Phylomania 2013

The year we moved to a three page program

University of Tasmania, School of Mathematics and Physics, 6-8 November

Public Talk

Wednesday, 6 November, 7pm, Stanley Burbury lecture theatre

Mike Steel, University of Canterbury, New Zealand
Mathematical Challenges in Finding the Tree of Life

Program

Thursday, 7 November

8.15am-9:00am	Registration and coffee
9:00am-9:10am	Welcome
9:10am-9:30am	**Andrew Francis**, University of Western Sydney, Australia *The inversion process in bacteria: distance metrics with group-theoretic models*
9:30am-9:50am	**Attila Egri-Nagy**, University of Western Sydney, Australia *Highways and byways in group-theoretic genome space*
9:50am-10:10am	**Michael Charleston**, University of Sydney, Australia *Fast Cophylogeny Mapping and Widespread Parasites*
10:10am-10:30am	**Greg Butler**, Concordia University, Canada *Using synteny in phylogenomics to resolve orthologs and paralogs*
10:30am-11:00am	Morning tea
11:00am-11:40am	**Mike Steel**, University of Canterbury, New Zealand *Ancestral reconstruction, lateral gene transfer, and the joys of leaping between trees*
11:40am-12:00pm	**Laura Boykin**, University of Western Australia *A practical guide to HPC Bayesian phylogenetic analyses: More chains or more replicates? GPU vs. CPU?*
12:00pm-12:20pm	**Rob Lanfear**, Australian National University *Estimating the Effective Sample Size of phylogenetic tree topologies from Bayesian MCMC analyses*
12:20pm-1:40pm	Lunch

1:40pm-2:00pm	**Sangeeta Bhatia**, University of Western Sydney, Australia *Group-theoretic formalization of the Double-cut-and-join model of chromosomal rearrangement*
2:00pm-2:20pm	**Stuart Serdoz**, University of Western Sydney, Australia *Studying bacterial inversion with random walks on groups*
2:20pm-2:40pm	**Stephen Crotty**, University of Adelaide, Australia *Model Misspecification due to Site Specific Rate Heterogeneity: how is tree inference affected?*
2:40pm-3:00pm	**Jonathan Mitchell**, University of Tasmania *Distinguishing convergence events on phylogenetic networks*
3:00pm-3:30pm	Afternoon Tea
3:30pm-3:50pm	**Michael Woodhams**, University of Tasmania *Hybridization Networks and Sequence Simulation*
3:50pm-4:10pm	**Tristan Stark**, University of Tasmania *Models of microsatellite evolution*
4:10pm-4:30pm	**Bennet McComish**, University of Tasmania *Microsatelite evolution in ancient and modern penguins*
4:30pm-4:50pm	**Subha Kalyaanamoorthy**, CSIRO Ecosystem Sciences, Canberra, Australia *Reconstructing ancestral sequences through a combined bioinformatics and molecular modelling approach*
7:30pm	**Phylomania 2013 dinner:** *Ciuccio, Salamanca Square*

Friday, 8 November

8:30am-9:10am	Coffee
9:10am-9:30am	**Lars Jermiin**, CSIRO Ecosystem Sciences, Canberra, Australia *Mixture models of nucleotide sequence evolution, and the evolution of yeast genomes*
9:30am-9:50am	**Thomas Wong**, CSIRO Ecosystem Sciences, Canberra, Australia *Augmenting phylogenetic data: Are we approaching the brick wall?*
9:50am-10:10am	**Greg Jordan**, University of Tasmania *Directional evolution of cell size in Proteaceae*
10:10am-10:30am	**Xia Hua**, Australian National University *Tracking the formation of a species assemblage over time: phylogenetic reconstruction of patterns of colonisation and speciation.*
10:30am-11:00am	Morning tea

11:00am-11:40am	**David Liberles**, University of Wyoming, USA *Mechanistic Models in Comparative Genomics*
11:40am-12:00pm	**Barbara Holland**, University of Tasmania *Assessing the fit of open and closed models of bacterial genome evolution*
12:00pm-12:20pm	**David Penny**, Massey University, New Zealand *Loss of information at deeper divergences*
12:20pm-1:40pm	Lunch and Poster Session **Guillaume Bernard, Cheong Xin Chan** University of Queensland, Australia *Inferring phylogenies with alignment-free methods*
1:40pm-2:00pm	**Karen Meusemann**, CSIRO Ecosystem Sciences, Australian National Insect Colletion, Canberra, Australia (on the behalf of the 1KITE consortium) *1000 Insect Transcriptomes – taking the next step in insect phylogenomics*
2:00pm-2:20pm	**Jeremy Sumner**, University of Tasmania *The squangles: The gift that just keeps on giving*
2:20pm-2:40pm	**Gillian Gibb**, Massey University, New Zealand *Why Fly When You Can Walk? The Evolution Of Flightlessness In Rails*
2:40pm-3:00pm	**Simon Hills**, Massey University, New Zealand *Marrying molecules and morphology in marine molluscs*
3:00pm-3:30pm	Afternoon Tea
3:30pm-3:50pm	**Chris Burridge**, University of Tasmania *Presence of a cryptic hybrid zone explains spatial variability in population genetic structuring of a colonial nesting seabird*
3:50pm-4:10pm	**Dorothy Steane**, University of Tasmania *Phylogenetic patterns of reproductive isolation in Eucalyptus*
4:10pm-4:30pm	**Catherine Byrne**, Tasmanian Museum and Art Gallery *Philosophical basis of modern phylogenetic inference. Do we really know what we are doing?*
4:30pm-4:50pm	**Barbara Schoenfeld**, University of Tasmania *Looking for patterns in the retention and loss of plastid genes*
4:50pm-6pm	Drinks and cheese

