



# Portobello '18

New Zealand Phylogenomics Conference

11-16 February, 2018  
Portobello, Otago Peninsula  
New Zealand



# 1 Welcome

Welcome to the 22nd Annual New Zealand Phylogenomics Meeting, the second to be held in Otago.

This meeting has been made possible through generous sponsorship from the Allan Wilson Research Theme at the University of Otago as well as the Otago Department of Mathematics and Statistics.

We gratefully acknowledge help and support from

1. Wayne Cameron and Portobello Community Inc.
2. Peter Simkins and the staff of the Penguin Café
3. Marguerite Hunter, Dept. Mathematics and Statistics, University of Otago
4. Dietrich Radel and the University of Canterbury Biomathematics Research Centre.

## Coronation Hall



The Coronation Hall was built in Portobello in 1912. Constructed originally to house the Portobello Road Board a 250 grant was made for its construction from the Government Coronation Fund in 1911. A portion of the small lagoon area adjacent to the road and Latham Bay was reclaimed to create the site the hall sits on today. In 1921 the kitchen addition was added. Today the hall is operated by a local committee who manage the bookings, cleaning and renovations of this valuable community asset.


<http://portobello.org.nz/our-community/portobello-coronation-hall/>

## 2 Information

- All talks will be held in Coronation Hall.
- Standard length talks are 15-20 minutes + 5 minutes discussion.
- Load up all talks onto the conference laptop before your session: contact Marnus or Lydia for help.
- An informal breakfast will be available in the Coronation Hall from 7:30am.
- Your registration includes breakfast, morning and afternoon teas, lunches, fish and chip dinner (Monday), boat launch (Tuesday), banquet (Thursday).
- There will be no beer tasting this year due to licensing issues, but you will be able to buy at least some Dunedin craft beer from the Portobello store or from the pub.
- Groceries etc. available from the Portobello store. Meals available from the pub, from 1908 Cafe Restaurant (make sure to reserve ahead), or Fish and Chip store.
- If you are in need of help, assistance, information, transport etc. talk to Mike Hendy, David Bryant or Marguerite Hunter.
- Limited wireless internet access is available in the hall. The password is "PHYLOBELLO15". This is running through a couple of mobile modems, so please don't download large files and please turn off system and software updates.

# Portobello

## CONFERENCE VENUE

 Coronation Hall

(approx 250m to the store)

## Store

 Portobello Store : 7.30am-8pm

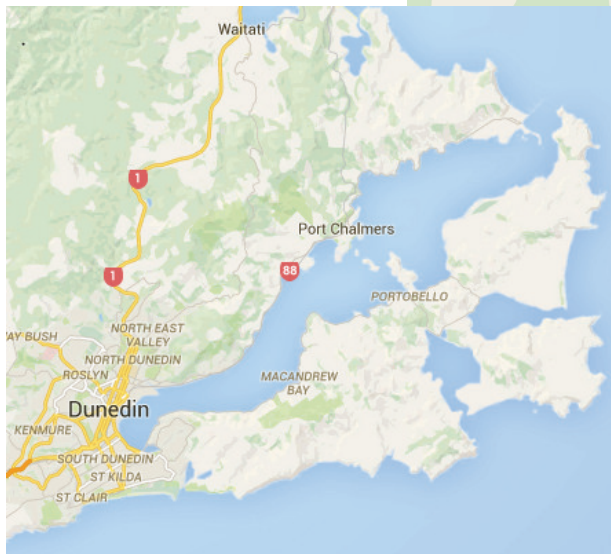
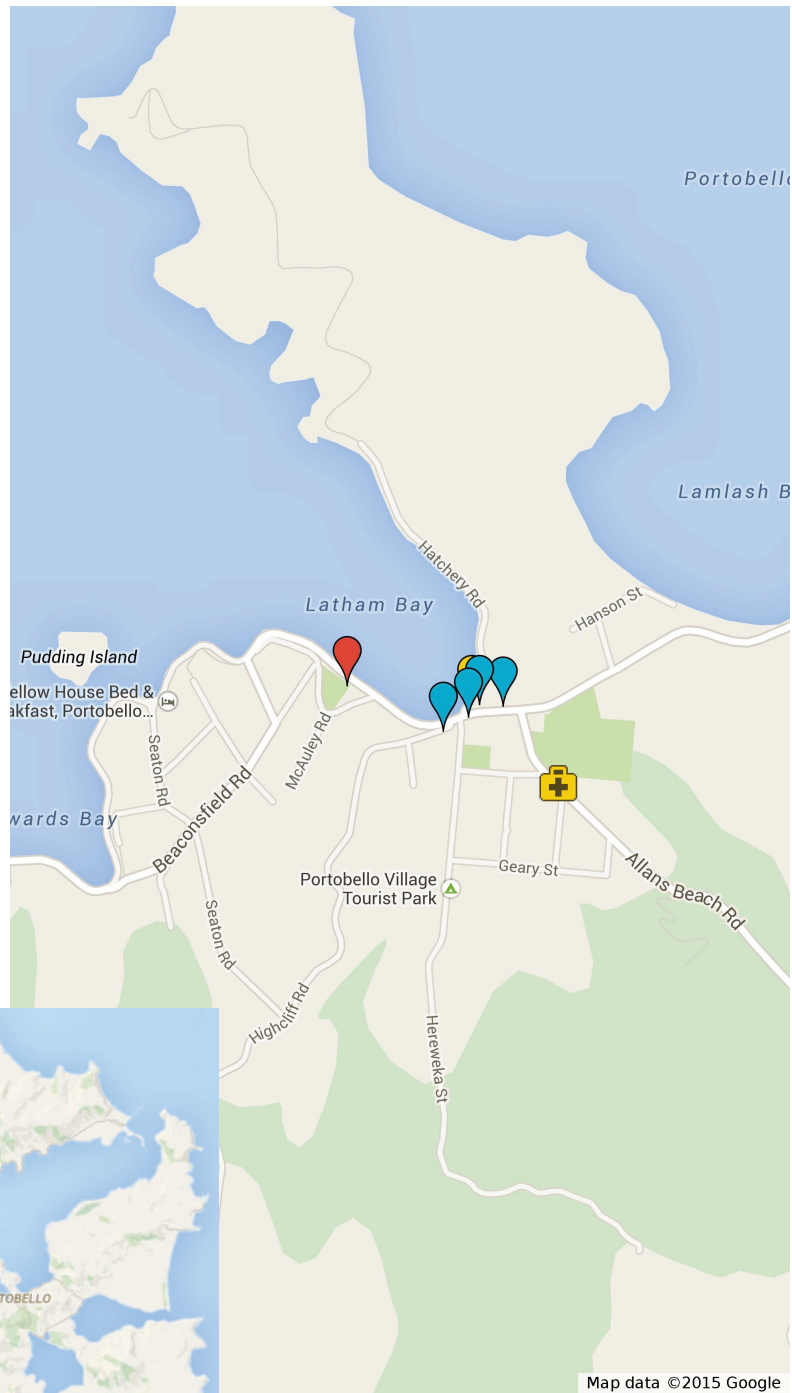
## Eating places

