Clinical Biochemistry 66 (2019) 63–75

Contents lists available at ScienceDirect

Clinical Biochemistry

journal homepage: www.elsevier.com/locate/clinbiochem

Proteome-wide onco-proteogenomic somatic variant identification in ER-positive breast cancer

Lampros Dimitrakopoulos[a,b,c,1], Ioannis Prassas[b,c,1], Anieta M. Sieuwerts[d], Eleftherios P. Diamandis[a,b,c,e], John W.M. Martens[d,**], George S. Charames[a,b,c,*]

[a] Department of Laboratory Medicine and Pathobiology, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada
[b] Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Joseph and Wolf Lebovic Health Complex, 600 University Avenue, Toronto, ON M5G 1X5, Canada
[c] Lunenfeld-Tanenbaum Research Institute, Sinai Health System, 600 University Avenue, Toronto, ON M5G 1X5, Canada
[d] Department of Medical Oncology and Cancer Genomics Netherlands, Erasmus MC Cancer Institute, Erasmus University Medical Center, Wytemaweg 80, 3015 CN Rotterdam, The Netherlands
[e] Department of Clinical Biochemistry, University Health Network, 190 Elizabeth Street, Toronto, ON M5G 2C4, Canada

ARTICLE INFO

Keywords:
Onco-proteogenomics
Variant peptides
Somatic mutations
Breast cancer
LC-MS/MS

ABSTRACT

Background: Recent advances in mass spectrometric instrumentation and bioinformatics have critically contributed to the field of proteogenomics. Nonetheless, whether that integrative approach has reached the point of maturity to effectively reveal the flow of genetic variants from DNA to proteins still remains elusive. The objective of this study was to detect somatically acquired protein variants in breast cancer specimens for which full genome and transcriptome data was already available (BASIS cohort).

Methods: LC-MS/MS shotgun proteomic results of 21 breast cancer tissues were coupled to DNA sequencing data to identify variants at the protein level and finally were used to associate protein expression with gene expression levels.

Results: Here we report the observation of three sequencing-predicted single amino acid somatic variants. The sensitivity of single amino acid variant (SAAV) detection based on DNA sequencing-predicted single nucleotide variants was 0.4%. This sensitivity was increased to 0.6% when all the predicted variants were filtered for MS "compatibility" and was further increased to 2.9% when only proteins with at least one wild type peptide detected were taken into account. A correlation of mRNA abundance and variant peptide detection revealed that transcripts for which variant proteins were detected ranked among the top 6.3% most abundant transcripts. The variants were detected in highly abundant proteins as well, thus establishing transcript and protein abundance and MS "compatibility" as the main factors affecting variant onco-proteogenomic identification.

Conclusions: While proteomics fails to identify the vast majority of exome DNA variants in the resulting proteome, its ability to detect a small subset of SAAVs could prove valuable for precision medicine applications.

Abbreviations: ACTB, actin cytoplasmic 1; AGC, automatic gain control; BSA, bovine serum albumin; CRMP1, dihydropyrimidinase-related protein 1; CTSH, procathepsin H; ERBB2, receptor tyrosine-protein kinase erbB-2; ESR1, estrogen receptor; eQTL, expression quantitative trait loci; FDR, false discovery rate; FPKM, fragments per kilobase of transcript per million mapped reads; GORASP2, golgi reassembly-stacking protein 2; HGVS, human genome variation society; HSPD1, 60 kDa heat shock protein mitochondrial; MRM, multiple reaction monitoring; NCE, normalized collision energy; NCL, nucleolin; PEP, posterior error probability; PGR, progesterone receptor; PRM, parallel reaction monitoring; PSM, peptide spectrum match; pQTL, protein quantitative trait loci; RPL26, 60S ribosomal protein L26; SAAV, single amino acid variation; SCX, strong cation exchange; SIFT, sorting intolerant from tolerant; SRM, selected reaction monitoring; TCGA, the cancer genome atlas; VPS35, vacuolar protein sorting-associated protein 35

* Corresponding author: G.S. Charames, Mount Sinai Hospital, Joseph & Wolf Lebovic Ctr., 600 University Avenue—Room 6–500, Toronto, ON M5G 1X5, Canada.
** Corresponding author: J.W.M. Martens, Erasmus MC Cancer Institute, Erasmus University Medical Center, Wytemaweg 80, 3015 CN Rotterdam, The Netherlands.
E-mail addresses: j.martens@erasmusmc.nl (J.W.M. Martens), george.charames@sinaihealthsystem.ca (G.S. Charames).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.clinbiochem.2019.01.005
Received 9 October 2018; Received in revised form 14 January 2019; Accepted 18 January 2019
Available online 23 January 2019

1. Introduction

As opposed to cancer cell lines that represent homogeneous cell populations, patient tumor specimens are characterized by intratumor heterogeneity arising from the dynamic clonal expansion of one transformed cell [1]. As a result, various cell subpopulations comprised of different genetic composition may exist within the same tumor. This heterogeneity inevitably hampers the successful detection of all variants that exist in a given tumor. This detection is further impeded by stromal contamination of the tumor tissue section. However, while at the DNA level sequencing technology now enables variant detection at the single cell level [1], mass spectrometry is not expected to allow the detection of variants expressed by small tumor cell subpopulations due to the limited analytical sensitivity.

Whole genome sequencing (WGS) has the ability to identify intronic and intergenic variants as opposed to whole exome sequencing (WES). Moreover, whole genome sequencing is more sensitive even for protein coding variant detection because of its more uniform coverage and better mapping approaches compared to whole exome sequencing [2]. Therefore, WGS variant data coupled with transcriptomic gene expression data and mass spectrometry-based proteomics can provide a better picture of how variants present at the DNA level are translated into modified proteins.

In the present study, shotgun proteomic data were obtained from 21 ER+ breast tissues for which WGS data and RNA sequencing data were available from previous studies [3,4]. The main goal of this study was to examine what portion of sequencing-predicted somatic variants could be detected at the proteome level by data-dependent MS/MS analysis. Subsequently, a customized FASTA database was built including variants called from sequencing raw data that were previously obtained from the same specimens in question, and the MS/MS spectra were searched against the entries contained in the database. Data analysis revealed the proportion of sequencing-predicted variant peptides detected at the proteome level, how this proportion increases when predicted variants are narrowed down to those most probable to be identified by shotgun mass spectrometry and the correlation between transcript abundance and variant protein expression.

2. Material and methods

2.1. Breast cancer tissue specimens

Twenty-one breast tissue specimens were analyzed in total. All specimens were ER-positive and invasive tumor cell percentage ranged from 50% to 98%. Detailed clinicopathological characteristics for the 21 tissues are summarized in Supporting Table 1. All samples were collected at the Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, The Netherlands.

2.2. DNA-RNA extraction and protein extraction and digestion

Tissue sections from the 21 breast cancer specimens were processed with RNA-Bee for RNA extraction and with the NucleoSpin Tissue kit for DNA extraction as previously described [5,6]. The extracted DNA and RNA was used for DNA-sequencing and RNA-sequencing experiments as previously reported [3,4] and the generated data were used in this project. The phenol-ethanol supernatant that remained after the RNA isolation with RNA-Bee was used for the extraction of the protein fraction of the samples according to a modified protocol from [7,8]. Specifically, proteins were precipitated from an aliquot of 80 μL from the organic fraction with 3 volumes of isopropanol (240 μL) and 10 min incubation at room temperature. Samples were centrifuged at 14,000 ×*g* for 15 min at 4 °C. Supernatant was removed and protein pellets were washed twice with 95% ethanol. Samples were sonicated with a probe sonicator (QSONICA, LCC., Newtown, CT) for 10 s at power 4 and the fine powder was allowed to settle to the bottom of the tube. Subsequently, most ethanol was removed carefully without drying the powder, which was resolubilized in 0.2% Rapigest SF (Waters, Milford, MA) in 50 mM ammonium bicarbonate. Solubilization was assisted with brief sonication and passing the samples through insulin syringes approximately 20 times. Samples were then diluted to 0.1% Rapigest SF with 50 mM ammonium bicarbonate and boiled for 5 min. Finally, the total protein content for each sample was determined with the Coomassie (Bradford) protein assay (Thermo Fisher Scientific).

Total protein yield ranged from 66.6 μg to 348.3 μg. Samples were then transferred to a new tube and proteins were reduced with 10 mM final concentration DL-Dithiothreitol for 40 min at 60 °C. Subsequently, they were alkylated with 20 mM final concentration iodoacetamide for 1 h in the dark, at room temperature. Trypsin (Trypsin from bovine pancreas, Sigma-Aldrich) was then added (1:20 trypsin to total protein ratio; 1 μg trypsin for every 20 μg of total protein as determined with the Bradford assay), and samples were incubated overnight at 37 °C. Following digestion, samples were centrifuged at 1000 rpm for 2 min and were acidified with trifluoroacetic acid at 1% final concentration to inhibit trypsin and cleave Rapigest. After 2 min centrifugation at 1000 rpm, samples were incubated at room temperature for 10 min and were finally centrifuged at 17,000 ×*g* for 30 min and supernatant was collected.

2.3. Strong cation exchange chromatography (SCX)

Digests were subjected to SCX chromatography to obtain the deepest proteome coverage possible. Due to the nature of the data-dependent acquisition of the shotgun proteomic approach (the 12 most abundant peptide species in a chromatographic run at any given time are being sampled from the mass spectrometer to be sequenced and then potentially identified), in order to identify more and especially low abundance peptides the reduction of the complexity of a given sample is necessary. Thus, more low abundance peptides are expected to be identified when all peptides in a given sample are equally distributed across various fractions. Chromatographic separation based on peptide charge is a typical and most effective way of achieving this peptide fractionation [9]. Specifically, strong cation exchange chromatography (SCX) was chosen and mobile phase A used was 0.26 M formic acid in 10% acetonitrile (pH 2–3) and mobile phase B was 0.26 M formic acid and 1 M ammonium formate in 10% acetonitrile (pH 2–3). Each sample was topped up until 500 μL with mobile phase A and was loaded onto a 500 μL loop connected to a PolySULFOETHYL A™ column (200 × 2.1 mm, 5 μm, 200 Å, The Nest Group, Inc., Southboro, MA). The SCX chromatography and fractionation were performed on an Agilent 1100 series HPLC system using a 60-min gradient, which was optimized to provide a uniform elution of peptides based on the absorption at 280 nm. Mobile phase B was introduced at 5 min and increased to 20% at 30 min and then to 100% at 50 min, where it remained until 55 min before going back to 0%. Fractions were collected every 6 min until the 24 min time point, every 3 min from the 24 min time point until the 30 min time point and every 1 min from the 30 min time point until the 50 min time point. Finally, two fractions were collected between the 50 min and 60 min time points. Twenty eight fractions were collected in total and according to the graphs obtained, peptides were eluted across 23 fractions (fractions 5–27) that were chosen for subsequent LC-MS analysis.

