# 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 Development | Multiple projects involving workflow development and feature addition including determination on whether workflows can be deployed into the cloud. | Joshua Allen<br><br>Mohith Manjunath<br><br>Weihao Ge<br><br>Raghid Alhazmy | The Mayo Clinic |
| CROPPS | GPU accleration and refactoring of code for gRNA design for CRISPR assays. | David Bianchi | Cornell Computer Science<br><br>Crop Science |
| NEAT | Software Development of sequence simulator with mutational models. | Josh Allen<br><br>Raghid Alhazmy | The Mayo Clinic<br><br>Ontario Cancer Research Center<br><br>The Broad Institute<br><br>University of Wyoming |
| Farm to Food Bank Mobile Application Development | 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

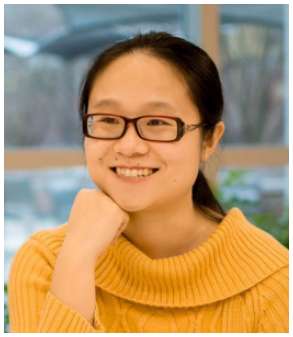
NCSA Genomics 'Best Original Undergraduate Research' Award

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<br><br>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<br><br>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<br><br>Weihao Ge | Prof. Sandra Rodriguez Zas |
| Metabolomics data analysis on microbiome | Understand gut microbiome products for personalized medicine. | Weihao Ge<br><br>Misael Lazaro<br><br>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**<br><br>Senior Research Programmer<br><br>BA Mathematics and English (2001)<br><br>MA English (2005)<br><br>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**<br><br>Research Scientist<br><br>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**<br><br>Research Programmer<br><br>Ph.D Aersospace Engineering (2014) | Quantum Computing in Biology and Chemistry |

| | | |
|---|---|---|
|  | **Weihao Ge**<br><br>B.S. Physics (2008)<br><br>M.S. Physics (2011)<br><br>Ph.D. Biophysics (2018) | Machine Learning in Epidemiology and Genomics. Biostatistics and Informatics. |
| | **Raghid Alhazmy**<br><br>Research Programmer<br><br>B.S. Biology (2021)<br><br>M.S. Bioinformatics (2023) | Genomic data analysis, Machine Learning, and Computational Biology. |
| | **Joao Paulo Gomes Viana** | |
| | **Misael Lazaro**<br><br>Academic Researcher<br><br>B.S. Biochemistry (2019)<br><br>M.S. Biochemistry (2023) | Genomics Data Analysis, Drug Development and Design, and Computational Biology |

## Graduate Students

| | | |
|---|---|---|
| | **Dhruvesh Shah**<br><br>School of Information | |
| | Yash Wasnik<br><br>Information | ### Racial Health Disparities<br><br>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

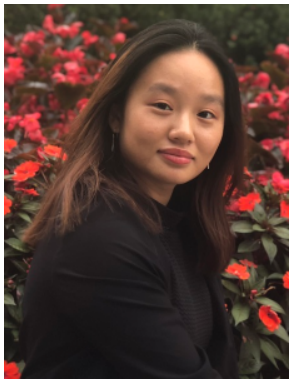
| | | |
|---|---|---|
| | | |

# Alumni

| | | |
|---|---|---|
| | **Katherine Kendig**<br><br>Associate Project Manager<br><br>B.A. Anthropology (2012)<br><br>M.F.A. Creative Writing (2017) | ## Project Management<br><br>Katherine is a project manager with the NCSA Industry Program, working primarily with biomedical partners.<br><br>She benchmarked the Sentieon variant calling software for the Mayo Grand Challenge: https://www.biorxiv.org/content/10.1101/396325v1<br><br>She has also contributed to NCSA's Public Affairs team, writing articles about NCSA and XSEDE research:<br><br>After the storm; Bringing supercomputing to psychology; DISSCO Tech; ECSS: Profiles in Consulting; NCSA Genomics; History was here |
| | **Ramshankar Venkatakrishnan**<br><br>Research Programmer<br><br>B.S. Electronics & Communications (2012)<br><br>M.S. Electrical & Computer Engineering (2015) | ## Phillips 66 and Hardware support<br><br>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.<br><br>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.<br><br>Ram also provides software and installation support for the HPC clusters at NCSA for a variety of clients. |
| | **Dan Lanier**, Research Programmer<br><br>B.S. Applied Mathematics (2008) | ## NCSA Industry<br><br>Dan supports biomedical partners in the NCSA Industry program.<br><br>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**<br><br>Research Programmer<br><br>BS Crop Sciences (2016)<br><br>MS Bioinformatics (2019) | ## Mayo Grand Challenge<br><br>Mr. K is working as a researcher in the Mayo Grand Challenge, which aims to drastically speed up the time for detection of genomic variants, and to extract more information from whole genome sequencing data.<br><br>## Genomic variant calling by assembly<br><br>Mr. K is focusing on a method to detect genomic variants by assembly.<br><br>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.<br><br>Mr. K is also working with Tiffany on the genomic analysis of HLHS for the Mayo Grand Challenge.<br><br>Poster: Variant Calling by Assembly<br><br>Poster: Reference-guided variant calling for non-repetitive sequences in Glycine Max |
| | **Brian Bliss**, Research Programmer | ## Data compression<br><br>Brian will be working on data compression for the Mayo Grand Challenge project. |
| | **Sushma Yellapragada**<br><br>Bachelor of Technology: Computer Science Engineering, Northcap University (2019)<br><br>M.S. Computer Science, UIUC (2022) | ## NEAT<br><br>Sushma is currently working on the NEAT project, contributing code and testing. |

