Identification of novel serological biomarkers for inflammatory bowel disease

using *E. coli* proteome chip

Chien-Sheng Chen1,7*, Sean Sullivan2,3*, Troy Anderson4*, Aik Choon Tan4,5*, Philip J. Alex2,
Steven R. Brant2, Carmen Cuffari3, Theodore M. Bayless2, Monica V. Talor6, C. Lynne Burek6,
Huan Wang2, Richard Li2, Lisa Wu Datta2, Yuqiong Wu2, Raimond L. Winslow4, Heng Zhu1§, and
Xuhang Li2§

Dept of Pharmacology & Molecular Sciences1, The Meyerhoff Inflammatory Bowel Disease
Center and GI Division/Dept of Medicine2, Dept of Pediatrics3, Institute of Computational
Medicine4, Dept of Oncology5, Dept of Pathology6, Johns Hopkins University School of Medicine,
Baltimore, MD 21205; Graduate Institute of Systems Biology and Bioinformatics7, National
Central University, Jhongli 32001, Taiwan

* These authors contributed equally to this work.

Running title: Novel serological IBD biomarkers by protein chip technology

§Address correspondence to:

Xuhang Li, Ph.D.
918 Ross Research Bldg, 720 Rutland Ave
GI Division, Dept of Medicine
Johns Hopkins University School of Medicine
Baltimore, MD 21205
Tel. 443-287-4804
Fax. 410-955-9677
E-mail: xuhang@jhmi.edu

Heng Zhu, Ph.D.
BRB 333, 733 Broadway
Dept of Pharmacology & HiT Center
Johns Hopkins University School of Medicine
Baltimore, MD 21205
Tel. 410-502-0878
Fax. 410-502-1872
E-mail: Heng.Zhu@jhmi.edu
ABSTRACT

Specific anti-microbial 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 E. coli proteome microarray as a novel high-throughput proteomic 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’s 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, SAM (significant analysis of microarray) analysis 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 are highly immunogenic in CD, only 19 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±4%; p<0.01), and CD from UC (accuracy: 80 ±2%; p<0.01), respectively. The Set 1 antibodies recognized three pairs of E. coli proteins: era vs ybaN, yhgN vs focA, and gabT vs ycdG and the Set 2 antibodies recognized yidX vs frvX. The specificity and sensitivity of Set 1 antibodies were 81±5% and 89±3%, respectively, while those of set 2 antibodies were 84±1% and 70±6%, respectively. Serum antibodies identified for distinguishing healthy controls vs UC were only marginal, since their accuracy, specificity and sensitivity were 66±5%, 69±5%, and 61±7%, respectively (p<0.04). Taken together, we have identified novel sets of serological biomarkers for diagnosis of CD vs healthy control and CD vs UC.
INTRODUCTION

Crohn’s disease (CD) 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 the clinics of IBD (5-7) see 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 oligomanna (anti-\textit{Saccharomyces cerevisiae}, ASCA), bacterial outer membrane porin C (OmpC), \textit{Pseudomonas fluorescens} bacterial sequence I2 (anti-I2), and most recently bacterial flagellin (CBir 1) (5-7, 13) All of these anti-microbial antibodies show preponderance in patients with CD. However, ASCA has been identified in up to 5% of patients with UC (13, 14).

In comparison, \textbf{perinuclear antineutrophil cytoplasm 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 auto-antibody 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 (ACCA), anti-laminaribioside IgG (ALCA), anti-manobioside IgG (AMCA), and antibodies against chemically synthesized (Σ) two major oligomannose epitopes, Man α-1,3 Man α-1,2 Man (ΣMan3) and Man α-1,3 Man α-1,2 Man α-1,2 Man (ΣMan4) (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.

