**Non-invasive Analysis of the Sputum Transcriptome Discriminates Clinical Phenotypes of Asthma**

**Xiting Yan, Jen-Hwa Chu, Jose Gomez, Maria Koenigs, Carole Holm, Xiaoxuan He, Mario F. Perez, Hongyu Zhao, Shrikant Mane, Fernando D. Martinez, Carole Ober, Dan L. Nicolae, Kathleen C. Barnes, Stephanie J. London, Frank Gilliland, Scott T. Weiss, Benjamin A. Raby,Lauren Cohn and Geoffrey L Chupp**

**Online Data Supplement**

Table of Contents

[**Non-invasive Analysis of the Sputum Transcriptome Discriminates Clinical Phenotypes of Asthma** 1](#_Toc409598164)

[Microarray Data Preprocessing 3](#_Toc409598165)

[*Sample Screening* 3](#_Toc409598166)

[*Normalization and Quality Control* 3](#_Toc409598167)

[*Batch Effect and RIN Number Adjustment* 3](#_Toc409598168)

[Identifying the Sputum TEA Clusters 5](#_Toc409598169)

[*Pathway based distance between samples* 5](#_Toc409598170)

[*Pathways driving each TEA cluster* 6](#_Toc409598171)

[*Pathways enriched for DEGs of each TEA cluster* 7](#_Toc409598172)

[*TH2 gene expression among the TEA clusters* 8](#_Toc409598173)

[*Sputum TEA classifier using blood gene expression* 8](#_Toc409598174)

[Validation of clinical phenotypes associated with TEA clusters in a second cohort 11](#_Toc409598175)

[*Asthma BRIDGE Cohort* 11](#_Toc409598176)

[*Sputum TEA cluster identification* 11](#_Toc409598177)

[REFERENCES 12](#_Toc409598178)

# Microarray Data Preprocessing

## *Sample Screening*

We have collected 190 visits by subjects including both asthma patients and healthy controls. All these subjects have total pack year 10, squamous cell %25% and have at least one array hybridized. Some of these visits were replicated visits by the same patient. Therefore, for the same subject, we only kept the visit with the highest RNA quality (RIN number) which resulted in 171 visits by subjects including 12 healthy normal controls and 159 asthma patients. For the 159 asthma patients, 108 were measured for the sputum gene expression profile using the Affymetrix HuGene 1.0 ST arrays. A flow chart showing the sample size for the sample screening steps is shown in Figure E1.

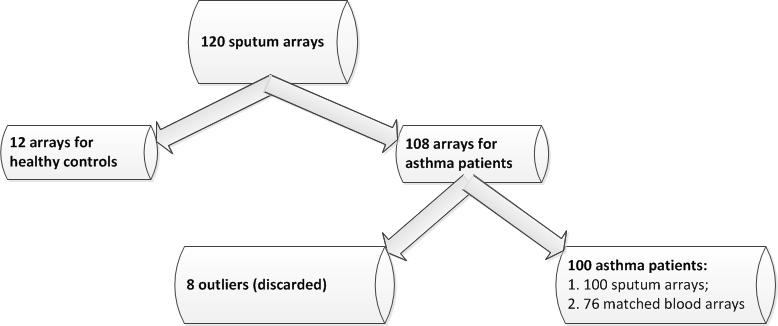


Figure E1. Flow chart showing the sample size for the sample screening steps.

## *Normalization and Quality Control*

The raw microarray data (CEL files) were corrected for background signal, quantile normalized and summarized using robust multiarray averageE[1](#_ENREF_1) (RMA) implemented by the XPS R package to obtain the log2 transformed normalized gene expression levels in both blood and sputum. Then principal component analysis (PCA) was applied to visualize the data and remove outlying arrays. To make sure the data has enough information for the batch effect adjustment, array batches with less than 3 arrays were removed from further analysis. In total, 8 arrays were removed based resulting in 112 sputum arrays for 100 asthma patients and 12 healthy controls.

## *Batch Effect and RIN Number Adjustment*

Visualization of the array data using PCA showed existed batch effect in the data (Figure E2A). ComBatE[2](#_ENREF_2) was applied to adjust the data for batch effect (Figure E2B). The batch effect adjusted data was further adjusted for RIN number using a linear regression model to remove the effect of RNA quality on the expression data. The three dimensional PCA plot of all the 112 sputum arrays after both the batch effect and RIN number adjustment is shown in Figure E3.

