Key Structural Patterns for miRNA Family Reconstruction

German Conference on Bioinformatics 2019

Cristian A. Velandia Huerto[1]

Advisor: Prof. Dr. *Peter F. Stadler*[1]
Collaborations: *Ali Yazbeck*[1]

[1] Inst.f.Informatik, Universität Leipzig, Leipzig, Germany.

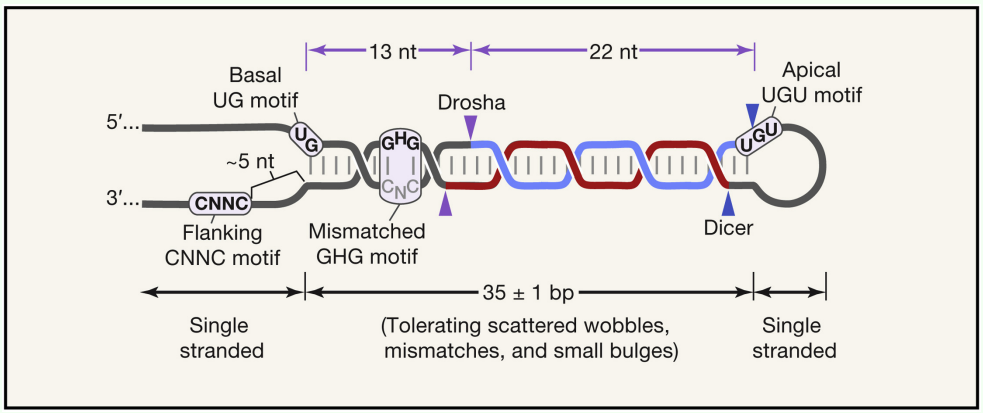
18.09.2019

Context: Do you know miRNAs?

What is a miRNA?

- a class of small RNA ($\sim$ 21-24 bases).
- Endogenous and single strand RNAs.
- Function: *Regulation* of gene expression (via Post-transcriptional gene silencing).
- Produced by microbes, sponges, metazoan, plants and viruses.
- Important role in development and physiology.
- Biogenesis pathway is different between plants and animals [Compartmentalized].

miRNA structure



Bartel David P, Cell, Volume 173, Issue 1, (2018)

- Trimming **Pri-**miRNA.
- Transport **Pre-**miRNA.
- Cleavage stem-loop.
- Release miRNA duplex.

Current approaches of detection

Experimental

- Northern blot.
- Microarrays.
- *In situ* hybridization.
- Amplification techniques.

Computational

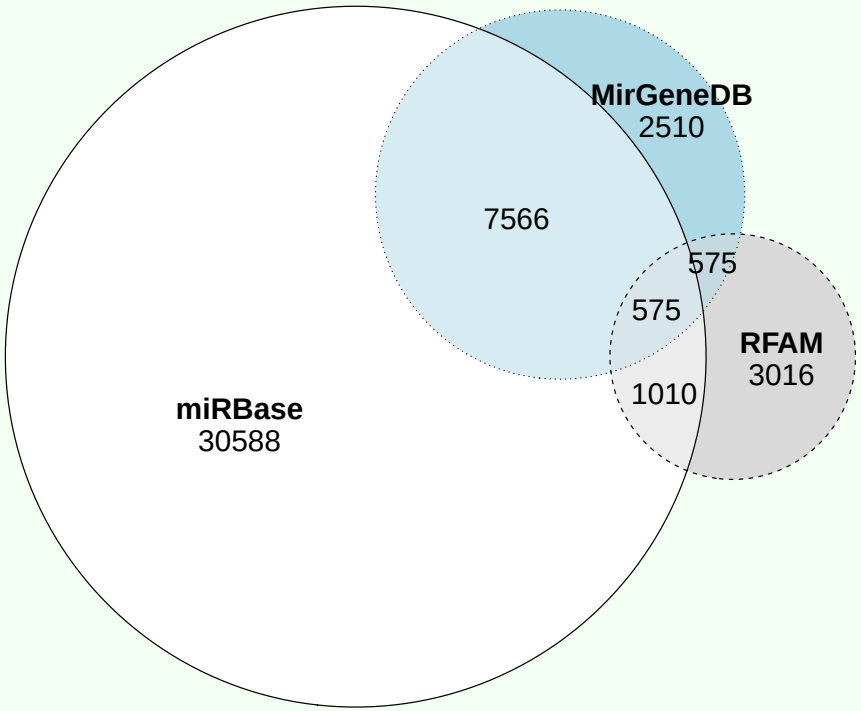
- Homology detection.
- *De novo* detection.

- Due experimental inability to detect all candidates: `Computational approaches`.
- Comprehensive understanding of evolution and functional adaptations of miRNAs requires a *Comprehensive annotation*.

Velandia-Huerto, CA. *et. al*, Evolution and Phylogeny of MicroRNAs-Protocols, Pitfalls, and Problems. Submitted manuscript (2019).

Current miRNA Annotations

Number of reported miRNA precursors

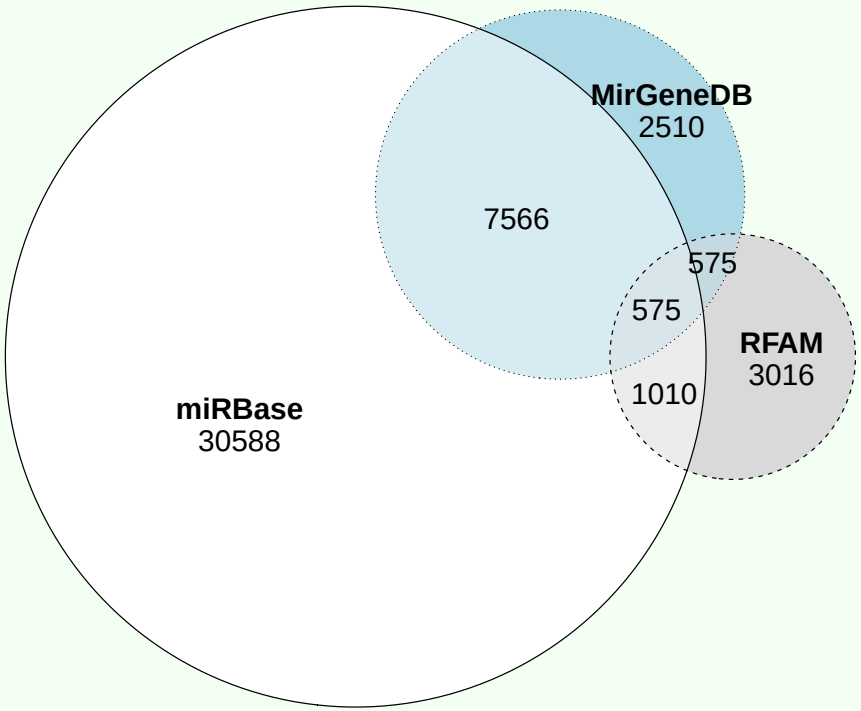


Updated August 30th, 2019

- Until now annotation of miRNA precursors are concentrated on three main databases: **miRBase** v.22 (38 589), **RFAM** v.14.1 (4 026) and **MirGeneDB** v.2 (10 076).

