Published in final edited form as:

J Proteome Res. 2017 February 03; 16(2): 842–851. doi:10.1021/acs.jproteome.6b00824.

Single Amino Acid Variant Profiles of Subpopulations in the MCF-7 Breast Cancer Cell Line

Zhijing Tan[†], Song Nie^{†,‡}, Sean P. McDermott^{§,||}, Max S. Wicha^{§,||}, and David M. Lubman^{*,†}

[†]Department of Surgery, University of Michigan, Ann Arbor, Michigan 48109, United States

[‡]Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99352, United States

[§]Department of Internal Medicine, Division of Hematology/Oncology, University of Michigan, Ann Arbor, Michigan 48109, United States

^{||}Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan 48109, United States

Abstract

Cancers are initiated and developed from a small population of stem-like cells termed cancer stem cells (CSCs). There is heterogeneity among this CSC population that leads to multiple subpopulations with their own distinct biological features and protein expression. The protein expression and function may be impacted by amino acid variants that can occur largely due to single nucleotide changes. We have thus performed proteomic analysis of breast CSC subpopulations by mass spectrometry to study the presence of single amino acid variants (SAAVs) and their relation to breast cancer. We have used CSC markers to isolate pure breast CSC subpopulation fractions (ALDH+ and CD44+/CD24– cell populations) and the mature luminal cells (CD49f–EpCAM+) from the MCF-7 breast cancer cell line. By searching the Swiss-CanSAAVs database, 374 unique SAAVs were identified in total, where 27 are cancer-related SAAVs. 135 unique SAAVs were found in the CSC population compared with the mature luminal cells. The distribution of SAAVs detected in MCF-7 cells was compared with those predicted from the Swiss-CanSAAVs database, where we found distinct differences in the numbers of SAAVs detected relative to that expected from the Swiss-CanSAAVs database for several of the amino acids.

Graphical Abstract

*Corresponding Author: dmlubman@umich.edu. Phone: 734-647-8834. Fax: 734-615-2088.

ORCID

David M. Lubman: 0000-0001-7731-0232

Notes

The authors declare no competing financial interest.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b00824.

Supplemental Table 1. Three subpopulations of SAAVs proteins and unique SAAVs. (XLSX)

Supplemental Table 2. Number of SAAVs in different proteins. Supplemental Table 3. 27 Cancer-related variant sites. Supplemental

Table 4. The predicted frequencies of DNA codons in common amino acids for *Homo sapiens*. (PDF)

Keywords

single amino acid variant; breast cancer; MCF-7 cell line; cell line subpopulation; proteomics

INTRODUCTION

A systematic discovery of mutations in protein amino acid sequence is a crucial step to improve our understanding of tumorigenesis and progression. Changes in the amino acid sequence of encoded proteins eventually affect protein stability, enzyme activity, and protein–protein interactions, which potentially cause different diseases.^{1–3} Detection of single-nucleotide polymorphisms (SNPs), especially nonsynonymous (nsSNPs) at the genome transcript level, will be important to understand the relationship between the genome and cancer development.^{4,5} These mutations at the genome level, though, may not be translated at the protein level so that detection of single amino acid variants (SAAVs) directly at the proteome level is vital to reveal the expression of such variants. The identification and characterization of SAAVs in human cancer are important for understanding tumorigenesis and progression and eventually may provide opportunities for clinical biomarkers for cancer early diagnosis and therapeutics.^{6,7}

Recent efforts have been devoted to database construction and SAAVs identification based on mass spectrometry detection. In one study, a mass-spectrometry-friendly database was developed for coding single nucleotide polymorphism (cSNP) identification, and 15 variant peptides were identified in lysosomes isolated from the MCF-7 breast cancer cell line.⁸ This was followed by other work that created a novel method to identify 629 nsSNPs, including 36 alternate SNP alleles not found in the NCBI or IPI protein databases from DU4475 breast tumor cells.⁹ Zeng and coworkers developed a novel method to discover SAAVs using a mass-spectrometry-based de novo sequencing algorithm where 248 protein variants were detected in paired normal prefrontal cortex and cerebellum tissue from three healthy human brains.¹⁰

More recent efforts to study SAAVs in the human proteome have included the creation of sophisticated human cancer proteome variants databases. For example, Li et al. creatively combined six sources including public databases HPI, COSMIC, OMIM, and TCGA and data from two groups' studies on large-scale cancer genome resequencing. The CanProVar database comprises 8570 cancer-related variants in 2921 proteins derived from six different sources and 41 541 noncancer specific variants in 30 322 proteins derived from the dbSNP database.¹¹ Subsequently, this group developed an integrated bioinformatics workflow to

detect variant peptides, and 204 variant peptides, including 5 peptides known for cancer-related mutations, were identified from three colorectal tumor specimens.¹² A more comprehensive variant-associated database was constructed by Song et al., where the humsavar database and CanProVar database were integrated into the UniProtKB/Swiss-Prot canonical protein database and 282 unique SAAVs sites were quantified in human liver tissues.¹³ In other work, 128 SAAVs paired with related canonical peptides using a customized database from neXtprot were identified.¹⁴

Other work by Yang et al. developed a cancer-related protein database with mutation information collected from several public sources including COSMIC, IARC, P53, OMIM, and UniProtKB. The newly developed database included over a million unique cancer-related mutations.⁷ A mass spec analysis of mutations in the MCF-7 breast cancer cell lines showed the presence of numerous mutated sites.¹⁵ Also, the group of Segura and coworkers developed a SAP database using data from the TCGA project and also related cancer-specific databases. They searched these databases using the Comet and Mascot search engines for MS data from several cell lines including MCF-7, where they found several hundred cancer-related related mutations.¹⁶

Other recent work has involved proteogenomic strategies for protein identification using custom databases using RNA-seq data and other genomic data. This strategy has been reviewed recently by Nesvizhskii and others.^{17–20} Several groups have generated such custom databases, where, for example, Krug and coworkers used a transcriptomic data set for HeLa cells with high-resolution MS/MS spectra and identified 451 variant peptides containing 401 nsSNVs.²¹ Sheynkman et al. also developed a customized SAP database from sample-matched RNA-Seq data and were able to detect 695 SAP peptides mapping to 504 nsSNV sites from their mass spec data.²² However, this strategy depends on an RNA-seq database for a specific sample and for a highly heterogeneous tumor or a cell line that has been passaged an unknown number of times, this database may not be transferable in terms of the expressed mutations.

