LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

DFG-Forschergruppe 1078
Natural selection in structured populations

Workshop - March 10 - 12, 2010
Lecture Hall B01.019, LMU BioCenter Martinsried

Next-Generation Sequencing: New Chances and Challenges for Evolutionary Genetics

Wednesday - March 10, 2010

09:00-09:15
Wolfgang Stephan – Introduction

09:15-10:15
Markus Schilhabel- Two Generations of Sequencing, and counting - Technologies and Applications

2nd Generation Sequencing. The technologies of FLX, SOLiD and Illumina, their applications and costs.
Where will it go from here? The 3rd generation and their methods.

10:15-10:45
Stephan Hutter – Demography and Selection in North-American *Drosophila melanogaster* inferred from genome wide SNP data

Next-gen sequencing allows for the creation of large-scale DNA polymorphism data within populations. Recently, 37 full genomes of *Drosophila melanogaster* coming from a natural population from North America have been sequenced. We use this data to assess single nucleotide polymorphism within this population. Focusing on the 22Mb X chromosome we extract a total of roughly 570 000 SNPs. This data forms the basis for a rigorous search of genomic regions that have currently undergone positive Darwinian selection. We apply statistical methods based on population genetics theory to uncover regions that show polymorphism patterns associated with recent selection. In order to make sure that these candidate regions are not false positives created by the demographic history of the population we furthermore estimate the most likely demographic scenario using an Approximate Bayesian Computation approach. This demographic model is then used as a neutral null model to evaluate the statistical significance of our putative regions under selection.

10:45-11:00 – *coffee*

11:00-12:00
Chris Wheat - A comparative analysis of Roche vs Illumina: transcriptome data assembly and implications

12:00-12:30
Miriam Linnenbrink - Expression variation at /B4galnt2/ is associated with a bleeding disorder in house mice – Which evolutionary forces govern this process?

The /B4galnt2/ gene is a conserved N-acetylgalactosaminyltranferase displaying a gastrointestinal epithelial cell-specific expression pattern in most mammals. In house mice, two highly divergent alleles segregate in natural populations and display complex signatures of selection. One allele class confers a gastrointestinal epithelial cell-specific expression pattern as seen in other mammals, while a second allele class confers a blood vessel expression pattern and is associated with a phenotype in mice that closely resembles a common human bleeding disorder, von Willebrand disease. Previous work showed that different local populations of /*Mus musculus domesticus*/ display dramatic differences in the frequency of

...

these alleles due to the recent action of natural selection. To further understand the nature of these forces, we are currently conducting a fine-scale analysis of allele frequencies across Europe. In addition, to determine whether alternative /B4galnt2/ alleles have been maintained by long-term balancing selection, we have performed a population survey of the gene region in a close relative of house mice, /*M. spretus*/. Our preliminary results indicate that although alternative /B4galnt2/ alleles have been maintained since the common ancestor of the /*M. musculus*/ species complex and /*M. spretus*/, widespread differentiation is present at /B4galnt2/ due to local selective pressures.

12:30-13:30 – *lunch break*

13:30-14:00
Annegret Werzner, Nicolas Svetec, Dirk Metzler, Wolfgang Stephan – Adaptation to a cold environment - QTL on the X-chromosome of *Drosophila melanogaster*

The subtropical African origin of *D. melanogaster* does not inhibit the species to spread and overwinter in temperate regions (McKenzie 1975). While colonizing Europe thermotolerance should be shaped by directional selection. One way to unravel the genome-wide distribution of candidate genes affecting the thermotolerance in *D. melanogaster* is the analysis of quantitative trait loci (QTL) (Mackay 2001, Morgan and Mackay 2006). Indeed, the European lines show significantly higher cold resistance compared to the ancestral African lines and overall five QTL positions could be located on the X-chromosome. One of these positions influences the cold resistance in opposing patterns for sexes. These QTL-sex interactions might preserve genetic variation within the population as males and females lack a consistent response to natural selection.

14:00-14:30
Stefan Laurent - Statistical evaluation of demographic models in *Drosophila melanogaster*

Southeast Asian populations of the fruit fly *Drosophila melanogaster* differ from ancestral African and derived European populations by several morphological characteristics. It has been argued that this morphological differentiation could be the result of an early colonization of Southeast Asia that predated the migration of *D. melanogaster* to Europe after the last glacial period (around 10,000 years ago). To investigate the colonization process of Southeast Asia, we collected nucleotide polymorphism data for more than 200 X-linked fragments and 50 autosomal loci from a population of Malaysia and measured several morphological traits (including ovariole number and egg size). We analyzed this new SNP dataset jointly with already existing data from an African and a European population by employing an Approximate Bayesian Computation approach. By contrasting different demographic models of these three populations, we do not find any evidence for an ancestral divergence between the African and the Asian populations and show that Asian and European populations of *D. melanogaster* share a non-African most recent common ancestor. By estimating posterior distributions for the parameters of our best model, we also show that Asian and European flies share a most recent common ancestor that existed about 4800 years ago (probably in the Middle East or Northeast Africa).

