technical notes Journal of proteome research

Comparative Shotgun Proteomics Using Spectral Count Data and Quasi-Likelihood Modeling

Ming Li,[†] William Gray,[†] Haixia Zhang,[‡,§] Christine H. Chung,[||,⊥] Dean Billheimer,[#] Wendell G. Yarbrough,[||,▽] Daniel C. Liebler,[‡,§] Yu Shyr,[†] and Robbert J. C. Slebos*[,‡,||]

Jim Ayers Institute for Precancer Detection and Diagnosis, Departments of Biostatistics, Biochemistry, Cancer Biology, Medicine, and Otolaryngology, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, and Department of Agricultural and Biosystems Engineering, University of Arizona, Tuscon, Arizona 85721

Received May 26, 2010

Abstract: Shotgun proteomics provides the most powerful analytical platform for global inventory of complex proteomes using liquid chromatography−tandem mass spectrometry (LC−MS/MS) and allows a global analysis of protein changes. Nevertheless, sampling of complex proteomes by current shotgun proteomics platforms is incomplete, and this contributes to variability in assessment of peptide and protein inventories by spectral counting approaches. Thus, shotgun proteomics data pose challenges in comparing proteomes from different biological states. We developed an analysis strategy using quasi-likelihood Generalized Linear Modeling (GLM), included in a graphical interface software package (QuasiTel) that reads standard output from protein assemblies created by IDPicker, an HTML-based user interface to query shotgun proteomic data sets. This approach was compared to four other statistical analysis strategies: Student *t* test, Wilcoxon rank test, Fisher's Exact test, and Poisson-based GLM. We analyzed the performance of these tests to identify differences in protein levels based on spectral counts in a shotgun data set in which equimolar amounts of 48 human proteins were spiked at different levels into whole yeast lysates. Both GLM approaches and the Fisher Exact test performed adequately, each with their unique limitations. We subsequently compared the proteomes of normal tonsil epithelium and HNSCC using this approach and identified 86 proteins with differential spectral counts between normal tonsil epithelium and HNSCC. We selected 18 proteins from this comparison for verification of protein levels between the individual normal and tumor tissues using liquid chromatography−multiple reaction monitoring mass spectrometry (LC−MRM-MS). This analysis confirmed the magnitude and direction of the protein expression differences in all 6 proteins for which reliable data could be obtained. Our analysis demonstrates that shotgun proteomic data sets from different tissue phenotypes are sufficiently rich in quantitative information and that statistically significant differences in proteins spectral counts reflect the underlying biology of the samples.

Keywords: LC−MS/MS • shotgun proteomics • multiple reaction monitoring (MRM) • head and neck carcinoma • Generalized Linear Model • spectral counting

Introduction

Over the past decade, shotgun proteomics has emerged as the most powerful analytical platform to characterize complex proteomes.[1,2] By combining multidimensional liquid chromatography−tandem mass spectrometry (LC−MS/MS) with database search and protein assembly algorithms, shotgun proteomics platforms surpass other MS-based proteomics systems in number and diversity of proteins identified and in dynamic range for detection. Nevertheless, shotgun proteomics using LC−MS/MS is essentially a sampling technique, in which probability of detection is a function of protein abundance and quantitation is assessed by counting the numbers of spectra that map to identified proteins.[3−5] However, random sampling of medium to low abundance proteins in shotgun analyses means that multiple replicate analyses are needed to establish the composition of complex proteomes.[6,7] Because shotgun analyses can represent complex proteomes in considerable depth, a key question is whether comparison of shotgun proteome inventories can reveal molecular characteristics of biologically distinct phenotypes.

This is more than a purely academic question, as high-throughput gene expression studies using microarrays have allowed a global view of molecular processes in complex biological systems.[8] Microarray-based gene expression studies have also revealed molecular differences between clinically distinct cancer phenotypes. For example, pioneering studies identified gene expression patterns that allowed molecular classification of lymphomas and breast cancers.[9,10] An extension of this approach led to the identification of gene expression signatures that predict response of lymph node-negative

* To whom correspondence should be addressed. Department of Cancer Biology, Vanderbilt University School of Medicine, U1213A Medical Research Building III, 465 21st Avenue South, Nashville, TN 37232-8575; Phone 615 322-3063, Fax 615 343-8372; e-mail: r.slebos@vanderbilt.edu.

[†] Department of Biostatistics, Vanderbilt University School of Medicine.

[‡] Jim Ayers Institute for Precancer Detection and Diagnosis, Vanderbilt University School of Medicine.

[§] Department of Biochemistry, Vanderbilt University School of Medicine.

[||] Department of Cancer Biology, Vanderbilt University School of Medicine.

[⊥] Department of Medicine, Vanderbilt University School of Medicine.

[#] University of Arizona.

[▽] Department of Otolaryngology, Vanderbilt University School of Medicine.

10.1021/pr100527g

Published on Web 06/29/2010

breast cancers to chemotherapy and endocrine therapy.[11,12] This latter approach has emerged as a prototype of a new generation of *in vitro* diagnostics for therapeutic decision-making in cancer.

However, gene expression profiles do not always correspond to proteome changes and do not capture many functional features of proteomes. Shotgun proteome analyses on standardized analytical platforms offer the prospect of characterizing collections of proteins that distinguish clinically relevant phenotypes. When applied to well-characterized biospecimens, proteome differences represent candidates for evaluation in a biomarker development pipeline.[13] One of the obstacles of employing shotgun proteomics in this context is the challenge of performing statistical comparisons of shotgun data sets. The characteristics of spectral count data and the stochastic representation of low abundance proteins present interesting challenges in statistical comparisons and identification of significant differences between complex proteomes.

Shotgun proteomic analyses yield spectral count MS/MS data, which are similar to Serial Analysis of Gene Expression (SAGE) mRNA tags, which can be modeled by maximum-likelihood techniques based on the Poisson distribution either using a frequentist approach or in a Bayesan framework.[14−16] However, the Poisson distribution is not very flexible for the empirical fitting of shotgun proteomics data and requires that the distribution mean and variance be equal.[17,18] We therefore developed a quasi-likelihood model, that allows over- or under-dispersion of spectral counts for the comparison of shotgun data sets,[19,20] using the F-test to compute p-values and applying the False Discovery Rate (FDR) method to correct for multiple hypothesis testing.[21] This modeling approach generates an estimate for the average spectral counts for each protein across analyses that can be used to compare two or more study conditions. Using a large data set of human proteins spiked into a yeast background, we demonstrate that quasi-likelihood modeling identifies proteomic differences under varying conditions. Using a large data set of human proteins spiked into a yeast background, we compared the performance of five different statistical tests to discern differences under varying conditions. The best results were obtained with GLM approaches and Fisher's Exact test. We analyzed shotgun proteomic data from normal tonsil epithelium and HNSCCs and identified proteins with possible differential expression between these two phenotypes (Figure 1). Expression differences were verified for a selection of proteins using LC−MRM-MS, including keratin 1 and 17, fatty acid-binding protein 5, guanine nucleotide-binding protein 6, cornifin 3 and ferritin light-chain. Many of the identified proteins had links to keratinocyte differentiation or were known as proteins involved in HNSCC carcinogenesis.

Methods

Biological Materials, Sample Preparation, and Analysis by LC−MS/MS. A yeast reference proteome was created for the Clinical Proteomic Technology Assessment for Cancer (CPTAC) network to determine variability in the performance of LC−MS/MS platforms across the CPTAC network. A mixture of 48 human proteins (Sigma Universal Protein Standard 1 (UPS1), Sigma-Aldrich, St Louis, MO) was spiked into the yeast reference proteome at 6 levels: 0, 0.24, 0.67, 2.54, 6.7, and 20 fmol/μg yeast protein as described by Paulovich.[22] The preparation and analysis of the yeast proteome spiked with an equimolar mixture of human proteins (Sigma UPS) at different concentrations was done as part of recently published CPTAC interlaboratory studies.[22] Preparation and processing of these samples was performed centrally at the National Institute for Standards and Technology (NIST) as described.[22] The proteins were reduced, the thiols were alkylated with iodoacetamide, and then the alkylated proteins were digested in solution with trypsin. Standard operating procedures for mass spectrometry analyses were implemented in detail, including HPLC and MS parameters.[22,23] For the current study, we used data from the CPTAC study obtained on two different LTQ-Orbitraps (3 replicates each) and one LTQ (3 replicates) from our laboratory at Vanderbilt University.

