

Validation of the Nanostring of the Banff Human Organ Transplant Gene Panel in Archival Data from Human Kidney Transplants and its Modelling Complexity

Rex Smith (✉ msmith@mgh.harvard.edu)

Massachusetts General Hospital

Research article

Keywords: Kidney, renal, transplantation, gene expression, statistics, modelling, classification, Nanostring, BHOT

Posted Date: September 15th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-70063/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published on March 19th, 2021. See the published version at <https://doi.org/10.1186/s12920-021-00891-5>.

Abstract

Background

RNA gene expression of renal transplantation biopsies is commonly used to identify rejection. Mostly done with microarrays, seminal findings describe and define the patterns of genes associated with types of rejection and non-rejection kidney allograft diagnoses. To make gene expression more accessible for pathology laboratories, the Molecular Diagnostics Working Group of the Banff Foundation for Allograft Pathology and NanoString Technologies partnered to create the Banff Human Organ Transplant Panel (BHOT), a gene panel set of 770 genes as a substitute for microarrays (~ 50,000 genes). The advantage of this platform is that gene expressions are quantifiable on formalin fixed and paraffin embedded archival tissue samples. This new technology, thus, makes gene expressions cheaper and accessible to more laboratories and investigators. The purpose of this report is to validate the BHOT panel as a surrogate for microarrays and test the accuracy of the modelled BHOT data.

Results

This limited NanoString gene set readily identifies renal rejection and non-rejection diagnostic patterns using *in silico* statistical analyses of seminal archival databases derived from renal transplant RNA expression arrays. Multiple modelling algorithms show a highly variable pattern within the error matrices per sample. The discrepancies within the error matrices are most likely related to the gene expression heterogeneity of samples within a given pathological diagnosis. This was confirmed by clustering the data into 8 groups, which modelled with fewer misclassifications.

Conclusion

This report validates gene expression of human renal allografts using the Banff Human Organ Transplant Panel as a surrogate for microarrays and confirms the its modelling complexity.

Background

RNA gene expression is now commonly used to find patterns of gene expression in renal transplants. Mostly done using microarrays on fresh tissue, many informative and seminal studies identified the dominant pattern of differential gene expressions associated with renal transplant diagnoses [1–9].

Allograft gene expression better defines endpoints and risk in renal allografts [10, 11], subtypes of ABMR [12], dysfunction [4, 13, 14], define a new diagnosis of C4d-negative ABMR [15, 16], and identify NK cells in ABMR [17]. In addition, in non-human primates, unsupervised principal components from Nanostring gene expression identified gene expression patterns that predicted outcomes including tolerance [18, 19].

Recent technology, NanoString nCounter, employs formalin fixed paraffin embedded tissue as the RNA source for gene expression studies, i.e., archival specimens [20]. Advantages of NanoString include gene expression analyses on the same tissue used for diagnosis, and more importantly, readily permits the expansion of gene expression studies to more laboratories that routinely make renal transplant diagnoses. NanoString uses synthetic oligonucleotides as targets for the RNA fragments derived by extraction of the RNA. NanoString gene panels employ only 770 gene targets and, therefore, are not gene discovery tools.

To promote gene expression in renal transplants, the Molecular Diagnostic Working Group of the Banff Foundation for Allograft Pathology and NanoString Technologies partnered to create a subset of microarray genes, the Banff Human Organ Transplant (BHOT) panel, which could be widely used by many laboratories, thereby, allowing widespread collaborations and clinical trials [21].

Optimally validation of the BHOT panel would be done by comparing the BHOT panel and microarrays on the same RNA, but such an experiment has not yet been done. The purpose of this report is to test *in silico* on archival data if the BHOT panel genes shows similar expression patterns corresponding to those previously reported using microarrays [1–9, 22]. Validation of the BHOT panel will encourage additional investigations of human renal allograft rejection. In addition, modelling studies were performed to test how well the BHOT expression data identifies the diagnostic classes.

Materials And Methods

Data. Downloaded text files of GSE data sets 30718 [6], 36059 [10, 23, 24], and 48581 [10, 25] from NCBI all derived from HU-133 plus 2 microarrays with their diagnostic annotations were first imported into excel. These seminal reports from these three databases established the gene expression patterns for T Cell Mediated Rejection (TCMR), Antibody Mediated Rejection (ABMR), and delayed graft function (Acute Kidney Injury, AKI) [6, 10, 23–25]. Gene names were checked due to gene label updates used in the Banff Human Organ Transplant gene set panel (BHOT). Microarrays contain many duplicate probes for individual genes, and the highest value per gene probe was chosen. These data were joined with the BHOT panel and excluded non-renal parenchymal and viral genes. Data were renormalized using the housekeeping probes with negligible effect. Coefficients of variation were calculated and the lowest % CV including housekeeping genes were deleted, leaving 667 genes and 764 samples [26]. Data were then \log_2 transformed.

Software. Analyses were performed using SAS/JMP 14.2 using linear models with validation, principal components, multiple logistic regression, K-means clustering with determination of cluster number or Python 3.7 with the scik-learn module (Pycaret 2.0). UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) clustering was performed in JMP14.2/R4.0. Graphing was performed with Graph Builder (JMP 14.2). Batch normalization was performed with JMP Genomics 9.0. From power calculations (power = 0.8), significance was set at a False Discovery Rate Adjusted P-value of 0.005 ($-\log_{10} = 2.3$). This was also applied to the significance of any mean difference.

Parsing of Genes. Three methods of parsing genes were used to identify multiple principal components that were used to optimally partition the diagnostic groups. The first method, supervised, finds the highest gene expressions by linear models within two comparison groups, TCMR, ABMR, AKI, MIXED, or NORMAL as compared to NO REJECTION. These derived gene sets are heterogenous (collinear and non-parametric) as multiple principal components with eigen values from 3 to > 100 were derived from each derived binary gene set. These principal components are called Pathological Based Principal Components (PBPC) [1–9, 27]. The second method, semi-supervised, used genes from CIBERSORT LM22, Blood Atlas, and KEGG, and NanoString Annotations for a specific cell type or immunological pathway. This method was inspired by Nanostring Advanced Analysis software, in which “scores” are created using singular value decomposition, a sparse principal component, of genes that identify a cell type or immunological pathway. Care was taken to eliminate genes, which cross react in other cell types or pathways. The genes within a cell type or pathway created one principal component with an eigen value > 5 and are called Cell Pathway Principal Components (CPPC). The third method, unsupervised (agnostic), derives multiple principal components with eigen values from 3 to > 100, from all genes without regard to a class or diagnosis and called Unsupervised Principal Components (UPC). Principal component analysis was chosen for data reduction due to the massive collinearity of individual gene expressions.

