Journal of proteome research

Article

Individual variability of protein expression in human tissues

Irena K Kushner, Geremy Clair, Samuel Owen Purvine,
Joon-Yong Lee, Joshua N. Adkins, and Samuel H. Payne

J. Proteome Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.jproteome.8b00580 • Publication Date (Web): 09 Oct 2018

ACS Publications

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Individual variability of protein expression in human tissues

Irena K. Kushner[1], Geremy Clair[1], Samuel Owen Purvine[1], Joon-Yong Lee[1], Joshua N. Adkins[1], Samuel H. Payne[1,2,*]

1. Biological Sciences Division, Pacific Northwest National Laboratory, Richland WA 99336
2. Current address: Biology Department, Brigham Young University, Provo UT 84602

Contact: sam_payne@byu.edu

Abstract

Human tissues are known to exhibit inter-individual variability, but a deeper understanding of the different factors affecting protein expression is necessary to further apply this knowledge. Our goal was to explore the proteomic variability between individuals as well as between healthy and diseased samples, and to test the efficacy of machine learning classifiers. In order to investigate whether disparate proteomics datasets may be combined, we performed a retrospective analysis of proteomics data from 9 different human tissues. These datasets represent several different sample prep methods, mass spectrometry instruments, and tissue health. Using these data, we examined inter-individual and inter-tissue variability in peptide expression, and analyzed the methods required to build accurate tissue classifiers. We also evaluated the limits of tissue classification by downsampling the peptide data to simulate situations where less data is available, such as clinical biopsies, laser capture microdissection or potentially single-cell proteomics. Our findings reveal the strong potential for utilizing proteomics data to build robust tissue classifiers, which has many prospective clinical applications for evaluating the applicability of model clinical systems.

Keywords: human variability, machine learning, classification, bioinformatics, data mining, data reuse

Introduction

Gene expression and the resultant interactions and activity of proteins define the cellular state. The characteristic set of proteins expressed in any specific set of cells governs its physical structure and biochemical functions. Tissues in the human body have a variety of cell types which work together to perform an integrated function, and the differences between tissues is the result of different protein expression patterns. When attempting to describe the molecular composition of a living system, measurements can be made at many different levels of granularity: whole organ, tissue, sub-tissue cellular clusters, or even single cells. It is important to build a knowledgebase at multiple levels of granularity, as it allows us to understand both the individual role of cell types and the emergent property of the whole organ. A proteomic characterization of these different levels has previously been explored for both liver and lung[1-4]. Gaining a deep understanding of tissue-specific expression is essential to understanding human health and disease, and is being explored widely at the level of transcription[5,6].

Although the proteome variability between specific tissues or organs has been reported for humans[7], non-human primates[8] and mice[9,10], there has not yet been an explicit exploration of the inter-individual variability of protein expression in the context of tissue comparisons. Thus, we know that each tissue has a different characteristic protein landscape, but it is unknown how much natural variability exists between individuals in a population. Being able to characterize the inter-individual variability is essential, not only for delineating between health and disease, but also for properly interpreting bioengineering and synthetic biology applications. For example, iPSC cells can be used to generate tissue-like cell populations through human-directed differentiation[11]. Are these synthetic constructs representative of a natural cell-type or tissue, especially in the context of inter-individual variability? Similarly, cell lines are often used as a proxy for their natural tissue when trying to understand the response to perturbation[12]. However,

since response is highly dependent on the cellular state, it is essential to know whether a cell line's proteome is representative, i.e. whether it fits within the nominal proteome variability of the target tissue as seen in individuals of a population.

Therefore, we performed a peptide-centric analysis of proteomics data from our historical archive by comparing data of up to 30 human samples from each of 9 tissues. Previously, re-analysis of historical datasets has been useful for insights[13-18]. Here we discovered significant inter-individual variability, but were able to train machine learning classifiers to accurately predict sample origin, i.e. what tissue the sample came from. Even for samples with highly similar expected protein compositions such as blood plasma, serum and cerebrospinal fluid, most classifier algorithms could predict the correct sample origin. As clinical applications like laser capture microdissection, imaging proteomics and single cell proteomics are likely to result in low proteome coverage, it is important to understand the practical limits and utility of sample source classifiers. We therefore tested the minimal number of peptides needed to produce an accurate classifier. Surprisingly, the proteome abundances were sufficiently distinct to successfully classify tissues with as few as 140 peptides. Finally, we compared the proteome of an immortalized cell line relative to the tissue of origin. For the cell line examined, the proteome is a poor match, underscoring the need to carefully vet model systems for their relevance[19].

Methods

Mass Spectrometry Data

Proteomics data were retrospectively gathered for nine human tissues: Blood Plasma, Blood Serum, Cerebrospinal Fluid (CSF), Liver, Monocyte, Ovary, Pancreas, Substantia Nigra, and

Temporal Lobe. The samples collected represent a mix of healthy and diseased tissue. Supplemental Table 1 gives a brief description of the data associated with each tissue and dataset. All of the original sample and data acquisition was done in accordance with the PNNL Institutional Review Board; all human subjects gave informed consent under their respective studies and are de-identified prior to analysis. Sample preparation methods used to generate peptide solutions prior to LC-MS/MS data acquisition may vary slightly between datasets, however, they are all based on a tissue homogenization, urea digest and C18 cleanup as described previously[20,21]. Blood-based samples did include protein depletion of the most highly abundant protein to improve dynamic range[22]. None of the samples used isotopic labeling or fractionation. The mass spectrometry data were produced by high-res Thermo Velos Orbitrap or Thermo QExactive instruments (San Jose, CA).

Peptide Identification and Quantification

All datasets were reanalyzed in a uniform manner, using the same protein sequence database and spectral identification algorithm. MSGF+ is a peptide/spectrum matching algorithm that identifies the best candidate peptide in the supplied database. The human protein sequence database was downloaded from Uniprot on April 12, 2017 and was supplemented with common contaminant proteins like trypsin and keratin. Mass spectra were identified using MS-GF+ (v2018.04.09)[23] with the following parameters: PrecursorMassTolerance, 20.0 ppm; fixed modification, alkylation of cysteines using carbamidomethyl; dynamic modification, oxidation of methionine; maximum modifications per peptide, 3; IsotopeError, -1,1; TargetDecoyAnalysis, true; FragmentationMethod, as written in the spectrum; Instrument, LCQ/LTQ (the parameter is inclusive for data generated for Velos and Q-Exactive instruments); Enzyme, PartTryp; NumTolerableTermini, 1;MinPeptideLength, 6; MaxPeptideLength, 50; and NumMatchesPerSpec, 1. MSGF+ output (in .mzID format) was converted to a tab-text format

where each row describes the best peptide for an MS/MS spectrum. All peptide/spectrum matches were filtered for quality using q-value ≤ 0.01. We next used MASIC[24] to generate selected ion chromatograms and chromatographic peak statistics for every precursor ion chosen for fragmentation. This chromatogram was used to calculate an MS1 intensity abundance value. The MASIC output was merged with MS-GF+ results by aligning MS2 scan number. This yields MS1 peak area quantitation values for all confidently identified peptides in the samples. For the purposes of classifier building, we worked exclusively with the peptide data and did not do any protein inference or protein level quantification.

