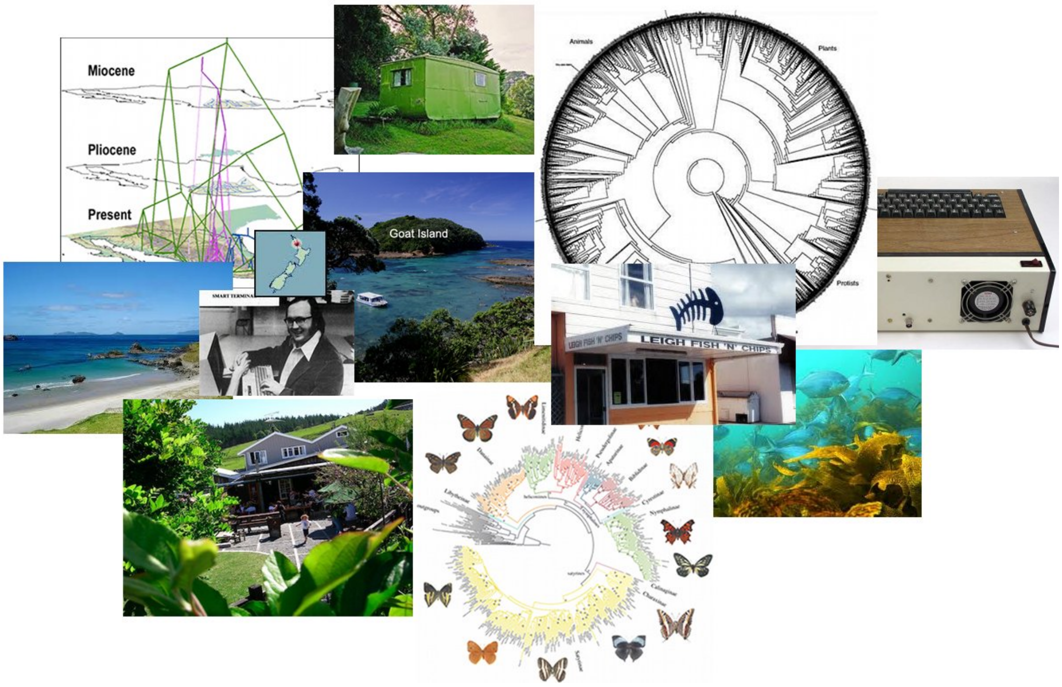

Leigh Sawmill 2011

...the cutting edge

The Annual New Zealand Phylogenetics Meeting

Sunday 6th – Friday 11th February 2011

Organisers

Stephane Guindon	s.guindon@auckland.ac.nz
Steffen Klaere	sklaere@maths.otago.ac.nz
Alexandra Miliotis	a.miliotis@auckland.ac.nz
David Bryant	david.bryant@otago.ac.nz
Dietrich Radel	D.Radel@math.canterbury.ac.nz (website)

Sponsors

The Department of Statistics at the University of Auckland
(http://www.stat.auckland.ac.nz/uoa/)

The Allan Wilson Centre for Molecular Ecology & Evolution
(http://www.allanwilsoncentre.ac.nz/)

List of Attendees

Alexei Drummond	Auckland	NZ	alexei@cs.auckland.ac.nz
Arndt von Haeseler	Vienna	Austria	arndt.von.haeseler@univie.ac.at
Barbara Holland	Hobart	Australia	barbara.holland@utas.edu.au
Barbara Schoenfeld	Palmerston North	NZ	b.schoenfeld@massey.ac.nz
Benjamin Redelings	Durham	USA	benjamin.redelings@nescent.org
Bennet McComish	Palmerston North	NZ	b.mccomish@massey.ac.nz
Bojian Zhong	Palmerston North	NZ	b.zhong@massey.ac.nz
Charles Pearce	Adelaide	Australia	charles.pearce@adelaide.edu.au
Charles Semple	Christchurch	NZ	c.semple@math.canterbury.ac.nz
Chieh-Hsi Wu	Auckland	NZ	cwu080@aucklanduni.ac.nz
Chris Simon	Welly/Storrs	NZ/USA	chris.simon@uconn.edu
Chuong Than	Michigan	USA	tvcuong@umich.edu
David Bryant	Dunedin	NZ	david.bryant@otago.ac.nz
David Penny	Palmerston North	NZ	d.penny@massey.ac.nz
Denise Kühnert	Auckland	NZ	denise.kuehnert@gmail.com
Ellen Nisbet	Adelaide	Australia	ellen.nisbet@gmail.com
Helen Shearman	Auckland	NZ	hshe018@aucklanduni.ac.nz
Jessica Leigh	Dunedin	NZ	jessica.w.leigh@gmail.com
Jessica Hedge	Edinburgh	UK	j.a.hedge@sms.ed.ac.uk
Joseph Heled	Auckland	NZ	jheled@gmail.com
Josh Collins	Palmerston North	NZ	j.collins@math.canterbury.ac.nz
Joshua Krissansen	Auckland	NZ	jkri012@aucklanduni.ac.nz
Julien Soubrier	Adelaide	Australia	julien.soubrier@adelaide.edu.au
Katharina Huber	Norwich	UK	katharina.huber@cmp.uea.ac.uk
Lars Jermiin	Canberra	Australia	lars.jermiin@csiro.au
Lindell Bromham	Sydney	Australia	lindell.bromham@anu.edu.au
Louis Ranjard	Auckland	NZ	l.ranjard@auckland.ac.nz
Marc Suchard	Los Angeles	USA	msuchard@ucla.edu
Mareike Fischer	Vienna	Austria	email@mareikefischer.de
Mathieu Blanchette	Montreal	Canada	blancem@mcb.mcgill.ca
Michael Charleston	Sydney	Australia	mcharles@it.usyd.edu.au
Michael Ott	Canberra	Australia	michael.ott@csiro.au
Mike Hendy	Palmerston North	NZ	m.hendy@massey.ac.nz
Mike Steel	Christchurch	NZ	mathmomike@gmail.com
Nicole Grünheit	Palmerston North	NZ	n.gruenheit@massey.ac.nz
Oliver Deusch	Palmerston North	NZ	o.deusch@massey.ac.nz
Peter Lockhart	Palmerston North	NZ	p.j.lockhart@massey.ac.nz
Peter Wills	Auckland	NZ	p.wills@auckland.ac.nz
Quentin Atkinson	Auckland	NZ	q.atkinson@auckland.ac.nz
Remco Bouckaert	Auckland	NZ	remco@cs.auckland.ac.nz
Rob Lanfear	Canberra	Australia	rob.lanfear@gmail.com
Russell Gray	Auckland	NZ	rd.gray@auckland.ac.nz
Ruth Bollongino	Mainz	Germany	bollongi@uni-mainz.de
Sandra Meid	Bonn	Germany	s.meid.zfmk@uni-bonn.de
Sebastian Böcker	Jena	Germany	sebastian.boecker@uni-jena.de
Sha Zhu	Christchurch	NZ	joe.zhu@pg.canterbury.ac.nz
Steffen Klaere	Dunedin	NZ	sklaere@maths.otago.ac.nz
Stephane Guindon	Auckland	NZ	s.guindon@auckland.ac.nz
Sven Herrmann	Norwich	UK	s.herrmann@uea.ac.uk
Tanja Stadler	Zurich	Switzerland	tanja.stadler@env.ethz.ch
Vincent Moulton	Norwich	UK	vincent.moulton@cmp.uea.ac.uk
Walter Xie	Auckland	NZ	walter@cs.auckland.ac.nz

