LMS Durham Symposium:
Mathematical Genetics
Monday 5th July - Thursday 15th July 2004

University of Durham, United Kingdom

List of abstracts

Shola Ajayi (University of Reading) Thursday 8th July 15:00

On calibrating estimates of the recombination parameter in a dynamic population

The size of a suitable estimate of the index of linkage disequilibrium is recognised as a signature of either an underlying genetic event or the dynamism of aspects of the demographic history of sampled chromosomes depending on the assumed model of evolution. We use simulation to study the precision of estimates of suitable parameters in both changing and constant pupulation scenarios.

Ellen Baake (University of Vienna) Wednesday 7th July 09:00

Ancestry in mutation-selection models

(Joint work with M. Baake, A. Bovier, J. Hermisson, H.-O. Georgii)

We consider the genetics of populations under the joint action of mutation and selection. To this end, mutation and reproduction are modelled as a multi-type branching process, of which we consider both the forward and the backward direction of time. The stationary state of the reversed process is the ancestral distribution, which turns out as a key for the study of mutation-selection balance. In particular, a general variational principle is available that relates the present population, the ancestral population, and the mean fitness. If the mutation process is reversible, mutation rates decay fast enough with distance to the target type, and the fitness function has certain symmetries, this maximum principle boils down to a low-dimensional problem that can often be solved explicitly.

Nick Barton (University of Edinburgh) Monday 12th July 09:00

Distinguishing causes of reduced diversity

Regions of genome with exceptionally low diversity are often attributed to "selective sweeps", in which fixation of a favourable mutation eliminates genetic variation. However, in structured populations there is an appreciable chance that long stretches of genome with low diversity will arise. Moreover, these have similar characteristics to those generated by selective sweeps. This may make it impossible to use sequence diversity alone to detect the action of selection.

Mark A. Beaumont (University of Reading) Thursday 8th July 10:00

Approximate Bayesian methods in genetic data analysis: some applications

A method of Bayesian computation is to compress the data into a list of summary statistics, s, simulate samples from the joint distribution P(Phi,S) of parameters and summary statistics, and then calculate the posterior distribution as the conditional distribution P(Phi | S=s). The degree of approximation depends on how well the data can be summarised, and

on how well the conditional distribution can be estimated. A brief overview is given of the various ways different researchers have tried to achieve this. One approach (Beaumont, Zhang & Balding, 2002) is to use regression methods to model P(Phi | S) in the vicinity of s. Twin goals here are a) to obtain a close to sufficient set of statistics, b) to use information from a large proportion of the points simulated from the joint distribution. I speculate on how these goals might be achieved. I describe recent applications of the method to the estimation of effective size in Wright-Fisher models, and to the estimation of recombination rate in the coalescent.

Matthias Birkner (Weierstrass Institute for Applied Analysis and Stochastics) Thursday 8th July 12:00

Spatial critically branching particle systems with state-dependent branching rate

The long-time behaviour of infinite, spatially distributed systems of critical branching particles is determined by the interplay of two competing mechanisms: fluctuations caused by thecritical (i.e. mean offspring per individual equals one) branching roughen the density landscape, driving the system towards local extinction, whereas the independent motion of particles has a smoothing effect. In the classical case of independent branching, stable long-time behaviour is possible iff the symmetrised individual motion is transient.

We consider a model where the branching rate of a particle is a function of the total number of particles at the given site, thus extending the classical model (corresponding to a linear function). We prove a comparison result for pairs of systems with ordered branching rate functions, showing that a larger branching rate function facilitates local extinction.

We study two natural examples in more detail: 1) In low dimensions, can we obtain stable long-time behaviour by strongly down-regulating the branching rate in crowded regions? The strongest conceivable down-regulation leads to "lonely branchers", where branching is only allowed whenever individuals are alone at their site. We conjecture that even this system will die out locally in low dimensions. We give some evidence by studying a caricature of the corresponding (spatially embedded) genealogical trees.

2) In high dimensions, where classical equilibria exist, we find a new regime of equilibria in the case of a quadratic branching rate function: Unlike the classical case, the variance of the number of individuals per site can become infinite.

Jochen Blath (University of Oxford) Tuesday 6th July 12:30

Coexistence in locally regulated competing populations

(Joint work with AM Etheridge and ME Meredith)

We propose two models of the evolution of a pair of competing populations. Both are lattice based. The first is a compromise between fully spatial models, which do not appear amenable to analytic results, and interacting particle system models, which don't, at present, incorporate all the competitive strategies that a population might adopt. The second is a simplification of the first in which competition is only supposed to act within lattice sites and the total population size within each lattice point is a constant. In a special case, this second model is dual to a branching-annihilating random walk. For each model, using a comparison with N-dependent oriented percolation, we show that for certain parameter values both populations will persist for all time with positive probability. We also present a number of conjectures relating to the role of space in the survival probabilities for the two

populations.

Leonardo Bottolo (Department of Statistics, University of Oxford) Thursday 8th July 15:00

The relationship between haplotype blocks and local recombination rates

(Joint with work the Oxford Statistics HapMap Analysis Group)

There has been considerable recent attention paid to the "block-like" structure of LD in much of the human genome. Such a simple structure of LD potentially allows considerable efficiency gains in disease association studies. There has been a plethora of methods for defining, describing, and characterising blocks.

In parallel with these observations has been the development of coalescent based statistical methods for estimating human recombination rates over scales of kilobases, from polymorphism data. These show a complex pattern of rate variation, with recombination hotspots a ubiquitous feature of the human genome, and considerable variation in recombination rates (by 3-4 orders of magnitude) outside the hotspots. Much human recombination occurs in hotspots, but around half of the recombination events occur outside them.

In this study we use fine-scale estimates of recombination rates to examine the structure of haplotype blocks, and in particular block boundaries, with respect to recombination rate. Whilst hotspots should break blocks, the typical spacing of hotspots (perhaps every 150-200kb on average) means that they cannot explain all gaps between blocks. Using a Bayesian hierarchical model, we thus also examine the extent to which block boundaries can be well explained by local genetic distance.

Reinhard Buerger (University of Vienna) Wednesday 7th July 11:30

A multilocus analysis of intraspecific competition and stabilizing selection on a quantitative trait

The equilibrium structure of an additive, diallelic multilocus model of a quantitative trait under frequency- and density-dependent selection is derived. The trait is under stabilizing selection and mediates intraspecific competition as induced, for instance, by differential resource utilization. It is assumed that stabilizing selection is weak, but the strength of competition may be arbitrary relative to it. Density dependence is caused by population regulation, which may be of a very general kind. Number and effects of loci are arbitrary and stabilizing selection is not necessarily symmetric with respect to the range of phenotypic values. All previously studied models of intraspecific competition for a continuum of resources known to the author reduce to a special case of the present model if overall selection is weak. Therefore, in this case our results are applicable as approximations to all these models. Our central result is the (nearly) complete characterization of the equilibrium structure in terms of all parameters. It is derived under the sole assumption that selection is weak enough relative to recombination to ignore linkage disequilibrium. In particular, necessary and sufficient conditions on the strength of competition relative to stabilizing selection are found that ensure the maintenance of multilocus polymorphism and the occurrence of disruptive selection. In this case, explicit formulas for the number of polymorphic loci at equilibrium, the allele frequencies, the genetic variance, and the strength of disruptive selection are obtained. For two loci, the effects of linkage are investigated analytically; for several loci, they are studied

numerically.

Carlos Bustamante (Cornell University) Monday 12th July 16:30

Comparative evolutionary genomics of Humans and Chimpanzees

(Joint work with Andrew Clark and Rasmus Nielsen)

A central goal of evolutionary genetics is quantifying the extent to which natural selection has shaped the genomes of closely related species. Of particular interest has been estimating the proportion of genetic differences between species that are due to adaptive evolution vis a vis the stochastic fixation of neutral or deleterious mutations. Due to advances in sequencing technology, we now have sufficient data to scan the genomes of certain groups of organims in order to find regions that harbor evidence of past or current adaptive evolution using predictions from population genetic theory. In this talk, I will discuss the statistical methodology we have used in collaboration with Andy Clark, Rasmus Nielsen, and Celera Diagnostics to study the comparative population genomics of Humans and Chimpanzees across 20,000 loci.

Marcella Capaldo (Dept of Statistics, Oxford) Wednesday 14th July 16:00

A disintegrationtheorem for infinite variance superprocesses

We outline some work, related to the talks by Mohle and Wakolbinger, which enlightens the connection between spatial continuous-state branching processes, generalized Fleming-Viot processes and Coalescents with multiple collisions. In particular, we present a generalization of Perkins' Disintegration Theorem, which allows to express a Dawson-Watanabe superprocess as a "Skew-product" of its total mass and a classical Fleming-Viot superprocess.

