caBIG™ cancer Biomedical Informatics Grid™ an initiative of the National Cancer Institute

Integrative Biomedical Research Design Patterns, HPC, Semantic Interoperability and Grid

Joel Saltz, MD PhD
Director Center for Comprehensive Informatics
Emory University

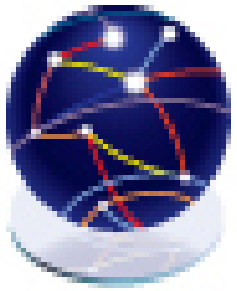
National Heart Lung and Blood Institute

EMORY UNIVERSITY

CVRG

Biomedical Informatics Consortia
What are these guys up to anyway?

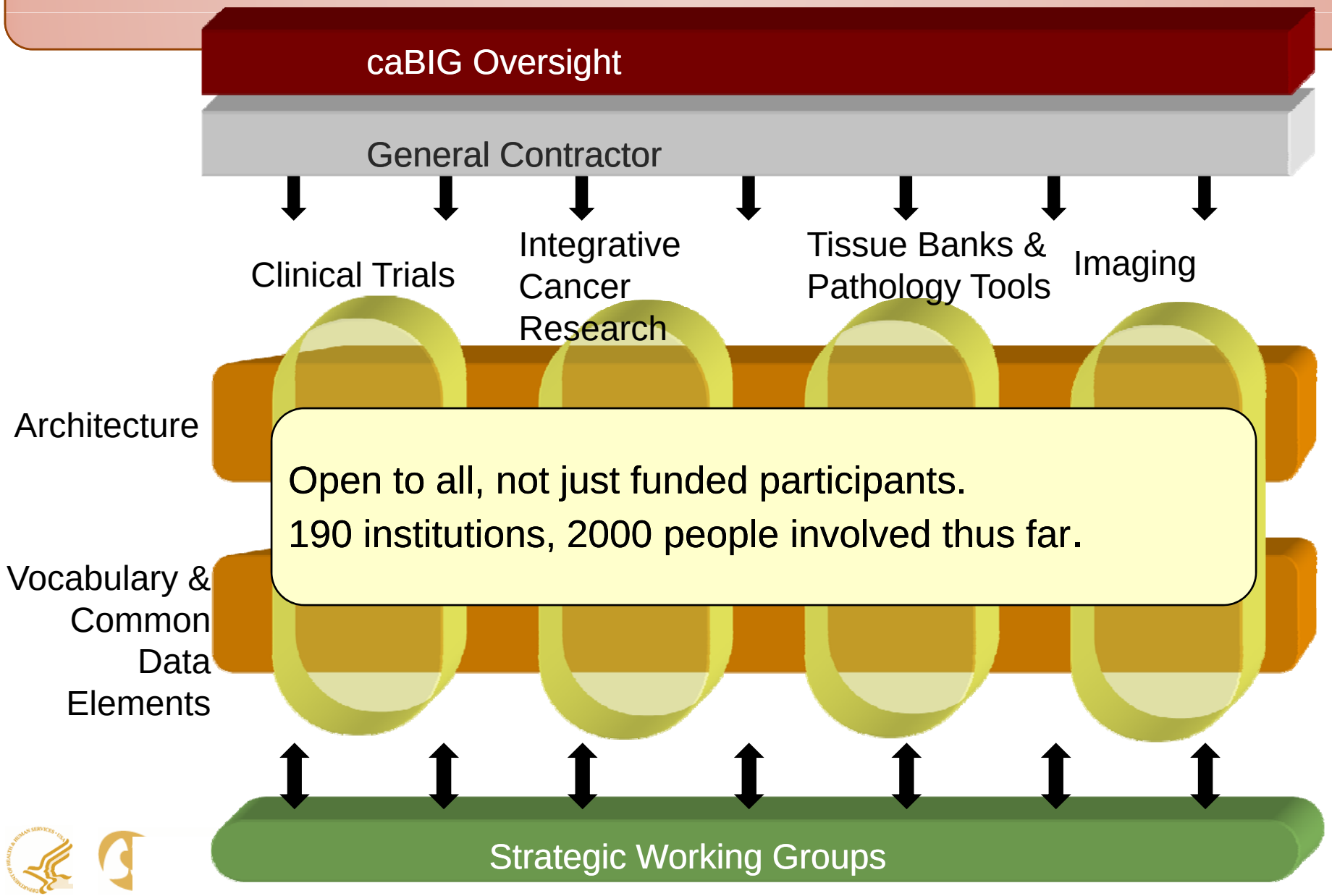
CTSA Clinical & Translational Science Awards



CVRG

CVRG

Example: caBIG Organization Structure



"Big" Design Patterns for Translational Research

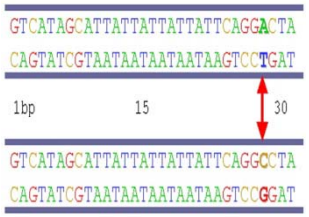
- Deep Integrative Analyses
- Multiscale Investigations that encompass genomics, epigenetics, (micro)anatomic structure and function

National Heart Lung and Blood Institute

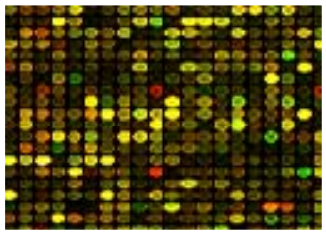
caBIG™ cancer Biomedical Informatics Grid™

CVRG

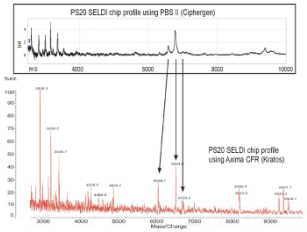
The Reynolds Study



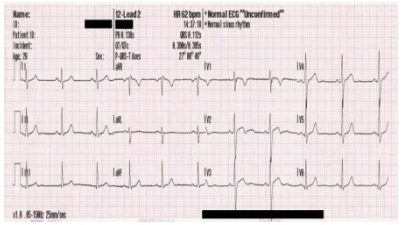
Genetic Variability



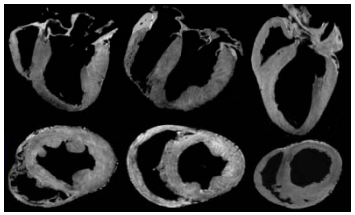
Gene Expression Profiling



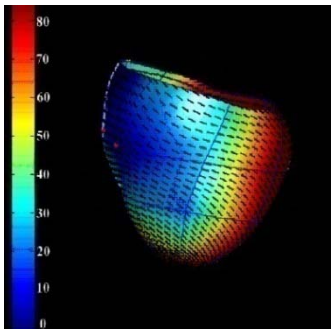
Protein Expression Profiling



Electrophysiological Data



Multi-Modal Imaging



Data Analysis And Modeling

- Prospective clinical research study – Deep Integrative Clinical Analysis
- Large subject cohort (~ 1,200) at high risk for sudden cardiac death
- All have:
  - CAD
  - LV dysfunction
  - received ICD placement
- Multi-scale data from each patient
- Challenge – discover biomarkers predictive of high risk
- Test biomarkers on novel (currently ~500) subject population

