

Identification of Novel Serological Biomarkers for Inflammatory Bowel Disease Using *Escherichia coli* Proteome Chip^{*}

Chien-Sheng Chen,^{a,b,c} Sean Sullivan,^{c,d,e} Troy Anderson,^{c,f} Aik Choon Tan,^{c,f,g} Philip J. Alex,^d Steven R. Brant,^d Carmen Cuffari,^e Theodore M. Bayless,^d Monica V. Talar,^h C. Lynne Burek,^h Huan Wang,^d Richard Li,^d Lisa Wu Datta,^d Yuqiong Wu,^d Raimond L. Winslow,^f Heng Zhu,^{a,i} and Xuhang Li^{d,j}

Specific antimicrobial antibodies present in the sera of patients with inflammatory bowel disease (IBD) have been proven to be valuable serological biomarkers for diagnosis/prognosis of the disease. Herein we describe the use of a whole *Escherichia coli* proteome microarray as a novel high throughput proteomics approach to screen and identify new serological biomarkers for IBD. Each protein array, which contains 4,256 *E. coli* K12 proteins, was screened using individual serum from healthy controls ($n = 39$) and clinically well characterized patients with IBD (66 Crohn disease (CD) and 29 ulcerative colitis (UC)). Proteins that could be recognized by serum antibodies were visualized and quantified using Cy3-labeled goat anti-human antibodies. Surprisingly significance analysis of microarrays identified a total of 417 *E. coli* proteins that were differentially recognized by serum antibodies between healthy controls and CD or UC. Among those, 169 proteins were identified as highly immunogenic in healthy controls, 186 proteins were identified as highly immunogenic in CD, and only 19 were identified as highly immunogenic in UC. Using a supervised learning algorithm (k -top scoring pairs), we identified two sets of serum antibodies that were novel biomarkers for specifically distinguishing CD from healthy controls (accuracy, $86 \pm 4\%$; $p < 0.01$) and CD from UC (accuracy, $80 \pm 2\%$; $p < 0.01$), respectively. The Set 1 antibodies recognized three pairs of *E. coli* proteins: Era versus YbaN, YhgN versus FocA, and GabT versus YcdG, and the Set 2 antibodies recognized YidX versus FrvX. The specificity and sensitivity of Set 1 antibodies were 81 ± 5 and $89 \pm 3\%$, respectively, whereas those of Set 2 antibodies were 84 ± 1 and $70 \pm 6\%$, respectively. Serum antibodies identified for distinguishing healthy controls versus UC were only mar-

ginal because their accuracy, specificity, and sensitivity were 66 ± 5 , 69 ± 5 , and $61 \pm 7\%$, respectively ($p < 0.04$). Taken together, we identified novel sets of serological biomarkers for diagnosis of CD versus healthy control and CD versus UC. *Molecular & Cellular Proteomics* 8: 1765–1776, 2009.

Crohn disease (CD)¹ and ulcerative colitis (UC) are chronic, idiopathic, and clinically heterogeneous intestinal disorders collectively known as inflammatory bowel disease (IBD) (1, 2). Although the distinction between UC and CD would seem clear based on the combination of clinical, endoscopic, and radiological criteria, indeterminate colitis is present in up to 10 and 20% of adult and pediatric patients with isolated colitis, respectively (3, 4).

Serological testing is a non-invasive method for diagnosing IBD and differentiating UC from CD (5–7). Several serological IBD biomarkers have been identified in the past decade, and some have been used in IBD clinics (5–7) (see the list below). Many of these antibodies are produced on intestinal exposure to normal commensal bacteria in genetically susceptible individuals. Although it is not known whether these antibodies are pathogenic or not, they are specific to patients with either CD or UC and may reflect a dysregulated immune inflammatory response to intestinal bacterial antigens (2, 8–10). Several experimental animal models of IBD have led to the theory that the pathogenesis of IBD is the result of an aberrant immune response to normal commensal bacteria in genetically susceptible individuals (11, 12). In fact, most of the major serological biomarkers being used in IBD clinics are antibodies to microbial antigens, including yeast oligomannose (anti-*Saccharomyces cerevisiae* (ASCA)), bacterial outer membrane porin C (OmpC), *Pseudomonas fluorescens* bacterial se-

From the Departments of ^aPharmacology and Molecular Sciences, ^bPediatrics, ^cOncology, and ^dPathology, ^eThe Meyerhoff Inflammatory Bowel Disease Center and Gastrointestinal Division, Department of Medicine, and ^fInstitute of Computational Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205 and ^gGraduate Institute of Systems Biology and Bioinformatics, National Central University, Jhongli 32001, Taiwan

Received, December 22, 2008, and in revised form, March 9, 2009
Published, MCP Papers in Press, April 7, 2009, DOI 10.1074/mcp.M800593-MCP200

¹ The abbreviations used are: CD, Crohn disease; IBD, inflammatory bowel disease; UC, ulcerative colitis; SAM, significance analysis of microarrays; HC, healthy control(s); ASCA, anti-*S. cerevisiae*; OmpC, outer membrane porin C; pANCA, perinuclear anti-neutrophil cytoplasmic antibody; kNN, k -nearest neighbors; SVM, support vector machines; k -TSP, k -top scoring pairs; GO, gene ontology.

quence I2 (anti-I2), and most recently bacterial flagellin (CBir 1) (5–7, 13). All of these antimicrobial antibodies show a preponderance in patients with CD. However, ASCA has been identified in up to 5% of patients with UC (13, 14).

In comparison, perinuclear anti-neutrophil cytoplasmic antibody (pANCA) with perinuclear highlighting was first described in 1990. Although generally considered an autoantibody, the specific antigenic stimulation for pANCA production remains unclear. This autoantibody is present in up to 70% of patients with UC and in up to 20% of patients with CD (6, 10). Recently a panel of five new anti-glycan antibodies have been identified, including anti-chitobioside IgA, anti-laminaribioside IgG, anti-mannobioside IgG, and antibodies against two major chemically synthesized (Σ) oligomannose epitopes, Man α -1,3 Man α -1,2 Man (Σ Man3) and Man α -1,3 Man α -1,2 Man α -1,2 Man (Σ Man4) (5, 13, 15). These new biomarkers serve as valuable complimentary tools to the available serological biomarkers mentioned above. Collectively these antibodies are not generally present in either children or adults with non-IBD disease and may represent serological markers of intestinal inflammation specific to UC or CD.

Although encouraging, none of the current commercially available biomarker tests/assays, including all of those mentioned above, can be used as stand alone tools in clinics and therefore are only recommended as an adjunct to endoscopy in diagnosis and prognosis of the disease (5, 7, 16). Therefore, additional specific and sensitive IBD biomarkers are needed as are prospective studies to assess the utility of current and newly identified biomarkers (5, 13). Proteomics technologies such as two-dimensional gel electrophoresis, various variations of mass spectrometry, and protein chip (array) technology are now proving to be powerful tools in biomarker discovery and are beginning to be utilized in IBD biomarker discovery (5, 17). These technologies enable robust and/or large scale and high throughput identification and analysis of differential protein expression when comparing disease with control. Blood-based (serum- or plasma-based) proteomics holds particular promises for biomarker discovery of various human diseases such as neurodegenerative diseases and cancers (18–20). Antigen microarrays are also powerful tools that allow high throughput serum analysis of aberrant immune responses in autoimmune diseases (21–23) as well as efficient discovery of biomarkers for infectious pathogens (24). Herein we describe the use of an *Escherichia coli* proteome microarray to characterize the differential immune response (serum anti-*E. coli* antibodies) among patients clinically classified as CD, UC, and healthy controls. We hypothesized that novel IBD-specific antimicrobial antibodies, particularly anti-*E. coli* antibodies, are present in IBD patients and are likely to be identified by screening the sera with *E. coli* protein arrays.

EXPERIMENTAL PROCEDURES

Patients and Serum Acquisition/Selection—Serum was obtained from 134 subjects in accordance with the policy of The Johns Hop-

kins Hospital Institutional Review Board. 66 patients had the diagnosis of CD, 29 patients were diagnosed with UC, and 39 subjects were non-IBD healthy controls (HC). The healthy controls and IBD patients were similar in age and sex distribution. The demographic and clinical characteristics of the patients are summarized in Table I. Clinical information was extracted from the written and electronic medical records. The diagnosis of CD and UC was established by standard clinical, radiographic, endoscopic, and histological criteria. Patients were classified as having CD based on the typical findings of skip lesions, deep linear or serpiginous ulcerations, cobblestoning, multiple noncaseating granulomas, transmural inflammation, small bowel involvement, structuring disease, or presence of fistulizing disease. The diagnosis of UC was considered if the colonic inflammation involved the rectum with or without proximal extension. The inflammation had to be continuous and be limited to the mucosa. There were no patients with proctitis enrolled in this study. The healthy controls consisted of individuals that had undergone colon cancer screening or screening for other non-IBD gastrointestinal diseases or any other immune diseases. The serum samples were obtained at the time of initial outpatient encounter, at the time of an endoscopy, or during hospitalization. The blood was collected into a serum separator tube (red top tube, BD Vacutainer) and spun down within 60 min of collection. Serum was removed, aliquoted, and stored in multiple at -80°C until assayed. All serum samples used in our study were selected from prescreening of at least 200 samples (50 UC, >100 CD, and 50 healthy controls) from a large pool of serum samples. Only samples with similar levels of immunoglobulins (determined by SDS-PAGE and Western blot using donkey anti-human IgG) were used in our study.

Fabrication of *E. coli* Proteome Chips—To facilitate the analysis of protein function in the bacterial proteomes, we constructed a protein chip that essentially covers the entire proteome of the *E. coli* K12 strain (25). Briefly 4,256 *E. coli* proteins were first purified using an ORF collection kindly provided by Mori and co-workers (26). *E. coli* cells first were grown overnight at 37°C in $2\times$ LB medium containing 30 $\mu\text{g}/\text{ml}$ chloramphenicol in a 96-well format and allowed to grow overnight. The overnight cultures were diluted to a final A_{600} of ~ 0.1 . After the cells were grown for ~ 3 h at 37°C , protein expression was induced with 1 mM isopropyl β -D-thiogalactoside for ~ 3.5 h. The liquid cultures were then harvested by centrifugation at 3,500 rpm for 5 min at 4°C . The pellets were stored at -80°C for future protein purification.