Niall Cardin (University of Oxford) Thursday 8th July 15:00

Estimation of recombination rates and inference from coalescent models

Recombination is a fundamental biological process that influences both long-term evolution as well as the distribution of genetic variation in natural populations. Learning about how recombination rates vary within a genome is therefore an important task, however, direct experimental estimation of the fine-scale structure of recombination rates is impracticable for genome-wide studies. Inference from population genetic data using coalescent models provides an atteractive alternative.

The coalescent with recombination provides a model of how genealogies change as you move spatially along a sequence. Ideally, for inference about recombination rates, we would use full likelihood methods. However, performing inference under such models is very computationally intensive. There is much interest, therefore, in producing models that approximate the coalescent under which inference is more computationally efficient.

We have developed two approximations to the coalescent process. One, based on the spatial algorithm of Wiuf and Hein, disallows certain types of coalescent event, generating a simple Markovian process along the sequence. The other, based on the models of Fearnhead and Donnelly and Li and Stephens, uses an approximation to marginal likelihoods in which a dynamic-programming approach can be used to calculate the likelihood of the data. These approaches are then joined through importance sampling to generate a potentially efficient method for estimating recombination rates and, more generally, learning about changes in marginal genealogy along a sequence. The ideas are illustrated through the example of a sample of a single pair of chromosomes.

Brian Charlesworth (University of Edinburgh) Tuesday 6th July 09:00

Evolution and variation in genomic regions with low recombination rates

Close linkage between sites subject to selection, or between neutral sites and selected sites, can have significant effects on patterns of evolution and variation. Models of population genetic processes that can generate such effects are reviewed. Their properties are related to data on molecular evolution and variation in low recombination genomic regions, where linkage is expected to be tight.

Taane Clark (University of Oxford) Thursday 8th July 15:00

Bayesian logistic regression using a perfect phylogeny

(Joint work with Maria De Iorio and Robert Griffiths)

Haplotype data capture the genetic variation among individuals in a population and among populations. An understanding of this variation and the ancestral history of haplotypes is important in genetic association studies of complex disease. We introduce a method for detecting associations between disease and haplotypes in a candidate gene region or candidate block with little or no recombination. A perfect phylogeny demonstrates the evolutionary relationship between single nucleotide polymorphisms (SNPs) in the haplotype blocks. Our approach extends the logic regression technique of Ruczinski *et al.* (2003) to a Bayesian framework, and constrains the model space to that of a perfect phylogeny. Environmental factors, as well as their interactions with SNPs, may be incorporated into the regression framework. We demonstrate our method on simulated data from a coalescent model, as well as data from a candidate gene study of smoking persistence.

Nicoleen Cloete (University of Auckland) Wednesday 14th July 16:15

MCMC for a distribution over ancestral selection graphs

(Joint work with Geoff Nicholls and David Scott)

In the abscence of selection effects, the genealogy of a random sample of a population of organisms can be represented as a rooted binary tree. The stochastic development of such a genealogy is modelled using the Kingman coalescent process. This process determines a probability distribution over rooted binary trees of fixed dimension.

Neuhauser and Krone gave a stochastic model generalising the Kingman coalescent in a natural way to include the effects of selection. This model determines a distribution over a class of graphs of randomly variable vertex number. Associated with the vertices are real scalar ages which cause a realisation of the Neuhauser Krone process to have random variable dimension. Our aim is to carry out Bayesian inference for the selection parameter of the model of Neuhauser and Krone, from alleleic data, using Markov Chain Monte Carlo. In this poster I describe an algorithm for estimating the selection parameter focusing on efficiency considerations.

Graham Coop (University of Oxford) Thursday 8th July 15:00

Full likelihood inference on gene trees under models of natural selection

(Joint work with Professor RC Griffiths)

The extent to which natural selection shapes diversity within populations is a key question for population genetics. Thus, there is considerable interest in quantifying the strength of selection. A full likelihood approach for inference about the selection coefficient of a single

selected site within an otherwise neutral fully linked sequence in a coalescent setting is described in this poster. The full information is used in this approach and hence it is often preferable to summary statistics. It has the following desirable qualities. It allows for the possibility of the hypothesis of selection to be tested in a likelihood ratio setting. The likelihood surface of the selection coefficient is also obtainable allowing the MLE to be found, and the distribution of ages of the mutations and clades under mutations is also calculable. The approach is general and can be used for any biallelic selection scheme. Selection is incorporated through modelling the frequency of the allelic classes stochastically back through time and then using a subdivided coalescent recursion. An importance sampling algorithm is then used to explore over coalescent tree space consistent with the data, under the infinite sites assumption.

Kevin Dawson (Rothamsted, BBSRC) Friday 9th July 11:30

A Bayesian approach to some model-based clustering problems in population genetics

(Joint work with Khalid Belkhir)

Some of the questions most frequently posed in population genetics can be formulated as clustering or assignment problems. That is, problems of assigning individuals or observations to categories, where in general the number and precise nature of these categories is uncertain. For example, assignment of individuals in a sample to source populations, on the basis of genetic marker data. (Another example is assignment to full-sib families.) Here, the parameter of interest is the partition of the set of sampled individuals, induced by their assignment to source populations (or to families).

In our Bayesian approach to this problem, we use a Markov chain Monte Carlo method to generate a large sample from the posterior distribution of this partition. However, in general it is not feasible to evaluate the evidence supporting each possible partition of the sample. I will discuss some methods for overcoming these serious computational and visualisation problems. I will also present some of graphics generated using the software package Partition (available from: http://www.univ-montp2.fr/~genetix/partition/partition.htm).

Angeles de Cara (ICAPB, Edinburgh) Friday 9th July 16:30

Models of evolution of assortative mating

(Joint work with N. Barton and M. Kirkpatrick)

We apply the multilocus technique to models of assortative mating, namely the preference-trait model and the similarity-driven mating. We extend the formalism to allow for modifiers of strength of assortment, and study their impact on the final polymorphism of the population. We compare and contrast haploid and diploid dynamics, and explore the feasibility of treating more than 2 loci.

Maria De Iorio (Imperial College) Thursday 8th July 09:30

Importance sampling on coalescent histories

(Joint work with RC Griffiths)

Stephens and Donnelly (2000) construct an efficient sequential importance sampling proposal distribution on coalescent histories of a sample of genes for computing the likelihood of a type configuration of genes in the sample. We present a characterisation of their importance sampling proposal distribution in terms of the diffusion process generator describing the distribution of the population gene frequencies. This characterisation leads to

a new technique for constructing importance sampling algorithms in a much more general framework when the distribution of population gene frequencies follows a diffusion process, by approximating the generator of the process.

Frantz Depaulis (Ecole Normale Supérieure) Friday 9th July 10:00

Population genetics of time structured data, an application on cave bear ancient DNA (Joint work with L. Orlando and C. Hänni)

New polymorphism datasets from time structured samples have arisen thanks to recent progresses in experimental and viral molecular evolution, and the sequencing of ancient DNA. Classical population genetics analysis tools do not take time factors into account, despite their possible impacts on analyses such as neutrality and population structure tests. Here, we describe a straightforward coalescent simulation algorithm adapted to take into account time structure in a DNA sequence dataset. We show that time structure can substantially bias the analyses. Even time structure that is a small compared to the age of the most recent common ancestor (MRCA) affects the distribution of polymorphism substantially, leading to more star like trees, with an excess of rare mutations and a deficit of linkage disequilibrium structure. When part of the sample is much older than the rest, a strong departure of the tests is detected in the opposite direction. Linkage disequilibrium and, to a minor extent, frequency spectrum are the two most affected statistics, whereas the topology of the tree (how symmetrical it is) is not affected. Time structure can also lead to apparent differentiation between two samples of different ages. We present an application on the extinct cave bear, for which a large time structured ancient DNA dataset is available. Despite a limited range of absolute dates in the sample, considering age structure changed the conclusion of several tests, generally making the data more likely under the neutral model. Interestingly, we found a significant negative correlation on a local population sample between the linkage disequilibrium and the distance between mutations. This particular analysis was not affected by time structure. Given the limited length of the sequence involved we regard this more likely reflects mutational effects rather than recombination.

Anna Di Rienzo (University of Chicago) Monday 12th July 17:00

Inferences about human demography based on multilocus analyses of noncoding sequences (Joint work with R. Hudson, A. Adams, B. Voight, L. Frisse, and Y. Qian)

Demography shapes neutral variation on a genome-wide scale. In order to make inferences about human population histories, we generated full re-sequencing data on 50 independent genomic segments unlinked to coding regions and conserved noncoding regions in three population samples originating from Africa, Asia and Europe. We model simple scenarios of growth and bottlenecks to estimate the portion of the parameter space that is consistent with multiple aspects of the sequence variation data, including frequency spectrum, polymorphism levels and linkage disequilibrium.

