A Deep Learning Model for Predicting NGS Sequencing Depth from DNA Sequence

Jinny Zhang

Rice University

Boyan Yordanov

Microsoft Research (United Kingdom)

Alexander Gaunt

Microsoft Research

Michael Wang

Rice University <https://orcid.org/0000-0001-7009-6958>

Peng Dai

Rice University

Yuan-Jyue Chen

Microsoft (United States) <https://orcid.org/0000-0003-3826-3677>

Kerou Zhang

Rice University

John Fang

Rice University <https://orcid.org/0000-0003-2372-0839>

Neil Dalchau

Microsoft Research (United Kingdom) <https://orcid.org/0000-0002-4872-6914>

Jiaming Li

Rice University

Andrew Phillips

Microsoft Research (United Kingdom) <https://orcid.org/0000-0001-9725-1073>

David Zhang (✉ dyz1@rice.edu)

Rice University

Article

Keywords: targeted high-throughput DNA sequencing, deep learning model, DLM, NGS sequencing depth, DNA probe sequence

Posted Date: July 9th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-37670/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Version of Record: A version of this preprint was published at Nature Communications on July 19th, 2021. See the published version at <https://doi.org/10.1038/s41467-021-24497-8>.

Abstract

Targeted high-throughput DNA sequencing is a primary approach for genomics and molecular diagnostics, and more recently as a readout for DNA information storage. Oligonucleotide probes used to enriching gene loci of interest have different hybridization kinetics, resulting in non-uniform coverage that increases sequencing costs and decreases sequencing sensitivities. Here, we present a deep learning model (DLM) for predicting NGS sequencing depth from DNA probe sequence. Our DLM includes a bidirectional recurrent neural network that takes as input both DNA nucleotide identities as well as the calculated probability of the nucleotide being unpaired. We applied our DLM to two different NGS panels: a designed 7,373-plex panel for DNA information storage, and a 39,145-plex panel for human single nucleotide polymorphisms. In cross-validation, our DLM predicts sequencing depth to within a factor of 3 with 99% accuracy for the designed panel, and 93% accuracy for the human panel. The same model is also effective at predicting the measured single-plex kinetic rate constants of DNA hybridization and strand displacement.

Introduction

With more than 3 billion DNA nucleotides in the haploid human genome, deep sequencing of the entire human genome for clinical applications is not economically feasible. Instead, researchers and diagnostic laboratories typically use targeted sequencing, in which a set of DNA hybridization probes is designed to bind and enrich the DNA regions of interest [1, 2]. However, the DNA oligonucleotide probes in a targeted sequencing panel typically all have different kinetics and thermodynamics of binding to their respective targets. Consequently, a naively designed and synthesized panel of DNA probes will result in grossly different enrichment efficiencies for different genetic loci.

The sensitivity of NGS to a locus is directly proportional to the number of NGS reads that contain the locus (the locus's *sequencing depth*). Nonuniformity of sequencing depth either reduces the sensitivity at low-depth loci, or necessitates additional sequencing to guarantee that all loci are sequenced to a minimum desired depth. Empirical optimization of an NGS panel's probe sequences and concentrations is time- and labor-consuming, but currently cannot be avoided. A computational method to predict the sequencing depth based on probe sequence could inform the selection of probe sets with higher uniformity and modulation of probe concentrations to achieve higher uniformity.

Here, we constructed a deep learning model (DLM) for predicting NGS sequencing depth for a given oligonucleotide probe, and characterized its performance on predicting the sequencing depths of two NGS panels, one with 39,145 probes against human single nucleotide polymorphisms, and one with 7,373 probes bearing artificially designed sequences for information storage [6]. Our DLM is based on a recurrent neural network (RNN) architecture to better captures both short-range and long-range interactions within the DNA probe sequence that can impact capture efficiency/speed.

The DNA biochemistry and biophysics literature contains several well-validated models of DNA structure, thermodynamics [8, 15], and kinetics [31–34]. Ignoring our extensive knowledge of DNA biophysics and relying only on DNA sequence information would likely lead to suboptimal DLM performance. Simultaneously, we want to avoid extensive feature construction and curation, as such expert systems are generally labor-intensive to build and exhibits low generalizability to adjacent problems. Consequently, we decided to take a middle ground where we utilized only a small number of global (oligonucleotide molecule-level) features and local (individual nucleotide-level) features that can be fully autonomously computed by the well-accepted DNA folding software Nupack [14].

