Changing Landscape for Biomedical Computation

HPC User Forum
Oxford, UK
September 29, 2016

Jack R Collins, Ph.D.
Advanced Biomedical Computing Center
Frederick National Lab for Cancer Research

Motivation
Government Initiatives

- Cloud First Policy (USA) (December 2010)
- Precision Medicine Initiative (January 2015)
- National Strategic Computing Initiative (July 2015)
- Cancer Moonshot (January 2016)

Motivation
Technology Considerations

- Advances in consumer electronics spur advances in experimental devices
  - Cameras and detectors
- Commoditization of Genomic Sequencing
  - Clinical, Microbiome, New Research App
- Nanotechnology
- ImmunoTherapy

- End of Free Performance (GHz)
  - GPGPU
  - Many-Core Processors
  - FPGA
- Cognitive Computing – Deep Learning

HPC and HPDA Workflow Convergence

- Alexander Szalay stated at SC15 that at Exascale everything becomes a Big Data problem.
  - The NCI-DOE collaboration has highlighted that many questions important to Cancer Biology are Grand Challenge problems requiring Peta to Exa-scale computation. (Initiated through the NSCI, PMI, and Cancer Moonshot)
- Cancer Biology still has much to learn before Clinical Oncology is Precise
  - Though much progress has been made, much more is needed
  - Therefore, collect and analyze as much data as possible to try to achieve a more holistic (Systems) view

Outline

- Computation for Precision Medicine
  - Genomics
  - Cryo-EM (Electron Microscopy)
  - Design for BioMaterials
  - Photo-Immunotherapy
- Using the Cloud
  - Data Transfer
  - Reproducibility
  - Information Security

Precision Medicine Workflows

"Doctors have always recognized that every patient is unique, and doctors have always tried to tailor their treatments as best they can to individuals. You can match a blood transfusion to a blood type — that was an important discovery. What if matching a cancer cure to our genetic code was just as easy, just as standard? What if figuring out the right dose of medicine was as simple as taking our temperature?" - President Obama, January 30, 2015

Sequencing and Mapping

Feature and Variant Detection and Annotation

Interpretation

HPC/HPDA is Integral to BioMedical Computing

- Precision Medicine requires high-turnaround computing
- Time to Solution Matters (Clinical Genome Workflows)
- Precision Medicine will also require knowledge beyond the genome
  - Function of Mutation (What causes the disease?)
  - Potential for Therapy (Prognosis for success? Right treatment?)
  - Reduce Toxicity (More specific targeting of tumor)
  - Environment (tumor, microbiome, exposure)

Computational landscape is becoming more heterogeneous and complex

- FPGAs
- GPGPUs
- Many-core processors
- Hierarchical memory and storage models
- Interconnect layers
- Cloud Resources

Precision Medicine
Feature and Variant Detection

- Streamline and expedite translation of actionable variants for a wide range of research and clinical applications
  - Tumor/Normal and Germline variant discovery
  - Data sets ranging from one to thousands of samples
  - Call SNPs, INDELs, and structural variants
  - Both mouse and human (and potentially other model organisms, as well)
  - Meet QC requirements

Precision Medicine Workflows

Optimizing the steps in workflow involves multiple computational platforms.

