National Cancer Institute

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.

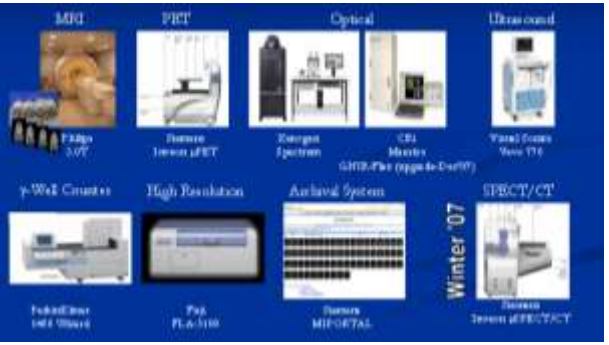
"Analyze this" *Nature Methods* ***8, 361 (May, 2011) Editorial***

"**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

Core Services (Exp)



Integrated Services (IT + Science)

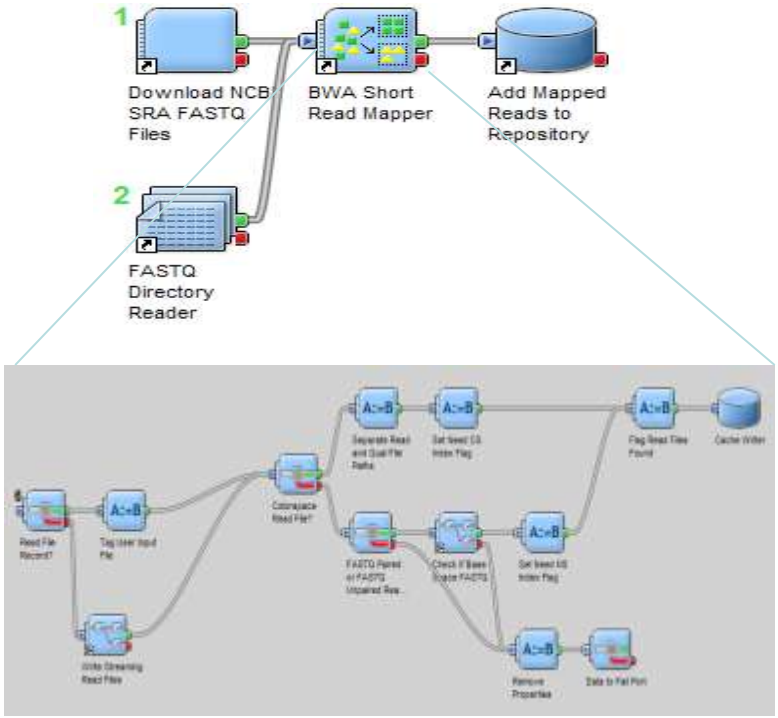
Mgmt

- Computational 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

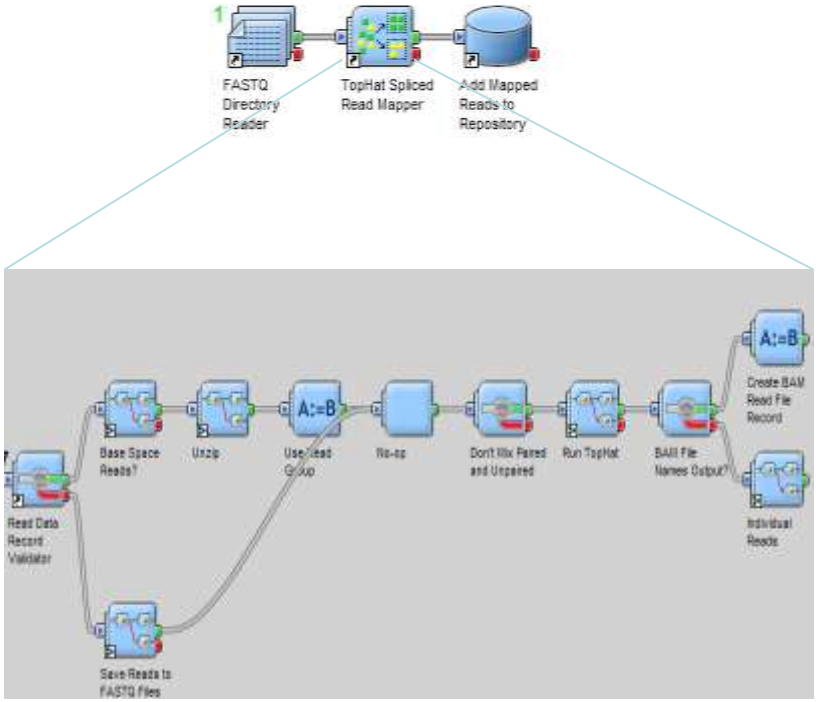
Workflow Automation

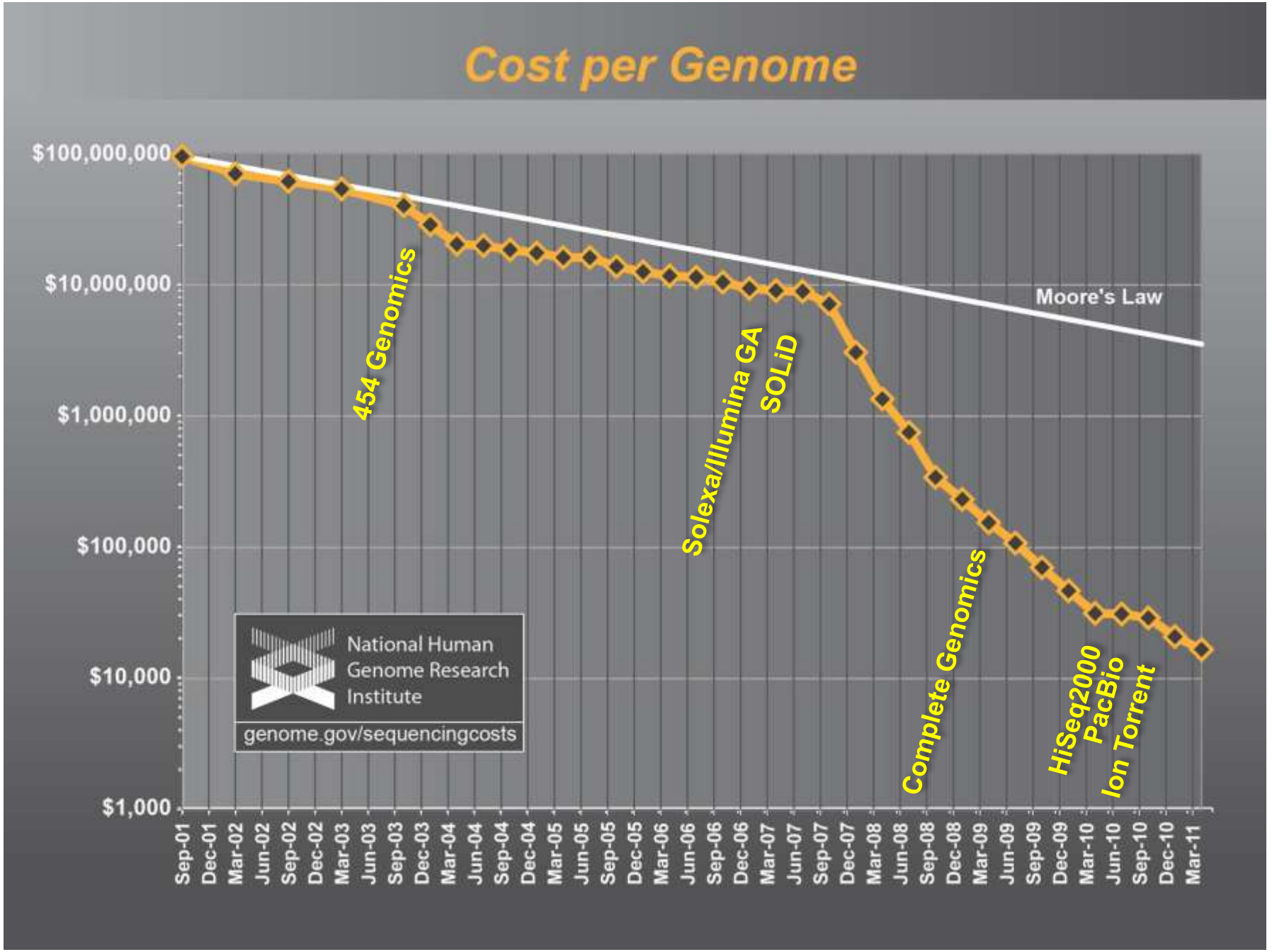
Build HPC into modules

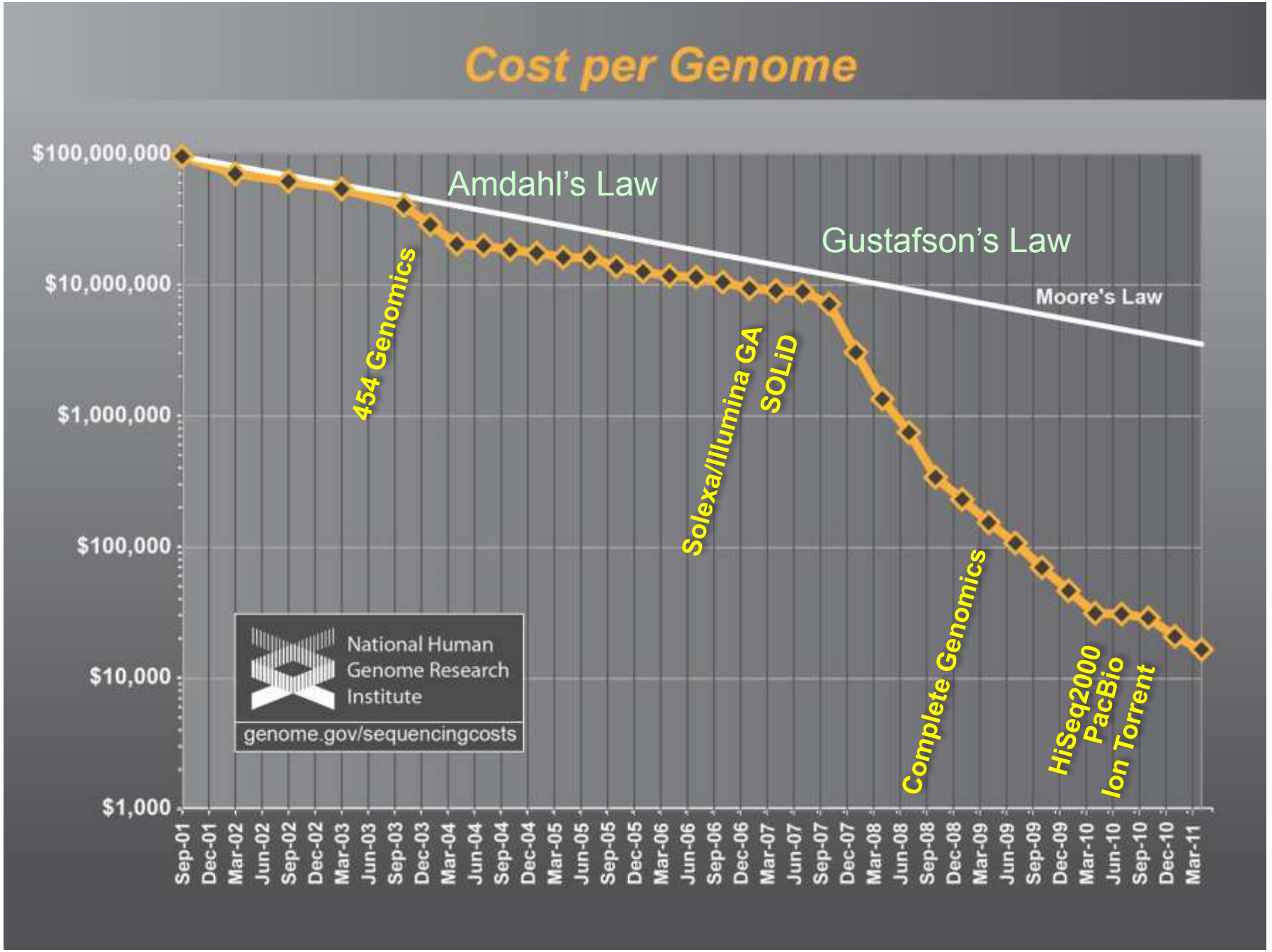
ChIPSeq



RNASeq




Cost per Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
Moore's Law
454 Genomics
Solexa/Illumina GA
SOLiD
Complete Genomics
HiSeq2000
PacBio
Ion Torrent
National Human Genome Research Institute
genome.gov/sequencingcosts
Sep-01
Dec-01
Mar-02
Jun-02
Sep-02
Dec-02
Mar-03
Jun-03
Sep-03
Dec-03
Mar-04
Jun-04
Sep-04
Dec-04
Mar-05
Jun-05
Sep-05
Dec-05
Mar-06
Jun-06
Sep-06
Dec-06
Mar-07
Jun-07
Sep-07
Dec-07
Mar-08
Jun-08
Sep-08
Dec-08
Mar-09
Jun-09
Sep-09
Dec-09
Mar-10
Jun-10
Sep-10
Dec-10
Mar-11


Cost per Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
Amdahl's Law
Gustafson's Law
Moore's Law
454 Genomics
Solexa/Illumina GA
SOLiD
