Journal of Transplantation Technologies & Research

First et al., J Transplant Technol Res 2017, 7:3
DOI: 10.4172/2161-0991.1000176

Research Article OMICS International

Analytical and Clinical Validation of a Molecular Diagnostic Signature in Kidney Transplant Recipients

Martin Roy First*, Deirdre Pierry, Michael McNulty, Sunil M Kurian, Stan Rose, Thomas Whisenant, Terri Gelbart, April Venzon, Nadia Bayat, Peter Lewis, John J Friedewald, Michael M Abecassis and Darren Lee

Comprehensive Transplant Center, Northwestern University, Chicago, USA

***Corresponding author:** Martin Roy First, Adjunct Professor of Surgery, Comprehensive Transplant Center, Northwestern University, 1201 W Wrightwood Ave, Chicago, Illinois 60614, USA, Tel: 7736778682; Fax: 7736778682; E-mail: roy@transplantgenomics.com

Received date: September 07, 2017; **Accepted date:** September 19, 2017; **Published date:** September 27, 2017

Abstract

Context: The TruGraf test is a blood-based assay that provides non-invasive, accurate assessment of adequacy of immunosuppression in kidney transplant recipients. TruGraf relies on gene-expression "signatures" that differentiate a state of Transplant eXcellence (TX, indicating adequately immunosuppressed) from not-TX.

Objective: To evaluate the performance of the TruGraf test.

Design: Analytical performance studies to characterize stability of RNA in blood during collection and shipment, analytical sensitivity (input RNA concentration), analytical specificity (interfering substances) and assay performance (clinical validity, and intra-assay, inter-assay, inter-laboratory reproducibility).

Results: Total RNA extracted from whole blood specimens collected in PAXgene Blood RNA tubes was stable up to 3 days at room temperature (stable RNA yield). Under routine ambient shipping conditions, storage and shipping temperatures did not affect results. However, specimen shipments exposed to temperatures >400°C or to ambient temperatures for >3 days were unacceptable for processing. Analytical sensitivity studies demonstrated tolerance to variation in RNA input (50 to 400 ng per 3' IVT (*in vitro* transcript] labeling reaction). Specificity studies using genomic DNA spiked into 3 'IVT reactions at 10-20% demonstrated negligible assay interference. The test was reproducible across operators, runs, reagent lots, and laboratories. External validation demonstrated that the TruGraf blood test accurately classified patients in 72% of 295 samples.

Conclusions: Analytical sensitivity, analytical specificity, robustness, quality control, and clinical validity of the TruGraf blood test were successfully verified, indicating its suitability for clinical use.

Keywords: Kidney transplant recipients; Gene expression signatures; Immunosuppression; TruGraf blood test

Introduction

The survival benefits of solid organ transplants in the United States are well documented [1]. Improvements in immunosuppression, better anti-infective agents, and improved ancillary care have resulted in significant improvements in short-term outcomes; however, there has been little improvement in long-term graft loss [2-4]. Data from the Organ Procurement Transplant Network/Scientific Registry of Transplant Recipients (OPTN/SRTR) 2015 Annual Data Report indicate that the kidney graft failure rate for deceased donor transplants is 4.8% at 6 months, 6.4% at 1 year, 14.6% at 3 years, 26.3% at 5 years and 52.84% at 10 years [5]. Over the past two decades, attrition rates in the first year post-transplant have shown significant improvement across all subgroups; however, there have been only minor improvements in attrition rates beyond the first year [3].

Post-transplant monitoring relies on serial serum creatinine (SCr) measurements and "protocol" or surveillance biopsies. However the SCr level is a highly insensitive indicator of the degree of damage in the kidney, has a lag time of weeks to months while on-going damage is occurring and changes are non-specific with regards to the cause of the injury. Surveillance biopsies are performed infrequently and with variable frequency after kidney transplantation. Biopsies, other than for-cause, are almost non-existent after two years, and when performed are invasive, expensive, and subject to inter-grader variability of approximately 30% [6]; therefore, performing invasive biopsies is not suitable for frequent monitoring. Currently, there is no validated test to measure or monitor the adequacy of immunosuppression, the failure of which may result in over-immunosuppression and opportunistic infections, or under-immunosuppression and acute rejection [7]. Recent reviews have highlighted the need for robust multicenter validation studies while underscoring the potential for biomarker monitoring of immunosuppressive therapy and transplant outcomes [8,9]. Molecular biomarkers have been studied in the graft, urine and blood of kidney transplant recipients [10-13].

Microarray analysis, a widely used technology for studying gene expression, is not routinely used as a diagnostic tool. Numerous studies have shown that microarray analysis results in improved diagnosis and risk stratification for conditions such as breast cancer [14].

In studies of diagnostic accuracy, the outcomes from a new test under evaluation ("new-test") is compared with outcomes from the reference standard, both measured in subjects who are suspected of

having the condition of interest [15]. New-test may refer to the methods used to obtain additional information on a patient's health status, including information from history and physical examination, laboratory tests, imaging tests, function tests and histopathology. The results of new-test may prompt clinical actions, such as further diagnostic testing, or the initiation or modification of treatment [15]. The term accuracy refers to the amount of agreement between the information from the new-test under evaluation and the reference standard. Diagnostic accuracy can be expressed in many ways, including sensitivity and specificity, likelihood ratios, diagnostic odds ratio, and the area under a receiver operator characteristic (ROC) curve [15]. Positive predictive value (PPV) and negative predictive value (NPV) should also be considered in evaluating the new-test [16].

