Predicting DNA hybridization kinetics from sequence

Jinny X. Zhang1,2†, John Z. Fang3, Wei Duan1, Lucia R. Wu1, Angela W. Zhang1, Neil Dalchau3, Boyan Yordanov3, Rasmus Petersen3, Andrew Phillips3 and David Yu Zhang1,2*†

Hybridization is a key molecular process in biology and biotechnology, but so far there is no predictive model for accurately determining hybridization rate constants based on sequence information. Here, we report a weighted neighbour voting (WNV) prediction algorithm, in which the hybridization rate constant of an unknown sequence is predicted based on similarity reactions with known rate constants. To construct this algorithm we first performed 210 fluorescence kinetics experiments to observe the hybridization kinetics of 100 different DNA target and probe pairs (36 nt sub-sequences of the CYCS and VEGF genes) at temperatures ranging from 28 to 55 °C. Automated feature selection and weighting optimization resulted in a final six-feature WNV model, which can predict hybridization rate constants of new sequences to within a factor of 3 with ~91% accuracy, based on leave-one-out cross-validation. Accurate prediction of hybridization kinetics allows the design of efficient probe sequences for genomics research.

Hybridization of complementary DNA and RNA sequences is a fundamental molecular mechanism that underlies both biological processes1–3 and nucleic acid analytic biotechnologies4–7. The thermodynamics of hybridization have been well studied, and algorithms based on the nearest-neighbour model of base stacking8,9 predict minimum free-energy structures and melting temperatures10,11 with reasonably good accuracy. In contrast, the kinetics of hybridization remain poorly understood, and no models or algorithms have been reported that accurately predict hybridization rate constants from sequence and reaction conditions (temperature and salinity). This knowledge deficiency has adversely impacted the research community by requiring either trial-and-error optimization of DNA primer and probe sequences for new genetic regions of interest, or brute-force use of thousands of DNA probes for target enrichment.

Predictive modelling of hybridization kinetics faces two main challenges. First, the kinetics of very few DNA sequences have been characterized directly, either in bulk solution12–16 or at the single-molecule level17–19. The primary reason for the lack of data is the cost of fluorophore-functionalized DNA oligonucleotides, which at roughly $200 per sequence becomes prohibitive for the hundreds of experiments needed to establish sequence generality. Second, the hybridization of complementary sequences can follow many different pathways20, rendering simple reaction models inaccurate for a large fraction of DNA sequences.

To create a sufficiently representative and sequence-general data set for developing a predictive model of hybridization kinetics, we experimentally characterized the kinetics of 210 individual hybridization reactions on 100 different pairs of complementary sequences. We were able to do this economically through the use of the X-probe architecture, in which universal fluorophore- and quencher-functionalized oligonucleotides are recycled across many different experiments.

From our experimental data we made three unexpected findings: (1) most hybridization reactions do not asymptotically reach more than 90% yield; (2) initial hybridization kinetics is generally uncorrelated with asymptotic yield; and (3) secondary structure in the middle of a DNA target sequence tends to more adversely affect hybridization kinetics. Additionally, we observed that structure-free DNA target/probe sequences generally tended to have faster hybridization kinetics, consistent with literature and our expectations, but even structure-free sequences exhibited more than one order of magnitude of variation in hybridization rate constants.

Based on our experimental data, we also constructed a new type of algorithm to predict DNA hybridization rate constants based on the target/probe sequence, called ‘weighted neighbour voting’ (WNV). WNV is inspired by but distinct from machine learning techniques such as k nearest Neighbour27 and kernel smoothing28. In WNV, each hybridization reaction is mapped to a set of bioinformatic feature values and can be considered a point in the high-dimensional feature space. Two hybridization reactions that are close in feature space are expected to exhibit similar kinetics. The rate constant of an unknown hybridization reaction is predicted based on the weighted average of observed rate constants of experimentally tested reactions, with weights dropping exponentially for reactions that are farther away in feature space. Under leave-one-out (LOO) cross-validation, our final WNV model predicts rate constants to within a factor of 2 for 80% of reactions, and within a factor of 3 for 91%. Next-generation sequencing (NGS) studies show a significant correlation (R² > 0.6) between the rate constants of DNA hybridization in single-plex versus multiplex, suggesting that the current work is a good starting point for the rational design and selection of DNA probes for highly multiplexed applications, such as target enrichment from genomic DNA4.

Experimental results

To systematically but economically characterize the hybridization kinetics of many different sequences we used the X-Probe architecture20, which makes use of universal fluorophore and quencher-labelled oligonucleotides (Fig. 1a). A universal fluorophore-labelled oligonucleotide was pre-hybridized to the probe and a universal
sequences of the targets were selected as close overlapping frames to systematically test the position effects of secondary structures. Red markers denote the sub-sequences (triplicate) of a hybridization reaction. All reactions proceeded in 5× PBS buffer. Supplementary Section 1 provides reproducibility studies and Supplementary structures predicted for the target sequences at 37 °C. Supplementary Table 1 presents the sequences of the 100 targets.

