**QUILTS: Sample Specific Protein Database Creation for the Identification of Variant, Novel and Alternatively Spliced Peptides from Cancer Proteomics.**

Ruggles, K.V.1, Tang, Z.1, Wang X.1, Teubl, J.1, Ovsiy, O.1, D. Fenyȍ1

*1New York University School of Medicine, New York, NY*

Corresponding Author: David Fenyȍ, Ph.D. tel: 212.263.2216, fax: 212.263.5995, email: david@fenyolab.org

**Running Title:** QUILTS: Software for Sample Specific Protein Database Creation

**Abbreviations:** AA, Amino Acid; BAM, Binary Alignment/Map ; BED, Browser Extensible Data; BWA, Burrows-Wheeler Alignment; CIMP, CPG island methylation phenotype; CIN, Chromosome Instability; DNA, Deoxyribonucleic Acid; MSI, Microsatellite Instability; NGS, Next Generation Sequencing; RNA, Ribonucleic Acid; SAM, Sequence Alignment/Map; SNP, Single Nucleotide Polymorphism; SNV, Single Nucleotide Variants; TCGA, The Cancer Genome Atlas;

**SUMMARY**

We have developed QUILTS, a proteogenomic integration software, which incorporates germline and somatic sequence variants, splice variants, novel expression, and fusion genes from DNA and RNA sequencing analysis into a reference proteome to create sample-specific protein sequence databases for mass spectrometry (MS) based protein identification. These databases allow for sensitive and specific identification of novel protein isoforms through inclusion of the full proteomic potential of each tumor while simultaneously minimizing database size. This tool is available for public use (quilts.fenyolab.org) and can be easily incorporated into current proteomics pipelines. To demonstrate the capabilities of QUILTS, we created tumor specific databases for 105 breast and 82 colorectal tumors from the Cancer Genome Atlas retrospective collection. We illustrate the number of predicted unique peptides based on nonsynonomous germline and somatic variant calls, unannotated alternative splicing and the translation of noncoding regions based on RNA-Seq junction data. An average of 13,315 and 13,886 germline and 668 and 2,710 somatic variant tryptic peptides were identified across the breast and colorectal tumors, respectively. Further, we identified an average of 65,277 breast and 35,755 colorectal unique peptides due to junction-based changes, comprised of a combination of unannotated alternative splicing events and the translation of noncoding (intronic, integenic, interexonic) gene regions. Additionally, we found that approximately 30% of the sequencing predicted protein changes are unverifiable by MS due to either tryptic peptide length or sequence homology. The results of the tumor database creation are presented with a focus on how the RNA-Seq and exome sequencing data can be used to detect single amino acid changes and the removal and addition of protein domains.

**INTRODUCTION**

Over the past decade, genomic studies have dominated the field of cancer biology, with major efforts from the Cancer Genome Atlas (TCGA), characterizing hundreds of tumors across 26 tumor types. DNA and RNA sequencing from these studies has provided tumor characterization including the identification of somatic mutations, amplifications and deletions, recombination events, disease marker expression, and significantly mutated genes ([1-10](#_ENREF_1)). Despite the tremendous strides made by cancer genomic studies in elucidating cancer biology, a significant gap between genome sequence and patient treatment still exists. In order to build upon the rich knowledge base genomic studies provide, large scale proteomic analysis offers a link between the genome and clinical prognosis and progression. Mass spectrometry (MS) based proteomic analysis is currently being completed on a subset of the TCGA tumors to determine cancer specific changes in protein expression, quantify post-translational modifications, particularly phosphorylation, and to identify novel protein isoforms and novel coding regions ([11](#_ENREF_11)). Tumor specific protein expression presents the opportunity for sensitive and specific drug targeting and biomarker development, the potential of which has not been fully realized.

Mass spectrometric identification of peptide sequences that are not contained in reference proteomes a non-trivial informatics challenge and it relies heavily on the quality of the protein sequence database used for identification. Specifically, databases with missing peptide sequences will fail to identify the corresponding peptides within the proteomic data, but inclusion of a large number of sequences in the search will decrease sensitivity. Therefore, an ideal database is both complete and small, containing the sequences of all proteins present in the sample and no other irrelevant sequences ([12](#_ENREF_12)). Given the diversity of protein isoforms in different cell types and the growing affordability of next generation sequencing (NGS) technology, it is advantageous to create sample-specific protein sequence databases for comprehensive peptide identification. RNA-Seq and genome sequencing information can be used to create these databases, incorporating variant proteins, alternatively spliced isoforms, and novel expression, as coded within the genome and transcriptome, allowing for the identification of sample specific peptides from the tandem MS analysis ([13-15](#_ENREF_13)).

The creation of sample-specific sequence databases is particularly important for cancer proteomics. Cancer patients acquire tumor specific somatic variants that can act as drivers of tumor progression influencing diagnosis, prognosis, and treatment strategies. These genetic rearrangements can differ quite drastically between tumors, even within the same tumor type ([16](#_ENREF_16)), highlighting the need for personalized database creation. Adding to genetic differences between tumors, large whole genome sequencing studies such as the 1000 genomes project have uncovered millions of germline variant differences between individuals ([17](#_ENREF_17)) and early cancer genetic studies identified mutations in rare, inherited alleles as important factors in cancer predisposition ([18](#_ENREF_18)). Therefore, both tumor and patient specific variant information should be included within an ideal database to identify inherited and acquired genomic variants contributing to tumor progression.