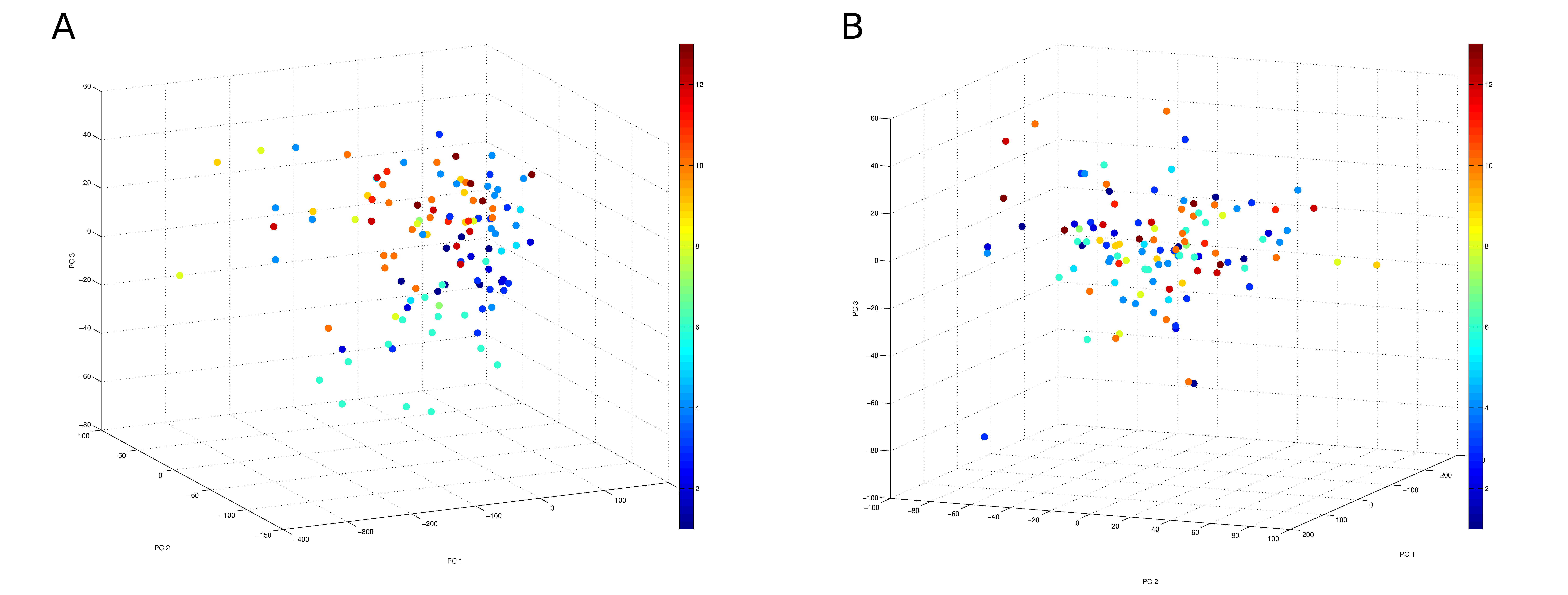


Figure E2: PCA plots for all the 112 sputum arrays (A) before the batch effect adjustment and (B) after the batch effect adjustment. Arrays were colored based on their batches so that arrays scanned at closer time have closer colors from the color bar.

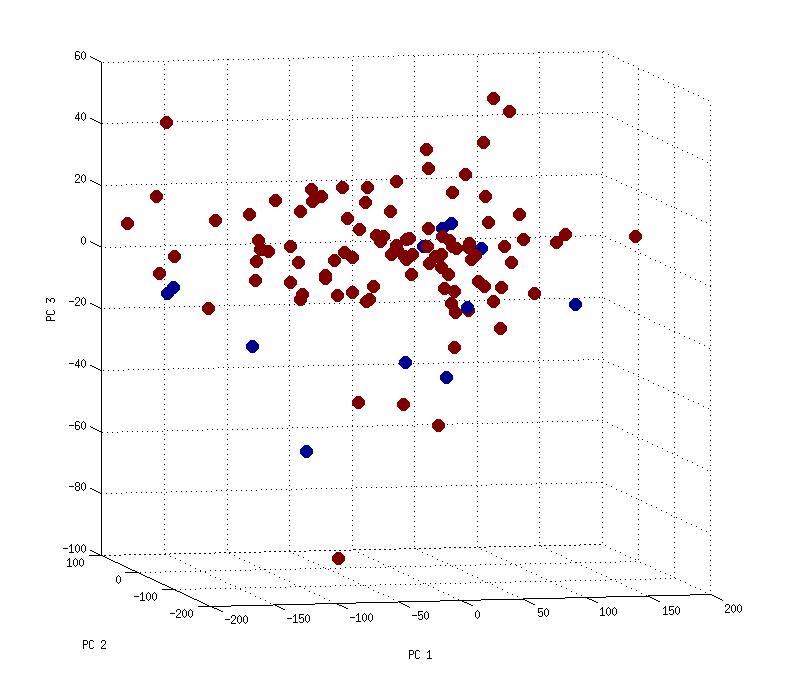


Figure E3: Three dimensional PCA plot for all the 112 sputum arrays. Red dots are asthma patients and blue dots are controls.

The principal component analysis was applied to the batch and RIN adjusted data and labelled the arrays based on the self-reported race (Figure E4). No significant population stratification effect was observed.

