Systematic transcriptome analysis reveals tumor-specific isoforms for ovarian cancer diagnosis and therapy

Christian L. Barrett[a,b], Christopher DeBoever[a,c], Kristen Jepsen[d], Cheryl C. Saenz[e], Dennis A. Carson[a,e,f,1], and Kelly A. Frazer[a,b,d]

[a]Moores Cancer Center, [b]Department of Pediatrics and Rady Children's Hospital, [c]Bioinformatics and Systems Biology, [d]Institute for Genomic Medicine, [e]Department of Medicine, and [f]Sanford Consortium for Regenerative Medicine, University of California, San Diego, La Jolla, CA 92093

Contributed by Dennis A. Carson, April 26, 2015 (sent for review February 10, 2015)

Tumor-specific molecules are needed across diverse areas of oncology for use in early detection, diagnosis, prognosis and therapy. Large and growing public databases of transcriptome sequencing data (RNA-seq) derived from tumors and normal tissues hold the potential of yielding tumor-specific molecules, but because the data are new they have not been fully explored for this purpose. We have developed custom bioinformatic algorithms and used them with 296 high-grade serous ovarian (HGS-OvCa) tumor and 1,839 normal RNA-seq datasets to identify mRNA isoforms with tumor-specific expression. We rank prioritized isoforms by likelihood of being expressed in HGS-OvCa tumors and not in normal tissues and analyzed 671 top-ranked isoforms by high-throughput RT-qPCR. Six of these isoforms were expressed in a majority of the 12 tumors examined but not in 18 normal tissues. An additional 11 were expressed in most tumors and only one normal tissue, which in most cases was fallopian or colon. Of the 671 isoforms, the topmost 5% (n = 33) ranked based on having tumor-specific or highly restricted normal tissue expression by RT-qPCR analysis are enriched for oncogenic, stem cell/cancer stem cell, and early development loci—including ETV4, FOXM1, LSR, CD9, RAB11FIP4, and FGFRL1. Many of the 33 isoforms are predicted to encode proteins with unique amino acid sequences, which would allow them to be specifically targeted for one or more therapeutic strategies—including monoclonal antibodies and T-cell–based vaccines. The systematic process described herein is readily and rapidly applicable to the more than 30 additional tumor types for which sufficient amounts of RNA-seq already exist.

ovarian cancer | RNA-seq | bioinformatics | diagnostics | therapeutics

Identifying molecules that are specific to tumors for use in early detection, diagnosis, prognosis, and therapeutic strategy design is both a primary goal and a key discovery challenge across diverse areas of oncology. Furthermore, the extent of inter- and intratumor heterogeneity indicates that multiple tumor-specific molecules will be needed for any of these applications (1–3). Although DNA alterations constitute the major focus of tumor-specific discovery efforts to date, in many respects mRNA is more attractive for this purpose. This is because RNA can (*i*) broadly reflect (malignant) cellular phenotypes, (*ii*) exist in thousands of copies per cell and thereby enable highly sensitive early detection and diagnostic assays, and (*iii*) sensitively and comprehensively reveal potential candidate antigens for monoclonal antibody targeting, vaccines, and adoptive immunotherapies (4–6). The efficacy of using mRNA for these purposes is highly dependent on the degree of tumor-specific expression.

One of the main themes of microarray-based experiments that have been undertaken during the last decade has been the discovery of tumor-specific "genes." Aside from the class of cancer-germ-line (aka cancer/testis) genes (7), few have been found. In retrospect, the "gene" concept critically hindered these efforts to discover tumor-specific expression because the word "gene" is a collective term for all mRNA isoforms expressed from a genomic locus. Malignant and normal tissue types can be distinguished by patterns of differential isoform use (8, 9), but when measured in aggregate at the "gene" level the isoform-specific differences are at best recognized as "gene overexpression" or "gene underexpression." Thus, mRNA expression is not commonly considered to be "tumor-specific", but "tumor-associated" (via overexpression). The distinction is important, for "tumor-specific" molecules are an ideal that is devoid of detection interpretation ambiguity and off-targeting. So although it has become increasingly clear that there are few, if any, "genes" only expressed in tumors, aside from fusion transcripts (10) the extent to which tumor-specific mRNA isoforms exist is unknown.

Transcriptome sequencing (RNA-seq) is a genomics technology whose principle purpose is to enable genome-wide expression measurements of mRNA isoforms—the level at which distinct tumor-specific mRNA molecules are to be found. To apply RNA-seq for the purpose of identifying mRNA isoforms that tumors express and normal tissues do not express, large databases of RNA-seq data from malignant and normal tissues are required. The Cancer Genome Atlas (TCGA; cancergenome.nih.gov) is an NIH-sponsored effort to study the RNA and DNA in 500 tumors for many cancer types, and the Genotype-Tissue Expression (GTEx) program (11) is an NIH-sponsored effort to study the RNA and DNA in thousands of samples from >50 distinct normal tissue sites. Both of these programs are multicenter efforts that are generating molecular profiling data at a rate, scale, and cost that almost certainly could not be borne by any

Significance

Identifying molecules that are specific to tumors for use in early detection, diagnosis, prognosis, and therapy is both a primary goal and a key discovery challenge across diverse areas of oncology. To discover ovarian tumor-specific molecules, we developed custom bioinformatics algorithms to analyze transcriptome sequence data of 296 ovarian cancer and 1,839 normal tissues and validated putative tumor-specific mRNA isoforms by RT–quantitative PCR. The results revealed multiple candidate diagnostic and therapeutic targets with unique sequences that were expressed in most of the cancers examined but not in normal tissues. The process we developed can be readily applied to identify diagnostic and therapeutic targets for any of the 30 or more tumor types for which large amounts of transcriptome data now exist.

Author contributions: C.L.B., K.J., D.A.C., and K.A.F. designed research; C.L.B., C.D., and C.C.S. performed research; K.J., C.C.S., and K.A.F. contributed new reagents/analytic tools; C.L.B. analyzed data; and C.L.B., D.A.C., and K.A.F. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: dcarson@ucsd.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1508057112/-/DCSupplemental.

