**TITLE:** (200 character limit)

Life Sciences Research Enabled by caBIG® Tools for Data Integration, Discovery and Analysis

**AUTHORS:**

Rashmi Srinivasa, Mervi Heiskanen, JP Marple, Karen Ketchum, Shine Jacob, Dan Kokotov, Anand Basu, Aris Floratos, Ian Fore, Michael Reich, Juli Klemm

With affiliations:

Mervi Heiskanen (NCI), JP Marple (5AM Solutions, Inc.), Karen Ketchum (Enterprise Solutions and Consulting), Shine Jacob (Enterprise Solutions and Consulting), Dan Kokotov (5AM Solutions, Inc.), Anand Basu (NCI), Aris Floratos (Columbia University), Ian Fore (NCI), Michael Reich (Broad Institute), Juli Klemm (NCI)

**ABSTRACT**: (2000 character limit)

The cancer Biomedical Informatics Grid® (caBIG®) is a collaborative network designed to accelerate the translation of discoveries from research to clinical care. This extensible informatics platform integrates diverse data types and supports interoperable software tools in clinical science, biospecimen management, imaging and discovery science. An institution can combine various caBIG® tools to form a comprehensive solution for data management as well as data integration, discovery and analysis. Data management solutions within the caBIG® Life Sciences domain include those for microarray data (caArray), biospecimens (caTissue Suite), nanoparticle data (caNanoLab) and Genome Wide Association Study data (caGWAS). To support the connection of data across these resources, there are integrative tools that allow the scientist to search data across different repositories connected to the grid, and analyze, integrate and visualize this data. caIntegrator2 allows users with no programming skills to set up a study-specific custom web portals that allow search and analysis across different data types. caBench-to-Bedside (caB2B) allows users to search array data, biospecimens, and nanoparticle data across instances of caArray, caTissue Suite, and caNanoLab, on the grid. In addition, both geWorkbench and GenePattern support integrative analysis and visualization of expression data, pathways and protein structures. Several of these tools have had recent releases: caIntegrator2 v1.1 supports lists of genes/subjects for query and analysis, and caBIO integration for pathway search. caArray v2.4 parses array data from Affymetrix, Illumina, Agilent, Nimblegen, Genepix and MAGE-TAB files, and data can be annotated using controlled vocabularies like the MGED Ontology. The NCI CBIIT caArray instance includes data from key NCI initiatives like TCGA, TARGET, MMHCC and ICBP, loaded using MAGE-TAB. geWorkbench 2.0 includes a comparative modeling pipeline and database of homology models for protein structures. caBIG®'s emphasis on a sSOA (semantic Service Oriented Architecture) ensures that a rich set of API methods is available to pass semantically meaningful data among applications. More information is available at <https://cabig.nci.nih.gov/>.