The Author(s). *BMC Genomics* 2016, **17**(Suppl 4):545
DOI 10.1186/s12864-016-2796-x

BMC Genomics

RESEARCH Open Access

Topic model-based mass spectrometric data analysis in cancer biomarker discovery studies

Minkun Wang[1,2], Tsung-Heng Tsai[1], Cristina Di Poto[1], Alessia Ferrarini[1], Guoqiang Yu[2] and Habtom W. Ressom[1*]

From IEEE International Conference on Bioinformatics and Biomedicine 2015
Washington, DC, USA. 9-12 November 2015

Abstract

Background: A fundamental challenge in quantitation of biomolecules for cancer biomarker discovery is owing to the heterogeneous nature of human biospecimens. Although this issue has been a subject of discussion in cancer genomic studies, it has not yet been rigorously investigated in mass spectrometry based proteomic and metabolomic studies. Purification of mass spectometric data is highly desired prior to subsequent analysis, e.g., quantitative comparison of the abundance of biomolecules in biological samples.

Methods: We investigated topic models to computationally analyze mass spectrometric data considering both integrated peak intensities and scan-level features, i.e., extracted ion chromatograms (EICs). Probabilistic generative models enable flexible representation in data structure and infer sample-specific pure resources. Scan-level modeling helps alleviate information loss during data preprocessing. We evaluated the capability of the proposed models in capturing mixture proportions of contaminants and cancer profiles on LC-MS based serum proteomic and GC-MS based tissue metabolomic datasets acquired from patients with hepatocellular carcinoma (HCC) and liver cirrhosis as well as synthetic data we generated based on the serum proteomic data.

Results: The results we obtained by analysis of the synthetic data demonstrated that both intensity-level and scan-level purification models can accurately infer the mixture proportions and the underlying true cancerous sources with small average error ratios ($< 7\ \%$) between estimation and ground truth. By applying the topic model-based purification to mass spectrometric data, we found more proteins and metabolites with significant changes between HCC cases and cirrhotic controls. Candidate biomarkers selected after purification yielded biologically meaningful pathway analysis results and improved disease discrimination power in terms of the area under ROC curve compared to the results found prior to purification.

Conclusions: We investigated topic model-based inference methods to computationally address the heterogeneity issue in samples analyzed by LC/GC-MS. We observed that incorporation of scan-level features have the potential to lead to more accurate purification results by alleviating the loss in information as a result of integrating peaks. We believe cancer biomarker discovery studies that use mass spectrometric analysis of human biospecimens can greatly benefit from topic model-based purification of the data prior to statistical and pathway analyses.

Keywords: Bayesian inference, Topic model, Purification, LC-MS, GC-MS, Extracted ion chromatogram, Metabolomics, Proteomics, Biomarker discovery

*Correspondence: hwr@georgetown.edu
[1] Department of Oncology, Georgetown University, 4000 Reservoir Rd NW, Washington D.C., USA
Full list of author information is available at the end of the article

Background

Identification of disease-related alterations in molecular and cellular mechanisms may reveal useful disease biomarkers. Discovery of clinically relevant biomarkers has potentially far reaching implications for disease management and patient treatment [1–4]. High-throughput omic technologies have facilitated the search for changes in the levels of various biomolecules (proteins, glycoproteins, metabolites, lipids, etc.) in biological samples [5, 6]. In particular, liquid (or gas) chromatography coupled with mass spectrometry (LC/GC-MS) has become an essential tool for profiling biomolecules in a variety of large-scale omic studies. Briefly, biomolecules are separated, fragmented, ionized and detected in LC/GC-MS instruments. Abundances of ions with various retention time and mass values are recorded for downstream data processing.

While the capability of high-throughput technology to yield comprehensive profiling and quantification offers a unique advantage in biomedical research, the heterogeneous nature of the biological samples poses a fundamental challenge in data analysis and interpretation. Specimens, such as tumor tissues and human blood, are typically mixtures of cells with distinct states and types, and usually only part of the constituent cell populations is relevant to the biological question of interest [7, 8]. In some cancer studies, heterogeneity is also observed within the malignant cell population, where multiple cancerous subtypes co-exist [9]. Ideally in a biomarker discovery study, one would perform between-group (cancer versus related disease, cancer versus healthy samples) differential expression analysis for type-specific constituents in samples [10]. However, biospecimens collected from patients usually exhibit some degree of heterogeneity. Moreover, the proportion of cancerous, other disease-related, and healthy components varies across individual samples pre-selected using pathological estimates. Therefore, the biomolecular measurements in expression profiles are attributed to multiple sites of origins with various mixture proportions. The cancerous profiles of interest are typically contaminated by other components, leading to unreliable results in differential analyses. Purification of samples is hence highly desired to remove the effects of heterogeneity.

Experimental methods for cleaning samples and isolating type-specific constituents are costly and time-consuming. Computational purification methods offer an attractive alternative that is inexpensive and efficient to implement, and can be applied to data already generated without any modifications on experimental procedures. Multiple approaches have been developed to deconvolute gene expression profiles in the past years, varying from linear regression based models [11, 12] to generative probabilistic models [13, 14]. Among these approaches, topic model based methods, e.g., ISOLATE [15] and ISOpure [8], showed promising performance in estimating the proportion of mixtures and inferring sample-specific purified profiles in heterogeneous genomic data. However, to the best of our knowledge, in omic studies involving quantitative analysis of proteins or metabolites, no such purification approaches have been applied to deal with the sample heterogeneity issue. With the increasing volume of these data generated by LC/GC-MS, it is necessary to implement the purification of data before downstream differential analyses. In this research, we first apply ISOpure, a topic model based purification approach to both synthetic and experimental data acquired from human sera and liver tissues by LC-MS and GC-MS, respectively. The purpose of this investigation is to test if sample heterogeneity issue in various biomolecular expression profiles can be addressed by adjusting ion intensities through topic models as in genomic studies. Also, we investigate the use of scan-level features, i.e. extracted ion chromatograms (EICs) instead of integrated peak intensities, to alleviate the information loss during the LC/GC-MS data preprocessing.

Methods

In this section, we introduce topic model-based intensity-level and scan-level purification methods. Assumptions and strategies in the topic models are elaborated. Mass spectrometric datasets from cancer biomarker discovery studies are described.

Intensity-level purification model

The LC/GC-MS instruments provide ion intensity values by counting the ions detected at specific m/z and retention time points. Due to the existence of heterogeneity, multiple constituents in the sample contribute to the observed expression profile. Therefore, we can model the expression profile of a heterogeneous sample $\boldsymbol{t}$ as a weighted mixture of expression profiles of multiple sources, including a cancerous origin $\boldsymbol{\gamma}$ and non-cancerous contaminants $\boldsymbol{\beta}$. The expression distribution for every biomolecule in each of the sources plays a role as a "topic" contributing to the mixed expression profile. Basically, each ion in the observed profile is associated with a specific topic, i.e. a multinomial distribution of ion counts over biomolecules, determined by the corresponding source profile. In this model, expression profiles refer to integrated peak intensities.

The purification procedure can be realized through a set of topic models, which are generative probabilistic models typically applied to text corpora mining. Specifically, each expression profile is characterized by a probability distribution across topics. Topics are probability distributions across biomolecules. These distributions can be inferred based on the analysis of a collection of expression profiles through topic models. These hierarchical Bayesian models are variants of latent Dirichlet allocation (LDA) [16], a topic model that can 1) infer the posterior probability of topics given observed profiles, and 2) estimate the parameters that generate the latent mixture proportion and topic panel. These topic models have been adapted and applied to gene expression profiles in genomic studies [8, 15].