Current miRNA Annotations

Number of reported miRNA precursors



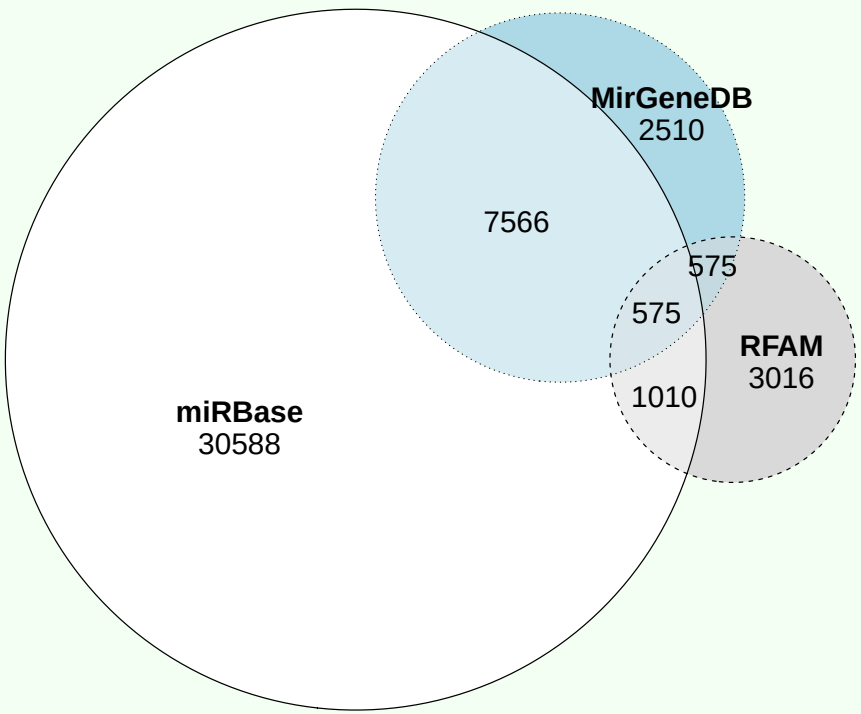
Updated August 30th, 2019

- Until now annotation of miRNA precursors are concentrated on three main databases: **miRBase** v.22 (38 589), **RFAM** v.14.1 (4 026) and **MirGeneDB** v.2 (10 076).
- Classification criteria are not the same

miRBase : seed homology
RFAM : Funcional, evolutionary conserved and evidence Secondary structure.
MirGeneDB : Collection of expression data.

Current miRNA Annotations

Number of reported miRNA precursors



Updated August 30th, 2019

- Until now annotation of miRNA precursors are concentrated on three main databases: **miRBase** v.22 (38 589), **RFAM** v.14.1 (4 026) and **MirGeneDB** v.2 (10 076).
- Classification criteria are not the same

  miRBase : seed homology
  RFAM : Funcional, evolutionary conserved and evidence Secondary structure.
  MirGeneDB : Collection of expression data.
- Different publications prone that **miRBase** report high number of false positives[1]. Despite this, is the most referenced miRNA database.

[1] Bastian Fromm, *et. al* .MirGeneDB 2.0: The metazoan microRNA complement. 2019. bioRxiv 258749; doi: https://doi.org/10.1101/258749

Current resolution on databases

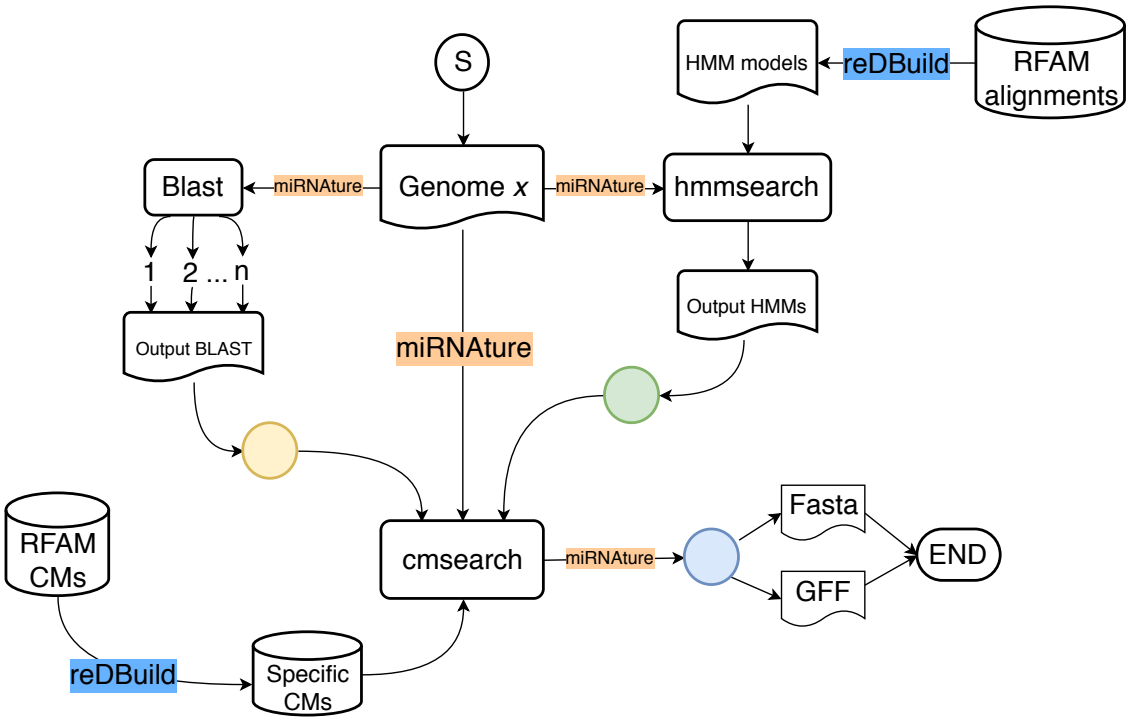
Table: A brief comparison between source miRNA databases

Database	Advantages	Disadvantages
miRBase	Annotation of mature sequences	Automatic family assigment based on seed
RFAM	Multiple alignments precursors	No mature sequences
miRGeneDB	Hand curated based on experimental data	Low number of species

Based on the last evidence. . .

