Cancer Molecular Analysis Project: Weaving a rich cancer research tapestry

Kenneth H. Buetow,[1] Richard D. Klausner, Howard Fine, Richard Kaplan, Dinah S. Singer, and Robert L. Strausberg[1]

National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892
[1]Correspondence: buetowk@pop.nci.nih.gov (K.H.B.), rls@nih.gov (R.L.S.)

The Cancer Molecular Analysis Project (CMAP) of the NCI is integrating diverse cancer research data to elucidate fundamental etiologic processes, enable development of novel therapeutic approaches, and facilitate the bridging of basic and clinical science.

The dawn of the post-genome sequence era brings the promise of a new day in cancer research, along with an unprecedented description of our genetic constitution and enthusiasm for industrial-scale approaches to previously vexing problems in cancer research. A myriad of emerging technologies are giving us insight into the transcriptome and the proteome. These are complemented by novel comprehensive approaches to manipulating biologic systems, such as chemical genetics. New imaging technologies give an extraordinary view of in vivo processes in real time. A key challenge facing biomedicine is how to maximize our efforts. These emerging technologies will fall short of their promise if the information they produce cannot be used to inform one another. It is not practical to catalog all the potential combinations and permutations of interest. The potential scientific insights from the synthetic examination of data are beyond enumeration.

Not surprisingly, early results from these technologies indicate that cancer is a disease of daunting complexity. Emerging from this apparent complexity is the understanding that cancer may actually arise from the combination of a finite number of fundamental molecular processes. Extracting these insights from the collections of observations generated by the modern technologies represents an almost overwhelming task. Translating these insights into novel therapeutic approaches confronts additional obstacles. Feeding back observations made in the clinical setting to refine our knowledge of these fundamental processes is formidable, yet critical.

The promise that molecularly targeted cancer therapy provides compels us to confront these challenges and explore the benefits of truly integrated biomedical observation sets. To this end, the NCI has undertaken the Cancer Molecular Analysis Project (CMAP). Its primary goal is to facilitate the identification and evaluation of molecular targets in cancer. It will serve as an infrastructure through which the threads of insight obtained from our new technological capacity to examine cancer will be interwoven.

Designing the CMAP

The synthesis of knowledge from disparate data is a multitiered challenge. At a practical level, the management, integration, visualization, and interpretation of such large data collections is a significant technical challenge. This task is further complicated at a conceptual level. First, the observations are routinely generated on discrete data sets, making it difficult to identify interactions between alternative types of observations. A more insidious barrier to integration is that each type of observation is generated by a different portion of the scientific community. Differences in these communities present barriers to communication of important findings. These barriers slow the rate at which important discoveries in one domain can be translated into another domain.

One means through which CMAP is addressing these challenges is the application of information technology. To facilitate information interpretation and integration, CMAP utilizes a novel bioinformatics infrastructure that spans the diverse fields of biomedical research that constitute cancer research. Graphical presentations of information and straightforward virtual analytic tools are provided. These tools draw from a state-of-the-art information system being developed by the NCI's Center for Bioinformatics (NCICB, http://ncicb.nci.nih.gov). This infrastructure is a "knowledge stack" composed of a vocabulary/ontology layer, a common data element layer, and biomedical object layer (http://ncicb.nci.nih.gov/core).

CMAP currently draws data from multiple resources. Genomic information is obtained from the National Center for Biotechnology Information (NCBI, http://ncicb.nlm.nih.gov), University of California Santa Cruz (http://genome.ucsc.edu), and the NCI's Cancer Genome Anatomy Project (CGAP, http://cgap.nci.nih.gov). Information on molecular pathways is obtained from BioCarta (http://www.biocarta.com). Functional classification of genes is obtained from the Gene Ontology Consortium (GO, http://www.geneontology.org). Information on gene expression comes from CGAP's serial analysis of gene expression (SAGE) data and from the NCI's Developmental Therapeutics Program's (DTP, http://dcp.nci.nih.gov) cDNA microarray evaluation of the NCI 60 cell lines (NCI60) used for drug screening. Molecularly targeted therapeutic agent information is obtained from the NCI's Cancer Therapy Evaluation Program (CTEP, http://ctep.nci.nih.gov). Preclinical efficacy of agents on the NCI60 is obtained from DTP. Information on clinical trials is obtained from CTEP and the NCI's Office of Cancer Communications. The CMAP data and infrastructure are publicly accessible (http://cmap.nci.nih.gov).