-  Portobello Hotel: 11am - late
-  Penguin Cafe Portobello: 8am-4pm
-  1908 Cafe Restaurant: 5pm - late
-  Ric's galley: 5.30-8pm, Wed - Sun

## Medical centre

 Otago Peninsula Medical Centre  
8.30am-5pm, Mo-Fr  
Tel: 03 4780880  
After hours: 03 4792900

No ATMs on the peninsula;  
EFTPOS usually available



### 3 Timetable

#### Sunday, February 11

---

---

17h00–20h00 Registration, Coronation Hall

---

---

#### Monday, February 12

---

---

8h30		Registration, Coronation Hall
9h00	Alex Gavryushkin	Fitness landscapes and incomplete data
9h25	Chris Simon	Parallel, episodic, and spectacular diversification of the microbial endosymbionts of cicadas
9h50	Ashar Malik	Exploring deep phylogenies using protein structure
10h15	<i>Morning tea</i>	
10h55	<i>Announcements</i>	
11h00	Alexandra Gavryushkina	Total-evidence analysis under stratigraphic range fossilised birth-death process
11h25	Jack Simpson	Combining phylogenetic trees into networks: What can be done with just two trees?
11h50	Daniel Huson	SplitsTree5 - New software for computing phylogenetic trees and networks
12h15	<i>Lunch break</i>	
14h10	Paul Gardner	Finding functionally significant genome variation
14h35	Maj Padamsee	Fungal endophytes associated with roots of <i>Agathis australis</i>
15h00	<i>Afternoon tea</i>	
16h00	Jonathan Klawitter	On shortest paths between phylogenetic networks under rSPR
16h25	Jeremy Sumner	Maximum likelihood distances for genome rearrangement models
18h00	Fish and chip dinner, either in the hall or at a local beach (weather permitting).	

---

---

## Tuesday, February 13

---

---

9h20	Jordan Douglas	Bayesian analysis and comparison of stochastic transcription models
9h45	Remco Bouckaert	Methods for Analysing a Deadly Disease Lurking in New Zealand Forests: a short history of Kauri Dieback
10h10	<i>Morning tea</i>	
11h00	Kristina Wicke	On the Shapley value of unrooted phylogenetic trees
11h25	Momoko Hayamizu	Counting the number of support trees for a binary phylogenetic network
11h50	Simone Linz	Characterizing the hybridization number for a set of phylogenies
12h15	<i>Lunch break</i>	
14h00	Charles Semple	When is a network captured by its path distances?
14h25	Jing Yang	The global dynamics of avian influenza H9N2 and the influence of poultry production and trade on its spread
14h50	David Bryant	Dinner products and the evolution of sets
15h15	<i>Afternoon tea</i>	
16h00	Harbour cruise, leaving Portobello wharf at 16h00, back at 18h30. Cash bar available on board.	
18h00	Free time (make your own dinner plans)	

---

---

## Wednesday, February 14th

Excursion day. Collect packed lunch from Coronation Hall between 8h00 and 9h30.

