A Bioinformatics Modeling Strategy for Efficient Migration of Candidate RNA Biomarkers to a Clinical Platform

OVERVIEW

Asuragen has developed bioinformatics tools for migrating candidate biomarkers discovered using a microarray platform to a clinically relevant, quantitative real-time PCR [reverse transcriptase (RT)-qPCR] platform. These tools take into account the fact that different platforms may measure the expression level for a gene differently due to platform-specific technical variance and noise. Without the consideration of such noise, an inefficient classifier may emerge from the platform migration. Successful migration requires the selection of biomarkers and models that will be most adaptable to platform transitions. This paper highlights the key steps needed to translate genomic signatures into clinically robust predictive models.

CONCLUSIONS

In this paper, we described our results for predicting those biomarkers with the highest cross-platform correlation and greatest likelihood for successful migration. Our data indicated that the migration of candidate biomarkers was affected by variance (i.e., expression difference) associated with both the discovery (microarray) and more diagnostic-ready (RT-qPCR) platforms.

In conclusion, bioinformatics modeling tools can enhance discovery efforts and the development and selection of biomarkers by improving the efficiency of biomarker migration, and by improving estimates of the number of samples needed to adequately power studies designed for such goals.

INTRODUCTION

Assay migration of expression markers from a high content platform (microarrays) to a low content platform (RT-qPCR) is a three step process that includes the identification of biomarker candidates using a discovery platform, migration of the biomarker candidates to a RT-qPCR platform, and validation of the biomarker classifier on a test set of samples (See Figure 1).

How many samples does the study need? Biomarker discovery is a prerequisite to any clinical test. In this step, it is critical to estimate the number of samples needed to generate a well powered predictive model. Sample size planning and statistical powering require that specific performance metrics [e.g., sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under the curve (AUC), etc.] be derived at the outset of the study in order to establish clear go or no-go criteria for each phase of the project. Once these aims are established, an initial microarray study is performed to determine if there are any significant differentially expressed genes between groups. If only small, insignificant differences in gene expression between groups are observed, then either more samples must be tested in order to increase the statistical power of candidate gene discovery or the study must be redesigned. An adequately powered study is then performed, enabling the selection of biomarkers for accurate classification of the discovery groups. When selecting biomarkers, it is important to consider the changes in model performance that accompany model and signature migration to a platform with inherently different detection properties.

Will the performance of the model change if it is migrated from microarray to RT-qPCR? The answer hinges upon the relative magnitudes of the technical noise from the two platforms. Correlation between RT-qPCR and microarray expression data is generally good, especially when identical transcript regions are targeted by both methods.[1] Probe sequence and relative transcript abundance have been observed to be important sources of variability between platforms.[2] However, in contrast to mRNA, prior studies have found weaker correlation of microRNA expression measurements between microarray and RT-qPCR[3], with low-expressing microRNAs in particular showing high variation between assays. In general, RT-qPCR measurements show better sensitivity and specificity for microRNA expression compared to microarrays.[3]

Our work addresses how the platform correlation and the noise between platforms can affect the power and the sample size estimates for both mRNA and microRNAs. This is important because the development of robust diagnostic assays relies upon the successful migration of biomarker candidates from high content discovery platforms, such as microarrays, to conventional diagnostic platforms such as RT-qPCR.

For example, to develop a binary classifier on a mRNA data set with the most significant genes exhibiting ~2-fold differential expression (~1 normalized log2-expression unit) and a microarray-estimated within-group standard deviation of ~0.5 normalized log2-expression units, recently developed methods suggest that a conservative estimate of the required number of samples is about 30 to generate a classifier whose accuracy is within 10% of the best possible if platform migration is not anticipated.[4] To get within 1% of maximal performance, the estimated number of samples that are needed increases to more than 50.[4] If the model is to be migrated to a different platform, however, and the differences in technical variance between the microarray and RT-qPCR platforms are taken into account, our methods yield a conservative estimate of 50 samples required for classifier development for accuracy within 10% of the theoretical maximum and 90 samples required for performance within 1% of optimum.[5]

This example of a conservative sample size estimate in anticipation of platform migration is derived from the assumption that the genes with the best performance on microarrays may exhibit larger technical variation when assayed by RT-qPCR. This assumption can, in some cases, be quite conservative, as many genes show higher technical noise when measured by

