Molecular diagnosis of kidney allograft rejection based on the Banff Human Organ Transplant gene panel: a multicenter international study

Dina Zielinski, Valentin Goutaudier, Marta Sablik, Gillian Divard, Olivier Aubert, Alexis Piedrafita, Fariza Mezine, Jessy Dagobert, Anais Certain, Blaise Robin, et al.

Molecular diagnosis of kidney allograft rejection based on the Banff Human Organ Transplant (B-HOT) gene panel: a multicenter international study

Dina Zielinski[1], Valentin Goutaudier[*1], Marta Sablik[*1], Gillian Divard[1,2], Olivier Aubert[1], Alexis Piedrafita[1,3] , Fariza Mezine[1], Jessy Dagobert[1], Anais Certain[1], Blaise Robin[1], Juliette Gueguen[4], Marion Rabant[1,5], Jean-Paul Duong van Huyen[5], Aurélie Sannier[1,6], Christine Randoux-Lebrun[7], Mehdi Maanaoui[8,9], Arnaud Lionet[8], Jean-Baptiste Gibier[10], Viviane Gnemmi[10], Moglie Le Quintrec[11], Bertrand Chauveau[12], Agathe Vermorel[13], Lionel Couzi[13], Oriol Bestard[14], Michelle Elias[2], Kevin Louis[2], Ivy A. Rosales[15,16], R. Neal Smith[15,16], Vanderlene L. Kung[17], Dany Anglicheau[18], Christophe Legendre[18], Arnaud Del Bello[3] , Edmund Huang[17], Benjamin Adam[19], Nassim Kamar[3], Robert B. Colvin[15,16], Michael Mengel[19], Carmen Lefaucheur[1,2], Alexandre Loupy[1,18#]

Corresponding author: Alexandre Loupy, Paris Institute for Transplantation and Organ Regeneration, 56 Rue Leblanc, Paris 75015. Email: alexandre.loupy@inserm.fr

* Valentin Goutaudier and Marta Sablik contributed equally as second authors.

1. Université Paris Cité, INSERM U970, Paris Institute for Transplantation and Organ Regeneration, Paris, France
2. Kidney Transplant Department, Saint-Louis Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France
3. Department of Nephrology and Organ Transplantation, Toulouse Rangueil University Hospital, Toulouse, France.
4. Néphrologie-Immunologie Clinique, Hôpital Bretonneau, CHU Tours, Tours, France.
5. Department of Pathology, Necker-Enfants Malades Hospital, Assistance Publique - Hôpitaux de Paris, Paris, France
6. Université de Paris, Assistance Publique-Hôpitaux de Paris (AP-HP), Service d'Anatomie et Cytologie Pathologiques, Hôpital Bichat, Paris, France
7. Department of Nephrology, Bichat Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France
8. Department of Nephrology, Service de Néphrologie, CHU de Lille, Hôpital Huriez, 59037, Lille, France
9. Inserm, CHU Lille, Institut Pasteur Lille, University in Lille, U1190 - EGID, 59000, Lille, France
10. Department of Pathology, Lille University Hospital, France
11. Department of Nephrology, Dialysis and Transplantation, Montpellier University Hospital, France
12. Department of Pathology, Bordeaux University Hospital, France
13. Department of Nephrology and Transplantation, Bordeaux University Hospital, France
14. Department of Nephrology and Kidney Transplantation, Vall d'Hebrón University Hospital, Barcelona, Spain
15. Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA
16. Center for Transplantation Sciences, Massachusetts General Hospital, Boston, MA
17. Department of Medicine, Division of Nephrology, Comprehensive Transplant Center, Cedars Sinai Medical Center, Los Angeles, CA
18. Department of Kidney Transplantation, Necker Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France
19. Department of Laboratory Medicine and Pathology, ECHA 5-411, University of Alberta, 11405 87 Avenue, Edmonton, Alberta T6G 1C9, Canada

List of abbreviations

AMR, antibody-mediated rejection

B-HOT, banff human organ transplant panel

BKV, BK virus

C4d, complement component C4d

DSA, donor-specific antibody

FFPE, formalin-fixed paraffin-embedded

HK, Housekeeping genes

HLA, human leukocyte antigen

IQR, interquartile range

KNN, K-Nearest Neighbors

LASSO, least absolute shrinkage and selection operator

LDA, linear discriminant analysis

NR, no rejection or injury related diagnosis

NRKI, non-rejection kidney injury

PCA, principal component analysis

PR-AUC, precision-recall area under the curve

RCC, reporter code count

QC, quality control

RLE, relative log expression

ROC-AUC, receiver operating characteristic area under the curve

SVM, support vector machine

TCMR, T cell-mediated rejection

ABSTRACT

Transcriptomic analysis of kidney biopsies has demonstrated potential to improve diagnosis of allograft rejection. Here, we developed a molecular assessment of antibody-mediated rejection (AMR) and T-cell-mediated rejection (TCMR) based on the Banff-Human-Organ-Transplant (B-HOT) consensus gene panel. Expression assays of formalin-fixed paraffin-embedded kidney biopsies from well-phenotyped cohorts were used to develop prediction models for AMR and TCMR and an automated report of gene expression-based diagnosis. The study population consisted of 950 kidney allograft biopsies from 10 transplantation centers in Europe and North America. The development cohort included 664 renal allograft biopsies split into a training (n=537) and test set (n=127), and two external validation cohorts (n=286). We performed gene selection using regularized regression and developed several different base models based on B-HOT expression data, which were combined into a single ensemble model for each rejection diagnosis. Model performance was assessed in the test set and the two external validation cohorts, showing good discriminative abilities (respective PR-AUC AMR=0.811, 0.891, 0.832 and TCMR=0.736, 0.810, 0.782). We identified challenging biopsies with histology below diagnostic thresholds for which gene expression-based probability can refine rejection diagnosis. This automated molecular diagnostic system shows potential for improving kidney allograft rejection diagnosis in routine practice and clinical trials.

INTRODUCTION

Diagnosis of kidney transplant rejection based on the international Banff Classification of allograft pathology has enabled improved diagnostic accuracy and patient stratification, aiding clinicians in treatment decisions[1]. Despite continuous improvements over the past 30 years, the Banff Classification has become highly complex, with most rejection-associated histological lesions using categorical scoring, which can result in information loss for continuous biological processes. Further, histology requires specialized expertise with limited sensitivity to assess the heterogeneity and dynamic range of injury and rejection and has considerable inter- and intra-observer variability[2,3].

