REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
14-03-2018	Final Report	19-Aug-2016 - 18-Dec-2017

4. TITLE AND SUBTITLE
Final Report: High Performance Computing for Faculty and Students at University of Houston - Downtown

5a. CONTRACT NUMBER
W911NF-16-1-0480

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
106012

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Houston-Downtown
One Main Street
Suite North 813
Houston, TX 77002 -1001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
68847-MA-REP.15

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Benjamin Soibam
UU	UU	UU	UU		19b. TELEPHONE NUMBER 713-226-5216

RPPR Final Report
as of 17-Apr-2018

Agency Code:

Proposal Number: 68847MAREP **Agreement Number: W911NF-16-1-0480**
INVESTIGATOR(S):

Name: Hong Lin
Email: linh@uhd.edu
Phone Number: 7132212781
Principal: N

Name: Ph.D Benjamin Soibam
Email: soibamb@uhd.edu
Phone Number: 7132265216
Principal: Y

Organization: **University of Houston-Downtown**
Address: One Main Street, Houston, TX 770021001
Country: USA
DUNS Number: 039674494 EIN: 746001399

Report Date: 18-Mar-2018 Date Received: 14-Mar-2018
Final Report for Period Beginning 19-Aug-2016 and Ending 18-Dec-2017
Title: High Performance Computing for Faculty and Students at University of Houston - Downtown
Begin Performance Period: 19-Aug-2016 **End Performance Period:** 18-Dec-2017
Report Term: 0-Other

Submitted By: Ph.D Benjamin Soibam Email: soibamb@uhd.edu
Phone: (713) 226-5216

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: **STEM Participants:**

Major Goals: 1. Acquisition of High Performance computing cluster (HPCC) at University of Houston-Downtown:

The main goal of this grant is to acquire a high-performance computing cluster at the University of Houston-Downtown. The HPCC will contain multiple nodes for running jobs, several terabytes of for storage, and high memory RAMs for running jobs. Faculty and students who need high computation resource in their research and coursework will have access to the cluster.

2. PI and co-PI undergo training sessions related to the usage of the cluster:

The PI and co-PI will receive training for the configuration of the clusters and the grid, the deployment of grid services, and the optimal use of the computing facilities. The PI and co-PI will attain hands on training on using high performance computing cluster and other related areas.

3. Use of the HPCC by faculty and students in research projects:

The instrument will be used for research purposes. Any Faculty or student, who is conducting research projects affiliated to UHD and requires high computing power, will have access to the cluster. One single project folder will be created for a faculty who uses the cluster for research purposes. Students who are working under that particular faculty can create sub-directories to store data and perform computation within the same project folder. However, students won't be able to directly access other subdirectories. Some identified research areas are large scale bioinformatics, computational brain wave modeling and analysis, molecular modeling and relativistic quantum mechanical calculations, robotics, large scale phylogenetic, population genetic analyses, process modeling/simulation/control etc.

4. Access of the HPCC by faculty and students in regular courses:

We believe experience in using HPC tools should not be limited to faculty and graduate students for research

purposes. We propose to integrate this computing resource while teaching undergraduate courses. The HPCC will be integrated into existing courses such Parallel Computing, Programming Language Concepts, Data Structures. This HPCC will also be an important resource for graduate program in Data Analytics at UHD. Students will have access to this HPCC for their projects and appropriate courses. This is an application-based master's program that will allow motivated and ambitious students to learn the statistical and computation tools to assemble, structure, and analyze large data sets.

Accomplishments: 1. Acquisition of the High Performance computing cluster (A DETAILED description of the components of the HPC is described at the end as a "report")

The high-performance computing cluster (HPC) was been successfully acquired and installed at the University of Houston-Downtown and has been functioning properly. Acquiring and installing a HPC cluster is no easy task, but the PI along with the information technology section of University of Houston-Downtown worked hard with the vendor and their engineers to craft the proper configuration of the cluster and its successful delivery and installation. The HPC was supposed to be delivered and installed at an earlier time. However because of Hurricane Harvey that hit Houston, the delivery was delayed. However, once things calmed down after Harvey, it was delivered and was successfully installed. Very briefly, there is one head/login node with 64 GB memory RAM and 32 computing nodes. Each computing node has 256GB memory with 24-cores (48 threads). The interconnects between different components is done using Omni-Path Networking (Line # 9) which allows fast communication. The cluster management is done via Bright Cluster which contains a nice interface for the administrator to manage and monitor the cluster. There is a 5 years warranty and parts replacement on the head node, and 3 years warranty and parts replacement on other components of the HPCC. The bright cluster management software will be used to manage the cluster.

