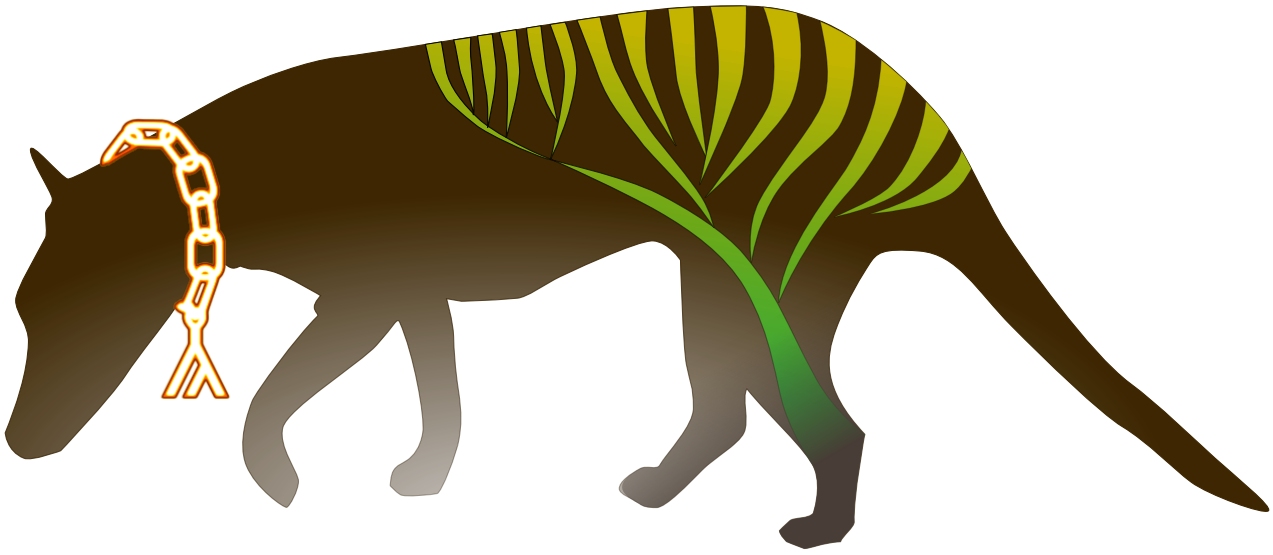
Phylömania 2019
It Göes to Eleven

Maths & Physics Building,
Sandy Bay Campus,
University of Tasmania, Hobart
November 20-22
2019



Sponsors

This year Phylomania is actively supported by AMSI — Australian Mathematical Sciences Institute, AustMS — the Australian Mathematical Society, and ANZIAM — Australia and New Zealand Industrial and Applied Mathematics. We are also extremely grateful to the support of Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers. These generous funding commitments have allowed us to invite several excellent national and international speakers to the conference and workshop. Thank you!

These **major sponsors** have allowed us to run on Wednesday 20th November,

The Workshop on Stochastic and Algebraic Models for Genome Evolution

Keynote Speakers

Professor Nadia El-Mabrouk

(Université de Montréal; supported by AMSI/AustMS/ANZIAM)

Nadia El-Mabrouk is Full Professor at the Computer Science Department of the University of Montreal. She holds a Ph.D. in theoretical Computer Science from the University Paris VII, obtained in 1996. She is member of the Centre de Recherche Mathematiques and the Robert Cedergren Centre for Bioinformatics and Genomics. Her expertise is in Computational Biology and her research focuses on developing algorithmic and mathematical methods for comparative genomics. She is regularly involved in the program committee of bioinformatics and computational biology conferences such as RECOMB, RECOMB-CG, ISMB, ECCB and WABI. She has published over 70 works including journal articles, refereed conference papers and several book chapters.

Distinguished Professor Seth Sullivant

(North Carolina State University; supported by AMSI/AustMS/ANZIAM)

Seth Sullivant received his PhD in 2005 from the University of California, Berkeley. After a Junior Fellowship in Harvard's Society of Fellows, he joined the department of mathematics at North Carolina State University in 2008 as an assistant professor. He was promoted to full professor in 2014 and distinguished professor in 2018. Sullivant's work has been honoured with a Packard Foundation Fellowship and an NSF CAREER award and he was selected as a Fellow of the American Mathematical Society. He helped to found the SIAM activity group in Algebraic Geometry where he has served as both secretary and chair. Sullivant's current research interests include algebraic statistics, mathematical phylogenetics, applied algebraic geometry, and combinatorics. He has published 55 papers and 2 books in these areas.

Dr Sophie Hautphenne

(University of Melbourne; supported by ACEMS)

Sophie Hautphenne is a Senior Lecturer in Applied Probability at the School of Mathematics and Statistics at The University of Melbourne, a Scientist at the Chair of Statistics in the Swiss Federal Institute of Technology, and an Associate Investigator at the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). Since April 2015, she is holding an ARC Discovery Early Career Researcher Award (DECRA) at the University of Melbourne. Sophie obtained her PhD in Mathematics from the Universite libre de Bruxelles in October 2009. Her fields of research are applied probability and stochastic modelling with a particular focus on branching processes, matrix analytic methods and epidemic models. Sophie is particularly interested in biological and ecological applications.

Phylomania 2019

	Wednesday	Thursday	Friday
8:50am	**Welcome**	Housekeeping	Housekeeping
	Gene (Re)Arrangements	**Networks (sort of)**	**Applied Phylogenetics**
9:00	⋆ ***Nadia El-Mabrouk*** ⋆ History of Gene Order	*Diane Donovan* Genotype/Phenotype Pipeline	*Lars Jermiin* Hawaiian Hoary Bats
9:20		*Yao-Ban Chan* Reconciling Gene Networks with Species Trees	
9:40	*Chad Clark* Algebraic Reversal-Deletion model	*Qiuyi Li* HIDTL	*Chris Burridge* Does the Virus Cross the Road?
10:00	*Venta Terauds* Circular Genome Distances		*Nick Fountain-Jones* Pathogen Phylogeography in Secretive Carnivores
10:20		*Matilda Brown* Leaf Surface as a Spatial Network	*Bennet McComish* Genome Selection in NT
10:40	MORNING TEA AND COFFEE		
	Algebraic Methods	**Combinatorics(?)**	**Sampling**
11:10	⋆ ***Seth Sullivant*** ⋆ Identifiability using Algebraic Matroids	*Arndt von Haeseler* Taboo Sequences	*Conrad Burden* Neutral Multi-Allele Wright Fisher sampling
11:30			
11:50	*Julia Shore* Tree Algebras!	*Mike Steel* Phylogenetic Diversity Indices	*Reed Cartwright* Weighted Reservoir Sampling
12:10pm	LUNCH		
	Modelling Gene Evolution	**Here Be Dragons**	**Weird & Wonderful**
1:40	⋆ ***Sophie Hautphenne*** ⋆ Markovian Binary Trees	*Joshua Stevenson* Rank-based Approaches for Phylogenetics	*Qian Feng* Recent recombinant DBLa in *Plasmodium falciparum*
2:00		*Michael Charleston* Exploring Split Space	*Allen Rodrigo* Heteroduplex Mobility Assay
2:20	*Albert Seowongsono* Tree Shape under Phase-Type Speciation Times	*Jeremy Sumner* Model-Specific Markov Embedding	*Katherine Turner* Topological Data Analysis of Phylogenetic Trees?
2:40	*Amanda Wilson* Retention of Gene Copies after Whole Genome Duplication	*Timothy (John) Hewson* Graded Rings of Markov Invariants	
3:00	AFTERNOON TEA AND COFFEE		
	Within-gene Coevolution		
3:30	*David Liberles* Amino Acid Substitutions to compare Compensatory Processes	*Kevin Downard* Darwin's Tree of Life is Numbered	*Mathieu Fourment* Probabilistic programming and fast Bayesian Inference
3:50		*Michael Hendriksen* A Cluster-Similarity Metric on Trees	*Stephen Crotty* Why did the model frighten the biologist?
4:10	*Tristan Stark* Amino Acid Substitution with Complete Linkage	*Andrew Francis* Space of Tree-based Phylogenetic Networks	*Benjamin Wilson* Pairwise Distances under a Weakened Four Point Condition
4:30	*Jiahao Diao* Evolution of Families of Gene Duplicates		
	PUB O'CLOCK		**Prizes and Closing**