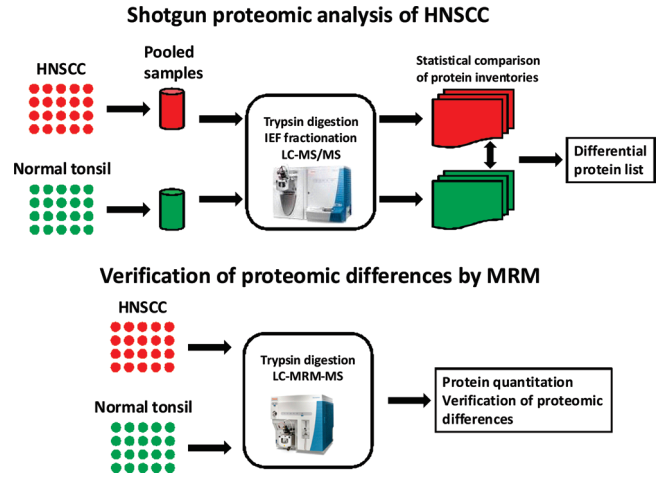

Figure 1. Outline of proteomic procedures for shotgun proteomic analysis of HNSCC and tonsil epithelium protein lysates and for protein quantitation analysis by MRM.

We analyzed a data set reflecting two clinically distinct phenotypes using 20 head and neck squamous cell carcinomas (HNSCC) and 20 normal tonsillectomy tissues from patients identified through the Head and Neck Tissue Repository at Vanderbilt University. This tissue repository was started in 2003 and is used to prospectively collect biological materials from all patients undergoing surgery in the head and neck area at Vanderbilt University. All tumors were obtained from newly diagnosed patients who, with a single exception, had not been previously treated by either chemotherapy or radiation. All tumors were Stage I−III, originated in the oral cavity (except for one oropharyngeal carcinoma) and were histologically classified by a certified pathologist. The normal tonsil epithelium was obtained from pediatric tonsillectomies performed at the Vanderbilt Children's Hospital. Informed consent was obtained from patients or their parents to use biological materials for research purposes under protocols approved by the Vanderbilt University Institutional Review Board. Patient information was kept in an anonymized database that only contained pertinent demographic information and links to the biological specimens. Researchers had no access to any information that could potentially identify individual patients.

All tissues were snap-frozen in liquid nitrogen within 30 min of surgical removal and kept at −80 °C until processing. Prior to sectioning, the tumor samples were macrodissected to achieve a minimum of 70% tumor cell content in the eventual specimen while epithelial cells from normal specimens were dissected away from lymphoid cells. The tissue was embedded in polyvinyl alcohol and three 60 μm slices were placed in separate centrifuge tubes. Polyvinyl alcohol was removed with two washes of 70% ethanol followed by a single wash with deionized water.

Mass spectrometry methods are described in detail in Supporting Information. In summary, tissues were used for protein purification and quantitation, followed by protein digestion using trypsin. The resulting peptides were separated on isoelectric focusing strips that were cut into 20 separate fractions. Each of these fractions was analyzed by a second separation on a liquid chromatography column, followed by MS/MS analysis on a LTQ-Orbitrap. Resulting Thermo "RAW" scans were converted to the universal mzML format and searched using the Myrimatch[24] search algorithm against either a yeast protein database (in the case of CPTAC data) or against a human protein database (for head and neck samples). Proteins were assembled using IDpicker software[25] reporting the minimal list of proteins that was able to explain all peptides that were successfully matched to MS-spectra. Filtering conditions were a minimum of 2 distinct peptides per protein for the yeast data sets, while an additional minimum of 10 identified spectra per protein was required for the head and neck data set. The full and unfiltered IDPicker output data set is provided as Supporting Information and includes a complete list of protein IDs, the number of distinct peptides observed per protein, the number of spectra observed per protein, percentage coverage and the full peptide sequences.

Quasi-Poisson Likelihood Model and Implementation in QuasiTel. Previous results have shown that a frequency-based analysis approach using the number of observed spectra (spectral counting) provides a rough measure of protein levels in complex protein mixtures, especially for more abundant proteins.[3,26,27] The goal of statistical analysis using this technique is to provide a list of proteins of interest for further study. In this approach, the spectra that are confidently assigned to peptide sequences using the above-described criteria are our primary outcome measure. The sum of all spectra linked to each of the proteins was then tabulated for each of the m replicate analyses of each biological sample. This resulted in a table consisting of a protein identifier in each row and a set of n columns representing each of the m replicate analyses for the r biological samples. Spectral count data obtained for several thousand proteins are assigned to this table and typically vary between 0 (no spectra observed) to hundreds of spectra observed.

Different methods have been developed to model spectral count data generated from shotgun proteomic analyses,[5] but currently, there is no "gold standard" statistical method for analyzing such data sets. We therefore developed a quasi-likelihood method and compared this method with four other statistical tests using a standardized data set of yeast spiked with human proteins.

To compare spectral counts for different clinical groups (such as normal vs tumor), we can model the data in a regression framework. Let Y denote spectral counts and x stand for group. Because Y represents spectral count, it is not appropriate to assume a Gaussian distribution; instead, generalized linear models (GLM) should be applied to handle such non-normal responses.[28] Specifically, the Poisson distribution, a distribution from the exponential family of distributions, is usually assumed for count data. The resulting model can be expressed as:

$$\mathrm{Log}(Y) = \beta_0 + \beta_1 X_1 + \dot{\varepsilon} \tag{1}$$

Equation 1 is fitted by maximizing the Poisson likelihood function, and the group effect can be assessed by testing the significance of β_1. When specifying a Poisson distribution, however, we put an equal mean-variance assumption on the data, which usually does not hold in the empirical fitting. To alleviate this assumption, we propose a quasi-Poisson maximum likelihood approach. Instead of claiming that Y is from a known distribution, we assume only knowledge of the first and second moments. In particular, for count data, we are able to specify the link and variance functions of the model, but we do not have a clear idea of an appropriate distribution form for the response. Thus, the important elements of the model specification are link and variance, with outcome less sensitive to distribution of the response, given a reasonable sample size.[29] For a quasi-Poisson method, the regression model can be specified as in eq 1. The fitting procedure is then analogous to fitting the model using Poisson likelihood.

Statistical properties of quasi-likelihood has been derived and presented[30] long ago. Yet, it is first time to be introduced to the shotgun data analysis field. To complete the method description, we rederive the quasi-likelihood Poisson model as follows: for the ith response Y_i (the spectral counts), we have $E(Y_i) = u_i$ and $\mathrm{Var}(Y_i) = \varphi V(u_i)$. Now define a score U_i as

$$U_i = \frac{Y_1 - u_1}{\varphi V(u_1)}$$

Then we have:

$$-E\left(\frac{\partial U_i}{\partial u_i}\right) = -E\left[\frac{-\varphi V(u_i) - (Y_i - u_i)\varphi V'(u_i)}{[\varphi V(u_i)]^2}\right]$$

From this follows:

$$E(U_i) = 0 \tag{2}$$

$$-E\left(\frac{\partial U_1}{\partial u_1}\right) = V(U_i) \tag{3}$$

We notice that the properties for score U_i shown in eq 2 and 3 are shared by the derivatives of the log-likelihood, which suggests that the integration U_i would serve as a good surrogate for likelihood. We can then define a log quasi-likelihood for Y_i as:

$$Q_i = \int_{y_i}^{u_i} \frac{y_i - t}{\varphi V(t)} \mathrm{d}t(t)$$

Then the log quasi-likelihood for all n observations will be

$$Q = \sum_{i=1}^{n} Q_t$$

We can thus verify that Q behaves just like a log-likelihood and that the estimation of β is obtained by maximizing Q. We summarize some features of the quasi-Poisson model as:

1. The usually asymptotic properties expected from maximum likelihood estimators also hold for quasi-likelihood based estimators.[30] Theoretically, these properties are assuring and desirable.

2. The quasi-likelihood Poisson method allows for the dispersion φ to be a free parameter. This parameter is useful in modeling overdispersion, which is typical of shotgun data.[31] The quasi-likelihood method can model the variation of such data with increased accuracy.

