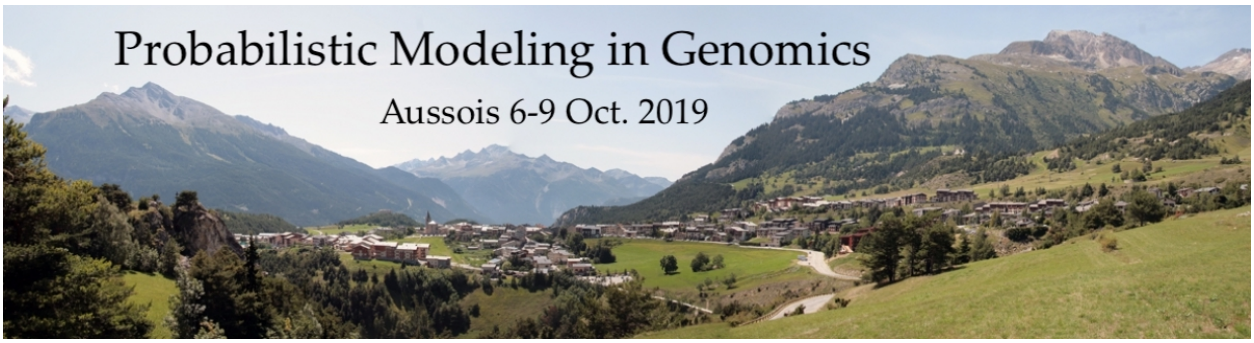

Table des matières

Oral presentations

Leveraging linkage disequilibrium in human population-genetic analyses (Ramachandran)

Ramachandran Sohini, Brown University, USA

Correlation among genotypes in human population-genetic datasets reflects certain aspects of population histories and complicates the localization of both adaptive and deleterious mutations. I will describe my view on how linkage disequilibrium can both complicate and enhace recent efforts to develop methods for localizing adaptive and disease-causing mutations, motivated by (1) incorporating summary statistics at various genomic scales into selection scans, (2) bridging the gap between polygenic and omnigenic complex traits, and (3) testing for differential genetic architecture for the same trait across ancestries.

Leveraging chromosomal linkage to infer natural selection from full-genome sequencing data (Friedlander)

Friedlander Eric <EricF2218@gmail.com> (1), Steinrücken Matthias <steinrue@uchicago.edu> (1) 1 - Ecology & Evolution, University of Chicago (United States)

We present a method for inferring natural selection from full-genome sequencing data obtained from a population with arbitrary population size history. Natural selection for a certain allele tends to increase its frequency within a population and leaves signatures in sequencing data which can be used to infer the target and strength of selection. Due to chromosomal linkage, selection also impacts the genetic variation in nearby neutral regions. Leveraging this signature, one can substantially increase the power of inference methods. However, changes in population size can yield patterns in the data that mimic the effects of selection. Therefore, the demographic history of the population has to be explicitly accounted for, which is not standard practice in most inference frameworks. Our framework is built on the two-locus Wright-Fisher diffusion that describes the haplotype frequency dynamics of two linked loci separated by a certain recombination distance. General explicit solutions are not known for the transition density of this diffusion when selection and recombination act simultaneously, as it requires solving a multidimensional system of partial differential equations. Thus, we use the so-called ``Method of Moments'', in which the moments of the transition density are expressed as solutions to ordinary differential equations. These moments can subsequently be used to compute the likelihood of the observed genetic variation in a sample from the population. A key technical challenge in this method is that the moments do not «close» in models with selection. Namely, moments of order n depend on those of order n+1, n+2, etc. We surmount this challenge by developing a novel method to estimate higher order moments from those of lower order, which can also be applied to the general problem of estimating allele frequency spectra for large samples from smaller samples. Using these efficient approximations to the dynamics of the diffusion and two-locus likelihood computations, we develop a composite likelihood framework for estimating the strength of selection in a population with arbitrary population size history from full-genome sequencing data. Furthermore, since we are working with a two-locus model, our framework can infer the strength of selection from segregating sites near the site under selection even if the selected variant is fixed in the sample. We demonstrate the accuracy and efficiency of the proposed methods on simulated data and show applications to the 1000 genomes dataset.

Modeling the Spatial Distributions of Rare Deleterious Alleles (Rice)

Rice Daniel <dpr@uchicago.edu> (1), Porras Christian <chrisporras1@uchicago.edu> (1), Novembre John <jnovembre@uchicago.edu> (1) 1 - Department of Human Genetics, University of Chicago (United States)

Evolutionary theory and recent large-scale population genomic data sets suggest that rare large-effect genetic variants can play an important role in the genetic basis of disease risk. Rare alleles tend to be young and, as a result, tend to be geographically localized. We are working to develop a better quantitative

understanding of the geographic distributions of rare alleles with the goal of improving population and complex-trait genetic methods. We will demonstrate how the theory of superprocesses can be used to model the geographic spread of a de novo deleterious mutation. In particular, we calculate the expected site frequency spectrum of a sample collected from one or more localized regions of a continuous spatial habitat. This result may aid in interpreting the genetic architecture inferred from GWAS studies and in assessing the validity of out-of-sample prediction across continuous space.

Transient Selection on Introgressed Neanderthal Alleles in Early Modern Humans (Peter)

Peter Benjamin <benjamin_peter@eva.mpg.de> (1) 1 - Max Planck Institute for Evolutionary Anthropology (Germany)

Gene flow from Neanderthals into modern humans has recently been shown to be a potent source of adaptive alleles. To understand the fine-scale timing of introgression events, we present admixfrog, an empirical Bayes method to identify introgressed fragments in low-coverage, contaminated DNA. In contrast to other Hidden-Markov-Model based approaches, Admixfrog uses an allele frequency- based approach and estimates most parameters directly from the data. The underlying model takes into account that ancient DNA is frequently ascertained to specific sites, low-coverage, and contaminated by a closely related population. In addition, available reference panels from archaic populations (i.e. Neanderthals or Denisovans) are typically small. Using simulations and testing on empirical data, we find that the method is able to accurately infer introgressed tracts at coverages as low as 0.1X, and with contamination rates exceeding 80%. When applying admixfrog to the earliest sequenced early modern humans (older than 25,000) years, we identify an excess of long (>1cM) introgressed fragments in all genomes, as well as an excess of sharing of such haplotypes between Western Eurasian genomes. These patterns suggest these individuals had numerous Neanderthal ancestors throughout their recent past, and that gene flow from Neanderthals into Western Eurasians was pervasive and ongoing until the extinction of Neanderthals. We also identify three candidate regions for adaptive introgression, where more than 80% of individuals carry a multi-cM introgressed haplotype. Remarkably, non of these regions went on to become fixed in present day population, suggesting that either the selective advantage was very temporally restricted, or that these loci are currently under balancing selection.

Bayesian Factor Analysis for the Inference of Population Genetic Structure from Temporal Samples (Jay)

Jay Flora <flora.jay@lri.fr> (1), Liégeois Séverine <sliegeois@yahoo.fr> (1), Demaille Benjamin <benjamin.demaille@lri.fr> (1), Francois Olivier <Olivier.Francois@imag.fr> (2) 1 - Université Paris-Sud, Université Paris-Saclay (France), 2 - Univ. Grenoble Alpes (France)

For many organisms, the number of temporal samples of DNA or ancient DNA (aDNA) has increased dramatically in recent years. In analyzing such data, a central question is to evaluate the relationships between sampled populations, or to determine which present-day population an ancient sample is closest to. To address this question, one of the most frequently-used algorithm is based on principal component analysis (PCA). When PCA is applied to temporal samples, time is, however, ignored during analysis. Some authors showed that time differences in samples can bias principal axes, creating sinusoidal shapes similar to those observed in spatial data. Alternative methods that combine ancient and modern samples by using projections on present-day samples also suffer from bias and shrinkage toward the center of the principal axes. Since those biases could lead to misinterpretations or to incorrect estimates of ancestry coefficients, it is important to propose methods to correct PCA when temporal samples are analyzed with this method. Here we introduce a new factor analysis (FA) method for describing ancestral relationships among samples collected at distinct time points in the past. Our motivation is to propose a factorial decomposition of the data matrix similar to a PCA, in which individual scores are corrected for the effect of allele frequency drift through time. Based on a diffusion approximation, our approach approximates allele frequency drift in a random mating population by a Brownian process. Using the Karhunen-Loeve theorem, we propose an equivalent representation of the factor model in which additional covariates, representing temporal eigenvectors, are introduced in the factor model. This representation makes use of informative Gaussian

prior distributions for the effect sizes of the temporal eigenvectors. Exact solutions for maximum a posteriori estimates of time-corrected factors can be obtained, and a fast algorithm based on random projections is proposed. We compared temporal FA with PCA and with PCA projections in coalescent and generative simulations of divergence and admixture scenarios. Distortions caused by temporal sampling were corrected by temporal FA in divergence scenarios. In admixture scenarios, estimates of ancestry coefficients were more accurate than those inferred from PCA. Next we applied temporal FA to study the evolution of hepatitis C virus in a patient infected by multiple strains, and to describe population structure for aDNA samples from ancient Europeans. The methods described in this abstract were implemented in the R package temporalFA available from a gitlab link.

A deep-learning approach for detecting selective sweeps based on the ancestral recombination graph (Mo)

Mo Ziyi <mo@cshl.edu> (1), Hejase Hussein <hijazi@cshl.edu> (1), Siepel Adam <asiepel@cshl.edu> (1) 1 - Simons Center for Quantitative Biology [Cold Spring Harbor] (United States)

Detecting signals of natural selection is a central problem in population genetics. Signals of selective sweeps provide insight into the nature of recent adaptation in modern humans and the sites of ongoing sweeps are likely to be associated with traits of interest. Traditionally, detecting selective sweeps involves designing summary statistics (such as Tajima's D) that capture spatial patterns of genetic diversity and haplotype structure in ways that are sensitive to perturbations created by selective sweeps. Recently, investigators have developed supervised machine learning methods, such as S/HIC, that improve prediction power by aggregating a collection of such summary statistics. However, even in combination, these low-dimensional summary statistics capture only a small portion of the information in sequence data. Here, we introduce a supervised machine learning method for selective sweep prediction that makes use of a much richer set of evolutionary features. These features are extracted from the ancestral recombination graph (ARG) of the samples, as inferred from sequence data by ARGweaver. Specifically, we summarize the ARG using the number of lineages remaining at discrete time points of the genealogies along the genome. Patterns in this feature set, such as outlier clusters of coalescent events and coalescent time to most recent common ancestry, can be learned to classify sweeps and predict selection coefficients. A deep learning model was trained on local genealogies inferred from single-population simulations that mimic the demographic history of CEU (European), YRI (Yoruba), and CHB (Chinese) populations. The model achieved an AUROC of 0.96 for classifying sweeps and neutral regions and a Pearson correlation of 0.88 between true and predicted selection coefficients in the simulated data. When applied to local genealogies inferred by ARGweaver on the 1000 Genomes dataset, our model was able to detect sweep signals and infer selection coefficients on the LCT locus in the CEU population, the NCOA1 locus in the YRI population, as well as other known examples of recent adaptation (e.g. pigmentation genes). By leveraging the abundance of information contained in ARGs, our method has the potential to greatly improve the power to detect selective sweeps and accurately infer selection coefficients.

Genealogical Inference from Thousands of Ancient and Modern Samples (Wohns)

Wohns Anthony <awohns@gmail.com> (1), Wong Yan <yan.wong@bdi.ox.ac.uk> (1), Mcvean Gil <gil.mcvean@bdi.ox.ac.uk> (1) 1 - Big Data Institute, University of Oxford (United Kingdom)

The thousands of ancient human genomes published over the last decade have provided important insights into ancestral demography and patterns of selection. Using these genomes together with the millions of publicly available modern genomes in a genealogical inference framework, such as tsinfer, provides novel insight into how ancient and modern samples are related. Such an approach could significantly aid our understanding of the forces shaping human genetic diversity. The tsinfer algorithm estimates the state and relative age of ancestral haplotypes to create genealogical topologies of contemporaneous samples; however, non-contemporaneous (ancient) genomes may be used if they are inserted as potential ancestral haplotypes at the correct relative age. Integration of ancient samples into tsinfer is thus dependent on estimating the age of nodes in an inferred genealogy. In this work, we first develop, implement, and test an importance sampling-based method for estimating node ages conditional on tree sequence topologies.

Second, we use these dated genealogies, combined with ancient samples, to improve estimates of mutation age. Finally, we infer genealogies where ancient haplotypes can serve as ancestors of modern samples. Using simulations, we demonstrate that the accuracy of both node and mutation age estimates is improved with both increasing sample sizes, from n=100 to 10,000, and population-scaled mutation rate, as well as with the inclusion of greater numbers of ancient samples. We demonstrate that the resulting tree sequences contain rich information on genomic descent from ancient samples, with preliminary results from a time series dataset of ancient genomes.

Inference under the exact coalescent with recombination (Mahmoudi)

Mahmoudi Ali <amahmoudi@student.unimelb.edu.au> (1), Balding David <david.baldin@unimelb.edu.au> (2), Chan Yao-Ban <yaoban@unimelb.edu.au> (3) 1 - School of Mathematics and Statistics, and Melbourne Integrative Genomics, The University of Melbourne (Australia), 2 - School of Mathematics and Statistics, School of BioSciences, and Melbourne Integrative Genomics, The University of Melbourne (Australia), 3 - School of Mathematics and Statistics, and Melbourne Integrative Genomics, The University of Melbourne (Australia)

A challenging problem in population genetics is to infer the full genealogical history of a sample of DNA sequences, otherwise known as the Ancestral Recombination Graph (ARG), under the coalescent with recombination. Inferring the ARG remains a problem since, for even a small number of DNA sequences, the state space of the ARG is large. Many different methods have been proposed to perform the inference, however, most of them have been limited to small datasets. One reason that these methods are not efficient for large sample sizes is because of the way they store and represent the genealogies, i.e., the data structure. Previous methods use a data structure in which each marginal tree is stored separately. This leads to inefficiencies, as neighboring trees in a genealogy share many parts. In order to gain efficiency and reduce processing time and storage capacity, taking these similarities into account is key. In 2016, an efficient data structure known as Tree Sequence Recording (TS) was introduced by Kelleher, Etheridge, and McVean to store the genealogical trees at each site. In this method, identical parts of consecutive trees are stored only once. More recently, an inference method, tsinfer, was proposed to infer whole-genome genealogies. This method leverages the features of TS and is applicable to large data sets. tsinfer infers the genealogical trees at each site, however, it is not a probabilistic inference model. Rather, it concentrates on compactly storing large datasets in a novel 'evolutionary encoding' format that enables more efficient access and processing of the data. In this work, we present a Markov chain Monte Carlo (MCMC) approach to perform probabilistic inference under the coalescent with recombination. Borrowing the idea of storing the genealogies with no repeated information from TS, we introduce a data structure to represent the full ARG. Under the infinite sites mutation model, we infer the full ARG and, unlike tsinfer, our method infers both genealogical trees and event times. Hence, the time to the most common ancestor, the ancestral state at each time, and the total branch length are obtained. We demonstrate the utility of our method by applying it to simulated datasets. Also, we compare our method with ARGweaver, state-of-the-art probabilistic method. Keywords: Statistical Genetics, The Coalescent with Recombination, Ancestral Recombination Graph, Population Genetics.

Improving inference of homologous recombination using state-of-the-art computational methods (Everitt)

Medina Aguayo Felipe <f.j.medinaaguayo@reading.ac.uk> (1), Everitt Richard <richard.g.everitt@gmail.com> (2), Didelot Xavier <xavier.didelot@gmail.com> (2) 1 - University of Reading (United Kingdom), 2 - University of Warwick (United Kingdom)

Recombination is a critical process in evolutionary inference, particularly when analysing within-species variation. In bacteria, despite being organisms that reproduce clonally, recombination commonly occurs when a donor cell contributes a small segment of its DNA. This process is typically modelled using an ancestral recombination graph (ARG). The ClonalOrigin model ([Didelot et al. 2010]) can be regarded as a good approximation of the ARG, in which recombination events are modelled independently given the clonal genealogy. Inference in the ClonalOrigin model is performed via a reversible-jump MCMC (rjMCMC)

algorithm, which attempts to jointly explore: the recombination rate, the number of recombination events, the departure and arrival points on the clonal genealogy for each recombination event, and the sites delimiting the start and end of each recombination event on the genome. However, as known by computational statisticians, the rjMCMC algorithm usually performs poorly due to the difficulty of proposing 'good' trans- dimensional moves. Recent developments in Bayesian computation methodology provide ways of improving existing methods and code, but are not well-known outside the statistics community. We present ideas based on sequential Monte Carlo (SMC) methodology that can lead to faster inference when using the ClonalOrigin model. (This is joint work with Felipe Medina Aguayo and Xavier Didelot.)

Inferring the landscape of recombination using recurrent neural networks (Kern)

Kern Andrew <adkern@uoregon.edu> (1), Adrion Jeffrey <jadrion@uoregon.edu> (1), Galloway Jared <jgallowa@uoregon.edu> (1) 1 - Institute of Ecology and Evolution, University of Oregon (United States)

Accurately inferring the genome-wide landscape of recombination rates in natural populations is a central aim in genomics, as patterns of linkage influence everything from genetic mapping to understanding evolutionary history. Here we describe ReLERNN, a deep learning method for accurately estimating a genome-wide recombination landscape using as few as four samples. Rather than use summaries of linkage disequilibrium as its input, ReLERNN considers columns from a genotype alignment, which are then modeled as a sequence across the genome using a recurrent neural network. We demonstrate that ReLERNN outcompetes existing methods and maintains high accuracy in the face of demographic model misspecification. We apply ReLERNN to natural populations of African Drosophila melanogaster and show that genome-wide recombination landscapes, while largely correlated among populations, exhibit important population-specific differences. Lastly, we connect the inferred patterns of recombination with the frequencies of major inversions segregating in natural Drosophila populations.

Reconstructing the genotypes of parents from siblings and close relatives (Qiao)

Qiao Ying <yq76@cornell.edu> (1), Jewett Ethan <ejewett@23andme.com> (2), Mcmanus Kimberly <kmcmanus@23andme.com> (2), Freyman Will <willf@23andme.com> (2), Blangero John <john.blangero@utrgv.edu> (3), Williams Amy <awilliams@cornell.edu> (1), Team The 23andme Research <research-team@23andme.com> (2) 1 - Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY (United States), 2 - 23andMe, Inc., Mountain View, CA (United States), 3 - Department of Human Genetics and South Texas Diabetes and Obesity Institute, School of Medicine, University of Texas of the Rio Grande Valley, Brownsville, TX (United States)

Children inherit two chromosome copies, one from each parent, with both formed via recombination. While each child inherits only half of the genomes of each parent, independent segregation and recombination are randomized such that n siblings will inherit on average a proportion of 1-(1/2)^n of both parents' genomes. Thus, the opportunity exists to reconstruct partial genomes of parents from a set of genotyped children. The successful reconstruction of ungenotyped parents has the potential to enable improved power in genome-wide association studies (GWAS) by adding more samples when the parents' phenotypic information is available. The inferred genotype data also makes it possible to resolve population origins of the segments of parents' genomes, and it can empower more precise relatedness inference. We developed a novel approach to reconstruct the genotypes of parents using a combination of family-based phasing of a set of siblings and identical by descent (IBD) sharing to other close relatives. To phase the siblings, we use a new extension of HAPI that enables joint phasing of siblings even without data from their parents. This joint phasing has low error and, when given data for ~8 or more siblings, often provides chromosome-scale haplotypes for the parents. For more moderate numbers of siblings, the phasing results in a number of multi-megabase long segments where which parent they belong to is ambiguous. We therefore leverage inferred IBD segments between the children and one or more their genotyped relatives to resolve the ambiguity. More specifically, since the IBD segments shared between the children and a given relative will generally be inherited from only one parent, this approach assigns the phased segments to the corresponding parent based on the IBD sharing. We tested this method on data from the San Antonio Mexican American Family Studies (SAMAFS) using 42 families with four siblings and seven families with

8-12 siblings. For the families with >= 8 children, our method inferred 90.3% of the two parents' phased haplotypes on average without using any other relatives, and with an error rate < 1e-3. In the largest SAMAFS family with 12 siblings, we reconstructed 95.2% of both parents' haplotypes with an error rate of 1.1e- 4. When analyzing the families with four children together with one of their second degree relatives, our approach reconstructed an average of 67.8% of parents' genotypes with an error rate < 1e-3. We also tested the method on families in the 23andMe dataset while leveraging IBD sharing information to their close relatives; this yielded similarly positive results that we will present. As large-scale datasets lead to the indirect recruitment of family data, our work holds promise to enable high-quality reconstruction of parent genotypes, opening the door to further analyses using inferred genotypes from individuals not directly collected.

How genetic risk for common disease changes with age (Jiang)

Jiang Xilin <xilin@well.ox.ac.uk> (1) (2), Holmes Chris <cholmes@stats.ox.ac.uk> (3) (1), Mcvean Gil <gil.mcvean@bdi.ox.ac.uk> (1) 1 - Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK (United Kingdom), 2 - The Wellcome Trust Centre for Human Genetics [Oxford] (United Kingdom), 3 - Department of Statistics [Oxford] (United Kingdom)

Genetics risk scores have great potential for disease prediction. However, most methods for estimating genetic risk assume that the effect is constant over age. Here, we present a framework for estimating how genetic risk changes with age and demonstrate that for many common diseases there is both age-dependent heterogeneity in how genetic risk affects future disease and, for a subset of diseases, multiple components of risk with distinct longitudinal profiles. To analyse longitudinal patterns of genetic risk, we use the proportional hazards model to estimate genetics effect sizes within age groups, conditioning on survival (i.e. no mortality, censoring or disease). We show that this approach is needed to avoid biases that arise in naive GWAS approaches, which are affected by the depletion of risk alleles in unaffected individuals over time and changes in baseline risk. We apply this model to the UK Biobank dataset, analysing 23 ICD-10 disease codes with prevalence > 1% and at least 20 independent associated variants. We use a Bayesian clustering approach on summary statistics to estimate latent curves and their posterior distributions, using spline functions to encourage smoothness in risk profiles over age and permutation tests to assess the evidence for distinct groups of variants with different age- related profiles. We identify 10 diseases with evidence for age-specific heterogeneity, including heart disease, skin cancer and gall-bladder diseases, several of which show evidence for more than one curve. We discuss biological processes that can result in such age-specific risk, notably gene-environment interactions, and the implications of these results for genetic prediction of risk.

kernelPSI: a powerful post-selection inference framework for nonlinear association testing in genome-wide association studies (Slim)

Slim Lotfi <lotfi.slim@mines-paristech.fr> (1) (2), Chatelain Clément <Clement.Chatelain@sanofi.com> (1), Azencott Chloé-Agathe <Chloe-agathe.Azencott@mines- paristech.fr> (2) (3), Vert Jean-Philippe <jpvert@google.com> (4) (2) 1 - Translational Sciences (France), 2 - CBIO - Centre for Computational Biology (France), 3 - Institut Curie (France), 4 - Google Brain (France)

We present the results of an extensive study, in which we demonstrate the use of kernelPSI[1] for genome-wide association studies (GWAS). kernelPSI is a statistical tool to perform post-selection inference (PSI) for nonlinear variable selection. The nonlinearity is modeled through quadratic kernel association scores, which are a quadratic form of the response vector. The latter scores allow the incorporation of nonlinear effects and interactions among covariates. In the context of GWAS, kernelPSI assesses the effect of a predetermined genomic region e.g. gene, or regulatory region, while simultaneously identifying the causal loci within. This can facilitate the downstream biological interpretation. Notably, the identification of causal loci is a major limitation of the sequence kernel association test[2] (SKAT), a state-of-the-art method for association testing of genomic regions. Moreover, in kernelPSI, we generalize the SKAT statistic to a broader family of association scores, hence providing a general and flexible framework to measure the association between a given locus and a phenotype of interest. Another added benefit of our framework in

comparison to SKAT is its greater statistical power, as shown in several experimental settings[1]. kernelPSI is a two-step approach: a number of putative loci are selected in a supervised manner in the first step, and their aggregated effect on the phenotype is tested in the second step. The selection step introduces a statistical bias in the subsequent hypothesis testing. For instance, if the most associated loci are selected in the first step, the significance of their overall effect is likely to be overestimated. To answer this problem, we develop a PSI methodology in order to derive valid empirical p-values. This is achieved thanks to a constrained sampling of replicates of the response vector. We then compare the statistics of the response to those of the replicates to obtain the desired p-values. In our case study, we apply kernelPSI to a set of obesity- related phenotypes such as body mass index (BMI), weight and fat distributions. Our selection of such phenotypes was motivated by the breadth of information available in large biobanks. Yet, kernelPSI can be applied to any continuous response. The theoretical foundations of kernelPSI have been published in a previous work[1], in addition to an eponymous R package[3] which implements our PSI framework with different association scores. [1] Slim, L., Chatelain, C., Azencott, C.-A., & Vert, J.-P. (2019). kernelPSI: a Post-Selection Inference Framework for Nonlinear Variable Selection. In K. Chaudhuri & R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning (Vol. 97, pp. 5857-5865). Long Beach, California, USA: PMLR. [2] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics, 89(1), 82-93. [3] Slim, L. (2019). kernelPSI: Post-Selection Inference for Nonlinear Variable Selection. Retrieved from https://cran.r-project.org/package=kernelPSI

Ongoing purifying and overdominant selection in the human genome (Wei)

Wei Xinzhu <aprilwei@berkeley.edu> (1), Nielsen Rasmu <rasmus_nielsen@berkeley.edu>, Pan Ziqing <panzq@berkeley.edu> 1 - UC Berkeley (United States)

The release of 500,000 genomes from the UK Biobank (UKB) provides unprecedented opportunities for studying ongoing selection using deviations from Hardy-Weinberg Equilibrium (HWE). Compared to the HWE expectation, we observe increased heterozygosity in British ancestry individuals in the UKB cohort genome-wide. Genotyping errors could potentially cause this observation. However, we find that nonsynonymous SNPs with low Minor Allele Frequency (MAF <5%) are enriched for excess heterozygosity compared to other SNPs of the same MAF. This observation cannot be explained by genotyping errors, but is likely due to purifying selection against recessive deleterious alleles. Perhaps surprisingly, low MAF archaic SNPs from Neanderthal admixture show depletion of excess heterozygosity, a pattern that can be replicated in simulations incorporating selection and addition, by mutation, of new deleterious alleles, for many generations after the time of admixture. We also find that high MAF (30-50%) nonsynonymous SNPs are enriched for excess heterozygosity compared to other SNPs of the same MAF. These SNPs are enriched for genes previously identified to be under balancing selection. They also show evidence of overdominance from decreased all-cause mortality in heterozygous individuals. Population genetic simulations under realistic parameter settings can recapitulate these observations. Our study demonstrates that analyzing patterns of deviations from HWE can be a powerful way to detect selection in large cohorts and that ongoing selection in humans is common.