Complete Genomics
HiSeq2000
PacBio
Ion Torrent
National Human Genome Research Institute
genome.gov/sequencingcosts
Sep-01
Dec-01
Mar-02
Jun-02
Sep-02
Dec-02
Mar-03
Jun-03
Sep-03
Dec-03
Mar-04
Jun-04
Sep-04
Dec-04
Mar-05
Jun-05
Sep-05
Dec-05
Mar-06
Jun-06
Sep-06
Dec-06
Mar-07
Jun-07
Sep-07
Dec-07
Mar-08
Jun-08
Sep-08
Dec-08
Mar-09
Jun-09
Sep-09
Dec-09
Mar-10
Jun-10
Sep-10
Dec-10
Mar-11

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

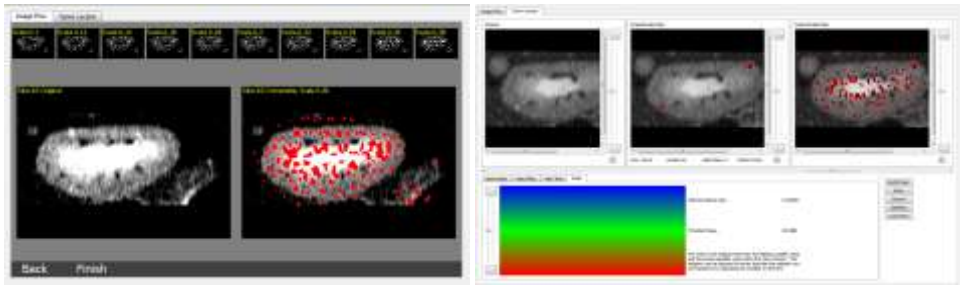
- “Time to solution” becomes critical when treatment decisions are based on technologies such as sequencing of tumor and patient.
- Must be “baked into” the decision support system workflow
- Simulations will be come part of the decision process within the next 10-15 years

Prototypical Oncology (Development) Workflow (coming ~~soon~~ now)

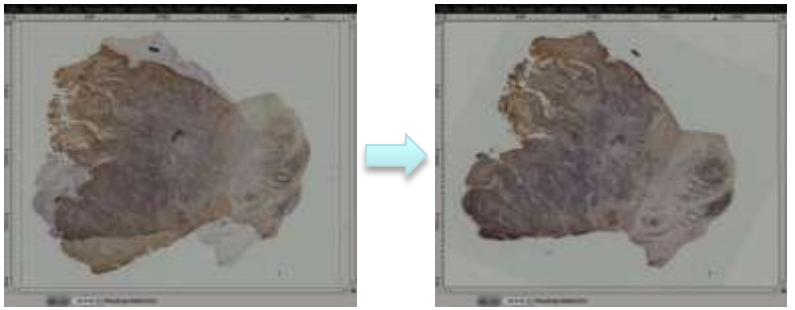
- Characterize tumor(s)
- Sequence tumor
- Target mutations
- Characterize at the molecular level
- Integrate all of the data
- Design and develop putative drug
- Test in animal models or clinical trials
- Treat patients

Characterize Tumor Images Analyzed

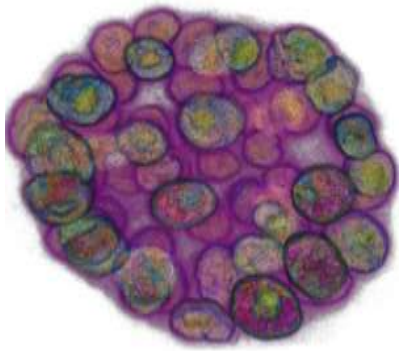
- Aperio Registration
- Quick2Insight
- Interface for SAIP met Segmentation
- Tumor Segmentation
- VDI



