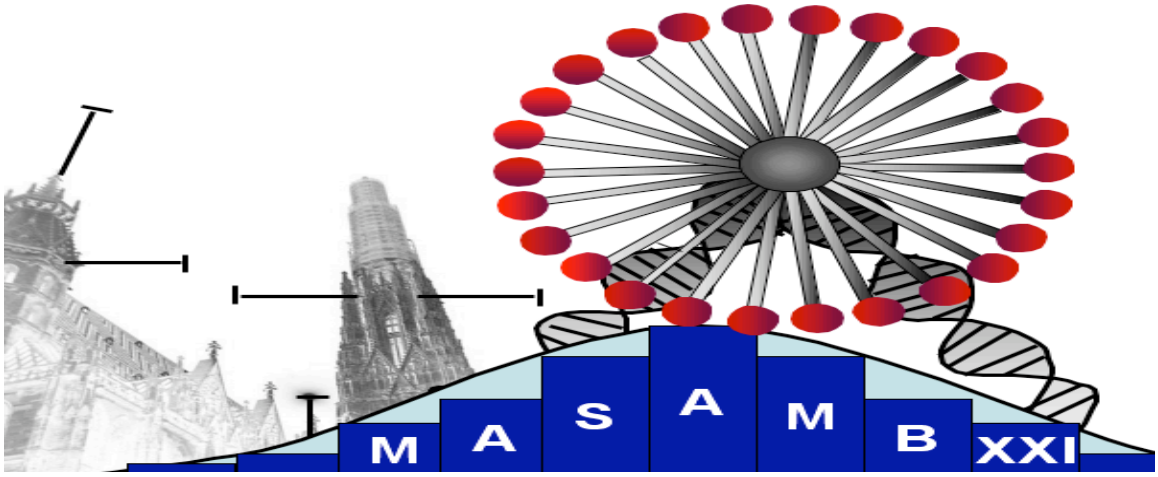

Programme for MASAMB 2011 (including abstracts)

MONDAY 11TH APRIL

11:00 - 13.45	**Registration**	
12:45 - 13.45	**Lunch**	
13:45 - 14:00	**Welcome and Introduction**	
SESSION 1 Population Genetics (NGS, SNPs and Structural Variation) CHAIR: Joachim Hermisson		
14:00 - 14:20	**Peter Arndt**, "Bayesian SNP polarisation"	**p. 3**
14:20 - 14:40	**Andreas Futschik**, "Optimal pooling strategies for SNP detection using next generation sequencing"	**p. 4**
14:40 - 15:00	**Luca Ferretti**, Sebastian Ramos-Onsins and Miguel Perez-Enciso "Population genomics from next generation sequencing data"	**p. 4**
15:00 - 15:20	**Ines Hellmann**, Olaf Thalmann, Anne Fischer and Linda Vigilant "Effects of sex-biased evolution on patterns of diversity in apes"	**p. 6**
15:20 - 15:40	**Botond Sipos**, Tim Massingham and Nick Goldman "Sequencing of repetitive genomic DNA aided by mutagenesis - a simulation study"	**p. 8**
15:40 - 16:00	**Lorenz Wernisch**, Klaudia Walter and Matt Hurles "Hierarchical Bayesian classifier for inferring genomic deletions"	**p. 9**
COFFEE BREAK		
SESSION 2 Evolutionary Genomics and Sequence Analysis CHAIR: Jonathan Bollback		
16:30 - 16:50	**Richard Goldstein**, Asif Tamuri and Mario Dos Reis "Characterising selection coefficients using from site- and time-dependent substitution models"	**p. 5**
16:50 - 17:10	**James Allen**, and Simon Whelan "Quantifying the effect of evolution and genomic alignment on de novo RNA gene prediction"	**p. 3**
17:10 - 17:30	**Anne Kupczok**, and Jonathan Bollback "Modeling the Evolutionary Dynamics of CRISPR spacers"	**p. 7**
POSTER SESSION + WINE		
20:00 - open end	**Conference Dinner at MARX Restauration**	

