Supplemental Table S4. Matches for the miR-125 Seed Heptamer and Its Shuffled Derivatives

| Seed Sequence | Matches in Human 3' UTRs |
|---|---|
| **CCCUGAG** | 663 |
| Ave. of 1000 random shuffles | 205 |
| **CGGACCU** | 83 |
| **CGCGUAC** | 6 |
| **CAGUGCC** | 708 |

The miR-125 heptamer is GC rich and therefore more prone to the vertebrate oligonucleotide-composition artifact of random shuffles than is the typical heptamer. On the whole, miRNA heptamer seeds have ~1.4 times as many hits to vertebrate UTR regions than do their randomly shuffled cohorts (Lewis et al., 2003). In the case of the miRanda study, when analyzing human sequences without considering conservation in other mammals, a total of 2,538,431 target sites were obtained for the real miRNAs, and an average of 2,033,701 were obtained for the shuffled cohorts of these miRNAs (John et al., 2004). The difference of ~500,000 predicted target sites, which corresponds to an average of more than 20 target sites per protein-coding gene, is difficult to rationalize as being solely the contribution of authentic targets. To the extent that the ~500,000 sites do not represent the contribution of authentic targets, they represent the contribution of false-positive predictions that were not accounted for by the shuffled cohorts, presumably because of the oligonucleotide-composition artifact. In other words, the matches to the shuffled cohorts underestimated the number of false-positive predictions in one genome by nearly 20%. Although this 20% underestimate might not seem large, it can have a large effect when propagated throughout the analysis, particularly when requiring multiple target sites in the same transcript (Supplemental Figure S1). The effect of artifactual signal above noise is compounded when requiring multiple sites in the same UTR, and depending on the algorithm used, the artifactual signal above noise can be increased further when considering higher scoring sites and/or conserved sites.