A molecular multi-gene classifier for disease diagnostics

Randolph Lopez Barrezueta[1,2], Ruofan Wang[3,4], Georg Seelig[2,5,6*]

[1] Department of Bioengineering, University of Washington
[2] Molecular Engineering & Sciences Institute, University of Washington
[3] Department of Biology, University of Washington
[4] Department of Microbiology, University of Washington
[5] Department of Electrical Engineering, University of Washington
[6] Paul G. Allen School for Computer Science & Engineering, University of Washington

*correspondence: gseelig@uw.edu

Abstract

Despite its early promise as a diagnostic and prognostic tool, gene expression profiling remains cost-prohibitive and challenging to implement in a clinical setting. Here, we introduce a molecular computation strategy for analyzing the information contained in complex gene expression signatures without the need for costly instrumentation. Our workflow begins by training a computational classifier on labeled gene expression data. This *in silico* classifier is then realized at the molecular level to enable expression analysis and classification of previously uncharacterized samples. Classification occurs through a series of molecular interactions between RNA inputs and engineered DNA probes designed to differentially weigh each input according to its importance. We validate our technology with two applications: a classifier for early cancer diagnostics and a classifier for differentiating viral and bacterial respiratory infections based on host gene expression. Together, our results demonstrate a general and modular framework for low-cost gene expression analysis.

Introduction

Gene expression changes are associated with every human disease. Monitoring such changes enables clinicians to perform diagnosis, evaluate therapeutic efficacy and predict disease recurrence[1-6]. Existing methods for high-throughput RNA detection such as RT-qPCR, microarrays or RNA sequencing can in principle be used to quantitatively monitor gene expression changes in diagnostic applications but remain cost-prohibitive in situations where recurrent monitoring or regular screenings are necessary[3, 7-9]. Moreover, the experimental complexity and the need for *in silico* computational analysis of the resulting data mean that such tests can only be performed in specialized laboratory settings. To overcome these limitations of complexity and cost it is necessary to develop instrument-free diagnostic tests that can be administered and interpreted directly at the point of care[10].

In the past two decades, researchers have found that peripheral gene expression (e.g. whole blood, platelets, exosomes, plasma or saliva) is consistently altered between cancer patients and healthy controls[5, 11-15]. For instance, relative quantitation of telomerase reverse transcriptase (hTERT) RNA in blood or serum has diagnostic and prognostic value in many different cancer types[13, 16-20]. Similarly, researchers have demonstrated that a classifier based on a patient's blood RNA profile can distinguish between bacterial and viral infections[10, 21]. Discriminating between these two groups is essential to address inappropriate prescription of antibiotics and combat antibiotic resistance. Importantly, early cancer diagnostics and combating antimicrobial resistance are just two examples of medical applications that would benefit from rapid and inexpensive gene expression diagnostics for use at home or the point of care.

Recent work in cell-free synthetic biology and DNA nanotechnology has demonstrated progress towards the goal of creating low-cost RNA diagnostics[22-26]. For example, Collins and collaborators developed a test for Zika virus by embedding a set of engineered molecular components for RNA sensing and signal amplification in a paper matrix[24]. Detection of the RNA marker is converted into a colorimetric signal that allows intuitive interpretation. However, to broaden the utility of such tests beyond applications where detection of a single marker is sufficient, it will be necessary to develop "molecular computation" technologies that can convert information encoded in multi-gene expression signatures into interpretable Yes/No answers.

Cell-free molecular circuits with dozens of interconnected components have been experimentally demonstrated and provide proof-of-principle that complex computation can be embedded in molecular substrates [27-33]. But rationally designed molecular circuits realized so far are not well-matched to diagnostic applications. For instance, it is often assumed that inputs take Boolean values (i.e. high or low)[27-30, 34], an assumption that is not naturally compatible with RNA inputs derived from a biological sample. In contrast, computational gene expression classifiers are commonly built using logistic regression, SVMs or neural network approaches that take better advantage of the information encoded in the actual levels of the biomolecules of interest[35-37]. Finally, inputs are typically short, unstructured oligonucleotides with carefully designed sequences rather than long biological RNAs with extensive secondary structure. To realize the potential of DNA computation for diagnostic applications it is thus necessary to develop molecular classifiers that operate directly on RNA inputs and produce a result rapidly and robustly[33].

Here, we address this challenge and demonstrate a framework for creating a DNA-based molecular "computer" capable of performing multi-gene classification (Fig. 1a). In our workflow, publicly available, labeled (e.g. bacterial infection vs. viral infection) gene expression data is first used to train an *in silico* linear classifier, specifically a support vector machine (SVM). During training, constraints are imposed to find the minimal set of genes that need to be considered for classification with a desired accuracy. The resulting model consist of a set of input features (i.e. the RNA transcripts), a positive or negative weight associated with each feature, and a set of mathematical operations (*i.e.* summation and comparison to a threshold) performed over these inputs. Once an optimal model has been obtained, a computational tool translates all parameters and mathematical functions into a novel class of DNA probes that realize the classifier at the molecular level. Below, we first test each molecular classifier component individually, starting with RNA detection and assignment of weights. Finally, we validate the entire workflow by implementing molecular classifiers for the two applications introduced above, namely early cancer diagnostics based on ratiometric detection of hTERT and distinguishing between bacterial and viral infections based on a panel of host genes.

Results

Detection of transcripts through assisted hybridization

The first step in our implementation of a molecular classifier is the detection of RNA transcripts (Fig. 1b). Initially, we pursued an approach using competitive hybridization (or "strand displacement") probes at room temperature (Supplementary Text 2, Supplementary Fig. 1). However, we found that the high degree of secondary structure in RNA transcripts severely limited probe binding efficiency. The use of computational tools for identifying unstructured stretches of RNA ameliorated the situation somewhat, but binding kinetics still varied widely (Supplementary Fig. 2). Moreover, the number of potential probe binding sites on a transcript was determined entirely by the secondary structure and could not be tuned at will which is incompatible with our molecular computation scheme, as detailed below.

To enable robust detection of a larger number of target regions within a transcript, we developed an assisted hybridization protocol. Specifically, we designed a two-stage reaction whereby an input sequence within the target RNA transcript (domain *a*) is thermally or chemically annealed to a hybridization probe consisting of two partially complementary strands (Fig. 1c). Additional helper strands (60 nt.) are included in the reaction; helper strands hybridize adjacent to the targeted region on the RNA to further help unfold its secondary structure and to prevent binding between the adjacent RNA regions and the single stranded domain of the hybridization probe. As a result of this initial reaction the longer probe strand becomes attached to the transcript and a short toehold (domain t1*) is exposed within that strand. Domain a* in the hybridization probe is partially double stranded (15 nt. single stranded and 15 nt. double stranded) and is complementary to the target sequence. Upon binding to its target, hybridization results in a maximum overall gain of 9 base pairs making this reaction thermodynamically favorable. Subsequently, a fluorescent reporter is added to the solution and reacts with the bound strand through toehold-mediated strand displacement, resulting in an increase in fluorescence. If the target RNA is not present, the translator probe reforms upon annealing and cannot interact with the fluorescent reporter.

Importantly, because of the two-stage design, the target sequence on the transcript is completely independent of the reporter sequence.