TUESDAY 12TH APRIL

SESSION 3 Gene expression and RNA-Seq
CHAIR: Christian Schlötterer

9:00 - 9:20	**Peter Glaus**, Antti Honkela and Magnus Rattray "Estimating differential expression of transcripts with RNA-seq by using Bayesian Inference"	**p. 5**
9:20 - 9:40	**Shengyu Ni**, "Gene expression comparison between RNA-seq and microarrays based on ranked genes list"	**p. 8**
9:40 - 10:00	**Simon Anders**, Alejandro Reyes and Wolfgang Huber "Studying alternative isoform regulation with RNA-Seq"	**p. 3**
10:00 - 10:20	**Alex Lewin**, and Ernest Turro "MMSEQ: Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads"	**p. 8**

COFFEE BREAK

SESSION 4 Genomics and Parameter Selection
CHAIR: Andreas Futschik

11:00 - 11:20	**Verena Zuber**, and Korbinian Strimmer "Improving biomarker discovery by taking account of correlation: the CAT-CAR approach"	**p. 9**
11:20 - 11:40	**Florian Frommlet**, Felix Ruhaltinger and Bogdan Malgorzata "Bayes optimal selection rules under sparsity with applications in GWAS"	**p. 4**
11:40 - 11:50	**Vote / Announcement of next MASAMB meeting**	

SHORT BREAK (group photograph, no cofee)

SESSION 5 Systems Biology I
CHAIR: Magnus Rattray

12:00 - 12:20	**Elisenda Feliu**, and Carsten Wiuf "Enzyme sharing as a cause of multistationarity in signaling systems"	**p. 4**
12:20 - 12:40	**Thomas Thorne**, and Michael Stumpf "Sequential Monte Carlo samplers for biological network inference"	**p. 9**
12:40 - 13:00	**James Hensman**, Magnus Rattray and Neil Lawrence "Functional nonparametric clustering of irregularly sampled gene expression time series"	**p. 6**

LUNCH BREAK

SESSION 6 Systems Biology II
CHAIR: David Kreil

14:00 - 14:20	**Emma Cooke**, Richard S Savage, Paul D W Kirk and David L Wild "Bayesian Hierarchical Clustering for Microarray Time Series Data with Replicates and Outlier Measurements"	**p. 3**
14:20 - 14:40	**Maxime Garcia**, Olivier Stahl, Pascal Finetti, Daniel Birnbaum, François Bertucci and Ghislain Bidaut "Biomarkers Discovery in Breast Cancer by Interactome-Transcriptome Integration"	**p. 5**
14:40 -15:00	**Julia Lasserre**, Alexander Zien, Klaus-Robert Müller and Martin Vingron "ProARTS: towards unfolding the structure of the sequence around human transcription start sites"	**p. 7**
15:00 - 15:20	**Alexey Stukalov**, and Jacques Colinge "Bayesian Inference of Protein Complexes from Mass Spectrometry Data"	**p. 9**
15:20 - 15:40	**Heather Harrington**, Gian Michele Ratto and Michael Stumpf "Spatio-temporal models of Erk1 and Erk2 in vivo"	**p. 6**
15:40 - 16:00	**Antti Honkela**, Neil Lawrence and Magnus Rattray "Hierarchical Gaussian Process Models of Gene Expression and Transcriptional Regulation"	**p. 7**

CLOSING STATEMENTS

Talk Abstracts

Quantifying the effect of evolution and genomic alignment on de novo RNA gene prediction

James Allen, and Simon Whelan

University of Manchester, 2 Lingard Road, Northenden, M22 4FN, UK

Non-coding RNA is biologically important and linked to a range of diseases, yet there are potentially many RNA genes of unknown function that do not belong to characterized RNA families. The de novo prediction of RNA genes is difficult because they are heterogeneous, and tend to have fewer restrictions than protein-coding genes. These problems are compounded by the lack of independent datasets for benchmarking predictions. We identify alignments of genes from 9 well-defined RNA families in UCSC genomic data for 32 vertebrate species, totaling 287 gene regions, and use randomization to create appropriate negative-control datasets. We evaluate three popular prediction programs (CMfinder, EvoFold, and RNAz), testing their ability to detect RNA genes and accurately determine gene boundaries. Our results show that RNA genes in genomic alignments have different evolutionary characteristics to structurally-aware RNA alignments, affecting gene prediction, and provide practical suggestions for the development and application of prediction programs.