Batch effects. UMAP was used to identify by clustering unknown anomalous effects (batch effects) in the archival data using genes with a coefficient of variation (CV) of $\leq 5\%$, which included the house keeping genes from BHOT. Genes with $CV \leq 5\%$ have little partitioning value. The three anomalous clusters were manually coded as categorical variables. Batch normalization was performed in JMP Genomics 9.2 using batch normalization.

Pathological diagnoses. Pathological diagnoses, derived from annotations of the downloaded databases, are categorical classes: T Cell Mediated Rejection (TCMR), Antibody Mediated Rejection (ABMR), Mixed (both TCMR and ABMR), NO REJECTION (NR), Acute Kidney Injury (AKI) which is defined as renal dysfunction unrelated to rejection and often occurring post transplantation, and Normal Native (NORMAL). These diagnostic classes are summary classes derived from the more complex classifications of renal allograft rejections per Banff classification schemes, which employ microscopic criteria, many of which cannot be evaluated, identified, or correlated with RNA expression [28, 29].

Definitions, abbreviations, and classification parameters are shown in supplemental data, Table 1A & B.

Results

Although the combined data were from the same array, batch effects can often skew data. It is unknown how many experiments were done to create the archived datasets, so that batch corrections cannot be done on individual experiments. To work around this problem UMAP clustering was performed on the genes with the lowest 5% coefficient of variation. Figure 1 shows a graph in which three clusters (A, B, & C, black arrows) were identified before batch correction. After batch normalization, one cluster remained (red

arrow). Such batch effects have a slight influence ($F = 0.02$) on classification accuracy (Pycaret classification, `compare_models` module), Supplemental Table 2, using all data and the target as DX (diagnosis). Batch normalized data was throughout.

Feature selection of the supervised (PBPC) and unsupervised (UPC) principal components using linear model, and active effects in multinomial logistic regression identified optimal sets of principal components that best partitioned the diagnostic groups with estimates and significance (Supplemental Table 3). Figure 2 shows a graph of the principal components (PCs) for the three different methods of gene selection. Pathologically based PCs (PBPCs, Fig. 2A) and the unsupervised PCs (UPC, Fig. 2C) readily partition the diagnostic groups. CPPC (Fig. 2B) identify immunologically interpretable patterns with a high PC for tubules in NORMALS, and a high endothelial PC in ABMR and MIXED but low in the other groups. Inflammatory cell types and mediators are highest in TCMR and MIXED rejections, known to contain inflammatory infiltrates and low in NORMALS and AKI.

The highest gene expressions per principal component (CPPC and PBPC) defined by PC loading tables were compared to the transcript patterns previously identified in microarrays called pathologically based transcripts (PBTs) [1–9, 22]. PBPC1 is dominated by genes identifying adaptive immunity, chemokine and cytokine signaling, cytotoxicity, T cell receptor signaling, toll-like receptor signaling, type 2 interferon, CD4 and CD8 T cells, and macrophages and found in PBTs (Type 2 interferon induced, cytotoxic lymphocyte induced, T cell, injury and injury repair transcripts). PBPC1, therefore, is an inflammatory signature that is associated with the inflammation commonly seen in TCMR and MIXED rejections. PBPC2, low in the AKI diagnoses, is low for genes in the cytokine signaling (JAK2) pathway, innate immunity, TH17 pathways, and toll-like receptor signaling and found in PBTs (late injury repair and type 2 interferon induced transcripts). PBPC3, high in AKI, identifies higher and different cytokines (CXCL13, 16, and CXCR6) and is found in PBTs (injury repair, endothelial, type 2 interferon induced, and decreased solute carrier transcripts). PBPC4, highest in ABMR and MIXED rejections, is dominated by the expression of endothelial genes and some CD4 cells and found in PBTs (endothelial and alloantibody induced transcripts), and is an endothelial pattern closely associated with antibody mediated rejections (ABMR and MIXED). PBPC5, high in AKI and TCMR, contains genes for B cells, complement, and innate immunity and is found in PBTs (B cell, macrophage, injury-related transcripts). PBPC6, lowest in the AKI and normal diagnoses is low for genes in innate immunity, type 2 interferon, and CD4s, and CD8 T cells and low in PBTs (injury related type 2 interferon inducible, T cell transcripts). PBPC7, is low for chemokine, T, B endothelial, and macrophage genes and low in PBTs (B cell, alloantibody induced, endothelial injury repair and type 2 interferon induced transcripts). PCPC8, highest in the normal diagnosis is high for glomerular, tubular, TH17 pathway, and tissue homeostasis genes and found in PBTs (solute carrier (high), alloantibody induced (low), endothelial (low), type 2 interferon induced (low)).

Within the unsupervised principal components UPC1, like PBPC1, identifies an inflammatory pattern highest in genes for adaptive immunity, chemokines, cytokines, cytotoxicity, innate immunity, toll-like receptor signaling, CD4 and CD8 T cells, and macrophages. UPC2, highest in NO REJECTION contains type 1 and 2 interferon related gene expressions, chemokine and cytokine, innate and toll-like receptor

related gene expressions without inflammatory cells and found in PBT interferon related transcripts. UPC3, highest in AKI, shows the highest gene patterns in cytokines, complement, innate immunity, oxidative stress, without markers for B, T, or macrophage cells and is found in PBTs (interferon and injury repair transcripts and solute carrier (low)). UPC4, high in ABMR and MIXED, identifies an endothelial pattern with many endothelial genes, adaptive immunity, chemokines, complement, cytokines, B cell, CD4, CD8, macrophage genes without any cytotoxicity signals and is found in PBTs (endothelial and alloantibody induced, B cell, type 2 interferon). UPC5, highest in the normal diagnosis includes gene signals for glomeruli, tubules, some innate, oxidative, TH17, TNF without type 2 interferon, plasma, CD4, CD8, or macrophage cells and is found in PBTs (high tubular, high endothelial, injury repair). UPC6, high in AKI, shows the greatest number of genes in adaptive, cytokine, complement, innate, and CD8, and macrophages and is found in PBTs (type 2 interferon, injury related, macrophage related transcripts). UPC7, highest in NO REJECTION, is highest for genes in chemokine, cytokine, innate, oxidative, TH17 pathways, with some markers for B, CD4, CD8, and macrophage cells and is found in PBTs (injury related, type 2 interferon related transcripts).

All the genes within these principal components (PBPC and CPPC) are described in many prior publications on gene expression in renal allografts [1–9, 22], thereby confirming the utility of the BHOT panel as a surrogate for microarrays.