In addition to these single nucleotide variants (SNVs), transcript isoforms due to alternative splicing or translation have also been implicated in tumor growth and progression ([19-21](#_ENREF_19)). Novel transcript isoforms resulting from coding of unannotated exon regions or fusion gene transcription provide prospective cancer specific proteins for biomarker development and drug targeting. Therefore, the incorporation of both genomic and transcriptomic into the protein sequence database is required for comprehensive tumor peptide identification

For this purpose, we have developed the software tool, QUILTS, which incorporates SNVs, novel splice junctions and fusion gene information for tumor specific database construction. Several similar tools have been developed to create customizable protein databases focusing on the incorporation of fusion genes ([22](#_ENREF_22)), SNVs ([13](#_ENREF_13), [15](#_ENREF_15), [23](#_ENREF_23)), *de novo* assembled transcripts([24](#_ENREF_24), [25](#_ENREF_25)), or splice-junction peptides ([23](#_ENREF_23), [26-29](#_ENREF_26)). QUILTS uses both DNA and RNA sequencing data to create the protein database, including the majority of potential genetic variation and rearrangements, while simultaneously maintaining a small database size. The number of sequences included in the protein database is limited by using only sequencing data for one sample at a time and by carefully selecting the number of required reading frame translations necessary for each genomic variation situation (as to not require 6 reading frame translations for every case). This allows for improved identification of novel protein isoforms identification in mass spectrometry-based cancer proteomic studies. We used a subset of the breast and colorectal sequencing data generated by the TCGA to demonstrate the sample specific databases creation using QUILTS software.

**EXPERIMENTAL PROCEDURES**

*Samples and Data Processing*

Available exome sequencing and RNA-Seq data in BAM file format were downloaded from The Cancer Genome Atlas (TCGA) data portal (https://cghub.ucsc.edu/) for human breast invasive carcinoma (BRCA) and colorectal carcinoma (CRC). The number of unique patients used for each tumor type was 105 and 82 respectively. Exome sequencing BAM files, all using Illumina methodology, were first converted back to FASTQ files using software Picard version 1.79 (http://picard.sourceforge.net) while RNA-Seq BAM files were converted to FASTQ files using in-house developed software and SAMtools version 0.1.19 ([30](#_ENREF_30)). Quality control analysis was completed on both sequencing types using FastQC version 0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Exome sequences were trimmed to 75 bps in length by tail trimming to increase read quality. These trimmed reads, in FASTQ format, were then aligned to the human reference genome version hg19 using the Burrows-Wheeler Alignment tool (BWA) version 0.7.3a-r367 ([31](#_ENREF_31)). SAMtools version 0.1.19 ([30](#_ENREF_30)) was used to convert the resulting SAM files into BAM files followed by sorting and indexing using Picard version 1.79. Mapped reads in the raw BAM files were then marked for duplicates, re-aligned locally, and base pair scores were re-calibrated using GATK version 2.6 ([32](#_ENREF_32), [33](#_ENREF_33)) and Picard version 1.79. Finally, software GATK version 2.6 was used to call variants (SNPs and indels) for both tumor (somatic) and patient (germline) samples.

RNA-Seq reads were trimmed by the last two base pairs to increase read quality. BWA alignment software version 0.7.3a-r367 ([31](#_ENREF_31)) alongside in-house developed software was used to remove contaminated sequences from sequencing adapters, mitochondrial and ribosomal DNA, enterobacteria phage phiX174, polyA and polyC. All cleaned and trimmed reads in FASTQ format were aligned to the human reference genome version hg19 using TopHat version 2.0.8 ([34-36](#_ENREF_34)) with –g 1, --bowtie2 (version 2.1.0.0), -M, -x 1, and --fusion-search settings to generate BAM files and junction files. We used all available sequencing data for each sample, which included one BAM file for each breast tumor, and up to 5 BAM files for each colorectal tumor. Read length differed between cancer types, with 75 base paired reads in breast and 50 base paired reads in colorectal.

*Software Design*

QUILTS is a Perl-based software which can take in up to 4 inputs for sample-specific database creation (**Fig. 1**). Currently, our web-accessible software is able to use either Ensembl or RefSeq as a reference for the hg19 proteome and genome. The possible input files for sample specific information include a BED file containing RNA-Seq predicted junctions, VCF files containing somatic variants and germline variants and a fusion file containing all predicted fusion genes. QUILTS is able to create FASTA files from sequencing data so long as at least 1 sample specific input file is included. Database assembly is outlined below using the TCGA tumors as an example.

*Step 1. Creation of Variant Peptide Database*

Details of variant location and nucleotide change were parsed out from the variant VCF file for subsequent incorporation into genomic sequences. Somatic variants were obtained by filtering out all germline variants found in the tumor variant calls, leaving only tumor specific variants. Based on intron/exon boundaries from Ensembl (version 70), sequences of annotated coding regions were extracted and variant changes were incorporated into these regions based on genomic location. The modified sequences were then translated to proteins and stored as a FASTA file (**Fig. 2H**). Stop codon removal and insertion due to single amino acid changes were accounted for and highlighted within file output.

