

Altered metabolite levels and correlations in patients with colorectal cancer and polyps detected using seemingly unrelated regression analysis

Chen Chen¹ · G. A. Nagana Gowda² · Jiangjiang Zhu³ · Lingli Deng⁴ · Haiwei Gu² · E. Gabriela Chiorean^{5,6,10} · Mohammad Abu Zaid⁵ · Marietta Harrison⁷ · Dabao Zhang¹ · Min Zhang^{1,8,9}  · Daniel Raftery^{2,10,11} 

Received: 17 June 2017 / Accepted: 30 August 2017 / Published online: 15 September 2017
© Springer Science+Business Media, LLC 2017

Abstract

Introduction Metabolomics technologies enable the identification of putative biomarkers for numerous diseases; however, the influence of confounding factors on metabolite levels poses a major challenge in moving forward with such metabolites for pre-clinical or clinical applications.

Objectives To address this challenge, we analyzed metabolomics data from a colorectal cancer (CRC) study, and used seemingly unrelated regression (SUR) to account for the effects of confounding factors including gender, BMI, age, alcohol use, and smoking.

Methods A SUR model based on 113 serum metabolites quantified using targeted mass spectrometry, identified 20 metabolites that differentiated CRC patients (n = 36),

patients with polyp (n = 39), and healthy subjects (n = 83). Models built using different groups of biologically related metabolites achieved improved differentiation and were significant for 26 out of 29 groups. Furthermore, the networks of correlated metabolites constructed for all groups of metabolites using the ParCorA algorithm, before or after application of the SUR model, showed significant alterations for CRC and polyp patients relative to healthy controls.

Results The results showed that demographic covariates, such as gender, BMI, BMI², and smoking status, exhibit significant confounding effects on metabolite levels, which can be modeled effectively.

Conclusion These results not only provide new insights into addressing the major issue of confounding effects in metabolomics analysis, but also shed light on issues related to establishing reliable biomarkers and the biological connections between them in a complex disease.

Electronic supplementary material The online version of this article (doi:10.1007/s11306-017-1265-0) contains supplementary material, which is available to authorized users.

✉ Min Zhang
minzhang@purdue.edu

✉ Daniel Raftery
draftery@uw.edu

¹ Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

² Northwest Metabolomics Research Center, Department of Anesthesiology and Pain Medicine, University of Washington, Seattle, WA 98109, USA

³ Department of Chemistry & Biochemistry, Miami University, Oxford, OH 45056, USA

⁴ Department of Electronic Science and Communication Engineering, State Key Laboratory for Physical Chemistry of Solid Surfaces, Xiamen University, Xiamen 361005, Fujian, China

⁵ Indiana University Melvin and Bren Simon Cancer Center, 535 Barnhill Drive, Indianapolis, IN 46202, USA

⁶ Department of Medicine, University of Washington, 825 Eastlake Ave East, Seattle, WA 98109, USA

⁷ Department of Medicinal Chemistry, Purdue University, West Lafayette, IN 47907, USA

⁸ Bioinformatics Center, School of Biomedical Engineering, Capital Medical University, Beijing 100069, China

⁹ Beijing Institute for Brain Disorders, Capital Medical University, Beijing 100069, China

¹⁰ Fred Hutchinson Cancer Research Center, 1100 Fairview Ave North, Seattle, WA 98109, USA

¹¹ Department of Chemistry, Purdue University, West Lafayette, IN 47907, USA

Keywords Seemingly unrelated regression · Colorectal cancer · Colorectal polyp · Metabolic profiling · Metabolomics · Targeted mass spectrometry · Clinical factors

1 Introduction

Colorectal cancer (CRC) affects both men and women of all ethnic groups around the world. As the second leading cause of cancer death, CRC is often found in people over 50 years old; an estimated 50,260 deaths is anticipated during 2017 in the United States alone (Siegel et al. 2017). It is widely accepted that the common risk factors of CRC include alcohol, colorectal polyps, family history, high-fat diets, smoking, and ulcerative colitis. CRC is highly treatable if detected at early stages, and therefore many screening tests (e.g., colonoscopy, flexible sigmoidoscopy, and stool-based tests) are performed for people at high risk (Lin et al. 2016). However, most of these tests are either invasive or expensive, which often make them difficult to implement in the clinic (Taylor et al. 2011). More recently, the U.S. FDA has approved Cologuard, a multi-target stool DNA test, and Epi proColon, a blood-based test for screening individuals at average-risk of CRC (Pickhardt 2016). The stool- and blood-based tests are increasingly becoming attractive primarily due to their non-invasive nature. Each of these tests, however, exhibits different levels of performance to detect CRC and the precursor lesions. In view of these challenges, only 40% of CRC patients are diagnosed and treated at an early stage (Stages I–II), where relatively high (80–90%) 5-year survival rates are common (Cancer Facts & Figures 2013). Therefore, it is crucial to develop sensitive, specific, inexpensive, and noninvasive methods that can be widely employed in regular clinical practice for people at high risk of CRC.

CRC, like most other cancers, is associated with altered cellular metabolism. For example, the well-known Warburg effect (Warburg 1956), involving increased glucose uptake and lactate production, has been exploited for cancer detection through PET scanning. Other metabolically related mechanisms are providing new strategies to develop novel therapies and detection modalities (Wise and Thompson 2010; Jain et al. 2012; Sreekumar et al. 2009; DeBerardinis and Chandel 2016; Gross et al. 2010; Ward and Thompson 2012; Munoz-Pinedo et al. 2012). The field of metabolomics simultaneously aims to identify altered metabolic pathways and mechanisms in diseases as well as robust and accurate biomarkers. To date, numerous metabolomics studies have focused on detecting biomarkers for CRC as well as altered metabolism (Qui et al. 2009; Nishiumi et al. 2012; Tan et al. 2013; Denkert et al. 2008; Chan et al. 2008; Ma et al. 2009; Li et al. 2013; Ritchie et al. 2010, 2013). A

few metabolomics studies have also focused on identifying metabolite biomarkers for polyp patients, who are considered at high risk for developing CRC (Eisner et al. 2013; Zhu et al. 2014). Depending on the biological samples and the analytical technique used in these studies, distinct sets of distinguishing biomarkers for CRC have been derived. It is, however, increasingly realized that the metabolite profiles are impacted by many factors such as gender, age, BMI, smoking, alcohol consumption, etc. The derived disease biomarkers are often confounded by one or more of these factors and hence often fail in validation studies. Thus, in general, the effect of confounding factors is a major challenge for metabolomics-based biomarker discovery research.

