Murray Patterson
1 Park Place, Room 636
Atlanta, Georgia 30303 (USA)
mpatterson30@gsu.edu
murraypatterson.github.io

Research Statement

My specialty is solving bioinformatics problems which require complex mathematical modeling and advanced knowledge of algorithmics. I have several years of experience addressing important biological problems in the area of sequence analysis such as haplotype assembly [6, 40, 32], and evolutionary molecular biology such as the correlated evolution of genes [17, 41], proteins [39, 18], and more recently, cancer progression [13, 45, 12]. Many of the modeling aspects of these works have been possible because of my solid background in the quantitative sciences and mathematics – particularly in combinatorics [34, 33] and algorithmics [35, 10], from both a theoretical and experimental standpoint. In addition to this, I have worked in several different departments, cultural settings and languages – each one providing a unique perspective on scientific research. All of these have given me the versatility to apply computer science concepts in very diverse environments and to very different fields of study. I believe that this makes me a good candidate for leveraging the full potential from an interdisciplinary field such as bioinformatics, and for bridging the several departments and disciplines necessary to do so.

The general workflow that I apply begins with identifying a relevant biological problem which could benefit from an algorithmic solution – often by talking to biologists and other bioinformaticians. This is followed by the detailed analysis of the data that could be useful for solving this problem, and then by proposing rigorous but realistic models. The final goal is an efficient and usable implementation for the aspects of the model which are computationally tractable. For those aspects of the model that are too computationally intensive to solve, often one can make simplifying assumptions to the model or apply algorithm engineering approaches which reflect the data being processed. My research has an impact on bioinformatics, because there are many relevant biological problems that need algorithmic solutions, but in particular because of my experience in every step of the process from the initial formulation all the way to a production-quality tool which solves this problem. Important practical problems which require complex mathematical modeling that are then made widely available to the community via usable implementations and scalable pipelines is something researchers could use more of – something I strive to make a reality. I now outline some of the main projects I am currently involved in, or have worked on in the past.

Haplotype Assembly. One of the best examples of my approach is a project on assembling haplotypes from sequencing reads data. The haplotypes of an individual are the various *copies* of its genome sequence. In the case of humans, which have two copies, one from each parent, knowing the haplotypes is informative for population genetics [8], and clinical decision-making [23]. The reason this is so important, is because the relative *orientation* of defects on the genome – whether a pair of genetic defects occurs on the same haplotype sequence or different ones – can mean the difference between a minor and a very major genetic disorder [44]. A sequencing read is a segment of one of the haplotype sequences of an individual. This is a great source of information because it provides

evidence that a set of genetic defects are together on some haplotype – the longer the reads, the larger the sets of defects linked. With a sizeable set of overlapping reads, one can then "assemble" them together to obtain entire haplotype sequences, hence finding the relative orientation of sets of genetic defects, specific to the sequenced individual, that are important for serious disorders. Typically, this is done by finding a *partitioning* of a set of reads into groups (one for each haplotype), in such a way that the number of inconsistencies among overlapping reads within each group is minimized. Note that, because of errors in the sequencing reads themselves, that this number of inconsistencies can never be zero – something that is theoretically possible in the absence of errors. Interestingly, it is because of these errors that this is a computationally difficult combinatorial problem [14].

The idea of assembling haplotypes from reads has existed for almost two decades [29], and there is even an early heuristic approach [38] to the problem. After this problem was formalized in [14] as the *minimum error correction* (MEC) problem, the next decade of research focused on various algorithmic approaches for the MEC problem [25, 11]. In recent years, the advent of long read technologies such as the Pacific Biosciences (PacBio) platform has resulted in a more than tenfold increase in sequencing read length. Because the sets of genetic defects that can be linked together has increased by an order of magnitude, this jump in read length has made haplotype assembly more practically relevant, sparking a renewed attention to methods for the MEC problem. While this problem remains computationally difficult in general, one can consider the distribution of the number of reads, *i.e.*, the read *coverage*, at any given location on the genome sequence to restrict the problem, providing a reasonably efficient solution to the MEC problem. This was the key observation I made that led to the first method and efficient proof-of-principle (`C++`) implementation for haplotype assembly from long reads called WhatsHap [40]. Since then, we have surrounded this core algorithm with an interface (`python`) that is easy to use and deals with standard file formats. WhatsHap is now a production-quality tool [32] that can be installed and used in just one line each. Most recently, we have devised an even more efficient implementation [6] of the core phasing algorithm by considering the distribution of the number of *errors* in the reads at any given location on the genome sequence to restrict the problem. This assumption can be made because PacBio reads exhibit a uniform error rate over the entire genome sequence [9]. Because this number of errors at a given location on the genome is usually much smaller than the read coverage at this location, this leads to a much more efficient algorithm, since it avoids exploring potential haplotype assemblies that already have an unrealistically elevated number of errors.

