7008–7021 *Nucleic Acids Research, 2010, Vol. 38, No. 20* *Published online 9 July 2010*
doi:10.1093/nar/gkq574

Meta-analysis of cancer gene expression signatures reveals new cancer genes, SAGE tags and tumor associated regions of co-regulation

Erşen Kavak[1,2,*], Mustafa Ünlü[1], Monica Nistér[3] and Ahmet Koman[1]

[1]Department of Molecular Biology and Genetics, Boğaziçi University, Istanbul, Turkey, [2]Department of Cell and Molecular Biology and [3]Department of Oncology-Pathology, Karolinska Institutet, Stockholm, 171 77 Stockholm, Sweden

Received December 9, 2009; Revised June 1, 2010; Accepted June 7, 2010

ABSTRACT

Cancer is among the major causes of human death and its mechanism(s) are not fully understood. We applied a novel meta-analysis approach to multiple sets of merged serial analysis of gene expression and microarray cancer data in order to analyze transcriptome alterations in human cancer. Our methodology, which we denote 'COgnate Gene Expression patterNing in tumours' (COGENT), unmasked numerous genes that were differentially expressed in multiple cancers. COGENT detected well-known tumor-associated (TA) genes such as TP53, EGFR and VEGF, as well as many multi-cancer, but not-yet-tumor-associated genes. In addition, we identified 81 co-regulated regions on the human genome (RIDGEs) by using expression data from all cancers. Some RIDGEs (28%) consist of paralog genes while another subset (30%) are specifically dysregulated in tumors but not in normal tissues. Furthermore, a significant number of RIDGEs are associated with GC-rich regions on the genome. All assembled data is freely available online (www.oncoreveal.org) as a tool implementing COGENT analysis of multi-cancer genes and RIDGEs. These findings engender a deeper understanding of cancer biology by demonstrating the existence of a pool of under-studied multi-cancer genes and by highlighting the cancer-specificity of some TA-RIDGEs.

INTRODUCTION

Cancer is still one of the most fatal diseases in the industrialized world. Cancer cells utilize an unbalanced state in the genome, epigenome and transcriptome to survive and proliferate, leading to the death of the host through multiple processes. Transcriptomes of cancers are being increasingly analyzed through the use of microarrays and other methods, including serial analysis of gene expression (SAGE) (1), SAGE data produced by new generation sequencing technologies (tag-Seq), (2) and RNA-Seq. This rich data cloud in turn lends itself well to cross-sectional studies that focus on identifying genes that are differentially expressed in multiple studies and multiple cancers. Rhodes *et al.* (3) first demonstrated that some gene expression changes are common to cancers. This concept was later extended and expanded (4). These studies showed that cancer gene expression patterns can sort cancer and normal tissue when used as a diagnostic signature. Results derived from such gene expression studies led to the first United States Food and Drug Administration approved gene expression signature based test, Mammaprint (5). Mammaprint predicts therapy outcome in breast cancer, providing a good example of the predictive power of gene expression signatures. In extending the concept further, another study (6) compared tissue- *and* cancer-specific gene expression to show that melanomas over-express more brain selective genes than other types of cancer, which the authors hypothesized might explain melanoma metastasis to the brain as a frequent outcome. Other studies utilizing similar approaches have shown that E2F transcription factor is likely mediating gene over-expression in most human cancers (7). In spite of the explanatory power of using gene expression signatures for diagnosis and classification, therapeutic targeting of even single genes is still a nascent field. Therefore, it is of great importance to identify novel cancer genes and gene variants as targets of therapy. Moreover, in spite of the popularity and availability of microarray technology, which has formed the basis of most earlier cross-sectional studies, sequencing based methods have the advantage of enabling analyses of both known and novel genes since they are not dependent on pre-selected probes. Information derived from

*To whom correspondence should be addressed. Tel: +46 703 832 668; Email: ersen.kavak@ki.se; ersenkavak@gmail.com sequencing based methods can serve to augment probe based methods by highlighting, for example, the involvement of differentially expressed splice-variants.

Co-regulation of proximate genes points to another, higher order control of gene expression. Cohen *et al.* (8) first elucidated the co-regulation of adjacent yeast gene pairs or triplets by comparing different data sets, such as cell cycle time course (14% of genes co-regulated) and sporulation (23% of genes co-regulated). Later, Caron *et al.* (9) defined regions of increased gene expression (RIDGEs) along chromosomes by analyzing SAGE libraries from different human tissues. This was followed by demonstration of large domains (10–30 genes) of co-regulation in *Drosophila*, which comprised 20% of all genes on the fruit-fly genome (10). And recently, Stransky *et al.* (11) identified RIDGEs within human bladder cancers and matched normals. Some of these RIDGEs were explained by chromosomal amplifications/deletions, but some of the RIDGEs were not. As data from developing high throughput methods accumulate, identification of genomic regions with similar regulation patterns becomes more useful in determining which, if any, transcriptional or epigenetic events may be involved in generating such co-regulation.

Considering the above studies and observations, we performed a meta-analysis of gene expression in cancer tissues along with matched normal tissues using integrated SAGE and microarray data. We found that over-expressed multi-cancer genes are significantly enriched for article annotations in spite of having a high ratio of not yet tumor-associated (NYTA) genes. In addition, we expanded our analysis to identify TA regions of increased gene expression (TA-RIDGEs) in comparison to Normal tissue-Associated RIDGEs (NA-RIDGEs). We borrowed the acronym RIDGE from Caron *et al.*'s concept, even though we studied regions of co-regulated differential expression. As well as showing the G/C richness of RIDGEs, we point to the distinct expression of a subgroup of the keratin gene family compared to epidermal differentiation complex (EDC) (12) expression in skin cancers. Finally we make all of our data publicly available in an online tool, oncoreveal (http://www.oncoreveal.org).

METHODS

Data collection and processing

In total 170 normal and 132 cancer SAGE libraries representing 32 different types of cancers, as well as 477 normal and 927 cancer microarray samples representing 37 different types of cancers were analyzed. In order to merge SAGE and microarray data, we mapped microarray probes and SAGE tags to Entrez Gene IDs. In total, this corresponds to 49 different types of tumors comprising four major types of human cancer, which are epithelial, hematological, central nervous system and connective tissue tumors. This data has been published in at least 32 separate articles (see Supplementary Tables S2 and 3 for lists and references). Of the 37 microarray data sets, 25 were downloaded from Oncomine. The rest was collected from the Entrez GEO database. In order to make the GEO data comparable to oncomine data, data was log2 transformed and median was set to 0 by subtracting 'median of all samples' from each data point, and SD was set to 1 by dividing each data point to standard deviation among all samples in the study. A Benjamini–Hochberg corrected *P*-value (*Q*-value) for multiple hypothesis testing (13) was calculated as the corrected *P*-value obtained from student's *t*-test as in (3). To maximize data retrieval, we used an iterative extraction process that scans through different Entrez Gene releases to convert Oncomine released gene symbols to Entrez Gene IDs (Supplementary Data). During pre-processing, probesets matching to multiple genes (in average 1%) or probesets matching to ESTs or non-matching probesets (in average 22%) were discarded (Supplementary Table S2B).

If multiple probesets matched to a single gene, we considered the average fold, *P*- and Q-values of multiple probesets that pointed in the same direction (over or under expression). If there were probesets which pointed in both directions, we considered that gene as both over and under-expressed in the cognate gene expression patterning in tumours (COGENT) procedure, which might be explained by multiple isoforms of the same gene.

Digital Gene Expression Display (DGED) tool was used to find differentially expressed SAGE tags between several different tumors and corresponding normals. DGED is one of the tools under the SAGE Genie (14) web platform founded to analyze CGAP data. The two parameters of output in DGED are the *F*- and *P*-values; which are statistical parameters to define the stringency of differential expression when comparing SAGE tag expression values in different samples as explained in (15). F parameter of DGED was chosen to be 1.5 to allow detection of mild and consistent changes. For SAGE COGENT analysis, long SAGE tags were trimmed to short SAGE tags and analyzed together with short SAGE libraries. For SAGE & microarray COGENT, short or long SAGE tags were converted to Entrez Gene IDs by using consensus tag mapping ('Results' section).

Tag-Seq data produces significantly more significant differences (e.g. ~10 000 differentially expressed genes versus ~500 differentially expressed genes) when compared to regular SAGE libraries due to very high tag counts. In order to accommodate this restriction, we calculated the average number of differential expression events in non-tag-Seq datasets and took the same number of differentially expressed genes or tags from tag-Seq datasets for microarray & SAGE COGENT or SAGE COGENT. We called this procedure rank selection.

RT–PCRs and tissue isolations

Brain tumor samples were collected at the Brain Surgery Department of Cerrahpaşa Hospital in Istanbul, Turkey with informed consents from patients and under permission from Cerrahpaşa Hospital's ethical committee. Non-tumor brain tissues were collected from epilepsy surgeries. All tumor tissues were diagnosed and non-tumor tissues verified to be normal by pathologists at Cerrahpaşa Hospital's Pathology Department. Tissues were fresh frozen within 30 min of removal from brain. RNA from brain tissues was extracted by using Qiagen Rneasy Lipid Tissue Midi Kit. First strand cDNA synthesis was performed with the Improm RT cDNA synthesis system. RT–PCRs were performed using Taq Polymerase (Fermentas). qRT–PCR reactions were carried out using Fast-start sybGreen kit (Roche, cat# 03003230001), and Roche Lightcycler 1.5. There is a high amount of *ACTB* or *GAPDH* up-regulation in tumors when compared to normals, as suggested by COGENT and verified in the independent central nervous system (CNS) tumor panel (Supplementary Figure S1). We therefore used S18 ribosomal RNA (Genbank accession: X03205) to normalize. The expression relative to the average normal varied between 0.67 and 1.28 in S18. This was a reasonable interval when compared to that of *ACTB* which was 0.87–2.60. For the primers used for PCRs, please see Supplementary Data.

