Banff International Research Station

for Mathematical Innovation and Discovery

New Mathematical Challenges from Molecular Biology
September 6 – 11, 2009

MEALS

*Breakfast (Buffet): 7:00 – 9:30 am, Sally Borden Building, Monday – Friday
*Lunch (Buffet): 11:30 am – 1:30 pm, Sally Borden Building, Monday – Friday
*Dinner (Buffet): 5:30 – 7:30 pm, Sally Borden Building, Sunday – Thursday
Coffee Breaks: 2nd floor lounge, Corbett Hall
***Please remember to scan your meal card at the host/hostess station in the dining room for each meal.**

MEETING ROOMS

All lectures will be held in Max Bell 159 (Max Bell Building accessible by walkway on 2nd floor of Corbett Hall). LCD projector, overhead projectors and blackboards are available for presentations. *Please note that the meeting space designated for BIRS is the lower level of Max Bell, Rooms 155-159. Please respect that all other space has been contracted to other Banff Centre guests, including any Food and Beverages in those areas.*

SCHEDULE

Sunday

16:00	Check-in begins (Front Desk – Professional Development Centre - open 24 hours)
17:30-19:30	Buffet Dinner

Monday

7:00-8:45	Breakfast
8:45-9:00	Introduction and Welcome to BIRS by BIRS Station Manager, Max Bell 159
9:00-9:40	Andy Clark *Next-generation sequencing and inference of gene conversion in tandem arrays*
9:50-10:30	Molly Przeworski *Causes and consequences of variation in human recombination*
10:30-11	Coffee Break
11:00-11:40	Rasmus Nielsen *Probabilities of identity by descent in the human genome*
11:30-13:00	Lunch
13:40-14:10	Anton Wakolbinger *When does Muller's ratchet click rarely?*
14:20-14:50	Guy Sella *Genomic signatures of adaptation in Drosophila*
15:00-15:30	Coffee Break
15:30-16:00	Ryan Gutenkunst *Inferring the joint demographic history of multiple populations*
16:10-16:40	Yun Song *Closed form sampling formulas for the coalescent with recombination*
17:30-19:30	Dinner
19:00-21:00	Reception – Max Bell Lounge (Fishbowl) FREE beverages

Tuesday

7-9	Breakfast
9-9:40	Jeff Jensen
	Identifying targets of selection in non-equilibrium populations
9:50-10:30	Allison Etheridge
	Evolution in a spatial continuum
10:30-11	Coffee Break
11-11:40	Bob Griffiths
	A Lambda-coalescent dual process in a Cannings model with genic selection
11:30-13:30	Lunch
13:00-13:30	Banff Centre Tour
13:30	Group Photo; meet on the front steps of Corbett Hall
13:40-14:10	Matthias Birkner
	Computing likelihoods under Lambda-coalescents
14:20-14:50	Ori Sarsayan
	A coalescent process with simulataneous multiple mergers
15-15:30	Coffee Break
15:30-16	Nathanel Berestycki
	The frequency spectra of Lambda-coalescents
16:10-16:40	Jason Schweinsberg
	The genealogy of branching Brownian motion with absorption
17:30-19:30	Dinner

Wednesday

7-9	Breakfast
9-9:40	Steve Evans
	Regulatory complex assembly, transcription, and traversal times of Markov chains
9:50-10:20	Paul Joyce
	A general extreme value theory model for adaptation of DNA sequeunces
10:30-11	Coffee Break
11-11:40	Rick Durrett
	Probability models of cancer development and progression
11:30-13:30	Lunch
	Free Afternoon – A bus will leave from the Professional Development Center at 1:15PM for Lake Louise returning at 6PM.One can walk along the lake or hike up a small amount of elevation to the Tea House
17:30-19:30	Dinner

Thursday

7-9	Breakfast
9-9:40	Bjarki Eldon
	A two locus diploid Moran model with recombination
9:50-10:20	Jay Taylor
	Genealogical consequences of fecundity variance polymorphism
10:30-11	Coffee Break
11:10-11:40	Frederick Matsen
	Polyhedral geometry and phyolgenetic tree inference on non tree metrics
11:30-13:30	Lunch
13:40-14:10	Martin Moehle
	Duality and asymptotics for a class of non-neutral discrete Moran models
14:20-14:50	Peter Pfaffelhuber
	The distributed genomes of bacterial populations
15-15:30	Coffee Break

15:30-	Informal discussions
18:00-20:00	BBQ – Donald Cameron – Outside Patio and Function Room 6

Friday

7:00-9:00	Breakfast
9:00-?	Informal Discussions – in the Banff Center or on hiking trails
11:30-13:30	Lunch

Checkout by 12 noon.