When the target and probe solutions were mixed, the solution quencher-labelled oligonucleotide was pre-hybridized to the target. Universal sequences and green regions are variable regions corresponding to the target or probe sequence. Fluorescence is initially high and decreases as the sequence hybridization reaction proceeds because the fluorophore (purple star) becomes localized to the quencher (black dot). A total of 100 different sub-sequences of the CYCS and VEGF genes were selected to be the target sequences. In this study, all target and probe sequences are 36 nt long (excluding universal regions). Then, 25 targets for each gene were chosen randomly with uniform distribution across the entire intron and exon region and another 25 targets were selected as close overlapping frames to systematically test the position effects of secondary structures. Red markers denote the sub-sequences of the genes selected as targets. Examples of secondary structures encountered in target sequences. Shown are predicted minimum free energy (mfe) structures predicted for the target sequences at 37 °C. Supplementary Table 1 presents the sequences of the 100 targets. Example kinetics traces (triplicate) of a hybridization reaction. All reactions proceeded in 5× PBS buffer. Supplementary Section 1 provides reproducibility studies and Supplementary Section 2 fluorescence traces for all 210 experiments.

We selected, as targets, 100 sub-sequences of the CYCS and VEGF genes, each target sub-sequence being 36 nucleotides (nt) long. Of the 50 targets for each gene, 25 were selected randomly with uniform position distribution across the gene and the other 25 were selected systematically so that the effects of secondary structure position could be examined (Fig. 1b,c).

Figure 1d shows triplicate kinetics traces for one hybridization reaction. A total of 210 hybridization experiments were characterized (100 reactions at 37 °C, 96 at 55 °C, 7 at 28 °C and 7 at 46 °C). There was very low experimental error in our fluorescence experiments; all triplicate data points agreed with each other to within 2%. To obtain maximally reliable experimental data for rate constant inference, we performed multiple experiments until determining a set of target and probe concentrations such that each hybridization reaction undergoes between two and ten half-lives within the 80–180 min observation time.

In all experiments, the concentration of the target was at least double that of the probe, to minimize the effects of slight pipetting variability. To ensure that the observed kinetics are primarily due to target/probe sequence rather than synthesis impurities, we experimentally observed kinetics for the hybridization of three sets of targets and probes, each as three separate syntheses from two different vendors (Integrated DNA Technologies and Sigma). The inferred hybridization rate constants for different syntheses showed minor variations, and were all consistent to within a factor of 2 (Supplementary Section 1).

Hybridization rate constant ($k_{TP}$) fitting

A simple two-state $T + P \rightarrow TP$ reaction model fails to reasonably fit the observed fluorescence kinetics. Notably, over 40% of the reactions asymptote to a final reaction yield of less than 85%, based on the positive control fluorescence where the target and probe are thermally annealed (Supplementary Section 1). We were surprised by the extent and reproducibility of the incomplete DNA hybridization yield, which may be due to misaligned hybridization or other nonspecific interactions between target and probe.

We considered three reaction models of hybridization to explain the kinetics data (Fig. 2a). Model H1 assumes that a fraction of the probes P are incapable of proper hybridization with target T or the accompanying fluorescence quenching. Model H2 assumes that all probe P is correctly synthesized, but that some fraction of the $T + P$ reaction undergoes an alternative pathway with rate constant $k_1$ to result in a state $TP_{bad}$ with high fluorescence. This frustrated state, $TP_{bad}$, may represent states in which T and P are co-localized by misaligned base pairs. Model H3 is a combination of models H1 and H2, wherein there exists both a fraction of bad P as well as the alternative pathway involving $TP_{bad}$.

For each of our 210 fluorescence kinetics experiments we performed fitting using each of the three models (Fig. 2b), finding parameters that minimize the sum-of-square relative error RE, where

$$RE = ((Data - Simulation)/Data).$$

The RE values of each hybridization
We next examined the systematically designed DNA target/probe sequence pairs for trends in $k_{\text{Hyb}}$ (Fig. 3d). The systematic sequences included 13 sets of three DNA target/probe pairs, each frame-shifted a small number of bases so that predicted secondary structure lies in the middle or 3′ regions of the target (Fig. 1c). We observed two interesting trends. First, the observed $k_{\text{Hyb}}$ values can vary greatly within a cluster: for example, targets 1–3 shows about 30-fold difference in hybridization rate constant, despite all three having similar standard free energies of folding. This indicates that the relative portion of secondary structures within a DNA sequence can have a large impact on kinetics. Second, targets with the secondary structure in the middle of the sequence (circles in Fig. 3d) tended to be slower to hybridize than targets with the structure at one end: in 8 of the 13 clusters, the target with central secondary structure was the slowest in each respective cluster.