To purify the fusion proteins, the frozen cell pellets were resuspended in phosphate lysis buffer containing 300 mM NaCl, 20 mM imidazole, CellLytic B, 1 mg/ml Lysozyme, 50 units/ml Benzonase, proteinase inhibitor mixture, and 1 mM PMSF. Along with nickel-nitrilotriacetic acid beads, the mixtures were incubated for 1.5 h at 4°C . After mixing, the resin-protein complexes were washed three times with Wash buffer I (50 mM NaH_2PO_4 with 300 mM NaCl, 10% glycerol, 20 mM imidazole, 0.01% Triton X-100 at pH 8) and three times with Wash buffer II (50 mM NaH_2PO_4 with 150 mM NaCl, 25% glycerol, 20 mM imidazole, 0.01% Triton X-100 at pH 8). Finally the fusion protein was eluted with elution buffer (50 mM NaH_2PO_4 , 150 mM NaCl, 25% glycerol, 250 mM imidazole, 0.01% Triton X-100, pH 7.5). All purified proteins were printed in duplicate onto FullMoon slides using a ChipWriter Pro (Bio-Rad) in a humidity-controlled chamber in a cold room (25). In addition, a dilution series of BSA and GST-His₆ proteins were always included on each chip as negative controls. To monitor success of the serum profiling reactions on the chips (quality and consistency of each serum sample (such as similar concentration of immunoglobulins)) and each protein chip, we included two positive controls on each chip as positive controls: 1) Ebna2, an Epstein-Barr virus-encoded antigen that is reactive in common serological assays for essentially every human being, and 2)

YLR-286C (endochitinase), a yeast protein that we found to be strongly and similarly recognized by every human serum we tested regardless of the disease (IBD). This protein was identified during our screening of a yeast protein chip (containing the entire yeast proteome) (27) with 50 human serum samples (25 healthy controls and 25 IBD patients) (data not shown). These two positive controls were similarly recognized (in terms of intensity) by all the serum samples used in this study.

Screen of *E. coli* Proteome Chip for Anti-*E. coli* Antibodies—The entire screening process, except for the washing steps as specified, was done at room temperature. *E. coli* protein chips stored at -80°C were thawed at room temperature (22°C) and blocked in Superblock Blocking Buffer (Pierce) for 1 h. The patient's serum was diluted (1:1,000) with blocking buffer in a total volume of 3 ml. The diluted serum was then applied to the chip entirely covering the surface. After 1-h incubation with gentle shaking on a rocker, the chip was rinsed once with 4 ml of TBS with 0.05% Tween 20 (TBS-T). The chip was then soaked in 4 ml of TBS-T, placed in a water bath, and washed for 10 min at 50°C with gentle horizontal agitation. This washing step was repeated twice. The chip was then cooled to room temperature. After removal of TBS-T, the chip was incubated for 1 h with the secondary antibody, a Cy3-labeled donkey anti-human IgA, IgG, and IgM (Jackson ImmunoResearch Laboratories) diluted at 1:400 in 3 ml of Superblock Blocking Buffer. The chip was then washed at 50°C in the same fashion as previously stated. After the final wash, the chip was rinsed in sterile water briefly and quickly spun at 2,000 rpm until dry prior to scanning. The chips were scanned with a GenePix array scanner (GenePix Pro 6.0 or GenePix 4200AL, Molecular Devices) at a wavelength of 536 nm. To achieve the best signal-to-noise ratio, many washing conditions with different stringencies were tested, including an increase of salt (0.5 or 1 M NaCl), addition of SDS (0.05 or 0.1%), change of washing temperature (22, 37, 40, or 50°C), and/or various combination of the conditions described above. The washing condition described here gave the best results among all conditions tested.

Protein Array Data Preprocessing—Each quantified sample array image was exported from GenePix (Molecular Devices) as a text file for preprocessing. The goal of preprocessing is to yield a feature of interest from each protein spot in the array that minimizes technical variability and maximizes the signal of interest. The ratio of the mean signal over the mean background signal for each protein spot was determined to be the best method of preprocessing. This method has the advantage that all features are normalized to their background signals. Thus, if a protein spot signal is artificially high due to an artifact on the slide the ratio will account for it. Furthermore this preprocessing method also normalizes the features across all arrays as the ratio is a standardized metric. The ratio represents the -fold change of the signal above background and can be interpreted as the degree of host serum reactivity to each spotted protein.

Univariate Significance Testing—Significance analysis of microarrays (SAM) (28) was used to determine proteins to which HC, CD, and UC groups of samples show a statistically significant immunogenic response. We used stringent criteria in the SAM and only called a protein significant when there was at least a 1.5-fold change difference between two phenotypes at 0% false discovery rate in 500 permutations.

Supervised Learning Algorithms—To construct the classifier in this study, we used three supervised learning methods. The algorithms implemented were *k*-nearest neighbors (*k*NN) (28), support vector machines (SVM), and the *k*-top scoring pairs (*k*-TSP) algorithm (29). The *k*-TSP algorithm was implemented using a publicly available executable program developed at the Institute for Computational Medicine of The Johns Hopkins University (29). SVM and *k*NN were

implemented using the R statistical programming language (packages e1071 and class for SVM and *k*NN, respectively).

Feature Selection—For *k*NN and SVM learning methods, SAM was applied to the training set for feature selection before the classifiers were trained on those data. The features selected in SAM were those that were found to be significant with a false discovery rate of 0. The *k*-TSP algorithm does not require feature reduction as it intrinsically selects the top scoring features. Parameters such as the number of nearest neighbors for *k*NN and the number of top scoring pairs for *k*-TSP were selected based on leave one out cross-validation performance on the training set. A script was written in Matlab to perform the cross-validation scheme and call executable files for the learning algorithms.

Cross-validation—10-fold cross-validation was performed on the data sets to obtain an unbiased estimation of the classification rate. In brief, one-tenth of the samples are randomly chosen to be the test set from the total number of samples, and the remaining nine tenths of the samples are defined as the training set. Each classifier is then trained on only the training set, including feature selection. Finally the trained classifiers are applied to the test set, and the number of correct classifications is recorded. This process is repeated 10 times to leave out and classify each patient sample as if they represented new data. The number of correct classifications divided by the number of total samples classified yields the unbiased estimate of correct classification rate.

Statistical Analyses—We used the open source statistical software R to perform the statistical analyses in this study. A *p* value <0.05 was regarded as significant.

RESULTS

Overall Strategy of Identifying IBD Serological Markers from *E. coli* Proteome Chips—Sera were collected from 134 individuals (29 HC, 66 CD, and 39 UC) from The Johns Hopkins Medical Institutes in this study (Table I). To identify potential biomarkers for IBD diagnosis, we profiled the antibody repertoire of the IBD patients using the *E. coli* proteome chips that each contained more than 4,200 individual proteins (see the schematic illustration of our strategy in Fig. 1). Because each protein was spotted in duplicate on the chip, we first analyzed the reproducibility of the duplicate of each protein. As shown in a scatter plot, the duplicate spots of each protein are highly correlated, indicating the good quality of the array manufacturing (supplemental Fig. S1; also see examples of the visual appearance of duplicate spots in Fig. 2). To recognize those reactive antibodies on the chips, we probed the chips with Cy3-labeled anti-human immunoglobulin antibodies. The immunogenic profiles of both the IBD patients and HC were acquired by the resulting fluorescent signals. CD *versus* UC *versus* HC can be distinguished by comparing the signal intensities between protein spots on the *E. coli* proteome chips (see the visual appearance of two representative chips probed with sera from CD and HC, respectively; Fig. 2). Two-level data analyses were performed with these immunogenic profiles (i) to identify differentially immunogenic responses among CD *versus* UC *versus* HC using SAM and gene ontology (GO) enrichment analysis and (ii) to construct robust classifiers to distinguish CD *versus* UC *versus* HC using the *k*-TSP method.

TABLE I
Demographic and clinical information of IBD patients and healthy controls
5-ASA, 5-aminosalicylic acid; AZA, azathioprine; 6-MP, 6-mercaptopurine.

Patient data/characteristics	CD (n = 66)	UC (n = 29)	HC (n = 39)
Gender: female (%)	55	53	43
Age: mean $\pm$ S.D. (yr)	36.7 $\pm$ 13.1	38 $\pm$ 14.5	47 $\pm$ 12.4
Age at diagnosis: mean (yr)	36.7	28	
Duration of disease (yr)	12.3	8.9	
Extraintestinal disease: n (%)	7 (11)	3 (10)	
Surgery: n (%)	46 (70)	4 (12)	
Ethnicity: n (%)			
African-American	29 (44)	10 (34)	
Caucasian	35 (53)	22 (75)	
Hispanic	2 (3)	0	
Smoking: n (%)			
Past or present	17 (26)	5 (17)	
Nonsmoker	40 (60)	18 (62)	
Unknown	9 (14)	9 (35)	
Medications: n (%)			
Antibiotics	23 (35)	4 (14)	
5-ASA	49 (74)	26 (90)	
Corticosteroids	16 (24)	9 (35)	
AZA/6-MP	26 (39)	17 (59)	
Methotrexate	0	1 (3)	
Infliximab	15 (23)	2 (7)	
Crohn disease subtype: n (total)			
Non-stricturing and non-penetrating	18 (27)		
Penetrating	26 (40)		
Stricturing	14 (21)		
Penetrating and stricturing	8 (12)		
Ulcerative colitis: n (total)			
Left-sided colitis		13 (45)	
Pancolitis		18 (62)	

Global Immunogenic Profiles of IBD against *E. coli*—Sera samples from HC (n = 39), patients with CD (n = 66), and patients with UC (n = 29) (Table I) were used to compare differences between HC and IBD immunogenic profiles. To investigate the differential global changes in immunogenic response to *E. coli* proteins among HC versus CD versus UC, we applied SAM as described under “Experimental Procedures” to the immunogenic profiles. For convenience, we term the *E. coli* proteins that were differentially recognized by serum antibodies from HC, CD, or UC as “*differentially immunogenic proteins*” throughout this study. Heat maps shown in Fig. 3, A–C, present a visual illustration of the differentially immunogenic proteins for each phenotype. 273 differentially immunogenic proteins were identified by SAM when comparing HC with CD samples. 81 proteins were highly immunogenic in CD samples, and 192 were highly immunogenic in HC samples (Fig. 3A). Conversely 188 proteins had different immunogenic responses in the IBD subtypes: 51 and 137 were highly immunogenic in UC and CD samples, respectively (Fig. 3B). When HC and UC samples were compared, only 27 and six proteins were discriminatory and highly immunogenic in HC and UC samples, respectively (Fig. 3C). A full list of the immunogenic *E. coli* proteins in Fig. 3, A–C, can be found in supplemental Tables S1–S3, respectively.