Though 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 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)

Proteomic technologies such as 2-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 to control. Blood-based (serum or plasma-based) proteomics hold 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 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 E 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 hypothesize that novel IBD-specific anti-
microbial antibodies, particularly anti-\textit{E. coli} antibodies, are present in IBD patients and are likely be identified by screening the sera with \textit{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 Hopkins Hospital Institutional Review Board. Sixty six patients had the diagnosis of Crohn’s disease (CD), 29 patients were diagnosed with ulcerative colitis (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 1. Clinical information was abstracted 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 individual undergone colon cancer screening or other non-IBD GI 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 minutes of collection. Serum was removed, aliquoted, and stored in multiple at -80\(^\circ\)C until assayed. All serum samples used in our study were selected from pre-screening 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 Igs) were used in our study.
Fabrication of *E. coli* proteome chips. To facilitate the analysis of protein function in the bacterial proteomes, we have 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 Dr. Mori and colleagues (26). *E. coli* cells first were grown overnight at 37 °C in 2x LB media containing 30 µg/ml chloramphenicol in a 96-well format and allowed to grow for overnight. The overnight cultures were diluted to a final OD$_{600}$ of ~0.1. After the cells were grown for ~3 hrs at 37 °C, and protein expression were induced with 1 mM isopropyl β-D-thiogalactoside (IPTG) for ~3.5 hrs. The liquid cultures were then harvested by centrifuge of 3500 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 re-suspended in phosphate lysis buffer, containing 300 mM NaCl, 20 mM imidazole, CelLytic B, Lysozyme (1 mg/mL), Benzonase (50 units/ml), proteinase inhibitor cocktail, and PMSF (1 mM). Along with Ni-NTA beads, the mixtures were incubated for 1.5 h at 4°C. After mixing, the resin-protein complexes were washed 3 times with Wash buffer I (50 mM NaH$_2$PO$_4$ with 300 mM NaCl, 10% glycerol, 20 mM imidazole, 0.01% Triton X-100, at pH 8) and 3 times with Wash buffer II (50 mM NaH$_2$PO$_4$ 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$_2$PO$_4$/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 are always included on each chip as negative controls. To monitor success of the serum profiling reactions on the chips, as well as the quality and consistency of the serum samples (such as similar concentration of immunoglobulins) and protein chips, we included two positive controls on each chip as a
positive control: 1), Ebna2, an Epstein-Barr virus (EBV)-encoded antigen that is reactive in common serological assays for essentially every human being; and 2), YLR-286C (endochitiase), a yeast protein which 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 (contains 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 (in intensity) recognized 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 one hour. The patient’s serum was diluted (1:1000) 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 hour incubation with gentle shaking on a rocker, the chip was rinsed once with 4 ml of Tris-buffered saline (TBS) with 0.05% Tween 20 (TBS-T). The chip was then soaked in 4 ml 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, G, and M (Jackson ImmunoLab) diluted at 1:400 in 3 mL 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 2000 rpm until dry prior to scanning. The chips were scanned with a GenePix array scanner (GenePix Pro 6.0 or GenePix 4200AL, Molecular Devices, PA) at wavelength of 536 nm. To achieve the best signal-to-noise ratio, many washing conditions with different stringencies had been tested, including 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 conditions described above. The washing condition described here gave best results among all conditions tested.

**Protein Array Data preprocessing.** Each quantified sample array image was exported from Genepix (Molecular Devices, CA) 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 for 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 analysis and only called a protein as significant with at least 1.5 fold change differences between two phenotypes at 0% False Discovery Rate in 500 permutations.

**Supervised learning algorithms.** To construct the classifier in this study, we employed three supervised learning methods. The algorithms implemented were k-Nearest Neighbors (kNN) (28), Support Vector Machines (SVM), and the k-Top Scoring Pairs Algorithm (k-TSP) (29). The k-TSP was implemented using a publicly available executable program (www.ccbm.jhu.edu/aboutus/news-ktsp.php) developed at the Institute for Computational
Medicine of Johns Hopkins University (29). SVM and kNN were implemented using the R statistical programming language (www.r-project.org), packages: e1071 and class for SVM and kNN, respectively.

**Feature Selection.** For kNN and SVM learning methods, SAM was applied to the training set for feature selection before the classifiers were trained on that data. The features selected in SAM were those that were found to be significant with a false discovery rate of zero. 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 kNN 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 executables for the learning algorithms.