Peter Donnelly (University of Oxford) Wednesday 7th July 12:30

The fine scale structure of recombination rate variation in the human genome

The nature and scale of recombination rate variation are largely unknown for most species. In humans, pedigree analysis has documented variation at the chromosomal level, and sperm studies have identified specific hotspots in which crossing-over events cluster. To address whether this picture is representative of the genome as a whole, we have developed

and validated a method for estimating recombination rates from patterns of genetic variation. From extensive single-nucleotide polymorphism surveys in European and African populations, we find evidence for extreme local rate variation spanning four orders in magnitude, in which 50% of all recombination events take place in less than 10% of the sequence. We demonstrate that recombination hotspots are a ubiquitous feature of the human genome, occurring on average every 200 kilobases or less, but recombination occurs preferentially outside genes

---

Alexei Drummond (University of Oxford) Thursday 8th July 17:00

Testing neutrality and modeling non-neutrality using temporally spaced sequence data

Many summary statistics have been developed to detect departures from neutral expectations. Here we apply a modification of Bayesian posterior predictive simulation to use summary statistics to test the goodness-of-fit of standard neutral models of evolution. The technique developed solves a number of problems inherent in previous tests of neutrality. Importantly, by employing a full model-based Bayesian analysis, our method can separate the effects of demography from the effects of selection. The method also allows multiple summary statistics to be used in concert. Situations in which analytical expectations and variances of summary statistics are not available are easily treated in our framework. This is especially useful for the analysis of temporally spaced data, which has great potential for testing neutrality but has not been used for this purpose largely because of limitations in available theory. We demonstrate the utility of our method on a number of datasets, and show that serially sampled datasets of RNA viruses frequently exhibit significant departures from neutrality, even after exponential growth is taken into account. We finish by suggesting a number of new Bayesian MCMC models that may be useful when selection is present, and demonstrate the effect of these new models on inference from real data sets.

Warren Ewens (University of Pennsylvania) Tuesday 13th July 12:00

Thoughts on the TDT

The transmission/disequilibrium test (TDT) has been very widely used as a test of linkage between a marker locus and a purported disease locus. Despite the extreme simplicity of the procedure, new questions continually arise about its use, and some of these will be discussed.

Greg Ewing (University of Auckland) Thursday 8th July 15:00

The structured coalescent with temporal sequence data using reversible jump MCMC

We present a Bayesian statistical inference approach for simultaneously estimating mutation rate, population sizes and migration rates in an island structured population, using temporal and spatial sequence data. We estimate migration history and rates from the DNA sequences taken at different times. We fit a model of the joint genealogy and migration process using MCMC over a space of trees labelled with migration events. Since the number and timing of events is unknown the MCMC must satisfy detailed balance between states in spaces of unequal dimension. A real HIV DNA sequence dataset with 2 demes, semen and blood, is used as an example to demonstrate the method by fitting asymmetric migration rates and different population sizes. This dataset exhibits a bimodal joint posterior distribution, with modes favouring different preferred migration directions.

Susanna Eyheramendy (University of Oxford) Thursday 8th July 15:00

Implementation of a Bayesian model in the selection of tagging SNPs

Increasing evidence shows that variation in the human genome reveals the genetic basis of many common diseases. Single nucleotide polymorphisms (SNPs) account for most of the variation (about ~ 90%) in a population. A major difficulty when trying to assess the risk of a disease based on SNPs is that the large number of SNPs, roughly 1 every 600 bases, results in unacceptably high costs for exhaustive genotyping. Therefore, researchers have put effort toward the selection of a small subset of SNPs (so-called tagging SNPs) that account for most of the variability in a genomic region. We implement, for each SNP from a set of SNPs, a Bayesian model that identifies a small subset of SNPs from such a set with which the SNP can be predicted. Based on these models, we develop an algorithm that selects a subset of the SNPs that capture most of the variation. To evaluate our method and to assess how well the selected tags are likely to perform in new haplotypes, not yet observed, we divide the haplotype samples in two sets. One set is used to select the tagging SNPs and the other set is used to measure the performance of the tagging SNPs in predicting haplotype variation. We demostrate the algorithm performance using the HapMap ENCODE regions. Our results confirm that a small proportion of SNPs is sufficient to capture the haplotypic variation in a population.

Adam Eyre-Walker (University of Sussex) Wednesday 14th July 09:30

The distribution of fitness effects of new mutations in humans and fruit flies

The distribution of fitness effects is central to understanding many problems in genetics and evolutionary biology, including such varied topics as the basis of human disease, the evolution of sex and the maintainance of genetic variation. I will describe a variety of projects in which we have attempted to estimate the distribution of fitness effects for both deleterious and advantageous mutations. Two approaches suggest the distribution of fitness effects for deleterious mutations can be described by a gamma distribution with a shape parameter of ~0.3 in humans and drosophila. We have also estimated the average strength of selection against deleterious mutations, although this is not known with any certainty. Theory suggests that the distribution of fitness effects of advantageous mutations is exponential. Using this fact, along an estimates of the proportion of rate amino acid substitutions which are adaptive (~25% in humans and flies), we estimate the average strength of selection acting upon adaptive mutations.

Daniel Falush (University of Oxford) Wednesday 14th July 10:00

Anthropological genetics, as taught by Helicobacter pylori

I will discuss the pleasures and pitfalls of making detailed demographic inferences using large DNA sequence datasets. The model will be inference of the pattern of spread of the bacterium Helicobacter pylori.

Paul Fearnhead (University of Lancaster) Monday 12th July 16:00

A review and comparison of population-based estimators of local recombination rates

There are a number of methods for estimating recombination rates from population data. These range from full-likelihood methods, which use all the information in the data but can be computationally prohibitive, to various quicker approximate likelihood methods.

Here we give an overview of these methods, compare based both on theoretical and

empirical results, and show how the methods can be used to draw important inferences about patterns of local recombination rates.

Yun Xin Fu (University of Texas at Houston) Saturday 10th July 09:00

Exact coalescent under the Wright-Fisher model

The Kingman coalescent was orignially formulated as a limiting process under the Wright-Fisher model, which is the most widely used population genetics model for reproduction. The Kingman coalescent heavily relies on the premise that population size is large and sample size is much smaller than population size. Whether sample size is too large compared to population size is rarely questioned in practice when applying statistical methods based on Kingman coalescent. In general the quality of Kingman coalescent as an approximation of the Wright-Fisher model is not well known. I will discuss the exact coalescent theory for the Wright-Fisher model and describe a simulation algorithm which is then used to study the property of the exact coalescent as well as its differences to Kingman coalescent. We show that Kingman coalescent differ from the exact coalescent by (1) shorter waiting time between successive coalescent events, (2) different probability of observing a topological relationship among sequences in a sample, and (3) slightly smaller tree length in the genealogy of a large sample. On the other hand, there is little difference in the age of the most recent common ancestor (MRCA) of the sample. The exact coalescent makes up the longer waiting time between successive coalescent events by having multiple coalescence at the same time. The most significant difference among various summary statistics of a coalescent examined is the sum of lengths of external branches, which can be more than 10% larger for exact coalescent than that for Kingman coalescent. As a whole, Kingman coalescent is a remarkably accurate approximation to the exact coalescent for sample and population sizes falling considerably outside the region that was originally anticipated.

David Goldstein (University College London) Wednesday 14th July 11:30

Haplotype mapping in pharmacogenetics

There is considerable enthusiasm for haplotype mapping in pharmacogenetics, but most studies reported to date have been haphazard and insufficient to comprehensively represent variation in the genes most likely to be relevant to drug safety and efficacy. Here I report on an analysis of patterns of genetic variation in 56 genes that metabolise or transport prescription medicines. Detailed analyses of 754 single nucleotide polymorphisms (SNPs) genotyped in two population samples (European and Japanese) provide a set of haplotype tagging SNPs that economically represent variation in most of the major enzymes that act on prescription drugs. These analyses provide a framework for systematic association studies in pharmacogenetics, and address a number of outstanding questions relating to haplotype mapping.

Following this, I provide a number of applications of haplotype mapping of variable drug response and disease predisposition, emphasizing the work that needs to be done to translate genotype-phenotype correlations into clinically useful diagnostics, and clinically useful leads concerning new therapeutic targets.

Robert Griffiths (University of Oxford) Wednesday 7th July 09:30

The genealogy of a mutation

The distribution of the frequency of a mutation arising in a population of genes can be modelled by a diffusion process on (0,1) with absorbing boundaries at 0,1 corresponding to loss or fixation of the mutant gene in the population. A classical approach to studying the genealogy of the mutation by Kimura is to take sample path averages of probabilities of events concerned with the mutation between the mutation arising and being lost or fixed in the population. Very general formulae can be derived for the age of the mutation known to be of frequency x in the population; and the frequency spectrum of the mutation, the number of mutant genes in a sample of n. A modern approach is to study the coalescent process underlying the diffusion process and then use combinatorial methods to obtain results about the genealogy. This talk is partly a review talk about the two approaches.