As shown in the Venn diagram in Fig. 3D, the immunogenic responses to 417 proteins were found to be different between HC and CD or UC. Of these 417 proteins, 169 proteins were identified as highly immunogenic in HC, 186 proteins were highly immunogenic in CD, and only 19 were highly immunogenic in UC. 44 proteins were highly immunogenic in both HC and IBD (CD or UC). Among these 44 proteins, six overlapped between HC and CD, and 38 overlapped between HC and UC. A full list of the immunogenic *E. coli* proteins in Fig. 3D can be found in supplemental Table S4. This demonstrates that UC and HC subjects share more common immunogenic profiles than CD subjects and HC. In general, our results indicate that many of the global immunogenic profiles of sera samples were systematically correlated with either healthy controls or IBD phenotypes and that it may be possible to discriminate sample class based on their immunogenic profile.

Protein Functional Enrichment Analysis—To delineate the immunogenic signatures of the healthy controls and IBD subtypes, we assigned the differentially immunogenic proteins to functional groups based on classification by gene ontology. Functional grouping of the 417 proteins was assigned by querying the databases EcoCyc and Kyoto Encyclopedia of Genes and Genomes as well as cross-checking with the

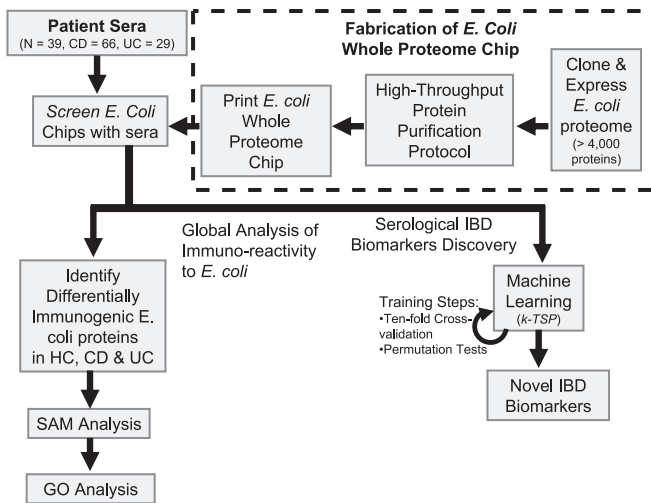


FIG. 1. Overall strategy for the identification of novel serological biomarkers for inflammatory bowel disease using *E. coli* whole proteome chip. To fabricate the whole proteome chip, we cloned and expressed >4,000 *E. coli* proteins. These proteins were purified using a high throughput protein purification protocol and printed onto Full-Moon slides using a ChipWriter Pro robot. 134 patient sera were collected from The Johns Hopkins Hospital for this analysis. These sera were screened by *E. coli* proteome chips. Two-level data analyses were performed: (i) global IBD analysis to identify differentially immunogenic proteins in HC, CD, and UC using SAM and GO enrichment analysis and (ii) serological IBD biomarker discovery using the *k*-TSP algorithm.

Affymetrix *E. coli* genome array annotation file. 338 of these 417 proteins were assigned to at least one GO term, and 78 hypothetical proteins have unknown annotations. We focused our enrichment analysis on six GO cellular component terms (membrane, cell wall, intracellular, macromolecular complex, periplasmic space, and cell projection). To assess whether the selected differentially immunogenic proteins were enriched in one of the GO terms, the hypergeometric statistical test was used to compute the probability of the number of proteins in each cellular component appearing by chance within the proteins highly immunogenic in HC (169), CD (185), and UC (18). Fig. 4 summarizes the enrichment analysis of those proteins that were immunogenic in HC and CD or UC. Antibodies against membrane proteins were highly enriched in HC samples ($p < 0.0001$). Interestingly antibodies against intracellular and macromolecular complex proteins were highly enriched in CD samples ($p < 0.05$), whereas those against cell wall proteins were highly enriched in UC samples ($p < 0.05$). Although 12% of the proteins that were found to be highly immunogenic in CD samples were located in the periplasmic space, their enrichment was not statistically significant ($p = 0.064$) for this IBD subtype. Cell projection proteins were not enriched in either healthy controls or IBD subtypes.

Machine Learning Analysis—Next we sought to construct optimal classifiers from the immunogenic response profiles to differentiate HC from the IBD subtypes (CD and UC) as well as to differentiate CD from UC. Upon successful construction of

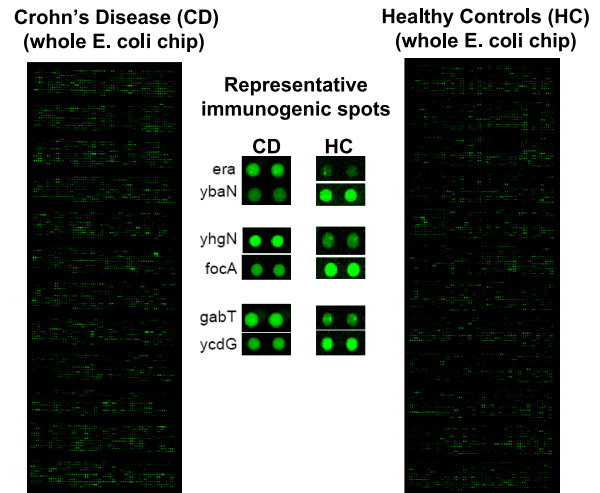


FIG. 2. Representative images of *E. coli* proteome chips probed by sera from CD and HC, respectively. Two *E. coli* proteome chips probed with sera from a CD patient (left panel) and HC (right panel) are shown. To identify the proteins that can be recognized by reactive serum antibodies, each *E. coli* protein chip was incubated with a serum from HC or CD as illustrated in Fig. 1. Cy3-labeled anti-human immunoglobulin antibodies were then probed on the chips, allowing visualization of immunoreactive protein spots. The immunogenic profiles of both the IBD patients and HC were acquired by the resulting fluorescent signals. Green spots are spots of *E. coli* protein in the chips detected by serum antibodies, representing immunogenic reactions. The intensity of the protein spots reflects immunogenicity of the proteins. The middle panels shows some representative images of immunogenic spots of three pairs of specific proteins (for more information on these proteins, see Fig. 5 and Tables I–III) from these proteome chips. Every *E. coli* protein is spotted in duplicate on the chip. CD versus UC versus HC can be distinguished by comparing the signal intensities between protein spots on the *E. coli* proteome chips.

these classifiers, the classification rules may result in the discovery of new robust biomarkers. Here we used *k*-TSP, a novel machine learning method, to discover a simple decision rules classifier from the immunogenic response profiles. The three top scoring pairs were identified as classifiers to differentiate HC samples from CD samples: if the immunogenic reactivity to Era is greater than that to YbaN, then CD or else HC, if YhgN is greater than FocA, then CD or else HC, and if GabT is greater than YdcG, then CD or else HC (see the representative examples of actual images of immunoreactive protein spots in Fig. 2). The algorithm uses a majority vote of the three pairs to classify a sample. Fig. 5A depicts the protein spot ratios for this classifier that separate the data between the two phenotypes where yellow represents a vote for CD and blue represents a vote for HC. Using the *k*-TSP classifier, 36 of 39 HC and 62 of 64 CD samples were correctly classified with an estimated 10-fold cross-validation accuracy of $86 \pm 4\%$ ($p < 0.01$). For distinguishing HC from UC samples, the *k*-TSP algorithm identified 11 feature pairs (Fig. 5B) with an estimated 10-fold cross-validation accuracy of $66 \pm 5\%$ ($p < 0.04$). A single feature pair of *k*-TSP classifier was identified for