3. The quasi-likelihood premise broadens our modeling possibilities for more real-world data types: when we do not have a clear idea about the distribution of the shotgun data, we still can model such data with knowledge only of link and variance.[29]

4. In addition, quasi-likelihood allows us to model data in a regression framework that is easily extended to model more complicated data from complicated experiments, such as repeated measurements, longitudinal data, etc. The quasi-likelihood method provides generally consistent estimates of regression coefficients even if the variance function is misspecified.[32]

Shotgun proteomic analyses typically identify thousands of proteins in complex cell or tissue lysates and hence generate multiple comparison artifacts. To simultaneously test thousands of proteins for differences requires a means to control the rate of statistical false discovery. Unadjusted P-values taken from single-inference procedures result in an increased rate of false positives. Family wise error rate methods are too conservative and have less application value for our purposes because they are not suitable for large-scale simultaneous hypotheses testing problems that arise from high-throughput technologies. To handle the complications presented by simultaneously testing thousands of proteins, we applied the False Discovery Rate (FDR) method.[21] The FDR method is a frequentist-based approach working with P-values (null hypothesis tail area).

Another complicating feature of shotgun proteomic data sets is the presence of large numbers of zero-values (i.e., no spectra observed for a given protein). For example, in comparing two data sets, it is not uncommon to encounter all zeros as spectral count values for a protein in one data set and nonzero values in the other. This causes the estimated standard error to be enormous and the corresponding Wald test statistics, based on estimated coefficients with standard errors, are not reliable. A conservative way to avoid this problem is to add one spectrum count to the data set containing only zeros, but this leads to data distortion at low spectral count levels. An alternative is to perform the comparison using nested models. When comparing models, the P-value of an F-test rather than a Chi^2 test is calculated following the quasi-likelihood method.[29] The "observed total spectrum count" in each run is added into the model as the offset. The offset serves as the "size" variable which determines the number of opportunities for proteins to occur, and by modeling such an offset, the spectral count data are normalized. The model generates quasi-p values for each of the protein entries in the data set and estimates an average spectral count (λ) across the replicate analyses. It also allows the specification of additional parameters, such as the distinction between subgroups within the analysis. We applied the above statistical methods to a data set with known "ground-truth", where different levels of an equimolar mixture of 48 human proteins were spiked into a yeast background. The full QuasiTel program code using the open-source statistical package R is provided as Supporting Information.

LC−MRM-MS Analyses. For LC−MRM-MS analyses, protein lysates from 20 normal and 20 HNSCC tissues were processed individually as described above with no additional IEF peptide fractionation. Tryptic peptides were resuspended in 0.1% formic acid to a final concentration of 0.44 mg/mL. A total of 18 proteins were selected for MRM analysis on the basis of differential spectral counts between normal epithelium and HNSCC. For each protein candidate, peptides were selected for MRM analysis giving priority to those peptides previously identified in the shotgun data set with high MS/MS spectral quality. Additional peptides were selected by *in silico* digestion, and all peptides included in MRM analysis were required to be 7 to 25 amino acids in length, be fully tryptic, and contain no ragged ends or motifs (NXT/S). Peptides containing cysteine and/or methionine residues were not excluded. Peptide uniqueness was confirmed by searching against the IPI human database (Version 3.56). Two to 4 unique peptides were selected for each protein and selected transitions for all candidate peptides were optimized in a trial LC MRM MS analysis consisting of a single normal and tumor samples, and transitions with high peak intensity and no interferences from background matrix were selected for further data collection. The complete list of MRM peptides, precursor masses and transitions is provided as Supporting Information. Two isotopically labeled peptides from β-actin (GYSFTTTAER, with $^{13}C/^{15}N$ labeled R), and annexin A1 (VLDLELK with $^{13}C/^{15}N$ labeled K, New England Peptide, LLC, Gardner, MA, >95% purity according to amino acid analysis) were each spiked into tissue digests at a concentration of 25 fmol/μL prior to MRM-MS analysis for data normalization.

All samples were analyzed on a triple quadrupole mass spectrometer (TSQ Quantum Ultra, Thermo Fisher Scientific, Waltham, MA) equipped with an Eksigent 1D Plus NanoLC pump (Eksigent Technologies, Dublin CA). The LC mobile phase consisted of Solvent A, an aqueous solution with 0.1% formic acid, and solvent B, acetonitrile with 0.1% formic acid. Peptides were separated on a packed capillary tip (Polymicro Technologies, 100 μm × 11 cm) with Jupiter C18 resin (5 μm, 300 Å, Phenomenex) using an in-line solid-phase extraction column (100 μm × 6 cm) packed with the same C18 resin using a frit generated with liquid silicate Kasil 1[33] as previously described.[34] Injections (2 μL) of a 0.44 mg/mL digest (based on protein concentration) were followed by a 10 min wash period with 100% A, then by elution with a gradient of 2−25% solvent B in 25 min, 25−50% solvent B in 20 min, and finally by elution with 50−90% solvent B in 10 min. Data acquisition was performed with an ion-spray voltage of 1200 V, capillary temperature 210 °C and skimmer offset −5 V. Both Q1 and Q3 were set at unit resolution (fwhm 0.7 Da), and Q2 at 1.5 mTorr. Scan width was 0.004 m/z and scan time 10 ms. Collision energies for each peptide are listed in the Supporting Information. Peak areas for each peptide were extracted and integrated using Pinpoint software (Thermo Fisher Scientific, Waltham, MA).

Criteria to accept or reject a MRM peptide measurement were: (1) all peptide MRM transitions were required to generate integrated precursor peak areas higher than 10^5; (2) at least three of the specified SRM transitions had a measured signal-to-noise greater than 3 in either normal or tumor samples; (3) peptide MRM transitions meeting the previous criteria were consistently observed in at least one of the tissue types; (4) retention time aligned for both precursor and individual SRM transitions; and (5) the relative intensities of fragment ions were consistent with those observed in ion trap MS/MS spectra (if available) and were consistent between different samples. Comparisons between labeled β-actin-normalized measure-

ments were made by Student's *t* test with Welch correction for unequal variance using Prism 5.0 (GraphPad Software, La Jolla, CA).

Results

To compare different statistical approaches to spectral count data, we employed an existing data set created within the framework of the CPTAC network was used for differential proteomic analysis. In this data set, a known set of 48 human proteins (Sigma UPS) was spiked in equimolar amounts into a yeast background (CPTAC Study 6).[23] The study employed reverse phase LC−MS/MS analyses of the samples on several Thermo LTQ and Orbitrap instruments. We compared spectral count data from triplicate analyses on one LTQ instrument and on two Orbitrap instruments at Vanderbilt using the CPTAC standard operating procedure for Study 6.[22] A total of 7 preparations were analyzed: yeast only (QC2, at 60 ng/μL) and a series of the following Sigma UPS spikes into 60 ng/μL yeast; 0.25 fmol/μL (A); 0.74 fmol/μL (B); 2.2 fmol/μL (C); 6.7 fmol/μL (D); and 20 fmol/μL (E). All data from the three instruments were integrated into a single protein assembly report requiring a minimum of 2 distinct peptides per protein at 2% peptide false discovery rate (FDR). The total number of MS/MS scans that were successfully matched to peptide sequences was 178 985, representing a total of 1942 protein groups, including identifications of 46 of the 48 spiked Sigma UPS proteins. In this context, the term "protein groups" denotes proteins that are indistinguishable based on identified peptide sequences as described previously.[35] The assembled protein list included 216 identifications to reversed-sequence proteins for an unacceptably high 22% protein FDR. We therefore required any protein to be identified by a minimum of at least 5 spectra; this modestly reduced the number of protein groups to 1437. This filter eliminated the vast majority of false identifications, yielding a 6.8% protein FDR, but resulted in no loss of any of the 46 identified Sigma UPS protein identifications. All subsequent analyses and statistical tests were performed on this data set of 1437 protein groups. The full IDPicker report and the filtered data set used for the statistical analyses are provided as Supporting Information.