Studying alternative isoform regulation with RNA-Seq

Simon Anders, Alejandro Reyes and Wolfgang Huber

European Molecular Biology Laboratory, Mayerhofstrasse 1, 69117 Heidelberg, Germany

In higher eukaryotes, most genes have several transcript isoforms, and the relative abundance of the isoforms of a gene may depend on tissue type and cellular state. Studying the regulation of relative isoform abundance is hence of importance. However, measured isoform abundance ratios often differ strongly between replicate samples (more so than whole gene expression values), and good estimation of their variance is crucial. Several recent publications ignored this issue, leading to loss of type-I error control and an excess of false positive findings.
To address this problem, we have developed a statistical analysis method that uses information borrowing across exons and genes for variance estimation and generalized linear models of the negative binomial family for testing. This allows for good detection power and reliable control of the false discovery rate. The method is being implemented as an R/Bioconductor package, which also offers facilities for result postprocessing and visualisation.

Bayesian SNP Polarisation

Peter Arndt

Max Planck Institute for Molecular Genetics, Ihnestr. 63/73, 14195 Berlin, Germany

With the availability of large amounts of population level data, for example on the position and frequency of SNPs, it is now feasible to study fitness effects of polymorphisms. However, since the ancestral state of a polymorphic site is in general not available, each SNP needs to be polarized, i.e. the ancestral state has to be inferred using data of related species. Here we present a Bayesian approach to SNP polarization, which is takes into account effects due to (i) different rates for the various mutation processes especially neighbor dependencies at CpG sites, (ii) parallel sequence evolution in the related species, and (iii) random genomic drift in the species under consideration. Our approach is superior to the generally used parsimony based methods.

Bayesian Hierarchical Clustering for Microarray Time Series Data with Replicates and Outlier Measurements

Emma Cooke, Richard S Savage, Paul D W Kirk and David L Wild

University of Warwick, MOAC DTC, Coventry House, Coventry CV4 7AL, UK

A key aim in systems biology is to link modelling of the interactions of system components with high throughput data. Time series experiments have become increasingly common, necessitating the development of novel analysis tools that capture the resulting data structure. We present a Bayesian hierarchical clustering algorithm for microarray time series that employs Gaussian process regression to capture the structure of the data. Using a wide variety of experimental data sets, we show that our algorithm consistently yields higher quality and more biologically meaningful clusters than current state-of-the-art methodologies. By using a mixture model likelihood, our approach permits a small proportion of the data to be modelled as outlier measurements which allows noisy genes to be grouped with other genes of similar biological function. Our method exploits replicate observations to inform a prior distribution of the noise variance, which enables the discrimination of additional distinct expression profiles.

Enzyme sharing as a cause of multistationarity in signaling systems

Elisenda Feliu, and Carsten Wiuf

Bioinformatics Research Centre, Aarhus University, C.F. Møllers Allé 8, 8000 Aarhus, Denmark

Multistationarity in biological systems is seen as a mechanism of cellular decision making. We consider biological systems where a signal is transmitted by phosphorylation and discuss emergence of multistationarity in small motifs that repeatedly occur in signaling pathways. Our motifs are built on a one-site modification cycle and account for features regarding the number of modification sites, and competition and non-specificity of enzymes.

We conclude that (a) multistationarity arises whenever a single enzyme catalizes the modification of two different but linked substrates; (b) multiple steady states require two opposing dynamics acting on the same substrate; (c) multistationarity in some of the systems occurs independently of the reaction rates.