To experimentally test this strategy, we designed hybridization probes to target three different regions of an mRNA coding for the fusion protein histone 2B Citrine (Citrine) as well as a control hybridization probe specific to GAPDH. For an initial test of the probe design with an unstructured target, a short oligonucleotide encoding the target sequence (30nM) was added to each probe at room temperature. As designed, addition of the target oligonucleotide resulted in increased signal from a downstream fluorescent reporter (Fig. 1d). In contrast, addition of *in vitro* transcribed Citrine RNA (30 nM) did not result in increased fluorescence, because the secondary structure of the RNA transcript hindered the strand displacement reaction. We then tested whether addition of the helper strands could aid hybridization between the RNA target and probe at room temperature, but we observed significant triggering for only one hybridization probe (Supplementary Fig. 3).

Subsequently, we implemented a thermal annealing strategy where the hybridization probe and corresponding helper strands were annealed with the Citrine RNA transcript before addition of the fluorescent reporter. Thermal annealing was performed by heating reactants to 70°C for 10 seconds and subsequently cooling down to 25°C at a rate of -1°C per 10 seconds. As expected, we observed a fluorescent response equivalent to the concentration of added transcript in all Citrine probes while the GAPDH probe showed no increased in fluorescence (Fig. 1e). We carried out the same reaction without addition of helper strands and we observed a lower fluorescence response across all conditions. These results suggest that the helper strands have a role in suppressing non-specific binding between single-stranded overhangs in the probe and single-stranded domains in the RNA target. We also observed very little increase in fluorescence in the case where no transcript was added. Moreover, we performed thermal annealing experiments in a background of cellular mRNA extracted from HEK-293 cells without observing any unspecific triggering. (Supplementary Fig. 4).

Since thermal annealing is not ideal for point-of-care diagnostic applications, we also implemented a chemical denaturing strategy for unfolding RNA targets. Following work by Shelton *et. al.*, we evaluated the use of Urea and subsequent addition of $MgCl^{2+}$ as a method to denature and renature nucleic acid base pairing[38]. We implemented this chemical annealing strategy by incubating a hybridization probe, helper strands and corresponding target in 6.4M urea for 15 minutes followed by incubation with Mg^{2+} for 15 minutes. We observed target-specific increase in fluorescence equivalent to thermal annealing conditions when adding the Citrine RNA transcript or a target oligonucleotide (Fig. 1f).

We note that this assisted hybridization strategy is quite distinct from earlier work in dynamic DNA nanotechnology that generally aimed to create fully autonomous systems that require minimal intervention from an experimentalist. However, we found that separating the detection reaction into an annealing step followed by a more conventional strand displacement-based reporter reaction improved not only the robustness of input detection but also dramatically accelerated it. Both features are crucial for designing a practical diagnostic test.

Molecular implementation of weights

In a gene expression classifier, RNA transcripts have varying levels of influence on the classifier outcome. *In silico*, every transcript is assigned a numerical weight capturing its importance (Fig. 2a). At the molecular level, we implemented these weights by designing multiple hybridization probes that target different regions within each RNA. For example, weights n=1, 2, N are realized by having 1, 2 or N distinct probes targeting the same transcript (Fig. 2b). Even though the targeted sequences on the transcript are different, each probe contains an identical output strand (domains t1*x* in Fig. 1c) which then triggers a fluorescent reporter. Every additional hybridization probe results in a proportional increase in the steady state fluorescence signal. The fluorescence due to $mRNA_1$ should thus be proportional to the product w_1*[$mRNA_1$] where w_1 is an integer weight and [$mRNA_1$] is the concentration of $mRNA_1$.

We implemented this set-up experimentally by designing reactions with 1, 2, 3 or 4 probes targeting contiguous regions on the Citrine transcript. To avoid saturation of the reporter complex, we operated the system in a regime where reporter and hybridization probes far exceeded the transcript concentration. We measured the fluorescence signal corresponding to the reporter complex before and after addition of the hybridized probe-RNA complexes until a steady state was reached (Fig. 2c). As expected, we found that the steady state signal was linearly proportional to the number of hybridization probes bound to the RNA transcript for all RNA concentrations tested, demonstrating that this mechanism can be used to assign an integer-valued weight to an RNA transcript (Fig. 2d).

Summation and thresholding

Building a complete linear classifier requires mechanism for summing up weights and comparing the sum to a threshold value to obtain the desired yes/no answer (Fig. 3a)[33, 39]. If there are multiple transcripts with different weights of the same sign, we can compute the sum of their contributions simply by using the same output sequence across all probes. The total concentration of output strands and thus the final fluorescence signal is then proportional to the sum w_1*[$mRNA_1$]+…+ w_N*[$mRNA_N$]. Weights with negative values can be implemented using a distinct output sequence for the negative probes. The sums of negative and positive weights in a classifier are then represented by the total concentrations of two distinct output strands.

To complete the summation, the individual sums of positive and negative weights – represented by (positive) concentrations of two distinct nucleic acids sequences – need to be subtracted from one another. Intuitively, such a subtraction can be realized as a chemical reaction whereby stoichiometric amounts of positive and negative output strands annihilate each other until only the majority species is left. The concentration of that species then is the final result of the summation over all weights. To implement such a stoichiometric annihilation reaction between two nucleic acid species of unrelated sequence, we take advantage of the cooperative hybridization mechanism ("annihilator" gate) introduced by Zhang[39, 40].

The final step in the molecular computation pipeline is to compare the result of the summation to a threshold value. In the simplest case, the threshold value is set to zero and the class a specific input sample belongs to is determined simply by the sign of the final sum. Non-zero threshold values can be realized by spiking the corresponding amount of negative or positive output strand into the reaction which biases the sum by a controlled amount.

Molecular thresholding of RNA transcripts

To experimentally test whether an "off-the-shelf" thresholding (or subtraction) element could be used in conjunction with our RNA detection scheme we created a DNA circuit consisting of three modules: a translator gate that connects the output strand from the assisted hybridization reaction to the threshold element, an "annihilator gate" and single-stranded reference oligonucleotide that together act as the threshold element and a catalytic reporter that amplifies any signal exceeding the threshold value to a constant level allowing for a Yes/No answer (Supplementary Fig. 5).

We tested this molecular thresholding system on three different transcripts (hTERT, EGFR, GAPDH) commonly used as biomarkers or reference genes for diagnostic purposes. To accommodate different RNAs only the hybridization probe and helper strands needed to be switched while all the other strand displacement components are retained, demonstrating modularity of the design. Each mRNA was individually transcribed *in vitro* from a cDNA template and quantified. For each transcript, we evaluated four experimental conditions using thermal annealing with varying ratios of transcript to reference oligonucleotide. Steady state fluorescence values were acquired two hours after addition of a catalytic amplifier and fluorescent reporter. With all three transcripts, we only observed an increase in fluorescence when the amount of transcript exceeded the amount of threshold.

A two-gene diagnostic classifier

For an experimental test of a full two-input classifier circuit, we selected hTERT, a cancer biomarker, as the target (associated with a positive weight) and GAPDH, a common internal reference gene in RT-PCR experiments as the reference RNA (associated with a negative weight) (Fig. 3b). Relative quantitation of HTERT to GAPDH in human plasma has been suggested as an early diagnostic and prognostic biomarker in human cancer[13, 16-20, 41, 42]. The thresholding (subtraction) and amplification reaction are performed exactly as above but instead of an external reference strand to set the threshold value, there now is an internal reference RNA associated with a negative weight that effectively sets a threshold (Supplementary Fig. 6).