2. PI and co-PI have successfully undergone training

PI Dr. Soibam and co-PI Dr. Lin have gone through 40 hours of training on theory and laboratory sessions on

High performance computing (HPC) at Rice University Ken Kennedy Institute during the period May 22-26, 2017. Topics ranging from MPI, OpenMP, Pthreads, Performance analysis (HPCToolkit and Jumpshot), accelerated computing using GPGPU (OpenCL and CUDA), and Parallel i/o were included in the training. They were able to access Rice's shared computing HPC infrastructure. PI and co-PI completed all hands on laboratory sessions supported by the instructor and lab assistants each day. This allowed PI and co-pies to have a practical understanding of various HPC tools. PI Dr. Soibam went through 40 hours of training on theory and laboratory sessions on Data Science at Rice University Ken Kennedy Institute during the period May 15-19, 2017. Topics ranging from Python, R, Deep Learning, supervised learning, unsupervised learning, data analytics on a computing cluster, bid data analytics were a part of the training sessions. PI was able to access Rice's shared computing HPC infrastructure. PI completed all hands on laboratory sessions supported by the instructor and lab assistants each day. This allowed PI to have a practical understanding of various data analytics tools and how they interface with computing cluster. Co-PI Dr. Lin also successfully completed Mplus Short Courses, Johns Hopkins Bloomberg School of Public Health, August 16-19, 2017. Topics includes Regression and mediation analysis using Mplus, and dynamic structural equation modeling of intensive longitudinal data using Mplus version 8.

3. Use of HPCC of faculty and students in Research Projects

The HPC cluster has been successfully used in research projects at UHD. Some of the research projects are:
- Computational Brain wave modeling and analysis
- Large-Scale Bioinformatics for identification of super-lncRNAs
- Large-Scale Bioinformatics for identifying breast cancer biomarkers
- Large scale biological network analysis and applications in cancer

In the coming months, more projects will use the cluster, some of them are
Phylogenetics and large-scale population genetic analysis, Computational Chemistry, Robotics etc.

4. Access of the HPCC by faculty and students in regular courses:

The HPC was integrated in a course on Predictive analytics course as a part of the Masters in Data Analytics program. Two teams of students requested access to the cluster. The first team worked on predictive analytics as a Football Coaching Aid. The team implemented various predictive analytics models such as Support vector machines, Decision Trees, Random Forest, and Deep Learning models to analyze historical NFL game data to

predict the outcome of a game, the number of yards gained or lost and discrete characterization of play types and directions. The second team worked on classifying person's sleep condition based on EEG waveforms recorded during sleep. In the project, several predictive models were used to predict whether a person has a sleep disorder that includes insomnia, bruxism, narcolepsy, sleep-disorder breathing.
In the coming months, depending on when courses are offered, other courses such as Parallel Computing, Programming Language Concepts, Data Structures will also allow students to use the HPC.

Products connected to the equipment:

Peer-reviewed Journals:

1. Wang, F, Dohogne, Z, Yang, J, Liu, Y, Soibam, B (2018). Predictors of breast cancer cell types and their prognostic power in breast cancer patients. BMC Genomics, 19, 1:137.
2. Robertson, MJ, Soibam, B, O'Leary, JG, Sampaio, LC, Taylor, DA (2018). Recellularization of rat liver: An in vitro model for assessing human drug metabolism and liver biology. PLoS ONE, 13, 1:e0191892.
3. Soibam, B (2017). Super-lncRNAs: identification of lncRNAs that target super-enhancers via RNA:DNA:DNA triplex formation. RNA, 23, 11:1729-1742.
4. Valenzuela, N, Soibam, B, Li, L, Wang, J, Byers, LA, Liu, Y, Schwartz, RJ, Stewart, MD (2017). HIRA deficiency in muscle fibers causes hypertrophy and susceptibility to oxidative stress. J. Cell. Sci., 130, 15:2551-2563.

Book Chapters:

1. Hong Lin, Introducing Scientific Measurement into Contemplative Practices, in: H. Lin, Q. Wang, D. Grimes (eds.), Empirical Studies of Contemplative Practices, Nova Science Publishers, 2018, to appear.

