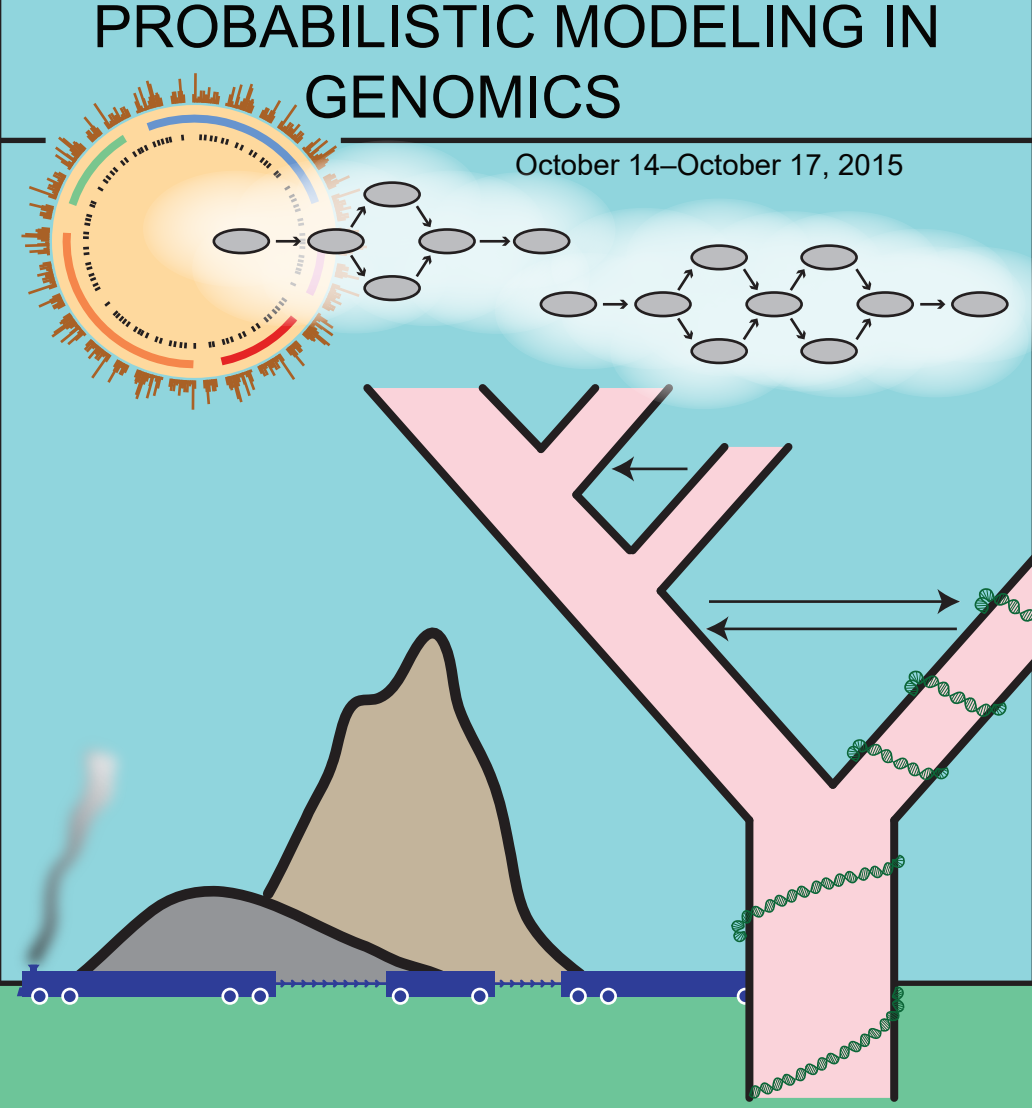
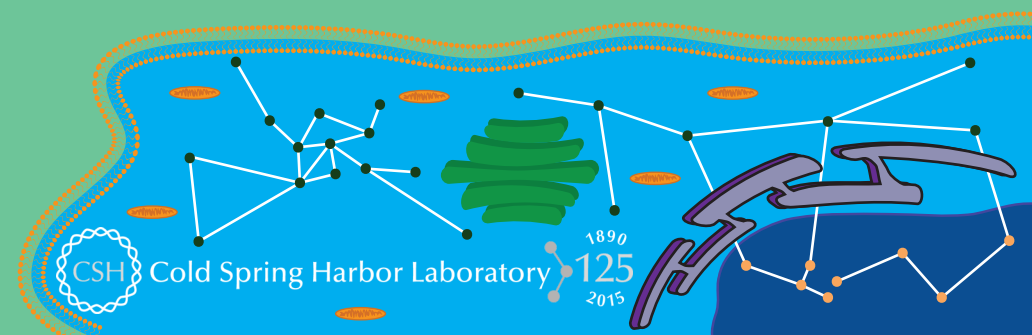
Abstracts of papers presented
at the 2015 meeting on

PROBABILISTIC MODELING IN GENOMICS

October 14–October 17, 2015





Cold Spring Harbor Laboratory
1890 125 2015

Abstracts of papers presented
at the 2015 meeting on

PROBABILISTIC MODELING IN GENOMICS

October 14–October 17, 2015

Arranged by

Barbara Engelhardt, *Princeton University*
Thomas Mailund, *Aarhus University, Denmark*
Adam Siepel, *Cold Spring Harbor Laboratory*

Contributions from the following companies provide core support for the Cold Spring Harbor meetings program.

Corporate Sponsors

Agilent Technologies
Bristol-Myers Squibb Company
Genentech
Life Technologies (part of Thermo Fisher Scientific)
New England BioLabs

Plant Corporate Associates

Monsanto Company

Cover Design by: Noah Dukler, Cold Spring Harbor Laboratory.

PROBABILISTIC MODELING IN GENOMICS
Wednesday, October 14 – Saturday, October 17, 2015

Wednesday	7:30 pm	**1** Demography and Admixture
Thursday	9:00 am	**2** Assembly and Variant Identification
Thursday	1:30 pm	**3** Poster Session
Thursday	4:00 pm	**Keynote Speaker**
Thursday	5:00 pm	*Wine and Cheese Party**
Thursday	7:30 pm	**4** Systems and Structural Biology
Friday	9:00 am	**5** Population Genetics and Natural Selection
Friday	2:00 pm	**6** Functional Genomics
Friday	5:00 pm	**Keynote Speaker**
Friday	6:00 pm	Banquet
Saturday	9:00 am	**7** Quantitative Genetics

---

* *Airslie Lawn*, weather permitting

Mealtimes at Blackford Hall are as follows:
Breakfast 7:30 am-9:00 am
Lunch 11:30 am-1:30 pm
Dinner 5:30 pm-7:00 pm
Bar is open from 5:00 pm until late

Printed on 100% recycled paper.

PROGRAM

WEDNESDAY, October 14—7:30 PM

SESSION 1 DEMOGRAPHY AND ADMIXTURE

Chairpersons: **John Novembre,** University of Chicago, Illinois
Yun Song, University of California, Berkeley / University of Pennsylvania, Philadelphia

THURSDAY, October 15—9:00 AM

SESSION 2 ASSEMBLY AND VARIANT IDENTIFICATION

Chairpersons: **Richard Durbin,** Wellcome Trust Sanger Institute, Hinxton, United Kingdom
David Haussler, HHMI, University of California, Santa Cruz

SESSION 3 POSTER SESSION

THURSDAY, October 15—4:00 PM

KEYNOTE SPEAKER

Michael I. Jordan
University of California, Berkeley

THURSDAY, October 15—5:00 PM

Wine and Cheese Party

THURSDAY, October 15—7:30 PM

SESSION 4 SYSTEMS AND STRUCTURAL BIOLOGY

Chairpersons: **Mona Singh,** Princeton University, New Jersey
Olga Troyanskaya, Princeton University, New Jersey

FRIDAY, October 16—9:00 AM

SESSION 5 POPULATION GENETICS AND NATURAL SELECTION

Chairpersons: **Rasmus Nielsen,** University of California, Berkeley
Amy Williams, Cornell University, Ithaca, New York

FRIDAY, October 16—2:00 PM

SESSION 6 FUNCTIONAL GENOMICS

Chairpersons: **Lior Pachter,** University of California, Berkeley
Sylvia Richardson, MRC Biostatistics Unit, Cambridge, United Kingdom

Genome-wide probabilistic dynamical modelling of transcription kinetics reveals patterns of RNA production delays
Antti Honkela, Jaakko Peltonen, Hande Topa, Iryna Charapitsa, Filomena Matarese, Korbinian Grote, Hendrik G. Stunnenberg, George Reid, Neil D. Lawrence, Magnus Rattray.
Presenter affiliation: University of Helsinki, Helsinki, Finland. 148

Multi-scale methods for detecting differences between multiple groups in high-throughput sequencing data and their application to small sample sizes
Heejung Shim, Zhengrong Xing, Ester Pantaleo, Francesca Luca, Roger Pique-Regi, Matthew Stephens.
Presenter affiliation: Purdue University, West Lafayette, Indiana. 149

FRIDAY, October 16—5:00 PM

KEYNOTE SPEAKER

Elizabeth A. Thompson
University of Washington, Seattle

"Modeling and inferring coancestry among multiple individuals across a chromosome."

FRIDAY, October 16

BANQUET

Cocktails 6:00 PM Dinner 6:45 PM

SESSION 7 QUANTITATIVE GENETICS

Chairpersons: **Jennifer Listgarten,** Microsoft Research New England, Cambridge, Massachusetts
Matthew Stephens, University of Chicago, Illinois

INFERRING POPULATION GROWTH RATES IN A STRUCTURED POPULATION

Mark Reppell, John Novembre

University of Chicago, Department of Human Genetics, Chicago, IL

Widely adopted methods for inferring population demography from genetic data, such as PSMC (Li and Durbin, 2011), dadi (Gutenkunst et al, 2009), and fastNeutrino (Bhaskar et al, 2014), assume either that samples are drawn from a panmictic population or from a very limited number of subpopulations. Population structure confounds signals of changing population size, and distorts demographic estimates in a manner determined by the sampling scheme of a study. In human studies, samples from broad geographic regions across which subtle population structure is likely to exist, such as Europe, are often treated as panmictic during analysis, likely resulting in inaccurate demographic estimates. Here, we propose a method to infer demography in populations with subtle substructure that can be represented with an island model. Our method combines information from sequence data in pairs of diploid samples for which we know whether they originate in the same or different subpopulations. At a locus, we leverage information about the structure of the underlying genealogy in addition to the site frequency spectrum to calculate a demographic likelihood. We use the Poisson Random Fields framework and deterministically calculate expected coalescent times with the result that our method is substantially more efficient than coalescent-based Monte Carlo integration.

Using simulations, we demonstrate the ability of our method to accurately infer population growth and size under scenarios with subtle structure, and contrast this with estimates from conventional methods that assume panmictic sampling. We demonstrate the robustness of our approach to model misspecification and genotyping error. Finally, we discuss how with real data our method can be combined with approaches such as PCA (Price et al, 2006) or fineStructure (Lawson et al, 2012) to first classify substructure in a sample, and then infer total population demography.

← Back

SMC++: DEMOGRAPHIC INFERENCE ON LARGE SAMPLES USING THE SEQUENTIALLY MARKOVIAN COALESCENT

Jonathan Terhorst[1], John A Kamm[1], Yun S Song[1,2,3]

[1]University of California, Berkeley, Department of Statistics, Berkeley, CA, [2]University of California, Berkeley, Computer Science Division, Berkeley, CA, [3]University of California, Berkeley, Department of Integrative Biology, Berkeley, CA

The Poisson random field (PRF) and pairwise sequentially Markov coalescent (PSMC) are two popular models for inferring population size histories. Each approach has advantages and drawbacks compared to the other. The PRF, by assuming independence among sites, cannot fully exploit modern whole-genome sequence data; in contrast, the PSMC uses linkage information between neighboring loci for enhanced power. On the other hand, while the PSMC is limited to a pair of haploid lineages, the PRF can in principle analyze an arbitrary number of individuals at once. The ability to scale to larger sample sizes is particularly important when reconstructing the recent history of a growing population, since there will be few recent coalescence events in a sample of size two.

In this work we present SMC++ (**s**equentially **M**arkov **c**oalescent **+ pl**enty of **u**nlabeled **s**amples), a new coalescent HMM-based method which combines the strengths of both the PRF and PSMC. The key innovations underlying this method are new theoretical results and an efficient algorithm for computing the expected sample frequency spectrum conditioned on the event that two distinguished lineages coalesce in a given time interval. Using this "conditioned SFS", we obtain a model which scales to multiple individuals while retaining many of the attractive features of the PSMC, including the ability to operate on unphased data. We study how the inclusion of additional samples improves our ability to reconstruct recent population size changes, as well as its effect on decoding the coalescence interval of the distinguished lineages *a posteriori*. We compare this performance to that of existing methods such as PSMC, MSMC, and DiCal. Finally, generalizations to the case of multiple populations are explored.

INFERENCE OF POPULATION SIZE TRAJECTORIES WITH TAJIMA'S COALESCENT

Julia A Palacios[1,2], Amandine Veber[3], John Wakeley[1], Sohini Ramachandran[2]

[1]Harvard University, Organismic and evolutionary biology, Cambridge, MA, [2]Brown University, Ecology and evolutionary biology, Providence, RI, [3]École Polytechnique, Centre de Mathématiques Appliquées, Palaiseau, France

Inferring and interpreting changes in effective population size over time (or "population size trajectories") is a core goal of population biology; genomic data allow such inference to be done from the deep past into the near present. Population sizes change due to shifting environmental pressures, migrations, population-founding events, and natural selection. Thus estimates of population size changes over time shed light on how historical events affect genetic diversity within a species, as well as the genetic response of a species to large-scale climatic events and to arrival in new environments.

Coalescent-based methods for inferring effective population size trajectories from sequence data in a Bayesian framework rely on the joint posterior distribution of model parameters and the coalescent genealogy that relates all the sequences sampled. In previous work focusing on inference of population size trajectories from sequential genealogies (Palacios et al., in press, Genetics), we implemented a novel Bayesian nonparametric approach for estimating population sizes with measures of uncertainty. We showed that our method outperforms recent likelihood-based methods that rely on discretization of the parameter space. In simulation of a variety of demographic models, our Bayesian approach produces point estimates four times more accurate than maximum likelihood estimation (based on the sum of absolute differences between the truth and the estimated values). Further, our method's credible intervals for population size as a function of time cover 90 percent of true values across multiple demographic scenarios, enabling formal hypothesis testing about population size differences over time.

Here, we present a new Bayesian approach for inference of population size trajectories from sequence data that relies on the joint posterior distribution of model parameters and the Tajima's genealogy. A Tajima's genealogy includes coalescent times and a ranked tree shape, and has a drastically smaller state space than that of a coalescent genealogy which includes coalescent times and a labeled topology. We provide a new Markov Chain Monte Carlo (MCMC) algorithm for sampling from our new proposed joint posterior distribution under the infinite sites and the Jukes-Cantor models for mutation. We compare the performance of our algorithm with the corresponding algorithms implemented in BEAST (Drummond and Rambaut, 2007). Further, we provide a new MCMC algorithm for sampling from the joint posterior distribution of model parameters and local Tajima's genealogies for inferring effective population size trajectories under the sequentially Markov coalescent model (SMC').

← Back

TREE CONSISTENT PBWTS AND THEIR APPLICATION TO RECONSTRUCTING ANCESTRAL RECOMBINATION GRAPHS AND DEMOGRAPHIC INFERENCE

Vladimir Shchur, Richard Durbin

Wellcome Trust Sanger Institute, Hinxton, United Kingdom

We present an efficient way to represent and process ancestral recombination graphs (ARGs). An ARG is a genealogical network which relates a sample of individual genomes by providing a local tree at each position, linked by ancestral recombinations that can be considered as prune-and-regraft operations on the local trees. A planar ordering of the leaves of local trees, together with a similarity function between adjacent leaves, provides a data structure for efficient and scalable storage, processing and inference of ARGs.

The positional Burrows-Wheeler transform PBWT is a representation of a set of haplotypes that supports very efficient data compression and fast haplotype matching. We introduce a modification of PBWT which we call a tree consistent PBWT, or tcPBWT for short, which has a natural and tractable connection with local coalescent trees and the efficient representation of ARGs described above. tcPBWT shows some improvement in compression rate compared to PBWT, which suggests that it has better consistency with the genealogy giving rise to the haplotypes. tcPBWT includes a great amount of planar ordering structure, so it can be used to efficiently generate possible ARGs, in time linear in the size of the data set. The distribution of topologies of inferred local trees is strongly influenced by the population structure of a sample. We present and discuss two summary statistics for population structure inference from ARGs derived via our tcPBWT approach, with their application to large data sets, such as the complete 1000 Genomes Phase 3 haplotypes.

← Back

DEMOGRAPHIC INFERENCE USING ARGWEAVER

Melissa J Hubisz[1,2], Adam Siepel[1]

[1]Cornell University, Biological Statistics and Computational Biology, Ithaca, NY, [2]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

Current methods for demographic inference are limited in the amount of data they can use or the type of models they consider. Most methods have at least one of the following limitations: allow only one or two populations, may not allow for recombination or patterns of linkage disequilibrium, allow for only a small number of samples, or use only summary statistics of the data. Here, we discuss the use of ARGweaver (Rasmussen et al, 2014) as a framework for demographic inference. ARGweaver makes use of both linkage disequilibrium and site patterns to infer ancestral recombination graphs (ARGs), and can be used genome-wide on many dozens of individuals. While ARGweaver was originally designed for a single constant-sized population, ARGs inferred by ARGweaver contain information about the true demography.

We first explore the potential for post-processing ARGs to infer demography. We find that ARGs generated from multiple-population data can reliably detect genomic segments with shared ancestry due to migration events or incomplete lineage sorting. These two scenarios can be distinguished by looking at the coalescence time of these lineages, which is specified by the ARG. We discuss the application of this approach to Neandertal and human sequence data, and find that introgressed regions can be identified with high power and low false-positive rate.

We then consider how demography can be inferred jointly with the ARG. We present updates to the ARGweaver model that allow for an arbitrary number of populations and migration bands, using a user-specified population tree. While this comes at a computational cost, it can reasonably be used genome-wide for at least three populations and several migration bands, and we discuss strategies for reducing the computational burden. This model can be used to flexibly infer population size changes and migration events, and should allow comparison of arbitrarily complex demographic scenarios, such as one vs several migration bands. Unlike previous methods, it simultaneously identifies the regions of the genome underlying the inference, and therefore can be used to identify and date introgressed haplotypes.

← Back

JOINT INFERENCE OF SAMPLE PEDIGREES, ADMIXTURE PROPORTIONS, AND MIGRATION RATES

Peter R Wilton[1], Pierre Baduel[1], Matthieu Landon[2], John Wakeley[1]

[1]Harvard University, Organismic and Evolutionary Biology, Cambridge, MA, [2]Harvard University, Systems Biology, Cambridge, MA

Different chromosomes in diploid sexual populations are inherited via different paths through a shared population pedigree that imposes a non-independence between loci that is not usually modeled in demographic inference. Previous work has shown that predictions about patterns of coalescence made without consideration of the shared pedigree are mostly accurate for panmictic, constant-sized populations that are at least moderately sized. We extend this work to consider how the population pedigree shapes patterns of genetic variation in the presence of population structure with migration between subpopulations. We show how the realized migration process constrains the movement of alleles between subpopulations, such that the effective migration rate fluctuates over time and in certain scenarios the pedigree affects coalescence farther into the past than in the panmictic case. We also show how coalescence time distributions from samples featuring identity by descent or recent admixture can be viewed as mixture distributions over different sample configurations, and we use this observation to develop a framework for inferring scaled mutation and migration rates jointly with features of the recent pedigree of small samples. This procedure is tested thoroughly by simulation and shown to accurately recover relatedness coefficients, admixture proportions, mutation rates, and migration rates.

← Back

THE HUMAN GENOME VARIATION MAP PILOT PROJECT

Benedict Paten[1], Adam Novak[1], Glenn Hickey[1], Sean Blum[1], Maciek Smuga-Otto[1], Karen Miga[1], Jerome Kelleher [2], Erik Garrison[3], Sasha W Zaranek[4], Heng Li[5], Stephen Keenan[6], Richard Durbin[3], Gil McVean[2], David Haussler[1]

[1]UC Santa Cruz Genomics Institute, 1156 High St, Santa Cruz, CA, [2]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, United Kingdom, [3]The Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom, [4]Curoverse, 212 Elm St, Somerville, MA, [5]Broad Institute, 415 Main St, Cambridge, MA, [6]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

Since researchers posted the first draft of the human genome on the Internet in 2000, the current model of human genetics in the genome era has not changed: genomics relies on a single, haploid set of 24 reference chromosomes to interpret human genomes. Essentially all novel sequencing data is analyzed by mapping the sequenced reads only to this one reference set of 24 chromosomes to identify variants. This leads to a phenomenon called reference allele bias—the tendency to over-report versions ("alleles") of the genome that are in the reference genome and miss non-reference alleles. Not only is this suboptimal, it is biased by genetic subpopulation: the current reference is a better reference for some subpopulations than others - it is not a universal reference for humanity. As a first step to remedy this deficiency, a limited set of alternative haplotypes have been added to the reference genome. But not only do these severely under represent human genetic diversity, they also overlap with the primary reference and create mapping ambiguity. As a result, they are ignored. We outline the human genome variation map project, an ambitious community collaboration of members of the Global Alliance for Genomics and Health to create a reference structure containing all common human variation, encoded as a single fundamental structure encoded as a graph. To date we have conducted a pilot evaluation in which prototype graphs have been constructed and evaluated for set of genomic regions. We will argue that these graphs can be used as a more effect substrate for genomic analyses, and that probabilistic models for read mapping, variant calling and phasing can all be naturally expanded to integrate this richer information source.

EFFICIENT CONSTRUCTION OF SPARSE STRING GRAPHS

Ilan Shomorony[1], Thomas Courtade[1], David Tse[2]
[1]UC Berkeley, EECS, Berkeley, CA, [2]Stanford University, EE, Stanford, CA

Emerging long-read sequencing technologies have the potential to generate less fragmented *de novo* genome assemblies by reducing ambiguity resulting from sequence repeats. In this context, assembly algorithms based on the framework of de Bruijn graphs seem unsuited to fully exploit the power of long reads. Instead, approaches based on finding maximal overlaps between pairs of reads, which are typically longer than de Bruijn k-mers, can often disambiguate the read data to a greater degree. Hence, read-overlap based approaches are expected to play a central role in the next generation of assemblers.

A fundamental challenge in any read-overlap approach is that one must identify which edges should and should not be included in the read-overlap graph. The paradigm of a string graph was introduced as a way to compactly represent the necessary read-overlap relationships by focusing on irreducible edges (i.e., those edges that cannot be eliminated through a process of transitive reduction). However, even after transitive edges are eliminated, the task remains to determine which edges/overlaps are "significant" enough to be in the final string graph. An overly-conservative pruning procedure leads to a dense graph with many spurious edges, rendering the task of extracting long contigs -- let alone obtaining a complete assembly -- very difficult. On the other hand, an aggressive pruning procedure will incorrectly remove "true" edges, leading to higher fragmentation of the assembly.

In this work, we study how the repeat statistics of the target genome affect how sparse a string graph can be, while remaining "sufficient" for assembly. For an appropriately defined notion of sufficiency, we derive conditions in terms of bridging of repeats under which a minimal sufficient string graph (i.e., a sufficient string graph with no spurious edges) can be constructed. Under a probabilistic read model, these conditions can be translated into coverage depth requirements that nearly match information-theoretic lower bounds. Finally, motivated by these fundamental insights and by recently proposed techniques to efficiently construct string graphs, we present a linear-time algorithm for the construction of a sparse string graph. Compared to existing algorithms, this method requires significantly lower sequencing coverage depths in order to construct a minimal sufficient string graph.

← Back

STRUCTURAL VARIANT DETECTION AND PHASING WITH LINKED-READS

Sofia Kyriazopoulou-Panagiotopoulou, Patrick Marks, Kristina Giorda, Heather Ordonez, Patrice Mudivarti, Grace Zheng, Michael Schnall-Levin
10X Genomics, 10X Genomics, Pleasanton, CA

We present algorithms for the detection and phasing of large-scale (>50 kb) structural variants using data generated by the GemCode platform from 10X Genomics. In the GemCode system, long molecules from 1ng of DNA are randomly partitioned into more than 100,000 individually-barcoded microfluidic droplets (~3Mbp/droplet). The resulting libraries are compatible with standard exome capture and short-read sequencing while maintaining valuable long-range information.

Our SV-detection algorithm first searches for all pairs of genomic loci with a significant number of common barcodes. This search is encoded as an efficient sparse matrix multiplication. Candidate loci from this first stage are filtered using a probabilistic method that models the partitioning and read generation processes. SV breakpoints are phased with respect to SNPs using a maximum likelihood approach that optimizes the conditional probability of the read and barcode support for each allele given a phasing configuration.

We applied our SV-detection and phasing algorithms on WGS data from the CEPH cell line NA12878 and the Genome in a Bottle Ashkenazi trio. In each sample, we detected tens of large-scale variants, including deletions, inversions, and complex events. We validated the vast majority (>75%) of these events with orthogonal approaches. Barcode information allowed us to confidently phase more than half of these large-scale variants. Trio data verified that our SV-phasing calls are consistent with inheritance patterns.

To evaluate our SV-detection algorithms on WES, we used the GemCode platform followed by exome capture on samples from the lung cancer cell line NCI-H2228 and the triple negative breast cancer cell line HCC38. In NCI-H2228, we confidently called and partially resolved the phase of the known EML4/ALK fusion. In HCC38, we identified several previously reported gene fusions, such as the LDHC/SERGEF fusion, as well as tens of novel large-scale SVs that were absent from the matched control sample.

In summary, linked-read data in combination with novel probabilistic algorithms allowed us to confidently detect and phase large structural variants from WGS and WES using limited starting material.

NEW GENOME REFERENCE STRUCTURES

Richard Durbin

Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom

The current genome reference model is to have a linear exemplar chromosome sequence, plus possibly a catalogue of variant sites with alternative non-reference alleles, plus possibly a collection of haplotype sequences given in terms of which variant at each site is present in which sequence. Analysis of data from new individuals typically involves mapping to a sequence that can be relatively distant from that of the sample, and rediscovery from scratch of many variant alleles already found in other samples and present in the catalogue. In the last few years the Genome Reference Consortium has introduced ALT versions of regions of chromosomes relatively distant from the primary chromosomes, but (a) these are not widely used, and (b) they only capture a small fraction of the known diversity. Recently a number of approaches have been investigated to improve this situation, ranging from graph references that include local variation in a more complex structure for mapping, to haplotype population models that represent long range linkage disequilibrium structure. I will discuss a number of these, with respect to data structures, algorithms, statistical interpretation, and their possible incorporation in new standards in the framework of the Global Alliance for Genomics and Health.

RESEQUENCING AGAINST A HUMAN WHOLE GENOME VARIATION GRAPH

Erik Garrison, Jouni Siren, Richard Durbin

Wellcome Trust Sanger Institute, Cambridge, United Kingdom

Despite the availability of whole genome sequences from many individuals in many species, when examining a new genome we only use a single reference genome as our basis for interpretation. As a result, our analyses are biased against the detection of any variation comprising significant differences from our reference. This prevents the effective use of resequencing in the context of high genetic diversity, or where variation of interest cannot be represented in the reference coordinate system. It also increases the cost of observing new genomes as all variation must be treated as novel.

We can reduce reference bias when we observe new genomes if the reference system against which we place new sequence data can incorporate multiple genomes representing known variation. One approach to do so is to construct the reference as a graph over sequence segments (nodes) and possible linkages between neighboring segments (edges). This simple model is capable of encoding a number of standard representations of genomic variation which could provide novel bases for resequencing, including multiple sequence alignments, partially ordered graphs, string graphs, and De Bruijn graphs.

We develop a software package (vg) and binary data formats to support resequencing operations using this variation graph as a reference system. As a proof of principle, we construct a whole genome variation graph from the human reference genome and the 5008 haplotypes in the 1000 Genomes Project's Phase 3 release. The resulting serialized, compressed graph is approximately the same size as the uncompressed human reference sequence, and can be constructed in several hours on a commodity server. To enable read alignment against this reference system, we store the graph and its k-mers in a disk-backed, sorted key-value store. We then align reads against the variation graph using a seed and extend alignment algorithm similar to that employed by standard hash-based read mappers, but generalized to work over partial orders. The result is an efficient basis for resequencing analysis against variation graphs constructed from many thousands of human genomes. Approximately 99% of the variation in a new sample may be directly mapped to known variants represented in this graph, simplifying the process of genotyping and allowing novel variant detection to focus on the remaining 1% of the variation. We are exploring alternative indexing systems using the Generalized Compressed Suffix Array (GCSA) approach, which generalises Burrows-Wheeler mapping to sequence graphs, and has the potential to support efficient mapping on a standard server without the disk-backed store.

THE RESURGENCE OF REFERENCE QUALITY GENOMES

Hayan Lee[1,2], James Gurtowski[1], Shinjae Yoo[3], Maria Nattestad[1], Shoshana Marcus[4], Sara Goodwin[1], Richard W McCombie[1], Michael C Schatz[1,2]

[1]Cold Spring Harbor Laboratory,, Cold Spring Harbor, NY, [2]Stony Brook University, Department of Computer Science, Stony Brook, NY, [3]Brookhaven National Laboratory, Computational Science Center, Upton, NY, [4]Kingsborough Community College, City University of New York, Brooklyn, Department of Mathematics and Computer Science, Brooklyn, NY

Several new 3rd generation long-range DNA sequencing and mapping technologies have recently become available that are starting to create a resurgence in genome sequence quality. Unlike their 2nd generation, short-read counterparts that can resolve a few hundred or a few thousand base-pairs, the new technologies can routinely sequence 10,000 bp reads or map across 100,000 bp molecules. The substantially greater lengths are being used to enhance a number of important problems in genomics and medicine, including de novo genome assembly, structural variation detection, and haplotype phasing.

Here we discuss the capabilities of the latest technologies, and show how they will improve the "3Cs of Genome Assembly": the contiguity, completeness, and correctness. We derive this analysis from (1) a meta-analysis of the currently available 3rd generation genome assemblies, (2) a retrospective analysis of the evolution of the reference human genome, and (3) extensive simulations with dozens of species across the tree of life.

We also propose a model using support vector regression (SVR) that predicts genome assembly performance using four features; read lengths(L) or coverage values(C) that can be used for evaluating potential technologies and genome size(G) and repeats(R) that present species specific characteristics. The proposed model significantly improves genome assembly performance prediction by adopting data-driven approach and addressing limitations of the previous hypothesis-driven methodology.

Overall, we anticipate these technologies unlock the genomic "dark matter", and provide many new insights into evolution, agriculture, and human diseases.

TRACTABLE HAPLOTYPE PHASING AND IMPUTATION WITH NONPARAMETRIC FRAGMENTATION-COAGULATION PROCESSES

Derek Aguiar[1], Lloyd T Elliott[3], Yee Whye Teh[3], Barbara E Engelhardt[1,2]

[1]Princeton University, Computer Science, Princeton, NJ, [2]Princeton University, Center for Statistics and Machine Learning, Princeton, NJ, [3]University of Oxford, Department of Statistics, Oxford, United Kingdom

Haplotype data are essential for many informative and essential analyses of genomic data, including allele specific expression and multilocus association mapping. But haplotype dependent analyses suffer from two problems that reduce their efficacy in genomics and medical research. First, haplotypes are often not known precisely: experimental methods to characterize haplotypes are prohibitively expensive, whereas computational approaches either use complex models that are intractable for large samples, or use oversimplified models for computational efficiency. Second, many haplotype-based analyses use haplotype sequences as a proxy for the latent genealogy (e.g., demographic history and association mapping); these analyses would greatly benefit from a representation that captures the rich genealogy of the sample in a computationally tractable and statistically robust framework. Here, we present a Bayesian nonparametric (BNP) and nonconjugate statistical model, computationally tractable posterior inference algorithms, and haplotype phasing and imputation methods that improve on the Li & Stephens (2003) PAC model by adding exchangeability, while maintaining the computational efficiency of HMM-based models. First, we estimate a biologically meaningful latent structure, called a *haplotype cluster graph*, which captures the evolutionary relationships between haplotypes using reference haplotypes; the haplotype partitioning process is exchangeable and thus does not depend on sampling order. Tractable genome-scale posterior inference of the haplotype cluster graph is achieved with maximization-expectation. Then, we exploit this compressed latent structure of reference haplotypes to make haplotype phasing and imputation of unphased genotypes tractable. We apply our methodology, *phaseME*, to compute complete genome-wide haplotype cluster graphs, haplotype phasings, and imputations using the 1000 Genomes Project data. We compare the results from phaseME with leading haplotype phasing and imputation software based on precision and computational resources to illustrate the benefits of model expressibility in this tractable framework.

IDENTIFYING THE TARGET OF SELECTIVE SWEEP.

Ali Akbari[1], Glenn Tesler[2], Roy Ronen[3], Yu Lin[4], Noah Rosenberg[5], Vineet Bafna[4]

[1]UC San Diego, Department of Electrical & Computer Engineering, La Jolla, CA, [2]UC San Diego, Department of Mathematics, La Jolla, CA, [3]UC San Diego, Bioinformatics Graduate Program, La Jolla, CA, [4]UC San Diego, Department of Computer Science & Engineering, La Jolla, CA, [5]Stanford University, Department of Biology, Stanford, CA

Methods for detecting the genomic signatures of natural selection have been heavily studied, and have been successful in identifying many selective sweeps. For the vast majority of these sweeps, the favored allele remains unknown, making it difficult to distinguish carriers of the sweep from non-carriers. Because carriers of ongoing selective sweeps are likely to contain a future most recent common ancestor, identifying them may prove useful in predicting the evolutionary trajectory – for example, in contexts involving drug-resistant pathogen strains or cancer subclones. In this research, we start with the development and analysis of a new statistic, the Haplotype Allele Frequency (HAF) score, assigned to individual haplotypes in a sample. The HAF score naturally captures many of the properties shared by haplotypes carrying an adaptive allele. We provide a theoretical model for the behavior of the HAF score under different evolutionary scenarios, and validate the interpretation of the statistic with simulated data. We develop an algorithm (PreCIOSS: Predicting Carriers of Ongoing Selective Sweeps) to identify carriers of the adaptive allele in selective sweeps, and we demonstrate its power on simulations of both hard and soft selective sweeps, as well as on data from well-known sweeps in human populations.

We also investigate the problem of identifying the favored mutation itself, without using functional information. We developed a tool to reconstruct possible genealogical scenarios, using clusters of recurrent mutations as guides to ancestral conserved haplotypes. We use a novel algorithm to select conserved haplotype segments that best separate carriers and non-carriers identified by PreCIOSS. In simulations, and on real data-sets with known sites under positive selection, our tool is able to short-list the favored mutation with high sensitivity and specificity. Applying this method on a 600 kbp (163 sites) region including the lactase (LCT) gene region centered at the known favored mutation in European (CEU) population, we obtained a small list of 3 candidates that includes the C/T-13910 (rs4988235) mutation, for which the T allele was found to be 100% associated with lactase persistence in the Finnish
population. Similar results were observed for other selective sweeps including PSCA in African (YRI), and ADH1B and EDAR in Asian (CHB+JPT) populations from the Hapmap2 project.

← Back

INFERRING RECENT DEMOGRAPHY SURFACES VIA HAPLOTYPE SHARING

Hussein Al-Asadi[1], Desislava Petkova[2], John Novembre[3], Matthew Stephens[3,4]

[1]University of Chicago, Evolutionary Biology, Chicago, IL, [2]University of Oxford, Wellcome Trust Centre of Human Genetics, Oxford, United Kingdom, [3]University of Chicago, Human Genetics, Chicago, IL, [4]University of Chicago, Statistics, Chicago, IL

Long stretches of sequence similarity between chromosomes are informative of recent ancestry and allow inference of demography in the recent past. We model the number of these stretches shared between pairs of individuals to infer a migration and population size surface for the recent past under a model of isolation by distance. Long stretches of sequence similarity correspond to long non-recombining (NR) segments. As these NR segments become larger, they become increasingly sensitive to recent coalescent events. Furthermore, long NR segments are only a function of recent demography, permitting us to simplify computation by ignoring demography in the distant past. In this work, we present a coalescent based method to model long NR segments. Since we are only interested in inferring demography in the recent past, we only consider NR segments greater than a threshold of 3 cM which can be inferred from long stretches of sequence similarity with almost perfect accuracy. Furthermore, we allow the user to look at specific length ranges depending on the time range of interest, for example [10cM, Inf) for the very recent past. We apply our method to array data from thousands of Europeans and find the inferred population size and migration surface reflects recent historical events.

A MODEL FOR LONG RANGE TRANSCRIPTIONAL REGULATION USING A SELF-AVOIDING WORMLIKE CHAIN APPROACH

Roee Amit[1,2]

[1]Technion - Israel Institute of Technology, Russell Berrie Nanotechnology Institute, Haifa, Israel, [2]Technion - Israel Institute of Technology, Department of Biotechnology and Food Engineering, Haifa, Israel

Transcriptional regulation-at-a-distance is a poorly understood phenomenon in Biology. Even though it is well-accepted that many non-gene coding sequences are capable of regulating the expression of genes, which are located hundreds to hundreds of thousands base-pairs away, there is no satisfying mechanistic explanation for this ubiquitous phenomenon. Here, we will show that polymer-physics considerations can provide a suitable explanation for the repression phenomenon termed "quenching", where a transcription factor is able to down-regulate expression just by seemingly being bound to the DNA. To do so, we numerically simulate chromatin as a thick chain with protrusions, corresponding to DNA occupied with transcription factors, and compute the resultant probability of cyclization or looping (i.e. when the two ends are in close vicinity to one another, given biologically-relevant boundary conditions). Our results show that protrusions placed near either of the chain's ends affect the probability of looping in the entropic regime, independently of looping length. These entropic effects are characterized by the reduction of the probability of looping, and can be explained by a form of "eclipse" or "occlusion" phenomenon, where the bound protein effectively interferes with the ability of one end of the chain to fully view or access the other. Our results indicate that the arrangement of protrusions, their total number, and whether or not they locally bend DNA can alter the looping inhibition or quenching effect over a wide-range as compared with the probability of looping for a thick chain without protrusions. This long range-effect does not emerge in the conventional wormlike chain model, and is a consequence of taking the volume of both the thick chain and the protrusions into consideration. Finally, our results imply that given these entropic effects, proteins can quench expression from extremely large distances, provided that a loop is formed between a close target site in their vicinity (<200 bp) to a promoter located far away.

A NOVEL MIXTURE MODEL OF RNA POLYMERASE II DYNAMICS

Joseph G Azofeifa[1], Mary A Allen[2], Robin D Dowell [1,2]

[1]University of Colorado, Computer Science, Boulder, CO, [2]University of Colorado, BioFrontiers Institute, Boulder, CO

Growing evidence from the ENCODE consortium and genome-wide association studies suggests that disease causing variation is primarily in non-protein coding regions of the genome. In addition, 73% of all GWAS single nucleotide polymorphisms correlate with enhancer chromatin marks such as H3K27ac, H3Kme1 and DNAse-hypersensitivity, highlighting the phenotypic consequences of altered transcriptional regulation. The majority of such enhancer regions are themselves transcribed to unstable RNA molecules, termed enhancer RNAs (eRNAs). Due to their instability, eRNAs are best detected by nascent transcription assays such as Global Run On-sequencing (GRO-seq). Unlike RNA-seq which quantifies steady state mRNA levels, GRO-seq measures areas of active transcription and thus serves as a unique read out of RNA Polymerase-II (Pol-II) dynamics. Hence we sought to develop a fully generative, probabilistic model of Pol-II behavior at both protein coding genes and cryptic non-coding areas of the genome.

Briefly, our model assumes that Pol-II binds DNA at key positions (transcription start sites or eRNAs) as a Gaussian distribution. Once loaded, Pol-II binds the forward or reverse strand as a Bernoulli trial and processes to some short distance as an exponential random variable. Finally, our model predicts that Pol-II will remain paused (with some probability) or elongate (with the remaining probability) as a homogenous Poisson point process, terminating after some genomic distance. Even with a complex mixture model composed of non-exponential family components, we derived the Expectation-Maximization algorithm to fit our model at many thousands of genomic loci in any GRO-seq dataset.

On a global average, we observed no differences in parameter estimates at either gene promoter or enhancer regions suggesting that Pol-II functions similarly regardless of a translated protein. We observe a strong correlation between our predicted Pol-II loading positions and enhancer marks like H3K27ac, H3K4me[1/3] and many transcription factors available through the ENCODE project. Strikingly upon activation of a transcription factor p53, we see a marked decrease in the variance of the Pol-II loading event over eRNAs containing a p53 binding motif.

