HPC for Computational Genomics

The NCSA Genomics Group is a host for research into the use of high performance computing (HPC) for primary genomics analyses, such as alignment, variant calling, genome assembly, and RNASeq. By its nature, this research is highly collaborative. Every member of our team is affiliated with multiple departments or campus initiatives. The student participants in this group serve as a bond between the campus faculty using computational genomics analyses in their research, and the NCSA experts in HPC, storage, networking, databases, etc. Together we enable the use of advanced computing infrastructure in computational genomics. Explore this page to find out who is involved, how we are connected, and what projects are currently ongoing.

Staff members from the software directorate and other groups within research consulting are frequently collaborators on our projects.

NCSA Press:

Crossing over, branching out: Meet the NCSA Genomics team

Engineering Open House Award

Collaborative efforts produce clinical workflows for fast, translational genetic analysis

Table of Contents:

- Active Projects
- Staff
- Graduate Students
- Undergraduate Students
- Alumni
- Other Collaborations

Active Projects

Project name	Project description	Genomics staff	Collaborators
Biomedical Pipeline Developme nt	Multiple projects involving workflow development and feature addition including determination on whether workflows can be deployed into the cloud.	Joshua Allen Mohith Manjunath Weihao Ge Raghid Alhazmy	The Mayo Clinic
CROPPS	GPU accleration and refactoring of code for gRNA design for CRISPR assays.	David Bianchi	Cornell Computer Science Crop Science
NEAT	Software Development of sequence simulator with mutational models.	Josh Allen Raghid Alhazmy	The Mayo Clinic Ontario Cancer Research Center The Broad Institute University of Wyoming
Farm to Food Bank Mobile Application Developme nt	Development of a mobile app for farmers to sell off-spec and extra produce to food banks.	Christina Fliege	Google

Christina Elizabeth Fliege

Technical Program Manager, National Center for Supercomputing Applications



NCSA Genomics, September 2017. Credit: Steve Deunsing



NCSA Genomics 'Best Original Undergraduate Research' Award



NCSA Genomics - Bluewaters tour, 17 June 2019

MINERVA	Developing an interface for combined genomic and diagnostic analysis for improved prediction, research and clinical interpretation of genomic variation.	David Bianchi Raghid Alhazmy	The Mayo Clinic
Investigator Portal	Developing a research compute portal for clinical diagnostics and genomics, where investigators can filter and query results against bioinformatic catalogs and applications, and generate re- useable datasets that can be viewed, analyzed and managed.	David Bianchi Misael Lazaro	The Mayo Clinic
GWAS Study of Dairy Cattle	Analyze genotype data of over 11 cow farms, to find the cows' susceptibility to diseases and underlying genomic variants	Joshua Allen Weihao Ge	Prof. Sandra Rodriguez Zas
Metabolomi cs data analysis on microbiome	Understand gut microbiome products for personalized medicine.	Weihao Ge Misael Lazaro David Bianchi	Prof. Issac Cann
genomic selection for maize and sorghum	Evaluate how much genomic variants will contribute to traits between maize and sorghum.	Weihao Ge	Prof. Alex Lipka

Staff

Research Interests				
	Joshua Allen Senior Research Programmer BA Mathematics and English (2001) MA English (2005) MS Bioinformatics (2019)	Sequence Simulation and Advanced Workflow Development. Modeling next-generation sequencing data, applying machine learning techniques to enhance models, writing production-ready code.		
	David Bianchi Research Scientist Ph.D Physical Chemistry (2022)	Metabolic Engineering, Synthetic Biology, Gene Regulation, Genomics Analysis, Digital Agriculture, Spatial /Multi Omics, Personalized Medicine, Research Software Engineering, High-Performance Computing, GPU Computing		
	Mohith Manjunath Research Programmer Ph.D Aersospace Engineering (2014)	Quantum Computing in Biology and Chemistry		