2.4. Shotgun LC-MS/MS

Before shotgun LC-MS/MS analyses, all fractions were desalted with Bond Elut OMIX C18 10 μL tips (Agilent). Peptides were eluted in 2.5 μL of 65% acetonitrile and 0.1% formic acid, while 0.1% formic acid was used to top up each sample to generate enough volume for 2 technical replicate injections. Samples were loaded onto in-house packed 3 cm long 5 μm particle C18 trap-columns from an EASY-nLC 1000 pump

(Thermo Fisher Scientific) 96-well microplate autosampler before being eluted to in-house packed 15 cm long 3 μm particle C18 analytical columns. Mobile phase A was 0.1% formic acid in water and mobile phase B 0.1% formic acid in acetonitrile. A five-step 60 min gradient was used: 1% to 5% of mobile phase B for 2 min, 5% to 35% for 47 min, 35% to 65% for 3 min, 65% to 100% for 1 min and 100% for 7 min. Flowrate was 300 nL/min for 53 min and 450 nL/min for 7 min. The liquid chromatography setup was coupled online to a Q Exactive™ Plus Hybrid Quadrupole-Orbitrap™ Mass Spectrometer (Thermo Fisher Scientific). Peptides were analyzed in data-dependent acquisition mode. A Full MS1 scan in the range 400 to 1500 m/z (resolution, i.e. a measure of the ability to distinguish two peaks of slightly different mass-to-charge ratios in a mass spectrum, 70,000) and AGC (automatic gain control, i.e. the maximum number of ions present at the ion trap at any given time) target 3×10^6 was followed by 12 data-dependent MS2 scans at 17,500 resolution and with AGC target 5×10^4, stepped NCE (normalized collision energy, i.e. the energy used by the mass spectrometer to produce the fragments of the precursor ions) 27 and 1.6 m/z isolation window (mass-to-charge range within which the selected ions pass through the mass filter. A dynamic exclusion window (precursor ions that have been selected for fragmentation within the selected time window are excluded from being selected again within this time window) of 45 s was enabled and precursor charge states of 1, > 4 and unassigned precursor charge states were excluded from MS2 fragmentation. MS repeatability was evaluated by running 10 fmol and 100 fmol BSA quality control samples between the samples. Specifically, 100 fmol BSA samples were analyzed in the beginning and at the end of each sample running sequence and 10 fmol BSA samples were analyzed in between every 3 sample fractions analyzed. Three specific peptide peaks were monitored in each BSA sample and was made sure that the respective areas under the curve exceeded a specific threshold to ensure the stability of the MS operation.

2.5. Mutation calling, gene expression analysis and database construction

Somatic variants and gene expression values (FPKM data values) used here have been described elsewhere [3,4]. Specifically, peripheral blood lymphocytes and adjacent normal breast tissues from a total of 560 breast cancer patients were used. In order to eliminate germline variants, sequencing data were obtained for these biospecimens along with an equal number of breast tumor samples (21 of these tumor samples were included in our study). During the raw sequencing data processing, the paired-end reads from all samples (normal and tumor) were aligned to the reference human genome (GRCh37) using Burrows-Wheeler Aligner (BWA). Subsequently, these aligned reads were used to call somatic variants using the CaVEMan algorithm (http://cancerit.github.io/CaVEMan/) which takes as input pairs of tumor/normal sequencing data and the output is the sum of variants which are expected to be somatic events and not found in normal tissues. Mutations detected by whole genome sequencing contained in our initial dataset were narrowed down only to those found in protein coding regions of the genome and in total 817 missense and nonsense mutations were included in a FASTA database file. All statistical analyses regarding the correlation between transcripts and proteins were performed with SPSS statistics, release 17.0.0 and Perseus software, version 1.5.6.0.

2.6. Database search

Shotgun proteomic raw files (one raw file with data for acquired spectra for each of the 23 fractions containing peptides for each sample analyzed) generated for the 21 breast cancer tissues were searched against a customized combined database consisting of the human reference database (UniProtKB reference proteome, 01/2016 release, containing 42,074 entries) and the 817 entries with the variant entries from the FASTA file constructed from the DNA sequencing data (42,891 entries in total). The search was executed with Proteome Discoverer software (version 1.4, Thermo Fisher Scientific) and Sequest HT was used as the search engine. Only spectra with precursor MS1 ions with masses 350–5000 Da were selected for the subsequent engine searches. Mass tolerance for precursor ions was set to 7 ppm and for fragment ions to 0.02 Da. Up to 2 missed cleavages were allowed and only peptides with 6–144 amino acids were considered. Cysteine carbamidomethylation was selected as a static modification and oxidation of methionine and deamidation of asparagine and glutamine as dynamic modifications. False discovery rate (FDR, the percentage of accepted false positive peptide identifications originating from the database search) was set to 1%, which is considered to be a relaxed threshold for the detection of variant peptides but it is acceptable for the detection of variant peptides that are subsequently validated by targeted proteomic assays.

2.7. MRM validation assays

Mutant peptides detected with the shotgun proteogenomic analysis were subsequently validated with targeted parallel reaction monitoring (PRM) assays utilizing isotopically labeled internal standards. This validation was performed in both new digests of the same samples in question as well as using the second technical replicate injection in which the respective mutant peptides had been identified during the shotgun experiment.

New digests were prepared as described above for the preparation of digests subjected to strong cation exchange chromatography, but this time no Offline-fractionation was employed before the targeted MRM assays. 30 μg of total protein were digested and following the addition of trifluoroacetic acid, 5 pmol from each heavy labeled standard peptide (JPT Peptide Technologies, Berlin, Germany) corresponding to the mutant peptides expected to be expressed in each sample according to the shotgun proteogenomic results were added to each sample. The same heavy labeled peptides were added in the second shotgun technical replicates in a final concentration of 20 fmol/uL before the samples were injected into the Q Exactive™ Plus Hybrid Quadrupole-Orbitrap™ Mass Spectrometer. These heavy labeled standard peptides exactly mimic the physical properties of the variant peptides and therefore exhibit identical behavior during their chromatographic separation, but can be distinguished by the mass spectrometer due to their difference in their mass-to-charge ratios. As a result, they serve as ideal standards for the targeted validation of the presence of the variant peptides identified with the shotgun approach.

2.8. Online database repositories search

Several cancer-associated sequencing databases were queried for the detected mutations to check whether they have been previously reported in the literature or they constitute novel mutations. Specifically, COSMIC (Catalogue of somatic mutations in cancer) [10], TumorPortal [11] and cBioPortal [12] online databases were searched for the variants detected at the proteome level. Two non-cancer databases were searched as well, namely the NHLBI GO Exome Sequencing Project (http://evs.gs.washington.edu/EVS/) and the ExAC browser [13].

3. Results

3.1. Database search results

Upon searching the MS/MS spectra against the FASTA database containing all the sequencing-predicted variants, we extracted the results with all peptides detected and all proteins inferred from these peptides. Subsequently, the first step was to search for evidence of translation of estrogen receptor, progesterone receptor and HER2 receptor transcripts and use this as a control check for the quality of our proteomic results. We sought to identify samples with at least one

peptide from each one of the three receptors having been detected and correlated this presence or absence with the pathological staining assay results available for the 21 tissue specimens. Supporting Table 1 shows the MS/MS search engine result for ESR1, PGR and ERBB2 peptides, the concordance with the immunohistochemical results and the total number of proteins identified in each sample. Specifically, in 16 of the 21 samples there is a concordance between the immunohistochemical staining and the mass spectrometric result for the detection of estrogen receptor, whereas for the progesterone receptor there is a concordance observed for the 17 of the 21 samples analyzed. Of note, there is not ERBB2 staining data available and as a result only the proteomic result is depicted in Supporting Table 1. Additionally, the total number of peptides and proteins identified across the 21 samples ranged from 42,720 and 3713 to 120,888 and 6369 respectively (median values: 80,748 peptides and 5188 proteins).

Whole-genome sequencing revealed a large number of mutations across all samples (95,998 for a median of 2603 point mutations per sample). However, only a fraction (817 point mutations resulting in 789 unique SAAVs or truncating mutations) of these mutations fell within the protein coding region. Out of those 789 mutations only 4 were detected at the protein level with our proteogenomic workflow. The breakdown of the number of mutations present in each sample genome-wide and among the protein coding sequence as well as those detected at the peptide level by MS/MS are depicted in Table 1.

3.2. Variant peptides detected

The database search matched in total four theoretical variant peptides that were suggested from the whole-exome sequencing data and included in the customized FASTA file with four experimentally acquired MS/MS spectra. These variant peptides were detected in four different samples and are surrogates of four different proteins. Nucleolin in sample PD6422, actin in sample PD4604, golgi reassembly-stacking protein 2 in sample PD9842 and isoform LCRMP-1 of dihydropyrimidinase-related protein 1 in sample PD9845. The mutations, the variant peptides and the most important attributes from the Proteome Discoverer software are shown in Table 2.

Ideally every shotgun proteogenomic endeavor should be accompanied with follow up targeted validation assays including internal standards. Therefore, the identified variant peptides were tracked with MRM assays in the same fractions that were analyzed in the initial discovery experiment as well as in new unfractionated digests of the same samples in question. While only one of the variant peptides was detected in the unfractionated digests (only the M219 V in sample PD9842, as expected because the lack of reduction of the sample complexity is not outweighed by the more sensitive nature of the targeted versus the shotgun proteomic approach), 3 of the 4 variant peptides were detected in the respective SCX fractions.

Validation of the triply charged AITIAGRINCPMYITK peptide in both the unfractionated digest and the SCX fraction failed, confirming that it was a false positive identification. Figs. 1, 2, 3 and 4 illustrate the MRM chromatograms for the four variant peptides and the original MS/MS spectra that led to their initial detection by the search engine.