Despite the critical need, currently, efforts to account for the contributions from numerous confounding factors in biomarker discovery are limited. Recently, in an effort to address this challenge, we utilized a powerful statistical approach, seemingly unrelated regression (SUR) (Zellner 1962; Aiken and West 1991; Ahrens 1971) to model the effects of such confounding factors, which thereby allowed an investigation of serum metabolites that could differentiate polyp patients and healthy controls (Chen et al. 2015). In the present study, we have extended this concept to unravel the confounding effects using a significantly larger pool of metabolites from three groups of subjects. We have used SUR to analyze metabolites measured using targeted mass spectrometry to compare patients with CRC or polyps and healthy controls. SUR analysis identified a number of metabolites, groups of biologically related metabolites and their correlations that are distinct as well as statistically significant among three sets of samples (e.g., CRC, patients with polyps, and healthy controls). Considering that conventionally used multivariate statistical analysis methods such as logistic regression and partial least squares-discriminant analysis are challenged by the large effects of confounding factors that can override subtle changes in metabolite levels due to diseases, SUR promises new avenues for addressing the dominant confounding effects in the metabolomics field.

2 Materials and methods

2.1 Patient samples

This study utilized metabolite data obtained from a previous study and the materials and methods are described in greater detail in the published report (Zhu et al. 2014). Briefly, a total of 158 serum samples from CRC patients ($n = 36$), polyp patients ($n = 39$), and healthy controls ($n = 83$), which represented a subset of the original samples that had full clinical information available were utilized for this study. Recruitment of patients and blood collections were made as per the approved Institutional Review Board protocols from

Purdue University and Indiana University School of Medicine. All consenting participants undergoing colonoscopy or CRC surgery were evaluated, and blood samples from the patients were obtained after overnight fasting and bowel preparation prior to their procedure. Each blood sample was allowed to clot for 45 min, centrifuged at 1500 g for 10 min, the serum (upper layer) was aliquoted into cryo-vials and stored at -80°C until analysis.

2.2 Solvents and chemicals

Acetonitrile, ammonium acetate, and acetic acid were purchased from Fisher Scientific (Pittsburgh, PA). Standard compounds used for confirming the measured metabolites were purchased from Sigma-Aldrich (Saint Louis, MO) or Fisher Scientific (Pittsburgh, PA) (Zhu et al. 2014). Isotope labeled internal standards (L-tyrosine- $^{13}\text{C}_2$ and sodium-L-lactate- $^{13}\text{C}_3$) were purchased from Cambridge Isotope Laboratories (Tewksbury, MA). The purities of unlabeled standards were $>95\%$, whereas the purities of the two ^{13}C labeled compounds were $>99\%$.

2.3 Sample preprocessing

Frozen samples were thawed at room temperature and 50 μL of each serum sample was placed in a 2 mL Eppendorf vial (Fisher Scientific). To precipitate proteins, methanol (150 μL) was added and the mixture was vortexed for 2 min and kept at -20°C for 20 min. Next, the mixture was centrifuged at 20,800 g for 10 min, and the supernatant was collected into a new Eppendorf vial. To recover any metabolites left in the protein pellet, each pellet was mixed with 300 μL methanol, vortexed for 10 min, centrifuged at 20,800 g for 10 min and the supernatant was pooled with that collected after the first centrifugation. The solution was then dried using a Vacufuge Plus evaporator (Eppendorf, Hauppauge, NY), reconstituted in a 500 μL solution of 5 mM ammonium acetate in 40% water/60% acetonitrile +0.2% acetic acid containing two labeled internal standards, 5.13 μM L-tyrosine- $^{13}\text{C}_2$ and 22.5 μM sodium-L-lactate- $^{13}\text{C}_3$ (Cambridge Isotope Laboratory, Tewksbury, MA). The samples were filtered through 0.45 μm PVDF filters (Phenomenex, Torrance, CA) and used for MS analysis. A quality control (QC), which was a pooled serum sample from all three groups of subjects was also processed and prepared following the same procedure as used for individual samples and analyzed with the patients samples once every 10 patient samples.

2.4 Targeted mass spectrometry analysis

All samples were subjected to LC-MS/MS analysis using two hydrophilic interaction chromatography SeQuant

ZIC-cHILIC columns (150 \times 2.1 mm, 3.0 μm particle size, Merck KGaA, Darmstadt, Germany) connected in parallel and an AB Sciex QTrap 5500 mass spectrometer (AB Sciex, Toronto, ON, Canada) equipped with an electrospray ionization (ESI) source. A total of 158 MRM transitions were targeted (99 in negative ion mode and 59 in positive ion mode), and 113 metabolites could be measured in the serum samples with sufficient signal-to-noise and very few or no missing data.

2.5 Data preprocessing

The original data set was filtered to obtain samples that had values for all of the demographical and clinical variables under investigation (i.e., no missing values), leading to 158 individuals with 36 CRC patients, 39 polyp patients and 83 healthy controls. The data were standardized by normalization to average QC values ($n \geq 5$ per batch) on a metabolite by metabolite basis to account for minor instrument drift during the 12 day acquisition.

