Repurposing DNA for information processing and storage

Randolph M. Lopez Barrezueta

A dissertation
submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

University of Washington

2018

Reading Committee:

Georg Seelig, Chair

Luis Ceze

Herbert Sauro

Program Authorized to Offer Degree:

Department of Bioengineering

University of Washington

Abstract

Repurposing DNA for information processing and storage.

Randolph M. Lopez Barrezueta

Chair of the Supervisory Committee:

Georg Seelig

Department of Electrical Engineering

DNA is the principal carrier of biological information. It is present in nearly all forms of life and it enables the transfer of genetic information between many generations. Beyond its natural role, DNA is also a programmable material for nanoscale engineering. In this work, we harnessed the modularity and programmability of synthetic DNA to build information processing and storage applications. In Chapter 1 and 2, we developed two approaches to build DNA-based molecular classifiers to interact with native RNA samples and perform rapid gene expression diagnostics. In Chapter 3, we explored DNA as a storage medium for digital information. Specifically, we explore how synthetic DNA can be manipulated to form long concatemers to facilitate nanopore sequencing readout, enabling a faster and real-time read-head for DNA storage. Altogether, this body of work demonstrates how DNA nanotechnology is a powerful tool to manipulate information at a nanoscale and it can enable the development of new applications spanning multiple scientific disciplines.

Table of contents

Introduction

Technologies associated with synthesis and sequencing of DNA have played a central role in advancing biological sciences over the past few decades. Specifically, automated DNA synthesis, available since the 1970s, along with molecular cloning techniques enabled modifying, copying and introducing new genetic sequences to reveal the workings of biology. In parallel, DNA started gaining interested beyond its role in biology but as a biopolymer with very useful characteristics. In 1981, Nadrian Seeman, considered the father of DNA nanotechnology, first proposed the construction of three-dimensional networks of synthetic nucleic acids[1]. In 1994, Leonard Adleman first demonstrated the use of DNA as a form of computation to solve the seven-point Hamiltonian path problem[2]. Since then, multiple new scientific fields have emerged that leverage DNA as a building block for manipulating information, interacting with biology and building complex molecular structures[3].

DNA is an exceptional material for molecular engineering. The predictability of nucleic acid base pairing and the large sequence space associated with it enables engineers to simulate and create scalable molecular systems. Arbitrary DNA strands can be evaluated *in silico* to derive secondary structure predictions that very closely approximate their actual behavior *in vitro*. The scalability of this engineering approach enabled researchers to build three-dimensional DNA nanostructures with outstanding precision[4], to assemble DNA circuits with tens of interconnected components[5, 6] and to create entirely synthetic genomes with millions of nucleic bases[7]. As the cost of synthesis and sequencing continues to drop, DNA engineering continues to expand into larger scales and new applications.

Using DNA to build computational modules that sense DNA inputs, process information and trigger an output signal has been a common theme in DNA nanotechnology. Specifically, DNA strand displacement has provided a design architecture for the construction of arbitrary computing modules. Despite an extensive body of work, this technology remains largely an academic pursuit with limited real-world applications. One natural substrate for the application of DNA computing is the analysis of biological samples which are made up of thousands of DNA and RNA sequences. One example of this application is the work by David Zhang et. al where DNA nanotechnology principles were harnessed to build dramatically better sensors for single-nucleotide polymorphisms[8, 9].

Gene expression analysis of RNA molecules is another promising real-world application for DNA nanotechnology. The relative quantities of thousands of RNA molecules in biological samples are commonly measured to understand cellular state. With the exception of translational variations and post translational modifications, biological changes are generally reflected in the upregulation or downregulation of a subset of genes. Capturing the state of RNA gene expression results in a snapshot of these biological changes. Building DNA-based computational systems to probe gene expression states have been a subject of interest in the field for many years[10, 11]. As gene expression profiling becomes more promising as a diagnostics tool, the need for better tools to measure gene expression changes has increased over time.

The goal of building a molecular classifier for gene expression profiling is to reduce the dimensionality of the readout to one or a few outputs. Without molecular computing, every RNA transcript level must be independently measured and then used as inputs for *in-silico* classification. In practice, this approach results in the development of diagnostics assay that require physical parallelization of tens or hundreds of measurements corresponding to each RNA transcript (e.g. quantitative PCR, microarrays or next-generation sequencing). Instead, a molecular classifier could in principle sense and process gene expression levels at the molecular level and report the output of the classification task. The result would be a simpler and more inexpensive implementation of gene expression profiling with applications in human health and beyond.

In Chapter 1 and Chapter 2, we explore two different approaches for the development of a molecular classifier for gene expression diagnostics. In Chapter 1, we implemented a design architecture where *in-silico* support vector machines (SVMs) can be accurately translated into a DNA implementation where hybridization and strand displacement carry out the classification task. In Chapter 2, we combined amplification and classification for gene expression profiling of RNA samples at low concentrations. In both cases, we started with a classification problem *in-silico* followed by compiling of DNA probes to solve a given classification task.

Another emerging application for DNA nanotechnology is the use of synthetic DNA to store digital information. Using DNA as a storage medium conveys multiple advantages over existing alternatives: it offers ultrahigh information density (hundreds of petabytes per gram), it can retain information for millions of years and writing and reading DNA will be relevant for the foreseeable future. In 2012, George Church et. al. demonstrated an important milestone by writing and reading 5.27 megabits using 54,898 oligonucleotides. Since then, DNA storage has gained steam and multiple groups have demonstrated increasing capabilities in terms of capacity, throughput and scalability[12-17]. In Chapter 3, we explored how nanopore sequencing is a promising sequencing platform to build a real-time read-head for DNA storage. In collaboration with the Molecular Information Systems Lab (MISL) at the University of Washington, we developed a strategy to maximize the throughput of nanopore sequencing for DNA storage using DNA assembly of short oligonucleotides.

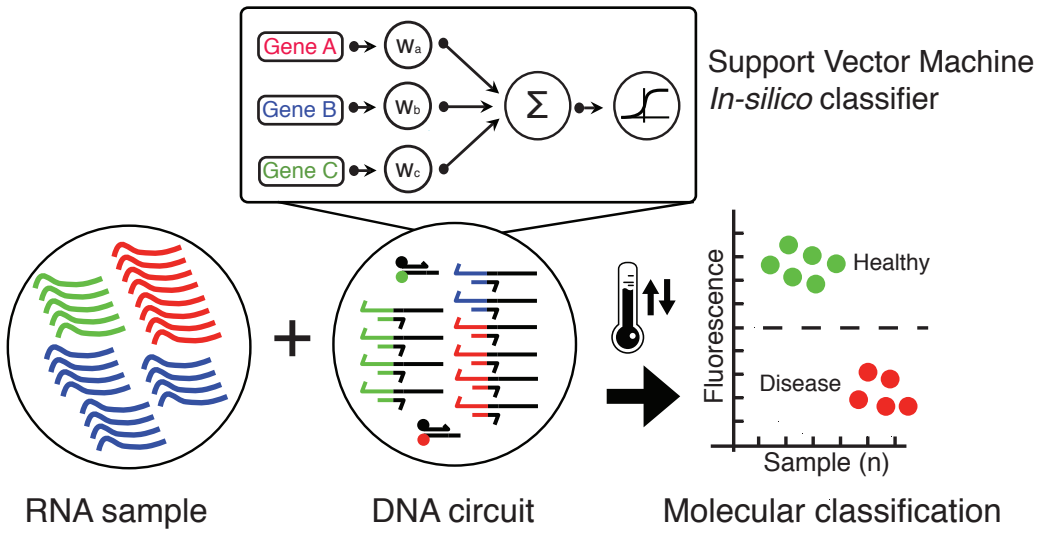
Through this body work, I hope to demonstrate how synthetic DNA is a powerful and versatile engineering material with immense opportunities in information processing and storage. As the cost associated with synthesizing and reading DNA continues to drop, we are currently scratching the surface of possibilities in DNA nanotechnology. It is my hope that this work will be a useful resource for those who will continue developing DNA nanotechnology and will inspire development of more real-world applications in this field.

Publication list

1) J. Bornholt, **R. Lopez**, D. Carmean, L. Ceze, G. Seelig, K. Strauss. A DNA-Based Archival Storage System. ASPLOS '16, 637-649 (2016).
2) Organick. L, S. Ang, Y. Chen, **R. Lopez** et al. Random access in large-scale DNA data storage. Nature Biotechnology 36, 242, doi:10.1038/nbt.4079 (2018).
3) **Lopez, R.**, Wang, R. & Seelig, G. A molecular multi-gene classifier for disease diagnostics. Nature Chemistry 10, 746-754 (2018).
4) **Lopez, R**. et al. DNA Assembly for Nanopore Data Storage Readout. Nature Communications. (submitted July 28th 2018).

Chapter 1

A molecular multi-gene classifier for disease diagnostics



Chapter 1: A molecular multi-gene classifier for disease diagnostics

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[19-24]. 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[21, 25-27]. 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[28].

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[23, 29-33]. For instance, relative quantitation of telomerase reverse transcriptase (hTERT) RNA in blood or serum has diagnostic and prognostic value in many different cancer types[31, 34-38]. Similarly, researchers have demonstrated that a classifier based on a patient's blood RNA profile can distinguish between bacterial and viral infections[28, 39]. 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[8, 40-43]. 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[41]. 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 [5, 6, 11, 44-47]. 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)[5, 6, 44, 45, 48], 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[49-51]. 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[11].

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 RNA transcripts using DNA strand displacement

Initially, we implemented a room temperature strand displacement cascade to detect RNA transcripts in solution. Specifically, we designed a two-stage cascade whereby an input sequence within the target RNA transcript (domain *a*) binds to a complementary probe via a toehold-exchange mechanism. As a result of this initial reaction the longer probe strand becomes attached to the transcript and a toehold (domain t1*) is exposed within that strand. In the subsequent strand displacement reaction, domains t1* x* in the bound probe strand interact with a universal fluorescent reporter resulting in an increase in fluorescent signal. Importantly, because of the two-stage design, the target sequence on the transcript is completely independent of the reporter

sequence. Different regions on a transcript can thus be targeted by changing the hybridization probe sequence but using the same fluorescent reporter.

Next, we considered how to select a specific target domain for probe binding within the much longer RNA transcript. In previous applications, single-stranded nucleic acids acting as inputs to DNA strand displacement cascades had generally been designed to have minimal secondary structure, because even limited structure in the toehold domain can reduce the reaction rate by orders of magnitude. In contrast, in our application the input sequence is constrained by a high degree of secondary structure which decreases the thermodynamic gain of completing displacement. In an initial attempt to minimize the impact of secondary structure we used Soligo, a software for rational design of antisense probes, to identify accessible target sites. For a first test, we used the sequence of a transcript coding for the fusion protein H2B-citrine. We designed a total of nine probe sequences, four of which (positions 204, 988, 1033 and 1078 counting from the beginning of the 5'UTR) were predicted to have low secondary structure and high target accessibility ($\Delta\Delta G_0 < -12$) to Soligo. To experimentally test our predictions and probe design, we synthesized the H2B-citrine RNA transcript using in-vitro transcription. We then combined this transcript with pre-assembled probe and reporter and followed the reaction at room temperature by monitoring changes in the fluorescence signal. Strand displacement probes and RNA transcripts were added at 50 nM and 30 nM respectively in the final mixture and in each experiment, a single probe was tested. We observed significant triggering in five out nine regions, including those that had better predicted target accessibility based on Soligo predictions**.**

Although these results demonstrate that the probe mechanism works as designed and that targeting regions with low structure is a promising strategy, there are downsides to this approach. Additionally, the number of probes that can be used to target a single transcript is limited and largely determined by the secondary structure of the transcript. Moreover, even if the target sited is accessible, reaction kinetics can vary between different sites and reaction rates can become limiting at low transcript concentrations.

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

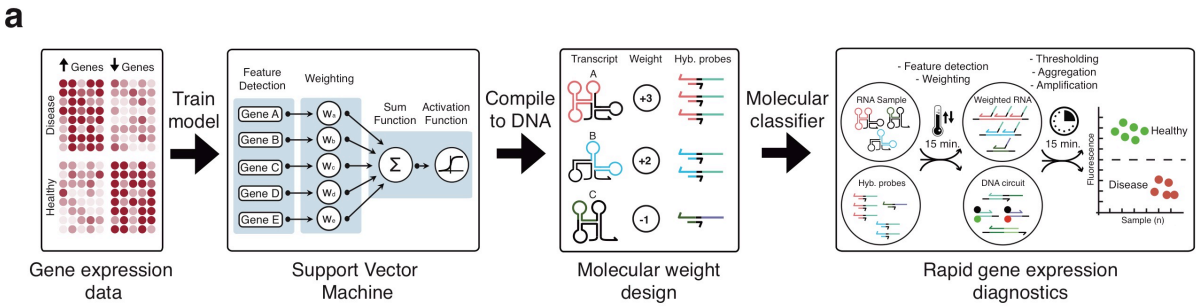
on the RNA to further help unfold its secondary structure and to prevent binding between the adjacent RNA regions and the single stranded domain of the hybridization probe. As a result of this initial reaction the longer probe strand becomes attached to the transcript and a short toehold (domain t1*) 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.

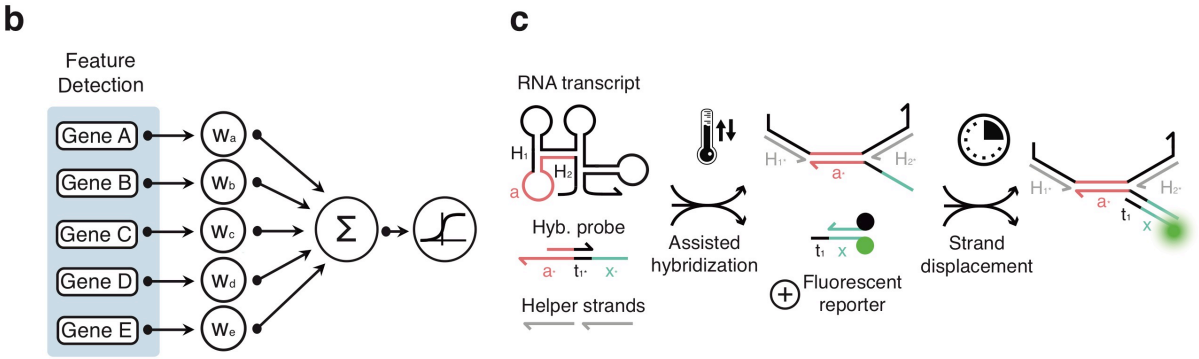
To experimentally test this strategy, we designed hybridization probes to target three different regions of an mRNA coding for the fusion protein histone 2B Citrine (Citrine) as well as a control hybridization probe specific to GAPDH. For an initial test of the probe design with an unstructured target, 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).

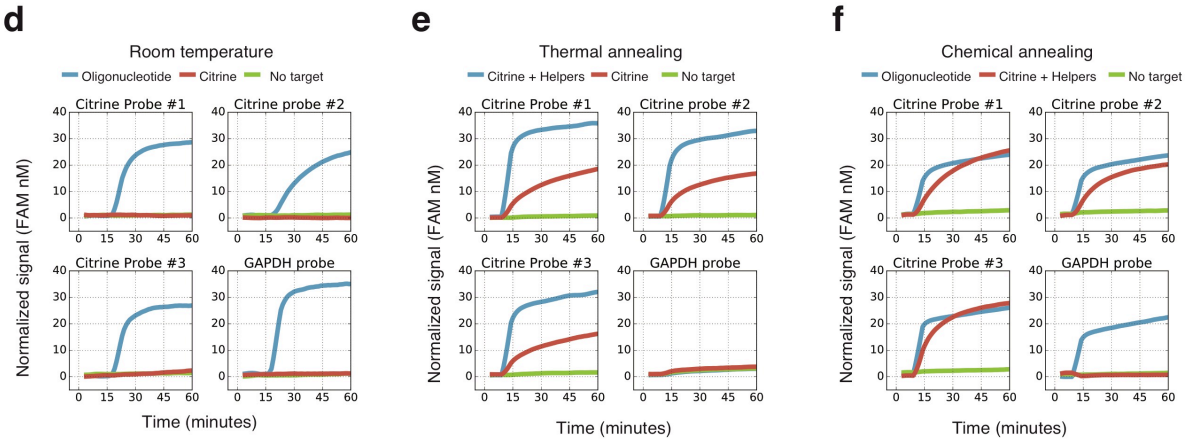
Subsequently, we implemented a thermal annealing strategy where the hybridization probe and corresponding helper strands were annealed with the Citrine RNA transcript before addition of the fluorescent reporter. Thermal annealing was performed by heating reactants to 70°C for 10 seconds and subsequently 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[52]. 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.


a
Genes
Genes
Disease
Healthy
Train model
Feature Detection
Weighting
Gene A
Gene B
Gene C
Gene D
Gene E
Sum Function
Activation Function
Compile to DNA
Transcript
Weight
Hyb. probes
A
B
C
+3
+2
-1
Molecular classifier
- Feature detection
- Weighting
- Thresholding
- Aggregation
- Amplification
RNA Sample
15 min.
Weighted RNA
15 min.
Hyb. probes
DNA circuit
Fluorescence
Healthy
Disease
Sample (n)
Gene expression data
Support Vector Machine
Molecular weight design
Rapid gene expression diagnostics


b
Feature Detection
Gene A
Gene B
Gene C
Gene D
Gene E
c
RNA transcript
Hyb. probe
Helper strands
Assisted hybridization
Fluorescent reporter
Strand displacement


d
Room temperature
Oligonucleotide
Citrine
No target
Citrine Probe #1
Citrine probe #2
Citrine Probe #3
GAPDH probe
Normalized signal (FAM nM)
Time (minutes)
e
Thermal annealing
Citrine + Helpers
Citrine
No target
f
Chemical annealing
Oligonucleotide
Citrine + Helpers
No target

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).