3.3. Gene expression correlation with mutant protein abundance

Only a small fraction of the sequencing-predicted variant peptides was detected at the proteome level. A variety of factors may have accounted for this low yield. Gene transcripts constitute the intermediate in the flow of genetic information from DNA to proteins. Hence, a first estimation for the expression of the mutant versions of the genes can be estimated using gene expression data which is derived from RNA-sequencing in our dataset [4]. For that purpose, we mined matched gene expression data available for the 21 tissues and ranked all transcripts with sequencing-predicted mutations in terms of decreasing abundance. Notably, for all transcripts no discrimination was made for any variant forms and all expression values mentioned below refer to the wild type and any variant forms taken into account altogether. Not surprisingly, all three transcripts with mutations predicted from the whole-genome sequencing data and detected by MS/MS (NCL in sample PD6422, ACTB in sample PD4604 and GORASP2 in sample PD9842) were ranked first among all the other transcripts with a sequencing-predicted mutation in the respective samples (Table 3). Specifically, log2-transformed FPKM values for each transcript were ranked in descending order for each sample. Nucleolin transcript was the most abundant transcript in sample PD6422 with 7.018 log2-transformed expression value, while actin and Golgi-reassembly stacking protein topped samples PD4604 and PD9842 with log2-transformed expression values of 8.531 and 4.904 respectively. It should be noted that log2-transformed expression values (regarding the WGS-predicted mutant transcripts only) ranged from 7.018 to −7.2219 in sample PD6422, from 8.531 to −7.136 in sample PD4604 and from 4.904 to −7.6103 in sample PD9842. Subsequently, we ranked all 20,188 genes in the 21 samples to conceptualize the levels of expression for those three transcripts across the full sample cohort. Table 4 shows the rank each of the 3 transcripts has in each sample, whereas Table 3 shows how all transcripts with sequencing-predicted mutations are ranked in samples PD6422, PD4604 and PD9842. Not surprisingly, all 3 transcripts ranked high in terms of abundance in all samples analyzed in the present study. When we checked the proteomic database search results for evidence of translation for the 3 proteins, an mRNA-protein expression correlation was revealed. At least one wild type peptide from NCL and GORASP2 proteins was detected in all samples, whereas at least one wild type peptide from actin was detected in all samples with the sole exception of sample 9894. All 20,188 transcripts were ranked in terms of descending log2-transformed FPKM values in each sample. The rank of the 3 transcripts of interest is reported for each sample. Values in bold represent the samples with the respective mutation detected at the DNA and protein level.

To further evaluate the relation between transcript and protein

Table 1
All somatic mutations detected from whole-genome sequencing, mutations in protein coding regions and mutations detected at the protein level in the 21 breast cancer tissues.

Sample	All WGS mutations	Protein coding mutations	MS/MS	Sample	All WGS mutations	Protein coding mutations	MS/MS
PD6418	2020	23	0	PD13626	1076	7	0
PD6422	5181	38	1	PD9842	3201	31	1
PD4607	25,953	203	0	PD9847	2229	19	0
PD4613	2195	27	0	PD9845	3147	31	1
PD4604	9323	64	1	PD14432	1870	19	0
PD7238	2217	21	0	PD14435	3716	34	0
PD13625	11,239	68	0	PD14437	4363	29	0
PD13623	3800	30	0	PD14439	1316	19	0
PD13631	2603	18	0	PD14441	4662	43	0
PD13629	2029	20	0	PD14442	1746	19	0
PD13619	2112	26	0	–	–	–	–

Table 2
Variant peptides detected with the proteogenomic workflow and the database search.

Mutation – sample	V404D – PD6422	D81N – PD4604	M219V – PD9842	V364M – PD9845
Protein name	Nucleolin (NCL)	Actin, cytoplasmic 1 (ACTB)	Golgi reassembly-stacking protein 2 (GORASP2)	Isoform LCRMP-1 of Dihydropyrimidinase-related protein 1 (CRMP1)
Peptide sequence	DTQDELKEVFEDAAEIR	YPIEHGIVTNWDNMEK	ISLPGQVAGTPITPLK	AITIAGRINCPMYITK
Area	4.80E + 07	2.42E + 08	6.59E + 06	–
q-value	0	0	0	0.001
Posterior error probability (PEP)	8.33E-07	8.40E-06	0.000212	0.0125
Xcorr	3.23	3.23	2.07	1.67
Intensity	8.70E + 05	6.37E + 06	2.02E + 05	1.35E + 06
Charge	3	3	2	2
m/z	669.98944	649.3096	796.4748	911.4864

Variant V364M failed to be validated as opposed to the other 3 variants that were successfully validated with targeted MRM assays.
*Area: The area under the chromatographic peak for the peptide in question.
*Intensity: Maximum number of ions reaching the detector per unit of time for the peptide in question.
*q-value: The minimal false discovery rate that the identification is considered correct.
*PEP: The posterior error probability (PEP) is the probability that the observed PSM is incorrect.
*Xcorr: Xcorr is a measure of the goodness of fit of experimental peptide fragments to theoretical spectra created from the sequence b and y ions; for doubly charged ions, identifications with Xcorr value equal or greater to 1.9 are considered true positives while for triply charged ions identifications with Xcorr values equal or greater to 2.3 are considered true positives.

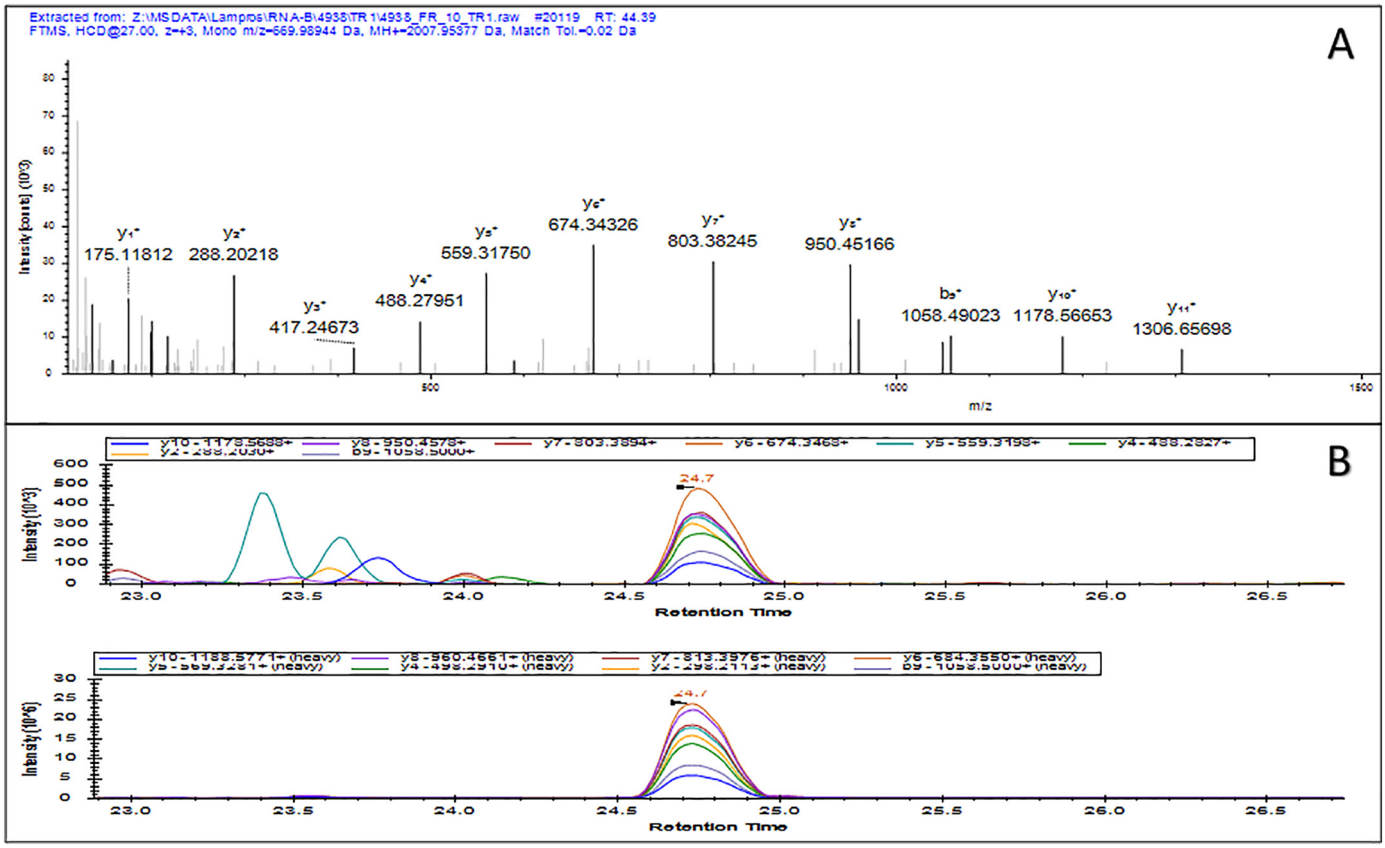

Fig. 1. Variant peptide DTQDELKEVFEDAAEIR (p.V404D) detected in the shotgun analysis and validated with the targeted MRM assay. (A) Annotated MS/MS spectrum. (B) Same relative transition intensities for both heavy isotopically labeled peptides and light peptides along with their co-elution confirm the presence of the peptide. Upper chromatogram represents the endogenous light peptide and lower chromatogram represents the heavy labeled spiked peptide standard.

abundance irrespective of the mutational status, we performed a global analysis of the correlation of mRNA-to-protein. For this correlation analysis, we focused on the 1740 mRNA-protein pairs that had non-zero FPKM and intensity area values across all 21 samples included in this study (Fig. 5). All transcript FPKM values and protein intensity area values were log2 transformed, and were as a result more normally distributed. 60.3% of the pairs were positively correlated (i.e. the greater the mRNA abundance, the greater the resulting protein abundance as well) and 14.9% were significantly correlated with 11.4% being both positively and significantly correlated according to Pearson coefficients and *p*-values (FDR 0.25 and Benjamini-Hochberg adjusted *p* value 0.037). Mean Pearson correlation coefficient was 0.09 (median 0.09). Even though for many pairs the positive correlation is not significant, Fig. 5 illustrates the global transcript-protein interplay across a cohort of 21 estrogen receptor-positive breast cancer tumors.