Josef Hofbauer (University College London) Saturday 10th July 12:00

Selection dynamics for a continuum of alleles

(Joint work with Ross Cressman)

When is a monomorphism (asymptotically) stable for the diploid selection model with a continuum of alleles?

Toby Johnson (University of Edinburgh) Monday 12th July 09:30

Multipoint linkage disequilibrium mapping using multilocus allele frequency data

I describe a likelihood based fine scale association mapping method for estimating the position of a disease predisposing gene relative to a battery of typed marker loci. The method uses multilocus allele frequency data from a sample of unrelated diseased individuals and from a sample of unrelated control individuals, that is, a case and control type design. This type of data could be obtained by typing DNA pools, which is a less expensive procedure than typing individuals separately. The method described uses a nonparametric model that makes it robust to the shape of the genealogy at the disease locus, and to the presence of phenocopies. It may be useful for mapping genes with a complex genetic basis. It can be implemented efficiently, making a multipoint analysis of a data set of a thousand markers feasible.

Paul Joyce (University of Idaho) Tuesday 6th July 12:00

Evaluating the theory of adaptive evolution: Do two Wrongs make a Wright?

Orr (2002, 2003) developed a series of general predictions for the adaptive evolution of molecular data. Since viruses can be evolved and sequenced is real time an experiment was developed to test the predictions of the Orr model. In particular, the prediction associated with the fitness rank of the next allele typically fixed by natural selection is investigated. The distribution of fitness rankings developed by Orr (2002) is not supported by the virus data. The data suggests that unequal mutation rates among transitions and transversions has a significant effect on the fitness rank distribution. Yet, mutation bias alone also does not explain the data. An adjusted version of the Orr model that accounts for mutation differences and selection differences adequately explains the data. Orr's model is based on the assumption that the probability that a mutant with selective advantage eventually fixes in the population is equal to 2s. In our experiment it is known that s is not small enough for this approximation to hold. Alternative models based on the work of Wahl 2003 that account for the effects of bottlenecks in experimental evolution are also investigated. This approach does not require that selection coefficients be small.

Despite the fact that the model assumptions are violated in (at least) two ways, the Orr model still provides a framework for testing the theory of adaptive evolution on real data.

Wilfrid Kendall (University of Warwick) Thursday 8th July 12:30

Multiresolution Ising models

(Joint work with RG Wilson)

We investigate a schematic phase transition diagram for the multiresolution Ising model. In this model nodes are connected in a regular tree-like structure, in which nodes on each layer of the tree are also connected to neighbours in a Euclidean grid. Nodes take values +/-1, and interact with spatial neighbours at one strength, and with parents and daughters at another strength. A reasonably complete schematic diagram can be derived for the case of "free boundary conditions"; work is in progress on the case of fixed boundary conditions.

The original motivation arose from work in statistical image analysis; however the methods and results may be suggestive for some problems in genealogical populations.

Steve Krone (University of Idaho) Tuesday 6th July 16:00

The role of spatial structure in Bacteriophage evolution

(Joint work with W. Wei, C. Coberly, H. Wichman)

A phage is a virus that infects (and multiplies within and lyses) bacterial cells. Phage-bacteria associations are well suited for studies in the experimental evolution of host-pathogen systems. Here, evolutionary and ecological processes occur on similar time scales and can be observed in real time. We investigate the effects of spatial structure on these evolutionary and ecological dynamics via interacting particle system models together with a coordinated series of experiments on agar plates. Comparison with the corresponding results for well-mixed liquid systems (and their ODE models) illustrates the impact of spatial structure on evolutionary processes.

Damian Labuda (University of Montreal, Sainte-Justine Hospital) Wednesday 14th July 12:00

X-linked markers tracing history of human populations

Nuclear DNA evidence complements both mtDNA and Y-chromosome diversity data in the genetic studies of human populations. A compound interrupted dinucleotide repeat from the dystrophin gene was analysed in a worldwide sample of 411 chromosomes. All of its variants fall within the formula n.n.n.n.X.m.Y, where the n's, X and Y denote the number of dinucleotides in the sequence GTnGGnGTnGGnGTX(ATGT)mGAY, while m describes the absence (0) or presence (2) of ATGT. By assuming that the internal motif ATGT appeared only once (supported by database search) and that the mutations within X occurred independently in the X.0.Y and X.2.Y lineages, we obtained the following tree that can be used to retrace human population history.The putative ancestral sequence GTXGAY (or 0.0.0.0.X.0.Y), diversified into the African-only haplogroups 0.0.0.0.X.0.Y > (4.1.0.0.X.0.Y > 4.2.0.0.X.0.Y) (5.1.0.0.X.0.Y > 5.1.3.1.X.0.Y > 5.1.3.2.X.0.Y), while that carrying ATGT motif diversified into haplogroups (((0.0.0.0.X.2.Y > ((5.1.0.0.X.2.Y > (5.2.0.0.X.2.Y) (5.1.3.1.X.2.Y) (5.1.5.1.X.2.Y))), which dispersed across all continents, including Africa. Altogether, we found 89 distinct haplotypes, with X and Y varying from 6 to 28 repeats. Sub-Saharan-Africans are the most diverse; 47% of their chromosomes are found among the exclusively African X.0.Y haplogroups representing the lineage that never left Africa. The

X.2.Y haplogroups represent the lineage that underwent range expansion within and outside of Africa. The haplogroup 5.1.0.0.X.2.Y is spread worldwide and accounts for half of all samples, 5.1.3.1.X.2.Y is preponderant in Eastern and Southern Asia, while 5.2.0.0.X.2.Y was found only among Amerindians and Europeans. In summary, we described an unusually robust nuclear marker to study human populations, which independently confirms a general out-of-Africa model known from mtDNA and Y-chromosome data; this is significant given 3 times greater historical depth of X-chromosome variability. IN addition it reveals a mosaic contribution of ancient endemic African lineages and new continentally shared African lineages to the present day Sub-Saharan populations. These results will be compared with those of an adjacent marker consisting of an 8Kb SNP-haplotype dys44 (AJHG 73:994-1015, 2003) that revealed the existence of unusually old lineages outside of Africa. In addition, we wish to discuss problems related to timing the genetic and population events, using both molecular clock (mutation and population variance in repeat lengths) as well as genetic clock based on recombinations.

Sabin Lessard (University of Montreal) Tuesday 6th July 09:30

The Ewens sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles

(Joint work with RC Griffiths)

The Ewens sampling formula, the probability distribution of a configuration of alleles in a sample of genes under the infinitely-many-alleles model of mutation, is proved by a direct combinatorial argument. The distribution is extended to a model where the population size may vary back in time. The distribution of age-ordered frequencies in the population is also derived, extending the GEM distribution of age-ordered frequencies in a constant sized population model. A connection is explored between the distribution of age-ordered frequencies and ladder indices and heights in an urn model. Finally, the genealogy of a rare allele is studied using the same combinatorial approach.

Laurence Loewe (University of Edinburgh) Saturday 10th July 11:30

Testing hypotheses about the importance of Muller's ratchet in mtDNA

Operation of Muller's ratchet has repeatedly been suspected in various biological systems and often it is implied that such systems are young or some other biological processes saved populations from extinction. Suspicions are usually based on lack of recombination, low population sizes, low fecundity and/or high mutation rates. Unfortunately, no formal model allowed precise quantification and hypothesis testing, mostly because predictions of the time between clicks of Muller's ratchet are difficult and computing intensive. Using individual-based simulation results from evolution@home, the first global computing system for evolutionary biology, I present a model that allows testing hypotheses about the evolutionary significance of Muller's ratchet. Investigations of mtDNA confirm earlier suspicions that a range of biologically realistic parameter combinations could lead to the extinction of humans and other mammalian populations over a period of 20 million years, especially for the relatively high mutation rates observed in pedigrees of human mitochondrial DNA. This complements a similar threat from mutations in nuclear DNA and suggests evaluation of unconventional explanations for the long-term persistence of endangered lines: (i) Current estimates of mutational effects may be skewed in non-trivial ways (eg. the high mutation rates observed may have deleterious selection coefficients of more than a few percent for most mutations). (ii) Rare recombination and/or other processes keep mitochondria from decaying. In any case, the results underscore the importance of

avoiding anthropogenic increase of mutation rates.