Graphical User Interface (GUI) for the Analysis of IDPicker Output. To facilitate the analysis of shotgun data sets summarized using IDPicker,[25] we coded the 5 statistical tests in the R-based program QuasiTel. QuasiTel allows direct import of IDPicker files for statistical comparisons by reading the IDPicker spectra-by-protein-by-group file and automatically loading the corresponding IDPicker summary file containing the number of confident identifications for each group used as offset in the model. This offset serves as a "normalization" factor when the numbers of confident IDs observed vary between replicates. The user can then select the sets of MS analyses to be compared and a minimum number of total spectral counts required across all groups for each protein. The latter increases the power of QuasiTel and eliminates proteins with too few spectral identifications to yield a meaningful comparison. The output file contains all protein IDs that fulfill the filtering criteria and lists spectral count numbers and model-generated rates (λ) for each group, as well as the crude p-value and FDR-corrected quasi-likelihood p-values for each of the statistical tests. It also calculates a protein dispersion parameter (φ), which relates to the variation in spectral counts across all measurements, as well as coefficients of variation for each of the groups. The dispersion parameter can be used to identify proteins with the greatest variation between replicate analyses across all groups, which may be of most interest as markers for individual properties of the samples. As an internal validation, the reported spectral counts are not taken directly from the input tables, but rather are reverse-calculated using the model-generated rates (λ) and provided offset numbers. The ratio of the group rates, expressed as base-2 log of the ratio of the rates ("rate ratio"), can be used as an estimate of the magnitude of the difference between the spectral counts and, by extension the expected protein levels, in each group.

Comparison of Statistical Tests Identifying Spiked Sigma UPS Proteins. We compared five statistical tests for their capacity to distinguish differences between the various spike levels of Sigma UPS in a yeast background. These tests included the Wilcoxon Rank test, Student's *t* test, Fisher's Exact test, and Generalized Linear Modeling using Poisson and quasi-likelihood approaches. We contrasted spectral count data from yeast-only LC−MS/MS analyses against data from yeast at each of the 5 spike levels and additionally created contrasts representing a 3-, 9-, and 27-fold difference in spiked protein levels on the yeast background (Table 1). For these comparisons, we used the 1437-protein group data set and list the number of true- and false-positive identifications of proteins with a minimum of 5 spectral counts across each of the comparisons that reached at least a 2-fold increase in the λ parameter. At the highest spike level (yeast-only against E), all tests performed similarly in identifying almost all of the 46 spiked proteins. This high level of identification remained in the yeast-only against D comparison with the two GLMs and Fisher's Exact test but Wilcoxon rank test and Student's *t* test identified fewer spiked proteins in this comparison. This pattern remained the same in the yeast-only against C spike level. At the two lowest spike levels, the quasi-likelihood model outperformed the other tests, albeit at the cost of an increased number of false-positives.

The patterns seen in the yeast-only against yeast plus spiked proteins results were also apparent when we created contrasts based on fold-differences in spiked protein levels. All tests were capable of distinguishing almost all true positives with 27-fold differences at the highest absolute concentration of spiked protein (B versus E). At a 3-times lower spike level the two GLMs and Fisher's Exact test remained powerful, while Wilcoxon rank and Student's *t* test identified fewer true positive proteins (A versus D). Similar results were obtained at the 9-fold spiked protein difference (C versus E and B versus D). At a 3-fold difference in spiked protein levels, the two GLMs were capable of identifying about half of the true-positive proteins that were spiked into yeast, while the other tests yielded lower numbers of true positives. Quasi-likelihood modeling outperformed Poisson modeling with a 3-fold difference at the lowest spiked protein levels (C versus D), although again at the cost of a higher number of false-positive identifications.

Comparative Analysis of HNSCC and Normal Tonsil Epithelium by Shotgun Proteomics. A small shotgun proteomic data set reflecting possible differences between HNSCC and normal tonsil epithelium was used to perform a shotgun proteomic comparison in a real-world scenario. Pools of 20 tumors and 20 normal tissues were subjected to analysis on a LTQ-Orbitrap in quadruplicate. The resulting data set consisted of a total of 52 956 scans that matched human peptide sequences with a 5% peptide FDR. Using the parsimonious protein assembly algorithm of IDPicker, the total number of protein groups identified in this data set was 2211 with a protein-level FDR of 6.7%. The numbers of protein groups

Table 1. Detection of Proteomic Differences in Yeast Sample with Spiked Human Proteins

	human/yeast spectral counts[a] (true positive/false positive)				
	quasi-likelihood	Poisson model	Fisher Exact test	Wilcoxon rank test	Student's *t* test
		Yeast versus yeast + spike			
E (20 fmol/μg yeast)	45/1	44/0	44/0	41/0	42/0
D 6.7 (fmol/μg yeast)	39/2	38/0	37/0	31/0	32/0
C 2.7 (fmol/μg yeast)	31/1	30/2	27/0	21/0	23/0
B 0.67 (fmol/μg yeast)	16/1	7/0	7/0	3/0	8/0
A 0.24 (fmol/μg yeast)	2/3	0/0	0/0	0/0	0/0
		27-fold difference			
B versus E	44/9	44/4	44/1	41/3	42/4
A versus D	38/1	37/0	36/0	31/0	32/2
		9-fold difference			
C versus E	42/10	42/6	42/2	39/2	40/2
B versus D	33/1	32/0	30/0	27/0	29/0
		3-fold difference			
D versus E	20/2	19/1	15/0	12/0	13/0
C versus D	14/3	8/0	8/0	7/0	12/0

[a] Spectral counts in the full data set which are at least 2-fold higher in the spiked sample with FDR-corrected *p*-value of less than 0.05.

Table 2. Top Ranked Proteins with Differential Spectral Counts between HNSCC and Normal Tonsil Epithelium

rank	IPI identifier	gene ID	normal[a]	HNSCC[a]	$^2\log(\lambda1/\lambda2)$	quasi *p*-value[b]
		Top ranked proteins with higher spectral counts in HNSCC				
1	IPI00007797.3	FABP5	0	22	34.99	0.00015
2	IPI00302944.3	COL12A1	3	38	4.02	0.00149
3	IPI00295400.1	WARS	3	24	3.36	0.00308
4	IPI00220327.3	KRT1	9 (1)	47 (31)	2.74	0.00306
5	IPI00450768.7	KRT17	43 (2)	202 (26)	2.59	0.00079
6	IPI00298860.5	LTF	22 (3)	77 (16)	2.17	0.0028
7	IPI00010951.2	EPPK1	18 (14)	63 (55)	2.17	0.00250
8	IPI00007244.1	MPO	8	27	2.11	0.00306
9	IPI00299263.5	ARFGAP3	3	10	2.09	0.02472
10	IPI00010800.2	NES	14	46	2.07	0.00414
11	IPI00738499.2	FTL	6	19	2.02	0.00308
		Top ranked proteins with higher spectral counts in normal tonsil epithelium				
1	IPI00082931.1	SPRR3	36	0	−33.54	0.00057
2	IPI00025084.3	CAPNS1	13	0	−33.52	0.00079
3	IPI00412546.3	CR1	13	0	−33.52	0.00079
4	IPI00165528.1	USP47	10	0	−33.14	0.00032
5	IPI00301250.6	EPS8L1	12	0	−30.52	0.01645
6	IPI00006034.1	CRIP2	11	0	−30.39	0.01554
7	IPI00375746.4	GBP6	35	1	−4.77	0.00478

[a] Total spectra observed for protein group (unique spectra for indicated protein if different from protein group total). [b] FDR corrected.

matched to HNSCC and normal tonsil epithelium were very similar, 1975 and 2125, respectively. The full list of protein group identifications is provided as Supporting Information. To eliminate proteins that had too few spectral counts to become statistically different, we further reduced the data set to protein groups that were identified by at least 5 spectra, resulting in a data set of 1733 protein groups, including 16 protein groups included as contaminants and 8 reverse-sequence proteins (0.9% protein-level FDR). A total of 1633 protein groups were observed in both normal tonsil epithelium and HNSCC while 75 protein groups were unique to normal epithelium and 22 protein groups were unique to HNSCC. A total of 1029 protein groups were observed in all 4 replicate analyses of both samples. The full data set with statistical analysis and filtering options is provided as Supplemental Table 1, Supporting Information.

The highest numbers of spectral counts for a single protein in both normal epithelium and HNSCC was observed for desmoyokin (AHNAK). This is an unusually large nuclear scaffolding protein of 5890 amino acids with an approximate MW of 700 kDa. AHNAK plays an important role in the formation of cytoskeletal structure, calcium homeostasis, and muscle regeneration.[36] We also found large numbers of spectra derived from several structural proteins such as plectin-1, myosin-9, and periplakin. The data set was tested for possible differences, which yielded a total of 86 protein database entries with a quasi *p*-value of less than 0.05 and at least a 4-fold difference in spectral counts ($\log_2$(rate1/rate2), or rate ratio, higher than 2 or lower than −2). This list includes a number of proteins with artificially low *p*-values either because no spectra were identified in one of the groups (reported with "cv" value of "NA"), or all nonzero measurements in one of the group were identical (reported with "cv" value of 0). These values serve to flag these proteins for manual evaluation of the data.