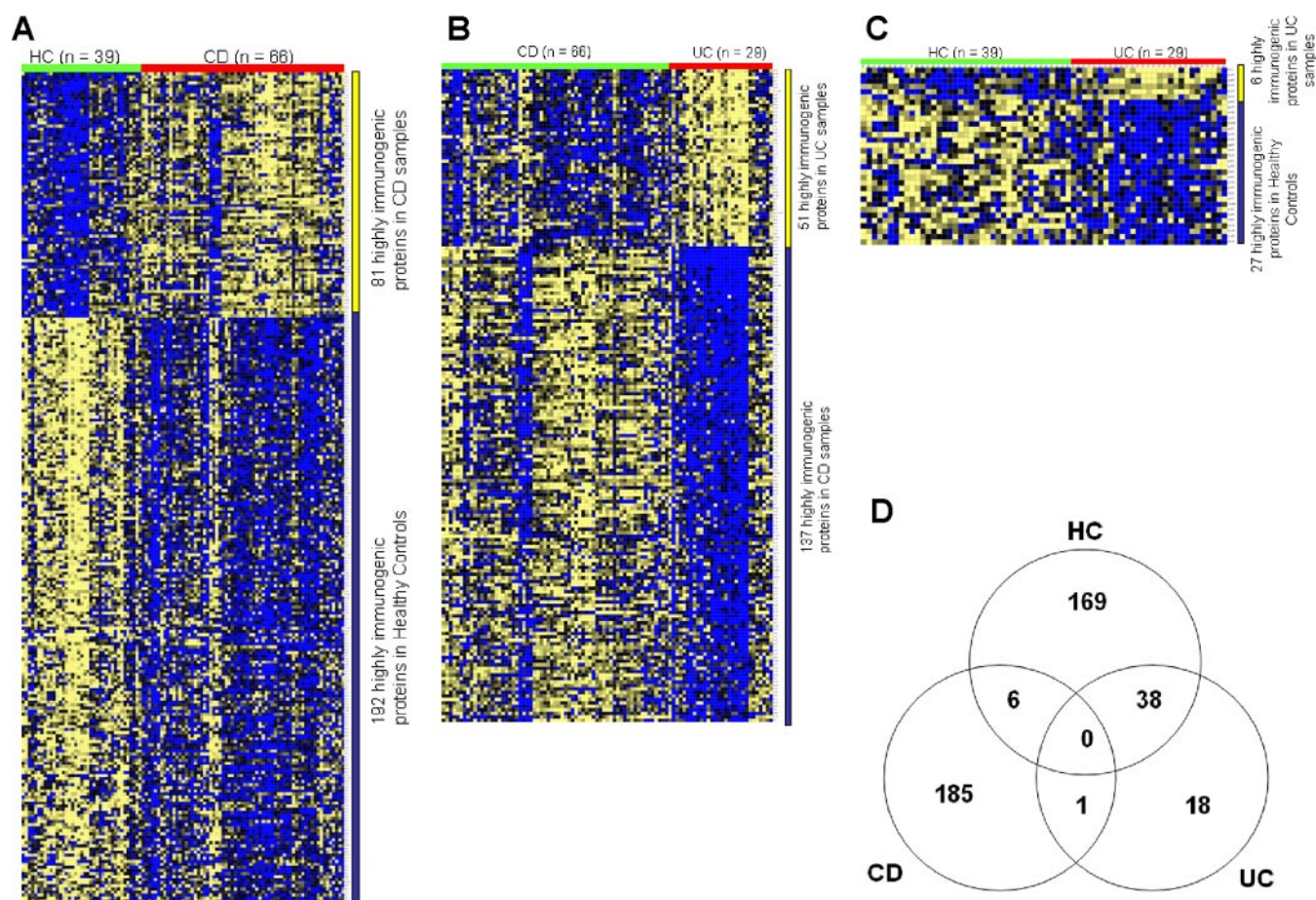


FIG. 3. Global immunogenic profiles of IBD patients' sera against *E. coli* proteins. A, heat map of 273 differentially immunogenic proteins between HC and CD samples identified by SAM. B, heat map of the 188 differentially immunogenic proteins between CD and UC samples identified by SAM. C, 33 differentially immunogenic proteins between healthy controls and UC samples as identified by SAM. Each row corresponds to a protein, and each column corresponds to a sample. The expression level for each protein is normalized across the samples such that the mean is 0 and the S.D. is 1. Yellow and blue colors indicate high and low immunogenic responses, respectively. D, Venn diagram of these differentially immunogenic proteins showing only limited overlap among HC versus CD versus UC.

differentiating CD from UC: if the immunogenic reactivity to FrvX is greater than that to YidX then UC or else CD (as illustrated in Fig. 5C; see representative examples of actual images of immunoreactive protein spots in supplemental Fig. S2). The estimated 10-fold cross-validation accuracy of this classifier is $80 \pm 2\%$ ($p < 0.1$).

We also compared the performance of *k*-TSP with SVM and *k*NN, two other commonly used learning algorithms, for each of the classification problems based on five runs of 10-fold cross-validation. Table II displays the results of 10-fold cross-validation for each of the three classifiers. As demonstrated in Table II, based on cross-validation, *k*-TSP performance meets or exceeds the performance of *k*NN and SVM for these classification problems. Because the cross-validation structure allowed each classifier to test on the same subsets of data as described under "Experimental Procedures," the performance of the three classifiers can be directly compared and tested for statistical significance by a simple Student's *t* test. The HC versus CD *k*-TSP classifier

outperformed the other methods in total classification performance ($p < 0.001$). For the remaining two classification problems, the *k*-TSP classifiers achieved nominally better but not statistically significant classification accuracy when compared with SVM and *k*NN classifiers.

From this study, we found that *k*-TSP performs much better than SVM and *k*NN in separating HC from CD. In addition, the ordering of the expression values within profiles are utilized in the *k*-TSP decision rules; therefore, the classifier is invariant to data preprocessing (29). Supplemental Fig. S3, A and B, shows that on their own the immunogenic responses to Era and YbaN (the top scoring pair in the HC versus CD *k*-TSP classifier) did not allow for class separation of the data; no threshold level would clearly separate HC from CD. However, the ratio of the two features (top scoring pair ratio) resulted in clear separation in the data, lending well to classification (supplemental Fig. S3C). Similar results were obtained when scatter plot analysis was done for the other two TSP pairs from the HC versus CD classifier (YhgN versus FocA and

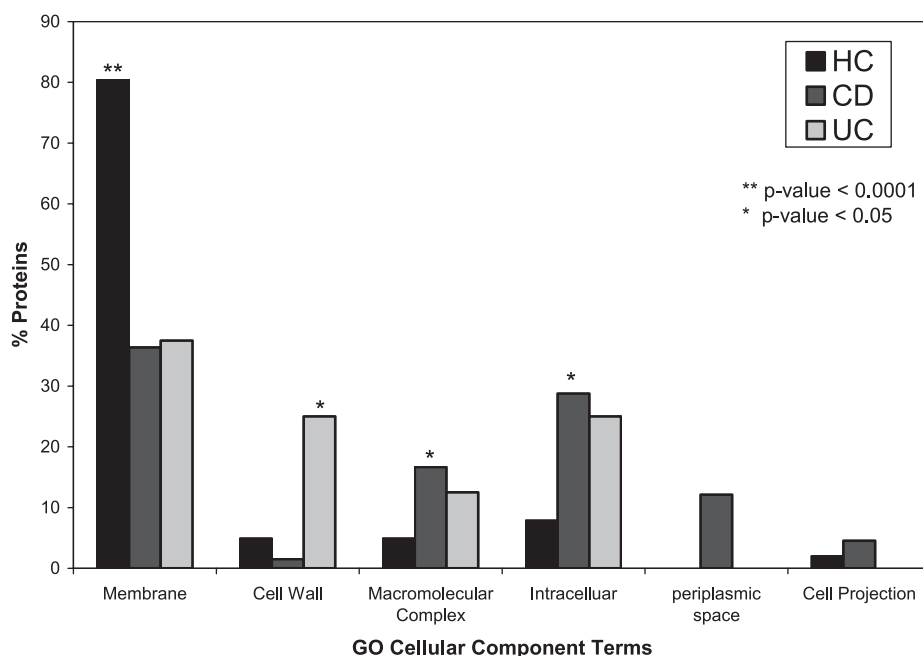


FIG. 4. Distribution of the cellular component terms in the highly immunogenic response proteins of HC, CD, and UC. Six gene ontology cellular component terms were examined. The cell projection term contains flagellum and fimbrium proteins. The main messages include the following. 1) ~80% of the highly immunogenic proteins were either membrane proteins in HC ($p < 0.0001$) compared with only ~37% of the top immunogenic proteins in CD patients (not statistically significant). 2) Conversely ~30% of the top immunogenic proteins in CD patients were intracellular proteins ($p < 0.05$) compared with only ~7% in HC (not statistically significant). 3) A significantly higher percentage of cell wall proteins (~26%) were immunogenic in UC ($p < 0.05$) compared with those in HC and CD (not significant). 4) A significant percentage of macromolecular complex proteins (~16%; $p < 0.05$) were immunogenic in CD compared with those in HC or UC (not statistically significant). No statistically significant enrichment of periplasmic space and cell projection proteins was found in HC, CD, and UC.

GabT versus YcdG, respectively; data not shown). This represents an advantage of *k*-TSP over other learning methods where interpreting the decision rules are easy and can facilitate follow-up study. It is important to note that SAM identified Era as the second best individual marker for up-regulation in CD; thus it appears that individual markers will not work well for classification, and this explains why *k*NN and SVM fail to match the performance of *k*-TSP as the relative feature levels within samples appear to be much more robust than the absolute feature levels across samples.

Robustness of the *k*-TSP Classifiers—To determine that class imbalance did not greatly affect the classification results, we performed additional analysis where samples were randomly discarded from the class with a greater total number of samples to equalize the class sizes. 10-fold cross-validation was performed as described above. The process was then repeated by discarding a different random set of samples. Table III shows the performance of each classifier given class balance in the training set. It demonstrates that *k*-TSP outperforms SVM and *k*NN in most instances whether or not the class size is balanced, further supporting the data presented in Table II.

Next to determine the significance of each classifier, a permutation test was performed by randomly shuffling the class labels while maintaining the same number of samples in each class. 10-fold cross-validation was carried out to yield a

classification rate for the permutation set. 100 permutations were performed to get a null distribution of expected classification rates by chance. The classification rate from the unpermuted data was then compared with the null distribution to determine significance. Table III shows the permutation test results for all the classification problems. For the *k*-TSP classifiers trained to differentiate between HC and CD samples as well as CD and UC samples, no permuted set achieved classification rates equal or superior to the original data out of 100 permutations. Thus, these classifiers were estimated to be significant at the $p < 0.01$ level. The *k*-TSP classifier built to differentiate HC and UC had four of 100 permutations achieve rates that matched or exceeded the original classifier; thus this classifier is near the typical significance threshold at $p < 0.05$.

