GENOMES TO STRUCTURE TO FUNCTION (AND MOVIES): ROLE OF HPC

Jack R. Collins, Ph.D.

Frederick National Laboratory for Cancer Research

HPC User Forum

September 16, 2014

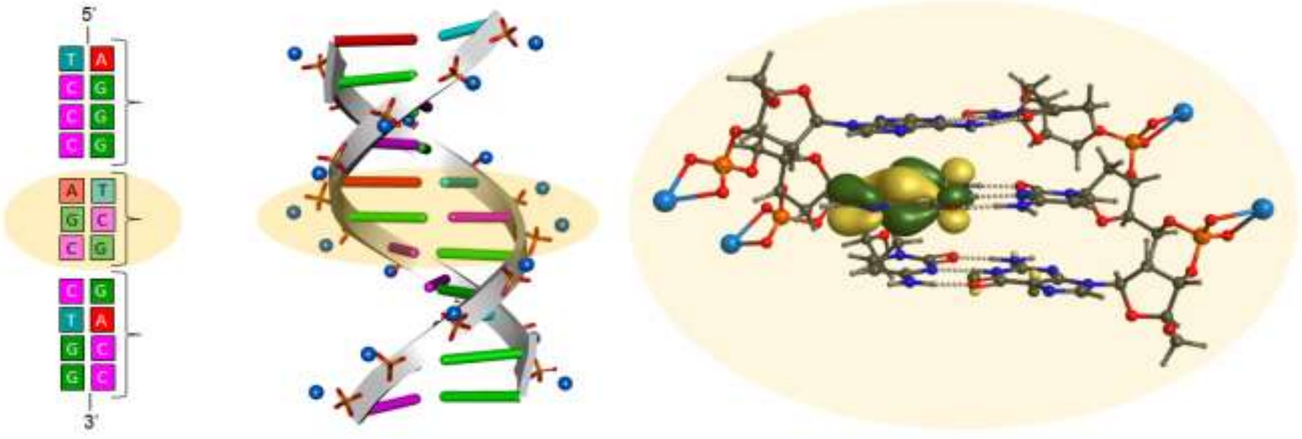
A Guiding View

Probably held by most HPC folks in this room

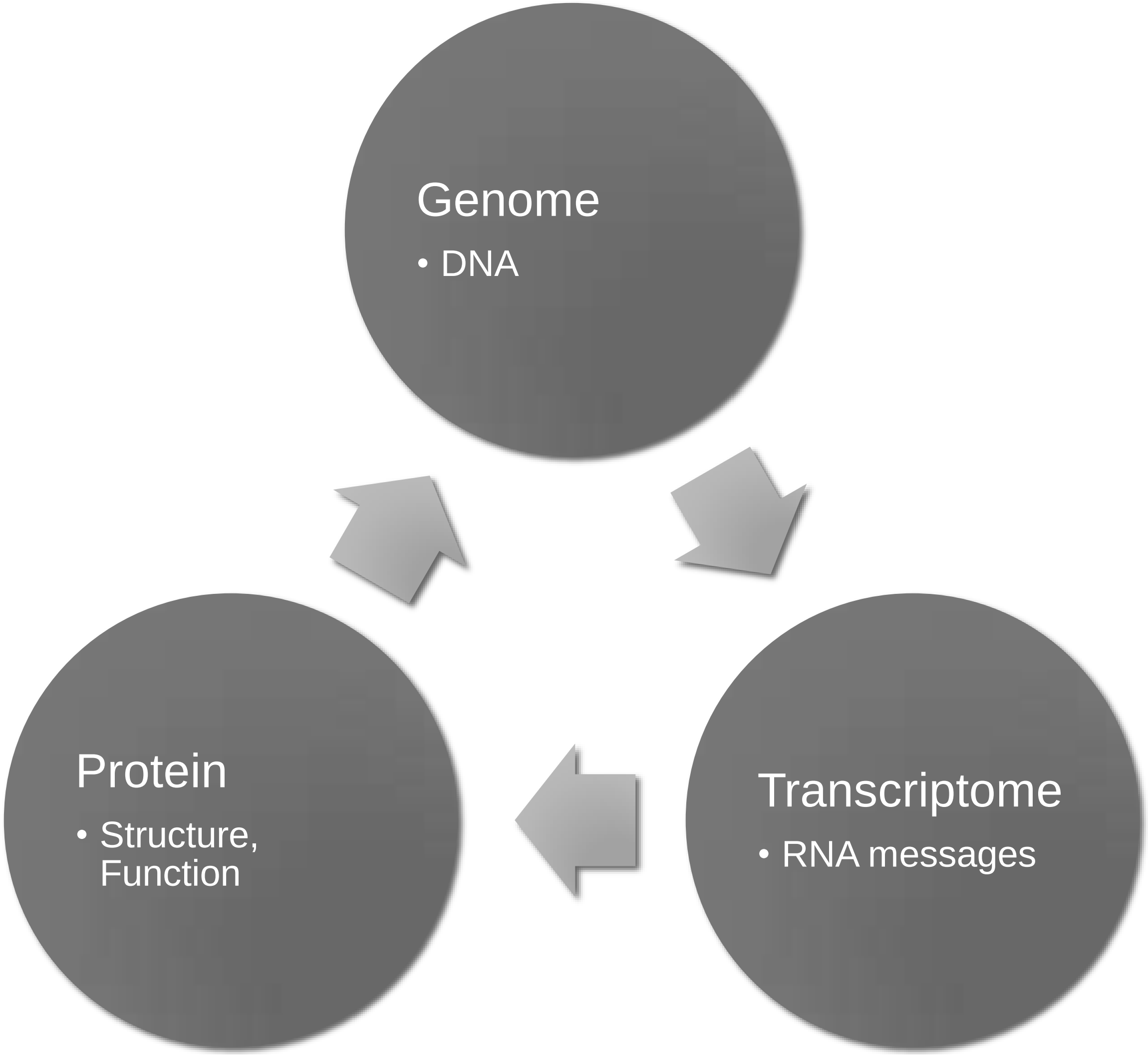
- “The more advanced the sciences have become, the more they have tended to enter the domain of mathematics, which is a sort of center towards which they converge. We can judge of the perfection to which a science has come by the facility, more or less great, with which it may be approached by calculation.” - Adolphe Quetelet
  - Edward Mailly, *Essai sur la vie et les ouv rages de Quetelet* in the *Annuaire de Vacadimie royale des sciences des lettres et des beaux-arts de Belgique* (1875) Vol. xli pp. 109-297 found also in "Conclusions" of *Instructions populaires sur le calcul des probabilités* p. 230
  - Wikiquote.org

Outline

- Changes in Biology / Life Sciences
- Data – Beyond Databases
- Examples merging simulation and experimental data
    - Ultra-high resolution structures
        - Xray
        - Electron Microscopy
    - Nanoparticles
        - Geno Nano-toxicity



