# 2

# Bioinformatics Approaches to Cancer Gene Discovery

**Ramaswamy Narayanan**

## Summary

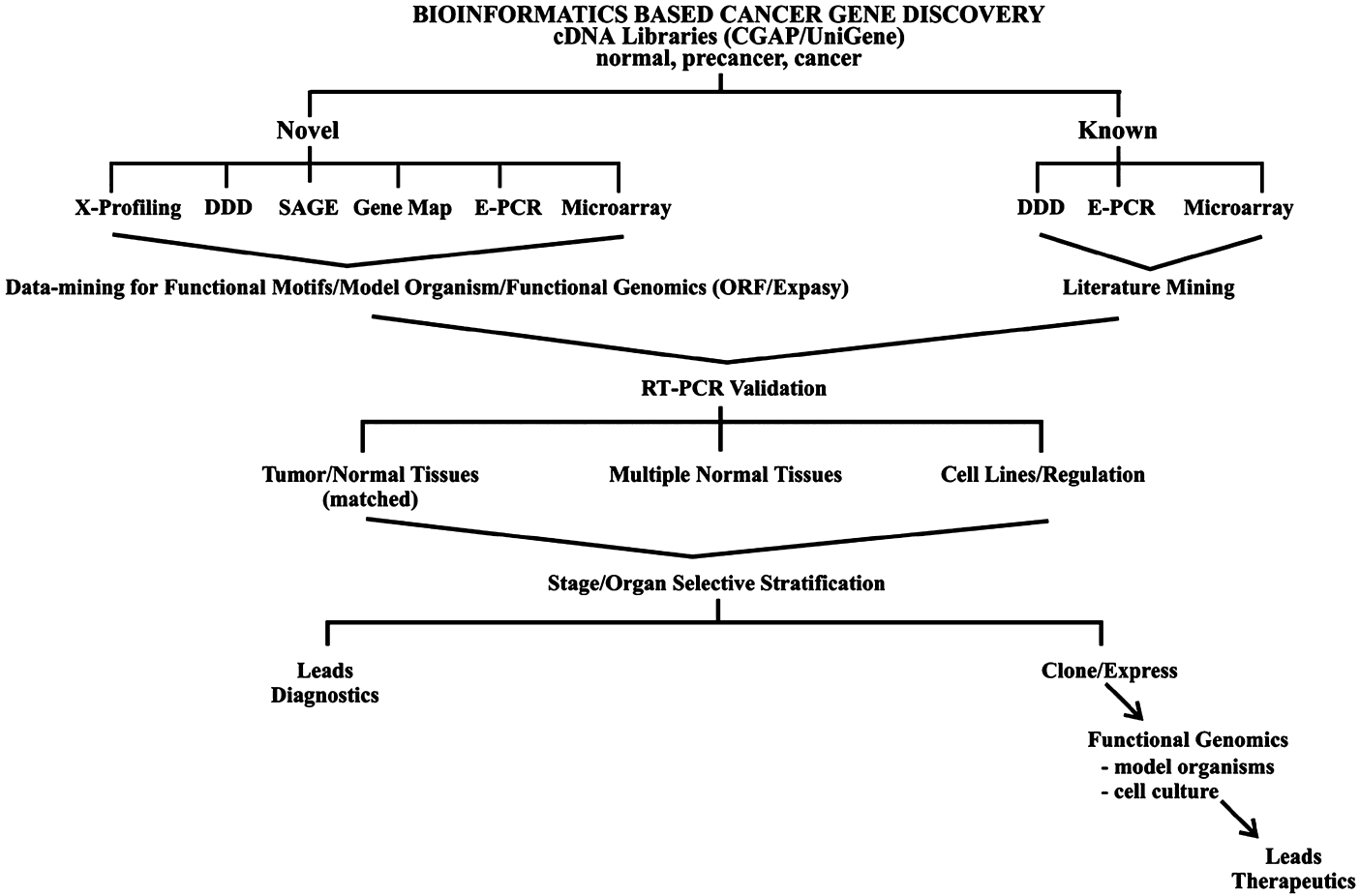
The Cancer Gene Anatomy Project (CGAP) database of the National Cancer Institute has thousands of known and novel expressed sequence tags (ESTs). These ESTs, derived from diverse normal and tumor cDNA libraries, offer an attractive starting point for cancer gene discovery. Data-mining the CGAP database led to the identification of ESTs that were predicted to be specific to select solid tumors. Two genes from these efforts were taken to proof of concept for diagnostic and therapeutics indications of cancer. Microarray technology was used in conjunction with bioinformatics to understand the mechanism of one of the targets discovered. These efforts provide an example of gene discovery by using bioinformatics approaches. The strengths and weaknesses of this approach are discussed in this review.

**Key Words:** Antisense; digital differential display; Down syndrome; genomics; solid tumors.

## 1. Introduction

The recent completion of the human genome sequencing efforts offers a new dimension in gene discovery approaches ***(1–4)***. From these vast numbers of new genes, new diagnostic and therapeutic targets for diseases such as cancer are predicted to emerge ***(5)***. Only a subset of genes is expressed in a given cell, and the level of expression governs function. High-throughput gene expression technology is becoming a possibility for analyzing expression of a large number of sequences in diseased and normal tissues with the use of microarrays and gene chips ***(6–9)***. A parallel way to initiate a search for genes relevant to cancer diagnostics and therapy is to data-mine the sequence database ***(10–14)***. A large number of expressed sequences from diverse organ-, species-, and disease-derived cDNA libraries are being deposited in the form of expressed sequence tags (ESTs) in different databases.

From: *Methods in Molecular Biology, vol. 360, Target Discovery and Validation Reviews and Protocols*
*Volume I, Emerging Strategies for Targets and Biomarker Discovery*
Edited by: M. Sioud © Humana Press Inc., Totowa, NJ


BIOINFORMATICS BASED CANCER GENE DISCOVERY
cDNA Libraries (CGAP/UniGene)
normal, precancer, cancer
Novel
Known
X-Profiling
DDD
SAGE
Gene Map
E-PCR
Microarray
DDD
E-PCR
Microarray
Data-mining for Functional Motifs/Model Organism/Functional Genomics (ORF/Expasy)
Literature Mining
RT-PCR Validation
Tumor/Normal Tissues (matched)
Multiple Normal Tissues
Cell Lines/Regulation
Stage/Organ Selective Stratification
Leads
Diagnostics
Clone/Express
Functional Genomics
- model organisms
- cell culture
Leads
Therapeutics

The Cancer Gene Anatomy Project (CGAP) database (http://cgap.nci.nih.gov/) of the National Cancer Institute (NCI) is an attractive starting point for cancer-specific gene discovery ***(12)***. The Human Tumor Gene Index was initiated by the NCI in 1997 with a primary goal of identifying genes expressed during development of human tumors in five major cancer sites: (1) breast, (2) colon, (3) lung, (4) ovary, and (5) prostate. This database consists of expression information (mRNA) of thousands of known and novel genes in diverse normal and tumor tissues. By monitoring the electronic expression profile of many of these sequences, it is possible to compile a list of genes that are selectively expressed in the cancers. Data-mining tools are becoming available to extract expression information about the ESTs derived from various CGAP libraries ***(10,13–15)***.