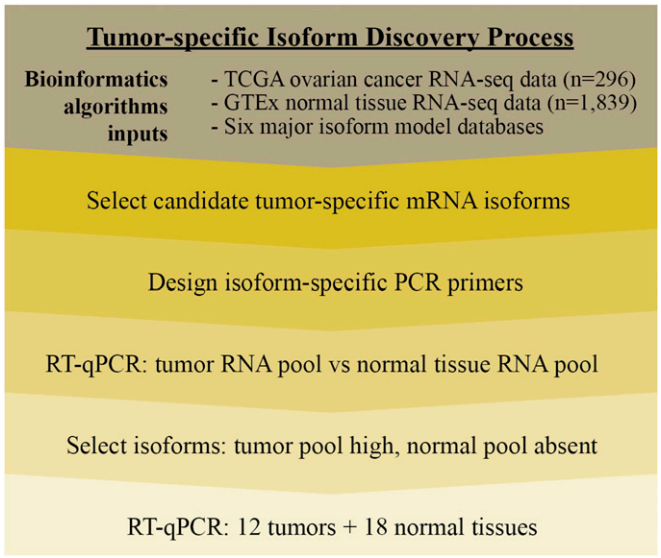

Fig. 1. Discovery process overview. We applied custom bioinformatics algorithms to large public databases of tumor and normal tissue RNA-seq data to rank-prioritize mRNA isoforms by likelihood of being tumor-specific. We then used RT-qPCR in two phases to confirm tumor-specific expression. First we performed RT-qPCR to analyze the RNA of four tumors pooled together vs. the RNA of four normal tissues pooled together. Then we selected the most likely tumor-specific isoforms based on expression profiles in these two pools. Final validation was RT-qPCR on individual tumor and normal tissues.

single entity. The primary intention of these efforts is to generate a public resource to catalyze leaps in progress across all aspects of cancer care, prevention, and therapy. The raw transcriptome data being produced by these efforts has tremendous discovery potential, but to date they have not been rigorously evaluated for tumor-specific molecules for diagnostic and therapeutic applications. Herein we report on our results to date in using these RNA-seq data to identify mRNA isoforms that are only expressed in high-grade serous ovarian adenocarcinomas (HGS-OvCa) and to evaluate the isoforms' potential for tumor biology insights and oncologic applications.

Results

The overall strategy of our tumor-specific isoform identification process (Fig. 1) is based on (*i*) computational algorithms we custom-developed for sensitive and accurate isoform identification, (*ii*) large databases of tumor and normal tissue RNA-seq data produced by TCGA and GTEx, and and (*iii*) high-throughput RT-qPCR experiments. As reported below, we first used our custom algorithms to efficiently process large amounts of RNA-seq data and applied one prioritization strategy to produce a list of mRNA isoforms rank-prioritized by likelihood of being tumor-specific. We then used our custom-developed software for automated design of isoform-specific PCR primers and performed RT-qPCR using pooled tumor RNA and pooled normal tissue RNA. For isoforms found to only be present in the tumor pool, we measured their expression by RT-qPCR in a larger set of non-pooled tumor and normal samples. The isoforms that were expressed across multiple tumors were then ranked based on whether they were expressed in zero, one, two, three, four, or more normal tissues and evaluated for oncologic applications.

Computational Pipeline for RNA-seq. The standard RNA-seq computational pipeline for organisms with a sequenced genome has three main components (Fig. 2*A*): (*i*) alignments of RNA-seq reads to the genome, (*ii*) an isoform model database, and (*iii*) an integration algorithm, whose input is the isoform model database and the read pair alignments and whose output is the expression level of the supplied isoforms. We developed a pipeline for isoform identification and expression level estimation that is distinguished by custom methodologies and software algorithms in each of these three components.

A major distinguishing feature of our approach to RNA-seq read alignment is our use of maximally sensitive alignment parameterizations coupled with nucleotide-resolution read-to-isoform correspondence verification. Such parameterizations enable the thorough detection of all RNA-seq read alignments spanning splice junctions, which are especially informative because they provide exon linkage information that can be crucial for accurate isoform identification. Current practice sets "minimum overhangs" of a read's alignment over a splice junction into an adjoining exon—often 8 bp or more—to guard against false genomic alignments. To maximally recover the information in RNA-seq reads, we consider alignments with even 1-bp overhangs, but then through nucleotide-resolution read-to-isoform correspondence verification we reject all read pair alignments that do not exactly match the human genome reference sequence. This approach has four consequences (Fig. 2*B*). First, we maximize the isoform identification information in each set of RNA-seq data. Second, we identify read pairs that do not correspond to any known isoform and prevent their subsequent use for isoform expression estimation. In practice, these rejected read pairs constitute 2–3% of the raw data and are indicative of the presence of isoforms that have not yet been discovered and incorporated into any public database (12). Third, we explicitly associate each read pair with a specific isoform or set of isoforms from which it could have been derived and then use this information in the final expression estimation stage. Owing to the high overlap of isoforms at a genomic locus, read pair alignments often overlap isoforms from which they both could and could not have been physically derived. In some RNA-seq computational protocols, this distinction is not addressed and read pair alignments are erroneously used to estimate the expression of isoforms from which they could not have been physically derived. As shown in Fig. S1 for an exemplar RNA-seq dataset, read-to-isoform correspondence verification markedly reduces the number of isoforms with which read pairs can be associated. Fourth, we explicitly associate read

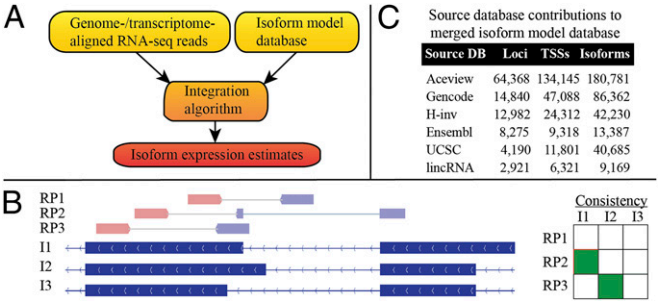

Fig. 2. RNA-seq bioinformatics. (*A*) Our custom RNA-seq computational pipeline broadly conforms to the standard three-component RNA-seq computational pipeline for organisms with a sequenced genome. (*B*) In our approach, we align read pairs (RP) with maximally sensitive parameterizations and use all known splice junctions, allowing even 1-bp splice junction "overhangs." Nucleotide-level read-to-isoform consistency analysis identifies and records the read pair-isoform tuples that are exactly concordant and filters out read pairs that are not exactly concordant with some known isoform (I). (*C*) We minimize isoform nonidentifications (false negatives) with our isoform model database that is a merger of the six major isoform model databases worldwide. Given the read pair-isoform tuples from *B*, we use a parsimony principle to subsequently minimize false isoform identifications (false positives).

pairs to isoforms to enable our strategy for minimizing both false positives and false negatives in RNA-seq experiments (discussed below).

A major distinguishing feature of our approach to isoform models is the use of a custom isoform model database that we created by merging all of the major isoform model databases (Fig. 2*C*). Although the use of only one particular isoform model database is standard in current RNA-seq computational protocols, doing so is a source of false negatives (13); if a particular isoform is not in the database, then the integration algorithm (Fig. 2*A*) cannot know about it and use it for expression estimation. By merging all major isoform model databases, our approach minimizes the possibility of such false negatives. Conversely, isoforms in a supplied isoform model database that are not actually expressed in a sample from which RNA-seq data were generated represent noise for the integration algorithm and can lead to the assignment of nonzero expression for unexpressed isoforms. To minimize the possibility of such false positives, we use the read-to-isoform verification information discussed above and our implementation of a greedy solution to the set cover problem (14) to identify the set of isoforms that most parsimoniously explains the RNA-seq read alignments. In effect, we create an isoform model database that is tailored to each RNA-seq experiment. As shown in Fig. S2, this tailoring reduces the number of isoforms from loci that are used as input to the integration algorithm.