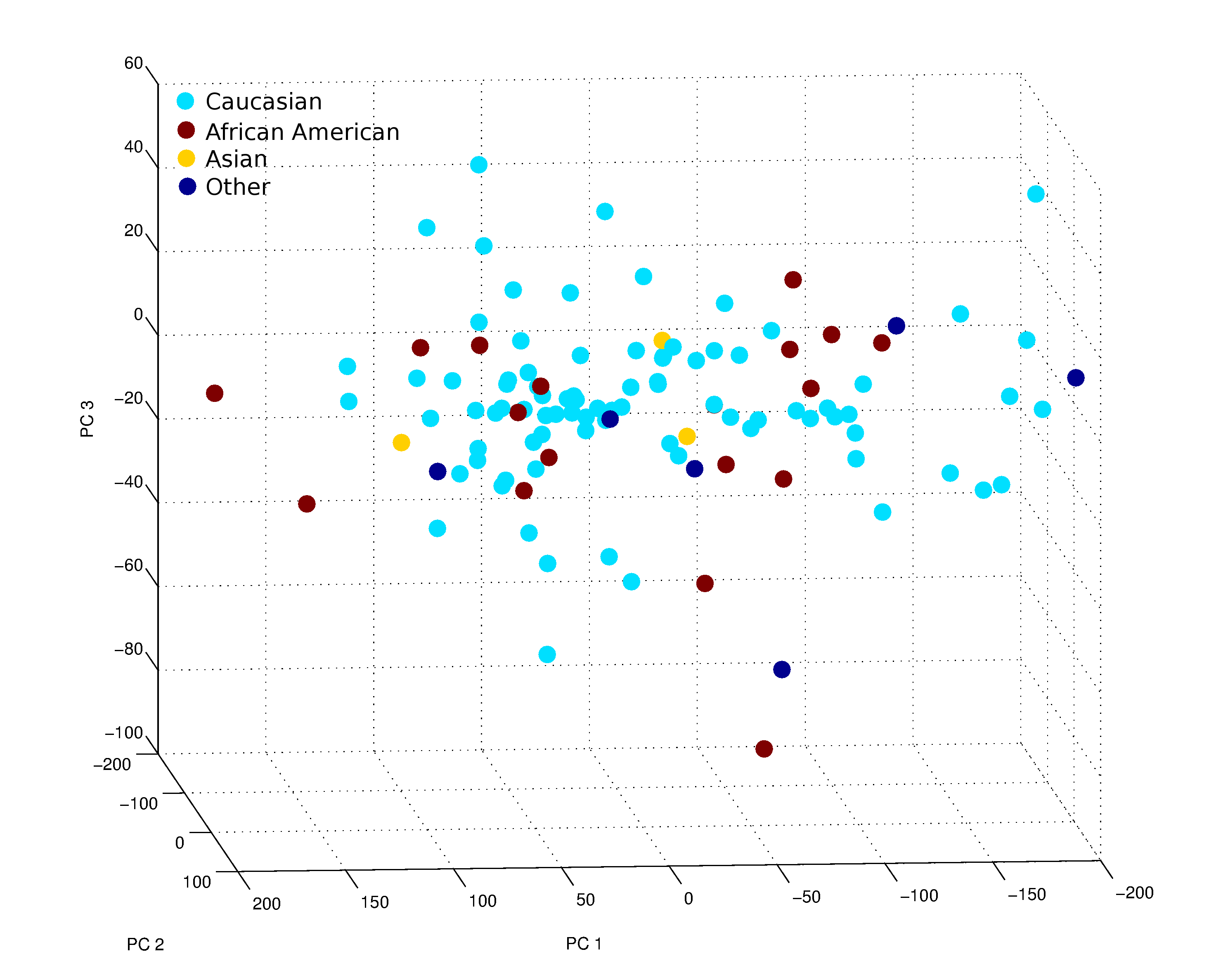


Figure E4: PCA plot for all the 112 sputum arrays labelled using the self-reported race.

# Identifying the Sputum TEA Clusters

## *Pathway based distance between samples*

To cluster the samples in an unbiased way, the distance between samples was calculated. One traditional way is to calculate the Euclidean distance between samples using genes with the largest variance. However, the Euclidean distance applies equal weights to all the genes regardless of the high variation in their expression levels. This may result in inaccurate measurement of the distance and errors in accurately identifying subtypes of asthma.

To overcome the methodological limitations of the traditional Euclidean distance, we developed a novel method to measure distances between samples that implicitly applies different weights by integrating genes in each pathway and calculating the distance by summarizing across different pathways. Known pathways were downloaded from MsigDBE[3](#_ENREF_3),[4](#_ENREF_4) to provide list of member genes in each pathway. Using the expression levels of genes in each pathway, a mixture Gaussian model (MCLUSTE[5](#_ENREF_5)) is fitted to identify any clusters or subtypes in the samples. Suppose the -th pathway identified clusters and the cluster assignment of sample is denoted as , where . The pathway-based distance between sample and is then calculated as

where is the number of pathways that assign the two samples into different clusters and is the total number of pathways that identified at least two clusters. Visualization of both the distance matrix (Figure 1B) and the connectivity criteriaE[6](#_ENREF_6) (Figure E5) showed that there seems to be three clusters. To identify the actual clusters, K means was then applied with the number of clusters set to 3. Compared to the clustering results by Euclidean distance, the pathway based distance matrix achieved lower internal criteria (data not shown).



Figure E5: The connectivity of the clustering results versus different numbers of clusters.