Modelling. Modelling is used to estimate how good variables describe the classification parameters, in this case, how accurately the principal components identify the diagnoses. Modelling programs assign a class or in this case a diagnosis to each sample based on the highest probability within the target diagnosis for a specific sample. Initially, multinomial logistic regression (JMP14.2) modelled the data (CPPC, PBPC, UPC and all three (ALL)) with the target as the diagnosis (DX). All four models created acceptable ROC curves, Fig. 3A, B, C, & D. However, the errors in the confusion matrices were substantial 43–60%, indicating that modelling poorly matched the DX, Table 1A. Also, problematic, the per sample error of 46.9% indicates per sample discrepancies showing that subtle variations in the data type engender different patterns of errors matching the DX. Reeve et.al. also, identified variations in misclassifications when clustering using archetypal analysis as compared to annotated diagnoses [30]. Individual samples are misclassified differently depending on the data set.

The classification assignment is derived from the highest probability per group, whether the highest probability is below or above 50%. The average probabilities were examined in misclassified and concordant samples, Fig. 4. Misclassified samples have lower average probabilities as compared to concordant samples suggesting that samples without a consistently high probability cause per sample variations in the error matrices [30].

One possibility for the complexity of modelling as shown by the high per sample misclassification patterns is that the annotated diagnosis is not a pure category and that gene expression heterogeneity exists within samples of a diagnostic class. To explore this, distributions were analyzed for all principal components by diagnosis. Figure 5 shows the distributions of all the principal components from the three

data sets (CPPC, UPC, and PBPC) for the six diagnostic classes for each sample (grey lines) and a group mean (black line). The grey lines show wide distribution patterns within a diagnosis, and the group mean shows biphasic distributions for the diagnoses AKI, TCMR, and MIXED, which are, therefore, mixtures of distributions. The red vertical line is the modal average of NORMAL. Vast heterogeneity is evident for all diagnoses.

To try to reduce the heterogeneity across samples, the samples were clustered using K Means, resulting in 8 clusters. Table 2 is a contingency table of DX vs clusters. NORMAL and AKI cluster similarly to DX, but the other diagnoses are widely distributed among the clusters, especially NO REJECTION. Again, using multinomial logistic regression (JMP14.2) modelled the data (CPPC, PBPC, UPC and all three (ALL)) with the target as the Clusters. Table 1B shows that the percent misclassification from the confusion matrices dropped to 6–17%, a dramatic improvement as compared to using the DX as the target, Table 1A. However, the per sample errors in misclassifications were 27.9%, which were much lower than 46.9% using the DX as the target.

To further refine and optimize the prior models, Pycaret, which uses Python and sciklearn modules, was used to test additional models. The Pycaret compare_models module permits comparison of multiple models to find the most optimal model by accuracy. The models tested are found in the supplemental Table 1A. The best model with the highest accuracy was also tested with the tuning, bagging, boosting, and blending with negligible improvements in accuracy (data not shown). Making an ensemble model of the top three models also did not show any improvement in accuracy. The optimal model for the data (All, CPPC, PBPC, UPC) with targets as DX and Clusters is shown in Table 3. Modeling all principal components (All) or the individual sets (CPPC, PBPC, and UPC) vs the target DX showed similar accuracies and precision with confusion matrix errors of 33–37% with per sample error rates of about 15%. Dramatic improvements in accuracies ($> .9$) was shown using Clusters as the target with dramatic reductions in the error rates.

Discussion

Findings show that the BHOT panel of genes recapitulates the diagnostic patterns from the archival data using these archival data sets (Fig. 2, supplementary Table 2). As the selected BHOT panel genes are derived from many microarray studies, it is not surprising that BHOT panel genes identify the expected patterns of rejection. Nevertheless, such *in silico* confirmation will encourage investigators to use this panel as it is easier than microarrays. All three methods of parsing genes created workable models with high average ROC scores. It is unclear which method of parsing the genes into principal components is easiest or is most suitable to create an efficient and standardized data analysis pipeline. Using PCs (PBPC) from sets of the highest genes between binary diagnostic comparisons is conceptually simple but engenders many principal components, which share collinearity and make feature selection for modelling both tedious and difficult. Using principal components of cell types and pathways is conceptually easier to understand immunologically. Creating unsupervised principal components is the easiest for feature selection and has an advantage that a latent variable or pathway may be present, which is not readily

identified by the first two gene selection methods [18, 19]. These three methods, including just finding the highest genes by t-tests, will likely vary between independently derived data sets because results are very dependent on the sample size of the data set, the balance of the classes employed, and the purity of the annotated class diagnosis.

Some investigators argue that gene expression models assign a more accurate diagnosis than the original pathological diagnosis and use such testing for clinical diagnosis. However, heterogeneity is present in the misclassification assignments per sample by different models [30] or data sets (CPPC, UPC, PBPC), (Table 1, Table 3) engendering discordant patterns of misclassifications. Therefore, changing the sample diagnosis by modelling may be premature. Model averaging or an ensemble of models does not solve this problem as a new error matrix is created, which still maintains the per sample variations in classifications. Arbitrarily, using only a high probability to assign a diagnostic classification solves part of the misclassification problem by reducing some misclassifications but limits model utility as many samples could remain unclassified [30]. To improve assignment of diagnoses of ambiguously modelled samples, additional clinical information such as some histological parameters, alloantibody, C4d, or time after transplant, could be incorporated (expert knowledge of prior probabilities) with gene expression to create a clinical pathological diagnosis [30, 31]. Although using such expert knowledge may allow assignment of some samples to a diagnosis, and make overall interpretation easier, interpreting such heterogeneous variables, absent in the model, is subjective and may work for some samples but not all.

Although clustering data independent of diagnosis makes a better model with fewer misclassifications, interpreting individual clusters remains problematic. It is better to find the best model for the data rather than find the best data for a model. Are these synthetic clusters just “toy” data, that models well but has no biological relevance? Some clusters resemble canonical diagnoses, but others do not. How do you assign a meaningful and clinically interpretable name to synthetic cluster? Nevertheless, creating more homogeneous groups of samples may identify clinically important groups of samples. Clustering expression rejection data can identify novel subgroups, not appreciated in the annotated classes [19, 27]. This is most important in the NO REJECTION diagnosis, which is the most heterogeneous by gene expression (Table 2) and the most frequent diagnosis. This diagnosis lacks evidence of rejection, and subjects usually have a preserved creatinine, yet the gene expression pattern within the NO REJECTION diagnosis is markedly heterogeneous as indicated by clustering. If some gene expression patterns in the NO REJECTION diagnostic category correlate with a subsequent clinical rejection or correlate with renal outcome, then analysis of gene expressions adds value to clinical decisions.