Saturday, 10 November

12pm-	**Phylomania 2013 bushwalk** (details TBA)

Abstracts

Guillaume Bernard, Cheong Xin Chan
University of Queensland, Australia

Inferring phylogenies with alignment-free methods (poster)
Joint work with Olivier Porion, James Hogan and Mark Ragan

Phylogenetic inference of biological sequences is usually based on multiple sequence alignment, in which whole sequences are aligned to the conserved positions across the set. Based on the implicit assumption of full-length contiguity of homologous sequences, this approach is sensitive to genetic recombination, shuffling and lateral genetic transfer, resulting in the loss of phylogenetic information. Alignment-free methods, in which short subsequences at a predefined length (e.g. k-mers, instead of whole sequences) are used for reconstructing sequence distance matrices, have recently been used to infer phylogenies and sites of lateral genetic transfer. However, the scalability and robustness of these methods remain to be investigated. Here, using simulated sequence data, we systematically assess the accuracy of phylogenetic inference using an alignment-free approach (based on similarity measure of k-mers using D_2 statistics) across three key evolutionary aspects: (a) sequence divergence, (b) among-site rate heterogeneity, and (c) G+C content biases, each across gene-length sequence sets of various sizes (8, 32 or 128). In comparison to phylogenies generated from the conventional alignment-based approach, we found that the alignment-free approach recovered accurate tree topologies at equal proportion in several cases, and at lower computational cost. Moreover, across published empirical data, performance of alignment-free approach appears promising. While the robustness of the alignment-free approach remains to be investigated across genome-scale data, our findings demonstrate the scalability and the potential use of alignment-free methods in large-scale phylogenomics.

Sangeeta Bhatia
University of Western Sydney, Australia

Group-theoretic formalization of the Double-cut-and-join model of chromosomal rearrangement

Establishing the minimal distance between genomes is a significant problem in computational genomics, as the distance is often used to establish evolutionary relationships including phylogeny. The "double cut and join" (DCJ) model of chromosomal rearrangement proposed by Yancopoulos *et. al.* [2005] has received attention as it can model inversions, translocations, fusion and fission on a multichromosomal genome which may contain both linear and circular chromo- somes. We realize the DCJ operator as a group action on the space of multichromosomal genomes. We study this group action, deriving some properties of the group and finding group-theoretic analogues for the key results in the DCJ theory. Our work translates the problem of finding the rearrangement distance between genomes into a problem in group theory.

Laura Boykin
University of Western Australia

A practical guide to HPC Bayesian phylogenetic analyses: More chains or more replicates? GPU vs. CPU?
Joint work with Peter Beerli, Ian Small, and Christopher Harris

Researchers with large genomic datasets are often faced with running phylogenetic analyses on high performance computing (HPC) systems. For example, biosecurity decisions often rely on identifying invasive species based on phylogenetic information and the decision must be made quickly (food spoilage). HPC systems can be used to aid in these invasive species identifications but the options available on the HPC machines can be daunting to those who are not familiar with the HPC architecture. The goal of this project is to provide useful information for integrating HPC architecture with common phylogenetic programs, such as MrBayes. We will assess whether more chains (heating) or more replicates is the best strategy to search the tree space. We will run these analyses on several different HPC platforms that might be available to researchers. The HPC facilities at iVEC will be utilized as we have access to a GPU machine (Fornax), CPU machine with the Infiniband switches (Epic) and a new CPU machine with the dragonfly interconnect switches (Magnus).

