PATROCLES: a database of polymorphic miRNA-mediated gene regulation in vertebrates

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

Nucleic Acids Research Advance Access – doi:10.1093/nar/gkp926

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.

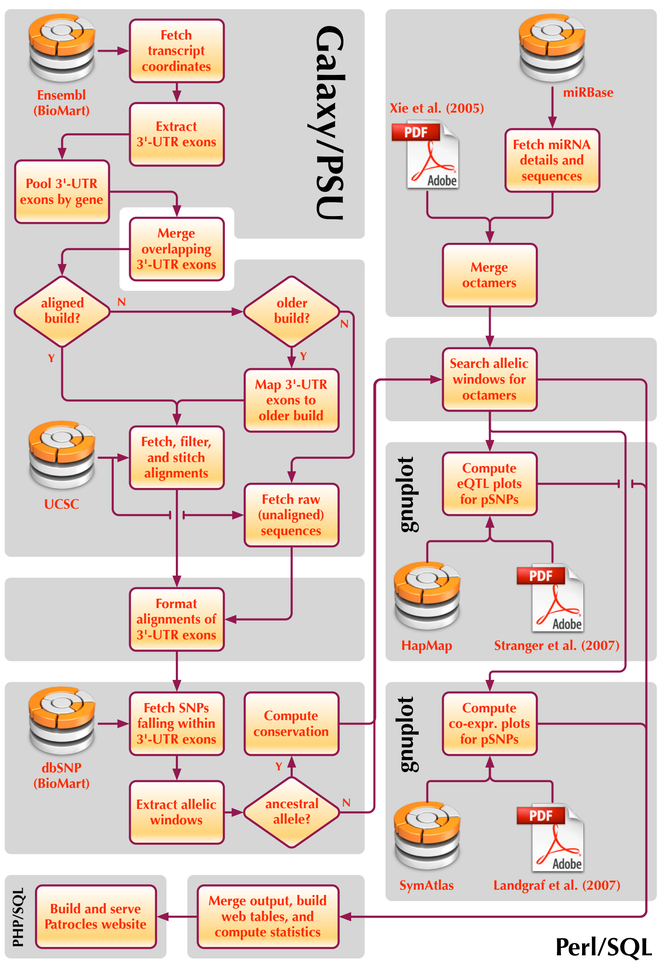

Figure 1: ***Pipeline for characterizing polymorphic targets.*** *Except for the steps performed remotely using the Galaxy server at Penn State, all computations are carried on locally through a combination of Perl scripts and SQL queries.*

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, polymorphic, or shifted. To improve sensibility without compromising specificity, heptamers are also considered as targets when conserved. 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.

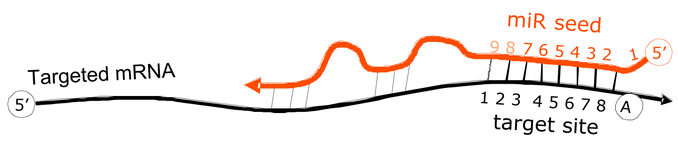

Figure 2: ***Generation of octamers from miRNAs.*** *Following Lewis* et al. *(2005), miRNA octamers correspond to the Watson-Crick reverse complement of nucleotides 2 to 8 of known miRNAs followed by an "A anchor" at their 3'-end. Conserved heptamers lack either the A anchor in 3' (7mer-m8) or the eighth seed match in 5' (7mer-A1). Whereas the same 540 octamers from Xie* et al. *(2005) are used for all PATROCLES species, miRNA octamers are species-specific and rely on miRBase content.*

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 to miRNA precursors and machinery genes. Any gene overlapping (even partially) such regions is considered affected and displayed 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

Currently, SNPs affecting targets, miRNA precursors or silencing machinery genes are compiled for six mammals and chicken, though with varying efficiency due to largely unequal amounts of input data (**Tables 1–3**).

	human	mouse	chimp	rat	dog	cow	chicken
3'-UTRs	24,319	21,911	16,576	12,798	7,640	12,954	11,208
SNPs in UTRs	136,159	126,589	6,365	11,275	3,761	3,891	14,385
pSNPs	31,995	24,590	664	1,639	429	363	1,680
miRNA genes	676	466	99	280	203	114	145
matures	676	484	90	285	176	114	123
matures*	170	117	1	58	1	8	9
octamers	683	466	74	274	135	83	89

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

	miRBase	Xie 2005	both
octamers	683	540	1,164
targets	375,054	323,833	661,187
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
polymorphic	3,295	1,944	4,970
shifted	837	741	1,526

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

	miRNAs	machinery
genes	377	52
SNPs	184	237
...in precursors	146	n.a.
in matures	26	n.a.
in seeds	12	n.a.
CNVs	158	17
eQTL	78	21

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

Characterization of octamers and targets

In human, the two octamer sets jointly define 1,164 unique octamers of which only 5% are common. Thus, at first glance, both sets appear to explore very distinct sequence domains. However, when considering heptamers and hexamers embedded within octamers, the concordance between the two sets increases markedly (**Fig. 4**).