Jonathan Marchini (University of Oxford) Monday 12th July 12:00

IN SILICO genotyping

(Joint work with the Oxford Statistics HapMap Analysis Group)

For a given collection of haplotypes typed at a set of L marker loci it is interesting to ask what we might observe if we were to type an additional marker locus at a position somewhere within the current set. To address this problem we have developed a method of simulating the likely configurations of such ungenotyped (or hidden) SNPs using an approximate coalescent model. Notably the method takes advantage of recently developed methods of estimating variation in recombination rate across the region of interest and is fast enough to simulate millions of possible configurations in a matter of seconds. The "hidden SNP simulator" has a natural application in assessing the completeness of HapMap coverage: how well do the patterns of LD of SNPs currently in the map capture LD for SNPs which have not yet been genotyped. It also leads to a natural measure of information content, and hence to a natural method for comparing possible sets of tagging SNPs. We show that the method performs well in two ways (a) using simulated datasets for which the true hidden SNP distribution is known, and (b) using the Encode data by artificially hiding known SNPs and comparing the simulated distribution to the truth. In addition, we describe the application of the hidden SNP simulator to (fine) mapping disease causing variants.

Gilean McVean (University of Oxford) Wednesday 14th July 12:30

Approximate models of recombination

Although simple to describe, performing full-likelihood inference under population genetic models that include both mutation and recombination is notoriously difficult. In the face of such a challenge it is possible either to approximate the likelihood (for example using composite-likelihood), or to develop simpler models under which inference is technically feasible. I will discuss some of these 'approximate' models of recombination, including Sved's (1971) model for linkage disequilibrium, the conditional haplotype distribution of Li and Stephens (2003), and a new model motivated by the spatial coalescent process of Wiuf and Hein (1999). I will compare models in terms of the predicted distribution of linkage disequilibrium, biological realism, and computational tractability.

Loukia Meligkotsidou (University of Lancaster) Wednesday 14th July 16:30

Maximum likelihood estimation of coalescence times in genealogical trees

(Joint work with Paul Fearnhead)

For the study of a population's evolutionary history a sample of DNA sequences of individuals in the population can be used. The similarity among these sequences shows the degree to which the individuals relate to each other. The evolutionary history of the sample can be represented by a rooted genealogical tree, with the root of the tree representing the most recent common ancestor of the sample. The tips of the tree correspond to different individuals in the sample and they are labeled with the respective DNA sequences. Going up the tree, coalescences of branches represent events when two individuals of a past generation share a common ancestor in the previous generation. Interest lies in estimating the coalescence times, in the past, and in particular the time to the most recent common ancestor of the sample.

We propose a new method for maximum likelihood estimation of coalescence times in genealogical trees. This maximum likelihood approach is based on a descretization of the support of coalescence times, which helps to approximate the likelihood fuction with a fine grid of points. An appropriately modified version of the Viterbi algorithm, originally introduced for the decoding problem in discrete-time hidden Markov models, is used to maximize the likelihood in a sequential manner.

Martin Möhle (University of Tübingen) Tuesday 6th July 11:30

Coalescent processes with simultaneous multiple collisions of ancestral lineages

A class of population models with non-overlapping generations and fixed population siz N is considered. It is assumed that the family size variables within each generation are exchangeable. A weak convergence criterion is established for a properly time-scaled ancestral process as N tends to infinity. It leads to a full classification of all coalescent generators for the class of neutral population models with exchangeable reproduction. The corresponding coalescent processes allow for simultaneous multiple mergers of ancestral lines. The Kingman coalescent appears if and only if triple mergers are asymptotically negligible in comparison with binary mergers. The dual process of a coalescent with simultaneous multiple collisions is characterised via its infinitesimal generator. Subclasses of the dual process of the coalescent with multiple collisions arise also in other context, for example as time-changed ratio processes in certain continuous mass stable branching models.

Simon Myers (University of Oxford) Wednesday 7th July 17:00

Do chimps share human recombination hotspots?

(Joint with W. Winkler, D. Reich, R. Bontrop, P. Donnelly, D. Altshuler)

Claudia Neuhauser (University of Minnesota) Wednesday 7th July 12:00

Ecological and evolutionary consequences of large-scale perturbations

(Joint work with Stuart Wagenius (Chicago Botanical Garden) and Eric Lonsdorf (University of Minnesota))

We will present an example that illustrates the importance of taking both ecological and evolutionary forces into account when predicting the dynamics of systems that are far away from equilibrium. This example concerns the consequences of large scale prairie fragmentation on a common flowering prairie plant, Echinacea angustifolia, the narrow-leafed purple coneflower. E. angustifolia is a typical prairie plant: self-incompatible, long-lived, no specialized seed dispersal, and generalist pollinators. Our field and theoretical studies show that due to the mating system, habitat reduction (i) shifts the mechanism for density dependence from intraspecific competition to limitation of availability of compatible pollen, causing a reduction in overall abundance, (ii) accelerates extinction, and (iii) results in a decline in inbreeding. We discuss the consequences for management of prairie fragments and prairie reserves.

Geoff Nicholls (University of Auckland) Thursday 8th July 16:30

Specifying and fitting a Dollo point process model of binary trait character evolution

We specify a stochastic model of binary-valued trait-character evolution. Although the model is a simple and natural model for cognate data, as far as I can see, no model of this type has been considered in the past. Because of the technical difficulty of fitting models,

authors have tended to avoid ab intitio model specification and fitting. The model is based on a point process-representation of the birth and death of traits. It is essentially a stochastic variant of Dollo parsimony. We fit the model to Indo-European cognate data using Markov chain Monte Carlo. The historical processes and modern observation practices which determine the data are complex. In an exploratory analysis, we identify some features of the data which raise difficulties for quantitative analysis. In particular, words generating cognate classes in the data are those words with descendants in two or more observed languages. We show that at least some of the difficulties may be overcome.

Richard Nicholls (QMW, University of London) Tuesday 13th July 16:00

Discrepancies between real population histories and the assumptions of mathematical models: problems and some solutions.

Although it is widely acknowledged that populations might not have reached drift-migration equilibrium since they were founded, mathematical models used in genetic analysis often gloss over this issue. Estimates of the time to equilibrium from stepping-stone and island models are reassuring on this issue, suggesting that relatively stable natural populations will have persisted long enough to reach equilibrium. I shall argue that this confidence is misplaced for many analyses, including those that use genetic data to draw inferences about the species demography. It is, however, possible to make use of comparisons between loci with different mutation rate to both assess the severity of the problem and to obtain more accurate estimates of the biological parameters. I shall demonstrate an application using data from endangered island skink populations. A conceptually similar trick is to make comparisons over shorter geographical scales. More troubling are the implications for phylogentic analyses and other inter-species comparisons, where the mutation rates are lower and the geographical scales are larger.

Rasmus Nielsen (Cornell University) Monday 12th July 12:30

Detecting selective sweeps from ascertained SNP data

(Joint work with Melissa Todd, Yuseob Kim, Carlos Bustamante and Andrew G. Clark)
With the availability of large scale SNp data, the question arises how best to identify regions of the genome that have been targets of recent selective sweeps. The two main problems encountered by statistical methods for detecting selection from this data are issues relating to the robustness of the underlying demographic model and problems relating to ascertainment biases causes by the SNP discovery and selection procedure. We will consider different approaches for addressing these problems and show applications on real data.

Nick Patterson (Broad Institute) Friday 9th July 09:30

How old is the most recent ancestor of two copies of an allele?

We apply diffusion theory to questions motivated by the coalescent, and study the exact distribution of the time to the most recent common ancestor (TMRCA) of k copies of an allele with population frequency f. We generalize a number of results of Kimura. In particular we show that the expected TMRCA of 2 copies of an allele with population frequency f is just 2 N f generations, where N is the effective population size. ~

Peter Pfaffelhuber (University of Munich) Wednesday 14th July 16:45

Genealogical trees including fossils

(Joint work with Andreas Greven and Anita Winter)
The Moran model and the Wright-Fisher diffusion, its diffusion limit are two standard models of neutral evolution. However most processes derived from these models carry only a limited part of the information that can be deduced from the evolution of the population. I introduce a stochastic process that takes values in the set of compact $\R$-trees. This process also includes all lines of ascent that have died at some time in the past. These trees can be obtained by a forward evolution and also by a backwards picture.

Jitka Polechova (University of Edinburgh) Wednesday 14th July 17:00

Speciation in sympatry? Evolution of associations between loci under disruptive selection and loci affecting assortative mating.

(Joint work with Nick Barton)

Though there have been many simulation studies, analytical models of speciation in sympatry have been rather rare. Recently, conditions for speciation in a two locus diploid model, based on the model originally proposed by Udovic (1980), were analytically derived by Gavrilets (2003). We confirm these results for the haploid model, and extend the analysis for more loci affecting the trait.