Chris Burridge
University of Tasmania

Presence of a cryptic hybrid zone explains spatial variability in population genetic structuring of a colonial nesting seabird

Factors responsible for spatial structuring of genetic variation among populations are varied, and most explanations centre on mechanisms that influence contemporary gene flow. However, in some instances there may be no obvious explanations for genetic structuring observed, or those invoked may reflect spurious correlations. A previous study of little penguins (*Eudyptula minor*) in southeast Australia documented low spatial structuring of genetic variation with the exception of colonies at the western limit of sampling, and their distinction was attributed to an intervening oceanographic feature or differences in breeding phenology. We conducted sampling across the entire Australian range, employing additional markers (12 microsatellites and mitochondrial DNA, 697 individuals, 17 colonies). The zone of elevated genetic structuring previously observed actually represents the eastern half of a hybrid zone, within which allele frequencies vary over much shorter spatial scales than elsewhere. Colonies separated by as little as 27 km in the zone are genetically distinguishable, while outside the zone homogeneity cannot be rejected at scales of up to 1400 km. The fine scale of genetic structuring within the zone, accompanied by a lack of environmental differences, linkage disequilibrium, or elevated inbreeding coefficient (F_{IS}), suggest neutral introgression following recent secondary contact of lineages, with gene flow yet to establish equilibrium allele frequency differences among colonies. This study highlights the potential contribution of hybrid zones to genetic structuring among populations, and the importance of sampling scale to detect these possible effects.

Greg Butler
Concordia University, Canada

Using synteny in phylogenomics to resolve orthologs and paralogs
Joint work with Christine Kehyayan

A long-standing problem in the functional annotation of proteins is to distinguish orthologs from paralogs. This problem spawned the field of phyogenomics, which applies phylogenetic techniques to address the problem. Now in the post-genomics era, we have the context of complete genomes in which to work, so we have investigated the use of genomic context, that is, synteny, to better resolve orthologs and paralogs.

We present our algorithm to determine "syntenic reciprocal best hits", a modified edge weight for similarity graphs, and the results of the application of the MCL Markov clustering algorithm to determine orthologous clusters across eight well-studied fungal genomes.

Catherine Byrne
Tasmanian Museum and Art Gallery

The philosophical basis of modern phylogenetic inference. Do we really know what we are doing?

"If science is not to degenerate into a medley of ad hoc hypotheses, it must become philosophical and must enter into a thorough criticism of its own foundations."
Alfred North Whitehead (1925: 25),
Science and the Modern World

The science of systematics over the last 20 years or so has undergone a revolution, in which vast amounts of molecular data have become readily available. Phylogenetic inference has now become the domain of mathematical algorithms such as maximum parsimony, maximum likelihood and Bayesian inference. Here I discuss critically and briefly current methods of testing systematic hypotheses. I will do this by scrutinising the philosophical basis of phylogenetic inference. Do we really know what we are doing?

Mike Charleston
University of Sydney, Australia

Fast Cophylogeny Mapping and Widespread Parasites

Cophylogeny mapping is the most intuitive method of reconciling evolutionary trees of ecologically linked groups of species. Hosts and parasites, hosts and pathogens, genes and species: answering questions about their coevolution is best done using a process of mapping one tree, the dependent P, into the other, the independent H, preserving what we already know about the associations of the two groups of extant taxa and minimising an overall cost function that is based on a small set of recoverable events.

Of course, the problem is hard. NP-Hard, in terms of the optimisation problem. The complexity arises as soon as host switches are permitted: without them, the problem can be solved in linear time, but with them, and if the divergence times in H are not fixed, the decision problem becomes NP-Complete.

But that's no excuse, we have to find solutions anyway. This talk is about how to do cophylogeny mapping, fast.

I will talk about new (and relatively new) methods of mapping P into H that take very

little time ($O(n \log n)$) and still get the optimal solution about 85% of the time on published data, and a method that builds on the method used in the program Jane (Versions 1-4; Libeskind-Hadas and colleagues) which maps nodes in P to edges in H to search the space of host timings and uses mapping of parasite nodes to host nodes. Mapping to host nodes permits solutions of the wide-spread parasite problem to be solved in $O(n^3)$ time if the host node timings are fixed, which means a long-standing problem in cophylogenetics, of how to deal with parasites occupying more than one host species, can be solved in reasonable time.

Stephen Crotty
University of Adelaide, Australia

Model Misspecification due to Site Specific Rate Heterogeneity: how is tree inference affected?

Phylogeneticists have long recognised that different sites in a DNA sequence can experience different rates of nucleotide substitution, and many models have been developed to accommodate this rate heterogeneity. But what happens when a single site exhibits rate heterogeneity along different branches of an evolutionary tree?

In this talk I'll discuss the notion of Site Specific Rate Heterogeneity (SSRH) and investigate a simple case, looking at the impact of SSRH on inference via maximum parsimony, neighbour joining and maximum likelihood.

