Postdoc Proposal

Shelby Solomon Darnell[1]

Abstract
The Panorama project is an NSF funded large collaborative project that brings together computing hardware and software scientists and engineers to work on a bespoke hardware and software to better manage pangenomic computing.

Keywords
Affective Genomics — Computational pangenomics — bespoke computing architecture — Adaptive Artificial systems — Causal inference

[1] *Department of Genetics, Genomics and Informatics, UTHSC, Memphis, TN, United States of America*
*__Corresponding author__: shelby@shelbydarnell.com

Contents

Introduction

My community is medically under-served and statistically underrepresented in the (health) sciences. Recently I started attending a meeting with an NSF project run by Prins a.o. to build a special pangenome supercomputer in collaboration with Cornell. Unsurprisingly, with 25 people attending, I am the only black person in the room. At a personal level, my young and non-obese sister has lost both kidneys because of complications from diabetes and hypertension, and I constantly ask myself 'why' or what are the causal factors leading to such disaster. Which further leads me to keep track of my own susceptibility to the same health issues, in order to avoid a similar disaster, while asking 'what could she have done differently to avoid the cataclysm?' What can I do with my computer science background to support designer medicine for anyone who needs it. These questions led me to engage Pjotr Prins and my interest in pursuing research in pangenomes.

According to HudsonAlpha, 'Pangenomes present all of the genes and DNA sequences within a species'. At the beginning of research and application of genetic information, once the human genome was fully sequenced, a single persons genome was used as the reference genome due to the time and expense of genetic sequencing. These reference genomes are used as comparators for disease or trait causes DNA changes and identifiers, cross species similarity, and other aspects of genomic research. When considering that certain disease and health issues are related to specific populations and different groups and people have different susceptibility to illness, having a single individuals genome representing all of humanity, the quality of the discovery is as limited as the diversity in a single person's genetic information. Hence the firm push, with funding for the development of as diverse a pangenome as possible, and the development of proper tools to support genomic research. Sucn a pursuit leads us to the human genome project.

The human genome project, funded by the **N**ational **H**uman **G**enome **R**esearch **I**nstitute (NHGRI), is pursuing many different avenues to meet their goal, of creating a pangenome that represents the world's diversity. Thankfully researchers at Cornell and UTHSC are developing a specialized computer with optimized algorithms for exploring pangenomes. Many avenues are being pursued while Cornell leads researchers toward this goal, some of which are: privacy preserving pangenomics, pangenome visualization, and broader impacts. Supporting the development of a pangenome computer and its complimenting bespoke software will enable the development of optimzed designer medicine and therapeutics in the future; hence, my excitement of pursuing this change in my career focus. These topics are mentioned as I intend to contribute to

these and expand of the same in this proposal.

In addition to working on topics about which I am passionate, one of the main purposes of a postdoc is to enable the all-around professional growth of a young researcher. Hence the results of an individual development plan will be combined with my research interests to determine the outcomes of this postdoctoral research.

1. Appointment

My appointment is with the University of Tennessee Health Sciences Center (UTHSC) in Memphis, Tennessee in the Genetic, Genomics, and Informatics group under the Panorama Project being done in conjunction with Christopher Batten at Cornell. My direct mentors will be Associate Professor Pjotr Prins (UTHSC), Associate Professor Christopher Batten (Cornell) and Professor Saunak Sen (UTHSC). The duration of the postdoc is up to two years, from March 2023 thru to end of February 2025.

2. Strengths & Weaknesses

Science careers has a tool to aide young researchers in building an individual development plan, called myIDP.

2.1 Strong skills

1. basic writing and editing
2. writing for nonscientists
3. speaking clearly and effectively
4. presenting to nonscientists
5. demonstrating workplace etiquette
6. complying with rules and regulations
7. maintaining positive relationships with colleauges
8. providing instruction and guidance
9. creating vision and goals
10. serving as a role model
11. careful recordkeeping practices
12. understanding of data ownership/sharing issues
13. demonstrating responsible authorship and publication practices
14. demonstrating responsible conduct in human research
15. how to maintain a professional network
16. technical skills related to my specific research area

2.2 Weak Skills & Deficiencies

1. contributing to institution (e.g. participate on committees)
2. demonstrating responsible conduct in animal research
3. how to negotiate
4. statistical analysis
5. seeking advice from advisors and mentors
6. developing/managing budgets
7. how to interview
8. delegating responsibilities
9. writing grant proposals
10. planning and organizing projects
11. navigating the peer review process