## Thursday, February 15th

---

---

9h00	Lina Herbst	On the Accuracy of Ancestral Sequence Reconstruction with Parsimony
9h25	Denise Kühnert	Bacterial phylodynamics: Can we disentangle bacterial transmission dynamics between hospitals and the community?
9h50	Mareike Fischer	Combinatorial views on persistent characters
10h15	<i>Morning tea</i>	
11h00	Marnus Stoltz	Some interesting properties of the Wright-Fisher Diffusion in one-dimension
11h25	Nick Matzke	Large state spaces and state-dependent speciation/extinction models: problems and prospects
11h50	Benny Chor	Ultra conserved protein elements
12h15	<i>Lunch break</i>	
14h10	Mike Steel	Species notions that combine a phylogenetic tree and phenotypic partitions
14h35	Stefan Grünewald	TBC
15h00	<i>Afternoon tea</i>	
16h00	Russell Gray and Adam Powell	Waves of history in Remote Oceania: language continuity despite population replacement in Vanuatu

---

---

18h30– Conference Banquet at Glenfalloch. Meet at Coronation Hall. Bus leaves 18:30

---

---

## Friday, February 16th

---

---

9h00	Michael Charleston	Taboo sequences
9h25	Walter Xie	Codon Substitution Model Implementation in BEAST 2
9h50	Graham Wallis	Split NZ: phylogeographic breaks in the South Island fauna and their causes
10h15	<i>Morning tea</i>	
11h00	Mike Hendy	SIMPLET: Split-Induced MP-Like Evolutionary Trees
11h25	Barbara Holland	Convergence-Divergence networks
11h50	Wrap up	

---

---

## 4 Submitted Abstracts

### Remco Bouckaert

(Center of Computational Evolution, U of Auckland, remco@cs.auckland.ac.nz)

#### *Methods for Analysing a Deadly Disease Lurking in New Zealand Forests: a short history of Kauri Dieback*

Phytophthora is a plant disease causing genus with over a 100 species that are responsible for damage to agriculture (in particular cocoa and potato) and forests. Phytophthora Agathidicida (aka PTA) threatens Kauri, the largest tree and culturally one of the most significant trees in New Zealand. Several theories exist for the diversification of PTA. 1: PTA diversified subsequent to its introduction in New Zealand in the 1950s at the Raetea nursery. 2: PTA diversified during a period of intense logging that accompanied European settlement (1850s-1950s). 3: PTA was endemic and diversified prior to 1850s. Using sequence data from PTA and *P. Infestans*, we use Bayesian phylogenetic techniques to distinguish between these theories employing a number of the latest methods. This talk will concentrate on these methods, in particular bModelTest, nested sampling and tree model adequacy implemented in BEAST.

### David Bryant

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

#### *Dinner products and the evolution of sets*

I will talk about some mathematics which has arisen out of our attempts to model environmental niches. Suppose that we model the 'niche' of a species as a convex set in some multi-dimensional space: the space might have dimensions for environmental features, or it might be completely abstract. How can we model the change of sets over time? or along a phylogeny? And how can we infer ancestral states tractably? We have made some progress on these problems using mathematical diversities, work which has led to us rewriting some textbook linear algebra.

### Michael Charleston

(University of Tasmania, michael.charleston@utas.edu.au)

#### *Taboo sequences*

In molecular phylogenetics it is routinely assumed that all intermediate sequences between two molecular sequences are equally valid: there are no "no-go areas", no "holes" in sequence space, that must be navigated. But if there are significant portions of the sequence space that are disallowed, there might be consequences to the accuracy of distances estimated on sequences in that space, and therefore consequences to phylogenetic inference.

Here, we consider the possibility that not all sequences are permitted — they are treated as "taboo". In this treatment we consider a very simple sequence space, as binary sequences, in which a short "00" motif is taboo.

We explore the combinatorial aspects of such sequence spaces, and determine some potential consequences to phylogenetic inference. These include whether Hamming distances can be realised within the space, whether Parsimony can work, and what is the effect on estimated dis-



tances in comparison with a Jukes-Cantor model of sequence evolution in which there are no taboo sequences.

[Joint work with Arndt von Haeseler]

## **Benny Chor**

(Tel Aviv University, benny@cs.tau.ac.il)

### *Ultra conserved protein elements*

#### Abstract

Ultra conserved DNA elements are long segments of consecutive nucleotides in a genomic sequence, that are shared exactly (conserved with 100% identity) among a set of three or more species. In 2004, ultra conserved DNA elements (UCDEs) were discovered, and sparked extensive follow up research. This discovery has challenged a number of accepted views on the relations between conservation, function, and essentiality, as many UCDEs are located in gene deserts, away from any protein coding genes, and have no known function.

We expand the notion of ultra conservation to proteins. Ultra conserved protein elements (UCPEs) are amino acid sequences that are shared exactly (conserved with 100% identity and no gaps) among a set of three or more species. We considered five mammalian species: human, mouse, rat, cow, and dog. We found over 2600 clusters of proteins (one per species) with ultra conserved protein element of length 80 amino acids or more. Proteins containing these elements are enriched for tens to hundreds of (human) gene ontology (GO) terms, for tens of KEGG pathways, and for tens of Pfam domains.

It is not clear how to measure the 'surprise' of having that many ultra conserved protein elements.

## **Jordan Douglas**

(University of Auckland, jdou557@aucklanduni.ac.nz)

### *Bayesian analysis and comparison of stochastic transcription models*

Transcription is a critical biological process which occurs in all living organisms. It involves copying the organism's genetic material into mRNA which can then be utilised by cellular machinery to produce protein. This is carried out by the protein RNA polymerase.

The main transcription pathway occurs in a three step cycle: first RNA polymerase translocates forward by one nucleotide, second the next complementary nucleotide binds to the protein, and third this nucleotide is added onto the end of the mRNA. However, in principle these steps could be reversible. A single cycle has three states and three reactions, so up to 6 rates may be required to model the forward and backward reactions. Different authors have made different assumptions around which rates are necessary to estimate. I wanted to perform a systematic comparison of all models to determine which parameters are necessary. In this talk I will describe how I have used Markov chain Monte Carlo approximate Bayesian computation (MCMC-ABC) as a means of parameter estimation and model selection.