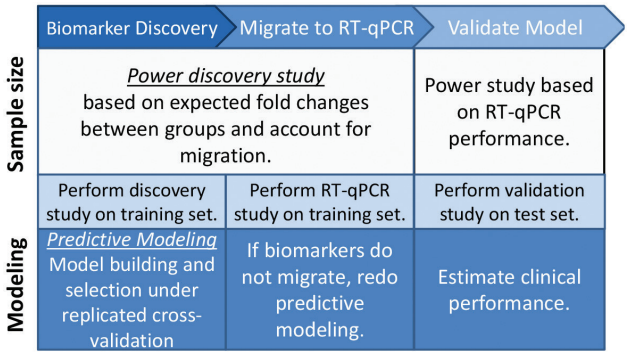

Figure 1. Overview of the assay migration process from microarrays to RT-qPCR. The additional noise from the qPCR platform must be considered when powering the original biomarker discovery study. Ideally, the migration would occur with the same samples used in the Biomarker Discovery phase. Finally, the model is validated on an independent sample set to establish clinical performance.

microarrays. While we generally recommend this conservative approach for sample size planning, we have found that it can be useful to estimate both "conservative" and "optimistic" bounds for migrated classifier performance, with the optimistic bounds (for which migrated performance will be superior to microarray performance) based on the opposite assumption that technical noise is much larger on the microarray platform than the RT-qPCR platform.

STUDY DESIGN

We have developed bioinformatics analysis methods and conducted several studies involving the migration of biomarker panels from a microarray platform to RT-qPCR platforms.

Data sets

Two data sets were used for this study:

1. An mRNA data set available in the public domain generated from 36 colon biopsies.[6,7]
2. A microRNA data set generated from 33 thyroid formalin-fixed paraffin-embedded (FFPE) specimens and processed at Asuragen.

Estimating required sample sizes for classifier construction on migrated platform

We applied bioinformatics methods to both data sets to estimate the target variance and to predict the number of samples required for training a classifier on a second platform.

Platforms

TaqMan RT-qPCR was used to assess expression levels of 67 mRNA and 34 microRNAs. Genome-wide mRNAs expression profiles were evaluated by HGU133 Plus 2.0 microarray (Affymetrix), while Agilent miRNA V.3 microarrays were used as the microRNA discovery platform.

Normalization

For the mRNA data set, data normalization was performed using RMA.[8] For the microRNA data set, data was normalized and filtered according to the standard in-house procedure[9] using the Variance Stabilization method.[10]

In order to align the microarray and PCR analysis, a normalizer-estimated normalization factor was subtracted from the RMA-normalized microarray expression values. This procedure aligned the microarray and RT-qPCR on similar scales except for a scale factor of -1. RT-qPCR data was reported as cycle threshold (Ct) values which are inversely proportional to the amount of target RNA whereas microarray signal intensity was positively correlated with concentration of target RNA.

RESULTS

Estimating the impact of platform migration on sample size estimates and model performance

Based on work by Dobbin et al. [4], we developed methods for determining the number of samples needed for a discovery study to build a classifier for a RT-qPCR platform. We estimated the number of samples required based on the relative magnitude of biological signal to platform migration-induced noise.

Figure 2 shows the results from theoretical sample size analysis which considered the influence of the attenuation factor. The "Attenuation Factor" indicates the ratio of a conservative estimate of the migrated standardized effect size to the array-estimated standardized effect size. Standardized effect size is defined as log-fold change divided by within-group standard deviation. This modeling shows that for classifiers based on biomarkers whose effect sizes are of similar magnitude to the noise levels associated with platform migration (roughly corresponding to the attenuation factor ~0.57 in the case plotted), the number of samples required to train a model to within a specified range of the optimal model performance (left panel) may be approximately double that estimated without taking migration into account. For example, with a standardized effect size of 2 (log-fold change twice as large as within-group deviation), the estimated required number of samples per group with an attenuation factor of 0.57 is 68, versus an estimated requirement of only 32 per group if no migration is required.

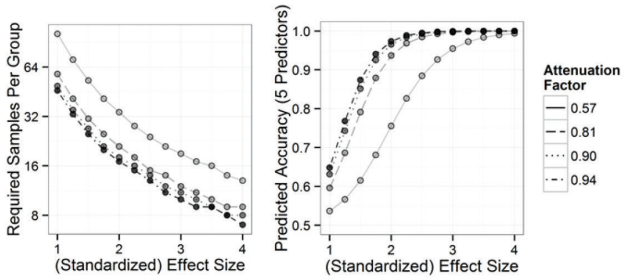

Figure 2. Theoretical modeling of sample number estimation based on the relative magnitude of biological signal to platform migration-induced noise. Left panel: Number of samples per group required to reach a standardized effect size. Right panel: Predicted classifier accuracies at varying standardized effect sizes and attenuation factor values where classifier is based on 5 biomarkers with 15 samples per group.

