CANCER GENOMICS & PROTEOMICS *11*: 13-24 (2014)

Cilia Gene Expression Patterns in Cancer

MAX SHPAK[1,2], MARCUS M. GOLDBERG[2] and MATTHEW C. COWPERTHWAITE[3]

[1]*NeuroTexas Institute, St. David's Health Care, Austin, TX, U.S.A.;*
[2]*Center for Systems and Synthetic Biology, University of Texas, Austin, TX, U.S.A.;*
[3]*Texas Advanced Computing Center, Austin, TX, U.S.A.*

Abstract. *Non-motile cilia are thought to be important determinants of the progression of many types of cancers. Our goal was to identify patterns of cilia gene dysregulation in eight cancer types (glioblastoma multiforme, colon adenocarcinoma, breast adenocarcinoma, kidney renal clear cell carcinoma, lung squamous cell carcinoma, lung adenocarcinoma, rectal adenocarcinoma, and ovarian cancer) profiled by The Cancer Genome Atlas. Among these types, 2.5-19.8% of cilia-associated genes were significantly differentially expressed (versus 5.5-32.4% dysregulation across all genes). In four cancer types (breast adenocarcinoma, colon adenocarcinoma, glioblastoma multiforme, and ovarian cancer), cilia genes were on average down-regulated (median fold change from –1.53–0.3), in the other four types, cilia genes were up-regulated (fold change=0.86-3.5). Pairwise comparisons between cancer types revealed varying degrees of similarity in the differentially expressed cilia genes, ranging from 7.1% (ovarian cancer and lung squamous cell carcinoma) to 65.8% (ovarian cancer and rectal adenocarcinoma). Hierarchical clustering and principal components analysis of gene expression identified glioblastoma multiforme, colon adenocarcinoma, breast adenocarcinoma; and kidney renal clear cell carcinoma, lung squamous cell carcinoma, lung adenocarcinoma, rectal adenocarcinoma, and ovarian cancer as sub-classes with similar dysregulation patterns. Our study suggests that genes involved in cilia biosynthesis and function are frequently dysregulated in cancer, and may be useful for identifying and classifying cancer types.*

Correspondence to: Max Shpak, NeuroTexas Institute, St. David's Healthcare, 1015 E, 32nd St. Suite 404, Austin, TX 78705, U.S.A. Tel: +1 5125448077, e-mail: shpak.max@gmail.com and Marcus Goldberg, Center for Systems and Synthetic Biology, University of Texas at Austin, 2500 Speedway, Austin TX 78712, U.S.A. E-mail: marcusmgoldberg@gmail.com

Key Words: Cancer, primary cilia, gene expression, patterns.

Cilia are organelles present on the surface of the majority of human cell types. There are two classes of cilia: motile and non-motile. Non-motile (primary) cilia are generally sensory organelles that are involved in signal transduction, response to chemicals in the external environment, and cellular growth and differentiation (1). Cilia also play an important developmental role in tissue and organ patterning, including cell adhesion/communication in the brain and the heart. Unsurprisingly, primary cilia are associated with a number of human diseases, or ciliopathies, such as polycystic kidney disease, Bardet–Biedel syndrome, and Meckel–Gruber syndrome (2).

Primary cilia have also been implicated in cancer, and tumorigenesis in particular, as a consequence of the involvement of primary cilia in cell-cycle regulation. At the molecular level, the most studied system is ciliary regulation of the hedgehog (Hh) signaling pathway, which has been shown to be abnormally activated in many different types of cancers (3-6) The presence of primary cilia can either increase or reduce tumorigenesis and cancer progression, depending on whether early oncogenic events occurred upstream or downstream of Hh activity, respectively. This dual role for primary cilia has major therapeutic implications because ciliogenesis inhibitors may enhance or reduce tumor growth depending on the type of cancer (7).

Changes in primary cilia are also observable at the histological and anatomical level. Specifically, the loss of primary cilia has been observed in a wide range of cancer types and has been shown to be associated with greater progression and poorer prognosis (6-11). Some types of cancer, such as basal cell carcinoma and medulloblastoma, retain their cilia (7), while others, such as breast and pancreatic cancer, lose their cilia (6, 12). These studies have suggested that primary cilia loss is not simply driven by increased replication rates of the cancer cells. Rather, cilia loss likely occurred through somatic mutation or other causes of dysregulation in the cancer genome. Further work is required to establish a causal link between cilia and cancer, since it is unclear whether changes in cilia genes are key to carcinogenesis, or simply a general consequence of gene dysregulation or enhanced mutation rates.

One way to investigate this question is to ask whether cilia gene dysregulation patterns are congruent with overall patterns of gene dysregulation in cancer. As there have been no studies that comprehensively examined dysregulation of ciliary gene expression in cancer, we focus specifically on the expression patterns of ciliary genes in eight common types of human cancer. The principal goals of the present study were a) to identify differentially expressed cilia genes in each cancer type, b) to use patterns of dysregulation across cilia genes as a criterion for classifying cancer, and c) to compare patterns of dysregulation in cilia genes to overall patterns of dysregulation across all genes.

Materials and Methods

Gene expression dataset. Gene expression data were collected from The Cancer Gene Atlas (TCGA, https://tcga-data.nci.nih.gov/tcga/) for eight different cancer types: invasive breast carcinoma (BRCA), colon adenocarcinoma (COAD), glioblastoma multiformae (GBM), kidney renal clear-cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), rectal adenocarcinoma (READ). This study only included cancer for which Level 2 expression data from Agilent custom gene expression microarrays was available for both tumor and control samples (healthy tissue samples from the same organ as the tumor, typically from different individuals rather than as paired samples). The Agilent G4502A custom microarray contains 244,000 probes that assay the expression of 21,232 genes. Table I summarizes the types and number of samples of each cancer that were included in the present study.

Table I. *A summary of cancer data sets from listed by cancer type, where the numbers in the last column are the number of tumor and control samples, respectively.*

Abbreviation	Cancer	n, (Tumor/Controls)
BRCA	Breast invasive carcinoma	431/35
COAD	Colon adenocarcinoma	169/7
GBM	Glioblastoma multiformae	517/19
KIRC	Kidney renal clear cell carcinoma	70/2
LUAD	Lung Adenocarcinoma	32/2
LUSC	Lung squamous cell carcinoma	135/5
OV	Ovarian serous cystadenocarcinoma	585/23
READ	Rectum adenocarcinoma	69/6

Ciliome dataset. In order to survey a broad set of genes associated with cilia structure and function, we obtained version 3.0 of the ciliary proteome database (freely available at http://v3.ciliaproteome.org/cgi-bin/index.php) (6, 12). This database integrates information from several lines of evidence including comparative genomics, mass spectroscopy and transcriptome analyses to produce a comprehensive data set of genes associated with cilia biosynthesis, regulation or function. We selected the reciprocal best hits (RBH) database, which only contains the proteins that were identified as RBHs *via* BLAST in the human genome and in the original species in which the gene was identified. To filter the RBH database, we used the default cut-off value of 1e-10 without any additional filters. This approach yielded a dataset of 1,920 ciliary proteins that overlap with the gene set covered in the expression probes; we note that since there is no direct experimental verification of cilia expression or function for all genes considered in this study, this is probably an overestimate of ciliome size (that is, it is unlikely that cilia genes make up over 9% of the human genome).