PROGRAMME

Sunday 6th Feb

18:00 – 20:00	**Registration and reception (Sawmill Café)**

Monday 7th Feb

08:30 – 09:00	**Registration**
09:00 – 09:20	**Greetings and opening**
09.20 – 10:20	**[Applied Phylogenetics I]** *Chair: Chris Simon* 09:20–09:40 Pete Lockhart *The rotting roots of flowering plants? [page 8]* 9:40–10:00 Bojian Zhong* *A formal quantitative test of convergent evolution [page 8]* 10:00–10:20 Michael Charleston *Ancient patterns of mimicry in Helconius butterflies [page 9]*
10:20 – 11:00	**Morning break**
11:00 – 12:00	**[Epidemiology]** *Chair: Mike Steel* 11:00–11:20 Tanja Stadler *Estimating epidemiological parameters from viral sequence data [page 10]* 11:20–11:40 Denise Kühnert* *Phylogenetic analysis under an epidemiological population model [page 10]* 11:40–12:00 Jessica Hedge* *Characterising the phylodynamics of a pandemic as it emerges [page 11]*
12:00 – 14:00	**Lunch**
14:00 – 15:00	**[Partitioned data, supertrees]** *Chair: Barbara Holland* 14:00–14:20 Cuong Than *Deep coalescence and species tree inference [page 12]* 14:20–14:40 Sebastian Böcker *FlipCut supertrees [page 12]* 14:40–15:00 David Bryant *Statistical MRP [page 13]*
15:00 – 15.30	**Afternoon break**
15:30 – 16.30	**[Diversity, conservation]** *Chair: Arndt von Haeseler* 15:30–15:50 Steffen Klaere *Some thoughts on measuring biodiversity [page 14]* 15:50–16:10 Helen Shearman* *Maximum Minimum Distance and more [page 14]* 16:10–16:30 Charles Semple *Approximability results in conservation biology [page 14]*
	Dinner – your own arrangements

Tuesday 8th Feb

09.20 – 10:20	**[Theoretical phylogenetics I]** *Chair: David Bryant* 09:20–09:40 Mike Steel *Sawing a tree off at the leaves [page 15]* 9:40–10:00 Sven Herrmann *On the split decomposition of a k-dissimilarity map [page 15]* 10:00–10:20 Vincent Moulton *Constructing and drawing planar split networks [page 16]*
10:20 – 11.00	**Morning break**
11:00 – 12:00	**[Methods and algorithms I]** *Chair: Mike Hendy (TBC)* 11:00–11:20 Joshua Collins* *Clustering character data into differing tree topologies [page 17]* 11:20–11:40 Michael Ott *New opportunities arising for phylogenetic tree inference from computer architecture and high performance computing [page 17]* 11:40–12:00 Jessica Leigh *The Big (Bayesian) picture in simulation-based testing [page 18]*
12:00 – 14:00	**Lunch**
14:00 – 15:20	**[Phylogeography]** *Chair: Lars Jermiin* 14:00–14.20 Louis Ranjard *Estimating dispersal range from species phylogenies [page 19]* 14:20–14:40 Ruth Bollongino *The origin of the first European farmers and their relation to indigenous hunter-gatherers [page 19]* 14:40–15:00 Chris Simon *Species or species swarms? Complex species boundaries result from repeated contact and gene exchange between recently diverged species of NZ cicadas [page 19]* 15:00–15:20 Remco Bouckaert *Substitution models with a relaxed attitude [page 20]*
15:20 – 15.50	**Afternoon break**
15:50 – 16.30	**[Language evolution]** *Chair: Alexei Drummond* 15:50–16:10 Quentin Atkinson *Language evolution in space and time [page 21]* 16:10–16:30 Russell Gray *Language evolution: From trees to networks [page 21]*
	Dinner – your own arrangements

Wednesday 9th Feb

All day	**Excursions**
18:30 – 20:00	**Dinner at the Sawmill Café**

Thursday 10th Feb

09.20 – 10:20	**[Theoretical phylogenetics II]** *Chair: Vincent Moulton* 09:20–09:40 Benjamin Redelings *Extending majority consensus trees to represent wandering taxa [page 22]* 9:40–10:00 Katharina Huber *Metrics on multi-labeled trees: interrelationships and diameter bounds [page 22]* 10:00–10:20 Mareike Fischer *Revisiting the question how many characters are needed to reconstruct the true tree [page 23]*
10:20 – 11.00	**Morning break**
11:00 – 12:00	**[Applied phylogenetics II]** *Chair: Marc Suchard* 11:00–11:20 Bennet McComish* *Recovering nucleotide substitution matrices from alignments [page 24]* 11:20–11:40 Ellen Nisbet *Solving mysteries in thoroughbred horse racing using phylogenetics [page 24]* 11.40–12.00 Oliver Deusch *Mutational dynamics and model misspecification in chloroplast genomes [page 25]*
12:00 – 14:00	**Lunch**
14:00 – 15:20	**[Rates and dates]** *Chair: Tanja Stadler* 14:00–14:20 David Penny *The J-shaped curve; integrating molecular evolution over different time scales [page 26]* 14:20–14:40 Stephane Guindon *New methods for estimating node ages from molecular data [page 26]* 14:40–15:00 Julien Soubrier* *The influence of rate heterogeneity among sites on the time dependency of molecular rates [page 27]* 15:00–15:20 Joseph Heled *That Beastly prior [page 27]*
15:20 – 15.50	**Afternoon break**
15:50 – 16.30	**[The Origins]** *Chair: David Penny* 15:30–15:50 Barbara Schönfeld* *Endosymbiosis – Yearning for a better life [page 28]* 15:50–16:10 Peter Wills *The origin of genes [page 28]*
17:00	**Best student talk prizes announcement**
17:30	**Drinks at the Sawmill Café**

