Automated Design of Probes for rRNA-targeted Fluorescence *In Situ* Hybridization (FISH) Reveals the Advantages of Dual Probes for Accurate Identification

Erik S. Wright (eswright@wisc.edu), L. Safak Yilmaz, Andrew M. Corcoran, Hatice E. Ökten, and Daniel R. Noguera (noguera@engr.wisc.edu)

Supplementary Information

Formamide Dissociation Profiles for Model Development. To build a dataset of perfect-match formamide dissociation profiles for model development, 36 formamide curves from a previous study (1) were directly used and 70 others were added by determining the formamide series of probes designed and tested for accessibility of target molecules in other studies (2, 3) (see Table S1 for the full list). Of the additional probes, 2 probes targeted *Escherichia coli* and 68 targeted one of four organisms studied in Okten et al. (3). Probe-conferred brightness was calculated using the mode of smoothed fluorescence histograms for *E. coli* probes (1) and the mean of a best-fitted normal distribution for others (3). The minor difference between the intensities calculated with the two methods was ignored since all profiles were normalized before curve-fitting.

Curve-fitting and Statistics. Previously established methods were used for curve-fitting during model development (4). In brief, parameters in Table S2 were obtained by best-fitting Equations S1 to S3 (see below) to normalized experimental profiles using NLINFIT routine of MATLAB Statistics Toolbox. Before curve-fitting, values that were less than 75% of the maximum signal intensity appearing at lower formamide concentrations were removed from each profile to prevent the influence of kinetic effects at low formamide conditions (4, 5). A specific γ-factor was used for each probe to align theoretical and experimental profiles on the vertical axis. Since this factor does not change the shape or the horizontal position of the theoretical curves, it does not affect curve-fitting, except for the loss of one degree of freedom for each probe's formamide profile, and is useful for minimizing the adverse effects of fluctuations due to kinetic factors and experimental uncertainty at the plateau of formamide profiles, as discussed in detail by Yilmaz et al. (4).

Updated mechanistic model of FISH. Mathematical modeling of FISH aims to predict hybridization efficiency, which is the ratio of probe-bound rRNA molecules ($[PR]$) to total rRNA ($[R]_o$) at equilibrium. This term is thermodynamically related to an overall equilibrium constant ($K_{overall}$) by Eq. S1, where $[P]_o$ represents the molar concentration of probe used (5). The overall equilibrium constant is further defined as a function of three Gibbs free energy changes (ΔG^o_i; i=1,2,3) representing the reactions for probe-target duplex formation, probe folding, and target folding, respectively. In addition, a linear free energy model is used to describe each ΔG^o value as a function of formamide concentration (1), thus giving the thermodynamic relationship in Eq. S2, where R is the ideal gas law constant ($1.99X10^{-3}$ kcal/mol K), and T is the hybridization temperature, typically 319.15 K (i.e., 46°C). In this equation, free energy values predicted for a hybridization buffer without formamide ($\Delta G^{o}_{i,0\%}$; i=1,2,3) are linearly adjusted to a given formamide concentration ([*FA*], typically expressed in percent by volume) using parameters termed *m*-values (1, 6). Therefore, hybridization efficiency can be obtained as a function of formamide using Eq. S1 and S2 with predicted free energy values and known probe concentration. Further, hybridization efficiency can be matched with experimentally obtained fluorescence intensity values using Eq. S3, where γ is a proportionality constant. Thus, model calibration depends on finding the best *m*-values that maximize the agreement between experimental and theoretical profiles of probe hybridization as a function of formamide concentration (1).

$$\frac{[PR]}{[R]_o} = \frac{[P]_o K_{overall}}{1+[P]_o K_{overall}} \tag{S1}$$

$$K_{overall} = \frac{\exp(-\frac{\Delta G^0_{1,0\%} + m_1[FA]}{RT})}{\left[1+\exp(-\frac{\Delta G^0_{2,0\%} + m_2[FA]}{RT})\right]\left[1+\exp(-\frac{\Delta G^0_{3,0\%} + m_3[FA]}{RT})\right]} \tag{S2}$$

$$\frac{I}{I_{\max}} = \gamma \frac{[PR]}{[R]_o} \tag{S3}$$

The initial model of formamide dissociation was calibrated using 27 probes and validated by 9 additional probes, all targeting *E. coli* (1). In this study, we expanded the model training set to a total of 106 probes by using five different organisms: *E. coli* (38 probes), *Rhodobacter sphaeroides* (10 probes), *Bacillus subtilis* (19 probes), *Saccharomyces* *cerevisiae* (19 probes), and *Stenotrophomonas maltophilia* (20 probes) (see full probe list in Table S1). In the original study, m_1 and m_3 values (Eq. S2) were defined as linear functions of probe length and ΔG^o_3, respectively, and the coefficients in the definition of these *m*-values were obtained by best-fitting the theoretical profiles to experimental ones. The other *m*-value, m_2, did not affect the fits significantly, as ΔG^o_2 was a positive or slightly negative value (i.e., probe self-complementarity was minimized), and therefore, it was assumed to be a constant consistent with the thermodynamics of DNA denaturation with formamide. The best-fitting *m*-values obtained with the new training set of 106 probes are shown in Table S2 in the retrained mechanistic model (RMM) column. The coefficients of *m*-value equations in RMM were similar to those in the original model and remained within the 95% confidence intervals originally provided, thereby validating the original model's accuracy with additional organisms. Profiles with the retrained model are provided in Figure S1.

We used two cross-validation tests to evaluate the actual predictive ability of our models. The first test was named leave-one-probe-out cross-validation (LOPOCV), and derived the predictive error for a probe by leaving it out of the training set. In this case the model training was performed with the remaining 105 probes, and the resulting model (a slightly different version of the actual model best-fitted to all 106 probes) was tested with the probe left out, which now served as an independent observation. The second validation was similar except that all probes targeting one of the five organisms were left out at once as the independent test set and the training was done with the remaining four organisms. This test, termed leave-one-organism-out cross-validation (LOOOCV), is more conservative since it leaves out a significant fraction of probes. Three statistical parameters from cross validation were used for comparing the predictive power of models. These are sum of squared errors (ε^2), coefficient of determination (R^2), and the average absolute deviation from the experimental melting point ($err([FA]_m)$), and they correspond to the parameters in curve fitting that are used to assess the goodness of new fits (Table S2). In addition, CV offsets, defined as the difference in melting point between best-fit and cross-validation curves (Fig. S2A), provided a metric of model convergence, with small CV offsets reflecting better convergence of the model to a robust set of best-fit parameters.

A comparison of RMM with the original model showed improvements in cross-validation ε^2 and R^2, although the average error in the prediction of melting points was higher. This indicated the new model is trained better to capture the shape of experimental curves, while the ability to predict the melting points was slightly lowered in return. Furthermore, large CV offsets indicated that RMM did not converge despite the use of a large dataset with five organisms. Specifically, 42 probes had CV offsets greater than 2% *FA* and 10 probes had CV offsets greater than 5% *FA* in LOOOCV. Therefore, we searched for other modeling alternatives to reduce the uncertainty in future predictions for other organisms, as described next.

A single-reaction model with a FISH-specific free energy definition for improved predictions. In a recent study with DNA microarray probes (4), we developed a highly predictive equilibrium formamide denaturation model by defining specific nearest neighbor parameters for the prediction of duplex free energy (here, ΔG^{o}_{1}). The other free energy changes that are related to probe and target structure were not included, and therefore, the simulation was driven by a single reaction. The nearest-neighbor rules were obtained as additional modeling parameters by curve-fitting, and they presumably characterized the average thermodynamic contribution of each nearest neighbor to all types of molecular interactions (probe-target, probe-probe, and target-target) within a unified reaction. We therefore reasoned that a set of FISH-specific rules that can capture formamide-driven nucleic acid interactions in FISH might also be effective and developed the nearest-neighbor model in Table S2. This was accomplished by first eliminating the terms for ΔG^{o}_{2} and ΔG^{o}_{3} from Eq. S2 and then deriving the formamide curves with ΔG^{o}_{1} alone. In this case ΔG^{o}_{1} was defined by a new set of best-fitted FISH-specific nearest neighbor parameters (Table S3). This model outperformed RMM in all three statistics of curve-fitting as seen in Table S2. However, the larger CV offsets indicated that this was not a converged model, and likely meant the model was over-fitted by the addition of 17 parameters (Table S3).

To solve the over-parameterization problem, we exploited an observed linear relationship (Figure S3) between fitted FISH-specific nearest neighbor free energies and the original *in solution* DNA/RNA parameters (7) to develop the single-reaction model (SRM) in Table S2. Instead of 17 nearest-neighbor parameters, SRM uses only two parameters that linearly convert the predicted $\Delta G^o{}_1$, calculated with *in solution* nearest neighbor parameters, to a FISH-specific free energy change. This linear free energy transformation in SRM (Table S2) results in SRM significantly reducing the magnitude of the $\Delta G^o{}_1$ value before its use in the simulation of formamide denaturation. In theory, this reduction reflects an average free energy penalty added to take into account the competition between duplex formation (DNA-RNA interactions) and target structure (RNA-RNA interactions), which is no longer represented directly in the model because of the elimination of $\Delta G^o{}_3$.

The SRM model was slightly worse than the nearest-neighbor model in curve-fitting criteria (but better than RMM) due to the lower number of parameters, but it was better in all aspects during cross-validation (Table S2). A particular strength of SRM is that it is a converged model, as indicated by small CV offsets, with none of the probes showing CV offsets greater than 2% *FA* in LOPOCV or greater than 5% *FA* in LOOOCV. Furthermore, SRM is the best model in predicting the melting point, with the error being less than 10% *FA* for the majority of probes (Fig. S2C), and it also captures the slope of formamide profiles better than RMM (Figs. S1 and S2B). As a final attempt in modeling, to determine whether the inclusion of $\Delta G^o{}_3$ also in a linearly transformed fashion would further improve predictions, we developed a double-reaction model

(Table S2). Since cross-validation tests showed no additional benefit of using this model, we chose SRM as the new predictor of formamide dissociation profiles for perfectly matched probes.

Using single-reaction model (SRM) as a predictor of mismatch stability. We compared RMM and SRM as potential predictors of formamide melting points for mismatched targets. Unlike with perfect-match duplexes, the available datasets are insufficient to develop mature mathematical models of dissociation profiles for mismatched hybrids due to the sheer number of different mismatch loops to be tested for a good coverage of mismatch thermodynamics. Since mismatch stability is a function of the internal loop created in the duplex structure and the adjacent base pairs (8), there are 192 different permutations for single mismatches alone. This number rises considerably when taking into account tandem mismatches, insertions, and deletions, thus requiring hundreds of probes for deriving reliable thermodynamic parameters as was recently done with DNA microarrays (4). For these reasons we sought a semi-quantitative and conservative mismatch predictor for FISH.

Previously, we generated a FISH dataset of 35 probes with a single mismatch to *E. coli* 16S rRNA by inserting different mismatches in 7 perfect-match parent probes (9). This dataset was useful for systematically evaluating different predictors of mismatch stability and for demonstrating that DNA/RNA mismatch effects could be approximated by averaging DNA/DNA and RNA/RNA parameters for mismatch loops.

The effect of a mismatch in FISH was measured with a metric called $\Delta[FA]_m$, which describes the decrease in melting point upon the insertion of a mismatch (i.e., the distance between formamide profiles of the perfect-match probe and the probe with a mismatch). In this study, we used the same metric (reversed in sign to maintain modeling convention (10)) and experimental dataset to evaluate the SRM and RMM models. As shown in Figure S4, the correlations between theoretical and experimental values were similar, with RMM giving slightly better predictions as indicated by a higher R^2. SRM also systematically underestimated the magnitude of $\Delta[FA]_m$ (Fig. S4A). This systematic reduction is apparently due to the lack of the free energy penalty for unfolding the target structure (ΔG^o_3), which is used by RMM but not SRM (Table S2). Considering RMM and SRM's comparable performance with the mismatched dataset, we selected SRM as the predictor for mismatched non-targets to take advantage of its substantially faster computation, because it does not require calculation of ΔG^o_3.

The previously recommended mismatch predictor for FISH was $\Delta\Delta G^o_1$, which represents the difference in ΔG^o_1 caused by mismatches (9). Consistently, the predictor in this study, $\Delta[FA]_{m,SRM}$, is a function of only $\Delta\Delta G^o_1$ and probe length as input variables, and it has a better correlation with data ($R^2 = 0.64$; Figure S4A) than $\Delta\Delta G^o_1$ alone ($R^2 = 0.59$; (9)). In addition, $\Delta\Delta G^o_1$ can capture the effect of complex mismatched conformations since it takes an average of extensively studied mismatch motifs in DNA/DNA and RNA/RNA duplexes. For instance, the effect of insertions and deletions (indels) and tandem mismatches are readily calculated with this approach. In practice, an observed $\Delta[FA]_m$ of less than -20% (i.e., more negative) shows strong destabilizing effects that make non-targets differentiable from targets by adjusting the stringency of the hybridization buffer (9). An observed $\Delta[FA]_m$ of -20% is approximately equivalent to a predicted value ($\Delta[FA]_{m,SRM}$) of -10% based on the trend line in Figure S4A. Accordingly, we defined qualitative thresholds based on non-target $\Delta[FA]_{m,SRM}$ as follows: Non-targets can be considered at very high risk of hybridization if $\Delta[FA]_{m,SRM} > -5\%$, high risk if $\Delta[FA]_{m,SRM}$ is between -10% and -5%, moderate risk if $\Delta[FA]_{m,SRM}$ is between -15% and -10%, low risk if $\Delta[FA]_{m,SRM}$ is between -20% and -15%, and no risk if $\Delta[FA]_{m,SRM}$ is below -20%.

Dual probe use to increase specificity and confidence level in target identification. A strategy systematically evaluated in this study was the use of two probes targeting the same organism as a means to increase specificity and minimize false-positive hybridizations. In principle, two probes with different labels can be simultaneously employed to identify target organisms based on double positive (overlapping) signal. The confidence in identification depends on the level of predictive errors of dissociation curves during design as well as the success in the minimization of the number of potential false positives (non-targets) for both probes. The latter is systematically handled in our approach by an algorithm that nearly exhaustively searches the space of known non-targets. Thus, unless predictions of melting points for both probes are off by a large margin, double positive results can be taken with a reasonable level of confidence. To statistically evaluate the accuracy of model predictions we derived a probability density function that defines expected predictive errors for the melting point of perfect matches. For this, we used LOOOCV results from model fits to conservatively estimate errors in organisms not included in the training sets and fitted a normal distribution on predictive errors of melting points ($\mathrm{err}[FA]_{\mathrm{m}}^{\mathrm{LOOOCV}}$) as shown in Figure S5. From this distribution it follows that there is a 71% chance that, using our hybridization model, the prediction for a probe is within 10% *FA* of the actual melting point, and therefore, the likelihood of two probes being predicted with less than 10% *FA* error is 50% (0.71 x 0.71 ~ 0.50) assuming that prediction error is independent from probe-to-probe.

The actual position of the predicted probe dissociation curves with respect to experimental observations can be obtained by hybridizing each probe with a pure culture or a mixed community sample, assuming that the community includes the target. Based on the above statistics, we expect to see the experimental dissociation profiles of both probes to be within 10% *FA* of the predicted curves half of the time. In such cases the user can identify the target organism based on double positives with high confidence. The statistical analysis also indicates that in an additional 22% of cases, at least one of the two probes will be underestimating the experimental melting point by more than 10% FA ($\mathrm{err[FA]}_{\mathrm{m}} < -10\%$; the other probe will either be within 10%

of prediction or also underestimate the melting point). Since an experimental profile located to the right of a theoretical profile (i.e., underestimated) is unlikely to be coming from a mismatch, these cases can also provide a reasonable level of confidence for the double positives provided that each probe is hybridized at a formamide concentration close to but less than its melting point. Therefore, in more than two thirds of all cases we expect the users to identify target organisms from double positives with reasonable confidence, without a need for experimental optimization with pure cultures or clones. This is a significant departure from the traditional approach of designing a single probe without mathematical modeling, where confidence is never high without optimization with pure cultures or clones.

It should be noted that in about a third of the above-defined good confidence cases, we expect the dissociation profiles of the two probes to be more than 10% *FA* distant from each other (e.g., when the actual melting point of one probe is 5% to the right of the design *FA* concentration and the other 10% to the left; probe-to-probe distance was evaluated based on one million random pairs of probes picked from the distribution in Figure S5). In such cases, successive hybridizations with two different formamide concentrations (close to the melting point of each probe) will be necessary to obtain double positives with an accurate identification of the target (11).

Detailed description of probe design algorithm. The probe design process (Figure S6) begins with a set of aligned RNA sequences arranged into target and non-target groups, which are user defined and at any taxonomic level. Here we describe the design algorithm using the example of genus-specific probes designed in this study, which used the complete RDP (10.30) database (12) of 2,459,684 bacterial and archaeal 16S rRNA gene sequences. The sequences were classified into 1,943 named genera, screened for chimeras using DECIPHER (13), and then used to form sets of tiled k-mers (27 nucleotides long) from all of the sequences belonging to each genus. These k-mers are overlapping with only one nucleotide separation between the start of adjacent k-mers. The frequency of the k-mer sequences at each site is determined by the fraction of each k-mer permutation in each target site position. The most abundant k-mers are recorded to reach at least 90% coverage of each target group. Use of the most abundant 90% of potential target sites minimizes the possibility of designing probes for spurious sequences (i.e., undetected chimeras and sequencing errors) at the expense of possibly neglecting minor real sequence variations. For the genus-level design described in this study, unclassified groups were neglected, as they are not necessarily phylogenetically coherent.

Once the initial k-mer set is created, additional constraints are implemented to improve the quality of the set to be used for probe design. For the probes designed in this study, sites in the sequence alignment were excluded from consideration if they required more than 4 permutations to represent at least 90% of the target group. Such sites represent variable regions of a sequence in which FISH probes could not provide adequate coverage of the group. Target sites were also eliminated if they were not included in at least 20% of the sequences in the group or were not located between *E. coli* positions 27 and 1,406 in the sequence alignment. These constraints were intended to ensure a holistic representation of sequence diversity within the target group since many sequences in the RDP database do not span the complete alignment.

For each of the remaining k-mers, probe design begins by estimating the required probe length at each k-mer location. Starting with a probe of 17 nucleotides in length and extending the 3'-end base-by-base, the optimal probe length is chosen such that all probe permutations representing a site are predicted with the single-reaction model (SRM) to be above a user-defined hybridization efficiency (HE; default 0.5) at the user-specified formamide concentration [*FA*] (default 35% *FA*). Potential probes are excluded if they cannot be represented by a single consensus sequence. This is a practical constraint that minimizes the cost incurred in the synthesis of multiple probes per target site and also effectively excludes insertions or deletions in target sites.

Four additional constraints are applied to the set of potential probes based on predictions of the recalibrated mechanistic model (RMM). These checks are intended to ensure accessibility of the target site in the RNA and applicability of the new model based on similarity to the probes used in its training set. First, the predicted $[FA]_m$ must be less than 70% *FA* to prevent users from attempting to design probes with overly large affinity that are not well represented in the training set. Second, the $[FA]_m$

predicted by both models must be within 15% for consistency. Third, the free energy of probe folding ($\Delta G^o{}_2$) must be greater than -3 kcal/mol to ensure that this side reaction does not dominate the hybridization. Finally, $\Delta G^o{}_{overall}$ must always be less than -10 kcal/mol (5) to ensure that $[FA]_m$ is calculable (i.e., ≥ 0) using RMM. In practice, these additional constraints eliminate around 10% of potential probes pre-selected with SRM.

Computational efficiency was a consideration in the implementation of the RMM checks. The most time consuming step in RMM is the determination of the stability of intramolecular secondary RNA-RNA interactions within the target molecule ($\Delta G^o{}_3$), which strongly influence target site accessibility (5). $\Delta G^o{}_3$ is calculated by taking the difference in free energy between two simulated foldings of the target domain using UNAFold (14), once with the target site prohibited from base pairing and again with the target site allowed to base pair. To circumvent the computational inefficiency associated with applying RMM to thousands of different sequences in a group, a single unambiguous consensus sequence that represents the most common permutation in each target site is created. To generate this consensus sequence the most abundant k-mers are concatenated into a single sequence that is used for $\Delta G^o{}_3$ calculations. In the case of multiple probe permutations per target site, the k-mer representing the target site is replaced to match the target permutation, and an additional RMM calculation is required for each target site. This process allows the target site to reflect every potential target permutation while the rest of the sequence reflects the majority consensus of the group.

A statistical analysis of this approach indicated only a small difference in the ΔG^{o}_{3} calculations (standard deviation of 1.6 kcal/mol) between the k-mer derived consensus sequence and real sequences when groups were defined at the genus level. The difference increases when groups are defined at higher taxonomic levels, reaching standard deviations of up to 2.7 kcal/mol at the phylum level. Thus, the k-mer derived consensus approach afforded the needed computational efficiency with reasonable accuracy, but higher errors are expected as the sequences in a group become more diverse. Importantly, the difference between the ΔG^{o}_{3} of individual sequences in the group and their group's consensus sequence was very small. The median ΔG^{o}_{3} of individual sequences was on average 1.0 kcal/mol less than that calculated using the consensus approach at the phylum level, while there was no difference at the genus level. Hence there is little practical difference between using the consensus approach and the median of all sequences in the group, except that the consensus approach only requires a single calculation. In order to maintain a high computational efficiency the algorithm only uses the consensus sequence approach in the calculation of ΔG^{o}_{3} when checking constraints with RMM.