Further characterization and classification of the ciliary proteins was obtained from the DAVID functional annotation and bioinformatics resource (http://david.abcc.ncifcrf.gov/gene2gene.jsp) (13). In order to determine whether dysregulated cilia genes are characterized by particular structural motifs and/or by their roles in specific functional pathways, enrichment analysis was applied to the gene set. DAVID's enrichment analysis tool combines information from multiple gene annotation ontologies (*e.g.* Gene Ontology: GO and Kyoto Encyclopedia of Genes and Genomes: KEGG) to identify clusters of genes with shared functional or structural properties, and determines whether a data set is enriched in that functional class relative to background frequencies (across the entire genome) using a Fisher's exact test. DAVID computes an enrichment score for the intersection among subclasses of genes characterized by a given function. The score of the intersection among these subclasses is computed from the geometric mean of the enrichment score within each subclass (see (14, 15)).

Statistical analysis. Apart from the DAVID enrichment analysis, all statistical analyses were performed in R (version 2.14) using the Bioconductor package (16), (http://www.bioconductor.org/). Level 2 gene-expression data were pre-processed by the TCGA, and included LOESS normalization (17), and transformation of expression levels onto a log base 2 scale. We independently fit linear models (18) to each set of transformed gene-expression measurements to compute standard error test statistics for contrasts between gene expression levels in tumor *versus* control samples. Empirical Bayes factors (posterior statistics derived from moderated t-distributions) were used to determine the significance of the difference in expression level of each gene in cancer *versus* control samples. The p-values derived from the empirical Bayes factors were adjusted for multiple comparison (over 21,232 genes) using the Benjamini–Hochberg false-discovery rate (FDR) correction. Significantly differentially expressed genes were defined as those with both a minimum two-fold change in expression level between cancers and controls, and a significance threshold of α=0.001 after the FDR correction to the *p*-values.

Those genes which showed significant dysregulation across different cancer types were selected to construct fold-change data matrices. Each gene was characterized by the difference between the mean value of its expression level across samples within a specific cancer and its mean expression level among the control samples. Pairwise Euclidean distances of these vectors of fold change values across cancers were computed, and hierarchical clustering was used to identify cancers with similar patterns of regulation among cilia genes and similarities in expression profiles.

Principal component analysis (PCA) was applied to the entire set of expression levels (*i.e.* all genes in the data set, including both those that are significantly dysregulated and those that are not) in order to characterize general patterns of covariation across fold change in expression levels, and to contextualize specific cancer types in gene space and cilia genes of interest in the cancer space. Because there are a total of 21,232 variables (genes) and only eight

observations (types of cancer), R-mode principal component vectors were computed using a singular value decomposition of the matrix of fold change in expression levels in order to reduce the dimensionality of gene space to eight independent vectors. Each cancer could then be quantified by its transformed score in this eight dimensional expression space. In contrast, Q-mode principal components were used to define gene expression levels in cancer space, and were computed directly from covariances in expression levels across cancer types.

Loadings of gene expression levels onto the first Q-mode principal component eigenvectors were obtained by computing the projections (inner products) of gene fold-change values across cancer types onto the first eigenvector. These loadings were used as a heuristic to capture the degree to which expression patterns of individual genes are concordant with the underlying global covariances in expression patterns across the majority of genes. K-means clustering (19) was carried out on cancers in the space of the first two R-mode principal component axes in order to locate cancer types in the reduced space of gene expression levels.

The first Q-mode principal component was also exploited to adjust the fold-change values in the expression data matrices. Subtracting the loadings of each gene onto the first Q-mode principal component allows one to distinguish expression patterns in specific cilia genes from overall statistical trends across the entire data set. Specifically, we used adjusted fold changes:

$$r'_{ij} = r_{ij} - \rho v_i \qquad \text{(Eq. 1)}$$

where r_{ij} is the fold change of probe j in cancer sample i, v_i is the ith element of the first principal component v (in Q-mode sample space), and ϱ is defined by the projection:

$$\rho = \sum_k r_{ik} v_k \Big/ \sum_k {v_k}^2 \qquad \text{(Eq. 2)}$$

This form of adjustment is analogous to the approach used to correct for allelic stratification data in (20), although here it is used to identify variation in patterns of gene expression that are not congruent with those across the genome as a whole, as opposed to co-variation in genotype or phenotype due to genealogical relatedness.

Hierarchical clustering of the eight cancer types was then carried out using Euclidean distances of the vectors of adjusted fold change expression values r_{ij}.

Results

Global gene expression patterns. We conducted a comprehensive comparison of differential gene expression across all cancer types surveyed, which is summarized in Table II. The proportion of differentially expressed genes ranged from 5.6% (LUAD) to 32.7% (OV). The differentially expressed genes were (on average) up-regulated with respect to controls in five cancer types (COAD, KIRC, LUAD, LUSC, READ) and down-regulated in three (BRCA, GBM, OV). The number of shared differentially expressed cilia genes was investigated among pairs of cancer types, and we found that the overlap in such genes across cancer types ranged from as high as 68.8% (in LUAD and READ, relative to LUAD) to as low as 7.3% (in LUAD and OV, relative to OV). The average proportion of dysregulated genes shared between any cancer pair was 31.8%.

The differentially expressed genes can be further classified according to the number of cancer types in which they are significantly dysregulated. Out of the 21,232 genes surveyed, the majority (13,660) were dysregulated in at least one cancer type. Considering only those that were differentially expressed in the majority (>4) of the cancer types, we found that there were 31 genes dysregulated across all eight cancer types surveyed, 315 shared across at least seven out of the eight (not necessarily the same seven), 543 shared across at least six, and 1,068 shared genes among at least five of the eight cancer types.

Hierarchical clustering of the cancer types by Euclidean distance in fold changes across all probes is shown in Supplementary Figure S1 (Supplementary data are available from the first author upon request), which identifies two subclusters: COAD, BRCA, GBM form a cluster distinct from the remaining five cancer types. The two subsets largely, but not entirely, correspond to cancer types where most differentially expressed genes are up-regulated *versus* those types in which they are down-regulated. The two counter examples to this pattern are COAD and OV. The differentially expressed genes in COAD are on average up-regulated, yet COAD clusters with BRCA and GBM (whose differentially expressed genes tend to be down-regulated). In contrast, the differentially expressed genes in OV are on average up-regulated, yet its expression profile clusters with cancer types whose genes are down-regulated on average. This indicates that the average direction of dysregulation is not necessarily a reliable indicator of overall similarity of gene expression profiles within a cancer type.

Ciliary gene expression patterns. In order to characterize gene dysregulation in the ciliome, we identified cilia genes among the global set of differentially expressed genes found in all cancer types (Table II). Genes identified by the cilia proteome database constitute 9.04% of the total number of genes surveyed in the study. In contrast, we observed that cilia genes constitute between 3.8% (LUSC) to 8.7% (KIRC) of all differentially expressed genes among the cancer types, with a mean proportion of 5.2% across all cancer types in the survey.

In all cancer types except COAD, the mean fold change in cilia gene expression was consistently in the same direction (up-regulation or down-regulation) as the mean fold change across all expressed genes. In contrast, the mean fold change among all differentially expressed genes in COAD was 1.15 (slightly more than two-fold up-regulation), while for the cilia genes, the mean fold change was –1.53.