Classifier Building and Testing

Using the datasets and identifications described above, we created classifiers to predict the tissue of origin. Classifiers were built with the Scikit-Learn machine learning library for Python (http://scikit-learn.org), which simplifies the interface to multiple machine learning methodologies. Below we describe the main steps used to build the classifiers on proteomics data. Full details of the method are available in the open source iPython notebook used to create the classifier, found on https://github.com/PNNL-Comp-Mass-Spec/Tissue_Classification.

As a first step, all identified and quantified peptides across all samples and all tissue types were loaded into a single dataframe and preprocessed via log2 transformation and median normalization to help counteract distribution skew and reduce the impact of outliers. Additionally, we filtered out peptides that were observed in fewer than five individuals within a single tissue. Since most Scikit-Learn classification algorithms do not handle missing values, these were imputed as half of the minimum observed abundance value. The resulting data table is called FullPeptideQuant and can be found in the GitHub repository with along with all the

code required to generate it. This table contains expression values for 68,623 peptides across the 253 individual datasets. The data from this table was randomly divided into training and testing datasets, containing 70% and 30% of the data respectively, and maintaining an even ratio of samples from each tissue.

We trained a classifier to predict the tissue of a given sample using a wide variety of algorithms: Logistic Regression, Multinomial Naïve Bayes, Random Forest, C-Support Vector Classification with linear kernel (SVC), Gradient Boosting, Gaussian Naïve Bayes, and K-Nearest Neighbors (KNN). Since classification algorithms generally benefit from standardization of data, all data was standardized using Scikit-Learn's StandardScaler method prior to being fed to the classifiers. This scaled the values of each feature (peptide) to become normally distributed, with a mean of 0 and standard deviation of 1. Because Multinomial Naive Bayes cannot handle negative values, data was translated to values between 0 and 1 via the MinMaxScaler method before being fed into this classifier. Classifier performance was measured by an accuracy score, obtained by dividing the number of correct predictions by the total number of predictions made. In the training phase, we estimated the performance of a classifier with 100 rounds of cross-validation. In each iteration, the training data was divided into two random subsets via a 70/30% split, preserving the relative number of samples per tissue. Then the classifier was trained on the larger subset and validated on the smaller subset. This process was repeated 100 times and all resulting accuracy scores were averaged. Each split of the data during the 100 rounds was completely independent.

We further refined the models by using feature selection methods and optimizing the models' hyper-parameters. Feature selection utilized combinations of the following three methods: SelectPercentile, PCA, and Recursive Feature Elimination (RFE). The SelectPercentile method removed peptides which did not vary significantly between tissue types, keeping only the

highest scoring percentage of features based on ANOVA F-value between tissues. We kept the top 25% of peptides, a procedure referred to in the results as 25P. The PCA method transformed the peptide abundance tables into a small set of principal components, where the number of components is equal to the number of training samples. RFE recursively considered smaller and smaller subsets of these components, determining the best features to keep and removing the rest. Coupled with these feature selection methods, we also optimized each machine learning methods' parameter space with ScikitLearn's GridSearch, which performs an exhaustive search over a user-specified set of possible parameter values to find the combination resulting in the highest cross-validation score. Once each classification model was finalized using the training data, they were used to predict tissue types for all samples in the test set.

Minimal classifiers

When training classifiers using only a subset of the data (results section Minimal Classifiers), we used the exact same train/test routines as outlined above. The only difference was that a fraction of the peptide dataframe was removed to simulate having less proteomics data. We used the following method to remove peptide data. For each tissue, we ranked the observed peptides by their average abundance across the training samples for that tissue. Only the peptides in the top percentile of average abundance were kept; this was done with each of the following percentile values: 0.0125, 0.0313, 0.0625, 0.125, 0.25, 0.5, 1, 2.5, 5, 10, 25, 50, 75, 90, and 100. In a practical sense, this removed each tissue's lowest observed peptides. After each round of peptide selection, classifiers were built using the reduced data frame. Peptides removed from the training set were then removed from the test set, and accuracy scores for the resulting predictions were recorded.

Data and Software Availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[25] partner repository with the dataset identifier PXD010271 and DOI:10.6019/PXD010271. Software used to execute the study and prepare the figures in the manuscript are available at our group's GitHub repository: https://github.com/PNNL-Comp-Mass-Spec/Tissue_Classification

Results

Inter individual variability in human tissue expression.

To understand the variability of protein expression between different individuals, we collected proteomic mass spectrometry data on whole tissue samples from different human subjects. Data represent nine tissues from ~30 subjects per tissue and up to three independent study sets (Figure 1). This cohort was created from a retrospective re-analysis of data from multiple experiments and projects performed in our lab over the last few years. Therefore, we expect there to be some batch effects, however, we believe this to be minimal in comparison to the differences between tissues. Moreover, mixing data from different projects represents a meaningful and realistic stress test to understanding variability. Our specific goal was to understand both the variability between individuals and also the differences between tissues. Of interest to us was whether samples from the same tissues have expression patterns that are reproducible and distinct when compared to the inter-individual variability.

Raw mass spectrometry data files were analyzed using MS-GF+ (PSM q-value cutoff 0.01); MS1 peak area intensities were used for quantitation via the MASIC program (see Methods). To simplify data processing and avoid the complications of protein inference, the results presented here are exclusively on peptide abundances.

Peptides displayed a high level of variation across tissues, with many peptides present in only a subset of the nine tissues. Figure 2 shows the variability of four peptides across the tissues and sample cohort. Two of the peptides display tissue-specific expression, a common situation as 40% of all peptides were only observed in a single tissue. Peptide DQTVSDNELQEMSNQGSK (clusterin, Uniprot ID P10909) is an example of a peptide with fairly consistent expression across different tissues; the only significant difference is the characteristic lower expression in the temporal lobe. 85.5% of all peptides showed a significant variation across tissues (ANOVA $P<0.01$, Bonferroni corrected), with most of these due to characteristic presence/absence observations between tissues.

To show the variability on a global scale, all quantified peptides were used as input to t-distributed Stochastic Neighbor Embedding (tSNE) to visualize the data (Figure 3). The plot shows that individual samples of a tissue generally cluster together, indicating that machine learning classification algorithms may be able to distinguish tissues based on peptide abundances. We next sought to understand the effect that disease may have on peptide expression, and how the magnitude of such a perturbation relates to the level expression differences between tissues. All quantified peptides for tissues having a mix of healthy and diseased datasets were again visualized with tSNE (Figure 4). For Substantia Nigra and CSF, the healthy and diseased samples are grouped together with no clear distinction. For blood plasma, ovary, pancreas, and liver there are potentially separate clusters for healthy and diseased samples, though they still group relatively close to each other.

Generalized classifiers for human tissues

Given the strong relationship shown between different samples from the same tissue, we wanted to explore whether proteomic mass spectrometry data can be used to uniquely identify different human tissues[26]. That is, given only the identified peptides and their abundances for a human tissue sample, can we accurately predict the sample origin? Ideally, such a classifier would be successful across a broad spectrum of samples, including diseased or healthy tissue, data derived from different mass spectrometry instruments, or samples prepared using different methods.

In order to build robust classifiers, we randomly split our data into separate training and testing datasets, and performed extensive cross-validation during training (see Methods). Using the Scikit-Learn machine learning library for Python, we created a classifier using seven different machine learning methods. Most classifiers achieved over 95% cross-validation accuracy (Figure 5). In general, when the classifiers misannotated tissues they tended to either confuse blood plasma, blood serum, and CSF, which are highly similar biofluids, or mistake substantia nigra and temporal lobe.

