The Impact of HPC and Data-Centric Computing in Cancer Research

Jack R. Collins, Ph.D.
Information Systems Program
Frederick National Laboratory for Cancer Research

July 5, 2012
HPC User Forum, London
Acknowledgements

• ISP Staff who did the work
• NCSA for computer time
• HPC User Forum
• Listeners
Practical Matters
What is “Big Data”? "Big Datasets"
- LHC experiment
- Many-dimensional time-series
- Large Grid or High-resolution Volumetric

"Lots of Data"
- Billions of photos or YouTube videos
- Tens of thousands of genomes

Computational needs can be very different but may change during the course of the analysis. If the data is “really big”, it will be impractical to move the data to the best computing platform. The computing infrastructure must adapt to the data analysis need.
Practical Matters
Why is “Big Data” different?

“Big Data” Concerns
- Storage is critical issue
- Security and Data Integrity are essential
- Easy Access to Data and Analysis
- Distributed Storage Environment
- Persistence, Versioning, and Retiring Data
- Metadata Repository
- Ownership and Responsibility (Stewardship)
- Business/Cost Model
New Reality?

Is HPC becoming commodity?
Does this change the message?

• What is HPC, today?
• HPC is increasingly becoming an appliance and part of the infrastructure required to analyze data, make decisions, and impact medical science on a daily and almost unnoticeable way.
Observations

- Traditional HPC experts often have a narrow view of the new applied user world.
- HPC should be “baked-in” to workflows to optimize “system/operational/business efficiency” of the process (that includes **humans** and software).
- Computing hardware (HPC) is cheap enough and increasingly indispensible to processes to be considered infrastructure.
**Observations**

- People and software drive innovation. HPC should be regarded as a *de facto* tool.
- “Cloud computing” will become portals to ubiquitous HPC infrastructure.
- HPC is about timely information and enabling “insight”. This will drive “systems biology”, “personalized medicine”, “nanoInformatics”, “Cancer Simulations”
- Need to re-examine the business model of HPC based on cost and impact.
In an era of stagnant funding, comparative analyses of methods and tool performance can help researchers do more with less.

“Analyze this” *Nature Methods* 8, 361 (May, 2011) Editorial

“One way of improving overall research efficiency is by analyzing and optimizing research methods and tools. Although published methods all presumably work at some level, **many have not been fully optimized**. A researcher developing a method only needs to optimize it sufficiently for his or her own use, and there is typically little incentive to go further. Although we strive to avoid this in the work we publish, many methods remain under-characterized and under-optimized.

Unfortunately, this situation can lead to gross inefficiencies when methods and tools are widely adopted. **It is all too common for researchers to waste considerable effort trying without success to implement an under-developed method or to use the wrong tool—or the right one in the wrong way—owing to insufficient performance data.**”
The dominant factor is data management and analysis.

### Integrated Services (IT + Science)

- **Compositional Analysis as a Service (CAaaS)**
- **Software as a Service (SaaS)**
- **Platform as a Service (PaaS)**
- **Infrastructure as a Service (IaaS)**
  - Compute
  - Storage
  - Networking
- **Software Kernel (OS, VMM)**
- **Firmware, HW**

### Core Services (Exp)

- **Mgmt**
Workflow Automation

Build HPC into modules

ChIPSeq

1. Download NCB SRA FASTQ Files
2. FASTQ Directory Reader

BWA Short Read Mapper
Add Mapped Reads to Repository

RNASeq

1. FASTQ Directory Reader
2. TopHat Spliced Read Mapper
Add Mapped Reads to Repository

Base Space Reads? Unzip Use Last Group No-oz Don't Mix Paired and Unpaired Run TopHat BAF File Names Output?

