Into the Deep

microRNA Detection using Next-Generation Sequencing Data

David Langenberger

Bioinformatics Group

Department of Computer Science

University of Leipzig

UNIVERSITÄT LEIPZIG

Next-Generation Sequencing

Massively Parallel Sequencing technologies

454 Life Sciences / Roche
(FLX Titanium Series)

1 million reads
400 bp

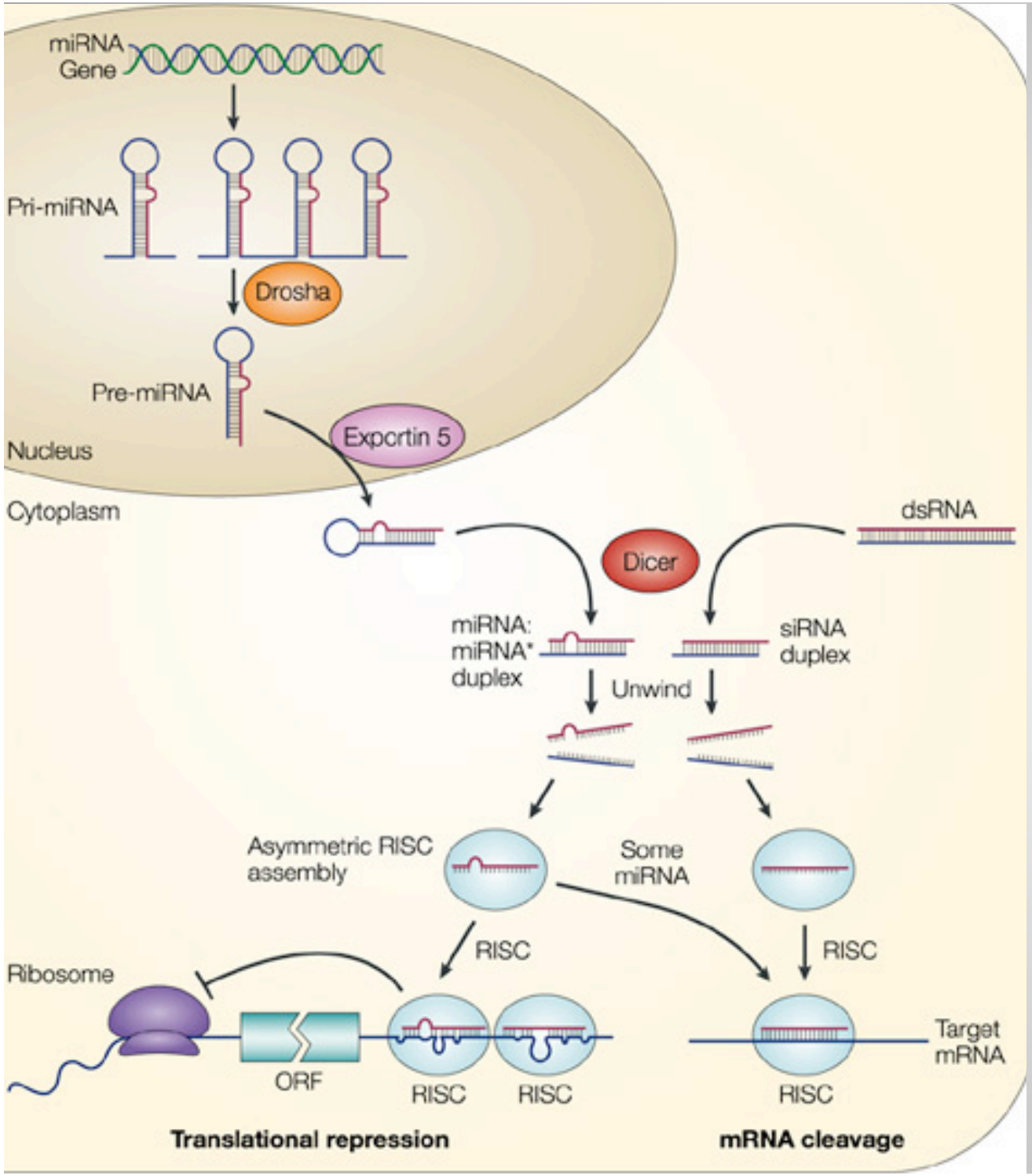
Solexa / Illumina
(Genome Analyzer)

40 million reads
35 bp

Applied Biosystems
(SOLiD System)