Attila Egri-Nagy
University of Western Sydney, Australia

Highways and byways in group-theoretic genome space
Joint work with Andrew Francis and Volker Gebhardt

Reconstructing phylogenetic trees requires establishing the distance between a pair of genomes by finding a shortest path between them. This distance metric provides a good first approximation, but other factors may also be part of the 'real distance'. For instance, the 'width' (the number of shortest paths) could also influence the course of a random or selection process. If genomes are represented by permutations, then the genome space becomes the Cayley graph of the group generated by evolutionary events and our question turns into an investigation of the set of all geodesic paths leading to a group element g. Using the prefix partial order (the weak order in Coxeter terminology), the set of group elements visited by geodesics forms an interval $[1, g]$. These intervals are graded partial orders with top and bottom elements, although they are not lattices in general. In this talk we describe how by using these intervals we can define refined equivalence relations that can further distinguish between elements of the same length.

Andrew Francis
University of Western Sydney, Australia

The inversion process in bacteria: distance metrics with group-theoretic models

The inversion process in bacteria is often used as a means to estimate evolutionary distance, avoiding interference from horizontal gene transfer. It can also be described using models

that involve group theory because each inversion can be regarded as an invertible action on the genome. The use of group theory brings with it an understanding of the process as a random walk on a Cayley graph, and the subsequent question of whether the minimal distance is the best metric after all.

Gillian Gibb
Massey University, New Zealand

Why Fly When You Can Walk? The Evolution Of Flightlessness In Rails
Joint work with Steve Trewick

It is a curious, even paradoxical, phenomenon that many birds are flightless, when for most of us birds are quintessentially flyers. For a flightless species to evolve from a flying ancestor there must be good ecological reasons and pathways for selection: options for being flightless must exist. The rails (Rallidae) are a cosmopolitan group of birds with a broad distribution through the world. Flightlessness has evolved independently multiple times within the group, often in association with the colonisation of islands and there are numerous examples in the New Zealand/Pacific region. Nearly all islands of the Pacific have been colonized by one or more rail lineages, and before human contact it is likely that most islands had endemic flightless rails. Evolution of flightlessness in rails has been shown to occur rapidly, and may be measured in 'generations rather than millennia'. Therefore rails are a highly appropriate case study of adaptive evolution, and an excellent group for the comparative analyses of flightlessness. While ecological reasons for loss of flight may be obvious, what sort of genetic adaptations might be underlying the evolution of flightlessness? We use comparative genomics of developmental and metabolic genes, coupled with a temporal-spatial framework to investigate the evolution of flightlessness in New Zealand rails.

Simon Hills
Massey University, New Zealand

Marrying molecules and morphology in marine molluscs
Joint work with James Crampton and Barbara Holland

Proper consideration of the evolution of life on earth must make a suitable phylogenetic evaluation of the 99% of organisms that are now extinct. Due to the lack of preserved DNA for the overwhelming majority of extinct organisms, the only data by which these evolutionary relationships can be assessed is morphological. However, independent phylogenetic analysis of molecular and morphological data usually reveals substantially different signals. Poor resolution, convergence and ecophenotypic variability plague phylogenetic reconstructions where molecular and morphological based analyses are compared.

Using robust molecular and morphometric datasets we explore these issues in the New Zealand marine mollusc genus *Alcithoe*. In evaluating disagreement between molecular and morphological data in *Alcithoe* we found that an ecological variable, maximum habitat depth of species, is correlated with a significant conflicting signal in the morphological dataset. We then examined if a phylogeny that is more similar to the molecular based reconstruction can be generated from the morphological data after filtering out characters that correlated with water depth.

Through this examination of molecular and morphological data with extant species we aim to identify an optimal set of morphological characters that will enable more accurate phylogenetic reconstruction and the inclusion of extinct species.

Barbara Holland
University Tasmania, Australia

Assessing the fit of open and closed models of bacterial genome evolution
Joint work with Nigel French, Patrick Biggs, Shoukai Yu

Multi-locus sequence typing (MLST) has long been a popular means of typing bacterial species. Traditionally 7 housekeeping genes are sequenced and each strain of bacteria is summarized by 7 numbers recording which unique sequence the strain has for each of the genes. With the advent of cheap genome sequencing this concept can easily be extended to larger numbers of genes giving allele profiles that may relate to many 100s of genes. The allele profile data has some interesting properties from a phylogenetic standpoint.

Bacteria evolve by a mixture processes that act vertically on the tree such as inheritance and mutation as well as processes that act horizontally across the tree. A mutation event or recombination event both have the same effect on the allele profile of changing the unique identifier for one of the genes. One question that arises is the extent to which the allele profile data fits the infinite alleles model. If sequences for each gene are long then mutation should usually introduce a sequence type that has not been seen before. With recombination it will depend on the extent to which the bacteria under consideration form a closed or open system.

In this talk I investigate the fit of closed and open models to allele profile data derived from 46 *Campylobacter jejuni* genomes.

Xia Hua
Australian National University

Tracking the formation of a species assemblage over time: phylogenetic reconstruction of patterns of colonisation and speciation.