Schedule

The conference is held at the Sandy Bay campus of University of Tasmania (UTAS).

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.

Every Day

- Welcome and Housekeeping at `8:50am`
- First session of talks from `9:00am` to `10:40am`
- Morning tea and coffee from `10:40am` to `11:10am`
- Second session of talks from `11:10am` to `12:10pm`
- Lunch from `12:10pm` to `1:40pm`
- Third session of talks from `1:40pm` to `3pm`
- Afternoon tea and coffee from `3pm` to `3:30pm`
- Fourth session of talks from `3:30pm` to `4:50pm`-ish.
- "Pub O'Clock" from then onwards

"Pub O'Clock" has limited options handy to campus, but our favourite drinking hole is Preachers on Knopwood Street in Battery Point, which is about a 30 minute walk. Someone will be able to help you find it!

Conference Dinner

This year the conference dinner is included in your registration (and invited speakers are covered), but if you are bringing a partner we will ask you to pay for them at $35 per person. The restaurant is BYO and we will attempt to secure a no-cost corkage for you.

The dinner is 7pm at Kathmandu Cuisine on the corner of Hampden Road and Francis Street, in Battery Point. There will be a walking bus and transport arranged as needed from Sandy Bay Campus (possibly via the pub).

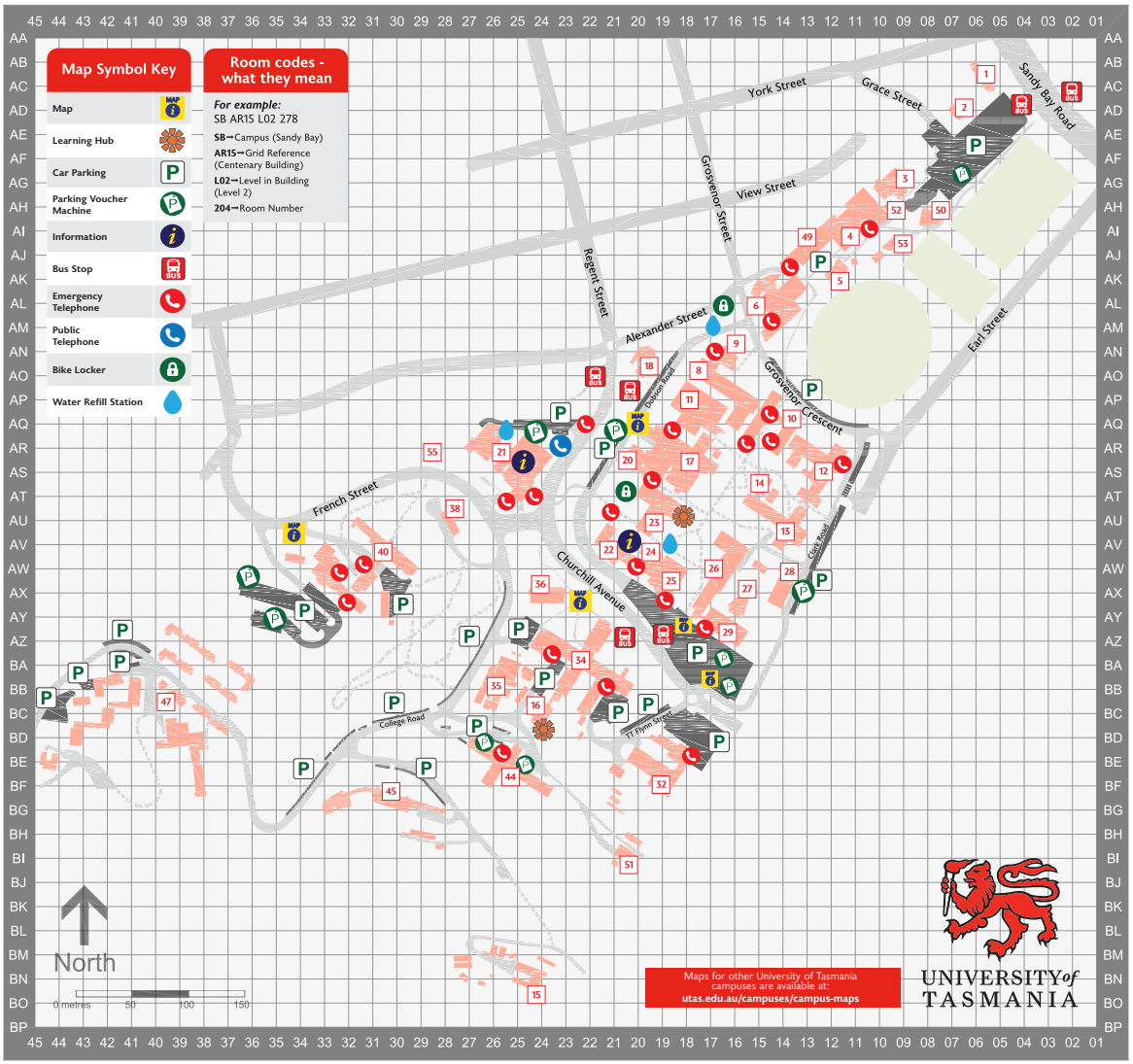
The Conference Outing on Saturday

- 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.