Finally, we examined whether specific mutations might have an

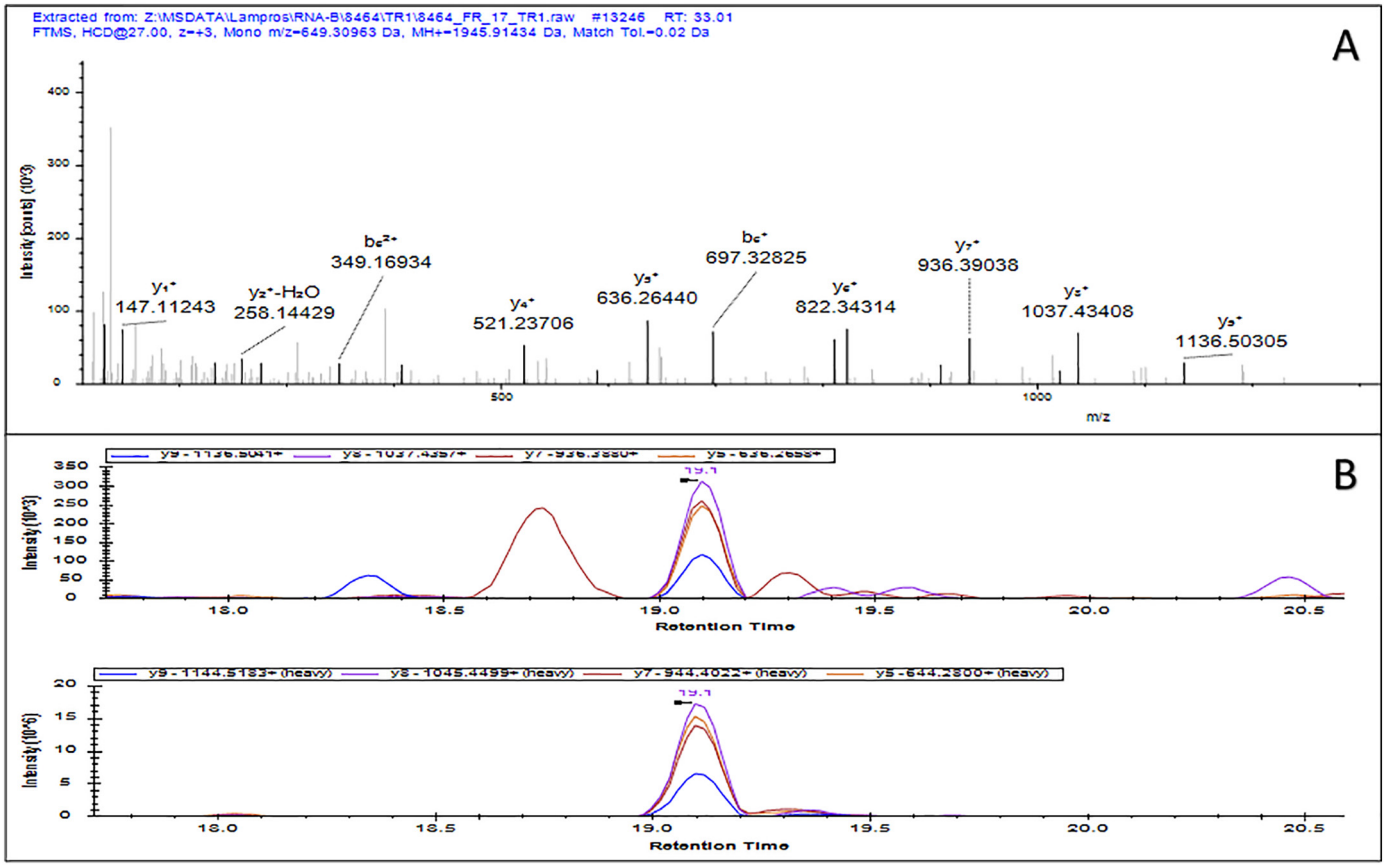

Fig. 2. Variant peptide YPIEHGIVTNWDNMEK (p.D81N) detected in the shotgun analysis and validated with the targeted MRM assay. (A) Annotated MS/MS spectrum. (B) Same relative transition intensities for both heavy isotopically labeled peptides and light peptides along with their co-elution confirm the presence of the peptide. Upper chromatogram represents the endogenous light peptide and lower chromatogram represents the heavy labeled spiked peptide standard.

effect on protein abundance, even though the respective variant peptides were not detected. For that purpose, 626 scatter plots representing each one of the transcript-protein pairs with at least one mutation detected by sequencing were created and were visually inspected for outliers. Specifically, log2 transformed FPKM values were plotted against log2 transformed summed intensity area values for all 21 samples.

Plots with readily visible outliers representing only the samples with the sequencing-predicted mutations and not having outliers from other samples were sought. Three such plots were selected and subsequent statistical analysis confirmed that in two of them, the data points that appeared to be outliers significantly deviated from the predicted regression line since the residual values for these data points exceeded by at least two times the standard deviation of all the remaining values. These two outliers were found in proteins HSPD1 and VPS35, while the data point in protein CTSH was not confirmed to be an outlier statistically. The scatter plots in Fig. 6 show the regression lines for the three transcript-protein pairs and the respective outlier data points. The two outliers in proteins HSPD1 and VPS35 may indicate that the mutation in these proteins has an effect on protein abundance and in these two cases may show that the variant proteins are less stable and are subsequently degraded.

3.4. Sensitivity of single amino acid variant (SAAV) detection

Three single amino acid variants were detected in the 21 breast tissue specimens out of 789 possible unique single amino acid variants, resulting in 0.4% sensitivity of sequencing-predicted somatic variant detection. However, not all of these mutations can be detected with the present workflow due to a number of inherent biological and technical limitations that hamper proteogenomic quests for variant peptides based on bottom-up proteomics.

To take into consideration these limitations and calculate a more realistic sensitivity of variant peptide detection, we narrowed the list with the 789 mutations down to those that are in theory detectable by shotgun proteogenomics. To this end, redundant variant peptides arising from various isoforms of the same protein were removed and both variant peptides originating from mutations generating a tryptic site were included. Following this step, all tryptic peptides starting with a proline or from which the succeeding amino acid is a proline were removed. Truncation mutations creating stop codons immediately after an arginine or lysine codon cannot be observed by mass spectrometry-based proteomics because even if the truncated protein is produced it will not have any unique peptide able to distinguish it from the wild type full length protein. Therefore, such mutations were removed as well. Finally, all variant peptides remained were filtered to keep only peptides with length that is suitable for shotgun mass spectrometric identification. Thus, only peptides with 6 up to 83 amino acids were included in the final list with all theoretical mass-spectrometry detectable peptides. Peptides with < 6 amino acids are not long enough to provide MS/MS spectra of good quality for reliable matching with the theoretical spectra and peptides longer than 83 amino acids have never been observed according to the Peptide Atlas Human peptide build (January 2016, 1202 samples, 1,166,164 peptides, PSM FDR = 0.0002). The final list contains 482 variant peptides that can be theoretically detected by MS/MS spectra matching. This reduction increases the sensitivity of variant peptide detection to 0.6%.

The next question we wanted to answer was how many of the proteins corresponding to these 482 peptides would be abundant enough to be detected in the respective samples with the sequencing-

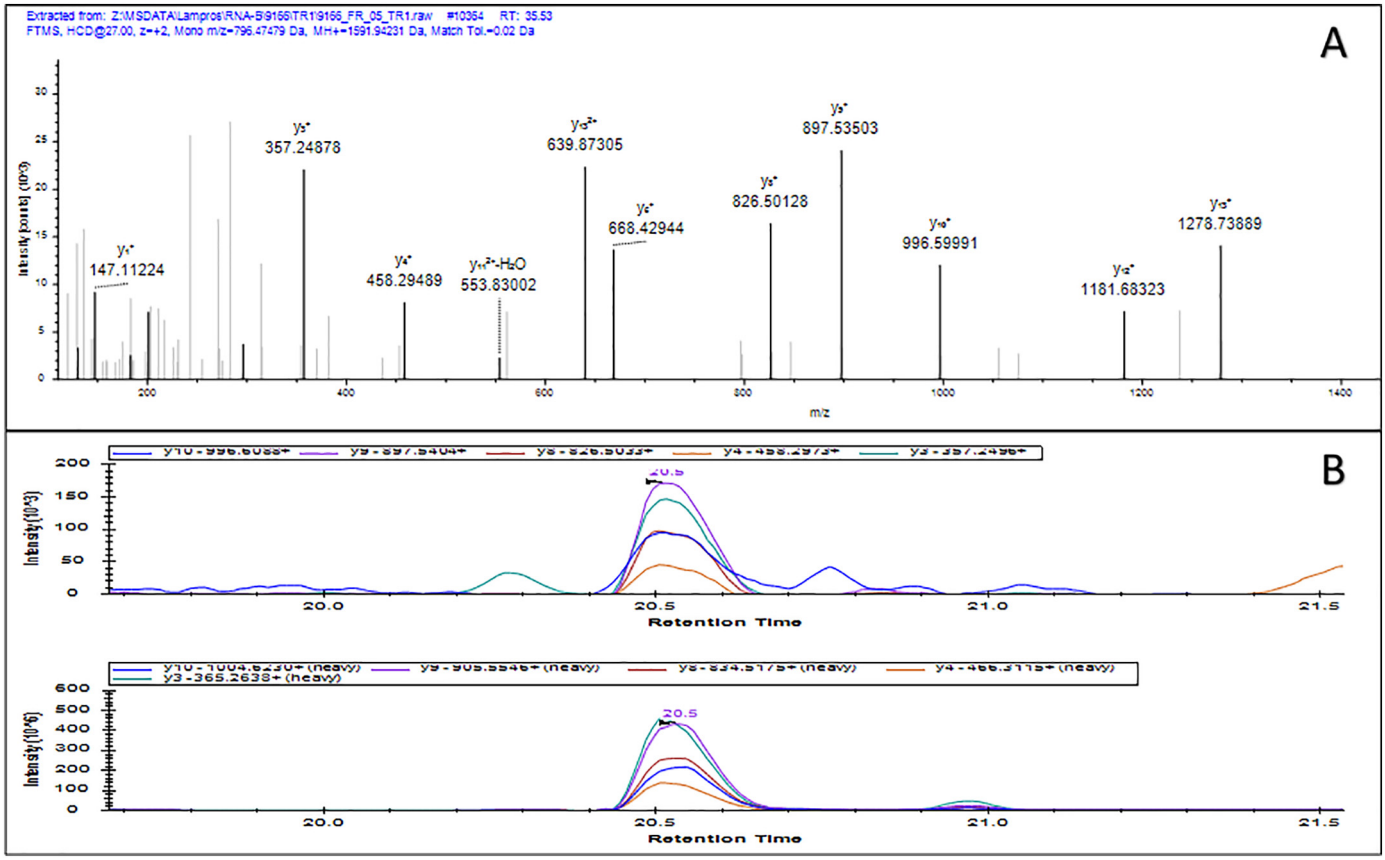

Fig. 3. Variant peptide ISLPGQVAGTPITPLK (p.M219V) detected in the shotgun analysis and validated with the targeted MRM assay. (A) Annotated MS/MS spectrum. (B) Same relative transition intensities for both heavy isotopically labeled peptides and light peptides along with their co-elution confirm the presence of the peptide. Upper chromatogram represents the endogenous light peptide and lower chromatogram represents the heavy labeled spiked peptide standard.

predicted mutation. We reasoned that when no other peptides (termed here "wild type" peptides) were detected for a given protein, the variant peptide would be highly unlike to be observed-due to low abundance, even if its characteristics would theoretically allow its mass spectrometric detection. After filtering the 482 proteins, we identified 103 with at least one wild type peptide detected in the respective samples. The wild type peptides identified for each of the 103 proteins ranged from only one peptide to 43, whereas the protein coverage ranged from 0.39% to 71.73% and the summed intensity area from 631,950 to 215,400,000,000. Not surprisingly, the 3 proteins with the variant peptides detected and validated were among the most abundant of the 103 proteins listed. Table 5 shows where exactly these proteins rank in terms of peptide summed intensity areas as well as their coverage, wild type peptide identifications and peptide-spectrum matches, whereas the complete list with all the 103 proteins is shown in Supporting Table 2. Additionally, Supporting Table 3 depicts the proportion of proteins detected with at least one wild type peptide from those that had at least one MS detectable variant peptide based on WGS data for each sample.