14:30-15:00
Lena Mueller - Gene expression variation in natural populations of *Drosophila*

Recent advances in genomic technologies have enabled global analyses of gene expression variation among individuals of a species. A surprising finding has been that many species harbour vast amounts of gene expression variation. The goal of this project is the evolutionary and functional analysis of genes showing significant expression variation within and between populations of the fruit fly *Drosophila melanogaster*. The objectives include: 1) microarray experiments to determine gene expression variation in two natural populations, 2) population genetic analyses of coding and regulatory sequences to elucidate the roles of selection, drift, and demography in the maintenance of gene expression variation, and 3) transgenic experiments to functionally test the effects of putative cis-regulatory variants on gene expression.

So far, we have used microarrays to identify genes showing significant expression differences between adult females from an African and a European population of *D. melanogaster*. The differentially-expressed genes include those with functions in insecticide resistance and proteolysis. Although several of the genes with the greatest expression divergence between populations show the same expression pattern in adult males, the vast majority show a difference only in females. This suggests that sex-specific regulatory adaptation has occurred in response to environmental change.

15:00-15:30 – *coffee*

15:30-16:00
Pleuni Pennings - How to detect population structure in American ants

The spatial structure of coevolutionary arms races of antagonistic species interactions are strongly shaped by the population structure and genetic diversity of the interacting species. We analyzed these population genetic parameters in three closely related ant species: the parasitic slavemaking ant *Protomognathus americanus* and its two host species *Temnothorax longispinosus* and *T. curvispinosus.* We sampled populations throughout the range of these species at six to eight microsatellite loci and an mtDNA sequence. We found high levels of genetic variation in all three species, only slightly less variation in the most common host *L. longispinosus*. Using Jost's D as a measure of differentiation we detected much stronger structuring in all species and less male-biased dispersal than previously thought. In my talk I will explain why D describes population differentiation better than Fst.

16:00-16:30
Tobias Pamminger - Induced anti-social parasite defense in ant colonies

Slavemaker ants, obligate social parasites, regularly conduct destructive raids to replenish their slave work force. During these raids, attacked host colonies suffer severe fitness costs due to the killing of nestmates and the loss of brood. Host species developed permanent defenses such as parasite recognition and they react aggressively towards slavemakers. Here we investigated the long-term response of free-living host colonies to a within-nest encounter with a slavemaker worker. We demonstrate not only that host ants effectively discriminate between slavemaker ants and conspecifics, but more surprisingly a short encounter with a dead slavemaker induces a long lasting, strong aggressive response to alien ants in general. This host response may be adaptive if an encounter with a slavemaker is a reliable indicator for an attack on the host colony in the near future. Induced host aggression represents a social immune response and indicates long-term social memory in the host ants. The induced elevated aggression was maintained over three days, but leveled off after 17 days. Costs associated with a higher aggression level might then be counter-balanced by a decreased risk of a parasite attack over time. This is the first example of an induced anti-social parasite defense in social insects.

16:30-17:00 - *tea*

17:00-18:00
Fritz Sedlazeck [1;2;3;4], Greg Ewing [2;5] & Arndt von Heaseler [1;2;3;4] - Next generation computing for next generation sequencing

[1]Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories, Vienna, Austria
[2]University of Vienna, Vienna, Austria
[3]Medical University of Vienna, Vienna, Austria
[4]University of Veterinary Medicine, Vienna, Austria
[5]Mathematics and BioSciences Group (MABS), Vienna, Austria

High throughput sequencing provides access to important information on genes, their function and genetic variation on genomes. One way of post processing is a reference guided assembly. For this approach, one needs an evolutionary closely related reference genome. However, in the end of 2009, only a few completely referenced organisms are available. Therefore, we have

also considered more distantly related organisms as reference candidates. In this study, we investigate the usability of the graphics card to gain speed and accuracy in mapping the reads. Further, we compare the performance of different assembly programs taking false positives into account. Here, we defined a false positive rate based on the number of misplaced or missing nucleotides per mapped read. This not only gives an indication of the possible missing SNPs, but it also takes the edge effect of local alignment into account.
We show that our approach is more accurate compared to currently used, and it achieves performance improvements of up to 112 fold, by using the graphics card.