GENOME-BASED CHARACTERIZATION OF INVASIVE, OTITIS ASSOCIATED AND CARRIAGE NON-TYPEABLE *HAEMOPHILUS INFLUENZAE* (NT*HI*) ISOLATES

Mara Barucco[1,2], Gabriella De Angelis[2], Monica Moschioni[2], Silvia Guidotti[2], Giulia Torricelli[2], Nicola Pacchiani[2], Mariagrazia Pizza[2], Stefano Censini[2], Marco Soriani[2], Alessandro Muzzi[2]

[1]Università di Pisa, Department of Physics "Enrico Fermi", Pisa, Italy,
[2]GSK Vaccines, Research Center, Siena, Italy

The huge genomic variability of non-encapsulated isolates of *Haemophilus influenzae* (NT*Hi*) is exploited by the pathogen to adapt to the host environment and reflects on the capability of this pathogen to cause invasive disease and to be associated with common pediatric diseases, including otitis media (OM) and with exacerbations of chronic obstructive pulmonary disease (COPD) in adults. This heterogeneity also impacts on the power to identify in preclinical studies, vaccine candidates inducing a cross-protective response versus infectious strains. Previously, a comprehensive analysis of NT*Hi* genomic diversity based on a global worldwide collection of isolates, revealed that the population structure of the species is constituted by distinct evolutionary clades, predominantly shaped by clonal evolution, with the majority of genetic information transmitted vertically within lineages. In the present study, we analyzed a panel of 103 isolates (provided by Ron Dagan), collected in the same hospital in Israel, between 2005-2012, divided in three groups depending on the disease outcome (35 otitis media, 34 invasive disease, 34 from healthy individuals). Based on whole-genome-sequencing (WGS) we looked for correlations between epidemiologic features and genetic characters. Mapping the sequenced genomes to a reference closed and annotated one, we identified specific missing genes, insertions, rearrangements and distribution of polymorphic sites, and we related this genetic variability to the functional annotation of the impacted genes. The identification of genes or their polymorphisms that correlate with epidemiologic features could reveal relevant peculiarities of the NT*Hi* population structure.

STRATEGIES FOR LEARNING UNDIRECTED GRAPHICAL MODELS FOR MIXED DATA TYPES

Andrew J Sedgewick[1,2], Ivy Shi[3], Rory M Donovan[1,2], Panayiotis V Benos[1,2]

[1]University of Pittsburgh, Computational and Systems Biology, Pittsburgh, PA, [2]University of Pittsburgh, Joint CMU-Pitt PhD Program in Computational Biology, Pittsburgh, PA, [3]University of Pittsburgh, Bioengineering, Pittsburgh, PA

Learned mixed graphical models (MGMs) provide both a network structure and a parameterized joint probability density over a combination of continuous and discrete variables, which are common in biomedical datasets. The network structure reveals the direct associations between variables and the joint probability density allows one to ask arbitrary probabilistic questions on the data. This information can be used for feature selection, classification and other important tasks. We studied the properties of MGM learning and applications of MGM models to high-dimensional data (both biological and simulated) and showed that MGMs can reliably uncover the underlying graph structure.and when used for classification or biomarker selection, their performance is comparable to currently available univariate methods (lasso and support vector machines). A mixed model learned over mRNA expression and clinical data from the Lung Genomics Research Consortium correctly recovered connections between the diagnosis of obstructive or interstitial lung disease, two diagnostic breathing tests, and cigarette smoking history. In addition our model suggested relevant mRNA markers that are linked to these three clinical variables. Finally, it identified CYP1A1 cytochrome gene as very important for the facilitation of the disease in people with smoking history.

A COALESCENT MODEL OF A SWEEP FROM A UNIQUELY DERIVED STANDING VARIANT

Jeremy J Berg, Graham Coop

University of California, Davis, Center for Population Biology, Davis, CA

The use of genetic polymorphism data to understand the dynamics of adaptation and identify the loci that are involved has become a major pursuit of modern evolutionary genetics. In addition to the classical ``hard sweep'' hitchhiking model, recent research has drawn attention to the fact that the dynamics of adaptation can play out in a variety of different ways, and that the specific signatures left behind in population genetic data may depend somewhat strongly on these dynamics. One particular model for which a large number of empirical examples are already known is that in which a single derived mutation arises and drifts to some low frequency before an environmental change causes the allele to become beneficial and sweeps to fixation. Here, we pursue an analytical investigation of this model, bolstered and extended via simulation study. We use coalescent theory to develop an analytical approximation for the effect of a sweep from standing variation on the genealogy at the locus of the selected allele and sites tightly linked to it. We show that the distribution of haplotypes that the selected allele is present on at the time of the environmental change can be approximated by considering recombinant haplotypes as alleles in the infinite alleles model. We show that this approximation can be leveraged to make accurate predictions regarding patterns of genetic polymorphism following such a sweep. We then use simulations to highlight which sources of haplotypic information are likely to be most useful in distinguishing this model from neutrality, as well as from other sweep models, such as the classic hard sweep, and multiple mutation soft sweeps. We find that in general, adaptation from a uniquely derived standing variant will be difficult to detect on the basis of genetic polymorphism data alone, and when it can be detected, it will be difficult to distinguish from other varieties of selective sweeps.

IDENTIFYING CAUSATIVE GENES: DECISION MAKING BASED ON THE BIRTHDAY MODEL

Yael Berstein, Shane E McCarthy, W. Richard McCombie

Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Woodbury, NY

Exome sequencing is a popular technique for the identification of disease-causing genes. A number of genes showing mendelian inheritance were identified through this method. However, it remains a challenge to leverage exome sequencing for the study of complex genetic disorders, e.g. schizophrenia and bipolar disorder. The genetic and phenotypic heterogeneity of the data is often a barrier to the detection of causative genes in complex genetic disorders. For example, the aggregation of different rare variants associated with a given disease can make the identification of causal genes statistically challenging. Here we propose a probabilistic model to predict causative rare variants. The model is based on general analysis of coincidences based on a popular probabilistic problem: the birthday problem. Analogically, we consider the probability of samples sharing a variant, as the chance of individuals sharing the same birthday. We evaluate through simulations the performance of our method for identifying causal rare variants in complex disorders. We investigated the effect of the parameters of our model, providing guidelines for its use and interpretation of the results. We implemented this probabilistic method to published data on autism spectrum disorder, hypertriglyceridemia, schizophrenia, and also on a current case-control study on bipolar disorder. Our method can detect rare variants, with minor allele frequency of 1% or less in 1000 genome, the Exome Variant Server, as well as in the affected population samples.

READ-BACKED ESTIMATES OF GENE CONVERSION RATES AND TRACT LENGTHS

Søren Besenbacher[1], The GenomeDenmark Consortium[2], Thomas Mailund[3]

[1]Aarhus University, Department of Molecular Medicine, Aarhus, Denmark, [2]Copenhagen Biocenter, GenomeDenmark, Copenhagen, Denmark, [3]Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark

In the past decade our understanding of recombination has improved substantially by analysing LD-patterns and finding recombination events in multi-generational families. Single generation (trio) data has sofar not been used find recombinations because independently observed SNVs can only be used to phase the child of the trio not the parents. Whole genome sequencing(WGS), however, makes it possible to use read-backed phasing to phase variants that are close together and thus find some recombinations using trio data. We report the first results of using read-backed phasing to study recombination. Instead of using existing software that can perform read-backed phasing we have develloped a new method that is specially tailored to studying recombinations. Our method first finds all pairs of heterozygous variants that occur on the same read or paired read. It then constructs haplotypes from these pairs and uses these haplotypes to find all pairs of variants that are informative about whether a phase-change occured between the two variants during transmission from parent to child. For each of these pairs we can observe whether an even or uneven number of phase-changes happened between the positions. We then use the pasimony principle and find the smallest number phase-changes necessary to explain the data. Using high-coverage WGS data from 50 danish trios sequenced using both paired-end and mate-pair sequencing with seven different insert sizes we find more than five hundred phase-changes between parents and offspring. We can observe that a third of these are accompanied by another phase-change close by and thus part of a non-crossover gene conversion. For the remaining phase-changes we do not know if they constitute a crossover(CO) or are part of a non-crossover(NCO). Half of the phase-changes are found in regions with a recombination rate above 1, compared to only 20% of all the informative pairs of variants. Observed NCO events show a strong allelic bias at heterozygous AT/GC SNVs with 64% (p=0.009) transmitting the GC allele.

ON ESTIMATING THE SHARED GENETIC BASIS OF COMPLEX PHENOTYPES BETWEEN POPULATIONS

Brielin C Brown[1], Noah Zaitlen[2]

[1]UC Berkeley, Computer Science, Berkeley, CA, [2]UC San Francisco, Lung Biology Center, San Francisco, CA

Many phenotypes vary in their distributions around the world, but the extent to which these differences are driven by genetic factors is unknown. Here we consider the problem of computing the correlation of causal variant effect sizes between two populations from summary statistics. This estimate captures the extent to which genetic and environmental modifiers alter effect sizes between populations. Previous approaches (Yang et al 2013 NG) rely on linear mixed models and have three disadvantages: they require primary genotypes and phenotypes, they are computationally expensive, and they assume that effect sizes are inversely proportional to the mean allele frequency between the two populations. Here, we provide a novel method for jointly estimating the genetic correlation (ρ_g) and heritability (h^2) of a phenotype in two populations from summary statistics with no assumption about the relationship between allele frequency and effect size. Our method is based on maximizing the likelihood of ρ_g, h^2_1, and h^2_2 given summary statistics and genotype covariances computed from a reference panel. We show that the distribution of the summary statistics is multivariate normal and that fitting only the diagonal of the covariance matrix yields an estimate in time O(NM), where N is the reference panel size and M is the number of SNPs.

Mixed-model approaches to estimating h^2 assume that effect size is inversely proportional to allele frequency. While violations of this assumption have been shown to cause bias (Speed et al 2012 AJHG), it is small in non-edge cases. We argue that this assumption is a non-starter when comparing two populations: it implies that an allele that is rare in one population but common in the other must necessarily have a larger effect size in the former population. We show that this can cause large differences in the estimated ρ_g in common situations: 0.38 when the true correlation is 0.5 and the genetic effects are randomly distributed across alleles. We show via simulation that our approach is unbiased even when only a small fraction of the variants are causal. It also extends naturally to partitioning via functional category: with a standard error of 0.09 when there are 50k individuals and 50k SNPs we are able to apply our h^2 estimator to smaller regions than prior approaches (Bulik-Sullivan et al 2015 NG). Finally, we apply our estimator to the Geuvadis dataset to determine the genetic correlation of gene expression between Yoruban and European populations. We find that when the heritability of expression is non-trivial the genetic correlation is high: $\rho_g = 0.782$ (0.017) across 1382 genes with $h^2 > 0.05$ in both populations.

GENOTYPE IMPUTATION WITH MILLIONS OF REFERENCE SAMPLES

Brian L Browning[1,2,3], Sharon R Browning[2]

[1]University of Washington, Medicine, Seattle, WA, [2]University of Washington, Biostatistics, Seattle, WA, [3]University of Washington, Genome Sciences, Seattle, WA

The pace of high-coverage, whole-genome sequencing is accelerating, and tens of thousands of individuals are being sequenced each year. This makes it possible to assemble reference panels of unprecedented size which can be used to impute very low frequency variants.

We present a new genotype imputation method that scales to reference panels with millions of individuals. The new imputation method, implemented in Beagle v4.1, is parallelized and memory efficient, making it ideally suited to multi-core computer processors. The method is based on the Li and Stephens model, but includes additional methodological innovations and computational optimizations to achieve fast, accurate, and memory-efficient imputation.

We compare Beagle with Impute2 and Minimac3 using 1000 Genomes Project data and simulated data. The three methods have very similar accuracy, but different memory requirements and different computation times. When imputing sequence data from 50,000 reference samples, Impute 2 required 110 GB of memory and could utilize only one CPU core at a time on our server due to memory constraints. Beagle's throughput when imputing from 50,000 reference samples was 600x greater than Impute2's throughput. When imputing 10Mb of sequence data from a 100,000 reference samples, Minimac3 required 7.5x more memory per computational thread than Beagle, used 50x more wall-clock time than Beagle, and used 7.3x more CPU time than Beagle. When using 100,000 reference samples, Beagle accurately imputed minor allele dose ($r2 \geq 0.8$) in variants with 15 or more copies of the minor allele in the reference panel ($MAF \geq 7.5 \times 10^{-5}$).

We demonstrate that Beagle's new imputation method scales to much larger reference panels by performing imputation using a simulated reference panel having 5 million samples and a mean variant density of one variant per 8 base pairs.

A MIXTURE MODEL FOR BIAS AND ERROR IN GENOMIC DATA REDUCES FALSE POSITIVE IDENTIFICATION OF HETEROZYGOTES

Reed A Cartwright[1,2], Steven H Wu[1], Rachel S Schwartz[1], David J Winter[1]

[1]Arizona State University, The Biodesign Institute, Tempe, AZ, [2]Arizona State University, School of Life Sciences, Tempe, AZ

Studying the process of *de-novo* mutation from deep-sequencing of related samples is a difficult task. Because *de novo*s are rare, artifacts generated by experimental and biological error tend to be more common than true positives. While *de novo*s can be identified through validation, this is a slow process. In order to estimate mutation rates on large datasets in an automated way, we need to develop new probabilistic models that can handle sources of false positives.

To this end, we have develop a new model for calculating genotype-likelihoods using a mixture of Dirichlet-multinomial distributions. This model allows us to fit (1) reference biases, (2) error rates, and (3) correlation of reads (over-dispersion) found in next-generation datasets. We have applied this model to mutation-accumulation experiments in *Tetrahymena thermophila* and determined that it can eliminate all false positive mutation calls.

In order to estimate genotype-likelihood models for human-sized datasets, we developed an expectation-maximization algorithm for rapidly fitting data to a mixture of Dirichlet-multinomials and estimating the Fisher information matrix of these estimates. This algorithm uses the observed information matrix to accelerate optimization producing quadratic convergence.

We evaluated our model on human data using heterozygous sites from the 1000-genomes CEU trio and haploid sites from the CHM1 cell line. We found that models containing two to four components provided the best fit to our data. Estimated model parameters were consistent across multiple sequencing runs (CEU trio 2010, 2011, 2012, and 2013 datasets) and genomic regions (chromosomes 10 and 21). Within the mixture, one component fits >95% of sites and has low reference bias, low error rates, and low over-dispersion. The minor components typically have higher bias, error, and over-dispersion.

After including sites that are called heterozygous in the CEU trio by simple methods but are not known to be polymorphic in humans, we find that a smaller proportion of sites fit the major component, indicating a likely increase in false positive heterozygotes in this dataset. These possible false positives were also enriched for regions with copy-number variation, indicating pseudo-heterozygotes produced by paralogous regions aligning to the same section of the reference genome. This result suggests that, without any validation or knowledge of segregating variation, it would be possible to use a mixture of Dirichlet-multinomials to identify low-quality heterozygous calls based on which component of the mixture they fall into.
These results are promising, and are already being used in open-source mutation-calling software available from the Cartwright Lab.

A STATISTICAL MODEL FOR DETECTING SIGNIFICANT CHROMATIN INTERACTION AT A FINE RESOLUTION FROM HI-C DATA

Mark A Carty[1,2], Alvaro Gonzalez[1], Lee Zamparo[1], Rafi Pelossof[1], Olivier Elemento[2], Christina Leslie[1]

[1]Memorial Sloan Kettering Cancer Center, Computational Biology, New York, NY, [2]Weill Cornell Medical College, Institute of Computational Biomedicine, New York, NY

We present a generalized linear model (GLM) to account for systematic source of variation in Hi-C read count data, including the dependence of random polymer ligation on genomic distance and GC content and mappability bias, while correctly modeling zero inflation and overdispersion. We use a hurdle negative binomial regression model for the Hi-C contact maps, which models the zero counts with a binomial logistic regression and uses a negative binomial regression to model the non-zero counts. Our model describes the background noise distribution that accounts for the majority of interaction bins that are supported by low read counts. We show that our model can be used to assign significance to chromatin interactions while reducing false positives relative to existing methods. Our approach can identify interactions at the sub-topological domain level from high-resolution Hi-C data, including CTCF-mediated loops, enhancer-promoter interactions, and aggregations of multiple promoters.

DETERMINANTS OF FINE-SCALE MUTATION RATES IN GERMLINE AND SOMA

Chen Chen[1,2], Joseph K Pickrell[1,2,4], Molly Przeworski[1,3,4]

[1]Columbia University, Department of Biological Sciences, New York, NY, [2]New York Genome Center,, New York, NY, [3]Columbia University, Department of System Biology, New York, NY, [4]co-supervised this project

Understanding the determinants of fine-scale mutation rates is of crucial importance for many efforts in human genetics, as well as in evolutionary biology. Recent sequencing of human tumors and of germline mutations in trios has led to the identification of a number of important factors, notably CpG methylation, expression levels, replication timing and GC content. Interestingly, however, some of the effects appear to differ between somatic and germline tissues. Notably, while mutation rates have been reported to increase with expression levels in tumors, no such effect has been detected in the germline, leading to speculation that error or repair processes may differ between tissues. These findings are hard to interpret, however, as, to date, each study has made different analysis decisions. Here, we use a single approach to analyze exome and whole genome sequence data from tumors alongside similar data on germline mutations. We discuss which factors appear to be important determinants of local mutation rates and how they differ between germline and other tissues.

SPECTRAL LEARNING ALGORITHMS FOR COMPARATIVE EPIGENOMICS

Jimin Song[1], Chicheng Zhang[2], Kamalika Chaudhuri[2], Kevin Chen[1]

[1]Rutgers, Genetics, Piscataway, NJ, [2]UCSD, Computer Science, La Jolla, CA

We have developed a new software program, Spectacle, for annotating chromatin states (e.g. enhancers, promoters etc.) from genome-wide chromatin mark data based on Hidden Markov Models (HMMs). Spectacle uses a novel approach to parameter estimation called "Spectral Learning" which we show is both faster than previous software that used the Expectation-Maximization (EM) algorithm and gives biologically more interpretable solutions, as judged by GWAS SNP enrichment. Importantly, the method is robust to class imbalance which is common in biological data sets and particularly severe in chromatin mark data.

Next, we consider chromatin marks from multiple human cell types or tissues. We model the relationship between multiple cell types by connecting the hidden states of the HMMs by a fixed tree of known structure. Since naive Spectral Learning algorithms result in time and space complexity exponential in the number of cell types, we exploit properties of the tree structure of the hidden states to provide novel spectral algorithms that are more computationally efficient. We show that our method, Spectacle-Tree, is much faster and has higher prediction accuracy for promoters than a previous method that used a variational approximation to the EM algorithm.

A COALESCENT HIDDEN MARKOV MODEL FOR INFERRING ADMIXTURE RELATIONSHIPS

Jade Yu Cheng*, Thomas Mailund*

Aarhus University, Bioinformatics Research Centre, Aarhus, Denmark

We develop a coalescence hidden Markov model for inferring parameters for admixture events. By tracing lineages in an admixed population and its source populations back in time and estimating the coalescence times of those lineages, we can infer the split time between the source populations, the time of admixture, and the admixture proportions.

A general admixture scenario involves extant populations A, B, C, and ancestral populations AC', BC', AC, BC, ABC. Population C is admixed from AC' and BC' which are related to A and B, respectively. Population AC is ancestral to A and AC'. Population BC is ancestral to B and BC'. Population ABC is ancestral to AC and BC. Our method infers the split times of A-AC', B-BC', and AC-BC, as well as the admixture time and the admixture proportions, which is α from C to AC' and $1-\alpha$ from C to BC'.

We validate the method using simulated sequences and apply the method to a number of polar bear and brown bear genomes to infer the complex population history of these species.

HLATX: A PIPELINE FOR GENOTYPING AND EXPRESSION ESTIMATION OF HLA GENES USING RNA-SEQ DATA

Jonatas E da Silva Cesar[1], Vitor R.C. Aguiar[1], Nikolaos I Panousis[2], Emmanouil T Dermitzakis[2], Diogo Meyer[1]

[1]University of Sao Paulo, Dept. of Genetics and Evolutionary Biology, São Paulo, Brazil, [2]University of Geneva, Dept. of Genetic Medicine and Development, Geneva, Switzerland

HLA genes encode glycoproteins expressed on the surface of cells and can be divided into two classes. Class I proteins present peptides generated from intracellular infectious agents and also regulate the activation of natural killer cells, while class II proteins present peptides derived from extracellular pathogens . GWAS studies have found SNPs within HLA genes to be associated to numerous autoimmune diseases, and there is strong evidence that an individual's HLA genotype will define resistance/susceptibility to infectious diseases (including HIV and hepatitis). The expression at HLA loci has also been shown to vary among individuals and alleles, and to play a role in defining phenotypes related to response to infections. These features highlight the importance of developing efficient tools to genotype assay the expression of HLA loci. Two main features of these genes pose difficulties to achieve these goals: HLA loci are extremely polymorphic, and are members of gene families with numerous paralogues. As a consequence, expression and genotyping data obtained from standard pipelines are often unreliable for HLA loci. Here, we present the hlaTX python package, which accurately genotypes and estimates expression for HLA-A, HLA-B, HLA-C Class I genes, and HLA-DQB1 and HLA-DRB1 Class II genes using RNAseq data. hlaTX first aligns reads to an index comprising all documented alleles, for the subset of exons which concentrate the highest polymorphism. The individual multi-gene genotype is inferred from the allele combination that maximizes the number of mapped reads with a correction to avoid false heterozygotes. Using this inferred genotype, a new index with the full gene sequence is constructed, and a new mapping round is performed accepting a higher number of mismatches. The resulting counts are submitted to an Expectation Maximization procedure to estimate the expression levels through a close to optimal likelihood model. Using data from the Geuvadis consortium, we show that hlaTX provides accurate estimates of HLA genotypes (>95% accuracy), and that estimates of expression are markedly different from those obtained using standard pipelines (correlation as low as X for HLA-B). We use simulations to show that this can be explained by the large number of reads which are discarded or poorly mapped when using the standard approach, and that by using an index that overcomes this problem hlaTX provides more accurate expression estimates.

A UNIFYING METHOD FOR MULTIPLE PHENOTYPE, MULTIPLE VARIANCE COMPONENT MIXED MODELS

Andrew Dahl[1], Jonathan Marchini[1,2]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [2]University of Oxford, Statistics, Oxford, United Kingdom

Mixed models have recently re-emerged as a state-of-the-art approach to control for relatedness, population structure and genome-wide polygenicity in human GWAS. Moreover, mixed models are becoming even more prominent as the field moves beyond this standard analysis, as they can be broadly used to decompose phenotypic variation into biologically meaningful components. Recently, single phenotype, single random effect mixed models have been generalized in two distinct directions. First, single phenotype mixed models with multiple random effects have been applied to partition heritability amongst functionally meaningful categories, to test gene-based association and to improve phenotype and breeding value estimation. Second, mixed models with multiple phenotypes and one random effect have been used to learn heritable phenotype networks, boost power in association studies and impute missing phenotype data.
We have generalized these approaches by developing the first efficient method to simultaneously model multiple phenotypes and multiple variance components. Furthermore, our method can utilize a wide range of likelihood penalty functions, providing additional statistical robustness, computational efficiency and biological interpretability. Even without penalization, our method can be run on tens of thousands of individuals, hundreds of variance components and dozens of phenotypes.
To demonstrate one possible application of our method, we performed a gene-based association test with multiple phenotypes while correcting for confounding structure. Specifically, we found that the FADS1 gene significantly associates with LDL and triglyceride traits in 5,163 samples from the NFBC19666 study (-log10(p)=8.19). We will also illustrate the performance of this method on simulated data.

STATISTICAL INFERENCE OF TRANSLATION DYNAMICS FROM RIBOSOME PROFILING DATA

Khanh Dao Duc, Yun S Song

UPenn, Mathematics, Philadelphia, PA

Providing detailed positional information of ribosomes on mRNA transcripts that are being translated, ribosome profiling technique gives insights on translation regulation. However, analytical tools for interpreting the data are still much in need of development and the relation between the ribosomal position and the translation dynamics remains elusive. We present here a new method to study these data and quantify the translational process for any specific gene, and more precisely estimate translocation rates and initiation rates. Our approach combines a probabilistic model of translation which takes into account the principal features of DNA translation and is based on the totally asymmetric exclusion process. By imbedding this model in an approximate Bayesian computation (ABC) framework, we developed an algorithm which jointly infers the translation rates from ribosome profiles, and thus determines the protein production rate associated with any specific gene. After showing the accuracy of our algorithm, we apply our methods to a data set of ribosome profiles in S. Cerevisiae. Inferring the translation rates, we then provide new insights on possible local interference between ribosomes, elongation speed position dependency and identification of genes associated with high or low production rate. Consequently, our method can be used as a generic and powerful tool to analyze data from ribosome profiling experiments and study translation.

ROBUST EFFECT SIZE ESTIMATION IN GENOMIC STUDIES: OVERCOMING WINNER'S CURSE

Gregory Darnell[1], Jenny Tung[2], Christopher Brown[3], Sayan Mukherjee[4], Barbara Engelhardt[5,6]

[1]Princeton University, Lewis-Sigler Institute, Princeton, NJ, [2]Duke University, Department of Evolutionary Anthropology, Durham, NC, [3]University of Pennsylvania, Department of Genetics, Philadelphia, PA, [4]Duke University, Departments of Computer Science, Statistical Science, Mathematics, Durham, NC, [5]Princeton University, Computer Science Department, Princeton, NJ, [6]Princeton University, Center for Statistics and Machine Learning, Princeton, NJ

Statistical methods for quantitative trait mapping often overestimate effect sizes, and study conclusions are seldom reproducible, a direct result of a phenomenon referred to as Winner's Curse [Zöllner et al., 2007]. In this work, we study the relationship between the effect size of single nucleotide polymorphisms (SNPs) and their respective minor allele frequency (MAF). We show the functional form of the SNP-MAF relationship for a number of QTL tests and explore how the test-specific functions affect bias in estimates of effect size. Even in the absence of a relationship between the SNP and response (i.e., the null hypothesis of no association), low MAF SNPs often produce phantom heteroskedastic effects, and introduce type 1 errors and upward bias in effect size estimates. We show how to mitigate this bias through use of the Student's t-statistic for implicit regularization. We further correct for Winner's Curse bias in genomics studies by explicitly modeling MAF in a Bayesian framework, which allows us to integrate over the biased standard errors in closed form. In simulated data application, our corrected approaches to univariate association mapping recover a site frequency spectrum that is closer to the null than previous studies. In real data application, our methods achieve greater enrichment of cis-regulatory elements in large eQTL study data.

A PROBABILISTIC FRAMEWORK FOR DATA-DIRECTED RNA SECONDARY STRUCTURE PREDICTION

Fei Deng, Mirko Ledda, Sana Vaziri, Sharon Aviran

University of California, Davis, Biomedical Engineering Department and Genome Center, Davis, CA

RNA structure plays important roles in processes such as mRNA translation, maturation, and decay. Yet, computational approaches to predicting structure from sequence alone are limited in accuracy. Recent experimental approaches to probe secondary structure, such as SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension), can be coupled with structure prediction algorithms to improve their performance. In such structure probing experiments, a chemical adduct modifies RNA residues in a structure-dependent manner. These modifications are then detected and quantified via reverse transcription reactions.

Following the success of SHAPE, advances in chemistry and genomics have spurred the development of new and forthcoming probing techniques as well as of massively parallel approaches to probing structure transcriptome-wide and in living cells. Despite their shared principles, these techniques differ in the types of structural information they extract and in the statistical properties of the data. Currently, existing data-directed structure prediction algorithms have been developed and optimized for only SHAPE. Such algorithms are heuristic and rely on estimated parameters, thus making it difficult to extend them to accommodate more complex scenarios or varied information sources.

Here, we have implemented a principled probabilistic framework for data-directed secondary structure prediction, recently outlined by Eddy. Our framework synthesizes a well-established free-energy minimization algorithm with likelihood-based calculations of probing data per considered structural contexts. As the approach derives explicitly from statistical modeling of the data, it is readily adaptable to a variety of probes and can account for complex interpretations of structural information. Since current SHAPE-directed structure prediction algorithms are limited in their ability to process such enhanced structural contexts, we have improved upon the dynamic programming paradigm to accommodate such contexts.

We test the performance of this approach using real data as well as model-based simulations. We find that proper integration of enhanced structural contexts leads to improved predictive ability. We also investigate the effects of data pre-processing strategies on the performance of this approach. Through such systematic explorations, we gain insights into how the statistical properties of the data drive structure prediction. These can in turn guide us in the design of improved structure prediction algorithms, refined statistical models, and pre-processing strategies for probing-directed structural analysis.

A BAYESIAN TO IDENTIFY VARIANCE QUANTITATIVE TRAIT LOCI

Bianca M Dumitrascu[1], Gregory Darnell[1], Julien Ayroles[1,4], Barbara Engelhardt[2,3]

[1]Princeton University, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ, [2]Princeton University, Computer Science, Princeton, NJ, [3]Princeton University, Center for Statistics and Machine Learning, Princeton, NJ, [4]Harvard University, Department of Organismic and Evolutionary Biology, Cambridge, MA

Phenotypic variation is the result of complex interplay between genotype and environment. Of central interest to the field of statistical genetics, mapping and quantifying the impact of genotype on phenotypic variation play an integral role in our understanding of complex traits, adaptation, and selection. To date, computational methods to identify genetic effects have focused on genetic variants or quantitative trait loci (QTL) with mean effects; many studies are underpowered to detect other types of association, and some believe that non-standard effects constitute the minority of genetic regulation of complex traits. We are interested in finding QTLs with variance effects, or heteroskedastic effects. To do this, we developed a robust Bayesian framework to test for heteroskedasticity (BTH) using a heteroskedastic linear regression model. Our model explicitly encodes the dependence between phenotype and the variance controlling genomic loci. We compute Bayes factors as the test statistic, integrating over uncertainty in all model parameters with appropriate regularizing priors. We developed fast approximation algorithms to scale the test to genome-wide application. Our method outperforms classical statistical tests for difference in variance across groups, such as the Levene and Brown-Forsythe tests, in diverse simulations, including loci with low minor allele frequency, mean effects, small samples, and imputed genotypes. As empirical validation of our method, we apply BTH to find variance QTLs in gene expression data and methylation data. We further explore the variance effects of known confounders such as sex and ethnicity on gene expression levels. Our results draw attention to the substantial but generally ignored role that variance effects plays in the genetics of gene regulation.

INFERRING EPISTASIS AND FITNESS LANDSCAPES FROM GENOMIC VARIATION WITHIN NATURAL BACTERIAL POPULATIONS

Daniel Falush

University of Swansea, Medical Sciences, Swansea, United Kingdom

In order to be propagated within bacterial populations with high recombination rates, genetic elements need be successful within particular clones but also to transmit effectively between clones. These dual forces lead to the selection of genes that are successful as part of coalitions. We describe statistical tests to identify these coalitions using multiple bacterial genome sequences. The approach can be thought of as a scan for epistasis, a top down way to estimating fitness landscapes or alternatively a GWAS for fitness. The main challenges for such scans are to identify an appropriate set of genomes and to control for confounding patterns of variation that arise due to clonal or population structure. I will describe the approaches we have taken in three species that present distinct challenges, namely Campylobacter coli, Vibrio parahaemolyticus and Helicobacter pylori, and outline unsolved methodological challenges in terms of identifying and interpreting epistatic interactions using population genomic data.

HLAASSOC: TESTS FOR ASSOCIATION BETWEEN HLA ALLELES AND DISEASES

Yanhui Fan[1], You-Qiang Song[2]

[1]The University of Hong Kong, Center for Genomic Sciences, Hong Kong, Hong Kong, [2]The University of Hong Kong, Department of Biochemistry, Hong Kong, Hong Kong

High-throughput and cost-effective HLA typing form genome-wide genotyping and next-generation sequencing data is feasible now. We present HLAassoc, a tool that capable of test associations between HLA alleles and diseases. Pearson's chi-squared test or Fisher's exact test can be performed on a 2 × 2 or 2 × n contingency table which contains the counts of allele(s) for a single locus in cases and controls. The Logistic regression and linear regression can include multiple covariates when testing for disease trait and quantitative trait, respectively. Then the p-values were adjusted by using multiple testing adjustment method such as the Bonferroni correction or false discovery rate (FDR) correction. The empirical p-values can also be calculated using permutation test, which randomly shuffled the phenotypes for individuals while keep the HLA alleles unchanged. HLAassoc is implemented in Python. HLAassoc is free, open-source software distributed under the GPLv2 open-source license and is available at https://github.com/felixfan/HLAassoc.

ACCURATE PREDICTION OF A-SITE LOCATION AND ROBUST MODELING OF TRANSLATION CONTROL WITH RIBOSOME PROFILING DATA

Han Fang[1,2,3], Max Doerfel[2], Gholson J Lyon[2], Michael C Schatz[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative BiologY, Cold Spring Harbor, NY, [2]Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Cold Spring Harbor, NY, [3]Dept. of Applied Math and Statistics, Stony Brook University, Stony Brook, NY

Ribosome profiling (Riboseq) has become a promising technique for monitoring protein translation in vivo. Several studies have suggested that expression measurements generated by Riboseq better explain the variance of protein abundance, relative to RNAseq data alone. However, there are very few methods available to jointly analyze Riboseq and RNAseq data in a systematic and standardized fashion. It is also unclear how to determine the A-site location on a read since Riboseq does not provide this information directly.

Here, we present a novel framework for joint analysis of Riboseq and RNAseq data. This framework includes an end-to-end solution from adapter trimming to expression quantification. We provide modules for between-sample normalization of expression levels, translation efficiency (TE) estimation, and ribosome A-site prediction. We show that with careful optimization of mapping, Cufflinks and Salmon are highly concordant on gene-level expression quantification on Riboseq data. Based on reads accumulating near the start codon, we used a SVM classifier to directly learn key features of A-site location along a Riboseq read. We observed strong dependencies of A-site location on both read length and codon offset, which might be due to variable effects of the digestion enzyme on individual ribosomes.

To demonstrate the effectiveness of this framework, we compared data from wild-type and knockout yeast strains involving ARD1, the catalytic subunit of the major N-terminal protein acetylation complex, NatA. This novel prediction method indeed has much higher accuracy for identifying A-site location than previous methods (0.85 vs. 0.64, 10-fold CV). Both Riboseq and RNAseq data showed that the mutant has significant reduction in expression of genes related to conjugation and mating (BH adjusted p-value: 3.54E-13). Genes involved in multi-organism processes have a significant reduction of TE (BH adjusted p-value: 5.4E-6), and the corresponding transcripts are mostly down regulated. Our method also identified 18 genes that have two-fold or more stop codon read-through in the mutant, relative to wild type. GO analysis suggested these read-through events are enriched in cytosolic large ribosomal subunit genes (BH adjusted p-value: 0.02). Ongoing work includes joint modeling of the translation initiation and elongation rates, which could yield a robust inference of protein synthesis rate.

LEARNING RNA STRUCTURE (ONLY) FROM STRUCTURE PROBING DATA

Cristina Pop*[1], Chuan Sheng Foo*[1], Rhiju Das[2], Daphne Koller[1]

[1]Stanford University, Computer Science Department, Stanford, CA,
[2]Stanford University School of Medicine, Department of Biochemistry, Stanford, CA
* equal contribution

The structure of an RNA molecule is critical to its function. Structured regions in an RNA molecule permit or impede the binding of proteins and small molecules, resulting in downstream effects on gene expression. While individual RNA structures are most accurately determined through low-throughput experimental means such as NMR spectroscopy or X-ray crystallography, recently introduced sequencing-based RNA structure-probing assays show promise as a high-throughput alternative. These assays reveal which nucleotides are paired and which are not, but cannot determine specific pairing partners. Structure-probing data have therefore been used in conjunction with computational methods in order to infer complete RNA structures.

We present a novel probabilistic method for RNA secondary structure prediction, CONTRAfold-SE, that models structure-probing data as observations of possibly unknown secondary structures. CONTRAfold-SE extends the CONTRAfold model for predicting secondary structure from sequence and can be learned from datasets containing only structure-probing data, or a mix of known structures and probing data; by contrast, CONTRAfold requires a set of complete structures to learn a model. Our method provides two key advantages. First, it obviates the need for heuristic treatments of the probing data (as in existing methods), such as thresholding the data to a binary value (reflecting whether a base is paired or not) or incorporating it in the energy model as a pseudo-energy term. Second, our probabilistic approach provides a principled framework for combining data from multiple structure-probing experiments. Each probing strategy has specific biases, and combining data obtained from small-scale experiments using different strategies has been shown to improve prediction accuracy.

We evaluated CONTRAfold-SE using a large collection of RNA structure probing data deposited in the RNA mapping database. We show that when predicting the structure of a novel sequence, CONTRAfold-SE is competitive with current methods, and outperforms CONTRAfold, whether using structure-probing data available for the novel sequence or just the sequence itself. We find that combining datasets using different structure-probing methods allows for cross-correction of errors in the data. We also trained CONTRAfold-SE on probing data alone and show that it is able to learn a reasonable model for secondary structure, suggesting that large, diverse sets of structure probing data could potentially be used to learn improved models for secondary structure in the future.

← Back

STATISTICAL MODELLING OF B CELL REPERTOIRES RESPONDING TO VACCINATION

Anna Fowler[1], Jacob D Galson[2], Marton Munz[1], Johannes Truck[2,3], Dominic Kelly[2], Gerton Lunter[1]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [2]University of Oxford, Oxford Vaccine Group, Department of Paediatrics, Oxford, United Kingdom, [3]University Children's Hospital, Paediatric Immunology, Zurich, Switzerland

B cell receptors (BCRs) are a component of the adaptive immune system that recognise and bind antigens. In order to have a healthy immune system, a diverse set of BCRs, capable of recognising many different antigens, is required. This BCR diversity is generated through a complex process of somatic recombination and hyper-mutation, thought to be capable of generating 10^{11} unique BCRs. NGS has allowed us to capture these somatic differences at the resolution of individual B cells, through targeted mRNA sequencing of the third complementary determining region (CDR3) which is the most variable region of the BCRs.

In response to stimulation, in this case vaccination, B cells that bind the antigen will proliferate and hyper-mutate. Through the statistical analysis of BCR repertoires collected pre- and post-vaccination in multiple individuals who have received the same vaccine, we aim to identify those CDR3 sequences that are vaccine-specific. This has the potential to provide a cheap and accurate correlate of immune protection, and add to our understanding of immunology.

We use a clustering algorithm to account for read errors and a small number of somatic mutations. The speed and simplicity of the clustering algorithm allows us to naively cluster samples from all subjects and all time points simultaneously, and provides a robust way to profile B cell cluster across time and subjects. To model the response, we use a mixture model and assume that in any single sample, these clusters are derived from B cells in one of three underlying states: not present in the sample; present, but not responding to a stimulus; or responding to a stimulus. B cells that are likely vaccine specific can then be identified as those that are quiescent prior to vaccination, and are likely to be in the responding state post vaccination.

In multiple data sets containing subjects vaccinated against different diseases, we identify small sets of B cells that are likely to be vaccine specific. These show evidence for an immune response in plasma cells peaking at day 7 post vaccination, as expected. They also show evidence for convergent evolution of the CDR3 in B cells responding to the same vaccine.

BAYESIAN SPARSE REGRESSION ANALYSIS DOCUMENTS THE DIVERSITY OF SPINAL INHIBITORY INTERNEURONS

Mariano I Gabitto[1], Ari Pakman[2], Jay B Bikoff[1,4], Larry F Abbott[1,3], Thomas M Jessell[1,4], Liam Paninski[1,2]

[1]Neuroscience, Neuroscience, New York, NY, [2]Statistics, Department of Statistics and Grossman Center for the Statistics of Mind, New York, NY, [3]Physiology, Department of Physiology and Cellular Biophysics, New York, NY, [4]HHMI,Kavli, Biochemistry, Howard Hughes Medical Institute, Kavli Institute for Brain Science Department of Biochemistry and Molecular Biophysics, New York, NY

Documenting the extent of cellular diversity is a critical step in defining the functional organization of tissues and organs. We have devised a sparse Bayesian framework that infers cell type diversity from partial or incomplete transcription factor expression data. This framework appropriately handles estimation uncertainty, can incorporate multiple cellular characteristics, and can be used to optimize experimental design. Through a focus on spinal V1 inhibitory interneurons, for which we have spatially mapped the expression of 19 transcription factors, we infer the existence of approximately 50 V1 cell types, most of which localize in compact spatial domains in the ventral spinal cord. Moreover, their expression profiles can be ordered into mutually exclusive clades that are informative for genetic targeting. Finally, we have validated inferred cell types by direct experimental measurement, establishing this Bayesian framework as an effective platform for cell type characterization in the nervous system and elsewhere.

