Mathematical & statistical foundations of MIMEAnTo.

The MIMEAnTo software processes MIME-generated data. Specifically, the input consists of base counts at each position, derived from next-generation sequencing after alignment, in the pool of selected- and non-selected RNA (see Fig. 1 herein and [1]). In the following, we will refer to bound and unbound samples, exemplarily for an RNA whose function is defined in terms of (protein-) binding. The software then (i) translates these counts into a quantitative effect associated with each mutation m at each nucleotide position i. This quantitative output is the relative effect of a particular mutation m at nucleotide position i on the dissociation constant with respect to e.g. protein binding, denoted by $\frac{\mathrm{Kd}_m}{\mathrm{Kd}_w}(i)$. This measure will be evaluated for all possible mutations and all nucleotide positions. Obviously, the binding site/region in the RNA is the area where mutations have the strongest impact on binding affinity. Secondly, (ii) the statistical significance of the $\frac{\mathrm{Kd}_m}{\mathrm{Kd}_w}(i)$ estimates is assessed. To do this, we use a re-sampling-like procedure that is already part of the input data.

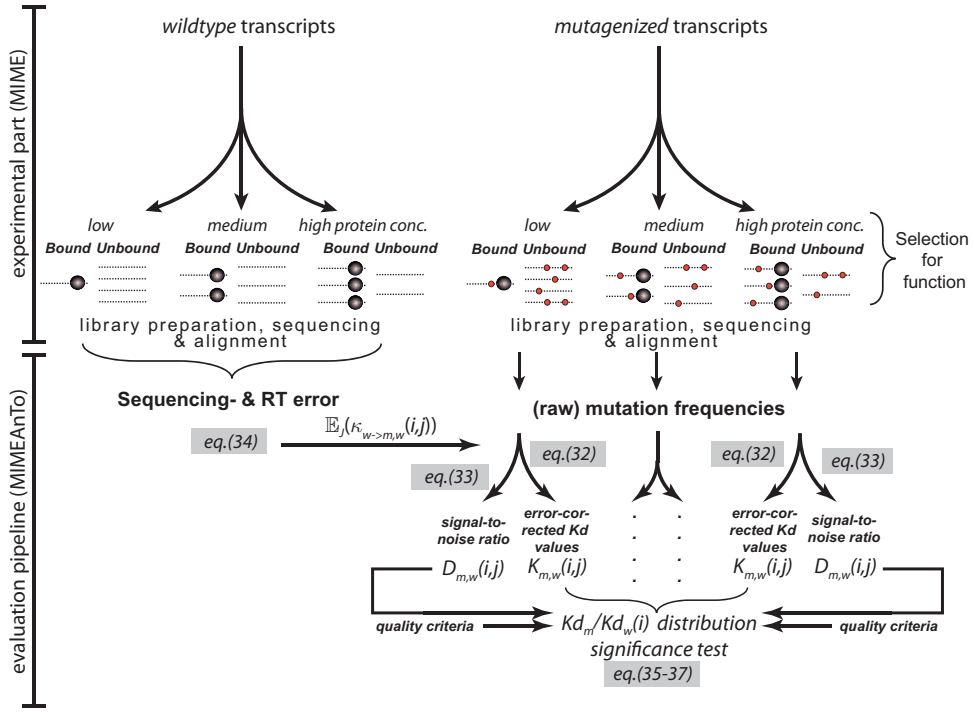

Figure 1. Mathematical analysis pipeline. Important quantities and equations are highlighted and can be found within this document. In brief: wild type libraries are sequenced after functional separation (bound/unbound). From the bound and unbound samples, position- and mutation specific detection errors $\mathbb{E}_J\left(\kappa_{w\to m,w}(i,j)\right)$ (combined RT- & sequencing error) are derived. They are subsequently used to correct re-sampled Kd estimates from the mutant libraries and to derive a signal-to-noise ratio for each position i, co-varying position j and mutation m. The signal-to-noise ratio, together with additional quality criteria can be used to detect and filter out unreliable/insufficient signals. The analysis pipeline yields the re-sampling distribution for $\frac{\mathrm{Kd}_m}{\mathrm{Kd}_w}(i)$ and its statistical ascertainment.

A sketch of the analysis pipeline is shown in Figure 1.
This document is organized as follows:

1. We derive the mathematical concepts that translate the sequencing output into relative Kd values.

Since the experimental procedure (sequencing and reverse transcription) may introduce a substantial number of falsely detected mutations (errors), these errors need to be quantified. The relative Kd values are corrected for these errors and for each mutation m and nucleotide position i, a signal-to-noise ratio $D_m(i)$ is derived. Note, that this procedure yields one point estimate for each mutation m at each nucleotide position i. Thus, although it would be possible to infer the effect of each mutation on binding, it is not possible to evaluate the statistical certainty of this estimate using this *simple* approach.

2. In order to assess the statistical certainty of the relative Kd estimates, a non-parametric method, based on re-sampling (which is inherent in the data), is employed. To accommodate the re-sampling procedure, the previously developed mathematical framework is extended. The results from the re-sampling analysis are reported by MIMEAnTo. The analysis pipeline referring to this *extended* approach is shown in Figure 1.

1 Mathematical & statistical concepts

1.1 Relation between nucleotide frequency and relative Kd values.

The basic reaction scheme underlying the competitive binding experiment that separates RNA by binding affinity is shown in Fig. 2 (left). In the graphic, the differentially colored ★ symbols indicate the presence of a particular mutation at a specific nucleotide position in the RNA.

