Algorithmic Challenges in Genomics (Spring 2016)

Final Program Report

Ron Shamir (Organizing Chair)

Background and goals

Computational biology, a.k.a. bioinformatics, has developed dramatically over the last two decades. It is by now a well-established discipline, with numerous undergraduate and graduate programs available around the world, many conferences, books, and scientific journals. Starting from strong roots in theoretical computer science, over the last decade there has been a dramatic expansion of the bioinformatics community that brought in many practitioners with roots outside computer science, e.g., in biology, physics, biochemistry, bioengineering and other disciplines. As a result, a significant part of the bioinformatics community drifted towards data-driven methods and often away from theoretically sound developments. Moreover, a plethora of new data types has brought about more emphasis on analysis and less on theory. The goal of the Simons Institute program was to **regain the theory-practice balance in bioinformatics**, by bringing together leaders and young scientists in the field with a strong interest in the algorithmic, methodological and theoretical aspects of computational biology.

Range and themes covered

As the span of bioinformatics has become extremely broad, we chose to focus on **genomics**, the field that deals with understanding the sequence, function and structure of genomes (the complete sequence of DNA present in each cell of an organism). Genomics is abuzz with huge data sets generated by international consortia and individual groups, with numerous novel experimental data types that require innovative analysis methodologies. New algorithms for data analysis are acutely needed today. Shortage of good, well founded algorithms constitutes a major bottleneck in utilizing genomics to advance our understanding in biology and medicine.

The Algorithmic Challenges in Genomics (ACG) program set out to focus on three closely related topics:

> **Computational Cancer Biology** is a rapidly expanding area, since high throughput sequencing (HTS) techniques have facilitated the sequencing of tens of thousands of tumor genomes, along with other types of information. Large international projects are collecting and organizing these data, but algorithms for analyzing the data are the bottleneck. Current analysis techniques combine graph theoretic and machine learning approaches. One line of analysis methodology builds on the rich combinatorial and algorithmic theory of genome rearrangements. Prominent novel challenges include handing of heterogeneous cancer samples, inference of cancer evolution, and single-cell analysis.
>
> **Regulatory Genomics and Epigenomics** aims to understand the way gene expression is controlled in cells by understanding the sequence elements and organization that govern this expression. Techniques here start from algorithms for motif discovery at the sequence level (a.k.a. "stringology") and also encompass interaction between diverse higher level elements affecting regulation, using statistical analyses. A plethora of novel array and HTS-based data types are part of this challenge.

Network Biology aims to understand the web of interactions among cellular components, which affect all activities and diseases. Only by understanding these interactions (among genes, proteins, RNAs, complexes and other molecules) can a higher level understanding of an organism's function, dynamics and development be achieved. Analysis techniques include graph algorithms, combinatorial optimization, machine learning and statistics.

Participants

The core of the program consisted of 31 long term participants (LTPs) and ten research fellows (RFs). All the fellows and about half the faculty spent the complete semester (four months) at the Simons Institute. Most of the remaining LTPs spent 1-2 months each at the Institute. In addition, there were about a dozen Visiting Graduate Students and Postdocs who participated in part or all of the semester. About half of the above were from the US and the rest from Europe and Asia.

In addition to the core faculty, research fellows and students, some 300 other researchers and students participated in four week-long workshops organized during the program. These are described below. Workshops and seminars were broadly announced and participation was open.

Activities

At the outset of the program, a four-day **boot camp** provided an introduction to the basics of modern computational biology, including fundamentals of string algorithms ("stringology"), algorithms for deep sequencing data, and an introduction to each of the three themes of the semester. These three themes were covered in depth by the three workshops that took place during the semester.

Each of the workshops brought together some 100 researchers from academia and industry, including computer scientists, biologists and medical researchers. Participants of the workshops ranged from graduate students and postdocs to senior faculty, interacting in an open and welcoming environment, with lively and significant discussions amongst all participants. An important component in each of the workshops was afternoon panel discussions, giving participants the opportunity to express thought-provoking personal perspectives on the field, and sometimes raising spirited debates on future directions and on the best approaches for evaluating progress. Responses from the attendees were highly enthusiastic. In some workshops, participants expressed interest in organizing a follow-up event in one or two years' time.

During the semester we also had three weekly seminar series: a joint **UCSF-Simons seminar**, organized by postdocs from UCSF and Simons fellows, who invited speakers from Bay Area universities and companies; an informal **whiteboard seminar**, in which ACG program participants described their ongoing research and open problems; and a **deep learning seminar**, initiated and organized by the program's fellows, wherein the participants read recent literature in the area together. The latter was also attended by participants of the concurrent program on "Counting Complexity and Phase Transitions." All seminars took place in a very open and informal atmosphere with a lively and fruitful exchange between the speakers and the audience.