Furthermore, in such a high-platform-noise (low-attenuation factor) case, the optimal model performance (right panel) targeted by this powering methodology may also be severely degraded relative to expectations ignoring migration: it is important to know whether the model being powered has any chance of providing the desired level of performance, and platform migration should be taken into account.

Predicting cross-platform correlation

We have identified expression level variance as an important factor in predicting cross-platform correlation. Figure 3 captures the basic relationships between microarray and RT-qPCR data in terms of correlation between platforms. The biological variance is directly related to the platform correlation: as the platform correlation decreases and approaches perfect correlation between platforms [1.0 Pearson correlation coefficients (PCC)], the standard deviation (representing the differential measurement between the groups) increases. The platform correlation improved substantially when the biological signal exceeded twice the level of noise associated with platform migration (as represented by red vertical lines in Figure 3). The blue line represents the predicted Pearson correlation coefficient (PCC). This illustrates the use of the bioinformatics modeling tool for the prediction of how well the discovery and RT-qPCR platforms will correlate based on platform noise and the variance of a gene. In support of results reported by others[3], we also observed better agreement between microarray and RT-qPCR for mRNA than for the miRNA data.

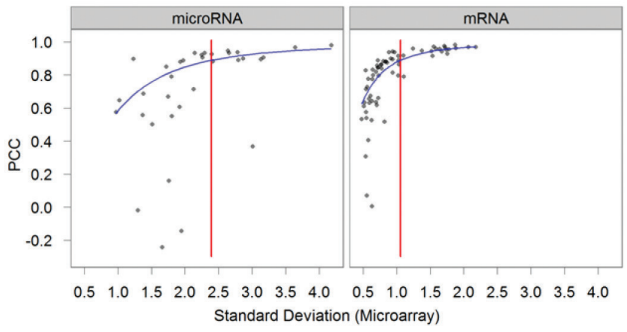

Figure 3. Important factors for predicting cross-platform correlation include expression level variation. Each point represents the Pearson correlation coefficient (PCC, absolute value) of microarray- and RT-qPCR log-expression values as a function of microarray log-expression standard deviation estimates. Red vertical lines are placed at twice the level of noise associated with platform migration. Blue lines illustrate PCC values predicted by modeling on the basis of standard deviation alone.

CONCLUSIONS

Biomarkers with the greatest differential measurement between groups (i.e. cancer versus benign) will have larger changes in expression levels, and such differences are more likely to be observed on another platform. This is especially important because a candidate biomarker determined to be highly significant from microarray analysis and hypothesis testing may be a poor candidate for a classifier on a RT-qPCR platform. This scenario arises if the candidate gene is significant only as a result of a modest effect size, combined with very small within-group variance. Our results show the trend of better platform correlation with larger variance. However, noisy platforms mute modest signals: therefore, to have an impact, this variance must be greater than the noise inherent in the platforms. We found that the correlation between microarray and RT-qPCR improves substantially when the biological signal exceeds twice the level of noise associated with platform migration. In our manuscript[5], we elucidate metrics to better predict whether a model (and its corresponding predictive power) can successfully migrate between platforms.

Bioinformatics modeling tools can enhance discovery efforts and the development and selection of biomarkers by improving the efficiency of biomarker migration, and by improving estimates of the number of samples needed to adequately power studies designed for such goals.

REFERENCES

1. Dallas, P.B., et al., BMC genomics, 2005. 6: p. 59.
2. Canales, R.D., et al., Nature biotech, 2006. **24**(9): p. 1115-22.
3. Chen, Y., et al., BMC genomics, 2009. 10: p. 407.
4. Dobbin, K.K. and R.M. Simon, Biostatistics, 2007. 8: p. 101-17.
5. Wylie, D. and A.T. Adai, 2013: In preparation.
6. Gyorffy, B., et al., PloS one, 2009. **4**(5): p. e5645.
7. Galamb, O., et al., Dis Markers, 2008. **25**(1): p. 1-16.
8. Irizarry, R.A., et al., Nucleic acids research, 2003. **31**(4): p. e15.
9. Davison, T.S., et al., Methods Enzymol, 2006. 411: p. 14-34.
10. Huber, W., et al., Bioinformatics, 2002. 18 Suppl 1: p. S96-104.

About Asuragen

Asuragen is a molecular diagnostics company using genomics to drive better patient management through best-in-class clinical solutions. The company uses a breadth of technologies and talent to discover, develop, and commercialize diagnostic products and clinical testing services with efficiency and flexibility both internally and for our companion diagnostic partners. Asuragen's products, services, and technologies drive countless patient management decisions across oncology, genetic disease and other molecular testing modalities.

ASURAGEN, INC. 2150 Woodward Street, Suite 100 • Austin TX 78744 • tel 512 681 5200 • fax 512 681 5201 • asuragen.com