Currently, there are 1.5 million ESTs in the CGAP database, of which 73,000 are novel sequences. These sequences also are subclassified into those derived from libraries of normal, precancerous, or cancer tissues. The CGAP database uses UniGene-based gene clustering. UniGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene) is an experimental system for automatically partitioning GenBank sequences into a nonredundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene. An essential weakness of the UniGene is that currently it does not allow for the identification of consensus or contig sequences of a gene. In addition, the UniGene database is continuously updated and hence the ESTs are often removed and reassigned to another UniGene. However, the contig or the consensus (Tentative Human Consensus, TC/THC) information can be obtained from other databases such as TIGR (http://www.tigr.org/). Multiple data-mining tools from the CGAP database can be used to facilitate the gene discovery.

## 2. Gene Discovery Strategy

An overall strategy for cancer gene discovery by using bioinformatics approaches is shown in **Fig. 1**. Survey sequencing of mRNA gene products can provide an indirect means of generating gene expression fingerprints for cancer cells and their normal counterparts. The cancer specificity of an EST can be predicted using multiple data-mining tools from the CGAP database. These tools include X-profiling, digital differential display (DDD), serial analysis of gene expression (SAGE), electronic-PCR, GeneMap, and microarray databases (*see* Chapters 1, 5, 6, Volume 1). For details of these tools, see the CGAP website.

---

Fig. 1. *(Opposite page)* Gene discovery strategy. A proposed approach to cancer gene discovery from the CGAP database is shown. Both novel and known ESTs are identified using multiple data-mining tools from this database. Further validation in the wet laboratory provides a rational for diagnostic and therapeutic target discovery.

Recently, a tool called digital analysis of gene expression has replaced DDD in the CGAP database; however, the DDD tool can still be accessed from the UniGene page. Briefly, these tools allow prediction of ESTs specific to each cancer type by comparing the occurrence of an EST between two pools of cDNA libraries in a statistically significant manner. The approaches to doing follow-up studies on novel or known ESTs may differ. The novel ESTs can be subject to additional data-mining to identify functional motifs by using the protein databases from the ExPASy (http://au.expasy.org/) server. For a novel EST, a hint of its function also is obtained from homologue-related information from a model organism database such as Homologene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db= homologene), whereas the known ESTs can be subjected to literature-mining to develop hints of relevance to specific cancer types. The chosen ESTs (hits) can be validated by reverse transcription (RT)-PCR by using cDNAs from normal and tumor tissues as well as appropriate cell lines for specificity and regulation. The expression specificity of the validated ESTs (leads) provides a rationale for developing a diagnostic application. The drug therapy use can be inferred using numerous techniques, such as antisense, small-interfering RNA, and so on.

## 3. Example of Cancer Gene Discovery

To identify solid tumor-specific genes, the DDD tool of the CGAP database was used. Both novel and known ESTs that are selectively up- or downregulated in six major solid tumor types (breast, colon, lung, ovary, pancreas, and prostate) were identified. DDD takes advantage of the UniGene database by comparing the number of times ESTs from different libraries were assigned to a particular UniGene cluster. Six different solid tumor-derived EST libraries (breast, colon, lung, ovary, pancreas, and prostate) with corresponding normal tissue-derived libraries were chosen for DDD (N = 110). To identify tumor- and organ-specific ESTs, all the other organ- and tumor-derived EST libraries (N = 327) were chosen for comparison with each of the six tumor types. The nature of the libraries (normal, pretumor, or tumor) was authenticated by comparison of the CGAP data with the UniGene database. Occasionally, the description of the EST libraries in CGAP and UniGene database do not show a match. Those few libraries showing discrepancies of definition between the two databases were excluded. The DDD was performed for each organ type individually. DDD was performed using ESTs from tumors (pool A) and corresponding normal organ (pool B) by using the online tool. The output provided a numerical value in each pool denoting the fraction of sequences within the pool that mapped to the UniGene cluster, providing a dot intensity. An example of DDD output for colon tumor-specific gene discovery by DDD is shown in **Fig. 2**.

| A colon tumors | B all others | Gene index | Gene Description | Fold |
|---|---|---|---|---|
| 0.00241<br>A>B | 0.00008<br>B<A | Hs.25640 | claudin 3 (CLDN3) | 30 (up regulated) |
| 0.00006<br>A>B | 0.00283<br>B<A | Hs.75792 | hemoglobin, alpha 1 (HBA1) | 47 (down regulated) |
| 0.00000<br>A>B | 0.00058<br>B<A | Hs.118843 | creatine kinase, muscle (CKM) | – |
| 0.00020<br>A>B | 0.00000<br>B<A | Hs.1545 | caudal type homeo box transcription factor 1 (CDX1) | + |
| 0.00031<br>A>B | 0.00000<br>B<A | Hs.75792 | ESTs | + |

Fig. 2. Example of DDD output. The numerical value in each box is the fraction of ESTs within the pool that mapped to the UniGene cluster (Hs.) shown. The dot is merely a visual aid that reflects the numerical values. If any pool participates in a statistically significant pairwise comparison with another pool, the relationship is indicated. "A>B" indicates a greater amount of ESTs found in colon cancer libraries versus other libraries for a particular gene. If the number of occurrences of an EST in a UniGene is zero, then the EST is predicted to be present only in the colon cancer libraries (shown as +/– fold). If the number is finite, then the ratio is shown as the -fold difference.

Fold differences were calculated by using the ratio of pool A:pool B. Statistically significant hits (Fisher exact test) showing >10-fold differences were compiled, and a preliminary database was created. Novel ESTs were compiled into a separate database. The UniGene database was accessed to establish an electronic expression profile (E-Northern) for each of the hits to facilitate tumor- and organ-selective gene discovery. The cytogenetic map position of the hits also was inferred from the UniGene page. A final database of ESTs that were upregulated, downregulated, and showed absolute differences (+/–) in the six tumor types was created. These hits were functionally classified into major classes of proteins by using gene ontology. Genes belonging to ribosomal proteins, enzymes, receptors, binding proteins, secretory proteins, and cell adhesion molecules were identified to be differentially expressed in these tumor types. A comprehensive database of hits was created, providing additional electronic expression data as well as novel ESTs that were thus identified ***(16)***. This database can be accessed on the World Wide Web at http://www.fau.edu/cmbb/publications/cancergenes.htm.