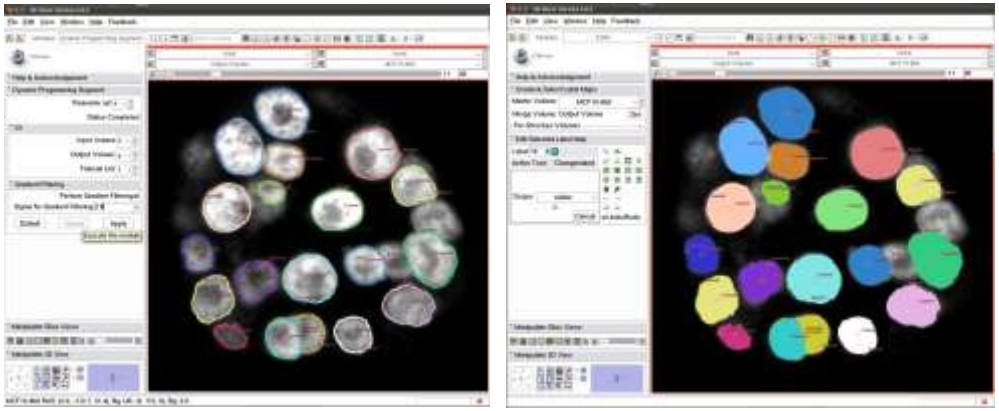
Common Imaging Tool Development for SAIP



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



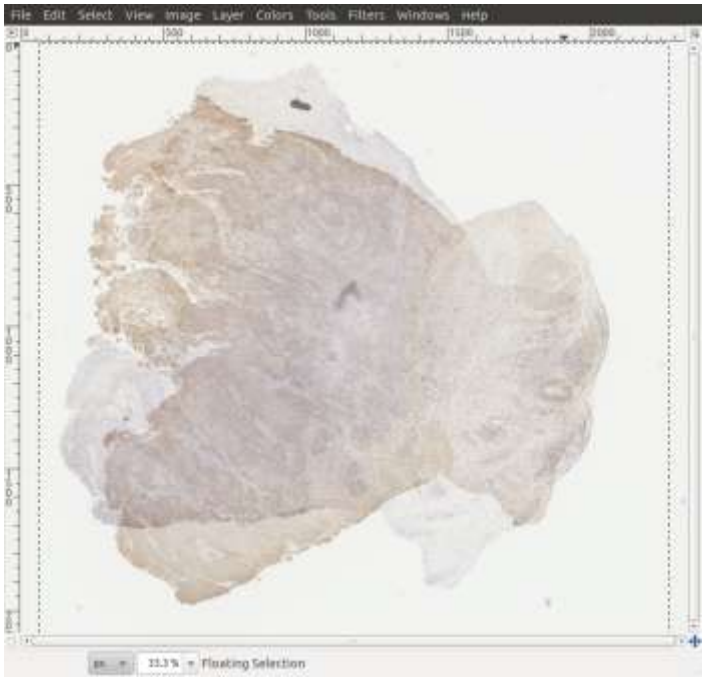
Automatic Visualization on 3D Biological Datasets