Although studies have identified numerous SAAVs from cancer cell lines,^{9,23} the presence of SAAVs in different subpopulations and the variant frequency for each amino acid in these cell lines is not known. One area where identifying unique SAAVs in subpopulations will be important is in the cancer stem cell (CSC) population. It has been shown that many cancers are initiated from a small subset of stem-like cells termed CSCs.^{24,25} These CSCs have the capacity to differentiate into multiple subpopulations with their own distinct protein expression,²⁶ and it might also be expected that they have their own unique sets of expressed SAAVs. Breast cancer stem cells (BCSCs) exist in distinct states; a mesenchymal-like (epithelial-mesenchymal transition) state, which expresses CD44+/CD24–, is primarily quiescent and localized at the tumor-invasive front. The other state is epithelial-like (mesenchymal–epithelial transition), which expresses intra-cellular aldehyde dehydrogenase (ALDH), is proliferative, and is typically located within the tumor.²⁷ Mature luminal cells are identified by cell-surface expression CD49f–EpCAM+.²⁸ There is extremely low overlap between the CD44+/CD24– and ALDH+ fractions in the MCF-7 breast cancer cell line, as reported using flow cytometry analysis.²⁷

In this work, we have detected SAAVs in three subpopulations from the MCF-7 breast cancer cell line using mass spectrometry followed by Swiss-CanSAAVs database searching. These subpopulations include the CD44+/CD24– and ALDH+ CSC populations and the CD49f–EpCAM+ mature luminal population. 374 unique SAAVs were identified in the MCF-7 cell line, in which 27 are cancer-related SAAVs. 135 unique SAAVs were found in the CSC populations compared with the mature luminal cells. The pattern and characterization of SAAVs in the Swiss-CanSAAVs database and MCF-7 cell line were analyzed and compared, in which it is shown that there are distinct differences in the expected and observed pattern of SAAVs. This is the first demonstration of unique SAAVs observed in three different subpopulations from the same cell line and importantly where a CSC population has its own distinct SAAVs pattern.

MATERIALS AND METHODS

Materials, Cell Sorting, Protein Extraction, and Tryptic Digestion Using Filter-Aided Sample Preparation (FASP)

The MCF-7 breast cancer cell line used in this work was purchased from ATCC (Manassas, VA). A description of the reagents used herein is the same as described in our previous work.²⁹ The procedures for cell sorting, protein extraction, and tryptic digestion using FASP for these samples also have been described in our previous work.²⁹

Nano LC–MS/MS Analysis

The peptide mixtures were analyzed by nano LC–MS/MS using an Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific, CA) in positive ion mode coupled to an EASY-nLC 1000 liquid chromatography system (Thermo Fisher Scientific, CA). A total sample of 50 000 cells was run eight times for two separate biological replicates to increase the protein coverage for each of the MCF-7 subpopulation samples. The peptides were separated on an Acclaim PepMap RSLC C18 column (25 cm column with 75 μ m inner diameter, packed 2 μ m particles, 100 Å, nanoViper) with a gradient of acetonitrile/H₂O containing 0.2% formic acid at a flow rate of 300 nL/min. The details of the gradient are as in our previous work.³⁰ The mass spectra data acquisition was performed using the Xcalibur software (Thermo Fisher Scientific, CA) in data-dependent mode. The ESI spray voltage was set at 2.5 kV. A full mass scan (m/z 400–1600) was performed, and the top 15 most intense ions in the full scan were chosen for MS/MS analysis. The normalized collision energy was set at 30% for high-energy collision dissociation (HCD) fragmentation. Dynamic exclusion settings were set as follows: repeat count 1, repeat duration 30 s, and exclusion duration 30 s.

Data Analysis

All raw data were searched against the Swiss-CanSAAVs database using SEQUEST HT in Proteome Discoverer 1.4 (Thermo Fisher Scientific, CA). The free Swiss-CanSAAVs database, which includes 87 733 amino acid variant sequences from 73 910 UniProtKB/Swiss-Prot canonical proteins, was downloaded from <http://bioanalysis.dicp.ac.cn/proteomics/Publications/SAAV/SAAV-Database.htm> on 01/21/2016. The searching parameters were set as follows: (1) cysteine carbamidomethylation (+57.021 Da) was set as

a fixed modification; (2) methionine oxidation (+15.995 Da) was set as dynamic modification; (3) the tolerance of precursor ions was set as 10 ppm and fragment ions were set as 0.5 Da, where the fragment ions were detected in the linear ion trap to increase the acquisition frequency of MS²; (4) a maximum of two missed cleavages were allowed; and (5) the shortest peptide length was set at six amino acids and the identified results were filtered using high confidence at 1% peptide-level false discovery rate (FDR). The identified peptide data were exported into the XML-formatted file from Proteome Discoverer software 1.4 and all peptides were combined together. The potential peptides with SAAVs sites were manually checked to remove false SAAVs that result from missed cleavage sites surrounding the centric variation site, where the real SAAV sites will be located in the resultant peptides. If multiple peptides share the same partial peptide sequences due to tryptic missed cleavages, only the shorter peptide sequence with the SAAV site was retained for the following SAAVs analysis. The protein function was annotated using Gene Ontology (GO) with QIAGEN'S Ingenuity Pathway Analysis (IPA) online software (<http://www.ingenuity.com/products/ipa>).