The goal of method validation in the molecular diagnostics laboratory is to ensure that a given test is ready for implementation in the clinical laboratory, and that the analytical performance of this Laboratory Developed Test (LDT) is comparable between testing laboratories.

To reach that goal, each step of the testing process must be carefully evaluated and documented. TruGraf (Transplant Genomics Inc., Mansfield, MA) is a blood-based assay that provides non-invasive, accurate detection of adequacy of immunosuppression in kidney transplant recipients. TruGraf relies on analysis of gene-expression "signatures" (reference profiles of the expression levels of many genes associated with a given phenotype) that can differentiate a state of Transplant eXcellence (TX, indicating adequately immunosuppressed) from not-TX. With the current standards for monitoring after kidney transplantation, significant tissue injury can progress for months to years without being detected or treated accordingly and result ultimately in graft failure and return to dialysis or death. Through differential diagnosis of TX versus not-TX, the TruGraf blood test provides a noninvasive tool to support physicians in maintaining levels of effective immunosuppression and to help guide personalized treatment plans. Significant healthcare savings could be realized by optimizing each individual patient's therapy by ensuring adequacy of immunosuppression to protect the function and prolong the survival of their graft. This manuscript provides the first description of the analytical validation studies supporting the use of the TruGraf blood test as performed in the Transplant Genomics Inc Clinical Laboratory Improvement Amendments (CLIA) Laboratory.

Materials and Methods

The various steps involved in performing the TruGraf blood test are described below. At a high level, blood samples are obtained from patients, and RNA is extracted, amplified and hybridized to DNA microarrays. Arrays are washed, stained and scanned to detect levels of hybridization of sample RNA to specific oligonucleotide probes. A proprietary algorithm is used to analyze the pattern of hybridization, compare the results with a reference dataset, and generate a qualitative result of "TX" or "not-TX". The results of the TruGraf blood test may be used by a physician, in the context of other clinical information available, to assess whether or not a kidney transplant recipient is adequately immunosuppressed.

RNA extraction, amplification and hybridization

Total RNA was extracted from PAXgene Blood RNA (IVD) tubes (Qiagen, Valencia, CA). PAXgene tubes were processed using PAXgene Blood micro RNA (miRNA) reagents on the QIAcube instrument (Qiagen, Valencia, CA) Total RNA yield and concentration were determined using the Nanodrop 8000 (Thermo Fisher Scientific). Samples were processed to remove globin RNA using the Ambion GLOBINclear Human kit (Thermo Fisher Scientific, Carlsbad, CA). Globin-reduced RNA quantity was determined using the Nanodrop 8000 and quality was determined using the Bioanalyzer RNA Nano system (Agilent Technologies, Santa Clara, CA) to generate an RNA integrity number (RIN) [17]. RNA yield and quality thresholds were established and used as acceptance criteria for downstream sample processing. The Affymetrix 3' IVT (*in vitro* transcript) PLUS labeling system was used to perform *in vitro* transcription and labeling reactions (3' IVT) on globin-reduced RNA with a reaction input of 200ng (Affymetrix, Santa Clara, CA). Samples were fragmented and a final pre-hybridization RNA quality check was performed on labeled cRNA as well as the fragmented final cRNA product. Hybridization cocktails were prepared with an input of 7.5 µg of biotin-labeled cRNA. Array hybridization and subsequent washing, staining and array scanning steps were completed on Affymetrix HG-U133+ arrays using the standard GeneTitan Gene Expression array workflow (Affymetrix, Santa Clara, CA). A whole assay control (WAC) consisting of RNA from a subject with a known TruGraf response processed from RNA extraction through GeneTitan processing was utilized in addition to no template (Nuclease free water) and Affymetrix External RNA controls (Poly A RNA, B2 Oligo and 20x Hybridization Controls) as in-process controls for the RNA labeling and hybridization reactions. Raw expression data files (CEL), an ASCII text file used by Affymetrix software, and generated by the GeneTitan were processed for Quality Control (QC) metrics using the Affymetrix Expression Console software (build 1.4.1.46, Affymetrix). Predefined specifications for yield, array data quality and control sample classifier results were used as acceptance criteria prior to sample data being analyzed on the TruGraf Classifier.

The analytical validation studies that were performed, including the sample source, the study design and the data evaluated, are summarized in Table 1.

Study	Sample source	Design summary	Data evaluated
Analytical Sensitivity - LOD	HeLa Control RNA	LOD testing was performed on a dilution series (4 dilutions of 3' IVT Labeling reaction input concentrations and 4 Hybridization reaction input concentrations) of HeLa Control RNA samples. Sample data analysis was performed on the Affymetrix Expression Console software.	In-rocess QC Points: NanoDrop and RIN values. Hyb QC Results – includes RLE values and signal boxplots, background levels. Labeling and Hyb Control acceptability, GAPDH signal intensity, GAPDH 3-5 Ratio, Pearson correlations. LOD – 4 sample input concentrations for both the IVT Labeling and hybridization reactions.