The very simplest model is of assortment amongst haploid individuals based on one locus, followed by selection against heterozygotes at a second locus. This behaves in essentially the same way as the Gavrilets/Udovic diploid model, revealing the same condition for "speciation" of $a + s > 1$, given that frequency-dependent selection is strong enough (a representing the assortment, s the strength of disruptive selection). This condition applies without linkage between the loci, the evolution of association is more feasible with limited recombination. We assess robustness of this prediction if traits happen to be determined by more loci, revealing how the limiting condition for the combined strength of assortative mating and disruptive selection changes, and how more loci alter the threshold strength of negative frequency-dependent selection stabilizing the polymorphic equilibria. When more then three or four loci are involved, the analysis gets rapidly intractable. Then we approximate the trait value assuming every bialelic locus has the same effect, such as only number of a concrete allele across all loci affects the phenotype, and assess stability of this solution.

Jonathan Pritchard (University of Chicago) Friday 9th July 09:00

Coalescent approaches to association mapping

(Joint work with Sebastian Zoellner)
We discuss our work on developing coalescent-based approaches to association mapping and LD-based fine mapping. In order to tackle the problem, we separate the inference into two stages. First, we use Markov chain Monte Carlo to sample from the posterior distribution of coalescent genealogies of all the sampled chromosomes without regard to phenotype. Then, averaging across genealogies, we estimate the likelihood of the phenotype data under various models for mutation and penetrance at an unobserved disease locus. The likelihood can be used to construct significance tests or Bayesian posterior distributions for location.

Molly Przeworski (Brown University) Wednesday 7th July 16:00

Directional positive selection on standing variation

(Joint work with Jeff Wall)

Susan Ptak (Max Planck Institute for Evolutionary Anthropology) Wednesday 7th July 16:30

Absence of the TAP2 human recombination hotspot in chimpanzees

(Joint work with Amy Roeder, Matthew Stephens, Yoav Gilad, Svante Paabo, Molly Przeworski)

Recent experiments using sperm typing have demonstrated that, in several regions of the human genome, recombination does not occur uniformly but instead is concentrated in hotspots of 1-2 kb. Moreover, the crossover asymmetry observed in a subset of these has led to the suggestion that hotspots may be short-lived on an evolutionary time-scale. To test this possibility, we focused on a region known to contain a recombination hotspot in humans, TAP2, and asked whether chimpanzees, the closest living relative of humans, harbor a hotspot in a similar location. Specifically, we used a new statistical approach to estimate recombination rate variation from patterns of linkage disequilibrium in a sample of 24 western chimpanzees (Pan troglodytes verus). This method has been shown to produce reliable results on simulated data and on human data from the TAP2 region. Strikingly, however, it finds very little support for recombination rate variation at TAP2 in the western chimpanzee data. Moreover, simulations suggest that there should be stronger support if there were a hotspot similar to the one characterized in humans. Thus, it appears that the human TAP2 recombination hotspot is not shared by western chimpanzees. This finding indicates that fine-scale recombination rates can change between very closely related species. This raises the possibility that rates differ among human populations, with important implications for linkage disequilibrium based association studies.

David Reich (Harvard Department of Genetics) Tuesday 13th July 09:30

High density admixture mapping to find genes for complex disease and application to multiple sclerosis

(Joint work with N Patterson, N Hattangadi, MW Smith, SJ O'Brien, PD Jager, MJ Daly, D Altshuler, JR Oksenberg, SL Hauser and DA Hafler)

Admixture mapping is a powerful way, in principle, to carry out whole-genome scans for disease genes. The method in theory requires ~100 times fewer markers than whole-genome haplotype mapping, but should have similar statistical power to detect disease variants that differ strikingly in frequency across populations. The key idea of admixture mapping is that near a disease gene, patient populations descended from the recent mixing of two or more ethnic groups should have an increased probability of inheriting the alleles from the group with greater disease susceptibility. Since gene flow occurred recently (in African and Hispanic Americans in the past 20 generations), recombination has not had much time to act and linkage disequilibrium should extend many centimorgans.

Admixture mapping has never previously been used to identify a gene associated with a disease. One reason is that there was no high density map with large frequency differences across populations. Here we report 3 results suggesting that admixture mapping should soon be a practical method for mapping genes for complex disease in African Americans.

• We have generated a map of 2,154 SNPs with average frequency difference of 57% between Africans and Europeans. These were chosen from ~450,000 SNPs with known Africans and European frequencies, and revalidated in a new set of 378 samples to confirm

their appropriateness for admixture mapping.

• We have developed methods that allow analysis of data from such a high density map. Applying the approach to two data sets we have collected in the laboratory, we show that strong admixture linkage disequilibrium extends, on average, 17 cM in African Americans. Power calculations show that ~2,000 markers in ~2,000 patients can provide high power to detect disease loci.

• Finally, we have applied whole-genome admixture mapping for the first time to a complex human disease (multiple sclerosis). We will report preliminary results from 756 SNPs genotyped in 712 cases and controls, which allow us to scan for admixture association.

Gesine Reinert (University of Oxford) Thursday 8th July 11:30

Small worlds - statistics

(Partly joint work with Andrew Barbour)
Small world models are networks consisting of many local links and fewer long range `shortcuts'. We consider some particular instances, and rigorously investigate the distribution of their inter--point network distances as well as some measures for local clustering. The motivating applications are metabolic networks.

Noah Rosenberg (University of Southern California) Monday 12th July 11:30

Estimating transposition rates using serial isolates of Mycobacterium tuberculosis

(Joint work with Peter Small, Mark Tanaka, and Anthony Tsolaki)
Serially sampled data can be used to estimate population-genetic parameters in studies of pathogens. For serial data, we develop transposition rate estimators under several transposition models. Estimates are then computed from genotypes of the element IS6110 in M. tuberculosis, using the Akaike information criterion to compare among models. Selection against excessive numbers of copies of the element is seen to have a strong effect on copy number.

François Rousset (Montpellier) Friday 9th July 12:00

Performance of maximum likelihood estimators of dispersal and mutation rates from gene frequency data

(Joint work with M. de Iorio, R. C. Griffiths and R. Leblois)
I will present simulation studies of the performance of maximum likelihood estimators in subdivided population models, using the algorithms presented by Mario de Iorio (this seminar). A comparison with existing software will be presented for a two population model. Next I will focus on stepping stone models and consider the sensitivity of the estimators to the mutation model and to the total size of the population.

Denis Roze (GEMI IRD) Friday 9th July 16:00

Inbreeding depression and the evolution of migration rates

Different factors potentially affect the evolution of migration rates in spatially structured populations: direct selective forces such as kin competition and the cost (in survival or energy) of migrating, and indirect forces, due to selection acting at other loci than the loci determining the propensity to migrate. For example, recessive deleterious mutations may select for increased migration in order to avoid inbreeding depression, while local

adaptations should select for lower migration rates. To study the effects of indirect selection and genetic architecture on the evolution of migration, I developed a model for a population subdivided into a very large number of finite demes. Using a QLE approximation, I derived approximations for the effect of a selected locus on the change in frequency at a linked migration modifier locus.

Stanley Sawyer (Washington University in Saint Louis) Tuesday 6th July 10:00

How can one tell in which direction evolution is going?

(Based on joint work with R. Kulathinal, C. Bustamante, and D. Hartl.)
In the long run, most biologists believe that the most important changes in organisms are due to the replacements of genes by new genes that do a better job for the organism.

However, many biologists believe that in large, established populations, most evolutionary change is, in contrast, due to the replacement of genes by slightly deleterious variants. The reason for this is that most mutations are harmful rather than helpful, and that mildly harmful mutations can replace a better, established gene by the chance effects of who mates with whom and who happens to survive. This process can take a long time for a large population, but most evolutionary change takes place on a long time scale. These chance effects could not establish a severely damaged gene that was important to its host, but could replace a good gene by a gene that was only slightly worse.

In this view, most evolution in large populations is downhill. Any improvement in the population is due to either (i) very rare mutations to significantly better genes, which then spread through the population very quickly, at which point the population begins moving downwards again from a higher plateau, or (ii) the large population is replaced by the descendants of an isolated small population in which a family of favorable mutations has been established by inbreeding and chance. An argument for either (i) are (ii) is that, in the fossil record, creatures appear not to change for long periods and then suddenly a noticeably different creature appears. For shorter time periods (millions of years rather than tens or hundreds of millions of years), there is not enough fossil evidence to tell definitely whether evolutionary change is continuous or else comes in bursts.

One way to address this question is from the distribution of DNA in contemporary populations. The distribution of a set of mutations within a population is different if the mutations are advantageous, deleterious, or selectively the same as the original variants. One can also use the number of established differences between two related species to gain additional information.