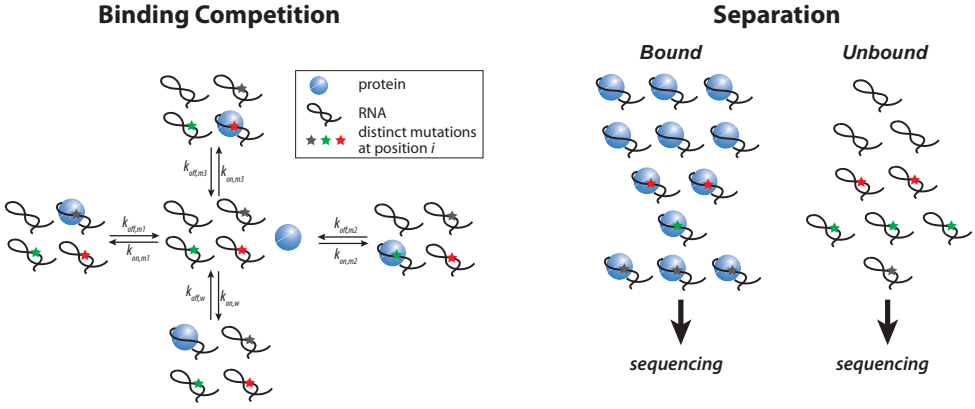

Figure 2. Reaction scheme underlying the competition experiment that selects RNA by binding affinity (left). Bound- and unbound RNA are separated, prepared for sequencing and sequenced to obtain mutation frequencies (right) at each position.

The mass-action kinetics describing this experiment are given by:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{S}_w(\mathrm{i,bound}) &= \mathcal{S}_w(i,\mathrm{unbound})\cdot B\cdot k_{\mathrm{on},w}(i) - \mathcal{S}_w(i,\mathrm{bound})\cdot k_{\mathrm{off},w}(i) && (1)\\
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{S}_w(i,\mathrm{unbound}) &= -\mathcal{S}_w(i,\mathrm{unbound})\cdot B\cdot k_{\mathrm{on},w}(i) + \mathcal{S}_w(i,\mathrm{bound})\cdot k_{\mathrm{off},w}(i) && (2)\\
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{S}_{m1}(i,\mathrm{bound}) &= \mathcal{S}_{m1}(i,\mathrm{unbound})\cdot B\cdot k_{\mathrm{on},m1}(i) - \mathcal{S}_{m1}(i,\mathrm{bound})\cdot k_{\mathrm{off},m1}(i) && (3)\\
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{S}_{m1}(i,\mathrm{unbound}) &= -\mathcal{S}_{m1}(i,\mathrm{unbound})\cdot B\cdot k_{\mathrm{on},m1}(i) + \mathcal{S}_{m1}(i,\mathrm{bound})\cdot k_{\mathrm{off},m1}(i) && (4)\\
&\vdots = \vdots\\
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{S}_{m3}(i,\mathrm{unbound}) &= -\mathcal{S}_{m3}(i,\mathrm{unbound})\cdot B\cdot k_{\mathrm{on},m3}(i) + \mathcal{S}_{m3}(i,\mathrm{bound})\cdot k_{\mathrm{off},m3}(i) && (5)
\end{aligned}$$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}B = {} & +\mathcal{S}_w(i,\mathrm{bound})\cdot k_{\mathrm{off},w}(i) + \mathcal{S}_{m1}(i,\mathrm{bound})\cdot k_{\mathrm{off},m1}(i) + \ldots + \mathcal{S}_{m3}(i,\mathrm{bound})\cdot k_{\mathrm{off},m3}(i)\\
& -\mathcal{S}_w(i,\mathrm{unbound})\cdot B\cdot k_{\mathrm{on},w}(i) \qquad (6)\\
& -\mathcal{S}_{m1}(i,\mathrm{unbound})\cdot B\cdot k_{\mathrm{on},m1}(i) - \ldots - \mathcal{S}_{m3}(i,\mathrm{unbound})\cdot B\cdot k_{\mathrm{on},m3}(i)
\end{aligned}$$

where $\mathcal{S}_w(i,\mathrm{bound})$ denotes the concentration of bound RNA carrying a wild type base at nucleotide position i and $\mathcal{S}_w(i,\mathrm{unbound})$ denotes the concentration of unbound wild type RNA. Correspondingly, the subscripts $m1,...,m3$ indicate the presence of one of the possible three mutations at nucleotide position i. The parameter B denotes the free protein and $k_{\mathrm{off}}(i), k_{\mathrm{on}}(i)$ denote the respective rates of dissociation and association.

In a steady state condition, i.e. if we allow sufficient time to obtain a binding equilibrium, the left hand side (the rate of change) of the equations becomes zero.

$$\begin{aligned}
0 &= \mathcal{S}_w(i,\text{unbound}) \cdot B \cdot k_{\text{on},w}(i) - \mathcal{S}_w(i,\text{bound}) \cdot k_{\text{off},w}(i) && (7)\\
0 &= -\mathcal{S}_w(i,\text{unbound}) \cdot B \cdot k_{\text{on},w}(i) + \mathcal{S}_w(i,\text{bound}) \cdot k_{\text{off},w}(i) && (8)\\
0 &= \mathcal{S}_{m1}(i,\text{unbound}) \cdot B \cdot k_{\text{on},m1}(i) - \mathcal{S}_{m1}(i,\text{bound}) \cdot k_{\text{off},m1}(i) && (9)\\
0 &= -\mathcal{S}_{m1}(i,\text{unbound}) \cdot B \cdot k_{\text{on},m1}(i) + \mathcal{S}_{m1}(i,\text{bound}) \cdot k_{\text{off},m1}(i) && (10)\\
\vdots &= \vdots
\end{aligned}$$