Multi-cancer genes

Python scipy (scientific python) and matplotlib libraries were used for the tests and graphs presented in Figure 3. We used geneRIF database to scan gene related articles which contains at least 1 geneRIF for ~10 000 genes (approximately half of the genes covered in this study). To define a cancer geneRIF; we scanned the geneRIFs for any of the 'tumor', 'carcinoma', 'cancer' or 'neopla' keywords.

When assigning a rank to genes to define the amount of change in single cancer-normal comparison, we used the lowest rank (highest change) in the case of multiple probesets/tags. We used the Entrez Homologene database to assign ortholog numbers.

To account for multiple hypotheses testing over genes for microarray data we performed a Benjamini–Hochberg correction as explained earlier. For SAGE data, we chose a relatively stringent *P*-value (0.02) for the analysis presented. To calculate the false discovery rate (FDR) of multiple hypotheses testing over samples we calculated the expected number of false positives by a randomization based approach which is based on randomly selecting X number of genes from Y number of cancer types; where X represents the number of genes altered in a certain cancer type.

RIDGEs

We assigned a score to each gene by using an expression matrix from the mean values of tumor and normal samples from the microarray studies. We did not consider SAGE data at this step in order to avoid possible biases from merging two different types of data distribution. The score is the average Spearman rank correlation coefficient (R_s) between the gene to be scored and each neighboring gene within a ± 21 window size. Genes with less than three expression data points were ignored in the calculations. To define NA-RIDGEs, we used the processed expression values from a wide histologically normal human tissue expression study (GSE2361) from Entrez GEO. (16)

To assign *P*-value thresholds to this average score (to assess which scores are significantly high), we calculated the average R_s score for a gene with N random genes from different chromosomes instead of neighboring genes, where N represents twice the window size. In order to optimize the window size, we repeated this procedure for window sizes 2–48. We then picked 21 as the optimum window size, because it maximizes the number of RIDGEs and the number of genes above the threshold does not increase any more after 21 (Supplementary Figure S3).

Human genome version 'hg18′ from UCSC genome browser was used in all analyses.

ICEBERG algorithm, zooming on RIDGEs and G/C content

In order to assign member genes to each TA-RIDGE we developed a double *P*-value approach to catch the iceberg like structure of the TA-RIDGE. The ICEBERG algorithm finds regions with at least one gene above a more stringent *P*-value (primary *P*-value) and detects the genes above a less stringent *P*-value (secondary *P*-value) in the close vicinity (we selected one window size upstream and two window sizes downstream as the proper neighborhood as a good separator of nearby RIDGEs, after manual curation of different possibilities). When working on differential expression of genes in TA-RIDGEs we first assigned a score of over-representation of differential expression for each cancer, which calculates how common differential expression would be in the same window size when the same number of genes as altered gene count are randomly selected from the corresponding probeset collection (array platform or all available SAGE to Entrez Gene ID mappings for short or long tag mappings). The co-regulation *P*-value was calculated with a similar randomization (e.g. arcs in RIDGE drawings represent co-regulations with a $P < 0.002$, please see Supplementary Data for a detailed explanation and the pseudocodes).

We calculated G/C content of the whole genomic region of RIDGEs for the data presented in Table 2. G/C or GC/CG content of the gene's genomic region or the promoter region (−2000:+250, relative to transcriptional start site) was calculated for the data presented in Supplementary Figure S9. To select non-RIDGEs of varying window sizes, we sought regions containing N consecutive genes with non-significant average Spearman scores ($P > 0.05$) where N stands for the window size of the non-RIDGE region.

The melanoma metastasis dataset was adapted from Entrez GEO record GSE8401 (17); which represents RNAs of fresh frozen tissue samples from either primary melanomas or melanoma metastasis samples.

RESULTS

We performed a gene and SAGE tag-centric meta-analysis of cancer gene expression data in order to determine which transcriptional units are common to multiple different

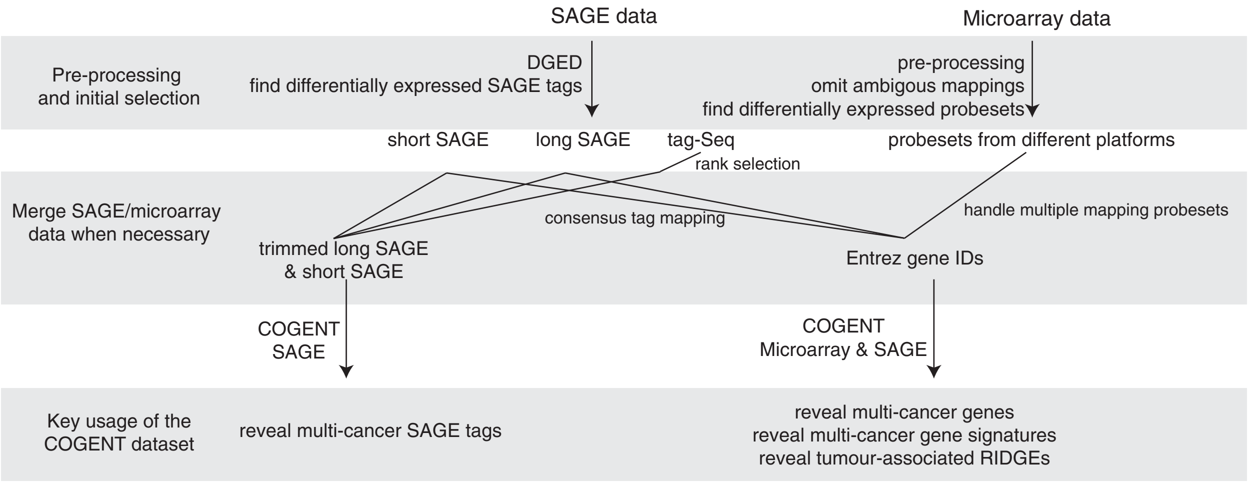

SAGE data
Microarray data
Pre-processing and initial selection
DGED
find differentially expressed SAGE tags
pre-processing
omit ambigous mappings
find differentially expressed probesets
short SAGE
long SAGE
tag-Seq
rank selection
probesets from different platforms
Merge SAGE/microarray data when necessary
consensus tag mapping
handle multiple mapping probesets
trimmed long SAGE & short SAGE
Entrez gene IDs
COGENT SAGE
COGENT Microarray & SAGE
Key usage of the COGENT dataset
reveal multi-cancer SAGE tags
reveal multi-cancer genes
reveal multi-cancer gene signatures
reveal tumour-associated RIDGEs

Figure 1. Study design.

cancers. We named our methodology cognate gene expression patterning in tumors (COGENT—Figure 1). One of the ways in which our findings differ significantly from previously published meta-analyses is that our datasets comprise all major types of cancer: epithelial, hematological, central nervous system and connective tissue tumors. Tumor names, classifications and datasets used are listed in Supplementary Tables S1–4. In discussing our results, change in cancer will mean the change relative to corresponding normal unless otherwise stated.

We developed several methods in order to efficiently use the data and to minimize false positives. Forty-three percent of the short tag (14 bp) mappings and 7% of the long tag (21 bp) mappings do not agree between single mappings of SAGE Map and best gene calls of SAGE Genie, the two most commonly used algorithms for SAGE tag mapping (for detailed comparison of the two algorithms; see Supplementary Data). In order to minimize false positives, we restricted our analyses to SAGE tags for which SAGE Map picked a single gene, which also was the same as SAGE Genie's best gene. We call this method consensus tag mapping. A total of 68 and 80% of genes can be detected for short and long SAGE tags, respectively, by consensus tag mapping. We further selected 13 genes, all of which were assigned by consensus tag mapping, and verified differential expression in an independent tumor–non-tumor brain tissue panel. RT–PCR and Q-RT–PCR results were consistent with the expression differences COGENT suggested (Figure 2, Supplementary Figure S1 and Table S5). We also developed another method, rank selection, in order to incorporate SAGE data produced by tag-Seq. Briefly, rank selection equates the amount of differential expression in tag-Seq data with the rest of datasets (See 'Methods' for a more detailed explanation).

Multi-cancer genes

In order to identify genes that are differentially expressed in several types of tumors we analyzed a dataset where we merged SAGE tags with microarray probesets using Entrez Gene IDs (See Figure 1 and 'Methods' section).

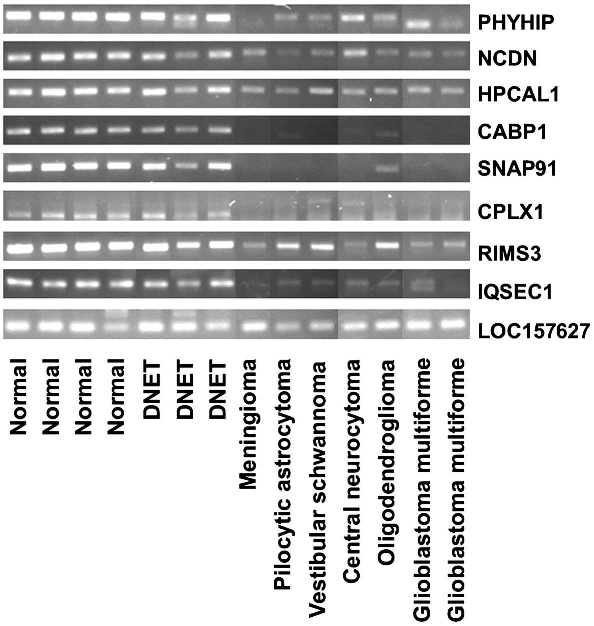


Figure 2. Verification of oncoreveal predictions in an independent panel of tumor and non-tumor brain samples. Verification of nine selected NYTA downregulated multi-cancer genes by RT–PCR. Verification of three selected NYTA upregulated multi-cancer genes is presented at Supplementary Table S1. The corresponding COGENT analysis for these genes is presented at Supplementary Table S5. DNET: Dysmbryoplastic NeuroEpithelial Tumor.

To obtain the results presented hereafter, we used a *Q*-value filter of 0.05 [Benjamini–Hochberg (13) corrected *P*-value from student's *t*-test as in (3)] for microarray probesets, and a *P*-value of 0.02 together with an *F*-value of 1.5 for SAGE tags [a Bayesian test described in (15)] for defining differential expression.