Conference papers, posters, presentations:

2. Hong Lin, V. Rajinikanth, Evaluation of Normal, Abnormal and Meditation EEG Signals Based on Amplitude Level and Entropy Value, Proceedings of 2018 the 8th International Workshop on Computer Science and Engineering (WCSE 2018), Bangkok, Thailand, June 28-30, 2018, to appear.
3. Hong Lin, Lalit, Grover, Thomas Wilson, Joseph Dao, Service Learning with Data Mining (Poster presentation), 2018 University of Houston-Downtown High-Impact Practices & Community Engagement Showcase, UHD, April 10, 2018.
4. Ruiz, E, Soibam, B Networks of Co-Expressed Protein Coding Genes and Long Non-Coding RNAs in Multiple Cancers. 29th HENAAC Conference. October, 2017, Pasadena, CA.
5. Singh, T, Soibam, B Identifying Genetic Markers to Improve Prognosis And Therapy of Melanoma Tumors 29th HENAAC Conference. October, 2016, Pasadena, CA.
6. Alhamadani, A, Soibam, B Predictive Analytics of Cell Types Using Single Cell Gene Expression Profiles. 29th HENAAC Conference. October, 2016, Pasadena, CA.

Others:
1. PI Dr. Soibam hosted one undergraduate student under the URAP (AEOP) under the educational outreach program by DoD. The undergraduate student performed research for a total of 300 hours in summer of 2017. The stipend of $3000 was sponsored by DoD.

2. PI Dr. Soibam hosted one high school student under the HSAP (AEOP) under the educational outreach program by DoD. The high school student performed research for a total of 300 hours in summer of 2017. The stipend of $3000 was sponsored by DoD.

3. PI Dr. Soibam receives Faculty Development Grant from the university of Houston-Downtown.

RPPR Final Report
as of 17-Apr-2018

Training Opportunities: PI Dr. Soibam and co-PI Dr. Lin have gone through 40 hours of training on theory and laboratory sessions on High performance computing (HPC) at Rice University Ken Kennedy Institute during the period May 22-26, 2017. Topics ranging from MPI, OpenMP, Pthreads, Performance analysis (HPCToolkit and Jumpshot), accelerated computing using GPGPU (OpenCL and CUDA), and Parallel i/o were included in the training. They were able to access Rice's shared computing HPC infrastructure. PI and co-PI completed all hands on laboratory sessions supported by the instructor and lab assistants each day. This allowed PI and co-pies to have a practical understanding of various HPC tools. PI Dr. Soibam went through 40 hours of training on theory and laboratory sessions on Data Science at Rice University Ken Kennedy Institute during the period May 15-19, 2017. Topics ranging from Python, R, Deep Learning, supervised learning, unsupervised learning, data analytics on a computing cluster, bid data analytics were a part of the training sessions. PI was able to access Rice's shared computing HPC infrastructure. PI completed all hands on laboratory sessions supported by the instructor and lab assistants each day. This allowed PI to have a practical understanding of various data analytics tools and how they interface with computing cluster. Co-PI Dr. Lin also successfully completed Mplus Short Courses, Johns Hopkins Bloomberg School of Public Health, August 16-19, 2017. Topics includes Regression and mediation analysis using Mplus, and dynamic structural equation modeling of intensive longitudinal data using Mplus version 8.

Results Dissemination: Faculty and students are being contacted about the availability of the HPC. They are being given accounts so that they can have access to the HPC.

Honors and Awards: PI Dr. Soibam hosted one undergraduate student under the URAP (AEOP) under the educational outreach program by DoD. The undergraduate student performed research for a total of 300 hours in summer of 2017. The stipend of $3000 was sponsored by DoD.

PI Dr. Soibam hosted one high school student under the HSAP (AEOP) under the educational outreach program by DoD. The high school student performed research for a total of 300 hours in summer of 2017. The stipend of $3000 was sponsored by DoD.

PI Dr. Soibam receives Faculty Development Grant (2017 and 2018) from the university of Houston-Downtown.

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Benjamin Soibam
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Co PD/PI
Participant: Hong Lin
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

ARTICLES:

RPPR Final Report
as of 17-Apr-2018

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: BMC Genomics
Publication Identifier Type: DOI Publication Identifier: 10.1186/s12864-018-4527-y
Volume: 19 Issue: 1 First Page #:
Date Submitted: Date Published: 2/1/18 12:00PM
Publication Location:
Article Title: Predictors of breast cancer cell types and their prognostic power in breast cancer patients
Authors: Fan Wang, Zachariah Dohogne, Jin Yang, Yu Liu, Benjamin Soibam
Keywords: Bioinformatics – Logistic regression – HER2 positive – Single cell sequencing
Abstract: We outline a predictive analytics pipeline to accurately predict 6 breast cancer cell types using single cell gene expression profiles. Instead of building predictive models using the complete human transcripts, the pipeline first eliminates predictors with low expression and low variance. A multinomial penalized logistic regression further reduces the size of the predictors to only 308, out of which 34 are long non-coding RNAs. Tuning of predictive models shows support vector machines and neural networks as the most accurate models achieving close to 98% prediction accuracies. We also find that mixture of protein coding genes and long non-coding RNAs are better predictors compared to when the two sets of transcripts are treated separately.