Results

Design of the Deep Learning Model. In the genomics field, DNA probe oligonucleotide lengths range between 50 nt and 150 nt. Thus, in designing our DLM, we considered that the model should be generalizable to DNA sequences of different lengths. To this end, NNs with a fixed number of input nodes, including conventional feed-forward NNs and convolutional NNs for image recognition [20], are not well-suited for DNA sequence inputs. Furthermore, from DNA thermodynamics and structure studies [14–17], we know that distal DNA nucleotides can hybridize to each other in secondary structures. These long-range interactions in DNA molecules are better captured by recurrent neural networks (RNNs), which have been applied commercially in speech recognition and natural language processing [18].

In brief, RNNs contain a number of internal hidden nodes, which are updated serially based on the ordered inputs and its current state values. RNNs have two primary implementations: long short-term memories (LSTMs) and gated recurrent units (GRUs). We chose to implement our DLM using GRUs because they have been reported to achieve similar performance using fewer computational resources [22]. Our DLM includes a total of four GRUs grouped into two sets: two GRUs for target sequence T , and two GRUs for probe sequence P . Although the target sequence is always the reverse complement of the probe sequence in our DLM model, we included separate GRUs for T and P both to ease the training of the model and to enable the DLM to be more generalizable to problems with asymmetric information on T and P , such as in the strand displacement kinetics that we discuss later.

Each of the two GRUs for each oligonucleotide (T or P) takes the sequence either in the direction from $5'$ to $3'$, or from $3'$ to $5'$. Unlike biological polymerization reactions which have a clear $5'$ to $3'$ directionality, the hybridization process is equally likely to initiate on either end. For RNNs and GRUs, the last inputs tend to have a larger influence on the final state values of the hidden nodes, so the design decision to include sequences in both directions is aimed at reducing input direction bias.

For each GRU, at every single nucleotide there are three input variables: (1) a binary bit indicating whether the nucleotide is a purine (A or G), (2) a binary bit indicating whether the nucleotide is “strong” (G or C), and (3) an analog Nupack-computed probability p_{unpaired} that the nucleotide is unpaired at the reaction conditions [14]. We chose to encode the identity of each nucleotide in two dimensions rather than a single dimension (e.g. A = 1, T = 2, C = 3, G = 4), in order to reflect the pairwise “distances” between any two

nucleotides, based on DNA biochemistry knowledge. The unpaired probability of each nucleotide reflects our bio-physical understanding that only unpaired nucleotides can participate in hybridization reactions; a paired nucleotide must first dissociate in order to allow new Watson-Crick base-pairing. p_{unpaired} is calculated using Nupack and considers the ensemble of all possible secondary structures that can be adopted by each DNA molecule, rather than just the minimum free energy structure.

Each GRU was designed to have 128 hidden nodes (h_t). All node values are initialized to 0, and updated based on each nucleotide's information. The hidden nodes of the RNNs represented potential patterns in the DNA sequence that the GRU could identify, and the final values of states after updating all nucleotides in the T and P sequences correspond to the presence or absence of those patterns. Thus, the number of hidden nodes in the RNN (currently 128) limited the maximum number of patterns that could be observed by the RNN. Preliminary studies showed similar prediction performance for GRUs with 128 internal states as for 256 internal states (data not shown), suggesting that 128 states were sufficient to capture the bulk of the patterns. Through the course of DLM training and weight updating through back-propagation, the GRU parameter weights were modified until they represented frequently observed patterns in the training data.

Downstream of the GRU, we used a conventional feed-forward neural network (FFNN) that takes as input the final state values of the hidden nodes of the GRUs (128 from $H^5 \rightarrow 3$ and 128 from $H^3 \rightarrow 5$). In addition to the hidden node values, the FFNN also takes as input 4 global features: the reaction temperature, the predicted standard free energy of folding of probe ($\Delta G^\circ(P)$) and Target ($\Delta G^\circ(T)$), and the predicted standard free energy of formation of the TP double-stranded DNA molecule ($\Delta G^\circ(TP)$). These global features were intended to capture properties of the $T + P$ reactions that were not easily revealed by the base pair probabilities. Thus, a total of 260 nodes were used as FFNN input. The FFNN network contained 2 hidden layers with 256 and 128 nodes, respectively; these values were picked arbitrarily based on our experience, and overall prediction performance did not appear to be sensitive to the dimensionality of the FFNN hidden layers.