B.S. Physics	
(2008)	
M.S. Physics (2011)	
Ph.D. Biophysics (2018)	
Raghid Alhazmy Genomic data analysis, Machine Learning, and Computational Biology.	
Research Programmer	
B.S. Biology (2021)	
M.S. Bioinformatics (2023)	
Joao Paulo Gomes Viana	
Misael Lazaro Genomics Data Analysis, Drug Development and Design, and Computational Biology	
Academic Researcher	
B.S. Biochemistry (2019)	
M.S. Biochemistry (2023)	

Graduate Students

	Dhruvesh Shah School of Information			
	Yash Wasnik	Racial Health Disparities Yazhuo is involved in Racial Health Disparities project and researches with machine learning and data science skills. Her work is to do statistical analysis and write codes to build a pipeline on health datasets in collaboration with team members.		
Undergraduate Students				

Alumni

	Katherine Kendig	Project Management
	Associate Project Manager	Katherine is a project manager with the NCSA Industry Program, working primarily with biomedical partners.
M CS	B.A. Anthropology (2012)	She benchmarked the Sentieon variant calling software for the Mayo Grand Challenge: https://www.biorxiv.org/content/10. 1101/396325v1
	M.F.A. Creative Writing (2017)	She has also contributed to NCSA's Public Affairs team, writing articles about NCSA and XSEDE research:
	g (2011)	After the storm; Bringing supercomputing to psychology; DISSCO Tech; ECSS: Profiles in Consulting; NCSA Genomics; H istory was here
	Ramshankar Venkatakrishnan	Phillips 66 and Hardware support
	Research Programmer	Ram is developing code for the Phillips 66 project with the Data Analytics team. The idea of the code is to use Machine Learning to determine the best price to sell their petroleum products. The model considers a vast array of parameters to make the decision.
	B.S. Electronics & Communications (2012)	Ram is working with the Innovative Systems Laboratory (ISL) at NCSA to create roofline model for a U250 Xilinx card using convolution as the code to plot the model.
	M.S. Electrical & Computer Engineering (2015)	Ram also provides software and installation support for the HPC clusters at NCSA for a variety of clients.
	<u>D</u> an Lanier, Research	NCSA Industry
	B.S. Applied	Dan supports biomedical partners in the NCSA Industry program.
301	Mathematics (2008)	Dan provides a complementary mix of expertise in HPC and mathematical data analysis to enable pharmaceutical, agricultural and medical companies to utilize the high performance computing resources at NCSA.
blocked URL	Matthew Kendzior	Mayo Grand Challenge
	Research Programmer	Mr. K is working as a researcher in the Mayo Grand Challenge, which aims to drastically speed up the time for detection
	BS Crop Sciences	of genomic variants, and to extract more information from whole genome sequencing data.
	MS Bioinformatics	Genomic variant calling by assembly
	(2019)	Mr. K is focusing on a method to detect genomic variants by assembly.
		He is employing the software Cortex-var, which constructs de-novo genome assembly on multiple sequencing samples, and then compares the resultant de Bruijn graphs to detect where they diverge, indicating a potential variant. This could be a good method for detecting novel variants, especially repeats and complex rearrangements in complex genomes, such as polyploid plants and cancer. Mr. K is using his strong background in genomics to interpret, clean-up and validate the output.
		Mr. K is also working with Tiffany on the genomic analysis of HLHS for the Mayo Grand Challenge.
		Poster: Variant Calling by Assembly
		Poster: Reference-guided variant calling for non-repetitive sequences in Glycine Max
	Brian Bliss, Research Programmer	Data compression
	riogrammer	Brian will be working on data compression for the Mayo Grand Challenge project.
20 500		
	Sushma Yellapragada	NEAT
	Bachelor of Technology:	Sushma is currently working on the NEAT project, contributing code and testing.
	Computer Science Engineering,	
	University (2019)	
	M.S. Computer Science, UIUC (2022)	