## *Pathways driving each TEA cluster*

For each of the KEGG pathway that identified more than one cluster using the Gaussian mixture model, a score was calculated to show how strong the given KEGG pathway was driving each TEA cluster. For a given TEA cluster and a given pathway , we first align the TEA clusters with the clusters identified by pathway . Suppose cluster has the largest number of overlapping samples with TEA cluster among all the clusters identified by pathway . Then the score is defined as the percentage of samples in TEA cluster that are not in cluster plus the percentage of samples in cluster that are not in TEA cluster . All the KEGG pathways were ranked based on this score and the top 5 KEGG pathways for each of the TEA clusters are shown in Table E1.

Table E1: The top 5 KEGG pathways that drive each TEA cluster. The score that was used to rank the pathways are also shown.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **TEA 1** | | **TEA 2** | | **TEA 3** | |
| ***Pathways*** | ***Score*** | ***Pathways*** | ***Score*** | ***Pathways*** | ***Score*** |
| KEGG\_NEUROACTIVE\_LIGAND\_RECEPTOR\_INTERACTION | 0.021 | KEGG\_PROGESTERONE\_MEDIATED\_OOCYTE\_MATURATION | 0.128 | KEGG\_NEUROACTIVE\_LIGAND\_RECEPTOR\_INTERACTION | 0.410 |
| KEGG\_PATHWAYS\_IN\_CANCER | 0.043 | KEGG\_VIBRIO\_CHOLERAE\_INFECTION | 0.128 | KEGG\_WNT\_SIGNALING\_PATHWAY | 0.457 |
| KEGG\_ADHERENS\_JUNCTION | 0.060 | KEGG\_EPITHELIAL\_CELL\_SIGNALING\_IN\_HELICOBACTER\_PYLORI\_INFECTION | 0.138 | KEGG\_INSULIN\_SIGNALING\_PATHWAY | 0.500 |
| KEGG\_TASTE\_TRANSDUCTION | 0.063 | KEGG\_ALANINE\_ASPARTATE\_AND\_GLUTAMATE\_METABOLISM | 0.150 | KEGG\_NEUROTROPHIN\_SIGNALING\_PATHWAY | 0.513 |
| KEGG\_TIGHT\_JUNCTION | 0.063 | KEGG\_CHEMOKINE\_SIGNALING\_PATHWAY | 0.150 | KEGG\_PROXIMAL\_TUBULE\_BICARBONATE\_RECLAMATION | 0.513 |

## *Pathways enriched for DEGs of each TEA cluster*

We identified the differentially expressed genes (DEGs) between each TEA cluster and controls to better understand the underlying biological mechanisms of each TEA cluster. Functional analysis of these DEGs was performed using GeneGo MetaCore which assesses the significant of enrichment of a collection of pathways in the DEGs. The list of top 10 enriched pathways can be found in Table E2. Since there are no significant DEGs under FDR<0.2 for TEA2, this analysis was not be performed for TEA2.

Table E2. List of the top 10 enriched pathways for TEA1 and TEA3 by GeneGO MetaCore.

|  |  |  |
| --- | --- | --- |
| **TEA 1** | | |
| **Maps** | **Total** | **pValue** |
| Histidine-glutamate-glutamine and proline metabolism/ Rodent version | 44 | 0.0044 |
| Colorectal cancer (general schema) | 29 | 0.0095 |
| Histamine metabolism | 10 | 0.0097 |
| Phospholipid metabolism p.2 | 12 | 0.0139 |
| Leukotriene 4 biosynthesis and metabolism | 14 | 0.0188 |
| DNA damage\_NHEJ mechanisms of DSBs repair | 19 | 0.0336 |
| Plasmalogen biosynthesis | 19 | 0.0336 |
| G-protein signaling\_Cross-talk between Ras-family GTPases | 20 | 0.0370 |
| Cell adhesion\_Gap junctions | 22 | 0.0441 |
| Huntington's disease (general schema) | 3 | 0.0454 |
| **TEA 3** | | |
| **Maps** | **Total** | **pValue** |
| Protein folding and maturation\_POMC processing | 30 | 5.8E-25 |
| Cytoskeleton remodeling\_TGF, WNT and cytoskeletal remodeling | 111 | 5.7E-08 |
| Development\_PIP3 signaling in cardiac myocytes | 47 | 1.2E-06 |
| Cytoskeleton remodeling\_Cytoskeleton remodeling | 102 | 1.4E-06 |
| Signal transduction\_AKT signaling | 43 | 9.0E-05 |
| Development\_Thrombopoietin-regulated cell processes | 45 | 1.2E-04 |
| Development\_IGF-1 receptor signaling | 52 | 2.7E-04 |
| Apoptosis and survival\_Role of PKR in stress-induced apoptosis | 53 | 2.9E-04 |
| Cell cycle\_Influence of Ras and Rho proteins on G1/S Transition | 53 | 2.9E-04 |
| Development\_WNT signaling pathway. Part 2 | 53 | 2.9E-04 |

## *TH2 gene expression among the TEA clusters*

Previous studies have identified the TH-2 inflammation high group of asthma patients based on the Th2 signature gene (IL-4, IL-5 and IL-13) expression levels in both sputum and blood. E[7-9](#_ENREF_7)We evaluated the mean Th2 signature gene expression levels of the TEA clusters in the sputum (Figure E6). As can be seen in Figure E6, there is a significant difference in the Th2 gene expression level among the TEA clusters, with TEA 1 showing the highest Th2 gene expression level (p value=4.2e-8). This is consistent with TEA cluster 1 having the highest FeNO level, ICS requirement, and risk of intubations. This shows consistent clinical and biological characteristics among the TEA clusters.