In the past two decades, whole transcriptome microarray-based studies of transplant biopsies have defined the molecular phenotypes of allograft rejection and molecular classifiers have demonstrated improved diagnosis of rejection and stratification of patients at high risk of graft failure. Gene expression analysis provides a highly reproducible mechanism-based evaluation that can detect cellular responses before they are visible from histology, thus more precisely capture disease activity and degree of injury.

Molecular profiling of kidney allograft biopsies was integrated into the Banff Classification in 2013[4] in order to improve diagnostic accuracy, in particular for antibody-mediated rejection[5–7]. However, microarray analysis has several barriers that limit its application in a clinical setting, including increased technical variation due to cDNA conversion, amplification, and labeling, unnecessary profiling of thousands of genes leading to large data processing times and storage costs, the need for a separate sample taken at the time of biopsy specifically for microarray analysis, and extended sample turn-around times due to central referral lab testing needs[7–9].

In 2019, the Banff Molecular Pathology Working Group defined and developed the Banff Human Organ Transplant (B-HOT) panel, a subset of 770 informative genes, which were demonstrated to be strongly associated with the relevant phenotypes in human allograft biopsies based on peer-reviewed studies[7]. Genes in the B-HOT panel cover the core pathways and processes related to tissue damage (inflammation, injury, viral infection), tissue rejection, and immune response. The B-HOT panel is and open-source gene list and is commercially available (NanoString Inc., Seattle, WA, USA, now part of Bruker Spatial Biology, Billerica, MA, USA) as a

standardized reagent kit to be operated on the NanoString nCounter™ platform. This assay is validated[10] for reproducible use in formalin-fixed, paraffin-embedded (FFPE) tissues samples and thus can be applied to the same biopsy sample used for routine histology. This enables direct comparison and integration of conventional histology and molecular diagnostics without any burden to the patient for procuring additional biopsy material. Furthermore, the nCounter™ platform is regulatory approved for clinical use, demonstrates inter-laboratory reproducibility, allows for simultaneous processing of multiple samples, facilitates decentralized clinical testing, while being cost effective[7]

Augmenting histology-based assessment of rejection through molecular diagnostics with continuous probabilistic scores can facilitate the assessment of challenging cases[11]. Gene expression based scores can also be used to detect early stage rejection in protocol biopsies that would otherwise go undetected[12]. Rather than the histopathological classification-based approach defined by semi-quantitative scoring of lesions and DSA assessment, expression-based diagnostics provide a continuous probability, allowing assignment of a level of confidence to the prediction beyond existing threshold-based systems.

The aims of this study were to 1) develop gene expression based predictive models for AMR and TCMR using the B-HOT panel in FFPE allograft biopsies from a large, well-phenotyped, multicenter cohort of kidney transplant recipients, 2) validate the generalizability of the model in additional external validation cohorts, and 3) create an automated diagnostic system requiring only raw gene expression count data to improve reproducibility, interpretability, and decentralized access.

METHODS

Study design and population

The multicenter international cohort study consisted of 950 kidney allograft biopsies, divided into a development and two external validation cohorts. The development cohort consisted of 664 renal biopsies collected from 603 patients in 8 different centers in France, Spain, and the United States, between 2004 and 2021. External validation cohorts included 243 biopsies from centers in the United States (Massachusetts General Hospital) and 43 biopsies from Canada (University of Alberta). This study was approved by local institutional review boards. Mixed rejection cases were excluded due to the co-occurrence of different inflammatory infiltrates, complicating the development of models that can accurately distinguish AMR and TCMR. Biopsies equivocal/suspicious for AMR or TCMR were also excluded as the aim is to improve diagnosis of these cases.

Clinical, biological and immunological data

At time of transplantation, clinical and biological data were collected related to i) recipient characteristics, ii) donor characteristics, iii) transplant characteristics, iv) immunological data, and v) immunosuppressive treatment. At time of kidney allograft biopsies, a standardized transplant assessment was performed, comprising i) clinical examination, ii) immunosuppressive treatment, iii) blood and urinary analyses for standard of care laboratory parameters, and iv) immunological phenotyping.

Kidney biopsy histological assessment

Kidney allograft biopsies were performed according to local centers' practice. Biopsy cores were formalin-fixed and paraffin-embedded for histological analysis. C4d staining was performed by immunohistochemistry on paraffin-embedded tissue or by immunofluorescence on frozen tissue according to local practices. Biopsies were assessed by local pathologists, and centrally reviewed according to the international and standardized Banff Classification for kidney allograft rejection[7,13]. To apply consistent rejection diagnosis across cohorts, we used the Banff 2019

Classification. In order to assess the impact of the Banff 2022 Classification[14] updates we performed a sensitivity analysis, testing model performance on the internal validation cohort and American validation cohort using the updated rejection labels. For these analyses, we grouped probable AMR and microvascular inflammation/injury, DSA-negative and C4d-negative under the diagnostic category of AMR. As the TCMR classification remained unchanged, model performance was not re-evaluated.

RNA extraction and expression assay

Biopsies from the development and validation cohorts were processed and assayed in their respective centers. Samples from the American cohort were processed as detailed in[12]. For all other cohorts, RNA was extracted from FFPE tissue section curls using the RNeasy FFPE Kit (Qiagen #73504) following the manufacturer's protocol. RNA was eluted in RNase-free water and samples were quantified using a Nanodrop 2000 and diluted to 20ng/µl. RNA samples were stored at -80C. Probes were hybridized to the Human Organ Transplant Panel (XT-CSO-HOT1-12) following the manufacturer's protocol (NanoString Technologies Inc.) and processed on the NanoString nCounter (MAX/FLEX system). Following imaging and counting, gene names were assigned to fluorescent probes using the Reporter Library File (NS_Hs_Transplant_v1.0).

Quality control and data normalization

Reporter Code Count (RCC) files containing raw count data for each biopsy were merged followed by quality control assessment and exclusion of outliers based on all samples in each cohort. Normalization was performed independently for each dataset – training, test, and each external validation cohort – to preserve separation of preprocessing and normalization across datasets **(Supplementary Figure 1)**. Housekeeping (HK) gene stability was calculated using the geNorm method[15]. We also performed a negative binomial regression for each rejection diagnosis (AMR or TCMR) and all housekeeping genes and excluded genes significantly associated with these outcomes of interest (Bonferroni adjusted $p < 0.05$). Housekeeping gene stability was assessed independently and excluded from normalization for each cohort. Expression data were corrected for technical variation with the remaining housekeeping genes using the remove unwanted

variation (RUV, k = 2) approach[16,17]. The number of unwanted factors (k) was selected empirically for each cohort by iterating over a range of values (k = 1–5) using PCA and RLE plots to evaluate reduction of technical variation, and clustering and differential expression analyses to assess preservation of biological signals. Normalization differences between centers were assessed based on the first principal component. PC1 scores were examined to ensure correlation with biological variables (rejection diagnosis) rather than technical factors (batch/center). The Kruskal-Wallis test was used to evaluate PC1 score associations, and the k value that best balanced removal of unwanted variation while preserving biological signal was chosen for each cohort.