The conference venue is Building #13 on the map below:

Sandy Bay Campus

	Grid Ref.	No.
301 Sandy Bay Road	AC06	1
6 Grace Street	AD07	2
Accommodation Services	BB39	47
Administration Building	AW21	22
Agricultural Science	BA22	34
Arts Lecture Theatre	AX16	27
Centenary Building	AR15	10
Central Science Laboratory	AR19	17
Chemistry	AR19	17
Child Care Centre (Lady Gowrie)	AG10	3
Child Care Cottage	AJ10	53
Christ College	BB39	47
Classics Museum, John Elliott	AX19	25
CODES - Centre for Ore Deposit and Earth Sciences	AS13	12
Co-op Bookshop	AS25	21
Community Health Clinic	AH10	52
Computing and Information Technology	AR15	10
Corporate Services Building	BE20	32
Cricket Pavilion	AK12	5
CSIRO	BF31	45
Earth Sciences	AS13	12
Education	AW31	40
Engineering	AP16	8
Engineering Workshop	AQ18	11
English Language Centre	AW31	40
Financial Services	BE20	32
Geography and Environment	AS13	12
Graduate Research Office	AC06	1
Herbarium, Tasmanian	AX24	36
Horticultural Research Centre	BN25	15
Human Resources	BE20	32
Humanities	AZ16	29
Hytten Hall	AW31	40
Information Technology Services (ITS)	BE20	32
John Fisher College	BB39	47
Law	AL14	6
Lazenby's Cafe	AX19	25
Library, Morris Miller	AU19	23
Life Sciences Building	BA22	34
Life Sciences Glasshouse	BB26	35
Mathematics	AT15	14
National Tertiary Education Union (Tas. Div.)	AS25	21
Newsagency and post office	AS25	21
Office of Research Services	AC06	1
Old Medical Sciences Building	BE26	44
Pharmacy	AR20	20
Physics	AU14	13
Plant Science	BA22	34
Psychology	AX17	26
Psychology Research Centre	AX14	28
Refectory (The Ref)	AS25	21
Research House	AU28	38
Research Office	AC06	1
Riawunna	AX17	26
Rugby Club	AI08	50
Sciences and Engineering, College of	AR15	10
Security	AD07	2
Social Sciences	AX17	26
Source Wholefoods	AR29	55
Steps Building	BI21	51
Student Administration	AU19	23
Surveying and Spatial Sciences	AN16	9
Tasmanian Institute of Agriculture (TIA)	BC24	16
Tasmanian Institute of Learning and Teaching (TILT)	AS25	21
Tasmanian School of Business and Economics (TSBE)	AR15	10
Tasmania University Union (TUU)	AS25	21
Trade Table Cafe	AR15	10
TUU Building	AS25	21
Unigym	AI11	4
University Apartments	BB39	47
University Centre	AX19	25
University Club	AO20	18
University College	AI13	49
University Foundation	AO20	18
Zoology	BA22	34



This map was last updated FEB 2018 | Infrastructure Services & Development | **utas.edu.au/isd**

Abstracts

Matilda Brown, School of Natural Sciences, University of Tasmania

Mapping the leaf surface as a spatial network (Short talk)

(Joint work with Barbara Holland, Greg Jordan)

The leaf epidermis is a mosaic of cells which forms the interface between the plant and its environment. The outer surface of these cells is preserved in minute detail by the cuticle, a waxy layer secreted by the epidermis. The cuticle is readily fossilised and can be used to estimate past vegetation and climate conditions. The links between cell size, shape and environment have previously been studied, but the cuticle also contains untapped spatial information. By representing the leaf as a spatial network graph, we can automate the extraction of cell alignment, arrangement and patterning traits. We demonstrate that these features can be used to identify different species of *Podocarpus*, and may have ecological significance.

Conrad Burden, Mathematical Sciences Institute, Australian National University

Sampling distributions from a neutral multi-allele Wright-Fisher diffusion with applications to estimating mutation rate matrices (Long talk)

The neutral Wright-Fisher model is a population genetics model in which each individual in the population inherits an allele type from a random parent in the previous generation. Neutral mutations between allele types are built into the model via a substitution rate matrix whose off-diagonal elements u_{ij} give the probability of a mutation from allele type i to allele type j in one generation. A convenient formalism is the diffusion limit in which the time between generations becomes infinitesimal, and the effective population size N becomes infinite in such a way that the the scaled mutation rate per continuum time, $Q_{ij} = Nu_{ij}$, remains finite. An obvious application is substitution of nucleotides at a neutral genomic site, in which case the allele types are $\{A, C, G, T\}$ and the rate matrix Q is 4×4 with small off diagonal elements which are typically $< 10^{-3}$. Under certain assumptions the rate matrix can be matched with the whole-population substitution matrix in phylogenetics.
We have addressed the following problem: Suppose a sample of given finite size is taken from a population which has evolved via a neutral Wright-Fisher diffusion process. What is the distribution of allele types across the sample? This distribution is the site-frequency spectrum of independent neutrally evolving sites, where the sites are chosen to be sufficiently separated so as to have independent coalescent trees due to recombination. We have solved for stationary [1, 3] and the time-dependent [4] distributions given initial conditions to first order in small mutations rates for an otherwise arbitrary Q. Methods used include direct solution of the forward-Kolmogorov equation, the backward generator, and the coalescent. We have also developed methods for inferring maximum likelihood estimates of rate matrices from site frequency data [5].
[1] Burden, C.J. and Tang, Y. 2016, Theor. Pop. Biol., vol. 112, pp. 22-32.
[2] Burden, C.J. and Tang, Y. 2017, Theor. Pop. Biol., vol. 113, pp. 23-33.
[3] Burden, C.J. and Griffiths, R.C. 2019, J. Math. Biol., vol 78, pp 1211-122.
[4] Burden, C.J. and Griffiths, R.C. 2019, J. Math. Biol., to appear.
[5] Vogl., C., Mikula, L. and Burden, C.J., 2019, in preparation.

Chris Burridge, School of Natural Sciences, University of Tasmania

Does the virus cross the road? Viral phylogeographic patterns among urban bobcat populations (Short talk)

(Joint work with Chris Kozakiewicz, Scott Carver, Nicholas Fountain-Jones (University of Tasmania), plus a cast of thousands)

Urban development has major impacts on connectivity among wildlife populations and is thus likely an important factor shaping pathogen transmission in wildlife. However, most investigations of wildlife diseases in urban areas focus on prevalence and infection risk rather than potential effects of urbanisation on transmission. Feline immunodeficiency virus (FIV) is a directly-transmitted retrovirus that infects many felid species and is a model for studying pathogen transmission at landscape scales. We reconstructed phylogenetic relationships among FIVLru isolates sampled from five bobcat populations in coastal southern California that appear isolated due to major highways and dense urban development. We found strong FIVLru phylogeographic structure among three host populations northwest of Los Angeles, largely coincident with host genetic structure. In contrast, relatively little FIVLru phylogeographic structure existed among two genetically-distinct host populations southeast of Los Angeles. Rates of FIVLru transfer among populations did not vary significantly, suggesting that the lack of phylogenetic structure southeast of Los Angeles may be a product of incomplete lineage sorting rather than frequent contemporary transmission among populations. Divergence dates among FIVLru lineages in several cases reflected historical urban growth and construction of major highways. Our results indicate that major barriers to host gene flow can also act as barriers to pathogen spread, suggesting potentially reduced susceptibility of these populations to outbreaks of novel pathogens.