Additional special activities included:

- **A field trip to Silicon Valley**, where we spent a day visiting **Agilent**, a leading genomic biotechnology company, and **23&Me**, a personal genomics start-up, for meetings, site visits and short talks by company and ACG program researchers.
- **Genomics Program Biotechnology Companies Day** at the Simons Institute, in which speakers from ten biotechnology and genomics companies (mostly start-ups) described their computational challenges and vision, and networked with the ACG program participants.
- An **integer linear programming (ILP) day**, organized by LTP Dan Gusfield, in which theoretical developments and applications of ILP to bioinformatics were discussed.

Beyond these structured events, the group of long-term visiting senior researchers, postdoctoral fellows and graduate students who were present throughout the semester was active daily in scientific interactions. As the goal of the program was to foster novel research and collaborations in the focus areas, the core participants were encouraged to present their current research work both informally within the workshops and seminars and at numerous informal meetings throughout the semester. Somewhat unexpectedly, the synergy with the parallel Institute program on Counting Complexity and Phase Transitions was quite intensive and several joint collaborations were initiated.

Outcomes

Overall, based on early assessments at the one-year anniversary, the program was highly successful. Participants applauded in particular the workshops (including the availability of talks on Youtube for broader exposure), and noted as the main achievement the many new interactions and collaborations that they forged. Some participants secured postdoctoral positions following contacts established at the program, and one student was hired by a company as a result of the Industry Day. New industry-academia collaborations were initiated. Naturally, some of the scientific work that started during the semester continues and will only fully bear fruit over time. Here are some notable highlights:

Direct research outcomes

- RF Thomas Sakoparnig and LTP Erik van Nimwegen, together with UCB faculty member Oscar Hallatschek, developed three (increasingly complex) models for describing prokaryotic evolution with a focus on high rates of recombination. The models led to an improved understanding of how diversity in natural prokaryotic populations is introduced and maintained [82].

- RF Yaron Orenstein and LTPs Carl Kingsford and Ron Shamir provided the first effective solution to the universal k-mer hitting set problem (finding a minimum-size set of k-mers that hits all possible L-long sequences). Many manifestations of this generic problem arise in the analysis of HTS data. The result [72] allows for more efficient analysis of these data, both in memory and in running-time, compared to the state of the art. A follow-up paper by the same team [63] provided a theoretical explanation for why certain fingerprinting/minimizer schemes have traditionally worked poorly, and used universal k-mers to design an improved minimizer scheme. This line of work advanced an interesting theoretical problem, explained theoretically the observed empirical performance of various minimizer schemes, and resulted in a better scheme based on universal k-mers that leads to faster algorithms for several genomics tasks.

- A new book on computational cancer genomics was initiated by LTPs Niko Beerenwinkel and Florian Markowetz, and major progress in writing it was achieved during the program.

- During the Network Biology workshop it became clear that several presentations were using the same generic transformation of the data to obtain state of the art results. To quote LTP Donna Slonim, "*The broad discussion of the advantage of diffusion based methods for calculating network distances for function prediction and disease gene discovery was a highly valuable summary and organization of disparate results in the field that had previously been considered separately. This synthesis is particularly valuable because it places the work in a conceptual and theoretical framework that suggests how to use such methods for future discoveries.*" LTP Roded Sharan teamed up with prominent participants in the workshop to write a review emphasizing the central role of network propagation in interpreting biological data and revealing novel genetic associations. This review has recently been published in Nature Reviews Genetics [18]. Several participants noted the impact of that discussion.

- LTPs Cenk Sahinalp and David Tse studied core computer science problems dealing with genomic data such as compression and encryption, accessing and comparing data in compressed and encrypted form, upper bounds on compression, analyzing of streaming genomic data, etc. [36].

- LTPs Ben Raphael, Ron Shamir and Roded Sharan, together with RF Meirav Zehavi, developed a new model for the copy number evolution problem in cancer and algorithms for its analysis [27].

- LTPs Sharan and Teresa Pryztycka have started a collaboration on network approaches for understanding the mutational landscape of cancer. They hypothesized that exploring the interplay between co-occurrence, mutual exclusivity, and functional interactions between genes can improve our understanding of the disease and help to uncover new relations between cancer driving genes and pathways. They developed a general framework for identifying modules with different combinations of mutation and interaction patterns, and by using Integer Linear Programming found optimally scoring sets of modules. The method helped identify functionally coherent modules that might be relevant for cancer progression, pairs of genes with potentially synergetic effects, and other interesting features [20].

- RF David Amar and LTP Ron Shamir developed the first method that can predict cancer type in metastases based on somatic mutation data alone [6]. Prior prediction methods needed to integrate additional data types to obtain such results.

- LTPs Niko Beerenwinkel and Ben Raphael studied models for tumor evolution. With Katharina Jahn and Jack Kuipers, both workshop visitors from Beerenwinkel's group, they finalized a new method for reconstructing tumor phylogenies from single-cell sequencing data [52]. Raphael and Beerenwinkel extended the method to test the infinite sites assumption using tumor sequencing data. Surprisingly, they found that the assumption is frequently violated, calling for more elaborate approaches to the tumor phylogeny problem [56].