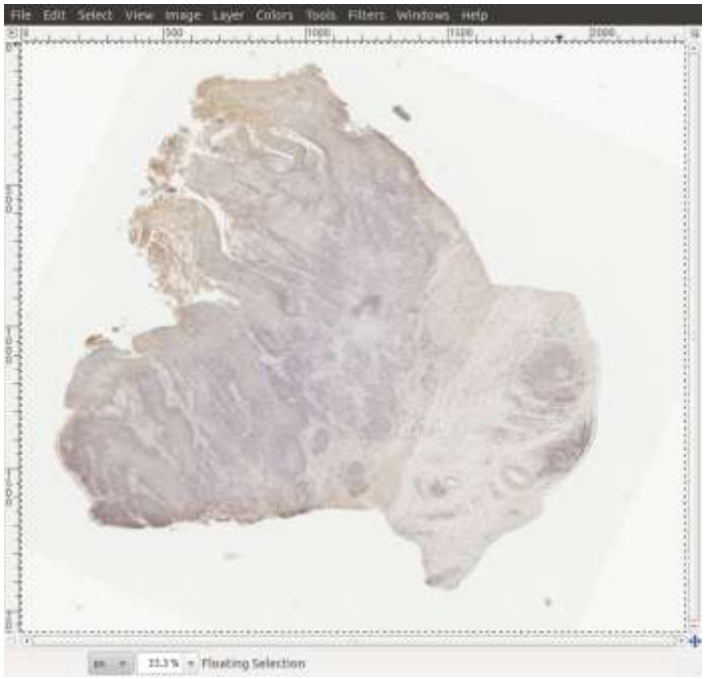


New Image Segmentation Module in Open Source Imaging Software 3D Slicer

Characterize Tumor
Pathologist Annotates Image

Characterize Tumor
Images Registered

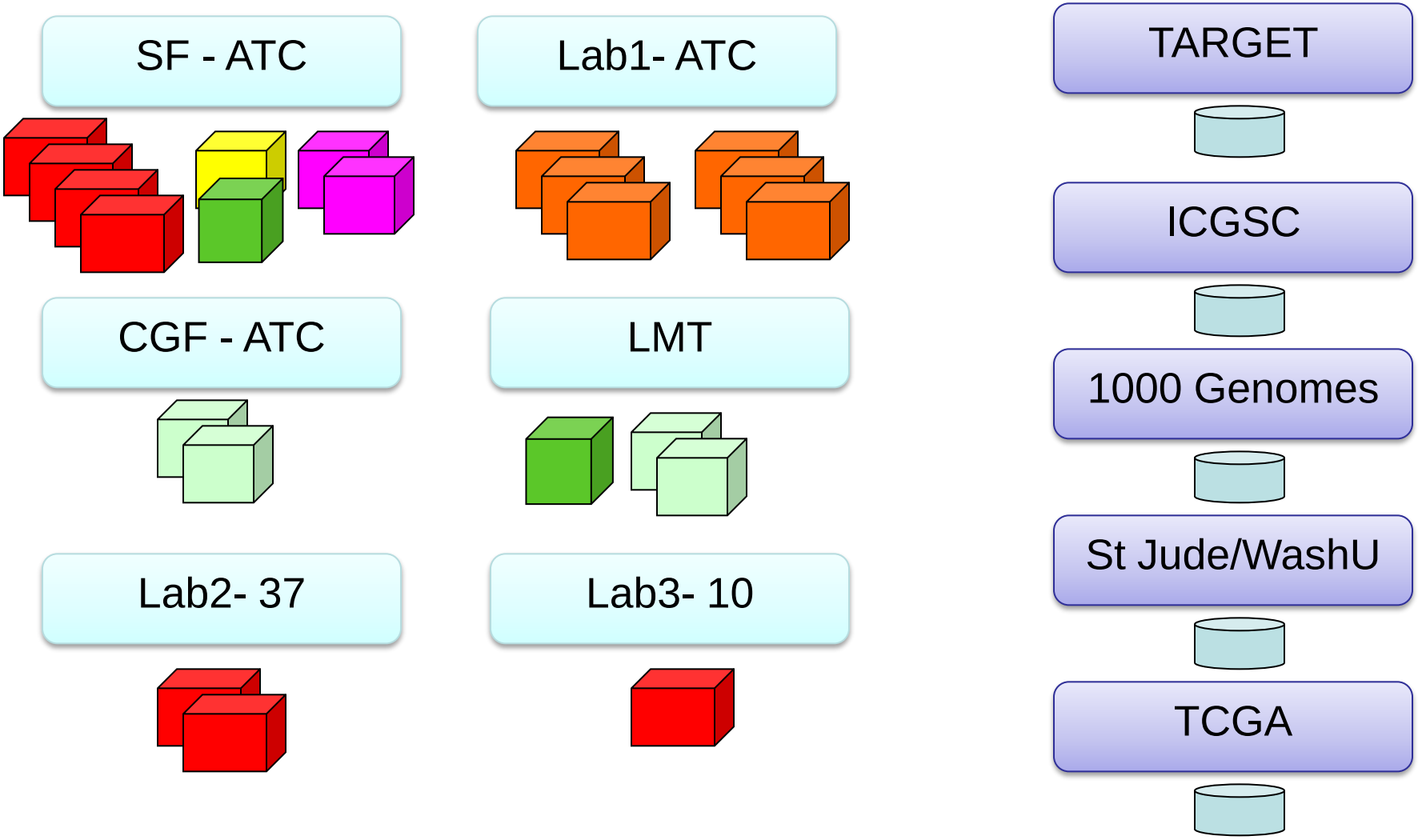




Optimizing People and Workflow
HPC Problem: Processing and Analysis

- 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

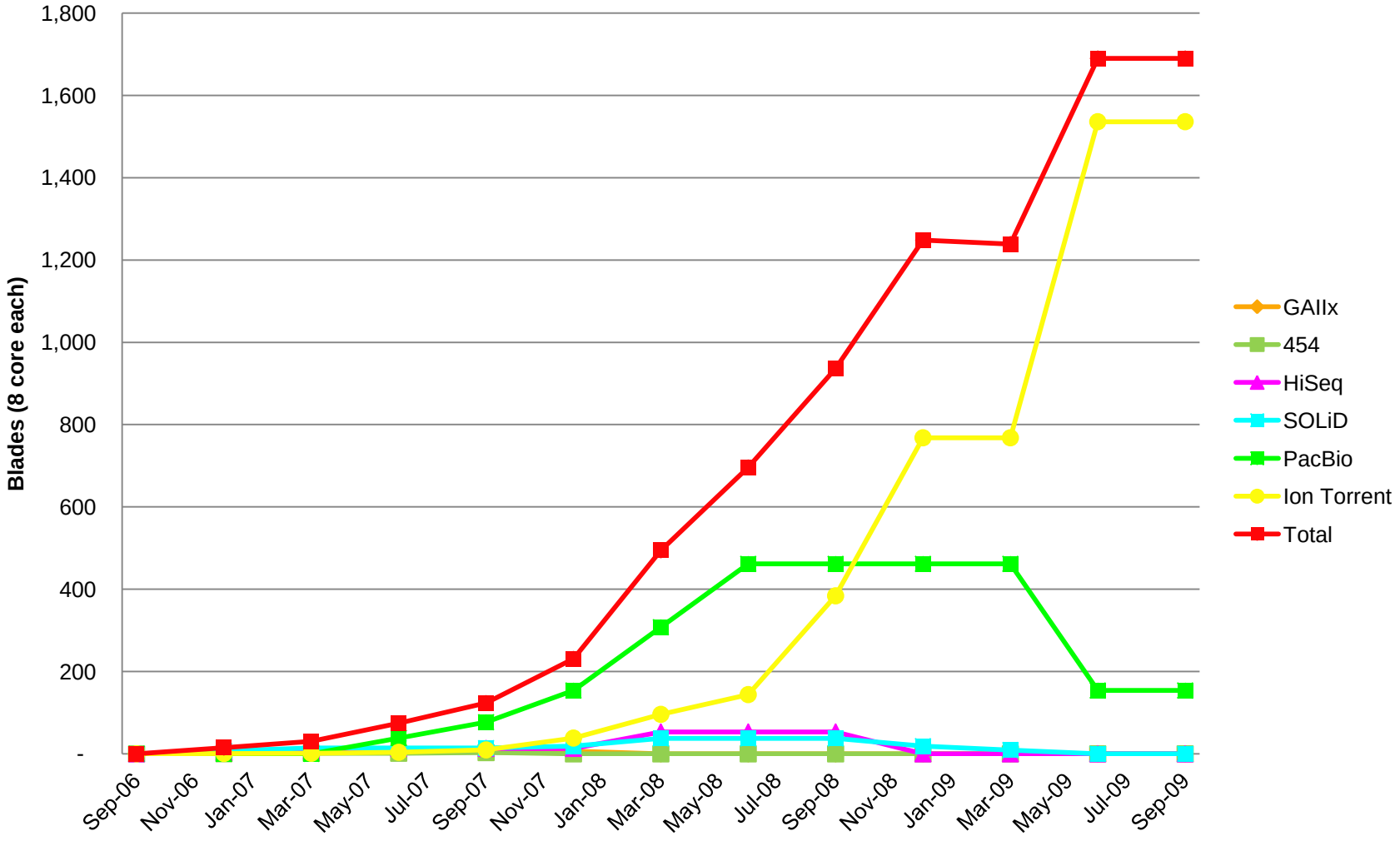


NGS Analysis

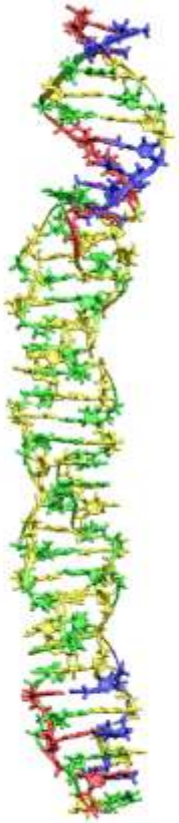
- **Dataset sizes are growing**
- **Time to generate the data is decreasing**
- **Analyses becoming more complex**
- **Data analyzed multiple times**

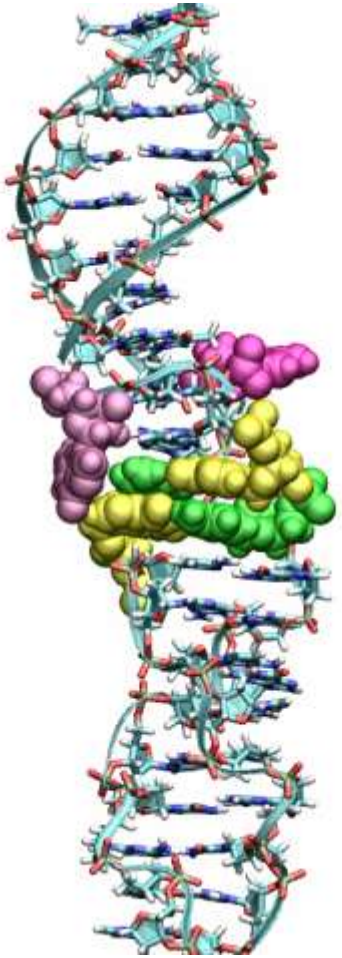
Analysis of NGS Data

(blades required to keep up with basic analysis of data)
(mapping and initial variation detection)



Targeting the Mutation



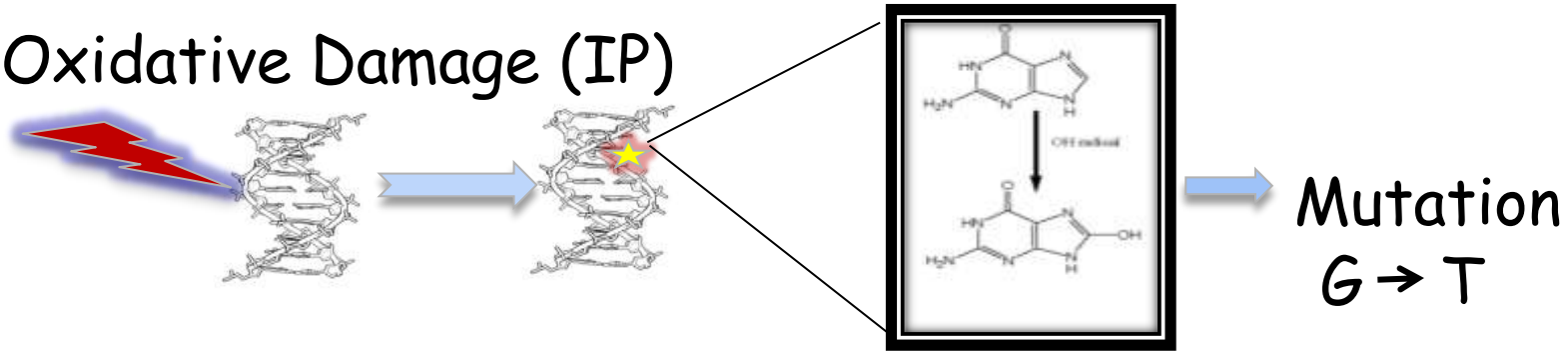


NCSA Resources for molecular elucidation in Biomarker/drug investigation

System	Local days using 16 cpus	ns/day	NCSA days using 96 cpus	ns/day
195000 atom 22ns	78.7	0.3	17.7	2.26
150000 atom 30ns	78.5	0.4	22.8	2.64

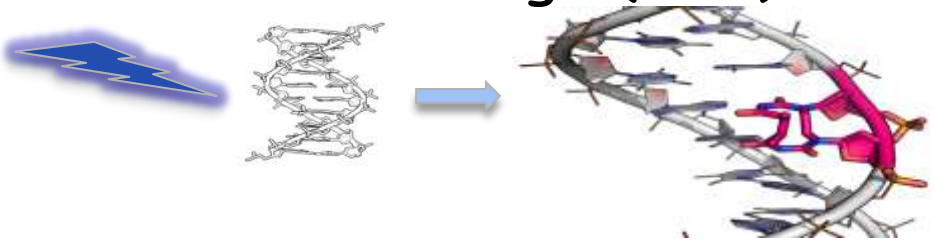
Modeling the Mechanism

Oxidative Damage (IP)



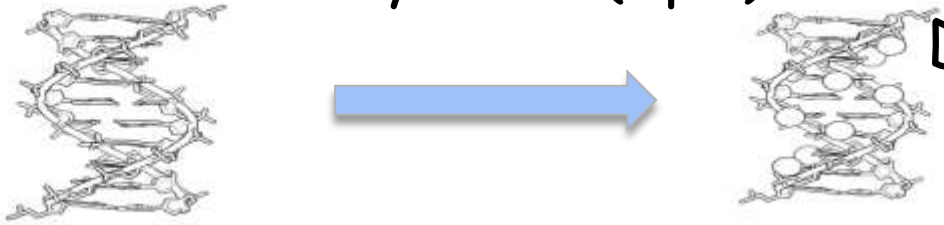
Mutation
G → T

Direct Photodamage (CPD)



Mutation
C → T

Methylation (CpG)