Table II. *A tally and summary statistics for differentially expressed genes for each cancer type. The second and fourth columns are the total number of genes that show significant dysregulation among the entire set and among cilia genes, respectively. The third and fifth columns are the median and mean (in parentheses) fold change between expression levels in the cancer and control samples for the total data set (third column) and the ciliary set (fifth).*

Cancer type	Differential expression, n	FC, median/mean (SD)	Ciliary differential expression, n	Ciliary FC, median/mean (SD)
BRCA	4456	–0.95/-0.08 (1.3)	231	–0.93/–0.22 (1.1)
COAD	4240	1.15/0.23 (2.5)	186	–1.53/–0.26 (2.5)
GBM	4027	–0.95/–0.16 (1.6)	225	–0.30/–0.25 (1.5)
KIRC	2117	3.20/3.26 (1.2)	185	2.93/3.19 (1.2)
LUAD	1191	3.30/3.40 (1.0)	48	3.55/3.51 (0.9)
LUSC	2280	2.62/2.30 (1.7)	86	2.44/2.17 (1.6)
OV	6949	–1.18/–0.29 (1.5)	354	–1.22/–0.56 (1.3)
READ	4231	1.88/0.85 (2.4)	197	1.77/0.86 (2.23)

Table III. *List of the cilia genes that are dysregulated in precisely 7, 6, and 5 of the 8 cancer types, respectively.*

Number of shared cancer types	Genes
7	*FMN2, MYH11, TMOD1*
6	*CENPE, CALU, CNN3, EIF3D, GANAB, FEN1, MPP2, NUDC, PDXK, RAB6A, STX1A, TMCC2*
5	*ABCF2, ADRM1, BCAT1, CNDP2, COG2, CSPG4, DGKI, DMXL2, DOCK7, EXOC7, FLNA, FXR1, GNAZ, GSTT2, ITSN1, MMAB, MYO5A, NOMO1, PA2G4, PADI2, RAB3A, RAB9A, SBDS, TMED4, TNPO1, WDR66, XPO1*

Further annotation of differentially expressed cilia genes using DAVID's functional clustering tool revealed enrichment of cilia genes in several pathways and structural roles. Specifically, among genes that were dysregulated in at least five out of eight cancer types, we found significant enrichment scores ($p<0.05$ following FDR correction) for those characterized by phospho-binding loop motifs. Smaller subsets of dysregulated cilia genes had significant enrichment scores identified with acetylation, endomembrane system, protein transport, phosphoprotein, and protein localization.

Similarity and variation in cilia gene expression patterns among cancer types. A comparatively limited subset of cilia genes (42) are dysregulated over a more than half of the cancer types in the survey. Table III lists genes according to the number of cancer types in which they are significantly dysregulated: 3 cilia genes are differentially expressed in seven of the cancer types, 12 differentially expressed in (exactly) six, and 27 differentially expressed in five. We remark that while there are no cilia genes dysregulated in all eight of the cancer types when expression levels are filtered by both a minimum two-fold change and $p<0.001$ as tests for significance, two genes (*MYH11* and *FMN2*) are dysregulated in all eight cancer types if significant dysregulation is determined solely by the p-value condition.

To contrast the shared pairwise similarity in ciliome dysregulation with shared similarity over all genes, we computed an enrichment score, a ratio that measures the fraction of cilia genes in cancer type i that are also dysregulated in type j with respect to the fraction of all shared dysregulated genes between the two cancer types. Specifically, let N_i and N_j be the total number of dysregulated genes in cancer types i and j, respectively, and n_i, n_j be the number of dysregulated cilia genes in cancer types i, j. We similarly defined N_{ij} and n_{ij} as the respective number of total dysregulated genes and the number of dysregulated cilia genes shared between i, j. The enrichment value of shared cilia genes E_{ij} is defined as an odds ratio, n_{ij}/n_i divided by N_{ij}/N_i, *i.e.*

$$E_{ij} = \frac{\frac{n_{ij}}{n_i}}{\frac{N_{ij}}{N_i}} = \frac{n_{ij}N_i}{n_i N_{ij}} \qquad \text{(Eq. 3)}$$

For example, BRCA and COAD share 60 dysregulated cilia genes, which make up 32.2% of the total number of dysregulated cilia genes (186) in COAD. Similarly, BRCA and COAD share 1,101 dysregulated genes overall, this intersecting set makes up 26% of the total set of 4,240 genes dysregulated in BRCA. The ratio of 32.2 to 26% gives an enrichment value of 1.24 for BRCA against COAD. An

Table IV. *Pairwise comparison of cilia genes differentially expressed in all cancer types. Each row i and column j represents a cancer type. The first entry in block i, j is the total number of cilia genes that are significantly dysregulated in both cancer i and j. The value in parentheses is the enrichment value Eij (as defined in equation 3), which compares the fraction of dysregulated cilia genes in cancer i that are also dysregulated in j to the fraction of dysregulated genes found overall in cancer i that are also dysregulated in cancer j.*

Cancer type	BRCA	COAD	GBM	KIRC	LUAD	LUSC	OV	READ
BRCA	×	60 (1.050)	75 (1.402)	33 (1.238)	11 (0.7633)	29 (0.9449)	116 (1.176)	62 (1.042)
COAD	60 (1.24)	×	64 (1.066)	32 (0.8423)	27 (0.9145)	35 (0.7563)	90 (1.050)	94 (1.012)
GBM	75 (1.30)	64 (0.8373)	×	26 (0.6884)	20 (0.8117)	32 (0.7771)	111 (1.129)	65 (1.001)
KIRC	33 (1.346)	32 (0.7745)	26 (0.8062)	×	17 (0.5603)	27 (0.6380)	41 (0.8467)	34 (0.7272)
LUAD	11 (0.9818)	27 (0.9954)	20 (1.125)	17 (0.6632)	×	21 (0.6386)	26 (0.8249)	33 (1.059)
LUSC	23 (1.299)	35 (0.8795)	32 (1.151)	27 (0.8070)	21 (0.6823)	×	59 (1.066)	52 (1.033)
OV	116 (1.197)	90 (0.9046)	111 (1.238))	41 (0.7929)	26 (0.6527)	59 (0.7895)	×	121 (0.9104)
READ	62 (1.159)	94 (0.9536)	65 (1.201)	34 (0.7451)	33 (0.9169)	52 (0.9499)	121 (0.9961)	×

Table V. *Cilia genes with highest absolute value loadings on Q-mode principal components axes 1 and combined loadings on principal component axes 1 and 2. Principal component vectors were computed from fold changes over the set of all genes.*

Top 5% of principal component 1 loadings	*FMN2, CENPE, CNN3, EIF3D, FEN1, MPP2, NUDC, PDXK, RAB6A, ABCF2, ADRM1, BCAT1, CNDP2, COG2, DMXL2, DOCK7, EXOC7, FLNA, FXR1, GNAZ, ITSN1, MYO5A, NOMO1, PA2G4, PADI2, RAB3A, RAB9A, SBDS, TMED4, TNPO1, WDR66, XPO1*
Top 5% of principal component 1+2 loadings	*FMN2, CENPE, CALU, CNN3, EIF3D, FEN1, MPP2, NUDC, PDXK, RAB6A, TMCC2, ABCF2, ADRM1, BCAT1, CNDP2, COG2, DMXL2, DOCK7, FLNA, FXR1, GNAZ, GSTT2, ITSN1, MMAB, MYO5A, NOMO1, PA2G4, RAB3A, RAB9A, SBDS, TMED4, TNPO1, WDR66, XPO1*

enrichment value close to 1 suggests that the overlap in differentially expressed cilia genes is consistent with the overall similarity among dysregulated genes. Values greater than unity suggest an enrichment for shared differentially expressed cilia genes, while values less than one suggest under-representation.