As a postdoc is usually 2-3 years long, a schedule of activities will be planned and evaluated by my mentors. This schedule will include learning activities, speaking engagements, research, writing, collaboration, and working on specific weak skills and deficiencies. Concerning the research component of the plan, three major aims will be

1. Differential privacy algorithm development and testing

Aim Aide in development of new pangenome layout algorithms

Aim Support broader impact initiatives

3. Background

3.1 Panorama Project

The Panorama project is a five year NSF funded effort to create the first integrated rack scale acceleration paradigm specifically for computational pangenomics. Christopher Batten and a team of seven primary investigators, including Pjotr Prins, currently lead this effort.

3.1.1 Problem

It has become necessary for computers to attempt analysis of massive datasets which need to be manipulated in irregular and rapidly changing ways while ensuring strict privacy guarantees. The ability to efficiently support large, sparse, dynamic yet private data for solving big complex problems on hetergeneous systems is one of the grand challenges in software/hardware systems research. Addressing this grand challenge is the Panarama project by way of the exploration of integrated rack scale acceleration for computational genomics. Integrated rack-scale acceleration refers to an emerging computer-systems paradigm that uses tens of tightly integrated computing nodes, each of which includes a mix of general-purpose processors and specialized accelerators interconnected with a special-purpose network. Computational pangenomics refers to a recent trend towards representing genomes, the genetic material of an organism, not as a single linear sequence of DNA base pairs but instead as an intricate network of sequences that efficiently represents the relationships between many individuals' genomes at once. Computational pangenomics naturally captures the trend towards big, sparse, dynamic, and private data and is thus a perfect application domain to explore heterogeneous software/hardware systems research.

3.1.2 Novelties

The project's novelties are: a truly cross-stack approach spanning applications, programming languages, compilers, architecture, security, and privacy including use of a one-of-a-kind Panorama prototype system; new hardware techniques to accelerate domain-specific computing and to unify heterogeneous systems; new software techniques to let programmers harness the performance advantages of heterogeneous systems; and new software/hardware techniques to make such heterogeneous systems more secure.

3.1.3 Impacts

The project's impacts are: to specifically enable computational biologists to better see the "genetic dark matter" of population-wide genomics which has been to date hidden, opening up new scientific discoveries; and to more generally enable future computer users to more easily take advantage of heterogeneous computer systems to solve large and complex problems. This project is also pursuing two broader impact initiatives. The first is an ambitious yet concrete initiative to increase participation of under-represented minority students in computer science by developing a low-level computer-systems module for a new four-week summer program targeting rising sophomores. The second involves specific plans to grow the open-source software/hardware ecosystem in the computational-biology and computer-systems communities.

3.1.4 Methods/Project Thrusts

The Panorama project includes a highly interdisciplinary team of researchers across four focus areas: applications (computational biology), programming languages & compilers, computer architecture, and security & privacy. The team is taking a holistic software/hardware co-design approach to explore five tightly interconnected research thrusts. The first three thrusts are structured from top-down across the computing stack. Thrust 1 investigates new computational pangenomics data structures and algorithms and will develop PanoBench, a new benchmark suite suitable for driving the remaining thrusts. Thrust 2 investigates new programming-language and compiler techniques [1, 2, 3]. Thrust 3 investigates new computer architectures with support for a whole-rack manycore with 1M+ cores and a partitioned global address space, unified array-based accelerators, and application-specific accelerator chiplets for computational pangenomics. The final two thrusts cut across both software and hardware. Thrust 4 investigates new security and privacy techniques including scalable secure computation on heterogeneous rack-scale systems, secure rack-scale resource management with auto-tuning, and differential privacy and homomorphic encryption for pangenomics. Thrust 5 involves holistically evaluating the research ideas in the other thrusts through the use of a one-of-a-kind Panorama prototype system.

3.2 Differential Privacy for Pangenomes

We implement a differentially private haplotype sampling method in a pangenomics toolkit. It projects ε-differentially private synthetic pangenome variation graphs out of pangenomes built from complete haplotype-resolved assemblies like those made in the Human Pangenome project. We generate ε-differentially private graphs from the human *m*ajor *h*istocompatibility *c*omplex (MHC), and use these to explore the effects of algorithm paramaters on output.