After all candidate probes are identified based on the perfect match hybridizations to the target group and the constraints described above, their potential for cross-hybridization with non-target groups is evaluated. This evaluation uses the difference in melting point created by the presence of mismatches, as estimated with SRM ($\Delta[FA]_{\mathrm{m,SRM}}$). The program calculates $\Delta[FA]_{\mathrm{m,SRM}}$ for all probe permutations with each non-target k-mer at the same position in the sequence alignment, and determines the non-target site with the least difference in melt points. All non-target groups with $\Delta[FA]_{\mathrm{m,SRM}} > -20\%$ are recorded as having the potential to cross-hybridize, and included in a specificity score calculation (Eq. S4), which penalizes the number of potential false positives (n) as well as the degree of cross-hybridization ($\Delta[FA]_{\mathrm{m,SRM}}$).

$$specificity\ score = \sum_{i=1}^{n}\left(-0.2 - 1.2^{\Delta[FA]_{m,SRM}}\right) \tag{S4}$$

The specificity score for a probe without any potential cross-hybridization is zero, and a worse probe will have a more negative specificity score. For example, a non-target predicted to cross-hybridize with a $\Delta[FA]_{\mathrm{m,SRM}}$ of -20% would decrease the specificity score by 0.23, whereas a non-target with a $\Delta[FA]_{\mathrm{m,SRM}}$ of 0% would decrease the specificity score by 1.20. In this way, non-targets with a very high chance of cross-hybridizing lower the specificity score by about 5-times more than that of non-targets with a low chance of being false positives. It is worth noting that $\Delta[FA]_{\mathrm{m,SRM}}$ is recorded for each non-target relative to the probe permutation with lowest $[FA]_{\mathrm{m,SRM}}$ predicted for the target group, as this permutation limits the concentration of formamide that can be used in the hybridization buffer. For this reason non-targets sometimes have predicted

$\Delta[FA]_{\mathrm{m,SRM}}$ that are positive, but these are treated as equivalent to 0% in the context of Eq. S4.

The algorithm's output for a FISH probe design is a list of candidate probes ranked by the specificity score. Probes with a score of 0 are ideal designs, without potential cross-hybridizations to any of the non-target groups. For probes with identical specificity scores, the ranking gives preference to those with the lowest number of permutations. The output also provides information on which non-target groups are predicted to have cross hybridizations, so the user can analyze these predictions in order to decide how to proceed with the designed probes. Non-target sites are provided so that the user knows what sequences to target with competitor oligonucleotide probes, if desired.

When the top-scoring probe has a specificity score that signals insufficient specificity, users can select the dual probe approach. In this case, two probes targeting the same group and labeled with different fluorophores would be used, and only the overlap in signal from both probes would constitute a positive identification. When designing dual probes, the algorithm completes an exhaustive search through all combinations of two probes and calculates the specificity score (Eq. S4) by using the minimum $\Delta[FA]_{\mathrm{m,SRM}}$ of the two probes. In addition, each of the two probes must be separated by at least 50 nucleotides to mitigate any potential influence of one probe over the hybridization of the second probe. Similar to the one probe design, the ranking of probes also favors probe sets with the least number of permutations when multiple sets have the same specificity score.

Demonstration of Dual Probe Design

The main text describes an example of dual probe design to detect *Xenorhadbus nematophila*. This example used the design output for dual probes (Figure S7), which resulted in fewer predicted cross-hybridizations than the single probe design (Figure S8).

Analysis of 16S rRNA Probes in probeBase

The main text presents a thermodynamics-based analysis of probes catalogued in probeBase (15). In this analysis, we identified probes that were originally designed to target a specific taxonomic group, but for which our re-evaluation of these probes showed that they had low or moderate coverage of their intended target group (Table S4). In addition, we also identified probes that had low specificity to their intended target group (Table S5). A graphical representation of specificity for probes in probeBase was also constructed (Figure S9). Finally, when applying our probe design approach to the design of phylum-level probes, four designs resulted in probes similar to existing phylum-level probes (Table S6).

REFERENCES

1. **Yilmaz LS**, **Noguera DR**. 2007. Development of thermodynamic models for simulating probe dissociation profiles in fluorescence in situ hybridization. Biotechnol. Bioeng. **96**:349–363.
2. **Yilmaz LS**, **Okten HE**, **Noguera DR**. 2006. Making all parts of the 16S rRNA of

Escherichia coli accessible in situ to single DNA oligonucleotides. Appl. Environ. Microbiol. **72**:733–744.

3. **Okten HE**, **Yilmaz LS**, **Noguera DR**. 2012. Exploring the in situ accessibility of small subunit ribosomal RNA of members of the domains Bacteria and Eukarya to oligonucleotide probes. Systematic and Applied Microbiology 1–11.
4. **Yilmaz LS**, **Loy A**, **Wright ES**, **Wagner M**, **Noguera DR**. 2012. Modeling formamide denaturation of probe-target hybrids for improved microarray probe design in microbial diagnostics. PLoS ONE **7**:e43862.
5. **Yilmaz LS**, **Noguera DR**. 2004. Mechanistic Approach to the Problem of Hybridization Efficiency in Fluorescent In Situ Hybridization. Appl. Environ. Microbiol. **70**:7126–7139.
6. **Myers JK**, **Nick Pace C**, **Martin Scholtz J**. 1996. Denaturant m values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. Protein Sci. **5**:981–981.
7. **Sugimoto N**, **Nakano S-I**, **Katoh M**, **Matsumura A**, **Nakamuta H**, **Ohmichi T**, **Yoneyama M**, **Sasaki M**. 1995. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry **34**:11211–11216.
8. **Sugimoto N**, **Nakano M**, **Nakano S-I**. 2000. Thermodynamics−Structure Relationship of Single Mismatches in RNA/DNA Duplexes. Biochemistry **39**:11270–11281.
9. **Yilmaz LS**, **Bergsven LI**, **Noguera DR**. 2008. Systematic evaluation of single mismatch stability predictors for fluorescence in situ hybridization. Environmental Microbiology **10**:2872–2885.
10. **Yilmaz LS**, **Parnerkar S**, **Noguera DR**. 2011. mathFISH, a Web Tool That Uses Thermodynamics-Based Mathematical Models for In Silico Evaluation of Oligonucleotide Probes for Fluorescence In Situ Hybridization. Appl. Environ. Microbiol. **77**:1118–1122.
11. **Wagner M**, **Amann R**, **Kämpfer P**, **Assmus B**, **Hartmann A**, **Hutzler P**, **Springer N**, **Schleifer K-H**. 1994. Identification and in situ Detection of Gram-negative Filamentous Bacteria in Activated Sludge. Systematic and Applied Microbiology **17**:405–417.
12. **Cole JR**, **Wang Q**, **Cardenas E**, **Fish J**, **Chai B**, **Farris RJ**, **Kulam-Syed-Mohideen AS**, **McGarrell DM**, **Marsh T**, **Garrity GM**, **Tiedje JM**. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research **37**:D141–D145.
13. **Wright ES**, **Yilmaz LS**, **Noguera DR**. 2012. DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. Appl. Environ. Microbiol. **78**:717–725.
14. **Markham NR**, **Zuker M**. 2008. UNAFold: software for nucleic acid folding and hybridization. Methods Mol. Biol. **453**:3–31.

15. **Loy A**, **Maixner F**, **Wagner M**, **Horn M**. 2007. probeBase--an online resource for rRNA-targeted oligonucleotide probes: new features 2007. Nucleic Acids Research **35**:D800–4.

16. **Fuchs BM**, **Wallner G**, **Beisker W**, **Schwippl I**, **Ludwig W**, **Amann R**. 1998. Flow cytometric analysis of the in situ accessibility of Escherichia coli 16S rRNA for fluorescently labeled oligonucleotide probes. Appl. Environ. Microbiol. **64**:4973–4982.

17. **Wallner G**, **Amann R**, **Beisker W**. 1993. Optimizing fluorescent in situ hybridization with rRNA-targeted oligonucleotide probes for flow cytometric identification of microorganisms. Cytometry **14**:136–143.

18. **Nielsen JL**, **Nguyen H**, **Meyer RL**, **Nielsen PH**. 2012. Identification of glucose-fermenting bacteria in a full-scale enhanced biological phosphorus removal plant by stable isotope probing. Microbiology **158**:1818–1825.

19. **Friedrich U**, **Van Langenhove H**, **Altendorf K**, **Lipski A**. 2003. Microbial community and physicochemical analysis of an industrial waste gas biofilter and design of 16S rRNA-targeting oligonucleotide probes. Environmental Microbiology **5**:183–201.

20. **Teira E**, **Reinthaler T**, **Pernthaler A**, **Pernthaler J**, **Herndl GJ**. 2004. Combining Catalyzed Reporter Deposition-Fluorescence In Situ Hybridization and Microautoradiography To Detect Substrate Utilization by Bacteria and Archaea in the Deep Ocean. Appl. Environ. Microbiol. **70**:4411–4414.

21. **Eilers H**, **Pernthaler J**, **Peplies J**, **Glockner FO**, **Gerdts G**, **Amann R**. 2001. Isolation of Novel Pelagic Bacteria from the German Bight and Their Seasonal Contributions to Surface Picoplankton. Appl. Environ. Microbiol. **67**:5134–5142.

22. **Xia Y**, **Massé DI**, **McAllister TA**, **Beaulieu C**, **Talbot G**, **Kong Y**, **Seviour R**. 2011. In situ identification of keratin-hydrolyzing organisms in swine manure inoculated anaerobic digesters. FEMS Microbiology Ecology **78**:451–462.

23. **Kubo K**, **Knittel K**, **Amann R**, **Fukui M**, **Matsuura K**. 2011. Sulfur-metabolizing bacterial populations in microbial mats of the Nakabusa hot spring, Japan. Systematic and Applied Microbiology **34**:293–302.

24. **Böckelmanna U**, **Manza W**, **Neub T**, **Szewzyka U**. 2000. Characterization of the microbial community of lotic organic aggregates (“river snow”) in the Elbe River of Germany by cultivation and molecular methods. FEMS Microbiology Ecology **33**:157–170.

25. **Eilers H**, **Pernthaler J**, **Glockner FO**, **Amann R**. 2000. Culturability and In Situ Abundance of Pelagic Bacteria from the North Sea. Appl. Environ. Microbiol. **66**:3044–3051.

26. **Harmsen H**, **Prieur D**, **Jeanthon C**. 1997. Group-Specific 16S rRNA-Targeted Oligonucleotide Probes To Identify Thermophilic Bacteria in Marine Hydrothermal Vents. Appl. Environ. Microbiol. **63**:4061–4068.

27. **Amann RI**, **Binder BJ**, **Olson RJ**, **Chisholm SW**, **Devereux R**, **Stahl DA**. 1990. Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. Appl. Environ. Microbiol. **56**:1919–1925.

28. **Arnds J**, **Knittel K**, **Buck U**, **Winkel M**, **Amann R**. 2010. Development of a 16S rRNA-targeted probe set for Verrucomicrobia and its application for fluorescence in situ hybridization in a humic lake. Systematic and Applied Microbiology **33**:139–148.

29. **Thayanukul P**, **Zang K**, **Janhom T**, **Kurisu F**, **Kasuga I**, **Furumai H**. 2010. Concentration-dependent response of estrone-degrading bacterial community in activated sludge analyzed by microautoradiography-fluorescence in situ hybridization. Water Research **44**:4878–4887.

30. **Kloep F**, **Manz W**, **RÃ ske I**. 2006. Multivariate analysis of microbial communities in the River Elbe (Germany) on different phylogenetic and spatial levels of resolution. FEMS Microbiology Ecology **56**:79–94.

31. **Bourne D**, **Holmes A**, **Iversen N**, **Murrell J**. 2000. Fluorescent oligonucleotide rDNA probes for specific detection of methane oxidising bacteria. FEMS Microbiology Ecology **31**:29–38.

32. **Neef A**, **Zaglauer A**, **Meier H**, **Amann R**, **Lemmer H**, **Schleifer KH**. 1996. Population analysis in a denitrifying sand filter: conventional and in situ identification of Paracoccus spp. in methanol-fed biofilms. Appl. Environ. Microbiol. **62**:4329–4339.

33. **Hallberg KB**, **Coupland K**, **Kimura S**, **Johnson DB**. 2006. Macroscopic Streamer Growths in Acidic, Metal-Rich Mine Waters in North Wales Consist of Novel and Remarkably Simple Bacterial Communities. Appl. Environ. Microbiol. **72**:2022–2030.

34. **Urzi C**, **La Cono V**, **Stackebrandt E**. 2004. Design and application of two oligonucleotide probes for the identification of Geodermatophilaceae strains using fluorescence in situ hybridization (FISH). Environmental Microbiology **6**:678–685.

35. **Eller G**, **Stubner S**, **Frenzel P**. 2001. Group-specific 16S rRNA targeted probes for the detection of type I and type II methanotrophs by fluorescence in situ hybridisation. FEMS Microbiol. Lett. **198**:91–97.

36. **Salzman NH**, **de Jong H**, **Paterson Y**, **Harmsen HJM**, **Welling GW**, **Bos NA**. 2002. Analysis of 16S libraries of mouse gastrointestinal microflora reveals a large new group of mouse intestinal bacteria. Microbiology (Reading, Engl.) **148**:3651–3660.

37. **Lücker S**, **Steger D**, **Kjeldsen KU**, **MacGregor BJ**, **Wagner M**, **Loy A**. 2007. Improved 16S rRNA-targeted probe set for analysis of sulfate-reducing bacteria by fluorescence in situ hybridization. Journal of Microbiological Methods **69**:523–

528.

38. **Manz W**, **Eisenbrecher M**, **Neu TR**, **Szewzyk U**. 1998. Abundance and spatial organization of Gram-negative sulfate-reducing bacteria in activated sludge investigated by in situ probing with specific 16S rRNA targeted oligonucleotides. FEMS Microbiology Ecology **25**:43–61.

39. **Bojesen AM**, **Christensen H**, **Nielsen OL**, **Olsen JE**, **Bisgaard M**. 2003. Detection of Gallibacterium spp. in Chickens by Fluorescent 16S rRNA In Situ Hybridization. Journal of Clinical Microbiology **41**:5167–5172.

40. **Mahmoud KK**, **McNeely D**, **Elwood C**, **Koval SF**. 2007. Design and Performance of a 16S rRNA-Targeted Oligonucleotide Probe for Detection of Members of the Genus Bdellovibrio by Fluorescence In Situ Hybridization. Appl. Environ. Microbiol. **73**:7488–7493.

41. **Devereux R**, **Kane MD**, **Winfrey J**, **Stahl DA**. 1992. Genus-and group-specific hybridization probes for determinative and environmental studies of sulfate-reducing bacteria. Systematic and Applied Microbiology **15**:601–609.

42. **Brinkmeyer R**, **Knittel K**, **Jurgens J**, **Weyland H**, **Amann R**, **Helmke E**. 2003. Diversity and Structure of Bacterial Communities in Arctic versus Antarctic Pack Ice. Appl. Environ. Microbiol. **69**:6610–6619.

43. **Küsel K**, **Pinkart HC**, **Drake HL**, **Devereux R**. 1999. Acetogenic and sulfate-reducing bacteria inhabiting the rhizoplane and deep cortex cells of the sea grass Halodule wrightii. Appl. Environ. Microbiol. **65**:5117–5123.

44. **Pirttila AM**, **Laukkanen H**, **Pospiech H**, **Myllyla R**, **Hohtola A**. 2000. Detection of Intracellular Bacteria in the Buds of Scotch Pine (Pinus sylvestris L.) by In Situ Hybridization. Appl. Environ. Microbiol. **66**:3073–3077.

45. **Moreno Y**, **Ballesteros L**, **García-Hernández J**, **Santiago P**, **González A**, **Ferrús MA**. 2011. Specific detection of viable Listeria monocytogenes in Spanish wastewater treatment plants by Fluorescent In Situ Hybridization and PCR. Water Research **45**:4634–4640.

46. **Rabus R**, **Fukui M**, **Wilkes H**, **Widdle F**. 1996. Degradative capacities and 16S rRNA-targeted whole-cell hybridization of sulfate-reducing bacteria in an anaerobic enrichment culture utilizing alkylbenzenes from crude oil. Appl. Environ. Microbiol. **62**:3605–3613.

47. **Boscaro V**, **Schrallhammer M**, **Benken KA**, **Krenek S**, **Szokoli F**, **Berendonk TU**, **Schweikert M**, **Verni F**, **Sabaneyeva E**, **Petroni G**. 2013. Unpublished. submitted.

48. **Meisinger DB**, **Zimmermann J**, **Ludwig W**, **Schleifer K-H**, **Wanner G**, **Schmid M**, **Bennett PC**, **Engel AS**, **Lee NM**. 2007. In situ detection of novel Acidobacteria in microbial mats from a chemolithoautotrophically based cave ecosystem (Lower Kane Cave, WY, USA). Environmental Microbiology **9**:1523–1534.

49. **Juretschko S**, **Loy A**, **Lehner A**, **Wagner M**. 2002. The Microbial Community Composition of a Nitrifying-Denitrifying Activated Sludge from an Industrial

Sewage Treatment Plant Analyzed by the Full-Cycle rRNA Approach. Systematic and Applied Microbiology **25**:84–99.

50. **Scheid D**, **Stubner S**. 2001. Structure and diversity of Gram-negative sulfate-reducing bacteria on rice roots. FEMS Microbiology Ecology **36**:175–183.

51. **Gunasekera TS**, **Dorsch MR**, **Slade MB**, **Veal DA**. 2003. Specific detection of Pseudomonas spp. in milk by fluorescence in situ hybridization using ribosomal RNA directed probes. J. Appl. Microbiol. **94**:936–945.

52. **Manz W**, **Amann R**, **Szewzyk R**, **Szewzyk U**, **Stenström TA**, **Hutzler P**, **Schleifer KH**. 1995. In situ identification of Legionellaceae using 16S rRNA-targeted oligonucleotide probes and confocal laser scanning microscopy. Microbiology (Reading, Engl.) **141 (Pt 1)**:29–39.

53. **Figuerola ELM**, **Erijman L**. 2007. Bacterial taxa abundance pattern in an industrial wastewater treatment system determined by the full rRNA cycle approach. Environmental Microbiology **9**:1780–1789.

54. **Kanagawa T**, **Kamagata Y**, **Aruga S**, **Kohno T**, **Horn M**, **Wagner M**. 2000. Phylogenetic Analysis of and Oligonucleotide Probe Development for Eikelboom Type 021N Filamentous Bacteria Isolated from Bulking Activated Sludge. Appl. Environ. Microbiol. **66**:5043–5052.

55. **Felske A**, **Akkermans AD**, **De Vos WM**. 1998. In situ detection of an uncultured predominant bacillus in Dutch grassland soils. Appl. Environ. Microbiol. **64**:4588–4590.

56. **Hugenholtz P**, **Tyson GW**, **Webb RI**, **Wagner AM**, **Blackall LL**. 2001. Investigation of Candidate Division TM7, a Recently Recognized Major Lineage of the Domain Bacteria with No Known Pure-Culture Representatives. Appl. Environ. Microbiol. **67**:411–419.

57. **Neef A**. 1997. Anwendung der in situ-Einzelzell-Identifizierung von Bakterien zur Populationsanalyse in komplexen mikrobiellen Biozönosen. Thesis.

58. **Meier H**, **Amann R**, **Ludwig W**, **Schleifer K-H**. 1999. Specific Oligonucleotide Probes for in situ Detection of a Major Group of Gram-positive Bacteria with low DNA G+C Content. Systematic and Applied Microbiology **22**:186–196.

59. **Mobarry BK**, **Wagner M**, **Urbain V**, **Rittmann BE**, **Stahl DA**. 1996. Phylogenetic probes for analyzing abundance and spatial organization of nitrifying bacteria. Appl. Environ. Microbiol. **62**:2156–2162.

60. **Manz W**, **Amann R**, **Ludwig W**, **Vancanneyt M**, **Schleifer KH**. 1996. Application of a suite of 16S rRNA-specific oligonucleotide probes designed to investigate bacteria of the phylum cytophaga-flavobacter-bacteroides in the natural environment. Microbiology (Reading, Engl.) **142 (Pt 5)**:1097–1106.

61. **Weller R**, **Glöckner FO**, **Amann R**. 2000. 16S rRNA-Targeted Oligonucleotide Probes for the in situ Detection of Members of the Phylum Cytophaga-Flavobacterium-Bacteroides. Systematic and Applied Microbiology **23**:107–114.

62. **Dohlen Von CD**, **Kohler S**, **Alsop ST**, **McManus WR**. 2001. Mealybug β-proteobacterial endosymbionts contain γ-proteobacterial symbionts. Nature **412**:433–436.

63. **K-H S**, **R A**, **W L**, **C R**, **N S**, **S D**. 1992. Pseudomonas: molecular biology and biotechnology. American Society for Microbiology.

64. **Jurgens G**, **Glöckner FO**, **Amann R**, **Saano A**, **Montonen L**, **Likolammi M**, **Münster U**. 2000. Identification of novel Archaea in bacterioplankton of a boreal forest lake by phylogenetic analysis and fluorescent in situ hybridization1. FEMS Microbiology Ecology **34**:45–56.

65. **Lin X**, **Wakeham SG**, **Putnam IF**, **Astor YM**, **Scranton MI**, **Chistoserdov AY**, **Taylor GT**. 2006. Comparison of Vertical Distributions of Prokaryotic Assemblages in the Anoxic Cariaco Basin and Black Sea by Use of Fluorescence In Situ Hybridization. Appl. Environ. Microbiol. **72**:2679–2690.

66. **Schuppler M**, **Wagner M**, **Schön G**, **Göbel UB**. 1998. In situ identification of nocardioform actinomycetes in activated sludge using fluorescent rRNA-targeted oligonucleotide probes. Microbiology (Reading, Engl.) **144 (Pt 1)**:249–259.

67. **Wagner M**, **Haider S**. 2012. New trends in fluorescence in situ hybridization for identification and functional analyses of microbes. Current Opinion in Biotechnology **23**:95–101.