RESULTS

Proteomic Analysis of MCF-7 Cells by LC-MS/MS

To investigate the SAAVs in three breast cancer CSCs, three different subpopulations of ALDH+ CSCs, CD44+/CD24-CSCs, and CD49f-EpCAM+ mature luminal cells were isolated via fluorescence-activated cell sorting (FACS) from an MCF-7 cell line. For each subpopulation, 50 000 cells were applied for protein extraction, and resultant tryptic digestion peptides were analyzed by LC-MS/MS on an Orbitrap Fusion mass spectrometer. A database search against the human UniProt database yielded ~3000 proteins from one typical RAW file. Compared with the results previously published,²⁹ more proteins were identified due to the use of the Orbitrap Fusion mass spectrometer and additional technical replicates. As many as 335 proteins with SAAVs were detected from all samples, of which 146, 158, and 238 proteins with SAAVs were detected from ALDH+, CD44+/CD24-, and CD49f-EpCAM+ subpopulations, respectively (Supplemental Table 1). A total of 3771, 3772, and 4943 proteins were detected from each of the ALDH+, CD44/CD24-, and CD49f-EpCAM+ subpopulations, respectively, which means that ~4% of proteins were identified with SAAVs in three subpopulations.

Large-Scale Identification of SAAVs

We analyzed our mass spectral data against the Swiss-CanSAAVs database. To ensure that we are reporting unique SAAV peptides, we manually checked each of the peptides to remove the peptides where the SAAV site is not located within the tryptic peptides. In the Swiss-CanSAAVs database, most peptides with SAAVs contain several K/R sites where up to eight may be present. As many as 374 SAAV sites in peptides from 335 proteins were identified from the MCF-7 cell line in the current analysis. Most proteins only possess a single variant site. 14, 7, and 2 proteins were found to possess two, three, and four SAAV sites, respectively. Only one protein actin, alpha 1, skeletal muscle had as many as six variant sites (data shown in Supplemental Table 2).

From the 87 733 known SAAVs in 15 609 proteins, there were 335 proteins identified with SAAVs in the MCF-7 samples, in which only 2% of the total proteins contain SAAVs (Table 1). For the 15 609 proteins in the Swiss-CanSAAVs database the average number of SAAVs for each protein is 5.6 compared with 1.1 average number of SAAVs for the MCF-7 cell line.

We also studied the three MCF-7 subpopulations where 238, 158, and 146 SAAVs were identified from CD49f–EpCAM+, CD44+/CD24–, and ALDH+, respectively (Supplemental Table 1). 56 of the 374 SAAVs were detected from all three subpopulations. The details of overlap between the three subpopulations are shown in Figure 1. The CD49f–EpCAM+ cells contained 143 SAAVs that were unique and not detected in the other subpopulations, including 2 breast cancer-related SAAVs and 12 other cancer-related SAAVs (Supplemental Table 1). ALDH+ contained 59 SAAVs that were distinct from the other populations, including 5 other cancer-related SAAVs, while CD44+/CD24– contained 59 specific SAAVs including 1 breast-cancer-related SAAVs and 4 other cancer-related SAAVs (Supplemental Table 1). It should be noted that in searching against the database the number of missed digestion sites was set as 2. Thus some redundant peptides were unavoidable. All redundant peptides were removed by a manual check of the results.

The MS/MS spectra for the variant peptide VFIEDVIR and canonical peptide VFIEDVSR from argininosuccinate synthase (ASS1) protein in MCF-7 cell line are shown in Figure 2. Both peptides share the same y1 ions, but a series of y2, y3, y4 y5, y6, and y7 ions indicating a + 26 Da mass shift corresponding to the substitution of Ser for Ile at position 65 was observed (Figure 2). These matched results of their MS/MS spectra confirm the variant site. It should be noted that most of the variant peptides and related canonical peptides are difficult to detect together due to the complexity of samples and low protein sequence coverage by current MS-based proteomics.⁶

Different Frequency of the Amino Acid Variants

We studied the frequency of variants for the 20 common amino acids in the Swiss-CanSAAVs database. For the identifiable amino acid variants, we found that 6959 of the A amino acid varied, while only 982 of the W amino acid varied among the 87 733 sites in the Swiss-CanSAAVs database. As shown in Table 2, each amino acid occupies a different ratio if we consider the whole frequency as 100%. The ratio of arginine (R) is 13.4%, which has the highest variant, while tryptophan (W) has a ratio of 1.1% compared with the whole database. The ratio of variant frequency for the other 18 amino acids varied from 2.2 to 7.9%. Also, for each of the amino acids we calculated the frequency change for the SAAVs in the MCF-7 cell line and found that the pattern of amino acid variant is different from the pattern in the Swiss-CanSAAVs database. The frequency of identifiable amino acid variants in the Swiss-CanSAAVs database is similar, as in previous investigations for disease-related amino acids where R and G are of higher frequency while W and Y are of lower frequency.³¹

The predicted number of SAAVs based on the SAAVs database is not consistent with that of the detected SAAVs in the MCF-7 cell line by mass spectrometry. The C, K M, N, O, S, and V amino acid residues were found to increase, while the amino acids F, H, P, and R decreased, whereas the remaining amino acid residues remained basically unchanged. We

also found that for most of frequent identifiable variants in the Swiss-CanSAAVs database and in the MCF-7 cell line that the top 32 and 23 variants were responsible for 50% of the SAAVs, respectively. In the Swiss-CanSAAVs database, the remaining 50% consists of 247 variants with 63, 27, and 17 of these being observed only one, two, and three times, respectively. However, in the MCF-7 cell line, the remaining 50% consists of 105 variants with 27, 26, and 16 of these being observed only one, two, and three times, respectively. There are a total 360 theoretical amino acid variant patterns; however, 102 and 276 amino acid variant patterns such as AW and AH do not exist in the Swiss-CanSAAVs database and MCF-7 cell line, respectively.