Inferring human biology from genetic association (Mc Vean)

Gil McVean, Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Genomics plc

Patterns of genetic association have revealed much about the biology underlying human traits and complex diseases. But how can we use such information systematically to learn about the processes - at molecular, cellular and tissue levels - that modulate risk? I will discuss some challenges, approaches, and solutions to the problems of integrating and interpreting data on such a vast scale. And how such information can be applied to diverse problems ranging from therapeutic target identification to quantifying individual risk.

Differential complex trait architecture across humans: epistasis identified in non-European populations at multiple genomic scales (Turchin)

Turchin Michael <michael_turchin@brown.edu> (1) (2), Ting Isabella <isabella_ting@brown.edu> (3) (2), Crawford Lorin <lorin_crawford@brown.edu> (4) (5) (2), Ramachandran Sohini <sramachandran@brown.edu> (1) (2) 1 - Department of Ecology and Evolutionary Biology, Brown University, Providence, RI (United States), 2 - Center for Computational Molecular Biology, Brown University, Providence, RI (United States), 3 - Department of Computer Science, Brown University, Providence, RI (United States), 4 - Center for Statistical Science, Brown University, Providence, RI (United States), 5 - Department of Biostatistics, Brown University, Providence, RI (United States)

Genome-wide association (GWA) studies have identified thousands of significant genetic associations in humans across a number of complex traits. However, the vast majority of these studies use datasets of predominantly European ancestry (Popejoy & Fullerton 2016). It has generally been thought that complex trait genetic architecture should be transferable across populations of different ancestries, but recent work has shown a number of differences in trait architecture across human ancestries, including heterogeneity in both the identified causal variants and estimated effect sizes (Martin et al. 2017, Wojcik et al. 2017). Here, we report further evidence that complex trait genetic architecture is fundamentally different among human ancestries by jointly leveraging pathway and epistasis analysis. Under the assumption that a given complex trait may have differential polygenic architectures across human ancestries, we hypothesize that human populations may also be enriched for differences in epistatic effects. However, since polygenic traits tend to have smaller GWA effect sizes, combining variants via pathway analysis may allow us to better reveal these signals. To accomplish this, we extend the concept of identifying marginal epistasis, moving from testing single variants (Crawford et al. 2017) to testing groups of variants for nonlinear association with a trait of interest. We apply our new method to multiple ancestries present in the UK Biobank (Sudlow et al. 2015) and explore multiple pathway-related interaction models. Using morphometric traits we find evidence for genome-wide epistasis in African and other non-European populations. We also find evidence that these trends exists on the SNP and gene levels as well. Results also indicate this may be due to increased heterozygosity in non-European populations. This suggests that non-European populations may be well-suited for identifying non-additive effects in human complex trait architecture; this also suggests further evidence that European populations -- predominantly used for epistasis studies -- may indeed be limited and inaccurate proxies for all human ancestries in complex trait research.

The effects of pleiotropy on polygenic adaptation (Hayward)

Hayward Laura <lauhayward@gmail.com> (1), Sella Guy <gs2747@columbia.edu> (1) 1 - Columbia University (United States)

Strong evidence suggests that adaptive changes to complex, quantitative traits, or polygenic adaptation, should be ubiquitous in many species including humans. Yet this mode of adaptation is still poorly understood, making for a substantial gap in evolutionary theory and limiting our ability to identify its footprints. Many complex, quantitative traits are thought to be under long-term stabilizing selection, with intermittent shifts in their optimal values; additionally, genetic variation that affects one trait often affects many others. In past work, we described the phenotypic and genetic response after a sudden change of environment induces an instantaneous shift in the optimum of a single trait under stabilizing selection. Here we generalize this work to account for the effects of pleiotropy after an instantaneous shift of a phenotypic optimum in an n-dimensional trait space. We find that the phenotypic dynamic is similar to the single trait case. Immediately after the shift, the mean phenotype approaches the new optimum rapidly, at a rate that was well approximated by Lande (1976). When most genetic variance before the shift is dominated by alleles of small and moderate effects (measured by a^2 - their squared effect size in the n-dimensional trait space), the Lande approximation holds more generally; but when large effect alleles contribute markedly to genetic variance, the dynamic can be more complex. Similar to the single trait case, we find that during the initial phase of rapid phenotypic adaptation moderate and large effect alleles have a similar contribution to the movement of the mean phenotype. During the period of equilibration that follows, the mean phenotype changes little, but the alleles that underlie the change exhibit turnover. The contribution of

moderate effect alleles largely supplants that of larger effect alleles. The multi-dimensional case differs from the single trait case in that an allele's response also depends on its projected effect in the direction of the shift. We measure this projection by r^2 - the ratio of its squared effect in that direction and its squared size in the n-dimensional trait space (r^2=1 in the single trait case). We derive a functional form for the dependence of allelic frequency change and contribution to phenotypic change as a function of the projection r^2. In lower dimensions, alleles with larger projections (r^2) contribute more to polygenic adaptation, but when the dimension is sufficiently high, both the proportion of both short- and long-term phenotypic change contributed by alleles with ratio r^2 follows a chi-squared distribution with one degree of freedom. In particular, the maximal contribution to phenotypic change arises from alleles whose squared effect in the direction of the shift equals its average squared effect in any direction (r^2=1).

Quantifying the Pre-Tumour and Tumour Evolutionary Processes From High Coverage Sequencing Data (Ang)

Ang Houle Armande <armande.anghoule@oicr.on.ca> (1) (2), Skead Kimberly <kimberly.skead@oicr.on.ca>, Lin Boxi <boxi.lin@oicr.on.ca>, Uzunovic Jasmina <jasmina.uzunovic@oicr.on.ca>, Denroche Rob <rob.denroche@oicr.on.ca>, Mehanna Pamela <pamela.mehanna@oicr.on.ca>, Agbessi Mawusse <mawusse.agbessi@oicr.on.ca>, Bruat Vanessa <vanessa.bruat@oicr.on.ca>, Boutros Paul <paul.boutros@oicr.on.ca>, Notta Faiyaz <faiyaz.notta@oicr.on.ca>, Wright Stephen <stephen.wright@utoronto.ca>, Stein Lincoln <lincoln.stein@oicr.on.ca>, Awadalla Philip philip.awadalla@oicr.on.ca 1 - Department of Molecular Genetics [Toronto] (Canada), 2 - Ontario Institute for Cancer Research [Canada] (Canada)

Cancer progression is an evolutionary process: somatic mutations can confer selective advantages to certain cells, eventually causing them to abnormally proliferate. Important aspects to consider when studying the evolutionary processes of cancer include the accurate characterization of somatic mutations. While positively selected alleles may lead to selective sweeps within the cancer population, and reduce the diversity of somatic variation, damaging mutations may occur and be swept along during the process. The signature of the selective sweeps may hinder the detection of negative selection within the cancer celL population. Our ability to capture these evolutionary signatures is dependent on depth of sequence coverage per tumour within a single patient. Average sequencing depths are often insufficient to detect mutations found at low variant allele frequency within a population of cancer cells, and sequencing at higher depths of coverage allows for a more complete picture of the full evolutionary spectrum. Here, we aim to accurately identify evolutionary processes in 245 cancer and pre-cancer samples sequenced with a minimal coverage of 100x from 5 cancer types. Using the read depth of somatic mutations to quantify the allelic frequency of a mutation within each population of cancer cells, we employ an Approximate Bayesian Computation framework to quantify the proportions of beneficial and deleterious mutations, and the timing of driver and passenger events throughout the life of a cancer. By applying these models and stratifying the mutational spectrum within patient samples, we capture non-neutral evolutionary processes previously undetected in tumours. We capture varying rates of damaging as well as beneficial mutations accumulating across patients within many tumour types. 40 samples overall show the presence of a combination of the accumulation of deleterious and beneficial mutations. Among tumours that depart from neutrality, we always capture a signature of negative selection, outlining the contribution of the accumulation of deleterious mutations in cancer. In a third of Barrett's esophagus biopsies, models of negative selection are the best fitting model which implicates that these samples are unlikely to escalate towards more severe carcinogenesis. In samples where the presence of deleterious selection is most likely, we observe genes having enrichments of damaging mutations which could be alternative targets of therapy. Our work demonstrates the heterogeneity of evolutionary processes within a cancer type, and further outlines the possibility of dubious inferences of cancer evolutionary processes caused by low sequencing depth.

Bayesian deconvolution of somatic clones and pooled individuals with expressed variants in single cells (Huang)

Huang Yuanhua <yuanhua@ebi.ac.uk> (1), Mccarthy Davis <dmccarthy@svi.edu.au> (2), Rostom Raghd <rr415@cam.ac.uk> (3), Teichmann Sarah <st9@sanger.ac.uk> (3), Stegle Oliver

<oliver.stegle@ebi.ac.uk> (1) 1 - European Bioinformatics Institute [Hinxton] (United Kingdom), 2 - St Vincent's Institute of Medical Research (Australia), 3 - Sanger Institute (United Kingdom)

Decoding the clonal substructures of somatic tissues sheds light on cell growth, development and differentiation in health, ageing and disease. However, approaches to systematically characterize phenotypic and functional variations between individual clones are not established. Here we present cardelino (https://github.com/PMBio/cardelino), a Bayesian method for inferring the clonal tree configuration and the identity of individual cells by modelling the expressed variants in single-cell RNA- seq (scRNA-seq) data. Critically, cardelino allows effective integration of information from imperfect clonal tree inferences based on bulk exome-seq data, and sparse variant alleles expressed in scRNA-seq data. Simulations validate the accuracy of our model and its robustness to the errors in the guide clone configuration. We applied cardelino to 32 human dermal fibroblast lines, identifying hundreds of differentially expressed genes between cells from different somatic clones. These genes are frequently enriched for cell cycle and proliferation pathways, indicating a key role for cell division genes in non- neutral somatic evolution. A similar problem is demultiplexing cells from pooled scRNA-seq experiments by using common genetic (similar to somatic) variants, a challenging task when genotype reference is not available. Here, we modified cardelino model and introduce a variational inference method (named Vireo), to efficiently and accurately demultiplex data from pooled experimental designs, supporting with partial or without any genotype information of the pooled samples.

Probabilistic approaches to inference of mutation rate and selection in cancer (Weghorn)

Weghorn Donate <dweghorn@crg.eu> (1) (2), Dietlein Felix <dietlein@broadinstitute.org> (3), Van Allen Eliezer <eliezer@broadinstitute.org> (3), Sunyaev Shamil <ssunyaev@rics.bwh.harvard.edu> (4) 1 - Universitat Pompeu Fabra [Barcelona] (Spain), 2 - Centre for Genomic Regulation, Barcelona (Spain), 3 - Department of Medical Oncology, Dana-Farber Cancer Institute, Boston (United States), 4 - Department of Biomedical Informatics, Harvard Medical School, Boston (United States)

Cancer is a highly complex system that evolves asexually under high mutation rates and strong selective pressures. Cancer genomics efforts have identified genes and regulatory elements driving cancer development and neoplastic progression. The detection of both significantly mutated (positive selection) and undermutated (negative selection) genes is completely confounded by the genomic heterogeneity of the cancer mutation rate. Here, I present an approach we recently developed in order to address mutation rate heterogeneity to increase the power and accuracy of selection inference. Using a hierarchical model, we infer the distribution of mutation rates across genes that underlies the observed distribution of the synonymous mutation count within a given cancer type. This enables the inference of the probability of nonsynonymous mutations under neutrality without additional parameters, however explicitly taking into account cancer-type-specific mutational signatures, which are known to be highly distinct. In addition to detecting an excess in the total number of mutations, we then augmented our test through integrating information at the single-nucleotide level, exposing a ‘selection mutational signature’. Based on a model that accounts for the extended sequence context (>5-mers) around mutated sites, this second component of the test identifies genes with an excess of mutations in unusual nucleotide contexts, which deviate from the characteristic context around neutrally evolving passenger mutations. I will show that the inclusion of this context test increases power to detect cancer driver genes particularly when the fraction of selected nucleotides on a gene is small. Using the combined test, we discovered a catalogue of well-known cancer driver genes as well as a long tail of novel candidate cancer genes with mutation frequencies as low as 1% and functional supporting evidence.

Comprehensive estimation of the tissue of origin of circulating cell-free DNA (Caggiano)

Caggiano Christa <christa@ucla.edu> (1), Celona Barbara <barbara.celona@ucsf.edu> (2), Garton Fleur <f.garton@imb.uq.edu.au> (3), Black Brian <brian.black@ucsf.edu> (4), Wray Naomi, Dahl Andy <andy.dahl@ucsf.edu> (5), Zaitlen Noah, <n.wray@imb.uq.edu> (3), <nzaitlen@mednet.ucla.edu> (5) 1 - Bioinformatics Program, University of California, Los Angeles (United States), 2 - Dept of Cardiology, University of California, San Francisco (United States), 3 - Institute of Molecular Biosciences, University of

Queensland (Australia), 4 - Dept of Cardiology, University of California, San Francisco (United States), 5 - Dept of Neurology, University of California, Los Angeles (United States)

Cell-free DNA (cfDNA) in the bloodstream originates from dying tissues. The analysis of cfDNA provides a non-invasive biomarker for diseases characterized by tissue-specific cell death. The purpose of this work is to create a statistical model that can accurately estimate which tissues are contributing to the presence of cfDNA in the blood. To do this, we leverage the distinct DNA methylation profile of each tissue type throughout the body and use this information to estimate the contribution of each of these tissues to the cfDNA mixture. Decomposing these mixtures, however, is difficult, as cfDNA of a disease-relevant tissue may only be present in the blood only in small amounts. Furthermore, many DNA methylation datasets, such as those from the ENCODE or ROADMAP projects, are whole genome bisulfite sequencing (WGBS), which are generally of low read depth (~10x). This low read depth means that methylation count data may be missing at a given DNA methylation site (CpG), or observed so infrequently as to be unreliable. Finally, accurately decomposing cfDNA mixtures requires a robust understanding of all possible tissue types that could potentially contribute to the mixture. This robust reference, however, is nearly impossible to assemble, as there are hundreds of distinct tissue types, and because the methylation state for a CpG in one tissue can vary. We developed an EM algorithm that estimates tissue type proportion from both WGBS cfDNA input and tissue reference data. Notably, our algorithm can handle missing and low count data and does not rely on CpG site curation. Our EM algorithm can also estimate an arbitrary number of ‚unknown' tissue type categories. We show in simulations that our algorithm can accurately estimate tissue of origin of cfDNA mixtures. Simulations also demonstrate that we can effectively estimate cfDNA originating from ‚rare' cell types. We also apply our EM algorithm to cfDNA from ALS patients. Our EM algorithm can detect differences between the tissue of origin of cfDNA from ALS patients and from healthy controls, illustrating that our algorithm can potentially identify clinical biomarkers for complex human diseases.

Identifying novel regulatory elements using RELICS, a statistical framework for the analysis of CRISPR regulatory screens (Fiaux)

Fiaux Patrick <pfiaux@ucsd.edu> (1) (2), Chen Hsiuyi <hschen@salk.edu> (1), Luthra Ishika <iluthra@salk.edu> (3), O'connor Carolyn <coconnor@salk.edu> (1), Mcvicker Graham <gmcvicker@salk.edu> (1) 1 - Salk Institute for Biological Studies (United States), 2 - University of California [San Diego] (United States), 3 - Simon Fraser University (Canada)

High-throughput CRISPR/Cas9 screens are a powerful new tool for the systematic discovery of regulatory elements in the human genome. In these regulatory screens, thousands of guide RNAs (gRNAs) are delivered to cells to target potential regulatory sequences for mutation, activation or inhibition. The cells are then sorted into high- and low-expression pools based on the expression of a target gene. While, these screens have the potential to perform unbiased discovery of regulatory elements, they generate noisy data and the performance of analysis methods has not been rigorously assessed. Here we describe RELICS, a statistical framework for Regulatory Element Identification in CRISPR Screens. RELICS models the observed guide counts in different expression pools with a generalized linear mixed model. This approach is very flexible, can jointly model multiple expression pools (beyond just high and low), incorporate variability across guides, and accommodate over- dispersion. To assess the performance of RELICS we have developed a simulation framework for generating CRISPR regulatory screen data and simulated 1000s of data sets under a wide variety of experimental and biological conditions. RELICS outperforms existing analysis methods on the simulated data and we have applied it to identify regulatory elements in several published datasets. In addition, we have applied RELICS to data from a paired-guide regulatory screen that we performed for GATA3 in Jurkat T cells. We identify a total of 23 putative regulatory elements within the 2MB targeted region surrounding GATA3. Notably 16 of the identified elements lie within the same topological associating domain as GATA3, but only 3 overlap enhancers predicted by ChromHMM.

Identification of complex methylation changes across developmental lineages using single-cell multi-omics (Frank)

Frank Max <max.frank@embl.de> (1) (2), Imaz-Rosshandler Ivan <iimaz@ebi.ac.uk> (3), Argelaguet Ricard <ricard@ebi.ac.uk> (3), Marioni John <marioni@ebi.ac.uk> (4) (5), Stegle Oliver <oliver.stegle@ebi.ac.uk> (3) (1) (2) 1 - European Molecular Biology Laboratory [Heidelberg] (Germany), 2 - German Cancer Research Center - Deutsches Krebsforschungszentrum [Heidelberg] (Germany), 3 - European Bioinformatics Institute [Hinxton] (United Kingdom), 4 - European Bioinformatics Institute [Hinxton] (United Kingdom), 5 - Cancer Research UK (United Kingdom)

The genomic sequence of an organism is nearly identical in all its cells and over its lifetime. Epigenomic marks, however, such as DNA-methylation are subject to drastic changes across different tissues and over the course of organism development. Recent technological advances, such as scNMT- seq[1], have made it possible to probe DNA-methylation as well as chromatin accessibility and transcriptome in the same cell. This offers unique opportunities to gain insight into mechanisms by which the epigenome shapes gene expression and influences cell fate. However, the analysis of these datasets poses major challenges: Firstly, a smaller number of cells can be assayed per experiment compared to conventional single-cell RNAseq. Secondly, a lower fraction of methylation sites is typically covered compared to bulk DNA-methylation measurements. This means that classical methods to test DNA- methylation and chromatin accessibility differences are underpowered to detect subtle changes between tissues or over the course of biological processes. Furthermore, most current tests are only able to detect differences between discrete and pre-defined cell populations, whereas single cell approaches allow for studying continuous processes such as cell differentiation and cell lineage development. To address this, we here present a two-step analysis approach: First, we align scarce scNMT measurements to large reference atlas maps based on conventional single-cell RNAseq. This allows us to precisely position cells along their developmental trajectories and characterize their cell type. Second, building on this reference mapping, we assess local changes in DNA methylation patterns across cells using a Gaussian process model that shares information between cells and neighbouring methylation sites. For each genomic region of interest, we train a model to describe methylation events, given their genomic position and every cell's mapped position in the reference atlas. By performing comparisons between models that account for different cellular dimensions extracted from the reference atlas and null models without atlas coordinates, the model allows to probe for lineage and cell-type specific DNA methylation changes. Importantly, this test does not rely on clustering of cell types or differentiation stages, which reflects the dynamic and continuous changes that cells undergo in-vivo. We make use of state of the art stochastic variational methods and employ a Bernoulli likelihood for the observed methylation data, allowing us to scale Gaussian process inference to large single-cell datasets To demonstrate the approach, we apply it to scNMT measurements of >1000 mouse embryonic stem cells and integrate it with a large single-cell RNAseq gastrulation atlas of 116.312 cells [2]. We pinpoint promoter and enhancer regions that show continuous methylation changes along key developmental lineages and across germ layers. Furthermore, our model can be used to accurately impute methylation events aiding in downstream analyses of the data, such as prediction of gene expression. Testing incremental changes of multiple molecular layers in parallel will become increasingly important with the rise of single-cell multi-omics studies. We present a modelling framework that can aid in the interpretation of such complex datasets. References: [1] Clark et al., Nature Communications(2018) [2] Pijuan-Sala, Nature(2019)

Characterizing chromatin landscape from aggregate and single-cell genomic assays using flexible duration modeling (Gabitto)

Gabitto Mariano <mgabitto@simonsfoundation.org> (1) 1 - Simons Foundation (United States)

ATAC-seq has become a leading technology for probing the chromatin landscape of single and aggregated cells. Distilling functional regions from ATAC-seq and other similar genomic technologies presents diverse analysis challenges, due to the relative sparseness of the data produced and the interaction of complex noise with multiple chromatin structure scales. Methods commonly used to analyze chromatin accessibility datasets are adapted from algorithms designed to process different experimental technologies, disregarding the statistical and biological differences intrinsic to the ATAC- seq technology. Here, we

present a Bayesian statistical approach that uses Hidden Semi-Markov models to better model the duration of functional and accessible regions, termed ChromA. We demonstrate the method on multiple genomic technologies, with a focus on ATAC-seq data. ChromA annotates the cellular epigenetic landscape by integrating information from replicates, producing a consensus de- noised annotation of chromatin accessibility. ChromA can analyze single cell ATAC-seq data, improving cell type identification and correcting many biases generated by the sparse sampling inherent in single cell technologies. We validate ChromA on multiple technologies and biological systems, including mouse and human immune cells and find it effective at recovering accessible chromatin, establishing ChromA as a top performing general platform for mapping the chromatin landscape in different cellular populations from diverse experimental designs. We will also discuss new work, not present in the early preprint to extend this model to CRISPR screens, single cell cut&run, DNA methylation and other genomic technologies aimed at chromatin state and function.

