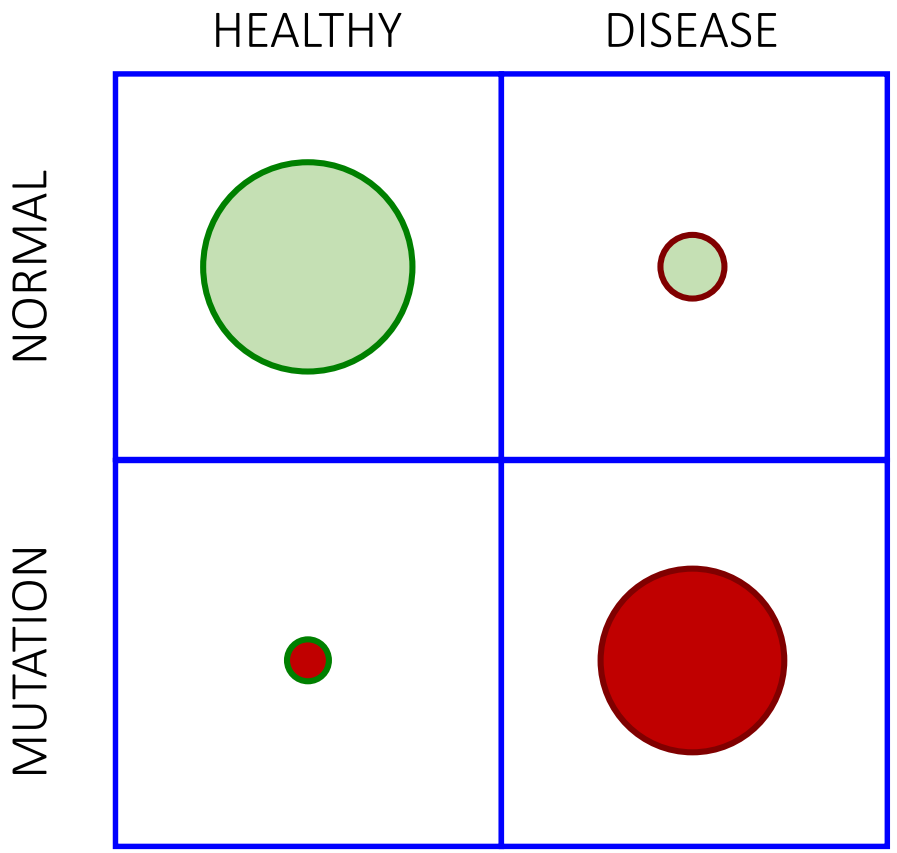
What is a miRNA?

Simon Rayner

AMG OUS/UiO

Genetics

GWAS - MONOGENIC DISORDERS



e.g. CFTRΔF508

deletion of three nucleotides spanning positions 507 and 508 of the CFTR gene on chromosome 7,

→ loss of the codon for phenylalanine (F).

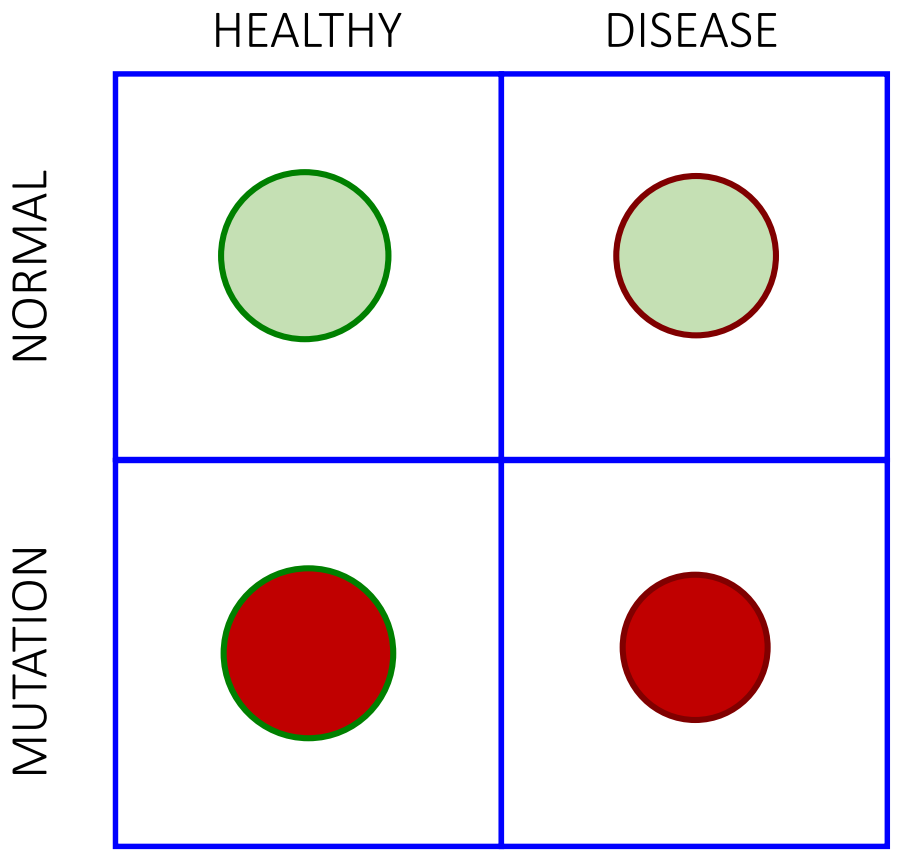
The CFTRΔF508 mutation produces an abnormal CFTR protein that cannot fold properly and which does not escape the endoplasmic reticulum for further processing.

Having two copies of this mutation is the most common cause of cystic fibrosis (CF) (2/3 CASES)

GENE PANELS

BreastNextTM gene panel- sequence 17 genes (**ATM, BARD1, BRCA1, BRCA2, BRIP1, CDH1, CHEK2, MRE11A, MUTYH, NBN, NF1, PALB2, PTEN, RAD50, RAD51C, RAD51D**, and **TP53**)

GWAS - POLYGENIC DISORDERS

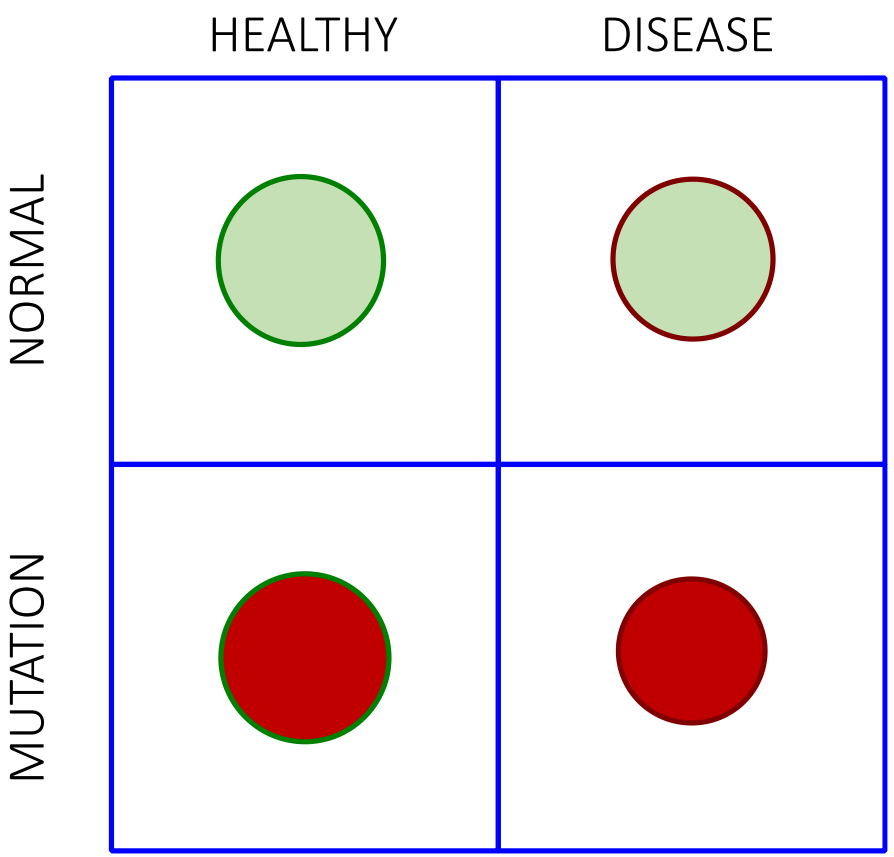


Once we start investigating multiple genes/polygenic disease, the effects can be less obvious and the analysis becomes more complicated.

If we add in the effect of a spectrum of phenotypes, then things become even more complex.

Now, we have to identifying meaningful associations across both genotype and phenotype spectrums.

GWAS - POLYGENIC DISORDERS



However, these studies still focus on the coding regions of the genome, which corresponds to about 2%

Non-coding genome

- Under slightly more relaxed selection pressure
- Many different features identified
- Regulatory roles

But where to start?

miRNAs

Identification of Novel Genes Coding for Small Expressed RNAs

Mariana Lagos-Quintana, Reinhard Rauhut, Winfried Lendeckel, Thomas Tuschl*

In *Caenorhabditis elegans*, *lin-4* and *let-7* encode 22- and 21-nucleotide (nt) RNAs, respectively, which function as key regulators of developmental timing. Because the appearance of these short RNAs is regulated during development, they are also referred to as small temporal RNAs (stRNAs). We show that many 21- and 22-nt expressed RNAs, termed microRNAs, exist in invertebrates and vertebrates and that some of these novel RNAs, similar to *let-7* stRNA, are highly conserved. This suggests that sequence-specific, posttranscriptional regulatory mechanisms mediated by small RNAs are more general than previously appreciated.

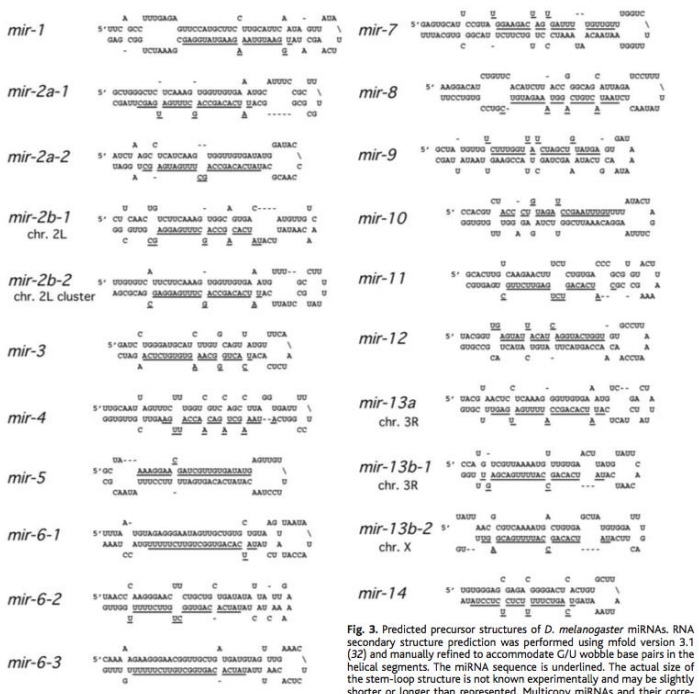


Fig. 3. Predicted precursor structures of *D. melanogaster* miRNAs. RNA secondary structure prediction was performed using mfold version 3.1 (32) and manually refined to accommodate C/U wobble base pairs in the helical segments. The miRNA sequence is underlined. The actual size of the stem-loop structure is not known experimentally and may be slightly shorter or longer than represented. Multicopy miRNAs and their corresponding precursor structures are also shown.