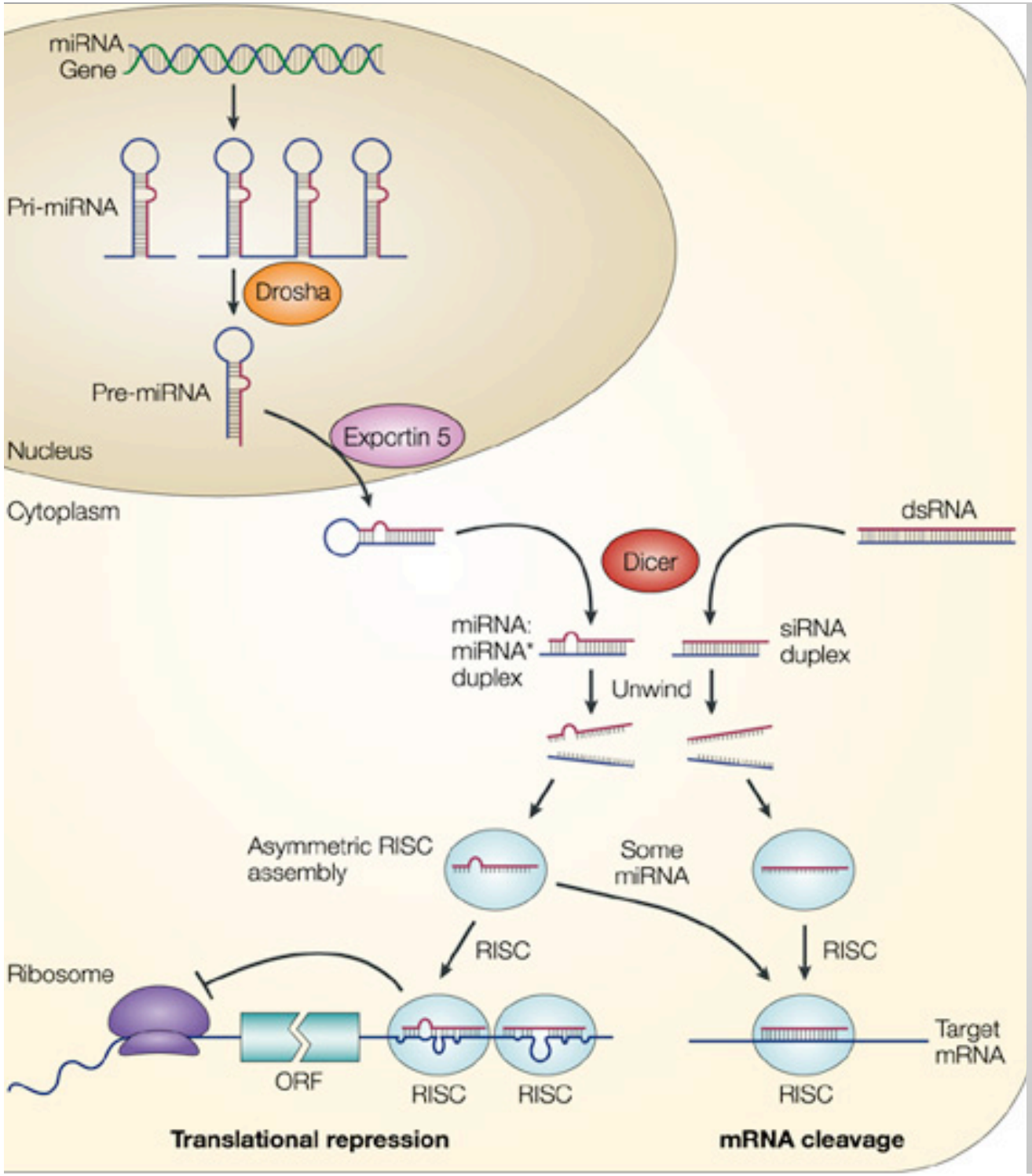
70 million reads
35 bp

microRNAs



Nature Reviews Genetics 5, 522-531, 2004

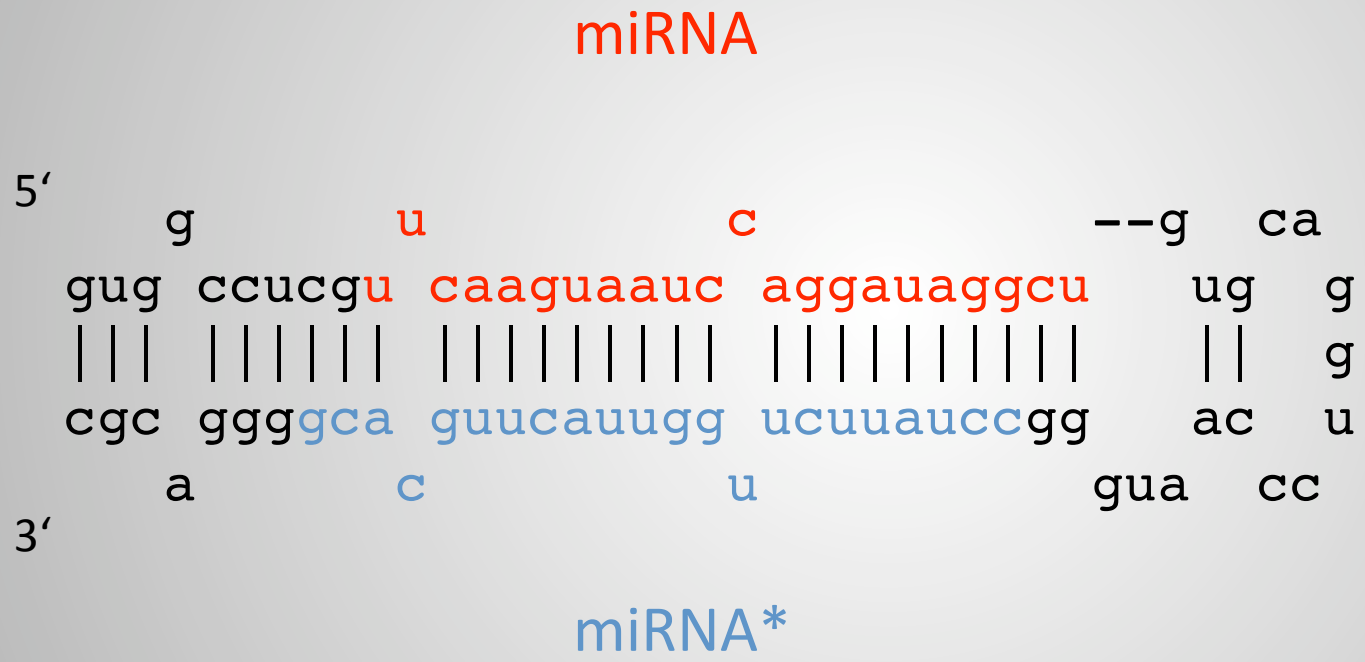
Information content



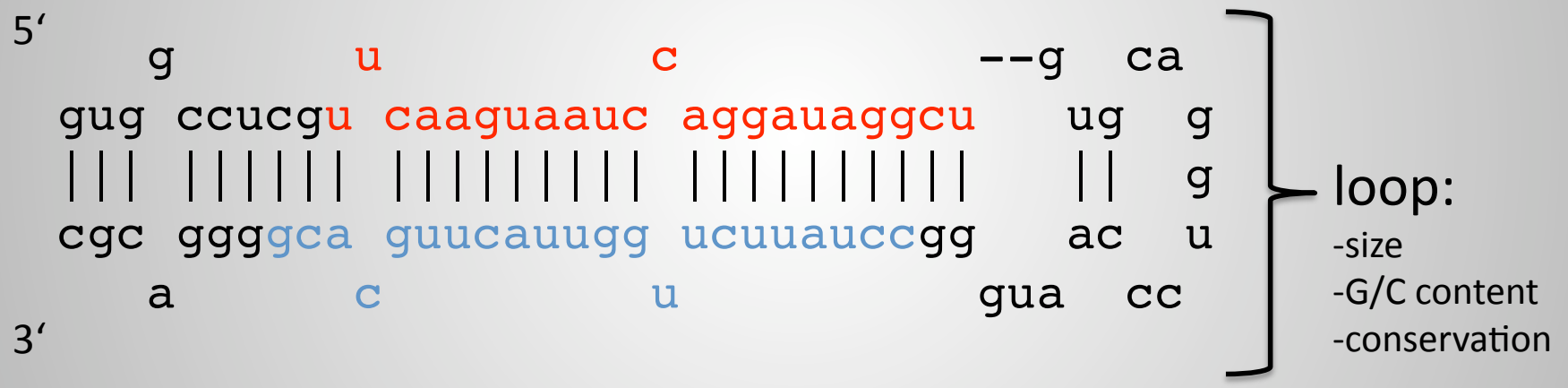
Nature Reviews Genetics 5, 522-531, 2004