## **Mareike Fischer**

(Greifswald University, email@mareikefischer.de)

### *Combinatorial views on persistent characters*

There are various ways to define and measure the fit of data on a given tree. For instance, for binary characters, a perfect fit of the data on the tree would correspond to a so-called binary perfect phylogeny, which allows for each character only one gain (and no loss), i.e. one change in total. A slightly less restrictive fit is given by a so-called binary perfect phylogeny with persistent characters. Persistent binary characters allow for a trait to be gained once and lost once, i.e. there can be up to two changes in total, but in a predetermined order.

Perfect phylogenies with persistent characters have recently been thoroughly studied in computational biology. However, some combinatorial problems remained unsolved. In my talk, I will answer some of these questions. In particular, I will characterize persistent characters in terms of Maximum Parsimony and the famous Fitch algorithm. Then I will show that the number of persistent characters is higher the more unbalanced a tree is (according to the Sackin index). Moreover, I will provide an upper bound on the number of characters together with their persistence status that are needed in order to uniquely determine a tree.

## **Paul Gardner**

(University of Otago, paul.gardner@otago.ac.nz)

### *Finding functionally significant genome variation*

Whole genome sequencing is now a routine operation for many research groups. The fundamental question that researchers wish to answer with each newly sequenced genome is “is there any interesting variation?” followed closely by “what does it do?”

We have been developing an approach based upon profile hidden Markov models that shows some promise. It captures significant non-synonymous variation within protein-coding genes, that alter either highly conserved residues and/or replace biochemically dissimilar residues. We have tested this on a range of datasets and have applied the method to pathogenic bacteria. We have identified several recurrent mutations that are strongly associated with adaptation to more invasive lifestyles.

In future work we hope to further generalise the approach to evaluate non-coding as well as synonymous variation that may also be functionally important.

## **Alexandra “Sasha” Gavryushkina**

(ETH Zurich, gavryushkina@gmail.com)

### *Total-evidence analysis under stratigraphic range fossilised birth-death process*

Recently, new methods for dating species phylogeny using the fossil record have been proposed. These methods are based on employing the fossilised birth-death process as a model for simultaneous processes of speciation and fossil sampling. A sequence of events: fossilisation, preservation, discovery, and inclusion in an analysis is modelled as a single fossil sampling event which occurs according to a Poisson process along lineages in a phylogeny. The existing applications of

this model do not allow multiple fossils to be assigned to a single species. However, paleontological data bases are often comprised of stratigraphic ranges which represent multiple fossil specimens of the same species. The direct modelling of the stratigraphic range data would facilitate a more accurate inference. We have extended the fossilised birth-death process to account for the stratigraphic range data. We have used this model (as an add-on for BEAST2) for a total-evidence analysis of simulated and empirical datasets to estimate dated phylogenies and speciation parameters using stratigraphic range data and morphological/molecular data. This method can also be used to infer transmission trees with multiple pathogen samples per patient or with known (or estimated) infection dates.

## **Alex Gavryushkin**

(University of Otago - Computer Science Dept, alex@gavruskin.com)

### *Fitness landscapes and incomplete data*

Despite fitness being a central concept in evolutionary biology, it is hardly possible to measure fitness for all genotypes in a natural population. Furthermore, the measurements are prone to statistical uncertainties and experimental noise. I will present an approach to make inferences about genetic interactions when the fitness landscape is incompletely determined. We demonstrate [1, 2, 3] that genetic interactions can often be inferred from fitness rank orders, where all genotypes are ordered according to fitness, and even from partial fitness orders, where only a fraction of the ranking is certain. We provide a complete characterisation of partial fitness orders that imply interactions. We also design a highly efficient algorithm to detect interactions from these kinds of data. Our methods apply to all common types of gene interaction and facilitate comprehensive investigations of diverse genetic interactions. I will demonstrate, for example, how these methods can be used to reveal non-trivial patterns of genetic interactions in HIV-1, the malaria-causing parasite *Plasmodium vivax*, the fungus *Aspergillus niger*, and the TEM-family of  $\beta$ -lactamase associated with antibiotic resistance. I will conclude with a strikingly unexpected application of our methods [3] to interspecies gut bacterial interactions where fitness is measured as that of the host.

[1] <https://doi.org/10.7554/eLife.28629>

[2] <https://doi.org/10.1101/180976>

[3] <https://doi.org/10.1101/232959>

## **Russell Gray**

(Max Planck Institute for the Science of Human History, gray@shh.mpg.de)

### *Waves of history in Remote Oceania: language continuity despite population replacement in Vanuatu*

The Austronesian languages of Vanuatu are notable for both their sheer number and their marked deviation from most other Oceanic languages. Their aberrant features include non-decimal numeral systems, rounded labial phonemes, dually articulated labial-velar phonemes, bilabial trills, dual exclusion of p and c phonemes, and serial verb constructions. Blust (2008) has argued that the presence of these linguistic features can only be explained by a wave of Papuan expansion into Remote Oceania that quickly followed the initial Austronesian expansion (around 3200 BP). Recent genomic analyses show that the earliest peoples reaching Remote Oceania associated with

Oceanic-speaking Lapita culture were almost completely East Asian, without detectable Papuan ancestry. Yet Papuan genetic ancestry is found across present-day Pacific populations, indicating that Papuan peoples have played a significant but largely unknown ancestral role. In this paper we (Russell Gray and Adam Powell) will outline what a combination of new ancient genome data and linguistic analyses can tell us about early Papuan and Austronesian language contact in Vanuatu. Our genome-wide data from 27 contemporary ni-Vanuatu demonstrate a subsequent and almost complete replacement of Lapita-Austronesian by Papuan ancestry. Despite this massive demographic change, incoming Papuan languages did not replace Oceanic languages.