We evaluated four classifiers with an hTERT weight of +1 and GAPDH weights of -1, -2, -3 and -4 (Fig. 3c). A sample containing both RNA transcripts was first combined with corresponding hybridization probes and helper strands. hTERT transcript was present at 15 nM while GAPDH transcript was titrated from 0 nM to 14 nM with all DNA circuits components added at higher, non-limiting concentrations. We further characterized a classifier response with an hTERT weight of +1 and GAPDH weight of -2 with a range of concentrations of each transcript (0nM to 20nM) (Fig. 3d,e). Overall, we evaluated 64 different experimental conditions where we recorded fluorescence levels for 2 hours after addition of strand displacement components. We only observed a significant increase in fluorescence in conditions when the amount of hTERT transcript was above the threshold set by the product of the GAPDH transcript concentration and weight, in agreement with the classifier design.

Training a multi-gene support vector machine

We next sought to scale up our molecular classifier framework. Discriminating between viral and bacterial infections using molecular gene expression classification is a promising application since it requires a rapid, cost-effective and self-contained process to be implemented in a clinical setting. In 2016, Tsalik *et. al.* developed a peripheral whole blood gene expression classifier with 130 genes to differentiate between bacterial infections, viral infections, non-infectious illness and healthy controls with 87% accuracy[10].

To build a molecular classifier, we first simplified the classification problem by distinguishing only between viral and bacterial infections. We used the publically available gene expression data corresponding to 115 viral infections and 70 bacterial infections for classifier training[10]. For each patient, gene expression values for 14,500 human genes were measured. We implemented a support vector machine (SVM) to determine the minimal set of genes and corresponding weights for this classification problem. This process involved iterating through multiple sets of features (genes) and associated weights until converging to a solution that resulted in the best classification outcome.

We trained an SVM algorithm with the following constraints: First we required a low number of genes (<10) to allow for the classifier to be implemented at the molecular level. Second, we constrained weights to integer values between -5 to +5. This choice was made to limit the number of probes for a single gene as well as the overall size of the classifier. Third, we made the misclassification penalty for bacterial samples 3 times higher than that for viral samples. This choice was made because the worst possible outcome is to incorrectly diagnose a bacterial infection as viral, delaying the use of antibiotics. Even though this classification model performed well in the validation set, it is important to note that a model with a higher number of features may be more robust when encountering gene expression variability absent in the training dataset. We selected 9 classifiers with at least 80% accuracy in the training set and validated them using a different gene expression data set[21]. We selected the classification model with the highest performance in the validation set to build a molecular classifier (Fig. 4a). The selected classifier correctly labelled 94% and 80% of bacterial and viral samples in the training set and 89% of bacterial and 90% of viral samples in the validation set (Fig. 4b,c).

A molecular implementation of the bacterial vs viral classifier

Next, we designed a molecular implementation of the bacterial vs. viral classifier. First, we selected regions in each transcript that consisted of individual exons that were at least 200 base-pairs long such that they could fit multiple hybridization probes. Due to the large number of transcripts and associated probes, we implemented a probe design tool for systematically generating the necessary DNA components for molecular classification. Each transcript was assigned a number of hybridization probes and helper strands, based on the weights learned *in silico*. Positive and negative transcripts were assigned hybridization probes with different output domains such that the concentrations of the positive and negative output strands represent the weighted sums of the respective RNA inputs, as described above. The complete DNA classifier consists of 20 hybridization probes and 14 helper strands (two for each transcript). A strand displacement cascade using two translator gates and two fluorescent reporters aggregate the signal generated by the hybridization module. Overall, the circuit consists of 62 different oligonucleotides.

Rather than performing the subtraction at the molecular level as we have done in the previous example, we chose to use two distinct fluorophores to read out the positive and negative output strands individually, which allowed us to more quantitatively characterize performance of individual classifier components. A fluorescent reporter containing a 6-FAM (Fluorescein) (FAM) and a quencher was associated with positive/bacterial transcripts while a fluorescent reporter containing a 6-Carboxyl-X-Rhodamine (ROX) and a quencher was associate with negative/viral transcripts (Fig. 5a). Upon reporter calibration, the fluorescence signal from the ROX reporter can be subtracted from the FAM reporter signal to obtain a normalized signal used for classification ([FAM]-[ROX] nM). Samples resulting in a normalized signal of [FAM]-[ROX] > 0 belong to the bacterial infection category while samples for which this signal is less than zero belong to the viral infection category.

After assembling the molecular classifier, we first used synthetic DNA oligonucleotide targets to individually test all 20 hybridization probes. Upon thermal annealing and subsequent strand displacement, we confirmed that each oligonucleotide target triggered the intended fluorescent channel with the expected signal intensity (corresponding to a unit weight) while the signal remained near background in the other channel (Fig. 5b). Subsequently, we tested the molecular classifier using in-vitro transcribed RNA species. After addition of each RNA transcript to the molecular classifier, we again measured the fluorescence response across both channels. For each transcript, we only observed significant increase in fluorescence in the expected channel. After calibration and subtraction of both channel fluorescence signals, we obtained a normalized signal for each transcript addition ([FAM] – [ROX] nM). We found this normalized signal to be proportional to the weight assigned to each gene suggesting that the molecular weight implementation was performed correctly (Fig. 5c).

Lastly, we tested our molecular classifier with samples containing RNA molecules matching the expression profiles from the training set microarray data. We selected 12 samples corresponding to six patients with viral and six patients with bacterial infections (Fig. 5d). We replicated the original gene expression profile by adding each cDNA amplicon based on its expected concentration as calculated from the microarray data. Each amplicon contained a T7 promoter for RNA transcription. Samples were then diluted to approximately 10 picomolar followed by in-vitro transcription which resulted in 1000x amplification (Fig. 5e). As expected, upon addition of each sample to the molecular classifier, we observed significant triggering in both fluorescence channels. All samples were classified correctly based on the normalized signal intensity**.** Furthermore, we found a strong correlation between the normalized signal intensity and the corresponding computational output for each sample as estimated using the corresponding SVM model (Fig 5f).

Discussion

We introduced a systematic framework for translating an *in silico* gene expression classifier into DNA circuitry. We confirmed the robustness of this framework by building two distinct classifiers with varying numbers of weights and inputs. Using our approach, any *in silico* classifier can in principle be converted into a molecular classifier, synthesized for rapid prototyping and experimentally validated.

We developed three novel building blocks to enable molecular computation with RNA transcripts as inputs. First, breaking up transcript detection into two separate steps, assisted hybridization and strand displacement, enabled us to robustly perform molecular computing with any RNA transcript as an input. Second, by varying the number of probes that hybridize to an RNA transcript we were able to differentially weigh the importance of transcripts. Third, by designing probes with shared output sequences we were able to compute the weighted sum of multiple transcript. So far, we have used these building blocks to create classifiers with up to seven distinct RNA inputs and up to five (positive or negative) probes per transcript. However, the size of the classifiers could in principle be scaled to tens or hundreds of targets with the number of weights only limited by the size of the transcripts. In principle, potential cross-talk between probes and incorrect targets becomes more likely when the number of probes is higher. Nevertheless, a thermodynamic simulation of these interaction can inform the selection of probes across the length of a target RNA transcript that exhibit little or no cross-talk.

