MUSC Bioinformatics Core (MBC)

Sean M Courtney[1,2], William B. Glen[1,2], E. Starr Hazard[1,3], Gary Hardiman[1,4]

[1]MUSC Bioinformatics, Center for Genomics Medicine, [2]Department of Pathology & Laboratory Medicine, [3]Library Science and Informatics, [4]Departments of Medicine & Public Health, Medical University of South Carolina (MUSC), Charleston SC

`[musc-bioinformatics/]$`

Mission

The MUSC Bioinformatics Core (MBC) provides state of the art bioinformatics analysis for genomic and epigenomic assays and supports the research goals of MUSC investigators. A range of bioinformatics services that have proven to be transformative in enabling investigators to advance disease-related research goals are offered.

Services Offered/Capabilities

Experimental Design
- Bioinformatics support for assistance with experimental design, choice of technological platform, data analysis and data quality control.

High-Performance Computing Resources
- Access to High-performance computing (HPCC) resources, and infrastructure for data storage, backup and mining.

Expression Microarray Technology
- Analyses of conventional Affymetrix and Illumina expression microarray platforms for genome wide analysis of gene expression.

Bioinformatics Support for High-Throughput Sequencing Assays
- High-throughput sequencing data analysis for Illumina HiSeq 2500 and Miseq platforms, whole genome sequencing (WGS), RNA sequencing (RNAseq), microRNA sequencing (miRNAseq), chromatin immunoprecipitation linked to massively parallel sequencing (ChIP-Seq), MethylC-sequencing, capture based sequencing (exome/focused panel) and metagenomics (16S rRNA gene sequencing).

Systems biology analyses
- Network-based approaches for analyzing high dimensional genomic data sets.

Training and Consultation: User Interface
- Consultation and training of students, postdoctoral fellows, investigators and technical staff regarding high-throughput sequencing methodologies and data analysis will be provided.

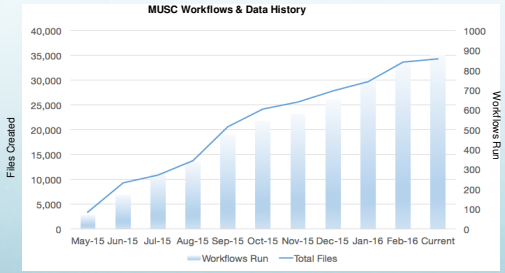
Core services are requested via a SPARC Request, after an initial consultation with Core staff.

MUSC HT Genomics at Scale

- Significantly improving local infrastructure, MUSC deployed the OnRamp BioInformatics Genomics Research Platform in May 2015
- OnRamp's GRP converges advanced genomics analysis, comprehensive data management, big data analytics and hyperscale servers
- This 'Big Data' solution utilizes hadoop software with automated data protection to seamlessly scale a 240 TB, 10-node server and storage infrastructure
- In the past 12 months since deployment, bioinformatics workflows have been greatly expanded, enabling a broad array of new analyses, including whole genome, exome, RNA variant, RNA fusion, ChIP and metagenomics
- More than 100TB of valuable genomic data has been generated, more than tripling the data generated in the prior 8 years at MUSC
- 900+ workflow analyses run to date
- The Core also maintains a traditional High Performance Cluster with 35TB of storage capacity.
- Additional state resources are available (including the high performance Palmetto Computing (HPC) resource at Clemson University).
- The MUSC Bioinformatics Core is focused on reducing barriers to utilization of bioinformatics tools to enable broad access and insights by MUSC investigators

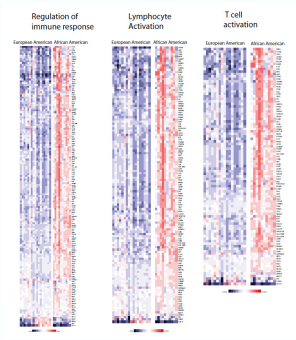
OnRamp Genomic Research Platform Capabilities

- Data cleaning, trimming and automated quality control
- RNA-Seq for single and paired-end reads
  - Gene Expression via Deseq2 and Tuxedo (Bowtie, Tophat, Cufflinks, HTseq)
  - RNA Fusion (ChimeraScan, Tophat-fusion)
  - *De Novo* Transcriptome Assembly (Trinity)
  - miRNAseq Comprehensive Analysis. Pipeline for microRNASequencing data (CAP-miRSeq)
  - RNA variant analysis
- DNA-Seq for single and paired-end reads (Picard, GATK)
  - Exome Seq Variant Calling
  - Whole Genome Seq Variant Calling
  - Advanced Variant Annotation and Filtering, including trio, tumor-normal (GEMINI)
- ChIP-seq for single and paired-end reads (MACS2, mosaics, HOMER)
- Metagenomics: 16S & shotgun (Superfocus, QIIME)
- Integrated reference databases: dbSNP, SnpEff, Clinvar, COSMIC, 1000 Genomes, OMIM, Ensembl, Refseq, CADD, ENCODE, GERP, KEGG, ESP, Polyphen, HPRD, FitCons, Pfam, SIFT, VISTA
- Hadoop Map/Reduce, Spark, R, Python, Perl, Bash and Java