Nature Reviews Genetics 5, 522-531, 2004

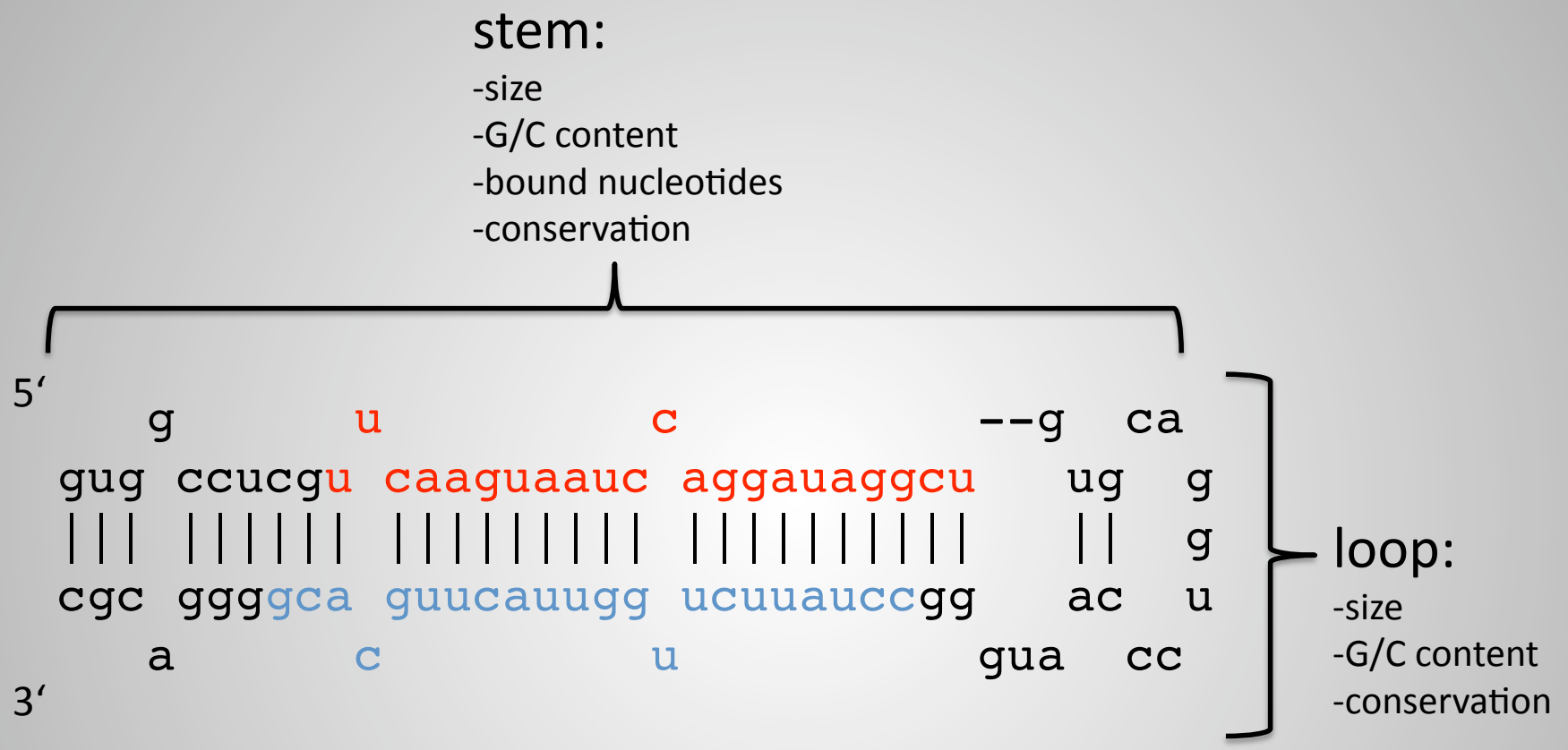
hairpin



hsa-miR-26a


5‘
g u c --g ca
gug ccucgu caaguaaauc aggauaggcu ug g
||| |||||| ||||||||| |||||||||| || g
cgc ggggca guucauugg ucuuauccgg ac u
a c u gua cc
3‘
loop:
-size
-G/C content
-conservation