If we solve eq. (7) for the unbound protein B, we obtain

$$B = \frac{\mathcal{S}_w(i,\text{bound}) \cdot k_{\text{off},w}(i)}{\mathcal{S}_w(i,\text{unbound}) \cdot k_{\text{on},w}(i)} = \frac{\mathcal{S}_w(i,\text{bound})}{\mathcal{S}_w(i,\text{unbound})} \cdot \text{Kd}_w(i), \tag{11}$$

which we can substitute back into eq. (9). Note, that we replace the subscript $m1$ with m in the following, because this equation is generic for all possible mutations (m denotes one possible mutation from the set of all possible mutations $\{m1, m2, m3\}$).

$$\begin{aligned}
0 &= \mathcal{S}_m(i,\text{unbound}) \cdot \frac{\mathcal{S}_w(i,\text{bound})}{\mathcal{S}_w(i,\text{unbound})} \cdot \text{Kd}_w(i) \cdot k_{\text{on},m}(i) - \mathcal{S}_m(i,\text{bound}) \cdot k_{\text{off},m}(i)\\
\mathcal{S}_m(i,\text{bound}) \cdot k_{\text{off},m}(i) &= \mathcal{S}_m(i,\text{unbound}) \cdot \frac{\mathcal{S}_w(i,\text{bound})}{\mathcal{S}_w(i,\text{unbound})} \cdot \text{Kd}_w(i) \cdot k_{\text{on},m}(i)\\
\frac{k_{\text{off},m}}{\text{Kd}_w \cdot k_{\text{on},m}}(i) &= \frac{\mathcal{S}_w(i,\text{bound})}{\mathcal{S}_w(i,\text{unbound})} \cdot \frac{\mathcal{S}_m(i,\text{unbound})}{\mathcal{S}_m(i,\text{bound})}\\
\frac{\text{Kd}_m}{\text{Kd}_w}(i) &= \frac{\mathcal{S}_w(i,\text{bound})}{\mathcal{S}_w(i,\text{unbound})} \cdot \frac{\mathcal{S}_m(i,\text{unbound})}{\mathcal{S}_m(i,\text{bound})}, && (12)
\end{aligned}$$

which denotes the impact of a particular mutation m (i.e. $A \to C$, $A \to G$ or $A \to U$ if the wild type base is adenosine) at position i in the RNA sequence on binding affinity.
Note: Unfortunately, the number of bound/unbound sequences $\mathcal{S}$ are not known, instead we derive NGS reads $\mathcal{R}$, subject to errors $\mathcal{X}$ (sequencing and RT-errors). In order to account for this fact, we have to consider the relation between 'reads' and 'sequence numbers', which is exploited in error correction and justification of the proceeding steps.

1.2 Relation between 'Reads' and 'Sequence Numbers'.

For any mutant nucleotide m at nucleotide position i, the number of NGS reads $\mathcal{R}_m(i)$ in the bound/unbound samples is related with the RNA sequence numbers after protein capturing $\mathcal{S}_m(i)$ via

$$\mathcal{R}_m(i) = \vartheta\left(\mathcal{S}_m(i) - \sum_{n \neq m} \mathcal{X}_{m \to n}(i) + \sum_{n \neq m} \mathcal{X}_{n \to m}(i)\right), \tag{13}$$

where $\mathcal{X}_{m \to n}(i)$ is a random variable that denotes the number of sequences that are in fact nucleotide m, but were falsely detected as some other base n. Likewise, $\mathcal{X}_{n \to m}(i)$ is a random variable indicating the number of sequences that were originally some other nucleotide n, but were detected as m due to RT- and sequencing errors. Parameter ϑ denotes the normalization (relative titration) factor (if applied), which guaranteed that equal amounts of protein-captured and non-captured sequences were used in the

NGS machinery. For the ease of reading, we skipped the indicator 'bound'/'unbound', since the equations apply in both cases. Consequently, we get

$$\mathcal{S}_m(i) \quad = \quad \frac{\mathcal{R}_m(i)}{\vartheta} + \sum_{n \neq m} \mathcal{X}_{m\to n}(i) - \sum_{n \neq m} \mathcal{X}_{n\to m}(i). \tag{14}$$

Since the wild type w is much more frequent in most samples (small per nucleotide mutation rate during RNA library preparation), i.e. $\mathcal{S}_w(i) \gg \mathcal{S}_m(i)$ and the error *probability* $\kappa_{n\to m}(i)$ (defined below) may not be vastly different for the distinct types of transitions $n \to m$, we have $\mathcal{X}_{m\to n}(i) \ll \mathcal{X}_{w\to m}(i) \gg \mathcal{X}_{k\to m}(i)$ for any $w \neq k \neq m$. Therefore, we can neglect all false detections, except those, where the wild type w was falsely detected as some mutant m. The expression simplifies accordingly:

$$\boxed{\mathcal{S}_m(i) \approx \frac{\mathcal{R}_m(i)}{\vartheta} - \mathcal{X}_{w\to m}(i)} \tag{15}$$

and likewise for the wild type

$$\mathcal{S}_w(i) \quad \approx \quad \frac{\mathcal{R}_w(i)}{\vartheta} + \sum_{m \in \mathbb{M}} \mathcal{X}_{w\to m}(i) \tag{16}$$

since $\kappa_{w\to m}(i) = \frac{\mathcal{X}_{w\to m}(i)}{\mathcal{S}_w(i)} \ll 1$ (the probability of false detection is small, typically $< 10^{-3}$ in Illumina machines) and $\vartheta \leq 1$ (only a fraction of the sample is taken for sequencing), the last equation simplifies further, as $\sum \mathcal{X}_{w\to m}(i) \ll \frac{\mathcal{R}_w(i)}{\vartheta}$:

$$\boxed{\mathcal{S}_w(i) \approx \frac{\mathcal{R}_w(i)}{\vartheta}.} \tag{17}$$