Simple Biology



Genome

- DNA

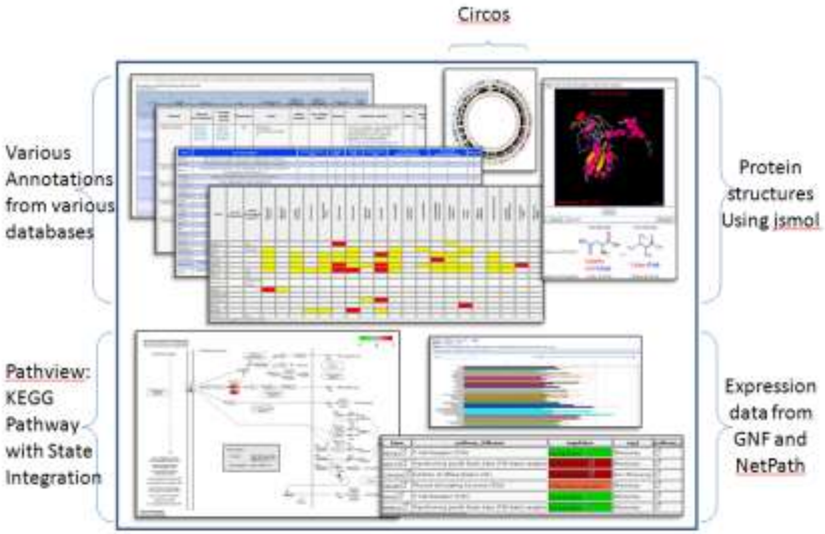
Transcriptome

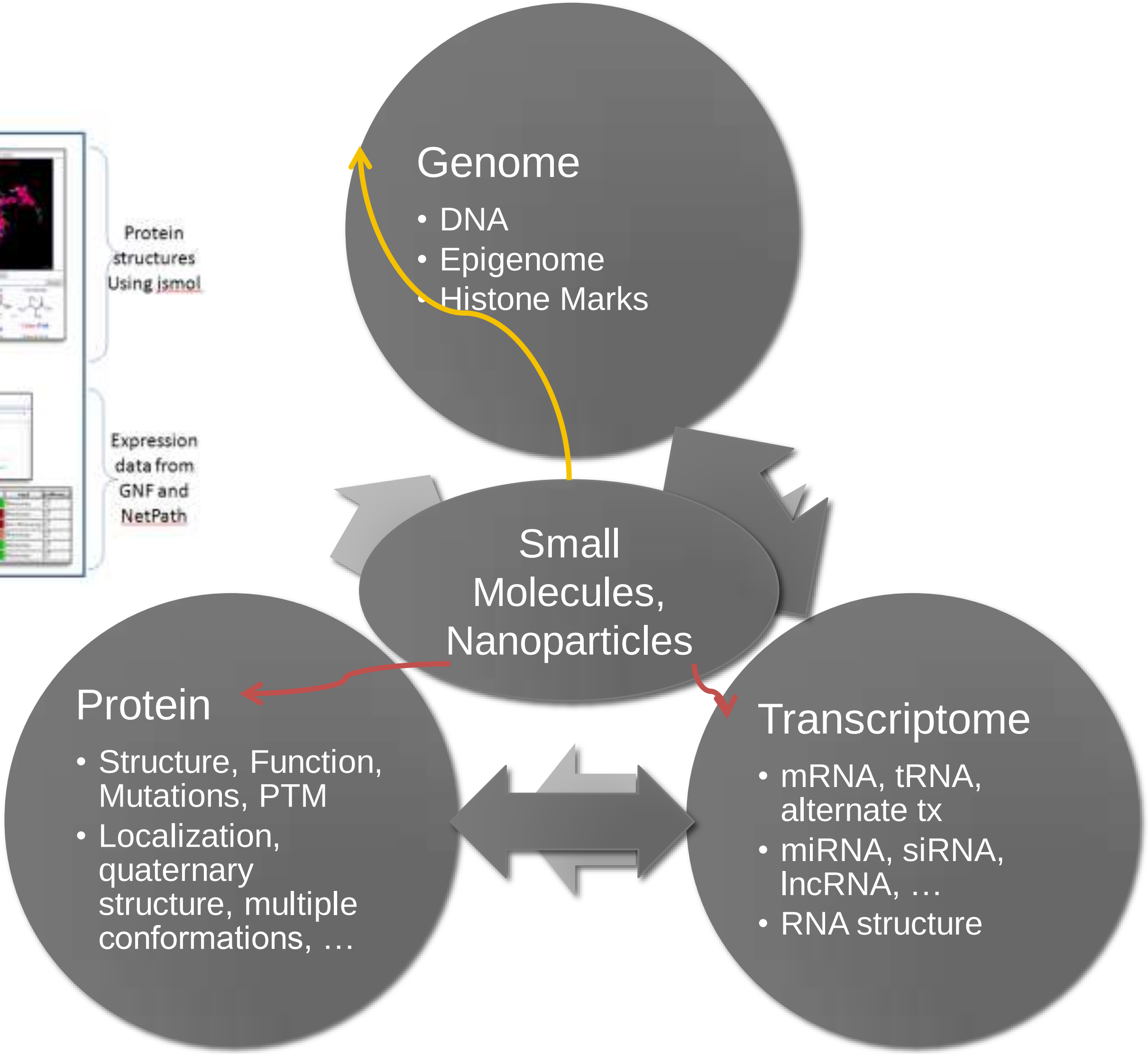
- RNA messages

Protein

- Structure, Function

Not So Simple Biology


Circos
Various Annotations from various databases
Protein structures Using jsmol
Pathview: KEGG Pathway with State Integration
Expression data from GNF and NetPath



