

not know unless we continue to move these new technologies into the clinic. Even if CCR5-targeted gene therapy is not the ultimate solution, human studies are certain to be highly informative with regard to HIV pathogenesis and human immunology.

ACKNOWLEDGMENTS

The authors wish to acknowledge amfAR, Project Inform, TAG and the AIDS Policy Project for supporting and stimulating cross-disciplinary discussion on the issues outlined in this commentary. The authors' work that contributed to this review was supported by the National Institute of Allergy and Infectious Diseases (RO1 AI087145 and K24AI069994 to S.G.D. and R37 AI40312 and DPI OD00329 to J.M.M.), the University of California, San Francisco (UCSF) Center for AIDS Research (P30 MH59037), the UCSF Clinical and Translational Science Institute (UL1 RR024131), the Harvey V. Berneking Living Trust and amfAR. J.M.M. is a recipient of the National Institutes of Health (NIH) Director's Pioneer Award Program, part of the NIH Roadmap for Medical Research.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at <http://www.nature.com/naturebiotechnology/>.

- Holt, N. *et al. Nat. Biotechnol.* **28**, 839–847 (2010).
- Siliciano, J.D. *et al. Nat. Med.* **9**, 727–728 (2003).
- Kuller, L.H. *et al. PLoS Med.* **5**, e203 (2008).
- Phillips, A.N., Neaton, J. & Lundgren, J.D. *AIDS* **22**, 2409–2418 (2008).
- Friedman, A.D., Triezenberg, S.J. & McKnight, S.L. *Nature* **335**, 452–454 (1988).
- Baltimore, D. *Nature* **335**, 395–396 (1988).
- Rossi, J.J., June, C.H. & Kohn, D.B. *Nat. Biotechnol.* **25**, 1444–1454 (2007).
- McCune, J.M. *Nature* **410**, 974–979 (2001).
- McCune, J.M. *Cell* **82**, 183–188 (1995).
- Hutter, G. *et al. N. Engl. J. Med.* **360**, 692–698 (2009).
- Moore, J.P., Kitchen, S.G., Pugach, P. & Zack, J.A. *AIDS Res. Hum. Retroviruses* **20**, 111–126 (2004).
- Glass, W.G. *et al. J. Exp. Med.* **203**, 35–40 (2006).
- DiGiusto, D.L. *et al. Sci. Transl. Med.* **2**, 36ra43 (2010).
- Shimizu, S. *et al. Blood* **115**, 1534–1544 (2010).
- Perez, E.E. *et al. Nat. Biotechnol.* **26**, 808–816 (2008).
- Hunt, P.W. *et al. J. Infect. Dis.* **194**, 926–930 (2006).

Microarrays in the clinic

Guy W Tillinghast

The MicroArray Quality Control (MAQC) consortium has evaluated methods for making clinically useful predictions from large-scale gene expression data.

Clinical application of gene expression microarrays¹ and other 'omics technologies is widely expected to usher in a new era of personalized medicine. But although DNA microarrays are beginning to be used in patient care^{2,3}, progress has been slow, in part because of analytic challenges and concerns about accuracy and reproducibility. In this issue, the MAQC consortium presents the results of a large study, MAQC-II⁴, to evaluate methods for building genomic classifiers—software programs that convert microarray profiles of an individual sample into a prediction, such as membership in a clinical class. The results show that microarray algorithms can be reliable enough to justify clinical application, at least within certain contexts. More broadly, the findings of MAQC-II on microarray classifiers may be useful for analyzing data from other high-throughput assays.

Existing clinical predictors have well-known limitations, especially with respect to complex diseases such as cancer. Given two individuals who present identical clinical parameters, one

may respond to a therapy whereas the other may not. In principle, genome-wide data should be able to discriminate between them. The most common goals of a clinical test are to make a diagnosis or to determine an appropriate therapy. In light of statistical considerations, these goals depend on the prevalence of a disease, suggesting that clinical DNA microarray tests will augment, and not supplant, other clinical information. Thus, a possible strategy would be to first use traditional clinical predictors to broadly identify patients who might benefit from a treatment, and to then use an expensive assay, such as a microarray, to eliminate those for whom the treatment is unlikely to be effective.