Literature reports and our own prior experience suggested that unstructured DNA sequences would hybridize more rapidly and with higher yield than structured ones. To see whether our experimental data are consistent with this observation, we plotted $k_{\text{Hyb}}$ and the bad fraction for only the hybridization reactions in which both the target and the probe have an ensemble (partition function) standard free energy $\Delta G^\circ$ of $>3$ kcal mol$^{-1}$, as predicted by Nupack for hybridization temperature and buffer conditions (Fig. 3e,f). The observed $k_{\text{Hyb}}$ values for these structure-free sequences are indeed faster than ‘typical’ sequences, with all $k_{\text{Hyb}} >1 \times 10^6$ M$^{-1}$ s$^{-1}$. Nonetheless, there is still significant variability in $k_{\text{Hyb}}$, ranging over more than 1 log. The asymptotic yield of the hybridization reactions is only slightly better for structure-free sequences than for other sequences.

**Predictive model construction**

**WNV model.** Our WNv model predicts the value of $k_{\text{Hyb}}$ for new hybridization reactions based on the similarity of the reaction to hybridization reactions with known rate constants (labelled instances). Each labelled instance makes a weighted vote of $\log_{10}(k_{\text{Hyb}})$, with instances that are more similar to the new reaction being weighted more heavily. The 210 hybridization reactions across 100 different target/probe pairs act as our initial database of labelled instances.

For each hybridization reaction, a number of features $f_i$ are calculated based on the sequences of the target and probe and the hybridization reaction temperature and buffer (Fig. 4b). A total of more than 50 different features were tested, of which 35 showed significant individual correlation with $k_{\text{Hyb}}$ (Supplementary Section 4). The disparity between two different hybridization reactions $j$ and $m$ is quantitated as distance $d_{jm}$, the Euclidean distance between the two hybridization reactions in feature space:

$$d_{jm} = \sqrt{\sum_i (f_i(j) - f_i(m))^2}$$

where $f_i(j)$ is the value of weighted feature $i$ for reaction $j$. Higher weights result in a wider feature dimension, which can potentially contribute more to the feature space distance (Fig. 4d).

From the database of hybridization experiments $m$ with known $k_{\text{Hyb}}(m)$ values, our WNv model makes the following prediction for $k_{\text{Hyb}}(j)$ of an unknown hybridization reaction $j$:

$$\log_{10}(k_{\text{Hyb}}(j)) = \frac{1}{Z_j} \sum_m 2^{-d_{jm}} \log_{10}(k_{\text{Hyb}}(m))$$

where $Z_j = \sum_m 2^{-d_{jm}}$ is the ‘partition function’ of the distances involving reaction $j$ (Fig. 4e). Figure 4f shows the relationship between feature space distance between a pair of hybridization
reactions (using our final feature list and weights) and their difference in observed $k_{\text{Hyb}}$ values.

The WNV model can be extended to any number of features. In general, the potential improvements in $k_{\text{Hyb}}$ prediction accuracy must be balanced against increased model complexity from having a large number of features. Additionally, the higher-dimensional feature space that accompanies an increased number of features makes the weight optimization significantly more difficult, due to the increased number of local fitness maxima. Through a series of computational optimization steps, we determined the optimal number of features to be 6: $nGp$, Pap, Temp, wPat, GavgMSR1 and Gb (see Supplementary Section 3 for the optimization methodology).

**Model performance.** To quantitate the overall performance of a particular WNV model (defined by its set of features and corresponding feature weights $w(i)$), we constructed the following 'Badness' metric:

$$\text{Badness} = 3 \times (1 - F2acc) + 3 \times (1 - F3acc) + 4 \times \text{RMSE}$$  (4)

where $F2acc$ is the fraction of all predicted reactions $j$ in which the predicted $k_{\text{Hyb}}(j)$ and the experimental $k_{\text{Hyb}}(j)$ agree to within a factor of 2, $F3acc$ is the fraction that agrees to within a factor of 3, and

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_j \left( \log_{10}(k_{\text{Hyb}}(j)) - \log_{10}(\hat{k}_{\text{Hyb}}(j)) \right)^2}$$  (5)

is the root-mean-square error of the logarithm of the hybridization rate constant (where $N = 210$ is the number of experiments).