Banff International Research Station

for Mathematical Innovation and Discovery

New Mathematical Challenges from Molecular Biology
September 6 – 11, 2009

ABSTRACTS

Nathanel Berestycki (Cambridge University)
The frequency spectra of Lambda-coalescents

Abstract: I will discuss some recent progress on the structure of the allelic partition associated with a Lambda-coalescent in the infinite alleles model and the frequency spectrum in the infinite sites model. In particular, in the case where the measure Lambda has a density which decreases as a power law near 0, we obtain a strong law of large numbers for both frequency spectra. The proof is essentially based on martingale arguments and a Tauberian theory of random partitions developed by Gnedin, Hansen and Pitman. Joint work with Julien Berestycki and Vlada Limic.

Matthias Birkner (Weierstrass-Institut fuer Angewandte Analysis und Stochastik)
Computing likelihoods under Lambda-coalescents

Abstract: We present and compare various importance sampling methods which allow to compute likelihoods of sequence data in the context of multiple merger coalescents, generalising results of Griffiths and Tavaré (1994), Stephens and Donnelly (2000) and Hobolth, Uyenoyama and Wiuf (2008) to the Lambda-case. We illustrate these methods using simulated and some real datasets. (Joint work with Matthias Steinrücken and Jochen Blath, TU Berlin)

Andrew Clark (Cornell University)
Next-generation sequencing and inference of gene conversion in tandem gene arrays

Abstract: Gene duplication and amplification are important evolutionary processes for tandemly arranged multigene families. In the 1980s, Tomoko Ohta, Thomas Nagylaki and others developed models to describe population genetic variation in tandem gene clusters under the processes of mutation, drift and gene conversion. Fitting these models and parameter estimation have been hindered by the empirical challenge to distinguish orthologous and paralogous gene copies in tandem gene arrays. Whereas tandem multigene repetitive arrays had been largely inaccessible or problematic to assemble with traditional sequencing techniques, the use of short read sequence data now allows us to perform such analyses across an entire gene array, and without requiring assembly. We do so using a novel approach based on properties of the multi-alignment to a single repeat unit. Such alignments confound orthologous and paralogous repeats, but algebraic manipulation allows us to distinguish these relations. We apply the method to investigate the evolutionary dynamics of four tandem arrays in Drosophila melanogaster: rDNA, Stellate, su(Ste), and the 1.688 satellite sequence. We determine that, in order to obtain results consistent with observed gene identities, the rate of gene conversion is nearly three orders of magnitude greater than the rate of mutation within each array, and additionally demonstrate that these results even hold true with much lower sequencing coverage. This method can therefore be applied to nearly all short read sequencing data sets from genomic DNA, making this a powerful technique for estimating rates of gene conversion and mutation. This is joint work with Melanie A. Huntley.

Rick Durrett (Cornell U.)
Probability models of cancer development and progression

Abstract: Cancer is the end result of a number of mutations. This raises the probability question of studying the waiting time until a prespecified sequence of mutations occurs in some individual in a population of cells. One is interested both in the case of a fixed population size or an exponentially growing tumor. I will describe results obtained in several papers that are joint work with Deena Schmidt, Jason Schweinsberg, Stephen Moseley, and John Mayberry.