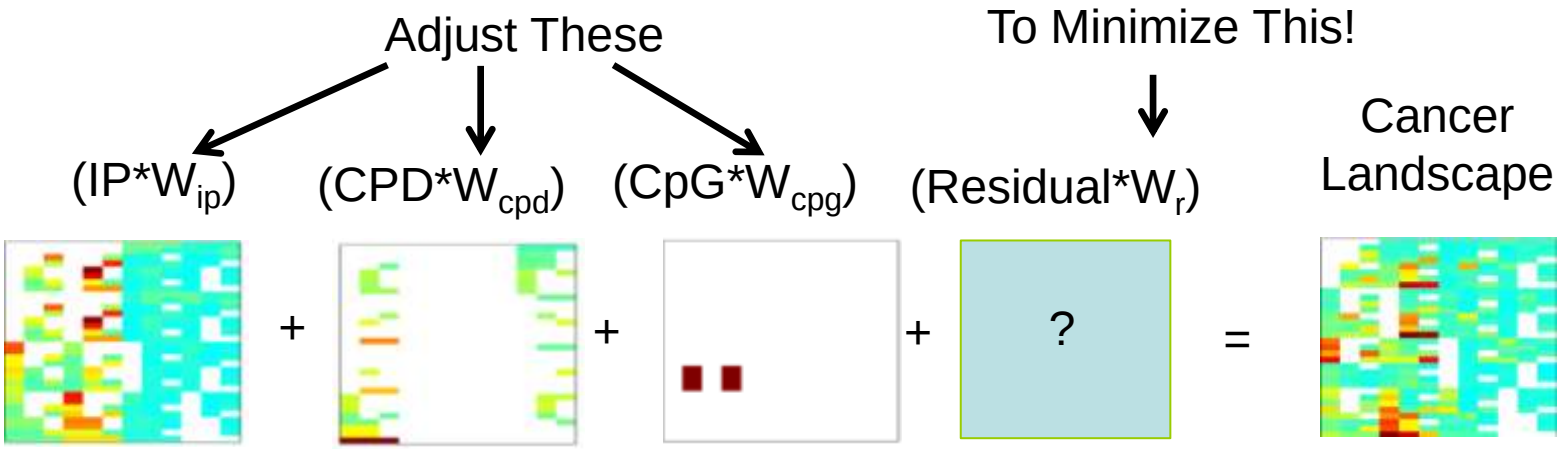


Deamination

Mutation
C →T

Calculate Contribution of Mechanism Templates to Cancer Landscapes

For each cancer:



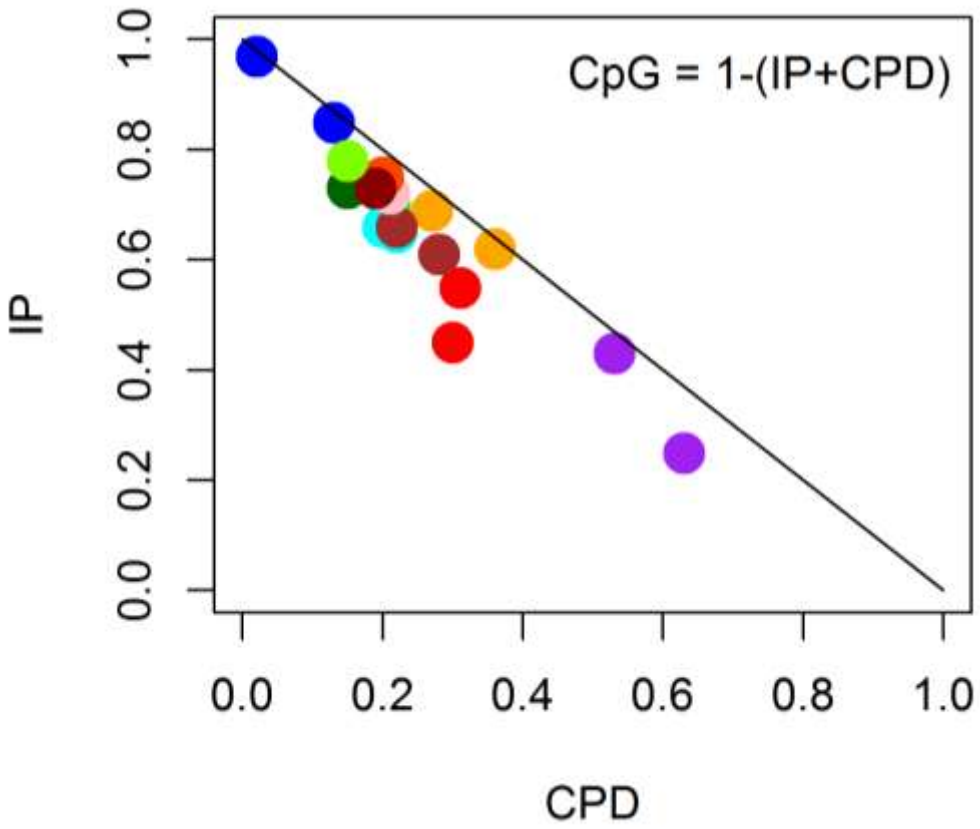
- Template weights are optimized based on an exhaustive search

Template Contributions

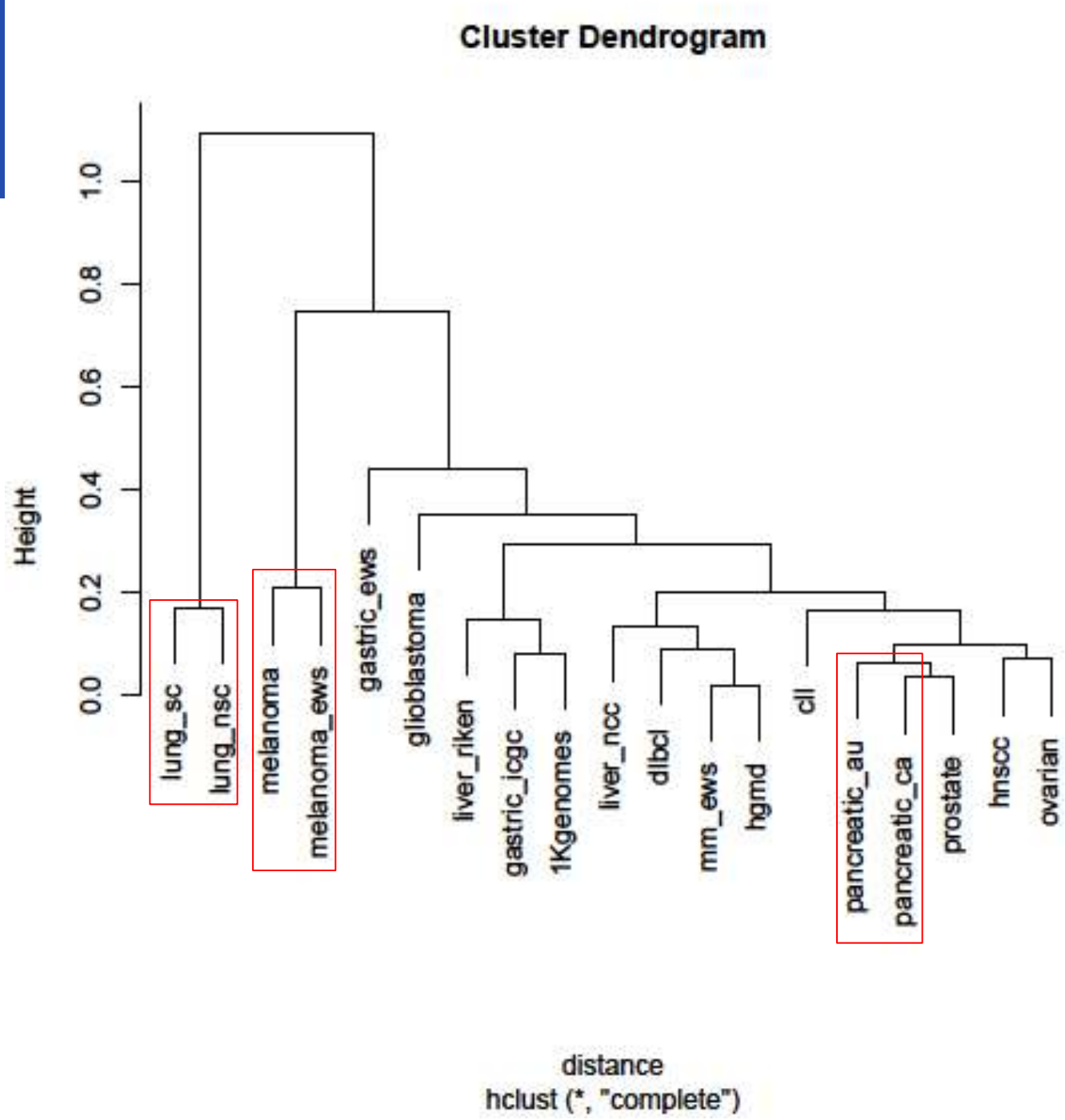
Calculated Contributions

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

Fitness Minimums



Cluster on Contributions



Template Contributions by Patient

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

liver_riken	**21.62%**	**36.63%**	**1.80%**	**39.95%**
liver_riken_RK001	24.24%	33.11%	1.77%	40.87%
liver_riken_RK002	24.99%	32.72%	1.78%	40.50%
liver_riken_RK003	23.89%	34.05%	1.79%	40.27%
liver_riken_RK006	22.90%	35.55%	1.81%	39.75%
liver_riken_RK010	22.88%	35.53%	1.81%	39.79%
liver_riken_RK015	22.73%	35.28%	1.79%	40.20%
liver_riken_RK023	21.69%	37.35%	1.20%	39.75%
liver_riken_RK024	21.70%	36.77%	1.81%	39.72%
liver_riken_RK026	21.17%	37.51%	1.81%	39.51%
liver_riken_RK029	21.14%	37.44%	1.81%	39.61%
liver_riken_RK034	21.21%	37.57%	1.82%	39.41%
liver_riken_RK042	21.16%	37.49%	1.81%	39.53%
liver_riken_RK046	21.55%	36.51%	1.80%	40.15%
liver_riken_RK063	21.62%	36.63%	1.80%	39.95%