Large medical cohorts with associations between genomes and phenotypes usually only provide controlled data access to trusted researchers. Our goal is to establish a standard whereby a they could release a fully-public transformation of this controlled data for global use by anyone. This would thus provide global biomedical utility without significant risk to study participants. We do so by applying tools from differential privacy.

Differential privacy is an approach to data release that which allows for the description of group characteristics without revealing information about single individuals. It quantifies privacy loss caused by the release of information, allowing us to reason about the risks that a particular data sharing model poses to individual privacy. We build on decades of work on differential privacy, which has yielded well-defined models of differentially private data publication [4]. Our work is similar to approaches that have been used for trajectory data release, but differs in the unique biomedical context and properties of the graph data structures we use. This has led to the need for an application-specific implementation.

Differential privacy provides tools that allow us to generate privacy-preserving synthetic databases out of real ones. We pursue this approach because we judge that it is easier to share static databases than organize access to differentially private queries. If we provide a tool to produce a differentially private synthetic pangenome, a genomics research project could release a pangenome that can be reused indefinitely by external researchers. Reuse of this database would not constitute increased privacy risk to individuals [4].

4. Individual Development Plan

Shelby Solomon's individual development plan is based on the strengths and weaknesses listed in section 2 and will build off of the background 3.

4.1 Technical & Research Aims

Aim Differential privacy algorithm development and testing

Concept To enable allowing a human pangenome graph to be constructed and publicly released using genome data from a diverse set of genome data by ensuring strong privacy for individuals.

Idea Work with Garrison on securing individual and group data with DP algorithms and integrate into PanoBench. Apply more artificial intelligence techniques to testing the efficacy of the already implemented algorithms.

Aim Aide in development of new pangenome layout algorithms

Concept These layout algorithms can provide effective visualization that reveals the detailed structure of regions of the human pangenome, which were completely invisible to genomics researchers before.

Idea Zhang and Garrison use stochastic gradient descent, a type of AI optimization with which I would love to experiment.

Aim Support broader impact initiatives

Concept All NSF grants must contain broader impact initiatives. This is due to the nature of work being funded by the taxpayer, it follows that projects should have as a partial focus 'broader impacts' or aspects that further things that are 'good for the people'.

Idea I have participated in many 'broader impacts' initiatives and guided many underrepresented students in the computing sciences, and look forward to supporting the same for the Panorama project by aiding in the mentoring and management of the low-level computer systems module for the 4-week summer program and support/grow the OSS ecosystem in computational biology.

4.2 Mentoring Plan

4.2.1 Mentorship Team

My mentorship team consists of Pjotr Prins, Erik Garrison and Christopher Batten. Pjotr's mentoring responsibilities include general software development leadership (Aims 1 & 2), open source software contribution (broader impacts Aim 3), and genetics/genomics guidance. I intend to work very closely with Erik's team as they are building the tools for which I see the application of artificial intelligence techniques being applicable. Christopher will mentor me for Aim 3, supporting broader impacts by including diverse undergraduate student in research.

4.2.2 Meeting and Evaluation Plan

I will have bi-weekly meetings with Pjotr either in person on online. Whereas I will be working closely with Erik's team and have almost daily meetings, while also scheduling bi-weekly evaluation meetings with Erik one-on-one. Christopher holds a weekly meeting about project progress that I will attend regularly, and give scheduled talks with updates about the progress of my work.

4.2.3 Training in Genomics

As a computer science PhD genetics, genomics and pangenomics are new areas to me. As such I will participate in educational activities, aka course work to support the research. The potential courses are listed in the appendix.

4.3 Postdoc timeline

References

[1] Muhammad Umar, Weizhe Hua, Zhiru Zhang, and G. Edward Suh. Softvn: Efficient memory protection via software-provided version numbers. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ISCA '22, page 160–172, New York, NY, USA, 2022. Association for Computing Machinery.

[2] Shaojie Xiang, Yi-Hsiang Lai, Yuan Zhou, Hongzheng Chen, Niansong Zhang, Debjit Pal, and Zhiru Zhang. Heteroflow: An accelerator programming model with decoupled data placement for software-defined fpgas. In *Proceedings of the 2022 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, FPGA '22, page 78–88, New York, NY, USA, 2022. Association for Computing Machinery.