Tumor-Specific Isoform Predictions from 2,135 RNA-seq Experiments. For our study we sought those mRNA isoforms most pervasively and exclusively expressed in HGS-OvCa. Using 296 curated TCGA RNA-seq datasets for HGS-OvCa, we first identified isoforms expressed in 90–100% of tumors. To capture even very lowly expressed transcripts, we used an expression level cutoff of 10^{-6} fragments per kilobase of transcript per million fragments mapped to define whether a transcript was expressed or not. This first filter yielded 117,108 isoforms (Fig. S3*A*). We next used the 1,839 GTEx RNA-seq datasets to count the number of normal tissues in which the average expression of each of these 117,108 isoforms was equal or higher. As shown in Fig. S3*B*, most of the isoforms expressed in 90–100% of the TCGA ovarian tumors were also expressed in many normal tissues. For each of the 22,082 isoforms that was equally or more highly expressed in at most one other tissue, we identified the normal tissue with the highest average expression and computed two statistics: (*i*) the Mann–Whitney *P* value associated with the two sets of expression values (i.e., tumor vs. normal) and (*ii*) the fold change of the average tumor expression over the average normal tissue expression. As shown by Fig. S3*C*, most of the 22,082 isoforms were not appreciably distinguished in their tumor expression from their "closest" normal tissue expression by average expression fold change or the distribution of expression values. Finally, we rank-prioritized the 22,082 isoforms by likelihood of being tumor-specific by sorting them by fold change and *P* value.

High-Throughput mRNA Isoform-Specific PCR Primer Design. The sequencing technology upon which this study is based has the limitation of only being applicable to ∼200–250 bp fragments of cDNA—restricting its ability to unambiguously identify mRNA isoforms that in the human genome are on average ∼2 kb. For this reason we used RT-qPCR to confirm the tumor-specific expression of mRNA isoforms that we rank prioritized by RNA-seq. To enable a large number of RT-qPCR experiments, we custom-developed software that could exhaustively identify and design primers for all unique amplicons of any target mRNA in the human genome. With this software we attempted to design primers for 671 of the topmost tumor-specific candidate mRNA isoforms. The number 671 was chosen so that we could perform our initial pooled screening (discussed below) with 11 384-well plate PCR experiments. To reach 671 it was necessary to attempt primer designs for the 1,230 topmost tumor-specific candidate mRNA isoforms—corresponding to a 54.6% primer design success rate. Of the unsuccessful attempts, 320 (26.0%) were due to the lack of a unique amplicon sequence in the target isoform and 239 (19.4%) were due to primer design failure. (Primer design failure can occur for reasons related to T_m requirements, forward and reverse primer compatibility, primer or amplicon sequence length constraints, and primer amplification of unintended products.)

Confirmation of Isoform Tumor-Specific Expression by RT-qPCR. We performed confirmatory RT-qPCR experiments (Fig. S4 and Table S1) using a two-phase approach. In phase 1 we used pooled RNA to efficiently filter out isoforms that were not expressed in tumors and/or were expressed in normal tissues. We formed a pool of four different tumor RNA samples and a pool of four different normal tissue RNA samples and then measured the expression of all 671 isoforms in both pools. As graphed in Fig. 3, we found that 66.2% ($n = 445$) of isoforms were detected in both pools, 18.2% ($n = 122$) were detected only in the tumor pool, 1.0% ($n = 7$) were detected only in the normal pool, and 14.5% ($n = 97$) were not detected in either pool. Furthermore, our experiments revealed the presence of novel isoforms that are not documented in any of the isoform model databases that we used to construct our custom isoform model database. In the group of isoforms found in both pools, 18.3% of reactions revealed one or two additional products. For the "tumor only" and "normal only" groups, the percentages were 5.7% and 0.4%, respectively.

In phase 2 we measured the expression of a subset of the isoforms in an expanded set of individual, nonpooled RNA samples. For the subset we selected isoforms that were detected in only the tumor pool, that were associated with a single peak melt curve, and that were at least moderately expressed [i.e., quantification cycle (C_q) <31–32]. (Low expression of an isoform in a pool could mean moderate expression in only one sample of the pool or low expression across all samples in the pool.) These selection criteria resulted in a subset constituting 86 of the 122

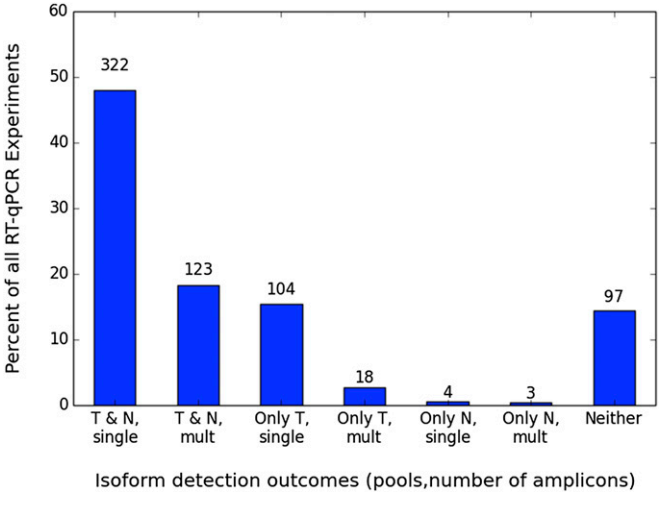

Fig. 3. Categories of pooled RNA RT-qPCR experiment outcomes. For candidate tumor-specific mRNA isoforms derived from RNA-seq-based analyses, we measured their expression by RT-qPCR in a pool of tumor (T) RNA samples and pool of normal tissue (N) RNA samples. The expression status of the isoforms in aggregate spanned all possible outcomes. By melt curve analysis, we observed instances in which just the target product (single) was amplified and instances in which multiple products (mult) were amplified—indicative of the presence of novel mRNA isoform structures. The number of isoforms in each category is displayed atop each bar.

isoforms only detected in the tumor pool. To expand the set of RNA samples we added an additional 8 tumor samples and an additional 14 normal tissue samples—for a total of 12 tumor samples and 18 normal tissue samples. We then measured by RT-qPCR the expression of the 86 isoforms in the 30 individual samples and then ranked the isoforms by the number of normal tissues in which they were expressed. The top-ranked 33 isoforms, shown in Fig. 4 and Table S2, constitute 5% of the original 671 isoforms investigated and either have tumor-specific or restricted normal tissue (normal-restricted) expression. The top six isoforms, or 0.9% of the original 671, were expressed in 6–12 of the 12 tumors and were undetectable in all 18 normal tissues examined. An additional 11 isoforms (1.6% of 671) were only observed in one normal tissue, which in most cases was either fallopian tube or colon. In the remaining 16 cases (2.4% of 671) in which the isoforms were present in two, three, or four normal tissues, fallopian tube and/or ovary were most consistently among the normal tissues.