Does it is possible to design a tool to annotate automatically miRNA candidates taking into account the last approaches?

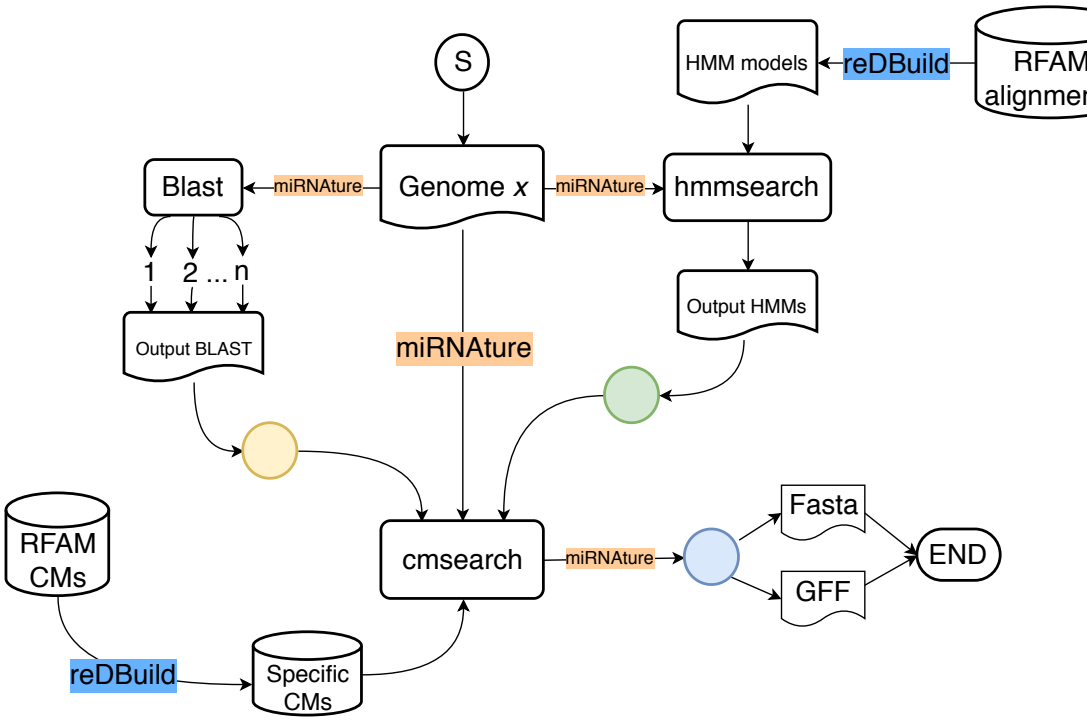
miRNAture

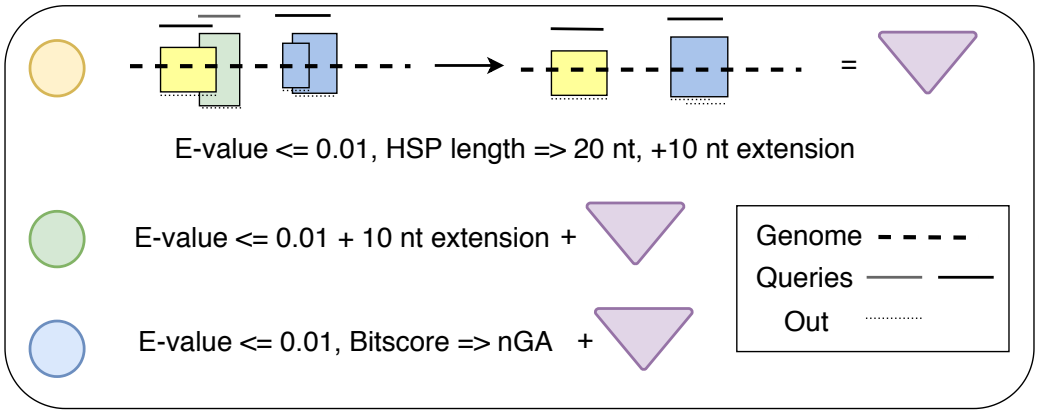


Complete Experimental Design

- 10 `blast` strategies
- 5289 sequences from 10 metazoa species
- 1100 Metazoan-specific Covariance-Models
- 21 subject Chordata genomes

miRNAture



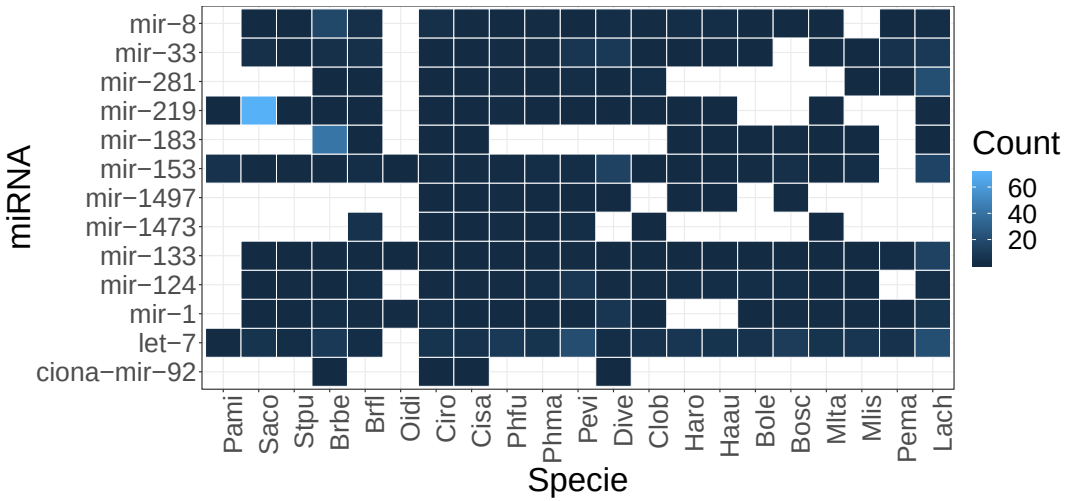


Experimental Design

- 10 `blast` strategies
- 5289 sequences from 10 metazoa species
- 13 Metazoan-specific Covariance-Models
- 21 subject Chordata genomes