Reed Cartwright, Arizona State University

Weighted Reservoir Sampling is a Poisson Process. (Short talk)

Reservoir sampling is a technique for sampling from a stream of observations and is useful for situations when the total number of observations in a dataset is either unknown or expensive to compute. In genomics, it is particularly useful for subsampling sequence data from FASTQ or BAM files in a single pass. Classically, reservoir sampling constructs an unweighted sample without replacement; however, recent advances have described strategies for sampling with weights and with replacement. Here, I present a novel interpretation of reservoir sampling as the outcome of a Poisson process and derive optimal algorithms for sampling with both weights and replacement.

Yao-ban Chan, School of Mathematics and Statistics / Melbourne Integrative Genomics, The University of Melbourne

An efficient algorithm for reconciliation of a gene network and species tree (Short talk)

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 present an efficient algorithm to solve the general problem of reconciliation between a gene network and species tree. Built on previous theoretical results (which we presented last year), this algorithm is fixed-parameter tractable in the level of the network, while remaining a practical solution to this problem.

Michael Charleston, School of Natural Sciences, University of Tasmania

Exploring split space with sub-flattenings (Short talk)

Phylogenetic inference is hindered by the size of tree space: for instance there are $(2n-5)!! = (2n-5)(2n-7)...(3)(1)$ unrooted binary trees with n leaves. Searching such a space for an optimal tree is challenging, despite considerable progress in speeding up searches over the last few decades. Constructing trees, e.g., based on pair-wise distances, is often achieved with a bottom-up, leaf-to-root approach, which makes "easy" decisions early on (e.g., which taxa are sister species), and leaves harder decisions until later. Trees can be thought of as collections of compatible splits, since the branches of a tree split the taxon set (the leaves) into two non-empty parts; a set of splits is compatible if they can all correspond to branches of the same tree. There are an exponential number of splits, but their number is still dwarfed by the number of possible trees. Hence, it is worthwhile to explore desirable splits to effectively reduce the search space by breaking it into smaller problems, top-down. Such a divide-and-conquer search tactic is the way in which, for example, sorting numbers can be achieved in $O(n \log n)$ time. I will present some initial thoughts on how we might use the newly-emerging "subflattening" method (Sumner, 2017 Sumner, J.G. Bull Math Biol (2017) 79: 619. https://doi.org/10.1007/s11538-017-0249-6) to provide a score of potential splits, and attempt to break up the phylogenetic inference problem from the top down.

Chad Clark, Centre for Research in Mathematics and Data Science, Western Sydney University

An Algebraic Reversal-Deletion Model for Bacterial Genome Rearrangement (Short talk)

Reversals are a major contributor to variation among bacterial genomes, with studies suggesting that reversals involving small numbers of regions are more likely than larger reversals. While reversals of neighbouring regions, otherwise known as inversions, have been accounted for in both the signed and unsigned case there has yet to be a model which also accounts for the process of deletion without insertion. In this talk, an algebraic model of the inversion-deletion process using the symmetric inverse monoid will be discussed along with exact algorithms for reconstructing the most recent common ancestor of two genomes arising by inversions and deletions. This is joint work with Andrew Francis (Western Sydney University), James Mitchell (University of Saint Andrews) and Julius Jonušas (TU Wien).

Stephen Crotty, University of Adelaide

Why did the model of sequence evolution frighten the biologist? (Short talk)

As multiple sequence alignments continue to grow in size, it is not unreasonable to expect that pattern heterogeneity within the alignments grows also. As such, increasingy complex models are required in order to minimise model misspecification. As use of these models become widespread, it may be naive to assume that well established information-theoretic model selection approaches will maintain their validity in all situations. By way of example, we construct theoretical and simulation-based arguments that suggest that model selection using information criteria is strongly biased towards partition models at the expense of mixture models. We then propose a hybrid mixture/partition model and investigate its performance on an empirical alignment of Cassava Brown Streak Virus isolates.

Jiahao Diao, School of Natural Sciences, University of Tasmania

Level-dependent QBD models for the evolution of a family of gene duplicates (Short talk)

(Joint work with Tristan L. Stark, David A. Liberles, Małgorzata M. O'Reilly, Barbara R. Holland)

A gene family is a set of evolutionarily related genes formed by duplication. Genes within a gene family can perform a range of different but possibly overlapping functions. The process of duplication produces a gene that has identical functions to the gene it was duplicated from with subsequent divergence over time. In this paper, we explore different models for the ongoing evolution of a gene family.
First, we consider a detailed model with multi-dimensional state-space which consists of binary matrices where rows of a matrix correspond to genes, columns correspond to functions, and the ijth entries record whether or not gene i performs function j. The large state space of this model makes it unsuitable for numerical analysis, but by considering the behaviour of this detailed model we can test the suitability of two alternative models with more tractable state-spaces.
Next, we consider a quasi-birth-and-death process (QBD) with two-dimensional states (n, m). The state (n, m) records the number $n = 1, 2, \ldots$ of genes in the family, and the number $m = 0, 1, \ldots, n$ of so called *redundant* genes (which are permitted to be lost). We contrast this to a level-dependent QBD with three-dimensional states (n, m, k) that record additional information $k = 1, \ldots, K$ which affects the transition rates.
We show that the model with two-dimensional states (n, m) is insufficient for meaningful analysis, while the model with the three-dimensional states (n, m, k) is able to capture the qualitative behaviour of the detailed model. We illustrate the fit between the level-dependent QBD and the original, detailed model, with numerical examples.

Diane Donovan, School of Mathematics and Physics, The University of Queensland

Mathematical Approaches to the Genotype/Phenotype Pipeline (Short talk)

(Joint work with Bevan Thompson, The University of Queensland; Kevin Burrage, Queensland University of Technology; Pamela Burrage, Queensland University of Technology; Emine Sule Yazici, Koc University, Turkey)

The premise for this talk is to promote discussions on how mathematical networks can be utilised to study the relationships between genotypes and phenotypes. We address the issue: can we pose new questions within this realm and ask how network theory can be applied to help resolve these questions? In this talk I will review the basic concepts of network theory and how these ideas might be applied to study genotypes.
I will embed this theory in the context of problems that have arisen in discussions with members of the UQ and QUT nodes of the ARC Centre for Plant Success in Nature and Agriculture.
Then I will give a brief discussion of mathematical approaches which have been suggest to investigate these problems.
If time permits I will introduce a novel approach to identifying the significance of information embedded in a genetic network.