Training and Validation the DLM on NGS read depth. Each of the two NGS datasets were used to independently train the DLM, and sequence depths were predicted in cross-validation for each NGS dataset individually. The reason for this is because each NGS experiment had a large number of different experimental variables (e.g. total number of reads, hybridization temperatures, experimental operator, sequencing instrument, etc.) that we felt were beyond the scope of the DLM. From a practical point of view, we expect that most users would aim to optimize probe sequence and concentration selective to improve uniformity within an NGS panel, rather than across different panels on different instruments.

For each NGS dataset, we randomly split the data into 20 classes, and predictions of each class (5% of total dataset) were obtained by a DLM trained on the remaining 19 classes (95% of total dataset), as shown in Fig. 2a. Thus, a total of 20 DLMs were used in the 20-fold cross validation predictions for evaluating prediction accuracy. There are roughly 300,000 weight parameters in the DLM (illustrated in Fig. 1c); these were preset via Xavier initialization (uniformly distributed weights with standard deviation

dependent on the number of parameters in a layer) in order to alleviate the vanishing gradient problem for deep NNs [36].

During training, we iteratively minimized the square error between the predicted and experimental log sequencing depth, using gradient descent with an Adam optimizer [37] to update the network weights. To minimize overfitting, we implemented an additional dropout layer after each hidden layer of the FFNN, in which 20% of parameters are randomly selected and prevented from updating in each training iteration. The DLM was implemented using Tensorflow [38], and DLM hyper-parameters include GRU hidden nodes (128), FFNN hidden nodes (256 and 128), batch size (999), learning rate (0.0001), and node dropout fraction (20%). We tried roughly 50 sets of different hyper-parameter values, and the values listed appear to yield the shortest training time and best predictive performance.

Fig. 2c summarizes the root-mean-square error (RMSE) of the DLM predicted values of $\log_{10}(\text{Depth})$ based on different sequences vs. the actual observed NGS read depth for a human single nucleotide polymorphism (SNP) panel comprising 39,145 probes synthesized as a pool by Twist Biosciences. Of the 39,145 probes, NGS results showed 0 reads on 1,105 probes. Our previous studies on Twist oligonucleotide pools suggest that the lack of sequencing reads for these may indicate difficulties with probe synthesis [43]. Consequently, we chose to exclude these probes with 0 observed NGS reads in order to eliminate the possibility of training against noise.

The DLM yields an average RMSE of roughly 0.30 on both the training classes and the test classes. For comparison, a naive model of predicting sequencing depth only based on the mean $\log_{10}(\text{Depth})$ of all observed sequences produces an RMSE of 0.41. Fig. 2d plots the comparison of predicted and measured sequencing depths for each sequence. From this figure, we see that a significant contributor to our DLM's RMSE is a subset of DNA sequences that are observed to have very low $\log_{10}(\text{Depth})$ (e.g. 0.3, corresponding to a depth of 2), but predicted to have $\log_{10}(\text{Depth})$ between 1 and 3.3. Our interpretation of this phenomenon is that there are a myriad possible reasons why probes may yield extremely low depth, e.g. poor probe synthesis yield, poor hybridization yield, sequence biases in nonspecific DNA binding to plasticware, and unintended crosstalk binding between different subsequences of the human genome, and poor bridge PCR efficiency during Illumina NGS that result in under-representation of reads. The DLM does not have sufficient training instances for each reason to be able to make confident predictions of very low sequencing depth.

Fig. 2e shows the DLM results applied to a second NGS panel comprising 7,373 DNA probes against non-biological DNA sequences intended for DNA information storage applications. As these sequences were procedurally generated to avoid known problematic DNA sequences, such as those with high or low G/C content (Fig. 2b), homopolymers, etc., there is much less variation in sequencing depth to begin with. Nonetheless, the DLM is effective at predicting sequencing depths (Fig. 2e) beyond a naive model (see also Supplementary Section S3).

