National Cancer Institute

Status Report:
caBIG® Accomplishments

National Cancer Advisory Board Meeting

September 8, 2010

Ken Buetow, Ph.D.

Director, Center for Bioinformatics and Information Technology

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

caBIG® at a glance

Community

- 2,300+ participants from more than 700 institutions
  - 56 NCI-designated Centers
  - 30 Community Centers
- 1000+ registrants for the 2010 caBIG® Annual Meeting
- 19 licensed Support Service Providers to sustain the biomedical community as they deploy caBIG® tools and technology (20% increase)
- 15 countries using caBIG® tools and technology to facilitate

Connectivity

- 78 applications supporting full continuum of biomedical research (10% increase)
- 149 “nodes” connected to National Grid via caGrid (25% increase)

Content

- 2.17 million biospecimens available through caGrid (82% increase)
- 4.76 million images stored in the National Biomedical Imaging Archive (28% increase)
- 39,952 microarray experiments available for research use on caGrid (60% increase)

Background

caBIG® Origins

- **circa 2002: NCAB charges NCI with developing a strategy to assist the Cancer Community with the increasingly unmanageable informatics challenges associated with clinical research, biospecimen management, and molecular analysis**

- **Late 2003 through 2004: NCI visits NCI-designated Cancer Centers to assess needs, capabilities, requirements, and resources**

- **2004: NCI launches caBIG® pilot program in NCI-designated Cancer Centers**

- **2007: NCI expands/adopts caBIG® as enterprise framework**

Partial List of NCI Programs Added since caBIG® Launch

- **Mouse Models in Human Cancer (MMHCC)**
- **Integrated Cancer Biology Program (ICBP)**
- **Interagency Oncology Taskforce**
- **Clinical Proteomic Technologies for Cancer Initiative**
- **Nanotechology Alliance**
- **Specialized Programs Of Research Excellence (SPOREs)**
- **Strategic Partnering to Evaluate Cancer Signatures (SPECS)**
- **Cancer Genetic Markers of Susceptibility (CGEMS)**
- **Glioblastoma Multiform Diagnostic Initiative (REMBRANDT)**
- **Clinical Trials Working Group**
- **Translational Research Working Group**
- **The Cancer Genome Atlas (TCGA)**
- **Therapeutically Applicable Research to Generate Effective Treatments (TARGET) Initiative**
- **Office of Biorepositories and Biospecimen Research (caHuB)**
- **NCI Community Cancer Centers Program**
- **(others)**

The Cancer Community Sought To*…

- **Manage/leverage local and publicly-accessible biomedical data for research**
- **Connect/streamline workflows**
- **Increase accuracy of processes**
- **Standardize/Streamline data collection**
- **Perform complex analysis across data sets**
- **Identify best practices**
- **Share data**

THE WISTAR INSTITUTE
TODAY'S DISCOVERIES – TOMORROW'S CURES

MAYO CLINIC

UCSF Helen Diller Family Comprehensive Cancer Center

Dan L. Duncan Cancer Center at Baylor College of Medicine

NCI·CC
A Cancer Center Designated by the National Cancer Institute

Jefferson | Kimmel Cancer Center NCI-designated

Memorial Sloan-Kettering Cancer Center

Dartmouth-Hitchcock
NORRIS COTTON CANCER CENTER
A National Cancer Institute Comprehensive Cancer Center

The Jackson Laboratory

H. LEE MOFFITT Cancer Center & Research Institute

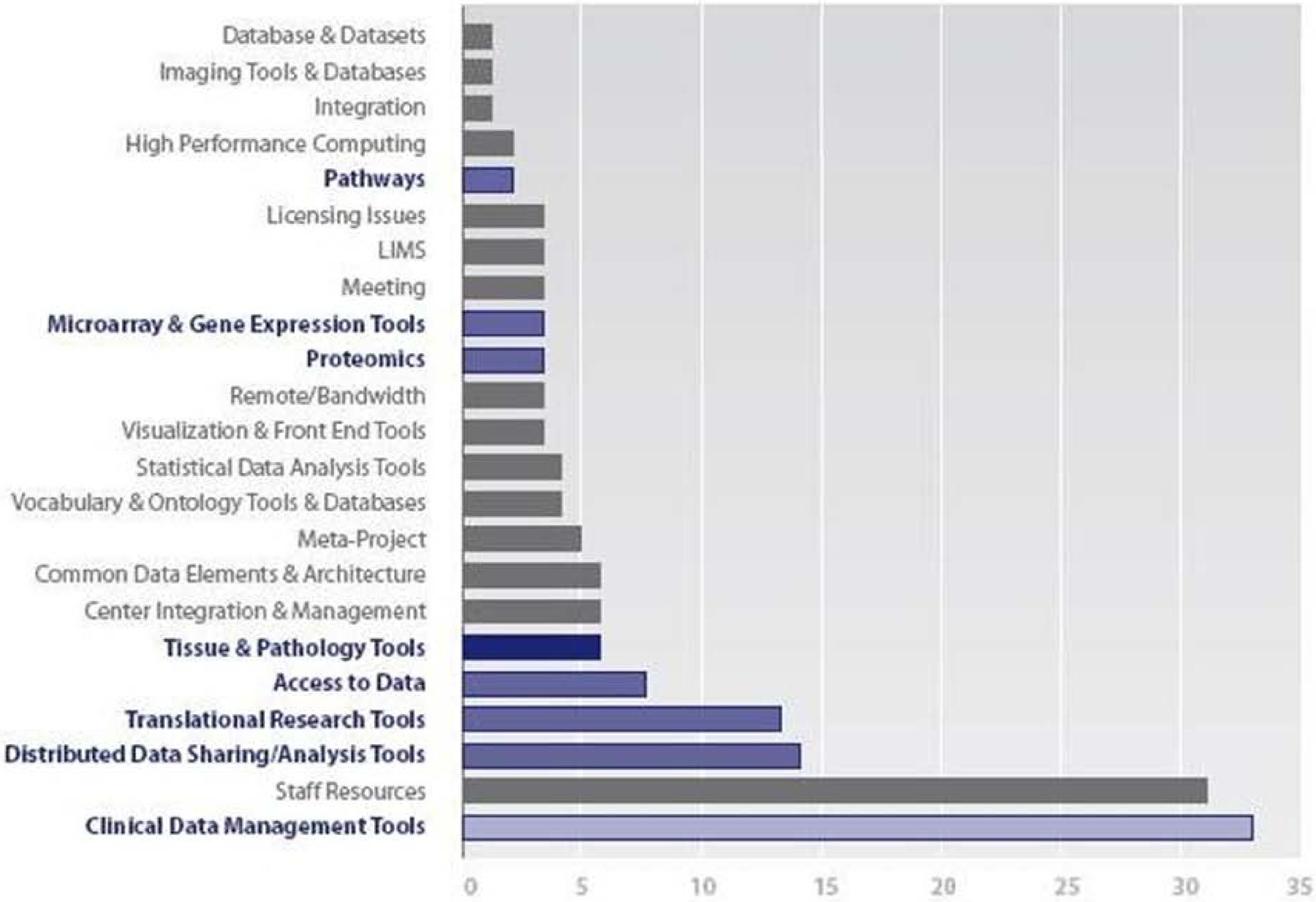
THE OHIO STATE UNIVERSITY
Comprehensive Cancer Center

DF/HCC
DANA-FARBER / HARVARD CANCER CENTER

JOHNS HOPKINS MEDICINE

** Based on 2004 Cancer Centers survey*

Cancer Center Needs



Cancer Center involvement identified priority areas for caBIG™

Cancer Community Landscape (circa 2003)

- **Integrated Systems**
  - Homegrown/ Commercial
  - Smooth navigation between applications
  - Difficult to expand/extend
  - Large IT staff
  - $10M’s invested
- **Heterogeneous Systems**
  - Complex mix of commercial and homegrown components (may be composed of dozens of components)
  - No common interfaces
  - Medium size IT staff
  - $1M’s invested
- **Informal/No Systems**
  - Use of productivity applications (e.g. Excel, Access)
  - Complex manual processes
  - Small or no IT staff
  - $100K’s invested

caBIG® Strategy at Launch

- **Community**
  - Establish an **open community** of participants from the spectrum of disciplines, geographies, types of institutions, etc.
  - Facilitate the work of others who are building capabilities
  - Adopt a **“federated”** model to allow local control of sharing and partnerships and to support individual labs and institutions
- **Content**
  - Facilitate access to rich primary data
  - Leverage existing academic and commercial software, wherever possible, to avoid unnecessary time and expense
  - Invest primarily in **open source** tools that the community does not have
- **Connectivity**
  - Recognize legacy IT systems to avoid “rip and replace” costs
  - Wherever feasible, make disparate applications compatible for “plug-and-play” compatibility and data-sharing through **standards-based interoperable infrastructure**

Community

caBIG® Operational Approach

Domain-level

Clinical Trials Management Systems Workspace (CTMS)

Integrative Cancer Research Workspace (ICR)

***In Vivo* Imaging Workspace (IMAG)**

Tissue Banks & Pathology Tools Workspace (TBPT)

Strategic-level

Strategic Planning Workspace (SP)

Training Workspace (D&T)

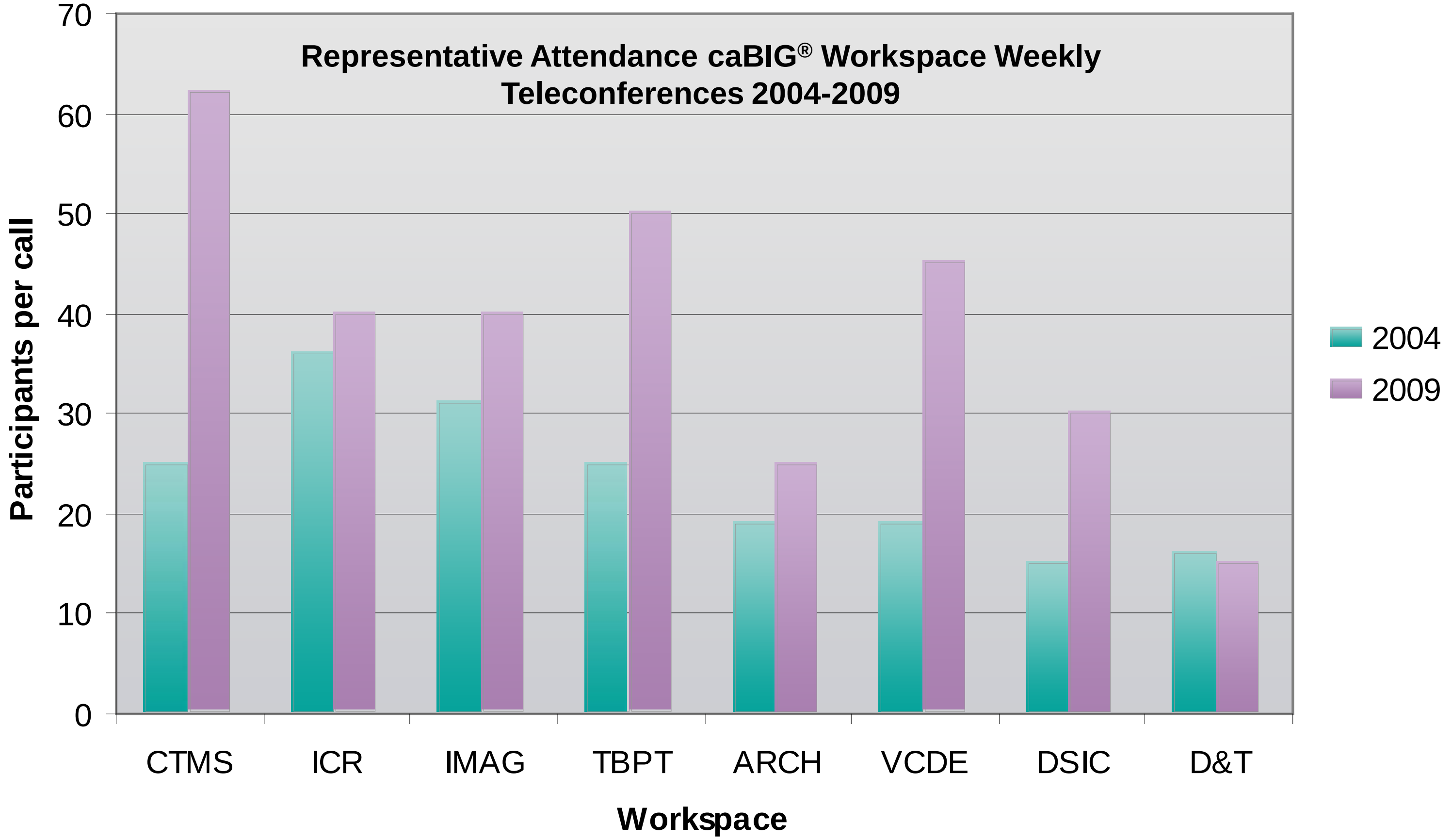
Data Sharing & Intellectual Capital Workspace (DSIC)

Cross-cutting

caBIG® Vocabularies and Common Data Elements Workspace (VCDE)

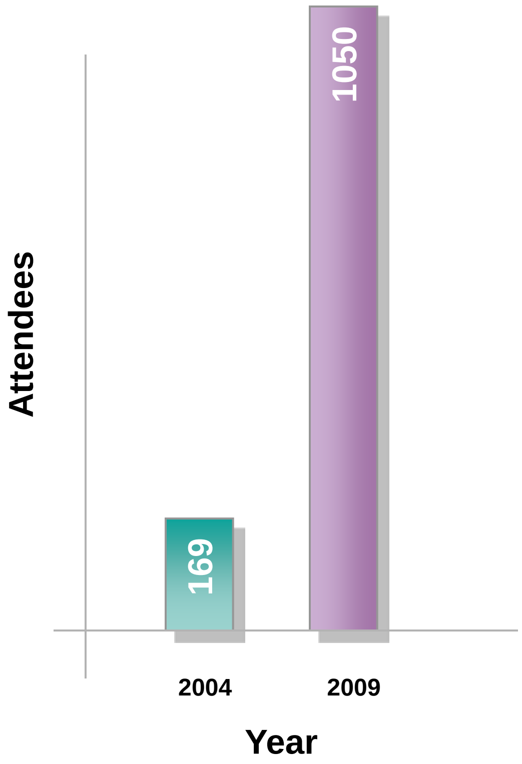
caBIG® Architecture Workspace (ARCH)

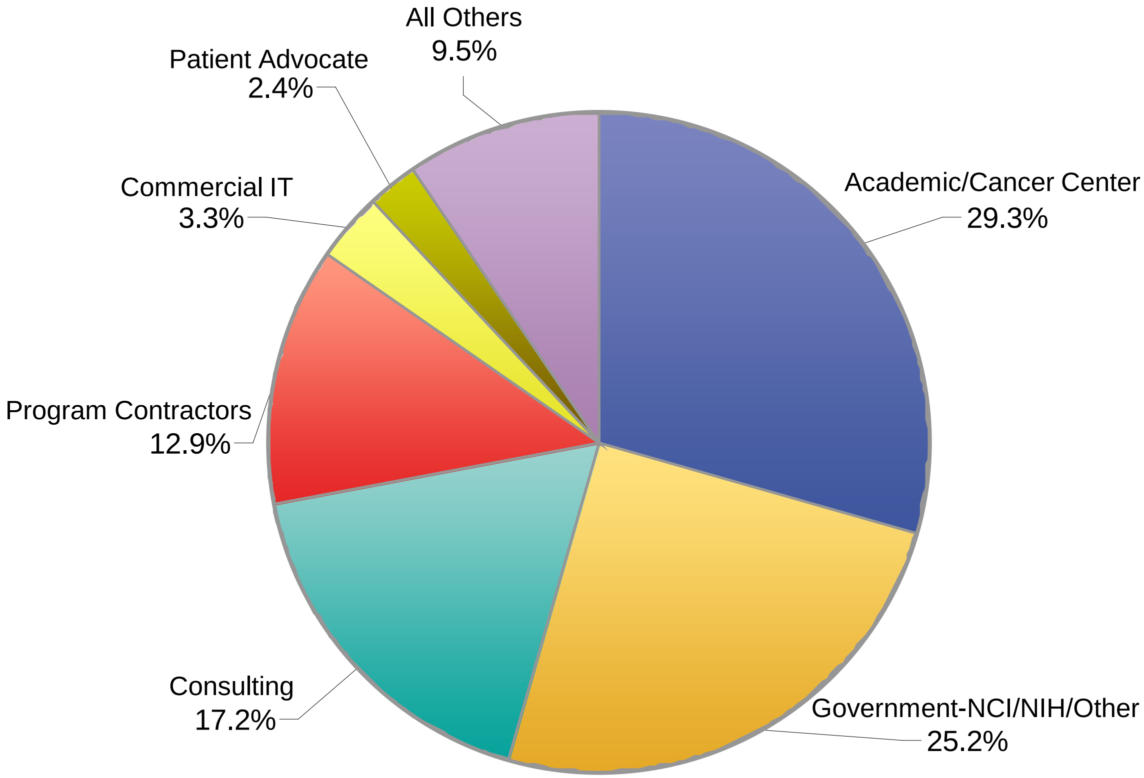
caBIG® and the Cancer Research Community



Representative Attendance caBIG® Workspace Weekly Teleconferences 2004-2009

caBIG® Annual Meeting Attendance





1050 people attended the 2009 caBIG® Annual Meeting, representing 425 organizations. 63% of those attendees, and 60% of their organizations were new in 2009 vs. 2008.