Subsequently, we went through all the 103 proteins to identify wild type peptides corresponding to the variant peptides predicted by DNA sequencing. These wild type peptides either differ from the variant peptides by only one amino acid (the substituted amino acid) or in the case that a tryptic site is created or lost they partially overlap with the variant peptide. In total, 9 such wild type peptides were detected in seven samples and are depicted in Table 6. Two of these wild type peptides correspond to the nucleolin and actin variant peptides detected in samples PD6422 and PD4604 samples respectively.

Lastly, we aimed to identify how many of the 93 breast cancer genes previously identified [3], were detected here at the proteome level. These results are summarized in Supporting Table 4.

3.5. Database repository search

None of the three mutations detected and validated at the proteome level has been deposited in the online databases searched. Specifically, no DNA mutations resulting in any of the three protein variants (NCL V404D, ACTB D81N and GORASP2 V364 M) are found in any of the tumor samples sequenced and deposited in TumorPortal, cBioPortal or the non-cancer GO Exome and the ExAC repository. However, the DNA mutations from which these three protein variants arise, (NCL 1211T > A, ACTB 241G > A and GORASP2 655A > G) are deposited in the COSMIC database [10] and the samples are the very same that were used in the present study to acquire the sequencing [3] and proteomic data. As a result and to the best of our knowledge, these amino acid variations are observed for the first time in the literature in this study.

4. Discussion

An important question pertinent to proteogenomic studies is what portion of sequencing-predicted somatic variants can be detected at the proteome level by data-dependent MS/MS analysis. On which biology-related and technology-related factors this depends we aimed to dissect in this paper. To serve that purpose, here we describe the effort for an integrated analysis of whole genome DNA sequencing (WGS) and RNA-sequencing data [3,4] with matched MS/MS raw data that were acquired here from the same 21 ER+ breast cancer tissues from which sequencing results were previously obtained.

All samples included in this study were immunohistochemically defined as estrogen receptor positive and this was confirmed with 16 of the samples having evidence for estrogen receptor expression at the protein level. Not surprisingly, the 5 samples with no ESR1 peptides

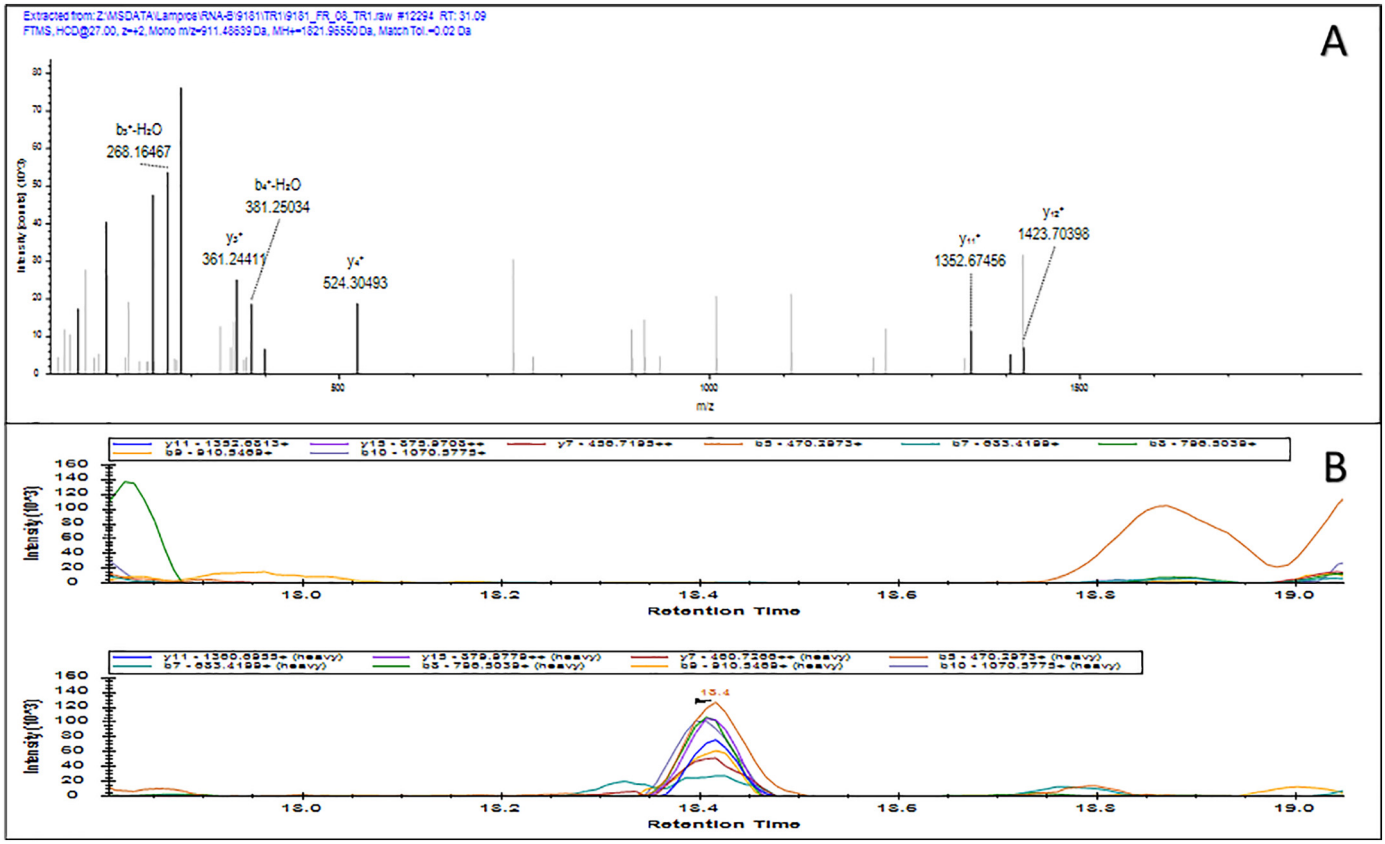

Fig. 4. Variant peptide AITIAGRINCPMYITK (p.V364M) detected in the shotgun analysis but failed to be validated with the targeted MRM assay. (A) Annotated MS/MS spectrum. (B) No peak for the endogenous light peptide confirm the absence of the peptide. Upper chromatogram represents the endogenous light peptide and lower chromatogram represents the heavy labeled spiked peptide standard.

detected were the 5 samples with the least overall protein identifications (with the exception of sample 8955).

Somatic mutations based on whole genome sequencing ranged from 1076 in sample 9152 to 25,953 in sample PD4607. Accordingly, seven variants fell within the protein coding regions in the former sample and 203 in the latter. From a total of 789 unique SAAVs across the 21 samples, four were detected from the MS/MS raw spectra obtained and searched against the customized database with all the reference protein entries and the variant entries appended. Two of the variants detected were triply charged and another two doubly charged. However, one of the doubly charged tryptic peptides in sample PD9845 included a cleavage which was missed and was expected to be identified with a +3 charge. Indeed, in the subsequent validation experiments with targeted multiple reaction monitoring assays this peptide failed to be validated as opposed to the other 3 peptides in samples PD6422, PD4604 and PD9842. The quality of the MS/MS spectrum that matched with the theoretical peptide that failed to be validated was very poor with only a few peaks of low intensity (Fig. 4). Thus, the low Xcorr value and the borderline q-value and PEP values for this peptide are justified (Table 2). Although the number of variant peptides identified in this study is not large, the percentage of variant peptides validated here by targeted MRM (75%) is comparable with the 45.9% previously reported [14]. Of note, the three variant peptides were detected in the three samples with the most total protein identifications and with relatively high invasive tumor cell percentage (80%, 80% and 98%).

The sensitivity of sequencing-predicted variant detection at the proteome level is an important index central to the concept of proteogenomics [15]. Even though not all variants identified by sequencing appear as gene products, the ability to detect as many variants possible among the ones actually translated is highly desirable, provided that meaningful clinical information can be extracted from such variants and the development of translational applications based on that information can be realized [16]. Here, the initial sensitivity of variant peptide detection was 0.4%, which after filtering out peptides that are theoretically impossible to be observed due to inherent technological limitations of the bottom-up proteogenomic approach employed was increased to 0.6%. Subsequent filtering to keep only proteins that are detectable at the protein level with at least one wild type peptide detected, further increased this theoretical sensitivity of SAAV detection to 2.9%. Our results are concordant with recent attempts to probe the sensitivity of sequencing-predicted variants at the proteome level. In the first large integrative onco-proteogenomic study reported, 796 single amino acid variants (both germline and somatic) were detected in a cohort of 86 colon and rectal cancer samples from matched RNA-sequencing and MS/MS data [15]. Four of the 796 variants were subsequently successfully validated by targeted PRM assays [15]. Sixty four of the variants were reported as somatic by TCGA and 101 by COSMIC [15]. In a similar approach in 169 high-grade serous ovarian tumors, a customized TCGA genomic sequencing-derived database multi-stage search revealed 1217 germline and somatic single amino acid variants at the proteome level [17]. A small subset of these variants were selected randomly and were validated when MS/MS spectra of synthetic peptide standards were obtained [17]. Two recent onco-proteogenomic studies in breast cancer, not only detected sequencing-predicted variants at the proteome level, but discriminated in germline and somatic mutations as well [18,19]. In one of them, 77 breast cancer samples were analyzed and 89 out of 4574 variants predicted from whole exome sequencing were detected at the proteome level resulting on 1.9% sensitivity [18]. Even though whole exome sequencing contributes to increased variant identification at the protein coding regions when compared to whole genome sequencing, the somewhat increased sensitivity of SAAVs compared to our study can be attributed to a

Table 3
All transcripts with a sequencing-predicted mutation ranked in terms of decreasing abundance in samples PD6422, PD4604 and PD9842.