Acknowledged Federal Support: **Y**

Report for Proposal No. 68847-RT-REP: High Performance Computing for Faculty and Students at University of Houston – Downtown

Benjamin Soibam, PI
Hong Lin, co-PI

The Equipment

The high-performance computing cluster (HPC) was been successfully acquired and installed at the University of Houston-Downtown and has been functioning properly. Acquiring and installing a HPC cluster is no easy task, but the PI along with the information technology section of University of Houston-Downtown worked hard with the vendor and their engineers to craft the proper configuration of the cluster and its successful delivery and installation. The HPC was supposed to be delivered and installed at an earlier time. However because of Hurricane Harvey that hit Houston, the delivery was delayed. However, once things calmed down after Harvey, it was delivered and was successfully installed.

Below is a table (Table 1) describing all the different components of the HPC. Very briefly, there is one head/login node with 64 GB memory RAM (Line#1) and 32 computing nodes (Line #3). Each computing node has 256GB memory with 24-cores (48 threads). The interconnects between different components is done using Omni-Path Networking (Line # 9) which allows fast communication. The cluster management is done via Bright Cluster which contains a nice interface for the administrator to manage and monitor the cluster.

Table 1: HPC cluster components

Line #	Item	unit price	Qty	Extension
1	Rackform U623.v6 (Head/Login Node) CPU: 2 x Intel Xeon E5-2623v4, 2.6GHz (4-Core, HT, 10MB Cache, 85W) 14nm RAM: 64GB (8 x 8GB DDR4-2400 ECC Registered 1R 1.2V DIMMs) Operating at 2400 MT/s Max	$5,272.91	2	$10,545.82
2	1 Rackform U623.v6 (StorX 1Z - Controller Node) CPU: 2 x Intel Xeon E5-2623v4, 2.6GHz (4-Core, HT, 10MB Cache, 85W) 14nm RAM: 128GB (8 x 16GB DDR4-2400 ECC Registered 2R 1.2V DIMMs) Operating at 2400 MT/s Max	$6,564.26	1	$6,564.26
3	Rackform R4422.v6.1 "24-cores [48-threads] / 256GB per Node with 42 month warranty CPU: 2 x Intel Xeon E5-2650v4, 2.2GHz (12-Core, HT, 30MB Cache, 105W) 14nm RAM: 256GB (8 x 32GB DDR4-2400 ECC Registered 2R 1.2V LRDIMMs) Operating at 2400 MT/s Max	$27,109.06	8	$216,872.48
4	1 Storform D59J.v3 "Storage 44 drive JBOD - ZFS on Linux"	$3,082.44	1	$3,082.44

5	1 StorX ZFS Pool "Performance	$2,372.76	1	$2,372.76
6	StorX ZFS Capacity	$451.61	12	$5,419.32
7	1 StorX ZFS Pool "FileStore/Archive"	$0.00	1	$0.00
8	StorX ZFS Capacity Device - vdev "FileStore/Archive"	$2,955.45	3	$8,866.35
9	Configurator "Omni-Path Networking"	$17,588.33	1	$17,588.33
10	1 Cisco Catalyst WSC2960L-48TS-LL Switch	$2,405.53	1	$2,405.53
11	Configurator "Bright Licensing" Omni-Path Switch: Intel Omni-Path 100Gbps Fabric Switch, 48-port QSFP28, 2PS,	$487.46	35	$17,061.10
12	1 Professional Services Tool "Bright/ZFS installation"	$1,792.11	1	$1,792.11
13	Professional Services Tool "Optional On-site deployment & Bright Cluster Training	$4,096.26	1	$4,096.26
14	Rack Integration Configurator Rack Integration: Medium Rack (11-50 Nodes) Rack: APC NetShelter SX 42U Oversized Rack Enclosure, 750mm (W) x 1200mm (D) - Shock Packaging Included (2000lbs) Network/Data Cables: 1/10GbE CAT6 RJ45 Cables Medium Rack (11-50 Nodes) Management Cables: No Item Selected InfiniBand Cables: No Item Selected Vertical PDUs: 4 x APC Metered Rack PDU, 30A/200-240V (NEMA L6-30P), 36x IEC 60320 C13 and 6x C19 Outlets, Zero U - 3m Cord	$8,330.77	2	$16,661.54
15	Workform 1000.v6 CPU: Intel Xeon E3-1225v5, 3.3GHz (4-Core, 8MB Cache, 14nm) 80W RAM: 16GB (2 x 8GB DDR4-2400 ECC Unbuffered 2R DIMMs) Operating at 2133 MT/s Max	$1,764.46	2	$3,528.92
16	iMac Desktop to connect to server	$1,774.00	1	$1,774.00
17	Seagate 8TB Enterprise Capacity 3.5 HDD V.5 (12Gb/s, 7.2K RPM, 256MB Cache, 4Kn)	$377.88	3	$1,133.64

