Identifying Biologically Relevant Mechanisms And Biomarkers Using Novel Bioinformatics Methods

Samer Hanoudi
Wayne State University

Follow this and additional works at: https://digitalcommons.wayne.edu/oa_dissertations

Part of the Bioinformatics Commons

Recommended Citation
https://digitalcommons.wayne.edu/oa_dissertations/3570

This Open Access Embargo is brought to you for free and open access by DigitalCommons@WayneState. It has been accepted for inclusion in Wayne State University Dissertations by an authorized administrator of DigitalCommons@WayneState.
IDENTIFYING BIOLOGICALLY RELEVANT MECHANISMS AND BIOMARKERS USING NOVEL BIOINFORMATICS METHODS

by

SAMER HANOUDI

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2022

MAJOR: COMPUTER SCIENCE

Approved By:

__________________________________________
Advisor Date

__________________________________________

__________________________________________

__________________________________________

__________________________________________
DEDICATION

This work is dedicated to my immediate family members. They always motivated me to finish my Ph.D.
ACKNOWLEDGEMENTS

I want to thank my advisor, Dr. Sorin Draghici for his valuable feedback and support. I also want to thank my dissertation committee members, Dr. Loren Schwiebert, Dr. Adi Tarca, and Dr. Suzan Arslanturk, for their assistance and availability to work with me. Finally, I want to thank my current and previous colleagues at the Intelligent Systems and Bioinformatics Laboratory.
# TABLE OF CONTENTS

Dedication ................................................................. ii
Acknowledgements ....................................................... iii
List of Tables ............................................................. vii
List of Figures ............................................................ ix
Chapter 1 Introduction .................................................. 1

Chapter 2 Novel T7 Phage Display Library Detects Classifiers for Sarcoidosis, Active Mycobacterium Tuberculosis Infection and cystic fibrosis .......... 7

2.1 Introduction ............................................................ 7

2.2 Materials and Methods ............................................. 12

2.2.1 Chemicals .......................................................... 12

2.2.2 Patient Selection .................................................. 12

2.2.3 Serum Collection ................................................ 13

2.2.4 Construction and Biopanning of T7 Phage Display cDNA Libraries 13

2.2.5 Microarray Construction and Immunoscreening ............... 13

2.2.6 Sequencing of Phage cDNA Clones .......................... 13

2.2.7 Data Acquisition and Pre-Processing ........................ 14

2.2.8 Statistical Analyses Sarcoidosis ............................... 15

2.2.9 Statistical Analyses Tuberculosis ............................. 16

2.2.10 Statistical Analyses Cystic Fibrosis .......................... 16

2.3 Results Sarcoidosis .................................................. 18

2.3.1 Complex Sarcoidosis (CSL) library detects unique antigens in the sera of Sarcoidosis .................................................. 18

2.4 Results Tuberculosis ................................................ 26
4.3.2 Mean value across all SNPs sets in \( r \) ........................................ 89

4.3.3 Comparisons for USA samples only ........................................... 92

4.4 Conclusion ..................................................................................... 93

Chapter 5 Platelets and renal failure in the SARS-CoV-2 ..................... 95

5.1 Introduction ................................................................................... 95

5.2 Methods ....................................................................................... 97

5.2.1 Study Design and Data Collection ............................................ 97

5.2.2 Specimen Collection and Testing .............................................. 97

5.2.3 Laboratory Data ....................................................................... 97

5.2.4 Statistical Analysis .................................................................. 98

5.3 Results .......................................................................................... 98

5.4 Discussion ..................................................................................... 102

Chapter 6 Conclusion ......................................................................... 104

References ......................................................................................... 106

Abstract .......................................................................................... 137

Autobiographical Statement ............................................................. 140
LIST OF TABLES

Table 1  Subjects demographics. ............................................. 20
Table 2  Subjects demographics. ............................................. 28
Table 3  10 Top Significant TB Clones. ................................. 35
Table 4  Subjects demographics. Age values are presented as means and variability in SD. \( N \) = Number of patients and percent are shown in parentheses. ............................................. 37
Table 5  14 Top Significant CF Clones. ................................. 43
Table 6  A comparison between the results of HighEdgeS, the classical approach, and GAGE as they would be used in practice. Three pathway analysis methods (ORA, SPIA, and LEGO) were applied on the genes yielded by HighEdgeS and the classical approach. GAGE uses the entire set of genes to directly identify pathways so ORA, SPIA and LEGO are not applicable for GAGE. The green background indicates the best results obtained for each dataset (each row). For the classical approach we selected the DE genes using an absolute fold change greater than 2 and FDR-corrected p-value less than 0.05. “No DE” means no genes met those thresholds. The performance was measured using the true positive rate (TPR) and false positive rate (FPR) for each data set. The results show that HighEdgeS yields the best TPR for all datasets. HighEdgeS also yield the best FPR for 2 out of the 3 datasets. Furthermore, for the Pdx1 dataset, even though the classical approach has a lower FPR values, its sensitivity is very low (not being able to identify any true positive (TP) pathway in 2 out of 3 cases, and identifying only one of the 3 true positive pathway in the remaining case). In contrast, HighEdgeS identified all true positive pathways in 2 out of 3 cases (when combined with ORA and SPIA) and 2 out of the 3 TP pathways when combined with LEGO. ............................................. 75
Table 7  The samples counts of the training data by region and country. .......... 81
Table 8  The samples counts of the testing data by region and country. .......... 85
Table 9  The mean values of accuracy, F1 score, sensitivity, and specificity across the 5 testing ethnic groups of the proposed classifier, the naive Bayes classifier, the Random Forest classifier, and the BIASLESS classifier. For each row, the maximum values of the F1 score are shown in the cells with green color. ............................................. 86
Table 10 The mean values of accuracy, F1 score, sensitivity, and specificity across the 5 testing ethnic groups. The maximum value of each assessment measure (column) is shown in the cells with green color.

Table 11 The mean values of accuracy, F1 score, sensitivity, and specificity for each group of SNPs and for samples collected only from the USA but from individuals belonging to different ethnic groups (i.e. Africa, America, Central South Asia, East Asia, and Europe). For each row, the maximum values of the F1 score (columns) are shown in the green cells.

Table 12 The mean values of accuracy, F1 score, sensitivity, and specificity across all sizes of groups of SNPs. The data includes samples collected only from the USA but from individuals belonging to different ethnic groups (i.e. Africa, America, Central South Asia, East Asia, and Europe). The maximum value of each assessment measure (column) is shown in the green cells.

Table 13 Clinical characteristics for the patients
Figure 1  PCA and Hierarchal clustering (option 1). (a) PCA plot along PC1 and PC2 generated with 1070 clones of four groups: (1) healthy control samples; (2) Sarcoidosis samples; (3) TB samples; and (4) Lung cancer. Biomarker clusters along the PC1 explain a variance of only 14%, while the variance along PC2 was about 13%. (b) The hierarchal clustering was applied on the healthy controls, sarcoidosis, TB patients and lung cancer with 1070 clones. No distinct clustering is observed when all clones are used.

Figure 2  PCA and Hierarchal clustering (option 1). (a) PCA plot along the PC1 and PC2 results when applied on 132 sarcoidosis clones (option 1). The PC1 explained 0.33 of variance, whereas PC2 explained 13% of the variance. As shown, the sarcoidosis samples are well separated from the lung cancer, TB controls and most healthy control samples. (b) Hierarchal clustering using the top 132 sarcoidosis clones (FDR<0.05). (c) PCA plot generated with the top 14 sarcoidosis clones. The PC1 explained 45% of the variance, whereas PC2 explained 16% of the variance. (d) Hierarchal clustering using the top 14 sarcoidosis clones. This figure demonstrates a good clustering performance with the 132 sarcoidosis clones and the top 14 significant sarcoidosis clones (FDR<0.01).

Figure 3  PCA and Hierarchal clustering (option 2). (a) PCA plot along PC1 and PC2 generated with 221 clones (FDR 0.01) of the four groups: (1) healthy control samples; (2) Sarcoidosis samples; (3) TB samples; and (4) Lung cancer. The PC1 explains 32% of the variance, while the variance along PC2 was 12%. (b) The hierarchal clustering was applied on the healthy controls, sarcoidosis, TB patients and lung cancer with 221 clones (FDR <0.01). (c) PCA plot along the PC1 and PC2 results when applied on the top 12 sarcoidosis clones. The PC1 explained 54% of the variance, whereas PC2 explained 14% of the variance. As shown, the sarcoidosis samples are well separated from the lung cancer, TB and most healthy control samples. (d) Hierarchal clustering using the top 12 sarcoidosis clones. This figure demonstrates a robust class separation using the top 12 sarcoidosis classifier clones.

Figure 4  Diagrammatic representation of significant clones from two approaches (option 1 and 2). (a) Illustrates the Venn diagram of 132 clones (FDR<0.05) from option 1 and 221 clones (FDR< 0.01) from option 2 and their intersection. (b) depicts the Venn diagram of the 14 classifiers clones from option 1 and 12 clones from option 2.
Figure 5  Classification to predict sarcoidosis from healthy controls, TB patients and LC patients on the testing set. (a) Performance of 132 clones on the testing set (option 1). (b) Performance of the top 14 classifier clones on the test set (option 1). The ROC curves demonstrate excellent classification performance with AUC of 0.947 with a sensitivity of 0.883 and specificity of 0.923. (c) Performance of 221 clones at the FDR=0.01 (option 2) on the testing set. (d) Performance of the top 12 clones (option 2) on the test set. The ROC curves demonstrate strong classification performance with AUC of 0.926 with sensitivity of 0.962 and specificity of 0.837.

Figure 6  PCA and Hierarchal clustering. (a) PCA score plots along PC1 and PC2 were generated with 1070 clones of three groups: (1) healthy control samples (black circles); (2) TB samples (green squares) and; (3) Sarcoidosis samples (blue triangle). Biomarker clusters along the PCA1 explain variance of only 0.15, while the variance along PC2 was about 0.14. (b) The hierarchal clustering was applied on the healthy controls (black labels), TB patients (green labels) and sarcoidosis (blue labels) with 1070 clones. (c) PCA score plots along the PC1 and 2 results when applied on 192 TB clones. The PC1 explained 0.26 of variance, whereas PC2 explained 0.15 of variance. As shown, the TB samples are well separated from the healthy controls and sarcoidosis samples. (d) Hierarchal clustering using only the highly significant 192 TB clones. The blue and black clusters include sarcoidosis and healthy controls, the green cluster includes all the TB samples except one. (e) PCA score plots along PC1 and 2 generated with top 10 highly significant clones. The PC1 explained 0.54 of variance, whereas PC2 explained 0.11 of variance. (f) Hierarchal clustering using 10 top significant TB clones. This figure demonstrates better clustering with 192 TB clones and the highly significant 10 TB clones (panels c, d, e, and f) when compared the clustering using all clones (panels a and b).

Figure 7  Heatmaps were generated based on 192 clones and the 10 highly significant clones from the data of 176 study subjects (45 healthy controls, 24 with TB, and 107 with sarcoidosis) (a,b). Each row represents a clone, while each column represents a study subject. As shown in Figure 2, most of TB samples clustered to the right side of heatmap plots, while sarcoidosis samples and healthy controls clustered on the left side of the plot, indicating different expression profiles.
Figure 8  Classification results to predict TB from healthy controls and sarcoidosis patients. (a) Performance of 192 clones on the test set. (b) Performance of the top 10 classifier clones on the test set. The ROC curves demonstrate excellent classification performance with AUC of 1 with sensitivity of 1 and specificity of 1. .................................................. 31

Figure 9  ROCs for the top 6 significant clones that are increased in TB sera compared to healthy controls and sarcoidosis. .................................................. 33

Figure 10  ROC for the top 4 significant clones that are decreased in TB sera compared to healthy controls and sarcoidosis. This figure demonstrates reasonable classification performance when the classification was applied on one clone. .................................................. 34

Figure 11  PCA and hierarchal clustering. (a) PCA score plots along the PC1 and 2 were generated with 1070 clones of three groups: 1) healthy control samples (yellow circle), 2) CF samples (blue triangle) and 3) LC samples (green square). Along the PCA1 explaining a variance of only 0.18 and along the PC2 of 0.12. (b) The hierarchal clustering was applied on the healthy controls (black labels), CF patients (red labels) and LC (blue labels) with 1070 clones. (c) PCA score plots along the PC1 and 2 results when applied on the highly significant 20 CF clones. The PC1 explained 0.49 of variance, whereas PC2 explained 0.09 of variance. As shown the CF samples are well separated from the healthy controls and LC samples. (d) Hierarchal clustering using only the highly significant 20 CF clones. The green cluster includes LC and healthy control samples (no CF samples), the magenta cluster includes all the CF samples, few healthy control and two LC samples. This figure demonstrates better clustering with the highly significant 20 CF clones (panels c and d) when compared with the clustering using all clones (panels a and b). .................................................. 38

Figure 12  Heatmap generated based on the 20 highly significant CF clones from the data of 111 study subjects (49 healthy controls, 31 with CF and 31 with LC). Each row represents a clone, while each column represents a study subject. As shown in this figure, most CF samples clustered to the left side of the heat map plot, while the LC samples and healthy controls clustered to the right side of the plot indicating different expression profiles. .................................................. 39
Figure 13  Shows the classification performance of the naive Bayes classifier. The classifier is to predict CF from LC and healthy control samples. (a) Performance of the classifier on the testing sets. Box plots indicate the AUC values (y-axis) when the classifier model was applied on the 1000 test sets. The x-axis is accumulating sets of clones. The accumulation of the clones starts with the most frequent clone and then one clone added at a time to reach 100 clones. (b) Performance of the classifier models on the validation set. As indicated the classifier models when they were built using the significant clones show a high AUC values on the testing sets as well as on the completely independent validating set. (c) The ROCs were generated from the average of the 1000 runs of the classifier models when applied on the validation set (randomly selected healthy controls, CF and LC) using the 20 highly significant CF clones. The ROC curve demonstrates an excellent classification performance with an average AUC of 0.973. These results indicate the excellent performance of the naive Bayes classifier on the 20 highly significant CF clones.

Figure 14  Shows the Pearson correlation of the identified biomarkers with clinical values. Scatter plots depicted correlation of the sweat chloride values with one clone (a) and aggregated 5 clones (b). Scatter plots depicted the correlation for BMI predicted with one clone (c) and aggregated 5 clones (d). Scatter plots depicted correlation of FEV1% with one clone (e) and aggregated 5 clones (f). The correlation values and p values are shown in the top right of each plot. The names of the clones are shown at the bottom of each plot.

Figure 15  The workflow of the proposed method, HighEdgeS. a) The global graph constructed from all interactions present in all KEGG pathways; b) The edge score histogram constructed from the input data. A change point analysis [86] is used to determine the beginning of the flat area of the curve. The selected edges will be those in the top 75% of the remaining scores. c) The global graph showing the edges with high scores in red; d) The subgraph with the high scoring edges only representing the putative mechanism involved in the given phenotype.
Figure 16  **Figure shows the results for the dataset Myd88.** The KO gene is shown in red in both panels. a) Mechanism results for HighEdgeS when we applied it on the Myd88 dataset. The results show two subgraphs. The one with the most genes shows the KO gene regulates the Cxcl1 gene and shows that the subgraph includes many Cxcl1 downstream genes. b) The results of the classical approach using DE genes. The Myd88 gene is connected to only one downstream gene. The result of the classical approach has many genes that are not connected indicating that the classical approach missed important interactions.

Figure 17  **Comparison between the results from HighEdgeS (left columns) and the classical approach (right columns) for the Myd88 dataset.** The three rows show the results of the pathway analysis methods ORA, SPIA, and LEGO. For each barplot, the x-axes show various values of the thresholds for each method. For the classical approach, the x-axes show various combinations of fold change and p-value thresholds. For the proposed approach the x-axes show the range from the point determined by change point analysis to the maximum value of edge scores. The y-axes in each graph show the scales for the false positive rate (FPR) and the true positive rate (TPR). The right y-axes in each graph represent the number of DE genes shown by the gray bars. The blue dashed line represents the change point safety margin from edge point analysis (the default for the proposed method). In each bar plot, the green bars represent the TPR and red bars represent the FPR for pathways ranking when the significance threshold $\alpha = 0.1$. True positive pathways are those containing the KO gene. The proposed method yields a perfect true positive rate of 100% in every case for its default threshold (blue dashed line). In contrast, the classical approach yields a TPR of less than 40% for all threshold combinations used with ORA (top right panel) and SPIA (middle right panel). The classical approach used with LEGO (lower right panel) yields a TPR varying between 0 and 93%. The figure also shows how the results of the classical approach depend very much on the combination of thresholds used for fold change and p-values, while the results obtained with the proposed method are much more stable.

Figure 18  **Figure shows the results for the dataset NeuroD1.** The KO gene is shown in red in both panels. a) The mechanism found using HighEdgeS shows NeuroD1 is regulating Ins2, Iapp, Gck and Ins1. b) The results were obtained using the classical approach for the same dataset. The classical approach results include the KO gene, but provides no explanation on how the suppression of this gene propagates further and affects the rest of the system.
Figure 19  **Comparison between the results from HighEdgeS (left columns) and the classical approach (right columns) for the Neurod1 dataset.**
The three rows show the results of the pathway analysis methods ORA, SPIA, and LEGO. For each barplot, the x-axes show various values of the thresholds for each method. For the classical approach, the x-axes show various combinations of fold change and p-value thresholds. For the proposed approach the x-axes show the range from the point determined by change point analysis to the maximum value of edge scores. The y-axes in each graph show the scales for the false positive rate (FPR) and the true positive rate (TPR). The right y-axes in each graph represent the number of DE genes shown by the gray bars. The blue dashed line represents the change point safety margin from edge point analysis (the default for the proposed method). In each bar plot, the green bars represent the TPR and red bars represent the FPR for pathways ranking when the significance threshold $\alpha = 0.1$. True Positive pathways are those containing the KO gene. The classical approach yields a TPR of 0 for all threshold combinations used with ORA (top right panel). For SPIA and LEGO, the classical approach yields 0 true positives for the more lenient thresholds, to the left of the x axis. For the more stringent thresholds, in this dataset there is only one DE gene left, the KO gene, which in turn identifies the true positive pathway. As explained in the text, the correct threshold is never known a priori which means that the classical approach may or may not identify the correct pathway, depending on the choice of the threshold. In contrast, the proposed method yields a perfect true positive rate of 100% in every case for its default threshold that is calculated automatically (blue dashed line).

Figure 20  **The comparison between the putative mechanisms constructed by HighEdgeS and the classical approach for the Pdx1 dataset.**
a.) The putative mechanism constructed by HighEdgeS. The graph shows the knocked out gene Pdx1 regulating 5 downstream genes.

b.) The results for the same dataset when the classical approach was applied. The results of the classical approach lack any connection among the genes and the KO. However, the results of the proposed approach show the interaction between the Pdx1 gene and the Ins1. In fact, the authors of the dataset discussed this interaction in their work indicating the role of the Ins1 in this condition.
Figure 21  **Comparison between the results from HighEdgeS (left columns) and the classical approach (right columns) for the Pdx1 dataset.**
The three rows show the results of the pathway analysis methods ORA, SPIA, and LEGO. For each barplot, the x-axes show various values of the thresholds for each method. For the classical approach, the x-axes show various combinations of fold change and p-value thresholds. For the proposed approach the x-axes show the range from the point determined by change point analysis to the maximum value of edge scores. The y-axes in each graph show the scales for the false positive rate (FPR) and the true positive rate (TPR). The right y-axes in each graph represent the number of DE genes shown by the gray bars. The blue dashed line represents the change point safety margin from edge point analysis (the default for the proposed method). In each bar plot, the green bars represent the TPR and red bars represent the FPR for pathways ranking when the significance threshold $\alpha = 0.1$. True Positive pathways are those containing the KO gene. The figure shows that the proposed approach yields a higher TPR than the classical approaches overall. For ORA, the classical approach fails to identify any of the 3 pathways containing the KO gene independently of the threshold combination used (top right panel). The classical approach used with SPIA identifies 1, 2 or 3 of the 3 true positive pathways depending on the threshold combination used (middle right panel). Finally, the classical approach used with LEGO identifies only one of the 3 true positive pathways and that only for about 1/3 of the range of thresholds explored. In contrast, the proposed approach used with ORA (top left) and SPIA (middle left) identifies all true positive pathways for any edge scores. When combined with LEGO (bottom left), the proposed approach identifies 2 out of the 3 true positives, still much better than the classical approach with LEGO (bottom right).  

Figure 22  **The workflow of the method.** On the training data $Tr$ we applied the equations 4.1, 4.2, and 4.3 to calculate the matrices $F$ for the 7 ethnic groups. Each row in $F$ stored the genotype frequencies for a particular SNP in a particular group. $P$ is the profile matrix for sample $y$. Each element $c$ in the profile matrix $P$, stored the frequency value from $F$ for each group that matched the observed genotype for sample $y$. The sum of the columns in $P$ is calculated and stored in the vector $S$. The class label of sample $y$ will be the name of the element with maximum value in $S$.  

---

71

---

82
Figure 23 The comparison of the performance of the proposed classifier (red),
the naive Bayes classifier (green), the Random Forest classifier (blue),
and the BIASLESS classifier (orange) on each set of SNPs. For each
ethnic group (i.e. Africa, America, Central South Asia, East Asia, and
Europe), the four classifiers were applied on sets of randomly selected
SNPs of various sizes \( r = \{10, 50, 75, 100, 250, 500, 750, 1000, 2500, 5000, 7500, \text{and } 7684\} \). Each box plot shows the values of each cri-
terion (e.g. sensitivity) across the five ethnic groups, when each clas-
sifier was applied on a particular number of SNPs set (e.g. 10). The
proposed classifier yielded the best F1 scores in 10 out of the 12 cases.

Figure 24 The comparison of the overall performance of the proposed classifier
(red), the naive Bayes classifier (green), the Random Forest classifier
(blue), and the BIASLESS classifier (orange) across all sets of SNPs.
For each ethnic group (i.e. Africa, America, Central South Asia, East
Asia, and Europe), the four classifiers were applied on sets of ran-
domly selected SNPs of various sizes \( r = \{10, 50, 75, 100, 250, 500, 750, 1000, 2500, 5000, 7500, \text{and } 7684\} \). Each box plot shows the performance summary of each measure for each method across the
entire set of SNPs. Each box plot includes 60 values (five populations
and 12 SNPs sets in each). The proposed classifier outperformed all
other approaches in terms of accuracy, F1 score, and sensitivity (pan-
els 1, 2 and 3, left to right). The mean specificity of the proposed
approach is very comparable with that of the random forest classifier
but the latter yielded a tighter range (last panel).