liver_ncc	**18.21%**	**40.82%**	**3.77%**	**37.21%**
liver_ncc_HX4	18.91%	40.33%	3.78%	36.98%
liver_ncc_HX5	18.25%	40.90%	3.78%	37.07%
liver_ncc_HX9	18.21%	40.82%	3.77%	37.21%
liver_ncc_HX10	18.90%	42.20%	1.89%	37.02%
liver_ncc_HX11	18.35%	43.03%	1.90%	36.72%
liver_ncc_HX12	18.99%	44.54%	1.96%	34.50%
liver_ncc_HX13	19.21%	39.07%	5.76%	35.96%
liver_ncc_HX14	19.21%	39.69%	5.12%	35.98%
liver_ncc_HX15	18.59%	41.02%	4.49%	35.91%
liver_ncc_HX16	19.19%	40.30%	4.48%	36.04%
liver_ncc_HX17	18.32%	41.07%	3.79%	36.82%

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

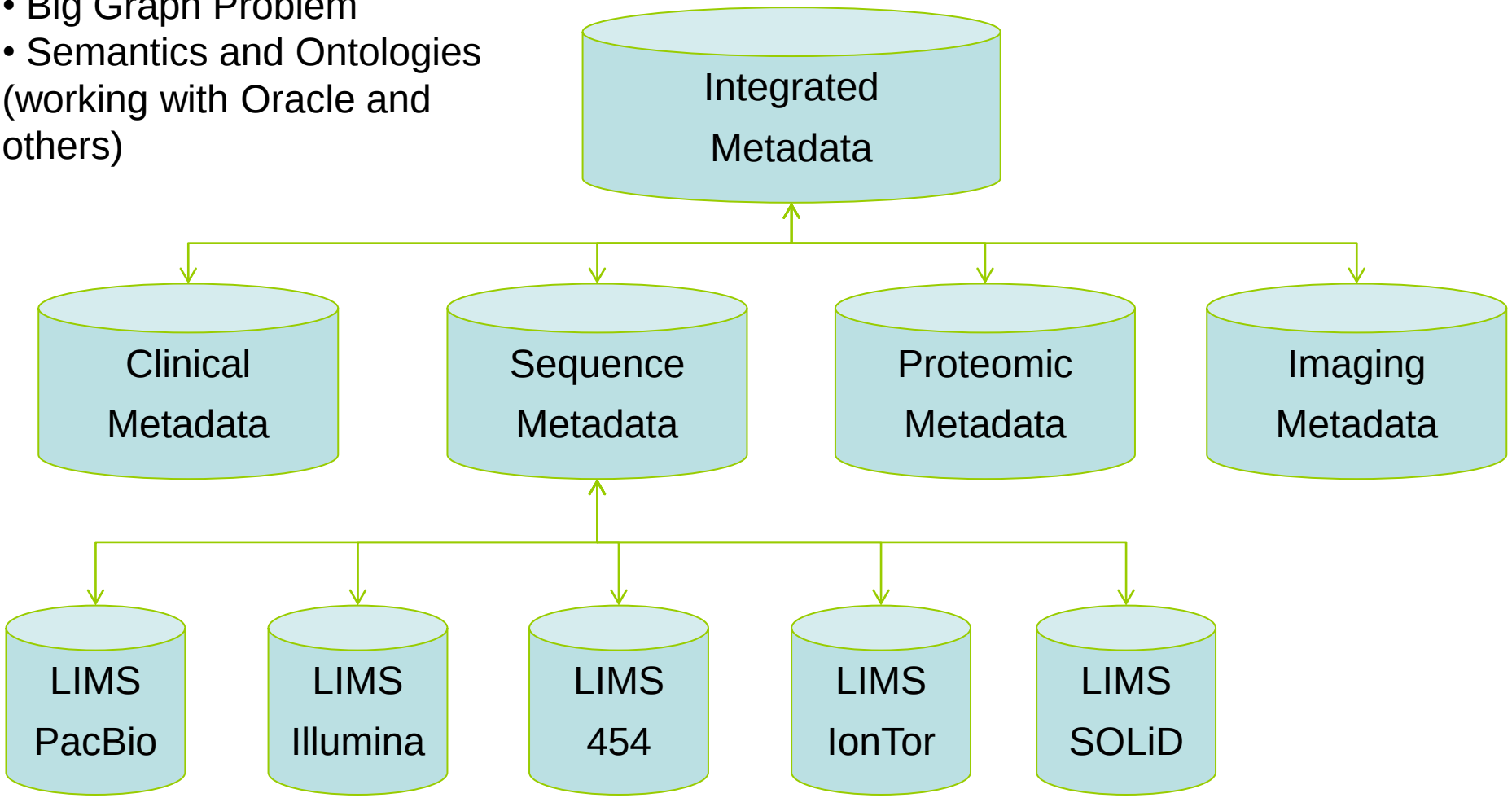
Gene	NCBI ID	Pphen-2-Pred	Pphen-2 Score		SIFT	
			HumDiv	HumVar	Prediction	score(median)
BRAF(D454E)	NP_004324.2	BEN	0.000 (sensitivity: 1.00; specificity: 0.00)	0.000 (sensitivity: 1.00; specificity: 0.00)	TOLERATED	1.00(3.18)
NCF2(H389Q)	NP_000424.2	PRD	**1.000 (sensitivity: 0.00; specificity: 1.00)**	**0.992 (sensitivity: 0.44; specificity: 0.97)**	TOLERATED	0.39(3.14)
G6PC3(Q272R)	NP_612396.1	BEN	0.005 (sensitivity: 0.97; specificity: 0.73)	0.005 (sensitivity: 0.97; specificity: 0.41)	TOLERATED	0.33(2.98)
KRAS(I36T)	NP_004976.2	BEN	0.009 (sensitivity: 0.97; specificity: 0.76)	0.044 (sensitivity: 0.94; specificity: 0.59)	**DAMAGING**	**0.02(3.04)**
PIK3CA(R524K)	NP_006209.2	BEN	0.000 (sensitivity: 1.00; specificity: 0.00)	0.002 (sensitivity: 0.99; specificity: 0.17)	TOLERATED	1.00(3.33)
SMAD4(E538K)	NP_005350.1	BEN	0.025 (sensitivity: 0.96; specificity: 0.80)	0.024 (sensitivity: 0.95; specificity: 0.54)	TOLERATED	0.21(3.22)
TP53(W23R)	NP_000537.3	POD	**0.999 (sensitivity: 0.14; specificity: 0.99)**	**0.997 (sensitivity: 0.24; specificity: 0.99)**	**DAMAGING**	**0.00(3.21)**
CYBB(S112P)	NP_000388.2	BEN	0.022 (sensitivity: 0.96; specificity: 0.80)	0.116 (sensitivity: 0.91; specificity: 0.67)	**DAMAGING**	**0.00(3.12)**
CYBA(Y72H)	NP_000092.2	PRD	**1.000 (sensitivity: 0.00; specificity: 1.00)**	**0.996 (sensitivity: 0.32; specificity: 0.98)**	TOLERATED	0.41(3.41)
CDKN2A(R87P)	NP_000068.1	PRD	**1.000 (sensitivity: 0.00; specificity: 1.00)**	**0.999 (sensitivity: 0.08; specificity: 1.00)**	TOLERATED	0.22(1.80)

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.