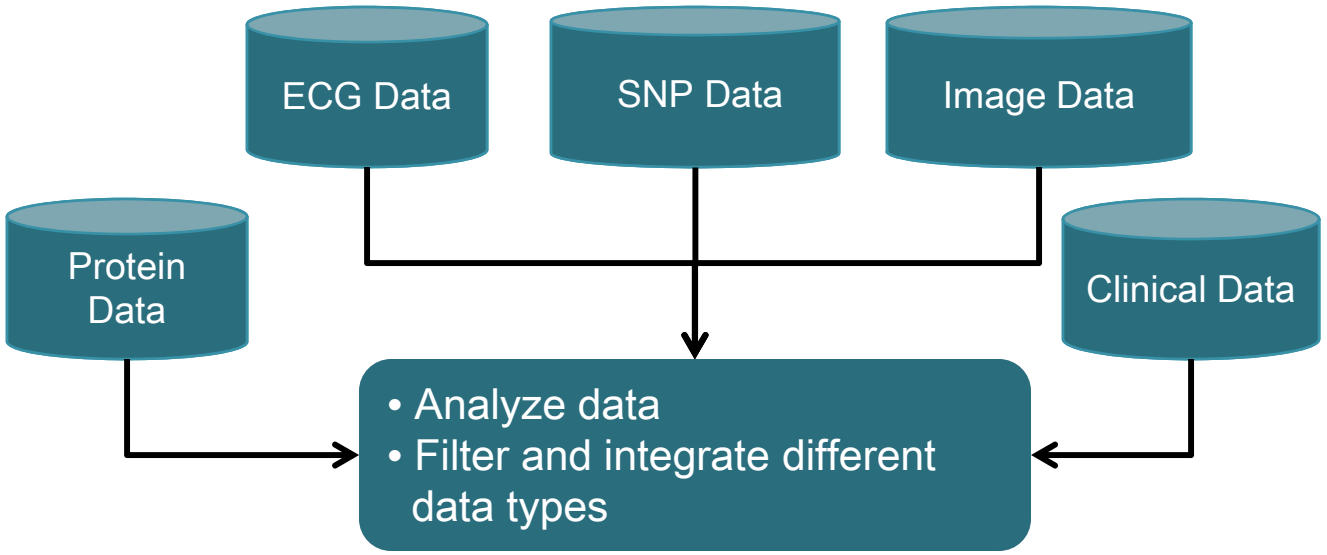
National Heart Lung and Blood Institute

caBIG cancer Biomedical Informatics Grid

CVRG

Data Analysis and Exploration: Multi-Scale Cardiovascular Data

- Investigate genotype-phenotype characteristics among a subset of patients in the Reynolds study
- Combine features across different levels of biological organization
  - SNP
  - mRNA
  - Protein
  - Imaging
  - Electrophysiology (ECG)
  - Clinical



CVRG: Primary Aims

- Support collaborative cardiovascular research
  - Integrative data analysis using heterogeneous, distributed resources
  - Securely share data and analysis methods with collaborators
  - Establish common set of services, data sources, vocabulary and common data elements for cardiovascular research community
  - Leverage caGrid, caBIG™, BIRN
  - Initial driving application is the Reynolds study -- an example of deep integrative clinical analysis
  - PI – Rai Winslow PhD, Center spans Hopkins, Emory, UCSD, Ohio State

National Heart Lung and Blood Institute

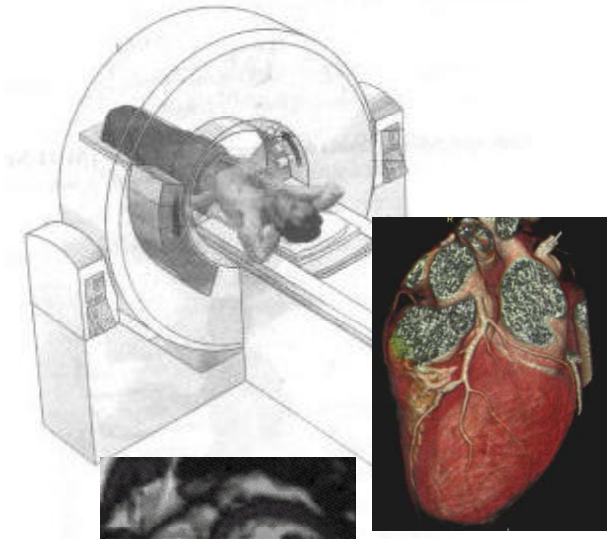
caBIG™ cancer Biomedical Informatics Grid™

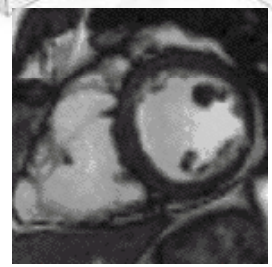
CVRG

Biomedical Informatics Services

- Security
- Semantic interoperability
- Data structure interoperability
- Interoperability with existing standards (e.g. HL-7, DICOM)
- Ability to compose services to create application
- Ability to efficiently invoke HPC services
- Efficient and expressive federated query

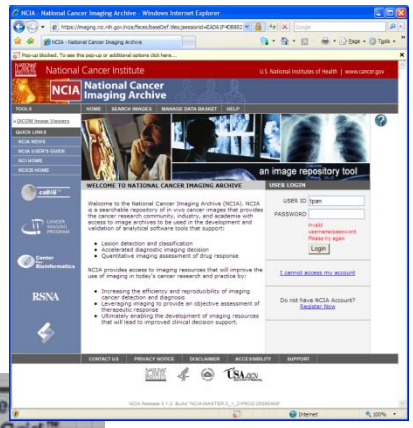
Image Management Workflow





Functional Image Analysis

(Virtual) PACS



AIM Annotation Database

National Heart Lung and Blood Institute

caBIG™ cancer Biomedical Informatics Grid™

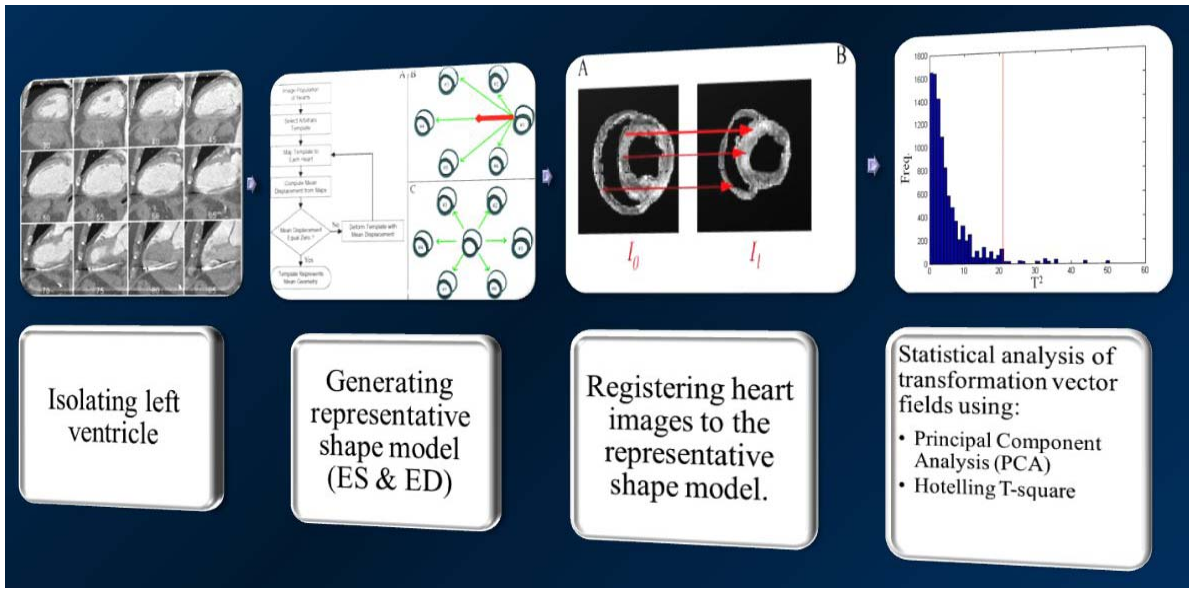
CVRG

CT Cardiac Shape Analysis Workflow

Heart Images (Human CT)

Isotropic preprocessing and segmentation to isolate left ventricle.

