# "Evolutionary Biology & the Theory of Computing" (Spring 2014) Final Program Report

Yun S. Song (Organizing Chair)

## Overall objectives and assessment of the program

Evolutionary biology is an intellectually rich field with a long history which has advanced remarkably through a synergistic interplay between deep understanding of biology and mathematical techniques, especially from probability and statistics. Over the past several decades, the role of computer science in studying biology has grown enormously, and computation has now become an indispensable part of the intellectual mix. Many current problems in evolutionary biology push the limits of computation, and new algorithmic insights are needed to make progress.

The main objectives of the Simons Institute program on "Evolutionary Biology & the Theory of Computing" were twofold:

1. To promote the interaction between theoretical computer scientists and researchers from the evolutionary biology, physics, and probability and statistics communities.

2. To encourage the participants to collaborate on identifying and tackling some of the most important theoretical and computational challenges arising from evolutionary biology.

One important tactical goal of the program was to provide theoretical computer scientists with the opportunity to understand the fundamental concepts and key questions in evolutionary biology.

This was an ambitious program that aimed to bring together experts from diverse disciplines and encourage interaction. The initial language barrier was not a major problem; a much bigger challenge was bridging the gap between disparate research interests to find common goals. Most mathematicians, statisticians and physicists participating in the program already had experience of collaborating fruitfully with evolutionary biologists. However, the cultural difference between theoretical computer scientists and the rest of the program participants was bigger than anticipated. Theoretical computer scientists prefer to work with simple models that can be understood in detail and for which they can prove rigorous theorems which may or may not generalize to more complex models. In contrast, evolutionary biologists put much more emphasis on biologically realistic models and inference methods for analyzing data. Another observed division was at the level of modeling. Specifically, most evolutionary biologists were focused on understanding the evolutionary mechanisms underlying population genetic variation, and the genetic basis of phenotypic traits and adaptation to new environments. On the other hand, a large fraction of theoretical computer scientists were more interested in understanding how systems evolve. Realizing these differences is a necessary step towards bringing the two communities closer, and the Simons Institute program provided a valuable opportunity for each community to become more aware of the other group's views and interests.

Despite the challenges just described, all participants found the program to be thought-provoking, and it was largely successful at meeting the aforementioned main objectives. The program's Research Fellows played a pivotal role in giving life to the program throughout the semester. Also, several junior theoretical computer scientists — especially Varun Kanade, Paul Valiant and

Greg Valiant — deserve credit for taking part in many of the activities described below and trying to reach out to the participants from other disciplines.

The reunion workshop (held July 27-29, 2015) provided an excellent opportunity to reflect on the program. In particular, there was a panel discussion to highlight some of the key open problems in theoretical biology, and to discuss how to bridge better the existing gap between biology and theoretical computer science. Paul Valiant voiced the view that the biology community is inward-looking and that it has largely ignored previous biology-related work by theoretical computer scientists. To increase impact and visibility, it was suggested that theoretical computer scientists should try to work on problems that biologists care about, and to publish in biology journals and have their papers peer reviewed by biologists. Furthermore, trying to tackle narrow, well-defined problems rather than aiming at overly general, ambitious goals might be a promising way to bring the two communities closer in the immediate future. Lastly, Chuck Langley pointed out that biology is currently awash with data, often of poor quality, and that biologists are struggling to process and make sense of it. Theoretical investigations that address this issue would be most welcome, as would work that characterizes the fundamental limits of what can be learned from data.

## Program activities and range of themes covered

The program started with a week-long Boot Camp, with introductory lectures given by theoretical computer scientists, biologists, mathematicians and physicists. A wide range of topics were covered, touching on the key themes of the program. Monty Slatkin's account of the historical development of evolutionary biology and Charles Marshall's lectures on "The Origin and Evolution of Life on Earth" were particularly well received. Christos Papadimitriou and Varun Kanade also gave excellent lectures on "Computational Views of Evolution" and "Evolution as Computational Learning" respectively; these lectures helped to clarify to the scientists from other disciplines the key questions in evolutionary biology that are of interest to theoretical computer scientists and how they go about thinking about them.

There were three workshops associated with the program. The first workshop was centered on statistical inference methods and computational challenges in large-scale population genomics in light of the recent explosion of DNA sequence data. The second workshop was closer to theoretical computer science, showcasing the key models and theories of evolution inspired by computational considerations, as well as highlighting research questions in evolutionary biology which might benefit from computational insights and methodology. The third workshop focused on new directions in probabilistic models of evolution, addressing a broad set of topics including the evolution of diseases and pathogens. Details on the outcomes of these workshops are provided in separate reports.

Three weekly activities were organized throughout the program to facilitate the convergence of backgrounds and research interests. A seminar series was held on Tuesdays, during which participants from diverse areas gave talks on their research. Every Thursday afternoon, there was an informal discussion session on "Ideas and Problems" related to evolutionary biology. Lastly, every Friday afternoon, there was a reading group dedicated to discussing classic and recent papers relevant to the main themes of the program. These activities encouraged participants to interact with each other on a daily basis, exchange ideas, and help each other learn complementary subjects.

## Notable successes of the program

There were several notable research results that arose during and as a direct consequence of the interaction between the participants of the program, a few of which are highlighted below. In the following summaries, program participant names are shown in boldface when mentioned for the first time.

