JCI insight

Assessing rejection-related disease in kidney transplant biopsies based on archetypal analysis of molecular phenotypes

Jeff Reeve, … , Philip F. Halloran, the MMDx-Kidney study group

JCI Insight. 2017;2(12):e94197. https://doi.org/10.1172/jci.insight.94197.

Clinical Medicine | Nephrology | Transplantation

Conventional histologic diagnosis of rejection in kidney transplants has limited repeatability due to its inherent requirement for subjective assessment of lesions, in a rule-based system that does not acknowledge diagnostic uncertainty. Molecular phenotyping affords opportunities for increased precision and improved disease classification to address the limitations of conventional histologic diagnostic systems and quantify levels of uncertainty. Microarray data from 1,208 kidney transplant biopsies were collected prospectively from 13 centers. Cross-validated classifier scores predicting the presence of antibody-mediated rejection (ABMR), T cell–mediated rejection (TCMR), and 5 related histologic lesions were generated using supervised machine learning methods. These scores were used as input for archetypal analysis, an unsupervised method similar to cluster analysis, to examine the distribution of molecular phenotypes related to rejection. Six archetypes were generated: no rejection, TCMR, 3 associated with ABMR (early-stage, fully developed, and late-stage), and mixed rejection (TCMR plus early-stage ABMR). Each biopsy was assigned 6 scores, one for each archetype, representing a probabilistic assessment of that biopsy based on its rejection-related molecular properties. Viewed as clusters, the archetypes were similar to existing histologic Banff categories, but there was 32% disagreement, much of it probably reflecting the […]

JCI insight

Assessing rejection-related disease in kidney transplant biopsies based on archetypal analysis of molecular phenotypes

Jeff Reeve,[1,2] Georg A. Böhmig,[3] Farsad Eskandary,[3] Gunilla Einecke,[4] Carmen Lefaucheur,[5,6] Alexandre Loupy,[5,7] Philip F. Halloran,[1,8] and the MMDx-Kidney study group[9]

[1]Alberta Transplant Applied Genomics Centre, Alberta, Canada. [2]Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, Alberta, Canada. [3]Division of Nephrology and Dialysis, Department of Medicine III, Medical University of Vienna, Vienna, Austria. [4]Department of Nephrology, Medizinische Hochschule Hannover, Hannover, Germany. [5]Paris Translational Research Center for Organ Transplantation, INSERM, UMR-S970, Paris, France. [6]Saint-Louis Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France. [7]Necker Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France. [8]Department of Medicine, Division of Nephrology and Transplant Immunology, University of Alberta, Edmonton, Alberta, Canada. [9]The MMDx-Kidney study group is detailed in the Supplemental Acknowledgments.

Conventional histologic diagnosis of rejection in kidney transplants has limited repeatability due to its inherent requirement for subjective assessment of lesions, in a rule-based system that does not acknowledge diagnostic uncertainty. Molecular phenotyping affords opportunities for increased precision and improved disease classification to address the limitations of conventional histologic diagnostic systems and quantify levels of uncertainty. Microarray data from 1,208 kidney transplant biopsies were collected prospectively from 13 centers. Cross-validated classifier scores predicting the presence of antibody-mediated rejection (ABMR), T cell–mediated rejection (TCMR), and 5 related histologic lesions were generated using supervised machine learning methods. These scores were used as input for archetypal analysis, an unsupervised method similar to cluster analysis, to examine the distribution of molecular phenotypes related to rejection. Six archetypes were generated: no rejection, TCMR, 3 associated with ABMR (early-stage, fully developed, and late-stage), and mixed rejection (TCMR plus early-stage ABMR). Each biopsy was assigned 6 scores, one for each archetype, representing a probabilistic assessment of that biopsy based on its rejection-related molecular properties. Viewed as clusters, the archetypes were similar to existing histologic Banff categories, but there was 32% disagreement, much of it probably reflecting the "noise" in the current histologic assessment system. Graft survival was lowest for fully developed and late-stage ABMR, and it was better predicted by molecular archetype scores than histologic diagnoses. The results provide a system for precision molecular assessment of biopsies and a new standard for recalibrating conventional diagnostic systems.

Role of funding source: We acknowledge the support of the Industrial Research Assistance Program. This research has been supported by funding and/or resources from University Hospital Foundation at the University of Alberta, Genome Canada, Canada Foundation for Innovation, and Roche Organ Transplant Research Foundation.

Conflict of interest: PFH holds shares in Transcriptome Sciences Inc., a company with an interest in molecular diagnostics.

Submitted: March 28, 2017
Accepted: May 5, 2017
Published: June 15, 2017

Reference information:
JCI Insight. 2017;2(12):e94197. https://doi.org/10.1172/jci.insight.94197.

Introduction

The drive to develop precision diagnostics arises from the acknowledged limitations of conventional diagnostic systems and disease classifications. This is illustrated by the diagnosis of rejection in kidney transplant biopsies, which is based on histologic lesions interpreted by empirically derived guidelines moderated by Banff consensus (1). There are 2 mechanisms of rejection: antibody-mediated (ABMR) and T cell–mediated (TCMR). The canonical lesions of TCMR are interstitial inflammation (i-score) and tubulitis (t-score), and those of ABMR are peritubular capillaritis (ptc-score), glomerulitis (g-score), and glomerular double contours (cg-score); both ABMR and TCMR can induce endothelial arteritis (v-score). Rejection diagnoses are defined by arbitrary thresholds and consideration of numerous rules and exceptions (1). Both the lesions and the diagnoses have limited reproducibility between pathologists (2, 3), and the current histology classes have borderline and ABMR-suspicious categories that leave

Table 1. Demographics and clinical features[A]

Patient Demographics (*n* = 1045)		All patients
Mean recipient age (range)		52 (18–86)
Recipient sex (% male)		559 (53%)
Ethnicity	European descent	522 (50%)
	Black	83 (8%)
	Other	100 (10%)
	Not available	340 (33%)
Primary Disease	Diabetic nephropathy	156 (15%)
	Hypertension/large vessel disease	54 (5%)
	Glomerulonephritis/vasculitis	37 (4%)
	Interstitial nephritis/pyelonephritis	23 (2%)
	Polycystic kidney disease	107 (10%)
	Others	468 (45%)
	Unknown etiology	200 (19%)
Mean donor age (range)		43 (0.03–85)
Donor sex (% male)		345 (49%)
Donor type (% deceased donor transplants)		692 (66%)
Clinical characteristics at time of biopsy (*n* = 1,208)		**All biopsies**
Median time of biopsy after transplant (TxBx) in days (range)		591 (1–11453)
Early biopsies (< 1 year)		507 (42%)
Late biopsies (≥ 1 year)		701 (58%)
Maintenance immunosuppression at biopsy, if recorded	Tacrolimus	712 (59%)
	Cyclosporine	192 (16%)
	Not on calcineurin inhibitors[B]	319 (26%)
Indication for biopsy	Primary nonfunction	10 (1%)
	Rapid deterioration of graft function	217 (18%)
	Slow deterioration of graft function	219 (18%)
	Stable impaired graft function	79 (7%)
	Investigate proteinuria/rejection/BK/creatinine	175 (14%)
	Delayed graft function	43 (4%)
	Others	443 (37%)
	Indication unknown	22 (2%)

[A]Local IRB numbers can be found in the Supplemental Acknowledgments. [B]Includes Everolimus (1), Sirolimus (25), Belatacept (11), Azathioprine (4), only corticosteroid (2), only corticosteroid/MMF (27), only MMF (5), no IS treatment (244).

some biopsies with indeterminate rejection status. The lesions are not highly disease specific and require arbitrary cutoffs. Furthermore, the Banff guidelines state that ABMR lesions cannot be used to diagnose ABMR without donor-specific antibodies (DSA) but fail to acknowledge that DSA can be missed for a variety of reasons (1, 4). The rules are empirically derived with the goal of achieving a reasonable trade-off between over- and underdiagnosis. Thus, in many biopsies, the rejection status is not known with a high degree of certainty, despite the commonly encountered phrase biopsy-proven rejection.

Replacing empirical histologic classifications with probabilistic data–driven molecular estimates of disease states offers many advantages (5–9). The most important of these is the ability to assign a level of confidence to each diagnosis, rather than requiring that all biopsies fulfilling a set of diagnostic criteria be grouped, with no provision for the strength of evidence. In addition, the molecular measurements are objective, highly reproducible, and require relatively small amounts of tissue. Molecular tests are usually derived by supervised analyses, in which conventional phenotypes (e.g., histologic diagnoses) are used to guide the development of molecular classifiers: machine learning–based algorithms that detect patterns of gene expression associated with the phenotypes. Our previous studies have been based on such classifiers comparing rejection vs. nonrejection phenotypes (10–14). However, we have subsequently found that methods that combine multiple classifiers provide better estimates than single classifiers (15), implying that a new method is needed to assemble the input from multiple classifiers into a single probabilistic assessment.