Finally to gauge the robustness of the classification rules discovered by the *k*-TSP method, we inspected the surrogate classifiers created during the 10-fold cross-validation procedure. Every loop of cross-validation creates a separate classifier used to predict the left out sample classes; these are called surrogate classifiers. Thus, for each problem of interest for which we performed 10-fold cross-validation in Table III, there were 50 classifiers to inspect (10 for each of the five runs). The percentage of the time that the rule from the final *k*-TSP classifier shows up in the 50 surrogate classifiers is an indicator of the robustness of that

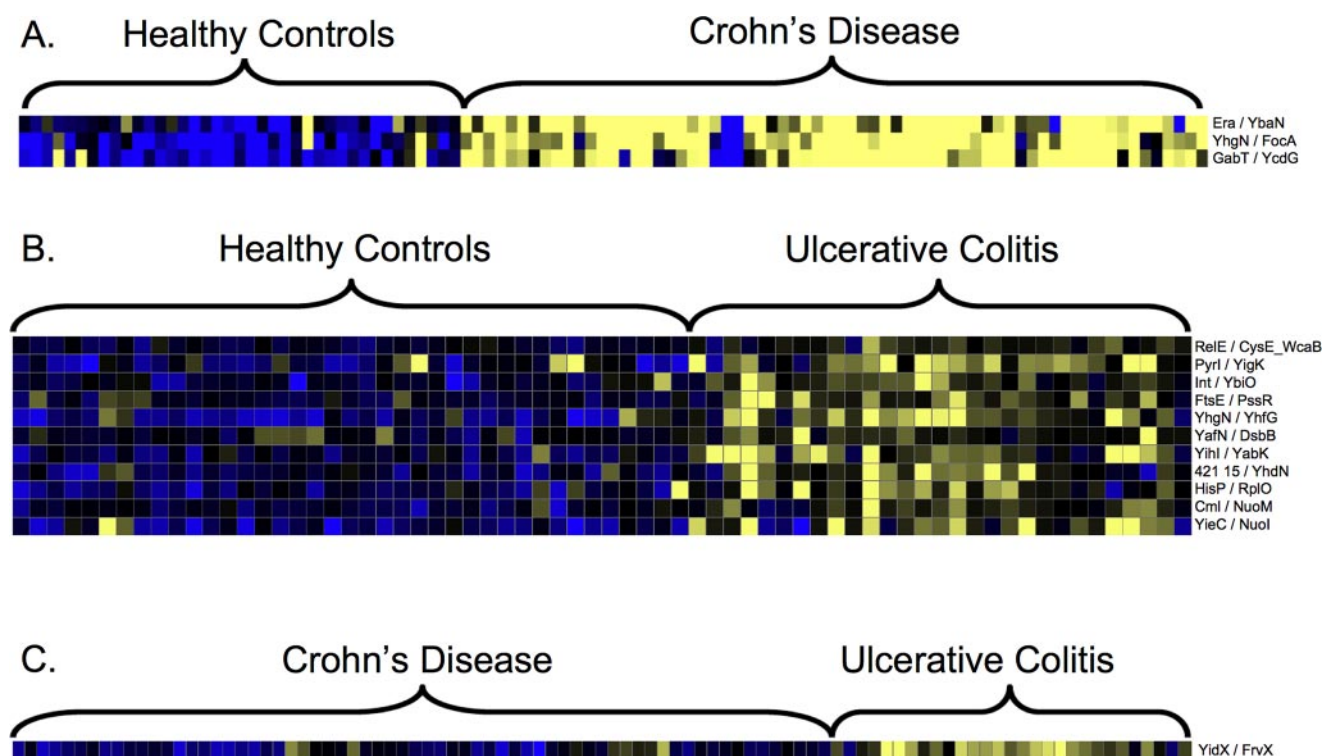


FIG. 5. *k*-TSP-identified top three pairs of biomarkers that can discriminate controls from CD patients. Each column represents the immunogenic reactivity by individual IBD patients or HC. Within a column, each row represents the ratio of the immunogenic reactivity of a top scoring pair of proteins. The expression values represented are the ratios of immunogenic reactivity (fluorescent signal or intensity) to protein X divided by the signals to protein Y, referred to as the TSP ratio (X and Y are example proteins). If the immunogenic reactivity of a patient to protein X was greater than the reactivity to protein Y, then the box will appear yellow (blue for vice versa) (see examples below). A depicts the classifier for HC versus CD (yellow, CD; blue, HC). For example, if Era is greater than YbaN, it will be a CD (yellow) or else a HC (blue). B displays the HC versus UC classifier (yellow, UC; blue, HC). For example, if RelE is greater than CysE/WcaB, it is a UC (yellow), or else it is classified as a HC (blue). C shows the CD versus UC classifier (yellow, UC; blue, CD). If FrvX is greater than or equal to YidX, it is a UC (yellow) or else a CD (blue). See the representative images of some of these protein pairs in Fig. 2 and supplemental Fig. S2.

rule. Table III shows that the pairs that showed up in the HC versus CD classifier as well as the UC versus CD classifier are fairly robust whereas the pairs in the HC versus UC classifier are not. Along with the permutation testing, this indicates that the HC versus CD and UC versus CD classifiers should perform well in independent testing whereas the HC versus UC classifier may not.

Stratifying CD Subtypes and Risk for Surgery—Certain antibody-based serological biomarkers (such as pANCA and ASCA) have shown promise in risk stratifying patients prior to instituting medical therapy or embarking on surgery. As an example, the presence of pANCA has been associated with the development of acute and chronic pouchitis after colectomy with ileal pouch-anal anastomosis. Similarly the presence of high titers of ASCA has been found to predict the occurrence of pouch complications and a more complicated disease course in Crohn disease (30, 31). To evaluate whether the new biomarkers we identified can be used to stratify CD and UC subtypes or risk for surgery, we used the Vienna classification to subtype patients with CD into the following behavior subtypes (Table I): penetrating/fistulizing, stricturing, penetrating/stricturing, and non-penetrating/non-stricturing.

Patients with UC were divided into those with left-sided disease (inflammation extending no further than the splenic flexure). Pancolitis was considered to be continuous inflammation from the rectum extending beyond the splenic flexure. We found by *k*-TSP analyses that these newly identified markers performed poorly in stratifying subtypes of CD or UC or risk for surgery most likely because of the smaller sample size of each disease type (data not shown; see sample sizes of CD or UC in Table I).

OmpC and FliC, Two of the Known Serological Markers, Performed Poorly—Although anti-OmpC and anti-CBir (FliC) have been recently considered two new IBD serological biomarkers, they were not picked up in our screening of the *E. coli* K12 proteome. Scatter plot (supplemental Fig. S4) analysis of *E. coli* OmpC and FliC demonstrates that neither allowed for class separation between control versus CD versus UC; no threshold level would clearly separate the data.

DISCUSSION

Protein microarrays have been demonstrated to be a powerful tool to identify biomarkers (24, 32–34). We present here the first study to identify serological biomarkers in human

TABLE II

Estimated 10-fold cross-validation classification rates of IBD using the three described classification methods

The reported rates are given in percentages and are the mean performance on all five runs of 10-fold cross-validation $\pm$ S.D. In parentheses are the numbers of samples in each subtype used for classification. Sp, specificity; Sn, sensitivity; PPV, positive predictive value; NPV, negative predictive value.

Method	Accuracy	Sp (HC)	Sn (CD)	PPV (CD)	NPV (HC)
	%	%	%	%	%
HC (39) vs. CD (66)					
k-TSP	86 $\pm$ 4	81 $\pm$ 5	89 $\pm$ 3	89	81
SVM	70 $\pm$ 2	66 $\pm$ 1	73 $\pm$ 2	79	59
kNN	63 $\pm$ 3	47 $\pm$ 7	73 $\pm$ 6	70	50
HC (39) vs. UC (29)					
k-TSP	66 $\pm$ 5	69 $\pm$ 5	61 $\pm$ 7	59	70
SVM	62 $\pm$ 5	58 $\pm$ 1	68 $\pm$ 12	55	71
kNN	60 $\pm$ 6	57 $\pm$ 2	64 $\pm$ 12	53	68
CD (66) vs. UC (29)					
k-TSP	80 $\pm$ 2	84 $\pm$ 1	70 $\pm$ 6	86	66
SVM	78 $\pm$ 3	82 $\pm$ 2	69 $\pm$ 9	86	63
kNN	78 $\pm$ 3	78 $\pm$ 4	61 $\pm$ 2	82	55

TABLE III

Permutation statistics for each pair of biomarkers

Top scoring pairs used for each classifier and the percentage of surrogate classifiers in which those pairs appear during 10-fold cross-validation (mean $\pm$ S.D., *p* value). CV, coefficient of variation.

<i>k</i> -TSP classifier	Accuracy		Estimated <i>p</i> value	Features in <i>k</i> -TSP classifier	Percent feature appearance
	10-fold CV	10-fold CV permutation			
	%				
HC vs. CD	86 $\pm$ 3	50 $\pm$ 8	<0.01	Era > YbaN = CD	90
				YhgN > FocA = CD	84
				GabT > YcdG = CD	64
HC vs. UC	66 $\pm$ 5	51 $\pm$ 9	<0.05	RelE > CysE/WcaB = UC	80
				Pylr > YjgK = UC	42
				Lnt > YbiO = UC	36
				FtsE > PssR = UC	36
				YhgN > YhfG = UC	20
				YafN > DsbB = UC	20
				Yihl > YabK = UC	26
				421 15 > YhdN = UC	24
				HisP > RplO = UC	16
				Cml > NuoM = UC	14
				YieC > NuoI = UC	12
UC vs. CD	80 $\pm$ 2	52 $\pm$ 6	<0.01	YidX > FrvX = UC	88

immunological diseases using a protein chip of whole prokaryotic proteome.

The significance of this study is 3-fold. First, the first proof of principle for the feasibility of the application of high density protein microarray/chip technology in the discovery of novel serological IBD biomarkers is presented. This study can serve as an example of similar proteomics approaches for hunting for serological biomarkers for other immune system-related diseases, such as autoimmune disorders. Second, this is also the first effort to examine human immune responses to the entire proteome of a microbial species under normal or any disease condition. It is surprising to learn that human circulating antibodies can recognize more than 400 *E. coli* proteins (Fig. 3D). Because it has been demonstrated that defective intestinal barrier function plays a central role in the pathogen-

esis of CD (35, 36), it is conceivable that in patients with CD commensal bacteria or their products could more readily penetrate intestinal epithelia. Therefore, it is less surprising that 185 of the *E. coli* proteins were recognized by sera from CD patients (Fig. 3D). However, it remains a mystery why there are a large number (185) of immunogenic *E. coli* proteins that were found to be specific in healthy controls whereas only 18 immunogenic proteins were found to be specific to UC. It is worth noting that in a previous report the development of colitis in T cell receptor α knock-out mice (a UC-like murine IBD model) was associated with restricted humoral responses to selected *E. coli* proteins (37). This report supports our finding that patients with UC had decreased immune responses to intestinal bacteria. Third, we identified a set of novel serological biomarkers that had >80% overall

accuracy and sensitivity in differentiating CD from HC or UC.