18	1 Intel 240GB DC S3520 Series 3D MLC (6Gb/s, 1.4 DWPD) 2.5" SATA SSD	$174.09	1	$174.09
			Total	**$319,938.95**

PI Dr. Soibam and co-PI Dr. Lin have gone through 40 hours of training on theory and laboratory sessions on High performance computing (HPC) at Rice University Ken Kennedy Institute during the period May 22-26, 2017. Topics ranging from MPI, OpenMP, Pthreads, Performance analysis (HPCToolkit and Jumpshot), accelerated computing using GPGPU (OpenCL and CUDA), and Parallel i/o were included in the training. They were able to access Rice's shared computing HPC infrastructure. PI and co-PI completed all hands on laboratory sessions supported by the instructor and lab assistants each day. This allowed PI and co-pies to have a practical understanding of various HPC tools. PI Dr. Soibam went through 40 hours of training on theory and laboratory sessions on Data Science at Rice University Ken Kennedy Institute during the period May 15-19, 2017. Topics ranging from Python, R, Deep Learning, supervised learning, unsupervised learning, data analytics on a computing cluster, bid data analytics were a part of the training sessions. PI was able to access Rice's shared computing HPC infrastructure. PI completed all hands on laboratory sessions supported by the instructor and lab assistants each day. This allowed PI to have a practical understanding of various data analytics tools and how they interface with computing cluster. Co-PI Dr. Lin also successfully completed Mplus Short Courses, Johns Hopkins Bloomberg School of Public Health, August 16-19, 2017. Topics includes Regression and mediation analysis using Mplus, and dynamic structural equation modeling of intensive longitudinal data using Mplus version 8.

Current Research Projects (also overlap with projects included in the original proposal) supported by the HPC

1. Computational Brain wave modeling and analysis

We have built a system for brain state analysis using electroencephalogram (EEG) data. High performance computing platform provides computational power to ensure real time processing of requests. In this project, the student work on feature extraction from EEG data and perform thorough testing to maximize the recognition rates of the brain state models. The student develop parallel machine learning algorithms, and implement the back-end programs to ensure real-time processing of the EEG data.
The challenge is to contemplate the rigorous time series analysis of brain waves to decipher trend, irregularities, cycles, seasonality and other variations among waves during different states. Therefore, feature extraction is an important part of EEG data analysis. The project aims to address the complexity of classification of brain waves data by modeling the major brain waves and achieve an efficient and predictable brain wave modeling system that has potential application in hospitality and clinical industry for self-controlled deep brain relaxation and early diagnosis of various brain abnormalities respectively. To ensure real-time processing of the physiological data and instant brain state classification, parallel programs have been used to perform dynamic machine learning algorithms. The results we have performed on the computing cluster indicate a big speedup when we use parallelism in the EEG modeling algorithms. More experiments will be performed on the cluster to examine the performance gain by parallelism as well as the feasibility of fine-tuning the models by incorporating multi-lateral physiological data.

2. Large-Scale Bioinformatics for identification of super-lncRNAs

Super-enhancers are characterized by high levels of Mediator binding and are major contributors to the expression of their associated genes. They exhibit high levels of local chromatin interactions and a higher order of local chromatin organization. On the other hand, lncRNAs can localize to specific DNA sites by forming a RNA:DNA:DNA triplex, which in turn can contribute to local chromatin organization. We characterize a new class of lncRNAs called super-lncRNAs that target super-enhancers and which can contribute to the local chromatin organization of the super-enhancers. First, we compiled 27 human cell and tissue types, which have publicly

