Mind the Gap 2010: Joining Theoretical and Empirical Population Genetics

Max-Planck-Institute for Evolutionary Biology, Plön
September 23 – September 24, 2010

Funded by the

VolkswagenStiftung

Contents

Program

Thursday, September 23, 2010

9:00		*Welcome*
9:05	John Pannell	*Possible links between sexual-system evolution and demographic processes in plants and animals*
9:45	Thorsten Reusch	*Do plants evolve differently?*
10:25	Stefan Laurent	*Aboveground plant populations are just the tip of the iceberg: seed banks and metapopulation in wild tomato species*
10:40	**Coffee break**	
11:15	Laura Rose	*Evolution of disease resistance in wild tomatoes*
11:55	Marcus Koch	*An* Arabidopsis *hybrid zone—genetic differentiation along a gradient reflects incomplete genetic amalgamation*
12:35	**Lunch**	
14:30	Alison Etheridge	*Modelling evolution in a spatial continuum*
15:10	Peter Pfaffelhuber	Discussion on *New models in population genetics*
16:00	**Coffee break**	
16:30	Torsten Günther	*Improved haplotype-based detection of selective sweeps in samples of unequally related individuals*
16:45	Cornelia Borck	*Linkage disequilibrium under soft selective sweeps*
17:00	Tanja Stadler	*Inferring the epidemic behavior of viruses from sequence data*
17:40		End of talks
19:00		**Conference dinner at Hotel Nordic**

Friday, September 24, 2010

9:00	Arne Traulsen	*Games & genes*
9:40	Lorens Imhof	*Phenotype switching and mutations in random environments*
10:20	**Coffee break**	
10:50	Duncan Greig	*Mysteries of the yeast life cycle*
11:30	Martin Kapun	*Experimental evolution of temperature adaptation in* Drosophila simulans
11:45	Florian Clemente	*Base composition evolution in putative neutrally evolving sequences in the* Drosophila melanogaster *subgroup*
12:00		**Lunch**
14:00	Bernhard Haubold	Discussion on *Indexes provide solutions to many computational problems in comparative genomics*
14:50	Asger Hobolth	*Analysis of whole genome population genetic data*
15:30		**Coffee break**
16:00	Dirk Metzler	*Jaatha: A fast composite likelihood approach to estimate demographic parameters*
16:40	Thomas Wiehe	*Measuring tree shape and using it as an evolutionary signature*
17:20		Closing remarks
17:25		**End of meeting**

Abstracts

Cornelia Borck: *Linkage disequilibrium under soft selective sweeps*

Hermisson and Pennings have shown that a selective sweep is likely to be founded not by a single but by several individuals, if the mutation rate to a selectively beneficial allele is sufficiently high. Such an event is called a soft sweep and the complementary event (the classical case) a hard sweep. I will show that the linkage disequilibrium pattern of a soft sweep differs substantially from that of a hard sweep due to haplotype structure.

Florian Clemente: *Base composition evolution in putative neutrally evolving sequences in the* Drosophila melanogaster *subgroup*