The first miRNAs were well characterized

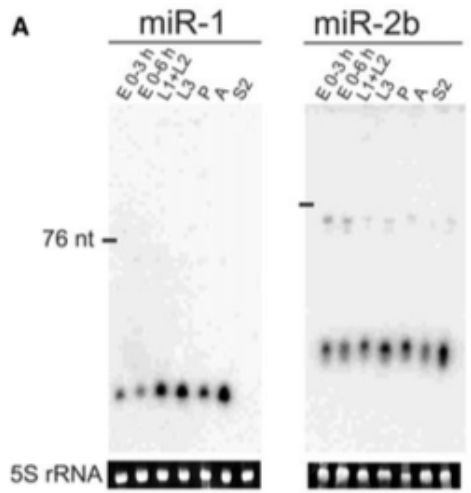
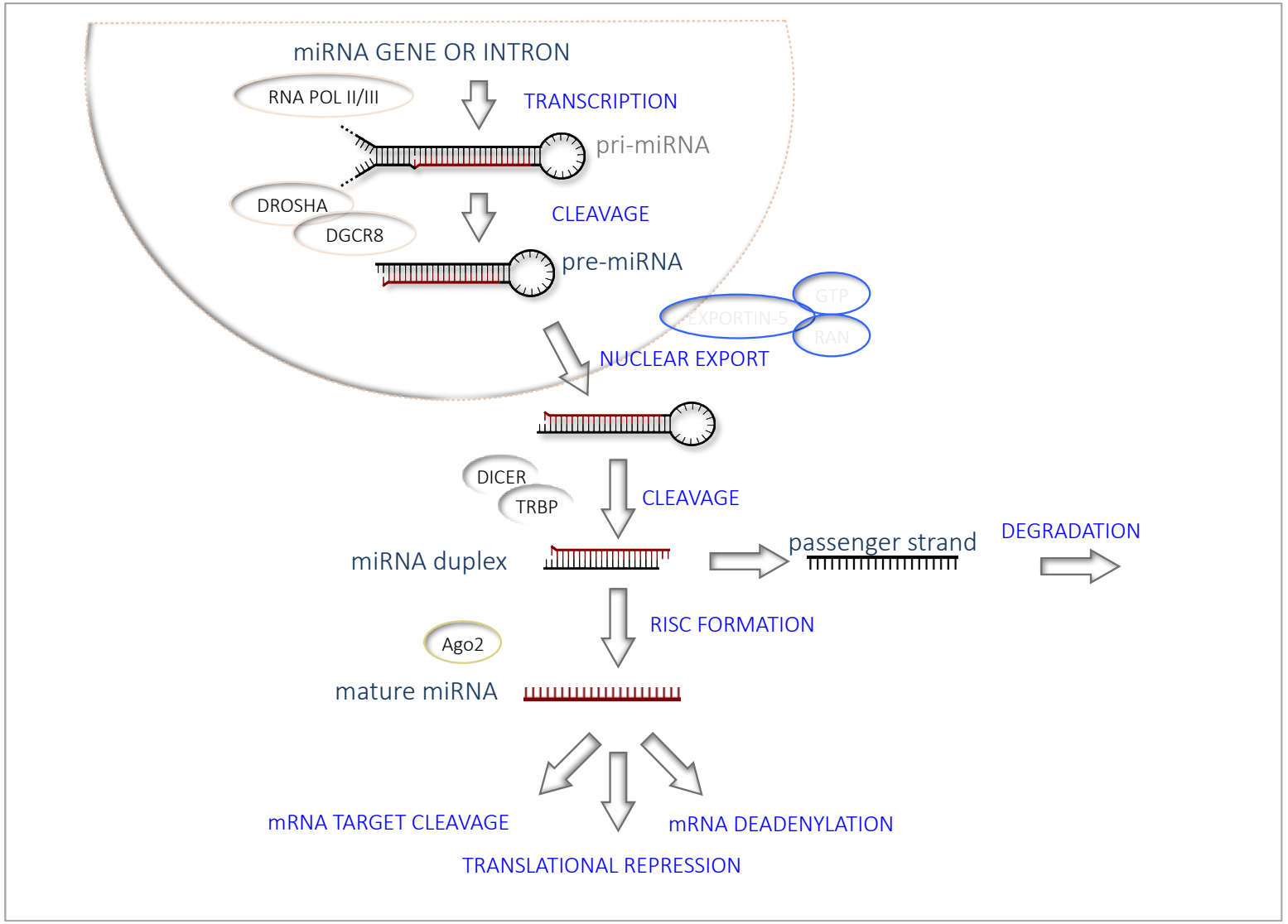
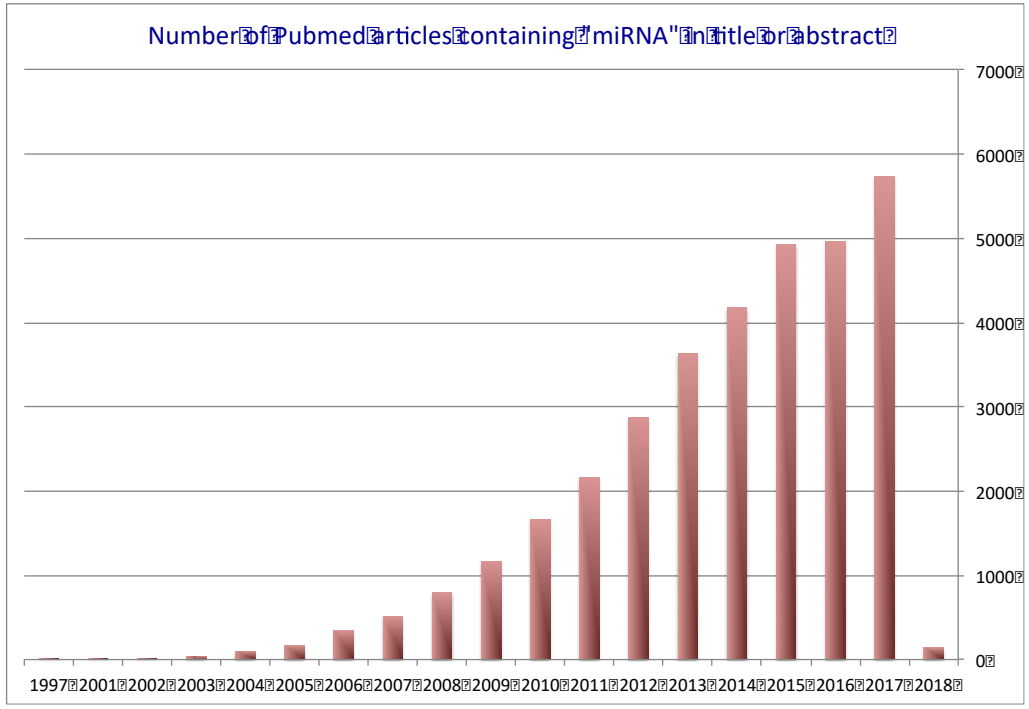


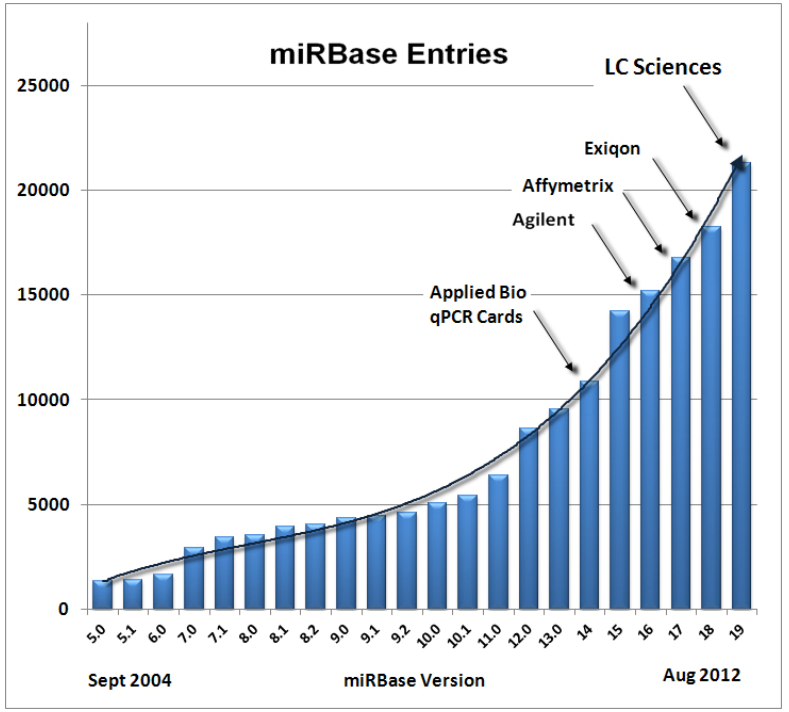
Fig. 1. Expression of miRNAs. Representative examples of Northern blot analysis are depicted (*21*). The position of 76-nt val-tRNA is indicated on the blots; 5S rRNA serves as a loading control. (**A**) Northern blots of total RNA isolated from staged populations of *D. melanogaster*, probed for the indicated miRNA. E, embryo; L, larval stage; P, pupa; A, adult; S2, Schneider-2 cells. (**B**) Northern blots of total RNA isolated from HeLa cells, mouse kidneys, adult zebrafish, frog ovaries, and S2 cells, probed for the indicated miRNA.