To carry out an analysis, one needs a statistical model for the sample frequencies of DNA changes as a function of mutation rates and amounts of selection. One then applies the statistical model to the sequence data and estimates parameters along with measures of statistical confidence of the parameter estimates. Unfortunately, the amount of data that is needed requires the use of computationaly intensive techniques (specifically, Markov Chain Monte Carlo).

The basic data that I am using consists of samples of DNA sequences from a variety of genes in two related species. The results show that most evolutionary change in a species of the fruit fly Drosophila is advantageous, but that most evolutionary change in a common weed, Arabidopsis, is disadvantageous. The reason for the difference is that the Arabidopsis

species studied is clonal. This makes it more difficult for the organism to get rid of mildly harmful mutations.

The benefit of this analysis is that one is able get get results not only for the proportion of beneficial changes to the populations, but also for the proportion of new mutations that are beneficial among those that are not immediately lethal, as well as the proportion of mutations that are beneficial among those that become common enough to be detected in small samples. These results could be model dependent, but they form an interesting approach to an important problem.

Stephen Schaffner (Broad Institute) Tuesday 13th July 12:30

Calibrating coalescent simulations of human genome sequence variation

---

Kristan Schneider (University of Vienna) Wednesday 14th July 17:15

The role of assortative mating in shaping genetic variation under frequency-dependent selection

A model of assortative mating and frequency-dependent selection on a quantitative trait is introduced and analyzed. In the sexual haploid population only females are under sexual selection, and may pay costs for being choosy. The trait is determined by a single locus with a finite number of alleles. We compare the model for different assumptions on the costs females have to pay. Moreover, we will assume that frequency-dependent selection is due to intraspecific competition mediated by stabilizing selection. Models of such kind have been used to demonstrate the possibility of sympatric speciation. We provide necessary and sufficient conditions in terms of the strength of assortment, intraspecific competition and stabilizing selection for the maintenance of polymorphism. We also investigate the occurrence of clustering of extreme types within the population.

Per Sjodin (EBC, Uppsala University) Thursday 8th July 15:00

The meaning and existence of an effective population size

(Joint work with Ingemar Kaj, Stephen Krone, Martin Lascoux and Magnus Nordborg)

We define the "coalescent effective population size" and argue that this is the relevant effective population size concept. To illustrate the idea we use simulations to show that it is necessary that e.g. population size fluctuations appear on a different time scale than coalescent events for the coalescent effective population size to exist.

Nick Smith (Lancaster University) Saturday 10th July 10:00

Bayesian inference of genome structure and application to base composition variation

(Joint work with Paul Fearnhead)

The human genome exhibits structure at many different scales (from single base pairs to whole chromosomes) and with regard to many different features (e.g. base composition, mutation rates, recombination, gene density and expression, repeat element density). In order to understand the evolution of the genome it is necessary to develop statistical tools for analysing genome structure. Here we use a novel Bayesian approach to infer discontinuities in base composition across long contiguous sequences from mammalian genomes. This Bayesian method appears to have a number of advantages over current techniques such as window-based ANOVAs and recursive segmentation methods. In particular, the Bayesian framework allows for powerful hypothesis testing and can be

readily extended to other genomic features.

Wolfgang Stephan (University of Munich) Monday 12th July 10:00

Localizing selective sweeps along a recombining chromosome

The theory of genetic hitchhiking predicts that the level of genetic variation is greatly reduced at the site of strong selection and increases as the recombinational distance from the site of selection increases. I am discussing how this pattern can be used to detect recent directional selection on the basis of DNA polymorphism data. Furthermore, I will describe the results from our genome scan of D. melanogaster to localize genes that have been subjected to selection.

Michael Turelli (University of California, Davis) Tuesday 13th July 17:00

Effects of genetic drift on variance components under a general model of epistasis

In general, genotypes at different loci interact non-additively to produce phenotypes. A major difficulty in understanding the consequences of these ubiquitous epistatic interactions is that there has been no simple mathematical formalism to describe them. Nick Barton and I have adapted our methods for analyzing general forms of multilocus selection to describe multilocus epistasis. Our formulation leads to relatively simple and intuitive expressions for the population mean and all variance components, namely additive, dominance and all levels of epistatic interaction. To illustrate the usefulness of our parameterization, we have investigated how the population mean and each variance component are expected to change when allele frequencies change because of genetic drift associated with a temporary reduction in population size (a "population bottleneck"). Over the past 15 years, a good deal of theoretical and experimental effort has gone into trying to understand these effects. By assuming that the drift-induced changes at different loci are statistically independent, we show that epistasis (without dominance) always increases the expected additive variance after a bottleneck above the value that would be expected for an additively determined character (with neither dominance nor epistasis). More generally, we show that for haploids or diploids without dominance, the expected value of every variance component is inflated by the existence of higher-order interactions (e.g., third-order epistasis inflates the expected additive-by-additive variance as well as the additive variance). No such general statements are possible with dominance, because with dominance alone (and no epistatic interactions), the expected additive variance can be either above or below the value expected from a purely additive model. We provide both analytical and numerical support for our independent-loci assumption. This assumption allows us to produce explicit formulas for how genetic drift changes the expected value and variance of the population mean and the expected values for all variance components. These results yield general conditions under which population bottlenecks will increase the expected additive genetic variance. These conditions can be expressed solely in terms of the genetic variance components in the base population if and only if there is no dominance. I will discuss the biological relevance of these calculations and suggest other possible applications of our description of epistasis.

Benjamin Voight (University of Chicago) Wednesday 14th July 17:30

Confounding from cryptic relatedness in association studies

(Joint work with Jonathan K. Pritchard)

In disease-association studies, it is well known that false positives can result from population structure. However, it is also true that kinship shared among cases and controls

increases the false positive rate (i.e. causes confounding). Although methods exist to address confounding due to relatedness if the genealogy is known, a less often recognized problem occurs if these relationships are cryptic, or unknown (Devlin and Roeder, 1999). For instance, because cases share a common trait which has a (partial) genetic basis, they are more likely to be related to one another than to random controls, resulting in extra variance of the usual test for association. Until now there has been little work to assess under what scenarios this type of confounding is likely to lead to inflated rates of false positives in practice. We take a population genetics based approach and model the relationship between the inflation factor due to confounding, δ, and the relatedness within a case-control sample. We derive novel equations that relate the inflation factor to observable genetic parameters: the relative recurrence risk ratio, λ_r, the population size from which the sample is drawn, and the sample size of the study. Analytic results show that, for outbred, randomly mating populations, δ is expected to be negligible. We apply our results and estimate the inflation factor for six phenotypes measured in the Hutterites, an inbred founder population with over 13,000 members about whom the genealogy is known. Our analytic results agree with empirical estimates of δ, which confirms the validity of our approach.

Arndt von Haeseler (Dűsseldorf) Thursday 8th July 09:00

IQPNNI: Moving fast through tree space and stopping in time

(Joint work with Le Sy Vinh)

We introduce a new tree reconstruction method that allows the reconstruction of a maximum likelihood phylogenetic tree for 1,000 sequences or more in acceptable time. By applying different approaches to change the branching pattern we evaluate the likelihood landscape and decrease the probability to get stuck in a local optimum. Finally, we use extreme value theory to determine when to stop the current search for a better tree.

John Wakeley (Harvard University) Tuesday 13th July 10:00

Natural selection and genetic drift in a subdivided population

(Joint work with Tsuyoshi Takahashi and Paul Slade)

The many-demes limit for selection and drift in an island model of population subdivision will be presented. The results relate the dynamics in a subdivided population to those in a panmictic population via a separation of timescales. Both forward-time dynamics of allele frequencies and backward-time dynamics of ancestral graphs will be discussed. Interesting biological conclusions may be drawn even from this very simple model.

Anton Wakolbinger (University of Frankfurt) Wednesday 7th July 10:00

Continuous-mass stable branching and its resampling counterpart

(Joint work with Matthias Birkner and Martin Moehle)

Stochastic models for the temporal evolution of a population consisting of two (or several) neutral types usually come in one of two flavours: One can either think of the superposition of independent branching processes, with a random total population size. Alternatively, one can fix the population size, impose a resampling mechanism and decree that the type is inherited from the designated "parent". In some cases there are well known results between the two classes of models: The ratio of one of two independent Feller branching diffusions

to the sum of the two is, after a time change depending only on the total mass process, a Wright-Fisher diffusion. We investigate analogous relations for stable continuous mass branching processes, and characterize the corresponding "ratio processes" as duals to a (subclass of) general coalescent processes with multiple mergers.

Jeff Wall (University of Southern California) Tuesday 13th July 11:30

Estimating recombination rates using three site likelihoods

We introduce a new method for jointly estimating crossing over and gene conversion rates using sequence polymorphism data. The method calculates probabilities for subsets of the data consisting of three segregating sites, then forms a composite likelihood by multiplying together the probabilities of many subsets. Simulations show that this new method performs better than previously proposed methods for estimating gene conversion rates, but that all methods require large amounts of data to provide reliable estimates.

