Miniworkshop
Multiple merger coalescents: population models and inference

Johannes Gutenberg University Mainz
2nd - 4th September 2019

Programme

Monday, 2nd September

14:00 Welcome

14:05 - 15:05 **Katrín Halldórsdóttir**: *Population genomics of highly fecund codfish - Going from samples to site frequency spectra -*

15.05 - 16:05 **Asger Hobolth**: *Phase-type distributions in population genetics*

16:05 - 16:35 Coffee break

16:35 - 17:35 **Maite Wilke Berenguer**: *The on-off-coalescent*

17:40 Wine and cheese

Tuesday, 3rd September

9.30 - 10:30 **Fabian Freund**: *Distinguishing coalescent models - which statistics matter most?*

10:30 - 11:00 Coffee break

11:00 - 12:00 **Bjarki Eldon**: *Experiments with the Schweinsberg model*

12:00 - 14:00 Lunch break

14:00 - 15:00 **Jere Koskela**: *Detecting multiple mergers from sequence data*

15:00 - 15:30 Coffee break

15:30 - 18:00 *Software-Afternoon*: **Bjarki Eldon, Fabian Freund, Jere Koskela**

19:00 Dinner

Wednesday, 4th September

9.30 - 10:30 **Martin Möhle**: *On the block counting process and the fixation line of the Bolthausen-Sznitman coalescent*

10:30 - 11:15 Coffee break

11:15 - 11:45 **Thibaut Sellinger**: *The Sequentially Markovian Beta Coalescent*

11:45 - 14:00 Lunch break

14:00 - 15:00 **Götz Kersting**: *Tree lengths for Λ-coalescents without a dust component and the site frequency spectrum of the Bolthausen-Sznitman coalescent*

15:00 - 16:00 Coffee and discussion

Abstracts

Population genomics of highly fecund codfish - Going from samples to site frequency spectra -

Katrín Halldórsdóttir (University of Iceland)

Introduction

As a biologist, I will focus my talk on the biology side of working with genetic data. First I will go through some terminology used in genetics and explain where the data are coming from. What is the advantage of having the whole genome sequenced instead of a single locus? I will introduce our model system, the codfish, their high- fecundity and discuss what high fecundity with skewed offspring distribution means.

Generation and nature of the data

Second, I will go through how the data are produced, the pros and cons of different techniques in sequencing. I will go through the format and main steps in generating site frequency spectra data from raw sequencing data and the nature of different datasets i.e. what is the right size of a sample under different biological questions. Next- generation sequencing can generate large amounts of data from a large number of individuals. But there are errors both in sequencing and in the alignment of raw sequence reads to a reference genome. For downstream analysis, we, therefore, use methods based on genotype likelihoods which are based on sequence and mapping quality scores.

Some examples of different statistical methods - Problems of the data

Finally, I will show some example of our own data. I will show how different filtering, sequencing coverage, different likelihood approaches and alignment methods can affect the final SNP sets that the site frequency spectrum is based on and can give different results. I will connect the results to the real biology of the organism behind the data, in our case the Atlantic cod, and how to estimate the fit of different models in the context of the nature of its life history, the burden of fisheries and possible population size.

Phase-type distributions in population genetics

Asger Hobolth (Aarhus University)

Probability modelling for DNA sequence evolution is well established and provides a rich framework for understanding genetic variation between samples of individuals from one or more populations. We show that the standard and more sophisticated coalescent models can be described in terms of the so-called phase-type theory, where complicated and tedious calculations are circumvented by the use of matrix manipulations. The application of phase-type theory in population genetics consists of describing the biological system as a Markov model by appropriately setting up a state space and calculating the corresponding intensity and reward matrices. Formulae of interest are then expressed in terms of these aforementioned matrices. We illustrate this by a number of examples with particular focus on the multiple merger coalescent. We believe that phase-type theory has great potential as a tool for analysing probability models in population genetics. The compact matrix notation is useful for clarification of current models, and in particular their formal manipulation and calculations, but also for further development or extensions. This is joint work with Mogens Bladt and Arno Siri-Jegeousse.

The on-off-coalescent

Maite Wilke Berenguer (Universität Bocum)

The seed bank coalescent is a coalescent structure that arises in populations with a "strong" seed bank effect and describes the genealogy of such a population where lineages can coalesce according to a Kingman-mechanism or migrate in and out of the seed bank independently at a constant rate. We introduce an extension of this model where both types of transitions can each be correlated leading to Λ-coalescences and simultaneous migrations in and out of the seed bank. We discuss the similarities between the mechanisms and a characterization of coming down from infinity for a special case of the on-off-coalescent (with Kingman coalescences).

Distinguishing coalescent models - which statistics matter most?