Figure 25 The comparison of the performance of the proposed classifier (red),
the naive Bayes classifier (green), the Random Forest classifier (blue),
and the BIASLESS classifier (orange) on sets of SNPs of various sizes.
Each of the four classifiers was applied on a number of randomly
selected SNPs \( r = \{10, 50, 75, 100, 250, 500, 750, 1000, 2500, 5000, 7500, \text{and } 7684\} \). Each box plot shows the values of each criterion for
samples collected only from the USA but from individuals belonging
to different ethnic groups (i.e. Africa, America, Central South Asia,
East Asia, and Europe). This assesses the ability of various approaches
to address the given problem, i.e. to correctly identify the ethnic
background of a sample. The proposed classifier yielded the best F1
scores in 11 out of the 12 cases.
Figure 26  The comparison of the overall performance of the proposed classifier (red), the naive Bayes classifier (green), the Random Forest classifier (blue), and the BIASLESS classifier (orange) across all sets of SNPs. The four classifiers were applied on sets of randomly selected SNPs of various sizes $r = \{10, 50, 75, 100, 250, 500, 750, 1000, 2500, 5000, 7500, \text{and } 7684\}$. Each box plot shows the performance summary of each measure for each method across the range of randomly selected SNPs. Each box plot shows the values of each criterion for samples collected only from the USA but from individuals belonging to different ethnic groups (i.e. Africa, America, Central South Asia, East Asia, and Europe). This assesses the ability of various approaches to address the given problem, i.e. to correctly identify the ethnic background of a sample. The proposed classifier outperformed all other approaches in terms of accuracy, F1 score, and sensitivity values. The Random Forest classifier yielded the best mean specificity.

Figure 27  GFR trend among COVID-19 subjects from day one to day eight of ICU stay. (A) Mean $\pm$ SD of GFR (y-axis) in COVID-19 patients without AKI from day one of admission plotted against time, showing the GFR trend. (B) The trend of GFR among COVID-19 patients with AKI from day1-8 of ICU-stay. There was a significant difference between AKI and non-AKI group in terms of GFR. The trend of GFR values during ICU admission among patients with COVID-19 infection for each day. Error bars plots to show the trend of GFR (y-axis) in COVID-19 patients without acute renal injury (A) and with acute renal injury (B) during ICU stay. Each error bars plot shows the data from each day. The trend line connects the means of each day, and the caps represent the SD. The GFR plot shows a continuous deterioration for patients with AKI starting from day 3, in contrast the GFR of patients without AKI shows continuous improvement.
Figure 28  Trend of GFR and MPV during ICU stay among patients with COVID-19 infection. (A) The GFR trend presented as line plot with error bars to show GFR (y-axis) in COVID-19 patients with and without acute renal injury in time-dependent manner (X-axis) during every 3 days of ICU stay. Patients with AKI (red) versus without AKI (black). B) MPV trend of patients with AKI (red) without AKI (black). Days of admission (x-axis). Each error bars plot shows the data of every 3 days. The dot represents the mean, and the cap represents the SD. The FDR p-values of each comparison are shown at the bottom of the plots. The GFR plot shows a deterioration after day 3 for patients with AKI. After 16 days, this trend showed upwards as some of the patients were either discharged or expired. The MPV plot shows a consistently higher level in AKI patients when compared to patients without AKI. After 16 days, this trend of MPV showed downwards. At each time point for GFR and MPV, the comparisons of AKI patients and without AKI were significant as shown of p-value on the bottom of each interval.
CHAPTER 1 INTRODUCTION

There is a substantial amount of new molecular and clinical data that is being generated. This made it possible to analyze that data to: i) identify reliable biomarkers, ii) identify biological mechanisms, iii) classify the population of an individual, and iv) identify the relevant clinical data in a particular disease. These developments will help treat and diagnose patients with various diseases including infectious diseases such as Tuberculosis and coronavirus 2 (SARS-CoV-2) as well as genetic disorders such as cystic fibrosis (CF).

Sarcoidosis is a granulomatous disease of unknown etiology [27]. Currently, we have limited tools to diagnose Tuberculosis [181]. Other methods led to the discovery of cystic fibrosis biomarkers but failed to be useful in clinical practice [165]. The unique features of the novel SARS-CoV-2 infection were unknown during the start of the pandemic, and there was an urgent need to understand the organs that might be compromised. One of the major challenges in life science research is understanding the mechanism involved in a given phenotype. When the correct mechanism is identified, that could benefit understanding the mechanisms of drug action and ultimately the treatment of a disease. Classifying the population of an individual accurately by using DNA information is an important problem in forensic evidence and evolutionary biology [136, 82, 30, 154, 75, 61]; however, working with fragmented DNA from degraded samples is commonly encountered when the samples come from crime scenes. Given the above challenges, I did my research in this area.

In, Chapter 2, I propose a biomarkers identification frameworks for Sarcoidosis, Tuberculosis (TB) and cystic fibrosis (CF) that uses data from T7 phage display cDNA platform.
Sarcoidosis is a systemic granulomatous disease of unknown etiology. Hypergamma-globulinemia and the presence of autoantibodies in Sarcoidosis suggest active humoral immunity to unknown antigen(s). We developed a complex cDNA library derived from tissues of sarcoidosis patients. Using a high throughput method, we constructed a microarray platform from this cDNA library containing large numbers of sarcoidosis clones. After selective biopanning, 1070 clones were arrayed and immunoscreened with 152 sera from patients with sarcoidosis and other pulmonary diseases. To identify the sarcoidosis classifiers, two statistical approaches were conducted: First, we identified significant biomarkers between sarcoidosis and healthy controls, and second identified markers comparing sarcoidosis to all other groups. We identified 14 clones in the first approach and 12 clones in the second approach discriminating sarcoidosis from other groups. We used the classifiers to build a naive Bayes model on the training set and validated it on an independent test set. The first approach yielded an AUC of 0.947 using 14 significant clones with a sensitivity of 0.93 and specificity of 0.88, whereas the AUC of the second option was 0.92 with a sensitivity of 0.96 and specificity of 0.83. These results suggest robust classifier performance. These results show that sarcoidosis is associated with a specific pattern of immunoreactivity that can discriminate it from other diseases.

Tuberculosis (TB) is caused by Mycobacterium Tuberculosis (MTB) and transmitted through inhalation of aerosolized droplets. Eighty-five percent of new TB cases occur in resource-limited countries in Asia and Africa and fewer than 40% of TB cases are diagnosed due to the lack of accurate and easy-to-use diagnostic assays. Currently, diagnosis relies on the demonstration of the bacterium in clinical specimens by serial sputum smear microscopy and culture. These methods lack sensitivity, are time consuming, expensive, and
require trained personnel. An alternative approach is to develop an efficient immunoassay to detect antibodies reactive to MTB antigens in bodily fluids, such as serum. Using sarcoidosis tissue, we developed a T7 phage cDNA library and constructed a T7 phage display cDNA platform. We immunoscreened our microarray platform with sera from healthy (n = 45), smear positive TB (n = 24), and sarcoidosis (n = 107) subjects. Using a student t-test, we identified 192 clones significantly differentially expressed at a False Discovery Rate (FDR) < 0.01. Among those clones, we selected the top ten most significant clones and validated them on an independent test set. The area under receiver operating characteristics (ROC) for the top 10 significant clones was 1 with a sensitivity of 1 and a specificity of 1. Sequence analyses of informative phage inserts recognized as antigens by active TB sera may identify immunogenic antigens that could be used to develop therapeutic or prophylactic vaccines, as well as identify molecular targets for therapy.

**Cystic fibrosis (CF)** is an autosomal recessive disorder affecting the cystic fibrosis transmembrane conductance regulator (CFTR). CF is characterized by repeated lung infections leading to respiratory failure. Using a high-throughput method, we developed a T7 phage display cDNA library derived from mRNA isolated from bronchoalveolar lavage (BAL) cells and leukocytes of sarcoidosis patients. This library was biopanned to obtain 1070 potential antigens. A microarray platform was constructed and immunoscreened with sera from healthy (n = 49), lung cancer (LC) (n = 31) and CF (n = 31) subjects. We built 1,000 naive Bayes models on the training sets. We selected the top 20 frequently significant clones ranked with student t-test discriminating CF antigens from healthy controls and LC at a False Discovery Rate (FDR) < 0.01. The performances of the models were validated on an independent validation set. The mean of the area under the receiver op-
erating characteristic (ROC) curve for the classifiers was 0.973 with a sensitivity of 0.999 and specificity of 0.959. Finally, we identified CF specific clones that correlate highly with sweat chloride test, BMI, and FEV1% predicted values. We show that CF specific serological biomarkers can be identified through immunoscreenings of a T7 phage display library with high accuracy, which may have utility in the development of molecular therapy.

In, Chapter 3, I present a method that can construct networks of genes that can be considered putative mechanisms. A major challenge in life science research is understanding the mechanism involved in a given phenotype. The ability to identify the correct mechanisms is needed in order to understand fundamental and very important phenomena such as mechanisms of disease, immune systems responses to various challenges, and mechanisms of drug action. The current data analysis methods focus on the identification of the differentially expressed (DE) genes using their fold change and/or p-values. Major shortcomings of this approach are that: i) it does not consider the interactions between genes; ii) its results are sensitive to the selection of the threshold(s) used, and iii) the set of genes produced by this approach is not always conducive to formulating mechanistic hypotheses. The putative mechanisms constructed by this approach are not limited to the set of DE genes, but also considers all known and relevant gene-gene interactions. We analyzed three real datasets for which both the causes of the phenotype, as well as the true mechanisms were known. We show that the method identified the correct mechanisms when applied on microarray datasets from mouse. We compared the results of our method with the results of the classical approach, showing that our method produces more meaningful biological insights.

In, Chapter 4, I am proposing a classification method that is able to analyze ge-
nomic data, more specifically SNP data, and assign an individual to a particular population/group. A current challenge in forensic evidence is to accurately classify samples using genomic data. Advances in high-throughput technologies made it possible to measure tens of thousands of SNPs from DNA, which in turn make it possible to uniquely identify individuals. However, fragmented DNA due to degradation is a common problem with samples from crime scenes. Here, we are presenting a classification method that is able to use SNPs from as little as 10% of the DNA in the human genome to identify the population background of a sample. We compared the performance of our method with three other classification methods: i) naive Bayes, ii) Random Forest, and iii) BIASLESS. The performance of each classifier was assessed using four criteria: accuracy, sensitivity, specificity, and F1 score. The classifier was developed and trained using data from the Human Genome Diversity Project and subsequently tested on a completely independent data set from the HapMap project. The accuracy, sensitivity, specificity, and F1 score values yielded by the proposed classifier were 0.963, 0.798, 0.983 and 0.827, respectively. The results show that the proposed method outperforms the existing methods.

In, Chapter 5, I present the findings of analyzing clinical data for 81 COVID-19 ICU patients. The coronavirus disease (COVID-19) is a highly transmittable viral infection caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). SARS-CoV-2 utilizes angiotensin converting enzyme 2 (ACE2) to gain entry into human cells. Activation of several proteases facilitates the interaction of viral spike proteins (S1) and ACE2 receptors. This leads to cleavage of host ACE2 receptors. COVID-19 disease encompasses a spectrum of systemic involvement far beyond respiratory failure alone. Several features of this disease, including the cause of acute kidney injury (AKI) and the hypercoagula-
ble state, remain poorly understood. Here, we show evidence that mean platelet volume (MPV) reflects platelet activation and activation of coagulation cascades and plays a major role in the development of acute renal failure. Furthermore, we show that the glomerular filtration rate (GFR) values deteriorated after day 3 for patients with acute renal injury (AKI).
CHAPTER 2 NOVEL T7 PHAGE DISPLAY LIBRARY DETECTS CLASSIFIERS FOR SARCOIDOSIS, ACTIVE MYCOBACTERIUM TUBERCULOSIS INFECTION AND CYSTIC FIBROSIS

2.1 Introduction

Sarcoidosis is a granulomatous disease of unknown etiology [27], yet the unifying environmental or genetic factors as initiators of this disease have not been found [71, 41, 47, 129]. Sarcoidosis affects multiple organs, such as lungs, skin, CNS and the eyes [27, 71, 26, 72]. Other immunological features include a shift towards T helper type1 response, lymphopenia or neutropenia, and in some cases increased production of autoantibodies [4, 177, 81, 29].

Sarcoidosis often coincides with other autoimmune disorders such as lupus erythematosus, vitiligo [177], autoimmune hepatitis, and Crohn’s disease (CD) [177, 113, 124, 140]. Several studies have suggested that the cellular and humoral responses associated with granuloma formation in this disease are the consequence of an exaggerated immune response to unknown antigens [56, 119]. Furthermore, subjects with sarcoidosis share several features, such as the presence of non-caseating granuloma, a lack of cutaneous reaction to tuberculin skin testing, and increased local and circulating inflammatory cytokines [27, 26, 72]. Interestingly, lack of responsiveness to PPD can also occur in other inflammatory diseases such as CD, rheumatoid arthritis (RA), or infectious diseases such as leprosy [129, 12, 118]. Pulmonary sarcoidosis and active pulmonary Tuberculosis (TB) share a number of clinical, radiological and histological similarities making differential diagnosis challenging.

Hypergammaglobulinemia, widely regarded as non-specific, is a frequent finding in
sarcoidosis that may suggest active humoral immunity to unknown antigen(s) [81]. Targeted studies evaluating humoral immunity in sarcoidosis have shown elevated IgG levels against components of various pathogens (mycobacterium tuberculosis and propionibacterium acne) [46, 51], as well as against several cellular components, a component intermediate filament protein, and others [62, 88, 170]. These data suggest the development of humoral responses against various antigens of different origins in this disease that can be profiled as diagnostics or to identify novel antigens contributing in the pathogenesis of the disease.

The prevalence of sarcoidosis is higher in the northern hemisphere. Furthermore, it has been reported that the incidence of sarcoidosis is increasing in the developing world and China [7, 98]. Therefore, the development of highly accurate diagnostic classifiers for the diagnosis of sarcoidosis has significance worldwide. To identify the sarcoidosis-associated antigens, we constructed four different T7 phage display cDNA libraries, two of which originated from sarcoid bronchoalveolar lavage (BAL) cells and white blood cells (WBCs). Two other cDNA libraries were derived from cultured human embryonic fibroblasts and splenic monocytes. We combined all 4 libraries into a complex sarcoidosis library (CSL). This novel complex library is custom made for the discovery of biomarkers of respiratory disorders, in particular for sarcoidosis [170, 168, 169, 171]. Recently, we have shown that our microarray technology detects specific classifiers for various respiratory diseases [170, 168, 169, 171]. Here, we tested the hypothesis that this technology is able to identify the specific classifiers for sarcoidosis in early stages within a large heterogeneous group of study subjects, including, heathy controls, Tuberculosis and lung cancer.

In addition to sarcoidosis, there is also a need for identifying biomarkers for other
diseases such as Tuberculosis (TB).

Tuberculosis (TB) remains a serious global health threat with 10 million new cases and 1.7 million deaths each year [93, 123]. Currently, we have limited tools available to diagnose active TB, predict treatment efficacy and the cure of Tuberculosis, or to detect the reactivation of a latent Tuberculosis infection, and assay the induction of protective immune responses through vaccination. A major obstacle to global control of TB remains inadequate case detection [181]. Efforts during the past decade to consistently diagnose and treat most infectious cases have slowed the TB incidence rate, but have not yielded substantial progress [181]. The existing TB diagnostic pipeline still does not have a simple, rapid, inexpensive point-of-care test [181]. Qualified Tuberculosis biomarkers are most urgently needed as predictors of reactivation, cure, and indicators of vaccine-induced protection [181].

Pulmonary Tuberculosis has clinical and pathological similarities with sarcoidosis. Sarcoidosis is a systemic granulomatous disease of unknown etiology with predominant involvement of the lungs, among other organs [42, 116, 163, 146]. Several studies have suggested that the cellular and humoral responses associated with granuloma formation in sarcoidosis are the consequence of an exaggerated immune response to specific Mycobacterium Tuberculosis (MTB) antigens [42, 19]. Sarcoidosis tissue has yielded MTB components including, ESAT6 and catalase-peroxidase (mKatG) [40]. Despite the presence of specific TB antigens in sarcoidosis lung tissues [19, 63, 128, 148], patients with sarcoidosis negatively respond to the tuberculin skin test and are considered to be anergic [112]. Additionally, sarcoidosis subjects rarely ever develop Tuberculosis. Lungs are highly involved both in sarcoidosis and TB. Resident alveolar macrophages (AMs) play an
important role in the pathogenesis and host defense of both diseases [42, 122, 134, 167]. It has been shown that AMs provide a reservoir for MTB and other slow growing organisms [128, 122, 134, 121]. Additionally, AMs play an integral role in autoimmunity and the initiation of fibrosis [122]. Based on this knowledge, we hypothesized that bronchioalveolar cells (BALs) of sarcoidosis subjects may harbor degradation products of specific pathogen(s), including MTB. We constructed four T7 phage display cDNA libraries, two of which originate from sarcoidosis BAL cells and white blood cells (WBCs), and two others derived from cultured human embryonic fibroblasts and splenic monocytes, and combined all four libraries into a complex library [170, 169]. We selected 1070 clones through biopanning and constructed a microarray platform with the selected clones. Previously, upon immunoscreening of this platform with sera from healthy controls, sarcoidosis and culture positive TB patients, we showed that we can detect highly sensitive and specific biomarkers for TB in the sera of subjects with culture positive MTB [170, 171]. In that study, the TB patients were smear negative but culture positive and at the time of sera collection, they were on treatment with anti-tuberculosis agents [170, 171]. To investigate whether our display library also detects specific biomarkers in sera from smear positive MTB patients and if these biomarkers differ from those of smear negative but culture positive TB, we immunoscreened T7 phage display libraries with sera of smear-positive TB patients. The objective of the present study was to identify the specific diagnostic biomarkers from the sera of TB patients who had active TB. We discovered reactive clones that distinguished sera from active TB patients from sarcoidosis patients and uninfected control sera with a high sensitivity and specificity.

In addition to TB, there is a tremendous need for developing reliable serum based
biomarkers for other diseases such as cystic fibrosis (CF).

Cystic fibrosis is an autosomal recessive disease caused by mutations in the gene encoding the cystic fibrosis transmembrane conductance regulator (CFTR) [166]. Currently, there are more than 1300 various mutations in CFTR gene that are known to cause the CF phenotype. The CF phenotype is characterized by chronic bacterial airway infections, neutrophilic inflammation with mucus in airways. Mutations in the CFTR gene affect the epithelial innate immune function in the lungs, resulting in exaggerated and ineffective airway inflammation that fails to eradicate pulmonary pathogens [25]. Bacterial infections in CF are characterized by organisms that have substantial genetic flexibility to evade phagocytic clearance and develop resistance to multiple antibiotics [25]. Repeated or chronic microbial infections are thought to be the major contributor to excessive inflammation leading to CF lung damage. In addition to chronic lung infections, CF subjects may exhibit exocrine pancreatic insufficiency, diabetes mellitus, and sexual organ dysfunction.

Circulating autoantibodies in CF sera have been widely reported, yet their significance is unknown [20, 15, 132]. Various proteins and protein degradation products have been explored as candidate biomarkers for clinical outcome, such as neutrophil elastase, IL-8 [37], and degradation products of lung surfactant protein SP-A [179, 37, 152, 153]. A variety of proteomic approaches exploited antigenic biomarkers that could provide candidates for the diagnosis of infection, prognostic indicators or vaccine development. Pedersen et al., used antibodies from CF patients to probe a protein array of body fluids prepared by two-dimensional gel electrophoresis for antigenic biomarker detection in Pseudomonas aeruginosa [132]. Others identified the outer membrane protein OprL as a seromarker for the initial diagnosis of Pseudomonas aeruginosa infection in CF patients [145].
Because the complex sarcoidosis library (CSL) represents a segment of the human lung microbiome, we hypothesize that it contains potential antigens relevant to CF. To test this, we immunoscreened our microarray platform with sera from healthy controls, CF and lung cancer patients using the power of antibody recognition present in human sera to discover potential serological biomarkers in CF.

2.2 Materials and Methods

2.2.1 Chemicals

All chemicals were purchased from Sigma-Aldrich (St. Louis, MO, USA) unless specified otherwise. LeukoLOCK filters and RNAlater were purchased from Life Technologies (Grand Island, NY, USA). The RNeasy Midi kit was obtained from Qiagen, (Valencia, CA, USA). The T7 mouse monoclonal antibody was purchased from Novagen (San Diego, CA, USA). Alexa Fluor 647 goat anti-human IgG and Alex Fluor goat anti-mouse IgG antibodies were purchased from Life Technologies (Grand Island, NY, USA).

2.2.2 Patient Selection

This study was approved by the institutional review board at Wayne State University and the Detroit Medical Center. Sera were collected from 4 groups: (1) healthy volunteers; (2) Sarcoidosis subjects; (3) smear positive pulmonary TB patients; (4) Cystic fibrosis subjects; and (5) sera from subjects with adenocarcinoma of the lungs (LC). All study subjects signed a written informed consent. All methods were performed in accordance with the human investigation guidelines and regulations by the IRB (protocol No = 055208MP4E) at Wayne State University. All sarcoidosis subjects were ambulatory patients. Sera from patients with Tuberculosis were obtained from the Foundation for Innovative New Diag-
nostics (FIND, Geneva, Switzerland). All TB patients had smear positive sputum.

2.2.3 Serum Collection

Using standardized phlebotomy procedures blood samples were collected and stored at -80 °C [170].

2.2.4 Construction and Biopanning of T7 Phage Display cDNA Libraries

We have used the same T7 phage display libraries as before [170, 169]. Briefly, T7 phage display libraries from BALs, WBCs, EL-1 and MRC5 were made to generate a complex sarcoid library (CSL) [170, 169]. Differential biopanning for negative selection was performed using sera from healthy controls to remove the non-specific IgG, and sarcoidosis sera for positive enrichment as described previously [170, 169].

2.2.5 Microarray Construction and Immunoscreening

A total of 1070 individually picked phage clones from the biopannings for microarray construction were the same as used in previous studies [170, 169]. The phage lysates were arrayed in quintuplicates onto nitrocellulose FAST slides (Grace Biolabs, OR, USA) using the ProSys 5510TL robot (Cartesian Technologies, CA, USA). The nitrocellulose slides were hybridized with sera and processed as described previously [170].

2.2.6 Sequencing of Phage cDNA Clones

Individual phage clones were PCR amplified using T7 phage forward primer 5’ GTTC-TAT CCGCAACGTATGG 3’ and reverse primer 5’ GGAGGAAAGTCGTTTTTTGGG 3’ and sequenced by Genwiz (South Plainfield, NJ, USA), using T7 phage sequence primer TGC-TAAGGACAACG TTATCGG. cDNA sequences of T7 phage clones obtained from Genwiz were translated into peptide/protein sequences using ExPASy translate tool. The length of each peptide clone is determined after the last amino acid of linker sequence (GDPNSS)
inserted in the frame of T7 phage till the stop codon of the sequence. Using NCBI protein BLAST site each identified sequence was used for further BLAST. For each peptide, we performed three BLASTs. First, the identified sequences were randomly blasted to the sequence data without indication of specific species. Second, we used random BLAST to the human genome and thirdly to the mycobacterium genome. We selected the proteins with the highest homology with our peptide sequence.