Our DLM in total contains over 300,000 parameters (e.g. node biases and node-node weights). The large number of parameters leads to potential concern regarding overfitting and model reproducibility. To address this, we next performed 15 independent DLM parameter initiation and training on the dataset, in order to characterize the reproducibility of the model (Fig. 3). All 15 DLMs consistently reached early-stop at roughly epoch 250 (Fig. 3a), and the predicted sequencing depths showed high pairwise concordance (Fig. 3bc). Across the 105 pairwise comparisons of the 15 DLMs, we observe a Pearson's r value of no less than 0.975. Consequently, we believe that our approach produces DLMs with fairly consistent predictions despite variations in parameter initialization and training.

DLM prediction of single-plex DNA hybridization and strand displacement rate constants. Based on our understanding of DNA and NGS, we believe that NGS read depth is primarily dependent on the yield and speed of DNA probe hybridization, and secondarily by instrument- and chemistry-specific biases. Consequently, our DLM should also be effective at predicting the rate constants of hybridization of DNA (Fig. 4a). To further challenge our DLM and to highlight the effectiveness of our DLM approach, we further applied the DLM to the prediction of a related DNA mechanism, strand displacement [24, 25] (Fig. 4b). Unlike NGS experiments in which thousands of DNA probes and targets are simultaneously hybridizing, for DNA hybridization and strand displacement rate constant prediction, we use time-based fluorescence data in which a single target and probe species are observed with high time and yield resolution (Fig. 4c, Supplementary Section S2. See also ref. [31] for additional explanation on hybridization reaction kinetics experimental details.

Experimental hybridization rate constants are taken from ref. [31], and experimental strand displacement rate constant data are collected for this work. Supplementary Section S2 describes details of how best-fit rate constants are fitted to fluorescence-based kinetics data. Fig. 4d shows the DLM prediction vs. experimental best-fit rate constants for 210 hybridization reactions and 211 strand displacement reactions. Here, we performed 100-fold leave-one-class-out (LOCO) prediction rather than 20-fold cross validation, because the smaller number of data points would lead to significant biases due to small sample sizes of the test set; see Supplementary Section S3 for details. The variation in rate constants observed is similar to the variation in NGS sequencing depths (4 logs), though the latter is possibly somewhat smaller due to the saturation of hybridization for the timescales of NGS hybrid-capture reactions.

The purpose of this sub-study was to see if the DLM could be used for non-NGS applications of nucleic acid molecular diagnostics, such as those based on qPCR [41] and electro-chemistry [27]. Importantly, our DLM was trained simultaneously on both the hybridization and the strand displacement rate constant datasets. Because the target T and probe P sequences for the hybridization reactions are identical to that of strand displacement reactions, the difference between the two is manifested only in the predicted probability of each nucleotide being unpaired. This information alone was enough to communicate to the DLM the distinction between hybridization and strand displacement, and no special case handling or neural network architecture modification was needed to accommodate strand displacement.

Contribution of Different Features to DLM Performance. The architecture of the DLM and the local and global features were initially decided based on our understanding of the behavior of DNA, rather than through knowledge-free exploration. Consequently, it is possible that some of the features are not directly relevant to NGS depth or hybridization/strand displacement rate constants. To test this hypothesis, we next constructed a series of DLMs in which different features were removed from the model (Fig. 5).

We found that 3 of the 4 global features, individually, had essentially no impact on any of the predictions. Both sets of local features (sequences and base probabilities) were important for some aspect of the DLM prediction, but it appears that the two are interchangeable for predicting the NGS depths of the human SNP panel. Examining the global features more closely, we note that temperature T is the same for all sequences within a panel, so it is tautological that the DLM cannot learn any impact from changing T .

The standard free energy of formation of the target-probe duplex E_{TP} likely did not matter because the lengths of all probes/targets were long enough that the probe binding was no longer limited by its thermodynamics. Finally, the standard free energy of folding of the target by itself E_T likely did not matter because it was not and could not be accurately calculated: whereas the probe P has a homogenous molecular population with a well-defined sequence, the target T is a heterogeneous mixture constructed through randomized physical fragmentation of human genomic DNA. Consequently, the 5' and 3' overhang sequences of the target are highly variable, and cannot be reflected as a single sequence. Prediction accuracies of all feature-reduced DLMs are summarized in Supplementary Section S4. The performance of even reduced models are in general much better than naive models (Supplementary Section S5).

Discussion

Targeted high-throughput sequencing of DNA has become a dominant method for biological and biomedical research, and furthermore is becoming standard of practice for cancer treatment [48]. More recently, targeted sequencing has been explored as a method for random-access readout of information stored densely and for long term in DNA [6]. Although DNA sequencing costs are exponentially decreasing over the years [49], poor sequencing uniformity would waste a large majority of reads sequencing high-depth targets redundantly and providing insufficient information on low-depth targets. Consequently, a strong need exists to rationally design NGS panels with high uniformity.