We use a modified topic model to purify the molecular expression profiles in cancer. Basically, three assumptions are made in developing the model. First, the source contaminants in each expression profile $\{\boldsymbol{t}_d\}_{d=1,\cdots,D}$ are coming from the control group $\{\boldsymbol{\beta}_m\}_{m=1,\cdots,M}$ (i.e., healthy, non-cancerous profiles, etc.). It is commonly observed that the cancerous tissues are surrounded by adjacent non-cancerous tissues, which are typically used as controls in differential expression analysis. Second, the corresponding cancerous origins $\{\boldsymbol{\gamma}_d\}_{d=1,\cdots,D}$ share an average cancer profile $\boldsymbol{\gamma'}$. Individual cancerous profile can be treated as a noisy version of the average cancer profile. Third, the average cancer profile $\boldsymbol{\gamma'}$ has similar patterns as non-cancerous profiles $\{\boldsymbol{\beta}\}$, except for some sites (biomolecules) which are differentially expressed between case and control groups. Thus, the cancerous profile can be treated as a similar non-cancerous profile with several sites altered.

The complete likelihood function in (1) describes how the profiles $\{\boldsymbol{t}_d\}_{(d=1,\ldots,D)}$ are generated. Specifically, we have two observable variables indicating D expression profiles in case group: $\{\boldsymbol{t}_d\}_{d=1,\cdots,D}$, $\boldsymbol{t}_d \in \mathbb{R}^N$, and M non-cancerous profiles in control group: $\{\boldsymbol{\beta}_m\}_{m=1,\cdots,M}$, $\boldsymbol{\beta}_m \in \mathbb{R}^L$. In our analysis, we normalize all profiles to have identical total ion counts of N and consider L biomolecules that are consistently detected in all the samples. For convenience, we represent the normalized profiles in two ways. Each heterogeneous cancer profile $\boldsymbol{t}_d$ is represented via N ions, with $t_{d,n} = \{1, 2, \cdots, L\}$ denoting the biomolecule corresponding to the n^{th} ion. Each non-cancerous profile $\boldsymbol{\beta}_m$ is represented via L biomolecules, with $\beta_{m,l}$ denoting the ion counts of the l^{th} biomolecule, and $\sum_{l=1}^{L} \beta_{m,l} = N$. The second expression can be further normalized by N to give a representation of multinomial distribution as a topic.

$$\begin{aligned}\mathcal{L}(\boldsymbol{t}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\gamma'} | \alpha, \boldsymbol{\beta}, \boldsymbol{\eta}, \kappa, \kappa') &\\ = p\left(\boldsymbol{\gamma'} | \boldsymbol{\beta}, \boldsymbol{\eta}, \kappa'\right) \cdot \prod_{d=1}^{D} p\left(\boldsymbol{\theta}_d | \alpha\right) \cdot p\left(\boldsymbol{\gamma}_d | \boldsymbol{\gamma'}, \kappa_d\right) &\\ \times \prod_{n=1}^{N} \left[p\left(z_{d,n} | \boldsymbol{\theta}_d\right) \cdot p\left(t_{d,n} | z_{d,n}, \boldsymbol{\theta}_d, \boldsymbol{\beta}, \boldsymbol{\gamma}_d\right)\right] &\end{aligned} \tag{1}$$

The model also incorporates the following latent variables: the average cancer profile $\boldsymbol{\gamma}' \in \mathbb{R}^L$, sample-specific pure cancer profiles $\{\boldsymbol{\gamma}_d\}_{d=1,\cdots,D}$, $\boldsymbol{\gamma}_d \in \mathbb{R}^L$, sample-specific mixture proportions of topics $\{\boldsymbol{\theta}_d\}_{d=1,\cdots,D}, \boldsymbol{\theta}_d \in \mathbb{R}^{M+1}$, and sample-specific topic indicators $\{\boldsymbol{z}_d\}_{d=1,\cdots,D}$, $\boldsymbol{z}_d \in \mathbb{R}^N$, $z_{d,n} = \{1, \cdots, M, M+1\}$. Their relationships with observations and parameters are given as below.

$$p\left(\boldsymbol{\theta}_d | \boldsymbol{\alpha}\right) = \text{Dirichlet}\left(\boldsymbol{\theta}_d | \boldsymbol{\alpha}, 1\right) \tag{2}$$

$$p\left(\boldsymbol{\gamma}' | \boldsymbol{\beta}, \eta, \kappa'\right) = \text{Dirichlet}\left(\boldsymbol{\gamma}' | \boldsymbol{\eta}^T \boldsymbol{\beta}, \kappa'\right) \tag{3}$$

$$p\left(\boldsymbol{\gamma}_d | \boldsymbol{\gamma}', \kappa_d\right) = \text{Dirichlet}\left(\boldsymbol{\gamma}_d | \boldsymbol{\gamma}', \kappa_d\right) \tag{4}$$

$$p\left(z_{d,n} | \boldsymbol{\theta}_d\right) = \text{Multinomial}\left(z_{d,n} | \boldsymbol{\theta}_d\right) \tag{5}$$

$$p\left(t_{d,n} | z_{d,n} \leq M, \boldsymbol{\theta}_d, \boldsymbol{\beta}, \boldsymbol{\gamma}_d\right) = \text{Multinomial}\left(t_{d,n} | \boldsymbol{\beta}_{z_{d,n}}\right) \tag{6}$$

$$p\left(t_{d,n} | z_{d,n} = M+1, \boldsymbol{\theta}_d, \boldsymbol{\beta}, \boldsymbol{\gamma}_d\right) = \text{Multinomial}\left(t_{d,n} | \boldsymbol{\gamma}_d\right) \tag{7}$$

The average cancer profile $\boldsymbol{\gamma}'$ is sampled from a Dirichlet distribution parameterized by a weighted mixture of non-cancerous profiles. Each pure cancer profile $\boldsymbol{\gamma}_d$ together with M contaminants $\{\boldsymbol{\beta}_m\}$ forms a sample-specific topic panel. The mixture proportion of topics determines $z_{d,n}$, indicating which source (i.e., γ_d or $\{\boldsymbol{\beta}_m\}$) each ion originates from. We infer the latent variables $\boldsymbol{\gamma}', \{\boldsymbol{\gamma}_d\}_{d=1,\cdots,D}, \{\boldsymbol{\theta}_d\}_{d=1,\cdots,D}$, and estimate the parameters using the two-step learning approach developed based on variational EM algorithm (ISOpure package [8], version 1.4). The graphical model representing the above topic model is shown in Fig. 1. This three-level model

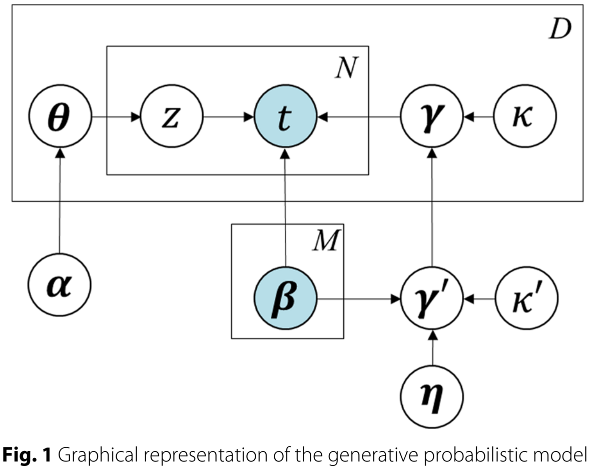

Fig. 1 Graphical representation of the generative probabilistic model. Hyperparameters $\boldsymbol{\eta}$, κ' together with sources of contaminants $\{\boldsymbol{\beta}_m\}$ determine an average cancer profile $\boldsymbol{\gamma'}$. Each of the D profiles is associated with a mixture proportion $\boldsymbol{\theta}_d$ (regularized by hyperparameter $\boldsymbol{\alpha}$) and a topic panel consisting of $\{\boldsymbol{\beta}_m\}$ and γ' (generated from the average cancer profile). Each of the N ions in a profile $t_{n,d}$ is sampled from a topic indicated by $z_{n,d}$

allows a single profile to be associated with multiple topics (i.e., cancerous and non-cancerous origins). Such property of the LDA-framed models enable more flexible representation in data structure than that by other unigram models or mixture of unigrams [16]. Also in contrast to linear regression models, these methods use a multinomial noise model that is a better fit to noise measurement in biomolecular expression data [13].