Homology on 21 chordata genomes

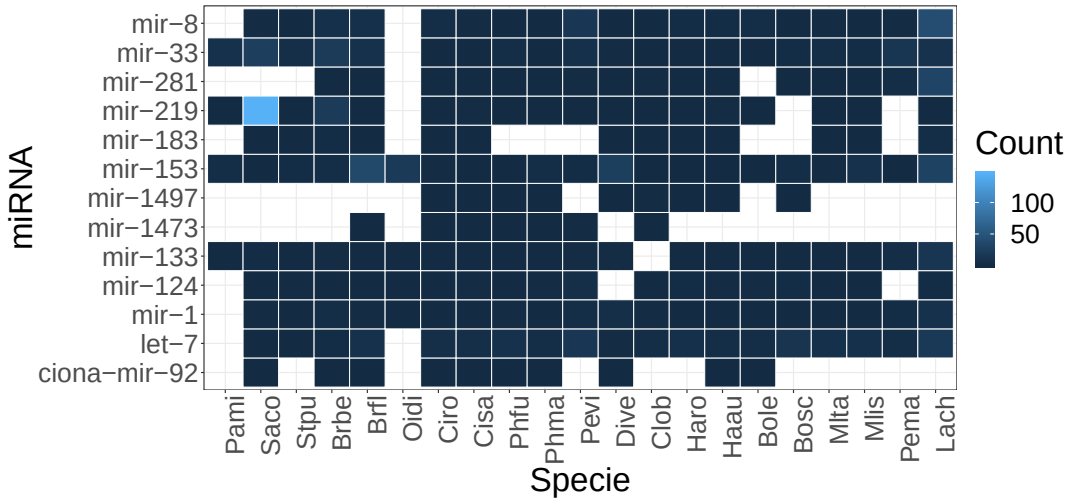
Default *cmsearch* predictions



Clade	**Families** (F)	**Loci** (L)	**Species** (S)	$L/(S * F)$
Echinodermata	8	20	2	1.25
Hemichordata	8	86	1	10.8
Cephalochordata	12	110	2	4.58
Tunicata	13	319	14	1.75
Vertebrata	10	113	2	5.65

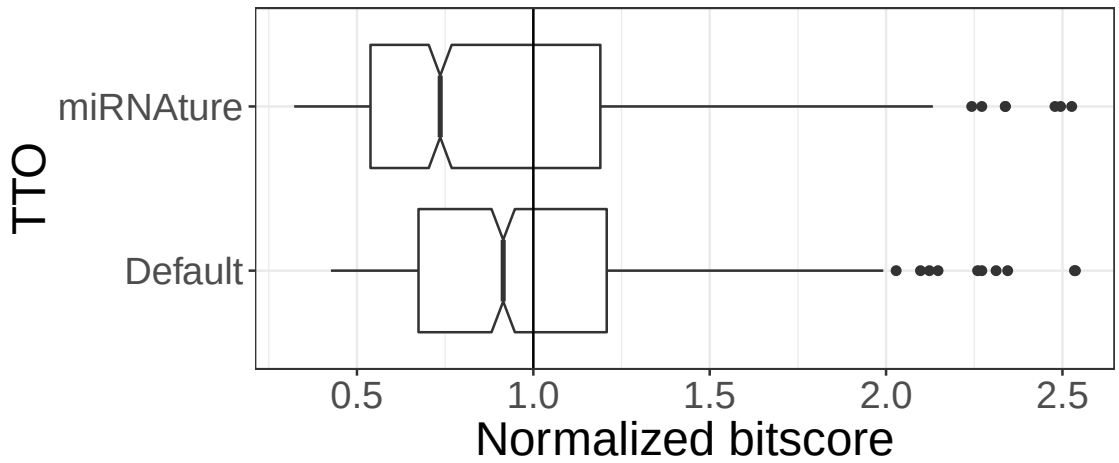
Homology on 21 chordata genomes

`miRNAture` predictions



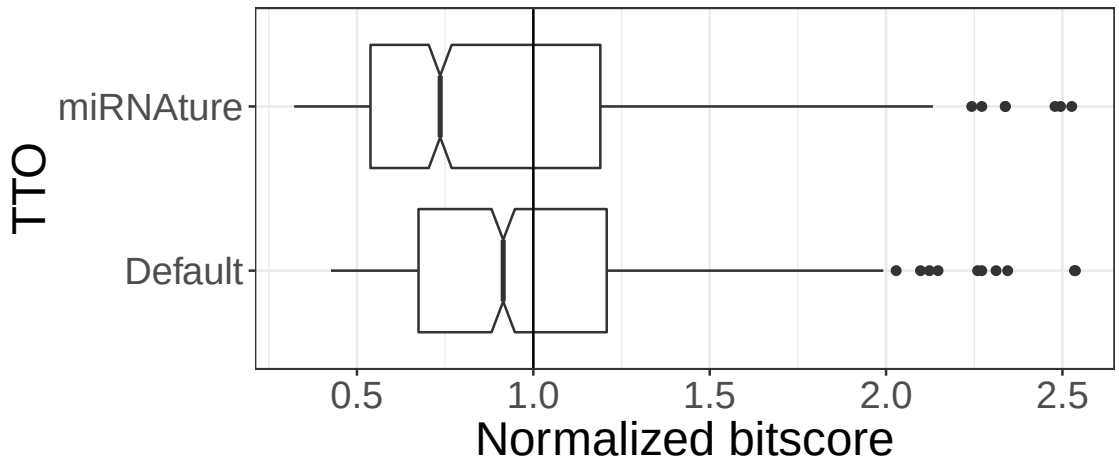
Clade	**Families** (F)	**Loci** (L)	**Species** (S)	$L/(S * F)$
Echinodermata	9	38	2	2.11
Hemichordata	10	188	1	18.8
Cephalochordata	12	143	2	5.96
Tunicata	13	435	14	2.39
Vertebrata	10	197	2	9.85