Bayesian nonparametric integration of multiple single-cell RNA-seq experiments (Archit)

Verma Archit <architv@princeton.edu> (1), Engelhardt Barbara <bee@princeton.edu> (1) 1 - Computer Science Department [Princeton] (United States)

Joint analysis of multiple single cell RNA-seq (scRNA-seq) data is often confounded by batch effects across experiments. Manifold alignment is an effective tool for representing and integrating multiple data sets to control the gene expression levels for confounding due to batch. A complete data integration procedure should provide: 1) mapping between high and low dimensional spaces that removes variation due to batch; 2) uncertainty estimates in the alignment of nonlinear manifolds; 3) reference-free regularization that preserves variation from sources other than batch; and 4) robust alignment when large portions of the subspaces are not shared. Current methods such as Seurat's Canonical Correlation Analysis and Mutual Nearest Neighbors are limited on one or more of these conditions. We propose manifold alignment with semi-supervised Gaussian Process Latent Variable Models (GPLVMs). We use GPLVMs with both unknown latent positions and known batch ‘fixed variables’ that are mapped to the high dimensional observation space of approximately 20,000 genes using Gaussian processes (GPs). We use a GPLVM with robust Student's t-distributed error and non- smooth Matern kernels to deal with the sparse, heavy-tailed nature of scRNA-seq observations. We demonstrate that our model meets all four conditions with simulated data and pancreas expression data from four different sequencing technologies. We use our method to correct for batch effect across scRNA-seq samples from 54 Yoruban individuals to discover cell type specific single cell expression quantitative trait loci (eQTLs). We also are able to recover the precise effects of batch and other fixed covariates on each gene using our model.

Phylogenetic modeling of epigenetic mark turnover uncovers genomic features that drive cis- regulatory evolution (Dukler)

Dukler Noah <ndukler@cshl.edu> (1), Huang Yi-Fei <yuh371@psu.edu> (2), Siepel Adam <asiepel@cshl.edu> (1) 1 - Simons Center for Quantitative Biology [Cold Spring Harbor] (United States), 2 - Huck Institutes of the Life Sciences [University Park] (United States)

With the proliferation of genome-wide functional assays (e.g. ChIP-seq, ATAC-seq, PRO-seq, etc.), there has been an explosion in the amount of available epigenetic data within and between species. Comparative analysis of epigenomic data provides the opportunity to improve our understanding of the evolution of regulatory elements. A wide variety of phylogenetic tools have been developed to study sequence evolution and have been used to identify loci associated with disease, development, and molecular phenotypes. However, there has been much less development of analogous tools to address the unique challenges of studying the evolution of the epigenome. Here, we introduce two new tools for evolutionary epigenomics, epiPhylo and phyloGLM, and apply them to histone mark data from the placental mammals to interrogate which features of gene expression, function, and cis-regulatory architecture influence cis-regulatory evolution. EpiPhylo combines a phylo-HMM with a negative binomial error model to jointly infer the location of cis-regulatory elements (CREs) and their evolutionary trajectories from noisy epigenomic data. We apply epiPhylo to previously published H3K27Ac data for nine placental mammals to call CREs, then test for

groups of CREs showing potential signatures of selection using phyloGLM. PhyloGLM uses a phylogenetic likelihood that describes the observed patterns of CRE gain and loss to jointly estimate the effects of a large number of CRE features on evolution rate. We investigate the effects of various genetic covariates (e.g. tissue specific expression, cross-tissue expression, and number of enhancers associated with a gene) for putative enhancers and promoters on turnover rate, and find significant associations for features related to both gene expression and cis- regulatory architecture, while observing differences between enhancers and promoters that may arise from differences in regulatory redundancy. Our results support the importance of both cross-tissue pleiotropy and enhancer redundancy in CRE evolution. Furthermore, we observe a divergent relationship between sequence and epigenetic conservation for two functional categories: transcriptional regulation and metabolism, which are classically used as proxies for dosage sensitivity. We propose that dosage sensitivity of target genes can partially explain the discrepancy between sequence and histone mark turnover rates of associated CREs. Together, epiPhylo and phyloGLM provide a rigorous phylogenetic framework for exploring explicit hypotheses about relationships between genomic context and trait evolution.

Exchangeable Variational Autoencoders for Genomic Data (Chan)

Chan Jeffrey <chanjed@berkeley.edu> (1), Spence Jeffrey <spence.jeffrey@berkeley.edu>, Song Yun <yss@berkeley.edu> 1 - UC Berkeley (United States)

Exchangeable-structured data is ubiquitous in biology and genomics. Common examples of exchangeable data include gene networks, single-cell data, and read sequence data. Outside of biology, exchangeable data can be seen in graphs, point clouds, and bootstrapped data. The prevalence of exchangeable data demands the development of methods which can leverage the permutation- invariance and conditionally i.i.d. nature of exchangeability to answer scientific questions. While recent work has adapted advances in machine learning towards exchangeable genomic data, little has been done in the context of Bayesian probabilistic modeling. In this work, we focus on the variational autoencoder (VAE) which marries both the Bayesian probabilistic modeling framework with deep learning to enable unsupervised learning. Recent work in biology has exploded with applications of the variational autoencoders including fields such as single-cell RNA-seq, drug response, and protein sequences. We develop an exchangeable variational autoencoder and showcase its efficacy in two key biological settings. First, we show how accounting for exchangeability in exchangeable VAEs can improve single-cell RNA-seq methods by directly accounting for the uncertainty in transcript abundances allowing for more precise downstream inferences. Much of the variance in the transcript abundances can be attributed to the alignment process, so using the exchangeable VAE to accurately account for the variance from multiple bootstraps of the alignment procedure greatly denoises the technical variation to enable better downstream analyses. Second, we show how our exchangeable VAE can be leveraged to improve variant calling in a completely unsupervised way which enables us to develop an accurate variant caller while also being able to utilize the trained generative model as a read simulator. Read simulators represent a great need in the genomics community as a high-quality read simulator would remove the data bottleneck of training methods on limited-amounts of experimental read data for a given problem setting. An improved variant caller that is comparable with that of GATK and DeepVariant in the supervised setting but also can be used in an unsupervised fashion would be useful for reference- free populations.

Can mutation-selection-drift balance on modifiers of mutation explain variation in mutation rates among human populations? (Milligan)

Milligan William <wm2377@columbia.edu> (1), Sella Guy <gs2747@columbia.edu> (1) 1 - Columbia University (United States)

Recent studies have shown mutation rates varied among human continental populations and hypothesized this resulted from genetic variation at loci that affect mutation rates, e.g. at genes involved in correcting replication or damage induced mutations. Here we demonstrate this hypothesis is not plausible. We assume

mutations at modifier loci - mutator alleles - arise a rate that depends on some target size and increase the genome-wide mutation rate. We further assume these alleles are selected against because individuals that carry them also carry a greater burden of deleterious, germline mutations (at other loci). We rely on a diffusion approximation to calculate the expected mean and variance of the mutation rate and validate our approximations against simulations. We find both quantities depend primarily on the target size at modifier loci and on the expected reduction in fitness per generation, s, caused by a mutator allele. We show that per unit target size, trade-offs in frequency and effect constrain mutator alleles, such that weakly selected weakly selected mutator alleles (2Ns~1) have the greatest impact on mutation rates. Under plausible assumptions about the target size affecting particular types of mutations, a single modifier locus would generate only minute variation in mutation rates. However, the cumulative effects of many modifier loci can be substantial. We then demonstrate modifier loci are unlikely to generate the observed mutation rate differences by mimicking the demographics settings in which human mutation rate variation arose. Specifically, we incorporate estimates of the split time between Europeans and Africans, as well as their past changes in effective population sizes. We simulate the allelic dynamics at modifier loci forward in time and use coalescent simulations to mimic the samples in which the variation in mutation rates were measured (plausibly assuming the bulk of polymorphic sites used in these measurements were neutral). We find genetic variation at modifier loci is unlikely to explain the observed variation in human mutation rates due to constraints imposed by selection and correlation in mutator allele frequencies between recently diverged populations. We also show incorporating additional selection against mutator alleles for their effect on somatic mutation rates and/or LD with selected loci would further reduce their impact on variation in mutation rates. Thus, our results support alternative explanations for the variation observed among human populations, e.g., that they resulted from changes in life history traits or in the environment that affect mutation rates.

The genomic view of diversification (Lambert)

Marin Julie <julie.marin@college-de-france.fr> (1), Achaz Guillaume <guillaume.achaz@mnhn.fr> (2) (1), Crombach Anton <anton.crombach@inria.fr> (3), Lambert Amaury <amaury.lambert@college-de-france.fr> (4) (1) 1 - Centre interdisciplinaire de recherche en biologie (France), 2 - Atelier de BioInformatique (France), 3 - INRIA (France), 4 - Laboratoire de Probabilités, Statistique et Modélisation (France)

The standard way of coupling discordant gene trees is to postulate the existence of a unique species tree where disagreements between gene trees are explained by incomplete lineage sorting (ILS) due to random coalescences of gene lineages inside the edges of the species tree. This paradigm, known as the multi-species coalescent (MSC), is constantly violated by the ubiquitous presence of gene flow revealed by empirical studies, leading to topological incongruences of gene trees that cannot be explained by ILS alone. I will argue that this paradigm should be revised in favor of a vision acknowledging the importance of gene flow and where gene histories shape the species tree rather than the opposite. We propose a new, plastic framework for modeling the joint evolution of gene and species lineages relaxing the hierarchy between the species tree and gene trees. We implement this framework in a probabilistic model called the gene-based diversification model based on coalescent theory, with four parameters tuning colonization, mutation, gene flow and reproductive isolation. We propose a fast estimation method based on the differences between gene trees and use it to evaluate the amount of gene flow in two empirical data-sets. This work should pave the way for approaches of diversification using the richer signal contained in genomic evolutionary histories rather than in the mere species tree.

Inference of ancient whole genome duplications and the evolution of the gene duplication and loss rate (Zwaenepoel)

Arthur Zwaenepoel, Yves Van de Peer VIB-UGent Center for Plant Systems Biology, Ghent University, Belgium

Gene tree - species tree reconciliation methods have been employed for studying ancient whole genome duplication (WGD) events across the eukaryotic tree of life. Most approaches have relied on using maximum likelihood trees and the maximum parsimony reconciliation thereof to count duplication events on specific branches of interest in a reference species tree. Such approaches do not account for uncertainty in the gene tree and reconciliation, or do so only heuristically. The effects of these simplifications on the inference of ancient WGDs are unclear. In particular the effects of variation in gene duplication and loss rates across the species tree have not been considered. Here, we developed a fully probabilistic approach for phylogenomic WGD inference, which allows to formally assess the statistical support for WGD hypotheses from alignments of multi-copy gene families while accounting for both gene tree and reconciliation uncertainty using a method based on the principle of amalgamated likelihood estimation. Both Maximum likelihood and Bayesian methods were implemented for inference under the model. Importantly, we implemented methods to estimate variation of duplication and loss rate across the species tree, using methods inspired by phylogenetic divergence time estimation. We applied our newly developed framework to ancient WGDs in land plants and investigate the effects of duplication and loss rate variation on reconciliation and gene count based assessment of these earlier proposed WGDs. Overall, our study provides another example of the power of probabilistic models of gene family evolution in evolutionary genomics.

The Cumulative Indel Model: fast and accurate statistical evolutionary alignment (De Maio)

De Maio Nicola <demaio@ebi.ac.uk> (1) 1 - European Bioinformatics Institute [Hinxton] (United Kingdom)

Sequence alignment is essential for phylogenetic and molecular evolution inference, as well as in many other areas of bioinformatics and evolutionary biology. Inaccurate alignments can lead to severe biases in most downstream statistical analyses. Statistical alignment based on probabilistic models of sequence evolution addresses these issues by replacing heuristic score functions with evolutionary model-based probabilities. However, score-based aligners and fixed-alignment phylogenetic approaches are still more prevalent than methods based on evolutionary indel models, mostly due to computational convenience. Here, I present new techniques for improving the accuracy and speed of statistical evolutionary alignment. The «cumulative indel model» approximates realistic evolutionary indel dynamics using differential equations. «Adaptive banding» reduces the computational demand of most alignment algorithms without requiring prior knowledge of divergence levels or pseudo-optimal alignments. Using simulations, I show that these methods lead to faster and more accurate pairwise alignment inference. The cumulative indel model and adaptive banding can therefore improve the performance of alignment and phylogenetic methods.

Posters

1 Why is diversity so low within the species? (Achaz)

Achaz Guillaume <guillaume.achaz@mnhn.fr> (1) (2) (3) Affiliations: 1 - Atelier de BioInformatique (France), 2 - Centre interdisciplinaire de recherche en biologie (France), 3 - Musee de l'Homme (France)

A basic prediction of the standard neutral model is that molecular diversity should scale linearly with the population size. However, there are strong evidences that diversity (1) is always many orders of magnitude lower from what is expected from the census size and (2) does not scale linearly with the species abundance. This strong discrepancy seems hardly compatible with demographic fluctuations, that nonetheless tend to lower diversity, and cannot be explained by structure, that tend to inflate diversity. We are now exploring models where selection lower diversity through genetic linkage, that are variants of the so-called genetic draft models. Under these types of models, diversity is expected to scale as a power function of the census size. Demography and structuration can be included in these models.

2 Signatures of replication timing, recombination and sex in the spectrum of rare variants on the human X chromosome and autosomes (Agarwal)

Agarwal Ipsita <ia2337@columbia.edu> (1), Przeworski Molly <mp3284@columbia.edu> (1) Affiliations: 1 - Columbia University (United States)

The sources of human germline mutations are poorly understood. Part of the difficulty is that mutations occur very rarely, and so direct pedigree-based approaches remain limited in the numbers that they can examine. To address this problem, we consider the spectrum of low frequency variants in a dataset (gnomAD) of 13,860 human X chromosomes and autosomes. X-autosome differences are reflective of germline sex differences, and have been used extensively to learn about male versus female mutational processes; what is less appreciated is that they also reflect chromosome-level biochemical features that differ between the X and autosomes. We tease these components apart by comparing the mutation spectrum in multiple genomic compartments on the autosomes and between the X and autosomes. In so doing, we are able to ascribe specific mutation patterns to replication timing and recombination, and to identify differences in the types of mutations that accrue in males and females. In particular, we identify C>G as a mutagenic signature of male meiotic double strand breaks on the X, which may result from late repair. Our results show how biochemical processes of damage and repair in the germline interact with sex-specific life history traits to shape mutation patterns on both the X chromosome and autosomes.

3 A New Isolation with Migration Model using whole-genome sequences (Ait)

Ait Kaci Azzou Sadoune <sadoune.aitkaciazzou@unifr.ch> (1), Leuenberger Christoph <christoph.leuenberger@unifr.ch> (1), Wegmann Daniel <daniel.wegmann@unifr.ch> Affiliations: 1 - Universite de Fribourg (Switzerland)

We present a new coalescent Hidden Markov Model to infer parameters of Isolation with Migration (IM) models while accounting for linkage disequilibrium. Due to the computational complexity of such methods, we don't consider the whole genealogies, but track some features of the genealogy only. In our case, we adopt the idea of Conditional Sampling Distribution (CSDs), used by dical2: we consider a particular haplotype h0 and track it back in time until it is absorbed (it coalesces) with the other haplotypes. Unlike dical2, however, we don't approximate the genealogies by unchanging trunks of ancestral lineages that extend indefinitely into the past. Instead, we account for coalescent events among the other lineages too, for which we propose an intuitive and computationally cheap formulation of the emission probabilities using two Markov discrete chains: 1) A vertical Markov chain that tracks the ancestral lineages back in time and allows us to calculate the probabilities of absorption at time t with a lineage from population j. 2) A horizontal Markov chain that tracks absorption states along chromosomes. As in dical2, we then use a leave-one-out composite likelihood to combine information from multiple haplotypes and infer parameters using numerical Baum-Welch optimization.

4 Genetics is an active learning algorithm for causal reconstruction of biological networks (Angeles-Albores)

Angeles-Albores David <dangeles@mit.edu> (1), Alm Eric <ejalm@mit.edu> (2) (3) (4), Thomson Matthew <mthomson@caltech.edu> (5) Affiliations: 1 - Department of Biological Engineering, MIT (United States), 2 - Center for Microbiome Informatics and Therapeutics, MIT (United States), 3 - Department of Biological Engineering, MIT (United States), 4 - Broad Institute [Cambridge] (United States), 5 - Department of Biology and Biological Engineering, Caltech (United States)

A common goal in genetics is the reconstruction of gene interactions into causal networks that provide insight into the inner workings of organisms. We can model genes in ON/OFF states, coupled with other genes via undirected edges. Such models are known as Markov Random Fields. Unfortunately, reconstruction of Markov fields from natural fluctuations is an NP-hard problem. The difficulty in reconstruction arises because sampling complexity grows exponentially as networks become strongly coupled (as biological networks are). This exponential growth is required to observe vanishingly rare high

energy states, which reveal otherwise invisible couplings. Thus, even with the enormous sequencing capacity of single-cell RNA-sequencing, we can only reconstruct small regulatory networks through observation. Instead, we must often settle for statistically satisfactory reconstructions. Reconstructions can become much more powerful if they are aided by targeted perturbations to the system. I will show that 'epistasis', as originally defined by Bateson in 1909, functions as a point-test for conditional independence. I will show that epistasis can be formalized and used to generate a mathematical theory that simultaneously describes a set of experiments to be performed and a computational algorithm to integrate information from these experiments into a complete network. In contrast with other methods such as maximum likelihood estimates, this theory becomes increasingly powerful and accurate as networks become tightly coupled.

5 Bait-ER: A Moran model for experimental evolution studies (Barata)

Barata Carolina <cdcbrb@st-andrews.ac.uk> (1), Borges Rui <ruiborges23@gmail.com> (2), Kosiol Carolin <ck202@st-andrews.ac.uk> (1) Affiliations: 1 - University of St Andrews [Scotland] (United Kingdom), 2 - Institute of Population Genetics, Vetmeduni Vienna (Austria)

Combining high-throughput next-generation sequencing with Experimental Evolution is a powerful approach to describe the evolution of allele frequencies in laboratory experiments. Researchers are thus able to re-sequence experimental populations to test evolutionary theory predictions. These Evolve and Re-sequence (E&R) experiments are especially useful for detecting signatures of short-term adaptation in the genome. One cost-effective way of estimating allele frequencies is to sequence pools of individuals (Pool-Seq). Though practical, pooled sequencing generates allele frequency variance due to finite sequencing depth. Moreover, most laboratory populations are small, where genetic drift plays a substantial role in determining the fate of most polymorphic sites. Accordingly, it is anything but trivial to test for selection in the genome and, particularly, to estimate selection coefficients. Despite numerous efforts to investigate allele frequency changes in such E&R experiments, most methods that aim at detecting selection still suffer from high false positive rates and low statistical power. For that, we have developed a fully Bayesian approach to estimate selection coefficients aimed at E&R allele frequency datasets. Our method is based on the Moran model of nucleotide evolution which allows for overlapping generations. We also employ a statistical test that uses Bayes Factors to compare two alternative non-nested models of evolution - neutrality and genetic drift vs positive selection. This hypothesis testing approach avoids the computational burden of simulating an empirical null distribution. We have tested our method using simulated data and compared its performance to that of other available software. It is comparable to other approaches in terms of computational time, which makes it suitable for genome-wide datasets. Testing was performed for various scenarios covering a wide range of experimental and demographic parameters. Our approach shows high accuracy even for complex demographic scenarios. Despite good overall performance, some allele frequency trajectories proved to be problematic, such as those where the effective population size is small and starting frequencies are low. Finally, we have analysed a Drosophila pseudoobscura time series dataset which looks at sexual selection in naturally polyandrous populations.

6 Detecting gene transfer within bacterial populations (Baumdicker)

Baumdicker Franz <baumdicker@stochastik.uni-freiburg.de> (1), Kelleher Jerome <jerome.kelleher@well.ox.ac.uk>, Kupczok Anne <akupczok@ifam.uni-kiel.de> Affiliations: 1 - Department of Mathematical Stochastics [Freiburg] (Germany)

Although bacteria reproduce clonally, it has become clear that a substantial fraction of genes has been transferred between bacterial species. However, the same mechanisms also lead to gene transfers between closely related strains of the same species. Here it is important to distinguish between homologous recombination replacing genetic material and gene transfer within the population that adds new genes to the genome of the recipient. Frequent transfer within a bacterial population can blur the signal of the clonal phylogeny and is hard to detect. Core genes will be mostly affected by homologous recombination, while accessory genes will be affected by both. Here we will focus on the transfer of accessory genes within populations. We present a method to distinguish gene transfers between and within populations, from whole genome data. In addition, we show how the ancestral graph created by gene transfer within the population can be simulated using a concept similar to gene conversion.

7 Improved prediction of site-specific mutation rates using k-mer pattern partition (Besenbacher)

Besenbacher Soren <besenbacher@clin.au.dk> (1) (2) Affiliations: 1 - Department of Molecular Medicine, Aarhus University (Denmark), 2 - Bioinformatics Research Centre, Aarhus University (Denmark)

Germline mutations are not uniformly distributed but occur with different rates at different positions in the human genome. The main factor influencing the rate at a specific position is the sequence context surrounding it. Models that use the sequence context at a given position to predict the mutation rate are very useful in both evolutionary studies and medical genetics. Modeling the mutation rate by estimating a rate for each possible k-mer, however, only works for small values of k since the data becomes too sparse for larger values of k. Here we propose a new method that solves this problem by grouping together k-mers with similar rates using IUPAC patterns. The proposed algorithm finds a set of IUPAC patterns so that each k-mer is matched by one and only one of the patterns. We refer to the method as k-mer pattern partition and we have implemented it in a software package called kmerPaPa. We use a large set of human de novo mutations to show that this new method leads to improved prediction of mutations rates and makes it possible to use longer sequence contexts than have previously been used to predict mutation rates. An added benefit of the model is that the results are easily interpretable and can reveal interesting patterns that are informative about the mutational processes that created the mutations. Because the model is trained on de novo mutations we expect the estimated mutation rates to be free from bias caused by selection and biased gene conversion. This means that the model can be used to disentangle the effects of mutation and selection on observed population genomics data. Additionally, the model can be used to improve frameworks for finding genes where de novo mutations cause disease.

8 Which birth-death models can account for competition in phylogenetic trees? (Biller)

Biller Priscila <pribiller@gmail.com> (1), Colijn Caroline <ccolijn@gmail.com> (1) Affiliations: 1 - Simon Fraser University (Canada)

Investigating how and why speciation and extinction processes vary over evolutionary time, geographical space and species groups is fundamental to understanding how ecological and evolutionary processes generate biological diversity. Birth-death models are one of the most common approaches to model diversification and estimate speciation and extinction rates. Although constant-rate birth-death models, because of their simplicity, have been used as a null model, different studies have pointed out that they are not adequate to capture the complexity and dynamics of speciation and extinction across the Tree of Life. Several biologically motivated extensions of birth-death models have been proposed in order to reflect structural properties of evolutionary trees inferred from real datasets. Our work investigates which model extensions could be an adequate null model for phylogenetic trees where competition plays an important role. We analyze the extent to which different birth-death models can capture the features of influenza viruses phylogenies, which are shaped by strong competition among the different strains.

9 Elastic net approach to spatially informed modelling of genetic variation (Bodde)

Bodde Marilou <mmb52@cam.ac.uk> (1), Durbin Richard <rd109@cam.ac.uk> (2) (1) Affiliations: 1 - Department of Genetics, University of Cambridge (United Kingdom), 2 - Sanger Institute (United Kingdom)

Marilou Bodde and Richard Durbin (mmb52@cam.ac.uk) Department of Genetics, University of Cambridge Modern genetic variation is shaped by many historic quantities and events, including population size, population structure, migration and local gene flow. Many inference approaches ignore spatial structure, but we know that for many species, including humans, geography and isolation by distance are extremely important. There are some approaches to modelling demography with spatial diffusion, such as EEMS to estimate effective migration surfaces, but here we will discuss a new approach. We are developing a spatially and temporally explicit data analysis approach based on a generative population model that aims to fit local population allele frequencies at different points in time using both modern and ancient DNA. The inference method is based on the 1980's elastic net approach, originally formulated to find good solutions

the travelling salesman problem [1,2]. This can be viewed as an optimization technique to fit elastic surfaces with noise to data. The change in allele frequencies, as a function of space and time, is affected differently by continuous local gene flow and big migration events and by explicitly representing these we hope to distinguish between the two. We will present the model and its application to human data from Europe and North Asia. References: 1. Durbin, R. & Willshaw, D. An analogue approach to the travelling salesman problem using an elastic net method.Nature326,689-691 (1987). 2. Durbin R., Szeliski, R. & Yuille, A. An analysis of the elastic net approach to the traveling salesman problem. Neural Computation1,348-358 (1989).

