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.

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:

- **Current People and Projects**
  - Mayo Grand Challenge: evaluating and streamlining genomics workflows/EpiQuant
  - Project Management
  - Data compression
  - NCSA Industry
  - Search Space Reduction
  - Mayo Grand Challenge
  - Genomic variant calling by assembly

- **Graduate Students**
  - Human Heredity & Health in Africa
  - Workflow management for variant calling
  - Nextflow Cortex Var Structural Variant Calling Workflow
  - Soybean Haplotype and Structural Variant Profiling and Analysis
  - Space Search Reduction and EpiQuant
  - Project Management
  - Mayo Grand Challenge Project
  - Racial Health Disparities

- **Undergraduate Students**
  - Resolving Racial Disparities by Applying Statistics on Complex, Multidimensional Datasets
  - Deployment of Variant Calling Workflows on Cloud Platform

- **High School Students**
- **Former Group Members**
  - Big data network transfers for genomics
  - Validation of Sentieon - the fast alternative to GATK
  - Convert Java-based GWAS code for Spark
  - Benchmarking performance and accuracy of genomic variant calling software
  - Validation and benchmarking on ParFu - a parallel file packaging utility
  - Parallelization of R
  - NCSA Industry
  - Benchmarking performance and accuracy of genomic variant calling software
  - Bioinformatics in the Cloud
  - Accurate and scalable GWAS algorithms
  - Data formats and data structures in computational genomics
  - Genomic variant calling by assembly
  - Evolution of molecular networks and persistence of organisms
  - NCSA Industry
  - Mutation profiles of cancer
  - Evolution of molecular networks and persistence of organisms
  - Workflow management comparisons

- **Other Collaborations**
  - HPCBio, Carver Biotechnology Center

Liudmila Sergeevna Mainzer
Technical Program Manager,
National Center for Supercomputing Applications
Research Assistant Professor, Institute of Genomic Biology
217-300-0568

NCSA Genomics, September 2017. Credit: Steve Deunsing

NCSA Genomics ‘Best Original Undergraduate Research’ Award
Current People and Projects

**Ramshankar Venkatakrishnan**
Research Programmer
- B.S. Electronics & Communications (2012)
- M.S. Electrical & Computer Engineering (2015)

**Mayo Grand Challenge: evaluating and streamlining genomics workflows/EpiQuant**

Ramshankar is working on computational improvements for the Mayo Grand Challenge, a genomics research project in partnership with the Mayo Clinic. Ram is rewriting Mayo's variant calling pipeline using the Cromwell/WDL workflow management.

Ram will also contribute his hardware expertise to the project, evaluating system architecture options to complement the team's software and coding improvements.

Ram 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.

Github: MayomicsVC Pipeline

**Katherine Kendig**
Associate Project Manager
- B.A. Anthropology (2012)
- M.F.A. Creative Writing (2017)

**Project Management**

Katherine is a project manager with the NCSA Industry Program, working primarily with biomedical partners.

She benchmarked the Sentieon variant calling software for the Mayo Grand Challenge: [https://www.biorxiv.org/content/10.1101/396325v1](https://www.biorxiv.org/content/10.1101/396325v1)

She has also contributed to NCSA’s Public Affairs team, writing articles about NCSA and XSEDE research:

- After the storm; Bringing supercomputing to psychology; DISSCO Tech; ECSS: Profiles in Consulting; NCSA Genomics; History was here