Biologic Basis and Applications of Candidate Tumor-Specific Molecules. Because the 33 mRNA isoforms in Fig. 4 are expressed in 6–12 of the 12 different tumors and have highly restricted or undetected normal tissue expression, they are of immediate and high interest for both understanding tumor biology and for oncologic applications. A complication that arises when interpreting isoform-level findings is that most isoforms of most genes have not been explicitly studied, and even small differences in mRNA or protein isoform primary sequence from a well-studied canonical isoform can alter the molecule's function, localization, lifetime, structure, and/or interaction network (15). With this caveat in mind, we highlight below isoforms that are likely to play a causative functional role in the malignant state and that have potential use for diagnosis and therapy.

Isoforms of genes related to oncogenesis, stem cells, and stem cell-like cancer cells. A structurally distinct mRNA isoform lAug10 of ETV4/PEA3 (Fig. 4) was expressed in all studied tumors and was detectable only in normal heart. ETV4 is a transcription factor that is active in developing embryos and adult tissues and that has a demonstrated transforming role in Ewing's tumors and prostate, ovarian, breast, and other solid tumors (16). The lAug10 isoform is incompletely known at the 3′ end, but enough of the transcript has been sequenced to reveal that lAug10 is the only ETV4 isoform with a truncated N-terminal amino acid sequence and a skipped exon 5. The functional implications of this distinguishing structure are unknown.

FOXM1 is a transcription factor that is both a potent oncogene and an important molecule for maintaining stem cell renewal (17). The gene is highly expressed across a broad range of different solid tumor types, including ovarian cancer. Integrated genomic analyses of ovarian cancer performed by TCGA found the FOXM1 regulatory network to be the most significantly altered in expression level across 87% of the 489 tumors studied. FOXM1 has multiple isoforms, two of which have been studied for their transforming potential (18). The study found that isoforms FOXM1b and FOXM1c both had transforming potential, and that FOXM1c was likely to be constitutively active because it was proteolytically processed to yield a short isoform without the N-terminal inhibitory domain. The lAug10 and gAug10/ENST00000536066 isoforms that we identify in Fig. 4 were neither of the isoforms studied, but interestingly both are short

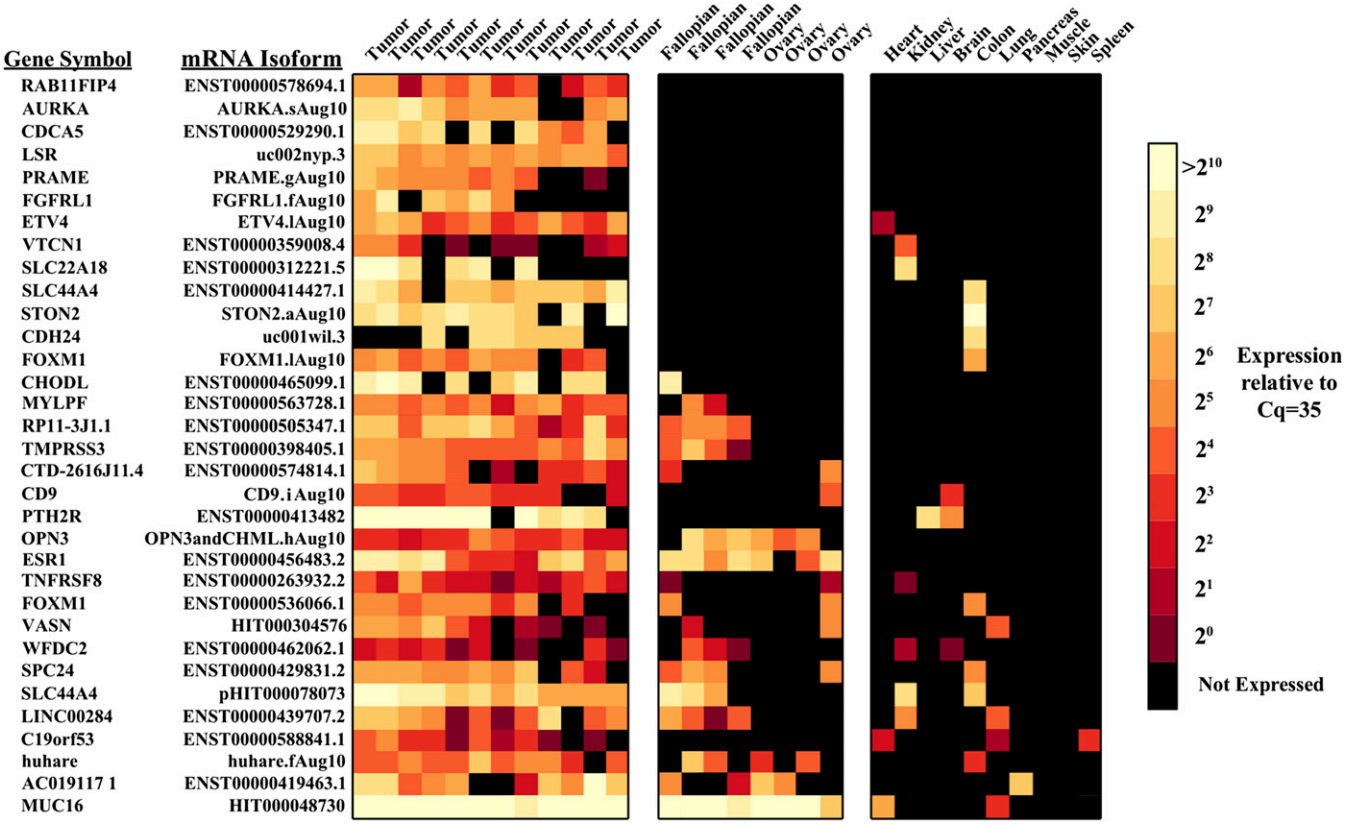

Fig. 4. RT-qPCR expression measurements for mRNA isoforms. In total, we selected 671 mRNA isoforms for tumor-specific confirmation RT-qPCR experiments. Using pooled RNA, we found a subset of them to be only expressed in the tumor RNA pool (Fig. 3). We selected 86 of these for a second set of RT-qPCR experiments with 12 tumor and 18 normal tissue RNA samples, which were not pooled. The 33 mRNA isoforms with tumor-specific and the most normal-restricted expression from the set of 86 are shown, constituting 5% of the original 671.

isoforms that are missing the N-terminal inhibitory domain. Thus, it may be that one or both of the FOXM1 isoforms that we identified are constitutively active transforming isoforms of FOXM1.