Kevin Downard, University of New South Wales, Sydney

Darwin's Tree of Life is Numbered. Resolving the Origins of Species by Mass (Short talk)

Synopsis: Here it is shown that mass trees can be successfully used to study relationships among a diverse set of species from across the animal kingdom.

Qian Feng, Melbourne Integrative Genomics, University of Melbourne

A new method for identifying recent recombinant DBLa sequences in the Malaria Parasite *Plasmodium falciparum* (Short talk)

(Joint work with Yaoban Chan, Heejung Shim. All authors are in Melbourne Integrative Genomics, University of Melbourne)

Background: The *var* genes of the malaria parasite *Plasmodium falciparum* encode the PfEMP1 antigen, which controls the ability of the parasite to evade the human immune response system. These genes are hyper-diverse, principally due to recombination between the many (60) genes per genome. The study of these genes is thus one core problem in current malaria research, with implications for future malaria interventions.
The evolution of *var* genes can be studied through a conserved part of one of their domains called DBLa tags. A first step to inferring the evolution of these genes is to identify which tags are recent recombinants (and their breakpoints), and which tags are the ancestors of these recombinants. While many methods have been developed to identify recombinants and breakpoints, they mostly work on aligned sequences, while the massive number of sequenced DBLa tags also requires an efficient method to detect recombinants.
Results: Here, we propose a distance-based procedure to detect recombinants in a large set of unaligned sequences. This procedure identifies breakpoints and ancestor/descendant triples using a jumping hidden Markov model, then detects recombinants based on distances calculated from a partial alignment. A statistical measure of support is also calculated via bootstrapping. We explore the accuracy of this method through extensive simulations. These results demonstrate that this novel approach enjoys high accuracy over a wide range of biologically realistic scenarios. We then apply this method to investigate a large dataset of DBLa tags from a high-transmission area of Ghana

Nick Fountain-Jones, School of Natural Sciences, University of Tasmania

Pathogen phylogeography in secretive carnivores (Short talk)

(Joint work with Simona Kraberger (Colorado State University), Roderick Gagne (Colorado State University), Daryl R. Trumbo (Colorado State University), Patricia Salerno (Colorado State University), W. Chris Funk (Colorado State University), Kevin Crooks (Colorado State University), Roman Biek (University of Glasgow), Mathew Alldredge (Colorado Parks and Wildlife), Ken Logan (Colorado Parks and Wildlife), Guy Baele (KU Leuven), Simon Dellicour (KU Leuven, University of Brussels), Holly B Ernest (University of Wyoming), Sue VandeWoude (Colorado State University), Scott Carver (University of Tasmania) and Meggan E. Craft (University of Minnesota))

Urban expansion can fundamentally alter wildlife movement and gene flow, but how urbanization alters pathogen phylogeography is poorly understood. We combine host and viral genomic data with landscape variables to examine the context of viral spread in puma from two contrasting regions with remarkably high resolution; one bounded by the wildland urban interface (WUI) and one rural with minimal anthropogenic development. We found landscape variables and host genomics only explained significant amounts of variation of FIV spread in the WUI population. The most important predictors of viral phylogeography also differed; host spatial proximity, host relatedness, and mountain ranges played a role in FIV spread in the WUI puma population, whereas unpaved roads were more important in the unbounded population. Here we show that anthropogenic landscape change can alter pathogen phylogeography, providing a more nuanced understanding of host-pathogen relationships to inform disease epidemiology and control in species of conservation concern.

Mathieu Fourment, University of Technology Sydney

Evaluating probabilistic programming and fast variational Bayesian inference in phylogenetics (Short talk)

(Joint work with Aaron Darling, University of Technology Sydney)

Recent advances in statistical machine learning techniques have led to the creation of probabilistic programming frameworks. These frameworks enable probabilistic models to be rapidly prototyped and fit to data using scalable approximation methods such as variational inference. In this work, we explore the use of the Stan language for probabilistic programming in application to phylogenetic models. We show that many commonly used phylogenetic models including the general time reversible substitution model, rate heterogeneity among sites, and a range of coalescent models can be implemented using a probabilistic programming language. The posterior probability distributions obtained via the black box variational inference engine in Stan were compared to those obtained with reference implementations of Markov chain Monte Carlo (MCMC) for phylogenetic inference. We find that black box variational inference in Stan is less accurate than MCMC methods for phylogenetic models, but requires far less compute time. Finally, we evaluate a custom implementation of mean-field variational inference on the Jukes-Cantor substitution model and show that a specialized implementation of variational inference can be two orders of magnitude faster and more accurate than a general purpose probabilistic implementation.

Andrew Francis, Western Sydney University

The space of tree-based phylogenetic networks (Long talk)

(Joint work with Mareike Fischer, Greifswald Universitat, Germany.)

Tree-based networks are a class of phylogenetic networks that attempt to formally capture what is meant by "tree-like" evolution. A given non-tree-based phylogenetic network, however, might appear to be very close to being tree-based, or very far. This talk will formalise a range of ways to define tree-based proximity for unrooted phylogenetic networks, one of which (based on the "nearest neighbour interchange (NNI)") gives rise to a notion of "tree-based rank". This provides a subclassification within the tree-based networks themselves, identifying those networks that are "very" tree-based. We also show that the class of tree-based networks as a whole is connected under the NNI, making it potentially searchable using NNI for sampling.

Sophie Hautphenne, School of Mathematics and Statistics at The University of Melbourne, Scientist at the Chair of Statistics in the Swiss Federal Institute of Technology, and Associate Investigator at the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)

An introduction to Markovian binary trees and their applications (**Keynote** talk)

Markovian binary trees (MBTs) form a particularly versatile and tractable class of continuous-time branching processes. In an MBT, the lifetime and reproduction epochs of each individual are controlled by an underlying transient continuous-time Markov chain. MBTs were initially developed to model evolutionary processes because they offer enough flexibility to account for non-constant speciation and extinction rates. They have also been applied to human and bird populations.
In this talk I will introduce MBTs and some of their properties. I will then discuss their statistical inference, in particular, the estimation of the parameters of the underlying Markov chain from the continuous observation of some populations using an EM algorithm. I will show how this method can be used to estimate trait-dependent speciation and extinction rates from phylogenetic trees when the individuals' traits are not observable. I will also highlight the current challenges and open questions in that research direction.

Michael Hendriksen, Western Sydney University

A Cluster-Similarity Metric on Phylogenetic Trees (Short talk)

Metrics on tree space are ubiquitous in phylogenetics, used from MCMC exploration to consensus trees. However, many of the common metrics suffer from issues – Robinson-Foulds has limited ability to distinguish distances due to a right skew in the distribution, and local operations such as NNI and SPR often result in neighbourhoods that have trees with wildly different hierarchies. We present a metric based on cluster similarity that does not suffer from these issues and has several qualities that make it uniquely suited to MCMC calculations. This metric uses a partial order that considers $RP(X)$ as a graded poset, and can be estimated in polynomial time.

