An alternative class of targets for microRNAs containing CG dinucleotide

Zheng Yan^{1,2*}, Bin Zhang^{2,3*}, Haiyang Hu^{2*}, Gangcai Xie, Song Guo, Philipp Khaitovich^{2§}, Yi-Ping Phoebe Chen^{1§}

¹Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Victoria, Australia

²Partner Institute for Computational Biology, 320 Yue Yang Road, Shanghai, China

³Graduate School of Chinese Academy of Sciences, 19 Yuquan Road, Beijing, China

*These authors contributed equally to this work

§Corresponding author

Email addresses:

PK: khaitovich@eva.mpg.de

YPC: phoebe.chen@latrobe.edu.au

24 **Abstract**

25 **Background**

26 MicroRNAs are endogenous ~23nt RNAs which regulate mRNA targets mainly
27 through perfect pairing with their seed region (positions 2-7). Several instances of
28 bulge UTR sequence can also be recognized by miRNA as their target. But such non-
29 Watson-Crick base pairings are incompletely understood.

30 **Results**

31 We found a group of miRNAs which had very few conservative targets while
32 potentially having a subclass of bulge message RNA targets. Compared with the
33 canonical target, these bulge targets had a lower negative correlation with the miRNA
34 expression, and either were downregulated in the miRNA overexpression experiment
35 or upregulated in the miRNA knock-down experiment.

36 **Conclusions**

37 We proved that the bulge target exists widely in certain groups of miRNAs and such
38 non-canonical targets can be recognized by miRNA. Incorporating these bulge
39 targets, combined with evolutionary conservation, will reduce the false-positive rate
40 of microRNA computational target prediction.

41 **Background**

42 MicroRNAs(miRNAs) are ~23 nucleotide RNAs that regulate eukaryotic gene
43 expression post-transcriptally [1]. miRNAs use base-pairing to guide RNA-induced
44 silencing complexes (RISCs) to specific message RNAs with fully or partly
45 complementary sequences, primarily in the 3' untranslated region [2]. The best
46 characterized features determining animal miRNA-target recognition are six-
47 nucleotide (nt) long seed sites, which perfectly complement the 5' end of the miRNA
48 (positions 2-7) [3]. This Watson-Crick seed pairing rule is sufficient on its own for

49 predicting conserved targets above the noise of false-positive predictions in most
50 miRNAs [4].

51 Most of the miRNA-target prediction algorithms rely heavily on seed rules and
52 evolutionary conservation [5, 6]. However, such strategies suffer from missing the
53 noncanonical target sites [7]. Several biological studies have functionally validated the
54 existence of imperfect binding sites [8–10].

55 Recently, Ago HITS-CLIP was used to precisely map the miRNA-binding sites in
56 both *Caenorhabditis elegans* [11] and mouse brains [7]. However, about one-quarter
57 of the total binding sites did not follow the classical seed rules in mouse brains [7].

58 Further analysis revealed that the miR-124, one of the most abundant miRNAs in Ago
59 complex in mouse brains, has plenty of noncanonical bulge sites. More recently, an
60 improved CLIP-seq method, CLASH (cross linking, ligation and sequencing of
61 hybrids), revealed around 60% of the seed interactions are noncanonical, containing
62 bulged or mismatched nucleotides [12].

63 Although these studies strongly suggest the existence of bulge sites, the general
64 features of their interactions with miRNAs are largely unknown, partly due to the
65 difficulty in determining how frequently such atypical sites are used in vivo and what
66 are the general rules to predict them.

67 Here, we analyze a group of highly conserved miRNAs in vertebrate, but with
68 relative fewer conservative target using the seed rule. Meanwhile, these miRNAs all
69 have a common feature, this being that their seed region contains CG dinucleotide
70 (hereafter refer as CG dimer). We found these potential miRNA regulatory sites have
71 a nucleotide bulge compared with a fully complementary sequence. This expands our
72 insight into miRNA-target interaction.

Results

MicroRNA containing CG dimer has fewer cononical targets

Evolutionary conservation has been widely used to identify miRNA-binding sites together with the seed rule. We searched for the orthologs of all the miRNAs annotated by miRbase (miRbase version 17) [13] using their mature sequence in the genomes of 23 species (Supplement Table 1). 1426 annotated miRNAs were divided into three categories, vertebrate, mammal and primate conservative miRNA. TargetScan [14] was used to look for the vertebrate conservative miRNAs' cononical targets. According to the target site conservative value (Table 1), a class of miRNAs containing CG dimer in their seed region have both very few cononical targets (t test, $p < 0.01$) and much fewer conservative target sites than the rest of the miRNAs (t test, $p < 0.01$). For the mammal and primate conservative miRNAs, miRNAs with CG dimer in their seed region also have much fewer conservative target sites (Supplementary Table 2, 3) (t test, $p < 0.01$).