Friday 11th Feb

08:55 – 09:00	**Penny ante prize announcement**
09:00 – 10:00	**[Applied phylogenetics III]** *Chair: Peter Lockhart* 09:00–09:20 Nicole Grünheit *Testing a hypothesis of Nunatak survival in the New Zealand Southern Alps [page 29]* 09:20–09:40 Charles Pearce *Some problems in post-mortem DNA profiling [page 30]* 09:40–10:00 Lars Jermiin *New methods of identifying fast-evolving sites in aligned sequence data [page 30]*
10:00 – 10:30	**Morning break**
10:30 – 11:50	**[Methods and algorithms II]** *Chair: Sebastian Böcker* 10:30–10:50 Rob Lanfear *Finding optimum partitioning schemes for the analysis of molecular sequence data [page 31]* 10:50–11:10 Mathieu Blanchette *Ancestral genome reconstruction and its uses toward annotating the human genome [page 31]* 11:10–11:30 Arndt von Haeseler MISFITS: *evaluating the goodness of fit between a phylogenetic model and an alignment [page 32]* 11:30–11:50 Barbara Holland *Phylogenies from* DArTs - *stochastic Dollo with censored data [page 32]*
11:50	**Closing comments, lunch**

ABSTRACTS

MONDAY

APPLIED PHYLOGENETICS I

Peter Lockhart
Title: *The rotting roots of flowering plants?*
Abstract: We have recently sequenced and analyzed sequences from the chloroplast genome of *Trithuria* - an unexpected basal angiosperm from dune lakes in Northland, New Zealand. This study raises some questions over the certainty often expressed in our understanding of the origins of flowering plants.

Bojian Zhong
Title: *A formal quantitative test of convergent evolution*
Abstract: Despite the modern evolutionary theory is widely accepted in biology, there are very few fundamental tests to support it. Here we concentrate on a formal quantitative test that sequences converge backwards through time. It predicts from evolutionary theory that ancestral sequences from different related groups (in our first test: monocots and eudicots) should be more similar to each other, than to any pairwise distance between members of these groups. We estimated the ancestral sequences of monocots (24 taxa) and eudicots (44 taxa) groups independently using 51 single chloroplast (chl) proteins, and compared the ancestral distance to the pairwise distance between any members of monocots and eudicots. We also concatenated 51 chl proteins to test this theory. From our preliminary analyses, most of individual chl proteins and concatenated data support the convergence hypothesis, which indicates that evolution could be quantitatively tested by applying formal tests. Other genes (such as mitochondrial and nuclear genes) and chl genes with deeper divergences can be evaluated in the next analyses. Joint work with Tim White, Mike Hendy and David Penny.

Michael Charleston

Title: *Ancient patterns of mimicry in Helconius butterflies*

Abstract: Within the Heliconius genus of butterflies there are about 75 species. Many of these species mimic others in the genus; the best studied such pair is the mimic *H. melpomene*, and its target *H. erato*. The mimicry is based on common wing patterns, where both target and mimic are unpalatable. These Müllerian mimicry complexes have been of major interest for nearly 150 years for their potential for having coevolved. Each of these two species divides into several races, whose monophyly has been debated for some time: initially the races were identified by morphology and geographic range, but subsequent molecular sequence data called these groups into some question. We examined the phylogenetic relationships among races based on published sequence data (COI, COII and two nuclear genes Mpi and Tpi). We used phylogenetic and coalescent-based methods to establish the most reliable levels of monophyletic groups within the two species – this appears to be at the level of biogeographic region, rather than race. By coding the mimicry associations accordingly, we analysed the cophylogenetic signal corresponding to the two trees. We found a highly significant level of congruence that we believe must come from a relatively long history of coevolution, based solely on mimicry.

EPIDEMIOLOGY

Tanja Stadler
Title: *Estimating epidemiological parameters from viral sequence data*
Abstract: Epidemiological processes leave a fingerprint in the pattern of genetic structure of virus populations. I present a new method to infer epidemiological parameters directly from viral sequence data. The method, being available within the Beast software package, is based on Bayesian phylogenetic inference assuming a birth-death model rather than the commonly used coalescent as the model for the epidemiological transmission of the pathogen. Using the birth-death model has the advantage that transmission and death rates are estimated independently, and therefore enables the estimation of key epidemiological parameters of the pathogen, like the basic reproductive number, using only sequence data. Our method yields a basic reproductive number of 2 for the Swiss HIV-1 epidemic, meaning that HIV is still not under control in Switzerland.

Denise Kühnert
Title: *Phylogenetic analysis under an epidemiological population model*
Abstract: Analysing the phylogenetics of a population of interest often implies population sizes to be assumed constant or exponentially growing. Since such assumptions are inappropriate for many populations, scientists search methods allowing for less restrictive assumptions on population sizes, e.g. skyline plots (Pybus, Rambaut, and Harvey, 2000).

Epidemiological methods have been suggested to be reconciled with phylogenetic approaches in order to account for evolutionary as well as ecological processes (e.g. Grenfell et al. 2004). This work in progress aims at a combined epidemiological phylogenetic analysis by incorporating the dynamics of an SIR model into Bayesian phylogenetic inference.

The advantage and difficulty of this approach is to model and estimate each of the changing numbers of susceptible (S) and infected (I) individuals over time. In a second step, this approach shall be incorporated into phylogeographic methods for the analysis of infectious diseases.

Jessica Hedge

Title: *Characterising the phylodynamics of a pandemic as it emerges*

Abstract: The 2009 H1N1 pandemic was unprecedented in its depth and coverage of surveillance and by June 2010, over 1800 virus genomes had been sequenced. We use Bayesian phylogenetics to estimate three important genetic and epidemiological parameters to charaterise the pandemic as it broke out (evolutionary rate, date of the outbreak and the basic reproductive ratio, R0). We show that after the sequencing of the first ~100 virus genomes in early June 2009, the accuracy of our estimates remains constant and the addition of further sequencing only improves precision. Similar results were found in an analysis of seasonal H1N1 outbreak in North America and suggests that sufficient genetic diversity is present in the first 100-150 genomes sequenced to characterise both outbreaks using these three parameters. We are currently addressing whether the accuracy of these estimates is provided by the initial sequence diversity or the number of sequences analysed.

