BIOMEDICAL INFORMATICS

Using BMI Data Warehouse and High Performance Cluster for Genomics Analysis

Nathan Salomonis, PhD
Assistant Professor, Biomedical Informatics, CCHMC
http://www.altanalyze.org

UNIVERSITY OF Cincinnati | Cincinnati Children's

Why we need the HPC?

- Process thousands of RNA-Seq samples processed per year (single-cell and bulk).
- Complex software setup and administered.
- Significant computational requirements for deeper analyses.
- Reduce analysis time from months to days.
- Human data must be secure and compliant.
- Data must be backed up and stored remotely.

Tools we frequently use on the HPC:

- RSEM
- AltAnalyze
- TopHat
- Cufflinks
- Homer
- GATK
- BEDTools
- Trinity
- samtools
- FASTQC
- R
- km
- Jellyfish
- vcftools
- Picardtools
- kallisto
- sailfish
- Miso
- DEXSeq
- rMATS

Example Project: Human AML

- Goal: Discover mutations in AML that impact splicing, tumorigenic and survival.
- Hundred of deeply sequenced RNA blood samples (n>800), microRNA-Seq (n>150), methylation arrays (n>150) from public repositories (GEO, CGHub, TARGET).
- Most samples have no mutational profile provided.

GEO Gene Expression Omnibus

Cancer Genomics Hub

TARGET

University of Cincinnati

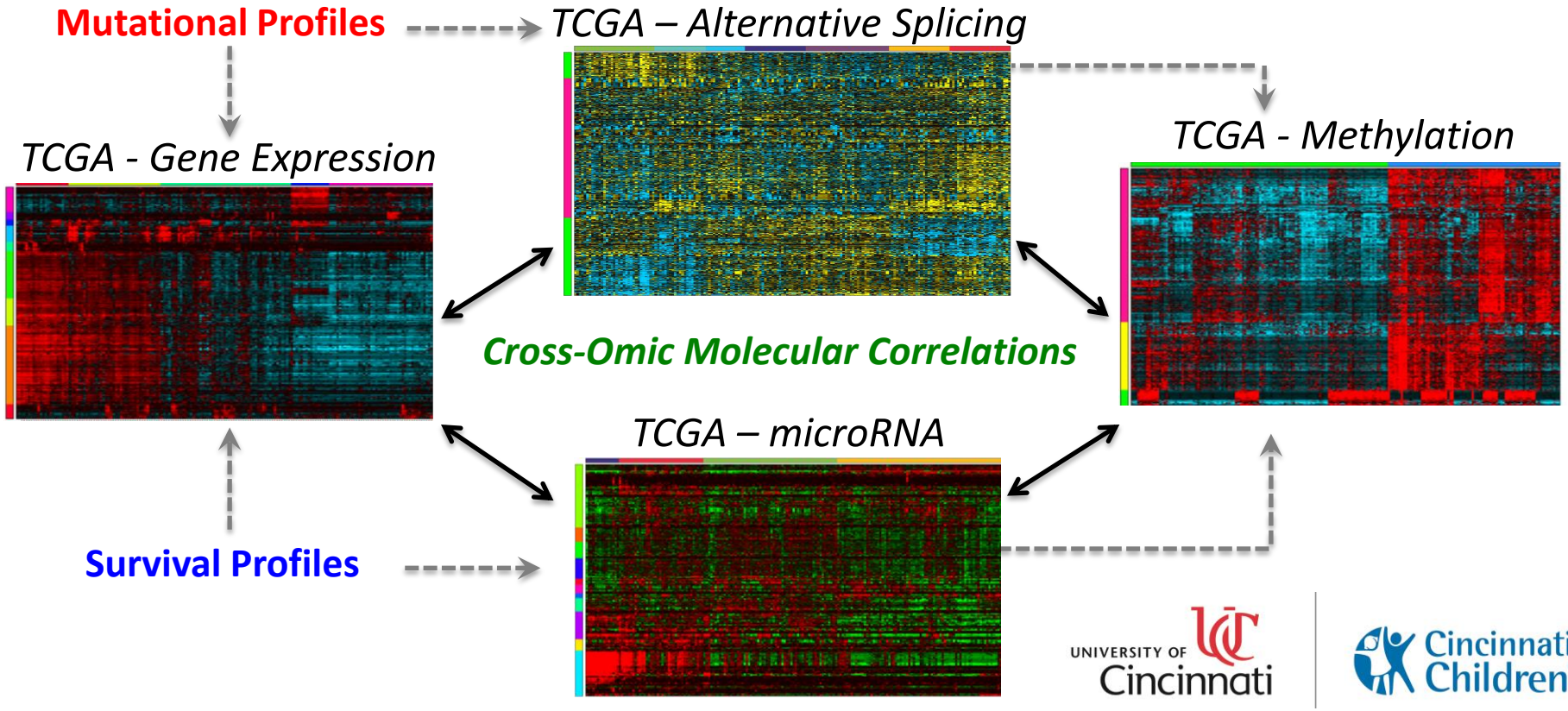
Cincinnati Children's

HPC is REQUIRED for Analysis!!