*The hypercycle and speed of evolution:* In 1978 Manfred Eigen along, with Peter Schuster, proposed a theory for the chemical origin of life — the hypercycle. The hypercycle is a continuous-time dynamical system where the rate of change of a type in a population depends not only on its concentration in the population, but also on a set of predecessors, which makes the equations have degree two or higher. Surprisingly, such dynamical systems neither have stable fixed points nor are they chaotic when there are more than four types. It was conjectured by Eigen and Schuster, and later proved by Hofbauer et al., that for more than four types, the dynamical system converges to a limit cycle in the interior of the simplex. This "existence" theorem is rather remarkable as it implies that the dynamical system is "well-behaved," since there is chaos in three or more dimensions. Indeed, the proof relies on the famous Poincaré-Bendixson theorem, which rules out chaos in two dimensions, implying that the dynamical systems arising in the hypercycle are effectively two-dimensional. The natural question about the time it takes to converge to the limit cycle remained a gaping hole in this literature for a good reason: we had no computational understanding of the Poincaré-Bendixson theorem. During the program, **Nisheeth Vishnoi** (EPFL) and **Christos Papadimitriou** (UC Berkeley) collaborated to investigate the computational complexity of the Poincaré-Bendixson theorem, and have resolved the computational complexity of this problem [81]. They also introduced the notion of an "approximate cycle" and proved an approximate Poincaré-Bendixson theorem guaranteeing that some orbits come very close to forming a cycle in the absence of approximate fixed points, a surprising fact that holds in all dimensions. However, the original question that they started with — i.e., can the limit cycle of the hypercycle (that is to say, life!) be approached in polynomial time? — remains open.

In the bigger scheme of understanding evolution and how life could have originated, Vishnoi has also been working on understanding the mixing time of stochastic evolutionary dynamics in finite populations. Such processes lie at the core of evolution and in the recent past (e.g., the stochastic version of Eigen's quasispecies model) have been used to model viral populations from the viewpoint of mutagenic drug design. Here, the time it takes for the population to reach a steady state is important both for the estimation of the steady-state structure of the population, as well as for determining the duration and strength of drug treatment. During the program, Vishnoi completed an important first paper on this problem where he proved that in the case of two genotypes, the underlying Markov chain mixes rapidly. More importantly, he made a connection with a discrete time dynamical system that is interesting in its own right [111]. To make progress on the general problem of proving rapid mixing when there are more than two genotypes, Vishnoi and graduate student **Piyush Srivastava** (UC Berkeley) collaborated extensively during the program, and developed a series of observations and understood obstacles in extending Vishnoi's previous results. Following the program, along with Ioannis Panageas, they were able to make use of the observations made during the program to prove a rapid mixing result for a broad class of such dynamics, thus resolving the central problem [79]. Technically, their result relies on a novel connection between Markov chains arising in such evolutionary dynamics and dynamical systems on the simplex. More generally, their result sheds light on how quickly life could have evolved.

*Consistency of phylogenetics methods:* Several theoretical problems in phylogenetics were fruit-

fully tackled during the program. One particularly important problem is reconstructing the ancestral state of a phylogenetic tree given the information about the state at the leaves of the tree. There are two very natural algorithms for accomplishing this: (i) maximum parsimony, which minimizes the number of state changes required to explain the data; and (ii) majority rule, which simply picks the majority among the states at the leaves of the tree. As one can construct specific trees on which one method outperforms the other, it would be useful to study the performance of these methods on random trees generated under widely-used evolutionary models. **Mike Steel** (University of Canterbury, New Zealand) spent a very productive month at the Institute where he initiated, completed, and submitted a paper with **Elchanan Mossel** (UC Berkeley) which addressed this problem. Using tools from probability theory such as coupling and the reflection principle, they were able to show that the majority rule is more accurate than maximum parsimony at reconstructing the state of an ancestor when evolutionary trees are drawn from the Yule model [73]. Steel also initiated and completed a new research project with **Sebastien Roch** (University of Wisconsin) on the consistency of multi-locus methods in phylogenetics [96]. More specifically, they studied a widely-used method known as *concatenation* which involves, as the name suggests, concatenating the sequences from a number of genes into a super-gene and pretending as if all the nucleotides in the entire super-gene arose independently according to a mutational process on a single phylogenetic tree. They showed formally that this procedure can lead to serious statistical issues. In particular, they proved rigorously that maximum likelihood estimation on concatenated data can be guaranteed to reconstruct an erroneous evolutionary history. Roch also initiated a new collaboration with **Constantinos Daskalakis** (MIT) on the effect of lateral gene transfer (LGT) in phylogenomic studies. LGT is problematic because it introduces cross-edges in what would otherwise be a tree structure across species, making the reconstruction of phylogenetic trees significantly more challenging. In previous work with Sagi Snir, Roch had shown that, under the assumption that LGT occurs at random along the tree of life, one can still recover the "tree signal" from the data in the presence of high levels of LGT. Daskalakis and Roch improved this result during the program by obtaining matching (up to constants) upper and lower bounds on the amount of LGT that is tolerable for reconstruction, and a manuscript on this work is in preparation [26]. The program also gave Roch a chance to complete an earlier project with Mossel on studying the trade-off between the number of loci and the length of each locus that is necessary to reliably reconstruct a species tree from a given fixed amount of sequence data. They showed that for a fixed amount of sequence data, it is always better to have a large number of short genes rather than a smaller number of longer genes [72].