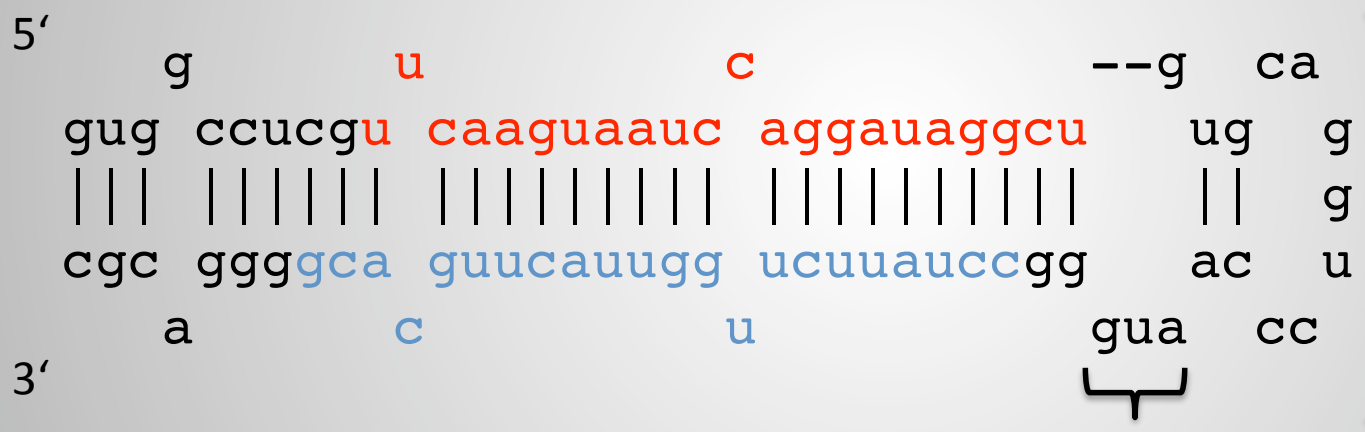
hsa-miR-26a


stem:
-size
-G/C content
-bound nucleotides
-conservation
5‘
g u c --g ca
gug ccucgu caaguaaauc aggauaggcu ug g
||| |||||| ||||||||| |||||||||| || g
cgc ggggca guucauugg ucuuauccgg ac u
a c u gua cc
3‘
loop:
-size
-G/C content
-conservation

hsa-miR-26a

stem:

-size
-G/C content
-bound nucleotides
-conservation



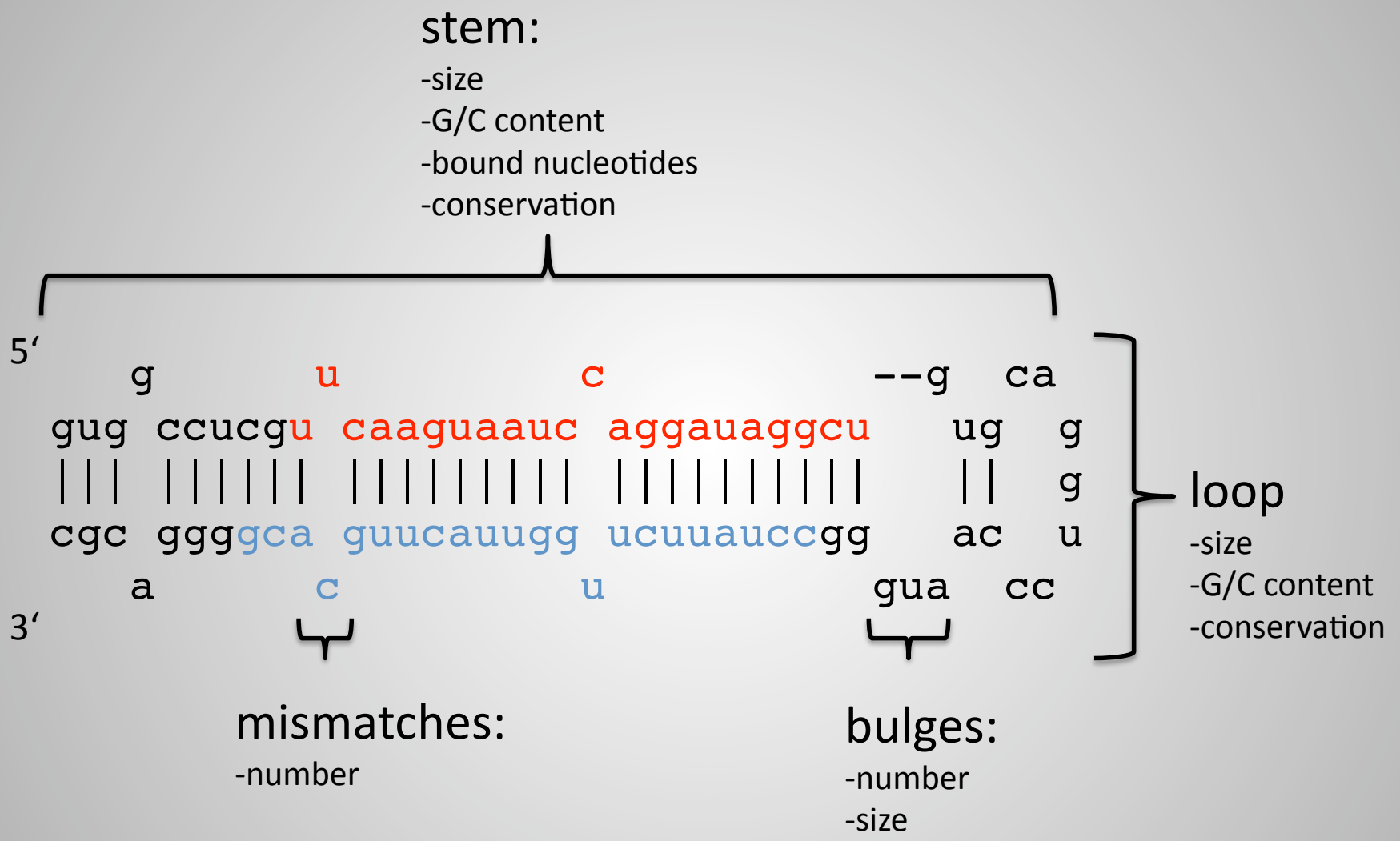
loop

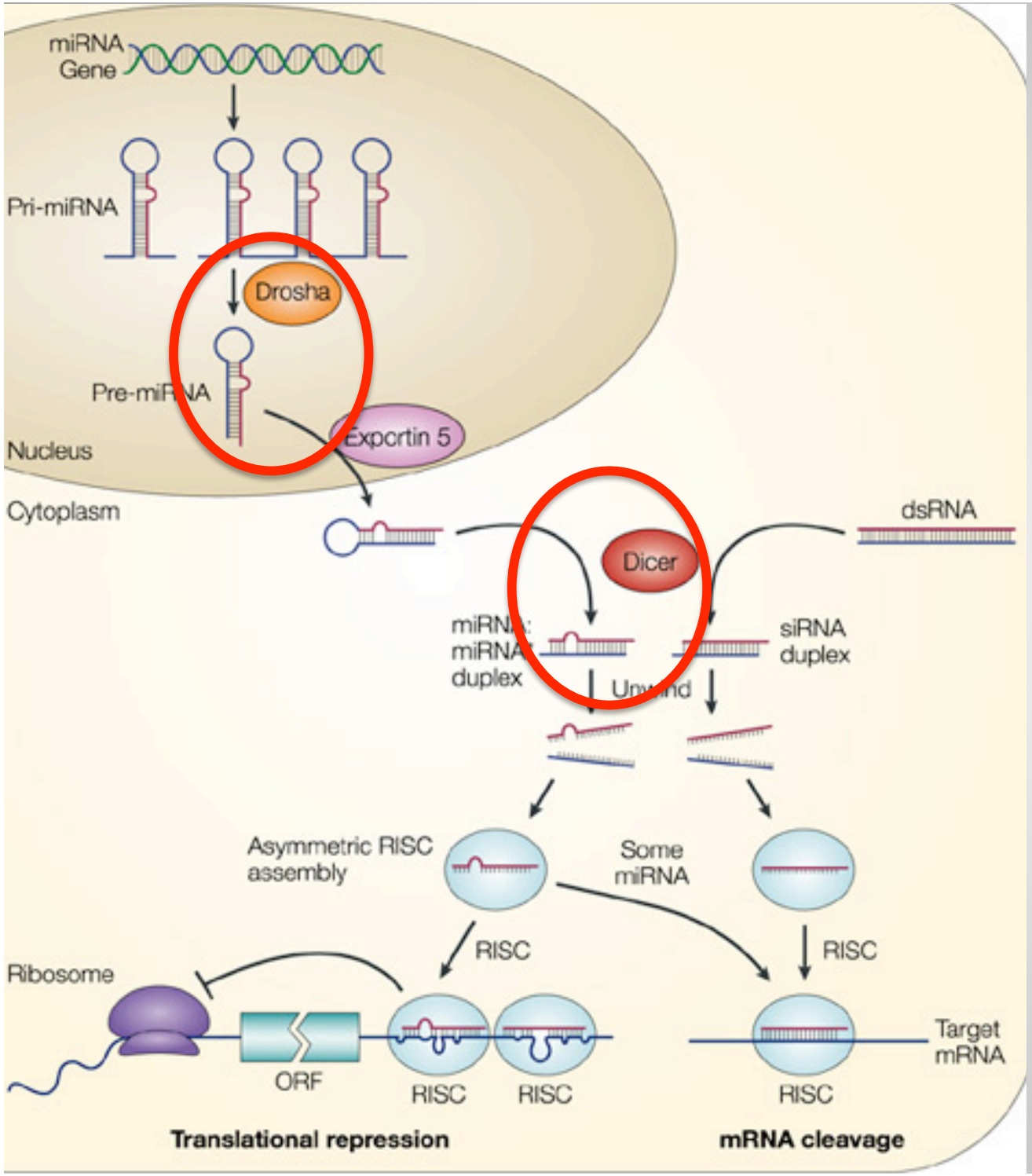
-size
-G/C content
-conservation

bulges:

-number
-size

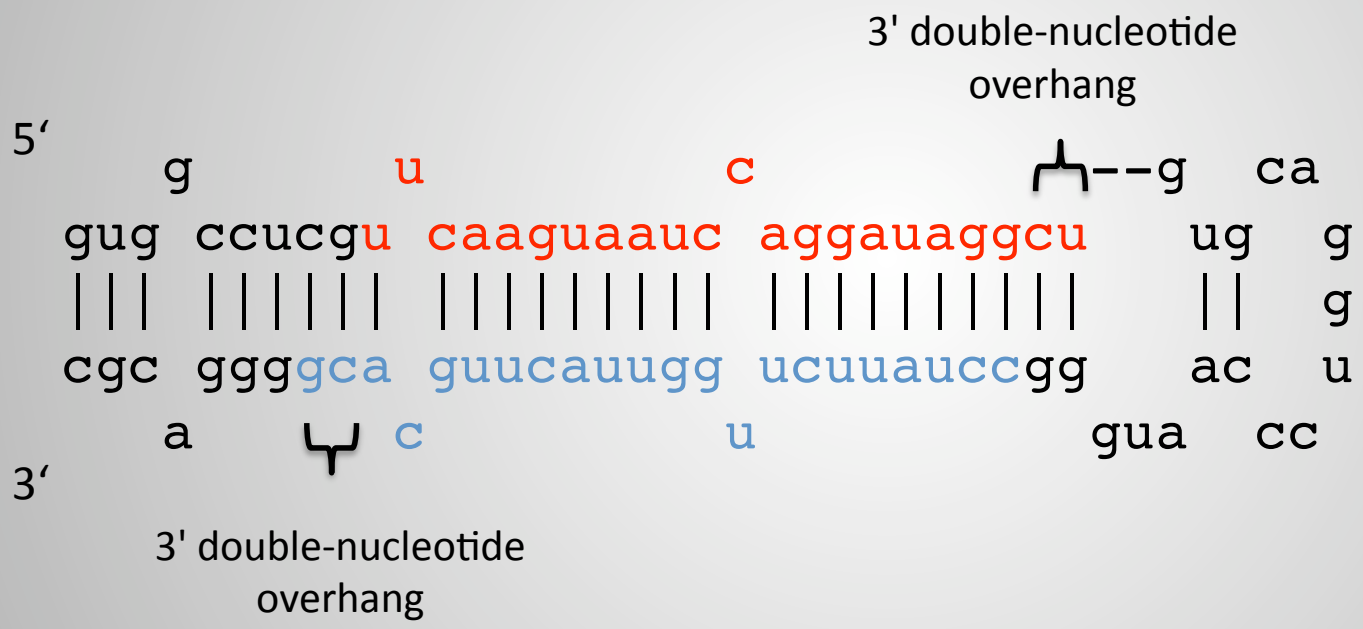
hsa-miR-26a


stem:
-size
-G/C content
-bound nucleotides
-conservation
5‘
g u c --g ca
gug ccucgu caaguaaauc aggauaggcu ug g
||| |||||| ||||||||| |||||||||| || g
cgc ggggca guucauugg ucuuauccgg ac u
a c u gua cc
3‘
loop
-size
-G/C content
-conservation
mismatches:
-number
bulges:
-number
-size
hsa-miR-26a



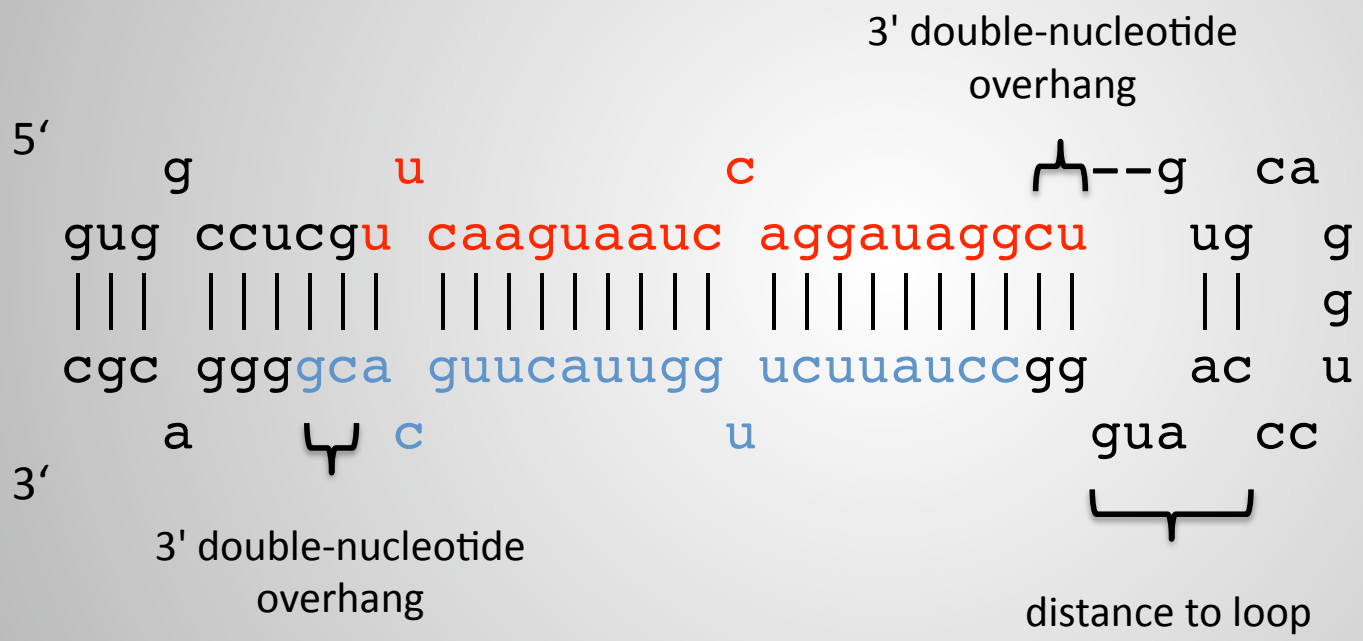
Nature Reviews Genetics 5, 522-531, 2004

