Published OnlineFirst May 5, 2015; DOI: 10.1158/1078-0432.CCR-14-3173

Biology of Human Tumors

Clinical Cancer Research

Creation of a Human Secretome: A Novel Composite Library of Human Secreted Proteins: Validation Using Ovarian Cancer Gene Expression Data and a Virtual Secretome Array

Vinod Vathipadiekal[1,2], Victoria Wang[2,3], Wei Wei[1,2], Levi Waldron[2,3,4], Ronny Drapkin[5], Michael Gillette[1,6], Steven Skates[2,7], and Michael Birrer[1,2]

Abstract

Purpose: To generate a comprehensive "Secretome" of proteins potentially found in the blood and derive a virtual Affymetrix array. To validate the utility of this database for the discovery of novel serum-based biomarkers using ovarian cancer transcriptomic data.

Experimental Design: The secretome was constructed by aggregating the data from databases of known secreted proteins, transmembrane or membrane proteins, signal peptides, G-protein coupled receptors, or proteins existing in the extracellular region, and the virtual array was generated by mapping them to Affymetrix probeset identifiers. Whole-genome microarray data from ovarian cancer, normal ovarian surface epithelium, and fallopian tube epithelium were used to identify transcripts upregulated in ovarian cancer.

Results: We established the secretome from eight public databases and a virtual array consisting of 16,521 Affymetrix U133 Plus 2.0 probesets. Using ovarian cancer transcriptomic data, we identified candidate blood-based biomarkers for ovarian cancer and performed bioinformatic validation by demonstrating rediscovery of known biomarkers including CA125 and HE4. Two novel top biomarkers (FGF18 and GPR172A) were validated in serum samples from an independent patient cohort.

Conclusions: We present the secretome, comprising the most comprehensive resource available for protein products that are potentially found in the blood. The associated virtual array can be used to translate gene-expression data into cancer biomarker discovery. A list of blood-based biomarkers for ovarian cancer detection is reported and includes CA125 and HE4. FGF18 and GPR172A were identified and validated by ELISA as being differentially expressed in the serum of ovarian cancer patients compared with controls. *Clin Cancer Res; 21(21); 4960–9. ©2015 AACR.*

Introduction

Epithelial ovarian cancer affects 23,000 women resulting in approximately 15,500 deaths in the United Sates per year (1). Because of the lack of symptoms of early-stage disease, approximately 75% of ovarian cancer patients present with disease involving the upper abdomen (FIGO stage III/IV) and only 30% of these patients survive 5 years beyond their diagnosis (2). In contrast, when ovarian cancer is diagnosed in the early stage, the prognosis is excellent with 5-year survival exceeding 90%. Hence, identification of early detection biomarkers specific for ovarian cancer could have a significant impact on mortality from ovarian cancer.

CA125 is the most widely studied serum biomarker for ovarian cancer. Although screening studies with CA125 tests, interpreted with a single threshold or serially, followed by ultrasound scans for women with a positive test have shown excellent specificity, the utility of CA125 as a biomarker for the early detection of ovarian cancer remains unproven largely due to its unknown sensitivity for early-stage disease in asymptomatic subjects. CA125 is not elevated in almost 50% of clinically detected stage I ovarian cancers (3) and is not expressed in approximately 20% of ovarian cancer (3). Specificity in the largest target population, postmenopausal women, is very high even though CA125 can also be elevated in common benign conditions, including uterine fibroids, benign ovarian tumors, pelvic endometriosis, follicular cysts, and cystadenoma, as these conditions are far more common in premenopausal women. And although CA125 is also elevated in women with other cancers such as pancreatic, and breast and lung metastatic to the peritoneum, the incidence of these malignancies is similar to ovarian cancer (4). HE4 is another biomarker that is increased in the serum of women with ovarian cancer (5). It has less sensitivity than CA125, although its specificity may be greater as HE4 appears to be less influenced by benign conditions.

[1]Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts. [2]Harvard Medical School, Boston, Massachusetts. [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts. [4]City University of New York, New York, New York. [5]Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts. [6]Broad Institute of MIT and Harvard, Cambridge, Massachusetts. [7]Biostatistics Unit, Massachusetts General Hospital, Boston, Massachusetts.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Current address for L. Waldron: Hunter College School of Urban Public Health, City University of New York, New York.

Corresponding Author: Michael Birrer, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114. Phone: 617-726-8624; Fax: 617-724-6898; E-mail: MBIRRER@mgh.harvard.edu

doi: 10.1158/1078-0432.CCR-14-3173

Published OnlineFirst May 5, 2015; DOI: 10.1158/1078-0432.CCR-14-3173

Translational Relevance

Large-scale genomic projects are providing extensive data on aberrant gene expression in different epithelial cancers. A major challenge is translating these data into clinically useful applications. We report a comprehensive "Secretome" and virtual array to support the identification of candidate blood-based biomarkers by using differential gene expression. Although we have demonstrated utility in the context of ovarian cancer transcriptional array data, this database can be applied to any cancer for which there are adequate gene-expression profiles of tumor and its normal counterpart (including RNAseq profiles) whereas the associated virtual array especially facilitates array-based experiments. Application of the "secretome array" to ovarian cancer transcriptome databases has rediscovered known biomarkers and identified novel candidates. This approach can accelerate biomarker discovery, leveraging genomic data to provide enriched candidate lists of potential blood-based proteins.

Little information on HE4's performance in prospective screening trials exists as only a single small screening trial has reported (6). To date, the evidence points to no single cancer biomarker being sufficiently sensitive for early-stage disease in asymptomatic women to meet the stringent criteria necessary as a first line test for the early detection of ovarian cancer. Additional serum biomarkers to detect ovarian cancer not expressing CA125 or to detect it earlier than CA125 are needed to identify a screening test that detects the full spectrum and earliest stages of the disease.

Transcriptomics has been widely used to identify differentially expressed genes and molecular signatures in many biologic processes (7–13). Transcription profiling studies have also been used to predict patients' survival in ovarian cancer (14). In addition, gene-expression changes and other genomic alterations can be correlated on a global level. Hence, transcriptomics can be extended to identify different types of biomarkers in many human cancers. However, sets of differentially expressed genes provide no intrinsic information about which are most likely to be reflected in the circulation. The identification of blood-based biomarkers would be greatly facilitated by a generic platform that identifies genes encoding proteins potentially found in the blood.

The "Secretome" and associated virtual array established in this study provides a platform for the identification of blood-based biomarkers for high-grade, advanced-stage serous ovarian tumors. The gene reference set and array are whole genome based, the latter using a commercially available expression platform that can be applied to any cancer for which there is adequate transcriptome data. As a proof of principle, we used expression profiling data generated from high-grade, advanced-stage serous ovarian cancer patient samples, normal ovarian surface epithelium (OSE), and normal fallopian tube epithelium (FTE). To prioritize candidates, we introduced a pathway-based biomarker identification approach as relevant secretome proteins might be interconnected with intracellular signaling pathways. These blood-based proteins were further filtered based upon their expression in normal organs and tissues. Our approach identified both established high-grade serous ovarian cancer biomarkers (including CA125 and HE4) and novel candidates. Two new markers (FGF18 and GPR172A) were validated at the mRNA levels using independent sets of microarrays and on the protein level in an independent cohort of serum samples.