The pattern of base composition is determined by four population genetic forces: mutation, biased gene conversion (BGC), selection, and genetic drift. The relative strength of these forces on putatively neutral evolving sequences has often been discussed but remains ambiguous. Here we consider both polymorphism (Shapiro *et al*., 2007) and divergence data (Singh *et. al*., 2009) to analyze these forces in *D. melanogaster* and its close relatives. Alignments with the outgroups *D. simulans*, *D. sechellia*, *D. erecta* and *D. yakuba* allow us not only to polarize polymorphism but also to classify the age of mutations. In order to gain the relative influence of base composition evolution, we compared fourfold degenerate sites to short introns. We find that AT/GC polymorphism is skewed towards an excess of AT low-frequency variants, indicating that GC is selectively favored. This asymmetry is weak in introns and pronounced at fourfold degenerate sites. While substitutions in introns are close to mutation-drift equilibrium, fourfold degenerate sites show an excess of GC→AT substitutions, suggesting non-equilibrium. The age classication of mutations reveals the existence of old polymorphism. On average, mutations in introns are older than mutations at fourfold degenerate sites. Overall, introns and fourfold degenerate sites show distinct patterns, suggesting the action of independent forces in base composition evolution. Therefore, common mechanisms, e.g., a shift in mutation bias or biased gene conversion as hypothesized in previous studies, seem relegated to a minor role. Furthermore, the presence of old polymorphism, especially in introns, provides evidence against a bottleneck in the *D. melanogaster* subgroup. The pattern in introns is close to that expected under mutation-drift equilibrium. The slight asymmetry of the AT/GC polymorphism may be explained by weak selection in favor of GC or a recent change in the mutation bias towards AT within *D. melanogaster*. On the other hand, the pattern at fourfold degenerate sites requires a complex form of selection: rather than a relaxation of selection due to a bottleneck, one or more shifts in codon usage bias seem likely. Moreover, our results show clearly that fourfold degenerate sites are inappropriate as a neutral reference to infer demography and selection. Short introns are closer to neutrality but also have to be considered with caution.

Alison Etheridge: *Modelling evolution in a spatial continuum*

Kingman's coalescent has been outstandingly successful as a tool in mathematical and statistical genetics. However, although it only applies to a very idealised biological situation and although it can be readily modified to incorporate some more realistic biological features, a satisfactory approach to modelling populations evolving in a spatial continuum has proved elusive, due in no small part to Felsenstein's 'pain in the torus'. On recent work with Nick Barton, IST Austria, we introduced a new framework for modelling the evolution of a population evolving in a spatial continuum which we will describe and, as time permits, explore in this talk.

Duncan Greig: *Mysteries of the yeast life cycle*

The yeast *Saccharomyces cerevisiae* has been used for thousands of years to make wine. In recent times, it has also become a favourite model organism for biology, in part because certain features of its life cycle make it particularly amenable to genetic and biochemical analysis. However, we know very little of how yeast lives outside the lab, and we do not understand why the different mechanisms and features of the life cycle evolved, or how they function in nature. I will review the yeast life cycle, point out some of major unanswered questions, and present some possible answers that require further analysis.

Torsten Günther: *Improved haplotype-based detection of selective sweeps in samples of unequally related individuals*

The increasing amount of genome information allows to address various questions regarding the molecular evolution and population genetics of different species. Such genome-wide datasets including thousands of individuals genotyped at hundreds of thousands of markers require time-efficient and powerful analysis methods. The presence of population structure introduces a bias into present population genetic tests of natural selection, which may confound results. Thus, modification of test statistics is necessary to introduce time-efficient and unbiased analysis methods. We present an improved haplotype-based test of selective sweeps in samples of unequally related individuals. For this purpose, we modified existing tests by weighting the contribution of each individual based on its uniqueness in the entire sample. In contrast to previous tests, this modified test is feasible even for large present and future datasets. The analysis of empirical data from humans and *Arabidopsis thaliana* reveals different results compared to previous tests. Additionally, we utilize forward-in-time simulations to estimate the sensitivity of such haplotype-based test statistics to complex demographic scenarios, such as population structure, population growth and domestication. Overall, the modified test leads to a slight increase of power to detect selective sweeps among all demographic scenarios.

Bernhard Haubold: Discussion on *Indexes provide solutions to many computational problems in comparative genomics*

Many computational challenges in comparative genomics are due to changes in scale: solutions that work on the scale of genes fail on the scale of genomes. In this discussion round I point out that there is a universal method to overcome such problems: indexing.

Indexing has a long tradition in book making and computer science. When constructing a book index, the aim is to make the text accessible via an alphabetically sorted list of key words and their locations. Similarly, a digital index allows access to a potentially large body of data by traversing a sorted representation of the data rather than the original data itself.