68. **Stoecker K**, **Bendinger B**, **Schöning B**, **Nielsen PH**, **Nielsen JL**, **Baranyi C**, **Toenshoff ER**, **Daims H**, **Wagner M**. 2006. Cohn's Crenothrix is a filamentous methane oxidizer with an unusual methane monooxygenase. Proc. Natl. Acad. Sci. U.S.A. **103**:2363–2367.

69. **Hesselmann RPX**, **Werlen C**, **Hahn D**, **van der Meer JR**, **Zehnder AJB**. 1999. Enrichment, Phylogenetic Analysis and Detection of a Bacterium That Performs Enhanced Biological Phosphate Removal in Activated Sludge. Systematic and Applied Microbiology **22**:454–465.

70. **West NJ**, **Schönhuber WA**, **Fuller NJ**, **Amann RI**, **Rippka R**, **Post AF**, **Scanlan DJ**. 2001. Closely related Prochlorococcus genotypes show remarkably different depth distributions in two oceanic regions as revealed by in situ hybridization using 16S rRNA-targeted oligonucleotides. Microbiology (Reading, Engl.) **147**:1731–1744.

71. **Kong Y**, **He M**, **McAlister T**, **Seviour R**, **Forster R**. 2010. Quantitative Fluorescence In Situ Hybridization of Microbial Communities in the Rumens of Cattle Fed Different Diets. Appl. Environ. Microbiol. **76**:6933–6938.

Table S1. Probes used in modeling.

	Target site position[c]			ΔG^o values at 46^oC					
Probe[a,b]	**5' end**	**3' end**	**Sequence (5'-3')**	ΔG^o_1	ΔG^o_2	ΔG^o_3	$\Delta G^o_{overall}$	$[FA]_{m,obs}$[d]	**Ref**[e]
E86-109	86	109	TCCGCCACTCGTCAGCAAAGAAGC	-27.4	0.2	-13.0	-14.1	42.6	C
E151-170	151	170	AGCTACCGTTTCCAGTAGTT	-22.2	0.1	-8.4	-13.4	33.6	A
Eco 181	181	198	CTTTGGTCTTGCGACGTT	-19.0	0.3	-4.2	-14.5	24.4	D
Eco 316	316	333	ACCGTGTCTCAGTTCCAG	-20.9	2.0	-5.8	-15.1	36.9	D
E392-410	392	410	CATACACGCGGCATGGCTG	-21.3	0.4	-4.5	-16.5	37.8	C
Eco 440	440	456	TCCCTTCCTCCCCGCTG	-26.2	4.6	-11.2	-15.0	60.0	D
Eco 541	541	558	CCGATTAACGCTTGCACC	-19.0	0.0	-2.2	-16.3	39.7	D
E615-634	615	634	GCAGTTCCCAGGTTGAGCCC	-27.8	0.2	-12.4	-15.1	47.7	C
E674-693	674	693	CACCGCTACACCTGGAATTC	-21.3	0.1	-7.3	-13.6	41.2	C
E774-792	774	792	TCTAATCCTGTTTGCTCCC	-21.5	1.3	-6.5	-14.9	24.0	C
E839-859	839	859	CCGGAAGCCACGCCTCAAGGG	-26.9	0.2	-11.8	-14.8	39.4	C
E886-901	886	901	TTGCGGCCGTACTCCC	-21.8	1.6	-7.3	-14.5	39.6	A
E958-980	958	980	GGTTCTTCGCGTTGCATCGAATT	-24.5	0.0	-6.2	-17.9	52.4	C
Eco 1042	1042	1059	GCCATGCAGCACCTGTCT	-23.7	0.4	-9.7	-13.7	31.1	D
E1088-1107	1088	1107	GCTCGTTGCGGGACTTAACC	-22.9	-0.1	-7.0	-15.4	33.7	C
E1136-1152	1136	1152	TTTGAGTTCCCGGCCGG	-21.9	0.7	-6.3	-15.4	35.8	C
E1218-1234	1218	1234	GCACGTGTGTAGCCCTG	-21.3	0.5	-7.0	-14.1	27.7	C
E1270-1291	1270	1291	ACTTTATGAGGTCCGCTTGCTC	-24.6	0.4	-3.5	-20.8	46.7	B
E1336-1356	1336	1356	CCACGATTACTAGCGATTCCG	-21.4	0.1	-4.4	-16.6	26.9	C
Eco 1410	1410	1427	GCAACCCACTCCCATGGT	-23.6	0.0	-9.2	-14.0	55.4	D
Eco 1482	1482	1499	TACGACTTCACCCCAGTC	-20.5	0.5	-4.7	-15.6	31.6	D
E255-278	255	278	CGCCTAGGTGAGCCGTTACCCCAC	-33.2	0.3	-12.4	-20.5	62.4	C
EUB338	338	355	GCTGCCTCCCGTAGGAGT	-25.1	-1.6	-9.2	-14.3	62.0	E
E1177-1201	1177	1201	TGACTTGACGTCATCCCCACCTTCC	-31.6	-0.7	-7.8	-22.9	69.4	B
E1425-1449	1425	1449	GGTTAAGCTACCTACTTCTTTTGCA	-25.7	0.6	-6.2	-19.3	30.7	C
E1449-1470	1449	1470	AGTGGTAAGCGCCCTCCCGAAG	-27.9	0.2	-10.1	-17.5	42.3	C

Table S1 cont.

	Target site position[c]			ΔG^o values at 46^{o}C					
Probe[a,b]	**5' end**	**3' end**	**Sequence (5'-3')**	ΔG^o_1	ΔG^o_2	ΔG^o_3	$\Delta G^o_{overall}$	$[FA]_{m,obs}$[d]	**Ref**[e]
E1457-1477	1457	1477	ATCACAAAGTGGTAAGCGCCC	-21.9	0.3	-8.3	-13.3	23.4	C
E483-506	483	506	CCGGTGCTTCTTCTGCGGGTAACG	-29.3	-0.2	-8.5	-20.3	63.5	C
E1416-1434	1416	1434	TTCTTTTGCAACCCACTCC	-21.1	1.5	-9.2	-11.8	31.5	C
E1471-1491	1471	1491	CACCCCAGTCATGAATCACAA	-21.3	0.5	-7.6	-13.5	31.7	C
E1-25	1	25	GAGCCATGATCAAACTCTTCAATTT	-21.8	1.0	-3.2	-18.5	40.7	C
E126-142	126	142	CCATCAGGCAGTTTCCC	-21.2	0.8	-7.8	-13.2	13.6	C
E577-596	577	596	TTAACAAACCGCCTGCGTGC	-21.2	0.2	-7.0	-13.9	31.7	C
E719-733	719	733	CGCCTTCGCCACCGG	-21.3	0.3	-1.4	-19.5	35.4	C
E56-75	56	75	CTGTTACCGTTCGACTTGCA	-21.5	1.8	-3.3	-18.2	37.0	C
E447-467	447	467	ATTAACTTTACTCCCTTCCTC	-21.1	1.8	-6.7	-14.4	28.8	C
E818-838	818	838	CACAACCTCCAAGTCGACATC	-21.0	1.3	-7.0	-13.9	33.2	C
E912-930	912	930	GCCCCCGTCAATTCATTTG	-22.5	0.6	-7.3	-15.0	32.6	C
Sac13-34	13	34	ATATGACTACTGGCAGGATCAA	-19.4	1.1	-5.3	-14.0	14.5	F
Sac109-135	109	135	ATAAACGATAACTGATTTAATGAGCCAT	-19.7	1.1	-6.2	-13.4	21.9	F
Sac149-173	149	173	ATTAGCTCTAGAATTACCACAGTTATA	-22.6	0.7	-7.8	-14.6	25.7	F
Sac231-247	231	247	AGTTGATAGGGCAGAAATT	-15.0	1.8	0.0	-14.5	6.8	F
Sac311-326	311	326	CTCAGGCTCCCTCTCC	-22.4	0.1	-8.1	-13.9	48.8	F
Sac358-378	358	378	AGGATTGGGTAATTTGCGCGC	-22.1	0.7	-8.7	-13.2	25.9	F
Sac583-600	583	600	CCAACCGGGCCCAAAGTTCAAC	-25.0	0.7	-11.5	-13.3	21.8	F
Sac720-741	720	741	CCTTGGCAAATGCTTTCGCAGTAG	-25.3	-0.6	-11.0	-13.5	45.8	F
Sac804-825	804	825	TCGGCATAGTTTATGGTTAAGA	-19.2	0.1	-3.9	-14.9	21.3	F
Sac844-859	844	859	AAGGTGCCGAGTGGGTCATTA	-23.2	0.1	-9.5	-13.3	46.9	F
Sac861-882	861	882	ACCCAAAGACTTTGATTTCTCG	-19.7	-0.7	-5.3	-13.5	25.7	F
Sac938-954	938	954	CCGCAGGCTCCACTCCT	-24.4	-0.3	-9.3	-14.5	43.5	F
Sac966-983	966	983	TGAGTTTCCCCGTGTTGAG	-21.5	1.2	-6.8	-14.6	24.7	F
Sac1108-1131	1108	1131	ACCACTATTTAGTAGGTTAAGGTCTC	-24.6	-0.1	-9.3	-14.8	24.3	F
Sac1154-1177	1154	1177	CATCGGCTTGAAACCGATAGTCCC	-26.2	-2.9	-9.4	-13.9	33.8	F
Sac1321-1347	1321	1347	CTAGGAATTCCTCGTTGAAGAGCAATA	-24.2	-0.9	-8.9	-14.3	35.9	F

Table S1 cont.

	Target site position[c]			ΔG^o values at 46^{o}C					
Probe[a,b]	**5' end**	**3' end**	**Sequence (5’-3’)**	ΔG^o_1	ΔG^o_2	ΔG^o_3	$\Delta G^o_{overall}$	$[FA]_{m,obs}$[d]	**Ref**[e]
Sac1420-1448	1420	1448	AGATCCTGAGGCCTCACTAAGCCATTC	-31.5	-1.2	-16.4	-13.8	54.5	F
Sac1446-1458	1446	1458	CTCCGCTCTGAGATGGAGTTGCCCCCT	-37.8	-0.6	-23.0	-14.0	70.7	F
Sac1464-1485	1464	1485	CCTCTAAATGACCAAGTTTGTCC	-22.2	0.7	-7.8	-14.2	37.3	F
Bc5-22	5	22	CCAGGATCAAACTCTCCG	-18.1	0.9	-4.0	-14.0	22.1	F
Bc26-41	26	41	CGCCAGCGTTCGTCCT	-21.8	0.6	-7.6	-14.0	12.2	F
Bc42-60	42	60	TTGCATGTATTAGGCACGC	-19.0	-0.3	-4.3	-14.1	24.1	F
Bc52-72	52	72	TCCGCTCGACTTGCATGTA	-21.4	0.7	-7.4	-13.8	29.6	F
Bc94-111	94	111	CGTCCGCCGCTAACATC	-20.1	0.7	-6.1	-13.8	21.2	F
Bc110-127	110	127	CCACGTGTTACTCACCCG	-21.3	-0.3	-6.8	-13.9	24.2	F
Bc365-387	365	387	AGACTTTCGTCCATTGCGGAAGA	-23.6	-2.4	-7.8	-13.4	28.0	F
Bc519-532	519	532	TACCGCGGCTGCTG	-17.4	0.0	-3.2	-13.8	9.1	F
Bc617-632	617	632	CCCTCCCCGGTTGAGC	-23.9	0.6	-9.9	-13.8	50.6	F
Bc658-673	658	673	CACTCTCCTCTTCTGC	-18.5	2.8	-4.7	-13.8	32.6	F
Bc697-713	697	713	CTCCACATCTCTACGCA	-18.5	0.0	-4.4	-13.7	39.0	F
Bc704-721	704	721	TGGTGTTCCTCCACATCT	-21.1	-0.1	-6.3	-14.3	42.0	F
Bc808-826	808	826	GCACTCATCGTTTACGGCG	-20.8	0.0	-5.9	-14.5	26.7	F
Bc983-1000	983	1000	AGGATGTCAAGACCTGGT	-18.2	0.3	-4.3	-13.6	25.8	F
Bc1038-1053	1038	1053	CACCACCTGTCACTCT	-18.7	0.2	-4.8	-13.6	25.2	F
Bc1066-1088	1066	1088	CCAACATCTCACGACACGAGCTG	-23.6	0.0	-9.2	-14.0	31.0	F
Bc1147-1167	1147	1167	TCACCGGCAGTCACCTTAGAG	-24.2	0.6	-10.4	-13.6	26.7	F
Bc1367-1387	1367	1387	CCCGGGAACGTATTCACCGCG	-25.5	-1.0	-10.1	-14.3	31.5	F
Bc1454-1472	1454	1472	ACCTTCGGCGGCTGGCTCC	-27.5	-0.3	-13.4	-13.5	45.0	F
Rh40-56	40	56	ATGTGTTAGGCCTGCCG	-20.3	0.6	-5.5	-14.6	23.1	F
Rh122-138	122	138	AGGGCACGTTCCCACGC	-22.5	-0.9	-7.4	-14.1	44.0	F
Rh389-406	389	406	CACGCGGCATGGCTAGAT	-20.2	0.7	-5.4	-14.6	30.2	F
Rh678-703	678	703	CTACGAATTTCACCTCTACACTCGGA	-25.4	0.4	-10.3	-14.8	41.9	F
Rh836-855	836	855	CACCGAACAGCATGCTGCC	-22.1	-0.6	-6.3	-15.0	28.2	F
Rh883-897	883	897	GGCCGTACTCCCCAG	-21.0	1.5	-6.9	-14.0	38.8	F

Table S1 cont.

	Target site position[c]			ΔG^o values at 46°C					
Probe[a,b]	**5' end**	**3' end**	**Sequence (5'-3')**	ΔG^o_1	ΔG^o_2	ΔG^o_3	$\Delta G^o_{overall}$	**[FA]$_{m,obs}$[d]**	**Ref[e]**
Rh939-954	939	954	CCACATGCTCCACCGC	-20.8	1.1	-6.2	-14.5	39.7	F
Rh1006-1029	1006	1029	ACTGAAGGAACCATCTCTGGAACC	-23.1	-1.2	-7.3	-14.5	28.4	F
Rh1292-1315	1292	1315	AGACCCCAATCCGAACTGAGACAG	-24.9	0.5	-11.2	-13.5	20.9	F
Rh1354-1374	1354	1374	TCACCGCGTCATGCTGTTACG	-23.7	-0.5	-8.6	-14.4	44.8	F
St87-105	87	105	CCACTCGCCACCCAGAGAG	-24.5	-0.4	-9.7	-14.1	60.0	F
St106-122	106	122	TATTCCTCACCCGTCCG	-21.1	3.3	-7.2	-13.9	47.4	F
St115-131	115	131	ATTCCGATGTATTCCTC	-16.5	1.5	-1.7	-14.7	24.6	F
St146-165	146	165	AAGTTTCCCTACGTTATCCC	-21.9	0.8	-7.4	-14.3	43.5	F
St178-195	178	195	TCACCCGTAGGTCGTATG	-19.9	-0.8	-5.8	-13.1	44.6	F
St210-227	210	227	TCAATCGCGCAAGGTCCG	-19.6	1.4	-6.1	-13.4	26.0	F
St282-302	282	302	CCTCTTAGACCAGCTACGGAT	-23.2	0.0	-9.3	-13.5	28.7	F
St411-433	411	433	CTTTACAACCCGAAGGCCTTCTT	-24.7	-0.7	-9.9	-13.9	30.1	F
St575-598	575	598	ACTTAAACGACCACCTACGCACGC	-25.0	0.6	-10.6	-14.2	22.3	F
St796-813	796	813	AGGGCGTGGACTACCAGG	-21.4	-0.1	-7.0	-13.9	40.9	F
St812-832	812	832	ATCCAGTTCGCATCGTTTAG	-20.3	1.0	-6.9	-13.3	39.1	F
St865-884	865	884	AGGCGGCGAACTTAACGCGT	-21.7	-0.9	-6.4	-14.3	21.6	F
St891-909	891	909	TTTCAGTCTTGCGACCGTA	-20.2	0.0	-6.8	-13.0	36.6	F
St947-971	947	971	CGTTGCATCGAATTAAACCACATAC	-20.5	0.6	-6.9	-13.4	28.4	F
St1105-1127	1105	1127	CAACTAAGGACAAGGGTTGCGCT	-22.8	0.3	-8.0	-14.5	34.4	F
St1172-1187	1172	1187	CCCCACCTTCCTCCGG	-24.2	1.4	-9.9	-14.2	63.8	F
St1309-1332	1309	1332	TCATGGAGTCGAGTTGCAGACTCC	-26.7	-3.0	-10.2	-13.5	41.6	F
St1406-1426	1406	1426	CAACAAACTCCCATGGTGTGA	-19.9	-0.3	-6.3	-13.0	34.5	F
St1453-1470	1453	1470	CGTGGCAAGCGCCCTCCC	-26.8	-1.6	-10.1	-15.1	39.6	F
St1470-1487	1470	1487	CCAGTCATCGGCCACACC	-23.0	0.6	-9.1	-13.7	44.3	F

[a] Probe names indicate target organism and target site positioning. E and Eco, *Escherichia coli*; Sac, *Saccharomyces cerevisiae*; Bc, *Bacillus subtilis*; Rh, *Rhodobacter sphaeroides*; St, *Stenotrophomonas maltophilia*.

[b] All *E. coli* probes were used in a previous modeling study (Yilmaz and Noguera, 2007) except for E1471-1491 and E1425-1449.

[c] In *E. coli* numbering.

[d] Experimental melting point (see text).

[e] Reference study where the probe is originally published. A, Yilmaz and Noguera (1); B, Yilmaz and Noguera (5); C, Yilmaz *et al.* (2); D, Fuchs *et al*. (16); E, Wallner *et al*. (17); F, Okten *et al.* (3).

Table S2. Model development, comparison, and statistics[a].

	Original model	***Retrained mechanistic model (RMM)***	**Nearest neighbor model**	***Single-reaction model (SRM)***	**Double-reaction model**
MODEL DESCRIPTION					
$\Delta G^{o}_{1,0\%}$ *(kcal/mol)*	ΔG^{o}_{1}	ΔG^{o}_{1}	$\Sigma\Delta G^{o}_{NN} + \Delta G^{o}_{ini}$	$0.26\Delta G^{o}_{1} - 6.5$	$0.33\Delta G^{o}_{1} - 5.6$
$\Delta G^{o}_{2,0\%}$ *(kcal/mol)*	ΔG^{o}_{2}	ΔG^{o}_{2}	na	na	na
$\Delta G^{o}_{3,0\%}$ *(kcal/mol)*	ΔG^{o}_{3}	ΔG^{o}_{3}	na	na	$0.31\Delta G^{o}_{3} - 2.0$
m_1 *(kcal/mol/% FA)*	$0.0095L+0.0697$	$0.0079L+0.0696$	$0.0032L+0.0116$	$0.0028L+0.0175$	$0.0029L+0.0170$
m_2 *(kcal/mol/% FA)*	0.1	0.1	na	na	na
m_3 *(kcal/mol/% FA)*	$0.0117\vert\Delta G^{o}_{3}\vert$	$0.0111\vert\Delta G^{o}_{3}\vert$	na	na	$0.0096\vert\Delta G^{o}_{3}\vert$
CURVE-FITTING [b,c]					
s^2	na	0.025	0.013	0.016	0.015
R^2	na	0.86	0.93	0.92	0.92
Mean $\vert err([FA]_m)\vert$ *(% FA)*	na	8.4	6.2	7.2	7.2
CROSS VALIDATION [c,d,e]					
ε^2	0.032	0.024\0.026	0.015\0.018	0.014\0.015	0.014\0.019
R^2	0.80	0.85\0.84	0.91\0.89	0.91\0.91	0.91\0.89
Mean $\vert err([FA]_m)\vert$ *(% FA)*	7.9	8.8\9.3	7.5\8.0	7.4\7.4	7.6\8
Mean $\vert CV\ offset\vert$ *(% FA)*	na	0.4\2	1.3\2.6	0.2\0.9	0.5\2.6
$\vert CV\ offset\vert > 2\%\ FA$[f]	na	5\42 (30)	23\55 (20)	0\9 (8)	4\30 (20)
$\vert CV\ offset\vert > 5\%\ FA$[f]	na	2\10 (9)	0\14 (9)	0\0	2\15 (10)

[a] na, not applicable.

[b] s^2, total residual squares divided by the degree of freedom; R^2 , coefficient of determination (4).

[c] *err([FA]*$_m$*)*, error in melting point prediction for best-fits (curve-fitting) or independent predictions (cross-validation) (4).

[d] *CV offset*, the distance between melting points of best-fit and cross-validation curve as defined in Figure S2A; ε^2, error squares averaged for all independent predictions during cross validation.

[e] LOPOCV \ LOOOCV results shown when available in this order (see text for details of cross-validation tests). Numbers in parentheses indicate number of *E. coli* probes that fell in the pertaining category during LOOOCV. Since the *E. coli* set is relatively large (38 probes; ~1/3 of the entire set) LOOOCV test with these probes is very conservative. Cross-validation of the original model is based on an average for all 79 probes that were not part of the training set of this model (1).

[f] Number of probes with CV offsets greater than the value mentioned.

Table S3. Nearest neighbor parameters.

		ΔG° at 46°C (kcal/mol)	
NN[a]	**DNA**[b]	**Original**[c]	**NN model**[d]
UU/AA	AA	0.12	0.09
GU/CA	AC	-0.91	-0.15
CU/GA	AG	-0.71	0.01
AU/TA	AT	-0.67	-0.36
UG/AC	CA	-1.34	-0.54
GG/CC	CC	-2.62	-0.78
CG/GC	CG	-1.27	-0.39
AG/TC	CT	-1.60	-0.58
UC/AG	GA	-1.29	-0.40
GC/CG	GC	-2.54	-0.62
CC/GG	GG	-1.90	-0.63
AC/TG	GT	-1.97	-0.77
UA/AT	TA	-0.40	-0.13
GA/CT	TC	-1.19	-0.41
CA/GT	TG	-0.67	-0.18
AA/TT	TT	-0.81	-0.26
ini[e]		3.14	-4.41

[a] Nearest neighbors. Base stacking is shown as RNA/DNA pairs in 5′-3′/3′-5′ direction.