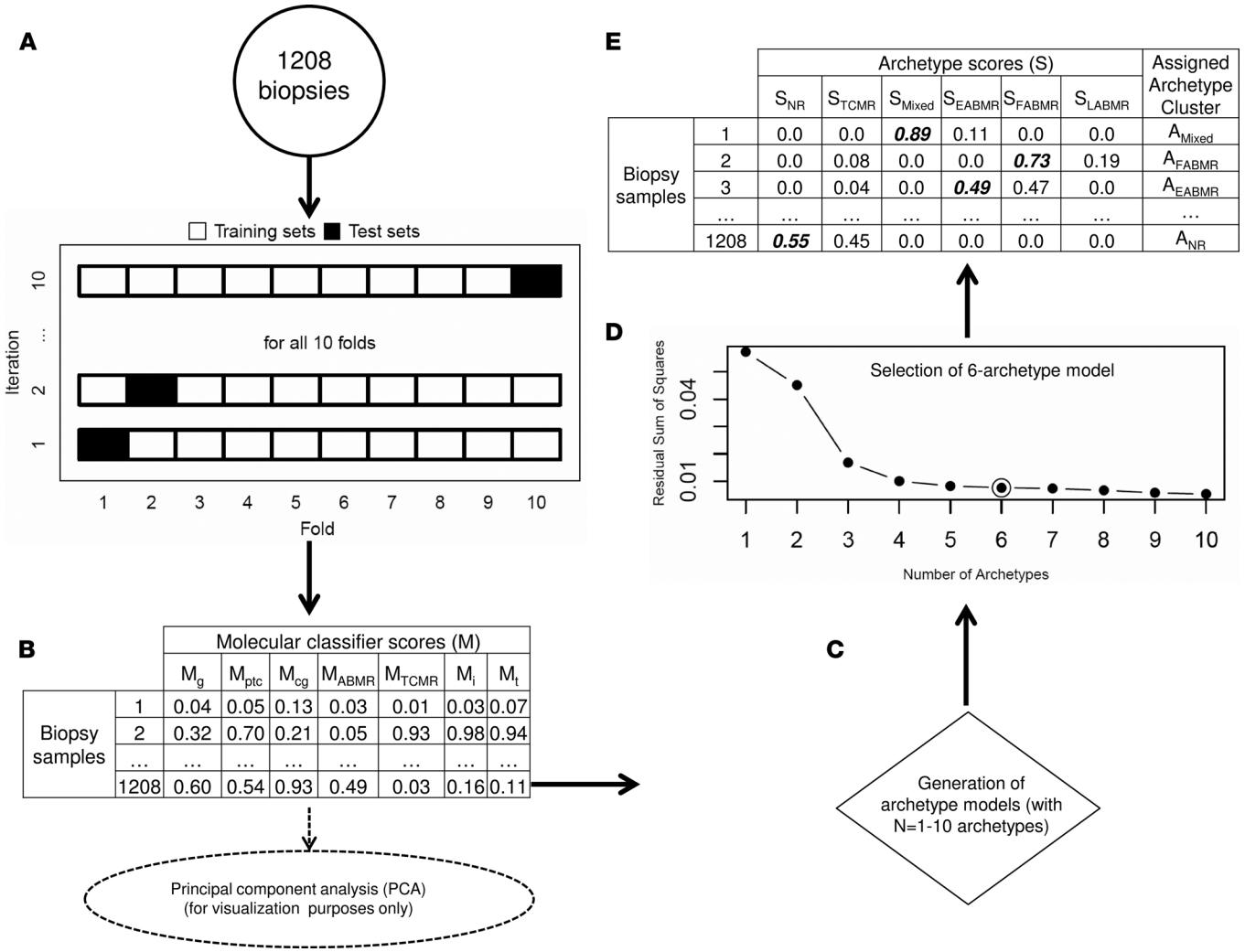

Figure 1. Classifier algorithm flowchart. A and **B** show how the base classifiers (TCMR, ABMR, i > 1, t > 1, g > 0, cg > 0, ptc > 0) were developed, while **C–E** show the archetypal analysis. For each of the 7 base classifiers: (**A**) 10-fold cross-validation is performed, randomly splitting the 1,208 biopsies into 10 folds of equal or near-equal size. For each of 10 iterations, 1 fold is left out as a test set (black box), and a classifier is developed using the remaining 9 folds (white boxes) as the training set. All aspects of classifier development, including probe set selection, are carried out from scratch within the training set samples at each iteration. The top 20 (by *P* value) differentially expressed probe sets comparing the binary phenotypes within the training set are selected as input features for the classifier. Twelve different classifier algorithms are developed in each training set, generating 12 scores for each test set sample (1 for each classifier algorithm). The median of these 12 is used as each test set sample's final score. This process is repeated over all 10 iterations, resulting in each biopsy being in a test set once and receiving a single value. This is repeated for each of the 7 base classifiers, resulting in a 1,208 × 7 matrix of classifier test set scores (**B**). This data is used as the input for both the principal component analysis (used for visualizing the multivariate distribution) and the archetypal analysis (**C–E**). We generated 10 archetype models (with n = 1–10 archetypes) (**C**). The residual sum of squares decreases with increasing numbers of archetypes (scree plot in **D**). We selected 6 archetypes (circled point in **D**) as the final archetypal model. (**E**) All biopsy samples are assigned a score for each of the 6 archetypes, and cluster assignments are made based on the highest score within that biopsy. The tables included show what typical data look like but do not represent actual results. S, archetype score; NR, no rejection; TCMR, T cell–mediated rejection; ABMR, antibody-mediated rejection; EABMR, early-stage ABMR; FABMR, fully developed ABMR; LABMR, late-stage ABMR; M, Molecular classifier scores; g, glomerulitis; cg, transplant glomerulopathy; ptc, peritubular capillaritis; i, interstitial inflammation; t, tubulitis

One such method is archetypal analysis (AA), which defines a limited set of extreme or pure phenotypes (i.e., archetypes) within a data set (16). It is unsupervised in that it chooses the archetypes based on patterns in the molecular data alone and does not use external phenotypic information (e.g., histologic diagnoses). An advantage of AA is that it describes each biopsy as a composite of the underlying archetypes, enabling precise, probabilistic assessments that retain the uniqueness of each sample, which is the ultimate goal of precision medicine. Furthermore, the use of AA in combination with dimensionality reduction methods such as principal component analysis (PCA) allows the characteristics of each sample to be visualized relative to all other samples in a reference set using 2- or 3-dimensional plots.

