National Cancer Institute

Getting Connected with caBIG®

LIFE SCIENCES DISTRIBUTION

The caBIG® Life Sciences Distribution is a collection of Life Sciences software tools designed to facilitate the discovery of the next generation of cancer diagnostics and therapeutics and realize the vision of Molecular, or Personalized Medicine. These integrated tools support a variety of capabilities from tracking and managing biospecimens to analyzing and integrating microarray data. Together, they enable cancer researchers to more easily organize, integrate, analyze, and share data from disparate sources across caGrid.

All components of the caBIG® Life Sciences Distribution can be plugged into the national caGrid backbone that connects caBIG® informatics resources across research organizations. The individual tools included in this bundle are free and are available for immediate download and use in organizations, and the bundling of these related and synchronous tools will facilitate easier, more streamlined adoption at host institutions.

This document provides an overview of the Life Sciences Distribution. It describes what the bundle is designed to do, its features and benefits, and the requirements for implementing the tools.

Capabilities and tools included in this bundle:

- Biobanking management system [caTissue Core]
- Virtual clinical data repository [Clinical Trials Object Data System (CTODS)]
- Genome-wide data management system [Cancer Genome-Wide Association Studies caGWAS)]
- *In vivo* image repository [National Cancer Imaging Archive (NCIA)]
- Microarray data management system [caArray]
- Microarray gene expression and sequence data management [geWorkbench]
- Cross application search and organization of data through the Life Sciences Distribution Browser [LSDB]
- caBIG®-compatible systems architecture [caGrid]

The Life Sciences Distribution was designed to facilitate the National Cancer Institute's (NCI) overarching goal to connect the people, institutions, and data in the cancer community through caBIG®. This collection of tools and capabilities is one of three "bundles" that have been designed to help support and streamline clinical trials, imaging, tissue banking, and integrative cancer research, and it provides the materials needed to join the secure caBIG® data-sharing framework.

Visit https://caBIG.nci.nih.gov/inventory for more detailed information and for access to caBIG® resources.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

Getting Connected with caBIG®

LIFE SCIENCES DISTRIBUTION

Tools	Description	Benefits
caArray	caArray is a microarray data management system that guides the annotation and supports the exchange of array data and also connects with analytical tools like geWorkbench.	• Provides both Web browser-based and programmatic access to microarray data • Facilitates integration of array data with other diverse data types including clinical, imaging, tissue, and functional genomics data through harmonization with relevant caBIG® models
caTissue Core	caTissue is a biobanking management tool to collect, manage, process, annotate, request, and distribute biospecimens and associated information.	• Provides Web browser-based and programmatic access to biospecimen data • Manages tissue, fluid, cell, and molecular biospecimen information • Allows users to find and request specimens needed for use in molecular correlative studies
geWorkbench	geWorkbench is a desktop bioinformatics platform that offers a comprehensive and extensible collection of tools for the management, analysis, visualization, and annotation of microarray-based gene expression and sequence data.	• Enables integrated access, analysis, and visualization of genomic data (gene expression, sequence, pathway, structure) • Enables sophisticated analysis of genomic data through the integration of visualization tools, external databases, and computational services
National Cancer Imaging Archive (NCIA)	NCIA is a searchable repository of *in vivo* cancer images, such as CT, MRI, and Digital X-rays. NCIA also contains annotation files (PDF, image markup) and annotation data provided by a curator.	• Serves as a platform for image data management and integration with other research data types, including clinical and genomic data • Enables development of imaging resources that will lead to improved clinical decision support • Accelerates diagnostic imaging decision making and quantitative imaging assessment of drug response
Life Sciences Distribution Browser (LSDB)	LSDB is a Web-based, free-form search tool that allows users to easily search across local and remote instances of LSD applications.	• Allows users to work with local data (data at their institution), NCICB data (data stored at the NCI), and publicly available data from across caGrid • Searches NCIA for images and caArray for microarrays with the same search criteria, providing scientists with the ability to quickly perform cross domain queries and validate research hypotheses
Cancer Genome-Wide Association Studies (caGWAS)	caGWAS allows researchers to integrate, query, report, and analyze a variety of data types from multiple sources including microarray, genomic, immuno-histochemistry, imaging, and clinical data through a single application.	• Facilitates rapid sharing of information and results analysis from various biomedical studies • Allows researchers and bioinformaticians to access and analyze clinical and experimental data across multiple clinical studies
Data Repository	**Description**	**Benefits**
Clinical Trials Object Data System (CTODS)	CTODS is a database and software system for storing and sharing clinical trials data in both identifiable and de-identified form.	• Enables a cancer research organization to utilize data from any in-house caBIG®-compatible Clinical Trials Data Management System (CDMS) or data source for non-clinical research • Enables a cancer research organization to share or access de-identified clinical trials data (data that have all patient identification information removed) over the national caGrid
Infrastructure	**Description**	**Benefits**
caGrid	caGrid is a service-oriented architecture and federation that connects caBIG®-compatible systems together regardless of where they are installed.	• Query across data resources installed in different locations • Automatically integrate comparable data from different sources • Create workflow pipelines for data retrieval and analysis using resources across the grid