**Brian Bliss**
Research Programmer

**Data compression**

Brian will be working on data compression for the Mayo Grand Challenge project.
<table>
<thead>
<tr>
<th><strong>Dan Lanier, Research Programmer</strong></th>
<th><strong>NCSA Industry</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S. Applied Mathematics (2008)</td>
<td>Dan supports biomedical partners in the NCSA Industry <a href="#">program</a>.</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Weihao Ge</strong></th>
<th><strong>Search Space Reduction</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S. Physics (2008)</td>
<td>Weihao is evaluating statistical methods for search space reduction in the analysis of GWAS data for genomic variant epistasis in association with disease to allow for faster, more meaningful results.</td>
</tr>
<tr>
<td>M.S. Physics (2011)</td>
<td>Her work is part of the CCBGM project &quot;Scaling the Computation of Epistatic Interactions in GWAS Data.&quot;</td>
</tr>
<tr>
<td>Ph.D. Biophysics (2018)</td>
<td>An assessment of true and false positive detection rates of stepwise epistatic model selection as a function of sample size and number of markers</td>
</tr>
<tr>
<td>advised by Dr. Eric Jacobsson</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Matthew Kendzior, Research Programmer</strong></th>
<th><strong>Mayo Grand Challenge</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>BS Crop Sciences (2016)</td>
<td>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.</td>
</tr>
<tr>
<td>MS Bioinformatics (2019)</td>
<td>Genomic variant calling by assembly</td>
</tr>
<tr>
<td></td>
<td>Mr. K is focusing on a method to detect genomic variants by assembly.</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Mr. K is also working with Tiffany on the genomic analysis of HLHS for the Mayo Grand Challenge.</td>
</tr>
<tr>
<td></td>
<td><strong>Poster: Variant Calling by Assembly</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Poster: Reference-guided variant calling for non-repetitive sequences in Glycine Max</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Joshua Allen</strong></th>
<th><strong>Graduate Students</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>BA Mathematics and English (2001)</td>
<td>Josh is involved with writing and testing code for the Mayomics project and has recently begun work in statistical analysis on the MGC 2 project.</td>
</tr>
<tr>
<td>MA English (2005)</td>
<td></td>
</tr>
<tr>
<td>MS Bioinformatics (2019)</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Education</td>
</tr>
<tr>
<td>--------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
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 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.  
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. |
| Dave Istanto       | B.S. Crop Sciences (2018)                                                | Nextflow Cortex_Var Structural Variant Calling Workflow  
Dave is responsible to develop a user-friendly and cluster-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 |
| Shubham Rawlani    | Bachelors in Electronics and Communication Engineering  Masters in Information Management | Space Search Reduction and EpiQuant  
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  
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 |
| Priya Balgi        | Bachelors in Information Technology Engineering  Masters in Information Management | Project Management  
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.  
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 |
<table>
<thead>
<tr>
<th><strong>Mingyu Yang</strong></th>
<th><strong>Mayo Grand Challenge Project</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>B.E. Network Engineering</td>
<td>Mingyu is working on optimize and test the performance of GABAC, which is a gene compression application.</td>
</tr>
<tr>
<td>M.S. Electrical and Computer Engineering</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Yazhuo Zhang</strong></th>
<th><strong>Racial Health Disparities</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>MS in Information Management</td>
<td>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.</td>
</tr>
</tbody>
</table>

**Undergraduate Students**

<table>
<thead>
<tr>
<th><strong>Dipro Ray</strong></th>
<th><strong>Resolving Racial Disparities by Applying Statistics on Complex, Multidimensional Datasets</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S. Computer Science (2020)</td>
<td>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.</td>
</tr>
<tr>
<td>Minor in Mathematics</td>
<td>In pursuit of this goal, Dipro also works on refining the statistical pipeline in a modular manner and chalk ing 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).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Tajesvi Bhat</strong></th>
<th><strong>Deployment of Variant Calling Workflows on Cloud Platform</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S. Computer Science (2020)</td>
<td>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.</td>
</tr>
<tr>
<td>Minor in Bioengineering</td>
<td></td>
</tr>
</tbody>
</table>

**High School Students**

We have several high school students working with our team to gain skills and complete projects in a real-world environment.
Sophia Torrellas

Sophia and Angelynn are 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 are running 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

Angelynn Huang

Former Group Members

Ellen Nie
B.S. Computer Science (2018)

Big data network transfers for genomics

Ellen is benchmarking the network transfers of genomic data across multiple sites. She wants to understand the limitations of modern network backbone for big data genomics, and to facilitate correct configuration of the endpoints to resolve those limitations. Ellen is looking at the sites of our collaborators in Toronto, South Africa, Sudan, and the UK.

Poster: Benchmarking and Optimization of Long Distance Big Data Transfers

Validation of Sentieon - the fast alternative to GATK

Ellen is also collaborating with OICR to validate the speed and accuracy of the new software package for genomic variant calling, called Sentieon DNASeq.