Bjarki Eldon (Harvard University)
A two locus diploid Moran model with recombination

Abstract: A two loci diploid Moran population model which allows for large number of offspring is considered. The main goal is to derive the ancestral process including recombination, but excluding selfing. Consider first a single locus model without selfing. In each timestep an offspring is created by choosing two gametes without replacement. Since parents always persist, a population of 2N gametes is composed of four `parent' gametes from the previous timestep, two `offspring' gametes, and 2N -6 gametes in the remaining N - 3 diploid individuals that didn't participate in the reproduction event of the previous timestep. Looking back in time, this means that a coalescent event can only occur when ancestral lines include lines from parents and the offspring. Assuming that one offspring arises each timestep, the ancestral process is the usual Kingman coalescent. If the number of offspring is proportional to the population size, we obtain multiple and simultaneous multiple merger coalescent processes. Since we allow only one pair of parents in each timestep, we can obtain at most four simultaneous mergers.

We include recombination and consider the ancestral process for a sample of size two at two loci. Forward in time, two diploid individuals are chosen to create gametes in the usual way. Two gametes, one from each parent, are then chosen uniformly at random to create a diploid offspring. A key characteristic of this model is that the only gametes that can change between timesteps are the offspring gametes. Convergence to a continuous-time ancestral process is obtained by scaling recombination to N timesteps, but with time rescaled in units of N^2 timesteps. The presence of parent gametes leads to a sparse rate matrix of the ancestral process compared to one obtained from Wright-Fisher reproduction. Correlations in coalescence times between the two loci are therefore essentially zero, which agrees only with the timescale of recombination.

Allison Etheridge (Oxford)
Evolution in a spatial continuum.

Abstract: Classical models for gene flow fail in (at least) three ways. First they cannot explain patterns in data observed over large scales, second they predict much more genetic diversity than is observed and third they assume that genetic loci evolve independently. In collaboration with Nick Barton we recently proposed a new framework for modelling populations evolving in a spatial continuum that addresses these issues. In this talk we describe the new framework and some asymptotic results for the genealogy of a sample from a population living on a two-dimensional torus. These results are joint work with Amandine Veber.

Steve Evans (University of California, Berkeley)
Regulatory complex assembly, transcription, and traversal times of Markov chains.

Abstract: A fundamental organizational part of the complex, highly organized biological processes we see around us takes place at a molecular level, as randomly diffusing proteins interact with individual gene loci to trigger the production of other molecular signals. Networks of such interactions can generate precise, reliable behaviors, despite the often large dependence of individual reactions on molecular concentrations and the inherent stochasticity of each interaction. One of the best studied examples of emergent reliability and precision among regulatory networks of gene expression comes from the dorsal-ventral patterning system in Drosophila. Patterning in this system is driven by localized, ventral activation of a uniformly expressed embryonic protein called Dorsal. The activated molecules diffuse dorsally and shuttle into the yet undifferentiated nuclei. Binding of Dorsal to regulatory DNA sequences in the nuclei activates a network of other transcription factors and signaling proteins. Five or more distinct domains of gene expression are delineated by this network, giving rise to distinct cell fates. The boundaries between different cell types are determined to single cell precision and are reliably made at each of the hundred odd cell-junctions along the length of each boundary, despite variations in environmental gene dose and protein concentration. We present a class of Markov models of regulatory complex assembly and transcription, and develop various analytical tools for the purposes of studying which network topologies are faster, less noisy, or more robust than others. This is joint work with Alistair Boettiger (U.C. Berkeley Biophysics), Peter Ralph (U.C. Berkeley Statistics and U.C. Davis Evolution and Ecology), and Michael Levine (U.C. Berkeley Molecular and Cell Biology).

Robert Griffiths (University of Oxford)
A Lambda-Coalescent Dual Process in a Cannings model with Genic Selection

Abstract: A coalescent dual process for a multi-type Cannings model in continuous time with genic selection can be derived using a generator approach. A graphical representation of the Cannings model identifies the dual as a strong dual process following typed lines backwards in time. The general limit population process as the population size tends to infinity is a process with genic selection that has Λ-jumps and possibly a diffusive component. The dual process is a Λ-coalescent with selection. Transition functions in the population process have a mixture representation in terms of the dual process transition functions. The Moran model with selection and its diffusion limit are special cases. The dual process then mirrors the Ancestral Selection Graph of Krone and Neuhauser (1997) and Neuhauser and Krone (1997), which allows one to reconstruct the genealogy of a random sample from a population subject to genic selection. The process extends a dual process construction in a Wright-Fisher diffusion in Barbour, Ethier and Griffiths (2000) and in the Moran model by Etheridge and Griffiths (2009). Joint research with Alison Etheridge.