PARTITIONED DATA, SUPERTREES

Cuong Than

Title: *Deep coalescence and species tree inference*

Abstract: A species tree in general can differ from the evolutionary relationships of genes sampled in the species. One of the major sources for species/gene tree discordance is incomplete lineage sorting, or deep coalescence. For a pair of a species tree and gene tree, we can measure the severity of deep coalescence by calculating the minimum number of extra gene lineages required to reconcile the gene tree within the species tree. In this talk, we will describe a dynamic programming algorithm for inferring species trees for a collection of gene trees by minimizing the deep coalescence cost. We then show that the minimizing deep coalescence criterion is not statistically consistent under the multispecies coalescent model. In particular, we show that for asymmetric four-leaf species trees and for species trees with at least five leaves, there exists a region of tree branch lengths over which the criterion produces an incorrect species tree estimates in the limit of infinitely many available gene trees. Joint work with Luay Nakhleh, Rice University, and Noah Rosenberg, University of Michigan.

Sebastian Böcker

Title: *FlipCut supertrees*

Abstract: Constructing a supertree of a given set of rooted input trees can be formalized in different ways, to cope with contradictory information in the input. In particular, there exist methods based on encoding the input trees in a matrix, and methods based on finding minimum cuts in some graph representing the incompatibility in the input. Evaluations have shown that matrix representation methods will compute supertrees of better quality but, unfortunately, the underlying problems are computationally hard. In contrast, graph-based methods compute a supertree in polynomial time, but supertrees are inferior in quality.

In my talk, I will present our novel FlipCut supertree method. This method combines the computation of minimum cuts from graph-based methods with a matrix representation method, namely Minimum Flip Supertrees. Here, the input trees are encoded in a 0/1/?-matrix, and we search for a minimum set of 0/1-flips such that the resulting matrix admits a directed perfect phylogeny. I will also show how to use edge weights, to weight the columns of the 0/1/?-matrix.

Initial evaluation are very promising, and FlipCut supertrees might indeed bridge the gap between Matrix Representation methods and graph-based methods.

David Bryant
Title: *Statistical MRP*
Abstract: In Mary Shelley's classic novel, Dr Frankenstein assembles his monster from abandoned bones and body parts. We have assembled a methodology for inferring phylogenies from multiple genes by following the same general principle. The method, which we call *Statistical MRP*, is fast, easy to implement, and, most importantly, statistically responsible. We construct a confidence set for the inferred supertree, a set which is empty when the null hypothesis of concordance is rejected.

Diversity, conservation

Steffen Klaere
Title: *Some thoughts on measuring biodiversity*
Abstract: In recent years, phylogenetic diversity has garnered extensive interest mainly from the theoretical side of the community. In particular, several approaches to quite complex optimization problems have been presented. Using our latest solver I will discuss the merits of the method and suggest alternative approaches. This is joint work with Bui Quang Minh and Arndt von Haeseler, Vienna and Felix Forest, Kew Gardens.

Helen Shearman
Title: *Maximum Minimum Distance and more*
The maximum minimum distance method (MMD) is intended to produce a subset of evolutionary units that maximises the spread of this subset across the tree. The method was designed to maximise the number of features captured by the subset when they evolve according to a model that allows features to arise and/or disappear through time. In this talk I look at in what circumstances this method captures more features than the subset found by maximising phylogenetic diversity (PD). I then propose a new method that combines many of the advantages of PD and MMD.

Charles Semple
Title: *Approximability results in conservation biology*
Abstract: Optimization problems arise in conservation biology. In the context of preserving species diversity, the problem is to maximize some given measure of diversity subject to certain constraints on resources. In this talk, we describe some recent approximability results for these problems. This is joint work with Magnus Bordewich (Durham University).

TUESDAY

THEORETICAL PHYLOGENETICS I

Mike Steel
Title: *Sawing a tree off at the leaves*
Abstract: Saw millers will tell you that the best way to fell a tree is to cut it off at the base. To the mathematician, however, it can seem just as natural to start at the top and saw off the leaves. Not surprisingly, this process carries with it some attendant hazards. In this talk I will describe how to safely wield the 'mathematical saw' on a phylogenetic tree, and thereby harvest not just leaves and stems, but a variety of low-hanging fruit.

Sven Herrmann
Title: *On the split decomposition of a k-dissimilarity map*
Abstract: A k-dissimilarity map on a finite set X is a function D assigning a real value to each subset of X with cardinality k. Such functions are commonly used to reconstruct evolutionary trees or networks. In this talk, I will explain how regular subdivisions of the kth hypersimplex can be used to obtain a canonical decomposition of a k-dissimilarity map into the sum of simpler k-dissimilarity maps arising from bipartitions (or splits) of X. In the special case k=2, this decomposition is the well-known split decomposition of a distance due to Bandelt and Dress. Furthermore, a characterisation those sets of splits that may occur in the resulting decompositions of k-dissimilarity maps will be given. This also gives a new proof of a theorem of Pachter and Speyer for recovering k-dissimilarity maps from trees. This is joint work with Vincent Moulton.

Vincent Moulton

Title: *Constructing and drawing planar split networks*

Abstract: Split networks are graph theoretical structures that generalize phylogenetic trees which are used in phylogenetics to visualize evolutionary data that supports conflicting phylogenetic signals. Split networks may be thought of as a tool to display split systems, or collections of bipartitions of a finite set, and may be generated using the SplitsTree package. Recently the NeighborNet approach for generating split networks has become rather popular, in part because it is guaranteed to generate a planar network. Even so, NeighborNet always places labels on the outside of the network, and there are certain split systems, so-called flat split systems, which can be displayed by planar networks only in case some labels are allowed to be placed inside the network too. Here we present some results on how to compute a minimal planar split network displaying a flat split system in polynomial time, provided the split system is given in a certain way. We also discuss how the networks generated by the algorithm compare with those generated by SplitsTree, and discuss their potential use in phylogeographic applications. This is joint work with Dr. Andreas Spillner, University of Greifswald

METHODS AND ALGORITHMS I

Joshua Collins

Title: *Clustering character data into differing tree topologies*

Abstract: It is possible to say individual character sites in a data sequence have definitely evolved on some tree. However, in some data sets, such as those arising from hybridisation, it may not always be the same tree. This talk will cover the details of a genetic algorithm that attempts to cluster such data into small sets of trees in a sensible way using various extensions of MP. At the end will be discussed the inherent shortcomings and possible improvements of the implementation.