	Angelo Santos	
	Yazhuo Zhang	Racial Health Disparities
	MS in Information Management	Yazhuo is involved in Racial Health Disparities project and researches with machine learning and data science skills. Her work is to do statistical analysis and write codes to build a pipeline on health datasets in collaboration with team members.
	Sijia Huo	Parallelization of R
	B.S. Mathematics & Computer Science (2018) second major in Statistics third major in Economics	Sijia is working with NCSA Faculty Fellow Dr. Zeynep Madak-Erdogan to introduce parallel R code into her research. Dr. Madak-Erdogan is exploring racial disparities in breast cancer occurrence through the lens of diet and nutrition.
	Ryan Chui	NCSA Industry
	B.S. Biochemistry (2016) M.S. Bioinformatics	Ryan performed software installation, benchmarking, and development for a variety of industry partners. To investigate how the training time for deep neural networks (DNN's) can be affected, Ryan worked with TensorFlow, Google's deep learning library, to perform multi-label classification on a data set.
	(2017)	He built an autoencoder – an unsupervised deep neural network - to extract salient features from the data.
	Department of Computer Science,	On Github:
	0100	EpiQuant: Hadoop, C, Tensorflow - epistasis software prototypes
		MLCC - multi-label cancer classification
		q2b - binary representation of nucleotides
		ptgz - parallel tar gzip
	Ionnio Zormono	Usage Analyzer - log analyzer for HPC schedulers
	B.S. Integrative Biology (2017)	Benchmarking performance and accuracy of genomic variant calling software Jennie collaborated to document our efforts in benchmarking variant calling on HPC systems. Jennie also participated in the debugging of the H3ABioNet GATK Germline Workflow.
		Bioinformatics in the Cloud
		Jennie is investigating the issues of portability, reproducibility and scaling of bioinformatics workflows in cloud
		Students Capitalize on Computational Genomics Research Using AWS
blocked URL	Angela Chen	Accurate and esclable CM/AS algorithms
	M.S. Statistics (2017)	Accurate and scatable GVVAS algorithms
	Department of Statistics, UIUC	Angela wrote a manuscript to demonstrate that her new stepwise epistatic model selection procedure has greater
	CompGen fellow	statistical power compared to other methods. However, the Java-based TASSEL5 cannot be easily parallelized across multiple nodes in a computational cluster, to run on modern, relevant datasets, which thend to be your large such as the Alzheimer's SNP panel.
	advised by Dr. Alexander Lipka	mouori, rolevant udiasets, which terre to be very large, such as the Alzheinner's SINF Paher.
	Khory Wagner	Khory provided the expertise in computer science to convert this Java code into C++ and parallelize
	advised by Dr. Vologymyr Kindratenko	it in HPC environment.

blocked URL blocked URL	Nainika Roy B.S. Molecular and Cellular Biology (2017) minor in Informatics and Chemistry SPIN fellow Junyu Li B.S. Molecular and	Data formats and data structures in computational genomics Genomic variant calling by assembly
	Cellular Biology (2017) minor in Computer Science SPIN fellow	Junyu worked with Mr. K in an interdisciplinary team, providing the expertise in math and computer science to automate the Cortex-var workflow and interpret the algorithm. Poster: Reference-guided variant calling for novel non-repetitive sequences in <i>Glycine max</i>
blocked URL	Noah Flynn B.S. Bioengineering, Mathematics (2017) minor in computer science SPIN fellow	Evolution of molecular networks and persistence of organisms
	Jacob Heldenbrand Research Programmer B.S. Biochemistry (2014) M.S. Bioinformatics (2016)	NCSA Industry Jacob supports biomedical partners in the NCSA Industry program. Jacob provides a complementary mix of expertise in HPC and bioinformatics data analysis to enable pharmaceutical, agricultural and medical companies to utilize the high performance computing resources at NCSA. Jacob and Azza Ahmed (Ph. D. candidate, University of Khartoum) are exploring and evaluating the use of Swift T for variant calling. Github: Swift T Variant Calling Guide: Downloading large datasets with SRA Toolkit
	Matthew Weber B.S. Molecular and Cellular Biology (2016) M.S. Bioinformatics (2018) Department of Crop Sciences, UIUC CompGen fellow advised by Dr. Matthew Hudson	Mutation profiles of cancer Mr. Weber is developing machine learning methods to effectively stratify cancers based on the statistical properties of mutations found in afflicted individuals. Cancer stratification is predictive of disease outcomes, drug response and drug metabolism. Effective computational approaches based on total data acquired to-date can make this process cheaper in the clinic. Matt collaborates with the Ontario Institute for Cancer Research to make sure his models are realistic. Paper: Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models Poster: Statistical models to capture mutational properties for NextGen Sequencing Data
	Aishwarya Raj B.S. Biochemistry (2019) minor in Bioinformatics Illinois Informatics Institute fellow	Evolution of molecular networks and persistence of organisms. Construct and compare gene, metabolic and signaling networks from organisms across the tree of life. The goal of the project is to provide support for the general framework of persistence strategies. It postulates that persistence is achieved by biological systems via a tradeoff of traits that serve either economy, flexibility, or robustness. In this project we want to determine and quantify the molecular mechanisms that underlie these persistence strategies. Will analysis of the biomolecular networks allow us to differentiate between organisms of differing economy, flexibility, and robustness, and subsequently classify unknown, newly discovered, or modified organisms within such predefined classes? Poster: Persistence Strategies in Biomolecular Network Architecture NCUR Slides: Architecture and Dynamics of Biomolecular Networks Facilitate Evolution of Persistence Strategies in Living Organisms