Supporting the Community

- **Knowledge Centers (KC)**
  - Domain expertise, technical and end-user documentation
  - Comprehensive, up-to-date installation packages for caBIG® tools
  - Administration of open source development of caBIG® tools
  - Collection and monitoring of defect reports, feature requests, and end-user requirements
  - Training materials for caBIG®
  - 6 Domain-focused Knowledge Centers: *caGrid, Vocabulary, Clinical Trials Management Systems, Molecular Analysis Tools, Tissue Banks, Data Sharing and Intellectual Capital*

- **Support Service Providers (SSP)**
  - Software application and technical support to end-users and IT professionals
  - 19 licensed SSPs to-date, quarterly solicitations

caBIG® Support Service Providers

- 5AM Solutions, Inc.
- Akaza Research
- Asclepius Solutions
- CTIS, Inc.
- E-SAC, Inc.
- Ekagra Software Technologies
- HealthCare IT, Inc. (HCIT)
- IMS, Inc.
- INFOTECH Soft, Inc.
- LabAnswer
- Moxie Informatics
- Persistent Systems
- Recombinant Data
- SAIC
- ScenPro, Inc.
- SemanticBits
- SRA Corporation
- TerpSys
- University of Utah

5AM | SOLUTIONS

Akaza research

PERSISTENT

Asclepius Solutions

CTIS | A HEALTH INFORMATICS COMPANY
Proven Solutions. Improved Results.

E-SAC
Enterprise Solutions And Consulting

ekagrasoft.com

HealthCare IT, Inc.

INFOTECHSoft
Innovative Healthcare Solutions

IMS

Recombinant
HEALTHCARE DATA WAREHOUSING

Lab Answer

TERPSYS®

Moxie Informatics
Know-How Solutions™

SemanticBits

SCENPRO

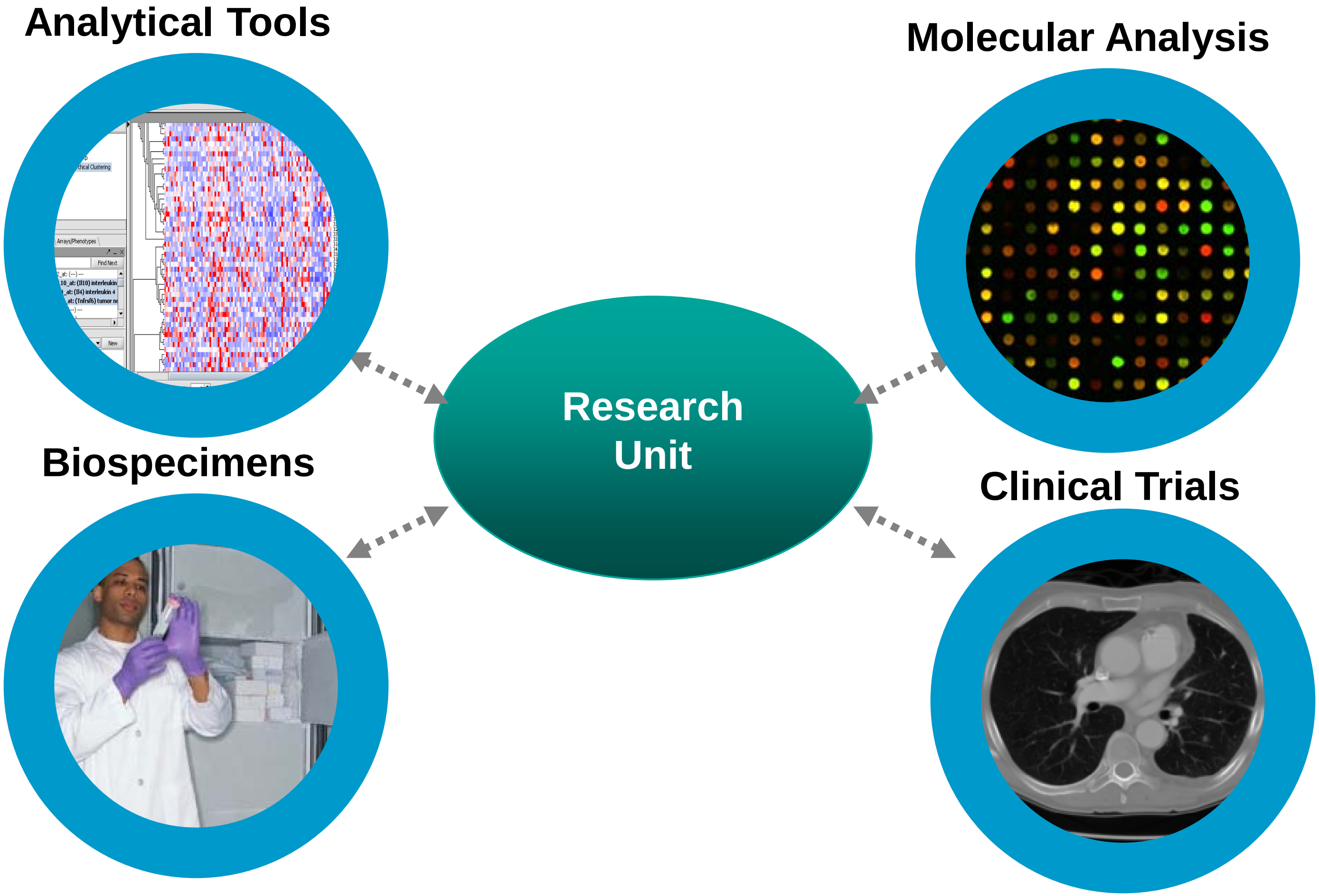
SRA
INTERNATIONAL, INC.

SAIC®
From Science to Solutions

HUNTSMAN
CANCER INSTITUTE
UNIVERSITY OF UTAH

Connectivity and Content

Supporting Individual and Institutional Needs



Analytical Tools

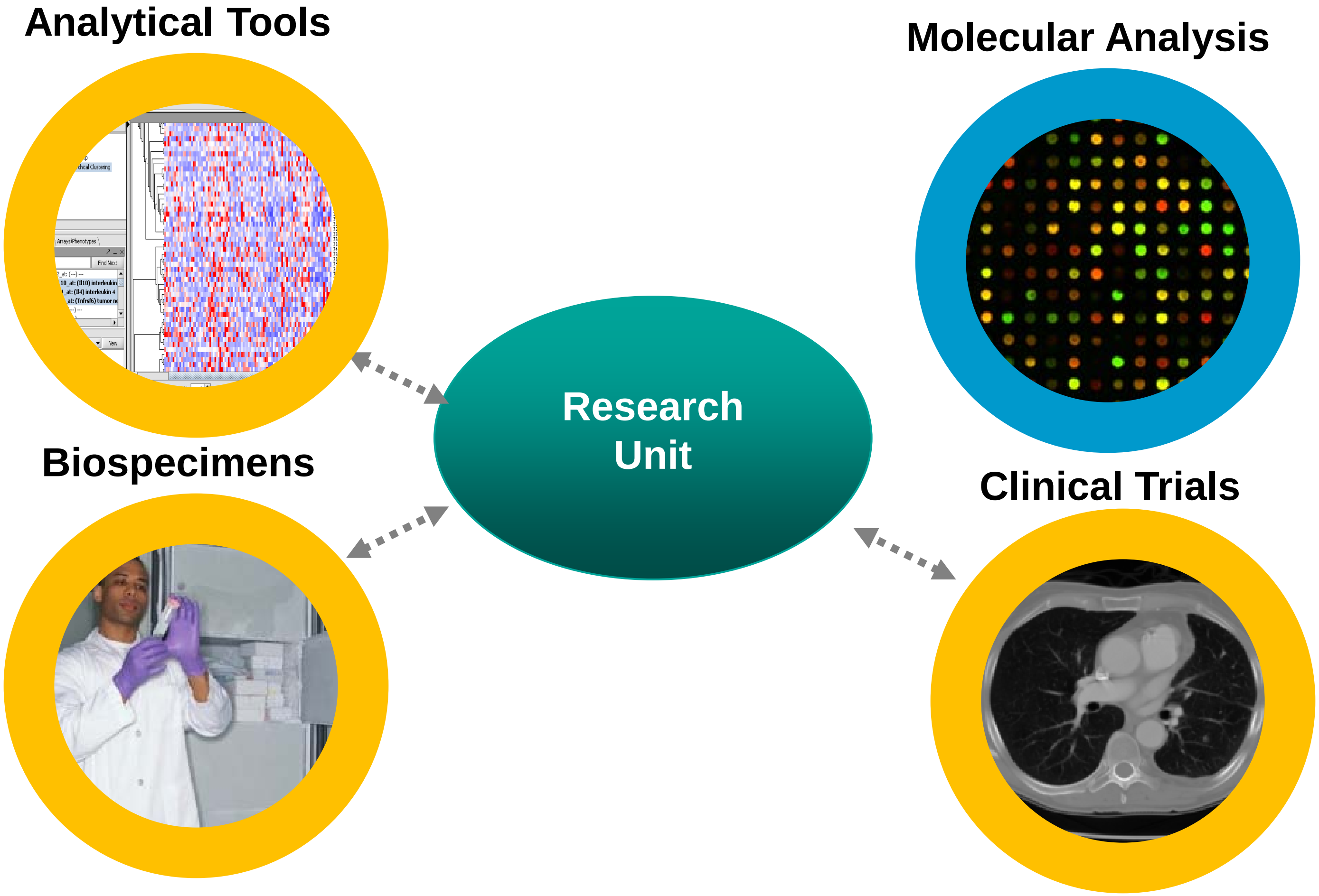
Molecular Analysis

Research Unit

Biospecimens

Clinical Trials

Supporting Individual and Institutional Needs



Analytical Tools

Molecular Analysis

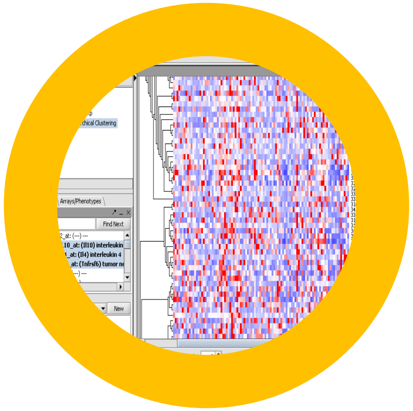
Research Unit

Biospecimens

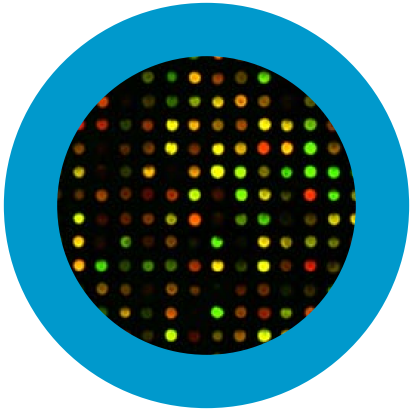
Clinical Trials

Supporting Individual and Institutional Needs

Analytical Tools



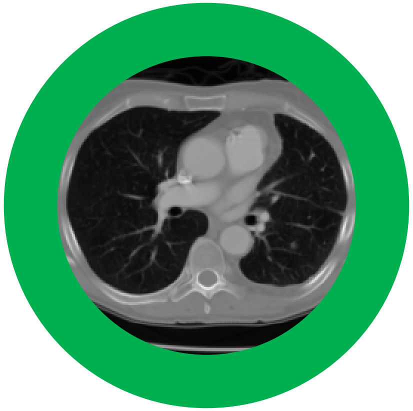
Molecular Analysis



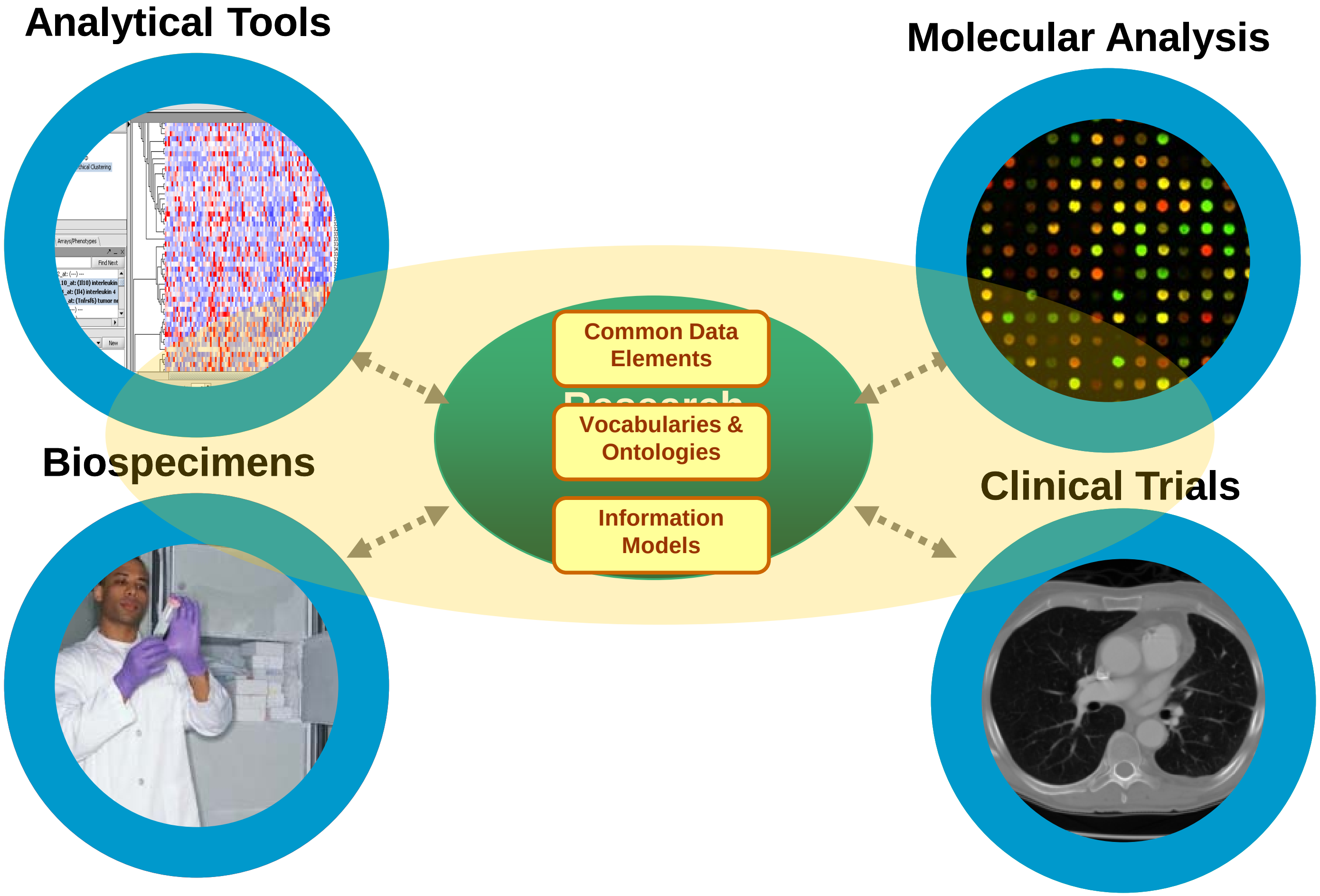
Research Unit

Biospecimens

Clinical Trials



Supporting Individual and Institutional Needs



Analytical Tools

Molecular Analysis

Common Data Elements

Research

Vocabularies & Ontologies

Information Models

Biospecimens

Clinical Trials

Clinical Trials Management Systems Workspace

Clinical Trials Management Systems Workspace (CTMS)