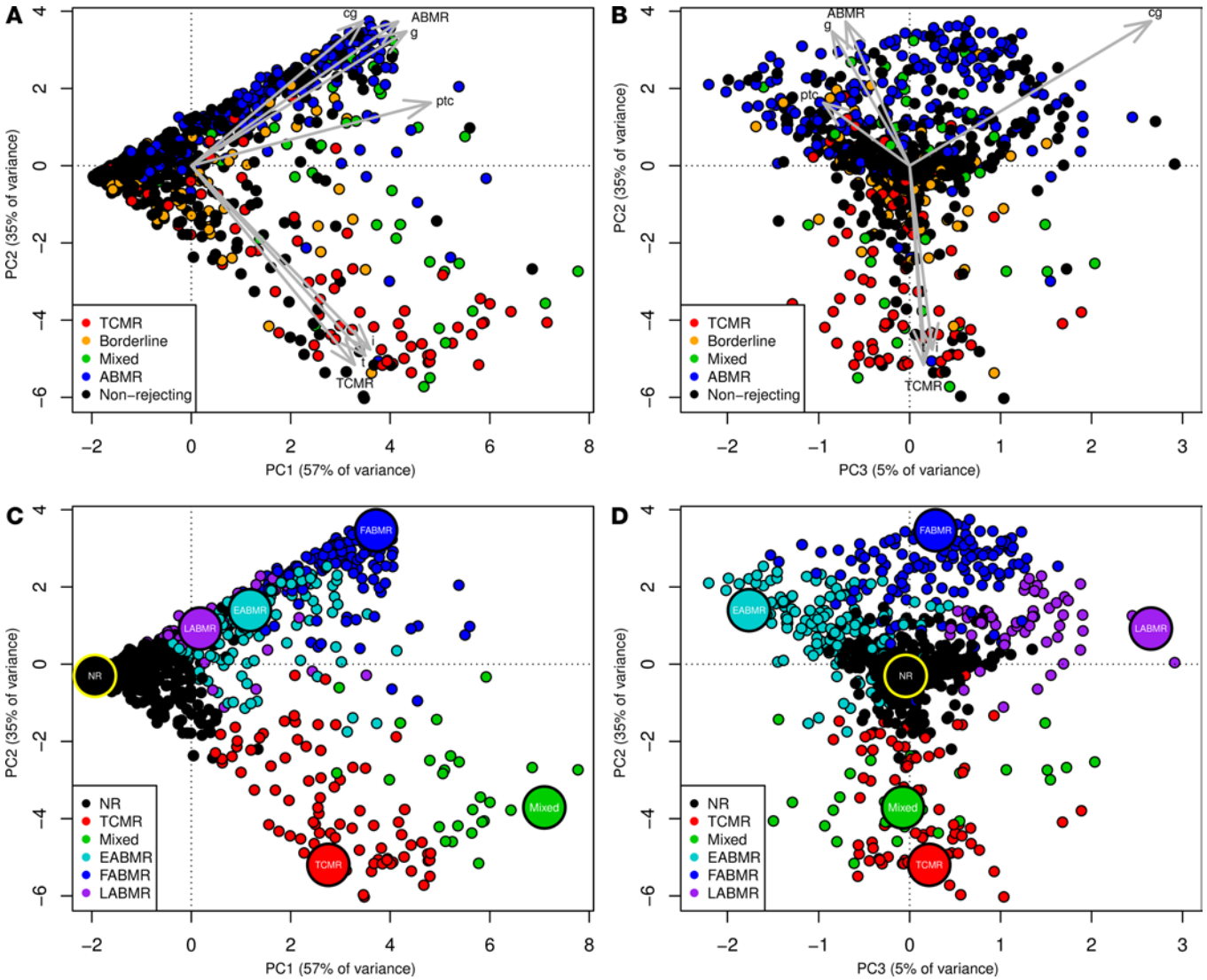

Figure 2. Principal component plots of the 1,208 reference set biopsies. The 1,208 × 7 matrix of base classifier scores (M_g, M_{ptc}, M_{cg}, M_i, M_t, M_{TCMR}, M_{ABMR}) was used to generate these PCA plots. (**A**) PC2 vs. PC1, with each biopsy (dot) colored by its histologic diagnosis. The superimposed arrows show the direction and relative magnitudes of the correlations between the 7 input variables (the molecular base classifiers) and the PC scores. (**B**) PC2 vs. PC3 from the same analysis. (**C** and **D**) The data points are identical to those in **A** and **B** but are now colored and labeled by archetype cluster. Locations of the 6 archetype centers are indicated by the large colored circles. NR, no rejection; TCMR, T cell–mediated rejection; ABMR, antibody-mediated rejection; EABMR, early-stage ABMR; FABMR, fully developed ABMR; LABMR, late-stage ABMR; g, glomerulitis; cg, transplant glomerulopathy; ptc, peritubular capillaritis; i, interstitial inflammation; t, tubulitis

In the present study, our goal was to develop a new system for assessing rejection-related disease states in kidney transplant biopsies based on molecular phenotypes. We sequentially applied supervised analysis to detect relevant molecular features and then unsupervised analyses to discover the disease classes. In order to sample the prevalent renal transplant population, we prospectively collected 1,208 unselected, kidney transplant biopsies from 13 international centers. Our goal was to develop a method for probabilistic assessment of rejection as a new biopsy diagnostic system and to compare this with the clinical, histological, and serological parameters.

Results

Population and biopsy characteristics

Population demographics (1,208 biopsies from 1,045 patients at 13 international centers; see Supplemental Acknowledgments) are shown in Table 1, and histologic diagnoses are shown in Table 2. ABMR ($C4d^+$ and $C4d^-$) was diagnosed in 215 biopsies, TCMR in 87, and mixed rejection in 41. Chronic

Table 2. Histologic diagnoses

ABMR-related (279)	ABMR	215 (18%)
	ABMR (suspected)	24 (2%)
	Transplant glomerulopathy (TG)	40 (3%)
Mixed (TCMR plus ABMR) (*n* = 41)		41 (3%)
TCMR-related (196)	TCMR	87 (7%)
	Borderline	109 (9%)
No rejection (692)	AKI	96 (8%)
	BK	37 (3%)
	Diabetic nephropathy	18 (1%)
	Glomerulonephritis	97 (8%)
	IFTA not otherwise specified	145 (12%)
	No major abnormalities (NOMOA)	274 (23%)
	Other[A]	25 (2%)

[A]Other include calcineurin inhibitor toxicity (CNIT), C4d deposition without morphologic evidence for active rejection, donor origin vascular disease, pyelonephritis, systemic infection/diarrhea, and bacterial infection.

active TCMR was never diagnosed. Some biopsies were assigned ABMR-related diagnoses: 24 ABMR suspected and 40 transplant glomerulopathy (TG). Borderline changes (canonical TCMR lesions below the threshold required for TCMR) were diagnosed in 109 biopsies. Overall, 516 had rejection or rejection-related diagnoses. The 692 nonrejecting biopsies should not be interpreted as normal, since this heterogeneous group includes many types of diseases and insults that occur in prevalent renal transplants, e.g., recurrent glomerulonephritis.

Supervised base classifiers

Figure 1 is a flowchart describing the analysis workflow, including generation of the base classifiers and their incorporation as input into the AA. We developed molecular classifiers (Figure 1) for estimating each of 7 rejection-related histologic phenotypes: diagnosis of TCMR, diagnosis of ABMR, and scores for the canonical lesions related to TCMR (i- and t-scores) and ABMR (ptc-, g-, and cg-scores). Details regarding the 7 base classifiers, each estimated using the median of scores from 12 machine learning algorithmic variants, as well as cut-offs and cross-validation methods are outlined in the Methods. Molecular base classifiers are designated by an M with a subscript indicating the feature they estimate, e.g., M_{ptc}.

The performance of the base classifiers in terms of predicting the 7 rejection-related phenotypes can be seen in Supplemental Figure 1; supplemental material available online with this article; https://doi.org/10.1172/jci.insight.94197DS1. Although there were overlaps in the molecular scores (generated using 10-fold cross-validation; Figure 1A) between histologic categories, the overall concordance between histology and classifiers was high, with AUCs ranging from 0.82–0.87. These molecular scores were used to populate the data matrix (1,208 samples × 7 classifier scores) depicted in Figure 1B.

Unsupervised AA

To derive a molecular classification of rejection that combines all base classifier results and minimizes reliance on the histologic classes, we set aside histologic diagnoses at this point and treated molecular rejection–related disease classification as an unsupervised clustering problem (Figure 1, C–E), using the data matrix from Figure 1B as the input.

The initial stage in AA involves developing models with different numbers of archetypes and choosing which to use as the final model. As with all clustering methods, the choice of cluster number is somewhat subjective. We tested models with 1–10 archetypes (Figure 1C) and chose 6 as the optimal number based on 3 separate criteria. The first was the commonly used "elbow" method (17), where the choice is based on the relative decrease in the residual sum of squares in a scree plot, as in Figure 1D. The second was by considering how the grouping of the biopsies changed when moving from 5 to 6 to 7 archetypes (the most plausible choices from the elbow method). Reducing the number to 5 mainly resulted in a merging of the molecular TCMR and mixed rejection groupings. Increasing to 7 resulted in a small new cluster similar to early-stage ABMR (EABMR) but with lower classifier scores for all ABMR-related phenotypes. The third was that 6 rejection-related groups were most congruous with existing beliefs regarding rejection-related states, which indicate 6 general classes: no rejection (NR), TCMR, EABMR, fully developed ABMR (FABMR), late-stage ABMR (LABMR); and mixed (18).

The final 6-archetype model assigns 6 scores to each biopsy, one for each archetype, with the scores summing to 1.0. We designate the 6 archetype scores by S with a subscript (e.g., S_{TCMR}, S_{Mixed}, etc.). Each biopsy was also assigned to a single archetype cluster based on its highest archetype score. For descriptive purposes, we will refer to the 6 archetype clusters by an "A" followed by a subscripted label reflecting the cluster's main histologic characteristics (e.g., the TCMR archetype cluster is A_{TCMR}).