Affine Registration



Generate Templates for end-systole and end-diastole phases

Register Heart Images to template and perform LDDMM mapping

Statistical Analysis (Principal Component Analysis)

National Heart Lung and Blood Institute

caBIG™ cancer Biomedical Informatics Grid™

CVRG

CALGB INTERSPORE ACRIN NCICB

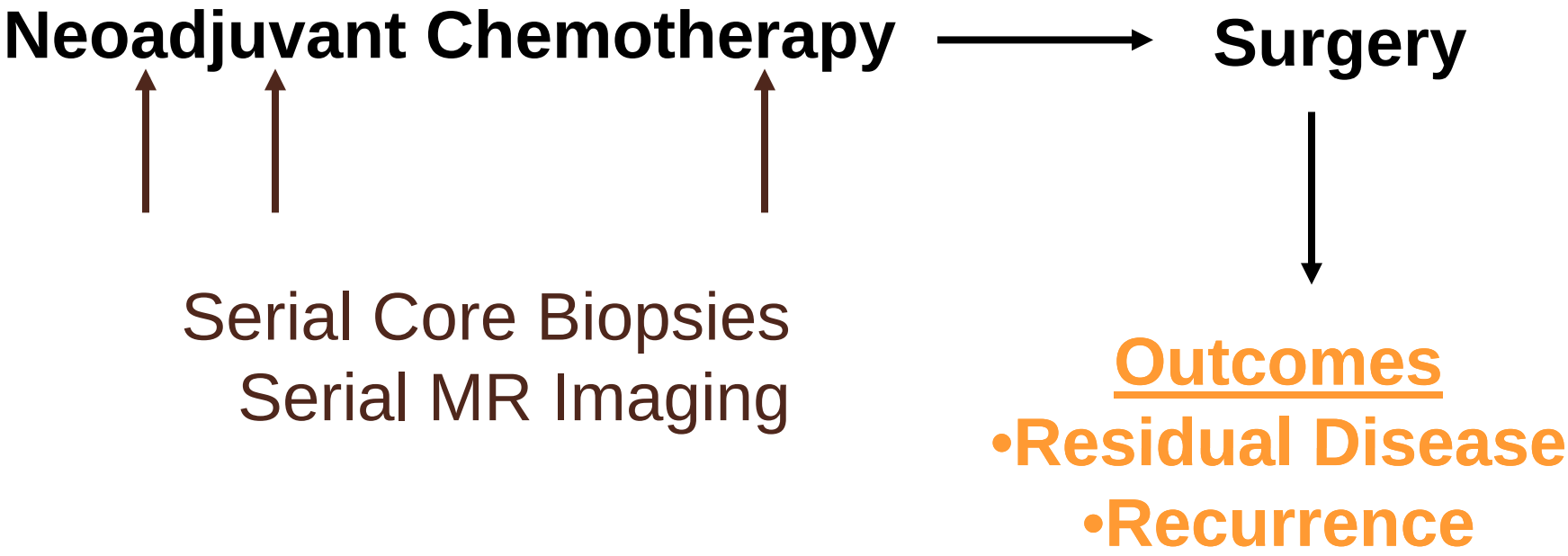
Investigation of
Serial studies to
Predict
Your
Therapeutic
Response with
Imaging and
And
moLecular analysis

I SPY WITH MY LITTLE EYE . . . . . . A BIO-MARKER BEGIN-ING WITH X . . . .