Genome

- DNA
- Epigenome
- Histone Marks

Small Molecules, Nanoparticles

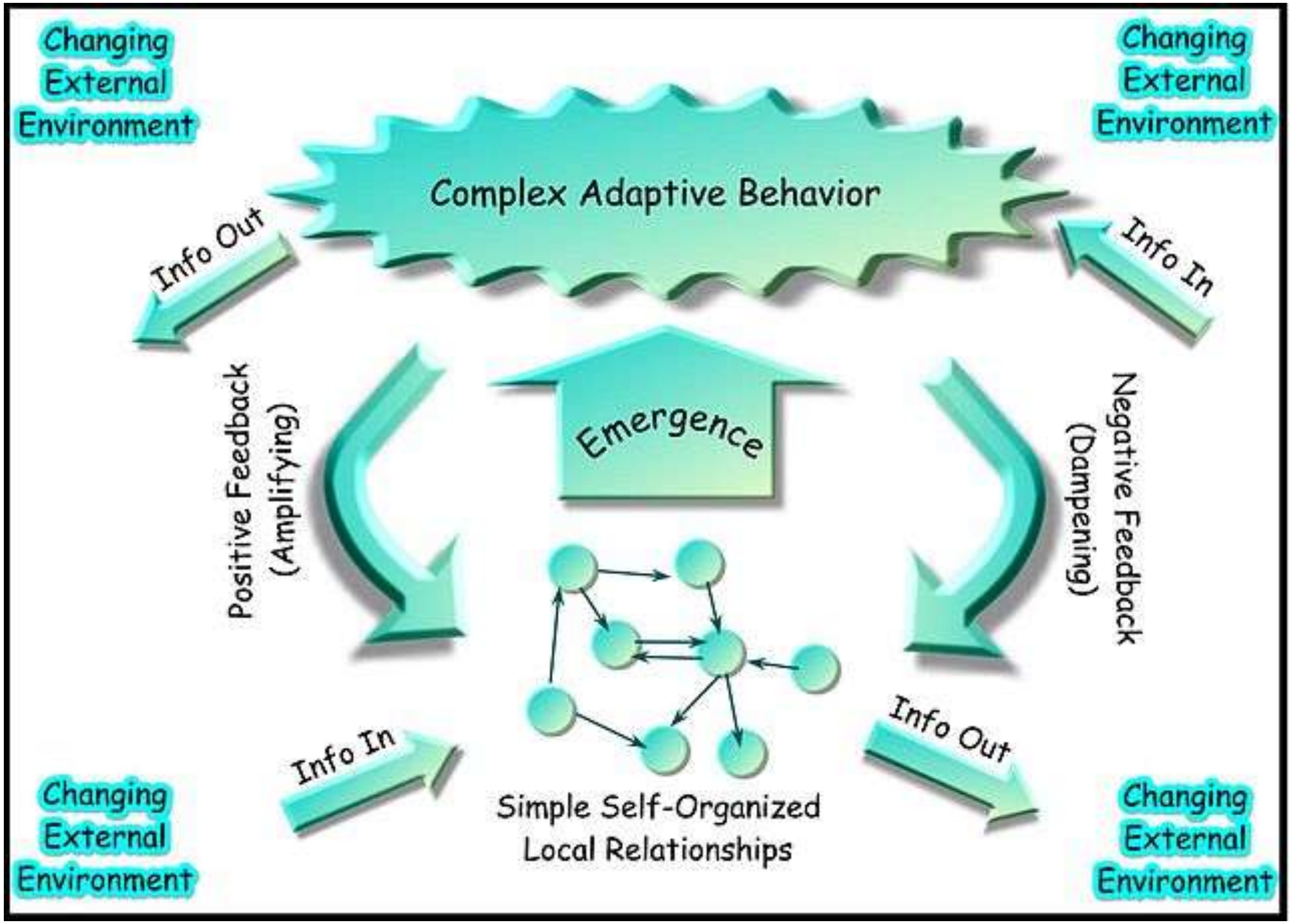
Protein

- Structure, Function, Mutations, PTM
- Localization, quaternary structure, multiple conformations, …

Transcriptome

- mRNA, tRNA, alternate tx
- miRNA, siRNA, lncRNA, …
- RNA structure

Closer to Reality

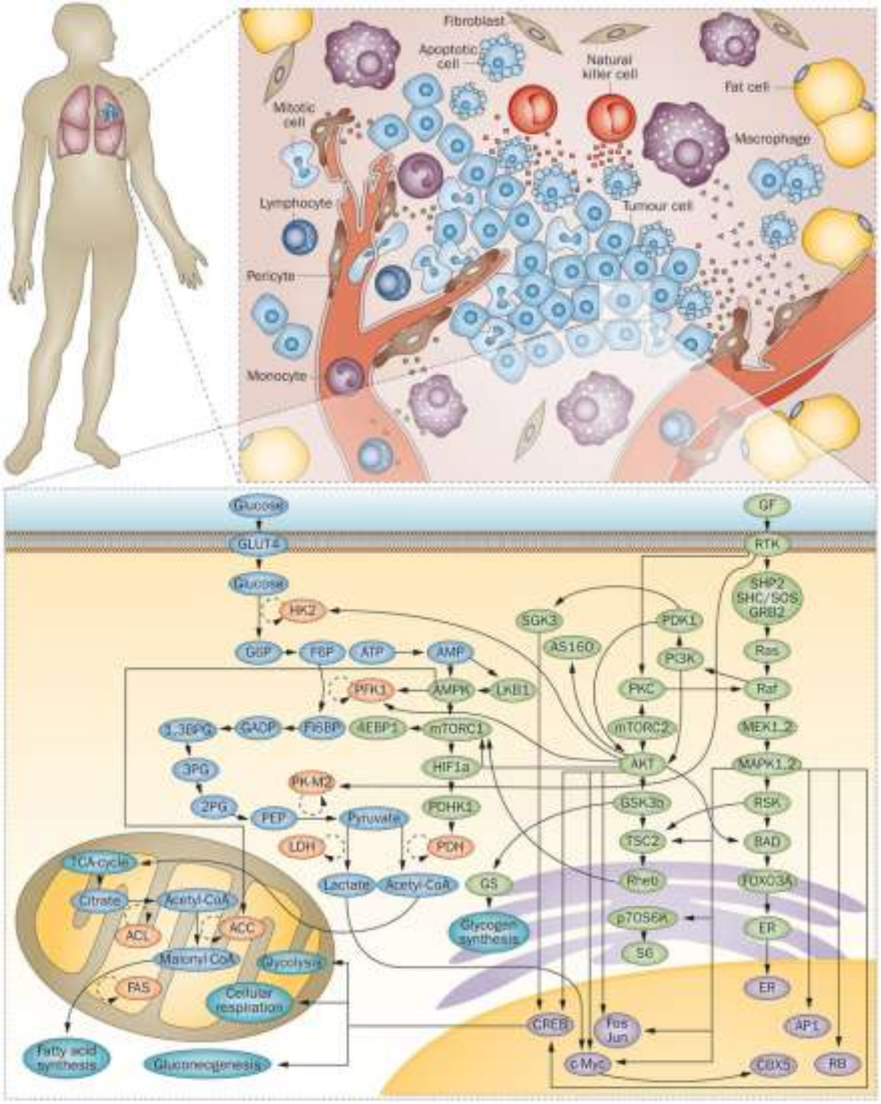


Changing External Environment
Changing External Environment
Complex Adaptive Behavior
Info Out
Info In
Emergence
Positive Feedback (Amplifying)
Negative Feedback (Dampening)
Info In
Info Out
Simple Self-Organized Local Relationships
Changing External Environment
Changing External Environment

Figure taken from: http://en.wikipedia.org/wiki/Complex_systems_biology

“The fact that cancerous cells can be inserted into an animal and not develop into a tumor, reinforces the theory that it is not the char- acteristics of the cells themselves, that result in cancer, but the properties emerging from the interaction between the cell and other response systems.” Knox Cancer Cell International 2010, 10:11

Beyond Databases: Application of cancer systems biology to decipher complex interactions in multiple dimensions



Cancer Systems Biology: a peek into the future of patient care?
Henrica M. J. Werner,
Gordon B. Mills &
Prahlad T. Ram
Nature Reviews Clinical Oncology 11, 167–176 (2014)
doi:10.1038/nrclinonc.2014.6

Beyond Databases:
Example from the microbiome

Vaginal microbe yields novel antibiotic

Nature, Erika Check Hayden, 11 September 2014

- Drug is one of thousands that may be produced by the human microbiome.
- "This is a great example of the **power of bioinformatics to not merely identify genes of interest from 'big data' 'omics, but to connect together cassettes of genes** to increase our fundamental understanding of how commensal bacteria maintain a healthy human microbiome," says microbial genomicist Derrick Fouts of the J. Craig Venter Institute in Rockville, Maryland: Quoted in Nature.