Outcomes of interest

The outcomes used in model development were based on histological assessment of renal allograft biopsy tissue according to the international Banff 2019 Classification[18]. Overall diagnostic categories included antibody-mediated (AMR) and T cell-mediated rejection (TCMR) as well as no rejection or injury related diagnosis (NR) and non-rejection kidney injury (NRKI) as described in detail in the results section. The same Banff Classification rules were applied to all reference biopsies to determine the rejection diagnosis and to develop gene expression based predictive models for AMR and TCMR.

Diagnostic model development and validation

The development cohort (n=664) was randomly partitioned (80:20) based on histological diagnosis into a training (n=537) and test set (n=127) with equal proportions of each histological diagnosis. Predictive models for AMR and TCMR were developed based on expression data in the development cohort. First, feature selection for each outcome was performed by repeated k-fold cross-validation (n=10; k=5) of LASSO regularization using all 758 endogenous B-HOT genes in the development cohort for AMR versus all other diagnoses and TCMR versus all other diagnoses. Rather than a single model based on one optimal metric, genes with non-zero coefficients in at least 25 out of 50 total iterations of the cross-validation procedure were retained. Our primary objective was to build robustness through consensus rather than optimize performance through increasingly complex models. To this end, several predictive base models were trained for each

outcome of interest based on each gene set, including standard and regularized logistic regression, linear discriminant analysis (LDA), and linear support vector machine (SVM). Hyperparameter optimization was performed for each base model by repeated k-fold cross-validation (n=10; k=5). Discrimination of each base model as well as the median score across all base models were evaluated using the area under the precision-recall curve (PR-AUC), area under the receiver operating characteristic curve (ROCAUC), log loss, and Brier score[19].

Performance of all base and ensemble models was assessed in the test set (n=127) and 2 external (n=243 and n=43) cohorts. We evaluated the predicted probabilities for all internal and external validation samples compared to expected rejection diagnosis. All discordant biopsies were then evaluated in the context of clinical and histological data. Model performance was evaluated using the Youden index, which balances sensitivity and specificity, minimizes misclassification, and avoids potential biases associated with fixed thresholds.

Given that biopsies within the intermediate probability range (25-75%) often reflect challenging cases, an arbitrary threshold of 50% was applied to address the uncertainty in this range when assessing inconsistencies between histology and gene expression-based probability estimates. Discrepancies between histology and gene expression-based probabilities were defined as follows: AMR histology with pAMR < 50%; TCMR histology with pTCMR < 50%; no rejection histology with pAMR and/or pTCMR > 50%.

Model development and validation details are reported according to the TRIPOD statement[20].

Development of an automated diagnostic report

We developed an automated diagnostic report (*HistoM*x report) to facilitate reproducibility and interpretability of gene expression-based diagnosis that requires only raw count data from biopsies assayed with the B-HOT panel.

An HTML file is generated using R Markdown[21] which is then converted into a paged PDF using wkhtmltopdf[22]. A test sample is normalized by multiplying by the quotient of the arithmetic mean of the geometric mean of housekeeping genes within each sample in the development cohort divided by the geometric mean of housekeeping genes in the new sample. Base models

for AMR and TCMR are then applied to normalized expression data to generate new predictions for the tested sample. Final probabilities represent the median of the base models and include a sample range indicating the confidence interval for the median. For comparison, we also include the interquartile range (IQR) of probabilities for AMR and TCMR in non-rejection biopsies in the development cohort, as well as an interpretation of the median probability. The report includes Principal Component Analysis (PCA), projecting the tested biopsy on PC1 and PC2 among the samples used in model development, as well as k-nearest neighbors (KNN) with histology-based diagnosis for the 25 closest biopsies based on Euclidean distance. Assay run information, as well as quality control, including binding density, fields of view, limit of detection (LoD), positive control linearity, percent of endogenous genes above the LoD, signal to noise (ratio of the geometric mean of housekeeping genes divided by the LoD), total housekeeping genes below the geometric mean of negative control genes, are included in each report with recommended cutoffs.

All statistical analyses were performed using R (version 4.0.5 R Foundation for Statistical Computing).

RESULTS

Patient and biopsy characteristics

The development cohort consisted of 664 allograft biopsies from 603 kidney transplant patients. The mean recipient age was 49.95 ± 1.12 years, with 401 (60.39%) males. A total of 117 patients (17.62%) had a prior kidney transplant and 17 (2.56%) were ABO incompatible. The median time between kidney transplantation and biopsy was 1 year (interquartile range (IQR) 0.25-1.08). Detailed baseline characteristics, including recipient, donor, and biopsy characteristics, histological lesions, and diagnoses are presented in **Table 1** and **Supplementary Tables 1-4**. Biopsies were split into a train (n=537) and test (n=127) set. All samples in the primary analysis include comprehensive clinical, demographic, and histology data, allowing necessary review of discrepancies and ultimately clinical validity of gene expression-based models.

Biopsy characteristics and related diagnoses

The development cohort consisted of 193 AMR (129 active AMR, 61 chronic active AMR, 3 chronic inactive AMR), 149 TCMR (101 acute TCMR, 48 chronic active TCMR), 127 NR (118 normal or minimal changes, 9 pristine), and 195 NRKI (113 Isolated IFTA ≥ 2, 25 BK virus nephropathy, 17 ATI without rejection, 13 glomerulonephritis (recurrent or de novo), 27 other). Banff lesion scores and diagnoses are shown in **Table 2**. The American validation cohort consisted of 243 biopsies (detailed cohort characteristics can be found in [12]): 95 chronic active AMR, 47 acute TCMR, 85 NR, and 16 NRKI and the Canadian validation cohort consisted of 43 biopsies: 25 AMR and 18 TCMR (**Supplementary Tables 7-8**). The internal test cohort closely resembles the development cohort (**Supplementary Tables 5-6**), while the external validation cohorts differ for certain patient and biopsy characteristics (**Supplementary Figures 2-3**). The Canadian cohort primarily included for-cause biopsies and younger recipients compared to the development and American validation cohorts. The American cohort showed a longer median time from transplant to biopsy, consistent with the higher incidence of chronic rejection cases. All biopsies were assayed using the B-HOT panel and the same QC and normalization procedures were applied independently to each cohort.