We chose to use this badness metric rather than RMSE only (that is, a least-squares fit) because we felt that it is more relevant for many applications involving the design of DNA oligonucleotide probes and primers. Rather than marginally improving the predictions of outlier sequences that are off by more than an order of magnitude, our badness metric instead emphasizes improving the fraction of predictions that are correct to within a factor of 3, or better yet within a factor of 2. Simultaneously, to allow efficient computational optimization of feature weights, the badness metric to be minimized cannot be locally flat, so RMSE is included as a component of badness. Use of different badness metrics will result in optimized feature weights that exhibit a different tradeoff between the magnitude and frequency of large prediction errors.

---

**Figure 3** | **Summary of observed hybridization kinetics.**

(a) Observed $k_{\text{Hyb}}$ value (model H3) for 96 targets at 37 and 55 °C. Four A/T targets were excluded from this because they were A/T-rich and did not stably bind to their probes at 55 °C. b) Most reactions did not reach completion, instead saturating between 60 and 100% yield. Yield was determined based on positive control experiments where target and probe were thermally annealed (Supplementary Section 1). We modelled incompleteness of hybridization as a "bad fraction" of probes that becomes kinetically trapped at a high fluorescence state. The best-fit bad fractions for the 96 targets at 37 and 55 °C are plotted. c) There appears to be no correlation between $k_{\text{Hyb}}$ and asymptotic yield. Blue and red dots show experiments at 37 °C and at 55 °C, respectively. d) Systematically designed target/probe sequences included 13 clusters, each comprising three targets. Within each cluster, the target sequences were shifted such that predicted secondary structure is present: (1) near the 5' end (plus symbols); (2) near the middle (circles); or (3) near the 3' end of the target (triangles). In 8 of 13 clusters, the target with structure in the middle was slowest. e) Because secondary structures are known to slow down kinetics, we examined the target/probe pairs in which both the target and the probe had a predicted ensemble (partition function) standard free energy ($\Delta G^\circ_{\text{Hyb}} > -3 \text{ kcal mol}^{-1}$) of greater than $-3 \text{ kcal mol}^{-1}$ at the experimental hybridization temperature, indicating minimal structure. At 37 °C and 55 °C, 29 of 100 reactions and 61 of 96 reactions satisfied this criterion, respectively. These reactions all have $k_{\text{Hyb}} \geq 1 \times 10^4 \text{ M}^{-1} \text{s}^{-1}$, but $k_{\text{Hyb}}$ values range over more than one order of magnitude. f) Minimal-structure targets exhibit significant variability of bad fraction, ranging between 0 and 25%.
One commonly held belief in the field is that the predicted secondary structure in the DNA target and probe sequences is highly inversely correlated with hybridization rate constants. We found this to be partially true: when the WNV model is constrained to the selection of only a single feature, the nGp feature (denoting the predicted ΔG° of the probe oligonucleotide based on Nupack at the hybridization temperature/buffer) emerged as the single best predictor of kHyb (Fig. 5a). Prediction using only nGp was accurate to within a factor of 2 for 61% of reactions and within a factor of 3 for 79% of reactions. However, prediction accuracy can be significantly improved by including more features in the WNV model.

Figure 5b,c shows the prediction accuracy of the best three-feature WNV model and the final six-feature WNV model. The six-feature WNV model is significantly better at prediction than the one-feature and three-feature models, with 80% accuracy within a factor of 2 and 91% accuracy within a factor of 3. The six features used were nGp, Pap, Temp, wPat, GavgMSR1 and Gb, with respective feature weights of 7.96, 15.12, 10.55, 4.44, 10.90 and 18.69 (Supplementary Section 4). Although nGp was the best single feature when considered in isolation, in the six-feature model its weight is the second smallest. This observation potentially suggests that the other features collectively hold information that overlaps with nGp.

To help the research community predict hybridization rate constants for DNA oligo probes and primers, we have constructed a web-based software tool, available at http://nablab.rice.edu/nabs/tools/kinetics. The software typically completes predicting kHyb within 30 s. It is currently seeded with the 210 hybridization experiment results performed in this Article and will be updated with additional hybridization experiment results in the future.

**Enrichment from human genomic DNA**

The human genome is over 3 billion nucleotides long, but the coding regions that form the exome collectively only span 1% of the genome. Within the 20,000 genes of the exome, typically there are only between 10 and 400 that are relevant to any particular disease. NGS\textsuperscript{2,23} is the preferred way to perform highly multiplexed analysis of many different DNA sequences within a sample. In NGS, anywhere between 1 million and 1 billion molecules are randomly sampled and the identities of the first 150–300 nt of each molecule...
The median probes (6-30 nt) are often used for these targeted phase enrichment using highly multiplexed hybridization by synthetic DNA oligonucleotide probes is desirable to sample multiple molecules (high read depth). Solid-phase enrichment using highly multiplexed hybridization by synthetic DNA oligonucleotide probes is often used for these targeted sequencing applications.