Materials and Methods

Generation of secretome array

The secretome was generated from eight databases, including secreted protein database (SPD), Uniprot secreted proteins, Signal Peptide Website (An Information Platform for Signal Sequences and Signal Peptides), Zhang database, GPRCDB (A Molecular-Specific Information System for G Protein-Coupled Receptors), and AmiGO (the Gene Ontology database; see Table 1 for details). Within each secretome source database, only human-specific proteins were searched. To create the virtual array, the gene identifiers provided by each database were mapped to Affymetrix human genome U133 Plus 2.0 probeset identifiers. BioMART-ENSEMBLE GENES 63, Homo sapiens genes GRCh37.p3 (http://www.biomart.org, June, 2011), and DAVID v6.7 (http://david.abcc.ncifcrf.gov/conversion.jsp, Sept. 21, 2011) were used to generate the identifier maps. Identifiers from each database were checked against each map, and the map containing the highest fraction of these identifiers was used. The DAVID map combined conversions for all Affymetrix 3′ arrays, so these were further narrowed down to 133 Plus 2.0 maps only, using the hgu133plus2.db Bioconductor package (v. 2.4.5). By this method, all identifiers were mapped to zero, one, or more Affymetrix probeset identifiers in one step. For each unique probeset identifier, a record was kept of which databases identified it, and of the original gene or protein identifiers mapped to it. All computations were performed in the R statistical environment v. 2.12.1 (R development Core Team, 2010).

Microarray data normalization and class comparison

We used two independently generated gene-expression datasets. Dataset A consisted of 10 microdissected ovarian cancer samples and 10 microdissected normal fallopian tube samples (15). Dataset B consisted of 53 microdissected ovarian tumor samples and 10 normal OSE samples. All the samples were profiled using Affymetrix Human Genome U133 Plus 2.0 arrays. The CEL files were background-corrected, normalized, and summarized using RMA (Bioconductor package affy) for the two datasets separately. Summarized expression data were filtered to contain probesets that are in the secretome. Differential gene-expression analysis between cancer and normal samples was carried out using LIMMA (Bioconductor package limma).

Pathway-based approach for biomarker discovery

PathwayStudio (Elsevier) software was used to identify biomarkers related to biologic pathways. This software uses a protein interaction database derived from the entire Medline abstract database. This type of analysis explores the global and systemic properties of the underlying molecular networks of the biomarker list generated for ovarian cancer and enables interpretation of the biologic significance of the gene list. This annotation can be used to prioritize candidate biomarkers for validation. The biomarker lists were imported to Pathway Studio and initially the algorithm "Find direct interactions" was used. This algorithm assembled a network of the molecules directly interacting in the imported gene list and allowed no additional objects to be added to the network. Subsequently, all the "direct interactions group" probesets were

Published OnlineFirst May 5, 2015; DOI: 10.1158/1078-0432.CCR-14-3173

Table 1. Databases used for secretome array generation

Database	No. of IDs	Map used	No. of Mapped	Source
spd	5,715	david_uniprotID	3,083	http://spd.cbi.pku.edu.cn/
Clark et al.	1,047	mart_genbank	946	http://genome.cshlp.org/content/13/10/2265/suppl/DC1
Diehn et al.	1,552	mart_unigene	1,013	Diehn et al. (ref. 20; Fig. 4)
Diehn et al.	842	mart_uniprotGeneName	678	Diehn et al. (ref. 20; Fig. 5)
Diehn et al.	285	mart_unigene	178	Diehn et al. (ref. 20; Fig. 6)
uniprot_secreted1	2,645	mart_uniprotAccession	2,402	http://www.uniprot.org/uniprot/?querysecreted+AND +organism%3A%22Homo+sapiens+%5B9606%5D%22&sortscore
Signal	500	david_uniprotID	447	http://www.signalpeptide.de/index.php
Zhang	3,243	mart_uniprotID	2,976	http://proline.bic.nus.edu.sg/spdb/download.html
gpcr.org_structure	212	mart_pdb	155	http://www.gpcr.org/7tm/?wicket:bookmarkablePage=:nl.ru.cmbi.mcsis.web.pages.proteinstructure.ProteinStructureOverviewPage
gpcr.org_family	1,333	mart_uniprotID	649	http://www.gpcr.org/7tm/proteinfamily/
AmiGO GO:0016020	10,868	mart_uniprotAccession	7,419	http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/gene-associations/gene_association.goa_human.gz?rev=HEAD
AmiGO GO:0005576	2,425	mart_uniprotAccession	1,900	http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/gene-associations/gene_association.goa_human.gz?rev=HEAD
Total	3,0667		21,846	

extracted and analyzed using the Fisher exact test to identify the statistically enriched pathway associated biomarkers.

Cancer selective expression approach to prioritize the candidate biomarkers

To identify markers that were uniquely expressed in ovarian tumors as opposed to genes that are ubiquitously expressed in many normal tissues or organs we used The Gene Expression Barcode resource (16). This database provides absolute measures of expression for the most annotated genes in essentially all normal tissue types and organs. This resource leverages information from the GEO and Array Express public repositories to build statistical models that convert data from a single microarray into expressed/unexpressed calls for each gene. The output of the algorithm is expressed as an average of 1's (expressed) or 0's (unexpressed) across tissues and is in the range of zero to one.

Validation of FGF18 and GPR172A

An independent gene-expression dataset (GEO Accession number: GSE26712) was used for bioinformatic validation. The dataset GSE26712 consists of 185 primary ovarian tumors and 10 normal OSE samples profiled using Affymetrix human U133A microarray (17).

Serum collection, FGF18 and GPR172A ELISA

Control ($n = 20$) and ovarian cancer patients' serum samples ($n = 20$, referred as "case" samples hereafter) were obtained from a previously published study by Early Detection Research Network (18) and the Department of Pathology, Massachusetts General Hospital respectively. Serum samples of both cohorts were collected following the same procedure as previously described (18). The control samples were collected from healthy, postmenopausal Caucasian women without apparent neoplastic disease and without active non-neoplastic disease (18). The "case" samples were collected from postmenopausal Caucasian women with high-grade, advanced-stage, serous ovarian cancer. FGF18 and GPR172A were quantified at protein level using a sandwich enzyme immunoassay technique. The ELISA kits were commercially obtained from My Biosource (FGF18 catalog number: MBS912811 and GPR172A catalog number: MBS702260) and used as per the manufacturer's instructions. All reagents used were supplied in the Kit. Each ELISA plate was read at 450 nm with the correction wavelength set at 540 nm.