USING ALLELE SPECIFIC EXPRESSION TO FIND EXPRESSION QTLS USING A BAYESIAN OVERDISPERSED POISSON GLM

Genna R Gliner[1], Yoson Park[2], Christopher Brown[2], Barbara Engelhardt[3]

[1]Princeton University, Operations Research and Financial Engineering Department, Princeton, NJ, [2]University of Pennsylvania, Department of Genetics, Perelman School of Medicine University of Pennsylvania, Philadelphia, PA, [3]Princeton University, Computer Science Department and Center for Statistics and Machine Learning, Princeton, NJ

Modeling the relationship between genetic regulatory elements and gene expression levels through allele specific expression (ASE) is a powerful approach to finding causal genetic variants. Current ASE detection methods use the relative abundance of expression of two alleles at a genetic locus in a heterozygous individual to search for statistically significant differential RNA-seq read counts across the two alleles. With a few exceptions, most methods do not share signal strength across tissues, samples, or coding genetic loci. We present a Bayesian model that applies an overdispersed Poisson linear model (OPLM) to identify ASE using haplotypes from phased genotype data and mapped read counts from RNA-seq data. The OPLM incorporates observations across all samples and all coding loci in an exon. Additionally, it includes a random effect to control for relatedness among samples. The model predictors include the set of local, phased, non-coding haplotypes used to identify SNPs that are most predictive of ASE signal; these SNPs will be candidate eQTLs. We also include phase uncertainty in our model to control for inaccuracies in phasing. We apply our model to simulated data and the Genotype Tissue Expression (GTEx) consortium data to show how we improve upon existing ASE detection methods and are able to fine-map cis-eQTLs through ASE.

A MECHANISTIC MODEL OF ASSORTATIVE MATING BY ANCESTRY IN AN ADMIXED POPULATION

Amy Goldberg[1], Ananya Rastogi[2], Noah A Rosenberg[1]

[1]Stanford University, Biology, Stanford, CA, [2]Indian Institute of Science Education and Research, Biology, Mohali, India

Empirical studies have suggested that spouses from admixed human populations are correlated in their ancestry components. To study the effects of assortative mating by admixture on the distribution of admixture levels in an admixed population, we generalize a two-sex mechanistic admixture model to permit flexible mating probabilities as a function of the admixture fractions of a mating pair. Under the model, we study the moments of the admixture fraction distribution for different models of mating, including both assortative and disassortative mating by ancestry. Through simulations of the model and analytical solutions in special cases, we demonstrate that the mean admixture level in an assortatively mating population is equivalent to that of a corresponding randomly mating population. The variance of admixture, however, increases with assortative mating and decreases with disassortative mating. For a constant admixture model, assortative mating changes the trajectory of the variance of the admixture fraction distribution over time. As the variance of admixture estimated from data on an admixed population is important in studying the timing of admixture and the properties of sex-biased admixture, we suggest that accounting for assortative mating can improve inferences from current admixture patterns about the history of the admixture process.

INTEGRATING GENE EXPRESSION AND CHROMATIN ACCESSIBILITY TO LEARN DIFFERENTIATION PROGRAMS OF HEMATOPOIESIS AND LEUKEMOGENESIS USING CONFIDENCE-RATED BOOSTING

Peyton Greenside[1], Jason Buenrostro[2,3], Ryan Corces-Zimmerman[4], Ravi Majeti[5], Howard Chang[3], Anshul Kundaje[2,6]

[1]Biomedical Informatics Training Program, Stanford School of Medicine, Stanford, CA, [2]Department of Genetics, Stanford School of Medicine, Stanford, CA, [3]Program in Epithelial Biology and the Howard Hughes Medical Institute, Stanford School of Medicine, Stanford, CA, [4]Program in Cancer Biology, Stanford School of Medicine, Stanford, CA, [5]Division of Hematology, Cancer Institute and Institute for Stem Cell Biology and Regenerative Me, Stanford School of Medicine, Stanford, CA, [6]Computer Science Department, Stanford Engineering, Stanford, CA

The diversity of cell types and lineages in the mammalian blood system is a result of cellular differentiation programs involving waves of activation and repression of regulatory elements by specific transcription factor complexes binding complex genome sequence grammars. Here, we develop a powerful learning framework based on confidence-rated Boosting algorithms to decipher hematopoietic differentiation programs by integrating regulatory sequence, gene expression and chromatin accessibility data across a multitude of normal and leukemic blood cell types. Our approach results in a highly interpretable set of regulatory programs in the form of Alternating Decision Trees that combine transcription factor expression dynamics and combinatorial sequence features to predict diverse trajectories of chromatin accessibility dynamics of regulatory elements that govern the differentiation process. We use a powerful margin-based scoring statistic to dissect regulatory heterogeneity encoded in the model by exposing the influence of individual sequence motifs, transcriptional regulators and combinations across all stages of differentiation and for key modules of regulatory elements that control transitions and bifurcation events between cell-types in the hematopoietic lineage. We recover established regulators of the hematopoietic lineage as well as propose novel transcriptional programs that govern each stage of differentiation. We also identify regulatory programs responsible for early stages of leukemogenesis.

A NEW METHOD FOR BAYESIAN HYPOTHESIS TESTING OF DEMOGRAPHIC HISTORIES

Ilan Gronau

Herzliya Interdisciplinary Center (IDC), Computer Science, Herzliya, Israel

High throughput sequencing has greatly improved our ability to investigate the evolutionary history of species using detailed demographic models. A popular approach for inferring parameters in these demographic models is by sampling genealogical histories at many short unlinked loci using a Markov chain Monte Carlo (MCMC) algorithm, e.g., IMa (Hey and Nielsen, 2007), BP&P (Yang and Rannala, 2010), and G-PhoCS (Gronau et al, 2011). The use of explicit coalescent models by these methods makes them powerful for inferring demographic parameters, but they are limited in their ability to assess the fit of the inferred model to data. We propose a novel and flexible statistical measure for model fit that is based on Bayes factors estimated using the samples generated by the MCMC algorithm. This method takes advantage of the strength of existing demography inference methods in exploring the space of plausible genealogies, and can be implemented through very minor adjustments to the existing source code and no modifications to the sampling algorithm itself.

The new Bayesian method is based on comparing the fit of two models M_1 and M_2 to data X through the Bayes factor, $BF(M_1:M_2) = P(X|M_1) / P(X|M_2)$. This is done by introducing an additional "null" model M_0 and computing the Bayes factors for each of the other models relative to M_0: $BF(M_1:M_2) = BF(M_1:M_0)/BF(M_2:M_0)$. The advantage of this approach is that it could be applied separately to each model (without consideration of any of the other alternative models), and that careful choice of the null model M_0 reduces the variance of the estimator for $BF(M_i:M_0)$. This is a particularly important feature of this method, when compared to the current state-of-the-art, which is based on variants of the harmonic mean estimator for P(X|M) (Newton and Raftery, 1994).

We implemented this technique in the sampling algorithm of G-PhoCS and show using simulated data that it results in a considerably more accurate measure of model fit compared to the harmonic mean estimator. We evaluate the sensitivity of our estimator in scoring different species tree topologies in the presence of gene flow and in providing support for a given admixture event. We conclude by demonstrating an application of these relative Bayes factors to re-examine the origin of domestic dogs and the admixture history of dogs and wolves.

INFORMATION-BASED CLUSTERING AND ESTIMATION OF PROBABILITIES OF FITNESS CONSEQUENCES ACROSS THE HUMAN GENOME.

Brad Gulko[1], Ilan Gronau[2], Adam Siepel[3]

[1]Cornell University, Cornell University, Graduate Field of Computer Science, Ithaca, NY, [2]Herzliya Interdisciplinary Center (IDC), Efi Arazi School of Computer Science, Herzliya, Israel, [3]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

We describe a new computational method for estimating probabilities of fitness consequences for point mutation at genomic positions that are associated with collections functional genomic properties. By considering selective pressure (S), functional properties (E), and genomic positions (I) as variables in joint distribution we identify combinations of functional properties, or 'functional fingerprints', that are most informative about selective pressures, $P(S \mid E)$. Our method derives a small number of intelligible fingerprints from diverse genomic properties over 14 cell-types, including transcription, splicing, transcription factor binding, DNA methylation, DNase I hypersensitivity, chromatin geometry and DNA melting temperature. We then aggregate selective pressure across observed genomic positions exhibiting each fingerprint to assess the probability that a point mutation would be associated with fitness consequence, $P(S \mid I)$. This probability serves as a readily interpretable score that both powerfully predicts *cis*regulatory elements and provides an intelligible explanation of functional properties in terms of the associated fingerprint.

The current work generalizes our previous work (Gulko et al., Nat. Gen. 2015), by extending the variety of functional assays and human cell types that are considered, while limiting model complexity by reducing the number of chromatin states. Chromatin states are identified via recursive subdivision of all genomic positions, minimizing the entropy of selective pressure conditioned on a single functional covariate at each binary subdivision. The result is a readily intelligible decision tree based on at a single covariate at each decision node. Leaves of the tree represent chromatin states, as defined by the aggregate covariate partitioning of all ancestral nodes. This process generates a joint identification of chromatin state and selective pressure, based on human polymorphism and primate divergence, for each position in the human genome.

Our new work demonstrates improved accuracy in identifying regulatory elements, increased genomic coverage and greater cell-type sensitivity, while simultaneously reducing model complexity and improving intelligibility. The computational framework scales well to large data sets while providing a clear path for increasing the breadth of cell-state specific functional variation, as well as the incorporation of more nuanced measures of selective constraint.

← Back

THE LINGERING LOAD OF ARCHAIC ADMIXTURE IN MODERN HUMAN POPULATIONS

Kelley Harris[1,2], Rasmus Nielsen[3,4,5]

[1]Stanford University, Genetics, Palo Alto, CA, [2]University of California Berkeley, Mathematics, Berkeley, CA, [3]University of California Berkeley, Integrative Biology, Berkeley, CA, [4]University of California Berkeley, Statistics, Berkeley, CA, [5]University of Copenhagen, Bioinformatics Centre, Copenhagen, Denmark

Founder effects and bottlenecks can damage fitness by letting deleterious alleles drift to high frequencies. This almost certainly imposed a burden on Neanderthals and Denisovans, whose genetic diversity was less than a quarter of the level seen in humans today. A more controversial question is whether the out-of-Africa bottleneck created differences in genetic load between modern human populations. Some previous studies concluded that it saddled non-Africans with potentially damaging alleles that could affect global disease incidence (Lohmueller et al 2009; Fu et al 2014), while other studies have inferred little difference in genetic load between Africans and non-Africans (Simons, et al 2014; Do et al 2015). However, no such studies, to our knowledge, have incorporated the potential fitness effects of introgression from Neanderthals into non-Africans. We present simulations showing that archaic introgression may have had greater fitness effects than the out-of-Africa bottleneck itself, saddling non-Africans with deleterious alleles that accumulated as nearly neutral variants in Neanderthals. Assuming that the exome experiences mutations with additive fitness effects drawn from a previously inferred gamma distribution (Eyre-Walker et al 2006), we predict the existence of strong selection against early Neanderthal-human hybrids. This is a direct consequence of mutation accumulation during a period of low Neanderthal population size that is thought to have lasted ten times longer than the out-of-Africa bottleneck (Pruefer et al 2014). The introgression scenario is well described by a series of two phases that are easy to summarize mathematically, the first phase characterized by selection against individuals who have more Neanderthal DNA than the population average. This quickly pushes the population into a second phase where all individuals have nearly the same Neanderthal admixture fraction and excess deleterious alleles are eliminated slowly and inefficiently. As a result of this inefficiency, the model predicts some transmission of deleterious Neanderthal variation to present-day non-Africans, but also predicts the depletion of Neanderthal DNA from conserved genomic regions as observed empirically by Sankararaman, et al (2014). Our results imply that this deficit of Neanderthal DNA from functional genomic regions can be explained without the action of epistatic reproductive incompatibilities between human and Neanderthal alleles.

← Back

COALESCENT TIMES AND PATTERNS OF GENETIC DIVERSITY IN SPECIES WITH FACULTATIVE SEX: EFFECTS OF GENE CONVERSION, POPULATION STRUCTURE AND HETEROGENEITY

Matthew Hartfield[1,2], Stephen I Wright[1], Aneil F Agrawal[1]

[1]University of Toronto, Department of Ecology and Evolutionary Biology, Toronto, Canada, [2]University of Aarhus, Bioinformatics Research Centre, Aarhus, Denmark

Many diploid organisms undergo facultative sexual reproduction. However, little is currently known concerning the distribution of neutral genetic variation amongst facultative sexuals except in very simple cases. Understanding this distribution is important when making inferences about rates of sexual reproduction, effective population size and demographic history. Here, we extend coalescent theory in diploids with facultative sex to consider gene conversion, selfing, population subdivision, and temporal and spatial heterogeneity in rates of sex. In addition to analytical results for two-sample coalescent times, we outline a coalescent algorithm that accommodates the complexities arising from partial sex; this algorithm can be used to generate multi-sample coalescent distributions. A key result is that when sex is rare, gene conversion becomes a significant force in reducing diversity within individuals, which can remove genomic signatures of infrequent sex (allelic sequence divergence, also known as the 'Meselson Effect') or entirely reverse the predictions. Our models offer improved methods for assessing the null, neutral model of patterns of molecular variation in facultative sexuals.

FAST AND ACCURATE APPROXIMATE INFERENCE OF TRANSCRIPT EXPRESSION FROM RNA-SEQ DATA USING BITSEQVB

Antti Honkela[1], James Hensman[2], Panagiotis Papastamoulis[3], Peter Glaus[4], Magnus Rattray[3]

[1]University of Helsinki, Helsinki Institute for Information Technology HIIT, Department of Computer Science, Helsinki, Finland, [2]University of Sheffield, Sheffield Institute for Translational Neuroscience (SITraN), Sheffield, United Kingdom, [3]University of Manchester, Faculty of Life Sciences, Manchester, United Kingdom, [4]University of Manchester, School of Computer Science, Manchester, United Kingdom

Assigning RNA-seq reads to their transcript of origin is a fundamental task in transcript expression estimation. Where ambiguities in assignments exist due to transcripts sharing sequence, e.g. alternative isoforms or alleles, the problem can be solved through probabilistic inference. Bayesian methods have been shown to provide accurate transcript abundance estimates compared to competing methods. However, exact Bayesian inference is intractable and approximate methods such as Markov chain Monte Carlo (MCMC) and Variational Bayes (VB) are typically used. While providing a high degree of accuracy and modelling flexibility, standard implementations can be prohibitively slow for large datasets and complex transcriptome annotations.

We propose a novel approximate inference scheme based on VB and apply it to the BitSeq model of transcript expression inference from RNA-seq data (Glaus et al., Bioinformatics 2012). Recent advances in VB algorithmics are used to improve the convergence of the algorithm beyond the standard Variational Bayes Expectation Maximisation (VBEM) algorithm. We apply our new BitSeqVB algorithm to simulated and biological datasets, demonstrating a significant increase in speed with only very small loss in accuracy of expression level estimation compared to original MCMC-based BitSeq. We carry out a comparative study against seven popular alternative methods and demonstrate that our new algorithm provides excellent accuracy and inter-replicate consistency while remaining competitive in computation time.

In addition to the new algorithmic developments and benchmarks, I will present new applications of BitSeq methods including pre-mRNA expression estimation from ribo-depleted RNA-seq and more informative differential expression analysis.

Pre-print with more information on BitSeqVB:
J. Hensman, P. Papastamoulis, P. Glaus, A. Honkela, and M. Rattray.
Fast and accurate approximate inference of transcript expression from RNA-seq data.
arXiv:1412.5995 [q-bio.QM]

EFFICIENT GENERALIZED LINEAR MIXED MODELS FOR THE GENETIC ANALYSIS OF COUNT-DERIVED AND BINARY PHENOTYPES

Danilo Horta, Oliver Stegle

[1]EMBL-EBI, Stegle Group, Cambridge, United Kingdom

The development of computationally efficient yet accurate models has received considerable attention in statistical genetics. In particular linear mixed models (LMMs) are now a well established tool and provide powerful control for population structure and relatedness, allow to aggregate across multiple causal variants in gene sets and can be used to leverage phenotype correlations between multiple (related) traits.

However, the vast majority of existing LMM approaches assume that phenotypes are continuous with Gaussian distributed residuals. This assumption is clearly violated in case/control studies but also in the context of a many sequencing-based phenotypes, such as Poisson distributed read count data or traits defined as the Binomial ratio of (typically small) count values. While generalized linear mixed models provide in principle an established solution to this problem, "exact" methods for parameter inference require expensive MCMC simulations and hence are not applicable to large cohorts. Consequently, non-Gaussian observation likelihood are in practice either ignored or one is left with methods that provide crude approximations to estimate the trait on a latent liability scale.

To address this, we here propose a highly effective deterministic algorithm QEP-LMM that enables near-exact marginalizing over the latent liability scale within the LMM framework. This model provides quadratic and in some instances even linear run-time complexity in the number of samples, thus enabling the analysis of datasets with tens of thousands of individuals in the context of genome-wide tests. We extensively compared our model with existing state-of-the-art tools (Gaussian LMM, GCTA, LTMLM and LEAP), both in terms of power to detect associations as well as accuracy for heritability estimation and phenotype prediction. Consistently across settings, we find substantial improvements over current approximate methods. Remarkably, we observe that QEP-LMM achieves near-identical performance to exact MCMC approaches for generalized LMMs at a runtime complexity that is comparable to a standard LMM. Finally, we provide empirical results to demonstrate practical utility of QEP-LMM in applications to data from the WTCCC and in the genetic analysis of splicing phenotypes in human LCLs.

INTEGRATION OF *DE NOVO* MUTATIONS WITH GENE EXPRESSION REGULATION NETWORKS IMPROVES RISK GENE DISCOVERY FOR DEVELOPMENTAL DISORDERS

Qiang Huang[1], Yufeng Shen[1,2,3]

[1]Department of Systems Biology, Columbia University Medical Center, Columbia University, New York, NY, [2]Department of Biomedical Informatics, Columbia University Medical Center, Columbia University, New York, NY, [3]JP Sulzberger Columbia Genome Center, Columbia University Medical Center, Columbia University, New York, NY

Recent large-scale exome sequencing studies of developmental disorders established that *de novo* mutations are major contributors to disease genetics. However, due to the scarcity of *de novo* coding mutation in each person and the large number of potential risk genes for a certain disease, comprehensive detection of risk genes based on *de novo* mutation only requires very large number of samples. Integrating statistical evidence form *de novo* mutations with functional genomics data and prior biological knowledge is a promising direction to improve our ability to detect novel candidate risk genes and interpret disease etiology. Several integrative methods have been developed based on different biological networks, such as protein-protein interaction (PPI) and gene co-expression networks. The basic idea of network methods is that one gene is more likely to be risk gene when it is connected with other risk genes in a network. However, it is not clear how the choice of networks combined with different integration models, produce the optimal results. In this study, we compiled *de novo* mutation data of more than 5500 cases from recently published studies on developmental disorders, compared different networks and algorithms to assess the performance. We leverage the large amount of samples to construct semi-gold standard results and simulate various scenarios for benchmarking performance. We show that well-constructed co-expression networks are generally better than PPI networks for discovering novel risk genes, primarily due to tissue specificity of interactions. To overcome the limitations of existing models and maximum the benefit of network information, we developed a new integration method based on a conditional random field model, MICRF (**M**utation **I**ntegration with **C**onditional **R**andom **F**ield). We weight edges by betweenness centrality to model functional bottlenecks. In our simulations, MICRF achieved better performance in various scenarios. In conclusion, we accessed the effect of different networks and the limitations of existing methods and built a general network model, which showed improvement in risk gene discovery for development disorders.

BAYESIAN INFERENCE OF GENE EXPRESSION DYNAMICS FROM TIME-COURSE NASCENT RNA SEQUENCING DATA

Yi-Fei Huang[1], André L Martins[1], Noah Dukler[2,1], Adam Siepel[1]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Weill Cornell Medical College, Tri-Institute Computational Biology and Medicine program, New York, NY

Nascent RNA sequencing technologies, such as GRO-seq, PRO-seq, and NET-seq, enable the study of elongation patterns of nascent RNAs on a genome-wide scale. By precisely mapping the genomic positions and orientations of RNA polymerase II in a time-course experiment, nascent RNA sequencing can reveal the dynamics of how nascent RNAs are generated and regulated in vivo. Because of the intrinsic stochasticity of both the transcription process itself and next-generation sequencing readouts, statistical models, especially hidden Markov models (HMMs), play a central role in the analyses of nascent RNA sequencing data. However, none of the existing HMMs explicitly model gene expression dynamics over time, which limits their power to provide a mechanistic description of gene regulation. In addition, most existing methods are applied separately to individual genes, which prohibits pooling of data across genes with similar patterns of regulation. We propose a novel Bayesian generative model, TimeHMM, to overcome the drawbacks of the existing models. Unlike existing HMMs, TimeHMM uses an inhomogeneous HMM defined over time to describe the patterns of gene expression at individual genes during a time-course experiment. To allow genes with similar patterns of regulation to share statistical strength, a sparse prior is defined over the inhomogeneous HMM which forces the observed data to be explained by a limited number of global up-regulation or down-regulation events. Markov chain Monte Carlo methods are used to perform a full Bayesian inference of the parameters and latent variables. From the sampled posterior distribution, it is possible to characterize the set of up- and down-regulation events, which genes are involved, at what times the events occurred, and to identify induced clusters of co-regulated genes. Overall, TimeHMM provides a novel framework to jointly infer global regulatory events and expression dynamics of individual genes based on GRO-seq and similar technologies. Simulations and time-course GRO-seq datasets are used to evaluate the power and robustness of TimeHMM.

A SYSTEMATIC EXPLORATION OF DEEP LEARNING METHODS FOR PREDICTING TRANSCRIPTION FACTOR BINDING FROM DNA SEQUENCE

Johnny Israeli[1], Irene M Kaplow[2], Avanti Shrikumar[2], Rahul Mohan[3], Anshul B Kundaje[2,4]

[1]Stanford University, Biophysics, Stanford, CA, [2]Stanford Univeristy, Computer Science, Stanford, CA, [3]Bellarmine College Preparatory School,, San Jose, CA, [4]Stanford University, Genetics, Stanford, CA

Interacting complexes of transcription Factors (TFs) bind combinatorial sequence grammars encoded in genomic regulatory elements. A number of machine learning methods have been developed to learn predictive sequence preferences of TFs from in-vitro and in-vivo TF binding assays. Support Vector Machines (SVMs) with string kernels for DNA sequence often provided state-of-the-art results for these problems. Recently, deep learning approaches such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have resulted in dramatic performance improvements for several related learning tasks in Natural Language Processing, Speech Processing and Computer Vision. Here, we develop a novel deep learning framework to learn from raw DNA sequence. Using extensive simulations of regulatory sequence, we evaluate the ability of deep CNNs and RNNs trained on raw sequence to capture different properties of regulatory sequences typical of transcription factor binding sites including probabilistic affinity to sequence motifs, positional and density distributions of motifs, combinatorial sequence grammars involving co-factor sequence preferences with spacing and order constraints. Our simulations are based on sequence properties derived from in-vitro HT-SELEX experiments and from in-vivo ChIP-seq binding sites of a diverse variety of TFs. We compare performance to various SVM based string kernel methods. For simpler simulations of binding preferences of individual TFs, deep CNNs and SVMs perform comparably. However, deep CNNs consistently outperform SVMs to learn from long regulatory sequences (500 bp to 1 Kb) with sparse signal, sequences containing positional biases of motifs or motifs of multiple TFs. For complex simulated motif grammars, deep RNNs achieve the highest accuracy, surpassing CNNs and significantly outperforming SVMs. Having learned general modeling principles from these simulations, we used deep neural networks to predict sequence preferences of TFs from HT-SELEX data and learn predictive models of in-vivo TF binding from ChIP-seq data. Our results demonstrate the superior performance of deep learning methods compared to SVMs for modeling regulatory sequence and TF binding preferences. Our study also provides novel insights into developing and exploring optimal deep learning architectures for DNA sequence.

← Back

COREPHASE: REDUCING MAXIMUM LIKELIHOOD PHASING PROBLEMS OF 2^{100} VARIABLES OR MORE TO EM-PRACTICAL SIZES WITHOUT LOSS OF OPTIMALITY VIA GRAPH-THEORETIC SYMMETRIES OF THE LIKELIHOOD FUNCTION

Douglas McErlean, Sorin Istrail

Brown University, Department of Computer Science, Providence, RI

Haplotype phasing is a problem of great practical importance, but also of considerable theoretical difficulty. Without specific haplotype data, genetic association studies are impossible, yet it has been proven that mathematically rigorous phasing, whether optimizing for the most parsimonious set of haplotypes or even just finding the likelihood of the best solution, is NP-hard. This forces phasing algorithms to compromise rigor in the name of practicality, as the exponential number of candidate haplotypes becomes intractable for anything beyond toy examples. Our work explores a new approach to this problem, by demonstrating that in practice we can exploit symmetries on the graph structure of the likelihood function to efficiently eliminate all but a small *core* of haplotypes while still guaranteeing the presence of fully optimal solutions. The algorithm to find this set of sufficient haplotypes is efficient in the size of its output, and we present theoretical evidence that the results should indeed be small with high probability. Finally, we test our implementation of this algorithm, CorePhase, on data from the 1000 Genomes project, demonstrating empirically that the results are in fact compact for real genome data, and can be extracted quickly. Notably, CorePhase was able to take practical phasing problems with 2^{100} possible haplotypes and reduce them to just a few thousand in a matter of minutes, putting the reduced problem well within the range where rigorous algorithms like Expectation Maximization can be efficiently applied.

DISCOVERY OF GENETIC HETEROGENEITY IN A CONTEXT OF PHYSIOLOGICAL HOMOGENEITY BY BIOLOGICAL DISTANCE CLUSTERING

Yuval Itan[1], Lei Shang[1], Lluis Quintana-Murci[2], Shen-Ying Zhang[1,2], Laurent Abel[3,1], Jean-Laurent Casanova[1,3,4]

[1]The Rockefeller University, Human Genetics of Infectious Diseases, New York, NY, [2]Institut Pasteur, Human Evolutionary Genetics, Paris, France, [3]INSERM, Human Genetics of Infectious Diseases, Paris, France, [4]Howard Hughes Medical Institute,, New York, NY

To determine the disease-causing allele(s) underlying primary human inborn errors, high-throughput genomic methods are applied and provide thousands of gene variants per patient. We recently reported a novel approach, the "human gene connectome" (HGC) – the set of all in silico-predicted biologically plausible routes and distances between all pairs of human genes, effective for prioritizing gene variants by biological distance from known disease-causing genes. However, there is currently no available method for automating the selection of candidate disease-causing mutant alleles in the absence of a known morbid gene in at least one patient with the disease of interest, posing a major bottleneck in the field in high-throughput clinical genomics. We hypothesized that within a cohort of patients with the same Mendelian disease, the cluster that contains the key disease-causing gene for each patient is the HGC-predicted biologically smallest cluster. We then developed and applied a Mendelian clustering algorithm, which estimates the biologically smallest HGC-predicted cluster that contains one allele per patient. By that we (i) approximated a solution for an NP-complete algorithmic problem (i.e. not possible to solve on a large scale by a computer), and (ii) estimated and statistically validated a set of disease-causing alleles in whole exome sequencing cohorts of Mendelian disease patients. The unbiased approach described above should facilitate the discovery of morbid alleles in patients with primary inborn errors that lack a genetic etiology.

THE HUMAN GENE DAMAGE INDEX: A NOVEL GENE-LEVEL APPROACH TO PRIORITIZE EXOME VARIATIONS

Yuval Itan[1], Lei Shang[1], Lluis Quintana-Murci[2], Shen-Ying Zhang[1,2], Laurent Abel[3,1], Jean-Laurent Casanova[1,3,4]

[1]The Rockefeller University, Human Genetics of Infectious Diseases, New York, NY, [2]Institut Pasteur, Human Evolutionary Genetics, Paris, France, [3]INSERM, Human Genetics of Infectious Diseases, Paris, France, [4]Howard Hughes Medical Institute,, New York, NY

The exome of a patient with a monogenic disease contains about 20,000 variations, of which only one or two are disease-causing. Variant- and gene-level in silico methods have been developed to select candidate variations prior to their experimental study. We aimed to develop a novel gene-level approach to predict whether specific variations in any given protein-coding gene may be disease-causing. We noticed that 58.32% of exome variants in the general population are found in only 2.42% of human genes. We thus developed the gene damage index (GDI): a novel genome-wide, gene-level estimate of accumulated mutational damage in the general population. We found correlations between GDI and selective evolutionary pressure, protein complexity, coding sequence length, and number of paralogs. We also showed that GDI better differentiates between truly disease-causing and falsely positive genes than three leading gene-level methods (genic intolerance, gene indispensability, and de novo excess). We further defined a high-GDI cutoff that can successfully filter out up to 60.62% of spurious variants from patients' exome highly mutated genes. Conversely, a low-GDI cutoff points to prime candidate mutations, in genes never or rarely mutated. This novel method should facilitate human genetic studies, especially for monogenic disorders.

THE MUTATION SIGNIFICANCE CUTOFF (MSC): A GENE-SPECIFIC APPROACH TO PREDICTING THE IMPACT OF HUMAN GENE VARIANTS

Yuval Itan[1], Lei Shang[1], Lluis Quintana-Murci[2], Shen-Ying Zhang[1,2], Laurent Abel[3,1], Jean-Laurent Casanova[1,3,4]

[1]The Rockefeller University, Human Genetics of Infectious Diseases, New York, NY, [2]Institut Pasteur, Human Evolutionary Genetics, Paris, France, [3]INSERM, Human Genetics of Infectious Diseases, Paris, France, [4]Howard Hughes Medical Institute,, New York, NY

The proposed thresholds of significance for current predictors of the biological impact of human genetic variants, such as CADD score, are identical for all genes across the genome. However, we found large differences between the predicted scores of the known pathogenic mutations of 4,015 disease-causing genes. As a result, the prediction rate for proven disease-causing alleles was found to be less than 40%. We thus estimated the 95% confidence intervals of the CADD scores for all known mutations in 17,765 human protein-coding genes. By inference from the subgroup of known disease-causing genes, we defined a gene-specific mutation significance cutoff (MSC) for each of these genes. We then validated the prediction power of this approach through both simulation and analyses of real data. In particular, the true positive prediction rate for a new set of proven disease-causing alleles increased from 34.10% for the regular CADD score to 95.15% with MSC. The MSC can be used to select candidate mutations in the exomes of patients with inborn errors of known or unknown disease-causing genes. The MSC greatly improves the predictive power of methods without gene-specific thresholds.

RECONSTRUCTING THE TEMPORAL PROGRESSION OF HIV-1 IMMUNE RESPONSE PATHWAYS

Siddhartha Jain, Joel Arrais, Narasimhan J. Venkatachari, Velpandi Ayyavoo, Ziv Bar-Joseph

Carnegie Mellon University, Pittsburgh, PA

Most methods for reconstructing signaling and regulatory response networks from high throughput data generate static models which cannot distinguish between early and late stages in the response. Some of the methods that can be used for dynamic analysis only utilize gene expression data and so cannot model the effects of host-pathogen and protein-protein interactions which play a major role in various immune and defense responses. Here we present TimePath, a new method that integrates time series and static datasets to reconstruct dynamic models of host immune response. TimePath integrates gene and miRNA expression data with interaction data using an Integer Programming (IP) based optimization function. The IP is used to jointly reconstruct the regulatory and signaling networks while constraining the pathways selected so that they reflect the temporal changes observed in the data . TimePath starts by initially selecting a large set of pathways that are rooted in source proteins (proteins that directly interact with the infecting agent's proteins) and end in gene that are differentially expressed (DE) at various time points in the response. This allows TimePath to model the impact of proteins that are only post-transcriptionally and / or post-translationally activated. Pathways for later DE genes are required to contain DE genes or miRNAs from earlier phases to explain their delayed response. Next, we use the IP to select a small subset of pathways that, together, explain the full set of DE genes. The resulting dynamic network generated by TimePath allows the assignment of pathways to specific temporal phases, the identification of key proteins and miRNAs controlling these phases and leads to testable hypotheses regarding the impact of various temporal perturbations.

We applied TimePath to model human response to HIV-1. For this we utilized time series gene and miRNA expression data in HIV-1 infected cells, data regarding interactions between HIV-1 proteins and human proteins and general human protein-protein and protein-DNA interaction data. The network constructed by TimePath for HIV-1 infection response included a significant number of proteins, miRNAs and pathways that were previously determined to play a role in HIV-1 infection response. Comparing the resulting network to prior methods for reconstructing response networks using knockdown data (KD) indicates that TimePath greatly improves upon these prior methods allowing us to both identify and assign a functional role to proteins whose KD impacts viral loads. In addition to identifying known proteins in the response, TimePath also predicted the involvement of novel proteins and was able to assign a temporal role to these proteins as part of the response process. We experimentally validated several of the time specific predictions. As we show, while blocking some proteins, at any phase, leads to reduction in viral load, blocking other proteins has a positive impact only if they are targeted at the specific phase predicted by TimePath validating the importance of temporal models for preventing and treating infections.

← Back

EXACT LIKELIHOODS FOR INFERENCE OF SELECTION FROM TIME SERIES GENETIC DATA

Ethan M Jewett[1], Yun S Song[2]

[1]UC, Berkeley, Statistics, Berkeley, CA, [2]UC, Berkeley, EECS, Berkeley, CA

Recently, many methods have been developed for inferring selection coefficients from time series data while accounting for finite population sizes and other biologically realistic phenomena. However, methods that model complex evolutionary and demographic processes often rely on approximations of the Wright-Fisher process that break down when allele frequencies are close to the boundary points 0 and 1, or which make assumptions about the magnitudes of selection coefficients, leading to biased estimates. Here, we devise a fast recursive approach to compute the exact likelihood of observing a sampled allele frequency trajectory under the Wright-Fisher model. In addition to improving inferences of selection coefficients, the method allows us to visualize the exact likelihood surface, providing insight into the accuracy with which selection coefficients can be inferred under different evolutionary scenarios. The method also allows us to benchmark other inference approaches against the the exact maximum likelihood. We find that very fast methods that ignore genetic drift can be almost as accurate as the exact likelihood, even when population sizes are small. Such methods are useful for performing genomic scans and for inferring selection parameters quickly when population sizes are large. Our analyses provide insights into the problem of inferring selection coefficients and the method we develop allows selection coefficients of arbitrary strengths to be inferred accurately.

GRAF – A TOOL SET TO QUICKLY FIND RELATED SUBJECTS FROM LARGE SETS OF GENOTYPE DATA OBTAINED WITH DIFFERENT METHODS

Yumi Jin, Michael Feolo, Stephen Sherry

National Institutes of Health, National Center for Biotechnology Information, Bethesda, MD

Over one million individuals in dbGaP database contain genotype data, but as a collection it is heterogeneous by having been submitted as separate studies obtained with different genotyping methods. Because of the size and constant growth of the collection, dbGaP curators face a substantial challenge in processing submissions – repeatedly identifying potential duplicate samples or cryptic relationships within and across hundreds of studies with every new submission. The Genetic Relationship and Fingerprinting (GRAF) software package was developed to meet this challenge.

GRAF first extracts individual genotypes for a defined set of 10,000 biallelic SNPs having a minor allele frequency ≥0.17 as reported by the 1000 Genome Project and covered by ≥80% of submitted genotyping methods. These variants are autosomal, ≥50,000 base pairs apart, and were selected to avoid complementary alleles (e.g. A/T, C/G) that cannot be correctly distinguished if reported in reverse orientation. Potential subject relationships are then inferred from the extracted genotypes using a pair-wise comparison.

GRAF uses two statistics, All Genotype Mismatch Rate (AGMR) and Homozygous Genotype Mismatch Rate (HGMR) to infer these relationships. AGMR is the genotype mismatch rate when all non-missing genotypes are compared, and HGMR is the mismatch rate when only homozygous genotypes are considered. For each pair of genotyped subjects, GRAF calculates the AGMR and HGMR values, and then infers a genetic relationship from a bivariate normal distribution model. Additionally, GRAF can compare submitted subject/sample relationships to those predicted by the program, and output both tabular results and graphical reports where outlier cryptic relationships can be easily spotted by visual inspection.

By using C++ bitwise operations and other performance optimizations, the relationship determination program (GRAF) can quickly find closely related subjects in large datasets. For example, it takes GRAF only 2.3 minutes to find all sample duplicates and closely related (up to 2nd degree) relatives in a dbGaP dataset of 12,000 samples. At scale, GRAF finds more than 200,000 pairs of duplicate samples or monozygotic twins in 19 hours for a pair-wise analysis of the entire dbGaP database: about one million samples genotyped for at least 10% of the fingerprint set.

GRAF is currently being used by dbGaP curators and submitters as a pre-submission and pre-analysis quality control tool to find and correct errors in genotype datasets, subject-sample mapping files and pedigree files. This presentation will describe GRAF and its functionality.

DETECTION OF TRANS-EQTLS BENEFIT FROM JOINT ANALYSIS IN MULTIPLE TISSUES WITH STATISTICAL FRAMEWORK THAT INCORPORATES TISSUE-SPECIFICITY AND MENDELIAN RANDOMIZATION.

Brian Jo[1], Ian McDowell[2], Andrew Taverner[1], Barbara Engelhardt[3]

[1]Princeton University, Quantitative and Computational Biology, Princeton, NJ, [2]Duke University, Computational Biology and Bioinformatics, Durham, NC, [3]Princeton University, Computer Science, Princeton, NJ

In expression quantitative trait loci (eQTL) studies, various statistical tests are used to identify putative genomic loci that regulate gene expression. Association tests in these studies benefit from the combination of larger sample sizes and multiple tissue samples per individual. For a number of reasons, cis-, or allele-specific, eQTL discovery has garnered the bulk of these benefits, while trans-, or distal, eQTL analyses, have not been able to take advantage of multi-tissue data for improved statistical power. In this work, we describe our modeling approaches that improve trans-eQTL analyses for large-scale eQTL studies with rich expression data. First, we explore the tissue-specificity of trans-eQTLs and use available and new methods to leverage multi-tissue expression data and analyses for gains in statistical power. Second, we adapt methods of Mendelian Randomization that exploit available cis-eQTLs for the purpose of recovering trans-effects in a formal statistical framework. Then, we apply our approaches to the Genotype-Tissue Expression (GTEx) data and explore the rich signal of trans-effects we recover from this data set in order to show that our framework is more effective at recovering trans-eQTLs than state-of-the-art methods.

CONFOUNDING FACTOR ESTIMATION THROUGH INDEPENDENT COMPONENT ANALYSIS (CONFETI)

Jin Hyun Ju[1,2,3], Sushila A Shenoy[2], Jason Mezey[1,2,4]

[1]Weill Cornell Medical College, Institute for Computational Biomedicine, New York, NY, [2]Weill Cornell Medical College, Department of Genetic Medicine, New York, NY, [3]Weill Cornell Graduate School of Medical Sciences, Physiology, Biophysics, and Systems Biology Graduate Program, New York, NY, [4]Cornell University, Department of Biological Statistics and Computational Biology, Ithaca, NY

Expression Quantitative Trait Loci (eQTL) studies have provided valuable insights on genetic mechanisms by investigating associations between genetic variations and gene expression levels. However, non-genetic factors influencing gene expression measurements, such as sample specific environmental effects or laboratory and procedure specific technical effects, often remain unaccounted for and could potentially obscure the interpretation of the data. Generally, such unobserved variables are referred to as hidden factors or confounding factors. The importance of correcting for confounding factors is increasing with the recent rise of consortium scale studies in which samples from various populations are processed across multiple laboratories and thus have more potential sources of variation.

Here we introduce CONfounding Factor Estimation Through Independent component analysis (CONFETI), a method based on Independent Component Analysis (ICA) to identify and correct the effects of confounding factors. ICA is a method that decomposes the data into non-Gaussian and statistically independent components, producing a representation of the generative structure of the observed data. CONFETI first projects the samples onto a lower dimensional space constructed by independent components, and estimates a sample similarity matrix using projections that are likely non-genetic. This information is subsequently used in a linear mixed model to test for associations between genotypes and phenotypes.