Studies of temporal patterns in the formation of species assemblages have generally focussed on taxa with a rich, continuous fossil record, or on extant species in islands of habitat with known histories. Given the growing availability of molecular and phylogenetic information for ever-increasing numbers of species, we should be able to extend the phyloinformatic approach to macroecology and macroevolution to a wider range of species assemblages. Here, we suggest some methods for estimating the timing of addition of species to assemblages which may prove useful. We show how whole-assemblage phylogenies estimated from publicly available data can be used to describe temporal patterns of the addition of species to assemblages, through both colonization and in situ speciation. The advantages of this approach is that it accounts for uncertainty in phylogenetic estimation and for uncertainty in deriving dates of biological events from nodes and branchlengths, it explicitly models sampling bias associated with detecting more recent speciation or colonization events, and it formally considers patterns of colonization and speciation within a hypothesis testing framework. We road test this approach on two data sets, New Zealand passerines and Madagascar squamates.

Lars Jermiin
CSIRO Ecosystem Sciences, Canberra, Australia

Mixture models of nucleotide sequence evolution, and the evolution of yeast genomes

Molecular phylogenetic studies of homologous sequences of nucleotides often assume that the evolutionary process was globally stationary, reversible, and homogeneous (SRH), and that the data can be modelled accurately using one or several site-specific, time-reversible rate matrices. However, a growing body of data suggests that evolution under globally SRH conditions is an exception, rather than a norm. To address this issue, we introduce a family of mixture models that considers heterogeneity in the substitution process across lineages (HAL) and heterogeneity in the substitution process across sites (HAS). We also introduce an algorithm for searching model space and identifying a model of evolution that is less likely to over- or under-parametrize the data. The merits of our algorithms are illustrated with an analysis of 42,337 second codon sites extracted from a concatenation of 106 alignments of orthologous genes encoded by the nuclear genomes of eight species of yeast. For this data set, our HAL-HAS model fits the data better than other models do. Parameter estimates for this model indicate not only a complex ancestral sequence but also a complex evolutionary process.

Greg Jordan
University of Tasmania

Directional evolution of cell size in Proteaceae
Joint work with R. J. Carpenter and T. J. Brodribb

Recently we showed evolutionary associations between cell sizes in different leaf tissues in Proteaceae. In some clades the cells of epidermis, mesophyll and xylem are all relatively large, in others these cells are relatively small. Furthermore, these associations appear to be ancient, reaching in some cases deep into the Cretaceous. Two complementary mechanisms underlie these links: a strong developmental link between genome size and cell size; and an adaptive link related to ecology - large cells are associated with open environments and small cells with closed forest. This adaptive link relates to functional co-ordination, with correlated cell sizes in disparate tissues leading to efficient allocation of construction costs. This link is particularly apparent for stomata and leaf veins. Small stomata create high maximum water demand, at least when densely packed. This demand needs to be supplied by small veins, which allow for a greater leaf-wide capacity to carry water.

Complicating our understanding of the significance of cell size are conflicting theories. One theory predicts that genome size (and therefore cell size) may have increased systematically through time. An alternative view is that cell size may have decreased with global drops in $p\mathrm{CO}_2$. We present phylogenetic and fossil evidence for the evolution of stomatal size in Proteaceae. Ancestral state analyses suggest that stomata in the Cretaceous and Paleogene were large ($\sim 45\mu$m long) and increased in a few clades of open environments and decreased in other clades, especially those of rainforest. The fossil evidence shows a general trend of increased stomatal size, both overall and within clades. The fossils show smaller stomatal sizes for ancient taxa than predicted by the ancestral state analyses. Thus, the fossil evidence is consistent with the idea of directional evolution of increasing stomatal and genome size, and contradicts both the ancestral state reconstructions and the CO_2 model.

Subha Kalyaanamoorthy
CSIRO Ecosystem Sciences, Canberra, Australia

Reconstructing ancestral sequences through a combined bioinformatics and molecular modelling approach

Tracing the molecular trails of substitutions between extinct and extant forms of life, in order to understand the underlying evolutionary processes, is often important to address the most intriguing questions of the present. Ancestral sequence reconstruction (ASR), a computational method that infers the sequence of ancestors using the phylogenetic relationships between the extant sequences, has become one of the most popular approaches to restore ancestral states. ASR has proved to be successful in different scientific areas: from protein engineering and pharmaceutical applications to the paleo-sciences. In this presentation, we will discuss the benefits and challenges faced when doing ASR and how molecular modelling approaches can effectively corroborate with the former in the identification of the most probable ancestral character states, which then later can be resurrected in laboratories. Preliminary results from the ancestral states of different enzymes reconstructed by our research team will be presented.