2.2.7 Data Acquisition and Pre-Processing

Following the immunoreaction, the microarrays were scanned in an Axon Laboratories 4100 scanner (Palo Alto, CA, USA) using 532 and 647 nm lasers to produce a red (Alexa Fluor 647) and green (Alexa Fluor 532) composite image. Cy5 (red dye) labeled anti-human antibody was used to detect IgGs in human serum that were reactive to peptide clones, and a Cy3 (green dye) labeled antibody was used to detect the phage capsid protein [170]. Using the ImaGene 6.0 (Biodiscovery) image analysis software, the binding intensity of each peptide with IgGs in sera was expressed as log2 (red/green) fluorescent intensities. These data were pre-processed using the limma package in the R language environment [169, 149, 175] and the normexp method was applied to correct the background [169, 150]. Within array normalization was performed using the LOESS method [170, 150, 191]. The scale method was applied to normalize between arrays [150, 191]. Intensity ratio of a clone in active TB samples divided by the same clone intensity ratio from healthy control samples were calculated to determine the fold change of a clone. To determine the fold change for CF clones, we also divided the intensity ratio of a clone in CF samples by the same clone intensity ration from healthy controls.
2.2.8 Statistical Analyses Sarcoidosis

To detect differentially expressed antigens for sarcoidosis, we applied a two-tailed t-test. To correct for multiple comparisons, we applied the false discovery rate (FDR) algorithm with a threshold of either 0.05 or 0.01 FDR [11]. All significant clones were sorted in an increasing order. We approach two statistical analyses using two-tailed t-tests. In option 1, we applied a t-test between sarcoidosis training samples versus healthy controls training samples. Out of the 52 sarcoidosis samples, 26 samples were randomly assigned to the training set and the other 26 samples to the testing set. The training and testing set for the 45 healthy controls were randomly assigned to 23 samples in training and 22 samples in test sets. In the testing set, we added 24 tuberculosis samples and 31 lung cancer samples. In option 2, we randomly split the samples from all groups in half. We assigned the first half of 23 control, 26 sarcoidosis, 16 lung cancer, and 12 tuberculosis samples to the training set. We assigned the second half of 22 healthy controls, 26 sarcoidosis, 16 lung cancer, and 12 tuberculosis samples to the testing set. We applied a t-test between sarcoidosis training samples versus healthy controls, lung cancer, and tuberculosis training samples to identify significant clones. For both options, we assessed the performance of classifiers clones, by applying principal component analysis (PCA), agglomerative hierarchical clustering (HC), heatmap, and naive Bayes classifier. The naive Bayes classifier model was built on the training samples to predict sarcoidosis samples from others (healthy controls, tuberculosis and lung cancer) samples and tested the classification model on the testing set (samples not used in the training set).
2.2.9 Statistical Analyses Tuberculosis

To detect differentially expressed antigens for TB, we applied a two-tailed t-test. In order to correct for multiple comparisons, we applied the false discovery rate (FDR) algorithm with a threshold of 0.01 FDR [11]. All significant clones were sorted in an increasing order. The top ten highly significant clones were considered as “classifier clones”. We randomly split the TB and healthy controls samples into: (i) training; (ii) test sets. Out of the 24 TB samples, 12 samples were randomly assigned to the training set and 12 samples to the testing set. The training and the testing sets for the 45 healthy controls were randomly assigned to 23 training and 22 test sets. A t-test was applied between TB-training samples versus healthy controls training samples. All 107 sarcoidosis samples were assigned to the testing set. To assess the performance of classifiers clones, we applied principal component analysis (PCA), agglomerative hierarchal clustering (HC), heatmap, and linear discriminant analysis (LDA). The LDA model was built on the training samples to predict TB samples from others (healthy controls and sarcoidosis) samples, and tested the classification model on the testing set (samples not used in the training set). We performed the classification on the classifiers clones. We applied principal component analysis (PCA), agglomerative hierarchal clustering (HC), and heatmap with all samples (training and testing). Those analyses were first applied to all clones (1070 clones), the significant clones, and then with the 10 highly significant classifier clones.

2.2.10 Statistical Analyses Cystic Fibrosis

To detect frequently differentially expressed antigens for CF we applied a two-tailed t-test. To evaluate the significant CF antigens identified with t-test, we applied principal
component analysis (PCA), agglomerative hierarchal clustering (HC), heatmap, and naive Bayes classifier. To avoid the problem of over-fitting, we randomly split the CF and healthy controls samples into: i) training, ii) test, and iii) validation sets. Out of the 31 CF samples, 21 samples were randomly assigned into training (10 samples) and testing (11 samples) sets. We repeated this process 1000 times to generate 1000 training and test sets. The remaining 10 CF samples were used as an independent validation set. The 1000 training and testing sets for the healthy controls were randomly selected from 33 out of 49 samples (16 training and 17 test set). Therefore, the number of samples for the validation set for healthy controls was 16. While 31 LC samples were randomly split into test (15 samples) and validation (16 samples) sets. For CF specific clones selection, we applied a t-test between the 1000 CF training sets vs. 1000 healthy control-training sets. To correct for multiple comparisons, we applied the false discovery rate (FDR) algorithm with a threshold of 0.01 FDR [170]. The frequency of each significant clone (FDR < 0.01) across all 1000 runs was calculated and sorted based on their frequency of occurrence. The top 20 clones were considered highly significant CF clones. We built a naive Bayes classifier on each of the 1000 training sets and tested the classifier model on the 1000 testing sets. Finally, the classification model was validated on a completely independent validation set. The range of clones starts with the most frequent clone followed by adding one clone at a time. We constructed the models on training sets and applied the model on the testing sets, as well as the validation set. Finally, we determined the correlation of biomarkers with body mass index (BMI) and % predicted forced expiratory volume (FEV1) of CF patients. We calculated combinations of 5 clones from the top set of markers. For each combination, the aggregated vector was calculated from the mean of 5 clones and the correlation between
the aggregated vector and BMI and FEV1% predicted was computed.

### 2.3 Results Sarcoidosis

#### 2.3.1 Complex Sarcoidosis (CSL) library detects unique antigens in the sera of Sarcoidosis

A panel of 1070 clones exhibits a limited ability to class separate sarcoidosis immune reactivity from other diseases. From two highly enriched pools of T7 phage cDNA libraries through biopanning of the CSL library, we randomly selected 1070 potential antigens [170, 168, 169]. This antigen panel was used to construct microarray platform that was immunoscreened with 152 sera from diverse study subjects that included: healthy controls (n = 45); sarcoidosis (n = 52), smear-positive TB patients (n = 24), and lung adenocancer (LC) patients (n = 31). The demographics of the study subjects are shown in (Table 1). Following immunoreaction, the microarray data were pre-processed and then analyzed as previously described [170, 168, 169]. To assess the performance of 1070 clones, we performed an unsupervised principal component analysis (PCA) using all 1070 clones with data from 152 study subjects. As shown in Fig. 1a, several healthy controls and sarcoidosis patients were clustered with TB and lung cancer groups. We also performed unsupervised hierarchical clustering (HC) with all 1070 clones on these 152 samples. We observed the magenta cluster with a mix of samples and lacks specific sub-clusters of sarcoidosis samples (Fig. 1b). Figures 1a and 1b show that using all 1070 clones lacks the ability to class separate the sarcoidosis samples from other samples.
Figure 1: PCA and Hierarchical clustering (option1). (a) PCA plot along PC1 and PC2 generated with 1070 clones of four groups: (1) healthy control samples; (2) Sarcoidosis samples; (3) TB samples; and (4) Lung cancer. Biomarker clusters along the PC1 explain a variance of only 14%, while the variance along PC2 was about 13%. (b) The hierarchical clustering was applied on the healthy controls, sarcoidosis, TB patients and lung cancer with 1070 clones. No distinct clustering is observed when all clones are used.
Table 1: Subjects demographics.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Control Subjects</th>
<th>Sarcoidosis Subjects</th>
<th>TB Subjects</th>
<th>Lung Cancers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (Mean ± SEM)</td>
<td>40.3 ± 7.5</td>
<td>30.6 ± 11.8</td>
<td>40.5 ± 8.5</td>
<td>62.8 ± 11.8</td>
</tr>
<tr>
<td>Gender, N (%)</td>
<td>45</td>
<td>52</td>
<td>24</td>
<td>31</td>
</tr>
<tr>
<td>Male</td>
<td>12 (26)</td>
<td>11 (21)</td>
<td>14 (58)</td>
<td>13 (42)</td>
</tr>
<tr>
<td>Female</td>
<td>33 (74)</td>
<td>41 (79)</td>
<td>10 (42)</td>
<td>18 (58)</td>
</tr>
<tr>
<td>Race, N(%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>31 (69)</td>
<td>49 (94)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>African</td>
<td>–</td>
<td>–</td>
<td>4 (17)</td>
<td>–</td>
</tr>
<tr>
<td>Caucasian</td>
<td>–</td>
<td>3 (6)</td>
<td>–</td>
<td>31 (100)</td>
</tr>
<tr>
<td>Asians</td>
<td>14 (31)</td>
<td>–</td>
<td>20 (83)</td>
<td>–</td>
</tr>
<tr>
<td>BMI (Mean ± SEM)</td>
<td>27 ± 3.8</td>
<td>28 ± 10.5</td>
<td>28 ± 6.9</td>
<td>28 ± 8.7</td>
</tr>
<tr>
<td>Neuro-ophthalmologic</td>
<td>NA</td>
<td>31 (60)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Lung</td>
<td>NA</td>
<td>48 (92)</td>
<td>24 (100)</td>
<td>31(100)</td>
</tr>
<tr>
<td>Skin</td>
<td>NA</td>
<td>36 (69)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Multiorgan</td>
<td>NA</td>
<td>45 (86)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>PPD</td>
<td>NA</td>
<td>Negative</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TB smear</td>
<td>NA</td>
<td>Negative</td>
<td>Positive</td>
<td>–</td>
</tr>
</tbody>
</table>

PPD (purified protein derivative)

Improvement of class separation by identifying significant clones (option1) and (option2)

To improve the classification performance, we applied two different t-tests: i) sarcoidosis training samples vs. healthy control training samples (option 1), and ii) sarcoidosis training samples vs. the rest (healthy controls, LC and TB samples (option 2). Two options resulted in two sets of differentially expressed clones. The first approach (option 1) at the threshold of False Discovery Rate (FDR)<0.05 identifies 132 and at the threshold of FDR (<0.01) identifies 14 significantly different clones between sarcoidosis and healthy controls.

Two PCA plots were constructed to determine if the 132 significant clones (FDR<0.05) and 14 clones (FDR<0.01) from option1 can improve the class separation of sarcoidosis immune reactivity from healthy controls, TB and lung cancer samples. As shown in Fig. 2a, using 132 significant clones improved the class separation of sarcoidosis subjects from all
other groups with a variance of 33% along the PC1 (Fig. 2a). Similarly, using hierarchical clustering showed an improved separation of sarcoidosis samples from all the others (Fig. 2b). Decreasing the FDR threshold to 0.01, we identified 14 highly significant clones differentially reactive in sarcoidosis versus healthy controls. When we constructed a PCA plot utilizing the 14 final clones from option 1, it resulted in a clear class separation of sarcoidosis samples from healthy controls, TB, and LC patients. The plot revealed distinct discrimination along the PC1 direction that could explain forty-five percent of variance (Fig. 2c). Similarly, HC algorithm was applied using 14 clones, we observed a distinct hierarchical linkage separating sarcoidosis samples from other samples (Fig. 2d).

**Option 2 selected clones aid to a robust class separation**

The statistical approach in option 2 yielded in 221 significant clones (FDR <0.01). This option represents a clinical setting example. To demonstrate the performance of the clones identified through option 2 (sarcoidosis samples versus all other samples), we applied PCA and hierarchical clustering. As shown in Fig. 3a, using 221 clones improved the class separation of sarcoidosis subjects from all other groups with a variance of 32% along the PC1. Similarly, when the hierarchical clustering algorithm was applied using 221 significant clones (FDR<0.01), we observed a distinct hierarchical linkage nearly perfectly separating the sarcoidosis patients from TB and well separation from LC and healthy controls (Fig. 3b). Furthermore, we sorted the clones based on the p-values and chose top 12 reactive clones in option 2 to construct PCA plot and hierarchical clustering. As shown in figures 3c and 3d, using the top 12 clones aided in the class separation of sarcoidosis subjects from all other groups with a variance of 54% along the PC1 (Fig. 3c). A distinct hierarchical linkage is well separating the sarcoidosis samples from all other samples. The
Figure 2: PCA and Hierarchical clustering (option 1). (a) PCA plot along the PC1 and PC2 results when applied on 132 sarcoidosis clones (option 1). The PC1 explained 0.33 of variance, whereas PC2 explained 13% of the variance. As shown, the sarcoidosis samples are well separated from the lung cancer, TB controls and most healthy control samples. (b) Hierarchical clustering using the top 132 sarcoidosis clones (FDR<0.05). (c) PCA plot generated with the top 14 sarcoidosis clones. The PC1 explained 45% of the variance, whereas PC2 explained 16% of the variance. (d) Hierarchical clustering using the top 14 sarcoidosis clones. This figure demonstrates a good clustering performance with the 132 sarcoidosis clones and the top 14 significant sarcoidosis clones (FDR<0.01).
clustering analysis using the top 14 clones using option 1 and the top 12 using option 2 show a robust clustering of sarcoidosis samples from the rest (healthy controls, TB and LC).

Venn diagram illustrates the significant clones yielded through two different statistical approaches as well as their intersection (Fig. 4a and b). Fig. 4a shows the Venn diagram of 132 clones (FDR< 0.05) from option 1 and 221 clones (FDR<0.01) from option 2. There were 112 shared clones between both options. Fig. 4b shows the Venn diagram of the classifier clones identified in two statistical approaches. As shown, among 14 classifiers in option 1 and 12 classifiers in option 2, six clones were common.

**Identification of classifiers to predict sarcoidosis**

To determine the classification performance of the identified clones using option 1 and 2, we applied the naive Bayes classification method using option 1 and option 2 significant clones. We also assessed the classification performance of the top 14 clones from option 1 and the top 12 clones from option 2. The classification models were trained on the training set and tested to classify sarcoidosis samples from other (healthy control, TB, and LC) on the testing set. As shown in Fig. 5a, the area under the curve (AUC) as a summary of the receiver operating curve (ROC) using the significant 132 clones (option 1) was 0.932 with a true positive (TP) of 24, a true negative (TN) of 71, false negative (FN) of 2 and false positive (FP) of 6. Next, we applied the classifier model on the test set using the top 14 clones from option 1 (FDR< 0.01). The results of this analysis are depicted in Fig. 5b, which shows an improved AUC of 0.947, when compared with the classification model of the 132 significant clones. Fig. 5c, shows the classification results of the 221 significant
Figure 3: PCA and Hierarchal clustering (option 2). (a) PCA plot along PC1 and PC2 generated with 221 clones (FDR 0.01) of the four groups: (1) healthy control samples; (2) Sarcoidosis samples; (3) TB samples; and (4) Lung cancer. The PC1 explains 32% of the variance, while the variance along PC2 was 12%. (b) The hierarchal clustering was applied on the healthy controls, sarcoidosis, TB patients and lung cancer with 221 clones (FDR <0.01). (c) PCA plot along the PC1 and PC2 results when applied on the top 12 sarcoidosis clones. The PC1 explained 54% of the variance, whereas PC2 explained 14% of the variance. As shown, the sarcoidosis samples are well separated from the lung cancer, TB and most healthy control samples. (d) Hierarchal clustering using the top 12 sarcoidosis clones. This figure demonstrates a robust class separation using the top 12 sarcoidosis classifier clones.
Figure 4: Diagrammatic representation of significant clones from two approaches (option 1 and 2). (a) Illustrates the Venn diagram of 132 clones (FDR < 0.05) from option 1 and 221 clones (FDR < 0.01) from option 2 and their intersection. (b) depicts the Venn diagram of the 14 classifiers clones from option 1 and 12 clones from option 2.
clones (option 2) representing an AUC under the ROC of 0.882 with TP of 25, TN of 40, FN of 1 and FP of 9. Similar to option 1, we applied the classification model on the test set using the top 12 clones from option 2. The results of this analysis are depicted in Fig. 5d, which shows an improved AUC of 0.926 when compared with the classification model of the 221 significant clones. These results suggest a robust classifier performance utilizing either the top 14 clones from option 1 or the top 12 clones from option 2.

2.4 Results Tuberculosis

2.4.1 Complex Sarcoidosis (CSL) Library Detects Unique Antigens in the Sera of Active Tuberculosis Patients

A panel of potential antigens was randomly selected from two highly enriched pools of T7 phage cDNA libraries through biopanning of the CSL library [170, 169]. The constructed microarray platform was immunoscreened with 176 sera (45 healthy controls, 24 smear-positive TB patients, and 107 sarcoidosis patients). The demographics of the study subjects are shown in Table 2. Following immunoreaction, the microarray data were preprocessed and then analyzed. First, we performed an unsupervised PCA using all 1070 clones with data from 176 study subjects. As shown in Figure 6a, several healthy controls and sarcoidosis patients are clustered together with TB subjects. We also performed unsupervised hierarchical clustering with all 1070 clones on these 176 samples. We observed the magenta cluster has a mix of samples and lacks specific sub-clusters of TB samples (Figure 6b). Next, we applied a two-tailed t-test and identified 192 clones that were differentially expressed in sera of smear-positive TB as compared to sarcoidosis patients and healthy controls at the FDR < 0.01. To determine whether the selected 192 significant
Figure 5: Classification to predict sarcoidosis from healthy controls, TB patients and LC patients on the testing set. (a) Performance of 132 clones on the testing set (option1). (b) Performance of the top 14 classifier clones on the test set (option1). The ROC curves demonstrate excellent classification performance with AUC of 0.947 with a sensitivity of 0.883 and specificity of 0.923. (c) Performance of 221 clones at the FDR=0.01 (option2) on the testing set. (d) Performance of the top 12 clones (option 2) on the test set. The ROC curves demonstrate strong classification performance with AUC of 0.926 with sensitivity of 0.962 and specificity of 0.837.
clones can improve the class separation of TB samples from healthy controls and sarcoidosis patients, we constructed a PCA plot. As shown in Figure 6c, there is a good separation of TB samples from sarcoidosis and healthy controls, in which twenty six percent of the variance was along PC1. Similarly, when we performed clustering using the 192 significant TB clones on all subjects, we observed a distinct hierarchical linkage clearly separating TB samples from healthy controls and sarcoidosis patients (Figure 6d). Furthermore, we constructed a PCA plot using 10 classifier clones that can differentiate TB patients from healthy controls and sarcoidosis patients. The result in Figure 6e shows a clear separation of TB samples from healthy controls and sarcoidosis patients. Fifty four percent of the variance was explained along PC1. Similarly, when performing the clustering algorithm using 10 TB classifier clones, we observed a distinct hierarchical linkage separating the TB patients from others (Figure 6f).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Control Subjects</th>
<th>Sarcoidosis Subjects</th>
<th>TB Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (Mean ± SEM)</td>
<td>40.3 ± 7.5</td>
<td>30.6 ± 11.8</td>
<td>40.5 ± 8.5</td>
</tr>
<tr>
<td>Race, N(%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>31 (69)</td>
<td>95 (89)</td>
<td></td>
</tr>
<tr>
<td>African</td>
<td></td>
<td></td>
<td>4 (25)</td>
</tr>
<tr>
<td>Caucasian</td>
<td>12 (11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asians</td>
<td>14 (31)</td>
<td>20 (75)</td>
<td></td>
</tr>
<tr>
<td>BMI (Mean ± SEM)</td>
<td>27 ± 3.8</td>
<td>28 ± 10.5</td>
<td>28 ± 6.9</td>
</tr>
<tr>
<td>Organ involvement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuro-ophthalmologic</td>
<td>NA</td>
<td>31 (29)</td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>NA</td>
<td>101 (94)</td>
<td>24 (100)</td>
</tr>
<tr>
<td>Skin</td>
<td>NA</td>
<td>46 (43)</td>
<td></td>
</tr>
<tr>
<td>Multiorgan</td>
<td>NA</td>
<td>65 (61)</td>
<td></td>
</tr>
<tr>
<td>PPD (purified protein derivative)</td>
<td>NA</td>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>TB smear</td>
<td>NA</td>
<td>Negative</td>
<td>Positive</td>
</tr>
</tbody>
</table>

Table 2: Subjects demographics.

Figure 7a displays a heatmap plot of the distinct expression features of 192 significant TB clones among the study subjects. The heatmap using top 10 significant TB clones
Figure 6: PCA and Hierarchal clustering. (a) PCA score plots along PC1 and PC2 were generated with 1070 clones of three groups: (1) healthy control samples (black circles); (2) TB samples (green squares) and; (3) Sarcoidosis samples (blue triangle). Biomarker clusters along the PCA1 explain variance of only 0.15, while the variance along PC2 was about 0.14. (b) The hierarchal clustering was applied on the healthy controls (black labels), TB patients (green labels) and sarcoidosis (blue labels) with 1070 clones. (c) PCA score plots along the PC1 and 2 results when applied on 192 TB clones. The PC1 explained 0.26 of variance, whereas PC2 explained 0.15 of variance. As shown, the TB samples are well separated from the healthy controls and sarcoidosis samples. (d) Hierarchal clustering using only the highly significant 192 TB clones. The blue and black clusters include sarcoidosis and healthy controls, the green cluster includes all the TB samples except one. (e) PCA score plots along PC1 and 2 generated with top 10 highly significant clones. The PC1 explained 0.54 of variance, whereas PC2 explained 0.11 of variance. (f) Hierarchal clustering using 10 top significant TB clones. This figure demonstrates better clustering with 192 TB clones and the highly significant 10 TB clones (panels c, d, e, and f) when compared the clustering using all clones (panels a and b).
Figure 7: Heatmaps were generated based on 192 clones and the 10 highly significant clones from the data of 176 study subjects (45 healthy controls, 24 with TB, and 107 with sarcoidosis) (a,b). Each row represents a clone, while each column represents a study subject. As shown in Figure 2, most of TB samples clustered to the right side of heatmap plots, while sarcoidosis samples and healthy controls clustered on the left side of the plot, indicating different expression profiles.

(classifiers) among study subjects is highlighted as a plot in Figure 7b.