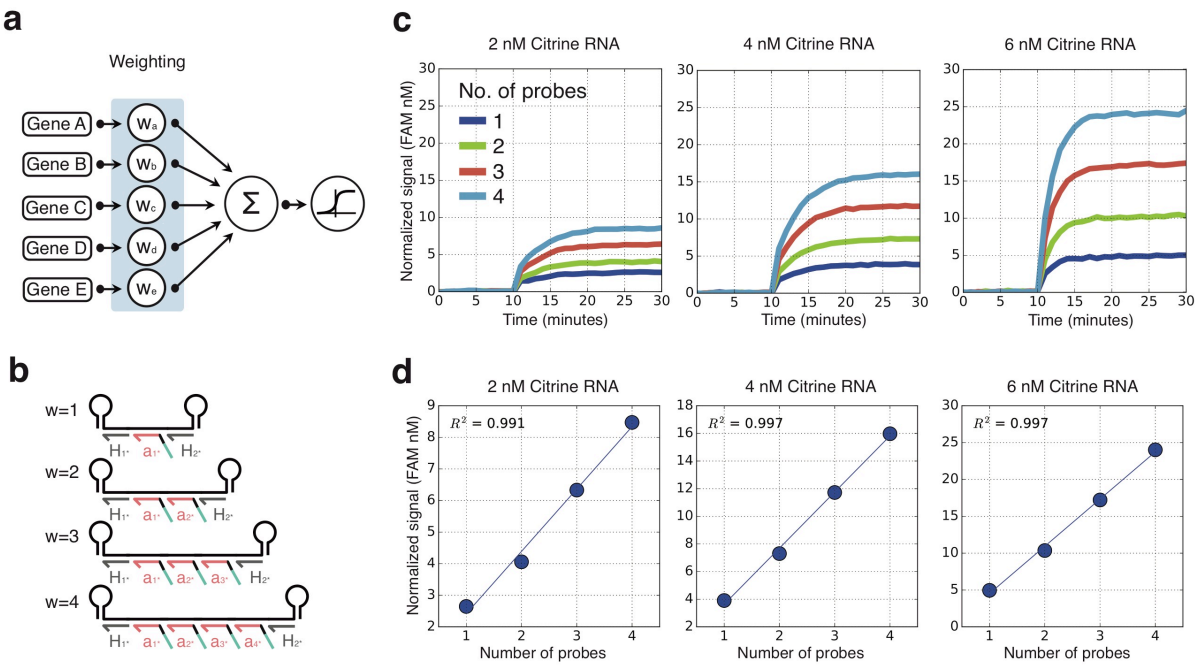


Figure 2 | Implementation of classifier weights by targeting of multiple adjacent regions in a transcript. a, Each transcript is assigned a weight reflecting its influence in the classifier decision. **b,** Each transcript is targeted with a number of probes equivalent to its classifier weight. By targeting 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.

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)[11, 53]. 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[53, 54]. 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[31, 34-38, 55, 56]. 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.

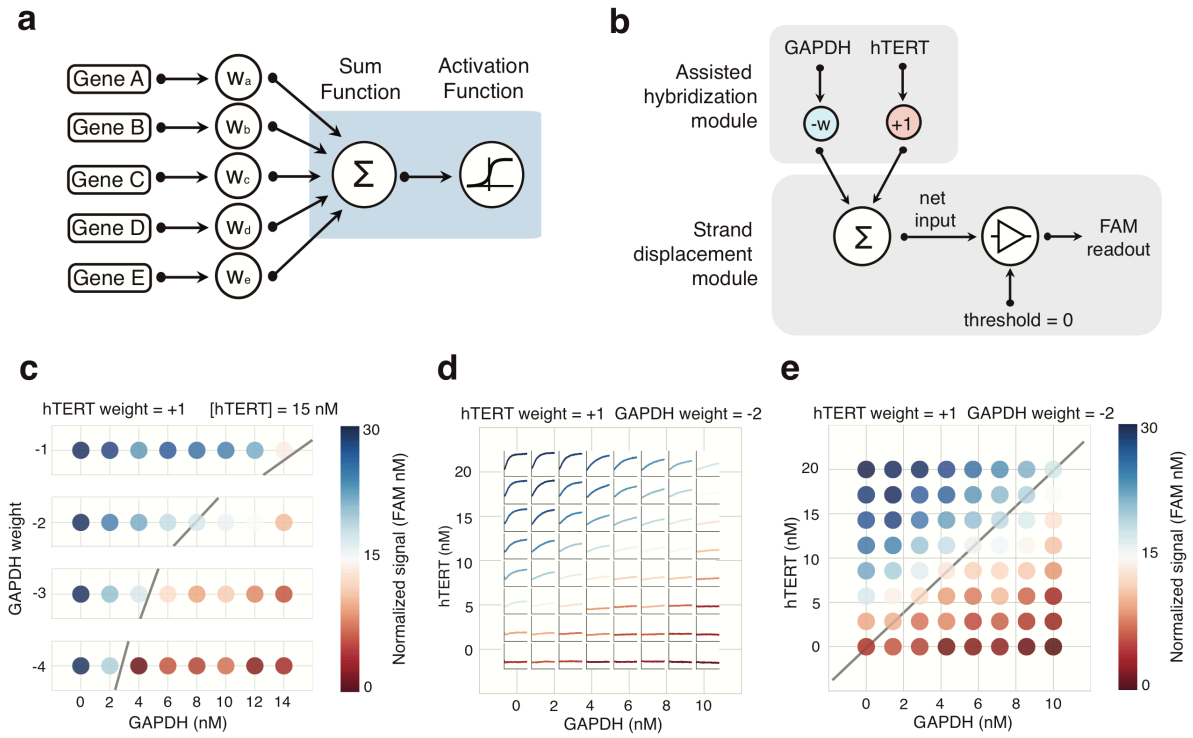


Figure 3 | Molecular implementation of a two-gene classifier for cancer diagnostics a, A sum and activation function are used to aggregate weighted gene expression information into a single, interpretable output. Upon transcript detection and scaling, a sum function calculates the resulting net input. 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.

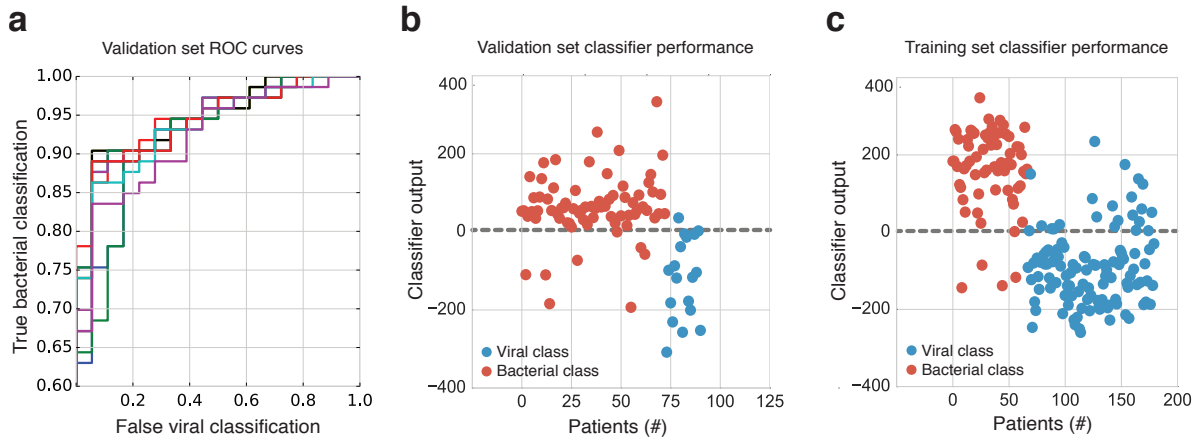
Training a multi-gene support vector machine

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[28].

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[28]. 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[39]. 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).

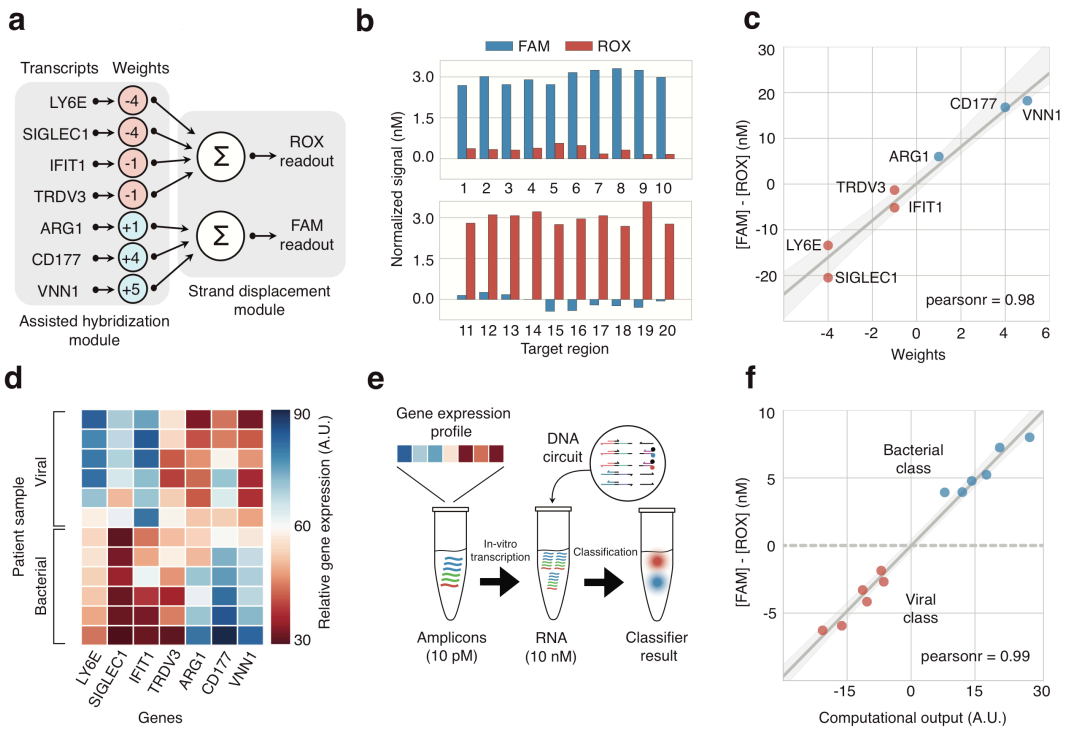


Figure 5 | A molecular classifier of host gene expression for respiratory infections diagnostics. a, Graphical representation of the viral vs. bacterial infection classifier. The classifier uses 7 genes. 20 hybridization probes assign weights ranging from -4 to +5 to each transcript. The weighted sums 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.

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

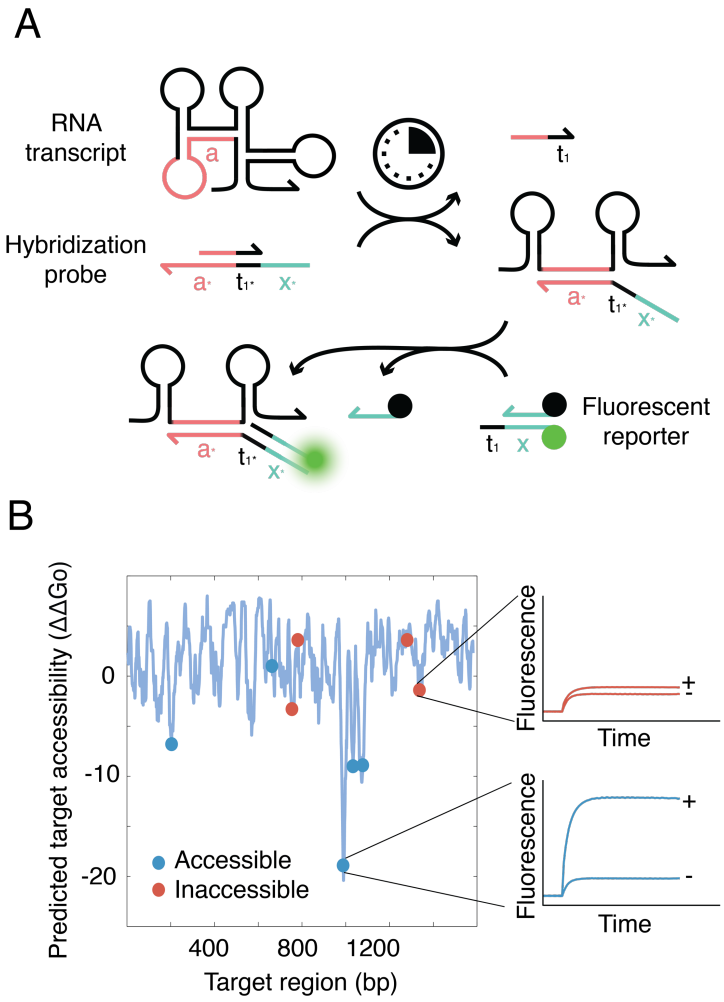
principle be scaled to tens or hundreds of targets with the number of weights only limited by the size of the transcripts. In principle, potential cross-talk between probes and incorrect targets becomes more likely when the number of probes is higher. Nevertheless, a thermodynamic simulation of these 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.

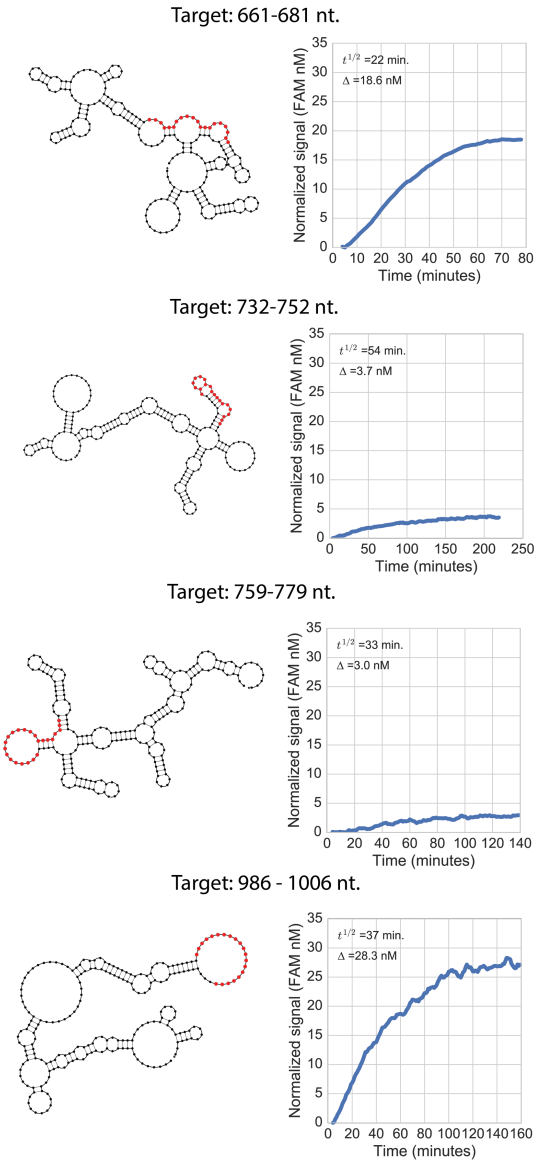
It should be noted however that the rate of the strand displacement reaction is highly dependent on the concentration of the RNA inputs, and including a pre-amplification step in the workflow would increase processing time. In this work, we demonstrated amplification of a mixture of cDNA amplicons in 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[23, 37]. 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[43, 57-60]. 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[40, 41].

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.

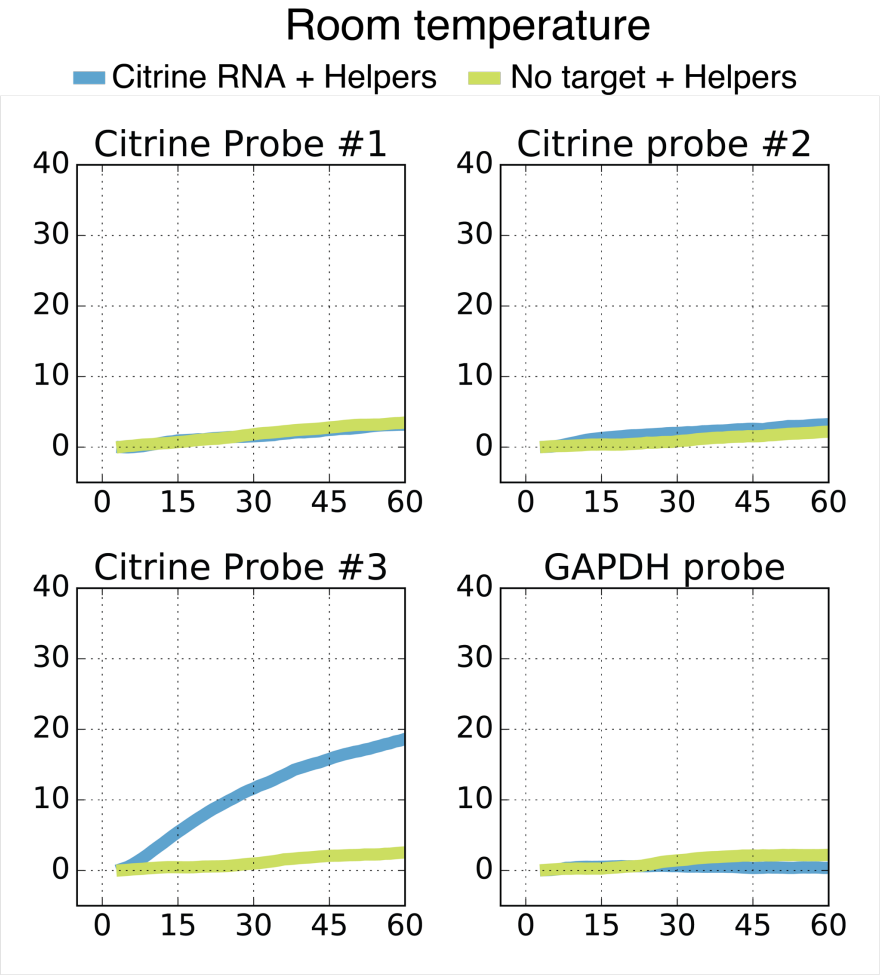
Supplementary figures



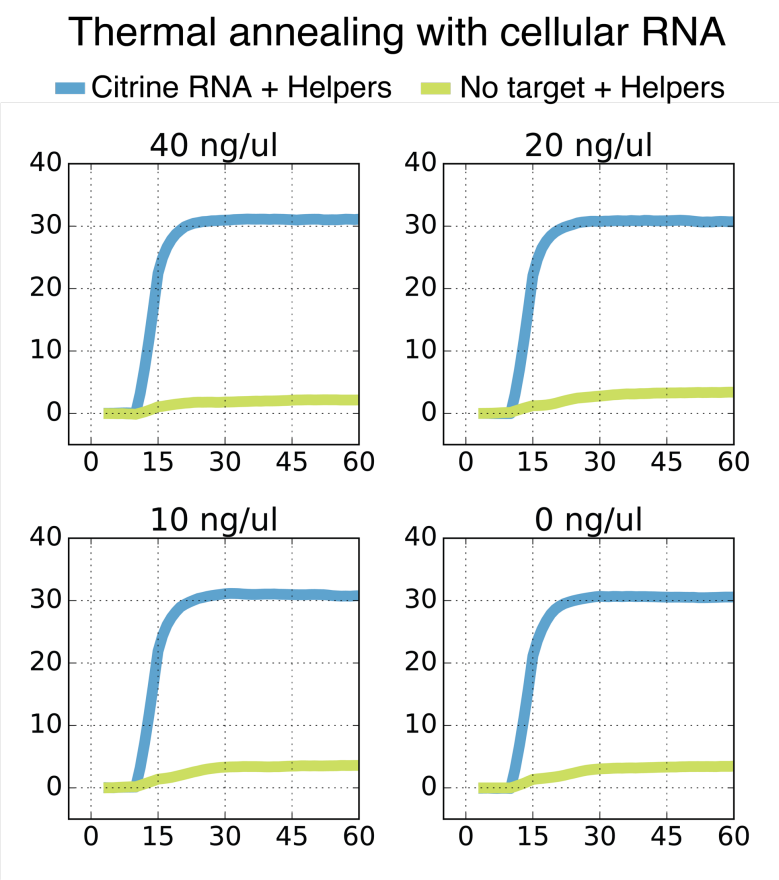
Supplementary figure 1 | Molecular overview of room temperature detection of RNA transcripts using DNA strand displacement.
a, A hybridization probe and a fluorescent reporter are used to detect an in-vitro transcribed RNA transcript in solution via DNA strand displacement. By modifying the sequence composition of domain 'a' in the hybridization probe, we targeted different regions in a transcript while using the same fluorescent reporter. **b,** We targeted strand displacement probes directly to an RNA transcript. We found that out of 9 different target regions, only 5 of them resulted in significant triggering of the strand displacement reaction due to secondary structure in the target RNA. Target sites with higher predicted accessibility (lower free energy) were found to trigger the strand displacement reaction.