- LTPs Beerenwinkel and Sharan and RFs Simona Cristea and Dana Silverbush developed a new method for integrating multi-omics measurements of tumors in order to prioritize cancer driver genes [9].

- Motivated by a talk from Dovetail Genomics during the Industry day, LTP David Tse and Jiaming Xu, a RF of the sister program Counting Complexity and Phase Transitions, asked how to effectively exploit the particular information in the new technology to complete sequence assembly. By exploiting these connections, they were able to find both information limits for this problem and efficient algorithms that come close to the limit [8].

- Inspired by discussions with local LTP Lior Pachter, LTP Erik van Nimwegen became interested in statistical inference of gene expression noise from fluorescent reporters. Erik realized it is a perfect pedagogical example for illustrating the differences between orthodox frequentist and Bayesian methods in statistical inference, and wrote a short paper on this topic, contrasting the two viewpoints [67].

- LTP Michael Waterman collaborated with Haiyan Huang (Berkeley Statistics Department) and Rachel Wang (Steins Fellow, Stanford). They extended their non-parametric correlation statistic based on gene expression patterns to capture patterns in time series [93]. Coexpression analysis is an important technique routinely used to infer gene regulatory interactions and aims to capture association patterns in gene expression levels under different conditions or time points. Temporal changes in gene expression may result in complex association patterns that require more sophisticated detection methods than simple correlation. For instance, the effect of regulation may lead to local and time-lagged associations. Furthermore, when comparing expression profiles from different individuals or even different species, the time points at which measurements were taken are usually not aligned nor directly comparable. The new measure addresses these issues.

- LTP Jean-Philippe Vert and Berkeley professor Sandrine Dudoit developed statistical methods for the analysis of single-cell RNA-seq data. They developed a new model that takes into account the specificities of the data, in particular the fact that many measures are missing, and produced publicly available software [81].

- LTP Jean-Philippe Vert and his student Marine Le Morvan benefited from numerous discussions with other participants to improve their work on the stratification of cancer patients from mutations in DNA, which resulted in the publication of their new algorithm NetNorM [57].

- RF Meirav Zehavi collaborated with Prof. Fedor Fomin of the sister program Counting Complexity and Phase Transitions, to solve several central problems on the computational complexity of the Terrain Guarding problem [7]. Meirav met Fomin for the first time at the Institute, and their collaboration has continued fruitfully over the year since the program ended, resulting in additional publications [32] and [33] as well as several ongoing projects.

- LTPs Cenk Sahinalp and David Tse, with Berkeley professor Tom Courtade, developed a new reference-free compressed representation for HTS data based on ”light” de novo assembly of reads, where each input read is represented as a node of a (compact) trie, and an iterative method to build a forest of such tries. Their algorithm is the first HTS data compression method to provide optimality guarantees, from both the combinatorial and information theoretic perspectives [35].

Other outcomes

- RF Iman Hajirasouliha noted that his participation in the program helped him in his future academic career. In his words, "*Beside my main research project at Simons and the amazing environment for collaborating and discussing ideas with colleagues, the Simons program helped me with my future career as a faculty. I was on the faculty job market during the time at Simons, and my mentor Ron Shamir, as well as many other senior faculty in the program, helped me a lot with my job application and interviews. I gave several practice and regular talks at Simons during the program and improved my presentation skills quite a bit*." Iman is now a faculty member in the Weil-Cornell Medical School.

- A student of Uwe Ohler found her current job as senior bioinformaticist at Natera during the Industry Day at the Simons Institute.

- RF Xiwei Zhang found her new postdoc advisor at the program. Xiwei says "*Another outcome for myself during my stay at Simons is that I met my current postdoc advisor, Prof. Nir Yosef, at the program. That's a big thing for my academic career."*

- During the Industry Day, RF Iman Hajirasouliha made contact with the company 10X Genomics. This evolved into a long-term collaboration and Iman is now developing methods for the analysis of long-read HTS data.

- Many participants commended the Industry Day. For example, LTP David Tse said *"The Biotech industry day was very good. It got me exposed to a broader range of problems (one of which led to a successful collaborative project)."*

- Several participants emphasized the value to their future research and way of thinking from exposure to deep learning methods in bioinformatics, both via the ad-hoc seminar and through other talks on the topic in the program and workshops.