[b] DNA nearest neighbors in 5′-3′ direction.

[c] Obtained for 46°C using ΔH° and ΔS° values from Sugimoto *et al.* (7), using ΔG°= ΔH° - T ΔS°.

[d] Best-fitting parameters with the nearest neighbor model (see Table S2).

[e] Free energy of initiation.

Table S4. Moderate and low coverage 16S probes published in literature.

Probe Name	***probeBase Accession Number***	***Sequence (5' to 3')***	***Target Group***	***Approximate Coverage***[a]	***Reference***
Str56	pB-03987	ATCCTGCGTTCTACTTGC	Streptococcus	0%	(18)
ALBO577	pB-00912	CCGAACCGCCTGCGCAC	Alcaligenes	1%	(19)
Eury806	pB-01172	CACAGCGTTTACACCTAG	Euryarchaeota	6%	(20)
CYT1438	pB-02589	CCGCTCCTTACGGTGACG	Cytophaga	10%	(21)
RUMs278	pB-03700	GTCCGGCTACCGATCGCG	Ruminococcaceae	13%	(22)
Cren537	pB-01171	TGACCACTTGAGGTGCTG	Crenarchaeota	14%	(20)
Pro60	pB-03989	CTCCCTTCACCGTTCGAC	Propionicimonas	16%	(18)
CYT1448	pB-02588	CTAGGCCGCTCCTTACGG	Cytophaga	16%	(21)
SFH646	pB-03885	CTCCCTGCCTCAAGTCCA	Sulfurihydrogenibium	18%	(23)
AERO1244	pB-01247	GCTTGCAGCCCTCTGTACGCG	Aeromonadaceae	19%	(24)
OCE232	pB-00614	AGCTAATCTCACGCAGGC	Oceanospirillum	30%	(25)
Thus438	pB-00314	GGGTTTCGTCCCGGGTTC	Thermus	30%	(26)
SRB385 (SRB)	pB-00300	CGGCGTCGCTGCGTCAGG	Desulfovibrionales	33%	(27)
VP403	pB-02645	CGAAGACCTTATCCTCCACG	Verrucomicrobium	33%	(28)
RHIZ1244	pB-02665	TCGCTGCCCACTGTCACC	Rhizobiales	35%	(29)
Ppu646	pB-01249	CTACCGTACTCTAGCTTG	Pseudomonas	56%	(30)
Mc1029	pB-00585	CCTGTGTCTTGGCTCCCGAA	Methylococcus	56%	(31)
PAR1457	pB-00278	CTACCGTGGTCCGCTGCC	Paracoccus	56%	(32)
SPA714	pB-02646	CCTTCGCCACTGGTCTTC	Spartobacteria	56%	(28)
LGC0355	pB-01212	GGAAGATTCCCTACTGCTG	Firmicutes	58%	(33)
Hydr540	pB-00187	TCGCGCAACGCTCGGGACC	Aquificales	60%	(26)
Geo/Blasto	pB-02584	CCATCCCCAGCCGGAAACC	Geodermatophilus	64%	(34)

Probe Name	***probeBase Accession Number***	***Sequence (5' to 3')***	***Target Group***	***Approximate Coverage***[a]	***Reference***
Mg1004	pB-00342	TACGATCTCTCACAGATT	Methylomicrobium	68%	(35)
Ver620	pB-01037	ATGTGCCGTCCGCGGGTT	Akkermansia	69%	(36)
DSM213	pB-00507	CATCCTCGGACGAATGCA	Desulfomicrobium	69%	(37)
DSV214	pB-00087	CATCCTCGGACGAATGC	Desulfomicrobium	69%	(38)
GAN850	pB-02644	TTGCTTCGAGAGCCATAC	Gallibacterium	71%	(39)
BDE525	pB-01570	GATCCCTCGTCTTACCGC	Bdellovibrio	72%	(40)
129(DSB129)	pB-00073	CAGGCTTGAAGGCAGATT	Desulfobacter	72%	(41)
ODB1021	pB-02636	GCGTCCCCTAAGGGAACT	Octadecabacter	74%	(42)

[a] Approximate coverage of sequences belonging to the target genus in the RDP database (version 10.28), or average coverage of all named genera belonging to the target group.

Table S5. Moderate and low specificity 16S probes published in literature.

Probe Name	*probeBase Accession Number*	*Sequence (5' to 3')*	*Specificity Score*	*Qualitative Specificity* [a]	*Reference*
LGC	pB-01040	TCACGCGGCGTTGCTC	626	low	(43)
E11	pB-01326	AGCCATGCAGCACCTGTCTC	405	low	(44)
Lmon	pB-03697	CTATCCATTGTAGCACGTG	320	low	(45)
SRB385Db	pB-00301	CGGCGTTGCTGCGTCAGG	291	low	(46)
MET1217	pB-00916	TTACGTGTGAAGCCCTGG	251	low	(19)
SRB385 (SRB)	pB-00300	CGGCGTCGCTGCGTCAGG	187	low	(27)
Lsinu_268	pB-03968	GCTAAAGATCGTAGCCTTGGTAA	173	low	(47)
SS_HOL1400	pB-01552	TTCGTGATGTGACGGGC	164	low	(48)
HoAc1402	pB-00183	CTTTCGTGATGTGACGGG	145	low	(49)
DSBAC357	pB-01320	CCATTGCGCAAAATTCCTCAC	136	low	(37)
DSBAC355	pB-00076	GCGCAAAATTCCTCACTG	129	low	(50)
PSE1284	pB-02540	GATCCGGACTACGATCGGTTT	103	low	(51)
G Rb	pB-00610	GTCAGTATCGAGCCAGTGAG	92	low	(25)
LEG705	pB-00193	CTGGTGTTCCTTCCGATC	88	low	(52)
Rhoc-1425	pB-01231	ACTACCTACTTCTGGTGA	53	low	(53)
KT13-231	pB-02593	ATCTAATCAAACGCGGGCC	52	low	(21)
DSV407	pB-00088	CCGAAGGCCTTCTTCCCT	47	low	(38)
G123T	pB-00170	CCTTCCGATCTCTATGCA	46	low	(54)
STEBA1426	pB-00266	ACTACCTACTTCTGGTGG	42	low	(49)
REX72	pB-01069	TGGGAGCAAGCTCCCAAAG	34	low	(55)
TM7905	pB-00600	CCGTCAATTCCTTTATGTTTTA	356	moderate	(56)
ALF968	pB-00021	GGTAAGGTTCTGCGCGTT	300	moderate	(57)

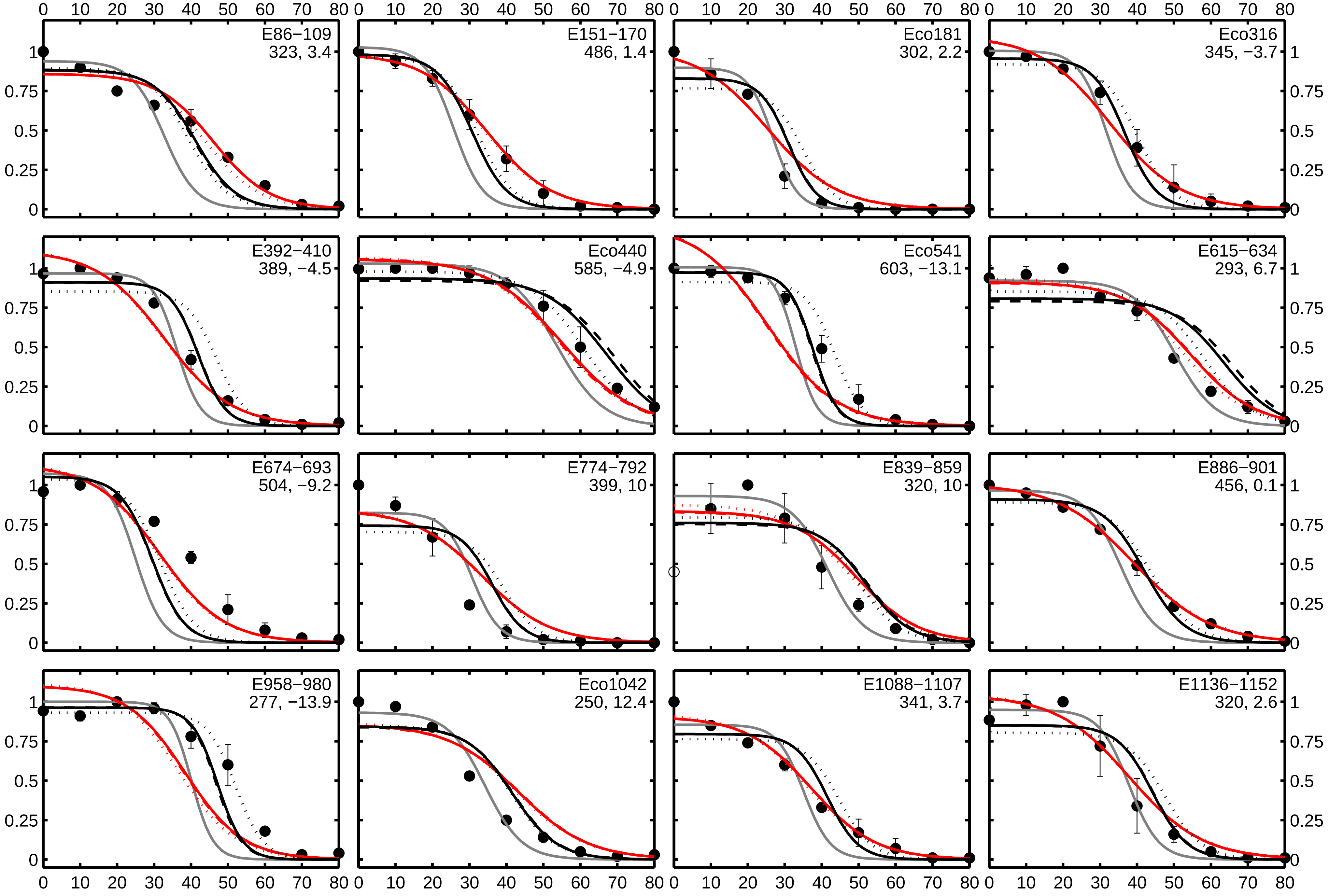
Probe Name	*probeBase Accession Number*	*Sequence (5’ to 3’)*	*Specificity Score*	*Qualitative Specificity* [a]	*Reference*
LGC354A	pB-00195	TGGAAGATTCCCTACTGC	183	moderate	(58)
Nso1225	pB-00248	CGCCATTGTATTACGTGTGA	159	moderate	(59)
CF319a	pB-00042	TGGTCCGTGTCTCAGTAC	155	moderate	(60)
CFB719	pB-00047	AGCTGCCTTCGCAATCGG	147	moderate	(61)
LGC353b	pB-01070	GCGGAAGATTCCCTACTGC	140	moderate	(55)
PARA739	pB-02664	GCGTCAGTATCGAGCCAG	101	moderate	(29)
DSV698	pB-00091	GTTCCTCCAGATATCTACGG	82	moderate	(38)
Lflag_268	pB-03967	GCTAAAGATCGAAGCCTTGGTAA	63	moderate	(47)
b886	pB-01024	TCAGGCGGTCGACTTCAT	63	moderate	(62)
Cte	pB-00378	TTCCATCCCCCTCTGCCG	55	moderate	(63)
SPH492	pB-00919	TAGCCGGAGCTTATTCTC	54	moderate	(19)
CREN569	pB-00789	GCTACGGATGCTTTAGG	53	moderate	(64)
EPSY549	pB-01321	CAGTGATTCCGAGTAACG	48	moderate	(65)
MNP1	pB-00628	TTAGACCCAGTTTCCCAGGCT	46	moderate	(66)
687(DSV687)	pB-00090	TACGGATTTCACTCCT	41	moderate	(41)
Lis-1255	pB-02605	ACCTCGCGGCTTCGCGAC	38	moderate	(67)
Cp1130-B	pB-01187	TTCCCGGCATTACCCGCT	33	moderate	(68)
RHW991	pB-00562	GTTCTCTTTCGAGCACTC	32	moderate	(69)
405_Syn	pB-00766	AGAGGCCTTCATCCCTCA	31	moderate	(70)
Cp1130-A	pB-01186	TTCCCGCCATTACGCGCT	29	moderate	(68)
EURY499	pB-00791	CGGTCTTGCCCGGCCCT	26	moderate	(64)
PAR1457	pB-00278	CTACCGTGGTCCGCTGCC	25	moderate	(32)

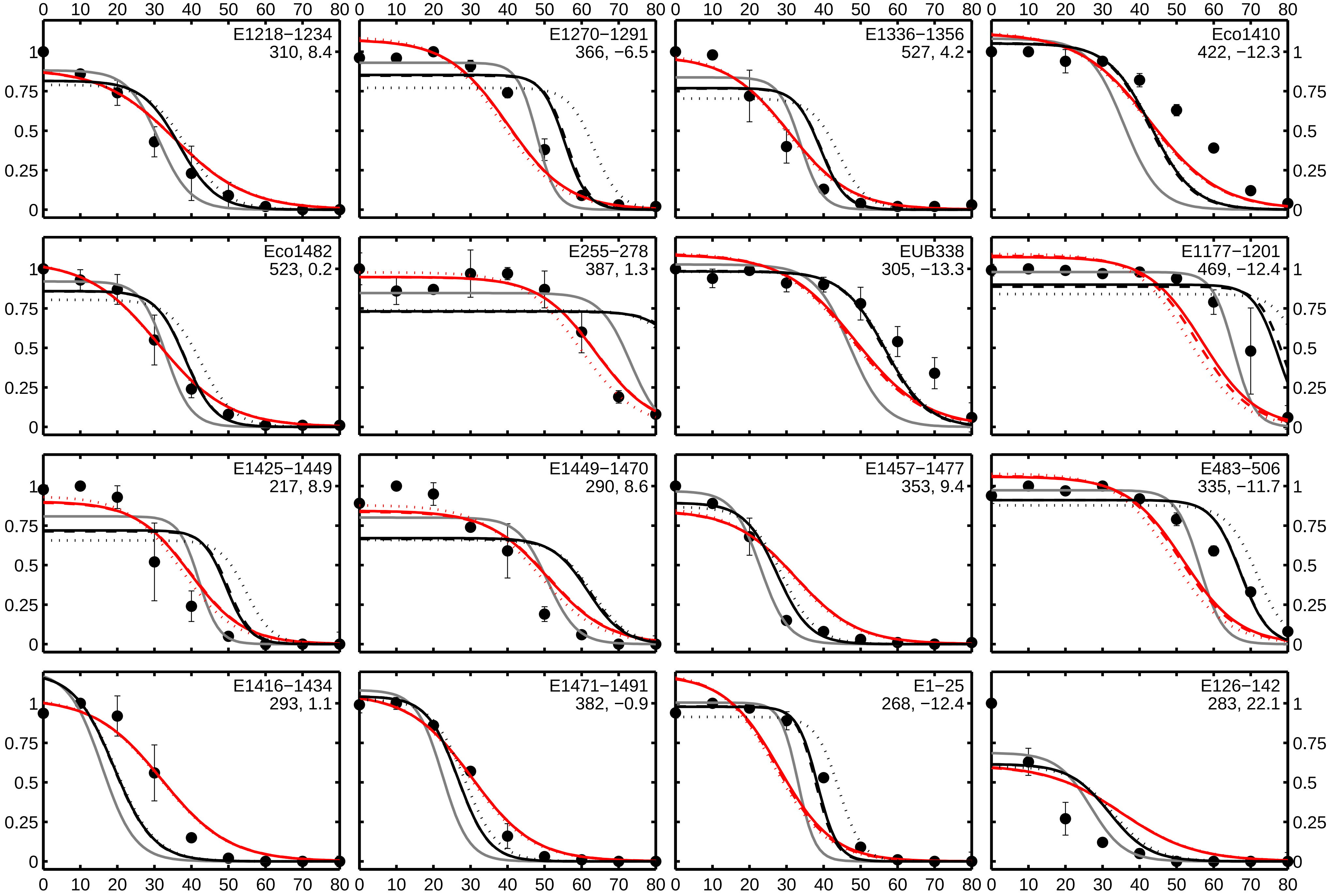
[a] Qualitative assessment of specificity relative to the breadth of the intended target group (see main text).

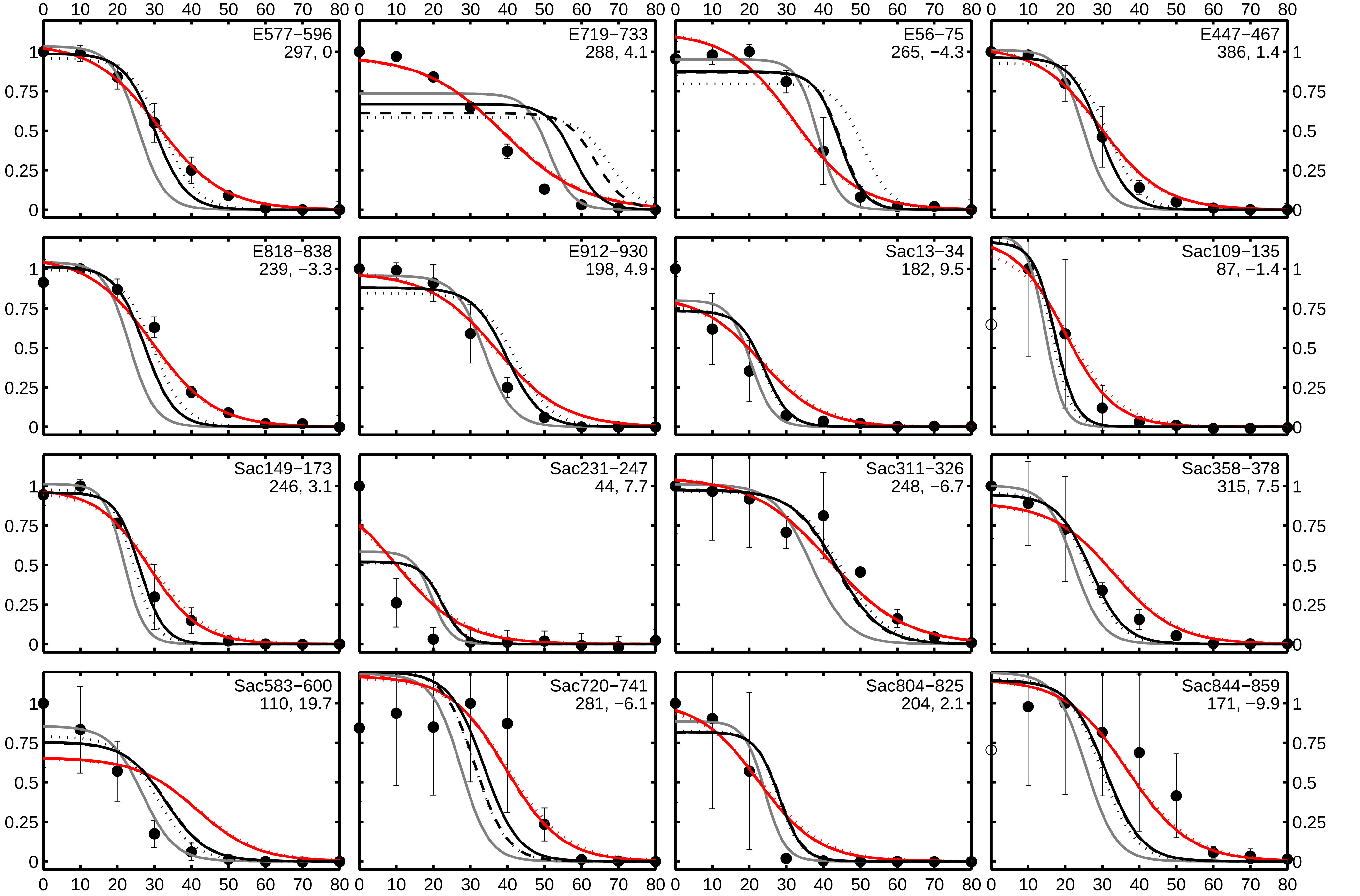
Table S6. Phylum-specific probes in probeBase (15) that were similar to those designed with DECIPHER.