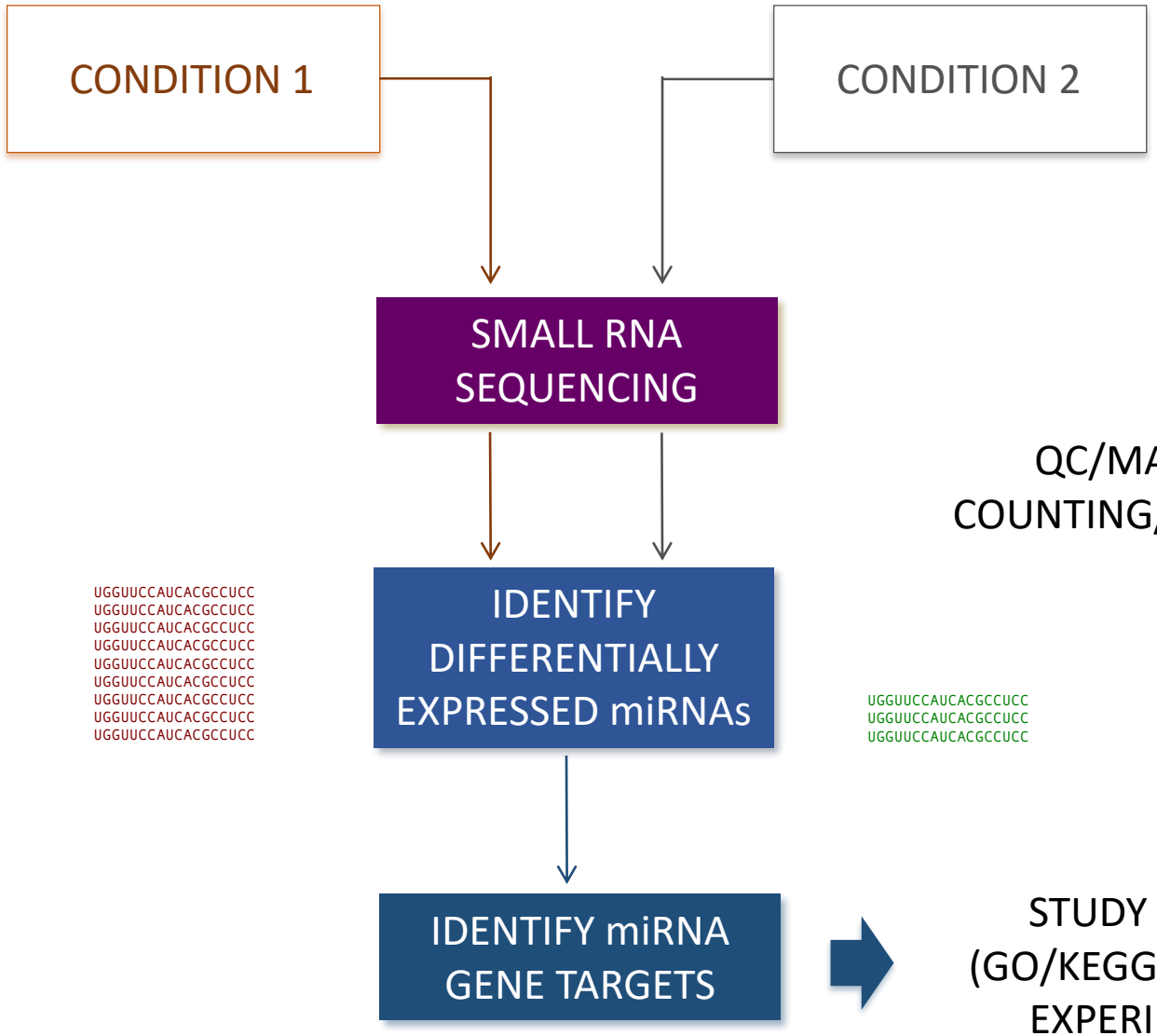
Determination of miRNA generation

miRNAs are a very popular research topic

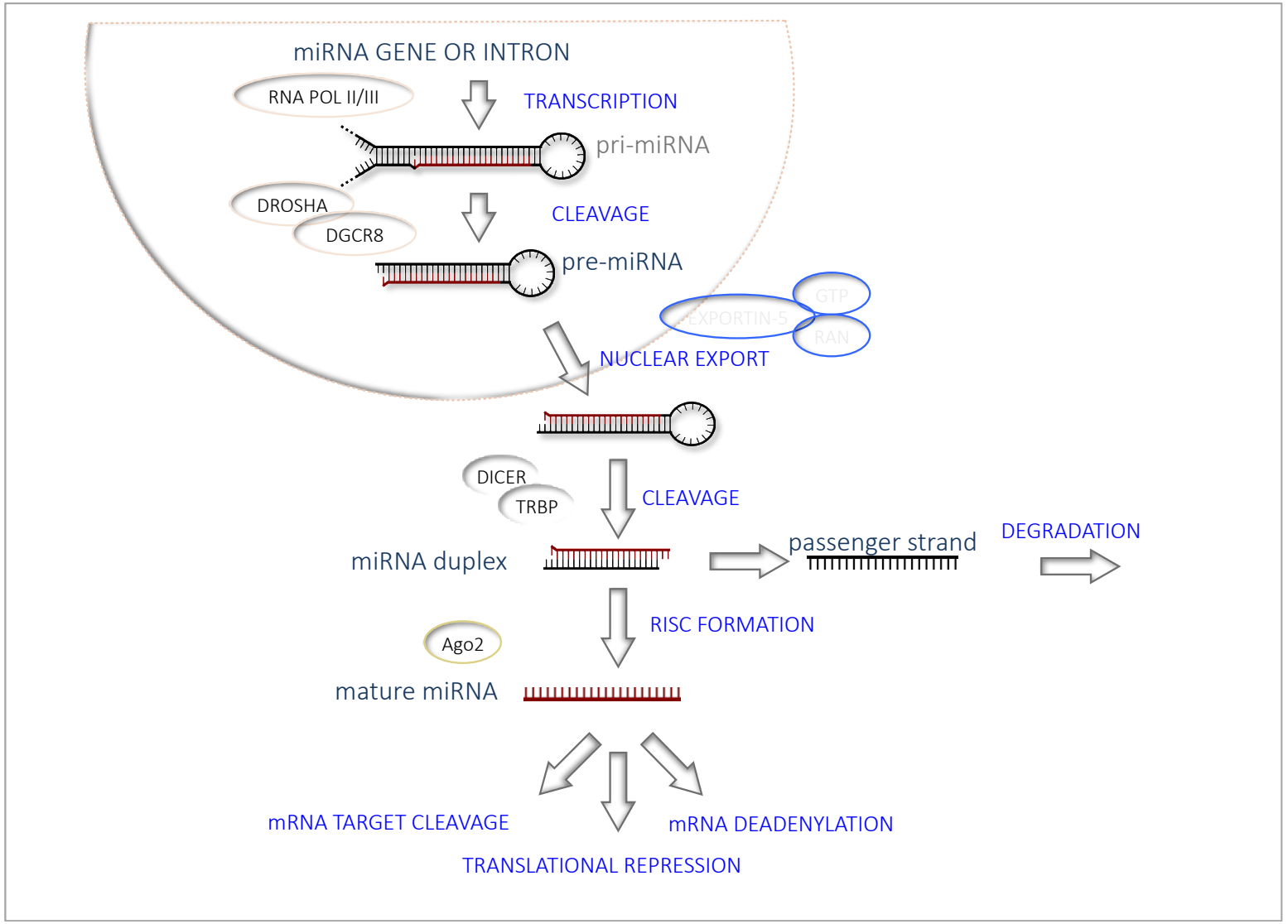




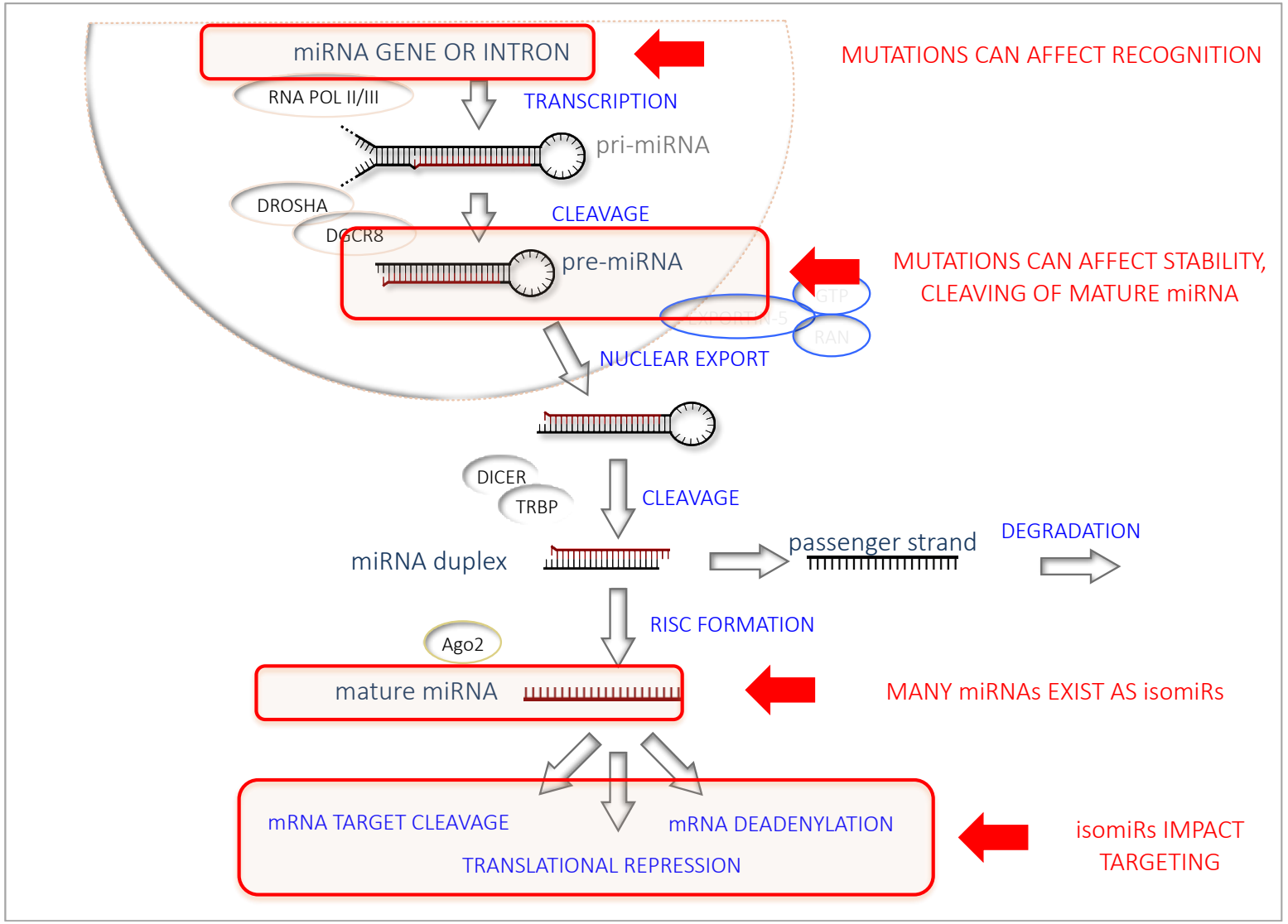
Discovery of new miRNAs is highly dependent on NGS



miRNAs are most commonly studied using NGS



Determination of miRNA generation



There are many points of deviation

ANNOTATION ISOFORMS

POPULATION TARGETING

HOW DO EACH OF THESE FACTORS AFFECT A miRNA STUDY?

ANNOTATION

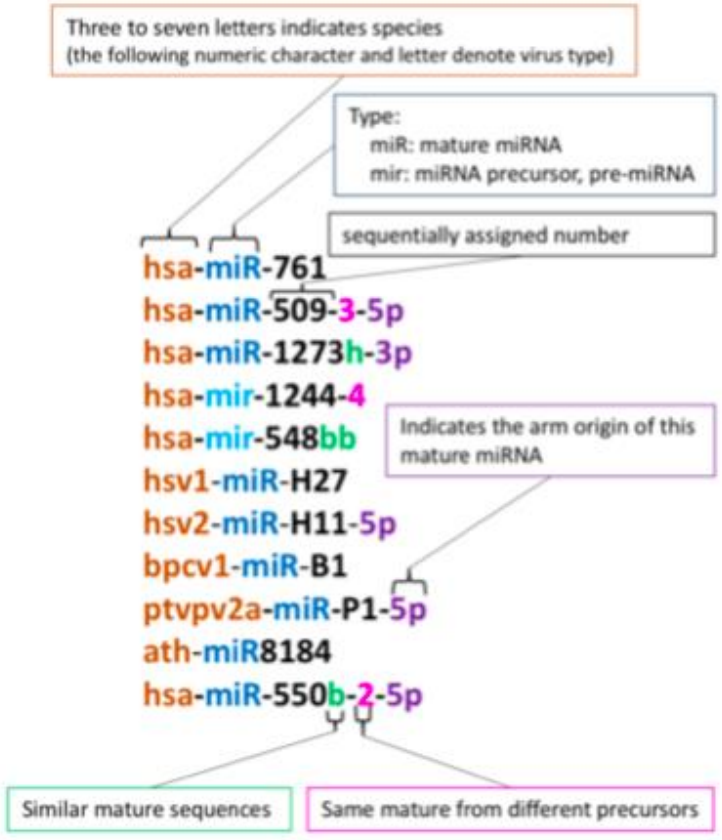
ISOFORMS

POPULATION

TARGETING

miRBASE ANNOTATION

```
chr11    .        miRNA_primary_transcript  2134134  2134209  .        -        .        Name=hsa-mir-483
chr11    .        miRNA    2134181  2134202  .        -        .        Name=hsa-miR-483-
5p;Derives_from=MI0002467
chr11    .        miRNA    2134142  2134162  .        -        .        Name=hsa-miR-483-
3p;Derives_from=MI0002467
```

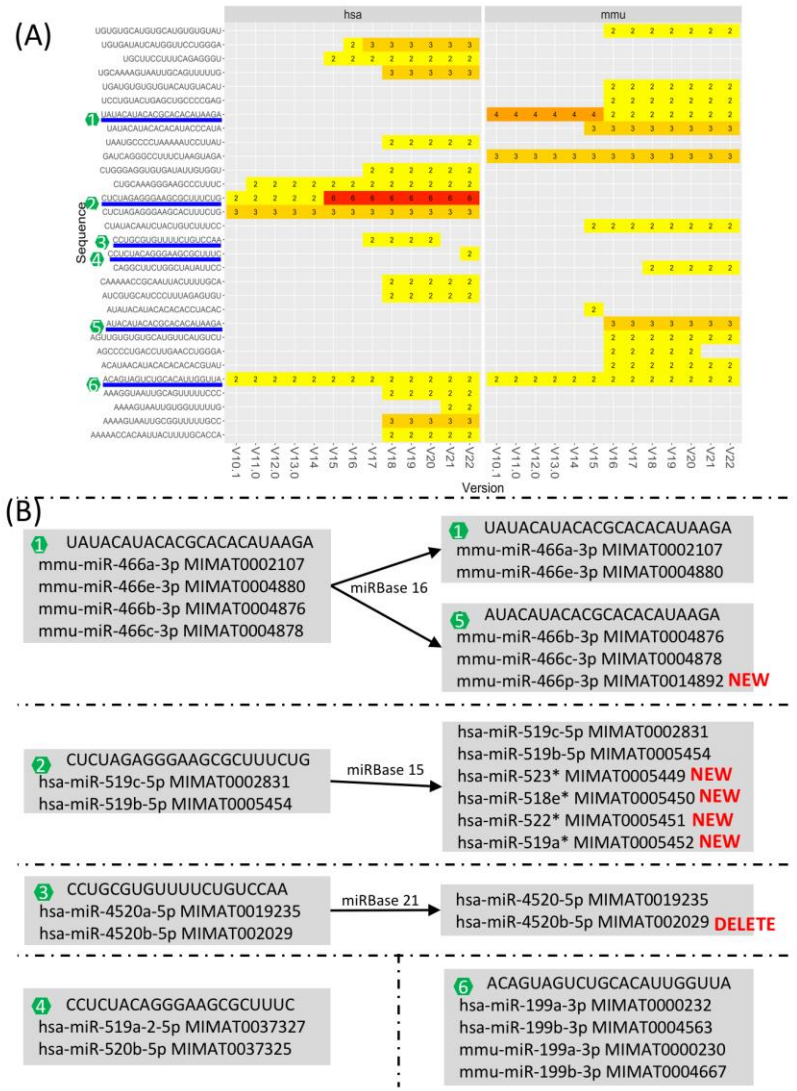


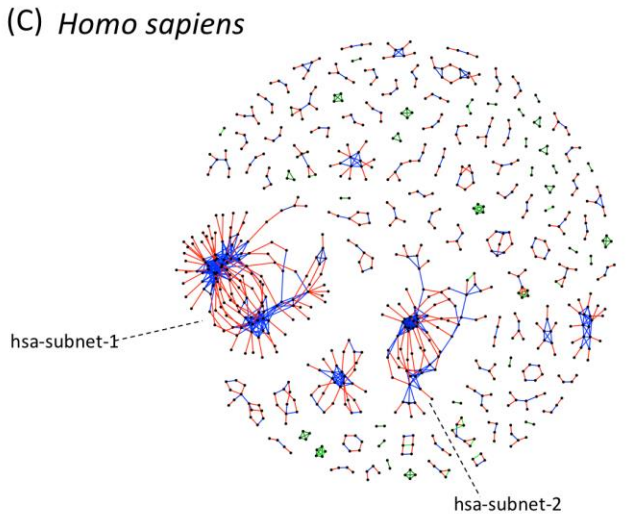
THE NAME TELLS US A LOT ABOUT A miRNA

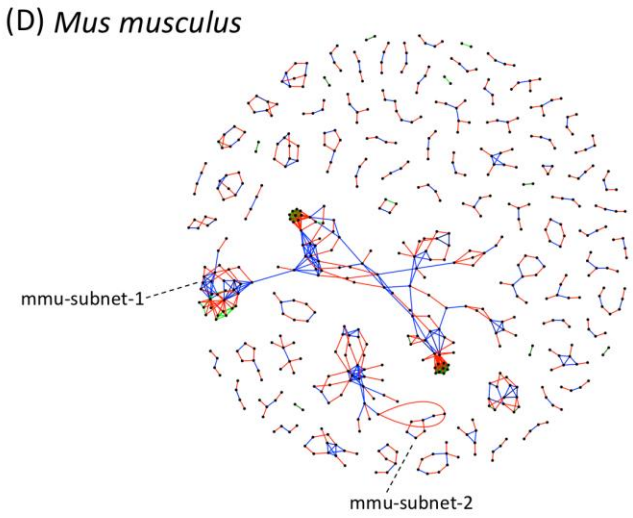
SAME miRNA SEQUENCE EXISTS IN DIFFERENT ENTRIES