*Demographic inference from allele frequency data:* One of the fundamental problems in evolutionary biology involves understanding the impact of population demography on the distribution of allele frequencies in the population. Much research activity is currently centered on inferring the population demography from the frequency of alleles in large genome samples drawn from the population. Despite the popularity of frequency spectrum-based inference methods, currently little is known about the information-theoretic limit on the estimation accuracy as a function of sample size. While previous work [8] — by program organizer **Yun Song** and Research Fellow **Anand Bhaskar** (UC Berkeley), revised during the program — has shown that one can uniquely recover details of the historical population demography given perfect allele frequency information (which is akin to having data from infinitely many sites in the genome), in practice, the finite length of the genome introduces sampling variance that can make it difficult to precisely infer details about historical population demography solely based on allele frequency data. To investigate this issue theoretically, Yun Song and program participant **Jonathan Terhorst** (UC Berkeley) initi-

ated a project during the program and showed that allele frequency data are not very informative about the deep history of populations [104]. In particular, if the population size has undergone a constriction in the past, say due to a migration bottleneck, the minimax error in estimating the historical population size at times more ancient than the bottleneck is at least $O(1/\log s)$, where $s$ is the number of independent polymorphic sites used in the analysis. This rate is exponentially worse than known convergence rates for many classical estimation problems in statistics. Another surprising aspect of their theoretical bound is that it does not depend on the number of sampled individuals. This means that, for a fixed number $s$ of polymorphic sites considered, using more individuals does not help to reduce the minimax error bound. Also during the program, Anand Bhaskar, Yun Song and Sebastian Roch initiated a collaboration to study the geometry of the distribution of allele frequencies in a genomic sample and characterize its dependence on the underlying population demographic model. They have several novel and unexpected results about the limits of demographic inference from allele frequency data that apply to any inference algorithm.

*Assortative mating in human populations:* A standard assumption in most population genetic analyses is that populations are well-mixed and individuals mate randomly. While this assumption is made for mathematical and computational convenience, little work has been done to study the extent to which this violation is violated in practice. During the program, Research Fellows **James Zou** (Harvard University) and **Sriram Sankararaman** (Harvard University) initiated a collaboration and realized that publicly available genomic data could be used to answer this question. Initially they found that, in admixed populations such as African-Americans and Latinos, the maternal and paternal genomes of an individual are significantly more similar than that of random couples. To understand this observation, they needed to infer the ancestries of the parents of an individual from the genotype of the individual. These ancestries can then be used to quantify the propensity for assortative (i.e., non-random) mating and to identify genetic loci that could mediate these patterns. Together with long-term participant **Eran Halperin** (Tel Aviv University), they developed a statistical model and method to estimate the genome-wide ancestral contributions of each parent of an admixed individual from the individual's genomic data [122]. The statistical model employed in this work consists of a pooled semi-Markov process and is related to factorial hidden Markov models. There are interesting statistical questions about efficient inference in these models, since the combinatorial constraints of the pooling make standard variational inference inapplicable.

To apply this model to better understand genomic data, they collaborated with groups at UCSF who had collected genotype and socio-economic data of Mexican and Puerto Rican individuals. By jointly analyzing the socio-economic and genomic data, they have been able to infer the relative contributions of genetic vs. socio-economic factors to non-random mating, and to identify specific subsets of the genome that are associated with these patterns of assortative mating [123]. They have found that genomic ancestry is a major factor in determining mating patterns, much more so than education level and other socio-economic factors. This project also involved collaboration with long-term participants Yun Song and Eran Halperin.

This line of work has helped to characterize the extent of non-random or assortative mating in human populations and has clearly underlined the importance of moving beyond the traditional assumptions of random mating in population genetics models. Taking this thread of research further, Halperin and Noah Zaitlen (UCSF) have been developing population genetic theory that can better account for assortative mating. In particular, they have found that estimation of parameters such as migration rates, recombination rates, and the dates of admixture are all affected by assortative mating, and they have derived analytic formulas to infer these parameters from sequence data under assortative mating models.

## Other research highlights of the program

**Nayantara Bhatnagar** (University of Delaware) initiated a collaboration with Erick Matsen and Robert Bradley from the Fred Hutchinson Cancer Research Center [10]. They are looking at statistical barriers to sequence alignment, in particular giving statistical explanations for the barrier encountered by most commonly used alignment programs in the "twilight zone" of sequence identity.

Research Fellow **Iain Mathieson** (Harvard) started a collaboration with Ken Wachter, professor of Demography at Berkeley. They are looking at the effect of genetic load on cognitive function and morbidity in the elderly. Wachter (along with **Steve Evans** and David Steinsalz) has done some theoretical work on this topic and has access to some suitable cohorts to investigate empirically. The plan is that they will directly test for an effect in these cohorts.

The program allowed **Gerton Lunter** (University of Oxford) to meet for the first time David Patterson of the Computer Science Division at Berkeley, to discuss a common benchmarking strategy for variant identification in genomics data. This visit culminated in their both participating in the Global Alliance for Genomics and Health (GA4GH, `http://genomicsandhealth.org`), in particular the Benchmarking and Reference Variation task teams, to develop common standards and protocols aimed at facilitating the exchange of genetics information.

**David Tse** (Stanford) had discussions with Yun Song and Anand Bhaskar that formed the seed for a project on developing algorithms for geographical localization of individuals from genotype data, and using such information to correct for spurious associations due to population stratification in genome-wide association studies. This project began in earnest during the Spring 2015 program on Information Theory at the Simons Institute (organized by Tse).