[3] Weizhe Hua, Muhammad Umar, Zhiru Zhang, and G. Edward Suh. Mgx: Near-zero overhead memory protection for data-intensive accelerators. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ISCA '22, page 726–741, New York, NY, USA, 2022. Association for Computing Machinery.

[4] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Postdoc Development Timeline		Year 1				Year 2			
		Quarters				Quarters			
		1	2	3	4	1	2	3	4
Research Plan	Differential Privacy AI test	■	■	■					
	AI Optimization for Layout					■	■	■	
	Broader Impact Work		■	■			■	■	
Career Plan	Present at UTHSC venues			■	■			■	■
	Present at Cornell venues			■	■			■	■
	Apply for NSF ExLENT*						■	■	
	Apply for NIH DS**		■	■	■	■	■	■	
	Submit/present at ICCGE***								■
	Attend/present at conferences			■	■			■	■

Figure 1. **Postdoc Timeline**
* - Experiential Learning for Emerging and Novel Technologies
** - Diversity Supplement
*** - International Conference on Computational Genomics and Evolution

Table 1. Potential Courses

TITLE	OFFERED BY	DESCRIPTION
Statistical Thinking and Data Analysis	MITOpenCourseware	Self explanatory
High-Dimensional Statistics	MITOpenCourseware	Intro to the finite sample analysis of high-dimensional statistical methods, state-of-the-art regression, matrix estimation, and PCA.
Genetics, Genomics & Informatics Seminar	UTHSC GGI	Discuss state-of-the-art research into genetics and genomics.
Foundations of AI in Healthcare I	UTHSC GGI	ID a biomedical/healthcare area of interest that may benefit from the application of AI/ML.
Foundations of AI in Healthcare II	UTHSC GGI	Explore modification and usage of ML algorithms.
Gene Structure and Function	UAH	Advanced studies of macromolecular structure and biological function of proteins and nucleic acids involved in the passage of genetic information and cellular response. Structural significance of viruses and molecular evolution included.
Biostatistics/AI	UAH with A&M	
Psychobiology Stress & Illness	UAH	Overview of physiological stress responses and their influence on health, behavior, and illness.
Bioinformatics I	UAH	Practical use in bioinformatics and X-ray crystallography
Bioinformatics II	UAH	Practical use in bioinformatics and applied genomics.
Microbial Genetics	UAH	Transmission, expression, and evolution of genes in microorganisms. Studies of chromosomes, plasmids, transposons, bacteriophages, and other genetic elements.
Advanced Molecular Techniques	UAH	Laboratory techniques in molecular biology including current methodology in genomics, proteomics, and RNA analysis.
Immunology	UAH	Innate, humoral, and cell-mediated immunity. Immune deficiencies and hypersensitivities. Autoimmunity, transplantation, and other genetic elements.

Table 2. Potential Courses

TITLE	OFFERED BY	DESCRIPTION
Algorithms in Bioinformatics	UAB	This course introduces various fundamental algorithms and computational concepts for solving questions in bioinformatics and functional genomics. These include graph algorithms, dynamic programming, combinatorial algorithms, randomized algorithms, pattern matching, classification and clustering algorithms, hidden Markov models and more. Each concept will be introduced in the context of a concrete biological or genomic application. A broad range of topics will be covered, ranging from genome annotation, genome reconstruction, microarray data analysis, phylogeny reconstruction, sequence alignments, to variant detection.
Next-generation Sequencing Data Analysis	UAB	This course is aimed to equip participants with the essential knowledge and skills required to begin analyzing next-generation sequencing data and carry out some of the most common types of analysis. The topics covered in-depth during this course are the analysis of RNA-Seq, ChIP-Seq data, ATACseq data, and Single-cell data, with an optional Variant Calling session. The sessions will also include Introduction to next-generation sequencing (NGS) technologies, common NGS data analysis issues, applications of sequencing technologies, introduction to bioinformatics file formats (e.g. FASTQ, bam, bed) and bioinformatics toolkits. At the end of this course, participants will have the expertise to perform these data analysis independently.
Visual Analytics for Bioinformatics	UAB	We will cover representation of different data types, concentrating on those generated by data-rich platforms such as next-generation sequencing applications, flow/mass cytometry, and proteomics, and will discuss the use of visualization techniques applied to assessing data quality and troubleshooting.
Bioinformatics specialization	UC San Diego via Coursera	Master bioinformatics software and computational approaches in modern biology.