**Cross-validation.** Ten-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 ten 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 (www.r-project.org/) to perform the statistical analyses in this study. P-value < 0.05 was regarded as significant.
RESULTS

Overall strategy of identifying IBD serological markers from E. coli proteome chips. Sera collected from 134 individuals (29 HC, 66 CD and 39 UC) from the Johns Hopkins Medical Institutes in this study (Table 1). 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 schematic illustration of our strategy in (Fig. 1). Since each protein was spotted in duplicate on the chip, we first analyzed the reproducibility of duplicate of each protein. As shown in scatter plot, the duplicate spots of each protein are highly correlated, indicating the good quality of the array manufacturing (Supplement Fig. S1; also see examples of 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 vs UC vs HC can be distinguished by comparing the signal intensities between protein spots on the E. coli proteome chips (see visual appearance of two representative chips probed with sera from CD and HC, respectively; Fig 2). Two-level of data analyses were performed with these immunogenic profiles (i) to identify differentially immunogenic responses among CD vs UC vs HC using Significance Analysis of Microarray (SAM) and Gene Ontology (GO) enrichment analysis; and (ii) to construct robust classifiers to distinguish CD vs UC vs HC using k-TSP method.

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 1) 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 vs CD vs UC, we employed SAM as described in the methods 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 manuscript. Heat maps shown in Fig. 3A-C present a visual illustration of the differentially immunogenic proteins for each phenotype. 273 differentially immunogenic proteins were identified by SAM when compared HC with CD samples. 81 proteins are highly immunogenic in CD samples and 192 are highly immunogenic in HC samples (Fig. 3A). Conversely, 188 proteins have different immunogenic responses in the IBD subtypes, 51 and 137 are highly immunogenic in UC and CD samples, respectively (Fig. 3B). When HC and UC samples are compared, only 27 and 6 proteins are discriminatory and highly immunogenic in HC and UC samples, respectively (Fig 3C). A full list of the immunogenic E. coli proteins in Fig. 3A-C can be found in Supplementary Tables S1-3, 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 are highly immunogenic in CD and only 19 in UC. 44 proteins were highly immunogenic in both HC and IBD (CD or UC). Among these 44 proteins, six overlap between HC and CD and 38 overlap between HC and UC. A full list of the immunogenic E. coli proteins in Fig. 3D can be found in Supplementary Table S4. This demonstrates that UC and HC subjects share more common immunogenic profiles than CD and HC. In general, our results indicate that much 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 (www.geneontology.org). Functional grouping of the 417 proteins were assigned by querying EcoCyc (http://ecocyc.org/) and KEGG (www.genome.ad.jp/kegg) databases, as well as cross-checked with Affymetrix E.
coli Genome Array annotation file (www.affymetrix.com). 338 of these 417 proteins were assigned to at least one gene ontology (GO) term, and 78 hypothetical proteins have unknown annotations. We focused our enrichment analysis on five 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 is 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 these proteins that are immunogenic in HC and CD or UC. Antibodies against membrane proteins are highly enriched in HC samples (p < 0.0001). Interestingly, antibodies against intracellular and macromolecular complex proteins are highly enriched in CD samples (p < 0.05), while those against cell wall proteins are highly enriched in UC samples (p < 0.05). Although 12% proteins that were found to be highly immunogenic in CD samples were located in periplasmic space, their enrichment was not statistically significant (p = 0.064) for this IBD subtype. Proteins located in cell projection term are not enriched in either healthy controls or IBD subtypes.

**Machine Learning Analysis.** Next, we seek to construct optimal classifiers from the immunogenic response profiles to differentiate HC from the IBD subtypes (CD and UC), as well as differentiate CD from UC. Upon successful construction of these classifiers, the classification rules may result in the discovery of new robust biomarkers. Here, we employed k-TSP, a novel machine learning method, to discover 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 > ybaN then CD or else HC, if yhgN is > focA then CD or else HC, and if gabT is > ycdG then CD or else HC (see representative examples of actual images of immuno-reactive 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 out of 39 HC and 62 out of 64 CD samples are correctly classified, with an estimated ten-fold cross-validation accuracy of 86±4% (p < 0.01). For distinguishing HC from UC samples, the k-TSP algorithm identifies eleven feature pairs (Fig. 5B) with an estimated ten-fold cross-validation accuracy of 66±5% (p < 0.04). A single feature pair of k-TSP classifier was identified for differentiating CD from UC: If the immunogenic reactivity to frvX is > yidX then UC or else CD, as illustrated in Fig. 5C (see representative examples of actual images of immuno-reactive protein spots in Fig. S2). This classifier has an estimated ten-fold cross-validation accuracy of 80±2% (p < 0.1).

