The Cancer Genome Anatomy Project: EST Sequencing and the Genetics of Cancer Progression [1]

David B. Krizman * †, *Lukas Wagner*‡, *Alex Lash*‡, *Robert L. Strausberg** and *Michael R. Emmert-Buck** §

*Cancer Genome Anatomy Project, Office of the Director, †Laboratory of Pathology, §Pathogenetics Unit, Laboratory of Pathology, National Cancer Institute; ‡National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD

Abstract

As the process of tumor progression proceeds from the normal cellular state to a preneoplastic condition and finally to the fully invasive form, the molecular characteristics of the cell change as well. These characteristics can be considered a molecular fingerprint of the cell at each stage of progression and, analogous to fingerprinting a criminal, can be used as markers of the progression process. Based on this premise, the Cancer Genome Anatomy Project was initiated with the broad goal of determining the comprehensive molecular characterization of normal, premalignant, and malignant tumor cells, thus making a reality the identification of all major cellular mechanisms leading to tumor initiation and progression ([Strausberg, R.L., Dahl, C.A., and Klausner, R.D. (1997). "New opportunities for uncovering the molecular basis of cancer." Nat. Genet., 16: 415–516.], www.ncbi.nlm.nih.gov/ncicgap/). The expectation of determining the genetic fingerprints of cancer progression will allow for 1) correlation of disease progression with therapeutic outcome; 2) improved evaluation of disease treatment; 3) stimulation of novel approaches to prevention, detection, and therapy; and 4) enhanced diagnostic tools for clinical applications. Whereas acquiring the comprehensive molecular analysis of cancer progression may take years, results from initial, short-term goals are currently being realized and are proving very fruitful.

Initial Goals of the Cancer Genome Anatomy Project

The first of the initial CGAP goals is to establish a Tumor Gene Index (TGI) to serve as a catalogue of all genes expressed in the cancer progression process, with special reference to tumor type and stage of progression (http://www.ncbi.nlm.nih.gov/ncicgap/EST/cgaplb.cgi; [1]). Establishment of the TGI is being done by constructing cDNA libraries from pathological tissue followed by high-throughput library sequencing. In general, this approach was used successfully to develop the Expressed Sequence Tag database (dbEST) by the National Center for Biotechnology Information (NCBI) and the I.M.A.G.E. consortium ([2,3]; http://www.ncbi.nlm.nih.gov/dbEST/index.html). The TGI uses the existing infrastructure of dbEST and amounts to a catalogue of all the genes that are expressed across the entire spectrum of cancer progression, with special attention to prostate, breast, ovarian, lung, and colon cancers. A secondary goal in developing the TGI was to identify the remaining members of the unique human gene set, represented by UniGene set of genes ([4,5]; http://www.ncbi.nlm.nih.gov/UniGene/index.html). In addition, as EST mapping proceeds and every UniGene cluster is eventually placed on the human transcript map, region-specific catalogues of genes will exist that can be matched to genomic "hotspots" correlating with specific cancer types and stages ([6–10]; http://www.ncbi.nlm.nih.gov/genemap/, http://www.ncbi.nlm.nih. gov/CCAP/NG/).

To develop the TGI and build on existing databases in parallel, two types of CGAP-specific cDNA libraries are currently in use. Standard bulk tissue cDNA libraries from RNA derived from large tumor tissues function primarily to create a general picture of those genes expressed in the tumor process in addition to driving the process of gene discovery to aid the UniGene effort. More than 80 bulk tissue cDNA libraries (normalized as well as non-normalized) from a wide range of tumor types and histologies have been sequenced, and to date more than 340,000 ESTs have been deposited in the TGI. In addition, more than 11,000 novel genes have been discovered thus far to supplement the UniGene set. Although these results have proven extremely useful, a serious drawback to the sequencing of bulk tissue cDNA libraries is the lack of gene expression information in the context of tumor biology. This is primarily due to cellular heterogeneity found in bulk tissue.