The gene expression data are heterogeneous within the original diagnostic classes because clustering all the principal components creates a better model with fewer misclassifications. This is mostly likely because pathological diagnoses are complex and critically dependent on microscopic findings that cannot be identified within a mixture of extracted RNAs. For example, tubulitis, which is mononuclear inflammation identified within tubules, is required for a diagnosis of TCMR but cannot be captured in a slurry of RNA. In addition, many of the Banff histological lesions used or required for allograft diagnoses

are also somewhat non-specific [10, 23, 32]. In addition, the rejection classes (TCMR, ABMR, and MIXED) used to categorize the data are summaries of rejection patterns, which have grades from clinically mild to clinically very serious, so that the annotated class diagnoses are heterogeneous and mere summary approximations.

Modelling creates additional challenges for investigators, who wish to use gene expression to inform diagnostic decisions. Diagnostic changes cannot be made solely by gene expression as different modelling algorithms or modelling slightly different sets of principal components do not uniformly assign a consistent pattern of per sample diagnosis. Additional information as covariates might include creatinine trajectory, proteinuria, time after transplant, prior diagnoses, or histological parameters might improve or alter modelling performance [33, 34]. This is problematic as investigators using different models or slightly different data sets could reassign diagnoses discordantly. Uncertainties could also arise in clinical trials using allograft transplant gene expression for classifications if contributors to the clinical trial assign variant gene expression classifications, depending on how the genes are analyzed or modelled. This problem also applies to aligning disparate studies investigating a similar hypothesis.

Creating additional categories by clustering more diagnostic categories, which makes modelling more consistent and lowers misclassification rates. But it is unclear if these additional categories represent biologically relevant diagnostic classes, or if they represent inconsequential minor variants, biopsy sampling error or evolving forms of established diagnoses. Only correlation of pathological diagnoses and gene expression patterns with the endpoint of renal allograft survival or subsequent rejections can resolve such discrepancies and identify the optimal and biologically relevant classes for clinical decision-making. This is likely best done by longitudinal analysis of a patient's samples.

Conclusion

In conclusion, these findings confirm the BHOT gene panel is a useful surrogate for microarrays to identify the expected gene expression patterns in human allografts and would encourage this easier gene expression method. These findings also confirm the complexity of modelling gene expressions and suggest that reassigning a diagnosis based solely on gene expression is not straightforward and possibly premature. This creates challenges for the application of gene expression for the classification of an individual patient's specimen for clinical decision making. Future analytical challenges facing investigators, who desire to use transplant gene expressions for clinical care include: 1) How and which genes are best and most efficiently parsed to create an efficient data analysis pipeline; 2) how is modelling is best performed to assign a diagnostic probability to a patient's sample; 3) what clinical and pathological parameters improve model performance; 4) how to resolve the heterogeneity of gene expression and pathological diagnoses into more homogeneous groups that permit the most accurate modelling and immunological interpretation, and, finally, determine if new and more homogeneous classes are biologically relevant.

Declarations

Competing interest

The author has no competing financial interest, including any with Nanostring Technologies

Funding

None

Declarations

Code available. Data in public domain.

Acknowledgments

The author thanks Drs. R. Abdi, J. Paik, V. Klepeis, A. Loussaint, C. Smith, and R. Colvin for their contributions.

References

1. Einecke G, Broderick G, Sis B, Halloran PF. Early loss of renal transcripts in kidney allografts: relationship to the development of histologic lesions and alloimmune effector mechanisms. *Am J Transplant.* 2007;7(5):1121–30.
2. Einecke G, Reeve J, Mengel M, Sis B, Bunnag S, Mueller TF, Halloran PF. Expression of B cell and immunoglobulin transcripts is a feature of inflammation in late allografts. *Am J Transplant.* 2008;8(7):1434–43.
3. Einecke G, Sis B, Reeve J, Mengel M, Campbell PM, Hidalgo LG, Kaplan B, Halloran PF. Antibody-mediated microcirculation injury is the major cause of late kidney transplant failure. *Am J Transplant.* 2009;9(11):2520–31.
4. Famulski KS, de Freitas DG, Kreepala C, Chang J, Sellares J, Sis B, Einecke G, Mengel M, Reeve J, Halloran PF. Molecular phenotypes of acute kidney injury in kidney transplants. *J Am Soc Nephrol.* 2012;23(5):948–58.
5. Famulski KS, Einecke G, Reeve J, Ramassar V, Allanach K, Mueller T, Hidalgo LG, Zhu LF, Halloran PF. Changes in the transcriptome in allograft rejection: IFN-gamma-induced transcripts in mouse kidney allografts. *Am J Transplant.* 2006;6(6):1342–54.
6. Famulski KS, Einecke G, Sis B, Mengel M, Hidalgo LG, Kaplan B, Halloran PF. Defining the canonical form of T-cell-mediated rejection in human kidney transplants. *Am J Transplant.* 2010;10(4):810–20.
7. Hidalgo LG, Einecke G, Allanach K, Halloran PF. The transcriptome of human cytotoxic T cells: similarities and disparities among allostimulated CD4(+) CTL, CD8(+) CTL and NK cells. *Am J Transplant.* 2008;8(3):627–36.
8. Hidalgo LG, Sellares J, Sis B, Mengel M, Chang J, Halloran PF. Interpreting NK cell transcripts versus T cell transcripts in renal transplant biopsies. *Am J Transplant.* 2012;12(5):1180–91.