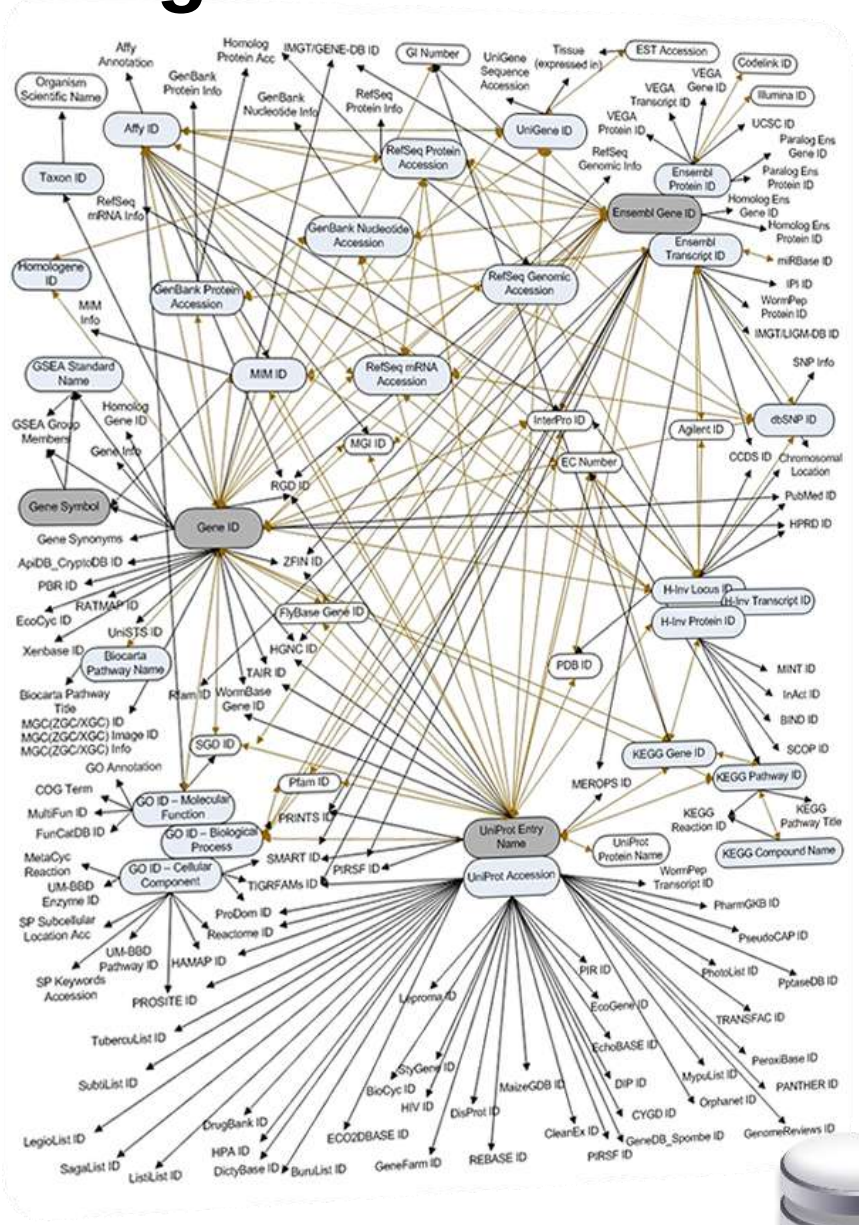
HumVar, consisted of all human disease-causing mutations from UniProtKB, together with common human nsSNPs (MAF>1%) without annotated involvement in disease, which were treated as non-damaging.

Mine and Integrate Multiple Data Sources "Systems Biology"

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



Integrate Context of Data



bioDBnet – **Bio**logical **D**ata**B**ase **Net**work

- 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

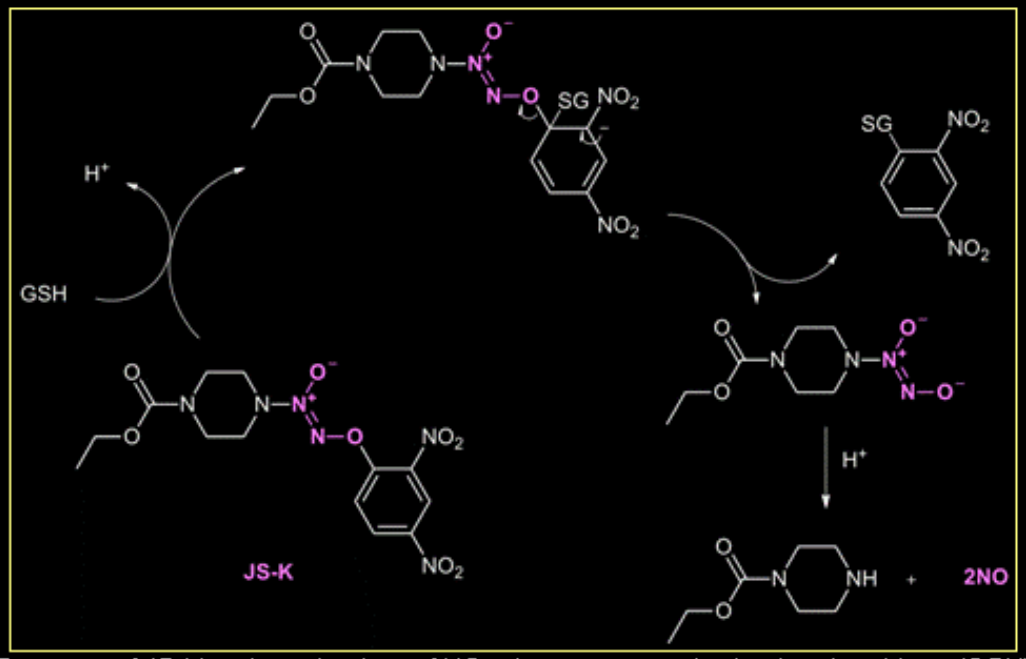
URL: http://biodbnet.abcc.ncifcrf.gov

Mudunuri U, Che A, Yi M, Stephens RM. *Bioinformatics*. 2009 Feb 15;25(4):555-6.

Characterize Therapies at Molecular Level

JS-K AND RELATED O^2-ARYLATED DIAZENIUMDIOLATES AS BROAD-SPECTRUM ANTICANCER AGENTS

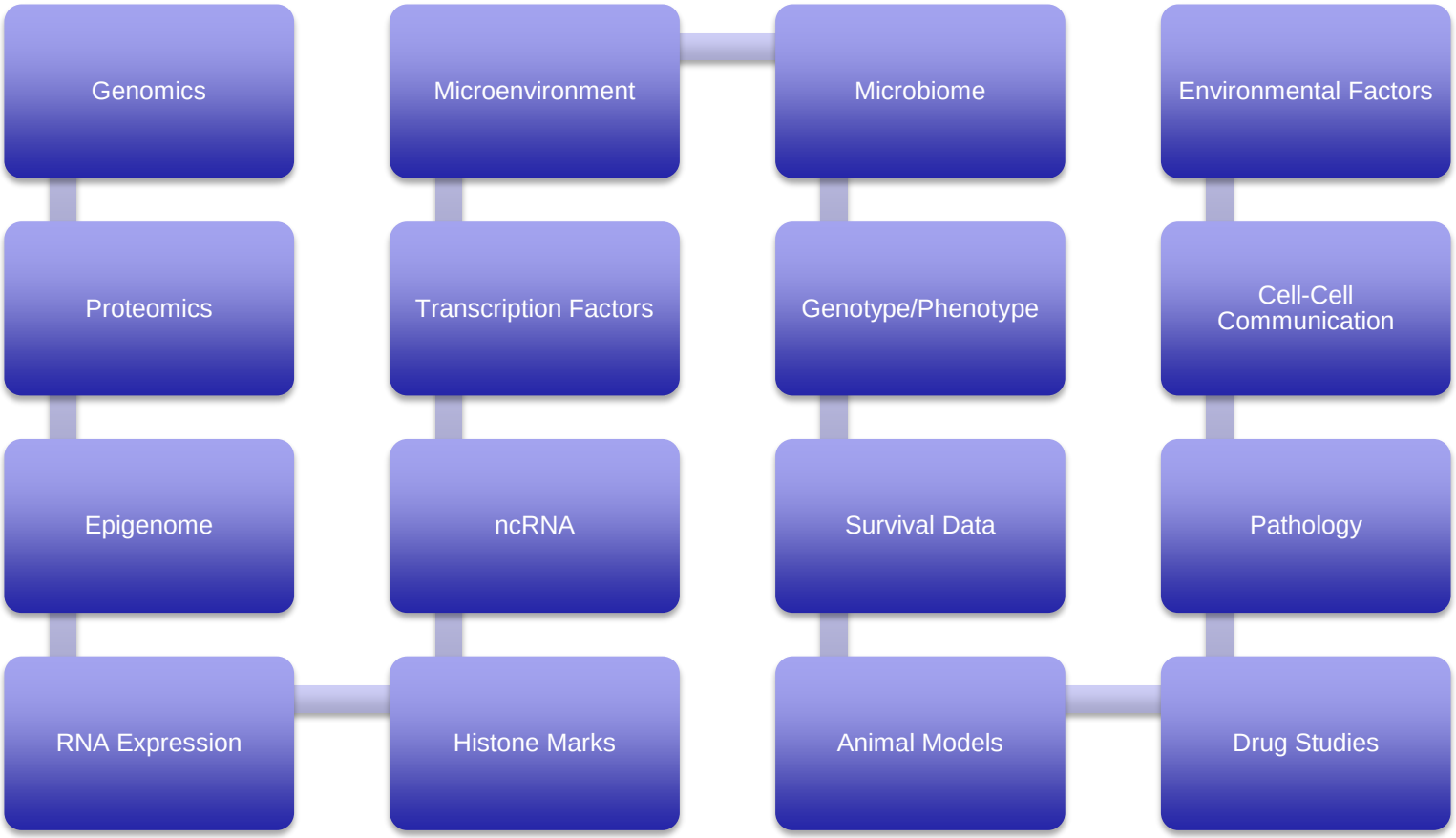
NO–releasing prodrugs of the O^2-arylated diazeniumdiolate class have shown themselves to be increasingly promising broad-spectrum anti-cancer drug candidates. Our lead compound JS-K has slowed tumor growth in several rodent models of cancer, including **leukemia**, **prostate cancer**, **multiple myeloma**, 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.