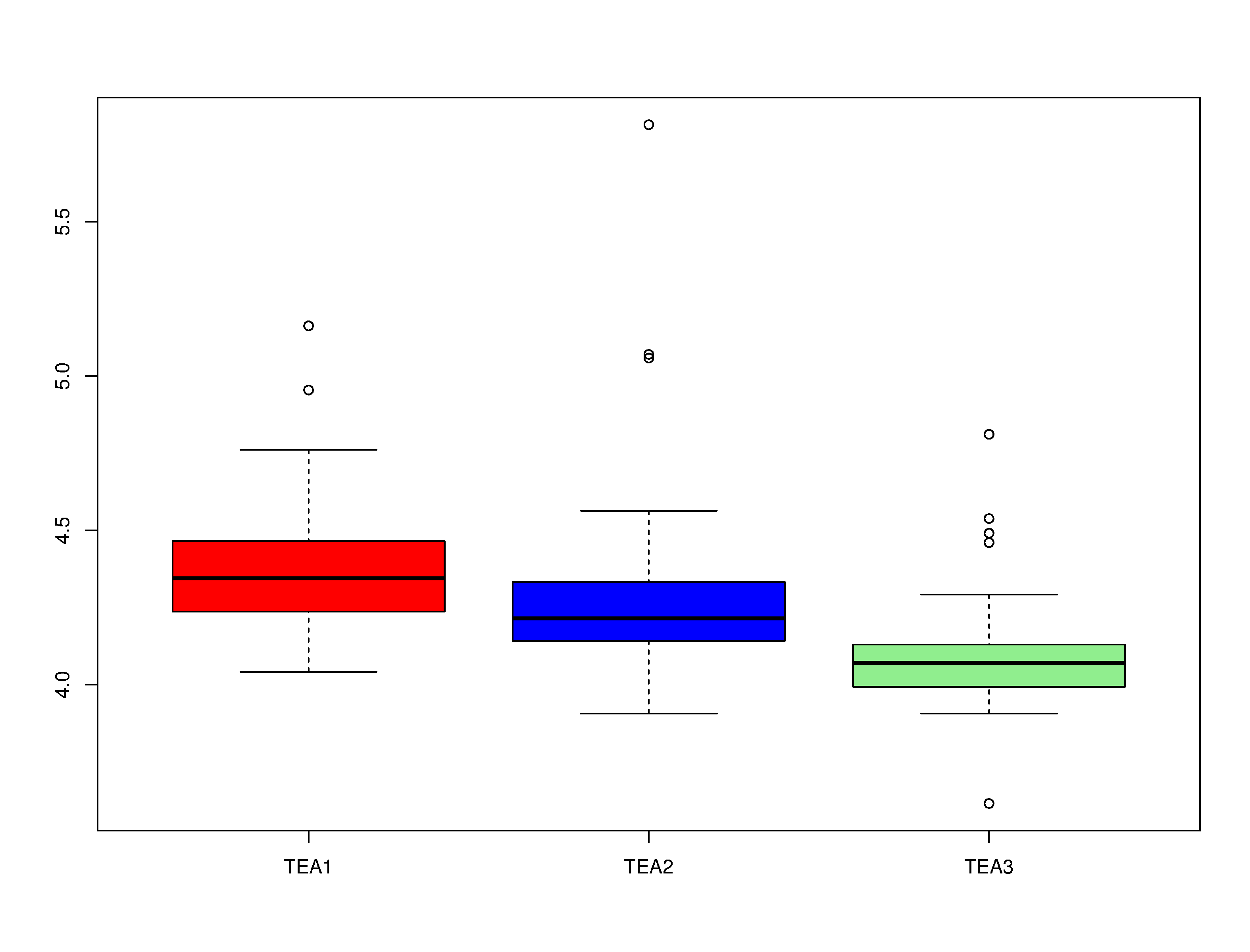


Figure E6. Boxplots of the mean TH2 gene expression in the three TEA clusters.