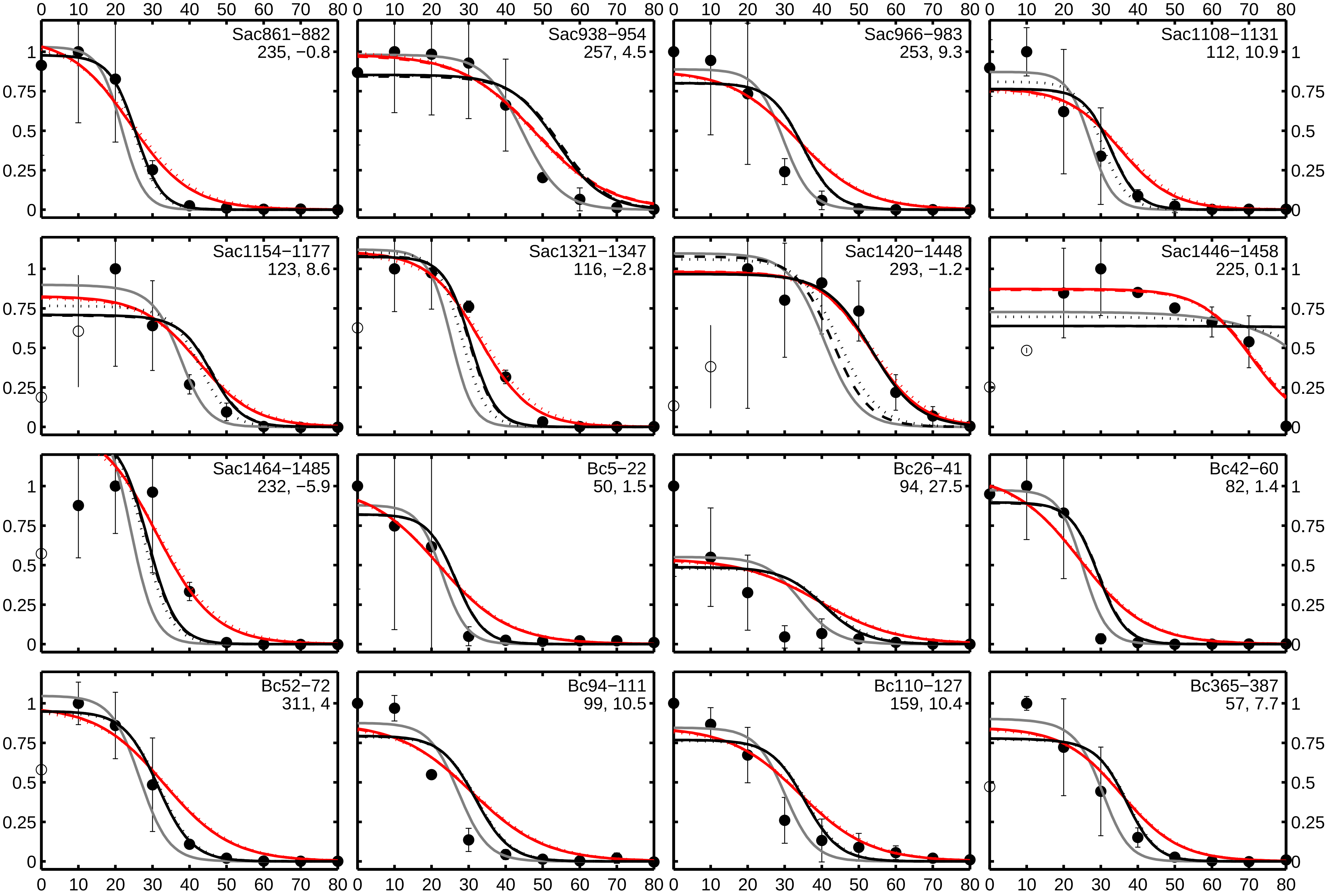
Probe Name	***probeBase Accession Number***	***Probe Sequence (5' to 3')***	***Probe Designed with DECIPHER (5' to 3')***	***Target phylum*** [a]	***Reference***
Bac1080	pB-02683	GCACTTAAGCCGACACCT	GGCACTTAAGCCGACACCT	Bacteroidetes	(71)
CREN512	pB-00788	CGGCGGCTGACACCAG	GCGGCGGCTGACACCAG	Crenarchaeota	(64)
LGC0355	pB-01212	GGAAGATTCCCTACTGCTG	YGGAAGATTCCCTACTGCTGC	Firmicutes	(33)
LGC354A	pB-00195	TGGAAGATTCCCTACTGC	YGGAAGATTCCCTACTGCTGC	Firmicutes	(58)

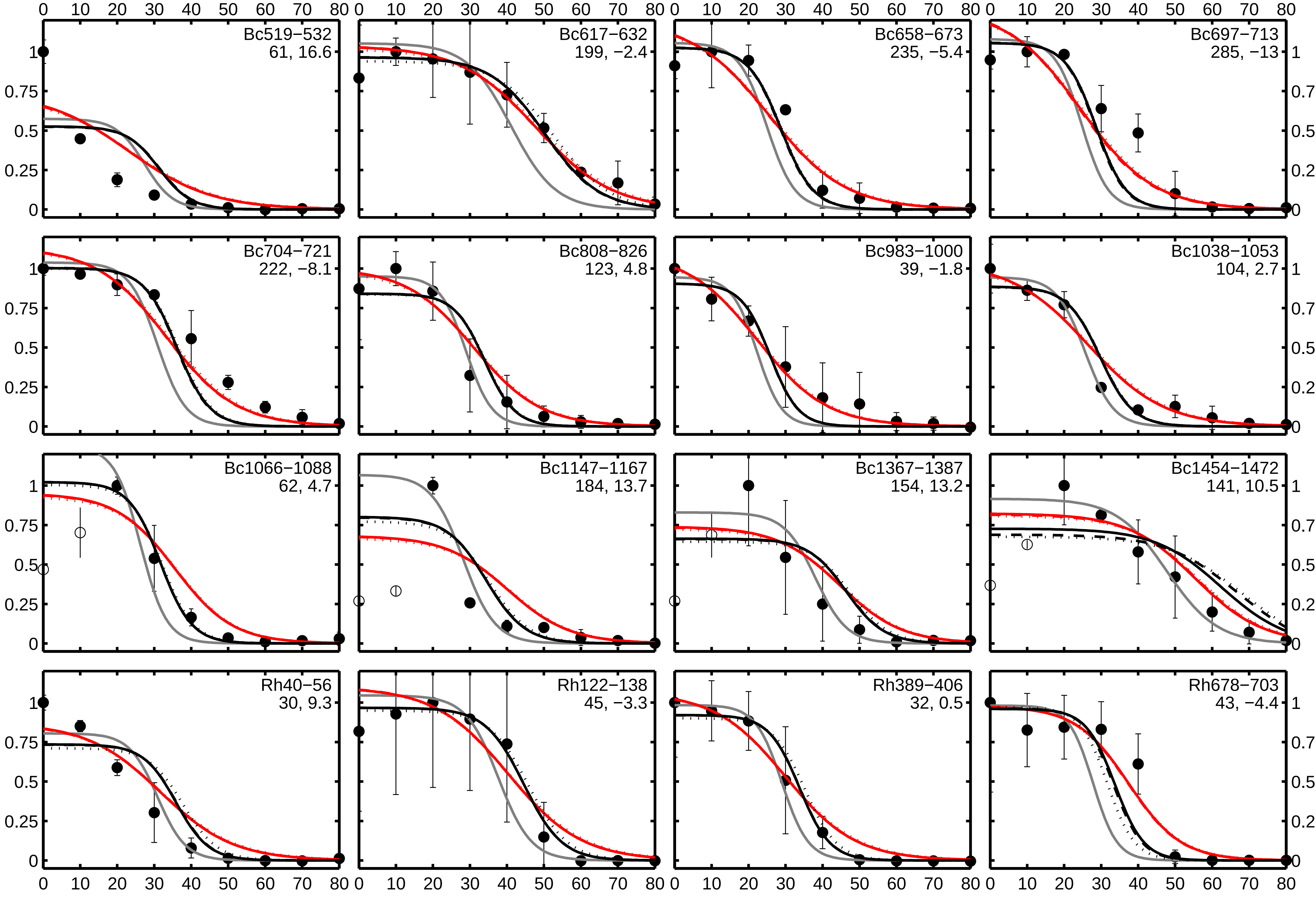
Figure S1: Experimental and theoretical formamide dissociation profiles. Circles are normalized experimental probe brightness values with open shapes indicating data excluded during curve fitting. Lines are theoretical profiles obtained with SRM (red), RMM (black), and the original model (grey). Solid, dashed, and dotted lines for SRM and RMM indicate best-fits, LOPOCV predictions, and LOOOCV predictions, respectively. Text on each graph indicates probe name followed by max brightness (Imax) and $errFA_m$ in SRM (best-fit). See text for details.

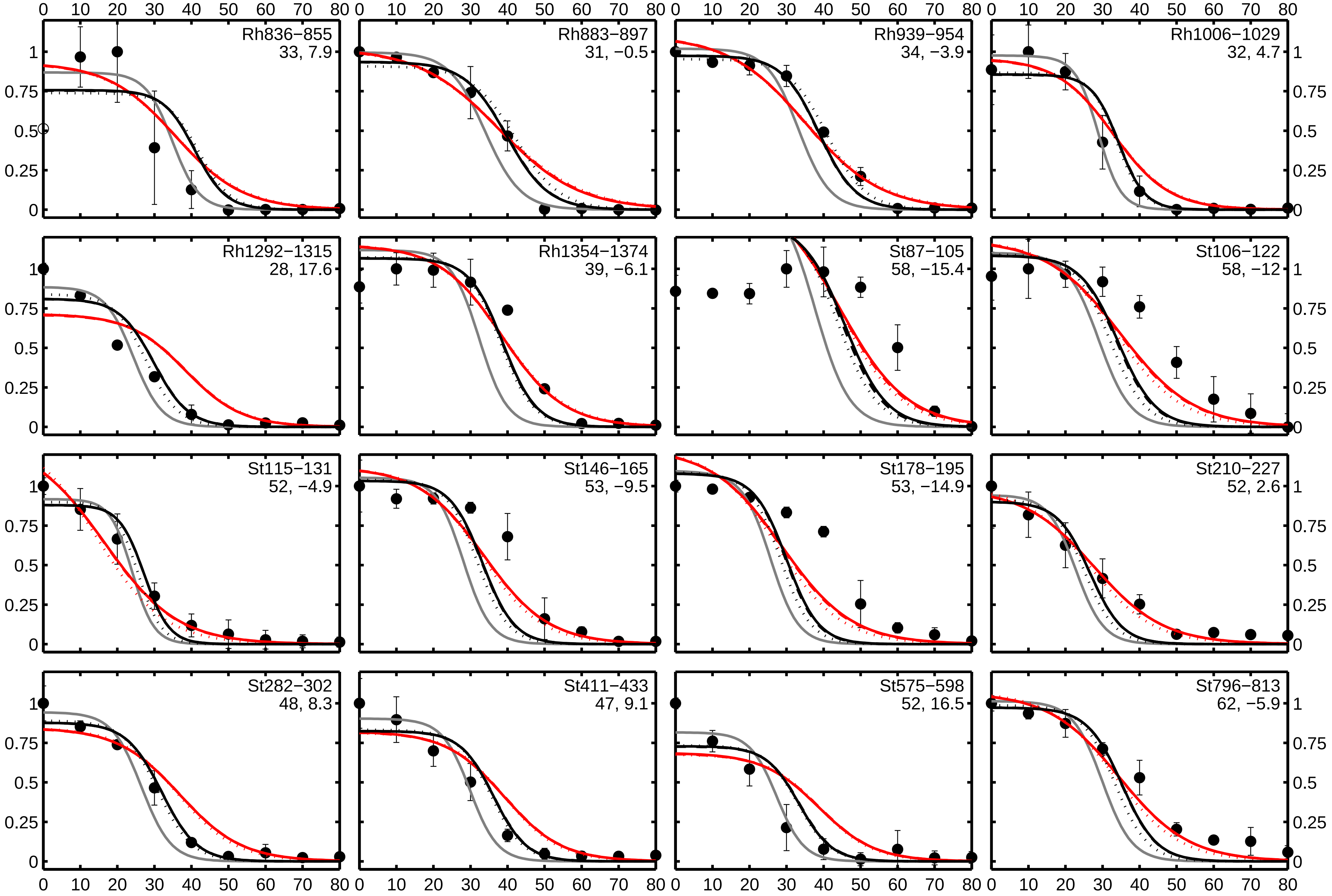

0 10 20 30 40 50 60 70 80
1
0.75
0.5
0.25
0
E86−109
323, 3.4
E151−170
486, 1.4
Eco181
302, 2.2
Eco316
345, −3.7
E392−410
389, −4.5
Eco440
585, −4.9
Eco541
603, −13.1
E615−634
293, 6.7
E674−693
504, −9.2
E774−792
399, 10
E839−859
320, 10
E886−901
456, 0.1
E958−980
277, −13.9
Eco1042
250, 12.4
E1088−1107
341, 3.7
E1136−1152
320, 2.6

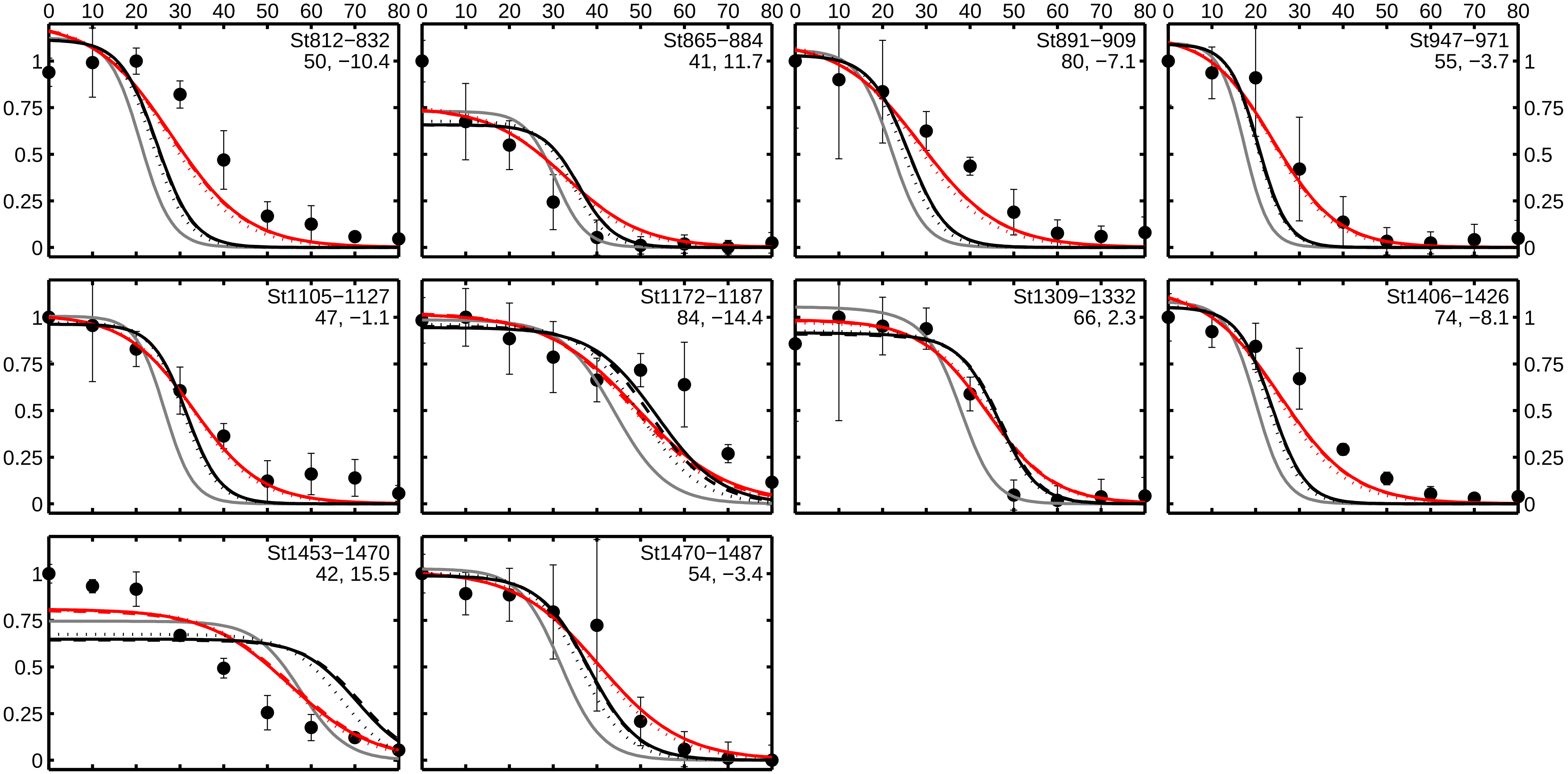

0 10 20 30 40 50 60 70 80
1
0.75
0.5
0.25
0
E1218−1234
310, 8.4
E1270−1291
366, −6.5
E1336−1356
527, 4.2
Eco1410
422, −12.3
Eco1482
523, 0.2
E255−278
387, 1.3
EUB338
305, −13.3
E1177−1201
469, −12.4
E1425−1449
217, 8.9
E1449−1470
290, 8.6
E1457−1477
353, 9.4
E483−506
335, −11.7
E1416−1434
293, 1.1
E1471−1491
382, −0.9
E1−25
268, −12.4
E126−142
283, 22.1


E577−596
297, 0
E719−733
288, 4.1
E56−75
265, −4.3
E447−467
386, 1.4
E818−838
239, −3.3
E912−930
198, 4.9
Sac13−34
182, 9.5
Sac109−135
87, −1.4
Sac149−173
246, 3.1
Sac231−247
44, 7.7
Sac311−326
248, −6.7
Sac358−378
315, 7.5
Sac583−600
110, 19.7
Sac720−741
281, −6.1
Sac804−825
204, 2.1
Sac844−859
171, −9.9
0
10
20
30
40
50
60
70
80
0
0.25
0.5
0.75
1


Sac861−882
235, −0.8
Sac938−954
257, 4.5
Sac966−983
253, 9.3
Sac1108−1131
112, 10.9
Sac1154−1177
123, 8.6
Sac1321−1347
116, −2.8
Sac1420−1448
293, −1.2
Sac1446−1458
225, 0.1
Sac1464−1485
232, −5.9
Bc5−22
50, 1.5
Bc26−41
94, 27.5
Bc42−60
82, 1.4
Bc52−72
311, 4
Bc94−111
99, 10.5
Bc110−127
159, 10.4
Bc365−387
57, 7.7
0
10
20
30
40
50
60
70
80
0
0.25
0.5
0.75
1


0 10 20 30 40 50 60 70 80
1
0.75
0.5
0.25
0
Bc519−532
61, 16.6
Bc617−632
199, −2.4
Bc658−673
235, −5.4
Bc697−713
285, −13
Bc704−721
222, −8.1
Bc808−826
123, 4.8
Bc983−1000
39, −1.8
Bc1038−1053
104, 2.7
Bc1066−1088
62, 4.7
Bc1147−1167
184, 13.7
Bc1367−1387
154, 13.2
Bc1454−1472
141, 10.5
Rh40−56
30, 9.3
Rh122−138
45, −3.3
Rh389−406
32, 0.5
Rh678−703
43, −4.4


Rh836−855
33, 7.9
Rh883−897
31, −0.5
Rh939−954
34, −3.9
Rh1006−1029
32, 4.7
Rh1292−1315
28, 17.6
Rh1354−1374
39, −6.1
St87−105
58, −15.4
St106−122
58, −12
St115−131
52, −4.9
St146−165
53, −9.5
St178−195
53, −14.9
St210−227
52, 2.6
St282−302
48, 8.3
St411−433
47, 9.1
St575−598
52, 16.5
St796−813
62, −5.9
0
10
20
30
40
50
60
70
80
0
0.25
0.5
0.75
1


St812−832
50, −10.4
St865−884
41, 11.7
St891−909
80, −7.1
St947−971
55, −3.7
St1105−1127
47, −1.1
St1172−1187
84, −14.4
St1309−1332
66, 2.3
St1406−1426
74, −8.1
St1453−1470
42, 15.5
St1470−1487
54, −3.4
0
10
20
30
40
50
60
70
80
0
0.25
0.5
0.75
1

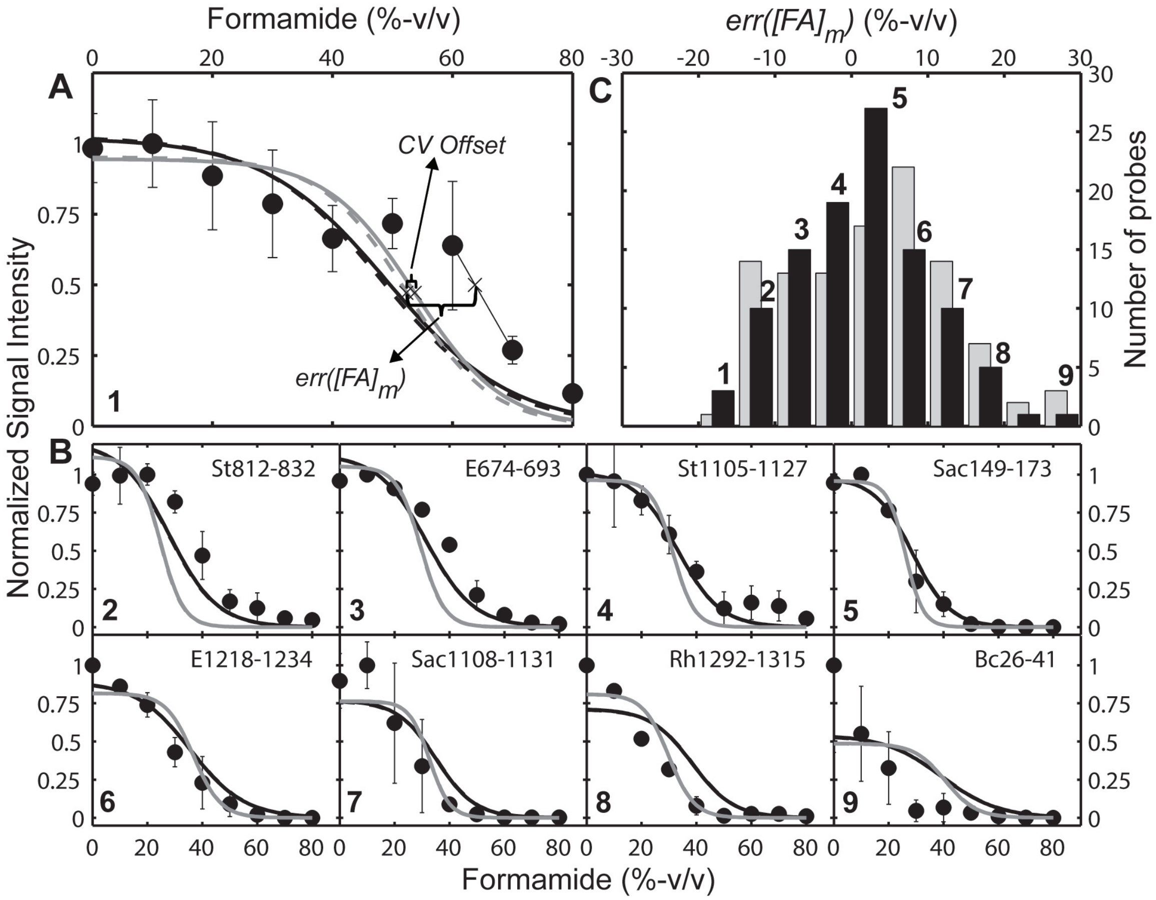


Figure S2. Formamide dissociation profiles and melting point predictions with selected models. **(A and B)** Best-fits (solid lines) and LOPOCV predictions (dashed lines) for RMM (grey) and SRM (black). Panel A illustrates error in melting point with respect to cross-validation as well as cross-validation offset. **(C)** Distribution of the melting prediction errors during LOPOCV for SRM (dark bars) and RMM (light bars) (also appears in Figure 1C). Probes in A and B are arbitrarily sampled from corresponding SRM bars in C according to the numbering indicated (see Figure S1 for all probes). LOPOCV predictions in A and B are generally not visible because of negligible CV offset.