New collaborations fostered

Numerous new collaborations were initiated during the program. Many of them already produced new research papers and others are ongoing. Here are just a few of them:

- Kingsford - Orenstein - Shamir
- Sahinalp - Tse - Courtade
- Beerenwinkel - Raphael
- Amar - Shamir - Vandin
- Beerenwinkel - Sharan - Silverbush - Cristea
- Strumfels - Ludington - Beerenwinkel
- Cowen - Ideker - Raphael - Sharan
- Raphael - Shamir - Zehavi - Sharan
- Vandin - Beerenwinkel - Cristea
- Vandin - Sharan
- Waterman - Huang
- Yosef - Amar - Shamir
- Zehavi - Formin
- Vandin - Amar - Shamir
- Beerenwinkel - Vandin - Cristea
- Sahinalp - Beerenwinkel
- Hajirasouliha - Popic - Batzoglou
- Ohler - Pollard - Ahituv
- Ohler - Orenstein - Berger
- Przytycka - Sharan
- Tse - Xu
- Pachter - van Nimwegen
- Vingron - Sharan - Raphael
- Weng - Sharan
- Wong - Brenner - Yu
- Zhang-Yosef

Lessons learned

Overall, the program was extremely successful. Its strongest points were the workshops, the many new contacts and interactions formed, and the newly initiated research projects. The Simons Institute facilities and staff were truly excellent and very instrumental for achieving the program goals. Somewhat expectedly, participation by experimentalists outside of the workshops was minor, due to their inability to leave their labs for extended periods. Similarly, involvement by many local researchers (from Berkeley, UCSF and Stanford) was limited, due to their other regular obligations. Mechanisms to improve these points could be considered for future programs. The Industry Day and the field trip were highly successful and could be repeated, but such activities require a lot of planning and good local contacts. Finally, given the size of the program, we feel in retrospect that it might have been useful to hold a weekly lunch for the LTPs and RFs, with the aim of speeding up the creation of new collaborations and increasing the pace of interactions.

Algorithmic Challenges in Genomics, Spring 2016

[1] S. ABADI, W. X. YAN, D. AMAR, and I. MAYROSE. A machine learning approach for revealing patterns underlying the 1CRISPR-Cas9 mechanism of action. Submitted.

[2] A. AGRAWAL, F. PANOLAN, S. SAURABH, and M. ZEHAVI. Simlutaneous Feedback Edge Set. In *Proceedings of the 27th International Symposium on Algorithms and Computation (ISAAC)*, pp. 5:1–5:13, 2016.

[3] A. AGRAWAL, S. SAURABH, R. SHARMA, and M. ZEHAVI. Kernels for Deletion to Classes of Acyclic Digraphs. *Journal of Computer and System Sciences (JCSS),* 2017. Accepted.

[4] D. AMAR, S. IZRAELI, and R. SHAMIR. Utilizing somatic mutation data from numerous studies for cancer research: proof of concept and applications. *Oncogene,* 2017.

[5] D. AMAR, S. IZRAELI, and R. SHAMIR. Cancer subtype classification using only somatic mutations: proof of concept and applications. Submitted.

[6] D. AMAR, R. SHAMIR, and D. YEKUTIELI. Extracting replicable associations across multiple studies: empirical Bayes algorithms for controlling the false discovery rate. Submitted.

[7] P. ASHOK, F. V. FOMIN, S. KOLAY, S. SAURABH, and M. ZEHAVI. Exact Algorithms for Terrain Guarding. In *Proceedings of the 33rd Annual Symposium on Computational Geometry (SoCG)*, 2017.

[8] V. BAGARIA, D. TSE, Y. WU, and J. XU. DNA seriation under the planted Hamiltonian path model. In *Information Theory and its Applications Workshop*, 2017.

[9] N. BEERENWINKEL, S. CRISTEA, R. SHARAN, and D. SILVERBUSH. CanOmics: Multi-Omics Data Integration Improves the Identification of Cancer Drivers. In preparation.

[10] N. BEERENWINKEL, S. CRISTEA, and F. VANDIN. Controlling the False Positive Rate for Mutual Exclusivity Testing. In preparation.

[11] N. BEERENWINKEL and F. MARKOWETZ. *Inferring tumor evolution*. Book in preparation.

[12] A. BELORKAR and L. WONG. GFS: fuzzy preprocessing for effective gene expression analysis. *BMC Bioinformatics,* 17, 17, pp. 169–184, 2016.

[13] S. BERGER, S. OMIDI, M. PACHKOV, P. ARNOLD, N. KELLEY, S. SALATINO, and E. VAN NIMWEGEN. Crunch: Completely Automated Analysis of ChIP-seq Data. *bioRxiv,* 042903, 2016.

[14] A. BOMERSBACH, M. CHIARANDINI, and F. VANDIN. An Efficient Branch and Cut Algorithm to Find Frequently Mutated Subnetworks in Cancer. In *Proceedings of the 16th Workshop on Algorithms in Bioinformatics (WABI)*, 2016.

[15] M. CECCARELLO, C. FANTOZZI, A. PIETRACAPRINA, G. PUCCI, and F. VANDIN. Clustering Uncertain Graphs. *arXiv preprint arXiv:1612.06675,* 2016.

[16] F. CHEN, S. WANG, X. JIANG, S. DING, Y. LU, J. KIM, S. C. ŞAHINALP, C. SHIMIZU, J. C. BURNS, V. J. WRIGHT, E. PNG, M. L. HIBBERD, D. D. LLOYD, H. YANG, A. TELENTI, C. S. BLOSS, D. FOX, K. LAUTER, and L. OHNO-MACHADO. PRINCESS: Privacy-protecting Rare disease International Network Collaboration via Encryption through Software guard extensionS. *Bioinformatics,* 33, 6, pp. 871–878, 2017.