## *Sputum TEA classifier using blood gene expression*

A logistic regression model was built to predict the TEA cluster of samples using blood gene expression data from validation cohort. Among the 100 samples included in the sputum TEA cluster analysis, 76 of them also have matched blood arrays, which were used as a training data set to build the logistic regression model. First, the sputum TEA cluster assignment of the 76 samples was obtained and the blood gene expression profiles of samples from each TEA cluster were compared to each other using the moderate Student’s t test. Genes with a nominal p values<0.001 were considered to be significant. Second, we merged the genes differentially expressed between each pair of TEA clusters and compare this merged gene set to the 19,601 genes probed on both the Affymetrix HuGene1.0 ST array and the Illumina HumanHT12 BeadChips. About 301 overlapping genes were overlapped. Finally, the data was standardized so that each gene has a mean of 0 and standard deviation of 1. Then a L1 regularized logistic regression model was fitted to the blood expression levels of these 301 genes which further selected 53 genes for the sputum TEA cluster classifying model using blood gene expression data. The list of these 53 genes and the top 10 enriched pathways by GeneGO MetaCore are shown in Table E3 and Table E4. The tuning parameter in LASSO was chosen using leave-one-out cross validation. The sputum TEA cluster classifier demonstrated an internal prediction accuracy of 89%.

Table E3. List of the 53 genes selected by LASSO for classifying patients with whole blood gene expression data.

|  |  |  |  |
| --- | --- | --- | --- |
| **transcript\_cluster\_id** | **Chromosome** | **Gene Symbol** | **Gene Name** |
| 7898594 | chr1 | HTR6 | 5-hydroxytryptamine (serotonin) receptor 6 |
| 7907396 | chr1 | C1orf10 | chromosome 1 open reading frame 105 |
| 7911017 | chr1 | SDCCAG8 | serologically defined colon cancer antigen 8 |
| 7915385 | chr1 | EDN2 | endothelin 2 |
| 7923131 | chr1 | DENND1B | DENN/MADD domain containing 1B |
| 7925728 | chr1 | OR6F1 | olfactory receptor, family 6, subfamily F, member 1 |
| 7940114 | chr11 | OR1S1 | olfactory receptor, family 1, subfamily S, member 1 |
| 7940989 | chr11 | ESRRA | estrogen-related receptor alpha |
| 7948155 | chr11 | OR5AP2 | olfactory receptor, family 5, subfamily AP, member 2 |
| 7949021 | chr11 | RCOR2 | REST corepressor 2 |
| 7950321 | chr11 | UCP3 | uncoupling protein 3 (mitochondrial, proton carrier) |
| 7950332 | chr11 | C2CD3 | C2 calcium-dependent domain containing 3 |
| 7950374 | chr11 | P4HA3 | prolyl 4-hydroxylase, alpha polypeptide III |
| 7951521 | chr11 | C11orf65 | chromosome 11 open reading frame 65 |
| 7967337 | chr12 | ABCB9 | ATP-binding cassette, sub-family B (MDR/TAP), member 9 |
| 7980213 | chr14 | PROX2 | prospero homeobox 2 |
| 8000899 | chr16 | ZNF768 | zinc finger protein 768 |
| 8000974 | chr16 | ZNF668 | zinc finger protein 668 |
| 8007794 | chr17 | ARL17A | ADP-ribosylation factor-like 17A |
| 8012028 | chr17 | ASGR2 | asialoglycoprotein receptor 2 |
| 8024934 | chr19 | ZNRF4 | zinc and ring finger 4 |
| 8028552 | chr19 | NFKBIB | nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, beta |
| 8030416 | chr19 | SCAF1 | SR-related CTD-associated factor 1 |
| 8037283 | chr19 | PSG4 | pregnancy specific beta-1-glycoprotein 4 |
| 8045208 | chr2 | POTEE | POTE ankyrin domain family, member E |
| 8053341 | chr2 | REG3A | regenerating islet-derived 3 alpha |
| 8061564 | chr20 | ID1 | inhibitor of DNA binding 1, dominant negative helix-loop-helix protein |
| 8062108 | chr20 | PROCR | protein C receptor, endothelial |
| 8066411 | chr20 | C20orf62 | chromosome 20 open reading frame 62 |
| 8076046 | chr22 | TMEM184B | transmembrane protein 184B |
| 8076219 | chr22 | SNORD83B | small nucleolar RNA, C/D box 83B |
| 8083605 | chr3 | RSRC1 | arginine/serine-rich coiled-coil 1 |
| 8083794 | chr3 | MYNN | Myoneurin |
| 8086048 | chr3 | TMPPE | transmembrane protein with metallophosphoesterase domain |
| 8091537 | chr3 | IGSF10 | immunoglobulin superfamily, member 10 |
| 8093852 | chr4 | MSX1 | msh homeobox 1 |
| 8097626 | chr4 | LOC441046 | glucuronidase, beta pseudogene |
| 8100519 | chr4 | TXNDC9 | thioredoxin domain containing 9 |
| 8104463 | chr5 | MARCH6 | membrane-associated ring finger (C3HC4) 6 |
| 8105842 | chr5 | CENPH | centromere protein H |
| 8114572 | chr5 | HBEGF | heparin-binding EGF-like growth factor |
| 8115871 | chr5 | FLJ16171 | FLJ16171 protein |
| 8124588 | chr6 | GPX6 | glutathione peroxidase 6 (olfactory) |
| 8125748 | chr6 | LYPLA2P1 | lysophospholipase II pseudogene 1 |
| 8136863 | chr7 | TMEM139 | transmembrane protein 139 |
| 8140909 | chr7 | ERVWE1 | endogenous retroviral family W, env(C7), member 1 |
| 8148655 | chr8 | TIGD5 | tigger transposable element derived 5 |
| 8150920 | chr8 | CYP7A1 | cytochrome P450, family 7, subfamily A, polypeptide 1 |
| 8150978 | chr8 | CA8 | carbonic anhydrase VIII |
| 8153043 | chr8 | ZFAT | zinc finger and AT hook domain containing |
| 8155696 | chr9 | FAM122A | family with sequence similarity 122A |
| 8166127 | chrX | GLRA2 | glycine receptor, alpha 2 |
| 8174648 | chrX | CXorf61 | chromosome X open reading frame 61 |