We attempted to optimize the models via feature selection and parameter tuning (see Methods). Several methods were used in combination and compared to the original model. The finalized models were subsequently used to predict the source tissue of the testing data. The most successful optimization method was with SelectPercentile down-selected peptide features and GridSearch parameter tuning (blue bar in Figure 5). With these two optimizations, all methods achieved 100% accuracy, except for Gaussian Naive Bayes. The two other combinations were a complex combination (see Methods). First, we use SelectPercentile to remove peptide features with low variance; second, perform a PCA to transform the peptide features into

principle components; third, perform recursive feature elimination on the principle components; optionally we also performed GridSearch. This is abbreviated in Figure 5 as 25P-PCA-RFE, and 25P-PCA-RFE-GS. With these last two combinations, the results were sometimes less accurate than the original models. This is likely due to overfitting on the training set. Indeed, given the relatively high number of peptide features compared to the number of samples, overfitting is likely in this style of hyper-parameter optimization. Thus, although feature selection and hyper-parameterization can be employed, we generally found that simple optimizations are more robust than complex multi-step processes.

Minimal classifiers

In practice, proteomics experiments may encounter situations where fewer peptides and proteins are identified than what was used to train the models above. This is prominent in targeted proteomics and extremely small samples from clinical biopsies, cell sorting or single cells[4,27-30]. Therefore, we next sought to understand the behavior of classification methods with less data, and to identify the minimal amount of data required to produce an accurate classifier.

To simulate samples with fewer peptide identifications, we iteratively downsampled our training and test sets by incrementally removing peptides and rebuilding classifier models. In each round of downsampling, peptides with the lowest abundances for each tissue were removed (as described in Methods). Training and testing the models was performed as described above; the only difference was the amount of peptide data available to the models. Performance on the testing data initially remained strong, and accuracy was fairly constant for most models down to about 5% of the originally observed peptides (Figure 6). At this level, ~5500 peptides, many models retained near-perfect accuracy. When training using only 5% to 2.5% of the data (~5,500 to ~2,800 peptides), four models remained above 90% accuracy in tissue classification.

The models created using Logistic Regression, Support Vector Classification, Random Forest, and K-nearest Neighbors all performed exceptionally well, and therefore have strong potential for future clinical applications, where LCM and cell sorting, or single cell proteomics currently achieve peptide identification rates in this range. Here, the most common error made was confusing temporal lobe for substantia nigra. At the extreme, three models maintained at least 90% accuracy down to 139 peptides, after which there was a significant loss of performance. When looking at current practical limits for SRM (~200-500 peptides), the best performing algorithm is SVC, which has ~94% accuracy. When trained on 284 peptides, the only mistake made by SVC was misclassifying about half the temporal lobe samples as substantia nigra. We note that for this level of accuracy, the classifier algorithms need peptide abundance data, as the classifiers were less successful when using only presence/absence information on observed peptide lists (Supplemental Figure 1).

We further investigated the 139 peptides which were descriptive of a sample's origin (Supplemental Figure 2). In this set are several expected tissue-specific proteins. Tissue samples from the substantia nigra and the temporal lobe of the brain both contained high amounts of the myelin basic protein (MBP_HUMAN), the glial fibrillary acidic protein (GFAP_HUMAN), and the myelin proteolipid protein (MYPR_HUMAN), well known structural components of the central nervous system and highly expressed in brain tissues. Similarly, tissue samples from the liver had two characteristic proteins: carbamoyl-phosphate synthase (CPSM_HUMAN) and fatty acid-binding protein (FABPL_HUMAN). Carbamoyl-phosphate synthase is part of the urea cycle and is known to be strongly expressed in liver tissue[31]; the fatty acid-binding protein plays a major role in the liver's uptake of cholesterol[32]. In addition to peptides from proteins which have a binary-like prediction capability, we note that the 139 peptides contain also come from common and highly abundant proteins such as actin, hemoglobin, tubulin, glyceraldehyde-3-phosphate dehydrogenase and histone proteins. The

peptides of which implicate specific paralogs and isoforms that can be used in combination to predict the source tissue of a proteomics sample.

How cohort size affects classifier accuracy

One simple method to improve the performance of machine learning classifiers is by having a larger number of training samples. However, it is both time- and cost- intensive to produce a large number of human samples and acquire the proteomics data. Therefore, it would be helpful to have an idea of how many samples is necessary to improve classifier accuracy. To test how well this would work with our proteomics data, we investigated how increasing the number of samples would improve the classification accuracy for different kinds of blood samples.

While most classifiers correctly distinguished plasma and serum, there was still some confusion because the expression profiles are very similar. Therefore, we gathered 30 additional datasets for both blood plasma and serum. Then we combined these with the plasma and serum training and test data to create a data frame containing 119 datasets: 59 plasma and 60 serum. To challenge the classification task, the data was aggressively filtered by removing half of the lowest observed peptides, resulting in a data frame with ~5000 peptides. Using this new data frame we built classifiers, incrementally increasing the number of samples in the training set (Figure 7). As expected, increasing training set size improved classifier performance on the test set. A significant improvement was seen between 5 and 15 samples, after which the rate of improvement was steady and did not yet show signs of diminishing returns. As was seen with the models used on minimal peptide sets, Logistic Regression and SVC were both very robust methods and achieved close to 90% accuracy.

Utility of cell lines as a proxy for tissue-based research

Cell lines are often used in biological studies as models for human cells. In recent years, a proliferation of different growth techniques has attempted to simulate a more accurate model and recapitulate organ/tissue function, including 3D culture, organoids, organs-on-a-chip, etc. Debate over whether these *ex vivo* cultures accurately represent their *in vivo* counterparts in their proteome[19] led us to explore whether cell lines could be used in place of their parent tissue in a classification setting. Such an application would be useful in cases where samples of human tissue are limited and cell lines are utilized for biomedical research.

Ten proteomics datasets for Huh-7 liver cell line cultures[33] were visualized via tSNE in the context of the other data (Figure 8). This cell line is a hepatocyte-derived carcinoma cell line, and therefore, might be assumed to be most similar to the liver tissue in our dataset. These liver cell line samples clustered together nearest substantia nigra and blood serum, and do not appear to cluster at all near the liver tissue. This difference is evident when looking at the most abundant peptides of the liver tissue compared to the Huh-7 cell line. Liver tissue has highly expressed proteins relating to the liver function in metabolism (carbamoyl-phosphate synthase, fatty acid binding protein, alcohol dehydrogenase, and fructose bisphosphate aldolase), while the cell line contains highly abundant proteins relating to cell cycle and growth (elongation factor, histones, tubulin, protein folding). These differently abundant proteins point to an expected diverged phenotype between the cell line and the primary tissue. Correspondingly, most classifiers did not classify the cell line as ‘liver’. Random Forest classified one sample as liver and the other 9 samples as either: blood plasma, ovary, or substantia nigra. All other classifiers tested failed to classify any of the Huh7 cell line samples as liver. These results corroborate previous findings that the proteome of cell lines varies significantly from their parent tissue[34,35].