Homology on 21 chordata genomes



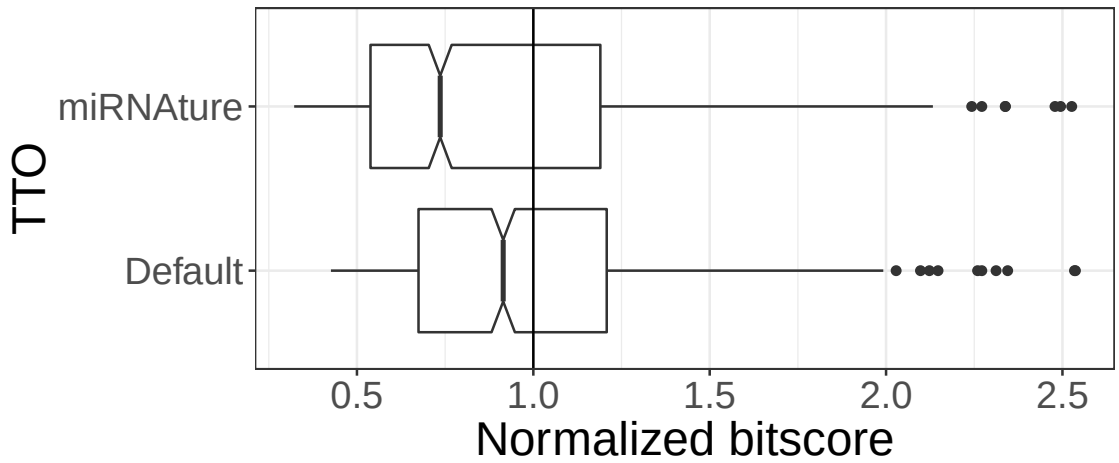
- Applying `miRNAture` the range of new candidates is increased.

Homology on 21 chordata genomes



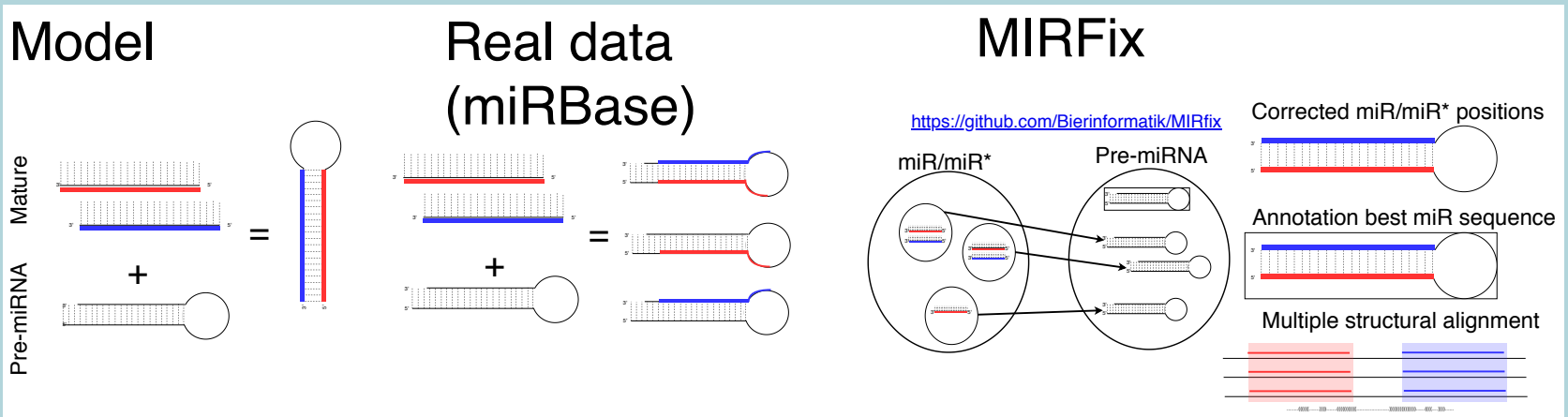
- Applying `miRNAture` the range of new candidates is increased.
- Also, with `miRNAture` the number of possible true positives is greater.

Homology on 21 chordata genomes



- Applying `miRNAture` the range of new candidates is increased.
- Also, with `miRNAture` the number of possible true positives is greater.
- Because of the increasing number of candidates, **additional filters have to be considered to improve sensitivity**.

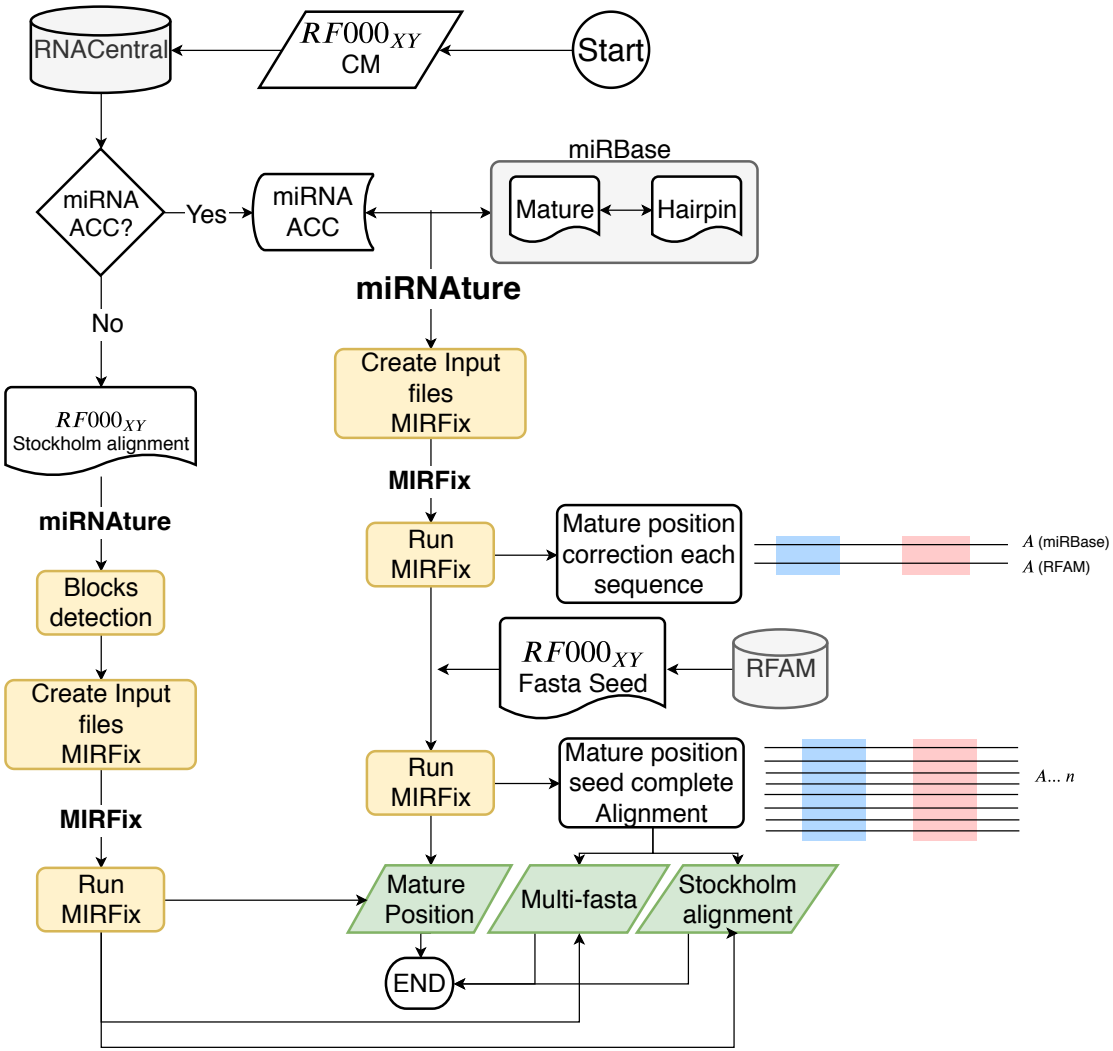
MIRFix



Correction of mature positions from miRBAse

- 48885 mature sequences (`miRBase v.21`)
- 38589 recursor sequences (`miRBase v.21`)