Colorectal cancer is a commonly diagnosed cancer in both men and women. In 2006, an estimated 106,680 new cases will be diagnosed, and 55,170 deaths from colorectal cancer will occur in the United States alone. About 75% of patients with colorectal cancer have sporadic disease, with no apparent evidence of having inherited the disorder. The remaining 25% of patients have a family history of colorectal cancer that suggests a genetic contribution, common exposures among family members, or a combination (http://www.nci.nih.gov/cancerinfo/pdq/genetics/colorectal). Although tumor suppressor genes such as deleted in colorectal cancer (DCC), adenomatous polyposis coli (APC), mutated in colorectal cancer (MCC), or oncogenes such as k-*ras* offer a promise for diagnosis of colon cancer, additional more specific markers are urgently needed to benefit colon cancer patients. With this in view, we chose two genes from this database predicted to be specific for colon tumors to test the validity of gene discovery by bioinformatics approaches.

## 4. Discovery of Colon Cancer-Specific Secreted Marker

To date a secreted marker for colon cancer diagnosis has not been identified. Hence, an attempt was made to predict and validate a colon cancer-specific EST that might harbor a signal peptide motif. Sixty-four UniGenes were identified to be upregulated in colon tumors by the DDD approach. Twenty-four of these UniGenes were found to be present only in colon tumor-derived cDNA libraries (CGAP, SAGE, and UniGene). One UniGene, Hs. 307047, which harbored a signal peptide motif, was chosen for further analysis. This UniGene currently has seven different ESTs as a part of the Unigene cluster. The SAGE tag (ACAGTAATGA) identified for this UniGene was also in a colon and gastric tumor-derived library. The TIGR Human Gene Index had a tentative Human Consensus (THC342146) comprising all of the seven different ESTs shown above with the longest EST being AA524300. The EST AA524300 was screened against the National Center for Biotechnology Information database for alu, vector, and bacterial contamination and was found to have no matches. There were significant matches against mouse and rat ESTs, all of them from colon-derived libraries. An *in situ* BLAST library-specific search at TigemNet (http://www.tigem.it) revealed that this EST was not present in any normal colon library. In addition, no match was detected against the Bodymap (http://bodymap.ims.u-tokyo.ac.jp) normal library collections. Comparison against the GeneMap (http://www.ncbi.nlm.nih.gov/genemap) suggested that the EST is located on chromosome 3q13.1. The translated sequence for this EST, compared against the Prosite signature database of ExPASy (http://www.expasy.ch/prosite) showed that this EST encodes a putative signal peptide, prokaryotic lipid binding sites, a prenyl group binding site and membrane glycosylphosphatidylinositol anchor sites. The prediction of a signal peptide sequence and the colon cancer

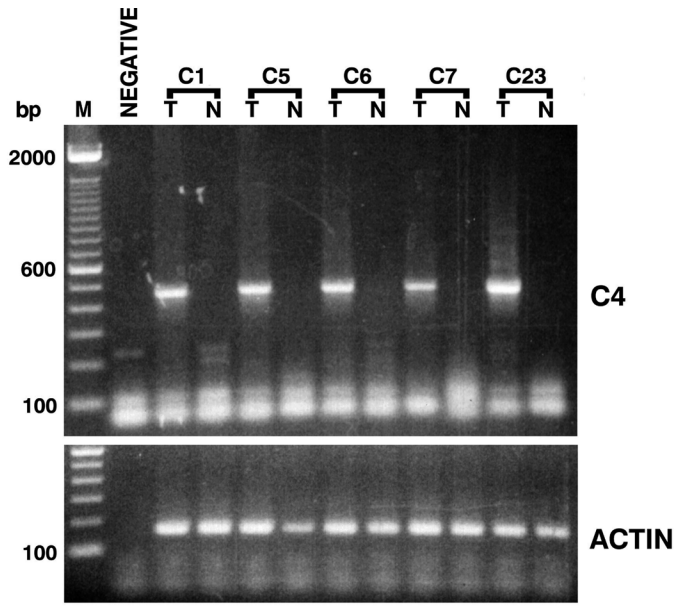



Fig. 3. Discovery of colon cancer-specific secreted marker. The cDNAs from matched sets of tumor (T) and normal (N) colon from five different patients were analyzed by RT-PCR for and for CCRG and actin. M, 100-bp ladder; negative, template minus control.

specificity allowed us to test a rationale of secreted marker discovery for colon cancer diagnosis. Preliminary RT-PCR analysis of a matched set of tumor and normal colon-derived cDNAs by using an exon-specific PCR primer pair detected a product in the tumor but not in the normal colon cDNA **(Fig. 3)**. A comprehensive RT-PCR based expression profiling revealed that this gene is expressed only in normal small intestine among many normal organs. Developmental expression in the fetal brain, kidney, and lung was seen. In addition, cDNAs from matched sets of tumor and normal tissues of breast, lung, ovary, pancreas, and prostate were negative for this EST expression, demonstrating that EST AA524300 is selectively upregulated in colon tumors but not in other major solid tumors. Furthermore, the EST expression was detected in cDNAs from cell lines of colon carcinomas, but not in the cell lines of breast, lung, ovary, pancreas, or prostate carcinomas. The EST expression was detected in three of three adenomas and three of three carcinomas of colon, but not in the polyp, which suggested that the putative gene encoded by this EST may be activated during early stages of colon cancer. Elevated colon carcinoma-related gene (CCRG) protein expression also was detected in the paraffin sections of colon tumors in comparison with the corresponding normal tissues ***(17)***.

The entire cDNA sequence of the *CCRG* is deposited in GenBank under accession AF323921. The putative open reading frame codes for 111 amino acids. The predicted molecular mass of the open reading frame of CCRG is 9.4 kDa, and its theoretical p*I* is 7.6. The C terminus has a unique cysteine-rich motif, 1CX11; 2CX8; 3CX1; 4CX3; 5CX10; 6CX1; 7CX1; 8CX9; 9C10C. A PFAM search (http://pfam.wustl.edu/) indicated that such a motif is present in keratin ultrahigh sulfur matrix proteins, metallothionine, and low-density lipoprotein-receptor-related proteins.