*Step 2. Creation of Junction Protein Database*

The RNA-Seq derived junction BED file contains predicted boundaries of adjoining exons (chromosome, intron start, intron end) for each sample. Splice junctions matching previously annotated boundaries (Ensembl version 70) were first filtered out, leaving only novel junction coordinates. These novel junctions were split into 3 categories, unannotated alterative splicing, completely novel junctions, and partially novel junctions. Unannotated alternative splicing occurs when two annotated exons are spliced in a previously undescribed way, creating a new isoform from known coding regions (**Fig. 2A**). Completely novel junctions is defined by two intronic, intergenic or interexonic regions being spliced together, creating a completely novel transcript from “non-coding” regions (**Fig. 2E, F**). Lastly, partially novel junctions occur when one intron/exon boundary is annotated while the connecting boundary is novel (**Fig. 2B-D**). The required frame translation for *in silico* protein synthesis is indication in Figure 2 for each scenario. Variant peptides are incorporated into the sequences of the alternatively spliced proteins as described in Step 1 (**Fig. 2H**).

*Step 3. Creation of Fusion Protein Database*

Predicted fusion gene output is translated by a 6 frame translation (**Fig. 1G**). Protein coding regions with more than 6 consecutive amino acids are included in the fusion protein database. Although QUILTS is capable of creating fusion-based databases, and this capability is available on for public use, no fusion data is presented in this manuscript.

**RESULTS**

The impressive efforts of the TCGA have provided publically accessible data for RNA and exome sequencing for thousands of tumors across several tumor types (https://tcga-data.nci.nih.gov/tcga/). We used sequencing analysis of 105 breast tumors and 82 colorectal tumors from this retrospective collection to demonstrate the capabilities of QUILTS in database creation. Breast tumors are comprised of 3 characterized subtypes, luminal, HER2, and basal-like, based on the expression of the oestrogen receptor (ER), progesterone receptor (PR) and HER2. Luminal tumors are ER positive, the HER2 group has amplification of the HER2 gene and the triple-negative group (also known as Basal-like) lack ER, PER and HER2 expression ([38](#_ENREF_38)). Additionally, three colorectal expression subtypes have been described: (1) microsatellite instability (MSI) and CPG island methylation phenotype (CIMP); (2) chromosomal instability (CIN), and (3) an invasive phenotype ([5](#_ENREF_5)). All breast and colorectal subtypes are represented within our tumor set, and are highlighted throughout our analysis.

*Variant Peptide Prediction*

Mass spectrometry based proteomics relies on enzymatic protein digestion, typically tryptic digestion, cleaving peptide chains at the carboxyl end of lysine (K) or arginine (R). Identification by MS methods requires specific chemical properties including low hydrophobicity and peptide length within approximately 6-30 amino acids. Additionally, if a peptide maps to more than one protein it is difficult to determine its protein of origin ([39](#_ENREF_39)) . This is a particular issue for homology of peptides containing single amino acid changes due to nonsynonomous gene mutations or peptides translated from novel DNA expression, since we are unable to confidently identify them as being sample-specific. For these reasons, only proteotypic tumor-specific changes with appropriate amino acid lengths (6-30) can be used to validate the sequencing predicted changes in MS/MS data. We have used these limitations to categorize the predicted proteomic changes either as verifiable or unverifiable by tandem MS. This allows us to determine the full mass spectrometric identification potential of each sample as well as peptide variants which cannot be verified using MS methods.

Both somatic and germline DNA sequencing was available for 97 breast and 74 colorectal tumors of the 105 and 82 total. For cases with only somatic variant data (8 of each tumor type), all DNA sequencing was classified as germline, as we were unable to differentiate between tumor specific and patient specific variation. Across breast tumors, the total number of predicted unique variant peptides within the appropriate length constraints ranged from 11,292 and 16,407 germline variants and 2 to 2,006 somatic variants per sample (omitting samples with only somatic variant calls). Similar analysis of the 82 colorectal samples found more inter-sample variation, consistent with previous analysis, demonstrating sequencing hypervariability in a portion of TCGA colorectal tumor samples ([5](#_ENREF_5)). We identified 7,368 to 27,315 predicted germline variants and 1 to 19,110 somatic variants across all tumor samples evaluated (**Fig. 3, Fig. S1**). Variation in the number of reads sequenced was unable to account for the clear differences in predicted somatic variant peptides across samples. The average number of DNA sequencing reads for colon (3.6e8) and breast (4.4e8) are on the same order of magnitude, and the number of reads was not correlated to the number of predicted variant trypic peptides (r2=0.0130 for breast; r2=0.0281 for colorectal). Since colorectal tumors had up to 5 sequencing files available, merging of multiple sequencing files may have resulting in increased sequencing error (**Fig. S3**). The remaining difference is likely explained by differences in tumor biology, as TCGA colorectal tumors have been shown to have a large variation in somatic mutations, with high mutation rates resulting from mutations in the DNA mismatch repair pathway ([5](#_ENREF_5)).

As for unverifiable peptides, on average, 31% of germline and 35% of somatic tryptic breast tumor peptides were filtered due to length constraints, being either <6 or >30 amino acids. Further, more than 3% of peptides supporting variant induced protein change were found to be homologous to the Ensembl reference proteome. Filtering by both peptide length and homology, the number of potential variants which can be identified by MS/MS in these tumors drops, on average, from 21,852 to 13,983 for breast and from 23,485 to 16,595 for colorectal (**Fig. 3, Fig. S1**).