Analytical Specificity Interference	HeLa Control RNA	RNA from the HeLa Control supplied with the 3' IVT was spiked with genomic DNA and processed thru array Hyb on the GeneTitan. Sample data analysis was performed on the Affymetrix Expression Console software.	In-process QC Points: NanoDrop and RIN values. Hyb QC Results – includes RLE values and signal boxplots, background levels, Labeling and Hyb Control acceptability, GAPDH signal intensity, GAPDH 3-5 Ratio, Pearson correlations Array CEL file data was analyzed on the Affymetrix Expression Console software. Resulting information about probeset intensity variation was used to evaluate effects of gDNA contamination on RNA specimen hybridization.
Accuracy (vs. biopsy results)	PAXGene Blood RNA from kidney transplant subjects	PAXgene Blood RNA from ~130 transplant subjects was obtained by the TGI CLIA Lab and processed thru the assay on the GeneTitan. Samples were randomized to one of several arrays in order to minimize processing bias. Molecular phenotype was compared to original histology results. In-process Hyb QC data was used to assess sample suitability.	In-process QC Points: NanoDrop and RIN values. Hyb QC Results – RLE values and signal boxplots, background levels. Labeling and Hyb Control acceptability, GAPDH signal intensity, GAPDH 3-5 Ratio, Pearson correlations. Array CEL file data was analyzed on the Affymetrix Expression Console software to generate Hyb QC data and on the TruGraf™ Classifier algorithm to generate IQ/IA scores.
Preanalytical Factors	PAXGene Blood RNA derived from 3 normal, non-transplant subjects (Sufficient blood was collected to allow for replicates of samples to be run).	Normal subject (NS) blood specimens in PAXgene tubes were obtained by the TGI CLIA Lab. Specimens were subject to varying preanalytic conditions and extracted. Downstream GLOBINclear, 3' IVT and array hybridization processing were performed on a single run.	In-process QC Points: NanoDrop and RIN values. Hyb QC Results – RLE values and signal boxplots, background levels, Labeling and Hyb Control acceptability, GAPDH signal intensity, GAPDH 3-5 Ratio, Pearson correlations. Array CEL file data was analyzed on the Affymetrix Expression Console software.
Reproducibility – Intra-assay	PAXGene Blood RNA derived from normal, non-transplant subjects (Sufficient blood was collected to allow for replicates of samples to be run).	Normal subject (NS) blood specimens in PAXgene tubes were obtained by the TGI CLIA Lab. Multiple replicates of 4 patient samples were processed on a single run.	In-process QC Points: NanoDrop and RIN values. Hyb QC Results – RLE values and signal boxplots, background levels, Labeling and Hyb Control acceptability, GAPDH signal intensity, GAPDH 3-5 Ratio, Pearson correlations. Array CEL file data was analyzed using RMA data from the Affymetrix Expression Console to generate hybridization metrics. Descriptive statistics were evaluated for reproducibility and precision.
Reproducibility – Intermediate Precision	PAXGene Blood RNA derived from normal, non-transplant subjects (Sufficient blood was collected to allow for replicates of samples to be run).	Normal subject (NS) blood specimens in PAXgene tubes were obtained by the TGI CLIA Lab. Replicates of 4 patient samples were run in duplicate on each of 8 separate runs. New reagent lots were rotated into the run schedule while holding the remaining reagent lots constant so that reagent effects could be pinpointed to the new reagent lot. At least 2 different lots of GeneTitan HG-U133+ GLOBINclear reagents, Qiagen PAXgene RNA (IVD), and 3'IVT Plus and Hyb/Wash/Stain reagents were used for this cohort.	In-process QC Points: NanoDrop and RIN values. Hyb QC Results – RLE values and signal boxplots, background levels, Labeling and Hyb Control acceptability, GAPDH signal intensity, GAPDH 3-5 Ratio, Pearson correlations. Array CEL files were analyzed using RMA data from the Affymetrix Expression Console to generate hybridization metrics. Descriptive statistics were evaluated for reproducibility and precision.
TruGraf™ Classifier Bioinformatics	Raw data files (.CEL)	Internal Validation – performed on ~120 data files from the original discovery data set. External / Independent Validation – 130 data files processed independent of the discovery set.	Array CEL files were analyzed using the TruGraf (v0.6) Classifier.
Abbreviations LOD: Limit of Detection; IVT: *In vitro* Transcription reaction; RIN: RNA Integrity Number; Hyb QC: Hybridization Quality Control; GAPDH: Glyceraldehyde 3-phosphate dehydrogenase; RLE: Relative logarithmic expression; CEL file: CEL is an ASCII text file used by Affymetrix software gDNA: genomic DNA			

Table 1: Summary of analytical validation studies.