[17] Z. CHEN, D. CHEN, H. DING, Z. HUANG, Z. LI, N. SEHGAL, A. FRITZ, R. BEREZNEY, and J. XU. Finding Rigid Sub-Structure Patterns From 3D Point-Sets. In *Proceedings of The 23rd International Conference on Pattern Recognition (ICPR)*, 2016. Conditionally accepted (minor revision) to Algorithmica.

[18] L. COWEN, T. IDEKER, B. RAPHAEL, and R. SHARAN. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics,* 2017. Presubmission inquiry submitted to Nature Biotechnology.

[19] K. CRONA, A. GAVRYUSHKIN, D. GREENE, and N. BEERENWINKEL. Inferring genetic interactions from comparative fitness data. Preprint, https://doi.org/10.1101/137372, 2017.

[20] P. DAO, Y.-A. KIM, D. WOJTOWICZ, S. MADAN, R. SHARAN, and T. M. PRZYTYCKA. BeWith: A Between-Within Method to Discover Relationships between Cancer Modules via Integrated Analysis of Mutual Exclusivity, Co-occurrence and Functional Interactions. *arXiv preprint arXiv:1704.08889,* 2017. Accepted to RECOMB 2017 and under revision in PLoS Computational Biology.

[21] D. DETOMASO and N. YOSEF. FastProject: A Tool for Low-Dimensional Analysis of Single-Cell RNA-Seq Data. *bioRxiv,* 043463, 2016. Under review.

[22] H. DING, J. GAO, and J. XU. Finding Global Optimum for Truth Discovery: Entropy Based Geometric Variance. In *Proceedings of The 32nd International Symposium on Computational Geometry (SoCG)*, 2016.

[23] H. DING, L. HU, L. HUANG, and J. LI. Capacitated Center Problems with Two-Sided Bounds and Outliers. In *Algorithms and Data Structures Symposium (WADS)*, 2017.

[24] H. DING, Y. LIU, L. HUANG, and J. LI. K-Means Clustering with Distributed Dimensions. In *Proceedings of*

The 33rd International Conference on Machine Learning (ICML), pp. 1339–1348, 2016.

[25] H. DING, L. SU, and J. XU. Towards Distributed Ensemble Clustering for Networked Sensing Systems: A Novel Geometric Approach. In *Proceedings of The 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, pp. 1–10, 2016.

[26] H. DING and J. XU. FPTAS for Minimizing the Earth Mover's Distance Under Rigid Transformations and Related Problems. *Algorithmica,* pp. 1–30, 2016.

[27] M. EL-KEBIR, B. RAPHAEL, R. SHAMIR, R. SHARAN, S. ZACCARIA, M. ZEHAVI, and R. ZEIRA. Copy-Number Evolution Problems: Complexity and Algorithms. Preprint, 2016. Workshop on Algorithms in Bioinformatics, submitted.

[28] M. EL-KEBIR, B. J. RAPHAEL, R. SHAMIR, R. SHARAN, S. ZACCARIA, M. ZEHAVI, and R. ZEIRA. Complexity and algorithms for copy-number evolution problems. *Algorithms for Molecular Biology,* 12, 2017.

[29] A. F. ERGUN. Buffer reordering for genomic data. In preparation.

[30] A. F. ERGUN, E. GRIGORESCU, E. SADEQI-AZER, and S. ZHOU. Periodicity with k-mismatches. In *Proceedings of the 21st International Workshop on Randomization and Computation (RANDOM)*, 2017.

[31] F. V. FOMIN, D. LOKSHTANOV, S. M. MEESUM, S. SAURABH, and M. ZEHAVI. Matrix Rigidity: Matrix Theory from the Viewpoint of Parameterized Complexity. In *Proceedings of the 34th International Symposium on Theoretical Aspects of Computer Science (STACS)*, 2017.

[32] F. V. FOMIN, D. LOKSHTANOV, F. PANOLAN, S. SAURABH, and M. ZEHAVI. Finding, Hitting and Packing Cycles in Subexponential Time on Unit Disk Graphs. In *Proceedings of the 44th International Colloquium on Automata, Languages and Programming (ICALP)*, 2017.

[33] E. M. GAYO, M. B. COLE, K. E. KOLB, Z. OUYANG, J. CRONIN, S. W. KAZER, J. ORDOVAS-MONTANES, M. LICHTERFELD, B. D. WALKER, N. YOSEF, A. K. SHALEK, and X. G. YU. Single-cell RNA-Seq Identifies A Highly Functional Dendritic Cell Subset In HIV-1 Elite Controllers. Preprint, 2016. Under review.

[34] M. GHANBARI, J. LASSERRE, and M. VINGRON. The Distance Precision Matrix: computing networks from nonlinear relationships. *arXiv preprint arXiv:1605.03378v2,* 2016.