Verification of Shotgun Proteomic Data by LC−MRM-MS. Results from global shotgun proteomic experiments yield long lists of proteins with varying degrees of evidence for differences in absolute levels. These results need to be verified by more targeted and quantitative technologies. Recently, LC−MRM-MS has been employed for this purpose,[13,15] and we chose this method to determine possible differences in protein levels for the 18 proteins in Table 1. This list includes all 11 proteins with higher spectral counts in HNSCC, with the exception of MUC5B, which was included as potential protein contaminant; and 7 proteins with the highest values in normal tonsil epithelium compared to HNSCC. We chose proteins for which high numbers of spectral counts were observed, but also proteins for which relatively few spectra were observed. Thus, the selected set of proteins formed a representative, albeit not comprehensive, selection of proteins for which the levels were potentially different between these two phenotypes. LC−MRM-MS was used to analyze the *individual* specimens that we used to generate the HNSCC and tonsil sample pools analyzed by LC−MS/MS.

Although LC−MRM-MS analyses frequently employ stable isotope labeled internal standards for the target peptides of

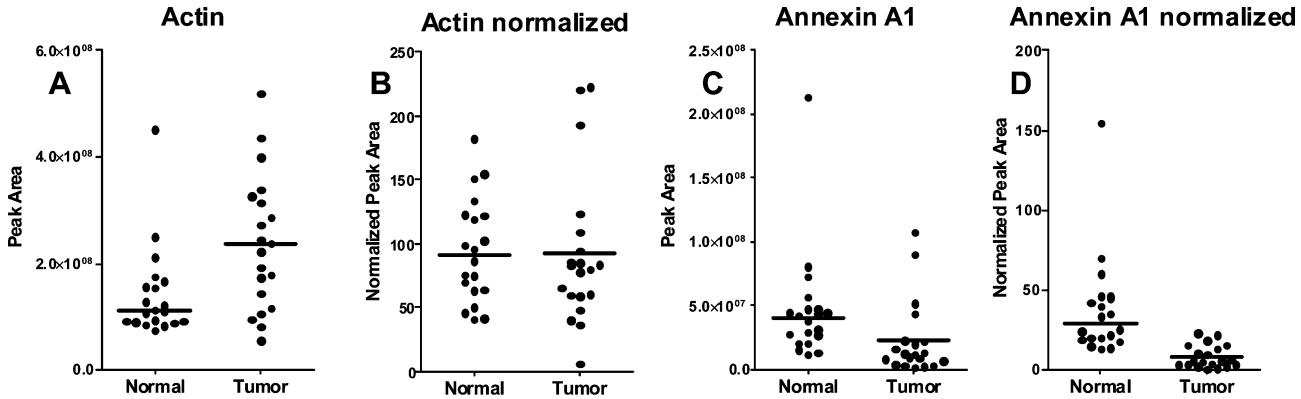

Figure 2. Comparison of MRM results obtained for actin and annexin-A1 on the 20 individual normal tonsil epithelia (normal) and 20 individual HNSCCs (tumor). (A) Results obtained for a single actin-specific peptide without normalizing the MRM measurements. (B) Results obtained with the same actin-specific peptide, normalized against an isotopically labeled version of the same sequence. (C) Measurements for annexin-A1 without normalization. (D) Measurements for annexin-A1, normalized against a labeled version of the identical annexin-A1 peptide. The actin example illustrates how an apparent difference in mean MRM measurement disappears after normalization to the identical labeled actin peptide, which indicates successful normalization for instrument variation. The annexin-A1 example shows that in contrast to actin, an apparent lower annexin-A1 level in HNSCCs remains after normalizing for instrument variation.

interest,[13,15,37] we employed instead a single isotope-labeled peptide standard found in two isoforms of the abundant cytoskeletal protein actin. We first measured the levels of endogenous actins using a single peptide common to the α- and γ-forms of actin using known levels of the identical isotopically labeled peptide spiked into each of the lysates. As shown in Figure 2, endogenous actin measurements appeared different between normal tonsil epithelium and HNSCC, but this apparent difference disappeared when normalizing against the spiked labeled peptide levels. To investigate the potential effect of normalization, we also took measurements of annexin-A1 using a spiked isotopically labeled peptide for the same protein (Figure 2). In this case, normalization of the annexin-A1 signals led to small adjustments in the observed values for individual cases but their overall ranking by peptide signal remained largely similar ($r = 0.81$, Spearman Rank correlation).

Of the 18 potential biomarker proteins initially selected for MRM, 6 proteins produced specific and reproducible signals in unfractionated lysate for at least a single peptide (Figure 3). All MRM-detectable proteins (KRT1, KRT17, FABP5, and FTL) with higher spectral counts in tumor versus normal in the shotgun data set showed a significant increase in MRM signals. The two MRM-detectable proteins (GBP6 and SPRR3) with higher spectral counts in normal tonsil epithelium were both verified by MRM to be significantly higher in normal compared to HNSCC. Expression of KRT1 has been described in keratinizing epidermis and is found in keratin pearls formed in well-differentiated HNSCCs,[38,39] such keratin agglomerations are normally not found in mucosal epithelium. Type II keratins form the largest cluster of protein groups in our analysis (Cluster 1) and KRT1 is detected by 56 peptide spectra of which 32 are unique for KRT1 and not shared with any other keratins. Of these 32 KRT1-specific peptides, only a single one was found in normal tonsillar epithelium, a finding that demonstrates that the differential protein findings are specific for KRT1. The other keratin, KRT17, is found in hyperproliferative squamous epithelia[39] and overexpression of KRT17 mRNA has been reported in HNSCC.[40] KRT17 is part of cluster 5 that includes a variety of Type I keratins, it is detected by a total of 245 spectra (protein group ID: F) of which 28 are specific for KRT17. Only 2 of the 28 KRT17-specific spectra were found in normal tonsillar epithelium compared to HNSCC, again reinforcing the notion that shotgun proteomics is highly specific in detecting proteomic differences. Of interest, KRT17 shares 92 spectra with protein group I in cluster 5, this is a Type I keratin that is not further characterized (IPI00240503.7). Only a single spectrum out of all 92 shared spectra was found in normal tonsillar epithelium, indicating that either KRT17 or the uncharacterized keratin was almost exclusively found in HNSCC. Likewise, the 22 spectra observed for FABP5 were only observed in HNSCC and not in tonsillar epithelium. FABP5 is a fatty acid-binding protein that was first described in psoriatic skin keratinocytes[41] and that appears to be involved in keratinocyte development.[42] Fatty acid-binding proteins are small cytoplasmic proteins that presumably fulfill roles in fatty acid uptake, transport and metabolism. FABP5 (E-FAPB) is upregulated by the tumor-associated antigen EpCAM in HNSCC,[43] while proteomic analysis of HPV-related HNSCCs revealed that FABP5 was one of the proteins upregluated in these tumors.[44] One of the proteins with lower spectral counts in HNSCC was GBP6. The presumed function of this protein is to bind guanylate and to hydrolyze GTP to both GDP and GMP but no data is available for this specific for of GBP. The other protein in that was found lower in HNSCC was SPRR3, or esophagin or cronifin-β. This is a small proline-rich cytoplasmic protein that is cross-linked to envelope protein of keratinocytes.[45] This protein was detected in oral epithelia and in well-differentiated squamous cell carcinomas of the skin or oral mucosa.[45] These results demonstrate that proteins that appear to be differentially expressed by shotgun analysis can be verified using the more quantitatively precise LC−MRM-MS approach on the individual sample lysates from which the pooled lysates for shotgun proteomics were generated.