PROBLEMATIC IN NGS OR MICROARRAY STUDIES

miRBASE ANNOTATION : HIGH SIMILARITY ENTRIES







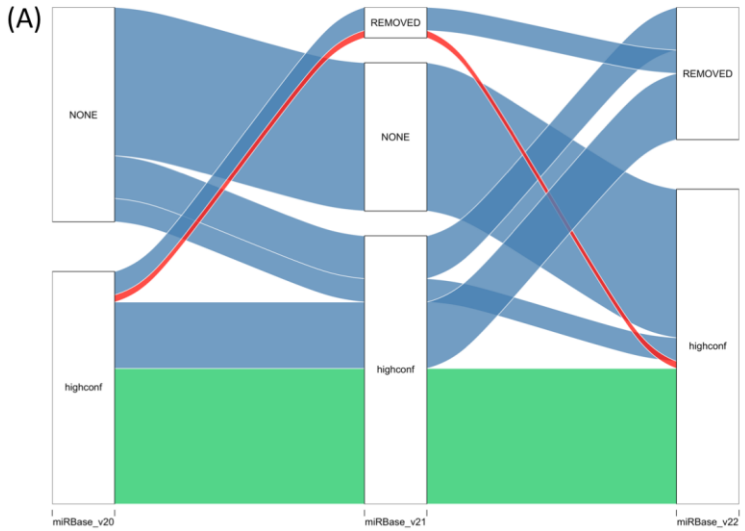
EVERY RELEASE HAS DUPLICATE SEQUENCES

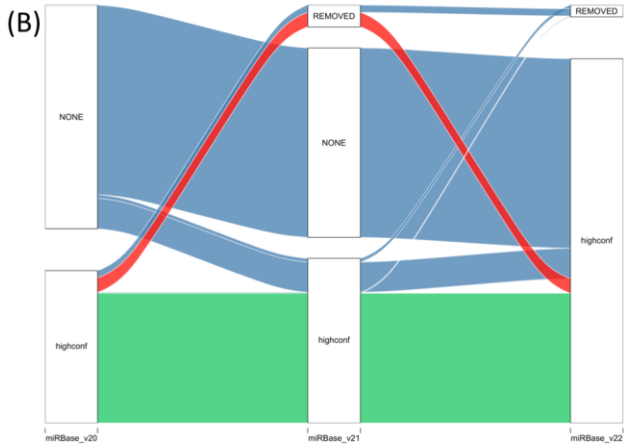
MORE WITH EVERY RELEASE

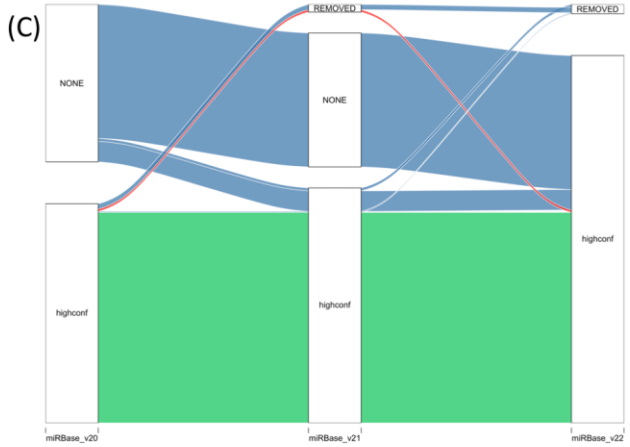
ALSO, HIGHLY SIMILAR SEQUENCES

COMPLICATES NGS AND MICROARRAY STUDIES

miRBASE ANNOTATION : HIGH CONFIDENCE SETS







ADDED → REMOVED → ADDED

PRESENT IN ALL THREE SETS

TO ADDRESS SOME OF THESE ANNOTATION PROBLEMS miRBASE RELEASED A HIGH CONFIDENCE ANNOTATION SET

3298 DISTINCT HAIRPIN PRECURSORS ACROSS THE THREE RELEASES, ONLY 925 (231 HUMAN HAIRPINS) ARE PRESENT ACROSS ALL THREE RELEASES

In Press: RNA Biology

ANNOTATION ISOFORMS

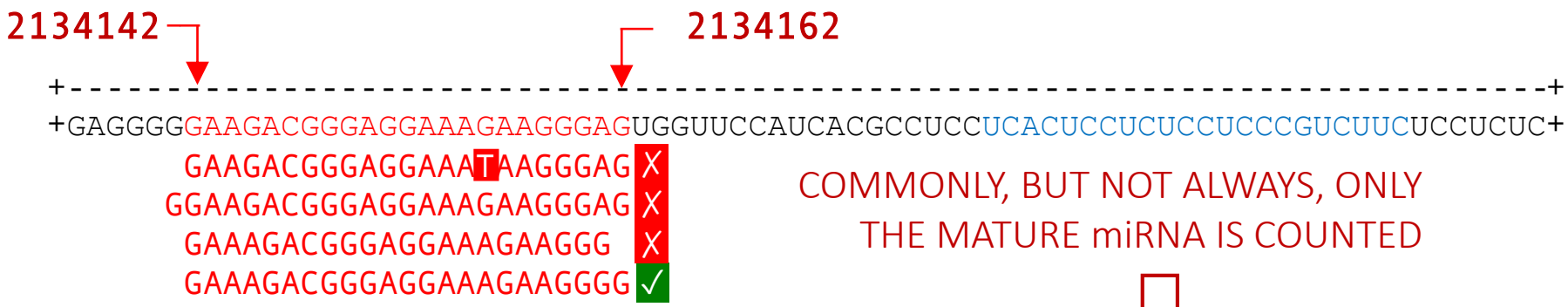
POPULATION TARGETING

isomiRs

```
chr11   .   miRNA_primary_transcript   2134134   2134209   .   -   .
        ID=MI0002467;Alias=MI0002467;Name=hsa-mir-483
chr11   .   miRNA   2134181   2134202   .   -   .
        ID=MIMAT0004761;Alias=MIMAT0004761;Name=hsa-miR-483-5p;Derives_from=MI0002467
chr11   .   miRNA   2134142   2134162   .   -   .
        ID=MIMAT0002173;Alias=MIMAT0002173;Name=hsa-miR-483-3p;Derives_from=MI0002467
```

```
chr11   .   miRNA_primary_transcript   2134134   2134209   .   -
chr11   .   miRNA   2134181   2134202   .   -   .
chr11   .   miRNA   2134142   2134162   .   -   .
```

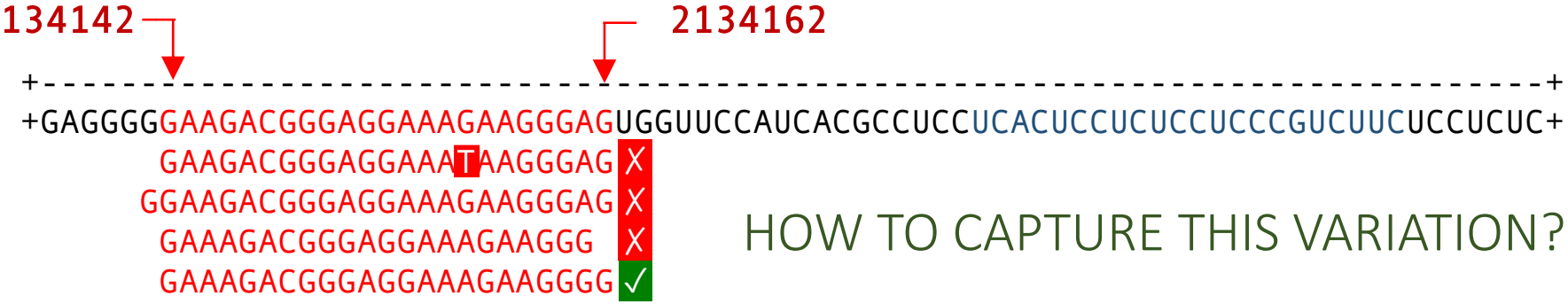
IN REALITY, THERE CAN BE MULTIPLE ISOFORMS (isomiRs) OF AN miRNA



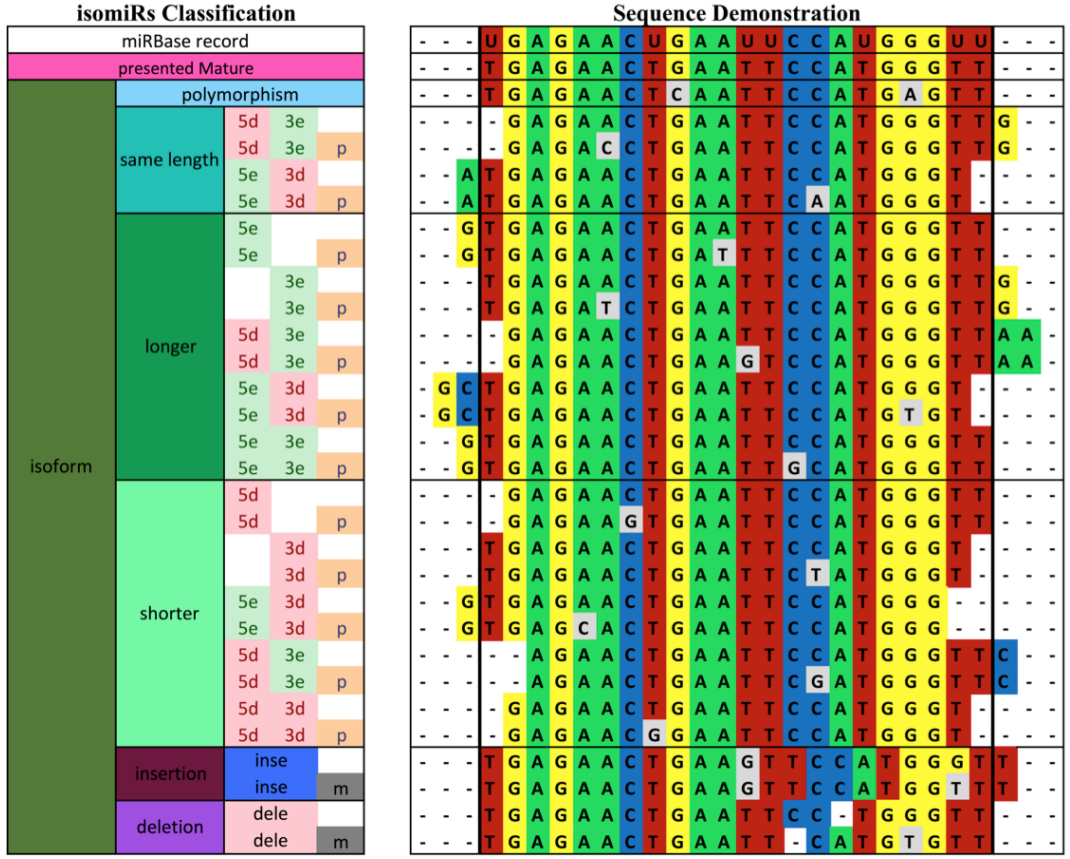
COMMONLY, BUT NOT ALWAYS, ONLY THE MATURE miRNA IS COUNTED

THE PRESENCE OF isomiRs COMPLICATES THE TARGETING PROCESS (SHIFTED SEED REGION)

isomiRs



HOW TO CAPTURE THIS VARIATION?



isomiR nomenclature	Description
mature	isomiR shares the identical sequence, same start and end position with the reference miRNA in miRBase.
poly	isomiR has same start and end position with mature miRNA, but has one or more internal nucleotide substitutions. (For the isomiRs groups defined below that contain additional modifications, this is annotated with a p).
sl-5e-3d sl-5e-3d-p	5' shift, but same read length as mature miRNA, (i.e. isomiR contains both a 5' extension and 3' deletion).

WHAT HAPPENS IF WE USE THIS NOMENCLATURE IN AN ANALYSIS?

Analysis using all different isomiR classifications

Same miRNAs
(> 1000 reads in at least one sample).