It has been found that the codon frequency usage in all coding sequences shows bias for gene expression, and likewise the base pair change also shows bias resulting in different frequencies between two amino acids such as RQ and QR in the database (Table 3). In most cases, the frequency of an amino acid variant relative to the frequency of the converse pattern is different. Some patterns show large differences such as RQ versus QR in the database and RH versus HR in the MCF-7 cell line (Figure 3A,B).

Gene Ontology Analysis

Gene Ontology (GO) analysis was performed using the Ingenuity Pathway Analysis (IPA) bioinformatics tool and database. For proteins with SAAVs from the MCF-7 cell line, most proteins were found to be located in the cytoplasm, nucleus, and plasma membrane (Figure 4A). The dominant types of proteins included cytokines and other proteins with no assignments (Figure 4B). Among the protein networks, we found that the proteins TP53, TRRAP, PTEN, and ERBB2 showed cancer-related variants (Figure 4C). BRCA2 has been reported as a marker that predisposes an individual for a high lifetime risk of breast cancer.³² A previous study has also shown that TP53 and PTEN mutations are responsible for an increased risk for the development of breast and other cancers.³³

Cancer-Related Proteins with SAAVs Sites

Cancer-related proteins with SAAVs may serve as candidate protein biomarkers or therapeutic targets. 95 unique SAAVs from 85 proteins related to cancer were identified based on manual searching against the CanProVar database (<http://bioinfo.vanderbilt.edu/canprovar/>). It should be noted that there is no variant information available for six proteins in the database search. Among 95 unique SAAVs, only 27 unique SAAVs in 19 proteins have cancer-related variants (Supplemental Table 3). There were three breast-cancer-related proteins identified from the CanProVar database, including Receptor tyrosine-protein kinase erbB-2 (ErbB-2), Cofilin-2 (CFL2), and 60S ribosome subunit biogenesis protein NIP7 homologue (NIP7).^{34–37}

DISCUSSION

In this study, we used mass spectrometry to study the MCF-7 cell line to search for cancer-related SAAVs. As many as 374 unique SAAVs sites from 335 proteins were identified in the MCF-7 cell line, where most proteins only have a single variant site and only 2% of proteins have SAAVs in the MCF-7 cell line. It should be noted that most peptides with

SAAVs in the Swiss-CanSAAVs database contain several trypsin missed K/R sites that result in false SAAVs peptides from the MCF-7 cell line. Despite the low occurrence in the Swiss-CanSAAVs database and in the MCF-7 cell line, these SAAVs may be important markers of cancer and factors in human disease.

Only 27 unique SAAV pairs are cancer-related variants, while the remaining 347 SAAVs were excluded for further consideration in terms of being cancer-related SAAVs. Some of these proteins are well-known cancer-related proteins, but the detected SAAV site is not the cancer-related site. For example, the N104S variant in *Ki-67* was detected in the MCF-7 cell line; however, the site is not related to any cancer based on the CanProVar database. Similarly, the T1388N variant was detected in the well-known breast cancer biomarker, that is, breast cancer type 2 susceptibility protein (BRCA2), but the specific site detected is not related to cancer.

For the three different subpopulations of the MCF-7 cell line, more SAAVs sites were detected in CD49f–EpCAM+ compared with CD44+/CD24– and ALDH+ subpopulations. Each subpopulation has a certain number of unique SAAVs, which shows distinct differences between these subpopulations. There are 59, 59, and 143 unique SAAVs in the CD44+/CD24–, ALDH+, and CD49f–EpCAM+ subpopulations, respectively. Compared with the mature luminal cells, 135 SAAVs were unique to the breast CSC, where there are 73 overlapping peptides between the CD44+/CD24– and ALDH+ cell subpopulations. All three different subpopulations have unique SAAVs, which illustrates the specificity of each subpopulation, which may result in heterogeneous breast cancer originating from different cells of origin.³⁸ Liu et al. identified a small population of breast cells that expressed both CD44+/CD24– and ALDH+ simultaneously using immunofluorescence, which suggests that some of the 73 overlapping SAAVs may derive from the same population or the CSC cells show the plasticity that allows the interchange between the ALDH+ and CD44+/CD24– cell subpopulation in two different states.²⁷

In more recent work it has been shown by sequencing studies that in BCSCs populations there is a dynamic conversion between differentiation states.³⁹ In the simplest CSC model, tumor cellular heterogeneity arises in part through differentiation of CSCs to generate more differentiated tumor cells. The differences between CSCs and their more differentiated progeny therefore are epigenetic rather than genetic, where recent work in BCSCs showed that in the genomic analysis of these cells that the majority of somatic mutations are shared between the CSCs and the bulk primary tumor.³⁹ If CSCs and their progeny share the same genetics, that is, mutation profile, then this DNA profile should be shared in CSCs and their progeny. Therefore, SAAV differences between these sub-populations must be generated at other levels, that is, transcriptional, RNA processing, translational, and so on, where these processes may themselves be epigenetically regulated. At this point the mechanism for this process is not known.