Timothy Hewson, School of Natural Sciences, University of Tasmania

Graded rings of Markov invariants (Short talk)

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.

Lars Jermiin, Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia; School of Biology and Environmental Sciences, University College Dublin, Belfield, Ireland; Earth Institute, University College Dublin, Belfield, Ireland

Phylogenetic analysis of genotype sequence data: a one-way ticket to paradise for Hawaiian Hoary bats (Long talk)

Single nucleotide polymorphism (SNP) data is becoming an increasingly important source of data for phylogenetic analysis, not the least because of the low cost and ease with which SNPs can be obtained. That said, serious challenges remain with respect to the collection, assembly, modelling and analysis of these data. Here, we present and discuss two aspects of this challenge. The first of these is that the Cormish-Bowden's (NAR 13:3012-3030 [1985]) nomenclature for incompletely specified nucleotides no longer applies. The second of these is that standard 4 x 4 Markov models of nucleotide substitution no longer apply. To remedy these challenges, we propose a new type of genetic data—genotype sequence data—and outline two Markov models of genotype substitution. One consequence following from using genotype sequence data is that homologous recombination no longer affects phylogenetic estimates. Another is that SNP data—which look like genotype sequence data—may be analysed using the new Markov models outlined herein. We apply the new models and phylogenetic methods to SNP data obtained from 24 Hawaiian Hoary bat samples and show that we are able to distinguish samples from different islands on the Hawaiian archipelago.

Qiuyi Li, The University of Melbourne

HIDTL, a new gene family evolution model (Long talk)

(Joint work with Yao-ban Chan, The University of Melbourne; Celine Scornavacca, Université Montpellier)

The evolution of gene families is an important aspect of molecular evolution and also crucial when inferring the relationships among species. Gene families evolve through a complex process involving evolutionary events such as speciation, gene duplication, horizontal gene transfer, and gene loss. Furthermore, when a population of individuals undergoes several speciations in a relatively short time, there can exist polymorphisms maintained throughout the time which eventually fix in different descendant lineages. This phenomenon is called incomplete lineage sorting (ILS). Due to these evolutionary processes, there are often topological differences between a gene tree and its corresponding species tree. Reconciliation methods are developed to explain these differences. Accurate gene and species reconciliation is fundamental to infer the evolutionary history of a gene family.
Any reconciliation method is built on a model of gene family evolution. A few gene family evolution models have been proposed over the last decade, for example the duplication-loss model, the locus tree model and the haplotype tree model. However, little attention has been paid to the presence of hemiplasy, which occurs when a newly created locus does not fix in all descendant species. In this talk, we review the existing models of gene family evoluation, and then introduce a new probabilistic gene family evolution model, HIDTL, which combines all the advantages of the existing models and additionally offers more flexibility by allowing hemiplasy. We compare HIDTL with the existing models to show that HIDTL can model more complex scenarios, and so should be used for testing the accuracy of reconciliation methods.

David Liberles, Temple University

A statistical analysis of clusters of amino acid substitutions to compare compensatory processes with directional selection (Long talk)

(Joint work with Peter B. Chi and Westin M. Kosater; David A. Liberles[1], Peter B. Chi[1,2], and Westin M. Kosater[1]; 1. Department of Biology and Center for Computational Genetics and Genomics, Temple University, Philadelphia, PA 19122, USA; 2. Department of Mathematics and Statistics, Villanova University, Villanova, PA 19085, USA)

Identifying positive directional selection as an indicator of functional change in proteins is one of the grand challenges in computational comparative genomics. Here, we present two statistical tests that examine the nature of the 3D clustering of amino acid substitutions along the lineage of a phylogenetic tree for the detection of positive directional selection. Using a parametric bootstrapping approach that controls for position solvent accessible surface area and contact number and examines a range of distances, the first test asks for statistical evidence for clustering against a random null model. The second test then uses the clustering model in the first test (consistent with compensatory processes) to test for positive directional selection when additional changes are deterministically added. The tests do not consider the number or rate of changes and are entirely based upon patterns of clustering.

Nadia El-Mabrouk, Département d'informatique et de recherche opérationnelle, Université de Montréal

A unifying approach for inferring the evolutionary history of gene order and content. (**Keynote** talk)

During evolution, genes are mutated, duplicated, lost and passed to organisms through speciation or Horizontal Gene Transfer (HGT). In addition, their organization in the genome is modified through inversions, transpositions, translocations and other rearrangement events. Understanding how gene order and content have evolved is essential for deciphering gene functions and interactions, with important biological implications. Ideally, all available information on gene sequence and organization should be considered in a single prediction method. However, gene sequence and gene order information are often considered separately. Indeed, inferring rearrangement events modifying gene organization is the purpose of the genome rearrangement field, while inferring losses, duplication and HGT events modifying gene content is the purpose of the gene tree – species tree reconciliation field. In this presentation, I will discuss this issue and present avenues for developing a unifying approach considering both gene orders and gene trees in the purpose of inferring the evolutionary history of gene repertoires.

Bennet McComish, Menzies Institute for Medical Research, UTAS

Is natural selection to blame for a vulvar cancer cluster among young Indigenous women? (Short talk)

Vulvar cancer is usually rare, and occurs most often in postmenopausal women. Among young (<50 years) Indigenous women living in some remote Aboriginal communities, however, the incidence of this malignancy is more than 70 times the national Australian rate for the same age group. Previously, we found that neither excess HPV incidence nor a particularly virulent strain of HPV could explain the very high incidence of vulvar cancer in this population. Reports that cases appeared to cluster in family groups suggested that a genetic susceptibility, either to the effects of HPV or another cause of vulvar cancer, may be involved in this cluster.
To investigate the role of genetic risk factors, whole genomes were sequenced for 29 cases, including four siblings, and 39 controls. We identified a set of variants present in at least 10 cases (including all four sibling cases) and no more than five controls, and used these variants to identify genes of interest, which were found to be enriched for genes known to be involved in cancer or response to HPV. In addition, we searched the sequence data for signatures of natural selection that may have led to increased susceptibility to HPV-induced carcinogenesis in this population.

Allen Rodrigo, The Australian National University

The use of the Heteroduplex Mobility Assay in phylogenetics and population genetics. (Short talk)

Heteroduplexes are double stranded DNA molecules where each strand comes from a different genetic variant in a mixture of haplotypes that have been allowed to disassociate and reanneal. Heteroduplexes migrate at a slower speed relative to homoduplexes on a gel. It has been shown that the speed of migration is inversely proportional to the genetic difference between the variants, and that these distances can be used to build phylogenetic trees. Here, we expand on these techniques, and show how we may estimate not only the trees themselves (with confidence equivalents), but also common population genetic descriptors including mismatch distributions and skyline plots. We validate our methods with real sequences, and compare them to other phylogenetic and population genetic approaches using sequence data.