We also compared the performance of k-TSP with SVM and kNN, two other commonly used learning algorithms, for each of the classification problems based on five runs of ten-fold cross-validation. Table 2 displays the results of ten-fold cross-validation for each of the three classifiers. As demonstrated in Table 2, based on cross-validation, k-TSP performance meets or exceeds the performance of kNN and SVM for these classification problems. Because the cross-validation structure allowed each classifier to test on the same subsets of data as described in the methods section, the performance of the three classifiers can be directly compared and tested for statistical significance by a simple student's t-test. The HC vs. 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 in classification accuracy when compared to SVM and kNN classifiers.
From this study, we found that k-TSP performs much better than SVM and kNN 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 pre-processing (29). Fig. S3-A&B (supplement data) shows that on their own, the immunogenic response to era and ybaN (the top scoring pair in the HC vs CD k-TSP classifier) do 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) results in clear separation in the data lending well to classification (Fig. S3-C). Similar results are true when scatter plot analysis were done for the other two TSP pairs from the HC vs CD classifier (yhgN vs focA and gabT vs 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 explains why KNN and SVM fail to match the performance of k-TSP as the relative feature levels within samples appear to be much more robust then 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 class with greater total number of samples in order to equalize the class sizes. 10-fold cross validation was performed as described. The process was then repeated by discarding a different random set of samples. Table 3 shows the performance of each classifier given class balance in the training set. It demonstrates that k-TSP outperforms SVM and kNN in most instances whether or not the class size is balanced, further supporting the data presented in Table 2.
Next, to determine the significance of each classifier, permutation test was performed by randomly shuffled the class labels while maintain the same number of samples in each class. 10-fold cross validation is carried out to yield a classification rate for the permutation set. 100 permutations were performed in order to get a null distribution of expected classification rates by chance. The classification rate from the un-permuted data is then compared to the null distribution to determine significance. **Table 3** 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 are estimated to be significant at the \( p < 0.01 \) level. The k-TSP classifier built to differentiate HC and UC had 4/100 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 ten-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 that we performed ten fold cross-validation in **Table 3**, there are 50 classifiers to inspect (10 for each of the 5 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 rule. **Table 3** shows that the pairs that show up in the HC vs. CD classifier as well as the UC vs. CD classifier are fairly robust while the pairs in the HC vs. UC classifier are not. Along with the permutation testing, this indicates that the HC vs. CD and UC vs. CD classifier should perform well in independent testing while the HC vs. 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 anastamosis. 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’s 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 1**): penetrating/fistulizing, stricturing, penetrating/structuring 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 due to the smaller sample size of each disease type (data not shown; see sample sizes of CD or UC in **Table 1**).