A future research direction is to integrate other sources of information in addition to sequencing reads in order to further boost both the quality and speed of haplotype assembly. One deficiency of sequencing reads is that, despite the fact that read lengths continue to increase with each new generation of sequencing technology, they still do not cover the entire genome sequence. On the other hand, haplotype reference panels [15] – a mosaic of haplotypes over the human genome sequence, compiled from several tens of thousands of individuals – for example, *do* cover the entire genome sequence. The methods that make use purely of reference panels take a statistical "walk" [28] over this mosaic from the beginning to the end of the sequence to determine the haplotypes. State-of-the-art statistical methods [30] make use of the latest panels [15], however what they can determine is limited to the population used to compile these panels. Rare genetic defects that are specific to an individual of study, for example, will only be found in the sequencing reads for this individual. So indeed, these two sources of information nicely complement each other, and approaches

that take both into account are already starting to appear [16], but this is only the beginning. One important resource that will help achieve this goal is a collaboration that I cultivated with the Genome in a Bottle (GIAB) Consortium [48], which began when they started using the WhatsHap tool [32]. We now communicate on a regular basis with one of the senior researchers of GIAB regarding the usage of our tools on the newest data that they make available, or what features we could add to our tools based on this data. So we are already well on the way to the integration of other information sources such as the one mentioned above – their continued communication and guidance in this process will be a great asset to this future research endeavor.

Another future direction of this work is the assembly of viral quasispecies. Computationally, this problem is similar to haplotype assembly, the additional challenge being that the number of haplotypes, the *ploidy*, is unknown a priori (one for each strain) and can be on the order of hundreds [2]. Recent methods have tackled this problem with overlap graphs [2], semi-definite programming [5] and tensor factorization [1], however, a machine learning technique such as correlation clustering [4] seems most promising, since it optimizes the clustering (*i.e.*, the haplotype assembly) *and* the number of such clusters (*i.e.*, the ploidy) *simultaneously*. Note that [5] infers the ploidy in a separate step.

Cancer Progression. Another project that I am involved in – due to its immediate and novel application, and the background in evolutionary biology I have which can advance this project – is evolutionary models of cancer progression. A tumor, when detected, contains a heterogeneous mixture of cells, which is a result of an intense and erratic process. This process, however, is an (extremely rapid) evolutionary process of cancer cells, called *clones*, which proliferate and differentiate from the founding clone of the cancer [37]. Understanding this clonal evolutionary history can help clinicians understand the cell heterogeneity in the tumors of various types of cancer and, more importantly, gives insights on how to devise therapeutic strategies [36, 46]. While this setting is quite different from classical phylogenetics, hence also the phylogenies we reconstruct for cancer, the goal is the same: to infer an evolutionary tree on genes and genomes, and it can therefore benefit from my background. In particular, my doctoral thesis concerns some of the mathematical aspects of gene rearrangement, which I later applied to a study on bacteria [41], later developing a more general framework for rearrangement and co-evolving genes [35, 17]. This background can provide insight to the genes and rearrangements that underpin somatic mutations [22]. I have also studied the co-evolution of proteins and metabolic function [18], later doing a more in depth study on the correlated evolution of genes in terms of enzyme presence, and genomes in terms of the metabolic network of the corresponding organism [39]. This study on an orthogonal view of evolution, and the resulting experience with systems biology, provides me another important line of attack on cancer progression research [3].

