***Adoption of caBIGTM Tools: Application to Acute Myelogenous Leukemia (AML) and Neuro-Oncology***

Data Access Plan

Rakesh Nagarajan, M.D., Ph.D.

660 S. Euclid Avenue, Campus Box 8118

Saint Louis, MO 63110

PHONE : 314-362-8859

FAX: 314-747-7999

EMAIL: rakesh@wustl.edu

Mark Watson, M.D., Ph.D.

660 S. Euclid Avenue, Campus Box 8118

Saint Louis, MO 63110

PHONE : 314-454-7919

FAX: 314-454-5208

EMAIL: watsonm@wustl.edu

**Background**

Nucleic acid microarrays (e.g. expression, comparative genomic hybridization [aCGH], and SNP profiling) are a powerful technology that is heavily utilized to unravel the molecular genetic basis of complex diseases such as tumorigenesis. At Washington University School of Medicine (WU) and the Alvin J. Siteman Cancer Center (SCC), these technologies have been leveraged to conduct almost 1,200 microarray expression experiments alone. These include data for over 11,500 chips derived from human and model organism (e.g. mouse, rat, zebrafish, and drosophila) tumor specimens from virtually every organ site. In addition, the cancer community at WU/SCC has conducted many experiments employing aCGH and SNP chips from NimbleGen and Affymetrix respectively. This rapid influx of large scale data sets has resulted in several major needs by our cancer research community. These include:

1. A centralized, standards-compliant repository for microarray data. This would not only enable users to:
	1. Access their data remotely.
	2. Facilitate re-analysis of data in other contexts (e.g. normal prostate tissues that were originally profiled against tumor specimens may be compared to normal specimens from other organ sites to identify tissue-specific markers).
	3. Facilitate meta-analyses. Repositories such as OncoMine (<http://www.oncomine.org>) have been used to conduct such meta-analyses successfully across many profiling experiments performed against the same tumor type (e.g. prostate tumor data sets generated by different research groups).
	4. Facilitate data sharing between collaborators prior to publication.
	5. Make their data available to the entire research community in a MIAME-compliant format after publication.
2. A standards-based biospecimen repository that contains quality assurance/quality control (QA/QC) data as well as associated pathology and clinical annotations that is **ALSO** semantically interoperable with the microarray repository described in #1 above.
3. Tools to analyze and visualize this microarray and biospecimen information in the context of genome annotation. While independent tools exist to analyze gene expression data alone (e.g. commercial tools such as SpotFire DecisionSite and Ingenuity Pathways as well as our own Function Express client) at WU/SCC, there are no tools which facilitate the entire analysis workflow (e.g. combine microarray data with biospecimen and clinicopathology information, filter, analyze, and visualize in the context of genome annotation). Furthermore, even basic user-friendly (i.e. for a bench scientist or translational researcher) tools to analyze SNP data do not exist to date.

**End Users**

The Alvin J. Siteman Cancer Center established the Bioinformatics Core in 2000 to support the informatics needs of the cancer research community. The Bioinformatics Core serves over 500 investigators across ~200 independent laboratories spread across the WU campuses. While our goal is to serve the needs described above to this entire community, we will engage two specific research groups that are engaged in complex translational research projects across three tumor types. These include:

1. Dr. Timothy Ley (Genomics of Acute Myelogenous Leukemia Program Project Grant [GAML-PPG])- Dr. Ley and his collaborators are integrating comprehensive clinical information, pathology and cytogenetics data, and profiling data sets (expression, aCGH, SNP, proteomics, and mutation profiling) to determine the molecular mechanisms underlying AML.
2. Dr. David Gutmann (Neuro-Oncology Research Group [NORG])- Dr. Gutmann and his collaborators are studying a wide range of peripheral and central nervous system tumors and are correlating molecular markers with clinical observations such as outcome and survival. The major tumor types being studied are juvenile pilocytic astrocytomas (JPAs) and malignant peripheral nerve sheath tumors (MPNSTs).

Details of each research group’s data sets to date (more data is being continuously generated) are provided in the following table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tumor Type** | **Patients** | **Samples** | **Expression Arrays** | **NimbleGen aCGH** | **SNP Arrays** |
| AML | 564 | 4,490 | 483 | 235 | 100 |
| JPA |  93 |  174 |  110 | 10 | 20 |
| MPNST | 38 | 38 | 38 | 8 | 38 |
| Oligodendroglioma | 7 | 7 | 7 |  0 | 0  |

**Proposal**

Our proposal to support the informatics needs of the GAML-PPG and the NORG is composed of the following five elements (See Figure 1):