**Eleazar Eskin** (UCLA) collaborated with Eran Halperin and James Zou to develop statistical techniques to deconvolve multiple cell types in epigenetic data. Their work provides a new capability for analyzing such data without the need for expensive reference panels.

**Oskar Hallatschek** (UC Berkeley) started a collaboration with **Joachim Hermisson** (University of Vienna) on adaptation in a spatially structured population and its consequences for the site frequency spectrum. Hermisson also initiated a collaboration with **Peter Pfaffelhuber** (University of Freiburg) to derive analytical results for the frequency of soft selective sweeps in spatially structured populations [44].

**Paul Valiant** led a popular open problems session where he demonstrated the challenges of simulating evolution on a computer through his attempts to evolve agents that can play Go. Agents receive a reward (fitness) based on their moves and can make local changes (mutations) to their algorithm.

## Impact of the program on the participants

Several participants, especially the Research Fellows, benefited significantly from the program, by being given a chance to meet scientists from several fields and at various stages of their career. The daily tea time at the institute provided an informal yet structured setting where participants could

get together and talk about their work. Several participants also noted that the ample coffee breaks during the week-long workshops were a big success in letting people follow up on the discussions that arose during the preceding workshop talks, and several collaborations arose spontaneously over these discussions.

Over the course of the program, the computer science theory community gained a more sophisticated understanding of the biological complexities of evolution due to exposure to ideas from other communities. The effects of this became evident, for example, in the work of theoretical computer scientist Vishnoi [111], which rigorously studies the mixing time of a very realistic and complicated Markov chain that commonly arises in evolutionary genetic models. The program can be viewed as a successful and important first step towards developing non-trivial connections across interdisciplinary boundaries in the study of evolution; it is to be hoped that the momentum initiated by the program will lead to a deepening of these connections over time.

## Evolutionary Biology and the Theory of Computing, Spring 2014


[1] Y. ABBASI-YADKORI, P. BARTLETT, and V. KANADE. Tracking Adversarial Targets. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 369–377, 2014.

[2] D. ALDOUS and W. HAN. Introducing Nash equilibria via an online casual game which people actually play. Preprint. 2015.

[3] D. ALDOUS, D. LANOUE, and J. SALEZ. The Compulsive Gambler Process. *Electronic J. Probability,* 20, 35, pp. 1–18, 2015.

[4] C. BANK, R. T. HIETPAS, J. D. JENSEN, and D. N. A. BOLON. A systematic survey of an intragenic epistatic landscape. *Molecular Biology and Evolution,* 32, pp. 229–238, 2015.

[5] N. H. BARTON, S. NOVAK, and T. PAIXÃO. Diverse forms of selection in evolution and computer science. *Proceedings of the National Academy of Sciences,* 111, 29, pp. 10398–10399, 2014.

[6] I. BEZÁKOVÁ, E. W. CHAMBERS, and K. FOX. Counting, Sampling, and Integrating Cuts in Bounded Treewidth Graphs. Submitted.

[7] I. BEZÁKOVÁ and Z. LANGLEY. Minimum Planar Multi-sink Cuts with Connectivity Priors. In *Mathematical Foundations of Computer Science 2014*, 94–105. Springer, 2014.

[8] A. BHASKAR and Y. S. SONG. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *The Annals of Statistics,* 42, 6, pp. 2469–2493, 2014.

[9] A. BHASKAR, Y. X. R. WANG, and Y. S. SONG. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research,* 25, 2, pp. 268–279, 2015.

[10] N. BHATNAGAR, R. BRADLEY, and E. MATSEN. A Statistical View of the Twilight Zone in Sequence Alignment. Preprint. 2015.

[11] N. BHATNAGAR, A. SLY, and P. TETALI. Decay of Correlations for the Hardcore Model on the d-regular Random Graph. *arXiv preprint arXiv:1405.6160,* 2014.

[12] Y. BRANDVAIN and G. COOP. Sperm should evolve to make female meiosis fair. *bioRxiv,* 2014.

[13] M. BUN and J. THALER. Dual Polynomials for Collision and Element Distinctness. *Electronic Colloquium on Computational Complexity (ECCC),* 22, 41, 2015.

[14] M. BUN and J. THALER. Hardness Amplification and the Approximate Degree of Constant-Depth Circuits. In *Proceedings of ICALP*, 2015. To appear.

[15] D. CHAKRABARTHY, S. KANNAN, and K. TIAN. Detecting Character Dependencies in Stochastic Models of Evolution. Journal of Computational Biology. Submitted.

[16] E. CHASTAIN, C. PAPADIMITRIOU, U. VAZIRANI, and A. LIVNAT. Population genetics as Multiplicative Weight Updates: the no-regret theorem. To be submitted.

[17] E. CHASTAIN, A. LIVNAT, C. PAPADIMITRIOU, and U. VAZIRANI. Algorithms, games, and evolution. *Proceedings of the National Academy of Sciences,* 111, 29, pp. 10620–10623, 2014.

[18] H. CHEN, J. HEY, and M. SLATKIN. A hidden Markov model for investigating recent positive selection through haplotype structure. *Theoretical Population Biology,* 99, pp. 18–30, 2015.

[19] Y.-T. CHEN and L. POPOVIC. Spatial epidemic process with recovery. *Preprint,* 2015.