Sequencing & Mapping

FPGA

Annotation

Large-scale parallelization and algorithm optimization

Interpretation

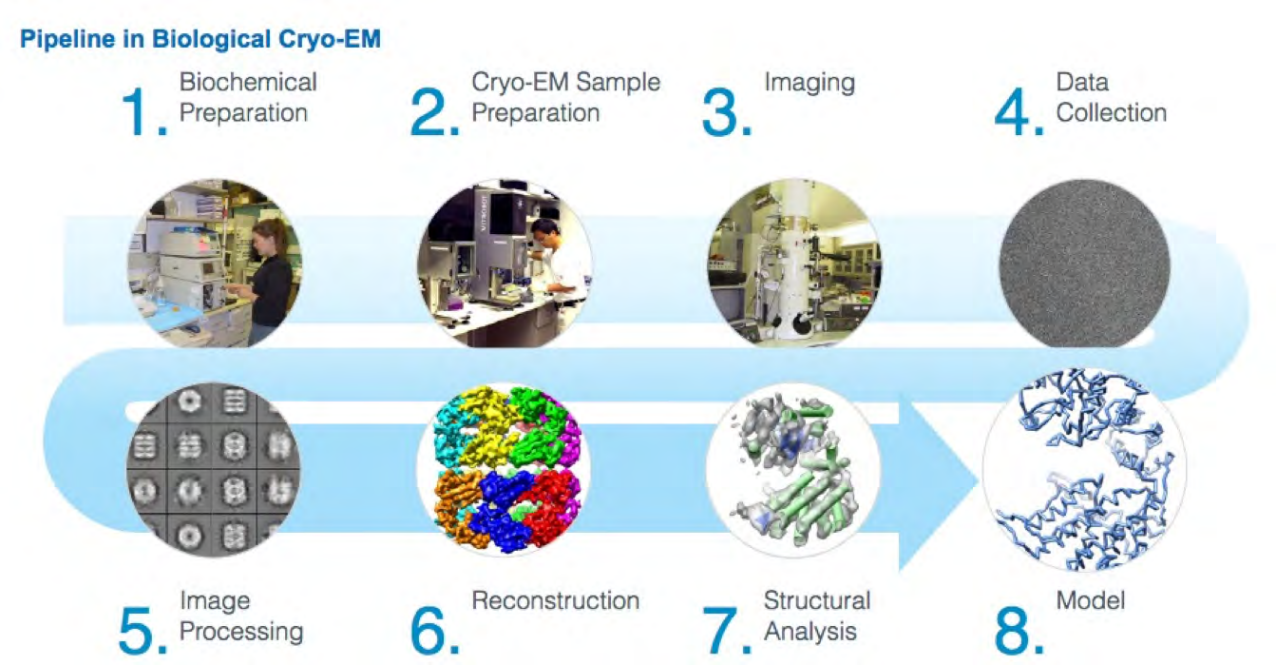
Machine Learning

GPGPU / KNL

Sequence Mapping Appliances
FPGA – Based (DRAGEN)

- The typical the analysis run time is around 24 hour per sample on cluster (48 cores)
- DRAGEN average run time is 3 minutes per exome when run singly, comparing 10 samples finished in 6 hours 25 minutes
  - Trio: 1hr 56mins
  - 10 samples: 6hr 25mins
  - 1000 samples (1 DRAGEN system): 31 days
  - 1000 samples (2 DRAGEN systems): 16 days
- Significant Advantages for Clinical (quick-turnaround) Analysis needed for Precision Medicine applications

Cryo-EM Workflows



Relion 1.4
TRPV1 Ion Channel

The size of the complete data set is close to **7 TB**.
35,645 particles picked from the raw EM images

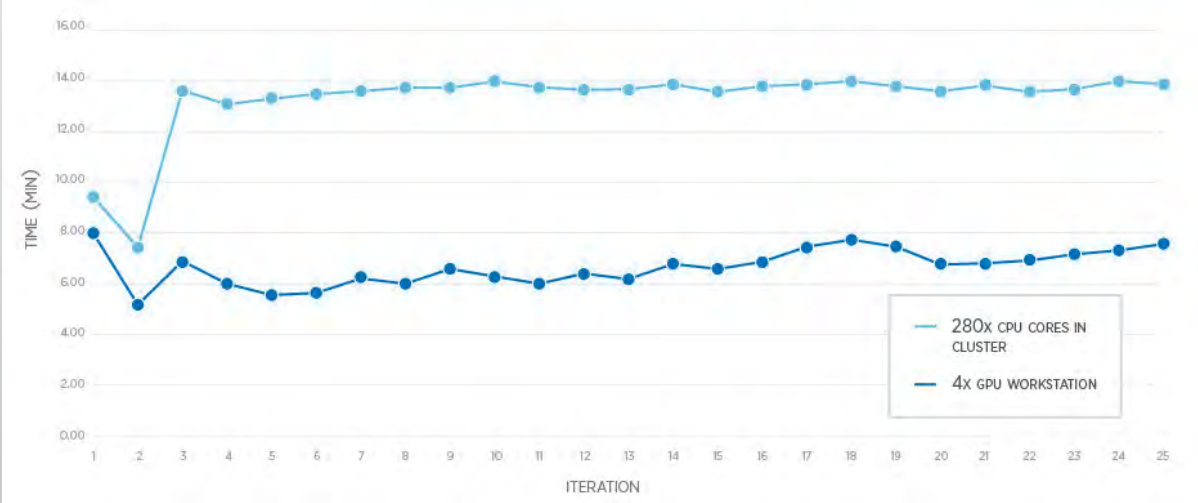
Cluster	Number of Nodes	Number of CPUs	Run Time (CPU hour)
FNLCR	4	64	500 – 800
Bridges	3	84	450 - 550

Image reconstruction
Deep Learning
Integration of Multiple Methods
SAXS/SANS

RELION 2 (beta)