Conclusion

As omics technologies enable a more expansive characterization of human tissues, especially at the scale of populations, it is important to understand the inherent variability of the proteome within a population. Our results show that proteomics data from a variety of individuals, instruments, and sample prep methods can be utilized together to create a robust tissue classifier capable of accurately predicting the origin of a sample. As our dataset included both healthy and diseased tissues, the classifiers were able to learn the common baseline for a tissue, regardless of health. For the tissues and diseases examined herein, it appears that diseases did not cause a significant proteomic change, relative to the difference between tissues. Although it was not explored in this manuscript, we believe that classifiers could be built to not only identify the tissue of origin, but also the relative health of a tissue. For these, it would be essential to gather a much larger sample set containing a variety of disease states per tissue.

The data used in this project was only from unlabeled and unfractionated bottom-up proteomics. It is important to note that a wide variety of other quantitative proteomics techniques exist and can be used in similar classification exercises. Although isobaric labeling techniques are broadly used, it is important to consider how labeling affects the ability to do meta-analyses across different experiments as shown here. Specifically, isobaric labeling is most often done in relation to a reference, creating a ratio based quantitative metric. Thus, merging data from different

experiments that utilized different references is statistically complicated. Label-free samples, on the other hand, are much simpler to merge across datasets and experiments.

In addition to a simple classifier of tissue origin, we explored two applications of classifiers in specific clinical settings: small sample sizes and the relevance of *in vitro* cell lines. Although much of academic research has access to large sample quantities, a growing number of clinical applications are likely to have only limited sample volumes, e.g. biopsies or laser-capture microdissection. Very recent work on single human cells is another exceptionally relevant and exciting area of science where classification would be valuable, yet where the number of identified peptides is low. To simulate the performance of a classifier in these settings, we downsampled the rich proteomics data and observed strong classifier performance even with <140 peptides - a range that is easily achievable by today's proteomic instrumentation platforms on even single cells.

In biomedical and pharmaceutical research, a primary need is to understand how cells/tissues/individuals will respond to various therapies. In recent years, a significant investment has been made to catalog cellular response to a wide variety of small molecule perturbations[36,37]. Yet, in order to understand how these massive libraries of cellular response relate to clinical applications, it is essential to understand how cell lines relate to human tissue. Using our classification algorithm, we confirm prior work on the level of transcriptomics that cell lines can bear little resemblance to their parent tissue[34] and more representative *in vitro* models should be pursued for cellular and molecular phenotypic studies.

SUPPORTING INFORMATION:

The following supporting information is available free of charge at ACS website http://pubs.acs.org

Supplemental Figure 1 - Classifier performance when ignoring peptide abundance. Using the same methods for building and testing classifiers, we modified the input data to contain only presence/absence of a peptide in a given tissue. Peptides are also iteratively downsampled, as was done in Figure 6. Performance of classifiers is less accurate when excluding peptide abundance.

Supplemental Figure 2 - Peptide abundance of the 139 peptides used to create minimal classifiers according to tissue. Each of the peptides is plotted with the values across all samples (before missing value imputation).

Supplemental Table 1 - Meta-data of the samples used in this classification project.

Acknowledgments

This work was supported by the IARPA FunGCAT program. The funders had no role in the design or interpretation of the experiment. Battelle operates the Pacific Northwest National Laboratory for the U.S. Department of Energy under contract DE-AC05-76RLO01830.

Author Contributions

I.K.K., G.C., S.H.P, J.N.A designed and executed the study. I.K.K., G.C., S.O.P., and J-Y.L. wrote software, processed and analyzed data. I.K.K and S.H.P wrote the manuscript with input from all authors.

The authors declare no conflicts of interests.

Figure Legends

Figure 1 – Samples used in this study. Tissue samples from nine different sites were obtained and used to acquire proteomic mass spectrometry data. In most cases, samples come from thirty different individuals, and represent a mix of healthy and diseased states.

Figure 2 - Examples of peptide variability across tissues. For four different peptides, non-imputed abundance values from the normalized data frame of train and test data were grouped by tissue. Peptides plotted: VLILGSGGLSIGQAGEFDYSGSQAVK (carbamoyl-phosphate synthase, Uniprot ID P31327), VLDSGAPIKIPVGPETLGR (ATP synthase subunit beta, mitochondrial, Uniprot ID P06576), DQTVSDNELQEMSNQGSK (clusterin, Uniprot ID P10909), TYFPHFDLSHGSAQVK (hemoglobin subunit alpha, Uniprot ID P69905). The first two peptides show presence/absence between tissues. The third and fourth peptides have the least and most variable abundance among peptides observed in at least 5 samples of each tissue.

Figure 3 - Grouping of individual samples. The peptide abundance dataframe was used as input to t-SNE. Samples from the same tissue generally cluster together in the plot.

Figure 4 - Grouping of healthy and diseased samples. Only tissues for which both healthy and diseased datasets are present were used as input for t-SNE. Data from diseased samples are marked by open circles; healthy samples are marked with closed circles.

Figure 5 - Test set accuracy following different optimizations. Four different methods for parameter tuning and feature selection were employed in combination to attempt to improve model performance. These results were plotted alongside the original models with no additional feature selection or parameter tuning. Gaussian Naive Bayes does not have parameters to tune,

so no grid search was performed following feature selection, i.e. its models for 25P-PCA-RFE and 25P-PCA-RFE-GS are identical.

Figure 6 - Classifier performance on limited peptide data. Seven machine learning algorithms were used to build tissue classification models on peptide data. Models were created for various iterations of downsampled data (see Methods). Numeric annotations represent the total number of peptides used in training at each step.

Figure 7- Serum and Plasma learning curves. For each training sample size n, classification models were trained to differentiate between blood serum and plasma. The reported accuracy score is an average of all 100 random trials.

Figure 8 - tSNE plot with liver cell line samples. Individual samples were clustered with t-SNE, as in Figure 3. Ten Huh-7 cell line samples had the tightest clustering of any group; all ten samples are so close together that they appear as a single gray dot in the t-SNE plot. Moreover, they do not appear to cluster near liver tissue samples.

References

1 Clair, G. *et al.* Spatially-Resolved Proteomics: Rapid Quantitative Analysis of Laser Capture Microdissected Alveolar Tissue Samples. *Scientific reports* **6**, 39223, doi:10.1038/srep39223 (2016).

2 Azimifar, S. B., Nagaraj, N., Cox, J. & Mann, M. Cell-type-resolved quantitative proteomics of murine liver. *Cell metabolism* **20**, 1076-1087, doi:10.1016/j.cmet.2014.11.002 (2014).

3 Ardini-Poleske, M. E. *et al.* LungMAP: The Molecular Atlas of Lung Development Program. *American journal of physiology. Lung cellular and molecular physiology* **313**, L733-l740, doi:10.1152/ajplung.00139.2017 (2017).

4 Zhu, Y. *et al.* Proteomic Analysis of Single Mammalian Cells Enabled by Microfluidic Nanodroplet Sample Preparation and Ultrasensitive NanoLC-MS. *Angewandte Chemie (International ed. in English)*, doi:10.1002/anie.201802843 (2018).

5 Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)* **348**, 648-660, doi:10.1126/science.1262110 (2015).

6 Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).

7 Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582-587, doi:10.1038/nature13319 (2014).

8 Lee, J. G. *et al.* A draft map of rhesus monkey tissue proteome for biomedical research. *PloS one* **10**, e0126243, doi:10.1371/journal.pone.0126243 (2015).

