DATABASE
The Journal of Biological Databases and Curation

Database, 2015, 1–8
doi: 10.1093/database/bav051
Database tool

Database tool

HCSD: the human cancer secretome database

Amir Feizi[1,2], Amir Banaei-Esfahani[1] and Jens Nielsen[1,2,3,*]

[1]Novo Nordisk Foundation Center for Biosustainability, Department of Biology and Biological Engineering, Chalmers University of Technology, SE-41296 Göteborg, Sweden, [2]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Fremtidsvej 3, DK-2970 Hørsholm, Denmark and [3]Novozymes A/S, Krogshoejvej 36, 2880 Bagsvaerd, Denmark

*Corresponding author: Tel: +46 31 772 3854; Fax: +46 31 772 3801; Email: nielsenj@chalmers.se

Citation details: Feizi,A., Banaei-Esfahani,A. and Nielsen,J. HCSD: the human cancer secretome database. *Database* (2015) Vol. 2015: article ID bav051; doi:10.1093/database/bav051

Received 23 October 2014; Revised 29 April 2015; Accepted 4 May 2015

Abstract

The human cancer secretome database (HCSD) is a comprehensive database for human cancer secretome data. The cancer secretome describes proteins secreted by cancer cells and structuring information about the cancer secretome will enable further analysis of how this is related with tumor biology. The secreted proteins from cancer cells are believed to play a deterministic role in cancer progression and therefore may be the key to find novel therapeutic targets and biomarkers for many cancers. Consequently, huge data on cancer secretome have been generated in recent years and the lack of a coherent database is limiting the ability to query the increasing community knowledge. We therefore developed the Human Cancer Secretome Database (HCSD) to fulfil this gap. HCSD contains >80 000 measurements for about 7000 nonredundant human proteins collected from up to 35 high-throughput studies on 17 cancer types. It has a simple and user friendly query system for basic and advanced search based on *gene name*, *cancer type* and *data type* as the three main query options. The results are visualized in an explicit and interactive manner. An example of a result page includes annotations, cross references, cancer secretome data and secretory features for each identified protein.

Database URL: www.cancersecretome.org.

Introduction

Cancer is currently seen as a cluster of complicated diseases with increasing prevalence globally (1). Understanding and curing cancer have entered a new phase with the advent of next generation sequencing and advanced proteomics (2). In particular, recent advances in both accuracy and scale of the measurements in proteomics using label-based methods (such as SILAC and iTRAQ) have revolutionized oncoproteomics (3). The cancer secretome, as a newly established subdiscipline of oncoproteomics, involves the detection, quantification and characterization of the secreted proteins (such as cytokines, growth factors etc.), shedome (shed receptors and proteases) and extracellular matrix components of a given type of cancer cell at a specific time point (4, 5). Many secreted proteins are linked to the hallmarks of cancer which are reliant on cell–cell adhesion and signaling (6, 7). Much analysis supports how these proteins in the tumor microenvironment control and regulate the cancer cell invasion and metastasis (8–11). Along with this, soluble factors in cancer secretome are promising for novel biomarkers and therapeutic targets for different types of cancers (6, 12–16). Accordingly, there has been increasing number of studies to analyze the cancer secretome resulting in rapid growth in data generation. For example, Wu and coworkers identified candidate serological biomarkers for various cancer types based on secretome analysis of 23 cancer cell lines (17). From 4584 nonredundant proteins identified in these cancer cell lines, they suggested between 6 and 137 marker candidates selective for each tumor type and 94 potential pan-cancer markers (proteins secreted by most cancer cell lines) and they verified several of the identified protein biomarkers (17). There are many other examples of the same kind of studies that have provided large amounts of data to be publically available (18–20). However, the lack of a specific database for cancer secretome data challenges researchers in the field to query community knowledge in terms of the time and efficiency. Therefore, designing a systematic and organized database to manage large volumes of unstructured cancer secretome data is in demand. To fulfil this important gap, we designed the Human Cancer Secretome Database (HCSD), a dynamic database with interactive web interface that provides the researchers with the opportunity to explore their protein of interest against the publicly available data on the human cancer secretome. HCSD has a simple and user-friendly query system for basic and advanced searches based on *gene name*, *data type*, and *cancer type* as the three main query options. The result pages are explicit and intractable. An example result page includes annotations, cross references, cancer secretome data and secretory features for each protein. Developing HCSD is an important bioinformatics solution to boost research in cancer secretome and tumor microenvironment.

Materials and methods

Data collection and preprocessing

To collect all relevant data from high-quality publications, a comprehensive literature survey was done searching the *Scopus* and *PubMed* database starting with the general keyword ‘cancer secretome’. To avoid accumulation of the false identifications that is frequent in proteomics data, we applied stringent selection criteria to filter out publications including: (i) to have standard workflow of one of the shotgun proteomics techniques (with biological and/or technical replications). (ii) Detailed description for each steps of the experimental design. (iii) Providing of all the parameters used in database searching and corresponding bioinformatics analysis. (iv) Having error estimation strategy (such as FDR). (v) Performing molecular/clinical validation experiment for the identified biomarkers. (vi) Providing supplementary detail information tables for identified portions in peptide and protein level. Applying these criteria total 35 high-throughput publications were selected as data source to collect the relevant data (see DATA SET menu in the web page).