Current commercial multiplex hybrid-capture panels generally use a very large number of synthetic probe oligonucleotides to fully tile or overlap-tile the genomic regions of interest (for example, 200,000 probes for whole-exome enrichment). Due to the large number of oligo species involved, the concentration of each species is thus necessarily quite low (tens of micromolar), resulting in hybrid-capture protocols that typically last at least 4 h and more frequently more than 16 h. Because of the varying hybridization kinetics of different probes (Fig. 3d), it is likely that many probes do not contribute significantly to hybridization yield and in fact slow down the hybrid-capture process by forcing lower concentrations of the fast-hybridizing probes.

To experimentally test this possibility, we first applied our hybridization rate constant prediction algorithm to all possible 36 nt probes to exon regions of 21 genes. Because the exon regions are typically 3,000 nt long, this corresponds to roughly 3,000 possible probes per gene. Predicted rate constants typically range over about two orders of magnitude (Supplementary Fig. 5), with the fast (≥95th percentile) probes typically being a factor of 3 faster than median probes (~20th percentile). NGS hybrid-capture enrichment typically uses probes longer than 36 nt (for example, Agilent SureSelect uses 120 nt probes), but there is probably a similar if not greater range of hybridization rate constants for longer probes due to the greater possibility of secondary structure and nonspecific interactions.

Next, we picked a total of 65 fast probes and 65 median probes across the exon regions of 21 different cancer-related genes. The expectation was that after a 24 h hybridization protocol, the fast and median probes would produce similar reads, but with a short 20 min hybridization protocol, the fast probes would exhibit significantly greater reads than median probes (Fig. 6a). Our library preparation protocol is summarized in Fig. 6b. All 130 probes are hybridized to the adaptor-ligated DNA simultaneously. However, the number of reads aligned to a particular probe is not directly proportional to its hybridization yield, due to well-documented sequencing bias. For example, some adaptor-ligated amplicons exhibit significant secondary structure and are less efficiently PCR-amplified during normalization, or less efficiently sequenced due to lower flow cell binding efficiency. For this reason, 15 fast and 15 median probes targeting four genes resulted in less than 100× sequencing depth and were excluded from subsequent analysis (Supplementary Section 5). We do not believe this affects the conclusions from our genomic DNA enrichment study.

Our comparison of reads for the 20 min hybridization library and the 24 h hybridization library indicates that the probes predicted to be fast on average exhibited both a twofold increase in reads in the 20 min library, and a twofold increase in the ratio of reads at 20 min versus 24 h. This is slightly worse than our algorithm’s predicted threefold difference between median and fast probes, but understandable given that our rate constant prediction algorithm was trained on single-plex hybridization rather than on multiplex hybridization. Our calibration experiments (Supplementary Section 5) indicate that the correlation constant between single-plex and multiplex $k_{hyb}$ values is approximately $r^2 = 0.6$.

Our results thus suggest that sparse hybrid-capture enrichment panels would produce faster kinetics at a significantly lower cost. Rather than fully tiling or overlap-tiling the genetic regions of interest, it would be better to use a higher concentration of a few probes with the fastest hybridization kinetics. Multiple probes are only needed insofar as biological genomic DNA may be fragmented and a different probe is needed to capture each fragment. With the notable exception of cell-free DNA, most genomic DNA from clinical samples is longer than 500 nucleotides.

**Figure 5 | Prediction accuracy of the WNV model using different numbers of features.**

(a) Prediction using a single feature, nGp, denoting the ensemble (partition function) standard free energy of the probe, as predicted by Nupack under the reaction conditions of interest. Top: Distribution of prediction error for $k_{hyb}$ in log10. Bottom: Predicted versus observed $k_{hyb}$ values. Each blue dot plots the predicted log10($k_{hyb}$) value versus the experimentally observed log10($k_{hyb}$) value for a single hybridization experiment. Each prediction was performed using a standard leave-one-out (LOO) approach: each $k_{hyb}$ prediction is based on 209 labelled instances (all reactions except the one to be predicted). The feature weights trained on all 210 data points (see Supplementary Section 3 for more details).

(b) Prediction using a three-feature WNV model, including nGp, Pap and temperature.

(c) Prediction using the final six-feature model.

Observed log10($k_{hyb}$), Predicted log10($k_{hyb}$), RMSE = 0.464, F3acc = 0.790. Prediction using a single feature, nGp, denoting the ensemble (partition function) standard free energy of the probe, as predicted by Nupack under the reaction conditions of interest. Top: Distribution of prediction error for $k_{hyb}$ in log10. Bottom: Predicted versus observed $k_{hyb}$ values. Each blue dot plots the predicted log10($k_{hyb}$) value versus the experimentally observed log10($k_{hyb}$) value for a single hybridization experiment. Each prediction was performed using a standard leave-one-out (LOO) approach: each $k_{hyb}$ prediction is based on 209 labelled instances (all reactions except the one to be predicted). The feature weights trained on all 210 data points (see Supplementary Section 3 for more details).