Discussion

Shotgun proteomic analyses are increasingly being applied to clinical specimens to investigate disease state, disease predictions, and response to treatment. A key objective of these studies is the identification of a set of proteins whose presence and expression levels distinguish the biological states under study. When comparing two biological states, it is reasonable

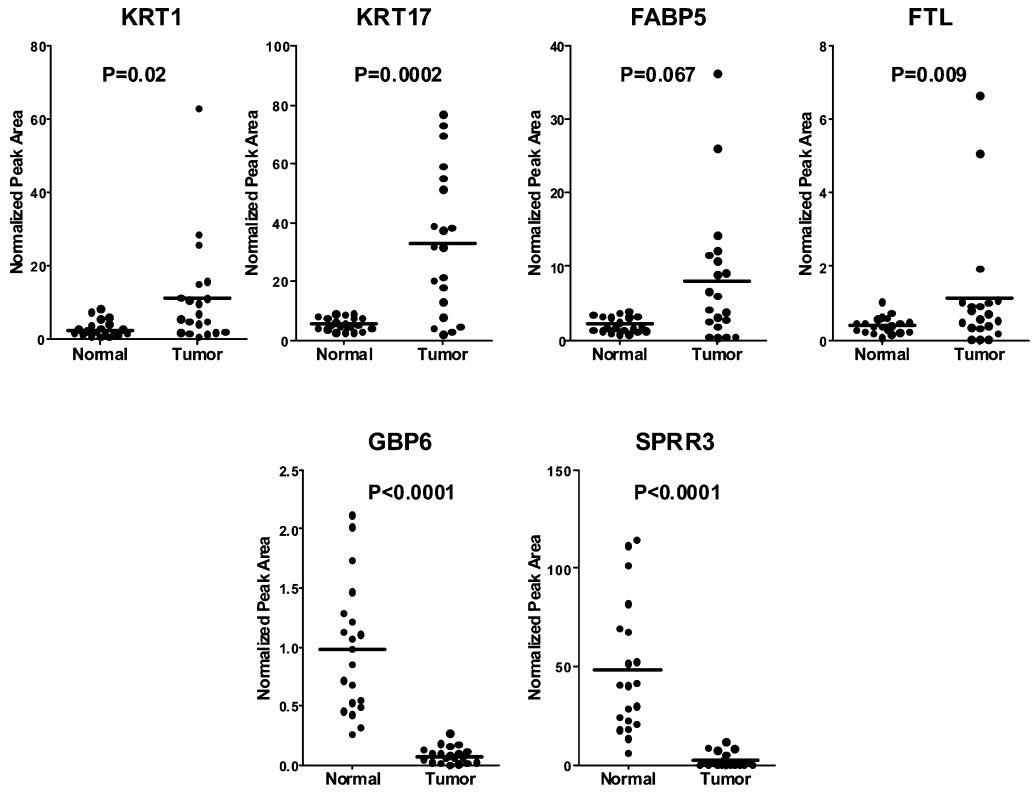

Figure 3. Examples of LC−MRM-MS analyses of a selected set of 6 proteins with large spectral count differences between HNSCC and normal tonsil epithelium. Unfractionated lysates were tested for 4 proteins with higher spectral counts in tumor versus normal tonsil epithelium in the shotgun data set (KRT1, KRT17, FABP5, and FTL) and 2 proteins with lower spectral counts in tumor versus normal tonsil epithelium (GBP6 and SPRR3). Results from these proteins confirmed the original findings in the shotgun proteomic data obtained from pooled samples. Indicated *p*-values were calculated using Student's *t* test.

to assume that the majority of proteins remain at more or less constant levels. Hence, the spectral counts of these proteins should also remain constant. Actual observations for specific proteins may vary depending on a variety of reasons related to the specifics of data-dependent spectral acquisition and peptide fractionation techniques. For instance, peptides from less abundant proteins may be masked by peptides from more abundant proteins if their precursor m/z and elution pattern is similar. Alternatively, peptides may not ionize well because of size or post-translational modifications or may be lost because they fall outside the IEF focusing range during peptide fractionation. Notwithstanding these specialized situations, the similar average behavior of peptides allows for a comparison of complex mixtures from different biological states that consist of similar sets of proteins, and spectral counting provides a convenient measure to assess potential differences in proteomes.

Using existing data from a CPTAC interlaboratory study on reproducibility of different shotgun proteome analysis approaches,[22,23] we tested several statistical comparison strategies to reveal biological differences between two large scale shotgun proteomic data sets. Such data sets have specific properties dictated by the nature of data-dependent MS sampling. The resulting MS/MS spectra are matched to protein-encoding sequences within the human genome by well-established methods.[24] These analyses generate a list of identified peptides and the numbers of times each of the peptides are observed. These count data can be modeled using GLM and other approaches. As we demonstrated previously, in complex protein mixtures such as tissue lysates, sampling of less abundant proteins becomes a chance event and multiple replicate analyses are necessary to increase coverage of the proteome.[7,23] In addition to the nonlinearity of count data at the lower and higher ends, comparisons of two shotgun proteomic data sets becomes difficult when peptides from a given protein are exclusively observed in only one of the two data sets. Due to the complexity of tissue lysates and properties of data acquisition, a large number of proteins are only identified by small numbers of peptides and may only be represented in a few of the replicate analyses or even be completely absent in one of two groups under comparison. This analysis shows that despite the above limitations, known differences in protein levels are detectable from the comparison of shotgun data sets given the proper statistical analysis tools.

The strategies described in this report can be used as a general framework for analysis of data from various shotgun proteomic experimental designs. The quasi-likelihood approach builds upon previous work using maximum likelihood-based methods, although it does not require identification probabilities from other modeling tools such as generated by ProteinProphet and used in analysis tools such as SASPECT.[15] The SASPECT package provides a test for comparing protein identifications between two groups. However, the test score is derived from the Boolean values for the presence or absence of a peptide identification rather than on the actual spectral counts. SASPECT relies on PeptideProphet confidence scores of each of the peptide identifications to account for error and is thus not appropriate for the analysis of pure spectral counts. A separate approach is to combine spectral counts and the number of protein observations in individual subjects into a "spectral index",[27] a method we did not consider because it is less appropriate for pooled data analysis. The spectral index

method is based on a combination of spectral count information and the number of specimens in which a certain protein is identified. Thus, it requires independent measurements on biologically independent specimens to create a meaningful spectral index calculation, something that can not be accomplished with a limited number of technical replicates on samples obtained by pooling a larger number of specimens for each of the phenotypes as outlined for the head and neck carcinoma data.

In this paper, we chose to explore methods that are not dependent on the analysis pipeline but rather can be applied to spectral count data, independently of how they were derived. With sufficient replicate analyses and summation of peptide identifications assigned to intact proteins the spectral count numbers are raised to such a degree that modeling of the observed counts becomes possible. Other tests, such as G-test, Fisher's Exact test, AC test, and Student's *t*-test, have also been used to detect differences in spectral counts between groups.[5,46] However, such methods are not designed specifically for spectral count data and since they do not take the distribution of the count data into consideration, they lack statistical power in general. The shortcomings of these other tests became apparent in the context of the CPTAC spikes in yeast lysate with known "ground-truth" that allowed us to compare the performance of a collection of different statistical tests in a present/absent setting and in a x-fold change setting, both at different spike levels. As expected, the nonparametric Wilcoxon rank test and Student's *t*-test were least capable of discerning the differences between the levels of spiked human proteins in yeast. Fisher's Exact test performed very well in most circumstances, but this approach lacks the option of including additional covariates in the comparisons. Poisson and quasi-likelihood modeling performed well in most situations, each with their specific strengths and shortcomings. An additional advantage of our modeling approach is that important covariates and/or factors describing known sources of variation (e.g., subgroups) may be included in the statistical model specification. For example, the CPTAC spike study was conducted at multiple sites and on multiple instruments, which could potentially contribute variability to the spectral counts. A Poisson or quasi-likelihood statistical model could easily accommodate site or instrument factors that would account for these variations, and result in more powerful tests. Indeed, the CPTAC analyses at Vanderbilt included both LTQ and Orbitrap instruments. We considered including an "instrument" factor to accommodate interinstrument variation. In this case, we found that interinstrument differences were not substantial after including the total spectral counts per run in the offset (data not shown). We elected to omit the instrument factor from the statistical analysis to streamline the presentation of results.

Protein identification from the CPTAC study 6 data set, whereby yeast was spiked with a mixture of human proteins provided a large data set that was acquired using standard operating procedures. For comparisons of yeast-only versus yeast plus spiked human proteins, Fisher's Exact, Poisson and quasi-likelihood modeling all performed similarly. However, in comparisons of fold-differences, the quasi-likelihood modeling generated a larger number of false positive results. Upon examination of the data, the reason for this difference appeared to be a greater sensitivity of quasi-likelihood modeling to homogeneously distributed spectral counts than the other two tests. In the spiked yeast data set, a larger number of proteins show such an even spectral count distribution and the majority of the false-positives were due to yeast proteins with small differences in absolute spectral counts that were very consistent between replicates. Such proteins are flagged in our analysis by cv values of "NA" or "0" and can be subjected to separate additional analyses. In a more variable data set such as the head and neck data set presented in this paper, this sensitivity to equal distribution of spectral counts among replicates does not lead to identifications of flagged proteins and provides an advantage over Fisher's Exact and Poisson tests where a single outlier value may result in a significant test that is not representative of the overall differences.