[35] T. GINART, J. HUI, K. ZHU, I. NUMANAGIC, T. COURTADE, C. ŞAHINALP, and D. TSE. Optimal Compressed Representation of High Throughput Sequence Data via Light Assembly. Submitted.

[36] T. GINART, K. ZHU, J. HUI, I. NUMANAGIC, D. TSE, T. COURTADE, and C. ŞAHINALP. Genomic Reads Forests for Compressed Representation of High Throughput Sequence Data. In *Proceedings of the Seventh RECOMB Satellite Workshop on Massively Parallel Sequencing (RECOMB-Seq)*, 2017.

[37] W. W. B. GOH, W. WANG, and L. WONG. Why batch effects matter in omics data, and how to avoid them. *Trends in Biotechnology,* 2017.

[38] W. W. B. GOH and L. WONG. Advancing clinical proteomics via analysis based on biological complexes: A tale of five paradigms. *Journal of proteome research,* 15, 9, pp. 3167–3179, 2016.

[39] W. W. B. GOH and L. WONG. Design principles for clinical network-based proteomics. *Drug discovery today,* 21, 7, pp. 1130–1138, 2016.

[40] W. W. B. GOH and L. WONG. Evaluating feature-selection stability in next-generation proteomics. *Journal of Bioinformatics and Computational Biology,* 14, 05, pp. 1650029:1–23, 2016.

[41] W. W. B. GOH and L. WONG. Integrating Networks and Proteomics: Moving Forward. *Trends in Biotechnology,* 34, 12, pp. 951–959, 2016.

[42] W. W. B. GOH and L. WONG. Protein complex-based analysis is resistant to the obfuscating consequences of batch effects---a case study in clinical proteomics. *BMC genomics,* 18, 2, pp. 142, 2017.

[43] W. W. B. GOH and L. WONG. Spectra-first feature analysis in clinical proteomics—A case study in renal cancer. *Journal of Bioinformatics and Computational Biology,* 14, 05, pp. 1644004:1–18, 2016.

[44] M. GOLUMBEANU. Dynamics of Proteo-Transcriptomic Response to HIV-1 Infection. In preparation.

[45] M. GOLUMBEANU. Transcriptome-wide identification of RNA-protein binding. In preparation.

[46] D. GUSFIELD. *Integer Programming In Computational Biology: An entry-level text for biologists*. Book in preparation.

[47] D. GUSFIELD. The matrix-chain multiplication problem revisited. In preparation.

[48] D. HAFEZ, A. KARABACAK, S. KRUEGER, Y.-C. HWANG, L.-S. WANG, R. P. ZINZEN, and U. OHLER. Predicting gene expression via semi-supervised assignments of enhancers to target genes. Revision submitted to Genome Biology.

[49] I. HAJIRASOULIHA. Leveraging tumor lineage trees to predict somatic structural variations, using paired-end

sequencing. Preprint, 2016.

[50] T. HANSEN and F. VANDIN. Finding Mutated Subnetworks Associated with Survival in Cancer. *arXiv preprint arXiv:1604.02467,* 2016.

[51] Z. HUANG, H. DING, and J. XU. Faster Algorithm for Truth Discovery via Range Cover. In *Algorithms and Data Structures Symposium (WADS)*, 2017.

[52] K. JAHN, J. KUIPERS, and N. BEERENWINKEL. Tree inference for single-cell data. *Genome Biology,* 17, 2016.

[53] A. JAVANMARD, A. BHASKAR, T. COURTADE, and D. TSE. Novel probabilistic models of spatial genetic ancestry with applications to stratification correction in genome-wide association studies. *Bioinformatics,* 33, 6, pp. 879–885, 2017.

[54] G. M. KAMATH, I. SHOMORONY, F. XIA, T. A. COURTADE, and D. N. TSE. HINGE: long-read assembly achieves optimal repeat resolution. *Genome research,* 27, 5, pp. 747–756, 2017.

[55] C. KOMUSIEWICZ, M. DE OLIVEIRA OLIVEIRA, and M. ZEHAVI. Revisiting the Parameterized Complexity of Maximum-Duo Preservation String Mapping. In *Proceedings of the 28th Annual Symposium on Combinatorial Pattern Matching (CPM)*, 2017.

[56] J. KUIPERS, K. JAHN, B. J. RAPHAEL, and N. BEERENWINKEL. A statistical test on single-cell data reveals widespread recurrent mutations in tumor evolution. Preprint, https://doi.org/10.1101/094722, 2016.

[57] M. LE MORVAN, A. ZINOVYEV, and J.-P. VERT. NetNorM: capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. *PLOS Computational Biology,* 2017.

[58] H. LEE, J. GURTOWSKI, S. YOO, M. NATTESTAD, S. MARCUS, S. GOODWIN, W. R. MCCOMBIE, and M. SCHATZ. Third-generation sequencing and the future of genomics. *bioRxiv,* 048603, 2016.