Table IV shows the values of E_{ij} for all pairs of cancer types. Averaged over all cancer pairs, the relative proportion of shared cilia genes was 0.95, suggesting a slightly lower representation of shared dysregulated cilia genes with respect to the proportion of all shared differentially expressed genes. BRCA shared the most differentially expressed cilia genes with the other cancer types, ranging from 1.35 (with respect to KIRC) to 0.98 (LUAD) enrichment (median=1.24). By comparison, LUAD shared the fewest differentially expressed cilia genes with other cancer types, ranging from 0.56 (KIRC) to 0.92 (READ) enrichment (median=0.76). Note that the shared fractions are asymmetric, *i.e.* $E_{ij} \neq E_{ji}$, because the denominator in each ratio is determined by the number of dysregulated genes specific to a particular cancer.

The extent of concordance in dysregulation patterns among genes across the different cancers was also quantified by their loadings on the first principal components axis. The first two of eight Q-Mode PCA axes account for 62.03 and 14.73% of the variance (Figure 1) in fold change expression level variation. Genes with large absolute loadings on the first PCA axis are those that are jointly dysregulated within several cancer types. We took the 1,062 genes (the upper 5%) that had the highest absolute value loadings on the first principal component axis, and found that 32 out of the 42 'majority-rule' dysregulated cilia genes were in this upper percentile of first PCA loading, with two additional cilia genes in the highest loadings for projections on principal component axes 1 and 2.

The relationship between statistically significant dysregulation and first PCA loading was not absolute, even among the three genes that were dysregulated in seven cancer types, one of them (*FMN2*) does not appear in the upper 5% of loadings on the first two principal component axes. Table V lists genes from the above set of 42 that have upper 5% loading scores on PCA axis 1 and jointly on PCA axes 1 and 2.

Genes identified in the ciliome have a broad range of specific roles in cells which can be characterized through functional annotation. Functional clustering (based on significant enrichment scores computed using the DAVID database) of the 42 genes that were significantly dysregulated in a majority of cancers shows significant enrichment with respect to cytoskeleton, mitosis and cytokinesis, protein transport, vesicle transport, and guanyl (GTP) binding pathways. The Supplementary Table SI (Supplementary data are available from the first author upon request) provides DAVID's functional

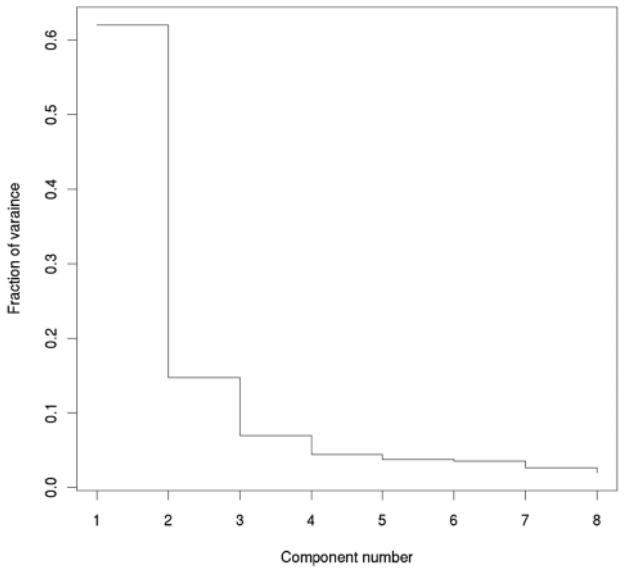

Figure 1. *Eigenvalues 1-8 derived from Q-mode principal component analysis on the data set of fold change over all genes in the eight cancer types. The eigenvalues are normalized to show the fraction of variance expressed by each principal component.*

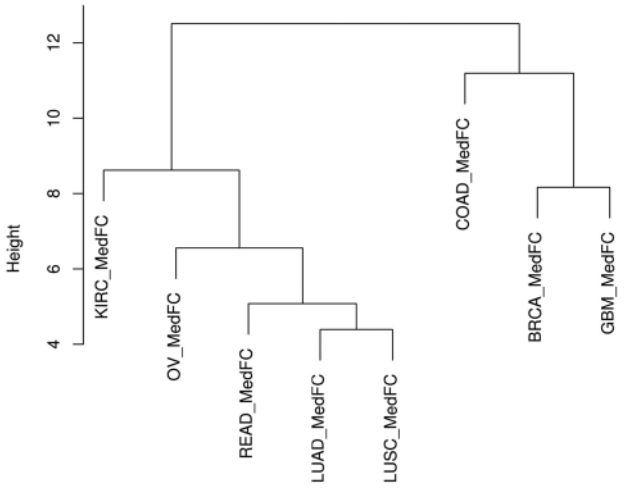

Figure 2. *Hierarchical clustering (dendrogram) of cancer types, derived from a Euclidean distance matrix using median fold change in the expression levels of the 42 cilia genes that are significantly dysregulated in at least five out of the eight types of cancers.*

characterization of the 42 cilia genes that are differentially expressed across five or more of the cancer types.

Clustering of cancer types with respect to gene expression. Hierarchical clustering of cancer types according to expression levels of the 42 significantly dysregulated cilia genes identifies two groups of cancer types as subtrees (Figure 2). The resulting topology of the dendrogram is largely consistent with that computed from all 21,232 genes in the sample (Supplementary Figure S1), apart from having OV as an outgroup to LUAD, LUSC, and READ rather than in a single node with READ. Meanwhile, if all cilia genes are included (Figure 3), most of the nodes and sub-trees in the dendrogram are entirely congruent with the tree structure derived from distances over all genes (the exception being COAD as an out-group to all other cancer types on a very short branch with respect to the BRCA and GBM pair). These dendrograms suggest that patterns of differential expression across all genes in cancer genomes are mirrored in the ciliome.

Figures 2, 3 and S1 also identify two clusters of cancer types in gene expression space with distinct dysregulation profiles, a result that is also supported by K-means clustering in PCA space (see below). The two clusters are COAD, BRCA and GBM, and KIRC, LUAD, LUSC, OV and READ, respectively. The dendrogram implies that there will be some cilia genes that are uniquely dysregulated in most of the cancer types in one subset but not the other, as well as a smaller subset dysregulated in both groups of cancer types. The Venn diagram shown in Figure 4 was constructed by applying a majority rule to identify subset-specific dysregulation, *i.e.* cilia genes that were dysregulated in at least two out of the three in the GBM group and at least three out of the five in the LUSC cluster. There were 130 cilia genes uniquely dysregulated in the subset containing GBM and 59 uniquely dysregulated in the subset containing LUSC, with 25 in the intersection.

Functional clustering using DAVID indicated significant enrichment in dysregulated cilia genes related to intracellular transport, vesicle mediated transport, and protein/macromolecule localization in the GBM cluster. Surprisingly, there were no statistically significant functional enrichments specific to the LUSC cluster (indicating that no particular functional subgroup was disproportionately represented), while the intersecting subset had weakly-significant enrichment for cytoskeletal components. Supplementary Tables SII and SIII (Supplementary data are available from the first author upon request) list the annotation sets and the enrichment scores for intersection of sets of genes that are dysregulated in the COAD, GBM and BRCA, and the KIRC, OV, READ, LUSC and LUAD classes, respectively.

Projecting fold changes of cilia gene expression levels for each cancer type onto the first two R-mode principal component axes gives clusters consistent with the dendrograms shown in Figure 2 and S1, *i.e.* the same subsets of three and five cancer types. This can be seen in the K-means clustering (for K=2 centers, in order to test consistency with the two subgroups identified through hierarchical clustering) in Figure 5.