While our work was in progress, several independent laboratories also identified this new family of genes. Holcomb et al. ***(18)*** identified a novel protein in a mouse model of allergic pulmonary inflammation that they called FIZZ1 (found in inflammatory zone). By performing a genomic screening, this group identified two other mouse homologs and two human homologs. mFIZZ2, found in intestinal crypt epithelium, had a human homolog identical to the original EST discovered in our work (AA524300) and was named hFIZZ1. Another group identified a protein in adipocytes that potentially linked obesity and insulin resistance to diabetes in a mouse model and named it resistin ***(19)***. Two other related proteins were identified in mice and humans, termed resistin-like molecule (RELM) alpha and beta ***(20)***. RELM alpha is prevalent in the stromal–vascular fraction of adipose tissue and lung, whereas PEΛM β is found in colonic epithelium. Steppan et al. ***(20)*** demonstrated that in mouse intestine RELM beta is predominant in proliferative epithelia at the base of the crypts and becomes diminished in nonproliferative differentiated epithelia that have migrated up from the crypt towards the luminal surface ***(20)***. Also, in a *min* mouse model, which is a model of familial adenomatous polyposis due to the harboring of a mutated APC gene ***(21)***, increased RELM beta mRNA expression was observed in tumors ***(20)***. Another study showed that RETNLB mRNA and protein expression was restricted to undifferentiated proliferating epithelium. These authors also detected RETNLB protein in the stools of human and mice ***(22)***. Another group ***(23)*** recently described identification of a set of novel genes by using representative difference analysis of myeloid cells from CCAAT/enhancer-binding protein δ knockout mice. A human homolog to one of the novel genes from this study was termed HXCP2, for a gene selectively upregulated in small intestine and colon. Amino acid homology studies indicate that hFIZZ1, RELM beta and HXCP2 are all 100% identical to the gene we discovered, CCRG. Based on all these results, The Human Genome Organization Gene Nomenclature Committee recently assigned the symbol RETNLB (resistin-like β) to the gene encoding CCRG/hFIZZ1/RELM betã/HXCP2. Other members of this newly described family of genes include resistin-like α, which shows specificity to adipocytes ***(24)*** and bronchial epithelial cells ***(25)*** and Resistin, which is specific to adipocytes ***(18)***.

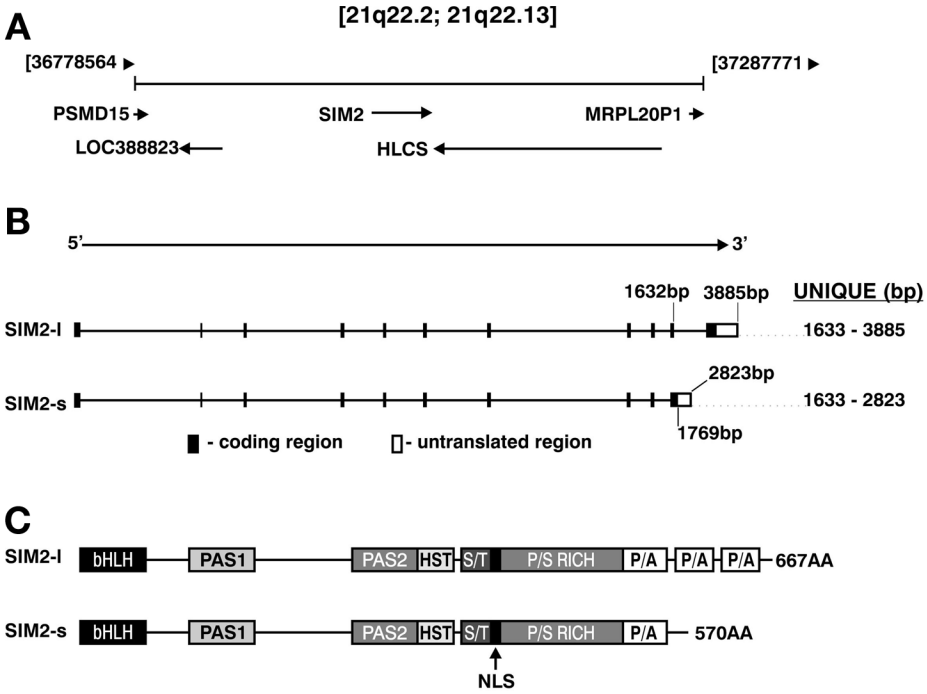



Fig. 4. Structure of human SIM2. **(A)** Location of the *SIM2* gene on a partial map (36,778,564–37,287,771 bp) of chromosome 21. HLCS, holocarboxylase synthetase; PSMD15, proteasome (prosome, macropain) 26S subunit, non-ATPase, 15; LOC388823, novel gene supported by EST alignment; MRPL20P1, mitochondrial ribosomal protein L20 pseudogene 1. **(B)** mRNAs for both isoforms of SIM2 (SIM2-l and SIM2-s) are shown. Boxes represent exons and lines represent introns. The unique 3′ sequences for each isoform are described. bp, base pair. **(C)** Schematic diagram of the conserved domains found in both SIM2 proteins. bHLH, basic helix-loop-helix; PAS1 and 2, PER/ARNT/SIM domains; HST, HIF1-α/SIM/TRH domain; S/T, Ser/Thr-rich region; NLS, nuclear localization signal; P/S RICH, Pro/Ser-rich region; P/A, Pro/Ala-rich region

## 5. Identification of Single-Minded 2 Gene as a Drug Therapy Target for Solid Tumors

Another EST predicted by the DDD tool to be upregulated in colon tumors belonged to UniGene cluster Hs. 146186, which was homologous to the Down syndrome-associated *SIM2* gene. Members of the human *SIM* gene family include *SIM1* and *SIM2*, which map to 6q16.3-q21 and 21q22.2, respectively ***(26,27)***. The *SIM2* locus spans over a 365-kilobase region on chromosome 21, as shown in **Fig. 4A** ***(28)***. In contrast to other species, the human *SIM2* gene exists in two distinct forms, the long and short forms (*SIM2-l* and *SIM2-s*) as shown in **Fig. 4B** ***(26)***. The original cDNA clone identified was *SIM2-s*, whereas genomic

sequencing subsequently identified *SIM2-l*. The *SIM2-s* mRNA includes a unique 3′ end, encoded by part of intron 10. This unique region, starting at the last nucleotide of exon 10 and including 1191 base pairs (bp) of intron 10, encompasses both coding and 3′ untranslated sequences. The 3′ end of *SIM2-l* mRNA is encoded by exon 11, which is separated from exon 10 by the complete intron 10 sequence (2528 bp). It is unclear whether these two isoforms are a consequence of an alternative use of the 3′ untranslated region contained within intron 10 of *SIM2* or are due to mispriming by using an A-rich sequence within this intron ***(26)***.

The SIM proteins belong to a family of transcription factors characterized by the basic helix-loop-helix (bHLH) and PAS (PER/ARNT/SIM) protein motifs ***(29)***. Regulation of transcription involves the dimerization of two bHLH-PAS proteins within the nucleus, DNA binding by the basic region of the bHLH, and interaction with the transcriptional machinery to modulate gene expression (reviewed in ref. ***29***). The N-terminal bHLH is the key motif for primary dimerization of the two proteins, whereas the PAS domain acts as the interface for selective protein partnering ***(29,30)***. The heterodimerizing partner choice of PAS proteins is highly specific and determines target gene regulation ***(31–33)***. The human SIM proteins are highly homologous to *Drosophila* and murine SIMs in their N-terminal domains, which contain a bHLH motif, two PAS domains, and the HST (HIF1-α/SIM/TRH) domain **(Fig. 4C)**. In the mouse, mSIM1 and mSIM2 can heterodimerize with ARNT, ARNT-2 or BMAL1 ***(31,32,34–36)***, whereas in humans, hSIM1 and hSIM2 can dimerize with either ARNT or ARNT-2 ***(37)***. A novel 23-amino acid nuclear localization signal was recently identified between residues 367 and 389 ***(38)***. This signal is in a region found in both forms of the human SIM2 protein.