[59] Y. LIU, H. DING, D. CHEN, and J. XU. Novel Geometric Approach for Global Alignment of PPI Networks. In *Proceedings of The 31st AAAI Conference on Artificial Intelligence (AAAI)*, pp. 31–37, 2017.

[60] Y. LIU, H. DING, Z. HUANG, and J. XU. Distributed and Robust Support Vector Machine. In *Proceedings of The 27th International Symposium on Algorithms and Computation (ISAAC)*, pp. 54:1-54:13, 2016.

[61] D. LOKSHTANOV, A. E. MOUAWAD, S. SAURABH, and M. ZEHAVI. Packing Cycles Faster Than Erdos-Posa. In *Proceedings of the 44th International Colloquium on Automata, Languages and Programming (ICALP)*, 2017.

[62] X. LU and S. KURIKI. Simultaneous confidence bands for contrasts between several nonlinear regression curves. *Journal of Multivariate Analysis,* 155, pp. 83–104, 2017.

[63] G. MARCAIS, D. PELLOW, D. BORK, Y. ORENSTEIN, R. SHAMIR, and C. KINGSFORD. Improving the performance of minimizers and winnowing schemes. *bioRxiv,* 104075, 2017. To appear in Intelligent Systems for Molecular Biology (ISMB), 2017.

[64] A. MARON-KATZ, D. AMAR, E. B. SIMON, T. HENDLER, and R. SHAMIR. RichMind: A Tool for Improved Inference from Large-Scale Neuroimaging Results. *PLOS One,* 11, 7, pp. e0159643:1–14, 2016.

[65] R. MIRAGAIA, X. ZHANG, V. SVENSSON, T. ILICIC, J. HENRIKSSON, and T. LONNBERG. Single-cell RNA sequencing reveals crosstalk of self-antigen expression and differentiation in thymic medullary epithelial cells. Scientific Reports, in revision.

[66] N. MUKHERJEE, L. CALVIELLO, A. HIRSEKORN, S. D. PRETIS, M. PELIZZOLA, and U. OHLER. Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nature Structural & Molecular Biology,* 24, pp. 86–96, 2017.

[67] E. VAN NIMWEGEN. Inferring intrinsic and extrinsic noise from a dual fluorescent reporter. *bioRxiv,* 049486, 2016.

[68] V. NTRANOS, G. M. KAMATH, J. M. ZHANG, L. PACHTER, and N. T. DAVID. Fast and accurate single-cell RNA-Seq analysis by clustering of transcript-compatibility counts. *Genome biology,* 17, 1, 2016.

[69] I. NUMANAGIĆ, J. K. BONFIELD, F. HACH, J. VOGES, J. OSTERMANN, C. ALBERTI, M. MATTAVELLI, and S. C. ŞAHINALP. Comparison of high-throughput sequencing data compression tools. *Nature Methods,* 5, pp. 1005–1008, 2016.

[70] S. OMIDI and E. VAN NIMWEGEN. Automated Incorporation of Pairwise Dependency in Transcription Factor Binding Site Prediction Using Dinucleotide Weight Tensors. *bioRxiv,* 078212, 2016. PLOS Computational Biology (2017, accepted).

[71] Y. ORENSTEIN, U. OHLER, and B. BERGER. Structural analysis of the compendium of RNA-binding proteins. In preparation.

[72] Y. ORENSTEIN, D. PELLOW, G. MARÇAIS, R. SHAMIR, and C. KINGSFORD. Compact universal k-mer hitting sets. Preprint. Submitted to WABI 2017.

[73] Y. ORENSTEIN, R. PUCCINELLI, R. KIM, P. FORDYCE, and B. BERGER. Joker de Bruijn: sequence libraries to cover all k-mers using joker characters. In revision for Cell Systems.

[74] Y. ORENSTEIN and R. SHAMIR. Modeling protein-DNA binding via high throughput in vitro technologies. In preparation.

[75] F. PANOLAN and M. ZEHAVI. Parameterized Algorithms for List K-Cycle. In *Proceedings of the 36th Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, pp. 22:1-22:15, 2016.

[76] J. PARK, D. BIANCHI, J. MARON, and D. SLONIM. Pathway centrality in protein interaction networks identifies functional mediators of pulmonary disease. Manuscript in preparation.

[77] M. C. PIETRAS, L. HAYDEN, J. MARON, D. BIANCHI, and D. SLONIM. TEMPO: analyzing differences in temporal patterns of gene expression. Manuscript in preparation.

[78] M. PIRKL, M. DIEKMANN, M. VAN DER WEES, N. BEERENWINKEL, H. FRÖHLICH, and F. MARKOWETZ. Inferring modulators of genetic interactions with epistatic nested effects models. *PLOS Computational Biology,* 13, 4, pp. e1005496:1–18, 2017.

[79] A. M. POOS, A. MAICHER, A. K. DIECKMANN, M. OSWALD, R. EILS, M. KUPIEC, B. LUKE, and R. KÖNIG. Mixed Integer Linear Programming based machine learning approach identifies regulators of telomerase in yeast. *Nucleic Acids Research,* 44, 10, pp. e93:1–9, 2016.