Fabian Freund (University of Hohenheim)
joint works with A. Siri-Jégousse (IIMAS, UNAM Mexico City), F. Menardo, S. Gagneux (Swiss Tropical and Public Health Institute Basel), K. Schmid, M. Vidal (University of Hohenheim)

Many inference methods developed to distinguish between multiple-merger and bifurcating genealogy models rely on the genetic information summarised in the allele frequency spectrum. With an Approximate Bayesian Computation approach based on random forests we analyse whether using further or different diversity statistics, e.g. measures of linkage disequilibrium, decrease inference errors compared to using only allele frequency information. This includes a new set of statistics, the minimal observable clade sizes, i.e. the minimal allele count > 1 of each sampled individual, for which we present some mathematical properties.
For a set of statistics distinguishing multiple-merger n-coalescents well from bifurcating genealogy models, we employ this Approximate Bayesian Computation scheme to select the best fitting model for a variety of samples of *Mycobacterium tuberculosis* and for one sample of the fungal crop pathogen *Exserohilum turcicum*.

Experiments with the Schweinsberg model

Bjarki Eldon (Museum für Naturkunde Berlin)

We consider the impact of large sample size on gene genealogies of samples drawn from natural highly fecund populations with sweepstakes reproduction. To this end we consider a modified Schweinsberg model of genetic reproduction, which incorporates a cut-off to the random number of juveniles contributed by a given individual. Depending on how the cut-off behaves relative to the population size, we obtain the Kingman-coalescent, a truncated Beta-coalescent, or the original Beta-coalescent of (Schweinsberg , 2003). The cut-off enables us to estimate the error of the coalescent approximation. The error estimates reveal that convergence can be very slow, and very small sample size can be sufficient to invalidate convergence, especially if the cut-off is of a specific form. However, the impact on inference will only be noticed at larger sample size than that at which convergence breaks down. We also investigate the applicability of the Schweinsberg model to inference from genetic and genomic data, and discuss possible further extensions, in particular to the genomic level. This is joint work with Jonathan A. Chetwynd-Diggle and Alison Etheridge.

Detecting multiple mergers from sequence data

Jere Koskela (University of Warwick)

Different coalescent models yield different predictions of observed DNA sequence diversity. An excess of singleton mutations is a prominent signal of a number of evolutionary forces, including high family size variance, and population growth. I will introduce a surprisingly simple, two dimensional summary statistic for SNP data; demonstrate that it distinguishes between these two scenarios with very high confidence; show that it is robust to some (but not all) biological confounders such as natural selection population structure; and discuss implications for an Icelandic cod data set in which an excess of singletons is clearly visible.

Software-Afternoon

Bjarki Eldon, Fabian Freund, Jere Koskela

The Software-Afternoon is an experimental format. Bjarki, Fabian and Jere agreed to help us gain an overview of the existing software for simulating and analysing data and of the particular features of these software. Furthermore, we ask: what would one like that such software achives? The question is thought as a discussion starter and we would be very glad to have a lively discussion on the topic. Participants are also welcome to bring their laptops and experiment with the software presented.

On the block counting process and the fixation line of the Bolthausen-Sznitman coalescent

Martin Möhle (Universität Tübingen)

The block counting process and the fixation line of the Bolthausen-Sznitman coalescent are analyzed. It is shown that these processes, properly scaled, converge in the Skorohod topology to the Mittag-Leffler process and to Neveu's continuous-state branching process respectively as the initial state tends to infinity. Strong relations to Siegmund duality, Mehler semigroups and self-decomposability are pointed out. Furthermore, spectral decompositions for the generators and transition probabilities of the block counting process and the fixation line of the Bolthausen-Sznitman coalescent are provided leading to explicit expressions for functionals such as hitting probabilities and absorption times. Extensions to exchangeable coalescents are discussed.

The Sequentially Markovian Beta Coalescent

Thibaut Sellinger (Technische Universität München)

Several methods based on the Sequential Markovian Coalescent (SMC) have been developed to use full genome sequence data to uncover population demographic history or life history trait. Those estimation can be of interest when generating a null model for selection tests. While these methods can be applied to all possible species, the underlying assumptions are those of a Wright-Fisher model. However, some species can be highly fecund, which violates the Wright-Fisher model. This phenomenon can lead to bias if not accounted for and be hard to detect when population size cannot be assumed constant in time. We here develop a novel SMC-based method to infer 1) the fecundity through a Beta-coalescent, and 2) the populations' past demographic history.

Tree lengths for Λ-coalescents without a dust component and the site frequency spectrum of the Bolthausen-Sznitman coalescent

Götz Kersting (Goethe Universität Frankfurt)

Branch lengths in Λ-coalescents are important from a biological point of view, because they have an impact on the number of mutations to be placed on the coalescent (e.g. in the infinite sites model). Up to now branch lengths have been investigated mainly for special classes, as the Kingman coalescent or Beta-coalescents. In this talk we present results on Λ-coalescents under the broad assumption that the coalescent lacks a dust component, e.g. for the total tree lengths or the total external lengths. The proofs rely on a new method for proving certain laws of large numbers, which appears promising for general decreasing Markov chains. A specific application concerns the site frequency spectrum of the Bolthausen-Sznitman coalescent. (Joint work with Christina Diehl)