The TruGraf classifier

The TruGraf classifier is based on a locked Support Vector Machine (SVM) model using the e1071 package, subsequently adapted for use in the R statistical computing environment. The Cohort 1/Validation Set was used to perform an internal validation of the classifier using the "leave-one-out" cross-validation and bootstrap resampling methods. External validation testing was performed on a second cohort of independent samples that were not used in the training dataset. Blood samples obtained from patients with biopsy-confirmed TX or not-TX phenotypes were used to perform this analysis (Table 2).

Samples	Discovery	Cohort 1	Cohort 2
TX	238	61	81
Not-TX	260	65	49
Total	498	126	130

Table 2: Discovery and validation data sets.

Results

Preanalytic conditions

One of the advantages of using the PAXgene Blood RNA system is that samples can be shipped overnight at ambient temperature without negative consequences for RNA quality. We examined the effects of potential shipping scenarios during the course of our validation in order to establish specimen acceptability criteria. The main indicator used to assess the quality of labeled RNA transcripts is the housekeeping gene Glyceraldehyde 3-phosphate dehydrogenase (GAPDH). This gene is expressed in most cell types at relatively consistent levels. The Affymetrix HG-U133+ chips include GAPDH probesets for use as assay control metrics. By comparing the signal intensities from the 3' probes to the 5' probes, it is possible to obtain a post-hybridization ("post-chip") measure of the integrity of labeled cRNA. If the resulting ratios are high, this indicates the presence of truncated transcripts. Using the GAPDH ratio as a quality indicator, we determined that samples that were subjected to elevated temperatures (>40°C) or extended periods at ambient temperatures (>3 days) prior to RNA extraction yielded degraded RNA unsuitable for downstream processing. Results from this cohort were used to establish specimen rejection criteria (specimens held at ambient temperature for >3 days, or specimens subjected to elevated temperatures, such as specimens that might get lost or delayed during shipping). Figure 1 shows the GAPDH ratio metrics for samples subjected to elevated preanalytic temperatures (sample aliquots E and F), as well as samples held at ambient temperature for longer than 3 days prior to extraction (sample aliquots C and D). The D, E and F samples are seen to be outliers when compared to other samples. The high ratios for these samples indicate poor RNA quality when compared to ratios for samples exposed to routine preanalytical conditions.

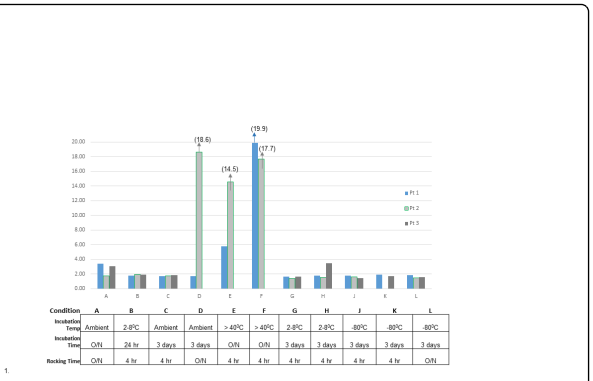

Figure 1: Pre-analytical effects of temperature and incubation length on RNA integrity showing glyceraldehyde 3-phosphate dehydrogenase (GAPDH) ratio metrics for samples subjected to elevated temperatures (E and F), and samples held at ambient temperature for >3 days prior to extraction (C and D).

In addition to GAPDH ratio data; the relative logarithmic expression (RLE) is a relative value measure of the deviation of a single array signal compared to a group average which is commonly used to assess variation in microarray studies. Low RLE values are characteristic of high reproducibility and unusually high values indicate outliers. 22 RLE signal data for various samples on this cohort confirm the limits of preanalytic conditions (Figure 2). In this figure, the RLE Boxplots for samples subjected to elevated preanalytic temperatures (sample aliquots E and F), as well as samples held at ambient temperature for longer than 3 days prior to extraction (sample aliquots C and D) showed greater signal variation when compared to samples exposed to routine preanalytical conditions.

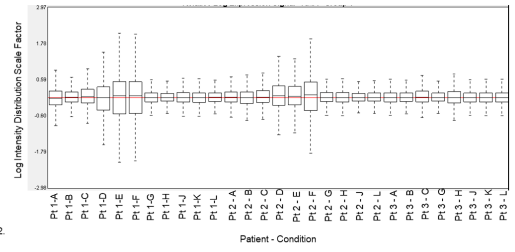

Figure 2: Relative log expression (RLE) Boxplots of pre-analytical samples illustrating that samples subjected to elevated temperatures (T4E and T4F), and held at ambient temperature for >3 days (T4D) had greater signal variation compared to routine samples. Abbreviations: Pt 1, Pt 2, Pt 3 = patients 1, 2, 3; A-L = replicate specimens.

Pearson's correlation data obtained by comparing signal data for all HG U133 Plus probesets again demonstrates the effect of subjecting samples to elevated temperatures or samples held for >3 days prior to extraction; r<0.900 compared to other samples in the cohort (Figure 3).

