The Bioinformatics Component within the Biostatistics and Bioinformatics Shared Resource Facility

Chi Wang[1,2], Jinze Liu[1,3], Hunter Moseley[1,4], Jinpeng Liu[1,3], and Heidi L. Weiss[1,2]

[1]B[2]SRF Markey Cancer Center, [2]Department of Biostatistics, [3]Department of Computer Science, [4]Department of Molecular and Cellular Biochemistry

MISSION

Based on MCC EAB recommendations, current and anticipated needs of MCC Research Program members and careful research for similar services/support from other NCI cancer centers, we propose the Bioinformatics Component within the Biostatistics and Bioinformatics Shared Resource Facility. Our missions are:

- Provides expert bioinformatics solutions on study design, computational processing, statistical analysis, and integration of high-throughput genomic, transcriptomic, and metabolomic data for all MCC members
- Build and maintain an infrastructure that enables the application of robust and timely biostatistics and bioinformatics analysis for investigators to both publish their work and obtain new funding
- Serve as a central point of contact and venue for collaboration with bioinformatics, computational biology, and systems biology specialists at UK who have additional expertises

PERSONNEL

Co-directed by Drs. Chi Wang and Jinze Liu, the Bioinformatics Component has 3 faculty and 1 staff with diverse expertise in microarray, next generation sequencing, and metabolomics data analysis.

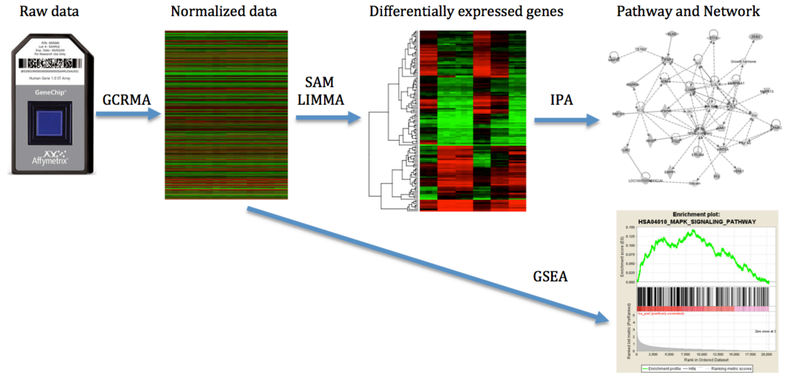
- Chi Wang, PhD (microarray and next generation sequencing)
- Jinze Liu, PhD (next generation sequencing)
- Hunter Moseley, PhD (Metabolomics)
- Jinpeng Liu, MS

SERVICES

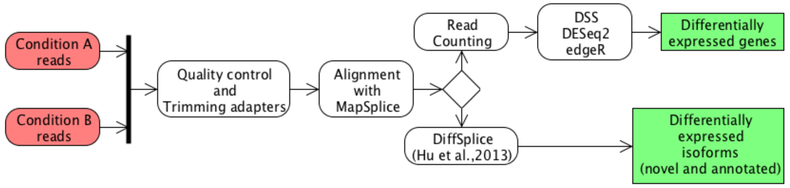
The Core aims to build and maintain robust and state-of-the-art analysis pipelines for analyzing, interpreting, and visualization of large-scale genomic, transcriptomic, and metabolomic data generated by Markey cancer research experiments. While these pipelines can be used for general purpose bioinformatics applications, they are specifically tailored to reveal mutations and complex behaviors of cancer genomes. We work closely with and provide custom bioinformatics solutions for MCC investigators. Our current services focus on the following areas. New services will be added depending on the demand.

- Microarray Data Processing and Analysis
  The core has developed a pipeline for microarray data processing and analysis, including data normalization, quality assessment, differential expression identification and visualization, and pathway/functional analysis.
- Next Generation Sequencing Data Processing and Analysis
  - ➢ DNA-seq data analysis with whole-genome sequencing or exome-sequencing
    The core has developed a pipeline for exome-sequencing data analysis, including data quality control, read alignment, variant calling, functional annotation and the identification of statistically significant variants differentiating across multiple groups.
  - ➢ RNA-seq data analysis
    The core has developed a pipeline for RNA-seq data analysis. The pipeline includes data quality control, read alignment, differential expression identification and visualization, and pathway/functional analysis. Besides gene expression analysis, we also support the discovery of novel alternative splicing as well as variant calling and fusion detection from RNA-seq data.
- Metabolomics Data Analysis
  The core provides informatics support for raw and intermediate data analysis of metabolomics datasets, especially stable isotope-resolved metabolomics datasets. Results of these analyses can feed into other biostatistical analyses provided by the core. Custom downstream metabolic modeling and relative flux analyses can be provided on a limited basis.
- Integrative Analysis of Multiple Genomics Datasets
  The core provides bioinformatics support to analyze the interaction or correlation across multiple genomic data.
- Genomic Data Mining
  The core utilizes genomic data repositories such as GEO, Oncomine, and TCGA to correlate genomic data from specific gene(s) of interest with clinical outcomes.
- Other Large-Scale Genomic Data Analysis
  The core provides bioinformatics support for other genomic experimental platforms such as the NanoString nCounter system.
- Grant-writing Support
  The core will help investigators with genomic study design, sample size/power calculation, data analysis plan, and writing of bioinformatics section.
- Training and Outreach
  The core will advertise services as they become available and work with investigators to establish new data analysis pipelines. The core will give informational seminars on supported analysis routines, and will host training series and workshops on commonly used bioinformatics tools, resources, and databases.