However, predicting DNA hybridization kinetics and efficiency is extremely difficult even for experts in single-plex settings [31]. In a complex multi-component system that is hybrid-capture target enrichment, prediction of sequencing depth becomes intractable for first-principles biophysical models. At the same time, the large size of NGS datasets renders the problem well-suited for machine learning.

Deep learning leverages large datasets to autonomously discover weak correlative features between inputs and outputs. This has led to deep learning becoming dominant in computer vision and other areas

with large available datasets. On the other hand, simpler statistical models (e.g. multivariate linear regression) remain dominant in the natural and biomedical sciences [50, 51] where well-curated data for specific problems are scarce. Expert system machine learning approaches based on extensive manual feature construction and curation takes an in-between approach, using expert knowledge to guide the construction of narrowly optimized prediction software, but are generally poor at generalization to similar problems.

In our DLM, we restricted our inputs to a limited set of global and local features that can be automatically computed based on DNA sequence, in order to avoid the trap of labor-intensive and problem-specific model construction. Given both DNA thermodynamics model inaccuracy/incompleteness [8] and the fact that we could not feasibly consider the intermolecular interactions from all 3+ billion nucleotides of the human genome, we believe that the Nupack-predicted base pair probability values likely have significant error. Future models for more accurately predicting base pair accessibility in a highly complex and heterogeneous solution could prove crucial to further improving the prediction accuracy of the DLM.

DNA sequence is rather unlike most other inputs for problems solved by deep learning networks. DNA molecules are known to have long-range interactions where distal DNA nucleotides bind to each other. Simultaneously, there are not orderly “grammar” rules such as in natural language processing [19] that can be readily discovered by neural networks. Distal and intermolecular DNA binding are essentially the effects of chemistry and does not necessarily need to conform to human common sense. These all contribute to the difficulty of building neural networks that accurately predict DNA behavior based on sequence.

Conversely, once a neural network architecture is established to “understand” DNA sequence, it could hold the potential to a large range of other nucleic acid-based problems. As a research example, a large range of non-coding RNA [52] have been discovered, and being able to predict their structures can provide insights to their function. As a biomedical example, codon optimization problems [53] for synthetic biology, including *de novo* construction of RNA-based drugs [54]. The incorporation of generalizable domain knowledge within deep learning architectures will be a key enabler for predicting behaviors for nucleic acids, given the impact of their sequences on form and function and the exponential number of possible sequences of given lengths.

Declarations

Acknowledgements. This work was funded by NIH grant R01HG008752 and U01CA233364 to DYZ. The authors thank Jianyi Nie for editorial assistance.

Author contributions. JXZ, BY, AG, AP, and DYZ conceived the project. PD, JXZ, YJC, KZ and JZF performed the experiments. ND performed hybridization reaction model fitting and selection. AG, BY, and MXW created and optimized the DLM. JL processed sequencing data. AP and DYZ wrote the manuscript with input from all authors.

Additional information. Correspondence may be addressed to AP (Andrew.Phillips@microsoft.com) or DYZ (dysz1@rice.edu). There is a patent pending on X-probes used in this work. There is a patent pending on the WNV model of rate constant prediction. MXW declares a competing interest in the form of consulting for Nuprobe. DYZ declares a competing interest in the form of consulting for Nuprobe, Torus Biosystems, and Avenge Bio.

Data Availability. The sequences of the DNA oligos used for the manuscript, the oligo concentrations used for fluorescence experiments, the best-fit rate constants, and the values of the manually-constructed features for the WNV model are included as an Excel file accompanying this manuscript. The original raw fluorescence data from our single-plex kinetics experiments are also included as a Zip file.

Code Availability. To allow readers to use the DLM for predicting rate constants, we provide our DLM software code and installation/usage instructions. These are attached with the submission. We also intend to make an online software tool available by the time of publication.