Figure 1E illustrates archetype scores and cluster assignments for 4 of the 1,208 samples, which have maximum scores in mixed, FABMR, EABMR, and NR, respectively, and are accordingly assigned to those clusters. However, much of the strength of AA lies in its probabilistic assessment,

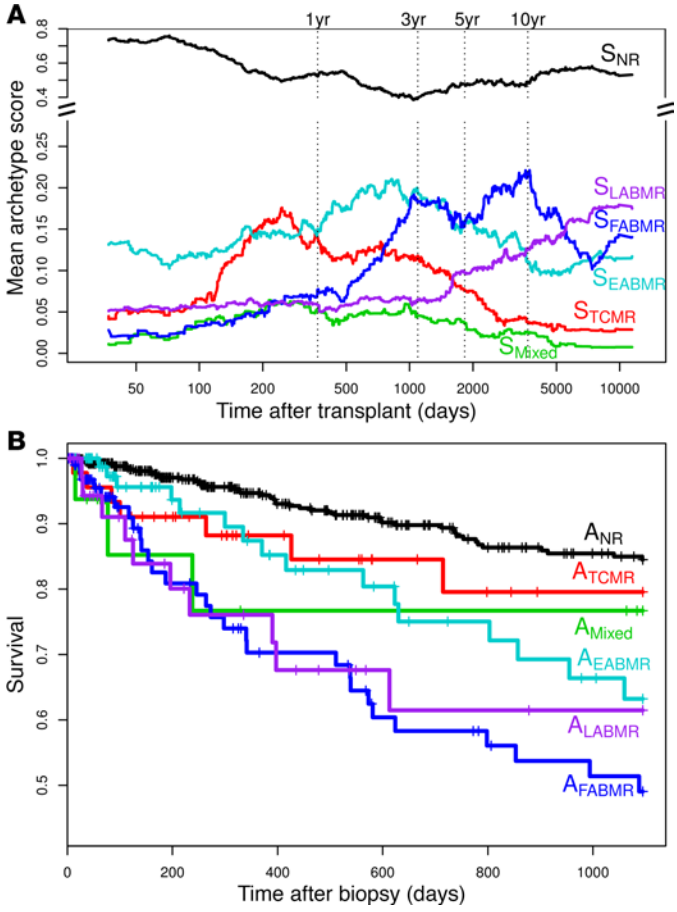

Figure 3. Archetype characteristics. (**A**) Trends in mean archetype scores (S) over time. The lines represent summarized aggregates of the mean archetype scores over (mostly) different patients' transplants rather than a true time course within individual patients. The plotted values are based on a sliding window of size n = 150 biopsies. Each line is the mean of that archetype's score in all 1,208 biopsies, not just the members of that archetype cluster. (**B**) Kaplan-Meier survival plot for the 6 archetype clusters. Vertical dash marks represent censored observations. A, archetype cluster; NR, no rejection; TCMR, T cell–mediated rejection; ABMR, antibody-mediated rejection; EABMR, early-stage ABMR; FABMR, fully developed ABMR; LABMR, late-stage ABMR.

since assigning single diagnostic categories to samples can be misleading. For example, sample 3 in Figure 1E is classified in the A_{EABMR} cluster but has nearly equal S_{EABMR} and S_{FABMR} scores. (The sum of the S_{EABMR}, S_{FABMR}, and S_{LABMR} scores can be considered as a total ABMR score.) The individual scores provide more detail than the cluster assignments regarding each biopsy. However, the cluster assignments are convenient for summarizing results and are therefore used in the presentation of some of our results.

Visualization using PCA

We used PCA to visualize the data matrix in Figure 1B that is used as the input for the AA. In the first 2 principal components (PCs), the biopsies distributed between 3 poles characterized by histologically defined nonrejection, TCMR, and ABMR (Figure 2A, which is colored by histologic diagnoses). The superimposed arrows indicate the direction and relative magnitude of the correlations between the 7 molecular input variables and the resulting PC scores, e.g., samples with high molecular TCMR and i- and t-classifier scores can be found at the bottom of the plot. Other combinations of molecular inputs and sample populations suggest that this is a typical and qualitatively robust pattern seen in kidney transplant biopsies (data not shown). The primary axis of variation (PC1) is always nonrejection to rejection (left to right in Figure 2A), while the secondary axis (PC2) is ABMR to TCMR (top to bottom in Figure 2A). Figure 2B represents the same PCA as Figure 2A but shows PC2 vs. PC3 (i.e., the view looking down the x axis of Figure 2A). It is clear that TCMR and ABMR represent continua with nonrejection and with each other, as opposed to discrete nonoverlapping entities.

Figure 2, C and D, show the same PCA results as Figure 2, A and B, but colored by the archetype cluster assignments (arrows omitted for clarity). Large circles indicate the locations of the 6 archetypes. PC3 now emerges as the continuum from early-stage (cg-poor) ABMR (A_{EABMR}) on the left through fully developed (A_{FABMR}) to late-stage (cg-dominant) ABMR (A_{LABMR}) on the right. PC3 comprised 5% of the variance in the data, compared with PC1 (57%) and PC2 (35%). Together, PC1, PC2, and PC3 account for 97% of the rejection-related variance in the 1,208 × 7 input matrix.

Summary of the main characteristics of each archetype cluster

The mean values for various clinical variables and histologic lesions are shown in Table 3, and the distributions of histologic diagnoses within archetype clusters are shown in Table 4.

A_{NR} (n = 774). Biopsies in A_{NR}, the largest cluster (median 395 days after transplant), showed the fewest rejection-related abnormalities and had relatively high mean GFR (46 cc/min). These represent the background level in nonrejecting indication biopsies in these centers. The proportion of patients positive for the clinical features was proteinuria, 0.54; DSA, 0.30; and panel reactive antibody (PRA), 0.55. Note the relatively high prevalence of DSA and PRA in this nonrejecting archetype. The proportion of biopsies with C4d staining was 0.10. The common histologic diagnoses in A_{NR} were no major abnormalities (NOMOA; 31%), atrophy-fibrosis (IFTA; 16%), and AKI (12%). The hyalinosis (ah-lesion) scores, which are due in part to calcineurin inhibitors, were higher in A_{NR} than in the A_{TCMR} and A_{Mixed} clusters, despite similar time after transplant (TxBx), compatible with some rejection being triggered by nonadherence (see below) (19).

Table 3. Values[A] of patient clinical variables and histologic lesion scores in the 6 archetype clusters in 1,208 biopsies

Variable	Archetype cluster					
	A_{NR} *n* = 774 (64%)	A_{TCMR} *n* = 81 (7%)	A_{Mixed} *n* = 27 (2%)	A_{EABMR} *n* = 139 (12%)	A_{FABMR} *n* = 136 (11%)	A_{LABMR} *n* = 51 (4%)
Median TxBx (days)	395	355	360	487	1,865	2,804
GFR (cc/min)	46	37	32	51	46	32
Proteinuria	0.54	0.48	0.74	0.52	0.79	0.86
DSA	0.30 (74)[B]	0.26 (7)	0.59 (5)	0.56 (7)	0.76 (11)	0.49 (6)
PRA	0.55	0.59	0.77	0.79	0.90	0.80
C4d	0.10 (190)	0.09 (14)	0.24 (10)	0.24 (28)	0.48 (33)	0.28 (8)
Donor age (years)	45	42	41	41	37	36
Recipient age (years)	52	46	42	52	45	49
g (glomerulitis)	0.16 (17)	0.16 (2)	0.62 (1)	0.95 (1)	1.32 (4)	0.98 (4)
ptc (capillaritis)	0.23 (22)	0.53 (6)	1.56 (0)	1.01 (2)	1.74 (3)	1.02 (5)
cg (double contours)	0.18 (17)	0.04 (3)	0.31 (1)	0.34 (3)	1.50 (6)	1.50 (5)
i (interstitial infiltrate)	0.28 (11)	1.85 (1)	2.31 (1)	0.62 (0)	0.50 (4)	0.69 (3)
t (tubulitis)	0.33 (14)	2.00 (1)	2.11 (0)	0.56 (0)	0.37 (3)	0.32 (4)
v (vasculitis)	0.02 (46)	0.27 (2)	0.46 (3)	0.10 (7)	0.16 (14)	0 (6)
ci (scarring)	1.11	1.55	1.04	1.23	1.64	1.85
ct (atrophy)	1.08	1.53	1.00	1.05	1.53	1.74
cv (intimal thickening)	1.02	0.90	0.92	0.83	1.24	1.14
ah (hyalinosis)	1.11	0.51	0.38	0.85	1.43	1.93

[A]Main table entries indicate means; time after transplant (TxBx) indicate medians. Proteinuria, DSA, PRA, and C4d are coded as positive/present = 1, negative/absent = 0. Therefore, the means for these variables indicate the proportion of biopsies that were positive/present. Missing values were excluded from the calculations. [B]Numbers in parentheses indicate the number of biopsies with missing values for the variables most often used in the diagnosis of rejection.

A_{TCMR} (n = 81). These kidneys, biopsied at median TxBx of 355 days, presented with a relatively low prevalence of proteinuria (0.48), DSA (0.26), PRA (0.59), and C4d staining (0.09) and a fairly low mean GFR (37 cc/min). The common histologic diagnoses in A_{TCMR} were TCMR (46%), BK virus (19%), and borderline rejection (9%). The histologic i-, t-, and v-lesion scores were high, although not as high as in A_{Mixed}, and the atrophy-scarring (ci and ct) scores were elevated.