Histological examination reveals that the prostate is a complex organ comprising of multiple cell types. Only 10% of the cells are epithelial in origin, whereas the remaining 90% of the organ comprises of inflammatory, fibroblastic, endothelial, and nervous cells. Yet it is the epithelium that gives rise to life-threatening prostate cancer [11,12]. Armed with this information, it is quite easy to understand why any attempts to determine a prostate epithelial–specific gene

profile would fail by sequencing a bulk tissue cDNA library from a normal prostate gland. Furthermore, bulk tumor tissue will undoubtedly contain inflammatory, structural, and endothelial cells regardless of the percent of tumor cells in the tissue as determined histologically. It was this realization that led to the development of laser capture microdissection (LCM) [13,14].

LCM is a process by which one is able to procure selected groups of cells, or even individual cells, from a heterogeneous population of cells in standard pathology preparations. The second type of cDNA library used in CGAP is constructed from RNA obtained by LCM [15,16]. These libraries make it possible for the first time, to perform large-scale, *in vivo* gene expression profiling from a specific cell type. To begin addressing the issue of gene expression and profiling in the process of prostate cancer progression, a total of 15 cDNA libraries have been constructed from normal epithelium, prostatic intraepithelial neoplasia (PIN) lesions, invasive tumor cells, and metastatic prostate lesions. Many of these libraries were constructed from cells dissected from the same patient and pathology preparation. More than 30,000 clones have been sequenced from these libraries, representing 5186 UniGene clusters. Not only are these sequences useful for prostate tissue–specific and prostate cancer stage–specific expression analysis, they are useful for gene discovery, as evidenced by the establishment of greater than 400 UniGene clusters. Thus these libraries possess the potential to discover weakly expressed, tissue-specific, and cell-specific transcripts not easily found in bulk tissue libraries.

CGAP bioinformatics

With the recent surge in genetic information available to the cancer researcher, it is apparent that useful bioinformatics packages need to be developed to address these issues. The CGAP Website has been actively pursuing this endeavor in trying to deliver tools that would allow the individual investigator to tease out interesting gene expression data from all of the cDNA libraries that currently exist in the TGI. One such function is Digital Differential Display (DDD). This utility uses the Fisher exact test [17] to compare one library to another, a pool of libraries to a single library, or a pool of libraries against another pool. In addition, all cDNA libraries that exist in dbEST can be used, not just those from CGAP. This allows for flexibility in designing an experiment in silica and many questions can be asked using this function.

For example, one may obtain a list of tissue-specific genes for the prostate by constructing several pools for the DDD to analyze (http://www.ncbi.nlm.nih. gov/ncicgap/ddd.html). This working example is entitled "Compare Stages of Prostate Cancer" and can be viewed on the URL listed above. Detailed instructions for using DDD can also be found at this site. To find tissue-specific genes, a control pool should consist of libraries specific to several tissues different from each other and from the tissue of interest, and the pools of interest should contain libraries which are as narrowly focussed as possible. In addition, a control pool should consist of several diverse libraries with many sequences. Choosing libraries too similar to each other for the control pool (for instance, several different libraries con-

Digital Differential Display (DDD)

DDD is a computational method for comparing sequence-based gene representation profiles among individual cDNA libraries or pools of libraries. The following table describes the pools that have been defined so far.

	Pool	Name	Lib ID(s)	ESTs
Edit...	A	Normal	281	3271
Edit...	B	Precancerous	282	3391
Edit...	C	Malignant	283	2819
Edit...	D	Control	162, 235	58697
New...				

Start Over

Figure 1. *DDD page showing the choice of pools used for this analysis.*

structed from brain tissue) would simply identify genes not expressed in brain tissue and not a superset of genes specific to prostate tissue. A similar difficulty would arise were the control pool to contain libraries with few ESTs. One would obtain genes not expressed in the small control pool, of which the tissue-specific genes would be a small fraction. Thus, we choose the libraries "Normalized infant brain 1NIB" (45472 ESTs) and "Soares fetal liver spleen 1NFLS S1" (29545 ESTs) for our control pool (Figure 1).