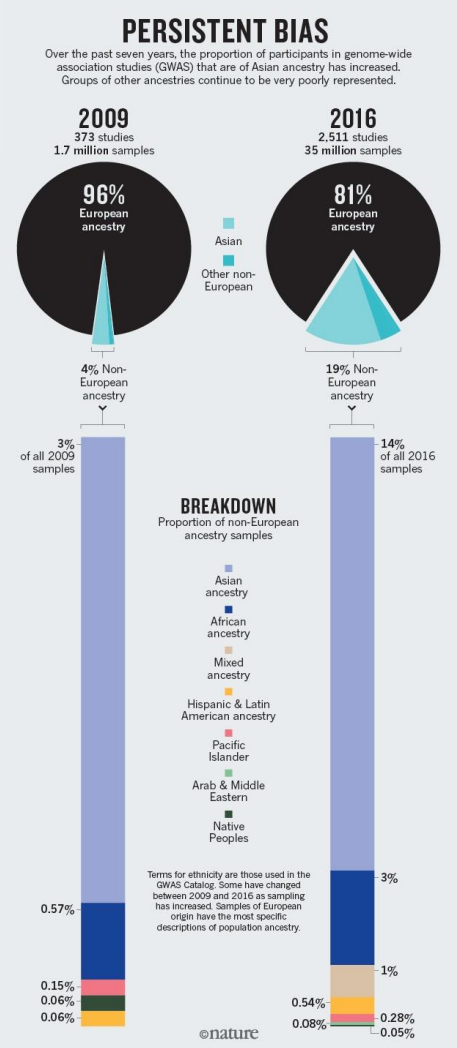
Many different types of isomiRs exist and vary between conditions

The reality is that things are more complicated

ANNOTATION ISOFORMS

POPULATION TARGETING

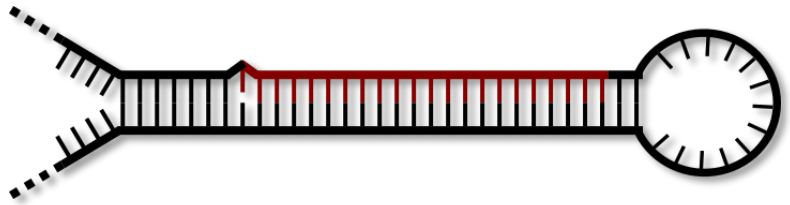
POPULATION VARIATION

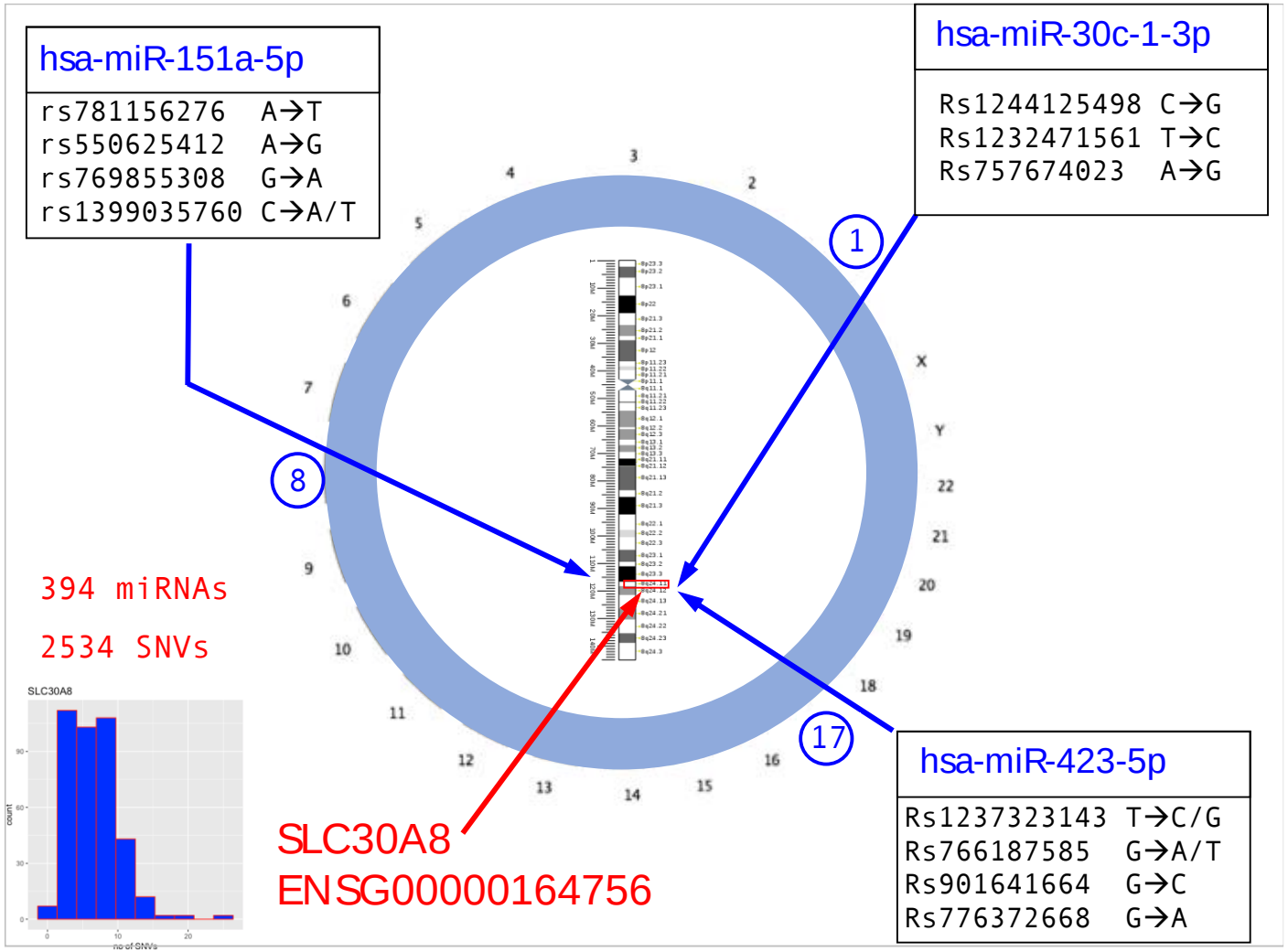


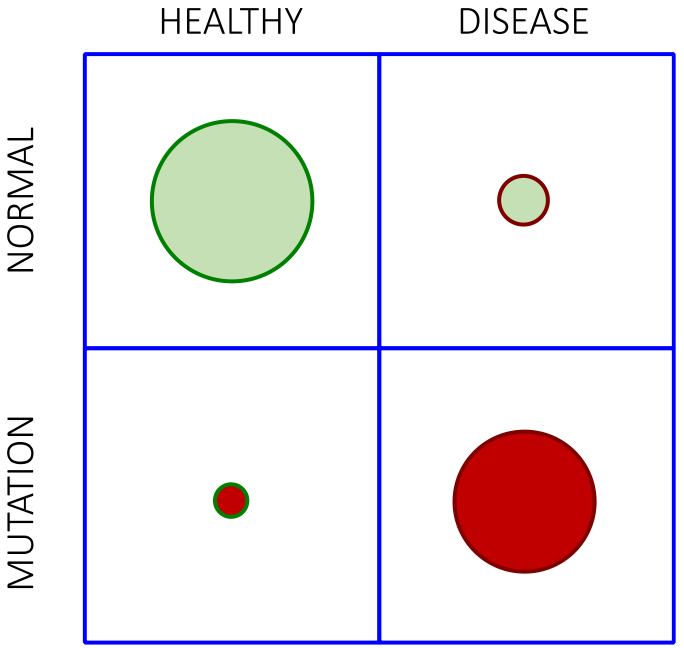
POPULATION BIAS EXISTS IN GWAS STUDIES

miRBase MAKES NO ACCOMMODATION FOR POPULATION VARIATION

FOR EXAMPLE, DO POPULATION SPECIFIC MUTATIONS LEAD TO ATTENUATION OF SPECIFIC miRNAs?




hsa-miR-151a-5p
rs781156276 A→T
rs550625412 A→G
rs769855308 G→A
rs1399035760 C→A/T
hsa-miR-30c-1-3p
Rs1244125498 C→G
Rs1232471561 T→C
Rs757674023 A→G
1
8
17
394 miRNAs
2534 SNVs
SLC30A8
SLC30A8
ENSG00000164756
hsa-miR-423-5p
Rs1237323143 T→C/G
Rs766187585 G→A/T
Rs901641664 G→C
Rs776372668 G→A


HEALTHY
DISEASE
NORMAL
MUTATION

ANNOTATION ISOFORMS

POPULATION TARGETING

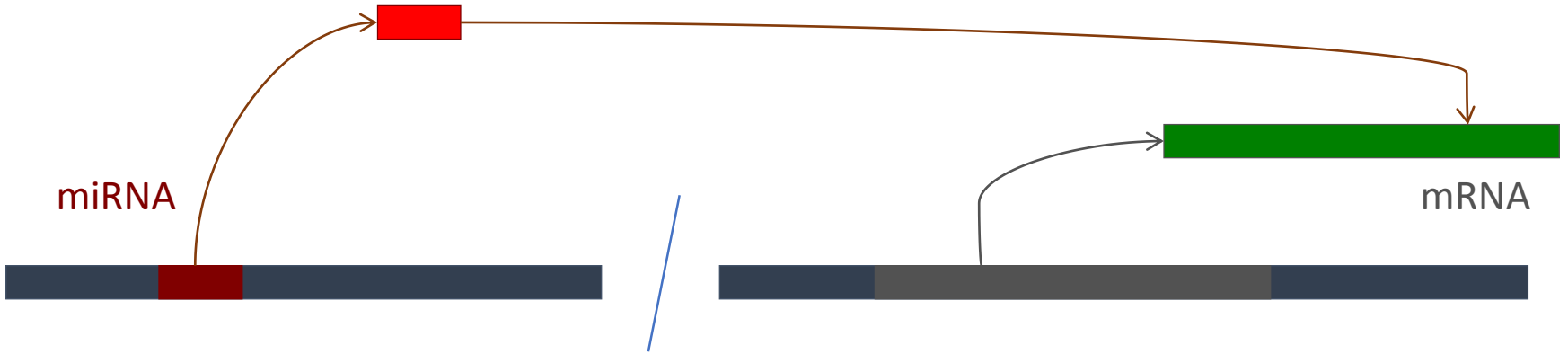
miRNA targeting

COMMON PERCEPTION: We know all there is to know

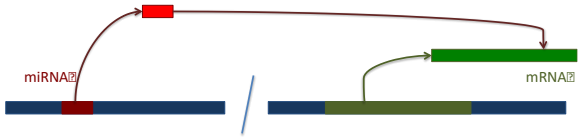
CONSEQUENCE: Not very trendy

REALITY: Poorly understood

KEY ASSUMPTION IN MOST miRNA STUDIES:
ONE TO ONE MAPPING BETWEEN miRNA and mRNA



miRNA targeting



1 to 1 mapping is a Gross oversimplification

One-to-Many / Many-to-One mapping between miRNA and mRNA

COMPLEX miRNA-mRNA TARGET NETWORKS

YIELD INSIGHT INTO POPULATION VARIATION AND DISEASE

There are many different prediction tools already…

FEATURES USED IN miRNA TARGET PREDICTION

Tool name	Seed match	Conservation	Free energy	Site accessibility	Target-site abundance	Machine learning	References
miRanda	X	X	X				Enright et al., 2003; John et al., 2004
miRanda-mirSVR	X	X	X	X		X	Betel et al., 2010
TargetScan	X	X					Lewis et al., 2005; Grimson et al., 2007; Friedman et al., 2009; Garcia et al., 2011
DIANA-microT-CDS	X	X	X	X	X	X	Maragkakis et al., 2009; Reczko et al., 2012; Paraskevopoulou et al., 2013
MirTarget2	X	X	X	X		X	Wang, 2008; Wang and El Naqa, 2008
RNA22-GUI	X		X				Hofacker et al., 1994; Miranda et al., 2006; Loher and Rigoutsos, 2012
TargetMiner	X	X	X	X	X	X	Bandyopadhyay and Mitra, 2009
SVMicrO	X	X	X	X	X	X	Liu et al., 2010
PITA	X	X	X	X	X		Kertesz et al., 2007
RNAhybrid	X		X		X		Rehmsmeier et al., 2004; Kruger and Rehmsmeier, 2006

TOOL AVAILABILITY AND USER FEATURES

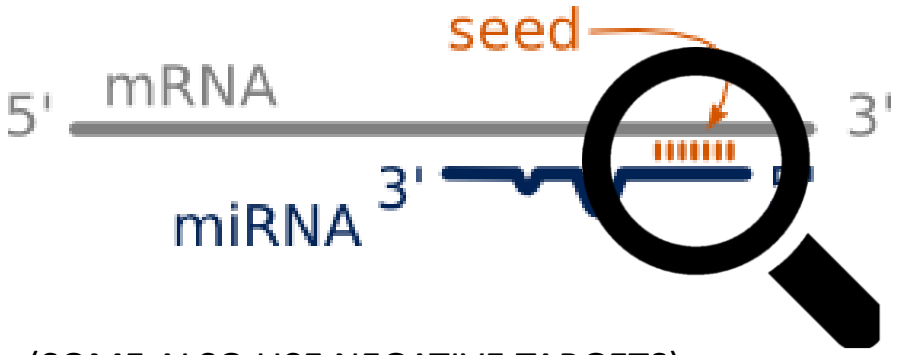
Tool name	Website	Online use	Source code	User adjustability	User-supplied data required	User level
miRanda	http://www.microrna.org/		X	X	Sequences	Advanced
miRanda-mirSVR	http://www.microrna.org/	X				All
TargetScan	http://www.targetscan.org	X				All
DIANA-microT-CDS	http://www.microrna.gr/microT-CDS	X				All
MirTarget2	http://mirdb.org	X		X		All
RNA22-GUI	https://cm.jefferson.edu/rna22v1.0/	X				Intermediate
TargetMiner	http://www.isical.ac.in/~bioinfo_miu/targetminer20.htm	X	X		Input file	Intermediate
SVMicrO	http://compgenomics.utsa.edu/svmicro.html	X	X		Sequences	Expert
PITA	http://genie.weizmann.ac.il/pubs/mir07/	X	X	X		All
RNAhybrid	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	X	X	X	Sequences	Advanced

All reviewed tools are freely available for academic use. All tools are actively maintained with updates in the past 5 years, with the exception of PITA and RNAhybrid.

doi.org/10.3389/fgene.2014.00023

miRNA targeting

KNOWN TARGETS:



(SOME ALSO USE NEGATIVE TARGETS)

EXTRACT HUMAN CRAFTED FEATURES:

E.G.