The information presented to the community is accessed through different high-level organizational views that help researchers approach the fully integrated dataset within a contextually familiar environment. Currently, there are four entry points to CMAP information: molecular profiles, molecular targets, targeted agents, and trials of targeted agents. These entry points roughly approximate the steps associated with selection,

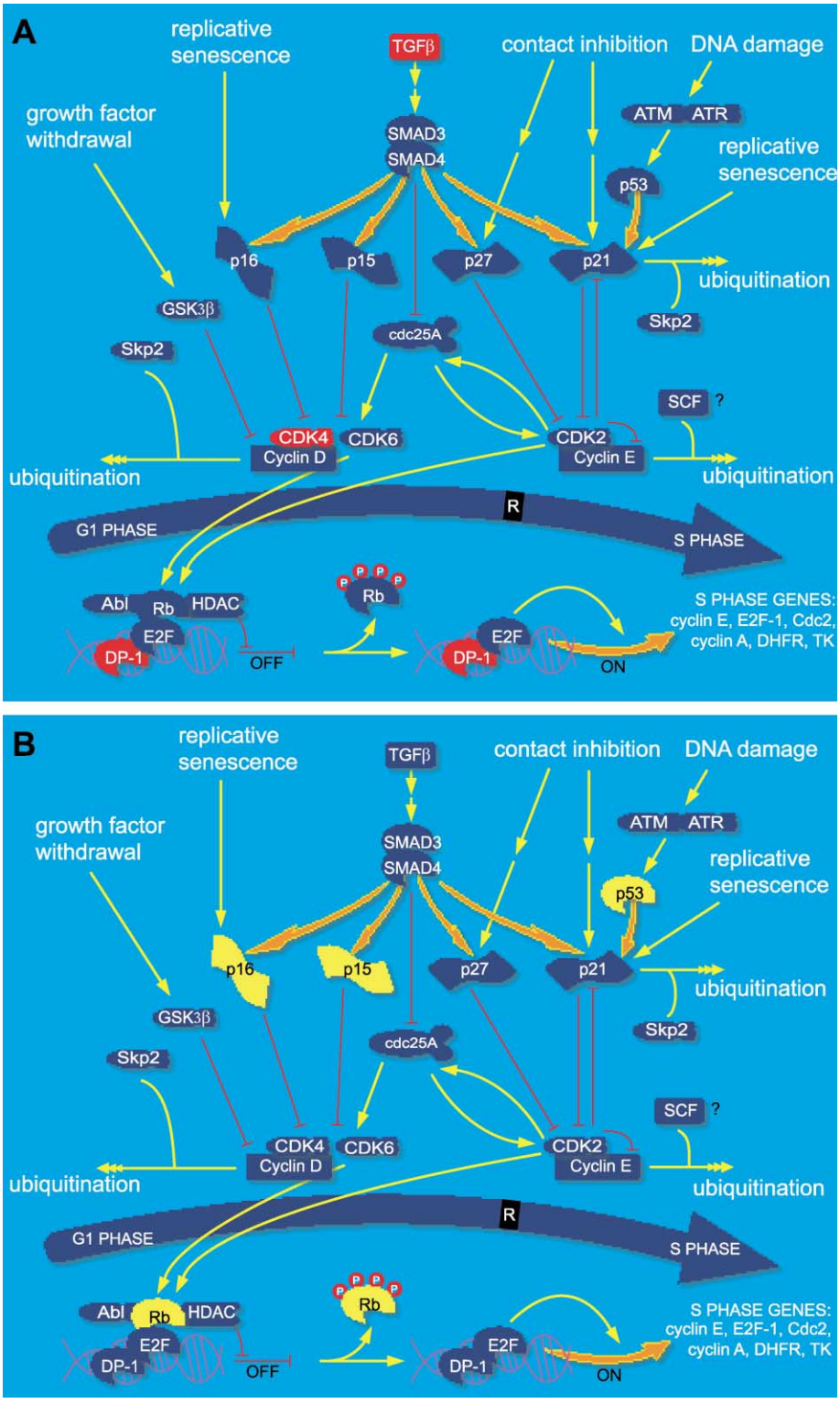

Figure 1. Expression differences and mutation profiles in tissue-specific pathway context

Cell cycle G1/S checkpoint pathway from BioCarta showing (**A**) genes with significant ($p < 0.05$) overexpression (red) in glioblastoma versus normal brain from CGAP SAGE data and (**B**) genes described in the literature to show alterations in glioblastoma (yellow).

development, and validation of a targeted therapy. The molecular profiles section presents the molecular description of tumors. The molecular targets section organizes and presents molecular information by functional ontologies and molecular pathways. The agents section provides access to information on molecularly targeted therapeutics. Finally, the trials section catalogs and describes NCI-supported clinical trials utilizing molecularly targeted agents.