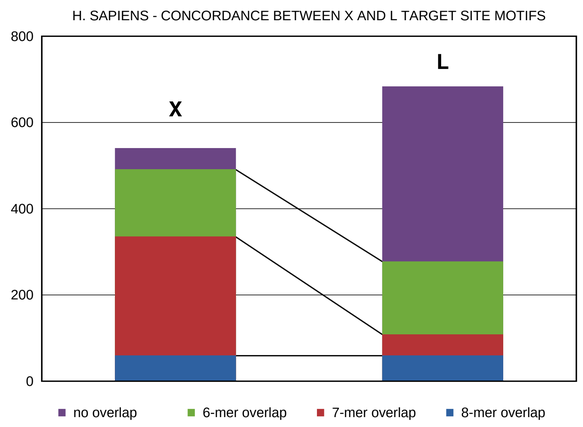

Figure 4: ***Concordance between the two octamers sets.*** *Number of octamers from Xie* et al. *(2005) (left column) and human miRNA octamers (right column) with overlapping octamer (blue), heptamer (red), hexamer (green) or without overlap (purple).*

To characterize PATROCLES targets, human octamers were split into three collections: (1) all unique octamers from Xie *et al.* (2005); (2) all unique miRNA octamers from miRBase; and (3) all unique miRNA* octamers from miRBase. Target abundance and conservation were then examined separately for each of these three collections (**Figs 5–6**).

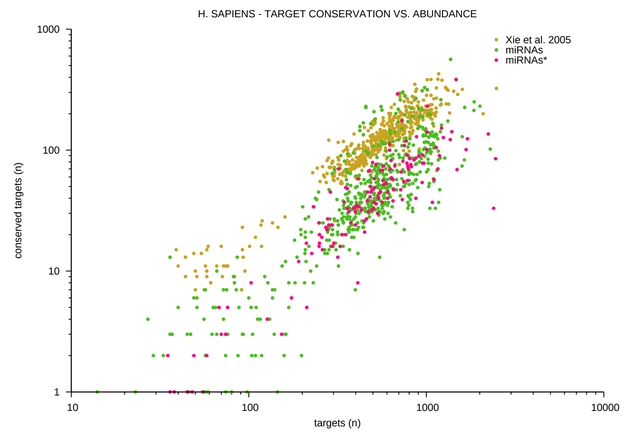

Figure 5: ***Target conservation vs. abundance.*** *For each octamer, the number of conserved targets is plotted as a function of the total number of targets. As expected from the protocol used for their identification, octamers from Xie* et al. *(2005) are distinctly more conserved than miRNA* octamers. In contrast, miRNA octamers are scattered, which indicates that they are diversely conserved. Note the logarithmic scale on both axes.*

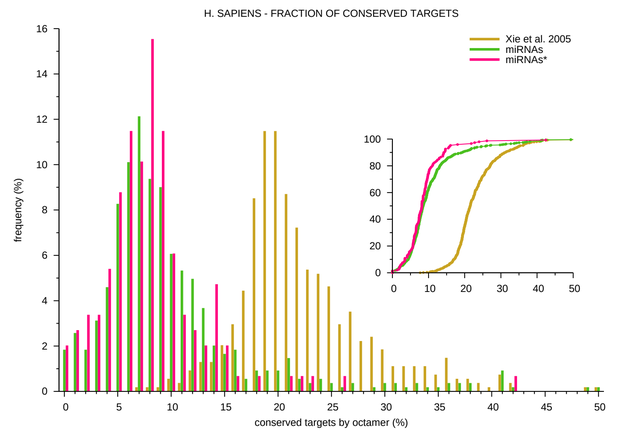

Figure 6: ***Fraction of conserved targets.*** *Proportions of conserved targets for the three octamer collections are shown either as distributions (main plot) or as cumulative curves (inset). Amongst miRBase-derived octamers, the proportion of conserved target sites is higher for miRNAs than for miRNAs*.*

Biological relevance of PATROCLES content

To validate PATROCLES predictions, we looked for signatures of purifying selection on pSNPs by simulating sets of pSNPs matching the original human set. To this end, the position of each true SNP was randomly shifted in the 3'-UTR space, yet respecting its trinucleotide context. The number of 'Patrocles events' obtained with the original set was then compared with the distribution of those compiled across 100 simulated SNP sets. As expected, we observed a strong signature of purifying selection against pSNPs that destroy conserved target sites (**Fig. 7**).

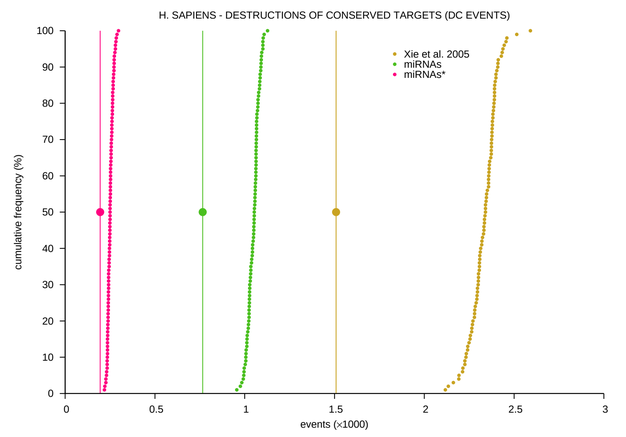

Figure 7: ***Selection on pSNPs destroying conserved targets.*** *Large dots: number of target destructions in the original SNP set. Small dots: cumulative frequency distribution obtained with 100 simulated SNP sets (see text). Evidence for purifying selection is found for each of the three octamer collections.*

Acknowledgments

This work was funded by grants from EU Framework 6 (CallimiR STREP, Epigenome NoE, Eadgene NoE), the Belgian Science Policy organisation (SSTC Genefunc, BioMAGNet PAI), the Fonds National de la Recherche Scientifique (FNRS), the Communauté française de Belgique (Game, BIOMOD ARC) and the University of Liège. C.C. is Chercheur Qualifié of the FNRS.