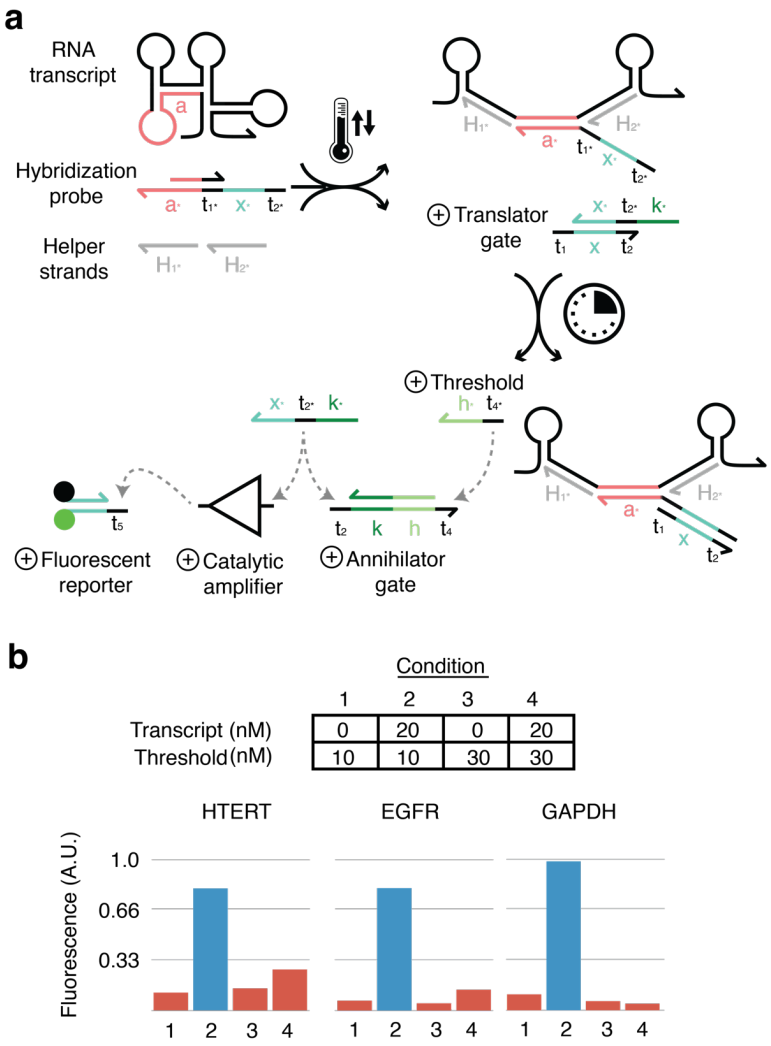
Supplementary figure 2 | Target secondary structure prediction and corresponding kinetic traces for strand displacement probes targeting different regions of a Citrine transcript at room temperature.
Minimum free energy structures corresponding to each target region and adjacent 100 nucleotides were obtained from Nupack. The highlighted red region corresponds to the target sequence for each hybridization probe. We evaluated each RNA target by using different hybridization probes under identical experimental conditions ([Reporter] = 30 nM, [Hyb. probe] = 30 nM and [Citrine RNA] = 30 nM). We observed a fluorescence response equivalent to the RNA input in probes targeting regions 661-681 and 986-1006 nt while the other two regions showed very little triggering.



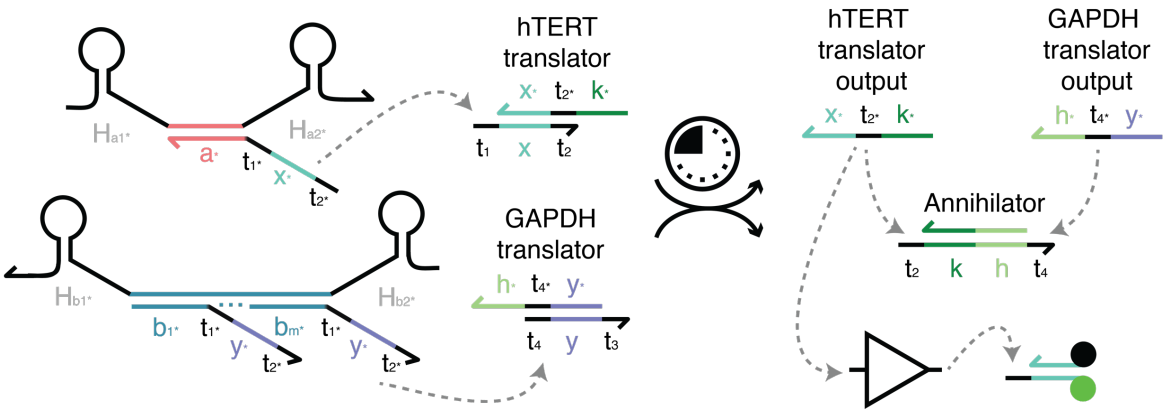
Supplementary figure 3 | Triggering of different hybridization probes using Citrine RNA and corresponding helper strands at room temperature.
Experiments were carried out with 50 nM of fluorescent reporter, 40 nM of hybridization probes, 30 nM of Citrine RNA and 400 nM of helper strands. We only observed a significant fluorescence response in Citrine Probe #3. These results indicate that addition of the helper strands is not enough to enable consistent triggering of the probes at room temperature.



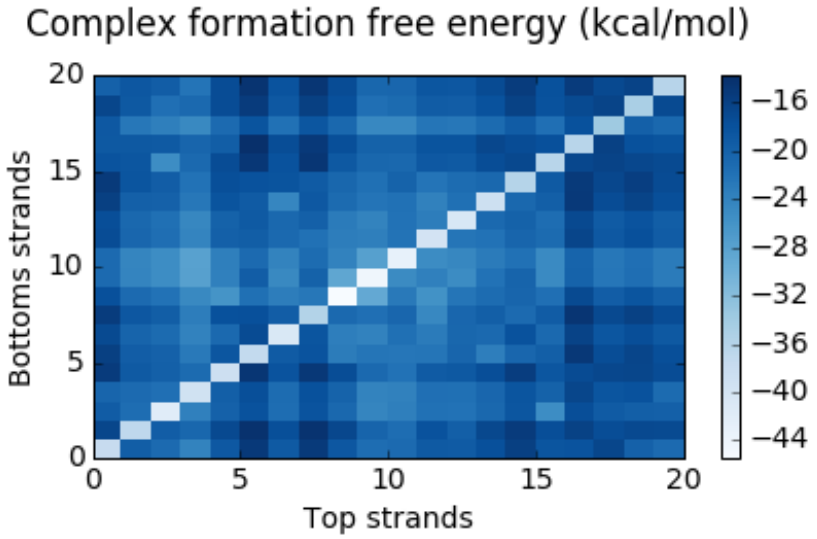
Supplementary figure 4 | Triggering of Citrine hybridization probe under different concentrations of cellular mRNA using thermal annealing.
Experiments were carried out with 50 nM of fluorescent reporter, 40 nM of hybridization probes, 30 nM of Citrine RNA and 400 nM of helper strands. Cellular mRNA was extracted from HEK293 human cell line and added to the annealing reaction at concentrations of 40, 20 and 10 ng/µl. The concentration of citrine RNA in the annealing reaction was 125 ng/µl. We observed no significant change in triggering under different concentrations of cellular mRNA.



Supplementary figure 5 | Absolute thresholding of RNA transcripts using a DNA strand displacement circuit.
a, Molecular overview of a thresholding/amplification cascade with RNA inputs. Upon thermal annealing, a translator gate is triggered based on the initial amount of RNA transcript. The output of the translator gate is consumed by an annihilator gate based on the concentration of a threshold species. If the concentration of a transcript is higher than that of a threshold species, a catalytic response triggers a fluorescent reporter. b, Steady state fluorescent signal after a two-hour reaction. All components were added in excess except for transcript and threshold species. A significant fluorescent response is observed when the concentration of transcript exceeds that of the threshold species (condition 2) and no significant fluorescent response is observed otherwise (conditions 1, 3 and 4).



Supplementary figure 6 | Molecular overview of the hTERT/GAPDH thresholding module.
Circuits were tested with varying number of GAPDH hybridization probes, corresponding to different GAPDH weights. A positive and a negative translator cascade the output signal from the assisted hybridization reaction of hTERT and GAPDH respectively. The outputs of the translator gates are consumed in equimolar manner by the annihilator gate. Only excess positive translator output that is not consumed in the annihilation reaction triggers a catalytic amplifier resulting in a fluorescent signal.



Supplementary figure 7 | Cross-talk analysis among viral/bacterial hybridization probes using thermodynamic simulations.
We used Nupack, a software for analysis and design of nucleic acid structures, to simulate hybridization among a set of 20 probes corresponding to viral/bacterial classifier. To check for cross-talk, we check hybridization between every bottom and top DNA strand corresponding to 20 probes (400 hybridization simulations). We found very little cross-talk between strands from different probes. As expected, hybridization between the correct pairs was highly favorable (diagonal pattern in the plot). In a larger classifier, this approach can be implemented to exclude probes with cross-talk.

Methods

DNA oligonucleotides
All DNA oligonucleotides were purchased from Integrated DNA Technology (IDT). Individual DNA oligonucleotides were suspended to 100 □M 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

For experiments using individual transcripts, kinetic fluorescence measurements were performed using a Horiba FluoroMax 3 spectrofluorometer and Hellma Semi-Micro 114F cuvettes. An external temperature bath maintained reaction temperature at 25°C. A four-sample changer was used so that time-based fluorescence experiments were performed in groups of four. For experiments related to comparison across multiple transcripts (e.g hTERT vs. GAPDH classifier, viral/bacterial classifier), 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

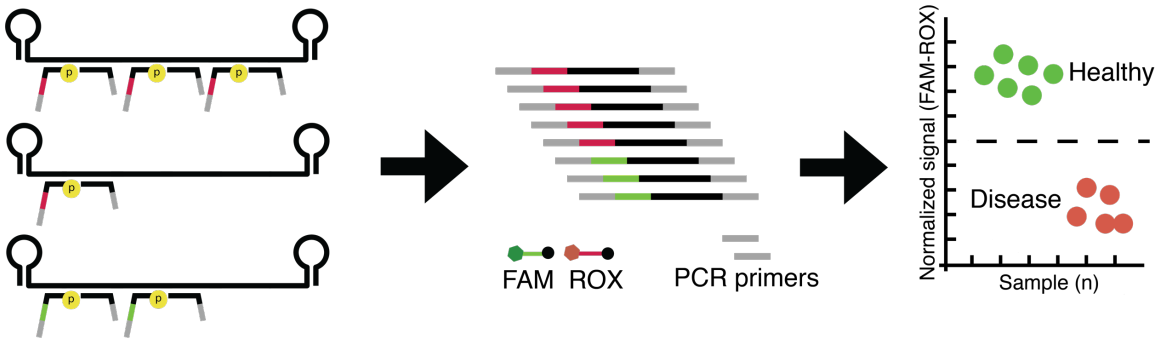
For training of the support vector machine algorithm, we obtained microarray data (NCBI GSE63990) for 273 ill patients and 44 healthy volunteers[28]. We processed the dataset by first selecting samples labelled only as bacterial or viral infections (70 and 115 samples respectively) and transforming the 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)[39]. 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.

Chapter 2

Combined amplification and molecular classification for gene expression diagnostics



Chapter 2: Combined amplification and molecular classification for gene expression diagnostics

Abstract

Gene expression profiling of clinical samples currently requires separate amplification and quantitation of each transcript. Here, we introduce a combined amplification and molecular classification strategy for gene expression classification. We adapted RASL (RNA-mediated oligonucleotide annealing, selection, and ligation) probes using an inexpensive ligation method and by incorporating a positive or negative barcode on each probe for molecular classification. Then, we designed 29 different probes for implementation of a classifier to distinguish among different human cancer cell lines. We demonstrated batch characterization of these probes using a next-generation sequencing platform. Finally, we analyzed and discussed how probe specific bias will impact the selection of a subset of probes to build a desired molecular classification system.

Introduction

Molecular gene expression analysis requires quantification of multiple RNA biomarkers in a given sample. In a biomarker discovery setting, this is currently performed using RNA-seq or microarray assays[21, 49, 50, 55, 61, 62]. These platforms enable the quantitation of thousands of different RNA biomarkers per sample. Once a subset of RNA biomarkers have been discovered and validated, gene expression analysis currently relies on quantitative PCR to measure a smaller number of RNA biomarkers relevant for a particular application. Quantitative PCR enables a faster and more cost-effective approach than RNA-seq or microarrays, but its implementation remains labor intensive and requires expensive instrumentation. A significant challenge is that each transcript requires a separate reaction for amplification and fluorescent measurement. Therefore, qPCR for gene expression analysis is not suitable for applications requiring inexpensive point-of-care screening or recurrent monitoring.

In Chapter 1, we demonstrated a strategy where gene expression classification is performed by a set of DNA probes without the need for measuring each individual biomarker. Therefore, any given panel of gene expression biomarkers can be classified in an individual reaction which drastically simplifies sample preparation requirements. However, the proposed workflow enabled classification of RNA samples at high concentration (picomolar to nanomolar). Most diagnostics applications require classification of RNA samples in the fentomolar range. Therefore, enabling both rapid amplification and molecular classification of RNA samples is necessary for a practical implementation of this technology.

There are several approaches for targeted RNA amplification including NASBA, rolling-circle amplification, DASL, RASL and multiplex PCR[63-68]. Upon evaluating and characterizing several of these methods, we decided to implement a molecular classification workflow using RASL probes. RASL (RNA-mediated oligonucleotide annealing, selection, and ligation) is commonly used for targeted RNA sequencing for the purpose of gene expression quantitation[67]. First, custom probe pairs are designed for each gene of interest. Each pair is designed to anneal to the target RNA in an adjacent manner such that the resulting gap can be ligated. Each probe also contains universal amplification overhangs. Experimentally, the probe pairs are hybridized to the

mRNA and separated from total RNA using oligo(dT)-biotin beads following by ligation and clean-up. PCR amplification of the ligated probes enables accurate amplification of low total RNA amounts (10 nanograms) for gene expression quantitation using next-generation sequencing.

In this work, we demonstrated a novel approach for combined amplification and molecular classification for gene expression analysis using modified RASL probes. We created RASL probes with an additional classification sequence domain targeting TaqMan reporters with two distinct fluorescence profiles. This enables a classification output to be activated at the same time as PCR amplification of different ligated probes. Since RASL probes typically display biased amplification profiles[67, 69], we created a strategy for batch testing many different probes in order to select those that would implement the best classification model. Finally, we analyzed the bias associated with each probe and discussed how a potential classifier can be designed based on this method.

Results

Overview of combined amplification and classification of RNA samples

We design a novel approach for building a molecular gene expression classifier by adapting RASL probes, commonly used for targeted RNAseq, into a combined amplification and classification workflow using a two-color fluorescent system. First, a relevant gene expression dataset is used to train a support vector machine to build a classification model that differentiates between two class labels (e.g. healthy vs. diseased). The resulting model consists of a list of relevant genes and corresponding weights capable of performing the classification. Next, a set of RASL probes are designed to target each of these transcripts. Each RASL probe contains either a positive or negative sequence barcode designed to trigger fluorescent reporters (FAM and ROX). Since each probe is expected to have a different amplification bias, multiple probes per gene are generated and characterized together. For this purpose, all probes are combined and evaluated on the presence of an RNA sample followed by next generation sequencing. This data is then used to quantify probe bias and determine a subset of probes capable of performing the classification task. Finally, this subset of probes can be used to implement a rapid gene expression diagnostics system where tens or hundreds of RNA transcripts can be amplified and serve as input for a classification problem.

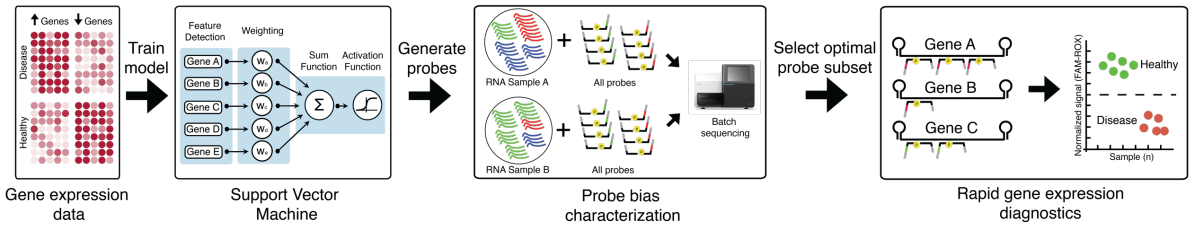


Figure 1 | Combined amplification and classification of RNA samples using modified RASL probes. An in-silico classifier is trained and validated on publicly available gene expression data. Then, multiple probes are designed and synthesized to target the set of genes in the classifier. This set of probes is hybridized, ligated and amplified using an RNA sample with known gene expression data. Using next-generation sequencing, the number of amplified probes can be counted in batch in order to determine probe specific amplification bias. Subsequently, this data informs the selection of an optimal subset of probes that will implement the desired classification model.