National Heart Lung and Blood Institute

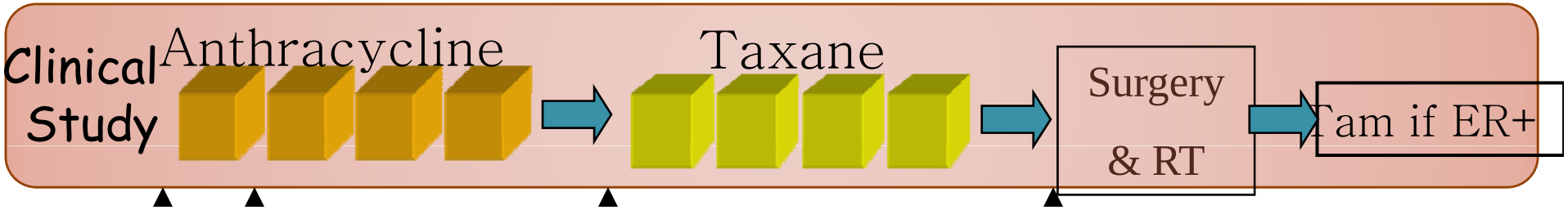
CVRG

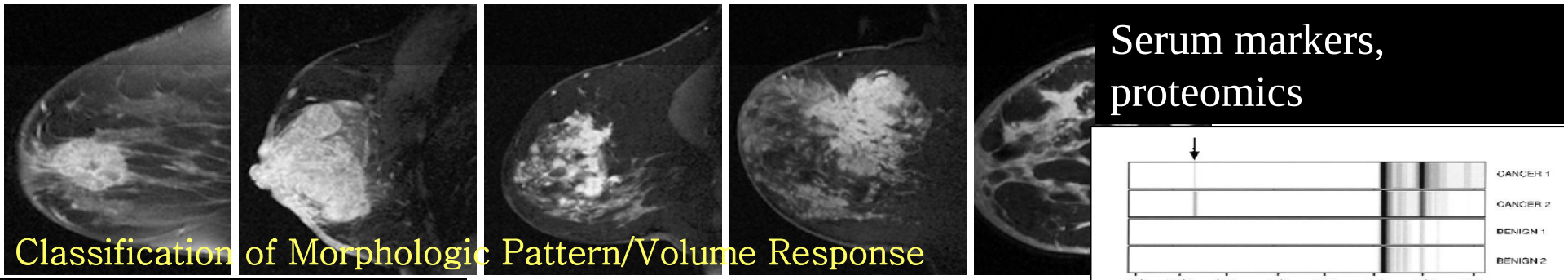
I SPY TRIAL Design



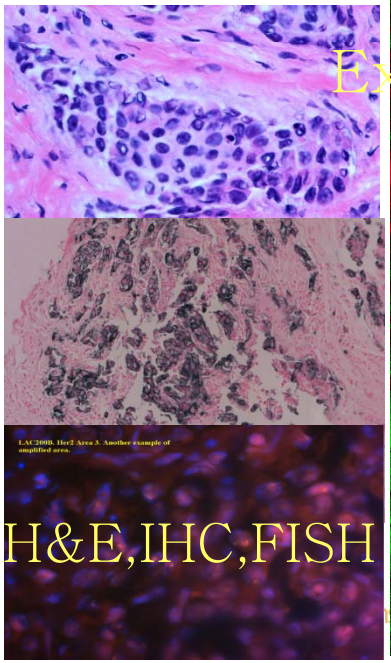
National Heart Lung and Blood Institute

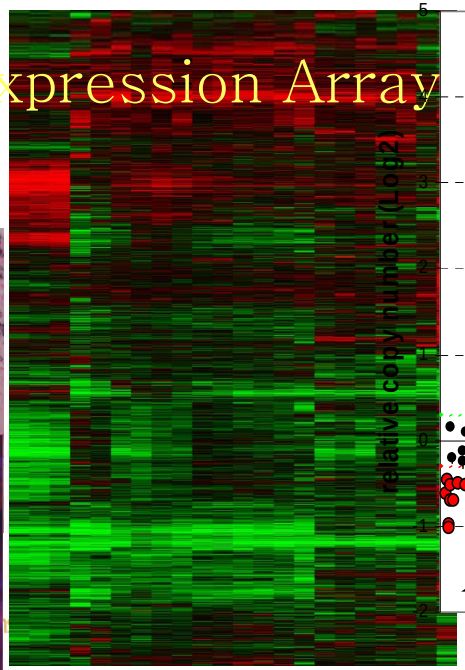
CVRG

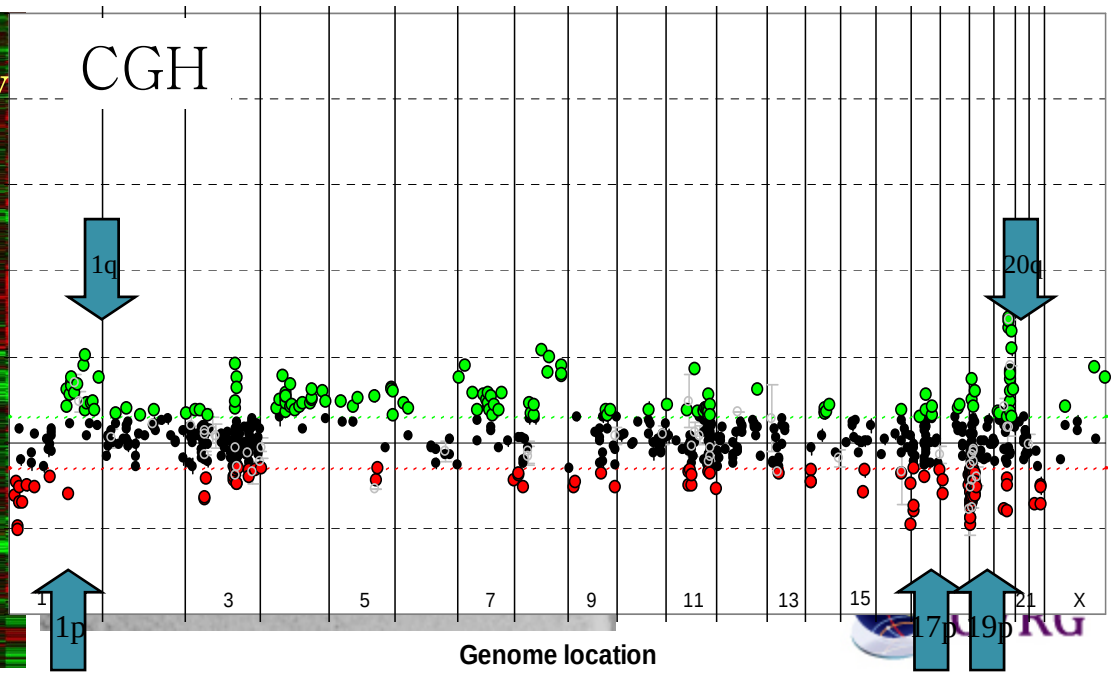




Tissue: Core or Surgical

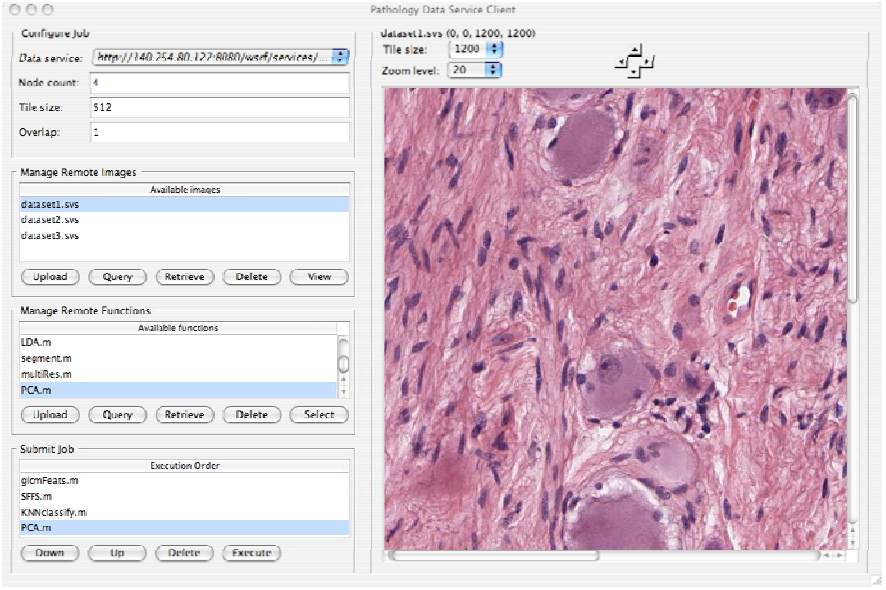






Pathology Coordinated Review

Multiheaded Microscope



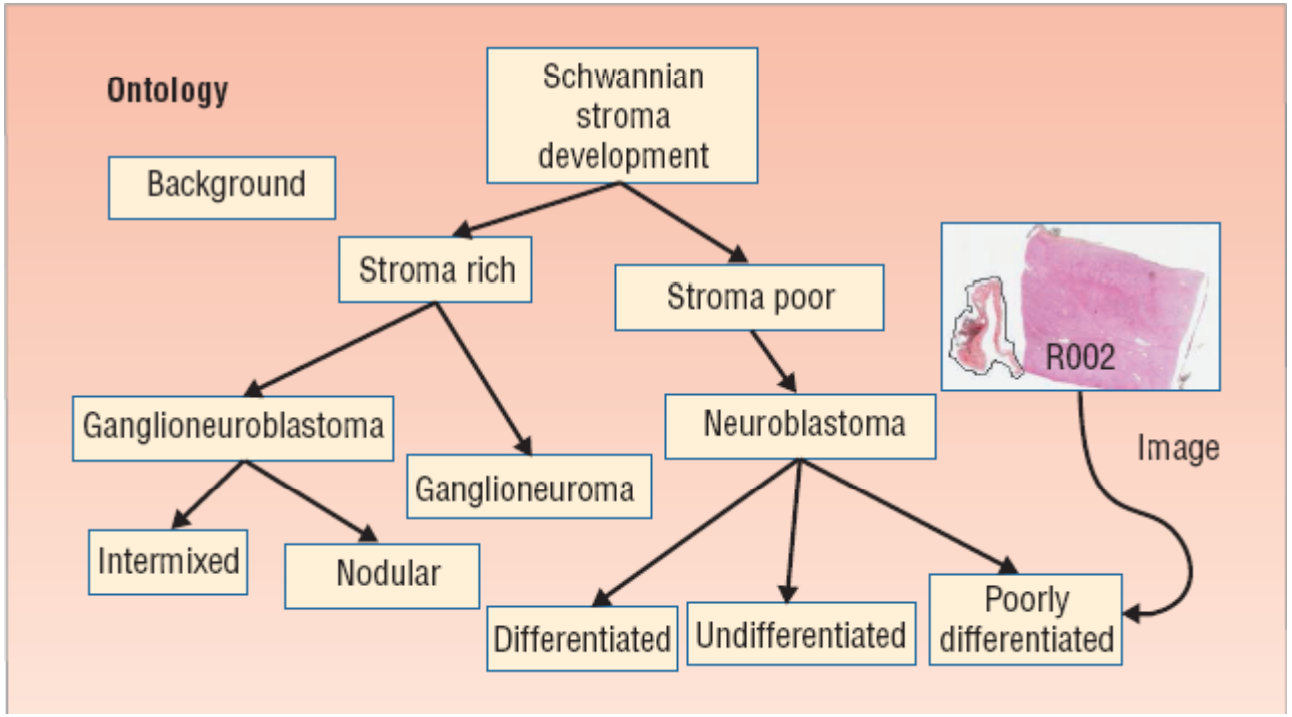
caMicroscope

National Heart Lung and Blood Institute

CVRG

Computer-assisted Histopathology

- Analyze images by computer
- Analyze the whole tissue, several slides
- Provide quantitative information to the pathologist
- Reduce inter- and intra-reader variability



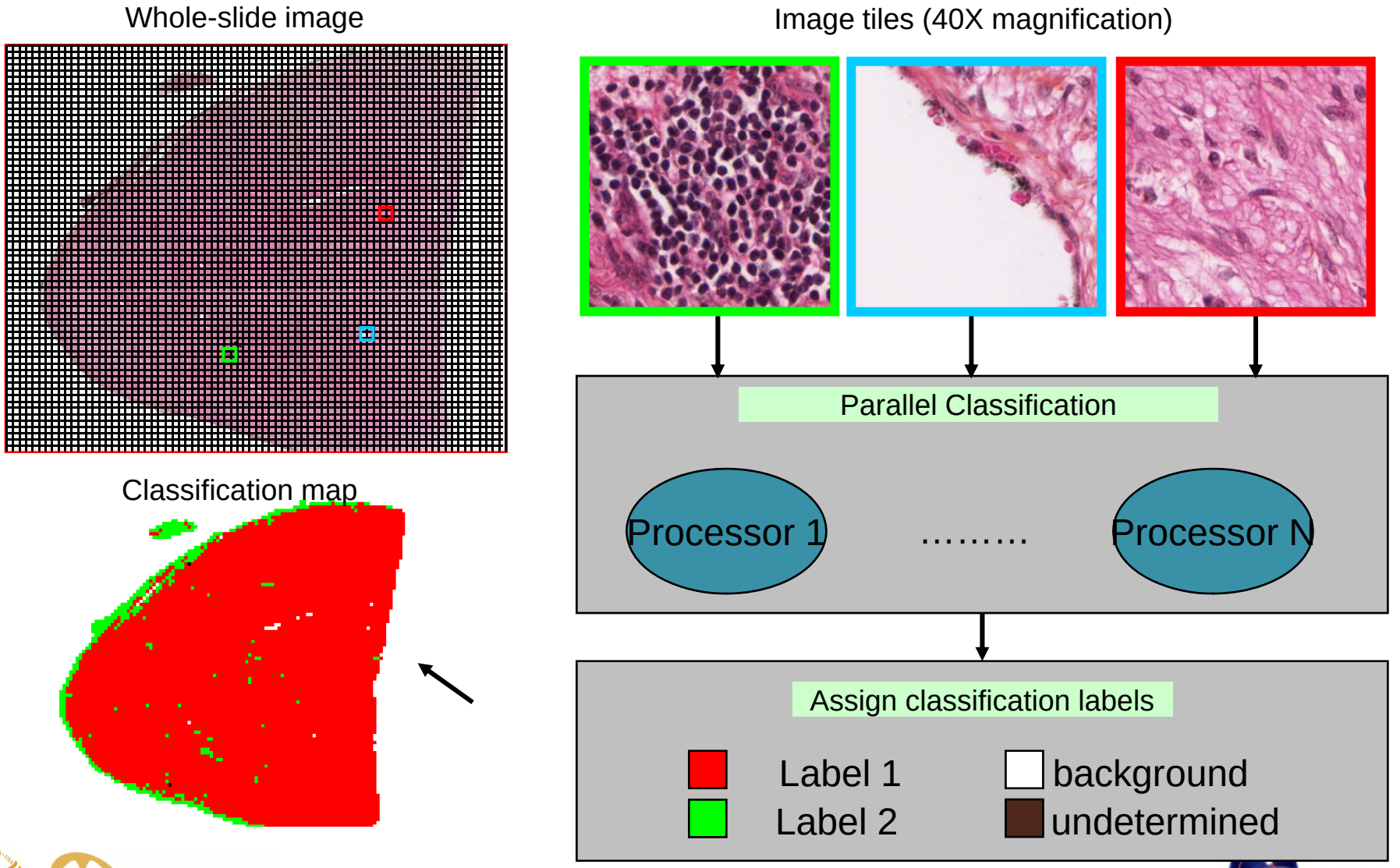
Morphological characterization of tissue used for prognosis