For the identification of protein variants in human cancer based on proteomics, several groups have performed extensive work on the exhaustive search of potential amino acid variants. For example, Li et al. developed a human Cancer Proteome Variation database based on searches that comprehensively integrates HPI, COSMIC, OMIM, and large-scale

mutational profiling studies on cancer genes and cancer genomes and a related bioinformatics workflow for SAAVs detection in shotgun proteomics.^{11,12} Although searching a large database such as the COSMIC database may result in additional SAAVs, it may also result in additional false-positives. In the current work we are devising custom databases based on the breast cancer cell lines in COSMIC for further searching. Also, the Swiss-CanSAAVs database comprehensively integrated a conventional UniprotKB/Swiss-Prot database (human) including 73 910 protein entries released on 11/25/2011 and CanProVar database (EnsemblV53, 97, 637 variant sequences) by Zou's group, which provides an opportunity for database searching for variant amino acid site analysis in human cancer samples.¹³ SAAVs detection was greatly improved in shotgun proteomics analysis of human cancer samples by removing redundant sequences and increasing the identification sensitivity in Swiss-CanSAAVs database searching. Our previous study applied this advanced database to identify and quantify variant peptides in pancreatic cancer patient's serum and detected 83 SAAV sites.⁴⁰

There are three possible sources of SAAVs including genomic nsSNPs, RNA-editing-dependent SAAVs, and dysfunction of protein editing, and they may be distinguished by proteogenomics analysis.^{23,41,42} However, SAAVs mostly derive from genomic nsSNPs that occur in coding regions, although they may be epigenetically controlled at some level.⁴³ The codon frequency usage in all coding sequences shows bias for gene expression based on the predicted Web site (<http://www.kazusa.or.jp/codon/>). It is well known that most amino acids are encoded by a set of synonymous trinucleotide units, that is, codons with a different third wobble position. Leucine (Leu) and arginine (Arg) are encoded by six codons with the first codon wobble position. Serine (Ser) possesses a complex codon pattern with all three wobble codon positions, while tryptophan (Trp) and methionine (Met) are each encoded by a single codon.⁴⁴ Codon usage shows extreme preferences, as shown in Supplemental Table 4. In most cases, if two amino acids share two of the same bases in codons, then they have a potential high frequency for variant.

In terms of the frequency of identifiable amino acids, each SAAV shows a different pattern in the SAAVs database with a resulting bias variant, where a few variants are high frequency. The amino acid variability is different in each case, including for the same pattern such as AT, TA, where the frequency of AT is two times that of TA. The total predicted frequency of A and T in *Homo sapiens* is 6.93 and 5.32%, respectively (Supplemental Table 4). Both are based on four codons, but with the first base different in all four codons. Theoretically, the number of AT and TA should be close and not a two-fold difference, which indicates that the amino acid variability is different and that bias selection exists. Additionally, the origin of various SAAVs, which include DNA mutations or RNA variants, as well as translational or other unknown post-transcriptional processes may also contribute to the amino acid variability bias.⁴⁵ RQ, AT, and PL are the top three variant patterns in the Swiss-CanSAAVs database. The SAAVs in the MCF-7 cell line show a different variant pattern compared with the Swiss-CanSAAVs database, where VI, LP, and AV are the top three patterns with the highest variant frequency in the MCF-7 cell line (Figure 3A,B). The LP variant is one of most harmful SAAVs based on the theoretic investigation of Peng et al. on disease-causing amino acid variants and their effect on protein stability and binding.⁴⁶

Among the 19 cancer-related SAAVs proteins detected in this work, several proteins have been investigated for their SAAV content. For example, the tumor suppressor p53 (TP53) is a key protein in cell defense against cancer. The key residue Leu348 is one of two key hydrophobic residues (another one is Leu-344) for maintaining the stabilization of tetramer interface due to the oligomerization shift toward the formation of stable dimers.⁴⁷ In the MCF-7 cell line, the variant of L348F as well as another two residues, D228H and G154D, was observed in the present study. However, some typical proteins with SAAV sites were excluded in the list such as BRCA2, where there are 93 potential SAAVs, but only one SAAV was detected in this study, which is not related to breast cancer. To better understand the presence of SAAVs, additional investigations on SAAVs in different breast cancer cell lines and as a function of passages will be needed.

In the current work, we have identified three breast-cancer-related proteins with SAAVs. Several of these have been shown to be related to breast cancer progression or proliferation. ErbB-2 works with ErbB-3 as an oncogenic unit to drive breast tumor cell proliferation.^{36,37} Cofilin-2 expression in primary breast cancer patients correlated with tumor grades, where Cofilin-2 expression increases dramatically, which is consistent with the grade of tumor.³⁴ It has been reported that *NIP7* gene expression at the RNA level decreases among bone metastasis in breast carcinomas.³⁵ However, there are no reports on 60S ribosome subunit biogenesis protein NIP7 homologue in breast cancer at the protein level.

CONCLUSIONS

We have detected SAAVs in three subpopulations from the MCF-7 cell line and uncovered 374 SAAVs sites including 3 breast-cancer-related SAAVs and 24 other cancer-related SAAVs. The SAAVs in three subpopulations overlap each other in some content. However, the CD44+/CD24-, ALDH +, and CD49f-EpCAM+ subpopulations each have 59, 59, and 143 unique SAAVs, respectively. We also compared identifiable SAAVs from the Swiss-CanSAAVs database to SAAVs from the MCF-7 cell line and found a different amino acid variant trend preference. The frequency with which each amino acid varies among the 19 amino acids shows a different pattern in the Swiss-CanSAAVs and MCF-7 cell line. The profile of SAAVs in the Swiss-CanSAAVs database and MCF-7 cell line was studied, which will be useful for understanding the status of SAAVs and its possible influence on human cancer. Further validation of cancer-related SAAVs in patients needs to be investigated in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Ramdane Harouaka in the Department of Internal Medicine, Dr. Jianhui Zhu and Dr. Mingrui An in the Department of Surgery at the University of Michigan, and Dr. Haidi Yin in the Department of Applied Biology and Chemical Technology at the Hong Kong Polytechnic University for a critical reading of the manuscript. This work was supported by the National Institutes of Health under grant number R01 GM49500. The MS data were generated by the Wayne State University Proteomics Core. The Proteomics Core is supported through the NIH Center Grant P30 ES 020957, P30 CA 022453 and the NIH Shared Instrumentation Grant S10 OD 010700.