Because the pools are easily edited, one may test to ensure that the results are independent of the choice of control pool by modifying the control pool at a later stage in the analysis.

Next, one chooses libraries for three prostate-specific pools; this is preferable to grouping the diverse libraries in a single pool because differences between the pools indicate the extent to which any gene is specific to normal, neoplastic, or preneoplastic tissue. We choose library "NCI_CGAP_Pr1, Microdissected, normal prostate epithe-

Statistically Significant Differences

A Normal	B Precan..	C Malign..	D Control	Gene index	Gene description
0.02018 ● A > D	0.03804 ● B > D	0.01809 ● C > D	0.00002 D < A D < B D < C	Hs.1548	Prostate specific antigen (APS)
0.01620 ● A > B A > D	0.00295 ○ B < A B > D B < C	0.01738 ● C > B C > D	0.00000 D < A D < B D < C	Hs.73487	Beta-microseminoprotein (prostate secreted) (MSMB)
0.01039 ● A > D	0.00944 ● B > D	0.01242 ● C > D	0.00000 D < A D < B D < C	Hs.1852	Prostatic acid phosphatase (ACPP)
0.00917 ● A > D	0.00944 ● B > D	0.01313 ● C > D	0.00002 D < A D < B D < C	Hs.108124	60S RIBOSOMAL PROTEIN L41
0.00887 ● A > D	0.01032 ● B > D	0.00816 ● C > D	0.00034 D < A D < B D < C	Hs.146398	Translation elongation factor 1-alpha-1 (EEF1A1)
0.00397 ○ A > D	0.00560 ● B > D	0.01029 ● C > D	0.00005 D < A D < B D < C	Hs.5174	Ribosomal protein S17 (RPS17)
0.00550 ● A > D	0.00973 ● B > D	0.00355 ○ C > D	0.00000 D < A D < B D < C	Hs.73098	Kallikrein 1 (renal/pancreas/salivary) {alternative products} (KLK1)

Figure 2. *Result page from the DDD analysis indicating 7 statistically significant prostate-specific transcripts. Many more transcripts were found, and those can be found on the DDD website as described in the text.*

lium" to exemplify normal tissue, "NCI_CGAP_Pr2, Microdissected, low grade prostatic intraepithelial neoplasia" to exemplify preneoplastic tissue, and "NCI_CGAP_Pr3, Microdissected, invasive prostate tumor" to exemplify neoplasia.

Because one of the pools is a diverse control, all prostate-specific genes expressed in one of the tissue-specific pools are listed; furthermore, differences between the pools are also listed. Note that comparing only prostate-specific normal and cancerous libraries would produce very few significant differences. Although the use of these as separate pools compared with a control pool displays differences that are not statistically significant, these genes would nonetheless amount to candidate genes involved in prostate cancer progression and may prove very useful to the cancer biologist as potential leads to experimental follow-up. Statistical significance is at the $P < = .05$ level for the Fisher exact test [17]. To see whether the differences found are due to idiosyncrasies of the libraries chosen, we can expand the pools by adding the

---

NCI CGAP   NCBI

xProfiler

This online tool displays the expression profiles of CGAP EST libraries.

Query: `OR[Lib.281,Lib.282,Lib.283]`   Submit

Query summary:

Pool	Query	Library Name	Sequences	UniGen
#A	Lib.281	NCI_CGAP_Pr1	5689	1360
#B	Lib.282	NCI_CGAP_Pr2	5688	1561
#C	Lib.283	NCI_CGAP_Pr3	5209	1250
#D	OR[Lib.281,Lib.282,Lib.283]			3362

Current query results:

OR[Lib.281,Lib.282,Lib.283] *(3362 UniGene clusters)*

#	UniGene id	Gene description	Lib.281 (n=3289)	Lib.282 (n=3405)	Lib.283 (n=2825)
1	Hs.1548	Prostate specific antigen	2% (66)	3% (129)	1% (51)
2	Hs.73487	Beta-microseminoprotein (prostate secreted)	1% (53)	<1% (10)	1% (49)
3	Hs.1852	Prostatic acid phosphatase	1% (34)	<1% (32)	1% (35)
4	Hs.108124	60S RIBOSOMAL PROTEIN L41	<1% (30)	<1% (32)	1% (37)
5	Hs.146398	Translation elongation factor 1-alpha-1	<1% (29)	1% (35)	<1% (23
6	Hs.77039	Ribosomal protein S3A	<1% (27)	<1% (12)	<1% (17
7	Hs.76317	Ribosomal protein L31	<1% (27)	<1% (6)	<1% (24
8	Hs.75362	H.sapiens DAP-3 mRNA	<1% (23)	<1% (12)	<1% (22
9	Hs.146411	Ribosomal protein S24	<1% (19)	<1% (11)	<1% (19
10	Hs.73098	Kallikrein 1 (renal/pancreas/salivary) {alternative products}	<1% (18)	<1% (33)	<1% (10
11	Hs.119252	TRANSLATIONALLY CONTROLLED TUMOR PROTEIN	<1% (18)	<1% (11)	<1% (14

Figure 3. *Display of a gene expression profile analysis from the following microdissected cDNA libraries: Lib.281 (NCI_CGAP_Pr1, microdissected normal epithelium), Lib.282 (NCI_CGAP_Pr2, microdissected preneoplastic lesion), and Lib.283 (NCI_CGAP_Pr3, microdissected invasive tumor).*

following: for normal libraries, "NCI_CGAP_Pr9 Microdissected, normal prostate epithelium," and "NCI_CGAP_Pr25, Cell line, normal prostate epithelial cell line"; for precancerous libraries, "NCI_CGAP_Pr4, Microdissected, high grade prostatic intraepithelial neoplasia" and "NCI_CGAP_Pr4.1, Microdissected, high grade prostatic intraepithelial neoplasia high grade"; and for cancerous libraries, "NCI-CGAP_Pr24, Cell line, invasive prostate tumor cell line (HPV immortalized)," "NCI_CGAP_Pr10, Microdissected, invasive prostate tumor," and "NCI_CGAP_Pr8, Microdissected, invasive prostate tumor."

There are two genes with significant differences between states in the small and large pools: prostate specific antigen (PSA) and beta-microseminoprotein (prostate secreted) (MSMB) (Figure 2). Although not statistically significant according to the Fisher test, we note that kallikrein also has different expression levels in normal and precancerous tissues, as do several ribosomal proteins. MSMB and kallikrein have been implicated in prostate cancer [18–21]. This suggests that UniGene clusters with similar expression profiles would be potential candidates for the molecular fingerprinting of the stages of prostate cancer. Note that UniGene cluster identifiers, although superficially very convenient as referents, are not guaranteed to be stable for archival purposes. This is because clusters may split or merge together with the addition of new sequences. Thus, it is safest to store the list of accession numbers in a cluster of interest.

The Fisher exact test, which is used to assess whether the difference in expression levels, is known to be conservative. It is therefore useful to have an independent tool to examine differences in expression level. The gene expression comparison utility (http://www.ncbi.nlm.nih.gov/ncicgap/EST/cgapqr.cgi) (Figure 3) does not attempt to assess statistical significance but can be used to identify which libraries have contributed sequences to a gene of interest. One difficulty particularly relevant in seeking novel ESTs is the observation that the Fisher exact test will not find a significant difference in expression levels for small clusters. The exact definition of small clusters depends on the total number of sequences in the pools being compared but, for instance, clusters of size 1 are never statistically significant. Thus, the Fisher exact test and the DDD interface to the test will tend to identify larger clusters and thus already characterized genes.

Two recent studies have demonstrated the usefulness of these prostate cDNA libraries [22,23]. A combination of computer-based analysis and laboratory analysis identified a number of genes from the prostate libraries within the TGI that have shown patterns of prostate-specific expression. The investigators suggest the procedure they used can be easily applied to the discovery of genes expressed in others organs or tumors.