Interestingly (or expectedly, depending on one's perspective) we found that genes that are over-expressed in more types of cancers (multi-cancer genes) are also more likely to have article annotations (assessed by having annotated geneRIF: author entered summary of the article; see (18) and http://www.ncbi.nlm.nih.gov/projects/GeneRIF/) or being annotated as cancer related

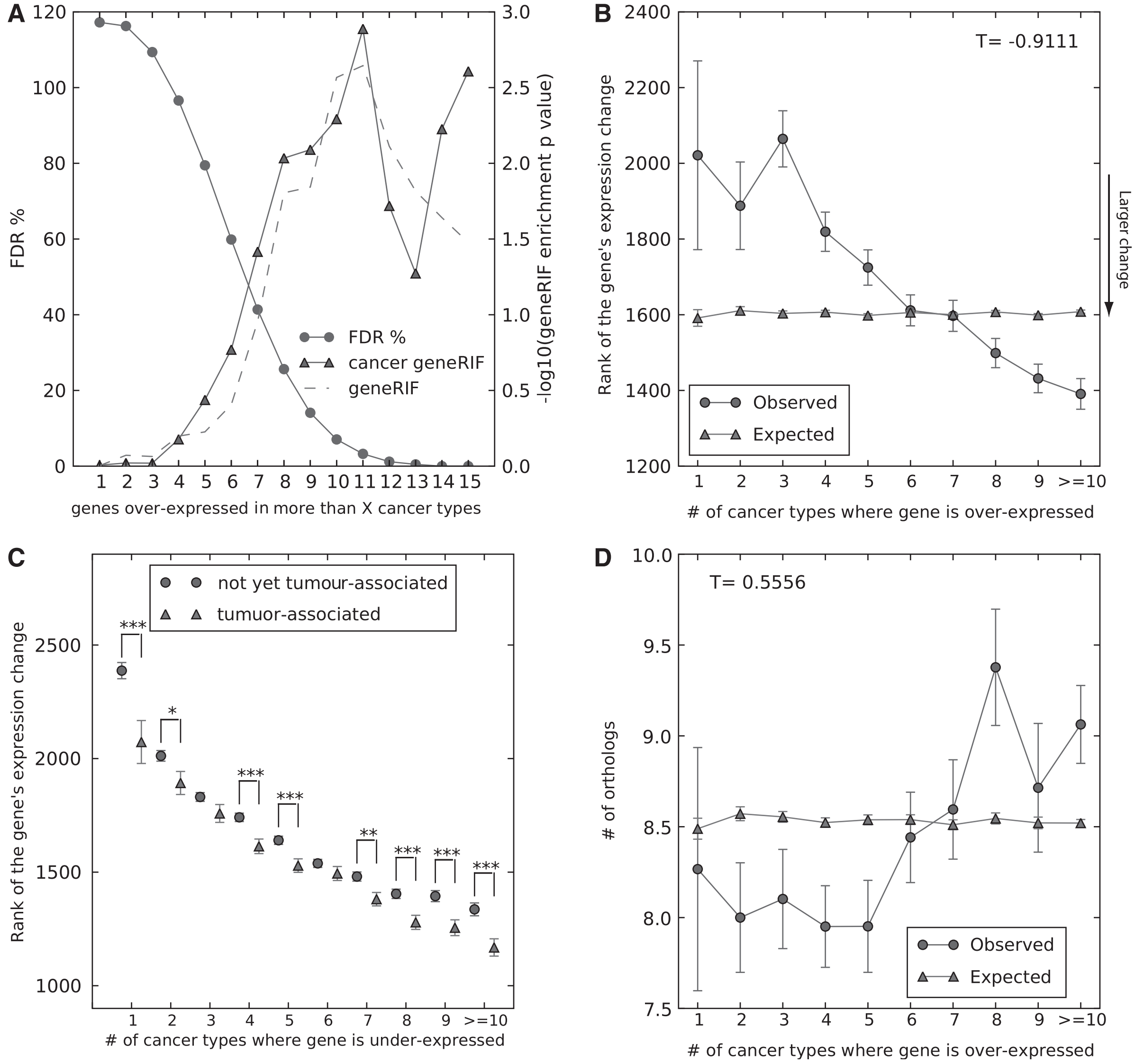


Figure 3. Multi-cancer genes. (**A**) Multi-cancer genes are more likely to have been studied and have cancer annotations. X axis indicates gene sets which are altered in more than a certain number of cancer types. Y axis indicates either FDR percentage or minus log *P*-value of the geneRIF enrichment. We used only those genes which were common to all platforms in this analysis. (**B**) Rank of a gene's expression change (in a single cancer-normal comparison) decreases with the number of cancer types it is over-expressed in. To define rank, each cancer-normal comparison data set is sorted by *P*-value for microarray data and first by *P*-value and then fold value for SAGE data. The minimum rank (highest change) of a gene was considered when multiple probesets mapped to a single gene. *T*-value is Kendall Tau B correlation coefficient between X axis values and means of each subgroup in the Y axis ($P = 0.00024$). Highly significant but smaller correlation exists when all data points rather than means are considered ($T = -0.09$, $P = 6e-41$). We used only those genes which were common to all platforms in this analysis. The expected distributions in B and D are the randomly shuffled versions of the observed values to control for the within group sample size effect, if any. (**C**) Rank of a gene's expression change in a single cancer type explains being associated with cancer. Ranks of the TA and not yet TA genes are significantly different for most of the X axis values (no. of cancer types), and when the data is not sub-divided by the X axis ($P < 0.000000$, Mann–Whitney U-test; data not shown). All genes are used in the analysis. (**D**) Over-expressed multi-cancer genes have more orthologs (homologs in different species). *T*-value is Kendall Tau B correlation coefficient between X axis values and means of each subgroup on the Y axis ($P = 0.025$). A significant but smaller correlation exists when all data points rather than means are considered ($T = 0.06$, $P = 1.46e-05$). We used only those genes which were common to all platforms in this analysis. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.005$. *P*-values are of two independent samples *t*-test. In all of the graphs, the point markers are means of the data, and the error bars indicate standard error.

(assessed by having cancer geneRIF based on filtering for several keywords; Figure 3A, Supplementary Figure S2A and Table S6). When FDR of the multi-cancer gene signature and enrichment probability of cancer geneRIFs are considered, being over-expressed in more than 10 cancer types seems like a reasonable threshold to define a multi-cancer gene set (randomization based FDR = 7%; geneRIF enrichment $P = 0.002$, cancer geneRIF enrichment $P = 0.005$, ranksums test; Figure 3A). There is not a similar enrichment of geneRIFs for under-expressed multi-cancer genes although FDR percent follows a similar pattern (Supplementary Figure S2A). Importantly, the genes which are common to all platforms ($N = 1896$) are significantly enriched for geneRIFs and cancer geneRIFs (ranksums test, $P < 1e-13$ for both). Therefore, genes common to all platforms were used where this platform bias would affect results.

Genes that show higher expression changes in single cancer types also tend to change more in multiple cancers (Figure 3B and Supplementary Figure S2B). Multi-cancer genes are thus more likely to be associated with cancer; and this is true for many which are familiar, well-studied cancer genes (*TP53*, *EGFR*, *ERBB2*, *VEGFA*, *PTGS2*, *BRCA1*, *CDKN2A*, *ESR1*, *PTEN*, *CCND1* and *HIF1A;* Supplementary Table S7A). We were however intrigued to also find numerous multi-cancer genes with no previous cancer annotation (e.g. *SYNCRIP*, *P4HA1*, *GART* and *ASCC3;* see Supplementary Table S7B for a representative list. For a more comprehensive analysis of multi-cancer genes, see www.oncoreveal.org or Supplementary Excel file). A total of 794 genes are over-expressed in more than 10 different types of cancers (FDR = 7.1%), but 437 of these do not have any cancer geneRIFs. A total of 212 genes are over-expressed in more than 13 types of cancers (FDR = 0.4%), 110 of them not being associated with cancer (See Supplementary Table S6 for detailed statistics). These NYTA genes differ in their rankings in single studies from TA multi-cancer genes. In general, a gene is probably already associated with cancer if it ranks high in a single type of cancer (Figure 3C and Supplementary Figure S2C), but multi-cancer genes that have lower rankings are less likely to have been studied. These results indicate the utility of COGENT, which enabled us to discover genes related to cancer in multiple cancer types, but which had not previously been linked with cancer due to their lower rankings in single studies.

We also functionally classified all multi-cancer genes which are differentially expressed in more than 10 cancer types. As expected, the entire set of over-expressed multi-cancer genes are significantly enriched (DAVID: Benjamini–Hochberg < 0.05) (19) for cancer related functions such as cell cycle and apoptosis (see Supplementary Table S8A for the top 50 enriched classes). In terms of signaling pathways, over-expressed multi-cancer genes are significantly enriched for TP53 and extra-cellular matrix–receptor interaction and focal adhesion genes (64 genes; ~8% of all) whereas under-expressed multi-cancer genes are enriched for ErbB signaling, calcium signaling and long-term potentiation genes (45 genes; ~10% of all) (Supplementary Table S8B,C). On the other hand, over-expressed NYTA multi-cancer genes are enriched for endoplasmic reticulum genes (46 genes; ~12% of all) and phosphoproteins (168 genes; ~39% of all) whereas under-expressed NYTA multi-cancer genes are enriched for alternatively spliced genes (132 genes; ~42% of all) and phosphoproteins (124 genes; ~40% of all) (Supplementary Table S8D, E). When the background gene set is set to TA multi-cancer genes rather than the whole genome, NYTA over-expressed multi-cancer genes are enriched only for endoplasmic reticulum genes (Supplementary Table S8F). COGENT detected 46 endoplasmic reticulum NYTA genes as over-expressed in a wide variety of tumors (data not shown). A similar comparison of under-expressed NYTA multi-cancer genes to TA ones did not reveal any significantly enriched functional category. This indicates that endoplasmic reticulum genes are not as studied as their importance in cancer would suggest.