Comparative genomics data is routinely stored in tables that are administered using relational database management systems such as MySQL. Indexing large tables can speed up data access by a factor of 1000 and more. But perhaps the best known application of computerized indexing is built into sequence comparison programs like BLAST. In this software a short query sequence is indexed to speed up the search for homologs in a long subject sequence. Not surprisingly, the converse, that is, to index the subject, leads to much greater increases in speed and underlies programs such as BLAT.

The data structures used to index genome-scale sequences have been the focus of intensive research in computer science over the past 15 years. Suffix trees, first made computationally tractable in the 1970's, have been the starting point of much of this work. They are perfect indexes in the sense that they track the positions of all possible "words" in a text, rather than the finite list of key words at the back of useful books. The disadvantage of suffix trees is their large memory requirement, which has lead to the development of two memory-efficient implementations: (i) Enhanced suffix arrays, invented a decade ago, which underlie popular programs such as the genome aligner `MUMmer`; and (ii) Ferragina-Manzini (FM) indexes, which are used in programs like the short read mapper `bowtie`.

So, next time you are confronted with a computational difficulties in comparative genomics, ask yourself: is there an index structure that might help overcome the impasse?

Asger Hobolth: *Analysis of whole genome population genetic data*

The DNA sequence of a whole genome can now be obtained at a relatively low cost, giving rise to population genetic data sets with few individuals and many loci. In this talk I will describe various ways of summarizing these data sets and discuss corresponding analysis tools. The main focus will be on work of my own and collaborators, but I will also mention work by others.

Lorens Imhof & Drew Fudenberg: *Phenotype switching and mutations in random environments*

Cell populations can benefit from changing phenotype when the environment changes. One mechanism for generating these changes is stochastic phenotype switching, whereby cells switch stochastically from one phenotype to another according to genetically determined rates, irrespective of the current environment, with the matching of phenotype to environment then determined by selective pressure. This mechanism has been observed in numerous contexts, but identifying the precise connection between switching rates and environmental changes remains an open problem. Here we introduce a simple model to study the evolution of phenotype switching in a finite population subject to random environmental shocks. We compare the successes of competing genotypes with different switching rates and obtain a complete characterization of how the optimal switching rates depend on the frequency of environmental changes in a symmetric setting. Our results explain why the optimum is relatively insensitive to fitness in each environment.

Martin Kapun, Viola Nolte, Robert Kofler, Pablo Orozco, Thomas Flatt, & Christian Schlötterer: *Experimental evolution of temperature adaptation in* Drosophila simulans

Drosophila simulans originated in Madagascar, colonized Africa and recently spread around the world. The colonization of new habitats and climate zones might thus have involved multiple adaptations to new environmental conditions. Various genome scans aiming to identify the underlying adaptations faced the problem to disentangle demography and selection. Here we use a complementary approach to understand thermal adaptation in *D. simulans* by experimental evolution selecting to segregating variation in natural outbred population from Northern Portugal.

A starter-population of *D. simulans* was split into two subsets of five replicate populations each, which are maintained at two different temperature extremes (10°C and 28°C, respectively). Using pooled samples from individuals from different generations, we trace adaptation on the genomic scale by second generation sequencing (Illumina GA IIx).

Marcus A. Koch & Roswitha Schmickl: *An* Arabidopsis *hybrid zone—genetic differentiation along a gradient reflects incomplete genetic amalgamation*

We describe a periglacial contact and suture zone of two *Arabidopsis* species: *Arabidopsis lyrata* and *Arabdiopsis arenosa*. Using chloroplast DNA sequence variation, microsatellite analysis, cytological data and morphometric data we elaborate a phylogenetic-evolutionary scenario of multiple introgression from *A. arenosa* into *A. lyrata*, polyploidization, migration and secondary contact of the two species. The spatial dimensions are approximately 120 x 30 kilometers spanning a region from the Eastern Austrian Forealps to the Danube river in the Wachau region and the adjacent Bohemian massif in the North. Along this geographical line we found considerable clinal genetic variation indicating either incomplete genetic amalgamation due to continuous and eventually reduced interpopulational geneflow or selection differs along this gradient since environmental conditions changes drastically from the higher-elevation limestone dominated forealps to lowland areas in the Wachau region dominated by silicious bedrocks. We will introduce various problems of the datasets such as differing mutational rates and changes in ploidy level.