Ryan Gutenkunst (Los Alamos National Labs)
Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data

Abstract: Building demographic models is a central problem in population genetics. Coalescent approaches have played an important role in these endeavors but suffer computationally as model complexity increases. We introduce an inference method based on the joint frequency spectrum of allelic variants within and between populations. For candidate demographic models (which potentially include selection), we numerically compute the expected spectrum using a diffusion approximation to the dynamics of the one locus, two-allele Wright-Fisher process. Our current implementation supports up to three simultaneous populations, and its efficiency lets us use conventional and parametric bootstrap resampling to estimate uncertainties for parameters and significance values for hypothesis tests. We apply our method to human expansion out of Africa and the settlement of the New World, inferring a model for African, European, East Asian, and Mexican populations. Combining our demographic model with a previously estimated distribution of selective effects for new amino acid mutations also accurately predicts the observed spectrum of nonsynonymous variants. Our methods and the models inferred from it offer new tools for studying the history and evolution of both our own species and others.

Jeff Jensen (UC Berkeley)
Identifying targets of positive selection in non-equilibrium populations

Abstract: Adaptation is a central focus of biology, although it can be difficult to identify both the molecular mechanisms and the agent of selection causing change. Methodologically, progress is being made towards this identification by combining both site frequency spectrum and linkage disequilibrium based hitchhiking predictions, in to a single framework. Experimentally, exciting applications of these newly developing statistics are underway. In the fly Drosophila miranda, these approaches have been used to characterize recent positive selection on a newly formed X chromosome, helping to elucidate the complicated acquisition of dosage compensation binding sites in the rapid transition from autosome to sex chromosome. In the mouse Peromyscus maniculatus, these statistics were utilized in order to localize a putatively selected region, with subsequent functional analyses demonstrating that a single amino acid deletion is in fact responsible for cryptic coat coloration. Thus, this combination of improved statistical techniques, genomic data, and functional genetics is just beginning to provide what will no doubt be a long list of 'textbook' examples of natural selection.

Paul Joyce (University of Idaho)
A General Extreme Value Theory Model for the Adaptation of DNA Sequences under Strong Selection and Weak Mutation

Abstract: Adaptive evolution refers to a genetic change that confers a strong selective advantage to individuals that inherit this change. Examples include genetic changes that make microbes drug-resistant, or insects pesticide-resistant. Recent theoretical studies of the adaptation of DNA sequences assume that the distribution of fitness effects among new beneficial mutations is exponential. This has been justified by using extreme value theory and, in particular, by assuming that the distribution of fitnesses belongs to the so-called Gumbel domain of attraction. However, extreme value theory shows that two other domains of attraction are also possible: the Frech?and Weibull domains. Distributions in the Frech? domain have right tails that are heavier than exponential, while distributions in the Weibull domain have right tails that are truncated. To explore the consequences of relaxing the Gumbel assumption, we generalize previous adaptation theory to allow all three domains. We find that, while the shape of the distribution of selection coefficients among beneficial mutations can vary dramatically across the domains of attraction, previous Gumbel-based predictions about the first step of adaptive walks are remarkably robust.

Frederick Matsen (Berkeley)
How can a single bad apple spoil the bunch? Using polyhedral geometry to understand phylogenetic inference on non tree metrics

Abstract: It is well known among phylogeneticists that adding an extra taxon (e.g. species) to a data set can alter the structure of the optimal phylogenetic tree in surprising ways. However, very little is known about this phenomenon, and nothing has been done from a mathematical standpoint. In particular, given starting data which perfectly fits a given tree, how "strange" does the data for the extra taxon have to be such that the resulting optimal tree for the combined data set does not contain the original tree? Also, what trees can be phylogenetically optimal when starting with a combined data set of this type? We are able to completely answer questions of this sort by applying the tools of polyhedral geometry to analyze optimal trees under the balanced minimum evolution (BME) criterion. This is joint work with Maria Angelica Cueto.