Another important finding of COGENT illustrates the dysregulation of ribosomal proteins. 121 ribosomal protein encoding genes are over-expressed at least in two types of tumors while 92% of these are also under-expressed in at least one tumor. A ribosomal protein gene subset consisting of *RPS16*, *RPL15*, *RPL31*, *RPS15A*, *RPL22* and *RPS6* was under-expressed in more than seven types of tumors, as well as being over-expressed in many tumors (Supplementary Excel file). *RPL18A* is an example of a ribosomal protein that is mainly under-expressed (under-expressed in six types of tumors) while being over-expressed in only two types (astrocytoma grade 2 and 3).

We also observed a trend of over-expressed multi-cancer genes having more orthologs (Figure 3D) whereas under-expressed multi-cancer genes having fewer orthologs (Supplementary Figure S2D). This result is not surprising considering that genes more strictly conserved throughout evolution are more likely to be involved in core-processes affected during tumorigenesis regardless of tissue context.

In addition to identifying multi-cancer genes, we isolated potentially important multi-cancer SAGE tags. This approach revealed SAGE tags that have probably not been included in single studies because users tend to ignore multiple mapping or ambiguously mapping tags. If such an ambiguous SAGE tag is differentially expressed in many types of tumors, identifying that tag's corresponding gene(s) might be worth the effort. For instance, seven over-expressed and 58 under-expressed SAGE tags occur in more than three types of tumors but do not reliably map to a gene by either SAGE Map or SAGE Genie (see detailed statistics in Table 1).

RIDGE analysis

Considering previous findings which indicate co-regulated genomic regions (11) in different cancers, we strove to identify multi-cancer regions of increased co-regulation (RIDGEs) in comparison to RIDGEs associated to variation among normal tissues (NA-RIDGEs). We used the Spearman rank correlation coefficient to compare expression of neighboring gene pairs as the co-regulation metric. To assign members to each RIDGE, we developed an algorithm which we named ICEBERG algorithm (See 'Methods' for detailed explanation). By using a primary $P = 0.002$ and secondary $P = 0.05$; the ICEBERG algorithm identified 81 TA-RIDGEs and 83 NA-RIDGEs. This corresponds to 10.9% of all the genes studied (2242 of 20433) being located in either TA-RIDGEs (1124 genes) or in NA-RIDGEs (1364 genes). A landscape of RIDGEs on two representative chromosomes is presented at Figure 4A. Roughly half of the clusters ($N = 45$) were both TA- and NA-RIDGEs, indicating co-regulation of these regions irrespective of the cancer status of the cell.

To investigate the presence of paralogs in RIDGEs, we defined a 'family gene' as a gene in a RIDGE which has at

Table 1. SAGE tags revealed by COGENT

	Over-expressed tags			Under-expressed tags		
No. of cancer types[a]	Single	2 and 3	>3	Single	2 and 3	>3
Tags with ambigous gene mapping	3558	1522	639	2974	1419	707
Tags which don't map to reliable 3′ends[b]	508	92	7	496	159	58
Tags which don't map to any existing cDNA sequence	19	2	0	23	8	2

[a] Being differentially expressed in this many number of cancer types when compared to corresponding normals.
[b] Tags which map to reliable 3′ends are used in routine SAGE data analysis. SAGE Map's polyadenylated mammalian gene collection (MGC), RefSEQ, mRNA or EST called tags and SAGE Genie's reliable 3′end calls (tags which are present at the most 3′ NlaIII site and come from a transcript with either polyA signal or polyA tail) were used to build reliable 3′ end tag sets.

least one more family member in the same RIDGE, inferring family membership from the similarity of gene symbols. In total 20% of the TA-RIDGEs consist of >50% 'family genes' (see Supplementary Figure S4A for the histogram). On the other hand, 40% of the TA-RIDGEs did not have any 'family genes'. This suggests that only a fraction of RIDGEs can be explained by larger groups of paralogs for which co-regulation might be a consequence of recent gene duplication events. Another useful score, 'normal co-regulation percentage' is the fraction of those genes on TA-RIDGEs which are also co-regulated among normal tissues ($P < 0.05$). According to the 'normal co-regulation percentage' density distribution, 30% of large TA-RIDGEs (RIDGEs containing ten or more genes) and 40% of all TA-RIDGEs display <20% normal co-regulation, i.e. they are tumor specific (See Supplementary Figure S4E–F). Figure 4B presents a selection of RIDGEs that vary in 'family gene' and 'normal co-regulation percentages'. Importantly, RIDGE analysis also revealed that 41% and 83% of the genes on the Y chromosome belong to a TA-RIDGE or an NA-RIDGE, respectively (Figure 4B).

Cancer specific RIDGEs tend to be less paralog dense when compared to global RIDGEs (i.e. TA-RIDGEs overlapping with NA-RIDGEs). In other words, paralog RIDGEs tend to be co-regulated both among normal tissues (NA-RIDGEs) and cancer (TA-RIDGEs) (chi square $P = 0.0284$ when paralog RIDGE threshold = 30% and normal co-regulation percentage = 30%; Supplementary Table S9).

Taken overall, these results suggest a prevalent mechanism of gene regulation in cancer related gene expression in multiple cancers, based solely on the physical location of genes in cancer-specific TA-RIDGEs.

A new subclass of keratin genes, an example finding by zooming on RIDGEs

We present one example finding of RIDGE analysis which points to a subgroup of seven keratin genes [*KRT5*, *KRT6* (A, B and C), *KRT14*, *KRT16* and *KRT17*] which are over-expressed in non-melanoma skin cancers (NMSCs), but are almost completely turned off in melanoma metastasis. This subgroup is over-expressed in squamous cell skin cancer (SCSC), as well as in basal cell skin cancer, but is significantly more downregulated in melanoma metastasis relative to primary melanoma (Figure 5B). In addition, the SPRR and S100A gene families which reside in the EDC are also regulated in approximately the same manner (Figure 5A and B). In fact, of all the genes in the entire EDC, over-expression in NMSCs is confined solely to S100A and SPRR genes (Figure 5A, lower panel).

Other examples and dynamics of RIDGEs

Some paralog TA-RIDGEs are also informative for cancer research. Metallothionein TA-RIDGE (Supplementary Figure S5), Histone 1 cluster TA-RIDGE (Supplementary Figure S6) and MHC Class II TA-RIDGE (Supplementary Figure S7) are differentially expressed in many cancers (see 'Discussion' section for further comments). Furthermore, a clear under-expression is observed in three different types of lymphomas at the MHC TA-RIDGE. Looking at the data from a different perspective, we analyzed the frequency of significant ($P < 0.01$) co-regulations among different cancers using all 81 TA-RIDGEs. As expected, different grades or states of the same cancers [e.g. astrocytoma grade 2 versus astrocytoma grade 3 or chronic phase and accelerated phase of chronic myeloid leukaemia (CML)] were among top co-regulated cancer types, as well as different types of hematological cancers (Supplementary Table S10).

Since most of the co-regulated regions of the genome we point out here are not well characterized, it was not possible to attach a significant functional meaning to the existence of non-paralog TA-RIDGEs. Nevertheless, we observed significant differential expression in tumors within many non-paralog TA-RIDGEs. For instance, Chr8_NAPRT1 (named after the leftmost gene in the RIDGE) at one end of the eighth chromosome is over-expressed in a variety of tumors including epithelial, CNS and hematological (Figure 4B and Supplementary Figure S8).

There is no commonality (i.e. enriched functional classes) among the genes at non-paralog RIDGEs other than high genomic G/C and GC/CG content. Chr8_NAPRT1 is a good example of such a TA-RIDGE with 63% average genomic G/C content that appears as a clear outlier on chromosome 8 (Supplementary Figure S8). Indeed, we found that G/C rich regions are associated with RIDGEs in general. When analyzing G/C content of RIDGEs (whole genomic region including intergenic regions), we used a previously established classification; isochores [(20) and Table 2] to define G/C content classes. TA-RIDGEs are