Development and validation of gene expression-based classifiers of AMR and TCMR

Gene selection was performed for each outcome using LASSO regression based on normalized counts for all endogenous B-HOT genes (n=758) resulting in 52 genes for AMR and 15 for TCMR (**Supplementary Figure 4**). The predicted probability of AMR or TCMR for each base model and the final ensemble score in the development cohort are shown in **Figure 1**, with AMR and TCMR model performance metrics shown in **Supplementary Table 9**. Ensemble predictions for biopsies in all validation cohorts are presented in **Supplementary Figure 5**. Principal Component Analysis of the ensemble scores for all biopsies in the development cohort are presented in **Supplementary Figure 6**.

As the aim was to provide continuous probabilities rather than binary diagnostic cutoffs, ensemble model performance was primarily evaluated using the Brier score which compares predicted probabilities to actual class labels and assesses both discrimination and calibration (a Brier score of 0 represents perfect accuracy and a score of 1 represents perfect inaccuracy), as well as the PR-AUC to evaluate model discrimination at a variety of thresholds. Model reproducibility and generalizability was evaluated in one internal and two external validation cohorts (**Table 3**). The median of the base models for each outcome was compared to the histology based diagnostic class. AMR metrics were as follows: Brier score (test=0.132; USA=0.148; Canada=0.169) and PR-AUC (test=0.811; USA=0.891; Canada=0.832) and TCMR metrics: Brier score (test=0.103; USA=0.125; Canada=0.186) and PR-AUC (test=0.736; USA=0.81; Canada=0.782). The final ensemble model consistently matched or outperformed individual base models, suggesting that combining predictions across diverse algorithms can improve model stability and generalizability (**Supplementary Tables 10-12**).

Discrepancy analysis

An arbitrary diagnostic cutoff of 50% applied to our B-HOT based model predictions resulted in discrepancies in 30% of biopsies in all validation biopsies and the following in individual cohorts: train=19%; test=22%; American=34%; Canadian=30%.

Discrepancies between histology and expression-based probabilities were primarily challenging cases for which molecular analysis could be useful, including abnormal histology

below current Banff diagnostic thresholds, non-specific lesions, tissue scarring, non-TCMR related inflammation and tubulitis, BKV nephropathy, AMR pathology without C4d or DSA, and post-treatment follow-up biopsies (**Figure 2**).

Automated reporting of gene-expression based rejection diagnosis

We provide examples of molecular reports for samples from the validation cohorts that demonstrate the potential clinical utility of gene expression profiling of allograft biopsies (**Figure 3**). The report accepts free text related to sample information as well as RNA quality control indicators. Since allograft injury and rejection represent continuous disease processes, *HistoMx* reports provide probabilistic scores rather than binary cutoffs to help address the uncertainty of current consensus based, categorial histology-based diagnosis. Prediction scores for each model (AMR and TCMR) are based on the median of each ensemble model. However, simply reporting the central tendency of each score ignores the high variability that can arise among different predictions for the same biopsy, which is primarily observed in the middle quartiles of predicted scores. Quantifying the uncertainty of each score can improve the interpretability of molecular based predictions. We thus return confidence intervals, which provide a range that estimates the reliability of the prediction as well as the interquartile range (IQR) of each score for biopsies with no injury or rejection diagnosis as a reference. While standard model validation allows removal of batch effects, as samples are normalized by cohort, the diagnostic report relies on an n+1 approach, normalizing and generating predictions for a single new sample. Spearman correlation coefficients between normalization approaches for each validation cohort were as follows for AMR: France=0.961; USA=0.963; Canada=0.922 and for TCMR: France=0.782; USA=0.868; Canada=0.655.

Banff Classification Updates

Diagnostic labels for 9/537 (1.7%) biopsies in the development cohort, 3/127 (2.4%) in the internal test cohort, and 6/243 (2.5%) in the American validation cohort, and 0/43 biopsies in the Canadian validation cohort were impacted by the updated Banff 2022 Classification[14] (**Supplementary Table 13**). We performed sensitivity analysis of the updated internal test cohort

and American cohort labels to evaluate the impact of the additional diagnostic categories of probable AMR and microvascular inflammation/injury, DSA-negative and C4d-negative. AMR model performance metrics remained stable (**Supplementary Tables 14 and 15)**.

DISCUSSION

In this study, we developed B-HOT gene expression based predictive models for antibody and T-cell mediated rejection based on a deeply phenotyped cohort of kidney transplant recipients and donors. Models showed good performance in both internal and external validation cohorts, demonstrating generalizability across clinical settings despite inter-cohort differences. Discrepancies between histology-based diagnosis and gene expression-based predictions were identified and were attributed to abnormal histology below current Banff consensus diagnostic thresholds, limited specificity of Banff lesions (e.g., glomerulitis in recurrent glomerulonephritis or inflammation and tubulitis in BKV nephropathy), advanced tissue scarring making Banff scoring challenging, AMR pathology without C4d or DSA, and post-treatment follow-up biopsies. These findings support the notion that histological and molecular profiling may capture distinct yet complementary aspects of graft pathology. In this context, areas of diagnostic uncertainty related to the Banff Classification have the greatest potential for molecular diagnostics to add value in clinical practice. The B-HOT panel is an~~d~~ open-source gene list and expression data feeding into the *HistoMx* classifiers described here can theoretically be generated on any platform. However, the clinical grade NanoString nCounter™ platform allows processing of formalin-fixed, paraffin-embedded (FFPE) tissue~~s~~ samples and thus can be applied to the same biopsy sample used for routine histology. This enables integration of molecular diagnostics into routine transplant pathology in a timely and cost-effective manner, and thus at scale adoption and clinical validation for defined context of use for molecular transplant biopsy assessment.

The B-HOT gene panel has been demonstrated to reliably capture the relevant genes and pathways associated with AMR and TCMR in renal allografts[12,23,24] and is further supported by the performance of the models developed in this and previous studies[25,26]. While the B-HOT panel consists of an expert-curated consensus gene set based on peer reviewed evidence, feature selection was performed to reduce collinearity as well as overfitting during model development. Associations are reproducible as long as biopsies are diagnosed with the same Banff rules, the same comparators are used, and the same model applied[27].

In addition to comparing predictions to expected histological diagnoses, evaluating the performance of models developed based on an uncertain ground truth requires re-assessing

cases showing discrepancies between histology and gene expression-based probabilities. A cohort with comprehensive clinical, demographic, and histology data is indispensable to review discrepancies as it provides the granularity needed to capture the complexity of allograft rejection[28]. Inconclusive discrepancies could be due to early stage rejection[12,29], ischemia-reperfusion injury in early post-transplant biopsies, incomplete histology, sampling error, or mRNA assay quality. We used a more stringent cutoff (50%) and observed a similar rate of discordance between *HistoMx* (30%) and the microarray based *MMDx* (37%)[30].