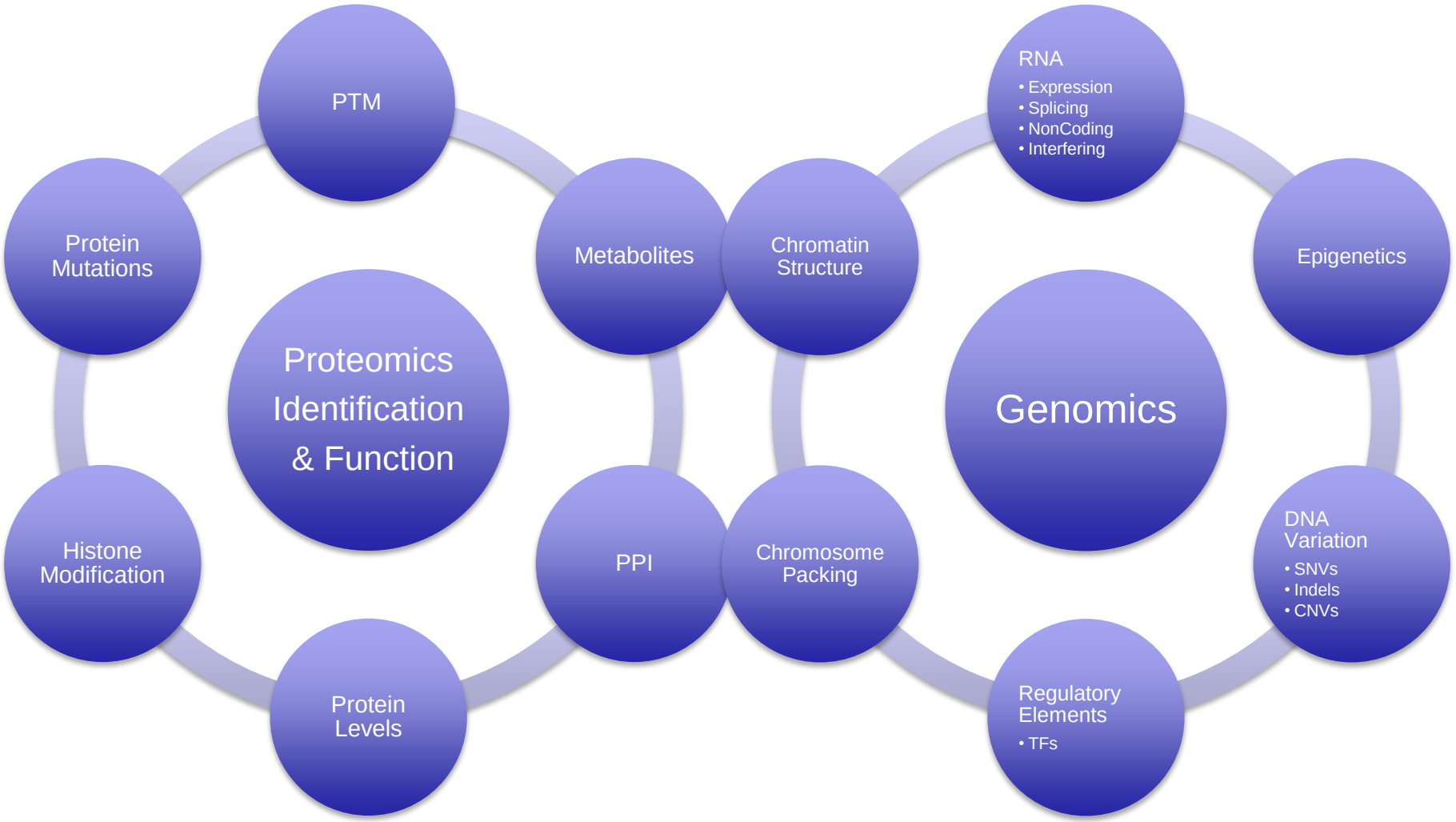
Structure of JS-K and mechanism of NO release upon activation by glutathione (GSH)

http://star.ncifcrf.gov/nitricoxide/

So, what are we missing?



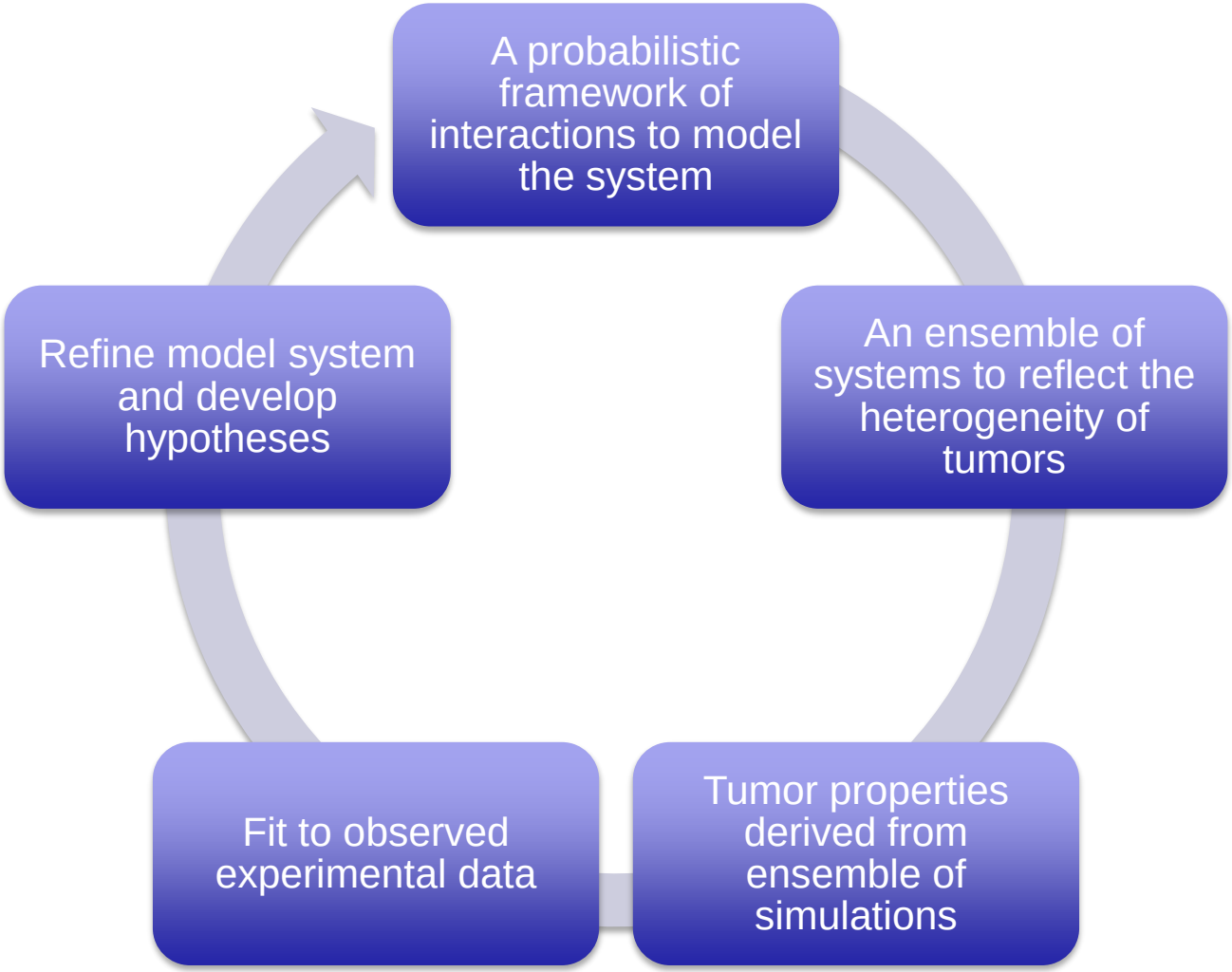
Need function of interacting systems



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



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?