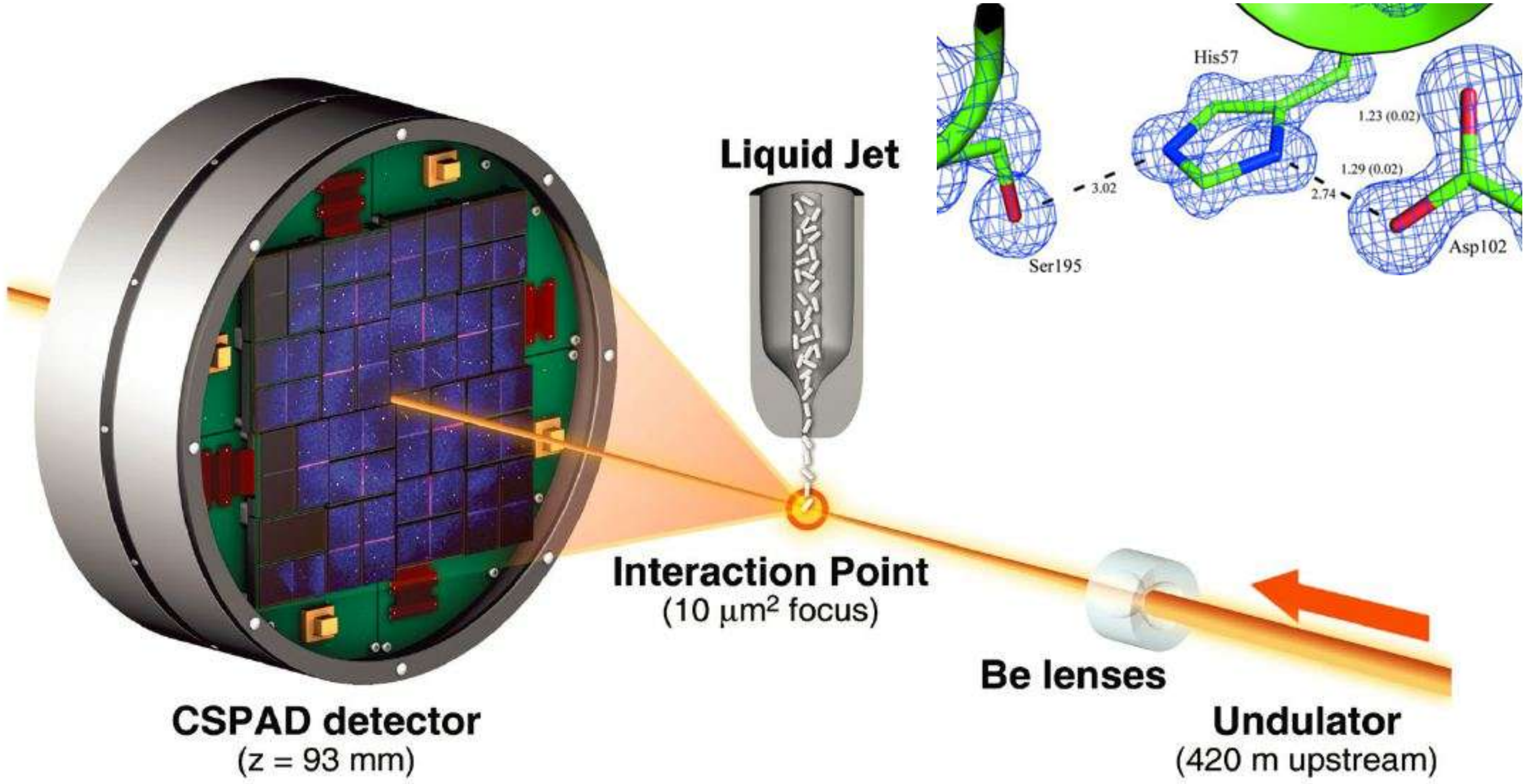
Developing Therapeutics: Functional understanding and drug development require 3D structures

Examples with a common theme:

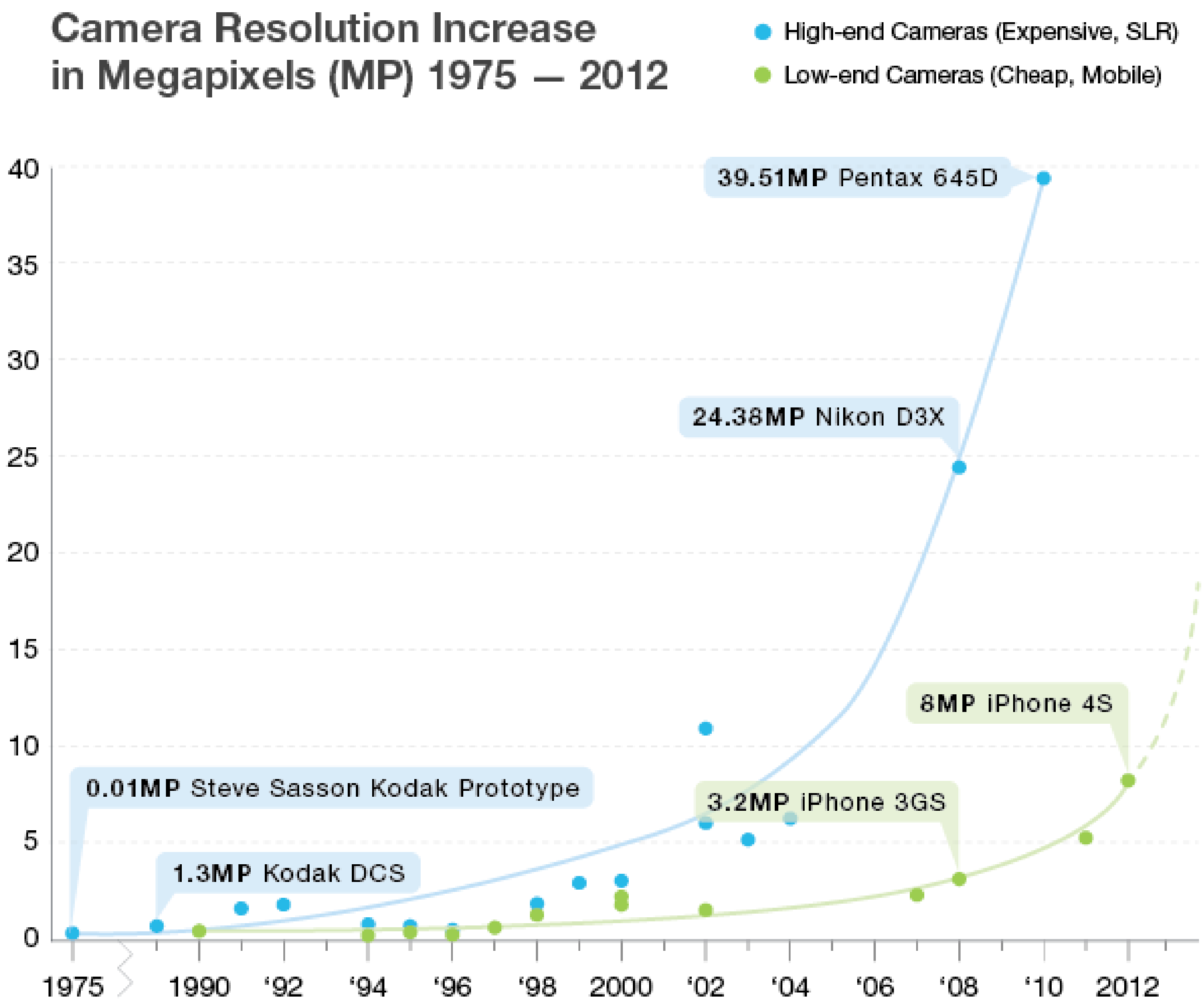
- BioXFEL: CXI detectors capture millions of high resolution images before merging at ~400Mb/image with current technologies. (Currently, 40TB/day : Next generation ~150TB/day)
- EM: Very large data sets (1.5Tb / set with current technologies)
- LVEM: Highly redundant sets (> 20,000 images per stack per view)

What’s technologies are fueling the advances in structural biology?