2.6 Data analysis using SUR

The SUR analysis that generalizes linear regression models for multiple response variables was originally proposed more than half a century ago (Zellner 1962). As illustrated in our recent paper (Chen et al. 2015), SUR consists of multiple regression equations, with each equation providing one response variable (the level of a metabolite). In this study, the SUR model was applied to test the effect of diagnosis on levels of metabolites by including the three sets of samples from CRC, polyp patients, and healthy controls. First, a SUR model was built to investigate how levels of all 113 metabolites were influenced by the following demographic variables: age, age 2 , gender, BMI, BMI 2 , smoking status, alcohol status, diagnosis, as well as the interactions between diagnosis and the other covariates. In all, a total of 15 covariates in each of the 113 regression equations was involved in the SUR model. The SUR model for each of the 113 metabolites can be written as:

$$Y_{ij} = \beta_{0j} + X_{i1}\beta_{1j} + X_{i2}\beta_{2j} + \dots + X_{ip}\beta_{pj} + \epsilon_{ij},$$

where Y_{ij} is the j^{th} metabolite level for the i^{th} individual, X_{ik} denotes the value of the k^{th} covariate of the i^{th} individual, and ϵ_{ij} is the error term which follows $N(0, \sigma_j^2)$. The error terms are independent across individuals, but may be correlated across the equations. Backward variable elimination (BVE) was then performed to remove insignificant covariates (i.e., those with p-value ≥ 0.05). Demographic covariates showing significant interactions with diagnosis were included in all linear regression models on each of the 113 metabolites in subsequent analysis. In addition to individual metabolites subjected to multiple linear regression analysis, we

also partitioned all metabolites into 29 biologically related groups based on their presence in predefined metabolic pathways (glycolysis, TCA cycle, purine metabolism, amino acid metabolism, etc.), and further applied SUR analysis to evaluate the effect of disease risk on different groups of metabolites. The ParCorA algorithm (Schafer and Strimmer 2005) was used to construct networks of metabolites in CRC, polyp patients and healthy subjects separately for all groups of metabolites both before and after using SUR to account for the effect of confounding variables. Using the ParCorA algorithm, the Pearson correlation coefficient between each pair of metabolites within the same group was calculated, and an edge was built between two metabolites if the p-value of the correlation was less than a pre-specified significance level ($p < 0.01$).

3 Results

This study utilized targeted LC-MS/MS based metabolite data previously obtained for human serum samples from 158 individuals with CRC ($n = 36$), patients with colorectal polyps ($n = 39$) or healthy controls ($n = 83$). Patient demographic and clinical information, including smoking and alcohol status are shown in Supporting Information Table S1. Using LC-MS/MS analysis, which involved utilizing a pre-established list of mass values for parent and fragment ions for each metabolite, relative concentrations of 158 different metabolites were measured. Of these, 113 metabolites were detected and quantified reliably as assessed by a lower coefficient of variation (CV) values for these metabolites (CV median 8%, ranging from 5 to 31%, with ~80% metabolites having $CV < 15\%$), and no missing values. The quantified metabolites represented diverse biochemical classes and over 25 distinct metabolic pathways.

3.1 SUR analysis

In the SUR analysis we performed BVE of the predictors, using the 113 metabolites and the demographic variables age, age², gender, BMI, BMI², smoking status, alcohol status, diagnosis, as well as the interactions between diagnosis and all the other seven covariates. BVE was employed to remove insignificant covariates prior to further analysis. Table 1 shows the results of BVE for predictors, which indicates, however, that all covariates were significant with p-values < 0.05 . Hence, all covariates were included in subsequent analysis. Using these predictors, each metabolite was analyzed separately using multivariate linear regression. The results of the linear regression analysis for testing the diagnosis and its interaction with other demographic variables for all 113 metabolites showed that 20 out of the 113 metabolites were significant

Table 1 The results of BVE for predictors including age, age², gender, BMI, BMI², smoking status, alcohol status, diagnosis, as well as the interactions between diagnosis and other covariates

Selected effects	p value
Age	0
Age ²	0
Gender	0
BMI	0
BMI ²	0
Smoking	0
Alcohol	9.45e-14
Diagnosis	0
Diagnosis × age	0
Diagnosis × age ²	0
Diagnosis × gender	0
Diagnosis × BMI	0
Diagnosis × BMI ²	0
Diagnosis × smoking	0
Diagnosis × alcohol	0

Table 2 Metabolites found to be significant ($p < 0.05$) for testing the diagnosis based on linear regression analysis using the predictors selected from BVE

Metabolite	p value	Adjusted p value
Alanine	1.53e-04	0.0022
Aspartic acid	5.09e-08	2.88e-06
2'-Deoxyuridine	0.0085	0.048
Glutamine	4.33e-04	0.0049
Histidine	8.30e-04	0.0078
Lysine	0.0042	0.028
Methionine	0.0017	0.014
Phenylalanine	0.0055	0.033
Alpha-ketoglutaric acid	0.0026	0.02
G16BP	0.0013	0.011
Glyceraldehyde	6.22e-10	7.03e-08
Cystathionine	3.88e-04	0.0049
Epinephrine	6.68e-04	0.0069
Hydroxyproline/Aminolevulinate	1.19e-06	2.23e-05
Hippuric acid	4.43e-06	7.15e-05
Linoleic acid	0.0052	0.033
Linolenic acid	0.0037	0.026
Orotate	1.15e-06	2.23e-05
Xanthine	2.13e-07	8.01e-06
Xanthosine	3.31e-07	9.34e-06

for diagnosis (Table 2, adjusted $p < 0.05$). Full results for the 113 metabolites are shown in Supporting Information Table S2. Among the 20 metabolites with adjusted $p < 0.05$, several had quite low p-values, with aspartic acid being the most significant, followed by glyceraldehyde, xanthine, xanthosine, orotate, hydroxyproline/aminolevulinate and hippuric acid (Table 2).

SUR analysis was then performed on subsets of the metabolites defined by their presence in 29 different metabolically related groups (Table 3). This SUR analysis showed that metabolites within the same group were significantly different among CRC, polyp patients and healthy controls for all 29 groups except three. The three groups that did not show significant differences were propionate metabolism consisting of four detected metabolites, bile acid metabolism consisting of four metabolites, and ketone bodies consisting of two metabolites. To further investigate the effects of diagnosis and demographic variables on the bile acid metabolism group, we built two SUR models, with the predictor being diagnosis only for one model and demographic variables only for the other model, to test the effects of diagnosis and demographic variables separately. The p-value for testing the diagnosis effect was 0.913 and

the p-value for testing the effect of demographic variables was 3.16×10^{-8} . Table 3 lists the 29 groups along with the number of metabolites in each group, the pathway each group represents and the corresponding p-value obtained from the SUR analysis. Names of the metabolites in each group are listed in the Supporting Information Table S3. Among the seven metabolites that showed highly significant differences in the individual metabolite analysis (Table 2), aspartic acid was present in 6 groups, glyceraldehyde and hippuric acid were present in one group each, and xanthine, xanthosine, orotate and hydroxyproline/aminolevulinic acid were present in two groups each, as shown in the Supporting Information Table S3. Interestingly, all the groups in which these highly significant metabolites were present (groups 3, 4, 5, 10, 15, 16, 17, 18, 19, 20 and 23 in