Despite this promise, DNA microarrays have not been rapidly adopted in clinical practice. One reason is the noise that results from analyzing thousands of genes, which can lead to false predictions. Consequently, microarrays have been criticized because studies of the same clinical groups using different microarray measurements or analytic methods have often yielded dissimilar lists of differentially expressed genes. A second concern is the inherent error in the technology. Error stems from high background at the bottom of the dynamic

range, saturation at the top of the dynamic range, and nonlinearity, at least with measurements of some transcripts.

Many statistical methods have been developed to address these challenges, including approaches for grouping samples and genes, data normalization schemes to allow meaningful comparisons across samples, multiple testing procedures to select differentially expressed genes and 'cross-validation' methods for using samples to train prediction algorithms while reducing bias. These methods are applied sequentially to transform massive data sets of raw microarray gene expression profiles into clinically useful classifiers (Fig. 1a). As the optimal combination of methods is difficult to determine, MAQC-II sought to evaluate approaches to building classifiers.

Clinical use of microarrays is particularly challenging owing to the variability of the arrays themselves and to the variability between patients and between laboratories performing the analyses. These effects fall under the rubric of 'batch effects' and cause false positives. Moreover, before MAQC-II, it had not been clear whether classifiers trained on an initial data set would be able to make accurate predictions based on completely independent samples collected at a later date.

The five-step process for building a classifier in MAQC-II involved designing the experiment, collecting microarray data, creating a predictive model, validating the model internally with the training samples and validating the model externally with new samples obtained independently from the training data. MAQC-II enlisted 36 teams of data analysts within government agencies, academia and industry. The teams were given six microarray data sets and charged with predicting 13 'endpoints' potentially relevant to clinical or preclinical applications. The data sets included toxicological studies of chemicals on rodents and expression profiles of human cancer patients. In total, the teams built >30,000 classifiers using hundreds of combinations of analytic methods. A team of referees comprising biostatisticians and experienced data analysts chose one 'candidate' model that was expected to have the best performance for each endpoint from among models nominated by each of the 36 teams.

Next, the consortium analyzed how well the models classified samples. Performance was measured using several metrics, but the one most familiar to clinicians is the receiver operating characteristic area under the curve (AUC), a metric that varies between 0 and 1, where 0.5 indicates performance no better than chance and 1 means that all samples are correctly classified and none misclassified. For most of the endpoints, the candidate

Guy Tillinghast is at the Riverside Cancer Care Center, Newport News, Virginia, USA.
e-mail: guy.tillinghast@rivhs.com