hairpin



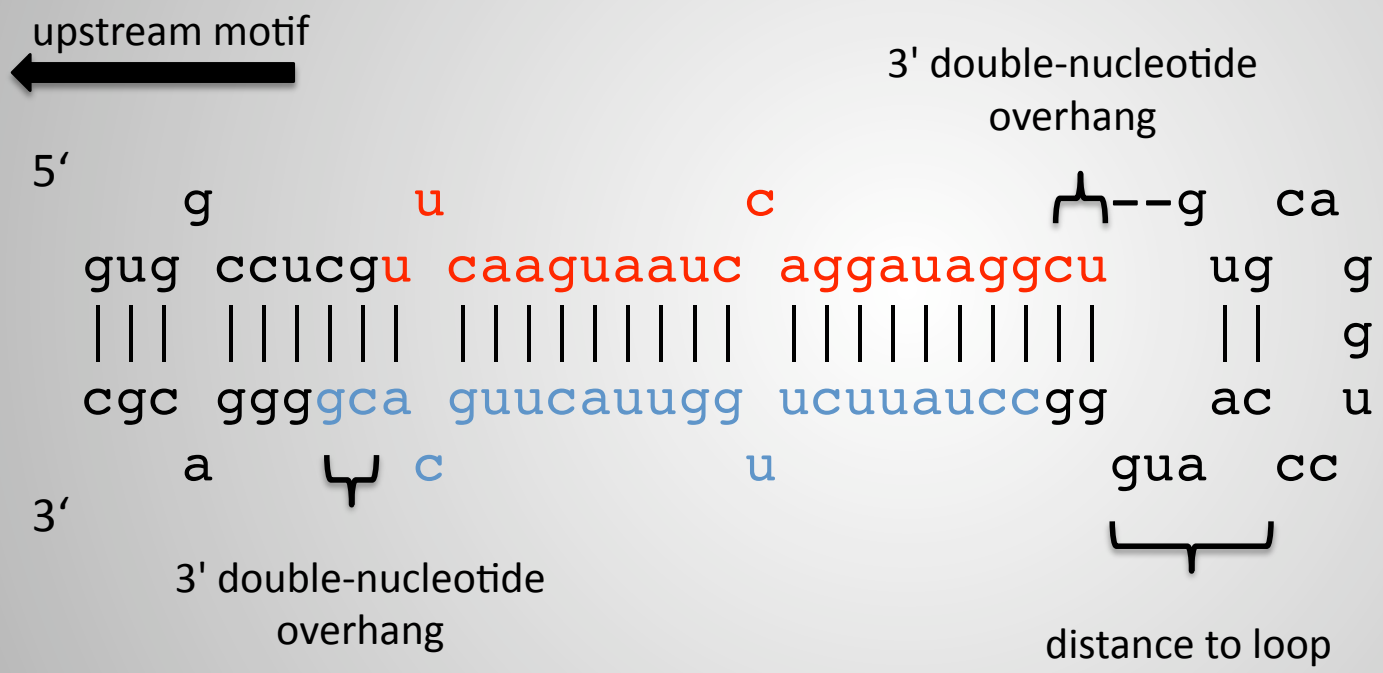
hsa-miR-26a

hairpin

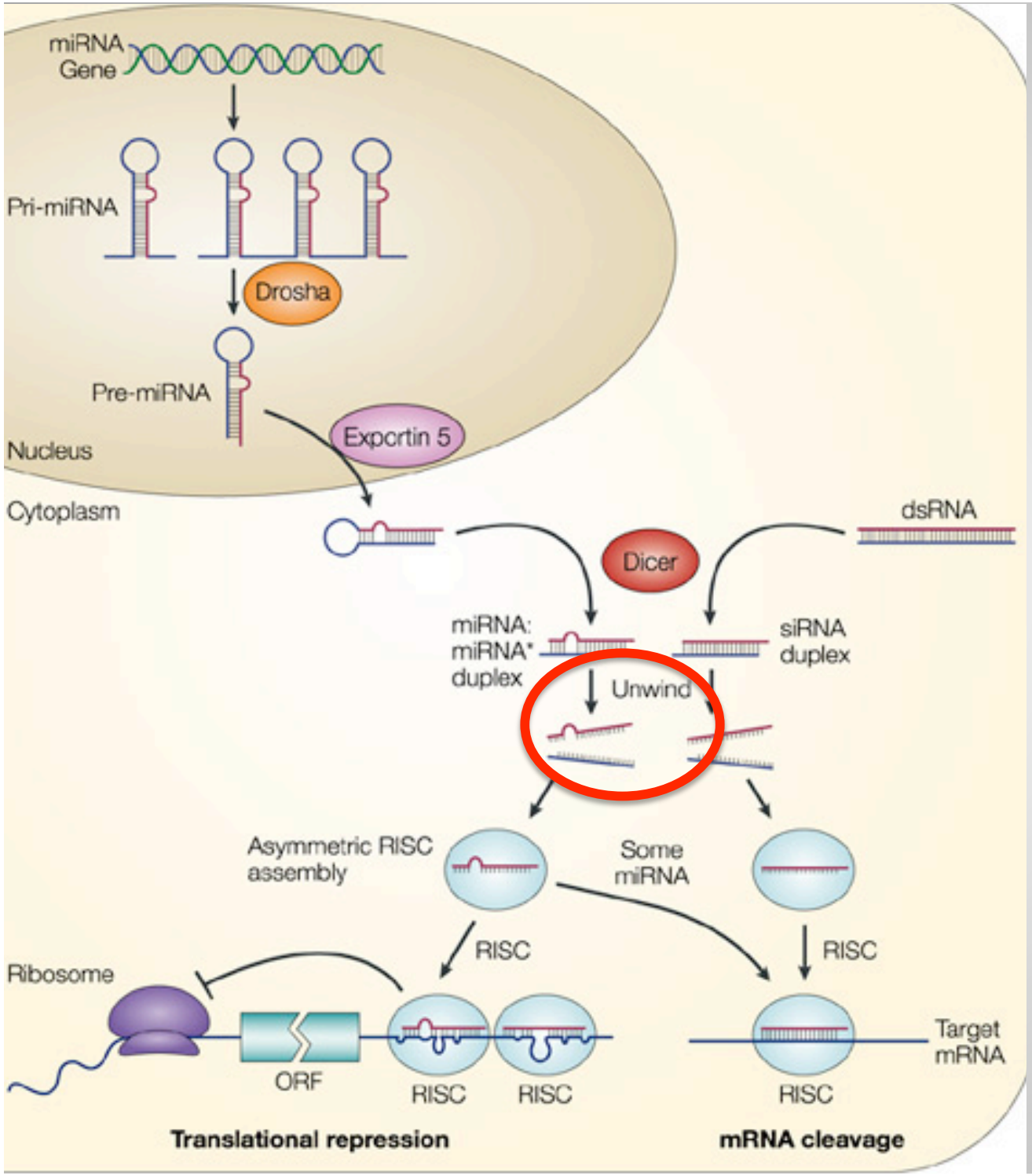


hsa-miR-26a

hairpin

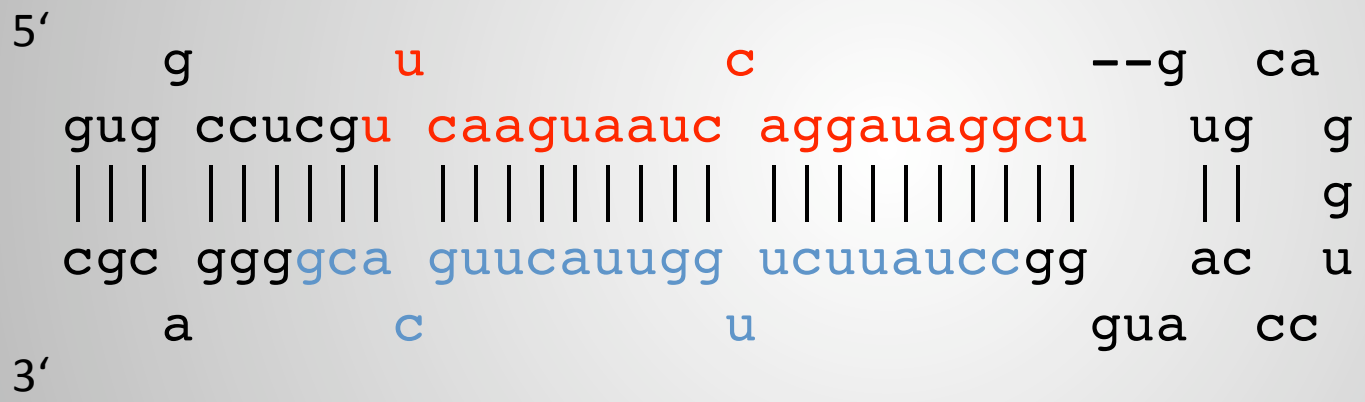