Patterns of Usage:

- **Complete end-to-end clinical trials management capability**
- ***Specific capability* to fill a gap in existing capabilities**
- ***Linkage tools* (caGrid, data standards) to connect existing in-house or commercial applications**
- ***Specifications* to build or modify in-house developed applications for interoperability**

caBIG® Capabilities:

- Collect clinical trials data - **C3D (Oracle Clinical)**
- Enroll, register, and track clinical trial participants across multiple sites - **C3PR**
- Create and manage clinical trial participant schedules and activities - **PSC**
- Create new protocols and track throughout the lifecycle of a study - **PLT**
- Store, browse,clinical laboratory data; share with other systems - **Lab Viewer**
- Collect and report adverse events - **caAERS**
- Connect systems and support clinical trials workflow integration; provide interoperability associated with SOA - **Integration Hub**
- Interoperate / Share data with 3rd party CDMS systems - **Clinical Connector**
- Exchange of clinical trials data across multiple systems - **CTODS**
- Help investigators comply with Federal registration requirements - **FIREBIRD**
- Recruit patients for clinical trials - **caMATCH**
- Provide common data elements and case report forms for standardization and reusability - **CRF Project**

Clinical Trials Management Systems Statistics

Institutions using caBIG® Clinical Trials Applications - by Stage:

Production	Preproduction	Pilot	Installation	Total
31	3	3	16	53

Additional CTMS Tools used by Cancer Centers and other Organizations

Tool	Total
Business Analysis Model (BAM)	8
Access defined semantics for protocol-driven research - Biomedical Research Integrated Domain Group model (BRIDG)	9*
Report clinical trials results to the NCI – Clinical Trials Reporting Program (CTRP)	37

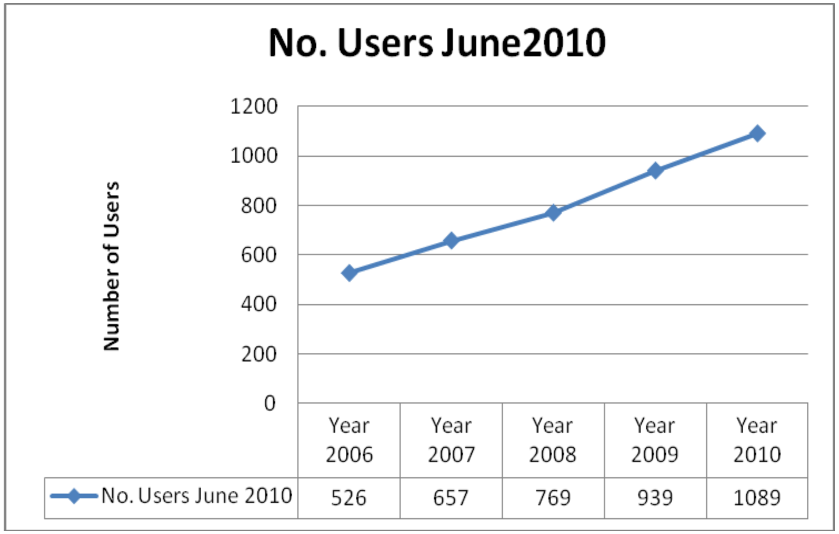
*Users include HL7 & CDISC, therefore reaching a large group

caBIG® Electronic Data Capture (EDC)

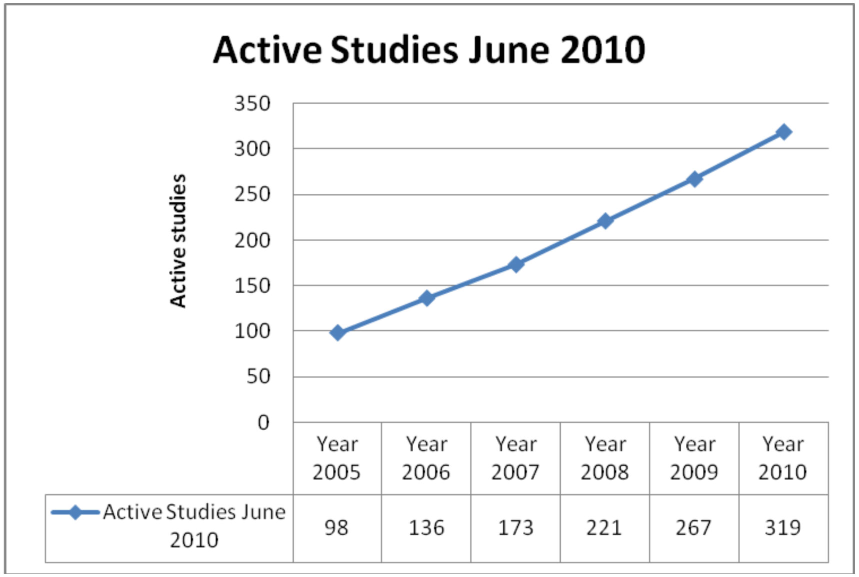
EDC Usage Statistics:

- **194 sites are using C3D**
  - 23 primary sites
  - 171 satellite sites
- **1000+ active C3D users across those sites**
- **1+ million Case Report Forms (CRFs) have been entered**
  - Over 600,000+ were populated automatically from other clinical systems
  - 150+ million individual data points have been collected
  - All CRFs are attached to CDEs in caDSR
- **342 trials use C3D**
  - *59 scientific publications have resulted from these trials to date*

C3D Usage



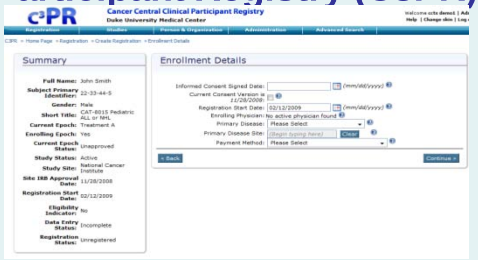
	Year 2006	Year 2007	Year 2008	Year 2009	Year 2010
No. Users June 2010	526	657	769	939	1089



	Year 2005	Year 2006	Year 2007	Year 2008	Year 2009	Year 2010
Active Studies June 2010	98	136	173	221	267	319

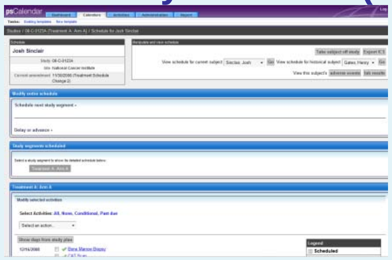
Sample Use: University of Arkansas: Managing Institutional Clinical Research

Cancer Central Clinical Participant Registry (C3PR)



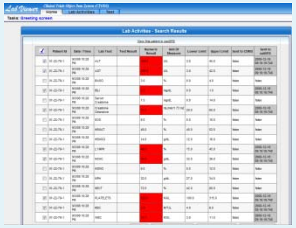
Eligibility is verified and patient is registered to a study

Patient Study Calendar (PSC)



Tracks the patient schedule throughout the study

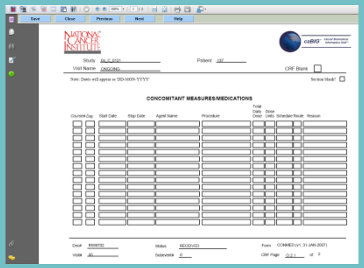
Lab Viewer



Identifies labs, loads them into the CDMS and AE system

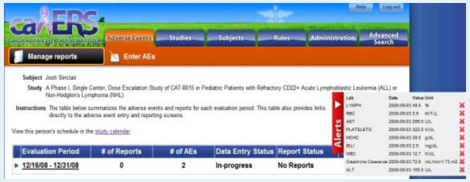
caXchange (Hub)

Vendor Clinical Data Management System



Clinical data is captured

Cancer Adverse Event Reporting System (caAERS)



Identifies and tracks adverse events and any associated schedule changes

(CTODS)

Clinical Trial Object Database System

vendor

caBIG®

Sample Use: AlphaVax Trial at Duke

- **Morse MA, Hobeika AC, Osada T, et. al. *An alphavirus vector overcomes the presence of neutralizing antibodies and elevated numbers of Tregs to induce immune responses in humans with advanced cancer.* J Clin Invest. 2010**

- ***"We report here that an alphavirus vector, packaged in virus-like replicon particles (VRP) and capable of efficiently infecting DCs, could be repeatedly administered to patients with metastatic cancer expressing the tumor antigen carcinoembryonic antigen (CEA) and that it overcame high titers of neutralizing antibodies and elevated Treg levels to induce clinically relevant CEA-specific T cell and antibody responses."***
- **caBIG® Capability:**
  - **caIntegration Hub (caXchange):** Collect and transmit clinical data between the participating trial sites
  - **C3D:** Collect participant clinical data

Integrative Cancer Research Workspace

Integrative Cancer Research Workspace (ICR)

Patterns of Usage:

- *Individual applications* to address specific analysis needs
- Access to *NCI-hosted instances* of applications online to manage data
- *“Software-as-a-service”* from 3rd party hosting partners
- *Multiple applications* independent to provide comprehensive analytical capabilities
- *Linked applications* to create custom analysis pipelines
- “Mix-and-match” analysis capabilities from the growing *NCI Enterprise Services* portfolio
- *Functional specifications* to develop or modify, or connect in-house applications

Integrative Cancer Research Workspace

Integrative Cancer Research Workspace (ICR)

caBIG® Capabilities:

- Perform genomic data with powerful workflows - GenePattern
- Perform integrated research across genomic data – geWorkbench
- Conduct high throughput genome analysis - Bioconductor
- Manage and annotate microarray gene expression data - caArray
- Develop custom, caBIG-compatible web portals to conduct integrative research - caIntegrator
- Integrated Cancer Genomics Viewer: Cancer Genome Workbench (CGWB)
- Query caGrid for analytical and data services - caBench-to-Bedside (caB2B)
- Electronically manage laboratory data - caELMIR
- Obtain biomedical annotations from curated data sources - Cancer Bioinformatics Infrastructure Objects (caBIO)
- Analyze proteomics data - LabKey/CPAS
- Query a database of genomic and proteomic annotations - Protein Information Resource (gridPIR)
- Manage proteomics laboratory information - Proteomics LIMS
- Capture proteomics experimental annotations - protExpress
- Classify complex fragment mixtures in mass spectroscopy - Q5
- Analyze microarray data in the context of pathways - Quantitative Pathway Analysis in Cancer (QPACA)
- Search a database of core pathways in human cancer - Reactome
- Analyze mass spectrometry proteomics data - RProteomics
- Make and share genomic annotations - SEED

Integrative Cancer Research Workspace

Integrative Cancer Research Workspace (ICR)

caBIG® Capabilities (continued):

- Analyze significant associations between genetic variations and disease - Cancer Genome-Wide Association Scan (caGWAS)
- Search a database of animal models for human cancer - Cancer Models Database (caMOD)
- Automatically annotate cancer related proteins - Cancer Molecular Pages
- Facilitate data sharing of nanoparticle information in cancer research - cancer Nanotechnology Laboratory (caNanoLab)
- Query across caBIG data services for clinical data - caTRIP
- Improve the quality of collected microarray data - chip artifact Correction (caCorrect)
- Perform statistical corrections to reduce systematic biases in microarray data - Distance Weighted Discrimination (DWD)
- Analyze of microarray data using gene annotation data - Function Express (caFE)
- Map gene connections between different approved genomic identifiers- GeneConnect
- Leverage the Gene Ontology (GO) to identify the biological processes - GOMiner™
- Identify differentially expressed biomarkers - omniBiomarker
- open source pathway database - Pathways Tools
- Screen and analyze RNA transcripts - TrAPSS
- Perform cluster modeling, visualization, and discovery - VIsual Statistical Data Analyzer (VISDA)

Different Capabilities Supporting Molecular Analysis

- **GenePattern**
  - Web-based genomic analysis platform provides more than 125 tools for gene expression analysis, proteomics, SNP analysis and common data processing tasks, including the creation of multi-step analysis pipelines (estimated13,000+ users at over 2,200 institutions)
- **geWorkbench**
  - Open-source software platform for genomic data integration, with more than 40 analysis and visualization tools for gene expression, sequences, protein structures, pathways, and other biomedical data. (estimated between 400-500 users)
- **Bioconductor**
  - Open-source, open-development platform for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, with more than 380 analytical modules available

Sample Use: Molecular Dissection of Colon Cancer

- **Sheffer M, Bacolod MD, et. al. *Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer* Proc Natl Acad Sci U S A. 2009;106(17):7131-6.**
- ***"To identify and characterize chromosomal abnormalities in colorectal cancer, we performed a statistical analysis of 299 expression and 130 SNP arrays profiled at different stages of the disease, including normal tissue, adenoma, stages 1-4 adenocarcinoma, and metastasis."***
- **caBIG® Capability:**
  - **caArray:** Share the large dataset with the community, associated with detailed sample annotations
  - **caIntegrator:** Provide community the capability to further analyze this unique data set
  - **caDSR**: Provide registration of clinical data elements

Tissue Banks & Pathology Tools Workspace

Tissue
Banks &
Pathology
Tools
Workspace
(TBPT)

Patterns of Usage:

- *One central biorepository* to manage multiple biospecimen collections
- *Separate local instances of biorepository* for each collection
- *Biorepository to securely share* samples among multiple collaborating organizations
- caBIG® biorepository as a back end *to existing commercial or legacy* biospecimen management systems in-house

caBIG® Capabilities:

- Collect, store, annotate, aliquot, search, and track distribution of many types of biospecimens -- **caTissue Suite**

Sample Use: Washington University at St. Louis

- **Washington University at St Louis manages more than 558,000 biospecimens in multiple core facilities across the university. Facilities managers see benefits in locating specimens, collecting detailed annotations, and managing multiple levels of consent tied to specific specimens.**
    - ***"caTissue allows us to collect much more detailed information about each sample, making it easier and more accurate to find exactly the samples requested by a researcher."*** Lisa Taylor-Reinwald, Neuropathology Core Lab Manager, Washington University in St Louis
    - ***"caTissue is much faster and more simpler to use than our home-grown system. We can locate and provide the samples to researchers with much more confidence that we had in the past."*** Gary Skolnik, Study Coordinator for Nutritional and Obesity Research, Washington University in St Louis