[20] D. DACHMAN-SOLED, V. FELDMAN, L.-Y. TAN, A. WAN, and K. WIMMER. Approximate resilience, monotonicity, and the complexity of agnostic learning. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 498–511, SIAM, 2015.

[21] T. H. DANG and E. MOSSEL. A Statistical Test for Clades in Phylogenies. *arXiv preprint arXiv:1407.7619,* 2014.

[22] S. DARUKI, J. THALER, and S. VENKATABSUBRAMANIAN. Streaming Interactive Verification in Data Analysis. Submitted to ESA. 2015.

[23] G. DASARATHY, R. NOWAK, and S. ROCH. Data Requirement for Phylogenetic Inference from Multiple Loci: A New Distance Method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 12, 2, pp. 422–432, 2015.

[24] C. DASKALAKIS, A. LIVNAT, C. PAPADIMITRIOU, and A. WU. A manuscript on epistasis and genetic diversity. In preparation.

[25] C. DASKALAKIS and G. KAMATH. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of COLT*, 2014.

[26] C. DASKALAKIS and S. ROCH. Species trees from gene trees in the presence of extensive lateral genetic transfer: a tight bound. To be submitted.

[27] A. DILTHEY, C. J. COX, Z. IQBAL, M. R. NELSON, and G. MCVEAN. Improved genome inference in the MHC using a population reference graph. *Nature Genetics,* 47, pp. 682–688, 2015.
[28] A. ETHERIDGE, A. VEBER, and F. YU. Rescaling limits of the spatial Lambda-Fleming-Viot process with selection. *arXiv preprint arXiv:1406.5884v1,* 2014.
[29] S. N. EVANS and A. HENING. Markov processes conditioned on their location at large random times. In preparation.
[30] S. N. EVANS and D. LANOUE. Recovering a tree from the lengths of subtrees spanned by a randomly chosen sequence of leaves. *arXiv preprint arXiv:1506.01091v1,* 2015. Submitted.
[31] F. FARNIA, M. RAZAVIYAYN, S. KANNAN, and D. TSE. Minimum HGR Correlation Principle: from Marginals to Joint Distribution. International Symposium on Information Theory. 2015.
[32] F. FARNIA, M. RAZAVIYAYN, and D. TSE. Inference and Feature Selection via Maximal Correlation. NIPS. 2015. Submitted.
[33] V. FELDMAN, W. PERKINS, and S. VEMPALA. On the complexity of random satisfiability problems with planted solutions. In *Proceedings of STOC*, 2015.
[34] V. FELDMAN, W. PERKINS, and S. VEMPALA. Subsampled Power Iteration: a Unified Algorithm for Block Models and Planted CSP's. *arXiv preprint arXiv:1407.2774,* July 2014.
[35] A. FERRER-ADMETLLA, M. LIANG, T. KORNELIUSSEN, and R. NIELSEN. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular biology and evolution,* 31, 5, pp. 1275–1291, 2014.
[36] N. A. FURLOTTE and E. ESKIN. Efficient multiple trait association and estimation of genetic correlation using the matrix-variate linear mixed-model. *Genetics,* 200, 1, pp. 59–68, 2015.
[37] B. H. GOOD and M. M. DESAI. The impact of macroscopic epistasis on long-term evolutionary dynamics. *Genetics,* 199, pp. 177–190, 2015.
[38] B. H. GOOD and M. M. DESAI. Deleterious Passengers in Adapting Populations. *Genetics,* 198, pp. 1183–1208, 2014.
[39] D. GUSFIELD and R. NIELSEN. Association Mapping for Compound Heterozygous Traits Using Phenotypic Distance and Integer Programming. Submitted for publication.
[40] D. GUSFIELD. *ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks*. MIT Press, 2014.
[41] O. HALLATSCHEK and D. S. FISHER. Acceleration of evolutionary spread by long-range dispersal. *Proceedings of the National Academy of Sciences,* 111, 46, pp. E4911–E4919, 2014.
[42] K. HARRIS and R. NIELSEN. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome research,* 24, 9, pp. 1445–1454, 2014.
[43] J. D. HERMAN, D. P. RICE, U. RIBACKE, A. A. DEIK, E. MOSS, D. NEAFSEY, M. M. DESAI, C. B. CLISH, R. MAZITSCHEK, and D. F. WIRTH. Genomic Evolutionary Approach Reveals Unexpected Metabolic Switch in Malaria Drug Resistance. *Genome Biology,* 15, 511, 2014.
[44] J. HERMISSON and P. PFAFFELHUBER. Fixation of strongly beneficial alleles under recurrent mutation in a structured population. Manuscript. In preparation.
[45] F. HORMOZDIARI, J. W. JOO, A. WADIA, F. GUAN, R. OSTROSKY, A. SAHAI, and E. ESKIN. Privacy preserving protocol for detecting genetic relatives using rare variants. *Bioinformatics,* 30, 12, pp. 204–211, 2014.
[46] F. HORMOZDIARI, E. KOSTEM, E. Y. KANG, B. PASANIUC, and E. ESKIN. Identifying causal variants at loci with multiple signals of association. *Genetics,* 198, 2, pp. 497–508, 2014.
[47] K. JAIN and W. STEPHAN. Response of polygenic traits under stabilizing selection and mutation when loci have unequal effects. *G3-Genes Genomes Genetics,* 5, pp. 1065–1074, 2015.
[48] P. A. JENKINS, P. FEARNHEAD, and Y. S. SONG. Tractable stochastic models of evolution for loosely linked loci. *Electronic Journal of Probability,* 20, 58, pp. 1–26, 2015.
[49] L. JIANG, P. LIU, C. BANK, N. RENZETTE, K. PRACHANRONARONG, L. S. YILMAZ, D. R. CAFFREY, K. B. ZELDOVICH, C. A. SCHIFFER, T. F. KOWALIK, J. D. JENSEN, R. W. FINBERG, J. P. WANG, and D. N. A. BOLON. Quantifying the realized and potential adaptive response of influenza neuraminidase in the presence of oseltamivir. In review.
[50] J. W. J. JOO, E. Y. KANG, N. FURLOTTE, B. PARKS, A. J. LUSIS, and E. ESKIN. Efficient and Accurate Multiple-Phenotypes Regression Method for High Dimensional Data Considering Population Structure. In *Research in Computational Molecular Biology*, pp. 136–153, Springer, 2015.
[51] V. KANADE and E. MOSSEL. MCMC Learning. In *Proceedings of the 28th Annual Conference on Learning Theory*, 2015. To appear.