Furthermore, we applied the LDA classification model and calculated the AUC values using 192 TB clones on the testing set (12 TB patients, 107 sarcoidosis patients, and 22 healthy controls). As shown in Figure 8a, the AUC under the ROC using 192 clones was one with no false positive and no false negative prediction. Next, we applied the classification model on the test set using 10 classifier clones. Figure 8b, shows that despite the reduction of clones to 10, the AUC under the ROC remained one, again with no false positive or false negative. These suggest robust classifier performance.
Figure 8: Classification results to predict TB from healthy controls and sarcoidosis patients. (a) Performance of 192 clones on the test set. (b) Performance of the top 10 classifier clones on the test set. The ROC curves demonstrate excellent classification performance with AUC of 1 with sensitivity of 1 and specificity of 1.
2.4.2 Characterization of the Ten TB Classifiers

Based on the results of training and test sets, we characterized the top 10 highly performing active TB clones through sequencing. After obtaining the sequences of clones, the Expasy program was used to translate the cDNA sequences to peptide/protein sequences [170, 169]. Protein blast using algorithms of the BLAST program were applied to identify the highest homology to identified peptides [170, 169]. The identified clones were blasted with human and MTB genomes and then selected those specific peptide sequences with the highest homology of amino acid sequence. All top 10 clones have the highest homology with TB sequences. Additionally, we compared these results with corresponding nucleotide sequences using nucleotide BLAST and determined the predicted amino acids in frame with T7 phage 10B gene capsid proteins. All of the 10 classifier clones are coded by the inserted gene fragments leading to out-of-frame peptides, therefore meeting the criteria of mimotopes [183] (Table 3). As sera of active TB patients reacted with these out-of-frame peptides, it is likely that these TB clones are produced as a result of altered reading frames or alternative splicing, as described in previous studies [170, 169, 183]. Table 3 shows the 10 most significant TB antigens, gene names, sensitivity, specificity, and FDR adjusted p-values. Figure 9 shows the ROC curves for six clones that are increased in TB, while Figure 10 shows ROC curves for four clones decreased in TB.

2.5 Results cystic fibrosis

2.5.1 Complex sarcoidosis library detects unique antigens in the CF sera.

A panel of potential antigens was randomly selected from two highly enriched pools of T7 phage cDNA libraries through biopanning of the CSL library [170]. A microarray
Figure 9: ROCs for the top 6 significant clones that are increased in TB sera compared to healthy controls and sarcoidosis.
Figure 10: ROC for the top 4 significant clones that are decreased in TB sera compared to healthy controls and sarcoidosis. This figure demonstrates reasonable classification performance when the classification was applied on one clone.
<table>
<thead>
<tr>
<th>Clone</th>
<th>Increased in Tuberculosis (TB)</th>
<th>Gene Name</th>
<th>p Value</th>
<th>p Value FDR</th>
<th>AUC</th>
<th>Sen</th>
<th>Spe</th>
</tr>
</thead>
<tbody>
<tr>
<td>P51_BP3_38</td>
<td>Polyketide synthase</td>
<td>Pks13, Rv3800c</td>
<td>4.7e-7</td>
<td>2.79e-5</td>
<td>0.98</td>
<td>1</td>
<td>0.97</td>
</tr>
<tr>
<td>P51_BP3_60</td>
<td>Hydrolase</td>
<td>Rv1723</td>
<td>1.62e-8</td>
<td>3.46e-6</td>
<td>0.95</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>P51_BP3_72</td>
<td>Ferredoxin</td>
<td>fdxA, Rv2007c</td>
<td>1.36e-9</td>
<td>7.27e-7</td>
<td>0.92</td>
<td>0.91</td>
<td>0.89</td>
</tr>
<tr>
<td>P51_BP3_131</td>
<td>Dihydroxy acid dehydratase</td>
<td>ilvD, Rv0189c</td>
<td>2.15e-8</td>
<td>3.84e-6</td>
<td>0.95</td>
<td>0.92</td>
<td>0.98</td>
</tr>
<tr>
<td>P51_BP3_137</td>
<td>Transketolase</td>
<td>TKT, Rv1449c</td>
<td>7.14e-8</td>
<td>9.72e-6</td>
<td>0.95</td>
<td>1</td>
<td>0.81</td>
</tr>
<tr>
<td>P197_BP4_1078</td>
<td>Signal peptidase</td>
<td>lepB, Rv2903</td>
<td>1.44e-7</td>
<td>1.36e-5</td>
<td>0.78</td>
<td>0.92</td>
<td>0.64</td>
</tr>
</tbody>
</table>

**Decreased in Tuberculosis (TB)**

<table>
<thead>
<tr>
<th>Clone</th>
<th>Gene Name</th>
<th>p Value</th>
<th>p Value FDR</th>
<th>AUC</th>
<th>Sen</th>
<th>Spe</th>
</tr>
</thead>
<tbody>
<tr>
<td>P51_BP3_334</td>
<td>TetR family transcriptional regulator</td>
<td>MRA_2532</td>
<td>4.02e-10</td>
<td>4.3e-7</td>
<td>0.98</td>
<td>1</td>
</tr>
<tr>
<td>P51_BP4_403</td>
<td>Menaquinone biosynthesis protein</td>
<td>menD, Rv0555</td>
<td>7.27e-8</td>
<td>9.71e-6</td>
<td>0.95</td>
<td>1</td>
</tr>
<tr>
<td>P51_BP4_497</td>
<td>Cobalamin biosynthesis protein</td>
<td>CobN, Rv2062c</td>
<td>1.11e-8</td>
<td>2.96e-6</td>
<td>0.88</td>
<td>0.92</td>
</tr>
<tr>
<td>P51_BP4_584</td>
<td>5-oxoprolinase</td>
<td>OplA, Rv0266c</td>
<td>5.82e-9</td>
<td>2.10e-6</td>
<td>0.94</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3: 10 Top Significant TB Clones.
platform was constructed and immunoscreened with 111 sera (49 healthy controls, 31 with CF and 31 with adenocarcinoma (LC) of the lungs. The demographics of the study subjects are shown in Table 4. Among the CF patients, 15 (48%) were genotyped as F508del homozygotes, 9 (29%) were heterozygotes for F508del, and 7 (23%) had various mutations such as G542X or 2789 + 5 GT0A/S489X and others (Table 4). Following immunoreaction, the microarray data were pre-processed and then analyzed. We applied a student t-test on 1,000 training sets (FDR < 0.01) between CF vs. healthy controls samples. A total of 599 clones appeared significant at least once. We calculated the frequency of each significant clone and selected the top 20 clones according to their significance and frequency. Furthermore, we performed an unsupervised PCA for all 1070 clones with data from 111 study subject sera. As shown in Fig. 11a, several LC and healthy controls are clustered together with the CF samples. To investigate whether the identified 20 highly significant CF clones can improve class separation of CF samples from LC and healthy controls, we constructed a PCA plot using only those clones (Fig. 11c). Using the 20 highly significant CF clones aided to a class separation of CF samples from LC and healthy controls. Forty nine percent of variance was explained along the PC1.

Next, we performed an unsupervised hierarchical clustering with all 1070 clones on 111 samples. We observed that the magenta cluster has a mix of samples and lacks specific sub-clusters of CF samples (Fig. 11b). In contrast, when the clustering algorithm was performed using the 20 highly significant CF clones on all samples, we observed a distinct hierarchical linkage, clearly demarcating CF samples from others (healthy controls and LC) (Fig. 11d). Distinct expression features of the 20 highly significant CF clones among study subjects are highlighted in a heatmap plot (Fig. 12).
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Control Subjects</th>
<th>CF Patients</th>
<th>Lung Cancer Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (Mean ± SEM)</strong></td>
<td>40.3 ± 11.5</td>
<td>31.7 ± 10.8</td>
<td>62.3 ± 11.9</td>
</tr>
<tr>
<td><strong>Race, N(%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>1 (3)</td>
<td>1 (3)</td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>25 (81)</td>
<td>30 (97)</td>
<td>31 (100)</td>
</tr>
<tr>
<td>Other</td>
<td>5 (16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoker (&gt;50 packs/yr.), N (%)</td>
<td>15 (30)</td>
<td>NA</td>
<td>6 (20)</td>
</tr>
<tr>
<td><strong>BMI (Mean ± SEM)</strong></td>
<td>28 ± 3.6</td>
<td>22.76 ± 0.61</td>
<td>24 ± 4.6</td>
</tr>
<tr>
<td><strong>Sweat Chloride values (mM/L)</strong></td>
<td>NA</td>
<td>103.31 ± 13.5</td>
<td>NA</td>
</tr>
<tr>
<td><strong>PFT Values (Mean ± SEM)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEVI (% predicted)</td>
<td>NA</td>
<td>59.30 ± 4.90</td>
<td>NA</td>
</tr>
<tr>
<td>FVC (% predicted)</td>
<td>NA</td>
<td>75.27 ± 4.49</td>
<td>NA</td>
</tr>
<tr>
<td>TLC (% predicted)</td>
<td>NA</td>
<td>101.47 ± 2.42</td>
<td>NA</td>
</tr>
<tr>
<td>DLCO (% predicted)</td>
<td>NA</td>
<td>89.77 ± 3.90</td>
<td>NA</td>
</tr>
<tr>
<td><strong>Gene Mutation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*Homogygous (Double mutation at F508 del)</td>
<td>NA</td>
<td>15 (48)</td>
<td>NA</td>
</tr>
<tr>
<td>**Heterogygous (Double mutation one with F508 del)</td>
<td>NA</td>
<td>9 (29)</td>
<td>NA</td>
</tr>
<tr>
<td>***Other (Double mutation with none at F508 del)</td>
<td>NA</td>
<td>7 (23)</td>
<td>NA</td>
</tr>
<tr>
<td><strong>Bacterial culture results</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudomonas (mucoid &amp; non-mucoid)</td>
<td>NA</td>
<td>21 (67)</td>
<td>NA</td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td>NA</td>
<td>13 (42)</td>
<td>NA</td>
</tr>
<tr>
<td>Aspergillus</td>
<td>NA</td>
<td>3 (1)</td>
<td>NA</td>
</tr>
<tr>
<td>Adenocarcinoma of Lung, N (%)</td>
<td>NA</td>
<td>NA</td>
<td>31 (100)</td>
</tr>
</tbody>
</table>

**Table 4: Subjects demographics.** Age values are presented as means and variability in SD. N = Number of patients and percent are shown in parentheses.
Figure 11: PCA and hierarchal clustering. (a) PCA score plots along the PC1 and 2 were generated with 1070 clones of three groups: 1) healthy control samples (yellow circle), 2) CF samples (blue triangle) and 3) LC samples (green square). Along the PCA1 explaining a variance of only 0.18 and along the PC2 of 0.12. (b) The hierarchal clustering was applied on the healthy controls (black labels), CF patients (red labels) and LC (blue labels) with 1070 clones. (c) PCA score plots along the PC1 and 2 results when applied on the highly significant 20 CF clones. The PC1 explained 0.49 of variance, whereas PC2 explained 0.09 of variance. As shown the CF samples are well separated from the healthy controls and LC samples. (d) Hierarchal clustering using only the highly significant 20 CF clones. The green cluster includes LC and healthy control samples (no CF samples), the magenta cluster includes all the CF samples, few healthy control and two LC samples. This figure demonstrates better clustering with the highly significant 20 CF clones (panels c and d) when compared with the clustering using all clones (panels a and b).
Figure 12: Heatmap generated based on the 20 highly significant CF clones from the data of 111 study subjects (49 healthy controls, 31 with CF and 31 with LC). Each row represents a clone, while each column represents a study subject. As shown in this figure, most CF samples clustered to the left side of the heat map plot, while the LC samples and healthy controls clustered to the right side of the plot indicating different expression profiles.
Next, we applied the classification model and calculated the AUC values on accumulating numbers of clones (see method section) on the test and validation sets. Figure 13a shows the AUC values for the test set. The lowest average AUC value for the test set was 0.956. Figure 13b graphically represents the performance of the classifier model when applied to the validation set. When we applied the classification model on the validation set, the lowest average AUC value was 0.926. These results clearly indicate that the classification model based on the accumulating number of significant clones when applied on the test and the validation sets have a very good classification performance. Finally, to assess if the identified highly significant CF clones provide a sound classification performance, we applied the naive Bayes classification algorithm with the highly significant CF clones to predict CF samples from healthy controls and LC samples. At the optimal threshold (highest true positivity with lowest false positivity for each of the 1000 runs), we could reliably predict CF from healthy controls and LC samples with a mean specificity of 0.959 (95% CI, 0.11-0.15) and a mean sensitivity of 0.999 (95% CI, 0.18-0.21). The mean AUC under the ROC for the classifier was 0.973 (95% CI, 0.07-0.094) (Fig. 13c).

2.5.2 Characterization of significant CF clones.

Based on the results of training and validation sets, we characterized the 20 highly performing clones through sequencing and identified which clones can predict sweat chloride tests, FEV1% predicted and body mass index (BMI). Among the 20 clones four CF reactive antigens comprise relatively large peptides, while 14 CF antigens are coded by the inserted gene fragments leading to out-of-frame-peptides, hereby meeting the definition of mimotopes [183] (Table 5). As CF sera reacted to these out-of-frame-peptides, it is likely that these clones represent CF antigens that are produced as a result of altered reading
Figure 13: Shows the classification performance of the naive Bayes classifier. The classifier is to predict CF from LC and healthy control samples. (a) Performance of the classifier on the testing sets. Box plots indicate the AUC values (y-axis) when the classifier model was applied on the 1000 test sets. The x-axis is accumulating sets of clones. The accumulation of the clones starts with the most frequent clone and then one clone added at a time to reach 100 clones. (b) Performance of the classifier models on the validation set. As indicated the classifier models when they were built using the significant clones show a high AUC values on the testing sets as well as on the completely independent validating set. (c) The ROCs were generated from the average of the 1000 runs of the classifier models when applied on the validation set (randomly selected healthy controls, CF and LC) using the 20 highly significant CF clones. The ROC curve demonstrates an excellent classification performance with an average AUC of 0.973. These results indicate the excellent performance of the naive Bayes classifier on the 20 highly significant CF clones.
frames or alternative splicing, as shown in previous studies [183, 101]. Table 5 shows the CF antigens, gene names, sensitivity, specificity and FDR adjusted p-value. Finally, we sought to determine whether any of the biomarkers correlate with sweat chloride test, BMI and FEV1% predicted value. Sweat chloride test, PFT and BMI values for CF subjects are shown in Fig. 14. Sweat chloride test is commonly used as a screening tool for CF diagnosis [58]. We found highest spearman correlation ($r = -0.54$) between sweat chloride values and the clone p51-BP3-113 (GEM-5047) (Fig. 14a). By combining this clone with four additional clones a higher correlation was reached ($r = -0.72$) (Fig. 14b). BMI is an important clinical measure among CF patients to predict exacerbation and decline of lung function testing [160]. We found highest spearman correlation ($r = -0.31$) between BMI and the P51-BP3-47 clone (dnaJ homolog) (Fig. 14c). By combining this clone with 4 other clones a higher correlation with BMI was reached ($r = -0.58$) (Fig. 14d). Additionally, we found the highest correlation ($r = -0.42$) between FEV1% predicted and clone P197-BP4-926 (Fig. 14e). The correlation value ($r = -0.6$) improved once we added 4 other clones (Fig. 14f). Seven out of 16 identified clones overlapped with highly specific and sensitive CF clones shown in Table 5. In addition, we identified 6 other clones with a significant correlation with sweat chloride test, BMI and FEV1% predicted values.
<table>
<thead>
<tr>
<th>Clone</th>
<th>Increased in Cystic Fibrosis vs Healthy Controls</th>
<th>NCBI Protein Number</th>
<th>p-Value</th>
<th>FDR p-Value</th>
<th>AUC (95% CI)</th>
<th>Sen (95% CI)</th>
<th>Spe (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P51_BP3_129</td>
<td>Chain A Pseudomonas Aeruginosa Metap, In Mn Form</td>
<td>4FO8</td>
<td>0.0069</td>
<td>0.032</td>
<td>0.964 (0.04-0.10)</td>
<td>0.9 (0.21-0.26)</td>
<td>0.97 (0.08-0.14)</td>
</tr>
<tr>
<td>P51_BP3_250</td>
<td>Beta-lactamase</td>
<td>GEM_5327</td>
<td>0.0125</td>
<td>0.033</td>
<td>0.908 (0.08-0.14)</td>
<td>0.8 (0.20-0.26)</td>
<td>0.93 (0.10-0.17)</td>
</tr>
<tr>
<td>P51_BP3_25</td>
<td>Histidine kinase</td>
<td>narX g2b</td>
<td>0.0017</td>
<td>0.031</td>
<td>0.875 (0.12-0.19)</td>
<td>0.71 (0.22-0.30)</td>
<td>0.95 (0.08-0.17)</td>
</tr>
<tr>
<td>P51_BP3_254</td>
<td>Outer membrane Porin</td>
<td>GEM_5047</td>
<td>0.0056</td>
<td>0.022</td>
<td>0.857 (0.14-0.20)</td>
<td>0.74 (0.26-0.29)</td>
<td>0.93 (0.10-0.16)</td>
</tr>
<tr>
<td>P51_BP3_47</td>
<td>dnaJ homolog</td>
<td>DNAJC10</td>
<td>0.0006</td>
<td>0.05</td>
<td>0.745 (0.08-0.14)</td>
<td>0.82 (0.20-0.21)</td>
<td>0.74 (0.08-0.14)</td>
</tr>
<tr>
<td>P51_BP3_252</td>
<td>γ-glutamyl traspeptidase</td>
<td>PS113-4947</td>
<td>0.0005</td>
<td>0.03</td>
<td>0.727 (0.06-0.11)</td>
<td>0.61 (0.12-0.14)</td>
<td>0.9 (0.11-0.15)</td>
</tr>
<tr>
<td><strong>P197_BP4_830</strong></td>
<td>TetR family transcriptional regulator</td>
<td>GEM_1794</td>
<td>3.44E-05</td>
<td>0.032</td>
<td>0.942 (0.06-0.09)</td>
<td>0.9 (0.21-0.22)</td>
<td>0.94 (0.10-0.12)</td>
</tr>
<tr>
<td><strong>P197_BP4_898</strong></td>
<td>AraC-family transcriptional regulator</td>
<td>APZ15_34865</td>
<td>0.0001</td>
<td>0.019</td>
<td>0.932 (0.07-0.09)</td>
<td>0.99 (0.20-0.26)</td>
<td>0.71 (0.08-0.16)</td>
</tr>
<tr>
<td><strong>P197_BP4_925</strong></td>
<td>HLA-DR alpha</td>
<td>HLA-DR</td>
<td>0.0076</td>
<td>0.037</td>
<td>0.931 (0.07-0.11)</td>
<td>0.99 (0.20-0.26)</td>
<td>0.81 (0.10-0.15)</td>
</tr>
<tr>
<td><strong>P197_BP4_1109</strong></td>
<td>Thoeredoxin like protein</td>
<td>TXNL1</td>
<td>0.001</td>
<td>0.045</td>
<td>0.924 (0.05-0.11)</td>
<td>0.89 (0.18-0.20)</td>
<td>0.87 (0.10-0.13)</td>
</tr>
<tr>
<td><strong>P197_BP4_1152</strong></td>
<td>NADH dehydrogenase subunit 1</td>
<td>MT-ND1</td>
<td>0.0007</td>
<td>0.03</td>
<td>0.895 (0.08-0.20)</td>
<td>0.91 (0.20-0.29)</td>
<td>0.78 (0.15-0.18)</td>
</tr>
<tr>
<td><strong>P197_BP4_834</strong></td>
<td>AMP-dependent synthetase</td>
<td>VL15_07170</td>
<td>0.0036</td>
<td>0.034</td>
<td>0.884 (0.10-0.20)</td>
<td>0.99 (0.20-0.26)</td>
<td>0.71 (0.13-0.18)</td>
</tr>
<tr>
<td><strong>P197_BP4_1114</strong></td>
<td>Peptide ABC transporter substrate binding protein</td>
<td>135_2059</td>
<td>0.0011</td>
<td>0.05</td>
<td>0.845 (0.13-0.17)</td>
<td>0.8 (0.26-0.26)</td>
<td>0.84 (0.14-0.16)</td>
</tr>
<tr>
<td><strong>P197_BP4_762</strong></td>
<td>Ketoacyl-ACP reductase</td>
<td>TabG</td>
<td>0.0011</td>
<td>0.05</td>
<td>0.826 (0.09-0.14)</td>
<td>0.79 (0.12-0.14)</td>
<td>0.87 (0.11-0.17)</td>
</tr>
</tbody>
</table>

Table 5: 14 Top Significant CF Clones.
2.6 Discussion

In the following three sub sections I will first discuss the results of Sarcoidosis in sub-section 2.6.1. Next, I will discuss the Tuberculosis results in sub-section 2.6.2. Then I will discuss the results of cystic fibrosis in sub-section 2.6.3.

2.6.1 Discussion for Sarcoidosis

Patients with sarcoidosis exhibit various immunological features including, a shift towards T helper type 1 response [147], lymphopenia or neutropenia, hypergammaglobulinemia, and in some cases increased autoantibodies [4, 177, 81, 29]. Several studies have suggested that the cellular and humoral responses associated with granuloma formation in this disease are the consequence of an exaggerated immune response to unknown antigens [56, 119]. Numerous studies found components (RNA, DNA) of pathogens including propionibacterium acnes and Mycobacterium tuberculosis in sarcoidosis tissues [56, 119, 46, 14, 117, 80].

Using the serological analysis of antigens by recombinant expression cloning (SEREX) as a basis for autoantibody discovery, we critically examined the relevant methods of biomarker discovery and developed an innovative immunoscreening to optimize the identification of specific autoantibodies [170, 107, 102]. To achieve this goal, a heterologous sarcoidosis library derived from mRNA of numerous sarcoidosis subjects were displayed on the T7 phage system [170, 171]. Furthermore, we used antibody recognition and random plaque selection during biopanning of the libraries to minimize the confounding effects of nonspecific antibodies. Recent evidence indicates that panels of biomarkers can achieve significantly higher accuracy than individual biomarkers [107, 89, 184, 22, 23].
Figure 14: Shows the Pearson correlation of the identified biomarkers with clinical values. Scatter plots depicted correlation of the sweat chloride values with one clone (a) and aggregated 5 clones (b). Scatter plots depicted the correlation for BMI predicted with one clone (c) and aggregated 5 clones (d). Scatter plots depicted correlation of FEV1% with one clone (e) and aggregated 5 clones (f). The correlation values and p values are shown in the top right of each plot. The names of the clones are shown at the bottom of each plot.
Our current data indicate that our technology detects sarcoidosis classifiers in early stages of this disease as compared to various other lung diseases. Important to note that the current sarcoidosis group differs from our previous study group. Sera were collected during initial diagnosis of sarcoidosis and none of the patients were treated with corticosteroids or other immunosuppressive medications. Furthermore, we performed two different statistical approaches to our data: Option 1, first detected the significant biomarkers between healthy controls vs. sarcoidosis; whereas option 2 chose the sarcoidosis clones by comparing sarcoidosis samples vs. all other groups. In both options, we used independent training and testing sets. Interestingly, 6 antigen clones were identical between options 1 and 2. Option 1 yielded in 8 unique clones, whereas option 2 yielded in 6 specific clones. The identification of autoantibodies in sarcoidosis is important, as they may contribute to the cause of disease. Furthermore, sarcoidosis specific immunoreactivity against identified antigen clones can be used to develop a direct ELISA test for detection of autoantibody in sarcoidosis as a diagnostic. Further works are needed to elucidate the role of identified antigen clones in clinical features and organ involvement of this disease.