Adapting RASL probes and workflow for molecular classification

In a typical RASL experiment, a set of two or three probe pairs are designed for every gene that is targeted. These probes are hybrid RNA/DNA such that the gap can be ligated using RNA ligase 2, a very efficient ligase enzyme. Each probe pair consists of a hybridization domain, targeting the RNA transcript, and conserved overhangs for PCR amplification. After probe amplification, next generation sequencing is used to quantified each probe count and perform differential gene expression analysis across different samples. Importantly, since each probe exhibits different ligation and amplification efficiency, the resulting data can only be used for relative expression difference between samples instead of absolute quantitation.

In order to enable inexpensive prototyping of our RASL-based molecular classification probes, we first modified the RASL probes by using probes made up entirely of synthetic DNA. These probes are about one order of magnitude cheaper to synthesize. In order to enable ligation of DNA probes on an RNA transcript target, we substituted RNA ligase 2 for SPLNTR, a novel enzyme that enables DNA-DNA ligation on RNA target. We compared both enzymatic ligation reactions and we observed a minimal decrease on ligation-amplification efficiency. Specifically, there is observable increase in non-specific ligation in the absence of target. However, using DNA probes resulted in an order of magnitude reduction in the cost of probe synthesis.

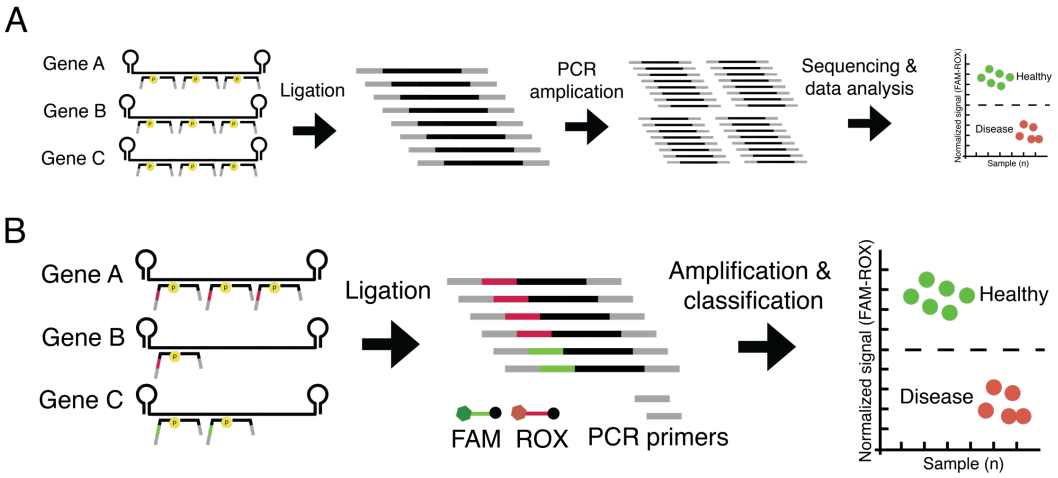
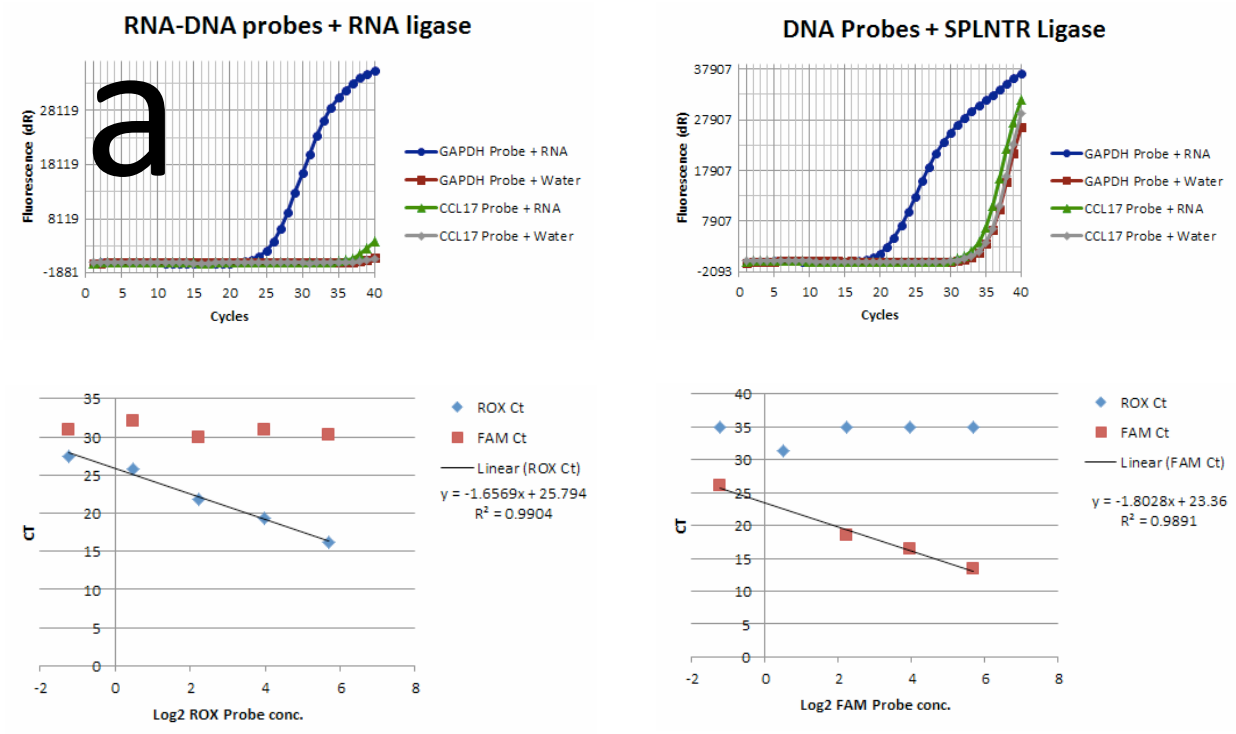


Figure 2 | Overview of existing RASL-seq and our modified method for molecular classification a, Overview of RASL-seq, a common method for targeted gene expression counting using next-generation sequencing. Multiple probe pairs are hybridized to their RNA targets followed by ligation of gaps. The ligated products can be amplified using common primers and then used as input for next generation sequencing. Sequencing data is then used for gene expression analysis. **b,** Our modified RASL probes contain a positive or negative barcode (red or green domain) associated with each probe. The number of probes that bind to each transcript can be varied to tune the effective amplification weight on each transcript. During amplification, two fluorescent reporters (FAM and ROX) are triggered based on the presence of each barcode. This fluorescent signal can then determine the classification outcome for a given sample.

Next, we modified RASL probes by incorporating a sequence barcode that associates each probe with either a positive or negative weight in the classifier. Specifically, a 25-nucleotide barcode triggers a FAM or ROX fluorescent reporter during PCR amplification. Any gene with a positive weight in the classifier is assigned a probe pair with the barcode triggering the FAM reporter and viceversa. The FAM or ROX reporters correspond to TaqMan probes. These fluorescent reporters are commonly used for sequence-specific fluorescent triggers for quantitative PCR. For

this work, they were designed to be triggered by either the positive or negative barcode in each probe. To test this, we two pairs of probes, each containing either a barcode triggering the FAM or ROX barcode. We amplified the probes after ligation using a TaqMan amplification master mix containing both reporters. We titrated each probe separately under absence of the other probe to check for the sensitivity range and potential cross-talk among fluorescence channels. As expected, we found a linear relationship between the log2 of the concentration of the probe and the fluorescence signal in the corresponding channel. We found no observable cross-talk.



Translating an in-silico classifier into RASL probes

We decided to test our approach for joint amplification and classification of RNA samples by building a cancer cell line molecular classifier. Cancer cell lines are easily available and can generate significant amounts of RNA for prototyping. We selected three cancer cell lines (HEK-293, MCF7, HEPG2) already available in our laboratory and retrieve their corresponding gene expression profile from the Cancer Cell Lien Atlas[70]. Using this gene expression profile, we built

an arbitrary classifier to distinguish HEK-293 from MCF7 and HEPG2. As in Chapter 1, we implemented a Support Vector Machine (SVM) training algorithm to select a number of transcripts and associated weights that would implement the classification task. The resulting classifier consisted of five different transcripts with corresponding weights from -5 to +4.

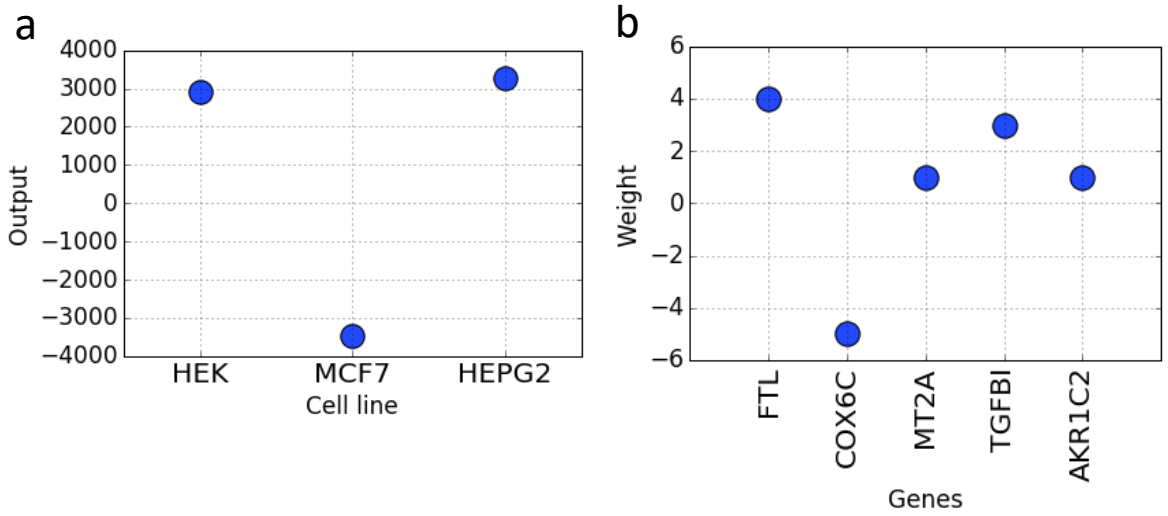


Figure 4 | Overview of *in-silico* cell line classifier to distinguish MCF7 vs. HEK and HEPG2. a, The classifier displayed very good performance to differentiate MCF7 against the other two cell lines. **b,** The necessary classification performance was achieved with only five transcripts and associated weights spanning from -5 to +4.

We decided to build a molecular implementation of this classifier by designing DNA probes that would hybridize to the corresponding transcripts. However, we expect the ligation and amplification of each probe to vary significantly based on previous characterizations of the RASL system. This problem is inherent to other approaches that rely on amplification of individual transcripts such as multiplex PCR or rolling circle amplification. Based on analyzing data associated with existing RASL experiments, we expected the probe associated bias to be within an order of magnitude measure by amplification efficiency. This probe associated bias would make it difficult to build a molecular classifier where each probe would contribute with a unit weight to the classification problem as we demonstrated in Chapter 1. Therefore, we decided to order six different probes for each transcript and then characterize how the bias would affect the implementation of a molecular classifier.

Probe design started by segmenting each RNA transcript sequence into 40 nt. segments, since each probe hybridizes to 20 nucleotides of the target. Probe sequences were filtered to have a GC content between 30% and 70%, a melting temperature between 60C and 85c and to contain either an 'A' or a 'T' at the donor base. Probes that satisfied these criteria were selected and the adaptor universal sequence containing the universal priming domain and the positive or negative barcode was added to the corresponding end. Subsequently, these probes were evaluated for self-complementary using NUPACK and those with high secondary structure were excluded. We also evaluate each probe per transcript to avoid any overlap in the target sequence. After meeting these conditions, we found at least 6 probes per transcripts except for transcript MT2A for which we only found 5 probes. Overall, we designed and synthesized 29 different probe pairs. In order to characterize probe bias, we carried out a batch experiment were the pooled probes were ligated, amplified and sequencing using RNA from three different cell lines as input. As with a

traditional RASL experiments, we expect that the sequencing counts of each probe would correspond to the gene expression differences between the cell lines adjusted by each probe bias.

After sequencing, we analyzed sequencing reads by counting each probe across the three different cell lines. First, we compared the sequencing counts of each probe with the gene expression profile in the database used to train the *in-silico* classifier. Overall, we found that the sequencing counts for each probe matched the relative gene expression differences in the database. Specifically, there is perfect agreement with the expected expression for gene FTL and AKR1C2 while for the other three genes there was some discrepancy with one of the cell lines. Despite these gene expression differences, we found that all probes targeting a given RNA target resulted in the same expression pattern with very few exceptions (Probe #5, #8, #23). Therefore, the probes are likely capturing actual gene specific expression differences between the cell lines based on the gene they are targeting.

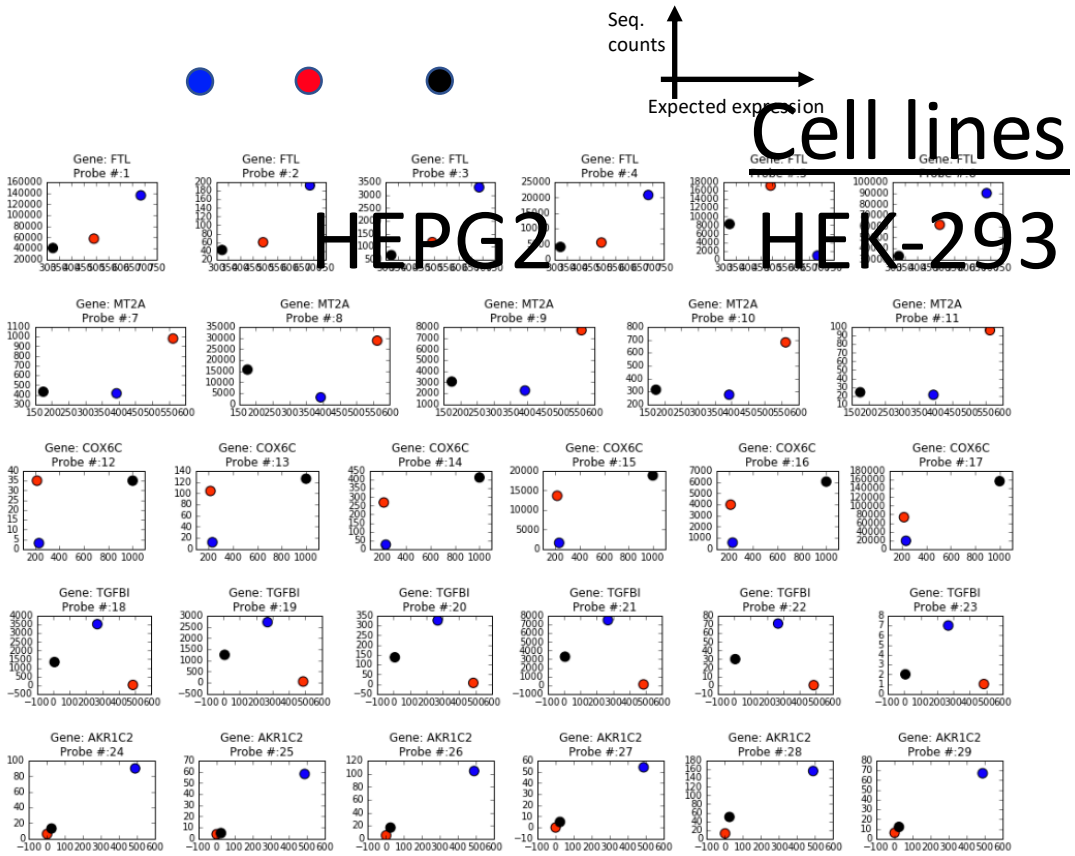


Figure 5 | Observed gene expression differences between batch modified RASL experiment and expected gene expression data from Cancer Cell Line Atlas. Each subplot corresponds to the sequencing counts for each probe (y-axis) and the corresponding expected gene expression for each cell line (x-axis). FTL and AKR1C2 resulted in sequencing counts that matched closely to the expected sequencing data.

We analyzed the sequencing read counts associated with different probes across the same gene target to evaluate probe bias. For example, Probe #1 and Probe #2 target gene FTL and they captured the same gene expression pattern but display sequencing counts that differ by three orders of magnitude. We estimated probe bias by assigning a baseline value of 1.0 to the probe with the maximum sequencing read count per given gene and comparing every other probe

targeting that gene. We found that 89% of probes fall within 3-order of magnitudes from the probe with highest sequencing counts. This is significantly higher bias than that observed in reported RASL experiments where usually probes fall within 1-order of magnitude from one another. One reason for this difference could be the use of SPLNTR ligase and DNA probes instead of RNA ligase. Even though we evaluated specificity and sensitivity between these two enzymes, it's possible that the bias between the probes is more pronounced using SPLNTR. Another possible explanation is that our criteria for probe selection (GC content, melting temperature) was no stringent enough to select probes with similar hybridization and ligation activity resulting in lower bias. Importantly, we compared the bias for a given probe across the three cell lines (HEK-293 vs. HEPG2 and HEK-293 vs. MCF7) and we found that it is remarkably consistent across all of them. This suggest that characterizing probe bias in a given RNA sample is sufficient to extrapolate its behavior in any sample. Next, we verified that the bias measured during the batch sequencing experiment would correspond to that of measuring the amplification profile in a quantitative-PCR experiment. For this purpose, we selected 11 probes corresponding to MT2A and COX6C and individually measured their efficiency in HEK-293 RNA. The resulting CT values from the qPCR experiment were compared to the sequencing counts for each probe. We found that the sequencing batch experiment is highly predictive of the bias of each individual probe measured via qPCR.