*Alternative Isoforms and Novel Expression*

Alternative splicing and coding of novel gene regions contributes even more complexity to the possible protein landscape of a tumor. The rise of RNA-Seq analysis has uncovered the prevalence of alternative splicing in gene transcription with up to 90% of genes having at least 2 transcripts ([40](#_ENREF_40)), highlighting the importance of such isoforms in cellular function. QUILTS considers several scenarios of RNA-Seq supported, novel intron-exon boundaries including unannotated splicing events in which known exons are combined in new ways (**Fig 2A**), partially novel splicing where a known exon is attached to intronic, interexonic, or intergenic gene regions (**Fig 2B-D**), or completely novel splicing where two noncoding genomic regions are connected (**Fig 2E,F**). We considered each of these classifications separately for both tumor types.

We first determined the full MS/MS verifiable potential of proteomic change resulting from novel translation junctions within each tumor. These changes are based on RNA-Seq predicted intron/exon boundaries are not annotated in the reference database (Ensembl version 70). Across the 105 breast tumors, between 1,141 and 5,621 unannotated splicing events, 15,267 to 54,989 partially novel events and between 15,343 and 38,467 completely novel exon events were identified. The colorectal data had consistently lower predicted junction peptides compared to that seen in the breast data. Colorectal analysis identified between 667 to 4,620 unannotated splicing events and 6,696 to 43,612 partially novel, and 4,949 to 35,501 completely novel exon events (**Fig. 4, Fig. S2**). In contrast to the DNA sequencing analysis, read number in RNA-showed some correlation with the number of predicted novel peptides (breast R2=0.272; colon R2=0.542). Therefore, the lower number of predicted novel variant peptides can be partially attributed to differences in the number of reads sequenced by RNA-Seq analysis, with a 6-folder high RNA read number in breast versus colon (**Fig S3**).

We found that, on average, 2.51% of unannotated alternative splicing, 1.37% of partially novel events, and 0.01% of completely novel events identified in breast tumors were homologous to the reference proteome. Likewise, colorectal tumors had 2.50% of unannotated alternative splicing, 1.33% of partially novel, and 0.02% of completely novel events have homology to annotated proteins (**Fig 4**). Further, 26.4% and 26.8% of peptides supporting unannotated alternative splicing were outside of the length constraints (>6, <30 amino acids) for breast and colorectal, respectively. In summary, approximately 70% of potential protein changes predicted by variant and junction data meet the qualifications for MS/MS measurement and use of these sample specific databases can assist in their identification.

*Most commonly mutated proteins*

Genomic studies including the TCGA have completed comprehensive analysis of these tumors, identifying significantly mutated genes within cancer types ([5](#_ENREF_5), [9](#_ENREF_9), [38](#_ENREF_38), [41](#_ENREF_41), [42](#_ENREF_42)) and across many different cancers ([8](#_ENREF_8), [21](#_ENREF_21)). We have used QUILTS database creation capabilities to identify the most common anticipated proteomic changes based on novel junction predictions across this tumor subset. As single nucleotide polymorphisms have been extensively studied in these cancers ([5](#_ENREF_5), [43](#_ENREF_43)), we focused specifically on instances of unannotated alternative splicing and partially novel expression, in which a known exon is connected to a region of noncoding DNA. Figure 5 shows each of the novel splicing events identified in both colorectal and breast, plotted by the percentage of colorectal and breast tumors in which it is predicted to occur. Alternative splicing in three genes, CASK, a calcilum/calmodulin dependent protein kinase, the scaffold protein SPTAN1, and FUBP3, was reported to occur in more than 90% of both tumors types, indicating either a shared cancer specific event, or lack of annotation for normal intron/exon boundaries in these genes (**Fig. 5A**). In comparison, 19 genes were found to have novel exon expression in more than 90% of the colorectal and breast tumors, including cell migration related genes TLN1, PRPF40A, and apoptotic regulators TRAF7 and BAX (**Fig 5B**).

In addition to junctions found with high frequency in both tumors types, we also highlighted changes occurring in only breast or colorectal as potential tumor type or tissue specific markers. A total of 459 (344 breast specific and 115 colorectal specific) genes with alternative splicing and 2000 (1712 breast specific, 288 colorectal specific) genes containing novel exons were found to have differential representation by tumor type (**Fig. 5**, red). The cell adhesion genes CDH5R and CDH17, the cytokine IL33, and the mucin genes MUC2 and MUC13 were among the differential colorectal specific novel junctions discovered (**Fig 5**). Changes occurring most frequently in breast include novel exons in MAPK8/JNK activator MAPK4K4, the tumor suppressor RNF40 (BRE1), and novel alternative splicing in the cell surface metalloprotease ADAM10 (**Fig. 5**). Although these differential changes are intriguing, evidence of their translation or regulatory impact is needed before conclusions can be made on their involvement in tumor progression. The presented custom databases created by QUILTS allows for the validation of these novel changes within MS/MS proteomic analysis, educating our junction annotation and, with luck, identifying cancer specific treatment targets.