Neuroblastoma – Shimada Classification
(Gurcan-OSU, Shimada – LA Children's)

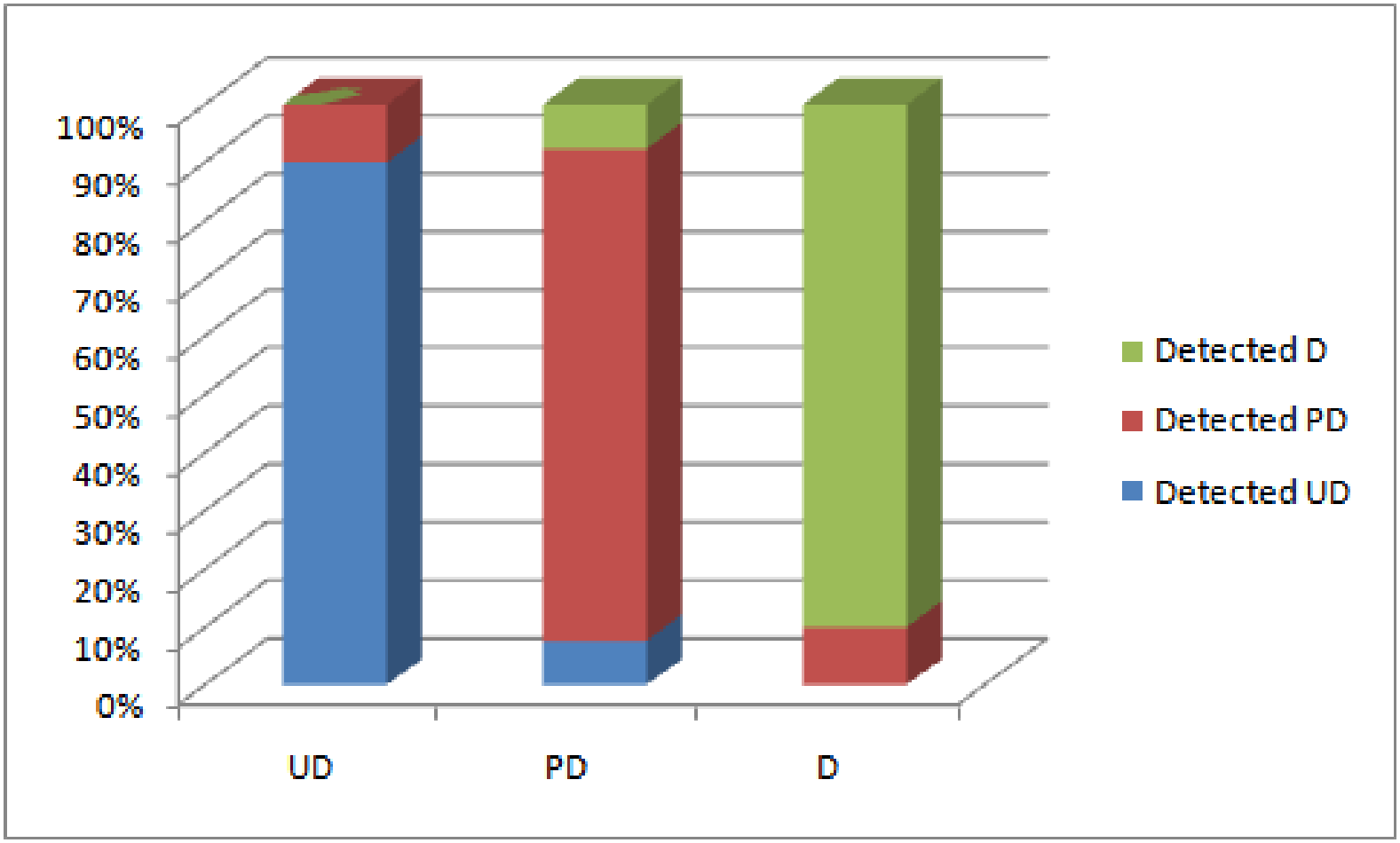
National Heart Lung and Blood Institute

CVRG

caMicroscope parallel processing caGrid/caOS/DataCutter



Example Algorithm Results: Neuroblastoma Grade of Differentiation



UD: Undifferentiating
PD: Partially differentiating
D: Differentiating

National Heart Lung and Blood Institute

caBIG™ cancer Biomedical Informatics Grid™

CVRG

Design Pattern Driven Requirements

- *Semantics: Design template involves deep integration of many types of information to synthesize knowledge*
- *Interoperability: Information drawn from commercial/enterprise systems e.g. health information records, PACS, Lab information systems, as well as genetic, genomic, epigenetic, microscopy databases*
- *HPC requirements arise from many sources: natural language processing, whole genome analyses, coordinated analysis of multiple types of molecular, image data*

Design Pattern Driven Requirements

- *Composition of computationally modest and HPC services – caGrid, caOS, DataCutter*
- *Composition of services written in multiple languages running in varied environments – Wings/Pegasus/Taverna/Introduce/gRAVI*
- *Workflow engines capable of efficient inter-service large scale data transfer, security delegation – New caOS Workflow Engine*
- Libraries of optimized components/services – *GPU/Cell DataCutter libraries for image analysis*
- *Integrated analysis/human review may require soft real time response*

Design Pattern Driven Requirements

- *Flexibility: ability to accommodate different data formats, different semantic classifications*
- *Interoperability: composition of caGrid, myGrid, BIRN, CVRG and unaffiliated web services*
- *Goal of caGrid Roadmap – plug and play workflow scripting environment, service level execution environment, fine grained execution environment*
  - *e.g. Taverna, caGrid, caOS, DataCutter;*
  - *Wings, Pegasus, Condor, DataCutter;*
  - *WEEP, caGRID, MPI*

National Heart Lung and Blood Institute

caBIG™ cancer Biomedical Informatics Grid™

CVRG

"Big" Design Patterns for Translational Research

- Deep Integrative Analyses
- Multiscale Investigations that encompass genomics, epigenetics, (micro)anatomic structure and function