A major concern of any proteomic study is the FDR (Flase discovery rate) control to prevent from inflation of false identifications. To obtain reliable results, 1% FDR should be applied on peptide and protein levels. When merging distinct datasets, which were analyzed separately, one has to take special care to avoid inflation of the FDR. This has been previously done by Schaab *et al.* (21). However, to apply such techniques in merging proteomics datasets, corresponding *P values* for the reported fold changes are necessary. Unfortunately, most of the available data sets have not included the raw data or the *P* values in their released data sets. This was a big challenge in designing HCSD, and to overcome to that, we therefore carefully collected the data based on the cut-off FDR reported in each paper and we did exclude all the proteins above used the cut-off. Also, the query results designed to be based on each study so the user can compare the results from different studies on a particular protein of interest and decide based on major votes. In line with this, we also provided more technical information on identification such as PSMs (peptide-spectrum matches) and the number of the unique peptides match in the results pages for each query. Doing this while it is not yet possible to fully resolve the inflation of false identification resulting from combining independent studies, the results will be reported study wise so the user can assess the reliability of the results by checking other supportive information from each study.

Next, the publications were categorized as *label-free* and *label-based* studies based on the proteomics techniques have been used to quantify the proteome. In *label-free* proteomics, the secretome of a specific cancer type is quantified without using a stable isotope containing compound and the peptide abundance is quantified by spectral counting. On the other hand, in *label-based* proteomics, stable isotope(s) is used for labeling and quantification of the peptides in the comparable samples. The *label-based* methods are less sensitive to the experimental biased than *label-free* methods (22, 23). Therefore, this categorization helps user to compare the results from two type of technology. Because of the difference in publishing the data from one paper to another, retrieving and processing data tables from various studies and merging them into a single database structure was time consuming. Missing information and the format of released data (PDF format) were also problematic in data collecting step. As most of the studies only report their data based on gene symbol or protein ID, for ID mapping, we used bioDBnet (24) to make data be searchable using different IDs in gene and protein levels. For each protein in the database the annotation data extracted from UniProt (25), ensemble (26) and Entrez (27) (Figure 1). An exclusive link for each record is provided to direct the user to its HPA cancer atlas (28) page. The HPA page provides the user with antibody-based protein profiling information for the protein of interest in 20 most common cancers. This allows user to compare the expression status of the collected proteins in HCSD based on quantitative methods against antibody-based staining data in HPA.

Databasing and interface design

We used a MYSQL relational database (version 5.5.8) to design and query HCSD database. The web interface implemented by *webpy* (www.webpy.org), a python based web framework. The *web.py* is in the public domain and it has been used by Google App Engine. To our knowledge; this is the first implementation of it in designing a bioinformatics database.It is as powerful as *Django* (https://www.

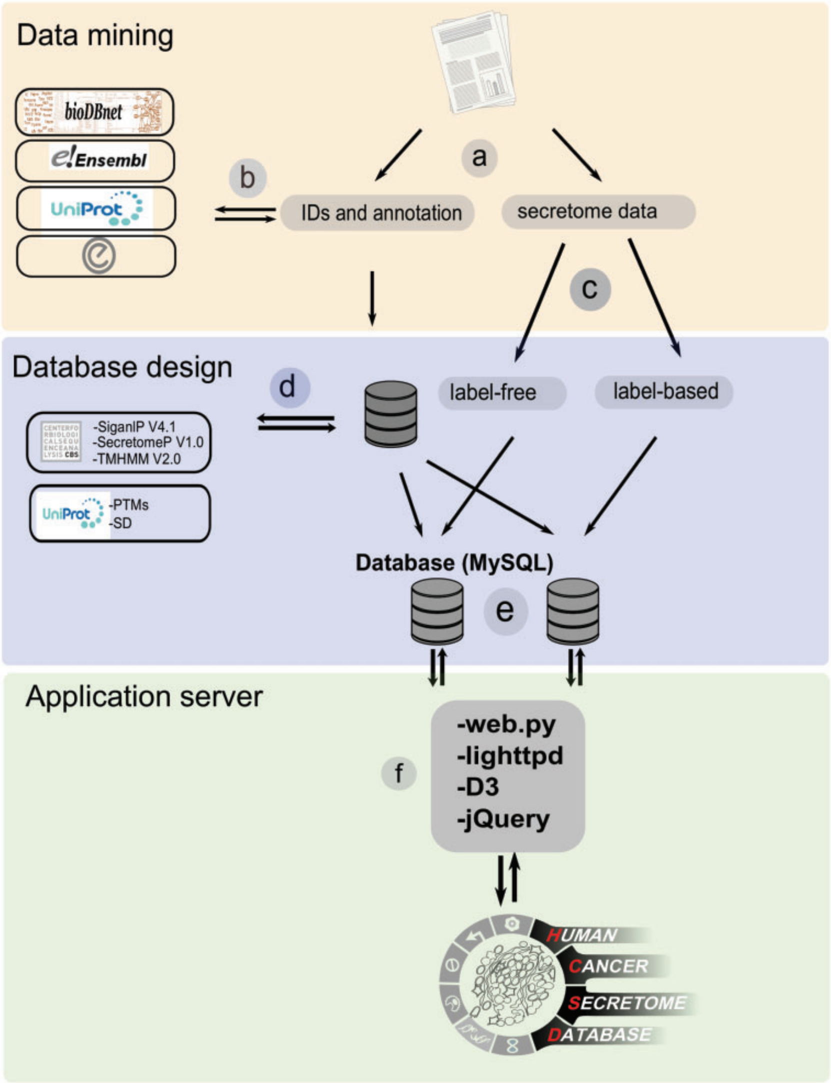


Figure 1. The workflow of HCSD design. (**a**) Appling the selection criteria, first all the cancer secretome data were collected and processed from literatures. (**b**) Then, I all the complementary annotation and cross references were obtainedfrom UniProt, Ensembl, bioDBnet and Entrez using thethe reported protein or gene IDs in the data tables. (**c**) Next, the secretory pathway features including signal peptide, transmembrane domains and nonclassic secretory proteins were predicted using CBS prediction servers (32–34). The secondary structures and PTMs information were retrieved from UniProt (35). (**d**) Based on the proteomics strategy used, the secretome data were divided into label-free and label-based studies. (**e**) The structured data tables used as input to MySQL to generate searchable data tables by end user (http://www.mysql.com/). (**f**) For the web server lighttpd is used to query the database (http://www.lighttpd.net/) and the web application and the interface were implemented using web.py (www.webpy.org), Javascript (http://en.wikipedia.org/wiki/JavaScript), jQuery (http://jquery.com/) and D3 (http://d3js.org/).