-		Cynthia Liu	Workflow management comparisons
-		B.S. Bioengineering (2019) minor in Computer	Cynthia worked to learn the Nextflow system for workflow management and to compare and contrast three competing workflow management options for bioinformatics in association with the work Ram is performing for the Mayo Grand Challenge.
	200		Poster: Comparative Analysis of Genomic Sequencing Workflow Management Systems
		Brian Rao B.S Integrative Biology	Brian wrote and tested the variant calling workflow code for the Mayo Grand Challenge. He focused on the accuracy and performance considerations of tumor variant detection in clinical settings.
		(2018)	
		Minor in Informatics	
all a	18 A.	Angelynn Huang	Angelynn contributed to benchmarking the performance and accuracy of Minimap2 (Li, 2018) - a program used for analyzing sequencing read data in genomics.
			Minimap2 maps the sequencing reads against the reference genome for the species. Currently, BWA MEM (Li, 2013) is the most widely used tool for this purpose, with Novoalign (Hercus and Albertyn, 2012) coming as a close second. However, recent research (Li, 2018) suggests that Minimap2 is equally accurate yet also faster than BWA MEM. Are these claims true? Can we validate them independently using our own measurements? Sophia and Angelynn ran tests in AWS to answer these questions.
			Poster : Minimap2_BWA MEM
			Spotlight: http://www.ncsa.illinois.edu/news/story/ncsa_student_spotlight_angelynn_huang_and_sophia_torrellas
blocked L	IRL	Sparsh Agarwal	Mayo Grand Challenge Project
		B.Tech + M.Tech in Biochemical Engineering and Biotechnology (2018)	He is working on Mayo Grand Challenge project that aims to detect genomic variants in humans responsible for HLHS disease by using Cortex-var software as the de novo assembler and variant caller.
		MS in Bioinformatics (2020)	
		Prakruthi Burra	Human Heredity & Health in Africa
		B. E. Computer Science (2018)	Prakruthi contributes to UIUC's work with the H3Africa Consortium. She is involved with projects on graph representations of genome assemblies and machine learning techniques applied to biological problems.
		M.S. Biological Sciences (2018)	
			Workflow management for variant calling
			Prakruthi is also implementing a variant calling workflow in Nextflow, an increasingly popular workflow manager. Prior to her workflow development work, she was briefly involved in testing the workflow developed for the Mayo Grand Challenge.
	- All Anno	Dave Istanto	Nextflow Cortex_Var Structural Variant Calling Workflow
Time t	68	B.S. Crop Sciences (2018)	Dave is responsible to develop a user-friendly and cluter-portable version of cortex_var workflow to detect large structural variants in given genomes using Nextflow workflow management language
			Soybean Haplotype and Structural Variant Profiling and Analysis
			Dave is responsible for both profiling of variants in 481 soybean lines, which later will be processed by correlating them to certain visible characteristics