National Heart Lung and Blood Institute

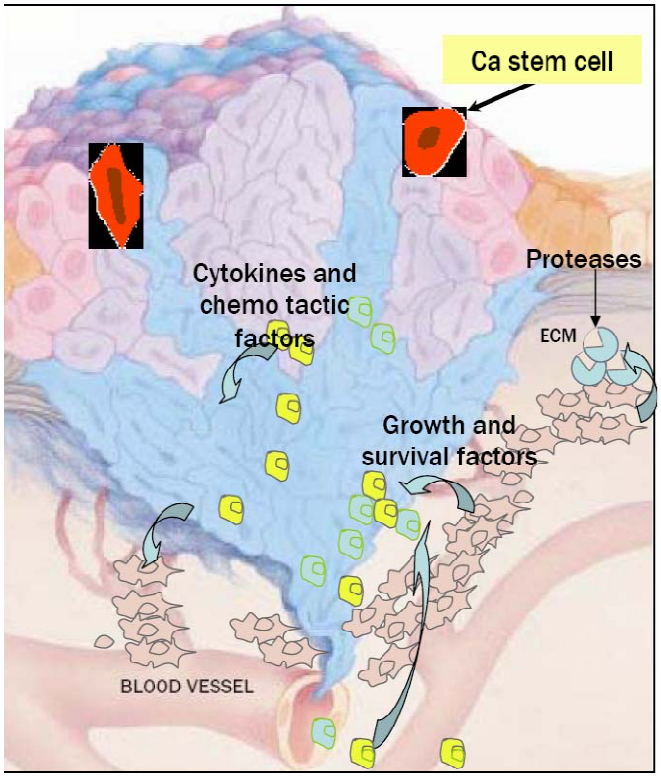
caBIG™ cancer Biomedical Informatics Grid™

CVRG

Tumor Microenvironment

- Cancer is a complex phenomenon
- A tumor is an organ
- Structural and functional differentiation within tumor
- Molecular pathways are time and space dependent
- "Field effects" – gradient of genetic, epigenetic changes
- *Experiments to elucidate integrate microscopy, high throughput genetic, genomic, epigenetic studies, flow cytometry, microCT, nanotechologies …*
- *Simulation is next frontier*

Tumors are organs consisting of many interdependent cell types



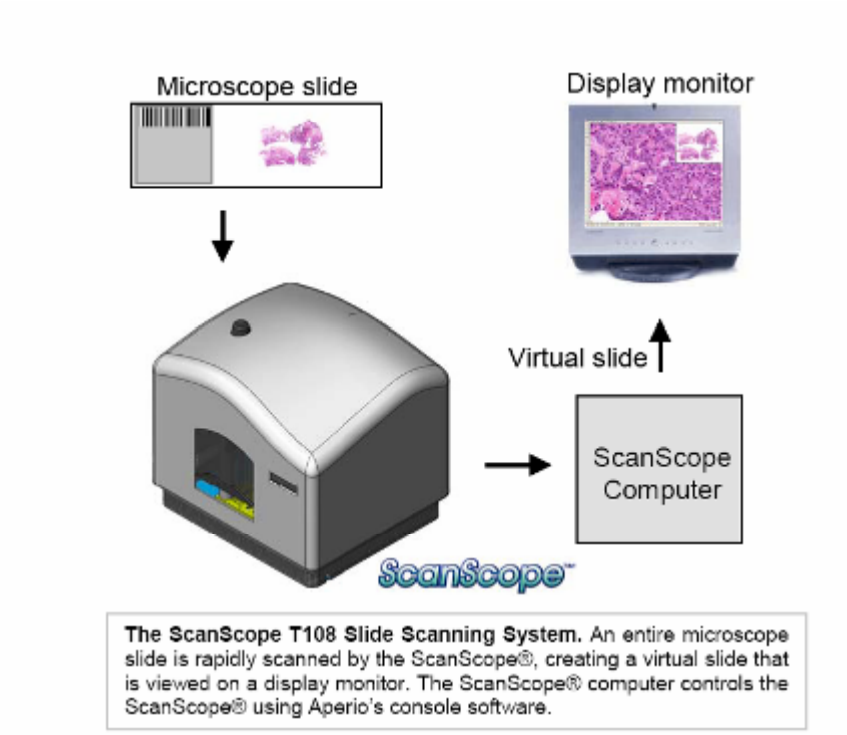
- **From John E. Niederhuber, M.D. Director National Cancer Institute, NIH**

National Heart Lung and Blood Institute

CVRG

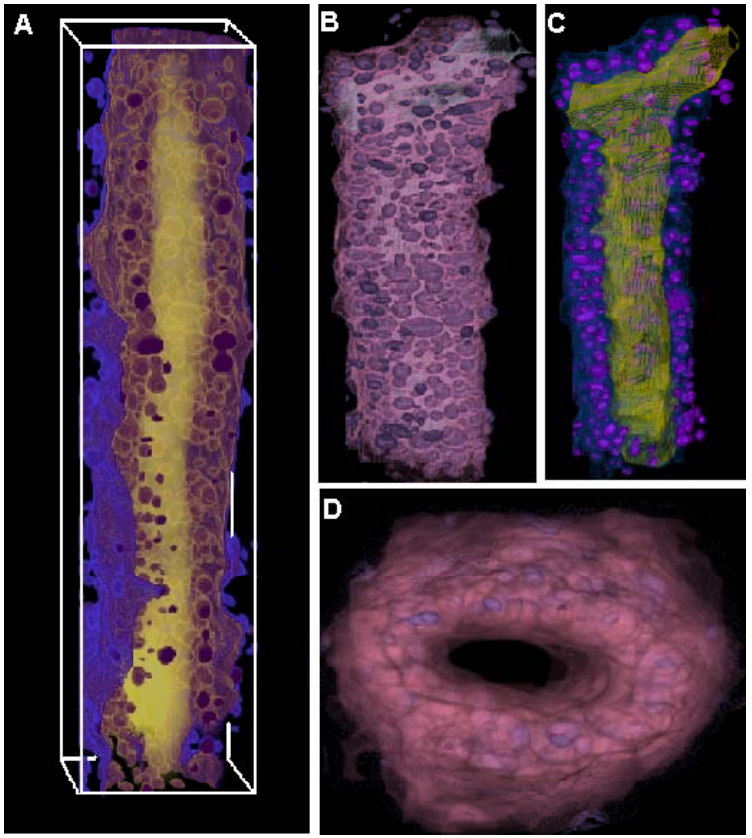
Tumor Microenvironment

Slide Scanning



The ScanScope T108 Slide Scanning System. An entire microscope slide is rapidly scanned by the ScanScope®, creating a virtual slide that is viewed on a display monitor. The ScanScope® computer controls the ScanScope® using Aperio's console software.

Ducts



National Heart Lung and Blood Institute

Imaging Team led by Raghu Marchiraj
Kun Huang OSU

CVRG

"GIS type service": Semantic Annotation and Spatial Reasoning

Ontology

- Endothelial cells touch blood vessel lumen
- Protein C is expressed only in endothelial cells

Instance Data

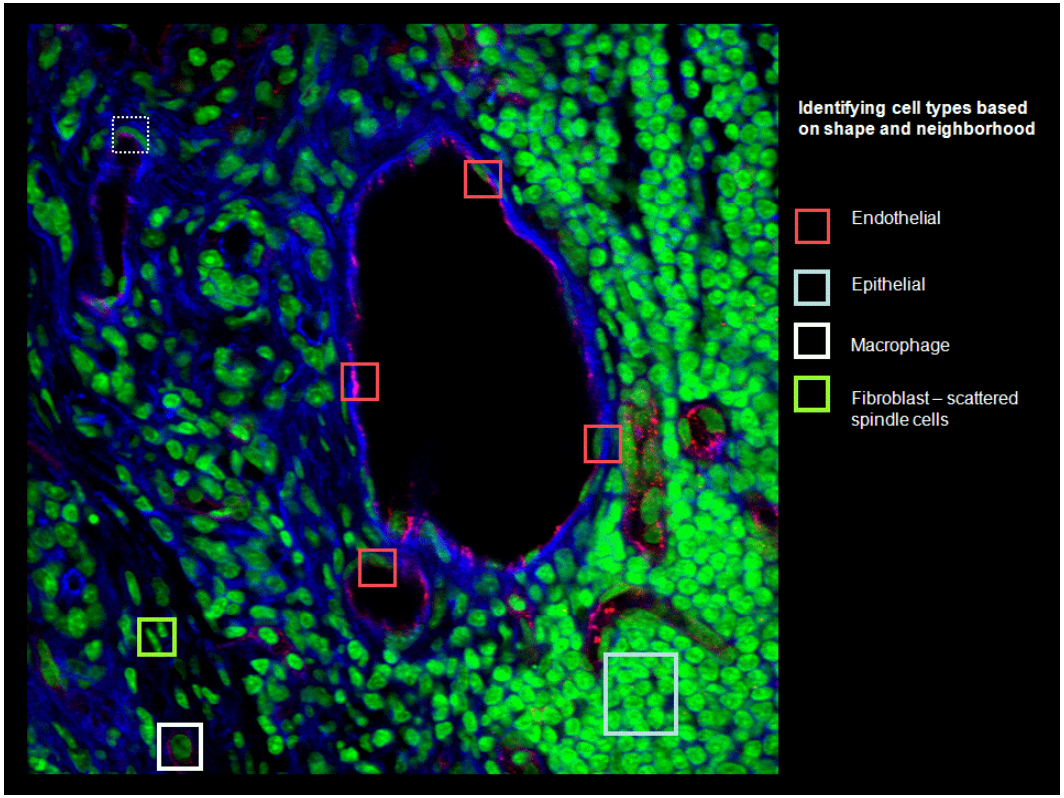
- Region A is a cell (from image analysis)
- Region A expresses protein C (from molecular assay)
- Region B (from expert markup)