Michael Ott

Title: *New opportunities arising for phylogenetic tree inference from computer architecture and high performance computing*

Abstract: Maximum Likelihood-based phylogenetic tree inference is considered to be extremely compute-and memory-intensive. In fact, runtime and memory requirements can render the analysis of large-scale datasets infeasible.

However, efficient and generic parallelisation approaches for the phylogenetic likelihood function are available that allow for overcoming both requirements as a limiting factor for phylogenetic analyses by distributing computations as well as data structures over multiple computers/nodes.

Additionally, recent developments in computer architecture offer new opportunities for applying more sophisticated models of sequence evolution (read: computationally expensive) with only little impact on overall runtime: due to the huge memory footprints, a large portion of the time required for inferring phylogenetic trees is spent on memory transfers.

As the performance gap of modern multicore processors and their memory subsystems continues to increase, additional computations can be hidden behind those memory transfers without affecting the total runtime.

Jessica Leigh

Title: *The big (Bayesian) picture in simulation-based testing*

Abstract: Statistical methods applied to many areas of the sciences are often assessed and marketed using simulation-based performance evaluation. This sort of framework can involve repeated simulation over a large number of combinations of values for relevant parameters, or the selection of a few "pet"parameter values. While the former approach to parameter selection can be inefficient and unwieldy, the second is far from objective and potentially dishonest. We have developed a Markov chain Monte Carlo sampling method to identify regions of parameter space where methods perform either well or poorly. Our method is similar to Approximate Bayesian Computation in that it does not involve the calculation of likelihoods, but samples from the probability distribution of interest, rather than an approximation thereof. In addition to describing our method, I will present results from its application to such diverse subject areas as population genetics and public health.

PHYLOGEOGRAPHY

Louis Ranjard
Title: *Estimating dispersal range from species phylogenies*
Abstract: The allopatric speciation model postulates that new species arise as a result of geographic isolation due to vicariance or dispersal. If biological dispersal is the main force responsible for population isolation and subsequent speciation, one can expect a correlation between the geographical location of the species and the corresponding phylogeny. The species tree could therefore convey precious information about the history of colonisation. The dispersal range of the species determines the amount of information recoverable: if species have a wide dispersal range, the correlation between geographic and phylogenetic distances is expected to be weak. On the other hand, if species have a limited dispersal range, adjacent locations have been colonised by closely related species, hence a strong association between geography and phylogeny. We simulate migration histories in a finite space and use a statistical approach to estimate the dispersal parameter of such model. Preliminary results about these simulations will be presented.

Ruth Bollongino
Title: *The origin of the first European farmers and their relation to indigenous hunter-gatherers*
Abstract: For already several years our group has been studying the molecular genetic background of the Neolithic Transition in Europe. Focusing on a broad perspective on prehistoric population movements in Eurasia, our main aim is to reveal the phylogeographic structure of mitochondrial lineages of human populations. It is still not fully understood how European hunter-gatherers are related to the first farming populations. Additionally, the analyses of selected genes like the "lactase persistence" help to understand the evolutionary adaptive processes that are caused by the immense changes of living conditions. Furthermore, we investigate the origin and spread of the four main domesticates (cattle, pigs, sheep and goat) to reveal a more detailed picture of the complex interactions of early farming societies.

Using Bayesian Serial SimCoal and Approximate Bayesian Computation (ABC) on a sample set widely spread over different geographical regions and from different time periods we simulated the development of prehistoric populations and their relation to each other. With regard to Next Generation Sequencing and new computational challenges we will present the current state of the projects and outline the future directions of this field of research.

Chris Simon
Title: *Species or species swarms? Complex species boundaries result from repeated contact and gene exchange between recently diverged species of NZ cicadas*
Abstract: Pleistocene interglacial contact between recently diverged species in the NZ grass-cicada "*Kikihia muta* complex" has resulted in a complex pattern of gene exchange across species boundaries. We used microsatellite loci to estimate current and past gene flow at secondary contact between lineages defined on the basis of courtship songs, morphological traits, and mitochondrial phylogenies. Our phylogeographic studies of the genus *Kikihia* identified 20 potential hybrid zones between species pairs that vary widely in their times of mtDNA divergence (between 20,000 and 3.5 million years). These well-supported molecular phylogenies, dated using Bayesian molecular relaxed-clock methods, are used as a temporal and spatial framework to understand species interactions. The mating song and female response of each species (controlling pre-zygotic isolation), has also been characterized throughout each species' range. This preliminary study examines the introgression of alleles at three contact zones involving four species (*Kikihia "nortwestlandica", K. "southwestlandica", K. muta muta, and K. "tuta"*). Two of these contact zones appear to be recent while the third contains a population of hybrid individuals that appears to have originated in a previous interglacial period. Joint work with Beth Wade, University of Connecticut.

Remco Bouckaert
Title: *Substitution models with a relaxed attitude*
Abstract:
Phylogeography analysis requires clock like trees for the more interesting types of inference. However, when the tree is mostly informed by sequence data (e.g. DNA sequences or language data) the rate of geographical dispersal may need to be reconciled with the rate of sequence evolution. One way to do this is to allow the geographical dispersal rate to be drawn from a distribution instead of being fixed at a point indicated by branch lengths in the tree. We study this relaxed clock approach for some geographical models as well as for some of the more popular substitution models like Jukes Cantor, HKY and GTR. These 'relaxed substitution models' turn out to have some nice mathematical properties. In many cases a closed form formula for the transition probabilities can be found which allows ease of implementation, and saving some computational effort compared to MCMC based approaches.

LANGUAGE EVOLUTION

Quentin Atkinson
Title: *Language evolution in space and time*
Abstract: Recent work in computational historical linguistics has successfully applied phylogenetic methods from biology to linguistic data to test hypotheses about language family relationships, chronology, and the tempo and mode of language evolution. However, relatively little attention has focused on explicitly modeling large-scale spatial processes of language change. Here I report results from collaborative research that uses tools from population genetics and phylogeography to analyze spatial information derived from comparative linguistic data. This work identifies clear spatial signal in the data that can be used to shed light on the origins of the world's major language families.