A_{Mixed} (n = 27). Biopsied at a median TxBx of 360 days, these kidneys shared features of TCMR and EABMR, and they had the lowest mean GFR (32 cc/min). A_{Mixed} biopsies had both histologic TCMR (i-, t-, and v-) and histologic ABMR (g- and ptc-) lesions, and a higher prevalence of DSA (0.59) than A_{NR} or A_{TCMR}. A_{Mixed} had the lowest hyalinosis score of any cluster, suggesting that this phenotype is triggered by nonadherence (19, 20). The common histologic diagnoses were rejection related: 44% TCMR, 33% mixed, and 11% ABMR.

A_{EABMR} (EABMR, n = 139). With a median TxBx of 487 days, these kidneys had relatively high mean GFR (51 cc/min), and many patients had DSA (0.56) and/or PRA (0.79). They had high ptc- and g-scores but low cg-scores and relatively little atrophy scarring (ci-score 1.23 and ct-score 1.05). The common histologic diagnoses were ABMR 35%, NOMOA 14%, and IFTA 14%. Some A_{EABMR} biopsies were called borderline (12%) or TCMR (8%), suggesting that EABMR is sometimes misclassified by histology because it has TCMR-like i-, t-, and v-lesions (18). The time-dependent histologic cg-lesion scores were lower than in the A_{FABMR} and A_{LABMR} clusters.

A_{FABMR} (FABMR, N = 136). Biopsies with A_{FABMR} presented at a median of 1,865 days (~5 years) after transplant, with good GFR (46 cc/min). DSA and PRA positivity were common in these patients, but 24% were DSA-negative and 10% were PRA-negative, respectively, by local standard-of-care testing, similar to a previous analysis (4). Eleven A_{FABMR} cases did not have DSA data recorded. A_{FABMR} biopsies had a high prevalence of C4d positivity (48%) and high scores for the ABMR histologic cg-lesions (1.50), ptc-lesions (1.74), and g-lesions (1.32), but low TCMR lesion scores. The most common histologic diagnosis was ABMR (62%), but 17 (13%) were called mixed by histology, reflecting the above-mentioned ambiguities caused by TCMR-like histologic lesions in pure ABMR. Eleven of these 17 were called TCMR based only on v-lesions, which are unreliable for diagnosing TCMR (20).

Table 4. Number of biopsies with each histologic diagnosis in the 6 archetype clusters (% of columns)

			Archetype cluster					
Histologic diagnosis		***n* = 1208**	**A_{NR}** *n* = 774	**A_{TCMR}** *n* = 81	**A_{Mixed}** *n* = 27	**A_{EABMR}** *n* = 139	**A_{FABMR}** *n* = 136	**A_{LABMR}** *n* = 51
ABMR-related *n* = 279	**ABMR**	215 (18%)	54 (7%)	2 (2%)	3 (11%)	48 (35%)	84 (62%)	24 (47%)
	ABMR suspected	24 (2%)	11 (1%)	1 (1%)	1 (4%)	3 (2%)	5 (4%)	3 (6%)
	TG	40 (3%)	16 (2%)	1 (1%)	0 (0%)	5 (4%)	11 (8%)	7 (14%)
Mixed		41 (3%)	3 (0%)	6 (7%)	9 (33%)	5 (4%)	17 (13%)	1 (2%)
TCMR-related *n* = 196	**TCMR**	87 (7%)	27 (3%)	37 (46%)	12 (44%)	11 (8%)	0 (0%)	0 (0%)
	Borderline	109 (9%)	79 (10%)	7 (9%)	1 (4%)	17 (12%)	3 (2%)	2 (4%)
Not rejection-related *n* = 516	**No major abnormalities (NOMOA)**	274 (23%)	240 (31%)	3 (4%)	0 (0%)	20 (14%)	6 (4%)	5 (10%)
	AKI	96 (8%)	90 (12%)	0 (0%)	0 (0%)	6 (4%)	0 (0%)	0 (0%)
	IFTA	145 (12%)	123 (16%)	2 (2%)	1 (4%)	12 (9%)	3 (2%)	4 (8%)
	GN	97 (8%)	77 (10%)	5 (6%)	0 (0%)	8 (6%)	5 (4%)	2 (4%)
	Diabetic nephropathy	18 (1%)	15 (2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (6%)
	BK nephropathy	37 (3%)	20 (3%)	15 (19%)	0 (0%)	1 (1%)	1 (1%)	0 (0%)
	Other	25 (2%)	19 (2%)	2 (2%)	0 (0%)	3 (2%)	1 (1%)	0 (0%)

A_{LABMR} (Late ABMR, n = 51). A_{LABMR} biopsies presented at a median of 2,804 days (~8 years) after transplant, with low GFR (32 cc/min) and a high prevalence of proteinuria (0.86). Only 49% had DSA, the lowest of the 4 ABMR/mixed clusters, similar to our previous estimates of late cg-dominant ABMR (4, 18). The common histologic diagnoses were ABMR (47%) and TG (14%). The cg-score was high (1.5), similar to A_{FABMR}, but the ptc- and g-scores were lower, more like those in A_{EABMR}. There was extensive scarring with ci at 1.9 and ct at 1.7. The ah score was 1.9, the highest of any cluster (in keeping with the association of high ah scores with chronic glomerular diseases and time after transplant; ref. 19).

Summary of discrepancies between histologic diagnoses and archetype clusters

Table 5 shows, within each histology diagnostic category, the number and proportion of biopsies assigned to each archetype cluster. The main areas where there were discrepant findings between histology and archetypes were as follows: i) histologic ABMR was 25% A_{NR}; ii) histologic TG was 60% molecular rejection, usually ABMR; iii) histologic mixed rejection was only molecular mixed (A_{Mixed}) 22% of the time, usually being molecular ABMR; iv) histologic TCMR was 43% molecular TCMR but was often NR, mixed, or ABMR; v) histologic borderline rejection, which is defined by TCMR-related i- and t-lesions, was usually molecular NR (A_{NR} 72%) and was more likely to be molecular ABMR (20%) than A_{TCMR} (6%). Of all 516 assigned histologic diagnoses related to rejection, 280 of 516 (54%) were molecularly discrepant.

Of 692 biopsies with no histologic rejection, 108 (16%) had molecular rejection. Of interest is the apparent molecular rejection (ABMR or TCMR) in 20 biopsies with histologic GN, and molecular ABMR in 3 biopsies with diabetic nephropathy. These suggest that rejection diagnoses can be missed when another histologic disease is present, perhaps because canonical rejection histologic lesions such as cg can be obscured by other diseases. Histologic BK virus nephropathy biopsies are 41% A_{TCMR}. This is mainly due to a convention by which some pathologists will not diagnose TCMR, despite typical lesions, when BK nephropathy is diagnosed (10), due to uncertainty as to whether the lesions are caused by rejection or virus infection.

Relationship between archetype scores and time after transplant

Figure 3A shows the change in the means of the 6 archetype scores (S) over TxBx. Since the sum of the scores at any given TxBx is 1.0, the mean score can also be interpreted as prevalence. Molecular nonrejection (A_{NR}) represented ~75% of early biopsies but fell to ~50% by the first year. Molecular TCMR (A_{TCMR}) was most common around 200 days and was rare after 7 years. Molecular mixed rejection (A_{Mixed}) was

Table 5. Discrepancies between histologic diagnoses and the archetype cluster assignments of nonrejection, TCMR, mixed, and all ABMR combined