Table 3 Results of SUR analysis on testing the effect of diagnosis in 29 groups of biologically related metabolites

Group number (# metabolites)	Biological connection/pathway	p value	Adjusted p value (BH)*
1 (9)	Glycolysis/gluconeogenesis	4.72e-05	9.74e-05
2 (6)	Valine/leucine/isoleucine biosynthesis	0.023	0.027
3 (10)	Alanine/aspartate/glutamate metabolism	1.28e-08	3.84e-08
4 (10)	Arginine/proline metabolism	2.97e-08	8.17e-08
5 (3)	Histidine metabolism	1.76e-09	8.30e-09
6 (4)	Propionate metabolism	0.099	0.11
7 (6)	Tyrosine metabolism	0.059	0.067
8 (10)	Glyoxalate and dicarboxylate metabolism	5.20e-05	1.01e-04
9 (6)	Phenylalanine/tyrosine/ tryptophan metabolism	5.47e-06	1.20e-05
10 (12)	Glycolysis/TCA cycle	1.44e-06	3.39e-06
11 (9)	Glycine/serine/threonine metabolism	9.21e-05	1.69e-04
12 (8)	TCA cycle	0.0035	0.0050
13 (8)	Glycolysis	1.35e-04	2.34e-04
14 (7)	Gluconeogenesis	3.25e-04	5.36e-04
15 (9)	Urea cycle	2.78e-09	1.02e-08
16 (19)	Amino acids	2.11e-12	2.32e-11
17 (51)	Amino acid metabolism	0	0
18 (18)	Nucleotide metabolism	1.41e-10	7.76e-10
19 (13)	Purine metabolism	1.44e-11	9.50e-11
20 (6)	Pyrimidine metabolism	6.58e-09	2.17e-08
21 (4)	Bile acid metabolism	0.21	0.22
22 (5)	Carbohydrate	0.006	0.0083
23 (10)	Lipids pathway	3.15e-07	8.00e-07
24 (4)	Vitamins	0.0028	0.0042
25 (3)	Microbial metabolism	0.0071	0.0094
26 (6)	Tryptophan metabolism	0.022	0.027
27 (7)	Methionine metabolism	0.0024	0.0038
28 (4)	Organic acids	0.021	0.027
29 (2)	Ketone bodies	0.13	0.14

See Supplemental Information Table S3 for metabolite groupings

*Benjamini–Hochberg FDR correction

Table 3) also exhibited very high levels of significance for diagnosis as shown by their very low p-values.

3.2 Correlation analysis

Differences in terms of correlated metabolite networks were also observed in CRC, polyp patients and healthy subjects for all groups of metabolites before and after using SUR to account for the effects of confounding variables. The number of significant correlations (connections) between metabolites varied between healthy individuals, patients with polys and CRC patients. It is interesting to observe that the complexity of constructed networks of metabolites in some groups was reduced from healthy subjects to polyp patients, and further reduced for CRC patients. For example, as shown in Fig. 1, the total number of connections of constructed networks of metabolites in Group 14 is 17, 13, and 7 for healthy subjects, polyp patients, and CRC patients, respectively. Overall, for

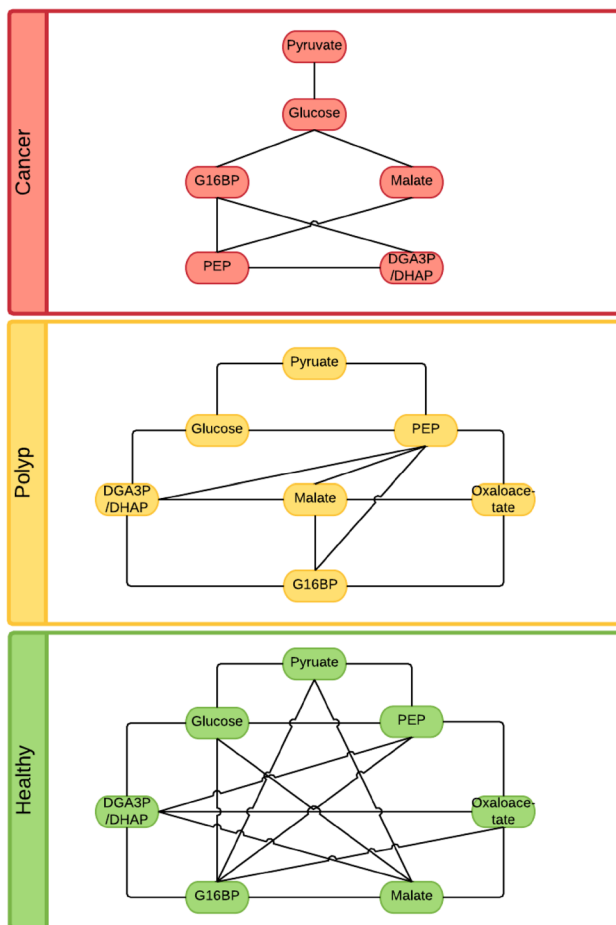


Fig. 1 Comparison of the number of significant correlations as indicated by the connections between metabolites for the constructed networks of Group 14 metabolites in cancer, polyp patients and healthy subjects. An overall reduction of the number of significant correlations is observed from healthy control to CRC

the Group 14 metabolites, there was a loss of correlation, in this case from an average of 0.36 for healthy subjects to an average of 0.24 for CRC patients, which results in a reduced number of connections. Of course, many of those correlations still exist, they are just smaller and therefore do not reach the cutoff value.

