PATROCLES: a database of polymorphic miRNA-mediated gene regulation

Samuel Hiard[1], **Denis Baurain**[2], **Wouter Coppieters**[2], **Xavier Tordoir**[2], **Carole Charlier**[2] **and Michel Georges**[2]

[1] Systems and Modeling, Montefiore Institute, University of Liège, Belgium.
[2] Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège, Belgium.
denis.baurain@ulg.ac.be, carole.charlier@ulg.ac.be

FINE-TUNING of gene expression by miRNAs requires a functional silencing pathway with many components. The corresponding sequence space (target 3'-UTRs, miRNA precursors and silencing machinery) is bound to suffer its toll of DNA sequence polymorphisms (DSPs) of which some have been demonstrated to alter phenotype. When functional, DSPs affecting miRNA-mediated post-transcriptional regulation are unlikely to create highly penetrant phenotypes. Instead they are expected to contribute to genetic variation of traits with complex inheritance. To assist in the identification of such DSPs we have mined public databases for Single Nucleotide Polymorphisms (SNPs), Copy Number Variants (CNVs) and expression QTL (eQTL) in the three sequence compartments involved in regulation by miRNAs. The result of our search is browsable via the PATROCLES website (`http://www.patrocles.org/`).

Methods

Three distinct pipelines ensure the identification of DSPs affecting the three compartments (see **Fig. 1** for polymorphic targets). SNPs are analyzed in all three pipelines, while CNVs and eQTL are only used for miRNA precursors and machinery genes.

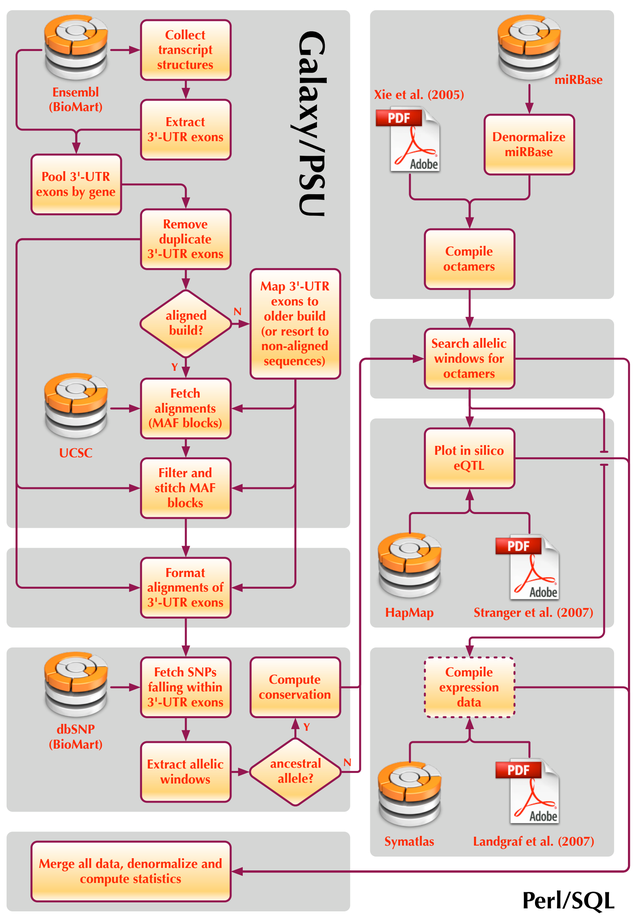

Figure 1: ***Pipeline for characterizing polymorphic targets.*** *Except for the steps performed remotely using the Galaxy server at Penn State, all computations are carried on locally through a combination of Perl scripts and SQL queries.*

SNPs

Target sites in 3'-UTRs are defined as ~1,200 octamers either complementary to the seed of known miRNAs (**Fig. 2**) or unusually frequent and/or conserved in 3'-UTRs (Xie *et al.*, 2005). First, the ancestral allele of each SNP falling in a 3'-UTR is identified by comparison with aligned orthologs. Encompassing octamers are then examined for potential targets, possibly conserved across species (**Fig. 3**). According to ancestrality and target conservation, Patrocles SNPs (pSNPs) are categorized as non-conserved destroyed, conserved destroyed, non-conserved created, conserved created (revertants), polymorphic, or shifted. The effect of SNPs falling in miRNA precursors is analyzed with RNAFOLD, whereas the effect of those falling in genes involved in miRNA biosynthesis or silencing machinery is extracted from ENSEMBL annotations.