To evaluate the performance of CONFETI, we generated multiple sets of synthetic eQTL data in which the ground truth is known to us. CONFETI successfully corrected the effects of simulated confounding factors and identified true hits more accurately compared to similar methods, such as SVA, PEER, and PANAMA. In human datasets, including data from the Genetic European Variation in Health and Disease (GEUVADIS) consortium, the Multiple Tissue Human Expression Resource (MuTHER) consortium, and the HapMap project, we found that CONFETI identified the most *cis*-hits in almost all cases. Additionally, more eQTL hits replicated between datasets with the same tissue types using CONFETI.

MOMI: A NEW METHOD FOR INFERRING DEMOGRAPHY AND COMPUTING THE MULTIPOPULATION SAMPLE FREQUENCY SPECTRUM

John A Kamm[1], Jonathan Terhorst[1], Yun S Song[1,2,3]

[1]UC Berkeley, Statistics, Berkeley, CA, [2]UC Berkeley, EECS, Berkeley, CA, [3]UC Berkeley, Integrative Biology, Berkeley, CA

The sample frequency spectrum (SFS) describes the distribution of allele counts at segregating sites, and is a useful statistic for both summarizing genetic data and inferring biological parameters. SFS-based inference proceeds by comparing observed and expected values of the SFS, but computing the expectations is computationally challenging when there are multiple populations related by a complex demographic history.

We are developing a new software package, momi (MOran Models for Inference), that computes the multipopulation SFS under population size changes (including exponential growth), population mergers and splits, and pulse admixture events. Underlying momi is a multipopulation Moran model, which is equivalent to the coalescent and the Wright-Fisher diffusion, but has computational advantages in both speed and numerical stability. Techniques from graphical models are used to integrate out historical allele frequencies. Automatic differentiation provides the gradient and Hessian, which are useful for searching through parameter space and for computing asymptotic confidence intervals.

Using momi, we are able to compute the exact SFS for more complex demographies than previously possible. The runtime of momi depends on the pattern of migration events, but for certain demographic histories, momi can scale up to tens to hundreds of populations. To demonstrate the utility of momi, we apply it to infer a model of human history involving archaic hominins (Neanderthal and Denisovan) and modern humans in Africa, Europe, East Asia, and Melanesia.

A DEEP NEURAL NETWORK APPROACH FOR PREDICTING THE EFFECTS OF GENETIC VARIANTS ON TRANSCRIPTION FACTOR BINDING

Irene M Kaplow[1], Ashley K Tehranchi[2], Johnny Israeli[3], Avanti Shrikumar[1], Rahul Mohan[4], Hunter B Fraser[2], Anshul Kundaje[1,4]

[1]Stanford University, Computer Science, Stanford, CA, [2]Stanford University, Biology, Stanford, CA, [3]Stanford University, Biophysics, Stanford, CA, [4]Stanford University, Genetics, Stanford, CA

The majority of disease-associated variants identified in genome-wide association studies (GWAS) are non-coding, and there is an enrichment for GWAS hits in regulatory elements. These non-coding variants often disrupt regulatory sequence elements thereby affecting local transcription factor (TF) binding. In order to identify variants affecting transcription factor binding, Tehranchi *et al.* (*in Review*) recently generated ChIP-seq data for five TFs – Jund, NFKB, Oct1, Pu.1, and Stat1, from a pool of LCLs from 60 Yoruban individuals and identified TF binding quantitative trait loci (bQTLs). Interestingly, a substantial proportion of the bQTLs were found to lie outside but in close proximity to canonical binding sites of these TFs, a property also shared by expression-QTLs, histone-modification-QTLs, DNase hypersensitivity QTLs, and fine-mapped disease-associated non-coding variants. Hence, the mechanism by which bQTLs affect *in vivo* TF binding by disrupting regulatory sequence remains poorly-characterized. In order to identify patterns in these mechanisms, we train a deep neural network to discriminate between SNPs in TF binding sites that act as bQTLs from those that do not by integrating local diploid sequence context, chromatin accessibility, nucleosome positioning and histone modifications as well as positional, density, and allelic properties of the variants. We use a multi-tasking formulation that allows joint training across bQTLs for multiple TFs, thereby modeling direct and indirect effects of variants on binding of TFs and their co-associating factors. We compare our deep learning approach to using support vector machines (SVMs) with string kernels. Our models provide new hypotheses about potential mechanisms through which variants affect regulatory molecular phenotypes.

← Back

QUANTIFYING AND MITIGATING THE EFFECT OF PREFERENTIAL SAMPLING ON PHYLODYNAMIC INFERENCE.

Michael D Karcher[1], Julia A Palacios[2,3,4], Trevor Bedford[5], Marc A Suchard[6,7,8], Vladimir N Minin[1,9]

[1]University of Washington, Department of Statistics, Seattle, WA, [2]Harvard University, Department of Organismic and Evolutionary Biology, Cambridge, MA, [3]Brown University, Department of Ecology and Evolutionary Biology, Providence, RI, [4]Brown University, Center for Computational Molecular Biology, Providence, RI, [5]Fred Hutchinson Cancer Research Center, Vaccine and Infectious Disease Division, Seattle, WA, [6]UCLA, Department of Human Genetics, Los Angeles, CA, [7]UCLA, Department of Biomathematics, Los Angeles, CA, [8]UCLA, Department of Biostatistics, Los Angeles, CA, [9]University of Washington, Department of Biology, Seattle, WA

Phylodynamics seeks to estimate effective population size fluctuations from molecular sequences of individuals sampled from a population of interest. One way to accomplish this task formulates an observed sequence data likelihood exploiting a coalescent model for the sampled individuals' genealogy and then integrating over all possible genealogies via Monte Carlo or, less efficiently, by conditioning on one genealogy estimated from the sequence data. However, when analyzing sequences sampled serially through time, current methods implicitly assume either that sampling times are fixed deterministically by the data collection protocol or that their distribution does not depend on the size of the population. Through simulation, we first show that, when sampling times do probabilistically depend on effective population size, estimation methods may be systematically biased. To correct for this deficiency, we propose a new model that explicitly accounts for preferential sampling by modeling the sampling times as an inhomogeneous Poisson process dependent on effective population size. We demonstrate that in the presence of preferential sampling our new model not only reduces bias, but also improves estimation precision. Finally, we compare the performance of the currently used phylodynamic methods with our proposed model through clinically-relevant, seasonal human influenza examples.

EFFICIENT GENOME SCALE COALESCENT SIMULATION FOR HUGE SAMPLE SIZES USING SPARSE TREES AND COALESCENCE RECORDS

Jerome Kelleher, Gilean McVean

University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom

A central challenge facing the analysis of genetic variation is to provide realistic genome simulation across millions of samples. Present day simulations do not scale well, or use an approximation that does not capture important linkage properties. We solve these problems by introducing sparse trees and coalescence records as the key units of genealogical analysis. Using these tools, we show that exact simulation of the coalescent with recombination for chromosome-sized regions over hundreds of thousands of samples is possible.

Our key innovation is to model genealogies as sparse trees, in which each node is an integer. In the simulation algorithm, each coalescence event is assigned a unique integer which corresponds to a tree node. The effect of this event is then encoded using a set of tuples (l, r, c_1, c_2, p, t), which state that, at time t, over the (half-open) genomic interval $[l, r)$, the parent of nodes c_1 and c_2 is p. This description of the state of the simulation drastically reduces memory requirements over existing methods, and also leads to a very concise description of the sequence of correlated trees output by the simulation.

We present an implementation of these ideas and show that it is substantially faster than simulators based on the sequentially Markov coalescent approximation for large sample sizes. For example, simulating the genealogies for 10^5 samples over a 100 megabase region with a recombination rate of 0.001 per site per $4N$ generations requires around 10 minutes and less than 1GB of RAM on a modern desktop computer. A similar simulation using the popular MaCS simulator required over 4 weeks on equivalent hardware (Layer et al. 2015, *Efficient genotype compression and analysis of large genetic variation datasets*). Storing the correlated trees output by coalescent simulations also presents a substantial challenge. In the example above, the genealogies consumed over 3.5TB of storage when encoded in Newick format. Using a file format based on coalescence records, the same information was encoded in around 90MB (giving an approximately 40,000 fold compression, in this particular example).

Interestingly, we find that that the number of recombination events that potentially affect the sample grows approximately quadratically with the scaled recombination rate. This is in stark contrast with the (big) ancestral recombination graph, in which the number of recombination events grows exponentially with recombination rate. We also discuss potential applications of the methods developed here to real data.

S/HIC: ROBUST IDENTIFICATION OF SOFT AND HARD SWEEPS USING MACHINE LEARNING

Andrew D Kern[1,2], Daniel R Schrider[1]

[1]Rutgers University, Department of Genetics, Piscataway, NJ, [2]Human Genetics Institute of New Jersey,, Piscataway, NJ

Detecting selective sweeps from whole genome sequencing data is a central problem for population genetics. However to date most methods have been hindered in their performance in realistic demographic scenarios. Moreover, over the past decade there has been a renewed interest in determining the importance of selection from standing variation in adaptation of natural populations, yet very few methods for inferring this model of adaptation at the genome scale have been introduced. Here we introduce a new method, S/HIC, that uses supervised machine learning to precisely infer the location of both hard and soft selective sweeps through a novel combination of spatially explicit summaries of genomic variation data. We show that S/HIC has unrivaled accuracy for detecting sweeps under demographic histories that are relevant to human populations. Moreover we show that S/HIC is uniquely robust among its competitors to model misspecification. Thus even if the true demographic model of a population differs catastrophically from that specified by the user, S/HIC still retains impressive discriminatory power. Finally we apply S/HIC to the case of resequencing data from human chromosome 18 in a European population sample and demonstrate that we can reliably recover selective sweeps that have been identified earlier using less specific and sensitive methods.

A PHYLOGENETIC FRAMEWORK TO STUDY THE EVOLUTION OF TRANSCRIPTIONAL REGULATORY NETWORKS

Christopher Koch[1], Alireza Fotuhi Siahpirani[1], Sushmita Roy[2,3]

[1]University of Wisconsin-Madison, Department of Computer Sciences, Madison, WI, [2]University of Wisconsin-Madison, Department of Biostatistics and Medical Informatics, Madison, WI, [3]University of Wisconsin-Madison, Wisconsin Institute for Discovery, Madison, WI

Transcriptional regulatory networks are largely responsible for gene regulation and information processing in cells, and their evolution has played a critical role in driving the morphological diversity of species. Tremendous strides have been made towards accurately reconstructing regulatory networks from gene expression data in a single organism. However, studying how regulatory networks evolve remains an open challenge, as most network reconstruction approaches have been attempted only for well-characterized organisms. We currently lack approaches to infer regulatory networks in non-model organisms. With the recent availability of comparative functional genomics datasets for multiple phylogenies, including yeast, fly and mammalian species, we have a unique opportunity to exploit these data to predict regulatory networks in non-model organisms. We propose a novel computational algorithm to learn regulatory networks across multiple species that imposes a phylogenetically motivated prior distribution on the graph structures. Our approach has the flexibility to incorporate species-specific information such as instances of transcription factor binding motifs. On simulated data generated from networks of known structure, we find that imposing a phylogenetic prior yields networks of better structure that additionally display phylogenetically coherent patterns of conservation and divergence. We apply our approach to expression datasets for six ascomycete yeast species. We find that our approach is able to recover more phylogenetically coherent networks that are enriched for known and literature-supported interactions. Additionally, our inferred networks reveal novel insights into the conservation of stress regulators. Our framework provides a promising approach to reconstructing regulatory networks across multiple species, enabling comparative studies between related species that can provide meaningful evolutionary insights.

← Back

INFERRING THE DISTRIBUTION OF FITNESS EFFECTS FOR A SPECIES GROUP USING MULTIPLE WHOLE-GENOME SEQUENCES.

Evan Koch[1], John Novembre[2]

[1]University of Chicago, Ecology and Evolution, Chicago, IL, [2]University of Chicago, Human Genetics, Chicago, IL

The distribution of fitness effects of new mutations (DFE) is often inferred using polymorphism data from single species. This approach parameterizes the DFE using the sample-frequency spectrum (SFS), which requires a substantial amount of sequence information from multiple individuals. We show that, if one assumes a group of species share a similar distribution of fitness effects, it is possible to infer a shared DFE using only patterns of heterozygosity in single genomes. This approach uses the fact that levels of polymorphism at genomic positions under purifying selection are dependent on the historical pattern of the effective population size. When a group of species differ sufficiently in their past effective population sizes, the differing levels of heterozygosity at constrained sites depend on the DFE. We estimate this history from single genomes using the Pairwise Sequentially Markovian Coalescent and numerically solve for the expected heterozygosity under a particular DFE in a diffusion model. This provides a link between individual genome sequences and levels of heterozygosity at constrained sites. We show that this can be used to find a DFE that best fits the pattern of deleterious variation in all species considered. This goal is motivated by the growing number genome projects sequencing only one or some small number of individuals from each species. For instance, the Avian Phylogenomics Project has sequenced individuals from 45 bird species, and it may be desirable to investigate differences in the DFE of various ecological types. We develop the above procedure for fitting a DFE and apply it to simulated data to determine how the number of species, differences in demographic history, and aspects of the DFE itself impact the recovery of its salient features.

DEEP LEARNING THE RELATIONSHIP BETWEEN CHROMATIN ARCHITECTURE, CHROMATIN STATE AND TRANSCRIPTION FACTOR BINDING

Chuan Sheng Foo[1], Johnny Israeli[2], Avanti Shrikumar[1], Anshul Kundaje[1,3]

[1]Stanford University, Computer Science, Stanford, CA, [2]Stanford University, Biophysics Program, Stanford, CA, [3]Stanford University, Genetics, Stanford, CA

Assays such as DNase-seq and MNase-seq that profile genome-wide chromatin accessiblity and nucleosome positioning have allowed comprehensive identification of regulatory elements (REs) and characterization of their local chromatin architecture. Different classes of active and repressed REs such as promoters, enhancers and insulators have been found to be associated with distinct combinations of histone modifications defining their chromatin state. Multivariate hidden Markov models are typically used to learn chromatin states from multiple histone modification ChIP-seq experiments for discovery and annotation of diverse REs. However, histone ChIP-seq experiments are time-consuming, costly and require large amounts of input material; limiting their applicability in rare cell types. Recently, the ATAC-seq assay was developed to simultaneously profile chromatin accessibility, nucleosomes and TF footprints at REs from low input samples based on direct in vitro transposition of sequencing adaptors into native chromatin. We train novel deep learning methods based on convolutional neural networks on a novel two-dimensional representation of ATAC-seq data to learn a direct mapping from chromatin architecture to chromatin state by leveraging subtle patterns in insert-size distributions. Using a multi-task, multi-modal formulation we integrate ATAC-seq data and DNA-sequence to simultaneously predict multiple histone modifications, chromatin state and CTCF ChIP-seq binding sites with high accuracy (80-90%). Models trained on a combination of DNase-seq and MNase-seq data also achieve similar high accuracy supporting a fundamental predictive mapping between chromatin architecture and chromatin state. We develop novel feature extraction and visualization methods to peer into the deep neural network models and identify predictive patterns such as nucleosomal asymmetry and TF footprints that are automatically learned from raw data. We explore the feasibility of cross-cell type prediction and determine the minimum sequencing depth requirements for predictive power. Our method could enable characterization of REs from low quantities of input material using a single assay, potentially enabling detailed regulatory maps in rare cell populations in primary tissue.

← Back

UNIFIED PROBABLISTIC FRAMEWORK FOR GENOME-WIDE CHARACTERIZATION OF HUMAN DISEASES

Young-suk Lee[1,2], Arjun Krishnan[2], Olga Troyanskaya[1,2,3]

[1]Princeton University, Computer Science, Princeton, NJ, [2]Princeton University, Lewis-Sigler Institute for Integrative Genomics, Princeton, NJ, [3]Simons Foundation, Simons Center for Data Analysis, New York, NY

Complex diseases are driven by multiple genetic changes and characterized by genome-wide perturbations of cellular pathways and functions. Gene expression profiling experiments have been potent in shedding light on the molecular pathology of diseases, with most studies typically focusing on a single disease and contrasting disease samples to their normal controls. However, such "one disease at a time" approaches disregard similarities and differences in pathological deregulations underlying different complex diseases and are thus unable to identify attributes unique to each particular disease, which is critical for developing targeted therapy. We have developed a unified probabilistic framework to quantify distinctive disease signals based on gene expression profiles. Leveraging thousands of disease-specific profiles from public repositories, this data-driven approach identifies distinctive molecular-level characteristics of each disease from both the functional and anatomical perspectives. Our framework can be used to distinguish between closely-related diseases, identify discerning genes and processes, associate rare-diseases to the nearest well-studied counterparts, and track the effectiveness of therapy.

Our method has several important characteristics:
1. Our approach is data-driven, and thus not susceptible to literature bias toward specific genes/research areas, and can also be used to effectively study genetically uncharacterized or poorly characterized diseases, including rare diseases.
2. The disease signals estimated by our method are consistent with the complex relationships among diseases, and sensitive enough to distinguish between closely related diseases and to detect molecular effects of treatment.
3. Since all diseases were analyzed simultaneously, our model for an individual disease represents what is unique to that disease in the context of all the other diseases. The disease models are biologically interpretable in terms of each disease's functional and anatomical contexts.

No curated set of genes are used in our data-driven approach and so our approach could easily be extendable to any human disease for which high-throughput data can be generated. We find that the most predictive genes identified by our method are significantly under-studied in the biomedical literature, demonstrating that many key biological processes underlying human pathophysiology are in fact in critical need of further investigation.

CAN YOU PREDICT THE SHAPE OF RIBOSOME PROFILES? MARGINAL PROBABILITY DENSITY ESTIMATION OF RIBOSOME FOOTPRINTS

Tzu-Yu Liu[1,2], Yun S Song[1,2,3,4]

[1]University of Pennsylvania, Departments of Mathematics and Biology, Philadelphia, PA, [2]University of California, Berkeley, Department of Electrical Engineering and Computer Sciences, Berkeley, CA, [3]University of California, Berkeley, Department of Statistics, Berkeley, CA, [4]University of California, Berkeley, Department of Integrative Biology, Berkeley, CA

The synthesis of protein is a process central to biology. While much has been studied in the transcription process, the dynamics of decoding the information in the mRNA transcripts into a sequence of amino acids is incompletely understood. Ribosome profiling is a useful experimental technique for studying translation dynamics and predicting protein synthesis. Applications of this technique have shown that ribosomes are not uniformly distributed along the mRNA transcript. Understanding how transcript-specific distribution arises is important for unraveling the translation mechanism. Here, we show that the mRNA sequence context is highly predictive of the shape of ribosome profiles.

We first apply wavelet analysis to represent the marginal distribution of ribosomes as a linear combination of basis functions. These basis functions form various scale of reconstruction of the original signal. By thesholding the number of scales, we decouple the global morphology from local patterns of ribosome marginal densities. This method enables us to remove the possible noise in the raw data. Then, by building sparse models to predict the marginal densities projected onto the subspace corresponding to each scale and using asymmetric kernel density estimation on the codon sequences to construct predictors, we identify the codon features that best associate with a given scale and estimate the extent to which they influence the ribosome densities.

Our results on Saccharomyces cerevisiae data show that the marginal ribosome densities can be predicted with high accuracy. This suggests that the ribosome distribution along a transcript is sequence dependent. More importantly, we find some codons dominating the global shapes, which have an impact not only on the ribosome density at its position but the neighboring positions. Since a ribosome occupies a space of more than one codon, this suggests that the translation speed is determined by several codons occupied by the ribosome. Besides, the possible interference among the ribosomes supports our model using asymmetric kernel density estimation on the sequence content. The proposed method has wide applications, including inferring isoform-specific ribosome footprints, designing transcripts with fast translation speeds, and understanding how the inferred fast and slow codons relate to the distinct stages of translation.

CONTRASTING REGIONAL ARCHITECTURES OF SCHIZOPHRENIA AND OTHER COMPLEX DISEASES USING FAST VARIANCE COMPONENTS ANALYSIS

Po-Ru Loh[1,2], Gaurav Bhatia[1,2], Alexander Gusev[1,2], Hilary K Finucane[3], Brendan K Bulik-Sullivan[2,4], Samuela J Pollack[1,2,5], Teresa R de Candia[6], Sang Hong Lee[7], Naomi R Wray[7], Kenneth S Kendler[8], Michael C O'Donovan[9], Benjamin M Neale[2,4], Nick Patterson[2], Alkes L Price[1,2,5]

[1]Harvard T.H. Chan School of Public Health, Epidemiology, Boston, MA, [2]Broad Institute, Medical and Population Genetics, Cambridge, MA, [3]Massachusetts Institute of Technology, Mathematics, Cambridge, MA, [4]Massachusetts General Hospital, ATGU, Boston, MA, [5]Harvard T.H. Chan School of Public Health, Biostatistics, Boston, MA, [6]University of Colorado Boulder, Psychology and Neuroscience, Boulder, CO, [7]University of Queensland, Queensland Brain Institute, Queensland, Australia, [8]Virginia Commonwealth University, Psychiatry and Human Genetics, Richmond, VA, [9]Cardiff University, MRC Centre for Neuropsychiatric Genetics and GenomicsCardiff, United Kingdom

Heritability analyses of GWAS cohorts have yielded important insights into complex disease architecture, and increasing sample sizes hold the promise of further discoveries. Here, we analyze the genetic architecture of schizophrenia in 49,806 samples from the Psychiatric Genomics Consortium, and nine complex diseases in 54,734 samples from the GERA cohort. For schizophrenia, we infer an overwhelmingly polygenic disease architecture in which ≥71% of 1Mb genomic regions harbor at least one variant influencing schizophrenia risk. We also observe significant enrichment of heritability in GC-rich regions and in higher-frequency SNPs for both schizophrenia and GERA diseases. In bivariate analyses, we observe significant genetic correlations (ranging from 0.18 to 0.85) among several pairs of GERA diseases; genetic correlations were on average 1.3x stronger than correlations of overall disease liabilities. To accomplish these analyses, we developed a fast algorithm, BOLT-REML, for multi-component, multi-trait variance components analysis that overcomes prior computational barriers that made such analyses intractable at this scale.

The overall framework of the BOLT-REML algorithm is Monte Carlo AI REML, a Newton-type iterative optimization of the (restricted) log likelihood with respect to the variance parameters sought. BOLT-REML begins a multi-variance component analysis by computing an initial estimate of each parameter using the single variance component estimation procedure of BOLT-LMM (Loh et al. 2015 Nat Genet), which is the only analysis possible with BOLT-LMM. Then, in each iteration, BOLT-REML rapidly approximates the gradient of the log likelihood using pseudorandom Monte Carlo sampling and the Hessian of the log likelihood using the average information matrix. The approximate gradient and Hessian produce a local quadratic model of the likelihood surface, which we optimize within an adaptive trust region radius—key to achieving robust convergence—to update the variance parameter estimates. These procedures allow BOLT-REML to consistently achieve convergence in $\approx O(MN^{1.5})$ time; in contrast, existing multi-component REML algorithms are less robust and/or require $O(MN^2+N^3)$ time (e.g., GCTA). For example, a six-variance component analysis of N=50K GERA samples typed at M=600K SNPs that would have required ~150 CPU hours and ~200 GB RAM using GCTA required only 16 CPU hours and 7 GB RAM using BOLT-REML.

← Back

FAST AND ACCURATE LONG-RANGE PHASING IN A UK BIOBANK COHORT

Po-Ru Loh[1,2], Pier Francesco Palamara[1,2], Alkes L Price[1,2,3]

[1]Harvard T.H. Chan School of Public Health, Department of Epidemiology, Boston, MA, [2]Broad Institute of Harvard and MIT, Program in Medical and Population Genetics, Cambridge, MA, [3]Harvard T.H. Chan School of Public Health, Department of Biostatistics, Boston, MA

Recent work has leveraged the unique genealogical structure and extensive genotyping (>30%) of the Icelandic population to provide accurate long-range phasing (LRP), enabling whole-genome sequence data to be effectively imputed using a subset of sequenced samples (Gudbjartsson et al. 2015 Nat Genet). Here, we develop a fast and accurate LRP method that enables this paradigm to be extended to outbred populations by harnessing long (>4 cM) identical-by-descent (IBD) tracts shared among distantly related individuals. We applied the method, Eagle, to N=150K samples (0.2% of the British population) from the UK Biobank interim release, and we determined that it is fast (~5 min per individual) and highly accurate (median switch error rate <0.5%). We further observed that in tests with 2% of genotypes randomly masked, Eagle imputed masked genotypes with high accuracy (r^2>0.75) down to a minor allele frequency of 0.1%; thus, sequencing of N=150K samples is sufficient for effective imputation down to 0.1% MAF in this population. We benchmarked Eagle against SHAPEIT2 (Delaneau et al. 2013 Nat Meth) and observed that SHAPEIT2 required ~3x as much time to phase the data in 10 batches of N=15K subsets, achieving ~1.5% median switch error. We estimate that SHAPEIT2 would require ~20x as much time as Eagle to phase all N=150K samples together; while this task was computationally infeasible, we ran SHAPEIT2 to completion on all samples for the first 40cM of chromosome 10 and observed accuracy similar to Eagle.

The Eagle algorithm has three main components. First, Eagle rapidly detects probable IBD tracts by identifying long regions of agreement at homozygous sites (i.e., IBS≥1), scoring identified regions using allele frequency and linkage disequilibrium information, and checking overlapping regions for consistency; Eagle uses the detected IBD to perform accurate initial long-range phasing in high-IBD regions. Second, Eagle performs local phase refinement in overlapping ~1 cM windows by detecting complementary haplotypes (among haplotypes inferred in the previous step) using locality-sensitive hashing; specifically, Eagle searches for long haplotypes consistent with each diploid individual and then searches for hash matches to the implied complementary haplotypes. Third, Eagle finalizes phase calls by running two iterations of HMM-based phasing using ~100 local reference haplotypes and aggressively pruning the search space to 100 states per position. All three steps are multithreaded and make use of bit operations to perform computations in 64-SNP blocks. We are releasing Eagle as open source software.

← Back

BASECALLING FROM RAW OXFORD NANOPORE DATA

Gerton Lunter

University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom

Oxford Nanopore's MinION platform has many interesting features, including its small size (it is operated from an ordinary laptop), its ability to sequence very long molecules and to physically select or reject molecules to sequence in real time. However, as expected from a single-molecule platform, the reads are fairly noisy, and this affects many applications.

The MinION device measures the current of H^+ ions through a pore that is partially blocked by a single-stranded DNA molecule. The specific configuration of DNA bases within the pore modulates this H^+ current. A special "motor" enzyme periodically and stochastically moves the DNA molecule through the pore, one base at a time, and the resulting stepwise changes in the H^+ current are decoded to reveal the original DNA sequence. The raw signal consists of current measurements at 5000 Hz, for each of the 500 pores. To deal with this large data volume, this signal is processed in real time to identify the intervals of constant current, referred to as "events". Event sequences are then uploaded to the cloud and converted to DNA sequences Nanopore's cloud basecaller, Metrichor.

Inevitably, the event caller makes mistakes, and in the absence of raw data these are difficult to correct downstream, leading to indel errors. Here I present a base caller that works directly off the raw signal. In addition to improved base qualities, the model in principle allows identification of deviations from the expected signals, to identify base modifications. I will present some initial results of this method to partially methylated *E. coli* sequences.

SIMULTANEOUSLY QUANTIFYING SCNA AND SVS IN CANCER GENOMES USING A PROBABILISTIC GRAPHICAL MODEL

Yang Li[1], Jian Ma[1,2]

[1]University of Illinois at Urbana-Champaign, Bioengineering, Urbana, IL, [2]University of Illinois at Urbana-Champaign, Institute for Genomic Biology, Urbana, IL

Genome aneuploidy, in which abnormal copy numbers of alleles are present, is a hallmark of cancer. A large number of tumor genomes are aneuploid and have undergone either large somatic copy number alterations (SCNAs) or even whole-genome duplications (WGD). Structural variations (SVs), including deletions, insertions, duplications and rearrangements, can further modify aneuploid cancer genomes into a mixture of rearranged genomic segments with extensive range of SCNAs. A comprehensive and precise characterization of these complex genomic changes is crucial in understanding the evolution of cancer genome and in interpreting cancer-specific functional genomic data.

Allele-specific SCNAs analysis has been performed for SNP array data and recently for NGS data as well. Separately, SV identification methods have also been developed for NGS data. It is essential to ask how SVs interact with SCNAs and how different SVs interact with each other. The answers to such questions can help unravel the detailed cancer genome structure. However, no integrative method specifically designed for simultaneously analyzing SVs and SCNAs has been developed.

We developed a novel algorithm called Weaver to simultaneously identify allele-specific SCNAs and SVs as well as their inter-connectivity in aneuploid cancer genomes. We first search for SVs based on NGS reads at base-pair resolution and build cancer genome graph, which is subsequently converted to a pair-wise Markov random field (MRF) -- a probabilistic graphical model for estimating joint probabilities. In the MRF, the allele-specific copy numbers are hidden states in nodes, and the observations contain all sequencing information, including coverage and linkage. Therefore, our goal of simultaneously quantifying SCNAs and SVs in allele-specific manner is formulated as an aggregated inference problem given sequencing data, which can be viewed as finding the maximum a posteriori (MAP) solution for MRF. Our simulation evaluation and extensive real data applications (on MCF-7, HeLa, and large-scale TCGA whole genome sequencing data) demonstrated that Weaver is highly accurate and can significantly refine the analysis of complex cancer genomes with novel capabilities. We expect that Weaver will be very useful to fully utilize the existing datasets in large-scale projects such as TCGA and ICGC.

GENOMIC-WIDE EVIDENCE OF INTER-CHROMOSOMAL LINKAGE DISEQUILIBRIUM.

Fabrizio Mafessoni, Kay Pruefer

Max Planck for Evolutionary Anthropology, Evolutionary Genetics, Leipzig, Germany

Studies in drosophila and yeast suggest that interactions between different genes often lead to non-linear phenotypic effects, a phenomenon defined as epistasis. Epistatic partners are likely candidates of joint natural selection, causing shifts in allele frequencies, and, more distinctively, associations between allelic variants at distant loci.

Here, we present a genome-wide scan for linkage between loci lying on different chromosomes, that cannot be explained by physical linkage. We find evidence of higher inter-chromosomal genotypic linkage disequilibrium in functional regions than in putatively neutral regions, suggesting a role of joint selection on epistatic partners. The genes involved are more often expressed in the same tissues. Alternative factors can create similar signals of long-distance linkage. Two of these factors, population structure and mapping artifacts, would not lead to the observed joint expression patterns and enrichment in genic regions. We also investigate whether undetected translocations drive such signal by analyzing local patterns of linkage disequilibrium.

Signatures of epistatic selection are expected to be short-lived, although non-random mating affects the rate of decay of linkage disequilibrium. We perform simulations to estimate the age of detectable adaptive events, and the required intensity of selection to lead to levels of inter-chromosomal linkage disequilibrium analogous to that observed.

HAPLOTYPE PHASING IN DOWN SYNDROME FAMILY TRIOS

Daniel Malmer[1], Robin Dowell[1,2,3]

[1]University of Colorado, Boulder, Department of Computer Science, Boulder, CO, [2]University of Colorado, Boulder, BioFrontiers Institute, Boulder, CO, [3]University of Colorado, Boulder, Department of Molecular, Cellular, and Developmental Biology, Boulder, CO

Haplotype phasing, the process of determining the precise sequence of alleles on homologous chromosomes, is an important early step in many association, imputation, and allele-specific expression studies. Leveraging pedigree information has been shown to vastly improve computational phasing methods, but no models currently allow for anueploid individuals within a typical euploid pedigree. Specifically, no methods exist to phase mother-father-child family trios where the child has Down syndrome (trisomy 21). Here we present a new model for phasing mother-father-child family trios that leverages family pedigree information and allows for a nondisjunction event in either of the parents.

The model is a factorial profile hidden Markov model (FPHMM) where each individual in the trio is associated with a separate HMM and the child HMM is dependent on the two parent HMMs. The FPHMM takes as input a file containing single-nucleotide polymorphism (SNP) data and a file containing mapped reads for each member of the family trio. Each position on the genome containing a SNP in any of the family members emits a state containing the phased nucleotides for each individual. Transition probabilities are calculated by leveraging paired- or single-end reads that span multiple heterozygous SNP locations as well as population haploblock frequencies. The child HMM's transition probabilities are also dependent on the states of the parents. Emission probabilities are calculated using read counts at SNP positions as well as nucleotide frequency priors.

The model works on both diploid and trisomy children belonging to a mother-father-child family trio. In future iterations, we plan to extend the model to leverage information from a typical euploid sibling. Additionally, work is being done to capture recombination events in the parents using a silent state in the child's HMM within the larger FPHMM.

We test on simulation data generated by a combination of existing tools, which generate DNA sequences of family pedigrees and simulate reads from next-generation sequencing technologies, and in-house scripts, which simulate the nondisjunction events in the parents. Additionally, we run the model on real SNP and read data from four family trios, each consisting of two typical parents and a child with Down syndrome.

TESTING DIRECTIONAL SELECTION ON POLYGENIC TRAITS USING ANCIENT DNA

Joseph H Marcus[1], Charleston Chiang[2], John Novembre[1]

[1]University of Chicago, Department of Human Genetics, Chicago, IL,
[2]University of California, Los Angeles, Department of Ecology and Evolutionary Biology,, Los Angeles, CA

In many cases, understanding the evolutionary basis of trait variation requires the ability to distinguish between models where a trait has recently undergone directional selection as opposed to stabilizing selection or simply has been neutrally evolving. Methods to distinguish amongst such models are still poorly developed, especially for highly polygenic traits. Recent progress in understanding the polygenic basis of trait variation, using genome-wide association studies (GWAS), and the increasing availability of ancient DNA (aDNA) samples provide new opportunities for understanding selection on polygenic traits. Here, we develop and compare several possible inferential procedures that intersect putative quantitative trait loci discovered via GWAS with aDNA data to test whether a phenotype has experienced directional selection on the basis of allele frequency change. We show that power exists even with small aDNA sample sizes when the signature of selection is distributed across many loci, and we apply this approach to investigate signatures of selection on height in humans, where directionality of selection may vary across populations. While aDNA studies in some species are rapidly scaling upwards in sample size, even small samples will be helpful for shedding light on the evolution of polygenic traits.

FAST PROBABILISTIC VARIANT INTEGRATION AND GENOTYPING USING EXACT ALIGNMENT OF *K*-MERS TO VARIANT GRAPHS

Jonas A Sibbesen, Lasse Maretty, Anders Krogh

University of Copenhagen, Department of Biology, Copenhagen, Denmark

Current approaches to discovery and genotyping of complex variants from high-throughput genome sequencing data generally employ a combination of orthogonal variant callers to ensure variant sensitivity and thus genotyping accuracy. However, approaches capable of integrating such sets of variant calls with the raw data (and possibly previously annotated variants) to produce a consistent set of genotypes with confidence estimates are missing.

We here present BayesTyper, a fully probabilistic approach to genotyping a population of individuals on a fixed set of arbitrarily complex variants (SNVs, insertions, deletions etc.). The method is based on exact alignment of read *k*-mers to a variant graph and hence has no intrinsic bias towards the reference sequence.

The algorithm first counts all non-singleton *k*-mers in the sequencing reads for each individual. Next, loci less than *k* nucleotides apart are joined to form a variant graph in which all possible haplotypes are enumerated except for larger graphs, where a heuristic is used to limit the number of haplotypes. Finally, the multiset containing all *k*-mers found in haplotypes are enumerated and combined with the corresponding sample table to provide a vector containing the occurrences of haplotype *k*-mers for each sample.

We model this observed count vector as generated by combining counts obtained from an individual's diplotype with counts originating from a carefully designed noise process. An individual's diplotype is in turn modelled as drawn from a shared population of haplotypes whose frequencies are modelled using a novel sparse prior. The posterior distribution over genotypes is inferred using collapsed Gibbs sampling of diplotypes, haplotype frequencies and noise states.

We demonstrate that our method is accurate as judged using both extensive simulations and real data from a large sequencing project. BayesTyper is available at www.github.com/bioinformatics-centre/BayesTyper.

We note that a part of this work was also presented at the Biological Sequence Analysis and Probabilistic Models conference 2014.

LEARNING TIME-VARYING GENE REGULATORY NETWORKS FROM GENE EXPRESSION DATA THROUGHOUT THE DEVELOPMENT OF DROSOPHILA USING CAUSAL INFERENCE ALGORITHMS.

Lenka Matejovicova[1], Caroline Uhler[1,2]

[1]IST Austria, Statistics Group, Klosterneuburg, Austria, [2]MIT, Institute for Data, Systems and Society, Cambridge, MA

Interactions between genes and gene regulation are crucial for the functioning of living organisms. The network representing these biological interactions can be characterized by a graph with directed edges describing direct transcriptional effects and weights of these edges representing the strength of these causal effects. Learning such networks by inferring directed graphs from observational data is a well-studied problem, but so far, methods have only been developed to estimate individual networks. In this work, we present a pipeline to learn a whole sequence of related networks in order to infer the gene regulatory network of Drosophila as it changes throughout the development.

To infer a single graph, we use a combination of the new Sparsest Permutation algorithm and our Permutation Equivalence Search. This combination has several qualities that make it well-suited for use in biological applications: first, the method selects for sparse graphs, which is also a natural property of biological networks and second, the method can take into account prior biological knowledge. The latter leads to a search that is more efficient and results in a network that satisfies the biological constraints. Due to the large number of interacting elements typical for non-trivial biological networks, many current algorithms seem to be computationally intractable or lead to many missing edges. Our algorithms have been proven to be consistent under strictly weaker conditions than previous algorithms and efficient enough to infer individual large directed acyclic graphs.

For learning multiple related networks from limited, non-independent datasets, we develop an extension of this method. Here, to gain necessary statistical power, we make an assumption that the networks change smoothly in time and we use also the data from closely related data-points, weighting them accordingly.

We apply the proposed new methods to learn time-varying transcriptional gene regulatory networks from observational gene expression data throughout the development of Drosophila. We compare our results to previous studies estimating the gene association graph, an undirected graph, and discuss our results with respect to experimentally validated pathways.

TIME-RESOLVED ESTIMATION OF THE EFFECTS OF LINKED SELECTION IN HUMAN EVOLUTION

Aaron J Sams, Philipp W Messer

Cornell University, Department of Biological Statistics and Computational Biology, Ithaca, NY

The pairwise sequentially Markovian coalescent (PSMC) model is increasingly being used as a standard tool for inferring a species' demographic history. The PSMC model is based on the insight that the distribution of local times since the most recent common ancestor (TMRCA) between two alleles in a diploid genome provides information about the history of population size changes over time. However, TMRCA distributions can also be affected by the effects of linked selection, such as background selection and hitchhiking during selective sweeps. In particular, TMRCA values should be reduced in genomic regions where selection is frequent and recombination rate is low. Here we show that the PSMC model can be used to quantify the impact of linked selection on patterns of genetic diversity by comparing the history of inferred effective population sizes between genomic regions with distinct levels of functional density. The time-resolution of PSMC further allows us to study how the effects of linked selection have varied over time, indicating historical periods in the evolution of a species during which selection has been more or less effective. We apply this approach to diploid genome sequences from a number of human individuals from Africa, Europe, and Asia and discuss how the out-of-Africa migration has influenced the effectiveness of selection between African and non-African populations.

INTEGRATING CODING AND REGULATORY ABERRATIONS TO CAPTURE DISEASE MECHANISMS USING A PROBABILISTIC GRAPHICAL MODEL

Aziz M Mezlini[1,2], Fabio Fuligni[1], Adam Shlien[1], Anna Goldenberg[1,2]

[1]Hospital for Sick Children, Genetics and Genome Biology, Toronto, Canada, [2]University of Toronto, Computer Science, Toronto, Canada

Majority of human diseases are complex, arising due to a multitude of factors. Identifying these factors is critical to understanding diseases and improving health care, yet it is a very difficult computational problem: low signal-to-noise ratio (only a few variants out of millions are likely to be causal), heterogeneity of reasons (e.g. coding, regulatory, epigenetic), epistasis (gene interaction patterns), etc.
We propose to combine two mostly complementary data sources: coding variants and gene expression. These two data sources are responsible for different kinds of protein aberrations. Combining them allows us to survey both coding and regulatory aberrations genome wide without underpowering the model. We developed a biologically motivated hierarchical factor graph model which efficiently combines these two sources of data. We use variant harmfulness and gene interactions as priors, to increase the likelihood of identifying the genes correctly. To our knowledge, this is the first work that takes into account complementarity of exome and gene expression data sources in a principled way, integrating variant harmfulness and gene interaction information in the inference process of the model. Our approach a) allows to integrate different data modalities; b) provides a principled way to aggregate rare (and common) variants; c) improves the power of detecting genes associated with a given disease; d) implicates proteins that have been affected in the population through regulatory mechanisms as well as the coding DNA sequence.
Our extensive simulations confirm that our method has superior sensitivity and precision compared to other methods that aggregate rare variants. We have tested our approach in a large breast cancer dataset as a proof of concept and found that our method is able to identify important breast cancer genes. Interestingly, we find genes that have DNA mutations or coding variants in some patients and gene expression aberrations in other patients, indicating that our method is able to effectively explain the disease in a larger number of patients.

