# Supplemental Material – "Structure-based Whole Genome Realignment Reveals Many Novel Non-coding RNAs"

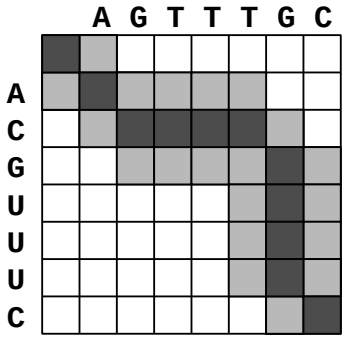
by Sebastian Will, Michael Yu, and Bonnie Berger

## A novel reference-based banding technique for multiple alignment

We introduce a banding technique for multiple sequence alignment. The general idea is to constrain the space of alignments considered to those that deviate by a small amount from a given reference alignment. The parameter $\Delta$ specifies the maximum amount of deviation. Informally, a residue or nucleotide in the reference alignment can be repositioned up to $\Delta$ positions up or downstream. In this way, banding enables a systematic "realignment" of a reference.

Our banding technique can be applied to any dynamic programming (DP) alignment algorithm that recursively computes optimal subalignments. Computation can be visualized over a DP matrix where each entry corresponds to a "cut", i.e. a split into a set of prefix subalignments and a set of suffix subalignments. A "trace" is a continuous path of matrix entries that correspond to a complete alignment. Supplementary Figure 1 illustrates the trace of a reference alignment $\mathcal{A}^{\mathrm{R}}$ and the set of all cuts within a deviation of 1 from $\mathcal{A}^{\mathrm{R}}$. Traces through these cuts correspond to complete alignments within a deviation of 1 from $\mathcal{A}^{\mathrm{R}}$. These alignments are exactly those considered under banding with $\Delta = 1$.

In the following sections we will formalize our banding technique under the notions of cuts and $\Delta$-deviation. We will describe banding for pairwise alignment and extend to progressive multiple alignment. Finally, we will describe the implementation of banding for the structure-based multiple RNA alignment tool LocARNA (Will et al. 2007).




**Supplementary Figure 1.** Matrix visualization of banding in a pairwise alignment of the sequences `ACGUUUC` and `AGTTTGC`. A cut corresponds to an element of an alignment trace. Consequently each cut $(i, j)$ is represented as its corresponding matrix entry (i,j). The trace of a reference alignment $\mathcal{A}^{\mathrm{R}}$ with alignment strings $\mathcal{A}^{\mathrm{R}}_1$=`AC---GUUUC` and $\mathcal{A}^{\mathrm{R}}_2$=`AGTTTG---C` is shown in dark grey, and the cuts within 1-deviation from $\mathcal{A}^{\mathrm{R}}$ in dark and light gray.

## Preliminaries

A *sequence* $S$ is a word over a fixed alphabet $\Sigma$. We reserve a special symbol '$-$'$\notin \Sigma$ to represent a *gap*. A word $T$ over the alphabet $\Sigma \cup \{-\}$ is an *alignment string*. $T_{|\Sigma}$ denotes the sequence obtained by removing all gaps from $T$. The *length of a word* $w$ is denoted by $|w|$ and its *$i$-th character* by $w[i]$. A *$q$-way (multiple) alignment* $\mathcal{A}$ *of length* $m =: |\mathcal{A}|$ is a matrix $\mathcal{A} \subseteq (\Sigma \cup \{-\})^{q \times m}$ with $q =: \mathrm{rows}(\mathcal{A})$ rows. Thus for all $1 \leq x \leq q$, the $x$-th row of $\mathcal{A}$, denoted by $A_x$, is an alignment string. $\mathcal{A}$ is an alignment of the sequences $S_1, \ldots, S_q$ if and only if (1) $S_x = A_{x|\Sigma}$ for all $1 \leq x \leq q$ and (2) no column in $\mathcal{A}$ is *gap-only*, i.e., it consists only of gaps. For example, a 3-way multiple alignment of the sequences $S_{\mathrm{ex1}} =$ `CTA`, $S_{\mathrm{ex2}} =$ `GTT` and $S_{\mathrm{ex3}} =$ `CGTT` is

$$\mathcal{A}_{\mathrm{ex}} = \begin{pmatrix} \texttt{C} & \texttt{-} & \texttt{-} & \texttt{T} & \texttt{A} \\ \texttt{-} & \texttt{G} & \texttt{T} & \texttt{T} & \texttt{-} \\ \texttt{C} & \texttt{G} & \texttt{T} & \texttt{T} & \texttt{-} \end{pmatrix}.$$

The rows of $\mathcal{A}_{\mathrm{ex}}$ are the alignment strings $A_{\mathrm{ex1}} =$ `C--TA`, $A_{\mathrm{ex2}} = -$`GTT`$-$ and $A_{\mathrm{ex3}} =$ `CGTT`$-$.

For a $q$-way alignment $\mathcal{A}$ and a $p$-tuple $\mathcal{I} = (x_1, \ldots, x_p)$ of distinct integers in $\{1, \ldots, q\}$, the *projection*

$\mathcal{A}\langle\mathcal{I}\rangle$ *of* $\mathcal{A}$ *onto* $\mathcal{I}$ is constructed by taking the matrix $(A_{x_1} \dots A_{x_p})^{\mathrm{T}}$ and deleting all gap-only columns. Intuitively, the projection is the subalignment implied by $\mathcal{A}$ on the sequences indexed by $\mathcal{I}$. $\mathcal{A}$ is *an alignment of* $\mathcal{A}^1$ *and* $\mathcal{A}^2$ if $\mathcal{A}^1 = \mathcal{A}\langle 1, \dots, q_1\rangle$ and $\mathcal{A}^2 = \mathcal{A}\langle q_1 + 1, \dots, q\rangle$ for some $1 \leq q_1 < q$. For example, $\mathcal{A}_{\mathrm{ex}}$ is an alignment of $\mathcal{A}^1_{\mathrm{ex}} = \mathcal{A}_{\mathrm{ex}}\langle 1\rangle = \begin{pmatrix} \texttt{C T A} \end{pmatrix}$ and $\mathcal{A}^2_{\mathrm{ex}} = \mathcal{A}_{\mathrm{ex}}\langle 2, 3\rangle = \begin{pmatrix} \texttt{- G T T} \\ \texttt{C G T T} \end{pmatrix}$.

For an alignment string $T$ and a position $j$, $0 \leq j \leq |T|$, the function

$$\mathrm{ctp}_T(j) := |(T[1 \dots j])_{|\Sigma}|$$

maps $j$ to a corresponding position in the sequence $T_{|\Sigma}$. Note that in the case where $T[j]$ is the gap symbol, $\mathrm{ctp}_T(j)$ points to the position left of the gap. The inverse of $\mathrm{ctp}_T$ is $\mathrm{ptc}_T$, mapping positions from $T_{|\Sigma}$ to $T$. For example, $\mathrm{ptc}_{A_{\mathrm{ex}2}}(3) = 4$, $\mathrm{ctp}_{A_{\mathrm{ex}1}}(5) = 3$, and $\mathrm{ctp}_{A_{\mathrm{ex}1}}(3) = 1$.