9. Hidalgo LG, Sis B, Sellares J, Campbell PM, Mengel M, Einecke G, Chang J, Halloran PF. NK cell transcripts and NK cells in kidney biopsies from patients with donor-specific antibodies: evidence for NK cell involvement in antibody-mediated rejection. *Am J Transplant*. 2010;10(8):1812–22.
10. Halloran PF, Chang J, Famulski K, Hidalgo LG, Salazar ID, Merino Lopez M, Matas A, Picton M, de Freitas D, Bromberg J, et al. Disappearance of T Cell-Mediated Rejection Despite Continued Antibody-Mediated Rejection in Late Kidney Transplant Recipients. *J Am Soc Nephrol*. 2015;26(7):1711–20.
11. Halloran PF, Pereira AB, Chang J, Matas A, Picton M, De Freitas D, Bromberg J, Seron D, Sellares J, Einecke G, et al. Microarray diagnosis of antibody-mediated rejection in kidney transplant biopsies: an international prospective study (INTERCOM). *Am J Transplant*. 2013;13(11):2865–74.
12. Halloran PF, Merino Lopez M, Barreto Pereira A. Identifying Subphenotypes of Antibody-Mediated Rejection in Kidney Transplants. *Am J Transplant*. 2016;16(3):908–20.
13. Einecke G, Kayser D, Vanslambrouck JM, Sis B, Reeve J, Mengel M, Famulski KS, Bailey CG, Rasko JE, Halloran PF. Loss of solute carriers in T cell-mediated rejection in mouse and human kidneys: an active epithelial injury-repair response. *Am J Transplant*. 2010;10(10):2241–51.
14. Venner JM, Famulski KS, Reeve J, Chang J, Halloran PF. Relationships among injury, fibrosis, and time in human kidney transplants. *JCI Insight*. 2016;1(1):e85323.
15. Haas M, Sis B, Racusen LC, Solez K, Glotz D, Colvin RB, Castro MCR, David DSR, David-Neto E, Bagnasco SM, et al. Banff 2013 meeting report: inclusion of C4d-negative antibody-mediated rejection and antibody-associated arterial lesions. *Am J Transplant*. 2014;14(2):272–83. PMID: 24472190.
16. Sis B, Jhangri GS, Bunnag S, Allanach K, Kaplan B, Halloran PF. Endothelial gene expression in kidney transplants with alloantibody indicates antibody-mediated damage despite lack of C4d staining. *Am J Transplant*. 2009;9(10):2312–23.
17. Yazdani S, Callemeyn J, Gazut S, Lerut E, de Loor H, Wevers M, Heylen L, Saison C, Koenig A, Thauinat O, et al. Natural killer cell infiltration is discriminative for antibody-mediated rejection and predicts outcome after kidney transplantation. *Kidney Int*. 2019;95(1):188–98.
18. Smith RN, Adam BA, Rosales IA, Matsunami M, Oura T, Cosimi AB, Kawai T, Mengel M, Colvin RB. RNA expression profiling of renal allografts in a nonhuman primate identifies variation in NK and endothelial gene expression. *Am J Transplant*. 2018;18(6):1340–50.
19. Smith RN, Matsunami M, Adam BA, Rosales IA, Oura T, Cosimi AB, Kawai T, Mengel M, Colvin RB. RNA expression profiling of nonhuman primate renal allograft rejection identifies tolerance. *Am J Transplant*. 2018;18(6):1328–39.
20. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol*. 2008;26(3):317–25.
21. Mengel M, Loupy A, Haas M, Roufosse C, Naesens M, Akalin E, Clahsen-van Gronigen MC, Dagobert J, Duong Van Huyen JP, Gueguen J, et al: **Banff 2019 Meeting Report: Molecular diagnostics in solid**

organ transplantation – Consensus for the Banff Human Organ Transplant (B-HOT) gene panel and open source multicenter validation. *American Journal of Transplantation* 2020:in press.

22. Alberta-Transplant-Applied-Genomics-Centre: **Gene Lists: Core PBT List (HG U133_2.0_Plus Arrays U133 Core PBTs.xlsx).**<https://www.ualberta.ca/medicine/institutes-centres-groups/atagc/research/gene-lists>. 2019.
23. Haas M. Molecular diagnostics in renal allograft biopsy interpretation: potential and pitfalls. *Kidney Int.* 2014;86(3):461–4.
24. Famulski KS, Reeve J, de Freitas DG, Kreepala C, Chang J, Halloran PF. Kidney transplants with progressing chronic diseases express high levels of acute kidney injury transcripts. *Am J Transplant.* 2013;13(3):634–44.
25. Eskandary F, Bond G, Regele H, Kozakowski N, Kikic Z, Wahrmann M, Haslacher H, Oberbauer R, Ramassar V, Halloran P, et al: **Late Antibody-Mediated Rejection in a Large Prospective Cross-Sectional Study of Kidney Allograft Recipients–Preliminary Results of the Screening Phase of the BORTEJECT Trial.** *Clin Transpl* 2014:189–195.
26. Wang Z, Lyu Z, Pan L, Zeng G, Randhawa P. Defining housekeeping genes suitable for RNA-seq analysis of the human allograft kidney biopsy tissue. *BMC Med Genomics.* 2019;12(1):86.
27. Reeve J, Bohmig GA, Eskandary F, Einecke G, Lefaucheur C, Loupy A, Halloran PF. group MM-Ks: **Assessing rejection-related disease in kidney transplant biopsies based on archetypal analysis of molecular phenotypes.** *JCI Insight* 2017, 2(12).
28. Roufosse C, Simmonds N, Clahsen-van Groningen M, Haas M, Henriksen KJ, Horsfield C, Loupy A, Mengel M, Perkowska-Ptasinska A, Rabant M, et al. A 2018 Reference Guide to the Banff Classification of Renal Allograft Pathology. *Transplantation.* 2018;102(11):1795–814.
29. Solez K, Colvin RB, Racusen LC, Haas M, Sis B, Mengel M, Halloran PF, Baldwin W, Banfi G, Collins AB, et al. Banff 07 classification of renal allograft pathology: updates and future directions. *Am J Transplant.* 2008;8(4):753–60.
30. Reeve J, Bohmig GA, Eskandary F, Einecke G, Gupta G, Madill-Thomsen K, Mackova M, Halloran PF. Group IM-KS: **Generating automated kidney transplant biopsy reports combining molecular measurements with ensembles of machine learning classifiers.** *Am J Transplant.* 2019;19(10):2719–31.
31. Madill-Thomsen K, Perkowska-Ptasinska A, Bohmig GA, Eskandary F, Einecke G, Gupta G, Halloran PF. Group MM-KS: **Discrepancy analysis comparing molecular and histology diagnoses in kidney transplant biopsies.** *Am J Transplant.* 2020;20(5):1341–50.
32. Amrouche L, Aubert O, Suberbielle C, Rabant M, Van Huyen JD, Martinez F, Sberro-Soussan R, Scemla A, Tinel C, Snanoudj R, et al: **Long-term outcomes of kidney transplantation in patients with high levels of preformed DSA: the Necker high-risk transplant program.** *Transplantation* 2017.
33. Einecke G, Reeve J, Gupta G, Bohmig G, Eskandary F, Bromberg J, Budde K, Halloran P. Factors associated with kidney graft survival in pure antibody-mediated rejection at the time of indication

biopsy: Importance of parenchymal injury but not disease activity. Am J Transplant. 2020. <https://doi.org/10.1111/ajt.16161>.

34. Loupy A, Aubert O, Orandi BJ, Naesens M, Bouatou Y, Raynaud M, Divard G, Jackson AM, Viglietti D, Giral M, et al. Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study. BMJ. 2019;366:l4923.