- SIZE OF SEED REGION
- MISMATCHES
- TYPE OF PAIRS
- FREE ENERGY
- …

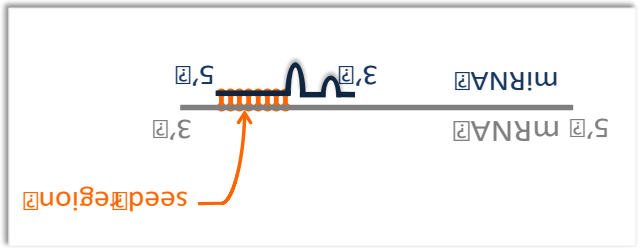
RULE BASED MODEL

```
e.g.
IF seed region > 7
+ IF GC pairs > X
+ IF free energy < Z
THIS IS A TARGET
```

MACHINE LEARNING MODEL

```
e.g.
SUPPORT VECTOR MACHINE
```

miRNA targeting



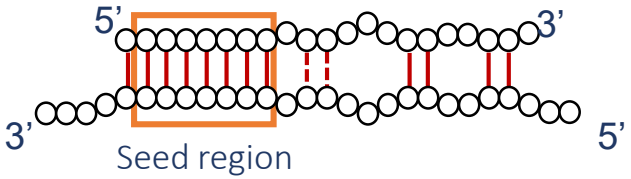
Seed region:

- First 6-8 nucleotides at the miRNA 5’ starting at nt 1 or 2

Types of binding sites:

Canonical

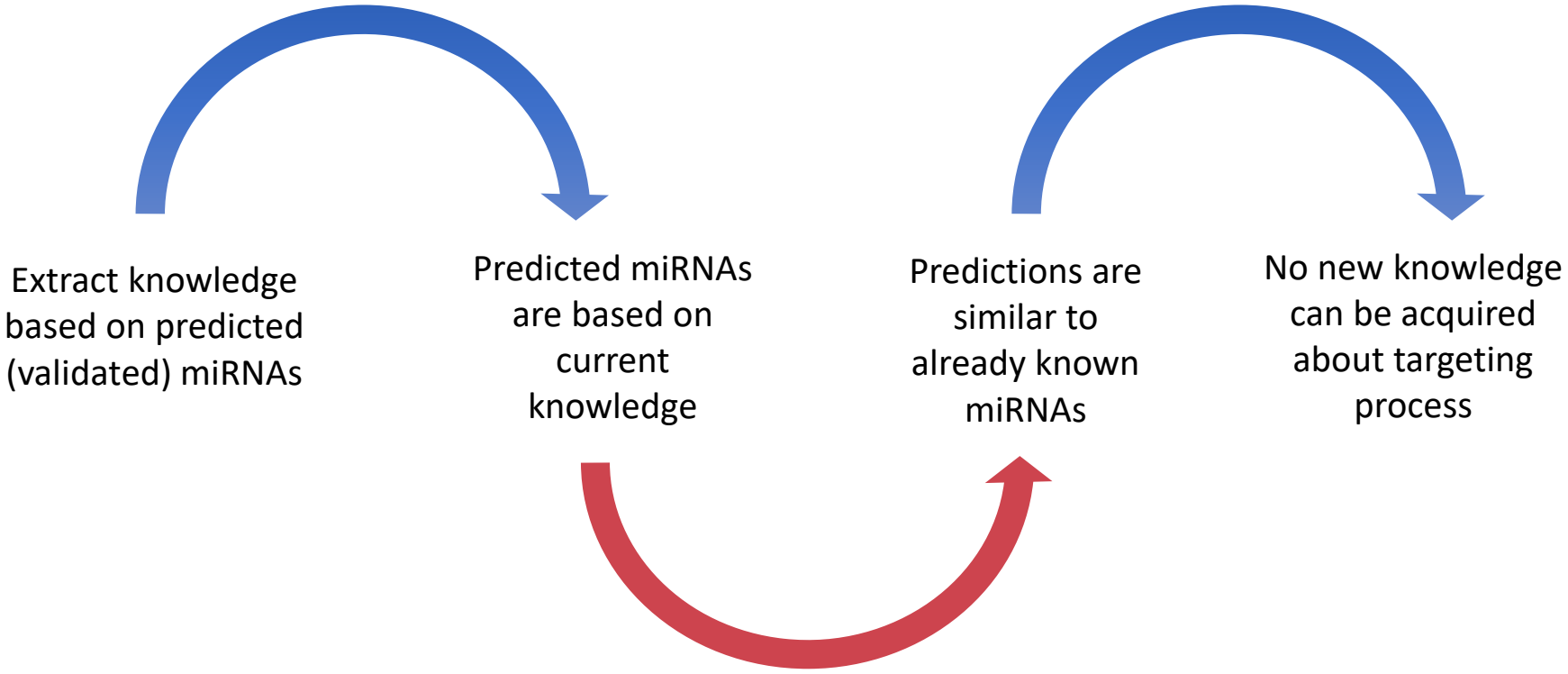
- Perfect pairing in seed region
  (6mer, 7mer, 8mer…)
- May have additional pairing



Non-canonical

- Imperfect pairing
  (wobbles, gaps, bulges…)
- Requires additional pairing

miRNA targeting



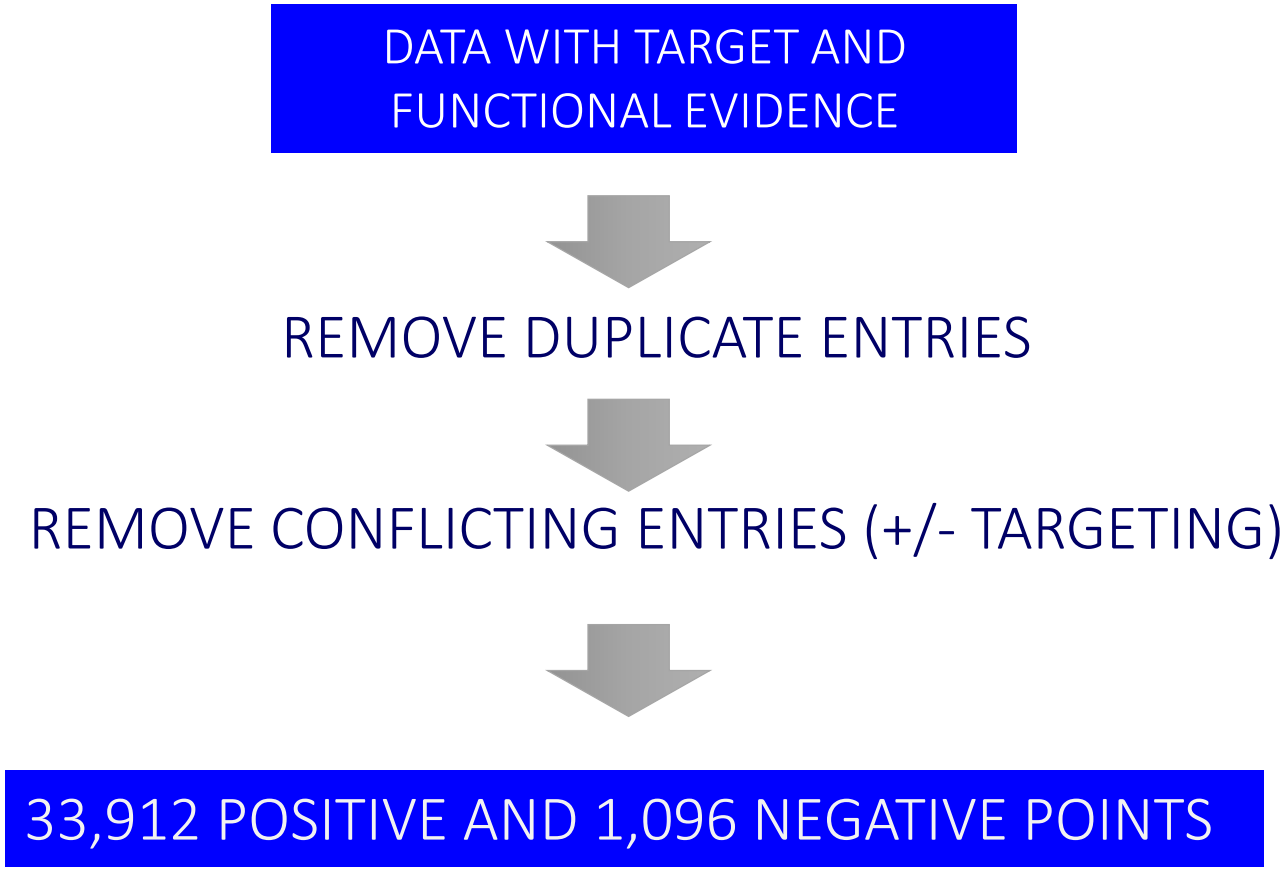
How can we improve on this?

- Don’t make any assumptions about the targeting process
- Use all the available raw data

Data Collection

NEED SIMILAR AMOUNTS OF POSITIVE AND NEGATIVE DATA

- POSITIVE: EVERYONE MEASURES THIS
- NEGATIVE: PEOPLE AREN'T SO INTERESTED IN THIS


DATA WITH TARGET AND FUNCTIONAL EVIDENCE
REMOVE DUPLICATE ENTRIES
REMOVE CONFLICTING ENTRIES (+/- TARGETING)
33,912 POSITIVE AND 1,096 NEGATIVE POINTS

miRNA TARGETING: CANDIDATE SITE SELECTION MODEL (CSSM)

THIS STILL DOESN'T GIVE US ENOUGH DATA

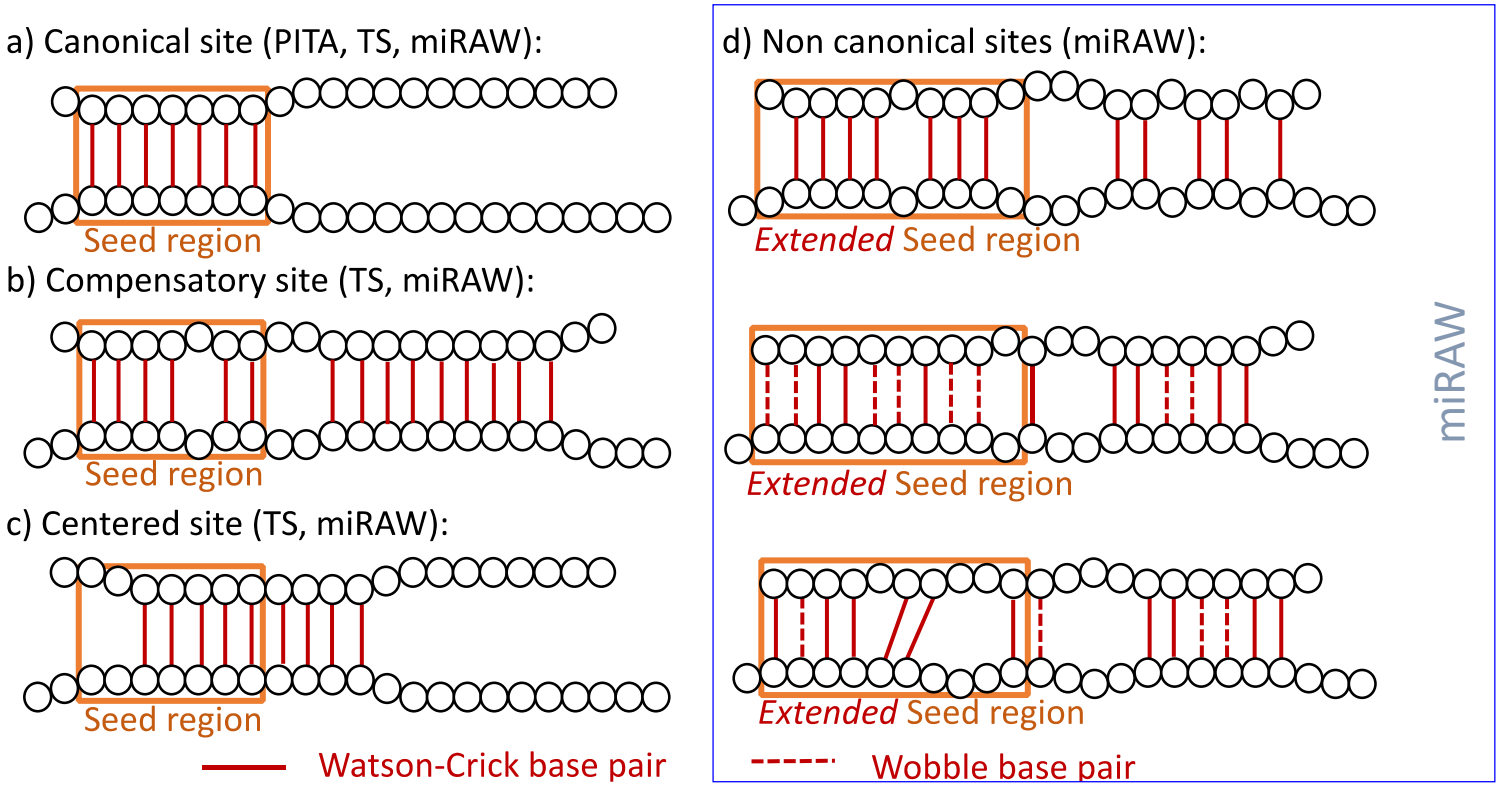
CSSM

(Relaxed/Greedy approach)