In relation to the inference of binding affinities, we can substitute the 'sequence numbers' with the NGS 'reads' (eqs. (15) and (17)) and obtain

$$\frac{\mathrm{Kd}_m}{\mathrm{Kd}_w}(i) \approx \frac{\mathcal{R}_w(i,\mathrm{bound})}{\mathcal{R}_w(i,\mathrm{unbound})} \cdot \frac{\mathcal{R}_m(i,\mathrm{unbound}) - \mathcal{X}_{w\to m}(i,\mathrm{unbound}) \cdot \vartheta(\mathrm{unbound})}{\mathcal{R}_m(i,\mathrm{bound}) - \mathcal{X}_{w\to m}(i,\mathrm{bound}) \cdot \vartheta(\mathrm{bound})}, \tag{18}$$

From the equation above, it is apparent that the relative Kd estimate is only reliable when $\mathcal{R}_m(i) \gg \mathcal{X}_{w\to m}(i) \cdot \vartheta$, i.e. when the signal $\mathcal{R}_m$ is larger than the *noise* $\mathcal{X}$.

1.3 Error correction.

In the following, we assume that the *noise* $\mathcal{X}$ is multinomially distributed, i.e. $\mathcal{X}_{w\to m}(i) \sim \mathcal{M}\big(\mathcal{S}_w(i), \kappa_{w\to m}(i)\big)$, as proposed elsewhere [2–4]. The expectation value for the number of mutations m of the multinomial distribution is *trials* $\times$ *success probability*. In our context this means that

$$\mathbb{E}\big(\mathcal{X}_{w\to m}(i)\big) = \quad \mathcal{S}_w(i) \cdot \kappa_{w\to m}(i) \approx \quad \frac{\mathcal{R}_w(i)}{\vartheta} \cdot \kappa_{w\to m}(i) \tag{19}$$

where $\kappa_{w\to m}(i)$ denotes the probability of falsely detecting a wild type residue at position i as mutation m.

Using the multinomial model, we may correct the Kd estimation proposed in eq. (18) for the *expected noise*, from eq. (19). This yields:

$$\frac{\mathrm{Kd}_m}{\mathrm{Kd}_w}(i) \quad \approx \quad \frac{\mathcal{R}_w(i,\mathrm{bound})}{\mathcal{R}_w(i,\mathrm{unbound})} \cdot \frac{\mathcal{R}_m(i,\mathrm{unbound}) - \mathbb{E}\big(\mathcal{X}_{w\to m}(i,\mathrm{unbound})\big) \cdot \vartheta(\mathrm{unbound})}{\mathcal{R}_m(i,\mathrm{bound}) - \mathbb{E}\big(\mathcal{X}_{w\to m}(i,\mathrm{unbound})\big) \cdot \vartheta(\mathrm{bound})} \tag{20}$$

$$= \quad \frac{\mathcal{R}_w(i,\mathrm{bound})}{\mathcal{R}_w(i,\mathrm{unbound})} \cdot \frac{\mathcal{R}_m(i,\mathrm{unbound}) - \kappa_{w\to m}(i) \cdot \mathcal{R}_w(i,\mathrm{unbound})}{\mathcal{R}_m(i,\mathrm{bound}) - \kappa_{w\to m}(i) \cdot \mathcal{R}_w(i,\mathrm{bound})} \tag{21}$$

$$\Rightarrow \boxed{\frac{\mathrm{Kd}_m}{\mathrm{Kd}_w}(i) \approx \frac{\frac{\mathcal{R}_m(i,\mathrm{unbound})}{\mathcal{R}_w(i,\mathrm{unbound})} - \kappa_{w\to m}(i)}{\frac{\mathcal{R}_m(i,\mathrm{bound})}{\mathcal{R}_w(i,\mathrm{bound})} - \kappa_{w\to m}(i)}}, \tag{22}$$

The computation of $\kappa_{w\to m}(i)$ will be explained later in the context of the *Statistical Evaluation* (next section).

1.4 Signal-to-noise ratio.

The 'signal-to-(expected)noise' ratio for each nucleotide position i and each mutant m can be computed according to:

$$D_m(i) \quad = \quad \frac{\mathcal{R}_m(i)}{\mathbb{E}\big(\mathcal{X}_{w\to m}(i)\big)\cdot\vartheta} \approx \frac{\mathcal{R}_m(i)}{\mathcal{R}_w(i)\cdot\kappa_{w\to m}(i)} \tag{23}$$

For all analysis, MIMEAnTo only evaluates data where the signal-to-noise ratio is above a user-defined threshold (see Fig. 1).

2 Statistical Evaluation.

Note, that the above described procedure yields one *point estimate* for the effect of each mutation m at each nucleotide position i on an RNA of length L, on binding (totalling $L \cdot 3$ estimates). Although it is possible to infer the effect of each mutation on binding from the *simple* method above, it is not possible to assess the *statistical certainty* of these estimates, unless vast numbers of repetition experiments were performed (which is expensive and time-consuming). In the following, we develop a non-parametric method to infer the statistical certainty of the relative Kd estimates. This method is based on jacknife-like re-sampling (which is inherent in the data). The re-sampling procedure will allow us to estimate the *probability distribution* of each relative Kd estimate from $L \cdot 3 \cdot N$ point estimates (where N refers to the number of re-samplings). In order to accommodate the re-sampling procedure, we extend the previously developed mathematical framework.
The results from the analysis below will be reported by MIMEAnTo. The analysis pipeline shown in Figure 1 refers to this *extended* approach.