**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 (**Fig. S4**) analysis of *E. coli* ompC and fliC demonstrates that neither allows for class separation of between control vs CD vs 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 immunological diseases using a protein chip of whole prokaryotic proteome.
The significance of this study is three-fold: First, it presents here the first proof of principle for the feasibility of application of high-density protein microarray/chip technology in the discovery of novel serological IBD biomarkers. This study can serve as an example of similar proteomic approaches for hunting serological biomarkers for other immune-related diseases, such as auto-immune disorders. Secondly, 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 \textit{E. coli} proteins (Fig. 3D). Since it has been demonstrated that defective intestinal barrier function plays a central role in the pathogenesis 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 \textit{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 are specific in healthy controls while only 18 immunogenic proteins are found to be specific to UC. It’s worth noting that in a previously report the development of colitis in T cell receptor alpha knockout mice (a UC-like murine IBD model) was associated with restricted humoral responses to selected \textit{E. coli} proteins (37). This report supports our finding that patients with UC have decreased immune responses to intestinal bacteria. Thirdly, we have identified a set of novel serological biomarkers that have >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 vs intracellular proteins in HC vs CD patients. Approximately 85% of the highly immunogenic proteins are either cell wall proteins or membrane proteins in HC, compared to only ~37% of the top immunogenic proteins in CD patients (Fig. 4; Supplement Tables 1-3). Conversely, ~30% of top immunogenic proteins in CD patients are intracellular
proteins compared to only ~7% in HC (Fig. 4; Supplement Tables 1-3). Furthermore, there is 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 of having 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 due to intestinal epithelial barrier, surface and membrane proteins of *E. coli* might be the primary antigens that are more accessible to the immune system, compared to intracellular proteins. In this case, immune system has adapted to the presence of luminal *E. coli*. In contrast, in CD patients, disrupted or compromised intestinal barrier (35, 36) may lead to the bacterium or its products across the gut luminal barrier. If the whole *E. coli* invades into the lamina propria, it will mostly likely be lysed by host immune system. Subsequently, *E. coli* components such as intracellular proteins that otherwise not 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 used 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 are only 6 overlapping proteins among 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 are identified here for discriminating CD from HC or UC have been previously described. 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 involves in the binding of GTP and nucleotide of cell cycle and can be found in intracellular membrane. In this study, an increased immunogenic response to era is associated with CD, identified by both SAM and k-TSP analyses. YbaN is predicted as a conserved inner membrane protein with unknown function. YhgN is predicted as an inner six transmembrane domains protein where the C-terminus is located in the periplasm (38). YcdG (also called rutG) is another predicted transmembrane with eleven 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 the 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 involve 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 (GABA) 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 a important protein in fructose-specific PEP-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 p-ANCA, ASCA, anti-OmpC, and 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 8 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’s 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: can we use the information to develop IBD-specific vaccines in the future?

**Implication of the novel IBD serological markers.** The newly identified biomarkers by k-TSP analysis have a particular impressive ~86% accuracy in differentiating CD from HC, with a specificity of ~81% and a sensitivity of ~89% (Table 2). In addition, k-TSP analysis yields an accuracy of ~80% in differentiating CD and UC, with a sensitivity of ~84% and specificity of ~70% (Table 2). These demonstrate that the sensitivity and specificity of these novel serological markers are comparable to those of combination of the multiple best-characterized IBD biomarkers (ASCA, pANCA, anti-OmpC, and anti-Cbir) (43, 44). More importantly, an identical performance can be achieved by using only the top 3 pairs of *E. coli* proteins for discriminating healthy controls vs CD, and one top pair of proteins for differentiating CD vs UC (Figs. 2 &5 and Tables 2 &3). These data provide a critical feasibility for 1) 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 *E. coli* protein array is not suitable for identifying serological biomarkers in differentiating UC from HC (only ~66% accuracy) (Tables 2 &3). Importantly, OmpC, an *E. coli* antigen for one of the widely studied current serological biomarker (anti-OmpC), was not picked up in our screen (Fig. S4A). Similarly, fliC, an *E. coli* flagellin protein equivalent the *Salmonella* flagellin (which is the antigen for anti-Cbir, another widely studied anti-bacterial antibody) did not show up in our analysis (Fig. S4A). These data would 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 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.
ACKNOWLEDGEMENTS

This work was supported by NIH-NIDDK grant (5R21DK77064), Broad Medical Research Program (IBD-0119R), and NIH/NIDDK KO1 (DK62264). Special thanks go to Mr. & Mrs. Morton Hyatt for their generous support for the on-going IBD research in Dr. Xuhang Li’s laboratory.
Reference List


markers in inflammatory bowel disease are associated with complicated disease behaviour, 

*Gut* 56, 1394-1403.


Table 1. Demographic and clinical information of IBD Patients and healthy controls.