djangoproject.com/) while it is much simpler to implement. The *lighttpd* (http://www.lighttpd.net/) were used as a fast and open source web server which security, speed, compliance and flexibility are all its characteristics comparing to other competitors. In the query page, two dynamic and searchable tables for *label-free* and *label-based* data were designed using DataTables plug-in of jQuary (https://www.datatables.net/). The result pages benefit from high quality visualization techniques to present the cancer secretome data and secretory features. For visualization, Javascripts and D3 (www.d3.org) were used upon *web.py*. HCSD is available at www.cancersecretome.org.

Querying HCSD

In order to query HCSD data, the user can start with quick search in the two interactive tables for the label-free and label-based data based on the gene of interest or information in other columns. Also, these tables are sortable for any columns of interest. We also designed an advanced query option for the user in order to query the protein/gene of interest to get more detail information. To do the advanced query, the user first needs to specify a gene symbol, UniProt or Ensemble gene ID in the query box. For example, if the target gene name is EGFR, the user can enter the EGFR in the query box (the first query field). The integrated autocomplete feature will let the user to choose the gene name or IDs in case of uncertainty. Next, the user has to select the cancer type of interest (or all the cancer types). The last option is to choose the data type which has three choices- the label-free, label-based and both options. Then, the user can submit the query to the server. The advanced query provides the user the possibility to combine various queries between the cancer types and quantification techniques. The result pages of label-*free* and *label-based* are similar in annotations and secretory features section (Figures 3 and 4), but they differ in secretome data results (Figure 2). For details explanation of the results pages see to the Figures 2–4.

Results

The structure of the HCSD

The HCSD structure was designed to fulfil four main goals (i) to provide a straightforward searchable depository for published data on different types of human cancer secretome, (ii) the ability to compare information across different secretome measurements (iii) to provide annotation; cross-references in both gene and protein level for each data points and (iv) prediction and visualization of the secretory features for each protein. Therefore, HCSD contains all the proteins (peptides) that are quantified so far to be (differentially) expressed in various cancer types secretome and at the same time provides annotation and predictions about their secretory type.

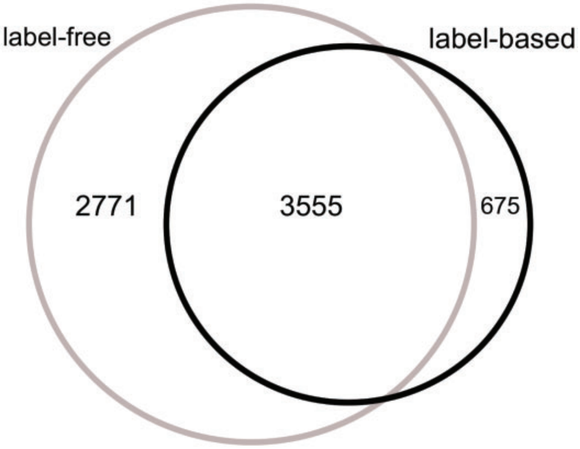
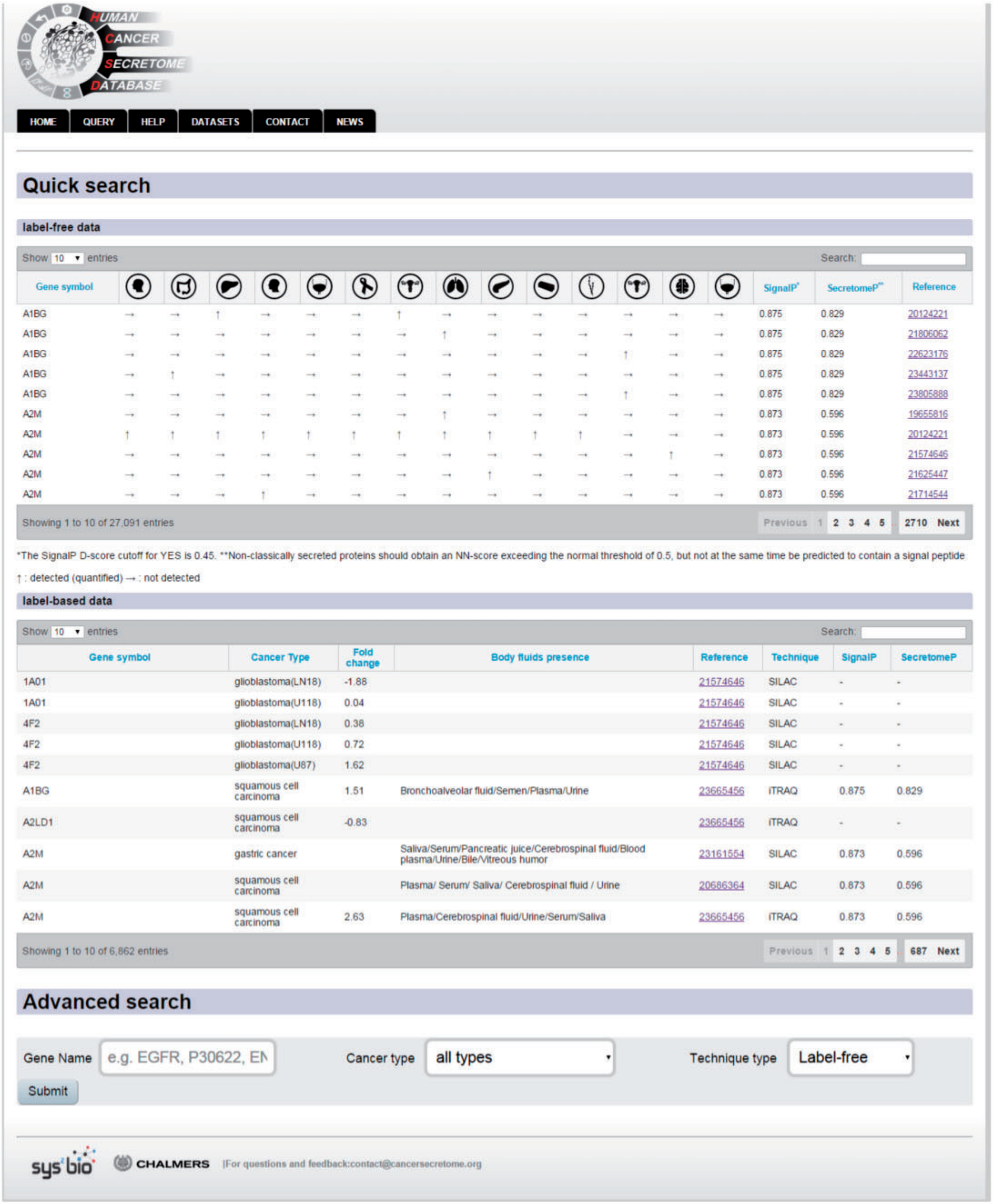