ESTIMATION OF PHYLOGENETIC BIRTH AND DEATH RATES FOR SMALL, NON-CODING RNAS IN THE PRESENCE OF ANNOTATION ERROR.

Jaaved Mohammed[1,2,3,4], Eric C Lai[3,4], Adam Siepel[1,3]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Cornell University, Department of Biological Statistics and Computational Biology, Ithaca, NY, [3]Tri-Institutional Training Program in Computational Biology and Medicine,,, New York, NY, [4]Sloan-Kettering Institute, Department of Developmental Biology, New York, NY

Despite the increased depth of whole-genome transcriptome datasets, accurate estimation of evolutionary birth and death rates for small, non-coding RNAs (sncRNAs) remains a challenge. Deep transcriptome datasets indicate many apparently novel, recently-evolved elements, some of which have borderline evidence. Unbiased estimates of birth and death rates therefore require methods that are capable of allowing for error in the annotation of low-evidence, putative nascent sncRNAs. This problem is especially apparent when transcriptomic datasets are heterogeneous across species.

Here, we present a phylogenetic probabilistic graphical model that accounts for annotation error in sncRNAs. This model permits estimation of both global and branch-specific rates of sncRNA gene birth and death, allowing for either global or leaf-specific error rates in the observed data. Given a phylogenetic tree and binary observations of presence or absence for a collection of sncRNA families, our method allows us (1) to predict individual edge-wise gain and loss events, (2) to infer overall birth and death rates; and (3) to infer leaf-wise error-rate parameters. We demonstrate the accuracy of our method with simulated and real data.

In addition, we apply our method to recent annotations of microRNAs (miRNAs) based on small-RNA sequence data from 12 *Drosophila* species. We estimate rates of miRNA birth and death using versions of our model that both allow and do not allow for annotation error, and show striking rate variation across miRNAs of varied biogenesis classes and genomic locale. We recommend the application of this method for other small, functional genomic elements, such as transcription factor binding sites, and other classes of short and long ncRNAs.

POPULATION STRATIFICATION CONTRIBUTION TO GENOMIC HERITABILITY

Gota Morota[1], Matthew L Spangler[1], Stephen D Kachman[2]

[1]University of Nebraska, Animal Science, Lincoln, NE, [2]University of Nebraska, Statistics, Lincoln, NE

The availability of the large volume of molecular markers has reshaped the landscape of statistical approaches to characterize population structure. Population stratification arises when the contributions of different ancestral populations to phenotype vary across subpopulations. Principal components analysis offers a means to infer population structures by associating the distribution of genetic variation against ethnic backgrounds, breed specifications, or geographical coordinates. Despite its popularity, it has little to offer in explaining the contribution of each eigenvector on phenotypic variation. Therefore, we sought to apply a reparameterized genomic best linear unbiased prediction (GBLUP) model to infer the impact of population stratification on the estimates of genomic heritability. Using GBLUP, a phenotype is regressed on eigenvectors extracted from a genomic relationship matrix, and genomic heritability is expressed as a function of regression coefficients for eigenvectors. The influence of eigenvectors was quantified with hot carcass weight in admixed beef cattle of unknown source populations, male flowering time in maize, and milk yield in homogeneous dairy cattle populations. The estimates of genomic heritability employing the entire eigenvectors were 0.42, 0.23, and 0.83 for beef cattle, maize, and dairy cattle, respectively. The exclusion of the top five eigenvectors decreased the estimates of genomic heritability, ranging from 0.38, 0.19, to 0.82. With the removal of the first 5, 10, or 20 eigenvectors, the proportion of reduction in genomic heritability relative to the total genomic heritability was highest in the beef cattle and maize data sets. Moreover, the impact of population structure defined as the first five eigenvectors explained 0.20, 0.21, and 0.06% of the total additive genetic variance. Underlying population structure may inflate genomic heritability estimates derived from all eigenvectors. The reductions in genomic heritability estimates are conjectured as the results of the correction of such a stratification. The findings revealed that GBLUP has a potential to advance the understanding of the genotype-to-phenotype map under population stratification by expanding the scope of possible approaches. We envision that these attempts to understand the extent to which population stratification contributes to quantitative genetic parameters, such as genomic heritability, will enhance the characterization of the genetic make-up of heterogeneous populations.

SCALABLE BAYESIAN KERNEL MODELS WITH VARIABLE SELECTION FOR TRAIT MAPPING

Lorin Crawford[1], Kris Wood[2], Sayan Mukherjee[1,3,4]

[1]Duke University, Statistical Science, Durham, NC, [2]Duke University, Cancer Biology and Pharmacology, Durham, NC, [3]Duke University, Computer Science, Durham, NC, [4]Duke University, Mathematics, Durham, NC

Nonlinear kernels are used extensively in regression models in statistics and machine learning since they often improve predictive accuracy. Variable selection is a challenge in the context of kernel based regression models. In linear regression the concepts effect size for the regression coefficients is very useful for variable selection. In this paper we provide an analog for the effect size of each explanatory variable for Bayesian kernel regression models when the kernel is shift-invariant, for example the Gaussian kernel. The key idea that allows an analog of the effect size is a random Fourier expansion for shift-invariant kernel functions. These random Fourier bases serve as a linear vector space in which a linear model can be defined and regression coefficients in this vector space can be projected onto the original explanatory variables. This projection serves as the analog for effective sample size. We apply this idea to problems in functional genomics as well as statistical genomics. We apply this framework to extract a signature of dysregulation from expression data of *BRAF* melanoma and expression quantitative trait (eQTL) mapping data.

DIMENSIONAL REDUCTION OF METAGENOMIC DATA BY FINDING ECOLOGICALLY EQUIVALENT SPECIES

Senthil Kumar Muthiah[1], Héctor Corrada Bravo[2], Eric V Slud[3], Mihai Pop[2]

[1]Bioinformatics and Genomics Program & Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD,
[2]Department of Computer Science & Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD,
[3]Department of Mathematics, University of Maryland, College Park, MD

Statistical inference of taxas ("species") observed in large-scale 16S metagenomic surveys is of considerable biomedical interest. The large number of species thus discovered (albeit with only a few dominating/abundant ones) and excess zeroes in the species count distributions, however, make it a challenge for performing statistical analyses. In this work, we mitigate these issues by aggregating counts of carefully chosen species that behave similarly to latent ecological factors and environmental processes. It is well known that the relative abundances of such ecologically equivalent/nearly-equivalent species are not necessarily influenced by changes in environmental conditions across local and regional scales, but their summed total abundance, however, is (Hubbell, 2001, Leibold & McPeek, 2006). We construct a Bayesian nonparametric model, and two posterior inference algorithms based on Gibbs sampling and Collapsed Variational Inference (Teh et al., 2007), which find clusters of species that satisfy this constraint. We subsequently create a reduced dataset that features these clusters as new units of analysis interest (termed "Equivalence Class Units") and their summed counts as new measurements across environmental conditions. In addition to allowing for an unbounded number of ECUs, and accounting for the compositional nature of sequencing data (Aitchison, 1982, Friedman & Alm, 2013), our prior can also incorporate known relationships among the species to be clustered (example, a taxonomic tree). Our approach is applicable to datasets with few thousands of species.

References:
1. Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological) 139–177.
2. Friedman, J., and Alm, E.J. (2012). Inferring correlation networks from genomic survey data. PLoS Computational Biology 8, e1002687.
3. Hubbell, S.P. (2001). The unified neutral theory of biodiversity and biogeography (Princeton University Press).
4. Leibold, M.A., and McPeek, M.A. (2006). Coexistence of the niche and neutral perspectives in community ecology. Ecology 87, 1399–1410.
5. Teh, Y.W., Kurihara, K., and Welling, M. (2007). Collapsed variational inference for HDP. In Advances in Neural Information Processing Systems, 1481–1488.

GENE AND NETWORK ANALYSIS OF COMMON VARIANTS REVEALS NOVEL ASSOCIATIONS IN COMPLEX TRAITS

Priyanka Nakka[1,2], Benjamin J Raphael[2,3], Sohini Ramachandran[1,2]

[1]Brown University, Ecology and Evolutionary Biology, Providence, RI, [2]Brown University, Center for Computational Molecular Biology, Providence, RI, [3]Brown University, Computer Science, Providence, RI

Genome-wide association studies (GWAS) have been used widely to identify loci associated with quantitative and complex traits. GWAS test the hypothesis that individual mutations of large effect generate phenotypes of interest. However, complex and quantitative traits are known to exhibit genetic heterogeneity on multiple levels: 1) the phenotype may be generated by multiple mutations within an associated gene; and 2) mutations in distinct genes within a pathway may interact and produce a phenotype. In both cases, GWAS are unlikely to uncover all associated mutations.

A common approach to address this missing heritability is to conduct gene and pathway analysis combining single nucleotide polymorphism (SNP) p-values to detect novel associations between genes and phenotypes of interest, and uncover biological pathways underlying complex phenotypes. However, popular methods for generating gene p-values from SNP p-values are computationally inefficient, produce imprecise p-values, are biased by gene length, and have low true positive rate at low false positive rates; these drawbacks strongly bias downstream pathway analysis.

Here we introduce a new method, PEGASUS (the Precise, Efficient Gene Association Score Using SNPs), for analytically generating gene p-values from SNP p-values that corrects for linkage disequilibrium (LD). We find that PEGASUS is unbiased by gene length and produces gene p-values with as much as 10 orders of magnitude higher precision than a competing method that models LD, while also running twice as fast. We use simulations to assess the accuracy of our method and find that it outperforms minSNP (using the smallest SNP p-value in a gene as the gene p-value) with 30% higher true positive rate when the false positive rate is fixed at 1%. We also find that PEGASUS gene results for real GWA datasets are enriched for known gene associations by as much as 2.8-fold in comparison to minSNP results. We use gene p-values from PEGASUS as input to the HotNet2 algorithm. We then identify networks of interacting genes harboring variants associated with several traits by adapting HotNet2 for use with common variants. We uncover interactions between genes previously associated with Waist-Hip Ratio and Ulcerative Colitis along with novel candidate genes; in contrast, existing methods do not identify key subnetworks previously associated with these phenotypes. We also identify subnetworks for Attention-Deficit/Hyperactivity Disorder, a phenotype that has no known genome-wide significant SNPs in GWAS.

← Back

COMPREHENSIVE GENOME AND TRANSCRIPTOME STRUCTURAL ANALYSIS OF A BREAST CANCER CELL LINE USING PACBIO LONG READ SEQUENCING

Maria Nattestad[1], Karen Ng[2], Sara Goodwin[1], Timour Baslan[1], Fritz Sedlazeck[1], James Gurtowski[1], Elizabeth Hutton[1], Marley Alford[1], Elizabeth Tseng[3], Jason Chin[3], Timothy Beck[2], Yogi Sundaravadanam[2], Melissa Kramer[1], Eric Antoniou[1], John McPherson[2], James Hicks[1], Michael Schatz[1], Richard McCombie[1]

[1]Cold Spring Harbor Laboratory, Quantitative Biology, Cold Spring Harbor, NY, [2]Ontario Institute for Cancer Research, Computational Biology, Toronto, Canada, [3]Pacific Biosciences, Bioinformatics, Menlo Park, CA

Genomic instability is one of the hallmarks of cancer, leading to widespread copy number variations, chromosomal fusions, and other structural variations in many cancers. The breast cancer cell line SK-BR-3 is an important model for HER2+ breast cancers, which are among the most aggressive forms of the disease and affect one in five cases. Through short read sequencing, copy number arrays, and other technologies, the genome of SK-BR-3 is known to be highly rearranged with many copy number variations, including an approximately twenty-fold amplification of the HER2 oncogene, along with numerous other amplifications and deletions. However, these technologies cannot precisely characterize the nature and context of the identified genomic events and other important mutations may be missed altogether because of repeats, multi-mapping reads, and the failure to reliably anchor alignments to both sides of a variation.

To address these challenges, we have sequenced SK-BR-3 using PacBio long read technology. Using the new P6-C4 chemistry, we generated more than 70X coverage of the genome with average read lengths of 9-13kb (max: 71kb). Using Lumpy as well as our novel assembly-based algorithms for analyzing split-read alignments, we have developed a detailed map of structural variations in this cell line. Taking advantage of the newly identified breakpoints and combining these with copy number assignments, we have developed an algorithm to reconstruct the mutational history of this cancer genome. From this we have characterized the amplifications of the HER2 region, discovering a complex series of nested duplications and translocations between chr17 and chr8, two of the most frequent translocation partners in primary breast cancers. We have also carried out full-length transcriptome sequencing using PacBio's Iso-Seq technology, which has revealed a number of previously unrecognized gene fusions and isoforms. Combining long-read genome and transcriptome sequencing technologies enables an in-depth analysis of how changes in the genome affect the transcriptome, including how gene fusions are created across multiple chromosomes. This analysis has established the most complete cancer reference genome available to date, and is already opening the door to applying long-read sequencing to patient samples with complex genome structures.

A COMPOSITE LIKELIHOOD APPROACH TO ESTIMATE MIGRATION RATES BETWEEN 2 POPULATIONS ON 4 SEQUENCES USING A COALESCENCE HIDDEN MARKOV MODEL

Svend V Nielsen, Thomas Mailund

Aarhus University, Bioninformatics Research Centre, Aarhus, Denmark

Given genetic data, our model aims to estimate migration rates and population sizes in a fixed number of epochs. By using 4 sequences it is possible to estimate separate migration rates for each direction. The 4 sequences are used to make two intrapopulation and one interpopulation alignment and each of these alignments gives one likelihood.
A likelihood is obtained by modelling the coalescent process with a hidden markov model where discretized coalescence times are the hidden states. In order to calculate the transition probabilities and emission probabilities of the HMM, continuous time Markov chains are constructed
The composite likelihood is augmented with priors on the parameters allowing Markov chain Monte Carlo methods to produce joint posterior distributions of the parameters. The chosen method is a parallelized Metropolis coupled MCMC but with adaptive proposals and adaptive temperatures. Good performance has been observed when increasing the number of jumps between chains to the maximum.
With 4 epochs, we are able to obtain very reasonable migration rates and population sizes of the first two epochs correctly on simulated data from a coalescence model. The last two epochs still suffers.

IDENTIFICATION AND CHARACTERIZATION OF CONSERVED SMALL ORFS IN ANIMALS

Sebastian D Mackowiak[1], Henrik Zauber[1], Chris Bielow[1,2], Denise Thiel[1], Kamila Kutz[1], Lorenzo Calviello[1], Guido Mastrobuoni[1], Nikolaus Rajewsky[1], Stefan Kempa[1], Matthias Selbach[1], Benedikt Obermayer[1]

[1]Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine, Berlin, Germany, [2]Berlin Institute of Health,, Berlin, Germany

There is increasing evidence that non-annotated short open reading frames (sORFs) can encode functional micropeptides, but computational identification remains challenging. Here we predict conserved sORFs in human, mouse, zebrafish, fruit fly and the nematode *C. elegans*. Isolating specific conservation signatures indicative of purifying selection on encoded amino acid sequence, we identify about 2000 novel sORFs in the untranslated regions of canonical mRNAs or on transcripts annotated as non-coding. Predicted sORFs show stronger conservation signatures than those identified in previous studies and are sometimes conserved over large evolutionary distances. Encoded peptides have little homology to known proteins and are enriched in disordered regions and short interaction motifs. Published ribosome profiling data indicate translation for more than 100 of novel sORFs, and mass spectrometry data gives peptidomic evidence for more than 70 novel candidates. We thus provide an integrated resource of conserved micropeptides for functional validation *in vivo*.

LEVERAGING DISTANT RELATEDNESS TO QUANTIFY HUMAN MUTATION AND GENE CONVERSION RATES

Pier Francesco Palamara[1,2], Laurent Francioli[3], Giulio Genovese[2], Peter Wilton[6], Alexander Gusev[1,2], Hilary Finucane[1,2], Sriram Sankararaman[2,4], Shamil Sunyaev[2,4], Paul deBakker[3], John Wakeley[6], Itsik Pe'er[7], Alkes Price[1,2,5]

[1]Harvard School of Public Health, Dept. of Epidemiology, Boston, MA, [2]Broad Institute, Program in Medical and Population Genetics, Cambridge, MA, [3]University Medical Center Utrecht, Dept. of Medical Genetics, Utrecht, Netherlands, [4]Harvard Medical School, Dept. of Genetics, Boston, MA, [5]Harvard School of Public Health, Dept. of Biostatistics, Boston, MA, [6]Harvard University, Dept. of Organismic and Evolutionary Biology, Cambridge, MA, [7]Columbia University, Dept. of Computer Science, New York, NY

The rate at which human genomes mutate is a central biological parameter that has many implications for our ability to understand demographic and evolutionary phenomena. We present a method for inferring mutation and gene conversion rates using the number of sequence differences observed in identical-by-descent (IBD) segments. The method relies on three key components. First, coalescent modeling is used to obtain the mean posterior time to most recent common ancestor (tMRCA) of IBD segments in an inferred demographic model. Second, we regress the observed sequence mismatches for several IBD length thresholds on the estimated tMRCA; the slope of this regression reflects the rate at which new mutations accumulate per generation time unit, while the genotyping error rate is captured by the intercept. Third, by quantifying how estimates vary as a function of allele frequency, we infer the probability that a site is involved in non-crossover gene conversion. We validate this method using extensive simulation, and apply it to 498 trio-phased sequenced Dutch individuals. We infer a point mutation rate of $1.66 \pm 0.04 \times 10^{-8}$ per base per generation, and a rate of $1.26 \pm 0.06 \times 10^{-9}$ for <20 bp indels. Our estimated average genome-wide mutation rate is higher than most pedigree-based estimates reported thus far, but lower than estimates obtained using substitution rates across primates. We infer the probability that a site is involved in non-crossover gene conversion as $5.99 \pm 0.69 \times 10^{-6}$, consistent with recent reports. We find that recombination does not have observable mutagenic effects after gene conversion is accounted for. We detect a strong enrichment for recent deleterious variation among mismatching variants found within IBD regions, and observe summary statistics of local IBD sharing to closely match previously proposed metrics of background selection, but find no significant effects of selection on our estimates of mutation rate. We detect no evidence for strong variation of mutation rates in a number of genomic annotations obtained from several recent studies.

← Back

LINGUISTICS OF MULTI-DOMAIN PROTEINS

Adam Seal, Vipul Periwal

National Institute of Diabetes and Digestive and Kidney Diseases, Laboratory of Biological Modeling, Bethesda, MD

Protein domains are segments of a protein sequence which have defined structural and functional characteristics. Proteins that contain more than one domain are called multi-domain proteins. Multi-domain proteins may link functions in cellular interaction networks. Bashton and Chothia provided evidence for rules underlying combinations of domains in proteins. Godzik and collaborators used graph theory to analyze such proteins, and more recently, deduced pairwise rules for domain combinations. Major trends that govern the relationships between domains are still not completely understood. Our approach looks at multi-domain proteins as a series of domains or "domain architectures". By applying techniques from computational linguistics to these domain architectures, we interpret the domains in a protein as words in a sentence, model relationships between domains as applications of rules in a link grammar, and infer the underlying stochastic grammar. We find examples in the biological literature that support the inferred rules, such as EGF and Laminin_G_2, I-set and Fn3, and CH and spectrin, among many others.

INFERRING LONG-RANGE REGULATION FROM CHROMATIN DATA

Malcolm G Perry, Boris Lenhard

Computational Regulatory Genomics, MRC Clinical Sciences Centre, Imperial College London, London, United Kingdom

Long range regulation of gene expression by distal enhancer elements is a key feature of Metazoan biology. Enhancer function is presumed to drive much of the non-coding conservation observed in Metazoa, and mutations in distal regulatory elements are increasingly recognised as important in human disease. Many methods exist for identifying enhancers, but they cannot easily be assigned to their target promoters, since chromatin looping allows enhancers to contact promoters over hundreds of kilobases away and cross other genes. Many studies simply assign enhancers to their nearest gene, although there are many known instances where this is incorrect.

We have developed a novel method for investigating long-range regulation using the correlation between histone marks and gene expression within topologically associating domains. These correlations show striking graphical patterns and allow statistical identification of the targets of long-range regulation. The inferred target genes show strong enrichment for developmental regulators, cell adhesion proteins and signalling receptors, which are categories associated with regions of high non-coding conservation known as Genomic Regulatory Blocks (GRBs). Large gene arrays, such as the protocadherin and olfactory receptor loci, are also identified by this method.

Our results show that large arrays of enhancers containing multiple genes act to regulate the expression of only one or two targets, and that the majority of genes are unresponsive to this regulation. This is consistent with current models of enhancer action but has previously been difficult to show on a genome-wide scale. Mapping these interactions also allows us to investigate the chromatin context of active enhancers, which are associated with histone variants and structural proteins.

The ability to predict domains of enhancers acting in concert alongside their targets will enable clearer interpretation of data from both GWAS and functional genomics experiments, and is an important step in moving towards a better understanding of the role of cis-regulatory elements in higher organisms.

TREES, ADMIXTURE: WHAT THE F...!

Benjamin M Peter, John Novembre

University of Chicago, Department of Human Genetics, Chicago, IL

Reich *et al.* (2009) introduced two simple admixture tests that use allele frequencies observed in three or four populations (the F3 and F4 tests). Both tests are based on simple statistics that satisfy some constraints if populations evolved according to a tree. Here, we show that these statistics have two equivalent interpretations:

Under the first interpretation, introduced by Reich *et al.*, genetic drift is modeled as a trait evolving on a phylogenetic tree. In this case, the test statistics have simple interpretations as branch lengths, and the tests have connections to elementary properties of phylogenetic trees. In particular, the three-population test can be interpreted as checking whether all branch lengths on a phylogeny are positive, and the classical maximum-likelihood (ML) test of treeness by Cavalli-Sforza and Piazza (1975) solves the same problem. When we compare the F3-test to the ML test we find in some settings that the ML test has higher power than F3.

The second interpretation is in terms of possible gene trees, where we derive expressions that allow us to examine the behavior of these admixture tests under a wide array of population genetic models. Under this model, the three-population test has an interpretation of testing whether a set of haploid individuals form a clade and are, on average, more closely related to each other than to individuals from other populations. We quantify this behavior and show under which condition the F3 test has power to detect admixture. Furthermore, we show that the four-population test may generate very high numbers of false-positives under population structure, and thus requires great care in applications.

This dual interpretation is explained because all tests are a statistic of the multidimensional site-frequency spectrum, which can be derived under both above models. Making these connections explicit serves both as a guide to interpretation for applied researchers and also highlights research avenues at the interface of population genetics and phylogenetics.

USING A SIGNAL OF EXTENDED LINEAGE SORTING TO DETECT POSITIVE SELECTION IN MODERN HUMANS SINCE THE SPLIT FROM NEANDERTAL AND DENISOVA

Stephane Peyregne[1], Christoph Theunert[1], Michael Dannemann[1], Michael Lachmann[2], Mark Stoneking[1], Kay Prüfer[1]

[1]Max Planck Institute for Evolutionary Anthropology, Genetics, Leipzig, Germany, [2]Santa Fe Institute,, Santa Fe, NM

Current methods to detect positive selection on the human lineage either rely on an unusually high proportion of functional changes compared to Great ape lineages (e.g. dn/ds) or unusual patterns of genetic diversity between individuals and populations (e.g. extended homozygosity, Tajimas D, Fst). However, the two types of tests cover very different time-frames: while the patterns of genetic diversity are most powerful during the selective sweep or shortly after, the unusual proportions of functional changes require recurrent events of selection over millions of years. Genome sequences of archaic humans (Neandertals and Denisovans) allow to investigate intermediate time-frames ranging from the split of modern humans from the archaics to the split of modern human populations.

We implemented a method to identify genomic regions that are likely to have undergone ancient selective sweeps. The method is based on a hidden Markov model that identifies regions in the genome where Neandertals and Denisovans fall outside of the present-day human variation (external regions). Regions that are unusually long are candidates for ancient selective sweeps. Through extensive simulations we test the power of the method to identify external regions, and to differentiate between positive selection and background selection. We show that our method detects older events of positive selection. Hence, we present an updated list of candidates that likely underwent positive selection on the modern human lineage since the split from Neandertal and Denisova.

A LIKELIHOOD APPROACH FOR DISTINGUISHING MUTAGENIC RECOMBINATION FROM NATURAL SELECTION ON GENOME-WIDE DIVERGENCE DATA

Tanya N Phung[1], Christian Huber[2], Kirk E Lohmueller[1,2]

[1]University of California, Los Angeles, Bioinformatics IDP, Los Angeles, CA, [2]University of California, Los Angeles, Department of Ecology and Evolutionary Biology, Los Angeles, CA

One important finding in population genetics in the 1990s was that genetic diversity positively correlates with recombination. Three possible mechanisms for this correlation include background selection (BGS), selective sweeps or mutagenic recombination. Understanding how these factors contribute to this correlation is key to establishing the amount and type of natural selection in the genome and understanding mutational processes. Though the correlation between diversity and recombination has been solidly established, evidence for a correlation between divergence and recombination has been mixed. Further, while it is thought that BGS can explain some of these patterns, a rigorous model-based comparison of these three mechanisms has not been performed. Here, we re-examine the relationship between divergence and recombination by comparing closely and distantly related species. To interpret these data, we develop a new likelihood-based framework that explicitly models both BGS and mutagenic recombination simultaneously. Our method is built upon the idea that BGS primarily affects divergence in regions of low recombination while mutagenic recombination affects regions of high recombination. We model BGS as a reduction in population size as a function of U, the deleterious mutation rate, and sh, the selective effect (Hudson and Kaplan 1995). To incorporate mutagenic recombination, we set the mutation rate of each window of the genome to $\mu_i = \mu + \varphi r_i$, where μ is the background mutation rate, r_i is the recombination rate of window i and φ is the scaling factor. Then, the composite likelihood for divergence (d) across the genome is the product of the likelihoods $L(\varphi, \mu, U, sh \mid d_i)$ for each window. We compute these likelihoods via Monte Carlo integration over coalescent genealogies that have been affected by background selection. Conditional on the genealogies, the probability of the observed divergence, including mutagenic recombination, follows a Poisson distribution. To maximize the likelihood, we search over a grid of values of φ, U, and sh. Likelihood ratio tests are used to compare models with and without background selection and mutagenic recombination. Preliminary analyses suggest that divergence between humans and distantly related species (ie. mouse) is correlated with recombination rate and that, contrary to previous intuition, background selection can contribute to this pattern. However, models also including mutagenic recombination yield a better fit, providing evidence of mutagenic recombination.

← Back

MODELING CONDITION SPECIFIC TRANSCRIPTION FACTOR BINDING WITH ATAC-SEQ

Roger Pique-Regi, Donovan Watza, Molly Estill, Francesca Luca

Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI

Deciphering the regulatory sequences which control gene transcription is a critical step in understanding both cellular and condition-specific regulatory programs encoded in the human genome. Transcriptional response is typically regulated by transcription factors (TFs) which are known to bind specific regulatory sequence motifs. Profiling across different environmental conditions the binding activity of these TFs can be quickly accomplished at a genome-wide scale with the recently developed technique ATAC-seq, which utilizes the Tn5 transposase to fragment and tag accessible DNA. When coupled with an advanced computational method such as CENTIPEDE binding models for TFs with known motifs can be generated across the genome. To date, there are no methods that efficiently incorporate the information provided by paired-end sequencing which allows both the identification of the library fragment length as well as the two cleavage locations that generated the fragment.
We have extended CENTIPEDE to utilize fragment length information to exploit the joint statistics of cleavage pairs. Our results indicate that paired-end sequencing provides a more informative footprint model for ATAC-seq libraries which leads to greater accuracy in predicting TF binding. These results were validated with ChIP-seq data (ENCODE Project) for multiple factors including CTCF, NRSF, NRF-1, and NFkB. We then assayed TF activity in lymphoblastoid cell-lines (LCLs) across multiple treatments (selenium, copper, retinoic acid, glucocorticoids and iron) for which we previously determined significant differences in gene expression levels. From our initial sequencing results we were able to resolve 383 actively bound motifs (Zscore > 5) across all conditions. We were also able to characterize 5236 regions that have significantly changed chromatin accessibility (FDR < 10%) in response to both copper and selenium. Ongoing analyses focus in integrating the changes in TF binding together with the transcriptional response. Our results demonstrate that ATAC-seq together with an improved footprint model are excellent tools for rapid profiling of transcription binding factor activity to study cellular regulatory response to the environment.

IDENTIFICATION OF CONTEXT-DEPENDENT GENE MODULES ACROSS MULTIPLE NETWORKS

Gerald Quon, Hyunghoon Cho, Bonnie Berger, Manolis Kellis

Massachusetts Institute of Technology, CSAIL, Cambridge, MA

Gene interaction networks help us understand how RNA and proteins organize themselves into functional units, pathways and modules within a set of cells and a particular context. Because of the large number of nodes and edges in these networks, as well as the complex, higher order structure typically present in these networks, an active area of research is the development of representations and models to identify gene modules and pathways in individual networks.

One of the most common types of gene interaction networks is the co-expression network, where each node represents a gene product, and an edge exists between two nodes if their expression levels co-vary across multiple biological samples. Given the wealth of gene expression profiles collected to date from e.g. The Cancer Genome Atlas and the GTEx projects, we are now able to construct co-expression networks for different tissues, under different contexts (healthy versus disease) and from different individuals. However, identification and interpretation of differences in network structure across multiple networks simultaneously is a challenging task and is not as well explored compared to the single network scenario, particularly when more than two networks are available.

We have developed a hierarchical Bayesian model named multi-network mixed-membership stochastic blockmodel (Multi-MMSB) that identifies gene modules in a manner such that transient and constitutive gene modules across networks are more easily distinguished. We first performed simulation experiments that showed Multi-MMSB to be superior at recovering gene modules compared to naive approaches that combine the networks into a single representative network before learning the modules. We then applied Multi-MMSB to a group of co-expression networks constructed from microarray profiles of asthma patients at three different stages: quiet, exacerbation, and two weeks after exacerbation. We discover a gene module associated with innate immune response and interferon signaling pathway that is only active during exacerbation, and another module associated with extracellular matrix disassembly that activates during exacerbation and remains active even after two weeks, suggesting the existence of a long-term molecular effect of an exacerbation event (p-value < 10E-4). Finally, we investigated the extent to which gene modules vary in structure across networks from different cancer types. These results support Multi-MMSB as a tool for systematically identifying differences in community structure across multiple networks.

LIKELIHOOD-BASED INFERENCE OF B CELL CLONAL FAMILIES

Duncan K Ralph, Frederick A Matsen

FHCRC, PHS, Seattle, WA

B cells develop via a Darwinian process of mutation and selection.
In order to understand these evolutionary dynamics, it is important to reconstruct the events by which the sequences came to be, that is, reconstructing the evolutionary tree for the sequences.
The first step for such work is to reconstruct which sequences came from the same naive B cells, that is, which cells form clonal families.
In this paper we describe and validate such a method, which is based on multi-hidden Markov Model (HMM) framework for B cell receptor (BCR) sequences.
Using this framework we can assign a likelihood to a collection of BCR sequences being members of a clonal family, and describe an agglomerative algorithm to find a maximum likelihood clustering.

ESTIMATING GENETIC RELATEDNESS IN A LARGE MULTI-GENERATIONAL DATASET OF 209 FAMILIES

Monica Ramstetter[1], Thomas Dyer[2], John Blangero[2], Amy Williams[1], Jason Mezey[1,3]

[1]Cornell University, Department of Biological Statistics and Computational Biology, Ithaca, NY, [2]University of Texas Health Science Center at San Antonio, Department of Genetics, San Antonio, TX, [3]Weill Cornell Medical College, Department of Genetic Medicine, New York, NY

Estimating genetic relatedness between pairs of individuals is crucial to a wide variety of analyses including genome-wide association studies (GWAS), heritability estimates, and identification of errors in pedigrees. Closely-related individuals must be removed from GWAS datasets and, more recently, from samples used to estimate heritability in order to avoid false signals and inflated heritability estimates. Additionally, identification of relatives is useful in forensics to help identify criminal suspects and to help find missing persons. Numerous methods exist for estimating relatedness between pairs of individuals, but few datasets with large pedigrees spanning multiple generations are available for analysis. Thus, researchers have assessed the performance of relationship inference methods using simulated genetic data that may not accurately mimic the genetic and haplotypic diversity or the patterns of relatedness in real data. Using data from the San Antonio Family Studies (SAFS), we explore the performance of several existing methods and heuristics which can be used to infer relatedness between pairs of individuals. The SAFS data consists of Illumina genotypes (660k and 1M arrays) for 2490 individuals in 209 families that span up to four generations. The evaluation includes identity-by-descent (IBD) detection methods (GERMLINE, FastIBD, Refined IBD, IBDLD, ISCA, and diCal-IBD) paired with heuristic methods which turn the IBD information into an estimate of degree of relatedness and type of relationship (parent-child, avuncular, etc.). We also examine standalone methods that estimate the degree of relatedness and/or type of relationship between individuals (ERSA, SNPduo, REAP, KING, RelateAdmix, Plink, kcoeff, and others). To assess performance, we consider power, precision, and recall for classifying the relationships in the SAFS dataset. We assume that the reported relationships in this curated dataset are true except for a few cases where several methods independently confirm a probable error. Our analysis also considers runtime, as many methods require several days to complete their analysis whereas others can finish in less than 5 minutes. Finally, we present possible modifications to improve upon existing methods, including the application of clustering methods to estimates of relatedness as well as Bayesian approaches.

IDENTIFYING CANDIDATE DRUG TARGETS : A NESTED ANOVA MODEL FOR QUANTIFYING TISSUE SPECIFICITY AND REGULATORY DIVERGENCE ACROSS SPECIES

Andrew Torck[1], Jiyoung Kim[2], Michael Q Zhang[3,4], Gregory Dussor[1], Theodore Price[1], Pradipta Ray[3]

[1]The University of Texas at Dallas, Brain and Behavioral Sciences, Richardson, TX, [2]University of Arizona, Pharmacology, Tucson, AZ, [3]The University of Texas at Dallas, Center for Systems Biology, Biological Sciences, Richardson, TX, [4]Tsinghua University, Synthetic and Systems Biology, Beijing, China

High-throughput pharmacological approaches to drug discovery typically involve identification of tissue specific gene transcripts (using information theoretic scores) in model species as putative low side-effect mRNA or protein targets . Divergent transcriptome evolution between humans and model species can cause false positives leading to significant wasted resources, dealt with by additionally profiling transcriptome of corresponding postmortem human tissue(s). Simple models of estimating transcriptome evolution in this framework involve establishment of a meaningful "testbench" of tissues/conditions in human and model species, and a notion of per-gene distance across species (often based on a simplistic correlation). However, there is no way to include multiple biological replicates / multiple model species (eg mouse+rat) into the analysis as safeguard against false negatives. Relative importance of various tissues in the testbench or missing data for tissue-species pairs also cannot be easily incorporated into the model.

We propose a nested ANOVA to characterize per-gene the "spread" of expression levels along each dimension (tissue, species). We perform hierarchical clustering on suitably normalized expression data along tissue and species dimensions separately, using a nested ANOVA model. Genes can then be clustered based on all the variance components at each level of the nested ANOVA across both tissue and species dimension, and clusters of genes with variance components consistent with a promising tissue-specific drug target (pharmacological constraints : tissue specificity identified by variance components along tissue dimension, conservation of gene expression between species identified by variance components along species dimension, tissue prioritization performed by differential constraints on variance in different tissues) can be probed further by studying regulatory evolution at promoters/enhancers. This approach scales to introduction of multiple biological replicates (for estimating regulatory plasticity), or of additional species, by simply requiring additional nested ANOVA steps. It can allow for missing data in some tissue-species pairs by making clustering distance function agnostic to specific tissue/species. Estimated variance components along species dimension help accurately gauge species divergence, without a restrictive linear correlation framework.

We apply this model to drug target discovery in chronic pain, using RNA-seq from mouse/rat nerve-injury models of dorsal root ganglia (DRG), and in-house RNA-seq of DRG in postmortem humans (with/without reported chronic pain). We integrate these with public RNA-seq data for other neural tissue, and tissues with potential side effects in pain therapy, to identify a set of candidate genes on which we will perform additional validation (schema at http://www.utdallas.edu/~prr105020/pmg.pdf)

THE IMMUTABLE METHYLOME : CHARACTERIZING REGIONS OF INVARIANT METHYLATION IN THE MAMMALIAN GENOME AS A COUNTERFOIL TO DISCOVERING TISSUE SPECIFIC METHYLATION PATTERNS

Pradipta Ray[1], Milos Pavlovic[1], Michael Q Zhang[1,2]

[1]The University of Texas at Dallas, Center for Systems Biology, Biological Sciences, Richardson, TX, [2]Tsinghua University, Synthetic and Systems Biology, Beijing, China

Recent works [Zhang et al,Gen Bio 2015; Yan et al, Sci Rep 2015] have introduced models for predicting mammalian whole-methylome CpG methylation (5-methylcytosine) status at single CpG site resolution with high >90% accuracy, based on various correlated genomic/epigenomic traits. This is significant improvement over state-of-the-art from recent years, at ~85% accuracy at CpG island/DNA fragment resolution [Das et al,PNAS 2006; Bock et al, PLoS Genetics 2006], using genomic traits as input. The trade-off involved pertains to a larger diversity (typically epigenomic) of input features: constraining such predictors to be applicable only to a limited number of resource-rich tissues/cell types where such datasets exist, requiring careful data standardization/design to avoid algorithmic artifacts/overfitting. Since DNA methylation is predictable using genomic features alone, we tested hypothesis that a significant fraction of the mammalian methylome is stably maintained across different cell types. Characterization of these regions provide a simple way to predict a significant part of the methylome, helping identify CpG sites that are condition/tissue-specifically methylated.

Based on 25 tissue/cell-type methylomes from Epigenome Roadmap consortium, we identified over 1/3 of CpG sites in the human methylome to be consistently methylated or demethylated across all cell types (eg. certain transposable elements are reliably methylated, promoters of housekeeping genes are reliably demethylated). Our method was conservative since low coverage sites for bisulfite-seq experiments were filtered out. Being cell-type invariant, these regions of invariant methylation are expected to have correlative (potentially causal) genomic features. We built a predictive max-margin model for classifying invariant vs cell type specific CpG sites. Our model helps identify species-specific universal spatially-contiguous methylation patterns. We interrogated these regions studying perturbations that change the underlying genome: DNA evolution between mouse and human - characterizing how such methylation invariant regions vary across species, and cancer methylomes in humans - characterizing how aberrant methylation patterns in cancer include epigenetic modifications in stably methylated regions. We characterized CpG 5-hydroxymethylcytosine (typical intermediate of the demethylation pathway) modifications contrasting its abundance in regions of invariant versus regions of variable methylation.

← Back

AN ALGORITHM TO IDENTIFY HIERARCHIES OF SIGNIFICANTLY MUTATED SUBNETWORKS IN CANCER

Mark D Leiserson[1,2], Matthew Reyna[1,2], Ben J Raphael[1,2]

[1]Brown University, Computer Science, Providence, RI, [2]Brown University, Center for Computational Molecular Biology, Providence, RI

A key challenge in cancer genomics is to distinguish the *driver* mutations that cause cancer from random, *passenger* mutations. Several statistical approaches have been introduced to predict genomic positions or genes that are drivers according to their frequency of mutation across samples. However, such approaches are challenged by the long tail of rarely mutated genes. Since driver mutations target cellular signaling and regulatory pathways, an alternative approach is to examine combinations of mutations in these pathways.

We present an algorithm to identify significant subnetworks of mutated genes in a biological interaction network, where significant subnetworks are defined *both* by the mutations within the genes in the subnetwork and by the local topology of interactions between genes. Our algorithm uses a heat diffusion process to combine both features, extending our earlier HotNet2 algorithm. The key innovations in the new algorithm are the derivation of a hierarchy of subnetworks across a range of thresholds and a statistical test to compare hierarchies under the null hypothesis that scores on genes are independent of their location on the network. This hierarchical approach better represents the relationships between subnetworks compared to the choice of a single threshold.