<table>
<thead>
<tr>
<th>Patient Data/Characteristics</th>
<th>CD (n = 66)</th>
<th>UC (n = 29)</th>
<th>HC (n = 39)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender: female %</td>
<td>55</td>
<td>53</td>
<td>43</td>
</tr>
<tr>
<td>Age: mean/st dev (yrs)</td>
<td>36.7 ± 13.1</td>
<td>38 ± 14.5</td>
<td>47 ± 12.4</td>
</tr>
<tr>
<td>Age at diagnosis: mean (yrs)</td>
<td>36.7</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Duration of disease (yrs)</td>
<td>12.3</td>
<td>8.9</td>
<td></td>
</tr>
<tr>
<td>Extraintestinal disease: n (%)</td>
<td>7 (11)</td>
<td>3 (10)</td>
<td></td>
</tr>
<tr>
<td>Surgery: n (%)</td>
<td>46 (70)</td>
<td>4 (12)</td>
<td></td>
</tr>
<tr>
<td>Ethnicity: n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>29 (44)</td>
<td>10 (34)</td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>35 (53)</td>
<td>22 (75)</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>2 (3)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Smoking: n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Past or present</td>
<td>17 (26)</td>
<td>5 (17)</td>
<td></td>
</tr>
<tr>
<td>Nonsmoker</td>
<td>40 (60)</td>
<td>18 (62)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>9 (14)</td>
<td>9 (35)</td>
<td></td>
</tr>
<tr>
<td>Medications: n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antibiotics</td>
<td>23 (35)</td>
<td>4 (14)</td>
<td></td>
</tr>
<tr>
<td>5-ASA</td>
<td>49 (74)</td>
<td>26 (90)</td>
<td></td>
</tr>
<tr>
<td>Corticosteroids</td>
<td>16 (24)</td>
<td>9 (35)</td>
<td></td>
</tr>
<tr>
<td>AZA/6-MP</td>
<td>26 (39)</td>
<td>17 (59)</td>
<td></td>
</tr>
<tr>
<td>Methotrexate</td>
<td>0</td>
<td>1 (3)</td>
<td></td>
</tr>
<tr>
<td>Infliximab</td>
<td>15 (23)</td>
<td>2 (7)</td>
<td></td>
</tr>
<tr>
<td>Crohn's Disease Subtype: n (total)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonstricturing and nonpenetrating</td>
<td>18 (27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Penetrating</td>
<td>26 (40)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stricturing</td>
<td>14 (21)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Penetrating and stricturing</td>
<td>8 (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ulcerative Colitis: n (total)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left Sided Colitis</td>
<td></td>
<td>13 (45)</td>
<td></td>
</tr>
<tr>
<td>Pancolitis</td>
<td></td>
<td>18 (62)</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Estimated ten-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 ten-fold cross validation ± the standard deviation. In parenthesis are the numbers of samples in each subtype used for classification. Sp = specificity, Sn = sensitivity, PPV= positive predictive value, NPV= negative predictive value.

<table>
<thead>
<tr>
<th>Healthy control (HC) (39) vs CD (66)</th>
<th>Method</th>
<th>Accuracy (%)</th>
<th>Sp (HC) (%)</th>
<th>Sn (CD) (%)</th>
<th>PPV (CD) (%)</th>
<th>NPV (HC) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k-TSP</td>
<td>86 ± 4</td>
<td>81 ± 5</td>
<td>89 ± 3</td>
<td>89</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>70 ± 2</td>
<td>66 ± 1</td>
<td>73 ± 2</td>
<td>79</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>kNN</td>
<td>63 ± 3</td>
<td>47 ± 7</td>
<td>73 ± 6</td>
<td>70</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Healthy control (HC) (39) vs UC (29)</th>
<th>Method</th>
<th>Accuracy (%)</th>
<th>Sp (HC) (%)</th>
<th>Sn (UC) (%)</th>
<th>PPV (UC) (%)</th>
<th>NPV (HC) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k-TSP</td>
<td>66 ± 5</td>
<td>69 ± 5</td>
<td>61 ± 7</td>
<td>59</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>62 ± 5</td>
<td>58 ± 1</td>
<td>68 ± 12</td>
<td>55</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>kNN</td>
<td>60 ± 6</td>
<td>57 ± 2</td>
<td>64 ± 12</td>
<td>53</td>
<td>68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CD (66) vs UC (29)</th>
<th>Method</th>
<th>Accuracy (%)</th>
<th>Sp (CD) (%)</th>
<th>Sn (UC) (%)</th>
<th>PPV (CD) (%)</th>
<th>PPV (UC) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k-TSP</td>
<td>80 ± 2</td>
<td>84 ± 1</td>
<td>70 ± 6</td>
<td>86</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>78 ± 3</td>
<td>82 ± 2</td>
<td>69 ± 9</td>
<td>86</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>kNN</td>
<td>78 ± 3</td>
<td>78 ± 4</td>
<td>61 ± 2</td>
<td>82</td>
<td>55</td>
</tr>
</tbody>
</table>
Table 3. **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 ± standard deviation, p-value).