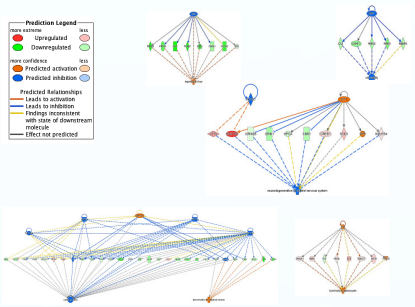
Available reference genomes
- Human: hg19, hg18, GRCh37
- Mouse: mm10, mm9, GRCm38, NCBIM37
- Rat: rn5, rn4, Rnor_6.0
- Dolphin: turTru1, turTru2
- Zebrafish: danRer7, Zv9, GRCz10
- Tetraodon: tetNig2, TETRA0DON8
- Atlantic Salmon: AGKD04
- Fugu: FUGU4, fr3
- Stickleback: BROADS1, gasAcu1
- Fly: dm3

Systems analysis of the prostate transcriptome



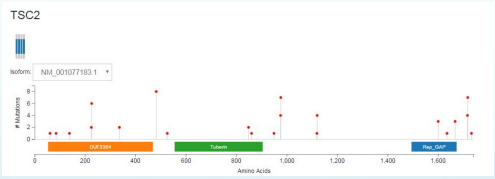
Differences in prostate transcriptional profiles between African American (AA) men and European American (EA) men. The study enrolled 27 subjects (10 AA and 17 EA men), who had selected surgical removal of the prostate (prostatectomy) as a definitive treatment for their prostate cancer. EA subjects (17 samples) were set as the control and AA subjects (10 samples) as the test for DE analyses to identify race-associated differences in prostate gene expression. Significant differences were detected in transcripts involved in 'Regulation of immune response', 'Lymphocyte Activation' and 'T cell activation'. Red and blue boxes colors depict relative over- and under-expression in AA relative to EA. The range of colors is between -115.3-fold and +115.3-fold and preserves qualitative relationships among individual values. All fold changes outside of this range have been truncated to ±115.3. Only transcripts found significant at the level $q < 0.1$ in the comparison, are shown (1).

Network analysis of mahi mahi RNA seq data



Network analysis. The Deepwater Horizon (DWH) oil spill contaminated the spawning habitats for numerous species of pelagic fish. Exposure to the water accommodated fraction (WAF) of oil from the spill has been shown to cause cardiac toxicity during early life stages across fish species. Network analysis of high-throughput RNA sequencing data from mahi mahi using Qiagen IPA predicted mechanisms by which WAF leads to hypertrophy of cardiomyocytes; degeneration of the eyes; cell viability; neurodegeneration of central nervous system & cardiogenesis and abnormal vertebral column (2).

Variant analyses with human cancer exome seq data



Lollipop-style mutation diagram to highlight genetic variations in the *TSC2* gene. Data were subjected to Illumina quality control (QC) procedures (>80% of the data yielded a Phred score of 30). Secondary analysis was carried out on the OnRamp Genomics Research Platform (OnRamp Bioinformatics, San Diego, CA).

An automated Exome seq workflow processed the data, including (1) data validation and quality control, (2) read alignment to the human genome (hg19) using bwa, (3) generation of VCFs with the genome analysis tool kit (GATK).

Bibliography

1. Hardiman G, Hazard SE, Savage SJ, Wilson RC, Smith MT, Hollis BW, Halbert CH Gattoni-Celli S. Systems analysis of transcriptomic differences in prostate tissue specimens between African American and White men. (In press)
2. Xu EGB, Mager E, Grosell M, Pasparakis C, Schlenker L, Stielitz J, Bennetti D, Hazard ES, Courtney S, Diamante G, Hardiman G, Schlenk D. Transcriptomic responses to Deepwater Horizon oil in Mahi-Mahi (Coryhaena hippurus) embryos reveals time and oil-dependent linkages between molecular initiating events and developmental toxicity. (Submitted)