The mathematical modeling is mass-action kinetics and thus steady states are solutions to a system of polynomial equations in the concentrations. By use of algebraic arguments, the conclusions are derived in full generality without restoring to simulations or random generation of parameters.

Population Genomics from Next Generation Sequencing Data

Luca Ferretti, Sebastian Ramos-Onsins and Miguel Perez-Enciso

CRAG and UAB, Dept. Ciencia Animal i dels Aliments, Facultat de Veterinaria, Universitat Autonoma de Barcelona, 08193 Bellaterra, Spain

Next generation sequencing technologies are the most promising source of data for population genetics. However, sequences obtained from pools and individuals at low coverage are unbalanced and prone to errors. We present estimators of variability and neutrality tests (Tajima's D, Fu and Li's D, Fay and Wu's H) for low coverage data, which deal with unequal read depth and sequencing errors. These statistics are particularly useful for pooled data. Since most information in low coverage data is contained in the frequency spectrum, we also discuss a class of optimized frequency-based neutrality tests that can be used for genomic applications. Finally, as an example of the usefulness of sequencing at low coverage for population genetic studies, we present applications to genome wide inference from pool data in Drosophila and in pigs.

Bayes optimal selection rules under sparsity with applications in GWAS

Florian Frommlet, Felix Ruhaltinger and Bogdan Malgorzata

Medical University Vienna, Spitalgasse 23, 1090 Vienna, Austria

We will present recent results on asymptotic optimality of multiple testing rules under sparsity given a certain decision theoretic framework. In particular we will show that procedures controlling the false discovery rate (FDR) adapt very well to unknown levels of sparsity. We provide precise conditions on the number of individuals, the number of tests, sparsity levels and FDR levels under which asymptotic Bayes-optimality holds. Similar results are shown for model selection in a multiple regression setting, where we introduce modifications of BIC which control the FDR. Our theoretical results are important for many applications in bioinformatics where the number of regressors is larger than the number of individuals. In particular we will discuss the performance of our model selection criteria in GWAS.

Optimal pooling strategies for SNP detection using next generation sequencing experiments

Andreas Futschik

Inst. f. Statistik, Univesity of Vienna, Universitaetsstrasse 5/9, 1010 Vienna, Austria

Being a cost effective strategy, it became popular to carry out sequencing experiments simultaneously on pools of individuals. As demonstrated for instance by Futschik and Schlötterer (2010) in the context of population genetic inference, this raises both opportunities and new statistical challenges.
The talk will present how to design pooling experiments in order to optimize the power of SNP detection in the presence of sequencing errors. While still controlling for type I errors, it turns out that separate sequencing of pools on a given number of lanes permits to detect SNP's whose frequency can be much lower than the probability of sequencing errors. When maximizing power, an important ingredient is the pool size which should neither be too small nor too large.

Biomarkers Discovery in Breast Cancer by Interactome-Transcriptome Integration

Maxime Garcia, Olivier Stahl, Pascal Finetti, Daniel Birnbaum, François Bertucci and Ghislain Bidaut

Centre de Recherche en Cancerologie de Marseille - CRCM U891 INSERM, 27 bd LeÃ Roure, 13273 Marseille, France

High-throughput gene-expression profiling technologies yeild genomic signatures to predict clinical condition or patient outcome. However, such signatures have limitations, such as dependency on training set, and lack of generalization. We propose a novel algorithm, ITI (Interactome-Transcriptome Integration, Garcia et al., in press) to extract a generalizable signature predicting breast cancer relapse by superimposition of a large-scale protein-protein interaction data over several gene-expression data sets. This method extends the Chuang et al. algorithm (2007), with the capability to extract a genomic signature from several gene-expression data sets simultaneously. It was trained with four breast cancer DNA microarray data sets and allowed the discovery of a breast cancer relapse signature constituted by 118 subnetworks that was generalizable on independant data. SVM classification shown an accuracy of 78% on Desmedt et al. dataset (2007). Several drivers genes were detected, including CDK1, NCK1 and PDGFB, some not previously linked to breast cancer relapse.