<table>
<thead>
<tr>
<th>k-TSP Classifier</th>
<th>Accuracy (%)</th>
<th>Estimated p-value</th>
<th>Features in k-TSP classifier</th>
<th>% feature appearance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HC vs. CD</strong></td>
<td>86 ± 3</td>
<td>50 ± 8</td>
<td>era &gt; ybaN = CD</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>yhgN &gt; focA = CD</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>gabT &gt; ycdG = CD</td>
<td>64</td>
</tr>
<tr>
<td><strong>HC vs. UC</strong></td>
<td>66 ± 5</td>
<td>51 ± 9</td>
<td>relE &gt; cysE/wcaB = UC</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>pyrI &gt; yjgK = UC</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>int &gt; ybiO = UC</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ftsE &gt; pssR = UC</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>yhgN &gt; yhfG = UC</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>yafN &gt; dsbB = UC</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>yihl &gt; yabK = UC</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>421#15 &gt; yhdN = UC</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>hisP &gt; rplO = UC</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cml &gt; nuoM = UC</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>yieC &gt; nuol = UC</td>
<td>12</td>
</tr>
<tr>
<td><strong>UC vs. CD</strong></td>
<td>80 ± 2</td>
<td>52 ± 6</td>
<td>yidX &gt; frvX = UC</td>
<td>88</td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

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 high-throughput protein purification protocol and printed onto FullMoon slides using a ChipWriter Pro robot. 134 patient sera were collected from the Johns Hopkins Hospital for this analysis. These sera were screen by *E. coli* proteome chips. Two-level of data analyses were performed: (i) global IBD analysis was performed to identify differentially immunogenic proteins in HC, CD and UC using Significance Analysis of Microarray (SAM) and Gene Ontology (GO) enrichment analysis; and (ii) serological IBD biomarkers discovery using k-TSP algorithm.

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 Crohn’s Disease (CD) patient (left panel) and a healthy control (HC) (right panel). 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. Middle panel shows some representative images of immunogenic spots of three pairs of specific proteins (see more information of these proteins in Fig. 5 and Tables 1-3) from these proteome chips. Every *E. coli* protein is spotted in duplicate on the chip. CD vs UC vs HC can be distinguished by comparing the signal intensities between protein spots on the *E. coli* proteome chips.
Fig. 3. Global immunogenic profiles of IBD patients’ sera against *E. coli* proteins. A. Heatmap of 273 differentially immunogenic proteins between Healthy controls (HC) and CD samples identified by SAM analysis. Yellow and blue colors indicate high and low immunogenic response, respectively. B. Heatmap of the 188 differentially immunogenic proteins between CD and UC samples identified by SAM analysis; and C. 33 differentially immunogenic proteins between Healthy controls and UC samples as identified by SAM analysis. 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 standard deviation is 1. Blue and yellow indicates high and low immunogenic proteins, respectively. D. Venn diagram of these differentially immunogenic proteins showing only limited overlapping among HC vs CD vs UC.

Fig. 4. Distribution of the Cellular Component terms in the highly immunogenic response proteins of healthy controls (HC), CD and UC. Six Cellular Component terms from the Gene Ontology were examined. Cell projection term contains flagellum and fimbrium proteins. The main messages include: 1) approximately 80% of the highly immunogenic proteins are either membrane proteins in HC (p < 0.0001), compared to only ~37% of the top immunogenic proteins in CD patients (not statistically significant); 2) conversely, ~30% of top immunogenic proteins in CD patients are intracellular proteins (p < 0.05) compared to only ~7% in HC (not statistically significant); 3) a significant higher percentage of cell wall proteins (~26%) are immunogenic in UC (p < 0.05) compared to those in HC and CD (not significant); and 4) a significant percentage of macromolecular complex proteins (~16%; p < 0.05) in CD compared to those in HC or UC (not statistically significant). No statistically significant enrichment of proteins of periplasmic space and cell projection were found in HC, CD and UC.
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 ratio of the immunogenic reactivity of a top scoring pair of proteins. The expression values represented are the ratio 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 $Y$ being example proteins). If the immunogenic reactivity of a patient to protein $X$ was greater than the reactivity to protein $Y$, the box will appear yellow, and blue for vice versa (see examples below). 