Tables

Table 1

SUMMARY OF ERROR IN CONFUSION MATRICES

A	Data	Target	Percent Errors	Percent per Sample Errors
	All	DX	42.8	
	CPPC	DX	48	46.9
	PBPC	DX	59.6	
	UPC	DX	54	
B	Data	Target	Percent Errors	Percent per Sample Errors
	All	Clusters	5.9	
	CPPC	Clusters	10.4	27.9
	PBPC	Clusters	8.6	
	UPC	Clusters	16.6	

**Percent Errors = Percent Errors
from Confusion Matrix**

**Percent Sample Errors = Percent of Samples
with a Single Discordance between
Confusion Matrices of a Target**

**From multinomial logistic regression, General Regression
Model, JMP 14.2. Parameters in Dictionary**

TABLE 2

CONTINGENCY TABLE DIAGNOSES VS CLUSTER

		Cluster									Total
		1	2	3	4	5	6	7	8		
DX	ABMR	Number	3	0	32	34	8	0	5	23	105
		Percent	2.9%	0.0%	30.5%	32.4%	7.6%	0.0%	4.8%	21.9%	
DX	AKI	Number	0	0	0	0	0	0	4	24	28
		Percent	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	14.3%	85.7%	
DX	MIXED	Number	5	0	3	13	4	0	2	1	28
		Percent	17.9%	0.0%	10.7%	46.4%	14.3%	0.0%	7.1%	3.6%	
DX	NOREJECTION	Number	12	4	190	56	28	4	32	188	514
		Percent	2.3%	0.8%	37.0%	10.9%	5.4%	0.8%	6.2%	36.6%	
DX	NORMAL	Number	0	0	0	0	0	22	0	0	22
		Percent	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	
DX	TCMR	Number	17	2	12	8	20	0	0	8	67
		Percent	25.4%	3.0%	17.9%	11.9%	29.9%	0.0%	0.0%	11.9%	
Total										764	

TABLE 3

Additional PyCaret Models

A	Data	Target	Model	Mean Accuracy	Ave Precision/Recall	Percent Errors	Percent Sample Errors
	All	DX	Extreme Gradient Boosting	0.74	0.79	36.5	
	CPPC	DX	Linear Discriminant Analysis	0.75	0.78	32.6	15.2
	PBPC	Dx	Extra Trees Classifier	0.75	0.76	35.3	
	PC	DX	Linear Discriminant Analysis	0.76	0.70	33.9	
B	Data	Target	Model	Mean Accuracy	Ave Precision/Recall	Percent Errors	Percent Sample Errors
	All	Clusters	Logistic Regression	0.93	0.95	6.9	
	CPPC	Clusters	Extra Trees Classifier	0.90	0.92	6.9	5.1
	PBPC	Clusters	Extreme Gradient Boosting	0.92	0.94	6.2	
	PC	Clusters	Logistic Regression	0.92	0.94	6.9	

Figures

Figure 1

UMAP CLUSTERING BEFORE AND AFTER BATCH NORMALIZATION

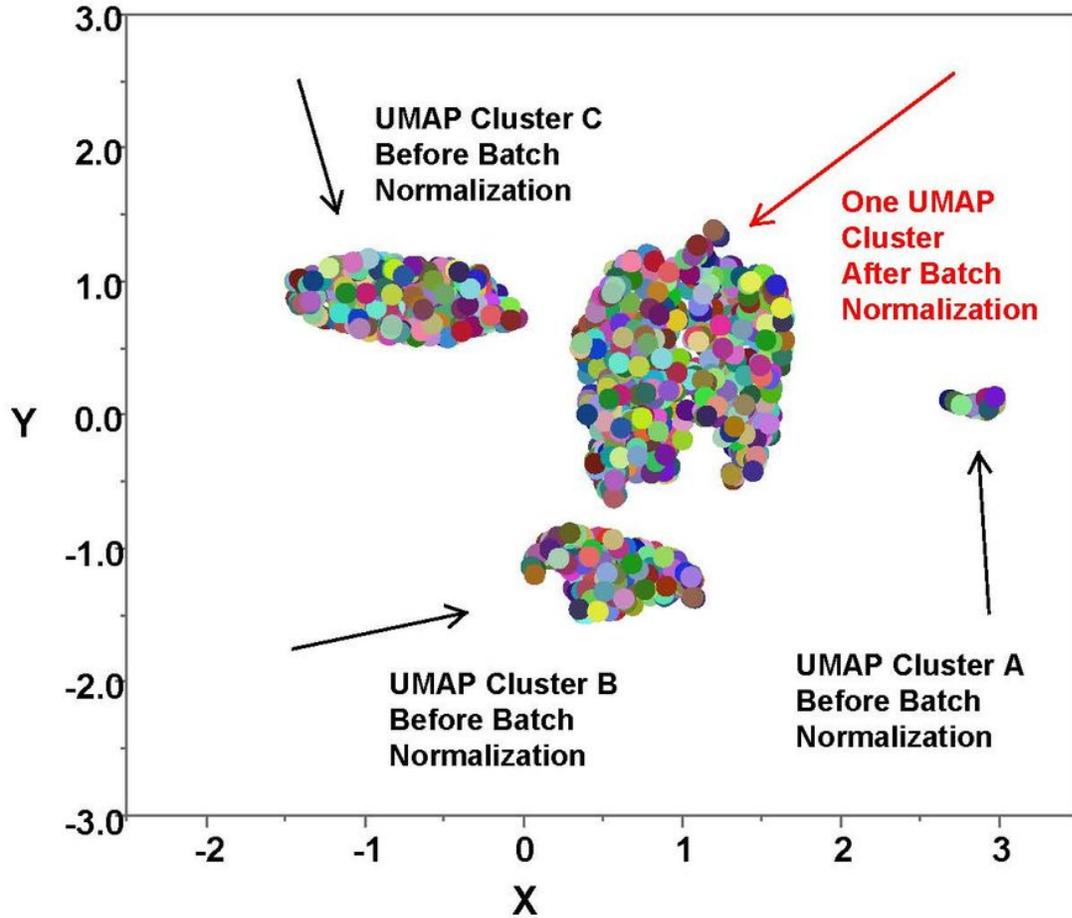


Figure 1

UMAP Clustering. UMAP clustering of data identified three groups, A, B, & C, each indicated by a black arrow. After batch normalization in JMP Genomics 9, repeat UMAP clustering created one group, red arrow. Each differently colored circle is one sample.