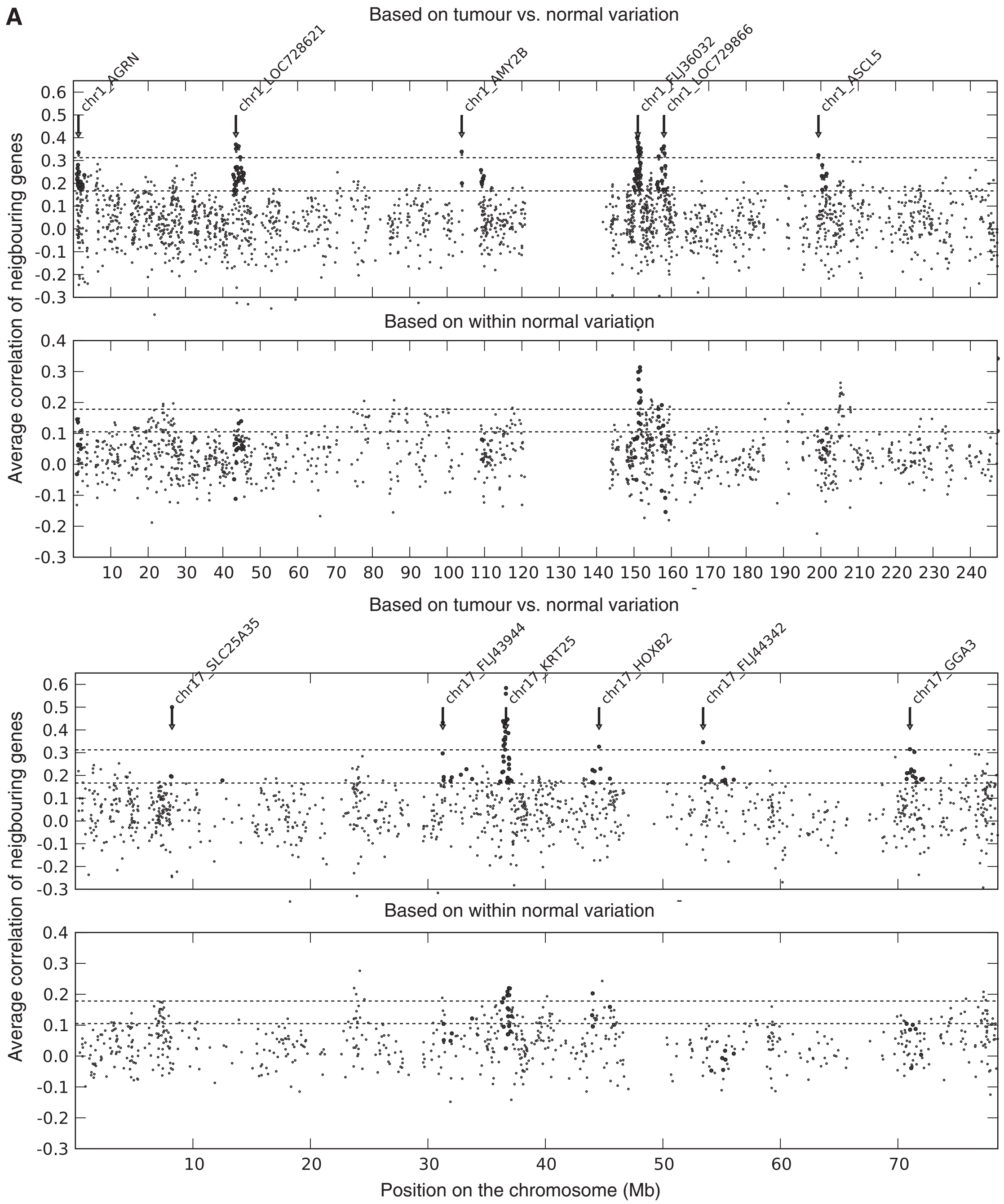


Figure 4. Overview and close up of TA-RIDGEs and NA-RIDGEs by using gene neighborhood correlation score. (**A**) Two representative chromosomes (upper—chr1; lower—chr17) which summarize the overall appearance of TA-RIDGEs and corresponding gene neighborhood correlation scores (average of R_s with ± 21 neighboring genes) with normal variation expression matrix. Darker dots were identified as being member of a TA-RIDGE by the ICEBERG algorithm. The TA-RIDGEs were named by using the chromosome they reside in and the leftmost member of the TA-RIDGE. Dashed lines represent the secondary *P*-value cut-off and primary *P*-value cut-off respectively from low to high in the Y axis. X axis represents position in megabases (MB). (**B**) Six representative TA-RIDGEs. Axes are as in (A). Family gene % (percentage of genes which have at least one family member in the same RIDGE) and normal co-regulation % (the % of genes at a TA-RIDGE that score above the *P* 0.05 threshold at the corresponding region from the normal variation expression matrix) are shown under each TA-RIDGE. Chr5_PCDHAC1, Chr17_Krt25 are examples of family gene dense TA-RIDGEs, which comprise 28% of all TA-RIDGEs. Chr16_HBQ1 is a non-family RIDGE which is common to cancer and normal context. chr7_C7orf28A and chr7_SSPO are two examples of cancer specific TA-RIDGEs, which comprise 40% of all TA-RIDGEs. Almost the entire chromosome Y is co-regulated among normal tissues and also in cancer and identified as one TA-RIDGE: chrY-RPS4Y1.

enriched for H3 class (highest G/C content, G/C > 53%) 3.8-fold relative to non-RIDGE windows (chi-square $P = 0.0129$). More strikingly, non-paralog TA-RIDGEs are enriched for H3 class 7.0-fold relative to non-RIDGE windows (chi-square $P = 0.0001$). Overall, 43% of the TA-RIDGEs correspond to high G/C content genomic regions (defined by H2 and H3 isochores: G/C% > 46%) which corresponds to a significant

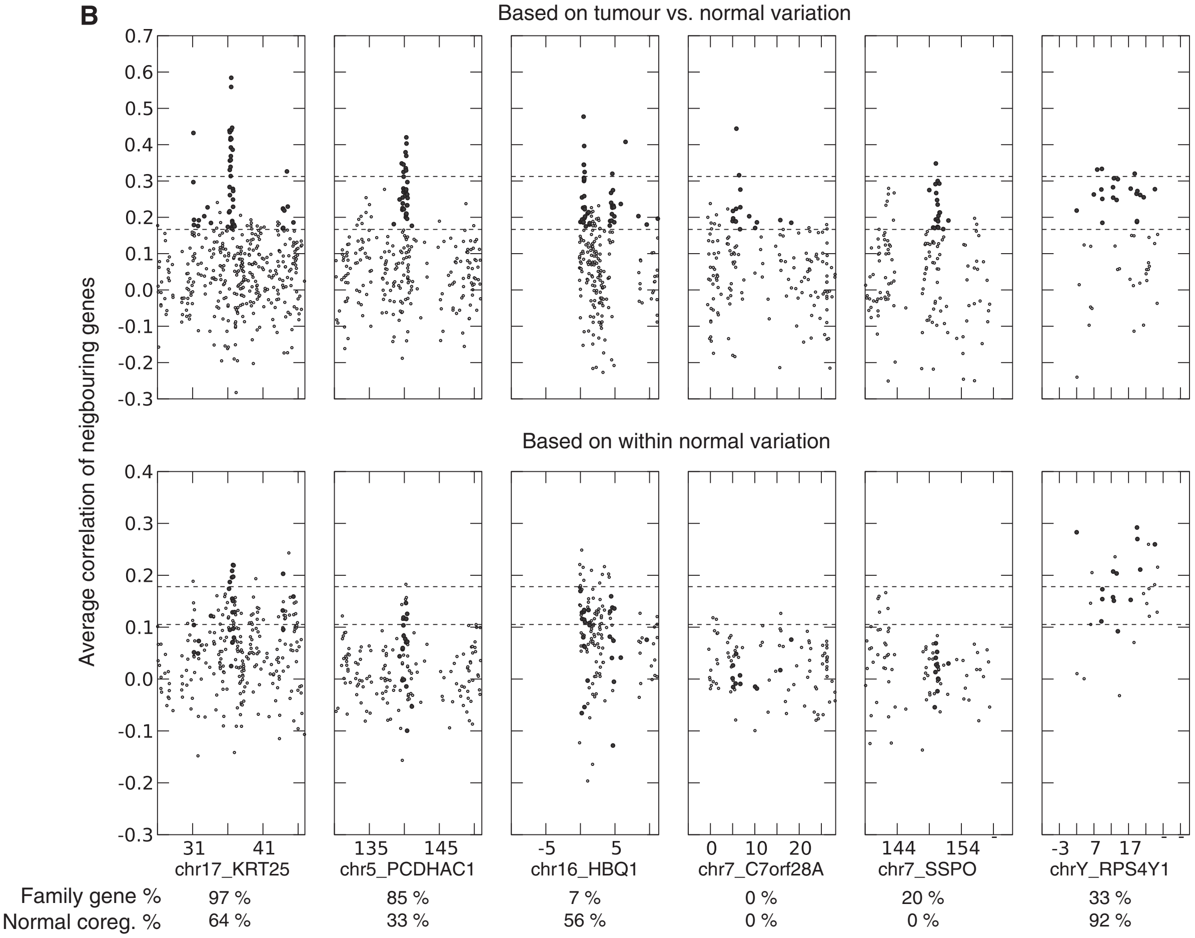


Figure 4. Continued.

1.72-fold enrichment when compared to non-RIDGE regions of varying window sizes (chi-square $P = 0.019$ when compared to non-RIDGEs with window size of 19 and $P = 0.0104$ when compared to non-RIDGEs with window size of 27). A similar enrichment was seen for NA-RIDGEs as well (Table 2). In addition, both the genomic regions and the promoter regions of genes in non-paralog TA-RIDGEs have significantly more GC/CG pairs when compared to either other RIDGEs or non-RIDGE regions (Supplementary Figure S9). Overall, though G/C richness is a general phenomenon that comprises 43% of the TA-RIDGEs and 40% of NA-RIDGEs, non-paralog RIDGEs tend to be more G/C and GC/CG rich than both non-RIDGE regions and other RIDGEs.

Online service for the analysis of multi-cancer genes and multi-cancer RIDGEs

We present a PHP/MySQL based online service in order for the research community to browse and filter the sets of multi-cancer genes and TA-RIDGEs (available at http://www.oncoreveal.org). Users are able to analyze both their own data and COGENT-SAGE, COGENT-microarray & SAGE data compiled from public domain and TA-RIDGEs interconnected with the COGENT browser module (Figure 6). Several filtering, sorting and exporting options are available at the web service, as well as the ability to overlay gene or tag lists. A unique property of oncoreveal is that users can create complex gene signatures by specifying complex queries. For instance, one can filter for all genes that are specifically over-expressed in more than five of glial cell cancers, while not under-expressed in any other cancer. Or, one can easily find cell cycle genes which are over-expressed in melanomas but not other types of skin cancers. Furthermore, users can visualize RIDGEs with the broad chromosome view (as in Figure 4) or the differential expression view (as in Figure 5).

DISCUSSION

In this study, we present a meta-analysis of gene expression in cancer by utilizing both SAGE and microarray data. The majority of significant differences were caught by microarray studies due to their larger sample size (148 139 microarray probesets versus 17 831 SAGE tags), but including SAGE data increased the overall significance of the observations as well as the tumor type coverage. For 13 tumor types, mostly CNS tumors, we had only SAGE data available.