Identification of bulge sites that pair to microRNA containing CG dimer

To uncover the possible bulge site, we allow one nucleotide insertion in every position in the seed region (Figure 1) for all the vertebrate conservative miRNAs. Using these artificial seed sequences, we find that only the bulge site inserted between CG dimer can increase the target number and conservation of the target sites (Figure 2). In contrast, the random bulge at the target binding site did not increase the conservation rate.

Transcriptome-wide evidence for miRNA repression through bulge target site

We used human age series mRNA and miRNA expression data [15] to quantify the transcriptome correlation between CG dimer miRNAs and their bulge target. The bulge target genes are significantly more negatively correlated to their miRNAs'

98 expression than the background (Wilcox test, $p < 0.01$) and for miR-191, the bulge
99 target even outperforms the seed target (Wilcox test, $p < 0.01$) (Figure 3).
100 We also use public data on transcriptome change after over-expression or knock-
101 down individual miRNAs from GEO. For miR-126, miR210 and miR-184, all the
102 bulge targets were significantly down-regulated after over-expression (Table 2) and in
103 the case of the knock-down experiment for miR-1204, the bulge target is also much
104 more highly expressed compared with the non-target gene (Wilcox test, $p < 0.01$).

105 **Free energies of CG bulge target duplexes are significantly lower than the** 106 **random bulges**

107 We compared the minimum free energy (MFE) between the canonical target, CG bulge
108 target and target with random bulges using RNAhybrid [16]. The non-canonical target
109 with a bulge between CG has a significantly lower MFE compared with the target
110 with random bulge (Wilcox test, $p < 0.05$, Figure 4).

111 **Validation of the bulge target site by the CLASH data**

112 To allow direct mapping of miRNA-target interactions, we use the CLASH dataset
113 [12] to validate our bulge target for the miRNAs containing CG dimer. Briefly, the
114 RNA molecules present in AGO-associated miRNA-target duplexes were partially
115 hydrolyzed, ligated, reverse transcribed and subjected to illumina sequencing.
116 Compared with the HITS-CLIP and PAR-CLIP dataset, CLASH technology
117 generated a group of reads which contain the miRNAs and their target site sequence
118 together (chimeric reads). In all the six independent CLASH experiments, we find 10
119 CG dimer miRNAs were detected in all the chimeric reads and 8 miRNAs had, in
120 total, 264 chimeric reads containing a bulge nucleotide between the CG dinucleotide
121 target site (Supplementary Table 4). For all miRNAs detected in the CLASH dataset,
122 the non-canonical interactions (G.U pairs, all possible one nt mismatch or bulge; non-
123 canonical seed) were about 1.7-fold more than the perfect seed base pairing. But

124 within the CG miRNA, only the bulge targets between CG dimer, which in
125 comparison to randomized sequences, showed strong enrichment among all the
126 interactions (Figure 5).

127 **Discussion**

128 The aim of this study is to identify the general features of miRNAs and their bulge
129 target interactions. We used the non-canonical miRNA target's interactome which
130 contains bulge nucleotide between CpG dinucleotide to test whether a bulge position
131 is random or has specific rules. This sub-class of miRNA was observed to have a few
132 seed targets and these seed targets are evolutionally less conservative, which makes
133 these miRNAs potentially non-canonical target rich. Multistep validation, which
134 included evolutionary, overexpression, correlation and CLASH analysis, supports the
135 reliability that between CpG in the seed region, there is a bulge containing a target
136 group.

137 Multiple studies have found the existence of a bulge target for miRNAs. In particular,
138 the PAR-CLIP and HIS-CLIP technique proved the abundance of non-canonical
139 targets binding to the RISC complex, and the bulge target was considered as one of the
140 major types within such a non-canonical target. However, the short of high-
141 throughput way to detect miRNA-target duplex makes it difficult to predict the
142 position of the hot spot of the bulge nucleotide. Thus, most computational algorithms
143 are restricted to predict only the seed target.

144 Notably, different miRNAs vary in the target interaction pattern. A previous study
145 [17] also showed that the composition of seed sequences is a major determinant of the
146 miRNA target pattern. However, of all the different kinds of non-canonical target
147 groups, even with the CLASH dataset, there are still no key features to distinguish any
148 miRNA group with a higher bulge target proportion.

149 Intriguingly, the CG bulge target genes are functionally enriched in synapse and
 150 neruon projection and development, indicating that fast evolved neuron cells are most
 151 sensitive to the pertubation of seed complementary base-pairing and most receptive in
 152 accomodating evolutionary innovation, such as bulge target recognition.
 153 Thus, a major novelty of this work is the identification of a sequence motif, CG
 154 dimer, in the seed region of miRNAs is strongly correlated to bulge targeting patterns.
 155 This seed variability issue should be taken into consideration when predicting targets
 156 for individual miRNAs.
 157 Overall, we found that the bulge targets were preferentially associated with the
 158 miRNAs containing CG dinucleotide in their seed region.