In Vivo Imaging Workspace

In Vivo
Imaging
Workspace
(IMAG)

Patterns of Usage:

- *Local* management of digital images
- NCI-hosted instance storing for private or public use
- *Novel imaging applications* using XIP framework
- Structured *annotation of images* with AIM
- *Virtual PACS* to connect commercial PACS systems with other image databases to share collections of images

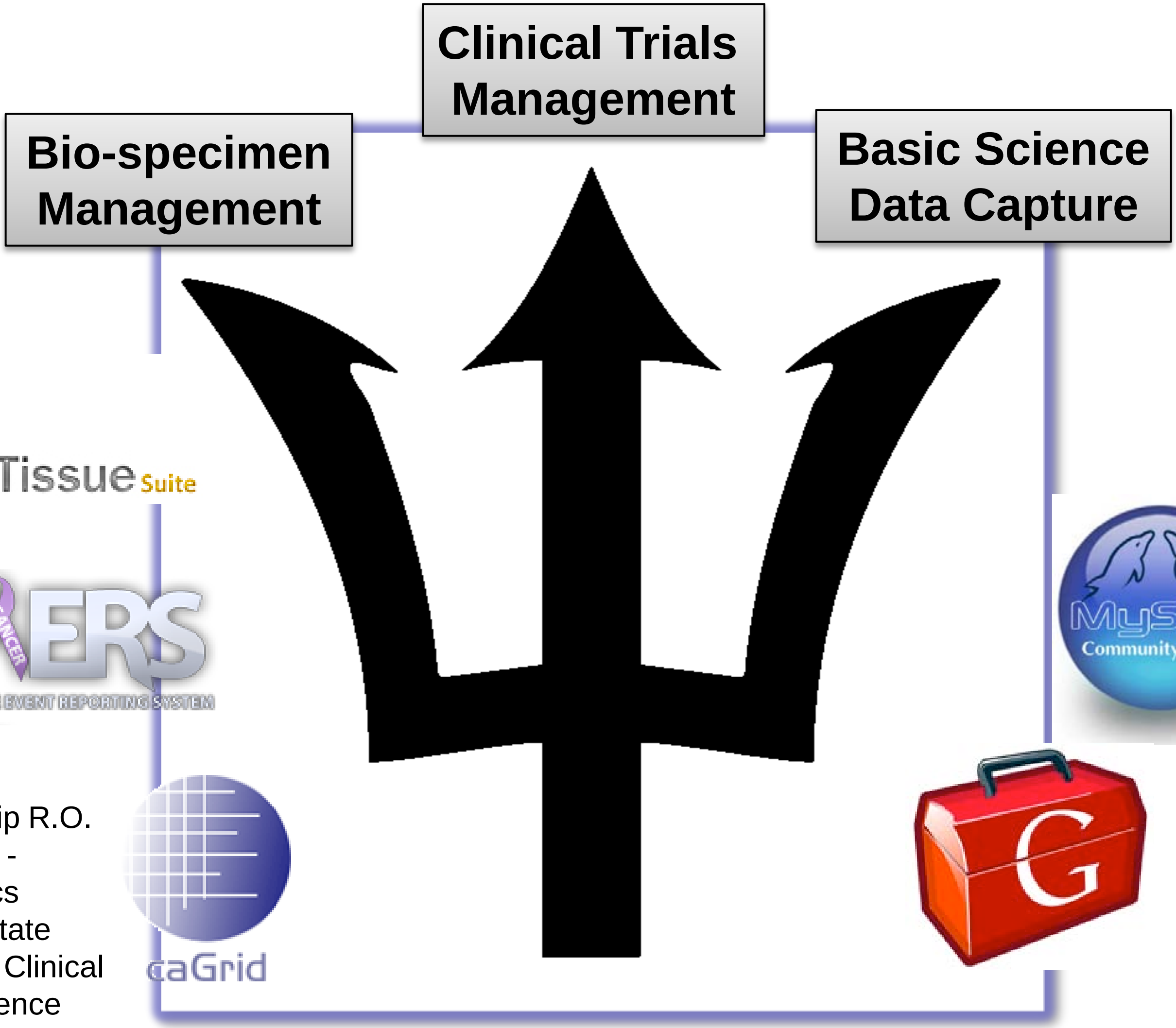
caBIG® Capabilities:

- Store, annotate and share DICOM format medical images -- **National Biomedical Image Archive (NBIA**)
- Capture radiologist's notes and share with colleagues using standards-based annotations -- **Annotation and Image Markup (AIM)**
- Capture digital images of pathology slides -- **caMicroscope**
- Develop and rapidly test novel image analysis algorithms -- **eXtensable Imaging Platform – XIP**
- Connect commercial PACS systems with other 3rd party image databases -- **Imaging Middleware/Virtual PACS**

Sample Use: Ohio State University

- **Connecting diverse sets of clinical and research data across the university and between OSU and external collaborators**
- **caBIG® Capability:**
  - **caGrid** – Grid technology is being applied to solve multiple data access problems including:
    - Creating the Translational Informatics and Data Management Grid (TRIAD) to allow researchers inside OSU to access basic and clinical research data across the university
    - Creating a federated National Clinical Trials metadata repository for all Clinical and Translational Science Aware (CTSA)-funded centers
    - Creating a federated data warehouse to connect OSU Hospital with Nationwide Children’s Hospital in Columbus Ohio as part of the Ohio Perinatal Research Network

TRITON



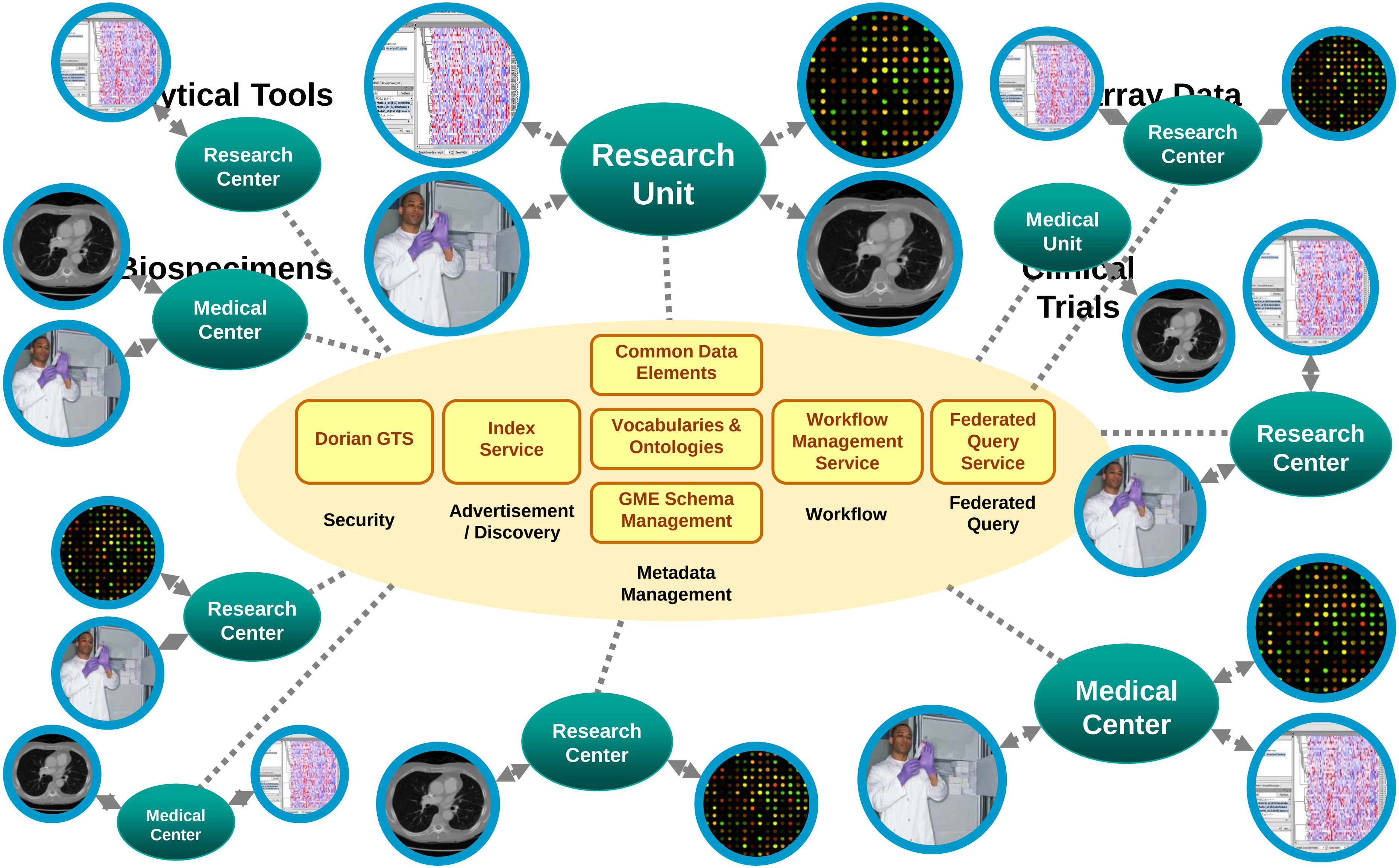
Slide courtesy of Philip R.O. Payne, PhD, Director - Biomedical Informatics Program, The Ohio State University, Center for Clinical and Translational Science

Sample Use: University of Alabama at Birmingham

- **Data Access and Sharing Initiative (DASI) project provides unified access to diverse sets of clinical and research data, along with other IT systems including radiology and billing across the Cancer Center, hospital, and entire University**
- **caBIG® Capabilities:**
  - **caGrid –** Serves as the underlying IT connection for virtually all research and clinical IT systems, linking the Cancer Center with the University basic research labs and the hospital. Initial project enables access to clinical outcomes data from the Hospital to basic researchers at the University labs
  - **caTissue –** Currently sharing more than 2200 samples on caGrid
  - **LabKey/CPAS –** Managing proteomics data for basic researchers
  - ***"It's fair to say that [caBIG®] is changing stuff all the way down from the wiring of the IT systems to the way researchers can find data, all the way up to the way senior management is thinking about not only doing individual research projects but also positioning the university going forward."*** Joan Hicks, Chief Information Officer, UAB Health System

Sample Use: Washington University at St Louis

- **Providing access to current and legacy in- and out-patient clinical data from more than 5 million patients seen at 13 hospitals and clinics across BJC Healthcare and linking multiple diverse in-house collections of biospecimens**
- **caBIG® Capabilities:**
    - **LexEVS** –80,000 terms harmonized to make clinical data interoperable in Clinical Investigation Data Exploration Repository (CIDER)
    - **caTissue** – 558,000 biospecimens across multiple core facilities managed by a central installation of caTissue
    - **caGrid** –CIDER database connected to caTissue and caArray
    - **caArray** –Data from almost 200 samples made available on caGrid
    - **caB2B** – Web-based queries to locate data in any connected repository



ytical Tools

Research Center

Biospecimens

Medical Center

Research Unit

rray Data

Research Center

Medical Unit

Clinical Trials

Common Data Elements

Dorian GTS

Security

Index Service

Advertisement / Discovery

Vocabularies & Ontologies

GME Schema Management

Metadata Management

Workflow Management Service

Workflow

Federated Query Service

Federated Query

Research Center

Research Center

Medical Center

Research Center

Medical Center

Cancer Knowledge Cloud

Grid Services Infrastructure

Cancer Knowledge Cloud
Grid Services Infrastructure

The Glioma Molecular Diagnostic Initiative

- **A national study through 2 NCI-funded brain tumor consortia**
- **More than 1000 patients with gliomas to be accrued**
- **Extensive prospective clinical data to be correlated with molecular data**

Objectives:

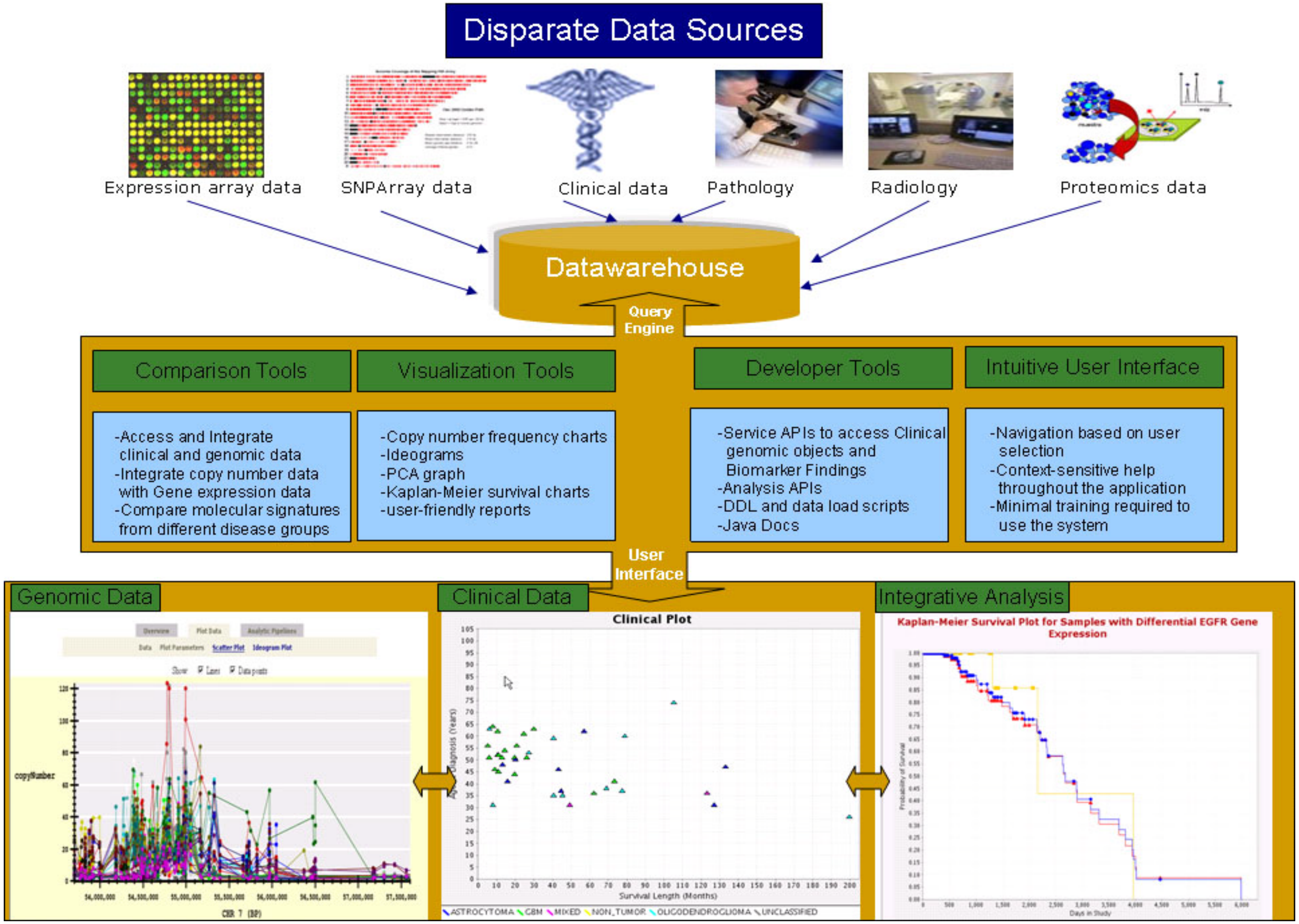
- Produce a biologically significant pathological classification of gliomas, with strong correlation to outcome of disease, that would allow for better prognostication and more informed therapeutic decision making.