Acknowledgement: Exxact corporation

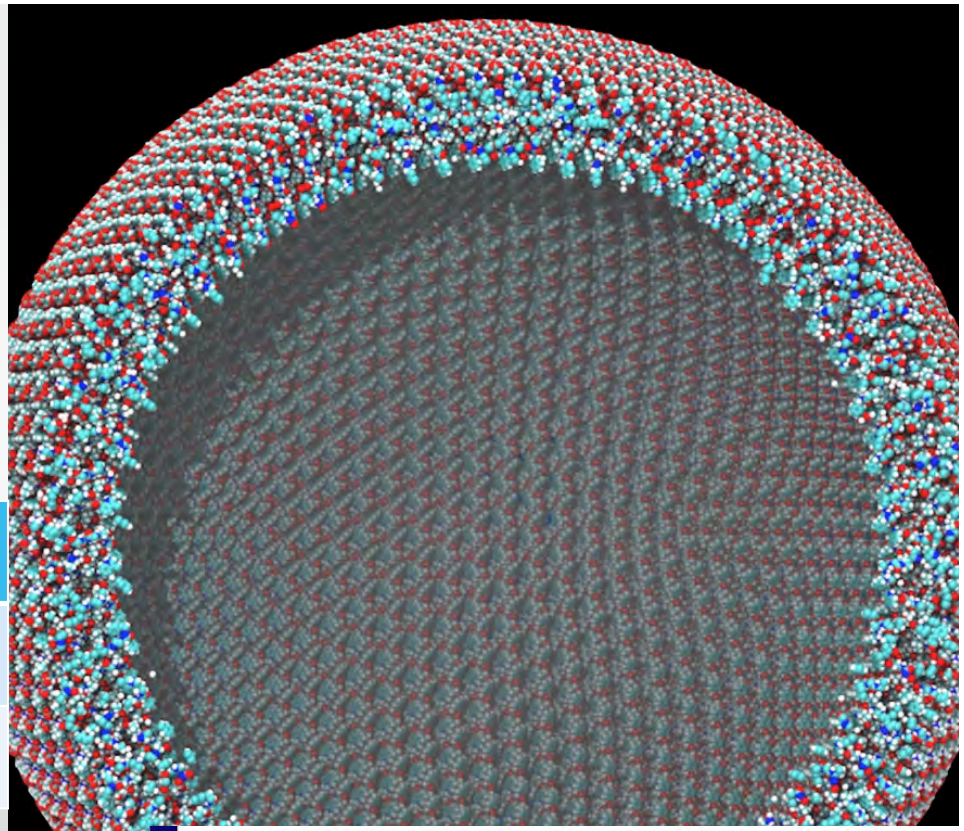
Relion Iteration GPU vs CPU Comparison



Molecular Design for Drug Delivery

1.3 Million Atoms	GROMACS 5.x	NAMD 2.11
1 GPU (8 Threads)	4 ns/day	2.2 ns/day
4 GPU (8 Threads)	11 ns/day	6 ns/day

Ubuntu 14.04, CUDA 7.5, 4Xeon(4) 3.6(3.8)GHz, 4TitanX



PNAS

Structural plasticity of a transmembrane peptide allows self-assembly into biologically active nanoparticles

Sergey G. Tarasov[a], Vadim Gaponenko[b], O. M. Zack Howard[c], Yuhong Chen[c], Joost J. Oppenheim[c], Marzena A. Dyba[a,d], Sriram Subramaniam[e], Youngshim Lee[b], Christopher Michejda[a,1], and Nadya I. Tarasova[c,2]

[a]Structural Biophysics Laboratory, National Cancer Institute, P.O. Box B, Frederick, MD 21702-1201; [b]Department of Biochemistry and Molecular Genetics, University of Illinois, 900 South Ashland, Chicago, IL 60607; [c]Cancer and Inflammation Program, National Cancer Institute, P.O. Box B, Frederick, MD 21702-1201; [d]SAIC-Frederick, Inc., National Cancer Institute, Frederick, MD 21702; and [e]Laboratory of Cell Biology, National Cancer Institute, 50 South Drive, Bethesda, MD 20892-8008

Edited* by Shuguang Zhang, Massachusetts Institute of Technology, Cambridge, MA, and accepted by the Editorial Board April 27, 2011 (received for review September 30, 2010)

Small motion dynamics within the electron density envelope reveal catalytic mechanism required for function.

MOPAC – PM7	Single Thread PGC	MKL (16Threads)	GPU (1 Tx)	GPU (4 Tx)
Chitinase (6759 atoms)	12800	98	12	4
Crambin (500 atoms)	7.2	0.2	0.2	0.15