Comparing the results of different hybridization protocols shows that the fast probes exhibited significantly greater reads than median probes (Fig. 6a). Our library preparation protocol is summarized in Fig. 6b. All 130 probes are hybridized to the adaptor-ligated DNA simultaneously. However, the number of reads aligned to a particular probe is not directly proportional to its hybridization yield, due to well-documented sequencing bias. For example, some adaptor-ligated amplicons exhibit significant secondary structure and are less efficiently PCR-amplified during normalization, or less efficiently sequenced due to lower flow cell binding efficiency. For this reason, 15 fast and 15 median probes targeting four genes resulted in less than 100× sequencing depth and were excluded from subsequent analysis (Supplementary Section 5). We do not believe this affects the conclusions from our genomic DNA enrichment study.

Our comparison of reads for the 20 min hybridization library and the 24 h hybridization library indicates that the probes predicted to be fast on average exhibited both a twofold increase in reads in the 20 min library, and a twofold increase in the ratio of reads at 20 min versus 24 h. This is slightly worse than our algorithm’s predicted threefold difference between median and fast probes, but understandable given that our rate constant prediction algorithm was trained on single-plex hybridization rather than on multiplex hybridization. Our calibration experiments (Supplementary Section 5) indicate that the correlation constant between single-plex and multiplex $k_{hyb}$ values is approximately $r^2 = 0.6$.

Our results thus suggest that sparse hybrid-capture enrichment panels would produce faster kinetics at a significantly lower cost. Rather than fully tiling or overlap-tiling the genetic regions of interest, it would be better to use a higher concentration of a few probes with the fastest hybridization kinetics. Multiple probes are only needed insofar as biological genomic DNA may be fragmented and a different probe is needed to capture each fragment. With the notable exception of cell-free DNA, most genomic DNA from clinical samples is longer than 500 nucleotides.
Hybridization experiments across 100 biological target sequences, we experimentally characterized the kinetics of 210 instances that is representative of the diversity of genomic DNA. The concentrations of the probes used for this study were intentionally selected to be 50 pM per probe so as to be similar to probe concentrations in commercial enrichment kits. At 50 pM concentrations, up to 200,000 probes can be used and the total oligo concentration would still be at a reasonable 10 µM. At the significantly (for example, 10x) higher individual probe concentrations that become feasible with a sparse coverage of target genetic regions, even the 20 min allotted here for hybridization could be further reduced, greatly speeding up the enrichment workflow from the current practice of 4–24 h.

**Discussion**

Here, the WNV framework and the rational design of features were combined with computational optimization of feature selection and feature weights, resulting in a final model that is capable of accurately predicting hybridization kinetics rate constants based on sequence and temperature information. The WNV model is highly scalable and easily incorporates new experimental data to provide improved predictions, without requiring model retraining. With every additional hybridization experiment and its accompanying fitted $k_{hyb}$ value, the six-dimensional feature space becomes denser, ensuring that on average a new hybridization experiment will be closer to an existing labelled instance.

To seed the model with a reliable initial database of labelled instances that is representative of the diversity of genomic DNA sequences, we experimentally characterized the kinetics of 210 hybridization experiments across 100 biological target sequences using fluorescence. The X-probe architecture allowed us to economically study kinetics for a reasonably large number of target sequences, but extra nucleotides of the universal arms may cause hybridization kinetics to differ slightly from that of a standard single-stranded probe. For example, there may be a systematic bias towards lower rate constants because of the reduced diffusion for the single-stranded probe. For example, there may be a systematic bias towards lower rate constants because of the reduced diffusion constants. Nonetheless, because all targets/probes use the same universal arm sequences, it is likely that the relative ordering of rate constants is preserved.

In this work we started with over 50 rationally designed features that we eventually pruned down to 6 in the final model. The high LOO validation accuracy of the WNV model indicates that these features capture a significant, if not majority, portion of the complexity of the hybridization process. Simultaneously, there remain pairs of experiments in our database with similar feature values but significantly different $k_{hyb}$ values. This implies the existence of undiscovered features that would distinguish these pairs of experiments; additional insight and creativity from the community in designing additional features would be welcomed.