1. caArray Adoption - caArray will be utilized to manage and annotate microarray expression data initially. When caArray begins to support SNP and aCGH chips, these data sets will be managed in caArray as well. Our current metadata model will be mapped onto the caArray model, and gaps (e.g. attributes in caArray that are missing in our data model) will be identified. These gaps will be filled by data curators within the SCC Bioinformatics Core working closely with end users within the GAML-PPG and NORG groups. While this task is manually intensive, we firmly believe it is a crucial step for sharing of data with an appropriate level of accompanying metadata. After mapping and data curation, we will import both the data (raw and processed) and metadata. Note that this process will eventually enable us to utilize caArray for ALL of our microarray data management needs (i.e. for all projects) although, only the GAML-PPG and NORG projects will be specifically supported within the scope of this proposal. Another benefit of this adoption is that for ALL **future** microarray experiments,



caArray will be utilized by the SCC Bioinformatics Core for data management, dissemination, and publication.

1. caTissue Adoption- All PPG- and NORG-related biospecimen information is currently stored in the SCC Tissue Procurement Facility’s (TPF) caTissue Core. We will migrate from caTissue Core to caTissue Suite when a production release is available.
2. Analytical Services- To analyze microarray expression data, we will adopt the services exposed by GenePattern (Consensus Clustering and Comparative Marker Selection) and geWorkBench (Hierarchical, k-means, and SOM clustering). Note that as part of the caBench-to-Bedside (caB2B) project, we are already integrating such services. As part of this proposal, we anticipate the need to install some of these services locally both for performance reasons as well as proprietary issues (i.e. investigators may not be comfortable having their ***pre-publication*** data sent across the internet for analysis at another organization or institution and the service providers may not be will/able to accommodate our analytical needs). Furthermore, we will adopt BioConductor’s DNAcopy for analysis of aCGH and SNP genotyping data. Finally, we will **develop** novel analytical services by wrapping other existing BioConductor/R packages and Affymetrix software for similar copy number analyses. These include PLASQ, GLAD, aCGH, RLMM, and Affymetrix Power Tools (DM and BRLMM algorithms). Note that as part of the caB2B project, we have already created an analytical service for identifying coexpressed genes (Template Matching Algorithm), thus demonstrating our competency in developing caGrid analytical services.

**Figure 1.** Each of the five elements of our proposal is displayed. These include caArray adoption and data migration, caTissue adoption and data migration, existing analytical service adoption as well as new analytical service creation, caGrid Workflow engine user interface use cases and requirements, and integration of the first four elements through caB2B.

1. Workflow Use Cases and Requirements- Microarray analyses are commonly composed of serial (acquire, filter, normalize, cluster), parallel (cluster using several algorithms), and iterative (perform iterations of data randomization and clustering to determine family-wise error rate [i.e. to correct for multiple comparisons]) steps. Performing such steps manually is very time consuming (See Figure 2). Such steps may be choreographed as one or more workflows to facilitate such analyses. In fact, caGrid has developed a workflow service that is based on Business Process Execution Language (BPEL). However, the current interface is too cumbersome and is not end user friendly (e.g. a BPEL document is required as input). Therefore, working with our end users in the GAML-PPG and NORG groups, we will define the use cases and requirements for visually describing workflows by examining the range of analyses that are performed by these groups (See Figure 3). These will be given to the caB2B/caGrid development team for design and implementation.

**Figure 2:** At the GeneChip Core, Affymetrix chips are scanned to generate image files (DAT). The Data Transfer Tool (DTT) is then used to import data into the GeneChip Operating System (GCOS), and GCOS is used to generate probe level data (CEL). Genotyping Analysis Software (GTYPE) uses DM/BRLMM to make genotyping calls and generates CHP files. Copy Number Analysis Tool (CNAT) uses the CHP files to calculate Copy Numbers (CNT) files. CHP files may be exported as TXT files. End users then need to run analysis tools that accept one or more file types individually and manually. Tools and file formats include dChip (CEL, CHP, and TXT), Partek (CEL and CNT), CNAG (CEL and CHP), and DNAcopy (TXT).

**Figure 3:** Using the CEL/TXT files generated by the GeneChip Core, end users may leverage many SNP analysis tools simultaneously and seamlessly through caB2B’s workflow interface. These SNP analysis tools will be exposed to the end-user as analytical services using caB2B.