Table E4. The top 10 pathways enriched for the 53 gene profile by GeneGO MetaCore.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Maps** | **Total** | **Genes** | **pValue** | **Min FDR** |
| [Immune response\_Lipoxins and Resolvin E1 inhibitory action on neutrophil functions](http://portal.genego.com/cgi/imagemap.cgi?id=2732) | 2/20 | I-kB, NFKBIB | 1.209E-03 | 1.050E-01 |
| [Impaired inhibitory action of lipoxins and Resolvin E1 on neutrophil functions in CF](http://portal.genego.com/cgi/imagemap.cgi?id=2693) | 2/23 | I-kB, NFKBIB | 1.603E-03 | 1.050E-01 |
| [Development\_ERBB-family signaling](http://portal.genego.com/cgi/imagemap.cgi?id=535) | 2/36 | I-kB, HB-EGF | 3.914E-03 | 1.343E-01 |
| [Immune response\_HMGB1/RAGE signaling pathway](http://portal.genego.com/cgi/imagemap.cgi?id=6111) | 2/50 | I-kB, NFKBIB | 7.453E-03 | 1.343E-01 |
| [Apoptosis and survival\_Role of PKR in stress-induced apoptosis](http://portal.genego.com/cgi/imagemap.cgi?id=6523) | 2/51 | I-kB, NFKBIB | 7.745E-03 | 1.343E-01 |
| [Development\_Role of IL-8 in angiogenesis](http://portal.genego.com/cgi/imagemap.cgi?id=3051) | 2/53 | I-kB, HB-EGF | 8.346E-03 | 1.343E-01 |
| [Immune response\_Role of PKR in stress-induced antiviral cell response](http://portal.genego.com/cgi/imagemap.cgi?id=6524) | 2/55 | I-kB, NFKBIB | 8.967E-03 | 1.343E-01 |
| [Immune response\_Gastrin in inflammatory response](http://portal.genego.com/cgi/imagemap.cgi?id=3136) | 2/60 | I-kB, HB-EGF | 1.061E-02 | 1.343E-01 |
| [Development\_EGFR signaling pathway](http://portal.genego.com/cgi/imagemap.cgi?id=443) | 2/65 | I-kB, HB-EGF | 1.237E-02 | 1.343E-01 |
| [Development\_Role of nicotinamide in G-CSF-induced granulopoiesis](http://portal.genego.com/cgi/imagemap.cgi?id=6425) | 1/8 | ID1 | 2.111E-02 | 1.343E-01 |

# Validation of clinical phenotypes associated with TEA clusters in a second cohort

## *Asthma BRIDGE Cohort*

1,542 subjects from the EVE Consortium of asthma genome-wide association studies were invited to participate in Asthma BRIDGE by completing a detailed phenotype questionnaire and providing tissue samples for gene expression profiling, including a subset of 981 who provided whole blood samples. Gene expression profiling was conducted using HumanHT12 BeadChips (Illumina, San Diego CA), employing a randomized-array allocation strategy to minimize potential confounding by technical batch effects. Following filtering of poor quality samples, 870 high quality whole blood expression profiles were available for analysis. Quantlie normalization and log2 transformation were conducted with the *lumi* (R package) (see <https://biolincc.nhlbi.nih.gov/home/>).E[10](#_ENREF_10) Illumina gene probes (Asthma BRIDGE cohort) were selected that matched Affymetrix HuGene 1.0 ST array probe sets (YCAAD cohort). Comparable information is available for 19,601 of the 22,148 probe sets on the Affymetrix HuGene 1.0 ST arrays.

## *Sputum TEA cluster identification*

The log2 transformed gene expression levels of the 53 genes chosen for the sputum TEA classifier model were obtained and standardized for each sample in the Asthma BRIDGE cohort. The expression levels were imported into the sputum TEA cluster classifier to predict the sputum TEA cluster assignment for each sample in the Asthma BRIDGE cohort. The predicted probability of a sample belonging to a particular TEA cluster was compared among the three TEA clusters (Figure E6). The comparison showed that samples in each TEA cluster are significantly more likely to belong to that cluster compared to the samples assigned to the other clusters (p value<2.2e-16 for all three TEA clusters). This quantitatively demonstrates that the three TEA clusters are well separated from each other in the Asthma BRIDGE cohort.