Results

Generation of secretome array

Genomic technologies provide a unique opportunity to globally identify potential candidate genes whose proteins could serve as plasma or serum markers of cancers. Numerous studies and databases report known and/or predicted secreted proteins, using complementary methodologies, which can aid such efforts. However, each resource is likely to be both incomplete and to contain incorrect predictions; in addition they are provided in inconsistent formats and use a variety of protein and gene identifiers. Thus, we collected and synthesized several databases from published papers providing lists of secreted proteins, transmembrane or membrane proteins, signal peptides, G-protein coupled receptor, or proteins existing in the extracellular region.

Our bioinformatics approach had three major phases as shown in Fig. 1. Phase I consisted of collecting the protein and gene information from different data sources (Table 1). All the databases and published articles were available in the public domain, and we searched only for human-specific proteins. Briefly, SPD has a collection of secreted proteins from Human, Mouse, and Rat proteomes, which also includes sequences from SwissProt, Trembl, Ensembl, and Refseq. We extracted 5715 UNIPROT ID entities from this database. Clark and colleagues (19) reported a database, Secreted Protein Discovery Initiative, for secreted and transmembrane proteins, which contained 1047 transcripts representing 1021 genes.

Diehn and colleagues (20) generated a database on membrane-secreted proteins, and expression of membrane-secreted genes in human malignancies and normal tissues. We used this data source to extract information for membrane/secreted proteins associated with human malignancies and normal tissues (1,552), tumor markers (842) and organ specific injury molecules (285). "UniProt" was used to extract 2645 UniprotID entities using the search terms "secreted" and organism: "*Homo sapiens*."

The website http://www.signalpeptide.de contains proteins with signal sequences and signal peptides grouped into Mammalia, Drosophila, Viruses, and Bacteria. Using the "advanced search" section of the website, we extracted all proteins for the organism "*Homo sapiens*," which identified 500 UniprotID entities. We also used signal peptide database (Zhang), which contains signal sequences for different species such as archae,

Published OnlineFirst May 5, 2015; DOI: 10.1158/1078-0432.CCR-14-3173

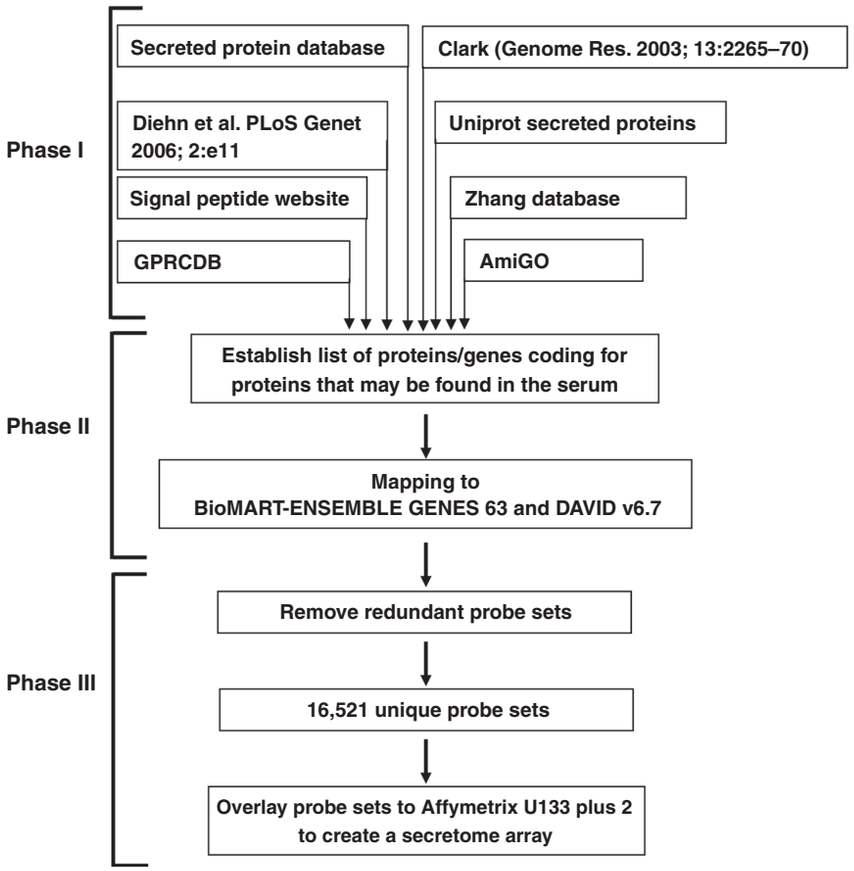

Figure 1.
Schematic overview of generation of Secretome Array.

prokaryotes, and eukaryotes. The search criteria used were Sequence type as "Signal peptide (DNA)" and narrowed down to the organism "*Homo sapiens*," which resulted in the identification of 3243 uniprotID entities.

G-protein–coupled receptors (GPCR) constitute a large and diverse family of proteins whose primary function is to transduce extracellular stimuli into intracellular. GPRCDB is a comprehensive database that stores large amounts of heterogeneous data on GPCRs. We downloaded all annotated protein structures and Class A Rhodopsin-like families from the GPCRDB, respectively, providing 212 PDB and 1333 UniprotID entities. The gene ontology (Homo sapiens Revision 1.9) provided 10,868 Uniprot Accession IDs associated with the GO terms for membrane and its related terms, and 2425 Uniprot Accession IDs associated with Extracellular Region and its related terms (see Supplementary Methods for details). This phase integrated data from diverse sources to create a comprehensive set of identifiers corresponding to potentially secreted human proteins, supported by varying numbers of sources.

Mapping of the secretome onto Affymetrix array

In phase II, we mapped the identifiers provided by the secretome generation database sources to Affymetrix probeset IDs. We checked each set of original identifiers against two competing resources for identifier mapping: BioMART-ENSEMBLE GENES 63 and DAVID v6.7, and used whichever allowed mapping to a greater number of the identifiers (Table 1). In phase III, we assembled the probesets and removed all redundant ones to generate the unique secretome virtual array. This identified 16,521 unique Affymetrix U133 plus 2 probesets. For each probeset identifier, a record was kept of which databases identified it, and which original gene or protein identifiers were mapped to it, providing provenance and a means to assess confidence in each probeset by the number of databases identifying it. Supplementary Table S1 provides detailed information on our secretome array. Out of 16,521 probesets in the Secretome Array, only 6 probesets were identified by all eight databases, 43 by 7 of 8 and increasing numbers for each smaller number of source databases. The 6 probesets with the highest score of eight corresponding to the genes *SFRP2*, *SEMA3F*, and *PGF*. SFRP2 gene encodes a member of the SFRP family that contains a cysteine-rich domain homologous to the putative Wnt-binding site of Frizzled proteins. SFRPs act as soluble modulators of Wnt signaling and it is a secreted protein (21). SEMA3F (semaphorins) are a family of proteins that are involved in signaling (22). PGF (placental growth factor) gene encodes a growth factor found in placenta, which is homologous to vascular endothelial growth factor and it is a secreted protein (23). All these genes are known secreted proteins.