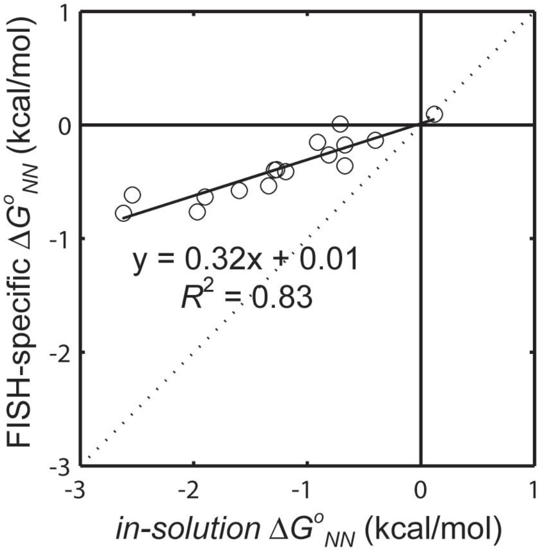


Figure S3. Relationship between original nearest-neighbor free energies (7) and best-fitting parameters used in nearest-neighbor model (Table S2). Dotted line indicates identity.

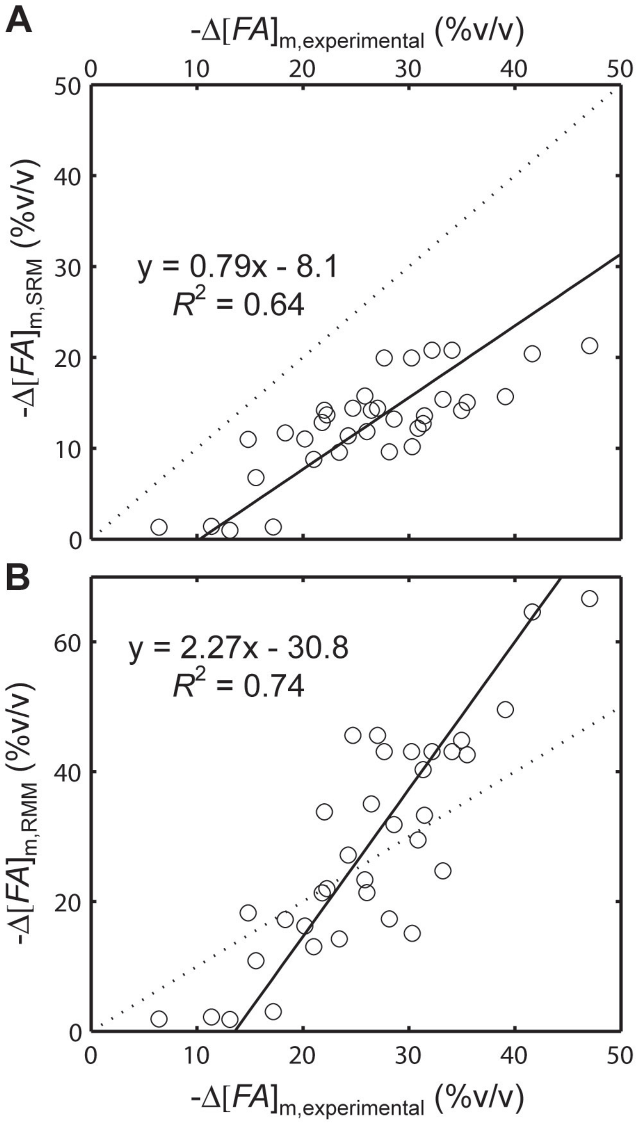


Figure S4. Correlation of experimental and theoretical Δ[*FA*]$_m$ values for SRM (A) and RMM (B). Panel A is the same as Figure 1D. Dotted lines indicate identity.

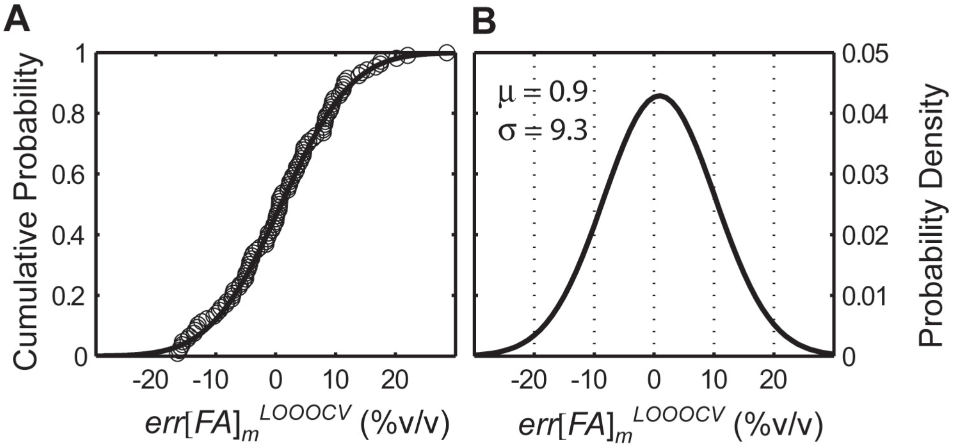


Figure S5. Derivation of a probability density function to describe predictive error on $[FA]_m$ based on Leave-One-Organism-Out cross validation (LOOOCV) tests. The distribution of predictive errors is approximately normally distributed (Shapiro-Wilk test for non-normality p = 0.723). **(A)** Cumulative distribution function (solid line) best-fitted to cumulative frequency of prediction errors (circles). **(B)** Corresponding probability density function with mean and standard deviation indicated in the upper left.

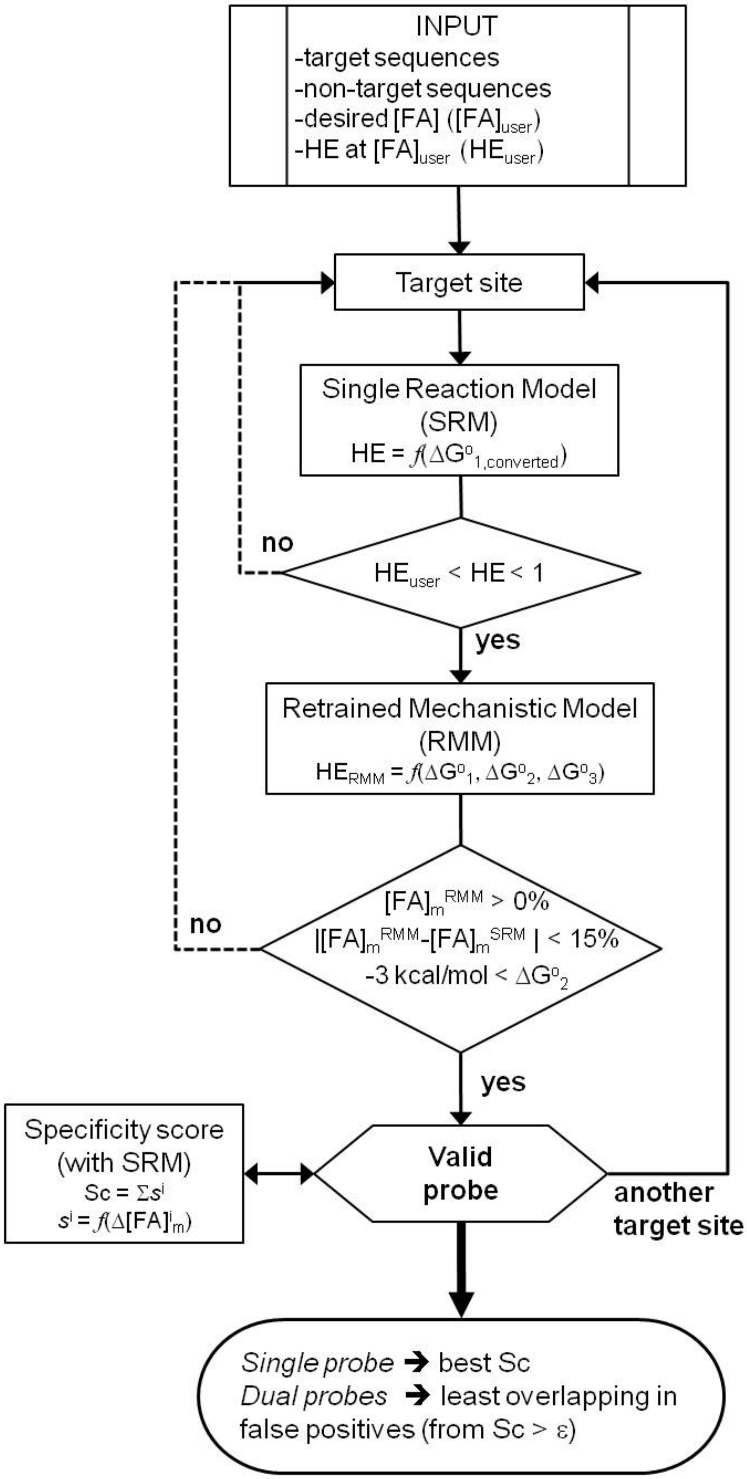
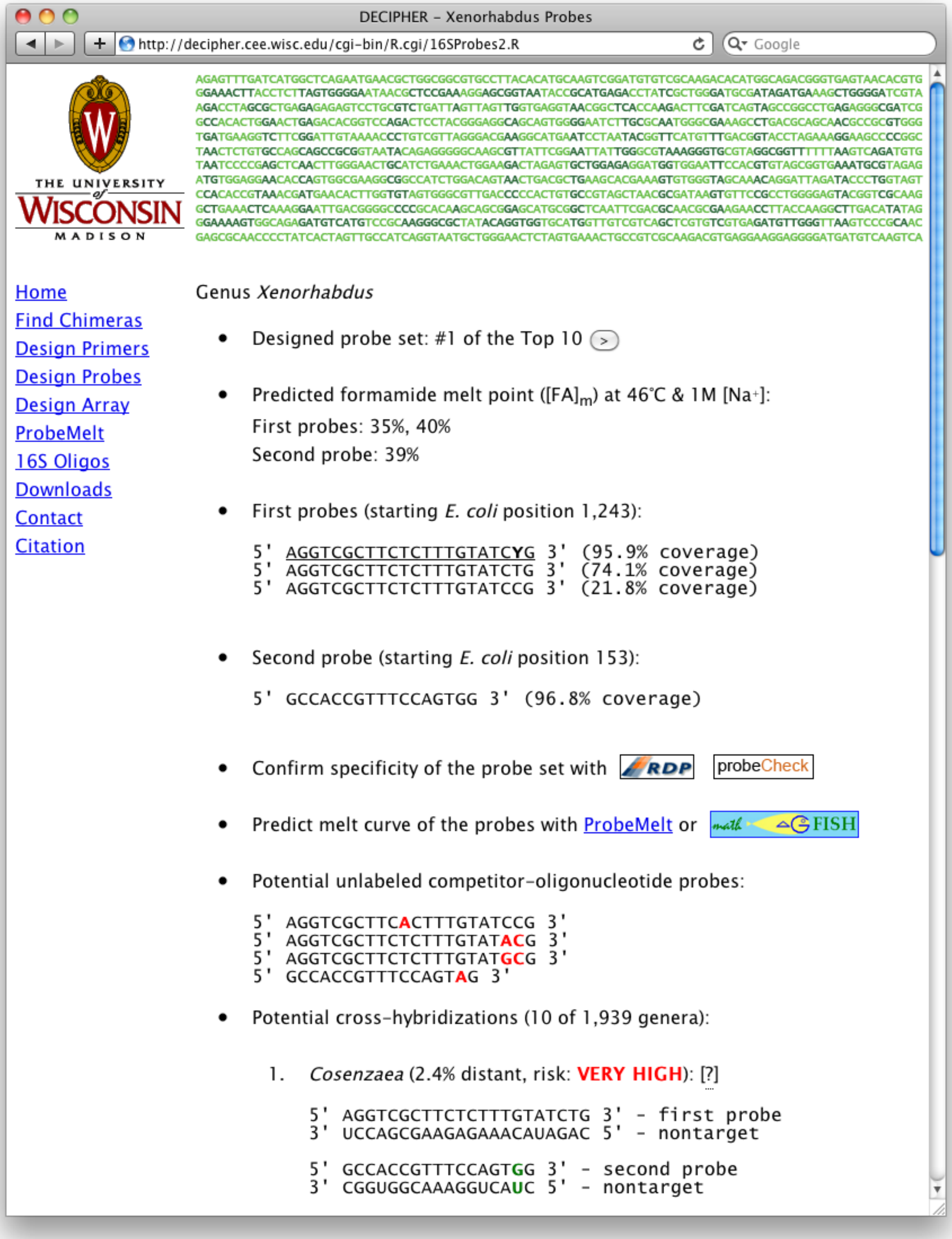

INPUT
-target sequences
-non-target sequences
-desired [FA] ([FA]$_{user}$)
-HE at [FA]$_{user}$ (HE$_{user}$)
Target site
Single Reaction Model
(SRM)
HE = $f(\Delta G^{o}_{1,converted})$
no
HE$_{user}$ < HE < 1
yes
Retrained Mechanistic Model
(RMM)
HE$_{RMM}$ = $f(\Delta G^{o}_{1}, \Delta G^{o}_{2}, \Delta G^{o}_{3})$
no
$[FA]_{m}^{RMM} > 0\%$
$|[FA]_{m}^{RMM}-[FA]_{m}^{SRM}| < 15\%$
-3 kcal/mol < ΔG^{o}_{2}
yes
Specificity score
(with SRM)
Sc = Σs^{i}
$s^{i} = f(\Delta[FA]^{i}_{m})$
Valid
probe
another
target site
Single probe ➔ best Sc
Dual probes ➔ least overlapping in
false positives (from Sc > ε)

Figure S6. Probe design algorithm. HE and *f* denote hybridization efficiency and a function, respectively. Solid feedback arrow indicates the use of a different target site, while dashed arrows indicate either a switch to a new target site or the change in the length of the existing target site. See text for details.


DECIPHER – Xenorhabdus Probes
http://decipher.cee.wisc.edu/cgi-bin/R.cgi/16SProbes2.R
Google
THE UNIVERSITY of WISCONSIN MADISON

```
AGAGTTTGATCATGGCTCAGAATGAACGCTGGCGGCGTGCCTTACACATGCAAGTCGGATGTGTCGCAAGACACATGGCAGACGGGTGAGTAACACGTG
GGAAACTTACCTCTTAGTGGGGAATAACGCTCCGAAAGGAGCGGTAATACCGCATGAGACCTATCGCTGGGATGCGATAGATGAAAGCTGGGGATCGTA
AGACCTAGCGCTGAGAGAGAGTCCTGCGTCTGATTAGTTAGTTGGTGAGGTAACGGCTCACCAAGACTTCGATCAGTAGCCGGCCTGAGAGGGCGATCG
GCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATCTTGCGCAATGGGCGAAAGCCTGACGCAGCAACGCCGCGTGGG
TGATGAAGGTCTTCGGATTGTAAAACCCTGTCGTTAGGGACGAAGGCATGAATCCTAATACGGTTCATGTTTGACGGTACCTAGAAAGGAAGCCCCGGC
TAACTCTGTGCCAGCAGCCGCGGTAATACAGAGGGGGCAAGCGTTATTCGGAATTATTGGGCGTAAAGGGTGCGTAGGCGGTTTTTTAAGTCAGATGTG
TAATCCCCGAGCTCAACTTGGGAACTGCATCTGAAACTGGAAGACTAGAGTGCTGGAGAGGATGGTGGAATTCCACGTGTAGCGGTGAAATGCGTAGAG
ATGTGGAGGAACACCAGTGGCGAAGGCGGCCATCTGGACAGTAACTGACGCTGAAGCACGAAAGTGTGGGTAGCAAACAGGATTAGATACCCTGGTAGT
CCACACCGTAAACGATGAACACTTGGTGTAGTGGGCGTTGACCCCCACTGTGCCGTAGCTAACGCGATAAGTGTTCCGCCTGGGGAGTACGGTCGCAAG
GCTGAAACTCAAAGGAATTGACGGGGCCCCGCACAAGCAGCGGAGCATGCGGCTCAATTCGACGCAACGCGAAGAACCTTACCAAGGCTTGACATATAG
GGAAAAGTGGCAGAGATGTCATGTCCGCAAGGGCGCTATACAGGTGGTGCATGGTTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAAC
GAGCGCAACCCCTATCACTAGTTGCCATCAGGTAATGCTGGGAACTCTAGTGAAACTGCCGTCGCAAGACGTGAGGAAGGAGGGGATGATGTCAAGTCA
```

Home
Find Chimeras
Design Primers
Design Probes
Design Array
ProbeMelt
16S Oligos
Downloads
Contact
Citation

Genus *Xenorhabdus*

- Designed probe set: #1 of the Top 10 >

- Predicted formamide melt point ($[FA]_m$) at 46°C & 1M [Na+]:
  First probes: 35%, 40%
  Second probe: 39%

- First probes (starting *E. coli* position 1,243):

```
5' AGGTCGCTTCTCTTTGTATCYG 3' (95.9% coverage)
5' AGGTCGCTTCTCTTTGTATCTG 3' (74.1% coverage)
5' AGGTCGCTTCTCTTTGTATCCG 3' (21.8% coverage)
```

- Second probe (starting *E. coli* position 153):

```
5' GCCACCGTTTCCAGTGG 3' (96.8% coverage)
```

- Confirm specificity of the probe set with RDP probeCheck

- Predict melt curve of the probes with ProbeMelt or mathFISH

- Potential unlabeled competitor-oligonucleotide probes:

```
5' AGGTCGCTTCACTTTGTATCCG 3'
5' AGGTCGCTTCTCTTTGTATACG 3'
5' AGGTCGCTTCTCTTTGTATGCG 3'
5' GCCACCGTTTCCAGTAG 3'
```

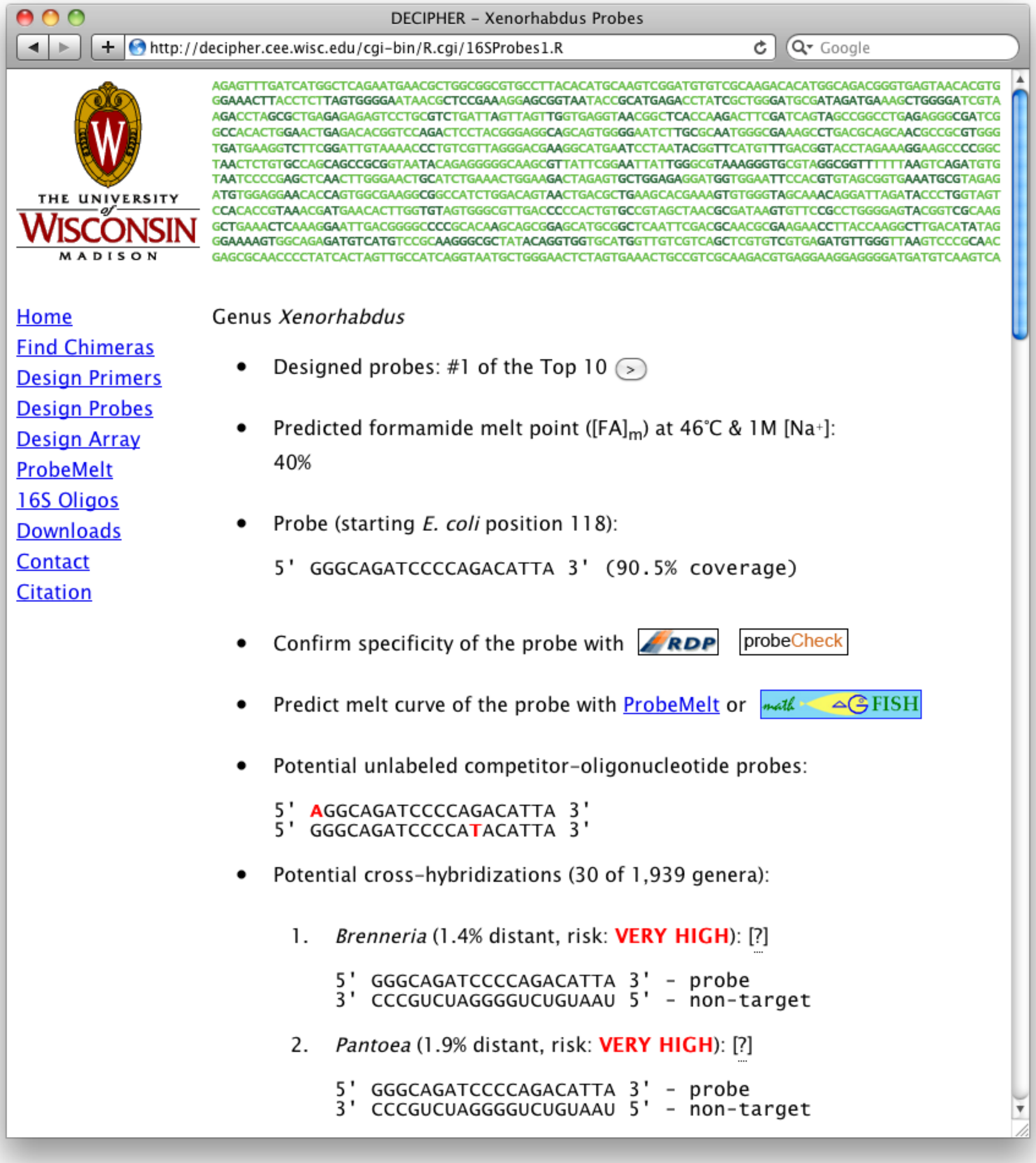
- Potential cross-hybridizations (10 of 1,939 genera):