Save Reads to FASTQ Files

Individual Reads
Cost per Genome

Amdahl’s Law

Gustafson’s Law

454 Genomics
Solexa/Illumina GA SOLiD
Complete Genomics
HiSeq2000 PacBio Ion Torrent

National Human Genome Research Institute

genome.gov/sequencingcosts
Cancer Focus

- In 2009, one person was expected to die from cancer every 56 seconds in the United States.
- In 2011, this number was 55 seconds.
HPC can ENABLE Diagnosis ➔ Treatment

- "Time to solution" becomes critical when treatment decisions are based on technologies such as sequencing of tumor and patient.
- Must be "baked into" the decision support system workflow
- Simulations will be come part of the decision process within the next 10-15 years
Prototypical Oncology (Development) Workflow (coming soon now)

• Characterize tumor(s)
• Sequence tumor
• Target mutations
• Characterize at the molecular level
• Integrate all of the data
• Design and develop putative drug
• Test in animal models or clinical trials
• Treat patients
Characterize Tumor Images Analyzed

- Aperio Registration
- Quick2Insight
- Interface for SAIP met Segmentation
- Tumor Segmentation
- VDI

Common Imaging Tool Development for SAIP

Automatic Visualization on 3D Biological Datasets

High-Res (up to 70k x 70k) Aperio Image Registration

New Image Segmentation Module in Open Source Imaging Software 3D Slicer
Characterize Tumor
Pathologist Annotates Image
Characterize Tumor Images Registered
Optimizing People and Workflow
HPC Problem: Processing and Analysis

- 300 Aperio Images
  - Up to 70k by 70k pixel resolution
  - 5~20 GB for each uncompressed image
- Insight Tool Kit Multi-resolution Image Registration Pipeline
- Sample statistics
  - 48 Cores 130GB memory 11~14 hours per image for registration based on full resolution images
- Desktop to view and compare images now requires high-performance GPUs and 32GB of memory.
Characterize Tumor at Genomic Level

- SF - ATC
- CGF - ATC
- Lab2- 37
- Lab1- ATC
- LMT
- Lab3- 10
- TARGET
- ICGSC
- 1000 Genomes
- St Jude/WashU
- TCGA

Platforms:
- 454
- SOLiD
- GAIlx
- HiSeq2k
- PacBio
- IonTorrent
- MiSeq
NGS Analysis

- Dataset sizes are growing
- Time to generate the data is decreasing
- Analyses becoming more complex
- Data analyzed multiple times
Analysis of NGS Data
(blades required to keep up with basic analysis of data)
(mapping and initial variation detection)
Targeting the Mutation
NCSA Resources for molecular elucidation in Biomarker/drug investigation

<table>
<thead>
<tr>
<th>System</th>
<th>Local days using 16 cpus</th>
<th>ns/day</th>
<th>NCSA days using 96 cpus</th>
<th>ns/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>195000 atom 22ns</td>
<td>78.7</td>
<td>0.3</td>
<td>17.7</td>
<td>2.26</td>
</tr>
<tr>
<td>150000 atom 30ns</td>
<td>78.5</td>
<td>0.4</td>
<td>22.8</td>
<td>2.64</td>
</tr>
</tbody>
</table>
Modeling the Mechanism

Oxidative Damage (IP)

Direct Photodamage (CPD)

Methylation (CpG)

Mutation $G \rightarrow T$

Mutation $C \rightarrow T$

Deamination
Calculate Contribution of Mechanism Templates to Cancer Landscapes

For each cancer:

\[(IP \times W_{ip}) + (CPD \times W_{cpd}) + (CpG \times W_{cpg}) + (Residual \times W_r) = \text{Cancer Landscape}\]

- Template weights are optimized based on an exhaustive search
## Template Contributions