Figure 2. The Venn diagram of the proteins measured in label-free and label-based studies (35 publications).

In eukaryotic cells, protein secretion is carried out either by the classic secretion pathway (having N-terminal signal peptide) or the non-classical pathway(s) (29, 30). It is valuable to know which processes the detected proteins potentially use for secretion in to the tumor microenvironment. Beside this, secretome analysis always is contaminated with proteins from cell debris or culture media that results in false identifications. To assist with these challenges, bioinformatics algorithms have been developed that can predict the secretory type of proteins from primary sequence based on signal peptide pattern, transmembrane domain or other motifs. These tools are extensively reviewed elsewhere (31). However, checking the reliability of the detection in secretome analysis is tightly depending on these tools, and therefore a secretory feature section is included in the results page for each protein query in order to give a summary of the predictions on signal peptide, transmembrane domain and nonclassical secretion signals using the most frequently used bioinformatics tools (Figure 1). Moreover, specific post-translational modifications (PTMs) are another characteristic of secretory proteins among which disulphide bonds and glycosylation sites (N-linked and O-linked) are the most specific. These information also can be visualized on protein sequence by the user in the result page (Figure 5).

An exclusive menu called ‘DATA SETS’ were designed which allows the user to get access and query the basic information about the publication used as data source. Each publication also has its own page which provides more details on the workflow and experimental design. The data set table provides hyperlinks to each publication PubMed page. The *study* column in the result page also



Quick search

label-free data

Show 10 entries Search:

Gene symbol															SignalP*	SecretomeP**	Reference
A1BG	→	→	↑	→	→	→	↑	→	→	→	→	→	→	→	0.875	0.829	20124221
A1BG	→	→	→	→	→	→	→	↑	→	→	→	→	→	→	0.875	0.829	21806062
A1BG	→	→	→	→	→	→	→	→	→	→	→	↑	→	→	0.875	0.829	22623176
A1BG	→	↑	→	→	→	→	→	→	→	→	→	→	→	→	0.875	0.829	23443137
A1BG	→	→	→	→	→	→	→	→	→	→	→	↑	→	→	0.875	0.829	23805888
A2M	→	→	→	→	→	→	→	↑	→	→	→	→	→	→	0.873	0.596	19655816
A2M	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	→	→	→	0.873	0.596	20124221
A2M	→	→	→	→	→	→	→	→	→	→	→	→	↑	→	0.873	0.596	21574646
A2M	→	→	→	→	→	→	→	→	↑	→	→	→	→	→	0.873	0.596	21625447
A2M	→	→	→	↑	→	→	→	→	→	→	→	→	→	→	0.873	0.596	21714544

Showing 1 to 10 of 27,091 entries Previous 1 2 3 4 5 … 2710 Next

*The SignalP D-score cutoff for YES is 0.45. **Non-classically secreted proteins should obtain an NN-score exceeding the normal threshold of 0.5, but not at the same time be predicted to contain a signal peptide

↑ : detected (quantified) → : not detected

label-based data

Show 10 entries Search:

Gene symbol	Cancer Type	Fold change	Body fluids presence	Reference	Technique	SignalP	SecretomeP
1A01	glioblastoma(LN18)	-1.88		21574646	SILAC	-	-
1A01	glioblastoma(U118)	0.04		21574646	SILAC	-	-
4F2	glioblastoma(LN18)	0.38		21574646	SILAC	-	-
4F2	glioblastoma(U118)	0.72		21574646	SILAC	-	-
4F2	glioblastoma(U87)	1.62		21574646	SILAC	-	-
A1BG	squamous cell carcinoma	1.51	Bronchoalveolar fluid/Semen/Plasma/Urine	23665456	iTRAQ	0.875	0.829
A2LD1	squamous cell carcinoma	-0.83		23665456	iTRAQ	-	-
A2M	gastric cancer		Saliva/Serum/Pancreatic juice/Cerebrospinal fluid/Blood plasma/Urine/Bile/Vitreous humor	23161554	SILAC	0.873	0.596
A2M	squamous cell carcinoma		Plasma/ Serum/ Saliva/ Cerebrospinal fluid / Urine	20686364	SILAC	0.873	0.596
A2M	squamous cell carcinoma	2.63	Plasma/Cerebrospinal fluid/Urine/Serum/Saliva	23665456	iTRAQ	0.873	0.596

Showing 1 to 10 of 6,862 entries Previous 1 2 3 4 5 … 687 Next

Advanced search

Gene Name: e.g. EGFR, P30622, EN Cancer type: all types Technique type: Label-free