- Only sequence provided for most samples. TCGA missing novel isoforms and poor alignment.
- 800 RNA-Seq samples = ~**400** days of compute time on a single high-end machine (16GB RAM). Possible in **10** days on the HPC (40 parallel jobs).
- Requires ~**25** TB of hard disk space, fast access with back-up. Data must be secure (genotypes).
- Combined analyses require **128** GB of RAM and a dozen CPUs (1 machine).
- Complex software required.

Novel Integrative Research Opportunities

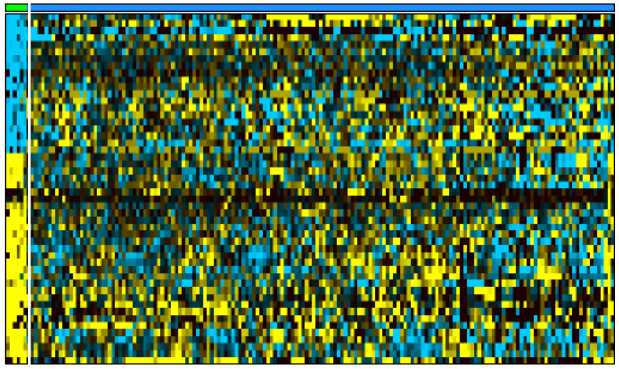
- Integrative models of gene expression, splicing, microRNA, mutations, methylation and prognosis.



Novel Integrative Research Opportunities

- Associate splicing signatures from TCGA to TARGET and uncharacterized AMLs (Leucegene) to find mutations.