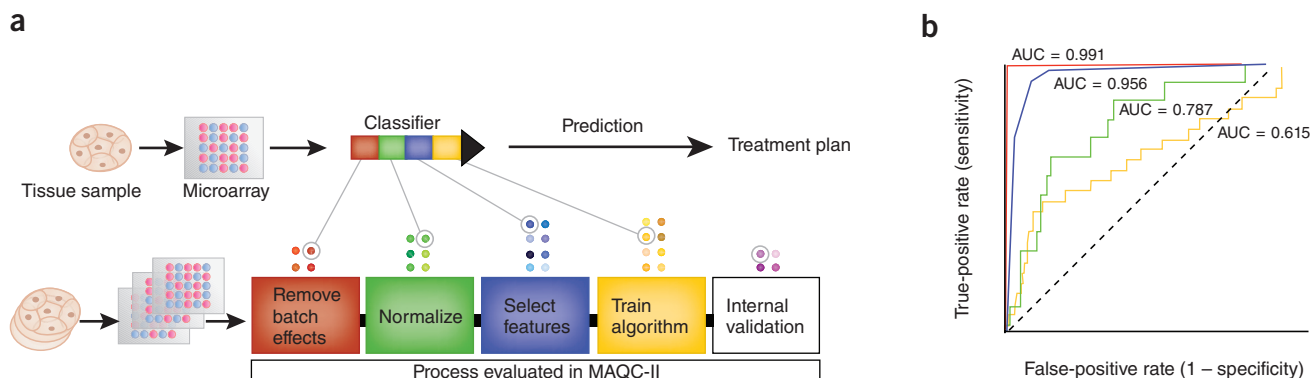


Figure 1 Using microarrays to make clinical predictions. **(a)** Current clinical decision-making processes can be refined by gene expression-based predictions generated by microarray classifiers (top). MAQC-II evaluated methods for constructing classifiers (bottom). Constructing a classifier from raw microarray data requires processing the data using a sequence of analytic steps (colored boxes). Many different approaches have been developed to solve each step (represented as dots above each box). In MAQC-II, >30,000 classifiers were constructed to test different combinations of analytic steps to predict 13 clinical and preclinical ‘endpoints’. **(b)** Curves showing the range of performance of classifiers developed for different data sets as part of MAQC-II. Performance is quantified using AUC. Data sets are characterized by the ratio of positive to negative samples in the cohort (P/N). Classifiers performed well for some endpoints, such as the sex of patients. The ~400 genes exclusively present on the Y chromosome made this an easy-to-predict positive control (red, training set P/N 1.44). The most difficult-to-predict endpoint was the overall survival of multiple myeloma patients, which has traditionally been difficult for other tests as well (orange, training set P/N 0.34). Classifiers for liver toxicity in rats (blue, training set P/N 0.58) and pathological complete remission in breast cancer (green, training set P/N 0.34) showed intermediate performance.

microarray-based classifiers performed far better than chance on the independent validation data set, with a range of 0.62–0.99. Moreover, the performance of the referee-selected candidate models was better than that of nominated models, suggesting that expert advice can enhance the modeling outcome.

Notably, classifier performance was found to depend heavily on the endpoint being predicted (Fig. 1b). However, it is evident from inspecting the data that there is a linear correlation between the AUC performance and the ratio of positive to negative samples in the cohort (‘training set P/N’). The composition of the training set is known to affect classification performance, and extreme imbalance, such as with the breast cancer and multiple myeloma endpoints (Fig. 1b, orange and green), may have adversely affected performance. Alternatively, the genetics of neuroblastoma and certainly the rodent data sets may be less variable and hence more tractable to modeling (Fig. 1b, blue). Moreover, genetic variation typically accumulates over time, making the genomes of the patients with breast cancer and multiple myeloma more variable than those with neuroblastoma and therefore less consistent with the reference genome from which the microarray platforms were constructed. These substantial differences in endpoints may have affected the validation AUC results.

Several findings from MAQC-II may help bring the technology closer to clinical use. Microarray experiments should be designed to minimize batch effects, such as those introduced by different laboratories or material lots. There should be a plan for detecting such

effects (e.g., by testing for unexpected genes that are expressed in different experimental conditions), and the same statistical test used to detect differentially expressed genes should be applied to all samples⁵. A gene that is differentially expressed in a pattern that matches the grouping of samples into batches should be examined closely and probably not used in a classifier.

Related to batch effects, quality control metrics should be used to distinguish variation in gene expression caused by laboratory artifact rather than by clinical phenotype. Quality control metrics are formulated to assess specific aspects of laboratory processing, such as RNA degradation or faulty equipment. These metrics can be used to adjust gene expression measurements or to identify problem microarrays. In the MAQC-II project, rather than adjusting measurements to account for laboratory noise, data analysts did not use samples that appeared to have quality control problems.

Several factors were found to influence classifier performance more than the type of algorithm used. One of these is the inherent difficulty of the biological phenomena being predicted. Another is the method for tuning the algorithm. Inexperience in tuning can be a major source of bias in the final classifier, especially if the predictive algorithm is not tuned for the population of interest. For example, in a population with low prevalence of a disease, it may be more desirable to have a test that makes few false predictions.