*Database size*

Minimizing database size is important in peptide identification, not only to reduce search time, but also because a larger database will result in greater false spectral matches. The effects of database size have been demonstrated, indicating that smaller databases result in more peptide identifications ([44-46](#_ENREF_44)). Figure 1 shows two possible Quilt outputs and their associated database size in the 105 breast tumors studied, quantified by unique peptide number. Creation of a sample specific database in which only tumors are processed alone results in a database containing approximately 640,000 unique peptides, only a 15% increase compared to the Ensembl reference database (558,811 peptides) (**Fig. 1A**). Combining variant and junction peptides across all 105 breast tumors results in a larger consensus database, containing 134,852 unique variant peptides and 1,799,638 junction peptides (**Fig. 1B**). This database, comprising of 2,493,301 unique peptides including the Ensembl reference proteome, is more than 4 times larger than the original Ensembl reference protein database. In addition to QUILTS’ sample specific databases, protein sequence databases using a 6-frame translation of the genome or with the inclusion of all dbSNP variant calls have also been discussed in the literature ([47-49](#_ENREF_47)). By our calculations, these methods result in 262,399,857 unique peptides associated with a 6-frame translation (**Fig. 1B**) and 858,934 with dbSNP inclusion (**Fig. 1D**), both of which are much larger than the sample specific output (**Fig. 1A**). Moreover, recent work by the 1000 genomes project has uncovered frequency-based genetic variation within the population ([17](#_ENREF_17)). Incorporation of the 1000 genome variants into a protein database using QUILTS results in 33,864 additional novel peptides (**Fig. 1C**), without any prior filtering, a comparatively minor increase.

**DISCUSSION**

The field of proteogenomics has seen incredible growth in the last 5 years, focusing on the integration of genomics and proteomics through peptide mapping and the use of sequencing data to attain a comprehensive view of the proteome. For this reason, we, and others ([23](#_ENREF_23), [27](#_ENREF_27)), have created tools to utilize NGS data to better annotate, identify, and quantify proteins. QUILTS is able to incorporate both DNA and RNA sequencing analysis to create comprehensive and sample specific protein sequence databases for MS/MS peptide identification, while minimizing database size to control for false discovery. The output of QUILTS, a FASTA formatted sequence database, can be used by common database search engines including X! Tandem ([50](#_ENREF_50)), MSGF+, Mascot([51](#_ENREF_51)), and Spectrum Mill, and is therefore easily incorporated into protein identification pipelines. We were able to control for database size by carefully considering each junction scenario for reading frame conservation, thereby minimizing the number of peptides considered. QUILTS can be easily used to create any number of sample specific databases based on Ensembl or RefSeq as a reference database.

For the purposes of database creation, we considered all RNA-Seq junction predictions without taking into account the number of reads supporting each junction. The reason behind this liberal inclusion was because we wanted the database to be complete, containing all possible proteomic changes, with the idea that the proteomic data can be used to filter through transcriptional false positives. Upstream filtering of junctions and variants based on read number can be used to reduce noise within the database and filtering based on 1000 genomes population frequencies or dbSNP identification can be used to remove known variants, depending on the requirements of the study.

Although our focus has been on the identification of novel proteins in cancer, the use of QUILTS can be expanded into other fields requiring sequence-based proteomic databases. This includes improving the annotation of gene structure in organisms lacking well annotated reference proteomes. Similar techniques have been applied to the maize genome (*Zea mays*)using splice graph methodology([14](#_ENREF_14)) and can be expanded to include other understudied organisms. Additionally, we plan to implement QUILTS database creation in studies focused on antibody discovery and venom proteomics, both of which require custom databases due to their hypervariability through recurrent gene recombination ([52-57](#_ENREF_52)). Lastly, RNA-processing and post-transcriptional modification have been shown to play an important role in neural plasticity and synapse formation (See Review([58](#_ENREF_58))). Sequencing-based protein databases can help to elucidate the effects of this RNA-editing on the brain proteome and in turn, its role in neuronal development.

QUILTS is freely available for public use at quilts.fenyolab.org, with the option of including intron/exon boundary information, variant calls and/or fusion genes to create protein sequence databases for tandem MS identification. The database output can then be used for novel, tumor specific peptide identification, informing both cancer biology and drug targeting. Further, QUILTS adds yet another layer of available data into the personalized medicine pipeline, supplementing the expansive genomic and transcriptomic analysis which has been collected over the past decade.

**ACKNOWLEDGMENTS**

This work was supported by funding provided by the National Institutes of Health through grant CA160035 and contract S13-068 from SAIC. This work has utilized computing resources at the High Performance Computing Facility of the Center for Health Informatics and Bioinformatics at the New York University Langone Medical Center.