Tetraspanin proteins are increasingly viewed as therapeutic targets because of their emerging key roles in tumor initiation, progression, metastasis, and sometimes angiogenesis (19). We identified isoform iAug10 of CD9/tetraspanin-29 that was expressed in 10 of 12 tumors and absent from all but one normal nongynecological tissue sample. CD9 is a cell surface marker for normal human embryonic stem cells and for cancer stem cells in non-small-cell lung carcinoma (20). It has various anti- and protumorigenic roles, with the latter including that of an oncogene in an ovarian cancer line (21). The varied and opposing roles of CD9 have been suggested to be a consequence of its different interaction partners in the plasma membrane (19). An additional and compatible reason, though, may be its multiple protein isoforms.

The lipolysis-stimulated lipoprotein receptor (LSR) is a gene that in basal-like triple-negative breast cancer cell lines is a biomarker of cells with cancer stem cell features and with a direct role in driving aggressive tumor-initiating cell behavior (22, 23). These observations are relevant to our study because of the discovery that basal-like breast cancers and ovarian serous cancers exhibit very similar mRNA expression programs and share critical genomic alterations (24)—indicating related etiology and therapeutic opportunities. At the gene level LSR is transcribed in multiple normal tissues, but our investigation revealed LSR isoform uc002nyp.3 to be expressed across all 12 tumors studied and undetectable in all 12 normal tissues studied. Intriguingly, because of this isoform's structure (Fig. 5*D*) it has dual therapeutic potential; its splice junction forms a unique amino acid sequence that is a predicted extracellular epitope and is computed to have a high binding affinity for three different MHC I alleles. Thus, this isoform has the potential of encoding a protein with one tumor-specific polypeptide that is both an antibody and T-cell target on ovarian cancer stem cells and that, if found to be expressed in breast basal-like tumors, could be relevant for multiple difficult tumor types.

Isoforms for early detection and monitoring of HGS-OvCa. The Papanicolaou test has recently been demonstrated to be a viable source of ovarian tumor cells (25). This observation allows for the possibility of an early ovarian cancer diagnostic test based on the detection of ovarian tumor-specific mRNA isoforms that are expressed in tumor cells that have disseminated to the cervix. For such an early detection strategy to work, one would need to identify mRNA isoforms that are only expressed in ovarian tumors and not in normal gynecologic tissues. Extensive experimental evidence (26–29) indicates that fallopian tube, and to a lesser extent the ovary, are the tissue(s) of origin of HGS-OvCa. Additionally, many studies (30–32) have demonstrated that expression profiles of tumors are more similar to those of their tissue of origin than to any other normal tissue, so for HGS-OvCa fallopian tube and ovary are the most stringent tissues against which to judge the tumor specificity of an mRNA isoform. As shown in Fig. 4, we found that 15 (2.2%) of our original starting set of 671 isoforms were not expressed in the ovary or fallopian tube, and so constitute an initial candidate set of mRNA isoforms upon which a new and innovative strategy for the early detection of ovarian can be developed.

Isoforms predicted to encode cell surface targets. The parathyroid hormone receptor 2 gene PTH2R encodes a class B (type II) G protein-coupled receptor (GPCR) that is predominantly expressed in endocrine and limbic regions of the forebrain and to a lesser extent in restricted cell types of peripheral tissues (33). Its function in nonbrain tissues and in cancer has not been studied. The mRNA isoform that we identified is highly expressed in 10 of the 12 tumors used herein (Fig. 4). The isoform is distinguished by its alternative first exon, which is predicted to retain a (likely cleaved) signal peptide (Fig. 5*A*). In addition to the signal peptide, the first exon would confer on the protein isoform a unique 12-aa sequence. Because the protein is a class B GPCR, its N-terminal sequence is expected to be extracellular and thus amenable to antibody targeting.

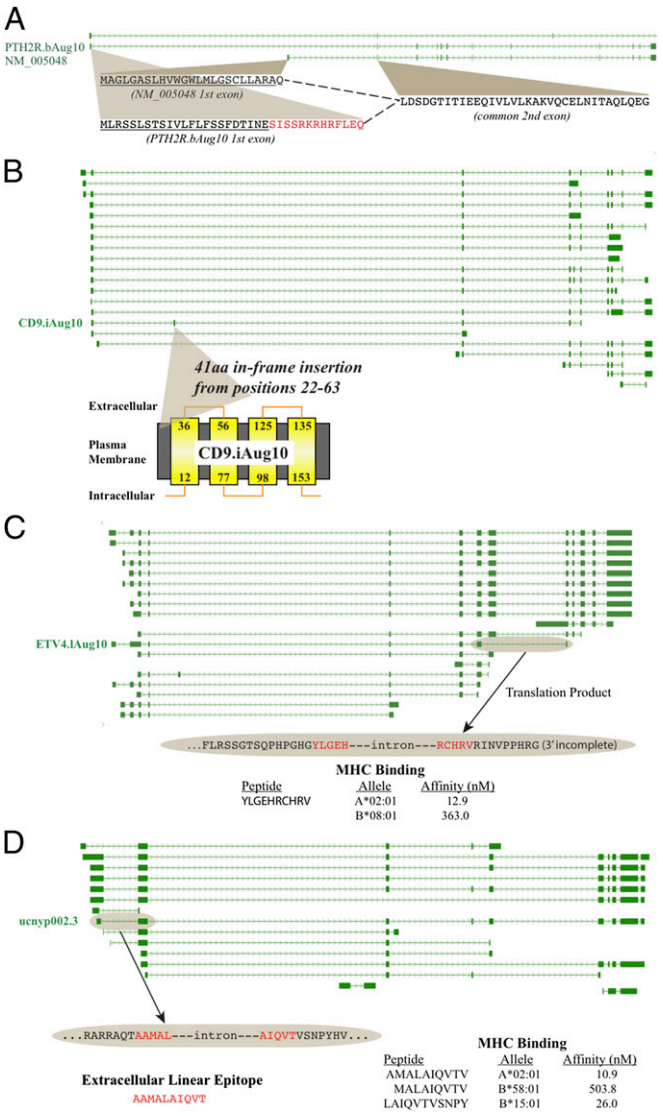

Fig. 5. Candidate protein therapeutic targets. (*A*) The candidate isoform PTH2R.bAug10 is distinguished from the canonical PTH2R isoform by its alternative first exon, which alters the N-terminal amino acid sequence. Both protein isoforms are predicted to contain signal peptides (that are likely cleaved). After signal peptide cleavage, the first exon of PTH2R.bAug10 would still retain a unique 12-aa sequence, which, because the protein is a class B GPCR, is expected to be extracellular and thus amenable for antibody targeting. (*B*) The candidate isoform CD9.iAug10 is distinguished by a unique exon, which is expected to add 41 uniquely distinguishing amino acids—some of which project into the extracellular environment and constitute a protein-specific antibody target. (*C*) The candidate isoform ETV4.lAug10 has a unique exon structure that creates a unique splice junction spanning amino acid sequence with high computed binding affinity to two common MHC 1 alleles. (*D*) The LSR mRNA isoform ucnyp002.3 contains a unique splice junction spanning amino acid sequence that is expected to reside in the extracellular domain of this plasma membrane protein and that also contains subsequences that are computed to have moderate to high binding affinity to multiple common MHC 1 alleles. Thus, the single amino acid sequence is amenable to two therapeutic modalities.