To discern possible differences in protein levels between normal tonsil epithelium and HNSCC we applied QuasiTel on a relatively small IDPicker data set obtained from pooled tissue specimens. Due to constraints on available MS time, we used a pooled lysate strategy whereby 20 individual protein lysates of equal protein amounts were combined into a single preparation that was subsequently analyzed by shotgun proteomics. This approach has the advantage of a large reduction in analysis time and focuses on proteomic differences that are common to the pooled samples. However, the pooling approach loses the capability to identify single outlier values from a small number of samples within the pool. This relatively small data set yielded statistically significant differences in protein spectral counts between normal tonsil epithelium and HNSCC pools. We chose to verify a subset of these proteins identified based on their differences in spectral counts between the two pooled samples. Of the 6 proteins for which reliable MRM data could be obtained, the general trend observed in the shotgun experiment was consistent for all of the proteins. The MRM signals for successfully measured peptides generally reflected the level of spectral counts in the shotgun data set. For instance, keratins 1 and 17 had high levels of counts in the HNSCC sample and MRM signals that were about 10-fold higher than for most of the other proteins. One potential drawback of MRM verification is that these studies are limited by the success at which proteotypic peptides can be selected and by the specificity of the signals that can be obtained for them using unfractionated cell lysates. The expansion of spectral libraries and better prediction tools for peptides suitable for MRM will improve these choices in the future.

Global comparisons of shotgun proteomic data sets such as described in this paper can provide important insight into the protein composition of biological specimens in terms of cellular location, molecular function or biological processes. However, these analyses are based on identifications of whole proteins and depend on the procedures used to obtain data sets of nonoverlapping protein sets. In IDPicker, a protein group consists of an aggregate of protein database entries that can only manually be distinguished on the basis of peptide identifications. With global analysis, the sum of all peptide identifications is used for comparisons between data sets, a comparison that ignores any differences at the peptide level. This feature can become problematic for example in the case of keratins where a large family of proteins shares multiple peptide sequences. Other examples include alternative forms as a result of post-translational modifications, protein processing, or changes in primary sequence dictated by alterations at the DNA level. Future enhancements of the IDPicker software package will need to address this complication of protein assembly so that subtle differences based on peptide sets can be revealed in global proteomic analyses.

Another limitation of the current study is that the normal comparison group consisted of pediatric cases, and thus, the comparison not only includes normal versus cancer differences but also potential differences between pediatric and adult tissues. This potential bias limits the value of the comparison and any of the described proteomic differences will need to be further validated using other protein detection strategies. Notwithstanding this aspect of the study, the major proteomic differences were all observed in proteins that were relevant for keratinocyte development and differentiation, while some proteins had already been identified as potentially upregulated in HNSCC versus normal epithelium. Further studies will be needed to confirm our proteomic findings in the context of a cancer/normal tissue comparison.

In conclusion, we have evaluated a panel of statistical methods for the assessment of differences in shotgun proteomic data sets. We have used this approach to detect and verify proteomic differences between normal tonsil epithelium and HNSCCs. The analysis approaches described for this study are adaptable to investigate potential proteomic differences in other large shotgun proteomic data sets.

Acknowledgment. We thank Dr. Dave Tabb and Dr. Lisa Zimmerman for helpful discussions and Brandee Brown and Misti Martinez for expert technical assistance. The research presented in this paper was supported by the National Cancer Institute Clinical Proteomic Technologies Assessment for Cancer program through National Institutes of Health Grant 1U24CA126479 and a generous gift from the Jim Ayers Foundation.

Supporting Information Available: Cluster information and supplementary tables. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) McDonald, W. H.; Yates, J. R. Shotgun proteomics and biomarker discovery. *Dis. Markers* **2002**, *18* (2), 99–105.
(2) Wu, C. C.; Yates, J. R. The application of mass spectrometry to membrane proteomics. *Nat. Biotechnol.* **2003**, *21* (3), 262–7.
(3) Liu, H.; Sadygov, R. G.; Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **2004**, *76* (14), 4193–201.
(4) Gao, J.; Opiteck, G. J.; Friedrichs, M. S.; Dongre, A. R.; Hefta, S. A. Changes in the protein expression of yeast as a function of carbon source. *J. Proteome Res.* **2003**, *2* (6), 643–9.
(5) Old, W. M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K. G.; Mendoza, A.; Sevinsky, J. R.; Resing, K. A.; Ahn, N. G. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* **2005**, *4* (10), 1487–502.
(6) Durr, E.; Yu, J.; Krasinska, K. M.; Carver, L. A.; Yates, J. R.; Testa, J. E.; Oh, P.; Schnitzer, J. E. Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. *Nat. Biotechnol.* **2004**, *22* (8), 985–92.
(7) Slebos, R. J.; Brock, J. W.; Winters, N. F.; Stuart, S. R.; Martinez, M. A.; Chambers, M. C.; Zimmerman, L. J.; Ham, A. J.; Tabb, D. L.; Liebler, D. C. Evaluation of Strong Cation Exchange versus Isoelectric Focusing of Peptides for Multidimensional Liquid Chromatography-Tandem Mass Spectrometry. *J. Proteome Res.* **2008**, *119* (8), 1531–7.
(8) Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **1995**, *270* (5235), 467–70.
(9) Sorlie, T.; Perou, C. M.; Tibshirani, R.; Aas, T.; Geisler, S.; Johnsen, H.; Hastie, T.; Eisen, M. B.; van de Rijn, M.; Jeffrey, S. S.; Thorsen, T.; Quist, H.; Matese, J. C.; Brown, P. O.; Botstein, D.; Eystein Lonning, P.; Borresen-Dale, A. L. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (19), 10869–74.
(10) van't Veer, L. J.; Dai, H.; van de Vijver, M. J.; He, Y. D.; Hart, A. A.; Mao, M.; Peterse, H. L.; van der Kooy, K.; Marton, M. J.; Witteveen, A. T.; Schreiber, G. J.; Kerkhoven, R. M.; Roberts, C.; Linsley, P. S.; Bernards, R.; Friend, S. H. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **2002**, *415* (6871), 530–6.
(11) Glas, A. M.; Floore, A.; Delahaye, L. J.; Witteveen, A. T.; Pover, R. C.; Bakx, N.; Lahti-Domenici, J. S.; Bruinsma, T. J.; Warmoes, M. O.; Bernards, R.; Wessels, L. F.; Van't Veer, L. J. Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics* **2006**, *7*, 278.
(12) Cronin, M.; Pho, M.; Dutta, D.; Stephans, J. C.; Shak, S.; Kiefer, M. C.; Esteban, J. M.; Baker, J. B. Measurement of gene expression in archival paraffin-embedded tissues: development and performance of a 92-gene reverse transcriptase-polymerase chain reaction assay. *Am. J. Pathol.* **2004**, *164* (1), 35–42.
(13) Rifai, N.; Gillette, M. A.; Carr, S. A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* **2006**, *24* (8), 971–83.
(14) Beissbarth, T.; Hyde, L.; Smyth, G. K.; Job, C.; Boon, W. M.; Tan, S. S.; Scott, H. S.; Speed, T. P. Statistical modeling of sequencing errors in SAGE libraries. *Bioinformatics* **2004**, *20* (1), i31–9.
(15) Whiteaker, J. R.; Zhang, H.; Zhao, L.; Wang, P.; Kelly-Spratt, K. S.; Ivey, R. G.; Piening, B. D.; Feng, L. C.; Kasarda, E.; Gurley, K. E.; Eng, J. K.; Chodosh, L. A.; Kemp, C. J.; McIntosh, M. W.; Paulovich, A. G. Integrated pipeline for mass spectrometry-based discovery and confirmation of biomarkers demonstrated in a mouse model of breast cancer. *J. Proteome Res.* **2007**, *6* (10), 3962–75.
(16) Choi, H.; Fermin, D.; Nesvizhskii, A. I., Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell. Proteomics* **2008**, *7*, 2373−2385.
(17) Cameron, A.; Trivedi, P. *Regression analsis of count data*; Cambridge University Press: Cambridge, 1998.
(18) Agresti, A. *Categorical data analysis*; John Wiley: New York, 2002.
(19) Breslow, N. *Test of hypotheses in overdispersed Poisson regression and other quasi-likelihood models*; Brown University: Providence, RI, 1990; Vol. 85.
(20) Faddy, M. J.; Bosch, R. J. Likelihood-based modeling and analysis of data underdispersed relative to the Poisson distribution. *Biometrics* **2001**, *57* (2), 620–4.
(21) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., Ser. B* **1995**, *57*, 289–300.
(22) Paulovich, A. G.; Billheimer, D.; Ham, A. J.; Vega-Montoto, L. J.; Rudnick, P. A.; Tabb, D. L.; Wang, P.; Blackman, R. K.; Bunk, D. M.; Cardasis, H. L.; Clauser, K. R.; Kinsinger, C. R.; Schilling, B.; Tegeler, T. J.; Variyath, A. M.; Wang, M.; Whiteaker, J. R.; Zimmerman, L. J.; Fenyo, D.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Mesri, M.; Neubert, T. A.; Reginier, F. E.; Rodriguez, H.; Spiegelman, C.; Stein, S. E.; Tempst, P.; Liebler, D. C., A CPTAC inter-laboratory study characterizing a yeast performance standard for benchmarking LC-MS Platform performance. *Mol. Cell. Proteomics* **2009**, *9*, 242−254.
(23) Tabb, D.; Vega-Montoto, L.; Rudnick, P.; Mulayath Variyath, A.; Ham, A.; Bunk, D.; LE, K.; Billheimer, D.; Blackman, R.; Cardasis, H.; Carr, S. A.; Clauser, K.; Jaffe, J.; Kowalski, K.; Neubert, T.; Regnier, F.; Schilling, B.; Tegeler, T.; Wang, M.; Wang, P.; Whiteaker, J.; Zimmerman, L.; Fisher, S.; Gibson, B.; Kinsinger, C.; Mesri, M.; Rodriguez, H.; Stein, S.; Tempst, P.; Paulovich, A.; Liebler, D.; Spiegelman, C. Repeatability and reproducibility in proteomic analyses by liquid chromatography - tandem mass spectrometry. *J. Proteome Res.* **2010**, *9* (2), 761–76.
(24) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **2007**, *6* (2), 654–61.
(25) Ma, Z. Q.; Dasari, S.; Chambers, M. C.; Litton, M. D.; Sobecki, S. M.; Zimmerman, L. J.; Halvey, P. J.; Schilling, B.; Drake, P. M.; Gibson, B. W.; Tabb, D. L. IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. *J. Proteome Res.* **2009**, *8* (8), 3872–81.
(26) Zybailov, B.; Coleman, M. K.; Florens, L.; Washburn, M. P. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal. Chem.* **2005**, *77* (19), 6218–24.
(27) Fu, X.; Gharib, S. A.; Green, P. S.; Aitken, M. L.; Frazer, D. A.; Park, D. R.; Vaisar, T.; Heinecke, J. W. Spectral index for assessment of differential protein expression in shotgun proteomics. *J. Proteome Res.* **2008**, *7* (3), 845–54.
(28) Nelder, J.; Wedderburn, R. Generalized linear model. *J. R. Stat. Soc.* **1972**, *132*, 370–84.