The C-terminal part of the human proteins is considerably divergent from other family members; however, there is still high conservation between hSIM1 and hSIM2 compared with mSIM1 and mSIM2, respectively ***(26)***. Both forms of human SIM2 contain a Ser/Thr-rich region (S348–T366), a Pro/Ser-rich region (P385–S503), a Pro/Ala-rich region (P504–P544) and a positively charged region between R367 and R382. In addition, the long-form protein contains two additional Pro/Ala-rich regions (P533–P596 and P611–P644) and a positively charged region (K559–R575). Domains rich in Ser/Thr and proline residues are present in both transcriptional repressors ***(39)*** and transcriptional activators ***(40,41)***, whereas Pro/Ser and Pro/Ala domains are characteristic of repressor motifs (reviewed in ref. ***42***).

At the time of discovery, the *SIM2* gene has not been linked to any cancer types. Reasoning that if the DDD prediction of specificity of *SIM2-s* could be validated in a cancer model, we would be able to link the Down syndrome gene *SIM2* with cancer and hence derive a novel cancer utility for an already known gene, further studies were undertaken.

The mRNA analysis of *SIM2-s* expression of a panel of tumor and normal human tissue-derived cDNAs by using RT-PCR showed tumor-specific expression of *SIM2-s* ***(43–45)***. In contrast to the bioinformatics prediction of colon tumor specificity, the SIM2-s-specific RT-PCR product also was seen in the pancreas and prostate tumor-derived cDNAs but not in the corresponding matched normal tissues. These results underscore the importance of wet laboratory validation of the bioinformatics prediction. Using a peptide-specific polyclonal antibody to the hSIM2-s unique region, a comprehensive immunohistochemical analysis of paraffin section-embedded tumor and normal tissues was undertaken. The majority of tumor sections analyzed stained positive. In colon and pancreatic specimens, early stage adenomas also showed hSIM2-s immunoreactivity. In prostate-related samples, hSIM2-s specific immunoreactivity was detected in almost all tumors of various Gleason scores and in prostatic intraepithelial neoplasia, but it was not detected in most stromal hyperplasia. Interestingly, in the benign prostatic hyperplasia (BPH) samples, some of the sections (20/36) showed positive staining. A subset of these retrospective BPH samples (6/6) that were matched with tumor specimens obtained from the same patient was positive for *SIM2-s* expression. It is tempting to suggest that h*SIM2-s* was activated in these BPH patients before clinical manifestation of prostate cancer. Currently, prostate-specific antigen (PSA) is the only indicator in use for prostate cancer ***(46)***, and additional markers are urgently needed. If validated with a larger cohort, the SIM2-s gene has the potential to become a predictor of risk of prostate cancer development. Assay systems similar to the systems used in PSA can be developed. In an independent prostate cancer profiling study using microarrays, *SIM2* was found to be upregulated in prostate cancer tissues, validating this gene's prostate cancer specificity ***(21)***. An example of colon tumor specificity of SIM2 consistent with the bioinformatics prediction is shown in **Fig. 5**. RT-PCR analysis of 14 different tumors and corresponding normal tissues from colon cancer patients showed the expression of *SIM2-s* gene in the tumors.

A potential drug therapy use of the SIM2-s gene is inferred using antisense knockout studies ***(44,45)***. In both colon (RKO) and pancreatic (CAPAN-1) cancer models, inhibition of SIM2-s expression by antisense resulted in apoptosis in vitro and in nude mice tumorigenicity models in vivo. The induction of apoptosis by the antisense was seen in tumor cells but not in normal renal epithelial cells, despite inhibition of *SIM2-s* expression ***(47)***. Whereas the antisense-treated RKO colon carcinoma cells did not undergo cell cycle arrest, several markers of differentiation were deregulated, including alkaline phosphatase activity, a marker of terminal differentiation. Protection of apoptosis and block of differentiation showed a correlation in the RKO model. In contrast, in normal renal epithelial cells the *SIM2-s* antisense treatment did not cause induction of differentiation. These results suggested that the targets of *SIM2-s* in tumor and

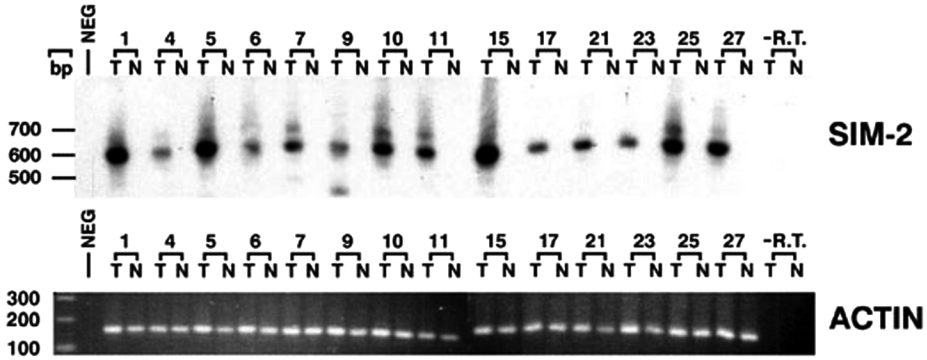



Fig. 5. Colon tumor specificity of the *SIM2-s* gene. Matched tumor and normal tissue-derived cDNAs from 14 independent colon carcinoma patients were analyzed for SIM2-s and actin gene expression by RT-PCR. The *SIM2-s*-specific PCR products were hybridized with an internal oligomer probe. M, 100-bp ladder; negative, template minus PCR control; RT, reverse transcriptase minus control.

normal cells may be different. This finding was consistent with an upregulation of a key stress response gene, Growth Arrest DNA-damage 45α in the tumor but not in the normal cells upon antisense treatment ***(47)***. The discovery of *SIM2-s* and its validation for drug therapy use offers one of the first examples of bioinformatics approaches to cancer gene discovery ***(48,49)***.