**REFERENCES**

1. TCGA (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615

2. Tameru, B., Nganwa, D., Bogale, A., Robnett, V., and Habtemariam, T. (2012) The Role of Computational Epidemiology and Risk Analysis in the Fight Against HIV/AIDS. *Journal of AIDS & clinical research* **3**

3. Kandoth, C., Schultz, N., Cherniack, A. D., Akbani, R., Liu, Y., Shen, H., Robertson, A. G., Pashtan, I., Shen, R., Benz, C. C., Yau, C., Laird, P. W., Ding, L., Zhang, W., Mills, G. B., Kucherlapati, R., Mardis, E. R., and Levine, D. A. (2013) Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67-73

4. TCGA (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England journal of medicine* **368**, 2059-2074

5. TCGA (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337

6. Ellis, M. J., Ding, L., Shen, D., Luo, J., Suman, V. J., Wallis, J. W., Van Tine, B. A., Hoog, J., Goiffon, R. J., Goldstein, T. C., Ng, S., Lin, L., Crowder, R., Snider, J., Ballman, K., Weber, J., Chen, K., Koboldt, D. C., Kandoth, C., Schierding, W. S., McMichael, J. F., Miller, C. A., Lu, C., Harris, C. C., McLellan, M. D., Wendl, M. C., DeSchryver, K., Allred, D. C., Esserman, L., Unzeitig, G., Margenthaler, J., Babiera, G. V., Marcom, P. K., Guenther, J. M., Leitch, M., Hunt, K., Olson, J., Tao, Y., Maher, C. A., Fulton, L. L., Fulton, R. S., Harrison, M., Oberkfell, B., Du, F., Demeter, R., Vickery, T. L., Elhammali, A., Piwnica-Worms, H., McDonald, S., Watson, M., Dooling, D. J., Ota, D., Chang, L. W., Bose, R., Ley, T. J., Piwnica-Worms, D., Stuart, J. M., Wilson, R. K., and Mardis, E. R. (2012) Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353-360

7. TCGA (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43-49

8. Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., Leiserson, M. D., Miller, C. A., Welch, J. S., Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J., and Ding, L. (2013) Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339

9. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615

10. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068

11. Ellis, M. J., Gillette, M., Carr, S. A., Paulovich, A. G., Smith, R. D., Rodland, K. K., Townsend, R. R., Kinsinger, C., Mesri, M., Rodriguez, H., and Liebler, D. C. (2013) Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer discovery* **3**, 1108-1112

12. Fenyo, D., Eriksson, J., and Beavis, R. (2010) Mass spectrometric protein identification using the global proteome machine. *Methods Mol Biol* **673**, 189-202

13. Wang, X., Slebos, R. J., Wang, D., Halvey, P. J., Tabb, D. L., Liebler, D. C., and Zhang, B. (2012) Protein identification using customized protein sequence databases derived from RNA-Seq data. *Journal of proteome research* **11**, 1009-1017

14. Castellana, N. E., Shen, Z., He, Y., Walley, J. W., Cassidy, C. J., Briggs, S. P., and Bafna, V. (2013) An Automated Proteogenomic Method Utilizes Mass Spectrometry to Reveal Novel Genes in Zea mays. *Molecular & cellular proteomics : MCP*

15. Li, J., Su, Z., Ma, Z. Q., Slebos, R. J., Halvey, P., Tabb, D. L., Liebler, D. C., Pao, W., and Zhang, B. (2011) A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Molecular & cellular proteomics : MCP* **10**, M110 006536

16. Stephens, P. J., McBride, D. J., Lin, M. L., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. A., Leroy, C., Edkins, S., Mudie, L. J., Greenman, C. D., Jia, M., Latimer, C., Teague, J. W., Lau, K. W., Burton, J., Quail, M. A., Swerdlow, H., Churcher, C., Natrajan, R., Sieuwerts, A. M., Martens, J. W., Silver, D. P., Langerod, A., Russnes, H. E., Foekens, J. A., Reis-Filho, J. S., van 't Veer, L., Richardson, A. L., Borresen-Dale, A. L., Campbell, P. J., Futreal, P. A., and Stratton, M. R. (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005-1010

17. Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073

18. Pomerantz, M. M., and Freedman, M. L. (2011) The genetics of cancer risk. *Cancer J* **17**, 416-422

19. Davuluri, R. V., Suzuki, Y., Sugano, S., Plass, C., and Huang, T. H. (2008) The functional consequences of alternative promoter use in mammalian genomes. *Trends in genetics : TIG* **24**, 167-177

20. Mardis, E. R., Ding, L., Dooling, D. J., Larson, D. E., McLellan, M. D., Chen, K., Koboldt, D. C., Fulton, R. S., Delehaunty, K. D., McGrath, S. D., Fulton, L. A., Locke, D. P., Magrini, V. J., Abbott, R. M., Vickery, T. L., Reed, J. S., Robinson, J. S., Wylie, T., Smith, S. M., Carmichael, L., Eldred, J. M., Harris, C. C., Walker, J., Peck, J. B., Du, F., Dukes, A. F., Sanderson, G. E., Brummett, A. M., Clark, E., McMichael, J. F., Meyer, R. J., Schindler, J. K., Pohl, C. S., Wallis, J. W., Shi, X., Lin, L., Schmidt, H., Tang, Y., Haipek, C., Wiechert, M. E., Ivy, J. V., Kalicki, J., Elliott, G., Ries, R. E., Payton, J. E., Westervelt, P., Tomasson, M. H., Watson, M. A., Baty, J., Heath, S., Shannon, W. D., Nagarajan, R., Link, D. C., Walter, M. J., Graubert, T. A., DiPersio, J. F., Wilson, R. K., and Ley, T. J. (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *The New England journal of medicine* **361**, 1058-1066