The CD9 isoform identified herein, which was expressed in 100% of the late-stage 296 TCGA tumors and in 10 of the 12 tumors (Fig. 4), contains a unique exon (Fig. 5*B*) that imparts upon the protein a unique, in-frame 41-aa sequence that encompasses the first two transmembrane regions of the protein and the extracellular domain between them—making it amenable to specific antibody targeting if expressed.

Isoforms predicted to encode epitopes for tumor vaccines. Although the C-terminal portion of the normal-restricted ETV4 isoform identified herein is incompletely known, the portion that is known reveals the isoform to have an exon-skipping event that is unique among all ETV4 isoforms—conferring on the resulting protein at least 14 unique amino acids (Fig. 5*C*). We analyzed the epitope potential of this region using a computational method (34) that has been recently validated by retrospective prediction against a large set of bona fide T-cell antigens that induced immune responses and were associated with tumor regression and long-term disease stability (35). We identified a 10-mer epitope centered directly over the unique splice junction and calculated it to have a very strong affinity (12.9 nM) for the HLA allele A*02:01 and a moderate affinity (363 nM) for the B*08:01 allele. Because the A*02:01 and B*08:01 alleles are among the most common HLA alleles in the Caucasian population of the United States (36), the ETV4 isoform is a strong candidate for immunotherapeutic application for ovarian cancer.

Discussion

We have developed a highly customized RNA-seq bioinformatics pipeline that is designed for isoform identification and that is distinct from standard approaches because of (*i*) its use of an isoform model database that is a merger of all isoform model databases available worldwide, (*ii*) its capability for maximally sensitive genome-wide read alignment, and (*iii*) the nucleotide resolution consistency analysis that is performed for every sequencing read–isoform combination. Furthermore, we developed a workflow for high-throughput, isoform-level RT-qPCR experiments that is distinguished by custom software for automated design of PCR primers that are specific to individual mRNA isoforms at complex genomic loci (i.e., loci in which no isoform may even have a uniquely distinguishing splice junction or exon). Both the RNA-seq pipeline and RT-qPCR infrastructure have been developed to the point of being highly automated, requiring for a new cancer type modest manual involvement and timeframes (~6–7 wk) to generate the level of analysis reported herein. We used our combined computational/experimental pipeline to generate detailed molecular hypotheses in the form of specific molecules (i.e., mRNA isoforms and/or the protein isoforms that they encode) with ovarian tumor-specific expression and with particular oncologic application(s). Importantly, the hypotheses we generated would not have been possible with gene-level analyses that by definition encompass numerous mRNA and protein isoforms in aggregate.

Analogous to the challenge of distinguishing driver from passenger mutations in cancer genomics (37), cancer transcriptomics must contend with the challenge of distinguishing those mRNA molecules that are important for the malignant phenotype from those that are not. We addressed this challenge by requiring the mRNA isoforms interrogated in this study to be expressed in 90–100% of the TCGA ovarian tumors, with the rationale being that a tumor-specific isoform that is present in most tumors is likely to be functionally important rather than due to a deregulation side effect. In support of this rationale, among the topmost 5% ($n = 33$) tumor-specific or normal-restricted isoforms are variants of genes that are demonstrated oncogenes, known to maintain the malignant state, have a direct role in driving aggressive tumor initiating cell behavior, or are necessary for maintaining a stem-cell phenotype. In addition to the cancer genomics goal of identifying driver mutations is the goal of identifying driver mutations that are "actionable." The RT-qPCR experiments revealed 15 mRNA isoforms that have the tumor specificity required for an early detection diagnostic of ovarian cancer. Additionally, at least 5 of the 33 mRNA isoforms confirmed by RT-qPCR to have tumor-specific or normal-restricted expression encode protein targets that have unique primary structures that would allow them to be specifically targeted by one or more therapeutic strategies, including monoclonal antibody therapy/chimeric T-cell generation, and peptide- or T-cell–based vaccines.

Beyond protein, mRNA itself has the potential to be a therapeutic target (38, 39). If proven to be so, mRNA has a great advantage over protein as a class of target molecule because MHC epitope and cell surface restrictions would not apply. However, like protein therapeutics, mRNA would need to be targeted isoform-specifically because of the high degree of identical nucleotide sequence among the isoforms from a genomic locus. Our study is pertinent to mRNA therapeutics because we demonstrate a feasible strategy for finding tumor-specific mRNA targets.

Herein we have proposed the idea—inspired by a DNA-based approach (25)—of an ovarian cancer detection test based on the detection of tumor-specific mRNA isoforms from malignant cells that have disseminated to the cervix and been collected during a Papanicolaou test. A strategy based on RNA and not DNA could have distinct advantages. Tumor types have characteristic expression profiles that are distinctive from both those of other tumor types and normal tissues. An approach based on RNAs that are broadly indicative of characteristic expression programs could be more robust because it would not rely on particular mutations but on a characteristic cancer cell expression phenotype. Furthermore, because somatic DNA mutations occur in one or a few copies per tumor cell and RNA isoforms can occur in hundreds to thousands of copies per cell, an assay based on mRNA is potentially much more sensitive. The first requirement for such a test is the enumeration of mRNA molecules that indicate the presence of an ovarian tumor. In our experiments, we identified isoforms that were expressed in most or all tumors and were not detected in any normal tissues. Furthermore, we identified additional isoforms that were expressed in most or all tumors and in only one normal tissue that, importantly, was not ovary or fallopian tube. These additional isoforms are also candidates for a detection test because, not being found in the gynecologic tissues tested, they would be indicative of tumor cells if detected in a Papanicolaou test.