(29) Faraway, J. *Extending linear model with R: generalized linear, mixed effects and nonparametric regression models*; Taylor and Francis: Boca Raton, FL, 2007.
(30) McCullagh, P. Quasi-likelihood functions. *Ann. Stat.* **1983**, *11*, 59–67.
(31) Pham, T. V.; Piersma, S. R.; Warmoes, M.; Jimenez, C. R. On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics* **2009**, *26* (3), 363–9.
(32) Moore, D. F.; Tsiatis, A. Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. *Biometrics* **1991**, *47* (2), 383–401.
(33) Cortes, H. J.; Pfeiffer, C. D.; Richter, B. E.; Stevens, T. Porous ceraminc bed supports for fused silica packed capillary columns used in liquid chromatography. *J. High Res. Chromatogr. Chromatogr. Commun.* **1987**, *10*, 446–8.
(34) Licklider, L. J.; Thoreen, C. C.; Peng, J.; Gygi, S. P. Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column. *Anal. Chem.* **2002**, *74* (13), 3076–83.
(35) Zhang, B.; Chambers, M. C.; Tabb, D. L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **2007**, *6* (9), 3549–57.
(36) Haase, H. Ahnak, a new player in beta-adrenergic regulation of the cardiac L-type Ca2+ channel. *Cardiovasc. Res.* **2007**, *73* (1), 19–25.
(37) Addona, T. A.; Abbatiello, S. E.; Schilling, B.; Skates, S. J.; Mani, D. R.; Bunk, D. M.; Spiegelman, C. H.; Zimmerman, L. J.; Ham, A. J.; Keshishian, H.; Hall, S. C.; Allen, S.; Blackman, R. K.; Borchers, C. H.; Buck, C.; Cardasis, H. L.; Cusack, M. P.; Dodder, N. G.; Gibson, B. W.; Held, J. M.; Hiltke, T.; Jackson, A.; Johansen, E. B.; Kinsinger, C. R.; Li, J.; Mesri, M.; Neubert, T. A.; Niles, R. K.; Pulsipher, T. C.; Ransohoff, D.; Rodriguez, H.; Rudnick, P. A.; Smith, D.; Tabb, D. L.; Tegeler, T. J.; Variyath, A. M.; Vega-Montoto, L. J.; Wahlander, A.; Waldemarson, S.; Wang, M.; Whiteaker, J. R.; Zhao, L.; Anderson, N. L.; Fisher, S. J.; Liebler, D. C.; Paulovich, A. G.; Regnier, F. E.; Tempst, P.; Carr, S. A. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol.* **2009**, *27* (7), 633–41.
(38) Xu, X. C.; Lee, J. S.; Lippman, S. M.; Ro, J. Y.; Hong, W. K.; Lotan, R. Increased expression of cytokeratins CK8 and CK19 is associated with head and neck carcinogenesis. *Cancer Epidemiol. Biomarkers Prev.* **1995**, *4* (8), 871–6.
(39) Chu, P. G.; Weiss, L. M. Keratin expression in human tissues and neoplasms. *Histopathology* **2002**, *40* (5), 403–39.
(40) Toyoshima, T.; Koch, F.; Kaemmerer, P.; Vairaktaris, E.; Al-Nawas, B.; Wagner, W. Expression of cytokeratin 17 mRNA in oral squamous cell carcinoma cells obtained by brush biopsy: preliminary results. *J. Oral Pathol. Med.* **2009**, *38* (6), 530–4.
(41) Madsen, P.; Rasmussen, H. H.; Leffers, H.; Honore, B.; Celis, J. E. Molecular cloning and expression of a novel keratinocyte protein (psoriasis-associated fatty acid-binding protein [PA-FABP]) that is highly up-regulated in psoriatic skin and that shares similarity to fatty acid-binding proteins. *J. Invest Dermatol.* **1992**, *99* (3), 299–305.
(42) Siegenthaler, G.; Hotz, R.; Chatellard-Gruaz, D.; Didierjean, L.; Hellman, U.; Saurat, J. H. Purification and characterization of the human epidermal fatty acid-binding protein: localization during epidermal cell differentiation in vivo and in vitro. *Biochem. J.* **1994**, *302* (Pt 2), 363–71.
(43) Munz, M.; Zeidler, R.; Gires, O. The tumour-associated antigen EpCAM upregulates the fatty acid binding protein E-FABP. *Cancer Lett.* **2005**, *225* (1), 151–7.
(44) Melle, C.; Ernst, G.; Winkler, R.; Schimmel, B.; Klussmann, J. P.; Wittekindt, C.; Guntinas-Lichius, O.; von Eggeling, F. Proteomic analysis of human papillomavirus-related oral squamous cell carcinoma: identification of thioredoxin and epidermal-fatty acid binding protein as upregulated protein markers in microdissected tumor tissue. *Proteomics* **2009**, *9* (8), 2193–201.
(45) Fujimoto, W.; Nakanishi, G.; Arata, J.; Jetten, A. M. Differential expression of human cornifin alpha and beta in squamous differentiating epithelial tissues and several skin lesions. *J. Invest. Dermatol.* **1997**, *108* (2), 200–4.
(46) Zhang, B.; VerBerkmoes, N. C.; Langston, M. A.; Uberbacher, E.; Hettich, R. L.; Samatova, N. F. Detecting differential and correlated protein expression in label-free shotgun proteomics. *J. Proteome Res.* **2006**, *5* (11), 2909–18.

PR100527G