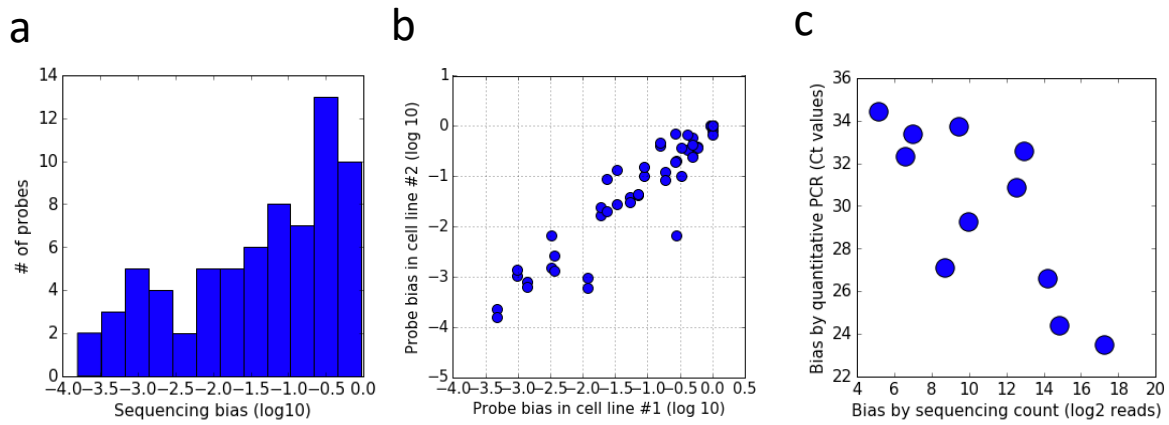


Figure 6 | Characterization of probe bias. a, To estimate probe bias, we calculated the difference in sequencing counts (log10) between every probe and the probe with the highest counts (normalized to zero). We performed this analysis across each gene and cell line in order to estimate the bias distribution of the probes. **b,** We found that the bias is probe specific and does not change across different cell lines **c,** Finally, we found that the bias observed during the batch sequencing experiment corresponded well to the amplification of each individual probe in a qPCR setting.

Despite observing over three orders of magnitude in probe bias in our sequencing experiment, we found that this bias is consistent across cell lines and it is highly predictive of individual probe behavior as measured by qPCR. Therefore, a viable alternative to build a molecular classifier is to first characterize a large subset of probes targeting relevant genes followed by selection of a subset probes that will implement the classification task. Importantly, the bias for each probe can be harnessed to assign weight values to each transcript. In contrast to our previous implementation, each probe consists of a random weight value that needs to be characterized prior to selecting the optimal subset of probes.

Discussion

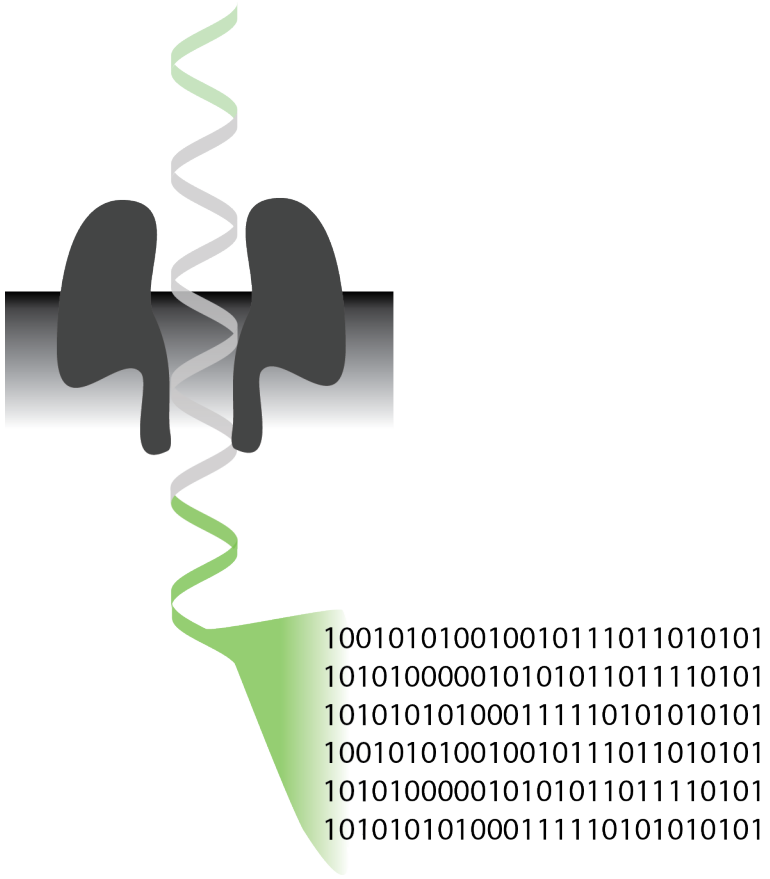
In this chapter, we demonstrated that RASL probes can be easily adapted for the purpose of building a molecular classifier. We modified RASL probes by enabling an inexpensive ligation protocol using SPLNTR and DNA probes. We included a positive or negative barcode in each probe to enable a two-color classification scheme during amplification and experimentally validated this approach using RNA from human cancer cell lines. Finally, we built a classification model *in-silico* to distinguish among three cancer cell lines. Then, we designed and synthesized corresponding RASL probes with barcodes associated to this classification problem. We tested all probes in batch using next generation sequencing in order to determine the amplification bias associated with each probe. We found that this bias was significant and made the implementation of a molecular classifier more challenging that we expected.

We expect that any targeted amplification method of RNA transcripts would result in significant bias relative to the original gene expression profile. To circumvent this challenge, characterization of a significant number of probes could be carried out in advance and then the correct subset of probes could be selected. Alternatively, each transcript could be targeted with both positive and negative probes such that a combination of these probes would implement the correct associated weight. Given the low cost of DNA synthesis and the possibility to fit tens of different probes within an RNA transcript, a subset of probes that implements the correct classification model should be within reach. Generating probes with a better bias distribution would require testing a smaller set of probes and it would facilitate the implementation of a molecular classifier.

Gene expression profiling is increasingly an important clinical metric for diagnosis a wide number of human diseases. Currently, this is carried out by amplifying and measuring each individual RNA transcript in a sample using qPCR. Instead, we demonstrated an approach for amplification and classification of samples in a single reaction. This would drastically reduce the complexity of gene expression profile and potentially enable a point-of-care solution to this type of diagnostics. Even though this approach would result in a more complex development stage, once a final subset of probes is found, the implementation is drastically simpler than the existing method.

Chapter 3

DNA Assembly for Nanopore Data Storage Readout



Chapter 3: DNA Assembly for Nanopore Data Storage Readout

Abstract

Synthetic DNA is becoming an attractive substrate for digital data storage due to its density, durability, and relevance in biological research. A major challenge in making DNA data storage a reality is that reading DNA back into data using sequencing by synthesis remains a laborious, slow and expensive process. Here, we demonstrate successful decoding of 1.67 megabytes of information stored in short fragments of synthetic DNA using a portable nanopore sequencing platform. We designed and validated a novel assembly strategy for DNA storage that drastically increases the throughput of nanopore sequencing. Importantly, our assembly strategy is generalizable to any application that requires high-throughput nanopore sequencing of small DNA amplicons.

Introduction

The digital revolution has resulted in the exponential growth of electronic data storage. During the past three decades, the amount of digital information has doubled every 2.5 years and it is expected to reach 375 exabytes by 2040[71, 72]. In anticipation of this significant increase in data storage demand, synthetic DNA has been widely studied as a promising storage medium for data archival[12-15, 17, 71, 73-76]. DNA offers ultrahigh information density capabilities in the order of hundreds of petabytes per gram[13, 14] and under proper conditions it can retain information for millions of years[75, 77]. Furthermore, technologies that enable high-throughput reading and writing of synthetic DNA have been rapidly evolving in parallel with the advent of the genomics revolution[78]. Nevertheless, several challenges remain unaddressed before DNA storage is able to meet the existing demand for data storage and become a cost-effective alternative to silicon.

DNA sequencing is used to read the information encoded in DNA back into digital bits. Currently, sequencing by synthesis (SBS), as commercialized by Illumina, is the leading technology for high-throughput sequencing[78]. In previous work, we demonstrated the ability to write over 200 megabytes of information in about 2 billion nucleotides while recovering all data without any bit errors using Illumina SBS technology[74]. Despite its low error rate and high throughput, SBS technology has several shortcomings in the context of DNA storage. In its current form, SBS technology is poorly fitted to end-to-end automation, requires bulky and expensive instrumentation and access to sequencing data is delayed until completion.

Nanopore sequencing, as commercialized by Oxford Nanopore Technologies (ONT), offers a sequencing alternative that is portable and automation-friendly, resulting in a better alternative for a real-time "read head' of a molecular storage system[79-83]. Specifically, ONT MinION is a four-inch long USB-powered device containing an array of 512 sensors, each connected to four biological nanopores, capable of producing up to 15 gigabases of sequencing output per flowcell. Each nanopore is built into an electrically resistant artificial membrane. During sequencing, a single strand of DNA passes through the pore resulting in a change in the current across the membrane. This electrical signal is processed in real time to determine the sequence identity of the DNA strand. In the context of DNA storage, real-time sequencing enables the ability to

sequence until sufficient coverage has been acquired for successful decoding without having to wait for an entire sequencing run to be completed **(Figure 1b)**.

Nanopore sequencing presents unique challenges to decoding information stored in synthetic DNA. In addition to a significantly higher error rate compared to SBS, nanopore sequencing of short DNA fragments results in significantly lower sequencing throughput due to slow DNA capture kinetics compared to read time. To address this, previous work by Yazdi et. al. demonstrated a DNA storage workflow where 17 unique large DNA fragments (~1000 bp.) encoding for a 3kB file were synthesized, and subsequently sequenced and decoded using ONT MinION platform[17]. However, existing scalable approaches for writing synthetic DNA rely on parallel synthesis of millions of short oligonucleotides (*i.e.* 100-200 bases in length) where each oligonucleotide contains a fraction of an encoded digital file. Sequencing of such short fragments results in significantly lower sequencing throughput in the ONT MinION, which limits its application as a reading method for DNA storage. Therefore, until now, there was no scalable and cost-effective end-to-end method to enable a DNA storage workflow using nanopore sequencing.

In this work, we demonstrate decoding of three files totaling 1.67 megabytes of digital information stored in 111,499 oligonucleotides using nanopore sequencing **(Figure 1a)**. Our work results in a 2-order of magnitude increase in demonstrated sequencing and decoding capacity using nanopore sequencing for a DNA storage application **(Figure 1b)**. Our real-time sequencing implementation for DNA storage provides a faster and more flexible alternative to decoding digital files encoded in DNA where sequencing can be carried until enough coverage has been acquired for decoding **(Figure 1c)**. To achieve this, we implemented a strategy that enables random access and assembly of a given DNA file stored in short oligonucleotides (150 bp.) into large DNA fragments containing up to 24 oligonucleotides (~5000 bp.). We evaluated Gibson Assembly and Overlap-Extension Polymerase Chain Reaction (OE-PCR) as suitable alternatives to iteratively concatenate and amplify multiple oligonucleotides in order to generate large sequencing reads. Furthermore, we implemented a new consensus algorithm capable of handling high error rates associated with nanopore sequencing. We demonstrated this approach by amplifying, assembling and sequencing 4 different files stored in DNA with significant improvements in capacity (bases/flowcell) and overall throughput (bases/second) **(Figure 1d)**. We were able to decode files with a minimum sequencing coverage as low as 22x, compared to 36x in our previous work[74]. Furthermore, our Gibson assembly concatenation strategy is generalizable to any amplicon sequencing application where higher nanopore sequencing throughput is desirable.

Results

Consecutive Gibson assembly for DNA storage file retrieval

The overall yield and quality of a nanopore sequencing run is dependent on the molecular size of the DNA to be sequenced. DNA molecules translocate through the pore at a rate of 450 bases/sec while it can take between 2 to 4 seconds for a pore to capture and be occupied by the next DNA molecule. Therefore, short DNA molecules result in a higher number of unoccupied pores over time which increases the rate of electrolyte utilization above the membrane. Ultimately, this

results in a faster loss in polarity and lower sequencing capacity and overall throughput. ONT estimates that the optimal DNA size to maximize sequencing yield is around 8 kilobases while the minimum size is 200 bases. Below 200 bases, event detection and basecalling is not possible. Therefore, a PCR-based random-access strategy for sequencing of the files encoded in 150 bp DNA oligonucleotides would have resulted in very low sequencing yield and limited decoding capabilities.

Thus, we developed an assembly method 'Consecutive Gibson Assembly' to assemble large sequencing reads from short amplicons for the purpose of enabling higher sequencing throughput from nanopore sequencing **(Figure 2a)**. We applied this strategy to enable random access and assembly of digital files encoded in oligonucleotide pools containing multiple files in millions of oligonucleotides. Random access is fundamental to the scalability of DNA storage eliminating the need to read all the data stored in a particular DNA pool. Each file in the oligo pool consisted of a set of 150-bases oligonucleotides with unique 20-nucleotide sequences at their 5' and 3' ends for PCR-based random-access retrieval (i.e. file ID) and a 110-nucleotide payload encoding for the digital information. Based on the number of amplicons to be assembled, separate PCR amplification reactions are carried out with primer sets designed to amplify a given file ID **(Figure 2b)**. Each primer set also contains sequentially overlapping overhangs necessary for the downstream assembly.

Overhang sequence design was performed using a nucleic acid thermodynamic simulation software (NUPACK[84]) to avoid primers with self-binding structures and cross-talk between orthogonal overhangs among other constraints **(Supplementary figure 2)**. First, we generated a random set of 30-nucleotide DNA sequences with GC content between 40% and 60%, while avoiding 4-nucleotide repeats of G or C. Subsequently, we used NUPACK to estimate intramolecular secondary structure probability and selected those sequences with minimal secondary structure. This selection step was necessary to enable PCR reactions with comparable amplification efficiencies, which facilitated the generation of the assembly fragments. Next, we evaluated our overhang primer set for orthogonality by estimating binding probability across all possible primer pairs. Based on this cross-talk mapping, we then selected a subset of primers that exhibit minimum cross-talk among all pairs. We repeated this process to create overhangs for every new file to be assembled since each file contained different primer sequences. We expect that selecting an overhang sequence subset with minimal cross-talk can result in higher efficiency in the assembly process. We used this DNA sequence design method to generate overhang sequences for each file assembly.

Upon PCR-amplification of each file with its corresponding priming pairs, PCR products are combined, purified using magnetic beads and combined into a Gibson assembly reaction. During the Gibson reaction, the exonuclease creates single-stranded 3´ overhangs that enable the annealing of fragments that share complementarity at one end while the polymerase fills in gaps within each annealed fragment. Finally, a DNA ligase seals nicks in the assembled DNA. The product of the Gibson assembly was then amplified using primers corresponding to unique sequences present at the ends of the assembly product. We implemented this approach to generate a 6-fragment assembly for each file, respectively. Using agarose gel electrophoresis, we found that the amplified Gibson Assembly product was the expected size (1,110 base pairs) with very small amounts of secondary products **(Figure 2c)**.

To generate even longer fragments, we performed a second Gibson Assembly iteration **(Figure 2d)**. The first assembly product was PCR-amplified using primers containing a second set of complementary overhangs separate PCR reactions. As with the first assembly, the PCR products were purified, combined into a Gibson assembly reaction and amplified using primers corresponding to unique sequences present at the ends of the assembly product. We implemented this approach to generate a 4-fragment second assembly resulting in 24-fragments of each individual oligonucleotide in the final product. Using agarose gel electrophoresis, we found that the amplified second assembly product was the expected size (4,590 base pairs). However, we also observed significant quantities of shorter DNA fragments that appear to correspond to smaller assembly sizes **(Figure 2e)**. The correct fragment size was gel extracted and used as input material for ONT nanopore sequencing.

We implemented Consecutive Gibson Assembly and nanopore sequencing on three files: a picture of the space shuttle (Shuttle, 115 kilobytes), a picture of the earth viewed from the Apollo 17 mission (Apollo, 1.5 megabytes) and a picture of Leonardo da Vinci Vitruvian Man (Vitruvian, 132 kilobytes). The Vitruvian Man had an encoding allowing for homopolymers in the DNA sequence while the other two files did not. The Space Shuttle file and the Apollo file were concatenated into 24-fragment assemblies while the homopolymer file was concatenated into a 6-fragment assembly. We successfully sequenced and decoded all these files using Illumina sequencing by synthesis.

Consecutive Gibson Assembly enables concatenation of DNA amplicons with identical primer regions into larger DNA fragments. The assembly size is determined by the number of fragments with unique complementary overhangs at each end. Furthermore, since the assembly product contains unique sequences at its end, a final amplification step can target the intended assembly. Furthermore, we demonstrated that this process can be iterative: the product of the first assembly can then become the initial fragment for a second assembly.

Payload concatenation using OE-PCR

Although Consecutive Gibson Assembly can concatenate multiple oligonucleotides, we found it difficult to assemble more than six fragments in a single reaction. We observed that the relative proportion of side products increased with the number of attempted fragments in the assembly. Since all fragments in the assembly contain a conserved file ID region, it is possible that base-pairing across this region resulted in the generation of these side products. Furthermore, the iterative Gibson assembly strategy requires significant sample preparation which hinders its applicability in an end-to-end DNA storage workflow.

As an alternative, we developed an additional assembly method “Overlap-Extension Polymerase Chain Reaction” (OE-PCR) **(Supplementary Fig. 3a).** In contrast with Consecutive Gibson Assembly, this assembly strategy requires that a DNA file is first synthesized into groups where each group has a unique forward ID and a unique back ID. In the first step of PCR, overlapping sequences between each DNA group can be created by using primers containing a 5’ overhang complementary to the molecule it is joined to. All amplified DNA groups are mixed together, and DNA groups with overlapping regions can be fused together via PCR with N cycles (N

equals to the number of DNA groups). Finally, the outermost primers are used to selectively amplify the full length of multiple-fused DNA.

We selected a text file corresponding to 365 Foreign Dishes (Dishes, 32 kB), a vintage book originally published in 1908. We encoded this file into 2,042 oligonucleotides and split into 10 groups (i.e., each group has about 206 oligonucleotides, and each group has its own unique IDs). We successfully used OE-PCR to assemble 10 fragments into a long DNA fragment with no visible side products **(Supplementary Fig. 3b)**. This greatly reduced the sample preparation process compared to the Consecutive Gibson Assembly approach since no gel-purification was required. The assembly product was purified using magnetic beads and then used as input material for ONT nanopore sequencing.

Nanopore sequencing and decoding

Each file was amplified, assembled and then sequenced in separate nanopore flow cells (R9.4 chemistry). Each sequencing run generated reads for 48 hours until completion. The Earth-Apollo file was sequenced using two MinION flowcells to generate enough reads for successful decoding. Sample preparation was done accordingly to ONT instructions for sequencing using $1D^2$ chemistry, which results in higher sequencing accuracy. $1D^2$ basecalling was performed after sequencing and resulted in 15-25% $1D^2$ reads , which were available for decoding. We analyzed these reads to evaluate the throughput, quality and decoding potential of each run.