Rob Lanfear
Australian National University

Estimating the Effective Sample Size of phylogenetic tree topologies from Bayesian MCMC analyses

In MCMC analyses, the number of independent samples from a chain is often lower than the total number of samples, because sequential samples from the MCMC can be autocorrelated. The Effective Sample Size (ESS) is a useful statistic when using or developing MCMC methods, because it estimates the number of independent samples we have from a given chain after accounting for autocorrelation. The ESS can therefore be used to assess the adequacy of Bayesian MCMC analysis, or to compare different MCMC approaches to the same problem. Standard methods exist to calculate the ESS for continuous parameters, and these are widely used in Bayesian phylogenetics. However, no methods exist for calculating the ESS of tree topologies, despite the fact that tree topologies are often of primary interest to those who use MCMC methods for phylogenetics. Here, I propose and compare two ways to estimate the ESS of tree topologies from MCMC analyses. This is very preliminary work and all comments and criticisms are very welcome.

David Liberles
University of Wyoming, USA

Mechanistic Models in Comparative Genomics

Computational genomics is now generating very large volumes of data that have the potential to be used to address important questions in both basic biology and biomedicine. Addressing these important biological questions becomes possible when mechanistic models rooted in biochemistry and evolutionary/population genetic processes are developed. Examples are described on problems involving the inference of duplicate gene retention mechanisms to apply in a gene tree/species tree reconciliation setting and in building a model for the parameterization of amino acid transition selective coefficients in protein evolution. A further brief discussion of differentiating between functional shifts and compensatory covariation at

different levels of biological organization is entertained. This description of mechanistic models is ultimately generalized towards future developments in computational genomics and the need for biological mechanisms and processes in biological models. Issues raised for discussion include how to validate models for mechanistic inference and how to insure that coarse graining of processes in simplifying models does not affect inference, even if the models fit the data well.

Bennet McComish
University of Tasmania

Microsatellite evolution in ancient and modern penguins

Microsatellites are short tandem repeat sequences that have been widely used as genetic markers in a variety of population genetic studies. The number of repeats at a locus is thought to change by slippage of DNA polymerases during replication, and these loci exhibit high levels of length polymorphism. Several models of microsatellite evolution have been developed, some of which take into account both replication slippage and point mutation. These models vary widely in complexity, but it is not clear that any of them succeed in capturing all of the relevant biological processes. I will present some new results from modern and ancient genome sequences of Adelie penguins, and discuss how these can be used in testing existing models of microsatellite evolution and developing new ones.

Karen Meusemann
CSIRO Ecosystem Sciences, Australian National Insect Colletion, Canberra, Australia
(on the behalf of the 1KITE consortium)

1000 Insect Transcriptomes – taking the next step in insect phylogenomics

The 1KITE project (1K Insect Transcriptome Evolution) is an international research initiative that aims to study all aspects of the evolutionary history of insects. To achieve this aim, scientists in the 1KITE project are sequencing and analysing the transcriptomes of more than 1,000 insect species encompassing all recognised extant insect orders. The project focuses on inferring robust phylogenetic trees for all major lineages of hexapods, but also encourages the development of new and advanced methods and software tools for data quality assessment, phylogenetic reconstruction, and molecular dating, and paves the way for numerous additional projects that will be feasible with the generated transcriptome data. Sequencing of the transcriptomes and sequence assembly was enabled and carried out by the BGI, Shenzhen, China. Altogether, 70 scientists from 8 nations – experts in insect morphology, taxonomy, systematics, paleontology, embryology, molecular biology, bioinfor matics, and scientific computing – are collaborating in 1KITE.

In this talk, a short introduction to the project will be given. The pipeline for phylogenetic analyses which we developed a first project that aims at inferring a phylogenetic backbone tree for the entire hexapod clade comprising around 150 taxa and 1,500 orthologous genes will be shortly outlined. This pipeline will be further developed and improved for subsequent analyses of taxonomic subgroups with up to 5,000 orthologous genes and finally analysing more than 1,000 insect transcriptomes. We focused on quality assessment of various steps in the analysis, e.g. alignment refinement and alignment masking, selection of "decisive datasets", or advanced data partitioning. In addition, we applied some rather rarely-used approaches in phylogenetic analyses, e.g. the use of protein domains for partitioning the data, and testing single phylogenetic hypotheses separately using "Four-

cluster Likelihood Mapping" (FcLM), introduced in the 1990's, in addition to common tree reconstruction methods. The talk will focus on the application of the FcLM approach on large-scaled datasets which aims to identify potential alternative signal that might not be obvious from common tree reconstruction methods with multiple taxa. Furthermore, we worked out approaches that try to cope more cautiously to discrimitate potential phylogenetic from other signal using large datasets. We expect to have an long-standing impact on entomological and phylogenetic research, however showing there is still plenty room for improvements on various levels of analyses.

Further information: `www.1kite.org`

Jonathan Mitchell
University of Tasmania

Distinguishing Convergence Events on Phylogenetic Networks

I will investigate whether small trees (two or three taxa) can be distinguished from more general networks which allow for convergence events between taxa (eg. two taxa may begin to converge due to hybridisation) by examining their probability distributions and the probability spaces they map to. I will use methods from algebraic geometry, in particular Groebner bases, to solve the polynomial equations which define the probability distributions.