10 Bayesian polymorphism-aware phylogenetic models accounting for allelic selection (Borges)

Borges Rui <ruiborges23@gmail.com> (1), Boussau Bastien <boussau@gmail.com> (2), Szollozi Gergely <sszolo@gmail.com> (3), Kosiol Carolin <ck202@st-andrews.ac.uk> (4) 1 - Institute of Population Genetics, Vetmeduni Vienna, Austria (Austria), 2 - Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5558, LBBE, F-69100, Villeurbanne, France (France), 3 - Department of Biological Physics, ELTE-MTA «Lendulet» Biophysics Research Group, Eotvos University, Pazmany P. stny. 1A, Budapest H-1117, Hungary (Hungary), 4 - Centre for Biological Diversity, University of St Andrews, St Andrews, Fife KY16 9TH, UK (United Kingdom)

Molecular phylogenetics has neglected polymorphisms within present and ancestral populations for a long time. Alternative models accounting for multi-individual data have nevertheless been proposed and are known as polymorphism-aware phylogenetic models (PoMo). PoMo adds a new layer of complexity to the standard nucleotide substitution models by accounting for the population- level (so far, genetic drift and mutations) processes to describe the evolutionary process. To do so, PoMo expands the standard substitution models to include polymorphic states, thereby naturally accounting for incomplete lineage sorting (ILS). Here, we extend PoMo to account for allelic selection and derive its stationary distribution. Our theoretical results constitute the basis of a new Bayesian framework. In this talk, we will discuss advantages of the Bayesian PoMo that allows the incorporation of more-complex processes such as the molecular clock and gene duplication and loss. In particular, we present the modeling of GC-biased gene conversion (gBGC), which is a recombination-associated process that prefers G/C alleles over A/T alleles, and a process that has been fairly ignored by phylogenetic inference. In phylogenetic terms, gBCG is a nucleotide usage bias, but as ILS, significantly impacts the accuracy of phylogenetic methods as it acts genome-wide. Here, we evaluate the joint impact of ILS and gBGC on tree inference using a polymorphism-aware model that is able to correct for both ILS and gBGC. More specifically, we simulated data sets where we combined scenarios of strong/weak ILS and gBGC. Phylogenetic inference was done using the polymorphism-aware phylogenetic models implemented in RevBayes. Not only we observed that the standard phylogenetic models perform poorly when estimating branch lengths, as they could not, in some cases, recover the true topology, especially for closely related populations. Implications to molecular dating are also presented with examples in real data sets.

11 Bits to Bases: Using Generative Models to Produce Synthetic Genetic Data (Burak)

Burak Yelmen, Aurelien Decelle, Linda Ongaro, Davide Marnetto, Francesco Montinaro, Corentin Tallec, Cyril Furtlehner, Luca Pagani, Flora Jay. Affiliation: Laboratoire de Recherche en Informatique, CNRS, Université Paris Sud, Orsay

Availability of genetic data has increased tremendously due to advances in sequencing technologies and reduced costs. The vast amount of genetic data is used in a wide range of fields, from medicine to evolution. However, the majority of the data held by private companies and government institutions are not accessible to researchers due to privacy issues. Using machine learning, we could generate synthetic genomes that successfully mimic the real ones but are not identical to any of them. We relied on two types of neural network architectures: (1) Generative Adversarial Networks, a breakthrough in the domain of computer vision allowing the generation of extremely realistic images; (2) Restricted Boltzmann Machines, capable of learning complex data distributions. We measured the quality of the generated genomes in terms of

population structure, linkage disequilibrium and haplotype diversity, and demonstrated that they provided an accurate representation of the real ones. Without duplicating any of the individuals, most key characteristics of the data were conserved. We also showed a drastic improvement compared to simpler Markovian models. A major application will be the conversion of private datasets to synthetic genomes that can then be made public without any privacy constraints. A direct implication is the increase in richness of public datasets, e.g. with populations still under-represented in genetic studies. To highlight the high potential of our approach we further demonstrated how synthetic genomes used in reference panels for imputation led to performances equally good as with real (hypothetically private) genomes.

12 Inferring Genotype-Environment Associations from Low-Depth Sequencing Data (Caduff)

Caduff Madleina <madleina.caduff@unifr.ch> (1) (2), Link Vivian <vivian.link@unifr.ch> (1) (2), Sonnenwyl Vincent <vincent.sonnenwyl@unifr.ch> (1), Leuenberger Christoph <christoph.leuenberger@unifr.ch> (1), Wegmann Daniel <daniel.wegmann@unifr.ch> (1) (2) Affiliations: 1 - Universite de Fribourg (Switzerland), 2 - Swiss Institute of Bioinformatics (SIB) (Switzerland)

The genetic basis of local adaptation can be revealed by identifying loci whose allele frequencies correlate with the environment. However, allele frequencies of neutral loci may also correlate in space as a result of isolation by distance. Disentangling the effects of environmental selection from the effects of neutral population structure is thus crucial when evidencing genotype- environment associations. A particularly powerful approach is to use latent factors that capture the residual genetic structure not explained by the environment. Latent factors and environmental effects are then estimated simultaneously. A problem with current such methods is that they assume genotypes to be known. However, this is rarely the case for data obtained by next-generation sequencing due to high error rates that prohibit confident genotype calling unless sequencing depth is very high. Given a fixed budget, only a limited number of samples can be sequenced at high depth, reducing the statistical power to infer genotype-environment associations. As a powerful alternative, we propose here to sequence many samples at low depth. For this we present a Bayesian method that properly accounts for genotype uncertainty while simultaneously inferring genotype-environment associations and population structure via latent factors. Our method also introduces a sparse prior reflecting the idea that only a small proportion of all loci contributes to local adaptation. The hierarchical formulation further allows us to account for linkage between loci and hence to aggregate information across loci in linkage disequilibrium with a causal locus. As we show with simulations, many samples sequenced at low depth always yields higher statistical power to infer genotype-environment associations than few samples sequenced at high depth. Indeed, we found that the power was maximized at an average sequencing depth of

13 Bayesian nonparametric inference of population trajectories via Tajima heterochronous n- coalescent. (Cappello)

Cappello Lorenzo <cappello@stanford.edu> (1), Palacios Julia A. <juliapr@stanford.edu> (1) Affiliations: 1 - Stanford University (United States)

The observed variation in gene samples allows to infer evolutionary parameters such as past population dynamics: it is common practice to model such a variation as a mutation process superimposed on a stochastic genealogy sampled from the Kingmann-coalescent. However, the state space of Kingman's genealogies grows superexpo- nentially; thus inference is computationally unfeasible already for small sample sizes. An alternative to Kingman coalescent has been proposed in the literature, the Tajiman-coalescent, which relies on a coarser resolution, reducing the state space substantially. Such process does not accomodate samples collected at different times, a situation that in applications is both real (e.g. ancient DNA, influenza viruses) and desirable (it reduces the variance of the estimate). In order to fill this gap, we introduce a new process, called Tajima heterochronous n-coalescent, define the exact likelihood of observed mutations given a Tajima's genealogies, and present a Bayesian nonparametric procedure to infer past population size. We propose also a new sampler to explore the space of Tajima's genealogies. We compare our procedure with state-of-art algorithms on simulated and real data-sets

14 Assessing the impact of demography and multinucleotide mutations on reference-free archaic admixture inference methods (Carlson)

Carlson Jedidiah <jed.e.carlson@gmail.com> (1), Hsieh Pinghsun <hsiehph@u.washington.edu> (1), Harris Kelley <harriske@uw.edu> (1) (2) Affiliations: 1 - Department of Genome Sciences, University of Washington (United States), 2 - Computational Biology Division, Fred Hutchinson Cancer Research Center (United States)

Advances in ancient DNA sequencing technology have produced high-quality reference genomes for Neanderthal and Denisovans, spurring the development of methods that infer the timing and extent of interbreeding between these archaic populations with homo sapiens and ultimately leading to the identification of genomic tracts of archaic admixture in extant human populations. More recently, several methods have estimated admixture contributions from unsampled ghost populations revealing new layers of complexity in the demographic and evolutionary history of human populations. Because ghost archaic populations inherently lack a reference genome, the methods used for inferring the presence of ghost admixture rely on the intuition that this signal will manifest as an abundance of private SNPs that tend to be closely-spaced and in high linkage disequilibrium. Such methods must therefore assume that this signal is uniquely identifiable from 1) patterns of variation caused by other demographic events not involving ghost admixture, and 2) patterns of variation caused by non-random or non-independent mutation processes, such as mutation hotspots or clustered multinucleotide mutations. In this study, we simulate data under a variety of realistic demographic scenarios using both a simple Poisson-distributed mutation model and a mutation model which accounts for non-independent multinucleotide mutations and regional mutation rate heterogeneity. We then apply various reference-free ghost admixture inference methods to these simulated datasets to evaluate the methods' sensitivity to fluctuations in the mutation model and demographic model parameters. We find that multiple methods are sensitive to both the demographic model parameters and the effects of multinucleotide mutations; in some scenarios, the false positive rate for archaic admixture is as high as 2%, suggesting that previously published estimates of archaic admixture may have been overestimated, particularly the putative signals of ghost archaic admixture where no archaic reference genome is available. Our results suggest that the modeling of mutational and demographic complexity is needed to accurately estimate the extent of ghost admixture, and we present strategies for distinguishing true ghost ancestry from variant clusters that only masquerade as such.

15 Epsilon-Genic Effects Bridge the Gap Between Polygenic and Omnigenic Complex Traits (Cheng)

Cheng Wei <wei_cheng1@brown.edu> (1) (2), Ramachandran Sohini <sohini_ramachandran@brown.edu> (1) (2), Crawford Lorin <lorin_crawford@brown.edu> (1) (3) (4) Affiliations: 1 - Center for Computational Molecular Biology, Brown University, Providence, RI, USA (United States), 2 - Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA (United States), 3 - Department of Biostatistics, Brown University, Providence, RI, USA (United States), 4 - Center for Statistical Sciences, Brown University, Providence, RI, USA (United States)

Traditional univariate genome-wide association studies generate false positives and negatives due to difficulties distinguishing causal variants from interactive variants (i.e., variants correlated with causal variants without directly influencing the trait). Recent efforts have been directed at identifying gene or pathway associations, but these are often computationally costly and hampered by strict model assumptions. Here, we present gene-Î, a new approach for identifying statistical associations between sets of variants and quantitative traits. Our key innovation is a recalibration of the genome-wide null model to include small-yet-nonzero associations emitted by interactive variants, which we refer to as epsilongenic effects. gene-Îefficiently identifies core genes under a variety of simulated genetic architectures, achieving up to ~90% true positive rate at 1% false positive rate for polygenic traits. Lastly, we apply gene-Îto summary statistics derived from six quantitative traits using European-ancestry individuals in the UK Biobank, and identify gene sets that are enriched in biological relevant pathways.

16 Evidence for a Paleolithic Back-to-Africa Migration (Cole)

Cole Christopher <ccole@well.ox.ac.uk> (1) (2), Zhu Sha (joe) <joe.zhu@bdi.ox.ac.uk> (3), Lunter Gerton <gerton.lunter@well.ox.ac.uk> (1) (2) Affiliations: 1 - MRC Weatherall Institute of Molecular Medicine (United Kingdom), 2 - The Wellcome Trust Centre for Human Genetics [Oxford] (United Kingdom), 3 - Oxford Big Data Institute (United Kingdom)

Background: Many details surrounding anatomically modern human's range expansion out of Africa (OoA) remain unclear. In part this is due to a lack of appropriate statistical models for inferring complex demographic histories in the ancient past. Methods: Here we introduce SMCSMC, an application of Sequential Monte Carlo to the sequentially Markovian Coalescent. We sequentially build samples from the posterior distribution of genealogical trees conditioned on genetic variation along the genome using a combination of importance sampling and periodic resampling. We generalize the Auxiliary Particle filter to continuous-time models in order to reduce bias in recent epochs. After obtaining a sample of trees, we use Variational Bayes to optimize parameters of the underlying demographic model. The method can be used with arbitrarily complex demographic models, so long as the model admits simulation of genealogies along the genome, for which we use the Sequential Coalescent with Recombination Model (SCRM). SMCSMC is available as a python package on the conda package manager. Results: We use SMCSMC to infer joint demographic histories in individuals from the Simons Genome Diversity Project (SGDP). We identify evidence for substantial migration from proto- Eurasians into the ancestors of modern Sub-Saharan Africans after the OoA split. Including parameters for asymmetric migration resolves an artefact in population size estimates of African populations first described in Li and Durbin 2011. All OoA populations which we tested (with the notable exception of the Karitiana and Surui) appear to have descended from the proto-Eurasian source population for the admixture event, while all sub-Saharan African groups, with the exception of the Khoisan peoples, appear to descend from the recipient population. The length distribution of the inferred resulting admixture tracts indicates that the event took place between 60-65 kya. We use coalescent simulation to explore our ability to recover back-migration and find that our inference is biased towards more recent events. We use ADMIXTOOLS and the isolated admixed tracts, covering between 4 and 6% of the genomes, to estimate a mixing proportion between 2 and 8%, depending on the population. Discussion: Inferring asymmetric migration in the period surrounding the OoA migration reveals a substantial admixture event of proto-Eurasian immigrants to sub-Saharan Africa, which explains and resolves a persistent artifact in single-population effective population size estimates in African subpopulations. The Khoisan, who diverged from other African populations prior to the OoA event, are an outgroup to the migration. The Karitiana and Surui are the only OoA groups we identified who do not act as a source for the admixture event, contesting the view that their ancestors diverged after the East/West Eurasian split. This opens the possibility of contributions to this group from an earlier branch of proto-Eurasians or possibly a separate migration out of Africa. The back-to-Africa migration we explore is consistent with earlier proposals to explain geographic distributions of Y chromosome and mitochondial haplogroups. This analysis represents the first whole-genome analysis to support this event and provide systematic characterization of its extent, magnitude, and timing.

17 Imputation of mother and fetus from sequence (Davies)

Davies Robert <robertwilliamdavies@gmail.com> (1), Chen Ruoyan <chenruoyan@genomics.cn> (2), Li Zilong <lizilong@bgi.com> (2), Liu Siyang <liusiyang@bgi.com> (2) Affiliations: 1 - Department of Statistics [Oxford] (United Kingdom), 2 - BGI-Shenzhen (China)

Large genome-wide association studies facilitate genetic research including more powerful locus discovery and more accurate polygenic scores. In recent years, non-invasive prenatal testing (NIPT) using cell-free DNA and low-coverage whole genome sequencing is increasingly become standard of care clinically. NIPT can therefore be a source of inexpensive genotypes for large scale GWAS. However, unlike traditional GWAS, NIPT contains information derived from three haplotypes-the maternal transmitted, maternal untransmitted, and fetal transmitted haplotypes. Here we present a dedicated method for NIPT imputation that simultaneously imputes both the maternal and fetal genomes. We present two versions of the method - an exact approach, and a sampling based approach that has linear computational complexity with

haplotype reference panel size. Assessing accuracy using high coverage trio data, we show this method outperforms existing methods for diploid low coverage whole genome sequence based imputation. We show that imputation accuracy is high at moderate coverage (at 4X, maternal r2 > 0.95, fetal r2 > 0.90), and reduces with decreasing coverage. We further show we are able to specifically impute each of the three haplotypes present in NIPT data. Finally we demonstrate potential for GWAS using the >1M person ~0.1X coverage Millionome study of Han Chinese females.

18 Inferring mutation spectrum histories from sample frequency spectra (Dewitt)

Dewitt William <wsdewitt@uw.edu> (1), Harris Kelley <harriske@uw.edu> (2) Affiliations: 1 - Department of Genome Sciences [Seattle] (United States), 2 - University of Washington [Seattle] (United States)

SNV spectra parameterized by triplet nucleotide context vary among human ancestry groups and among great ape species. The triplet-resolved sample frequency spectrum (3-SFS) encodes information about triplet-specific mutation rate histories, but standard demographic inference methods assume a constant and context-agnostic mutation rate to infer effective population size histories and migration parameters from the SFS. Here, we explore prospects for inference of mutation spectrum histories from the 3-SFS, first considering sequential inference procedures that condition on demography. We model the 3-SFS in a diffusion setting with a time-inhomogeneous boundary flux of input mutations in each triplet component. In a coalescent setting we model the expected 3-SFS as a nonlinear functional of the triplet mutation spectrum history. Using coalescent and forward simulation approaches, we study the effects of population structure on SNV spectra and ask if mutational pulses caused by a segregating mutator allele are distinguishable from those that occur due to population-wide environmental causes. We use these tools to help interpret previously-reported evidence of an ancient TCC>TTC mutation pulse in Europeans and South Asians.

19 Toward more realistic sequentially Markov coalescent models (Dutheil)

Dutheil Julien <dutheil@evolbio.mpg.de> (1) Affiliations: 1 - Max Planck Institute for Evolutionary Biology (Germany)

While population genomic data sets and approaches are used to address increasingly diverse biological questions for a wide range of species, there is a need for population genetic models that can extract biological signal from such data. The sequentially Markov coalescent (SMC) is a first-order Markov approximation to the sequential coalescent, which allows computationally efficient demographic inference with complete genome data. The SMC is a unique framework that successfully exploits the information contained in marginal genealogies, as well as their linkage along the genome. While several developments focused on the implementation of more complex demographic scenarios, for instance allowing for ancestral population structure, existing models so far consider the process to be sequentially homogeneous. This assumption is at odds with our knowledge of the biology of genomes, as mutation and recombination rates can notoriously be highly variable along the genome. I will present here a new framework, termed Çintegrative SMCÈ (iSMC), which extends the original SMC to allow parameters (such as the mutation and recombination rates) to vary along the genome. These heterogeneous parameters are modeled using a prior distribution and an autocorrelation parameter, their variation along the genome being modeled as a Markov process. As a result, the iSMC process is a Markov-modulated Markov chain that can be handled similarly to the original SMC. I will show that iSMC can successfully infer the demography together with the recombination and mutation landscapes. I will more specifically discuss the challenges of applying such models to a broader range of organisms, including ancient genomes and non-Primate species.

20 Reconstructing complex evolutionary and demographic histories (Eriksson)

Eriksson Anders <aeriksson75@gmail.com> (1) Affiliations: 1 - Department of Medical & Molecular Genetics, King's College London, SE1 9RT London (United Kingdom)

Since their first appearance in Africa, only around 300,000 years ago, anatomically modern humans have spread across the world into a wide range of environments with different climates, pathogens, flora and

fauna. As a result, they have also gone through several major demographic and cultural transitions, in particular over the last 15,000 years, that would have entailed major changes to diet, pathogen exposure and other environmental variables that are today associated with individual variation in many physiological traits and susceptibility to disease. A detailed understanding of these processes is important for a wide range of fields, including evolutionary biology, biological anthropology and genomic medicine, and the availability of whole genome sequences from many thousands of individuals from different populations around the world offers unprecedented opportunities to answer these questions. However, current tools provide very limited resolution to the timing of the inferred processes and are also sensitive to the confounding effects of past demographic events, in particular admixture between past populations as the result of long-range migrations. Together these factors have prevented linking past selection and demographic events to environmental and climatic causes during this time period, thereby limiting our ability to understand how these factors have shaped recent human evolution and population histories. I will present a flexible and powerful spatially explicit framework for inferring past demographic process and natural selection, that enables explicitly taking the age and geographic locations of past and present samples into account, as well as linking demography and natural selection to paleoclimate and terrain. Exploiting recent advances in analyses of ancient DNA and large whole genome sequence datasets from diverse present-day populations, this framework brings a number of advantages: it can explicitly account for long-range gene flow, such as during the spread of agriculture with Neolithic farmers and the gene flow from the Asian steppes into Europe during the Bronze age; estimate the geographic origin and spread of advantageous alleles during spatial selective sweeps and to identify adaptive and maladaptive gene flow in past admixture events; and formally place the mysterious ÇghostÈ populations, so common in ancient DNA analysis, in space and time and thus linking them to the archaeological record. Not only will this framework give us the statistical power to disentangle complex evolutionary and demographic processes, but also to formally test hypothesis of the environmental and cultural processes that have been driving them. This will be important for wide range of fields, including evolutionary and medical genetics, biological anthropology and archaeology.

21 Estimating the conditional risk of psoriatic arthritis in the UK Biobank (Fadil)

Fadil Chaimaa <chaimaa.fadil@trinity.ox.ac.uk> (1) (2), Mcvean Gil <gil.mcvean@bdi.ox.ac.uk> (2) (3) Affiliations: 1 - Department of Statistics, University of Oxford, Oxford (United Kingdom), 2 - Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford (United Kingdom), 3 - Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford (United Kingdom)

Early diagnosis and treatment of rheumatic diseases such as rheumatoid arthritis, psoriatic arthritis and systemic lupus erythematosus have been hindered by heterogeneous clinical presentations and progression outcomes. This raises the need to better characterise the phenotypic subgroups underlying this heterogeneity. We address this question by using the UK Biobank, a longitudinal phenotypic and genetic database, to investigate the role played by co-morbidities and their underlying genetic architecture on disease susceptibility and prognosis. We specifically focus on estimating the conditional risk of psoriatic arthritis (PsA) to be able to define a population of patients at increased risk of developing and sustaining PsA. PsA is a chronic immune-mediated disease, often described as a progression of psoriasis, an inflammatory skin disease affecting more than 3% of the world population. It also shows widespread musculoskeletal inflammation, a characteristic of inflammatory arthritides. In order to pull apart the heterogeneity within PsA cases and elucidate the conditional role played by co- morbidities in modulating the risk for PsA, we rely on the UK Biobank to identify the different patterns of HLA associations within PsA patients and compare these to characteristic MHC associations found in psoriasis and inflammatory arthritides We rely on a cohort of 982 PsA patients, 7787 psoriasis patients and 86289 patients diagnosed with inflammatory arthritis conditions. Among the entire spectrum of ICD- 10 hospital episodes, psoriasis (RR=4.0) and rheumatoid arthritis (RR=1.6) stand out as the co- morbidities with highest risk for PsA, re-emphasising the conception of PsA as a co-occurence of psoriasis and arthritis disorders. While psoriasis is ubiquitous among PsA patients, we identify on the other hand a subset of PsA patients with no arthritis diagnosis preceding their PsA diagnosis, suggesting the existence of multiple pathways through which PsA can be triggered.

22 Fast and accurate identity-by-descent inference despite haplotype and phasing errors (Freyman)

Freyman Will <willf@23andme.com> (1), Mcmanus Kimberly <kmcmanus@23andme.com> (1), Shringarpure Suyash <sshringarpure@23andme.com> (1), Jewett Ethan <ejewett@23andme.com> (1), Auton Adam <aauton@23andme.com> (1) Affiliations: 1 - 23andMe, Inc. (United States)

Estimating the genomic location and length of identical-by-descent (IBD) segments among related individuals is a central step in many genetic analyses. Because IBD segments are broken up by meiotic recombination they are expected to be longer for close relatives. However, long IBD segments are more likely to be impacted by haplotype and phasing errors compared to short segments. This makes accurate inference of phased IBD among close relatives particularly challenging. Here we present a method based off the positional Burrows-Wheeler transform (PBWT) and a probabilistic hidden Markov model (HMM) to make fast and accurate IBD estimates. We use haplotype data simulated over pedigrees to explore the performance of our algorithm against other IBD inference approaches for both distant and close relatives. Additionally we calculate our method's false positive rate and power to detect IBD segments of varying lengths.