We developed an automated diagnostic system that generates a report detailing sample quality and molecular scores for the probability of AMR and TCMR. The cohort used in model development consists of a clinically representative population of kidney transplant biopsies, which is critical to address the limited challenge bias in which a 'sick versus well' design is employed, leading to poor generalizability in a clinical setting. Multi-disciplinary expertise was implemented at every stage of model development and validation to address both potential risks and benefits for patients and ensure interpretability by clinicians in routine practice.

Models were trained using Banff 2019 diagnostic labels. Sensitivity analysis applying Banff 2022 Classification to the internal test and American validation cohorts showed stable classifier performance. However, the Banff 2022 Classification does not yet clearly define whether the novel categories – microvascular inflammation/injury, DSA-negative and C4d-negative and probable AMR – should be included under AMR in molecular classifier development. Further research is needed to clarify their molecular profiles and overlap with established AMR signatures, and dedicated studies are required to validate classifier performance in these novel categories.

This study has several limitations: 1) Models were developed using only FFPE samples, which have the advantage of using the same sample from routine diagnosis. While there is a robust association between platforms, further studies are needed to address the known technical differences in tissue sample preservation, such as RNA later versus FFPE[31]. 2) Models were evaluated using standard multi-sample QC and normalization. In a decentralized clinical setting, the use of panel standards to correct for technical variation when processing individual samples are desirable [32]. 3) The development cohort contained nearly twice as many active compared to

chronic rejection, which can limit model accuracy for these cases. While there is currently no consensus on treatment of active versus chronic rejection, histology complexity increases with time post-transplant, due to co-occurrence of chronic damage and active conditions.

In conclusion, we developed and validated the performance of B-HOT gene panel-based predictive models for kidney allograft rejection. We embedded these models into an automated reporting system to facilitate their use in routine clinical practice. This first version of *HistoMx* molecular classifiers will be regularly updated with multidisciplinary collaborative input and major updates to the Banff Classification. Compared to other commercially available products, the advantage of the here described molecular diagnostic system, is that it can be operated in a local molecular pathology lab on the same biopsy material as processed for routine histology. Furthermore, the system is built on the consensus B-HOT panel, which is also used in the international Banff consensus for histologic diagnosis, together allowing for decentralized but standardized histo-molecular diagnostic work-up, for example in the setting of multicenter studies or to support patients referred between centers. The combination of the B-HOT panel with automated reporting advances molecular transplant diagnostics into an accessible and cost-effective methodology that will enable collaborative efforts necessary to generate the appropriate context of use. To fully realize this potential, prospective and randomized studies of standardized interventions comparing additional molecular testing to current standard of care would be highly valuable in further validating and optimizing use of these tools to improve the outcomes of patients living with a kidney transplant. The decentralized approach described here using routine FFPE biopsies represents a critical enabler to this end.

AUTHOR CONTRIBUTIONS

AL, DZ, VGoutaudier, MS, MMengel, and RC designed the study, performed data analysis, and wrote the manuscript. DZ, VGoutaudier, JG, GD, MS, OA, JD, BR, FM, AC, MR, JPDVH, AS, CRL, BC, AV, LC, MLQ, MMaanaoui, ALionet, JBG, VGnemmi, OB, KL, IR, RNS, RC, VLK, DA, EH, ADB, NK, BA, MMengel, CL, AL contributed to data acquisition and interpretation. All authors reviewed the manuscript.

CONFLICT OF INTEREST

The authors declare no competing interests.

DATA AND CODE AVAILABILITY

All minimum data and code to reproduce the figures are deposited into the synapse public repository https://doi.org/10.7303/syn66272947 and are freely available. A sign-in process is required to access the data. Full source data are available from the corresponding author upon request.

REFERENCES

1. Loupy, A., Mengel, M. & Haas, M. Thirty years of the International Banff Classification for Allograft Pathology: the past, present, and future of kidney transplant diagnostics. *Kidney Int.* **101**, 678–691 (2022).
2. Furness, P. N. *et al.* International variation in histologic grading is large, and persistent feedback does not improve reproducibility. *Am. J. Surg. Pathol.* **27**, 805–810 (2003).
3. Furness, P. N., 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.* **60**, 1998–2012 (2001).
4. Haas, M. *et al.* Banff 2013 meeting report: inclusion of c4d-negative antibody-mediated rejection and antibody-associated arterial lesions. *Am. J. Transplant* **14**, 272–283 (2014).
5. Halloran, P. F., Famulski, K. S. & Reeve, J. Molecular assessment of disease states in kidney transplant biopsy samples. *Nat. Rev. Nephrol.* **12**, 534–548 (2016).
6. Loupy, A. *et al.* Molecular microscope strategy to improve risk stratification in early antibody-mediated kidney allograft rejection. *J. Am. Soc. Nephrol.* **25**, 2267–2277 (2014).
7. Mengel, M. *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. *Am. J. Transplant* **20**, 2305–2317 (2020).
8. Haas, M. Molecular diagnostics in renal allograft biopsy interpretation: potential and pitfalls. *Kidney international* vol. 86 461–464 (2014).
9. Allanach, K. *et al.* Comparing microarray versus RT-PCR assessment of renal allograft biopsies: similar performance despite different dynamic ranges. *Am. J. Transplant* **8**, 1006–1015 (2008).
10. Adam, B. *et al.* Multiplexed color-coded probe-based gene expression assessment for clinical molecular diagnostics in formalin-fixed paraffin-embedded human renal allograft tissue. *Clin. Transplant.* **30**, 295–305 (2016).
11. Madill-Thomsen, K. S. & Halloran, P. F. Precision diagnostics in transplanted organs using microarray-assessed gene expression: concepts and technical methods of the Molecular Microscope® Diagnostic System (MMDx). *Clin. Sci. (Lond.)* **138**, 663–685 (2024).
12. Rosales, I. A. *et al.* Banff Human Organ Transplant transcripts correlate with renal allograft pathology and outcome: Importance of capillaritis and subpathologic rejection. *J. Am. Soc. Nephrol.* **33**, 2306–2319 (2022).
13. Yoo, D. *et al.* An automated histological classification system for precision diagnostics of kidney allografts. *Nat. Med.* **29**, 1211–1220 (2023).
14. Naesens, M. *et al.* The Banff 2022 Kidney Meeting Report: Reappraisal of microvascular inflammation and the role of biopsy-based transcript diagnostics. *Am. J. Transplant* **24**, 338–349 (2024).
15. Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, RESEARCH0034 (2002).
16. Bhattacharya, A. *et al.* An approach for normalization and quality control for NanoString RNA expression data. *Brief. Bioinform.* **22**, (2021).
17. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
18. Loupy, A. *et al.* The Banff 2019 Kidney Meeting Report (I): Updates on and clarification of criteria for T cell- and antibody-mediated rejection. *Am. J. Transplant* **20**, 2318–2331 (2020).
19. Huang, Y., Li, W., Macheret, F., Gabriel, R. A. & Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Inform. Assoc.* **27**, 621–633 (2020).

20. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* **350**, g7594 (2015).
21. *Rmarkdown: Dynamic Documents for R*. (Github).
22. Wkhtmltopdf. https://wkhtmltopdf.org.
23. Smith, R. N. *et al.* Utility of Banff Human Organ Transplant gene panel in human kidney transplant biopsies. *Transplantation* **107**, 1188–1199 (2023).
24. Beadle, J. *et al.* Application of the Banff Human Organ Transplant Panel to kidney transplant biopsies with features suspicious for antibody-mediated rejection. *Kidney Int.* (2023) doi:10.1016/j.kint.2023.04.015.
25. Giarraputo, A. *et al.* Banff Human Organ Transplant consensus gene panel for the detection of antibody mediated rejection in heart allograft biopsies. *Transpl. Int.* **36**, 11710 (2023).
26. Giarraputo, A. *et al.* Relevance of the Banff human organ transplant consensus gene panel for detecting antibody and T-cell-mediated rejection of kidney allografts. *Kidney Int. Rep.* **9**, 2290–2294 (2024).
27. Halloran, P. F. *et al.* Review: The transcripts associated with organ allograft rejection. *Am. J. Transplant* **18**, 785–795 (2018).
28. Halloran, P. F. *et al.* The molecular phenotype of kidney transplants. *Am. J. Transplant* **10**, 2215–2222 (2010).
29. Sablik, M. *et al.* Microvascular inflammation of kidney allografts and clinical outcomes. *N. Engl. J. Med.* (2024) doi:10.1056/NEJMoa2408835.
30. Madill-Thomsen, K. *et al.* Discrepancy analysis comparing molecular and histology diagnoses in kidney transplant biopsies. *Am. J. Transplant* **20**, 1341–1350 (2020).
31. Toulza, F. *et al.* Technical considerations when designing a gene expression panel for renal transplant diagnosis. *Sci. Rep.* **10**, 17909 (2020).
32. Wallden, B. *et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med. Genomics* **8**, 54 (2015).

FIGURES AND TABLES

Table 1: Baseline patient characteristics in the study population (n=603 kidney transplant patients).

Recipient characteristics	
Age (years), mean (SD)	49.6 (14.5)
Sex male, n (%)	362 (60.0%)
Cause of end-stage renal disease, n (%)	
Autosomal dominant polycystic kidney disease	85 (14.1%)
Diabetes	46 (7.6%)
Glomerulonephritis	168 (27.9%)
Polycystic kidney disease	1 (0.2%)
Tubulo-interstitial	58 (9.6%)
Vascular	37 (6.1%)
Unknown etiology	136 (22.6%)
Other	72 (11.9%)
Donor characteristics	
Age (years), mean (SD)	53.9 (17.0)
Sex male, n (%)	285 (55.1 %)
Deceased donor, n (%)	468 (77.7%)
Expanded criteria donor n (%)	199 (33.8%)
Transplant baseline characteristics	
Prior kidney transplant, n (%)	111 (18.4%)
Cold ischemia time (hours), mean (SD)	15.5 (10.0)
Delayed graft function, n (%)	114 (18.9%)
HLA-A/B/DR mismatches, mean (SD)	3.4 (1.5)
DSA at time of transplant, n (%)	138 (22.9%)
ABO incompatible transplant, n (%)	14 (2.3%)

Table 2. Biopsy characteristics in the development cohort (n=664 kidney allograft biopsies).

Time from transplant to biopsy (months), median (IQR)	12.0 (10.0)
Indication of biopsy, n (%)	
Protocol	364 (54.8%)
Clinically indicated	288 (43.4%)
Banff diagnosis, n (%)	
Active AMR	129 (19.4%)
Acute TCMR	101 (15.2%)
ATI without rejection	17 (2.6%)
BK virus nephropathy	25 (3.8%)
Chronic active AMR	61 (9.2%)
Chronic active TCMR	48 (7.2%)
Chronic inactive AMR	3 (0.5%)
Glomerulonephritis (recurrent or de novo)	13 (2.0%)
Isolated IFTA ≥ 2	113 (17.0%)
Normal or minimal changes	118 (17.8%)
Other	27 (4.1%)
Pristine	9 (1.4%)
Banff lesion scores	
Glomerulitis (g) score, n (%)	
0	456 (68.7%)
≥1	198 (29.8%)
Peritubular capillaritis (ptc) score, n (%)	
0	389 (58.6%)
≥1	230 (34.6%)
Double contour (cg) score, n (%)	
0	577 (86.9%)
≥1	70 (10.5%)
Interstitial inflammation (i) score, n (%)	
0	462 (69.6%)
≥1	189 (28.5%)
Tubulitis (t) score, n (%)	
0	365 (55.0%)
≥1	286 (43.1%)
Vasculitis (v) score, n (%)	
0	554 (83.4%)
≥1	58 (8.7%)
Vascular fibrous intimal thickening (cv) score, n (%)	

0	177 (26.7%)
≥1	386 (58.1%)
Interstitial fibrosis (ci) score, n (%)	
0	194 (29.2%)
≥1	454 (68.4%)
Tubular atrophy (ct) score, n (%)	
0	189 (28.5%)
≥1	459 (69.1%)
Interstitial fibrosis/tubular atrophy (IFTA) score, n (%)	
0	199 (30.0%)
≥1	411 (61.9%)
Total inflammation (ti) score, n (%)	
0	365 (55.0%)
≥1	239 (36.0%)
Inflammation in areas of IFTA (i-IFTA) score, n (%)	
0	306 (46.1%)
≥1	203 (30.6%)
Mesangial matrix expansion (mm) score, n (%)	
0	459 (69.1%)
≥1	133 (20.0%)
Arteriolar hyalinosis (ah) score, n (%)	
0	211 (31.8%)
≥1	399 (60.1%)
C4d score, n (%)	
No	444 (66.9%)
Yes (≥1 by IHC or ≥2 by IF)	47 (7.1%)
Allograft function	
Serum creatinine (umol/L), mean (SD)	179 (145)
eGFR (mL/min/1.73 m2), mean (SD)	42.8 (20.7)

Figure 1: Predicted probabilities of AMR or TCMR in rejection and non-rejection biopsies.