For a pairwise alignment $\mathcal{A}$ of the sequences $S_1$ and $S_2$, the *cut of* $\mathcal{A}$ *at column* $i$ is the pair

$$c = (c_1, c_2) = (\mathrm{ctp}_{A_1}(i), \mathrm{ctp}_{A_2}(i)).$$

That is, $\mathcal{A}$ can be cut at column $i$ into a prefix and a suffix alignment such that the prefix is an alignment of the sequences $S_1[1 \dots c_1]$ and $S_2[1 \dots c_2]$. The cuts of $\mathcal{A}^2_{\mathrm{ex}}$ at the columns from 0 to 4 are $(0,0), (0,1), (1,2), (2,3)$, and $(3,4)$, respectively. Note that $\mathcal{A}$ is uniquely described by its *set of cuts* $\mathrm{cuts}(\mathcal{A})$.

## $\Delta$-deviation from a reference alignment

Define the *distance between pairwise cuts* $c = (c_1, c_2)$ *and* $c' = (c'_1, c'_2)$ as their Manhattan distance

$$\|c - c'\|_1 = |c_1 - c'_1| + |c_2 - c'_2|.$$

Define the deviation of a cut $c = (c_1, c_2)$ from a reference alignment $\mathcal{A}^{\mathrm{R}}$ as the distance to the closest cut in $\mathcal{A}^{\mathrm{R}}$

$$\mathrm{d}_{\mathcal{A}^{\mathrm{R}}}(c) = \min_{c^{\mathrm{R}} \in \mathrm{cuts}(\mathcal{A}^{\mathrm{R}})} \|c - c^{\mathrm{R}}\|_1.$$

$c$ is said to be within $\Delta$-*deviation* from $\mathcal{A}$ if $\mathrm{d}_{\mathcal{A}^\mathrm{R}}(c) \leq \Delta$. For two pairwise alignments $\mathcal{A}$ and $\mathcal{A}^\mathrm{R}$ over the same sequences, define the *deviation of* $\mathcal{A}$ *from* $\mathcal{A}^R$ as the largest deviation of any cut in $\mathcal{A}$ from $\mathcal{A}^\mathrm{R}$

$$\mathrm{d}_{\mathcal{A}^\mathrm{R}}(\mathcal{A}) = \max_{c \in \mathrm{cuts}(\mathcal{A})} (\mathrm{d}_{\mathcal{A}^\mathrm{R}}(c)).$$

Generalizing to multiple alignments, define the *deviation of* $\mathcal{A}$ *from* $\mathcal{A}^R$ as the maximum deviation over all projections to pairwise subalignments ("maximum-of-pairs"). That is,

$$\mathrm{d}_{\mathcal{A}^\mathrm{R}}(\mathcal{A}) = \max \Big\{ \mathrm{d}_{\mathcal{A}^\mathrm{R}\langle i,j\rangle}(\mathcal{A}\langle i,j\rangle) \mid 1 \leq i < j \leq q \Big\}.$$

In this way, if the deviation of $\mathcal{A}$ from $\mathcal{A}^\mathrm{R}$ is limited by at most $\Delta$, then so is the deviation for all projections to pairwise subalignments. $\mathcal{A}$ is said to be within $\Delta$-*deviation* from $\mathcal{A}^\mathrm{R}$ if $\mathrm{d}_{\mathcal{A}^\mathrm{R}}(A) \leq \Delta$.

## Banding in pairwise alignment algorithms

For two input sequences $S_1$ and $S_2$, consider a DP algorithm, like the classic Needleman-Wunsch, that computes a pairwise alignment $\mathcal{A}$ by recursively computing the scores of prefix subalignments $M_{ij}$. For a given reference alignment $\mathcal{A}^\mathrm{R}$ and deviation parameter $\Delta$, the goal of our banding technique is to find the optimal $\mathcal{A}$ under the constraint that $\mathrm{d}_{\mathcal{A}^\mathrm{R}}(\mathcal{A}) \leq \Delta$. We achieve this goal by restricting computation to the $M_{ij}$ that correspond to a set of cuts, denoted by $\mathcal{C}_2(\mathcal{A}^\mathrm{R}, \Delta)$, that are within $\Delta$-deviation of $\mathcal{A}^\mathrm{R}$. That is,

$$\mathcal{C}_2(\mathcal{A}^\mathrm{R}, \Delta) := \left\{ (i,j) \in \{0,\ldots,|S_1|\} \times \{0,\ldots,|S_2|\} \;\middle|\; \exists c^\mathrm{R} \in \mathrm{cuts}(\mathcal{A}^\mathrm{R})\colon \|c^\mathrm{R} - (i,j)\|_1 \leq \Delta \right\}.$$

This restriction is sufficient because $\mathrm{cuts}(\mathcal{A}) \subseteq \mathcal{C}_2(\mathcal{A}^\mathrm{R}, \Delta)$ if and only if $\mathrm{d}_{\mathcal{A}^\mathrm{R}}(\mathcal{A}) \leq \Delta$.

$\mathcal{C}_2(\mathcal{A}^\mathrm{R}, \Delta)$ consists of elements $(i, \underline{j}_i) \ldots (i, \overline{j}_i)$ for each row $0 \leq i \leq |S_1|$. Therefore, it is convenient to describe it in terms of the boundaries $\underline{j}_i$ and $\overline{j}_i$. For each $1 \leq i \leq |S_1|$, optimizing over all pairs of cuts $(i,j)$ and $(i',j')$ of $\mathcal{A}^\mathrm{R}$, $\underline{j}_i$ is defined as the minimum of all $j' - \Delta$ if $i' = i$ and $i'$ if $j' - j \leq \Delta$, but set to 0 if this is negative; $\overline{j}_i$, as the maximum of all $j' + \Delta$ if $i' = i$ and $i'$ if $j - j' \leq \Delta$, but limited to $|S_2|$ if larger. Note that we can compute this in linear time, since the number of cases is $O(\Delta|S_1|)$.

## Banding in progressive multiple alignment

We devise a progressive alignment scheme to solve the problem of multiple alignment in deviation $\Delta$ from a reference alignment $\mathcal{A}^{\mathrm{R}}$. The elementary operation, or a single step, of this scheme is computing an alignment $\mathcal{A}$ of two alignments $\mathcal{A}^1$ and $\mathcal{A}^2$ restricted by a set of permissible cuts $\mathcal{C}(\mathcal{A}^{\mathrm{R}}, \Delta)$. Let $\mathcal{A}^1$ and $\mathcal{A}^2$ be alignments of respective sequences $S_1, \ldots, S_{\mathrm{rows}(\mathcal{A}^1)}$ and $S_{\mathrm{rows}(\mathcal{A}^1)+1}, \ldots, S_q$. W.l.o.g. let $\mathcal{A}^{\mathrm{R}}$ be a multiple alignment of the sequences $S_1, \ldots, S_q$.

We perform a single progressive alignment step by aligning two "consensus" sequences $\hat{S}_1$ and $\hat{S}_2$ of respective alignments $\mathcal{A}^1$ and $\mathcal{A}^2$. We denote the optimal alignment of $\hat{S}_1$ and $\hat{S}_2$ by $\mathcal{A}_{\mathrm{p}}$. Then, $\mathcal{A}_{\mathrm{p}}$ *induces* a multiple alignment $\mathcal{A}$ of $\mathcal{A}^1$ and $\mathcal{A}^2$, write $\mathcal{A} := [\![\mathcal{A}_{\mathrm{p}}; \mathcal{A}^1, \mathcal{A}^2]\!]$, which is optimal, among all alignments of $\mathcal{A}^1$ and $\mathcal{A}^2$, due to the sum-of-pairs score.