TCGA Splicing Signatures:
U2AF1 Mutant Patients

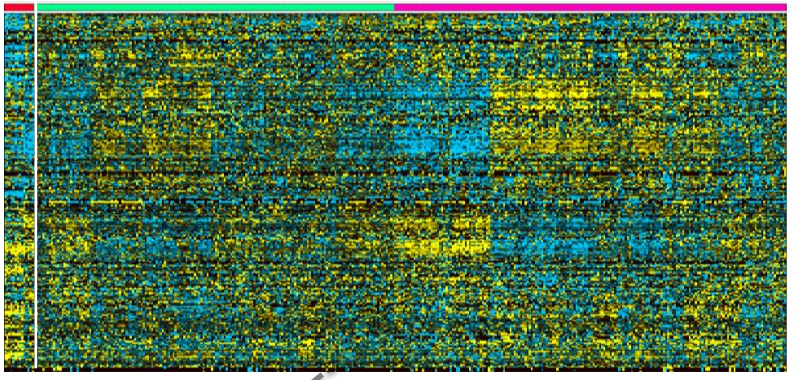


U2AF1 Mutant Patients

Machine Learning Supervised Classification of Mutational Profiles

U2AF1 Predicted

Leucegene Matching Splice Signature



Visualization and Validation of Predicted Patient Mutations

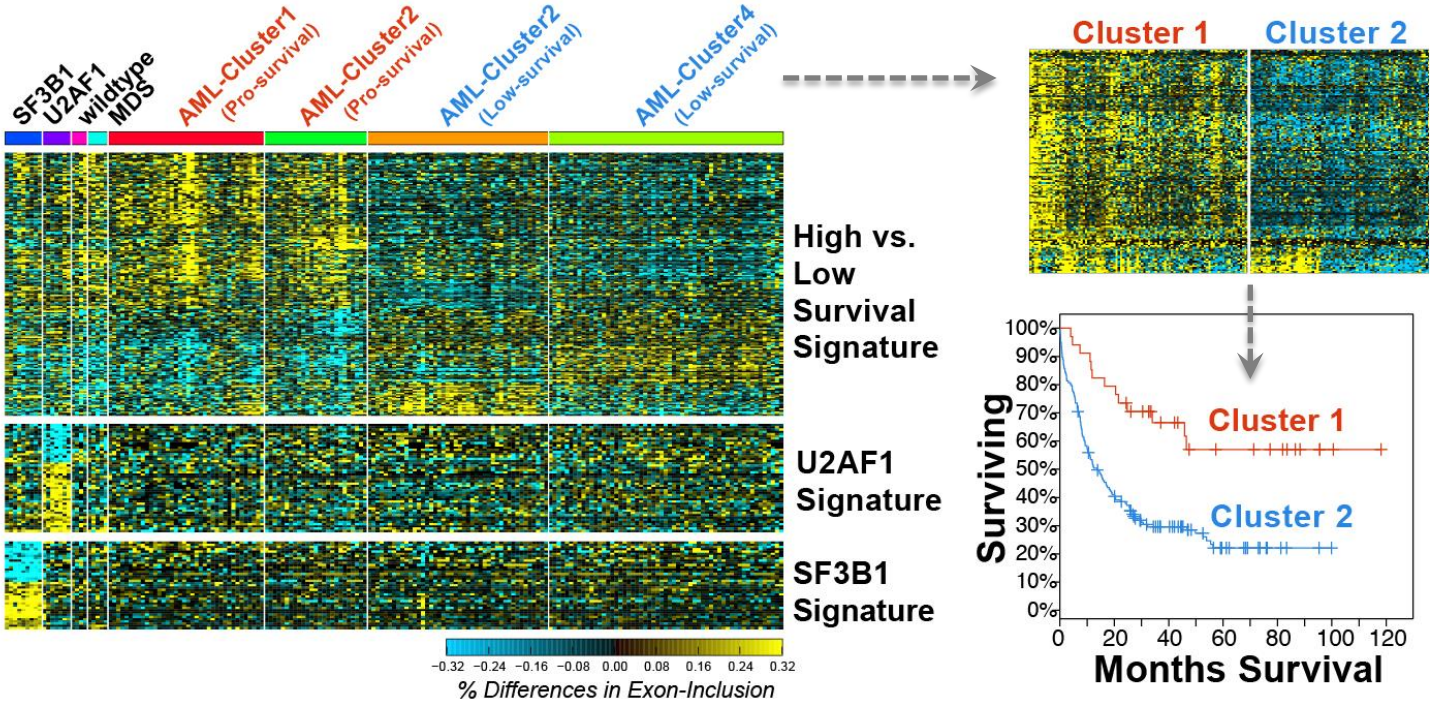
200GB
Data Visualized

UNIVERSITY OF Cincinnati

Cincinnati Children's

Novel Integrative Research Opportunities

- Find de novo splicing signatures and associate with prognosis.



UNIVERSITY OF Cincinnati

Cincinnati Children's

BIOMEDICAL INFORMATICS

Genomic Analysis and High Performance Computing

Prakash Velayutham, MS

Team Lead, Linux and High Performance Computing

Biomedical Informatics

Kevin Sandy

Sr. Systems Analyst

Biomedical Informatics

UNIVERSITY OF Cincinnati

Cincinnati Children's

Agenda

- Introduction to HPC
- HPC Infrastructure overview
- Applications and customers
- Genomics using HPC
- Workflow tools
- Q & A

HPC and Linux Team

Kevin

Carmen

Jason

Mark

Introduction to HPC

- Why use it?
- How to get access?

Why use the HPC?

Local machines generally have limited resources

- Processors
- Memory
- Storage
- Time

Why use the HPC?

Focus on what you need to accomplish

- No need to compile software and dependencies
- Approximately 400 software packages / versions currently available

Why use the HPC?

Scale out

- With MPI, jobs can run on multiple nodes simultaneously
- With job dependencies, independent steps can be run simultaneously

How to use the HPC?

- Email help@bmi.cchmc.org to have your account setup
- Access can be via:
  - NoMachine (NX)
  - Citrix (in progress)
  - SSH
- Data volumes can be mounted to your Windows or Mac computer for easy access

HPC Infrastructure at CCHMC

- We have 3 different HPC environments
  - Clinical Exome
    - Restricted access
  - Research production and development HPC
    - Available to
      - **all CCHMC personnel**
      - **UC and other external collaborators**

HPC Infrastructure at CCHMC

- Clinical Exome
  - 96 cores
  - 10G ethernet
  - Inside CCHMC network
  - Strictly for clinical purposes
  - CLIA/CAP compliant

HPC Infrastructure at CCHMC

- Production research HPC
  - Currently at ~700 cores
  - Mostly HP blades
  - Cores range from 4 – 16 per node
  - RAM ranges from 8G – 256G per node
  - 2 - Tesla (K10) compute nodes for GPU computing
  - 10G ethernet
  - Direct connection to Isilon high performance storage cluster

HPC Infrastructure at CCHMC

- Development research HPC
  - Currently at ~600 cores
  - Older HP blades
  - 4 or 8 cores per node
  - 1G ethernet