Convert Java-based GWAS code for Spark

In a project described below (Accurate and scalable GWAS algorithms) we are improving performance of a stepwise epistatic model selection for Genome-Wide Association Studies. The method itself works well, but the current Java implementation is way too slow for modern data sizes.

We would like to deploy this Java code on Spark, to see if the necessary performance gains could be obtained.

A successful student applicant will use Java Spark API to adapt the current code for a Spark platform that is being deployed at NCSA ISL2.0. This code will be validated for correctness in collaboration with a student statistician from the lab of Dr. Lipka, who developed this statistical method.

Poster: Scaling the Computation of Epistatic Interactions in GWAS Data
<table>
<thead>
<tr>
<th>Tiffany Li</th>
<th><strong>Benchmarking performance and accuracy of genomic variant calling software</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S. Integrative Biology (2018)</td>
<td>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. 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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Validation and benchmarking on ParFu - a parallel file packaging utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sijia Huo</th>
<th><strong>Parallelization of R</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S. Mathematics &amp; Computer Science (2018)</td>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ryan Chui</th>
<th><strong>NCSA Industry</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S. Biochemistry (2016)</td>
<td>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. He built an autoencoder – an unsupervised deep neural network - to extract salient features from the data.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>On Github:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>EpiQuant: Hadoop, C, Tensorflow - epistasis software prototypes</td>
<td></td>
</tr>
<tr>
<td>MLCC - multi-label cancer classification</td>
<td></td>
</tr>
<tr>
<td>q2b - binary representation of nucleotides</td>
<td></td>
</tr>
<tr>
<td>ptgz - parallel tar gzip</td>
<td></td>
</tr>
<tr>
<td>Usage Analyzer - log analyzer for HPC schedulers</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Jennie Zermeno</th>
<th><strong>Bioinformatics in the Cloud</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S. Integrative Biology (2017)</td>
<td>Jennie is investigating the issues of portability, reproducibility and scaling of bioinformatics workflows in cloud infrastructure by instantiating containerized versions of workflows.</td>
</tr>
</tbody>
</table>

| 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. |

<p>| Students Capitalize on Computational Genomics Research Using AWS |</p>
<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>Department and Major</th>
<th>Advisor</th>
<th>Project Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela Chen</td>
<td>Accurate and scalable GWAS algorithms</td>
<td>Department of Statistics, UIUC, CompGen fellow</td>
<td>Dr. Alexander Lipka</td>
<td>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. 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 provided the expertise in computer science to convert this Java code into C++ and parallelize it in HPC environment.</td>
</tr>
<tr>
<td>Khory Wagner</td>
<td>Data formats and data structures in computational genomics</td>
<td>Department of Statistics, UIUC, CompGen fellow</td>
<td>Dr. Volodymyr Kindratenko</td>
<td></td>
</tr>
<tr>
<td>Nainika Roy</td>
<td>Genomic variant calling by assembly</td>
<td>Department of Statistics, UIUC, CompGen fellow</td>
<td>Mr. K</td>
<td>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 <em>Glycine max</em></td>
</tr>
<tr>
<td>Junyu Li</td>
<td>Evolution of molecular networks and persistence of organisms</td>
<td>Department of Statistics, UIUC, CompGen fellow</td>
<td>Mr. K</td>
<td></td>
</tr>
<tr>
<td>Noah Flynn</td>
<td>Evolution of molecular networks and persistence of organisms</td>
<td>Department of Statistics, UIUC, CompGen fellow</td>
<td>Mr. K</td>
<td></td>
</tr>
<tr>
<td><strong>Jacob Heldenbrand</strong></td>
<td><strong>NCSA Industry</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research Programmer</td>
<td>Jacob supports biomedical partners in the NCSA Industry program.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.S. Biochemistry</td>
<td>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.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2014)</td>
<td>Jacob and Azza Ahmed (Ph. D. candidate, University of Khartoum) are exploring and evaluating the use of Swift T for variant calling.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M.S. Bioinformatics</td>
<td>Github: Swift T Variant Calling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2016)</td>
<td>Guide: Downloading large datasets with SRA Toolkit</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Matthew Weber</strong></th>
<th><strong>Mutation profiles of cancer</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S. Molecular and</td>
<td>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.</td>
</tr>
<tr>
<td>Cellular Biology</td>
<td>Paper: Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models</td>
</tr>
<tr>
<td>(2016)</td>
<td>Poster: Statistical models to capture mutational properties for NextGen Sequencing Data</td>
</tr>
<tr>
<td>M.S. Bioinformatics</td>
<td><a href="https://cropsci.illinois.edu">Department of Crop Sciences, UIUC</a></td>
</tr>
<tr>
<td>(2018)</td>
<td>CompGen fellow</td>
</tr>
<tr>
<td></td>
<td>advised by Dr. Matthew Hudson</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Aishwarya Raj</strong></th>
<th><strong>Evolution of molecular networks and persistence of organisms</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>B.S. Biochemistry</td>
<td>Construct and compare gene, metabolic and signaling networks from organisms across the tree of life.</td>
</tr>
<tr>
<td>(2019)</td>
<td>The goal of the project is to provide support for the general framework of persistence strategies.</td>
</tr>
<tr>
<td>minor in Bioinformatics</td>
<td>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?</td>
</tr>
<tr>
<td><a href="https://informatics.uiuc.edu">Illinois Informatics Institute</a> fellow</td>
<td>Poster: Persistence Strategies in Biomolecular Network Architecture</td>
</tr>
<tr>
<td></td>
<td>NCUR Slides: Architecture and Dynamics of Biomolecular Networks Facilitate Evolution of Persistence Strategies in Living Organisms</td>
</tr>
<tr>
<td>Cynthia Liu</td>
<td>Workflow management comparisons</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------------------------</td>
</tr>
</tbody>
</table>
| B.S. Bioengineering (2019)  
minor in Computer Science | 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. |