Identification of potential secreted biomarkers for ovarian cancer using the secretome

We applied the secretome virtual array to gene-expression databases of ovarian cancer and normal controls to identify genes that are differentially expressed. Figure 2 demonstrates the schematic overview of our approach to identify detection biomarkers

Published OnlineFirst May 5, 2015; DOI: 10.1158/1078-0432.CCR-14-3173

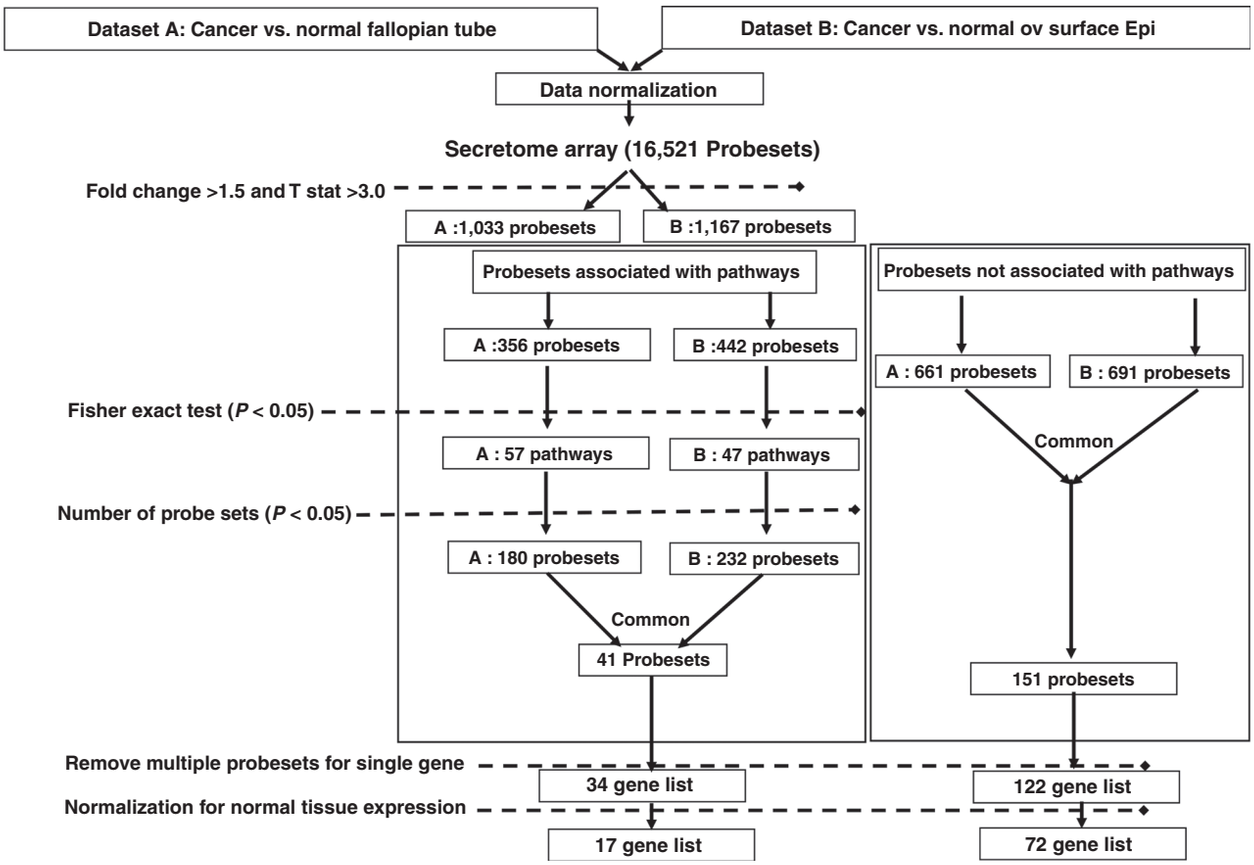

Figure 2.
Schematic overview of ovarian cancer biomarker discovery.

candidates for ovarian cancer. On the basis of the recent evidence on alternative sites of the origin for ovarian cancer, we generated two different gene lists for biomarkers based on the comparison between high-grade serous ovarian cancer gene-expression data with, both normal fallopian tube and normal OSE (24, 25). All expression data were generated from microdissected tissue samples. A two-group comparison was conducted using LIMMA to generate a list of differentially expressed probesets between tumor and normal tissue. This list was filtered through the secretome array (16,521 probesets), which yielded 1033 probesets (fold change $\geq$1.5, T >3.0) that were upregulated in the ovarian cancer in compared with normal FTE (List A in Fig. 2; List 1 in Supplementary Table S4). Independent analysis using a cohort of ovarian cancer expression data compared with normal OSE identified 1,167 upregulated probesets in cancer (List B in Fig. 2; List 2 in Supplementary Table S4).

Bioinformatic validation of the secretome array

To validate the secretome array, we searched our differentially expressed gene list for previously characterized ovarian cancer biomarkers. Review of the literature revealed six biomarkers that have been described as potential blood-based biomarkers for ovarian cancer. These include CA125 (3), HE4 (26), prostasin (27), osteopontin (28), VEGF (29), and IGFBP2 (30). All of these proteins were found to be statistically significantly overexpressed in cancer compared with normal epithelium in our list (Table 2).

Filtering of gene lists based upon pathway association

To filter our gene list further, we used pathway identification. Our goal for this analysis was to use the secretome data to gain functional insights pertaining to the roles of these proteins in biologic processes and use that information to better prioritize the

Table 2. Validation of secretome array by previously identified potential serum-based biomarkers

Gene name	Probe set	Fold change cancer vs. FTE	T Stat	Fold change cancer vs. OSE	T stat
Cancer antigen 125 (CA-125)	220196_at	2.09	2.36	2.8	3.15
Human epididymis protein 4 (HE4)	203892_at	10.5	11.27	1.4	1.38
Prostasin	202525_at	2.78	4.78	2.65	5.0
Osteopontin	1568574_x_at	1.56	5.5	1.98	3.2
Vascular endothelial growth factor (VEGF)	212171_x_at	1.55	5.3	2.45	4.6
Insulin-like growth factor–binding protein-2	202718_at	1.67	1.62	5.1	3.8

Published OnlineFirst May 5, 2015; DOI: 10.1158/1078-0432.CCR-14-3173

lists. The biomarker gene list A (1033 probesets, fold change ≥1.5, T >3.0) was imported to PathwayStudio software. The algorithm identified 356 probesets, which network for the molecules that are directly interacting. Subsequently, a Fisher exact test was used to identify pathways that are statistically enriched in the 356 probesets identifying 57 pathways ($P < 0.05$) involving 180 unique probesets. The pathways identified include Focal Adhesion Regulation, VEGFR → NFATC signaling, GFR → NCOR2 signaling, FGFR → RUNX2 signaling, among others; a detailed list is provided in Supplementary Table S2. Similar analysis was carried out on gene list 2 (1167 probesets, fold change ≥1.5, T > 3.0) for identifying biomarkers associated with pathways. The algorithm identified 422 probesets, which identify molecules that are directly interacting. The Fisher exact test using 422 probesets identified 47 pathways ($P < 0.05$) involving 232 unique probesets. The pathways identified include Frizzled R → CTNNB signaling, Activin R → SMAD2/3 signaling, VasopressinR2 → CREB/ELK-SRF/AP-1/EGR signaling, EDG3/5 → AP-1/ELK-SRF signaling, and Notch → TCF3 signaling. A detailed list is provided in Supplementary Table S3. We hypothesized that genes found in both lists would reflect more profound biology and that a common list would give us more robust biomarkers for validation. This analysis identified 41 probesets (List 3) common in Lists 1 and 2 (Supplementary Table S4).