Stefan Laurent: *Aboveground plant populations are just the tip of the iceberg: seed banks and metapopulation in wild tomato species*

Wild tomato species, which originated in western South America and the Galapagos Islands, are found in a wide range of habitats, and have thus to cope with various abiotic (e.g. temperature fluctuations, drought) and biotic stresses (e.g. attack by pathogens and herbivores). In order to study molecular signatures of adaptation, we develop here first a demographic model taking into account the life history traits of such species, namely seed bank, spatial structure of populations and range expansion. Our aim is to explain the discrepancies between the high effective population size inferred from genetic data and the very small census population sizes observed in nature. Metapopulation structure with restricted migration among demes and seed banks are two well known mechanisms that increase effective population size. Here we estimate parameters of seed banks (germination rate and maximum life expectancy of seeds) and of metapopulation (migration rates) using an Approximate Bayesian Computation framework.

Dirk Metzler: *Jaatha: A fast composite likelihood approach to estimate demographic parameters*

Given population genetic data from related populations or species we aim to estimate parameters like population split times, population growth rates and migration rates. Maximum-Likelihood and Bayesian methods based on importance sampling and MCMC are expected to give the best results but often need several months of computer run-time for just one dataset. Moreover, some dataset require models which are not incorporated in the available implementations of these methods, and the development of such a software package may also take years due to the complex data structures involved. We discuss alternative heuristics that are fast and easy to implement.

John Pannell: *Possible links between sexual-system evolution and demographic processes in plants and animals*

The remarkable diversity of plant sexual systems points to repeated evolutionary transitions between contrasting strategies. One frequent such shift is the evolution of self-fertilisation. Another is the transition between hermaphroditism and dioecy (the possession of separate sexes). In my presentation, I will explore the implications of demographic fluctuations, operating at a number of spatial scales, for transitions from dioecy to self-compatible hermaphroditism. Although I will focus particularly on a series of detailed studies of an uncharismatic European herb, my main aim will be to illustrate the striking parallels displayed by plants and animals that have undergone the same evolutionary transitions for apparently similar reasons—selection for reproductive assurance when mating partners are scarce. For subsequent discussion, I will ask to what extent plant and animal systems in which shifts in both sexual systems and population structure may have coincided provide fertile material for population-genetic analysis of signatures of demographic processes.

Peter Pfaffelhuber: Discussion on *New models in population genetics*

Population genetics has inspired an impressive body of research in mathematics over the past decade. However, only if empiricists are aware of the structure of the models coming out of this research can the new ideas find their way into mainstream biology. Here we will discuss two recent remarkable models that are concrete enough for transfer to empirical studies:

1. Usually, populations under natural selection are studied using a Wright-Fisher model in which the number of offspring per individual depends on the parental genotype. Brunet and Derrida have recently introduced a contrasting model in which each of N individuals in a population produces the same number of gametes. These inherit their fitness from their parents and the fittest N gametes survive to reproduce.

2. While real populations live in continuous space, population geneticists usually model populations as occupying discrete islands. The new spatial model by Barton and Etheridge overcomes this gap between model and

reality by placing populations in two-dimensional continuous space. Every individual can leave offspring in a circular area centered on the parent's location.

Both models are mathematically tractable and ready to be applied to population genetic measurements.