- Through the informatic mining of extensive molecular/genetic data to be able to find new molecular targets for therapy that would be relevant both to the disease type and patient status.

- Produce a **publicly accessible** database containing all the aforementioned data, that will contain the analysis tools necessary for all ends of the research spectrum (clinicians, chemists, geneticist, bioinformaticians, etc) to profit from the wealth of information stored, produce new models of disease, and envision new targeted therapies.

Slide courtesy of Dr. Howard Fine, NeuroOncology Branch, CCR

Data Integration via Rembrandt



Disparate Data Sources

Comparison Tools

- -Access and Integrate clinical and genomic data
- -Integrate copy number data with Gene expression data
- -Compare molecular signatures from different disease groups

Visualization Tools

- -Copy number frequency charts
- -Ideograms
- -PCA graph
- -Kaplan-Meier survival charts
- -user-friendly reports

Developer Tools

- -Service APIs to access Clinical genomic objects and Biomarker Findings
- -Analysis APIs
- -DDL and data load scripts
- -Java Docs

Intuitive User Interface

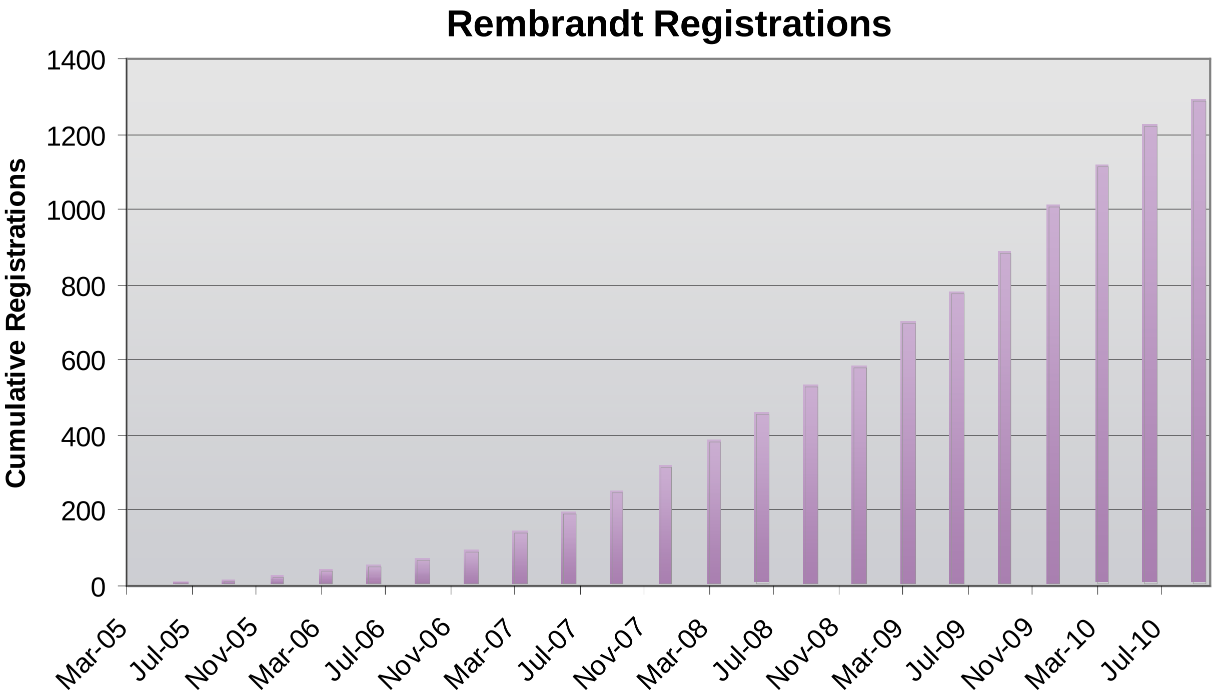
- -Navigation based on user selection
- -Context-sensitive help throughout the application
- -Minimal training required to use the system

6/21/2006

Slide courtesy of Dr. Howard Fine, NeuroOncology Branch, CCR

Sample use: Access to REMBRANDT Data

- **caBIG® provides a bioinformatics framework and web portal for researchers to integrate clinical and functional genomics data from the REMBRANDT project (http://caintegrator-info.nci.nih.gov/rembrandt).**



Sample Use: : PANVAC Trial at Duke

- **PANVAC - VF Vaccine for the Treatment of Metastatic Pancreatic Cancer**
  - By injecting immunity-producing drugs directly into the tumor, the body develops an increased ability to recognize the tumor, and subsequent metastases, as foreign and attack them directly.
  - This study is designed to study the safety and efficacy of PANVAC-VF in combination with Granulocyte-macrophage colony-stimulating factor (GM-CSF) versus best supportive care or palliative chemotherapy
- **caBIG® Capability:**
  - **caIntegration Hub (caXchange):** enables the collection and transmission of clinical data between the participating trial sites

Duke Multi-Center Vaccine Study

- PANVAC Trial
  - Grid-enabled
  - Secure
  - Accessible

Slide courtesy of Robert P. Annechiarico, Director of Cancer Center Information Systems, Duke Comprehensive Cancer Center

Duke Comprehensive Cancer Center
THE PROMISE OF RESEARCH-THE POWER OF HEALING
NCI CCC

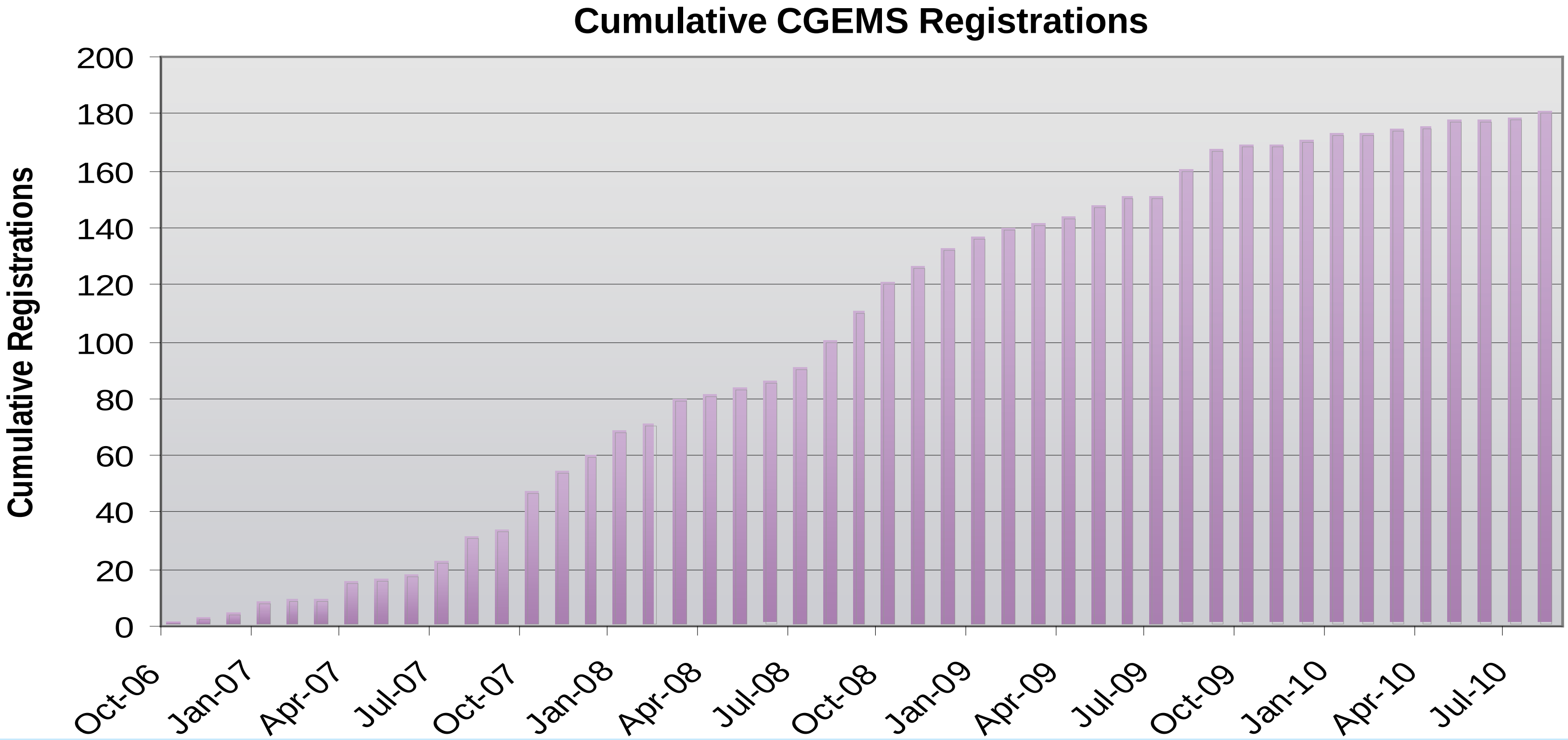
Sample Use: Director’s Challenge Lung Study

- **Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma. Shedden K, Taylor JM, Enkemann SA, et al. *Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study*. Nat Med. 2008 Aug;14(8):822-7. Epub 2008 Jul 20.**
- ***“Here we report a large, training-testing, multi-site, blinded validation study to characterize the performance of several prognostic models based on gene expression for 442 lung adenocarcinomas. The hypotheses proposed examined whether microarray measurements of gene expression either alone or combined with basic clinical covariates (stage, age, sex) could be used to predict overall survival in lung cancer subjects…This study also provides the largest available set of microarray data with extensive pathological and clinical annotation for lung adenocarcinomas. ”***
- **caBIG® Capability:**
  - **caArray:** Share the large dataset with the community, associated with detailed sample annotations
  - **caIntegrator:** Provide the community the capability to further analyze this unique data set
  - **caDSR**: Provide registration of clinical data elements

Sample Use: Reuse of Cancer Genetic Markers of Susceptibility (CGEMS) Data

Through caIntegrator, caBIG® enables researchers to access data from the CGEMS project. More than 25 publications have arisen from mining CGEMS data.

- **Saetrom P, Biesinger J, Li SM, et. al. *A risk variant in an miR-125b binding site in BMPR1B is associated with breast cancer pathogenesis*. Cancer Res. 2009 Sep 15;69(18):7459-65.**



Sample Use: Managing Medical Images

- **caBIG® Usage: National Biomedical Image Archive (NBIA)**
  - 4.76+ million images publicly available through NCI instance of NBIA
    - RIDER/LIDC/IDRI – 1,766,820
    - CT Colonography – 941,774
    - Virtual Colonoscopy – 686,257
    - Rembrandt VASARI – 593,260
    - FDA Phantoms – 420,938
    - TCGA – 113,953
    - RTOG 0522 Trial – 206,665
    - PET Lymphoma Study – 28,462
  - Head and Neck radiation collaboration with NCI Radiation Research and American College of Radiology
  - Organizations may choose to install their own instance of NBIA and store their research images locally
    - Washington University at St Louis – 12 million images
    - Fox Chase Cancer Center – 10,000 images
    - Pediatric Brain Tumor Consortium – 5,500
    - NCRI – 985
    - University of Colorado
    - NIAMS OsteoArthritis Initiative – 14.4+ million images

Sample Use: RIDER and LIDC

- **van Ginneken B, Armato SG 3rd, de Hoop B, et. al. *Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study.* Med Image Anal. 2010 Dec;14(6):707-22. Epub 2010 Jun 4.**
  - *“Results show a substantial performance difference between algorithms, and demonstrate that combining the output of algorithms leads to marked performance improvements.”*
- **22 additional secondary publications based on Reference Image Database to Evaluate Therapy Response (RIDER) and Lung Image Database Consortium (LIDC) data**
- National Biomedical Image Archive (NBIA) serves as database to make available a high-quality annotated set of chest images to researchers, for algorithm validation and training
  - *“NBIA was absolutely essential to provide access to large set of highly curated images to researchers in diverse locations. These reference images leads to creation of more accurate tools, which can measure changes in tumors more accurately, and at an earlier stage of treatment, allowing us to get patients off treatment faster or change the treatment sooner if it’s not effective.”* **Charles Meyer, Ph.D., Lab Director, Digital Image Processing Laboratory, Department of Radiology, University of Michigan**

Sample Use: Biospecimens

- **Usage: 2.17+ million biospecimens available through caTissue, at least 1.46 million biospecimens are available for sharing via caGrid**
- **34 organizations in production usage, including:**

🟩 Washington University St. Louis	558494	Loma Linda University	8204
🟩 Fox Chase Cancer Center	288803	Ohio State University Medical Center	7273
University of California San Francisco	281338	Christiana Care	5013
🟩 Thomas Jefferson University - Kimmel Cancer Center	236458	University of Arkansas	4472
🟩 University of Michigan	215653	🟩 University of Alabama Birmingham	2200
🟩 Indiana University - Regenstrief	189105	Vanderbilt-Ingram Cancer Center	1813
Dana Farber Cancer Center	128798	Georgetown University - Lombardi Cancer Center	1405
Johns Hopkins University	67644	University of Arizona	1000
🟩 University of Pittsburgh	50444	Baylor - Duncan Cancer Center	469
Emory University	39973	🟩 M.D. Anderson Cancer Center	164
Louisiana Cancer Consortium	29276	🟩 University of Iowa - Holden Cancer Center	150
🟩 University of Pennsylvania	14616		
Edinburgh Experimental Cancer Medicine Center	14000		
University of Nebraska - Eppley Cancer Center	10198		
🟩 Yale	8570		

🟩 Green = samples on the grid

Cooperative Banking Committee for the Melanoma Grid

Ding L, Ellis MJ, Li S, Larson DE, (Watson M) et. al. *Genome remodelling in a basal-like breast cancer metastasis and xenograft*. Nature. 2010 Apr 15;464(7291):999-1005.