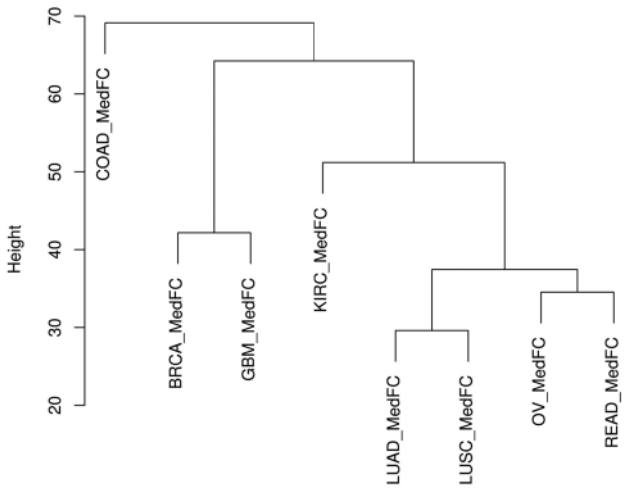

Figure 3. *Hierarchical clustering (dendrogram) of cancer types, derived from the matrix of Euclidean distances of fold change values across the set of all 1,920 cilia genes.*

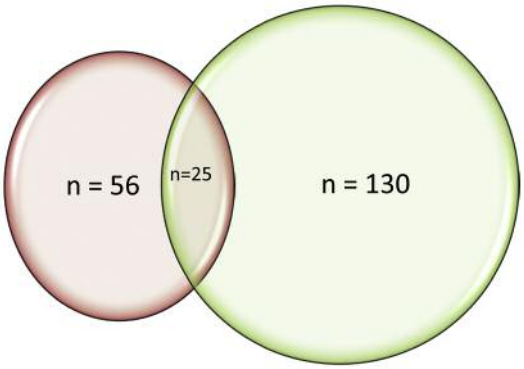

Figure 4. *Venn diagram showing the number of cilia genes uniquely dysregulated in the cluster [COAD, GBM, BRCA], in the cluster [KIRC, OV, READ, LUSC, LUAD], and in both sets (intersection). The larger subset of genes is found in the COAD group.*

When the fold change values of cilia genes were adjusted by subtracting their projections onto the first Q-mode principal component axis (from the eight PCA eigenvectors computed from the set of all 21,232 genes), we obtain a set of measurements that captures variational properties of gene expression levels that are not concordant with overall tendencies and directions among the entire set of genes. The resulting topology of the dendrogram (Figure 6) differs from that in Figures 2 and S1. For instance, READ and OV no longer share a common node, nor do GBM and BRCA.

Clustering of cilia gene expression across cancer types. The 42 cilia genes that are dysregulated in at least five of the cancer types are hierarchically clustered in Figure 7. While groups of genes form clearly defined sub-trees with short branches, indicating very similar patterns of dysregulation across cancer types, these subsets do not correspond to specific functional pathways identified using DAVID. For example, cytoskeleton genes *FMN2*, *CNN3*, and *TMOD1* are dysregulated in cancer types that appear in different subtrees. The same is observed in mitosis and cytokinesis genes *BCAT1*, *NUDC*, and *DOCK7*, and among membrane protein genes *CSPG4* and *MPP2*.

Discussion

The present study identified genes with hypothesized cilia-related function that were dysregulated in cancers, among them, 42 that were dysregulated in at least 5 of the cancer types surveyed. The comparatively weak pairwise overlap may suggest that different sets of cilia genes are dysregulated in different cancers, but to a certain extent it also reflects the number of independent tumor samples available for each cancer, since larger tumor datasets possessed greater statistical power to detect for differential gene expression. For instance, LUAD had the smallest sample sizes for both cancer and control, while OV had the largest, so not surprisingly, we identified the fewest dysregulated genes in LUAD and the most in OV.

More generally, the results of this study suggest that dysregulation of gene expression in the ciliome mirrors broader patterns of the entire cancer genome. This can be seen most clearly from the fact that dendrograms of cancer types derived from Euclidean distances in expression levels across all genes (Supplementary Figure S1) have essentially the same topology as those based on cilia genes, including the more restricted set with significant fold change across a majority of cancer types (Figure 2). Moreover, there is no evidence that dysregulation of the ciliome plays a disproportionate role in cancer, since, if anything, the fraction of significantly dysregulated cilia genes relative to all dysregulated genes tends to be smaller than the fraction of identified cilia genes relative to the entire set, as can be seen from the ratios in Table II. Even the fraction of pairwise shared cilia genes was generally slightly smaller than the pairwise shared genes from the overall pool (Table IV).

Nevertheless, differential expression of particular cilia genes is an important aspect of cancer biology. There are a number of cilia genes that are dysregulated across multiple cancers, including three that were differentially expressed in seven of the eight surveyed. For instance, dysregulation of *MYH11* has been identified in previous expression array studies as contributing to colorectal cancer (21, 22), while *TMOD1* has been found to be strongly antigenic in pancreatic and ovarian cancer (23). Similarly, *FMN2* has been recognized for its role