2.6.2 Discussion for Tuberculosis

Standard methods to diagnose TB and to monitor response to treatment rely on sputum microscopy and culture. The current CDC/NIH roadmap emphasizes the need for the development of new TB biomarkers as alternative methods [123]. Recently, a tremendous effort has been put forward elucidating the antibody responses to MTB antigens, which has implications for the development of new antigens to diagnose and monitor successful treatment, as well as to develop effective vaccination [91]. Most other studies searching for TB antigens have identified unspecific markers primarily involving host response such
as C-reactive protein, serum amyloid A and other non-specific markers [31, 1].

In view of this background, we hypothesized complex library derived from sarcoidosis tissue may harbor degradation products of MTB antigens and these antigens can be used as a bait to specifically and selectively bind to antibodies present in sera from active TB subjects. Our microarray platform identified 10 highly significant TB clones that can discriminate TB patients from healthy controls and sarcoidosis patients. All of these clones are TB specific and related to bacterial growth of M. Tuberculosis and its metabolism (Table 3). Among the 10 TB specific phage peptides, six out-of-frame peptides were increased in sera of active TB patients (Table 3).

We detected these novel antigens using a heterologous library derived from sarcoidosis subjects. Lungs are highly exposed to numerous bacteria and our library is predominantly derived from sarcoidosis BAL cells and WBCs containing diverse immune cells, including macrophages that were exposed to various pathogens. We postulate that the CSL represents a segment of the lung microbe containing diverse antigens for TB, sarcoidosis, and cystic fibrosis [170, 169, 171].

There are various applications of a phage display. In the current work, we used a phage display for the discovery of TB biomarkers. The same system can be applied to identify novel markers for multi-drug resistance in TB, which is becoming a major issue in TB treatment. Additionally, phage displays can be used for the development of specific targeted therapies [187]. The phage display technology and immunoscreening have utilities not only in identifying diagnostic biomarkers, but also may enable us to develop a novel targeted therapy utilizing the peptide sequences (mimotopes) as vehicles to deliver specific drugs. The identified sequences can be used to develop peptide/protein-coated magnetic
nanoparticles for clinical testing or for applications in drug delivery [141]. Additionally, this technology might enable us to discover unknown epitopes targeting specific bacterial antigens leading to immunogenicity and antibody production in TB subjects, as well as providing us with a better understanding of host immune defenses in TB subjects. For instance, TB sera were less reactive to some of the identified clones (TetR, menD, CobN, and OplA), these clones are less likely to be used for diagnostic purposes. However, these clones can be used to develop a new vaccine and to boost immunity against TB infection. Furthermore, this microarray platform can be hybridized to detect IgA in sputum of TB patients that may have clinical values. Moreover, antibody detection in the sera of patients has a potential value in clinical practice, as it is non-invasive and requires a minimal amount of blood or other bodily fluids.

The lack of sensitivity and specificity and cross-reactivity of biomarkers with other diseases dampened the enthusiasm in TB biomarker discovery studies. However, our study shows excellent sensitivity and specificity, not only as compared to healthy controls but also to another granulomatous disease. Other studies using gene expression profiling between TB and sarcoidosis found 94% similarities [90, 108]. Our system has the advantage of detecting TB clones with high sensitivity and specificity and is based on an immune reaction rather than gene expression. The detection of this immune reaction, in the form of antibodies, relies on a complex interaction between antigen presenting cells, T cells and B cells that leads to specific antibody production in response to a TB infection. Highly specific biomarkers may have a potential role as candidate antigens in the development of novel vaccination for TB or for multidrug resistant bacterial infections.
2.6.3 Discussion for cystic fibrosis

CF is characterized by a self-perpetuating cycle of airway obstruction, chronic bacterial infection, and vigorous inflammation that results in bronchiectasis, progressive obstructive lung disease, and marked shortening of life expectancy. Despite having identical cystic fibrosis transmembrane conductance regulator genotypes, individuals with F508del homozygous CF demonstrate significant variability in the severity of pulmonary disease and infection. Non-invasive serological biomarkers that can aid to monitor disease progression or evaluate response to therapy would be extremely valuable. Several groups attempted to identify specific biomarkers to predict inflammation in CF using various biofluids such as sputum, BAL and serum [153, 165]. Most of these methods led to the discovery of a series of markers or expression signatures but failed to be useful in clinical practice [165]. In view of this background, we applied a novel high throughput technology to overcome the current gap by constructing phage-protein microarrays in which peptides were derived from a unique sarcoidosis cDNA library and expressed as a phage fusion protein. Through immunoscreening and rigorous statistical analysis, we identified 20 highly significant CF clones as biomarkers that are able to discriminate between CF and healthy controls as well as lung cancer sera. One important issue in biomarker discovery is the validation of biomarkers and sample selection. To overcome this issue, we randomly assigned samples into 1000 training sets instead of using one training set. Then, we compared the healthy controls and CF samples for each pair of such random sets. The ranking of the top 20 clones was based on the significance and frequency of each clone (how many times each clone appears significant at FDR < 0.01).
Environmental stresses including cigarette smoking, hypoxia, and chronic inflammation have also been implicated in reducing the cystic fibrosis transmembrane conductance regulator (CFTR) function [138, 18]. Additionally, subjects with smoking related chronic obstructive lung disease (COPD) can develop a similar clinical phenotype with recurrent respiratory infections, mucus inspissation and airway obstruction that is attributed to acquired CFTR deficiency [162]. In unsupervised HC, we have seen few false positive classifications that might have been due to the selection of control groups (lung cancer and healthy controls), who had a significant smoking history. Interestingly, these subjects had more than 50 pack years of smoking history.

To our knowledge, no previous study used phage display technology to detect CF serum biomarkers. We detected novel antigens for CF using a heterologous library derived from sarcoidosis subjects. Lungs are highly exposed to numerous bacteria and our library is predominantly derived from sarcoidosis BAL cells and WBCs containing diverse immune cells, including macrophages that were exposed to various pathogens. Hence, we postulate that the CSL represents a segment of the lung microbiome containing diverse antigens including CF specific antigens, sarcoidosis and TB specific antigens [170, 171]. The phage display technology and immunoscreening has utilities not only in identifying of diagnostic biomarkers, but also may enable us to develop a novel targeted therapy utilizing the peptide sequences (mimotopes) as vehicles to deliver specific drugs. For instance, among highly significant clones, we found a sequence peptide homologous to histidine kinase (narX) with high specificity and sensitivity. Bacterial histidine kinases are promising targets for the development of antibacterial therapy. Currently efforts have been made to identify specific compounds targeting the inhibition of histidine kinase as antibacterial
therapy [10]. Additionally, this technology might enable us to discover unknown epitopes targeting specific bacterial antigens leading to immunogenicity and antibody production in CF subjects, as well as providing us with a better understanding of host immune defenses in CF subjects. Furthermore, this microarray platform can be hybridized to detect IgA in sera or saliva of CF patients that may have clinical values.

In summary, we have developed a novel T7 phage display library derived from BALs and leukocytes of patients with sarcoidosis that displays a significant segment of the potential antigens that can recognize IgG antibodies in CF sera with high accuracy. Furthermore, we have identified a set of CF clones that highly correlate with clinical measures such as, sweat chloride values, BMI and FEV1. Microarray and immunoscreening has a value in clinical practice in antibody detection as it is non-invasive and requiring a minimal amount of blood. The identified sequences can be used to develop peptide/protein-coated magnetic nanoparticles for clinical testing or for applications in drug delivery [142]. The present study describes a novel approach to identify CF biomarkers. Further studies with a larger cohort group of patients and/or longitudinal studies are needed to investigate the role of these antigens in CF, their mechanism of action and their utilities in drug design and monitoring of therapy.
CHAPTER 3 IDENTIFYING BIOLOGICALLY RELEVANT PUTATIVE MECHANISMS IN A GIVEN PHENOTYPE COMPARISON

3.1 Introduction

Identifying the mechanism involved in a particular phenotype is an essential step in understanding the biological phenomena that lead to the phenotype. Mechanism identification became more and more feasible with the availability of high-throughput biological data, which makes it possible to measure the expression of thousands of genes at once, and with the expanding knowledge of interactions between biological entities such as genes and proteins.

In biology, the concept of mechanism has several meanings [126]. In this chapter, we are interested in the causal mechanism which is defined as “a step-by-step explanation of the mode of operation of a causal process that gives rise to a phenomenon of interest.” [126]. Henceforth, we will use the term mechanism to refer to the causal mechanism. We also use the qualifier “putative” because the mechanisms we identify here are not mechanistically proven, but rather proposed mechanisms compatible with all gene expression changes measured throughout the system.

Many of the existing approaches to identify biological mechanisms focus on the selection of differentially expressed (DE) genes. One basic approach is to consider the expression fold change between two groups (i.e. disease vs. healthy). Typically an arbitrary threshold on such fold change (FC) is applied, and a gene will be considered DE when its absolute FC value is greater than the chosen threshold [33, 76, 176]. Other approaches use genes expression to calculate a p-value for each gene representing the probability of obtaining the observed expression change just by chance. This p-value is then used alone or
in conjunction with the FC value to determine if a gene is DE or not. The results obtained by these approaches are very sensitive to the selection of the FCs and p-values thresholds [131]. An important question here is related to the number of biological replicates needed and which RNA-seq differential expression tool should be used in order to identify the DE genes. Schurch and others [157] evaluated 11 tools on an RNA-seq experiment data with 48 biological replicates [109, 43, 65, 178, 182, 5, 105, 94, 151, 92, 173, 97, 96]. The false discovery rate was $\lesssim 5\%$ for 9 out of the 11 tools for all numbers of replicates. Regardless of the tools used, these approaches only provide a set of DE genes. For most purposes, identifying a set of DE genes is useful but is far from providing an understanding of the underlying phenomena [39].

Several methods were developed to provide an understanding of the underlying biological mechanisms by building regulatory networks from scratch from gene expression data, [55, 17, 110, 194, 66] or by integrating gene expression with other molecular data types such as transcription binding sites, and protein-protein interaction (PPI) data [185, 195, 103]. These methods have the advantage of being able to go beyond the existing known pathways and discover completely new mechanisms. On the other hand, these methods are burdened by the need to re-discover and re-create every time networks corresponding to well understood phenomena, already described by existing gene signaling networks such as those contained in KEGG [78, 79], Reactome [28], and Pathway Commons [21].

This was the motivation behind developing a plethora of pathways analysis methods [83, 115, 85, 59, 13, 174, 36, 106, 125, 35, 77]. These methods avoid the burden of having to re-discover well understood mechanisms by using the existing databases of
known pathways as the starting point and attempting to identify the mechanisms under-
ly the given phenotype by identifying which of the known pathways are significantly
impacted by the given phenotype [115]. These methods are very valuable but many times,
they cannot provide a full understanding of the underlying phenomena due to the sheer
size of some of the existing pathways. For instance, many KEGG pathways [79, 78] can
have well over one hundred genes (e.g. “Pathways in cancer” has 397 genes/nodes and
1107 edges/interactions). If one or more of these large pathways are identified as signifi-
cantly impacted in a given phenotype, the researchers are still at a loss as to which part of
such a large pathway is the real key to the given phenomenon. To relate to the previous ex-
ample, most experiments comparing cancer samples vs. normal tissue are likely to identify
the human “pathways in cancer” (hsa05200) as being significant without really helping
the understanding of the mechanism underlying the particular type of cancer studied in
that experiment.

In this chapter we describe a method able to identify which specific gene-gene in-
teractions and signals from an existing pathway may constitute the putative mechanism
associated with a phenotype of interest. The approach presented here takes into account
both the classical factors, i.e. the expression changes and statistical significance of genes,
as well as the knowledge captured by all existing known pathways. In section Materials
and Methods (3.2) we describe the method in detail, and in section Results (3.3) we
present the results of the application of the method on several datasets. The results of
the proposed method show it is capable of identifying the correct mechanism in all three
mouse datasets included in the chapter. We compare the results of our method with the
results of the classical approach that identifies mechanisms based only on the genes’ fold
changes and p-values, showing that our method goes beyond what can be obtained with the existing methods. The gene list generated with the proposed method ranked the KEGG pathways with the KO gene similarly or higher than the gene list of the classical approach when combined with the methods including ORA [84, 38], SPIA [174], and LEGO [36]. We applied the three methods ORA, SPIA, and LEGO, to show that the comparison with the classical approach is not favoring a particular pathway analysis method. We also show that the proposed approach is less sensitive to the choice of parameters compared with the classical methods of using FCs and p-values.

3.2 Materials and Methods

Here we are proposing a method, henceforth referred to as HighEdgeS, that is able to identify putative mechanisms in phenotype comparison experiments by incorporating gene expression values and previously known interactions among genes. This is achieved in the following four steps: First, the method builds a global graph that consists of all known interactions among genes. Here, we used interaction knowledge from the KEGG pathway database, which was downloaded on Dec 2016, although other pathway databases can be used. The second step is to assign scores to all edges in the global graph by taking into account the measured expression change of the genes connected to each edge, as well as the statistical significance of such change. To calculate gene expression measurements, we used custom chip definition files (CDFs) for the preprocessing of the microarray data [155]. Given two genes A and B, connected by an edge in the global graph, let us consider $FC_A$ and $FC_B$ as the expression changes of A and B, respectively, and $p_A$, $p_B$ the probability of observing $FC_A$, $FC_B$ just by chance. For each dataset, the FC value ($FC_A$) for gene A was
calculated by comparing the expression values in KO samples vs. normal samples. The p-value for gene A ($p_A$) was also calculated between the same groups (i.e. KO samples vs. normal samples) using a t-test. The edge between A and B will have a score calculated as follows:

$$EdgeScore_{AB} = |FC_A| \cdot (1 - p_A) + |FC_B| \cdot (1 - p_B) \quad (3.1)$$

After computing the score for each edge, the third step is to determine the “most important” edges. Choosing an arbitrary threshold on the edge scores would have led to the same problem of sensitivity to parameters of the classical approaches. Therefore, we use a change point analysis [86] on the distribution of edge scores to automatically determine the critical edge score threshold based on the observed data. Intuitively, the change point can be thought of as the inflection point of the distribution of the edge scores after which the distribution becomes flat. An example is shown in Fig. 15b.

Since we observed that the threshold determined by the change point analysis may still include a large number of edges, which may introduce some false positives, we decided to consider only the top 75% of the edges in this interval, thus leaving a safety margin. The 75% was chosen in order to avoid selecting a large number of genes, based on the typical distribution of the edge scores. No changes to this threshold were made after the initial choice. The interval of edge scores used to select edges is shown by the red arrow in Fig. 15b. In the final step of this approach a graph representing the putative mechanism related to the phenotype is constructed from the high-score edges selected as above. The final result of the proposed analysis method is a graph with the following properties: a) many (but not all) nodes are genes with high fold changes and/or significant p-values,
Figure 15: **The workflow of the proposed method, HighEdgeS.** a) The global graph constructed from all interactions present in all KEGG pathways; b) The edge score histogram constructed from the input data. A change point analysis [86] is used to determine the beginning of the flat area of the curve. The selected edges will be those in the top 75% of the remaining scores. c) The global graph showing the edges with high scores in red; d) The subgraph with the high scoring edges only representing the putative mechanism involved in the given phenotype.
and b) each edge corresponds to a known biological signal (the initial graph is constructed using all known interactions from KEGG). The value here is identifying the small subset of interactions from the larger subsets of interactions described by curate pathways. Notably, the graphs constructed by this approach are generally different from the subgraphs that could be obtained based on fold-change alone, p-values alone or fold changes and p-values together.

The main problem when proposing a new analysis method is posed by the lack of gold standards for validation. In this case, our goal is to investigate whether the proposed analysis method is able to correctly identify the mechanism(s) involved in a given phenotype. In order to perform this type of validation we chose three datasets comparing a wild-type with a phenotype derived by knocking out (KO) a single gene. A knock-out dataset is a dataset where a gene is intentionally disrupted, inactivating it completely (or nearly so). In all these situations, we know the true cause of the phenomenon (the KO of the target gene), as well as the initial mechanism that causes the phenotype (the genes that are immediately connected to the KO gene).

A good analysis method should identify the KO gene as central in the resulting mechanism, along with the interactions and processes downstream of the KO gene. However, since our approach uses an edge analysis, it may be seen more likely to yield networks of genes, which may be seen as an unfair advantage with respect to the classical approach that selects differentially expressed (DE) genes regardless of their connectivity. In order to level the field and ensure an unbiased comparison, we also compared the results of the two approaches using a classical pathway analysis that aims to identify the pathways that are implicated in a given phenotype.
We applied HighEdgeS, as well as the classical approach on three knock-out (KO) datasets. The KO genes in these datasets were Myd88, NeuroD1, and Pxd1. The GEO[9] accession numbers for these datasets are GSE22873, GSE6030, and GSE29048, respectively. We compared the proposed approach with the classical approach by considering the entire list of genes provided by each method and using them to perform a pathway analysis, more specifically the over-representation analysis (ORA) [84, 38], SPIA [174], and LEGO [36]. We also applied GAGE [106] on all genes of the experiment. In order to also compare the results of our method with a method that is not dependent on threshold selection, we used GAGE [106]. GAGE also ranks gene sets similarly to ORA, SPIA and LEGO, but its input is the entire set of genes in a dataset (i.e. not only the DE genes). In summary, the input to GAGE is: i) the entire set of genes in an experiment and ii) a list of gene sets (pathways with gene names only no interactions). GAGE’s output is one list of ranked gene sets/pathways. The results from GAGE were compared with the results obtained with the threshold automatically calculated by the proposed method. In both cases, we can calculate a false positive rate (FPR) and a true positive rate (TPR) for each data sets. For the methods SPIA and LEGO we used the R implementation versions 2.22.0, 1.0 and 2.20.1 respectively. In essence, we ask which of the known pathways are significantly enriched in the genes found by each of the two approaches (i.e. HighEdgeS and the classical). We chose to perform this comparison because, in principle, even a result in which the KO gene is not reported as relevant, but all the other genes related to the phenomenon are, should be considered a meaningful result. Pathways were considered significant when their FDR corrected p-values were < 0.1.

The assessment was performed considering that the pathways where the KO gene was
present were truly causal for the phenotype and should be reported as significant (positives), whereas pathways without the KO gene should not be significant (negatives). Given this assumption, we can compute the number of True Positives, True Negatives, False Positives, and False Negatives obtained by analyzing each list of genes.

For HighEdgeS the list of relevant genes was the list of genes connected to at least one of the high-score edges, while for the classical approach it was the list of genes identified as DE by the choice of thresholds. In each case, we calculated the True Positive Rate (TPR) and False Positive Rate (FPR). Since the results of the classical approach are sensitive to the chosen thresholds, we calculated these measures for an entire range of thresholds. The positive likelihood ratio was computed across this entire range to determine the threshold that produced the best result within each dataset. For each dataset we considered the threshold that produces the best positive likelihood ratio in order to present the classical approach in the best possible light. In reality, using the fold change and p-value approach will always produce results inferior to the ones reported here because the optimal threshold cannot be known in advance for any given dataset.

Thresholds were generated as follows: for the classical approach the range starts from $\log_2|FC| = 0.5$ and $-\log_{10}(p) = 0.5$ (least stringent), to $\max(\log_2|FC|)$ and $-\log_{10}(p) = 5$ (most stringent) for each dataset. Note that the classical approach uses two variables: FC and p-value and requires an arbitrary threshold for each. In contrast, HighEdgeS does not require the user to choose any thresholds.

At each interval in the threshold range of HighEdgeS, the method was applied and the genes that were connected to the high score edges were used to rank pathways using ORA, SPIA and LEGO. The TPR and the FPR were calculated each time. A similar process
was applied for the classical approach using the somewhat arbitrary, but commonly used, combinations of fold-change and p-value thresholds.

It is important to note that HighEdgeS returns a network, while the classical approach returns a set of genes. In order to make the results comparable of the two approaches: we built a network by retrieving all the interactions from KEGG connecting any pair of genes resulting from the classical approach. The comparison highlights the fact that in spite of the fact that both approaches use the same type of information (fold changes, p-values and known gene signals from KEGG), HighEdgeS is able to produce a sequence of interactions as the putative mechanism for the process involved in the phenotype, whereas the classical approach yields mostly disconnected genes. Note that the same values of $p_A$, $p_B$, $FC_A$, and $FC_B$ were used in both methods (i.e. the classical approach and HighEdgeS). Finally, because we used KEGG as a reference, HighEdgeS can only detect edges, and therefore genes, that are present in KEGG. In order to keep the comparison as fair as possible, for the classical approach we used only genes that belong to at least one KEGG pathway, as it would be unfair to compare connectivity if we were to use also genes that are disconnected because they do not belong to any KEGG pathway.

3.3 Results

3.3.1 Myd88 knock-out

In this Myd88 knock-out dataset, the authors were interested in the effect of knocking out the Myd88 gene in the liver [192], and the involvement of Myd88 in an inflammatory response. The authors of the dataset reported that Myd88 is essential for Tnf survival mechanisms [192]. Further, they also reported the activation of NfkB (aka Rela, shown in
Fig. 16a) by Myd88. From this dataset, HighEdgeS constructed a graph consisting of 35 edges and 36 genes where the KO gene plays a central role, as it can be seen in Fig. 16a. In this result, HighEdgeS found many connected related genes from the same family, such as Cxcl, Tlr, and Tab. This graph shows the effect of Myd88 on Rela (Nfkb) and on Tnf, the same interactions reported in the experiment that produced the dataset.

The classical approach identified 20 edges and 44 genes. Although Myd88 was among them, the graph generated from the 42 genes, shown in Fig. 16b, shows that most genes are not connected. Thus the classical results cannot really be used as a proposed putative mechanism. Among the genes identified by the classical approach, only the gene Cxcl1 was connected to the KO gene. Furthermore, the classical approach failed to identify the gene Rela, which was part of the known mechanism described by the authors of the experiment. Although the graph built using the results of the classical approach contains several interactions (shown in the top left corner of Fig. 16b), none of the genes were connected to the known cause of phenomenon, the Myd88 gene.