### Calculated Contributions

<table>
<thead>
<tr>
<th>Sample</th>
<th>CPD</th>
<th>IP</th>
<th>CpG</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>gastric_icgc</td>
<td>18.68%</td>
<td>33.15%</td>
<td>8.44%</td>
<td>39.73%</td>
</tr>
<tr>
<td>gastric_ews</td>
<td>19.14%</td>
<td>28.71%</td>
<td>15.95%</td>
<td>36.20%</td>
</tr>
<tr>
<td>lung_sc</td>
<td>9.16%</td>
<td>59.91%</td>
<td>1.41%</td>
<td>29.52%</td>
</tr>
<tr>
<td>lung_nsc</td>
<td>1.40%</td>
<td>68.03%</td>
<td>0.70%</td>
<td>29.86%</td>
</tr>
<tr>
<td>pancreatic_ca</td>
<td>13.02%</td>
<td>44.64%</td>
<td>4.34%</td>
<td>38.00%</td>
</tr>
<tr>
<td>pancreatic_au</td>
<td>12.78%</td>
<td>42.59%</td>
<td>5.48%</td>
<td>39.16%</td>
</tr>
<tr>
<td>liver_riken</td>
<td>21.45%</td>
<td>36.94%</td>
<td>1.19%</td>
<td>40.41%</td>
</tr>
<tr>
<td>liver_ncc</td>
<td>16.94%</td>
<td>43.28%</td>
<td>2.51%</td>
<td>37.27%</td>
</tr>
<tr>
<td>melanoma</td>
<td>27.89%</td>
<td>22.63%</td>
<td>2.11%</td>
<td>47.37%</td>
</tr>
<tr>
<td>melanoma_ews</td>
<td>33.86%</td>
<td>13.44%</td>
<td>6.45%</td>
<td>46.25%</td>
</tr>
<tr>
<td>dlbcl</td>
<td>13.13%</td>
<td>43.31%</td>
<td>9.19%</td>
<td>34.37%</td>
</tr>
<tr>
<td>mm_ews</td>
<td>13.76%</td>
<td>40.66%</td>
<td>8.13%</td>
<td>37.45%</td>
</tr>
<tr>
<td>hgmd</td>
<td>13.58%</td>
<td>40.73%</td>
<td>7.41%</td>
<td>38.28%</td>
</tr>
<tr>
<td>1Kgenomes</td>
<td>17.06%</td>
<td>37.16%</td>
<td>6.70%</td>
<td>39.08%</td>
</tr>
<tr>
<td>glioblastoma</td>
<td>14.79%</td>
<td>50.71%</td>
<td>4.93%</td>
<td>29.57%</td>
</tr>
<tr>
<td>cll</td>
<td>11.15%</td>
<td>41.83%</td>
<td>2.79%</td>
<td>44.23%</td>
</tr>
<tr>
<td>hnscc</td>
<td>9.31%</td>
<td>45.32%</td>
<td>7.45%</td>
<td>37.92%</td>
</tr>
<tr>
<td>prostate</td>
<td>11.87%</td>
<td>45.62%</td>
<td>5.00%</td>
<td>37.51%</td>
</tr>
<tr>
<td>ovarian</td>
<td>9.09%</td>
<td>47.24%</td>
<td>4.24%</td>
<td>39.43%</td>
</tr>
<tr>
<td>mean</td>
<td>9.09%</td>
<td>47.24%</td>
<td>4.24%</td>
<td>37.98%</td>
</tr>
</tbody>
</table>
Cluster on Contributions

Cluster Dendrogram

- lung_sc
- lung_nsc
- melanoma
- melanoma_ews
- gastric_ews
- glioblastoma
- liver_riken
- gastric_logc
- 1Kgenomes
- liver_ncc
- dbcl
- mm_ews
- hgd
- cl
- pancreatic_au
- pancreatic_ca
- prostate
- hnscc
- ovarian

Distance
hclust (*, "complete")
Template Contributions by Patient

- Individual liver tumors show very similar patterns of template contributions.

<table>
<thead>
<tr>
<th></th>
<th>liver_riken</th>
<th>liver_ncc</th>
</tr>
</thead>
<tbody>
<tr>
<td>liver_riken</td>
<td></td>
<td></td>
</tr>
<tr>
<td>liver_riken_RK001</td>
<td>24.24% 33.11% 1.77% 40.87%</td>
<td>18.91% 40.33% 3.78% 36.98%</td>
</tr>
<tr>
<td>liver_riken_RK002</td>
<td>24.99% 32.72% 1.78% 40.50%</td>
<td>18.25% 40.90% 3.78% 37.07%</td>
</tr>
<tr>
<td>liver_riken_RK003</td>
<td>23.89% 34.05% 1.79% 40.27%</td>
<td>18.21% 40.82% 3.77% 37.21%</td>
</tr>
<tr>
<td>liver_riken_RK006</td>
<td>22.90% 35.55% 1.81% 39.75%</td>
<td>18.90% 42.20% 1.89% 37.02%</td>
</tr>
<tr>
<td>liver_riken_RK010</td>
<td>22.88% 35.53% 1.81% 39.79%</td>
<td>18.35% 43.03% 1.90% 36.72%</td>
</tr>
<tr>
<td>liver_riken_RK015</td>
<td>22.73% 35.28% 1.79% 40.20%</td>
<td>18.99% 44.54% 1.96% 34.50%</td>
</tr>
<tr>
<td>liver_riken_RK023</td>
<td>21.69% 37.35% 1.20% 39.75%</td>
<td>19.21% 39.07% 5.76% 35.96%</td>
</tr>
<tr>
<td>liver_riken_RK024</td>
<td>21.70% 36.77% 1.81% 39.72%</td>
<td>19.21% 39.69% 5.12% 35.98%</td>
</tr>
<tr>
<td>liver_riken_RK026</td>
<td>21.17% 37.51% 1.81% 39.51%</td>
<td>18.59% 41.02% 4.49% 35.91%</td>
</tr>
<tr>
<td>liver_riken_RK029</td>
<td>21.14% 37.44% 1.81% 39.61%</td>
<td>19.19% 40.30% 4.48% 36.04%</td>
</tr>
<tr>
<td>liver_riken_RK034</td>
<td>21.21% 37.57% 1.82% 39.41%</td>
<td>18.32% 41.07% 3.79% 36.82%</td>
</tr>
</tbody>
</table>
# Assessing the Impact of Mutations