We define a set of *permissible cuts* $\mathcal{C}(\mathcal{A}^{\mathrm{R}}, \Delta)$ for the alignment $\mathcal{A}_{\mathrm{p}}$ of $\hat{S}_1$ and $\hat{S}_2$, such that the set of $\mathcal{A}_{\mathrm{p}}$ where $\mathrm{cuts}(\mathcal{A}_{\mathrm{p}}) \subseteq \mathcal{C}(\mathcal{A}^{\mathrm{R}}, \Delta)$ describe exactly the alignments whose induced alignments are in $\Delta$-deviation from the reference alignment, i.e.,

$$\mathrm{cuts}(\mathcal{A}_{\mathrm{p}}) \subseteq \mathcal{C}(\mathcal{A}^{\mathrm{R}}, \Delta) \text{ iff } \mathrm{d}_{\mathcal{A}^{\mathrm{R}}}([\![\mathcal{A}_{\mathrm{p}}; \mathcal{A}^1, \mathcal{A}^2]\!]) \leq \Delta.$$

Due to the definition of the deviation as "maximum-of-pairs", we can compute this set as the intersection of pairwise cut sets $\mathcal{C}(\mathcal{A}^{\mathrm{R}}, x, y, \Delta) \subseteq \{0, \ldots, |\hat{S}_1|\} \times \{0, \ldots, |\hat{S}_2|\}$ $(1 \leq x \leq \mathrm{rows}(\mathcal{A}^1), \mathrm{rows}(\mathcal{A}^1)+1 \leq y \leq q)$ that guarantee the $\Delta$-deviation for the pairwise alignment of sequences $S_x$ and $S_y$, i.e.,

$$\mathrm{cuts}(\mathcal{A}_{\mathrm{p}}) \subseteq \mathcal{C}(\mathcal{A}^{\mathrm{R}}, x, y, \Delta) \text{ iff } \mathrm{d}_{\mathcal{A}^{\mathrm{R}}\langle x,y\rangle}([\![\mathcal{A}_{\mathrm{p}}; \mathcal{A}^1, \mathcal{A}^2]\!]\langle x, y\rangle) \leq \Delta.$$

Let $y' = y - \mathrm{rows}(\mathcal{A}^1)$ denote the index of sequence $S_y$ in $\mathcal{A}^2$. The set $\mathcal{C}(\mathcal{A}^{\mathrm{R}}, x, y, \Delta)$ from the alignment strings ${A^1}_x$, ${A^2}_{y'}$, ${A^{\mathrm{R}}}_x$, ${A^{\mathrm{R}}}_y$ is generated as follows. A cut $c_p$ of $A_p$ corresponds to an *induced cut* $c = (c_1, c_2)$ *of* $[\![\mathcal{A}_p; \mathcal{A}^1, \mathcal{A}^2]\!]\langle x, y\rangle$, which is defined by

$$c = [\![c_p; {A^1}_x, {A^2}_{y'}]\!] = (\mathrm{ctp}_{{A^1}_x}(c_1), \mathrm{ctp}_{{A^2}_{y'}}(c_2)).$$

For each cut $c$ of the alignment $\mathcal{A}^{\mathrm{R}}\langle x, y\rangle$, we generate the set of cuts $c'$ in distance $\Delta$. Then for each such cut $c'$, we generate the cuts $c_p$, where $c' = [\![c_p; {A^1}_x, {A^2}_{y'}]\!]$, and add them to the set $\mathcal{C}(\mathcal{A}^{\mathrm{R}}, x, y, \Delta)$. As in the pairwise case, this is computed in terms of boundaries $\underline{j}_i$ and $\overline{j}_i$ for "matrix rows" $1 \leq i \leq \mathrm{rows}(\mathcal{A}^1)$. Supplementary Figure 2 provides examples of cut sets $\mathcal{C}(\mathcal{A}^{\mathrm{R}}, x, y, \Delta)$.

Multiple alignments generated by progressive alignment built on this *strict* definition of permissible cuts $\mathcal{C}(\mathcal{A}^{\mathrm{R}}, \Delta)$ have at most deviation $\Delta$ from the reference alignment. However, this strategy can fail to produce an alignment, for similar reasons that alignment cannot guarantee success in constructing an optimal alignment, because of misalignments in earlier progressive steps. We emphasize that such potential inconsistencies cannot be avoided in a *progressive* alignment method that guarantees the maximum deviation $\Delta$. As long as such events are rare (like we observe), they can be tolerated for screening applications. In cases, where this behavior is unwanted, we can instead apply a variant of the algorithm that *relaxes* the constraints, thereby ignoring violations due to previous steps.

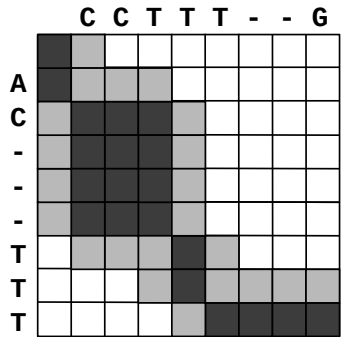
### Relaxed cut sets.

For applications that require guaranteed success, we propose a relaxation of the method that avoids inconsistencies by relaxing the distance constraints in an optimal way. By DP, we compute a relaxed cut set $\mathcal{C}_{\mathrm{relaxed}}(\mathcal{A}^{\mathrm{R}}, \Delta)$ that has a size limited by $\Delta$ and minimizes the distance to the sets $\mathcal{C}(\mathcal{A}^{\mathrm{R}}, x, y, \Delta)$. $\mathcal{C}_{\mathrm{relaxed}}(\mathcal{A}^{\mathrm{R}}, \Delta)$ is computed as set of cuts in $\Delta$- deviation from an alignment $\mathcal{A}'$ of $\hat{S}_1$ and $\hat{S}_2$. $\mathcal{A}'$ minimizes $\sum_{c \in \mathrm{cuts}(\mathcal{A}')} \mathrm{cost}(c)$, where the cost of a cut of $\mathcal{A}'$ is defined by

$$\mathrm{cost}(c) := \sum_{\substack{1 \leq x \leq \mathrm{rows}(\mathcal{A}_1),\\ \mathrm{rows}(\mathcal{A}_1) < y \leq \mathrm{rows}(\mathcal{A})}} \min_{c' \in \mathcal{C}(\mathcal{A}^{\mathrm{R}}, x, y, 0)} \|c - c'\|_1.$$

The alignment $\mathcal{A}'$ is obtained by traceback from the DP matrix C evaluating $\mathrm{C}(0,0) = \mathrm{cost}(0,0), \mathrm{C}(i,0) = \mathrm{cost}(i,0) + \mathrm{C}(i-1,0), \mathrm{C}(0,j) = \mathrm{cost}(0,j) + \mathrm{C}(0,j-1)$, and

$$\mathrm{C}(i,j) = \mathrm{cost}(i,j) + \min\{\mathrm{C}(i-1,j-1), \mathrm{C}(i-1,j), \mathrm{C}(i,j-1)\}$$