1. Integration- caB2B will be utilized to acquire microarray and biospecimen data from caArray and caTissue Core (and later caTissue Suite) and will then be used to analyze and visualize these data sets. Using the workflow interface built as a result of #4 above, users may perform complex analyses that utilize computational services as well as annotation services. For example, “find all amplified and deleted genes as well as those genes whose expression is perturbed in AML/JPA in samples with poor outcome but NOT in those with good outcome, and show me these genes in the context of a literature-based gene network.” This would require:
	1. caArray for acquisition of expression/aCGH/SNP data
	2. caTissue Core/Suite for acquisition of outcomes information associated with biospecimens used for the array analysis.
	3. analytical tools capable of finding genes with perturbed expression (Comparative Marker Selection) and identifying copy number changes (BioConductor/R/Affymetrix tools)
	4. caFE for literature-based gene network information

**Data Sharing**

Biospecimen and microarray data that are part of this project will be shared with other institutions and organizations through the deployment of caTissue Core/caTissue Suite and caArray instances, respectively, on publicly available servers. We currently manage and run >50 servers within the SCC Bioinformatics Core where the vast majority of servers are publicly accessible. In fact, we are one of the few proponents within the caBIGTM community of a SINGLE caTissue database and application server instance architecture where a single database behind the institutional firewall is connected to by a server running the web application within a de-militarized zone (DMZ). This server in the DMZ is publicly accessible and runs the caTissue instance using secure socket layer (SSL). For data sets where protected health information (PHI) is not involved (e.g. microarray data), our servers (database and application) our fully accessible through the Wide Area Network (WAN) as well. Examples include <http://bioinformatics.wustl.edu> (production instance), <http://geneconnect.wustl.edu> (production instance), and <http://catisuecore.wustl.edu> (our caTissue demo site). All of our production servers have been available >95% of the time since the inception of the SCC Bioinformatics Core.

Data sets that are anticipated to be shared include:

* Microarray-based gene expression data- Raw (e.g. Affymetrix DAT, CEL, and DTT files) and processed files (e.g. Affymetrix MAS 5.0 analysis results)
* Microarray-based whole genome genotyping (SNP) data- Raw (e.g. Affymetrix DAT, CEL, and DTT files) and processed files (e.g. Affymetrix GTYPE/CNAT/CNAG analysis results)
* aCGH data- Processed files (e.g. NimbleGen processed results)
* Higher order analytical results from caB2B (e.g. clustering/classification results and cross-tool genomic segments that are amplified/deleted, or have loss of heterozygosity)
* De-identified, non-PHI biospecimen and related patient information (e.g. tissue site, clinical diagnosis, pathological status, and associated annotation information as they become available through caTissue Suite)

The following regulatory and proprietary issues are anticipated:

* IRB approval to share de-identified biospecimen information- To facilitate this step, Mark Watson, the director of our TPF, is completing a revised document for IRB review and approval. This [document](http://cabigcvs.nci.nih.gov/viewcvs/viewcvs.cgi/caties_adp_washu/IRB%20Requirements%20Packet/07Revised%2099-0573.pdf) (click to view) has been authored and made available as part of our caTissue adoption project. This will allow the TPF to use caTissue Core and in the future caTissue Suite for its biospecimen informatics needs. It will also enable this data to reside on a networked server that is accessible from any WAN IP using SSL. If the IRB denies our request for a single instance of caTissue, we will deploy a de-identified instance outside of the institutional firewall. This feature will be available within caTissue Suite v1.1 which will be available in March 2008. For the purposes of this project, if the IRB denies our request for a single instance of caTissue, we will deploy a de-identified instance and migrate biospecimen data relevant to the GAML-PPG and NORG groups.
* There are no anticipated regulatory (i.e. HIPAA) issues with microarray expression and aCGH data sets. For genome-wide SNP profiling data sets, it has been proposed that a complete SNP profile may constitute PHI. This issue is under active debate nationally. Thus, this issue needs to be addressed by the DSIC workspace for resolution. In addition, our local IRB will be notified of these issues, and they will need to set policies regarding such data. For example, Dr. Tim Ley has recently modified the IRB document for the GAML-PPG to include explicit consent to share genome-wide sequence information.
* There are no proprietary issues with regard to published microarray (expression, SNP, and aCGH) data sets. Some of the data sets referenced above have been already published, but are not easily accessible or well annotated using current systems. These will be made more accessible through publicly available caArray and caTissue Core/Suite servers as stated above. Note that the SAME servers will be used by both Washington University investigators (both members and non-members of the GAML-PPG and NORG) and investigators from the broader caBIGTM and biomedical research communities.

Data shared through the publicly available caArray and caTissue Core/Suite servers will be available indefinitely as these will ALSO be the same servers through which all future data sharing of such data sets will occur for the entire SCC Bioinformatics Core. Thus, as stated above, it is our intention to leverage and test the GAML-PPG and NORG use cases for data sharing and to apply the same process for ALL future microarray and biospecimen data sets generated by both cancer and non-cancer groups. For example, beyond the GAML-PPG and NORG specific data sets, our institution has well over 1,800 microarray data sets from human tumor biospecimens that will be represented in both caTissue Core / Suite and caArray. Many of these data sets have been previously published or have little or none of the proprietary constraints discussed above. Thus, specific work using GAML-PPG and NORG data sets will pave the way for the sharing of even larger cohorts of combined biospecimen and genomic data.