**Poster: Comparative Analysis of Genomic Sequencing Workflow Management Systems**

<table>
<thead>
<tr>
<th>Brian Rao</th>
<th>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.</th>
</tr>
</thead>
</table>
| B.S Integrative Biology (2018)  
Minor in Informatics |  

**Other Collaborations**

| Dr. Matthew Hudson | HPCBio, Carver Biotechnology Center  
http://hpcbio.illinois.edu/ |
|---------------------|------------------------------------------------------------------|
| Bioinformatics  
Crop Science |  

<table>
<thead>
<tr>
<th>Dan Wickland</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ph.D. Informatics (2019)</td>
</tr>
<tr>
<td>Dr. Daniel Katz</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Computer Science</td>
</tr>
<tr>
<td>Github: Swift Variant Calling</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Azza Ahmed</th>
<th>Madak-Erdogan Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Science</td>
<td>Systems Biology of Estrogen Signaling</td>
</tr>
<tr>
<td>University of Khartoum</td>
<td></td>
</tr>
<tr>
<td>advised by Dr. Faisal Fadieelmola</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dr. Zeynep Madak-Erdogan</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Food Science &amp; Human Nutrition</td>
<td></td>
</tr>
<tr>
<td>Madak-Erdogan Lab</td>
<td>Systems Biology of Estrogen Signaling</td>
</tr>
<tr>
<td>• NCSA Faculty Fellow 2017-2018</td>
<td>• Understanding Breast Cancer Disparities in African-American Women</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Brandi Smith</th>
<th>H3Africa Consortium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ph.D. Food Science and Human Nutrition (2021)</td>
<td></td>
</tr>
<tr>
<td>bioinformatics workflows in the cloud</td>
<td>• custom genotyping chip for African populations</td>
</tr>
<tr>
<td>H3Africa bioinformatics node accreditation</td>
<td></td>
</tr>
<tr>
<td>Morgan Taschuk</td>
<td>OICR</td>
</tr>
<tr>
<td>---------------</td>
<td>-----</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• production infrastructure for primary genomics analyses</td>
</tr>
<tr>
<td></td>
<td>• reproducibility of research in cancer genomics</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Paul Hatton</th>
<th>University of Birmingham</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPC / Visualisation</td>
<td></td>
</tr>
</tbody>
</table>

[OICR logo]

[University of Birmingham logo]