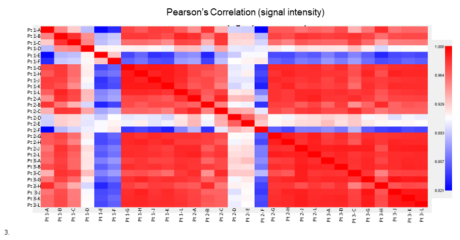

Figure 3: Pearson's correlation scores for pre-analytical samples comparing signal data for all HG U1333 Plus probesets again demonstrates the effect of preanalytical conditions. Samples T4E and T4F and sample T4D showed decreased similarity to aliquots subjected to routine preanalytical conditions. Abbreviations: Pt 1, Pt 2, Pt 3 = patients 1, 2, 3; A-L = replicate specimens.

Analytical sensitivity: RNA input

Two limit of detection (LOD) reactions were tested. Group T1A consisted of a set of 4 HeLa Control samples, run in duplicate designed to span the 3' IVT reaction input range from 50 ng to 400 ng of globin-reduced RNA. Affymetrix recommends a 3' IVT process input RNA range of 50 ng to 500 ng per 3' IVT reaction. Samples in this group had their hybridization reaction inputs normalized to 7.5 ug per reaction, creating a group of samples whose results were informative of the effects of differing RNA labeling reaction concentrations.

Group T1B consisted of a set of 4 HeLa Control samples, run in duplicate designed to span a hybridization reaction input range from 3 ug to 10 ug of fragmented, labeled cRNA. Affymetrix recommends a hybridization reaction input of 7.5 ug per hybridization reaction (the range of inputs for this group would therefore represent 40 to 133% of the recommended range). Samples in this group had their 3' IVT reaction inputs set at 200 ng/reaction, creating a group of samples whose results were informative of the effects of differing hybridization reaction concentrations. Review of the GAPDH 3' to 5' Ratio housekeeping gene metric for samples in this cohort demonstrated values ranging from 1.08 to 1.29 for these samples, indicating high quality labeled cRNA. QC analysis of sample data (using Affymetrix Expression Console software build 1.4.1.46) allowed for visualization of clear outliers, especially for the hybridization reaction extremes. Both RLE mean data and RLE signal boxplots indicate that hybridization reaction inputs of 3 µg and 10 µg per reaction result in outliers (Figures 2 and 3). Review of the signal intensity data for 3'IVT labeling and hybridization controls was similarly informative as decreased performance was seen for the higher concentration reactions for both labeling and hybridization reactions.

Analytical specificity: genomic DNA

The Affymetrix Gene Expression analysis workflow has several different purification steps designed to eliminate interfering substances. The RNA extraction process removes heme and DNA as a part of RNA purification process; the globin reduction process removes globin RNA and the purification step at the conclusion of the 3'IVT Labeling process removes unincorporated label and "left over" reagents, yielding pure labeled cRNA. Genomic DNA was tested as a potential interfering substance that might be present as a result of deviations from the standard RNA extraction process. We designed the samples processed to test the effects of 10-30% (by reaction input) genomic DNA spiked into the labeling reaction. The effects of the varied reaction input conditions were assessed through QC metrics. Probeset signal intensity variance, as expressed by CV% was reviewed for TruGraf classifier – informative probesets. These data indicate that 10-20% genomic DNA contamination of sample RNA does not interfere with individual probeset intensity. When the percentage of contaminating genomic DNA reached 30% individual probeset signal intensities showed increased variance as seen by higher CVs (Figure 4).

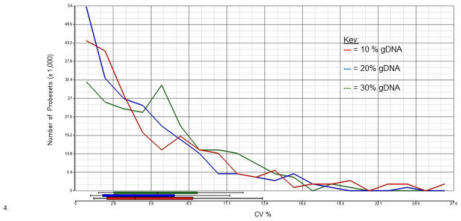

Figure 4: Intermediate precision effects on control probeset intensities indicating that probeset intensities are relatively consistent across lot-to-lot reagent changes. Raw probeset intensities are normalized as a part of routine data analysis as an additional means of controlling lot-to-lot variability.

Accuracy and reportable range

Independent clinical validation of the performance of the TruGraf assay was completed on a total of 295 patient samples collected in 3 sample cohorts of 126, 130 and 39 samples. Comparison of the TruGraf molecular phenotype measured from blood was made to the histological phenotype reported for a tissue biopsy collected at the same time as the corresponding blood sample was drawn. Results are shown in Table 3. Accuracy of the TruGraf blood test was 72% (95% confidence interval+0.01%). The sensitivity of 78% and positive predictive value (PPV) of 89% indicate that a "true TX" will be identified as TX positive in a high proportion of the intended clinical patient population using the TruGraf test.

Raw Data			
True Phenotype	n	TruGraf TX	TruGraf not TX
TX	163	127	36
Not TX	132	47	85
TOTAL	295	174	121
Statistics*	(n = 295)		
Accuracy	72% (212/295)		
Sensitivity	78%		

Specificity	65%
PPV	89%
NPV	45%

Table 3: Accuracy statistics *population prevalence of TX assumed to be 7.

Assay reproducibility

CEL file data for samples tested on the precision cohorts was analyzed using the Affymetrix Gene Console software. PLIER (Probe Logarithmic Intensity Error) analysis was performed to generate QC metric data. Data for the internal and external RNA Controls was used to assess technical performance of sample processing [18]. Descriptive statistics for external RNA Controls were used to assess precision for hybridization (Hyb) (20X Hyb Controls) and labeling (PolyA IVT Controls). Distribution statistics for average GAPDH signal intensity as well as the GAPDH 3' to 5' ratio were also used as internal sample metrics.