Estimating differential expression of transcripts with RNA-seq by using Bayesian Inference

Peter Glaus, Antti Honkela and Magnus Rattray

The University of Manchester, Oxford Road, Kilburn Building, room G33, Manchester M13 9PL, UK

High-throughput sequencing enables expression analysis at the level of individual transcripts. The analysis of transcriptome expression levels and differential expression estimation requires a probabilistic approach to properly account for ambiguity caused by shared exons and finite read sampling as well as the intrinsic biological variance of transcript expression.

We present a Bayesian approach for estimation of transcript expression level from RNA-seq experiments. Inferred relative expression is represented by MCMC samples from the posterior probability distribution of a generative model of the read data. We propose a novel method for differential expression analysis across replicates which propagates uncertainty from the sample-level model while modelling biological variance using an expression-level dependent prior. We demonstrate the advantages of our method using a RNA-seq dataset (Xu G. et al., RNA 2010) with technical and biological replication for both studied conditions.

Characterising selection coefficients using from site- and time-dependent substitution models

Richard Goldstein, Asif Tamuri and Mario Dos Reis

MRC National Institute for Medical Research, The Ridgeway, London NW7 1AA, UK

The distribution of selection coefficients is an important unresolved problem in population genetics. Standard models of amino acid substitutions are not able to address this issue adequately as they consider the selective constraints acting on each location at each point in evolutionary time to be the same, differing at most by a multiplicative constant. Recent models of sequence change have relaxed these assumptions. We describe new mutation-selection models that better represent the evolutionary process, and can characterise the site- and time-dependent selective constraints acting on the protein sequence. By applying these models to mitochondrial proteins, we observe a multi-modal distribution of selection coefficients of possible and accepted mutations, and can describe how these distributions depend on the local structure. We can also characterise the nature and changes in the selection acting on influenza proteins before and after host shift events.

Spatio-temporal models of Erk1 and Erk2 in vivo

Heather Harrington, Gian Michele Ratto and Michael Stumpf

Imperial College London, Biochemistry Building, London SW7 2AZ, UK

Mitogen activated protein kinase (MAPK) signalling cascades are pivotal elements of many eukaryotic signal transduction networks. Because of their fundamental importance they have also attracted considerable attention from modellers and a host of papers has recently been published focussing on different aspects of the MAPK signalling dynamics; this is particularly true for the Erk/Mek system, which has become the canonical example for MAPK signalling systems.
Erk exists in many different isoforms, of which the most widely studied are Erk1 and Erk2. These isoforms – which differ only subtly at the sequence level – show radically different trafficking between cytoplasm and nucleus. Here we use spatially resolved data on Erk1/2 and Mek1/2 to develop and analyse spatio-temporal models of these MAPK cascades; and we discuss how sensitivity analysis can be used to discriminate between different mechanisms. We are especially interested in understanding why two such similar proteins should co-exist in the same organism, as their functional roles appear to be different. Our models allow us to elucidate some of the factors governing the interplay between the phosphorylation and de-phosphorylation processes and the localization of ERK1/2 and Mek1/2 in different cellular compartments. Our analysis also allows us to illustrate the specific challenges that are raised by these increasingly emerging spatially resolved cellular signalling data. The results obtained from our spatially structured models compare well against recent single cell proteomic data on Erk/Mek system.

Effects of Sex-Biased Evolution on Patterns of Diversity in Apes

Ines Hellmann, Olaf Thalmann, Anne Fischer and Linda Vigilant

The Mathematics and BioSciences Group, Max F. Perutz Laboratories GmbH, Dr. Bohr-Gasse 9, 1030 Vienna, Austria

Apes show striking variability in the reproductive strategies that underlie their social systems. For example, behavioral studies suggest that most offspring in a gorilla group are sired by one dominant male, while bonobos are relatively promiscuous. Furthermore, observational data also suggest that apes have a strong sex bias in their migration patterns. Here we investigate potential long-term effects of reproductive strategies and female-biased migration on levels of diversity by comparing sequence data from mitochondria, autosomes, X- and Y-chromosome. To do so, we sequenced ten 5kb loci on the X and Y in ~10 bonobos, eastern, western and central male chimpanzees as well as ~10 male mountain and lowland gorillas, and then combined this with previous studies of autosomal and mitochondrial data.
We have too little data to detect a significant departure from the X/A of 0.75, which would be expected given no sex biases. However, the general pattern of X/A-ratios appears to be dominated by female biased migration rather than a big difference in reproductive variance between males and females.