The hybridization reactions experimentally characterized in the work were all performed in 5× PBS buffer and all target and probe sequences were 36 nt long. These experiment constraints were designed to reduce the diversity of hybridization reactions, to ease the training of the WNV model. We plan to expand our experimental studies to vary these conditions, to allow the WNV model to accurately account for buffer conditions and probe lengths. We suspect that longer DNA target/probe systems will exhibit even

**Figure 6 | Comparison of probes predicted to possess median versus fast hybridization kinetics for enrichment from human genomic DNA.**

(a) Diagram showing the workflow of the enrichment process. The hybridization reactions experimentally characterized in the work were all performed in 5× PBS buffer and all target and probe sequences were 36 nt long. These experiment constraints were designed to reduce the diversity of hybridization reactions, to ease the training of the WNV model. We plan to expand our experimental studies to vary these conditions, to allow the WNV model to accurately account for buffer conditions and probe lengths. We suspect that longer DNA target/probe systems will exhibit even

(b) Comparison of probes predicted to possess median versus fast hybridization kinetics for enrichment from human genomic DNA. The concentrations of the probes used for this study were intentionally selected to be 50 pM per probe so as to be similar to probe concentrations in commercial enrichment kits. At 50 pM concentrations, up to 200,000 probes can be used and the total oligo concentration would still be at a reasonable 10 µM. At the significantly (for example, 10x) higher individual probe concentrations that become feasible with a sparse coverage of target genetic regions, even the 20 min allotted here for hybridization could be further reduced, greatly speeding up the enrichment workflow from the current practice of 4–24 h.

**Discussion**

Here, the WNV framework and the rational design of features were combined with computational optimization of feature selection and feature weights, resulting in a final model that is capable of accurately predicting hybridization kinetics rate constants based on sequence and temperature information. The WNV model is highly scalable and easily incorporates new experimental data to provide improved predictions, without requiring model retraining. With every additional hybridization experiment and its accompanying fitted $k_{hyb}$ value, the six-dimensional feature space becomes denser, ensuring that on average a new hybridization experiment will be closer to an existing labelled instance.

To seed the model with a reliable initial database of labelled instances that is representative of the diversity of genomic DNA sequences, we experimentally characterized the kinetics of 210 hybridization experiments across 100 biological target sequences using fluorescence. The X-probe architecture allowed us to economically study kinetics for a reasonably large number of target sequences, but extra nucleotides of the universal arms may cause hybridization kinetics to differ slightly from that of a standard single-stranded probe. For example, there may be a systematic bias towards lower rate constants because of the reduced diffusion constants. Nonetheless, because all targets/probes use the same universal arm sequences, it is likely that the relative ordering of rate constants is preserved.

In this work we started with over 50 rationally designed features that we eventually pruned down to 6 in the final model. The high LOO validation accuracy of the WNV model indicates that these features capture a significant, if not majority, portion of the complexity of the hybridization process. Simultaneously, there remain pairs of experiments in our database with similar feature values but significantly different $k_{hyb}$ values. This implies the existence of undiscovered features that would distinguish these pairs of experiments; additional insight and creativity from the community in designing additional features would be welcomed.

The hybridization reactions experimentally characterized in the work were all performed in 5× PBS buffer and all target and probe sequences were 36 nt long. These experiment constraints were designed to reduce the diversity of hybridization reactions, to ease the training of the WNV model. We plan to expand our experimental studies to vary these conditions, to allow the WNV model to accurately account for buffer conditions and probe lengths. We suspect that longer DNA target/probe systems will exhibit even...
more variability in hybridization kinetics; conversely, shorter DNA binding (for example, 10 nt) may exhibit less variability in $k_{\text{hyb}}$. Additionally, with genomic DNA targets, the long-range secondary structure and the fragmentation pattern of genomic DNA targets should also be considered. New features will probably be needed for such expanded models.

Multiplex hybrid-capture panels for enriching target regions from genomic DNA are commonly used in targeted sequencing for scientific and clinical studies. In the absence of reliable kinetics prediction software, researchers and companies have taken a brute-force probe design approach, using fully tiled or overlapping-tiled probes to cover genetic loci of interest. Although this approach ensures the presence of at least some fast-binding probes, it is both expensive (in terms of synthesis and quality control of thousands of probes) and results in slower workflows. Accurately predicting multiplexed hybridization kinetics will enable precision design of sparse, high-performance probe panels for target enrichment.

Data availability. Sequences used for all experiments are provided in the Supplementary Data Tables. Raw fluorescence traces are plotted in the Supplementary Information and exact numerical data are available upon request. Calculated feature values are provided in the Supplementary Information.

Software availability. We have constructed a web-based software tool, available at http://nablab.rice.edu/nabtools/kinetics, that computes the predicted rate constant of a hybridization reaction given the sequence and temperature. The software typically completes the prediction of $k_{\text{hyb}}$ within 30 s. It is currently seeded with the 210 hybridization experiment results performed in this Article and will be updated with additional hybridization experiment results in the future.