An intriguing observation in our study is the difference in the immunogenicity of surface/membrane *versus* intracellular proteins in HC *versus* CD patients. Approximately 85% of the highly immunogenic proteins were either cell wall proteins or membrane proteins in HC compared with only ~37% of the top immunogenic proteins in CD patients (Fig. 4 and supplemental Tables S1–S3). Conversely ~30% of top immunogenic proteins in CD patients were intracellular proteins compared with only ~7% in HC (Fig. 4 and supplemental Tables S1–S3). Furthermore there was no overlap among the top immunogenic *E. coli* surface/membrane proteins among the three distinct populations (HC, CD, and UC; see Fig. 3D). This suggests that the host immunological response to *E. coli* is drastically different between HC and CD patients. The mechanism involved in these immunogenic differences is not clear at this moment. We postulate that, in immunologically healthy hosts where *E. coli* are largely confined to the luminal side of the gut because of the intestinal epithelial barrier, surface and membrane proteins of *E. coli* might be the primary antigens that are more accessible to the immune system compared with intracellular proteins. In this case, the immune system has adapted to the presence of luminal *E. coli*. In contrast, in CD patients, a disrupted or compromised intestinal barrier (35, 36) may lead to the bacterium or its products crossing the gut luminal barrier. If the whole *E. coli* invades into the lamina propria, it will mostly likely be lysed by the host immune system. Subsequently *E. coli* components such as intracellular proteins that otherwise would not be seen by the intestinal immune system in the lamina propria are presented by antigen-presenting cells (such as macrophages or dendritic cells). This may dramatically alter the previously adapted immune system that is only accustomed to the luminally exposed *E. coli*, resulting in an overwhelming production of antibodies against these intracellular *E. coli* proteins. The consequences of these immune responses include recruitment of various inflammatory immune cells such as neutrophils, dendritic cells, and lymphocytes to lamina propria or between colonic epithelial cells, leading to dysregulated mucosal inflammation. This hypothesis may also explain why there were only six overlapping proteins among the 354 top immunogenic proteins recognized by HC and CD patients (Fig. 3D).

Biological Significance of the Novel IBD Serological Biomarkers—None of the serum antibody biomarkers that were identified here for discriminating CD from HC or UC have been described previously. Although most of the antigens (*E. coli* proteins) responsible for generation of these marker antibodies have not been well characterized, their identity and function can be predicted based on their sequence information. Among the proteins in the *k*-TSP classifier, Era, YbaN, YhgN, FocA, GabT, and YcdG (Fig. 5A), for discriminating CD from HC, Era is a GTP-binding protein that is involved in the binding of GTP and nucleotides of the cell cycle and can be found

in the intracellular membrane. In this study, an increased immunogenic response to Era was associated with CD as identified by both SAM and *k*-TSP analyses. YbaN is predicted to be a conserved inner membrane protein with unknown function. YhgN is predicted to be an inner six-transmembrane domain protein of which the C terminus is located in the periplasm (38). YcdG (also called RutG) is another predicted transmembrane protein with 11 helices; the C terminus of the protein is located on the cytoplasmic side of the inner membrane (38). This protein is predicted to be involved in pyrimidine utilization in *E. coli* where it may function as a proton-driven uracil uptake system (39). FocA, an inner membrane protein, is a putative formate transporter that may be involved in both formate uptake and efflux. Disruption of the FocA gene confers resistance to hypophosphite, a toxic formate analogue (40). GabT, 4-aminobutyrate aminotransferase, is a well characterized protein and acts as the initial enzyme of the 4-aminobutyrate degradation pathway in *E. coli* (41). Among the pair of proteins (FrvX and YidX) that were identified to be discriminatory between CD and UC, FrvX is an important protein in the fructose-specific phosphoenolpyruvate-dependent sugar phosphotransferase system (42), and YidX is a predicted lipoprotein the function of which is currently unknown.

Like all the previously identified serological (antibody) biomarkers, including pANCA, ASCA, anti-OmpC, anti-I2, and anti-CBir, the pathological or functional consequences of having these newly identified circulating antibodies are unclear. From the available information on these eight proteins described above, we are unable to explain why they were found, among >4,200 proteins in the *E. coli* proteome, to be discriminatory between HC and CD or between CD and UC. Although the circulating antibodies against specific microbes can be used as biomarkers, it is most likely that these antibodies are made for some specific purposes/functions, either physiological or pathological. It may take a while for us to eventually have an answer. If at least some of the antibodies are pathological, an interesting question will be whether we can use the information to develop IBD-specific vaccines in the future.

Implication of the Novel IBD Serological Markers—The biomarkers newly identified by *k*-TSP analysis had a particularly impressive ~86% accuracy in differentiating CD from HC with a specificity of ~81% and a sensitivity of ~89% (Table II). In addition, *k*-TSP analysis yielded an accuracy of ~80% in differentiating CD and UC with a sensitivity of ~84% and specificity of ~70% (Table II). These demonstrate that the sensitivity and specificity of these novel serological markers are comparable to those of a combination of the multiple best characterized IBD biomarkers (ASCA, pANCA, anti-OmpC, and anti-CBir) (43, 44). More importantly, an identical performance was achieved by using only the top three pairs of *E. coli* proteins for discriminating healthy controls *versus* CD and one top pair of proteins for differentiating CD *versus* UC (Figs. 2 and 5 and Tables II and III). These data provide a critical

feasibility for 1) a validation study using additional larger cohorts of IBD patients and controls and 2) future development of novel assay kits for diagnosis of CD and UC. However, it is necessary to point out that our current approach screening an *E. coli* protein array is not suitable for identifying serological biomarkers in differentiating UC from HC (only ~66% accuracy) (Tables II and III). Importantly OmpC, an *E. coli* antigen for one of the widely studied current serological biomarker (anti-OmpC), was not picked up in our screen (supplemental Fig. S4A). Similarly FliC, an *E. coli* flagellin protein equivalent to *Salmonella* flagellin (which is the antigen for anti-CBir, another widely studied antibacterial antibody) did not show up in our analysis (supplemental Fig. S4A). These data suggest that anti-OmpC and at least the antibody against *E. coli* FliC are not robust serological biomarkers for IBD.

In conclusion, we have presented here the first demonstration that using a protein array to screen circulating disease-specific antibodies is a robust, effective, and high throughput approach for discovery of novel biomarkers of IBD. This approach can be readily applied to screen serological biomarkers of various autoimmune diseases and/or even infectious diseases.

* This work was supported, in whole or in part, by National Institutes of Health Grants 5R21DK77064 and KO1 DK62264 from the NIDDK. This work was also supported by Broad Medical Research Program Grant IBD-0119R and by Mr. and Mrs. Morton Hyatt (to X. L.).

§ The on-line version of this article (available at <http://www.mcponline.org>) contains supplemental material.

© These authors contributed equally to this work.

† To whom correspondence may be addressed: Dept. of Pharmacology and HiT Center, Johns Hopkins University School of Medicine, BRB 333, 733 Broadway, Baltimore, MD 21205. Tel.: 410-502-0878; Fax: 410-502-1872; E-mail: Heng.Zhu@jhmi.edu.

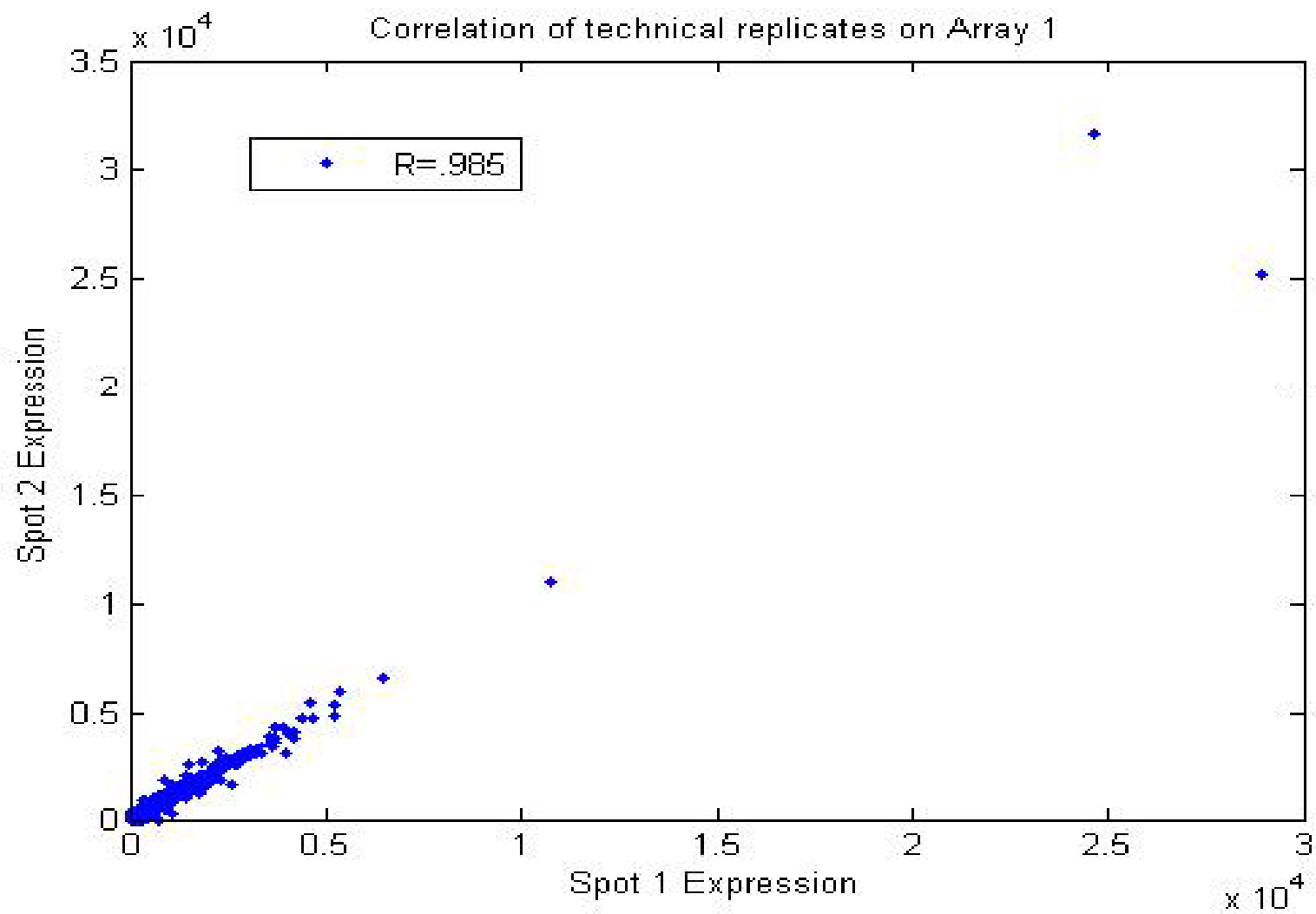
‡ To whom correspondence may be addressed: GI Division, Dept. of Medicine, Johns Hopkins University School of Medicine, 918 Ross Research Bldg., 720 Rutland Ave., Baltimore, MD 21205. Tel.: 443-287-4804; Fax: 410-955-9677; E-mail: xuhang@jhmi.edu.