Functional nonparametric clustering of irregularly sampled gene expression time series

James Hensman, Magnus Rattray and Neil Lawrence

The University of Sheffield, SITraN, 385A Glossop Road, Sheffield S10 2HQ, UK

Bayesian nonparametric clustering -- the infinite mixture model -- has been successfully used to cluster gene expression profiles (e.g. Rasmussen et al., IEEE-TCBB 2009, Medvedovic and Sigavanesan Bioinformatics, 2002), revealing informative structures regarding gene interaction. In these works, the gene expression is measured at regular intervals, and the measurements are treated as a vector: temporal correlations are not modeled.

We propose an extension to the infinite mixture model where temporal relations are modeled by a Gaussian process. This allows for clustering where data are not sampled at regular intervals or data are missing: our model deals with different experimental repeats with different temporal sampling schemes; and with experimental repeats in different species where time has been scaled so as to account for e.g. different species' embryonic development rates.

Further, we can include a simple ODE model of mRNA transcription/decay to cluster genes by their transcription rate profiles while simultaneously learning unknown model parameters.

Hierarchical Gaussian Process Models of Gene Expression and Transcriptional Regulation

Antti Honkela, Neil Lawrence and Magnus Rattray

University of Helsinki, Gustaf HÃ€llstrÃ¶min katu 2 b, P.O. Box 68, 00014 Helsinki, Finland

Biological systems are inherently dynamic and time series data provide great insight to understanding them. Such data are most naturally modelled in continuous-time framework that can be directly applied to data with diverse or uneven sampling. Gaussian processes provide a convenient tool for specifying priors over latent continuous-time functions in such models. Such models have previously been proposed for example with a differential equation model of gene regulation.

We propose extending these models with a hierarchical Gaussian process that allows modelling diverse experimental setups, such as mixed longitudinal/cross-sectional designs and phylogenetic structure. In the linear differential equation model, this approach can improve performance over our previous work (Honkela et al., PNAS 2010) even in simple transcription factor target ranking, and provides flexibility for modelling more complex data, such as the multi-species Drosophila data of Kalinka et al. (Nature 2010).

Modeling the Evolutionary Dynamics of CRISPR spacers

Anne Kupczok, and Jonathan Bollback

IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) is an adaptive heritable immune system found in Eubacteria and Archaea. The spacers between the repeats represent viral/plasmid targeting sequences and the system functions in an analogous way to the eukaryotic siRNA system. The length and content of the spacer array varies considerably among individuals within species (suggesting a rapid arms race) and it has been suggested that there is a selective cost, in the absence of parasites, associated with maintaining these arrays.
Therefore, the rate at which spacers are gained and lost from these arrays provides insight into the evolutionary dynamics of host-parasite interactions. To this end we model spacer insertion and deletion as a continuous-time Markov process along the phylogeny. The maximum-likelihood framework allows us to estimate the overall rates of gain and loss of spacers, relative evolutionary rates of viral and plasmid spacers, and how these differ among bacterial species. We evaluate different models by simulation and analyse bacterial data sets.