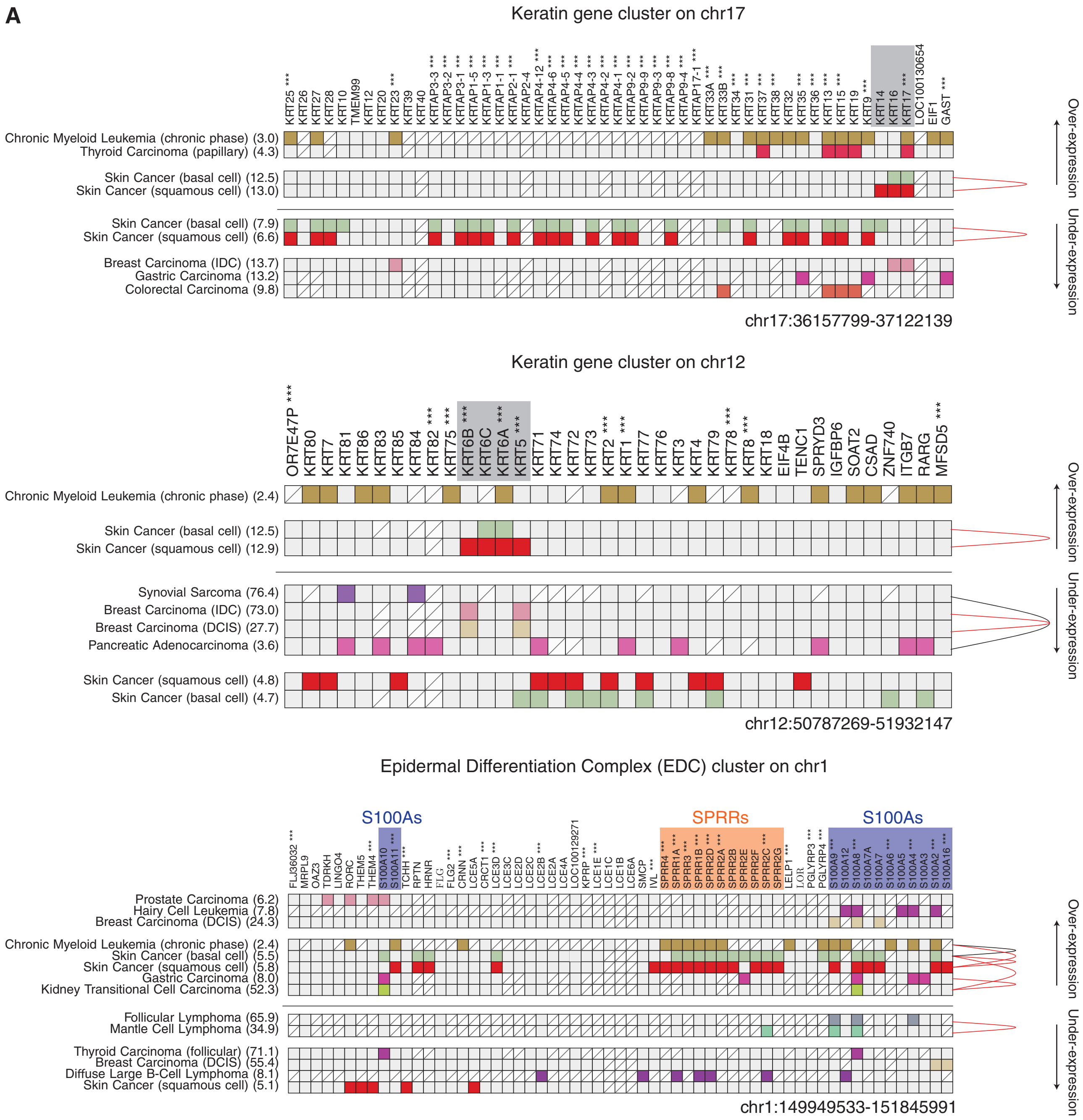


Figure 5. Regulation of Keratin, SPRR and S100A gene families in skin cancer and melanoma metastasis. (**A**) There are two large keratin TA-RIDGEs on the genome on chr12 and chr17 almost all of which are down-regulated in squamous cell and basal cell skin cancer when compared to normal skin. However; KRT5, KRT6A, KRT6B and KRT6C (on chr12, shaded gray) and KRT14, KRT16 and KRT17 (on chr17, shaded gray) are up-regulated in SCSC and four of them are also upregulated in basal cell skin cancer. The TA-RIDGE which corresponds to the EDC that contains SPRRs (shaded red) and S100As (shaded blue) is mostly upregulated in both types of NMSC. All three gene families are over-expressed in chronic phase CML as well. Different colors represent different tumors. Each colored box represents differential expression with the merged COGENT data set (Q < 0.05 for microarray, $P < 0.02$ for SAGE). Light gray boxes represent non-differential expression. Cross-dashed boxes represent missing data. Cancers are sorted by the fold enrichment of differential expression over the expected random differential expression, as explained in Supplementary Data. Fold enrichment values are stated in parentheses beside the tumor names. Only cancers that are significantly (randomization based test: $P < 0.01$) enriched for differential expression are shown. The arcs between rows represent significant co-differential expression events (red arcs: $P < 0.001$, black arcs: $P < 0.002$). Cancers are divided into two by the presence or absence of a significant co-regulation with another cancer. Genes annotated with *** are significantly ($P < 0.05$) co-regulated with the neighbors. Close-ups of RIDGEs with different filters can be visualized at www.oncoreveal.org. (**B**) SPRR genes, S100A genes and Keratin genes which are over-expressed in SCSC (red squares in A) are among the top down-regulated genes in melanoma metastasis samples when compared to primary melanoma samples. Each bar in the bar graph represents the fold change between average expression values of two classes. Error bars represent the average of standard deviations over all possible fold changes between two classes divided by size of all possible comparisons. ***$P < 0.001$ by Mann–Whitney U test. Inset: distribution of fold changes presented in the main graph. Boxes represent inter-quartile range. Whiskers span 1.5 times inter-quartile range.

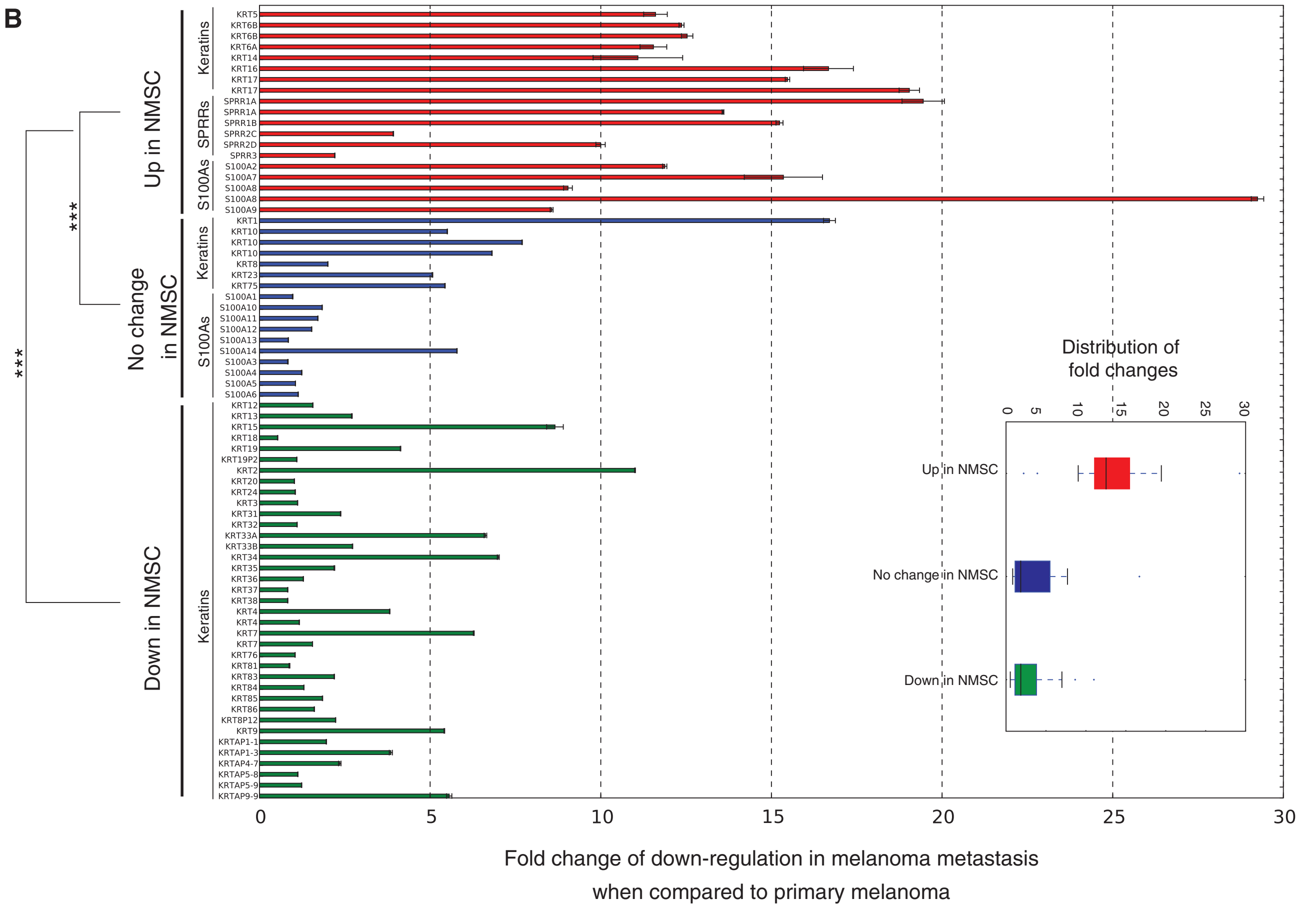


Figure 5. Continued.