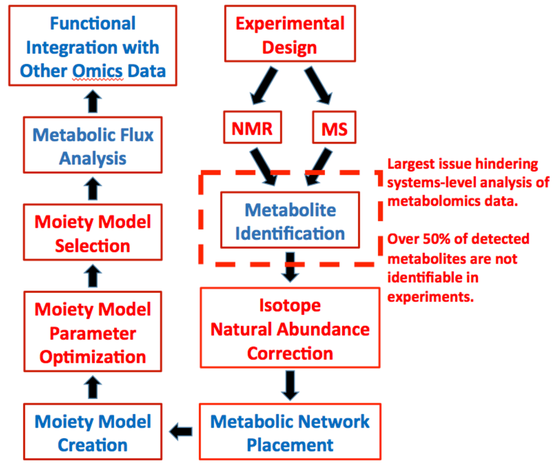
EXAMPLE PIPELINES



Example 1. Microarray data analysis pipeline developed by Dr. Chi Wang. We have developed R and JAVA scripts to implement the pipeline and to efficiently utilize available software such as Bioconductor, Ingenuity Pathway Analysis and GSEA.



Example 2. Next generation sequencing RNA-seq data analysis pipeline developed by Dr. Jinze Liu. We have developed a pipeline including numerous novel computational methods for the analysis of RNA-seq data. The pipeline takes raw sequencing reads from experimental samples under different conditions. These reads will first go through quality control and adapter trimming followed by the alignment to the reference genome. The pipeline may identify differentially expressed genes through EdgeR based on gene read count. Alternatively, the pipeline also identifies differentially expressed isoforms using DiffSplice, a graph-based method that allows the discovery of novel isoforms and mutations.



Example 3: Metabolomics Data Analysis Pipeline developed by Dr. Hunter Moseley. We are implementing a full data analysis pipeline for stable isotope resolved metabolomics experimental data. This pipeline starts with raw data analysis and reduction followed by metabolite identification, natural abundance correction, and placement. This is followed by metabolic modeling at the level of functional groups (chemical moieties) for atomic tracing and flux analysis of stable isotopes through cellular metabolism. Finally, the pipeline feeds into other omics-level data streams for integration.

COLLABORATIVE WORKS WITH MCC INVESTIGATORS

Over the past 2 years, we have collaborated with more than 15 MCC investigators from all Research Programs (CS, RR, DT and CP) in *in vivo*, biospecimen, clinical and population-based genomic studies utilizing different biostatistical and bioinformatics platforms. Our collaborations have led to many publications in high-quality journals.

MCC Investigator	Program	Published journal
Peter Zhou	CS	Cancer Cell
Peter Zhou	CS	Cell Reports
Peter Zhou	CS	Oncogene
Natasha Kyprianou	CS	PLOS ONE
Chunming Liu	CS	J. Biological Chemistry
Vivek Rangnekar	CS	J. Cellular Physiology
Suleiman Massarweh	DT	Future Oncology
Tianyan Gao	CS	Gastroenterology

BIOINFORMATICS METHODOLOGICAL WORKS

Our faculty members are actively developing novel bioinformatics methods to meet the computational and analytical challenges in dealing with complex high-throughput data. Over 20 papers have been published. Below are some of our recent publications.

- Wu H*, **Wang C*** and Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq. Biostatistics, 14(2): 232-43, 2013. *Authors with equal contribution
- Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR, Monroy A, Kuan PF, Hammond SM, Makowski L, Randell SH, Chiang DY, Hayes DN, Jones C, Liu Y, Prins JF, **Liu J**. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. Nucleic Acids Res, 41(2): e39, 2013.
- Carreer W, Flight R, and **Moseley H**. A computational framework for high-throughput isotopic natural abundance correction of omics-level ultra-high resolution FT-MS datasets. Metabolites, 3: 853-866, 2013.