Martin Moehle (University of Tuebingen)
Duality and asymptotics for a class of non-neutral discrete Moran models

Abstract: A Markov chain X with finite state space $\{0,...,N\}$ and tridiagonal transition matrix is considered where transitions from i to i-1 occur with probability $(i/N)(1-p(i/N))$ and from i to i+1 with probability $(1-i/N)p(i/N)$, where $p:[0,1] \to [0,1]$ is a given function. It is shown that, if p is continuous with $p(x) \leq p(1)$ for all $0 \leq x \leq 1$, then for each N a dual process Y to X (with respect to a specific duality function) exists if and only if 1-p is completely monotone with $p(0)=0$. A probabilistic interpretation of Y in terms of an ancestral process of a multi-type Moran model with a random number of types is presented. It is shown that under weak conditions on p the process Y, properly time-and space-scaled, converges to an Ornstein-Uhlenbeck process as N tends to infinity. The asymptotics of the stationary distribution of Y is studied as N tends to infinity. Examples are presented involving selection mechanisms. (Joint work with Thierry Huillet.)

Rasmus Nielsen (UC-Berkeley)
Probabilities of identity by descent and selection in the human genome.

Abstract: There has recently been considerable interest in detecting natural selection in the human genome. Selection will usually tend to increase Identity By Descent (IBD) among individuals in a population, and many methods for detecting ongoing positive selection on new mutations indirectly take advantage of this. In this talk we show that excess IBD sharing is a more general property of natural selection, enabling detection of selection that is otherwise difficult to detect, such as selection acting on standing genetic variation. We use a recently developed method for identifying IBD segments among individuals from genome-wide data to scan populations from the new HapMap phase 3 project for regions that have been under strong very recent selection. The HLA region is by far the region showing the most extreme signal, suggesting that much of the recent selection acting on the human genome has been immune related and acting on HLA loci. As equilibrium overdominance does not tend to increase IBD, we argue that this type of selection is not the only selection acting in the human MHC region.

Peter Pfaffelhuber (Freiburg)
The distributed genome of bacterial populations

Abstract: The genome of bacteria is less stable than the genome of eucaryotes. In particular, it is an empirical observation that bacteria from the same population carry different genes. We study a model where new genes are introduced from the environment and can be lost along ancestral lines. In addition, we randomize the genealogy according to Kingmans coalescent. This mutation model, the infinitely many genes model, appears to be new in the population genetic literature. We obtain some mathematical results which fit well with empirical data. This is joint work with Franz Baumdicker and Wolfgang Hess, University of Freiburg.

Molly Przeworski (Chicago)
Causes and consequences of variation in human recombination

Abstract: Recombination is a fundamental process that helps to align chromosomes, ensure proper disjunction and maintain genome integrity. These roles impose a number of constraints on the number and placement of recombination events on each chromosome. In humans, errors in the recombination process can lead to aneuploidy and chromosomal rearrangements, highly deleterious outcomes. Yet in spite of its essential roles, recombination varies markedly among humans. To characterize this variation and understand some of its consequences, we have focused on the analysis of dense, genome-wide genotyping data collected in a large pedigree of Hutterite individuals. Here, I will present two of our main findings: (i) Extensive variation in fine-scale recombination patterns among humans, notably in the use of recombination "hotspots". (ii) Evidence for the requirement for one crossover per chromosome rather than per arm to ensure proper disjunction, with additional crossovers occurring in proportion to physical length. Interestingly, we find that the requirement is not absolute, as chromosome 21 seems to be frequently transmitted properly in the absence of a crossover in females. This finding that raises the possibility of a back-up mechanism aiding in its correct segregation. This work is joint with Graham Coop (UC Davis) and Carole Ober (U Chicago).

Ori Sargsyan (Harvard)
A Coalescent Process with Simultaneous Multiple Mergers for Approximating the Gene Genealogies of Many Marine Organisms

Abstract: We describe a forward-time haploid reproduction model with a constant population size that includes life history characteristics common to many marine organisms. We develop coalescent approximations for sample gene genealogies under this model and use these to predict patterns of genetic variation. Depending on the behavior of the underlying parameters of the model, the approximations are coalescent processes with simultaneous multiple mergers or Kingman's coalescent. Using simulations, we apply our model to data from the Pacific oyster and show that our model predicts the observed data very well. We also show that a fact which holds for Kingmans coalescent and also for general coalescent trees that the most-frequent allele at a biallelic locus is likely to be the ancestral alleleis not true for our model. Our work suggests that the power to detect a sweepstakes effect in a sample of DNA sequences from marine organisms depends on the sample size. Joint work with John Wakeley.