Because of its affordability and widespread use, most techniques infer the clonal evolutionary history of cancer based on next-generation bulk sequencing data [19, 31, 43]. Since bulk-sequencing data is already at a fairly coarse resolution, these models, mostly for the sake of simplicity, make the *infinite sites assumption* (ISA), *i.e.*, each genomic site undergoes at most one mutation during the clonal evolution of a cell. This implies that the evolutionary history is a *perfect phylogeny*, allowing efficient algorithms for its inference [24]. While *single cell sequencing* (SCS) technologies provide the level of resolution that we need to better understand clonal evolution, it remains an expensive task, and is still plagued by high dropout and false-negative rates. Nonetheless, with the rapid improvement of this technology, coupled with decreasing prices, tools based on SCS data are beginning to appear [26, 42].

While these methods still make the ISA, even more recent studies have shown evidence of widespread recurrence and loss of mutations in tumor evolution [7, 27], thanks to the resolution of SCS. One reason that the above methods can still make the ISA is because the dropout rate of SCS is still sufficiently high that the amount of missing data allows the fitting of a perfect phylogeny. But as the technology rapidly improves, this will no longer be possible either, and so a more general model beyond perfect phylogeny is needed.

In [45], we proposed some phylogenetic models which slightly relax the perfect phylogeny – allowing a limited amount of recurrence or loss of mutation. One of the better models, the Dollo-k model – a restriction of the more general model [20] – is versatile because it allows a number (k) of independent losses of each mutation in the inferred phylogeny, and so we implemented a tool for inferring cancer progression which solves an *integer linear programming* (ILP) formulation of this model [13]. Now we are developing an approach which infers Dollo-k phylogenies using Simulated Annealing [12]. The resulting tool outperforms state-of-the-art tools for phylogeny inference from SCS data that take [47] or do not take [26] into account mutation losses. Most recently, we have added to our model the possibility of mutation-dependent false-negative rates – due to the various read coverage and (mutation) expression levels of different genomic loci – the first tool to consider this. A future direction of this work could be to complement this inference with information about the metabolism based on (single cell) RNA-seq data. This would bolster the SCS data, but could also provide new insights.

Gene Rearrangement. In [41], I developed an approach for most parsimoniously inferring gene adjacencies in the internal nodes of a phylogenetic tree from the genes adjacent in the sequences of the extant species at the leaves – a generalization of Fitch's algorithm [21] – effectively reconstructing ancestral genomes, which I applied to a study on cyanobacteria. The simplifying assumption that leads to an efficient algorithm is that adjacencies are independent – in the few cases where this causes a conflict, we can efficiently *linearize* [35] the constructed ancestral genome. We later combined this into a more general and easy-to-use framework [17]. Interestingly, this framework and background could be used to understand gene rearrangement in cancer progression [22], providing new insights.

Summary. With my background in algorithmics which I now apply to the latest problems in bioinformatics, coupled with my experience in a variety of settings, I am well-versed in taking on the challenges of this field that spans biology, mathematics, computer science, health and translational medicine. Because the application of algorithms and computation to the field of biology, medicine, etc., is relatively new, there is a need for mathematical models, methods and tools. For this reason, I have become very adept at designing a good model for a biological problem, requiring abstract reasoning and a high-level understanding of the problem, independent of any programming language or implementation. Consequently, I have also developed the skills to then take such a model, and make its mathematical devices work well in practice, through empirical algorithmics, software engineering, and efficient scalable implementation. These skills intersect a variety of computer science, mathematics and engineering concepts, allowing me to teach this wide variety of courses. What is more important however, is that biologists and clinicians are also new to algorithms and computation, and so I have gained the deep understanding of these topics and the experience needed to convey their concepts to this more general audience. Moreover, I have several ongoing research projects with important applications, which have gained plenty of momentum in terms of publications and worldwide collaborators. The future advancements and directions of these and related projects are

equally exciting and challenging, and I look forward to engaging students in these research-level activities, and also the collaboration with researchers and institutions to carry this out.

References

[1] Soyeon Ahn, Ziqi Ke, and Haris Vikalo. Viral quasispecies reconstruction via tensor factorization with successive read removal. *Bioinformatics*, 34(13):i23–i31, 2018.