2.1 Re-sampling Procedure

Instead of using the estimates derived above directly, MIMEAnTo uses re-sampling statistics in order to evaluate the *significance* of any relative Kd estimate. The basic idea is to determine binding affinities $\mathrm{Kd}_{m,w}(i,j)$ for each combination of residues (i,j), where the first residue i is mutated and the second residue j is in the wild type configuration (thus having no effect on the *relative* Kd estimate). This allows to re-estimate the effect of a mutation m at position i N-times, i.e. for each i, we can go through all $j \neq i$ (all pairs of residues), see Fig. 3 (left panel).

Since less sequence fragments will cover both i and j the further j lies away from i (see Fig. 3, left panel), the procedure is highly similar to a classic jack-knife re-sampling procedure in which a re-estimation is performed each time after removing one individual from the pool of samples. The re-sampling will then give a non-parametric and unbiased probability distribution of the estimate; –in our case $\mathrm{Kd}_m/\mathrm{Kd}_w(i)$.

Analogous to the previous section, **for each pair** (i,j), we get:

$$\mathrm{K}_{m,w}(i,j) = \quad \frac{\mathrm{Kd}_{m,w}}{\mathrm{Kd}_{w,w}}(i,j) = \quad \frac{\mathcal{S}_{w,w}(i,j,\text{bound})}{\mathcal{S}_{m,w}(i,j,\text{bound})} \cdot \frac{\mathcal{S}_{m,w}(i,j,\text{unbound})}{\mathcal{S}_{w,w}(i,j,\text{unbound})}, \tag{24}$$

where $\mathrm{K}_{m,w}(i,j)$ is a re-sampling estimate of $\mathrm{Kd}_m/\mathrm{Kd}_w(i)$, given the pair of positions (i,j).

In order to accommodate the re-sampling method we will have to extend the previously described method. We will follow the presentation as before. As in the previous section, the number of bound/unbound sequences $\mathcal{S}$ are not known and we have the number of NGS reads $\mathcal{R}$ instead, which are subject to errors $\mathcal{X}$ (sequencing and RT-errors). In order to account for this fact, we have to update the relation between 'reads' and 'sequence numbers' for pairs of residues (i,j).

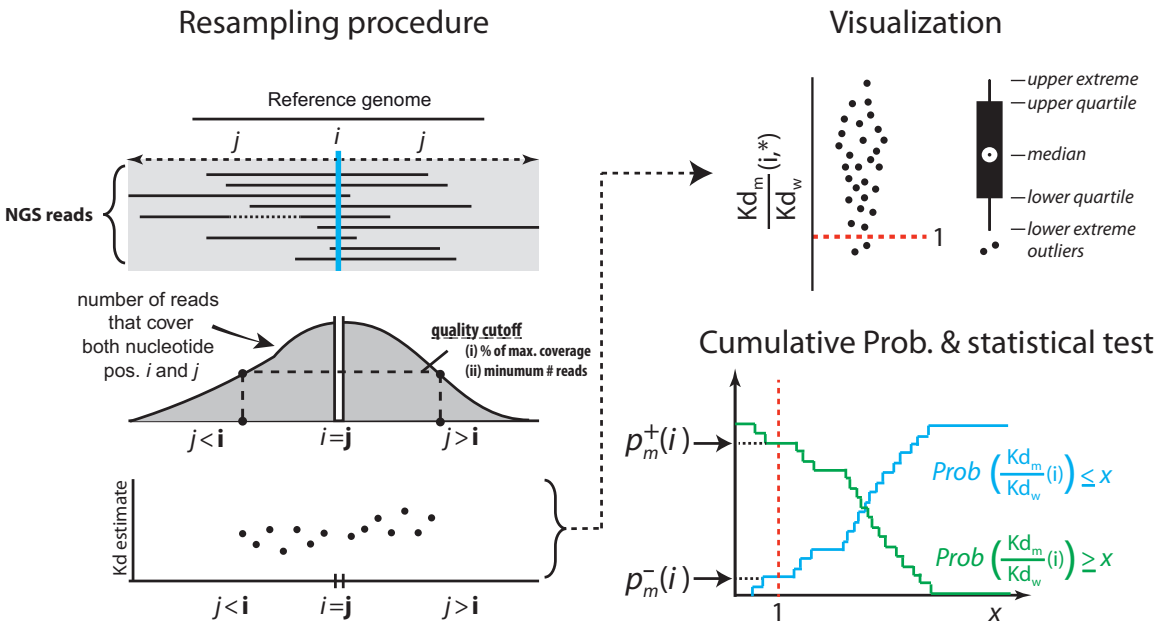

Figure 3. **Left:** Schematic of re-sampling procedure. In order to re-sample the relative Kd value $\mathrm{Kd}_m/\mathrm{Kd}_w(i)$ for mutation m and nucleotide position i, values are computed for pairs of positions (i,j), where position i is mutated and position j is wild type. Each pair of positions (i,j) yields a Kd estimate (bottom). Since the number of sequence reads that span position i and j decreases as the distance between i and j increases (middle), this is analogous to a jacknife procedure. **Right:** The re-sampling distribution for $\mathrm{Kd}_m/\mathrm{Kd}_w(i)$ can be visualized (top) and non-parametric tests (bottom) can be performed in order to detect whether a mutation m at position i significantly increases binding (with p-value: $p_m^+(i)$) or whether it significantly decreases binding in relation to the wild type (with $p_m^-(i)$)