Compared with existing methods for gene expression analysis, our approach is well-suited for inexpensive and rapid examination of clinical samples (Supplementary Table 5). Because of its experimental simplicity, our workflow is fast: the combined reaction time for the assisted hybridization module and strand displacement reaction was under 20 minutes with no additional time required for computational analysis and data interpretation. More fundamentally, the amount of work required to perform gene expression classification using our framework is independent of the number of genes in the assay. The complexity of RT-qPCR experiments, the current gold-standard for gene expression profiling in the clinic, in contrast scales linearly with the number of genes being analyzed. The DNA-based classification workflow thus dramatically reduces the need for liquid handling making it a good fit for point-of-care applications. RNA sequencing and barcoded RNA hybridization (Nanostring) also allow for multiplexed gene expression analysis in a single reaction but require expensive instrumentation or consumables. In contrast, we perform expression analysis by harnessing DNA computation while relying on inexpensive instrumentation: a thermocycler and a fluorescence reader. Finally, all alternative approaches provide information about the expression of individual genes in a panel, while our approach aggregates this information at the molecular level and provides a single, easy-to-interpret diagnosis, enabling fast turnaround.

It should be noted however that the rate of the strand displacement reaction is highly dependent on the concentration of the RNA inputs, and including a pre-amplification step in the workflow would increase processing time. In this work, we demonstrated amplification of a mixture of cDNA amplicons in the low picomolar range using in vitro transcription before molecular classification. However, RNA transcripts are typically present at attomolar or femtomolar concentrations in tissue and blood RNA samples[5, 19]. Other amplification strategies, such as rolling circle amplification or loop mediated isothermal amplification, will need to be explored for further amplification and may be more suited for point of care applications[26, 43-46]. Moreover, the output of the classification can be measured using a different readout system such as a paper based substrate or a colorimetric reaction to further increase sensitivity or simplify readout of results[23, 24].

Still, by demonstrating a robust and modular approach for instrument-free analysis of complex gene expression signatures, our work closes an important gap in the existing toolbox for

engineering affordable point-of-care diagnostics. The number of clinical studies examining how variations in peripheral gene expression are associated with disease diagnostics, monitoring and prognosis is ever increasing, and the use of molecular computation for gene expression analysis suggests a path towards translating this academic knowledge into future diagnostics.

Methods

DNA oligonucleotides
All DNA oligonucleotides were purchased from Integrated DNA Technology (IDT). Individual DNA oligonucleotides were suspended to 100 μM and stored in water. Fluorophore and quencher-labelled oligonucleotides were ordered HPLC purified, except for FAM-labelled oligonucleotides. Unlabeled oligonucleotides were unpurified.

Hybridization probe preparation
Hybridization probes consisted of annealed complex of two DNA oligonucleotides: a 21-nt bottom strand and a 56-nt top strand. The strands were mixed stoichiometrically with 30% excess of the bottom strand and then thermally annealed: heated to 98C for 10 seconds and cooled uniformly from 98°C to 25°C over the course of 73 minutes.

Hybridization probes for the viral/bacterial classifier
40 oligonucleotides (top and bottom strands) corresponding to 20 hybridization probes were order using IDT 25 nmole DNA Plate Oligo synthesis normalized to 100uM on IDT LabReady buffer. For purification, 20 top strands and 20 bottom strands were pooled together respectively and purified as a mixture using 12% Urea 19:1 acrylamide: bisacrylamide gel (SequaGel UreaGel System. National Diagnostics). Subsequently, gel bands were visualized using ultraviolet light with a fluorescent backplate, and then cut out and eluted into 1 ml 1X TAE, 12.5 mM Mg^{++} for 12 hours. Concentrations were calculated by measuring absorbance at 260 nm (Eppendorf Biophotometer plus) and using IDT-specified extinction coefficient.

Strands displacement probe preparation
Strand displacement probes (translators, reporters, catalytic amplifiers and annihilator gates) consisted of annealed complexes of two or more DNA oligonucleotides. The strands were mixed stoichiometrically with 10% excess of the target binding strand for the translator, catalytic amplifier gate and annihilator gate. Subsequently, DNA complexes were thermally annealed: heated to 98C for 10 seconds and cooled uniformly from 98°C to 25°C over the course of 73 minutes. After annealing, individual probes were purified using a 12% non-denaturing PAGE gel as described above.

Cellular mRNA preparation
Cellular mRNA was extracted from HEK-293 (ATCC 30-2003) human cell line using a magnetic isolation kit for mRNA (NEB Next Poly(A) mRNA Magnetic Isolation kit #E7490). Cellular mRNA was aliquoted and stored in nuclease free water with RNAse inhibitor (NEB) at -80°C until needed.

RNA target preparation

Amplicons corresponding to RNA target sequences were generated by PCR amplification of HEK-293 cDNA or human genomic DNA (ThermoFisher Catalog number 4312660). Amplification of each target was carried out with a corresponding forward primer containing a T7 RNA polymerase promoter sequence (5-TAATACGACTCACTATAGGG-3). After amplification, each product was visualized on a 1.5% agarose gel and the correct band was excised and processed with a gel extraction kit (QIAGEN catalog number 28704). RNA targets were generated using T7 RiboMAX™ Express Large-Scale RNA Production System (Promega). Purification of RNA targets was carried out using a phenol/chloroform extraction protocol. Final RNA concentrations were determined using absorbance at 260 nm and estimated extinction coefficient for the corresponding single stranded RNA. RNA was aliquoted and stored in nuclease free water with RNAse inhibitor (NEB) at -80°C until needed.

Time-course fluorescence measurements
Kinetic fluorescence measurements were performed using a fluorescence plate reader for higher measurement throughput (Biotek Synergy HTX). Thermal annealing and strand displacement reactions were carried out in 1X TAE, 12.5 mM Mg^{++}.

Fluorescence normalization
Arbitrary fluorescence units were converted to concentrations using a calibration curve of each reporter complex. To create a calibration curve, annealed reporter complex stock was suspended in 1X TAE/Mg^{++} and an initial baseline fluorescence signal was recorded. That was followed by stepwise addition of known concentrations of reporter triggering strands. After each trigger strand addition, the steady state was recorded.

Viral/Bacterial SVM training and validation
For training of the support vector machine algorithm, we obtained microarray data (NCBI GSE63990) for 273 ill patients and 44 healthy volunteers[10]. We processed the dataset by first selecting samples labelled only as bacterial or viral infections (70 and 115 samples respectively) and transforming the microarray gene expression ratios by logarithm of base 2 to estimate biological expression levels. We trained an SVM algorithm (classifier with a linear kernel) on this data set to distinguish between viral and bacterial classes using the svm.LinearSVC function from Python library sklearn. We used a squared hinge loss function with L1 norm while iterating through multiple penalty parameters to obtain SVM classifiers with varying number of features. We found 9 models that employed less than 10 genes while maintaining a classification accuracy of 80% or higher in the training set. We evaluated these classifiers using a different microarray dataset (NCBI GSE6269) where they performed similarly well (AUC > 0.90)[21]. Finally, we selected the classifier with the highest AUC value for experimental implementation.

Computational tool for generating hybridization probes from the in silico classifier
First, we generated an input file containing each transcript sequence and their corresponding weights from the *in silico* classifier. A python script sliced the transcript sequence to generate helper strands (first and last 60 nts.), hybridization targets (30 nt. each) and hybridization probes. Hybridization probes were generated with either a positive or negative sequence domain based on the classifier weight. The output of this script contains each component sequence (helper, top strand hybridization probe, bottom strand hybridization probe and target sequence) and name.