The resulting algorithm is generally applicable across different gene scores and networks. We demonstrate an application of the algorithm to analyze a Pan-Cancer dataset of somatic mutation data from The Cancer Genome Atlas (TCGA) using several protein-protein interaction networks. The algorithm identified groups of genes overlapping known cancer signaling pathways, as well as novel groups of interacting genes. In comparison to a similar analysis with the HotNet2 algorithm, our new algorithm reveals additional genes and subnetworks in the hierarchy. We anticipate that our algorithm will be useful for analyzing additional datasets of germline and somatic variants, using a variety of other interaction networks such as tissue-specific networks or regulatory networks.

A GENETICAL GENOMICS APPROACH TO IDENTIFY CUANTITATIVE EXPRESION LOCI INVOLVED IN THE DEFENSE RESPONSE TO A FUNGAL PATHOGEN IN SUNFLOWER

Maximo L Rivarola[1,3], Federico Ehrenbolger[1,3], Carla Filippi[1,3], Jeremias Zubrzycki[1,3], Julio Di Rienzo[2], Paula Fernandez[1,3], Sergio Gonzlez[3], Carla Maringolo[1,3], Facundo Quiroz[1,3], Diego Cordez[1,3], Diego Alvarez[1,3], Alejandro Escande[1,3], Esteban Hopp[3], Ruth Heinz[1,3], Veronica Lia[1,3], Norma Paniego[1,3]

[1]Instituto Nacional de Tecnologia Agropecuaria, Biotecnology, Castelar, Argentina, [2]Universidad Nacional Cordoba, Agrarias, Cordoba, Argentina, [3]Consejo Nacional de Investigaciones Científicas y Técnicas, Ciencias Naturales, Buenos Aires, Argentina

Defense response to the fungal pathogen *Sclerotinia sclerotiorum* in plants is a complex trait with many layers of regulation. We are taking a genetical genomics approach on an association mapping set of sunflower recombinant inbred lines (RILs) showing phenotypical diversity (resistant or susceptible) in response to this fungal attack. On one hand, we performed microarray analysis on our previously designed chip and quantitative de-novo RNA-seq at three different times after fungal infection. On the latest dataset, we evaluated different assembly algorithms and strategies to create a transcriptomic reference for the digital counting task, along with its gene ontology (GO) term annotation. Statistical analysis of microarray data using a linear mixed model and a generalized linear model (GLM) for our RNA-seq data was performed so as to evaluate contrasts in a more efficient manner taking into account genotype, time, lane of sequencer, and infection. This comprehensive strategy, still in progress, led us to detect a collection of several dozens of genes with differential expression and dozens of GO terms enriched in infected sunflower florets at different time points. As we expected, if not for the randomized block design and GLM modeling which enabled us to distinguish the fungal effect, most of the up or down regulated loci are strongly associated to the phenology state or the genetic background of the inbred line. To validate these results, 58 of these candidate genes were assayed by quantitative real-time PCR, of these 19 were validated. On the other hand, QTL mapping was carried out using linkage (123 RILs) and association (135 RILs) mapping from the INTA genetic breeding program. Hence we identified 13 candidate loci with significant contribution to the fungal resistance response and 32 major defense QTL. It is our interest to enhance the integration of the approaches described here combining genetic and genomic resources, in particular through functional enrichment, SNP discovery, and future eQTL analysis, to reach interesting leads on significant loci to fungal disease resistance in sunflower.

← Back

DECIPHERING MUTATIONAL SIGNATURES IN CANCER WITH THE HIERARCHICAL DIRICHLET PROCESS

Nicola D Roberts, Peter J Campbell

Wellcome Trust Sanger Institute, Cancer Genome Project, Hinxton, United Kingdom

Somatic mutations stem from a variety of underlying processes, including replication error, exposure to chemical or physical mutagens, and defective DNA repair. Each process results in a characteristic distribution or 'signature' of mutation classes, and these signatures can be deciphered from patterns in somatic mutation catalogs of cancer genomes. Previous studies have relied upon non-negative matrix factorization to identify mutational signatures, but this method lacks a probabilistic framework for modeling relationships between samples and calculating formal statistics. These limitations are overcome with a non-parametric Bayesian approach using the hierarchical Dirichlet process (HDP).

The HDP method for mutational signatures analysis defines a flexible tree of Dirichlet processes to reflect the relationships between different samples. Samples may be grouped by any number of pertinent factors, such as cancer type, germline genotype, mutagen exposure, or patient of origin (may have multiple metastases or subclones from the same individual). The posterior sampling process borrows information across samples and groups to identify shared signatures, while simultaneously quantifying differences between groups.

Here, I illustrate the utility of the HDP method with a pan-cancer analysis of the TCGA catalog, an intra-patient analysis of prostate cancer metastases and subclones, and demonstrate how to simultaneously discover new signatures and match data to known signatures by conditioning on a previous dataset. The mutational signatures identified include those associated with UV radiation, smoking, APOBEC activity, and POLE mutations. All methods presented are available as an open-source R package here: https://github.com/nicolaroberts/hdp

NAIVE BAYES CLASSIFICATION FOR IDENTIFYING GENOMIC SITES UNDER NEUTRAL EVOLUTION, HARD SWEEPS, AND SOFT SWEEPS

Stephen Rong[1,2], Lauren Alpert[1,2], Sohini Ramachandran[1,2]

[1]Brown University, Department of Ecology and Evolutionary Biology, Providence, RI, [2]Brown University, Center for Computational Biology, Providence, RI

Supervised-learning methods based on multiple summary statistics and simulated training data have been developed to detect signals of hard selective sweeps (i.e. positive selection from *de novo* mutations) in genomic data. However, it has been argued that other types of selection — such as soft selective sweeps (i.e. positive selection from standing variation), negative selection, balancing selection, and polygenic selection — may be just as or more pervasive than hard sweeps in shaping observed patterns of genetic variation in multiple species. Previously, we developed a novel Naive Bayes framework for classification of single-nucleotide polymorphisms (SNPs) in population genomic data as neutral sites or hard selective sweeps, which has the advantages over similar methods of giving probabilistically interpretable results and accounting for dependencies among various statistics for hard selective sweeps. Here, we extend our Naive Bayes framework to classification between multiple types of selection. We present preliminary results for multi-level classification of SNPs as neutral sites, hard selective sweeps, or soft selective sweeps, focusing on the question of whether the three classes are separable using our Naive Bayes framework and how the partitioning of the three classes depends on the set of summary statistics used.

INFERENCE OF CLONAL GENOTYPES FROM SINGLE CELL SEQUENCING DATA

Andrew Roth[1,2], Andrew McPherson[1,3], Alexandre Bouchard-Côté[4], Sohrab Shah[1,5,6]

[1]BC Cancer Agency, Department of Molecular Oncology, Vancouver, Canada, [2]University of British Columbia, Graduate Bioinformatics Training Program, Vancouver, Canada, [3]Simon Fraser University, School of Computing Science, Burnaby, Canada, [4]University of British Columbia, Department of Statistics, Vancouver, Canada, [5]BC Cancer Agency, Michael Smith Genome Sciences Centre, Vancouver, Canada, [6]University of British Columbia, Department of Pathology and Laboratory Medicine, Vancouver, Canada

Single cell sequencing is emerging as a powerful method to study the clonal population structure of tumours. Advances in cell isolation coupled with improvements in the the efficiency of DNA amplification from single cells have made high throughput profiling of single cell genomes a feasible assay for identifying genomic clones and studying their evolution.

Current protocols for single cell sequencing generate imperfect data with missing values, highly biased allelic counts at heterozygous loci, and false measurements of genotypes due to sequencing multiple cells. Because of these errors, measurements from a single cell do not accurately represent a clonal genotype. We suggest that aggregating measurements from multiple cells from the same clonal population would improve inference of clonal genotypes. To do this we need to know the clonal population of origin for each cell, which is of course not possible without knowing the genotype composition of the population. Thus we have the problem of jointly inferring clonal genotypes and assigning cells to clonal populations.

We have developed a novel statistical model and a variational Bayes inference method to address this problem. Using allelic measurements from a fixed set of loci across a population of single cells, the model solves this joint inference problem, while also estimating the unknown number of clones. Our model is able to handle missing data and allelic dropout in a principled way. In addition, we are able to identify measurement which are the result of sequencing multiple cells. Furthermore, the model is more flexible than previous approaches, allowing input data from multiple discrete data-types with an arbitrary number of states.

We introduce a novel hybrid strategy to generate realistic synthetic datasets for benchmarking purposes. Using this ground truth data we show that our model outperforms alternative methods with respect to both clustering and clonal genotype prediction accuracy.

We demonstrate the utility of our model with three real world data-sets: six childhood acute lymphoblastic leukemia tumours, a multiply passaged xenograft series and three multiply sampled high grade serous ovarian cancer (HGSOC) tumours.

← Back

A BAYESIAN HIERARCHICAL SPARSE FACTOR MODEL FOR COMPLEX EXPERIMENTS IN GENETICAL GENOMICS

Daniel Runcie[1], Sayan Mukherjee[2], RJ Cody Markelz[3]

[1]University of California Davis, Plant Sciences, Davis, CA, [2]Duke University, Statistical Science, Durham, NC, [3]University of California Davis, Plant Biology, Davis, CA

Genome-wide gene expression data can provide rapid insight into mechanisms underlying responses to genetic variants or environmental perturbations. Gene expression datasets generated from complex multi-factor experiments are increasingly common in medicine, evolutionary biology, and agriculture. Robust models for identifying sets of genes associated with particular factors in experiments with correlated samples are lacking. Here, we propose a Bayesian model that aims to uncover gene expression signatures of the response to particular experimental treatments, such as plant density, in experiments with design features like sub samples, split plots or repeated measures, or with the experimental units sharing an arbitrary covariance such as from a pedigree or a structured population. We assume that genes function within semi-independent modules (factors), and that these factors are latent traits that vary according to the experimental treatments, genetic backgrounds, and any additional micro-environmental variation. This implies a factor structure for the gene expression covariation that is highly structured – both across genes and among samples – and that can be characterized by a relatively low number of parameters, which we enforce with biologically-motivated sparsity-inducing priors. To account for sub-samples or other pseudo-replication in the experimental design, we employ an efficient hierarchical mixed effect model simultaneously for the observed gene expression traits and the underlying latent traits. The advantages of this approach are two-fold. First, the mixed effect model permits modeling of the raw data, rather than on means of subsamples, and so can leverage the among gene correlation structure. Second, the factors themselves can be explored to provide biological intuition into the mechanisms driving biological responses to the experimental factors by inspecting functional or pathway classifications of genes in the modules. We demonstrate our approach on a large RNAseq dataset from *Brassica rapa*.

EFFECTS OF ADAPTIVE NEANDERTAL INTROGRESSION AT THE OAS LOCUS ON THE MODERN HUMAN INNATE IMMUNE RESPONSE

Aaron J Sams[1], Yohann Nedelec[2,3], Anne Dumain[2], Vania Yotova[2], Philipp W Messer[1], Luis B Barreiro[2,4]

[1]Cornell University, Department of Biological Statistics & Computational Biology, Ithaca, NY, [2]CHU Sainte-Justine Research Center, Department of Genetics, Montreal, Canada, [3]University of Montreal, Department of Biochemistry, Montreal, Canada, [4]University of Montreal, Department of Pediatrics, Montreal, Canada

It is now clear the ancestry of all individuals living outside of sub-Saharan Africa is composed of roughly two percent Neandertal ancestry. Yet, it remains largely unclear to what extent this contribution from Neandertals impacts modern human biology, and further, to what extent it may have provided adaptive genetic variation to modern human populations. The immune system is one physiological system that harbors higher than typical amounts of genetic variation in order to provide a flexible set of responses to infection. Here we use coalescent simulation and population genetic approaches to demonstrate a signal of adaptive introgression in the 2'-5' oligoadenylate synthetase (OAS) gene cluster region of chromosome 12. The adaptive region encodes for three active OAS enzymes (OAS1-3) that are involved in the innate immune response to viral infection. In order to evaluate the functional consequences of the adaptive haplotype we infected primary macrophages and peripheral blood mononuclear cells from people with and without the Neandertal haplotype with a panel of viruses and viral-syntetic ligands. Our results show that people with the Neandertal-like haplotype show marked functional differences in the transcriptional regulation of OAS1 and OAS3 in response to virtually all viral agents tested, which illuminate the phenotypic effects of Neandertal haplotypes into the regulation of innate immune responses in modern human populations.

MANTA: RAPID DETECTION OF STRUCTURAL VARIANTS AND INDELS FOR CLINICAL SEQUENCING APPLICATIONS

Xiaoyu Chen[1], Ole Schulz-Trieglaff[2], Richard Shaw[2], Bret Barnes[1], Felix Schlesinger[1], Anthony J Cox[2], Semyon Kruglyak[1], Christopher T Saunders[1]

[1]Illumina, Inc, Bioinformatics, San Diego, CA, [2]Illumina Cambridge Ltd, Bioinformatics, Little Chesterford, United Kingdom

We describe Manta, a method to discover structural variants and indels from next generation sequencing data. Manta is optimized for rapid clinical analysis, calling structural variants, medium-sized indels and large insertions on standard compute hardware in less than a tenth of the time that comparable methods require to identify only subsets of these variant types: for example NA12878 at 50x genomic coverage is analyzed in less than 20 minutes. Manta can discover and score variants based on supporting paired and split-read evidence, with scoring models optimized for germline analysis of diploid individuals and somatic analysis of tumor-normal sample pairs. Call quality is similar to or better than comparable methods, as determined by pedigree consistency of germline calls and comparison of somatic calls to COSMIC database variants. Manta consistently assembles a higher fraction of its calls to basepair resolution, allowing for improved downstream annotation and analysis of clinical significance. We provide Manta as a community resource to facilitate practical and routine structural variant analysis in clinical and research sequencing scenarios.

Manta source code and Linux binaries are available from http://github.com/Illumina/manta

SPECIES TREE INFERENCE WITH POLYMORPHISM-AWARE PHYLOGENETIC MODELS

Dominik Schrempf[1,2], Bui Quang Minh[3], Nicola De Maio[4], Arndt von Haeseler[3,5], Carolin Kosiol[1]

[1]Institut für Populationsgenetik, Vetmeduni, Vienna, Austria, [2]Vienna Graduate School of Population Genetics, Vetmeduni, Vienna, Austria, [3]Center for Integrative Bioinformatics, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Vienna, Austria, [4]Nuffield Department of Medicine, Oxford University, Oxford, United Kingdom, [5]Bioinformatics and Computational Biology, University of Vienna, Vienna, Austria

The availability of genome-scale inter- and intraspecies data leads to new opportunities in phylogenetics to improve tree accuracy and resolution as well as to take important steps towards understanding the process of speciation.

We present a novel maximum likelihood implementation of a Polymorphism-Aware Phylogenetic Model (PoMo, De Maio et al., MBE 2013) that can do both, parameter estimation and species tree inference for genome-wide data of a moderate number of species while still allowing for many individuals per species. It 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. Thereby, a Moran process is used to model genetic drift. Although a single phylogeny — the species tree — is considered, PoMo naturally accounts for incomplete lineage sorting because ancestral populations can be in a polymorphic state.

A large scale simulation study with four different scenarios for small and large trees (incomplete lineage sorting, anomaly zone, recent radiation and trichotomy) as well as applications to great ape data (12 populations in total, Prado-Martinez et al., 2013) show that PoMo is fast while being more accurate than other state-of-the-art methods (De Maio, Schrempf, and Kosiol, Syst. Biol. 2015). PoMo also efficiently and accurately estimates evolutionary parameters relevant for GC variation along mammalian genomes from exome-wide alignments of four great ape species.

Recently, we have derived a reversible version of PoMo and implemented it into IQ-Tree (Nguyen et al., 2015), an efficient and easy-to-use software package. We observe a reduction of runtime by a factor of 50 and no loss of accuracy. This demonstrates that PoMo is suitable to infer large scale phylogenies from population data.

MIXED GRAPHICAL MODELS FOR ANALYSIS OF MULTI-MODAL GENOMIC AND CLINICAL VARIABLES

Andrew J Sedgewick[1,2], Joseph D Ramsey[3], Peter Spirtes[3], Clark Glymour[3], Panayiotis V Benos[1]

[1]University of Pittsburgh, Computational and Systems Biology, Pittsburgh, PA, [2]Joint Carnegie Mellon-University of Pittsburgh PhD Program in Computational Biology,, Pittsburgh, PA, [3]Carnegie Mellon University, Philosophy, Pittsburgh, PA

Graphical models are an important tool for biomedical research because they can intuitively represent the underlying structure of complex, multivariate probability distributions found in biological data. Learned models can be used for classification, biomarker selection, and functional analysis. These models are often designed to handle only one type of data, however, and this limits their applicability to a large class of biological datasets with both continuous and discrete variables. To address this issue, we developed new methods for directed network recovery over mixed discrete and continuous data, and compared them to existing causal discovery algorithms. Our method first learns an undirected mixed graphical model (MGM) superstructure and then uses that as a starting point for PC-Stable and GES, two well-known causal network search algorithms. We tested our methods on simulated datasets of continuous and categorical variables generated from a variety of network structures which included both cycles and scale-free topologies. Our methods, MGM-PCS and MGM-GES outperformed existing methods for both overall and directed edge recovery on these simulated data. When applied breast cancer data from The Cancer Genome Atlas (TCGA), our methods recovered relevant connections between RNA-seq variables and clinical variables for hormone receptor status and PAM50 subtype label.

THE PROBLEM OF AND SOLUTION TO ACCURATELY ASSESS HIGHLY POLYMORPHIC REGIONS ON HTS RELATED STUDIES.

Fritz J Sedlazeck[1,2], Naoki Osada[3], Michael C Schatz[1], Arndt von Haeseler[2]

[1]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, [2]Max F Perutz Laboratories, Center for Integrative Bioinformatics Vienna, Vienna, Austria, [3]National Institute of Genetics, Mishima, Japan

The advent of high throughput sequencing (HTS) has boosted the variety of sequencing projects related to molecular biology and medicine. Mapping reads to a reference genome is one of the fundamental steps in HTS related analysis, including for variant identification, transcript abundance estimation, and many others. However, mapping reads from a heterogeneous sample to a reference genome can lead to biased results and inability to identify the alternate alleles i.e. “reference mapping bias”. We studied this bias in several of the most widely used read mapping algorithms including BWA/BWA-MEM, Bowtie/Bowtie2, TopHat/TopHat2, and our own NextGenMap, in an F1 cross from inbred lines of D. melanogaster Mel 6 x D. melanogaster RAL774 where a precise catalog of heterozygous positions could be determined from the parental reference genomes. By examining regions with different rates of heterozygosity, we show that both SNP calling and transcript abundance analysis were highly skewed against the alternate alleles proportional to the frequency of heterozygosity, including completely mis-analyzing certain regions as having allele-specific expression. In contrast, we show that the highly sensitive mapper like NextGenMap are less affected by reference bias and thus more suitable for reliable analyses of HTS data for polymorphic samples.

PHASING FOR MEDICAL SEQUENCING USING RARE VARIANTS AND LARGE HAPLOTYPE REFERENCE PANELS

Kevin Sharp[1], Olivier Delaneau[2], Jonathan Marchini[1,3]

[1]University of Oxford, Dept. of Statistics, Oxford, United Kingdom, [2]University of Geneva, Département de Génétique et Développement, Geneva, Switzerland, [3]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom

There is growing recognition of the importance, in clinical settings, of accurate estimation of haplotypes from high coverage sequencing of single samples. Although such haplotypes can be estimated as an imperfect mosaic of high quality haplotypes (copying states) from a reference panel, they include large numbers of low frequency variants that are hard to phase. Consequently, selection of the most informative set of copying states in a computationally tractable way is becoming a fundamental problem for the field.
Despite their challenges, rare variants can also be phase informative: sharing of such rare variants between two individuals is more likely to arise from a recent common ancestor and, hence, also more likely to indicate similar shared haplotypes. Our method exploits this idea to select a small set of highly informative copying states. When combined with a widely used Hidden Markov Model phasing algorithm (SHAPEIT2), we obtain significant gains in phasing accuracy over the current selection approach, as well as a significant improvement in speed.
We tested our method by phasing two regions of chromosome 20, comprising 48MB, for each of two high coverage (130x) trio parents of European ancestry using a reference panel comprising 7,510 UK10K haplotypes. Our method of choosing states consistently improved accuracy compared to the current version of SHAPEIT2. Averaged over individuals and 20 different runs, our method achieved a 14.1% improvement in switch error rate (SER) over SHAPEIT2, but required no MCMC iterations. When SHAPEIT2 was also run without MCMC iterations, the improvement was 39.4%.
To obtain further improvements, we used our approach to perform a single step rephasing of the UK10K panel. This increased the improvement in SER over SHAPEIT2 both with and without MCMC iterations to 26.4% and 48.1% respectively.
Finally, our new version of SHAPEIT permits the rare variant haplotype selection method to be used in combination with phase-informative paired end reads. Combining these sources of information led to a total reduction in SER over baseline SHAPEIT2 (without MCMC) of 52.5%.
Our single step rephasing of chromosome 20 for the UK10K panel took 15.1 hours using 12 CPU cores. While phasing a panel from scratch would require an iterative extension to our approach, these results represent a proof of concept that rare variant sharing patterns can be utilised to phase the entire 100,000 Genomes Project dataset.

EXPLAINING MISSING HERITABILITY USING GAUSSIAN PROCESS REGRESSION

Kevin Sharp[1,2], Wim Wiegerinck[1], Alejandro Arias-Vasquez[2,3], Barbara Franke [2,3], Cornelis A Albers[2,3], Hilbert J Kappen[1]

[1]Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands, [2]Radboud University Medical Centre, Department of Human Genetics, Nijmegen, Netherlands, [3]Radboud University Medical Centre, Department of Psychiatry, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands

For many traits and common human diseases, causal loci uncovered by genetic association studies account for little of the known heritable variation. 'Missing heritability' might lie in the effect of non-additive interactions between multiple loci, but this has been difficult to test using existing parametric approaches. We employed a non-parametric, Bayesian method, based on Gaussian Process Regression, for identifying associated loci in the presence of interactions of arbitrary order. On both simulated and real datasets we demonstrate that the method has considerable power to detect high-order interactions and explain missing heritability.

As a proof of principle, we analysed 46 quantitative yeast phenotypes. Whereas detected pairwise QTL-QTL interactions accounted for little of the variance (a median of 3% per trait), we found that over 70% of the total known missing heritability could be explained using common genetic variants, many without significant marginal effects. Interestingly, the availability of biological replicates significantly improved the power to identify such loci and, hence, to explain variance.

These results already represent a significant advance in approaches to understanding the missing heritability problem with potentially important implications for studies of complex, quantitative traits. Importantly, however, features of the algorithm can be exploited to permit application to datasets incorporating much larger numbers of putative QTLs. In particular, the most computationally expensive steps consist of large numbers of independent computations that conform to the SIMT parallel computation model of Graphics Processing Units (GPUs). We also describe such an implementation. Using a single Nvidia Tesla M2090 GPU we already achieve two orders of magnitude improvement in wall-clock time per iteration over a serial implementation. This indicates the potential of the approach for application to other model organisms and human datasets.

LEARNING CONTEXT-SPECIFIC REGULATORY LEXICONS USING WORD EMBEDDINGS

Avanti Shrikumar[1], Rahul Mohan[2], Johnny Israeli[3], Anshul Kundaje[1,4]

[1]Stanford University, Dept. of Computer Science, Stanford, CA, [2]Bellarmine College Preparatory School, Dept. of Computer Science, San Jose, CA, [3]Stanford University, Biophysics Program, Stanford, CA, [4]Stanford University, Dept. of Genetics, Stanford, CA

Combinatorial "grammars" of transcription factor (TF) binding sites encoded in the genomic sequence of regulatory elements are a primary determinant of their context-specific regulatory activity. These sequence grammars consist of specific combinations of subsequences exhibiting complex co-occurrence density distributions as well as spatial and positional relationships. The most commonly used representation for genomic sequence is the bag-of-words model that is incapable of encoding such complex grammars. Hence, machine learning methods that use these naive representations of regulatory sequence for predicting tissue-specificity of regulatory elements are likely to underperform. Recently, Recurrent Neural Networks (RNNs) have achieved impressive performance on analogous learning tasks in Natural Language Processing (NLP) that require learning grammatical relationships between words. An important ingredient in the success of RNNs is the development of "word embeddings" which are rich representations capable to capturing complex lexical relationships. Word embeddings represent words as vectors such that semantically similar words are close in the resulting vector space. Inspired by the success of word embeddings in NLP, we developed two types of "word embeddings" for genomics: the first is an embedding of k-mers for learning tasks that use raw sequence as input, and the second is an embedding of in-vivo TF binding events for tasks that use TF ChIP-seq binding data as input. We compared several embedding techniques ranging from SVD on a co-occurrence matrix to GloVe and word2vec. We demonstrate that, when used with a random forest to predict the tissue-specific activity of regulatory elements, the best k-mer embeddings perform better than a bag-of-kmers or bag-of-motifs representation while providing over a 40x reduction in dimensionality. We also find that an RNN trained using the k-mer embeddings successfully learns regulatory grammars on simulated data. We further show that an RNN trained using TF embeddings to predict the tissue-specificity of regulatory elements performs better than approaches that do not leverage the embeddings. Finally, by visualising the embedded space and the activity of nodes in the RNN, we are able to gain insights into the learned regulatory grammars.

PROBABILISTIC MODELS OF SINGLESAMPLE AND MULTISAMPLE VARIANT CALLING

Suyash S Shringarpure[1], Armin Pourshafeie[2], Carlos D Bustamante[1]

[1]Stanford University, Genetics, Stanford, CA, [2]Stanford University, Physics, Stanford, CA

Variant calling is an important step in the analysis of NGS data, since it affects the results of downstream analyses. Singlesample calling and multisample calling are two main methods for SNP discovery from genomes of many individuals. While singlesample calling is efficient and parallelizable, it is limited in extracting information from only a single sample. This has been seen to decrease its performance at SNP discovery, particularly with low-coverage sequence data. Multisample calling was developed to address this shortcoming by sharing information across many samples using Hardy-Weinberg equilibrium. These methods have been extensively used to discover variants in human and non-human genome sequences. While many empirical observations of the characteristics of these models exists, there are no models that describe the general behavior of these models with varying coverage, sample size and minor allele frequency (MAF).

We developed Bayesian models of singlesample and multisample SNP calling that characterize the performance of these methods. We use our models to show how the performance of these methods depends on factors like coverage, sample size and MAF. Using the SNP quality phred score as a metric, we show that multisample quality score has a more direct dependence on SNP MAF than the singlesample quality score. We use our model to explain some previously reported empirical observations about the performance of these models:

- For SNP discovery in low-coverage data, multisample calling has higher power than singlesample calling.
- For SNP discovery in high-coverage data, both methods have high power.
- In large samples multisample calling has less power than singlesample calling for identifying variants with very low frequency (MAF < 0.001).

Our models suggest that while multisample calling is often better than singlesample calling for SNP discovery, this is not always true, particularly for very rare variants in large samples. We use our models to provide guidelines on scenarios under which one method is expected to perform better than the other. We also use our models to show how prior information about variants from reference panels such as 1000 Genomes can be used to improve the performance of singlesample SNP calling.

MODELING LINKAGE DISEQUILIBRIUM AND MUTATION TO ESTIMATE TIME TO THE COMMON ANCESTOR FOR A BENEFICIAL ALLELE

Joel Smith[1], Matthew Stephens[2], John Novembre[3]

[1]University of Chicago, Ecology and Evolution, Chicago, IL, [2]University of Chicago, Human Genetics and Statistics, Chicago, IL, [3]University of Chicago, Human Genetics, Chicago, IL

A key principle for allele age estimation is that due to recombination and mutation, the ancestral haplotype of the focal allele decays at some rate every generation. We provide a method to exploit this process and jointly infer the time to the common ancestor (TMRCA) of a postively selected allele as well as its ancestral haplotype following a selective sweep. We do so using a hidden Markov model, allowing us to sum over uncertainty in recombination events off of the ancestral haplotype. This framework uniquely models the accumulation of derived mutations on the ancestral haplotype, the length distribution of the ancestral haplotype, and background haplotype diversity to infer the time to the common ancestor. For parameter values typical of human populations, our method provides root mean squared error values of ~80 generations for TMRCA values up to 1000 generations gold. Applications from the 1000 Genomes human data include Eurasian alleles associated with lactase persistence and skin pigmentation.

IMPROVED D-STATISTIC FOR LOW-COVERAGE DATA AND ADMIXTURE GRAPHS.

Samuele Soraggi[1], Carsten H Wiuf[1], Anders Albrechtsen[2]

[1]Copenhagen University, Department of Mathematics, Copenhagen, Denmark, [2]Copenhagen University, The Bioinformatics Centre, Copenhagen, Denmark

Next Generation Sequencing provides a massive quantity of data, and it is widely analyzed in population genetics research. Since many NGS datasets are sequenced at low coverage, SNP and genotype calling have high uncertainty. This might affect statistical methods that make use of genotypes. A commonly used tool for detecting ancient admixture events is the D-statistic.

The D-statistic makes use of the patterns between alleles in four different groups of individuals in order to show the direction of a gene flow or to assess the correctness of a phylogeny of four populations in the configuration (H1(H2(H3,H4))), where H1 is an outgroup. For low-depth sequencing the D-statistic is highly susceptible to errors deriving from the SNP and genotype calling. For low-depth sequencing where genotype calling is not possible the method relies on sampling one allele from a single individual from each group at each site to evaluate ABBA and BABA patterns. This sampling procedure ignores much of the information in the data and only works for one individual from each group. Moreover, the D-statistic does not allow to infer information on more complex phylogenies, for example having multiple admixture events or more than four groups of individuals.

We have implemented in ANGSD a version of the D-statistic that does not require genotype calling but is still able to utilize all the information and allows for multiple individuals for each of the four groups. Using real data we evaluate the method's power and false positive rate using different topologies. For example (African (CEU (Han Chinese, Native Americans))) is used to show a wrong phylogeny and the evidence of admixture from CEU to Native Americans. We show that the weighted D-statistic has more power than the previous method of sampling only one allele per site and is still powerful with the use of low-coverage data. We also further develop and generalize the theory regarding the admixture graph, a structure that extends the capability of the D-statistic in order to model complex histories of populations.

INFERRING DEMOGRAPHIC HISTORY OF MULTIPLE POPULATIONS BASED ON THE SITE FREQUENCY SPECTRUM

Vitor C Sousa[1,2], Isabel Alves[1,2], Isabelle Dupanloup[1,2], Laurent Excoffier[1,2]

[1]University of Berne, CMPG, Institute of Ecology and Evolution, Bern, Switzerland, [2]Swiss Institute of Bioinformatics, Computational Population Genetics, Lausanne, Switzerland

The Site Frequency Spectrum (SFS) summarizes the distribution of allele frequencies across populations, and it has been shown to be useful to infer past demographic events. Previously, we have developed a composite likelihood approach to infer parameters based on the SFS implemented in the program fastsimcoal2. This procedure relies on coalescent simulations to approximate the expected SFS under complex demographic models. Here, we extend this approach by considering an exact solution of the expected SFS based on the conditional expectations of coalescent branch lengths, which is suited for models with piecewise changes in population sizes, admixture events and multiple populations. This procedure is implemented within a conditional maximization algorithm to find the parameters that maximize the likelihood. As expected, under simple models, this approach leads to a better approximation of the SFS and more accurate parameter estimates than our previous method based on drawing random coalescent trees. We also compare our results to those of Chen (2012) [Theoretical Population Biology 81(2):179-195], who computed the exact SFS for simple scenarios under a coalescent framework. We apply these approaches to date and quantify the number of major admixture events between archaic hominins (Neanderthal and Denisovan), and modern human populations from Europe, Asia and Oceania.

A DEPENDENCE-AWARE COMPOSITE FRAMEWORK FOR IDENTIFYING AND LOCALIZING HARD SELECTIVE SWEEPS

Lauren A Sugden, Sohini Ramachandran

Brown University, Ecology and Evolutionary Biology, Providence, RI

To understand how human populations have evolved in response to selective forces, we need statistical methods that can effectively mine the patterns of observed genetic variation in diverse present-day genomes. Here, we introduce a novel framework for detecting hard selective sweeps, which leave behind three major genomic signatures: long-range haplotype blocks, altered site frequency spectra, and population differentiation. Many statistics exist to detect each of these signatures, and recently, composite methods have been introduced that detect selection by combining multiple statistics. Composite methods generally gain power, but statistics measuring similar signatures can introduce bias. Our method combines multiple statistics from across all three genomic signatures in a classification framework that returns probabilistically interpretable results, deals naturally with loci with undefined statistics, and accounts for the correlations among component statistics, using a machine-learning tool called an Averaged One-Dependence Estimator (AODE). Our approach is the first to explicitly model the dependence between pairs of statistics, thus minimizing bias, and allowing us to better understand how sweep parameters affect observed patterns of genomic variation.

Our classifier infers the probability that a locus has undergone a hard sweep based on joint component statistic distributions learned from extensive demographic simulations of neutrally evolving loci and hard sweeps with varying strengths and ages. In simulated data, we show that this classifier vastly outperforms other methods in detection and localization of sweep signals, in some cases reducing false positive predictions by seven-fold compared to the current state-of-the-art method. Our classifier performs particularly well when identifying completed sweeps and fast sweeps, which have great biological significance. In data from the 1000 Genomes Project, we recover known sweep regions, with high scores localized near previously validated adaptive mutations, including the genes *DARC* in West Africans, *EDAR* in East Asians, and *SLC24A5* in Europeans. We also show that the dependencies modeled in the AODE framework are necessary for the detection of signals within some genes, including *CD36* in West Africans, which harbors malaria resistance alleles. We apply our classifier to an exome dataset from the San population in Southern Africa, varying the demographic scenarios used for training in order to assess the effect of demographic misspecification. Our methods produce fewer false positives and negatives compared to existing approaches, thus identifying promising targets for experimental validation.

MULTIVARIATE APPROACHES INCREASE POWER IN GENOME-WIDE ASSOCIATION STUDIES

Michael C Turchin[1], Matthew Stephens[1,2]

[1]University of Chicago, Human Genetics, Chicago, IL, [2]University of Chicago, Statistics, Chicago, IL

For over a decade human geneticists have been conducting genome-wide association studies (GWAS) to elucidate the underlying genetic variants most important to both mendelian and complex traits. However, almost all GWAS analyses are performed in a univariate framework – they look for associations between genetic variation and a single phenotype, one at a time, even when the study measured multiple phenotypes. This is despite the fact that many papers have shown that taking a multivariate approach can increase power to detect significant genetic associations. For example, using a multivariate analysis of GWAS summary data from the Global Lipids Consortium (Teslovich et al. 2010), Stephens 2013 identified an additional 18 genetic loci associated with lipid levels compared with the published univariate findings.

Motivated by these results, we conducted multivariate analyses on the recently published updated global lipids data (Willer et al. 2013). Encouragingly, in the updated data, we find that the 18 additional genetic loci identified in Stephens 2013 are now considered 'genome-wide significant' in univariate analyses, confirming that the multivariate approach can increase power with little or no increase in false positive rate. Further, applying the multivariate analysis approach to the recently published data we identify an additional 74 putatively associated loci beyond the 157 found in Willer et al. 2013.

We are also applying this approach to other studies where data on multiple phenotypes are available. For example, we performed multivariate analysis of GWAS summary data from the 2010 and 2014/5 Genetic Investigation of ANthropometric Traits (GIANT) studies. As in the lipid analyses, preliminary multivariate analyses of these data identifies additional putative associations. These results serve both to add to our understanding of the genetic architecture underlying each of these phenotypes, and, we hope, to encourage researchers to perform multivariate analyses of other studies.

Stephens M (2013) A Unified framework for association analysis with multiple related phenotypes. PLoS ONE 8(7): e65245

Teslovich TM et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466(7307): 707-13

Willer CJ et al. (2013) Discovery and refinement of loci associated with lipid levels. Nat Genet 45(11): 1274-83

MATRIX ADAPTIVE SHRINKAGE: MODELING GENETIC EFFECTS ACROSS MULTIPLE SUBGROUPS

Sarah M Urbut[1], Gao Wang[1], Matthew Stephens[1,2]

[1]University of Chicago, Department of Human Genetics, Chicago, IL,
[2]University of Chicago, Department of Statistics, Chicago, IL

Variation in gene expression is an important mechanism underlying susceptibility to complex disease. The simultaneous genome-wide assay of gene expression and genetic variation allows the mapping of the genetic factors that underpin individual differences in quantitative levels of expression (expression QTLs; eQTLs). However, most studies are performed in a single tissue framework and fail to correlate the effect of genetics across multiple tissue types. Thus available methods are limited in their ability to jointly analyze data on all tissues to maximize power, while simultaneously allowing for both qualitative and quantitative differences among eQTLs present in each tissue. Here we introduce a method to efficiently mapping expression QTLs in large numbers of diverse cell-types and tissues from RNA-seq data.

We aim to combine information across tissues to fully acknowledge the multi-tissue nature of a SNP and report an effect size rather than simply binary outcome to compare among SNPs called active within a tissue or among tissues. This allows us to capture distinct variation in effect sizes within and between subgroups and to describe these `patterns of sharing.' We assume that each eQTL belongs to a group; within each group, the tissues exhibit characteristic patterns of sharing which can be captured by considering the covariance structure of the genetic effects among tissues. To mathematically capture these genetic effects, we apply a simple Bayesian Mixture mixture model, in which we assume all the gene-snp pair effect vectors arise from a mixture of a finite number of multivariate normal distributions. Each component of the mixture is defined by the prior covariance matrix from which the vector of standardized effect sizes is drawn. Choosing these directional covariance matrices is non-trivial as we hope that the combination will capture all the patterns of sharing present. Our novel approach provides a list of data-sensitive covariance matrices whose distinct diagonal and off-diagonal elements capture a wide array of patterns of sharing. Furthermore, providing a grid of scaling factors allows for data-adaptive 'shrinkage' according to the scaled covariance matrices which maximize the likelihood of the data set. The resulting model preserves tissue-specific effects while giving us insight about the distribution of of heterogeneity - that is, effects of opposite sign - and provides us with novel insight about overall patterns of sharing in the data-set.

← Back

A RANDOM MEASURE BASED FRAMEWORK FOR IDENTIFYING PATTERNS OF CLUSTERING IN LOCALIZED mRNA AND PROTEIN DISTRIBUTIONS

Jonathan H Warrell*[1,2], Anca F Savulescu*[1,2], Robyn Brackin*[1], Musa M Mhlanga[1,2]

[1]Council for Scientific and Industrial Research, Gene Expression and Biophysics Group, Pretoria, South Africa, [2]University of Cape Town, Division of Chemical Systems and Synthetic Biology, Faculty of Health Sciences, Cape Town, South Africa

We introduce a random measure based framework which permits the detection and comparison of clustering patterns in both point processes and continuous valued random measures. The framework is used to investigate the localization of corresponding mRNAs and proteins in polarizing mouse fibroblasts from fluorescence in situ hybridization (FISH) and immunofluorescence (IF) data respectively, and identify protein localization patterns which may be determined by pre-established mRNA localization and localized translation. Micropatterning is used to allow fine-grained comparisons of distributions, and a dependency of the patterning of many mRNA and protein distributions tested on the microtubule organizing center (MTOC) position is observed. The random measure framework permits statistical techniques and indices associated with Ripley K analysis to be applied outside a point process context. Specifically, we provide a discrete convolution based estimator of a general form of the Ripley K function, which is asymptotically consistent for both point process and continuous valued random measure cases. Further, we generalize statistical indices such as the Ripley H and L functions and clustering index by considering a general formulation of complete spatial randomness provided by completely random measures, and show how algorithms can be provided to estimate these quantities when samples from the appropriate random measures can be generated. In this respect, we consider gamma processes, marked Poisson processes and non-parametric sampling strategies as alternative ways of defining a general null hypothesis of complete spatial randomness.