References

1. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet.* 2006; 7:61–80. [PubMed: 16824020]
2. Yates CM, Sternberg MJE. The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein protein interactions. *J Mol Biol.* 2013; 425(21):3949–3963. [PubMed: 23867278]
3. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks N, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, Menzies A, Mould C, Parker A, Stevens C, Watt S, Hooper S, Wilson R, Jayatilake H, Gusterson BA, Cooper C, Shipley J, Hargrave D, Pritchard-Jones K, Maitland N, Chenevix-Trench G, Riggins GJ, Bigner DD, Palmieri G, Cossu A, Flanagan A, Nicholson A, Ho JWC, Leung SY, Yuen ST, Weber BL, Seigler HF, Darrow TL, Paterson H, Marais R, Marshall CJ, Wooster R, Stratton MR, Futreal PA. Mutations of the *BRAF* gene in human cancer. *Nature.* 2002; 417(6892):949–954. [PubMed: 12068308]
4. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, Defazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR. Patterns of somatic mutation in human cancer genomes. *Nature.* 2007; 446(7132):153–158. [PubMed: 17344846]
5. Cole C, Krampis K, Karagiannis K, Almeida JS, Faison WJ, Motwani M, Wan Q, Golikov A, Pan Y, Simonyan V, Mazumder R. Non-synonymous variations in cancer and their effects on the human proteome: workflow for NGS data biocuration and proteome-wide analysis of TCGA data. *BMC Bioinf.* 2014; 15(1):28.
6. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, Davies SR, Wang S, Wang P, Kinsinger CR, Rivers RC, Rodriguez H, Townsend RR, Ellis MJC, Carr SA, Tabb DL, Coffey RJ, Slebos RJC, Liebler DC, Cptac N, et al. Proteogenomic characterization of human colon and rectal cancer. *Nature.* 2014; 513(7518):382–387. [PubMed: 25043054]
7. Yang X, Lazar IM. XMAN: A *Homo sapiens* mutated-peptide database for the MS analysis of cancerous cell states. *J Proteome Res.* 2014; 13(12):5486–95. [PubMed: 25211293]
8. Schandorff S, Olsen JV, Bunkenborg J, Blagoev B, Zhang Y, Andersen JS, Mann M. A mass spectrometry-friendly database for cSNP identification. *Nat Methods.* 2007; 4(6):465–466. [PubMed: 17538625]
9. Bunger MK, Cargile BJ, Sevinsky JR, Deyanova E, Yates NA, Hendrickson RC, Stephenson JL. Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. *J Proteome Res.* 2007; 6(6):2331–2340. [PubMed: 17488105]
10. Su ZD, Sheng QH, Li QR, Chi H, Jiang X, Yan Z, Fu N, He SM, Khaitovich P, Wu JR, Zeng R. *De novo* identification and quantification of single amino-acid variants in human brain. *J Mol Cell Biol.* 2014; 6(5):421–433. [PubMed: 25007923]
11. Li J, Duncan DT, Zhang B. CanProVar: a human cancer proteome variation database. *Hum Mutat.* 2010; 31(3):219–228. [PubMed: 20052754]
12. Li J, Su ZL, Ma ZQ, Slebos RJC, Halvey P, Tabb DL, Liebler DC, Pao W, Zhang B. A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Mol Cell Proteomics.* 2011; 10(5):M110.006536.
13. Song C, Wang F, Cheng K, Wei X, Bian Y, Wang K, Tan Y, Wang H, Ye M, Zou H. Large-scale quantification of single amino-acid variations by a variation-associated database search strategy. *J Proteome Res.* 2014; 13(1):241–248. [PubMed: 24237036]
14. Hwang H, Park GW, Kim KH, Lee JY, Lee HK, Ji ES, Park SKR, Xu T, Yates JR, Kwon KH, Park YM, Lee HJ, Paik YK, Kim JY, Yoo JS. Chromosome-based proteomic study for identifying novel

protein variants from human hippocampal tissue using customized neXtProt and GENCODE databases. *J Proteome Res.* 2015; 14(12):5028–5037. [PubMed: 26549206]