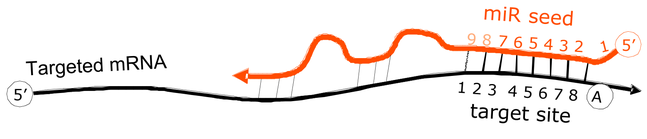

Figure 2: ***Generation of octamers from miRNAs.*** *Following Lewis et al. (2005), miR octamers correspond to the Watson-Crick reverse complement of nucleotides 2 to 8 of known miRNAs followed by an "A anchor" at their 3'-end. Whereas the same 540 octamers from Xie et al. (2005) are used for all* PATROCLES *species, miR octamers are species-specific and rely on miRBase contents.*

CNVs and eQTL

Respectively available for human, mouse and rat, or only for human, CNV and eQTL coordinates obtained through database and literature mining are mapped on miRNA precursors and machinery genes. Any gene overlapping (even partially) such regions is considered affected and flagged in PATROCLES.

```
1. human: A   ...TTTGGTG[A]AACCAAC... => ancestral allele
   human: G   ...TTTGGTG[G]AACCAAC... =>   derived allele
   chimp      ...TTTGGTG[A]AACCAAC... =>   sibling species

2.    rat     ...TTTGGTG[A]AACAAAC...
    mouse     ...CTTGGTG[A]AACAAAC...

3.    dog     ...TTTGGTG[A]AACTAAC...
    cow       ...TTTGGTG[A]AACTAAC...

(3/3)            TTTGGTG[A]
(3/3)             TTGGTG[A]A
(3/3)              TGGTG[A]AA
(3/3)               GGTG[A]AAC
(2/3) not in dog/cow  gtg[a]aacc
(2/3) not in dog/cow   tg[a]aacca
(2/3) not in dog/cow    g[a]aaccaa      => hsa-miR-29b-2*
(2/3) not in dog/cow     [a]aaccaac
```

Figure 3: ***Target identification and conservation.*** *UCSC aligned block from the 3'-UTR of human gene* ENSG00000151136 *centered on SNP* rs2241183 *(in brackets). The ancestral allele (A) has been identified by comparison with the chimp ortholog. When no sibling sequence is available, a candidate allele is considered ancestral if conserved in at least one ortholog from each of three groups (e.g., primates, rodents and other mammals). A sliding window is then used to search for octameric targets in both allelic variants. Each octamer is simultaneously screened for conservation using the same criterion as for ancestrality. The lower part of the figure shows the eight octamers of the A-variant, among which the first four are conserved, the seventh being the only octamer that corresponds to a target, though not conserved here.*

Results

PATROCLES content statistics

Currently, polymorphic targets are available for five mammals and chicken, though to varying extent due to largely unequal amounts of input data (**Tables 1–3**).

	human	mouse	rat	cow	dog	chicken
3'-UTRs	24,319	21,911	12,798	12,954	7,640	11,208
SNPs in 3'-UTRs	136,147	126,230	9,534	3,909	2,465	14,769
pSNPs	31,995	24,523	1,376	365	293	1717
miRNA precursors	676	466	280	114	203	145
matures	676	484	285	114	176	123
matures*	170	117	58	8	1	9
octamers	683	466	274	83	135	89

Table 1: ***Comparative statistics across species.***

	miRBase	Xie 2005	both
octamers	683	540	1164
targets	375,024	323,812	661,137
conserved	40,715	74,435	104,725
affected	26,719	20,679	45,119
NC destroyed	10,328	7,392	16,954
C destroyed	959	1,546	2,266
NC created	11,244	9,006	19,301
C created	58	50	104
polymorphic	3,295	1,944	4,970
shifted	837	741	1,526