This algorithm allowed us to generate a gene list, associated with direct interacting networks and biologic pathways. Alternatively, there may be molecules that are not found within pathways (due to the lack of intervening genes on the primary lists), which still have the potential to serve as ovarian cancer biomarkers. These molecules will be filtered out in the above analysis. Hence, we generated a second gene list using probesets not associated with pathways. There were 691 and 661 probesets filtered out from Lists 1 and 2, respectively. We searched for common probesets present in these filtered out probesets and identified 151 unique probeset biomarker list (List 4; Supplementary Table S5).

A cancer-selective expression approach to prioritize candidate biomarkers

To identify biomarkers most likely to be uniquely elevated in the blood of ovarian cancer patients, we filtered our lists according to gene expression in normal tissues or organs and prioritized those with low expression in normal organs and tissues. The Gene Expression Barcode Resource Database provides absolute measures of expression for all the probesets in a variety of tissues and organs. We extracted the expression data for each gene in our gene lists from The Gene Expression Barcode Resource Database against 38 normal tissues and organs, including the major organs such as liver, kidney, ovary, spleen, thyroid, and lungs. We averaged the expression levels for all normal tissues to a single reference value (average gene-expression bar code) for each probeset, expressed these in the range 0 to 1. A number close to 1 indicates that the probeset is almost certainly expressed in the tissue and a number of 0 or close to 0 indicates that the probeset is probably not expressed in that tissue. Figure 3 displays the expression pattern of probesets for Lists 3 and 4 on all major normal tissues and organs evaluated. In the 41 gene lists, two probesets (*BAK1* and *TSPAN17*) had a gene-expression bar code of zero signifying there was no detectable expression in any of the normal tissues

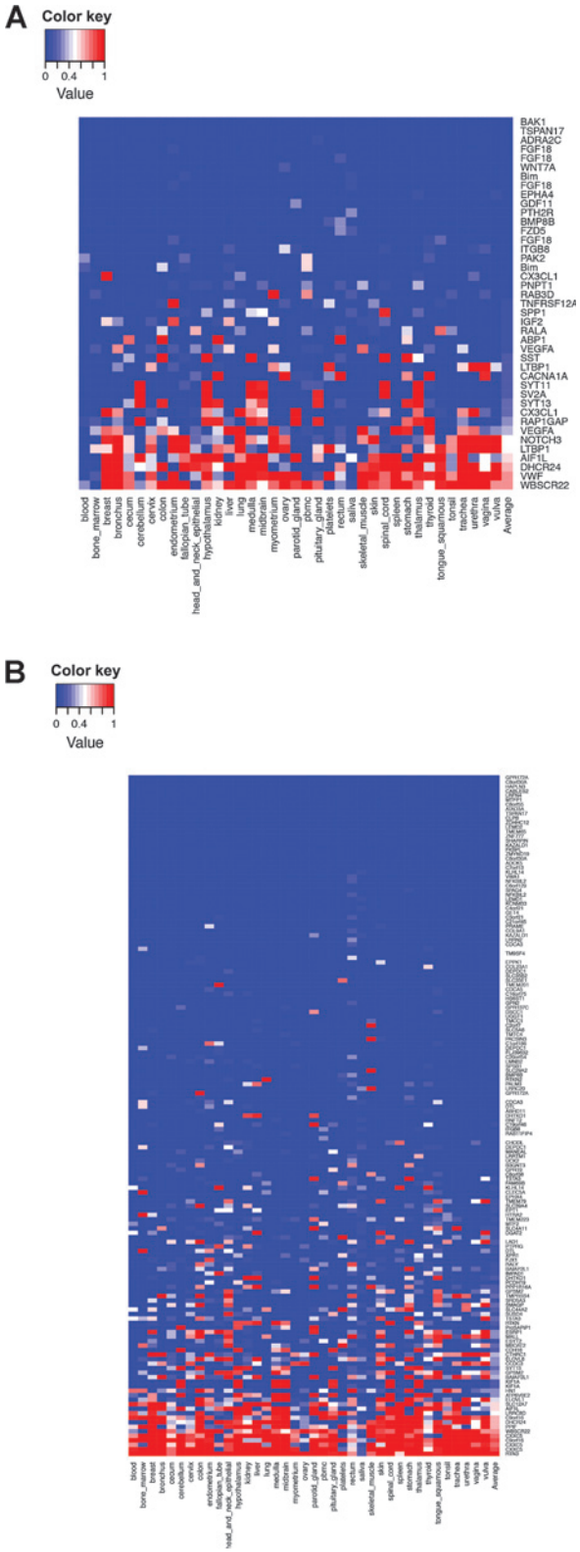

Figure 3.
Heatmap demonstrating expression level of each probesets in the list 3 (A) and list 4 (B) in various normal organs and tissues: Blue represents that the probeset has low or no expression and red represents that the probeset is almost certainly expressed in that tissue type.

Published OnlineFirst May 5, 2015; DOI: 10.1158/1078-0432.CCR-14-3173

evaluated. Twenty-four probesets have a gene-expression bar code <0.10 which includes four probesets corresponding to the *FGF18* gene. In the 151 gene list, 27 probesets such as *GPR172A*, *C8orf30A*, *HAPLN3*, *CABLES2*, *LRFN4*, *MTFP1*, *C8orf55*, *ATAD3A*, *TSPAN17*, *CLPB*, *ZDHHC12*, *LEMD2*, *TMEM65*, *ZNF777*, *SHARPIN*, *KAZALD1*, *FKBPL*, *ZMYND19*, *C8orf30A*, *ADCK5*, *C7orf13*, *KCNMB3*, *GET4*, *C21orf45*, *TM9SF4*, *SLC35B2*, and *HS6ST1* had a bar code of zero, whereas 104 probesets had a bar code <0.10.

We hypothesized that a blood-based biomarker should be expressed at relatively low levels in normal tissues to improve the background to tumor ratio. Hence, we ranked the biomarker lists based on their low-expression gene-expression bar code in normal tissues and organs. We found that *FGF18* had four probesets with high expression in cancer and low expression in normal tissues and organs (low average gene-expression bar code) from List 3. *GPR172A* is the gene that had two probesets with high expression in cancer and low expression in normal tissues. We selected both these molecules for validation.