The samples processed on the Intra-run precision run were processed with the same lot numbers of reagents throughout the processing workflow. The samples were hybridized on the same Affymetrix HG-U133 Plus array plate. This created a group of 16 samples for use in gathering baseline statistics on within-run variability. All 16 samples on this run demonstrated control intensities within ± 2.5 S.D. of the mean intensities. Table 4 demonstrates the CV's for all external RNA controls were less than 10% indicating a high degree of reproducibility.

Control	PolyA-LYS	PolyA-PHE	PolyA-THR	PolyA-DAB	Hyb bioB	Hyb bioC	Hyb bioD	Hyb Cre
With-in Run	9.5%	6.6%	6.7%	3.4%	4.1%	3.0%	1.7%	0.6%
Between Run	14.4%	10.8%	12.1%	7.3%	4.5%	3.5%	2.5%	1.4%

Table 4: Average coefficient of variation (CV) for external RNA controls (within-run and between run).

Samples processed as part of the intermediate precision cohort were processed on 8 separate runs utilizing several reagent lots (including HG-U133 Plus lots) systematically changed to capture variation throughout the processing workflow. Review of mean intensity values for the Poly A and hybridization controls displayed the appropriate pattern of increasing signal values; reflecting the increase in relative concentration of the controls. All samples on these runs demonstrated control intensities within +2.5 S.D. of the mean intensities (Table 5). The CVs for the PolyA Controls (external RNA Labeling Controls) were <15%, with the LYS Control (present in copy number ratio of 1:100,000) showing the highest CV and the DAP Control (present in copy number ratio of 1:6,667) showing a CV below 10%. External hybridization RNA control CVs were less than 5% indicating a high degree of reproducibility; the highest CV was seen for the bioB Control, which is present at a concentration of 1.5 pM.

Figure 5 displays a graphical view of variation of representative housekeeping and control probeset intensities. Note that probeset intensities are fairly tightly clustered regardless of lot-to-lot reagent changes.

Validation Study	Qiagen RNA Kit	Ambion GLOBINclear	Affy 3' Kit	AFFY HWS Kit	HG-U133 Array	Affy Wash A	Affy Wash B
T5	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1
T6A	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1
T6B	Lot 2 Qiagen RNA	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1
T6C	148052863	Lot 2 Ambion GLOBINclear	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1
T6D	Baseline/Lot1	Baseline/Lot1	Lot 2 Affy 3' IVT Reagents	Baseline/Lot1	Baseline/Lot1	Lot 2 Wash Buffers	Lot 2 Wash Buffers
T6E	Baseline/Lot1	Baseline/Lot1	Lot 2 Affy 3' IVT Reagents	Baseline/Lot1	Lot 2 Arrays	Lot 2 Wash Buffers	Lot 2 Wash Buffers
T6F	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Lot 2 Wash Buffers	Lot 2 Wash Buffers
T6H	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Baseline/Lot1	Lot 2 Arrays	Lot 2 Wash Buffers	Lot 2 Wash Buffers

T6J	Baseline/Lot1	Baseline/Lot1	Lot 3 Affy 3' IVT Reagents	Baseline/Lot1	Lot 3 Arrays	Baseline/Lot1	Baseline/Lot1

Table 5: Reagent lot schedule for intermediate precision control runs.

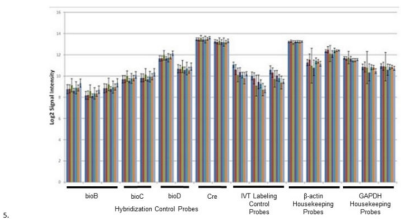

Figure 5: Probeset signal intensity variance, expressed by CV%, genomic DNA contamination of sample RNA.

TruGraf classifier development and validation

The TruGraf classifier algorithm (version 0.6) is a proprietary software package developed for use in the TGI automated bioinformatics pipeline. The TruGraf classifier is based upon previously published data. The current algorithm version has been locked, validated and implemented in TGI's CLIA laboratory workflow in the R statistical computing environment (version 3.1.2) [11-13]. The input for the software is an individual .CEL file generated by the Affymetrix GeneTitan instrument. Within the software, the data from the .CEL file is converted to a list of normalized gene expression values (signals) which correlates with the amount of RNA detected by each probeset on the Affymetrix GeneChip DNA microarray for the sample being analyzed. The values generated for a specific group of probesets present in the locked classifier are used by a locked Support Vector Machine (SVM) model (implemented from the e1071 R package version 1.6-6) which was trained on a discovery dataset (498 samples total) to generate a phenotypic classification / interpretation of Transplant eXcellence (TX) or not-TX for the sample.

Description of input and output data files or information in each process step

Input files are .CEL files generated by the Affymetrix GeneTitan which scans and processes Affymetrix HG-U133+ GeneChips. Output files from the Classifier algorithm are tab-delimited text files with 3 fields per sample processed: Sample ID, TX or not-TX interpretation and GAPDH 3-5' ratio.