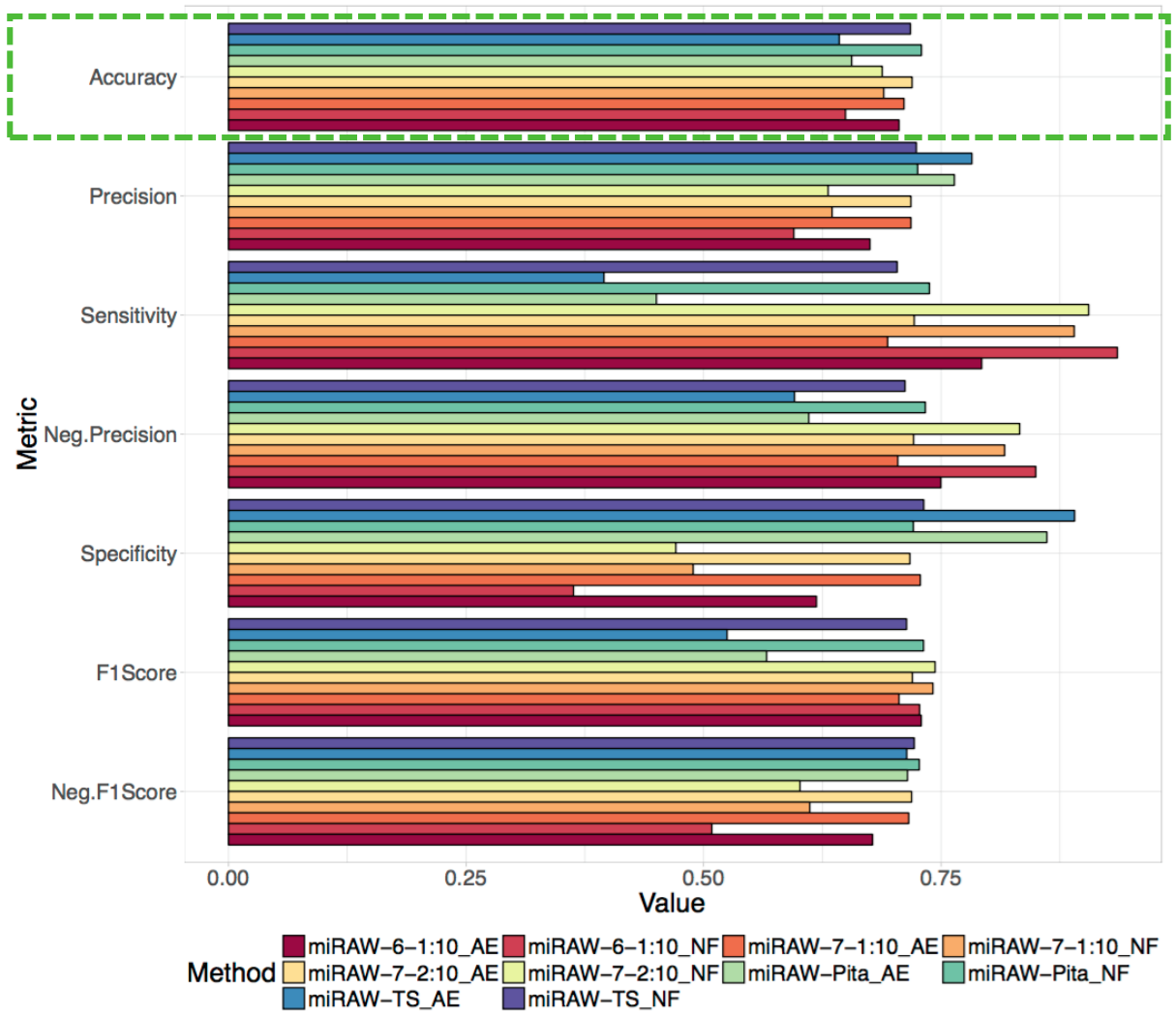
Extended seed region (10)

Requires 6 or 7 matching pairs

Allow mismatches, wobble pairs, bulges…



How does the model perform?



miRAW-TS-NF
miRAW-TS_AE
miRAW-Pita_NF
miRAW-Pita_AE
miRAW-7-2:10_NF
miRAW-7-2:10_AE
miRAW-7-1:10_NF
miRAW-7-1:10_AE
miRAW-6-1:10_NF
miRAW-6-1:10_AE

$$\text{ACCURACY} = \frac{TP + TN}{TP + FP + FN + TN}$$

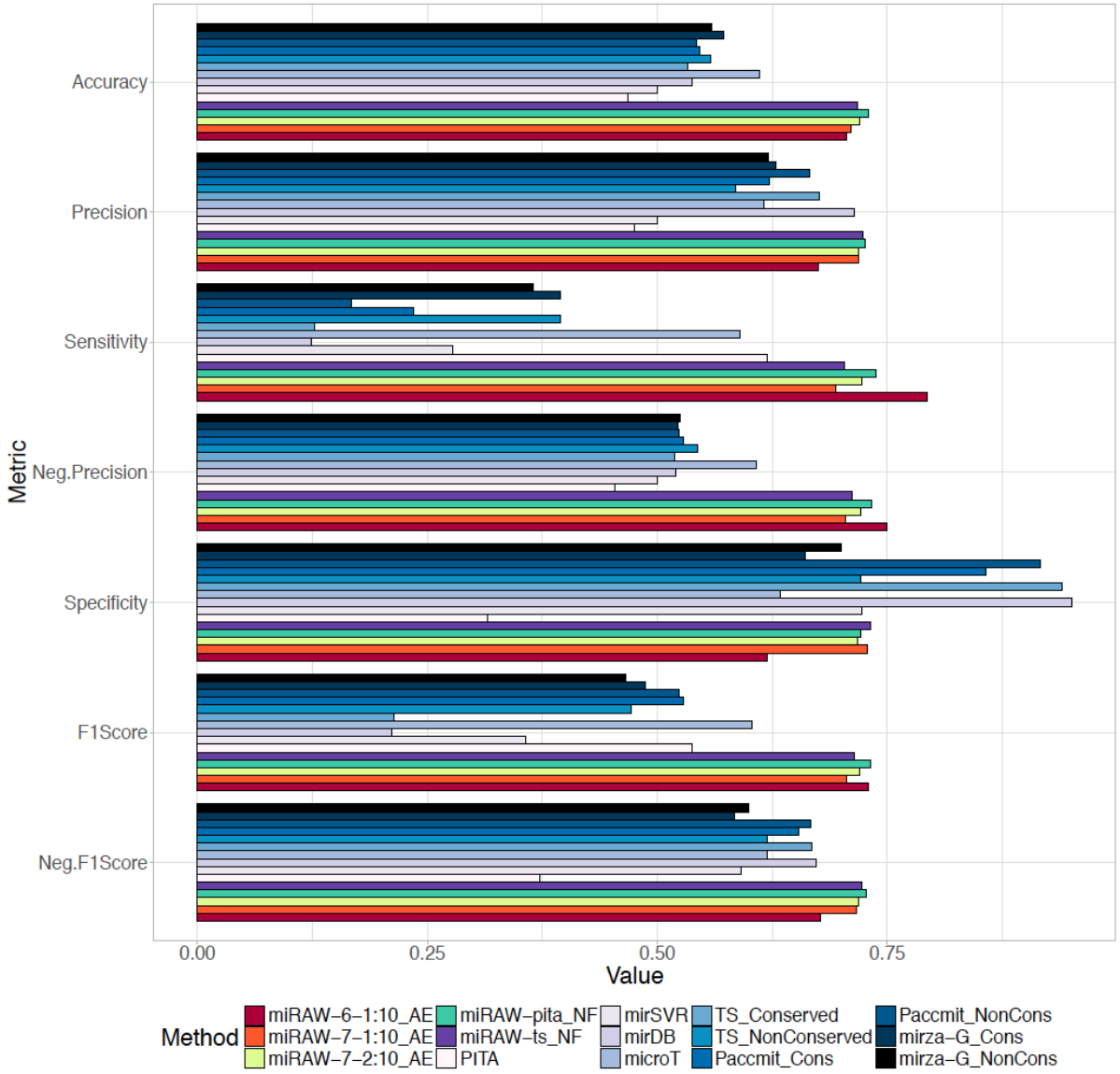
TP = correctly identified miRNA binding site
TN = correctly identified site where miRNA does not bind

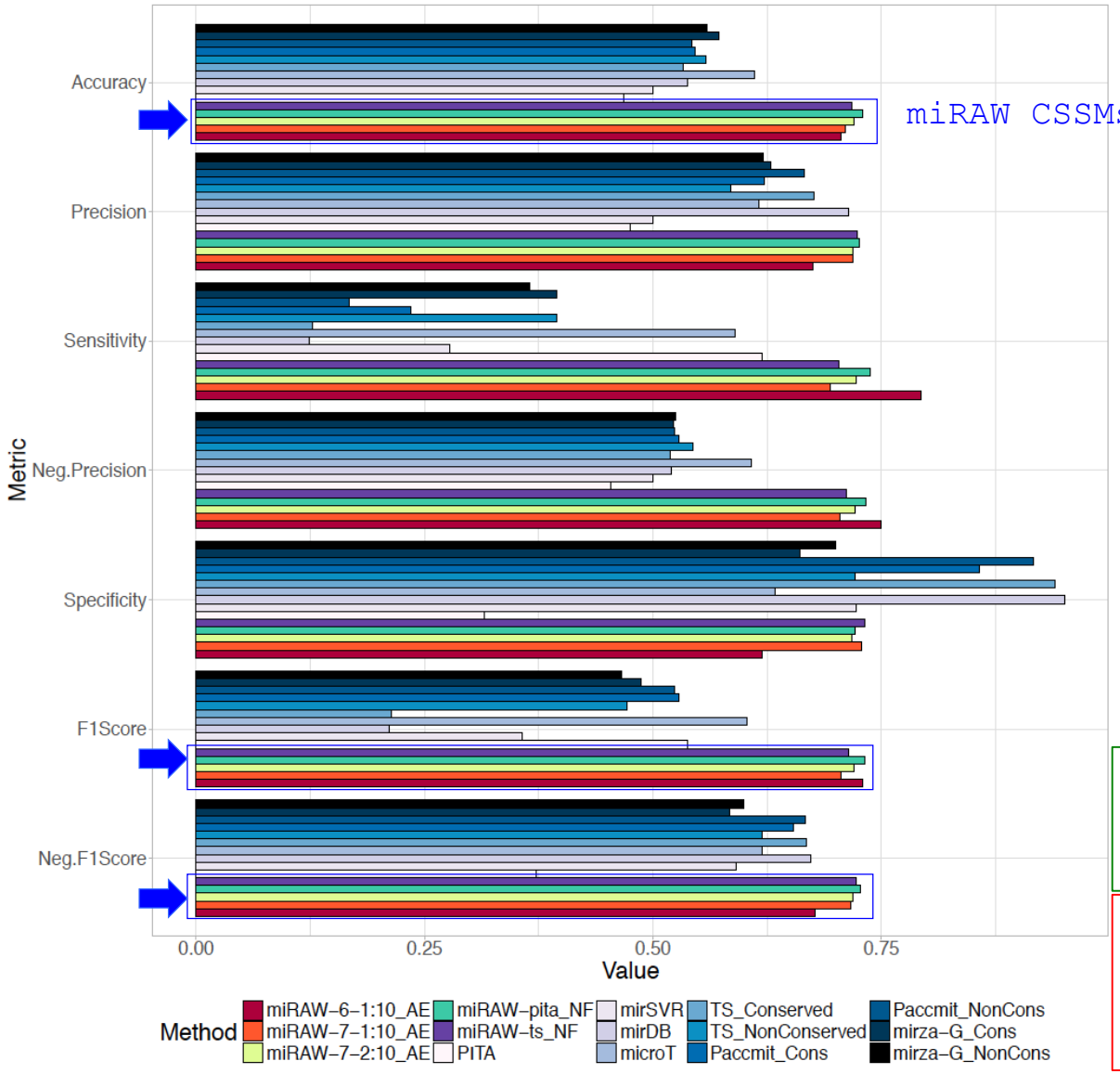
A MEASURE OF THE OVERALL PERFORMANCE OF THE MODEL

FP = we say an miRNA binds to a 3'UTR – it doesn't
FN = we say an miRNA doesn't bind to any 3'UTR. It does

THE CANONICAL CSSMs HAVE HIGHER ACCURACY IN THE ABSENCE OF FILTERING
THE NON-CANONICAL CSSMS HAVE HIGHER ACCURACY IN PRESENCE OF FILTERING