[52] V. KANADE, E. MOSSEL, and T. SCHRAMM. Global and Local Information in Clustering Labeled Block Models. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM)*, pp. 779–792, 2014.
[53] V. KANADE and J. THALER. Distribution-Independent Reliable Learning. In *Proceedings of the 27th Annual Conference on Learning Theory*, pp. 3–24, 2014.
[54] J. KELLEHER, A. ETHERIDGE, and N. BARTON. Coalescent simulation in continuous space: Algorithms for large neighbourhood size. *Theoretical population biology*, 95, pp. 13–23, 2014.
[55] J. KIM, E. MOSSEL, M. Z. RÁCZ, and N. ROSS. Can one hear the shape of a population history? *Theoretical Population Biology*, 100, pp. 26–38, 2015.
[56] S. KRYAZHIMSKIY, D. P. RICE, E. JERISON, and M. M. DESAI. Global Epistasis Makes Adaptation Predictable Despite Sequence-Level Stochasticity. *Science*, 344, pp. 1519–1522, 2014.
[57] R. KUMAR and L. POPOVIC. Large deviations with averaging of jump-diffusion processes. *Preprint*, 2014. In review for Stochastic Processes and their Applications.
[58] J. B. LACK, C. M. CARDENO, M. W. CREPEAU, W. TAYLOR, R. B. CORBETT-DETIG, K. A. STEVENS, C. H. LANGLEY, and J. E. POOL. The Drosophila Genome Nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics*, 199, 4, pp. 1229–1241, 2015.
[59] G. I. LANG and M. M. DESAI. The spectrum of adaptive mutations in experimental evolution. *Genomics*, 104, pp. 412–416, 2014.
[60] S. A. LANGLEY, G. H. KARPEN, and C. H. LANGLEY. Nucleosomes shape DNA polymorphism and divergence. *PLoS genetics*, 10, 7, pp. e1004457, 2014.
[61] I. LAZARIDIS, N. PATTERSON, A. MITTNIK, G. RENAUD, S. MALLICK, P. H. SUDMANT, J. G. SCHRAIBER, S. CASTELLANO, K. KIRSANOW, C. ECONOMOU, and OTHERS. Ancient human genomes suggest three ancestral populations for present-day Europeans. *arXiv preprint arXiv:1312.6639*, 2013.
[62] M. LIANG and R. NIELSEN. The lengths of admixture tracts. *Genetics*, 197, 3, pp. 953–967, 2014.
[63] A. LIVNAT. A manuscript on the source of creativity in evolution. To be submitted.
[64] A. LIVNAT and C. PAPADIMITRIOU. A note on finite populations. To be submitted.
[65] A. LIVNAT, C. PAPADIMITRIOU, and U. VAZIRANI. Algorithms, games and evolution: the sequel. To be submitted.
[66] A. LIVNAT and C. PAPADIMITRIOU. Sex as an algorithm: evolution under the lens of computation. Under review, CACM.
[67] A. LIVNAT, C. PAPADIMITRIOU, A. RUBINSTEIN, G. VALIANT, and A. WAN. Satisfiability and evolution. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 524–530, IEEE, 2014.
[68] S. LUO and J. MATTINGLY. Scaling limits of a model for selection at two scales. Preprint. 2015.
[69] S. MANGUL, N. C. WU, N. MANCUSO, A. ZELIKOVSKY, R. SUN, and E. ESKIN. Accurate viral population assembly from ultra-deep sequencing data. *Bioinformatics*, 30, 12, pp. 329–337, 2014.
[70] I. MATHIESON and G. MCVEAN. Demography and the age of rare variants. *PLoS Genetics*, 10, e1004528, 2014.
[71] I. MOLTKE, M. FUMAGALLI, T. S. KORNELIUSSEN, J. E. CRAWFORD, P. BJERREGAARD, M. E. JØRGENSEN, N. GRARUP, H. C. GULLØV, A. LINNEBERG, O. PEDERSEN, and OTHERS. Uncovering the genetic history of the present-day greenlandic population. *The American Journal of Human Genetics*, 96, 1, pp. 54–69, 2015.
[72] E. MOSSEL and S. ROCH. Distance-based species tree estimation: information-theoretic trade-off between number of loci and sequence length under the coalescent. RANDOM. 2015. To appear.
[73] E. MOSSEL and M. STEEL. Majority rule has transition ratio 4 on Yule trees under a 2-state symmetric model. *Journal of theoretical biology*, 360, pp. 315–318, 2014.
[74] D. B. NEALE, J. L. WEGRZYN, K. A. STEVENS, A. V. ZIMIN, D. PUIU, M. W. CREPEAU, C. CARDENO, M. KORIABINE, A. E. HOLTZ-MORRIS, J. D. LIECHTY, and OTHERS. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome biology*, 15, 3, pp. R59, 2014.
[75] J. NEIDHART, I. G. SZENDRO, and J. KRUG. Adaptation in tunably rugged fitness landscapes: The Rough Mount Fuji Model. *Genetics*, 198, 2, pp. 699–721, 2014.
[76] I. E. OCHS and M. M. DESAI. The competition between simple and complex evolutionary trajectories in asexual populations. *BMC evolutionary biology*, 15, 1, pp. 55, 2015.