Spatial Rule

- touches(Region B, Region A) – algorithmically evaluates to true

Spatial and Ontological Inference

- Region A is an endothelial cell
- Region B is a blood vessel



National Heart Lung and Blood Institute

CVRG

Mouse Placenta: Understand function of Rb gene

Letters to Nature

Nature **421**, 942-947 (27 February 2003) | doi: 10.1038/nature01417

Extra-embryonic function of Rb is essential for embryonic development and viability

Lizhao Wu[1], Alain de Bruin[1], Harold I. Saavedra[1], Maja Starovic[2], Anthony Trimboli[1], Ying Yang[3], Jana Opavska[1], Pamela Wilson[1,4], John C. Thompson[4], Michael C. Ostrowski[4,5], Thomas J. Rosol[5,6], Laura A. Woollett[7], Michael Weinstein[4,5], James C. Cross[2], Michael L. Robinson[3,5,8] and Gustavo Leone[1,4,5]

The retinoblastoma (*Rb*) gene was the first tumour suppressor identified[1]. Inactivation of *Rb* in mice results in unscheduled cell proliferation, apoptosis and widespread developmental defects, leading to embryonic death by day 14.5 (refs 2–4). However, the actual cause of the embryonic lethality has not been fully investigated. Here we show that loss of Rb leads to excessive proliferation of trophoblast cells and a severe disruption of the normal labyrinth architecture in the placenta. This is accompanied by a decrease in vascularization and a reduction in placental transport function. We used two complementary techniques—tetraploid aggregation and conditional knockout strategies—to demonstrate that *Rb*-deficient embryos supplied with a wild-type placenta can be carried to term, but die soon after birth. Most of the neurological and erythroid abnormalities thought to be responsible for the embryonic lethality of *Rb*-null animals were virtually absent in rescued *Rb*-null pups. These findings identify and define a key function of Rb in extra-embryonic cell lineages that is required for embryonic development and viability, and provide a mechanism for the cell autonomous versus non-cell autonomous roles of Rb in development.

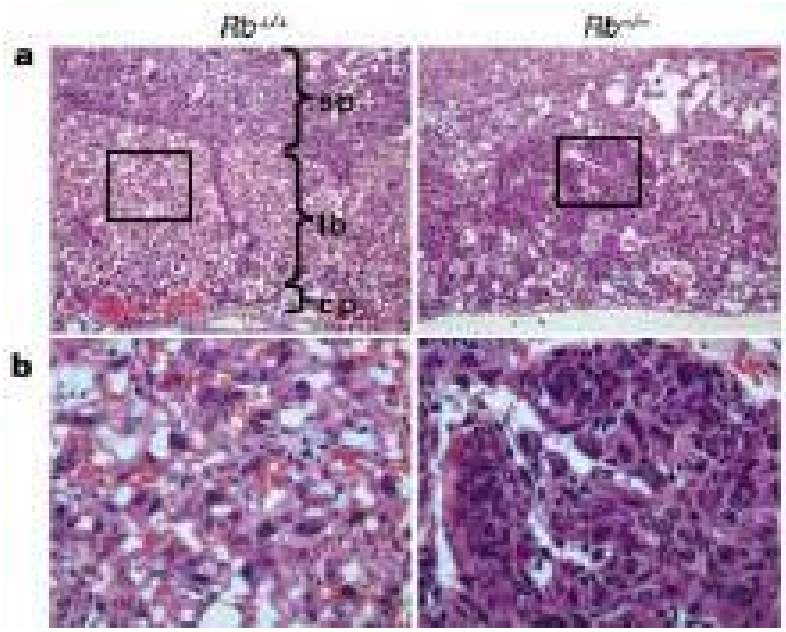
National Heart Lung and Blood Institute

caBIG™ cancer Biomedical Informatics Grid™

CVRG

Wild vs Mutant

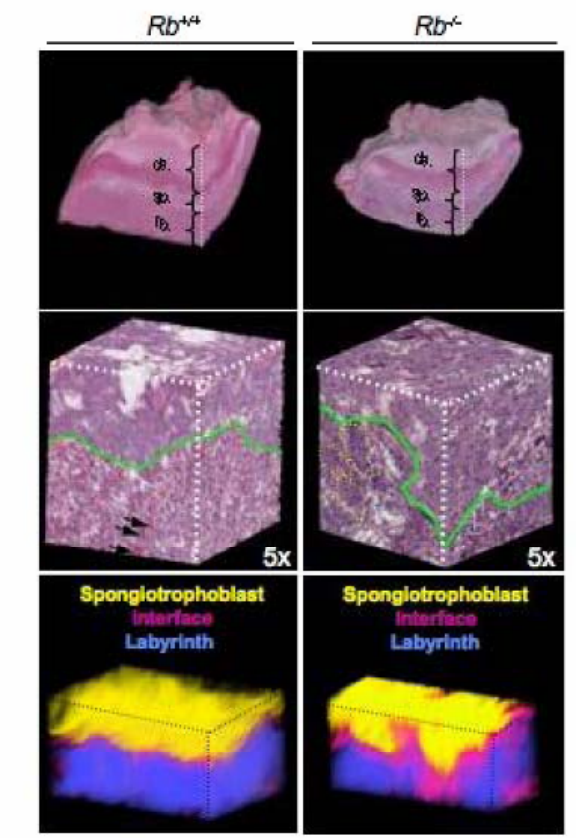
Wild type - Labyrinth neat, well-ordered, maternal blood sinusoids and trophoblasts evenly dispersed among fetal blood cells.



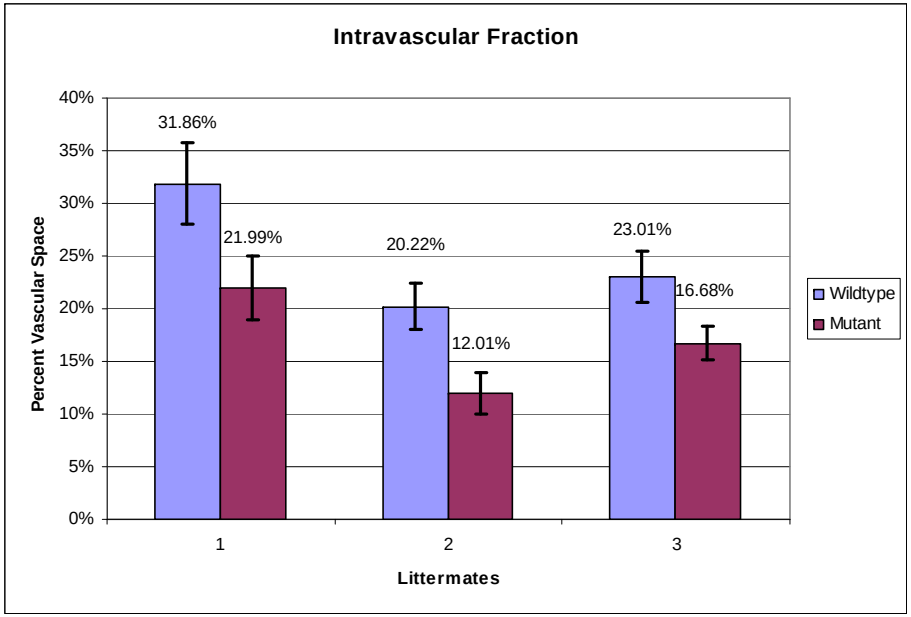
Mutant - Trophoblasts grow wildly, clump together and disrupt fetal and maternal cells layers necessary for proper embryonic growth