References

1. L. Mamanova, A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, D. J. Turner, Target-enrichment strategies for next-generation sequencing. *Nature methods*, 7(2), 111 (2010).
2. Gnirke, A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, Brockman, S. Gabriel. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature biotechnology*, 27(2), 182 (2009).
3. I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann (2016).
4. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature*, 521(7553), 436-444 (2015).
5. Y. Zhang, S. X. Chen, P. Yin, Optimizing the specificity of nucleic acid hybridization. *Nature chemistry*, 4(3), 208 (2012).
6. Ceze, J. Nivala, K. Strauss, Molecular digital data storage using DNA. *Nature Reviews Genetics*, 20(8), 456-466 (2019).
7. R. Wu, J. S. Wang, J. Z. Fang, E. R. Evans, A. Pinto, I. Pekker, Y. Zhang, Continuously tunable nucleic acid hybridization probes. *Nature methods*, 12(12), 1191 (2015).
8. Wang, D. Y. Zhang. Simulation-guided DNA probe design for consistently ultraspecific hybridization. *Nature Chem* 7, 545–553 (2015).
9. R. Wu, S. X. Chen, Y. Wu, A. A. Patel, D. Y. Zhang, Multiplexed enrichment of rare DNA variants via sequence-selective and temperature-robust amplification. *Nature biomedical engineering*, 1(9), 714 (2017).
10. Seelig, D. Soloveichik, D. Y. Zhang, E. Winfree, Enzyme-free nucleic acid logic circuits. *science*, 314(5805), 1585-1588 (2006).

11. Qian, E. Winfree, Scaling up digital circuit computation with DNA strand displacement cascades. *Science*, 332(6034), 1196-1201 (2011).
12. M. Cherry, L. Qian, Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks. *Nature*, 559(7714), 370 (2018).
13. M. Dirks, N. A. Pierce, A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of computational chemistry*, 24(13), 1664-1677 (2003).
14. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. Khan, N. A. Pierce, NUPACK: analysis and design of nucleic acid systems. *Journal of computational chemistry*, 32(1), 170-173 (2011).
15. SantaLucia, D. Hicks. The Thermodynamics of DNA Structural Motifs. *Ann. Rev. Biochem.* 33, 415-440 (2004).
16. Zuo, X., Cui, G., Merz, K. M., Zhang, L., Lewis, F. D., & Tiede, M. (2006). X-ray diffraction "fingerprinting" of DNA structure in solution for quantitative evaluation of molecular dynamics simulation. *Proceedings of the National Academy of Sciences*, 103(10), 3534-3539.
17. H. Turner, Thermodynamics of base pairing. *Current opinion in structural biology*, 6(3), 299-304 (1996).
18. Graves, A. R. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). IEEE (2013).
19. Cambria, B. White, Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2), 48-57 (2014).
20. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105) (2012).
21. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
22. Chung, C. Gulcehre, K. Cho, & Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
23. Y. Zhang, E. Winfree, Robustness and modularity properties of a non-covalent DNA catalytic reaction. *Nucleic acids research*, 38(12), 4182-4197 (2010).
24. Y. Zhang, Towards domain-based sequence design for DNA strand displacement reactions. In *International Workshop on DNA-Based Computers* (pp. 162-175). Springer, Berlin, Heidelberg (2010, June).
25. C. Simmel, B. Yurke, H. R. Singh, Principles and applications of nucleic acid strand displacement reactions. *Chemical reviews*, 119(10), 6326-6369 (2019).
26. Taylor, , Wakem, M., Dijkman, G., Alsarraj, M., & Nguyen, M. (2010). A practical approach to RT-qPCR—publishing data that conform to the MIQE guidelines. *Methods*, 50(4), S1-S5.