ISCF, Ubuntu 14.04, CUDA 6, 4Xeon(4) 3.6(3.8)GHz, 4TitanX

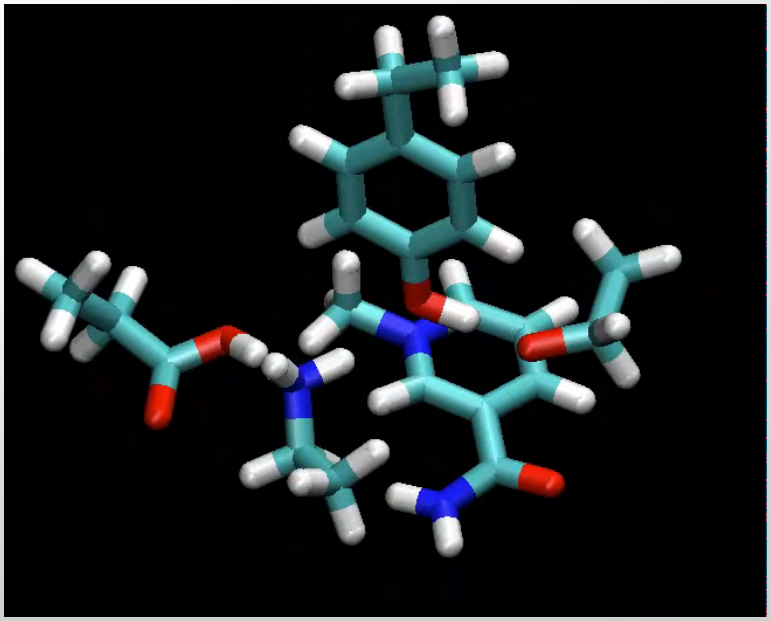
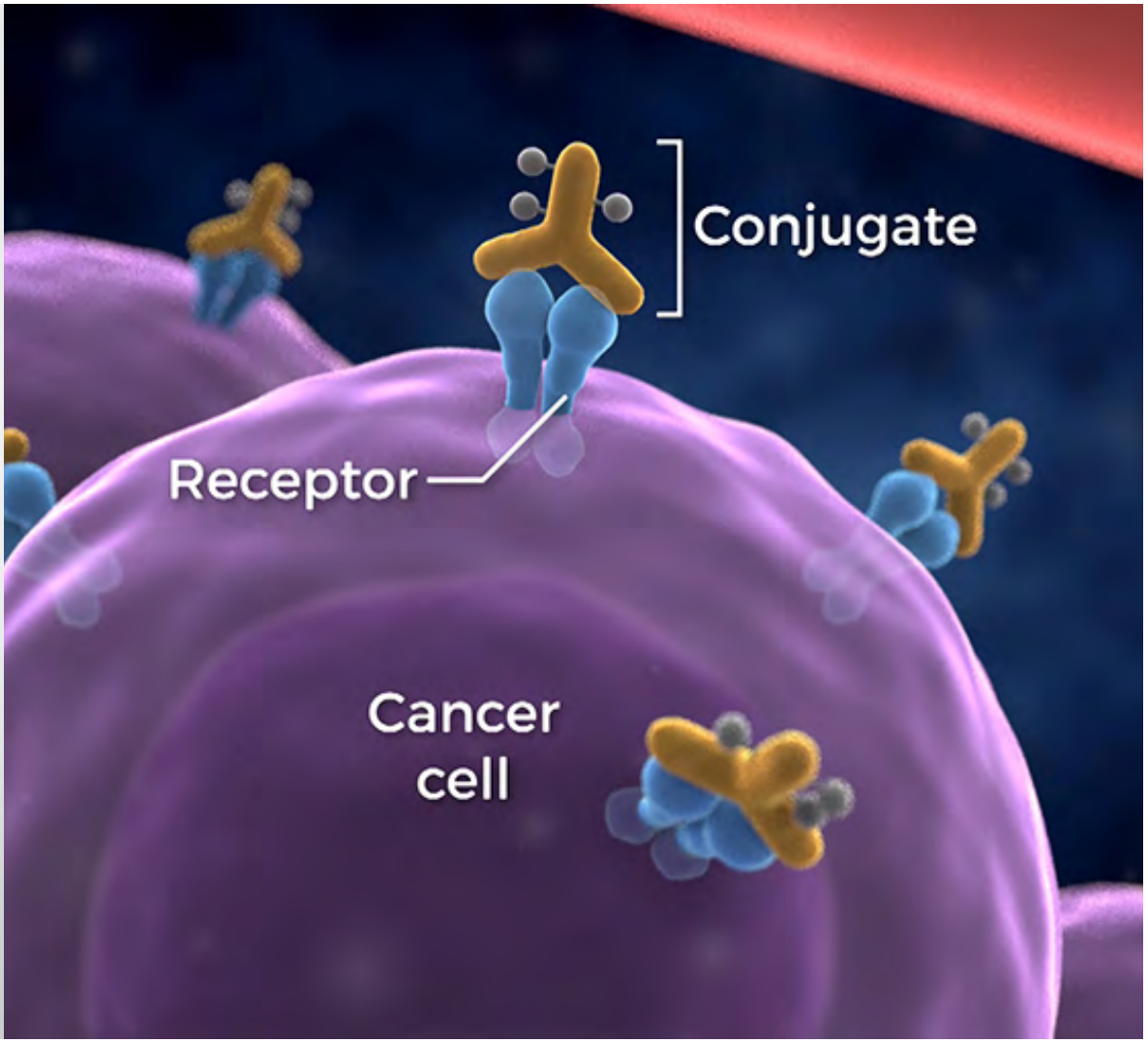
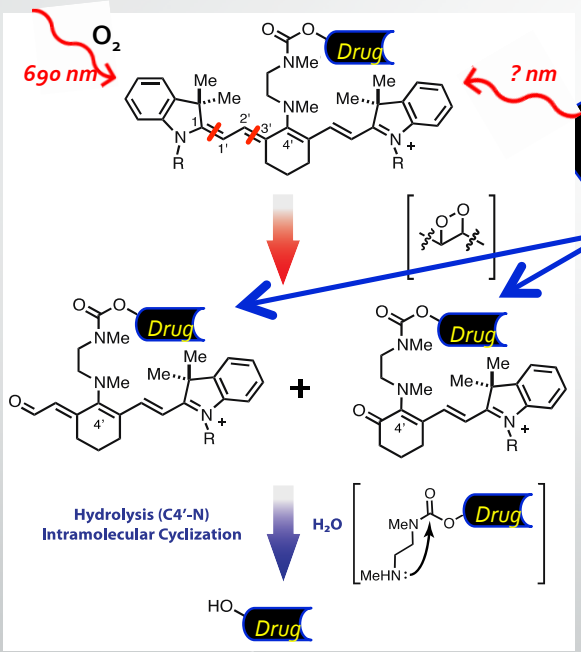


Photo-Immunotherapy



Computation to Assist Design of Molecular Agents for Oncology



Quantum Chemistry used to:

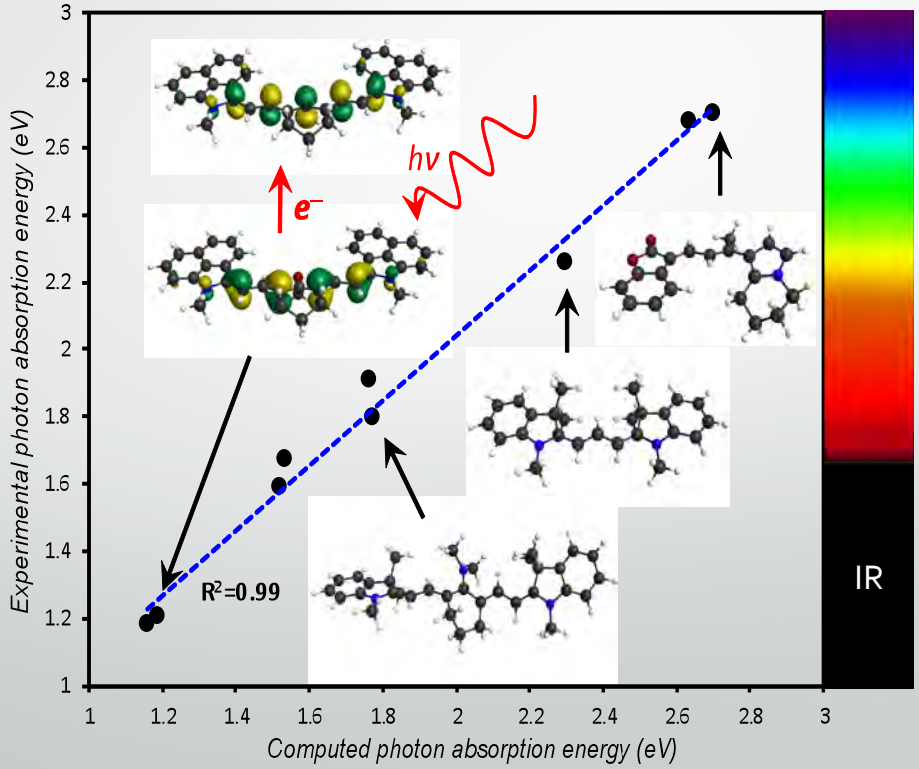
* *Predict activation wavelength of future delivery prototypes*

* *Explain why only two breaks occur*

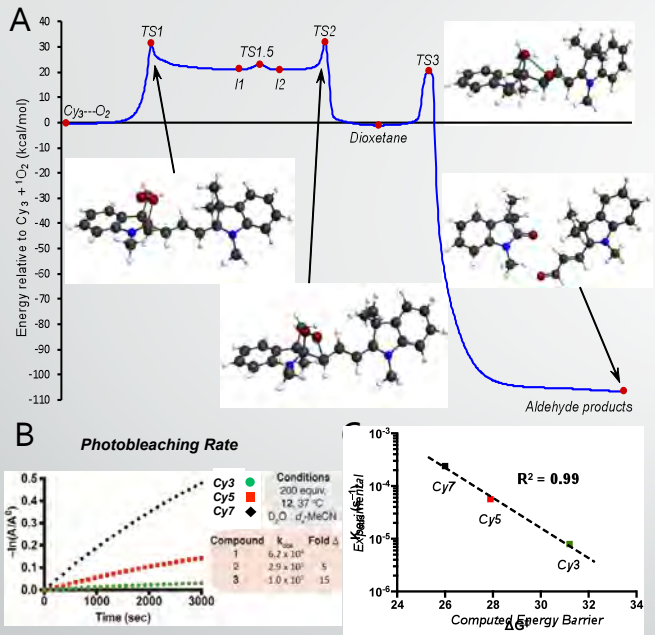
Goal is to assist in design of more effective delivery prototypes

Photo-Activated Drug Delivery

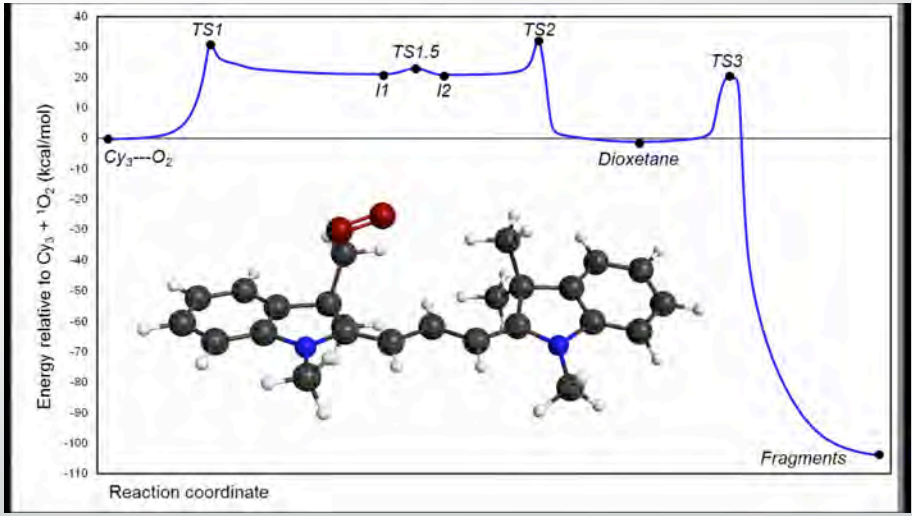
Infra-red and NIR -> Deeper Penetration



Kinetics of Photobleaching predicted via full reaction pathway calculation



Photobleaching is related to the lifetime of the drug in the system and overall stability for use as a therapeutic.