[77] J. OTWINOWSKI and J. KRUG. Clonal interference and Muller's ratchet in spatial habitats. *Physical biology,* 11, 5, pp. 056003, 2014.
[78] T. PAIXAO, G. BADKOBEHE, N. H. BARTON, C. DOLGAN, D. C. DANG, T. FRIEDRICH, P. K. LEHRE, D. SUDHOLT, and B. TRUBENOVA. A unifying framework for evolutionary processes. J. Theor. Biol. In review.
[79] I. PANAGEAS, P. SRIVASTAVA, and N. K. VISHNOI. Evolutionary Dynamics in Finite Populations Mix Rapidly. Submitted.
[80] C. PAPADIMITRIOU. Algorithms, complexity, and the sciences. *Proceedings of the National Academy of Sciences,* 111, 45, pp. 15881–15887, 2014.
[81] C. H. PAPADIMITRIOU and N. K. VISHNOI. On the Poincaré-Bendixson Theorem and Computational Complexity. Submitted.
[82] P. A. PAPAKONSTANTINOU, J. XU, and Z. CAO. Bagging by Design (on the Suboptimality of Bagging). In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
[83] P. A. PAPAKONSTANTINOU and G. YANG. Cryptography with Streaming Algorithms. In *CRYPTO*, pp. 55–70, August 2014.
[84] S.-C. PARK and J. KRUG. The c-record process and evolution in epistatic rough Mount Fuji fitness landscapes. Preprint. 2015.
[85] S.-C. PARK, J. NEIDHART, and J. KRUG. Greedy adaptive walks on a correlated fitness landscape. Preprint. 2015.
[86] S.-C. PARK, I. G. SZENDRO, J. NEIDHART, and J. KRUG. Phase transition in adaptive walks on the rough Mount Fuji fitness landscape. *Phys. Rev.,* 91, 042707, 2015.
[87] B. M. PETER and M. SLATKIN. The effective founder effect in a spatially expanding population. *Evolution,* 69, 3, pp. 721–734, 2015.
[88] P. PFAFFELHUBER and L. POPOVIC. How spatial heterogeneity shapes multi-scale biochemical reactions. *Journal of the Royal Society Interface,* 12, 104, 2015.
[89] L. POPOVIC and M. RIVAS. Cherries and parameter inference on multi-type Yule trees. Journal of Mathematical Biology. In review.
[90] L. POPOVIC and M. RIVAS. The coalescent point process of multi-type branching trees. *Stochastic Processes and their Applications,* 124, 12, 2014.
[91] F. RACIMO, M. KUHLWILM, and M. SLATKIN. A test for ancient selective sweeps and an application to candidate sites in modern humans. *Molecular biology and evolution,* 31, 12, pp. 3344–3358, 2014.
[92] F. RACIMO, S. SANKARARAMAN, R. NIELSEN, and E. HUERTA-SÁNCHEZ. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics,* 16, pp. 359–371, 2015.
[93] P. L. RALPH and G. COOP. Convergent Evolution During Local Adaptation to Patchy Landscapes. *bioRxiv,* pp. 006940, 2014.
[94] D. P. RICE, B. H. GOOD, and M. M. DESAI. The evolutionarily stable distribution of fitness effects. *Genetics,* 200, pp. 321–329, 2015.
[95] A. RIMMER, H. PHAN, I. MATHIESON, Z. IQBAL, S. TWIGG, A. WILKIE, G. MCVEAN, and G. LUNTER. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet,* 46, 8, pp. 912–918, 2014.
[96] S. ROCH and M. STEEL. Likelihood-based tree reconstruction on a concatenation of alignments can be positively misleading. *Theoretical Population Biology,* 100, pp. 56–62, 2015.
[97] S. ROCH and T. WARNOW. On the Robustness to Gene Tree Estimation Error (or lack thereof) of Coalescent-Based Species Tree Methods. *Syst Biol,* 2015. Doi:10.1093/sysbio/syv016.
[98] J. G. SCHRAIBER, S. N. EVANS, and M. SLATKIN. Bayesian inference of natural selection from allele frequency time series. In preparation.
[99] P. R. STAAB, S. ZHU, D. METZLER, and G. LUNTER. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics,* 31, 10, pp. 1680–1682, 2015.
[100] M. STEEL. Tracing evolutionary links between species. *American Mathematical Monthly,* 121, 9, pp. 771–792, 2015.
[101] M. STEEL and J. D. VELASCO. Axiomatic opportunities and obstacles for inferring a species tree from gene trees. *Systematic Biology,* 63, 5, pp. 772–778, 2014.
[102] J. H. SUL, T. RAJ, S. DE JONG, P. I. DE BAKKER, S. RAYCHAUDHURI, R. A. OPHOFF, and B. HAN. Accurate and fast multiple-testing correction in eqtl studies. *American Journal of Human Genetics,* 96, 6, pp. 857–