2.2 Relation between 'Reads' and 'Sequence Numbers' for re-sampling relative Kd values.

The number of NGS reads $\mathcal{R}_{m_1,w}(i,j)$ with a mutation in the first site and a wild type in the second m_1, w is related to the RNA sequence numbers after protein capturing $\mathcal{S}_{m_1,w}(i,j)$ via

$$\begin{aligned}\mathcal{R}_{m_1,w}(i,j) \;=\; \vartheta\Big(&\mathcal{S}_{m_1,w}(i,j) - \\ &\sum_{n_1\neq m_1}\mathcal{X}_{m_1\to n_1,w}(i,j) - \sum_{m_2\in\mathbb{M}_2}\mathcal{X}_{m_1,w\to m_2}(i,j) - \sum_{n_1\neq m_1}\sum_{m_2\in\mathbb{M}_2}\mathcal{X}_{m_1\to n_1,w\to m_2}(i,j) \\ &+ \sum_{n_1\neq m_1}\mathcal{X}_{n_1\to m_1,w}(i,j) + \sum_{n_2\neq w}\mathcal{X}_{m_1,n_2\to w}(i,j) + \sum_{n_1\neq m_1}\sum_{n_2\neq w}\mathcal{X}_{n_1\to m_1,n_2\to w}(i,j)\Big),\end{aligned} \tag{25}$$

where e.g. $\mathcal{X}_{m_1\to n_1,w}(i,j)$ denotes the number of sequences with the mutant nucleotide pair $(i = m_1, j = w)$ where the first position i was randomly detected as some other nucleotide $(m_1 \to n_1)$. In analogy with section 1.2, we can assume $\sum_{n_1\neq m_1}\mathcal{X}_{m_1\to n_1,w} \ll \mathcal{X}_{w\to m_1,w} \gg \mathcal{X}_{k_1\to m_1,w}$. Also, since double mutated sequences are assumed to be much less abundant than single mutated sequences, we assume $\sum_{m_2\in\mathbb{M}_2}\mathcal{X}_{m_1,w\to m_2}(i,j) \gg \sum_{n_1\neq m_1}\sum_{m_2\in\mathbb{M}_2}\mathcal{X}_{m_1\to n_1,w\to m_2}(i,j)$ and $\sum_{n_2\neq w}\mathcal{X}_{m_1,n_2\to w}(i,j) \gg$

$\sum_{n_1 \neq m_1} \sum_{n_2 \neq w} \mathcal{X}_{n_1 \to m_1, n_2 \to w}(i,j)$. Therefore, eq. (25) simplifies to:

$$\begin{aligned} \mathcal{R}_{m_1,w}(i,j) &\approx \vartheta\Big(\mathcal{S}_{m_1,w}(i,j) + \mathcal{X}_{w \to m_1,w}(i,j) + \sum_{m_2 \in \mathbb{M}_2} \mathcal{X}_{m_1,w \to m_2}(i,j) + \sum_{n_2 \neq w} \mathcal{X}_{m_1,n_2 \to w}(i,j)\Big), \\ &\approx \vartheta\Big(\mathcal{S}_{m_1,w}(i,j) + \mathcal{X}_{w \to m_1,w}(i,j)\Big), \end{aligned} \tag{26}$$

where the last step is motivated by the fact that the mutation rate in the library is usually low and thus RNAs carrying a mutation at either i or j are less abundant than the double wild type. From here we get

$$\mathcal{S}_{m_1,w}(i,j) \approx \frac{\mathcal{R}_{m_1,w}(i,j)}{\vartheta} - \mathcal{X}_{w \to m_1,w}(i,j). \tag{27}$$

In analogy, we get

$$\begin{aligned} \mathcal{S}_{w,w}(i,j) &\approx \frac{\mathcal{R}_{w,w}(i,j)}{\vartheta} + \sum_{m_2 \in \mathbb{M}_2} \mathcal{X}_{w,w \to m_2}(i,j) + \sum_{m_1 \in \mathbb{M}_1} \mathcal{X}_{w \to m_1,w}(i,j) \\ \mathcal{S}_{w,w}(i,j) &\approx \frac{\mathcal{R}_{w,w}(i,j)}{\vartheta}. \end{aligned} \tag{28}$$

Combining the equations (27) and (28) with eq. (24), we get

$$\mathrm{K}_{m_1,w}(i,j) \approx \frac{\mathcal{R}_{w,w}(i,j,\text{bound})}{\mathcal{R}_{w,w}(i,j,\text{unbound})} \cdot \frac{\mathcal{R}_{m_1,w}(i,j,\text{unbound}) - \vartheta(\text{unbound}) \cdot \mathcal{X}_{w \to m_1,w}(i,j,\text{unbound})}{\mathcal{R}_{m_1,w}(i,j,\text{bound}) - \vartheta(\text{bound}) \cdot \mathcal{X}_{w \to m_1,w}(i,j,\text{bound})}, \tag{29}$$

which is in analogy to eq. (18).