ProARTS: towards unfolding the structure of the sequence around human transcription start sites

Julia Lasserre, Alexander Zien, Klaus-Robert Müller and Martin Vingron

Max-Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany

Characterizing promoters is fundamental to solve TSS annotation, but despite much effort to build powerful detectors, TSSs and promoters are generally not well understood. One reason for this might be the heterogeneity of the arrangement, around the TSS, of the sequence elements which are responsible for transcriptional regulation. To tackle this hurdle, we suggest a new approach that relies on energy profiles and on classification through Support Vector Machines. Promoters are first clustered according to their energy profiles. Sequences of one particular cluster are then classified against TSS-free sequences and the remaining promoters. Our method's performance compares with the state-of-the-art ARTS from Sonnenburg etal, confirming that relevant motifs are learnt. However, unlike prior work, it is set up in an original manner that allows the use of classification to find cluster-specific words. Based on those, we observe cluster-specific arrangements of kmers around the TSS, as a step towards categorizing promoters.

MMSEQ: Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads

Alex Lewin, and Ernest Turro

Imperial College London, Dept Epidemiology and Biostatistics, St Mary's Campus, Norfolk Place, London W2 1PG, UK

Second-generation sequencing of RNA samples (RNA-seq) allows researchers to quantify the abundance of transcripts in biological samples with greater sensitivity than microarrays, which suffer from non-specific hybridisation and saturation biases. RNA-seq, however, poses new statistical and computational challenges. Sequence reads are in general much shorter than transcripts and may therefore map to multiple isoforms of the same gene or even to multiple genes sharing a subsequence. Moreover, paired-end reads and reads that straddle one or more exon junctions complicate the mapping process. There is a need for methods that deal with isoform-level quantitation using all mappable reads in a principled, scalable and user-friendly way.

We present a novel pipeline and methodology for simultaneously estimating isoform expression and allelic imbalance in diploid organisms using RNA-seq data. We achieve this by modelling the expression of haplotype-specific isoforms. If unknown, the two parental isoform sequences can be individually reconstructed. A new statistical method, MMSEQ, deconvolves the mapping of reads to multiple transcripts (isoforms or haplotype-specific isoforms). Each read is mapped to a set of reference transcripts and the total number of reads mapping to each set is counted. The contribution from multi-mapping reads to the expression signals is then disaggregated using expectation-maximisation and Gibbs sampling to provide expression estimates and accompanying standard errors for each isoform and gene. Our software can take into account non-uniform read generation and works with paired-end reads. We show results of the method applied to a data set consisting of 61 HapMap individuals, and an F1 hybrid mouse brain data set.

Gene expression comparison between RNA-seq and microarrays based on ranked genes list

Shengyu Ni

Partner Institute for Computational Biology, C/O Abt. Vingron, Ihnestr. 73, 14195, Berlin, Germany

With the recent improvements in the efficiency, quality, and cost of genome-wide sequencing, RNA-seq is now an alternative way for high-throughput studies of gene expression. However, RNA-seq might confront unnecessary barrier where a new experiment need to compare results to a database, where the reference experiments in the database are done by microarray, for example to ascertain the impact of uncharacterized perturbations on the cell. We present here a new measure “R2KS” which enables experiments from different technologies to be compared. The measure is then applied to tissue specific gene expression data and shows highly correlation for the same tissue from different platform, we also discussed the technology bias shown by our measure.

Sequencing of repetitive genomic DNA aided by mutagenesis - a simulation study

Botond Sipos, Tim Massingham and Nick Goldman

EMBL-EBI, Welcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

The increased throughput from Next Generation Sequencing (NGS) technologies has transformed large-scale de novo genome sequencing into a practical research tool. However, it is hard to obtain high-quality assemblies for genomes harbouring a significant amount of repetitive DNA, for example recent paralogues or CNVs, because the limited length of sequence reads results in assembly ambiguities.

Using realistic simulation experiments, we explore the feasibility of a "Sequence Analysis by Mutagenesis" approach, relying solely on PCR-based experimental preparation, to resolve difficult regions. This simple idea, pioneered with traditional sequencing technologies, relies on comparing the sequences of randomly mutated - hence less repetitive - copies of the original problematic molecule, and could in principle solve the problem of genome assembly from short-read data. Emerging microfluidics technologies, like the one marketed by RainDance Technologies, may make experiments using this strategy practical in the near future.