868, 2015.
[103] P. TATARU, J. A. NIRODY, and Y. S. SONG. diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals. *Bioinformatics,* 30, 23, pp. 3430–3431, 2014.
[104] J. TERHORST and Y. S. SONG. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. In *Proc. Natl. Acad. Sci. U.S.A.*, volume 112, pp. 7677–7682, 2015.
[105] J. TERHORST and Y. S. SONG. Multi-locus analysis of genomic time series data from experimental evolution. *PLoS Genetics,* 11, 4, pp. e1005069, 2015.
[106] J. THALER. Lower Bounds for the Approximate Degree of Block-Composed Functions. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 22, pp. 150, 2014.
[107] M. V. TROTTER, D. B. WEISSMAN, G. I. PETERSON, K. M. PECK, and J. MASEL. Cryptic genetic variation can make "irreducible complexity" a common mode of adaptation in sexual populations. *Evolution,* 68, pp. 3357–3367, 2014.
[108] G. VALIANT and P. VALIANT. An Automatic Inequality Prover and Instance Optimal Identity Testing. In *Proceedings of IEEE FOCS*, 2014.
[109] G. VALIANT and P. VALIANT. Instance Optimal Learning. *arXiv preprint arXiv:1504.05321v1,* 2015.
[110] P. VALIANT. Evolvability of Real Functions. *ACM Transactions on Computation Theory (TOCT),* 6, 3, pp. 12, 2014.
[111] N. K. VISHNOI. The Speed of Evolution. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1590–1601, 2015.
[112] H. P. DE VLADAR and N. BARTON. Stability and response of polygenic traits to stabilizing selection and mutation. *Genetics,* 197, 2, pp. 749–767, 2014.
[113] Z. WANG, J. H. SUL, S. SNIR, J. A. LOZANO, and E. ESKIN. Gene-Gene interactions detection using a two-stage model. *Journal of Computational Biology,* 2015.
[114] J. L. WEGRZYN, J. D. LIECHTY, K. A. STEVENS, L.-S. WU, C. A. LOOPSTRA, H. A. VASQUEZ-GROSS, W. M. DOUGHERTY, B. Y. LIN, J. J. ZIEVE, P. J. MARTÍNEZ-GARCÍA, C. HOLT, M. YANDELL, A. V. ZIMIN, J. A. YORKE, M. W. CREPEAU, D. PUIU, S. L. SALZBERG, P. J. DE JONG, K. MOCKAITIS, D. MAIN, C. H. LANGLEY, and D. B. NEALE. Unique Features of the Loblolly Pine (Pinus taeda L.) Megagenome Revealed Through Sequence Annotation. *Genetics,* 196, pp. 891–909, 2014.
[115] D. B. WEISSMAN. Stress-induced variation can cause average mutation and recombination rates to be positively correlated with fitness. In *ALIFE 14: The Fourteenth Conference on the Synthesis and Simulation of Living Systems*, volume 14, pp. 43–44, 2014.
[116] D. B. WEISSMAN and O. HALLATSCHEK. The rate of adaptation in large sexual populations with linear chromosomes. *Genetics,* 196, 4, pp. 1167–1183, 2014.
[117] A. WOLLSTEIN and W. STEPHAN. Adaptive Fixation in Two-Locus Models of Stabilizing Selection and Genetic Drift. *Genetics,* 198, 2, pp. 685–697, 2014.
[118] M. A. YANG, K. HARRIS, and M. SLATKIN. The projection of a test genome onto a reference population and applications to humans and archaic hominins. *Genetics,* 198, 4, pp. 1655–1670, 2014.
[119] W. Y. YANG, A. PLATT, C. W. CHIANG, E. ESKIN, J. NOVEMBRE, and B. PASANIUC. Spatial localization of recent ancestors for admixed individuals. *G3,* 4, 12, pp. 2505-18, 2014.
[120] A. ZIMIN, K. A. STEVENS, M. W. CREPEAU, A. HOLTZ-MORRIS, M. KORIABINE, G. MARÇAIS, D. PUIU, M. ROBERTS, J. L. WEGRZYN, P. J. DEJONG, D. B. NEALE, S. L. SALZBERG, J. A. YORKE, and C. H. LANGLEY. Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. *Genetics,* 196, pp. 875–890, 2014.
[121] D. ŽIVKOVIĆ, M. STEINRÜCKEN, Y. S. SONG, and W. STEPHAN. Transition densities and sample frequency spectra of diffusion processes with selection and variable population size. *Genetics,* 200, 2, pp. 601–617, 2015.
[122] J. ZOU, E. HALPERIN, and S. SANKARARAMAN. Inferring parental genomic ancestries using pooled semi-Markov processes. Proceedings of ISMB. 2015. To appear.
[123] J. Y. ZOU, D. S. PARK, E. G. BURCHARD, D. G. TORGERSON, M. PINO-YANES, Y. S. SONG, S. SANKARARAMAN, E. HALPERIN, and N. ZAITLEN. A genetic and socio-economic study of mate choice in Latinos reveals novel assortment patterns. Under review.