Orthogonal to these entry points is the ability to determine the cancer context through which information is obtained and integrated. Information retrieval spans the continuum from all cancers of all types to specific histological subtypes of cancer of a given tissue type. This variability enables two different lines of inquiry. When queried from a histologically specific perspective, it is possible to discover molecular heterogeneity within a cancer type. Alternatively, a query that aggregates over multiple cancer types facilitates the identification of commonalties among their molecular architectures.

In silico hypothesis generation through CMAP

Each of the individual sections provides opportunities to discover patterns, view information in biological contexts, and/or integrate across the various cancer research disciplines. While obviously not a replacement for actual experimentation, the infrastructure facilitates hypothesis generation from integration of compiled crossdisciplinary data resources. This potential is perhaps best demonstrated by a data-mining example.

To identify potential molecular targets in glioblastoma using the CMAP infrastructure, one can start by looking at molecular pathways which have genes expressed either in the brain or in glioblastoma based on SAGE data. Among these is the G1/S checkpoint pathway, one commonly felt to be critical to abnormal cell growth. Examination of this pathway ideogram allows one to identify genes that show significant differences in gene expression (Figure 1A). The ideograms show four genes (ATR, p15, p16, CDK4, TFDP1, and TGFβ1) whose expression is increased in glioblastoma relative to normal brain. Interestingly, with the exception of TGFβ1, these genes are either the same as (p15, p16) or adjacent to (p53, RB1) genes within the pathway that have been described in the literature to show mutations in glioblastomas (Figure 1B). The specific expression information for all 26 genes within the pathway can be examined (Figure 2). TGFβ1, TFDP1, and CDK4 show highly significant overexpres-

Target	SAGE Data		
	Normal	Cancer	P<
TGFB1			8.92e-09
CDKN2A			4.49e-02
CDC25A			1.00e+00
TP53			6.07e-01
CCNE1			1.00e+00
CDK2			3.79e-01
SKP2			5.70e-01
CDK6			4.49e-01
CDKN2B			1.03e-02
RB1			1.00e+00
MADH4			5.20e-01
ATR			4.84e-02
GSK3B			7.89e-01
TFDP1			1.30e-07
CCNA1			5.28e-01
DHFR			8.49e-01
HDAC1			7.58e-02
CDK4			6.99e-09
E2F1			4.49e-01
ABL1			5.36e-01
CDKN1A			9.39e-02
ATM			6.51e-01
MADH3			1.00e+00
CDKN1B			2.35e-01
CDC2			9.82e-05
KITLG			1.00e+00

Figure 2. Virtual RNA dot blot presentation of SAGE expression data of genes in the G1/S checkpoint pathway

Tissue	EST Data		SAGE Data	
	Normal	Cancer	Normal	Cancer
ALL TISSUES				
adipose			--	--
adrenal cortex	--		--	--
adrenal medulla	--		--	--
bone			--	--
bone marrow			--	--
brain				
cerebellum		--		
cerebrum		--		--
cervix			--	--
colon				
ear		--	--	--
endocrine			--	--
esophagus			--	--
eye			--	--
gastrointestinal tract			--	--
genitourinary			--	--
germ cell	--		--	--
head and neck			--	--
heart		--		--
kidney				--
liver				--
lung				--
lymph node			--	--
mammary gland				
muscle			--	--
nervous			--	--
ovary				
pancreas				
pancreatic islet		--	--	--

Figure 3. Virtual northern blot of EST and SAGE expression data for CDK4 from numerous human tissue samples

sion in tumor when compared to normal ($p < 0.00001$). Among these, only CDK4 has been annotated to have an agent for which it is a target, flavopirodiol. Flavopirodiol is observed to be the subject of 10 relevant trials, although none are specifically targeted at glioblastoma.

To further evaluate the efficacy of the CDK4 target and flavopirodiol as an agent, the CMAP infrastructure can be used to assemble additional information. First, a virtual Northern blot can be generated that shows the expression of CDK4 in other tissues. CDK4 does not appear to have high expression levels in any essential organ tissues, suggesting that potential toxicity might be low (Figure 3). Inspection of CDK4 expression in individual SAGE libraries indicates that its composite high expression masks variability in expression between libraries (Figure 4). This variability is reflected as well in the expression patterns of CNS cell lines that are part of the NCI 60 screening panel (Figure 4). Examining the results of growth inhibition studies shows a relationship between relative expression of CDK4 and the magnitude of growth inhibition in the CNS lines (Figure 4).