**SIFT:**  
http://sift.jcvi.org/

**PolyPhen-2:**  
http://genetics.bwh.harvard.edu/pph2/

**BEN:** Benign;  
**PRD:** Probably-Damaging(high-confidence);  
**POD:** Possibly-Damaging

<table>
<thead>
<tr>
<th>Gene</th>
<th>NCBI ID</th>
<th>Phe-2 Pred</th>
<th>Phe-2 Score</th>
<th>SIFT Prediction</th>
<th>Score(median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRAF(D454E)</td>
<td>NP_004324.2</td>
<td>BEN</td>
<td>0.000</td>
<td>0.000 (sensitivity: 1.00; specificity: 0.00)</td>
<td>TOLERATED 1.00(3.18)</td>
</tr>
<tr>
<td>NCF2(H389Q)</td>
<td>NP_00424.2</td>
<td>PRD</td>
<td>1.000</td>
<td>0.992 (sensitivity: 0.44; specificity: 0.97)</td>
<td>TOLERATED 0.39(3.14)</td>
</tr>
<tr>
<td>G6PC3(Q272R)</td>
<td>NP_612396.1</td>
<td>BEN</td>
<td>0.005</td>
<td>0.005 (sensitivity: 0.97; specificity: 0.41)</td>
<td>TOLERATED 0.33(2.98)</td>
</tr>
<tr>
<td>KRAS(136T)</td>
<td>NP_004976.2</td>
<td>BEN</td>
<td>0.009</td>
<td>0.044 (sensitivity: 0.94; specificity: 0.59)</td>
<td>TOLERATED 0.21(3.22)</td>
</tr>
<tr>
<td>PIK3CA(R524K)</td>
<td>NP_006209.2</td>
<td>BEN</td>
<td>0.000</td>
<td>0.002 (sensitivity: 0.99; specificity: 0.17)</td>
<td>TOLERATED 1.00(3.33)</td>
</tr>
<tr>
<td>SMAD4(E538K)</td>
<td>NP_005350.1</td>
<td>BEN</td>
<td>0.025</td>
<td>0.024 (sensitivity: 0.95; specificity: 0.54)</td>
<td>TOLERATED 0.21(3.22)</td>
</tr>
<tr>
<td>TP53(W23R)</td>
<td>NP_000537.3</td>
<td>POD</td>
<td>0.999</td>
<td>0.997 (sensitivity: 0.24; specificity: 0.99)</td>
<td>DAMAGING 0.02(3.04)</td>
</tr>
<tr>
<td>CYB(B8112P)</td>
<td>NP_000388.2</td>
<td>BEN</td>
<td>0.022</td>
<td>0.116 (sensitivity: 0.91; specificity: 0.67)</td>
<td>DAMAGING 0.00(3.12)</td>
</tr>
<tr>
<td>CBYA(Y72H)</td>
<td>NP_0000092.2</td>
<td>PRD</td>
<td>1.000</td>
<td>0.996 (sensitivity: 0.32; specificity: 0.98)</td>
<td>TOLERATED 0.41(3.41)</td>
</tr>
<tr>
<td>CDKN2A(R87P)</td>
<td>NP_000068.1</td>
<td>PRD</td>
<td>1.000</td>
<td>0.999 (sensitivity: 0.08; specificity: 1.00)</td>
<td>TOLERATED 0.22(1.80)</td>
</tr>
</tbody>
</table>

**HumDiv**, was compiled from all damaging alleles with known effects on the molecular function causing human Mendelian diseases, present in the UniProtKB database, together with differences between human proteins and their closely related mammalian homologs, assumed to be non-damaging.