Brighter Light: *x-ray lasers can generate ultra-high resolution structures*



Better Detectors / Cameras



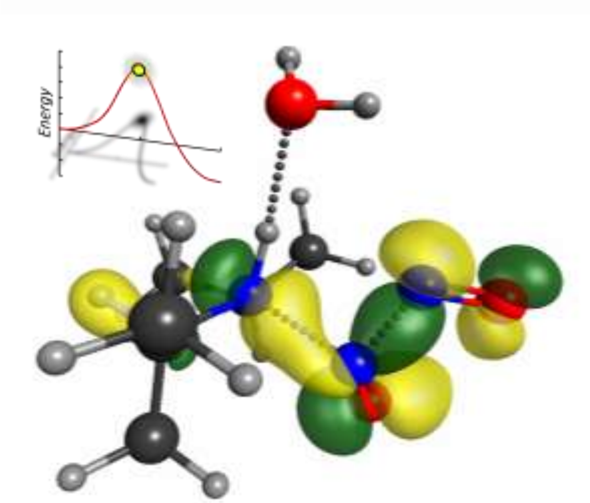
Science benefits from consumer and astronomy applications driving increased sensitivity and speed (Just like GPU advances are helping push HPC.)

X-ray imaging of biomolecules

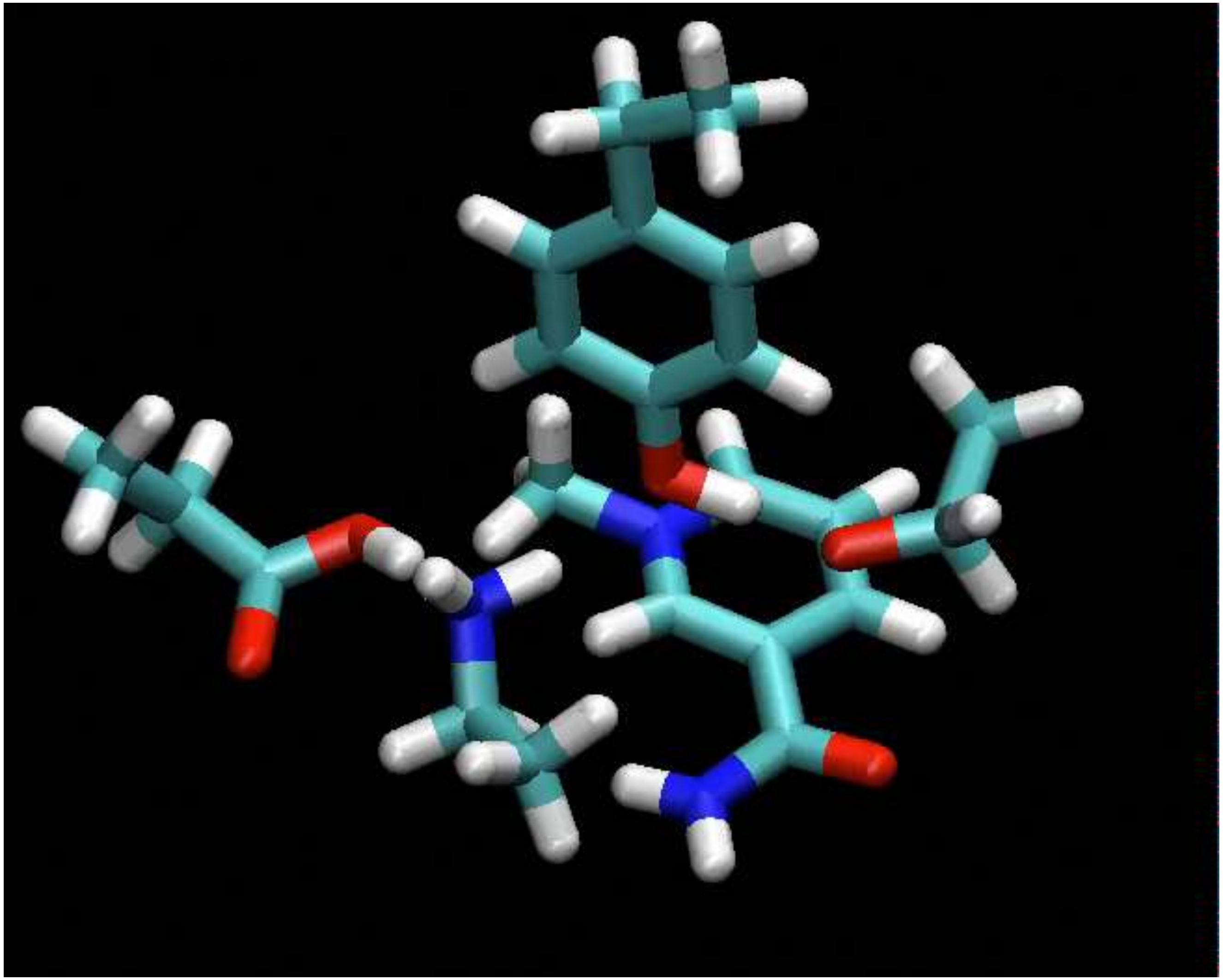
- "With the new bioimaging technique developed in the BioXFEL center, we will be able **to analyze crystals 1,000 times smaller** than the ones we can use now," Lattman said. "These are crystals we could never use before and, in fact, may not have known existed. A whole new universe of drug targets will become accessible for study as a result."
- "The techniques the BioXFEL center will develop could shorten the process of determining protein structure from years to days," said Ourmazd of the University of Wisconsin-Milwaukee. "This **will rely heavily on mathematical algorithms** we and others are developing to deduce structure from millions of ultralow-signal snapshots."
- A key advantage is that it will let scientists **see the motions of molecules for the first time**. "Most biological processes require movements within the molecules involved," Lattman said.
- http://www.buffalo.edu/ubreporter/featured-stories.host.html/content/shared/university/news/ub-reporter-articles/stories/2013/lattman_bioxfel.detail.html#sthash.E3SHsnSy.dpuf

Merging QM with Experiment to explore "*chemical resolution*"

- **Ultra-high resolution** data contains finer details including the positions of protons - important to function of proteins and to aid in NMR refinement.
- Some of these processes can not be described using stationary models but can be revealed by refining the structure using a combination of quantum mechanical tools and careful matching of the electron density data.