Scan-level purification model

Here, we extend the topic model to utilize the scan-level measurements instead of the integrated peak intensities. During LC/GC-MS data preprocessing, ion intensity is obtained by integrating the scan-level measurements of a detected chromatographic peakl within a specified retention time (RT) interval. This integration or truncation, however, inevitably brings in variances which interfere with original sample heterogeneity. Therefore, we propose to investigate LC/GC-MS data purification with scan-level measurements based on extracted ion chromatogram (EIC), which preserves scan-level peak shape information. We hypothesize that purification at the scan level leads to more accurate results and offers the opportunity to extend the model to characterize both ion abundance and peak shape.

After ion tracing and missing value interpolation, we can obtain a list of EICs for each sample. EIC is characterized by its retention time (corresponding to multiple scans), mass value, and ion abundance. In this scenario, the observed data $\{\boldsymbol{t}_d\}$ (same for $\{\boldsymbol{\beta}_m\}$) consists of multiple EIC peaks. Each peak is represented by ion abundances across S scans with a certain elution profile shape $\mathcal{F}(\cdot)$ as shown in Fig. 2. Using these scan-level features, we model each EIC peak as shown in Eq. (8):

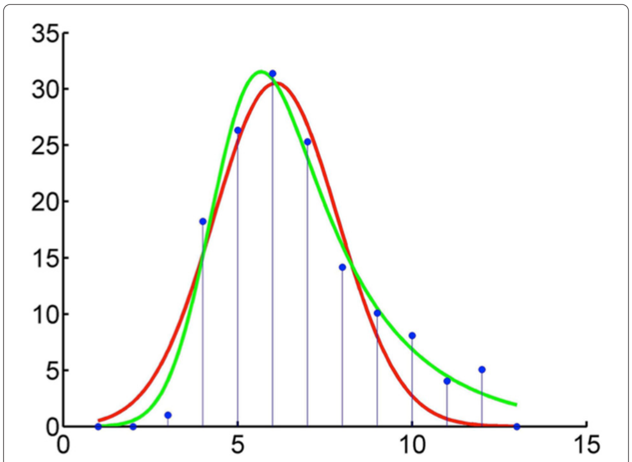

Fig. 2 Extracted ion chromatography and peak shape function. Example of Gaussian (*red*) and exponentially modified Gaussian (*green*) peak shapes fitted to an experimental EIC involving 13 scans (*blue*)

$$t_{d,n}(s) = x_{d,n} \cdot \delta_{d,n}(s) \cdot \mathcal{F}\left(s, \boldsymbol{\phi}_{d,n}\right) + e_{d,n}(s),\ s = 1, \cdots, S \tag{8}$$

where, $x_{d,n}$ is the ion abundance for n^{th} compound of d^{th} sample; $\delta_{d,n}(s)$ is a latent indicator to model the missing scans; the chromatographic peak shape is characterized by the exponentially modified Gaussian (EMG) function [17] parameterized by $\boldsymbol{\phi}$, as described in Eq. (9), and $e_{d,n}(s)$ is the random noise.

$$\mathcal{F}(s, \boldsymbol{\phi}) = \frac{1}{2}\zeta \exp\left(\frac{1}{2}\zeta\left(2\mu + \zeta\sigma^2 - 2s\right)\right) \times \left(1 - \operatorname{erf}\left(\frac{\mu + \zeta\sigma^2 - s}{\sqrt{2}\sigma}\right)\right),\ \boldsymbol{\phi} \doteq \{\mu, \zeta, \sigma\} \tag{9}$$

We hypothesize that the data heterogeneity in $t_{d,n}$ corresponds to the shape of the EIC (characterized by $\boldsymbol{\phi}$) as well as ion abundance $x_{d,n}$.

We extend the purification model we used for integrated peaks by adding a lower layer to characterize the scan-level information as illustrated in Fig. 3. The three assumptions are maintained in this model in terms of the dependancy of ion abundance variables. That is, Eqs. (2)–(7) still hold for ion abundances $\boldsymbol{x}_t$, $\boldsymbol{x}_\beta$, $\boldsymbol{x}'_\gamma$, and $\boldsymbol{x}_\gamma$. We assume error terms in intensity measurements in Eq. (8) are independent random variables generated by a normal distribution with conjugate prior following an inverse-Gamma distribution:

$$e_{d,n}(s)|\sigma^2_{e_d} \sim \mathcal{N}\left(0, \sigma^2_{e_d}\right), \quad \sigma^2_{e_d} \sim \mathcal{IG}\left(a_e, b_e\right). \tag{10}$$

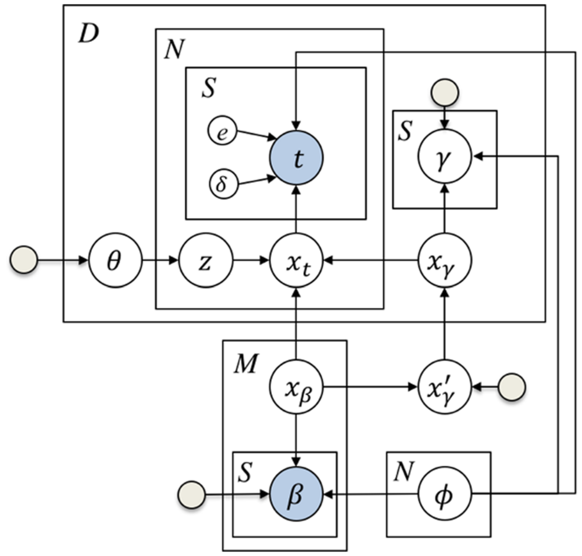

Fig. 3 Graphical representation of the scan-level topic model. A lower layer to characterize the scan-level information is added. Ion abundances $\boldsymbol{x}_t$, $\boldsymbol{x}_\beta$, $\boldsymbol{x}'_\gamma$, and $\boldsymbol{x}_\gamma$ together with peak shape (parameterized in $\boldsymbol{\phi}$) determined the observed feature list $\boldsymbol{t}$, $\boldsymbol{\beta}$

The missing scan indicator variable $\delta_{d,n}(s)$ follows a Bernoulli distribution, parameterized by q_d with a prior of Beta distribution:

$$p\left(\delta_{d,n}(s)|q_d\right) = \text{Bernoulli}\left(\delta_{d,n}(s)|q_d\right), \quad p\left(q_d|a_q, b_q\right) = \text{Beta}\left(q_d|a_q, b_q\right). \tag{11}$$

The observed data point therefore follows the distribution:

$$t_{d,n}(s)|x_{td,n}, q_d, \phi_{d,n}, \sigma_{e_d}^2 \sim q_d \mathcal{N}\left(x_{td,n}\mathcal{F}(s, \phi_{d,n}), \sigma_{e_d}^2\right) + (1 - q_d)\mathcal{N}\left(0, \sigma_{e_d}^2\right). \tag{12}$$