Table 4 summarizes the number of connections in the constructed networks of metabolites, as well as the average value of the correlation coefficient between metabolites in CRC, polyp patients and healthy subjects for all groups of metabolites both before and after using SUR to account for the effects of confounding variables. The results show that while the constructed networks of metabolites for a few groups remain the same before and after using SUR to account for the effects of confounding variables, the networks for most groups exhibited differences before and after using SUR to account for the effects of confounding variables. For example, while the constructed network of metabolites for Group 15 has 18, 9 and 5 connections in healthy controls, polyp patients and CRC before using SUR (see Fig. 2), respectively, the number of connections changed to 14, 10 and 9 after using SUR to account for the effects of confounding variables (Fig. 2). In addition, using SUR to account for the effects of confounding variables had an effect on the average value of correlation between metabolites in CRC, polyp patients and healthy subjects for most groups of metabolites. For Group 15, the average value of correlation between metabolites in CRC, polyp patients and healthy subjects was 0.18, 0.11 and 0.18, respectively, before using SUR, while the values change to 0.15, 0.16, and 0.18 after using SUR (Table 4). While the effect of SUR on some metabolite groups was observed to be sizeable, the overall effect of SUR on the median number of connections and median correlation values across all groups was relatively small.

4 Discussion

Advances in metabolomics technologies have enabled the analysis of a wider pool of metabolites with high fidelity and led to the identification of numerous metabolites that are putative biomarkers for the prediction or the diagnosis of numerous diseases. However, confounding effects pose a major challenge for reproducibility, validation and ultimately the utility of such potential biomarkers for pre-clinical or clinical applications. This challenge, in general, represents a major bottleneck in biomarker discovery for the metabolomics field. Therefore, it is increasingly becoming a critical priority that such confounding factors be taken into account in the effort to identify and validate reliable biomarkers. In the current study, we have used a SUR regression model to analyze metabolite data from

Table 4 Number of connections in the constructed networks of metabolites and average value of correlation between metabolites in CRC (C), polyp patients (P) and healthy subjects (H) for 29 groups of biologically connected metabolites before and after using SUR to account for the effects of confounding variables

Group number	Total # of connections						Average value of correlation					
	Before SUR			After SUR			Before SUR			After SUR		
	C	P	H	C	P	H	C	P	H	C	P	H
1	16	22	28	18	21	27	0.31	0.35	0.39	0.29	0.36	0.38
2	3	4	7	2	4	7	0.18	0.18	0.11	0.15	0.14	0.11
3	10	10	10	10	9	12	0.16	0.07	0.10	0.14	0.079	0.080
4	8	10	22	7	10	20	0.14	0.14	0.15	0.14	0.16	0.16
5	1	1	1	1	1	1	0.16	0.13	0.13	0.10	0.23	0.14
6	1	2	5	1	3	5	0.03	0.20	0.26	0.03	0.14	0.28
7	2	4	6	2	5	6	0.16	0.29	0.20	0.16	0.23	0.19
8	8	10	13	9	9	13	0.14	0.082	0.094	0.12	0.10	0.076
9	4	6	10	4	6	10	0.14	0.25	0.27	0.13	0.26	0.27
10	21	41	42	23	40	41	0.23	0.34	0.30	0.21	0.35	0.29
11	8	4	7	10	2	10	0.15	-0.0057	0.008	0.12	0.012	0.0033
12	3	21	17	3	18	16	0.11	0.34	0.25	0.11	0.35	0.24
13	13	17	21	13	16	21	0.30	0.35	0.41	0.28	0.36	0.39
14	7	13	17	9	12	16	0.27	0.34	0.38	0.24	0.36	0.36
15	5	9	18	9	10	14	0.18	0.11	0.18	0.15	0.16	0.18
16	57	68	92	53	72	81	0.26	0.25	0.23	0.25	0.27	0.23
17	264	288	464	269	280	433	0.16	0.14	0.13	0.13	0.14	0.13
18	28	36	28	26	32	28	0.087	0.14	0.067	0.10	0.10	0.071
19	16	14	16	12	13	14	0.086	0.11	0.088	0.094	0.080	0.099
20	1	2	1	1	2	0	0.073	0.14	-0.0060	0.068	0.11	-0.004
21	2	1	3	2	2	3	0.071	0.11	0.030	0.067	0.11	0.024
22	2	3	5	2	3	5	0.025	0.030	0.032	-0.017	0.055	0.047
23	8	15	9	6	14	12	0.13	0.19	0.12	0.17	0.17	0.11
24	0	0	1	1	0	1	0.22	0.030	-0.0072	0.17	0.041	0.015
25	0	2	1	1	2	0	-0.20	0.27	0.071	-0.11	0.22	0.082
26	6	3	7	8	4	6	0.31	0.19	0.18	0.27	0.13	0.17
27	4	7	5	6	5	5	0.21	0.071	0.11	0.15	0.070	0.085
28	1	0	2	1	0	2	0.051	0.12	0.10	0.067	0.081	0.11
29	1	1	1	1	0	1	0.86	0.28	0.66	0.88	0.34	0.66
Median	5	7	9	6	6	10	0.16	0.14	0.13	0.14	0.14	0.13

CRC patients, polyp patients and healthy controls while accounting for the effects of some major confounding variables including gender, BMI, age, and alcohol/smoking use. Previously, SUR analysis was applied for genome-wide association (Saint-Pierre et al. 2011) and nutritional epidemiology studies (Carroll et al. 2006). We recently proposed that the SUR model could be useful for metabolomics applications, and demonstrated how SUR analysis of NMR detected metabolites could differentiate polyp patients from healthy controls (Chen et al. 2015). In the present study, we have used a larger pool of metabolites (>100) obtained using highly sensitive targeted LC-MS. The SUR model was applied to test diagnosis by including three levels of the diagnosis variable: CRC, polyps or healthy control. Even after correction for the potentially confounding clinical factors, many metabolites (20) were

found to be significant for the diagnosis (see Table 3). These results signify the magnitude of the confounding effects as well as the ability of the SUR models to account for the contributions of such effects.