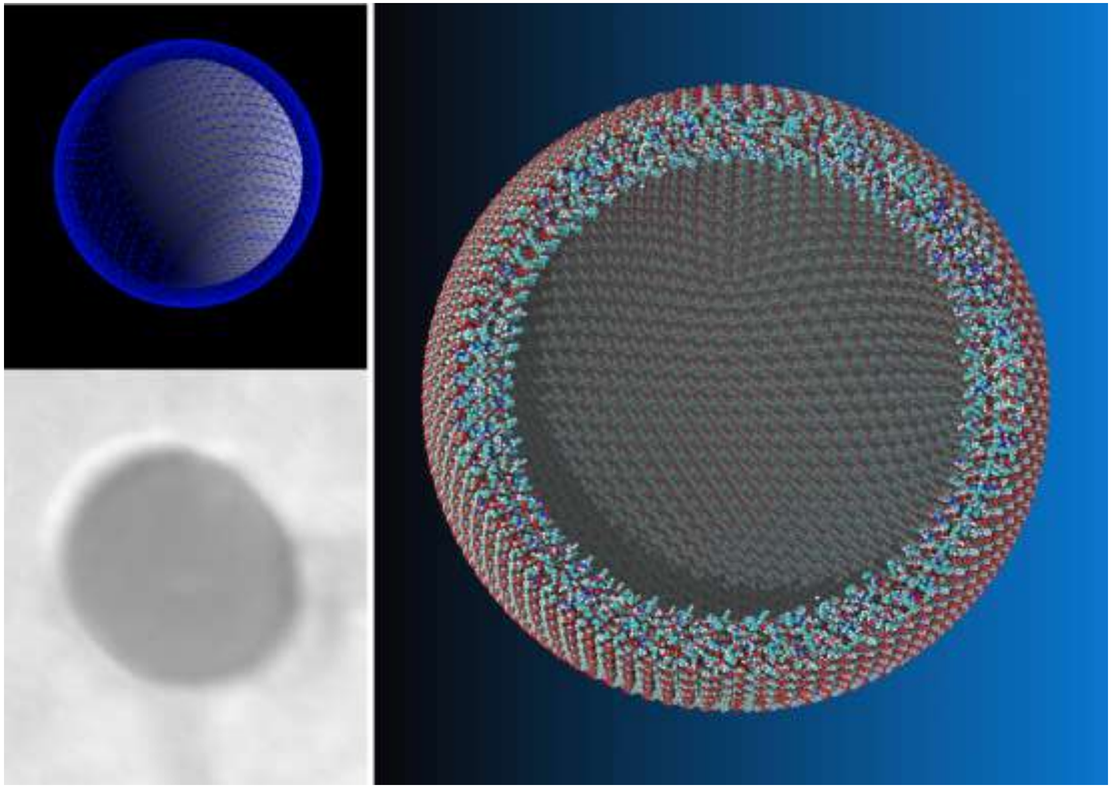


QMRx: *Small motion dynamics within the electron density envelope reveal distribution related to catalytic mechanism required for function.*



Structural Analysis integrating EM and QM tools for large molecular aggregates

Use of high contrast (low voltage) EM for the 3D reconstruction of a complex nanomaterial without preprocessing of the sample or the use of staining agents.



Clockwise from bottom left: 1) Electron microscopy image of self assembled nanoparticle with a hydrodynamic radii ~ 22.5nm. 2) Intermediate model. 3) The final model (right) contains 670,000 atoms.

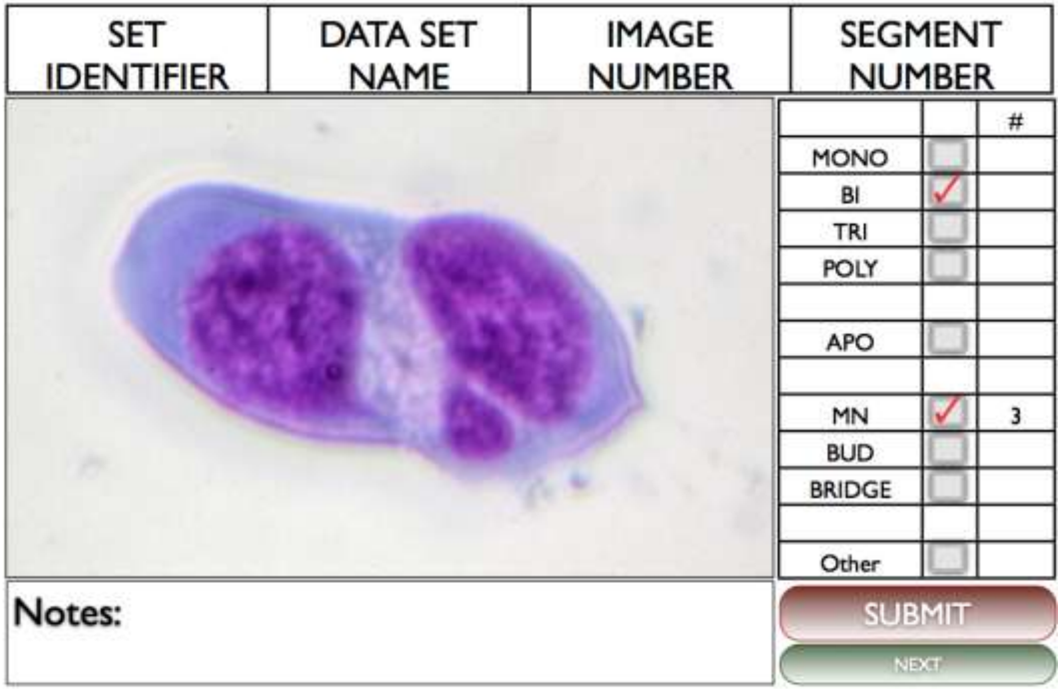
Nanoparticle simulation and FDA approval

Incorporate the relevant risk characterization information, hazard identification, exposure science, and risk modeling and methods into the safety evaluation of nanomaterials.	Risk Characterization	Targeted research in FDA-regulated product areas of potential nanotechnology applications **where risk characterization information** would help **to enhance the understanding of hazard identification**, exposure science, and **risk modeling**.
Evaluate risk assessment approaches for risk management.	Risk Assessment	Enhance state of knowledge and **scientific evidence to support potential development of generalized class-based approaches** to risk assessment of FDA-regulated products containing nanomaterials.
Integrate and standardize risk communication within the risk management framework	Risk Communication	Improve risk communication associated with FDA regulated product areas that either contain nanomaterials or product areas otherwise relevant to nanotechnology

Geno-Nano-Toxicity

- Recent studies show the *in vitro* micronucleus assay to be a powerful tool in the study of nanoparticle-induced genotoxicity. ABCC developed procedures that facilitate the use of high contrast images to improve the quantitative annotation of micronucleus assay images.

Automated workflow with feature extraction can facilitate the access to archived data providing an extra benefit to re-evaluate results and to facilitate the compilation of training sets.