For the 1.5 MB Earth-Apollo file, two sequencing runs generated a total of 267,152 $1D^2$ reads for analysis and decoding. The size distribution of nanopore reads corresponded closely to the gel-extracted input DNA, indicating that most reads corresponded to the full assembly **(Figure 3a)**. We aligned each read to the expected payloads and we found that on average each read contributed to 18 alignments, close to the ideal maximum of 24 **(Figure 3b)**. This resulted in a 43X average sequencing coverage for 102,084 reference payloads **(Figure 3c)**. For comparison, 4.4M reads would have been necessary to achieve the same sequencing coverage without any assembly. We found a wide distribution in the quality score of all sequencing reads by analyzing the Phred quality score estimated by ONT **(Figure 3d)**. The average Phred quality score per read was 15.33, corresponding to an estimated error rate probability of 2.93%. Additionally, we estimated the overall error rate based on our alignment and we found an error rate of 6.87%. For insertions and deletions, we found no significant bias across bases. However, substitution error rates were significantly higher between purines (A to G and G to A) and pyrimidines (C to T and T to C) **(Figure 3e)**. We found a similar error distribution across all files sequenced.

We performed an equivalent sequencing analysis for the other three files **(Supplementary Fig. 4, 5 and 6)** and then we attempted decoding each file using the $1D^2$ reads as inputs to an improved DNA codec[74]. First, we performed multiple sequential alignments of the front and back file ID in each read and then selected the sequence in between, which should correspond to the payload sequence. Then, the payload sequences were clustered based on similarity[76], followed by determining a consensus sequence for each cluster, and error-correction across consensus sequences to recover the original file.

We employed a new consensus algorithm that improves upon the algorithm from our previous work[74]. In the new algorithm consensus sequence is recovered via a process where pointers for payload sequences are maintained and moved from left to right, and at every step of the process the next symbol of the sequence is estimated via a plurality vote. For the payload sequences that agree with plurality, the pointer is moved to the right by 1. But for the sequences that do not agree with plurality, the algorithm classifies whether the reason for the disagreement is a single deletion, an insertion, or a substitution. This is done by looking at the context around the symbol under consideration. Once this is estimated, the pointers are then moved to the right accordingly. The key difference between our new algorithm and our previous implementation[74] is that in cases when disagreements cannot be classified, we do not drop respective payload sequences from consideration, but attempt to bring them back at later stages. This allows us to successfully decode from notably lower coverages because more information about the sequences was used overall.

We successfully decoded both non-homopolymer files (Apollo and SpaceShuttle files) concatenated using a 24-fragment sequential Gibson assembly. We were also successful decoding the Dishes assembled using the 10-fragment OE-PCR strategy. By subsampling the number of reads used for decoding, we found that minimum coverages of 23x, 22x and 27x were sufficient to decode the Apollo, Shuttle and Dishes file, respectively. Nevertheless, despite having a coverage of 166x, we were unable to decode the homopolymer Vitruvian Man file. Even though the overall error rate was similar to that of the non-homopolymer files, we found that nanopore sequencing tends to underestimate the length of homopolymer runs, leading to correlated deletion errors across payload sequences that cannot be corrected by the consensus algorithm. 24.3% of all original sequences corresponding to the Vitruvian Man file contained a homopolymer run of length 5 or more. However only 57% of reads corresponding to those sequences had a run of such length.

We found that increasing the input DNA size resulted in significant improvements in the sequencing throughput. To evaluate this improvement, we quantify sequencing throughput across 5 sequencing runs of equivalent quality. When comparing sequencing runs based on the input DNA fragment size, we found a modest reduction in the number of reads as the fragment size increased **(Figure 5a)**. Despite this modest reduction in total reads, we found a 7-fold increase in sequencing throughput across the fragment size range we evaluated (between 600 and 4700 base pairs) **(Figure 5b)**. This increase in sequencing yield was necessary to enable the successful decoding of a 1.5 MB file while sequencing using only two MinION flowcells.

Discussion

We demonstrated megabyte-scale decoding of digital files encoded in synthetic DNA using a portable nanopore sequencer. In combination with our robust encoding scheme, our assembly framework enabled sequencing and decoding of a 1.5-megabyte file using only two ONT MinION flowcells. We also demonstrated that this method is compatible with a PCR-based random-access strategy for DNA storage. Despite introducing a modest increase in preparation work, our assembly strategy can increase the effective coverage of nanopore sequencing in other applications that require sequencing of short DNA amplicons.

Both assembly methods described in this paper enabled concatenation of synthetic oligonucleotides into larger DNA fragments with unique sequences and their ends. This feature enabled selective amplification of the final assembly product. Iterative Gibson assembly is compatible with generalizable file architecture where each oligonucleotide in a given file is synthesized with the same file ID at each end. Therefore, the synthesis of a given file is agnostic to its later assembly step. In contrast, OE-PCR requires each file to be synthesized in groups containing unique overhang sequences based on the structure of the final assembly. Despite limiting the applicability of this assembly to files encoded with the necessary overhang sequence or to groups of files to be sequenced together, this strategy resulted in better assembly results and a simpler sample preparation workflow. Other strategies such as Golden Gate assembly or rolling circle amplification are also potential alternatives to enable longer sequencing reads from short oligonucleotides used for DNA storage[85].

The future of DNA storage will likely depend on technologies that enable cost-efficient and fast reading and writing of digital information using synthetic DNA. While sequencing by synthesis remains widely used for its low error rate and reliability, nanopore sequencing is being rapidly adopted in applications that require long DNA sequencing reads[86], high degree of portability[79], or real-time sequencing[80]. Given its portability and low-cost, a nanopore-based read-head for DNA storage has the potential to democratize this data storage technology outside of the research arena. Furthermore, recent advances in solid-state nanopore technology promise to further improve the cost and scale of nanopore sequencing.

Methods

Assembly sequence design. Overhang sequence design was performed using a nucleic acid thermodynamic simulation software NUPACK v3.0.5. Secondary structure analysis was performed using 'pairs' function from NUPACK at 65°C with each overhang primer. Resulting secondary structure predictions are ranked by probability of self-pairing and primers with low secondary structure are selected. Next, we evaluated cross-talk among all possible pairs of overhang primers using 'pairs' function from NUPACK at 25°C, at 10nM of each overhang primer. Next, pairs with an equilibrium binding of 1nM or above were sequentially eliminated until finding a subset of primers with minimal cross-talk.

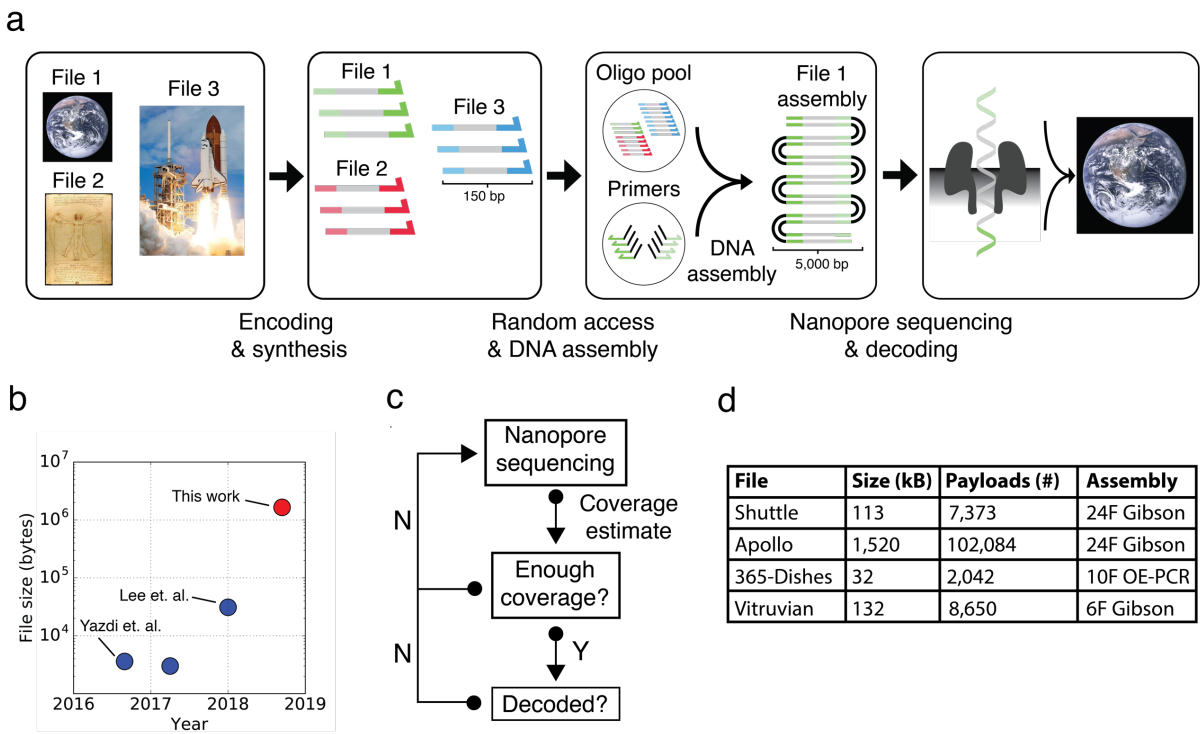
PCR amplification. Each pool of synthetic DNA was rehydrated in 1X TE buffer. PCR based random access and overhang addition of a given file from a DNA pool was performed as following: mix 10 ng of ssDNA pool, 2 μM of forward and reverse primer, 25 μL of 2x Kapa HiFi enzyme mix, 1.25 μL of EvaGreen dye 20X and 20 μL of molecular grade water. All PCR reactions were carried out on a quantitative PCR instrument where amplification was stopped before the reaction reached a plateau phase. All primers were ordered from Integrated DNA Technologies. Assembly products were amplified using an equivalent protocol was diluted 10X.

Gibson assembly. PCR reactions were cleaned-up using KAPA Pure beads for 1st assembly Gibson reactions. PCR reactions were gel-purified using a 1% agarose gel and NEB Monarch DNA Gel Extraction kit for 2nd assembly Gibson reactions to select out any side products from the 1st assembly. Gibson reactions were carried out using NEB Gibson Assembly Master Mix where the fragments were combined at 200 ng of DNA and incubated at 50°C for 1 hour.

OE-PCR. First, each individual fragment was amplified with the following PCR protocol. In a 20 uL reaction, 1 uL of 10 nM single-stranded DNA pool was mixed with 1 uL of 10 uM of the forward primer and 1 uL of 10 uM of the reverse primer, 10 uL of 2X KAPA HIFI enzyme mix, 1 uL of 20X EVA Green, and 6 uL of molecular biograde water. The reaction followed a thermal protocol: (1) 95˚C for 3 min, (2) 98˚C for 20 sec, (3) 62˚C for 20 sec, (4) 72˚C for 15 sec. The PCR reaction was performed on a quantitative PCR (qPCR) instrument and stopped before reaction reached a plateau phase. The length of amplified product was confirmed using a Qiaxcel fragment analyzer, and the sample concentration was measured by Qubit 3.0 fluorometer.

Next, all amplified fragments were mixed with equal molar ratio (1.5 ng for each fragment) together with 10 uL of 2X KAPA HIFI enzyme mix in a total of 20 uL reaction. The reaction followed a standard PCR thermal protocol for N cycles (N is the total number of fragments). After that, 1 uL of the amplified product was mixed with 1 uL of 10 uM of the forward primer, 1 uL of 10 uM of the reverse primer, 10 uL of 2X KAPA HIFI enzyme mix, and 7 uL of molecular biograde water. The mixture followed the same thermal protocol as above and stopped before reaching a plateau phase. The final product was verified using a Qiaxcel fragment analyzer.

ONT MinION sequencing. Final assembly products were amplified in multiple PCR reactions to generate 1 μg of DNA after gel purification. Purification was carried out with a NEB Monarch DNA Gel Extraction kit, followed by column clean-up with Sigma-Aldrich Sephadex G-25 to remove excess salt. Finally, KAPA Pure beads were used to concentrate the final DNA library to 45μL for nanopore sequencing. DNA purity and concentration were verified using Thermo Fisher Qubit and NanoDrop instruments. Sequencing sample preparation was carried using $1D^2$ ligation kit LSK-308 and MinION flowcells R9.4. Each file assembly was sequenced in a separate flowcell and sequencing was carried out for 48 hours.



File	Size (kB)	Payloads (#)	Assembly
Shuttle	113	7,373	24F Gibson
Apollo	1,520	102,084	24F Gibson
365-Dishes	32	2,042	10F OE-PCR
Vitruvian	132	8,650	6F Gibson

Figure 1 | Overview of the DNA data storage workflow. a, The encoding process starts with mapping multiple digital files into 150-nucleotide DNA sequences, which are sent for synthesis. Each file has unique sequence addresses at the 5' and 3' end of each oligonucleotide for random access retrieval. Using PCR primers containing complementary overhang sequences, a specific file can be amplified and concatenated into long double-stranded DNA molecules suited for ONT Nanopore sequencing. Upon sequencing, a subset of reads with high accuracy are used to decode the selected file. **b,** Sequence-until diagram. Nanopore sequencing enables real-time coverage estimation for decoding of digital files store in DNA. This enables the user to generate reads until coverage is enough for successful decoding. Upon decoding, a different file can be sequenced in the same flowcell or the sequencing run can be stopped and resumed later on. **c,** 4 different files encoded in DNA were amplified, assembled and sequenced using ONT MinION platform. We implemented two different assembly strategies: OE-PCR and Gibson assembly. **d,** Our assembly strategy enabled successful decoding of 1.67 MB of digital information stored in DNA using nanopore sequencing. In comparison with previous work with DNA storage and nanopore sequencing, this represents a 2-order of magnitude improvement in information decoded.

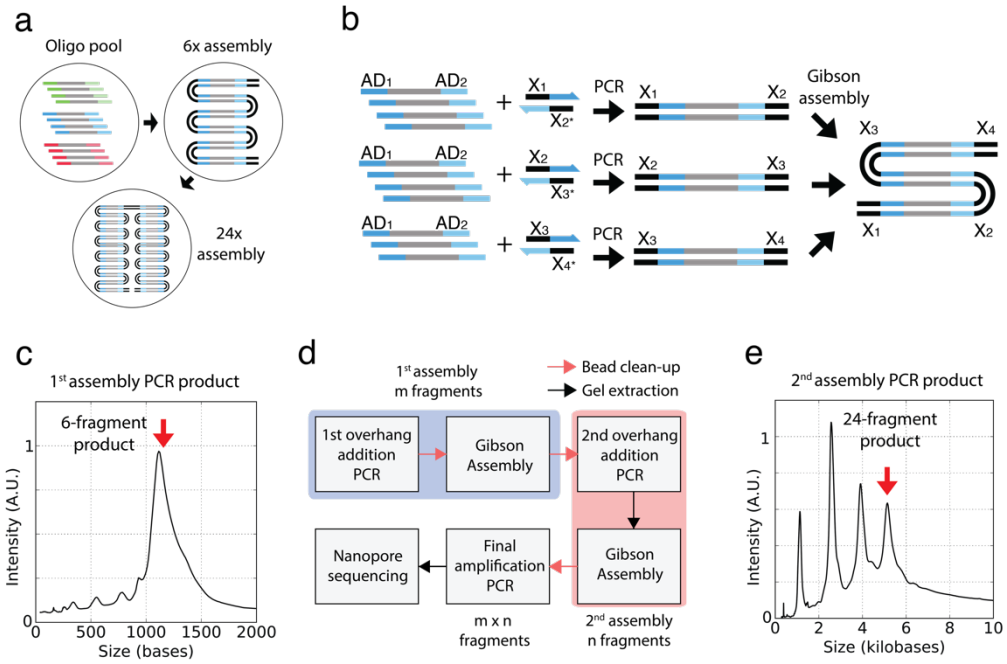
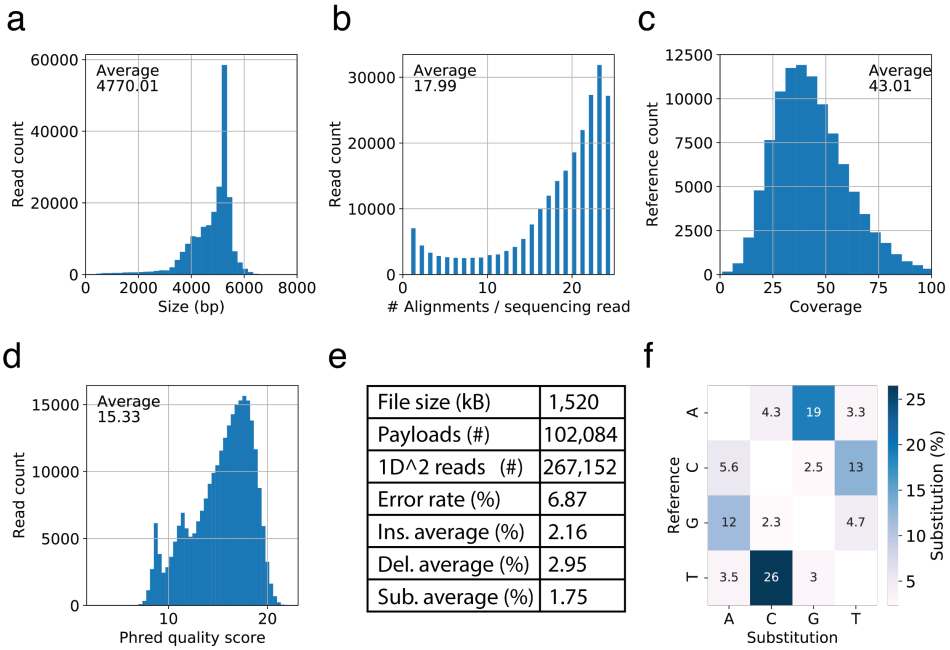


Figure 2 | Random access and sequential Gibson strategy for DNA storage. a, Our sequential Gibson assembly strategy starts enables random access of particular file from a pool of DNA oligonucleotides. First, the selected file is assembled into a 6-fragment assembly following by a second 4-fragment assembly. The resulting DNA product corresponds to a concatenation of 24 payloads of the selected file. **b,** First, a given file is PCR amplified using primers specific to the its address (AD_1 & AD_2) and containing overlapping overhang sequences (X_n). For a three-fragment assembly, three separate PCR amplification reactions are carried out each containing primer pairs with different overhangs based on a sequential assembly design (X_1 & X_2*, X_2 & X_3* and X_3 & X_4*). Upon amplification, products are purified and combined into a Gibson assembly reaction where the resulting assembled product is generated. **c,** Gel electrophoresis size distribution corresponding to a PCR amplification of a 6-fragment 1st Gibson assembly. Expected band size was 1,110 bp. **d,** Schematic representation of a two-fold assembly process. The product of the first assembly (m fragments) is amplified with distinct overhangs and the products are mixed in a Gibson assembly reaction (n fragments) to create a final assembly product (m x n fragments). **e,** Gel electrophoresis size distribution corresponding to a PCR amplification of a 24-fragment 2nd Gibson assembly. Expected band size was 4,590 bp.