9 Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174-1189, doi:10.1016/j.cell.2010.12.001 (2010).

10 Geiger, T. *et al.* Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Molecular & cellular proteomics : MCP* **12**, 1709-1722, doi:10.1074/mcp.M112.024919 (2013).

11 Zhang, D. *et al.* Highly efficient differentiation of human ES cells and iPS cells into mature pancreatic insulin-producing cells. *Cell research* **19**, 429-438, doi:10.1038/cr.2009.28 (2009).

12 Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science (New York, N.Y.)* **313**, 1929-1935, doi:10.1126/science.1132939 (2006).

13 Callister, S. J. *et al.* Comparative bacterial proteomics: analysis of the core genome concept. *PloS one* **3**, e1542, doi:10.1371/journal.pone.0001542 (2008).

14 Payne, S. H. *et al.* The Pacific Northwest National Laboratory library of bacterial and archaeal proteomic biodiversity. *Scientific data* **2**, 150041, doi:10.1038/sdata.2015.41 (2015).

15 Robin, T., Bairoch, A., Muller, M., Lisacek, F. & Lane, L. Large-scale reanalysis of publicly available HeLa cell proteomics data in the context of the Human Proteome Project. *Journal of proteome research*, doi:10.1021/acs.jproteome.8b00392 (2018).

16 Martens, L. & Vizcaino, J. A. A Golden Age for Working with Public Proteomics Data. *Trends in biochemical sciences* **42**, 333-341, doi:10.1016/j.tibs.2017.01.001 (2017).

17 Shatsky, M. *et al.* Bacterial Interactomes: Interacting Protein Partners Share Similar Function and Are Validated in Independent Assays More Frequently Than Previously Reported. *Molecular & cellular proteomics : MCP* **15**, 1539-1555, doi:10.1074/mcp.M115.054692 (2016).

18 Matic, I., Ahel, I. & Hay, R. T. Reanalysis of phosphoproteomics data uncovers ADP-ribosylation sites. *Nature methods* **9**, 771-772, doi:10.1038/nmeth.2106 (2012).

19 Gillet, J. P. *et al.* Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 18708-18713, doi:10.1073/pnas.1111840108 (2011).

20 Piehowski, P. D. *et al.* Sources of technical variability in quantitative LC-MS proteomics: human brain tissue sample analysis. *Journal of proteome research* **12**, 2128-2137, doi:10.1021/pr301146m (2013).

21 Wang, S. *et al.* Quantitative proteomics identifies altered O-GlcNAcylation of structural, synaptic and memory-associated proteins in Alzheimer's disease. *The Journal of pathology* **243**, 78-88, doi:10.1002/path.4929 (2017).

22 Liu, T. *et al.* Evaluation of multiprotein immunoaffinity subtraction for plasma proteomics and candidate biomarker discovery using mass spectrometry. *Molecular & cellular proteomics : MCP* **5**, 2167-2174, doi:10.1074/mcp.T600039-MCP200 (2006).

23 Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature communications* **5**, 5277, doi:10.1038/ncomms6277 (2014).

24 Monroe, M. E., Shaw, J. L., Daly, D. S., Adkins, J. N. & Smith, R. D. MASIC: a software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC-MS(/MS) features. *Computational biology and chemistry* **32**, 215-217, doi:10.1016/j.compbiolchem.2008.02.006 (2008).

25 Vizcaino, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic acids research* **44**, D447-456, doi:10.1093/nar/gkv1145 (2016).

26 Dammeier, S. *et al.* Mass-Spectrometry-Based Proteomics Reveals Organ-Specific Expression Patterns To Be Used as Forensic Evidence. *Journal of proteome research* **15**, 182-192, doi:10.1021/acs.jproteome.5b00704 (2016).

27 Abbatiello, S. E. *et al.* Large-Scale Interlaboratory Study to Develop, Analytically Validate and Apply Highly Multiplexed, Quantitative Peptide Assays to Measure Cancer-Relevant Proteins in Plasma. *Molecular & cellular proteomics : MCP* **14**, 2357-2374, doi:10.1074/mcp.M114.047050 (2015).

28 Ippoliti, P. J. *et al.* Automated Microchromatography Enables Multiplexing of Immunoaffinity Enrichment of Peptides to Greater than 150 for Targeted MS-Based Assays. *Analytical chemistry* **88**, 7548-7555, doi:10.1021/acs.analchem.6b00946 (2016).

29 Zhu, Y. *et al.* Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells. **9**, 882, doi:10.1038/s41467-018-03367-w (2018).

30 Budnik, B., Levy, E., Harmange, G. & Slavov, N. Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *bioRxiv*, doi:10.1101/102681 (2018).

31 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science (New York, N.Y.)* **347**, 1260419, doi:10.1126/science.1260419 (2015).

32 Huang, H. *et al.* Human FABP1 T94A variant enhances cholesterol uptake. *Biochimica et biophysica acta* **1851**, 946-955, doi:10.1016/j.bbalip.2015.02.015 (2015).

33 Sainz, B., Jr., TenCate, V. & Uprichard, S. L. Three-dimensional Huh7 cell culture system for the study of Hepatitis C virus infection. *Virology journal* **6**, 103, doi:10.1186/1743-422x-6-103 (2009).

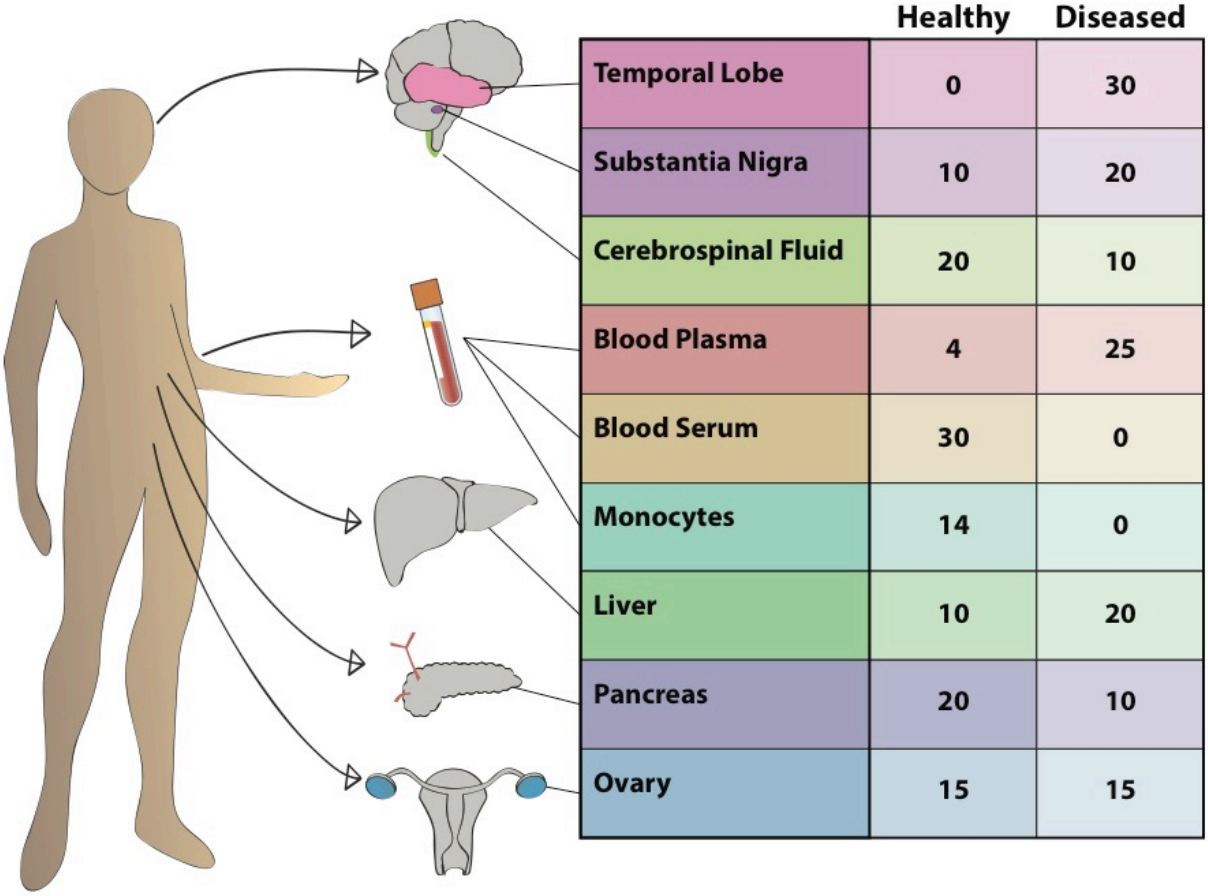
34 Pan, C., Kumar, C., Bohl, S., Klingmueller, U. & Mann, M. Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions. *Molecular & cellular proteomics : MCP* **8**, 443-450, doi:10.1074/mcp.M800258-MCP200 (2009).