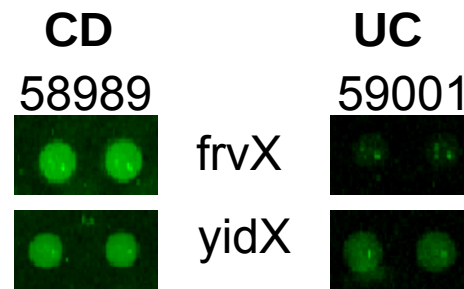
REFERENCES

- Podolsky, D. K. (2002) Inflammatory bowel disease. *N. Engl. J. Med.* **347**, 417–429
- Xavier, R. J., and Podolsky, D. K. (2007) Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448**, 427–434
- Geboes, K., and De Hertogh, G. (2003) Indeterminate colitis. *Inflamm. Bowel Dis.* **9**, 324–331
- Guindi, M., and Riddell, R. H. (2004) Indeterminate colitis. *J. Clin. Pathol.* **57**, 1233–1244
- Li, X., Conklin, L., and Alex, P. (2008) New serological biomarkers of inflammatory bowel disease. *World J. Gastroenterol.* **14**, 5115–5124
- Peyrin-Biroulet, L., Standaert-Vitse, A., Branche, J., and Chamailard, M. (2007) IBD serological panels: facts and perspectives. *Inflamm. Bowel Dis.* **13**, 1561–1566
- Vermeire, S., Vermeulen, N., Van Assche, G., Bossuyt, X., and Rutgeerts, P. (2008) (Auto)antibodies in inflammatory bowel diseases. *Gastroenterol. Clin. North Am.* **37**, 429–438, vii
- Strober, W., Fuss, I. J., and Blumberg, R. S. (2002) The immunology of mucosal models of inflammation. *Annu. Rev. Immunol.* **20**, 495–549
- Blumberg, R. S., Saubermann, L. J., and Strober, W. (1999) Animal models of mucosal inflammation and their relation to human inflammatory bowel disease. *Curr. Opin. Immunol.* **11**, 648–656
- Papp, M., Altörjay, I., Norman, G. L., Shums, Z., Palatka, K., Vitalis, Z., Foldi, I., Lakos, G., Tumpek, J., Udvardy, M. L., Harsfalvi, J., Fischer, S., Lakatos, L., Kovacs, A., Bene, L., Molnar, T., Tulassay, Z., Miheller, P., Veres, G., Papp, J., and Lakatos, P. L. (2007) Seroreactivity to microbial components in Crohn's disease is associated with ileal involvement, noninflammatory disease behavior and NOD2/CARD15 genotype, but not with risk for surgery in a Hungarian cohort of IBD patients. *Inflamm. Bowel Dis.* **13**, 984–992
- Wirtz, S., and Neurath, M. F. (2007) Mouse models of inflammatory bowel disease. *Adv. Drug Deliv. Rev.* **59**, 1073–1083
- Xavier, R., and Podolsky, D. K. (2005) Commensal flora: wolf in sheep's clothing. *Gastroenterology* **128**, 1122–1126
- Papp, M., Altörjay, I., Dotan, N., Palatka, K., Foldi, I., Tumpek, J., Sipka, S., Udvardy, M., Dinya, T., Lakatos, L., Kovacs, A., Molnar, T., Tulassay, Z., Miheller, P., Norman, G. L., Szamosi, T., Papp, J., and Lakatos, P. L. (2008) New serological markers for inflammatory bowel disease are associated with earlier age at onset, complicated disease behavior, risk for surgery, and NOD2/CARD15 genotype in a Hungarian IBD cohort. *Am. J. Gastroenterol.* **103**, 665–681
- Sandborn, W. J. (2004) Serologic markers in inflammatory bowel disease: state of the art. *Rev. Gastroenterol. Disord.* **4**, 167–174
- Ferrante, M., Henckaerts, L., Joossens, M., Pierik, M., Joossens, S., Dotan, N., Norman, G. L., Altstock, R. T., Van Steen, K., Rutgeerts, P., Van Assche, G., and Vermeire, S. (2007) New serological markers in inflammatory bowel disease are associated with complicated disease behaviour. *Gut* **56**, 1394–1403
- Kornbluth, A., and Sachar, D. B. (2004) Ulcerative colitis practice guidelines in adults (update): American College of Gastroenterology, Practice Parameters Committee. *Am. J. Gastroenterol.* **99**, 1371–1385
- Alex, P., Gucsek, M., and Li, X. (2009) Applications of proteomics in the study of inflammatory bowel diseases: current status and future directions with available technologies. *Inflamm. Bowel Dis.* **15**, 616–629
- Goldknopf, I. L. (2008) Blood-based proteomics for personalized medicine: examples from neurodegenerative disease. *Expert Rev. Proteomics* **5**, 1–8
- Maurya, P., Meleady, P., Dowling, P., and Clynes, M. (2007) Proteomic approaches for serum biomarker discovery in cancer. *Anticancer Res.* **27**, 1247–1255
- Veenstra, T. D., Conrads, T. P., Hood, B. L., Avellino, A. M., Ellenbogen, R. G., and Morrison, R. S. (2005) Biomarkers: mining the biofluid proteome. *Mol. Cell. Proteomics* **4**, 409–418
- VanMeter, A., Signore, M., Pierobon, M., Espina, V., Liotta, L. A., and Petricoin, E. F., 3rd (2007) Reverse-phase protein microarrays: application to biomarker discovery and translational medicine. *Expert Rev. Mol. Diagn.* **7**, 625–633
- Cekaite, L., Hovig, E., and Sioud, M. (2007) Protein arrays: a versatile toolbox for target identification and monitoring of patient immune responses. *Methods Mol. Biol.* **360**, 335–348
- Hamelinck, D., Zhou, H., Li, L., Verweij, C., Dillon, D., Feng, Z., Costa, J., and Haab, B. B. (2005) Optimized normalization for antibody microarrays and application to serum-protein profiling. *Mol. Cell. Proteomics* **4**, 773–784
- Steller, S., Angenendt, P., Cahill, D. J., Heuberger, S., Lehrach, H., and Kreutzberger, J. (2005) Bacterial protein microarrays for identification of new potential diagnostic markers for *Neisseria meningitidis* infections. *Proteomics* **5**, 2048–2055
- Chen, C. S., Korobkova, E., Chen, H., Zhu, J., Jian, X., Tao, S. C., He, C., and Zhu, H. (2008) A proteome chip approach reveals new DNA damage recognition activities in *Escherichia coli*. *Nat. Methods* **5**, 69–74
- Saka, K., Tadenuma, M., Nakade, S., Tanaka, N., Sugawara, H., Nishikawa, K., Ichiyoshi, N., Kitagawa, M., Mori, H., Ogasawara, N., and Nishimura, A. (2005) A complete set of *Escherichia coli* open reading frames in mobile plasmids facilitating genetic studies. *DNA Res.* **12**, 63–68
- Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R. A., Gerstein, M., and Snyder, M. (2001) Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105
- Tusher, V. G., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 5116–5121
- Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L., and Geman, D. (2005)

- Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* **21**, 3896–3904
30. Dendrinou, K. G., Becker, J. M., Stucchi, A. F., Saubermann, L. J., LaMorte, W., and Faray, F. A. (2006) Anti-Saccharomyces cerevisiae antibodies are associated with the development of postoperative fistulas following ileal pouch-anal anastomosis. *J. Gastrointest. Surg.* **10**, 1060–1064
 31. Vasilias, E. A., Kam, L. Y., Karp, L. C., Gaiennie, J., Yang, H., and Targan, S. R. (2000) Marker antibody expression stratifies Crohn's disease into immunologically homogeneous subgroups with distinct clinical characteristics. *Gut* **47**, 487–496
 32. Li, B., Jiang, L., Song, Q., Yang, J., Chen, Z., Guo, Z., Zhou, D., Du, Z., Song, Y., Wang, J., Wang, H., Yu, S., Wang, J., and Yang, R. (2005) Protein microarray for profiling antibody responses to Yersinia pestis live vaccine. *Infect. Immun.* **73**, 3734–3739
 33. Hudson, M. E., Pozdnyakova, I., Haines, K., Mor, G., and Snyder, M. (2007) Identification of differentially expressed proteins in ovarian cancer using high-density protein microarrays. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 17494–17499
 34. Horn, S., Lueking, A., Murphy, D., Staudt, A., Gutjahr, C., Schulte, K., König, A., Landsberger, M., Lehrach, H., Felix, S. B., and Cahill, D. J. (2006) Profiling humoral autoimmune repertoire of dilated cardiomyopathy (DCM) patients and development of a disease-associated protein chip. *Proteomics* **6**, 605–613
 35. Gibson, P. R. (2004) Increased gut permeability in Crohn's disease: is TNF the link? *Gut* **53**, 1724–1725
 36. Laukoetter, M. G., Nava, P., and Nusrat, A. (2008) Role of the intestinal barrier in inflammatory bowel disease. *World J. Gastroenterol.* **14**, 401–407
 37. Mizoguchi, A., Mizoguchi, E., Tonegawa, S., and Bhan, A. K. (1996) Alteration of a polyclonal to an oligoclonal immune response to cecal aerobic bacterial antigens in TCR alpha mutant mice with inflammatory bowel disease. *Int. Immunol.* **8**, 1387–1394
 38. Daley, D. O., Rapp, M., Granseth, E., Melén, K., Drew, D., and von Heijne, G. (2005) Global topology analysis of the Escherichia coli inner membrane proteome. *Science* **308**, 1321–1323
 39. Loh, K. D., Gyaneshwar, P., Markenscoff Papadimitriou, E., Fong, R., Kim, K. S., Parales, R., Zhou, Z., Inwood, W., and Kustu, S. (2006) A previously undescribed pathway for pyrimidine catabolism. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5114–5119
 40. Suppmann, B., and Sawers, G. (1994) Isolation and characterization of hypophosphite-resistant mutants of Escherichia coli: identification of the FocA protein, encoded by the pfl operon, as a putative formate transporter. *Mol. Microbiol.* **11**, 965–982
 41. Dover, S., and Halpern, Y. S. (1972) Utilization of gamma-aminobutyric acid as the sole carbon and nitrogen source by Escherichia coli K-12 mutants. *J. Bacteriol.* **109**, 835–843
 42. Reizer, J., Michotey, V., Reizer, A., and Saier, M. H., Jr. (1994) Novel phosphotransferase system genes revealed by bacterial genome analysis: unique, putative fructose- and glucoside-specific systems. *Protein Sci.* **3**, 440–450
 43. Bossuyt, X. (2006) Serologic markers in inflammatory bowel disease. *Clin. Chem.* **52**, 171–181
 44. Austin, G. L., Herfarth, H. H., and Sandler, R. S. (2007) A critical evaluation of serologic markers for inflammatory bowel disease. *Clin. Gastroenterol. Hepatol.* **5**, 545–547

Fig. S1





IF $\text{frvX} \geq \text{yidX}$ THEN CD, ELSE UC

Fig. S2

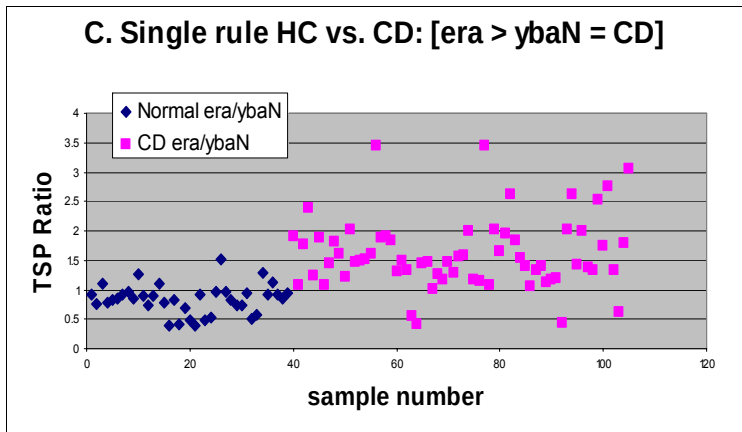
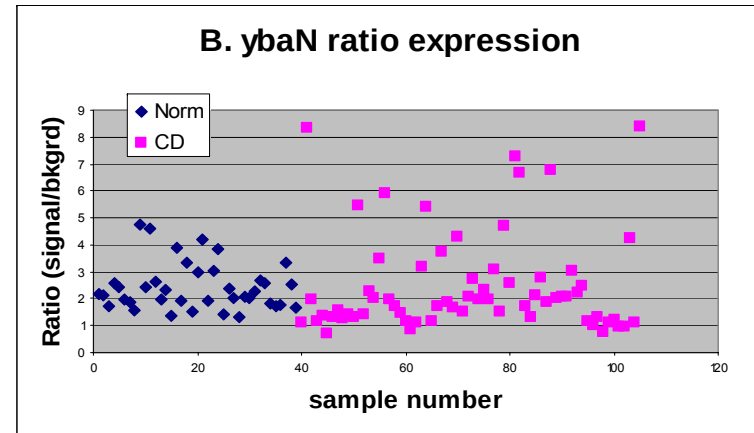
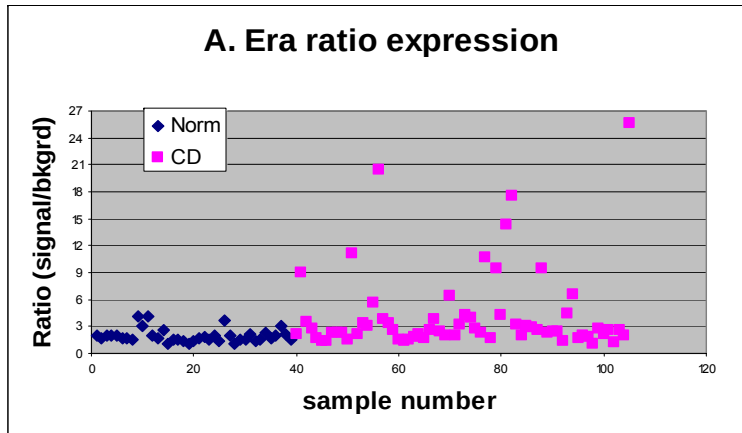
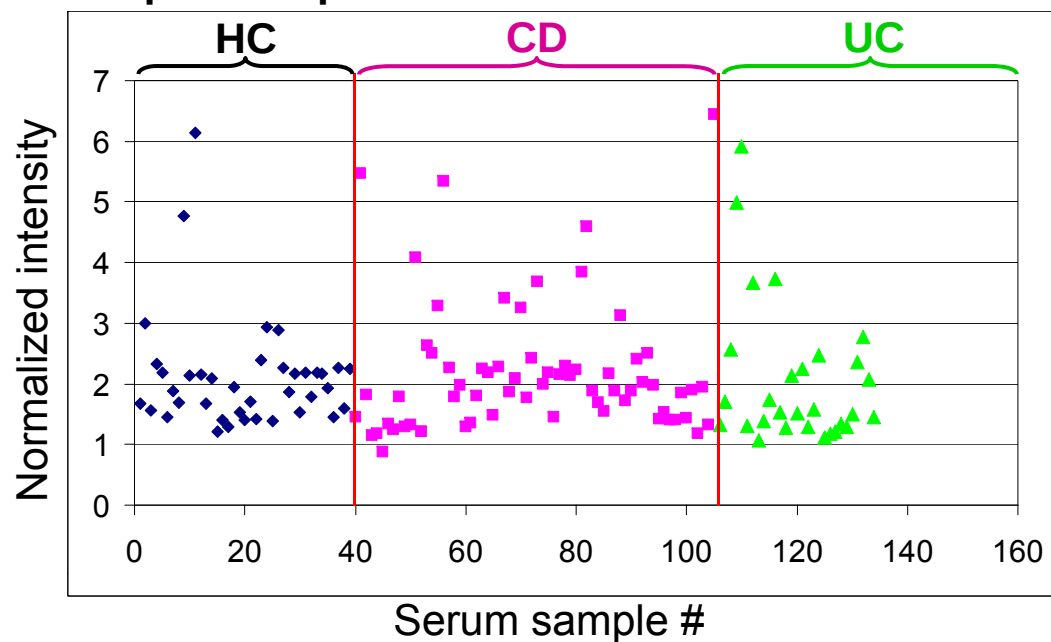


Fig. S3

A. ompC ratio plot



B. fliC ratio plot

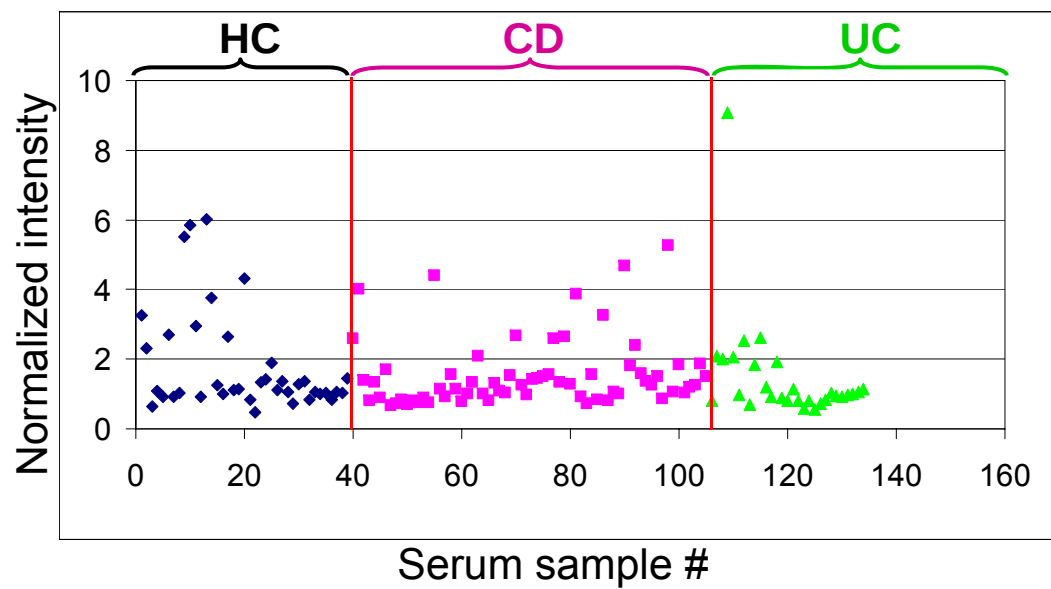


Fig. S4