The peak shape parameters $\boldsymbol{\phi}$ are considered to have a normal distribution and its detailed priors are described in [17]. The extended model contains variables that are mutually coupled, providing no analytical form for the posterior distribution in calculation. As a variational approximation, we can split the model into two components: 1) mixture model of underlying ion abundances, and 2) scan-level feature generation. We adopt a two-phase approach to iteratively update the latent variables and estimate the parameters between the two parts. Specifically, we use a Markov chain Monte Carlo (MCMC) sampling method [17] to infer the peak shape model parameters of the second part (i.e., ion abundance $\boldsymbol{x}_t$, $\boldsymbol{x}_\beta$, and shape function parameters ϕ). We then treat $\boldsymbol{x}_t$, $\boldsymbol{x}_\beta$ as observed variables to implement the inference on the first part using the same algorithm [8] employed in the intensity-level purification. Once converged, the model outputs the sample-specific mixture proportion $\boldsymbol{\theta}$, pure ion abundance $\boldsymbol{x}_\gamma$, shape function parameters $\boldsymbol{\phi}$ and related parameters. After purification is performed, ion intensity may be calculated by applying peak detection algorithms [18, 19] to the pure EIC peaks $\{\boldsymbol{\gamma}_{d,n}\}$ recovered using Eq. (8).

Mass spectrometric datasets

The experimental data were acquired by analyses of tissue and blood samples from patients with hepatocellular carcinoma (i.e., HCC, case group) and liver cirrhosis (control group) [1–4]. HCC is a highly heterogeneous disease both at the molecular and clinical levels [20]. Whereas all patients in this study were diagnosed with liver cirrhosis, about half of them were also diagnosed with HCC. Contamination occurs due to the influence from cirrhotic constituents in HCC samples. In this study, we used GC-MS data acquired by analysis of metabolites in 15 tissues and LC-MS data acquired by analysis of proteins in sera from 116 subjects.

GC-MS based metabolomic dataset

Fifteen liver tissues were collected from 10 participants recruited at MedStar Georgetown University Hospital. As shown in Fig. 4, the tissues were collected from 5 HCC cases (5 tumor and 5 adjacent cirrhotic tissues) and 5 patients with liver cirrhosis. Samples were profiled through Agilent 7890A gas chromatography coupled with LECO's time-of-flight mass spectrometer to characterize the metabolome alterations associated with HCC development in cirrhotic patients. We identified 559 metabolites after preprocessing the GC-MS raw data by ChromaTOF GC software with True Signal Deconvolution package (Leco Corporation). Two types of purification are investigated on the data. One is to purify HCC profiles by removing contaminants from cirrhotic profiles. The other is to purify adjacent cirrhotic profiles by reducing the impact of the profiles attributed to HCC.

LC-MS based proteomic dataset

We acquired 116 proteomic data by analysis of sera from 57 HCC cases and 59 patients with liver cirrhosis recruited from the hepatology clinics at MedStar Georgetown University Hospital. Following depletion and digestion, proteins extracted from sera were injected into a 3000 Ultimate nano-LC system interfaced to LTQ Orbitrap Velos and TSQ Vantage mass spectrometers in untargeted and targeted analyses, respectively. Proteins were

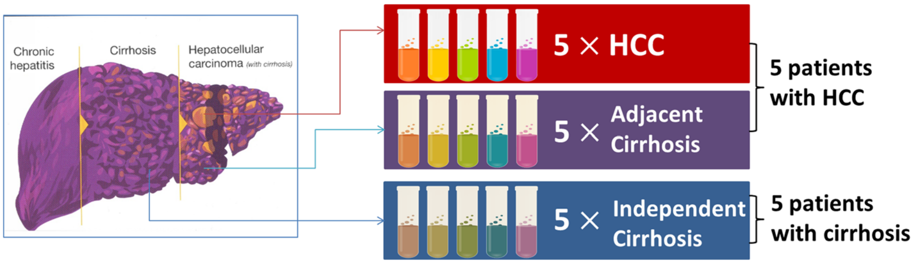
Chronic hepatitis
Cirrhosis
Hepatocellular carcinoma (with cirrhosis)
5 × HCC
5 × Adjacent Cirrhosis
5 × Independent Cirrhosis
5 patients with HCC
5 patients with cirrhosis

Fig. 4 Fifteen tissue samples collected from 10 subjects (5 HCC cases and 5 cirrhotic controls). Five tumor and five adjacent cirrhotic tissues were obtained from the 5 HCC cases. Additional 5 cirrhotic tissues were obtained from the 5 independent subjects with liver cirrhosis

identified and quantified by MaxQuant [21] and Skyline [22] in preprocessing untargeted and targeted LC-MS data, respectively. Finally, 101 proteins that were consistently identified across 116 samples were selected as intensity-level features in expression profiles (i.e., $L = 101$). All profiles were normalized to the mean total-ion-counts at $N = 1.68 \times 10^8$. It is still not clear how the development of tumor in liver directly affect the alterations in blood. We hypothesize that there are some impacts from cirrhotic constituents contributing to the HCC profile in serum. The contamination may occur in an indirect way through, for example, secreted biomolecules instead of adjacent tissue cells. We apply the purification to remove the influence from cirrhotic contaminants.

Synthetic datasets

Before applying the models to experimental data, we generated synthetic datasets by artificially mixing real LC-MS data on both intensity and scan levels, and evaluated the model based on its performance of deconvolving the mixed data. We generated synthetic data based on the 116 LC-MS profiled serum proteomic dataset. We assume here that human sera are homogeneous specimens. Hence we can mix them to simulate heterogeneous cancer profiles. Figure 5 shows the generative process of 30 synthetic cancer profiles with contamination, following the steps below:

(i) Average the profiles of HCC group, $\{\boldsymbol{\lambda}_s\}_{s=1,\cdots,57}$, to obtain an average cancer profile $\boldsymbol{\gamma}'$, which is close to the real cancerous profile for HCC.

(ii) Sample 30 individual pure cancer profiles $\{\boldsymbol{\gamma}_d\}_{d=1,\cdots,30}$ from a Dirichlet distribution, as in (4), parameterized by $\boldsymbol{\gamma}'$ and $\kappa_d = \frac{1}{\min_l(\gamma'_l)}$.

(iii) Randomly select a subset of cirrhotic profiles $\{\boldsymbol{\beta}_m\}_{m=1,\cdots,M}$ ($M = 9$ in this simulation) as sources of contamination. Normalize them into form of multinomial distribution.

(iv) Combine M cirrhotic profiles with each of the pure cancer profiles to create 30 topic panels, each consisting of $M + 1 = 10$ profiles.