Figure 2

FIGURE 2 A

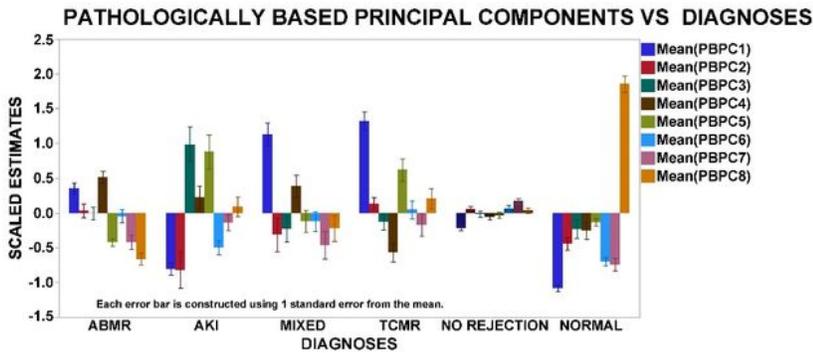


FIGURE 2 B

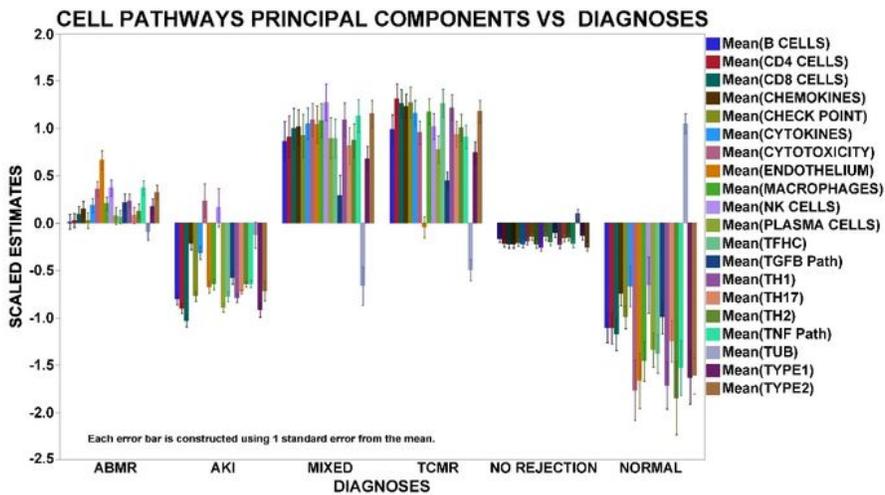


FIGURE 2 C

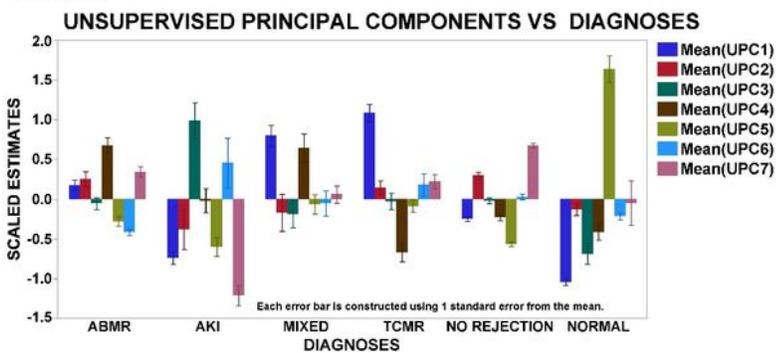
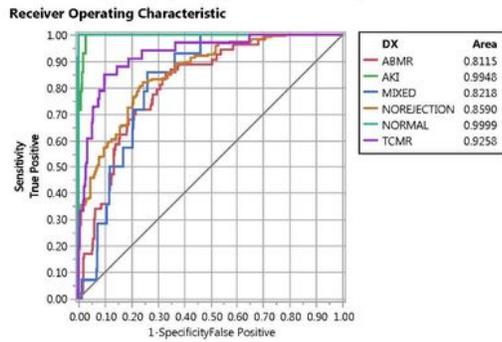


Figure 2

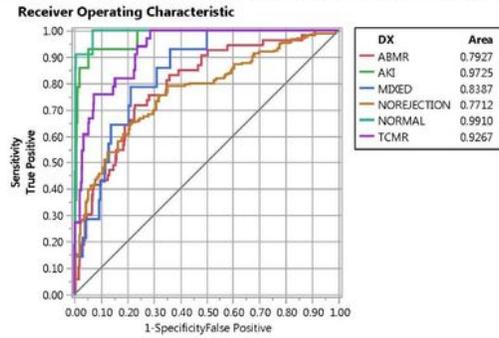
Scaled estimates vs Diagnosis for each of the three types of derived principal components: Pathologically Based (PBPC); Cell Pathways (CPPC); and unsupervised (UPC). Error bar is one SEM.

Figure 3

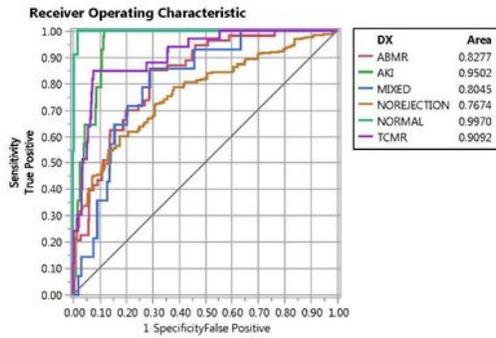
A Validation Multinomial Logistic Regression All Data Target= Diagnosis (DX)



B Validation Multinomial Logistic Regression CPPC Data Target= Diagnosis (DX)



C Validation Multinomial Logistic Regression PBPC Data Target= Diagnosis (DX)



D Validation Multinomial Logistic Regression UPC Data Target= Diagnosis (DX)

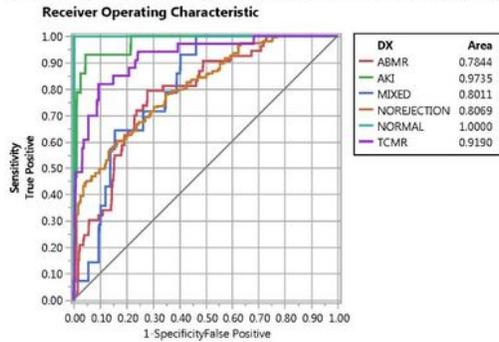


Figure 3

ROC graphs of validated multinomial logistic regression. Target is diagnosis (DX). A: Data = All PC (CPPC, PBPC, & UPC); B: Data = CPPC; C: Data = PBPC; D: Data = UPC.