159 **Methods**

160 **miRNA sequences and 3'UTR sequence alignments**

161 Mature miRNA sequences were obtained from the miRBase website
 162 (<http://www.mirbase.org>) [13]. miRNAs are considered to be conserved if they share
 163 the same mature sequence in different groups of species, vertebrate, mammal and
 164 primates. Genomic coordinates of Ensembl human genes (hg18) were used to extract
 165 the human 3'UTR sequences and the corresponding aligned sequences from the 28-
 166 species alignment (MAF file) available at the UCSC Table browser. Only protein
 167 coding genes were included in the database and when several mRNA isoforms were
 168 reported for the same Ensembl gene ID, only the one with the longest 3'UTR
 169 sequence was used in the analysis.

170

171 **Predictions of seed and bulge target for conservative CG dimer miRNAs**

172 The seed sequences for the CG dimer miRNAs were extracted to find three types of
 173 targets. Any coding gene's 3'UTR containing a perfect complementary sequence was

174 defined as a seed target. For the bulge target, we allowed one extra nucleotide to exist
175 between CG dimer. Randomly inserted single nucleotide seed sequences were used as
176 control. The occurrences of the homologous target sites in different species were
177 summed up for seed, bulge and control separately as the conservation rates.

178 **miRNAs and target expression correlation analysis**

179 12 human brain prefrontal cortex samples' miRNA (GSE29356) and coding gene
180 transcriptome data (GSE22570) were used to check expression correlation. Both the
181 Pearson and Spearman methods were used to calculate the correlation. For miRNA
182 overexpression in vitro, miR-126 in LM2 breast cancer cell (GSE23905), miR-184 in
183 SY5Y (GSE26545) and miR-210 in MCF-7 cells and MDA-MB-231 cells
184 (GSE25162) were downloaded from GEO and for the miRNA knockdown
185 experiment, miR-126 in MDA-MB-231 cells and miR-1204 in SUM159PT
186 (GSE37185). The Mann-Whitney-Wilcoxon Test was used to test the seed and bulge
187 target expression change in the transfection experiment.

188 **Confirm bulge target with CLASH dataset**

189 The miRNA-mRNA interaction sequences were downloaded from the journal's website in
190 the supplementary Data section [12]. In this published CLASH dataset, the
191 crosslinked RNA-induced silencing complex (RISC) in HEK293 cells were
192 immunoprecipitated. The miRNA and cognate mRNA target transcripts were ligated
193 and sequenced together. The chimeric reads containing vertebrate conservative CG
194 dimer miRNAs were extracted. The bulge target was recognized if there is one extra
195 nucleotide between CG dimer.

196 **Competing interests**

197 The authors declare that they have no competing interests.

198

199 **Authors' contributions**

200 Z.Y and B.Z performed the computational analyses. Y.P.C and Z.Y designed the
201 research and wrote the paper.

202

203 **Acknowledgements**

204 We thank the members of the Khaitovich and Chen laboratories for helpful
205 discussions. We thank Xiaowei Wang for the suggestion. This work was supported by
206 La Trobe University Postgraduate Scholarship.

207

208

209 **Figures**

210 **Figure 1 - miRNA and Seed/bulge target duplex model**

211 The miR-184 seed sequence was used to illustrate a canonical target match, and a
212 non-canonical target match with a bulge nucleotide between the CG dinucleotide.

213 **Figure 2 - Targets with Bulge between CG dimer have higher conservative rate**

214 Two different target site conservation rates in 23 vertebrate compared with a random
215 bulge target.

216 **Figure 3 - CG dimer miRNA suppression of the bulge target expression.**

217 Cumulative distribution of correlation coefficient between miR-191 and target
218 expression level.

219 **Figure 4 - Mimium Free energy between seed pair, CG bulge and Random**
 220 **bulge.**
 221 Boxplot of the distrubution of the canonical target, noncanonical bulge interaction and
 222 random bulge target.

223

224 **Figure 5 - CLASH data validation o f CpG containing miRNA and bulge target**
 225 **interaction.**
 226 Proportion of canonical seed interaction, noncanonical bulge interaction among
 227 CLASH chimeras, scramble and shuffle data.

228

229

230 **Tables**

231 **Table 1 - Vertebrate conservative miRNA TargetScan result**
 232 All the conservative miRNAs in 23 vertebrate species listed in Supplementary Table 1
 233 were used to calulate their canonical target conservative rates by TargetScan. The
 234 results were sorted from lowest conservative rate to highest.