Table 2: ***Targets and pSNPs in human genes.***

	miRNAs	machinery
genes	377	51
SNPs	184	237
...in precursors	136	n.a.
in matures	36	n.a.
in seeds	12	n.a.
CNVs	158	17
eQTL	78	21

Table 3: ***DSPs in human miRNAs and machinery genes.***

Characterization of PATROCLES targets

To evaluate the validity of PATROCLES targets, we assembled three collection of human octamers as following: (1) all unique miRNA* octamers from miRBase (controls; n=148); (2) all unique miRNA octamers found on the same precursors (n=106); (3) all unique octamers from Xie et al. (2005) not corresponding to any known miRNA (n=422). Target and pSNPs data pertaining to these octamers were then analyzed (**Figs 4–8**).

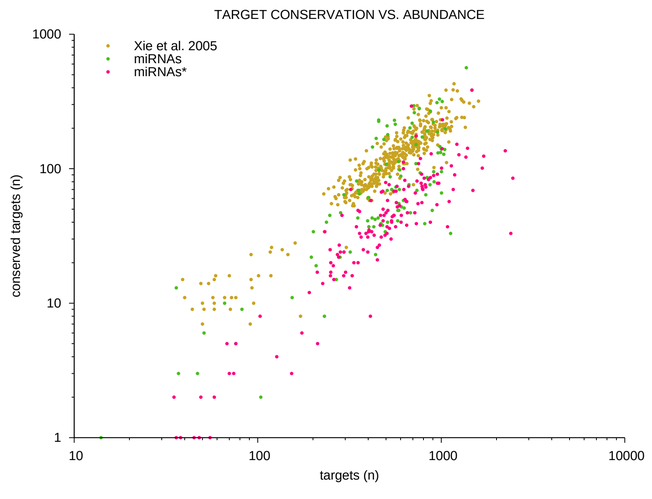

Figure 4: ***Target conservation and abundance.*** *For each octamer, the number of conserved targets is plotted as a function of the total number of targets. As expected from the protocol used for their identification, octamers from Xie et al. (2005) are distinctly more conserved than miRNA* octamers. In contrast, miRNA octamers are scattered, which indicates that they are diversely conserved. Note the logarithmic scale on both axes.*

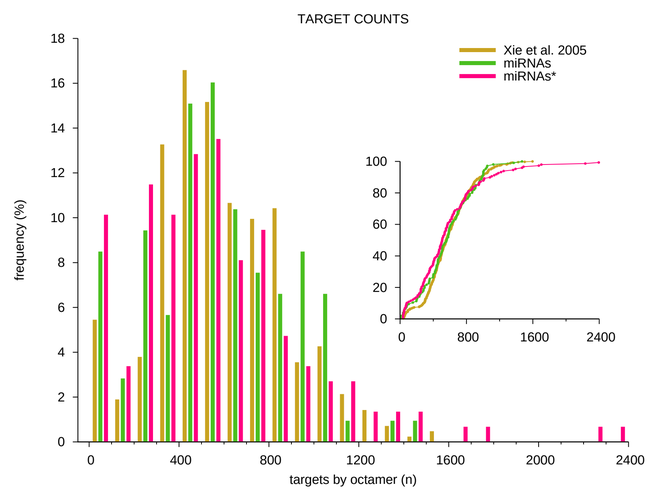

Figure 5: ***Comparative abundance of targets.*** *Total counts for the three collections are shown either as distributions (main plot) or cumulative curves (inset). Note the shared bulge of scarce (<100) targets, the excess of common (900–1100) miRNA targets, as well as the excess of very common (>1600) miRNA* targets, along with a depletion in the modal area (400-700).*