David Penny
Massey University, New Zealand

Loss of information at deeper divergences

It has been shown by Mossel and Steel that simple Markov models lose information at the deepest divergences; and that the fall off is exponential at deeper times. However, that does not mean that there is no information left, for example, the three-dimensional structure of proteins may still retain information at deeper divergences. Biologists still want to estimate these deeper divergences and thus it is a significant question to find other sources of information. Several suggestions are offered for a formal analysis. Firstly, we probably expect that where there is a real Gamma distribution of rates that information may be retained for longer. Secondly, the inference of ancestral sequences at deeper divergences appears quite robust, and there is some evidence that this may help recover deeper divergences. Thirdly, it is increasingly possible to infer three-dimensional structures, and these may retain information longer. Fourthly, an approach of weighting, not of characters, but of the partitions they are consistent with, might help. Fifthly, possibly gene order information might be helpful. Several examples of such approaches will be presented, and a challenge issued to solve some of these fundamental issues.

Barbara Schoenfeld
University of Tasmania

Looking for patterns in the retention and loss of plastid genes

Over the course of plastid evolution most genes present in the common ancestor of these photosynthetic organelles have been lost from the plastids' genomes. This process of gene loss has resulted in quite diverse patterns of gene presence and absence in the different

lineages. However, the selection pressures that govern the loss or retention of genes in the organelle remain unknown. I present here the results of my attempts to identify pairs of genes whose presence or loss are coordinated, based on a data set compiled from 198 fully sequenced plastid genomes and a reconstruction of their gene loss history. This will hopefully help to identify functional associations between genes that underpin gene loss dynamics.

Stuart Serdoz
University of Western Sydney, Australia

Studying bacterial inversion with random walks on groups

Applying parsimony to the inversion distance problem involves using the least number of inversions necessary to explain the difference between two genomes. We investigate this principle by modelling the inversion process as a random walk on the Cayley graph of the group that arises from the model. As expected, for small distances parsimony is a fine assumption however this gets worse however as the distances get longer. We derive some properties of the longer term behaviour of the model, including the mixing time and the expected distance in the limit.

Tristan Stark
University of Tasmania

Models of Microsatellite evolution

Microsatellites are DNA repeat sequences which are frequently used as genetic markers due to their high levels of polymorphism and relative abundance. In order to make accurate inferences using microsatellite data it is necessary to have mathematical models which describe their evolution in time. The most frequently used models are continuous-time Markov chains with a one-dimensional state space, these only track one feature of their evolution, the repeat number. It is well established that point mutations play an important role in the evolution of microsatellites, but current models are limited in the way they handle such mutations. We propose a model with a two-dimensional state space that tracks both the repeat number and the buildup of impurities due to point mutations.

Dorothy Steane
University of Tasmania

Phylogenetic patterns of reproductive isolation in Eucalyptus
Joint work with Matthew Larcombe, Rebecca Jones, Dean Nicolle, Barbara Holland, René Vaillancourt, Brad Potts

Reproductive isolation is a fundamental characteristic of speciation. This study assesses the phylogenetic basis of postmating reproductive isolation in *Eucalyptus*. *Eucalyptus globulus* pollen was applied to 100 eucalypt species from 13 taxonomic sections, mainly from the most speciose and commercially important subgenus, *Symphyomyrtus*. Data on pre-dispersal (the number of hybrids produced) and post-dispersal (their survival at one year) compatibility were combined with phylogenetic data (based on over 8000 DArT markers) to identify patterns in hybridisation, and assess the relative importance of pre- and post-dispersal barriers

to reproductive isolation. Clade I (including *E. globulus*) and clade II are closely related and had a higher proportion of hybridising taxa (92%) than the phylogenetically more distal clades III and IV (10%), suggesting a significant reproductive barrier. In general, hybrid compatibility declined with increasing genetic distance. However, pre-dispersal compatibility showed a leptokurtic decay, declining rapidly and then slowing, while post-dispersal compatibility declined slowly in a more linear fashion. The results indicate that both pre- and postzygotic barriers are influencing postmating isolation, and may be driven by different processes, possibly including both natural selection and drift. Comparing the results with the most recent dated phylogeny indicates that complete reproductive isolation may take 21-31 mya in Eucalyptus. The study has practical implications for tree breeding, and for quantifying the genetic risk that *E. globulus* plantations pose to indigenous *eucalypt* populations in Australia.

Mike Steel
University of Canterbury, New Zealand

Ancestral reconstruction, lateral gene transfer, and the joys of leaping between trees

In part 1, I will present some recent results with Olivier Gascuel on how accurately we can expect to predict ancestral states at the interior nodes of a tree.