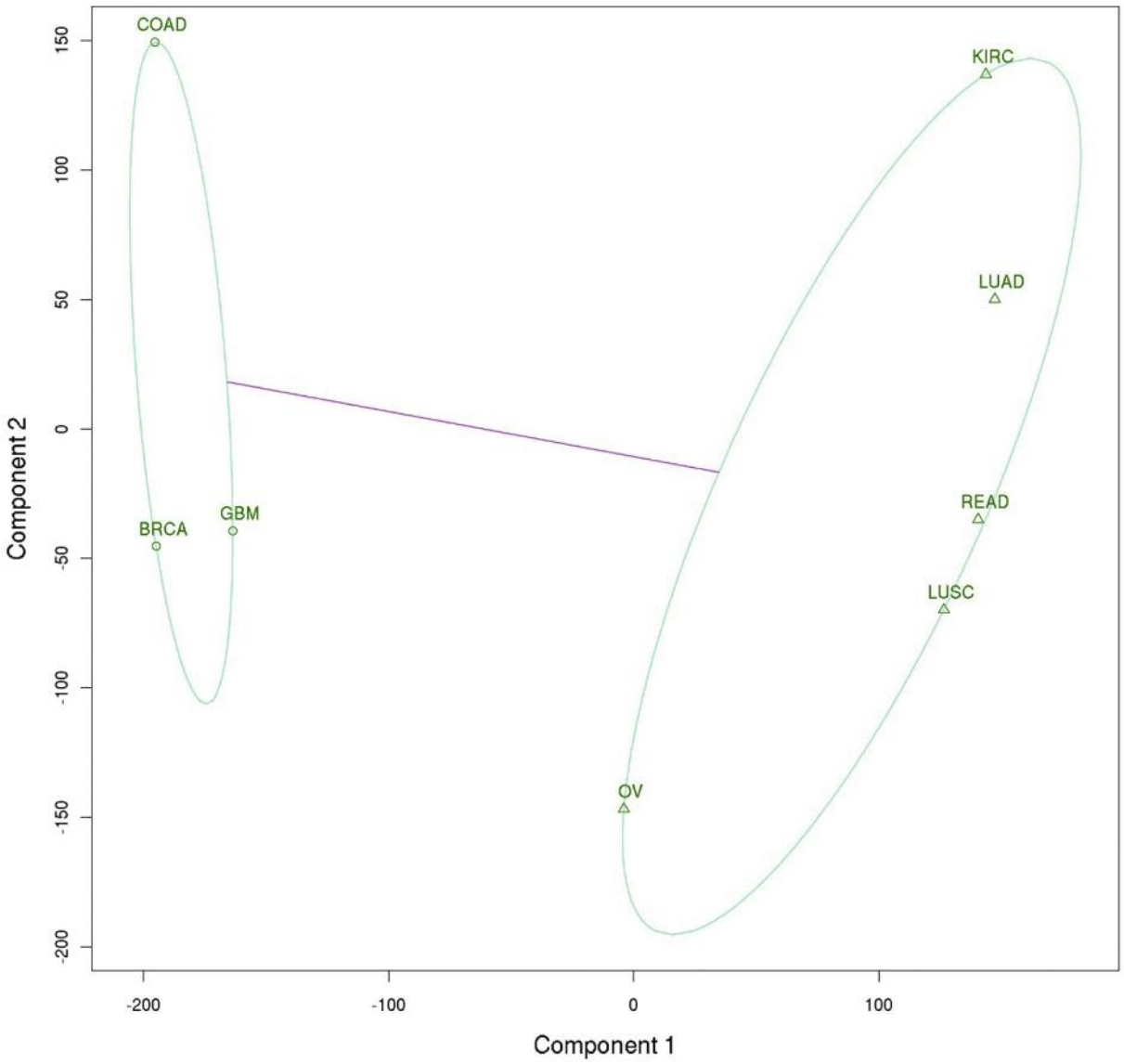

Figure 5. *K-Means clustering of cancer types. Each cancer type's coordinate values (and the Euclidean distances among points representing cancer types) were defined by their loadings on R-mode principal components 1 and 2 (i.e. R-mode principal component analysis on fold change in expression level across all cilia genes). It can be seen that the two principal components separate the cancer types into the same clusters as the hierarchical clustering shown in Figures 2 and 3.*

in a tumor-suppressor pathway, specifically, in preventing degradation of the cell-cycle inhibitor *P21*, thereby controlling the rate of cell division (24).

We note that dysregulation of particular cilia genes is not necessarily consistent across multiple cancers. For instance, the *FMN2* gene was significantly up-regulated relative to the control samples in some types of cancer while being significantly down-regulated in others. This accounts for the fact that of the 42 dysregulated cilia genes, only 32 had loadings on the upper 5% of projection scores on the first principal component axis. Essentially, differential gene expression in opposing directions accounts for weak loadings of the remaining 10 genes (including *FMN2*).

Since the first principal component captures most of the co-variation in gene expression, the goal of adjusting fold changes values by subtracting their regression on the first PCA axis was to identify those variational aspects of differential expression that are not following the overall trends in the genome. Consequently, the fact that the resulting dendrogram (Figure 6) on the adjusted scores had a different

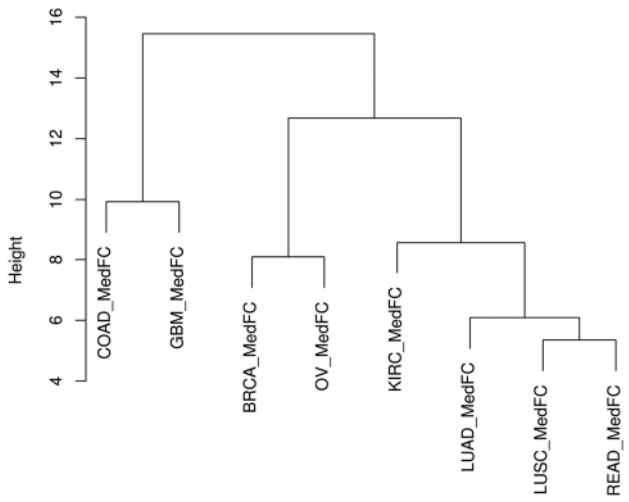

Figure 6. *Hierarchical clustering (dendrogram) of cancer types on adjusted fold changes of expression levels in the 42 significantly dysregulated cilia genes that are shared across cancers. The dendrogram was computed from Euclidean distances among adjusted gene expression values, defined as the raw fold change for each gene expression level across the different cancer types, minus the projection of these values onto the first Q-mode principal component (PCA in 'sample space'), as defined in equations 1-2.*

topology (particularly with regard to pairwise associations) than the clusters derived from unadjusted scores indicates the existence of dysregulation patterns that are specific to particular subsets of cilia genes among cancers. Strong dysregulation over a comparatively small subset of genes may also account for the reason why clusters defined by Euclidean distance on the unadjusted scores (specifically, the subsets of five and three cancer types) were not fully congruent with subsets defined by whether the average direction of dysregulation tended to be positive or negative.

Genes identified in the ciliary proteome were by no means closely related to one another in either their functional pathways or their expression patterns. For instance, when genes were clustered by their expression levels across cancers (Figure 7), those with significant enrichment scores for cytoskeletal function (*e.g. TMOD1, CNN3, FMN2*) are scattered throughout the dendrogram, as opposed to being members of a single, congruently dysregulated subset. The same was true for genes associated with protein transport (*e.g. FLNA, CNDP2*), mitosis/cytokinesis (*e.g. BCAT1, NUC, DOCK7*), cell membrane proteins (*CSPG4, MPP2*) and a wide range of other categories and pathways. Presumably, this is due to the fact that while oncogenesis often involves dysregulation of genes in similar functional roles (*e.g.* cell signaling, cell-cycle regulation, adhesion) , they need not be the same set of genes in each role. For example, *BCAT1* and *NUDC* are both expressed during mitosis. The former is dysregulated in (among others) GBM, but not in *BRCA*, despite the fact that the two cancer types share similar overall expression patterns. Similarly, the plasma membrane gene *CSPG4* is differentially expressed in OV but not in GBM or KIRC, while another plasma membrane gene *MPP2* is dysregulated in GBM but not in KIRC.

A potential caveat to consider in cluster analysis is whether the similarities and associations in expression patterns reflect differences in gene expression in cancer, or just differential gene expression across the tissue types that gave rise to cancer. In other words, could it be that the difference in expression patterns between GBM and OV simply reflects gene expression in glial *versus* ovarian cells? To some degree, this concern is addressed by the fact that we define phenotype not by raw expression level but by the fold change in expression between cancer and controls (albeit not necessarily samples from the same patient). We can further address this concern by noting that cluster analysis of gene expression profiles across normal, healthy tissue types does not give the same dendrograms as those we found in cancer genomes.

The dendrogram in Figure 8 is modified after Figure 1 in (25), which was computed by hierarchical clustering of Euclidean distances of gene expression levels from different tissue types (we only show those tissues/organs corresponding to the tumor samples and to the control). Their results indicate that healthy ovary, lung, and breast tissue shared a common node, indicating similar expression profiles. In contrast, our dendrograms in Supplementary Figure S1 and Figures 2 and 3 suggest that cancers of these tissues did not have similar expression profiles, either overall or among cilia genes. Furthermore, among the cancers surveyed, OV and READ share the highest proportion of dysregulated genes while healthy ovarian tissues do not have expression profiles similar to healthy colon cells. These contrasts suggest that the differences in gene expression between tumor cells and the control tissue samples are typically specific to the cancer rather than the source tissues.