[2] Jasmijn A. Baaijens, Amal Zine El-Aabidine, Eric Rivals, and Alexander Schönhuth. De novo assembly of viral quasispecies using overlap graphs. *Genome Research*, 27(5):835–848, 2017.

[3] Mehmet G. Badur and Christian M. Metallo. Reverse engineering the cancer metabolic network using flux analysis to understand drivers of human disease. *Metabolic Engineering*, 45:95–108, 2018.

[4] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2004.

[5] Somsubhra Barik, Shreepriya Das, and Haris Vikalo. QSdpR: Viral quasispecies reconstruction via correlation clustering. *Genomics*, 2017.

[6] Stefano Beretta, Murray Patterson, Simone Zaccaria, Gianluca Della Vedova, and Paola Bonizzoni. HapCHAT: Adaptive haplotype assembly for efficiently leveraging high coverage in long reads. *BMC Bioinformatics*, 19(1):252, 2018.

[7] David Brown, Dominiek Smeets, Borbála Székely, et al. Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. *Nature Communications*, 8:14944, 2017.

[8] Sharon R. Browning and Brian L. Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.

[9] Mauricio O. Carneiro, Carsten Russ, Michael G. Ross, Stacey B. Gabriel, Chad Nusbaum, and Mark A. DePristo. Pacific Biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, 13(1):375, 2012.

[10] Cedric Chauve, Ján Maňuch, and Murray Patterson Roland Wittler. Tractability results for the consecutive-ones property with multiplicity. In *Combinatorial Pattern Matching: 22nd Annual Symposium (CPM)*, volume 6661 of *Lecture Notes in Computer Science (LNCS)*, pages 90–103, 2011.

[11] Zhi-Zhong Chen, Fei Deng, and Lusheng Wang. Exact algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics*, 29(16):1938–1945, 2013.

[12] Simone Ciccolella, Mauricio Soto Gomez, Murray Patterson, Gianluca Della Vedova, Iman Hajirasouliha, and Paola Bonizzoni. Inferring cancer progression from single cell sequencing while allowing loss of mutations. *bioRxiv*, 268243, 2018.

[13] Simone Ciccolella, Mauricio Soto, Murray Patterson, Gianluca Della Vedova, Iman Hajirasouliha, and Paola Bonizzoni. gpps: An ILP-based approach for inferring cancer progression with mutation losses from single cell data. In *IEEE Computational Advances in Bio and medical Sciences, 8th International Conference (ICCABS)*, 2018. to appear.

[14] Rudi Cilibrasi, Leo van Iersel, Steven Kelk, and John Tromp. On the complexity of several haplotyping problems. In *Algorithms in Bioinformatics, 5th International Workshop (WABI)*, pages 128–139, 2005.

[15] The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48:1279–1283, 2016.

[16] Olivier Delaneau, Jonathan Marchini, and The 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 genomes project haplotype reference panel. *Nature Communications*, 5(3934), 2014.

[17] Wandrille Duchemin, Yoann Anselmetti, Murray Patterson, Yann Ponty, Sèverine Bérard, Cedric Chauve, Celine Scornavacca, Vincent Daubin, and Eric Tannier. DeCoSTAR: Reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. *Genome Biology and Evolution*, 9(5):1312–1319, 2017.

[18] Mohammed El-Kebir, Tobias Marschall, Inken Wohlers, Murray Patterson, Jaap Heringa, Alexander Schönhuth, and Gunnar W. Klau. Mapping proteins in the presence of paralogs using units of coevolution. *BMC Bioinformatics*, 14(15):S18, 2013.

[19] Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J. Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015.

[20] J.S. Farris. Phylogenetic analysis under Dollo's law. *Systematic Biology*, 26(1):77–88, 1977.

[21] W.M. Fitch. Toward defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology*, 20:406–416, 1971.

[22] Simon A Forbes, David Beare, Prasad Gunasekaran, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(Database issue), 2015.

[23] Gustavo Glusman, Hannah C. Cox, and Jared C. Roach. Whole-genome haplotyping approaches and genomic medicine. *Genome Medicine*, 6(9):73, 2014.