Independent validation of candidate biomarkers for ovarian cancer

To validate our top biomarker candidates (FGF18 and GPR172A), we looked at the mRNA expression level of these molecules in ovarian cancer specimens using a publically available independent gene-expression database (GSE26712). Both molecules were found to be overexpressed in serous ovarian tumor samples ($n = 185$) in comparison with their expression in normal OSE ($n = 10$) in a statistically significant manner ($P < 0.001$; Fig. 4A and B). Because these candidate biomarkers were identified using the secretome array, we expected to find them in the blood. We tested serum samples from women with advanced-stage ovarian cancer by ELISA for blood levels of FGF18 and GPR172A and compared that with normal age–matched controls. In the ovarian cancer group, levels of both molecules were found to be significantly increased in comparison with the control group. Serum FGF18 level increased 1.9-fold in the ovarian cancer group in comparison with the control group ($P < 0.0001$, Fig. 4C). There was a 2.9-fold increase in GPR172A serum level observed in the ovarian cancer group in comparison with the control group

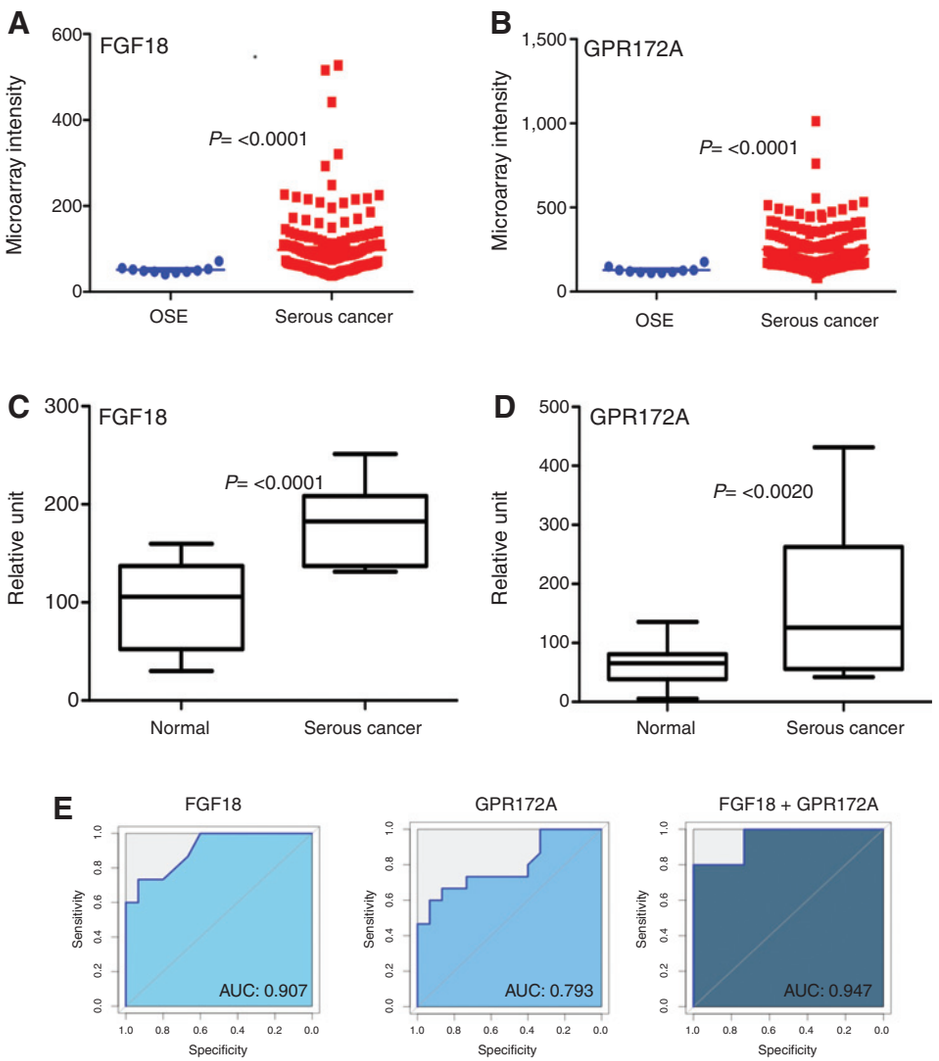

Figure 4.
Validation of the candidate biomarkers in the secretome array. A and B, mRNA expression levels of FGF18 (A) and GPR172A (B) in the validating dataset GSE26712. Graph shows microarray gene-expression intensity of FGF18 and GPR172 in ovarian cancer and normal controls. C and E, ELISA assays for FGF18 and GPR172A. The graph shows the ELISA assay performed for FGF18 (C) and GPR172A (D) proteins on serum samples of ovarian cancer patients and normal controls. The data are shown as mean ± SEM. The sensitivity and specificity of FGF18 and GPR172A as indicator of sample original (normal or ovarian cancer) were calculated by a receiver operating characteristic (ROC) curve (E).

Published OnlineFirst May 5, 2015; DOI: 10.1158/1078-0432.CCR-14-3173

($P = 0.0020$, Fig. 4D). In addition, FGF18, GRP172A and the accumulation of both markers present decent sensitivity and specificity to indicate the origin of serum specimens (normal or ovarian cancer) in the cohort used for ELISA (Fig. 4E). These results demonstrate the potential value of the secretome array in translating genomic data into the discovery of blood-based biomarkers. Furthermore, FGF18 and GPR172A appear to be novel biomarkers for ovarian cancer and warrant further evaluation.

Discussion

The secretome is a subset of the proteome consisting of proteins secreted by living cells through signal peptide, exosome or proteins shed from the surface of living cells. These proteins constitute an important class of molecules, encoded by approximately 10% of the human genome (31). Proteins of the secretome have been demonstrated to play important roles in tumorigenesis, and are therefore of increasing interest as a means to identify and characterize potential diagnostic and prognostic biomarkers, as well as therapeutic targets. Blood-based biomarkers are particularly useful as they are easy to obtain and quantify. Thus, finding novel methods to more efficiently identify potential blood-based markers is a critical need. High-throughput technologies based on genomic and transcriptomic data represent a huge yet still underexplored resource for biomarker discovery in cancer. A platform/algorithm that can use genomic data to more effectively identify blood-based candidate biomarkers would be of great value to the biomarker development community and facilitate clinical translation of large-scale molecular profiling experiments.