**Supplementary Figure 2.** Cut sets for banding in the case of multiple progressive alignment. Cut sets for sequences $x$ and $y$ of $\mathcal{A}^{\mathrm{R}}$. Intersecting such sets for all pairs of sequences $x$ and $y - \mathrm{rows}(\mathcal{A}^1)$ from the respective input alignments $\mathcal{A}^1$ and $\mathcal{A}^2$ yields a cut set for one step of the progressive multiple alignment algorithm. Cut sets for $\Delta = 0$ (dark gray), $\Delta = 1$ (dark+light gray). There, the rows $x$ and $y$ of $\mathcal{A}^{\mathrm{R}}$ are ${A^{\mathrm{R}}}_x$=`AC--TTT-` and ${A^{\mathrm{R}}}_y$=`-CCTT-TG`; the corresponding rows of the input alignments are ${A^1}_x$=`AC---TTT` and ${A^2}_{(y-\mathrm{rows}(\mathcal{A}^1))}$=`CCTTT--G`.

for $1 \leq i \leq |\mathcal{A}_1|, 1 \leq j \leq |\mathcal{A}_2|$. Finally, we set $\mathcal{C}_{\mathrm{relaxed}}(\mathcal{A}^{\mathrm{R}}, \Delta) := \mathcal{C}(\mathcal{A}', \Delta)$. Clearly, a heuristic based on this relaxation will not *guarantee* $\Delta$-deviation from $\mathcal{A}^{\mathrm{R}}$. However, by construction, it will favor low deviation and limit the computational cost and deviation in each progressive alignment step by $\Delta$.

## Simultaneous Alignment and Folding in LocARNA in Limited Deviation of a Reference Alignment

We review details of the LocARNA score and algorithm and show the lifting of the novel banding technique to LocARNA's simultaneous alignment and folding DP algorithm.

### Structure

A *base pair* is a pair $a = (i, j) \in \mathbb{N}^2$. We call $i =: a^\ell$ its *left end* and $j =: a^{\mathrm{r}}$ its *right end*. An *(RNA) structure $P$ for length $n$* is a set of *base pairs* $(i, j)$, $1 \leq i < j \leq n$, where no two different base pairs share a common end, i.e., for all $(i, j), (i', j') \in P : i = i' \implies j = j'$ and $j \neq i'$. We call $P$ *crossing* iff there

exist two base pairs $(i,j),(i',j') \in P$ such that $i < i' < j < j'$. Otherwise, $P$ is called *non-crossing* or *nested*. We discuss only non-crossing structure.

### Similarity and Simultaneous Alignment and Folding

Following (Hofacker et al. 2004) and (Will et al. 2007), we define a *sequence-structure similarity score for an alignment $|\mathcal{A}|$ and an RNA structure $P$ for length $|\mathcal{A}|$*. In the case of a pairwise alignment $\mathcal{A}$, this similarity score is of the form

$$\begin{aligned}
&\text{simscore}(S_1, S_2, A_1, A_2, P) = \\
&\sum_{\substack{(i,j)\in P \\ A_1[i]\neq -, A_1[j]\neq - \\ A_2[i]\neq -, A_2[j]\neq -}} \tau^{S_1,S_2}(\text{ctp}_{A_1}(i), \text{ctp}_{A_1}(j), \text{ctp}_{A_2}(i), \text{ctp}_{A_2}(j)) \qquad \text{(structural similarity)} \\
&+ \sum_{\substack{1\leq i\leq n, \\ i \text{ unpaired in } P, \\ A_1[i]\neq -, A_2[i]\neq -}} \sigma^{S_1,S_2}\big(\text{ctp}_{A_1}(i), \text{ctp}_{A_2}(i)\big) + \sum_{k>0} \gamma(k) N_k^{A_1,A_2} \qquad \text{(sequence similarity,affine gap cost)}
\end{aligned}$$

where $\sigma^{S_1,S_2}$ is a sequence similarity and $\tau^{S_1,S_2}$ is a structural similarity function, $\gamma(k) = \gamma_o + k\gamma_e$, and $N_k^{A_1,A_2}$ is the number of maximal subsequences of $k$ gaps in $A_1$ and $A_2$. For the definition of $\sigma^{S_1,S_2}$ and $\tau^{S_1,S_2}$ confer (Will et al. 2007). We generalize this to the $q$-way case by sum-of-pairs, i.e., we define for $q \geq 2$,

$$\text{simscore}(\mathcal{A}, P) = \sum_{1\leq x<y\leq q} \text{simscore}(S_x, S_y, A_x, A_y, P).$$

Given sequences $S_1, \ldots, S_q$, the problem of *simultaneous alignment and folding* (SA&F) asks for the alignment $\mathcal{A}$ of $S_1, \ldots, S_q$ and structure $P$ for length $|\mathcal{A}|$ that maximize $\text{simscore}(\mathcal{A}, P)$.

An efficient algorithm to solve this specific problem for the pairwise case ($q = 2$) was introduced in (Hofacker et al. 2004) and significantly improved in (Will et al. 2007). Whereas LocARNA (Will et al. 2007) (and therefore our implementation) supports affine gap cost, we keep presentation simple, by

describing only linear gap cost, where each gap costs $\gamma$ (i.e. $\gamma(k) = k\gamma$).

The pairwise LocARNA-algorithm has parameters $(n, m, \sigma, \tau, \gamma)$, where $n$ and $m$ are sequence lengths, $\sigma$ denotes sequence similarity, $\tau$ structural similarity, and $\gamma$ gap cost. We assume for our description w.l.o.g that the algorithm aligns two sequences $S_1$ and $S_2$ of respective lengths $n$ and $m$. The algorithm evaluates, for $M_{i\,i-1;k\,k-1} = 0$, the recursion

$$M_{i\,j;k\,l} = \max \begin{cases} M_{i\,j-1;k\,l-1} + \sigma(j,l); M_{i\,j-1;k\,l} + \gamma; M_{i\,j;k\,l-1} + \gamma \\ \max_{j'l'} M_{i\,j'-1;k\,l'-1} + D_{j'\,j;l'\,l} \end{cases} \tag{1}$$

$$D_{i\,j;k\,l} = M_{i+1\,j+1;k-1\,l-1} + \tau(i,j;k,l)$$

for $1 \leq i < j \leq n$ and $1 \leq k < l \leq m$. The matrix entries $M_{i\,j;k\,l}$ are defined as the maximal similarity score of an alignment of subsequences $S_1[i \ldots j]$ and $S_2[k \ldots l]$. $D_{i\,j;k\,l}$ is the maximal similarity score of such an alignment where base pairs $(i, j)$ and $(k, l)$ are matched.

In this way, the pairwise LocARNA-algorithm solves the alignment problem for sequence $S_1$ and $S_2$ when parametrized by $(|S_1|, |S_2|, \sigma^{S_1,S_2}, \tau^{S_1,S_2}, \gamma)$. The maximal sequence-structure similarity is obtained as $M_{1\,1;n\,m}$ and the actual alignment is obtained by trace back from the DP matrices.