Thorsten Reusch: *Do plants evolve differently?*

Somatic mutations are an underappreciated source of genetic variation within multi-cellular organisms. The resulting genetic mosaicism should be particularly abundant in large clones of vegetatively propagating plants. Little is known on the abundance and ecological correlates of genetic mosaicism in field populations, despite its potential evolutionary significance. Because sexual reproduction restores genetic homogeneity, the prevalence of genetic mosaicism should increase with increasing clonality. This was tested in populations of the ecologically important marine angiosperm Zostera marina, ranging from Portugal to Finland. Genetic mosaics were detectable as complex microsatellite genotypes at two hypervariable loci that revealed additional mosaic alleles, suggesting the presence of multiple divergent meristematic cell lineages within the same plant module. As predicted, the proportion of mosaic genotypes was negatively correlated with clonal richness, and thus, sexual reproduction. The neutral mutations observed at microsatellite markers suggest the possibility of degradation of traits involved in sexual reproduction no longer selected for. On the other hand, the identified genetic variation may also be correlated with adaptive genetic variation within clones, compensating for the lack of meiotic mutations in the near-absence of sexual reproduction.

Laura Rose: *Evolution of disease resistance in wild tomatoes*

I will describe our investigations of the evolutionary history of five genes in a defense signaling pathway in wild tomatoes. Pathway theory predicts that genes that function downstream in the pathway, which serve as convergence points for upstream signals, should show greater evolutionary constraint. We find that two of the upstream genes evolve under strong evolutionary constraint, while the other genes, which operate further downstream in the pathway, show evidence of balancing selection. This counterintuitive observation may be likely in pathways involved in pathogen defense. Pathogens may specifically target downstream positions in resistance pathways to manipulate or nullify host resistance. However, plants also express pathogen specific receptors that function upstream in resistance pathways and activate the resistance responses upon pathogen detection. Therefore, it is likely that genes throughout defense pathways serve as targets for coevolution between hosts and pathogens.

Tanja Stadler: *Inferring the epidemic behavior of viruses from sequence data*

I present a method which allows to infer directly epidemiological and evolutionary parameters from virus sequences using a Bayesian method. The epidemiological model has a parameter for transmission and becoming-non-infectious and therefore explicitly describes the epidemiological process, in contrast to the previously used coalescent which merely captures the changes in effective population size. The model is implemented as a prior distribution in the Beast software package. The epidemiological model can be used together with any of the evolutionary models.

Using our method, we analyzed several HIV-1 subtype B sub-epidemics in Switzerland. In particular, we calculated the basic reproductive number R_0, which determines the spread of an epidemic, to be between 1.05 and 4.62.

The advantage of the presented R_0 estimation is that the method only relies on sequence data. Existing R_0 estimation methods are based on the initial population growth or on the coalescent and rely on a good estimate of the length of infection time, which is problematic as this time span is very variable for HIV.

Arne Traulsen: *Games & genes*

In the past years, evolutionary game theory has benefited a lot from adopting concepts and ideas from population genetics. The focus shifted from infinite populations to stochastic effects in finite population games. Now, quan-

tities such as fixation probabilites are also routinely addressed in evolutionary game theory. Evolutionary games are often viewed as a phenotype based approach to evolutionary change, since they do not take the intricacies of genetic architecture into account. Typically, evolutionary game theory considers pairwise interactions. While most researchers have interacting organisms in mind, e.g. in the context of behavioral ecology or social interactions, it could also be the two alleles at one locus in a diploid genome. Based on this perspective, many results from evolutionary game theory can be transferred to population genetics (and vice versa). Several examples will be presented where evolutionary game theory and population genetics can constructively interact with each other, which can provide new perspectives in both fields.

Thomas Wiehe: *Measuring tree shape and using it as an evolutionary signature*

Although tree shape measures have attracted the attention of theoreticians at least since the 1980s, they have so far found relatively little use in practical methods of phylogenetics or population genetics. In this talk, I will give a short review of tree shape measures, discuss weaknesses and strengths, and show an application in population genetics.