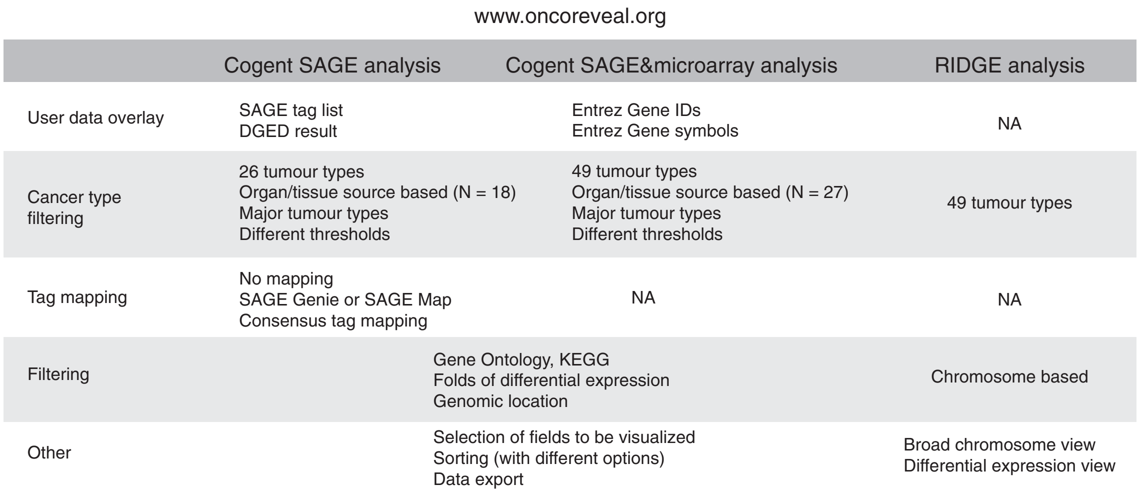
Table 2. GC rich regions are more likely to be co-regulated

	L1	L2	H1	H2	H3
TA-RIDGEs (81)	6.2	21.0	29.6	37.0	6.2
Non-paralog TA-RIDGEs (family gene % < 20) (45)	8.9	22.2	26.7	31.1	11.1
NA-RIDGEs (83)	3.6	27.7	28.9	30.1	9.6
Non-RIDGEs (window size 19) (494)	5.9	29.8	37.0	25.5	1.8
Non-RIDGEs (window size 27) (291)	5.2	33.7	35.7	24.1	1.4

Values show frequency of RIDGEs among all RIDGEs. Letter designation represents isochore classification by G/C content. G/C content is calculated by considering the whole genomic region of the RIDGE. L1 < 37%, 37% ≤ L2 < 41%, 41% ≤ H1 < 46%, 46% ≤ H2 < 53%, H3 ≥ 53%. Numbers in the parentheses stand for number of RIDGEs.

Our findings highlight many multi-cancer genes that have not yet been studied, which is particularly important due to the fact that over-expressed genes often present feasible drug targets and under-expressed multi-cancer genes may represent important gene therapy candidates. These NYTA multi-cancer genes change significantly between cancer and normal tissues, but they rank lower than already discovered multi-cancer genes. At first glance, this might be interpreted to imply that NYTA multi-cancer genes are less important in cancer. This conclusion would be wrong. It is entirely possible that genes with a smaller magnitude of change in expression level with small variance among many cancer types may play as important a role in tumorigenesis as high magnitude changes with small variance. At the same time, the small magnitude of change in expression helps to explain why these genes have not yet been studied as cancer genes.

Enrichment of phosphoproteins and alternatively spliced genes among NYTA multi-cancer genes is informative but not surprising due to the established variability and disturbance of cellular machinery in cancer. COGENT also points to 46 over-expressed NYTA multi-cancer genes which are functionally located in the endoplasmic reticulum. Although the endoplasmic reticulum is not known to be involved in cancer, related ER genes have been directly or indirectly associated to multiple cancer related pathways such as apoptosis or angiogenesis (21). Changes in the endoplasmic reticulum may also contribute to more global effects on cell surface properties important for immune recognition of cancer



www.oncoreveal.org

	Cogent SAGE analysis	Cogent SAGE&microarray analysis	RIDGE analysis
User data overlay	SAGE tag list DGED result	Entrez Gene IDs Entrez Gene symbols	NA
Cancer type filtering	26 tumour types Organ/tissue source based (N = 18) Major tumour types Different thresholds	49 tumour types Organ/tissue source based (N = 27) Major tumour types Different thresholds	49 tumour types
Tag mapping	No mapping SAGE Genie or SAGE Map Consensus tag mapping	NA	NA
Filtering	Gene Ontology, KEGG Folds of differential expression Genomic location		Chromosome based
Other	Selection of fields to be visualized Sorting (with different options) Data export		Broad chromosome view Differential expression view

Figure 6. Online service; www.oncoreveal.org. Users can analyze COGENT SAGE, COGENT microarray & SAGE and TA-RIDGEs by using oncoreveal. Short SAGE tags or DGED results for COGENT–SAGE and Entrez Gene IDs or symbols can be used as input. Users can filter altered genes by cancer type (with any of the three different classifications) or number of cancer types in which change occurred. Several different filtering options from commonly used data sources such as gene ontology and different visualization, sorting and export options are available.

cells. Over-expression of another class, ribosomal genes, is generally attributed to an increased cell proliferation rate as a result of increased metabolic activity in tumors, and generally deemed functionally irrelevant. Viewed in this light, the downregulation of these particular genes may be pointing to an altogether different effect. Although there are several findings which suggest functional relevance for the differential expression of ribosomal proteins, such as regulation by p53, RB, MYC and PTEN (22), COGENT's pointing to occasional under-expression of ribosomal genes among multiple cancer types remains engrossing and intriguing.

Co-regulated regions have been studied in many different contexts including cancer. To our knowledge, ours is the first study that compares TA- and NA-RIDGEs in order to determine cancer-specific TA-RIDGEs. One concern is to accurately assign the boundaries of a RIDGE, which we approached by using the ICEBERG algorithm. Another is that using different expression matrixes will lead to identification of different RIDGEs. For instance, 27 of 81 TA-RIDGEs were also identified as RIDGEs in the study by Stransky *et al.* (11) (at least one gene was commonly identified as being significantly co-regulated). Common RIDGEs included MHC Class II, S100A family and Histone cluster 1. There are several hypotheses that might explain the occurrence of RIDGEs, such as genomic amplifications/deletions, epi-genetic regulations and a regional lack of DNA repair. In the bladder cancer context of Stransky *et al.* (11), genomic amplifications/deletions detected by the comparative genomic hybridization (CGH) method turned out to explain some fraction of the RIDGEs, but not all. In this study, we point to a significant association between RIDGEs and G/C and GC/CG rich regions. This finding is consistent with previous SAGE experiments which showed that G/C rich regions are highly expressed in normal tissues (23). In addition, G/C richness of RIDGEs might reflect several different underlying regulation mechanisms such as chemical fragility, DNA motifs or methylation of DNA through GC/CG dinucleotides.

Keratins are intermediate filament proteins which are well established epithelial cell markers and EDC genes are epidermal differentiation markers. Down-regulation of differentiation markers in a more cancerous state is expected due to stem cell-like state of tumors. However, over-expression in NMSCs when compared to normal skin and differential down-regulation in melanoma metastasis of the genes which are over-expressed in NMSC suggests new functional roles for seven keratin genes of separate functional classes [*KRT5* and *KRT6* are type II epithelial keratins and *KRT14*, *KRT16* and *KRT17* are type I epithelial keratins (24)], SPRRs and the S100A gene family. Keratins are starting to be associated with diseases (24) which suggests new functions such as K17′s effect on cell growth (25) or K8′s and 18′s necessity for melanoma invasion (26). SPRRs are keratinocyte differentiation markers (27) which are known to contribute to the formation of the cornified envelope of skin cells (28). Although SPRRs' role in tumorigenesis is not known, they are over-expressed under stress conditions which might explain their up-regulation in NMSCs (28). One member of the family, SPRR3, acts like a tumor suppressor on esophageal squamous cell carcinoma cells (29). S100 proteins are already associated with cancer, such as S100A2 (30) and S100A4 (31) which have been associated with tumor cell motility and metastasis. A melanoma metastasis to primary melanoma comparison might reflect the differences between real melanoma cells and other skin components due to selection in the metastasis site and probable mixture of cell types at the primary source. However, this probability does not decrease the importance of the presented co-regulation of different gene families. At most it might affect further interpretation of this finding, if true.

MHC Class II TA-RIDGE resides on chromosome VI. Although the view on MHC Class II proteins is that they are generally not expressed on tumor cells, they are known to be over-expressed in several tumors (32) such as gliomas, and some tumor cells can be recognized by CD4+ T cells due to the presence of MHC Class II based antigen presentation (33). Furthermore, to the best of our knowledge, the under-expression of MHC Class II proteins in lymphomas is not an established fact, in spite of supporting evidence such as lack of detection of HLA-DR in diffuse large cell lymphoma patients as a sign of shorter survival (34). The histone RIDGE also resides on the same chromosome. There is a considerable amount of data about the effect of post-translational modifications of histone proteins such as deacetylation or methylation which in turn epi-genetically regulate gene transcription. However, there are few examples of histone proteins which are related to cancer directly (35). Indeed, none of the histone genes show a cancer geneRIF and the majority of the histone genes do not have any geneRIFs at all (38 of 47 histone genes). Histone proteins are expressed from four different clusters in the human genome (Histone clusters 1–4). In total, 66 histone genes reside at histone cluster 1. Every class of five different histone proteins (H1-H2A, H2B, H3 and H4) is expressed from this cluster. In histone cluster 1, mainly histone 1 (e.g. HIST1H1T) and histone 2 (e.g. HIST1H2BC, HIST1H2AC and HIST1H2BD) genes are over-expressed in cancer tissues (Supplementary Figure S6). The significant alteration of histones in multiple tumors is likely to be functionally relevant in tumorigenesis.

Approximately half of the RIDGEs did not arise from paralog genes. One possibility is that the differential expression events at these RIDGEs are just side-effects of a previously important key tumorigenic event. Another possibility is that only some of the genes within these RIDGEs contribute to the key processes during tumorigenic transformation. For instance, *TNF* is co-regulated with its neighbors (chr6_CDSN). Some other TA genes within the non-paralog RIDGEs with most cancer geneRIFs are *RHOA*, *MAPK3*, *CYP2E1*, *RELA*, *GJA1*, *DAPK1*, *CCKBR*, *FASN*, *GPX1* and *HSPB1*. It might be possible that if a gene is being regulated together with a well known cancer gene, it is also functionally important in tumorigenesis.

