Structure-based Whole Genome Realignment Reveals Many Novel Non-coding RNAs

Sebastian Will[1⋆] , Michael Yu[1⋆] and Bonnie Berger[1⋆⋆]

Math. Department and CSAIL, MIT, 77 Massachusetts Ave, Cambridge, MA, USA

Recent genome-wide computational screens that search for conservation of RNA secondary structure in whole genome alignments (WGAs) have predicted thousands of structural non-coding RNAs (ncRNAs). The sensitivity of such approaches, however, is limited due to their reliance on sequence-based whole-genome aligners, which regularly misalign structural ncRNAs. This suggests that many more structural ncRNAs may remain undetected. Structure-based alignment, which could increase the sensitivity, has been prohibitive for genome-wide screens due to its extreme computational costs. Breaking this barrier, we present the pipeline REAPR (RE-Alignment for *de novo* Prediction of structural ncRNA) that realigns whole genomes based on RNA sequence *and* structure and then evaluates the realignments for potential structural ncRNAs with a ncRNA predictor such as RNAz 2.0. For efficiency of the pipeline, we develop a novel banding realignment algorithm for the RNA multiple alignment tool LocARNA. This allows us to perform very fast structure-based realignment within limited deviation of the original multiple alignment from the WGA. We apply REAPR to the complete twelve Drosophila WGAs to predict ncRNAs across all these Drosophila species. Compared to direct prediction from the original WGA at the same False Discovery Rate (FDR), we predict twice as many high-confidence ncRNA candidates in *D.melanogaster* while less than doubling the run-time. As a novelty in ncRNA prediction, we control the FDR, going beyond the usual *a posteriori* FDR estimation. Applying the sequence-based alignment tool Muscle for realignment, we demonstrate that structure-based methods are necessary for effective prediction of originally misaligned ncRNAs. Comparing to recent screens of Drosophila and ENCODE we show that REAPR outperforms the widely-used *de novo* predictors RNAz, EvoFold, and CMfinder. Finally, we reveal, with high confidence, a putative structural motif in the long ncRNA roX1 of *D.melanogaster*, known to regulate X chromosome dosage compensation in male flies. Interestingly, we recapitulate the Drosophila phylogeny, based on co-predicted ncRNAs across all fly genomes.

⋆ Joint first authors

⋆⋆ Corresponding author, E-mail: bab@mit.edu