Bibliography

1. Vargas, J.D. & Lima, J.A.C. Coronary artery disease: a gene-expression score to predict obstructive CAD. *Nat. Rev. Cardiol*, 243-244 (2013).
2. Veer, V.t.L.J., Dai, H., Vijver, V.M.J. & He, Y.D. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 530-536 (2002).
3. Blank, P.R. et al. Cost-effectiveness analysis of prognostic gene expression signature-based stratification of early breast cancer patients. *Pharmacoeconomics* **33**, 179-190 (2015).
4. Myers, M.B. Targeted therapies with companion diagnostics in the management of breast cancer: current perspectives. *Pharmgenomics Pers Med*, 7-16 (2016).
5. Rotunno, M. et al. A Gene Expression Signature from Peripheral Whole Blood for Stage I Lung Adenocarcinoma. *Cancer Prev Res* **4**, 1599-1608 (2011).
6. Lunnon, K., Sattlecker, M. & Furney, S.J. A blood gene expression marker of early Alzheimer's disease. *J Alzheimers Dis.* **33**, 737-753 (2013).
7. Koscielny, S. Why most gene expression signatures of tumors have not been useful in the clinic. *Sci. Transl. Med.* **2** (2010).
8. Sotiriou, C. & Piccart, M.J. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nat. Rev. Cancer* **7**, 545-553 (2007).
9. Cassarino, D.S., Lewine, N., Cole, D. & Wade, B. Budget impact analysis of a novel gene expression assay for the diagnosis of malignant melanoma. *J Med Econ.* **17**, 782-791 (2014).
10. Tsalik, E.L. et al. Host gene expression classifiers diagnose acute respiratory illness etiology. *Sci. Transl. Med.* **8** (2016).
11. Best, M.G. et al. RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. *Cancer Cell* **28**, 666-676 (2015).
12. Yuan, T., Huang, X., Woodcock, M., Du, M. & Dittmar, R. Plasma extracellular RNA profiles in healthy and cancer patients. *Sci. Rep.* **6** (2016).
13. Dasí, F. et al. Real-time quantification in plasma of human telomerase reverse transcriptase (hTERT) mRNA: a simple blood test to monitor disease in cancer patients. *Lab. Invest.* **81**, 767-769 (2001).
14. Zhang, L. et al. Salivary Transcriptomic Biomarkers for Detection of Resectable Pancreatic Cancer. *Gastroenterology* **138**, 949 (2009).
15. Zhang, L. et al. Development of transcriptomic biomarker signature in human saliva to detect lung cancer. *Cell Mol Life Sci* **69**, 3341-3350 (2012).
16. Kyo, S., Takakura, M., Fujiwara, T. & Inoue, M. Understanding and exploiting hTERT promoter regulation for diagnosis and treatment of human cancers. *Cancer Sci.* **99**, 1528-1538 (2008).
17. Lledo et al. Real time quantification in plasma of human telomerase reverse transcriptase (hTERT) mRNA in patients with colorectal cancer. *Colorectal Dis* **6**, 236-242 (2004).
18. March-Villalba, J.A. et al. Cell-Free Circulating Plasma hTERT mRNA Is a Useful Marker for Prostate Cancer Diagnosis and Is Associated with Poor Prognosis Tumor Characteristics. *PLoS ONE* (2012).

19. Miura, N., Nakamura, H., Sato, R. & Tsukamoto, T. Clinical usefulness of serum telomerase reverse transcriptase (hTERT) mRNA and epidermal growth factor receptor (EGFR) mRNA as a novel tumor marker. *Cancer Sci.* **97**, 1366-1373 (2006).
20. Terrin, L. et al. Relationship between tumor and plasma levels of hTERT mRNA in patients with colorectal cancer: implications for monitoring of neoplastic disease. *Clin. Cancer Res.* **14**, 7444-7451 (2008).
21. Ramilo, O., Allman, W., Chung, W., Mejias, A. & Ardura, M. Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Blood* **109**, 2066-2077 (2007).
22. Chen, S.X. & Seelig, G. An Engineered Kinetic Amplification Mechanism for Single Nucleotide Variant Discrimination by DNA Hybridization Probes. *J. Am. Chem. Soc* **138**, 5076–5086 (2016).
23. Pardee, K., Green, A.A., Ferrante, T. & Cameron, D.E. Paper-based synthetic gene networks. *Cell* **159**, 940-954 (2014).
24. Pardee, K. et al. Rapid, low-cost detection of Zika virus using programmable biomolecular components. *Cell* **165**, 1255-1266 (2016).
25. Jung, C. & Ellington, A.D. Diagnostic applications of nucleic acid circuits. *Acc. Chem. Res* **47**, 1825-1835 (2014).
26. Gootenberg, J.S. et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356**, 438-442 (2017).
27. Qian, L. & Winfree, E. Scaling Up Digital Circuit Computation with DNA Strand Displacement Cascades. *Science* **332**, 1196-1201 (2011).
28. Qian, L., Winfree, E. & Bruck, J. Neural network computation with DNA strand displacement cascades. *Nature* **475**, 368-372 (2011).
29. Seelig, G., Soloveichik, D., Zhang, D. & Winfree, E. Enzyme-Free Nucleic Acid Logic Circuits. *Science* **314**, 1585-1588 (2006).
30. Chen, Y.-J. et al. Programmable chemical controllers made from DNA. *Nat. Nanotechnol.* **8**, 755-762 (2013).
31. Genot, A.J., Fujii, T. & Rondelez, Y. Scaling down DNA circuits with competitive neural networks. *J. R. Soc. Interface* **10**, 20130212 (2013).
32. Franco, E. et al. Timing molecular motion and production with a synthetic transcriptional clock. *Proc Natl Acad Sci U S A* **108** (2011).
33. Mills, A.P. Gene expression profiling diagnosis through DNA molecular computation. *Trends Biotechnol* **20**, 137-140 (2002).
34. Green, A.A. et al. Complex cellular logic computation using ribocomputing devices. *Nature* **548**, 117-121 (2017).
35. Brown, M.P.S., Grundy, W.N. & Lin, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* **97**, 262-267 (2000).
36. Abusamra, H. A comparative study of feature selection and classification methods for gene expression data of glioma. *Procedia Comput Sci* **23**, 5-14 (2013).
37. Liu, H., Li, J. & Wong, L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform.* **13**, 51-60 (2002).
38. Shelton, V.M., Sosnick, T.R. & Pan, T. Applicability of Urea in the Thermodynamic Analysis of Secondary and Tertiary RNA Folding. *Biochemistry* **38**, 16831-16839 (1999).

39. Zhang, D. & Seelig, G. DNA-Based Fixed Gain Amplifiers and Linear Classifier Circuits. *LNCS* **16**, 176-186 (2010).
40. Zhang, D. Cooperative Hybridization of Oligonucleotides. *J. Am. Chem. Soc* **133**, 1077-1086 (2011).
41. Dasí, F. et al. Real-time quantification of human telomerase reverse transcriptase mRNA in the plasma of patients with prostate cancer. *Ann. N. Y. Acad. Sci.* **1075**, 204-210 (2006).
42. Yang, Y.J., Chen, H., Huang, P., Li, C.H. & Dong, Z.H. Quantification of plasma hTERT DNA in hepatocellular carcinoma patients by quantitative fluorescent polymerase chain reaction. *Clin Invest Med* **34** (2011).
43. Lizardi, P.M., Huang, X., Zhu, Z. & Bray-Ward, P. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nature Genet* **19**, 225-232 (1998).
44. Zhao, W., Ali, M.M., Brook, M.A. & Li, Y. Rolling circle amplification: applications in nanotechnology and biodetection with functional nucleic acids. *Angew Chem Int Ed Engl.* **47**, 6330-6337 (2008).
45. Notomi, T., Okayama, H. & Masubuchi, H. Loop-mediated isothermal amplification of DNA. *Nucleic acids Res* **28**, e63 (2000).
46. Tomita, N., Mori, Y., Kanda, H. & Notomi, T. Loop-mediated isothermal amplification (LAMP) of gene sequences and simple visual detection of products. *Nat. Protoc.* **3**, 877-882 (2008).