Participants

#	Name	Email
1	Altrock, Philipp	altrock@evolbio.mpg.de
2	Baake, Ellen	ebaake@techfak.uni-bielefeld.de
3	Bataillion, Thomas	tbata@daimi.au.dk
4	Birkner, Matthias	birkner@mathematik.uni-mainz.de
5	Blath, Jochen	blath@mail.math.tu-berlin.de
6	Borck, Cornelia	cornelia.borck@stochastik.uni-freiburg.de
7	Caliebe, Amke	caliebe@medinfo.uni-kiel.de
8	Chain, Freddy	chain@evolbio.mpg.de
9	Chen, Wei	wchen@zoologie.uni-kiel.de
10	Clemente, Florian	florian.clemente@vetmeduni.ac.at
11	Depperschnidt, Andrej	depperschmidt@stochastik.uni-freiburg.de
12	Eldon, Bjarki	eldon@stats.ox.ac.uk
13	Erin, Noemie	erin@evolbio.mpg.de
14	Etheridge, Alison	etheridg@stats.ox.ac.uk
15	Ewing, Greg	gregory.ewing@univie.ac.at
16	Gokale, Chaitanya	gokhale@evolbio.mpg.de
17	Greig, Duncan	d.greig@evolbio.mpg.de
18	Günther, Torsten	torsten.guenther@uni-hohenheim.de
19	Hellmann, Ines	ines.hellmann@univie.ac.at
20	Hobolth, Asger	asger@birc.au.dk
21	Huang, Weini	huang@evolbio.mpg.de
22	Hustedt, Thiemo	thustedt@gmx.de
23	Hutter, Stefan	hutter@zi.biologie.uni-muenchen.de
24	Hutzenthaler, Martin	hutzenthaler@biologie.uni-muenchen.de
25	Imhof, Lorens	limhof@uni-bonn.de
26	Kapun, Martin	capoony@gmail.com
27	Klassmann, Alexander	alex.klassmann@koeln.de
28	Kluth, Sandra	sandra.kluth@gmx.de
29	Koch, Marcus	marcus.koch@urz.uni-heidelberg.de
30	Kosiol, Carolin	ckosiol@gmail.com
31	Krug, Joachim	krug@thp.uni-koeln.de
32	Laurent, Stefan	laurent@zi.biologie.uni-muenchen.de
33	Metzler, Dirk	metzler@biologie.uni-muenchen.de
34	Millinski, Manfred	millinski@evolbio.mpg.de
35	Naduvilezhath, Lisha	Lisha@biologie.uni-muenchen.de
36	Nothnagel, Michael	nothnagel@medinfo.uni-kiel.de
37	Pannell, John	john.pannell@plants.ox.ac.uk
38	Papkou, Andrei	apapkou@zoologie.uni-kiel.de
39	Pfaffelhuber, Peter	peter.pfaffelhuber@stochastik.uni-freiburg.de
40	Reed, Floyd	reed@evolbio.mpg.de

continued on next page

continued from previous page

#	Name	Email
41	Reusch, Thorsten	treusch@ifm-geomar.de
42	Rose, Laura	rose@bio.lmu.de
43	Schlötterer, Christian	christian.schloetterer@vu-wien.ac.at
44	Schulenburg, Hinrich	hschulenburg@zoologie.uni-kiel.de
45	Sheppard, Anna	asheppard@zoologie.uni-kiel.de
46	Stadler, Tanja	tanja.stadler@env.ethz.ch
47	Stephan, Wolfgang	stephan@zi.biologie.uni-muenchen.de
48	Sturm, Anja	asturm@math.uni-goettingen.de
49	Tautz, Diethard	tautz@evolbio.mpg.de
50	Traulsen, Arne	traulsen@evolbio.mpg.de
51	Veber, Amandine	amandine.veber@math.u-psud.fr
52	Vogl, Claus	claus.vogl@vu-wien.ac.at
53	Wakolbinger, Anton	wakolbin@math.uni-frankfurt.de
54	Wiehe, Thomas	twiehe@uni-koeln.de
55	Wittmann, Meike	meike.wittmann@googlemail.com
56	Wolf, Andreas	wolf@medinfo.uni-kiel.de
57	Wu, Bin	bin.wu@evolbio.mpg.de
58	Zivkovic, Daniel	daniel.zivkovic@gmx.at