Bayesian Inference of Protein Complexes from Mass Spectrometry Data

Alexey Stukalov, and Jacques Colinge

CeMM-Research Center for Molecular Medicine GmbH, Lazarettgasse 14/AKH BT 25.3, 1090 Vienna, Austria

Affinity-purification assays combined with Mass Spectrometry (AP/MS) constitute a powerful and widely used technology for the discovery of protein-protein interactions and protein complexes. However, the interpretation of AP/MS data is complicated by the presence of contaminants and non-specific binders, and the absence of some true interactions. Moreover, existing protein complex prediction methods were developed for genome-wide datasets with a large number of purification experiments. They are not applicable to targeted studies that focus on a limited number of protein complexes.

To overcome these limitations we have developed a new Bayesian checkerboard bi-clustering method. The Bayesian model parameters are fit by an advanced MCMC parallel sampler. The result is a probability distribution over plausible protein complexes, which are recurrently observed in AP/MS experiments. By utilizing both protein co-occurrence information and semi-quantitative MS data, we are able to improve predictions accuracy over methods that only use one type of data.

Sequential Monte Carlo samplers for biological network inference

Thomas Thorne, and Michael Stumpf

Imperial College London, Biochemistry Building, Imperial College, London SW7 2AZ, UK

We present Sequential Monte Carlo methods that allow us to infer network structures in Systems Biology. These methods offer computational and conceptual advantages over standard MCMC approaches which we highlight in two contexts: (i) we develop an SMC procedure that allows us to infer dynamical Bayesian networks from gene expression data at the whole genome scale; (ii) in an approximate Bayesian computation (ABC) setting we calibrate models of network evolution against available protein-protein interaction data; here we employ ideas from spectral graph theory to compare simulated and real networks without incurring loss of information through the use of extraeneous summary statistics. This allows us to employ model selection approaches directly, and we analyse a number of models of network evolution in light of the available network data.
In both applications we are able to illustrate the use of massively parallel GPU architectures that SMC based samplers are readily adapted to.

Hierarchical Bayesian classifier for inferring genomic deletions

Lorenz Wernisch, Klaudia Walter and Matt Hurles

MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 0SR, UK

Predicting discrete outcomes utilizing several prediction methods is a standard problem in bioinformatics and genomics. If a gold standard is available, it can be used for calibration of a joint predictor combining the various methods. However, if a gold standard is not available or not good enough, we suggest a comprehensive statistical model where the true state of the predicted variable is considered latent.
As an example, we analyse deletion predictions for the Pilot 1 data (whole genomic sequence data on 179 samples) of the 1000 Genomes Project, where ten different prediction methods and four different external validation methods (eg. PCR, aCGH) need to be combined into one coherent set. Our model based on variational methods provides a joint predictor, which is superior to each individual predictor, without the need of a gold standard, but using external validation data. The model also produces posterior distributions for parameters of interest, such as a false discovery rate for each predictor.

Improving biomarker discovery by taking account of correlation: the CAT-CAR approach

Verena Zuber, and Korbinian Strimmer

University of Leipzig, IMISE, Härtelstraße 16-18, 04107 Leipzig, Germany

Due to the high dimensionality of genomic data variable selection is a vital part of many biostatistical analyzes. While most traditional approaches ignore the correlation structure among predictor variables we argue here that taking account of correlation is beneficial, e.g., for ranking genes in differential expression, or for finding biomarkers in a regression setup. Specifically, we introduce two novel scores, the CAT and the CAR score, that explicitly take account of correlation in classification and linear regression, respectively. Here, we show that squared CAR scores are a natural measure of variable importance that provide a canonical ordering of the explanatory variables and encourage grouping of correlated predictors. In computer simulations we demonstrate that CAR scores are a highly effective means for variable selection with a prediction error and true and false positive rates that compares favorable with elastic net, boosting and other modern regression approaches. We illustrate the approach by analyzing a data set on disease progression in diabetes as well as gene expression data investigating the effect of aging in the brain. R packages for computing CAT and CAR scores are available from CRAN.