available data on lncRNA expression profiles from the NIH epigenome roadmap (http://www.roadmapepigenomics.org/mapping) as well as super-enhancer coordinates (http://bioinfo.au.tsinghua.edu.cn/dbsuper/). To identify lncRNAs that target super-enhancers (which we term super-lncRNAs) in each cell or tissue type, we used triplexator to identify triplex-forming sites by expressed lncRNAs (fpkm > 0.5) in the super-enhancer (SE) sequences. Only the lncRNAs that are active in the tissue of interest were considered. Our procedure to identify super-lncRNAs. An equal number of random sequences with comparable lengths were extracted from the human genome, and triplex-forming sites by lncRNAs on these sequences were also obtained. To obtain statistically meaningful super-lncRNAs, a logistic regression model (Materials and Methods) based on the frequency of binding sites of an lncRNA in the actual SEs and a random set was used to detect lncRNAs, which were a significant source of targets. Using this approach, we identified 442 unique super-lncRNAs transcripts that target super-enhancers in at least one of 27 cell or tissue types. Super-lncRNAs in a particular cell line or tissue type passed a *P*-value <0.05 cutoff and targeted at least 3% of super-enhancers in that particular cell line or tissue. The HPC cluster was used for all the genomic, bioinformatics analysis that required high power memory intensive computational tasks. This work would not have been feasible without the HPC acquired through the DoD grant.

3. Large-Scale Bioinformatics for identifying breast cancer biomarkers

Comprehensive understanding of intratumor heterogeneity requires identification of molecular markers, which are capable of differentiating different subpopulations and which also have clinical significance. One important tool that has been addressing this issue is single cell RNA-Sequencing (scRNASeq) that allows the quantification of expression profiles of transcripts in individual cells in a population of cancer cells. Using the expression profiles from scRNASeq, current studies conduct analysis to group cells into different subpopulations using clustering algorithms. In this study, we explore scRNASeq cancer data from a different perspective. We focus on scRNASeq data originating from cancer cells pertaining to a particular cancer type, where the cell type or the subpopulation to which each cell belongs is known. We investigate if the “cell type” of a cancer cell can be predicted based on the expression profiles of a small set of transcripts. We outline a predictive analytics pipeline to accurately predict 6 breast cancer cell types using single cell gene expression profiles. Instead of building predictive models using the complete human transcripts, the pipeline first eliminates predictors with low expression and low variance. A multinomial penalized logistic regression further reduces the size of the predictors to only 308, out of which 34 are long non-coding RNAs. Tuning of predictive models shows support vector machines and neural networks as the most accurate models achieving close to 98% prediction accuracies. We also find that mixture of protein coding genes and long non-coding RNAs are better predictors compared to when the two sets of transcripts are treated separately. A signature risk score originating from 65 protein coding genes and 5 lncRNA predictors is associated with prognostic survival of TCGA breast cancer patients. This association was maintained when the risk scores were generated using 65 PCGs and 5 lncRNA separately. We further show that predictors restricted to a particular cell type serve as better prognostic markers for the respective patient subtype.

The HPC cluster was used for all the genomic, bioinformatics analysis that required high power memory intensive computational tasks. This work would not have been feasible without the HPC acquired through the DoD grant.

4. Large scale biological network analysis and applications in cancer

According to the National Cancer Institute, approximately 38.5 percent of men and women will be diagnosed with cancer of any site in their lifetime. In an effort to reduce this percentage and further cancer research, a gene network approach can be implemented to suggest core pathways that are conserved across multiple cancer types and build confidence in gene interactions based on co-expression of protein-coding genes and long non-coding RNAs. Protein-coding genes are significant since they encode for specific proteins, which in turn will