When we applied the methods ORA, SPIA, and LEGO to the list of genes obtained using the best choice of parameters from the classical approach, we obtained a TPR value of 0.31 and an FPR of 0.08 from ORA, a TPR value of 0.31 and an FPR of 0.01 from SPIA, and a TPR value of 0.93 and an FPR of 0.04 from LEGO. In contrast, while HighEdgeS obtained a TPR value of 1 and a FPR value of 0.13 when ORA was applied, a TPR value of 1 and a FPR of 0.10 from SPIA, and a TPR value of 1 and a FPR of 0.19 from LEGO (see Fig. 17). The TPR yielded by GAGE was 1, and the FPR was 0.26. The TPR of GAGE is similar to the results of HighEdgeS. This shows that applying the methods ORA, SPIA, and LEGO on the mechanism found by HighEdgeS retrieves every single pathway truly involved with
Figure 16: Figure shows the results for the dataset Myd88. The KO gene is shown in red in both panels. a) Mechanism results for HighEdgeS when we applied it on the Myd88 dataset. The results show two subgraphs. The one with the most genes shows the KO gene regulates the Cxcl1 gene and shows that the subgraph includes many Cxcl1 downstream genes. b) The results of the classical approach using DE genes. The Myd88 gene is connected to only one downstream gene. The result of the classical approach has many genes that are not connected indicating that the classical approach missed important interactions.
the phenotype (i.e. containing the KO gene), while yielding a number of false positives which is comparable with the classical analysis. The TPR value of 1 for HighEdgeS from this dataset is similar to the TPR from GAGE. However, the FPR values for the proposed method from ORA, SPIA and LEGO are better than FPR value of GAGE.

3.3.2 NeuroD1 knock-out

In this experiment, the authors extracted mRNA from the mouse pineal gland, and then confirmed the KO using qRT-PCR [120]. NeuroD1 mRNA is highly abundant in the adult rat pineal gland and has been found to be involved with the regulation of insulin [120].

When applied to this dataset, HighEdgeS constructs a network of 6 genes connected with 5 edges, shown in Fig. 18a. In this network, the NeuroD1 gene is among the genes of the proposed putative mechanism, and has four genes immediately downstream of it: Ins1, Ins2, Iapp, and Gck. The presence of the Ins1 and Ins2 genes in this mechanism confirms the regulatory action of NeuroD1 on those genes. Studies demonstrated the effects of insulin on melatonin synthesis in the pineal gland through crosstalk between noradrenergic and insulin pathways [133]. This crosstalk also involves the gene Gck, involved in most glucose metabolism pathways [60].

When we applied the classical approach to the same dataset, we obtained a graph consisting of 22 genes and one edge, shown in Fig. 18b. These results include a relatively small number of genes, yet when we inspect the graph constructed from these genes, we notice that the KO gene is not connected to any other gene. Thus, in this dataset as well, the classical approach is unable to yield any plausible mechanism for the given phenotype. Although the KO gene is shown in the result, this graph fails to propose any reasonable explanation on how the KO of NeuroD1 influenced the rest of the system.
Figure 17: **Comparison between the results from HighEdgeS (left columns) and the classical approach (right columns) for the Myd88 dataset.** The three rows show the results of the pathway analysis methods ORA, SPIA, and LEGO. For each barplot, the x-axes show various values of the thresholds for each method. For the classical approach, the x-axes show various combinations of fold change and p-value thresholds. For the proposed approach the x-axes show the range from the point determined by change point analysis to the maximum value of edge scores. The y-axes in each graph show the scales for the false positive rate (FPR) and the true positive rate (TPR). The right y-axes in each graph represent the number of DE genes shown by the gray bars. The blue dashed line represents the change point safety margin from edge point analysis (the default for the proposed method). In each barplot, the green bars represent the TPR and red bars represent the FPR for pathways ranking when the significance threshold $\alpha = 0.1$. True positive pathways are those containing the KO gene. The proposed method yields a perfect true positive rate of 100% in every case for its default threshold (blue dashed line). In contrast, the classical approach yields a TPR of less than 40% for all threshold combinations used with ORA (top right panel) and SPIA (middle right panel). The classical approach used with LEGO (lower right panel) yields a TPR varying between 0 and 93%. The figure also shows how the results of the classical approach depend very much on the combination of thresholds used for fold change and p-values, while the results obtained with the proposed method are much more stable.
Figure 18: **Figure shows the results for the dataset NeuroD1.** The KO gene is shown in red in both panels. a) The mechanism found using HighEdgeS shows NeuroD1 is regulating Ins2, Iapp, Gck and Ins1. b) The results were obtained using the classical approach for the same dataset. The classical approach results include the KO gene, but provides no explanation on how the suppression of this gene propagates further and affects the rest of the system.
When we applied ORA, SPIA, and LEGO to the networks obtained by the two approaches, the results of HighEdges yielded a TPR of 1 and a FPR of 0.08 from ORA, a TPR of 1 and FPR of 0.12 for SPIA, and TPR of 1 and FPR of zero for LEGO. The results of the classical approach yielded a TPR of zero and a FPR of zero for ORA, a TPR of 1 and FPR of 0 from SPIA, and TPR of 1 and FPR of 0 from LEGO. When we applied GAGE on the Neurod1 dataset, it yielded a TPR of 0, and a FPR of 0.04.

This means that no pathway involved with the phenotype was reported as significant when analyzing with the list of genes obtained from the classical approach, when ORA was applied yet with SPIA and LEGO the TPR was one. At the same time, the analysis of SPIA, LEGO, and ORA the enrichment analysis of the results yielded by HighEdgeS is able to identify as significant the only pathway truly involved with the phenotype (see Fig. 19). The TPR value of 1 for HighEdgeS in combination with any of the ORA, SPIA and LEGO is substantially better than the TPR from GAGE. In essence, the proposed method is able to correctly find the only pathway containing the knockout gene regardless of the method used to identify the significant pathways. At the same time, GAGE fails to identify the causal pathway in this data set.

3.3.3 Pdx1 knock-out

In this dataset the authors were interested in the effect of the knock-out of the transcription factor Pdx1 and its role in the duodenum [24]. When we applied HighEdgeS to analyzed this dataset, it produced a connected graph with 12 genes and 11 edges. When applied to this dataset, the classical approach reported 18 DE genes, including the KO gene. However, no known interaction was found among these genes, as shown in Fig. 20b, showing that the classical approach failed again to propose a meaningful mechanism able
Figure 19: Comparison between the results from HighEdgeS (left columns) and the classical approach (right columns) for the Neurod1 dataset. The three rows show the results of the pathway analysis methods ORA, SPIA, and LEGO. For each barplot, the x-axes show various values of the thresholds for each method. For the classical approach, the x-axes show various combinations of fold change and p-value thresholds. For the proposed approach the x-axes show the range from the point determined by change point analysis to the maximum value of edge scores. The y-axes in each graph show the scales for the false positive rate (FPR) and the true positive rate (TPR). The right y-axes in each graph represent the number of DE genes shown by the gray bars. The blue dashed line represents the change point safety margin from edge point analysis (the default for the proposed method). In each bar plot, the green bars represent the TPR and red bars represent the FPR for pathways ranking when the significance threshold $\alpha = 0.1$. True Positive pathways are those containing the KO gene. The classical approach yields a TPR of 0 for all threshold combinations used with ORA (top right panel). For SPIA and LEGO, the classical approach yields 0 true positives for the more lenient thresholds, to the left of the x axis. For the more stringent thresholds, in this dataset there is only one DE gene left, the KO gene, which in turn identifies the true positive pathway. As explained in the text, the correct threshold is never known a priori which means that the classical approach may or may not identify the correct pathway, depending on the choice of the threshold. In contrast, the proposed method yields a perfect true positive rate of 100% in every case for its default threshold that is calculated automatically (blue dashed line).
to explain the given phenotype.

When ORA was applied on the list of genes produced by HighEdgeS, yielded a TPR of 1 and an FPR of 0.14, a TPR of 1 and FPR of 0.25 from SPIA, and a TPR of 0.66 and FPR of 0.17 from LEGO. The list of genes produced by the classical approach yielded a TPR of zero and an FPR of 0.01 or ORA, a TPR of 1 and FPR of zero from SPIA, and a TPR of zero and FPR of zero from LEGO (see Fig. 21). The TPR yielded by GAGE was 0.66, and the FPR was 0.15.

In this case, although the classical approach correctly found the KO gene among the DE genes, the presence of other genes, not relevant to the phenomenon, negatively impacted the results of ORA, de facto hiding the real phenomenon. Interestingly, HighEdgeS produced the same result independently of the value used for the safety margin, indicating that the method is very stable with respect to the choice of the parameters. The results of SPIA on the list of genes from the classical approach were perfect (no false positive) and identified all true pathways. When combined with ORA and SPIA, HighEdgeS yielded a perfect TRP of 1, identifying all 3 pathways containing the knockout gene. When combined with LEGO, the proposed method identified 2 of the 3 causal pathways, GAGE only identified 2 of the 3 causal pathways, as well.

3.4 Discussion

Identifying the mechanisms involved in a disease is an important step toward understanding diseases and developing effective treatments. Many approaches that aim to identify these mechanisms use a selection based on thresholds on FC or p-value to find genes of interest in the given phenotype. These approaches ignore the interactions between genes,
Figure 20: The comparison between the putative mechanisms constructed by High-EdgeS and the classical approach for the Pdx1 dataset. a.) The putative mechanism constructed by HighEdgeS. The graph shows the knocked out gene Pdx1 regulating 5 downstream genes. b.) The results for the same dataset when the classical approach was applied. The results of the classical approach lack any connection among the genes and the KO. However, the results of the proposed approach show the interaction between the Pdx1 gene and the Ins1. In fact, the authors of the dataset discussed this interaction in their work indicating the role of the Ins1 in this condition.
Figure 21: **Comparison between the results from HighEdgeS (left columns) and the classical approach (right columns) for the Pdx1 dataset.** The three rows show the results of the pathway analysis methods ORA, SPIA, and LEGO. For each barplot, the x-axes show various values of the thresholds for each method. For the classical approach, the x-axes show various combinations of fold change and p-value thresholds. For the proposed approach the x-axes show the range from the point determined by change point analysis to the maximum value of edge scores. The y-axes in each graph show the scales for the false positive rate (FPR) and the true positive rate (TPR). The right y-axes in each graph represent the number of DE genes shown by the gray bars. The blue dashed line represents the change point safety margin from edge point analysis (the default for the proposed method). In each bar plot, the green bars represent the TPR and red bars represent the FPR for pathways ranking when the significance threshold $\alpha = 0.1$. True Positive pathways are those containing the KO gene. The figure shows that the proposed approach yields a higher TPR than the classical approaches overall. For ORA, the classical approach fails to identify any of the 3 pathways containing the KO gene independently of the threshold combination used (top right panel). The classical approach used with SPIA identifies 1, 2 or 3 of the 3 true positive pathways depending on the threshold combination used (middle right panel). Finally, the classical approach used with LEGO identifies only one of the 3 true positive pathways and that only for about $1/3$ of the range of thresholds explored. In contrast, the proposed approach used with ORA (top left) and SPIA (middle left) identifies all true positive pathways for any edge scores. When combined with LEGO (bottom left), the proposed approach identifies 2 out of the 3 true positives, still much better than the classical approach with LEGO (bottom right).
and they are very sensitive to threshold selection [44, 45]. Here we described a method that overcomes these limitations, and produces results that are more biologically relevant and less sensitive to the choice of parameters, when compared to the results of the classical approach.

The classical approach uses two different measures, fold change and p-values.

The fold changes are often considered on a log scale and the p-values are usually corrected for multiple comparisons with methods such as FDR. However, the user is still required to make two choices: for the log fold change and for the FDR-corrected p-value. As Figs. 17, 19, and 21 show, the results can vary dramatically depending on the choices made for these parameters. The proposed method has a single parameter, the edge scores, and the approach calculates the best value for this threshold in each case. Thus, the user does not need to make any choices whatsoever. The results in Figs. 17, 19, and 21 (blue line in each left panel) show this calculated threshold provides better results than most of the choices based on fold change and p-values in the classical approach.

In order to compare the performance of our method with the performance of the classical approach, we analyzed three datasets where the underlying cause of the given phenotype was known. The proposed method constructed mechanisms having the KO gene in a central role in all datasets. Since the goal of our method is to identify the mechanism in a phenotype comparison, the only way in which such a method can be truly validated is to analyze data from experiments satisfying two conditions: i) the cause of the phenotype is known, and ii) the cause of the phenotype is related to a single gene, rather than a combination of genes, or gene-environment interaction(s). This is why we only looked for knockout (KO) datasets. Furthermore, the proposed method is using gene-gene
interactions from the signaling pathways of the KEGG database. Thus, to evaluate our method with any dataset, the dataset must come from a KO experiment and the knocked-out gene must be connected with other gene(s) in the signaling pathways of KEGG. At this time, mouse is the only organism that has data from single cause experiments, that are of sufficient quality, and that involve sufficiently well annotated genes to allow us to prove the mechanisms found are correct. Even though the validity of the method could only be demonstrated on mouse data, there is nothing specific to mouse in the analytical approach used and the method can be applied to any organism, as well as it is reasonably well annotated.

In all three datasets illustrated here, the proposed method was able to identify interactions that both: i) were known to be related to the phenomenon, and ii) constitute a plausible explanation for how the effect of the KO gene was propagated in the rest of the system. At the same time, the classical approach mostly produced disconnected graphs that were not helpful to explain other changes in the organism. A quick look at the results of HighEdgeS might suggest that it was simply reporting all the genes connected with a gene of a large fold change (FC). This is not the case. In the Myd88 dataset, the Myd88 gene is directly connected to 35 downstream genes and 13 upstream genes in the KEGG global graph. In the results of the Myd88 dataset HighEdgeS reported only 10 downstream genes and only 2 upstream genes as shown in Fig. 16a. This is a concrete example demonstrating that the proposed method does not report genes just because they are connected with a gene of a large FC.

Table 6 summarizes the comparison between the results of HighEdgeS, the classical approach, and GAGE as they would be used in practice. Three pathway analysis methods
(ORA, SPIA, and LEGO) were applied on the genes yielded by HighEdgeS and the classical approach. GAGE uses the entire set of genes to directly identify pathways so ORA, SPIA and LEGO are not comparable to GAGE. For the classical approach we selected the DE genes using an absolute fold change greater than 2 and FDR-corrected p-value less than 0.05. For the Myd88 and Neurod1 datasets, no genes meet these thresholds showing the limitations of this approach. The performance was measured using the true positive rate (TPR) and false positive rate (FPR) for each data set. The results show that HighEdgeS yields the best TPR for all datasets. HighEdgeS also yields the best FPR for 2 out of the 3 datasets. It is important to note that a perfect TPR (or sensitivity) can always be obtained if the results include all pathways. This, of course, will come with a very poor FPR since all but the true positives pathways will be false positives. Conversely, a perfect FPR (or specificity) can always be obtained if the results do not include any pathway. This however, will be associated with a TPR of zero since no true positives will be included in the results. This is precisely what happens for the classical approach for the Pdx1 dataset. The classical approach has a lower FPR values, only because its sensitivity is very low (not being able to identify any true positive (TP) pathway in 2 out of 3 cases, and identifying only one of the 3 true positive pathways in the remaining case). In contrast, HighEdgeS identified all true positive pathways in 2 out of 3 cases (when combined with ORA and SPIA) and 2 out of the 3 TP pathways when combined with LEGO.

It is important to note that some of the existing pathway annotations may have some annotation bias. For instance, the biochemical pathways in KEGG are acknowledged to have some bias toward metabolism[127]. Also, the degree to which a given genome is annotated in various databases including KEGG is another important potential limitation
Table 6: A comparison between the results of HighEdgeS, the classical approach, and GAGE as they would be used in practice. Three pathway analysis methods (ORA, SPIA, and LEGO) were applied on the genes yielded by HighEdgeS and the classical approach. GAGE uses the entire set of genes to directly identify pathways so ORA, SPIA and LEGO are not applicable for GAGE. The green background indicates the best results obtained for each dataset (each row). For the classical approach we selected the DE genes using an absolute fold change greater than 2 and FDR-corrected p-value less than 0.05. “No DE” means no genes met those thresholds. The performance was measured using the true positive rate (TPR) and false positive rate (FPR) for each data set. The results show that HighEdgeS yields the best TPR for all datasets. HighEdgeS also yield the best FPR for 2 out of the 3 datasets. Furthermore, for the Pdx1 dataset, even though the classical approach has a lower FPR values, its sensitivity is very low (not being able to identify any true positive (TP) pathway in 2 out of 3 cases, and identifying only one of the 3 true positive pathway in the remaining case). In contrast, HighEdgeS identified all true positive pathways in 2 out of 3 cases (when combined with ORA and SPIA) and 2 out of the 3 TP pathways when combined with LEGO.
of this approach. More specifically, since this approach relies on existing annotations, the proportion of well studied genes in a given genome is important. Better results are more likely to be obtained for data from human or other common model organisms such as mouse, rat, drosophila, etc. Since the interactions that we are using to get the results of our method are limited to the genes in KEGG, this might be a limitation to this approach. Nevertheless, annotation bias will affect any type of analysis approach. We expect that the results provided by the method proposed here will become more and more accurate as the sources of various annotations - including KEGG - become more accurate themselves and start to reduce some of their existing intrinsic bias.
CHAPTER 4  POPULATION CLASSIFICATION USING SNP DATA

4.1  Introduction

Classifying the population of an individual by using DNA information is an important problem in forensic evidence and evolutionary biology [136, 82, 30, 154, 75, 61]. Population classification using biological samples became more and more feasible with the availability and the advances of high-throughput biological data.

Currently, it is possible to measure thousands of single nucleotide polymorphisms (SNPs), and short tandem repeat (STR). Many approaches have been developed to identify samples from different ethnic groups by using various genomic data such as SNPs [139, 3] and STRs [154]. SNP data have been shown to overcome the technical problems of STR profiling [82], and are known to be the best markers for the use in forensics [130]. Classification approaches that can identify the population that an individual belongs to based on genomic data are already available. Across the entire genome, a set of 45 unlinked autosomal SNPs were identified as suitable for universal human identification [130]. An alternative multiplex assay with 52 SNPs is also available for human identification [154].

Creating such an assay with 50-100 SNPs that are validated for forensic analysis can be difficult, and may require far more DNA sample than is needed for STR systems [16]. The ForenSeq Kintelligence kit targets 10,230 SNP markers and was designed specifically for forensic genetic genealogy [53]. Another kit that was developed by Verogen is the ForenSeq DNA Signature Prep Kit which uses 56 SNPs that are not commonly used in current kits [52]. Applied Biosystems developed a set of PCR amplicons called Precision ID Ancestry Panel using 165 autosomal markers (SNPs) [6].
There are three major problems that are facing this industry: i) sample contamination, ii) DNA quality, and iii) mixture of more than one person’s genome. In the area of forensic science, genotyping failures when working with fragmented DNA from degraded samples is commonly encountered when the samples come from crime scenes. This is because current SNP microarrays require large amounts of genomic DNA, which must not be heavily fragmented. These requirements are often not satisfied by DNA samples found at crime scenes [82]. Many strategies were attempted to repair DNA degradation [137, 32]. However, current repair strategies are limited and require further investigations [2]. Furthermore, sometimes DNA samples may include only a small fragment of DNA e.g. from a single chromosome. In such situations, the methods above could not be used.

Several methods were developed to analyze genetic data. Rannala et al., presented a statistical method that used multilocus genotype data, and a Bayesian approach to identify individuals who are immigrants or have recent immigrant ancestry in a population [144]. Yang et al., developed the BIASLESS method which is an analysis able to distinguish between samples from closely related ethnic populations [190]. The clustering method STRUCTURE [136] is able to assign individuals to a population and to infer the population structure. Other extensions of the STRUCTURE method were published in [48, 49, 70]. The software fastSTRUCTURE [139] is the latest extension of the STRUCTURE method and is two order of magnitude faster than the original STRUCTURE.

In general, clustering methods are highly sensitive to the number selected clusters [135, 159]. Selecting more than the ideal number of clusters will result in some samples being assigned to a new cluster, and selecting too few will result in samples from two cluster mixing into one. EIGENSTRAT was proposed to address some of the computational cost
limitation of STRUCTURE, by correcting for population stratification [135]. Other methods such as FRAPPE [172] and ADMIXTURE [3] were developed to estimate individual admixtures.

There are thousands of criminal cases that could be resolved by using DNA samples with advanced forensic methods (The DNA Doe Project dnadoeproject.org). While there are many methods available to unravel human evolutionary history from genomic data [156], not many methods are available to assign a sample from an individual to one of several predefined populations [30]. Among those focused on forensics, Kayser and de Knijff discussed several methods that improve forensic science through advances in genomics [82]. For example, they discussed the application of DNA-based prediction of human eye color and the genetic substructure of human populations. Few private companies are working with customers (e.g. law enforcement) to learn more about the identity of unknown persons who left traces of DNA at a crime scene. OTHRAM, Inc. is developing approaches to get quality sequencing methodology and building software algorithms to identify the genetic data for a person of interest that correspond to degraded samples. Furthermore, OTHRAM is working on developing models for sample prediction (similar to the one proposed in this chapter).

In this chapter, we are introducing a classifier that is able to analyze genomic data, more specifically SNP data, and assign an individual to a particular population/ethnic group. Random SNPs across the genome were used in this analysis. We show that the proposed analysis is less dependent on the selection of a set of SNPs which makes it better suited for high density genotyping assays. Because this approach does not require a specific panel of SNPs which may be spread out throughout the entire genome, this method
does overcome the limitation of the current approaches that require a large amount of DNA [16]. Another limitation of the current approaches is that the sample type needs to be known. For example, the ForenSeq Kintelligence kit requires the sample to be from hair, bone, teeth, blood, semen, or buccal swabs [53].

We show that the proposed method performs well, yielding an accuracy of 96%, when applied on only 10% of the genome. This is in contrast with the existing methods which require the availability of data from the whole genome. The details of the method are described in section 4.2. The proposed method is compared in section 4.3 with two popular classification algorithms: the naive Bayes classifier [114] and the Random Forest (RF) classifier [100]. In addition to these two popular methods, we also compare the proposed method with the BIASLESS method [190], which was developed specifically for the classification of individuals from different ethnic groups.

4.2 Methods

Here, we are proposing a method that is able to classify population samples using SNP data. The proposed method was trained on data from the Human Genome Diversity Project (HGDP) [75] from the CEPH website. Subsequently, the method was tested on completely independent data from the HapMap project [57], which is currently available on the NCBI website.

The training data have the following 7 groups of ethnic origin: \( g1 = \{\text{Africa, America, Central South Asia, East Asia, Europe, Middle East, Oceania}\} \) (see Table 7).

The data were stored in a set of matrices \( T_r \) for training (see Fig. 22). This set contains a matrix for each population. Each element in these matrices stored the observed genotype
Table 7: The samples counts of the training data by region and country.

<table>
<thead>
<tr>
<th>Region/Ethnicity</th>
<th>Count</th>
<th>Country</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>159</td>
<td>Angola</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Botswana Or Namibia</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Central African Republic</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Congo</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kenya</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lesotho</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Namibia</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nigeria</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Senegal</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>South Africa</td>
<td>2</td>
</tr>
<tr>
<td>America</td>
<td>31</td>
<td>Colombia</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mexico</td>
<td>24</td>
</tr>
<tr>
<td>Central South Asia</td>
<td>50</td>
<td>China*</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pakistan</td>
<td>40</td>
</tr>
<tr>
<td>East Asia</td>
<td>90</td>
<td>Cambodia</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>China**</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Japan</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Siberia</td>
<td>15</td>
</tr>
<tr>
<td>Europe</td>
<td>88</td>
<td>Europe</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>France</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Russia</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Russia-Caucasus</td>
<td>14</td>
</tr>
<tr>
<td>Middle East</td>
<td>146</td>
<td>Algeria-Mzab</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Israel-Carmel</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Israel-Central</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Israel-Negev</td>
<td>47</td>
</tr>
<tr>
<td>Oceania</td>
<td>33</td>
<td>Bougainville</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>New Guinea</td>
<td>16</td>
</tr>
</tbody>
</table>

Name of population: *Uygur *Yi, Daur, Mongola, Lahu, and Han

for a particular SNP (rows) in a particular sample (column).