Histologic diagnosis			Archetype cluster (% of row)				Discrepancies between molecular diagnoses (clusters) in each histologic diagnostic category[A] (=% of row)
		***n* = 1208**	**A_{NR}** *n* = 774	**A_{TCMR}** *n* = 81	**A_{Mixed}** *n* = 27	**All ABMR** *n* = 326	
ABMR-related *n* = 279	**ABMR**	215	54 (25%)	2 (1%)	3 (1%)	**156 (72%)[B]**	59/215 (27%)
	ABMR suspected	24	11 (46%)	1 (4%)	1 (4%)	**11 (46%)**	13/24 (54%)
	TG	40	**16 (40%)**	1 (2%)	0 (0%)	23 (58%)	24/40 (60%)
Mixed *n* = 41	**Mixed**	41	3 (7%)	6 (15%)	**9 (22%)**	23 (56%)	32/41 (78%)
TCMR-related *n* = 196	**TCMR**	87	27 (31%)	**37 (43%)**	12 (14%)	11 (13%)	50/87 (57%)
	Borderline	109	79 (72%)	**7 (6%)**	1 (1%)	22 (20%)	102/109 (94%)
All histologic rejection related		*516*	*190 (37%)*	*54 (10%)*	*26 (5%)*	*246 (48%)*	*280/516 (54%)*
Not rejection-related *n* = 692	**No major abnormalities (NOMOA)**	274	**240 (88%)**	3 (1%)	0 (0%)	31 (11%)	34/274 (12%)
	AKI	96	**90 (94%)**	0 (0%)	0 (0%)	6 (6%)	6/96 (6%)
	IFTA	145	**123 (85%)**	2 (1%)	1 (1%)	19 (13%)	22/145 (15%)
	GN	97	**77 (79%)**	5 (5%)	0 (0%)	15 (15%)	20/97 (21%)
	Diabetic nephropathy	18	**15 (83%)**	0 (0%)	0 (0%)	3 (17%)	3/18 (17%)
	BK nephropathy	37	**20 (54%)**	15 (41%)	0 (0%)	2 (5%)	17/37 (46%)
	Other	25	**19 (76%)**	2 (8%)	0 (0%)	4 (16%)	6/25 (24%)
All histologic nonrejection related		*692*	*584 (84%)*	*27 (4%)*	*1 (0.1%)*	*80 (12%)*	*108/692 (16%)*
						Total discrepancies:	**387/1,208 (32%)**

[A]Considering the histologic diagnosis of borderline as "suspicious for TCMR" (1), and therefore discrepant when it is A_{NR}, A_{Mixed}, or all ABMR; ABMR suspected as discrepant when it is A_{NR}, A_{TCMR}, or A_{Mixed}. [B]The bolded numbers indicate the agreement between the histologic assessment and the molecular assessment. All others are discrepancies. =% of, percent of total number analyzed. TG, transplant glomerulopathy; TCMR, T cell–mediated rejection; NOMOA, no major abnormalities; AKI, acute kidney injury; IFTA, atrophy-fibrosis with no explanatory disease state; GN, glomerulonephritis; MMDx, Molecular Microscope Diagnostic System.

never common (*n* = 27) and had a long peak extending from ~200–1,000 days. Some molecular EABMR (A_{EABMR}) was detected in the first months (usually in patients transplanted with positive DSA; ref. 21), but the main peak was around 900 days before becoming uncommon in late biopsies. Molecular FABMR (A_{FABMR}) started to become common after 2 years, peaked at ~10 years, and decreased in prevalence thereafter. Molecular LABMR (A_{LABMR}) was rare before 4 years but became progressively more common from that point onward, particularly after 10 years.

Graft failures after biopsy

Figure 3B shows Kaplan-Meier estimates of death-censored graft survival after biopsy in the 879 patients with follow-up data available (1 random biopsy per patient). Failures were least common in molecular NR (A_{NR}), followed by molecular TCMR (A_{TCMR}). Early failures within 1 year after biopsy were common in 3 of the 4 molecular ABMR clusters (A_{Mixed}, A_{FABMR}, and A_{LABMR}) but uncommon for A_{EABMR}. We calculated bootstrap corrected C-statistics (analogous to AUCs) for predicting 3-year survival using the archetype clusters, archetype scores, or Banff rejection–related diagnostic categories (each biopsy being defined as one of ABMR, TCMR, mixed, borderline, and NR). The C-statistics were: histologic diagnoses (0.60), archetype cluster assignments (0.65), and archetype scores (0.73). The *P* values for comparing these estimates were: archetype scores vs. either histologic diagnoses or clusters, 3×10^{-6}; diagnoses vs. clusters, 0.06.

Discussion

We analyzed 1,208 indication biopsies from 13 centers with the goal of developing a new molecular classification for rejection in kidney transplant biopsies. Seven molecular rejection–related scores were generated using supervised analysis based on histologic diagnoses and lesion scores. To avoid excessive reliance on any one classifier method, each of the 7 scores was defined as the median of the output from 12 different classifier algorithms. These molecular scores, assigned as test set results via cross-validation, were then used

as inputs for an unsupervised AA that produced 6 clusters corresponding to NR, TCMR, mixed rejection, and early-stage, fully developed, and LABMR. Molecular TCMR was predominant early, becoming rare by 10 years. Molecular ABMR occupied a continuum of molecular space and was composed of 3 partially differentiated subtypes that probably represent phases in the natural history of ABMR. These peak at median times of approximately 500; 1,900; and 2,800 days after transplant, respectively. In addition, a subtype of relatively early biopsies (A_{Mixed}, median ~1 year) had both TCMR and ABMR characteristics, as well as histologic evidence of nonadherence. Progression to graft failure after biopsy was common in ABMR and was better predicted by the archetype scores than by histologic diagnoses or archetype clusters, most likely due to the higher information content of continuous scores. The 32% discrepancy rate with histology was comparable with previous analyses of the uncertainty in histology based on the limited reproducibility of lesions and diagnoses. The results provide a critical assessment of the rejection states in the prevalent renal transplant population and highlight potential opportunities for diagnostic improvement.

The strength of the present analysis is that it combines the base classifiers in an unsupervised analysis without being constrained by existing histologic classes, letting the molecular data itself show the patterns of variation. Once these patterns/clusters were established, their characteristics in terms of histologic and clinical phenotypes were summarized. There is no single unsupervised clustering method that is "best" for all data sets, and the choice should be based on experience with the clinical phenotypes (domain-specific knowledge) and the particular objectives of the study. After evaluating several alternatives, we selected AA as our method of choice. K-means clustering, while popular, generates only discrete clusters, not probabilities. In addition, the clustering of k-means and related methods tends to place too much emphasis on PC1 and PC2 at the expense of lower PCs (based both on previous reports and observations from our own data; ref. 22). Note the importance of PC3 in this analysis in revealing the continuum in ABMR. An additional advantage of AA is that the archetype locations are less affected by the density of samples, in particular areas of multivariate space, making it more resistant to variation in the specific case mix of the population being analyzed (16).

The previously developed and independently validated molecular classifiers for ABMR (12, 13) and TCMR (10, 11), plus the lesion classifiers described in this paper, can now be merged into the AA diagnostic system to assess rejection-related states in new biopsies. When new samples become available, their gene expression data can be fed directly into the fixed classifiers from this study to generate AA-based molecular assessments. There are several reasons why this system will improve estimates of disease states compared with our earlier classifiers. First, the sample size (1,208 biopsies) is larger than in the earlier papers (403 for both TCMR and ABMR). Second, in looking at large numbers of individual classifier scores, there are occasionally samples that seem anomalous, e.g., high ABMR with low rejection scores. Combining different perspectives by using classifiers for several different aspects of rejection minimizes the influence of outlier results. Likewise, for samples that are difficult to call because of indeterminate molecular phenotypes, the combination of more "molecular opinions" helps to clarify matters. Third, even within a single base classifier, we found that using the median of 12 diverse classifier methods generated results that were more consistent than those based on a single method, as we did in the earlier papers.

An unexpected benefit of the unsupervised combined molecular phenotype approach used here is that it finds phenotypes that are not directly accessible through the use of simple ABMR/TCMR/rejection base classifiers. An ABMR classifier built from a comparison of ABMR vs. everything else in this data set correlates strongly with what turns out to be the FABMR (A_{FABMR}) archetype but is unable to find many of the LABMR (A_{LABMR}) group because they, on average, have fairly weak pure ABMR signals. Detecting this phenotype requires an approach more subtle than averaging, or even weighting a set of ABMR genes. A_{Mixed} provides another such example. This represents a fairly uncommon phenotype that we could conceivably have named the nonadherent cluster. These examples illustrate how AA can discover emergent properties of the raw molecular data that could not be found using simpler univariable approaches.

While several factors probably contribute to the discrepancies between molecular and histologic diagnoses, the poor agreement between pathologists in assigning diagnoses in abnormal biopsies (2, 3, 10) suggests that errors in conventional phenotyping play a large role. Some errors are due to questionable guidelines: the pathologist seeing tubulitis and interstitial infiltrate in ABMR will diagnose "mixed", despite the fact that these are nonspecific lesions in inflammatory renal diseases (23, 24). Endothelial arteritis (presence of v-lesions) is also ambiguous, occurring in TCMR, ABMR, and renal injury (20), yet it can be used to diagnose either TCMR or ABMR. Much weight is placed on DSA, although this occurs in many

indication biopsies with no molecular rejection (25). Other discrepancies are the result of inconsistencies associated with assigning DSA status, which is complicated by considerations of specificity, IgG subclass, and complement binding. Some discrepancies may reflect sample heterogeneity since the biopsy core that is read for histology is different from the core assessed by microarray, although our analysis of molecular scores in samples divided in half suggests that core-to-core variation in molecular scores is low (26). Some discrepancies will reflect variability in the molecular scores, although we have minimized this by using the median of 12 classifier methods for each of the 7 base classifiers (see Methods). Finally, dividing a continuum of phenotypes into distinct categories inevitably leads to apparent discrepancies based solely on where samples fall in relation to the thresholds imposed by the decision rules: it is better to retain the actual, probabilistic scores for each biopsy.