* These authors contributed equally to this work

EFFECTS OF STOCHASTIC CO-REGULATION AND HIERARCHY ON THE STRUCTURE OF ATTRACTORS AND STEADY-STATE DISTRIBUTIONS OF *IN SILICO* GENE REGULATION NETWORK MODELS

Jonathan H Warrell[1,2], Musa M Mhlanga[1,2]

[1]Council for Scientific and Industrial Research, Gene Expression and Biophysics Group, Pretoria, South Africa, [2]University of Cape Town, Division of Chemical Systems and Synthetic Biology, Faculty of Health Sciences, Cape Town, South Africa

We propose a class of Boolean stochastic network models incorporating local rules with a property we describe as 'tiered co-regulation', and various degrees of global hierarchy and feedback to which the local rules may be tightly or loosely coupled. We are motivated in proposing this class of models by recent empirical work which suggests that co-regulation of gene expression in transcription factories is ubiquitous, and can be both stochastic and hierarchical, with stochasticity resulting from the fact that indeterminacy of nuclear 3D chromatin conformation in individual cells determines which transcription factories are formed and hence the degree of co-regulation. We use mean-field based theoretical approaches and *in silico* simulations to investigate emergent global properties of networks in the class proposed. Through such analyses, we show that 1) tiered co-regulation rules lead to networks with structured (hierarchical) noise at steady-state (using a probabilistic graphical model framework to analyse distribution / noise structure), 2) increasing local tiered co-regulation and limiting global feedback levels both increase the robustness / stability of a hierarchical network, and 3) local tiered co-regulation and hierarchical structure lead to networks with new attractors / steady-states which are 'close by' when the network is deformed by for instance fixing the expression levels of a fixed number of genes. We suggest the closeness of attractors / steady-states following network deformation may have functional relevance in the context of changes in gene expression resulting from signalling and cell differentiation, since it has been demonstrated that signalling events can alter co-regulation via changing chromatin conformation, and that observed gene expression dynamics are consistent with cell differentiation models involving movement between attractors in transcription factor expression networks. Further, we show through simulations that similar properties hold in a class of Markov Jump process models with analogous co-regulation rules and structure.

RECONSTRUCTING DYNAMICS OF BLOOD CELL DIFFERENTIATION FROM HIGH THROUGHPUT GENETIC BARCODING EXPERIMENTS

Jason Xu[1], Peter Guttorp[1], Janis L Abkowitz[2], Cynthia E Dunbar[3], Vladimir N Minin[1,4]

[1]University of Washington, Statistics, Seattle, WA, [2]University of Washington, Hematology, Seattle, WA, [3]National Institute of Health, National Heart, Lung, and Blood Institute, Bethesda, MD, [4]University of Washington, Biology, Seattle, WA

The human body regulates cell composition of blood through *hematopoiesis*, a complex system in which self-renewing hematopoietic stem cells (HSCs) produce mature blood cells. HSCs specialize or differentiate through a series of progenitor cells in a process whose dynamics and structure are largely unknown. Understanding the details of the hematopoietic system is a fundamental problem in systems biology, and progress in this area will also shed light on other areas of basic biology. For example, further advances in hematopoiesis research will yield insights into cellular interaction mechanisms, cell lineage programming, and characterization of cellular phenotypes during cell differentiation. Moreover, understanding hematopoiesis is clinically important. Stem cell transplantation is a mainstay of cancer therapy, and all blood cell diseases including myeloproliferative disorders and myelodysplasia are caused by malfunctions in some part of the hematopoiesis process.

Hematopoiesis research was one of the earliest successes of mathematical modeling in cell biology. However, findings from existing mathematical models, dictated by the available data as well as computational limitations, have not resolved long standing questions about patterns and sizes of the clones descended from individual HSC lineages. Recently emergent experimental techniques now allow for DNA labeling/barcoding of individual stem cells, allowing for lineage tracking through time by barcode identification. Although data from such procedures have already generated important insights, new statistical tools are needed to fully utilize the power of barcoding experiments. We propose a canonical branching process model of hematopoiesis and develop statistical algorithms to fit this stochastic model to the barcoded lineage data. To overcome computational challenges involved in this task, we develop new algorithms to efficiently handle large longitudinal datasets generated by experimentalists studying hematopoiesis in rhesus macaques. Our statistical methodology will allow hematologists to answer many questions about the hematopoietic system, and to use high resolution barcoding data to rigorously explore and compare alternative models of hematopoiesis.

ORTHOCLUST: A MULTI-LAYERS NETWORK FRAMEWORK FOR CLUSTERING HIGH-THROUGHPUT BIOLOGICAL DATA ACROSS SPECIES

Koon-Kiu Yan[1,2], Daifeng Wang[1,2], Mark Gerstein[1,2,3]

[1]Yale University, Program of Computational Biology and Bioinformatics, New Haven, CT, [2]Yale University, Department of Molecular Biophysics and Biochemistry, New Haven, CT, [3]Yale University, Department of Computer Science, New Haven, CT

Multi-layers network is a mathematical notion where multiple layers of networks are concatenated to form an interconnected structure. While it has been widely interested in computational social science to represent different social circles of an individual (e.g. Facebook network, LinkedIn network, and Twitter network form a multi-layer network), such notion is of particular interest to biology because of the existence of multiple relational connections (e.g. co-expression, genetic interactions), multiple types of regulation operating at different time-scales (e.g. transcriptional regulation, post-translational phosphorylation), as well as corresponding networks across different species. As a step to leverage the mathematical framework for mining biological data, we developed OrthoClust, a novel orthology-based multi-layer network framework for the simultaneous clustering of networks from multiple species. OrthoClust integrates the co-association networks of individual species with the orthology relationships of genes between the species, arrives at a multi-layer network that is unique to biological application. As an example, we applied OrthoClust to construct modules for C. elegans, D. melanogaster and human using RNA-Seq data from the ENCODE consortium. The resultant expression modules are defined in a cross-species fashion – a cluster could consist of genes from multiple species and could be either conserved or species-specific. Modules were then used for inferring putative functions of ncRNAs based on "guilt-by-association"; worm, fly and human ncRNAs anchored to the same module may potentially have analogous functions. Furthermore, the concept of orthology-based meta-clustering illustrated by OrthoClust can be generalized to take into account of the dynamics of expression profiles across species via a state space model, resulting at the dynamical interplay between the conserved and species specific modules.

CHARACTERIZING THE TRANSFER OF CRISPR ARRAYS

Joy Y Yang[1], Mark B Smith[2], Martin F Polz[3], Eric J Alm[2,3,4]

[1]MIT, Computational and Systems Biology, Cambridge, MA, [2]OpenBiome, Microbiology, Medford, MA, [3]MIT, Civil and Environmental Engineering, Cambridge, MA, [4]MIT, Bioengineering, Cambridge, MA

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and their Cas (CRISPR associated) genes are present across a wide range of Archaea and Bacteria genomes. These sequences have been fairly recently proven to act as an adaptive immune system against phage. [1] CRISPR systems are suspected to be transferrable across widely diverged lineages [2], and in fact, CRISPR subtypes are hypothesized to have emerged through horizontal gene transfer and recombination. [3]

These conclusions are drawn mainly through analyzing Cas gene and repeat similarity as well as neighboring transposon sequences. This type of analysis is effective for identifying distant transfers and quantifying transfer frequency. However, the identification of recent transfer events and transfer functionality is not yet well characterized, and can be better addressed by analyzing shared spacer content.

In order to further quantify the occurrence of array transfers, we first calculate the average proportion of genomes that carry each gene (correcting for the phylogenetic covariance as a result of non-random sampling); then fit hidden markov models to these adjusted proportions over each genome. We can then calculate the posterior probability of an array lying in a transferred genomic island. Maintenance of these transferred arrays hints at the possibility that phage may have host range more broad than previously believed.

Finally, we observe that single identical spacers are often shared between bacteria at a wide range of 16S distances. Whether this is the result of highly beneficial spacers that have been transferred and maintained or the result of independent acquisition is not entirely clear. However, in either case, such spacers can provide insights into biases in this defense system as well as the susceptibilities in their phage targets.

1. Barrangou, R., Fremaux, C., Deveau, H., et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science, 2007, 315, 5819, 1709-1712.
2. Bhaya, D., Davison, M. and Barrangou, R. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu.Rev.Genet., 2011, 45, 273-297.
3. Makarova, K. S., Haft, D. H., Barrangou, R., et al. Evolution and classification of the CRISPR-Cas systems. Nat.Rev.Microbiol., 2011, 9, 6, 467-477.

← Back

MIXED-LAYERED NETWORK MODELING FOR GENE EXPRESSION ACROSS INDIVIDUALS AND TISSUE TYPES

Shuo Yang[1], Dana Pe'Er[2,3], Itsik Pe'Er[1,3]

[1]Columbia University, Department of Computer Science, New York, NY,
[2]Columbia University, Department of Biological Sciences, New York, NY,
[3]Columbia University, Department of Systems Biology, New York, NY

Expression quantitative loci (eQTLs) have been extensively studied, as they provide an attractive functional interpretation to sequence variation. Yet, many traditional eQTL studies had been mainly focused on *cis*- analysis of each transcript, without drawing on the complete picture of the regulatory repertoire. Furthermore, such studies standardly remove genomewide effects from data as a preprocessing step, interpreting them as artifactual batch effects rather than genetic or other biological signal. In this work, we devise a mixed-layered regression model, considering polygenic multi-loci eQTL regulation of gene expression. The model includes several layers of variables, where successive layers are connected through logistic functions whose parameters are inferred through regression with sparsity regularization. A key feature of our model is allowing *trans*-SNPs to indirectly affect a particular target gene through an intermediate hidden layer, which we call *cell environment variables*. We encourage these variables to affect gene expression in a cell-type-specific fashion. Variables in this hidden layer are designed to implicitly describe unmeasured quantities in the sampled cells, such as chromatin state, activity of a cellular pathway, level of a small molecule, or unknown mechanisms, all of which affect expression in a cell-type-specific manner. The flexibility of the model allows additional features: genomewide effects that are due to individual genotypes (population stratification) and samples (experimental, environmental and other sources) can be modeled directly in a non-linear fashion, which we show is required by real data. Additional features include prior assumptions regarding *cis*-eQTLs can be incorporated; and multiple tissues can be modeled as mixtures of cell types whose characteristic expression profiles are related according to a known hierarchy. We solve our model using stochastic gradient descent with back-propagation, and we test our modeling in GTEx dataset. We present computational challenges arising from such analysis endeavor, along with strategies to mitigate them.

COMPUTATIONAL METHODS FOR ANALYZING IMMUNOGLOBULIN HEAVY CHAIN GENES

Shishi Luo, Jane Yu, Yun S Song

University of California, Berkeley, Computer Science, Berkeley, CA

Antibodies, which are determined by the genetic data of their corresponding B-cell receptors, must be capable of recognizing a diverse and large set of antigens in order to adequately fend against infectious diseases. Consequently, there exists a tremendous number of different B-cell receptors, and this magnitude captures the staggering variability and complexity of the immunoglobulin (IG). Clearly, a more feasible mechanism of characterizing and assessing antibody variation is needed. Our work intends to understand some of the underlying determinants that contribute to this great diversity, namely the immunoglobulin heavy chain variable region (IGHV) locus. This locus codes for the part of the antibody that recognizes foreign antigens, and appears to have undergone numerous gene duplication and diversification processes. The result is that IGHV haplotypes (instances of the IGHV locus) vary not only by single nucleotide polymorphisms, but also in the copy number of gene segments and the ordering of gene segments. To date, complete haplotype sequencing and analysis of the human IG heavy-chain V genes has only been reported by Watson et al.\cite{watson} and Matsuda et al.\cite{matsuda}, illustrating the sparse work done in this area and the promise of this new research. Our work proposes computational methods to overcome this challenge so that we may be able to explore genotypic differences between individuals. Our work also analyzes the IGHV locus of a pedigree of sixteen individuals as well as data of 281 individuals of various ethnicities, elucidating interesting and complex comparisons among individuals. Through analysis of this data, we also propose novel IGHV alleles as well as present an assessment of variation never before previously quantified.

INFERRING A HAPLOTYPE REFERENCE PANEL FOR *PLASMODIUM FALCIPARUM* GENOMIC VARIATION FROM FIELD SAMPLES WITH MIXED INFECTIONS

Sha Zhu[1], Pf3k consortium [2,3,4]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [2]Broad Institute, Cambridge, MA, [3]University of Oxford,, Oxford, United Kingdom, [4]Wellcome Trust Sanger Institute,, Cambridge, United Kingdom

Genetic variation within the malarial parasite *Plasmodium falciparum* is known to affect key phenotypes including drug resistance and risk of severe disease. Advances in technology and experimental protocol mean that obtaining high coverage genome sequencing data from routine blood samples taken in the field is now possible. However, interpreting such data is difficult because of high rates of mixed infections and highly variable data quality.

To provide a framework for analysing genetic variation in *P. falciparum*, the Pf3k project is working to build a global reference map of genome sequence and tools that can enable rapid analysis of data from field samples, with 2,512 samples available to date. Here, we describe methods developed within the project for inferring the structure and identity of strains present in a sample by combining a reference panel of known haplotypes with data from an additional sequencing experiment. In particular, we describe Monte Carlo methods for inferring haplotypes present in a sample that generalise techniques developed for diploid samples, but which can cope with multiple strains and the over-dispersion of allele counts that results from experimental protocol. We use simulations and analysis of controlled mixtures of known genomes to assess the accuracy of the method.

By applying these methods to data from two geographical regions with distinct malarial epidemiologies we characterise how levels of mixed infection, diversity, geographical differentiation and haplotype structure vary in the parasite. In particular, we find that a minor of samples (115 out of 228 cases from Southeast Asia, and 99 out of 478 cases from Ghana) present evidence of infection by a single parasite strain and multiple cases with over two strains. Our results demonstrate the feasibility of inferring genome-wide patterns of haplotype structure in malarial parasites taken from clinical field samples and establish a resource for driving the development of new approaches for integrating population genetic and epidemiological modelling.

BAYESIAN VARIANT-BASED PATHWAY ENRICHMENT ANALYSIS USING GWAS SUMMARY STATISTICS

Xiang Zhu[1], Matthew Stephens[1,2]

[1]University of Chicago, Department of Statistics, Chicago, IL, [2]University of Chicago, Department of Human Genetics, Chicago, IL

Carbonetto and Stephens (2013) developed a multiple-SNP modeling approach that integrated pathway enrichment analysis with variant prioritization in enriched pathways, and demonstrated its potential to yield novel biological insights into complex human traits and diseases. The method, however, is limited by the requirement of individual-level genotype and phenotype data, which are not widely available for large GWAS. In contrast, single-SNP association summary statistics are often released in public domain. Here we present a new Bayesian method for multiple-SNP pathway analysis that relies solely on GWAS summary statistics and linkage disequilibrium (LD) structure inferred from a public reference panel. This strategy is justified by the fact that, under standard genetic association model, the posterior of multiple-SNP effect sizes depends on individual-level GWAS data only through single-SNP summary statistics and sample LD. Our method adopts a recently proposed large-scale Bayesian regression model for GWAS summary statistics (Zhu and Stephens, ASHG 2015). Unlike in previous work where each SNP was treated equally likely to be associated with the phenotype a priori, the new method allows the prior probability of each SNP being causal to depend on its membership of a pathway so that potential enrichment of associations within the pathway can be captured. A parallel algorithm using mean-field variational approximation is developed to ensure scalability for posterior inference in genome-wide applications. On GWAS summary statistics of four common diseases from the WTCCC studies and 3,160 pathways from eight web databases, our method obtains results comparable to the analysis that used individual-level data (Carbonetto and Stephens, 2013). For rheumatoid arthritis and type 1 diabetes, the most enriched pathway is primarily contributed by a set of genes within the major histocompatibility complex region (posterior probabilities = 1, 0.97). The top-ranked pathways that show strong support for enrichment in Crohn's disease are IL12-mediated signaling, cytokine signaling, IL23-mediated signaling and immune system (BF = 2.63e9, 7.41e8, 5.25e8, 1.74e6). Two pathways, incretin and glucagon-like peptide-1 regulation, show moderate evidence for enrichment in type 2 diabetes (BF = 38.90, 33.88).

A GENETIC AND SOCIO-ECONOMIC STUDY OF MATE CHOICE IN LATINOS REVEALS NOVEL ASSORTMENT PATTERNS

James Zou[1], Danny Park[2], Esteban Burchard[2], Dara Torgerson[2], Maria Pino-Yanes[2], Yun Song[3], Sriram Sankararaman[4], Eran Halperin[5], Noah Zaitlen[2]

[1]Microsoft Research, Research, Cambridge, MA, [2]UCSF, Medicine, San Francisco, CA, [3]UC Berkeley, Computer Science, Berkeley, CA, [4]Harvard University, Medicine, Boston, MA, [5]Tel Aviv University, Computer Science, Tel Aviv, Israel

Understanding the patterns and drivers of nonrandom mating in human populations has important implications for genetics and medicine as well as for economics and sociology. Previous studies have provided evidence for assortative mating in diverse populations and have suggested that genetic assortment is significantly smaller than educational assortment in non-Hispanic whites. Nevertheless, there remain important open questions regarding assortment patterns (especially in non-European populations), the relationship between assortment based on genomic factors and socio-economic factors (such as income), and biological mechanisms driving such assortment.

In this work we addressed these questions by jointly analyzing detailed socio-economic attributes and high-quality genotype data associated with large cohorts of Mexicans and Puerto Rican individuals sampled from multiple geographic locations. We developed a novel computational method, ANCESTOR, which uses a probabilistic model of pooled semi-Markov processes to accurately infer the genomic ancestry (fraction of the genome derived from European, Native American and African ancestries) of each parent from just the offspring's genotype. We applied ANCESTOR to our richly phenotyped Latino cohorts To quantify genomic assortment in the absence of parental genotypes.

In ethnically homogeneous Latino communities, we quantified genomic ancestry as a key axis of assortment. Partners are much more likely to share similar genomic ancestries than random individuals. Consistent with this, partners are more closely related—equivalent to between third and fourth cousins in Mexicans and Puerto Ricans—than expected based on random mating. Our analysis showed that socio-economic factors only explain a small portion of the genomic ancestry similarity between partners. Strikingly, after normalizing by population background, partners are more similar in genomic ancestries than in education levels. We found that Puerto Rican couples are especially correlated in genes involved in facial development (above and beyond genome-wide similarities), suggesting that similarity at these genes represents an axis of assortment. We replicated our findings across multiple locations. Our analysis integrated population genomics with quantitative social sciences to address fundamental questions about mate selection.

← Back

ASSESSING THE VARIABILITY OF PROTEIN FAMILY ABUNDANCE IN THE HUMAN GUT MICROBIOME

Patrick H Bradley[1], Katherine S Pollard[1,2]

[1]UCSF, J. David Gladstone Institutes, San Francisco, CA, [2]UCSF, Div. of Biostatistics, Institute for Human Genetics, Institute for Computational Health Sciences, San Francisco, CA

The human gut microbiome is a diverse ecosystem, harboring microbes capable of performing a wide array of biochemical functions. The full complement of microbially-encoded functions also varies between individual hosts. Gene families that remain relatively constant could include those necessary for microbial fitness as well as for successfully colonizing the human GI tract, while those that are highly variable could be subject to drift or biogeographical variation, or could be involved in adaption to specific niches. However, we have previously lacked quantitative methods to assess this variability. Here, we develop a statistical test to identify gene families that are significantly more or less variable than expected across individuals. This method controls for between-study variation and preserves correlations between gene families via resampling-based multiple testing. We then apply this test to shotgun-sequenced stool samples from healthy individuals drawn from three different populations. We also assess whether the variable and invariable gene families that we identify can be explained by their estimated phylogenetic distribution. Several two-component signaling pathways and PTS transporters, despite being predicted to have a relatively restricted phylogenetic distribution, appear to have low variability across individuals; conversely, gene families involved in flagellar assembly appear to have high variability despite being more broadly distributed across bacteria. Many biological pathways, including central carbon metabolism, lipopolysaccharide biosynthesis, and bacterial secretion systems, appear to have both “core” and variable components, the latter of which may relate to differences in lifestyle and host interaction. Additionally, many of the protein families that we identify as variable also co-vary across individuals but do not appear to correlate with measured clinical variables, indicating a potential role for unmeasured variables in driving the functional composition of the gut microbiome.

A PROBABILISTIC MODEL FOR INTEGRATING GENOME-SCALE DATA WITH TREE-LIKE DEPENDENCIES

John A Capra[1], Dennis Kostka[2]

[1]Vanderbilt University, Biological Sciences, Nashville, TN, [2]University of Pittsburgh, Developmental Biology, Pittsburgh, PA

Precise spatiotemporal control of gene expression is essential for proper differentiation and development. A complex array of dynamic biochemical events, including transcription factor (TF) binding, histone modification, and DNA methylation interact to regulate RNA expression across cellular transitions. As cells differentiate through these transitions they trace hierarchical trajectories, and these relationships can be represented with a lineage tree that links progenitor cells to one or more descendant lineages. Recent experimental efforts have generated thousands of genome-wide profiles of functional marks and gene expression from hundreds of cell types. These data offer the promise of a systems-level understanding of the molecular drivers behind cellular state transitions. However, appropriate probabilistic models that integrate these diverse data and take the lineage relations of assayed cell types into account are still needed. Therefore, we developed a statistical framework, based on Ornstein-Uhlenbeck type processes, for modeling the dynamics of genome-wide data that explicitly accounts for the tree-like dependencies between different cellular contexts and flexibly integrates data generated by different genome-wide assays, e.g. for histone modifications, TF binding, and gene expression.

Using recent genome-scale epigenetic modification and gene expression data from 16 stages of blood cell development, we show that taking cell-type dependencies into ac-count improves on common analysis strategies and enables new analyses. For DNA methylation our model improves the inference of missing data, and for gene expression data it results in higher power and decreased false positive rates in tests for differential expression. Our statistical framework also enables identification of genomic loci with significant shifts in epigenetic state or gene expression over specific transitions in blood differentiation. Finally, using our approach to remove correlations induced by develop-mental relationships facilitates the application of causal inference algorithms to build a regulatory network that includes epigenetic modifications and gene expression. The resulting network recapitulates many known regulatory relationships between hematopoietic transcription factors and suggests several specific epigenetic modifications that may mediate these interactions.

In summary, our integrative OU-based framework enables robust unbiased identification of significant changes in genomic data across related cellular contexts and provides a convenient platform for investigating the molecular interactions that drive gene expression dynamics.

TRANSCRIPTIONAL REGULATION INFERENCE FROM CHROMATIN ACCESSIBILITY AND GENE EXPRESSION MEASUREMENTS IN INNATE LYMPHOID CELLS

Emily R Miraldi[1,2,3], Maria Pokrovskii[2], Jason A Hall[2], Nicholas Carriero[1], Dan R Littman[2], Richard A Bonneau[1,3]

[1]Simons Foundation, Simons Center for Data Analysis, New York, NY, [2]New York School of Medicine, Kimmel Center for Biology and Medicine of the Skirball Institute, New York, NY, [3]New York University, Center for Genomics and Systems Biology, Courant Institute of Mathematical Sciences, New York, NY

Innate lymphoid cells (ILCs) compose a newly discovered and relatively rare population of immune cell lineages (ILC1, ILC2, ILC3) that play many important roles, including early defense against infections, tissue homeostasis and repair, and autoimmune disease. These diverse ILC physiological functions require coordination of complex gene expression patterns, involving thousands of genes. Our goal here is the inference of transcriptional regulatory networks (TRNs) to model these gene expression responses as multivariate functions of transcription factor (TF) activities. Microarray and RNAseq technologies enable systematic measurement of tens of thousands of genes, but inference of TRNs from gene expression data is nontrivial because (1) most datasets are in the underdetermined regime (there are many more TFs (e.g., predictors) than samples), and (2) TF gene expression is an inexact proxy for TF activity, as gene transcripts often require downstream processing (e.g., translation, chemical modification) prior to activation (i.e., affecting target gene expression). The Inferelator is an algorithm for time-series gene expression data that uses sparse regression to infer parsimonious networks of TRNs. This algorithm was originally applied to infer TRNs in prokaryotes, but was recently adapted for the more complex mammalian setting, to build a TRN for Th17-cell development in vitro. Inference of the Th17 TRN required both computational developments and integration of epigenetic data in addition to gene expression. Specifically, ChIPseq data was required; ChIPseq provides information about where specific TFs physically bind DNA and thus which genes they likely regulate. While ChIPseq experiments are feasible for highly abundant and/or culturable cells, the scarcity of ILCs precluded application of this technique here. In this work, we develop methods to integrate a newly developed technology, assay for transposase-accessible chromatin (ATACseq), which can be combined with TF motif information and gene expression data, to gain ChIPseq-like information for multiple TFs, from a single experiment involving orders of magnitude fewer cells. The newest Inferelator version provides a unified framework to integrate prior knowledge of TF-gene interactions and was validated in bacteria. Here, we adapt this methodology to integrate the ATACseq information as a prior in the mammalian setting. First, we validate our integrative method for TRN inference from ATACseq and gene expression data in the more well-studied context of in vitro Th17 cells, for which we also generated ATACseq data and for which a gold-standard network based on TF knockout and ChIPseq data exists. We then apply our methods to ILCs and use the resulting TRNs to understand ILC regulation in host health and pathophysiology.

← Back

SHARING AND SPECIFICITY OF CO-EXPRESSION NETWORKS ACROSS 35 HUMAN TISSUES

Emma Pierson[1], Alexis Battle[2], Sara Mostafavi[3]

[1]Oxford University, Computer Science, Oxford, United Kingdom, [2]Johns Hopkins University, Computer Science, Baltimore, MA, [3]University of British Columbia, Statistics, Medical Genetics, Vancouver, Canada

Tissue-specificity, in which cells perform different functions despite possessing identical DNA, is achieved partially through tissue-dependent mechanisms of gene regulation, including epigenetic modification and transcriptional and post-transcriptional regulation. These complex programs of control produce different gene expression programs across tissues, with most genes showing statistically significant differential expression. To understand the regulation of tissue-specific gene expression, we used a large compendium of gene expression data across 35 human tissues generated by the GTEx project. This data provides an opportunity for deriving shared and tissue specific gene regulatory networks on the basis of co-expression between genes. However, a small number of samples are available for a majority of the tissues, and therefore statistical inference of networks in this setting is highly underpowered. To address this problem, we developed a new algorithm for inferring tissue-specific gene co-expression networks that uses a hierarchy of tissues to share data between related tissues. We show that this transfer learning approach increases the accuracy with which networks are learned. Applying this model to the GTEx dataset, we make several observations about patterns of tissue specificity. First, we observe that tissue-specific transcription factors are hubs that preferentially connect to genes with tissue specific functions. Second, we observe that genes with tissue-specific functions lie at the peripheries of our networks. In our ongoing work, we apply the same model to transcript-level data, identifying patterns of tissue-specificity that are mainly evident at the transcript level.

THE LENGTH AND DISTRIBUTION OF ADMIXTURE TRACTS

Mason Liang, Rasmus Nielsen

UC Berkeley, Departments of Integrative Biology and Statistics, Berkeley, CA

The distribution of admixture tracts in admixed populations provides information about the process of admixture, in particular the number and timing of admixture events. We develop theory that can be used to predict distributions of admixture tract lengths under a variety of models. We show that the distribution of admixture tracts in the genome is more complicated than previously assumed and often is not well-approximated by a Markov process. We also develop theory that described the variance of of admixture proportions among individuals in a population as a function of the admixture time(s), and show how these results can be used for estimation purposes. Finally, we extend existing theory on admixture linkage disequilibrium (LD) to the case of there-point LD, and use these results to develop a new methods for improved estimation of admixture times and for estimation of admixture times in scenarios with more than one admixture event.

DEMOGRAPHIC-AWARE INFERENCE OF THE STRENGTH OF PURIFYING SELECTION BASED ON HAPLOTYPE PATTERNS

Diego Ortega Del Vecchyo[1], Kirk E Lohmueller[1,2], John Novembre[3]

[1]University of California, Los Angeles, Bioinformatics, Los Angeles, CA, [2]University of California, Los Angeles, Ecology and Evolutionary Biology, Los Angeles, CA, [3]University of Chicago, Human Genetics, Chicago, IL

The extent of purifying selection in a population plays a key role in determining the amount of genetic and possibly phenotypic variation in a population. Recent genome sequencing studies with large sample sizes in humans capture a vast amount of putatively deleterious rare variants, providing an important source of information to analyze how selection impacts those variants. To estimate the strength of purifying selection, we have developed a novel likelihood-based method that uses the lengths of pairwise haplotype identity by state among rare-variant carrying haplotypes. Our method conditions on the present-day frequency of the allele and is based on theory predicting that, under constant population sizes, the alleles under purifying selection are on average younger than neutral alleles and should have higher average levels of haplotype identity among variant carriers. We developed a computational framework to obtain the probability distribution of the lengths of pairwise haplotype identity given a certain selection coefficient, demographic scenario and present-day allele frequency. The method performs two integrations: one over all possible allele frequency trajectories given a certain selection coefficient and present-day allele frequency using a fast importance-sampling algorithm and another integration over all pairwise coalescent times given a certain allele frequency trajectory. We find the maximum likelihood estimate of the selection coefficient given a set of pairwise haplotypic identity lengths using a grid-search approach. Simulations indicate that our method provides unbiased estimates of selection in scenarios with both constant and variable population sizes. We provide an example of how to apply this method to estimate the selection coefficient of variants in different functional categories from sequence data of 1577 Sardinian individuals.

ANCESTRY LOCALIZATION FROM GENOTYPE DATA UNDER GENERAL PROBABILISTIC MODELS OF SPATIAL EVOLUTION AND A CORRECTION FOR POPULATION STRATIFICATION IN GENOME-WIDE ASSOCIATION STUDIES

Adel Javanmard[1,2], Anand Bhaskar[3], Thomas Courtade[4], David Tse[1]

[1]Stanford University, Electrical Engineering, Stanford, CA, [2]University of Southern California, Marshall School of Business, Los Angeles, CA, [3]Stanford University, Genetics, Stanford, CA, [4]University of California, Berkeley, Electrical Engineering and Computer Sciences, Berkeley, CA

Genome-wide association studies (GWAS) can be confounded by hidden population structure when population ancestry is correlated with both the genotype and the trait. Popular methods to correct for population ancestry in GWAS first project the genotypes into a low dimensional space that approximately captures the ancestry-dependent component of genetic variation, and then remove this ancestry-genotype correlation from the data. These ancestry correction procedures typically employ linear methods such as principal components analysis (PCA). However, while linear methods like PCA have been used for both geographic localization of samples and for correcting for population structure in GWAS, there is a lack of a sufficiently rich probabilistic model to justify the use of PCA.

We revisit this problem by first defining a general flexible probabilistic model of allele frequency evolution that generalizes several previously developed parametric models of spatial genetic variation such as SPA (Yang et al., Nature Genetics 2012) and SpaceMix (Bradburd et al., bioRxiv 2015). In our model, individuals are sampled from some geographic space, where the allele frequency at each SNP at any geographic location is drawn from some arbitrary stationary stochastic process that represents the complex patterns of historical population structure and migration.

We develop a localization algorithm that uses the correlation between genotypes in the sample to infer their ancestral locations. Our algorithm can be viewed as a form of manifold learning, and attempts to place individuals in the geographic space while respecting the genotype correlations between genetically similar individuals. Our algorithm is oblivious to the minutiae of the probabilistic model, a feature that distinguishes it from algorithms developed for parametric models such as SPA and SpaceMix. On simulated and real data from 95 human subpopulations, we observe a 25% reduction in localization error compared to PCA.

We then develop a hypothesis testing procedure for detecting association between the trait and genotype which uses the inferred ancestral locations of the sample individuals to correct for spatial variation in allele frequencies. Our procedure has higher power and lower type-I error relative to popular PCA-based methods like EIGENSTRAT on both real and synthetic GWAS datasets.

INFERRING LOCAL ANCESTRY BY JOINTLY ANALYZING ADMIXED SAMPLES

Amy L Williams

Cornell University, Biological Statistics & Computational Biology Dept, Ithaca, NY

Local ancestry inference identifies the population of origin of DNA at each position in an admixed individual's genome. This information is necessary to perform admixture mapping, an approach to disease and trait association mapping that identifies loci with significant deviations in ancestry proportions among a set of admixed individuals that have the trait of interest. Local ancestry information has also been widely used to study human demographic history, and because rare variants correlate with local ancestry, local ancestry may be useful in addressing fine-scale population stratification in rare variant association studies.

Numerous methods exist for inferring local ancestry, but the accuracy of their inference in Latino and other groups is not ideal for case-only admixture mapping and has potential to be improved. The most accurate local ancestry methods require panels of unadmixed individuals to help build models of the ancestral populations that contribute ancestry to the admixed individuals. A challenge for such methods that rely heavily on panels is that ancestral populations may only exist in admixed form, with related unadmixed populations being drifted from those ancestors. An alternative to requiring panels is to attempt to leverage information from within the admixed samples themselves. The method RFMix performs an iterative approach that, beginning with unadmixed panels, also utilizes admixed individuals to infer local ancestry. RFMix has improved results compared to other approaches, but may not fully capture the latent information within admixed samples because of its initial dependence upon panels for inference in the first iteration.

We introduce MIX-HAPI, a method that extends the haplotype inference framework HAPI-UR to infer local ancestry. The algorithm uses global ancestry estimates as a prior probability for local ancestry and leverages all input samples with any combination of admixed or unadmixed individuals. Our initial evaluation using simulated data shows that MIX-HAPI is extremely accurate in inferring local ancestry in the presence of unadmixed individuals that derive from a different population than the true ancestral group. The approach can be iterated via expectation-maximization in order to better infer local ancestry in regions that are initially uncertain. We are currently evaluating the method on simulated data, including Latinos, with variable numbers and population origins of unadmixed individuals included in the analysis. Because unadmixed Native American groups may be differentially related to the ancestors of Latinos, this method holds promise to improve inference accuracy in real data and has application to other admixed population groups.

MULTIDIMENSIONAL SCALING (MDS) ANALYSIS, SPECTRAL DECOMPOSITION AND COALESCENT THEORY

Ivan Levkivskyi[1], Anna-Sapfo Malaspinas[2]

[1]Institute for Theoretical Physics, Physics, Zürich, Switzerland, [2]Institute of Ecology and Evolution, Biology, Bern, Switzerland

Population structure plays an important role in determining the evolutionary history of a group. In recent years, the unprecedented increase in sequencing data has opened up a wide range of possibilities to investigate population histories – provided one can handle such large amounts of data. Methods based on non-parametric multidimensional statistics (more specifically principal components analysis, PCA) were first applied to genetic data more than 30 years ago. PCA has since become a standard tool in population genetics owing in particular to the low computational demand of such analyses. In this work, we investigate a related statistical approach, namely multidimensional scaling (MDS).

Following a recent study by McVean [1], we first derive analytical results for two and three populations without migration that relate the Euclidean distance between points on the MDS plots with pairwise coalescent times. We then perform coalescent simulations and study the rate at which simulated data converges towards theoretical predictions as a function of the number of SNPs. Finally we generalize the approach to an arbitrary number K of populations.

[1] McVean, G., 2009. A Genealogical Interpretation of Principal Components Analysis. PLoS Genet 5, e1000686. doi:10.1371/journal.pgen.1000686

APPROXIMATING SWEEPS IN THE ARG FOR INFERENCE

Jeremy J Berg, Graham Coop

University of California, Davis, Center for Population Biology, Davis, CA

A great deal of analytical work has focused on describing the joint effect of selection and recombination on sequence variation, and on the way genealogies surrounding a selective sweep differ from those of the neutral coalescent. However, methods for accurately simulating the joint impact of selection and recombination on sequence variation do not necessarily lend themselves easily to inference, and methods for detecting or characterizing selective sweeps have therefore mostly relied on simple summary statistics.

The full history of a set of DNA sequences can be captured by a structure known as the ancestral recombination graph (ARG), which catalogs the set of recombination and coalescent events ancestral to the sampled chromosomes. Recent methods development projects have produced a number of approaches to infer all or most of the structure of the ARG as a route to inferring various parameters regarding the evolutionary history of the population.

In this work we consider the influence of sweeps on the ARG as a collection of recombination events and simultaneous multiple mergers which are correlated along the sequence and whose distribution depend on the variety of sweep which has occurred. We use this approximation to simulate a recurrent sweep model in the background of the standard coalescent with recombination, and we take steps toward explicit inference of sweeps within models of the ancestral recombination graph by considering how to interpret our point process approximation for selective sweeps within the sequentially Markov coalescent.

DISENTANGLING TRANSCRIPTIONAL HETEROGENEITY AMONG SINGLE-CELLS: A BAYESIAN APPROACH

Catalina Vallejos[1,2], Sylvia Richardson[1], John Marioni[2]

[1]University of Cambridge, MRC Biostatistic Unit, Cambridge, United Kingdom, [2]University of Cambridge, MRC Biostatistic Unit, Cambridge, United Kingdom, [3]European Bioinformatics Institute, EMBL, Hinxton, United Kingdom

The revolution of transcriptomics — moving from bulk samples to single-cell resolution — can provide novel insights into the regulation of gene expression e.g. by identifying differentially expressed genes between two cell types or by using gene expression profiles in order to characterise distinct (and potentially unknown) sub-populations of cells. To analyse single-cell RNA-seq data, methods that have been formulated for bulk experiments cannot be directly applied. In particular, successful normalisation strategies for bulk RNA-seq (scRNA-seq) datasets can lead to unstable results at single-cell level and the effect of technical variation — reflected in weak correlations among technical replicates — is often ignored by such approaches.

BASiCS (Bayesian Analysis of Single-Cell Sequencing data) [1] is a hierarchical Bayesian model that deals with normalisation, the quantification of technical variation and other downstream analyses (e.g. identifying highly and lowly variable genes) simultaneously, borrowing information from intrinsic transcripts and technical spike-in genes. In this work, we extend BASiCS to include other downstream analyses that help functional characterization.. First, we focus on the detection of differentially expressed genes between two pre-specified populations of cells (defined by experimental conditions or cell-types). In contrast to most related literature, we do not only examine changes in the overall expression level between the samples, instead we also study changes in cell-to-cell biological heterogeneity. Evidence of a gene being differentially expressed is quantified by means of Bayes Factors and a decision rule is calibrated using the spike-in genes and permutation experiments. Secondly, we explore cases where the aim is to characterise unknown sub-populations of cells by clustering cells according to their gene expression profile and selecting markers for representing such sub-populations. We illustrate our methods using simulated and real datasets.

[1] Vallejos, Marioni and Richardson (2015). BASiCS: Bayesian Analysis of Single-Cell Sequencing data. PLoS Computational Biology 11(6): e1004333

LEARNING HIGH DIMENSIONAL CAUSAL NETWORKS FROM OBSERVATIONAL DATA USING MULTIPLE GENETIC INSTRUMENTS

Benjamin Frot[1], Luke Jostins[2], Gilean McVean[1,2]

[1]University of Oxford, Department of Statistics, Oxford, United Kingdom,
[2]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom

Mendelian randomisation, an instrumental variable method, has been successful at establishing causal relationships in epidemiology. However, current methods have limited use on datasets comprising many genetic variants and measurements (*e.g.* from gene expression or biobank studies). In particular, the assumption that instruments act on the outcome only through the intermediate phenotype (no-pleiotropy) is hard to verify in practice. Moreover current methods cannot deal with many correlated measurements.
Here, we ask when and how it is both valid and possible to estimate a latent linear causal network over a set of correlated variables without prior knowledge about the validity of genetic instruments. Our starting point is the graphical lasso (GLASSO) estimator of Friedman et al., which is a popular tool to learn sparse Gaussian Graphical Models (GGM) in the absence of confounding. The Low Rank plus Sparse decomposition (LRPS) of Chandrasekaran et al. extends GLASSO by decomposing the precision matrix as a sum of a sparse and a low-rank matrix, while the Sparse Conditional GGM (SCGGM) model of Zhang L et al. extends GLASSO by modelling the relationship between genetic instruments and measurements. Building on LRPS and SCGGM we suggest a tractable penalised likelihood estimator which relies on the convexity of the l_1 and nuclear norms. In the absence of genetic variants, our method is equivalent to LRPS while, in the absence of confounders, its behaviour is closer to SCGGM.
First, we show that the problem of learning the statistical model over the correlated measurements is well-defined in a wide range of situations that correspond to distinct biological problems. With low pleiotropy, consistency is achieved in the presence of arbitrary confounding. In contrast, with high pleiotropy, causal relationships can still be identified as long as confounders have a low rank structure. We validate the new method through extensive simulations.
To assess the value of the new method we analyse two large population studies of human gene expression. After selecting a subset of 1,000 genes with known eQTLs in both data sets and fitting both GLASSO and the new method, we measure biological relevance of the inferred graph by the enrichment of edges in Gene Ontology (GO) terms and the ability to predict the GO terms of node from its neighbours. We find that edges identified by the new method are six times more enriched in GO terms and twice as predictive as those inferred by GLASSO. Moreover, by analysing all 10,000 genes simultaneously, we demonstrate that the improved ability of the new method to model latent variables results in 12-fold greater consistency between the subset and full data analyses compared to GLASSO.