HPC Infrastructure at CCHMC

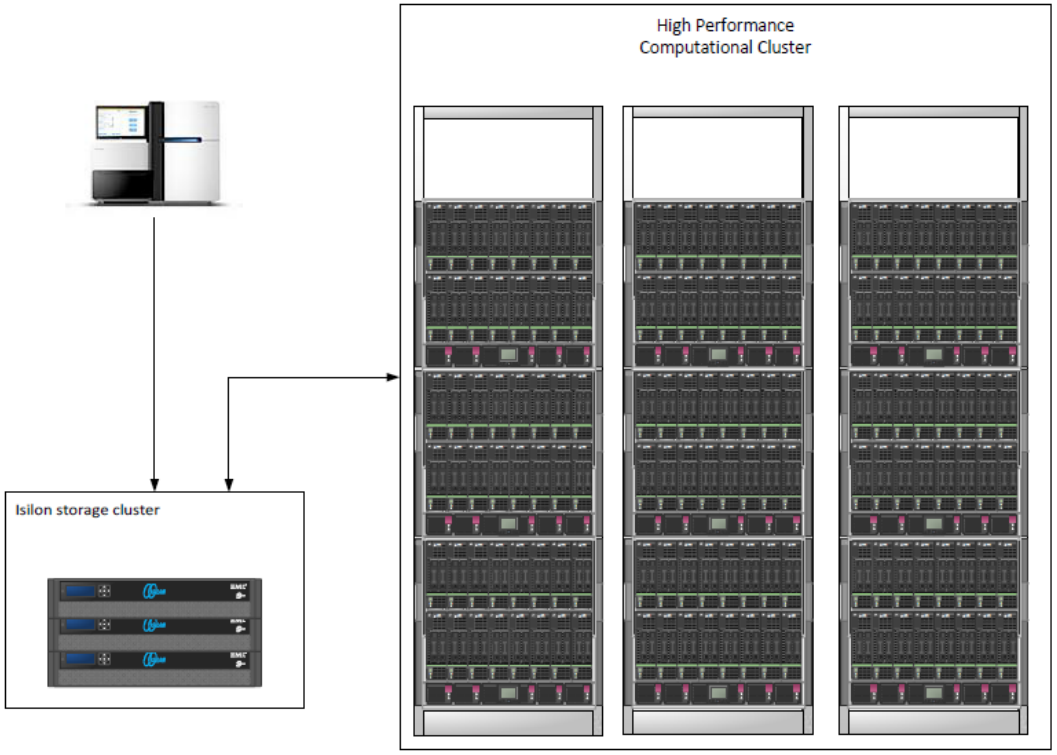
	2014	2015	Increase
Total jobs	912268	1608997	76%
Total job hours	724215	1181501	63%
Jobs / hour	104	183	76%
Average job time	~47 minutes	~44 minutes	

- Users: > 50
- Applications: > 300

Some of the Research Areas

- Genomics
- Metagenomics
- Protein docking, folding and structure prediction
- Natural language processing
- Functional neuroimaging
- Molecular dynamics
- Pharmacodynamics
- Large scale rendering

Genomic Applications using HPC



Genomic Applications using HPC

- Output data from the sequencers are stored in the Isilon storage cluster.
- Offline Base Caller – Base calling and QSEQ formatted output.

Genomic Applications under HPC

- Demultiplexing and “bcl to fastq” conversion is done using home-grown scripts.
- Further downstream analysis conducted by individual researchers per their needs.
- FASTQ files are available for users to download to run through their own analysis process.

Common Genomics Software Used

- BWA, Bowtie – Sequence alignment
- Affy Power Tools – To analyze and work with Affymetrix GeneChip® arrays
- bamtools – Tools to work with BAM and SAM files
- bedtools, plink – Genomic analysis tools
- R/Bioconductor

Common Genomics Software Used

- Kallisto, RSEM, sailfish, Trinity - RNA-Seq
- Mothur, QIIME, LEfSe, MetaPhlAn, PhyloPhlAn – Metagenomics
- MACS – ChIP-Seq
- miRanda, miRDeep2 – miRNA experiments
- vcftools

Workflow Tools

- Workflow tools let you create a pipeline.
- Connects to a cluster in the backend.
- Determines and manages job dependencies automatically.
- Either a thick client or web-based.

Workflow Tools

- LONI – Java-based thick client.
- Galaxy – Web-based workflow software. We have a local instance.
- GenePattern – Broad institute – Broad user community.
- AltAnalyze – Command-line and GUI available.

Other Research Tools

- Linux/HPC Team also manages the following tools.
  - Strand NGS
  - SAS
  - Genome Browser (https://gb.research.cchmc.org)
  - Mascot (https://research.cchmc.org/mascot/)