SET IDENTIFIER	DATA SET NAME	IMAGE NUMBER	SEGMENT NUMBER

		#
MONO	□	
BI	☑	
TRI	□	
POLY	□	
APO	□	
MN	☑	3
BUD	□	
BRIDGE	□	
Other	□	

Notes:

Role/Challenges for HPC

- Challenge of integrating “Big Data” into the Enterprise HPC infrastructure to enable workflows using heterogeneous technologies. (NoSQL, Hadoop, Graph Analytics, Literature, etc.)
  - System may need to be “tuned/balanced” differently
- Challenge of integrating heterogeneous computational technologies (CPU, Big Memory, Accelerators - GPGPU, Phi, FPGA) to work together efficiently.
  - System may need to be designed differently
- Challenge of efficient software to effectively make use of the heterogeneous HPC infrastructure.
  - Software may need to be redesigned and rewritten
- Challenge of integrating skilled HPC people to catalyze adoption/innovation using HPC computational resources.

Acknowledgements

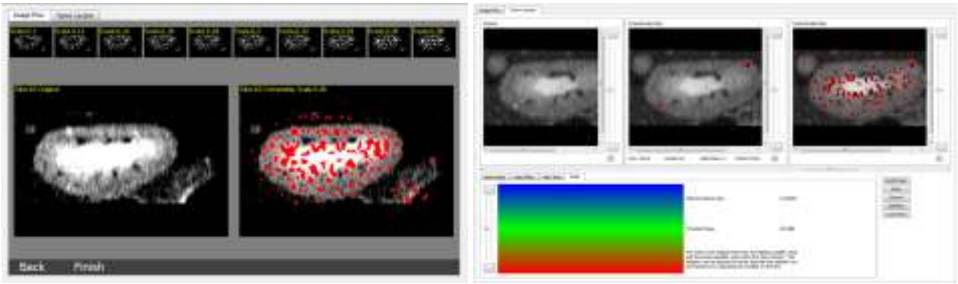
- Raul Cachau, Ph.D.
- Yanling Liu, Ph.D.
- Joe Ivanic, Ph.D.
- Brian Luke, Ph.D.
- Uma Mudunuri

Just for Fun: A Life Science User/Researcher*

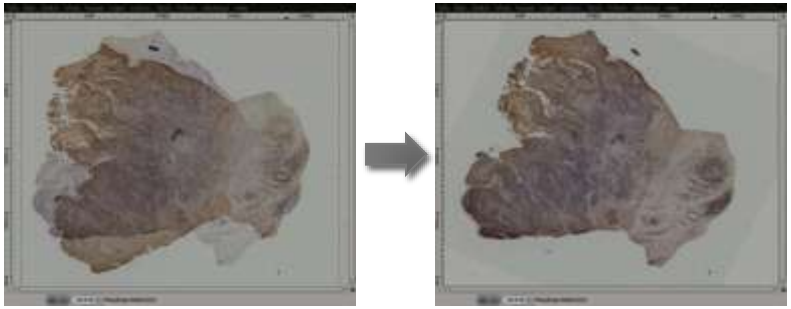
*Not representative of all, but not uncommon

- Preferred programming model?
  - R, Matlab, or Python (maybe Java)
- Algorithm/Code may not be “FLOPs” dependent
  - Often involves integer or character or mixed
- Uses a Mac because “it works”
  - Generally doesn’t want to be bothered with the details of “how” it works but wants it to work and solve their problem when they need it.
- Will spend money on generating lots of data, students, postdocs
  - And often worry about what to do with the data afterward
- Generally prefers Open Source software
  - Many of the applications change rapidly with little or no support
  - Wants to play with it on their laptop before production
- Hear that GPUs (Phi, etc.) can make my application run faster
  - Can you port my script?
- Is willing to use “Cloud” because you don’t have to wait for IT to provision a system (and there are community scripts and AMIs that make it relatively easy)
- Wants to stay up to date with the latest cutting edge science
  - This was published yesterday and I want to use it on the HPC system

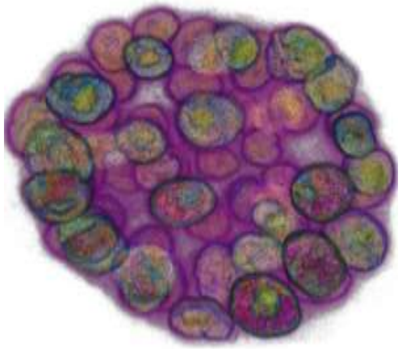
Merging Enterprise and HPC: Optimizing People and Workflows



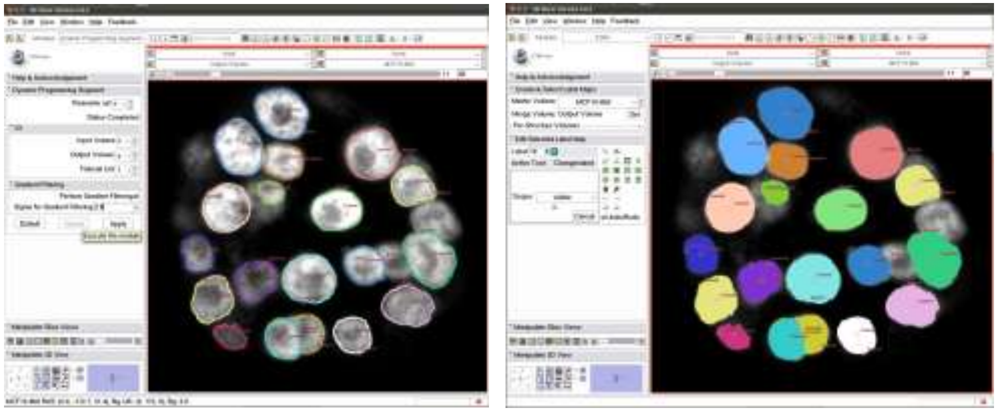
Common Imaging Tool Development for SAIP