Submit

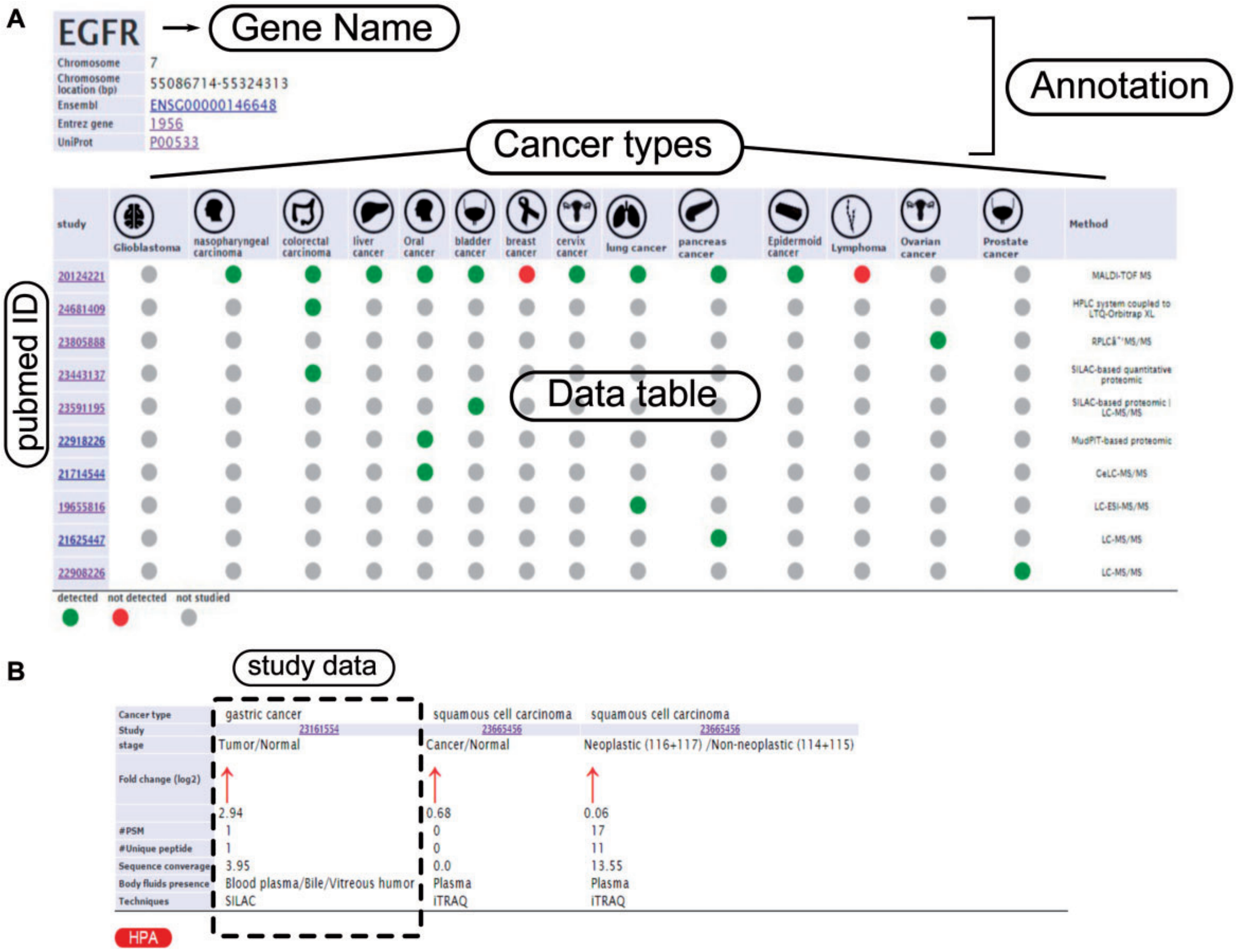
sys bio CHALMERS |For questions and feedback:contact@cancersecretome.org

Figure 3. Example of the query page showing the search options.

directs the user to PubMed page of the corresponding publication.

The statistics of the HCSD

From 87 496 total measurements stored in HCSD, ~85% are derived from label-free on 14 cancer types. So far, the label-based cancer secretome analysis has been mainly performed on 5 cancer types (Supplementary Tables S1 and S2). The Lung cancer secretome is the most studied cancer and includes ~11% of the total data (Supplementary Tables S1 and S2). From 7001 unique proteins in HCSD, 6326 are measured in label-free (with 1148 being transmembrane proteins) and 4230 being measured in label-



EGFR	
Chromosome	7
Chromosome location (bp)	55086714-55324313
Ensembl	ENSG00000146648
Entrez gene	1956
UniProt	P00533

study	Glioblastoma	nasopharyngeal carcinoma	colorectal carcinoma	liver cancer	Oral cancer	bladder cancer	breast cancer	cervix cancer	lung cancer	pancreas cancer	Epidermoid cancer	Lymphoma	Ovarian cancer	Prostate cancer	Method
20124221	⚪	🟢	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢	🔴	⚪	⚪	MALDI-TOF MS
24681409	⚪	⚪	🟢	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	HPLC system coupled to LTQ-Orbitrap XL
23805888	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	🟢	⚪	RPLCâ"'MS/MS
23443137	⚪	⚪	🟢	⚪	⚪	⚪	⚪	⚪		⚪	⚪	⚪	⚪	⚪	SILAC-based quantitative proteomic
23591195	⚪	⚪	⚪	⚪	⚪	🟢					⚪	⚪	⚪	⚪	SILAC-based proteomic | LC-MS/MS
22918226	⚪	⚪	⚪	⚪	🟢	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	MudPIT-based proteomic
21714544	⚪	⚪	⚪	⚪	🟢	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	CeLC-MS/MS
19655816	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	🟢	⚪	⚪	⚪	⚪	⚪	LC-ESI-MS/MS
21625447	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	🟢	⚪	⚪	⚪	⚪	LC-MS/MS
22908226	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	🟢	LC-MS/MS