35 Shi, J., Wang, X., Lyu, L., Jiang, H. & Zhu, H. J. Comparison of protein expression between human livers and the hepatic cell lines HepG2, Hep3B, and Huh7 using SWATH and MRM-HR proteomics: Focusing on drug-metabolizing enzymes. *Drug metabolism and pharmacokinetics* **33**, 133-140, doi:10.1016/j.dmpk.2018.03.003 (2018).

36 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

37 Keenan, A. B. *et al.* The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. *Cell systems* **6**, 13-24, doi:10.1016/j.cels.2017.11.001 (2018).

Figure 1



	Healthy	Diseased
Temporal Lobe	0	30
Substantia Nigra	10	20
Cerebrospinal Fluid	20	10
Blood Plasma	4	25
Blood Serum	30	0
Monocytes	14	0
Liver	10	20
Pancreas	20	10
Ovary	15	15

Figure 2

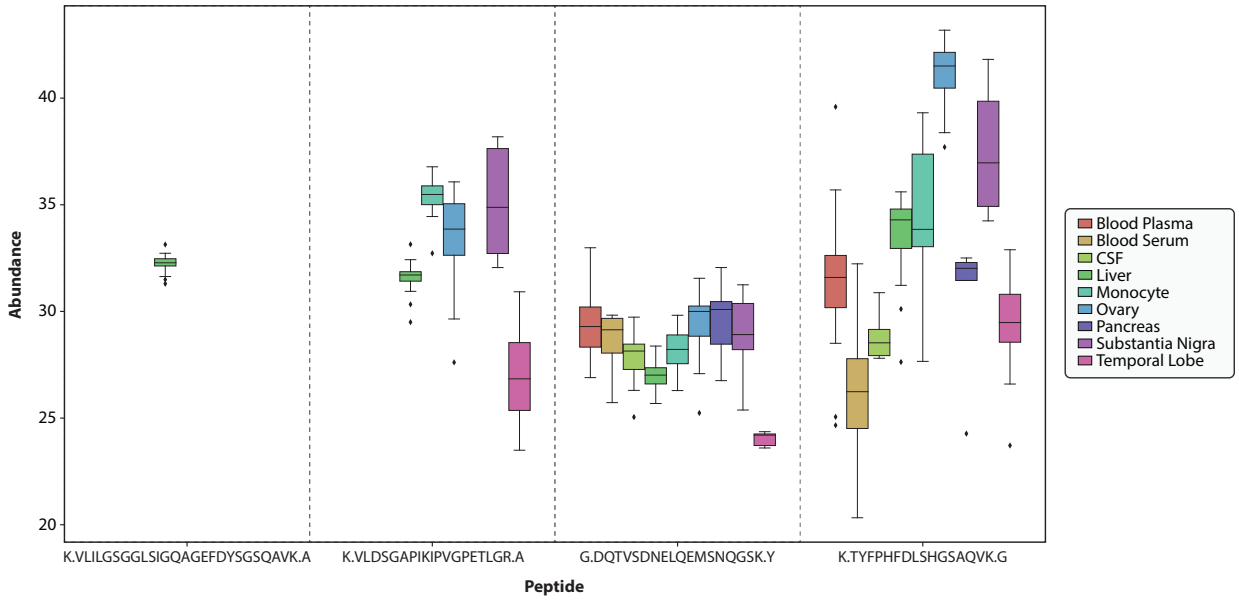

Abundance
40
35
30
25
20
K.VLILGSGGLSIGQAGEFDYSGSQAVK.A
K.VLDSGAPIKIPVGPETLGR.A
G.DQTVSDNELQEMSNQGSK.Y
K.TYFPHFDLSHGSAQVK.G
Peptide
Blood Plasma
Blood Serum
CSF
Liver
Monocyte
Ovary
Pancreas
Substantia Nigra
Temporal Lobe

Figure 3

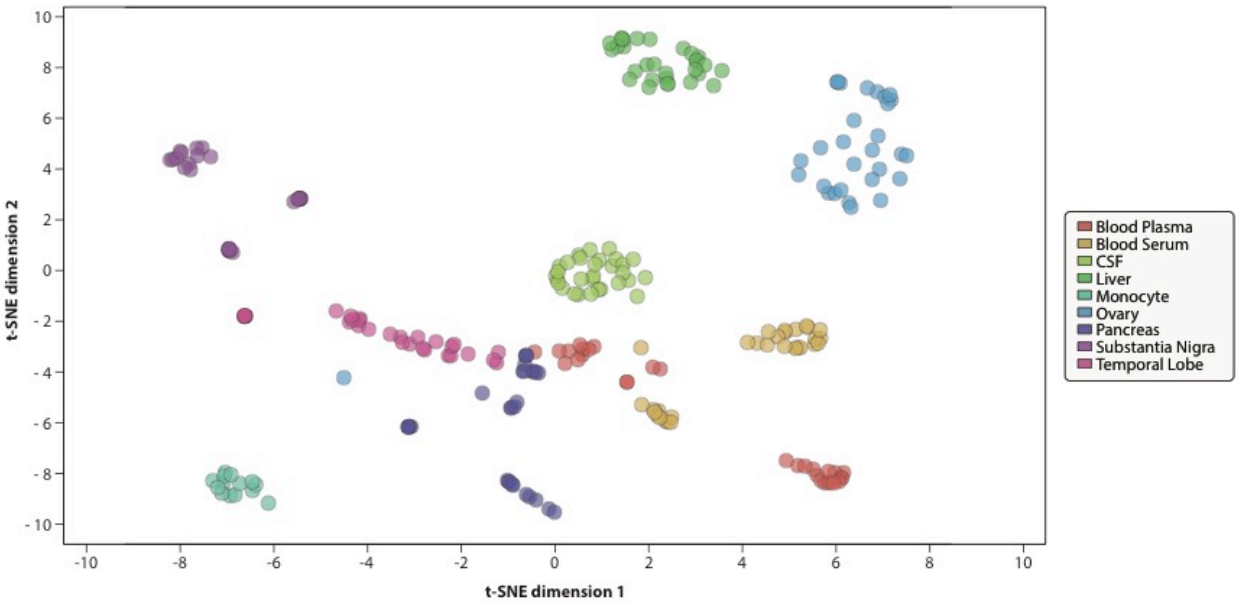

t-SNE dimension 2
t-SNE dimension 1
10
8
6
4
2
0
-2
-4
-6
-8
-10
-10
-8
-6
-4
-2
0
2
4
6
8
10
Blood Plasma
Blood Serum
CSF
Liver
Monocyte
Ovary
Pancreas
Substantia Nigra
Temporal Lobe