Bioinformatic Validation – Internal and External Validation

The Internal validation of the classifier was performed with two methods: (i) leave-one-out cross-validation and (ii) bootstrap resampling. The leave-one-out cross-validation method iteratively removes each sample from the population, trains the model on the remaining samples, and predicts the class of the excluded sample. The accuracy is then calculated as a percentage of correctly classified samples in the discovery dataset. Each bootstrap resampling is a random subset of samples from the discovery dataset where some samples were represented more than once. Each bootstrapped dataset is then used to train a model that is tested on the samples not found in the bootstrap resampling population. On average, ~63% of the samples in the discovery dataset will be found in a given bootstrap iteration and the testing results are used to calculate an estimated error based on the number resampling iterations (i.e., 500. The "632plus" adjustment is then applied to the estimated error) [19].

The External validation was performed on two independent cohorts of TX and not-TX samples that were not used in the training dataset (Table 2). Each validation run of the bioinformatic pipeline used a set of sequentially named (i.e., S001 to S126) .CEL files as input and delivered the output as a tab-delimited text file with three fields for each of the samples. These fields are the sample name (i.e., S001 to S126), the classification of the sample, and an internal (non-reportable) score associated with the classification.

Result comparisons were based on use of the histologically determined phenotype from the sample's matched biopsy. Additionally, the algorithm installation in the TGI CLIA Lab was tested by running CEL files from the validation cohort #1 and the independent cohort #2 (Table 2) and comparing the results output to the previous algorithm output generated during the internal validation. Interpretation calls and score outputs of the external validation testing were compared to results generated during internal testing. Additional statistics (accuracy, sensitivity, and specificity) were calculated for the entire validation cohort and compared to the previous results (Table 6). Each of these metrics were calculated for the validation cohorts using TX or not-TX as the reference range.

Error rates were determined by comparing molecular phenotype results to the histological phenotypes of a matched biopsy obtained at the time of PAXgene blood sampling. Error rates for each of the validation cohorts were calculated using the following equation: 100% - % calculated accuracy).

Metric	Discovery	Cohort 1	Cohort 2
Accuracy	77.4%	77.0% (97/126)	71.5% (93/130)
Sensitivity	69.2%	70.9%	82.4%

Specificity	74.0%	83.3%	54.6%
Note: Accuracy table for the discovery cohort is not given because the accuracy for this cohort is the average of 100 bootstrap iterations each of which have their own accuracy metrics			

Table 6: TruGraf classifier validation statistics.

Figure 6 illustrates theTruGraf™ Classifier Bioinformatics Validation Schema. Error rates were determined by comparing molecular phenotype results to the histological phenotypes of a matched biopsy obtained at the time of PAXgene blood sampling; internal validation–23%, validation cohort 1–22.6% and independent validation cohort 2–30.2%.

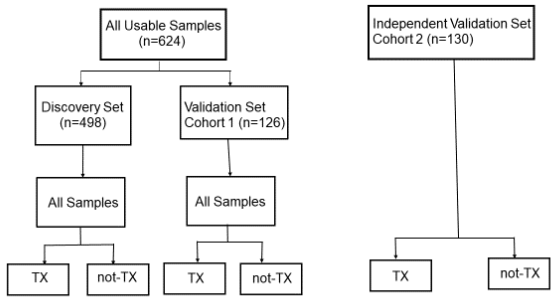

Figure 6: TruGraf classifier development and validation schema. Result comparisons were based on the TruGraf test result compared to the histological phenotypes from kidney transplant biopsy obtained at the same time as the blood specimen.

Discussion

In this report, we have demonstrated both the analytical validity and clinical validity of the TruGraf blood test and the robust nature of the assay utilizing a number of different metrics. The goal of method validation in the molecular diagnostics laboratory is to ensure that a given test is ready for implementation in the clinical laboratory. To reach that goal, multiple steps in the testing process have been carefully evaluated and documented. Clinical validity has been demonstrated based on a high correlation between TruGraf results and the current gold standard of care for assessing adequacy of immunosuppression, namely histology on tissue samples collected by biopsy of the transplanted kidney.

Predictive performance of the assay has been demonstrated in multiple cohorts, and at a level relevant to how the test will be used in the clinical situation. TruGraf is a blood-based assay that provides a non-invasive and an accurate assessment of adequacy of immunosuppression in kidney transplant recipients. TruGraf relies on analysis of gene-expression signatures that profile the expression levels of many genes associated with a given phenotype, thereby differentiating a state of Transplant eXcellence (TX, indicating adequately immunosuppressed) from not-TX. With the current standards for monitoring after kidney transplantation, significant tissue injury can progress for a prolonged period of time without being detected or treated accordingly, and result ultimately in graft failure and return to dialysis or death. TruGraf blood testing allows noninvasive serial monitoring of kidney transplant patients to detect indicators of adequacy of immunosuppression, which previously was only possible with insensitive trailing biomarkers of damage already done, or invasive procedures not suitable for serial monitoring. Many transplant centers perform invasive protocol biopsies with the inherent assumption that a patient will benefit, when, in fact, approximately 80% of these patients, all of whom have no signs of renal dysfunction, are determined to be immune quiescent. The primary intended use for the TruGraf test will thus be on patients with stable renal function after transplantation to confirm (rule in) a blood gene expression profile consistent with a state of sufficient or over-immunosuppression (TX). Considered in the context of the rest of the clinical information available to the physician, a TX result may support a decision to avoid costly and invasive protocol biopsies on the vast majority of patients who would not benefit from them. We have previously described the potential economic benefits of using the test in place of protocol biopsies [20]. Physicians will be able to use TruGraf results in combination with other laboratory test results and other clinical findings to help develop an individualized treatment plan based on each patient's unique biology and immune activity levels.