PD6422		PD4604		PD9842	
Transcript	FPKM	Transcript	FPKM	Transcript	FPKM
NCL	7.018	ACTB	8.5309	GORASP2	4.9041
GSDMB	6.0946	IFI16	5.3574	NEK10	4.2383
ANKMY1	5.6982	KIAA1109	4.5253	TTC27	3.7903
EPN3	5.6462	CFH	4.3644	OBSL1	3.0045
SMC1A	5.3371	ANKRD17	4.1798	FAM122A	2.9306
SYNM	4.8455	IGSF6	3.6887	CELSR2	2.6991
MDC1	4.7697	PARD3	3.6318	RC3H1	2.4251
SFRS12	4.3343	TMEM168	3.6276	WDR78	2.3234
FAT1	4.0855	NEGR1	3.5681	ZNF736	2.1939
CELSR3	3.7234	MPP6	3.5257	MAP3K1	2.0738
TPRG1	3.7058	MICAL2	3.1198	AHCTF1	1.8306
HELB	3.4165	RRAGC	3.108	ZNF407	1.5552
WDR33	3.3927	VPS13A	3.073	SUOX	1.4564
DDX21	3.2749	ARHGEF11	2.9911	TJP3	1.2403
GFM2	2.953	DCK	2.9864	FAM126A	1.004
ITGAL	2.9524	PCDHGA12	2.8497	ARHGEF10L	0.9081
BMP1	2.8016	SCMH1	2.7686	MYO1F	−0.1703
NEURL4	2.7699	CHUK	2.5513	PREX2	−0.299
PTEN	2.603	EGFR	2.4327	MAST3	−0.8218
VPS13C	2.1612	ABCC1	2.3872	CXorf27	−2.6674
DIS3L	1.6881	TP53	2.3364	E2F8	−3.3775
ZFHX4	0.6188	AIM1	2.3061	CACNA1A	−3.8364
TPBG	0.4058	CENPJ	1.9522	DNAH6	−4.1182
SCYL3	0.2184	TTN	1.8265	FAT2	−7.6103
C9orf167	–0.046	RET	1.7408		
THG1L	−0.6111	GREB1L	1.4232		
KIF1A	−2.6145	ZNF473	1.2722		
IGDCC4	−2.6969	PROM1	1.1375		
ZNF648	−3.308	KIF4A	0.9142		
TBX21	−3.5168	ZFP112	0.8294		
DSCAM	−4.5739	ZNF175	0.5966		
CALY	−4.5819	EVPL	0.5655		
ADAM29	−4.7977	ARAP3	0.3632		
DUSP9	−5.5505	ADAMTS16	0.2631		
PRR23C	−6.4056	NAP1L2	−0.265		
ACOT12	−6.5304	C1orf61	−0.7046		
FSTL5	−6.772	TYRP1	−0.9239		
WDR87	−7.2219	ARHGEF4	−1.9043		
		ACADL	−2.4034		
		C8orf47	−2.7459		
		PRSS36	−2.8042		
		SCN4A	−2.8734		
		KIF19	−3.5665		
		LAMB4	−4.3209		
		GABRR1	−4.4488		
		FER1L6	−5.1713		
		DLGAP3	−5.3463		
		PKHD1	−5.5919		
		C12orf50	−5.6195		
		CCDC135	−5.6429		
		CXorf30	−5.8672		
		CNGB3	−6.3154		
		PNPLA1	−6.578		
		ZFP42	−7.136		

number of factors. Specifically, more extensive offline fractionation, longer gradient during the data dependent acquisition of MS/MS spectra, the three biological replicates processed and the lack of targeted validation may account for the increased sensitivity observed. The second study did not span a large number of individual tumors but on the contrary focused on two individual breast-cancer-patient-derived xenografts (PDX), one derived from a luminal and one from a basal breast tumor [19]. However, in this case very deep proteomic coverage was sought by extensive offline reversed-phase fractionation (15–25 fractions) and 48 process replicates that were obtained with two different mass spectrometers [19]. The matched whole genome sequencing data provided the sample-specific customized databases and the searches conducted revealed 5 somatic variant peptides in the basal and

Table 4
NCL, ACTB and GORASP2 transcripts ranked in terms of abundance in the 21 breast tissues (Out of a total of 20,188 transcripts).

Sample	NCL	ACTB	GORASP2	Sample	NCL	ACTB	GORASP2
PD6418	53rd	55th	3381st	PD13626	101st	27th	2142nd
PD6422	**140th**	29th	3926th	PD9842	90th	51st	**1265th**
PD4607	74th	57th	1626th	PD9847	69th	56th	1113rd
PD4613	49th	35th	2893rd	PD9845	106th	45th	1242nd
PD4604	158th	**66th**	2779th	PD14432	104th	37th	2353rd
PD7238	89th	34th	1601st	PD14435	123rd	62nd	3597th
PD13625	152nd	18th	1642nd	PD14437	118th	36th	2230th
PD13623	94th	103rd	1427th	PD14439	117th	73rd	3207th
PD13631	148th	27th	1458th	PD14441	185th	93rd	2652nd
PD13629	133rd	20th	1571st	PD14442	226th	65th	1566th
PD13619	158th	34th	1360th	–	–	–	–

For all protein coding gene transcripts the total amount was taken into account. No discrimination was made for the wild type and the mutant form for each transcript.

140 somatic variant peptides in the luminal sample [19]. Taking under consideration that the somatic variants predicted by WGS were 5371 and 11,973 respectively, the sensitivities of variant detection were 0.1% for the basal tumor and 1.2% for the luminal tumor revealing a remarkable difference between the two breast cancer intrinsic subtypes. The authors attributed this difference in differential translation and increased protein degradation effects in the basal tumor. Our study revealed a sensitivity that was in between these two values (0.4% before and 0.6% after adjusting for detectability by MS). However, our cohort was comprised of 21 ER-positive breast tumors that are predominantly luminal (all except two samples) [20], therefore this lower sensitivity might be attributed to the absence of targeted validation (which is expected to lower the sensitivity) and the much deeper coverage of the xenograft study [19].

When gene transcripts in each sample were ranked in terms of abundance, the transcripts of NCL, ACTB and GORASP2 ranked first among the transcripts with a sequencing-predicted mutation in the samples that the respective variant peptide was detected, while they exhibited a uniform pattern of expression across the 21 samples with slight between sample variations and irrespective of mutational status, when all transcripts were ranked in each sample. Interestingly, nucleolin and actin transcripts ranked among the 250 most abundant transcripts in all 21 samples. Taking under consideration the high coverages and intensity areas that characterized the three proteins in the samples their surrogate variant peptide was detected, a trend towards a positive correlation between transcript and protein abundance and variant peptide detection can be inferred.

The high protein abundance of nucleolin and actin was further ascertained when apart from their variant peptides detected, another 25 wild type tryptic peptides were detected for both of them. Remarkably, the wild type peptides corresponding to the variant peptides were detected as well in both cases (they are identical except for the altered amino acid), thus likely denoting a heterozygous status for these mutations. Heterozygogocity most likely also explains the remaining seven wild type peptides detected in this study corresponding to variant peptides predicted based on somatic variants present in the particular specimen. Since the wildtype is detected but the predicted variant is not, one would assume that in most of these cases the mutant proteins are posttranslationally degraded or the mutant transcript destabilized, because in case they are expected to be produced in similar to wild type protein amounts, the high similarity of these variant peptides to their wild type counterparts should have allowed their shotgun mass-spectrometric detection. An alternative explanation is that the wild type detection is due to stromal contamination of the specimen. To test the former hypothesis, we calculated the mRNA-to-protein correlation across the 21 samples. The global correlation focused only on the most abundant transcripts and proteins that had non-zero values for all

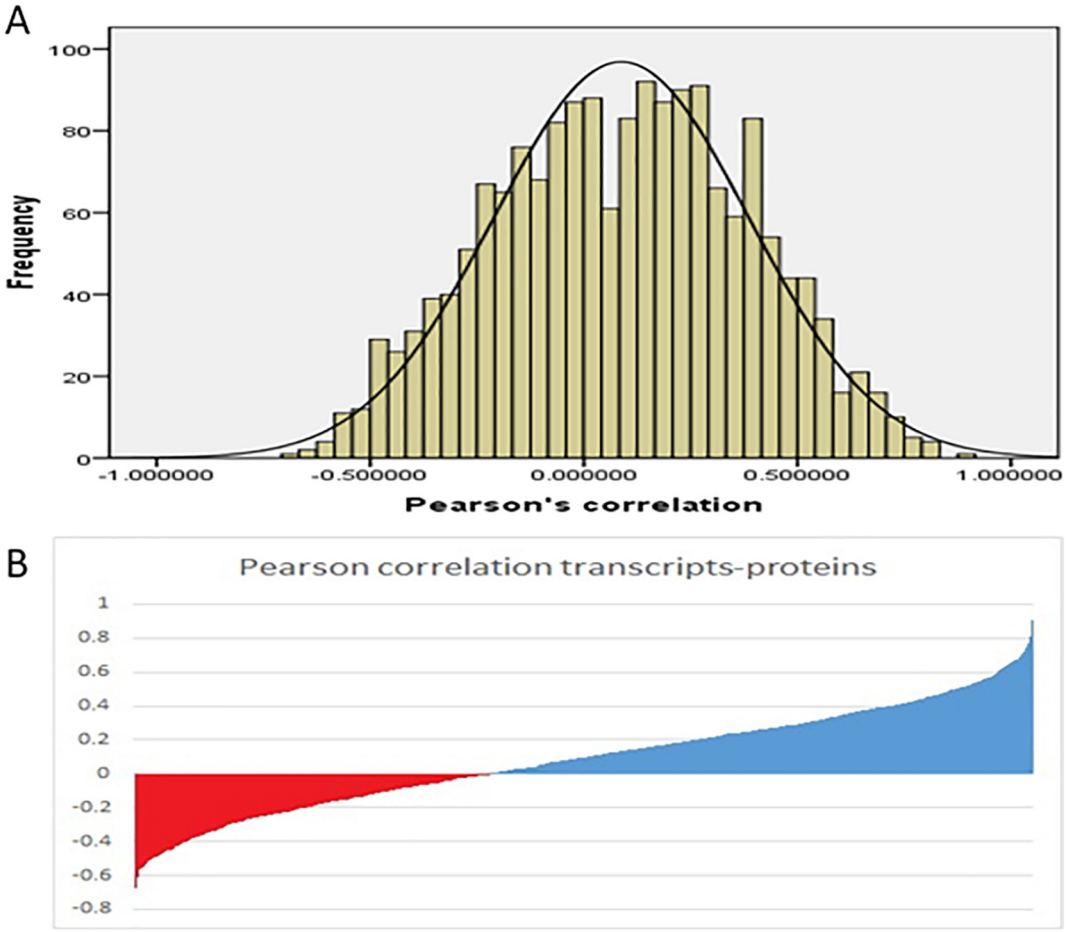

Fig. 5. Correlations between mRNA and protein abundance across the 21 breast tissue samples. (A) Frequency histograms for Pearson correlation coefficients. (B) 60.3% of the transcript-protein pairs exhibited positive Pearson correlation.