## **Momoko Hayamizu**

(The Institute of Statistical Mathematics / JST PRESTO, hayamizu@ism.ac.jp)

### *Counting the number of support trees for a binary phylogenetic network*

A rooted binary phylogenetic network  $N$  on a set  $X$  of leaves is called a tree-based network if  $N$  contains a spanning tree  $T'$  that is a subdivision of a rooted binary phylogenetic tree  $T$  on the same leaf-set  $X$ . If such trees exist for  $N$ , then  $T'$  is called a support tree for  $N$ , whereas  $T$  is called a base tree of  $N$ . In this talk, I will outline a linear time algorithm for computing the number of support trees for an arbitrary rooted binary phylogenetic network  $N$ , which returns a non-zero number if and only if  $N$  is tree based.

## **Mike Hendy**

(University of Otago, mhendy@maths.otago.ac.nz)

### *SIMPLET: Split-Induced MP-Like Evolutionary Trees*

Maximum Parsimony (MP) has been an early popular phylogenetic method, but is hampered by some limitations. We propose an adaption to MP, "SIMPLET", which is designed to address these limitations. SIMPLET is based on interpreting the frequencies of splits induced by an alignment of nucleotide sequences under an i.i.d. model with each substitution having equal weight.

## **Lina Herbst**

(Ernst-Moritz-Arndt-University Greifswald, lina.herbst@uni-greifswald.de)

### *On the Accuracy of Ancestral Sequence Reconstruction with Parsimony*

We examine a mathematical question concerning the reconstruction accuracy of the Fitch algorithm for reconstructing the ancestral sequence data of the most recent common ancestor given a phylogenetic tree and sequence data for all taxa under consideration. In particular, in the case of the symmetric four-state substitution model, we answer affirmatively a conjecture of Li, Steel and Zhang which states that for any ultrametric phylogenetic tree and a symmetric model, the Fitch parsimony method using all terminal taxa is more accurate, or at least as accurate, for ancestral state reconstruction than using any particular terminal taxon. This conjecture had so far only been answered for two-state characters by Fischer and Thatté. Here, we focus on answering the biologically more relevant case with four states, which corresponds to ancestral sequence reconstruction from DNA or RNA data. (Joint work with Mareike Fischer)

## **Barbara Holland**

(University of Tasmania, barbara.holland@utas.edu.au)

### *Convergence-Divergence networks*

Authors: J Mitchell, B Holland\* (presenting), J Sumner

Over the last few years we have been working on a class of models we call 'Convergence-Divergence models'. These allow for traditional speciation events where species diverge from a common ancestor but they also allow species to become more similar again. In this talk I will

(1) introduce the model,

(2) discuss potential areas of application: morphological convergence, modelling gene content, introgression

(3) discuss issues around identifiability with an intriguing link to the molecular clock in the three-taxon case

(For people who saw this talk at Phylomania I promise there are some new results for 4 taxa)

## **Daniel Huson**

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

### *SplitsTree5 - New software for computing phylogenetic trees and networks*

SplitsTree4, written and published in 2006, is an interactive tool for computing phylogenetic trees and networks. While still widely used, the program has many limitations by today's standards. For example, it was designed for small datasets, the user interface is dated, visualization techniques are limited and most algorithms are not parallelized. In this talk we will present SplitsTree5, a reimplement of SplitsTree that aims at addressing these issues. In addition, the new program will also incorporate many features of Dendroscope3, in particular rooted phylogenetic networks, of PopArt, in particular haplotype networks, as well as PCoA and related methods. SplitsTree5 is designed around a workflow graph that explicitly represents all data and algorithms. It provides an automatically generated and updated methods section for publication. To illustrate the work being done to improve and parallelize algorithms, using SplitsTree5, it takes less than two minutes to read in 10,000 trees on 600 taxa and compute and draw a 10%-consensus network for them. SplitsTree5 is written in Java using JavaFX and will be released as open source. In this talk we will demonstrate an alpha version of SplitsTree5. This is joint work with David Bryant and Daria Evseeva.

## **Jonathan Klawitter**

(University of Auckland, jo.klawitter@gmail.com)

### *On shortest paths between phylogenetic networks under rSPR*

The rSPR (rooted Subtree Prune and Regraft) operation has recently been generalised from phylogenetic trees to phylogenetic networks. The space of phylogenetic networks under rSPR can be visualised by a graph in which each vertex represents a phylogenetic network and two vertices are joined by an edge precisely if they are one operation apart. The rSPR distance of two phylogenetic networks is then defined as the length of a shortest path between these two networks in this graph. In this talk we will discuss the rSPR distance and the behaviour of shortest paths between two phylogenetic networks.

## **Denise Kühnert**

(University Hospital Zurich, denise.kuehnert@gmail.com)

### *Bacterial phylodynamics: Can we disentangle bacterial transmission dynamics between hospitals and the community?*

The field of phylodynamics has arisen focused on virus epidemics. Bacterial phylodynamics have only become feasible due to advances in sequencing technologies. We perform a simulation study to evaluate the potential of whole genomic sequence analysis of nosocomial outbreaks. *Staphylococcus aureus* serves as a model organism, since its epidemic dynamics are relatively well-understood. Can we tell if transmission is mainly driven by the hospital or the community?