Received 31 October 2016; accepted 21 September 2017; published online 6 November 2017

References
1. Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. Science 286, 950–952 (1999).
2. Kornberg, A. & Baker, T. A. DNA Replication (Freeman, 1992).
3. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120, 15–20 (2005).
4. Izkovit, S. & van Oudenaarden, A. Validating transcripts with probes and imaging technology. Nat. Methods 8, S12–S19 (2011).
5. Lockhart, D. J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. 14, 1675–1680 (1996).
6. Guninie, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27, 182–189 (2009).
7. Khodakov, D., Wang, C. & Zhang, D. Y. Diagnostics based on nucleic acid sequence variant profiling: PCR, hybridization, and NGS approaches. Adv. Drug Delivery Rev. 105, 3–19 (2016).
8. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 288, 911–940 (1999).
9. SantaLucia, J. & Hicks, D. The thermodynamics of DNA structural motifs. Annu. Rev. Biochem. 33, 415–440 (2004).
10. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31, 3406–3415 (2003).
11. Zadeh J. N. et al. NUPACK: analysis and design of nucleic acid systems. J. Comput. Chem. 32, 170–173 (2011).
12. Morrison, L. E. & Stols, B. M. Sensitive fluorescence-based thermodynamic and kinetic measurements of DNA hybridization in solution. Biochemistry 32, 3095–3104 (1993).
13. Reynaldo, L. P., Vologodskii, A. V., Neri, B. P. & Lyamichev, V. I. The kinetics of oligonucleotide replacements. J. Mol. Biol. 297, 511–520 (2000).
14. Zhao, D. Y. & Winfree, E. Control of DNA strand displacement kinetics using toehold exchange. J. Am. Chem. Soc. 131, 17303–17314 (2009).
15. Outridge, T. E., Sulc, P., Romano, F., Doye, J. P. & Louis, A. A. DNA hybridization kinetics: zippering, internal displacement and sequence dependence. Nucleic Acids Res. 41, 8886–8895 (2013).
16. Schreck, J. S. et al. DNA hairpins destabilize duplexes primarily by melting rather than by inhibiting hybridization. Nucleic Acids Res. 43, 6181–6190 (2015).
17. Cisse, I., Kim, H. & Ha, T. A rule of seven in Watson–Crick base-pairing of mismatched sequences. Nat. Struct. Mol. Biol. 19, 623–627 (2012).
18. Jungmann, R. et al. Single-molecule kinetics and super-resolution microscopy by fluorescence imaging of transient binding on DNA oligomers. Nano Lett. 10, 4756–4761 (2010).
19. He, G., Li, J., Qi, H., Cui & Guo. X. Direct measurement of single-molecule DNA hybridization dynamics with single-base resolution. Angew. Chem. Int. Ed. 55, 9036–9040 (2016).
20. Wang, J. S. & Zhang, D. Y. Simulation-guided DNA probe design for consistently ultra-specific hybridization. Nat. Chem. 7, 545–553 (2015).
21. Gao, Y., Wolf, L. K. & Georgiadis, R. M. Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. Nucleic Acids Res. 34, 3370–3377 (2006).
22. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. Trends Genet. 30, 418–426 (2014).
23. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and is impact on genomics. Cell 155, 27–38 (2013).
24. Chalambakis, C. S. & et al. Performance comparison of four exome capture systems for deep sequencing. BMC Genomics 15, 449 (2014).
25. Clark, M. M. et al. Performance comparison of exome DNA sequencing technologies. Nat. Biotechnol. 29, 908–914 (2011).
26. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure. J. Cell-free DNA comprises an in vivo nucleosome footprint that informs is tissues-of-origin. Cell 164, 57–68 (2016).
27. Denoeux, T. A k-nearest neighbor classification rule based on Dempster–Shafer theory. IEEE Trans. Syst. Man Cybern. 25, 804–813 (1995).
28. Wand, M. P. & Jones, M. C. Kernel Smoothing (CRC Press, 1994).

Acknowledgements
The authors thank S.X. Chen for assistance with NGS sequence alignment. This work was funded by National Institutes of Health grant R01HG008752 to D.Y.Z.

Author contributions
J.X.Z., I.R.W. and D.Y.Z. conceived the project. J.X.Z. and A.W.Z. performed the experiments. N.D. and A.P. performed hybridization rate constant prediction. J.X.Z., L.R.W. and D.Y.Z. performed MLR model construction and optimization. N.D., B.Y. and R.P. performed MLR model construction and optimization. D.Y.Z. wrote the manuscript with input from all authors.

Additional information
Supplementary information is available in the online version of the paper. Reprints and permissions information is available online at www.nature.com/reprints. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to D.Y.Z.

Competing financial interests
There is a patent pending on the X-probes used in this work, and a patent pending on the WNV model of hybridization rate constant prediction.