Sample use: Skin SPORE

- **Halaban R, Zhang W, Bacchiocchi A, et. al. *PLX4032, a selective BRAF(V600E) kinase inhibitor, activates the ERK pathway and enhances cell migration and proliferation of BRAF melanoma cells*. Pigment Cell Melanoma Res. 2010 Apr;23(2):190-200. Epub 2010 Feb 10.**
  - *"The persistently active ERK1/2 triggered downstream effectors in BRAF(WT) melanoma cells and induced changes in the expression of a wide-spectrum of genes associated with cell cycle control."*
- **Rubinstein JC, Tran N, Ma S, Halaban R, Krauthammer M*. Genome-wide methylation and expression profiling identifies promoter characteristics affecting demethylation-induced gene up-regulation in melanoma.* BMC Med Genomics. 2010 Feb 9;3:4.**
  - *"We show that the combination of promoter CpG content and methylation level informs the ability of decitabine treatment to up-regulate gene expression."*
- **caBIG® Capabilities:**
  - **caTissue:** Manage biospecimens and associated annotations
  - **caArray:** Collect, manage and annotate gene expression data from microarrays
  - **caIntegrator:** Makes integrated data available to researchers across the SPORE

Sample Use: Cooperative Groups Breast Cancer Data Mart

- **Yuan, X and Anderson, S.** ***Meta-analysis methodology for combiningtreatment effects from Cox proportional hazard models with different covariate adjustments.*** **Biometrical Journal 52 (2010) 4, 1–19**

  (2 more in preparation)

  - Data from 54 studies about women with breast cancer, from 6 co-operative groups.
  - Collection of 65,191 patient records in the database provides unprecedented collection of structured data for collaborative research

Sample Use: The Cancer Molecular Analysis Portal

http://cma.nci.nih.gov

- Enables users to access, search, visualize, and integrate genomic data with corresponding clinical information
- Helps find novel correlations between data and observations that would be difficult or impossible to find using conventional analytical tools and methods
- Provides access to, and facilitate analysis of, data from other research studies such as REMBRANDT*, TARGET* GSK Expression, COSMIC mutations, and JHU mutations

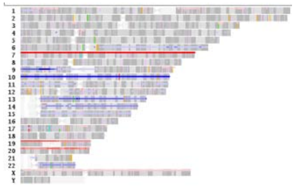
* The Cancer Genome Atlas (TCGA)
* Repository of Molecular Brain Neoplasia Data (REMBRANDT)
* Therapeutically Applicable Research to Generate Effective Treatments (TARGET)

Cancer Genome Workbench for Integrated Cancer Data @ http://cgwb.nci.nih.gov

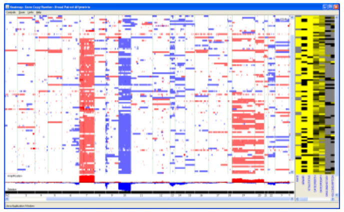
Project	Disease	Genomic Data				
		Copy Number	Gene Expression	Methylation	Clinical	Mutation
TCGA	**GBM, Ovarian**	+	+	+	+	+
TSP	**lung**	+			+	+
TARGET	**ALL, NB**	+	+		+	+
JHU	**GBM, Pan**					+
Rembrandt	**GBM**	+				
GSK cell line	**>30 tissues**	+	+			
COSMIC	**>30 tissues**					+

Viewers for individual & integrated genomic Data

Landscape

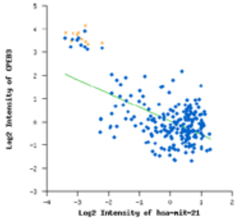


Heatmap

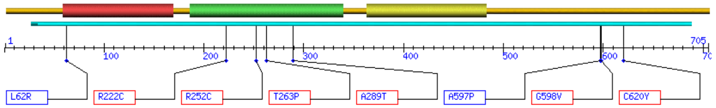


Genome

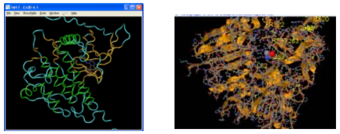
Correlation Plot



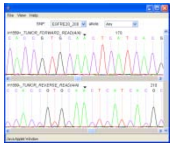
Protein



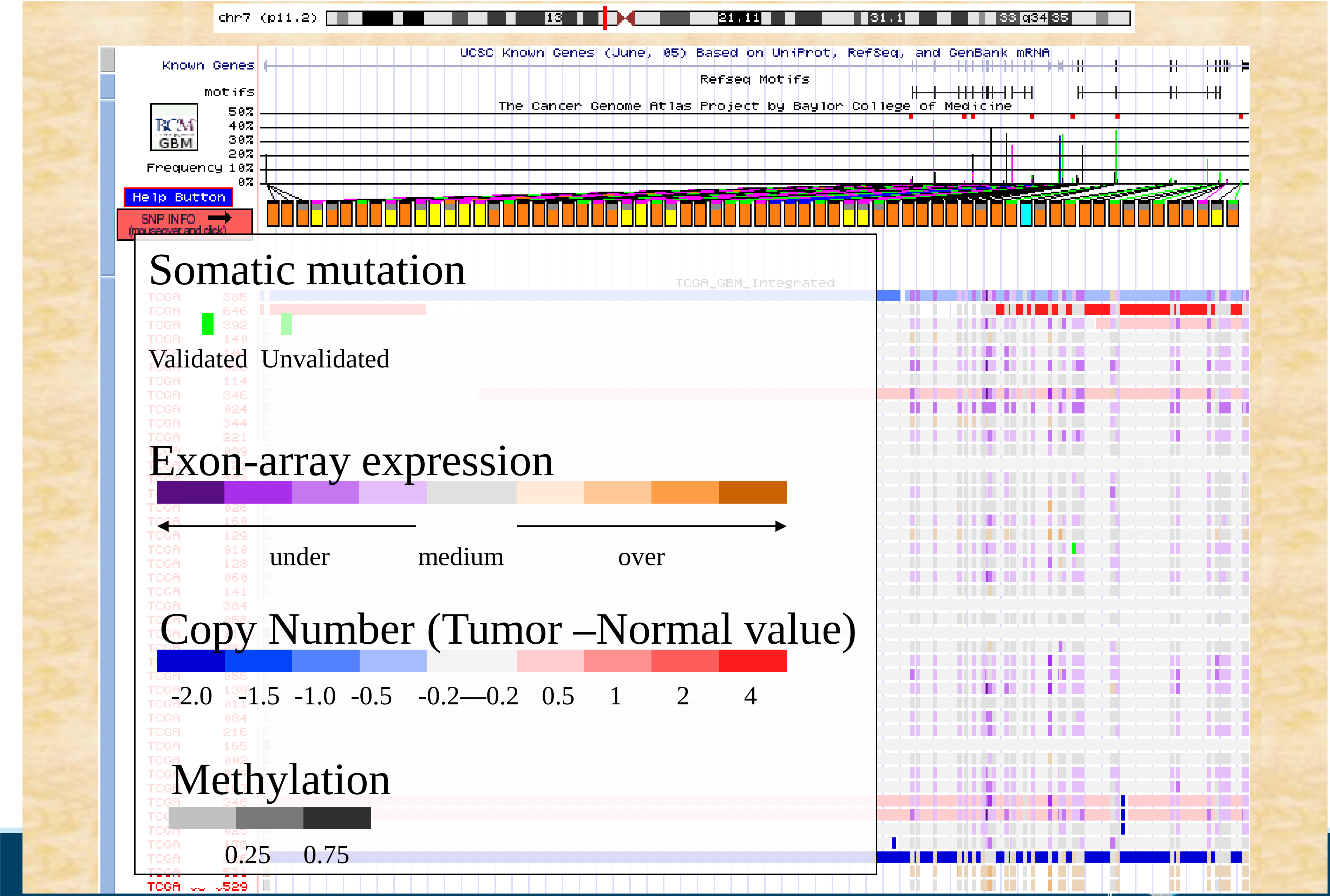
3D Structure Viewer



Trace



Comprehensive Genomic Summary



Somatic mutation

Validated Unvalidated

Exon-array expression

under medium over

Copy Number (Tumor –Normal value)

-2.0 -1.5 -1.0 -0.5 -0.2—0.2 0.5 1 2 4

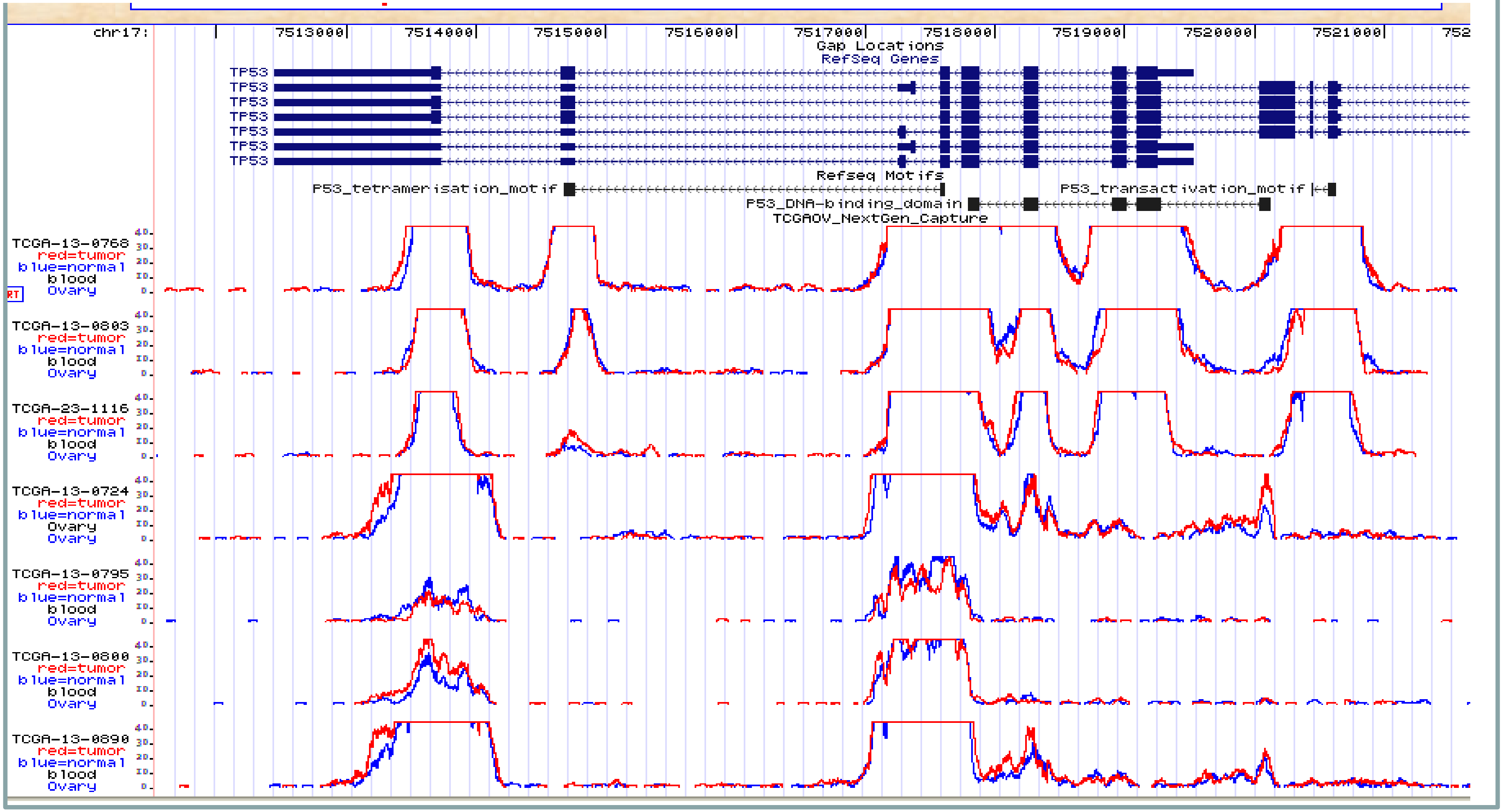
Methylation

0.25 0.75

Putative Somatic Mutations can be Manually Reviewed Eg: Frameshift Mutation in EGFR in Paired Tumor/Normal

SNP: EGFRP5_154i
allele: Any

Next-gen sequencing p53 analysis

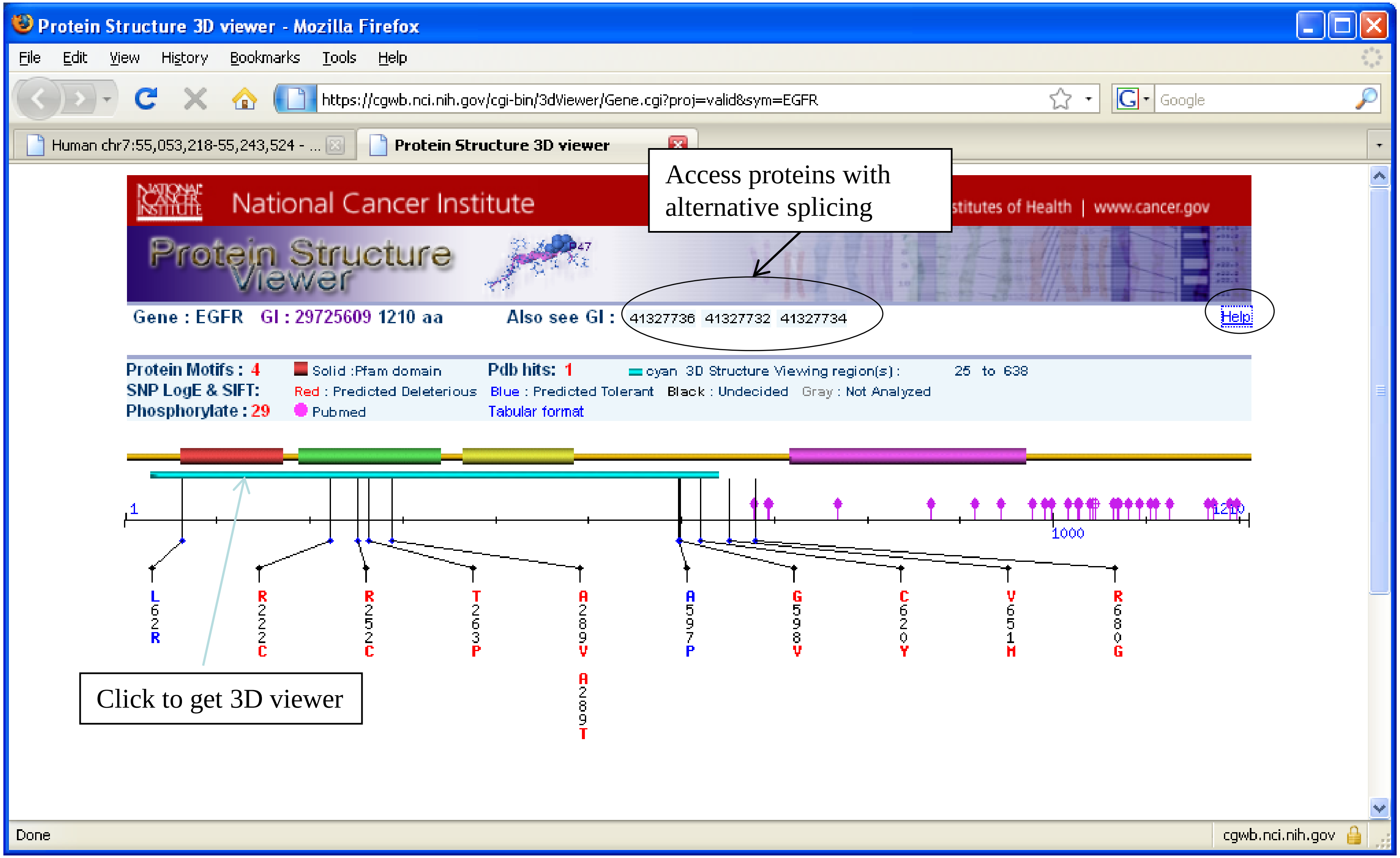


Sample coverage maximized to 40x, height represents coverage

Red line=coverage for a tumor sample

Blue line=coverage for its matching normal

Protein structure view of EGFR mutations



https://cgwb.nci.nih.gov/cgi-bin/3dViewer/Gene.cgi?proj=valid&sym=EGFR

Gene : EGFR GI : 29725609 1210 aa Also see GI : 41327736 41327732 41327734

Protein Motifs : 4 ▬ Solid :Pfam domain Pdb hits: 1 ▬ cyan 3D Structure Viewing region(s) : 25 to 638