We also remark that the cilia genes themselves have products that have a wide range of functional roles. Among them are cytoskeletal genes that regulate mitosis and cytokinesis (*e.g. BCAT1, DOCK7, NUDC, CENPE*), calcium-binding proteins involved in endoplasmic reticulum function (*e.g.* CALU), plasma membrane proteins (*MPP2* and *CSPG4*), guanyl binding proteins (including several RAB genes), and a number of genes involved in protein and vesicle transport (*e.g. FLNA, CNDP2*). Most of these functions are not directly associated with primary cilial structure as such. Consequently, it is not clear whether observed loss or modification of primary cilia (3, 5, 8) is a contributing cause to carcinogenesis or a secondary manifestation due to dysregulation of genes that drive other processes in cell biology, quite apart from their possible roles in the ciliome.

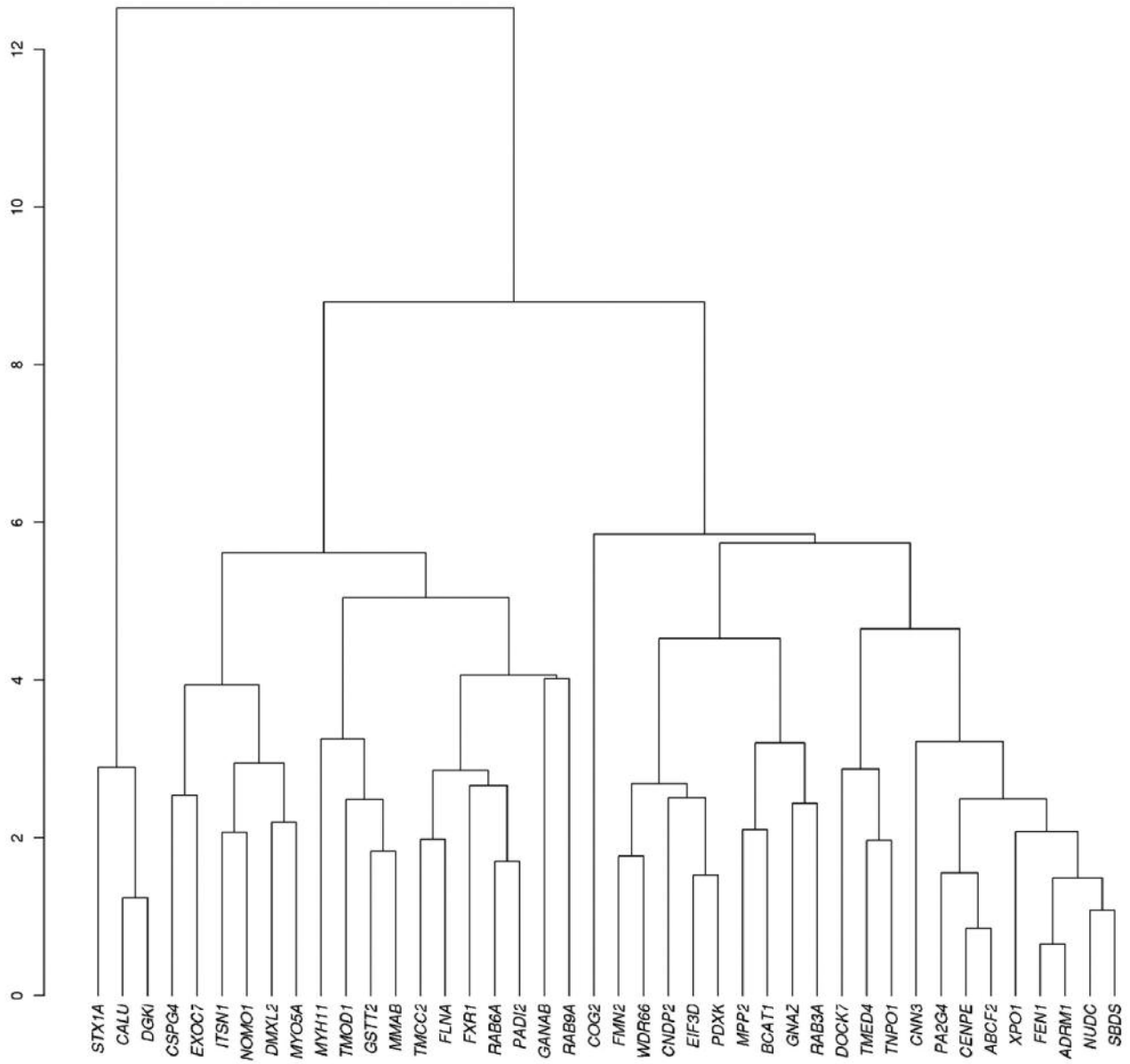

Figure 7. *Hierarchical clustering (dendrogram) of the 42 cilia genes that are significantly dysregulated in at least five types of cancer, clustered by their similarity of expression patterns across cancer types. Each gene was represented as a vector of length 8, the values of which were fold change of expression levels of that gene in every cancer type. The hierarchical clustering was derived from a distance matrix computed from the Euclidean distances across all pairs of genes.*

A notable absence from the set of dysregulated ciliary genes are those that are known to be associated with the hedgehog signaling pathway. Out of the 1,092 cilia genes, only five were functionally-related to the hedgehog pathway. Among these, two were dysregulated in more than a single type of cancer. One of these, *RAB23*, is an oncogene involved in GTPase-mediated signal transduction and has been identified in a number of developmental abnormalities (26) and in cancer. The other, *FKBP8*, is an immunophilin involved in protein trafficking and neuronal development (27). Dysregulation of both *RAB23* (28) and *FKBP8* (29) has been documented in a number of cancer types, particularly primary gastric cancer. More generally, since disruption of the hedgehog pathway has been previously identified as being of significant importance in a number of cancer types (1, 30), the absence of *SHH*, *DHH* genes and most downstream elements from the analysis are a consequence of the scope of the cilia proteome database's search criteria and working definitions.

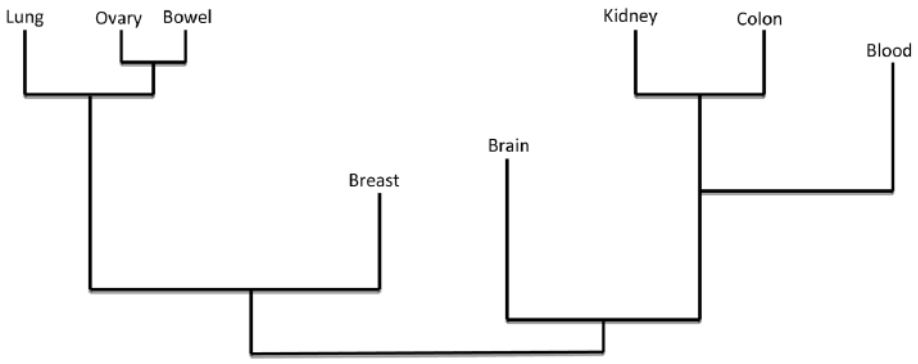

Figure 8. *Dendrogram showing the relationship among healthy tissue types, clustered by gene expression level (based on Figure 1 in Shyamsundara (25). Compared to the hierarchical clustering of fold change in gene expression derived from cancers of these same organs and tissues in Figures 2 and 3 and Supplementary Figure S1.*

Conclusion

Regardless of whether their dysregulation is cause or consequence in oncogenesis, changes in the expression level of cilia genes is closely related to dysregulation across the genome and the breakdown of a wide range of cellular structures and physiological processes. Consequently, further functional studies of cilia genes and their targeted pathways can identify useful markers of cancer progression, and may even serve as suitable drug targets for therapy.