Plots of individual base model and ensemble (median) probabilities for each biopsy in the development cohort training subset (n=537). Each panel includes biopsies with the indicated histology-based diagnosis of non-AMR (n=381), non-TCMR (n=417), AMR (n=156), and TCMR (n=120). Dashed lines represent the median ensemble score for all biopsies in each group.

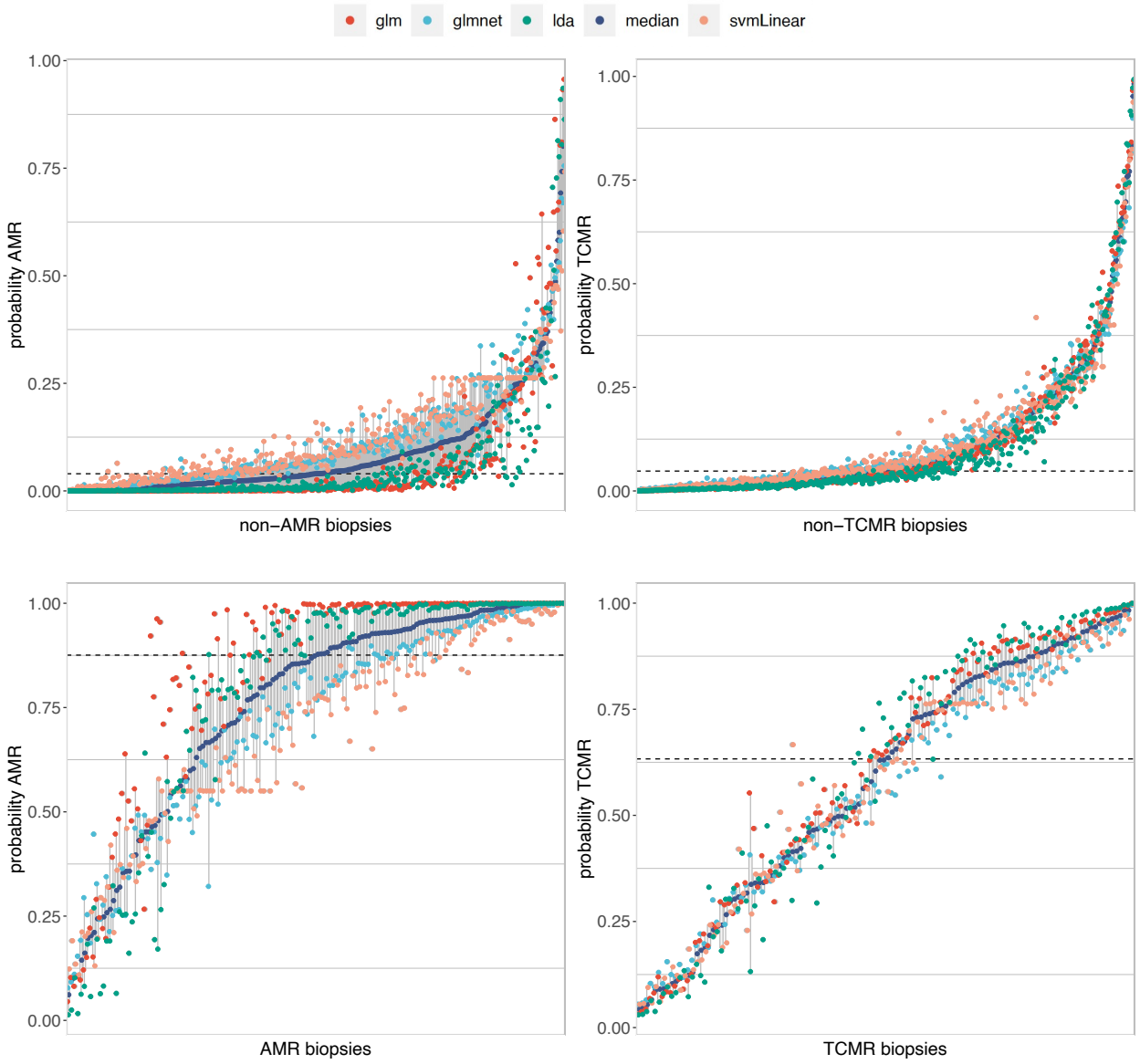


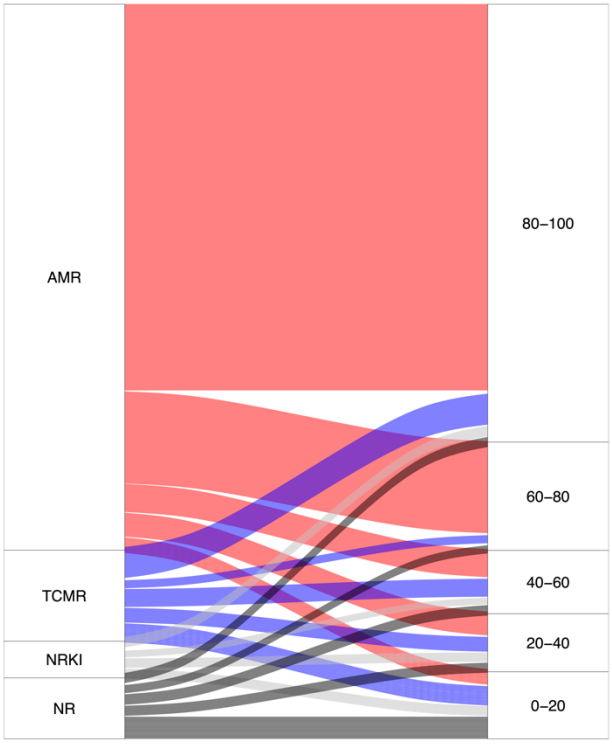
Table 3: Performance of rejection models in internal and external validation cohorts.