## **Simone Linz**

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

### *Characterizing the hybridization number for a set of phylogenies*

Throughout the last decade, we have seen much progress towards characterizing and computing the minimum hybridization number for a set of phylogenetic trees. Roughly speaking, this minimum quantifies the number of hybridization events needed to explain a set of trees by simultaneously embedding them into a phylogenetic network. From a mathematical viewpoint, the notion of agreement forests is the underpinning concept for almost all results that are related to calculating the minimum hybridization number for two trees. However, despite various attempts, characterizing this number in terms of agreement forests for more than two trees remains elusive. In this talk, we first discuss a new characterization to compute the minimum hybridization number in the space of tree-child networks. Subsequently, we show how this characterization extends to the space of all rooted phylogenetic networks. (Joined work with Charles Semple.)

## **Ashar Malik**

(Massey University, a.j.malik@massey.ac.nz)

### *Exploring deep phylogenies using protein structure*

The primary step in the process of characterising a novel protein sequence is by comparison with those already characterised. Similarity based functional characterisation and determination of an evolutionary origin can become a non-trivial problem for significantly diverged proteins. Protein structure, on the other hand, is considered to be conserved over longer evolutionary timescales. An evolutionary signal lost from the sequence may therefore still be retained within the conserved structure. Current times are seeing an exponential growth in protein structural data, presenting a unique opportunity to explore deep evolutionary questions. However a poor understanding of protein structural evolution prevents usage of classical sequence-based phylogenetic methods. Empirical distance-based methods have been employed, however the lack of a method to gauge the robustness of evolutionary relationships inferred limits their traction. A novel molecular dynamics-based bootstrap approach is presented which allows for the quantitative assessment of evolutionary relationships inferred from protein structural phylogenies. The novel ability to associate a measure of robustness with inferences allows protein structure-based phylogenies to uncover answers to deep evolutionary questions. In this talk I will present the framework of the method, some results and limitations of this method.

## **Nicholas Matzke**

(ANU and U of Auckland, nick.matzke@anu.edu.au)

### *Large state spaces and state-dependent speciation/extinction models: problems and prospects*

Discrete methods for phylogenetic biogeography, where species' geographic ranges are modelled as a series of presences and absences in discrete areas, have become very popular, but are subject to fairly strict computational limits, because the size of the state space grows at  $2^{(\text{number of areas})}$ , and matrix exponentiation becomes very slow above about 1500 states. In addition, it has become clear that the currently-used biogeography models, such as Dispersal-Extinction-Cladogenesis (DEC) and modifications such as DEC+J (adding jump dispersal), while valid models, are assuming a Yule process – complete sampling and no extinction – a ridiculous assumption. The class of models known as state-dependent speciation (SSE) models can handle extinction, but they use numeric integration to calculate the likelihood and are likely to have even more severe speed (and perhaps accuracy) problems with large numbers of states; however, the question of how speed and accuracy scale with the number of states has not been seriously explored. I present results demonstrating that DEC and DEC+J are special cases of the Cladogenetic SSE (ClASSE) model, explore the scaling issue, and explore options for improving computational speed.

## **Maj Padamsee**

(Landcare Research, padamseem@landcareresearch.co.nz)

### *Fungal endophytes associated with roots of *Agathis australis**

Kauri (*Agathis australis*, Araucariaceae) is restricted in distribution to the Northern tip of the North Island of New Zealand. Living to 1,500 years or more and having trunks up to 3 m diam., *A. australis* exerts enormous influence on surrounding forest composition and structure and provides varying habitat niches for complex fungal communities. However, information on the diversity of fungi associated with *A. australis* is sparse. Additionally, since the 1970s these trees have been under threat from the exotic invasive pathogen, *Phytophthora agathidicida* that causes kauri dieback. Our study aimed to characterise the fungal root endophytic community of *A. australis*. We obtained root samples from three locations and 32 kauri trees in the Waitakere ranges, Auckland. We isolated over 300 cultures from kauri roots and obtained DNA sequences from all of the fungal cultures. We clustered the sequences as OTUs and assigned identities using BLAST. The 354 sequences clustered into 34 OTUs with 63 singletons using a 97% cut-off. We compared the OTUs on the basis of location and disease status. The results give us an insight into the diversity of fungi associated with *A. australis* and suggest the possible impacts of *P. agathidicida* on the fungal community.

## **Charles Semple**

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

### *When is a network captured by its path distances?*

To what extent is an edge-weighted network determined by the path-length distances between its leaves? It is well known that the path-length distances between each pair of leaves of an edge-weighted tree is sufficient to determine (capture) the tree uniquely. This result dates back to Zaretskii (1965) and Buneman (1974), and underlies many widely-used reconstruction methods including Neighbor-Joining. Does this sufficiency extend to networks? In this talk, we explore this question and discuss some recent results for the class of tree-child networks.