A. depicts the classifier for HC vs CD (yellow = CD, blue = HC). For example, if era is $>$ ybaN, it will be a CD and shows as yellow, or else a HC (blue). 

B. displays HC vs. UC classifier (yellow = UC, blue = HC). For example, if relE is $>$ cysE_wcaB, it is a UC and shows as yellow, or else it is classified as a HC (blue). 

C. shows the CD vs. UC classifier (yellow = UC, blue = CD). If frvX is $\geq$ yidX, it is a UC (yellow), or else a CD (blue). See representative images of some of those protein pairs in Fig. 2 & Fig. S2.
Patient Sera (N = 39, CD = 66, UC = 29)

Screen E. Coli Chips with sera

Print E. coli Whole Proteome Chip

High-Throughput Protein Purification Protocol

Clone & Express E. coli proteome (> 4,000 proteins)

Fabrication of E. Coli Whole Proteome Chip

Global Analysis of Immuno-reactivity to E. coli

Identify Differentially Immunogenic E. coli proteins in HC, CD & UC

SAM Analysis

GO Analysis

Serological IBD Biomarkers Discovery

Machine Learning (k-TSP)

Training Steps:
• Ten-fold Cross-validation
• Permutation Tests

Novel IBD Biomarkers

Fig. 1
Crohn’s Disease (CD) (whole E. coli chip)  Healthy Controls (HC) (whole E. coli chip)

Representative immunogenic spots

<table>
<thead>
<tr>
<th></th>
<th>CD</th>
<th>HC</th>
</tr>
</thead>
<tbody>
<tr>
<td>era</td>
<td><img src="era_chip.png" alt="Green Spot Image" /></td>
<td><img src="era_chip.png" alt="Green Spot Image" /></td>
</tr>
<tr>
<td>ybaN</td>
<td><img src="ybaN_chip.png" alt="Green Spot Image" /></td>
<td><img src="ybaN_chip.png" alt="Green Spot Image" /></td>
</tr>
<tr>
<td>yhgN</td>
<td><img src="yhgN_chip.png" alt="Green Spot Image" /></td>
<td><img src="yhgN_chip.png" alt="Green Spot Image" /></td>
</tr>
<tr>
<td>focA</td>
<td><img src="focA_chip.png" alt="Green Spot Image" /></td>
<td><img src="focA_chip.png" alt="Green Spot Image" /></td>
</tr>
<tr>
<td>gabT</td>
<td><img src="gabT_chip.png" alt="Green Spot Image" /></td>
<td><img src="gabT_chip.png" alt="Green Spot Image" /></td>
</tr>
<tr>
<td>ycdG</td>
<td><img src="ycdG_chip.png" alt="Green Spot Image" /></td>
<td><img src="ycdG_chip.png" alt="Green Spot Image" /></td>
</tr>
</tbody>
</table>
Fig. 3

A. 182 highly immunogenic proteins in Healthy Controls

B. 61 highly immunogenic proteins in CD samples

C. 61 highly immunogenic proteins in UC samples

D. Venn diagram showing overlap of immunogenic proteins in Healthy Controls (HC), CD, and UC samples.

HC (n = 39)  CD (n = 86)  UC (n = 29)

169  6  0

185  1  18

38
**Fig. 4**

![Graph showing GO Cellular Component Terms]

- **Membrane**: **HC**, **CD**, **UC**
- **Cell Wall**: **HC**, **CD**, **UC**
- **Macromolecular Complex**: **HC**, **CD**, **UC**
- **Intracellular**: **HC**, **CD**, **UC**
- **Periplasmic space**: **HC**, **CD**, **UC**
- **Cell Projection**: **HC**, **CD**, **UC**

**Key**:
- **** p-value < 0.0001
- * p-value < 0.05

% Proteins