[80] V. POPIC and S. BATZOGLOU. Privacy-Preserving Read Mapping Using Locality Sensitive Hashing and Secure Kmer Voting. *bioRxiv,* 046920, 2016. Under review.

[81] D. RISSO, F. PERRAUDEAU, S. GRIBKOVA, S. DUDOIT, and J.-P. VERT. ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data. *bioRxiv,* 125112, 2017.

[82] T. SAKOPARNIG, C. FIELD, and E. VAN NIMWEGEN. The dominance of recombination in E. coli genome evolution: why clonal ancestry cannot be recovered from genomic data. In preparation.

[83] D. SEN, J. KAMINSKI, R. A. BARNITZ, M. KURACHI, E. J. WHERRY, N. YOSEF, and W. N. HAINING. The Epigenetic Landscape of T Cell Exhaustion. Preprint, 2016. Under review.

[84] R. SHAMIR, M. ZEHAVI, and R. ZEIRA. A Linear-Time Algorithm for the Copy Number Transformation Problem. In *Proceedings of the 27th Annual Symposium on Combinatorial Pattern Matching (CPM)*, pp. 16:1-16:13, 2016.

[85] I. SHOMORONY, T. COURTADE, and D. TSE. Fundamental Limits of Genome Assembly under an Adversarial Erasure Model. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications,* 2, 2, pp. 199–208, 2016.

[86] I. SHOMORONY, G. M. KAMATH, F. XIA, T. A. COURTADE, and N. T. DAVID. Partial DNA assembly: a rate-distortion perspective. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pp. 1799–1803, 2016.

[87] I. SHOMORONY, S. KIM, T. COURTADE, and D. TSE. Information-Optimal Genome Assembly via Sparse Read-Overlap Graphs. *Bioinformatics,* 32, 17, pp. i494–i502, 2016.

[88] E. SZCZUREK and N. BEERENWINKEL. Linear effects models of signaling pathways from combinatorial perturbation data. *Bioinformatics,* 32, 12, pp. i297–i305, 2016.

[89] R. THOMAS, S. THOMAS, A. K. HOLLOWAY, and K. S. POLLARD. Features that define the best ChIP-seq peak calling algorithms. *Briefings in Bioinformatics,* 18, 3, pp. 441–450, 2016.

[90] F. VANDIN and N. BEERENWINKEL. FDR control for TiMEx. In preparation.

[91] S. VELASCO, M. M. IBRAHIM, A. KAKUMANU, G. GARIPLER, M. A. AL-SAYEGH, A. HIRSEKORN, F. ABDUL-RAHMAN, R. SATIJA, U. OHLER, S. MAHONY, and E. O. MAZZONI. A multi-step transcriptional and chromatin state cascade underlies motor neuron programming from embryonic stem cells. *Cell Stem Cell,* 20, pp. 205–217, 2017.

[92] A. WAGNER, A. REGEV, and N. YOSEF. Uncovering the vectors of cellular identity with single cell genomics. *Nature Biotechnology,* 2016.

[93] R. WANG, H. HUANG, and M. WATERMAN. Generalized correlation measure using count statistics for time-course data. Submitted to Bioinformatics.

[94] S. WHALEN, R. M. TRUTY, and K. S. POLLARD. nhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics,* 48, pp. 488–496, 2016.

[95] L. YANG, Y. ORENSTEIN, A. JOLMA, J. TAIPALE, R. SHAMIR, and R. ROHS. DNA shape readout

specificities of different transcription factor families. In preparation.

[96] L. YANG, Y. ORENSTEIN, A. JOLMA, Y. YIN, J. TAIPALE, R. SHAMIR, and R. ROHS. Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Molecular Systems Biology,* 13, 2, pp. 910, 2017.

[97] M. YE, X. ZHANG, G. C. RACZ, Q. JIANG, and B. M. MORET. NEMo: An Evolutionary Model With Modularity for PPI Networks. *IEEE Transactions on Nanobioscience,* 16, 2, pp. 131–139, 2017.

[98] S. ZACCARIA, M. EL-KEBIR, G. W. KLAU, and B. J. RAPHAEL. The Copy-Number Tree Mixture Deconvolution Problem and Applications to Multi-sample Bulk Sequencing Tumor Data. In *International Conference on Research in Computational Molecular Biology (RECOMB)*, pp. 318–335, 2017.

[99] M. ZEHAVI. Parameterized Approximation Algorithms for Packing Problems. In *Theoretical Computer Science (TCS)*, vol. 648, pp. 40–55, 2016.

[100] M. ZEHAVI. A randomized algorithm for long directed cycle. *Information Processing Letters,* 116, 6, pp. 419–422, 2016.

[101] X. ZHANG, V. SVENSSON, J. K. KIM, J. C. MARIONI, and S. A. TEICHMANN. Scoring the isoform variation in single cell RNA-seq data. In preparation.