In part 2, I will describe a second project on species tree reconstruction when genes have evolved under a model of random lateral gene transfer (LGT). A typical question is: 'could we reconstruct a species tree on (say) 200 species from lots of gene trees, if each gene has been laterally transferred into other lineages, on average, ten times?' Another is 'can LGT lead to inconsistent tree estimation?' Our analysis involves a curious connection to random walks on cyclic graphs.

Jeremy Sumner
University of Tasmania

The squangles: The gift that just keeps on giving

The squangles are a set of (two, three or four?) *Markov invariants* valid for quartet phylogenetic trees under the general Markov model. These invariants are homogeneous degree 5 polynomial functions of phylogenetic sequence data comprising 66,744 monomial terms. For a given quartet topology these functions can be organised such that two of the four form *phylogenetic* invariants (i.e. vanish identically on "perfect" data for that tree), and hence can be utilized to extract evolutionary relationships. Their existence and explicit polynomial form has been established since 2008(ish), with the implementation of a least squares method presented in Holland *et. al.* (2012). I will discuss the remarkable properties of this least squares method, and clarify (for the first time) where the squangles fit into the general scheme of phylogenetic invariants (as studied by Allman, Rhodes, Eriksson, Darisma *et. al.*). Remarkably, it turns out the squangles occur as "tripod" invariants and hence, as phylogenetic invariants, can be understood using the general scheme of discrete symmetries I discussed at Phylomania 2012.

Michael Woodhams
University of Tasmania

Hybridization Networks and Sequence Simulation

I have a program to generate random phylogenetic trees with hybridization and introgression events. The program can simulate Dollo process data on the network, or invoke the Filo sequence simulation program to simulate DNA or amino acid sequences. The program has many configuration options, including restricting hybridizations to closely related species, and allowing speciation and hybridization rates to vary with time.

Thomas Wong
CSIRO Ecosystem Sciences, Canberra, Australia

Augmenting phylogenetic data: Are we approaching the brick wall?

There are three ways to augment phylogenetic data: we can increase the number of sequences in the alignment, we can increase the number of sites in the alignment, or we can increase the number of sequences as well as the number of sites in the alignment. The first two options have drawbacks over the third option. In this seminar, I focus on phylogenetic accuracy when the number of taxa increases. Although it is widely believed that increasing the number of taxa can improve the phylogenetic accuracy, our results, which are consistent with a conjecture published in Syst. Biol. (53: 638-643), show that (i) the average internal edge length decreases as the number number of sequences goes up, and (ii) the chance of recovering a correct internal edge is negatively correlated with the length of that edge. Fortunately, our results also show that increasing the number of sites in the alignment has a positive effect on phylogenetic accuracy.

List of participants

Guillaume Bernard
University of Queensland, Australia

Sangeeta Bhatia
University of Western Sydney, Australia

Laura Boykin
University of Western Australia, Australia

Anna Bruniche-Olsen
University of Tasmania

Chris Burridge
University of Tasmania

Greg Butler
Concordia University, Canada

Catherine Byrne
Tasmanian Museum and Art Gallery

Michael Charleston
University of Sydney, Australia

Stephen Crotty
University of Adelaide, Australia

Attila Egri-Nagy
University of Western Sydney, Australia

Sarah Fayed
University of Tasmania

Mathieu Fourment
University of Sydney, Australia

Andrew Francis
University of Western Sydney, Australia

Gillian Gibb
Massey University, New Zealand

Nick Ham
University of Tasmania

Simon Hills
Massey University, New Zealand

Barbara Holland
University of Tasmania

Xia Hua
Australian National University

Melissa Humphries
University of Tasmania

Peter Jarvis
University of Tasmania

Lars Jermiin
CSIRO Ecosystem Sciences, Canberra, Australia

Rebecca Jones
University of Tasmania

Greg Jordan
University of Tasmania

Subha Kalyaanamoorthy
CSIRO Ecosystem Sciences, Canberra, Australia

Saan Ketelaar-Jones
University of Tasmania

Rob Lanfear
Australian National University

David Liberles
University of Wyoming, USA

Bennet McComish
University of Tasmania

Karen Meusemann
CSIRO Ecosystem Sciences, Canberra, Australia

Jonathan Mitchell
University of Tasmania

Malgorzata O'Reilly
University of Tasmania

David Penny
Massey University, New Zealand

Ben Rohrlach
University of Adelaide, Australia

Barbara Schoenfeld
University of Tasmania

Stuart Serdoz
University of Western Sydney, Australia

Tristan Stark
University of Tasmania

Dorothy Steane
University of Tasmania

Mike Steel
University of Canterbury, New Zealand

Jeremy Sumner
University of Tasmania

Thomas Wong
CSIRO Ecosystem Sciences, Canberra, Australia

Michael Woodhams
University of Tasmania

James Worth
University of Tasmania

Cheong Xin Chan
University of Queensland, Australia