23 Identifying eQTLs from Single-Cell RNA-seq Using a Topic Modeling Framework (Gewirtz)

Gewirtz Ariel <gewirtz@princeton.edu> (1), Engelhardt Barbara <bee@princeton.edu> (2) Affiliations: 1 - Quantitative and Computational Biology [Princeton] (United States), 2 - Computer Science Department [Princeton] (United States)

Understanding how genotypic variation affects downstream phenotypes such as gene expression and disease is vital for developing effective, personalized treatments. However, most existing expression quantitative trait loci (eQTL) association studies use bulk RNA-seq data, which potentially confounds results due to averaging gene expression levels over a mixture of heterogeneous cell types. Single-cell RNA-seq (scRNA-seq) produces fine-grained cellular-level transcription information, yet statistical methodology to integrate single cell RNA-seq data with single nucleotide polymorphisms (SNPs) is critically lacking. Current scRNA-seq studies generally combine gene counts across all cells from an individual and use this 'pseudo-bulk' count as input to a single-gene, single-SNP eQTL association test. However, due to statistical properties of scRNA-seq data such as sparsity and heavy tails, the assumptions behind basic association methods like linear regression do not hold. Controlling for cell type heterogeneity in pseudo-bulk methods produces artifacts that may confound associations as in bulk RNA-seq studies. Additionally, testing groups of correlated genes and SNPs instead of a one-to-one association test provides more comprehensive and robust insight into genetic drivers of transcriptional variation and better captures the highly complex transcriptional regulatory process. In this study, we develop a novel topic modeling framework that (i) uses raw scRNA-seq UMI counts, avoiding addition of spurious signals through data normalization or transformation, (ii) tests for association between genotype and single-cell transcriptional data, (iii) exploits correlations among multiple genes and SNPs to increase sensitivity and robustness, and (iv) addresses issues related to tissue heterogeneity. Topic models are currently used in the single-cell field to cluster cells and perform dimensionality reduction; our approach adapts these models to allow two or more data modalities (naively, SNPs and genes) and to account for cell type labels and proportions when association testing. We apply our model to a scRNA- seq data set of 484,072 cell-type-labeled PBMC cells from 119 individuals with germline SNP information to identify eQTLs, explore their potential mechanisms, and investigate their variation among cell types and individuals.

24 Accurate genotyping in polymorphic repetitive loci using k-mer count profiles (Gibling)

Gibling Heather <heather.gibling@oicr.on.ca> (1) (2), Ang Houle Armande <armande.anghoule@oicr.on.ca> (1) (2), Simpson Jared <Jared.Simpson@oicr.on.ca> (1) (3), Awadalla Philip <philip.awadalla@oicr.on.ca> (1) (2) Affiliations: 1 - Ontario Institute for Cancer Research [Canada]

(Canada), 2 - Department of Molecular Genetics, University of Toronto (Canada), 3 - Department of Computer Science, University of Toronto (Canada)

In highly repetitive loci, accurate identification of genomic variants using short-read sequencing can be difficult due to reads mapping to more than one region, which can affect downstream analyses regarding polymorphisms and gene expression. The difficulty is amplified when different variants have a high sequence similarity. We are developing a probabilistic method to accurately call genotypes in repetitive loci using a k-mer count profile approach. Counts of k-mers present in sequencing reads of a sample are compared to k-mer count profiles from known alleles and a Poisson distribution of expected counts determines the probability of observing the reads from an allele given the k-mer count profiles. To assess effectiveness, we called alleles for the highly polymorphic gene PRDM9, which has 36 known alleles that differ by arrangements of minisatellite-like zinc finger (ZnF) repeats. Our method is able to accurately call haploid PRDM9 alleles: using 100bp paired-end reads simulated at 100X coverage with 0% and 0.1% sequencing error rates, we observe average F1-scores of 0.999 and 0.980, respectively. Even when reducing the simulated coverage to 20X with sequencing error rates of 0% and 0.1%, we can still achieve high accuracy, with F1-scores of 0.979 and 0.942, respectively. We are currently developing an approach for incorporating the distance between k-mers on paired-end reads given the expected distance from a Gaussian distribution which will help distinguish between diploid genotypes. Initial tests have resulted in F1-scores of 0.983 for calling diploid PRDM9 genotypes from 100bp paired-end reads simulated at 100X coverage with a 0% sequencing error rate. This approach is extendable to several other repetitive or polymorphic regions of the genome, such as the Cytochrome P450 (CYP) genes involved in drug metabolism, which we are exploring to better predict drug response given allelic variability. Our tool will provide a resource for better characterization of variants that are traditionally difficult to ascertain using current short-read sequencing approaches.

25 Demographic Model Selection with Deep Learning (Gladstein)

Gladstein Ariella <aglad@med.unc.edu> (1), Schrider Daniel <drs@unc.edu> Affiliations: 1 - The University of North Carolina at Chapel Hill (United States)

Genome sequences contain clues about evolutionary history, including natural selection, population size changes, and gene flow between different populations. Answering each of these questions not only informs our view of natural history, but also of the forces shaping genetic diversity within populations such as pathogens and other environmental selective pressures. One of the key challenges in population genomics is to infer demographic histories on the basis of genome-sequence data, in part because such information is a prerequisite for many downstream analyses. Here we compare a number of approaches for inferring demographic histories, focusing primarily on two- population models with varying degrees and scenarios of gene flow. We examine a number of previously described methods as well as a novel one that uses a Convolutional Neural Network (CNN) that discriminates between demographic models on the basis of chromosome-scale sequence alignments. Instead of using classic population genetic statistics, the CNN learns the informative features from the sequence alignments. We train, validate, and test the CNN on coalescent simulations, evaluating a variety of neural network architectures, training set sizes, alignment lengths, and demographic model complexities. We find that the CNN requires comparatively few simulations to attain a high accuracy of demographic model discrimination, and often matches or exceeds the accuracy of existing methods. We believe this work has the potential to drive a paradigm shift in the methodology of population genetics research: instead of using statistics designed for particular population genetics questions, we can directly use the raw sequence alignment data to answer a wide range of model classification or parameter inference questions.

26 Evolution of germline mutation rate in great apes (Goldberg)

Goldberg Michael E. <goldmich@uw.edu> (1), Harris Kelley <harriske@uw.edu> (1) Affiliations: 1 - University of Washington [Seattle] (United States)

Although the germline mutation rate is classically regarded as a fixed parameter of the evolutionary process, recent studies of human and ape genetic variation have shown that the mutation rate and spectrum can

evolve rapidly. The relative mutation rates of different three-base-pair genomic motifs differ significantly among great ape species, suggesting the recent fixation of unknown modifiers of DNA replication fidelity. To shed light on what these modifiers might be, we measured the relative mutabilities of all three-base-pair motifs in specific compartments of the genome (such as endogenous retroviruses and late-replicating regions) that we expect to be targeted by known mutational processes. Using genetic diversity data from 88 great apes, we measured the covariation of mutational spectra between compartments and species, finding evidence of compartment-specific mutational processes that are largely conserved across the ape phylogeny. In contrast, however, species-specific recombination hotspots demonstrate divergent mutational signatures as a result of the rapid evolution of their function. The differences between compartment-specific mutational signatures are robust to regional variation in mapping quality or nucleotide content. These compartment-specific signatures layer with species-specific signatures to create rich mutational portraits: for example, orangutan endogenous retroviruses contain an identifiable mixture of an orangutan-specific signature and a signature that we hypothesize is due to hydroxymethylation of retrovirus-derived DNA. Strikingly, western chimpanzees have a different mutation spectrum from other subspecies of chimps, and the difference between western and non-western chimps closely resembles the difference between repetitive and nonrepetitive DNA. Our results suggest that rapidly evolving mutation rate modifiers tend to act broadly in trans across the whole genome, whereas cis regulators of mutation in specific genomic compartments are highly conserved between species.

27 Detecting archaic adaptive introgression using convolutional neural networks. (Gower)

Gower Graham <graham.gower@gmail.com> (1), Racimo Fernando <fracimo@bio.ku.dk> (1) Affiliations: 1 - University of Copenhagen (Denmark)

Analysis of sequencing data from Neanderthals and Denisovans has revealed several episodes of admixture occurred between modern and archaic hominin lineages during the Pleistocene. Archaic hominins were present in Eurasia long before modern humans, and thus likely had more time to adapt to local conditions. Recently, numerous studies have shown evidence for beneficial variants that were introduced into the modern human gene pool from archaic humans, and were later positively selected - a process known as adaptive introgression. However, there are no explicit frameworks for jointly modelling introgression and positive selection, and inferring parameters of interest, like the strength of selection on beneficial archaic alleles. Machine learning frameworks such as convolutional neural networks (CNNs) are increasingly being applied to problems in genomics, including the detection of positive selection. CNNs do not require the specification of an analytical model of allele frequency dynamics, and have outperformed alternative methods for classification and parameter estimation tasks in population genetics. Thus, CNNs are potentially well suited to the identification of adaptive introgression, and estimation of related parameters such as the strength of selection. We simulated a demographic model reflecting the history of modern humans as they migrated out of Africa, followed by admixture with Neanderthals, with various modes of selection on a region of the genome, including adaptive introgression and selective sweeps from de novo mutations. We then trained a CNN on genotype matrices derived from the simulated archaic and modern human genomes, to distinguish selection from neutrality and classify the mode of selection. Our CNN architecture exhibits good performance on simulated data, even for genotype matrices scaled down to low-resolution images. We then tested a range of additional parameters relevant to the characteristics of empirical datasets, such as unphased genotypes and varying degrees of data missingness. Finally, we applied our trained CNN to predict and model adaptive introgression in the 1000 Genomes Project dataset, in order to detect candidates for selection and understand their adaptive history.

28 Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding (Hanghoej)

Hanghoej Kristian <k.hanghoej@bio.ku.dk> (1), Moltke Ida <ida@binf.ku.dk> (1), Andersen Philip Alstrup <philipalstrup@hotmail.com> (1), Manica Andrea <am315@cam.ac.uk> (2), Sand Korneliussen Thorfinn <thorfinn.sand@gmail.com> (1) Affiliations: 1 - University of Copenhagen (Denmark), 2 - University of Cambridge (United Kingdom)

The estimation of relatedness between pairs of possibly inbred individuals from high- throughput sequencing data has previously not been possible for samples where we cannot obtain reliable genotype calls, as in the case of low-coverage data. We introduce ngsRelateV2 that takes into account the possibility of individuals being inbred by estimating the nine condensed Jacquard coefficients. Based on linear combinations of these, we calculate a series of unbiased compound statistics such as relatedness, kinship and inbreeding coefficients. The method accounts for genotype uncertainty making it particularly well suited for low coverage sequencing data. We demonstrate the even for complicated pedigree scenarios, compound statistics remain highly accurate. The software is available as an open source C/C++ program and hosted at https://github.com/ANGSD/ngsRelate. To facilitate easy analysis, the program is able to work directly on the most commonly used container formats for raw sequence (BAM/CRAM) and summary data (VCF/BCF). The program is threaded and scales linearly with the number of cores allocated to the process.

29 Predicting the short-term success of human influenza A variants with machine learning (Hayati)

Hayati Maryam <mhayati@sfu.ca> (1), Colijn Caroline <ccolijn@gmail.com> (1), Biller Priscila <pribiller@gmail.com> Affiliations: 1 - Simon Fraser University (Canada)

Seasonal influenza viruses are constantly changing, and produce a different set of circulating strains each season. These small genetic changes can accumulate over time and result in antigenically different viruses. Accordingly, this may prevent the body's immune system to recognize those viruses. Due to the rapid mutations in the hemagglutinin gene, vaccines against seasonal influenza have to be updated frequently. This requires choosing strains to include in the updates to maximize the vaccines' benefits, according to estimates of which strains will be circulating in upcoming seasons. This is a challenging prediction task. In this paper we use longitudinally sampled phylogenetic trees based on hemagglutinin sequences, together with counts of epitope site polymorphisms in hemagglutinin, to predict which influenza strains are likely to be successful. We extract small groups of taxa (subtrees) and use a suite of features of these subtrees as key inputs to the machine learning tools. Using a range of training and testing strategies, including training on H3N2 and testing on H1N1, we find that successful prediction of the future expansion of small subtrees is possible from these data, with accuracies of 0.71- 0.85 and AUC 0.75-0.9.

30 Modeling dynamics of circulating tumor DNA for detecting resistance to targeted therapies: a phylogenetic approach (Herbach)

Herbach Ulysse <ulysse.herbach@inria.fr> (1) Affiliations: 1 - Inria Nancy - Grand Est (France)

Targeted therapies represent a real advance in the treatment of patients with cancer. Most of these therapies are kinase inhibitors and require precise analysis of tumor DNA mutations to ensure the absence of primary resistance. Although tumours are often genetically heterogeneous with the presence of many subclones, they release circulating cell-free DNA (cfDNA) that can be directly extracted from basic blood samples: as sensitivity of measurements improves, such liquid biopsies increasingly appear as a mirror of tumour heterogeneity. In this work, we describe a promising statistical approach to analyze longitudinal cfDNA data, with the purpose of gaining a deeper understanding of the mechanism by which resistance develops in specific patients. While addressing the now classic problem of reconstructing the associated phylogenetic tree, this approach also describes production of cfDNA from the temporal dynamics of cells, in order to best exploit the longitudinal structure of the data.

31 Identification of rare variants predisposing to kidney cancer (Hubert)

Hubert Jean-Noel <hubertjn@fellows.iarc.fr> (1) Affiliations: 1 - International Agency for Research on Cancer (France)

Investigating the genetic predisposition to cancer has large implications. Available data today allow for a more detailed examination of the contribution of rare variants to specific cancer types, which should improve our understanding of the pathogenic mechanisms at play in different cancer types. In addition, such

analyses contribute to better assess the role of rare genetic variation in the 'missing heritability' of complex diseases. Despite poor public awareness, it is estimated that kidney cancer will be the seventh most frequent diagnosed cancer in 2019 in the US, generating significant and rising healthcare costs. The recent GWAS effort has led to the identification of 13 kidney cancer risk loci, accounting for around 10% of the disease heritability. Aiming to further elucidate genes associated with kidney cancer, we acquired 482 exomes from cancer patients with a high risk of genetic predisposition. We performed rare variants tests using both internal (i.e., available in-house) and external (e.g., UK Biobank exomes) ancestry-matched controls, which in particular allowed the identification of genes able to influence genome stability and proliferation capability. Our analyses show the possibility of identifying rare variants predisposing to kidney cancer from a panel enriched in high-genetic risk cases.

32 Evaluating Neanderthal admixture time estimates (Iasi)

Iasi Leonardo Nicola Martin <leonardo_iasi@eva.mpg.de> (1), Peter Benjamin <benjamin_peter@eva.mpg.de> (1) Affiliations: 1 - Max Planck Institute for Evolutionary Anthropology (Germany)

An emergent finding in evolutionary genetics is that gene flow between distinct populations is much more common than previously thought. A question of great interest is when this admixture happened, particularly in the case of gene flow between Neanderthals and modern humans. The most- widely used approaches are based on a recombination clock, i.e. they measure the decay of the length of introgressed fragments over time. Most models make fairly strong assumptions about the data, such as that the recombination rate is known, and that gene flow happened over a very short period of time. Here we present a simulation study where we test the effect of some of these assumptions, to evaluate our knowledge of when Neanderthal gene flow happened, based on present-day genetic data. To allow for ongoing gene flow, we introduce a model where migration times follow a gamma distribution, in which case the admixture tract length distribution has a closed form. However, we find that even if migration persists over thousands of generations, the effect on admixture time inference is small, suggesting an inherent limit to the accuracy that can be obtained when estimating the time of gene flow. Moreover, we find that particularly the accuracy of the recombination map has by far the highest impact on inferred admixture times. Even small deviations can lead to an underestimate of admixture times up to 60%. Based on these results, we find that most attempts to estimate the timing of gene flow from present-day data are likely seriously underpowered. Analyses incorporating ancient DNA (both to calibrate the recombination clock and to have more recent admixture times), will be required to resolve this issue.

33 An efficient method for inferring pedigrees (Jewett)

Jewett Ethan <ejewett@23andme.com> (1), Mcmanus Kimberly <kmcmanus@23andme.com> (1), Freyman Will <willf@23andme.com> (1), Blakkan Cordell <cblakkan@23andme.com> (1), Mountain Joanna <jmountain@23andme.com> (1), Auton Adam <aauton@23andme.com> (1) Affiliations: 1 - 23andMe, Inc. (United States)

Pedigree inference is an important problem in genetics with applications that include rare disease mapping, genetic risk prediction, the validation of self-reported relationships, and the reconstruction of genealogies to address population-genetic questions. We present a composite likelihood method for inferring pedigrees from pairwise identity by descent (IBD) data. The method is similar to constructive approaches that build a pedigree for a set of relatives by iteratively adding one individual at a time. We demonstrate how methodological differences with previous methods can improve accuracy and run time. The new method runs quickly and accurately, providing accurate estimates for large and sparsely genotyped pedigrees.

34 A fast genome chopper to detect strong species decline (Kerdoncuff)

Kerdoncuff Elise <elise.kerdoncuff@mnhn.fr> (1) (2), Achaz Guillaume <guillaume.achaz@mnhn.fr> (3) (4) (5), Lambert Amaury <amaury.lambert@college-de-france.fr> (6) (4) 1 - Institut de Systematique, Evolution, Biodiversite UMR 7205 (France), 2 - Centre interdisciplinaire de recherche en biologie (France),

3 - Atelier de BioInformatique (France), 4 - Centre interdisciplinaire de recherche en biologie (France), 5 - Musee de l'Homme (France), 6 - Laboratoire de Probabilites, Statistique et Modelisation (France)

Only 5% of described species have a conservation status. Methods used to assess conservation status cannot be generalized to all species.Using coalescent theory, we developed a new method to study demography based on the length of compatible blocks along the genome, i.e. blocks of nucleotides within which recombination events are not detectable. From whole-genome data of multiple individuals in a population, we can chop a chromosome into compatible blocks in seconds. Lengths of compatible blocks depend on the frequency of recombination events which is influenced by the ancestral history of the population. Using the distribution of block lengths, we can discriminate a constant population and a declining one. This method can infer a very recent decline of a population from DNA sequences. It could be a new tool to assess conservation status in a wide range of species.

35 A systematic search for intronic elements (Landen)

Landen Gozashti and Russell Corbett-Detig, University of California Santa Cruz

Introns are sequences interrupting genes that must be removed from mRNA before translation, and are a hallmark of eukaryotic genomes. They likely play important roles in genome evolution, but have poorly understood origins (Simmons et al. 2015). Many species exhibit major intron loss events, which probably occur through RNA mediated homologous recombination of cDNA (Lee and Stevens 2016). In contrast, some species exhibit prolific intron gain. Micromonas pusilla, an aquatic picophytoplankton, probably exhibits the most notable recent case of intron gain. Intronic sequences known as introner elements (IEs) colonized the M. pusilla genome in astounding quantities, likely through a mechanism involving DNA transposition (Huff et al. 2016). Contrary to canonical introns, introner elements exhibit conserved sequences and lengths. Similar phenomena are known to exist in fungi (van der Burgt et al 2012; Wu et al. 2017). Although introner elements are known to exist in some species, no study has conducted a systematic search for them. We developed a computational pipeline for intron element detection and implemented it on all Genbank and Refseq assemblies with valid genome annotations available through NCBI (Geer et al. 2010). We report putative novel IE discoveries in several species. Our results suggest that transposons may generate introns on genomic scales in a subset of lineages by co-opting pre-existing splice sites or encoding their own.

36 Inferring fluctuating population size and selection with phylogenetics codon models (Latrille)

Latrille Thibault <thibault.latrille@ens-lyon.org> (1), Lartillot Nicolas <nicolas.lartillot@univ- lyon1.fr> Affiliations: 1 - Laboratoire de Biometrie et Biologie Evolutive - UMR 5558 (France)

Selection in protein-coding sequences can be detected based on multiple sequence alignments using phylogenetic codon models. Mechanistic approaches, grounded on population- genetics first principles, have been recently developed. These so-called mutation-selection models explicitly formalize the interplay between mutation, selection and drift, and return an estimate of the amino-acid fitness landscape, considered static along the phylogeny. They were recently proposed as a null (nearly-neutral) model against which to test for the presence of adaptation (Rodrigue, Lartillot MBE 2016, Bloom, 2016). However, these models rely on the assumption of multiplicative fitness landscapes (no epistasis) and constant population size; they also ignore polymorphism in extant species, with only one sequence representing the whole population. As a result, they return potentially biased estimates. We propose an extended mutation-selection model relaxing some of these assumptions, by accommodating for fluctuating population size and fluctuating mutation rate along the phylogeny, and by modeling polymorphism in extant species. The resulting mechanistic framework allows for a reconstruction of long-term trends in population size along the phylogeny. Simultaneously, it offers a better background for detecting adaptation across large clades, by correcting for local changes in the relative strength of selection and random drift. Finally, our work also points to important theoretical questions about how coding sequences respond to changes in effective population size and to selection.

37 Leverage pleiotropic effects from genome-wide association studies using frequentist and Bayesian sparse group models (Lefranc)

Lefranc Alexandre <alexandre.lefranc@univ-pau.fr> (1), Liquet Benoit <benoit.liquet@univ- pau.fr> (1) Affiliations: 1 - Laboratoire de Mathematiques et de leurs Applications [Pau] (France)

Results from genome-wide association studies (GWAS) suggest that complex diseases are often affected by many variants with small effects, known as polygenicity. Bayesian methods provide attractive tools for identifying signal in data where the effects are small but clustered. For example, by incorporating biological pathway membership in the prior they are able to integrate the ideas of gene set enrichment to identify groups of biologically significant genetic variants. Accumulating evidence suggests that genetic variants may affect multiple different complex diseases, a phenomenon known as pleiotropy. Method: In this work we propose frequentist and Bayesian statistical method to leverage pleiotropic effects and incorporate prior pathway knowledge to increase statistical power and identify important risk variants. We offer novel feature selection methods for the group variable selection in multi-task regression problem. We develop methods using both penalised likelihood methods and Bayesian spike and slab priors to induce structured sparsity at a pathway, gene or single- nucleotide polymorphism (SNP) level. We implement Gibbs sampling algorithms for the Bayesian analysis and an alternating direction method of multipliers (ADMM) algorithm for the penalised regression methods. The performance of the proposed approaches are compared to state-of-the-art variable selection strategies on simulated data sets. Result: The penalised likelihood approaches are computationally efficient using alternating direction method of multipliers algorithm. These approaches perform reasonably well in variable selection but the reconstructed signal is underestimated. The multivariate Bayesian sparse group selection with spike and slab prior performed the best in terms of signal recovery. The Bayesian method provides a natural method for quantifying the variability of the estimated coefficients. Simulation results suggest that when computationally possible the Bayesian estimators should be used. Conclusion: The developed statistical approaches is applied for enriching our insights about the genetic mechanisms of thyroid and breast cancer types. The analysed data come from case-control studies including 3766 SNPs from 337 genes from 10 non-overlapping gene pathways. The thyroid cancer data set includes 482 cases and 463 controls. The breast cancer data set includes 1172 cases and 1125 controls.

38 Go low with ATLAS: maximizing population genetic insight from minimal sequencing depth (Link)

Link Vivian <vivian.link@unifr.ch> (1) (2), Kousathanas Athanasios <athanasios.kousathanas@unil.ch> (3) (2), Hofmanova Zuzana <zuzana.hofmanova@unifr.ch> (1) (2), Reyna Carlos <carlos.reyna@unifr.ch> (1) (2), Pochon Zoe <zoe.pochon@unifr.ch> (1), Blocher Jens <jbloeche@students.uni-mainz.de> (4), Leuenberger Christoph <christoph.leuenberger@unifr.ch> (5), Burger Joachim <jburger@uni-mainz.de> (4), Wegmann Daniel <daniel.wegmann@unifr.ch> (1) (2) Affiliations: 1 - Departement de Biologie, Universite de Fribourg (Switzerland), 2 - Swiss Institute of Bioinformatics (SIB) (Switzerland), 3 - Institute of Bioinformatics, Universite de Lausanne (Switzerland), 4 - Paleogenetics Group, University of Mainz (Germany), 5 - Departement de Mathematiques - Universite de Fribourg (Switzerland)

Many methods in population genomics rely on called genotypes as input. However, especially at low depth, calling genotypes is error-prone, thus uncertain genotypes are usually filtered out based on the genotype quality. But filtering causes biases, often leading to an underestimation of genetic diversity. Alternatively, the issue of genotyping uncertainty can be solved using a probabilistic approach in which hierarchical parameters (e.g. genetic diversity) are inferred by integrating over all possible genotypes at each locus. We here present three tools, based on this philosophy, that quantify key evolutionary quantities of genetic diversity directly from sequence alignments of individuals or populations. First, we quantify heterozygosity within genomic windows under Felsenstein's 1981 substitution model. Second, we measure the pairwise genetic distance between individuals, which can also be used to infer relatedness or perform a Multidimensional Scaling Analysis without genotype calls. Third, we quantify population structure by inferring the deficit in heterozygous genotypes as measured by the inbreeding coefficient. Using simulations as well as downsampling experiments of real data, we show that all these methods perform well even at

very low mean sequencing depth, often at or below 1x. As such, these methods allow to invest in more samples rather than higher depth, and hence to increase statistical power when characterizing evolutionary processes. However, all methods based on genotype likelihoods require these to accurately reflect the genotype uncertainty. For this, base sequencing quality scores must be carefully recalibrated, for which we present a new method particularly suited for low-depth data that does not rely on reference genome data, but exploits homozygous or conserved regions in the genome. All our tools are implemented in our well-documented and user-friendly program ATLAS, which can readily be used in combination with other tools such as ANGSD or GATK. ATLAS is particularly suited for ancient samples that have generally low endogenous DNA content and are affected by Post-Mortem Damage (PMD), a process that causes the replacements of cytosine with thymine and leads to mutations that are not reflective of a sample's diversity. While PMD is usually addressed by removing or down-weighting potentially damaged data, ATLAS explicitly accounts for PMD in the genotype likelihoods, enabling an unbiased and more powerful comparisons between ancient and modern samples. To illustrate the power of ATLAS, we used it to infer the origin of 18 soldiers from a colossal Bronze-age battlefield in norther Germany. This battlefield, which involved thousands of warriors, challenges the view of a lack of large-scale social organization in norther Europe during that era.