FIGURE 4

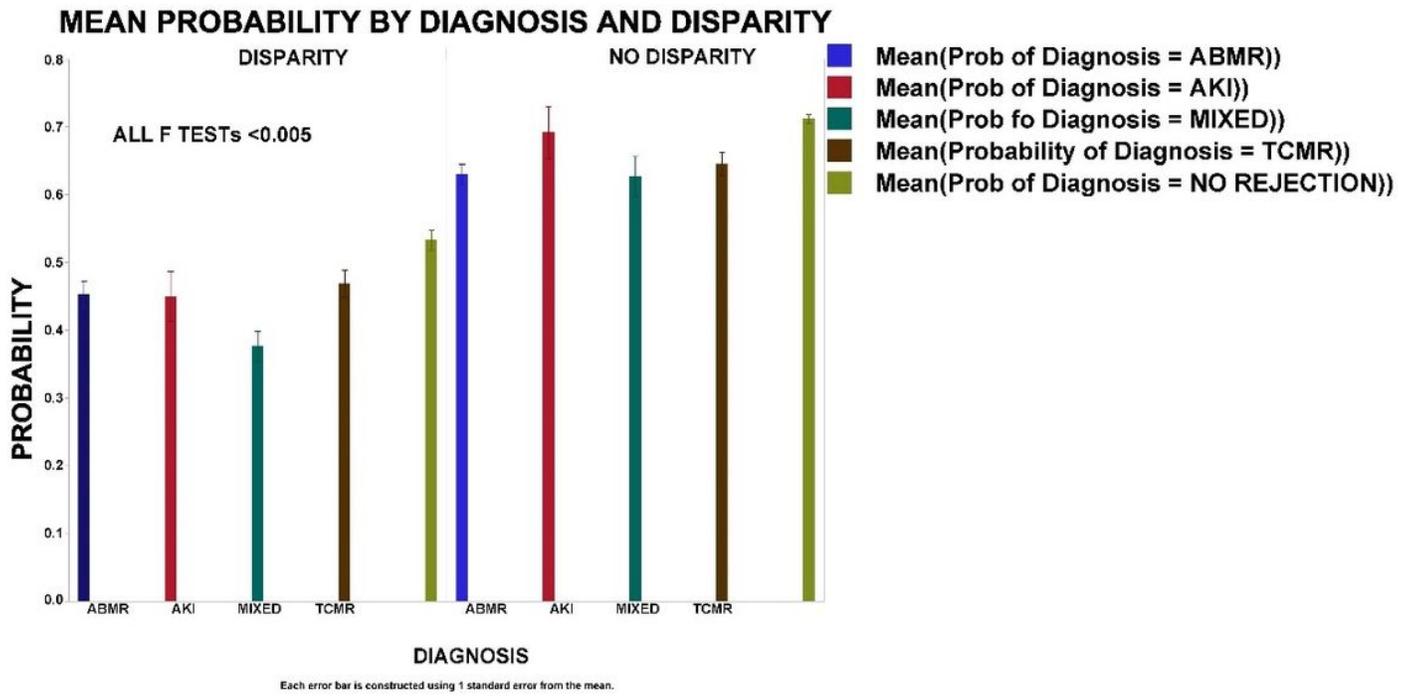


Figure 4

Mean Probability vs Disparity. Using the regression in 3A, the regression creates a column for the probability for each of the six diagnoses for each sample. The highest probability is assigned as most likely DX. The disparities between the annotated DX and the most likely DX creates the confusion matrix. No disparity: DX = most likely Dx. Disparity: DX does not equal most likely DX.

FIGURE 5

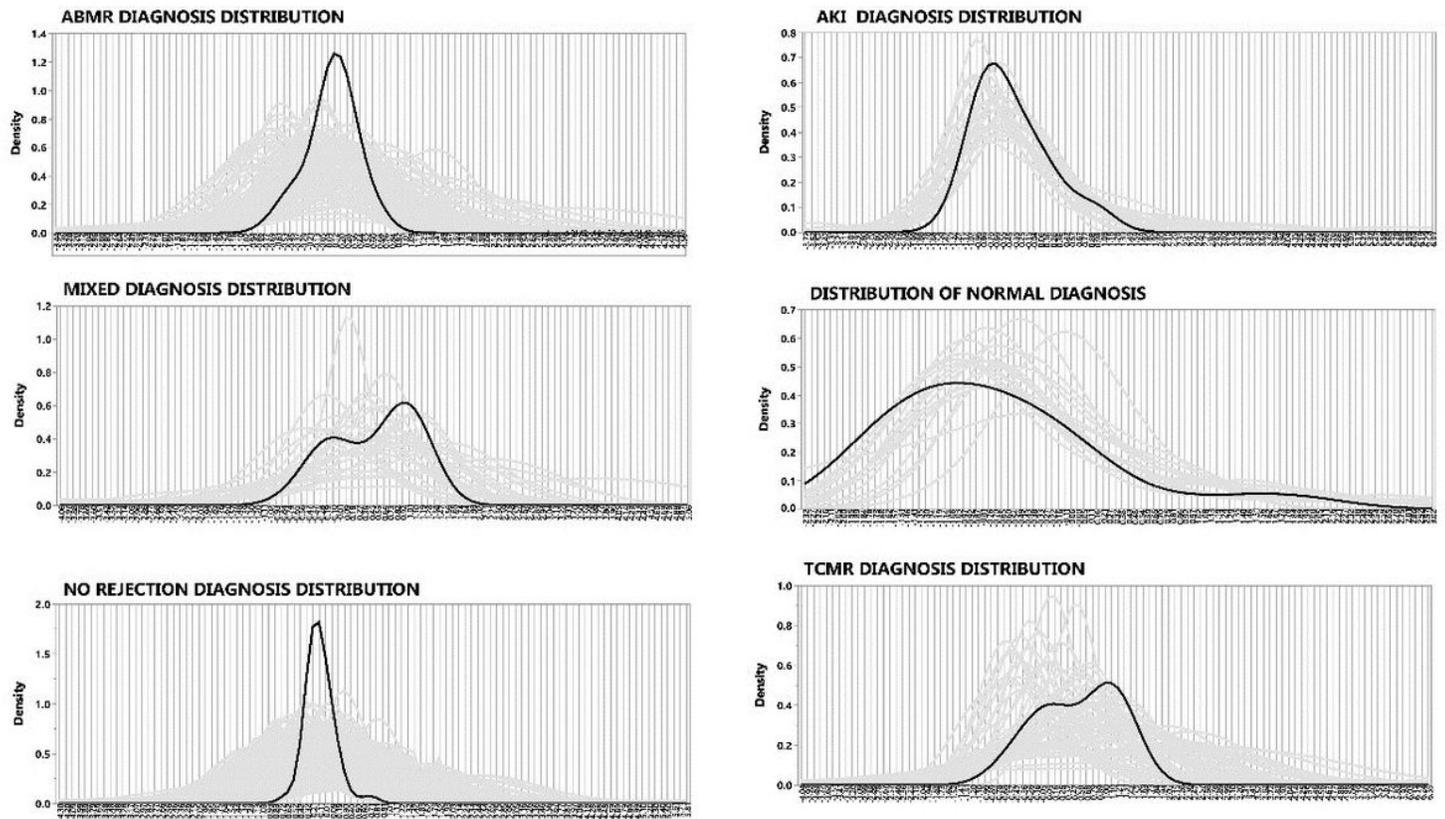


Figure 5

Distribution of all PC (CPPC, PBPC, & UPC) for each of the six annotated diagnoses.