**HumVar**, consisted of all human disease-causing mutations from UniProtKB, together with common human nsSNPs (MAF>1%) without annotated involvement in disease, which were treated as non-damaging.
Mine and Integrate Multiple Data Sources
“Systems Biology”

- Big Graph Problem
- Semantics and Ontologies (working with Oracle and others)

Integrated Metadata

- Clinical Metadata
  - LIMS PacBio
- Sequence Metadata
  - LIMS Illumina
- Proteomic Metadata
  - LIMS 454
- Imaging Metadata
  - LIMS IonTor
- Imaging Metadata
  - LIMS SOLiD
bioDBnet – Biological Data Base Network

- Integrates 28 widely used biological databases
- Current version has 170 nodes and 612 edges
- Tools for conversions within and between different types of biological identifiers
- Generates extensive annotation reports within a few seconds

URL: http://biodbnet.abcc.ncifcrf.gov

Characterize Therapies at Molecular Level

JS-K AND RELATED O²-ARYLATED DIAZENIUMDIOLATES AS BROAD-SPECTRUM ANTI-CANCER AGENTS

NO-releasing prodrugs of the O²-arylated diazeniumdiolate class have shown themselves to be increasingly promising broad-spectrum anti-cancer drug candidates. Our lead compound JS-K has slowed tumor growth in several rodent models of cancer, including leukemia, prostate cancer, multiple myeloma, and liver cancer. Its second-generation analog, PABA/NO, acted with a potency similar to that of cisplatin in an in vivo model of ovarian cancer. JS-K has proven active in blocking angiogenesis in vitro and in vivo, inhibiting tumor cell invasiveness, and synergizing with cytarabine, bortezomib, arsenite, and cisplatin. So far, lead compounds JS-K and PABA/NO have shown the ability to significantly slow tumor growth in vivo with no evidence of toxicity being observed at therapeutic doses.
So, what are we missing?

- Genomics
- Proteomics
- Epigenome
- RNA Expression
- Microenvironment
- Transcription Factors
- ncRNA
- Histone Marks
- Microbiome
- Genotype/Phenotype
- Survival Data
- Animal Models
- Environmental Factors
- Cell-Cell Communication
- Pathology
- Drug Studies
Need function of interacting systems

- Proteomics
  - Identification & Function
  - PTM
  - Metabolites
  - Chromatin Structure
  - PPI
- Genomics
  - RNA
    - Expression
    - Splicing
    - Non-Coding
    - Interfering
  - Epigenetics
    - DNA Variation
      - SNVs
      - Indels
      - CNVs
  - Regulatory Elements
    - TFs
- Protein Mutations
- Histone Modification
- Protein Levels
- Chromosome Packing

SAIC-Frederick, Inc.
A subsidiary of Science Applications International Corporation
To understand cancer, we need to simulate a tumor ➔ A “Virtual Tumor”

• Tumors are Heterogeneous
  – Cell population distribution
• Cell-Cell Communication is Important
• Intracellular Communication is Important
• Cellular processes are stochastic
  – 23 pairs of chromosomes (not moles)
• This leads to a high-dimensional dynamic, stochastic, complex system to simulate
One approach along with lots of data

A probabilistic framework of interactions to model the system

An ensemble of systems to reflect the heterogeneity of tumors

Tumor properties derived from ensemble of simulations

Fit to observed experimental data

Refine model system and develop hypotheses
Is this an HPC problem?

- Needs lots of compute in short amount of time using lots of data
- Not necessarily FP intensive
- High human interactivity so quick turnaround
- Computation of ensembles of systems
What would my HPC computer look like?

- Lots of memory bandwidth.
- Many lookups, compares, and branches per clock tick. Not just FP.
- Ingest data from LARGE databases (I/O)
- Scale as I need to reduce time to solution or grow model
- Scale FP as models evolve
- Software libraries that efficiently use the hardware
- Lots of capacity to run ensemble simulations in parallel so results can be aggregated to calculate distributions in timely manner
To be continued ... 

- Questions?
- Comments?
- Answers?