caBIG®
cancer Biomedical Informatics Grid®

Features

caArray

- Web-based forms for researcher submission and annotation of data consistent with MIAME guidelines
- Bulk data import of MAGE-TAB files
- Group-based permission scheme including settings for publishing data to the public domain
- Search and navigate features to readily discover and extract data of interest
- Submission and retrieval of Affymetrix, GenePix, and Illumina native expression array files, Affymetrix and Illumina native SNP files

caTissue Core

- Tracking of multiple specimens and refined materials (RNA, DNA, Protein) from the same participant
- Role-based permission scheme for repository personnel and researchers
- Tracking of quality assurance, distribution, derivation, and aliquotting of biospecimens
- Flexible storage container structure

geWorkbench

- Analysis and visualization tools for microarray-based gene expression profiling data from a variety of systems including Affymetrix MAS5/GCOS, Matrix format (geWorkbench), RMAExpress, and GenePix
- Analysis and visualization tools for gene and protein sequence data (FASTA); plug-in components including filter and normalize, promoter analysis, regulatory networks, protein structure, raw and differential expression, enrichment analysis, annotation, sequence analysis, and pattern discovery
- Support for pathways (BioCarta), gene ontologies, networks, and patterns based on regular expressions
- Integrated access to many external data sources and computational services (e.g., GoldenPath at Santa Cruz, NCBI BLAST, BioCarta diagrams through caBIO)
- Front-end to caArray and access provider to caGrid-enabled computational resources

NCIA

- Searchable repository of *in vivo* cancer images along with key annotations
- Access to image archives and imaging resources
- Facilitates development and validation of analytical software tools that support lesion detection and classification software, accelerate diagnostic imaging decisions, and quantify imaging assessment of drug response

LSDB

- Ability to query across caArray and NCIA applications installed locally or remotely
- Provides a front-end onto the local installation of LSD, so that institution users can easily access and query across their LSD applications from a single point

caGWAS

- Standardized model to represent SNP genotype data, SNP association findings, population frequency data, and clinical phenotype
- Support for search and retrieval of GWAS findings in the context of genes or chromosomal regions of interest
- Allows users to load GWAS studies and provides powerful search and download capabilities for large or small datasets

NOTE - Due to recent developments in forensic science, it is recommended that human specimen data stored in caGWAS not be connected to the caGrid.

Features

CTODS

- Based on open standards and standards-based tools designed to enable the cancer research community to share, interpret, and integrate de-identified information
- Consistent with the Biomedical Research Integrated Domain Group (BRIDG) model that underpins data interchange standards and technology solutions that enable harmonization between the biomedical/clinical research and healthcare arena

Features

caGrid

- Nationally-deployed, standards compliant, data and analysis grid from which any caBIG®-compatible system can plug into or draw
- Business Process Execution Language (BPEL) workflow engine for data analysis pipeline construction and execution
- Federated, cross-domain data mining and integration
- Standardized programming interfaces for application developers

BUNDLE REQUIREMENTS

The caBIG® Life Sciences Distribution is multi-platform, and it will run on any appropriately powered and configured systems capable of running the underlying software infrastructure. Windows, Linux, and MacOSX operating systems are all supported. Check the caBIG® tools Web page (https://cabig.nci.nih.gov/tools) for the most up-to-date information on the system requirements outlined below.

SUPPORTING SOFTWARE

- Apache Ant
- Apache Axis
- Java Development Kit (JDK)
- JBoss Application Server
- MySQL Database
- Hibernate
- Common Security Module: CSM
- Globus, part of caGrid installation
- Other software: MIRC T29-a and Cedara I-Response Workstation (IRW) (for NCIA), Castor (for CTOM)

RESOURCES

- Specific tool information: https://cabig.nci.nih.gov/tools
- caGrid information: https://cabig.nci.nih.gov/workspaces/Architecture/caGrid
- Overview of caBIG®: http://cabig.cancer.gov
- Detailed information about caBIG®, including training compatibility, etc: https://cabig.nci.nih.gov
- For general information about "Getting Connected with caBIG®": https://cabig.nci.nih.gov/getting_connected

CONTACT

caBIGconnect@cancer.gov

NIH Publication No. 08-6382
Printed March 2008
Online Version Updated October 2008