Figure 4

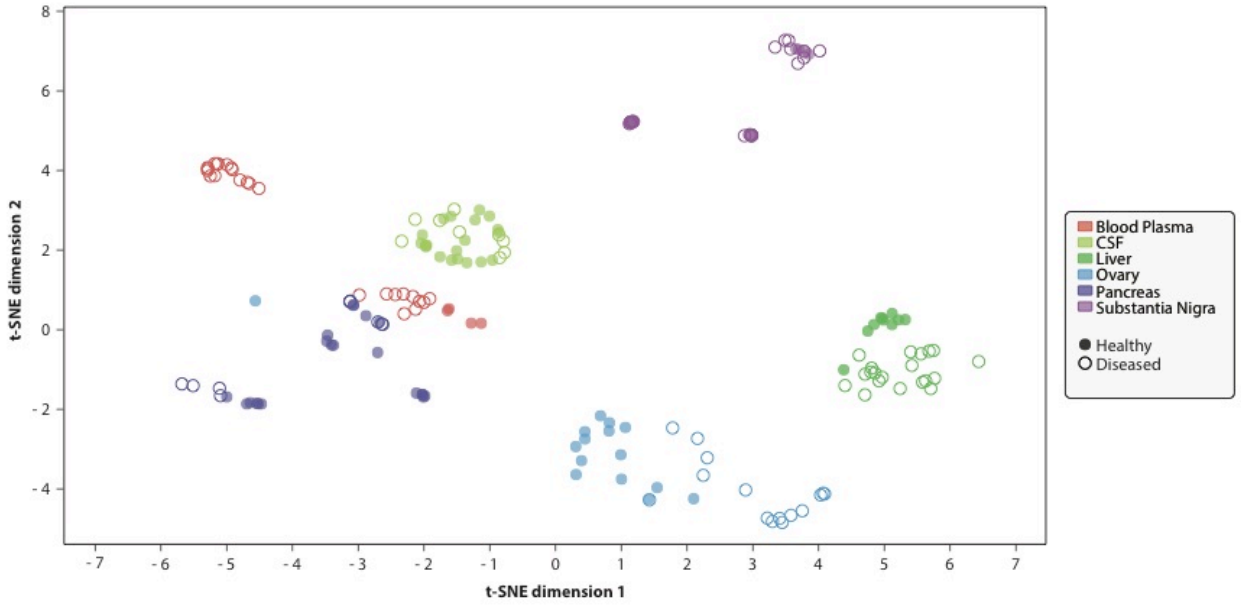

t-SNE dimension 2
t-SNE dimension 1
8
6
4
2
0
-2
-4
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
Blood Plasma
CSF
Liver
Ovary
Pancreas
Substantia Nigra
Healthy
Diseased

Figure 5

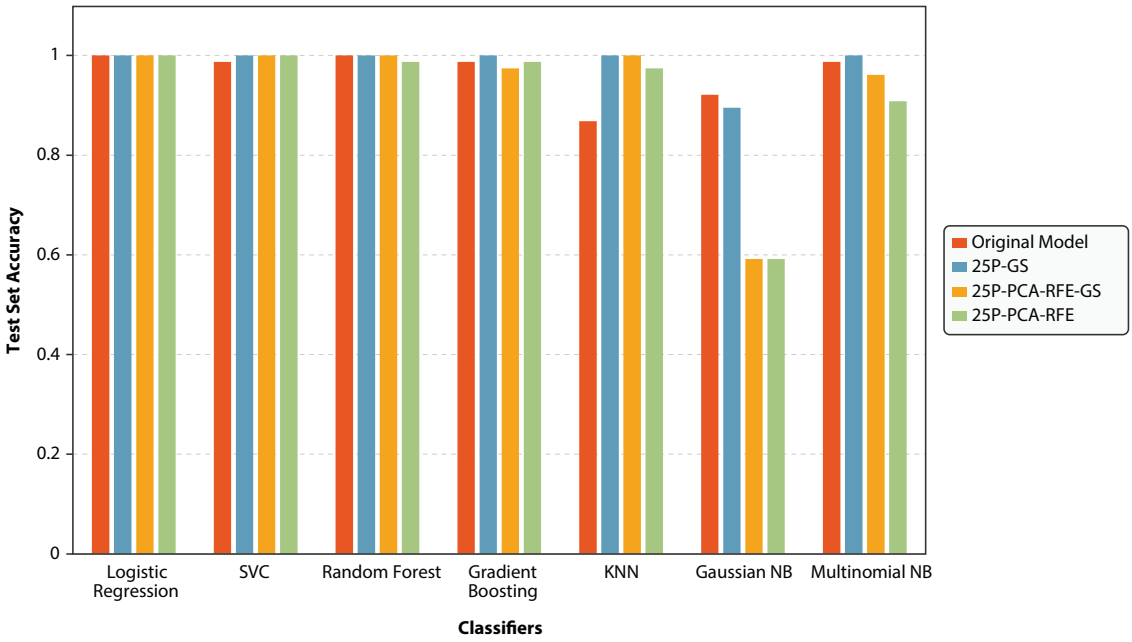

Test Set Accuracy
1
0.8
0.6
0.4
0.2
0
Logistic Regression
SVC
Random Forest
Gradient Boosting
KNN
Gaussian NB
Multinomial NB
Classifiers
Original Model
25P-GS
25P-PCA-RFE-GS
25P-PCA-RFE