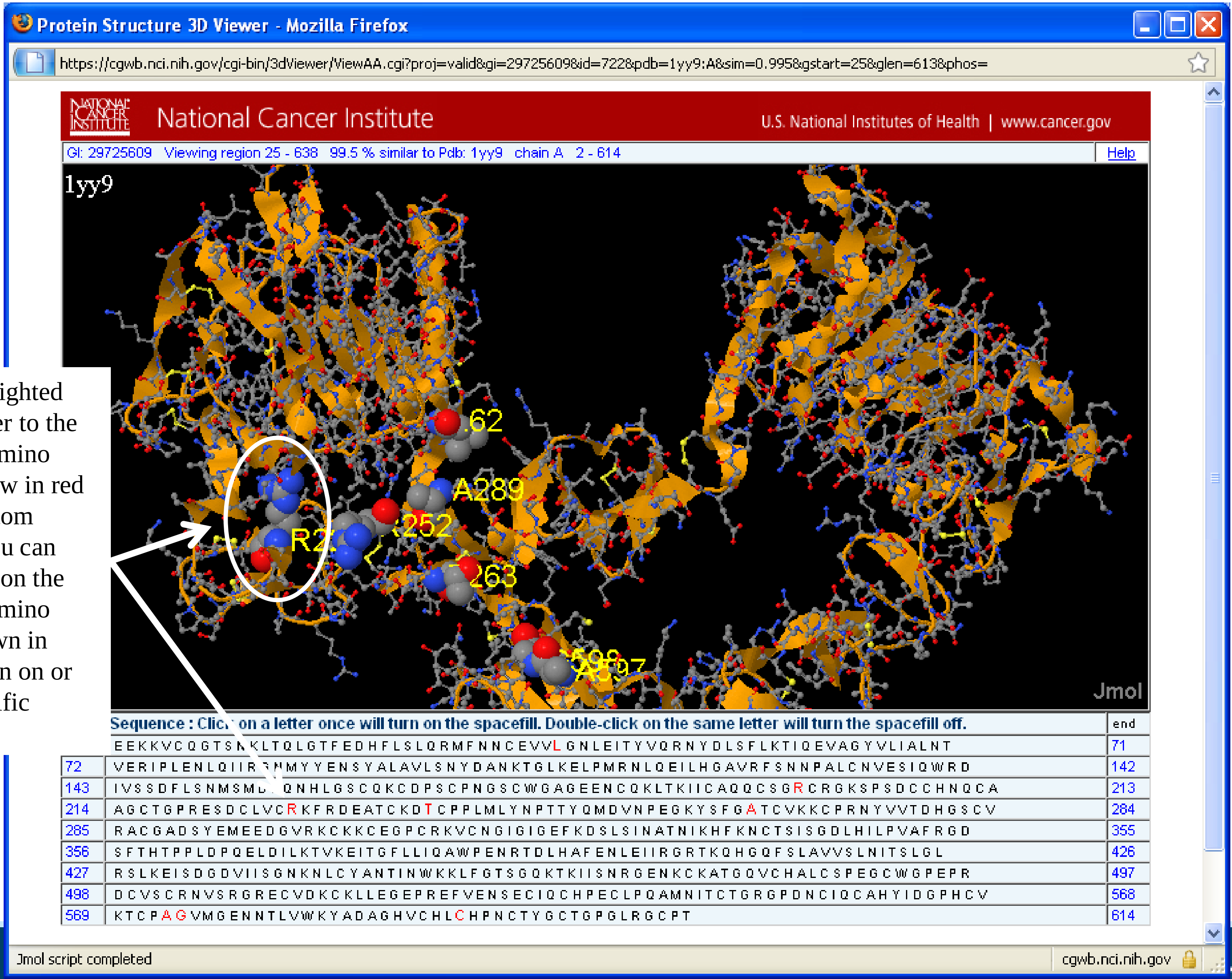
SNP LogE & SIFT: Red : Predicted Deleterious Blue : Predicted Tolerant Black : Undecided Gray : Not Analyzed

Phosphorylate : 29 ● Pubmed Tabular format

Access proteins with alternative splicing

Click to get 3D viewer

3D Structure Viewer

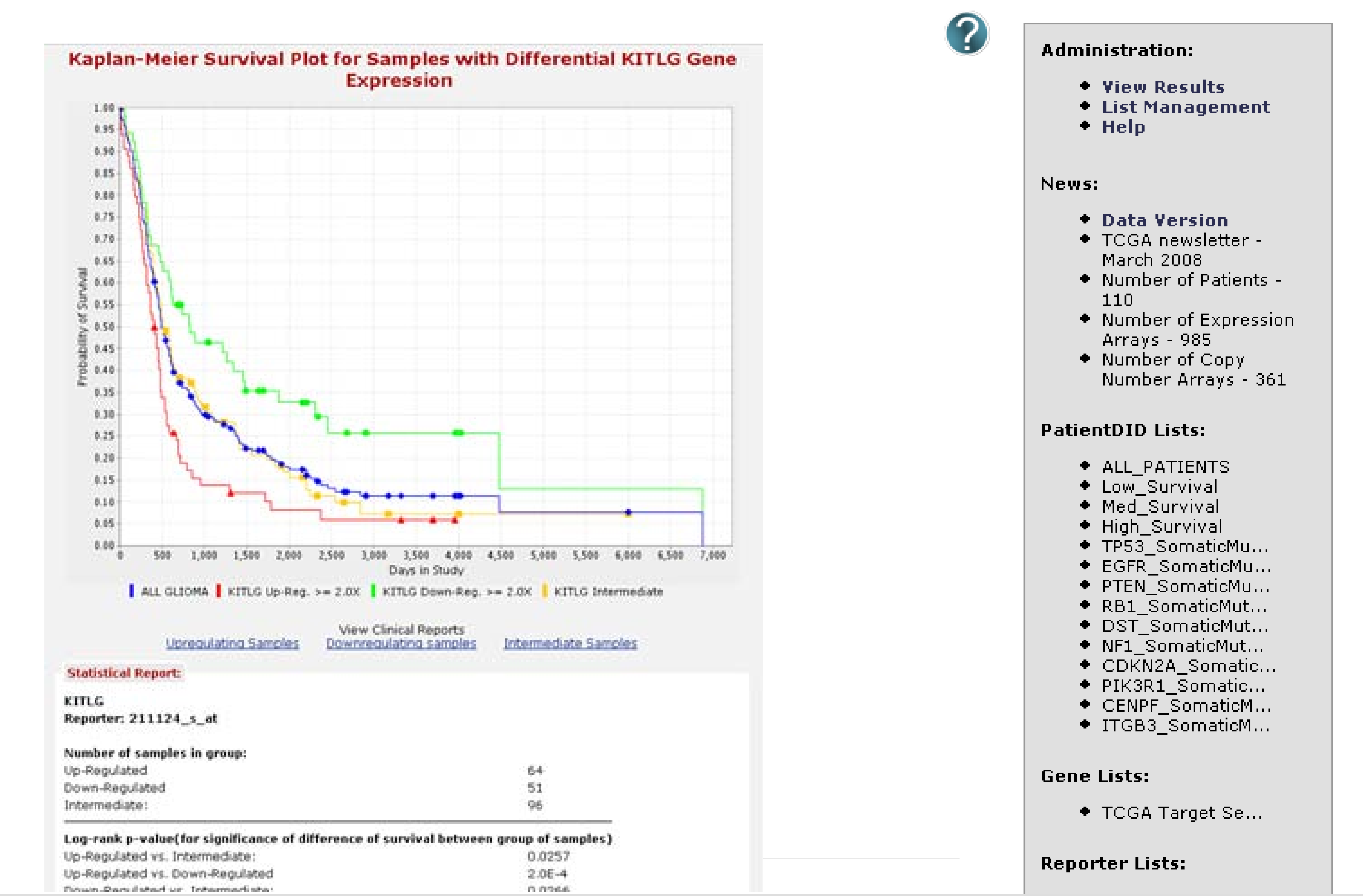


GI: 29725609 Viewing region 25 - 638 99.5 % similar to Pdb: 1yy9 chain A 2 - 614 | Help

	Sequence : Click on a letter once will turn on the spacefill. Double-click on the same letter will turn the spacefill off.	end
	E E K K V C Q G T S N K L T Q L G T F E D H F L S L Q R M F N N C E V V L G N L E I T Y V Q R N Y D L S F L K T I Q E V A G Y V L I A L N T	71
72	V E R I P L E N L Q I I R G N M Y Y E N S Y A L A V L S N Y D A N K T G L K E L P M R N L Q E I L H G A V R F S N N P A L C N V E S I Q W R D	142
143	I V S S D F L S N M S M D F Q N H L G S C Q K C D P S C P N G S C W G A G E E N C Q K L T K I I C A Q Q C S G R C R G K S P S D C C H N Q C A	213
214	A G C T G P R E S D C L V C R K F R D E A T C K D T C P P L M L Y N P T T Y Q M D V N P E G K Y S F G A T C V K K C P R N Y V V T D H G S C V	284
285	R A C G A D S Y E M E E D G V R K C K K C E G P C R K V C N G I G I G E F K D S L S I N A T N I K H F K N C T S I S G D L H I L P V A F R G D	355
356	S F T H T P P L D P Q E L D I L K T V K E I T G F L L I Q A W P E N R T D L H A F E N L E I I R G R T K Q H G Q F S L A V V S L N I T S L G L	426
427	R S L K E I S D G D V I I S G N K N L C Y A N T I N W K K L F G T S G Q K T K I I S N R G E N K C K A T G Q V C H A L C S P E G C W G P E P R	497
498	D C V S C R N V S R G R E C V D K C K L L E G E P R E F V E N S E C I Q C H P E C L P Q A M N I T C T G R G P D N C I Q C A H Y I D G P H C V	568
569	K T C P A G V M G E N N T L V W K Y A D A G H V C H L C H P N C T Y G C T G P G L R G C P T	614

Big, highlighted atoms refer to the mutated amino acids (show in red in the bottom panel). You can also click on the mutated amino acid (shown in red) to turn on or off a specific mutation

Gene expression analyses related to clinical outcome



Kaplan-Meier Survival Plot for Samples with Differential KITLG Gene Expression

View Clinical Reports

Upregulating Samples Downregulating samples Intermediate Samples

Statistical Report:

KITLG
Reporter: 211124_s_at

Number of samples in group:

Up-Regulated	64
Down-Regulated	51
Intermediate:	96

Log-rank p-value(for significance of difference of survival between group of samples)

Up-Regulated vs. Intermediate:	0.0257
Up-Regulated vs. Down-Regulated	2.0E-4
Down-Regulated vs. Intermediate:	0.0266

Administration:

- View Results
- List Management
- Help

News:

- Data Version
- TCGA newsletter - March 2008
- Number of Patients - 110
- Number of Expression Arrays - 985
- Number of Copy Number Arrays - 361

PatientDID Lists:

- ALL_PATIENTS
- Low_Survival
- Med_Survival
- High_Survival
- TP53_SomaticMu...
- EGFR_SomaticMu...
- PTEN_SomaticMu...
- RB1_SomaticMut...
- DST_SomaticMut...
- NF1_SomaticMut...
- CDKN2A_Somatic...
- PIK3R1_Somatic...
- CENPF_SomaticM...
- ITGB3_SomaticM...

Gene Lists:

- TCGA Target Se...

Reporter Lists:

caBIG® In the Future

Expanding Across Sectors and Internationally

15 countries engaged with and/or using caBIG® tools and technologies, including:

- United Kingdom
- Latin America
- India
- China
- Mexico, Brazil, Uruguay, Argentina, Chile
- Czech Republic
- The Netherlands
- Germany
- Finland
- Jordan
- Pakistan
- Australia
- New Zealand

- **United Kingdom:** NCI collaboration with National Cancer Research Initiative (NCRI) focused on the use of caGrid technology to connect researchers and enable exchange of research data.
- **Latin America:** Latin American Cancer Pilot Program began patient enrollment for two clinical trials on molecularly characterized stage II and III breast cancer patients, using a broad suite of caBIG® tools to ensure that researchers can compare data across partner sites.
- **China:** Duke University Comprehensive Cancer Center and Beijing University Cancer Hospital launched a collaboration using caBIG® capabilities to conduct first clinical trials in China where all patients are registered electronically.
- **India:** NCI engaged with the All India Institute of Medical Sciences, the Center for Development of Advanced Computing, and the Tata Memorial Hospital of Mumbai over the use of grid computing for managing clinical trials data.

Next Generation: caBIG® 2.0

caBIG® capabilities are evolving to support the rapidly changing needs of Cancer Centers as they move to EHRs and genomically-guided medicine:

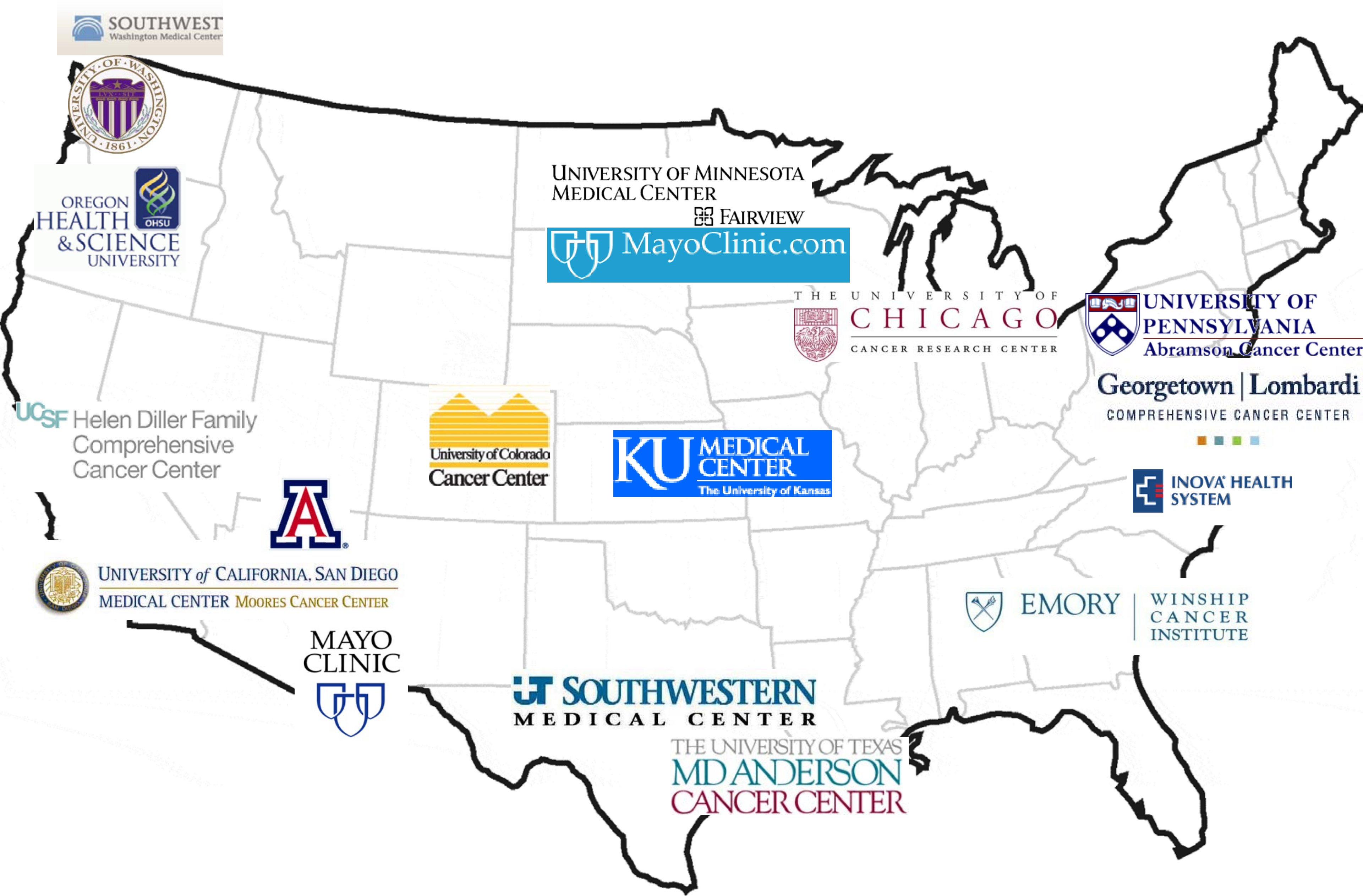
- Collaborate with ASCO on caEHRs
- Capitalize on tech advances: cloud computing, Service-Oriented Architecture, mobile devices
- Lower barriers: make it easier and easier for all stakeholders to use caBIG® research capabilities
- Leverage increasingly mature collection of publicly-available open source infrastructure