Russell Gray
Title: *Language evolution: From trees to networks*
Abstract: Historical linguistics and evolutionary biology share a fascination in trees. However, in both disciplines it has long been recognized that horizontal transmission is an important process that is not captured in pure tree models. In recent years evolutionary biologists have developed explicit phylogenetic methods to take into account horizontal gene transfer. In this talk I will apply these phylogenetic networks methods to infer the frequency of borrowing during the evolution of Indo-European and Polynesian languages. Some of the results are quite surprising.

THURSDAY

THEORETICAL PHYLOGENETICS II

Benjamin Redelings
Title: *Extending majority consensus trees to represent wandering taxa*
Abstract: Biologists commonly use the majority consensus tree to visually summarize Bayesian posterior distributions on evolutionary tree shape. This summary combines full splits that individually have strong support into a single topology that may have multifurcations. In order to reveal hidden structure in tree distributions, we extend the majority consensus to represent supported partial splits (of only some leaf taxa) by introducing a new tree structure in which each branch may have a range of attachment locations.

Katharina Huber
Title: *Metrics on multi-labeled trees: interrelationships and diameter bounds*
Abstract: Multi-labeled trees or MUL-trees, for short, are trees whose leaves are labeled by elements of some non-empty finite set X such that more than one leaf may be labeled by the same element of X. This class of trees includes phylogenetic trees and tree shapes. MUL-trees arise naturally in, for example, biogeography and gene evolution studies and also in the area of phylogenetic network reconstruction. In this talk we introduce novel metrics that may be used to compare MUL-trees, most of which generalize well-known metrics on phylogenetic trees and tree shapes. These metrics can be used, for example, to better understand the space of MUL-trees or to help visualize collections of MUL-trees. In addition, we describe some relationships between the MUL-tree metrics that we present and also give some novel diameter bounds for these metrics.

Mareike Fischer

Title: *Revisiting the question how many characters are needed to reconstruct the true tree*

Abstract: The question of how many sequence sites are required to recover the evolutionary relationship of the underlying species accurately is important for phylogeneticists. It is known that a particularly challenging problem for phylogenetic methods arises when a rapid divergence event occurred in the distant past, which leads to long pending branches and a short internal branch in the corresponding phylogenetic tree.

While most previous approaches tackling this problem considered only 2-state models, we investigate the scenario based on all four (DNA) character states. Particularly, we analyze a binary unrooted 4-taxon phylogenetic tree with a short interior edge and pending edges of multiple lengths. In my talk, I will present an optimal branch length of the interior edge in this case and I will explain how many characters are at least needed to reconstruct the 'true' tree.

APPLIED PHYLOGENETICS II

Bennet McComish

Title: *Recovering nucleotide substitution matrices from alignments*

Abstract: The nucleotide substitution rate matrix, *Q*, is a key parameter of molecular evolution. *Q* is often assumed to be constant over a phylogenetic tree, but in reality there are a number of mutational processes that are known to affect nucleotide composition by altering *Q*. Variations in *Q* are therefore key to understanding fundamental processes of molecular evolution, and are also likely to have major effects on phylogeny reconstruction. There are several different methods for inferring *Q*, but none have been applied on a large scale. I will present some preliminary results of the estimation of *Q* from large alignments.

Ellen Nisbet

Title: *Solving mysteries in thoroughbred horse racing using phylogenetics*

Abstract: We have retrieved ancient DNA from two historic Thoroughbred horses and used phylogenetics to solve two of the longest running mysteries in racing history. The greatest racehorse ever known was Eclipse who was born during the solar eclipse of 1764 and never lost a race. His body was used for the first animal autopsy, his skeleton preserved and is now housed in the Royal Veterinary College. However there has been controversy over the authenticity of his skeleton ever since. The 1880 Epsom Derby was won by Bend Or. In one of the great controversies of racing, a groom claimed that Bend Or and another horse, Tadcaster, had been swapped during training as yearlings. The groom was sacked, the result stood and Bend Or now resides at the Natural History Museum, London. Eclipse and Tadcaster were both extremely popular at stud, and the vast majority of racehorses today are descendants. We have sequenced the mitochondrial D-loop DNA and the *ASIP* and *MC1R* coat colour gene from each skeleton, comparing results with over 1000 living thoroughbred horse to confirm or deny identity. Our results challenge the ancestry of many living thoroughbred horses. Joint work with Mim Bower, University of Cambridge.

Oliver Deusch

Title: *Mutational dynamics and model misspecification in chloroplast genomes*

Abstract: A major controversy in plant evolution is the rooting of the angiosperm phylogeny with different studies having obtained inconsistent results regarding the nature of the most basal angiosperm. Analyses are typically performed on a dataset of ~60 genes present in all lineages where genes are aligned, concatenated and phylogenetic analyses are carried out. This approach is prone to model misspecification with concatenated datasets sometimes favouring an evolutionary model not supported by a single gene. Model fitting can even be problematic when individual genes are analyzed as the same gene may show lineage specific mutational dynamics. We hypothesize that non-homogeneous covarion-like patterns of evolution are possibly explained by positional effects of genes at different locations in the chloroplast genome.

I present results from a study investigating variation in the mutational dynamics of different chloroplast lineages (including non-pine conifer genomes obtained by Illumina GAII sequencing in our group) and their relationship to spatial patterns of gene rearrangement. I discuss how features of the chloroplast genome impact on model fitting with time reversible stationary models.

David Penny
Title: *The J-shaped curve; integrating molecular evolution over different time scales*
Abstract: The so-called molecular clock of molecular evolution showed that the rate of neutral mutations per generation also equalled the long-term of sequence change. However, there is now more interest on the 'apparent' acceleration of short-term rates, or the 'apparent' acceleration of rates when speciation is occurring. We have extended the long-term calculations into the 'twilight zone' whilst fixation or loss of mutations is still occurring. Knowing population size and structure means that we can determine that the real mutation rate does not change at shorter times. There are interesting and predictable short-term effects from population structures modelled as different forms of connected graphs. We find a range of effects from those of population size, population structure, and speciation. Only preliminary work has been done on analytical solutions, but the possibilities are endless. Joint work with Chris Tuffley, Tim White and Mike Hendy

Stephane Guindon
Title: *New methods for estimating node ages from molecular data*
Abstract: Hierarchical Bayes modeling provides a suitable statistical framework for the estimation of divergence times using molecular sequences. The inference generally relies on the Metropolis-Hastings technique, which sometimes displays poor mixing behavior (i.e., the model parameters are difficult to estimate). The first part of this talk introduces a new approach that combines a new model of evolution of the rate of evolution to an approximation of the likelihood function. This method leads to convenient mathematical simplifications that considerably improve the estimation process.