235 **Table 2 - Functional analysis of CpG miRNA and their bulge targets**
 236 The transcriptome data after overexpression or knockdown studies. The Mann-
 237 Whitney-Wilcoxon Test was used to test the seed and bulge target expression change
 238 compared with random genes.

239 **Additional files**

240 **Supplementary Table 1 – All the 23 species which were used for miRNA and**
 241 **target analysis.**
 242 All the miRNAs and their target conservation scores were calculated based on these
 243 23 species. File was in Excel format.

244

245 **Supplementary Table 2 – Mammal conservative miRNA TargetScan result**

246 All the miRNAs which have same mature sequence in the mammal species were

247 tested were sorted according to their target site conservation rate. File was in Excel

248 format.

249

250 **Supplementary Table 3 – Primate conservative miRNA TargetScan result**

251 All the miRNAs which have same mature sequence in the primate species we tested

252 were sorted according their target site conservation rate. File was in Excel format.

253 **Supplementary Table 4 – CLASH chimeras reads which the miRNA-target**
 254 **duplex were sequenced**

255 All the CG miRNAs which were detected together with their bulge targets between

256 CG dimer are listed in this table. File was in Text format.

257

258

259

260

261

262 **References**

263 1. Bartel DP: **MicroRNAs: Genomics, Biogenesis, Mechanism, and Function.** *Cell*

264 2004:281–297.

265 2. Brodersen P, Voinnet O: **Revisiting the principles of microRNA target**

266 **recognition and mode of action.** *Nat Rev Mol Cell Biol* 2009, **10**:141–148.

267 3. Bartel DP: **MicroRNAs: Target Recognition and Regulatory Functions.** *Cell*

268 2009:215–233.

- 269 4. Brennecke J, Stark A, Russell RB, Cohen SM: **Principles of microRNA-target**
270 **recognition.** *PLoS Biol* 2005, **3**:0404–0418.
- 271 5. Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC: **Evolution,**
272 **biogenesis, expression, and target predictions of a substantially expanded set of**
273 **Drosophila microRNAs.** *Genome Res* 2007, **17**:1850–1864.
- 274 6. Friedman RC, Farh KKH, Burge CB, Bartel DP: **Most mammalian mRNAs are**
275 **conserved targets of microRNAs.** *Genome Res* 2009, **19**:92–105.
- 276 7. Chi SW, Hannon GJ, Darnell RB: **An alternative mode of microRNA target**
277 **recognition.** *Nat Struct Mol Biol* 2012, **19**:321–327.
- 278 8. Ha I, Wightman B, Ruvkun G: **A bulged lin-4/lin-14 RNA duplex is sufficient for**
279 **Caenorhabditis elegans lin-14 temporal gradient formation.** *Genes Dev* 1996,
280 **10**:3041–3050.
- 281 9. Vella MC, Choi E-Y, Lin S-Y, Reinert K, Slack FJ: **The C. elegans microRNA**
282 **let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR.** *Genes*
283 *Dev* 2004, **18**:132–137.
- 284 10. Didiano D, Hobert O: **Perfect seed pairing is not a generally reliable predictor**
285 **for miRNA-target interactions.** *Nat Struct Mol Biol* 2006, **13**:849–851.
- 286 11. Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang TY, Pasquinelli AE, Yeo
287 **GW: Comprehensive discovery of endogenous Argonaute binding sites in**
288 **Caenorhabditis elegans.** *Nat Struct Mol Biol* 2010, **17**:173–179.

- 289 12. Helwak A, Kudla G, Dudnakova T, Tollervey D: **Mapping the human miRNA**
290 **interactome by CLASH reveals frequent noncanonical binding.** *Cell* 2013,
291 **153**:654–665.
- 292 13. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase:**
293 **microRNA sequences, targets and gene nomenclature.** *Nucleic Acids Res* 2006,
294 **34**:D140–D144.
- 295 14. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by**
296 **adenosines, indicates that thousands of human genes are microRNA targets.** *Cell*
297 2005, **120**:15–20.
- 298 15. Somel M, Liu X, Tang L, Yan Z, Hu H, Guo S, Jiang X, Zhang X, Xu G, Xie G,
299 Li N, Hu Y, Chen W, Pääbo S, Khaitovich P: **MicroRNA-driven developmental**
300 **remodeling in the brain distinguishes humans from other primates.** *PLoS Biol*
301 2011, **9**:e1001214.
- 302 16. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R: **Fast and effective**
303 **prediction of microRNA/target duplexes.** *RNA* 2004, **10**:1507–1517.
- 304 17. Wang X: **Composition of seed sequence is a major determinant of microRNA**
305 **targeting patterns.** *Bioinformatics* 2014, **30**:1377–1383.

306

307

Bulge target

Seed target

Minimum Free Energy

variable

bulge

random bulge

seed