B3LYP-PCM geometry optimization: 24 cores, 1.5 hours, minimal memory
ORMAS-PT2-PCM computation of S0 -> S1 excitation energy: 16 core, 7 hours, 200 Gbytes
B3LYP geometry optimization: 16 cores, 8 hours, minimal memory
ORMAS-PT2-PCM S0 -> S1 excitation energy: 32 cores, 62 hours, 1.3 Tbytes

Cloud Computing – Initial Assessment

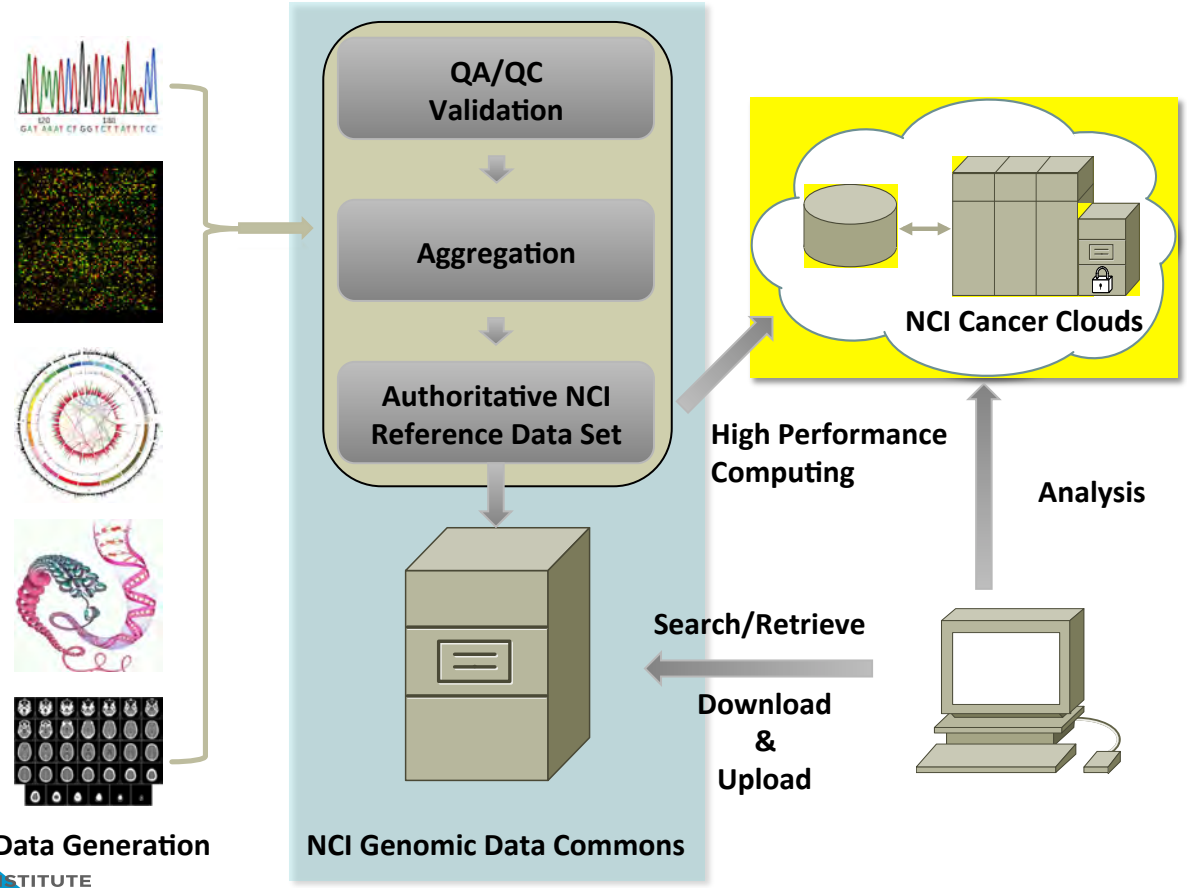
- Collaboration
- Reproducibility
- Data Transfer
- Containers

Test Data Analysis capabilities:
-Ability to support several languages
-Test **reliability, accuracy**, **reproducibility**
-Test **usability**
-Test **scalability**

Cloud Drivers

- Assuming the 2.5 PB TCGA data set
  - Downloading TCGA data at 10 Gb/sec would take ~23 days
  - Only large institutions have the ability to utilize this data
  - These data will continue to grow at an increasing rate