The top left of Fig. 22 shows the 7 training matrices, and the bottom left of Fig.22 shows a sample $y$ from the testing data.

The location of SNPs were the variables used in the analysis. To be able to apply the analysis on the training and testing data, we selected the SNPs that were measured in both the training and the testing sets from chromosome one and had the same two observed nucleotides. We defined $SNP_{total}$ as the total number of SNPs used in the analysis.

Each SNP had 4 possible genotypes, two homozygous and two heterozygous. To avoid the ambiguity of the heterozygous genotype, we sorted the genotype in each SNP. For
Figure 22: The workflow of the method. On the training data $Tr$ we applied the equations 4.1, 4.2, and 4.3 to calculate the matrices $F$ for the 7 ethnic groups. Each row in $F$ stored the genotype frequencies for a particular SNP in a particular group. $P$ is the profile matrix for sample $y$. Each element $c$ in the profile matrix $P_y$ stored the frequency value from $F$ for each group that matched the observed genotype for sample $y$. The sum of the columns in $P$ is calculated and stored in the vector $S$. The class label of sample $y$ will be the name of the element with maximum value in $S$. 

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>SNP_1</td>
<td>c[1,1]</td>
<td>c[1,2]</td>
<td>c[1,3]</td>
<td>c[7684,1]</td>
<td>c[7684,2]</td>
<td>c[7684,3]</td>
<td>c[7684,4]</td>
</tr>
<tr>
<td>America</td>
<td>SNP_2</td>
<td>c[1,1]</td>
<td>c[1,2]</td>
<td>c[1,3]</td>
<td>c[7684,1]</td>
<td>c[7684,2]</td>
<td>c[7684,3]</td>
<td>c[7684,4]</td>
</tr>
<tr>
<td>Central South Asia</td>
<td>SNP_3</td>
<td>c[1,1]</td>
<td>c[1,2]</td>
<td>c[1,3]</td>
<td>c[7684,1]</td>
<td>c[7684,2]</td>
<td>c[7684,3]</td>
<td>c[7684,4]</td>
</tr>
<tr>
<td>East Asia</td>
<td>SNP_4</td>
<td>c[1,1]</td>
<td>c[1,2]</td>
<td>c[1,3]</td>
<td>c[7684,1]</td>
<td>c[7684,2]</td>
<td>c[7684,3]</td>
<td>c[7684,4]</td>
</tr>
<tr>
<td>Europe</td>
<td>SNP_5</td>
<td>c[1,1]</td>
<td>c[1,2]</td>
<td>c[1,3]</td>
<td>c[7684,1]</td>
<td>c[7684,2]</td>
<td>c[7684,3]</td>
<td>c[7684,4]</td>
</tr>
<tr>
<td>Middle East</td>
<td>SNP_6</td>
<td>c[1,1]</td>
<td>c[1,2]</td>
<td>c[1,3]</td>
<td>c[7684,1]</td>
<td>c[7684,2]</td>
<td>c[7684,3]</td>
<td>c[7684,4]</td>
</tr>
<tr>
<td>Oceania</td>
<td>SNP_7</td>
<td>c[1,1]</td>
<td>c[1,2]</td>
<td>c[1,3]</td>
<td>c[7684,1]</td>
<td>c[7684,2]</td>
<td>c[7684,3]</td>
<td>c[7684,4]</td>
</tr>
</tbody>
</table>

For a sample $y$ from the testing data, the algorithm of the proposed method selects the frequency values from all groups in $F$ that match the observed allele for that SNP.
instance, if the observed heterozygous genotype was “CA”, we sorted the letters, and “AC” was the actual value stored in the training (Tr) and the testing data; then we combined the two heterozygous genotypes into one. Thus, the analysis was performed with 3 genotypes (e.g. AA, AC, CC) instead of 4 (e.g. AA, AC, CA, CC).

We calculated the frequency matrices \( F \) for each group in \( g1 \) by using the equations: 4.1, 4.2, and 4.3. The frequency matrices \( F \) had \( SNP_{total} \) rows and three columns (one for each genotype). Each element in the frequency matrices stored the frequency of one of the three genotypes for each SNP in each group in \( g1 \) (see the top right of Fig.22).

\[
F[g1_j][SNP_i, homo1] = \frac{\text{number of times } SNP_i = \text{homo1}}{\text{len}} \tag{4.1}
\]

\[
F[g1_j][SNP_i, hetero] = \frac{\text{number of times } SNP_i = \text{hetero}}{\text{len}} \tag{4.2}
\]

\[
F[g1_j][SNP_i, homo2] = \frac{\text{number of times } SNP_i = \text{homo2}}{\text{len}} \tag{4.3}
\]

Where \( SNP_i \) is the SNP id at index \( i = 1,2,3,...,SNP_{total} \), \( g1_j \) is a group name at the index \( j = 1,2,3...|g1| \), \( \text{len} \) is the number of training samples in each group, and \( \text{homo1, hetero, and homo2} \) are the observed genotypes in the data (e.g. AA, AC, CC).

To predict the class label for a sample \( y \) from the testing data, we calculated the profile matrix \( P \) for each sample \( y \). We selected the matched frequency of the observed genotype for sample \( y \) from the frequency matrices \( F \) for each SNP in each group of \( g1 \) (see equation 4.4). Thus, each SNP in \( P \) (each row) has 7 values, one from each group (see the bottom right of Fig.22). In the matrix \( P \), the sum of each group was calculated, and the group name with the maximum sum was assigned as the predicted class label of sample \( y \).
\[
\begin{aligned}
P[SNP_i, g_{1j}] = & \begin{cases} 
F[g_{1j}][SNP_i, homo1] & \text{if } Ts[g_{1j}][SNP_i, y] = homo1 \\
F[g_{1j}][SNP_i, hetero] & \text{if } Ts[g_{1j}][SNP_i, y] = hetero \\
F[g_{1j}][SNP_i, homo2] & \text{if } Ts[g_{1j}][SNP_i, y] = homo2
\end{cases}
\end{aligned}
\] (4.4)

Where \(P\) is the profile matrix, \(SNP_i\) is the SNP id at index \(i = 1, 2, 3, \ldots, SNP_{total}\), \(g_{1j}\) is a group name at the index \(j = 1, 2, 3, \ldots |g_1|\), \(y\) is a sample in the testing matrices \(Ts\), and \(homo1, hetero,\) and \(homo2\) are the observed genotypes in the data (e.g. AA, AC, CC).

Note that the classification model was built on the training data which has 7 ethnic groups \(g_1\), and was tested on independent data which had only five ethnic groups \(g_2 = \{\text{Africa, America, Central South Asia, East Asia, Europe}\}\).

### 4.3 Results

The goal of any classification algorithm is to accurately predict the class label of a sample. To assess the performance of the proposed method, we compared its results with three popular classification methods i) the naive Bayes classifier ii) the Random Forest (RF) classifier, and iii) BIASLESS [190]. We used the naive Bayes classifier available in the \texttt{e1071} R package [114] and the Random Forest classifier available in the \texttt{randomForest} R package [100]. When we built the Random Forest classification models, we set the parameters \texttt{importance} and \texttt{proximity} to \texttt{TRUE}.

We applied the BIASLESS classifier available on the website of the package [189] on the training data. To get the BIASLESS model with the highest accuracy, we kept the parameters for cross-validation and marker selection to their default values.

To determine the performance of the four methods (including the proposed one) using different numbers of SNPs, we applied each of the four classification models on a different
number of randomly selected SNPs \( r = \{10, 50, 75, 100, 250, 500, 750, 1000, 2500, 5000, 7500, SNPs_{total}\} \). For the first index of \( r \), we randomly selected 10 SNPs without replacement out of the \( SNPs_{total} \); then we built the models of the four classifiers using the training data with only the 10 selected SNPs. The built models were then used to perform the classification on the testing data. Note that the last element in \( r \) has all SNPs i.e. the SNPs were not randomly selected in this particular case.

We used the SNPs for which the RefSNP genotypes were on the forward strand. The number of SNPs that are common in the training and the testing sets is \( SNPs_{total} = 7684 \). The training matrices \( (Tr) \) have a total of 597 samples distributed as follows: Africa \( (n=159) \), America \( (n=31) \), Central South Asia \( (n=50) \), East Asia \( (n=90) \), Europe \( (n=88) \), Middle East \( (n=146) \), and Oceania \( (n=33) \). The testing data include a total of 1,115 samples distributed between the ethnic groups as follows: Africa \( (n=488) \), America \( (n=71) \), Central South Asia \( (n=83) \), East Asia \( (n=234) \), and Europe \( (n=239) \) (see Table 8 for count by country).

<table>
<thead>
<tr>
<th>Region/Ethnicity</th>
<th>Count</th>
<th>Country</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>488</td>
<td>USA</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kenya</td>
<td>254</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nigeria</td>
<td>163</td>
</tr>
<tr>
<td>America</td>
<td>71</td>
<td>USA</td>
<td>71</td>
</tr>
<tr>
<td>Central South Asia</td>
<td>83</td>
<td>USA</td>
<td>83</td>
</tr>
<tr>
<td>East Asia</td>
<td>234</td>
<td>China</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>USA</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Japan</td>
<td>82</td>
</tr>
<tr>
<td>Europe</td>
<td>239</td>
<td>USA</td>
<td>162</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Italy</td>
<td>77</td>
</tr>
</tbody>
</table>

For each classification model, we calculated the True Positive (TP), the True Negative (TN), the False Positive (FP), and the False Negative (FN) values. The total number of positive and negative samples varied based on the number of testing samples in each
group. For example, when we applied a learned classification model on the testing samples for the Africa group, the number of positive samples was the number of samples in the Africa group in the testing data (488). To calculate the number of negative samples for the Africa group, we subtracted the number of positives (488) from the total number of testing samples (1115). Thus, the number of negative samples for the Africa group was 627. We compared the results of the four classifiers on the five testing groups \( g^2 \). For each group, we calculated the following four criteria: accuracy, F1 score, sensitivity, and specificity for different numbers of randomly selected SNPs. Each classifier was applied to the samples of each group on 12 different numbers of selected SNPs.

To summarize the performance of each classifier, we calculated the mean value across the five ethnic groups for each criterion. The results of the mean value for each of the 12 randomly selected SNP sets are discussed in sub-section 4.3.1 (see Fig. 23, and Table 9), and the results of the mean value across all of the 12 randomly selected SNP sets are discussed in sub-section 4.3.2 (see Fig. 24, and Table 10).

Table 9: The mean values of accuracy, F1 score, sensitivity, and specificity across the 5 testing ethnic groups of the proposed classifier, the naive Bayes classifier, the Random Forest classifier, and the BIASLESS classifier. For each row, the maximum values of the F1 score are shown in the cells with green color.
Figure 23: The comparison of the performance of the proposed classifier (red), the naive Bayes classifier (green), the Random Forest classifier (blue), and the BIASLESS classifier (orange) on each set of SNPs. For each ethnic group (i.e. Africa, America, Central South Asia, East Asia, and Europe), the four classifiers were applied on sets of randomly selected SNPs of various sizes $r = \{10, 50, 75, 100, 250, 500, 750, 1000, 2500, 5000, 7500, \text{and} \ 7684\}$. Each box plot shows the values of each criterion (e.g. sensitivity) across the five ethnic groups, when each classifier was applied on a particular number of SNPs set (e.g. 10). The proposed classifier yielded the best F1 scores in 10 out of the 12 cases.

Table 10: The mean values of accuracy, F1 score, sensitivity, and specificity across the 5 testing ethnic groups. The maximum value of each assessment measure (column) is shown in the cells with green color.

<table>
<thead>
<tr>
<th></th>
<th>Acc</th>
<th>F1</th>
<th>Sen</th>
<th>Spc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>0.963</td>
<td>0.827</td>
<td>0.798</td>
<td>0.983</td>
</tr>
<tr>
<td>naive Bayes</td>
<td>0.955</td>
<td>0.744</td>
<td>0.707</td>
<td>0.984</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.934</td>
<td>0.537</td>
<td>0.488</td>
<td>0.990</td>
</tr>
<tr>
<td>BIASLESS</td>
<td>0.899</td>
<td>0.548</td>
<td>0.492</td>
<td>0.965</td>
</tr>
</tbody>
</table>

4.3.1 Mean value at each SNPs set in $r$

To demonstrate the performance of the four classifiers on each set of SNPs in $r$, we calculated the mean of each criterion (e.g. sensitivity) across the five testing groups ($g_2$) for each classifier. Thus, this approach yielded 12 mean values and 12 box plots for each assessment measure from each classifier (see Fig. 23 and Table 9). The proposed classifier yielded the maximum mean accuracy value of 0.991, the maximum mean F1 score value of 0.952, and the maximum mean sensitivity value of 0.935. The Random Forest classifier
Figure 24: The comparison of the overall performance of the proposed classifier (red), the naive Bayes classifier (green), the Random Forest classifier (blue), and the BIASLESS classifier (orange) across all sets of SNPs. For each ethnic group (i.e. Africa, America, Central South Asia, East Asia, and Europe), the four classifiers were applied on sets of randomly selected SNPs of various sizes $r = \{10, 50, 75, 100, 250, 500, 750, 1000, 2500, 5000, 7500, \text{ and } 7684\}$. Each box plot shows the performance summary of each measure for each method across the entire set of SNPs. Each box plot includes 60 values (five populations and 12 SNPs sets in each). The proposed classifier outperformed all other approaches in terms of accuracy, F1 score, and sensitivity (panels 1, 2 and 3, left to right). The mean specificity of the proposed approach is very comparable with that of the random forest classifier but the latter yielded a tighter range (last panel).
Figure 25: The comparison of the performance of the proposed classifier (red), the naive Bayes classifier (green), the Random Forest classifier (blue), and the BIASLESS classifier (orange) on sets of SNPs of various sizes. Each of the four classifiers was applied on a number of randomly selected SNPs $r = \{10, 50, 75, 100, 250, 500, 750, 1000, 2500, 5000, 7500, 7684\}$. Each box plot shows the values of each criterion for samples collected only from the USA but from individuals belonging to different ethnic groups (i.e. Africa, America, Central South Asia, East Asia, and Europe). This assesses the ability of various approaches to address the given problem, i.e. to correctly identify the ethnic background of a sample. The proposed classifier yielded the best F1 scores in 11 out of the 12 cases.

yielded the maximum mean specificity value of 0.999 (see Fig. 23 and Table 9). The proposed classifier yielded the best F1 scores in 10 out of the 12 cases.

The maximum mean sensitivity, accuracy, and $F1$ scores values for the proposed classifier were higher than the maximum mean sensitivity, accuracy, and $F1$ scores values for the naive Bayes classifier, the Random Forest classifier and the BIASLESS classifier. The assessment of the four classifiers by calculating the maximum mean values at each of the 12 randomly selected SNPs set, indicated that the proposed classifier outperformed the other three classifiers.

4.3.2 Mean value across all SNPs sets in $r$

To summarize the performance of each classifier across the five groups and across all sets of SNPs in the range $r$, the mean of each measure was calculated across the range $r$. 
<table>
<thead>
<tr>
<th>Proposed</th>
<th>naiveBayes</th>
<th>RF</th>
<th>BIASLESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1 score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 26: The comparison of the overall performance of the proposed classifier (red), the naive Bayes classifier (green), the Random Forest classifier (blue), and the BIASLESS classifier (orange) across all sets of SNPs. The four classifiers were applied on sets of randomly selected SNPs of various sizes \( r = \{10, 50, 75, 100, 250, 500, 750, 1000, 2500, 5000, 7500, \text{ and } 7684\} \). Each box plot shows the performance summary of each measure for each method across the range of randomly selected SNPs. Each box plot shows the values of each criterion for samples collected only from the USA but from individuals belonging to different ethnic groups (i.e. Africa, America, Central South Asia, East Asia, and Europe). This assesses the ability of various approaches to address the given problem, i.e. to correctly identify the ethnic background of a sample. The proposed classifier outperformed all other approaches in terms of accuracy, F1 score, and sensitivity values. The Random Forest classifier yielded the best mean specificity.
Table 11: The mean values of accuracy, F1 score, sensitivity, and specificity for each group of SNPs and for samples collected only from the USA but from individuals belonging to different ethnic groups (i.e. Africa, America, Central South Asia, East Asia, and Europe). For each row, the maximum values of the F1 score (columns) are shown in the green cells.

<table>
<thead>
<tr>
<th>SNPs</th>
<th>Proposed Acc</th>
<th>naive Bayes Acc</th>
<th>Random Forest Acc</th>
<th>BIASLESS Acc</th>
<th>Proposed F1</th>
<th>naive Bayes F1</th>
<th>Random Forest F1</th>
<th>BIASLESS F1</th>
<th>Proposed Sen</th>
<th>naive Bayes Sen</th>
<th>Random Forest Sen</th>
<th>BIASLESS Sen</th>
<th>Proposed Spc</th>
<th>naive Bayes Spc</th>
<th>Random Forest Spc</th>
<th>BIASLESS Spc</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.796</td>
<td>0.381</td>
<td>0.349</td>
<td>0.913</td>
<td>0.324</td>
<td>0.341</td>
<td>0.911</td>
<td>0.378</td>
<td>0.372</td>
<td>0.922</td>
<td>0.793</td>
<td>0.342</td>
<td>0.34</td>
<td>0.914</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.876</td>
<td>0.625</td>
<td>0.605</td>
<td>0.939</td>
<td>0.689</td>
<td>0.616</td>
<td>0.955</td>
<td>0.868</td>
<td>0.516</td>
<td>0.95</td>
<td>0.842</td>
<td>0.503</td>
<td>0.96</td>
<td>0.842</td>
<td>0.461</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>0.909</td>
<td>0.732</td>
<td>0.696</td>
<td>0.956</td>
<td>0.899</td>
<td>0.646</td>
<td>0.957</td>
<td>0.872</td>
<td>0.518</td>
<td>0.943</td>
<td>0.849</td>
<td>0.535</td>
<td>0.47</td>
<td>0.942</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.914</td>
<td>0.743</td>
<td>0.72</td>
<td>0.961</td>
<td>0.906</td>
<td>0.752</td>
<td>0.972</td>
<td>0.887</td>
<td>0.533</td>
<td>0.973</td>
<td>0.844</td>
<td>0.55</td>
<td>0.449</td>
<td>0.942</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0.947</td>
<td>0.842</td>
<td>0.805</td>
<td>0.977</td>
<td>0.928</td>
<td>0.752</td>
<td>0.972</td>
<td>0.887</td>
<td>0.533</td>
<td>0.973</td>
<td>0.844</td>
<td>0.55</td>
<td>0.449</td>
<td>0.942</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.969</td>
<td>0.919</td>
<td>0.902</td>
<td>0.978</td>
<td>0.945</td>
<td>0.808</td>
<td>0.771</td>
<td>0.976</td>
<td>0.55</td>
<td>0.989</td>
<td>0.851</td>
<td>0.556</td>
<td>0.515</td>
<td>0.934</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>0.969</td>
<td>0.926</td>
<td>0.904</td>
<td>0.977</td>
<td>0.941</td>
<td>0.792</td>
<td>0.764</td>
<td>0.976</td>
<td>0.55</td>
<td>0.989</td>
<td>0.851</td>
<td>0.556</td>
<td>0.515</td>
<td>0.934</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.967</td>
<td>0.921</td>
<td>0.897</td>
<td>0.975</td>
<td>0.941</td>
<td>0.79</td>
<td>0.774</td>
<td>0.967</td>
<td>0.57</td>
<td>0.532</td>
<td>0.988</td>
<td>0.578</td>
<td>0.527</td>
<td>0.934</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>2500</td>
<td>0.975</td>
<td>0.941</td>
<td>0.924</td>
<td>0.981</td>
<td>0.951</td>
<td>0.817</td>
<td>0.804</td>
<td>0.973</td>
<td>0.588</td>
<td>0.993</td>
<td>0.868</td>
<td>0.569</td>
<td>0.49</td>
<td>0.957</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>0.976</td>
<td>0.943</td>
<td>0.926</td>
<td>0.982</td>
<td>0.954</td>
<td>0.823</td>
<td>0.812</td>
<td>0.975</td>
<td>0.604</td>
<td>0.906</td>
<td>0.848</td>
<td>0.553</td>
<td>0.489</td>
<td>0.941</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>7500</td>
<td>0.977</td>
<td>0.945</td>
<td>0.93</td>
<td>0.982</td>
<td>0.954</td>
<td>0.824</td>
<td>0.814</td>
<td>0.975</td>
<td>0.617</td>
<td>0.911</td>
<td>0.849</td>
<td>0.55</td>
<td>0.49</td>
<td>0.936</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>7684</td>
<td>0.977</td>
<td>0.945</td>
<td>0.93</td>
<td>0.982</td>
<td>0.954</td>
<td>0.824</td>
<td>0.814</td>
<td>0.975</td>
<td>0.617</td>
<td>0.911</td>
<td>0.849</td>
<td>0.55</td>
<td>0.49</td>
<td>0.936</td>
<td>0.49</td>
<td></td>
</tr>
</tbody>
</table>

Table 12: The mean values of accuracy, F1 score, sensitivity, and specificity across all sizes of groups of SNPs. The data includes samples collected only from the USA but from individuals belonging to different ethnic groups (i.e. Africa, America, Central South Asia, East Asia, and Europe). The maximum value of each assessment measure (column) is shown in the green cells.

<table>
<thead>
<tr>
<th>Proposed</th>
<th>naive Bayes</th>
<th>Random Forest</th>
<th>BIASLESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc</td>
<td>0.958</td>
<td>0.822</td>
<td>0.799</td>
</tr>
<tr>
<td>F1</td>
<td>0.921</td>
<td>0.733</td>
<td>0.71</td>
</tr>
<tr>
<td>Sen</td>
<td>0.886</td>
<td>0.547</td>
<td>0.506</td>
</tr>
<tr>
<td>Spc</td>
<td>0.845</td>
<td>0.526</td>
<td>0.473</td>
</tr>
</tbody>
</table>

The best mean sensitivity, accuracy, and F1 score values for the proposed classifier are higher than the mean sensitivity, accuracy, and F1 scores values for the naive Bayes classifier, the Random Forest classifier, and the BIASLESS classifier. The best mean specificity value is obtained by the Random Forest classifier. Assessing the four classifiers by calcu-
lating the mean value of accuracy and F1 scores across the range \( r \), shows the proposed classifier outperformed the other three classifiers.