Author's Contributions

The project was conceived by Matthew Cowperthwaite. Preliminary data processing was carried out by Marcus Goldberg, most of the statistical analysis was done by Max Shpak. All three authors contributed to the writing of the manuscript.

Acknowledgements

We thank Jason Pan for assistance with obtaining and processing the data and John Wallingford for helpful advice and comments. This work was supported by a grant from the St. David's Foundation Impact Fund.

References

1 Tukachinsky H, Lopez LV and Salic A: A mechanism for vertebrate hedgehog signaling: recruitment to cilia and dissociation of SuFu–Gli protein complexes. Jour Cell Bio *191*: 415-428, 2010.

2 Badano JL, Mitsuma N, Beales PL and Katsanis N: The ciliopathies: an emerging class of human genetic disorders. Annu Rev Hum Genom Hum Genet *7*: 125-148, 2006.

3 Breunig JJ, Sarkisian MR, Arellano JI, Morozov YM, Ayoub AE, Sojitra S, Wang B, Flavell RA, Rakic P and Town T: Primary cilia regulate hippocampal neurogenesis by mediating sonic hedgehog signaling. Proc Natl Acad Sci USA *105*: 13127-13132, 2008.

4 Cervantes S., Lau J, Cano DA, Borromeo-Austin C and Hebrok M: Primary cilia regulate Gli/Hedgehog activation in pancreas. Proc Natl Acad Sci USA *107*: 10109-10114, 2010.

5 Louvi A and Grove E: Cilia in the CNS: The quiet organelle claims center stage. Neuron *69*: 1046-1060, 2011

6 Hassounah NB, Bunch TA and McDermott KM: Molecular pathways: The role of primary cilia in cancer progression and therapeutics with a focus on hedgehog signaling. Clin Cancer Res *18*: 2429-2435, 2012

7 Han YG, Kim HJ, Dlugosz AA, Ellison DW, Gilbertson RJ and Alvarez-Buylla A: Dual and opposing roles of primary cilia in medulloblastoma development. Nat Med *15*: 1062-1065, 2009.

8 Hou S and Han Y-G: Primary cilia and cancer. *In*: K.L. Tucker and T. Caspary, eds., Cilia and Nervous System Development and FunctionSpringer-Verlag, New York pp. 209-228, 2013.

9 Moser, JJ., Fritzler MJ and Rattner JB: Primary ciliogenesis defects are associated with human astrocytoma/glioblastoma cells. BMC Cancer *9*: 448, 2009.

10 Schraml P, Frew IJ, Thoma CR, Boysen G, Struckmann K, Krek W and Moch H: Sporadic clear cell renal cell carcinoma but not the papillary type is characterized by severely reduced frequency of primary cilia. Mod Pathol *22*: 31-36, 2008.

11 Seeley ES, Carrière C, Goetze T, Longnecker DS and Korc M: Pancreatic Cancer and precursor pancreatic intraepithelial neoplasia lesions are devoid of primary cilia. Cancer Res *69*: 422-430, 2009.

12 Yuan K, Frolova N, Xie Y, Wang D, Cook L, Kwon YJ, Steg AD, Serra R and Frost AR: Primary cilia are decreased in breast cancer: analysis of a collection of human breast cancer cell lines and tissues. Jour Histochem Cytochem *58*: 857-870, 2010.

13 Huang da W, Sherman BT and Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*: 44-57, 2009.

14 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA *102*: 15545-1550, 2005.

15 Irizarry RA, Wang C, Zhou Y and Speed TP: Gene set enrichment made simple. Stat Methods Med Research *18*: 565-575, 2009.
16 Reimers M and Carey VJ: Bioconductor: an open source framework for bioinformatics and computational biology. Methods Enzymol *411*: 119-134, 2006.
17 Smyth GK and Speed TP: Normalization of cDNA microarray data. Methods *31*: 265-273, 2003.
18 Smyth GK: Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol *3*: Article 3, 2004.
19 Hartigan JA and Wong MA: A K-means clustering algorithm. Appl Stat *28*: 100-108, 1979.
20 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA and Reich D: Principal components analysis corrects forstratification in genome-wide association studies. Nat Genet *38*: 904-909, 2006.
21 Kruhøffer M, Jensen JL, Laiho P, Dyrskjøt L, Salovaara R, Arango D, Birkenkamp-Demtroder K, Sørensen FB, Christensen LL, Buhl L, Mecklin JP, Järvinen H, Thykjaer T, Wikman FP, Bech-Knudsen F, Juhola M, Nupponen NN, Laurberg S, Andersen CL, Aaltonen LA and Ørntoft TF: Gene expression signatures for colorectal cancer microsatellite status and HNPCC. Brit Jour Cancer *92*: 2240-2248, 2005.
22 Alhopuro P, Phichith D, Tuupanen S, Sammalkorpi H, Nybondas M, Saharinen J, Robinson JP, Yang Z, Chen LQ, Orntoft T, Mecklin JP, Järvinen H, Eng C, Moeslein G, Shibata D, Houlston RS, Lucassen A, Tomlinson IP, Launonen V, Ristimäki A, Arango D, Karhu A, Sweeney HL and Aaltonen LA: Unregulated smooth-muscle myosin in human intestinal neoplasia. Proc Natl Acad Sci USA *105*: 5513-5518, 2008.
23 Gnjatic S, Ritter E, Büchler MW, Giese NA, Brors B, Frei C, Murray A, Halama N, Zörnig I, Chen YT, Andrews C, Ritter G, Old LJ, Odunsi K and Jäger D: Seromic profiling of ovarian and pancreatic cancer. Proc Natl Acad Sci USA *107*: 5088-5093, 2010.
24 Yamada K, Ono M, Perkins ND, Rocha S and Lamond AI: Identification and Functional Characterization of FMN2, a Regulator of the Cyclin-Dependent Kinase Inhibitor p21.Molec Cell *7*: 922-933, 2013.
25 Shyamsundar R, Kim YH, Higgins JP, Montgomery K, Jorden M, Sethuraman A, van de Rijn M, Botstein D, Brown PO and Pollack JR: A DNA microarray survey of gene expression in normal human tissues. Genome Bio *6*: R22.
26 Wang Y, Ng EL and Tang BL: Rab23: what exactly does it traffic? Traffic *7*: 746-750, 2006.
27 Bulgakov OV, Eggenschwiler JT, Hong DH, Anderson KV and Li T: FKBP8 is a negative regulator of mouse sonic hedgehog signaling in neural tissue. Development *131*: 2149-2159, 2004.
28 Hou Q, Wu YH, Grabsch H, Zhu Y, Leong SH, Ganesan K, Cross D, Tan LK, Tao J, Gopalakrishnan V, Tang BL, Kon OL and Tan P: Integrative genomics identifies RAB23 as an invasive mediator gene in diffuse-type gastric cancer. Cancer Res *68*: 4624-4630, 2008.
29 Oh JH, Yang JO, Hahn Y, Kim MR, Byun SS, Jeon YJ, Kim JM, Song KS, Noh SM, Kim S, Yoo HS, Kim YS and Kim NS: Transcriptome analysis of human gastric cancer. Mamm Genome *16*: 942-954, 2006.
30 di Magliano MP: Hedgehog signalling in cancer formation and maintenance. Nat Rev Cancer *3*: 903-911, 2013.

Received January 11, 2014
Revised January 23, 2014
Accepted January 24, 2014