The second part of the talk focuses on the distinction between models that describe rate trajectories and those that describe rates averaged along branches. Assuming that a Brownian process describes the rate trajectory, it is possible to derive the exact distribution of rates averaged along branches. Properties of this distribution will be discussed. In particular, the prior correlation of average rates across edges leads to interesting predictions from a biological perspective.

Julien Soubrier

Title: *The influence of rate heterogeneity among sites on the time dependency of molecular rates*

Abstract: Recent work has suggested that molecular rates appear to accelerate towards the present in a curvi-linear relationship informally described as the 'rates curve' (Ho et al., 2005; Penny, 2005; Endicott et al., 2009; Henn et al., 2009). If correct, the implications of this bias are manifold, and include the inability to accurately date evolutionary events in the recent past, such as human evolution, domestication and epidemiological events.

Multiple potential causes of the rates curve have been suggested, including selection, demographic factors and substitution saturation (Subramanian, 2009; Navascués, Emerson, 2009; Soares et al., 2009).

We investigate the impact of among-site rate heterogeneity on the time dependency of molecular rates, using mathematical models and simulated sequences.

With theoretical models, we show that a rates curve is explicitly predicted in a simple situation where sequences are comprised of two classes of sites, evolving under the same substitution model except for different rates. We also explore the resulting curve, the functions of the ratio between fast and slow sites, and the proportion of sites.

Simulated datasets also produce a rates curve with standard analytical programs, such as BEAST, across a range of time periods. Using these simulations, we are able to show the impact of such bias on both rate estimates and node age estimates over a wide range of time.

Joseph Heled

Title: *That Beastly prior*

Abstract: The use of fossil evidence to calibrate divergence time estimation has a long history. More recently Bayesian MCMC has come to dominate phylogenetic inference and fossil evidence has been re-interpreted as the specification of prior distributions on the divergence times of calibration nodes. These so-called ``soft calibrations" have become widely used but the statistical properties of calibrated tree priors in a Bayesian setting has not been carefully investigated. Here we clarify that calibration densities, such as those defined in BEAST 1.5, do not represent the marginal prior distribution of the calibration node. We illustrate this with a number of analytical results on small trees. We also describe an alternative construction for a calibrated Yule prior on trees that allows direct specification of the marginal prio distribution of the calibrated divergence time, with or without the restriction of monophyly. This method requires the computation of the Yule prior conditional on the height of the divergence being calibrated. Unfortunately, a practical solution for multiple calibrations remains elusive. Our results suggest that direct estimation of the prior induced by specifying multiple calibration densities should be a prerequisite of any divergence time dating analysis.

The Origins

Barbara Schönfeld
Title: *Endosymbiosis – Yearning for a better life*
Abstract: Endosymbiosis, the merger of two initially independent organisms is a complex process. It is on its genomic level characterized by extreme gene loss, reduction and specialization in function.

Endosymbiotic relationships have always played an essential role in eukaryote evolution, most prominently with the symbiogenesis of mitochondria and plastids. These milestones in the evolution of eukaryotic life lie deep in time, making a detailed reconstruction of the processes leading to organelle formation difficult. However there are also younger endosymbioses that allow eukaryotic hosts to utilize prokaryotic metabolic functions such as nitrogen fixation and amino acid synthesis to their advantage. I discuss some recently initiated whole genome studies.

Peter Wills
Title: *The origin of genes*
Abstract: What features of the physical world make it possible for genes to exist as units of heredity? The banal, uninformative answer to this question is "natural selection", which consigns biology, just as Rutherford would have it, to the domain of stamp collecting, a heap of arbitrary detail concerning historical events. Following a different path, Schrödinger tried to delineate what sort of atomic-level structure could, in principle, serve as a gene; and drew attention to the information-storage capacity and thermal stability of what Delbrück had called a "quasi-periodic crystal". In the subsequent development of molecular biology, ideas like "information" and "code", taken from the theory of computation, were adopted uncritically as providing a new *lingua franca* for explanations in biology. Now that the physical origin of many genes is an irrelevant detail of their existence, especially those that originate through commercial or military decision-making, it is necessary to ask whether events that were constrained primarily by purely symbolic relationships, rather than molecular interactions, were somehow involved at the beginning, as nucleic acid sequences first came to represent functionalities that conferred selective value on them. I will discuss these problems in relation to the phylogeny of the amino-acyl tRNA synthetases and other enzymes that all organisms share, tracing the origin of life not to ancient terrestrial events, but to the intrinsic possibility of autocatalytic functionality and the variable consequences of finely differentiated structures interacting with one another.

FRIDAY

APPLIED PHYLOGENETICS III

Nicole Grünheit
Title: *Testing a hypothesis of Nunatak survival in the New Zealand Southern Alps*
Abstract: Two competing hypotheses have been used to explain plant distributions in the European alps. The *Nunatak* (from Inuit *nunataq*) hypothesis postulates that in areas covered by permanent ice during Pleistocene glacial periods, ice free microhabitats nevertheless existed which supported plant life. These ice age refugia acted then as a source for recolonisation of other areas once the ice sheet retreated. In contrast, the *tabula rasa* hypothesis postulates total extinction within glaciated areas, survival in peripheral refugia, and then postglacial re-immigration from these refugia (Stehlik, 2003). Similar questions exist over the response of New Zealand plants to past climate change. Wardle (1988) has suggested that during Pleistocene glacial periods plants were eliminated from much of the Southern alps (except perhaps in the Seaward Kaikoura and Fiordland mountains).

In contrast, and while acknowledging the importance of dispersal for explaining plant distributions, McGlone, et al. (2001) notes the rapid rate of recolonisation of plant distributions following the last glacial maxima, and has questioned whether this observation suggests *nunatak* survival for some species (McGlone, 1985). *Nunatak* survival has been speculated for *Pachycladon enysii*, a high altitude alpine herb restricted to rocky bluffs in the South Island mountains of New Zealand. Based on present day distributions and ecological preferences, Heenan and Mitchell (2003) have hypothesized *insitu* survival for this species and extermination of *Pachycladon fastigiatum*, at high elevation in the New Zealand Southern Alps during Pleistocene glacial periods. To test this hypothesis we are undertaking phylogenetic analyses of chloroplast wide SNP markers. These have been determined using GAII Illumina (GAIIx) 75 base single read sequencing.

Heenan, P.B. and Mitchell, A.D. (2003) Phylogeny, biogeography and adaptive radiation of *Pachycladon* (Brassicaceae) in the mountains of South Island, *New Zealand Journal of Biogeography*, **30**, 1737 - 1749.