File size (kB)	1,520
Payloads (#)	102,084
1D^2 reads (#)	267,152
Error rate (%)	6.87
Ins. average (%)	2.16
Del. average (%)	2.95
Sub. average (%)	1.75

Figure 3 | Nanopore sequencing analysis for 1.5 MB Apollo file. Two MinION flowcells generated 267,152 $1D^2$ reads of the 24-fragment Gibson assembly of the Apollo file. **a,** Base pair size of sequencing reads matches closely with the assembly size of 4,590 bp. **b,** We aligned each reference payload sequence to the sequencing reads. Each sequencing read resulting in an average of 17.99 alignments to different payloads. Ideally, each read should have 24 alignments. **c,** We found an average sequencing coverage of 43X per payload. **d,** We estimated raw sequencing quality by analyzing the average Phred quality score in each read. **e,** Based on the reads that aligned to payloads, we calculated the average percent error for each base for insertions, deletion and substitutions. **f,** Substitution comparison across different bases revealed strong bias in between purines and pyrimidines.

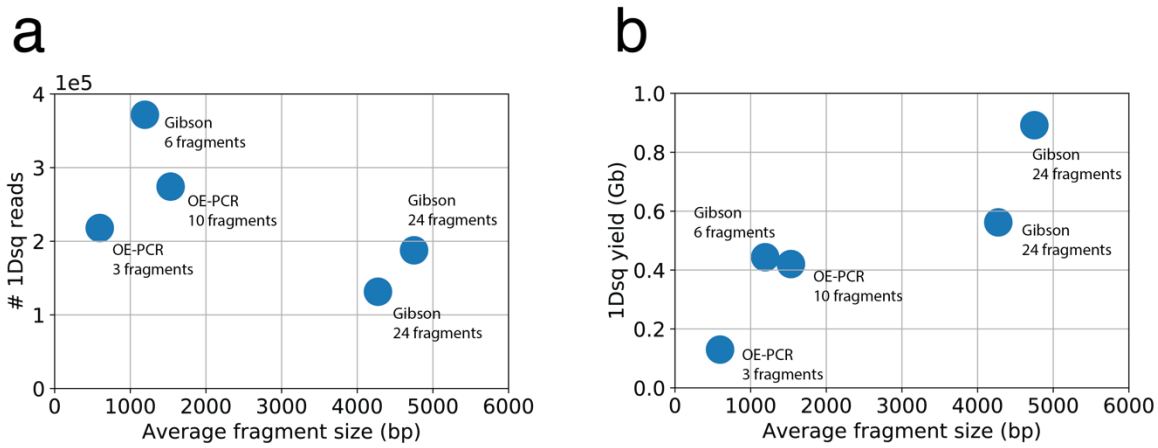
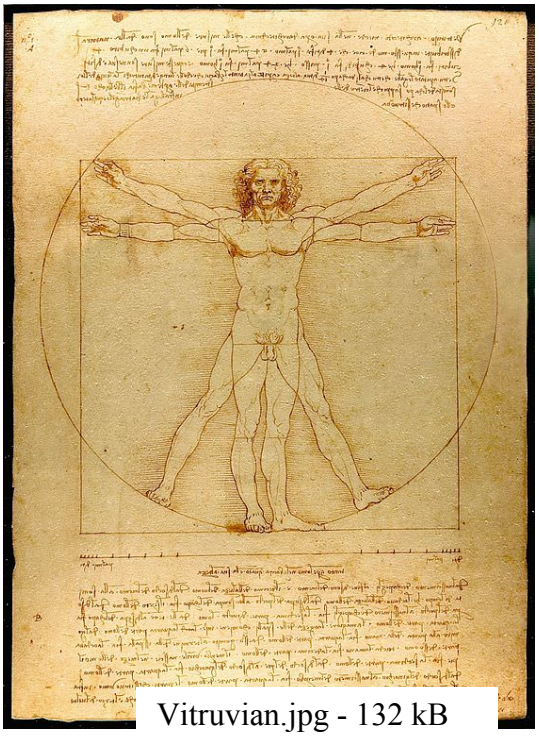


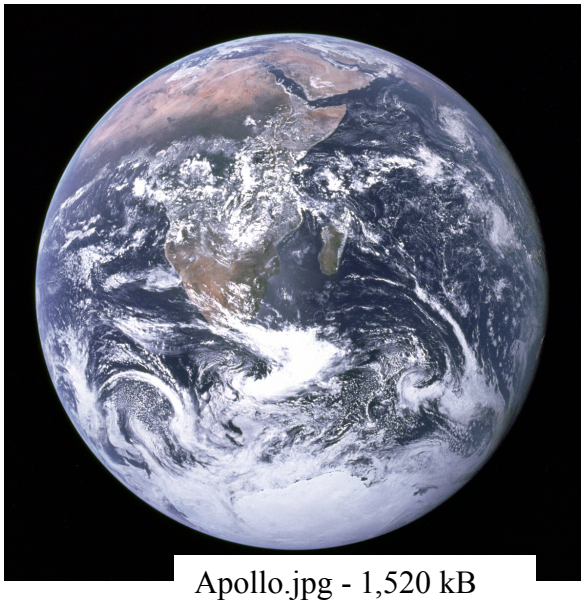
Figure 5 | Throughput comparison among sequencing runs with different DNA fragment sizes. a, We compared 5 runs of equivalent sequencing quality to understand how input DNA size affects sequencing throughput. We found a modest decrease in the number of $1D^2$ reads as the input DNA size increased. **b,** However, we found that the overall $1D^2$ yield in bases increased by a factor of 7-fold between a 600bp and 4700bp input DNA size.

Supplementary Figures

Shuttle.jpg - 113 kB



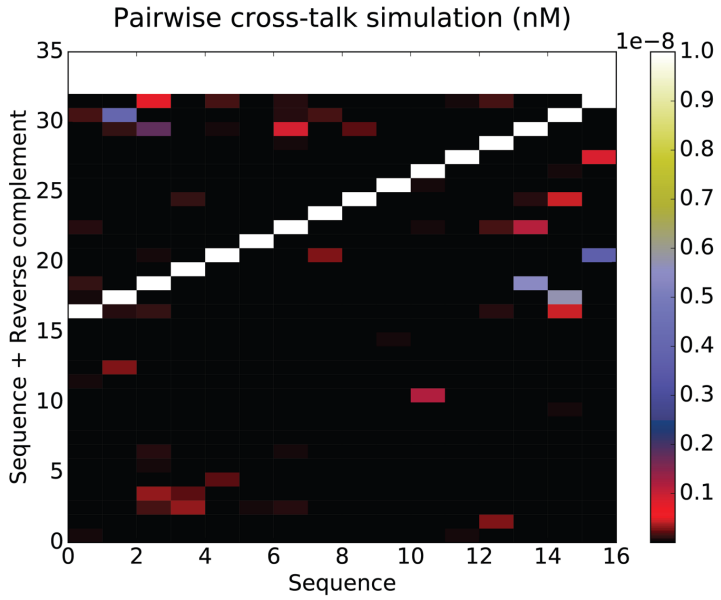
Vitruvian.jpg - 132 kB



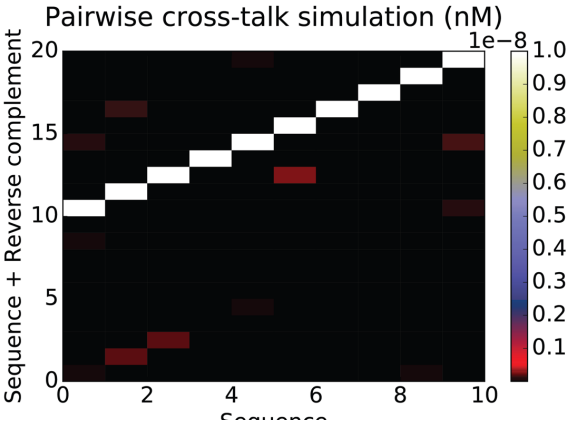
Apollo.jpg - 1,520 kB

Supplementary figure 1 | Three pictures in JPEG format encoded in DNA and sequenced using ONT Nanopore sequencing.

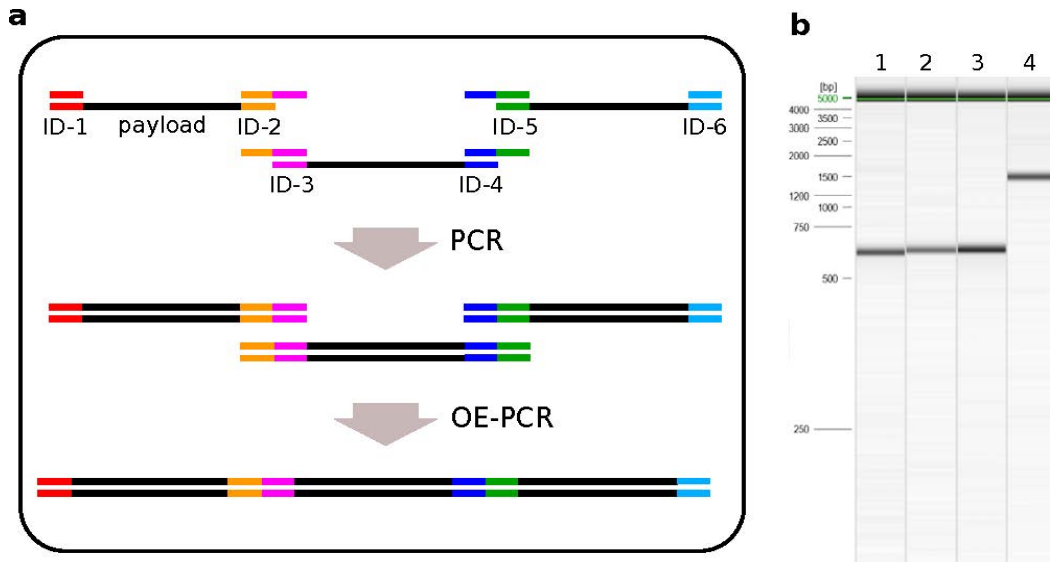
a



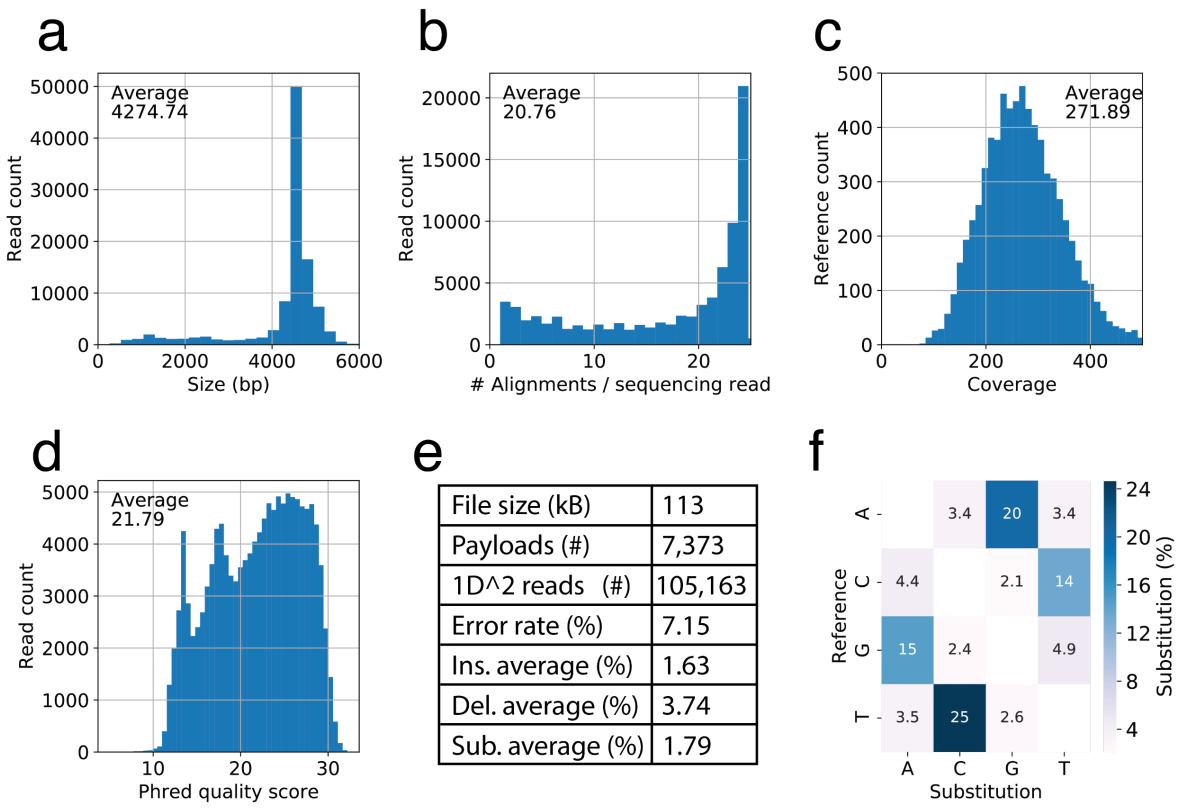
b



Supplementary figure 2 | Cross-talk simulation for overhang sequences. a, We used a nucleic acid thermodynamic simulation software package to estimate bound equilibrium concentrations for every pair with a starting concentration of 10nM at 25°C. The diagonal line of white data points corresponds to binding between each sequence and its reverse complement. **b,** Then, we implemented an algorithm for sequentially removing overhang sequences with unintended binding. Threshold for unintended interaction was varied based on the number of final fragments necessary for assembly (threshold was set to 1nM for this example).

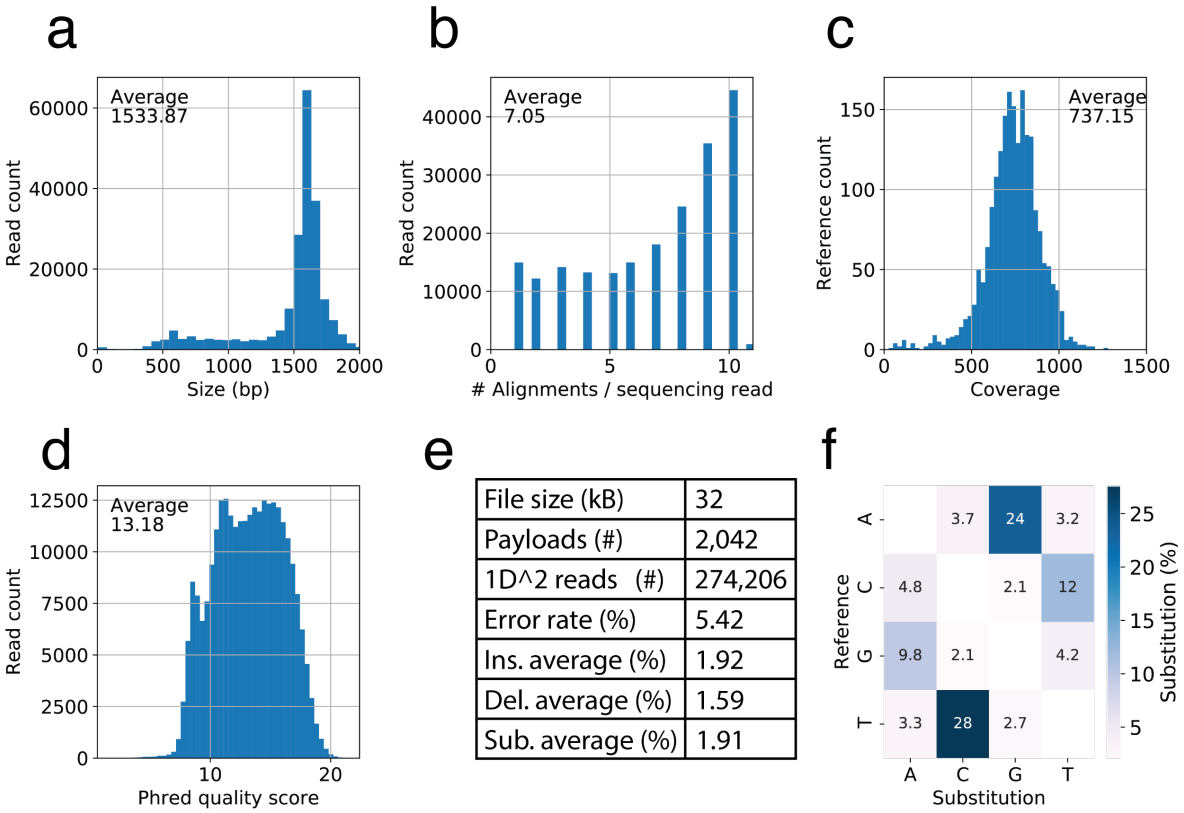


Supplementary figure 3 | Random access and OE-PCR strategy for DNA storage a, First, all the payloads associated with a particular file are split into multiple groups. Each group is given unique file IDs. To retrieve a particular file, each group is amplified with primers containing overhangs overlapping with each adjacent group. Subsequently, all the groups can be combined and amplified using primers corresponding to the end (e.g. ID-1 and ID-6) to form the assembly product using overlap-extension PCR. **b,** Bioanalyzer traces corresponding to the assembly of group 1-2-3 (lane 1), group 4-5-6 (lane 2), group 7-8-9 (lane 3) and group 1 through 10 (lane 4). We found no observable by-products in the assembly of up to 10 groups demonstrating the scalability of this approach.