By constructing a secretome through gene ontology, we aimed to provide a powerful tool for body-fluid (e.g., serum and ascites)–based biomarker discovery in cancer. The secretome database described in this article was generated by systematic review of all relevant publically available databases and publications (Table 1). One of the challenges was to select the appropriate databases and ensure that all genes that encode proteins ultimately found in the blood were included. We opted to be inclusive even at the expense of having genes that encode proteins rarely found in the blood, anticipating that subsequent filtering and prioritization could be performed after the differential expressed genes were determined. To this end, we used several gene ontology groups interrogating all possible mechanisms for extracellular protein release. It is important to note that our secretome array consequently includes several membrane, cytoplasmic, mitochondrial, and even nuclear proteins that do not have distinct extracellular release mechanisms according to prediction algorithms. These proteins have been *de facto* detected through LC/MS-based proteomic studies and included into databases such as the Human Plasma PeptideAtlas (32). This effort has, therefore, ensured the inclusion of potential biomarkers with atypical mechanisms of secretion. The final "Secretome Virtual Array" assembled by integrating this information provided a list of 16,521 Affymetrix probesets representing transcriptomics data for secreted proteins.

One potential limit of LC/MS-based proteimic biomarker identification is the low sensitivity to detect proteins with low-expression levels. Conversely, the recently developed mass spectrometry for reliable quantification of analytes of low abundance such as parallel reaction monitoring (PRM) or selected reaction monitoring (SRM)–based mass spectrometry requires a predefined set of peptides/proteins before investigation. Similar situation lies in affinity reagent–based proteomic methodologies, which grant high sensitivity and accuracy to screen significant amount of specimens, but only for limited number of targets. Our database provides an opportunity to investigators to reduce the shotgun proteomics related complexity in biomarker discovery by focusing on proteins prescreened through large-scale genomic studies. Different screening criteria can be used alone or integrated, including differential gene-expression and gene ontology analysis used in this study. This could include prognostic/survival impact and DNA copy number to select candidate proteins with biologic significance, and thus high potential for further evaluation using body fluid samples. Of importance, the universal application of the secretome should be noted as it can be used with array-based data (as seen in this study), but also RNAseq and even whole-genome sequencing. By minimizing the number of putative markers to a more manageable scale, low-throughput yet highly sensitive approaches such as multiplex ELISA, antibody array, reverse-phase protein array (RPPA) or targeted mass spectrometry–based assays can be applied to larger numbers of biologic samples as robust tools to identify novel biomarkers, which have low abundance and are less likely to be detected by LC/MS-based approaches. To facilitate this process, we have established a public website where appropriate tools are available to allow the use of the secretome by clinical researchers without a bioinformatics background.

As proof of principle, a comprehensive list of genes with altered expression in ovarian cancer was applied to the "secretome array." This list was generated through comparing the transcriptome of laser-capture microdissected (LCM) ovarian tumors and two possible origins of ovarian epithelial cancer, OSE, and FTE (33) from healthy donors. Considering the heterogeneity of normal and neoplastic tissues of ovary and fallopian tubes, the utilization of microdissected specimen minimized the interference from nonepithelial cells and improved the accuracy of differential expression analysis. The validation of the secretome array was 2-fold: (i) multiple previously reported ovarian cancer biomarkers, including the highly credentialed markers CA125 and HE4, are found on the list, and (ii) demonstration of increased expression of two new markers (FGF18 and GPR172A) in the blood of ovarian cancer patients compared with normal women. These latter two markers were chosen based upon a strongly positive discovery signal and the availability of commercial grade ELISA assays. This two-tier validation provides a compelling evidence that the secretome array can assist in the identification of circulating biomarkers from genomically derived candidate lists.

It is well accepted that the neoplastic secretome actively controls various stages of carcinogenesis such as tumor initiation, differentiation, invasion, metastasis, and angiogenesis. The activation of specific oncogenic pathways driven by secreted proteins makes them potential therapeutic targets against tumor progression. Considering this, we introduced a "Pathway-Based" approach to investigate our secretome gene list. We identified 180 and 232 biomarker probesets reflecting established that were upregulated in ovarian cancer compared with normal OSE and normal fallopian tube epithelial cells, respectively. FGF18, one of the circulating biomarkers identified in this study, has recently been demonstrated to have prognostic significance in ovarian cancer. Functional studies of FGF18 have further revealed its role in ovarian tumorigenesis as well as its oncogenic influence on ovarian tumor vasculature and tumor-associated macrophages (34). Several therapeutic approaches against FGF signaling have been developed, including receptor

Published OnlineFirst May 5, 2015; DOI: 10.1158/1078-0432.CCR-14-3173

tyrosine kinase inhibitors, receptor-neutralizing antibodies, and pan-FGF ligand traps (35), making FGF18 inhibition a potential therapeutic option for patients with high circulating FGF18 level.

We acknowledge that the transcriptome-based discovery of secreted biomarkers may not be comprehensive. First, the success of this approach requires correlation between levels of transcript and of the corresponding protein something that has been demonstrated to hold for only a subset of genes. Thus, it is likely that the secretome will include false positives (transcript increased but not protein) and miss other circulating proteins (transcript unchanged but protein increased). Second, the secretome and associated array cannot predict the change of extracellular biomarker levels due to posttranslational cleavage or altered cellular transportation activity (release rate). Third, our approach did not evaluate the contribution of tumor stroma to the ovarian tumor secretome. This problem can be solved by mapping our secretome database to a differentially expressed gene list generated by comparing microdissected ovarian tumor stroma and normal ovarian stroma. Finally, it is worth to note that our small-scale ELISA validation study has validated the elevated expression of FGF18 and GPR174A *in sera* from high-grade serous ovarian cancer patients, but is not sufficient to prove them as markers to predict ovarian cancer. For diagnostic ovarian cancer biomarker discovery and validation, it will be essential to use more stringent, larger-scale studies using serum specimens from benign gynecologic disease as controls. Nevertheless, the construction of the "secretome virtual array" provides valuable resources for screening any particular gene list for proteins that are more likely to be found in the circulation. Limited number of candidate biomarkers identified through the secretome array would be suitable for measurement in plasma/serum from cases and controls using sensitive, accurate, and highly specified technologies such as ELISA, RPPA, and PRM- or SRM-based targeted mass spectrometry. The power of secretome array is further amplified by its general applicability (to any gene-expression database of any derivation), ease of application, and ability to provide multiple levels of filtering. This approach provides a means to translate the large publically available genomic databases into subsets of candidate markers likely to be found in the circulation, accelerating the vital work of identifying clinically relevant blood-based biomarkers.

Disclosure of Potential Conflicts of Interest

S. Skates is a consultant/advisory board member for Abcodia. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: V. Vathipadiekal, W. Wei, M. Birrer
Development of methodology: V. Vathipadiekal, L. Waldron, M. Birrer
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): R. Drapkin, M. Birrer
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): V. Vathipadiekal, V. Wang, W. Wei, L. Waldron, R. Drapkin, M. Gillette, S. Skates, M. Birrer
Writing, review, and/or revision of the manuscript: V. Vathipadiekal, W. Wei, L. Waldron, R. Drapkin, M. Gillette, S. Skates, M. Birrer
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): V. Vathipadiekal, M. Birrer
Study supervision: M. Birrer
Other (contributed to all aspects of this project): M. Birrer

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 8, 2014; revised March 18, 2015; accepted April 19, 2015; published OnlineFirst May 5, 2015.