blocked LIBI	Shubham Rawlani	
		Space Search Reduction and EpiQuant
	Electronics and Communication Engineering	Shubham is involved in data analysis part where he writes code for data wrangling, extraction and cleaning to ease out the evaluation of statistical algorithms in the analysis of GWAS data for genomic variant epistasis
	Masters in Information Management	Shubham is also involved in benchmarking the EpiQuant project and will collaborate to improve the scalability by testing on different datasets and nodes to achieve efficient results
blocked URL	Priya Balgi	Project Management
	Bachelors in Information Technology Engineering	Priya is responsible for assisting in execution of Project Management tasks. Additionally, she performs genomics workflow testing using bash scripting in HPC environment and is developing a website using GitHub Pages/Jekyll for creation & auto-maintenance of project documentation.
	Masters in Information Management	She also lead a student group of 8 for representing NCSA industry research during the Engineering Open House where the Genomics group won the Second Best Original Under Graduate Research Award and will also represent NCSA Industry research at the BioIT World Conference.
		Poster: NCSA Industry Research
blocked URL	Mingyu Yang	Mayo Grand Challenge Project
	B.E. Network Engineering	Mingyu is working on optimize and test the performance of GABAC, which is a gene compression application.
	M.S. Electrical and Computer Engineering	
	Yazhuo Zhang	Racial Health Disparities
	MS in Information Management	Yazhuo is involved in Racial Health Disparities project and researches with machine learning and data science skills. Her work is to do statistical analysis and write codes to build a pipeline on health datasets in collaboration with team members.
	Dipro Ray	Resolving Racial Disparities by Applying Statistics on
	B.S. Computer Science (2020)	Complex, Multidimensional Datasets
	Minor in Mathematics	Dipro is working on turning a proof-of-concept prototype, of a statistical pipeline to analyze health data, into a well- structured open source package that is very portable, containerized and deployable through the cloud (like AWS), making such critical software available to researchers and collaborators with only a few commands.
T		In pursuit of this goal, Dipro also works on refining the statistical pipeline in a modular manner and chalking out key design decisions for its implementation, and improving the package's computational efficiency (by making use of the host computer's architecture and resources)."
blocked URL	Tajesvi Bhat	Deployment of Variant Calling Workflows on Cloud Platform
	B.S. Computer Science (2020) Minor in Bioengineering	Tajesvi is working on this that project aims to deploy variant calling workflows implemented using systems such as WDL and Nextflow in AWS and other cloud services.





Other Collaborations

blocked URL	Dr. Matthew Hudson Bioinformatics Crop Science Dan Wickland Ph.D. Informatics (2019)	HPCBio, Carver Biotechnology Center
blocked URL	Dr. Daniel Katz Computer Science	NCSA Scientific Software and Applications Portable variant calling workflow in Swift
	Computer Science University of Khartoum advised by Dr. Faisal Fadlelmola	Github: Swift Variant Calling
	Dr. Zeynep Madak-Erdogan Food Science & Human Nutrition	Madak-Erdogan Lab Systems Biology of Estrogen Signaling • NCSA Faculty Fellow 2017-2018 • Understanding Breast Cancer Disparities in African-American Women

	Brandi Smith Ph.D. Food Science and Human Nutrition (2021)	 H3Africa Consortium bioinformatics workflows in the cloud custom genotyping chip for African populations H3Africa bioinformatics node accreditation
	Morgan Taschuk Bioinformatics blocked URL	 OICR production infrastructure for primary genomics analyses reproducibility of research in cancer genomics
blocked URL	Paul Hatton HPC / Visualisation blocked URL	University of Birmingham
blocked URL	Nahil Sobh	UIUC Beckman Institute
DIOCKEO URL	Umberto Ravaioli	DIDC ECE, BECKMAN Institute
	Cyberinfrastructure, ECE	Biosketch
	Lynn Hassan Jones	UIUC
	Radiology	Resume