39 Demographically explicit scans for genetic barriers (Lohse)

Laetsch Dominik <dominik.laetsch@ed.ac.uk> (1), Aeschbacher Simon <simon.aeschbacher@uzh.ch> (2), Martin Simon <Simon.Martin@ed.ac.uk> (3), Lohse Konrad <klohse@ed.ac.uk> (1) Affiliations: 1 - Institute of Evolutionary Biology [Edinburgh] (United Kingdom), 2 - Department of Evolutionary Biology and Environmental Studies (Switzerland), 3 - University of Cambridge (United Kingdom)

Genome scans for outliers of divergence have largely been based on one dimensional summary statistics, such as dxy or Fst which suffer from a number of well known limitations. Perhaps most fundamentally, such statistics are too coarse to allow for any quantitative link between sequence variation and the population level processes that give rise to it. Thus the interpretations of outlier scans are generally verbal and ignore the large variation inherent in the coalescent process. Here I describe a composite likelihood approach that uses blockwise sequence variation to quantify heterogeneity in both effective gene flow and effective population size along the genome. Tests on simulated data and a reanalysis of divergence between Heliconius melpomene and H. cydno show that this model-based framework i) has greater power than Fst scans to identify genetic barriers that have arisen in the presence of gene flow, ii) is less sensitive to false positives, in particular regions of reduced genetic diversity due to background and positive selection unrelated to speciation and iii) can accommodate variation in recombination rates. Model based scans are flexible and provide estimates of the genome wide distribution of compound parameters of interest which are more easily interpretable in terms of the interplay between demography and selection than simple measures of sequence divergence.

40 The Impact of Population Demography on the Joint Allele Frequency Spectrum of Closely Related Species (Muller)

Muller Rebekka <rebekka.muller@math.uu.se> (1), Kaj Ingemar <ingemar.kaj@math.uu.se> (1), Mugal Carina Farah <carina.mugal@ebc.uu.se> (1) Affiliations: 1 - Uppsala University (Sweden)

Genome-wide polymorphism data of a population can be summarized in the so-called allele frequency spectrum (AFS) that records the frequency distribution of derived alleles. In literature both diffusion-based approaches and coalescent theory are used to represent the AFS. Starting from a diffusion-based approach the non-equilibrium AFS can be derived using a Poisson stochastic integral. The time dependence plays an important role especially when modelling the joint AFS of closely related species, since in this case the frequency spectrum is composed of shared ancestral polymorphisms in the populations arising through mutations before speciation and of lineage-specific polymorphisms appearing after the speciation event. So far, using diffusion-based approaches, the joint AFS can only be approximated numerically without an explicit formula. We use a time-dependent Poisson random field model to derive an analytical representation of the joint AFS of closely related species. Moreover we address the typical assumption of

equal population sizes of the related species by allowing for different population demographies and incorporating gene flow between the populations. We investigate the impact of possible patterns of asymmetry that result from the different population sizes.

41 Flexible Markov random field priors for birth-death phylogenetic tree models (Magee)

Magee Andrew <andyfmagee@gmail.com> (1), Hohna Sebastian <hoehna@lmu.de>, Leache Adam <leache@uw.edu>, Minin Vladimir <vminin@uci.edu> Affiliations: 1 - Department of Biology, University of Washington (United States)

Studying variation in the processes of speciation and extinction enables researchers to examine the patterns and processes that shape the diversity of life on earth. Birth-death processes provide a model-based framework in which such studies can be accomplished by asking questions about changes in the birth rate of lineages in a phylogenetic tree. Early approaches to studying temporal variation in birth rates using birth-death process models faced difficulties due to both unsampled taxa and the limited number of birth rate functional forms that could be considered in an analysis. Despite the development of approaches that ameliorate some of these concerns, the development of a truly flexible time-varying birth-death process model remains an open question. We use a piecewise-constant birth-death process model, combined with both Gaussian Markov random field (GMRF) priors and Horseshoe Markov random field (HSMRF) priors, to approximate arbitrary changes in birth rate through time. We implement these models in the statistical phylogenetic software platform RevBayes, allowing us to jointly estimate birth-death process parameters, phylogeny, and nuisance parameters in a Bayesian framework. We test both GMRF and HSMRF models on a variety of simulated diversification scenarios, and then apply them to a species-level and an epidemiological dataset. We find that both models are capable of inferring variable diversification rates and of correctly rejecting variable models in favor of effectively constant models, and that in general the HSMRF-based model enjoys higher precision than its GMRF counterpart, without sacrificing much accuracy. Applied to a macroevolutionary dataset of the Australian gecko family Pygopodidae, our models detect a speciation-rate decrease in the last 12 million years. Applied to an infectious disease phylodynamic dataset of sequences from HIV subtype A in Ukraine, our models detect a complex pattern of variation in the rate of infection.

42 SigNet: Identifying mutational processes in cancer using neural networks (Maretty)

Maretty Lasse <lasse.maretty@clin.au.dk> (1), Besenbacher Soren <besenbacher@clin.au.dk> (1) Affiliations: 1 - Aarhus University (Denmark)

Cancer is caused by the accumulation of mutations arising from multiple processes such as exposure to different mutagens and defects in DNA repair mechanisms, and information about such processes carries both etiological and prognostic relevance. The mutational processes contributing to a set of cancer samples can be identified and characterised by factoring a mutation matrix, which contains the number of observed mutations for each mutation type for every sample, using non-negative matrix factorization (NMF). In the NMF approach, the mutation type is typically defined using only the immediate sequence context and substitution type as the number of parameters grows exponentially in the number of genomic features used to define the type. However, the activity of different processes is known to vary across a range of additional genomic features such as longer range sequence context, transcription level, and replication timing, and hence potentially valuable information for discriminating different processes is lost in the standard NMF approach. We here propose a new variant of NMF in which the latent mutational susceptibility of a genomic site for a given mutational process is modelled as a nonlinear function of the features (parameterized by a neural network) instead of being estimated explicitly for each mutation type. This parameterization renders the number of parameters linear in the number of features (for a fixed size neural network and number of processes) and hence enables inclusion of an arbitrary number of features. Maximum likelihood and Bayesian estimation in this model have been implemented in a software package called SigNet. We explore the model's performance using different sequence context sizes on whole genome sequencing data from 2,500 cancer samples and compare with the classical NMF approach using different metrics. Preliminary results show that using a longer sequence context enables identification of more mutational signatures. We

expect the model to both find use as a tool for discovering new mutational processes and for use as a mutational background model for detection of cancer drivers.

43 How the quantitative genetics toolbox can help evolutionary physiology? A case study of the parasitoid wasp venom. (Mathe-Hubert)

Mathe-Hubert Hugo <hugomh@gmx.fr>, Monrolin Marie <marie.monrolin@hotmail.fr>, Le Goff Isabelle <isabelle.legoff@inra.fr>, Poire Marylene <marylene.poirie@inra.fr>, Malausa Thibaut <tmalausa@inra.fr> (1) Affiliations: 1 - Institut Sophia Agrobiotech [Sophia Antipolis] (France)

Physiology, being the link between the genome and selected phenotypic traits, must be studied in an evolutionary perspective to better understand i) selection pressures that shape it and ii) the constraints it imposes on the evolution of phenotypic traits and genomes. Here we studied how the venom of an endoparasitoid wasp evolved in response to a host shift. Since eggs of endoparasitoids perform their development within their hosts, their physiology is involved in close antagonistic interactions with the physiology of their hosts. These hosts die when the parasitoids successfully develop. One of the critical components of this interaction is the venom injected by mothers, along with the egg in each host. This venom not only prevents the destruction of the developing eggs by the host immune system, but also it tunes the host physiology to maximise its nutritive value for the developing offspring. Using an experimental evolution studying the change in the venom composition in response to a host shift, I illustrate the use of quantitative genetics tools such as the multivariate QST, the breeding value, the G matrix, and the selection gradients. I used these tools to identify the venom components which amount changed in response to the host shift, to assess whether these changes were adaptive or not, whether they were plastic or genetic, and whether there is some individual variability in the reaction norm. Finally, these tools identified which venom components affected the extinction dynamic of the experimental populations.

44 Testing for Hardy-Weinberg equilibrium in structured populations using genotype or low depth next generation sequencing data (Meisner)

Meisner Jonas <jonas.meisner@bio.ku.dk> (1), Albrechtsen Anders <albrecht@binf.ku.dk> (1) Affiliations: 1 - Bioinformatics Centre, Department of Biology, University of Copenhagen (Denmark)

Testing for deviations from Hardy-Weinberg equilibrium (HWE) is a common practice for quality control in genetic studies. Variable sites violating HWE may be identified as technical errors in the sequencing or genotyping process, or they may be of particular evolutionary interest. Large scale genetic studies based on nextâgeneration sequencing (NGS) methods have become more prevalent as cost is decreasing but these methods are still associated with statistical uncertainty. The large scale studies usually consist of samples from diverse ancestries that make the existence of some degree of population structure almost inevitable. Precautions are therefore needed when analysing these data set, as population structure causes deviations from HWE. Here we propose a method that takes population structure into account in the testing for HWE, such that other factors causing deviations from HWE can be detected. We show the effectiveness of PCAngsd in low depth NGS data, as well as in genotype data, for both simulated and real data set, where the use of genotype likelihoods enables us to model the uncertainty.

45 Variable prediction accuracy of polygenic scores within an ancestry group (Mostafavi)

Mostafavi Hakhamanesh <hsm2137@columbia.edu> (1), Harpak Arbel <ah3586@columbia.edu> (1), Conley Dalton <dconley@princeton.edu> (2), Pritchard Jonathan <pritch@stanford.edu> (3), Przeworski Molly <mp3284@columbia.edu> (1) Affiliations: 1 - Columbia University (United States), 2 - Princeton University (United States), 3 - Stanford University (United States)

Fields as diverse as human genetics and sociology are increasingly using polygenic scores based on genome-wide association studies (GWAS) for phenotypic prediction. However, recent work has shown that polygenic scores have limited portability across groups of different genetic ancestries, restricting the contexts in which they can be used reliably and potentially creating serious inequities in future clinical

applications. Recent discussion about this challenge has focused primarily on the impact of differences in linkage disequilibrium patterns and allele frequencies across human populations that arose from their distinct demographic and recombination histories. Here, using the UK Biobank, we show that prediction accuracy can differ markedly even across groups with highly similar ancestry. First, focusing on BMI, blood pressure and years of schooling as examples, we illustrate that prediction accuracy varies across groups that differ by characteristics such as age, sex, and socio-economic status, even when they share similar ancestry. We further demonstrate that the prediction accuracy of a polygenic score depends on whether the GWAS is conducted among unrelated individuals or within sibling pairs, even when both analyses are matched to have similar sampling noise. We derive analytic results that clarify how indirect parental effects (Ôgenetic nurtureÕ) and assortative mating can lead to such differences in the prediction accuracies of polygenic scores. Our findings highlight both the complexities of interpreting polygenic scores and underappreciated obstacles to their broad use.

46 Using time-dependent Poisson random field models for polymorphism-aware expression of dN/dS (Mugal)

Mugal Carina Farah <carina.mugal@ebc.uu.se> (1), Kaj Ingemar <ingemar.kaj@math.uu.se> (1) Affiliations: 1 - Uppsala University (Sweden)

The ratio of non-synonymous over synonymous sequence divergence, dN/dS, is a widely-used estimate of the non-synonymous over synonymous fixation rate ratio r, which measures the extent to which natural selection modulates protein sequence evolution. Its computation is based on a phylogenetic approach and computes sequence divergence of protein-coding DNA between species, traditionally using a single representative DNA sequence per species. This approach ignores the presence of polymorphisms and relies on the indirect assumption that new mutations fix instantaneously, an assumption which is generally violated and reasonable only for distantly related species. The violation of the underlying assumption leads to a time-dependence of sequence divergence, and biased estimates of r in particular for closely related species, where the contribution of ancestral and lineage-specific polymorphisms to sequence divergence is substantial. Recent efforts to jointly analyse polymorphism and divergence are so far frequently based on stationary Poisson random field models, which assume that lineage sorting is completed between species. To address the impact of incomplete lineage-sorting, we use a time-dependent Poisson random field model and derive an analytical expression of dN/dS as a function of divergence time and sample size. This mathematical treatment enables us to show that the joint usage of polymorphism and divergence data can assist the inference of selection for closely related species. Moreover, our analytical framework provides the basis for unbiased estimation of r for closely related species.

47 Whole-genome simulations within population-scale pedigrees (Nelson)

Nelson Dominic <nelson.dominic@gmail.com> (1), Kelleher Jerome <jerome.kelleher@well.ox.ac.uk> (2), Ragsdale Aaron <aaron.ragsdale@mail.mcgill.ca> (1), Mcvean Gil <gil.mcvean@bdi.ox.ac.uk> (2), Gravel Simon <simon.gravel@gmail.com> (1) Affiliations: 1 - McGill University and Genome Quebec Innovation Centre, Montreal, Quebec, Canada (Canada), 2 - Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK (United Kingdom)

With the advent of increasingly high-performance genetic simulation software, large cohorts can now be simulated to aid in the understanding of demographic and evolutionary history, and in the discovery of disease associations. However, as simulated cohorts become larger and more complex, they require more sophisticated models in order to reflect realistic patterns of relatedness and diversity. Coalescent simulators have been extensively used for this purpose due to their computation efficiency and well-developed mathematical theory. However, coalescent theory exhibits significant distortions of sample relatedness and the distribution of IBD when sample size is large or when simulating long regions. The msprime coalescent simulation software has recently been extended to allow Wright-Fisher simulations, which do not share these biases, and allow large whole-genome datasets to be generated. But in spite of these improvements the Wright-Fisher model remains a highly idealized representation of real human pedigrees, which are shaped by complex effects such as assortative mating, inbreeding, and isolation-by-distance. To better

understand what effects these have on present-day diversity, we further extend msprime to allow simulations to take place within a pre-specified pedigree. This has several advantages. First, simulations can make use of an increasing number of large genealogical datasets, some of which contain several million individuals, and which provide detailed insights into recent human evolution. Second, pedigrees with desired characteristics can be generated separately in order to isolate the effects of a particular pedigree structure. In either case pedigrees of any size can be used, with simulations continuing under the Wright-Fisher or coalescent models once the founders of the pedigree have been reached. We present here the results of a simulation study performed using a population-scale pedigree for the province of Quebec, and show how it differs from simulations performed in a panmictic population with a similar high-level demographic history. We further present a preliminary investigation of the effects of inbreeding and outbreeding on the distribution of IBD in large cohorts.

48 Genetic algorithm for demographic inference from the allele frequency spectrum (Noskova)

Noskova Ekaterina <ekaterina.e.noskova@gmail.com> (1), Ulyantsev <vl.ulyantsev@gmail.com> (1), Dobrynin Pavel <pdobrynin@gmail.com> (2) (3)

Affiliations: 1 - ITMO University, St. Petersburg, Russia (Russia), 2 - Smithsonian Conservation Biology Institute, Center for Species Survival, National Zoological Park, Washington, D.C., USA (United States), 3 - Theodosius Dobzhansky Center for Genome Bioinformatics, Saint Petersburg State University, St. Petersburg, Russia (Russia)

Understanding the roles of demography and selection in the formation of species and divergence of populations are central problems in population genetics. Records of population history are imprinted in the genomes of individuals within species and can be inferred using a variety of algorithmic and statistical methods. Recently, with the increasing generation of whole genome data from populations through next-generation sequencing (NGS) technologies, it has become possible to explore complex and parameter-rich demographic histories, which includes such events as migration, population splits and changes in the effective size of populations over time. The allele frequency spectrum (AFS) - the joint distribution of allele frequencies in one or more populations, is one of the most convenient and popular presentations for summarizing genetic information across the genome. Much research has been devoted to the analysis of the allele frequency spectrum and its dependence on the demographic history of populations. This has led to several methods for simulating the expected allele frequency spectrum from a proposed demographic model, such as those implemented in the programs ,a,i and moments. These methods simulate AFS under a variety of researcher-specified demographic models and estimate the best model and associated parameters using likelihood-based optimizations. However, such algorithms are based on local search algorithms, which have some limitations and can be ineffective in practice (i.e., they made not find the global optimum). Currently, there are no known algorithms to perform global searches of demographic models with a given AFS. The genetic algorithm is one of the most efficient heuristic algorithms for global searches of complex and rich parameter space. It is based on the principle of evolution and its versatility has led to its wide application, including the reconstruction of phylogenetic trees, ancestral genome composition inference, and evolutionary biology in general. We developed a new method based on the genetic algorithm for unsupervised demographic model inference from an observed allele frequency spectrum. This method uses the existing solutions for simulating an allele frequency spectrum from a given demographic model, namely ,a,i or moments. Our method supports up to 3 populations and has been implemented in the software, GADMA (Genetic Algorithm for Demographic Model Analysis). The effectiveness of the method was tested on three empirical data sets: modern humans, Euphydryas gillettii butterflies and Scotobleps gabonicus frogs. In each example, we found that GADMA inferred a demographic model close to or even better than the one that was previously reported. Moreover, GADMA is able to infer multiple demographic models at different local optima close to the global one, providing a larger set of possible scenarios to further explore demographic history.

49 Estimating Coalescent Root-Subtrees (Otto)

Otto Moritz <moritz_otto@gmx.net> (1) (2), Wiehe Thomas <twiehe@uni-koeln.de> (1) Affiliations: 1 - Institut fur Genetik, Universitat zu Koln (Germany), 2 - Mathematisches Institut, Universitat zu Koln (Germany)

In population genetic applications, it is often of interest to estimate properties of a genealogical tree from a limited number of single nucleotide polymorphisms (SNPs). In case of binary coalescent trees of finite size, one such goal is to determine the left and right subtrees of the root and the respective leaf subsets. A simple estimate uses a version of 2-means clustering which maximizes the Hamming distance between the two cluster centroids. However, this solution unnecessarily neglects available information. We define an improved estimator, conditioning on the sample allele frequency of the available SNPs. As an application we show how our estimator can be used to calculate 'topological linkage disequilibrium' between two genetic loci [1]. [1] Wirtz J , Rauscher M, Wiehe T (2018) Theo Pop Biol 124:41

50 Efficient variance components analysis across millions of genomes (Pazokitoroudi)

Pazokitoroudi Ali <alipazoki@ucla.edu> (1), Wu Yue <wuyue0715@ucla.edu> (1), S. Burch Kathryn <kathy.s.burch@gmail.com> (2), Hou Kangcheng <kangchenghou@gmail.com> (3), Zhou Aaron <aaronzhouqian@ucla.edu> (1), Pasaniuc Bogdan <pasaniuc@ucla.edu> (4) (5) (3), Sankararaman Sriram <sriram.sankararaman@gmail.com> (4) (5) (1) Affiliations: 1 - Department of Computer Science, UCLA, Los Angeles, California (United States), 2 - Bioinformatics Interdepartmental Program, UCLA, Los Angeles, California (United States), 3 - Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, UCLA, Los Angeles, California (United States), 4 - Department of Computational Medicine, David Geffen School of Medicine, UCLA, Los Angeles, California (United States), 5 - Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, California (United States)

Variance components analysis has emerged as a powerful tool to probe the genetic basis of complex traits, with applications ranging from heritability estimation to association mapping.While the ability to fit flexible variance component models to large-scale datasets is essential to obtain accurate and novel insights into genetic architecture, fitting such models requires scalable algorithms. Approaches for estimating variance components typically search for parameter values that maximize the likelihood or the restricted maximum likelihood (REML). Despite a number of algorithmic improvements computing REML estimates of the variance components on data sets such as the UK biobank (around 500,000 individuals genotyped over a million common SNPs) remains challenging. The reason is that methods for computing these estimators typically perform repeated computations on the input genotypes. Here, we present a new algorithm for multiple variance components estimation which is a randomized version of Haseman-Elston regression. Our proposed algorithm is accurate and highly efficient -- requiring only a few hours to estimate hundreds of variance components on a million individuals genotyped at a million SNPs. Across a wide range of simulations, the ability of our method to fit multiple variance that account for frequency and LD-dependent effects yields unbiased estimates of genome-wide SNP heritability. We illustrate the utility of our method by analyzing 22 complex traits across about 300,000 individuals in the UKBiobank. Relative to our estimates, other methods that can be applied at scale, such as stratified LD-score regression and SumHer, yield SNP heritability estimates that are higher by 2.5% and 25 % on average. Furthermore, we partitioned heritability across bins of minor allele frequency (MAF) and LD and observed that, SNPs with lower levels of LD tend to have higher heritability enrichment for both common (seven-fold more enrichment on average over 22 traits) as well as low-frequency SNPs (eight-fold more enrichment on average over 22 traits) across all traits consistent with reports of the impact of negative selection on these traits. Comparing the quartile with the lowest LD score to the quartile with the highest LD score, height showing similar increase in heritability enrichment for common and low-frequency SNPs (7.7 fold and 5.6 fold for low-frequency and common SNPs) while systolic blood pressure shows a greater increase in the low-frequency SNPs relative to common SNPs (18 fold for low-frequency vs 5 fold for common SNPs).

51 Modeling ancient DNA damage to estimate present-day DNA contamination (Peyregne)

Peyregne Stephane <stephane_peyregne@eva.mpg.de> (1), Peter Benjamin <benjamin_peter@eva.mpg.de> (1) Affiliations: 1 - Max Planck Institute for Evolutionary Anthropology (Germany)

Present-day DNA contamination is a recurrent issue in ancient DNA studies because of the low DNA content in historical and archaeological material. Here, we present a method to quantify the amount of present-day DNA contamination. This method is based on a Hidden Markov Model of the patterns of post-mortem damage along ancient DNA fragments. As DNA degrades, breaks in the DNA lead to single-stranded regions that are susceptible to hydrolytic damage, resulting in the accumulation of nucleotide substitutions through time. Our Hidden Markov Model makes use of these substitutions to map single-stranded regions along the sequence of an ancient DNA fragment. To distinguish between sequences from ancient and modern sources, we compare this model to one that assumes a uniform distribution of substitutions along the sequence, as expected from polymorphisms or sequencing errors for present-day DNA fragments. This method has two main advantages over other approaches that aim to quantify the amount of present-day DNA contamination. First, we do not rely on prior knowledge of the genetic relationship between contaminant and endogenous DNA, which makes it possible to estimate contamination even when this relationship is unknown or when they do not differ substantially (e.g. ancient DNA from early modern humans). Second, we show that the method has the power to estimate contamination from less than 10,000 sequencing reads, making it particularly useful for analysing badly preserved specimens with low coverage data.