21. Balmain, A., Gray, J., and Ponder, B. (2003) The genetics and genomics of cancer. *Nat Genet* **33 Suppl**, 238-244

22. Frenkel-Morgenstern, M., Lacroix, V., Ezkurdia, I., Levin, Y., Gabashvili, A., Prilusky, J., Del Pozo, A., Tress, M., Johnson, R., Guigo, R., and Valencia, A. (2012) Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome research* **22**, 1231-1242

23. Wang, X., and Zhang, B. (2013) customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics*

24. Adamidi, C., Wang, Y., Gruen, D., Mastrobuoni, G., You, X., Tolle, D., Dodt, M., Mackowiak, S. D., Gogol-Doering, A., Oenal, P., Rybak, A., Ross, E., Sanchez Alvarado, A., Kempa, S., Dieterich, C., Rajewsky, N., and Chen, W. (2011) De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. *Genome research* **21**, 1193-1200

25. Evans, V. C., Barker, G., Heesom, K. J., Fan, J., Bessant, C., and Matthews, D. A. (2012) De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nature methods* **9**, 1207-1211

26. Sheynkman, G. M., Shortreed, M. R., Frey, B. L., and Smith, L. M. (2013) Discovery and Mass Spectrometric Analysis of Novel Splice-junction Peptides Using RNA-Seq. *Molecular & cellular proteomics : MCP* **12**, 2341-2353

27. Woo, S., Cha, S. W., Merrihew, G., He, Y., Castellana, N., Guest, C., Maccoss, M., and Bafna, V. (2013) Proteogenomic Database Construction Driven from Large Scale RNA-seq Data. *J Proteome Res*

28. Xing, X. B., Li, Q. R., Sun, H., Fu, X., Zhan, F., Huang, X., Li, J., Chen, C. L., Shyr, Y., Zeng, R., Li, Y. X., and Xie, L. (2011) The discovery of novel protein-coding features in mouse genome based on mass spectrometry data. *Genomics* **98**, 343-351

29. Omenn, G. S., Yocum, A. K., and Menon, R. (2010) Alternative splice variants, a new class of protein cancer biomarker candidates: findings in pancreatic cancer and breast cancer with systems biology implications. *Disease markers* **28**, 241-251

30. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079

31. Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760

32. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303

33. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498

34. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36

35. Kim, D., and Salzberg, S. L. (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome biology* **12**, R72

36. Trapnell, C., Pachter, L., and Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111

37. The Cancer Genome Atlas

38. TCGA (2012) Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70

39. Craig, R., Cortens, J. P., and Beavis, R. C. (2005) The use of proteotypic peptide libraries for protein identification. *Rapid communications in mass spectrometry : RCM* **19**, 1844-1850

40. Pal, S., Gupta, R., and Davuluri, R. V. (2012) Alternative transcription and alternative splicing in cancer. *Pharmacology & therapeutics* **136**, 283-294

41. Corley, C. D., Lancaster, M. J., Brigantic, R. T., Chung, J. S., Walters, R. A., Arthur, R. R., Bruckner-Lea, C. J., Calapristi, A., Dowling, G., Hartley, D. M., Kennedy, S., Kircher, A., Klucking, S., Lee, E. K., McKenzie, T., Nelson, N. P., Olsen, J., Pancerella, C., Quitugua, T. N., Reed, J. T., and Thomas, C. S. (2012) Assessing the continuum of event-based biosurveillance through an operational lens. *Biosecur Bioterror* **10**, 131-141

42. (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43-49

43. (2012) Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70

44. Jeong, K., Kim, S., and Bandeira, N. (2012) False discovery rates in spectral identification. *BMC bioinformatics* **13 Suppl 16**, S2

45. Eriksson, J., and Fenyo, D. (2004) The statistical significance of protein identification results as a function of the number of protein sequences searched. *Journal of proteome research* **3**, 979-982

46. Falth, M., Skold, K., Svensson, M., Nilsson, A., Fenyo, D., and Andren, P. E. (2007) Neuropeptidomics strategies for specific and sensitive identification of endogenous peptides. *Molecular & cellular proteomics : MCP* **6**, 1188-1197

47. Risk, B. A., Spitzer, W. J., and Giddings, M. C. (2013) Peppy: proteogenomic search software. *Journal of proteome research* **12**, 3019-3025

48. Bunger, M. K., Cargile, B. J., Sevinsky, J. R., Deyanova, E., Yates, N. A., Hendrickson, R. C., and Stephenson, J. L., Jr. (2007) Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. *Journal of proteome research* **6**, 2331-2340

49. Bitton, D. A., Smith, D. L., Connolly, Y., Scutt, P. J., and Miller, C. J. (2010) An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome. *PloS one* **5**, e8949

50. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466-1467

51. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567

52. Fischer, N. (2011) Sequencing antibody repertoires: the next generation. *mAbs* **3**, 17-20

53. Prabakaran, P., Zhu, Z., Chen, W., Gong, R., Feng, Y., Streaker, E., and Dimitrov, D. S. (2012) Origin, diversity, and maturation of human antiviral antibodies analyzed by high-throughput sequencing. *Frontiers in microbiology* **3**, 277

54. Escoubas, P., Quinton, L., and Nicholson, G. M. (2008) Venomics: unravelling the complexity of animal venoms with mass spectrometry. *Journal of mass spectrometry : JMS* **43**, 279-295