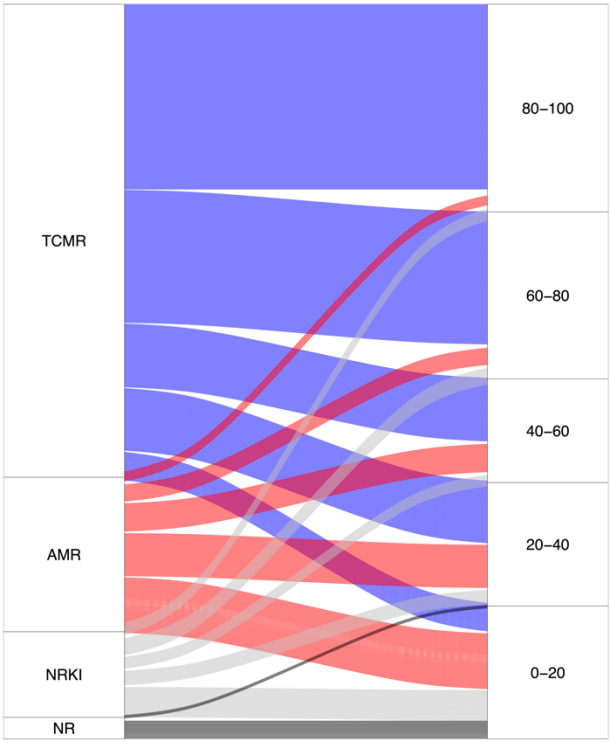
AMR and TCMR ensemble model performance metrics in the internal test (France) and external validation (USA and Canada) cohorts. The PR (Precision-Recall) baseline represents the proportion of positive samples in each cohort. Model performance of class assignment was determined by using Youden's index as the threshold based on the development cohort training subset ROC curve. PPV: positive predictive value; TPR: true positive rate; TNR: true negative rate; FDR: false detection rate; FPR: false positive rate, FI, F1 score which computes the average of precision and recall.

	AMR			TCMR		
	France	**USA**	**Canada**	**France**	**USA**	**Canada**
Brier score	0.132	0.148	0.169	0.103	0.125	0.186
log loss	0.078	0.075	0.094	0.07	0.054	0.108
PRAUC	0.811	0.891	0.832	0.736	0.81	0.782
PR baseline	0.291	0.391	0.581	0.228	0.193	0.419
ROCAUC	0.864	0.9	0.831	0.891	0.939	0.776
Youden index	0.348	0.348	0.348	0.296	0.296	0.296
Balanced accuracy	0.8	0.792	0.71	0.796	0.657	0.678
accuracy	0.795	0.831	0.744	0.835	0.864	0.698
PPV	0.612	0.935	0.719	0.618	0.938	0.667
TPR	0.811	0.611	0.92	0.724	0.319	0.556
TNR	0.789	0.973	0.5	0.867	0.995	0.8
FDR	0.388	0.064	0.281	0.382	0.062	0.333
FPR	0.211	0.027	0.5	0.133	0.005	0.2
F1	0.698	0.739	0.807	0.667	0.476	0.606

Figure 2. Histology-based diagnosis and predicted gene expression based probability.

Sankey diagrams depicting histology-based diagnosis and gene expression based predicted probability for AMR (left panel) and TCMR (right panel) for each biopsy in all validation cohorts (n=413). Tables describe observed discordances between histology-based rejection diagnoses and predicted probabilities and explanations for these discrepancies. AMR=antibody-mediated rejection; TCMR=T-cell mediated rejection; NRKI=non-rejection kidney injury; NR=non-rejection.





Discrepancy	Explanation
(+) AMR histology and (-) AMR HistoMx	post-treatment response, no DSA and/or C4d, ischemia-reperfusion injury
(-) AMR histology and (+) AMR HistoMx	post-treatment non-response, early-stage rejection
AMR histology below Banff thresholds and (+) AMR HistoMx	DSA+ isolated v lesions, MVI and DSA/C4d status, MVI>2 in presence of recurrent or de novo glomerulonephritis

Discrepancy	Explanation
(+) TCMR histology and (-) TCMR HistoMx	post-treatment response, isolated v lesions, i-IFTA, ischemia-reperfusion injury
(-) TCMR histology and (+) HistoMx TCMR	post-treatment non-response, early-stage rejection, BKV
TCMR histology below Banff thresholds and (+) TCMR HistoMx	borderline TCMR, i-IFTA

Figure 3. Examples of *HistoMx* automated molecular reports. *Left panel:* a protocol biopsy 12 months post-transplantation with histology-based T-cell mediated rejection showing no signs of AMR or TCMR based on gene expression-based profiling. *Right panel:* a *for-cause* biopsy with histology based chronic active antibody-mediated rejection (negative for both C4d and DSA), supported by a very high probability of AMR based on gene expression profiling.

// **Sample info:**

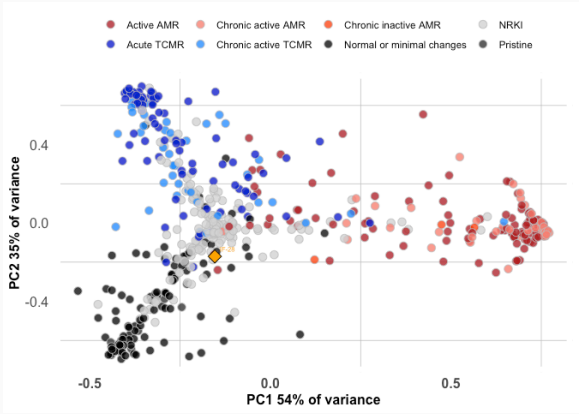
Biopsy indication	Protocol M12	Histology	Acute TCMR (g0 ptc0 i0 t0 v3 ci3 ct3 ifta3)

// **Gene expression based probabilities:**

Diagnosis	IF-28	Sample range	Normal range	Interpretation
AMR	7.1	1.3 - 13.9	0.5 - 6.4	unlikely
TCMR	2.3	1.3 - 3.7	1.5 - 8.5	unlikely

Diagnosis based scores represent the probability (0-100%) that a biopsy has an expression profile similar to reference samples with a given histology based rejection diagnosis (AMR or TCMR). 'Sample range' is the confidence interval for the predicted probabilities of the tested sample. 'Normal range' is the middle 50% of scores in biopsies with no rejection or injury diagnosis.

// **Principal Components Analysis (PCA) of molecular scores:**



// **Sample info:**

Biopsy indication	For cause	Histology	Chronic active AMR (g2 ptc1 cg1 i1 t0 v0 C4d0 DSA-)

// **Gene expression based probabilities:**

Diagnosis	NS-492	Sample range	Normal range	Interpretation
AMR	97.9	95.9 - 100	0.5 - 6.4	very high probability
TCMR	2.3	0 - 2.7	1.5 - 8.5	unlikely

Diagnosis based scores represent the probability (0-100%) that a biopsy has an expression profile similar to reference samples with a given histology based rejection diagnosis (AMR or TCMR). 'Sample range' is the confidence interval for the predicted probabilities of the tested sample. 'Normal range' is the middle 50% of scores in biopsies with no rejection or injury diagnosis.

// **Principal Components Analysis (PCA) of molecular scores:**