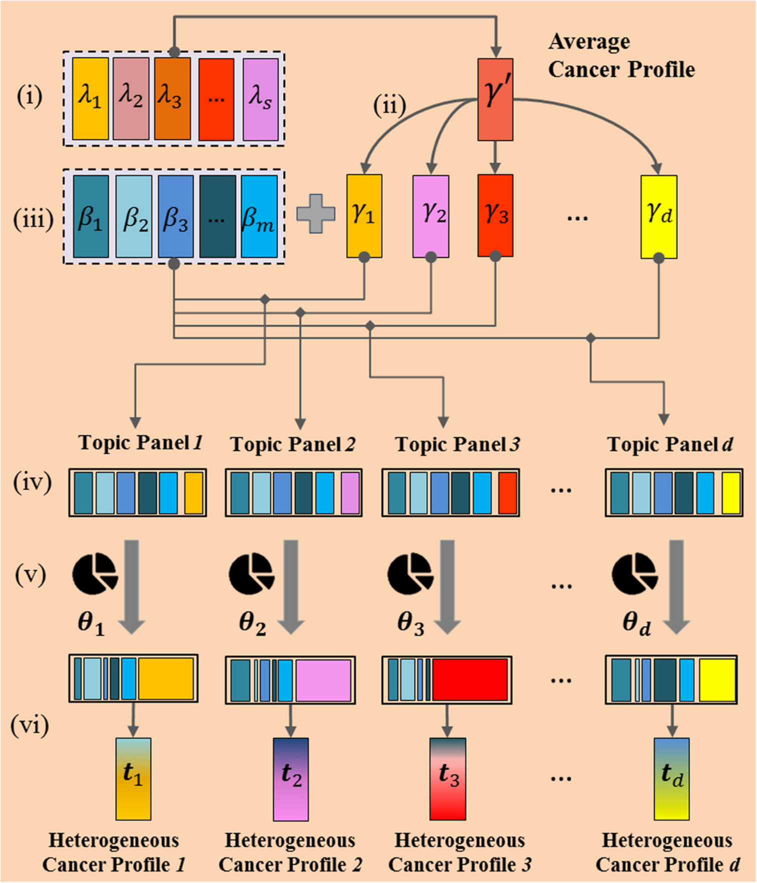

Fig. 5 Generative process of heterogeneous cancer profiles. (*i*) average cancer profiles in case group; (*ii*) generate sample-specific pure cancer profile; (*iii*) select sources of contaminants in control group; (*iv*) form topic panels; (*v*) generate sample-specific mixture proportions; (*vi*) generate synthetic cancer profiles

(v) Sample 30 mixture proportions $\{\boldsymbol{\theta}_d\}_{d=1,\cdots,30}$ from a Dirichlet distribution, as in (2), parameterized by $\boldsymbol{\alpha} = [1, \cdots, 1, 5]$, which is uniform for the first nine constituents (contaminants) and with a larger value assigned to last constituent (cancer origin). This ensures a larger proportion of cancerous component in final cancer profile.

(vi) Sample a topic indicator $z_{d,n}$ from $\boldsymbol{\theta}_d$ using (5), and sample a $t_{d,n}$ from $\boldsymbol{\beta}_z$ if $z \leq M$ or $\boldsymbol{\gamma'}_d$ otherwise, as in (6), (7). Repeat the sampling for $N = 1.68 \times 10^8$ times to generate a synthetic cancer profile $\boldsymbol{t}_d$.

Each of these 30 heterogeneous cancer profiles is a mixture of a pure cancer profile and multiple contaminants. The intensity-level purification procedure will help retrieve the pure cancer profile and estimate the sample purity as well as proportions of contaminants. Similar to intensity-level simulation, we generated heterogeneous dataset using scan-level features, i.e. EICs, exported from Skyline [22]. Corresponding to 101 proteins, 187 peptides with 561 scan-level features were extracted in each of the 116 samples. Each feature contains 60 scans representing a chromatographic peak as illustrated in Fig. 6. We followed the same steps (i-vi) except that we average and blend EIC peaks instead of protein intensities. Finally, 30 heterogeneous cancerous samples, each characterized by a list of 561 EICs, are generated.

Evaluation methods

We evaluated the performances of our proposed models on both synthetic and real experimental LC/GC-MS datasets in consideration of the following three goals: 1) to test on intensity level if the model can reasonably estimate the proportion of mixtures in each of the synthetic profiles and recover the pure cancer profiles underneath; 2) to demonstrate if the scan-level purification model gives more accurate estimation on synthetic data; 3) to investigate the benefits of using these models to purify samples from cancer patients collected in our previous differential analysis studies.

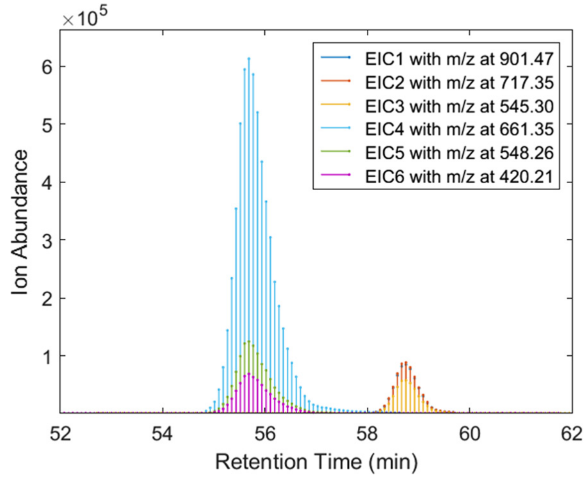

Fig. 6 Extracted ion chromatograms from LC-MS based serum proteomic data. Extracted ion chromatogram is characterized by *m*/*z*, retention time, and ion abundance

Outputs of intensity-level model include the sample-specific mixture proportions $\{\boldsymbol{\theta}^*_d\}$, pure cancer profiles $\{\boldsymbol{\gamma}^*_d\}$, and the estimated average cancer profile $\boldsymbol{\gamma'}^*$. Whereas, we expect outputs of sample-specific mixture proportion $\{\boldsymbol{\theta}^*_d\}$, pure ion abundance $\{\boldsymbol{x}^*_\gamma\}$, peak shape function parameters $\boldsymbol{\phi}^*$ from extended model. For synthetic datasets, we compare the estimated proportions of mixtures $\{\boldsymbol{\theta}^*_d\}$ with the true ones ($\{\boldsymbol{\theta}_d\}$) used to generate the synthetic data. Estimation error ratio for a single sample is defined in Eq. (13).

$$\xi_d\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}\right) = \frac{\|\boldsymbol{\theta}^*_d - \boldsymbol{\theta}_d\|_1}{\|\boldsymbol{\theta}_d\|_1} \times 100\,\%,\ d = 1, \cdots, 30 \quad (13)$$

Different from point-wise intensities, the scan-level estimation error ratio for a single sample is defined in Eq. (14)

$$\xi_d\left(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}\right) = \frac{\left\| \sum_{s=1}^{S} \left[\boldsymbol{\gamma}^*_d(s) - \boldsymbol{\gamma}_d(s)\right] \right\|_1}{\left\| \sum_{s=1}^{S} \boldsymbol{\gamma}_d(s) \right\|_1} \times 100\,\%,\ d = 1, \cdots, 30 \quad (14)$$

For experimental datasets, we evaluated the performances in multiple aspects including statistical significance of the candidate biomarkers, ROC curves in distinguishing the biological groups, and pathway analysis results.

Results and discussions

Synthetic datasets

We applied current model and the extended model to the synthetic intensity-level and scan-level LC-MS datasets, respectively. By incorporating peak detection algorithms, we can further compare the purification performances between the two topic models.