Conclusions and Future CGAP Goals

The DDD and gene expression comparison utilities are tools that currently exist to analyze CGAP-generated data. The CGAP website has historically been dynamic and is in continuous flux according to the data present in the TGI; thus all utilities are subject to continual improvements and upgrades. The example outlined in this article focused on prostate cancer. The immediate CGAP goal is to complete construction and sequencing of analogous cDNA libraries from microdissected cells representing all stages of ovarian, lung, colon, and breast cancers. Thus, analysis of the gene expression profiles of these first 5 cancers will undoubtedly render unprecedented bioinformation to the cancer community. More tumors will likely be added to this list once these 5 are completed.

A future goal for the analysis of gene expression in cancer progression is the development and use of serial analysis of gene expression (SAGE) cDNA libraries from cancer tissue [24]. A number of these libraries have recently been constructed and sequenced by CGAP, and utilities to analyze these data are starting to emerge (http://www.ncbi.nlm.nih.gov/SAGE/). Due to the larger amount of data that can be obtained by sequencing SAGE libraries, greater statistical significance to computer-generated gene expression analysis can be ascribed to these analyses. However, because these libraries were generated from bulk tumor tissue and not microdissected cells, direct comparison of tumor to preneoplastic or normal cellular states cannot be made. Thus, an ideal gene expression analysis of cancer progression might be the application of SAGE technology to microdissected cells.

In addition to expanding and improving the usefulness of gene expression profiles generated from sequencing cDNA libraries, CGAP has recently committed to the study of cancer progression at the genomic level by establishment of the Genetic Annotation Initiative (GAI) and the Cancer Chromosome Aberration Project (cCAP). The GAI goal is to discover and catalogue single nucleotide polymorphisms in cDNA sequence (SNPs) that correlate with cancer initiation and progression, whereas the goal of cCAP is to develop a set of tools that will allow for the expedient definition and detailed characterization of chromosomal alterations associated with cancer initiation and progression.

Expansion to model organisms is beginning to take shape within CGAP as well. Establishment of the mouse TGI will take place in the near future that will mirror the current human TGI in that both bulk tumor tissue and microdissected cells will be used to generate cDNA libraries for high-throughput sequencing. Like the human TGI, the 5 cancers of focus for the mouse are prostate, breast, lung, colon, and ovarian cancers.

In conclusion, the CGAP encompasses an entire approach to understanding cancer at the molecular level. Even in its infancy it shows great promise for uncovering important gene expression changes involved in cancer initiation and progression. An example for discovering such changes has been outlined here. In the near future one could envision that as the TGI grows linearly, possibilities for bioinformatics could expand exponentially. With addition of the new CGAP initiatives discussed here, the National Cancer Institute optimistically looks forward to uncovering the molecular changes that lead to cancer initiation and progression.

References

[1] Strausberg RL, Dahl CA, and Klausner RD (1997). New opportunities for uncovering the molecular basis of cancer. *Nat Genet* **16**, 415–516.

[2] Boguski MS, Lowe TM, and Tolstoshev CM (1993). dbEST-database for "expressed sequence tags." *Nat Genet* **4**, 332–333.

[3] Lennon G, Auffray C, Polymeropoulos M, and Soares MB (1996). The I.M.A.G.E. Consortium: An integrated molecular analysis of genomes and their expression. *Genomics* **33**, 151–152.

[4] Boguski M, and Schuler G (1995). ESTablishing a human transcript map. *Nat Gen* **10**, 369–371.

[5] Schuler G (1997). Pieces of the puzzle: Expressed sequence tags and the catalog of human genes. *J Mol Med* **75**, 694–698.

[6] Hudson TJ, Stein LD, Gerety SS, Ma J, Castle AB, Silva J, Slonim DK, Baptista R, Kruglyak L, Xu SH, et al (1995). An STS-based map of the human genome. *Science* **270**, 1945–1954.