Jason Schweinsberg (University of California at San Diego)
The genealogy of branching Brownian motion with absorption

Abstract: Motivated by recent work of Brunet, Derrida, Mueller, and Munier concerning the genealogy of a population in the presence of selection, we consider branching Brownian motion in which particles experience a negative drift and are killed upon reaching zero. We show that the number of particles, properly scaled, converges to a continuous-state branching process. We also show that the genealogy of the particles can be described by a coalescent process known as the Bolthausen-Sznitman coalescent. This is joint work with Julien Berestycki and Nathanael Berestycki.

Guy Sella (Hebrew University)
Genomic Signatures of Adaptation in Drosophila

Abstract: A growing body of research over the past decade is challenging our view of molecular evolution by suggesting that a substantial fraction of the divergence between species may in fact be adaptive and that these adaptations may frequently influence patterns of polymorphism. The most compelling evidence comes from studies in Drosophila, where both McDonald-Kreitman based estimates of the rate of adaptive evolution and analysis of genome-wide patterns of polymorphism, believed to be associated with selective sweeps, suggest widespread and frequent adaptations. I shall describe recent work aimed at better characterizing the signatures of selective sweeps in the Drosophila simulans genome with the intention of inferring genome-wide adaptive parameters in this and other species.

Yun Song (UC Berkeley)

Closed-form sampling formulas for the coalescent with recombination

Abstract: For a given population genetics model, the probability of observing a sample of DNA sequences plays a fundamental role in various applications, but closed-form sampling formulas are generally very difficult to obtain. In particular, when recombination is involved, obtaining an analytic formula for the sampling distribution has so far remained an intractable problem. In this talk, I will focus on the two-locus case with an arbitrary model of mutation and show that it is possible to obtain useful closed-form sampling formulas when the recombination rate is large. I will discuss the accuracy of our asymptotic sampling formula for a wide range of recombination rates and suggest a concrete application in the context of the composite-likelihood method. Lastly, using our sampling formula, I will describe a simple sufficient condition for a given two-locus sample configuration to have a finite maximum likelihood estimate (MLE) of the recombination rate. This condition is the first analytic result on the classification of the MLE, and is instantaneous to check in practice, provided that one-locus probabilities are known. Joint work with Paul A. Jenkins.

Jay Taylor (Oxford)
The Genealogical Consequences of Fecundity Variance Polymorphism

Although analyses of natural selection mainly focus on differences in mean fecundity or survival, selection can also act on fecundity variance. In this talk I will describe some of the genealogical consequences of both within- and between-generation fecundity variance polymorphism. Perhaps the most surprising result is that in these models, there are infinitely many combinations of genotype-dependent fecundity distributions that have the same diffusion approximation but distinct coalescent processes, i.e., ancestral processes and allele frequency dynamics are not in one-to-one correspondence even when the existence of a moment dual suggests that they should be. These results make use of a
generalization of the structured coalescent processes introduced by Kaplan, Darden and Hudson (1988). I will also describe a graphical representation of these coalescents and show that for sufficiently large differences in the fecundity variance of different genotypes, the number of branches in the graph can undergo finite time blow-up.

Anton Wakolbinger (Goethe-University Frankfurt)

When does Muller's ratchet click rarely?

Abstract: A classical version of Muller's ratchet (Haigh 1978) has the three parameters N, λ and s, where N is the size of the (haploid) popluation, λ is the (stepwise) mutation rate, and s is the decrease of fitness per mutation. For large N s and N λ, the size of the best class "in equilibrium" is approximately N $\exp(-\lambda/s)$. We present a couple of heuristic arguments indicating that the size of N s $\exp(-\lambda/s)$ is critical for whether the losses of the best class (the "clicks of the ratchet") happen regularly or rarely on the scale N $\exp(-\lambda/s)$. This builds on joint work with Alison Etheridge and Peter Pfaffelhuber.