## 6. Identification of SIM2-s Targets by Microarray Technology

The transcription factor function of the SIM family of proteins ***(26)*** suggests a regulatory role. In addition, from the role of SIM in the midbrain an inference can be made of its potential involvement in differentiation ***(50)***. The precise molecular targets of the SIM proteins, however, are not known. The inhibition of SIM2-s expression in the RKO cells induces pronounced apoptosis within 14–24 h ***(44)***. Hence, we used global gene expression analysis to dissect the molecular targets that are affected in the antisense-treated cells. RKO colon carcinoma cells were treated in vitro with either the control or the antisense drug (100 n*M*) and at 10, 14, 18, and 24 h, RNA from the treated cells was analyzed using the Affymetrix U133A (largely known genes) Human Genome array. To develop better reliability, each RNA was analyzed using duplicate chips (chip replication) from two independent experiments (biological replicate). For experimental design, Minimum Guidelines for Experimental design, MGED guidelines (http://www.mged.org/) were followed and the entire data set point can be viewed from the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/) with the accession E-MEXP-101.

The Human Genome U133 (HG-U133) set, consisting of two GeneChip arrays (A and B), contains almost 45,000 probe sets representing more than

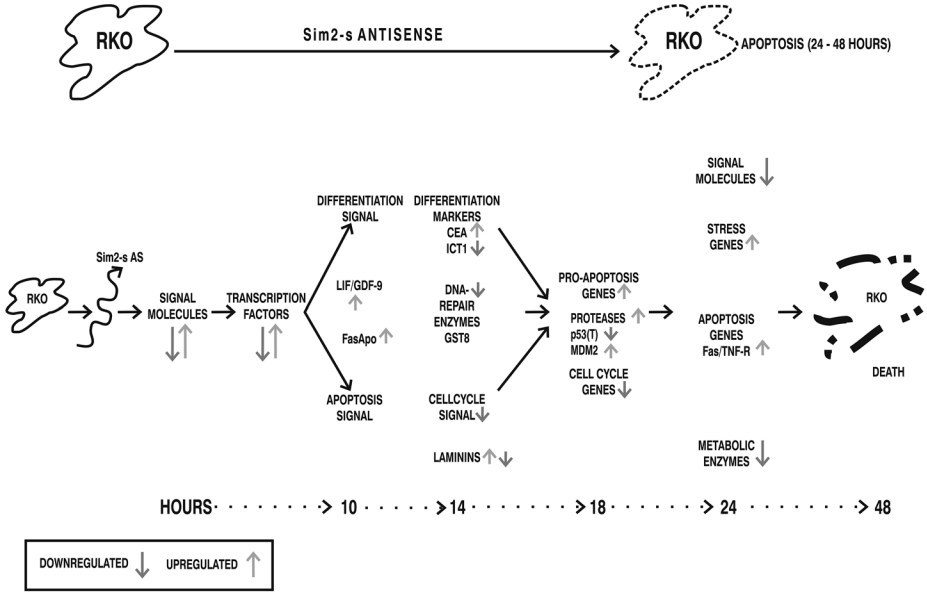



Fig. 6. Mechanism of the *SIM2-s* antisense from the GeneChip analysis. RKO colon carcinoma cells were treated with SIM2-s antisense and at the indicated time, total RNA was analyzed for a global gene expression profile with an Affymetrix U133 A and B microarray. The GeneChip output from U133A (known genes) was used to build the preliminary model as shown.

39,000 transcripts derived from approx 33,000 well-substantiated human genes. The chip output was subjected to statistical filtering (100% concordance of the hits from chip and biological replication), *p* values (>0.05) and -fold changes (>2-fold). The filtered hits were subjected to gene ontology by using GeneSpring (http://www.silicongenetics.com), and a list of genes belonging to distinct families was generated. From the pattern of the unique gene expression profile of the antisense-treated cells, a preliminary working model was generated **(Fig. 6)**. In general, a cascade of gene expression changes was seen that included early perturbation of signal transduction molecules and transcription factors. This cascade was followed by induction of differentiation signals such as leukemia inhibitory factor and growth differentiation factor. In addition, a key apoptotic signal (FasApo) was activated. Induction of differentiation markers succeeded the differentiation signals. At a later time-point, the *SIM2-s* antisense caused activation of proapoptotic genes and downregulation of proteases and tumor p53. This was accompanied by downregulation of cell cycle genes. At 24 h of treatment with the *SIM2-s* antisense, signs of stress were apparent, indicated by upregulation of stress response genes; this upregulation was accompanied by a downregulation of metabolic enzymes.

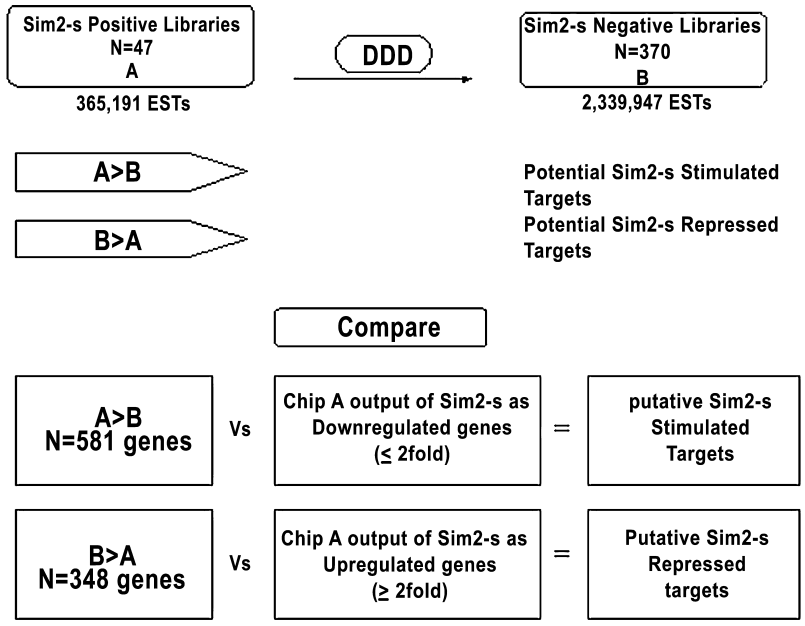



Fig. 7. DDD filter to identify SIM2-s stimulated and repressed genes. SIM2-s-positive libraries (colon, pancreas, and prostate tumor-derived) were included in pool A and SIM2-s-negative libraries (breast, ovary, and lung tumor-derived) were included in pool B. The DDD tool was used to identify SIM2-s coexpressed targets.

Efforts are currently underway to identify direct targets regulated by SIM2-s by means of bioinformatics approaches by using DDD **(Fig. 7)**. The basis for this filter is that genes expressed in tumors where SIM2-s is activated (coexpressed genes) must encompass the putative SIM2-s targets. The SIM2-s-positive (pool A) cDNA UniGene libraries can be compared with SIM2-s negative (pool B) libraries from the CGAP database by using the DDD tool from CGAP to identify SIM2-s coexpressing genes. Genes that are elevated in SIM2-s-positive libraries (A>B) are predicted to encompass SIM2-s-stimulated genes. Correspondingly, genes that are downregulated in the SIM2-s-positive libraries (B>A) are predicted to encompass potential SIM2-s-repressed genes. Among the A>B list of 581 genes, key apoptotic, survival, signal transduction molecules as well as *SIM2-s* were identified, suggesting a potential use of this filter. The antisense fingerprint of genes from the time-course experiments can be compared using this DDD filter. Comparison of A>B output with the fingerprint of antisense downregulated genes from the chip output is predicted to identify SIM2-s-stimulated targets. Conversely, comparison of B>A output with the fingerprint of antisense upregulated genes is predicted to identify SIM2-s

repressed targets. Such a strategy has been recently used to identify interleukin-8 coexpressed genes ***(51)***.