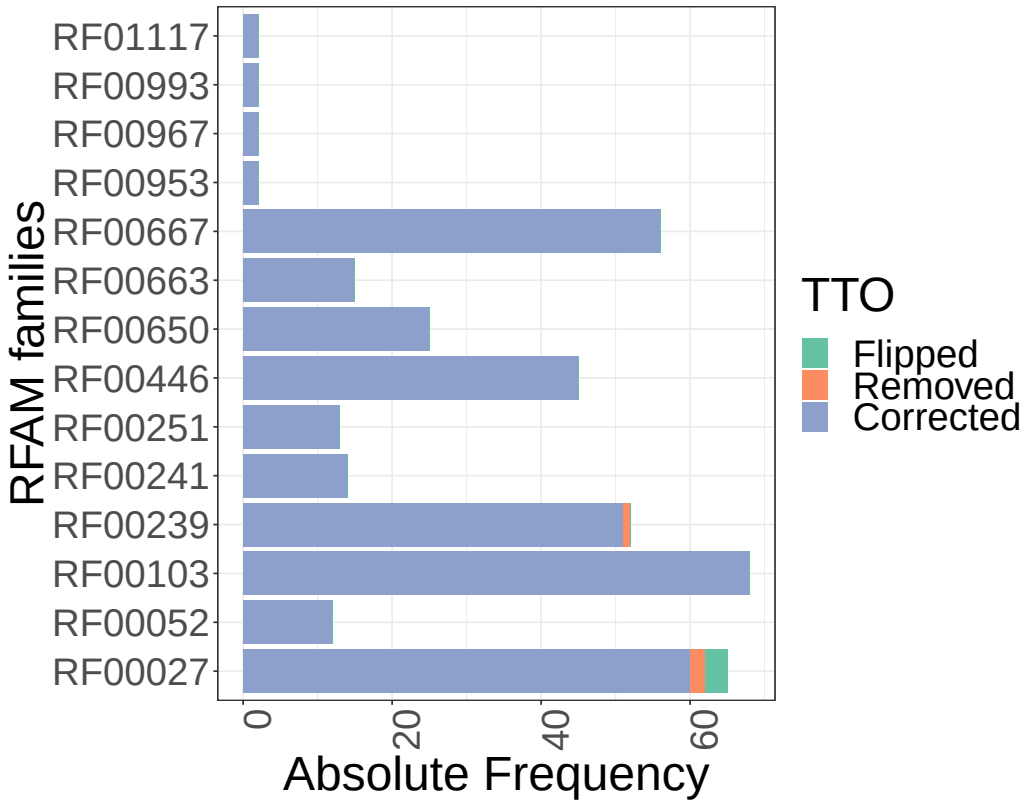
MIRFix + miRNAture



RFAM Mature annotation

- 462 miRNA-specific CM from RFAM 14.1.
- $\sim$ 69.05% (319) CM have at least one sequence represented on `miRBase`
- $\sim$ 30.95% (143) mature prediction were based on the block detection

Annotation of mature sequences RFAM sequences



*subset of families

RFAM Families: 5074*

Mature annotation: 4964

Discarted: 59

Flipped: 51

let-7 (RF00027): 70

Mature annotation: 59

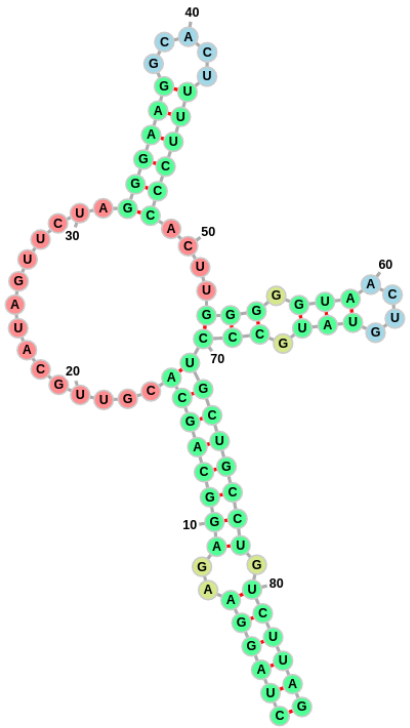
Discarted: 2

Flipped: 3

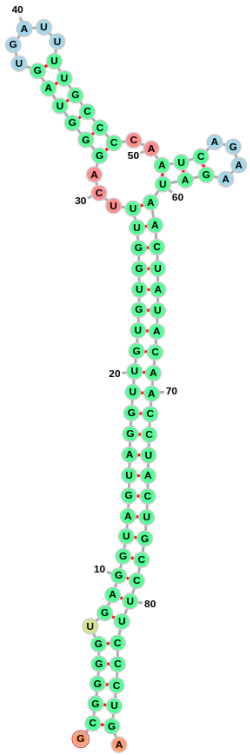
Annotation of mature sequences RFAM sequences

Removed Sequences from RF00027 (let-7)

Macaca mulatta,
AANU01169061.1/33054-33138



Monodelphis domestica ,
AAFR03057393.1/4201-4115



Did not possible to annotate mature sequences.

- Now is possible to annotate the position of mature sequences on miRNA precursors
- Based on the mature sequence, the miRNA is discarted or accepted
- Precursors could be corrected, based on the position of their mature (s) sequence (s)

Annotation and expansion of a *canonical* miRNA

Annotation mature sequences mir-1497 (RF00953)

```
# STOCKHOLM 1.0
B1566810522#1  cugguauaAGCGUAACCUCUUGUAAAUCUCCAACACUAAGCUUGGUuacagcuc-UGUUGAAGAAUUGCAGGUGGUAGGUGCUaucca
B1566808146#0  uggcuguuAGCGCAACCACUUGUGAAUCUCCAACACUCGGCUUUGUccacgacuuUGUUGAAGAAUUGCAGGUGGUAGGUGCUuucaa
#=GC SS_cons   ........(((((.((((((((((.(((.(((((...(((((.....))))).)))))..)))))..))).))).)))))))..)))))).....
```

Mature inferred from conservation block, no reported mature sequences.