detected	not detected	not studied
🟢	🔴	⚪

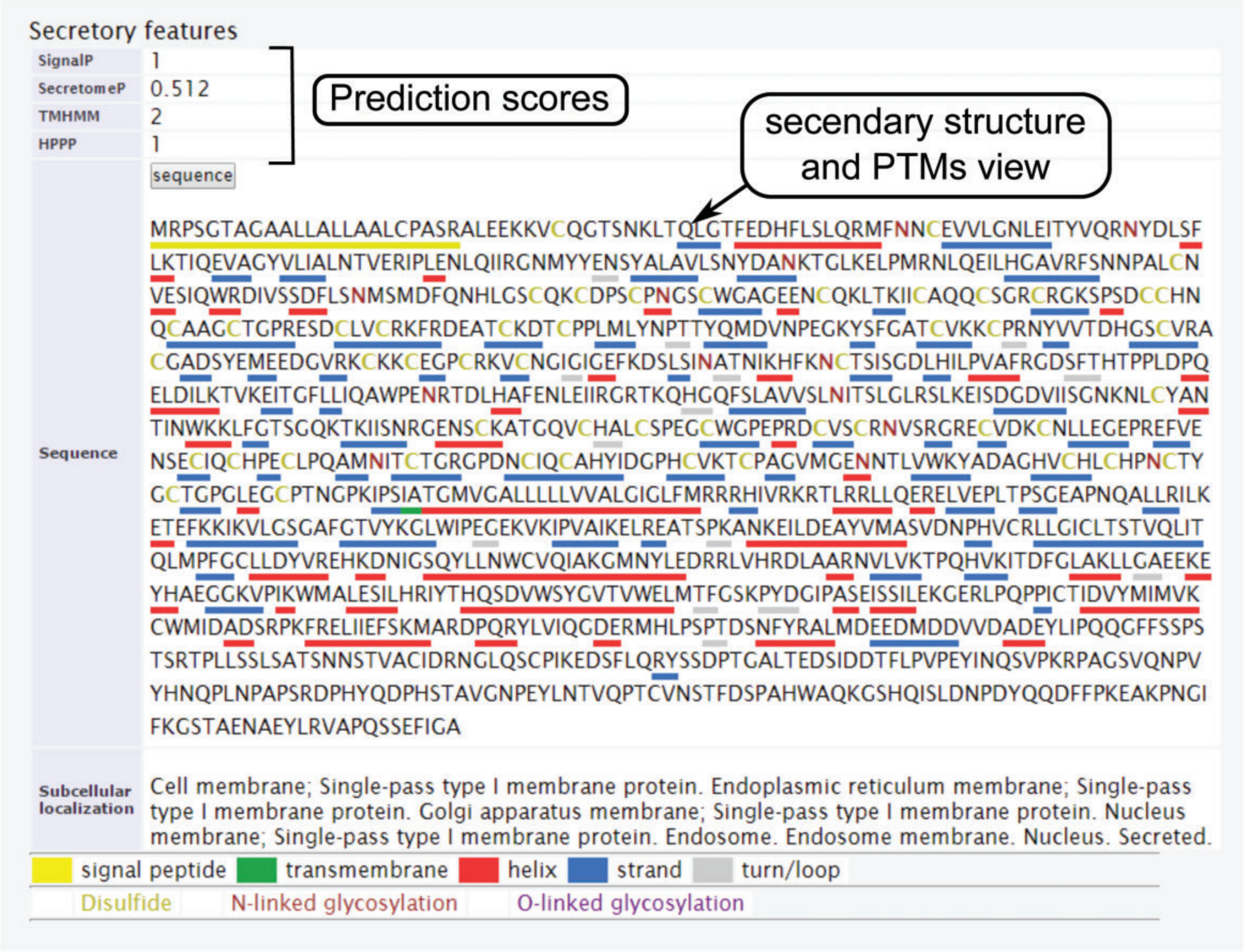
Cancer type	gastric cancer	squamous cell carcinoma	squamous cell carcinoma
Study	23161554	23665456	23665456
stage	Tumor/Normal	Cancer/Normal	Neoplastic (116+117) /Non-neoplastic (114+115)
Fold change (log2)	↑	↑	↑
	2.94	0.68	0.06
#PSM	1	0	17
#Unique peptide	1	0	11
Sequence converage	3.95	0.0	13.55
Body fluids presence	Blood plasma/Bile/Vitreous humor	Plasma	Plasma
Techniques	SILAC	iTRAQ	iTRAQ

Figure 4. Example from the result pages for label-free and label-based studies. In the result page, the first section (**a**) provides the annotations such as gene name, description, chromosomal location and cross references ID to the Ensembl (http://www.ensembl.org/), Entrez (http://www.ncbi.nlm.nih.gov/), and UniProt (www.uniprot.org/). In case of label-free search, exploring all type of cancers will be visualized as a table with the cancer type icons in the header. The first column contains hyperlinked PubMed IDs. For each cancer type column, the protein of interest is detected (green spot), not detected (red spot) or not studied (grey spot). The last column specify the proteomics method used in the study. (**b**) In the case of label-based data, the result table header includes cancer types and the follow up information including the cancer stages, quantified fold change, number of the PSMs, number of the unique peptides, sequence coverage and body fluid presence will come as additional rows.

based (with 534 transmembrane proteins).These two datasets share 3555 proteins (Figure 2). In general, most of the proteins detected in different cancer types secretome are secreted by nonclassical secretion pathways (Figure S1). In total, 1413 nonredundant proteins are detected to be secreted by classic secretion pathway in 14 cancer types from 21 label-free, while this number for nonclassical secreted proteins is 4,945. These numbers in label-based studies are 840 (classic) and 3409 (non-classic) proteins (Supplementary Tables S1 and S2). Most of the cancer secretome data was generated on cancer cell lines. In 35 publications used to design HCSD, 70 cancer cell lines were used to study the cancer secretome (Supplementary Table S3). In case that the authors did not include the cancer type of the cell lines they used, we included the corresponding cancer type.