samples and was comparable to previously reported correlation coefficients [15,17,18]. When individual transcript-protein pairs were examined to find samples having a sequencing-predicted mutation and significantly deregulated protein abundance when compared to the rest of the samples, the results revealed two such proteins (HSPD1 and VPS35). VPS35 was detected in sample 9922 with 4 unique wild type peptides. The sequencing-predicted mutation is a truncating mutation (p.S395*). Only one of the four peptides is located before the position of the truncation and could originate from the truncated protein as well. Indeed, that peptide was found to be approximately 1.5 orders of magnitude more abundant than any of the other three peptides. This suggest the truncated protein is likely produced. HSPD1 protein was found to be significantly downregulated in sample 9144 as well and 15 unique wild type peptides were detected in that sample, however the variant peptide is bound to be missed by current bottom-up proteomics due to a proline following the tryptic site. Nevertheless, the significantly reduced protein abundance could be attributed to (proteasomal) degradation of the variant HSPD1 protein. Both wild type and the corresponding variant peptides of NCL and ACTB proteins were detected and the two proteins were among the most abundant ones as already mentioned. Thus, it is well justified that no outliers were spotted in the transcript-protein scatter plots representing these two proteins. As far as the remaining 7 proteins in which the wild type peptides (corresponding to the variant ones) were detected and based on the protein abundance and the MS "friendliness" of the variant peptides, the detection of the latter would be expected, no outliers representing the samples with the respective mutations were observed. As a result, we can draw the conclusion that the variant proteins in these samples are either not translated or translated and degraded. In both cases the wild type proteins could compensate for this absence of the variant proteins. The possibility that the variant proteins are translated, but either their lower abundance when compared with their wild type counterparts or the slight difference the variant peptide has compared to the wild type one in combination with the nature of peptide sampling based on MS1 spectra in a shotgun proteogenomic experiment hamper their detection, cannot be excluded as well.

Finally, none of the three mutations reported here has ever been observed before, either at the DNA or protein level. The stringent variant calling filtering criteria from the WGS data and the proteomic detection and validation of these variants provide indisputable evidence for their existence.

Our main goal was to identify factors that affect the sensitivity of variant peptide detection. From a biological point of view a variant detected at the DNA level may not be observed at the proteome level for a variety of reasons. Firstly, it goes without saying that the more abundant a given variant peptide is, the more probable it is to be detected. Of note, each of the 93 breast cancer genes previously identified [3], is detectable at the protein level in approximately 1 out of 3 samples (median value) and approximately 40% of breast cancer genes (median value) are detectable at the protein level in any given sample (Supporting Table 4). While some known driver genes highly mutated in cancer were detected in very few or none of the 21 samples in this study (e.g. BRAF, BRCA1, BRCA2, TP53), some others were abundant enough at the proteome level to be detected in almost all samples (e.g. CDH1, GATA3, KRAS, PIK3CA). This data shows that some important genes may be entirely missed at the proteome level even if they are

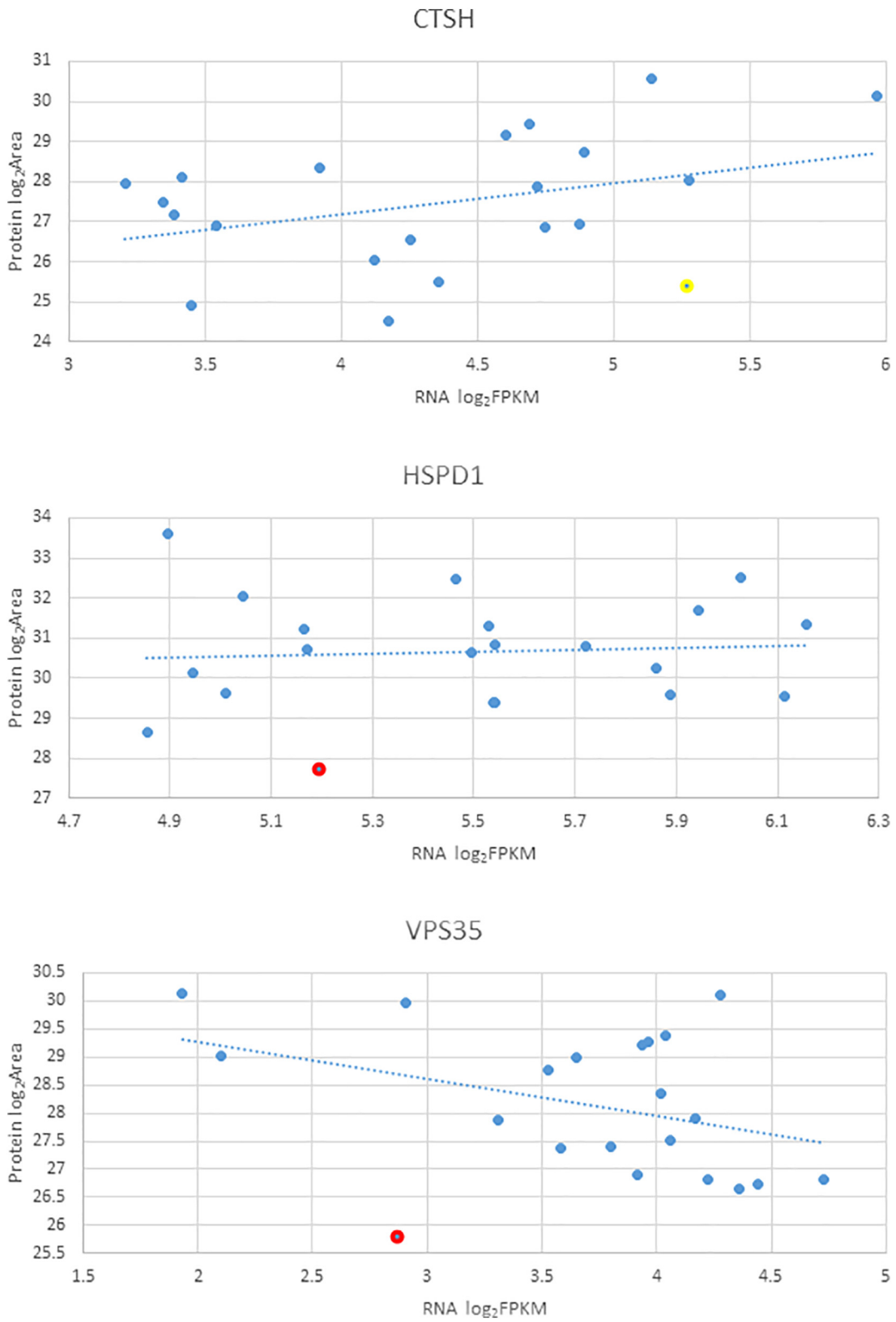

Fig. 6. Regression analysis between mRNA and protein abundance for three transcript-protein pairs. (A) Data point at (5.263, 25.415, yellow colored) for protein CTSH (p.R111*) was not found to be at least two standard deviations below the best-fit line and failed to be flagged as an outlier. (B) Data point at (5.192, 27.740, red colored) for protein HSPD1 (p.R268H) was found to be more than two standard deviations below the best-fit line and was flagged as an outlier. (C) Data point at (2.863, 25.806, red colored) for protein VPS35 (p.S395*) was found to be more than two standard deviations below the best-fit line and was flagged as an outlier. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
NCL, ACTB and GORASP2 proteins were among the 103 most abundant proteins with at least one sequencing-predicted mass spectrometry "friendly" variant peptide and one wild type peptide detected with the MS/MS analysis.

Protein	Coverage	Peptide #	PSMs #	Intensity area	Rank
ACTB	71.73%	25	637	2.15E+11	1st
NCL	37.32%	25	80	3.35E+09	4th
GORASP2	19.91%	5	13	4.36E+08	17th

Rank refers to the protein position when all 103 proteins were ranked in descending summed intensity areas.

translated due to abundance reasons. However, it is still encouraging that a significant number of known driver genes may be detectable at the proteome level. Secondly, some variant proteins may be translated but they may be unstable or targeted for degradation by the cell machinery [21]. Additionally, some variant proteins may never appear as protein gene products due to negative regulation at the transcriptional or the translational level (e.g. eQTLs and pQTLs) [22–24]. Finally, even though RNA-sequencing may sound as a better option for proteogenomic experiments since RNA is more proximal to the proteins, the correlation between transcript and protein abundance is not absolute [24]. In the case of variants, here we have shown that there seems to be a degree of association between transcript abundance and variant protein expression and variant peptide detection, but studies focusing on the investigation of the extent of such correlation remain to be done.

When it comes to technological limitations that affect the detection of variant peptides, they are mainly inherent limitations of bottom-up proteomics approaches utilized in onco-proteogenomic studies as well [25]. Specifically, ion suppression observed when peptide species highly similar to the variant peptides (e.g. wild type peptides) compete for detection or there is competition with other co-eluting matrix components is one of the main factors that hamper the detection of variant peptides [26]. Additionally, biases against the identification of highly hydrophobic of hydrophilic peptides and peptides exhibiting poor ionization or fragmentation behavior add up to the limiting factors. Although in relatively complex matrices it is not trivial to predict the detectability of any given peptide, it has been demonstrated that in simple matrices the majority of peptides present can be classified as either detectable or not detectable with only a very small minority of the peptides being characterized by intermediate detectability [27]. It would be rational to hypothesize that the same principles govern the detection of variant peptides and therefore certain variant peptides are bound to be missed by proteogenomic studies. Finally, false negatives of the database search is another source of missed proteogenomic identifications as well. Of note, since there is no way to always reliably distinguish between the biological or technological origin of the unobserved proteogenomic events at the proteome level, their actual biological significance will remain unknown [17].