Acknowledgements. We thank Yuan-Jyue Chen, Sifang Chen, Gourab Chatterjee and David Yu Zhang for their support and helpful discussion. This work was supported by NSF grants CCF-171449 and CCF-1317653.

Author contributions. R.L. and G.S. designed experiments and wrote the paper. R.L. and R.W. performed the experiments.

Supplementary Materials. Supplementary information includes supplementary text, figures and tables.

Code availability. The computer code corresponding to the computational sections of this work are available in the following GitHub repository: https://github.com/rmlb/classifier_probegen/ or from the corresponding author upon request.

Data availability. The characterization data and experimental protocols for this work are available within this manuscript and its associated Supplementary Information, or from the corresponding author upon request.

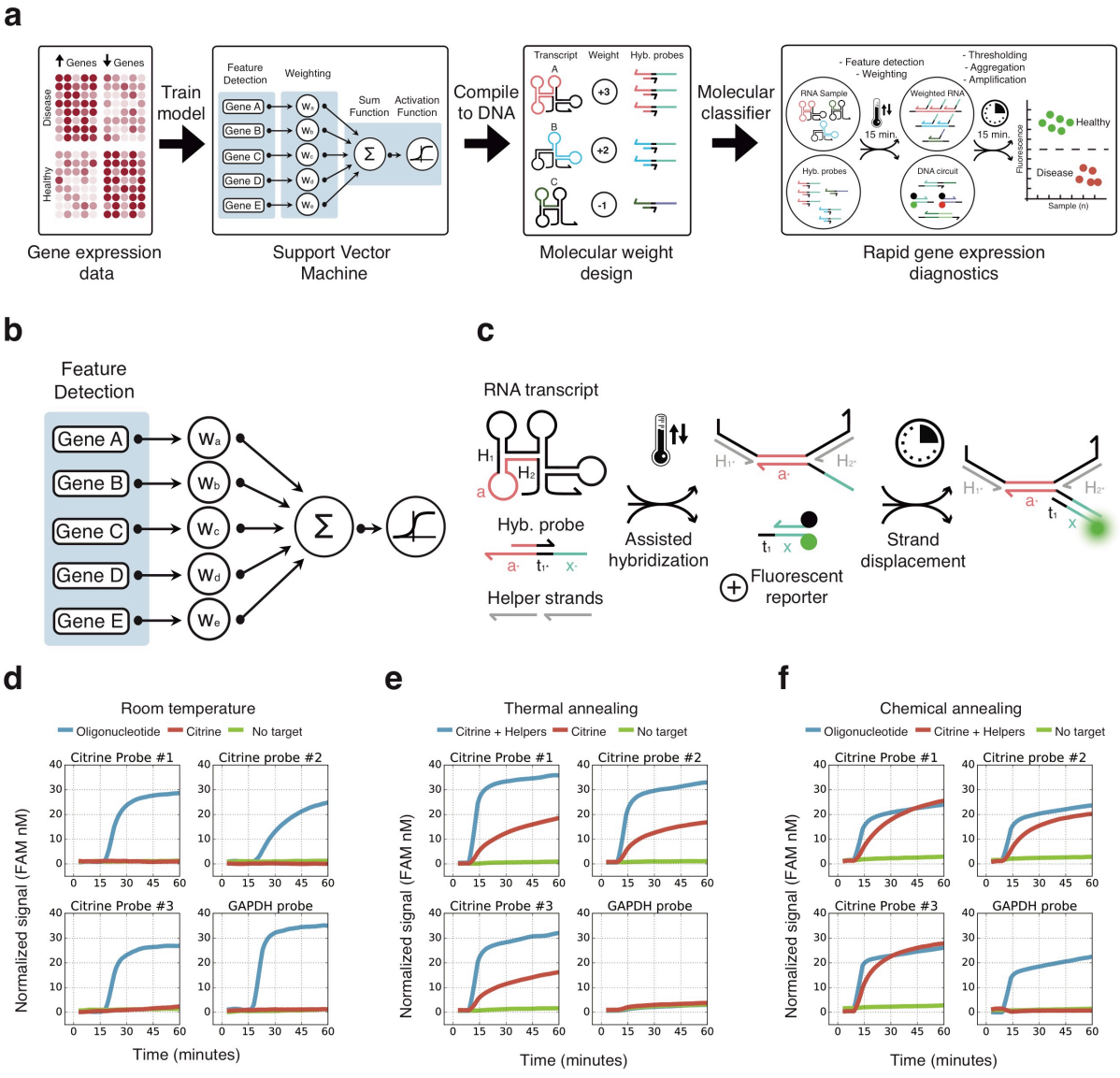

Figure 1 | A universal framework for rapid prototyping of molecular classifiers for gene expression diagnostics. a, An in-silico classifier is trained and validated on publicly available gene expression data. The weights and other characteristics of the in *silico* classifier are then translated into DNA complexes that realize the classifier at the molecular level. Finally, the molecular classifier is tested with RNA targets and a diagnosis is obtained. **b,** As a first step towards creating a molecular gene expression classifier, we developed a systematic approach for detecting specific RNA transcripts with DNA strand displacement cascades **c,** The molecular mechanism for coupling DNA-based circuits with endogenous RNA transcripts consists of two reaction steps. First, a hybridization probe and helper strands are hybridized to the target site using chemical or thermal annealing. Subsequently, a fluorescent reporter is added to the reaction and binds to the product of the assisted hybridization reaction via strand displacement. **d,** We tested the RNA detection reaction by designing 3 hybridization probes targeting different regions in Citrine and a probe targeting a region in GAPDH. At room temperature, the addition of Citrine transcript (30 nM) resulted in no significant triggering in all probes. As a positive control, we added a target oligonucleotide (30 nM) for each probe that resulted in the expected fluorescence response. **e,** Experimental results corresponding to the thermal annealing protocol where each probe was annealed with Citrine RNA and corresponding helper strands before addition of the fluorescent reporter. All Citrine probes were triggered by the Citrine RNA while the GAPDH probe resulted in no fluorescence response. Without inclusion of the helper strands, Citrine probes resulted in a diminished fluorescence response. **f,** Experimental results corresponding to the chemical annealing protocol where each probe was incubated with Citrine RNA and corresponding helper strands in Urea and subsequently in $MgCl_2$. We observed the expected fluorescence response with addition of an oligonucleotide target or Citrine transcript.