| | | |
|---|---|---|
| | **Angelo Santos** | |
| | **Yazhuo Zhang**<br><br>MS in Information Management | ## Racial Health Disparities<br><br>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**<br><br>B.S. Mathematics & Computer Science (2018)<br><br>second major in Statistics<br><br>third major in Economics | ## Parallelization of R<br><br>Sijia is working with NCSA Faculty Fellow Dr. Zeynep Madak-Erdogan to introduce parallel R code into her research.<br><br>Dr. Madak-Erdogan is exploring racial disparities in breast cancer occurrence through the lens of diet and nutrition. |
| | **Ryan Chui**<br><br>B.S. Biochemistry (2016)<br><br>M.S. Bioinformatics (2017)<br><br>Department of Computer Science, UIUC | ## NCSA Industry<br><br>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.<br><br>He built an autoencoder – an unsupervised deep neural network - to extract salient features from the data.<br><br>On Github:<br><br>EpiQuant: Hadoop, C, Tensorflow - epistasis software prototypes<br><br>MLCC - multi-label cancer classification<br><br>q2b - binary representation of nucleotides<br><br>ptgz - parallel tar gzip<br><br>Usage Analyzer - log analyzer for HPC schedulers |
| | **Jennie Zermeno**<br><br>B.S. Integrative Biology (2017) | ## Benchmarking performance and accuracy of genomic variant calling software<br><br>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.<br><br>## Bioinformatics in the Cloud<br><br>Jennie is investigating the issues of portability, reproducibility and scaling of bioinformatics workflows in cloud infrastructure by instantiating containerized versions of workflows.<br><br>Students Capitalize on Computational Genomics Research Using AWS |
| blocked URL | **Angela Chen**<br><br>M.S. Statistics (2017)<br><br>Department of Statistics, UIUC<br><br>CompGen fellow<br><br>advised by Dr. Alexander Lipka | ## Accurate and scalable GWAS algorithms<br><br>Angela and Khory collaborated to improve the scalability and parallelization of the statistical software TASSEL5, widely used for conducting genome wide association studies (GWAS) in plants.<br><br>Angela wrote a manuscript to demonstrate that her new stepwise epistatic model selection procedure has greater 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 tend to be very large, such as the Alzheimer's SNP panel. |
| | Khory Wagner<br><br>advised by Dr. Vologymyr Kindratenko | Khory provided the expertise in computer science to convert this Java code into C++ and parallelize<br><br>it in HPC environment. |

| | | |
|---|---|---|
| blocked URL | **Nainika Roy**<br><br>B.S. Molecular and Cellular Biology (2017)<br><br>minor in Informatics and Chemistry<br><br>SPIN fellow | Data formats and data structures in computational genomics |
| blocked URL | **Junyu Li**<br><br>B.S. Molecular and Cellular Biology (2017)<br><br>minor in Computer Science<br><br>SPIN fellow | Genomic variant calling by assembly<br><br>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.<br><br>Poster: Reference-guided variant calling for novel non-repetitive sequences in *Glycine max* |
| blocked URL | **Noah Flynn**<br><br>B.S. Bioengineering, Mathematics (2017)<br><br>minor in computer science<br><br>SPIN fellow | Evolution of molecular networks and persistence of organisms |
| | **Jacob Heldenbrand**<br><br>Research Programmer<br><br>B.S. Biochemistry (2014)<br><br>M.S. Bioinformatics (2016) | NCSA Industry<br><br>Jacob supports biomedical partners in the NCSA Industry program.<br><br>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.<br><br>Jacob and Azza Ahmed (Ph. D. candidate, University of Khartoum) are exploring and evaluating the use of Swift T for variant calling.<br><br>Github: Swift T Variant Calling<br><br>Guide: Downloading large datasets with SRA Toolkit |
| | **Matthew Weber**<br><br>B.S. Molecular and Cellular Biology (2016)<br><br>M.S. Bioinformatics (2018)<br><br>Department of Crop Sciences, UIUC<br><br>CompGen fellow<br><br>advised by Dr. Matthew Hudson | Mutation profiles of cancer<br><br>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.<br><br>Paper: Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models<br><br>Poster: Statistical models to capture mutational properties for NextGen Sequencing Data |
| | **Aishwarya Raj**<br><br>B.S. Biochemistry (2019)<br><br>minor in Bioinformatics<br><br>Illinois Informatics Institute fellow | Evolution of molecular networks and persistence of organisms<br><br>Construct and compare gene, metabolic and signaling networks from organisms across the tree of life.<br><br>The goal of the project is to provide support for the general framework of persistence strategies.<br><br>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<br>classes?<br><br>Poster: Persistence Strategies in Biomolecular Network Architecture<br><br>NCUR Slides: Architecture and Dynamics of Biomolecular Networks Facilitate Evolution of Persistence Strategies in Living Organisms |