The distribution of the archetype scores over time offers insights into the natural history of rejection disease states in organ transplants, confirming the general features of the 6 histology-based classes and consistent with the observations in other cohorts (27–29). Biopsies in the early period predominantly show nonrejection, except for a very early peak in A_{EABMR} that probably reflects ABMR caused by preexisting DSA. TCMR dominates later in the first year but gradually becomes rare, possibly reflecting adaptive changes in donor-specific T cell responsiveness related to immune checkpoints (30). Recent nonadherence often presents as late TCMR or mixed TCMR plus early ABMR around 12–24 months (19, 31). A period of nonadherence may also trigger DSA and subclinical ABMR, which can present years later in apparently adherent patients (27). The striking interaction between TxBx and the diagnosis of early-stage, fully developed, and late-stage molecular ABMR has implications for understanding the natural history of ABMR. The peak times at which patients with early-stage, fully developed, and LABMR present with biopsy indications (500; 1,900; and 2,800 days) suggest stages of disease progression, but also imply that ABMR is often relatively silent. Thus, some EABMR is probably missed until presenting as FABMR, and some FABMR must be silent until presenting as LABMR. Moreover, it is striking that EABMR rarely presents after 10 years. These findings raise fundamental questions about the DSA/ABMR problem: what is the natural history of DSA? Can DSA and ABMR spontaneously disappear? What determines the pathogenicity of DSA? How can silent ABMR phenotypes be detected in the clinic, and if they can, how should they be managed? The use of the archetype method to assess each biopsy's molecular rejection phenotype will be useful in addressing such unknowns.

The present approach is of general interest for improving existing diagnostic systems and disease classifications via precision assessments. Histologic and molecular features are never truly specific (14) because disease processes share mechanisms with the nonspecific response to wounding, including innate immunity and microcirculation remodeling. Since many cases are diagnostically difficult, as acknowledged in the histology guidelines by specifying "borderline" and "ABMR suspected" categories, diseases (here specified as rejection states) should be assessed probabilistically. These probabilities are best evaluated by combining the output from multiple machine learning–based algorithms. This will provide the most precise evaluation of each biopsy, and as such should be used to inform clinical decisions.

Methods

Population and biopsy processing. The 1,208 biopsy samples were collected from 1,045 patients (Table 1). Biopsy samples were run on Affymetrix hgu219 PrimeView microarray chips. Of these samples, 529 have been published previously using the older Affymetrix hgu133plus2 chip (http://www.ncbi.nlm.nih.gov/geo/; accession numbers GSE36059 and GSE48581). One hundred and seventy-four of the earlier biopsies no longer had RNA available and were therefore not included in the present study. The details of microarray expression data are posted on the Gene Expression Omnibus (GEO) website (http://www.ncbi.nlm.nih.gov/geo/; accession number GSE98320).

The 7 base classifiers. Interquartile range (IQR) filtering was first performed to reduce the total number of probe sets analyzed from the original 49,495 to roughly half: 24,693. This is a nonspecific filtering method for eliminating low variance probe sets that are likely, a priori, to be noninformative. It does not make use of any information regarding the samples' phenotypes and is, therefore, unbiased in that regard.

AA used 7 molecular scores as inputs. These scores were based on classifiers designed to estimate the probability of each of the following phenotypes: TCMR (samples diagnosed with histologic TCMR vs. all other biopsies); ABMR (histologic ABMR [$C4d^-$ and $C4d^+$] vs. all other biopsies; note that for both these classifiers, "all other biopsies" included mixed rejection; high i-lesions (i-lesions > 1 vs. those ≤ 1); high

t-lesions (t-lesions > 1 vs. those ≤ 1); high g-lesions (g-lesions > 0 vs. those = 0); high cg-lesions (cg-lesions > 0 vs. those = 0); and high ptc-lesions (ptc-lesions > 0 vs. those = 0).

The cutoffs for the lesion classifiers were chosen to reflect those used by the Banff system for making diagnoses, although there is ongoing debate as to how the g- and ptc-scores should be combined.

Each of the 7 classifiers was built as follows. Biopsies were divided randomly into 10 groups or "folds" of approximately equal size (Figure 1A). Each fold was left out in turn and used as the test set for the classifiers trained in the remaining 9 of 10 folds (the training set), resulting in a 10-fold cross-validation (CV) of the data. In each training set, 12 different classifier algorithms were used to generate a probability between 0.0 and 1.0 in each test set sample. The median of the 12 classifier scores was used as the final output for each test set sample. We reasoned that this ensemble approach provided a more stable score than using any single method. The 12 classifiers (and associated functions in the "caret" package of "R" were: linear discriminant analysis (lda), regularized discriminant analysis (rda), mixture discriminant analysis (mda), flexible discriminant analysis (fda), gradient boosting machine (gbm), radial support vector machine (SVMR), linear support vector machine (SVML), random forest (rf), C5.0, neural networks (nnet), Bayes glm (bayesglm), and generalized linear model elastic-net (glmnet). The TCMR and ABMR classifiers are similar to those used in our earlier publications (10, 12), which were based on a smaller data set and used only 1 classifier method (lda), rather than the median of 12. It is important to note that, in all cases, no information of any sort leaked from a sample into data that was used to generate its scores. For example, probe set selection was done from scratch within each training set fold in the cross-validations.

After completing the 10-fold assessment of the 1,208 biopsies, an additional set of 12 classifiers for each of the 7 phenotypes was built using all 1,208 biopsies. These were the final models — the ones to be used on future data to generate input for the AA.

AA. In choosing a method for the unsupervised component of this study, we had 4 main requirements: i) both probabilistic and discrete estimates of cluster membership had to be generated; ii) the clusters should conform at least approximately with the currently accepted Banff diagnostic rejection categories; iii) the probabilities should align reasonably with the distributions in the PCA plots, e.g., a sample in Figure 2A at ($x = 3$, $y = -5$) should have a higher TCMR probability than one at ($x = 1$, $y = -2$) because it is farther away from the nonrejecting biopsies at the far left of the plot; and iv) the same method that was used to assign cluster scores/clusters to the 1,208 reference set samples could be used for assigning scores/clusters in future, unknown samples.

After testing several alternative unsupervised methods, we selected AA (22, 32) rather than clustering methods such as k-means and partitioning around medoids because AA allows each sample to be assigned a probability of membership in each archetype/cluster, in addition to a class assignment based on the cluster with the highest score. Although there are versions of k-means that allow for probabilistic class membership (fuzzy k-means clustering), they share with k-means the characteristic that distance (and therefore probability) is measured relative to the center of each cluster rather than based on the extremal points as in AA — a characteristic we think is more in keeping with an intuitive interpretation of the distributions as visualized in the PCA plots of Figure 2. Related to clustering, AA is a method that describes data in terms of combinations of a fixed number of archetypal samples. These lie toward the edges of the multivariate data distribution and describe a boundary encapsulating a large proportion of the data set. The archetypes are hypothetical rather than real samples, being caricatures of idealized extreme phenotypes. Each sample in the population gets a score between 0.0 and 1.0 for each archetype, under the constraint that the scores across all archetypes within each individual sum to 1.0. It is an unsupervised method, since the algorithm is given only the raw data, not any descriptive or outcome-related information (e.g., diagnoses). It has been used in a wide variety of fields, including sports analytics (33), astrophysics (34), marketing (35), and bioinformatics/medicine (16, 36, 37). Although it is common to show archetype results on a PCA plot as we do here, AA is conceptually very different from PCA. For instance, it employs no dimensionality reduction and gives equal weight to each of the original (scaled) input variables. PCA is used in connection with AA for visualization purposes only and should be considered only a rough approximation of the actual data (7-dimensional in our case) on which the AA is based.

Archetypes were assigned using the "archetypes" (22) package in "R". This archetype model is then used for generating both the archetype clusters and scores in the 1,208 set and clusters/scores in future biopsy samples (once they have had their base molecular scores assigned by the 7 fixed base classifiers).

Statistics. Selection of the top 20 probe sets used in each fold of cross-validation for the base classifiers was based on a Bayesian *t* test implemented in the "topTable" function of the "R" limma package. Cox

regressions and associated bootstrap-corrected C-statistic estimates were implemented using the "cph" and "validate" functions, respectively, from the "R" rms package.