27. Das, , Ivanov, I., Montermini, L., Rak, J., Sargent, E. H., & Kelley, O. (2015). An electrochemical clamp assay for direct, rapid analysis of circulating nucleic acids in serum. *Nature chemistry*, 7(7), 569.
28. Langmead, S. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9:357-359 (2012).
29. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, 25(14), 1754-1760 (2009).
30. S. Wang, D. Y. Zhang, Simulation-guided DNA probe design for consistently ultraspecific hybridization. *Nature chemistry*, 7, 545 (2015).
31. X. Zhang, J. Z. Fang, W. Duan, L. R. Wu, A. W. Zhang, N. Dalchau, B. Yordanov, R. Petersen, A. Phillips, D. Y. Zhang, Predicting DNA hybridization kinetics from sequence. *Nature Chemistry*, 10, 91-98 (2018).
32. Y. Cheng, H. Chen, J. Morrison, Kinetics of DNA replication in a dicentric X chromosome formed by long arm to long arm fusion. *Human genetics*, 56(1), 71-79 (1980).
33. P. Reynaldo, A. V. Vologodskii, B. P. Neri, V. I. Lyamichev, The kinetics of oligonucleotide replacements. *Journal of molecular biology*, 297(2), 511-520 (2000).
34. Y. Zhang, E. Winfree, Control of DNA strand displacement kinetics using toehold exchange. *Journal of the American Chemical Society*, 131(47), 17303-17314 (2009).
35. Srinivas, T. E. Ouldrige, P. Sulc, J. M. Schaeffer, B. Yurke, A., Louis, E. ... Winfree, On the biophysics and kinetics of toehold-mediated DNA strand displacement. *Nucleic acids research*, 41(22), 10641-10658 (2013).
36. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256) (2010).
37. P. Kingma, Vol. 5. (2015) (ICLR).
38. Abadi, M., Barham, , Chen, J., Chen, Z., Davis, A., Dean, J., & Kudlur, M. (2016, November). Tensorflow: a system for large-scale machine learning. In *OSDI* (Vol. 16, pp. 265-283).
39. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from over-fitting. *The Journal of Machine Learning Research*, 15(1), 1929-1958 (2014).
40. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, N. Arnheim, Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732), 1350-1354 (1985).
41. Higuchi, C. Fockler, G. Dollinger, R. Watson, Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Nature Biotechnology*, 11(9), 1026 (1993).
42. Khodakov, C. Wang, D. Y. Zhang, Diagnostics based on nucleic acid sequence variant profiling: PCR, hybridization, and NGS approaches. *Advanced drug delivery reviews* (2016).

43. Pinto, S. X. Chen, D. Y. Zhang. Simultaneous and stoichiometric purification of hundreds of oligonucleotides. *Nature communications*, 9(1), 1-9 (2018).
44. Castanotto, J. J. Rossi, The promises and pitfalls of RNA- interference-based therapeutics. *Nature*, 457(7228), 426 (2009).
45. Deng, C. C. Wang, K. W. Choy, Q. Du, J. Chen, Q. Wang, T. Tang, Therapeutic potentials of gene silencing by RNA interference: principles, challenges, and new strategies. *Gene*, 538(2), 217-227 (2014).
46. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, F. Zhang, Multiplex genome engineering using CRISPR/Cas systems. *Science*, 1231143 (2013).
47. A. Doudna, E. Charpentier, The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213), 1258096 (2014).
48. Meldrum, M. A. Doyle, R. W. Tothill, Next-generation sequencing for cancer diagnostics: a practical perspective. *The Clinical Biochemist Reviews*, 32(4), 177 (2011).
49. R. Mardis, A decade's perspective on DNA sequencing technology. *Nature*, 470(7333), 198-203 (2011).
50. Cronin, C. Sangli, M. L. Liu, M. Pho, D. Dutta, A. Nguyen, D. Watson, Analytical validation of the Oncotype DX genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor-positive breast cancer. *Clinical chemistry*, 53(6), 1084-1091 (2007).
51. D. Cohen, L. Li, Y. Wang, C. Thoburn, B. Afsari, L. Danilova, R. Hruban, Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, eaar3247 (2018).
52. Beermann, M. T. Piccoli, J. Viereck, T. Thum, Non-coding RNAs in development and disease: background, mechanisms, and therapeutic approaches. *Physiological reviews*, 96(4), 1297-1325 (2016).
53. Presnyak, N. Alhusaini, Y. H. Chen, S. Martin, N. Morris, N. Kline, ... , J. Collier, Codon optimality is a major determinant of mRNA stability. *Cell*, 160(6), 1111-1124 (2015).
54. C. Burnett, J. J. Rossi, RNA-based therapeutics: current progress and future prospects. *Chemistry & biology*, 19(1), 60-71 (2012).