Intensity-level purification

We obtained an average error ratio of mixture proportion $\bar{\xi}_d(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ at 2.33 %, indicating a good characterization of original proportions. The comparison of proportion parameters for the first six profiles is depicted in Fig. 7 using radar charts and scatter plots. As shown in the figure, the estimation in each profile has captured consistent patterns as the ground truth in each of the 10 components. We achieved an average correlation coefficient between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}^*_d$ at 0.975. The model accurately recognized those non-cancerous constituents contributed as small as 5 % in each sample. The proportion of cancerous origin is overestimated in some samples due to the

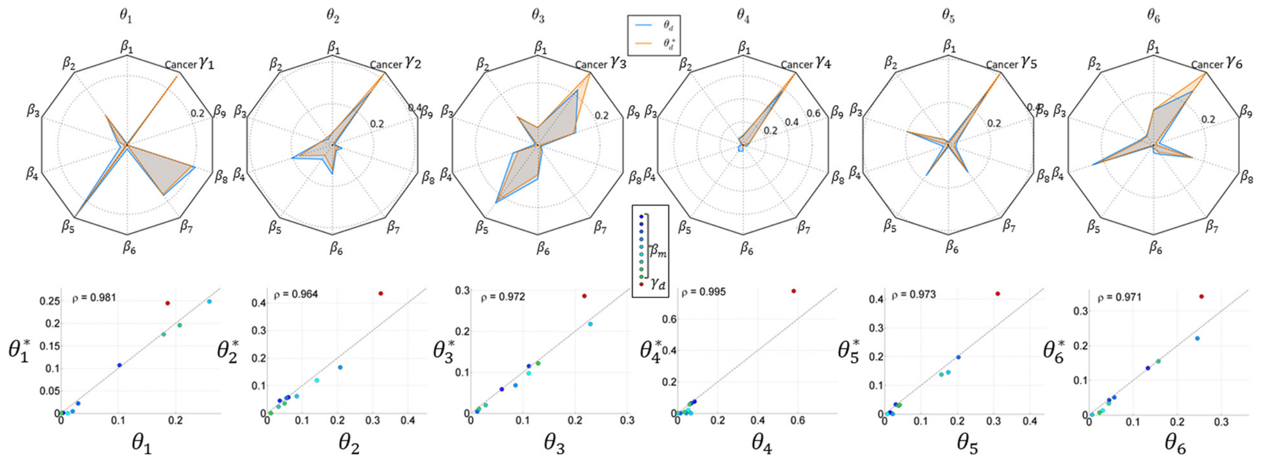

Fig. 7 Similarity evaluation on $\boldsymbol{\theta}$. Comparison between estimated $\boldsymbol{\theta}^*$ and true mixture proportions $\boldsymbol{\theta}$ for the first six profiles. *Top*: radar charts with 10 spokes, each representing a source in topic panel. The proportion of each source is depicted by the length of lines with color (*orange* for estimation $\boldsymbol{\theta}^*$ and *blue* for ground truth $\boldsymbol{\theta}$). *Bottom*: scatter plots of corresponding proportions in ground truth $\boldsymbol{\theta}$ and estimation $\boldsymbol{\theta}^*$. The correlation coefficients ρ are given on the left-top

smaller contributions from the contaminants. The differences between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_d^*$ are also related to the recovered pure cancer profiles $\{\boldsymbol{\gamma}_d^*\}$. Similarly, we have the average estimation error ratio for sample-specific pure cancer profiles $\bar{\xi}_d(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}) = 6.51$ %, which is smaller than $\bar{\xi}_d(\boldsymbol{t}, \boldsymbol{\gamma}) = 16.57$ %, i.e., the error ratio between unpurified cancer profile and true cancer profile. Figure 8 shows scatter plots of 101 proteins in unpurified cancer profile $\{\boldsymbol{t}_d\}_{d=1,\cdots,6}$ versus true cancer profile (blue) and in purified cancer profile versus true cancer profile (orange). The average correlation coefficient increases from 0.986 to 0.999 after purification. The effects of purification are illustrated in Fig. 9 by projecting the high-dimensional (dim = 101) profiles onto their top three principal components. We observe that the purified cancer profiles were more distant from non-cancerous profiles, and regularized towards an average cancer profile.

Scan-level purification

We first evaluated the purification power in the case of scan-level features. The average estimation error ratio of mixture proportions is 3.57 % by Eq. (13). In terms of recovering the underneath pure feature list, we achieved the average estimation error ratio for sample-specific pure cancerous feature list $\bar{\xi}_d(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}) = 3.12$ %, which is smaller than $\bar{\xi}_d(\boldsymbol{t}, \boldsymbol{\gamma}) = 9.61$ %, i.e., the error ratio between unpurified cancerous feature list and ground truth. The purification with scan-level features works to some extent but it is also interesting to prove the extended model works in a more accurate way than intensity-level topic model. To allow intensity-level purification model to handle scan-level synthetic dataset, we employed peak detection algorithms (i.e., through successive convolution with a 4th order Savitzky-Golay smoothing filter and a first-order derivative of a Gaussian kernel with window width of 25 scans, standard deviation of 3) to transfer EIC peaks into intensities using area under curve. The same algorithm is applied for transferring purified peak list resulted from the extended model. We obtained a greater distance of mixture proportion with $\bar{\xi}_d^{\mathcal{I}}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ at 7.23 % if using intensity-level purification model, compared to half ($\bar{\xi}_d^{\mathcal{S}}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 3.57$ %) achieved by extended scan-level purification model.

LC-MS based proteomic dataset

We treated all 59 cirrhotic profiles as origins of contaminants to purify 57 HCC profiles. We plotted these profiles using their first three principal components in Fig. 10.

Similar to the simulation result, we observed a clearer distinction between HCC and cirrhotic profiles after purification. To further understand the improvements, we carried out the following analyses on both purified and unpurified profiles.

Firstly, in statistical analysis, the most relevant proteins with differential intensities between HCC cases and cirrhotic controls were selected using t-test, and the associated p-values were adjusted based on multiple testing correction (FDR $\leq$ 0.05). We found 43 proteins with significant change in expression between the two groups. The number of reported significant proteins under the same testing method increased from 43 to 75 after purification. The majority of the proteins identified in original profiles (40 out of 43) remained significant after purification. If purified based on scan-level features, the number of significant proteins also increased to 69, among which 38 and 61 are overlapped with unpurification and intensity-level purification results, respectively.

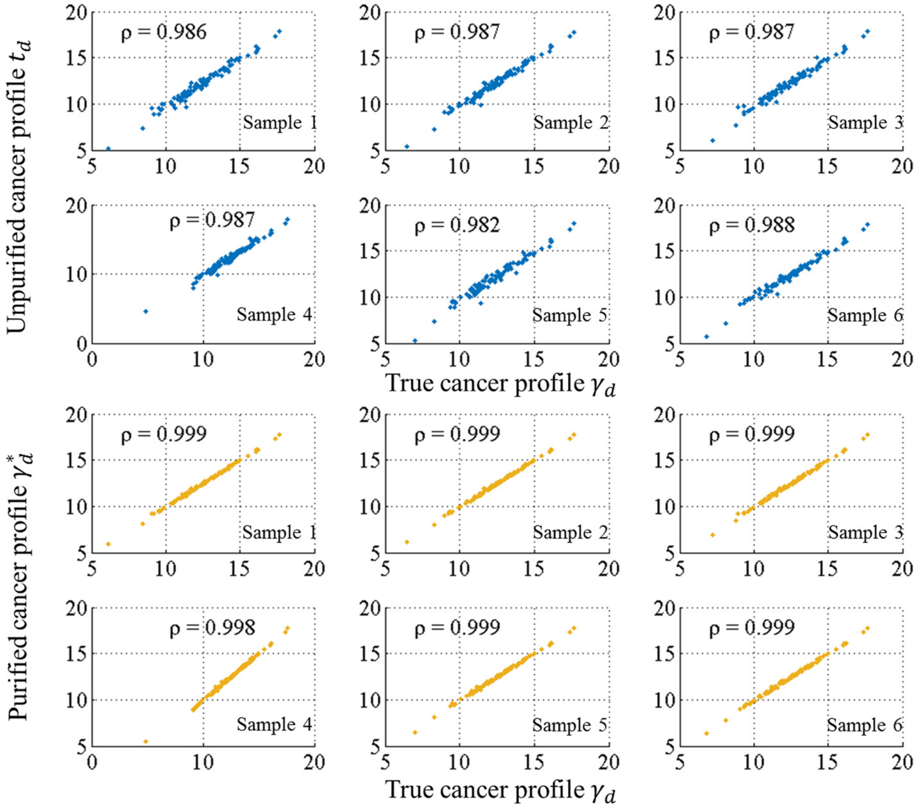

Fig. 8 Similarity evaluation on γ. The first six out of 30 scatter plots of unpurified cancer profiles versus true cancer profiles (*blue*) and corresponding scatter plots of purified cancer profiles versus true cancer profiles (*orange*). The correlation coefficients ρ between each pair of profiles are given on the left-top