15. Garin-Muga A, Corrales FJ, Segura V. Proteogenomic analysis of single amino acid polymorphisms in cancer research. *Adv Exp Med Biol.* 2016; 926:93–113. [PubMed: 27686808]
16. Vegvari A. Mutant Proteogenomics. *Adv Exp Med Biol.* 2016; 926:77–91. [PubMed: 27686807]
17. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat Methods.* 2014; 11(11):1114–25. [PubMed: 25357241]
18. Sheynkman GM, Shortreed MR, Cesnik AJ, Smith LM. Proteogenomics: Integrating next-generation sequencing and mass spectrometry to characterize human proteomic variation. *Annu Rev Anal Chem.* 2016; 9(1):521–45.
19. Kumar D, Bansal G, Narang A, Basak T, Abbas T, Dash D. Integrating transcriptome and proteome profiling: Strategies and applications. *Proteomics.* 2016; 16(19):2533–2544. [PubMed: 27343053]
20. Alfaro JA, Sinha A, Kislinger T, Boutros PC. Oncoproteogenomics: cancer proteomics joins forces with genomics. *Nat Methods.* 2014; 11(11):1107–13. [PubMed: 25357240]
21. Krug K, Popic S, Carpy A, Taumer C, Macek B. Construction and assessment of individualized proteogenomic databases for large-scale analysis of nonsynonymous single nucleotide variants. *Proteomics.* 2014; 14(23–24):2699–2708. [PubMed: 25251379]
22. Sheynkman GM, Shortreed MR, Frey BL, Scalf M, Smith LM. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J Proteome Res.* 2014; 13(1):228–240. [PubMed: 24175627]
23. Lichti CF, Mostovenko E, Wadsworth PA, Lynch GC, Pettitt BM, Sulman EP, Wang Q, Lang FF, Rezeli M, Marko-Varga G, Vegvari A, Nilsson CL. Systematic identification of single amino acid variants in glioma stem-cell-derived chromosome 19 proteins. *J Proteome Res.* 2015; 14(2):778–786. [PubMed: 25399873]
24. Clarke MF, Dick JE, Dirks PB, Eaves CJ, Jamieson CH, Jones DL, Visvader J, Weissman IL, Wahl GM. Cancer stem cells—perspectives on current status and future directions: AACR Workshop on cancer stem cells. *Cancer Res.* 2006; 66(19):9339–9344. [PubMed: 16990346]
25. Reya T, Morrison SJ, Clarke MF, Weissman IL. Stem cells, cancer, and cancer stem cells. *Nature.* 2001; 414(6859):105–111. [PubMed: 11689955]
26. Pattabiraman DR, Weinberg RA. Tackling the cancer stem cells - what challenges do they pose? *Nat Rev Drug Discovery.* 2014; 13(7):497–512. [PubMed: 24981363]
27. Liu S, Cong Y, Wang D, Sun Y, Deng L, Liu Y, Martin-Trevino R, Shang L, McDermott SP, Landis MD, Hong S, Adams A, D'Angelo R, Ginestier C, Charafe-Jauffret E, Clouthier SG, Birnbaum D, Wong ST, Zhan M, Chang JC, Wicha MS. Breast cancer stem cells transition between epithelial and mesenchymal states reflective of their normal counterparts. *Stem Cell Rep.* 2014; 2(1):78–91.
28. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat ML, Gyorki DE, Ward T, Partanen A, Felleppa F, Huschtscha LI, Thorne HJ, Fox SB, Yan M, French JD, Brown MA, Smyth GK, Visvader JE, Lindeman GJ. Aberrant luminal progenitors as the candidate target population for basal tumor development in *BRCA1* mutation carriers. *Nat Med.* 2009; 15(8):907–913. [PubMed: 19648928]
29. Nie S, McDermott SP, Deol Y, Tan Z, Wicha MS, Lubman DM. A quantitative proteomics analysis of MCF7 breast cancer stem and progenitor cell populations. *Proteomics.* 2015; 15(22):3772–83. [PubMed: 26332018]
30. Yin H, Tan Z, Wu J, Zhu J, Shedden KA, Marrero J, Lubman DM. Mass-selected site-specific core-fucosylation of serum proteins in hepatocellular carcinoma. *J Proteome Res.* 2015; 14(11):4876–84. [PubMed: 26403951]
31. Casadio R, Vassura M, Tiwari S, Fariselli P, Martelli PL. Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome. *Hum Mutat.* 2011; 32(10):1161–1170. [PubMed: 21853506]
32. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, Bishop DT, Weber B, Lenoir G, Chang-Claude J, et al. Genetic heterogeneity and penetrance analysis of the *BRCA1* and *BRCA2* genes in breast cancer families. *Am J Hum Genet.* 1998; 62(3):676–689. [PubMed: 9497246]
33. Walsh T, Casadei S, Coats KH, Swisher E, Stray SM, Higgins J, Roach KC, Mandell J, Lee MK, Ciernikova S, Foretova L, Soucek P, King MC. Spectrum of mutations in *BRCA1*, *BRCA2*,

- CHEK2*, and *TP53* in families at high risk of breast cancer. *Jama-J Am Med Assoc.* 2006; 295(12): 1379–1388.
34. Luo D, Wilson JM, Harvel N, Liu J, Pei L, Huang S, Hawthorn L, Shi H. A systematic evaluation of miRNA:mRNA interactions involved in the migration and invasion of breast cancer cells. *J Transl Med.* 2013; 11(1):57. [PubMed: 23497265]
 35. Savci-Heijink CD, Halfwerk H, Koster J, van de Vijver MJ. A novel gene expression signature for bone metastasis in breast carcinomas. *Breast Cancer Res Treat.* 2016; 156(2):249–259. [PubMed: 26965286]
 36. Holbro T, Beerli RR, Maurer F, Koziczak M, Barbas CF, Hynes NE. The ErbB2/ErbB3 heterodimer functions as an oncogenic unit: ErbB2 requires ErbB3 to drive breast tumor cell proliferation. *Proc Natl Acad Sci U S A.* 2003; 100(15):8933–8938. [PubMed: 12853564]
 37. The Breast Cancer Linkage Consortium. Cancer risks in BRCA2 mutation carriers. *JNCI-J Natl Cancer I.* 1999; 91(15):1310–1316.
 38. Skibinski A, Kuperwasser C. The origin of breast tumor heterogeneity. *Oncogene.* 2015; 34(42): 5309–16. [PubMed: 25703331]
 39. Klevebring D, Rosin G, Ma R, Lindberg J, Czene K, Kere J, Fredriksson I, Bergh J, Hartman J. Sequencing of breast cancer stem cell populations indicates a dynamic conversion between differentiation states in vivo. *Breast Cancer Res.* 2014; 16(4):R72. [PubMed: 24998755]
 40. Nie S, Yin H, Tan Z, Anderson MA, Ruffin MT, Simeone DM, Lubman DM. Quantitative analysis of single amino acid variant peptides associated with pancreatic cancer in serum by an isobaric labeling quantitative method. *J Proteome Res.* 2014; 13(12):6058–6066. [PubMed: 25393578]
 41. Ruggles KV, Tang ZJ, Wang XY, Grover H, Askenazi M, Teubl J, Cao S, McLellan MD, Clauser KR, Tabb DL, Mertins P, Slebos R, Erdmann-Gilmore P, Li SQ, Gunawardena HP, Xie L, Liu T, Zhou JY, Sun SS, Hoadley KA, Perou CM, Chen X, Davies SR, Maher CA, Kinsinger CR, Rodland KD, Zhang H, Zhang Z, Ding L, Townsend RR, Rodriguez H, Chan D, Smith RD, Liebler DC, Carr SA, Payne S, Ellis MJ, Fenyo D. An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Mol Cell Proteomics.* 2016; 15(3):1060–1071. [PubMed: 26631509]
 42. Menon R, Im H, Zhang E, Wu SL, Chen R, Snyder M, Hancock WS, Omenn GS. Distinct splice variants and pathway enrichment in the cell-line models of aggressive human breast cancer subtypes. *J Proteome Res.* 2014; 13(1):212–227. [PubMed: 24111759]
 43. Wu JR, Zeng R. Molecular basis for population variation: from SNPs to SAPs. *FEBS Lett.* 2012; 586(18):2841–2845. [PubMed: 22828278]
 44. Chaney JL, Clark PL. Roles for synonymous codon usage in protein biogenesis. *Annu Rev Biophys.* 2015; 44:143–66. [PubMed: 25747594]
 45. Matuschek C, Lehnhardt M, Gerber PA, Poremba C, Hamilton J, Lammering G, Orth K, Budach W, Bojar H, Bolke E, Peiper M. Increased CD44s and decreased CD44v6 RNA expression are associated with better survival in myxofibrosarcoma patients: a pilot study. *Eur J Med Res.* 2014; 19(1):6. [PubMed: 24491153]
 46. Peng YH, Alexov E. Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding. *Proteins: Struct, Funct, Genet.* 2016; 84(2):232–239. [PubMed: 26650512]
 47. Hanagiri T, Shinohara S, Takenaka M, Shigematsu Y, Yasuda M, Shimokawa H, Nagata Y, Nakagawa M, Uramoto H, So T, Tanaka F. Effects of hyaluronic acid and CD44 interaction on the proliferation and invasiveness of malignant pleural mesothelioma. *Tumor Biol.* 2012; 33(6):2135–2141.