Discussion

How secreted proteins or peptides from cancer cells remodel the tumor microenvironment in favor of the metastasis is a pivotal research interest in the tumor biology. Cancer cell secretome profiling is a promising approach to find potential body fluid-accessible cancer biomarkers and therapeutic targets, however mining the increasing data from different labs is a big challenge which affects the efficiency of selecting useful candidates and results in the accumulation of redundant and false identified proteins. HCSD (www.cancersecretome.org) was developed as a database to store and query publically available human cancer secretome data to bypass these challenges. It provides the researchers to have access to all the high-throughput data from studies in this field together with the needed detail



Secretory features

SignalP	1
SecretomeP	0.512
TMHMM	2
HPPP	1

Sequence

```
MRPSGTAGAALLALLAALCPASRALEEKKVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEVVLGNLEITYVQRNYDLSF
LKTIQEVAGYVLIALNTVERIPLENLQIIRGNMYYENSYALAVLSNYDANKTGLKELPMRNLQEILHGAVRFSNNPALCN
VESIQWRDIVSSDFLSNMSMDFQNHLGSCQKCDPSCPNGSCWGAGEENCQKLTKIICAQQCSGRCRGKSPSDCCHN
QCAAGCTGPRESDCLVCRKFRDEATCKDTCPPLMLYNPTTYQMDVNPEGKYSFGATCVKKCPRNYVVTDHGSCVRA
CGADSYEMEEDGVRKCKKCEGPCRKVCNGIGIGEFKDSLSINATNIKHFKNCTSISGDLHILPVAFRGDSFTHTPPLDPQ
ELDILKTVKEITGFLLIQAWPENRTDLHAFENLEIIRGRTKQHGQFSLAVVSLNITSLGLRSLKEISDGDVIISGNKNLCYAN
TINWKKLFGTSGQKTKIISNRGENSCKATGQVCHALCSPEGCWGPEPRDCVSCRNVSRGRECVDKCNLLEGEPREFVE
NSECIQCHPECLPQAMNITCTGRGPDNCIQCAHYIDGPHCVKTCPAGVMGENNTLVWKYADAGHVCHLCHPNCTY
GCTGPGLEGCPTNGPKIPSIATGMVGALLLLLVVALGIGLFMRRRHIVRKRTLRRLLQERELVEPLTPSGEAPNQALLRILK
ETEFKKIKVLGSGAFGTVYKGLWIPEGEKVKIPVAIKELREATSPKANKEILDEAYVMASVDNPHVCRLLGICLTSTVQLIT
QLMPFGCLLDYVREHKDNIGSQYLLNWCVQIAKGMNYLEDRRLVHRDLAARNVLVKTPQHVKITDFGLAKLLGAEEKE
YHAEGGKVPIKWMALESILHRIYTHQSDVWSYGVTVWELMTFGSKPYDGIPASEISSILEKGERLPQPPICTIDVYMIMVK
CWMIDADSRPKFRELIIEFSKMARDPQRYLVIQGDERMHLPSPTDSNFYRALMDEEDMDDVVDADEYLIPQQGFFSSPS
TSRTPLLSSLSATSNNSTVACIDRNGLQSCPIKEDSFLQRYSSDPTGALTEDSIDDTFLPVPEYINQSVPKRPAGSVQNPV
YHNQPLNPAPSRDPHYQDPHSTAVGNPEYLNTVQPTCVNSTFDSPAHWAQKGSHQISLDNPDYQQDFFPKEAKPNGI
FKGSTAENAEYLRVAPQSSEFIGA
```

Subcellular localization: Cell membrane; Single-pass type I membrane protein. Endoplasmic reticulum membrane; Single-pass type I membrane protein. Golgi apparatus membrane; Single-pass type I membrane protein. Nucleus membrane; Single-pass type I membrane protein. Endosome. Endosome membrane. Nucleus. Secreted.

🟨 signal peptide 🟩 transmembrane 🟥 helix 🟦 strand ⬜ turn/loop

Disulfide N-linked glycosylation O-linked glycosylation

Figure 5. The secondary structure, secretory pathway features, subcellular localization and PTMs information. Querying both label-free and label-based studies, the second part of the result page is specified for the prediction scores of the secretory features and visualization of the PTMs and secondary structure information. The secretory features include scores of SingalP (33) (for signal peptide), TMHMM (for transmembrane domain) (32), SecretomeP (32, 34) (for nonclassical secretion), and HPPP (for human plasma membrane proteins) (36). The last row of the table shows the subcellular localization data. The PTMs are color coded. The color code legend for PTMs and secondary structure information will appear below the table.

information in terms of the functional annotation and secretory type for each protein. It also allows exploring previously used workflows, cell lines, validated biomarkers and clinical surveys. HCSD can be used extensively by tumor biologist to find their target secreted factor in specific or various cancer types with all the annotations and sequence bioinformatics analysis of the primary sequence and secondary structure information of the target proteins. All this will facilitate the oncoproteomics studies in future.

Supplementary Data

Supplementary data are available at *Database* Online.

Acknowledgements

The authors memorize Aaron Swartz (the developer of the *webpy*) who dedicated his short life to the open-source community. This work was funded by Novo Nordisk Foundation and Knut and Alice Wallenberg foundation.

Funding

The Novo Nordisk Foundation; Knut and Alice Wallenberg Foundation. Funding for open access charge: Chalmers University of Technology Library.