  1. *Cosenzaea* (2.4% distant, risk: **VERY HIGH**): [?]

```
5' AGGTCGCTTCTCTTTGTATCTG 3' - first probe
3' UCCAGCGAAGAGAAACAUAGAC 5' - nontarget

5' GCCACCGTTTCCAGTGG 3' - second probe
3' CGGUGGCAAAGGUCAUC 5' - nontarget
```

Figure S7. Screenshot of the 16S Oligos page for targeting the genus *Xenorhabdus* using two probes.



```
AGAGTTTGATCATGGCTCAGAATGAACGCTGGCGGCGTGCCTTACACATGCAAGTCGGATGTGTCGCAAGACACATGGCAGACGGGTGAGTAACACGTG
GGAAACTTACCTCTTAGTGGGGAATAACGCTCCGAAAGGAGCGGTAATACCGCATGAGACCTATCGCTGGGATGCGATAGATGAAAGCTGGGGATCGTA
AGACCTAGCGCTGAGAGAGAGTCCTGCGTCTGATTAGTTAGTTGGTGAGGTAACGGCTCACCAAGACTTCGATCAGTAGCCGGCCTGAGAGGGCGATCG
GCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATCTTGCGCAATGGGCGAAAGCCTGACGCAGCAACGCCGCGTGGG
TGATGAAGGTCTTCGGATTGTAAAACCCTGTCGTTAGGGACGAAGGCATGAATCCTAATACGGTTCATGTTTGACGGTACCTAGAAAGGAAGCCCCGGC
TAACTCTGTGCCAGCAGCCGCGGTAATACAGAGGGGGCAAGCGTTATTCGGAATTATTGGGCGTAAAGGGTGCGTAGGCGGTTTTTTAAGTCAGATGTG
TAATCCCCGAGCTCAACTTGGGAACTGCATCTGAAACTGGAAGACTAGAGTGCTGGAGAGGATGGTGGAATTCCACGTGTAGCGGTGAAATGCGTAGAG
ATGTGGAGGAACACCAGTGGCGAAGGCGGCCATCTGGACAGTAACTGACGCTGAAGCACGAAAGTGTGGGTAGCAAACAGGATTAGATACCCTGGTAGT
CCACACCGTAAACGATGAACACTTGGTGTAGTGGGCGTTGACCCCCACTGTGCCGTAGCTAACGCGATAAGTGTTCCGCCTGGGGAGTACGGTCGCAAG
GCTGAAACTCAAAGGAATTGACGGGGCCCCGCACAAGCAGCGGAGCATGCGGCTCAATTCGACGCAACGCGAAGAACCTTACCAAGGCTTGACATATAG
GGAAAAGTGGCAGAGATGTCATGTCCGCAAGGGCGCTATACAGGTGGTGCATGGTTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAAC
GAGCGCAACCCCTATCACTAGTTGCCATCAGGTAATGCTGGGAACTCTAGTGAAACTGCCGTCGCAAGACGTGAGGAAGGAGGGGATGATGTCAAGTCA
```

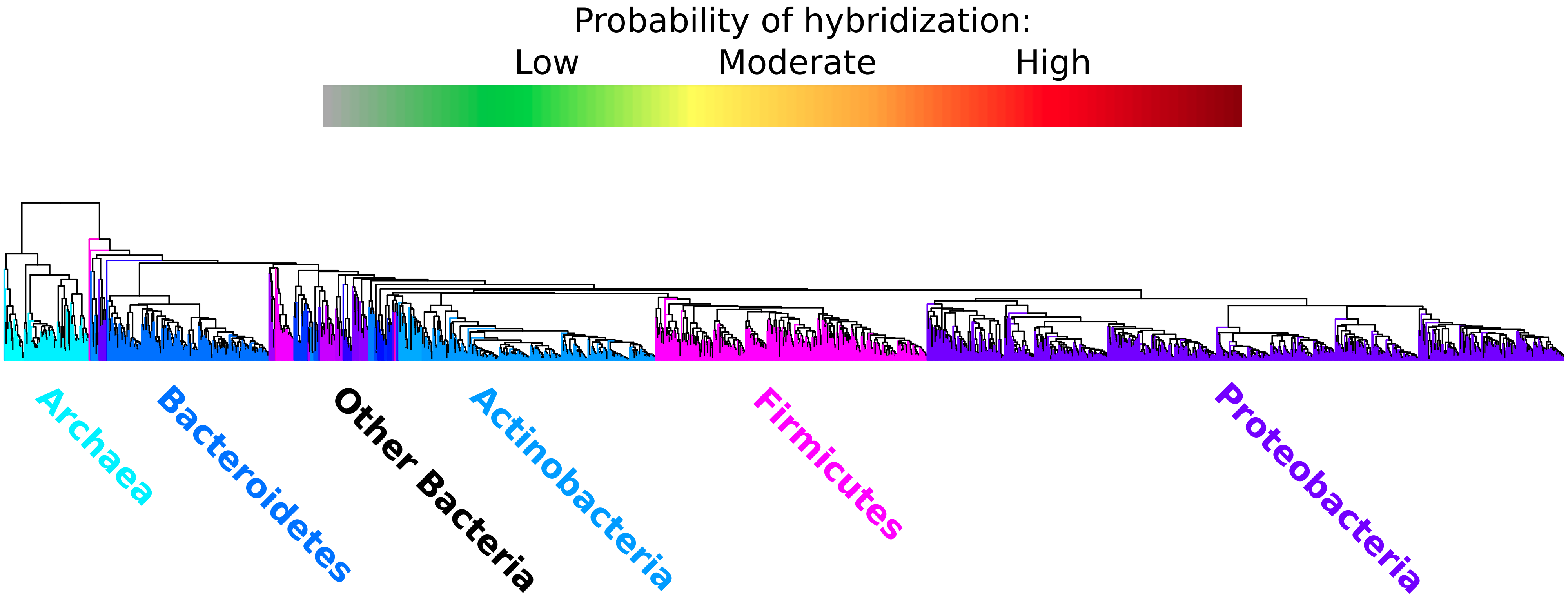
Genus *Xenorhabdus*

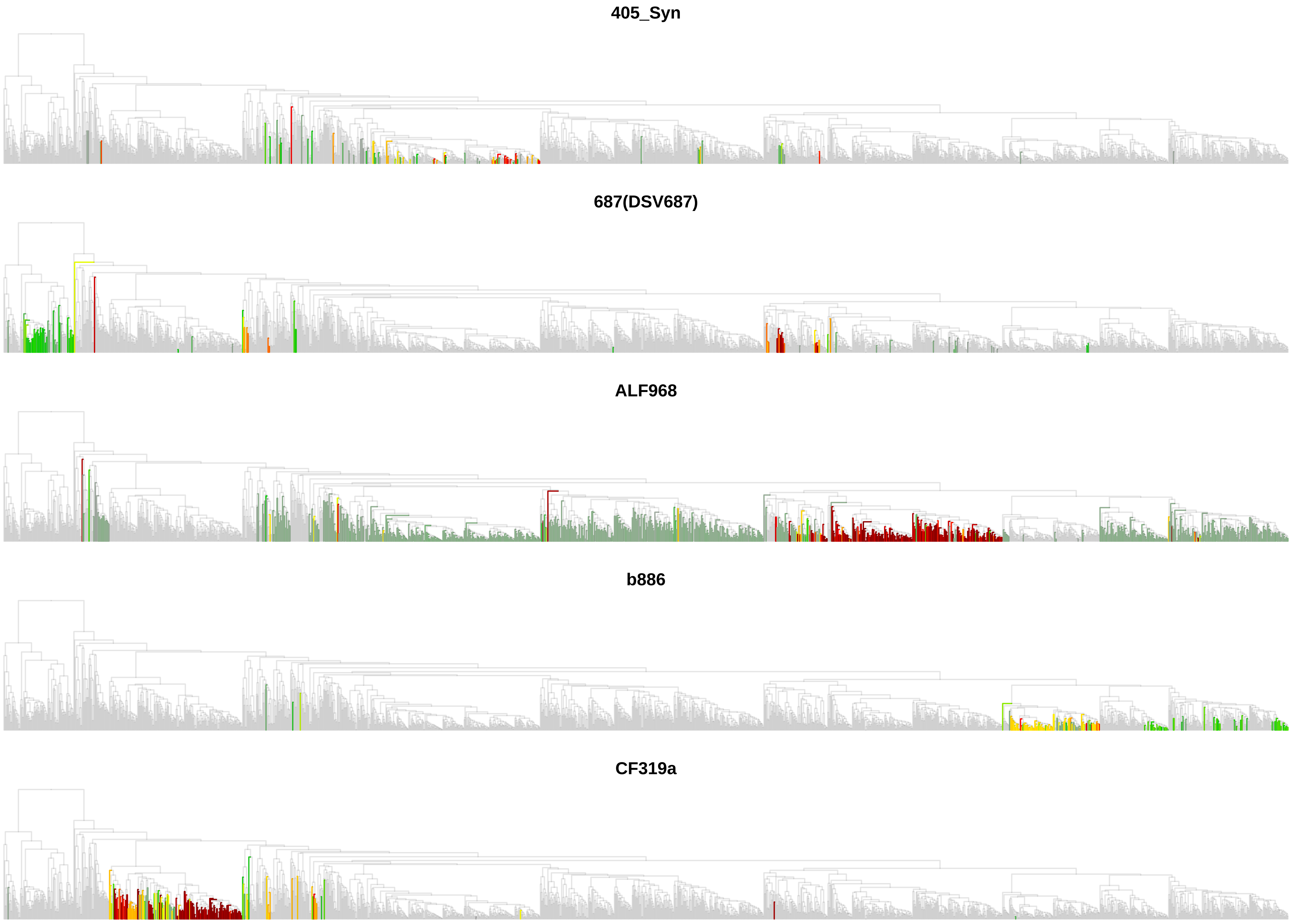
- Designed probes: #1 of the Top 10 >
- Predicted formamide melt point ($[FA]_m$) at 46°C & 1M [Na+]:
  40%
- Probe (starting *E. coli* position 118):
  ```
  5' GGGCAGATCCCCAGACATTA 3' (90.5% coverage)
  ```
- Confirm specificity of the probe with RDP probeCheck
- Predict melt curve of the probe with ProbeMelt or mathFISH
- Potential unlabeled competitor-oligonucleotide probes:
  ```
  5' AGGCAGATCCCCAGACATTA 3'
  5' GGGCAGATCCCCATACATTA 3'
  ```
- Potential cross-hybridizations (30 of 1,939 genera):

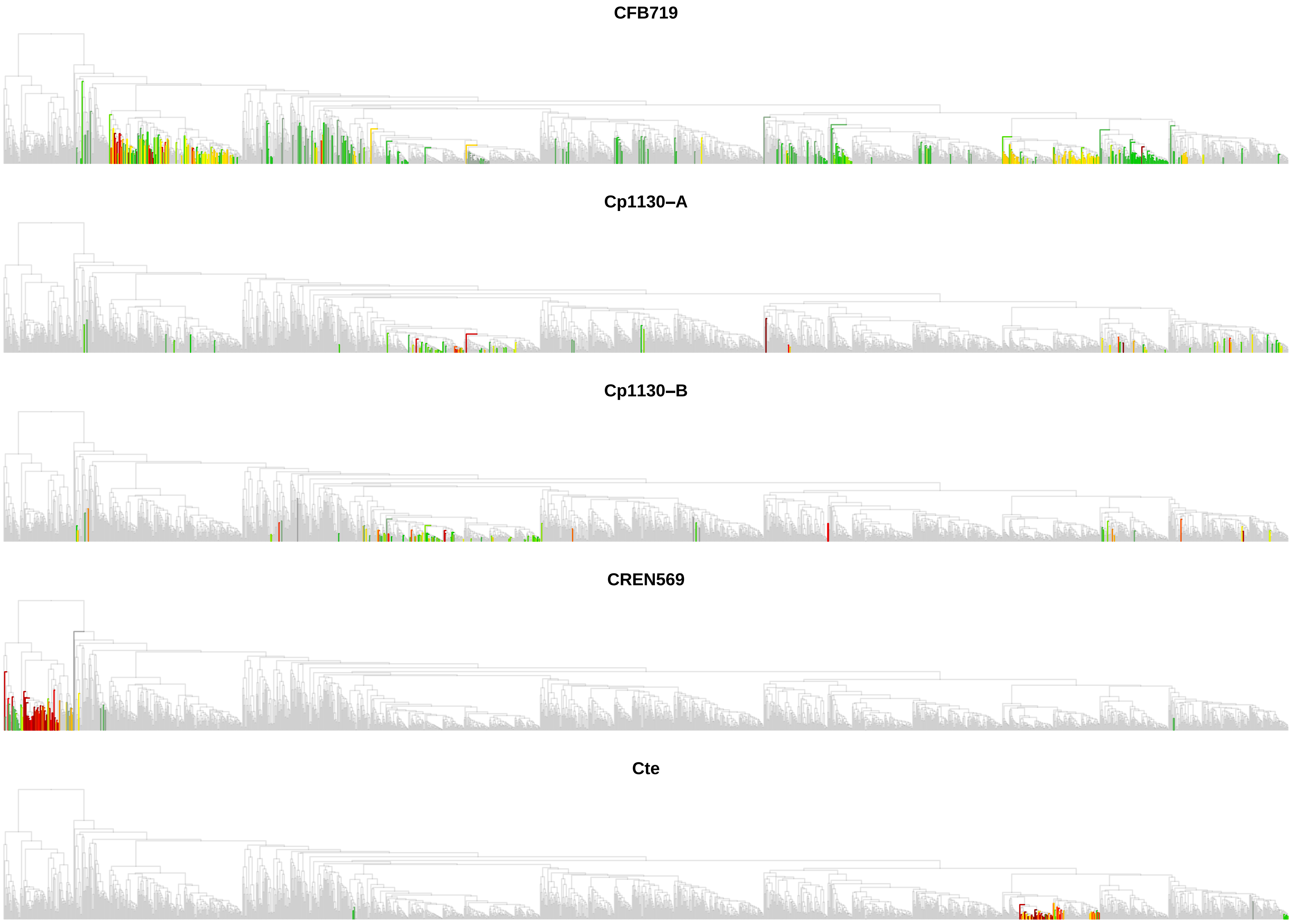
  1. *Brenneria* (1.4% distant, risk: **VERY HIGH**): [?]
     ```
     5' GGGCAGATCCCCAGACATTA 3' - probe
     3' CCCGUCUAGGGGUCUGUAAU 5' - non-target
     ```
  2. *Pantoea* (1.9% distant, risk: **VERY HIGH**): [?]
     ```
     5' GGGCAGATCCCCAGACATTA 3' - probe
     3' CCCGUCUAGGGGUCUGUAAU 5' - non-target
     ```

Figure S8. Screenshot of the 16S Oligos page for targeting the genus *Xenorhabdus* using one probe.

Figure S9: Selected phylogenetic trees showing probability of hybridization with each probe in probeBase. Color of each leaf represents the probability of hybridizing with sequences in each genus. See main text for details.