The same algorithm can be employed in a progressive alignment scheme to compute multiple alignments (Hofacker et al. 2004; Will et al. 2007). There the algorithm computes an alignment $\mathcal{A}$ of two alignments $\mathcal{A}^1$ and $\mathcal{A}^2$. For this reason the algorithm is parametrized by $(|\mathcal{A}^1|, |\mathcal{A}^2|, \sigma^{\mathcal{A}^1,\mathcal{A}^2}, \tau^{\mathcal{A}^1,\mathcal{A}^2}, \gamma)$. Details on how to construct $\sigma^{\mathcal{A}^1,\mathcal{A}^2}$ and $\tau^{\mathcal{A}^1,\mathcal{A}^2}$ according to the sum-of-pairs idea are given in (Hofacker et al. 2004) and (Will et al. 2007).

### Extending the banding technique to simultaneous alignment and folding

Adapting the banding technique to the LocARNA algorithm, we change the semantics of the matrix entries $M_{i\,j;k\,l}$ and $D_{i\,j;k\,l}$ such that they contain the maximal score only over subalignments of alignments

with limited deviation $\Delta$ from $\mathcal{A}^{\mathrm{R}}$. Due to this definition, we need to compute entries $M_{i\,j;k\,l}$ only if the optimal alignment can be derived from an alignment of subsequences $S_1[i]\dots S_1[j]$ and $S_2[k]\dots S_2[l]$, i.e., only if the cuts $(i-1,k-1)$ and $(j,l)$ are in $\mathcal{C}_2(\mathcal{A}^{\mathrm{R}},\Delta)$. $D_{i\,j;k\,l}$ needs to be computed only if $i$ can be matched to $k$ and $j$ can be matched to $l$, i.e., $(i,k),(i-1,k-1),(j,l)$, and $(j-1,l-1)$ are in $\mathcal{C}_2(\mathcal{A}^{\mathrm{R}},\Delta)$. Furthermore, the computation of $M_{i\,j;k\,l}$ is restricted to indices $i$ and $k$, where $i-1$ and $k-1$ can match (with the exception of $(i,k)=(1,1)$).

## Sensitivity for annnotations in *D. melanogaster* through successive REAPR steps

We further examined the annotation sensitivity of the REAPR pipeline step-by-step. As each step removes some genomic regions from further consideration as ncRNAs, sensitivity progressively decreases through the pipeline. Table 1 shows, for each step, the number of annotations overlapping the genomic regions still in consideration after the step is completed and the associated percentage loss in sensitivity. Both REAPR with LocARNA realignment at $\Delta = 20$ and the control pipeline without realignment, differing only in the last step, are shown. A substantial fraction of the overall loss in sensitivity occurs during the first step, i.e. slicing the WGA into windows. This loss results from REAPR being an RNAz-based pipeline. In this step, windows that do not have more than one sequence meeting certain gap and base composition criteria are removed. In general, an annotated ncRNA is not covered by a window only if the WGA does not align it well to genomic regions of one or more other organisms. Thus, it is fair to assume that most of these annotated RNAs are lost because they are misaligned by the WGA at a non-local scale. With the current methods, such RNAs cannot be found. In the sets `miscRNA` and `ncRNA`, there is an even larger drop in the annotations that overlap with stable loci. This confirms that many of these annotated RNAs do not form stable structures or form only a few local structures. The final step from annotations

in stable loci to those overlapping predictions is the only step that is affected by realignment in REAPR. Even with realignment, there is still a significant loss for the weakly structural sets `miscRNA` (32%) and `ncRNA` (54%). However, for the remaining sets `Rfam`, `miRNA`, and `tRNA` we observe high sensitivities between 85% and 94%. For all sets, the loss is significantly reduced.

## Co-predicted ncRNAs recapitulate the *Drosophila* phylogeny

We estimated a phylogeny of *Drosophila* based on co-predicted ncRNAs across fly species (Supplementary Fig. 4). For every pair of genomes, we counted the number of times the two genomes were represented in the same high-confidence prediction. Note that a locus alignment regularly contains sequences from only a subset of the twelve fly genomes, since often the sequences of some genomes are either deleted or weakly aligned in the locus region; consequently, such genomes are removed in the first pipeline stage (Methods). Naturally, from a given locus, we can predict ncRNA only for the genomes present in the locus. As background, we count how often the two genomes are part of the same stable locus alignment. The negative log-odds of these frequencies are distances betweeen the genomes. The effect of normalizing with these particular background frequencies is that the distances are based only on conservation of structure and not influenced by other factors, in particular the length of genomes, the number of stable RNAs in the genomes, or the co-occurrence of genomes in the same stable locus or general locus, which are more obviously related to evolutionary similarity.

Although the phylogeny is constructed straightforwardly from our predictions, we observe remarkable agreement with the FlyBase *Drosophila* phylogeny (Clark et al. 2007). Our only topological difference is the association of *D.willistoni* with the subgenus *Drosophila* instead of the larger *Sophophora*. In the light of recent discussion by (Bhutkar et al. 2008) on the position of *D.willistoni*, this may well reflect a true relationship on the level of structural RNA.

Since previous computational screens for ncRNAs aimed at predicting candidates in a reference organ-

**Table 1.** Sensitivity to *D. melanogaster* annotations through successive steps of RNAz-based pipelines. Entries indicate the number of annotations overlapping the genomic regions that are still in consideration after each step. Arrows indicate the percentage loss from step to step. The first two steps are characteristic of RNAz-based pipelines regardless of realignment, whereas the third step is specific to the REAPR pipeline with realignment. In parenthesis, we provide annotation numbers and losses specific to the control pipeline without realignment.

| | **WGA** | | **windows** | | **stable loci** | | **REAPR Δ20** (control) |
|---|---|---|---|---|---|---|---|
| **Rfam** | 664 | -31% | 460 | -20% | 369 | **-15%** | **315** |
| | | | | | | (-28%) | 267 |
| **miRNA** | 191 | -17% | 159 | -15% | 135 | **-6%** | **127** |
| | | | | | | (-10%) | 121 |
| **miscRNA** | 390 | -32% | 266 | -45% | 145 | **-32%** | 99 |
| | | | | | | (-45%) | 80 |
| **tRNA** | 292 | -11% | 259 | -16% | 218 | **-14%** | 187 |
| | | | | | | (-33%) | 147 |
| **ncRNA** | 198 | -11% | 177 | -40% | 106 | **-54%** | 49 |
| | | | | | | (-70%) | 32 |

ism (e.g., *D. melanogaster* or *H. sapiens*), they evaluated only regions of the alignment that are represented in the reference genome. In contrast, we screened without this bias, allowing us to study the fly phylogeny based on predicted structural RNAs.

## Unsurprisingly Weak Correlation of RNAz Scores and Expression Levels

We plotted the expression level and RNAz score of high confidence predictions that have an expression level of at least 50 reads (Supplementary Fig. 5). The relationship is visually weak, and the Spearman rank correlation is only 0.19. While a strong correlation would have been interesting, we find this weak correlation to be unsurprising. We believe that the primary reason that could have driven a strong correlation would have been if each of these measures was positively correlated to the separate notion of a locus's functional significance. After all, a locus can function as an ncRNA only if it is expressed. However, while these relationships may hold to some degree on a qualitative level, they may not hold very strongly on a quantitative level. For example, an ncRNA with an important regulatory role could have a higher or lower expression level or a higher or lower RNAz score than other loci. As another example, a housekeeping gene that has a low RNAz score and is in fact not an ncRNA could be constitutively expressed at high levels in the cell.