| | | |
|---|---|---|
| | **Cynthia Liu**<br><br>B.S. Bioengineering (2019)<br><br>minor in Computer Science | ## Workflow management comparisons<br><br>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.<br><br>Poster: Comparative Analysis of Genomic Sequencing Workflow Management Systems |
| | **Brian Rao**<br><br>B.S Integrative Biology<br><br>(2018)<br><br>Minor in Informatics | 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. |
|  | **Angelynn Huang** | Angelynn contributed to benchmarking the performance and accuracy of Minimap2 (Li, 2018) - a program used for analyzing sequencing read data in genomics.<br><br>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.<br><br>Poster : Minimap2_BWA MEM<br><br>Spotlight: http://www.ncsa.illinois.edu/news/story/ncsa_student_spotlight_angelynn_huang_and_sophia_torrellas |
| blocked URL | **Sparsh Agarwal**<br><br>B.Tech + M.Tech in Biochemical Engineering and Biotechnology (2018)<br><br>MS in Bioinformatics (2020) | Mayo Grand Challenge Project<br><br>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. |
| | **Prakruthi Burra**<br><br>B. E. Computer Science (2018)<br><br>M.S. Biological Sciences (2018) | ## Human Heredity & Health in Africa<br><br>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.<br><br>## Workflow management for variant calling<br><br>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. |
| | **Dave Istanto**<br><br>B.S. Crop Sciences (2018) | ## Nextflow Cortex_Var Structural Variant Calling Workflow<br><br>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<br><br>## Soybean Haplotype and Structural Variant Profiling and Analysis<br><br>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 URL | **Shubham Rawlani**<br><br>Bachelors in Electronics and Communication Engineering<br><br>Masters in Information Management | ## Space Search Reduction and EpiQuant<br><br>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<br><br>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**<br><br>Bachelors in Information Technology Engineering<br><br>Masters in Information Management | ## Project Management<br><br>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.<br><br>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.<br><br>Poster: NCSA Industry Research |
| blocked URL | **Mingyu Yang**<br><br>B.E. Network Engineering<br><br>M.S. Electrical and Computer Engineering | ## Mayo Grand Challenge Project<br><br>Mingyu is working on optimize and test the performance of GABAC, which is a gene compression application. |
| | **Yazhuo Zhang**<br><br>MS in Information Management | ## Racial Health Disparities<br><br>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**<br><br>B.S. Computer Science (2020)<br><br>Minor in Mathematics | ## Resolving Racial Disparities by Applying Statistics on Complex, Multidimensional Datasets<br><br>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.<br><br>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**<br><br>B.S. Computer Science (2020)<br>Minor in Bioengineering | ## Deployment of Variant Calling Workflows on Cloud Platform<br><br>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. |

| | | |
|---|---|---|
| | **Tiffany Li**<br><br>B.S. Integrative Biology (2018)<br><br>minor in Computer Science | ## Benchmarking performance and accuracy of genomic variant calling software<br><br>Tiffany collaborates to document our efforts in benchmarking variant calling on HPC systems. We have run variant calling experiments on 500 genomes in parallel, on Blue Waters, to identify performance bottlenecks when using the GATK best practices workflow.<br><br>We have also tested a number of alternative software, such as Isaac, Genalice, and Sentieon, as well as Dragen - a hardware solution. Tiffany is documenting the pros and cons of each of these excellent approaches in a separate manuscript.<br><br>## Validation and benchmarking on ParFu - a parallel file packaging utility<br><br>Tiffany is also involved in testing and benchmarking of ParFu, an MPI tool for creating or extracting directory tree archives written by Dr. Craig Steffen, who works in the Blue Waters team.<br><br>Github: Parfu Archive Tool |

# Other Collaborations

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

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