Thursday - March 11, 2010

09:00-10:00

Casey Bergman - Population genomics of *Drosophila* transposable elements

Transposable elements (TEs) are dynamic and abundant mutagenic agents that are one of the major forces that influence genome structure and evolution. The fruitfly, *Drosophila melanogaster*, has long been a model species to discover the fundamental mechanisms of TE mobilization and their consequent impact on genome biology and evolution. I will briefly review results on TE abundance and distribution gained from the complete sequencing of the *D. melanogaster* genome. I will then show how comparing paralogous copies of TEs at different locations within a single genome sequence can provide insight into the age distribution of TE insertions that can be tested using population genomic data. Using a coalescent-based age-of-allele test, I will argue that the majority of retrotransposon insertions in a North American population of *D. melanogaster* are observed at frequencies predicted by their age since insertion under neutrality. Lastly, I will show how population genomic data from Roche/ 454 shotgun sequencing of multiple strains of *D. melanogaster* can be used to extend results on TE abundance from the reference genome and provide insight into the insertion preferences of multiple TE families.

10:00-10:30

Martin Hutzenthaler – A coalescent in continuous space

The Kingman coalescent provides a description of the genealogical relationship amongst a set of genes in a panmictic population of constant size. The Lambda-coalescent extends this model to populations whose genealogy exhibits multiple mergers. Especially some marine species like oyster or cod are argued to have multiple mergers in their genealogy. The corresponding forwards in time model of the Kingman coalescent and of the Lambda-coalescent are the Fleming-Viot process and the Lambda-Fleming-Viot process, respectively. These models are for unstructured populations. In recent work Etheridge and Barton introduced an analogue hereof for populations distributed across a spatial continuum like the two-dimensional plane. This model and some of its properties are subject of this talk.

10:30-10:45 – *coffee*

10:45-11:45

Alexis Stamatakis - Rapid Sequential and Parallel Evolutionary Placement of Short Sequence Reads

We present an Evolutionary Placement Algorithm (EPA) for the rapid assignment of sequence fragments (short reads) to edges of a given phylogenetic tree under the maximum-likelihood (ML) model.
The accuracy of the algorithm is evaluated on several real-world data sets and compared to placement by pair-wise sequence comparison, using edit distances and BLAST. We test two versions of the placement algorithm. One version is slow, but more accurate, because edge length optimization at the insertion position is done after the insertion of each short read while the other version is faster because edge lengths are only approximated at the insertion position.
For the slow version, additional heuristic techniques are explored that yield almost the same run time as the fast version with only a small loss of accuracy. When those additional heuristics

are employed, the run time of the more accurate algorithm is comparable to that of a simple BLAST search for data sets with a high number of short query sequences. Moreover, the accuracy of the EPA is significantly higher, in particular when the sample of taxa in the reference topology is sparse or inadequate. Our algorithm, which has been integrated into RAxML, therefore provides an equally fast but more accurate alternative to BLAST for tree-dependent assignment of the evolutionary origin of short sequence reads.

11:45-12:15
LISHA NADUVILEZHATH - Estimating parameters of different demographic models

We have developed a novel composite likelihood method to estimate demographic paramters such as divergence time, population mutation rate, and migration. Different simulation scenarios and comparisons to other methods show that our method gives not only reliable results but is also fast.

12:15-13:30 – *lunch break*

13:30-14:00
CORNELIA BORCK - Linkage Disequilibrium under Soft Sweeps

If the mutation rate θ to a selectively beneficial allele is sufficiently high it becomes likely, that a selective sweep is caused not by a single but by several individuals. Such an event is called a soft sweep and the complementary event a hard sweep, the classical case. I will describe the standard linkage deviation $\sigma_D^2$of two neutral loci in a neighbourhood of the selected locus under soft sweeps.

14:00-14:30
HILDEGARD UECKER - Soft selective sweeps in structured populations

A "soft selective sweep" refers to adaptation from multiple mutational origins. We describe a method to derive the probability of soft sweeps in a structured population. Various scenarios (continent-island, two islands) are discussed.

14:30-15:00
GREGORY EWING - Simulating Selection with the Coalescent.