## **Chris Simon**

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

### *Parallel, episodic, and spectacular diversification of the microbial endosymbionts of cicadas*

Insects with amino-acid poor diets harbor bacterial endosymbionts that supply needed amino acids and vitamins. In plant-sucking bugs, they often come in pairs. One member of the pair typically produces 8 essential amino acids while the other produces two. In all large plant sucking bugs the slowly evolving dominant endosymbiont is *Sulcia*, but the partner endosymbiont varies among families. In cicadas, it's called *Hodgkinia*. At the NZ phylogenetics meeting last year, I reviewed widespread and unprecedented *Hodgkinia* lineage splitting within individual cicadas across the family Cicadidae. Two genera examined in detail revealed that the splitting was tree-like and was followed by genome degradation/pseudogenization. Multiple species in a single genus of cicadas may or may not share the same set of *Hodgkinia* lineages. New, more detailed within and among population studies include more cicada species in *Tettigades*, and *Magiicada* and two sister species from Australia. We found that some *Hodgkinia* lineage splits persist for 30 to 40



million years while others are very recent and differ among individuals within populations. Within the genus *Tettigades*, we see the parallel evolution of many *Hodgkinia* lineages from one in at least ten different *Tettigades* taxa over the last five or so million years.

## **Jack Simpson**

(University of Canterbury, jrs149@uclive.ac.nz)

### *Combining phylogenetic trees into networks: What can be done with just two trees?*

The study of evolution has been typically conceptualized in simple mathematical structures known as trees. However, it has become increasingly clear that not all evolutionary events can be expressed by trees alone. Non-tree-like evolutionary events, such as hybridization and horizontal gene transfer, demand a more general structure that can express the combination of different trees. Networks are the natural mathematical extension of trees but they come at the cost of the 'nice' tree structure. What is wanted is a network flexible enough to describe all observed evolutionary relationships but still confined by the 'important' properties of trees. In 2015 Francis and Steel introduced the important class of phylogenetic networks known as tree-based networks. A network is tree-based if it can be constructed from an underlying tree by adding only edges between the tree edges. It is known that every vertex in a tree-based network is covered by an underlying base tree. In this talk, I will look at how many trees are required to cover every edge in a tree-based network. Then I will explore what can be done with the sub-class of tree-based networks in which every vertex and edge can be covered by precisely two trees.

## **Mike Steel**

(University of Canterbury, mathmomike@gmail.com)

### *Species notions that combine a phylogenetic tree and phenotypic partitions*

Biologists have long argued as to how best one might define the concept of 'species'. Increasingly, phylogenies have played an important role in this debate. However, phylogeny alone does not suffice, as phenotypes also play an important role (e.g. sister taxa on a tree may be so different that they count as different as species). A recent paper developed a novel approach for describing two well-defined notions of 'species' based on a phylogenetic tree and a phenotypic partition. In this talk, I describe some further combinatorial properties of this approach and describe an extension that allows an arbitrary number of phenotypic partitions to be combined with a phylogenetic tree for these two species notions (joint work with Anica Hoppe and Sonja Türpitz).

## **Marnus Stoltz**

(University of Otago, stoltzstep@gmail.com)

### *Some interesting properties of the Wright-Fisher Diffusion in one-dimension*

The Wright-Fisher model approximates gene frequency in a population of a specific loci comprised of two possible allele types (*A* and *a*). The model (named after Sewall Wright and Ronald Fisher) assumes that generations do not overlap (for example, annual plants have exactly one generation per year) and that each copy of the gene found in the new generation is drawn independently at

random from all copies of the gene in the old generation. The diffusion processes associated to equations of Wright-Fisher type in one spatial dimension has a number of interesting properties. For example, in some cases the associated heat equation on the interval  $[0, 1]$  simply vanishes on the boundary. We consider various aspects of this problem, in particular the boundary conditions that will ensure existence and uniqueness. As well as an analysis of the infinitesimal generators of the  $C_m$ -semigroups and their adjoints.

## **Jeremy Sumner**

(University of Tasmania, jsumner@utas.edu.au)

### *Maximum likelihood distances for genome rearrangement models*

Bacterial evolution is commonly modelled via gene rearrangements, with evolutionary distance taken to be the minimum number of rearrangements needed to convert one genome into another. Recent work has suggested that maximum likelihood estimates (MLEs) of time elapsed are a better proxy for true evolutionary distance. In this talk, I will present results which provide significant theoretical support to this claim. By applying techniques from group representation theory, we have significantly reduced the computational complexity of the required calculations. This has allowed us to compare the properties of minimum distances to MLEs for genomes with up to eleven regions under several distinct rearrangement models.

Joint work with Venta Terauds.

## **Graham Wallis**

(Univ of Otago, g.wallis@otago.ac.nz)

### *Split NZ: phylogeographic breaks in the South Island fauna and their causes*

The beech gap of South island was first documented in 1926 and subsequently generalized to several species. The general pattern is low diversity/endemism in the central (Canterbury) region, separating regions of high diversity to the north and south. Possible causes of this pattern include glaciation (recent) and tectonic rifting (ancient).

A review of 13 genera of alpine/montane birds and insects shows a repeated phylogeographic split between northern and southern forms at about 2 Ma, centred on what was a narrow glacial alpine neck in the early Pleistocene. South Island alpine taxa have preserved a vicariant genetic 'memory' of the first of many glaciations, usually resulting in speciation. The geological neck would have been most intensely glaciated, leaving disjunct refugial populations to north and south. These could have met again during interglacials, perhaps hybridizing on contact, possibly following density trough population dynamics.

Under a typical model of allopatric speciation, mountain ranges separate lowland populations, which evolve into separate lineages either side of, and parallel to the range. In the case of glaciation affecting alpine taxa, phylogeographic breaks occur at right angles (transverse) to the range. This pattern suggests a creative role for glaciation leading to speciation on temperate mountain systems worldwide.