Specie	Loci	Annotation
B. schlosseri	2	`miRNAture`
C. robusta	1	`RFAM`
C. savignyi	1	`RFAM`
C. oblonga	3	`miRNAture`
D. vexillum	1	`miRNAture`
H. aurantium	1	`miRNAture`
H. roretzi	1	`miRNAture`
P. fumigata	2	`miRNAture`
P. mammilata	1	`miRNAture`
Total	13	

Using `miRNAture` was possible to detect 11 new candidates on *Tunicata* clade.

What about the position of their mature sequences?

Annotation and expansion of a *canonical* miRNA

Annotation mature sequences mir-1497 (RF00953)

```
# STOCKHOLM 1.0
B1566810522#1  cugguauaAGCGUAACCUCUUGUAAAUCUCCAACACUAAGCUUGGUuacagcuc-UGUUGAAGAAUUGCAGGUGGUAGGUGCUaucca
B1566808146#0  uggcuguuAGCGCAACCACUUGUGAAUCUCCAACACUCGGCUUUGUccacgacuuUGUUGAAGAAUUGCAGGUGGUAGGUGCUuucaa
#=GC SS_cons   ........(((((.((((((((((.(((.((((((...(((((.....))))))..))))))..))).))))))))))..))))).....
```

Mature inferred from conservation block, no reported mature sequences.

Inclusion and validation of detected miRNAs by *miRNAture*

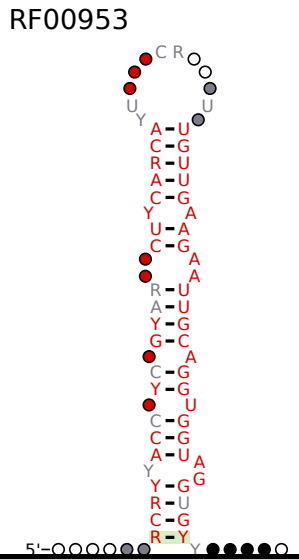
```
# STOCKHOLM 1.0
H1568379444#11  CUC------CCGCAUUACCACCUGUACAUCUCCAGCAUUU--CUCG---G-------UUUGUUGAAGAAUUGCAGGUGGUAGGUGCCGGAGA----
H1568379806#10  uauu-------ACGUCACCUCCGGCAACGCUUCAACAUUA--UACGcuuG-------UUUGUUGAAGAAUUGCAGGUGGUAGGUGUagauac----
H1568383613#13  ----------CGCAUCACCACCUGUACAUCUCCAGCAUUU--GUugaa-G-------UUUGUUGAAGAAUUGCAGGUGGUAGGUGC----------
H1568376381#12  CUC------CCGCAUUACCACCUGUACAUCUCCAGCAUUU--CUCG---G-------UUUGUUGAAGAAUUGCAGGUGGUAGGUGCCGGAGA----
B1566810522#1   GCUGG-UAUAAGCGUAACCUCUUGUAAAUCUCCAACACUA--AGCUUGGUUACAGCUC-UGUUGAAGAAUUGCAGGUGGUAGGUGCUAUCC-----
B1566808146#0   -CUGGCUGUUAGCGCAACCACUUGUGAAUCUCCAACACUC--GGCUUUGUCCACGACUUUGUUGAAGAAUUGCAGGUGGUAGGUGCUUUCA-----
H1568382511#3   UUG------AAACAUUACCUCCGGCAGCGCUUCAGCAUUU--GCCAGU---------UUUGUUGAAGAAUUGCAGGUGGUAGGUGUUUAUGC----
H1568384316#2   UUG------AAACAUUACCUCCGGCAGCGCUUCAGCAUUU--GUCAGU---------UUUGUUGAAGAAUUGCAGGUGGUAGGUGUUUAUGC----
H1568377768#5   GCUGG-UAUAAGCGUAACCUCUUGUAAAUCUCCAACACUA--AGCUUGGUUACAGCUC-UGUUGAAGAAUUGCAGGUGGUAGGUGCUAUCCa----
H1568375669#4   -CUGGCUGUUAGCGCAACCACUUGUGAAUCUCCAACACUC--GGCUUUGUCCACGACUUUGUUGAAGAAUUGCAGGUGGUAGGUGCUUUCAagccg
H1568380600#7   ------UCUGUGCGUUACCAUCAGCAACCCUUCAGCAUUG--AACG---G-------UUUGUUGAAGAAUUGCAGGUGGUAGGCGCCACAAA----
H1568375805#6   ------UCUGUGCGUUACCAUCAGCAACCCUUCAGCAUUG--AACG---G-------UUUGUUGAAGAAUUGCAGGUGGUAGGCGCCACAAA----
H1568379608#9   c-------UCAGCACCACUAUCGGCAAGUCUUCAGCAagauaCUUG---G-------UAUGUUGAAGAAUUGCAGGUGGUAGGUGCUUUgu-----
H1568383969#8   ------UCUGUGCGUUACCAUCAGCAACCCUUCAGCAUUG--AACG---G-------UUUGUUGAAGAAUUGCAGGUGGUAGGCGCCACAAA----
#=GC SS_cons    ..........((((.(((.((.(((((..((.(((((.....................)))))..))..)))))..)).))))..))))..........
```