There are a number of hard limitations to the approach for tumor-specific isoform identification and validation. These hard limitations are due to the "short read" nature of RNA-seq data and to the great extent to which mRNA isoforms at a genomic locus share exons and splice junctions. RNA-seq reads represent, essentially, 200–250 contiguous base pairs of processed mRNA. Because most mRNAs are much longer than 250 bp, RNA-seq reads cannot provide the information that links distant exons and that is often necessary for unambiguous identification of the source mRNA isoform. Our RNA-seq computational procedure was designed for maximum accuracy in identifying those isoforms that were, and were not, represented in an RNA-seq dataset. To achieve this goal, we minimized false negatives by merging all of the major isoform model databases and then developed nucleotide-level correspondence and parsimony algorithms to minimize false positives. Nonetheless, determining which isoforms generated a set of RNA-seq reads is an inference problem that will always be error-prone and because of this no isoform identification procedure will be completely accurate. However, even if one were able to identify the mRNA isoforms underlying an RNA-seq dataset with complete accuracy, there is a severe limitation on the rate at which their expression can be confirmed by PCR. To confirm an mRNA isoform one must design PCR primers that amplify a uniquely distinguishing nucleotide sequence. At complex genomic loci this is a very challenging

task because of the extent to which exons and splice junctions are shared among isoforms. A major component of our study is the algorithms that we developed for automated design of isoform-specific PCR primers. Even with our specialty software we found that we could only design primers for ~55% of isoforms, meaning that almost half of the isoforms that we predicted by RNA-seq to be tumor-specific could not be investigated by RT-qPCR. Furthermore, for ~25% of the isoforms for which we could design primers, melt curve analysis revealed the presence of multiple PCR products (often two or three)—indicating the presence of new isoforms. These observations are compatible with recent transcriptome sequencing experiments that have reported on new isoform discovery rates (12, 40, 41). That RT-qPCR discovers isoforms at a higher rate attests to its higher sensitivity and lack of library preparation procedures.

As opposed to the hard limitations that exist for our approach, there are three "soft" limitations that could be readily addressed to potentially improve our tumor-specific isoform identification rate. First, we used only two metrics to rank-prioritize isoforms by likelihood of being tumor-specific. The output of our RNA-seq computational procedures has six metrics. Additionally, our procedures have three threshold values that have not been optimized. We expect that the use of more or other metrics for rank prioritization and of optimized threshold values will yield additional results of the same qualitative nature as reported herein. Second, ovary and fallopian tube were the most common normal tissues in which isoforms were expressed (Fig. 4). As the tissue of origin and primary tumor site, these are exactly the normal tissues in which a tumor-expressed isoform is most likely to be expressed. Unfortunately, these are also exactly the normal tissues for which we had the fewest normal control RNA-seq datasets (three ovary and one fallopian tube). Thus, our ability to negatively filter tumor-expressed isoforms was limited. The GTEx project is actively sequencing ovary and fallopian tube, so this soft limitation will diminish in the near future. Third, we did not account for the known expression subtypes of HGS-OvCa (42–44), but instead sought mRNA isoforms that were expressed in all tumor subtypes (i.e., 90–100% of the 296 TCGA tumors). Incorporating subtype classification into our procedures could yield tumor subtype-specific mRNA isoforms.

We note that additional experiments will be required for the proposed applications of the tumor-specific isoforms that we identified. Tumor cells that disseminate to the cervix or into the bloodstream may down-regulate the isoforms that are expressed in primary tumors, so for utility in a Papanicolaou test-based early detection diagnostic or in identifying circulating tumor cells the continued expression of isoforms in these nonprimary tumor sites will need to be confirmed. Additionally, mRNA expression does not always equate to protein expression, so for the protein isoforms with therapeutic target potential their expression and cellular localization in tumor cells will need to be experimentally confirmed.

In summary, we have developed and rigorously evaluated a systematic process for identifying tumor-specific mRNA isoforms that leverages the large and growing public databases of tumor and normal tissue RNA-seq data. We have quantified the rate at which tumor-specific isoforms can be identified for HGS-OvCa and have demonstrated that they have the potential to provide the specificity needed for extremely specific diagnostics and therapeutics. Our findings are relevant in a larger context because the procedures we developed can be readily and rapidly applied to any of the 30 or more tumor types for which large amounts of RNA-seq data now exist.

Methods and Materials

RNA-seq Bioinformatics. We created a custom isoform model database by merging the six major isoform model databases available worldwide. We used the set of all isoform splice junctions from our custom database and lenient parameterizations to perform highly sensitive genome-wide alignment of RNA-seq paired-end reads. We then performed an alignment-filtering step to remove spurious alignments that can be generated by using lenient parameterization. To filter, we analyzed each read pair alignment to determine whether or not its implied cDNA fragment was a contiguous subsequence of any mRNA isoform(s). We then use the filtered read alignments to compute the subset of our custom isoform model database that most parsimoniously accounted for the filtered alignments. In effect, we created a tailored isoform model database for each RNA-seq dataset. Finally, we converted read pair genome alignments to transcriptome alignments and explicitly used the strict correspondence between read pairs and isoforms to compute isoform-level expression.

RT-qPCR. We performed RT-qPCR experiments according to Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines (45), which among other criteria include the use of multiple references for intersample comparison and the calculation of PCR efficiencies for quantification. Tumor RNA was obtained from the University of California, San Diego Moores Cancer Center Biorepository and commercially (Origene). Normal tissue RNA was obtained commercially (Biochain and Origene).

Further details are provided in *SI Methods and Materials*.

ACKNOWLEDGMENTS. The Iris and Matthew Strauss Center for the Early Detection of Ovarian Cancer and Colleen's Dream Foundation Kicking for the Dream supported this study. RT-qPCR was conducted at the Institute for Genomic Medicine Genomics Center of the University of California, San Diego, with support from NIH–National Cancer Institute Grant P30CA023100.