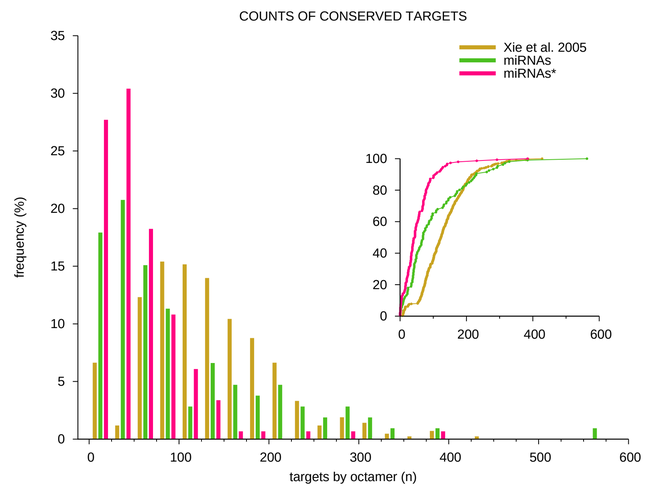

Figure 6: ***Comparative abundance of conserved targets.*** *While miRNA* octamers are the less frequently conserved and Xie octamers the most, the distribution of conserved miRNA octamers is intermediate and of a more complex shape.*

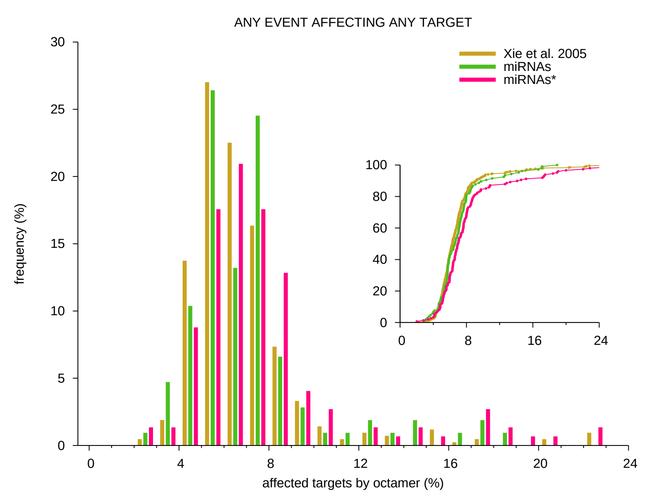

Figure 7: ***Comparative abundance of pSNPs.*** *Xie octamers are less affected than others. Among octamers derived from miRNA precursors, true miRNAs are less affected than miRNAs*. This suggests that* PATROCLES *targets are indeed under selection.*

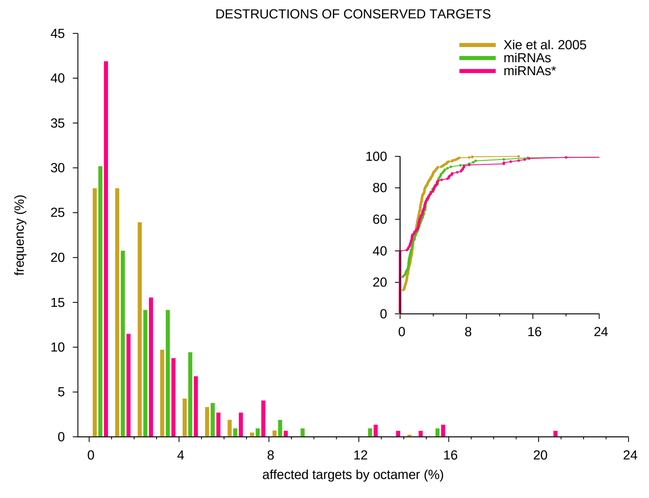

Figure 8: ***pSNPs destroying conserved targets.*** *In spite of a left shift due to scarcity of conserved targets, comparison of the three collections indicates that true targets are under selection.*

Acknowledgments

This work was funded by grants from the EU "CallimiR" STREP project, the Belgian Science Policy organisation (SSTC Genefunc PAI), from the Communauté Française de Belgique (Game & Biomod ARC) and from the University of Liège. C.C. is Chercheur Qualifié from the FNRS.