← Back

A PRIOR-BASED INTEGRATIVE FRAMEWORK FOR INFERRING TRANSCRIPTIONAL REGULATORY NETWORKS

Alireza Fotuhi Siahpirani[1,2], Sushmita Roy[1,3]

[1]University of Wisconsin, Madison, Wisconsin Institute for Discovery, Madison, WI, [2]University of Wisconsin, Madison, Computer Sciences, Madison, WI, [3]University of Wisconsin, Madison, Biostatistics and Medical Informatics, Madison, WI

Regulatory network inference is a long-standing problem in studies of gene regulation. Recent comparative analysis of expression-based network inference methods in eukaryotic systems shows that expression alone is weak predictor of regulatory edges. In this work we develop a novel probabilistic graphical model based network reconstruction algorithm that is able to integrate diverse types of regulatory evidences. Probabilistic graphical models provide a powerful framework to represent regulatory networks as they capture both the structure (who regulates whom) and the logic (parametric functions that define the regulatory relationships). We develop several computational metrics to assess network quality based on their agreement with known gold standards and predictive power of expression.

We define a prior distribution of the graph structure that is parametrized to incorporate different types of evidences that support the presence of a functional regulatory interaction. Our prior based framework specifies the prior probability of the presence of an edge based on these different evidences. The different types of regulatory evidences that can be integrated using our approach includes sequence specific motif instances on the promoter of a target gene, presence of a TF binding on the target gene, as well as genome-wide expression levels from TF knockouts. We combined these different sources of evidence into a single logistic prior value for the presence of an edge. Given the prior on the graph and measured expression data from a large number of biological samples, our goal is to find the most likely graph structure given the data. We apply our network inference algorithm within a stability selection framework; which is a sampling based approach to obtain confidence on the inferred edges by generating different subsamples from the data, learning separate model structures followed by aggregating the number of times edges have been seen in any subsample.

We assess the performance of our method on publicly available data sets from S.cerevisiae. We find that adding sequence motifs as prior greatly improves the quality of inferred networks. Furthermore, our networks are significantly better predictive of expression compare to purely motif-based network underscoring the importance of integrating sequence and expression-based datasets for inferring networks. Finally, we apply our approach to predict condition-specific regulatory networks and prioritize important regulators associated with stress response.

← Back

GENOME-WIDE PROBABILISTIC DYNAMICAL MODELLING OF TRANSCRIPTION KINETICS REVEALS PATTERNS OF RNA PRODUCTION DELAYS

Antti Honkela[1], Jaakko Peltonen[2,3], Hande Topa[2], Iryna Charapitsa[4], Filomena Matarese[5], Korbinian Grote[6], Hendrik G Stunnenberg[5], George Reid[4], Neil D Lawrence[7], Magnus Rattray[8]

[1]University of Helsinki,, Helsinki, Finland, [2]Aalto University,, Espoo, Finland, [3]University of Tampere,, Tampere, Finland, [4]Institute for Molecular Biology,, Mainz, Germany, [5]Radboud University Nijmegen,, Nijmegen, Netherlands, [6]Genomatix Software GmbH,, Muenchen, Germany, [7]University of Sheffield,, Sheffield, United Kingdom, [8]University of Manchester,, Manchester, United Kingdom

Genes with similar transcriptional activation kinetics can display very different temporal mRNA profiles due to differences in transcription time, degradation rate and RNA processing kinetics. Understanding these differences is essential for regulatory network inference, where incorrect models can lead to mis-attribution of some regulatory relationships. Recent studies have shown that a splicing-associated RNA production delay can be significant, but the prevalence of such delays is unknown.
We introduce a joint probabilistic dynamical model of transcriptional activation and mRNA accumulation which can be used for inference of transcription rate, RNA production delay and degradation rate given genome-wide data from high-throughput sequencing time course experiments. We combine a linear delay differential equation model with a non-parametric statistical Gaussian process modelling approach allowing us to capture a broad range of activation kinetics. The linearity of the differential equation model allows us to analytically derive a joint Gaussian process covariance for the continuous-time input and response functions, and to marginalise these functions out exactly. We use Bayesian Hamiltonian Monte Carlo sampling to obtain a full posterior distribution over all model parameters allowing us to quantify the uncertainty in estimates of the kinetic parameters.
We apply the model to data from estrogen receptor (ER-α) activation in the MCF-7 breast cancer cell line. We use RNA polymerase II (pol-II) ChIP-Seq time course data to characterise transcriptional activation and mRNA-Seq time course data to quantify mature transcripts. We find that 11% of genes with a good signal in the data display a delay of more than 20 minutes between completing transcription and mature mRNA production. The genes displaying these long delays are significantly more likely to be short. We also find a statistical association between high delay and late intron retention in pre-mRNA data, indicating significant splicing-associated production delays in many genes.

Pre-print with more details:
A. Honkela et al.
Genome-wide modelling of transcription kinetics reveals patterns of RNA production delays.
arXiv:1503.01081 [q-bio.GN]

← Back

MULTI-SCALE METHODS FOR DETECTING DIFFERENCES BETWEEN MULTIPLE GROUPS IN HIGH-THROUGHPUT SEQUENCING DATA AND THEIR APPLICATION TO SMALL SAMPLE SIZES

Heejung Shim[1], Zhengrong Xing[2], Ester Pantaleo[3], Francesca Luca[4,5], Roger Pique-Regi[4,5], Matthew Stephens[2,3]

[1]Purdue University, Department of Statistics, West Lafayette, IN, [2]University of Chicago, Department of Statistics, Chicago, IL, [3]University of Chicago, Department of Human Genetics, Chicago, IL, [4]Wayne State University, Center for Molecular Medicine and Genetics, Detroit, MI, [5]Wayne State University, Department of Obstetrics and Gynecology, Detroit, MI

Identification of differences between multiple groups in molecular phenotypes measured by high-throughput sequencing assays is frequently encountered in genomics applications. Common problems include eQTL mapping using RNA-seq and detecting differences in chromatin accessibility across conditions using ATAC-seq. Those high-throughput sequencing data provide high-resolution measurements on how traits vary along the whole genome in each sample. However, typical analyses fail to exploit the full potential of these high-resolution measurements, instead aggregating the data at coarser resolutions, such as genes, or windows of fixed length.

Previously, we developed wavelet-based methods, WaveQTL, that fully exploit the high-resolution data, and demonstrated the potential of WaveQTL over a simple window-based approach for associating genetic variants with sequence-based molecular phenotypes. We used a normal model for WaveQTL, being equivalent to assuming that read count at each base follow normal distribution, resulting in potential limitations at small sample sizes and for low library sequencing depths. Motivated by this, we have developed "multi-scale" methods, multiseq, that model "the count nature of the sequence data directly".

Applying multiseq to ATAC-seq data measured on three Copper treated and three control LCL samples to detect differences in chromatin accessibility, we find that multiseq performs well in small sample size compared to WaveQTL: at FDR=0.05 the number of signals detected by WaveQTL is less than 8% of that detected by mutiseq. We also show that two methods have different powers over multiple ranges of read counts: difference is more substantial for lower read counts. Further, comparing multiseq with DESeq2, commonly used window-based approach for count data, we demonstrate the substantial advantages of multiseq over the window method: at FDR=0.05 multiseq identified 23 times more differences than DESeq2 and, importantly, identified signals with different biological interpretation than those identified by DESeq2. In brief, the main advantage of multiseq is that due to "multi-scale" nature, multiseq easily captures signals varying in their scale (narrow or broad) while window-based methods are well powered to detect signals that occur on a single scale determined by the length of the window.

FROM GWAS TO EPIGENOME-WIDE ASSOCIATION STUDIES: CORRECTING FOR CONFOUNDING FACTORS

Jennifer Listgarten

Microsoft Research New England, Cambridge, MA

Understanding the genetic underpinnings of disease is important for screening, treatment, drug development, and basic biological insight. Genome-wide associations, wherein individual or sets of genetic markers are systematically scanned for association with disease are one window into disease processes. Naively, these associations can be found by use of a simple statistical test. However, a wide variety of confounders lie hidden in the data, leading to both spurious associations and missed associations if not properly addressed. These confounders include population structure, family relatedness, cell type heterogeneity, and environmental confounders. I will discuss the state-of-the art approaches for conducting these analyses, in which the confounders are automatically deduced, and then corrected for, by the data and model.

MODELING HIGH-DIMENSIONAL PHENOTYPES IN GENETICS : UNCOVERING BRAIN NETWORKS AND GENE NETWORKS

Jonathan Marchini, Victoria Hore, Lloyd Elliott

University of Oxford, Department of Statistics, Oxford, United Kingdom

Genetic association studies have yielded a wealth of biologic discoveries. However, these have been mostly carried out by analyzing one trait and one SNP at the time, thus often failing to capture the underlying complexity of traits. In order to move beyond simple GWAS, joint analysis of complex, high-dimensional phenotypes represents an important extension of phenotype-genotype associations with great potential. I will present two new methodologies for the analysis of very high-dimensional traits.

The first involves analysis of resting state fMRI brain imaging data as a phenotype. Such datasets, which are 4D in nature (3D brain location through time), are collected to investigate functional connectivity between different cortical regions. After pre-processing the data is reduced to a network for each individuals that measures the connection strength between regions. Interest then lies in understanding the genetic basis of these network connectivity matrices (or netmats). We have developed a Bayesian hierarchical model of connectivity that uses genetic kinship between individuals to help uncover heritable network substructure. Applied to data from the Human Connectome Project this model is able to uncover significant connectivity substructure between individuals.

Secondly, we have developed a general Bayesian framework for decomposing matrices and tensors of multi-tissue gene expression datasets into sparse latent factors, where latent factors consist of networks of co-varying genes. We use the individual scores vector of each factor (or component) as a phenotype in a GWAS to identify genetic variants that drive the gene network associated with that component. We have applied our method to data from the EuroBATS project which consists of RNA sequencing on 845 related individuals in adipose, skin and LCLs. Our method uncovers many sparse components that exhibit strong statistical and biological significance. These networks involve regulation of MHC Class I and Class II genes, Type I interferon response to virus, histone RNA processing, regulation of zinc finger gene clusters and genes involved in master regulator of gene expression in adipose tissue.

A TWO-WAY MIXED-MODEL METHOD FOR JOINT ASSOCIATION MAPPING, WITH APPLICATION IN *ARABIDOPSIS-XANTHOMONAS* PATHOSYSTEM

Miaoyan Wang[1], Hana Lee[2], Chris Meyers[2], Fabrice Roux[3], Joy Bergelson[2], Mary Sara McPeek[1]

[1]University of Chicago, Department of Statistics, Chicago, IL, [2]University of Chicago, Department of Ecology & Evolution, Chicago, IL, [3]INRA/CNRS Laboratory of Plant-Microbe Interactions, Institute National de la Recherche Agronomique, Toulouse, France

We consider the problem of how to test for genetic association and for gene-gene interaction in a genome-wide association study that involves two interactive organisms. Existing approaches to association mapping typically consider only a single organism type and aim to identify genes that are marginally associated with the phenotype of interest. However, in a host-pathogen interactive system, the response (e.g., infection) often depends on the specific pairing of host and pathogen. Conducting separate GWASs on host and on pathogen is less attractive than analyzing them jointly, because host-alone or pathogen-alone genetic attributes presumably represent only a fraction of the phenotypic variation, and may not unravel the full genetic architecture of a complex trait that depends on the specific pairing of both. In such cases, it is desirable to systematically model the trait and perform GWAS mapping jointly by taking advantage of the genome sequences available for both partners of the system.

Our solution is to develop a two-way mixed-effect model with multiple variance components. We use empirical genetic relatedness matrices (GRMs) to account for the background genetic heterogeneity due to the polygenic contributions from both organisms and their inter-genome interaction. In additional to the host GRM and the pathogen GRM, we choose to use a G * G interaction kernel matrix which we define to be a Hadamard (i.e., element-wise) product of these two GRMs, for the purpose of modeling inter-genome interaction. We propose various score tests, depending on the specific goal, for assessing genetic association effects of individual SNPs and/or SNP pairs. Our methods have been applied to the *Arabidopsis-Xanthomonas* dataset collected in Joy Bergelson's lab. In this context, we consider both the Gaussian and the Binomial-like two-way mixed-effect models using a quasi-likelihood framework. Because many SNP sites are present in only a subset of the sampled *Xanthomonas* strains, we extend the calculation of the empirical GRM for *Xanthomonas* to allow for this feature. We test for association of the trait with individual SNPs and with SNP pairs that include one SNP from each organism. We find that the infection response in the *Arabidopsis-Xanthomonas* pathosystem has a clear host-pathogen specificity, i.e., certain *Arabidopsis* SNPs exhibit strong genetic effects only when paired with certain *Xanthomonas* SNPs.

← Back

FALSE DISCOVERY RATE (FDR) METHODOLOGY, FIRST PUT FORWARD BY BENJAMINI AND HOCHBERG, AND FURTHER DEVELOPED BY MANY AUTHORS - INCLUDING STOREY, TIBSHIRANI, AND EFRON - IS NOW ONE OF THE MOST WIDELY USED STATISTICAL METHODS IN GENOMICS, AMONG OTHER AREAS OF APPLICATION

Matthew Stephens

University of Chicago,, Chicago, IL

A typical genomics workflow consists of i) estimating thousands of effects, and their associated p values; ii) feeding these p values to software (e.g. the widely used qvalue package) to estimate the FDR for any given significance threshold. In this talk we take a fresh look at this problem, and highlight two deficiencies of this standard pipeline that we believe could be improved. First, current methods, being based directly on p values (or z scores), fail to fully account for the fact that some measurements are more precise than others. Second, current methods assume that the least significant p values (those near 1) are all null - something that initially appears intuitive, but will not necessarily hold in practice. We suggest simple approaches to address both issues, and demonstrate the potential for these methods to increase the number of discoveries at a given FDR threshold. We also discuss the connection between this problem and shrinkage estimation, and problems involving sparsity more generally.

AN EFFICIENT LINEAR MIXED MODEL BASED ON MUTUAL INFORMATION FOR IDENTIFYING LOCI INVOLVED IN GENE-GENE AND GENE-ENVIRONMENT INTERACTIONS

Alexander I Young[1], Fabian Wauthier[1,2], Peter Donnelly[1,2]

[1]University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, [2]University of Oxford, Department of Statistics, Oxford, United Kingdom

There is a major open question as to how important gene-gene and gene-environment interaction effects are in the genetic architecture of human diseases and traits. The controversy remains unresolved partly due to a lack of powerful methods for detecting these effects and partly due to the lack of suitably sized datasets. The imminent availability of large population based studies, including biobanks, will for the first time offer the sample size required to properly address this question. The mutual information between a phenotype and a genetic variant is a general measure of their dependence, which captures all of the phenotypic information carried by the variant, not just the linear dependence. We develop a new test statistic based on the mutual information, which can be decomposed into the evidence for additive, dominance, and variance controlling effects. The variance contribution captures the marginal signal that the variant is involved in an interaction. We show theoretically that, under certain conditions, our test is the most powerful for detecting a locus with both additive and non-additive effects. Previous methods for detecting loci likely to be involved in interactions have not taken advantage of the increased power and control of population structure offered by mixed models. Our model incorporates a low rank random effect to adjust for population structure and polygenic additive effects. We provide a novel algorithm to fit the model whose core computations scale linearly with sample size, making it feasible to analyze biobank scale datasets. Furthermore, simulations on real genetic data show that our method has increased power over existing methods to detect loci involved in interactions. We also introduce a new visualisation tool, a generalization of the traditional Manhattan Plot, which stacks the evidence for additive, dominance and variance controlling effects respectively at each SNP. In particular, it highlights genomic regions with significant evidence for involvement in interactions. We apply our method to search for loci involved in gene-environment interactions affecting BMI in the UK Biobank, which has data on ~150,000 individuals genotyped at ~800,000 SNPs. We replicate the previously known variance controlling effect of the FTO locus on BMI, and we find evidence that a previously unreported locus without an additive effect on BMI is involved in gene-environment interactions.

A LATENT VARIABLE MODEL FOR SURVIVAL TIME PREDICTION WITH CENSORING AND DIVERSE PREDICTORS.

Shannon R McCurdy[1], Annette M Molinaro[2], Lior Pachter[3,4,5]

[1]UC Berkeley, California Institute for Quantitative Biosciences, Berkeley, CA, [2]UCSF, Department of Neurological Surgery, San Francisco, CA, [3]UC Berkeley, Electrical Engineering and Computer Science, Berkeley, CA, [4]UC Berkeley, Mathematics, Berkeley, CA, [5]UC Berkeley, Molecular and Cell Biology, Berkeley, CA

In the era of precision medicine, we are presented with a wealth of potential predictors for clinical outcomes such as survival time. These predictors can be genomic, epigenomic, clinical, and histopathological in nature, and additionally can be very high dimensional. Moreover, disease populations can be heterogeneous, with distinct clinical outcomes for subpopulations. This wealth of data presents a modeling challenge: what are sensible ways to combine and distill this wealth of data into a predictive model for a clinical outcome such as survival time?

We propose a latent variable model that combines factor analysis for various exponential family data types and a Cox Proportional Hazards model for continuous survival time and informative or non-informative censoring. The factor and Cox models are connected through low-dimensional latent variables. This dimensional reduction limits the number of model parameters. We also incorporate L_1 sparsity constraints for various model parameters. The expected values for the latent variables can be used for sample visualization. This model provides a predictive framework for survival time. We show simulation results.

NOTES

NOTES

NOTES

NOTES

NOTES

Participant List

Dr. Derek Aguiar
Princeton University
daguiar@cs.princeton.edu

Mr. Ali Akbari
UC San Diego
alek0991@gmail.com

Mr. Hussein Al-Asadi
University of Chicago
halasadi@uchicago.edu

Dr. Roee Amit
Technion
amitroee4@gmail.com

Dr. Georgios Athanasiadis
Aarhus University
athanasiadis@birc.au.dk

Mr. Joseph Azofeifa
University of Colorado - Boulder
joseph.azofeifa@colorado.edu

Dr. Brunilda Balliu
Stanford University School of Medicine
bballiu@stanford.edu

Dr. Mara Barucco
Università di Pisa
mara.baruc@gmail.com

Dr. Panayiotis (Takis) Benos
University of Pittsburgh
benos@pitt.edu

Mr. Jeremy Berg
University of California, Davis
jeremy.jackson.berg@gmail.com

Dr. Yael Berstein
Cold Spring Harbor Laboratory
yberstei@cshl.edu

Dr. Søren Besenbacher
Aarhus University
besenbacher@clin.au.dk

Dr. Anand Bhaskar
Stanford University
anand.bhaskar@gmail.com

Dr. Patrick Bradley
Gladstone Institutes @ UCSF
patrick.bradley@gladstone.ucsf.edu

Mr. Brielin Brown
UC Bekeley
brielin.brown@gmail.com

Dr. Brian Browning
University of Washington
browning@uw.edu

Ms. Nuala Caomhanach
American Museum of Natural History
ncaomhanach@amnh.org

Dr. John Capra
Vanderbilt University
tony.capra@vanderbilt.edu

Dr. Reed Cartwright
Arizona State University
cartwright@asu.edu

Mr. Mark Carty
Memorial Sloan Kettering Cancer Center
mac449@cornell.edu

Prof. Kevin Chen
Rutgers University
kcchen@dls.rutgers.edu

Dr. Chen Chen
Columbia University
cc3499@columbia.edu

Ms. Jade Cheng
Aarhus University
ycheng@cs.au.dk

Dr. Michele Clamp
Harvard University
mclamp@g.harvard.edu

Prof. Thomas Courtade
UC Berkeley
courtade@berkeley.edu

Dr. Jônatas da Silva César
University of São Paulo
jonataseduardo@gmail.com

Mr. Andrew Dahl
University of Oxford
andywdahl@gmail.com

Ms. Ana Damljanovic
Seven Bridges Genomics
ana.damljanovic@sbgenomics.com

Dr. Khanh Dao Duc
UPenn
daoduc@berkeley.edu

Mr. Gregory Darnell
Princeton University
gdarnell@princeton.edu

Dr. Laura DeMare
Genome Research/Molecular Case Studies
ldemare@cshl.edu

Dr. Fei Deng
University of California, Davis
feideng@ucdavis.edu

Mr. Kushal Dey
University of chicago
kshldey@gmail.com

Dr. Robin Dowell
University of Colorado
robin.dowell@colorado.edu

Mr. Noah Dukler
Cold Spring Harbor Labs
ndukler@cshl.edu

Ms. Bianca Dumitrascu
Princeton University
biancad@princeton.edu

Dr. Richard Durbin
The Wellcome Trust Sanger Inst.
rd@sanger.ac.uk

Prof. Barbara Engelhardt
Princeton University
bee@princeton.edu

Dr. Daniel Falush
University of Swansea
danielfalush@googlemail.com

Dr. Yanhui Fan
The University of Hong Kong
felixfanyh@gmail.com

Mr. Han Fang
Cold Spring Harbor Laboratory
hanfang.cshl@gmail.com

Mr. Pierre Faux
University of Liège
pierre.faux@ulg.ac.be

Dr. Nicole Figurro
hfjdjd
nicole.figueroa16@hotmail.com

Mr. Dmytro Fishman
University of Tartu
dmytrofishman@gmail.com

Dr. Barrett Foat
Monsanto Company
barrett.foat@monsanto.com

Mr. Chuan Sheng Foo
Stanford University
csfoo@cs.stanford.edu

Mr. Alireza Fotuhi Siahpirani
University of Wisconsin-Madison
fotuhisiahpi@wisc.edu

Dr. Anna Fowler
Oxford University
afowler@well.ox.ac.uk

Mr. Benjamin Frot
University of Oxford
benjamin.frot@stats.ox.ac.uk

Dr. Audrey Fu
University of Idaho
audreyqyfu@gmail.com

Mr. Mariano Gabitto
Columbia University
mig2118@columbia.edu

Mr. Erik Garrison
Wellcome Trust Sanger Institue
eg10@sanger.ac.uk

Prof. Anthony Gitter
University of Wisconsin-Madison
gitter@biostat.wisc.edu

Ms. Genna Gliner
Princeton University
genna@princeton.edu

Ms. Amy Goldberg
Stanford University
agoldb@stanford.edu

Dr. Anna Goldenberg
SickKids Research Institute/University of Toronto
anna.goldenberg@utoronto.ca

Dr. Steve Goldstein
University of Wisconsin - Madison
sgoldstein@wisc.edu

Ms. Peyton Greenside
Stanford University
pgreens@stanford.edu

Dr. Ilan Gronau
Herzliya Interdiscipinary Center (IDC)
ilan.gronau@idc.ac.il

Dr. Brad Gulko
Cornell University/CSHL
bgulko@cshl.edu

Dr. Kelley Harris
Stanford University
kelleyh@stanford.edu

Dr. Matthew Hartfield
University of Toronto
matthew.hartfield@utoronto.ca

Dr. David Haussler
University of California, Santa Cruz Genomics Institute
haussler@soe.ucsc.edu

Dr. Glenn Hickey
UCSC
glenn.hickey@gmail.com

Dr. Antti Honkela
University of Helsinki
antti.honkela@hiit.fi

Dr. Danilo Horta
EMBL-EBI
horta@ebi.ac.uk

Dr. Chiaowen Hsiao
The University of Chicago
joyce.hsiao1@gmail.com

Dr. Yifei Huang
Cold Spring Harbor Laboratory
yihuang@cshl.edu

Dr. Qiang Huang
Columbia University
qh2159@cucm.columbia.edu

Ms. Melissa Hubisz
Cold Spring Harbor Laboratory
mhubisz@cshl.edu

Dr. Iuliana Ionita-Laza
Columbia University
iulianalaza@icloud.com

Mr Johnny Israeli
Stanford University
johnnyisraeli@gmail.com

Prof. Sorin Istrail
Brown University
sorin_istrail@brown.edu

Dr. Yuval Itan
The Rockefeller University
yitan@rockefeller.edu

Mr. Siddhartha Jain
Carnegie Mellon University
pcptheorem@gmail.com

Dr. Ethan Jewett
UC, Berkeley
ejewett@gmail.com

Dr. Yumi Jin
NIH/NLM/NCBI
jinyu@ncbi.nlm.nih.gov

Mr. Brian Jo
Princeton University
bj5@princeton.edu

Dr. Michael Jordan
Univeraity of California, Berkeley
jordan@stat.berkeley.edu

Mr. Jin Hyun Ju
Weill Cornell Graduate School
jij2009@med.cornell.edu

Mr. John Kamm
UC Berkeley
jkamm@stat.berkeley.edu

Ms. Irene Kaplow
Stanford University
ikaplow@stanford.edu

Mr. Michael Karcher
University of Washington
michael.d.karcher@gmail.com

Mr. Keffy Kehrli
Stony Brook University
keffy.kehrli@stonybrook.edu

Dr. Jerome Kelleher
University of Oxford
jerome.kelleher@well.ox.ac.uk

Dr. Andrew Kern
Rutgers Univeristy
kern@biology.rutgers.edu

Mr. Christopher Koch
University of Wisconsin-Madison
ckoch3@wisc.edu

Mr. Evan Koch
University of Chicago
emkoch@uchicago.edu

Dr. Anders Krogh
University of Copenhagen
krogh@binf.ku.dk

Dr. Anshul Kundaje
Stanford University
akundaje@stanford.edu

Dr. Sofia Kyriazopoulou-Panagiotopoulou
10X Genomics
sofia@10xgenomics.com

Mr. Young-suk Lee
Princeton University
youngl@princeton.edu

Ms. Hayan Lee
Cold Spring Harbor Laboratory
hlee@cshl.edu

Dr. Jennifer Listgarten
Microsoft Research New England
jennl@microsoft.com

Dr. Tzu-Yu Liu
University of Pennsylvania
liutzuyu@gmail.com

Dr. Po-Ru Loh
Harvard T.H. Chan School of Public Health
loh@hsph.harvard.edu

Ms. Mengyin Lu
University of Chicago
mengyinlu228@gmail.com

Ms. Mengyin Lu
University of Chicago
mengyinlu228@gmail.com

Dr. Gerton Lunter
University of Oxford
gerton.lunter@well.ox.ac.uk

Dr. Jian Ma
University of Illinois at Urbana-Champaign
jianma@illinois.edu

Dr. Thomas MacCarthy
Stony Brook University
thomas.maccarthy@stonybrook.edu

Mr. Fabrizio Mafessoni
Max Planck for Evolutionary Anthropology
fabrizio_mafessoni@eva.mpg.de

Dr. Thomas Mailund
Aarhus University
mailund@birc.au.dk

Dr. Anna-Sapfo Malaspinas
University of Bern
annasapfo@gmail.com

Mr. Daniel Malmer
University of Colorado, Boulder
daniel.malmer@colorado.edu

Dr. Jonathan Marchini
University of Oxford
marchini@stats.ox.ac.uk

Mr. Joseph Marcus
University of Chicago
jhmarcus@uchicago.edu

Mr. Lasse Maretty
University of Copenhagen
lassemaretty@binf.ku.dk

Ms. Lenka Matejovicova
IST Austria
lenka.matejovicova@gmail.com

Dr. Shannon McCurdy
UC Berkeley
smccurdy@berkeley.edu

Dr. Philipp Messer
Cornell University
philipp.messer@gmail.com

Mr. Aziz Mezlini
University of Toronto
aziz.mezlini@utoronto.ca

Dr. Emily Miraldi
Simons Foundation, NYU School of Medicine, NYU
emiraldi@nyu.edu

Mr. Jaaved Mohammed
CSHL
jm889@cornell.edu

Dr. Gota Morota
University of Nebraska–Lincoln
morota@unl.edu

Mr. Hakhamanesh Mostafavi
Columbia University
hsm2137@columbia.edu

Dr. Sara Mostafavi
University of British Columbia
mostafavi.sara@gmail.com

Dr. Sayan Mukherjee
Duke University
sayan@stat.duke.edu

Dr. Swagatam Mukhopadhyay
Human Longevity, Inc.
smukhopadhyay@humanlongevity.com

Dr. Senthil Kumr Muthiah
University of Maryland, College Park
smuthiah@umiacs.umd.edu

Ms. Priyanka Nakka
Brown University
priyanka_nakka@brown.edu

Dr. Maria Nattestad
Cold Spring Harbor Laboratory
mnattest@cshl.edu

Dr. Rasmus Nielsen
University of California, Berkeley
rasmus@binf.ku.dk

Mr. Svend Nielsen
Aarhus University
svn@cs.au.dk

Mr. Adam Novak
UC Santa Cruz Genomics Institute
anovak@soe.ucsc.edu

Dr. John Novembre
University of Chicago
jnovembre@uchicago.edu

Dr. Benedikt Obermayer
Max-Delbrück-Center for molecular Medicine
benedikt.obermayer@mdc-berlin.de

Mr. Diego Ortega Del Vecchyo
University of California, Los Angeles
diegoortega@ucla.edu

Prof. Lior Pachter
UC Berkeley
lpachter@math.berkeley.edu

Dr. Julia Palacios Roman
Harvard University and Brown University
julia.pal.r@gmail.com

Dr. Pier Francesco Palamara
Harvard University
ppalama@hsph.harvard.edu

Dr. Benedict Paten
UCSC
benedict@soe.ucsc.edu

Dr. Vipul Periwal
NIDDK
vipulp@niddk.nih.gov

Mr. Malcolm Perry
Imperial College London
m.perry13@imperial.ac.uk

Dr. Benjamin Peter
University of Chicago
benj.pet@gmail.com

Mr. Stephane Peyregne
Max Planck Institute for Evolutionary Anthropology
stephane_peyregne@eva.mpg.de

Ms. Tanya Phung
University of California, Los Angeles
tnphung@ucla.edu

Dr. Roger Pique-Regi
Wayne State University
rpique@wayne.edu

Mr. Milos Popovic
Seven Bridges Genomics
milos.popovic@sbgenomics.com

Dr. Gerald Quon
MIT
gerald.quon@gmail.com

Dr. Goran Rakocevic
Seven Bridges Genomics
goran.rakocevic@sbgenomics.com

Dr. Duncan Ralph
fhcrc
dkralph@gmail.com

Dr. Sohini Ramachandran
Brown University
sramachandran@brown.edu

Ms. Monica Ramstetter
Cornell University
kameradin@gmail.com

Dr. Pradipta Ray
The University of Texas at Dallas
pradiptaray@utdallas.edu

Dr. Mark Reppell
University of Chicago
mreppell@uchicago.edu

Dr. Matthew Reyna
Brown University
matthew_reyna@brown.edu

Dr. Sylvia Richardson
MRC Biostatistics Unit
sylvia.richardson@mrc-bsu.cam.ac.uk

Dr. Maximo Rivarola
National Institute of Agronomical Sciences
rivarola.maximo@inta.gob.ar

Ms. Nicola Roberts
Wellcome Trust Sanger Institute
nr3@sanger.ac.uk

Mr. Stephen Rong
Brown University
stephen_rong@brown.edu

Mr. Andrew Roth
BC Cancer Agency Research Centre
andrewjlroth@gmail.com

Dr. Sushmita Roy
UW Madison
sroy@biostat.wisc.edu

Dr. Daniel Runcie
University of California Davis
deruncie@ucdavis.edu

Dr. Aaron Sams
Cornell University
as2847@cornell.edu

Dr. Christopher Saunders
Illumina
csaunders@illumina.com

Mr. Dominik Schrempf
Institute of Population Genetics, Vetmeduni Vienna
dominik.schrempf@gmail.com

Mr. Andrew Sedgewick
University of Pittsburgh
ajsedgewick@gmail.com

Dr. Fritz Sedllazeck
Cold Spring Harbor Laboratory
fritz.sedlazeck@gmail.com

Dr. Kevin Sharp
University of Oxford
sharp@stats.ox.ac.uk

Dr. Vladimir Shchur
The Wellcome Trust Sanger Institute
vs3@sanger.ac.uk

Dr. Heejung Shim
Purdue University
hjshim@gmail.com

Dr. Ilan Shomorony
UC Berkeley
ilanshom@gmail.com

Ms. Avanti Shrikumar
Stanford University
avanti.shrikumar@gmail.com

Dr. Suyash Shringarpure
Stanford University
suyashs@stanford.edu

Mr. Jonas Sibbesen
University of Copenhagen
jasi@binf.ku.dk

Prof. Adam Siepel
Cold Spring Harbor Laboratory
asiepel@cshl.edu

Dr. Mona Singh
Princeton University
mona@cs.princeton.edu

Mr. Joel Smith
University of Chicago
joelsmith@uchicago.edu

Dr. Yun Song
University of California, Berkeley
yss@berkeley.edu

Dr. Samuele Soraggi
University of Copenhaghen
samuele@math.ku.dk

Dr. Vitor Sousa
University of Bern
vitor.sousa@iee.unibe.ch

Prof. Matthias Steinruecken
University of Massachusetts, Amherst
steinruecken@schoolph.umass.edu

Dr. Matthew Stephens
University of Chicago
stephens999@gmail.com

Dr. Lauren Sugden
Brown University
lauren_alpert@brown.edu

Mr. Lei Sun
University of Chicago
lsunchina@gmail.com

Mr. Jonathan Terhorst
UC Berkeley
terhorst@stat.berkeley.edu

Dr. Elizabeth Thompson
University of Washington
eathomp@uw.edu

Dr. Elizabeth Thompson
University of Washington
eathomp@uw.edu

Dr. Olga Troyanskaya
Princeton University
ogt@princeton.edu

Mr. Robert Tunney
UC Berkeley
robert.tunney@gmail.com

Mr. Michael Turchin
University of Chicago
mturchin20@uchicago.edu

Ms. Sarah Urbut
University of Chicago
surbut@uchicago.edu

Mr. Wei Wang
University of Chicago
wei.wang0610099@gmail.com

Ms. Miaoyan Wang
University of Chicago
miaoyan@galton.uchicago.edu

Dr. Gao Wang
University of Chicago
wangow@gmail.com

Dr. Doreen Ware
Cold Spring Harbor Laboratory USDA ARS
ware@cshl.edu

Dr. Jonathan Warrell
Council for Scientific and Industrial Research
jonathan.warrell@gmail.com

Ms. Melanie Weber
Cold Spring Harbor Laboratory
mweber@cshl.edu

Dr. Amy Williams
Cornell University
alw289@cornell.edu

Mr. Peter Wilton
Harvard University
peterrwilton@gmail.com

Mr. Zhengrong Xing
University of Chicago
zhengrong@galton.uchicago.edu

Mr. Jason Xu
University of Washington
jasonxu@uw.edu

Dr. Koon-Kiu Yan
Yale University
koon-kiu.yan@yale.edu

Mr. Shuo Yang
Columbia University
shuo@cs.columbia.edu

Dr. Joy Yang
Massachusetts Institute of Technology
yangjy@mit.edu

Mr. Alexander Young
University of Oxford
ay@well.ox.ac.uk

Ms. Jane Yu
University of California, Berkeley
janeyu@eecs.berkeley.edu

Dr. Yiqiang Zhao
China Agricultural University
yiqiangz@cau.edu.cn

Mr. Xiang Zhu
University of Chicago
xiangzhu@uchicago.edu

Dr. Sha Zhu
University of Oxford
joe.zhu@well.ox.ac.uk

Dr. James Zou
MIT and Microsoft Research
jzou@fas.harvard.edu

VISITOR INFORMATION

EMERGENCY	CSHL	BANBURY
Fire	**(9) 742-3300**	**(9) 692-4747**
Ambulance	**(9) 742-3300**	**(9) 692-4747**
Poison	**(9) 542-2323**	**(9) 542-2323**
Police	**(9) 911**	**(9) 549-8800**
Safety-Security	**Extension 8870**	

Emergency Room **Huntington Hospital** 270 Park Avenue, Huntington	**631-351-2300** **(1037)**
Dentists Dr. William Berg Dr. Robert Zeman	 **631-271-2310** **631-271-8090**
Doctor MediCenter 234 W. Jericho Tpke., Huntington Station	**631-423-5400** (**1034**)
Drugs - 24 hours, 7 days Rite-Aid 391 W. Main Street, Huntington	**631-549-9400** (**1039**)

Free Speed Dial
Dial the four numbers (********) from any **tan house phone** to place a free call.

GENERAL INFORMATION

Books, Gifts, Snacks, Clothing, Newspapers
BOOKSTORE 367-8837 (hours posted on door)
Located in Grace Auditorium, lower level.

Photocopiers, Journals, Periodicals, Books, Newspapers
Photocopying – Main Library
Hours: 8:00 a.m. – 9:00 p.m. Mon-Fri
10:00 a.m. – 6:00 p.m. Saturday
***Helpful tips* – Use PIN# 62435** to enter Library after hours.
See Library staff for photocopier code.

Computers, E-mail, Internet access
Grace Auditorium
Upper level: E-mail and printing
STMP server address: mail.optonline.net
To access your E-mail, you must know the name of your home server.

Dining, Bar
Blackford Hall
Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00
Bar 5:00 p.m. until late (Cash Only)
Helpful tip - If there is a line at the upper dining area, try the lower dining room

Messages, Mail, Faxes, ATM
Message Board, Grace, lower level

Swimming, Tennis, Jogging, Hiking

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m. Two tennis courts open daily.

Russell Fitness Center

Dolan Hall, east wing, lower level
PIN#: **Press 62435 (then enter #)**

Concierge

On duty daily at Meetings & Courses Office.
After hours – From tan house phones, dial x8870 for assistance

Pay Phones, House Phones

Grace, lower level; Cabin Complex; Blackford Hall; Dolan Hall, foyer

CSHL's Green Campus

Cold Spring Harbor Laboratory is pledged to operate in an environmentally responsible fashion wherever possible. In the past, we have removed underground oil tanks, remediated asbestos in historic buildings, and taken substantial measures to ensure the pristine quality of the waters of the harbor. Water used for irrigation comes from natural springs and wells on the property itself. Lawns, trees, and planting beds are managed organically whenever possible. And trees are planted to replace those felled for construction projects.

Two areas in which the Laboratory has focused recent efforts have been those of waste management and energy conservation. The Laboratory currently recycles most waste. Scrap metal, electronics, construction debris, batteries, fluorescent light bulbs, toner cartridges, and waste oil are all recycled. For general waste, the Laboratory uses a "single stream waste management" system, removing recyclable materials and sending the remaining combustible trash to a cogeneration plant where it is burned to provide electricity, an approach considered among the most energy efficient, while providing a high yield of recyclable materials.

Equal attention has been paid to energy conservation. Most lighting fixtures have been replaced with high efficiency fluorescent fixtures, and thousands of incandescent bulbs throughout campus have been replaced with compact fluorescents. The Laboratory has also embarked on a project that will replace all building management systems on campus, reducing heating and cooling costs by as much as twenty-five per cent.

Cold Spring Harbor Laboratory continues to explore new ways in which we can reduce our environmental footprint, including encouraging our visitors and employees to use reusable containers, conserve energy, and suggest areas in which the Laboratory's efforts can be improved. This book, for example, is printed on recycled paper.

1-800 Access Numbers

AT&T	**9-1-800-321-0288**
MCI	**9-1-800-674-7000**

Local Interest

Fish Hatchery	631-692-6768
Sagamore Hill	516-922-4447
Whaling Museum	631-367-3418
Heckscher Museum	631-351-3250
CSHL DNA Learning Center	x 5170

New York City

Helpful tip -

Take Syosset Taxi to Syosset Train Station ($9.00 per person, 15 minute ride), then catch Long Island Railroad to Penn Station (33rd Street & 7th Avenue). Train ride about one hour.

TRANSPORTATION

Limo, Taxi

Syosset Limousine	516-364-9681 (**1031**)
US Limousine Service	800-962-2827,ext:3 **(1047)**
Super Shuttle	800-957-4533 (**1033**)
To head west of CSHL - Syosset train station	
Syosset Taxi	516-921-2141 (**1030**)
To head east of CSHL - Huntington Village	
Orange & White Taxi	631-271-3600 (**1032**)

Trains

Long Island Rail Road	822-LIRR
Schedules available from the Meetings & Courses Office.	
Amtrak	800-872-7245
MetroNorth	800-638-7646
New Jersey Transit	201-762-5100

Ferries

Bridgeport / Port Jefferson	631-473-0286 **(1036)**
Orient Point/ New London	631-323-2525 **(1038)**

Car Rentals

Avis	631-271-9300
Enterprise	631-424-8300
Hertz	631-427-6106

Airlines

American	800-433-7300
America West	800-237-9292
British Airways	800-247-9297
Continental	800-525-0280
Delta	800-221-1212
Japan Airlines	800-525-3663
Jet Blue	800-538-2583
KLM	800-374-7747
Lufthansa	800-645-3880
Northwest	800-225-2525
United	800-241-6522
US Airways	800-428-4322