This weak correlation indicates that expression alone does not serve as a good proxy for benchmarking ncRNA prediction nor as sufficient validation of function. Using expression as a benchmark can identify which ncRNA predictors are enriched for high expression, but not necessarily which ones better capture the total set of ncRNAs in terms of measures like sensitivity or FDR.

## REAPR Predictions in Long ncRNAs

Table 2 lists the long non-coding RNAs from the FlyBase annotation set `ncRNA` that overlap with high-confidence predictions of REAPR in *D. melanogaster*. We show single annotated ranges of long non-coding RNAs together with the names of overlapping (merged) high-confidence predictions (sorted by ncRNA names). The latter names are composed of the WGA block coordinates and the locus ID of all merged loci. The names correspond to the online supplement table of REAPR ($\Delta = 20$) high confidence predictions in *D. melanogaster* (http://reapr.csail.mit.edu/Fly/reapr20_dm2.bed). The column labeled "#" shows the number of REAPR predictions in the long ncRNA. For example, the table reports roX1-RA to contain the two predictions X 3665964 3708413.92 and X 3665964 3708413.185 from Figure 6 of the main text. Several of the long ncRNAs, e.g. CR43334-RA (7 overlaps) and CR43314-RD (8 overlaps), have remarkable overlap with our predictions, indicating a high degree of structure. All of the reported RNAs from the annotation set are bona fide long ncRNAs that are longer than 200nt and frequently even several thousand nucleotides long. (We have removed two shorter annotations snmRNA:158-RA and snmRNA:254 with overlaps.) The reported ncRNAs are thus much longer than typical REAPR predictions.

**Table 2.** REAPR predictions of long ncRNAs from FlyBase annotation set ncRNA.

| Name (FB id) | # | REAPR High-confidence Predictions |
|---|---|---|
| 7SLRNA:CR32864-RA (FBtr0081624) | 1 | 3R_2612115_2651106.89 |
| 7SLRNA:CR42652-RA (FBtr0302398) | 1 | 3R_2612115_2651106.195 |
| CR18854-RB (FBtr0302441) | 1 | 2L_9782784_9790968.35 |
| CR18854-RC (FBtr0079919) | 1 | 2L_9782784_9790968.35 |
| CR31044-RA (FBtr0085391) | 3 | 3R_24984168_25046151.102$3R_24984168_25046151.225<br>3R_24984168_25046151.227<br>3R_24984168_25046151.106$3R_24984168_25046151.228 |
| CR31044-RC (FBtr0303427) | 3 | 3R_24984168_25046151.102$3R_24984168_25046151.225<br>3R_24984168_25046151.227<br>3R_24984168_25046151.106$3R_24984168_25046151.228 |
| CR31846-RB (FBtr0303312) | 2 | 2L_13387774_13508614.347$2L_13387774_13508614.152<br>2L_13387774_13508614.153$2L_13387774_13508614.348 |
| CR32028-RA (FBtr0076660) | 1 | 3L_8411516_8446956.296 |

| | | |
|---|---|---|
| CR32205-RA (FBtr0075000) | 2 | 3L_19372126_19491209.268<br>3L_19372126_19491209.48 |
| CR32218-RA (FBtr0304110) | 1 | 3L_19816292_19853995.16 |
| CR32658-RA (FBtr0073626) | 1 | X_11923641_11960675.4$X_11923641_11960675.63 |
| CR32658-RB (FBtr0073627) | 1 | X_11923641_11960675.4$X_11923641_11960675.63 |
| CR32658-RC (FBtr0073628) | 1 | X_11923641_11960675.4$X_11923641_11960675.63 |
| CR32690-RB (FBtr0303456) | 1 | X_10034733_10067079.27 |
| CR32730-RA (FBtr0071016) | 3 | X_6954857_7000161.26$X_6954857_7000161.61<br>X_6954857_7000161.62$X_6954857_7000161.27<br>X_6954857_7000161.30 |
| CR33963-RA (FBtr0100004) | 2 | X_12470049_12532543.49 X_12470049_12532543.53 |
| CR42646-RA (FBtr0302350) | 2 | 2R_847136_982050.115 2R_847136_982050.32 |
| CR42745-RA (FBtr0303214) | 2 | 3R_22009768_22077220.483<br>3R_22009768_22077220.483 |
| CR42767-RA (FBtr0303397) | 1 | X_9400682_9414674.19$X_9400682_9414674.3 |
| CR42839-RA (FBtr0303861) | 1 | 3R_11819947_11843432.41 |
| CR42858-RA (FBtr0304027) | 1 | 3R_5648820_5695010.123 |
| CR42859-RA (FBtr0304029) | 1 | 2L_2239473_2263557.8 |
| CR42862-RA (FBtr0304059) | 1 | 3L_207751_217833.24$3L_207751_217833.7 |
| CR43159-RA (FBtr0305611) | 1 | 3L_15267930_15485878.721$3L_15267930_15485878.348 |
| CR43238-RA (FBtr0306291) | 1 | 3R_26407990_26693611.313$3R_26407990_26693611.893 |
| CR43264-RA (FBtr0306548) | 1 | X_5660354_5679586.91$X_5660354_5679586.170 |
| CR43270-RA (FBtr0306851) | 1 | 3L_22540117_22613045.229$3L_22540117_22613045.110 |
| CR43280-RA (FBtr0306721) | 1 | 3L_18363299_18565897.554 |
| CR43283-RB (FBtr0306729) | 3 | 3R_7066873_7158976.162$3R_7066873_7158976.22<br>3R_7066873_7158976.163$3R_7066873_7158976.23<br>3R_7066873_7158976.25 |
| CR43299-RA (FBtr0306829) | 1 | X_15439105_15532352.240 |
| CR43299-RB (FBtr0306830) | 2 | X_15439105_15532352.98<br>X_15439105_15532352.240 |
| CR43299-RC (FBtr0306831) | 2 | X_15439105_15532352.240<br>X_15439105_15532352.98 |
| CR43301-RA (FBtr0306833) | 3 | 3R_5161283_5189245.5 3R_5161283_5189245.6<br>3R_5161283_5189245.58$3R_5161283_5189245.8 |
| CR43301-RB (FBtr0306834) | 3 | 3R_5161283_5189245.5 3R_5161283_5189245.6<br>3R_5161283_5189245.58$3R_5161283_5189245.8 |