Comparison with other prediction tools





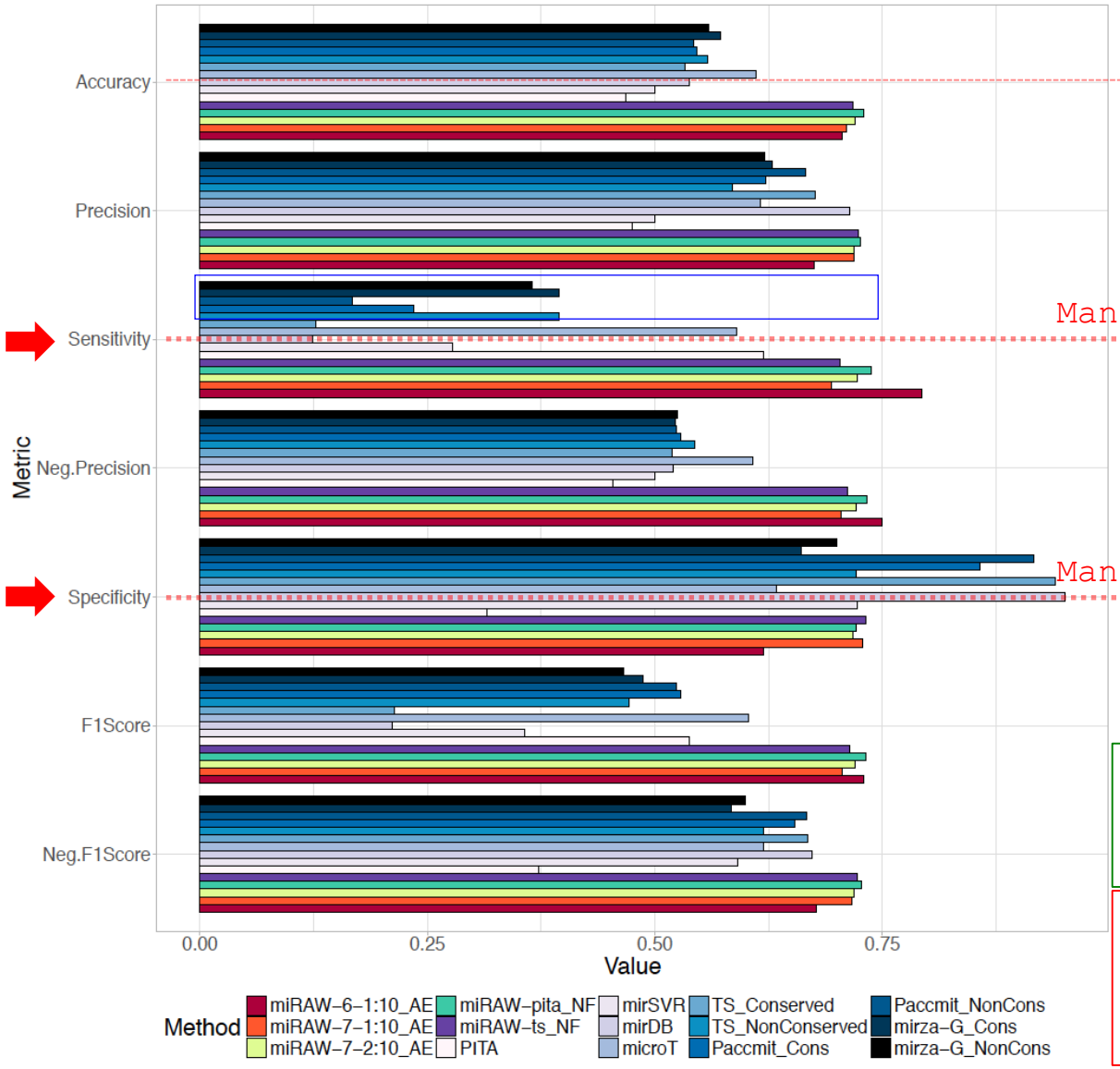
miRAW CSSMs

$$\frac{TP+TN}{TP+FP+FN+TN}$$

miRAW's OVERALL PERFORMANCE IS SIGNIFICANTLY BETTER

TP = correctly identified miRNA binding site
TN = correctly identified site where miRNA does not bind

FP = we say an miRNA binds to a 3'UTR – it doesn't
FN = we say an miRNA doesn't bind to any 3'UTR. It does



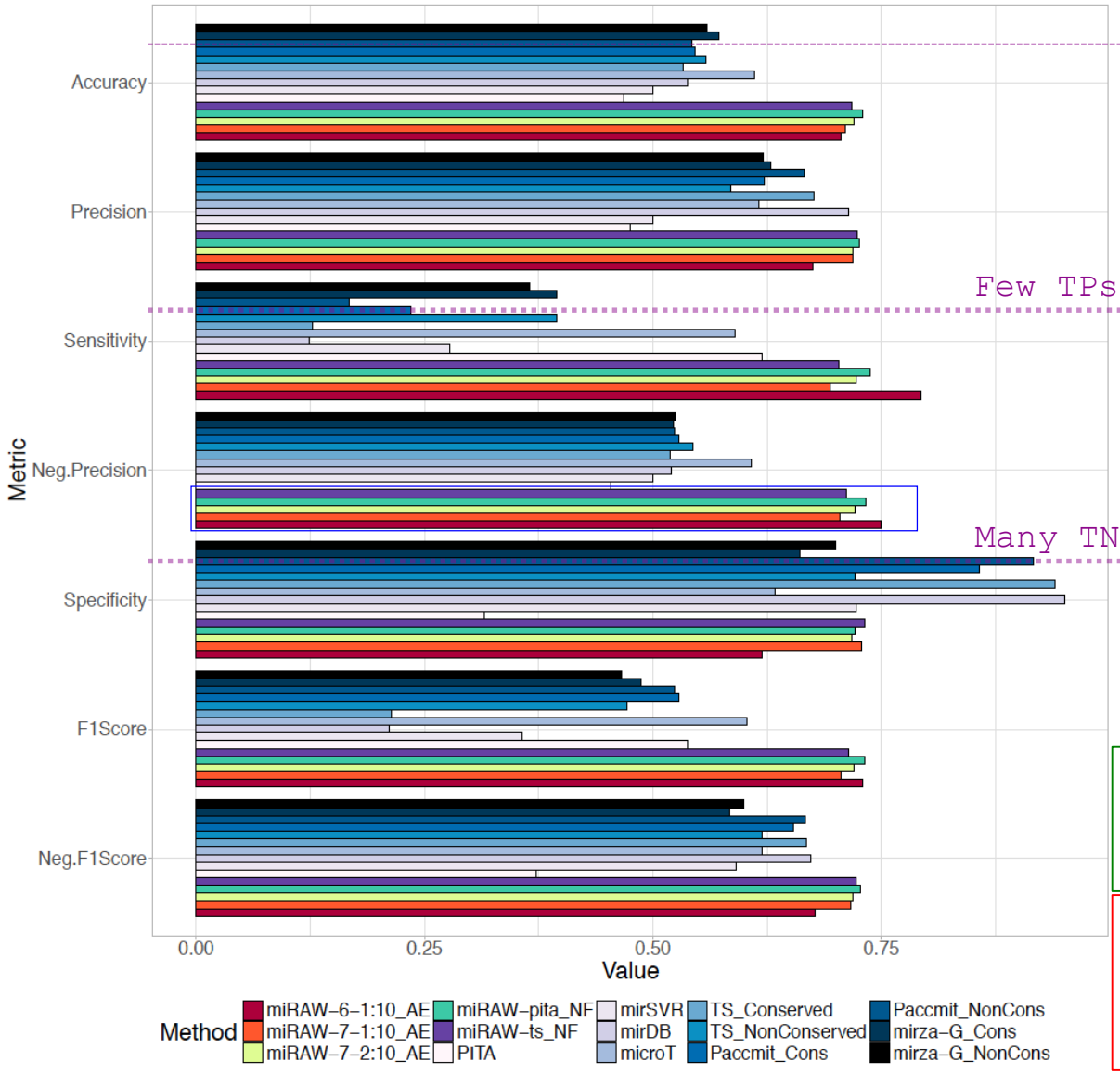
mirDB

Many FNs/Few TPs $\frac{TP}{TP+FN}$

Many TNs/Few FPs $\frac{TN}{TN+FP}$

TP = correctly identified miRNA binding site
TN = correctly identified site where miRNA does not bind

FP = we say an miRNA binds to a 3'UTR – it doesn't
FN = we say an miRNA doesn't bind to any 3'UTR. It does



Paccmit

Few TPs/Many FNs

$$\frac{TP}{TP+FN}$$

Many TNs/Few FPs

$$\frac{TN}{TN+FP}$$

TP = correctly identified miRNA binding site
TN = correctly identified site where miRNA does not bind

FP = we say an miRNA binds to a 3'UTR – it doesn't
FN = we say an miRNA doesn't bind to any 3'UTR. It does

- miRAW outperforms existing target prediction tools
- Discovers existing canonical targets, but identifies many non-canonical ones that are missed by knowledge based models
- We made (almost) no assumptions about what is important - Insufficient data → restrict model
- Hardest part is building the dataset – lack of data standards

PLOS | SUBMISSION

miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts

Albert Pla[1], Xiangfu Zhong[1,2], Simon Rayner[1,2*]

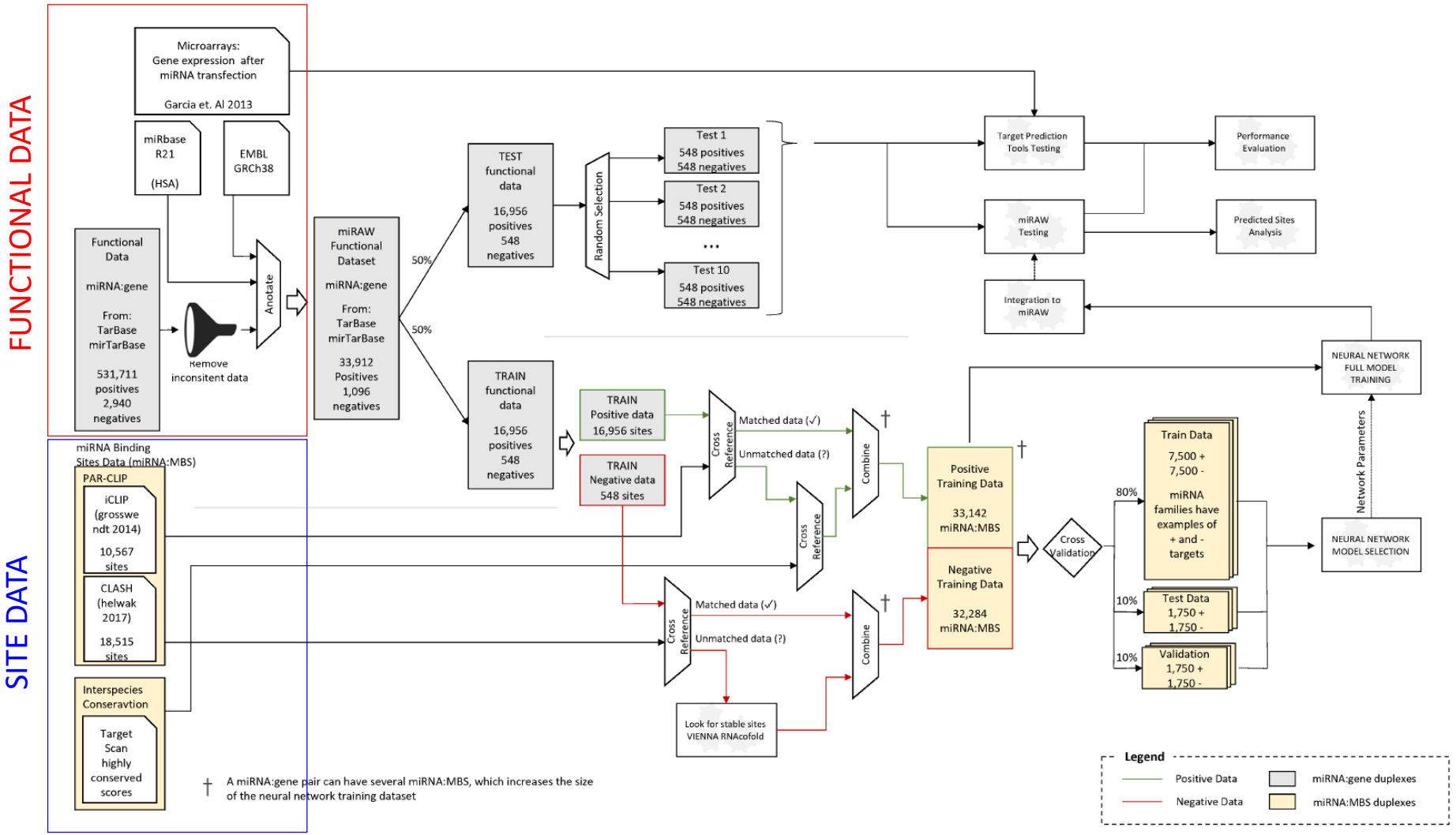
1 Department of Medical Genetics, University of Oslo, Oslo, Norway
2 Avdeling for Medisinsk Genetikk , Oslo University Hospital, Oslo, Norway

a.p.planas@medisin.uio.no, joey.zhong.cn@gmail.com, *simon.rayner@medisin.uio.

In press PLOS Comp Bio

Data and source code available at:
https://bitbucket.org/account/user/bipous/projects/MIRAW

Data preparation



Ongoing work

Developing an experimental platform allowing us to directly investigate miRNA-mRNA interactions of interest (compare shotgun vs directed sequencing)

Generate more negative data

Better understanding of targeting process