2.3 Error correction.

Once again, we assume that the *noise* $\mathcal{X}$ is multinomially distributed, i.e.:
$\mathcal{X}_{w \to m_1,w}(i,j) \sim \mathcal{M}\big(\mathcal{S}_{w,w}(i,j), \kappa_{w \to m_1,w}(i,j)\big)$. In our context this means that

$$\mathbb{E}\big(\mathcal{X}_{w \to m_1,w}(i,j)\big) = \mathcal{S}_{w,w}(i,j) \cdot \kappa_{w \to m_1,w}(i,j) \approx \frac{\mathcal{R}_{w,w}(i,j)}{\vartheta} \cdot \kappa_{w \to m_1,w}(i,j) \tag{30}$$

Correspondingly we get,

$$\mathrm{K}_{m_1,w}(i,j) \approx \frac{\frac{\mathcal{R}_{m_1,w}(i,j,\text{unbound})}{\mathcal{R}_{w,w}(i,j,\text{unbound})} - \kappa_{w \to m_1,w}(i,j)}{\frac{\mathcal{R}_{m_1,w}(i,j,\text{bound})}{\mathcal{R}_{w,w}(i,j,\text{bound})} - \kappa_{w \to m_1,w}(i,j)} \tag{31}$$

$$\boxed{\mathrm{K}_{m_1,w}(i,j) \approx \frac{\frac{\mathcal{R}_{m_1,w}(i,j,\text{unbound})}{\mathcal{R}_{w,w}(i,j,\text{unbound})} - \mathbb{E}_J\big(\kappa_{w \to m_1}(i)\big)}{\frac{\mathcal{R}_{m_1,w}(i,j,\text{bound})}{\mathcal{R}_{w,w}(i,j,\text{bound})} - \mathbb{E}_J\big(\kappa_{w \to m_1}(i)\big)}} \tag{32}$$

which can be used to re-compute $\mathrm{Kd}_m/\mathrm{Kd}_w(i)$.

2.4 Signal-to-noise ratio.

The 'signal-to-(expected)noise' ratio for mutant nucleotide position i and co-varying wild type position j can be computed according to:

$$\begin{aligned} D_{m_1,w}(i,j) &= \frac{\mathcal{R}_{m_1,w}(i,j)}{\mathbb{E}\big(\mathcal{X}_{w\to m_1,w}(i,j)\big)\cdot\vartheta} \approx \frac{\mathcal{R}_{m_1,w}(i,j)}{\mathcal{R}_{w,w}(i,j)\cdot\kappa_{w\to m_1,w}(i,j)} \\ &\approx \frac{\mathcal{R}_{m_1,w}(i,j)}{\mathcal{R}_{w,w}(i,j)\cdot\mathbb{E}_J\big(\kappa_{w\to m_1}(i)\big)} \end{aligned} \tag{33}$$

For all analysis, we only evaluate data where the signal-to-noise ratio was above a user-defined threshold (see quality criteria in MIMEAnTo).

2.5 Estimation of error probability κ

If experiments with non-mutated RNA are conducted in parallel (see also Fig. 1), we have $\mathcal{S}_{w,m_2}(i,j) = \mathcal{S}_{m_1,w}(i,j) = \mathcal{S}_{m_1,m_2}(i,j) = 0$ and can thus estimate the probability of falsely detecting a wild type residue at position i as some mutant. In fact, the re-sampling scheme allows us to compute statistical properties of the error probability

$$\mathbb{E}_J\big(\kappa_{w\to m_1}(i)\big) \approx N^{-1}\cdot\sum_{j\in J}\frac{\mathcal{R}_{m_1,w}(i,j)}{\mathcal{R}_{w,w}(i,j)} \tag{34}$$

where N denotes the number of co-varying positions $j \in J$ that have sufficient read coverage. Thus, along the same lines as the re-sampling scheme for the relative Kd, we can estimate a confidence range for the error probability $\kappa_{w\to m}(i)$.

Note: The range of estimates for $\kappa_{w\to m_1,w}(i,*)$ and their coefficient of variation can be assessed to justify the use of the maximum likelihood estimate $\mathbb{E}_J\big(\kappa_{w\to m_1}(i)\big)$ in eq. (32) and eq. (33). The coefficient of variation in the re-sampling distribution of $\kappa_{w\to m}(i)$ should be small. Note also that $\mathbb{E}_J\big(\kappa_{w\to m_1}(i)\big)$ can be more reliably estimated from the data than $\kappa_{w\to m_1,w}(i,j)$ (law of large numbers) and that due to the linearity of the expectation value, we may use $\mathbb{E}_J\big(\kappa_{w\to m_1}(i)\big)$ instead of $\kappa_{w\to m_1,w}(i,j)$ in eqs. (32) and (33).

2.6 Quality criteria.

1. In order to statistically evaluate whether a mutation m at position i significantly affects binding, we first apply eq. (32) for all position $j \in J$ that have a sufficient signal-to-noise ratio (see eq. (33)). "Sufficient" is defined as follows: If the ratio is below a user-supplied threshold both in the bound and unbound samples, the corresponding Kd estimate $\mathrm{K}_{m_1,w}(i,j)$ is discarded. If the signal is below the threshold at either the bound- or unbound samples, the respective estimate is tagged as either being a lower- or upper estimate of $\mathrm{K}_{m_1,w}(i,j)$ and assigned (imputed) the value of the median $\mathrm{K}_{m_1,w}(i,*)$ estimate. This has the following reason: If a mutation strongly decreases Kd (increases binding), all sequences carrying this mutation may be bound, and none- or too little amounts of sequence may remain unbound. Thus, $\mathrm{K}_{m_1,w}(i,j)$ may not be accurately determined and we can assume that $\mathrm{K}_{m_1,w}(i,j)$ may in fact be lower than estimable.
   Note: The latter point illustrates why MIME works best when performed with different protein concentrations (see Fig. 1, upper part): I.e. if a mutation increases binding, then at high protein concentrations, all RNAs carrying this mutant may be bound. Subsequently, the signal-to-noise ratio in the unbound sample may be too low. The opposite holds true when a mutation decreases binding. Thus, low protein concentrations, i.e. $B \ll \mathrm{Kd}_w$ are more suitable for quantifying the effect of mutations that increase binding, and high protein concentrations $B \gg \mathrm{Kd}_w$ are best for

