating evolutionary distance as a maximum likelihood estimate (MLE) of time elapsed may be applied to models with any selection of rearrangements and probabilities thereof. Further, unlike primitive estimators such as minimal reversal distance, the MLE distance does not necessarily exist for every pair of genomes under a given model, meaning that we can distinguish the cases in which the genomes are not related. Whilst the factorial complexity of the problem remains, limiting the 'exact' calculation of MLEs to genomes with a relatively small number of regions (for the moment, at least), we show that the **existence** or otherwise of an MLE for a given pair of genomes may be predicted quite easily, without the need to calculate the entire likelihood function.

Kristina Wicke, University of Greifswald

On non-binary treebased networks
(Joint work with Mareike Fischer, Michelle Galla, Lina Herbst & Yangjing Long)

Phylogenetic networks are a generalization of phylogenetic trees allowing for the representation of non-treelike evolutionary events such as hybridization.
A special class of phylogenetic networks are so-called *treebased* networks, which can be constructed from a phylogenetic tree by adding additional edges.
In this talk, we will mainly be concerned with non-binary treebased networks. However, we first revisit some established results for binary treebased networks, before extending these results to non-binary ones. Here, we focus in particular on the notion of so-called *universal treebased networks*.

Benjamin Wilson, Lateral GmbH

Embedding trees in hyperbolic space

(Joint work with Matthias Leimeister, Lateral GmbH)

The curvature of hyperbolic space makes it a natural place to embed trees and tree-like graphs. For instance, binary trees of any depth can be drawn in the hyperbolic plane with arbitrarily low distortion, i.e., such that the spacial distance between embedded vertices approximates their distance in the graph. This talk will describe a methodology for performing gradient descent in hyperbolic space and survey recent results in machine learning in hyperbolic space that concern the embedding of trees and the related problem of the determination of an embedding from pair-wise distances ("multi-dimensional scaling").

Michael Woodhams, University of Tasmania

Closure in Codon Models

Closure is a property of rate matrices in a model, without which it is mathematically inconsistent to use the model when model parameters change with time. Closed DNA models are the Lie Markov models, which have been extensively discussed at earlier Phylomanias.
Codon models calculate transition probabilities between different codons, accounting for the genetic code and DNA transition probabilities. Due to the irregularity of the genetic code, we cannot make a closed codon model without very many parameters. However, Muse-Gaut codon models are built on top of a DNA model, and we can chose a closed DNA model.
We use numerical simulations to study whether Muse-Gaut models based on closed or non-closed DNA models are more accurate in the presence of time varying parameters.

List of Attendees

Brown, Matilda . . . . . matilda.brown@utas.edu.au
Bryant, David . . . . . david.bryant@otago.ac.nz
Bui, Minh . . . . . m.bui@anu.edu.au
Burden, Conrad . . . . . conrad.burden@anu.edu.au
Burridge, Chris . . . . . chris.burridge@utas.edu.au
Chan, Yao-ban . . . . . yaoban@unimelb.edu.au
Charleston, Michael . . . . . michael.charleston@utas.edu.au
Chernomor, Olga . . . . . o.chernomor@gmail.com
Cherryh, Caitlin . . . . . caitlin.cherryh@anu.edu.au
Darling, Aaron . . . . . aaron.darling@uts.edu.au
Diao, Jiahao . . . . . jiahao.diao@utas.edu.au
Downard, Kevin . . . . . kevin.downard@unsw.edu.au
Feng, Qian . . . . . fengq2@student.unimelb.edu.au
Fischer, Mareike . . . . . email@mareikefischer.de
Hahn, Matthew . . . . . mwh@indiana.edu
Hein, Jotun . . . . . hein@stats.ox.ac.uk
Hewson, John . . . . . timothy.hewson@utas.edu.au
Holland, Barbara . . . . . barbara.holland@utas.edu.au
Jordan, Greg . . . . . greg.jordan@utas.edu.au
Lanfear, Rob . . . . . rob.lanfear@anu.edu.au
McComish, Bennet . . . . . bennet.mccomish@utas.edu.au
McInerney, Tim . . . . . tim.mcinerney@anu.edu.au
Mehta, Rohan . . . . . rsmehta@stanford.edu
Mitchell, Jonathan . . . . . jonathanmitchell88@gmail.com
Murakami, Yukihiro . . . . . y.murakami@tudelft.nl
Naser-Khdour, Suha . . . . . suha.naser@anu.edu.au
Neeman, Teresa . . . . . terry.neeman@gmail.com
O'Reilly, Małgorzata . . . . . malgorzata.oreilly@utas.edu.au
Pérez, Cassius . . . . . cassiusmanuelperez@gmail.com
Rorlach, Ben . . . . . adam.rohrlach@adelaide.edu.au
Shore, Julia . . . . . julia.shore@utas.edu.au
Stevenson, Joshua . . . . . joshua.stevenson@utas.edu.au
Stoltz, Marnus . . . . . stoltzstep@gmail.com
Sumner, Jeremy . . . . . jsumner@utas.edu.au
Sykes, Janan . . . . . janan.sykes@utas.edu.au
Terauds, Venta . . . . . venta.terauds@utas.edu.au
Tupper, Paul . . . . . paulfredtupper@gmail.com
Wicke, Kristina . . . . . kristina.wicke@web.de
Wilson, Benjamin . . . . . benjamin@lateral.io
Woodhams, Michael . . . . . michael.woodhams@utas.edu.au

Index