Julia Shore, School of Natural Sciences, University of Tasmania

Tree algebras: Fun matrix sets with relevance to phylogenetics. (Short talk)

(Joint work with Dr. Jeremy Sumner, Prof. Barbara Holland)

Our recent work in testing origin of life hypotheses utilised a method of generating rate matrices from phylogenetic trees by setting the rate of change between two taxa as the parameter assigned to the most recent common ancestor of those taxa. This methodology has resulted in intriguing matrix sets. It has so far been found that if a unique rate parameter is assigned to each node in a bifurcating rooted tree, the resulting matrix space forms an abelian matrix algebra (a matrix vector space closed under matrix multiplication). The task now is to determine what kind of spaces are generated by trees with non-unique parameters assigned to its nodes.

Albert Soewongsono, School of Natural Sciences, University of Tasmania

Tree Shape Statistics of Tree0s Generated Using Phase-Type Distributed Times to Speciation (Short talk)

(Joint work with Barbara Holland, Małgorzata O'Reilly (School of Mathematics and Physics, UTAS))

This talk will present some preliminary findings in examining tree balance statistics for trees generated using a Coxian Phase Type (PH) distribution of waiting times until speciation. Some earlier results (Hagen *et al.*, 2015) have tried to fit a model that matches with empirical tree data by analysing their tree balance statistics. One of those models is by applying a speciation rate that decreases over species age. This was done by imposing Weibull distribution with shape parameter less than one for speciation time. The biological motivation for using the Weibull was the assumption that a species can be viewed as a collection of i.i.d. large populations. The simulation done using that model suggests results that match thousands of empirical trees in terms of their balance.However, viewing those sub-populations as being i.i.d. may not be biologically reasonable. Here, we will be using PH distribution, specifically Coxian PH distribution to analyse the problem. The justification in using PH is due to its denseness in the field of all positive-valued distributions and using Coxian PH because every acyclic PH distribution has a Coxian PH representation. The early observations using simulations with PH type show a prospective direction towards fitting to empirical tree data.

keywords: Phase-type distribution, tree balance statistics, fully-resolved rooted trees.

Tristan Stark, School of Natural Sciences, University of Tasmania; Department of Biology and Center for Computational Genetics and Genomics, Temple University

Characterising amino acid substitution with complete linkage of sites on a lineage. (Short talk)

(Joint work with David Liberles)

Amino acid substitution models are commonly used for phylogenetic inference, for ancestral sequence reconstruction, and for the inference of positive selection. All commonly used models explicitly assume that each site evolves independently, an assumption that is violated by both linkage and protein structural and functional constraints. Here we attempt to address complete linkage by considering a population-genetic model for the evolution of linked sites. The model is a composite of several Moran models tracking the evolution of a population with fixed number of alleles at some (potential) linkage block. When a mutation arises, a 2-allele Moran model is started to track the genotypes of the population, another, linked mutation may arise on the background of the already-segregating allele, at which point a 3-allele Moran model is started. This process continues until the eventual fixation of some allele. We have analysed this model directly for very-small populations to investigate the effect of scenarios of complete linkage on the estimation of fitness parameters from substitution data, and we are working on developing an approximation to allow tractable analysis for large populations.

Mike Steel, Biomathematics Research Centre, University of Canterbury

Combinatorial properties of phylogenetic diversity indices (Short talk)

(Joint work with Kristina Wicke)

Phylogenetic diversity indices provide a formal way to apportion 'evolutionary heritage' across species. Two natural diversity indices are Fair Proportion (FP) and Equal Splits (ES). FP is also called 'evolutionary distinctiveness' and, for rooted trees, is identical to the Shapley Value (SV), which arises from cooperative game theory. In this talk, I consider the extent to which FP and ES can differ, characterise tree shapes on which the indices are identical, and study the equivalence of FP and SV and its implications in more detail.
We also define and investigate analogues of these indices on unrooted trees (where SV was originally defined), including an index that is closely related to the Pauplin representation of phylogenetic diversity.

Joshua Stevenson, University of Tasmania

Investigating rank-based approaches for inferring phylogenies (Short talk)

Flattenings are matrices constructed using site-pattern counts from an alignment. These matrices provide a way of identifying 'true' *splits*, and hence the true evolutionary tree, via the evaluation of their rank. The size of these matrices, exponential in the number of taxa, introduces a computational challenge. This challenge led to the development of so-called *subflattenings* (Sumner, 2017), which exhibit analogous rank properties but have smaller dimensions (quadratic in the number of taxa). The construction of subflattenings involves representation theory and the application of a similarity transformation in which some choices are involved.
This talk summarises my Honours thesis, which explores some algebraic concepts related to these matrices, and the practical implications of some possible choices in the construction of subflattenings.

Seth Sullivant, North Carolina State University, Department of Mathematics

Identifiability in Phylogenetics Using Algebraic Matroids (**Keynote** talk)

Identifiability is a crucial property for a statistical model since distributions in the model uniquely determine the parameters that produce them. In phylogenetics, the identifiability of the tree parameter is of particular interest since it means that phylogenetic models can be used to infer evolutionary histories from data. In this paper we introduce a new computational strategy for proving the identifiability of discrete parameters in algebraic statistical models that uses algebraic matroids naturally associated to the models. We then use this algorithm to prove that the tree parameters are generically identifiable for 2-tree CFN and K3P mixtures. We also show that the k-cycle phylogenetic network parameter is identifiable under the K2P and K3P models.

Jeremy Sumner, School of Natural Sciences, University of Tasmania

The Markov embedding problem from an algebraic perspective (Short talk)

(Joint work with Michael Baake)

A Markov matrix is said to be embeddable if it admits a representation as the exponential of a rate (generator) matrix. Equivalently, a Markov matrix is embeddable if and only if it arises as a probability transition matrix for a (finite-state) continuous-time Markov chain. In this work, we revisit the Markov embedding problem from an algebraic perspective with a focus on model-specific variants; in particular, we take examples from nucleotide substitution models commonly used in phylogenetics. Algebraically, we show that the centralizer of the Markov matrix plays a pivotal role in controlling the solution space.

Venta Terauds, Discipline of Mathematics, University of Tasmania

Symmetry and redundancy: choosing the right algebra for circular genome distance computations (Long talk)

(Joint work with Jeremy Sumner)

The maximum likelihood estimate of time elapsed (MLE) has many advantages over other estimates of evolutionary distance for circular genomes under rearrangement models. However, even when one utilises the representation theory of the symmetric group algebra to convert the combinatorial computations into matrix ones, the calculation of MLEs retains a factorial complexity.
We show that the appropriate theoretical setting for the MLE computations is in fact not the symmetric group algebra but a smaller algebra. By incorporating the symmetry of the genomes and the model into the structure of the algebra, we remove redundancy in the calculations and make a commensurate gain in efficiency.