Figures

Figure 1

Overview of targeted NGS panel workflow and deep learning model (DLM) for predicting sequencing depth. (a) DNA probes (P) are designed for hybridizing to target sequences (T) for subsequent solid-phase separation and enrichment. In NGS, each read corresponds to a randomly sampled DNA molecule from the enriched library, and NGS reads are bioinformatically aligned to the probe sequences using standard algorithms and software. See Supplementary Section S1 for further details on the NGS experimental and bioinformatic workflow. (b) Overview of NGS read depth prediction method. (c) The DLM consists of 4 recurrent neural networks implemented as gated recurrent units (GRUs) and 1 feed-forward fully-connected neural network (FFNN). Each GRU has 128 internal state nodes. The final GRU node values for the target DNA sequence T and for the probe DNA sequence P from 51 to 31 ($H_5 \rightarrow 3$ and $H_5 \rightarrow 3$) are summed; likewise $H_3 \rightarrow 5$ and $H_3 \rightarrow 5$. These two hidden state node sum vectors are then concatenated into a vector of P T P 256 node values, serving as input to the FFNN. Additionally, 4 global parameters also serve as input to the FFNN, bringing the total inputs to 260. The output was a single node corresponding to the log predicted read depth $\log_{10}(\text{depth})$ for the DNA target sequence.

Figure 2

Training and cross-validation results of the DLM on predicting NGS read depth. (a) Data pre-processing and training pipeline. Raw FASTQ NGS reads were aligned to the target sequences [28, 29], and the read depth of each template sequence is counted. Separately, local and global features for each target/probe are computed using Nupack [14]. The complete dataset was split into 20 classes, and each class was predicted based on the DLM training results on the other 19 classes. (b) Histogram of GC content of the probe sequences from two panels. The human single nucleotide polymorphism (SNP) panel contains a total of 39,145 probes, and contained a greater variability in G/C content. Among the human SNP panel, 1,105 probes had a experimental sequencing depth of 0, and were excluded from training and prediction, resulting in 38,040 data points for analysis. The synthetic DNA panel used for DNA information storage contains a total of 7,373 probes, and were procedurally designed and have a tighter G/C content distribution. No synthetic DNA probes were excluded from analysis. (c) Summary of observed root-mean-square error (RMSE) for the human SNP panel. Each point corresponds to the prediction results for one of the 20 validation classes; the naive model predicts $\log_{10}(\text{Depth})$ based on the mean of the training set.

(d) Predicted vs. observed sequencing depth for the human SNP panel. Plotted are the aggregated results for all 20 validation classes. Shaded in dark gray are the zones where the predicted and the observed read depth agreed to within a factor of 2; light gray shows agreement to within factor of 3. F2err and F3err denotes the fraction of sequences with depth predicted to beyond a factor of 2 and 3, respectively. (e) Predicted vs. observed sequencing depth for the synthetic DNA panel.

Figure 3

Reproducibility of DLM predictions. (a) Plot of Tensorflow Loss vs. training epoch, 15 independent DLM training runs. At above 250 epochs, the training set Loss consistently continues to decrease but the validation set Loss start increasing, indicating overfitting. Consequently, we halt DLM training after 250 epochs. (b) Comparison of read depth prediction results on the human SNP panel from two DLMs independently trained. Each DLM was initiated with different weight parameters. See Supplementary Section S3 for additional comparison results. (c) Deviation between predicted read depths from 2

independently trained DLMs. (d) Summary of pairwise comparison results for 15 independently trained DLMs using the human SNP panel data (105 total comparisons).

Figure 4

Applying our DLM to prediction of single-plex hybridization and strand displacement rate constants. Schematic of (a) hybridization and (b) strand displacement reactions. (c) Sample kinetics traces of a hybridization and a strand displacement reaction. Reaction yields inferred through fluorescence; see Supplementary Section S2 for experimental and data processing details. (d) Accuracy of DLM predictions of hybridization (blue) and strand displacement (orange) rate constants. Due to the small number of data points here (421), we implemented prediction based on leave-one-out rather than 20-fold cross validation, due to the large expected variation in small validation classes. Note that strand displacement and hybridization reaction parameters were co-trained using the same DLM, and predictions were likewise using a single DLM. (e) Comparing DLM prediction performance to a previous expert system machine learning approach based on weighted neighbor voting [31]. Prediction results are similar.

Figure 5

Assessing importance of different components of the DLM to prediction accuracy, measured by RMSE. We tested different DLM that lack some features, in order to assess which ones meaningfully contribute to prediction performance. Highlighted in red are the applications in which a reduced DLM model performs significantly worse than our default DLM. Columns highlighted in blue indicate the reduced DLMs with essentially identical performance to our default DLM. While some features are redundant for the specific applications considered here, we note that most features improve performance in at least one application.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [NNKSuppv1.pdf](#)