### 4.3.3 Comparisons for USA samples only

To demonstrate the performance of the proposed classifier on a practical example, we evaluated its performance on samples collected from the same country. This models the real-life problem we are trying to address, which is to identify the ethnic background for a given sample found at a crime scene. For this purpose, for each type of classifier, we built a model on all samples from the training data and evaluated using only the validation samples collected in the USA. Similar to sub-sections 4.3.1 and 4.3.2, we compared the results of the proposed classifier with the naive Bayes classifier, the Random Forest classifier, and the BIASLESS classifier. We show the classification results at each SNPs set in \( r \) (see Fig. 25 and Table 11) and across all SNPs sets in \( r \) (see Fig. 26, Table 12).

The results of each SNPs set (from the USA samples) are similar to the results of sub-section 4.3.1. The proposed classifier yielded the maximum mean accuracy value of 0.977, the best mean F1 score value of 0.945, and the best mean sensitivity value of 0.93. The Random Forest classifier yielded the maximum mean specificity value of 0.997 (see Fig. 25 and Table 11). The proposed classifier yielded the best mean F1 scores in 11 out of the 12 cases.

The results across the SNPs sets only from samples collected in the USA are similar to the result of sub-section 4.3.2. The mean values of accuracy, F1 score, sensitivity, and specificity were averaged across all sizes of groups of SNPs. This yielded one mean value and one box plot from each classifier for each assessment measure. The proposed classifier yielded the best mean accuracy value of 0.938, the best mean F1 score value of 0.822, and
the best mean sensitivity value of 0.799. The Random Forest classifier yielded the best mean specificity value of 0.982 (see Fig. 26 and the green cells in Table 12).

Assessing the four classifiers by calculating the values of accuracy, F1 score, and sensitivity across the range $r$, shows the proposed classifier outperformed the other three classifiers when the four classifiers were tested only with samples collected in the USA.

4.4 Conclusion

Identifying the population of an individual by using biological samples is an important problem in forensic evidence and evolutionary biology. In this study, we are introducing a classifier that is able to perform an accurate identification of the population of a sample based only on SNP data from about 10% of the genome. We compared the performance of the proposed classifier with three other classification algorithms i) the naive Bayes classifier ii) the Random Forest (RF) classifier, and iii) the BIASLESS classifier.

The analysis results for each of the 12 SNPs sets show that the proposed classifier yielded the maximum mean accuracy, F1 score, and sensitivity, values when compared with the other three classifiers. Box plots of these results are shown in Fig. 23.

In addition to the results on each of the 12 individual SNPs sets, we also compared the results from the analysis performed across the 12 SNPs sets. Box plots of these results are shown in Fig. 24, and the mean values are shown in Table 9. These results also showed that the proposed classifier yielded the maximum mean sensitivity, accuracy, and F1 score values when compared with the other three classifiers. The Random Forest classifier yielded the maximum specificity value

We demonstrated the performance of the proposed classifier on a practical example.
The real-world application would involve identifying the ethnic background for a sample collected at a crime scene. Therefore, we tested the classifiers only on samples collected in the US but involving five ethnic groups. Every classifier was used to predict the ethnic group for each sample. The proposed classifier outperformed the naive Bayes classifier, the Random Forest classifier and the BIASLESS classifier when assessed with accuracy, F1 score, and sensitivity. The specificity result of the Random Forest classifier was better than the other three classifiers.
CHAPTER 5 PLATELETS AND RENAL FAILURE IN THE SARS-COV-2

5.1 Introduction

The coronavirus disease 19 (COVID-19) is a highly transmittable viral infection caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [164]. Most infected individuals exhibit dry cough, dyspnea, and bilateral ground glass opacities on radiographic images [69]. While the critical importance of respiratory derangement in COVID-19 is well appreciated, several clinical features of COVID-19 infection remain poorly understood including the increased incidence of thromboembolic disease despite being on prophylactic anticoagulation an acute renal injury requiring renal replacement therapy. For instance, based on our observations, several patients rapidly decompensated days after stabilization of oxygenation during or after cessation of mechanical ventilation. Findings of micro and macrothrombi are common in these patients [188]. Others have documented cerebral ischemic infarcts in these subjects [193]. Microvascular thrombosis appears to be one of the defining feature of the novel SARS-CoV-2 infection [104]. Formation of vascular thrombi in association with progressive severe endothelial injury in COVID-19-infected subjects may determine case fatality and renal dysfunction in this disease.

SARS-CoV-2 is single-stranded positive-sense RNA virus [50]. Similar to other CoV, during viral entry into the host cell the spike proteins (S) on the envelope of SARS-CoV-2 are cleaved into S1 and S2 subunits [95]. Through the S1 receptor-binding domain (RBD), S1 directly binds to the peptidase domain (PD) of the metallocarboxyl peptidase angiotensin receptor (ACE) 2 to gain entry into human cells [95, 68, 99]. ACE receptors
(ACE and ACE2) are expressed in almost all tissues and ACE2 is expressed predominantly on the alveolar epithelial type II cells and capillary endothelial cells [64, 87]. ACE and ACE2 regulate vascular tone, fibrinolysis and the coagulation cascade [158, 54]. ACE activity increases vascular tone and activates the coagulation cascade by promoting the generation of angiotensin (Ang) II.

Activation of the renin-angiotensin-aldosterone system (RAAS) plays a critical role in renal injury [161]. Angiotensin II is a potent vasoconstrictor and activates the coagulation cascade by increasing platelet activation and size, inducing microvascular thrombosis [158, 74].

The coagulation cascade, fibrinolytic system and the complement system closely collaborate to control a wide range of biological and pathological functions including immune responses to pathogens, cell migration and tissue hemostasis. A variety of microorganisms can trigger endothelial injury, alternative complement activation and hence activate the coagulation pathway perpetuating inflammation and organ dysfunction. We hypothesized that platelet activation through multiple mechanisms plays an important role in COVID-19 associated coagulopathy and acute kidney injury (AKI). Our data indicate significant increase incidence of AKI in critically ill patients due to COVID-19 infection. In this chapter, we show evidence that mean platelet volume (MPV) reflects platelet activation and activation of coagulation cascades and plays a major role in the development of acute renal failure.
5.2 Methods

5.2.1 Study Design and Data Collection

We analyzed the data for adult patients with laboratory-confirmed COVID-19 infection admitted to the intensive care units (ICU) at Harper University Hospital in Metro Detroit between March 15 to April 20, 2020. Eighty-one adults (18 years of age or older) were identified. Informed consent was waived, and researchers analyzed only identified (anonymized) data. Data from the institution’s electronic medical record were obtained through a research form. We obtained demographic data, information on clinical symptoms or signs at presentation, laboratory, and radiologic results. Furthermore, we reviewed the peripheral blood smears of patients during ICU admission. All laboratory tests were performed at the discretion of the treating physicians. This study was reviewed and approved by the Detroit Medical Center and the Wayne State University Institutional Review Boards.

5.2.2 Specimen Collection and Testing

Clinical specimens for COVID-19 diagnostic testing were obtained in accordance with the Centers for Disease Control and Prevention (CDC) guidelines. Nasal swabs were performed in all patients using a real-time RT-PCR assay developed by the CDC at the Michigan State Public Health Laboratory to detect the virus.

5.2.3 Laboratory Data

All laboratory data were part of routine clinical data obtained during hospitalization and ICU admission. The following data were analyzed: mean platelet volume (MPV) and Glomerular filtration rates (GFR). The GFR data were calculated based on creatinine
clearance, gender and race.

5.2.4 Statistical Analysis

Data are presented as line plots with error bars between two groups i) acute kidney injury (AKI) presence and ii) AKI absence. Independent unpaired t-tests were used measure the significance difference between the variables of the clinical data. Acute kidney injury is defined as per guidelines of American Society of Nephrology [180]. For all analyses, p-values $<0.05$ were considered significant. For the GFR data, a trend analysis was performed on the data from each day for up to 8 days. Furthermore, trend analysis was performed on the daily data for GFR and MPV. Each interval represents data from every 3 days except for the last interval, which includes all data beyond 16 days. We compared the presence and absence of AKI and tested the significance of each interval. A similar analysis was applied on the daily MPV data. Statistical analyses were performed using the programing language R.

5.3 Results

During early March 2020 through April 15, 2020, we reviewed the charts of 81 subjects who were positive for COVID-19. Among the 81 subjects, 41 (51%) were female. Mean age was $60 \pm 13.8$ y (range 24-95 y). Almost all subjects in our study had hypertension (95%), over half (57%) had diabetes mellitus (DM) and many had a combination of both (55%). Patients demographics are shown in Table 13. The mean body mass index (BMI) was $33 \pm 8.9 \ kg.m^{-2}$. Overall mortality in our study was 65%. Among study subjects, 66 patients (81%) developed acute renal injury and 49% required renal replacement therapy.

Acute renal injury (AKI) has been associated with worse outcomes in patients with
Table 13: Clinical characteristics for the patients

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>All Patients (N = 81)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (Mean ± SD)</td>
<td>60±13.8</td>
</tr>
<tr>
<td>BMI (Mean ± SD)</td>
<td>32±11.9</td>
</tr>
<tr>
<td>Gender, N (%)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>41 (50.9)</td>
</tr>
<tr>
<td>Male</td>
<td>40 (49.1)</td>
</tr>
<tr>
<td>Race, N (%)</td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>63 (78%)</td>
</tr>
<tr>
<td>Caucasian</td>
<td>3 (3.7%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1 (1.7%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>14 (17.3%)</td>
</tr>
<tr>
<td>Smoking history, N (%)</td>
<td>14 (6%)</td>
</tr>
<tr>
<td>Coexisting disorder, N (%)</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>75 (92%)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>46 (57%)</td>
</tr>
<tr>
<td>Chronic kidney disease</td>
<td>19 (23%)</td>
</tr>
<tr>
<td>Cardiovascular diseases</td>
<td>19 (23%)</td>
</tr>
<tr>
<td>Respiratory diseases</td>
<td>13 (16%)</td>
</tr>
<tr>
<td>Cancer</td>
<td>2 (2%)</td>
</tr>
</tbody>
</table>

COVID-19, therefore we analyzed laboratory data based on the presence or absence of AKI. Figure 27 shows the GFR trends during 8 days of ICU stay between two groups: absence of AKI (Figure 27A) and presence of AKI (Figure 27B). A student t-test was performed to detect differences in outcome variables of selected laboratory values between the group with and without AKI. Figure 27 shows error bar plots illustrating mean ± SD of the relevant laboratory data comparing subjects, who developed AKI versus, who did not. In order to further determine the kinetic of GFR variation in a time-dependent manner during ICU admission, we collected the daily GFR values of all study subjects and compare the kinetics of GFR based on the presence and absence of AKI. Figure 28A shows plots for GFR in a time-dependent manner based on presence and absence of AKI. As expected, there was a significant difference between GFR values of subjects with AKI versus no AKI, and deterioration in most patients occurred within 4-7 days of ICU stay. Similarly, most
Figure 27: GFR trend among COVID-19 subjects from day one to day eight of ICU-stay. (A) Mean ± SD of GFR (y-axis) in COVID-19 patients without AKI from day one of admission plotted against time, showing the GFR trend. (B) The trend of GFR among COVID-19 patients with AKI from day1-8 of ICU-stay. There was a significant difference between AKI and non-AKI group in terms of GFR. The trend of GFR values during ICU admission among patients with COVID-19 infection for each day. Error bars plots to show the trend of GFR (y-axis) in COVID-19 patients without acute renal injury (A) and with acute renal injury (B) during ICU stay. Each error bars plot shows the data from each day. The trend line connects the means of each day, and the caps represent the SD. The GFR plot shows a continuous deterioration for patients with AKI starting from day 3, in contrast the GFR of patients without AKI shows continuous improvement.
Figure 28: Trend of GFR and MPV during ICU stay among patients with COVID-19 infection. (A) The GFR trend presented as line plot with error bars to show GFR (y-axis) in COVID-19 patients with and without acute renal injury in time-dependent manner (X-axis) during every 3 days of ICU stay. Patients with AKI (red) versus without AKI (black). B) MPV trend of patients with AKI (red) without AKI (black). Days of admission (x-axis). Each error bars plot shows the data of every 3 days. The dot represents the mean, and the cap represents the SD. The FDR p-values of each comparison are shown at the bottom of the plots. The GFR plot shows a deterioration after day 3 for patients with AKI. After 16 days, this trend showed upwards as some of the patients were either discharged or expired. The MPV plot shows a consistently higher level in AKI patients when compared to patients without AKI. After 16 days, this trend of MPV showed downwards. At each time point for GFR and MPV, the comparisons of AKI patients and without AKI were significant as shown of p-value on the bottom of each interval.

Patients with AKI died despite renal replacement therapy. Only six patients, who showed improvement of GFR after 16 days, survived. We performed similar analysis using MPV data. As shown in figure 28B, COVID-19 subjects with AKI exhibited significantly larger platelets based on MPV. It is important to mention that all subjects with COVID-19 have higher MPV values (when compared with normal value), however, in the AKI group the MPV values were significantly higher. MVP values in our study patients did fluctuate day by day but on average, a higher MVP was identified with the presence of AKI.
5.4 Discussion

Our study showed that 81% of critically ill COVID-19 patients developed AKI during their ICU course and half of these patients required renal replacement therapy. Furthermore, the mortality among the critically ill cohort of COVID-19 patients, who developed AKI was significantly higher. Previously, a large study reported the AKI incidence of 17% in critically ill patients admitted to ICU due to H1N1 influenza infection [111]. The incidence of AKI in bacterial sepsis and ARDS has been reported between 23% and 41% during ICU stay [8, 143]. Our data indicate that the prevalence of AKI in COVID-19 infected individuals was substantially higher as compared to AKI incidence due to H1N1 or bacterial sepsis. In COVID-19 infection multiple mechanisms may lead to microcirculatory alterations such as endothelial injury, autonomic nervous system response, activation of the coagulation cascade and RAAS pathway.

Soon after the onset of the COVID-19 pandemic, reports of increased activation of the coagulation pathway and thrombotic events emerged [34, 67]. Patients may present with acute pulmonary embolism, deep venous thrombosis, acute ischemic events, or exhibit clotting of arterial and venous lines.

SARS-CoV-2 utilizes the novel ACE 2 receptor to gain entry into human cells [99]. This process leads to shedding of host ACE2 receptor and the loss of ACE2 function [67]. Alteration of ACE2 function have been associated with chronic kidney disease and diabetes nephropathy [186]. Several studies have shown that hypertension, diabetes and obesity are associated with dysregulation of Ang II and ACE2 [73]. It is important to note that 75% of our patients had hypertension and 58% had diabetes mellitus. The SARS-CoV-2 infection and the loss of function of ACE2 may potentiate the existing prothrombotic state
of subjects with diabetes and hypertension. The unique features of COVID-19 infection are therefore, the loss of ACE2 function and increased Ang II activity promote severe endothelial dysfunction, platelet activation, an increased burden of microvascular thrombosis, and subsequent microcirculatory compromise in many organs including lungs, heart, CNS, and kidney associated with fatal outcomes.
CHAPTER 6 CONCLUSION

First, I presented a framework to detect novel antigens for Sarcoidosis, Tuberculosis and cystic fibrosis using a library derived from sarcoidosis subjects. Lungs are highly exposed to numerous bacteria. Our library is predominantly derived from sarcoidosis BAL cells and WBCs containing diverse immune cells, including macrophages that were exposed to various pathogens. The phage display technology and immunoscreening has utilities not only in identifying diagnostic biomarkers, but also may enable us to develop a novel targeted therapy utilizing the peptide sequences (mimotopes) as vehicles to deliver specific drugs.

Second, I proposed a new method that identified a putative mechanism, and I applied it on 3 real KO data sets. In all three datasets illustrated, the proposed method was able to identify interactions that both: i) were known to be related to the phenomenon, and ii) constitute a plausible explanation for how the effect of the KO gene was propagated in the rest of the system. At the same time, the classical approach mostly produced disconnected graphs that were not helpful to explain other changes in the organism.

Third, I proposed a new population classification method. To demonstrate the performance of the proposed method, I compared its performance with the naive Bayes classifier, the Random Forest classifier, and the BIASLESS classifier. I built the models on samples from the training data and tested on an independent testing data. The proposed classifier outperformed the naive Bayes classifier, the Random Forest classifier and, the BIASLESS classifier when assessed with accuracy, F1 score, and sensitivity.

Fourth, I presented the findings of analyzing ICU clinical data for 81 COVID-19 patients.
The GFR and MPV data on a daily basis and every 3 days are presented in trend analysis plots. I compared the trend values of GFR and MPV between patients with AKI and without AKI. The GFR plot shows a deterioration after day 3 for patients with AKI. The MPV values were significantly higher with AKI patients. These results suggest that SARS-CoV-2 plays a major role in the development of acute renal failure.
REFERENCES


[28] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, Bijay Jassal, Steven Jupe, Lisa Matthews, Bruce May, Stanislav Palatnik, Karen Roth-
fels, Veronica Shamovsky, Heeyeon Song, Mark Williams, Ewan Birney, Henning Hermjakob, Lincoln Stein, and Peter D'Eustachio. The Reactome pathway knowl-

[29] Peggy Cuilliere-Dartigues, Marie-Caroline Meyohas, Pierre Balladur, Norbert Claude Gorin, and Paul Coppo. Splenic sarcoidosis: An unusual aetiology of agranulocyto-


[34] Pierpaolo Di Micco, Vincenzo Russo, Novella Carannante, Michele Imparato, Stefano Rodolfi, Giuseppe Cardillo, and Corrado Lodigiani. Clotting factors in covid-19:


[88] Andrew J. Kinloch, Ylva Kaiser, Don Wolfgeher, Junting Ai, Anders Eklund, Marcus R. Clark, and Johan Grunewald. In situ humoral immunity to vimentin in hla-


[101] Ho-Sheng Lin, Harvinder S. Talwar, Adi L. Tarca, Alexei Ionan, Madhumita Chatterjee, Bin Ye, Jerzy Wojciechowski, Saroj Mohapatra, Marc D. Basson, George H. Yoo, Brian Peshek, Fulvio Lonardo, Chuan-Ju G. Pan, Adam J. Folbe, Sorin Draghici,


[112] Duncan Maru and Sunil Sharma. Tuberculin skin test among pulmonary sarcoidosis patients with and without tuberculosis: Its utility for the screening of the two


[131] Kuang-Hung Pan, Chih-Jian Lih, and Stanley N Cohen. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA mi-


[137] Carsten M Pusch, Ian Giddings, and Michael Scholz. Repair of degraded duplex DNA from prehistoric samples using Escherichia coli DNA polymerase I and T4 DNA


[163] Zhimin Song, Lisa Marzilli, Brian M. Greenlee, Edward S. Chen, Richard F. Silver, Frederic B. Askin, Alvin S. Teirstein, Ying Zhang, Robert J. Cotter, and David R. Moller. Mycobacterial catalase-peroxidase is a tissue antigen and target of the adap-


[170] Harvinder Talwar, Rita Rosati, Jia Li, Dana Kissner, Samiran Ghosh, Félix Fernández-Madrid, and Lobelia Samavati. Development of a t7 phage display library to detect


ABSTRACT

IDENTIFYING BIOLOGICALLY RELEVANT MECHANISMS AND BIOMARKERS USING NOVEL BIOINFORMATICS METHODS

by

SAMER HANOUDI

May 2022

Advisor: Dr. Sorin Draghici

Major: Computer Science

Degree: Doctor of Philosophy

There is a tremendous need to analyze molecular and patient clinical data to identify biomarkers, biological mechanisms, or to simply classify samples accurately. Issues such as: i) limited tools to diagnose many diseases, ii) not considering biological interactions, or iii) damaged DNA samples could cause a challenge in identifying valuable insights. In this work, I try to address these issues by developing different bioinformatic frameworks.

First, I present three frameworks to identify i) Sarcoidosis biomarkers, ii) Tuberculosis biomarkers and iii) Cystic fibrosis (CF) biomarkers. I identified Sarcoidosis biomarkers, and used them to classify Sarcoidosis samples from non-Sarcoidosis (healthy controls, Tuberculosis, and lung cancer) with sensitivity of 0.92 and specificity of 0.88. I identified 10 TB biomarkers and applied them to classify TB samples versus non-TB (healthy controls and sarcoidosis). The area under the receiver operating characteristics (ROC) curve for the top 10 biomarkers was 1 with a sensitivity of 1 and a specificity of 1. I identified 20 CF biomarkers and used them to classify CF from non-CF (healthy controls and lung cancer). The mean area under the ROC curve for the CF biomarkers was 0.97 with a sensitivity of
0.99 and specificity of 0.95.

Second, I present a method that can construct networks of genes that can be considered putative mechanisms. A major challenge in life science research is understanding the mechanism involved in a given phenotype. The putative mechanisms constructed by this approach are not limited to the set of DE genes, but also considers all known and relevant gene-gene interactions. We analyzed three real datasets for which both the causes of the phenotype, as well as the true mechanisms were known. We show that the method identified the correct mechanisms when applied on microarray datasets from mouse. We compared the results of our method with the results of the classical approach, showing that our method produces more meaningful biological insights.

Third, I propose a classification method that is able to analyze genomic data and assign an individual to a particular population/group. A current challenge in forensic evidence is to classify samples using genomic data accurately. Fragmented DNA due to degradation is a common problem with samples from crime scenes. The proposed classification method can use SNPs from as little as 10% of the DNA in the human genome to identify the population background of a sample. I compared the performance of the proposed method with three other classification methods: i) naive Bayes, ii) Random Forest, and iii) BIASLESS. The accuracy, sensitivity, specificity, and F1 score values yielded by the proposed classifier were 0.963, 0.798, 0.983, and 0.827, respectively. The results show that the proposed method outperforms the existing methods.

Finally, I present the findings of analyzing clinical data for 81 COVID-19 ICU patients. The coronavirus disease (COVID-19) is a highly transmissible viral infection caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). I show evidence that
mean platelet volume (MPV) reflects platelet activation and activation of coagulation cascades and plays a major role in the development of acute renal failure. Furthermore, I show that the glomerular filtration rate (GFR) values are deteriorating after day three for patients with acute renal injury (AKI). Such findings will help with the treatment.
AUTOBIOGRAPHICAL STATEMENT

Education

- Ph.D. Candidate Computer Science, Wayne State University, Detroit, MI. (May 2012-current)
- MS, Biomedical Engineering, Wayne State University, Detroit, MI. (Sep 2010 - May 2012)
- BS, Computer Science, Al-Mansour University, Baghdad, Iraq (Sep 1992 - May 1996)

Peer-reviewed journal articles


7. Samer Hanoudi, Harvinder Talwar, Sorin Draghici, and Lobelia Samavati. “Autoantibodies Against Cytoskeletons and Lysosomal Trafficking in Sarcoidosis Discriminate Sarcoidosis From Healthy Controls, Tuberculosis and Lung Cancers” (Accepted November/30/2021)