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 softwares 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. Strengths & Weaknesses

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

1.1 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

1.2 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

2. Individual Development Plan

Shelby Solomon's individual development plan is based on the strengths and weaknesses listed in section 1 and will build off of the background **??**.

2.1 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.

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

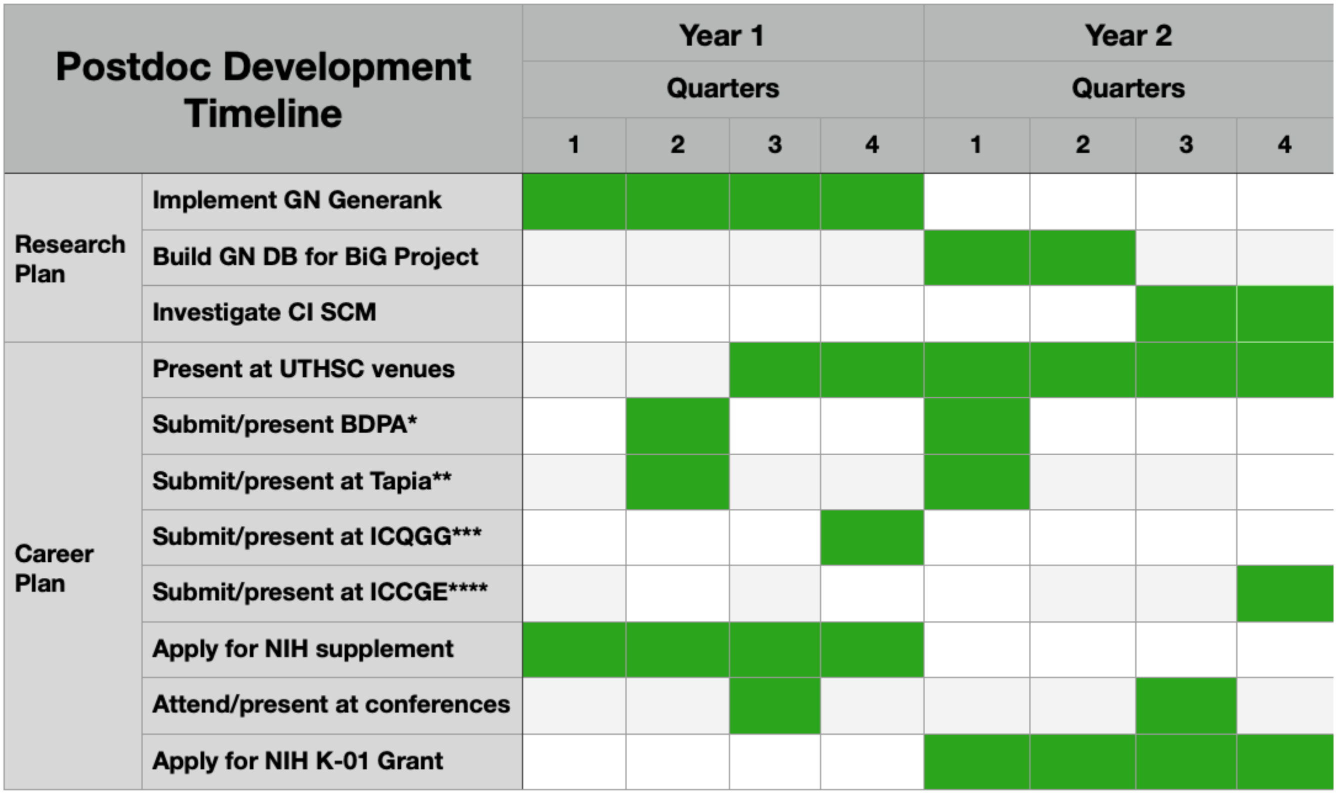
Concept All NSF grants must contain broader impact initiatives. This is due to the nature of work being funded by the taxpayer, it follows that projects should have as a partial focus 'broader impacts' or aspects that further things that are 'good for the people'.

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.

2.2 Postdoc timeline

References

A. Potential Courses



Postdoc Development Timeline		Year 1				Year 2			
		Quarters				Quarters			
		1	2	3	4	1	2	3	4
Research Plan	Implement GN Generank	🟩	🟩	🟩	🟩				
	Build GN DB for BiG Project					🟩	🟩		
	Investigate CI SCM							🟩	🟩
Career Plan	Present at UTHSC venues			🟩	🟩	🟩	🟩	🟩	🟩
	Submit/present BDPA*		🟩			🟩			
	Submit/present at Tapia**		🟩			🟩			
	Submit/present at ICQGG***				🟩				
	Submit/present at ICCGE****								🟩
	Apply for NIH supplement	🟩	🟩	🟩	🟩				
	Attend/present at conferences			🟩				🟩	
	Apply for NIH K-01 Grant					🟩	🟩	🟩	🟩

Figure 1. **Postdoc Timeline**

* - Black Data Processing Associates, bdpa.org

** - ACM/CMD-IT Tapia Celebration of Diversity in Computing

*** - International Conference on Quantitative Genetics and Genomics

**** - International Conference on Computational Genomics and Evolution

Table 1. Possible Coursework

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. Possible Coursework

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.