Cancer Cloud Pilots in Context



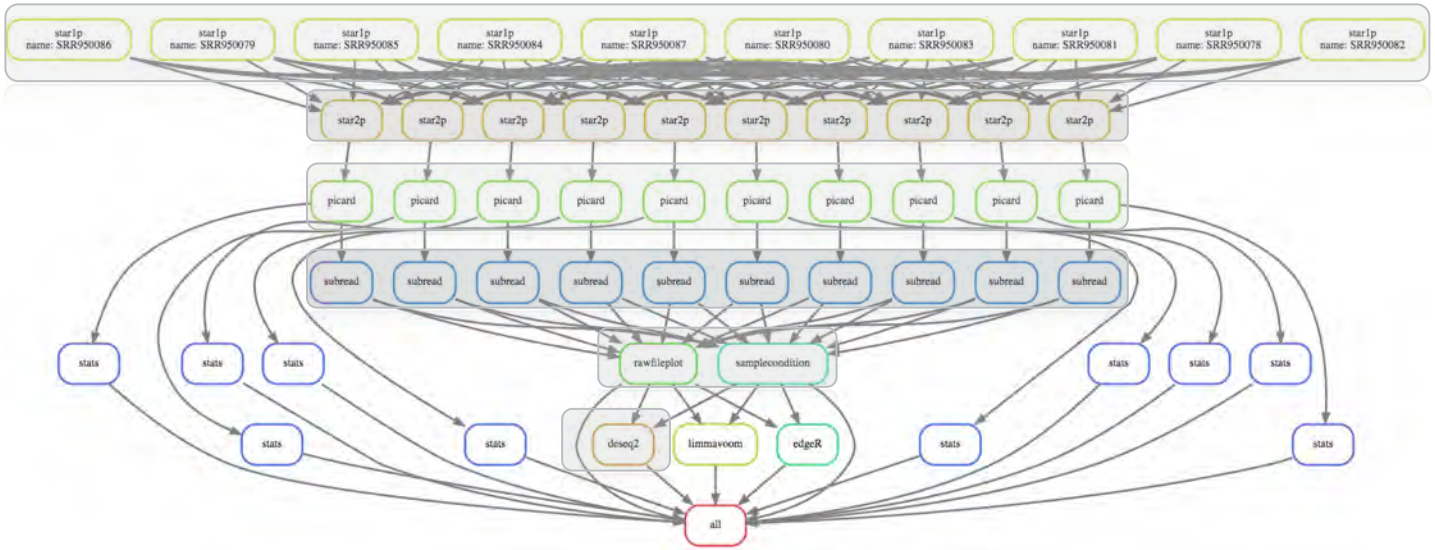
Objective of NCI Clouds Pilots evaluation

Test Data Access- TCGA data and local data

Test Data Analysis capabilities:
-Ability to support several languages
-Test **reliability, accuracy, reproducibility**
-Test **usability**
-Test **scalability**

BENCHMARK PIPELINE = data analysis pipelines used by the CCBR run on NIH's Biowulf cluster

RNA-seq workflow



Mean runtime of the SBG pipeline: 8.15h
Mean runtime of Benchmark pipeline: 6.93h

Benchmark (Biowulf)

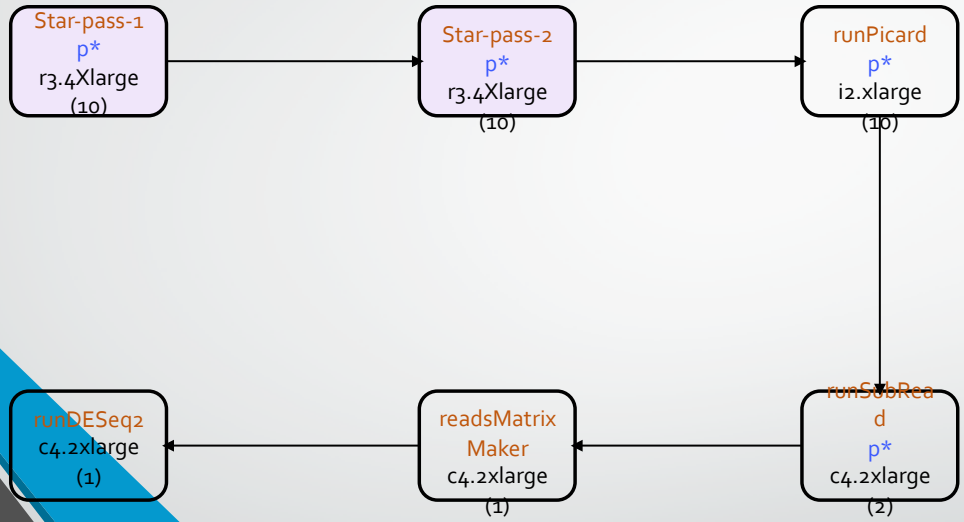
	Time (hours)
Min Total Runtime	4.08
Mean Total Runtime	6.93
Max Total Runtime	10.22