The I-SPY TRIAL (Investigation of Serial studies to Predict Your Therapeutic Response with Imaging And moLecular analysis):

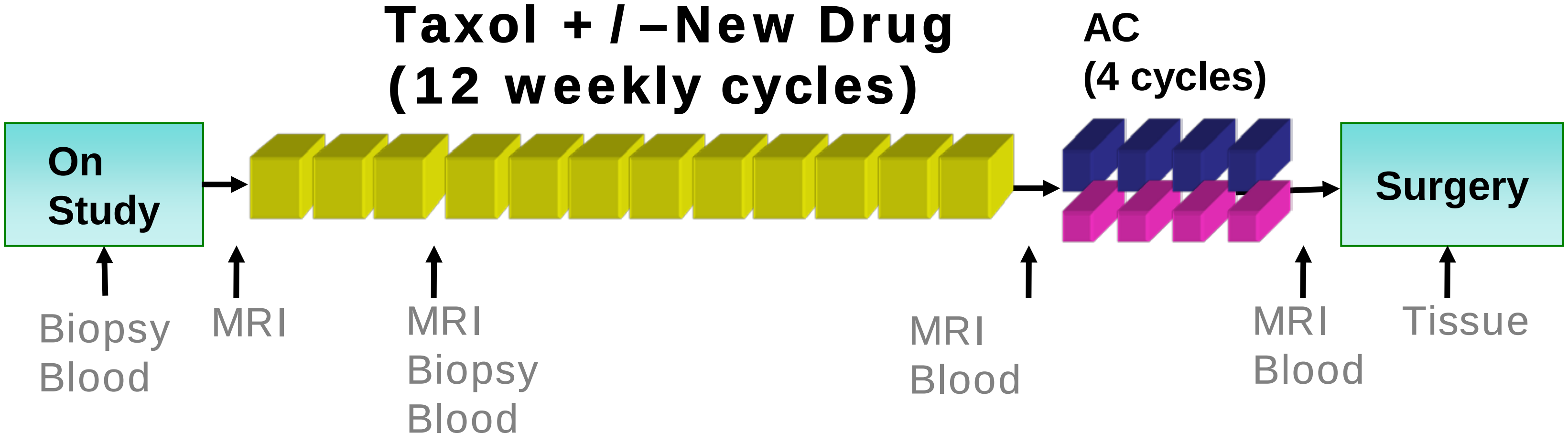
A national study to leverage biomarkers
in predicting response to combinatorial therapy for
women with Stage 3 breast cancer.

(PI Laura Esserman, UCSF)

Projected I-SPY 2 study sites



I-SPY Adaptive Trial Outline



Taxol + / – New Drug
(12 weekly cycles)

AC
(4 cycles)

On Study

Surgery

Biopsy
Blood

MRI

MRI
Biopsy
Blood

MRI
Blood

MRI
Blood

Tissue

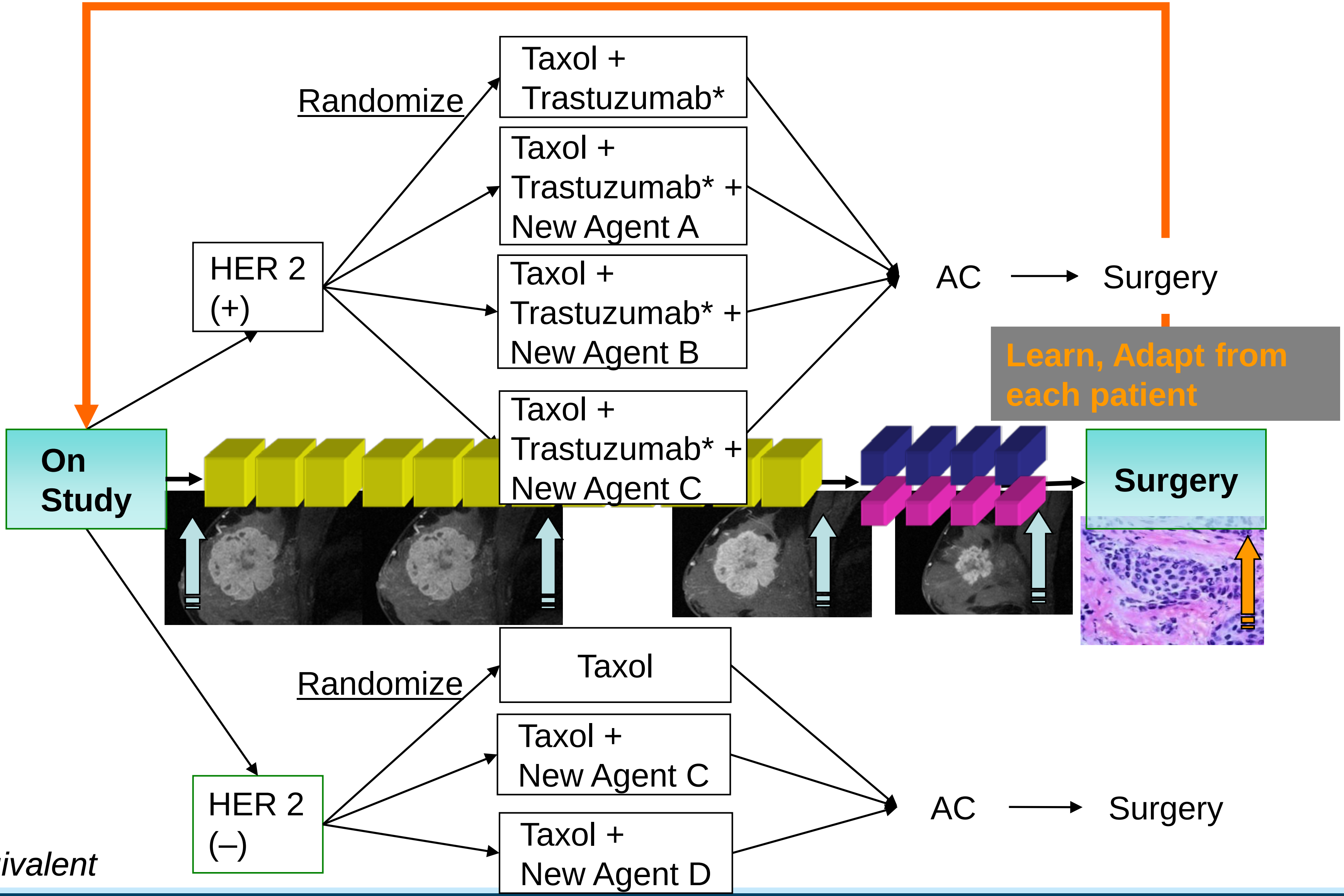
Accrual: Anticipate 800 patients over 3–4 years

Enroll ~20 patients per month

Participating Sites: 15–20 across US and Canada

I-SPY Adaptive Trial:

Introduce several new agents for a given profile



On Study

HER 2 (+)

Randomize

Taxol + Trastuzumab*

Taxol + Trastuzumab* + New Agent A

Taxol + Trastuzumab* + New Agent B

Taxol + Trastuzumab* + New Agent C

AC → Surgery

Learn, Adapt from each patient

Surgery

HER 2 (–)

Randomize

Taxol

Taxol + New Agent C

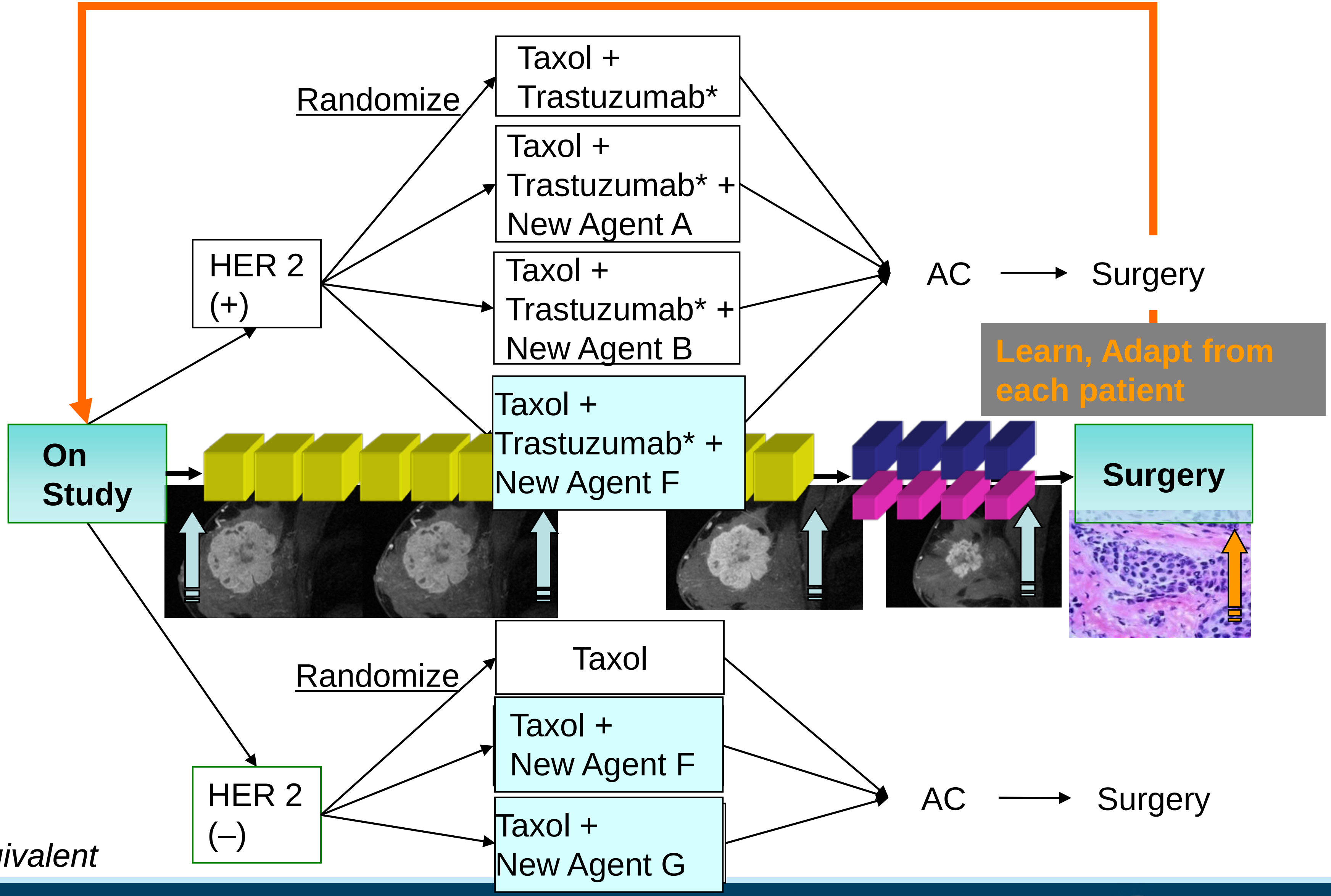
Taxol + New Agent D

AC → Surgery

*Or Equivalent

I-SPY Adaptive Trial:

Introduce several new agents for a given profile



On Study

HER 2 (+)

Randomize

Taxol + Trastuzumab*

Taxol + Trastuzumab* + New Agent A

Taxol + Trastuzumab* + New Agent B

Taxol + Trastuzumab* + New Agent F

AC → Surgery

Learn, Adapt from each patient

Surgery

HER 2 (–)

Randomize

Taxol

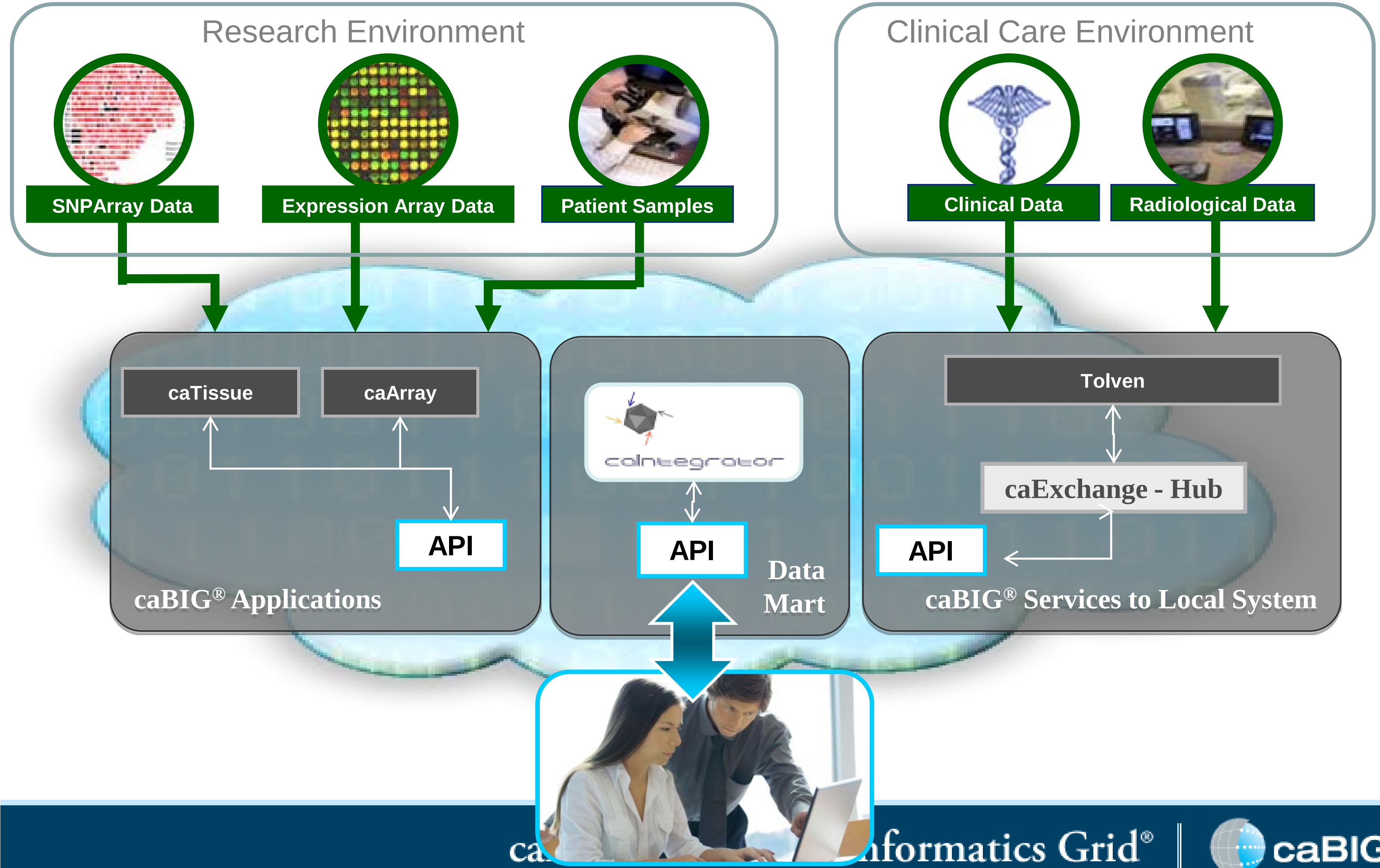
Taxol + New Agent F

Taxol + New Agent G

AC → Surgery

**Or Equivalent*

I-SPY TRIAL IT Infrastructure



Research Environment

SNPArray Data

Expression Array Data

Patient Samples

Clinical Care Environment

Clinical Data

Radiological Data

caTissue

caArray

API

caBIG® Applications

caIntegrator

API

Data Mart

Tolven

caExchange - Hub

API

caBIG® Services to Local System