| | | |
|---|---|---|
| CR43305-RA (FBtr0306839) | 1 | 2R_18491053_18500732.174$2R_18491053_18500732.44 |
| CR43314-RA (FBtr0306897) | 6 | 2L_11811119_11982002.263$2L_11811119_11982002.581<br>2L_11811119_11982002.264<br>2L_11811119_11982002.583<br>2L_11811119_11982002.588<br>2L_11811119_11982002.291$2L_11811119_11982002.622<br>2L_11811119_11982002.292 |
| CR43314-RC (FBtr0306898) | 7 | 2L_11811119_11982002.263$2L_11811119_11982002.581<br>2L_11811119_11982002.264<br>2L_11811119_11982002.583<br>2L_11811119_11982002.588<br>2L_11811119_11982002.596<br>2L_11811119_11982002.291$2L_11811119_11982002.622<br>2L_11811119_11982002.292 |
| CR43314-RD (FBtr0306899) | 8 | 2L_11811119_11982002.257$2L_11811119_11982002.575<br>2L_11811119_11982002.263$2L_11811119_11982002.581<br>2L_11811119_11982002.264<br>2L_11811119_11982002.583<br>2L_11811119_11982002.588<br>2L_11811119_11982002.596<br>2L_11811119_11982002.291$2L_11811119_11982002.622<br>2L_11811119_11982002.292 |
| CR43334-RA (FBtr0306918) | 7 | 3L_588730_632529.148$3L_588730_632529.58<br>3L_588730_632529.149$3L_588730_632529.59<br>3L_588730_632529.154 3L_588730_632529.155<br>3L_588730_632529.71 3L_588730_632529.80<br>3L_588730_632529.82 |
| CR43344-RA (FBtr0306973) | 3 | 2L_18445565_18479823.25$2L_18445565_18479823.63<br>2L_18445565_18479823.27<br>2L_18445565_18479823.28$2L_18445565_18479823.65 |
| CR43372-RA (FBtr0307362) | 1 | 3R_17341791_17399115.142$3R_17341791_17399115.26 |
| Hsromega-RA (FBtr0084057) | 1 | 3R_17087193_17178558.50 |
| Hsromega-RB (FBtr0084058) | 1 | 3R_17087193_17178558.50 |
| Hsromega-RC (FBtr0084059) | 1 | 3R_17087193_17178558.50 |
| RNaseMRP:RNA-RA (FBtr0091662) | 1 | 3R_19531299_19557312.83$3R_19531299_19557312.84$-<br>3R_19531299_19557312.85 |
| RNaseP:RNA-RA (FBtr0085775) | 2 | 3R_26952344_27057668.350<br>3R_26952344_27057668.351 |
| Yu-RA (FBtr0302365) | 1 | 2R_19983214_20031920.130 |
| bxd-RA (FBtr0083342) | 1 | 3R_12509464_12623559.369 |
| bxd-RB (FBtr0083343) | 1 | 3R_12509464_12623559.369 |
| bxd-RC (FBtr0083344) | 1 | 3R_12509464_12623559.369 |
| iab-4-RA (FBtr0083362) | 1 | 3R_12623559_12822774.463$3R_12623559_12822774.93 |
| pncr004:X-RA (FBtr0091949) | 1 | X_18977524_19011215.134 |

| | | |
|---|---|---|
| pncr009:3L-RA (FBtr0091951) | 2 | 3L_19372126_19491209.146<br>3L_19372126_19491209.147 |
| pncr017:3R-RA (FBtr0091956) | 1 | 3R_27588420_27598275.32 |
| roX1-RA (FBtr0070634) | 2 | X_3665964_3708413.185 X_3665964_3708413.92 |
| roX1-RB (FBtr0070635) | 2 | X_3665964_3708413.185 X_3665964_3708413.92 |
| roX2-RA (FBtr0073514) | 2 | X_11419270_11426141.5$X_11419270_11426141.8<br>X_11419270_11426141.9 |
| roX2-RB (FBtr0073515) | 2 | X_11419270_11426141.5$X_11419270_11426141.8<br>X_11419270_11426141.9 |
| snRNA:7SK-RA (FBtr0091929) | 1 | 3R_3298203_3312554.8$3R_3298203_3312554.41 |
| sphinx-RA (FBtr0111044) | 1 | 4_963017_1010501.96$4_963017_1010501.31 |
| sphinx-RB (FBtr0111045) | 1 | 4_963017_1010501.96$4_963017_1010501.31 |
| sphinx-RC (FBtr0111046) | 1 | 4_963017_1010501.96$4_963017_1010501.31 |

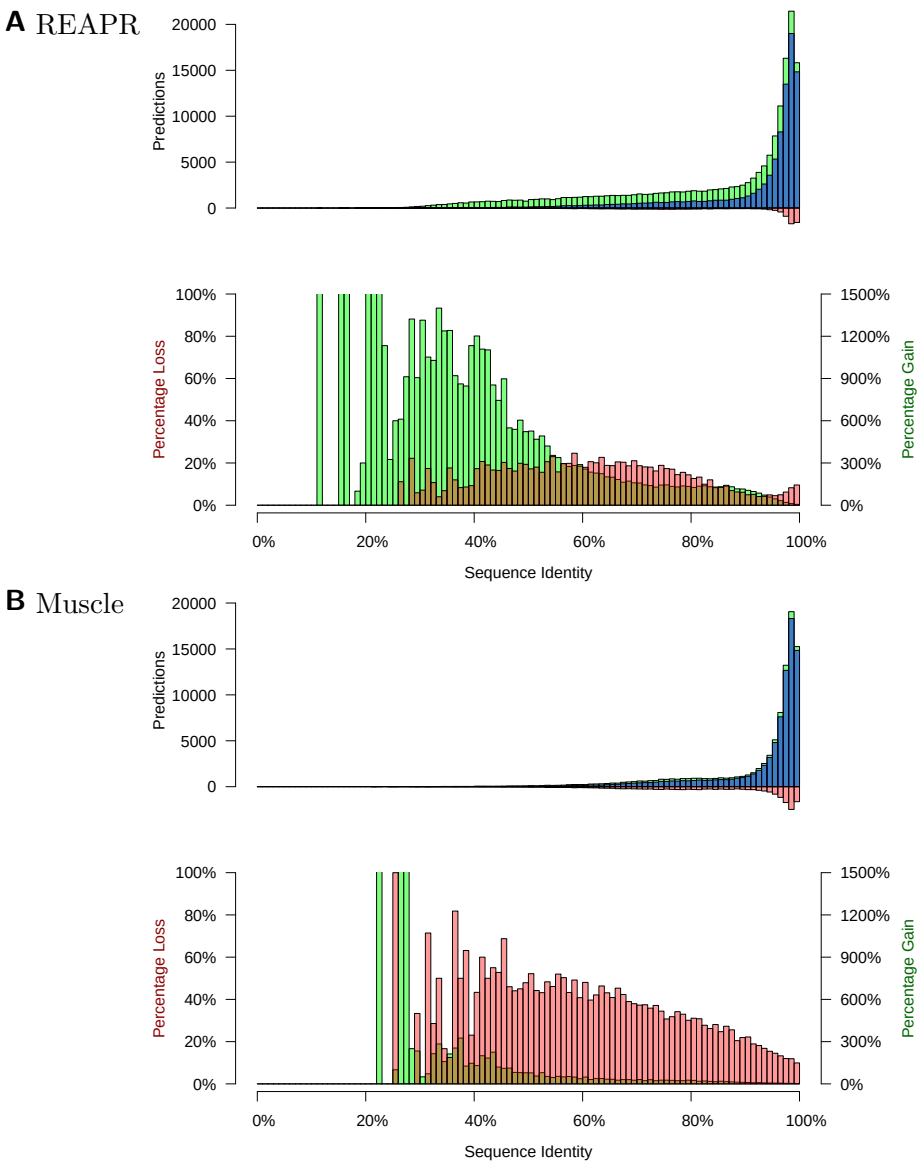
## References

Bhutkar, A., Schaeffer, S. W., Russo, S. M., Xu, M., Smith, T. F., and Gelbart, W. M., 2008. Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes. *Genetics*, **179**(3):1657–80.