The results of MAQC-II highlight two priorities for future work. First, the field needs rigorous standards for reporting the steps

used to develop a classifier, its parameters of use and the appropriate quality metrics. Examples in the literature² may provide useful starting points. A classifier submitted for publication or for regulatory approval should specify how to use it to classify new samples—for example, the normalization and batch effect correction procedures to perform, the essential quality control checks and how to handle quality control flaws. The final report of a prediction algorithm should provide the variance (that is, standard error) of the performance measure as well as an estimation of the bias. A prediction report based on analysis of an individual patient sample should be accompanied by a report of quality metrics and their normal values and a report of batch effect measures that could provide a clinician with a sense of whether a microarray is within the range of the samples for which the test was developed⁵.

Second, methods are needed to combine microarray predictions with existing clinical decision-making tools, such as nomograms (a graphical chart for performing calculations). In constructing a nomogram, it will be necessary to determine how to balance the data from a microarray classifier with traditional clinical predictors. In addition, approaches should be developed to handle variability. For instance, the microarray chips used in MAQC-II have already been replaced by newer versions.

A key observation of MAQC-II—namely, that some endpoints seem inherently more predictable than others, regardless of the analytic methods used—suggests that gene expression microarrays may not capture a

sufficiently rich snapshot of disease physiology. In such cases, complementary technologies, which measure mRNA expression, protein levels, genetic mutation, copy number variation, gene silencing or regulatory RNA expression, could be considered. Alternatively, the best technology may vary by tumor type. High-throughput sequencing, in particular, offers advantages over microarrays in that coverage of the genome is less biased and the dynamic range is larger⁶. With luck, the results of MAQC-II will be useful for shepherding

other high-throughput technologies toward the clinic as well.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. DeRisi, J.L., Iyer, V.R. & Brown, P.P. *Science* **278**, 680–686 (1997).
2. Dumur, C.I. *et al. J. Mol. Diagn.* **10**, 67–77 (2008).
3. Buyse, M. *et al. J. Natl. Cancer Inst.* **98**, 1183–1192 (2006).
4. The MicroArray Quality Control (MAQC) consortium. *Nat. Biotechnol.* **28**, 827–838 (2010).
5. Luo, J. *et al. Pharmacogenomics J.* **10**, 278–291 (2010).
6. Schuster, S.C. *Nat. Methods* **5**, 16–18 (2008).

flank the targeted gene in the genome. Warner *et al.*¹ purchased the oligos, which were made on a programmable microarray. Next, using a clever cloning strategy, they appended the oligos to DNA elements that modulate gene expression. Attaching the targeting oligos to the strong $P_{\text{tetO-1}}$ promoter created a DNA cassette that was expected to upregulate the targeted gene after incorporation into the genome. Conversely, attaching the targeting oligo to a weak ribosome binding site produced a DNA cassette that downregulated the targeted gene. An antibiotic resistance gene allowed selection for the genetic modifications. As a result of the DNA synthesis and manipulation steps, Warner *et al.*¹ created two libraries of linear DNA fragments, each with 4,077 DNA cassettes pooled together in a single tube.

These libraries of DNA oligonucleotides were used to modify the *E. coli* genome by means of recombineering, a homologous recombination-based method in *E. coli* expressing λ phage recombination factors (λ gam, bet and exo)⁵. Growth on antibiotic medium selects for successful recombinants, and the sites of recombination are determined by homology of the targeting oligos to genomic regions flanking each gene.

The resulting collections of modified *E. coli* strains were then challenged by growth in environmental conditions of interest. Warner *et al.*¹ measured the relative fitness of each

Shaking up genome engineering

KA Tipton & John Dueber

A new method generates genome-scale modified bacteria with unprecedented ease.