Katharine Turner, Mathematical Sciences Institute, Australian National University

Topological data analysis of phylogenetic trees? (Long talk)

I come as a representative of the field of topological data analysis, where a phylogenetic tree is considered an example of a labelled merge tree. In the hope of building bridges, I will describe some of the distances considered in topological data analysis for both labelled and unlabelled merge trees. No background in topology will be assumed.

Arndt von Haeseler, University of Vienna and Medical University of Vienna

Evolution of sequences with taboos (Long talk)

(Joint work with Cassius Manuel, Stephan Pfannerer and inspired by discussion with Mike Charleston)

Models of sequence evolution typically assume that all sequences are possible. However, restriction enzymes that cut DNA at specic recognition sites provide an example where carrying a recognition sequence can be lethal. Motivated by this observation, we studied sequence evolution that takes taboo subsequence into account.
We discuss conditions under which the resulting sequence space stays connected and we show that for biological taboo sequences the resulting sequence space is connected. We give a simple example of taboo words, where the sequence space is not connected.
Finally we show how to model sequence evolution in the presence of taboo sequences.

Amanda Wilson, Temple University

Modeling Probabilities of Retention of Gene Copies after Consecutive Whole Genome Duplication Events (Short talk)

(Joint work with David A Liberles, Temple University. Tristan Stark, Temple University)

Gene duplications are a major mechanism in allowing protein and pathway diversification. They create genes that lack selective constraint, which provides opportunity for the genes to accumulate mutations under positive or neutral selection, allowing the genes to take on new or more specialized functions (neofunctionalization, subfunctionalization). Whole genome duplications are common among fish and plant species, and are often the result of meiotic errors. In the past we often attributed the retention of a gene to the gene duplicability hypothesis, that some genes are inherently more likely to be retained. However, a recent study showed unexpected results in the Atlantic salmon genome that appeared to suggest that the probabilities of a gene being retained after one whole genome duplication event is independent of the probability that a gene is retained after a second whole genome duplication event. Here, we construct four models for different hypotheses that explain retention of gene copies that result from consecutive gene duplication events. The four models include an independence hypothesis, a revised gene duplicability hypothesis, a novel mutational opportunity hypothesis, and a hybrid model of the latter two models. The novel mutational opportunity hypothesis explains that the probability of the retention of a gene after consecutive whole genome duplication events is affected by mechanisms that allowed for these genes to be retained after previous whole genome duplication events. For all of the models, we incorporate time between duplication events, t_1, and time since the most recent duplication event, t_2, as important variables for determining the conditional probabilities of gene-specific retention. We also consider the number of modular functional units a gene has. These models can be used together with statistical model testing to identify the best supported biological hypothesis. This ultimately will entail applying these models to real world data sets in organisms that have had relatively recent consecutive gene duplication events at varying times in history to see which hypothesis is most likely to give rise to observed data across t_1 and t_2 values.

Benjamin Wilson, Lateral R&D Pty Ltd

Joint-estimation of pairwise distances under a weakened 4PC (Long talk)

We propose an algorithm for distance estimation that uses hyperbolic geometry to jointly estimate the pairwise distances subject to a weakening of the four point condition. The algorithm represents the taxa by a point configuration in a space of constant negative curvature. The points are iteratively rearranged such that the distance between any pair of points accounts for the site differences between the corresponding taxa under the chosen mutation model. The rearrangement is achieved via standard gradient-based optimisation and may be seen as a differentiable approximation of the maximum likelihood tree search. In this talk, I'll demonstrate the approach via animated examples, before presenting the results of a first evaluation of its performance.

List of attendees

Bochenek, Nicholas . . . . . nb31@utas.edu.au
Brown, Matilda . . . . . matilda.brown@utas.edu.au
Burden, Conrad . . . . . conrad.burden@anu.edu.au
Burridge, Chris . . . . . chris.burridge@utas.edu.au
Cartwright, Reed . . . . . cartwright@asu.edu
Chan, Yao-ban . . . . . yaoban@unimelb.edu.au
Charleston, Michael . . . . . michael.charleston@utas.edu.au
Clark, Chad . . . . . chad.clark@westernsydney.edu.au
Crotty, Stephen . . . . . stephen.crotty@adelaide.edu.au
Diao, Jiahao . . . . . jiahao.diao@utas.edu.au
Donovan, Diane . . . . . dmd@maths.uq.edu.au
Downard, Kevin . . . . . kevin.downard@unsw.edu.au
Feng, Qian . . . . . fengq2@student.unimelb.edu.au
Fountain-Jones, Nick . . . . . nick.fountainjones@utas.edu.au
Fourment, Mathieu . . . . . mathieu.fourment@uts.edu.au
Francis, Andrew . . . . . a.francis@westernsydney.edu.au
Hautphenne, Sophie . . . . . sophiemh@unimelb.edu.au
Hendriksen, Michael . . . . . m.hendriksen91@gmail.com
Hewson, Timothy . . . . . timothy.hewson@utas.edu.au
Hodge, Terrell . . . . . terrell.hodge@wmich.edu
Holland, Barbara . . . . . barbara.holland@utas.edu.au
Huttley, Gavin . . . . . gavin.huttley@anu.edu.au
Jarvis, Peter . . . . . peter.jarvis@utas.edu.au
Jermiin, Lars . . . . . lars.jermiin@anu.edu.au
Li, Qiuyi . . . . . qiuyi.li1@unimelb.edu.au
Liberles, David . . . . . daliberles@temple.edu
Liu, Qin . . . . . qin.liu@utas.edu.au
El-Mabrouk, Nadia . . . . . mabrouk@iro.umontreal.ca
McComish, Bennet . . . . . bennet.mccomish@utas.edu.au
O'Reilly, Małgorzata . . . . . malgorzata.oreilly@utas.edu.au
Rodrigo, Allen . . . . . allen.rodrigo@anu.edu.au
Shore, Julia . . . . . julia.shore@utas.edu.au
Soewongsono, Albert . . . . . albertchristian.soewongsono@utas.edu.au
Stark, Tristan . . . . . tristan.stark@utas.edu.au
Steel, Mike . . . . . mathmomike@gmail.com
Stevenson, Joshua . . . . . joshua.stevenson@utas.edu.au
Sullivant, Seth . . . . . smsulli2@ncsu.edu
Sumner, Jeremy . . . . . jsumner@utas.edu.au
Sykes, Janan . . . . . janan.sykes@utas.edu.au
Terauds, Venta . . . . . venta.terauds@utas.edu.au
Thompson, Bevan . . . . . mahthomp@maths.uq.edu.au
Turner, Katharine . . . . . katharine.turner@anu.edu.au
von Haeseler, Arndt . . . . . arndt.von.haeseler@univie.ac.at
Wilson, Amanda . . . . . tuj69282@temple.edu
Wilson, Benjamin . . . . . benjamin@lateral.io
Woodhams, Michael . . . . . michael.woodhams@utas.edu.au
Yates, Luke . . . . . luke.yates@utas.edu.au