[24] Dan Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21(1):19–28, 1991.

[25] Dan He, Arthur Choi, Knot Pipatsrisawat, Andan Darwiche, and Eleazar Eskin. Optimal algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics*, 26(12):i183–i190, 2010.

[26] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome Biology*, 17(1):86, 2016.

[27] Jack Kuipers, Katharina Jahn, Benjamin J. Raphael, and Niko Beerenwinkel. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Research*, 27:1885–1894, 2017.

[28] Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

[29] Ross Lippert, Russell Schwartz, Guiseppe Lancia, and Sorin Istrail. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Briefings in Bioinformatics*, 3(1):23–31, 2002.

[30] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, et al. Reference-based phasing using the haplotype reference consortium panel. *Nature Genetics*, 48:1443–1448, 2016.

[31] S. Malikic, A.W. McPherson, N. Donmez, and S.C. Sahinalp. Clonality inference in multiple tumor samples using phylogeny.

Bioinformatics, 31(9):1349–1356, 2015.

[32] Marcel Martin, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schönhuth, and Tobias Marschall. WhatsHap: fast and accurate read-based phasing. *bioRxiv*, 085050, 2016.

[33] Ján Maňuch and Murray Patterson. The complexity of the gapped consecutive-ones property problem for matrices of bounded degree. *Journal of Computational Biology*, 18(9):1243–1253, 2011.

[34] Ján Maňuch, Murray Patterson, and Cedric Chauve. Hardness results on the gapped consecutive-ones property problem. *Discrete Applied Mathematics*, 160(18):2760–2768, 2012.

[35] Ján Maňuch, Murray Patterson, Roland Wittler, Cedric Chauve, and Eric Tannier. Linearization of ancestral multichromosomal genomes. *BMC Bioinformatics*, 13(19):S11, 2010.

[36] A. Sorana Morrissy, Livia Garzia, David J. H. Shih, et al. Divergent clonal selection dominates medulloblastoma at recurrence. *Nature*, 529(7586):351–357, 2015.

[37] Peter C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[38] Alessandro Panconesi and Mauro Sozio. Fast Hare: A fast heuristic for single individual SNP haplotype reconstruction. In *Algorithms in Bioinformatics: 4th International Workshop (WABI)*, pages 266–277, 2004.

[39] Murray Patterson, Thomas Bernard, and Daniel Kahn. Correlated evolution of metabolic functions over the tree of life. *bioRxiv*, 093591, 2016.

[40] Murray Patterson, Tobias Marschall, Nadia Pisanti, Leo van Iersel, Leen Stougie, Gunnar W. Klau, and Alexander Schönhuth. WhatsHap: Weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology*, 22(6):498–509, 2015.

[41] Murray Patterson, Gergely Szöllősi, Vincent Daubin, and Eric Tannier. Lateral gene transfer, rearrangement, reconciliation. *BMC Bioinformatics*, 14(15):S4, 2013.

[42] Edith M. Ross and Florian Markowetz. Onconem: inferring tumor evolution from single-cell sequencing data. *Genome Biology*, 17(1):69, 2016.

[43] F. Strino, F. Parisi, M. Micsinai, and Y. Kluger. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Research*, 41(17):e165, 2013.

[44] Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J. Topol, and Nicholas J. Schork. The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215–223, 2011.

[45] Gianluca Della Vedova, Murray Patterson, Raffaella Rizzi, and Mauricio Soto. Character-based phylogeny construction and its application to tumor evolution. In *Unveiling Dynamics and Complexity: 13th Conference on Computability in Europe (CiE)*, volume 10307 of *Lecture Notes in Computer Science*, pages 3–13, 2017.

[46] Jiguang Wang, Emanuela Cazzato, Erik Ladewig, et al. Clonal evolution of glioblastoma under therapy. *Nature Genetics*, 48(7):768–776, 2016.

[47] Hamim Zafar, Anthony Tzen, Nicholas Navin, Ken Chen, and Luay Nakhleh. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biology*, 18(1):178, 2017.

[48] Justin M. Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, 32(3):246–251, 2014.