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



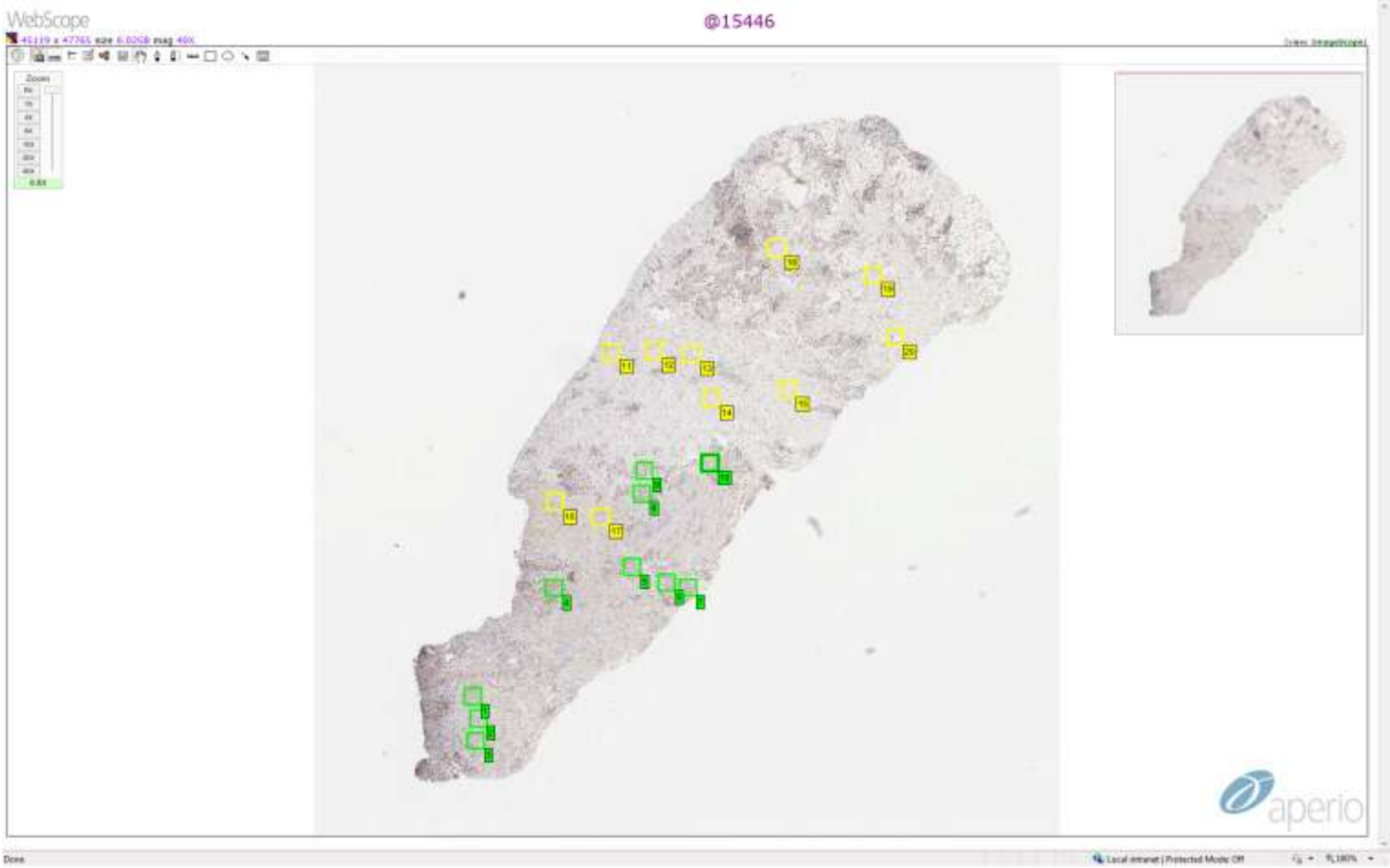
Automatic Visualization on 3D Biological Datasets



New Image Segmentation Module in Open Source Imaging Software 3D Slicer

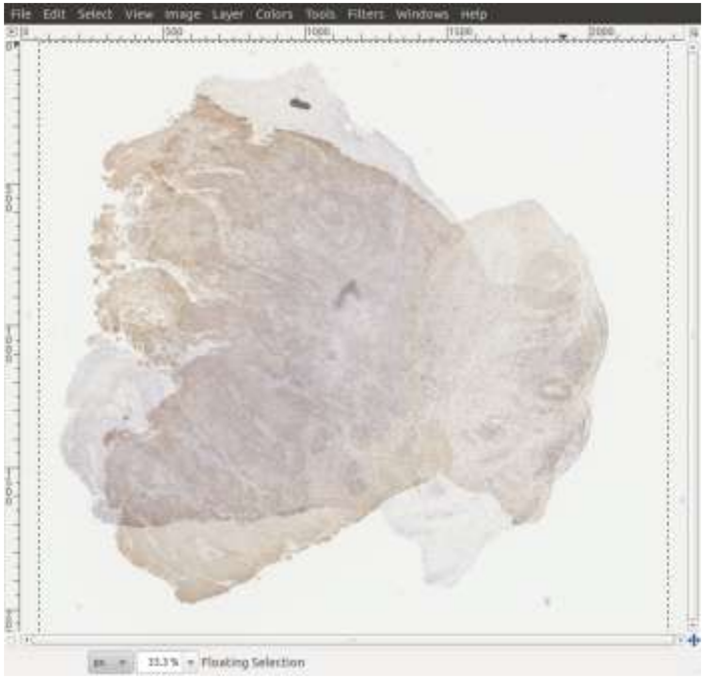
Optimizing People and Workflows

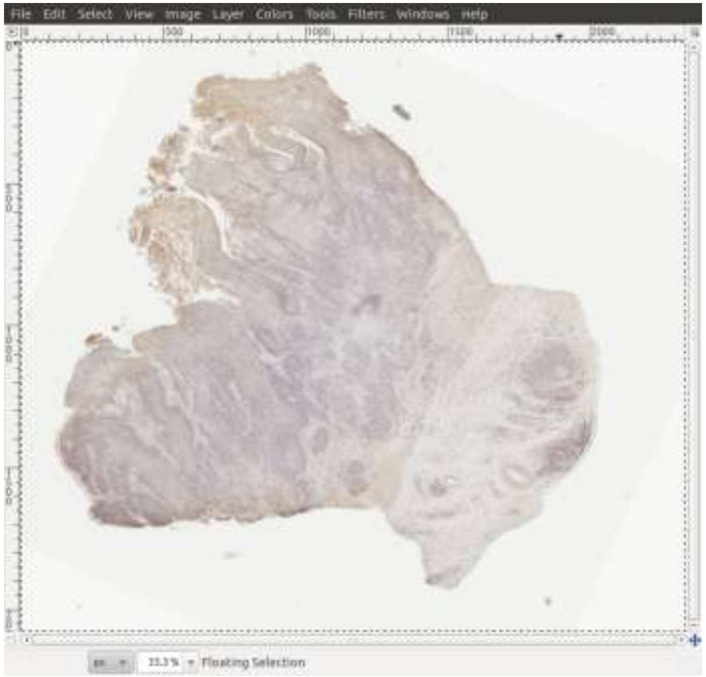
Pathologist Annotates Image


WebScope
@15446
45119 x 47765 size 6.02GB mag 40X
Zoom
0.8X
11
12
13
14
20
aperio
Done
Local intranet | Protected Mode: Off
100%

Optimizing People and Workflows

HPC Problem: Processing and Analysis


File Edit Select View Image Layer Colors Tools Filters Windows Help
500
1000
1500
2000
px
33.3 %
Floating Selection


File Edit Select View Image Layer Colors Tools Filters Windows Help
500
1000
1500
2000
px
33.3 %
Floating Selection

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
- NCSA system
  - Brute Force full resolution registration requires 720 cores 1.8TB mem for 240 hours
  - Modification to use low resolution + full resolution refinement requires 180 cores 450GB mem for 40 hours
- Result: **What would have been an impractical / impossible problem for the pathologists with the tools they had was solved fairly easily by a practical application of HPC and domain expertise**.

What would my HPC computer look like?

- Lots of memory bandwidth.
- Many lookups, compares, and branches per clock tick. Not just Flops.
- Ingest data from LARGE databases (I/O)
- Scale as I need to reduce time to solution or grow model or model complexity evolves
- 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

From a presentation given at HPC
User forum 2 1/2 years ago.

Software?

- Programming model?
- Skilled programmers?
- More efficient utilization of the CPU
    - Beyond a few % of theoretical peak