Figure 1. Venn diagram showing the overlap of SAAVs in three subpopulations, ALDH+, CD49f-EpCAM+, and CD44+/CD24-cells derived from MCF-7 cell line.

Figure 2.
Representative MS/MS spectra of (A) variant peptide and related (B) canonical peptide.

Figure 3. Frequency of the top eight SAAV's pairs from Swiss-CanSAAVs database (A) and MCF-7 cell line (B).

Figure 4.

GO annotations and interactions of associated proteins with SAAVs using IPA online software. Comparison of cellular component (A) and biological process (B) and network of 22 proteins with SAAVs site (C) via GO annotations.

Table 1Number of SAAVs Detected in MCF-7 Cell Line and Predicted Number Based on SAAVs Database^a

original amino acid	number of SAAVs detected	predicted number based on SAAVs database
A	29	30
C	14↑	10
D	19	18
E	21	20
F	2↓	8
G	31	29
H	3↓	11
I	17	16
K	19↑	12
L	22	23
M	15↑	11
N	16↑	13
P	12↓	24
O	20↑	12
R	40↓	50
S	30↑	27
T	23	22
V	30↑	26
W	4	4
Y	7	8

^aNumber increase or decrease is marked with upward arrow or downward arrow.

Table 2

Amino Acid Variant Trend Preference in Swiss-CanSAAVs Database and MCF-7 Cell Line

	Swiss-CanSAAVs		MCF-7	
	<u>ratio of variant trend (100% percent)</u>		<u>ratio of variant trend (100% percent)</u>	
	prevariant	after variant	prevariant	after variant
Ala	4.7	7.9	7.7	8.8
Cys	4.5	2.8	3.7	2.7
Asp	3.7	4.7	5.0	4.5
Glu	3.6	5.4	5.6	7.8
Phe	3.1	2.2	0.5	2.7
Gly	4.6	7.7	8.5	4.0
His	5.2	2.9	0.8	5.1
Ile	5.0	4.2	4.5	6.7
Lys	4.6	3.3	5.0	3.7
Leu	6.9	6.1	5.8	5.1
Met	3.9	2.9	4.0	1.9
Asn	4.2	3.5	4.2	5.3
Pro	5.3	6.4	3.2	5.9
Gln	4.8	3.3	5.3	4.5
Arg	8.6	13.4	10.9	7.0
Ser	8.2	7.1	8.0	8.3
Thr	6.6	6.0	6.4	5.6
Val	7.5	6.9	8.0	8.3
Trp	2.4	1.1	1.1	0.5
Tyr	2.5	2.3	1.9	1.6

Table 3

Top 20 of the Pattern of Amino Acid Variants in Swiss-CanSAAVs Database and MCF-7 Cell Line

	<u>Swiss-CanSAAVs</u>		<u>MCF-7</u>	
	<u>ratio of variant (%)</u>		<u>ratio of variant (%)</u>	
number	pattern	ratio	pattern	ratio
1	RQ	2.8	VI	4.0
2	AT	2.7	LP	3.2
3	PL	2.5	AV	2.9
4	RH	2.4	DN	2.7
5	EK	2.4	AT	2.7
6	AV	2.3	GS	2.7
7	RC	2.1	TA	2.7
8	GR	2.1	IV	2.4
9	VI	1.9	SA	2.4
10	RW	1.9	KE	2.1
11	LP	1.8	KR	2.9
12	PS	1.6	MT	1.9
13	VM	1.6	ND	1.9
14	GS	1.6	EK	1.9
15	DN	1.5	GR	1.9
16	IV	1.5	PL	1.9
17	TA	1.4	SN	1.6
18	NS	1.3	VA	1.6
19	TI	1.3	AS	1.6
20	TM	1.2	CY	1.6