55. Rodriguez de la Vega, R. C., Schwartz, E. F., and Possani, L. D. (2010) Mining on scorpion venom biodiversity. *Toxicon : official journal of the International Society on Toxinology* **56**, 1155-1161

56. Ueberheide, B. M., Fenyo, D., Alewood, P. F., and Chait, B. T. (2009) Rapid sensitive analysis of cysteine rich peptide venom components. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 6910-6915

57. Scheid, J. F., Mouquet, H., Ueberheide, B., Diskin, R., Klein, F., Oliveira, T. Y., Pietzsch, J., Fenyo, D., Abadir, A., Velinzon, K., Hurley, A., Myung, S., Boulad, F., Poignard, P., Burton, D. R., Pereyra, F., Ho, D. D., Walker, B. D., Seaman, M. S., Bjorkman, P. J., Chait, B. T., and Nussenzweig, M. C. (2011) Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* **333**, 1633-1637

58. Kiebler, M. A., Scheiffele, P., and Ule, J. (2013) What, where, and when: the importance of post-transcriptional regulation in the brain. *Frontiers in neuroscience* **7**, 192

**FIGURE LEGENDS**

**Figure 1. QUILTS database creation. (A)** QUILTS is able to generate tumor specific protein databases, through the incorporation of germline variant calls from normal patient tissue, somatic variant calls, junction prediction and fusion genes from tumor tissue and a reference genome and protein database (either Ensembl or RefSeq). Across the 105 TCGA breast tumors studied, we found the average size of the sample specific databases to be approximately 640,000 unique peptides. (**B**) A composite database can also be created, using all unique variant and junction peptides across all breast tumors. Protein database size (measured by unique peptide number) through the incorporation of all variants in the dbSNP variant database (**C**), all variants identified in the 1000 genomes project ([17](#_ENREF_17)) (**D**) or with a 6-frame translation of the entire human genome (**E**).

**Figure 2. QUILTS processing of different variant scenarios.** QUILTS treats each potential splicing situation differently, in terms of how many frames are translated into protein. Junction-based changes include unannotated alternative splicing with conserved exon boundaries (**A**), truncation of an exon (**B**), and elongation of an exon within an intron (**C**), all with conserved reading frame. For elongation of an exon in the intergenic space (**D**) and novel exon expression (**E**), frame translation is determined based on whether or not the novel exon boundary is up or downstream of the annotated exon. Junctions showing completely novel expression require a 6 reading frame translation (**F**). Fusion genes (**G**) are also translated in 6 frames. Single amino acid changes resulting from germline or somatic variants require only 1 reading frame **(H**), though situations with insertion or deletion of nucleotides are treated as junction changes as indicated above (**B-D**).

**Figure 3. Predicted proteomic change based on single nucleotide variants.**  Number of somatic (**A**) and germline (**B**) variants predicted to result in novel peptides in each breast (n=105) and colorectal (n=82) tumor surveyed. Black bars designate variant tryptic peptides which result from a single amino acid change, but which still has homology to the reference genome. Gray bars indicates variant tryptic peptides which are <6 or >30 amino acids long. Maroon bars show the number of unique somatic variant peptides and green, the number of germline predicted variant peptides which are verifiable by MS/MS for each sample.

**Figure 4. Peptides supporting alternative splicing and novel expression in breast and colorectal samples.** Number of unannotated alternative splicing (**A**) , partially novel expression (**B**) and completely novel expression events (**C**) predicted to result in novel peptides in each breast (n=105) and colorectal (n=82) tumor surveyed. Black designates junction tryptic peptides which result from a novel intron/exon boundary, but which still has homology to the reference genome. Gray indicates junction tryptic peptides which are <6 or >30 amino acids long. Orange bars show the number of unique unannotated alternatively spliced peptides, blue indicates the number of partially novel expression events and purple, the number of completely novel expression events which are verifiable by MS/MS for each sample.

**Figure 5. Differential proteomic changes in colorectal and breast tumors.** Each unannotated alternative splicing event (**A**) and novel exon event (**B**) were plotted as the percent of samples containing the event in colorectal versus the percent in breast. Events which have differential percentages between colorectal and breast (>30% difference) are highlighted in red by the corresponding annotated gene. Labeled grey dots are events found in more than 90% of both colorectal and breast samples. Labeled red dots are the top 20 breast or colorectal genes associated with each event, based on the percent difference between the two tumor types.

**Supplementary**

**Figure S1.** Boxplots showing the mean number of predicted variant peptides and number of DNA sequencing reads for breast and colorectal samples corresponding to (**A**) germline, (**B**) somatic single nucleotide variants.

**Figure S2.** Boxplots showing the mean number of predicted novel peptides for breast and colorectal samples corresponding to unannotated alternative splicing (**A**), partially novel expression (**B**) and completely novel expression (**C**) in breast and colorectal tumors.

**Figure S3**. Comparison of sequencing reads and predicted novel trypic peptide number. Boxplot showing the mean number of total DNA sequencing reads (**A**) and the number of somatic RNA-seq reads for breast (**B**) and colorectal (**C**) tumors. The number of total DNA sequencing reads versus the number of predicted variant peptides for each sample (**D**). The number of total somatic RNA-Seq reads versus the number of predicted somatic variant peptides for each sample (**E**).