Oncoreveal has several novel utilities when compared to similar existing tools such as Oncomine. Oncoreveal allows an unlimited size of user gene/tag list overlays. It allows SAGE data analysis in addition to the SAGE-microarray merged data analysis by allowing users to use or omit consensus tag mapping. Moreover, oncoreveal enables RIDGE analysis by offering basic filtering options as a starting point as well as detailed visualization options. In our opinion, having most of the data presented in this study available, browsable and filterable online makes this data as informative as possible and is likely to lead to new discoveries which we have not had an opportunity to reveal with our analysis.

In conclusion, our study revealed a set of previously unstudied multi-cancer genes that are differentially expressed in as many tumors as established cancer genes. We also investigated TA-RIDGEs in comparison with NA-RIDGEs which appeared to coincide with GC rich regions of the human genome. In addition to presenting several examples of RIDGE analysis, we present an online tool, oncoreveal, for researchers to analyze multi-cancer genes/tags and TA-RIDGEs. We believe that, especially with the availability of new generation high-throughput methods, analysis of RIDGEs, multi-cancer genes and multi-cancer gene signatures is now more feasible and will add considerably to our knowledge about tumorigenesis, which will eventually lead to efficient diagnosis and treatment.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Dr Rickard Sandberg for fruitful discussions. In addition, the authors thank Prof. Mustafa Uzan regarding help with the collection of brain tissues. Finally, the authors would like to thank anonymous reviewers for their valuable comments.

FUNDING

Turkey's council of scientific research (TUBITAK, grant numbers 104T027, 105T301 and 106T153); Scientific Research Projects (BAP); The Swedish Cancer Society; The Swedish Research Council and The Cancer Society in Stockholm. Funding for open access charge: Bogazici University.

Conflict of interest statement. None declared.

REFERENCES

1. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
2. Morrissy,A.S., Morin,R.D., Delaney,A., Zeng,T., McDonald,H., Jones,S., Zhao,Y., Hirst,M. and Marra,M.A. (2009) Next-generation tag sequencing for cancer gene expression profiling. *Genome Res.*, **19**, 1825–1835.
3. Rhodes,D.R., Yu,J., Shanker,K., Deshpande,N., Varambally,R., Ghosh,D., Barrette,T., Pandey,A. and Chinnaiyan,A.M. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl Acad. Sci. USA*, **101**, 9309–9314.
4. Xu,L., Geman,D. and Winslow,R.L. (2007) Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics*, **8**, 275.
5. Bueno-de-Mesquita,J.M., van Harten,W.H., Retel,V.P., van't Veer,L.J., van Dam,F.S., Karsenberg,K., Douma,K.F., van Tinteren,H., Peterse,J.L., Wesseling,J. *et al.* (2007) Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER). *Lancet Oncol.*, **8**, 1079–1087.
6. Axelsen,J.B., Lotem,J., Sachs,L. and Domany,E. (2007) Genes overexpressed in different human solid cancers exhibit different tissue-specific expression profiles. *Proc. Natl Acad. Sci. USA*, **104**, 13122–13127.

7. Rhodes,D.R., Kalyana-Sundaram,S., Mahavisno,V., Barrette,T.R., Ghosh,D. and Chinnaiyan,A.M. (2005) Mining for regulatory programs in the cancer transcriptome. *Nat. Genet.*, **37**, 579–583.
8. Cohen,B.A., Mitra,R.D., Hughes,J.D. and Church,G.M. (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.*, **26**, 183–186.
9. Caron,H., van Schaik,B., van der Mee,M., Baas,F., Riggins,G., van Sluis,P., Hermus,M.C., van Asperen,R., Boon,K., Voute,P.A. *et al.* (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, **291**, 1289–1292.
10. Spellman,P.T. and Rubin,G.M. (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. *J. Biol.*, **1**, 5.
11. Stransky,N., Vallot,C., Reyal,F., Bernard-Pierrot,I., de Medina,S.G., Segraves,R., de Rycke,Y., Elvin,P., Cassidy,A., Spraggon,C. *et al.* (2006) Regional copy number-independent deregulation of transcription in cancer. *Nat. Genet.*, **38**, 1386–1396.
12. Volz,A., Korge,B.P., Compton,J.G., Ziegler,A., Steinert,P.M. and Mischke,D. (1993) Physical mapping of a functional cluster of epidermal differentiation genes on chromosome 1q21. *Genomics*, **18**, 92–99.
13. Hochberg,Y. and Benjamini,Y. (1990) More powerful procedures for multiple significance testing. *Stat. Med.*, **9**, 811–818.
14. Boon,K., Osorio,E.C., Greenhut,S.F., Schaefer,C.F., Shoemaker,J., Polyak,K., Morin,P.J., Buetow,K.H., Strausberg,R.L., De Souza,S.J. *et al.* (2002) An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11287–11292.
15. Lal,A., Lash,A.E., Altschul,S.F., Velculescu,V., Zhang,L., McLendon,R.E., Marra,M.A., Prange,C., Morin,P.J., Polyak,K. *et al.* (1999) A public database for gene expression in human cancers. *Cancer Res.*, **59**, 5403–5407.
16. Ge,X., Yamamoto,S., Tsutsumi,S., Midorikawa,Y., Ihara,S., Wang,S.M. and Aburatani,H. (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, **86**, 127–141.
17. Xu,L., Shen,S.S., Hoshida,Y., Subramanian,A., Ross,K., Brunet,J.P., Wagner,S.N., Ramaswamy,S., Mesirov,J.P. and Hynes,R.O. (2008) Gene expression changes in an animal melanoma model correlate with aggressiveness of human melanoma metastases. *Mol. Cancer Res.*, **6**, 760–769.
18. Mitchell,J.A., Aronson,A.R., Mork,J.G., Folk,L.G., Humphrey,S.M. and Ward,J.M. (2003) Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annu. Symp. Proc.*, 460–464.
19. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
20. Costantini,M., Clay,O., Auletta,F. and Bernardi,G. (2006) An isochore map of human chromosomes. *Genome Res.*, **16**, 536–541.
21. Moenner,M., Pluquet,O., Bouchecareilh,M. and Chevet,E. (2007) Integrated endoplasmic reticulum stress responses in cancer. *Cancer Res.*, **67**, 10631–10634.
22. Ruggero,D. and Pandolfi,P.P. (2003) Does the ribosome translate cancer? *Nat. Rev. Cancer*, **3**, 179–192.
23. Versteeg,R., van Schaik,B.D., van Batenburg,M.F., Roos,M., Monajemi,R., Caron,H., Bussemaker,H.J. and van Kampen,A.H. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.*, **13**, 1998–2004.
24. Gu,L.H. and Coulombe,P.A. (2007) Keratin function in skin epithelia: a broadening palette with surprising shades. *Curr. Opin. Cell Biol.*, **19**, 13–23.
25. Kim,S., Wong,P. and Coulombe,P.A. (2006) A keratin cytoskeletal protein regulates protein synthesis and epithelial cell growth. *Nature*, **441**, 362–365.
26. Hendrix,M.J., Seftor,E.A., Seftor,R.E., Gardner,L.M., Boldt,H.C., Meyer,M., Pe'er,J. and Folberg,R. (1998) Biologic determinants of uveal melanoma metastatic phenotype: role of intermediate filaments as predictive markers. *Lab. Invest.*, **78**, 153–163.
27. Gibbs,S., Fijneman,R., Wiegant,J., van Kessel,A.G., van De Putte,P. and Backendorf,C. (1993) Molecular characterization and evolution of the SPRR family of keratinocyte differentiation markers encoding small proline-rich proteins. *Genomics*, **16**, 630–637.
28. Martin,N., Patel,S. and Segre,J.A. (2004) Long-range comparison of human and mouse Sprr loci to identify conserved noncoding sequences involved in coordinate regulation. *Genome Res.*, **14**, 2430–2438.
29. Zhang,Y., Feng,Y.B., Shen,X.M., Chen,B.S., Du,X.L., Luo,M.L., Cai,Y., Han,Y.L., Xu,X., Zhan,Q.M. *et al.* (2008) Exogenous expression of Esophagin/SPRR3 attenuates the tumorigenicity of esophageal squamous cell carcinoma cells via promoting apoptosis. *Int. J. Cancer*, **122**, 260–266.
30. Diederichs,S., Bulk,E., Steffen,B., Ji,P., Tickenbrock,L., Lang,K., Zanker,K.S., Metzger,R., Schneider,P.M., Gerke,V. *et al.* (2004) S100 family members and trypsinogens are predictors of distant metastasis and survival in early-stage non-small cell lung cancer. *Cancer Res.*, **64**, 5564–5569.
31. Tarabykina,S., Griffiths,T.R., Tulchinsky,E., Mellon,J.K., Bronstein,I.B. and Kriajevska,M. (2007) Metastasis-associated protein S100A4: spotlight on its role in cell migration. *Curr. Cancer Drug Targets*, **7**, 217–228.
32. Marsman,M., Jordens,I., Griekspoor,A. and Neefjes,J. (2005) Chaperoning antigen presentation by MHC Class II molecules and their role in oncogenesis. In Vande Woude,G.F. and Klein,G. (eds), *Advances in Cancer Research*, Vol. 93. Academic Press, Oxford, UK, pp. 129–158.
33. King,J., Waxman,J. and Stauss,H. (2008) Advances in tumour immunotherapy. *QJM*, **101**, 675–683.
34. Miller,T.P., Lippman,S.M., Spier,C.M., Slymen,D.J. and Grogan,T.M. (1988) HLA-DR (Ia) immune phenotype predicts outcome for patients with diffuse large cell lymphoma. *J. Clin. Invest.*, **82**, 370–372.
35. Ho,C.C., Cheng,C.C., Liu,Y.H., Pei,R.J., Hsu,Y.H., Yeh,K.T., Ho,L.C., Tsai,M.C. and Lai,Y.S. (2008) Possible relation between histone 3 and cytokeratin 18 in human hepatocellular carcinoma. *In Vivo*, **22**, 457–462.