## **Kristina Wicke**

(University of Greifswald, kristina.wicke@web.de)

### *On the Shapley value of unrooted phylogenetic trees*

The Shapley value, a solution concept from cooperative game theory, has recently been considered for both unrooted and rooted phylogenetic trees. Here, we focus on the Shapley value of unrooted trees and first revisit the so-called split counts of a phylogenetic tree and the Shapley transformation matrix that allows for the calculation of the Shapley value from the edge lengths of a tree. We show that non-isomorphic trees may have permutation-equivalent Shapley transformation matrices and permutation-equivalent null spaces. This implies that estimating the split counts associated with a tree or the Shapley values of its leaves does not suffice to reconstruct the correct tree topology. We then turn to the use of the Shapley value as a prioritization criterion in biodiversity conservation and compare it to a greedy solution concept. Here, we show that for certain phylogenetic trees, the Shapley value may fail as a prioritization criterion, meaning that the diversity spanned by the top  $k$  species (ranked by their Shapley values) cannot approximate the total diversity of all  $n$  species.

Joint work with Mareike Fischer.

## **Walter Xie**

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

### *Codon Substitution Model Implementation in BEAST 2*

The codon substitution model is firstly promoted as a more accurate model in the phylogenetic analysis for the evolution of protein-coding DNA sequences by Goldman & Yang 1994, which is a nucleotide substitution model considering the knowledge of the genetic code and synonymous and nonsynonymous substitutions. Then, the set of more general models, M-series (M0-M13), are published by Yang et al. 2000.

M0 and M1 including four types of equilibrium codon frequencies are recently available in a BEAST 2 package. These two models can estimate nonsynonymous/synonymous rate ratio that is very useful to understand selective pressure at the protein level.

## **Jing Yang**

(The University of Auckland and Beijing Normal University, Judyssister@163.com)

### *The global dynamics of avian influenza H9N2 and the influence of poultry production and trade on its spread*

The H9N2 avian influenza virus is considered a potential pandemic viral strain, posing a threat to public health and the poultry industry. However, studies on the underlying factors contributing to the evolution, genetic diversity and migration of H9N2 virus are still lacking. We investigate the spatial distribution of the avian-origin H9N2 virus at a global scale. Using Bayesian Markov Chain Monte Carlo inference framework, we estimated the genetic diversity, evolutionary rates and phylogenetic history of the virus. We used a new accurate approximation to the structured coalescent model to investigate the level of evidence for poultry trade being a driving factor in virus movement. Results indicate that domestic poultry may be the dominant host of H9N2 virus in developing countries below 35degree N, possibly resulting from poor poultry management prac-

tices. The phylogeny of H9N2 groups into three clusters which are driven by geographic origins. Higher evolution rates and population sizes of the virus occur in Asia, where domesticated chicken are thought to be the prevailing host sustaining viruses persistence and evolution. And we need to further evaluate the evidence for poultry trade as the driver of the global spatial distribution of H9N2 virus.

## 5 Participants

Remco Bouckaert	remco@cs.auckland.ac.nz
David Bryant	dbryant@maths.otago.ac.nz
Michael Charleston	michael.charleston@utas.edu.au
Benny Chor	benny@cs.tau.ac.il
Jordan Douglas	jdou557@aucklanduni.ac.nz
Alexei Drummond	alexei@cs.auckland.ac.nz
Mareike Fischer	email@mareikefischer.de
Paul Gardner	paul.gardner@canterbury.ac.nz
Alexandra Gavryushkina	gavryushkina@gmail.com
Alex Gavryushkin	alex@gavruskin.com
Russell Gray	gray@shh.mpg.de
Stefan Grünewald	stefan@picb.ac.cn
Momoko Hayamizu	hayamizu@ism.ac.jp
Mike Hendy	mhendy@maths.otago.ac.nz
Lina Herbst	lina.herbst@uni-greifswald.de
Barbara Holland	barbara.holland@utas.edu.au
Daniel Huson	daniel.huson@uni-tuebingen.de
Jonathan Klawitter	jo.klawitter@gmail.com
Denise Kühnert	denise.kuehnert@gmail.com
Simone Linz	s.linz@auckland.ac.nz
Peter Lockhart	p.j.lockhart@massey.ac.nz
Catherine Macken	c.macken@auckland.ac.nz
Ashar Malik	a.j.malik@massey.ac.nz
Nick Matzke	nick.matzke@anu.edu.au
Roman Oliynyk	roli573@aucklanduni.ac.nz
Maj Padamsee	padamseem@landcareresearch.co.nz
Adam Powell	powell@shh.mpg.de
Charles Semple	charles.semple@canterbury.ac.nz
Chris Simon	chris.simon@uconn.edu
Jack Simpson	jrs149@uclive.ac.nz
Mike Steel	mathmomike@gmail.com
Marnus Stoltz	stoltzstep@gmail.com
Jeremy Sumner	jsumner@utas.edu.au
Leonna Szangolies	p.j.lockhart@massey.ac.nz
Lydia Turley	lydiamturley@gmail.com
Graham Wallis	g.wallis@otago.ac.nz
Bevan Weir	weirb@landcareresearch.co.nz
Kristina Wicke	kristina.wicke@web.de
Walter Xie	walter@cs.auckland.ac.nz
Jing Yang	Judyssister@163.com
Rong Zhang	rzha419@aucklanduni.ac.nz

*Conference organisers:* David Bryant, Mike Hendy, Marguerite Hunter