Wild Type vs Mutant: Analysis of Entire Placenta

3-D Reconstruction



Quantitative tissue analysis



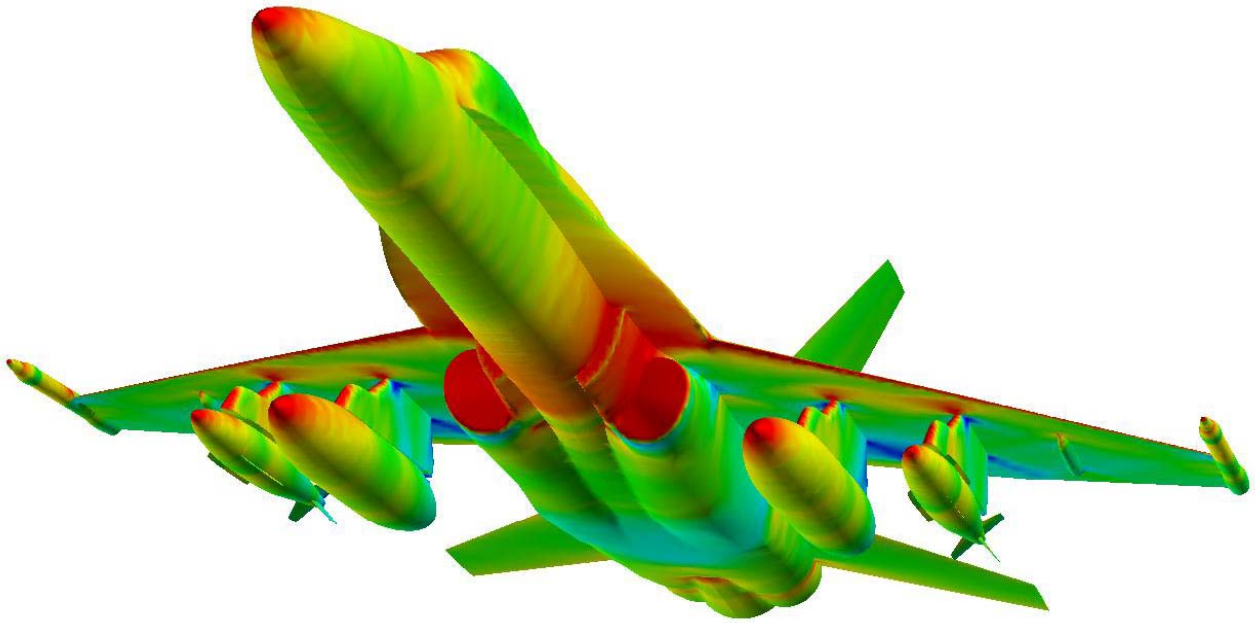
National Heart Lung and Blood Institute

CVRG

Design Pattern Driven Requirements for Multiscale

- Complex, hierarchical annotation of microanatomic structures; molecular composition: "*ducts, a specific duct, epithelial cells surrounding a specific duct, a particular epithelial cell in the neighborhood of a particular duct, the nucleus of a specific epithelial cell in the neighborhood of a particular duct …*"
- Spatial/semantic queries: *What is the morphological/molecular effect on cell type 1 if we make a genetic change in cell type 2*
- Algorithm annotation and composition: *Interoperability critically dependent on semantic modeling of application domain*
- *Interplay between spatial and molecular data underlies increasing fraction of biomedical research studies – "GIS type" service*

When are we going to get serious about simulation?



National Heart Lung and Blood Institute

caBIG™ cancer Biomedical Informatics Grid™

CVRG

caGrid Roadmap planning process
Engage the Computer Science Systems Software Community

Data and Analytic Services – Present and Future

- Easy integration of existing database systems, High-performance Grid Nodes, multi-core systems, on-demand computing, data intensive computing, parallel database and file systems.

Workflows and Orchestration

- Interoperability between different workflow execution environments; hierarchical workflow systems; HPC and large scale data support

Federated Query

- Semantic, federated, spatial query support

Semantic Infrastructure

- Semantic annotations for services, relationship between semantics and data structures, systematic curation vs community freedom, semantic query support.

Security

- Security middleware support for complex organizations, complex workflows. Compliance with regulatory guidelines

Acknowledgments

The caGrid team:caGrid 1.0: Scott Oster, Stephen Langella, Shannon Hastings, David Ervin, Ravi Madduri, Tahsin Kurc, Frank Siebenlist, Ian Foster, Krishnakant Shanbhag, Peter Covitz

The caOS team: Renato Ferreira, Shannon Hastings, Umit Catalyurek

Parameterized workflow project: Mary Hall, Yolanda Gil, Ewa Deelman, Tahsin Kurc, Vijay Kumar, Varun Ratnaker, Jihie Kim

OSU Imaging Algorithm Team: Raghu Machiraju, Metin N. Gurcan Ph. D. , Kun Huang Ph.D, Kishore Mosaliganti, Lee Cooper, Antonio Ruiz, Olcay Sertel

The Imaging Informatics/HPC team: Tony C. Pan M.S., Ashish Sharma Ph.D., Manuel Ujaldon Ph.D. Olcay Sertel Antonio Ruiz, Vijay Kumar Sivaramakrishnan Narayanan,

Umit V. Catalyurek Ph.D

CVRG: Rai Winslow PhD (PI), Project 1 team: Mark Ellisman, Tahsin Kurc, Justin Permar, Steven Granite, Jeff Grethe, Anthony Kolasny, Tony Pan, Justin Permar

Tumor Microenvironment PIs: Mike Ostrowski, Gustavo Leone

Advanced Technology Consortium, QARC, ITC: TJ Fitzgerald, Jim Purdy, Walter Bosch

Eliot Siegel, Paul Mulhorn, Michael McNitt-Gray, all SMEs and participants in the caBIG in-vivo imaging workspace

12 years of virtual microscope: Alan Sussman, Umit Catalyurek, Tahsin Kurc, Henrique Andrade, Renato Ferreira ….

Carole Goble and the myGrid team

caGrid Teragrid Team Members

- **geWorkbench (Columbia University)**
  - Christine Hung (ch2514@columbia.edu)
  - Kiran Keshav (keshav@c2b2.columbia.edu)
- **caGrid (Ohio State University)**
  - Scott Oster (oster@bmi.osu.edu)
  - Stephen Langella (langella@bmi.osu.edu)
- **caGrid/TeraGrid (Argonne National Laboratory)**
  - Ravi Madduri (madduri@mcs.anl.gov)
- **TeraGrid (Argonne National Laboratory)**
  - Stuart Martin (smartin@mcs.anl.gov)
- **TeraGrid (Texas Advanced Computing Center)**
  - Stephen Mock (mock@tacc.utexas.edu)
- **Management**
  - Aris Floratos (Columbia University)
  - Krishnakant Shanbhag (Argonne National Laboratory)
  - Michael Keller (Booz Allen Hamilton)
  - Patrick McConnell (Duke University)
  - Nancy Wilkins-Diehr (San Diego Supercomputer Center)

National Heart Lung and Blood Institute

caBIG™ cancer Biomedical Informatics Grid™

CVRG

CVRG Acknowledgements

- **Department of Biomedical Informatics, The Ohio State University**
  - Joel Saltz
  - Tahsin Kurc
  - Justin Permar
  - Tony Pan
  - Stephen Langella
- **Center for Research in Biological Systems, University of California, San Diego**
  - Mark Ellisman
  - Jeff Grethe
  - Ramil Manansala
- **Institute for Computational Medicine, Johns Hopkins University**
  - Raimond L. Winslow
  - Michael I. Miller
  - J. Tilak Ratnanather
  - Stephen J. Granite
  - Anthony Kolasny
  - Aaron Lucas
  - Kyle Reynolds
  - Tim Brown
  - Bryan Schwam
  - David Hopkins

National Heart Lung and Blood Institute

caBIG™ cancer Biomedical Informatics Grid™

CVRG

National Heart Lung and Blood Institute

caBIG™ cancer Biomedical Informatics Grid™

CVRG