Conflict of interest. None declared.

References

1. Siegel,R., Ma, J., Zou, Z.,*et al*. (2014) Cancer statistics, 2014. *CA Cancer J. Clin.*, **64**, 9–29.
2. Omenn,G.S. (2014) Strategies for genomic and proteomic profiling of cancers. *Stat. Biosci.*, 1–7.
3. Jain,K.K. (2014) Applications of Biotechnology in Oncology. New York Heidelberg Dordrecht London, Springer.
4. Karagiannis,G.S., Pavlou,M.P. and Diamandis,E.P. (2010) Cancer secretomics reveal pathophysiological pathways in cancer molecular oncology. *Mol. Oncol.*, **4**, 496–510.

5. Paltridge,J.L., Belle,L. and Khew-Goodall,Y. (2013) The secretome in cancer progression. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, **1834**, 2233–2241.
6. Makridakis,M. and Vlahou,A. (2010) Secretome proteomics for discovery of cancer biomarkers. *J. Proteomics*, **73**, 2291–2305.
7. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
8. Kessenbrock,K., Plaks,V. and Werb,Z. (2010) Matrix metalloproteinases: regulators of the tumor microenvironment. *Cell*, **141**, 52–67.
9. Whiteside,T. (2008) The tumor microenvironment and its role in promoting tumor growth. *Oncogene*, **27**, 5904–5912.
10. Mbeunkui,F. and Johann D.J. Jr (2009) Cancer and the tumor microenvironment: a review of an essential relationship. *Cancer Chemother. Pharmacol.*, **63**, 571–582.
11. Barderas,R., Mendes,M., Torres,S. *et al.* (2013) In-depth characterization of the secretome of colorectal cancer metastatic cells identifies key proteins in cell adhesion, migration, and invasion. *Mol. Cell. Proteomics*, **12**, 1602–1620.
12. Pavlou,M.P. and Diamandis,E.P. (2010) The cancer cell secretome: a good source for discovering biomarkers? *J. Proteomics*, **73**, 1896–1906.
13. Stastna,M. and Van Eyk,J.E. (2012) Secreted proteins as a fundamental source for biomarker discovery. *Proteomics*, **12**, 722–735.
14. Caccia,D., Zanetti Domingues,L., Miccichè,F. *et al.* (2011) Secretome compartment is a valuable source of biomarkers for cancer-relevant pathways. *J. Proteome Res.*, **10**, 4196–4207.
15. Ralhan,R., Masui,O., DeSouza,L.V. *et al.* (2011) Identification of proteins secreted by head and neck cancer cell lines using LC-MS/MS: Strategy for discovery of candidate serological biomarkers. *Proteomics*, **11**, 2363–2376.
16. Planque,C., Kulasingam,V., Smith,C.R. *et al.* (2009) Identification of five candidate lung cancer biomarkers by proteomics analysis of conditioned media of four lung cancer cell lines. *Mol. Cell. Proteomics*, **8**, 2746–2758.
17. Wu,C.-C., Hsu,C.-W., Chen,C.-D. *et al.* (2010) Candidate serological biomarkers for cancer identified from the secretomes of 23 cancer cell lines and the human protein atlas. *Mol. Cell. Proteomics*, **9**, 1100–1117.
18. Lawlor,K., Nazarian,A., Lacomis,L. *et al.* (2009) Pathway-based biomarker search by high-throughput proteomics profiling of secretomes. *J. Proteome Res.*, **8**, 1489–1503.
19. Pocsfalvi,G., Votta,G., De Vincenzo,A. *et al.* (2011) Analysis of secretome changes uncovers an autocrine/paracrine component in the modulation of cell proliferation and motility by c-Myc. *J. Proteome Res.*, **10**, 5326–5337.
20. Loei,H., Tan,H.T., Lim,T.K. *et al.* (2012) Mining the gastric cancer secretome: identification of GRN as a potential diagnostic marker for early gastric cancer. *J. Proteome Res.*, **11**, 1759–1772.
21. Schaab,C., Geiger,T., Stoehr,G. *et al.* (2012) Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Molecular & Cellular Proteomics*, **11**, M111. 014068.
22. Bankevich,A., Nurk,S., Antipov,D. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.
23. Bantscheff,M., Lemeer,S., Savitski,M.M. *et al.* (2012) Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.*, **404**, 939–965.
24. Mudunuri,U., Che,A., Yi,M. *et al.* (2009) bioDBnet: the biological database network. *Bioinformatics*, **25**, 555–556.
25. Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, bar009.
26. Flicek,P., Amode,M.R., Barrell,D. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.
27. Maglott,D., Ostell,J., Pruitt,K.D. *et al.* (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
28. Uhlen,M., Oksvold,P., Fagerberg,L. *et al.* (2010) Towards a knowledge-based human protein atlas. *Nat. Biotechnol.*, **28**, 1248–1250.
29. Nickel,W. (2010) Pathways of unconventional protein secretion. *Curr. Opin. Biotechnol.*, **21**, 621–626.
30. Rothman,J.E. and Orci,L. (1992) Molecular dissection of the secretory pathway. *Nature*, **355**, 409–415.
31. Caccia,D., Dugo,M. and Callari,M. (2013) Bioinformatics tools for secretome analysis. *Biochim. Biophys. Acta*, **1834**, 2442–2453.
32. Emanuelsson,O., Brunak,S., von Heijne,G. *et al.* (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–971.
33. Petersen,T.N., Brunak,S., von Heijne,G. *et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.
34. Bendtsen,J.D., Jensen,L.J., Blom,N. *et al.* (2004) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.*, **17**, 349–356.
35. Wu,C.H., Apweiler,R., Bairoch,A. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
36. Farrah,T., Deutsch,E.W., Omenn,G.S. *et al.* (2011) A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell. Proteomics*, **10**, M110. 006353.