405_Syn
687(DSV687)
ALF968
b886
CF319a



CFB719

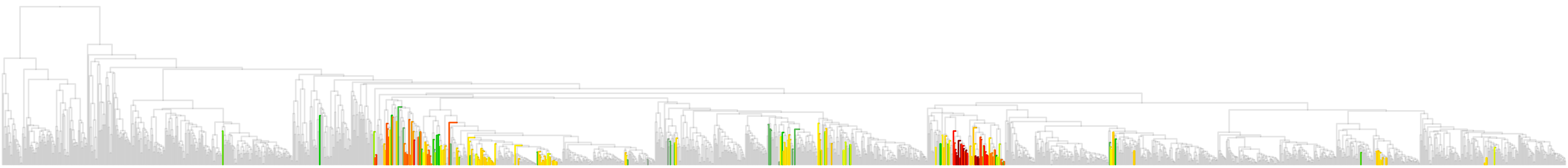
Cp1130–A

Cp1130–B

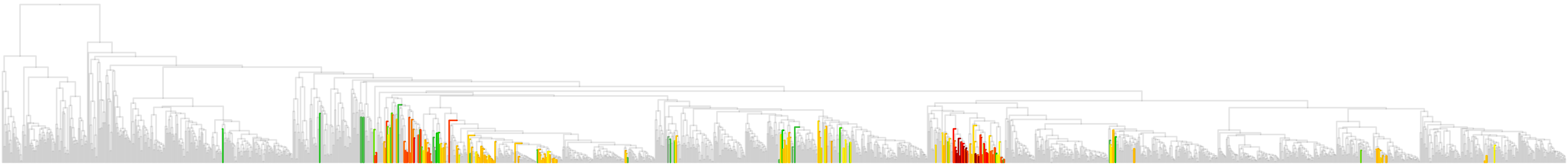
CREN569

Cte

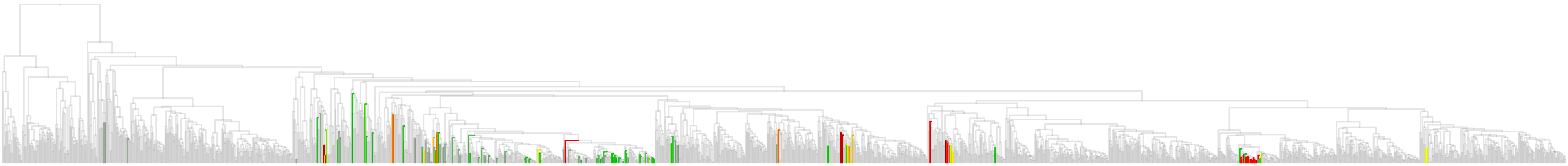
DSBAC355



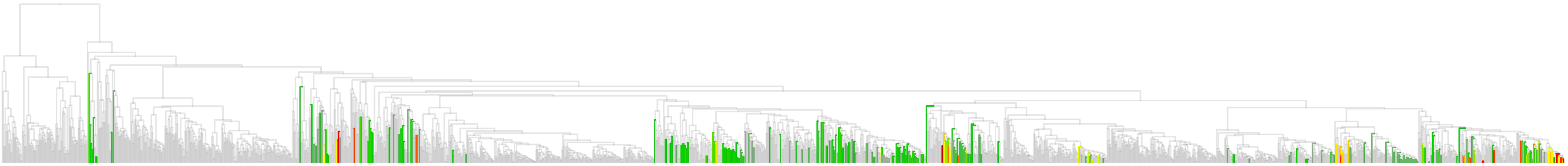
DSBAC357



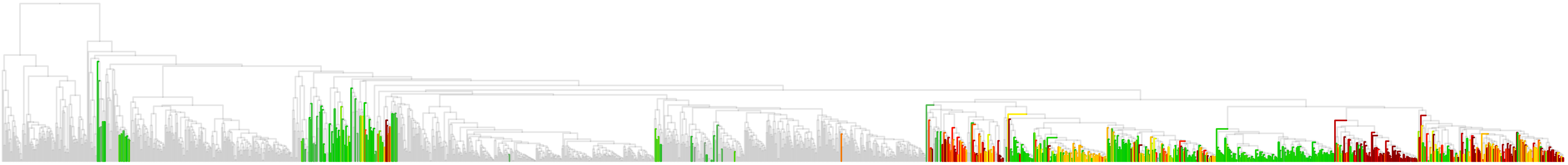
DSV407



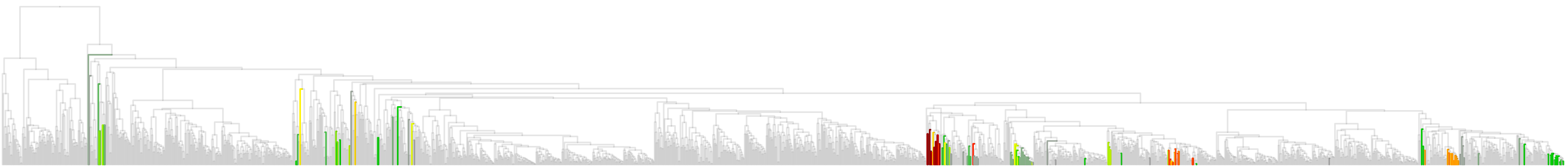
DSV698



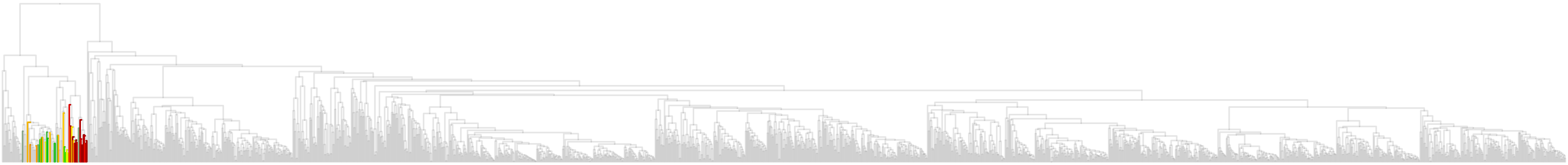
E11



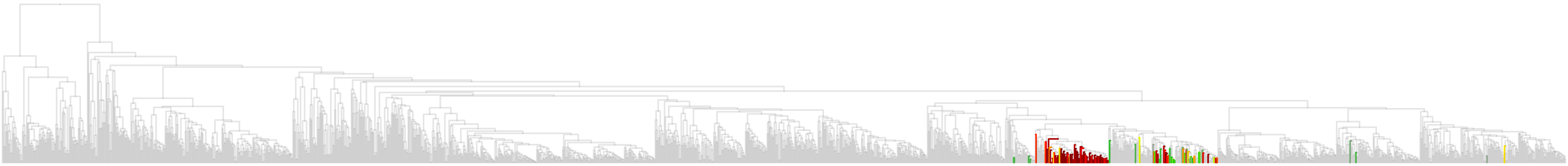
EPSY549



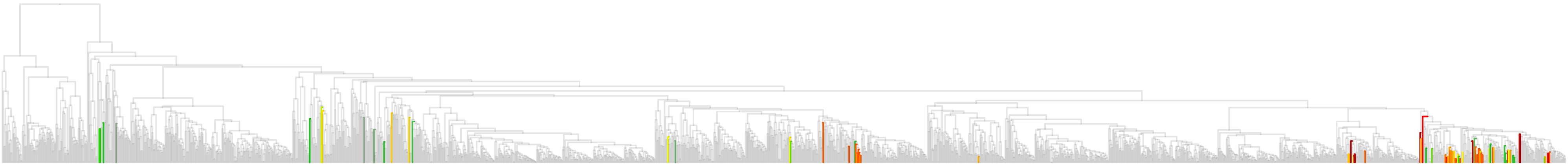
EURY499



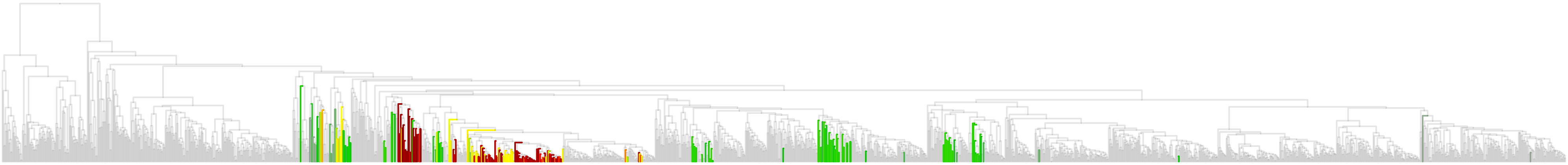
G Rb

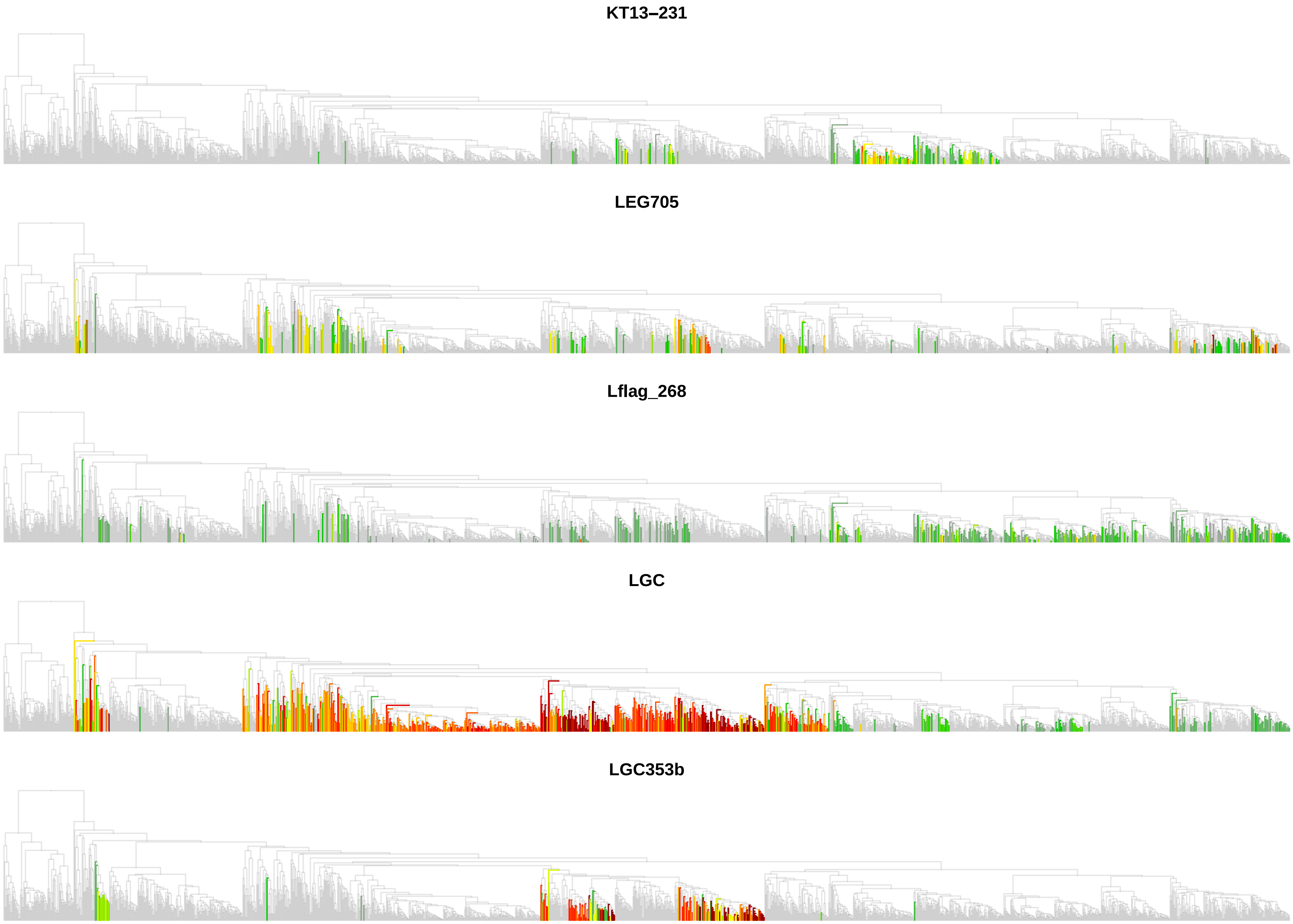


G123T



HoAc1402





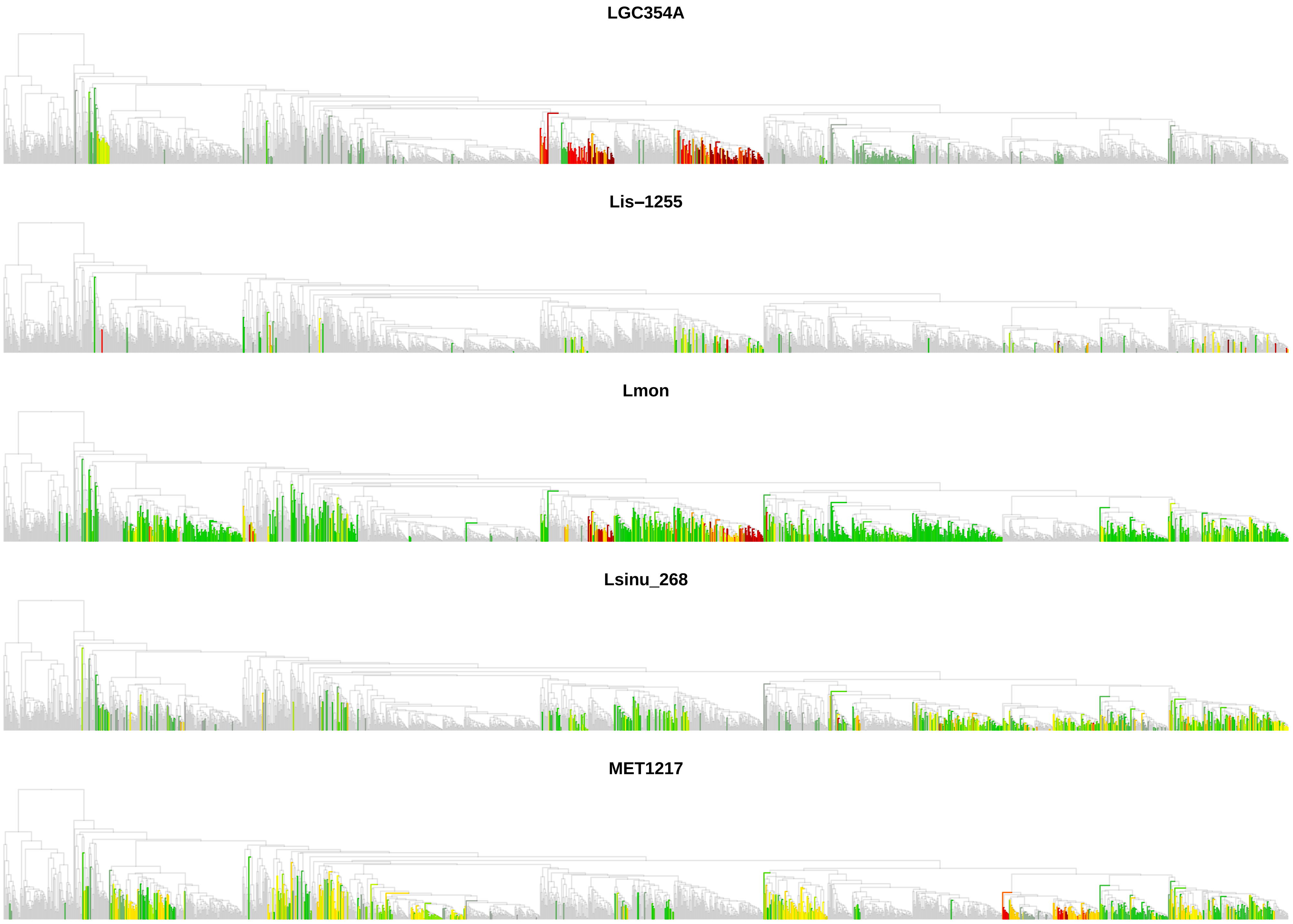
KT13–231

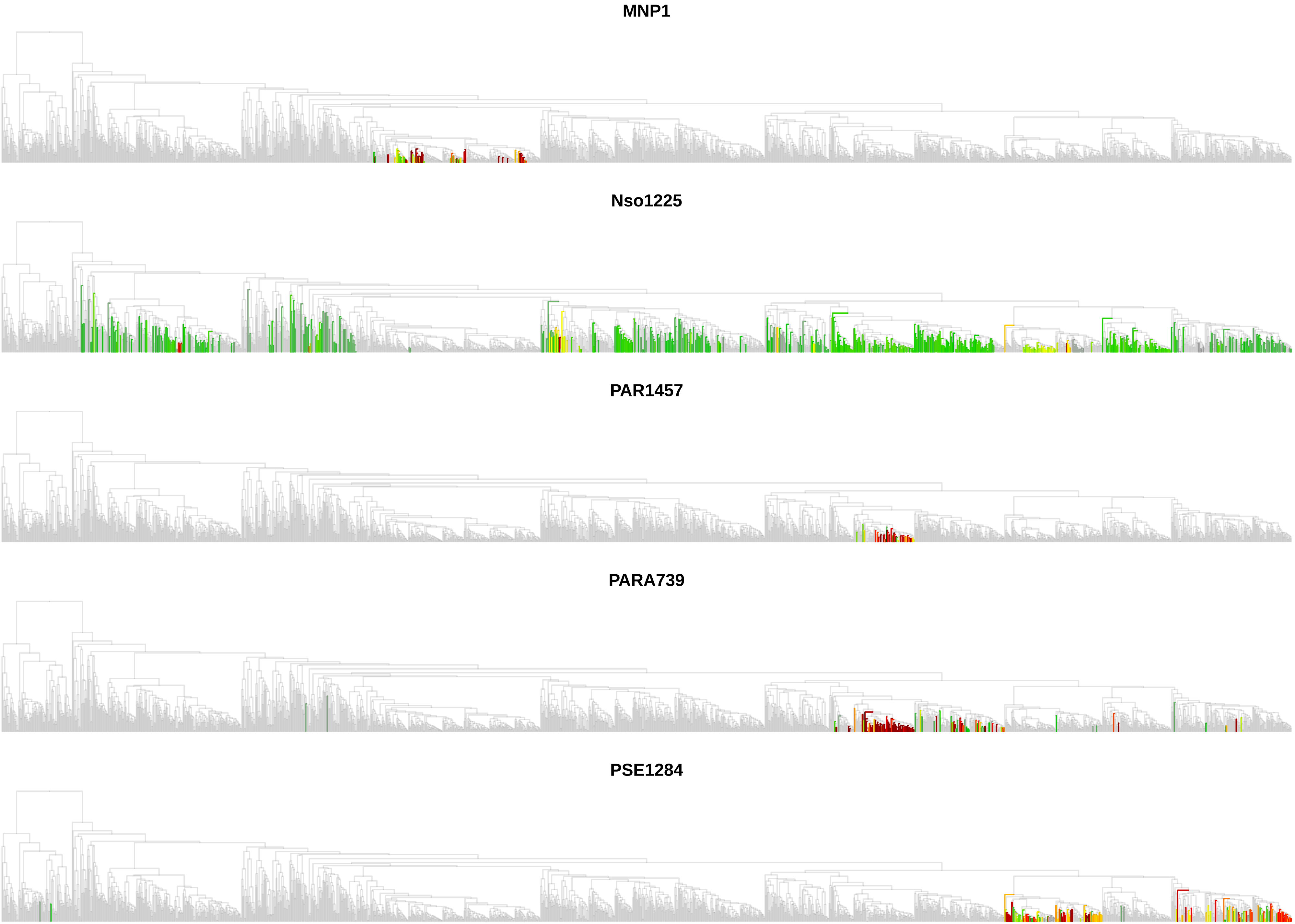
LEG705

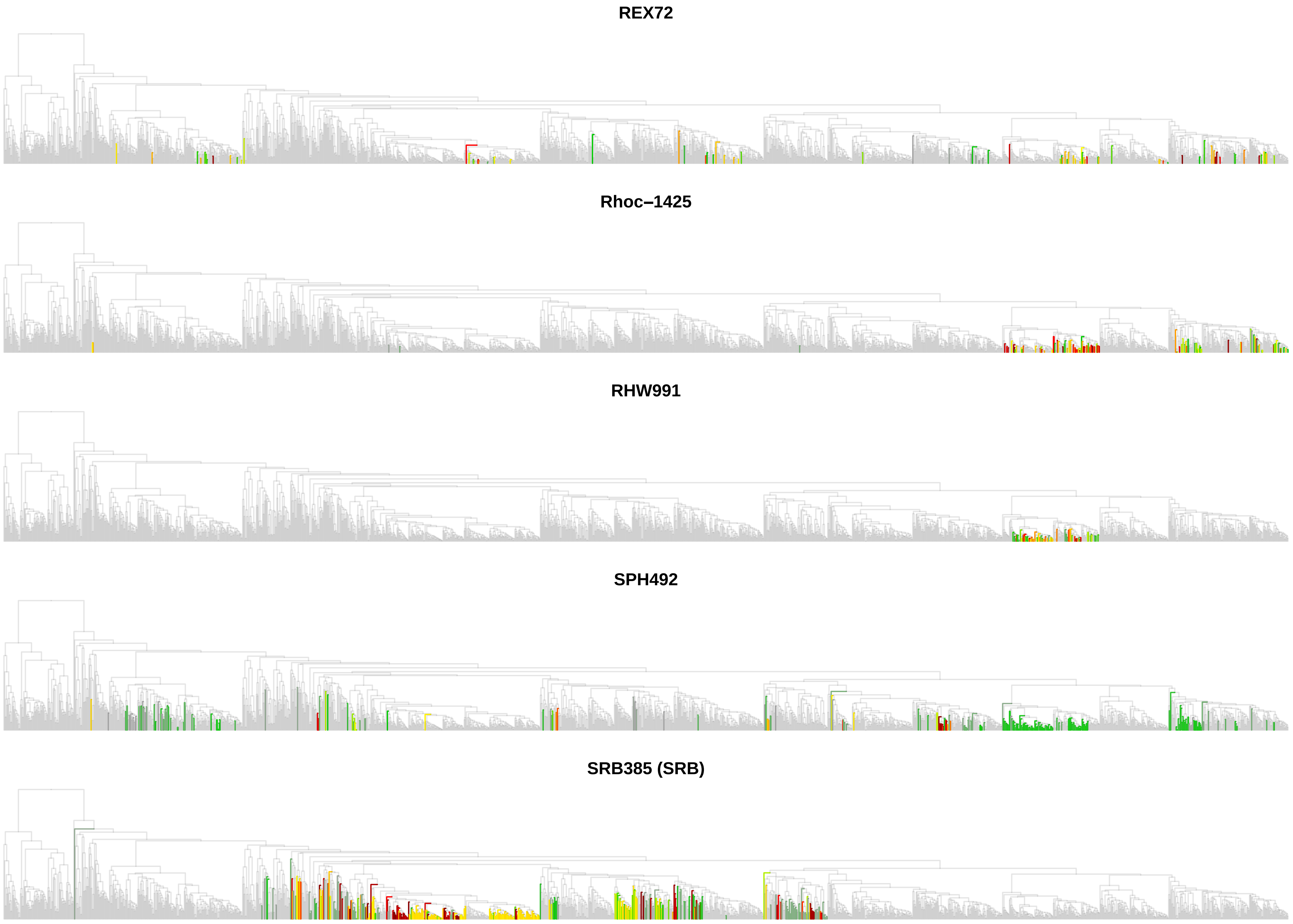
Lflag_268

LGC

LGC353b


LGC354A
Lis–1255
Lmon
Lsinu_268
MET1217


MNP1
Nso1225
PAR1457
PARA739
PSE1284



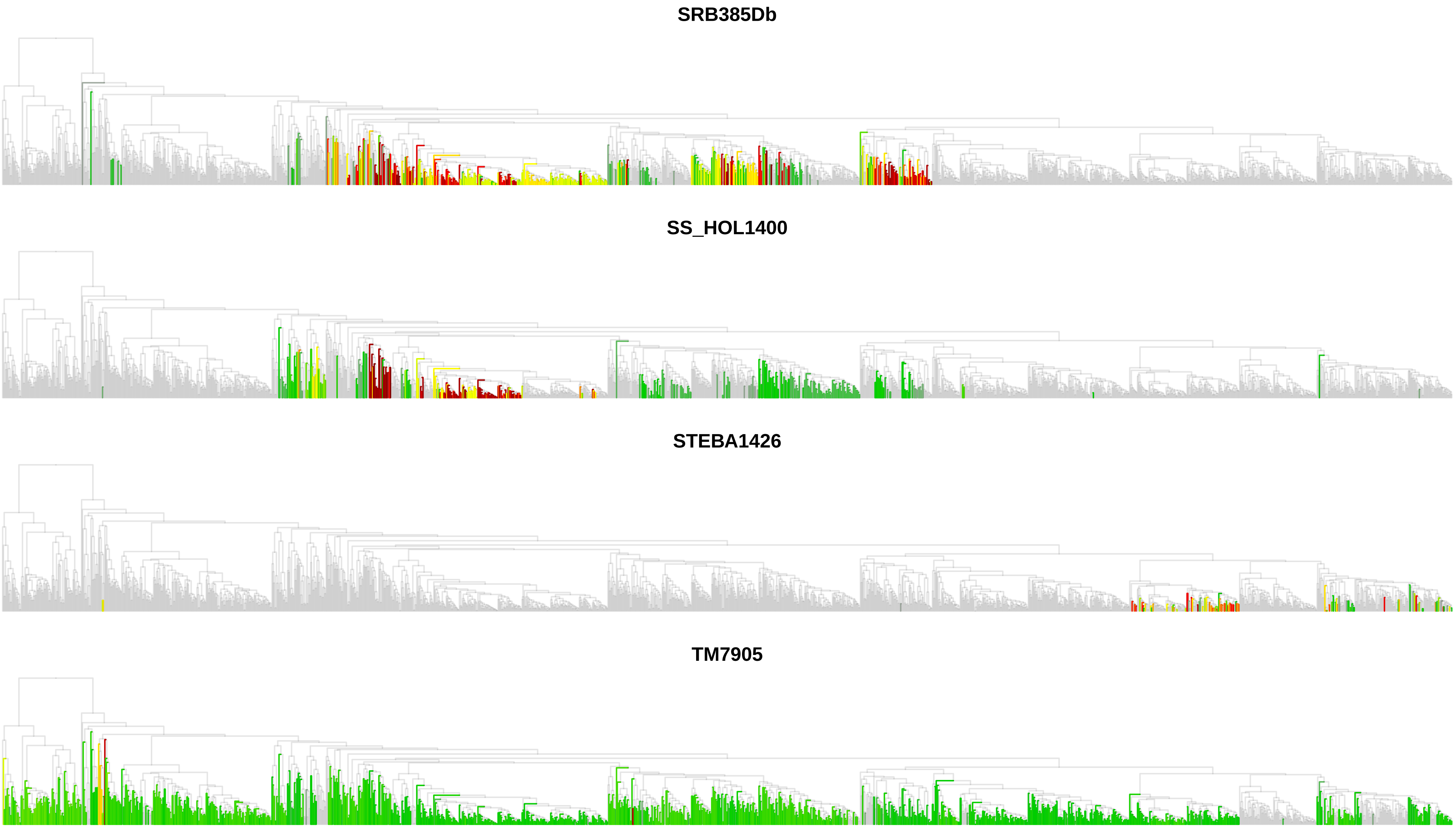
REX72

Rhoc–1425

RHW991

SPH492

SRB385 (SRB)



SRB385Db

SS_HOL1400

STEBA1426

TM7905