Annotation and expansion of a *canonical* miRNA

Inclusion and validation of detected miRNAs by *miRNAture*

```
# STOCKHOLM 1.0
H1568379444#11  CUC------CCGCAUUACCACCUGUACAUCUCCAGCAUUU--CUCG---G-------UUUGUUGAAGAAUUGCAGGUGGUAGGUGCCGGAGA----
H1568379806#10  uauu-------ACGUCACCUCCGGCAACGCUUCAACAUUA--UACGcuuG-------UUUGUUGAAGAAUUGCAGGUGGUAGGUGUagauac----
H1568383613#13  ----------CGCAUCACCACCUGUACAUCUCCAGCAUUU--GUugaa-G-------UUUGUUGAAGAAUUGCAGGUGGUAGGUGC----------
H1568376381#12  CUC------CCGCAUUACCACCUGUACAUCUCCAGCAUUU--CUCG---G-------UUUGUUGAAGAAUUGCAGGUGGUAGGUGCCGGAGA----
B1566810522#1   GCUGG-UAUAAGCGUAACCUCUUGUAAAUCUCCAACACUA--AGCUUGGUUACAGCUC-UGUUGAAGAAUUGCAGGUGGUAGGUGCUAUCC-----
B1566808146#0   -CUGGCUGUUAGCGCAACCACUUGUGAAUCUCCAACACUC--GGCUUUGUCCACGACUUUGUUGAAGAAUUGCAGGUGGUAGGUGCUUUCA-----
H1568382511#3   UUG------AAACAUUACCUCCGGCAGCGCUUCAGCAUUU--GCCAGU---------UUUGUUGAAGAAUUGCAGGUGGUAGGUGUUUAUGC----
H1568384316#2   UUG------AAACAUUACCUCCGGCAGCGCUUCAGCAUUU--GUCAGU---------UUUGUUGAAGAAUUGCAGGUGGUAGGUGUUUAUGC----
H1568377768#5   GCUGG-UAUAAGCGUAACCUCUUGUAAAUCUCCAACACUA--AGCUUGGUUACAGCUC-UGUUGAAGAAUUGCAGGUGGUAGGUGCUAUCCa----
H1568375669#4   -CUGGCUGUUAGCGCAACCACUUGUGAAUCUCCAACACUC--GGCUUUGUCCACGACUUUGUUGAAGAAUUGCAGGUGGUAGGUGCUUUCAagccg
H1568380600#7   ------UCUGUGCGUUACCAUCAGCAACCCUUCAGCAUUG--AACG---G-------UUUGUUGAAGAAUUGCAGGUGGUAGGCGCCACAAA----
H1568375805#6   ------UCUGUGCGUUACCAUCAGCAACCCUUCAGCAUUG--AACG---G-------UUUGUUGAAGAAUUGCAGGUGGUAGGCGCCACAAA----
H1568379608#9   c-------UCAGCACCACUAUCGGCAAGUCUUCAGCAagauaCUUG---G-------UAUGUUGAAGAAUUGCAGGUGGUAGGUGCUUUgu-----
H1568383969#8   ------UCUGUGCGUUACCAUCAGCAACCCUUCAGCAUUG--AACG---G-------UUUGUUGAAGAAUUGCAGGUGGUAGGCGCCACAAA----
#=GC SS_cons    ..........((((.(((.((.(((((..((.(((((.....................)))))..))..))))).)).))))..))))..........
```

Specie	Loci	Annotation	Validated
B. schlosseri	2	miRNAture	2
C. robusta	1	RFAM	1
C. savignyi	1	RFAM	1
C. oblonga	3	miRNAture	3
D. vexillum	1	miRNAture	1
H. aurantium	1	miRNAture	0
H. roretzi	1	miRNAture	1
P. fumigata	2	miRNAture	2
P. mammilata	1	miRNAture	1
Total	13		12



Conclusions

Based on the miRNAture + MIRFix combination

- An improved miRNA annotation could be reached combininig different homology searches + SS validation. **As implemented on** `miRNAture`.
- Annotation of miR and miR* (matures) sequences has to be considered when a *canonical* miRNA is going to be annotated.
- The combination of the approaches opens the door for create new multiple alignments, new covariance models and even scale for a new way to classify miRNA families.

Thanks!, Vielen Dank!, Obrigado!, ¡Gracias!

Peter F. Stadler

Ali Yazbeck

Clara I. Bermúdez

Adriaan Gittenberger

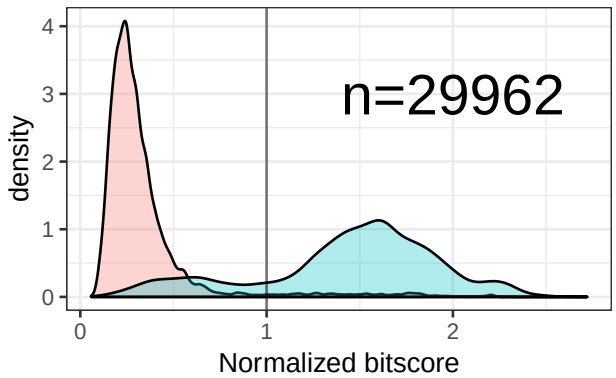
Federico D. Brown

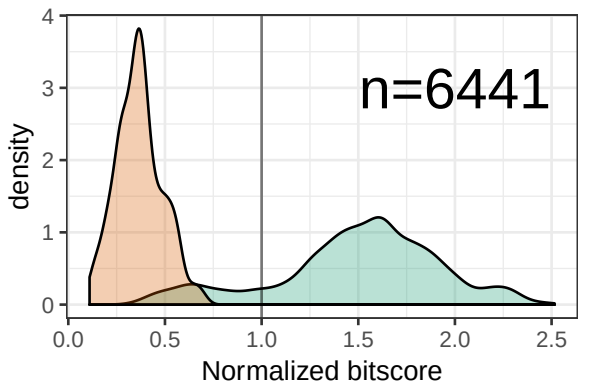
Getting DNA, RNA data

CM re-definition scores

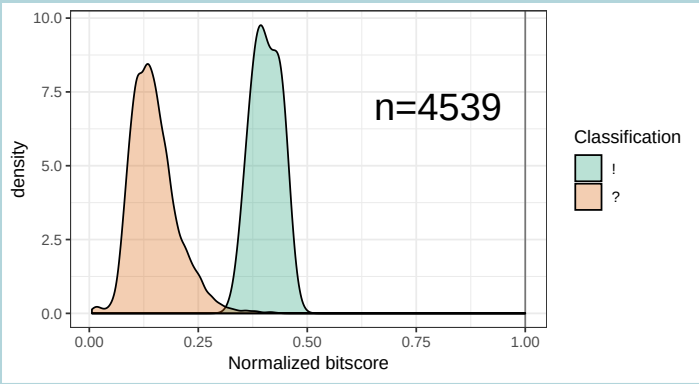
Plotting `cmseach` results

Positive

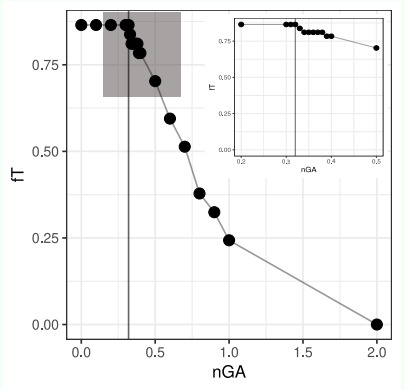




Negative



Haro data evaluated and *nGA*



$nGA = 0.32$

Assumption: Haro candidates are true candidates.