Figure 6

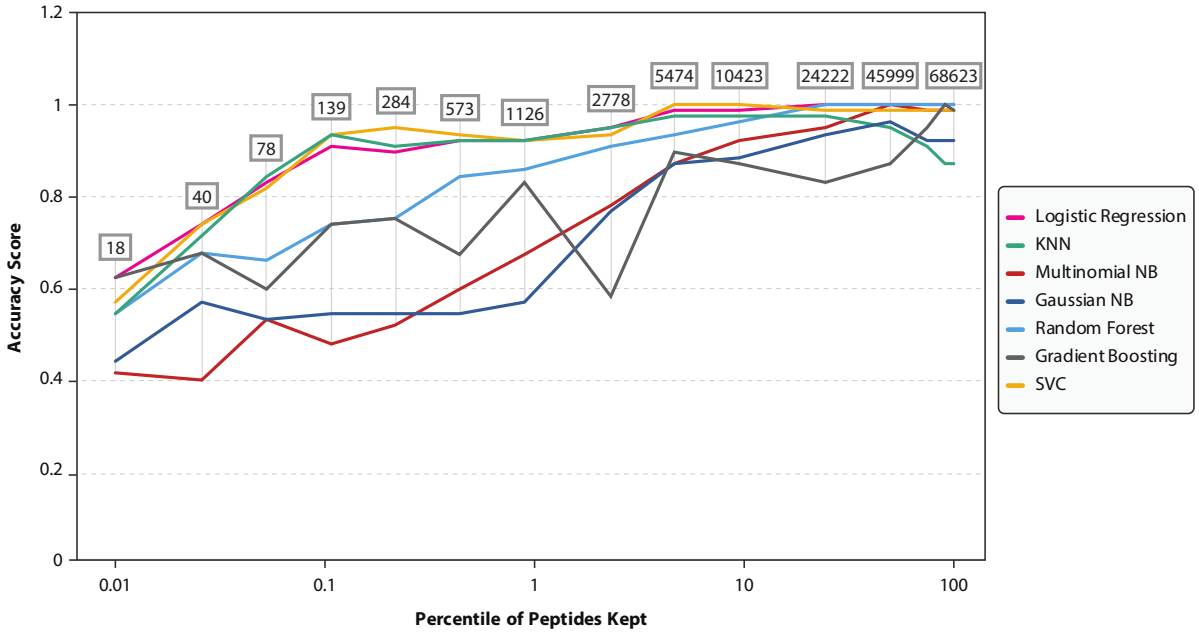

1.2
1
0.8
0.6
0.4
0.2
0
Accuracy Score
0.01
0.1
1
10
100
Percentile of Peptides Kept
18
40
78
139
284
573
1126
2778
5474
10423
24222
45999
68623
Logistic Regression
KNN
Multinomial NB
Gaussian NB
Random Forest
Gradient Boosting
SVC

Figure 7

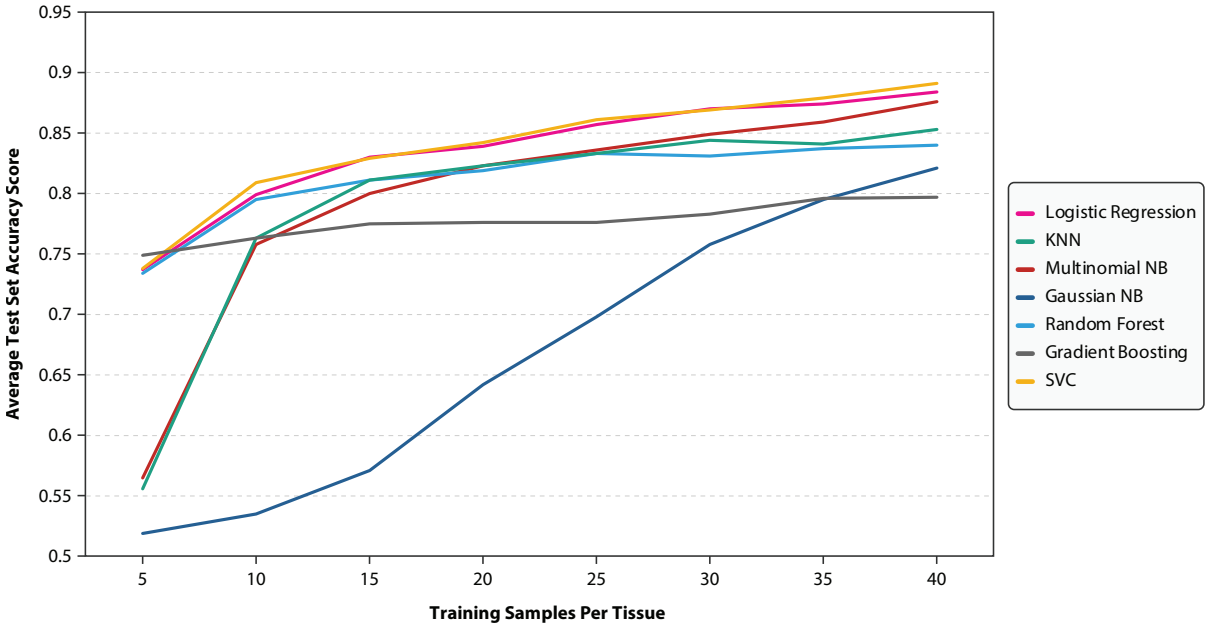
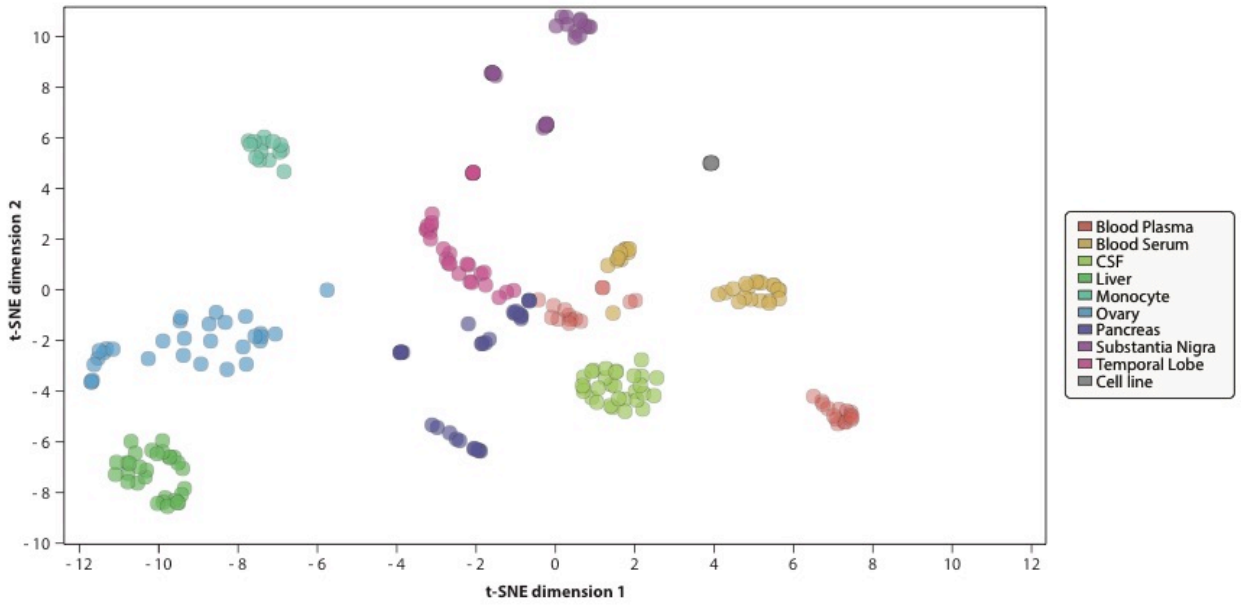

0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Average Test Set Accuracy Score
5
10
15
20
25
30
35
40
Training Samples Per Tissue
Logistic Regression
KNN
Multinomial NB
Gaussian NB
Random Forest
Gradient Boosting
SVC

Figure 8


t-SNE dimension 2
t-SNE dimension 1
10
8
6
4
2
0
-2
-4
-6
-8
-10
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
Blood Plasma
Blood Serum
CSF
Liver
Monocyte
Ovary
Pancreas
Substantia Nigra
Temporal Lobe
Cell line

For TOC Only