McGlone, B.S. (1985) Plant biogeography and the late Cenozoic history of New Zealand, *New Zealand Journal of Botany*, **23**, 723 - 749.

McGlone, B.S., Duncan, R.P. and Heenan, P.B. (2001) Endemism, species selection and the origin and distribution of the vascular plant flora of New Zealand, *Journal of Biogeography*, **28**, 199 - 216.

Stehlik, I. (2003) Resistance or emigration? Response of alpine plants to the ice ages, *Taxon*, **52**, 499 - 510.

Wardle, P. (1988) Effects of glacial climates on floristic distribution in New Zealand. A review of the evidence, *New Zealand Journal of Botany*, **26**, 541 - 555.

Charles Pearce
Title: *Some problems in post-mortem DNA profiling*
Abstract: A standard problem in studies involving non-recent DNA is degradation/fragmentation. This can occur even with material only a few years old, let alone what we refer to as Ancient DNA. In this talk we consider
(a) the impact of environmental factors on the quality of post-mortem DNA profiling, and
(b) a specific study involving linking three living individuals with a putative ancestor who died a century ago.
This is joint work with Maciej Henneberg, University of Adelaide.

Lars Jermiin
Title: *New methods of identifying fast-evolving sites in aligned sequence data*
Abstract: Rate-heterogeneity across sites is a widely recognized problem in molecular phylogenetic studies. Attempts to overcome the problem involve deleting the fast-evolving sites from the data before the phylogenetic analysis and/or using phylogenetic methods that assume rate-heterogeneity across sites can be modeled by a distribution of probabilities. Each of these approaches has its own advantages and limitations. Here, we describe and compare 4 metrics that may be used to identify fast-evolving sites. One of these metrics (i.e., the probability of compatibility) turned out to be particular good at identifying fast-evolving sites - when these sites are removed from the data, the result is an increase in the consistency of the alignment, and another phylogeny.

METHODS AND ALGORITHMS II

Rob Lanfear

Title: *Finding optimum partitioning schemes for the analysis of molecular sequence data*

Abstract: Models of molecular evolution usually represent crude simplifications of the ways that DNA sequences change over time. These simplifications can limit the accuracy of the inferences we make from DNA sequences. One common technique for overcoming these limitations is to estimate different models of molecular evolution for different subsets of sites (such as different codon positions) in a given DNA dataset - an approach known as partitioning. However, as the size and complexity of DNA datasets increases, it becomes increasingly difficult to choose an appropriate partitioning scheme for a given dataset. The aim is to define enough partitions to capture meaningful differences in molecular evolution between sites, while avoiding the definition of too many partitions and consequent over-parameterisation. This task can be difficult because the number of possible partitioning schemes can be extremely large even for modest DNA datasets. I present some straightforward approaches and software which can be used find optimum partitioning schemes for DNA datasets. Joint work with Simon Ho, University of Sydney.

Mathieu Blanchette

Title: *Ancestral genome reconstruction and its uses toward annotating the human genome*

Abstract: With the number of sequenced vertebrate genomes rapidly growing, the exciting prospect of being able to accurately infer ancestral genomes becomes within reach. In this presentation, I will discuss how ancestral DNA sequences can be inferred and how they can be then used to help addressing some key questions in genomics. Reconstructing ancestral sequences poses a number of algorithmic challenges. I will first describe some of our work on aligning orthologous sequence and inferring ancestral sequences, focusing on the accurate identification of insertions and deletions. Next, I will discuss how one can take advantage of the availability of inferred ancestral sequences to help at three important tasks: (i) identify non-coding sites under selection in the human genome; (ii) improve the detection of transcription factor binding sites; and (iii) determine the target gene(s) of long-range enhancers. Evolution has been conducting site-specific functionality assays for hundreds of millions of years. The ability to decipher the results of these experiments has and will continue to provide us with a wealth of information about our genome and the impact of mutations.

Arndt von Haeseler
Title: MISFITS: *evaluating the goodness of fit between a phylogenetic model and an alignment*
Abstract: As models of sequence evolution become more and more complicated, many criteria for model selection have been proposed, and tools are available to select the best model for an alignment under a particular criterion. However, in many instances the selected model fails to explain the data adequately as reflected by large deviations between observed pattern frequencies and the corresponding expectation. We present MISFITS, an approach to evaluate the goodness of fit. MISFITS introduces a minimum number of "extra substitutions" on the inferred tree to provide a biologically motivated explanation why the alignment may deviate from expectation. These extra substitutions plus the evolutionary model then fully explain the alignment. We illustrate the method on several examples and then give a survey about the goodness of fit of the selected models to the alignments in the PANDIT database.

Barbara Holland
Title: *Phylogenies from* DArTs - *Stochastic Dollo with censored data*
Abstract: Diversity Array Technologies (DArTs) are a relatively new kind of molecular marker system that seem like they could be usefully applied to phylogenetics (a few papers have already explored this). Like marker systems such as AFLP and RFLP, the method produces presence absence data, but unlike these methods it is very unlikely for shared presences to occur by chance. The basic idea is as follows. One or a small number of genomes are selected to form the genomic representation. Two enzymes are used to cut the DNA from these genomes at certain recognition sites (a rare 6bp recognition site and a more frequent 4bp recognition site). Fragments of DNA whose ends have been cut by two rare recognition sites are amplified. These fragments, which are said to form the genomic representation, are arranged on a microchip. Other genomes can then be checked to see which fragments within the genomic representation have copies of their own sequence. For each other genome that is compared to the genomic representation this results in a binary sequence that indicates presence (1) or absence (0) of each of the fragments. The first obvious advantage of this approach is that it creates a representation of the whole genome rather than just a few genes. This alleviates the problem of picking a small set of genes that may not be representative of the evolutionary history of the species. The second advantage is that in comparison to an individual site, long fragments of DNA are very unlikely to be similar due to chance. So if two species share a fragment it is vastly more likely that they share it due to common ancestry rather than due to a chance similarity. To use these data for phylogenetics it would be useful to develop a likelihood equivalent of Dollo parsimony (in which characters can be lost multiple times but gained only once), such models have already been explored in the context of language evolution and gene content evolution. However, another complicating issue is the censoring effect created by only being able to see those fragments that were in the original genomic representation, i.e. fragments that are shared by a group of species but that are not present in the original species used to make the genomic representation are missing from the data. Joint work with Dorothy Steane.