Detected metabolites in this study span more than 25 metabolic pathways, including major pathways such as the TCA cycle, glycolysis, gluconeogenesis, amino acid metabolism, urea cycle, purine, pyrimidine and nucleotide metabolism, carbohydrate and lipids pathways. To test the differentiating ability of such metabolic pathways, we have also investigated how major metabolic pathways are affected by the confounding variables. We therefore grouped metabolites based on their known association with specific metabolic pathways and subjected each group to SUR analysis. In line with the performance of the individual metabolites (Table 2), the low p-values observed for most of the biological groups indicate

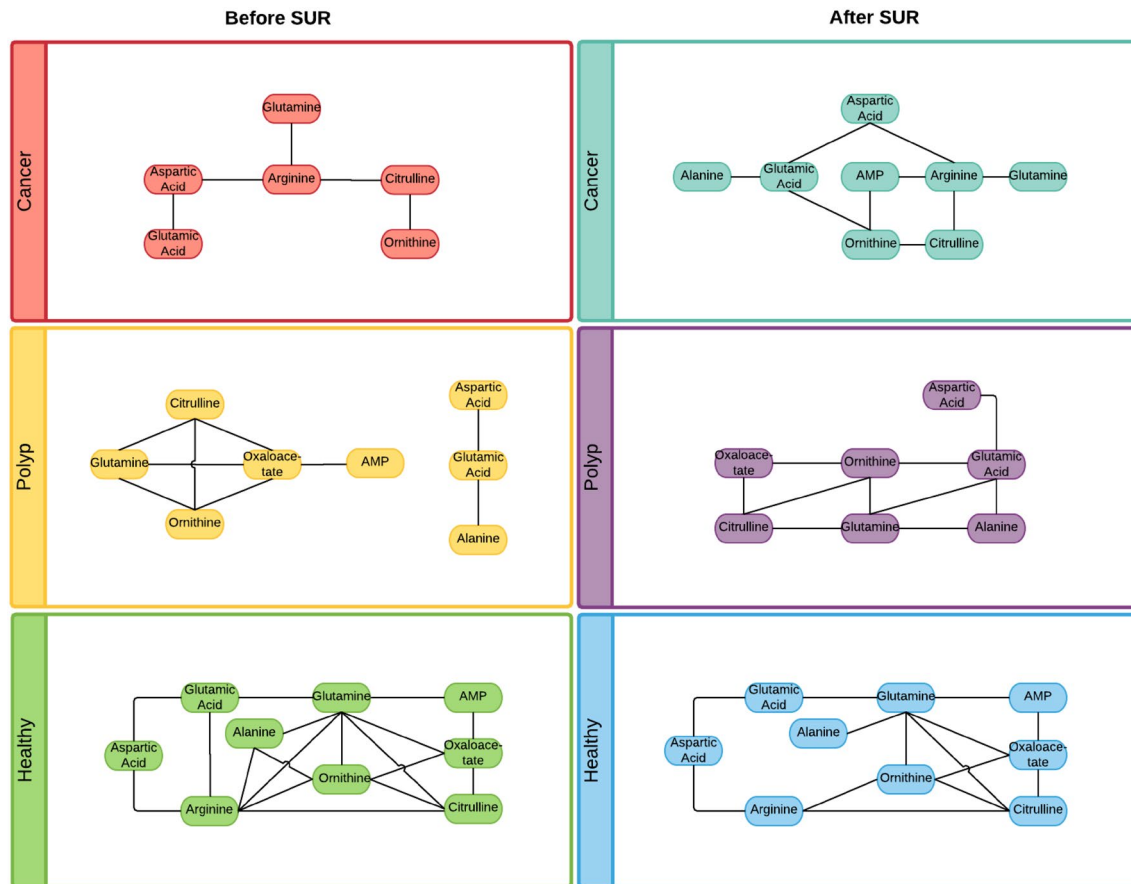


Fig. 2 Comparison of the number of significant correlations as indicated by the connections between metabolites for the constructed networks of Group 15 metabolites in cancer, polyp patients and healthy

subjects. SUR analysis affects a number of the significant correlations between metabolites, but the overall pattern of higher correlation between metabolites in healthy controls continues to hold

the strong differentiating ability of the groups of metabolites (Table 3).

Colorectal polyps represent a major risk factor for the later development of CRC. The ability to distinguish patients with polyps from healthy controls as well as CRC is critical for the prediction and diagnosis of CRC. With the goal of visualizing metabolic distinctions among CRC, polyp patients and healthy controls, we constructed networks of correlated metabolites using different groups of metabolites and the ParCorA algorithm. Calculation of the correlations was performed before and after applying the SUR model. The results show that both the number of connections and the average value of correlations were different among CRC, polyp patients and healthy controls, which indicate differences in the altered metabolism in the patients and control groups (Table 4). In some cases, the results were dramatically different (see Fig. 1). In addition, after SUR analysis, both the number of connections and the average correlations were altered for all but a few groups of metabolites, which indicates the extensive influence of confounding variables on metabolite levels.

The metabolites that were found to be significant ($p < 0.05$) for testing the diagnosis (Table 2) represent many metabolic pathways including glycolysis, TCA cycle, amino acids, nucleotide and lipids pathways. Alteration of these pathways is well known in cancer. For example the Warburg effect that affects glycolysis and TCA cycle pathways for energy generation in cancer cells is a well-known characteristic of cancer metabolism (Warburg 1956). Individually, 17 metabolites that were found to be significant for testing the diagnosis, based on SUR analysis, were previously shown to have links with CRC and colorectal polyp (Zhu et al. 2014). It was shown that aspartic acid, glutamine, lysine, methionine, histidine, hippuric acid, alpha-ketoglutarate, glyceraldehyde, hydroxyproline/aminolevulinic acid, linoleic acid, linolenic acid and 2'-deoxyuridine distinguish CRC and healthy controls as well as CRC and polyps. Similarly, cystathionine was shown to be associated with CRC and healthy controls; alanine, xanthine and orotate were shown to distinguish CRC and polyps; and glucose-1,6-bisphosphate (G16BP) was shown to distinguish between polyp patients and healthy controls.

Interestingly, among the metabolites detected by SUR analysis (Table 2), three metabolites including phenylalanine, xanthosine and epinephrine were unique in that their association with polyps is shown here for the first time. The fact that these metabolites were not found to be significant in the absence of SUR analysis (Zhu et al. 2014) indicates that the three metabolites are strongly influenced by the demographical variables. The association of phenylalanine indicates altered amino acid metabolism in cancer and is in accordance with many other amino acids that are associated with the disease. Xanthosine is associated with purine metabolism and indicates altered purine metabolism that is further supported by the observed significant changes in xanthine, which is also associated with purine metabolism. It may be noted that about half of the metabolites that were previously shown to be associated with CRC (Zhu et al. 2014) lost their significance after SUR analysis, which further signifies the magnitude of contributions from the demographic covariates. Nevertheless, some of this reduction may be due to the smaller number of samples analyzed in this study.