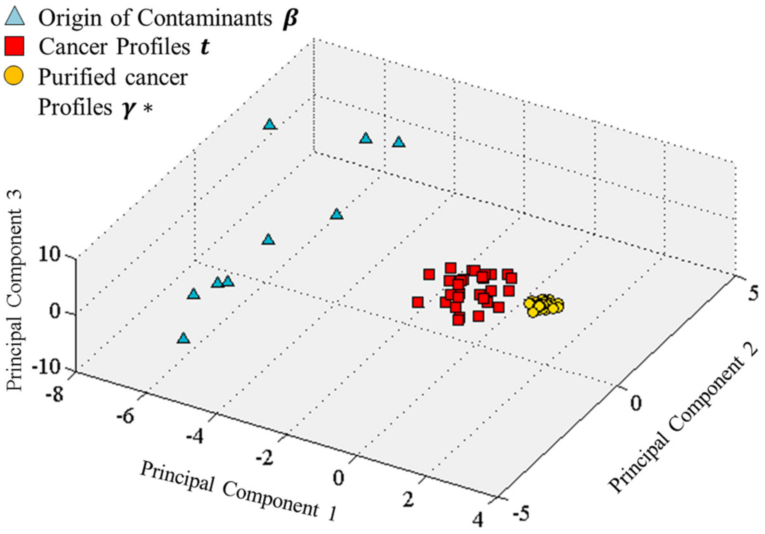

Fig. 9 PCA analysis on simulated dataset. Thirty cancer profiles $\{\boldsymbol{t}_d\}$ (*red square*), 30 purified cancer profiles $\{\boldsymbol{\gamma}^*_d\}$ (*yellow circle*), and 9 sources of cirrhotic contaminants $\{\boldsymbol{\beta}_m\}$ (*blue triangle*)

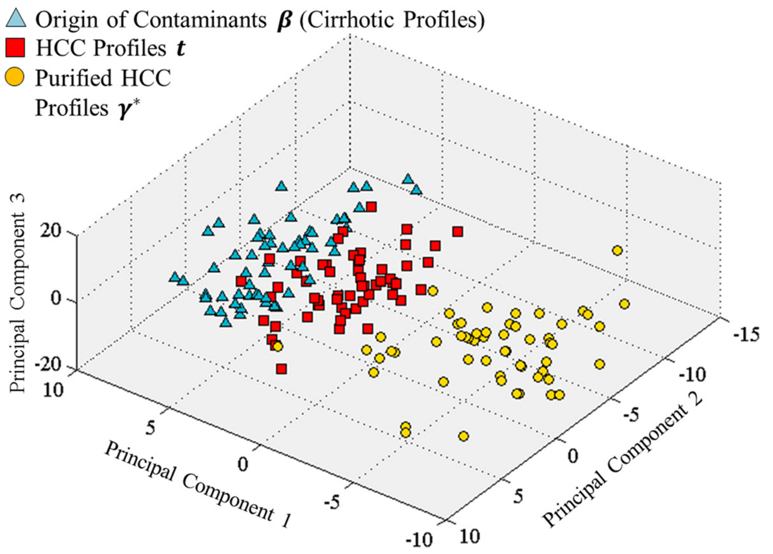

Fig. 10 PCA analysis on proteomic dataset. Fifty seven HCC profiles $\{\boldsymbol{t}_d\}$ (*red square*), 57 purified HCC profiles $\{\boldsymbol{\gamma}^*{}_d\}$ (*yellow circle*), and 59 sources of cirrhotic contaminants $\{\boldsymbol{\beta}_m\}$ (*blue triangle*)

Figure 11a, b, and c show ROC curves for each of the 43, 75, and 69 significant proteins, respectively. A bootstrap method (1000 bootstrap replicates) was used to compute the 95 % confidence interval (CI) of the area under each ROC curve. After intensity-level and scan-level purification we respectively achieved an average AUC of 0.793 (with 95 % CI at [0.700, 0.863]) and 0.811(with 95 % CI at [0.719, 0.890]), both higher than 0.706 (with 95 % CI at [0.606, 0.795]) for original biomarkers. More powerful biomarkers were selected after scan-level purification.

Finally, we used DAVID [23] (version 6.7) to identify significant signaling pathways, where the UniProt IDs of the significant proteins were mapped to the KEGG [24] database. As shown in Table 1, three pathways were reported from the original list of significant proteins. Following intensity-level and scan-level purifications, we found peroxisome proliferator-activated receptor (PPAR) signaling pathway with five and six significant proteins involved in addition to the three pathways (complement and coagulation casades, systemic lupus erythematosus, and prion disease) identified without purification. This is interesting in light of previous reports linking cancer and PPARs expressed in human liver [25].

GC-MS based metabolomic dataset

Heterogeneity issue is more intuitive in tissue samples, where the contaminations originate from the neighboring non-homogeneous cells. We first purified the HCC profiles $\{\boldsymbol{t}_d\}_{d=1,\cdots,5}$ using independent cirrhotic profiles $\{\boldsymbol{\beta}_m\}_{m=1,\cdots,5}$ as the sources of contamination. Without

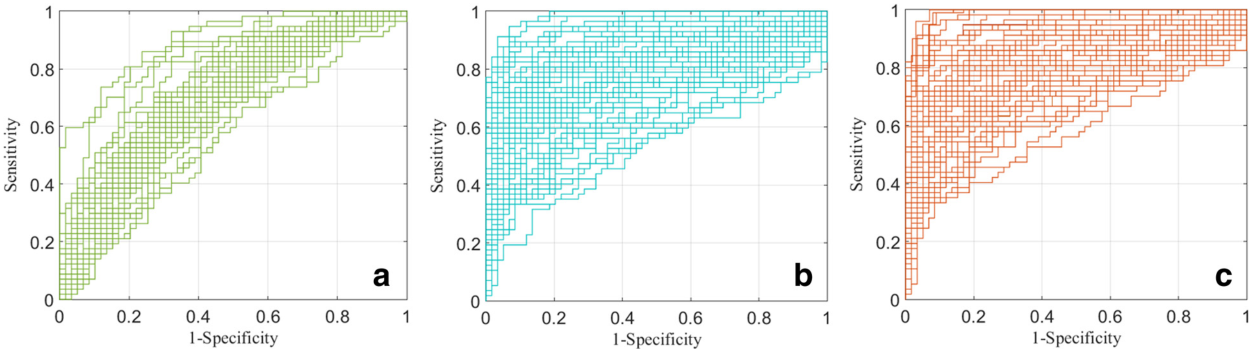

Fig. 11 ROC curves of significant proteins. **a** ROC curves for each of 43 significant proteins before purification ($\overline{\text{AUC}} = 0.706$, 95 % CI [0.606, 0.795]). **b** ROC curves for each of 75 significant proteins after intensity-level purification ($\overline{\text{AUC}} = 0.793$, 95 %CI [0.700, 0.863]). **c** ROC curves for each of 69 significant proteins after scan-level purification ($\overline{\text{AUC}} = 0.811$, 95 % CI [0.719, 0.890])

Table 1 Signaling Pathways (number of significant proteins involved in the pathway)