52 Estimation of relatedness in ancient populations (Popli)

Popli Divyaratan <divyaratan_popli@eva.mpg.de> (1), Peyregne Stephane <stephane_peyregne@eva.mpg.de> (1), Skov Laurits <laurits_skov@eva.mpg.de> (1), Iasi Leonardo <leonardo_iasi@eva.mpg.de> (1), Grote Steffi <steffi_grote@eva.mpg.de> (1), Meyer Matthias <mmeyer@eva.mpg.de> (1), Herraez David Lopez <david_lopez@eva.mpg.de> (1), Hajdinjak Mateja <mateja_hajdinjak@eva.mpg.de> (1), Slon Viviane <viviane_slon@eva.mpg.de> (1), Kelso Janet <kelso@eva.mpg.de> (1), Paabo Svante <paabo@eva.mpg.de> (1), Peter Benjamin <benjamin_peter@eva.mpg.de> (1) Affiliations: 1 - Max Planck Institute for Evolutionary Anthropology (Germany)

Identifying related individuals is one of the key applications of genetics, as related individuals need to be removed for many analyses, and we may learn about social and cultural practices. In ancient DNA studies, this is often difficult because the genotypes determined from ancient individuals are typically sparse, suffer from ascertainment bias and may partly derive from present-day DNA contaminating the experiments. Most current approaches to estimate relatedness between individuals are allele-frequency based, and have significant drawbacks for the analysis of ancient DNA in that they require intermediate or high (>1x) coverage, large reference panels from related populations or large number of diploid called sites. Here, we present a method to infer relatedness from low- coverage data. Our method models the fact that every genome is a mosaic of fragments inherited from various ancestors, and closely related individuals will share large chunks of their chromosomes identical by descent (IBD). We develop a Hidden Markov Model to identify IBD fragments shared between pairs of ancient samples, and use this model to infer their degree and nature of relatedness. We evaluate this method on simulations and apply it to 31 DNA libraries prepared from 15 Neanderthal specimens. These libraries have been used to capture 713,000 informative sites; the coverage per specimen ranges from 0.01 to 1.58. Preliminary results suggest that three of the specimens come from the same individual, and that one specimen is closely related to four other specimens. Although further work is required to elucidate the exact relatedness among these individuals, we show that relatedness inference is feasible even for data with coverages significantly less than 1x.

53 What generates diversity in regions of low recombination? (Pouyet)

Gilbert Kim J. <kim.gilbert@iee.unibe.ch> (1), Pouyet Fanny <fanny.pouyet@gmail.com> (1), Peischl Stephan <stephan.peischl@bioinformatics.unibe.ch> (2), Excoffier Laurent <laurent.excoffier@iee.unibe.ch> (1) Affiliations: 1 - University of Bern (Switzerland), 2 - Interfaculty Bioinformatics Unit (UNIBE) (Switzerland)

Genome-wide variation in recombination impacts neutral genetic diversity since selection has a greater impact on linked polymorphism when sites occur in regions of low recombination(Charlesworth et al. 1995). Overall, diversity varies accordingly with recombination rate due to linkage with negatively selected sites. This phenomenon, termed background selection, is a driving force in humans impacting up to 85% of the genome (Pouyet et al., 2018). However, in regions of extremely low recombination diversity is larger than expected under a model of background selection only. We thus investigate whatother evolutionary processes could be affecting diversity in these regions. Previous work has investigated the possible role of associative over-dominance in driving neutral genetic diversity and its interaction with background selection (Charlesworth et al.2009, Nordborg & Charlesworth 1996). Associative over-dominance (AOD)corresponds to the increase in neutral diversity when these neutral variants are linked to partially or fully recessive deleterious alleles (i.e. selection acts against mutant homozygotes only, or more strongly). However, much of this work is limited either by single-locus models or only cases of no recombination. Here, we combine both simulations and analyses of human data to examine the process ofAOD. Simulations clearly show AOD to be a process occurring in regions of low recombination, and we proceed to investigate whether similar regions exist in the human genome. We found candidates such as the major histocompatibility complex (MHC) that is 3Mb on chromosome 6 where diversity is even higher is low versus high recombination regions, suggesting that AOD occurs in humans. Charlesworth B, Betancourt AJ, Kaiser VB, Gordo I (2009) Genetic recombination and molecular evolution.Cold Spring Harbor Symposia on Quantitative Biology, vol. LXXIV. Nordborg M, Charlesworth B (1996) The effect of recombination on background selection.Genet. Res.67: 159-174. Pouyet F, Aeschbacher S, Thiery A, Excoffier L (2018)Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences.eLife, 7:e36317.

54 New models to infer spatiotemporal patterns of adaptation and migration (Racimo)

Racimo Fernando <fracimo@bio.ku.dk> (1) Affiliations: 1 - University of Copenhagen (Denmark)

Evolutionary genomics has - in the last two decades - unearthed a rich history of population dynamics while studying species across the planet, including complex patterns of divergence, migration and admixture among differentiated groups. Yet genome-wide studies of selection often assume simple dynamics (e.g. a 3-population tree) or aim to control for complex dynamics without explicitly modeling them (e.g. using the genome-wide covariance matrix). This prevents these rich historical insights from bearing on our understanding of past adaptive events in the organisms we study. We have developed several new methods to explicitly account for complex population dynamics while looking for loci with footprints of positive selection, and applied them to present-day and ancient genomic datasets. These include programs that can use admixture graphs and latent mixed-membership models to pinpoint exactly where and when in the history of a species a particular selective event took place, while explicitly modeling migration and admixture processes. Finally, we have also been working on modeling ancestry and selection as spatiotemporal dynamic processes - borrowing insights from environmental and paleoclimatic research to uncover the drivers of migration and adaptation (e.g. temperature, vegetation or pathogens), while accounting for the fact that these drivers may have also changed over time and space.

55 Inferring deep population structure in Africa using linkage disequilibrium (Ragsdale)

Ragsdale Aaron <aaron.ragsdale@mail.mcgill.ca> (1), Gravel Simon <simon.gravel@mcgill.ca> (2) Affiliations: 1 - Department of Human Genetics, McGill University (United States), 2 - Department of Human Genetics, McGill University (Canada)

Throughout history, populations have expanded and contracted, split and merged, and exchanged migrants. Because these events affect contemporary genetic diversity, we can learn about history by comparing predictions from evolutionary models to genetic data. We developed an approach to rapidly compute predictions for a wide range of diversity measures for many populations with complex demography, including for patterns of shared linkage disequilibrium between populations, and we derived unbiased estimators for these same statistics from unphased sequencing data. These methods are packaged together in a likelihood-based inference framework for multi-population demographic inference. Multi-population linkage disequilibrium statistics are informative about deep population structure and archaic admixture, even when there is no available genetic data from diverged human lineages. Using our approach, we find evidence for substantial and possibly long-lasting admixture from a deeply diverged lineage within Africa. We further infer multi-population demographic models for a large set of diverse African populations, which reveals population structure that predates the split of Eurasian and African populations. Our results underline the need for demographic models that better describe population structure within Africa, which can strongly affect predicted patterns of linkage disequilibrium and genetic diversity. More broadly, we highlight that by studying a wide variety of diversity statistics we can assess the robustness of commonly used evolutionary models and build more informed models of demographic history.

56 Fast computation and duality for tree sequence statistics (Ralph)

Ralph Peter <plr@uoregon.edu> (1), Kelleher Jerome <jerome.kelleher@well.ox.ac.uk> Affiliations: 1 - University of Oregon (United States)

To every statistic of the allele frequency spectrum corresponds a statistic of tree shape that is its conditional expectation given the population's tree sequence, averaging over infinite-sites mutations. I will describe a general framework to define and efficiently compute a very general class of such mutations, for both genome sequence and for tree shape. This allows us to compute statistics of very large genomic datasets quickly, such as pairwise divergences between tens of populations in windows along the genome, or GWAS on many traits at Biobank scale. The duality also makes it possible to identify where in time the signal from statistics along the genome derive from.

57 An improved recalibration model for accurately estimating genetic diversity from low and ancient sequencing data (Reyna)

Reyna Carlos <carlos.reyna@unifr.ch> (1) (2), Link Vivian <vivian.link@unifr.ch> (3) (2), Wegmann Daniel <daniel.wegmann@unifr.ch> Affiliations: 1 - Departement de Biologie, Universite de Fribourg (Switzerland), 2 - Swiss Institute of Bioinformatics (SIB) (Switzerland), 3 - Universite de Fribourg (Switzerland)

There is growing evidence that genetic diversity can be accurately inferred from low-depth next-generation sequencing data if genotyping uncertainty is properly accounted for. Methods that do so, however, assume that the quality scores provided are accurate, i.e. that they properly reflect the associated sequencing error rates. Yet, this is rarely the case since raw base quality scores provided by sequencing machines are typically biased. Proper recalibration of the quality scores is therefore essential for most downstream analysis, and in particular when working with noisy data usually obtained from ancient samples. Current algorithms for quality score recalibration usually require a set of sites with known genotypes. This is typically obtained by using a reference sequence and masking all sites known to be variable in a population or species. But that knowledge is often lacking, especially when working with non-model organisms or ancient samples for which modern diversity might not be reflective. To overcome this short-coming, we recently introduced a new method that 1) only requires a set of sites at which the sample is known to be homozygous (but not its genotype) and 2) is readily extended to additional covariates beyond the original quality score, for instance the position within the sequencing read, the nucleotide contex or mapping quality. Examples of regions for which a sample can be assumed homozygous include the mtDNA, the X-chromosome in male mammals, or sites highly conserved among species. Here we assessed the power of our method by downsampling sequencing data of ancient human genomes to various depth, from which we then estimated heterozygosity after recalibrating the quality scores. As faulty quality scores affect diversity estimates at

low-depth more strongly, consistent estimates are reflective of proper recalibration. Using such comparisons we show that recalibrating quality scores based on sites highly conserved among mammals as reflected by the RS-Score (also called GERP score) results in very accurate estimates of genetic diversity. However, that accuracy was strongly affected by the number of homozygous sites used to learn recalibration parameters, with too few sites resulting in noisy recalibration and too many sites implying less stringent conservation and hence overestimation of the error rates. When using about 10 Mb (corresponding to an RS-score of 3.9), however, heterozygosity was estimated very accurately down to a mean sequencing depth below 1x. We finally used this method to evidence lower genetic diversity of mesolithic hunter-gatherer individuals compared to neolithic farmers. By comparing diversity in coding versus non-coding regions we further evidence differences in the strength of purifying selection acting within these populations

58 Inferring runs of homozygosity from low coverage (ancient) DNA data (Ringbauer)

Harald Ringbauer, John Novembre, Matthias Steinrücken. Affiliations: University of Chicago

The ancient DNA revolution has delivered spectacular new insights into human population history. Here we present work on a novel method to detect long runs of homozygosity (ROH) for such data. These blocks are the genetic signposts of consanguineous matings, and as such, the frequency and length distribution of ROH blocks yields insight into recent population structure. For high coverage present-day datasets, one can identify ROH by scanning for regions that lack heterozygote markers. But this strategy frequently fails for ancient individuals: The typically low read depth (<5x) makes reliable diploid genotype calling infeasible. To overcome this limitation, our new method makes use of linkage disequilibrium information from a panel of modern reference haplotypes under a Hidden Markov Model (HMM). It scans for long stretches of genome that are copies from various single haplotypes in the reference panel. We tested an implementation of the method, termed HAPSBURG (Haplotype Block Sharing by uninterrupted recent Genealogy), on simulated data. Our results show that it works for coverage down to about 0.5x for a commonly used aDNA data type (1240K SNP capture data). I will also present first example applications to data from low coverage ancient humans, which demonstrate the ability of HAPSBURG to robustly identify ancient individuals that are offspring of consanguineous matings.

59 Site-specific detection of adaptive evolution in protein-coding DNA using a Bayesian mutation- selection model (Rodrigue)

Rodrigue Nicolas <nicolas.rodrigue@carleton.ca> (1), Lartillot Nicolas <nicolas.lartillot@univ- lyon1.fr> (2) Affiliations: 1 - Dept. of Biology, Carleton University (Canada), 2 - Laboratoire de Biometrie et Biologie Evolutive - UMR 5558 (France)

Statistical modeling of the long-term evolution of protein-coding DNA is an active area of research in molecular phylogenetics. Recent works have adopted a mutation-selection framework, whereby the substitution process is specified from a set of parameters controlling a point-mutation process, and another (potentially site-specific) set controlling the probability of fixation of mutations. We have previously discussed using such a framework as a more relevant null model against which to test for features of molecular evolution (Rodrigue, Philippe & Lartillot, PNAS, 2010), and have used simulations to illustrate the potential of the approach when interested in uncovering genes exhibiting global signatures of adaptation (Rodrigue & Lartillot, MBE, 2017). Here, we present recent extensions along these lines, which allow for site-specific detection of adaptation within the mutation-selection framework. We illustrate the potential of the approach with simulations, and an application on a previously well-studied dataset.

60 Using positional information for predicting transcription factor binding sites (Romero)

Romero Raphael <raphael.romero@umontpellier.fr> (1) (2), Marin Jean-Michel <jean-michel.marin@umontpellier.fr> (1), Lebre Sophie <sophie.lebre@umontpellier.fr> (3) (1), Lecellier Charles <charles.lecellier@igmm.cnrs.fr> (4), Brehelin Laurent <brehelin@lirmm.fr> (2) Affiliations: 1 - Institut

Montpellierain Alexander Grothendieck (France), 2 - Laboratoire d'informatique de Robotique et de Microelectronique de Montpellier (France), 3 - Univ. Paul-Valery-Montpellier 3 (France), 4 - Institut de genetique moleculaire de Montpellier (France)

Transcription factors (TF) play a central role in the mechanism of tran- scription. These proteins bind the DNA sequence at particular binding sites. Binding sites are resumed in probabilistic model known as binding motifs or Position Weight Matrix [1]. Such motif can be used to compute binding affinities and to identify potential binding sites of the associated TF. However this approach has usually low accuracy, with lot of false positives. In order to provide more accurate predictions of TF binding sites, we recently proposed a method that uses the fact that TFs do not bind DNA in an isolated way but in combination with others TFs. This method, named TFcoop [2], bases its prediction upon the binding affinity of the target TF as well as any other TF identified as cooperating with the target. Given a set of positive and negative sequences obtained from ChIP-seq experiments, TFcoop uses a logistic model trained with a LASSO regularization [3] for selecting the cooperating TFs. The approach outperforms the classical TF binding prediction methods and allows the identification putative cooperating TFs. Here we introduce a more refined method that complements the TF binding affinity with positional information of the potential binding sites. Our aim is to study the importance of such information for predicting TF binding. In order to consider several subsequences of the original sequence while avoiding prohibitive computing time, we developed a segmentation algorithm based on a lattice. The selected subsequences are used to create new features that are added to the logistic model. In addition, by centering the sequences on the binding site of the targeted TF, this segmentation algorithm enables us to consider the relative position between TFs' binding sites. This information is particularly relevant for specific 1biological questions, and can be used in different classification problems like cell type specific TF binding. The relative position information already pointed out cell-type specific cooperativity between TFs in some experiments. We are currently exploring 12 TFs on 90 ChIP-seq experiments. References [1] Wyeth W. Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics, 5(4):276- 287, April 2004. [2] Jimmy Vandel, Oceane Cassan, Sophie Lebre, Charles-Henri Lecellier, and Laurent Brehelin. Probing transcription factor combinatorics in different promoter classes and in enhancers. March 2018. [3] Trevor Hastie, Sami Tibshirani, and Jerome Friedman. Elements of Statis- tical Learning: data mining, inference, and prediction. 2nd Edition., 2009.

61 Distinguishing pedigree relationships using multi-way identity by descent sharing and sex-specific genetic maps (Sannerud)

Sannerud Jens <jgs267@cornell.edu> (1), Qiao Ying <yq76@cornell.edu> (1), Williams Amy <awilliams@cornell.edu> (1) Affiliations: 1 - Cornell University, Department of Biological Statistics and Computational Biology (United States)

Pedigree structures capture data of wide utility to geneticists,including information fortrait mapping, heritability estimation, and the study of parent-of-origin effects. Traditionally, pedigrees had to be self-reported, limiting their application to cohorts where relationships were explicitly recorded. With the rise of biobank-scale studies, the opportunity to infer numerous unreported pedigrees latent within large datasets has become manifest. Most relatedness inference methods produce the degree of relatedness between two individuals; however, this degree is ambiguous with regard to the specific pedigree that relates the samples. Although it is simple to separate the two possibilities for first-degree relatives (parent-child and full sibling pairs), for even one degree higher relationship distance Ñ second- degree relationships Ñ current methods have a greatly reduced capacity for inference. We present a method that can not only discriminate between pedigree relationships for pairs of a given degree, but can also report the sex of the unsampled individual that links the pair. This method, CREST (Classification of Relationship Types) relies on distinct patterns of autosomal shared identical by descent (IBD) segments to classify pairs within a relationship degree. We focus on classifying second-degree relatives, specifically half-sibling (HS), grandparent-grandchild (GP), and avuncular (AV) pairs, and further label GP and HS relationships as paternal or maternal. To classify the relationship type, CREST leverages IBD sharing between the second-degree pair x1 and x2, and a more distant mutual relative y of 3rd to 5th degree with x1 and x2. For each pair and relative, we compute the ratio Ri = IBD(x1, x2, y)/IBD(xi, y), where i = 1,2 and IBD(a, ..., z) denotes the length of IBD sharing among all individuals a, ..., z. The ratio will differ in accordance with x1 and x2Õs

relationship type. We perform classification using kernel density estimators (KDEs) trained on these ratios from simulated data. To classify the GP and HS relationships as paternal or maternal, we leverage the key insight that the male and female genetic maps differ profoundly, implying that sex-specific patterns of recombination will be detectable using the IBD segments shared by x1 and x2. We model the familial history of recombinations that generated these shared IBD segments to compute the probability of the shared IBD under either a male or female genetic map. To validate CREST, we used real pedigree samples from the Generation Scotland (GS) dataset, which contains 848 GP, 381 HS, and 6599 AV relatives (total 7828 pairs). For those pairs where at least one member shares >=12.5% of their (diploid) genome IBD to other mutual relatives, CREST correctly identifies 92.9% of GP (65/70), 96.1% of HS (73/76), and 89.3% of AV pairs (826/925). Furthermore, the method correctly infers the parental sex of 95.0% of GP (806/848, receiver operating characteristic area under the curve [ROC AUC] = 0.991) and 99.0% of HS relatives (377/381, ROC AUC = 0.990). Notably, analyses of the X chromosome for two of the mistaken HS inferences indicate that the original data likely misidentified their relationship as paternal, and we confirmed with GS representatives that one of these pairs is mislabeled. Further analyses are underway for the remaining pairs. We anticipate CREST will find utility in large datasets where it will be able to reliably recover the underlying pedigrees, empowering a range of other analyses.

62 Common pitfalls in the analysis of scRNA-seq data (Sarkar)

Sarkar Abhishek <aksarkar@alum.mit.edu> (1), Lu Mengyin <mengyinlu228@gmail.com> (2), Stephens Matthew <mstephens@uchicago.edu> (2) (1) Affiliations: 1 - Department of Human Genetics, University of Chicago (United States), 2 - Department of Statistics, University of Chicago (United States)

The development of single cell RNA sequencing (scRNA-seq) technology has enabled investigation of heterogeneity between individuals cells, dynamics of cellular processes, and the biological pathways underlying cellular differentiation. However, a number of features of scRNA-seq data continue to cause confusion among practitioners applying published methods to analyze new data, as well as researchers developing and benchmarking new methods. For example, concepts such as dropout and zero inflation are neither precisely nor consistently defined in the literature. Here, we propose a principled framework which separates biological variability from the measurement process. This framework allows us to reframe old questions, such as (1) whether scRNA-seq data are zero-inflated, and (2) what distribution scRNA-seq data follows, in a new light. Our approach gives new insight into common analysis problems such as normalization and data transformation.

63 Mathematical properties of coalescence times in a diploid model of a consanguineous population (Severson)

Alissa Severson, Shai Carmi, Noah Rosenberg. Affiliations: Department of Genetics, Stanford University

Consanguineous unions, in which mating pairs share a recent common ancestor, produce offspring whose two genomic copies possess increased sharing of long segments inherited identical-by-descent (IBD). In a population, consanguinity increases the rate at which IBD segments pair within individuals to produce runs of homozygosity (ROH). The extent to which such unions affect IBD sharing between rather than within individuals, however, is not immediately evident from within-individual levels. To study this relationship, we recently developed a coalescent model that uses the fact that the time to the most recent common ancestor (TMRCA) for a pair of genomes at a specific locus is inversely related to IBD sharing between the genomes in the neighborhood of the locus. The model considers a set of mating pairs in a diploid population, treating the fraction of consanguineous unions as a parameter, and it enables the study of IBD sharing for a pair of genomes sampled either within an individual or in different individuals. Here, we derive the variance of coalescence times in the model, studying its dependence on the frequency of consanguinity and the kinship coefficient of consanguineous relationships. To derive the full distributions of the TMRCA, we introduce a separation-of-time-scales approach that treats consanguinity analogously to mathematically similar phenomena such as partial selfing. We evaluate the separation-of-time-scales approach by comparing its coalescence time distributions to simulations from the exact discrete-time Markov chain. The results extend

the potential to make predictions about ROH and IBD in relation to demographic parameters of diploid populations.

64 ngsPSMC: genotype likelihood-based PSMC for analysis of low coverage NGS data (Shchur)

Shchur Vladimir <vlshchur@gmail.com> (1), Sand Korneliussen Thorfinn <thorfinn.sand@gmail.com> (2), Nielsen Rasmus <rasmus_nielsen@berkeley.edu> (3) Affiliations: 1 - National Research University Higher School of Economics (Russia), 2 - University of Copenhagen (Denmark), 3 - UC Berkeley (United States)

Effective population size is one of the major characteristics of any population. It can reveal substantial information about population history including previous bottlenecks, expansions etc,which is of importance for understanding the pattern of genetic variation in the population. PSMC (Li, Durbin 2011) is one of the most used methods for estimating effective population sizes. However, in some cases (for low coverage genomes, in particular in the study of non-model species) its application is rather limited, because SNPs cannot be called reliably. We present a new version of PSMC, calledngsPSMC, which does not require SNP calling and instead works with genotype likelihoods. This method substantially expands the scenarios under which effective populaiton sizes can be reliably estimated. Our implementation also includes new features, including linear decoding of coalescent process (following Harris et al 2014), multithreading, and new parametrisation (cubic spline instead of 'pattern' sharing).

65 New features for polymorphism-aware phylogenetic models (Schrempf)

Schrempf Dominik <dominik.schrempf@gmail.com> (1), Borges Rui <ruiborges23@gmail.com> (2), Minh Bui Quang <m.bui@anu.edu.au> (3), Kosiol Carolin <ck202@st-andrews.ac.uk> (2) (4) Affiliations: 1 - Department of Biological Physics, Lorand University Budapest (Hungary), 2 - Institute of Population Genetics, Vetmeduni Vienna (Austria), 3 - Ecology and Evolution, Research School of Biology, Australian National University (Australia), 4 - Centre for Biological Diversity, University of St Andrews, St Andrews, Fife KY16 9TH, UK (United Kingdom)

Molecular phylogenetics has neglected polymorphisms within present and ancestral populations for a long time. Recently, multispecies coalescent based methods have increased in popularity, however, their application is limited to a small number of species and individuals. We have introduced a polymorphism-aware phylogenetic model (PoMo), which overcomes this limitation and scales well with the increasing amount of sequence data. PoMo circumvents handling of gene trees and directly infers species trees from allele frequency data. PoMo extends any DNA substitution model and additionally accounts for polymorphisms in the present and in the ancestral population by expanding the state space to include polymorphic states. It is a selection-mutation model which separates the mutation process from the fixation process. PoMo naturally accounts for incomplete lineage sorting because ancestral populations can be in a polymorphic state. Our method can accurately and time- efficiently estimate the parameters describing evolutionary patterns for phylogenetic trees of any shape (species trees, population trees, or any combination of those). We have implemented our PoMo approach as software package IQ-TREE-POMO with several new features: (i) a search for the statistically best-fit mutation model (ModelFinder), (ii) the ability to allow mutation rate variation across sites (e.g., gamma distribution), assessment of branch support values (bootstrapping and jackknifing), (iv) simulator of sequences evolving under PoMo (bmm-simulate). Applications using great ape data sets will be presented. In particular, the new genome-wide data set of seven baboon populations (genus Papio) present a unique opportunity to apply our method to a primate clade that involves more complex processes than those usually assumed by phylogenetic models. The history of Papio includes episodes of introgression or admixture among genetically distinct lineages. We will discuss the effect of this complex history on genome-wide phylogenetic inference with PoMo as well as other approaches. We will also present new estimates of divergence times and mutation rates.