Differentiating metabolites in serum identified based on the present SUR analysis were also compared with biomarker candidates obtained in a number of other, earlier studies that used the same type of biospecimen, serum. Supplemental Information Table S4 provides a list of differentiating metabolites that are the same and different from those identified in the present study. Studies by Li et al. (Li et al. 2013) and (Ritchie et al. 2010, 2013) were focused on serum lipids and hence none of the biomarkers identified had any commonality with those in the present study. In addition to our own prior work (Zhu et al. 2014), three studies identified several aqueous metabolites that differentiated colorectal cancer from controls. In the GC-MS based study by Nishiumi et al., one out of a total four differentiating metabolites was same as that identified by SUR analysis (Nishiumi et al. 2012); in the study by Qiu et al., four out of thirty-three differentiating metabolites were the same (Qiu et al. 2009); and in the study by Tan et al., seven out of seventy-two metabolites were same as that identified by SUR analysis (Tan et al. 2013). Interestingly, phenylalanine, one of the three significant metabolites revealed after SUR analysis was also identified by two other studies, and xanthosine was identified by one other study (see Supplementary Table S4). However, a large number of differentiating metabolites from the earlier studies were different. The use of multiple analytical platforms that detect different pools of metabolites can explain such differences to some extent. Nevertheless, SUR analysis eliminated many metabolites including 3-hydroxybuturate, fumarate, creatinine, and proline that were shown to be differential metabolites by many studies (including our own prior study); in some cases SUR revealed other metabolites to have differentiating potential

as discussed above. These findings further highlight the contribution of potential confounding effects in metabolomics studies.

5 Conclusion

Our results indicate that demographic covariates such as gender, BMI, BMI², and smoking status exhibit significant confounding effects on metabolite levels as well as correlations between metabolites in the same biological group, which need to be taken into account in order to obtain a better representation of the effects of disease on metabolite levels. Importantly, biomarker discovery can be confounded by one or more of these factors. In this work, we used SUR to model how metabolites levels as well as their correlations are affected by several clinical variables, and investigated the ability of serum metabolites to differentiate CRC, polyp patients and healthy controls after accounting for the effects of confounding variables. These results may promise new avenues for addressing the often dominant confounding effects in the metabolomics field.

Acknowledgements The authors gratefully acknowledge the support of the Cancer Care Engineering (CCE) project, a joint effort between the Oncological Sciences Center (Purdue Center for Cancer Research, NCI P30CA023168) in the Purdue University Discovery Park and the Indiana University Melvin and Bren Simon Cancer Center (NCI P30CA082709). Support for the CCE project is gratefully acknowledged from the Walther Cancer Foundation, NIH (UL1RR025761), DOD (USAMRMC (CDMRP) W81XWH-008-1-0065, 9107003) and the Regenstrief Foundation. Additional financial support from NIH (R03CA211831 to N.G.), the Walther Cancer Foundation Bioinformatics grant, and the Cancer Center Support Grant P30CA015704-40) is also gratefully acknowledged.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

Ethical statement Recruitment of patients and blood collections was made with written informed consent as per the approved Institutional Review Board protocols from Purdue University and Indiana University School of Medicine.

References

- Ahrens, H. (1971). *Multivariate analysis*. Krishnaiah Paruchuri R. (Ed.) New York: Academic Press Inc.
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: testing and interpreting interactions*. Newbury Park, CA: Sage Publications, Inc.
- Cancer Facts & Figs. 2013. American Cancer Society: Atlanta, GA, 2013. <https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2013.html>. Accessed on June 14, 2017.