References

1. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. CA Cancer J Clin 2012;62:10–29.
2. Siegel R, DeSantis C, Virgo K, Stein K, Mariotto A, Smith T, et al. Cancer treatment and survivorship statistics, 2012. CA Cancer J Clin 2012;62:220–41.
3. Jacobs I, Bast RC Jr. The CA 125 tumour-associated antigen: a review of the literature. Hum Reprod 1989;4:1–12.
4. Sjovall K, Nilsson B, Einhorn N. The significance of serum CA 125 elevation in malignant and nonmalignant diseases. Gynecol Oncol 2002;85:175–8.
5. Chudecka-Glaz A, Rzepka-Gorska I, Wojciechowska I. Human epididymal protein 4 (HE4) is a novel biomarker and a promising prognostic factor in ovarian cancer patients. Eur J Gynaecol Oncol 2012;33:382–90.
6. Hertlein L, Stieber P, Kirschenhofer A, Furst S, Mayr D, Hofmann K, et al. Human epididymis protein 4 (HE4) in benign and malignant diseases. Clin Chem Lab Med 2012;50:2181–8.
7. Barlund M, Forozan F, Kononen J, Bubendorf L, Chen Y, Bittner ML, et al. Detecting activation of ribosomal protein S6 kinase by complementary DNA and tissue microarray analysis. J Natl Cancer Inst 2000;92:1252–9.
8. Wang T, Hopkins D, Schmidt C, Silva S, Houghton R, Takita H, et al. Identification of genes differentially over-expressed in lung squamous cell carcinoma using combination of cDNA subtraction and microarray analysis. Oncogene 2000;19:1519–28.
9. Xu J, Stolk JA, Zhang X, Silva SJ, Houghton RL, Matsumura M, et al. Identification of differentially expressed genes in human prostate cancer using subtraction and microarray. Cancer Res 2000;60:1677–82.
10. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer [see comments]. Nat Genet 1996;14:457–60.
11. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680–6.
12. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
13. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.
14. Mok SC, Bonome T, Vathipadiekal V, Bell A, Johnson ME, Wong KK, et al. A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: microfibril-associated glycoprotein 2. Cancer Cell 2009;16:521–32.
15. Verhaak RG, Tamayo P, Yang JY, Hubbard D, Zhang H, Creighton CJ, et al. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. J Clin Invest 2013;123:517–25.
16. McCall MN, Uppal K, Jaffee HA, Zilliox MJ, Irizarry RA. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. Nucleic Acids Res 2011;39:D1011–5.
17. Bonome T, Levine DA, Shih J, Randonovich M, Pise-Masison CA, Bogomolniy F, et al. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. Cancer Res 2008;68:5478–86.
18. Skates SJ, Horick NK, Moy JM, Minihan AM, Seiden MV, Marks JR, et al. Pooling of case specimens to create standard serum sets for screening cancer biomarkers. Cancer Epidemiol Biomarkers Prev 2007;16:334–41.
19. Clark HF, Gurney AL, Abaya E, Baker K, Baldwin D, Brush J, et al. The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: a bioinformatics assessment. Genome Res 2003;13:2265–70.

20. Diehn M, Bhattacharya R, Botstein D, Brown PO. Genome-scale identification of membrane-associated human mRNAs. PLoS Genet 2006;2:e11.
21. Melkonyan HS, Chang WC, Shapiro JP, Mahadevappa M, Fitzpatrick PA, Kiefer MC, et al. SARPs: a family of secreted apoptosis-related proteins. Proc Natl Acad Sci U S A 1997;94:13636–41.
22. Kolodkin AL, Matthes DJ, Goodman CS. The semaphorin genes encode a family of transmembrane and secreted growth cone guidance molecules. Cell 1993;75:1389–99.
23. Maglione D, Guerriero V, Viglietto G, Ferraro MG, Aprelikova O, Alitalo K, et al. Two alternative mRNAs coding for the angiogenic factor, placenta growth factor (PlGF), are transcribed from a single gene of chromosome 14. Oncogene 1993;8:925–31.
24. Levanon K, Crum C, Drapkin R. New insights into the pathogenesis of serous ovarian cancer and its clinical impact. J Clin Oncol 2008;26:5284–93.
25. Karst AM, Drapkin R. Ovarian cancer pathogenesis: a model in evolution. J Oncol 2010;2010:932371.
26. Hellstrom I, Raycraft J, Hayden-Ledbetter M, Ledbetter JA, Schummer M, McIntosh M, et al. The HE4 (WFDC2) protein is a biomarker for ovarian carcinoma. Cancer Res 2003;63:3695–700.
27. Mok SC, Chao J, Skates S, Wong K, Yiu GK, Muto MG, et al. Prostasin, a potential serum marker for ovarian cancer: identification through microarray technology. J Natl Cancer Inst 2001;93:1458–64.
28. Kim JH, Skates SJ, Uede T, Wong KK, Schorge JO, Feltmate CM, et al. Osteopontin as a potential diagnostic biomarker for ovarian cancer. JAMA 2002;287:1671–9.
29. Oehler MK, Caffier H. Diagnostic value of serum VEGF in women with ovarian tumors. Anticancer Res 1999;19:2519–22.
30. Flyvbjerg A, Mogensen O, Mogensen B, Nielsen OS. Elevated serum insulin-like growth factor-binding protein 2 (IGFBP-2) and decreased IGFBP-3 in epithelial ovarian cancer: correlation with cancer antigen 125 and tumor-associated trypsin inhibitor. J Clin Endocrinol Metab 1997;82:2308–13.
31. Pavlou MP, Diamandis EP. The cancer cell secretome: a good source for discovering biomarkers? J Proteomics 2010;73:1896–906.
32. Farrah T, Deutsch EW, Aebersold R. Using the Human Plasma PeptideAtlas to study human plasma proteins. Methods Mol Biol 2011;728:349–74.
33. Kindelberger DW, Lee Y, Miron A, Hirsch MS, Feltmate C, Medeiros F, et al. Intraepithelial carcinoma of the fimbria and pelvic serous carcinoma: evidence for a causal relationship. Am J Surg Pathol 2007;31:161–9.
34. Wei W, Mok SC, Oliva E, Kim SH, Mohapatra G, Birrer MJ. FGF18 as a prognostic and therapeutic biomarker in ovarian cancer. J Clin Invest 2013;123:4435–48.
35. Turner N, Grose R. Fibroblast growth factor signalling: from development to cancer. Nat Rev Cancer 2010;10:116–29.

Published OnlineFirst May 5, 2015; DOI: 10.1158/1078-0432.CCR-14-3173

Clinical Cancer Research

Creation of a Human Secretome: A Novel Composite Library of Human Secreted Proteins: Validation Using Ovarian Cancer Gene Expression Data and a Virtual Secretome Array

Vinod Vathipadiekal, Victoria Wang, Wei Wei, et al.

Clin Cancer Res 2015;21:4960-4969. Published OnlineFirst May 5, 2015.