SBG - Amazon

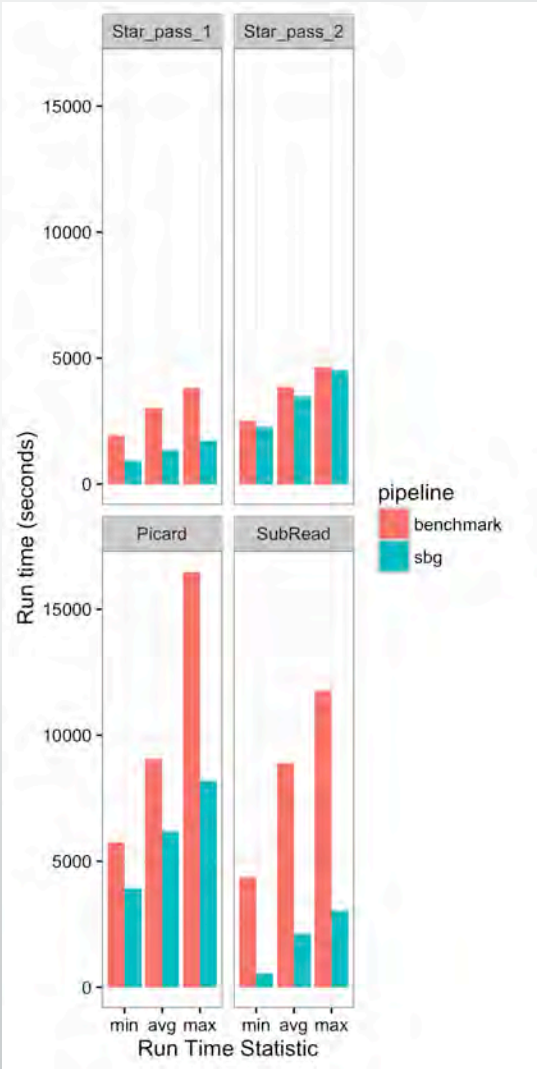
Run	Time (hours)	Price ($)	Completed
1	5.75	64.94	Y
2	5.75	65.55	Y
3	5.75	65.55	Y
4	5.75	65.55	Y
5	6.00	66.97	Y
6	6.00	65.43	Y
7	6.00	65.43	Y
8	7.00	64.01	Y
9	14.50	64.01	Y
10	19.00	67.42	Y
11	92.50	141.14	N (aborted)

Cloud outperforms Benchmark in tool-wise CPU times

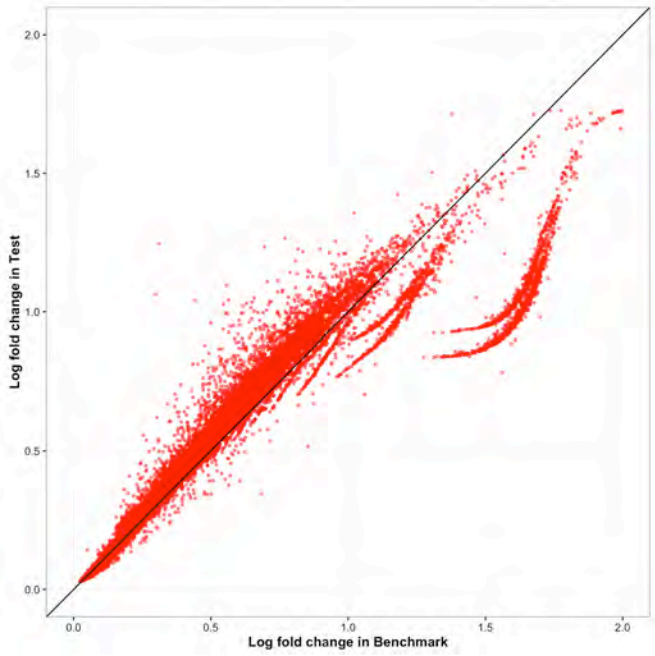
Data transfer times outweigh Cloud pipeline

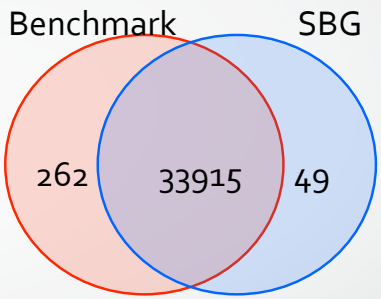


*run in parallel mode



Reproducibility





Overlap in the list of genes reported by DESeq2 in the Benchmark and SBG pipelines

(1) Number of genes reported in the final DESeq2 report are not identical.
Reported log fold changes between Benchmark and SBG pipelines are not concordant.

Heterogeneity leads to Increased Complexity and Challenges

Opportunities

- Ability to optimize individual steps of the workflow
- Significant gains in performance are seen
- Enable analyses not available before
  - More data
  - More complex data and integrated methods
  - Machine Learning

Challenges

- Cultural Challenges in the organization
- Information Security Challenges and increased planning
- Software Modernization
- Reproducible Results
- Skilled Computational Scientists, Developers, and Analysts
- When to go to Cloud? When not?
- Information Security

Optimized Code and Algorithms

- Code Libraries
  - More efficient
  - Faster time to solution
  - Reproducible
  - Correct algorithm translation to software
- Better Algorithms
  - Rethink the algorithm
  - Design for scaling
- Computers will be required to design better code if we are to achieve the potential of these systems
  - Computer Systems are more complex and Computers will program them

Acknowledgements

- Genomics: Yongmei Zhou, Justin Lack, Tin Li (Edico Genomics)
- Cryo-EM: Jianghai Zhu, Raul Cachau
- Biomaterials: Raul Cachau
- QM: Joe Ivanic
- Cloud: Anjan Purkayastha, Fathi Elloumi