perform critical functions. Although there has not been extensive research conducted on long non-coding RNAs, they are important regulators of gene expression; thus, remain a significant component of an organism's genome. With the use of a network search algorithm, neXus, significant active subnetworks can be discovered based on differential expression data of the co-expressed protein-coding genes and the long non-coding RNAs from multiple cancer types. The intent for this research project is to find compelling modules which are differentially regulated with similar expression patterns across eight different cancer types based on the genomic data of co-expressed protein-coding genes and long non-coding RNAs. There have been various studies which utilize the neXus network search algorithm; however, it has not been applied to various cancer types using co-expressed data of protein-coding genes and long non-coding RNAs.To deliver on this research project's intent, data was first gathered from a publicly available database, The Cancer Genome Atlas Project. The eight cancer types utilized include: breast invasive carcinoma, head and neck squamous cell carcinoma, kidney renal clear cell carcinoma, liver hepatocellular carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, prostate adenocarcinoma, and thyroid carcinoma. Protein-coding gene expression data and long non-coding RNA data from patients that did not have tumors (normal patients) and patients that exhibited having a malignant tumor (cancer patients) was used for this research project. Before the generation of networks could be made, the gene expression data had to be filtered. After matching the patients' co-expressed data using Excel, the p-values for each gene in all of the cancers was obtained, as well as the false discovery rate (FDR). The 2995 genes that had an FDR of less than 0.05 and were common across all of the eight cancer types were considered significant. In order to generate networks for each cancer type, a MATLAB script was implemented, which required the genes of interest and the gene expression text files for normal and cancer patients. This script searched for an absolute correlation between the differentially expressed data, using the cutoffs of 0.2 and 0.3. After receiving the four output files from this script for all of the cancer types, the subnetworks began to be generated using the neXus algorithm. In order for a subnetwork to be generated, all of the genes contained in the subnetwork must be differentially expressed between normal and cancer patients and there must be a strong interaction between the neighboring genes in the subnetwork among the multiple cancer types. This algorithm utilized two cutoffs, Depth First Search (DFS) and Clustering Coefficient (CC). The DFS cutoff determined the amount of interaction the genes had to meet to determine how deep the subnetwork would be, while the CC cutoff determined the connectedness of the subnetwork. The growth of a subnetwork begins with a seed gene and has other genes (nodes) added to its functional neighborhood if the path confidence of the node from the seed gene is above a certain threshold. The high network score of a subnetwork depends on the number of genes and interaction; therefore, if the high network score constraint or the clustering coefficient constraint is not met, the growth of the subnetwork is terminated. As a result, there was a total generation of 3072 active subnetworks after using eleven different parameters. A closer analyzation of a subnetwork from a conservative parameter, DFS of 0.8 and CC of 0.3, showed that it was comprised of 59 genes. Distinguishably, the genes of the subnetwork were most highly expressed in cancer patients than in normal patients and were strongly correlated to each other across the eight cancer types. In addition, numerous gene functions within the subnetwork are known to cancer pathways. Although there was not a close analyzation of the rest of the subnetworks, this data presents promising results. Therefore, this research project will be continued with further analyzation of the generated subnetworks from different parameters. Overall, the data suggests that subnetworks of co-expressed protein-coding genes and long non-coding RNAs has the potential to advance cancer research and reduce the lifetime risk of developing cancer statistic.

The HPC cluster was used for all the genomic, bioinformatics analysis that required high power memory intensive computational tasks. This work would not have been feasible without the HPC acquired through the DoD grant.

5. Integration of HPC in regular courses:

The HPC was integrated in a course on Predictive analytics course as a part of the Masters in Data Analytics program. Two teams of students requested access to the cluster. The first team worked on predictive analytics as a Football Coaching Aid. The team implemented various predictive analytics models such as Support vector machines, Decision Trees, Random Forest, and Deep Learning models to analyze historical NFL game data to predict the outcome of a game, the number of yards gained or lost and discrete characterization of play types and directions. The second team worked on classifying person's sleep condition based on EEG waveforms recorded during sleep. In the project, several predictive models were used to predict whether a person has a sleep disorder that includes insomnia, bruxism, narcolepsy, sleep-disorder breathing.

Products connected to the equipment:

Peer-reviewed Journals:

1. Wang, F, Dohogne, Z, Yang, J, Liu, Y, Soibam, B (2018). Predictors of breast cancer cell types and their prognostic power in breast cancer patients. BMC Genomics, 19, 1:137.
2. Robertson, MJ, Soibam, B, O'Leary, JG, Sampaio, LC, Taylor, DA (2018). Recellularization of rat liver: An in vitro model for assessing human drug metabolism and liver biology. PLoS ONE, 13, 1:e0191892.
3. Soibam, B (2017). Super-lncRNAs: identification of lncRNAs that target super-enhancers via RNA:DNA:DNA triplex formation. RNA, 23, 11:1729-1742.
4. Valenzuela, N, Soibam, B, Li, L, Wang, J, Byers, LA, Liu, Y, Schwartz, RJ, Stewart, MD (2017). HIRA deficiency in muscle fibers causes hypertrophy and susceptibility to oxidative stress. J. Cell. Sci., 130, 15:2551-2563.

Book Chapters:

1. Hong Lin, Introducing Scientific Measurement into Contemplative Practices, in: H. Lin, Q. Wang, D. Grimes (eds.), Empirical Studies of Contemplative Practices, Nova Science Publishers, 2018, to appear.

Conference papers, posters, presentations:

2. Hong Lin, V. Rajinikanth, Evaluation of Normal, Abnormal and Meditation EEG Signals Based on Amplitude Level and Entropy Value, Proceedings of 2018 the 8th International Workshop on Computer Science and Engineering (WCSE 2018), Bangkok, Thailand, June 28-30, 2018, to appear.
3. Hong Lin, Lalit, Grover, Thomas Wilson, Joseph Dao, Service Learning with Data Mining (Poster presentation), 2018 University of Houston-Downtown High-Impact Practices & Community Engagement Showcase, UHD, April 10, 2018.
4. Ruiz, E, Soibam, B Networks of Co-Expressed Protein Coding Genes and Long Non-Coding RNAs in Multiple Cancers. 29th HENAAC Conference. October, 2017, Pasadena, CA.
5. Singh, T, Soibam, B Identifying Genetic Markers to Improve Prognosis And Therapy of Melanoma Tumors 29th HENAAC Conference. October, 2016, Pasadena, CA.
6. Alhamadani, A, Soibam, B Predictive Analytics of Cell Types Using Single Cell Gene Expression Profiles. 29th HENAAC Conference. October, 2016, Pasadena, CA.