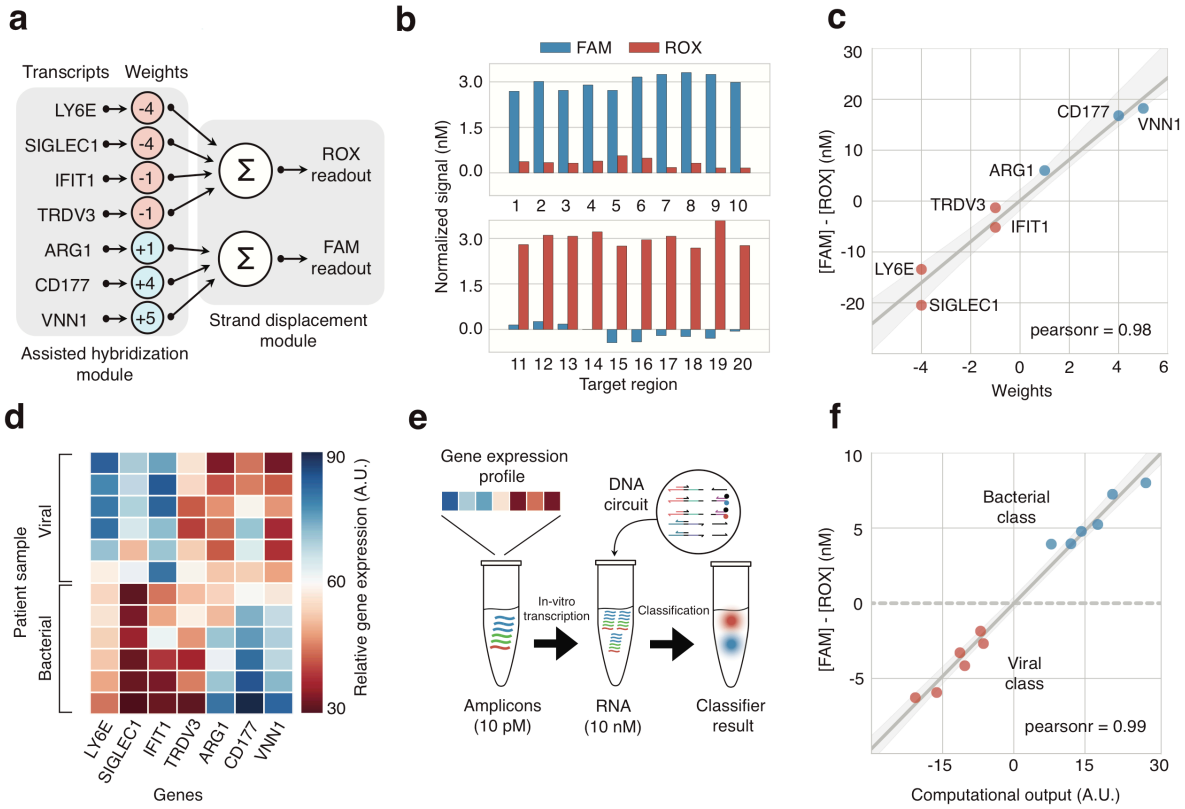

Figure 2 | Implementation of classifier weights by targeting of multiple adjacent regions in a transcript. a, Each transcript is assigned a weight reflecting its influence in the classifier decision. **b,** Each transcript is targeted with a number of probes equivalent to its classifier weight. By targeting probes to neighboring regions, only a single pair of flanking helper strands is necessary for each transcript hybridization event. **c,** Probe binding was characterized through fluorescence kinetics experiments. Initial fluorescence values correspond to quenched reporter in solution. After 10 minutes, annealed probe-transcript complexes are added to the solution resulting in an increase in fluorescence proportional to the number of hybridization probes (1, 2, 3 or 4). Reactions were carried out with 50 nM of reporter, 40 nM of combined hybridization probe and different concentrations of Citrine transcript **d,** Steady state fluorescence response corresponding to 1, 2, 3 or 4 hybridization probes targeting the H2B-Citrine RNA transcript. As expected, we observed a linear relationship between the number of hybridization probes and the fluorescence response across a range of Citrine RNA concentrations.

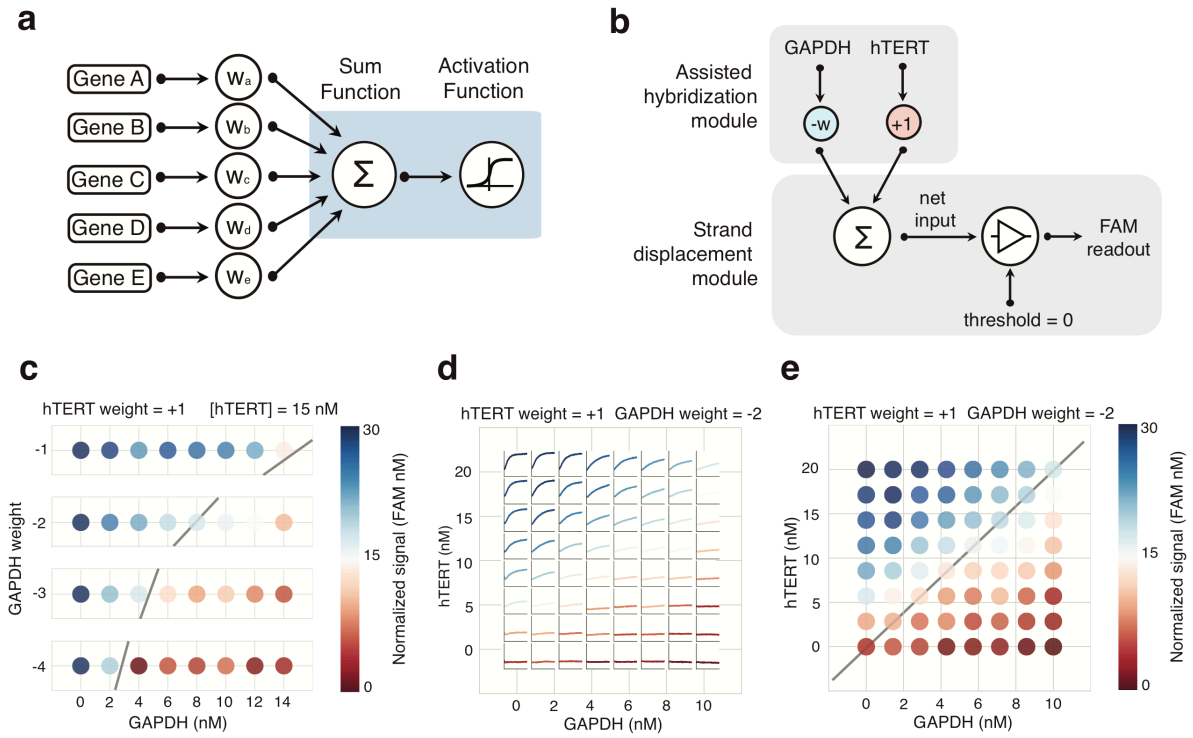

Figure 3 | Molecular implementation of a two-gene classifier for cancer diagnostics a, A sum and activation function are used to aggregate weighted gene expression information into a single, interpretable output. Upon transcript detection and scaling, a sum function calculates the resulting net input. If the net input is higher than a threshold, an activation function produces a catalytic response. **b,** Graphical representation of the hTERT/GAPDH molecular classifier with variable negative weights for GAPDH and a weight of +1 for hTERT. **c,** Final state fluorescence measurements after 2 hours corresponding to four classifiers with varying GAPDH weights. Grey line indicates ideal thresholding boundary. Reactions were carried out with 50 nM of reporter, 100 nM of helper strands and 30 nM of catalytic amplifier, annihilator, translators and hybridization probes. **d,** 2-hour fluorescence measurements after addition of strand displacement components corresponding to a +1 hTERT / -2 GAPDH molecular classifier. **e,** End point fluorescence measurements after 2 hours corresponding to a +1 hTERT / -2 GAPDH molecular classifier.