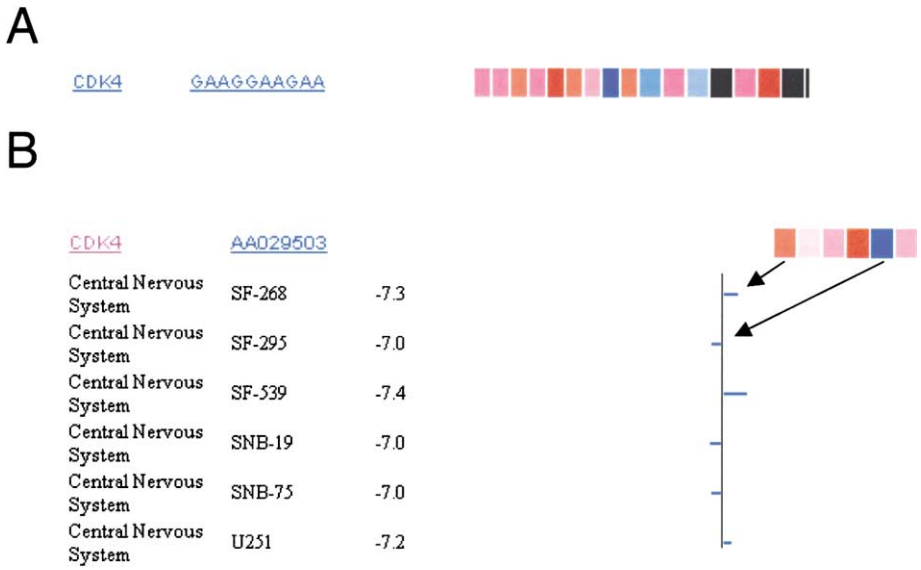

Figure 4. Interrelationship of selected candidate target gene expression and growth inhibition

Relative expression of CDK4 in (**A**) individual SAGE libraries for the displayed tag contributing to the virtual dot blot composite score and (**B**) individual CNS cell lines within the NCI60 screening panel. The lower panel shows the relative growth inhibition (GI50) for the cell lines with respect to mean growth inhibition across all lines, with bars to the left showing less than average inhibition and bars to the right showing greater than average inhibition. Arrows indicate the correspondence between relative CDK4 expression and growth inhibition.

These results suggest that assays assessing expression of CDK4 may be an important requisite component of any successful trial.

CMAP into the future

The above example suggests that potentially useful information can already be drawn from CMAP infrastructure. Insight is limited, though, to the information resources upon which CMAP draws. Other NCICB efforts will broaden this base substantially. More specifically, the NCICB has built tools to collect animal model data being generated within NCI's Mouse Models of Human Cancer Consortium (MMHCC). Cancer researchers can use this infrastructure to add their own models to the collective. The MMHCC will coordinate the curation of the data submitted. This will permit users to select among the degree of vetting associated with data they choose to explore. Similar infrastructure has been developed to allow cancer researchers to submit gene expression data. Plans are under way to introduce chemical genomics and proteomics data.

Informatics can only partially accomplish the task of data integration. It is clear that some categories of data may not be generated in the course of standard cancer investigation. More critically, observations of different types must be made against common biologic specimens in order to meaningfully explore their integration. To bridge these gaps, CMAP will generate new datasets. More specifically, CMAP will generate gene expression information using cDNA microarrays to measure RNA levels from tumors. On this same collection of tumor samples, key genes will be tested for the presence of somatic mutations and constitutional genetic variation. Additional data resources will be generated as the need for bridges is identified.

The CMAP effort is still in its early stages. Nevertheless, interesting information already can be gleaned. Gaps in our knowledge as well as connections become obvious. Not unexpectedly, errors, omissions, and contradictions in the data emerge when viewed in this framework. Most of these will turn out to be experimental errors. However, these contradictions may challenge old assumptions and suggest new experiments that could alter existing paradigms.

It has been suggested that the primary challenge facing cancer research is the weaving of the disparate threads of cancer biology. Application of information technology to integrated data sources, such as that being undertaken by CMAP, may be an important first step in constructing the necessary loom.

Further reading

Klausner, R.D. (2002). The fabric of cancer biology–Weaving together the strands. Cancer Cell *1*, 3–10.

Strausberg, R.L., Greenhut, S.F., Grouse, L.H., Schaefer, C.F., and Buetow, K.H. (2001). In silico analysis of cancer through the Cancer Genome Anatomy Project. Trends Cell Biol. *11*, S66–S71.

Weber, B.L. (2002). Cancer genomics. Cancer Cell *1*, 37–47.