Others:

1. PI Dr. Soibam hosted one undergraduate student under the URAP (AEOP) under the educational outreach program by DoD. The undergraduate student performed research for a total of 300 hours in summer of 2017. The stipend of $3000 was sponsored by DoD.

2. PI Dr. Soibam hosted one high school student under the HSAP (AEOP) under the educational outreach program by DoD. The high school student performed research for a total of 300 hours in summer of 2017. The stipend of $3000 was sponsored by DoD.

3. PI Dr. Soibam receives Faculty Development Grant from the university of Houston-Downtown.

Planned Projects to supported by the HPC in the immediate future:

1. Phylogenetics and large-scale population genetic analysis: Large-scale phylogenetic and population genetic analyses that would benefit from being run on a computing cluster. Currently, these analyses are carried out on a desktop computer and run for several days each. Our lab is moving toward collecting high throughput genomic data (rather than single gene sequences), which will increase analysis time and data storage exponentially on a desktop making it nearly impossible to do this kind of project without a computing cluster. All of the software we currently use is compatible with parallel computing clusters. These large datasets are difficult to store on desktop computers as well, so additional storage space will be required. Examples of current projects that would benefit from the computing cluster include phylogenetic analyses of turtles in the genus *Rhinoclemmys*, phylogeographic analysis of bats in the tribe *Lasiurini*, and population genetics of bowhead whales. Dr. Baird has previous experience using computing clusters to conduct phylogenetic analyses during her graduate studies at the University of Texas. Experiments conducted for different biological conditions and systems will regularly generate high throughput genomic data. An availability of a cluster can analyze large volumes of different data sets quick and simultaneously.

2. Computational Chemistry: Plutonium and uranium are two actinides regarded as biological hazards which are known to be toxic and exposure occurs. The current treatment for actinide poisoning is *chelation therapy* which involves the use of sequestering agents that bind actinides by forming actinide complexes which are then excreted from the body. A series of novel sequestering agents have been designed and developed for potential use in the treatment of actinide poisoning8. Prof. Benavides is proposing to study the plutonium and uranium complexes formed with some of these novel sequestering agents using a computational approach that includes molecular modeling of the actinide complexes followed by quantum mechanical calculations. The studies will allow Prof. Benavides and her undergraduate research students to determine the equilibrium geometries, molecular energies (computed IR spectra), and other molecular properties for various actinide clusters, which will then be compared with single x-ray structures in order to assess the validity of their molecular models. Figure on the left shows an example of one of the plutonium complex the we are interested in studying which is formed between Pu(IV) and the designed ligand 1-hydroxypridin-2-one. These proposed studies use quantum mechanical calculations9 on actinide clusters which require memoryand processing power that exceeds that capacity of a personal computer. For this reason, Benavides and her group have been unable to pursue these studies. The acquisition of a high-performance computing cluster will provide vast amount of memory and processing power that will render these studies feasible.

3. Robotics: The high-performance computing cluster will be used for teaching and research areas related to Robotics in UHD. High performance computing is a key to build robots that can sense and respond to signals in the unstructured environments. A variety of robotic algorithms have heavy computational demands to process a large amount of data. For example, a single camera mounted on a robot has produced one image with 24-bit color 640 x 480 pixels, which contributes about 900 K Bytes data. If a video sequence has 10 seconds and each second has 24 frames, the produced data size is about 216 M Bytes. The parallel computing and architecture provided by high performance computing cluster enable our students to study the complicated algorithms and systems of robots effectively and efficiently. Computing cluster will be used for

multiple levels in parallel computing including sensor level, functional module level, multi-robot system level, and algorithm level. Additionally, the requested storage resources can store a larger number of videos and images acquired by robots for student study in class and projects.

4. Opportunities for local high school teachers and/or students: High school students selected in the HOUSTON PREP program hosted at UHD will be introduced to High Performance Computing. They will run several computing applications in a cluster environment as well as on high-end windows workstation the same set of applications. Then they will compare the time taken to run the applications on both systems. Concepts of HPC systems will be introduced to them by giving them these kinds of assignments.