Through differential diagnosis of Transplant eXcellence, TruGraf provides a noninvasive tool to support physicians in maintaining effective levels of immunosuppression and help guide personalized treatment plans. In the process, patients will be spared unnecessary protocol biopsies, the healthcare system will realize significant economic benefits, and the ability to intervene early with therapies to fend off clinical acute rejection may provide the added benefit of improving long term outcomes.

Conclusion

In conclusion, the analytical sensitivity, analytical specificity, robustness, quality control and clinical validity of the TruGraf assay were successfully verified in these studies.

Acknowledgement

The authors of this manuscript have the following conflicts of interest to disclose: SMK, JJF and MMA are founding scientists and have stock ownership in Transplant Genomics Inc. MRF, DP, MM, SR, AV, NB, PL, and DL are full-time employees at Transplant Genomics Inc. The other authors have no conflicts of interest to disclose.

References

1. Rana A, Gruessner A, Agopian VG, Khalpey Z, Riaz IB, et al. (2015) Survival benefit of solid-organ transplant in the United States. JAMA Surg 150: 252-259.
2. Matas AJ, Gillingham KJ, Humar A, Kandaswamy R, Sutherland DER, et al. (2008) 2202 kidney transplant recipients with 10 years of graft function: what happens next? Am J Transplant 8: 2410-2419.

3. Lamb KE, Lodhi S, Meier-Kriesche HU (2011) Long-term renal allograft survival in the United States: a critical reappraisal. Am J Transplant 11: 450-462.
4. Montgomery RA (2014) One kidney for life. Am J Transplant 14: 1473-1474.
5. Hart A, Smith JM, Skeans MA, Gustafson SK, Stewart DE, et al. (2017) OPTN/SRTR 2015 annual data report: Kidney. Am J Transplant 17: 21-116.
6. Mengel M, Sis B, Halloran PF (2007) SWOT analysis of Banff: strengths, weaknesses, opportunities and threats of the international Banff consensus process and classification system for renal allograft pathology. Am J Transplant 7: 2221-2226.
7. Waldmann H (2014) Drug minimization in transplantation. Curr Opin Organ Transplant 19: 331-331.
8. Lo DJ, Kaplan B, Kirk AD (2014) Biomarkers for kidney transplant rejection. Nature Rev Nephrol 10: 215-225.
9. Willis JC, Lord GM (2015) Immune biomarkers: The promises and pitfalls of personalized medicine. Nat Rev Immunol 5: 323-329.
10. Suthanthiran M, Schwartz JE, Ding R, Abecassis M, Dadhania D, et al. (2013) Urinary-cell mRNA profile and acute cellular rejection in kidney allografts. N Engl J Med 369: 20-31.
11. Flechner S, Kurian S, Salomon D, Steven RH, Sharp SM, et al. (2004). Kidney transplant rejection and tissue injury by gene profiling of biopsies and peripheral blood lymphocytes. Am J Transplant 4: 1475-1489.
12. Kurian SM, Williams AN, Gelbart T, Campbell D, Mondala TS, et al. (2014) Molecular classifiers for acute kidney transplant rejection in peripheral blood by whole gene expression profiling. Am J Transplant 14: 1164-1172.
13. Modena B, Kurian SM, Gaber LW (2016) Gene expression in biopsies of acute rejection and interstitial fibrosis/tubular atrophy reveals highly shared mechanisms that correlate with worse long-term outcomes. Am J Transplant 16: 1982-1988.
14. Glas AM, Floore A, Delshaye LJMJ, Witteveen AT, Pover RCF, et al. (2006) Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC Genomics 7: 278.
15. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Clin Chem 49: 1-6.
16. Brenner H, Gefeller O (1997) Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med 16: 981-991.
17. Schroeder A, Mueller O, Stocker S, Ragg T (2006) The RIN: An RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Bio 7: 3.
18. Copois V, Bibeau F, Bascoul-Mollevi C (2007) Impact of RNA degradation on gene expression profiles: Assessment of different methods to reliably determine RNA quality. J Biotechnol 127: 549-559.
19. Efron B, Tibshirani R (1997) Improvements on cross-validation: The .632 Bootstrap Method. J Am Stat Assoc 92: 548-560.
20. First MR, Lee D, Lewis P, Rose S (2017) An economic analysis of the cost effectiveness of blood gene expression profiling in kidney transplant recipients. J Health and Med Econ 3: 2017.