- Carroll, R. J., Midthune, D., Freedman, L. S., & Kipnis, V. (2006). Seemingly unrelated measurement error models, with application to nutritional epidemiology. *Biometrics*, *62*(1), 75–84.
- Chan, E. C. Y., Koh, P. K., Mal, M., Cheah, P. Y., Eu, K. W., Backshall, A., Cavill, R., Nicholson, J. K., & Keun, H. C. (2008). Metabolic profiling of human colorectal cancer using high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy and gas chromatography mass spectrometry (GC/MS). *Journal of Proteome Research*, *8*(1), 352–361.
- Chen, C., Deng, L., Wei, S., Nagana Gowda, G. A., Gu, H., Chiorean, E. G., Abu Zaid, M., Harrison, M. L., Pekny, J. F., Loehrer, P. J., Zhang, D., Zhang, M., & Raftery, D. (2015). Exploring metabolic profile differences between colorectal polyp patients and controls using seemingly unrelated regression. *Journal of Proteome Research*, *14*(6), 2492–2499.
- DeBerardinis, R. J., & Chandel, N. S. (2016). Fundamentals of cancer metabolism. *Science Advances*, *2*(5), e1600200.
- Denkert, C., Budczies, J., Weichert, W., Wohlgenuth, G., Scholz, M., Kind, T., Niesporek, S., Noske, A., Buckendahl, A., Dietel, M., & Fiehn, O. (2008). Metabolite profiling of human colon carcinoma - deregulation of TCA cycle and amino acid turnover. *Molecular Cancer*, *7*(1), 72.
- Eisner, R., Greiner, R., Tso, V., Wang, H., & Fedorak, R. N. (2013). A machine-learned predictor of colonic polyps based on urinary metabolomics. *BioMed Research International*, *2013*, 303982.
- Gross, S., Cairns, R. A., Minden, M. D., Driggers, E. M., Bittinger, M. A., Jang, H. G., Sasaki, M., Jin, S., Schenkein, D. P., Su, S. M., Dang, L., Fantin, V. R., & Mak, T. W. (2010). Cancer-associated metabolite 2-hydroxyglutarate accumulates in acute myelogenous leukemia with isocitrate dehydrogenase 1 and 2 mutations. *The Journal of Experimental Medicine*, *207*(2), 339–344.
- Jain, M., Nilsson, R., Sharma, S., Madhusudhan, N., Kitami, T., Souza, A. L., Kafri, R., Kirschner, M. W., Clish, C. B., & Mootha, V. K. (2012). Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science*, *336*(6084), 1040–1044.
- Li, F., Qin, X., Chen, H., Qiu, L., Guo, Y., Liu, H., Chen, G., Song, G., Wang, X., Li, F., Guo, S., Wang, B., & Li, Z. (2013). Lipid profiling for early diagnosis and progression of colorectal cancer using direct infusion electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Rapid Communications in Mass Spectrometry*, *27*(1), 24–34.
- Lin, J. S., Piper, M. A., Perdue, L. A., Rutter, C. M., Webber, E. M., O'Connor, E., Smith, N., & Whitlock, E. P. (2016). Screening for Colorectal Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *The Journal of the American Medical Association*, *315*(23), 2576–2594.
- Ma, Y.-L., Qin, H.-L., Liu, W.-J., Peng, J.-Y., Huang, L., Zhao, X.-P., & Cheng, Y.-Y. (2009). Ultra-high performance liquid chromatography mass spectrometry for the metabolomic analysis of urine in colorectal cancer. *Digestive Diseases and Sciences*, *54*(12), 2655–2662.
- Munoz-Pinedo, C., El Mjiyad, N., & Ricci, J. E. (2012). Cancer metabolism: current perspectives and future directions. *Cell Death and Disease*, *3*, e248.
- Nishiumi, S., Kobayashi, T., Ikeda, A., Yoshie, T., Kibi, M., Izumi, Y., Okuno, T., Hayashi, N., Kawano, S., Takenawa, T., Azuma, T., & Yoshida, M. (2012). A novel serum metabolomics-based diagnostic approach for colorectal cancer. *PLoS ONE*, *7*(7), e40459.
- Pickhardt, P. J. (2016). Emerging stool-based and blood-based non-invasive DNA tests for colorectal cancer screening: the importance of cancer prevention in addition to cancer detection. *Abdominal Radiology*, *41*, 1441–1444.
- Qiu, Y., Cai, G., Su, M., Chen, T., Zheng, X., Xu, Y., Ni, Y., Zhao, A., Xu, L. X., Cai, S., & Jia, W. (2009). Serum metabolite profiling of human colorectal cancer using GC-TOFMS and UPLC-QTOFMS. *Journal of Proteome Research*, *8*(10), 4844–4850.
- Ritchie, S., Ahiahonu, P., Jayasinghe, D., Heath, D., Liu, J., Lu, Y., Jin, W., Kavianpour, A., Yamazaki, Y., Khan, A., Hossain, M., Su-Myat, K., Wood, P., Krenitsky, K., Takemasa, I., Miyake, M., Sekimoto, M., Monden, M., Matsubara, H., Nomura, F., & Goodenowe, D. (2010). Reduced levels of hydroxylated, polyunsaturated ultra long-chain fatty acids in the serum of colorectal cancer patients: implications for early screening and detection. *BMC Medicine*, *8*(1), 13.
- Ritchie, S. A., Tonita, J., Alvi, R., Lehotay, D., Elshoni, H., Myat, S., McHattie, J., & Goodenowe, D. B. (2013). Low-serum GTA-446 anti-inflammatory fatty acid levels as a new risk factor for colon cancer. *International Journal of Cancer*, *132*(2), 355–362.
- Saint-Pierre, A., Kaufman, J. M., Ostertag, A., Cohen-Solal, M., Boland, A., Toye, K., Zelenika, D., Lathrop, M., de Vernejoul, M. C., & Martinez, M. (2011). Bivariate association analysis in selected samples: application to a GWAS of two bone mineral density phenotypes in males with high or low BMD. *European Journal of Human Genetics*, *19*(6), 710–716.
- Schafer, J., & Strimmer, K. (2005). An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks. *Bioinformatics*, *21*(6), 754–764.
- Siegel, R., Miller, K. D., & Jemal, A. (2017). Cancer statistics, 2016. *CA - Cancer Journal for Clinicians*, *67*, 7–30.
- Sreekumar, A., Poisson, L. M., Rajendiran, T. M., Khan, A. P., Cao, Q., Yu, J., Laxman, B., Mehra, R., Lonigro, R. J., Li, Y., Nyati, M. K., Ahsan, A., Kalyana-Sundaram, S., Han, B., Cao, X., Byun, J., Omenn, G. S., Ghosh, D., Pennathur, S., Alexander, D. C., Berger, A., Shuster, J. R., Wei, J. T., Varambally, S., Beecher, C., & Chinnaiyan, A. M. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, *457*(7231), 910–914.
- Tan, B., Qiu, Y., Zou, X., Chen, T., Xie, G., Cheng, Y., Dong, T., Zhao, L., Feng, B., Hu, X., Xu, L. X., Zhao, A., Zhang, M., Cai, G., Cai, S., Zhou, Z., Zheng, M., Zhang, Y., & Jia, W. (2013). Metabonomics identifies serum metabolite markers of colorectal cancer. *Journal of Proteome Research*, *12*(6), 3000–3009.
- Taylor, D. P., Cannon-Albright, L. A., Sweeney, C., Williams, M. S., Haug, P. J., Mitchell, J. A., & Burt, R. W. (2011). Comparison of compliance for colorectal cancer screening and surveillance by colonoscopy based on risk. *Genetics in Medicine*, *13*(8), 737–743.
- Warburg, O. (1956). On the origin of cancer cells. *Science*, *123*, 309–314.
- Ward, P. S., & Thompson, C. B. (2012). Metabolic reprogramming: A cancer hallmark even warburg did not anticipate. *Cancer Cell*, *21*(3), 297–308.
- Wise, D. R., & Thompson, C. B. (2010). Glutamine addiction: A new therapeutic target in cancer. *Trends in Biochemical Sciences*, *35*(8), 427–433.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, *57*(298), 348–368.
- Zhu, J., Djukovic, D., Deng, L., Gu, H., Himmati, F., Chiorean, E. G., & Raftery, D. (2014). Colorectal cancer detection using targeted serum metabolic profiling. *Journal of Proteome Research*, *13*(9), 4120–4130.