[7] Gyapay G, Schmitt K, Fizames C, Jones H, Vega-Czarny N, Spillett D, Muselet D, Prud'Homme JF, Dib C, Auffray C, Morissette J, Weissenbach J, and Goodfellow PN (1996). A radiation hybrid map of the human genome. *Hum Mol Genet* **5**, 339–346.

[8] Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Hudson TJ, et al (1996). A gene map of the human genome. *Science* **274**, 540–546.

[9] Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tomé P, Hui L, Matise TC, McKusick KB, Beckmann JS, Bentolila S, Bihoreau M-T, Birren BB, et al (1998). A physical map of 30,000 human genes. *Science* **282**, 744–746.

[10] Mitelman F, Mertens F, and Johansson B (1997). A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nat Genet* **15**, 417–474.

[11] Murphy GP, Busch C, Abrahamsson PA, Epstein JI, McNeal JE, Miller GJ, Mostofi FK, Nagle RB, Nordling S, Parkinson C, et al (1994). Histopathology of localized prostate cancer. Consensus Conference on Diagnosis and Prognostic Parameters in Localized Prostate Cancer. *Scand J Urol Nephrol* **162**, [Suppl] 7–42.

[12] Epstein JI (1996). Pathologic evaluation of prostatic carcinoma: critical information for the oncologist. *Oncology* **10**, 527–534.

[13] Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, and Liotta LA (1996). Laser Capture Microdissection. *Science* **274**, 998–1001.

[14] Bonner RF, Emmert-Buck M, Cole K, Pohida T, Chuaqui R, Goldstein S, and Liotta LA (1997). Laser Capture Microdissection: Molecular analysis of tissue. *Science* **278**, 1481–1483.

[15] Krizman DB, Chuaqui RF, Meltzer PS, Trent JM, Duray PH, Linehan WM, Liotta LA, and Emmert-Buck MR (1996). Construction of a representative cDNA library from prostatic intraepithelial neoplasia (PIN). *Cancer Res* **56**, 5380–5383.

[16] Peterson LA, Brown M, Carlisle AJ, Kohn E, Liotta LA, Emmert-Buck MR, and Krizman DB (1998). An improved method for microdissected cDNA libraries using papillary serous ovarian carcinoma cells. *Cancer Res* **58**, 5326–5328.

[17] Seigel S (1956) *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York.

[18] Heeb MJ, and Espana F (1998). Alpha2-macroglobulin and C1-inactivator are plasma inhibitors of human glandular kallikrein. *Blood Cells Mol Dis* **24**, 412–419.

[19] Saedi MS, Hill TM, Kuus-Reichel K, Kumar A, Payne J, Mikolajczyk SD, Wolfert RL, and Rittenhouse HG (1998). The precursor form of the human kallikrein 2, a kallikrein homologous to prostate-specific antigen, is present in human sera and is increased in prostate cancer and benign prostatic hyperplasia. *Clin Chem* **44**, 2115–2119.

[20] Tsurusaki T, Koji T, Sakai H, Kanetake H, Nakane PK, and Saito Y (1998). Cellular expression of beta-microseminoprotein (beta-MSP) mRNA and its protein in untreated prostate cancer. *Prostate* **35**, 109–116.

[21] Hyakutake H, Sakai H, Yogi Y, Tsuda R, Minami Y, Yushita Y, Kanetake H, Nakazono I, and Saito Y (1993). Beta-microseminoprotein immuno-reactivity as a new prognostic indicator of prostatic carcinoma. *Prostate* **22**, 347–355.

[22] Vasmatzis G, Essand M, Brinkmann U, Lee B, and Pastan I (1998). Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proc Natl Acad Sci USA* **95**, 300–304.

[23] Brinkmann U, Vasmatzis G, Lee B, Yerushalmi N, Essand M, and Pastan I (1998). PAGE-1, an X chromosome-linked GAGE-like gene that is expressed in normal and neoplastic prostate, testis, and uterus. *Proc Natl Acad Sci USA* **95**, 10757–10762.

[24] Velculescu VE, Zhang L, Vogelstein B, and Kinzler KW (1995). Serial analysis of gene expression. *Science* **270**, 484–487.