Without purification	Intensity-level purification	Scan-level purification
Complement and coagulation cascades (13)	Complement and coagulation cascades (18)	Complement and coagulation cascades (19)
Systemic lupus erythematosus (5)	Systemic lupus erythematosus (6)	PPAR signaling pathway (6)
Prion diseases (4)	PPAR signaling pathway (5)	Systemic lupus erythematosus (4)
-	Prion diseases (4)	Prion diseases (4)

purification, none of the 559 metabolites passed the statistical test as significant (FDR adjusted p-value $\leq$ 0.05). However, seven metabolites were identified as significant after the profiles were purified. For the adjacent cirrhotic profiles $\{\boldsymbol{\psi}_{\boldsymbol{d}}\}_{d=1,\cdots,5}$, we applied the model to remove contaminations from any neighboring cancerous cells. We expected to observe that the purified adjacent cirrhotic profiles became close to independent cirrhotic profiles. The dissimilarity, defined in (8), between independent and adjacent cirrhotic profiles is $\bar{\xi}(\boldsymbol{\psi}, \boldsymbol{\beta}) = 28.3$ %, and goes down to $\bar{\xi}\,(\boldsymbol{\psi}^*, \boldsymbol{\beta}) = 24.9$ % after purification. The improvements are less substantial compared to the previous datasets, presumably due to the limited sample size and potential overfitting issue.

Conclusions

In this paper, we investigate topic model-based inference methods to computationally address heterogeneity issue in samples analyzed by LC/GC-MS. The topic model gives a probabilistic explanation on the corpus of LC/GC-MS based profiles on both integrated peak and scan-level ion intensity levels. The performances of our models in estimating mixture proportion and retrieving underlying true cancer profile are evaluated through well-designed synthetic data. We observed that incorporation of scan-level features gives more accurate purification results by alleviating the loss in information caused as a result of integrating peak intensity values. Through GC-MS metabolomic and LC-MS proteomic datasets we acquired from tissues and blood samples, respectively, we showed the benefit of applying topic-model based purification of the data prior to statistical and pathway analyses. Specifically, we observed improved discrimination between case and control groups and biologically meaningful pathway analysis results. Future studies will focus on cross-validation of the findings either computationally through mass spectrometric data from large-scale cancer biomarker discovery studies or by using ground-truth information from pathology reports and literature survey.

Abbreviations

EIC, extracted ion chromatogram; GC, gas chromatography; HCC, hepatocellular carcinoma; LC, liquid chromatography; MS, mass spectrum; RT, retention time

Declarations

This article has been published as part of BMC Genomics Vol 17 Suppl 4 2016: Selected articles from the IEEE International Conference on Bioinformatics and Biomedicine 2015: genomics. The full contents of the supplement are available online at http://bmcgenomics.biomedcentral.com/articles/supplements/volume-17-supplement-4.

Funding

Publication of this article was funded in part by NIH Grants R01GM086746 and U01CA185188 awarded to HWR.

Authors' contributions

MW developed the methods and implemented the algorithms. MW, THT, and HWR designed the study. CDP and AF performed the sample collection, preparation, and data acquisition. GY and HWR supervised the analysis. MW and HWR wrote the manuscript with contributions from all other authors. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Oncology, Georgetown University, 4000 Reservoir Rd NW, Washington D.C., USA. [2] Department of Electrical and Computer Engineering, Virginia Tech, 900 N Glebe Rd, Arlington, VA, USA.

Published: 18 August 2016

References

1. Xiao JF, Varghese RS, Zhou B, Nezami Ranjbar MR, Zhao Y, Tsai TH, Di Poto C, Wang J, Goerlitz D, Luo Y, et al. LC–MS based serum metabolomics for identification of hepatocellular carcinoma biomarkers in Egyptian cohort. J Proteome Res. 2012;11(12):5914–23.
2. Tsai TH, Wang M, Di Poto C, Hu Y, Zhou S, Zhao Y, Varghese RS, Luo Y, Tadesse MG, Ziada DH, et al. LC–MS profiling of N-glycans derived from human serum samples for biomarker discovery in hepatocellular carcinoma. J Proteome Res. 2014;13(11):4859–68.
3. Tsai TH, Song E, Zhu R, Di Poto C, Wang M, Luo Y, Varghese RS, Tadesse MG, Ziada DH, Desai CS, et al. LC-MS/MS-based serum proteomics for identification of candidate biomarkers for hepatocellular carcinoma. Proteomics. 2015;15(13):2369–81.
4. Ranjbar MRN, Luo Y, Di Poto C, Varghese RS, Ferrarini A, Zhang C, Sarhan NI, Soliman H, Tadesse MG, Ziada DH, et al. GC-MS based plasma metabolomics for identification of candidate biomarkers for hepatocellular carcinoma in Egyptian cohort. PloS ONE. 2015;10(6): 0127299.
5. Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool opportunities and potential limitations. Mol Cell Proteomics. 2004;3(4):367–78.
6. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010;11(1):31–46.
7. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366(10):883–92.
8. Quon G, Haider S, Deshwar AG, Cui A, Boutros PC, Morris Q. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. Genome Med. 2013;5(3):29.
9. Garcia-Closas M, Hall P, Nevanlinna H, Pooley K, Morrison J, Richesson DA, Bojesen SE, Nordestgaard BG, Axelsson CK, Arias JI, et al. Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. PLoS Genet. 2008;4(4):1000054.

10. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ. Cell type–specific gene expression differences in complex tissues. Nat Methods. 2010;7(4):287–9.
11. Wang N, Gong T, Clarke R, Chen L, Shih IM, Zhang Z, Levine DA, Xuan J, Wang Y. Undo: a bioconductor r package for unsupervised deconvolution of mixed gene expressions in tumor samples. Bioinformatics. 2015;31(1):137–9.
12. Lawson CL, Hanson RJ. Solving Least Squares Problems. vol. 161. Englewood Cliffs, NJ: Prentice-hall; 1974.
13. Qiao W, Quon G, Csaszar E, Yu M, Morris Q, Zandstra PW. Pert: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. PLoS Comput Biol. 2012;8(12):1002838.
14. Posekany A, Felsenstein K, Sykacek P. Biological assessment of robust noise models in microarray data analysis. Bioinformatics. 2011;27(6): 807–14.
15. Quon G, Morris Q. ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. Bioinformatics. 2009;25(21):2882–9.
16. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
17. Nezami Ranjbar MR, Tadesse MG, Wang Y, Ressom HW. Bayesian normalization model for label-free quantitative analysis by lc-ms. IEEE/ACM Trans Comput Biol Bioinforma. 2015;12(4):914–27.
18. Wang M, Yu G, Mechref Y, Ressom HW. GPA: An algorithm for LC/MS based glycan profile annotation. In: the Proceedings of International Conference on Bioinformatics and Biomedicine (BIBM). Shanghai, China: IEEE; 2013. p. 16–22.
19. Monroe ME, Shaw JL, Daly DS, Adkins JN, Smith RD. Masic: A software program for fast quantitation and flexible visualization of chromatographic profiles from detected lc–ms (/ms) features. Comput Biol Chem. 2008;32(3):215–7.
20. Nault JC, Villanueva A. Intratumor molecular and phenotypic diversity in hepatocellular carcinoma. Clin Cancer Res. 2015;21(8):1786–88.
21. Cox J, Mann M. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26(12):1367–72.
22. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics. 2010;26(7):966–8.
23. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2008;4(1):44–57.
24. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
25. Tachibana K, Yamasaki D, Ishimoto K, Doi T. The role of PPARs in cancer: PPAR Res. vol. 2008. Article ID 102737; 2008, p. 15. doi:10.1155/2008/102737.