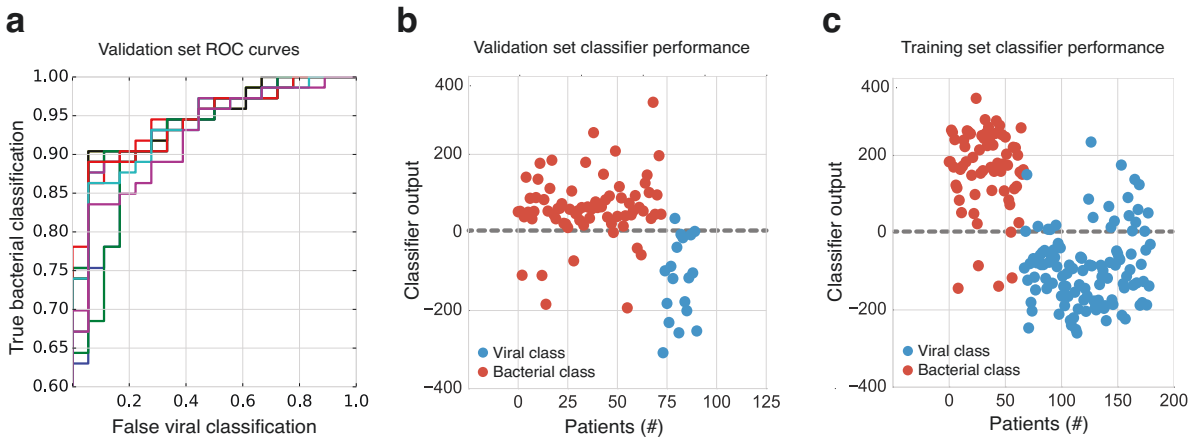

Figure 4 | *In silico* training of a minimal linear classifier to discriminate viral from bacterial infections based on host gene expression data. a, ROC curves illustrate the diagnostic ability of a binary classifier system as the threshold is varied. ROC curves correspond to the classification performance in the validation set from 10 classification models selected from the training phase. We used the classification model with the highest AUC in the validation dataset to build a molecular classifier. **b,** Performance of the selected classifier in the validation set where 89% and 90% of bacterial and viral samples were labeled correctly. **c,** Performance of the selected classifier in the training set where 94% and 80% of bacterial and viral samples were labeled correctly.

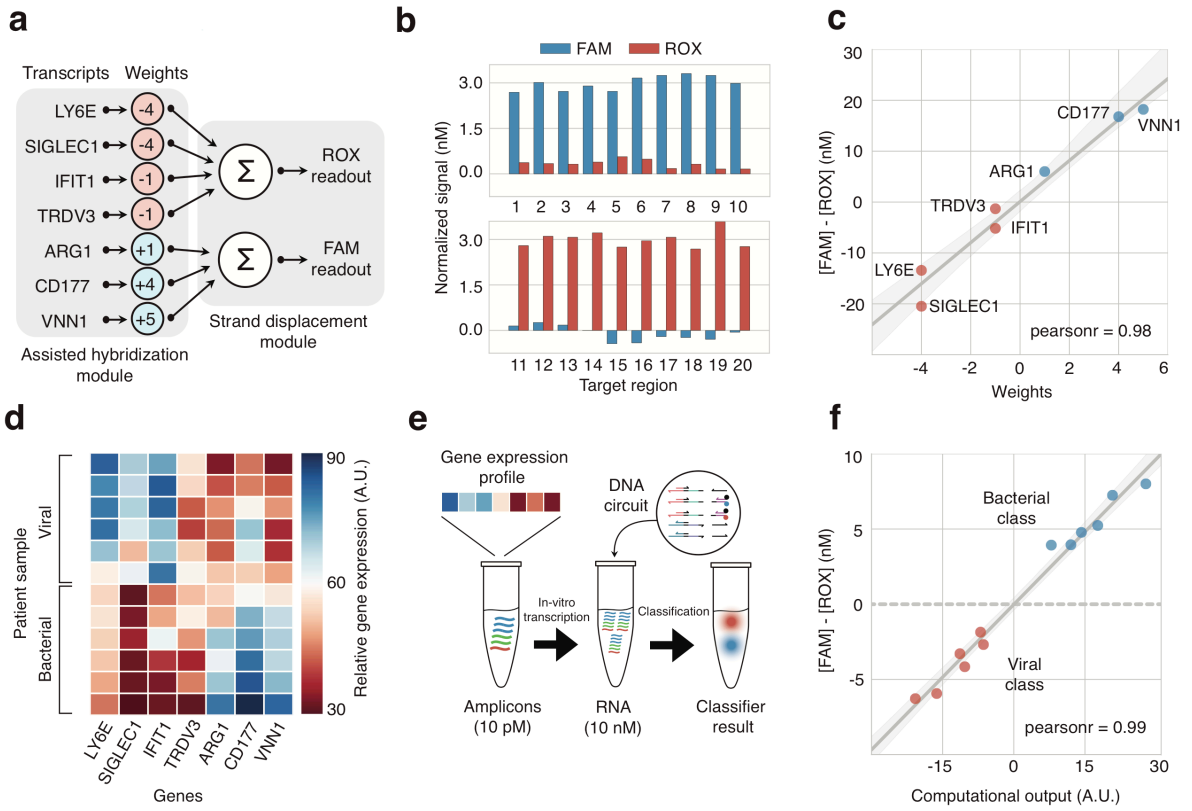

Figure 5 | A molecular classifier of host gene expression for respiratory infections diagnostics. a, Graphical representation of the viral vs. bacterial infection classifier. The classifier uses 7 genes. 20 hybridization probes assign weights ranging from -4 to +5 to each transcript. The weighted sums of all transcripts with positive and negative weights are independently measured using two spectrally distinct reporters. **b,** As an initial test, we added 20 oligonucleotides (3nM) corresponding to the target sequences of each hybridization probe individually and measured the fluorescence response across both channels. Targets 1-10 corresponded to transcripts with positive weights (FAM) while targets 11-20 corresponded to transcripts with negative weights (ROX). As expected, each target resulted in specific triggering of the assigned reporter with almost no crosstalk. **c,** The molecular classifier was tested using *in vitro* transcribed RNA transcripts. Addition of each transcript resulted in a fluorescence signal proportional to the weight associated with a transcript. **d,** Gene expression data for 6 bacterial and 6 viral samples selected from the training set to validate the molecular classifier. **e,** Gene expression patterns for each sample were replicated by mixing gene amplicons containing T7 RNA polymerase promoter sequences in the ratios expected from the microarray data. Subsequently, the samples were *in vitro* transcribed resulting in production of RNA molecules with approximately 1000X amplification. Upon addition of the molecular classifier, fluorescence signals were recorded across both channels and a classification value was recorded. **f,** All samples were classified correctly by the molecular classifier: a positive normalized signal was obtained for bacterial class samples and a negative for viral class samples. The normalized fluorescence signal matches the estimated computational SVM output, reflecting the correct implementation of the weights in a sample containing multiple RNA transcripts.