Bruce Weir (North Carolina) Tuesday 13th July 09:00

The genetic architecture of human chromosome 20

(Joint work with Lon Cardon and Bill Hill)

There are growing numbers of very dense SNP datasets for human chromosomes, one of which is for chromosome 20 (Ke et al, Human Mol Genet 13:577-588, 2004). These datasets are allowing a new examination of patterns of genotypes and association throughout the genome, and they suggest care in using patterns of association to map disease genes. As a simple example, the practise of using departures from Hardy-Weinberg to point to genotyping errors ignores the correlation of Hardy-Weinberg test statistics due to linkage disequilibrium and the subsequent clustering of the markers showing such departures. Another advantage of dense datasets is that empirical distributions of the population-structure parameter F_{ST} can be obtained, and the predicted long upper tail of the distribution of estimates over loci can be confirmed. The data sets are also allowing statements to be made about higher-order measures of allelic association.

David Welch (University Of Auckland) Thursday 8th July 15:00

Reconstucting host genealogy and parasite history - inference using an integrated stochastic model

(Joint work with GKN Nicholls, AG Rodrigo, W Solomon)

In many situations, the genealogy of a parasite closely follows the genealogy of its host species. If the parasite is passed only vertically (from parent to child at birth), the correspondence is perfect. If it is passed only horizontally (between any two members of a population at any time), the two histories diverge immediately. In this poster, I present a new stochastic model of a parasite that is transmitteed at birth and via contact. The model gives rise to a coalescing and branching graph structure similar Krone and Neuhauser's Ancestral Selection Graph. I discuss how the graph, which represents the coupled host-virus genealogy, can be reconstructed using a Bayesian approach and Markov Chain Monte Carlo techniques. The problem is motivated with an example of a cat population hosting the Feline Immunodeficiency Virus.

John Whittaker (Imperial College) Tuesday 6th July 16:30

Predicting the functional consequences of amino acid polymorphisms using hierarchical

Bayesian models

(Joint work with Claudio Verzilli, Nigel Stallard and Daniel Chasman)

Genetic polymorphisms that lie in the coding regions of DNA may havephenotypic effects, for example by influencing disease susceptibility. Detection of deleterious mutations is hampered by the large number of candidate sites present; therefore methods to prioritise the most promising sites are needed. To this end, a possible approach is to use structural and sequence-based information of the encoded protein to predict whether a mutation at a particular site is likely to disrupt the functionality of the protein itself. We propose a hierarchical Bayesian Multivariate Adaptive Regression Spline (BMARS) model for supervised learning in this context and assess its predictive performance using data from mutagenesis experiments on Lac repressor and Lysozyme proteins.

Hilde Wilkinson-Herbots (University College London) Tuesday 13th July 16:30

Fst and the "effective level of gene flow" in models of subdivided populations.

(Joint work with R. Ettridge)

Using the structured coalescent model, values of Fst have been calculated for various models of subdivided populations. As Fst is still commonly used by population geneticists to infer the "effective level of gene flow", the relationship between the effective level of gene flow and the actual migration rate is also of interest. Our focus will be on the effect that unequal migration rates can have on Fst, on the strength of its dependence on the mutation rate, and on the effective level of gene flow. We will also illustrate how the ancestral population size affects the value of Fst in a model of a population split.

Scott Williamson (Cornell University) Saturday 10th July 09:30

Non-stationary population genetic models with selection: theory and inference

(Joint work with Carlos Bustamante)

Non-stationary demographic processes, such as recent population growth, can have a large impact on patterns of polymorphism and divergence in natural populations. This effect is particularly relevant to statistical methods which aim to detect and quantify natural selection from sequence data because selection and population growth can have similar effects on patterns in the data. Disentangling the effects of selection and demography is therefore a major challenge for population genetics. Here we address this problem by developing population genetic models that incorporate both natural selection and non-stationary demographic processes, arriving at predictions for overall levels of polymorphism, the frequency spectrum of polymorphic sites, and divergence between species. Further, we apply a Poisson Random Field framework for statistical inference using the frequency spectrum. If data are available from putatively neutral genomic regions (e.g. non-coding regions or pseudogenes), then this framework can be used to correct for the effect of demography on the frequency spectrum. Monte Carlo methods are used to address how robust this method is to different types of demographic forces. We demonstrate the application of this method using a large data set of human Single Nucleotide Polymorphisms.

Daniel Wilson (University of Oxford) Thursday 8th July 15:00

Diversifying selection and functional constraint: estimating the dN/dS ratio for gene sequences in the presence of recombination

(Joint work with Gil McVean)

A popular way to model natural selection in gene sequences is by way of the dN/dS ratio, which is the relative rate of non-synonymous to synonymous mutations in the evolutionary history of a sample of sequences.

Traditionally the dN/dS ratio has been used as a summary statistic in phylogenetics, but the mutation models of Goldman and Yang (1994) and Nielsen and Yang (1998) promoted dN/dS to the status of a parameter, which they called omega.

Treating dN/dS as a mutation parameter is a useful way to model natural selection as a form of mutational bias. Values of omega less than one are interpreted as functional constraint, values greater than one can be interpreted as diversifying selection, and omega equal to one corresponds to selective neutrality. Estimating the value of omega for a sample of gene sequences allows inference to be performed on the mode of selection acting on a region, or particular sites within that region.

Maximum likelihood methods that estimate omega are in widespread use, and have been applied to many organisms. However, their common assumption of complete linkage along the sequence has been shown to lead to false positives when that assumption is violated by the presence of recombination.

We present work-in-progress on a new method that aims, not only to estimate omega in the presence of recombination, but to co-estimate the recombination rate itself. We demonstrate its application in estimating a constant omega along a sequence, and in estimating site-wise omega's, and discuss briefly the direction we hope to take this research.

Ian Wilson (University of Aberdeen) Friday 9th July 12:30

Genealogies and population structure.

This talk details ongoing work on coalescent based inference using the software BATWING) to look at the association between genealogy and phenotype. This is done treating phenotpye as analogous to location in structured populations.

Carsten Wiuf (University of Aarhus) Tuesday 6th July 17:00

An algorithm for dividing genotypes into blocks: An extension of Hudson and Kaplan's R_M to genotypic data.

Hudson and Kaplan (1985) introduced the four-gamete test to test for the presence of recombination in a sample of DNA sequences. Further, they came up with a lower bound R_M to the actual number of recombination events experienced in the sample's history. R_M+1 can be interpreted as the minimum number of topologies required to explain the sequences. They assumed an infinite-site model as the data generating process.

Today many data sets spanning large chromosomal regions are generated. Because of limited resources, these data sets often consist of unphased genotypes, rather than haplotypes. That is, if an individual is heterozygote for a SNP with alleles 0 and 1 it is not known which of an individuals two chromosomes harbor the 0 allele, respectively the 1 allele.

In this talk, I discuss a test similar to the four-gamete test and introduce a lower bound R_M^g to the number of recombination events experienced in the sample's history. R_M^{g+1} can be interpreted as a lower bound to the number of topologies required to explain the sample. In addition, various results, theoretical and simulated, about incompatibilities in a sample of genotypes are presented. If time allows, an analysis of a real data set is shown, using the above measures.

Ziheng Yang (University College London) Thursday 8th July 16:00

Comparison of synonymous and nonsynonymous rates to detect selection in protein-coding genes

The difference in synonymous and nonsynonymous substitution rates (dS and dN) provides a measure of selective pressure at the protein level. A dN/dS ratio greater than one means that nonsynonymous mutations offer a fitness advantage and are fixed in the population at a higher rate than synonymous mutations. In this talk I will discuss codon-based substitution models designed to detect positive selection affecting a few amino acids in a protein.

Xu-Sheng Zhang (University of Edinburgh) Friday 9th July 17:00

The frequency distribution of genes affecting quantitative traits in populations under natural selection

(Joint work with Bill Hill)

The distributions of frequencies and effects of segregating genes affecting quantitative traits are important in many evolutionary and genetic processes within populations. These are hard to obtain although information is becoming available with QTL mapping. These depend on many factors, including the mutation rate, the distribution of the effects of mutants, and the population size. Mutant genes were assumed to have an effect on the trait and fitness and thus were under joint stabilizing and pleiotropic selection with mutation-selection balance in a finite population. Linkage and epistasis were ignored. The frequencies of segregating genes is expected to become very heavily weighted to extremely low values when the genes have a large effect on either the trait or fitness. Consequently, much of the variation in the trait is expected to be contributed by genes which are nearly neutral for fitness and with intermediate or small effect on the trait.