Study approval. All biopsies were collected in protocols approved by the IRB in each of the 13 centers. The study was registered at clinicaltrials.gov (NTC1299168).

Author Contributions

JR contributed data analysis and manuscript writing\reviewing. GAB contributed biopsies, acquired data, and assisted with manuscript development and editing. FE contributed biopsies, acquired data, and assisted with manuscript development and editing. GE contributed biopsies, acquired data, and assisted with manuscript development and editing. CL contributed biopsies, acquired data, and assisted with manuscript review. AL contributed biopsies, acquired data, and assisted with manuscript review. PFH was the principal investigator and assisted with manuscript writing\reviewing. MMDx-Kidney study group contributed biopsies, acquired data, and assisted with manuscript editing.

Acknowledgments

See Supplemental Acknowledgments for consortium details.

We gratefully acknowledge the support of the Industrial Research Assistance Program. This research has been supported by funding and/or resources from University Hospital Foundation at the University of Alberta, Genome Canada, Canada Foundation for Innovation and Roche Organ Transplant Research Foundation. PFH held a Canada Research Chair in Transplant Immunology until 2008 and currently holds the Muttart Chair in Clinical Immunology. Funding for this research supported by Genome Canada, Canada Foundation for Innovation, and Roche Organ Transplant Research Foundation.

Address correspondence to: Philip F. Halloran, Alberta Transplant Applied Genomics Centre, #250 Heritage Medical Research Centre, University of Alberta, Edmonton, AB T6G 2S2, Canada. Phone: 780.492.6160; Email: phallora@ualberta.ca.

1. Loupy A, et al. The Banff 2015 Kidney Meeting Report: Current Challenges in Rejection Classification and Prospects for Adopting Molecular Pathology. *Am J Transplant*. 2017;17(1):28–41.
2. Furness PN, et al. International variation in histologic grading is large, and persistent feedback does not improve reproducibility. *Am J Surg Pathol*. 2003;27(6):805–810.
3. Furness PN, Taub N, Convergence of European Renal Transplant Pathology Assessment Procedures (CERTPAP) Project. International variation in the interpretation of renal transplant biopsies: report of the CERTPAP Project. *Kidney Int*. 2001;60(5):1998–2012.
4. Halloran PF, Famulski KS, Chang J. A Probabilistic Approach to Histologic Diagnosis of Antibody-Mediated Rejection in Kidney Transplant Biopsies. *Am J Transplant*. 2017;17(1):129–139.
5. Szolovits P, Pauker SG. Categorical and Probabilistic Reasoning in Medical Diagnosis. *Artif Intell*. 1978;11(3):115–144.
6. Doust J. Diagnosis in General Practice. Using probabilistic reasoning. *BMJ*. 2009;339:b3823.
7. Kim KI, Simon R. Probabilistic classifiers with high-dimensional data. *Biostatistics*. 2011;12(3):399–412.
8. Widiger TA, Samuel DB. Diagnostic categories or dimensions? A question for the Diagnostic And Statistical Manual Of Mental Disorders--fifth edition. *J Abnorm Psychol*. 2005;114(4):494–504.
9. Woodbury MA, Clive JM. Continuous and discrete global models of disease. *Mathematical Modelling*. 1986;7(5–8):1137–1154.
10. Reeve J, et al. Molecular diagnosis of T cell-mediated rejection in human kidney transplant biopsies. *Am J Transplant*. 2013;13(3):645–655.
11. Halloran PF, et al. Potential impact of microarray diagnosis of T cell-mediated rejection in kidney transplants: The INTERCOM study. *Am J Transplant*. 2013;13(9):2352–2363.
12. Sellarés J, et al. Molecular diagnosis of antibody-mediated rejection in human kidney transplants. *Am J Transplant*. 2013;13(4):971–983.
13. Halloran PF, et al. Microarray diagnosis of antibody-mediated rejection in kidney transplant biopsies: an international prospective study (INTERCOM). *Am J Transplant*. 2013;13(11):2865–2874.
14. Halloran PF, Venner JM, Famulski KS. Comprehensive analysis of transcript changes associated with allograft rejection: Combining universal and selective features [published online ahead of print January 19, 2017]. *Am J Transplant.* https://doi.org/10.1111/ajt.14200.
15. Halloran PF, Famulski KS, Reeve J. Molecular assessment of disease states in kidney transplant biopsy samples. *Nat Rev Nephrol*. 2016;12(9):534–548.
16. Hart Y, et al. Inferring biological tasks using Pareto analysis of high-dimensional data. *Nat Methods*. 2015;12(3):233.
17. Thorndike RL. Who Belongs in the Family? *Psychometrika*. 1953;18:267–276.

18. Halloran PF, Merino Lopez M, Barreto Pereira A. Identifying Subphenotypes of Antibody-Mediated Rejection in Kidney Transplants. *Am J Transplant*. 2016;16(3):908–920.
19. Einecke G, Reeve J, Halloran PF. Hyalinosis Lesions in Renal Transplant Biopsies: Time-Dependent Complexity of Interpretation. *Am J Transplant*. 2017;17(5):1346–1357.
20. Salazar ID, Merino López M, Chang J, Halloran PF. Reassessing the Significance of Intimal Arteritis in Kidney Transplant Biopsy Specimens. *J Am Soc Nephrol*. 2015;26(12):3190–3198.
21. Aubert O, et al. Antibody-mediated rejection due to pre-existing vs. de novo DSA in kidney allograft recipients [published online ahead of print March 2, 2017]. *J Am Soc Nephrol*. https://doi.org/ 10.1681/ASN.2016070797.
22. Eugster MJA, Leish F. From Spider-Man to Hero - Archetypal Analysis in R. *J Stat Softw.* 2009;30(8):1–23.
23. Iványi B, Marcussen N, Olsen S. Tubulitis in primary vascular and glomerular renal disease. *Pathol Res Pract*. 1995;191(12):1245–1257.
24. Berden AE, et al. Tubular lesions predict renal outcome in antineutrophil cytoplasmic antibody-associated glomerulonephritis after rituximab therapy. *J Am Soc Nephrol*. 2012;23(2):313–321.
25. Eskandary F, et al. Diagnostic Contribution of Donor-Specific Antibody Characteristics to Uncover Late Silent Antibody-Mediated Rejection-Results of a Cross-Sectional Screening Study. *Transplantation*. 2017;101(3):631–641.
26. Madill-Thomsen KS, Wiggins RC, Eskandary F, Bohmig GA, Halloran PF. The effect of cortex/medulla proportions on molecular diagnoses in kidney transplant biopsies: rejection and injury can be assessed in medulla [published online ahead of print February 22, 2017]. *Am J Transplant*. https://doi.org/10.1111/ajt.14233.
27. Wiebe C, et al. Rates and determinants of progression to graft failure in kidney allograft recipients with de novo donor-specific antibody. *Am J Transplant*. 2015;15(11):2921–2930.
28. Naesens M, et al. The histology of kidney transplant failure: a long-term follow-up study. *Transplantation*. 2014;98(4):427–435.
29. Hricik DE, et al. Multicenter validation of urinary CXCL9 as a risk-stratifying biomarker for kidney transplant injury. *Am J Transplant*. 2013;13(10):2634–2644.
30. Venner JM, Famulski KS, Badr D, Hidalgo LG, Chang J, Halloran PF. Molecular landscape of T cell-mediated rejection in human kidney transplants: prominence of CTLA4 and PD ligands. *Am J Transplant*. 2014;14(11):2565–2576.
31. Sellarés J, et al. Understanding the causes of kidney transplant failure: the dominant role of antibody-mediated rejection and nonadherence. *Am J Transplant*. 2012;12(2):388–399.
32. Cutler A, Breiman L. Archetypal Analysis. *Technometrics*. 1994;36:338–347.
33. Eugster MJA. Performance profiles based on archetypal athletes. *Int J Perform Anal Sport*. 2012;12(1):166–187.
34. Chan BHP, Mitchell DA, Cram LE. Archetypal analysis of galaxy spectra. *Mon No R Astron Soc*. 2003;338:790–795.
35. Li S, Louviere J, Carson R, Wang P. Archetypal analysis: A new way to segment markets based on extreme individuals. In: *A Celebration of Ehrenberg and Bass: Marketing Knowledge, Discoveries and Contribution ANZMAC 2003 Conference Proceedings*. Adelaide; 2003:1674–1679.
36. Thøgersen JC, Mørup M, Damkiær S, Molin S, Jelsbak L. Archetypal analysis of diverse Pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. *BMC Bioinformatics*. 2013;14:279.
37. Korem Y, et al. Geometry of the Gene Expression Space of Individual Cells. *PLoS Comput Biol*. 2015;11(7):e1004224.