66 A 100,000 Genome Project haplotype reference panel (Shi)

Shi Sinan <sinan.shi@stats.ox.ac.uk> (1), Hu Sile <sile.hu@stats.ox.ac.uk> (1), Marchini Jonathan <jonathan.marchini@well.ox.ac.uk> (2), Myers Simon <simon.myers@stats.ox.ac.uk> (1) Affiliations: 1 - Department of Statistics [Oxford] (United Kingdom), 2 - Regeneron Pharmaceuticals, Inc (United States)

The 100,000 Genomes Project aims to sequence 100,000 genomes from around 70,000 people from the UK.It is expected that the use of high coverage sequencing will produce an almost complete characterization of the genetic variation in the project participants and will constitute the largest human genetic variation resource ever collected in the UK, and maybe the world. One of our main research goals is to create an accurate haplotype reference panel for use in genotype imputation. We have created a preliminary haplotype reference panel using called genotypes in 28,893 participants. Sequencing data was mapped and genotypes were called using the central processing pipelines developed by 100,000 Genomes Project. This resulted in a dataset consisting of ~230 million SNPs across the autosomes. The dataset consists of a diverse set of ancestries with percentages self reporting as White, Asian, Black and Mixed ancestry of 69.2%, 8.7%, 2.3% and 2.1% respectively. Project participants consist of probands for rare diseases and their close relatives, so the dataset as a whole contains large amount of related individuals. For example, 60.67% of the 28,893 participants have at least one first degree relative also in the study, which greatly aids phasing. We used a new phasing program SHAPEIT4 to phase the genotypes at an overlapping set of 820,548 SNP sites included in the HRC reference panel on chr20. We assessed phasing performance using 200 trio parents, phased without their children, but together with 28,693 other samples. The majority (81%) of these trio parents reported White British ancestry and had a median switch error rate of 0.75%. The phasing was carried out without use of any relatedness information. We also directly compared the resulting 57,786 phased haplotypes to the HRC reference panel (64,976 haplotypes) in terms of imputation performance, by imputing genotypes into 10 individuals of European ancestry, based on genotypes on Illumina 1M-Duo3_C genotyping array, and comparing the results to genotypes derived from high-coverage sequencing. At variants with frequency 0.01% we obtained a mean imputation r2 of 0.65 and 0.75 using the HRC and 100,000 Genomes reference panels respectively. We will also report comparisons of phasing methods that use read information and relatedness and how this translates into downstream imputation performance, and the utility of imputing the UK Biobank dataset using the 100,000 Genomes reference panel.

67 Decoding of Neural Network Basecallers for Nanopore Sequencing (Silvestre-Ryan)

Silvestre-Ryan Jordi <jordisr@berkeley.edu> (1), Holmes Ian <ihh@berkeley.edu> (1) Affiliations: 1 - University of California, Berkeley (United States)

In nanopore sequencing, as on the Oxford Nanopore Technologies platform (ONT), a single- stranded DNA molecule is threaded through a protein nanopore embedded in a synthetic membrane across which an electric potential is applied. As the DNA passes through the pore, it perturbs the flow of ions in a sequence-dependent manner. This time series of current measurements can then be base called, yielding the DNA sequence. Factors such as measurement noise and variability in the rate of DNA translocation complicate decoding of current to sequence, motivating the use of machine learning techniques. In general, each nucleotide spends a random length of time in the pore, and so may generate zero, one, or many current samples during its transition. The segmentation of the time series (i.e. its partitioning into individual chunks, each of which corresponds to the transition of a single nucleotide through the pore) is unknown, complicating the use of neural networks that would map the time series of current samples to a DNA sequence. This is addressed by existing base callers via Connectionist Temporal Classification (CTC), a method borrowed from speech recognition. Under CTC, the outputs of the network are identified with transitions in a finite state machine, and the base caller defines a posterior probability of any given DNA sequence (marginalizing over all possible segmentations of the time series) which can be calculated through dynamic programming. A final decoding step returns the most likely base called sequence, whether by Viterbi decoding or some other type of search. We set out to evaluate the impact of the decoding algorithm on base calling accuracy. To this end we have developed our own base calling software PoreOver, which implements a recurrent neural network base caller and associated CTC decoding algorithms. We assess the use of a beam search in decoding and find it improves base calling accuracy

compared with a best-path Viterbi search. We additionally extend this beam search decoding method for the more general task of base calling a single consensus sequence from two reads, and make use of a heuristic to make the probabilistic search more tractable. Consensus base calling is especially relevant to ONT's 1D2 sequencing protocol, in which the template and complementary strands are successively passed through the pore, generating two current samples for the same sequence. We tested our optimized dynamic programming algorithm on 1D2 nanopore reads, yielding a modest improvement in accuracy over decoding each read individually. Finally, we apply our decoding algorithms to the flip-flop CTC models used in the latest generation of ONT base callers.

68 Distinguishing signals of admixture from demography (Skov)

Skov Laurits <lauritsskov2@gmail.com> (1), Iasi Leonardo <leonardo_iasi@eva.mpg.de> (1), Popli Divyaratan <divyaratan_popli@eva.mpg.de> (1), Peyregne Stephane <stephane_peyregne@eva.mpg.de> (1), Peter Benjamin <benjamin_peter@eva.mpg.de> (1) Affiliations: 1 - Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig (Germany)

Introgression of archaic haplotypes into non-African populations has been well-studied; the main signal include long, divergent haplotypes and a shift in allele frequency towards ancient populations (D-statistic). Archaic hominins are known to have inhabited Africa at the same time as anatomically modern humans, and gene flow between these populations is a possibility. However, no ancient DNA form archaic populations in Africa has yet been found and thus traditional methods for detecting introgressing haplotypes using an archaic reference genome cannot be applied. While methods for detecting archaic admixture in the absence of archaic reference genomes has been developed, these methods typically require an unadmixed population for comparison or prior knowledge of the population demography. Typically it is unknown whether these assumptions hold, and it is difficult to assess and signals generated by local population structure or population size changes which can mimic signals of admixture. Here we present an approach to validate signals of admixture when archaic reference genomes are absent and the demography of the population of interest is unknown. We infer introgressed haplotypes using a Hidden Markov model (HMM) and perform posterior predictive checking on summary statistics relevant to admixture. We apply our method to simulated data and show that this approach accurately recovers signals of admixture, when admixture is present, and does not recover signals of admixture when admixture is absent. We also apply the approach to archaic admixture into non-African populations and show that the admixture is well supported. Finally we show that signals of admixture between African population and African Archaic hominins are much harder to recover.

69 Evidence of deep-lineages in African genealogies (Speidel)

Speidel Leo <speidel@stats.ox.ac.uk> (1), Myers Simon <myers@stats.ox.ac.uk> (1) Affiliations: 1 - University of Oxford, Department of Statistics (United Kingdom)

Recent progress in genealogy estimation has made inferred genealogies of many thousands of samples possible (Kelleher et al., 2018; Speidel et al., 2019). In our recent paper (Speidel et al., 2019), we have demonstrated the wide-ranging utility and power of genealogy-based inferences across many population genetic applications, e.g., for inferring past demographic histories, evidence of selection, or introgression. In this presentation, I will focus on apparent signatures of unknown hominids in African genealogies, indicating recent contact with groups distantly diverged from Neanderthals or Denisovans and hence separate to the introgression signal observed in non-Africans. By randomly reshuffling tree topologies while fixing ages of coalescences (and demographic histories), we quantify the extent to which observed genealogies differ from panmictic scenarios. We observe a large excess of deep lineages unique to African populations, while a majority of such lineages are explained by Neanderthal or Denisovan introgression in non-Africans.

70 Deep imputation of tensors with structural missingness via exchangeability (Spence)

Spence Jeffrey <spence.jeffrey@berkeley.edu> (1), Batra Sanjit <sanjitsbatra@berkeley.edu> (1), Fischer Jonathan <jrfischer@berkeley.edu> (1), Song Yun <yss@berkeley.edu> (1) Affiliations: 1 - University of California, Berkeley (United States)

Biological data can often be represented as multidimensional arrays or tensors. For example, expression levels may be measured for a set of genes across both tissues and individuals, or a number of different assays may be performed across the genome in many cell types. These data are often plagued by structural missingness: entire rows or columns may be missing if a given assay has not been performed or if a given tissue was not assayable. If these missing entries could be accurately imputed, then it would be possible to obtain the results of unperformed experiments, such as measuring gene expression in difficult to assay tissues, like the brain, by using expression levels in easier to assay tissues such as whole blood. Unfortunately, imputing these missing entries is difficult, and most previous approaches assume a low-dimensional linear latent structure. Biological data are often highly nonlinear, however, and with massive datasets it should be possible to relax these assumptions. We developed a method based on deep convolutional neural networks, capable of capturing the nonlinear relationships between the missing and observed entries. We leverage the exchangeable structure present in many biological datasets, to overcome problems of identifiability using a novel permutation-equivariant scheme. We apply our method to a number of different real datasets and demonstrate that out method largely outperforms existing methods and baselines.

71 Bayesian interaction and difference detection in Hi-C data using generalized additive models and fused lasso (Spill)

Spill Yannick <yannick.spill@unistra.fr> (1) (2)1 - Biotechnologie et signalisation cellulaire (France), 2 - Centro de Regulacion Genomica (Spain)

3C-like experiments, such as 4C or Hi-C, have been fundamental in understanding genome organization. Thanks to these technologies, it is now known, for example, that Topologically Associating Domains (TADs) and chromatin loops are implicated in the dynamic interplay of gene activation and repression, and their disruption can have dramatic effects on embryonic development. However, the analysis of Hi-C experiments is both statistically and computationally demanding. Most methods are hindered by the high noise, large quantities of data and inadequate modelling of spatial dependency. In this talk, I will present a new way to represent Hi-C data, which leads to a more detailed classification of paired-end reads and, ultimately, to a new normalization and interaction detection method. This method, called Binless, uses a generalized additive model framework, and makes extensive use of the sparse fused lasso regression in a Bayesian setting. Binless is resolution-agnostic, and adapts to the quality and quantity of available data. Using a large-scale benchmark, I demonstrate that Binless is able to call interactions with higher reproducibility than other existing methods.

72 Modeling maintenance of functional redundancy using tRNA genes (Thornlow)

Thornlow Bryan <bthornlo@ucsc.edu> (1), Ridgley Trevor <tridgley@ucsc.edu> (1), Corbett- Detig Russ <russcd@gmail.com> (1) Affiliations: 1 - University of California, Santa Cruz (United States)

Transfer RNA (tRNA) genes are essential for the production of all proteins across all forms of life. However, the forces governing their maintenance and evolution are poorly understood. Primate genomes contain roughly 400 tRNA genes, encoding 47 different anticodons. Given that many of these genes are exactly identical in their nucleotide sequences, they must be at least somewhat functionally redundant. Theory suggests that functional redundancy, in which the fitness of an individual is dependent on only one copy of a given gene, can be maintained indefinitely when the germline and somatic mutation rates are sufficiently high and vary among genes. Our previous work shows that many tRNA genes experience germline mutation rates roughly tenfold greater than the genome-wide average as a result of transcription-associated mutagenesis, and we expect that these factors contribute to tRNA gene family evolution. However, other

duplicate gene families, such as ribosomal RNAs and oncogenes, may be conserved in many copies for cumulative function and for regulation of high-fidelity processes, respectively. In short, there are many possible explanations for our hundreds of deeply conserved human tRNA genes. To identify the processes that drive tRNA gene family evolution, we have built a forward-in-time individual-based simulation to study the evolution of tRNA genes at population scale. Our simulator incorporates both germline and somatic mutation rates, which are dependent on the transcription rate of each gene. We have also incorporated several fitness functions, representing functional redundancy, and stabilizing selection towards an optimal level of total expression, among others. We present our versatile simulation tool, which can easily be used to study other duplicate gene families, and we fit our simulation results to the demographic history and tRNA distributions of the great apes, demonstrating that high germline and somatic mutation rates are important factors in maintenance of functional redundancy.

73 Using two-loci statistics for inferring the properties of recent bottlenecks and founder events in human history (Tournebize)

Remi Tournebize1,2, Priya Moorjani1,2. Affiliations: 1: Department of Molecular and Cell Biology, University of California, Berkeley, USA. 2: Center for Computational Biology, University of California, Berkeley, USA

Founder events, whereby a new population is formed by a subset of individuals from a larger one, have played a critical role in shaping genetic diversity in humans. Founder events can occur due to geographical separation (e.g. in Finns) or cultural separation (for instance, due to endogamy as seen in Ashkenazi Jews or South Asians). Founder events reduce genetic variation in populations, decrease the efficacy of selection to remove deleterious variants and increase the risk of recessive diseases. To characterize founder events in humans, we introduce a novel method to infer the age and intensity of founder events. This method uses the autocorrelation in allele sharing (ASC) across the genome between pairs of individuals to recover signatures of past bottlenecks. We show that ASC decays exponentially with genetic distance, with the rate of decay proportional to the age of the founder event and the amplitude inversely proportional to the intensity of the bottleneck (a measure of the effective population size and duration of the bottleneck). We further show that the variance in ASC can help to disentangle the estimation of the bottleneck size and duration. Using coalescent simulations, we demonstrate that ASC provides reliable estimates of the age and intensity of founder events under a range of demographic scenarios. We illustrate our method by applying it to a large dataset of South Asians, containing 2,800 individuals from over 260 ethnolinguistic groups. We infer that most of the founder events in South Asia postdate the admixture between Ancestral North Indians (ANI) and Ancestral South Indians (ASI) that occurred in the past 4,000 years. The estimated intensity of founder events in South Asians suggests that many groups may have high rates of recessive diseases and that recessive disease mapping efforts in South Asia can help to reduce disease burden in the subcontinent. Our method uses diploid genotypes for inference and unlike other available approaches does not require phased data, which makes it applicable to datasets with small sample sizes and ancient genomes.

74 Efficient simulation of introgression, admixture and local ancestry (Tsambos)

Tsambos Georgia <gtsambos@student.unimelb.edu.au> (1) (2), Ralph Peter <plr@uoregon.edu> (3), Kelleher Jerome <jerome.kelleher@well.ox.ac.uk> (4), Leslie Stephen <stephen.leslie@unimelb.edu.au> (1) (2), Vukcevic Damjan <damjan.vukcevic@unimelb.edu.au> (1) (2) Affiliations: 1 - University of Melbourne (Australia), 2 - Melbourne Integrative Genomics (Australia), 3 - University of Oregon (United States), 4 - Big Data Institute (United Kingdom)

To assess the performance of methods in population genetics, we often wish to simulate realistic genetic datasets while retaining detailed information about the history of the simulated genomes. This is especially important when the consequence of admixture on patterns of genetic diversity is of primary interest. Many existing methods can infer the ancestral origin of chromosomal segments. However, it is difficult to simulate chromosomes for which the true origin of those segments is known; existing approaches are approximate and ad-hoc. Recent advances implemented in the software msprime (Kelleher et al. 2016) and SLiM (Haller et al. 2018) allow us to efficiently record genetic information using a succinct tree sequence data structure, which provides unprecedented detail about the genealogy of the sample. However, for the purposes of

studying admixture and ancestry, this detail can be overwhelming and difficult to analyse. We are often most interested in the ancestral population that particular genomic segments have been inherited from (i.e. the local ancestry of the sample). Recovering this information from the overall genealogies is challenging. In this work, we will outline a method that combines these existing state-of-the-art tools with a processing step to efficiently extract local ancestry information. The simulation procedure combines a forward-in-time step to simulate admixture with a backwards-in-time step to simulate genetic diversity in the ancestral populations. These techniques allow the user to track local ancestry in large simulations under realistically complex demographic scenarios, with minimal computational overhead. To illustrate the usefulness of this procedure, we will also show how it might be used to benchmark the performance of ancestry inference methods on various admixed populations, and to assess the degree of incomplete lineage sorting. More broadly, we anticipate that this procedure will make it easier to explore the impact of complex demographic hypotheses on detailed patterns of genetic diversity.

75 From Summary Statistics to Individual Level Data: Correcting for Genetic Drift within GWAS (Tutert)

Tutert Marcus <marcus.tutert@stx.ox.ac.uk> (1), Hinch Robert <robert.hinch@gmail.com> (2) (1), Mcvean Gil <gil.mcvean@bdi.ox.ac.uk> (2) (1) Affiliations: 1 - Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford (United Kingdom), 2 - Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford (United Kingdom)

Genome Wide Association studies (GWAS) have been used widely to identify variants that influence complex traits and diseases. However, individual level data from GWAS is typically masked from public view, for reasons of privacy, consent, and computational burden. Instead, study information can be captured through sharing of summary-statistics (SS), typically effect sizes, and their standard errors. SS methods for downstream analyses such as imputation, fine-mapping, and estimation of heritability, often employ the use of external population reference panels, like those from the 1000 Genomes, to model linkage disequilibrium (LD) structure in the GWAS population. However, genetic differences between the GWAS study and reference panel populations can result in substantial inaccuracy. Here, we describe a maximum likelihood estimator of Fst between a GWAS population and external reference population that uses only SS data from the GWAS study, specifically the standard errors, which are closely related to the allele frequencies in the study population. We use the Nichols- Balding beta-binomial model for genetic drift in allele frequencies, augmented with a Gaussian copula approach to model correlation in allele frequencies among nearby variants that arises through linkage disequilibrium. Moreover, we explicitly model error, such as arising through poor imputation, genotyping error or the influence of covariates. Parameters are estimated using maximum likelihood. We evaluate the accuracy and robustness of the estimator through simulations, finding that the Pearson r2 correlation between the simulated and inferred Fst is greater than 0.93, and apply the approach to empirical data sets where individual and summary-level data are both available. Finally, we discuss how these estimates have the potential to improve downstream GWAS analyses from summary statistics.

76 UK Biobank participants that moved 20 km from their birthplace have on average higher socioeconomic status and improved health (Williams)

Woods Ian <iwoods@ithaca.edu> (1), Williams Amy L. <alw289@cornell.edu> (2) Affiliations: 1 - Department of Biology, Ithaca College (United States), 2 - Department of Computational Biology, Cornell University (United States)

The UK Biobank (UKB) is a rich repository of genotype and phenotype data comprising nearly half a million people, and has seen widespread adoption in numerous past and ongoing studies. Given its magnitude and data access policies, many studies have constructed polygenic scores using UKB effect size estimates, highlighting the importance of the UKB, but also raising potential concerns about the reliability of these estimates. Most notably, subtle population structure may confound UKB effect sizes, and may do so in complicated ways depending on the phenotype under study. We performed an analysis of the demographic characteristics of the UKB samples, focusing especially on patterns of migration. As might be expected, the

place of residence (POR) of over 84% of the participants is within 20 km of one of the 21 assessment centers (94% within 25 km), indicating highly localized participant sampling. However, only 51% of the participants were born within 20 km of their assessment center (57% within 25 km), leading to a partitioning of the dataset into groups we term 'stayers' and 'movers'. We performed this partitioning, defining movers as those with distance between their place of birth (POB) and POR of >20 km, yielding 55.8% of samples labeled as stayers and 44.2% as movers. The dispersal among movers is highly variable with a mean POB-POR distance of 160 km, and a standard deviation of 128 km; by contrast, stayers have a very small mean POB-POR distance of 6.6 km (standard deviation 4.9 km). Given this division, we analyzed potential phenotypic correlations with mover status, and found that it is correlated with a range of phenotypes, including numerous positive physical health outcomes (lower BMI, higher overall health rating), higher educational attainment (EA), income, and somewhat increased standing height, whereas interpersonal relationship satisfaction is lower in these samples. As individuals that have migrated may have slightly different genotypes than those that have not, we performed genome-wide association studies (GWAS) on 329,746 individuals of white British ancestry for 11 phenotypes. These GWAS included the top 40 principal components, sex, and SNP chip as covariates. We then performed a second set of GWAS that included these same covariates together with mover status, and we compared these to the GWAS without mover status. The effect size estimates of the top SNP in each genome-wide significant locus (P < 5e-8) changed for several phenotypes. For example, the absolute value of the effect sizes shrank by an average of 14.2% for EA, 13.7% for income, and 1.6% for standing height (P â‰¤ 1e-18 for difference in effect sizes for these phenotypes). Moreover, the number of loci with genome-wide significant SNPs reduced by 48% for EA (from 40 to 21) and 48% for income (21 to 11), though remained the same for standing height (461 in both analyses). These results highlight the difficulty in reliably inferring effect sizes in the presence of population structure'"particularly for phenotypes that may be correlated with migrant status'"adding to a growing body of evidence that subtle bias can remain in estimated effect sizes even when using state-of-the-art techniques to correct for such structure.

77 Inferring tree sequences from large DNA datasets: problems and solutions (Wong)

Wong Yan <yan.wong@bdi.ox.ac.uk> (1), Kelleher Jerome <jerome.kelleher@well.ox.ac.uk>, Wohns Anthony Wilder <wilder.wohns@merton.ox.ac.uk> (1), Mcvean Gil <gil.mcvean@bdi.ox.ac.uk> (1) Affiliations: 1 - Big Data Institute, University of Oxford (United Kingdom)

We have recently developed a scalable algorithm, tsinfer, to infer the genealogical history of thousands or millions of genomes within a species. The algorithm is based on the idea of reconstructing ancestral genetic haplotypes (ancestors). A major benefit of this approach is that it is grounded in biological reality: we know such genetic ancestors must have existed at specific points in the past, although we may not know much about them. We outline the general logic of the ancestral inference process, in particular the matching of samples and ancestral haplotypes using the Li and Stephens copying process, and show how it accurately infers ancestry in real-world datasets. We will discuss how thinking about ancestral haplotypes leads to natural extensions of our methods to allow for missing data, fragmentary samples,ancient samples and to account for sequencing (and other) errors. We also describe ongoing work to infer the times at which genetic ancestors existed in the past, and improvements of the method to more precisely track the genomic locations and date of recombination breakpoints. Finally, we discuss how inferred tree sequences can be used to address key problems in population genetics.

78 A novel statistical method for identifying combinatorial regulatory elements via deconvolution of multiplexed CRISPR regulatory screens in single-cells (Zhou)

Zhou Jessica <jlz014@eng.ucsd.edu> (1) (2), Mcvicker Graham <gmcvicker@salk.edu> (2) (1) Affiliations: 1 - Bioinformatics and Systems Biology Program, University of California San Diego (United States), 2 - Salk Institute for Biological Sciences (United States)

Regulation of gene expression is critical for maintaining functional biological processes, and dysregulation can affect physiology and drive disease onset. However, the identity of regulatory sequences, their target genes and their effects on targets are largely unclear. CRISPR-based genome editing experiments for

screening regulatory elements have recently emerged as a new approach for identifying regulatory elements and their gene targets. However, most regulatory screens to date have focused on the effects of individual regulatory elements and overlooked how combinations of multiple regulatory elements affect gene expression. Several technical challenges further confound our ability to identify regulatory elements and study their behavior. Guide RNAs (gRNAs) used for directing CRISPR perturbations have variable efficiency and are prone to off-target effects. Additionally, high-throughput pooled CRISPR regulatory screens are only scalable to screening for regulatory elements of a relatively small number of genes at once, and yield simple readouts. These concerns must be addressed to improve accuracy and scope of regulatory element discovery. Recent technological advances have enabled large scale, multiplexed CRISPR regulatory screens at single-cell resolution. These experiments make it possible to generate transcriptomic readouts of CRISPR experiments at a high-throughput, allowing the effects of genomic sequence perturbations to be measured on multiple genes at once and increasing power for regulatory element discovery. In particular, these experiments have the potential to measure combinatorial effects of multiple regulatory elements on the expression of a target gene (or genes). In response to the increasing prevalence of these multiplexed, single-cell CRISPR regulatory screens, I am developing a new statistical analysis method that leverages the power and complexity of the datasets they generate to identify regulatory sequences and their targets. This method will account for relevant experimental variables by modeling guide efficiency and off-target effects, as well as the sparsity of single-cell sequencing data. Notably, my method will be capable of identifying combinations of regulatory elements that act redundantly or synergistically to influence expression of their targets.