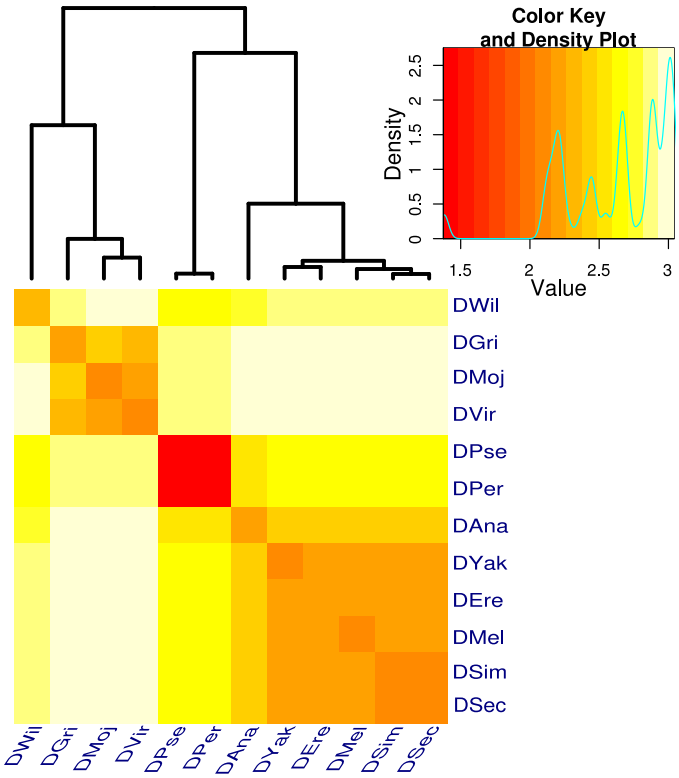
Clark, A. G., Eisen, M. B., Smith, D. E., and MacCallum, I., 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**(7167):203–18.

Hofacker, I. L., Bernhart, S. H., and Stadler, P. F., 2004. Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**(14):2222–7.

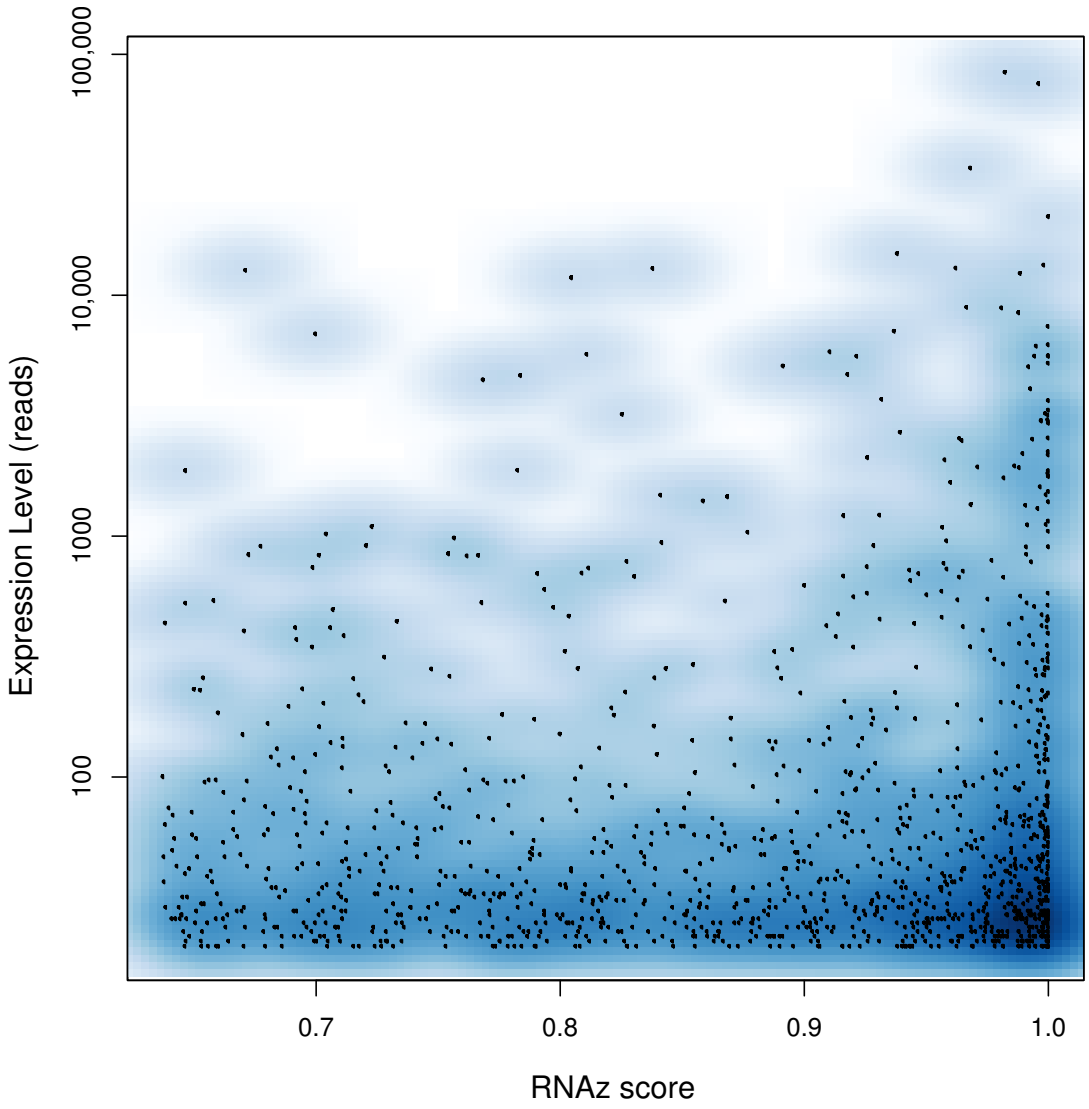
Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R., 2007. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology*, **3**(4):e65.




**Supplementary Figure 3.** A normalized representation of Figures 3A-B, which is also reproduced here for comparison. The number of novel (green) or lost (red) predictions at a given sequence identity was divided by the total number of predictions from the original WGA (blue + red) at that identity. Note the difference scales used for percentage loss and gain. For both REAPR and Muscle realignment, the percentage loss is more uniformly distributed across sequence identities than the absolute count of lost predictions. The percentage gain shows REAPR's greater ability to identify low identities than Muscle. (A) REAPR, $\Delta = 20$. (B) Muscle.




**Supplementary Figure 4.** Distances of *Drosophila* genomes and phylogeny due to co-predictions of ncRNA across all genomes (at $\Delta = 20$). Heat map of the log odds (co-prediction over background frequency, see text) and dendrogram by `heatmap.2` of the R-package `gplots`.




**Supplementary Figure 5.** Density scatterplot of the RNAz score vs. the expression level in high-confidence predictions from REAPR with $\Delta = 20$.