## 7. Potential Drawbacks

Although the aforementioned two examples provide a proof of concept for cancer gene discovery by using bioinformatics approaches, there are several areas requiring caution.

1. Most of the data sets for ESTs in the CGAP database encompass bulk tissue-derived cDNAs. These tissues are often contaminated with surrounding normal areas as well as necrotic regions that may contribute to both false positive and false negative results. In the future, use of laser capture-microdissected tissues should help alleviate this problem.
2. The diverse data-mining tools of the CGAP database use the UniGene clustering as a basis for partitioning the ESTs. However, due to the dynamic nature of the UniGene clustering, the member ESTs often are reassigned or withdrawn from the cluster. This reassignment or withdrawal may change the predicted expression specificity of an EST.
3. The sources of cDNAs from EST library versus SAGE library are often different; hence, different tools may identify different ESTs. Thus, the choice of a well validated library is essential for more effective gene discovery.
4. Multiple databases such as UniGene, SAGE, and Microarray are available for predicting electronic expression; however, the results from these databases may not show a correlation in the sources of tissues.
5. In the two examples discussed, the *CCRG/RETNLB* gene validation was consistent with the DDD prediction of colon cancer specificity. In contrast, the *SIM2* gene was specific not only to colon but also to prostate and pancreatic cancers. Thus, depending on the user's definition of stringency of target selection, the *SIM2-s* gene might or might not have scored positive.
6. The wet laboratory validation of the chosen EST for expression specificity often involves the use of patient-derived tissues. It is often difficult to define the degree of normalness in the surrounding normal tissues. Further patient-to-patient variation in gene expression contributes to loss or gain of gene expression. Hence, a statistically significant number of samples needs to be analyzed for a chosen EST before establishing the specificity of expression.
7. One of the drawbacks of GeneChip-based experiments is that the chip output can be very large (thousands of genes). Development of multiple filters at the computational level (bioinformatics) and at the biological level (system) is crucial to reducing the number of potential gene targets.

## 8. Conclusions

The completion of the human genome sequencing efforts promises to offer new ways to discover genes with novel diagnostic and therapeutic potential for diverse

diseases. Data-mining the cancer genome allows us to rapidly discover cancer genes as we have shown with two specific examples. Microarray technology and bioinformatics approaches can be used in conjunction to facilitate target(s) discovery and to clarify the mechanism. Currently, considerable false positives and false negatives are encountered due to the nature of the cDNA libraries. However, better integration of the databases and improvements in the quality of the EST libraries in the future will greatly improve the gene discovery process.

## Acknowledgments

I thank members of my laboratory for valuable contributions and Jeanine Narayanan for editorial assistance.

## References

1. Andrade, M. A. and Sander, C. (1997) Bioinformatics: from genome data to biological knowledge. *Curr. Opin. Biotechnol.* **8,** 675–683.
2. Cavalli-Sforza, L. L. (2005) The Human Genome Diversity Project: past, present and future. *Nat. Rev. Genet.* **6,** 333–340.
3. Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. (1998) New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282,** 682–689.
4. Robbins, R. J. (1996) Bioinformatics: essential infrastructure for global biology. *J. Comput. Biol.* **3,** 465–478.
5. Fannon, M. R. (1996) Gene expression in normal and disease states–identification of therapeutic targets. *Trends Biotechnol.* **14,** 294–298.
6. Elek, J., Park, K. H., and Narayanan, R. (2000) Microarray-based expression profiling in prostate tumors. *In Vivo* **14,** 173–182.
7. Heller, R. A., Schena, M., Chai, A., et al. (1997) Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA* **94,** 2150–2155.
8. Khan, J., Bittner, M. L., Saal, L. H., et al. (1999) cDNA microarrays detect activation of a myogenic transcription program by the PAX3-FKHR fusion oncogene. *Proc. Natl. Acad. Sci. USA* **96,** 13,264–13,269.
9. Lockhart, D. J., Dong, H., Byrne, M. C., et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14,** 1675–1680.
10. Lal, A., Lash, A. E., Altschul, S. F., et al. (1999) A public database for gene expression in human cancers. *Cancer Res.* **59,** 5403–5407.
11. Nacht, M., Ferguson, A. T., Zhang, W., et al. (1999) Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer. *Cancer Res.* **59,** 5464–5470.
12. Strausberg, R. L., Dahl, C. A., and Klausner, R. D. (1997) New opportunities for uncovering the molecular basis of cancer. *Nat. Genet.* **15,** 415–416.
13. Wheeler, D. L., Chappey, C., Lash, A. E., et al. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **28,** 10–14.