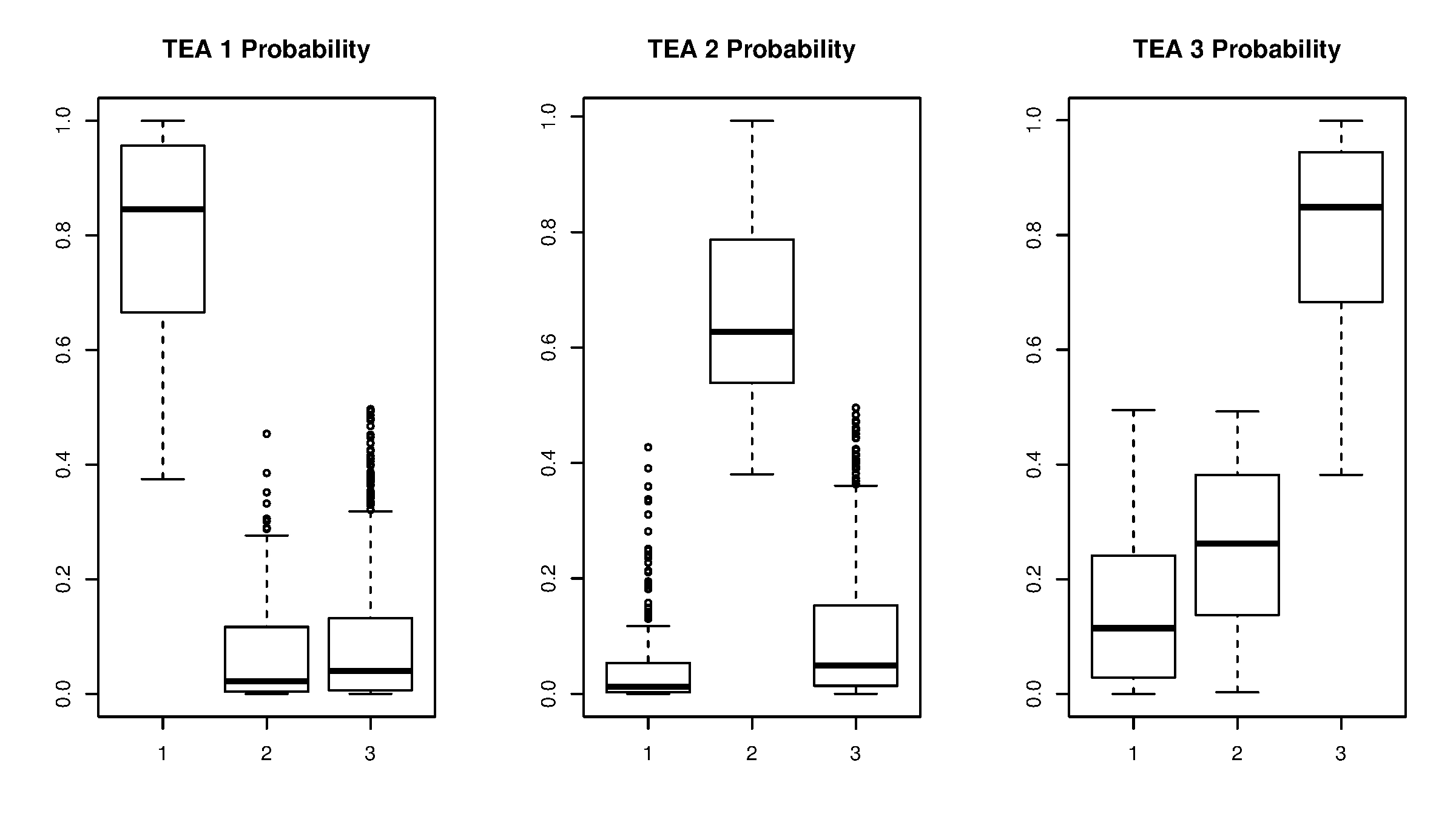
****

Figure E7. Boxplot of the predicted probability of samples belonging to each TEA cluster 1 (left panel), TEA cluster 2 (middle panel) and TEA cluster 3 (right panel), respectively. X-axis of the boxplots show the TEA cluster assignment of the samples.

# REFERENCES

E1. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic acids research 2003;31:e15.

E2. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8:118-27.

E3. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics 2003;34:267-73.

E4. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 2005;102:15545-50.

E5. Fraley C, Raftery AE. MCLUST: Software for model-based cluster analysis. J Classif 1999;16:297-306.

E6. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics 2005;21:3201-12.

E7. Peters MC, Mekonnen ZK, Yuan S, Bhakta NR, Woodruff PG, Fahy JV. Measures of gene expression in sputum cells can identify TH2-high and TH2-low subtypes of asthma. The Journal of allergy and clinical immunology 2014;133:388-94.

E8. Woodruff PG, Boushey HA, Dolganov GM, et al. Genome-wide profiling identifies epithelial cell genes associated with asthma and with treatment response to corticosteroids. Proceedings of the National Academy of Sciences of the United States of America 2007;104:15858-63.

E9. Woodruff PG, Modrek B, Choy DF, et al. T-helper type 2-driven inflammation defines major subphenotypes of asthma. American journal of respiratory and critical care medicine 2009;180:388-95.

E10. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics 2008;24:1547-8.