15:00-15:30 – *coffee*

15:30-16:00
MAREIKE JOGLER, HELGE SIEMENS, HONG CHEN AND JÖRG OVERMANN -
Bacterial speciation - planktonic freshwater bacteria as a model system

The role of recombination, adaptation and selection in shaping bacterial diversity was elucidated by searching for different ecotypes within groups of closely related bacterial lineages (up to 100 % 16S rRNA gene sequence identity). Members of the family Sphingomonadaceae constituted an abundant fraction of the Alphaproteobacteria in the oligotrophic, alpine Walchensee and the mesotrophic, prealpine Starnberger See. Of these, two phylogenetically tight subgroups of Sphingomonadaceae, relatives of Sandarakinorhabdus limnophila as well as the novel lineage G1A, were identified by a seasonal clone library as the dominant Sphingomonadaceae. These two dominant groups were found to be physiologically active throughout the year by DGGE. In parallel, a large number of Sphingomonadaceae could be recovered in pure culture by a high throughput cultivation approach followed by a PCR based Sphingomonadaceae screening. Among them were 8 isolates corresponding to the S. limnophila-cluster whereas 65 isolates were related to the G1A-cluster. Based on their dominance and year-round activity in situ and the availability of several independent pure culture isolates, these two bacterial groups will be used as a model system to elucidate the presence of different ecotypes by qPCR. Their specific ecological niches will be identified by physiological characterisation and in situ growth test.

16:00-16:30

Hong Chen - Analysis of the population structure of aquatic Sphingomonadaceae by MLSA

In order to elucidate the role of recombination, adaptation and selection in shaping bacterial diversity, the family Sphingomonadaceae (Alphaproteobacteria) was used as the model group. First a multilocus sequence analysis (MLSA) approach which targets a set of 9 housekeeping genes (atpD, dnaK, EF-G, EF-Tu, gap, groEL, gyrB, recA, rpoB) in Sphingomonadaceae was established. Then they were used to elucidate the population structure and the significance of recombination events.
The new primers for these housekeeping genes were designed based on all available genome sequences of 5 strains of Sphingomonadaceae and 2 strains of the genus Erythrobacter (closest phylogenetic relative to the Sphingomonadaceae). Subsequently, 96 strains of Sphingomonadaceae were isolated from Starnbergersee and Walchensee, and subjected to MLSA analyses. Based on their rRNA gene sequences, these strains fall into 16 different groups. While 16S rRNA gene sequences were identical for certain members of one 16S rRNA group, the concatenated tree of all 9 housekeeping genes revealed the presence of a significantly larger divergence between the different strains. Most significantly, MLSA revealed the presence of distinct subclusters among individual 16S rRNA groups, suggesting different selection pressure between subclusters and the existence of distinct evolutionary units despite the identical or very similar 16S rRNA gene sequences.

16:30-17:00 – *tea*

17:00-18:00

Carsten Wiuf - ABC: A useful tool for analysis of large-scale data sets

Statistical tools for analyzing large-scale biological data sets are indispensable and in high demand. Bayesian approaches have been widely popular because they confer computational tractability through MCMC and other simulation techniques. However, despite of this, the posterior density and/or the likelihood of the data is often intractable, either because of the large amounts of data or because interesting models are highly complex. One solution circumventing this problem is Approximate Bayesian Computation (ABC); ABC operates on summary data to make broad inferences with less computation than might be required if all available data were analyzed. In recent years ABC has become a very widespread and popular tool in computational biology and genetics.
The talk will review ABC and problems encountered when using ABC for inference, such as how best to summarize the data. As an example I will discuss an application of ABC in the context of protein interaction networks and discuss ABC in relation to genomics data.

Friday - March 12, 2010

09:00-10:00

A.W. Nolte - Next generation sequencing as a means to access to the evolutionary genomics of non model organisms

Our understanding of the genomic basis underlying early evolutionary divergence is still limited. Progress in this field is driven by technological advances that permit access to genomic
data which includes both sequencing technology and the availability of sequenced genomes. While the former give an unprecedented access to raw sequences as such, the latter provide a backbone for many analysis approaches. Obviously, genetic analysis in model organisms is greatly fostered by the wealth of available resources. However, a vast number of evolutionary processes that attract the attention of evolutionary biologists are not observed in model organisms. Here, I will present two studies that use next generation sequencing to learn about the process of natural hybridization and the genomics of largely unexplored species of fishes. Two different study systems will be discussed that illustrate
1) how next generation sequencing can serve to integrate information from unexplored species with that of model organisms with fully sequenced genomes and

2) that next generation sequencing provides enough data to detect rare signals that would be difficult to identify using small scale approaches.

10:00-10:30
JUSTYNA WOLINSKA - **Parasites of Daphnia: how transmission mode affects the population genetic structure**

10:30-10:45 – *coffee*

10:45-11:45
LAUREN MCINTYRE - RNA-seq: Sample Size, Coverage and Experimental Design

11:45-12:45 – *lunch break*

12:45-14:00 - *PI Meeting, PhD Students and PostDocs meet with invited speakers*

14:00-15:00
SERGEY NUZHDIN - Genome-wide resequencing to detect local adaptation