14. Zhang, L., Zhou, W., Velculescu, V. E., et al. (1997) Gene expression profiles in normal and cancer cells. *Science* **276,** 1268–1272.
15. Schmitt, A. O., Specht, T., Beckmann, G., et al. (1999) Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res.* **27,** 4251–4260.
16. Scheurle, D., DeYoung, M. P., Binninger, D. M., Page, H., Jahanzeb, M., and Narayanan, R. (2000) Cancer gene discovery using digital differential display. *Cancer Res.* **60,** 4037–4043.
17. DeYoung, M. P., Damania, H., Scheurle, D., Zylberberg, C., and Narayanan, R. (2002) Bioinformatics-based discovery of a novel factor with apparent specificity to colon cancer. *In Vivo* **16,** 239–248.
18. Holcomb, I. N., Kabakoff, R. C., Chan, B., et al. (2000) FIZZ1, a novel cysteine-rich secreted protein associated with pulmonary inflammation, defines a new gene family. *EMBO J.* **19,** 4046–4055.
19. Steppan, C. M., Bailey, S. T., Bhat, S., et al. (2001) The hormone resistin links obesity to diabetes. *Nature* **409,** 307–312.
20. Steppan, C. M., Brown, E. J., Wright, C. M., et al. (2001) A family of tissue-specific resistin-like molecules. *Proc. Natl. Acad. Sci. USA* **98,** 502–506.
21. Su, A. I., Cooke, M. P., Ching, K. A., et al. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. USA* **99,** 4465–4470.
22. He, W., Wang, M. L., Jiang, H. Q., et al. (2003) Bacterial colonization leads to the colonic secretion of RELMbeta/FIZZ2, a novel goblet cell-specific protein. *Gastroenterology* **125,** 1388–1397.
23. Kubota, T., Kawano, S., Chih, D. Y., et al. (2000) Representational difference analysis using myeloid cells from C/EBP epsilon deletional mice. *Blood* **96,** 3953–3957.
24. Blagoev, B., Kratchmarova, I., Nielsen, M. M., et al. (2002) Inhibition of adipocyte differentiation by resistin-like molecule alpha. Biochemical characterization of its oligomeric nature. *J Biol Chem.* **277,** 42,011–42,016.
25. Teng, X., Li, D., Champion, H. C., and Johns, R. A. (2003) FIZZ1/RELMalpha, a novel hypoxia-induced mitogenic factor in lung with vasoconstrictive and angiogenic properties. *Circ. Res.* **92,** 1065–1067.
26. Chrast, R., Scott, H. S., Chen, H., et al. (1997) Cloning of two human homologs of the *Drosophila* single-minded gene SIM1 on chromosome 6q and SIM2 on 21q within the Down syndrome chromosomal region. *Genome Res.* **7,** 615–624.
27. Dahmane, N., Charron, G., Lopes, C., et al. (1995) Down syndrome-critical region contains a gene homologous to *Drosophila* sim expressed during rat and human central nervous system development. *Proc. Natl. Acad. Sci. USA* **92,** 9191–9195.
28. Frazer, K. A., Tao, H., Osoegawa, K., et al. (2004) Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res.* **14,** 367–372.
29. Crews, S. T. and Fan, C. M. (1999) Remembrance of things PAS: regulation of development by bHLH-PAS proteins. *Curr. Opin. Genet. Dev.* **9,** 580–587.

30. Taylor, B. L. and Zhulin, I. B. (1999) PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol. Mol. Biol. Rev.* **63,** 479–506.
31. Ema, M., Morita, M., Ikawa, S., et al. (1996) Two new members of the murine Sim gene family are transcriptional repressors and show different expression patterns during mouse embryogenesis. *Mol. Cell Biol.* **16,** 5865–5875.
32. Michaud, J. L., DeRossi, C., May, N. R., Holdener, B. C., and Fan, C. M. (2000) ARNT2 acts as the dimerization partner of SIM1 for the development of the hypothalamus. *Mech. Dev.* **90,** 253–261.
33. Pongratz, I., Antonsson, C., Whitelaw, M. L., and Poellinger, L. (1998) Role of the PAS domain in regulation of dimerization and DNA binding specificity of the dioxin receptor. *Mol. Cell. Biol* **18,** 4079–4088.
34. Moffett, P. and Pelletier, J. (2000) Different transcriptional properties of mSim-1 and mSim-2. *FEBS Lett.* **466,** 80–86.
35. Probst, M. R., Fan, C. M., Tessier-Lavigne, M., and Hankinson, O. (1997) Two murine homologs of the *Drosophila* single-minded protein that interact with the mouse aryl hydrocarbon receptor nuclear translocator protein. *J. Biol. Chem.* **272,** 4451–4457.
36. Swanson, H. I., Chan, W. K., and Bradfield, C. A. (1995) DNA binding specificities and pairing rules of the Ah receptor, ARNT, and SIM proteins. *J. Biol. Chem.* **270,** 26,292–26,302.
37. Ooe, N., Saito, K., Mikami, N., Nakatuka, I., and Kaneko, H. (2004) Identification of a novel basic helix-loop-helix-PAS factor, NXF, reveals a Sim2 competitive, positive regulatory role in dendritic-cytoskeleton modulator drebrin gene expression. *Mol. Cell. Biol.* **24,** 608–616.
38. Yamaki, A., Kudoh, J., Shimizu, N., and Shimizu, Y. (2004) A novel nuclear localization signal in the human single-minded proteins SIM1 and SIM2. *Biochem. Biophys. Res. Commun.* **313,** 482–488.
39. Madden, S. L., Cook, D. M., Morris, J. F., Gashler, A., Sukhatme, V. P., and Rauscher, F. J., III (1991) Transcriptional repression mediated by the WT1 Wilms tumor gene product. *Science* **253,** 1550–1553.
40. Franks, R. G. and Crews, S. T. (1994) Transcriptional activation domains of the single-minded bHLH protein are required for CNS midline cell development. *Mech. Dev.* **45,** 269–277.
41. Mermod, N., O'Neill, E. A., Kelly, T. J., and Tjian, R. (1989) The proline-rich transcriptional activator of CTF/NF-I is distinct from the replication and DNA binding domain. *Cell* **58,** 741–753.
42. Hanna-Rose, W. and Hansen, U. (1996) Active repression mechanisms of eukaryotic transcription repressors. *Trends Genet.* **12,** 229–234.
43. DeYoung, M. P., Scheurle, D., Damania, H., Zylberberg, Z., and Narayanan, R. (2002) Down's syndrome-associated Single Minded gene as a novel tumor marker. *Anticancer Res.* **22,** 3149–3158.
44. DeYoung, M. P., Tress, M., and Narayanan, R. (2003) Identification of Down's syndrome critical locus gene SIM2-s as a drug therapy target for solid tumors. *Proc. Natl. Acad. Sci. USA* **100,** 4760–4765.

45. DeYoung, M. P., Tress, M., and Narayanan, R. (2003) Down's syndrome-associated Single Minded 2 gene as a pancreatic cancer drug therapy target. *Cancer Lett.* **200,** 25–31.
46. Goolsby, M. J. (2001) Use of PSA measurement in practice. *J. Am. Acad. Nurse Pract.* **13,** 246–248.
47. Aleman, M. J., DeYoung, M. P., Tress, M., Keating, P., Perry, G. W., and Narayanan, R. (2005) Inhibition of single minded 2 gene expression mediates tumor-selective apoptosis and differentiation in human colon cancer cells. *Proc. Natl. Acad. Sci. USA* **102,** 12,765–12,770.
48. Ratan, R. R. (2003) Mining genome databases for therapeutic gold: SIM2 is a novel target for treatment of solid tumors. *Trends Pharmacol. Sci.* **24,** 508–510.
49. Touchette, N. (2003) Mouse-to-mouse revelation: genome yields cancer drug target. Genome News Network. http://www.genomenewsnetwork.org/articles/04_03.mouse.shtml. April 18, 2003.
50. Nambu, J. R., Lewis, J. O., Wharton, K. A., Jr., and Crews, S. T. (1991) The *Drosophila* single-minded gene encodes a helix-loop-helix protein that acts as a master regulator of CNS midline development. *Cell* **67,** 1157–1167.
51. Benbow, L., Wang, L., Laverty, M., et al. (2002) A reference database for tumor-related genes co-expressed with interleukin-8 using genome-scale in silico analysis. *BMC Genomics* **3,** 29.