quantifying the effect of mutations that decrease binding (Kd_w denotes the dissociation constant of the wild type RNA).
Note: We recommend to use signal-to-noise ratios $D_{m_1,w}(i,j)$ of at least 2 (the *true* signal is at least twice the expected error) in MIMEAnTo. Note that a signal-to-noise ratio below 2 would mean that one tries to read *meaning* into *noise*. The value of the minimum signal-to-noise > 2 should be guided by careful consideration between data *quality* and *quantity*. I.e. the higher $D_{m_1,w}(i,j)$, less evaluable positions are to be expected, where one can predict an exact Kd value. To this end, $D_{m_1,w}(i,j) > 10$ may be unrealistic in most applications. A strong MIME signal should furthermore be robust against a range of $D_{m_1,w}(i,j)$ values.

2. For the re-sampling procedure as illustrated in Fig. 3 (left), only positions j are evaluated where the total number of sequence fragments covering both i and j has at least a user-defined '% of the maximum coverage' (middle panel in Fig. 3, left). Secondly, the total number of reads that have to be available (middle panel in Fig. 3, left) has to exceed a value stated by the user ('minimum coverage criterium'). Both criteria together ensure that each re-sampling is based on a sufficient number of reads and thus provides meaningful estimates.
Note: If the NGS coverage is generally very high ($\geq 1 \cdot 10^6$ per position), then the '% of the maximum coverage' criterium will be more relevant. If the coverage is rather moderate ($\approx 1 \cdot 10^5$ per position), then the 'minimum coverage criterium' will be more relevant. A rule-of-thumb to select the 'minimum coverage criterium' can be derived as follows: Let us assume that the mutation rate in the library is μ on average. We would like to evaluate only mutations m_1 at position i that are present in at least 100 reads. Thus, the 'minimum coverage criterium' should be [minimum coverage] $\geq \frac{100}{\mu}$.

3. The $\mathrm{K}_{m_1,w}(i,j)$ values then give rise to an empirical distribution (see Fig. 3, left (lower panel) and right (upper panel)). A minimum number of re-samplings is required in order to reconstruct the empirical distribution with sufficient confidence (see Fig. 3, lower panel on the right).
Note: In order to statistically evaluate the re-sampling distribution in terms of a p-value (next section), a 'minimum number of $\mathrm{K}_{m_1,w}(i,j)$ estimates' has to fulfill the quality criteria. I.e. if $p < 0.05$, then the number of evaluable $\mathrm{K}_{m_1,w}(i,j)$ values should be $\gg 20$. In general we recommend to use a 'minimum number of estimates' criterium of ≥ 50.

2.7 Statistical test.

The statistical test can subsequently be performed on the re-sampling distribution, see Fig. 3 (lower panel on the right): To test whether a mutation at position i significantly increases Kd/decreases binding, i.e. $\mathcal{H}_0 : \mathrm{K}_m(i) \leq 1$, $\mathcal{H}_1 : \mathrm{K}_m(i) > 1$, the raw p-value (= probability of type I error/false rejection of the null hypothesis) can be computed according to:

$$p_m^-(i) \quad = \quad \frac{\#\mathrm{K}_{m,w}(i,*) \leq 1}{\#\mathrm{K}_{m,w}(i,*)}, \tag{35}$$

where # denotes the 'number of estimates' and '*' indicates that all N positions $j \in J$ are evaluated that pass the quality criteria (previous section). To test if mutation m at position i increases binding, the p-value is calculated according to:

$$p_m^+(i) \quad = \quad \frac{\#\mathrm{K}_{m,w}(i,*) \geq 1}{\#\mathrm{K}_{m,w}(i,*)}. \tag{36}$$

Note that any threshold (e.g. 2-fold increase/decrease, etc.) can be tested.

When several nucleotide positions i are assessed, test corrections need to be performed. All p-values reported by MIMEAnTo are corrected by Benjamini-Hochberg false discovery rate method (BHFDR) [5],

which proceeds as follows: All p-values are sorted in ascending order and p-values are corrected according to

$$\tilde{p}_k = p_k \cdot K/k \tag{37}$$

where p_k and $\tilde{p}_k$ denote the k-smallest raw- and corrected p-value and K denotes the total number of p-values computed.
There is a significant impact of mutation m at nucleotide position i, if $\tilde{p} < \alpha$.
Note: The significance level α (prob. of type I error/*false positive*) is an input parameter in MIMEAnTo. We recommend to use $\alpha = 0.05$.

References

1. Smyth RP, Despons L, Huili G, Bernacchi S, Hijnen M, et al. (2015) Mutational interference mapping experiment (mime) for studying rna structure and function. Nat Methods 12: 866–872.

2. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N (2011) Shorah: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics 12: 119.

3. Prosperi MCF, Prosperi L, Bruselles A, Abbate I, Rozera G, et al. (2011) Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. BMC Bioinformatics 12: 5.

4. Prabhakaran S, Rey M, Zagordi O, Beerenwinkel N, Roth V (2013) Hiv haplotype inference using a propagating dirichlet process mixture model. IEEE/ACM Trans Comput Biol Bioinform .

5. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B Methodological 57: 289–300.