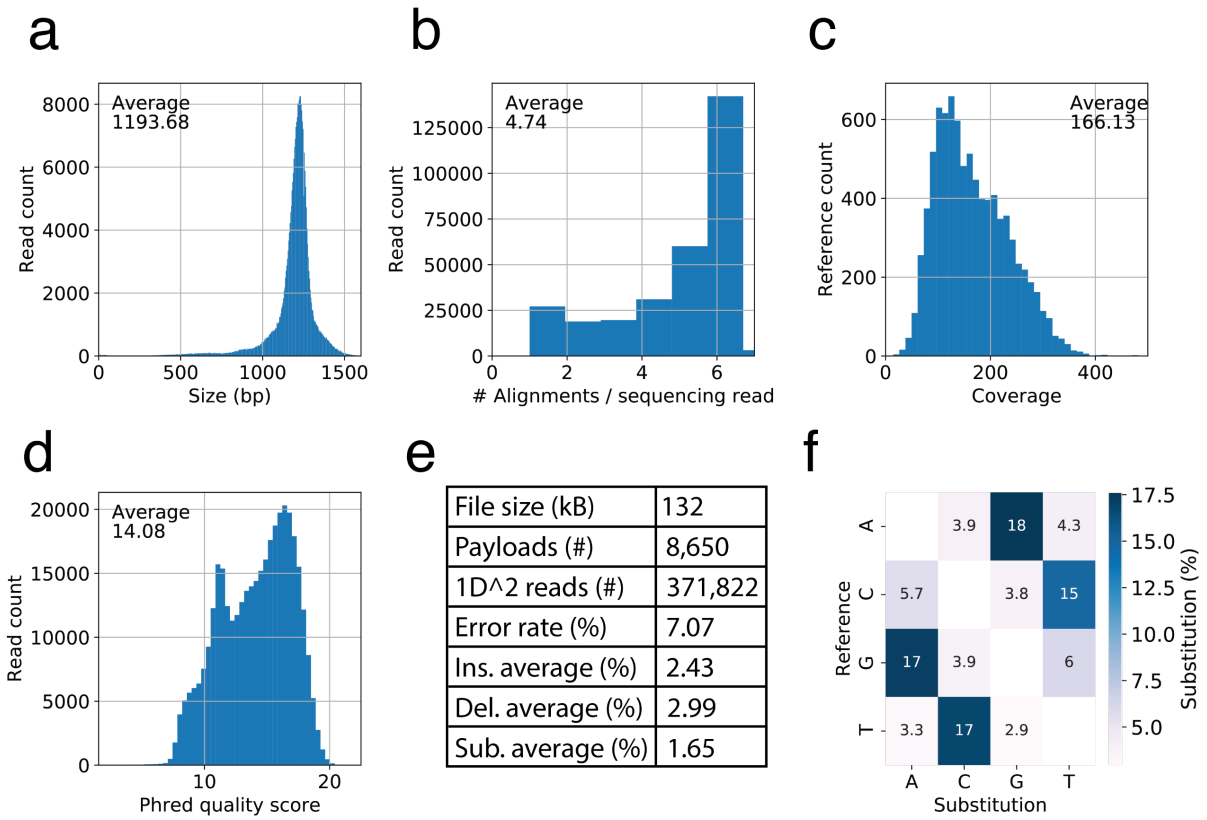
File size (kB)	113
Payloads (#)	7,373
1D^2 reads (#)	105,163
Error rate (%)	7.15
Ins. average (%)	1.63
Del. average (%)	3.74
Sub. average (%)	1.79

Supplementary figure 4 | Sequencing analysis for 113 kB Shuttle file. One MinION flowcells generated 105,163 $1D^2$ reads of the 24-fragment Gibson assembly **a,** Base pair size of sequencing reads matches closely with the assembly size of 4,590 bp. **b,** We aligned each reference payload sequence to the sequencing reads. Each sequencing read resulting in an average of 20.76 alignments to different payloads. Ideally, each read should have 24 alignments. **c,** We found an average sequencing coverage of 271x per payload. **d,** We estimated raw sequencing quality by analyzing the average Phred quality score in each read. **e,** Based on the reads that aligned to payloads, we calculated the average percent error for each base for insertions, deletion and substitutions **(f)** Substitution comparison across different bases revealed strong bias in between purines and pyrimidines.



File size (kB)	32
Payloads (#)	2,042
1D^2 reads (#)	274,206
Error rate (%)	5.42
Ins. average (%)	1.92
Del. average (%)	1.59
Sub. average (%)	1.91

Supplementary figure 5 | Sequencing analysis for 32 kB Dishes file. One MinION flowcells generated 274,206 $1D^2$ reads of the 10-fragment OE-PCR assembly **a,** Base pair size of sequencing reads matches closely with the assembly size of 1,500 bp. **b,** We aligned each reference payload sequence to the sequencing reads. Each sequencing read resulting in an average of 7.05 alignments to different payloads. Ideally, each read should have 10 alignments. **c,** We found an average sequencing coverage of 737x per payload. **d,** We estimated raw sequencing quality by analyzing the average Phred quality score in each read. **e,** Based on the reads that aligned to payloads, we calculated the average percent error for each base for insertions, deletion and substitutions **(f)** Substitution comparison across different bases revealed strong bias in between purines and pyrimidines.



File size (kB)	132
Payloads (#)	8,650
1D^2 reads (#)	371,822
Error rate (%)	7.07
Ins. average (%)	2.43
Del. average (%)	2.99
Sub. average (%)	1.65

Supplementary figure 6 | Sequencing analysis for 132 kB Vitruvian file. One MinION flowcells generated 371,822 $1D^2$ reads of the 6-fragment Gibson assembly **a,** Base pair size of sequencing reads matches closely with the assembly size of 1,110 bp. **b,** We aligned each reference payload sequence to the sequencing reads. Each sequencing read resulting in an average of 4.74 alignments to different payloads. Ideally, each read should have 6 alignments. **c,** We found an average sequencing coverage of 166x per payload. **d,** We estimated raw sequencing quality by analyzing the average Phred quality score in each read. **e,** Based on the reads that aligned to payloads, we calculated the average percent error for each base for insertions, deletion and substitutions **(f)** Substitution comparison across different bases revealed strong bias in between purines and pyrimidines.

References

1. Seeman, N.C. Nucleic acid junctions and lattices. *Journal of theoretical biology* **99** (1982).
2. Adleman, L.M. Molecular computation of solutions to combinatorial problems. *Nature* **266**, 1021-1024 (1994).
3. Seeman, N.C. DNA in a material world. *Nature* **421**, 427-431 (2003).
4. Douglas, S.M., Dietz, H., Liedl, T., Högberg, B. & Nature, G.-F. Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* (2009).
5. Qian, L. & Winfree, E. Scaling Up Digital Circuit Computation with DNA Strand Displacement Cascades. *Science* **332**, 1196-1201 (2011).
6. Qian, L., Winfree, E. & Bruck, J. Neural network computation with DNA strand displacement cascades. *Nature* **475**, 368-372 (2011).
7. al., D.G.e. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52-56 (2010).
8. 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).
9. Chen, S.X., Zhang, D.Y. & Seelig, G. Conditionally fluorescent molecular probes for detecting single base changes in double-stranded DNA. *Nature chemistry* **5**, 782-789 (2013).
10. Lim, H.-W. et al. Biomolecular computation with molecular beacons for quantitative analysis of target nucleic acids. *Biosystems* **111**, 11-17 (2013).
11. Mills, A.P. Gene expression profiling diagnosis through DNA molecular computation. *Trends Biotechnol* **20**, 137-140 (2002).
12. Bornholt, J., Lopez, R., Carmean, D.M. & Sigops …, C.-L. A DNA-based archival storage system. *ACM SIGOPS ...* (2016).
13. Church, G.M., Gao, Y. & Kosuri, S. Next-Generation Digital Information Storage in DNA. *Science* (2012).
14. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science (New York, N.Y.)* **355**, 950-954 (2017).
15. Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77 (2013).
16. Organick, L., Ang, S.D., Chen, Y.J., Lopez, R. & bioRxiv, Y.-S. Scaling up DNA data storage and random access retrieval. *bioRxiv* (2017).
17. Yazdi, H.S.M., Gabrys, R. & Milenkovic, O. Portable and error-free DNA-based data storage. *Scientific Reports* **7** (2017).
18. Bornholt, J. et al. A DNA-Based Archival Storage System. *A DNA-Based Archival Storage System*, 637-649 (2016).
19. Vargas, J.D. & Lima, J.A.C. Coronary artery disease: a gene-expression score to predict obstructive CAD. *Nat. Rev. Cardiol*, 243-244 (2013).
20. 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).
21. 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).

22. Myers, M.B. Targeted therapies with companion diagnostics in the management of breast cancer: current perspectives. *Pharmgenomics Pers Med*, 7-16 (2016).
23. Rotunno, M. et al. A Gene Expression Signature from Peripheral Whole Blood for Stage I Lung Adenocarcinoma. *Cancer Prev Res* **4**, 1599-1608 (2011).
24. Lunnon, K., Sattlecker, M. & Furney, S.J. A blood gene expression marker of early Alzheimer's disease. *J Alzheimers Dis.* **33**, 737-753 (2013).
25. Koscielny, S. Why most gene expression signatures of tumors have not been useful in the clinic. *Sci. Transl. Med.* **2** (2010).
26. 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).
27. 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).
28. Tsalik, E.L. et al. Host gene expression classifiers diagnose acute respiratory illness etiology. *Sci. Transl. Med.* **8** (2016).
29. 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).
30. Yuan, T., Huang, X., Woodcock, M., Du, M. & Dittmar, R. Plasma extracellular RNA profiles in healthy and cancer patients. *Sci. Rep.* **6** (2016).
31. 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).
32. Zhang, L. et al. Salivary Transcriptomic Biomarkers for Detection of Resectable Pancreatic Cancer. *Gastroenterology* **138**, 949 (2009).
33. Zhang, L. et al. Development of transcriptomic biomarker signature in human saliva to detect lung cancer. *Cell Mol Life Sci* **69**, 3341-3350 (2012).
34. 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).
35. 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).
36. 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).
37. 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).
38. 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).
39. 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).
40. Pardee, K., Green, A.A., Ferrante, T. & Cameron, D.E. Paper-based synthetic gene networks. *Cell* **159**, 940-954 (2014).
41. Pardee, K. et al. Rapid, low-cost detection of Zika virus using programmable biomolecular components. *Cell* **165**, 1255-1266 (2016).

42. Jung, C. & Ellington, A.D. Diagnostic applications of nucleic acid circuits. *Acc. Chem. Res* **47**, 1825-1835 (2014).
43. Gootenberg, J.S. et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356**, 438-442 (2017).
44. Seelig, G., Soloveichik, D., Zhang, D. & Winfree, E. Enzyme-Free Nucleic Acid Logic Circuits. *Science* **314**, 1585-1588 (2006).
45. Chen, Y.-J. et al. Programmable chemical controllers made from DNA. *Nat. Nanotechnol.* **8**, 755-762 (2013).
46. Genot, A.J., Fujii, T. & Rondelez, Y. Scaling down DNA circuits with competitive neural networks. *J. R. Soc. Interface* **10**, 20130212 (2013).
47. Franco, E. et al. Timing molecular motion and production with a synthetic transcriptional clock. *Proc Natl Acad Sci U S A* **108** (2011).
48. Green, A.A. et al. Complex cellular logic computation using ribocomputing devices. *Nature* **548**, 117-121 (2017).
49. 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).
50. Abusamra, H. A comparative study of feature selection and classification methods for gene expression data of glioma. *Procedia Comput Sci* **23**, 5-14 (2013).
51. 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).
52. 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).
53. Zhang, D. & Seelig, G. DNA-Based Fixed Gain Amplifiers and Linear Classifier Circuits. *LNCS* **16**, 176-186 (2010).
54. Zhang, D. Cooperative Hybridization of Oligonucleotides. *J. Am. Chem. Soc* **133**, 1077-1086 (2011).
55. 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).
56. 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).
57. 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).
58. 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).
59. Notomi, T., Okayama, H. & Masubuchi, H. Loop-mediated isothermal amplification of DNA. *Nucleic acids Res* **28**, e63 (2000).
60. 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).
61. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C. & Lossos, I.S. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* (2000).

62. Chen, J.J. & Chen, C.H. Microarray gene expression. *Encyclopedia of Biopharmaceutical …* (2003).
63. Burchill, S.A., Perebolte, L., Johnston, C. & Top, B. Comparison of the RNA-amplification based methods RT–PCR and NASBA for the detection of circulating tumour cells. *British journal of …* (2002).
64. Deng, R., Zhang, K., Sun, Y., Ren, X. & Li, J. Highly specific imaging of mRNA in single cells by target RNA-initiated rolling circle amplification. *Chemical Science* **8**, 3668-3675 (2017).
65. Stougaard, M., Juul, S., Andersen, F.F. & Knudsen, B.R. Strategies for highly sensitive biomarker detection by Rolling Circle Amplification of signals from nucleic acid composed sensors. *Integrative Biology* **3**, 982 (2011).
66. Takahashi, H., Matsumoto, A., Sugiyama, S. & Kobori, T. Direct detection of green fluorescent protein messenger RNA expressed in Escherichia coli by rolling circle amplification. *Analytical biochemistry* (2010).
67. Larman, B.H. et al. Sensitive, multiplex and direct quantification of RNA sequences using a modified RASL assay. *Nucleic acids research* **42**, 9146-9157 (2014).
68. Ståhlberg, A. et al. Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. *Nucleic acids research* (2016).
69. Li, H., Qiu, J. & Fu, X.-D.D. RASL-seq for massively parallel and quantitative analysis of gene expression. *Current protocols in molecular biology* **Chapter 4**, 9 (2012).
70. Barretina, J., Caponigro, G., Stransky, N. & Venkatesan, K. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* (2012).
71. Zhirnov, V., Zadegan, R.M., Sandhu, G.S. & materials, C.-G.M. Nucleic acid memory. *Nature materials* (2016).
72. Alharthi, A., Krotov, V. & Bowman, M. Addressing barriers to big data. *Business Horizons* **60**, 285-292 (2017).
73. Yazdi, S., Yuan, Y., Ma, J., Zhao, H. & reports, M.-O. A rewritable, random-access DNA-based storage system. *Scientific reports* (2015).
74. Organick, L. et al. Random access in large-scale DNA data storage. *Nature Biotechnology* **36**, 242 (2018).
75. Grass, R.N., Heckel, R. & Chemie …, P.-M. Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. *Angewandte Chemie …* (2015).
76. Rashtchian, C. et al. in Clustering Billions of Reads for DNA Data Storage 3360-3371 (2017).
77. Puddu, M., Paunescu, D., Stark, W.J. & nano, G.-R.N. Magnetically recoverable, thermostable, hydrophobic DNA/silica encapsulates and their application as invisible oil tags. *ACS nano* (2014).
78. Shendure, J. et al. DNA sequencing at 40: past, present and future. *Nature* **550**, 345-353 (2017).
79. Castro-Wallace, S.L., Chiu, C.Y., John, K.K. & reports, S.-S.E. Nanopore DNA sequencing and genome assembly on the International Space Station. *Scientific reports* (2017).
80. Hoenen, T., Groseth, A. & infectious …, R.-K. Nanopore sequencing as a rapidly deployable Ebola outbreak tool. *Emerging infectious …* (2016).

81. Johnson, S.S., Zaikova, E., Goerlitz, D.S. & of …, B.-Y. Real-time DNA sequencing in the Antarctic dry valleys using the Oxford Nanopore sequencer. *Journal of* ... (2017).
82. Laver, T., Harrison, J., O'Neill, P.A. & detection …, M.-K. Assessing the performance of the oxford nanopore technologies minion. *Biomolecular detection* ... (2015).
83. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nature methods* **13**, 751-754 (2016).
84. Zadeh, J.N., Steenberg, C.D. & Bois, J.S. NUPACK: analysis and design of nucleic acid systems. *NUPACK: analysis and design of nucleic acid systems* (2011).
85. Li, C. et al. INC-Seq: accurate single molecule reads using nanopore sequencing. *GigaScience* **5**, 34 (2016).
86. Jain, M., Koren, S., Miga, K.H., Quick, J. & Nature …, R.-A.C. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature* ... (2018).

Conclusion

In this body of work, we described the development of two independent platforms where we employed DNA nanotechnology as a foundational tool to solve important engineering challenges. In the past, DNA-based technologies, such as polymerase chain reaction or DNA sequencing, were fundamental for the advancement of human health and biological sciences. We expect that the work described here lays down the foundation for future avenues of growth in the field. Importantly, the applications described here are highly interdisciplinary. They required understanding of principles in disciplines as diverse as molecular biology, genomics, biophysics, chemistry, computer architecture and machine learning. It is important for researchers in the field to look beyond DNA nanotechnology into other disciplines where potential applications exist. We have only scratched the surface when it comes to the potential of DNA nanotechnology, and biological engineering at large, to transform existing technological paradigms and improve the human condition.

Acknowledgements

This work would have not been possible without the support of many generous, patient and intelligent people. In academia, I have found an incredible community that fosters critical thinking, skepticism and collaboration. Despite the many challenges that plague our community, I finish graduate school with the conviction that scientists and academics must continue expanding their field of influence in society. Given the current state of affairs, we have the responsibility to translate the scientific method outside science in order to improve the world around us.

I found the culture in the Seelig Lab to be particularly enriching and nurturing. I was fortunate to be given an unusual amount of freedom to pursue my own ideas and to make my own mistakes. The entire journey transformed me into a better scientist, citizen and human being. Yuan Chen, Gourab Chatterjee, Benjamin Groves and Alexander Rosenberg were senior graduate students and postdocs that gave me a significant amount of their time for mentoring and teaching. I am incredibly grateful to them. In graduate school, I also found an incredible group of friends that became an emotional support throughout this journey. Sundipta Rao, Sifang Chen, Sumit Mukherjee, Arjun Khakhar and Chuhern Hwang were the human foundation that allowed me to pursue my scientific work.

I am incredibly grateful to my advisor Dr. Georg Seelig for the opportunity to work in his group. Georg is constantly listening and thinking about new ideas. His openness is matched by healthy skepticism that generally motivates his students to test their ideas. This led to a lab culture of self-motivating individuals willing to try new things without being paralyzed by a fear of failure. I also want to thank Dr. Luis Ceze and Dr. Karin Strauss for letting me join them in their journey from Computer Science to the messy world of wet lab experiments. Their ability to transition, almost seamlessly, from discipline to disciple is both humbling and inspiring. Their entrepreneurial approach to science has proven to be successful at connecting dots that others would have missed. I would also like to thank Dr. James Carothers, Dr. Suzie Pun, Dr. Barry Lutz, Dr. Wendy Thomas and Dr. Herbert Sauro for being part of my examination committees

throughout graduate school. Their criticism and feedback were necessary to become a better scientist.

Finally, I would like to thank my mom. The odds were against us from the beginning. I grew up in a collapsing society, without a father and with very limited financial means. Yet, my mom proved to be the most invaluable asset I have ever had. Her core belief was that education was the best possible investment of time and money. This required sacrificing short-term rewards for long-term potential. In practical terms, this meant that close to half of our household income was dedicated to my education. This financial and emotional investment kept pushing me to persevere and prove myself worthy of such effort. It was a shot in a million. Yet, as I finish writing my PhD dissertation and I talk to her over the phone, she still asks me: "What are you going to study next?".