1. Farhangfar CJ, Meric-Bernstam F, Mendelsohn J, Mills GB, Lucio-Eterovic AK (2013) The impact of tumor heterogeneity on patient treatment decisions. *Clin Chem* 59(1):38–40.
2. Swanton C (2012) Intratumor heterogeneity: Evolution through space and time. *Cancer Res* 72(19):4875–4882.
3. Marusyk A, Almendro V, Polyak K (2012) Intra-tumour heterogeneity: A looking glass for cancer? *Nat Rev Cancer* 12(5):323–334.
4. Adamia S, et al. (2014) A genome-wide aberrant RNA splicing in patients with acute myeloid leukemia identifies novel potential disease markers and therapeutic targets. *Clin Cancer Res* 20(5):1135–1145.
5. Lupetti R, et al. (1998) Translation of a retained intron in tyrosinase-related protein (TRP) 2 mRNA generates a new cytotoxic T lymphocyte (CTL)-defined and shared human melanoma antigen not expressed in normal cells of the melanocytic lineage. *J Exp Med* 188(6):1005–1016.
6. Rousseaux S, et al. (2013) Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci Transl Med* 5(186):186ra66–186ra66.
7. Coulie PG, Van den Eynde BJ, van der Bruggen P, Boon T (2014) Tumour antigens recognized by T lymphocytes: At the core of cancer immunotherapy. *Nat Rev Cancer* 14(2):135–146.
8. David CJ, Manley JL (2010) Alternative pre-mRNA splicing regulation in cancer: Pathways and programs unhinged. *Genes Dev* 24(21):2343–2364.
9. Venables JP, et al. (2009) Cancer-associated regulation of alternative splicing. *Nat Struct Mol Biol* 16(6):670–676.
10. Annala MJ, Parker BC, Zhang W, Nykter M (2013) Fusion genes and their discovery using high throughput sequencing. *Cancer Lett* 340(2):192–200.
11. Lonsdale J, et al.; GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45(6):580–585.
12. Mercer TR, et al. (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* 30(1):99–104.
13. Wu P-Y, Phan JH, Wang MD (2013) Assessing the impact of human genome annotation choice on RNA-seq expression estimates. *BMC Bioinformatics* 14(Suppl 11):S8.
14. Chvatal VA (1979) A greedy heuristic for the set-covering problem. *Math Oper Res* 4(3):233–235.
15. Weatheritt RJ, Gibson TJ (2012) Linear motifs: Lost in (pre)translation. *Trends Biochem Sci* 37(8):333–341.
16. Oh S, Shin S, Janknecht R (2012) ETV1, 4 and 5: An oncogenic subfamily of ETS transcription factors. *Biochim Biophys Acta* 1826(1):1–12.
17. Teh M-T (2012) FOXM1 coming of age: Time for translation into clinical benefits? *Front Oncol* 2:146.
18. Lam AKY, et al. (2013) FOXM1b, which is present at elevated levels in cancer cells, has a greater transforming potential than FOXM1c. *Front Oncol* 3:11.
19. Hemler ME (2014) Tetraspanin proteins promote multiple cancer stages. *Nat Rev Cancer* 14(1):49–60.
20. Zhao W, Ji X, Zhang F, Li L, Ma L (2012) Embryonic stem cell markers. *Molecules* 17(6):6196–6236.
21. Hwang JR, et al. (2012) Upregulation of CD9 in ovarian cancer is related to the induction of TNF-α gene expression and constitutive NF-κB activation. *Carcinogenesis* 33(1):77–83.
22. Leth-Larsen R, et al. (2012) Functional heterogeneity within the CD44 high human breast cancer stem cell-like compartment reveals a gene signature predictive of distant metastasis. *Mol Med* 18:1109–1121.

23. Reaves DK, et al. (2014) The role of lipolysis stimulated lipoprotein receptor in breast cancer and directing breast cancer cell behavior. *PLoS ONE* 9(3):e91747.
24. Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490(7418):61–70.
25. Kinde I, et al. (2013) Evaluation of DNA from the Papanicolaou test to detect ovarian and endometrial cancers. *Sci Transl Med* 5(167):167ra4.
26. Lee Y, et al. (2007) A candidate precursor to serous carcinoma that originates in the distal fallopian tube. *J Pathol* 211(1):26–35.
27. O'Shannessy DJ, et al. (2013) Gene expression analyses support fallopian tube epithelium as the cell of origin of epithelial ovarian cancer. *Int J Mol Sci* 14(7): 13687–13703.
28. Kim J, et al. (2012) High-grade serous ovarian cancer arises from fallopian tube in a mouse model. *Proc Natl Acad Sci USA* 109(10):3921–3926.
29. Kessler M, Fotopoulou C, Meyer T (2013) The molecular fingerprint of high grade serous ovarian cancer reflects its fallopian tube origin. *Int J Mol Sci* 14(4):6571–6596.
30. Marquez RT, et al. (2005) Patterns of gene expression in different histotypes of epithelial ovarian cancer correlate with those in normal fallopian tube, endometrium, and colon. *Clin Cancer Res* 11(17):6116–6126.
31. Sproul D, et al. (2012) Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. *Genome Biol* 13(10):R84.
32. Ge X, et al. (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics* 86(2):127–141.
33. Dobolyi A, Dimitrov E, Palkovits M, Usdin TB (2012) The neuroendocrine functions of the parathyroid hormone 2 receptor. *Front Endocrinol (Lausanne)* 3:121.
34. Nielsen M, et al. (2007) NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS ONE* 2(8):e796.
35. Fritsch EF, et al. (2014) HLA-binding properties of tumor neoepitopes in humans. *Cancer Immunol Res*, 10.1158/2326-6066.CIR-13-0227.
36. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR (2011) Allele frequency net: A database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res* 39(Database issue):D913–D919.
37. Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res* 39(17):e118.
38. Zangi L, et al. (2013) Modified mRNA directs the fate of heart progenitor cells and induces vascular regeneration after myocardial infarction. *Nat Biotechnol* 31(10): 898–907.
39. Zhou J, Shum K-T, Burnett JC, Rossi JJ (2013) Nanoparticle-based delivery of RNAi therapeutics: Progress and challenges. *Pharmaceuticals (Basel)* 6(1):85–107.
40. Lin Y, et al. (2012) An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res* 40(17):8460–8471.
41. Howald C, et al. (2012) Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res* 22(9):1698–1710.
42. Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* 474(7353):609–615.
43. Tothill RW, et al.; Australian Ovarian Cancer Study Group (2008) Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res* 14(16):5198–5208.
44. Verhaak RGW, et al.; Cancer Genome Atlas Research Network (2013) Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J Clin Invest* 123(1): 517–525.
45. Bustin SA, et al. (2009) The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* 55(4):611–622.
46. Mailman MD, et al. (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 39(10):1181–1186.
47. Thierry-Mieg D, Thierry-Mieg J (2006) AceView: A comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* 7(Suppl 1):S12.1–S12.14.
48. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35(Database issue):D61–D65.
49. Hsu F, et al. (2006) The UCSC known genes. *Bioinformatics* 22(9):1036–1046.
50. Harrow J, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* 22(9):1760–1774.
51. Yamasaki C, et al.; Genome Information Integration Project And H-Invitational 2 (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res* 36(Database issue):D793–D799.
52. Trapnell C, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511–515.
53. Dobin A, et al. (2012) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1): 15–21.
54. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) *Introduction to Algorithms* (MIT Press, Cambridge, MA).
55. Roberts A, Pachter L (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* 10(1):71–73.
56. Untergasser A, et al. (2012) Primer3—new capabilities and interfaces. *Nucleic Acids Res* 40(15):e115.
57. Qu W, et al. (2012) MFEprimer-2.0: A fast thermodynamics-based program for checking PCR primer specificity. *Nucleic Acids Res* 40(Web Server issue):W205–W208.
58. Dwight Z, Palais R, Wittwer CT (2011) uMELT: Prediction of high-resolution melting curves and dynamic melting profiles of PCR products in a rich web application. *Bioinformatics* 27(7):1019–1020.
59. Savitzky A, Golay MJE (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem* 36(8):1627–1639.
60. Rao X, Lai D, Huang X (2013) A new method for quantitative real-time polymerase chain reaction data analysis. *J Comput Biol* 20(9):703–711.
61. Hellemans J, Mortier G, De Paepe A, Speleman F, Vandesompele J (2007) qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol* 8(2):R19.