5. Conclusions

Overall, it is expected that variant peptides abundant enough to be identified with a shotgun proteogenomic approach will be subsequently validated in a targeted fashion in the same fractions that were initially detected, while they may or may not be detected in the non-enriched samples. On the contrary, peptides not discovered during the shotgun proteogenomic experiment, will be entirely missed, even though they might have been abundant enough to be detected with a targeted approach. Additionally, variant peptide detection depends on the bioinformatics specifics of each proteogenomic study, namely the nature of nucleotide sequencing raw data (whole genome, exome, RNA-sequencing), the filtering criteria used to call variants and create the customized sample-specific database and the search algorithm used to match MS/MS spectra with the theoretical ones.

To recapitulate, we conclude that the sensitivity of variant peptide detection in a proteogenomic quest largely depends on two factors which are the abundance of the respective transcripts and proteins and the "compatibility" of any given variant peptide to be detected by mass spectrometry. Even though many recent proteogenomic studies have shed light on the sensitivity of variant peptide detection and to the extent that mRNA and protein abundances correlate [15,17–19,28], whether the correlation between variant transcripts and variant proteins is similar and the overall view of the flow of protein coding variants from DNA to RNA to proteins remains to be elucidated.

Competing interests

The authors declare that they have no competing interests.

Research funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics approval

This study used coded tumor tissues in accordance with the Code of Conduct of the Federation of Medical Scientific Societies in the Netherlands (http://www.federa.org/codes-conduct). In addition, the protocol to study biological markers associated with disease outcome was approved by the medical ethics committee of the Erasmus Medical Center Rotterdam, The Netherlands (MEC 02.953).

Authors' contributions

LD, IP, EPD and GSC conceived and designed the experiments. AMS and JWMM provided the tissue sections. LD performed the experiments. LD analyzed the data and together with IP wrote the manuscript. All the authors have accepted responsibility for the entire content of this submitted article and approved submission.

Table 6
Nine wild type peptides corresponding to predicted variant peptides were detected in 7 samples.

Sample	Protein	Mutation	Wild type peptide	Variant peptide
PD6418	PTRF	p.V52M	SDQVnGVLVLSLLDK	SDQVNGMLVLSLLDK
PD6422	NCL	p.V404D	VTQDELKEVFEDAAEIR	DTQDELKEVFEDAAEIR
PD4607	HSP90B1	p.E130K	LISLTDENALSGNEELTVK	LISLTDENALSGNK(ELTVK)
PD4607	TSNAX	p.E197K	VTPVDYLLGVADLTGELmR	VTPVDYLLGVADLTGK(LMR)
PD4607	ACSL3	p.E569Q	WLCTGDIGEFEPDGCLK	WLCTGDIGEFQPDGCLK
PD4604	ACTB	p.D81N	YPIEHGIVTNWDDMEK	YPIEHGIVTNWDNmEK
PD7238	FUT11	p.S239P	HIPVDSYGK	HIPVDPYGK
PD13625	ARAF	p.A601T	TQADELPACLLSAAR	TQADELPACLLSTAR
PD14435	SLC12A7	p.E106K	LANYTNLSQGVVEHEEDEESR	LANYTNLSQGVVEHEEDK(ESR)

NCL and ACTB variant peptides (highlighted) were detected and validated with the targeted MRM assays. n = deamidated asparagine, m = oxidized methionine.

Acknowledgements

We would like to thank Serena-Nik Zainal for kindly providing the processed genomics and transcriptomics data.

Data availability

The mass spectrometry proteomics data and the search results output files have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD007797. Reviewer account details (http://www.ebi.ac.uk/pride/archive/login): Username: reviewer00978@ebi.ac.uk, Password: ObctSqW7. MRM raw data have been deposited to the PeptideAtlas SRM Experiment Library (PASSEL) with the dataset identifier PASS01094. Username and password for reviewer access (https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/PASS_View): Username: PASS01094, Password: KX7787isc.

References

[1] F. Beca, K. Polyak, Intratumor heterogeneity in breast cancer, Adv. Exp. Med. Biol. 882 (2016) 169–189.
[2] B. Royer-Bertrand, C. Rivolta, Whole genome sequencing as a means to assess pathogenic mutations in medical genetics and cancer, Cell. Mol. Life Sci. 72 (8) (2015) 1463–1471.
[3] S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L.B. Alexandrov, S. Martin, D.C. Wedge, et al., Landscape of somatic mutations in 560 breast cancer whole-genome sequences, Nature 534 (7605) (2016) 47–54.
[4] M. Smid, F.G. Rodriguez-Gonzalez, A.M. Sieuwerts, R. Salgado, W.J. Prager-Van der Smissen, M.V. Vlugt-Daane, A. van Galen, S. Nik-Zainal, J. Staaf, A.B. Brinkman, et al., Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration, Nat. Commun. 7 (2016) 12910.
[5] A.M. Sieuwerts, M.E. Meijer-van Gelder, M. Timmermans, A.M. Trapman, R.R. Garcia, M. Arnold, A.J. Goedheer, H. Portengen, J.G. Klijn, J.A. Foekens, How ADAM-9 and ADAM-11 differentially from estrogen receptor predict response to tamoxifen treatment in patients with recurrent breast cancer: a retrospective study, Clin. Cancer Res. 11 (20) (2005) 7311–7321.
[6] J. Liu, A.M. Sieuwerts, M.P. Look, M. van der Vlugt-Daane, M.E. Meijer-van Gelder, J.A. Foekens, A. Hollestelle, J.W. Martens, The 29.5 kb APOBEC3B deletion polymorphism is not associated with clinical outcome of breast cancer, PLoS One 11 (8) (2016) e0161731.
[7] R.B. Braakman, A.M. Sieuwerts, A. Umar, Shotgun proteomics on tissue specimens extracted with Acid guanidinium-thiocyanate-phenol-chloroform, Methods Mol. Biol. 1293 (2015) 115–122.
[8] K.G. Kline, B. Frewen, M.R. Bristow, M.J. Maccoss, C.C. Wu, High quality catalog of proteotypic peptides from human heart, J. Proteome Res. 7 (11) (2008) 5055–5061.
[9] K.C. Chan, H.J. Issaq, Fractionation of peptides by strong cation-exchange liquid chromatography, Methods Mol. Biol. 1002 (2013) 311–315.
[10] S.A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C.G. Cole, S. Ward, E. Dawson, L. Ponting, et al., COSMIC: somatic cancer genetics at high-resolution, Nucleic Acids Res. 45 (D1) (2017) D777–D783.
[11] M.S. Lawrence, P. Stojanov, C.H. Mermel, J.T. Robinson, L.A. Garraway, T.R. Golub, M. Meyerson, S.B. Gabriel, E.S. Lander, G. Getz, Discovery and saturation analysis of cancer genes across 21 tumour types, Nature 505 (7484) (2014) 495–501.
[12] J. Gao, B.A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S.O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, et al., Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal, Sci. Signal. 6 (269) (2013) pl1.
[13] M. Lek, K.J. Karczewski, E.V. Minikel, K.E. Samocha, E. Banks, T. Fennell, A.H. O'Donnell-Luria, J.S. Ware, A.J. Hill, B.B. Cummings, et al., Analysis of protein-coding genetic variation in 60,706 humans, Nature 536 (7616) (2016) 285–291.
[14] H. Sun, C. Chen, B. Lian, M. Zhang, X. Wang, B. Zhang, Y.X. Li, P. Yang, L. Xie, Identification of HPV integration and gene mutation in HeLa cell line by integrated analysis of RNA-Seq and MS/MS data, J. Proteome Res. 14 (4) (2015) 1678–1686.
[15] B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M.C. Chambers, L.J. Zimmerman, K.F. Shaddox, S. Kim, et al., Proteogenomic characterization of human colon and rectal cancer, Nature 513 (7518) (2014) 382–387.
[16] J.A. Alfaro, A. Sinha, T. Kislinger, P.C. Boutros, Onco-proteogenomics: cancer proteomics joins forces with genomics, Nat. Methods 11 (11) (2014) 1107–1113.
[17] H. Zhang, T. Liu, Z. Zhang, S.H. Payne, B. Zhang, J.E. McDermott, J.Y. Zhou, V.A. Petyuk, L. Chen, D. Ray, et al., Integrated proteogenomic characterization of human high-grade serous ovarian cancer, Cell 166 (3) (2016) 755–765.
[18] P. Mertins, D.R. Mani, K.V. Ruggles, M.A. Gillette, K.R. Clauser, P. Wang, X. Wang, J.W. Qiao, S. Cao, F. Petralia, et al., Proteogenomics connects somatic mutations to signalling in breast cancer, Nature 534 (7605) (2016) 55–62.
[19] K.V. Ruggles, Z. Tang, X. Wang, H. Grover, M. Askenazi, J. Teubl, S. Cao, M.D. McLellan, K.R. Clauser, D.L. Tabb, et al., An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer, Mol. Cell. Proteomics 15 (3) (2016) 1060–1071.
[20] B. Weigelt, F.C. Geyer, J.S. Reis-Filho, Histological types of breast cancer: how special are they? Mol. Oncol. 4 (3) (2010) 192–208.
[21] W.K. Kim, M. Park, M. Park, Y.J. Kim, N. Shin, H.K. Kim, K.T. You, H. Kim, Identification and selective degradation of neopeptide-containing truncated mutant proteins in the tumors with high microsatellite instability, Clin. Cancer Res. 19 (13) (2013) 3369–3382.
[22] K.V. Ruggles, K. Krug, X. Wang, K.R. Clauser, J. Wang, S.H. Payne, D. Fenyo, B. Zhang, D.R. Mani, Methods, tools and current perspectives in proteogenomics, Mol. Cell. Proteomics 16 (6) (2017) 959–981.
[23] A. Battle, Z. Khan, S.H. Wang, A. Mitrano, M.J. Ford, J.K. Pritchard, Y. Gilad, Genomic variation. Impact of regulatory variation from RNA to protein, Science 347 (6222) (2015) 664–667.
[24] Y. Liu, A. Beyer, R. Aebersold, On the dependency of cellular protein levels on mRNA abundance, Cell 165 (3) (2016) 535–550.
[25] R. Aebersold, M. Mann, Mass-spectrometric exploration of proteome structure and function, Nature 537 (7620) (2016) 347–355.
[26] A. Furey, M. Moriarty, V. Bane, B. Kinsella, M. Lehane, Ion suppression; a critical review on causes, evaluation, prevention and applications, Talanta 115 (2013) 104–122.
[27] Y.F. Li, R.J. Arnold, H. Tang, P. Radivojac, The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics, J. Proteome Res. 9 (12) (2010) 6288–6297.
[28] T. Kislinger, B. Cox, A. Kannan, C. Chung, P. Hu, A. Ignatchenko, M.S. Scott, A.O. Gramolini, Q. Morris, M.T. Hallett, et al., Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling, Cell 125 (1) (2006) 173–186.