Systematic approaches to mutate and characterize the function of every gene in a microbe have been hampered by the need to manually create thousands of separate strains through tedious genetic manipulation. In this issue, Warner *et al.*¹ describe an approach to create and characterize rationally modified versions of almost every gene in *Escherichia coli*. Using this strategy, the authors quickly zero in on genes that influence industrially relevant traits, such as tolerance to toxins in a biofuel feedstock. The method enables single genome modifications to be probed rapidly and comprehensively and correlated to a phenotype, yielding information that lays a foundation for gene mapping and for engineering strains with desired phenotypes.

Until now, systematic phenotyping of mutants in yeasts^{2,3} and *E. coli*⁴ has been accomplished by Herculean manual efforts to create thousands of mutant strains, each with a different single-gene knockout. Although the resulting strain collections have proven valuable, it remains a challenge to create, on a genome scale, new collections of mutants for targeted applications or to control gene expression levels using a strong promoter, an inducible promoter or a low-efficiency ribosome binding site.

In contrast, the method of Warner *et al.*¹—trackable multiplex recombineering (TRMR), pronounced ‘tremor’ (Fig. 1)—offers a fast and cheap approach for creating collections of mutants. Impressively, the authors were able to

construct libraries containing up- and down-regulated versions of 96% of the genes in the *E. coli* genome in one week at a materials cost of ~\$1 per targeted gene.

The first step in TRMR is to obtain thousands of 189-base-pair oligonucleotides that target and uniquely identify every *E. coli* gene. Each of these oligos consists of a barcode tag unique to a gene and regions of homology that

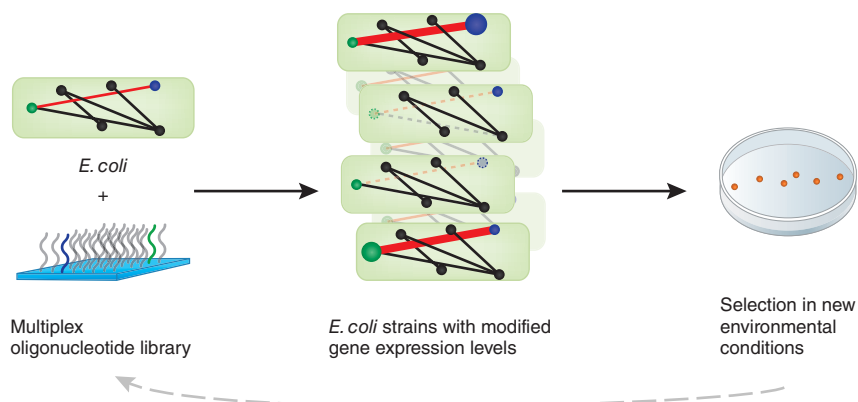


Figure 1 TRMR enables genome-scale selection of rational modifications to the expression of single genes. A multiplex library of oligonucleotides is synthesized to encode a unique barcode tag and regions of homology flanking individual target genes in the *E. coli* genome (left). A series of cloning steps generates linear DNA fragments that contain sequences necessary for up- or downregulating the expression of each target gene. *E. coli* are transformed with this library of linear fragments to create a collection of genetically modified strains (middle, green cells containing a modified genetic network). The modifications alter the functional linkages between genes. (Lines in the networks represent linkages, with thickness being the strength of the link. Circles represent genes, with translucency and a dashed outline representing attenuated expression). The *E. coli* strain collection is grown on medium containing an environmental challenge of interest (right). The identities and relative abundances of individual survivors are determined by sequencing colonies using universal primer sequences. Alternatively, survivors are determined in bulk by microarray analysis of the barcode tags. Importantly, the basic TRMR strategy is amenable to rapid iteration such that the most promising gene modifications are used to seed subsequent cycles of mutation and selection (dotted arrow).

KA Tipton and John Dueber are at the University of California Berkeley, Berkeley, California, USA.
e-mail: jdueber@berkeley.edu