hsa-miR-26a



Nature Reviews Genetics 5, 522-531, 2004

hairpin



- miRNA–miRNA* thermodynamic asymmetry

hsa-miR-26a

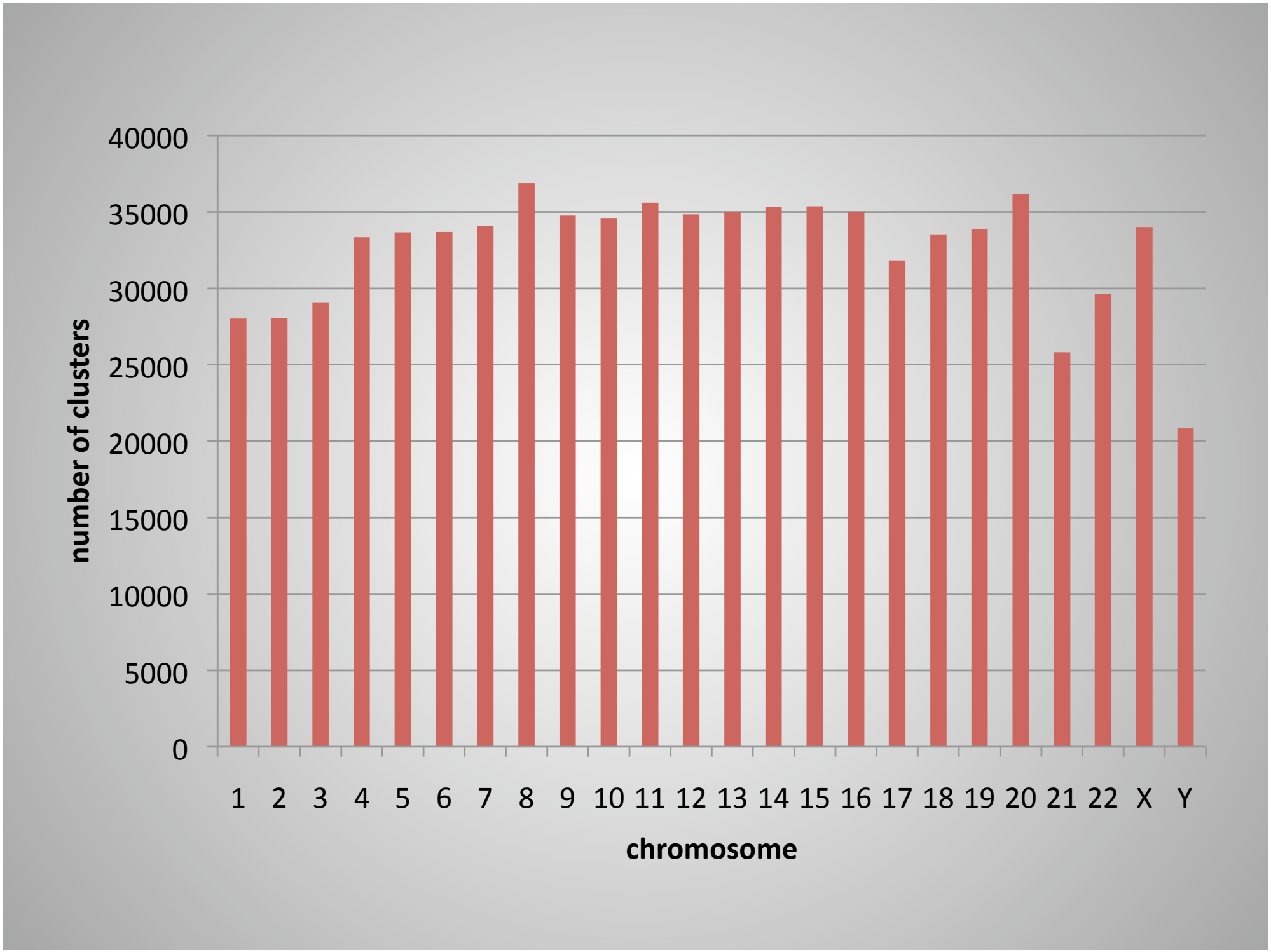
Data

355,453 reads

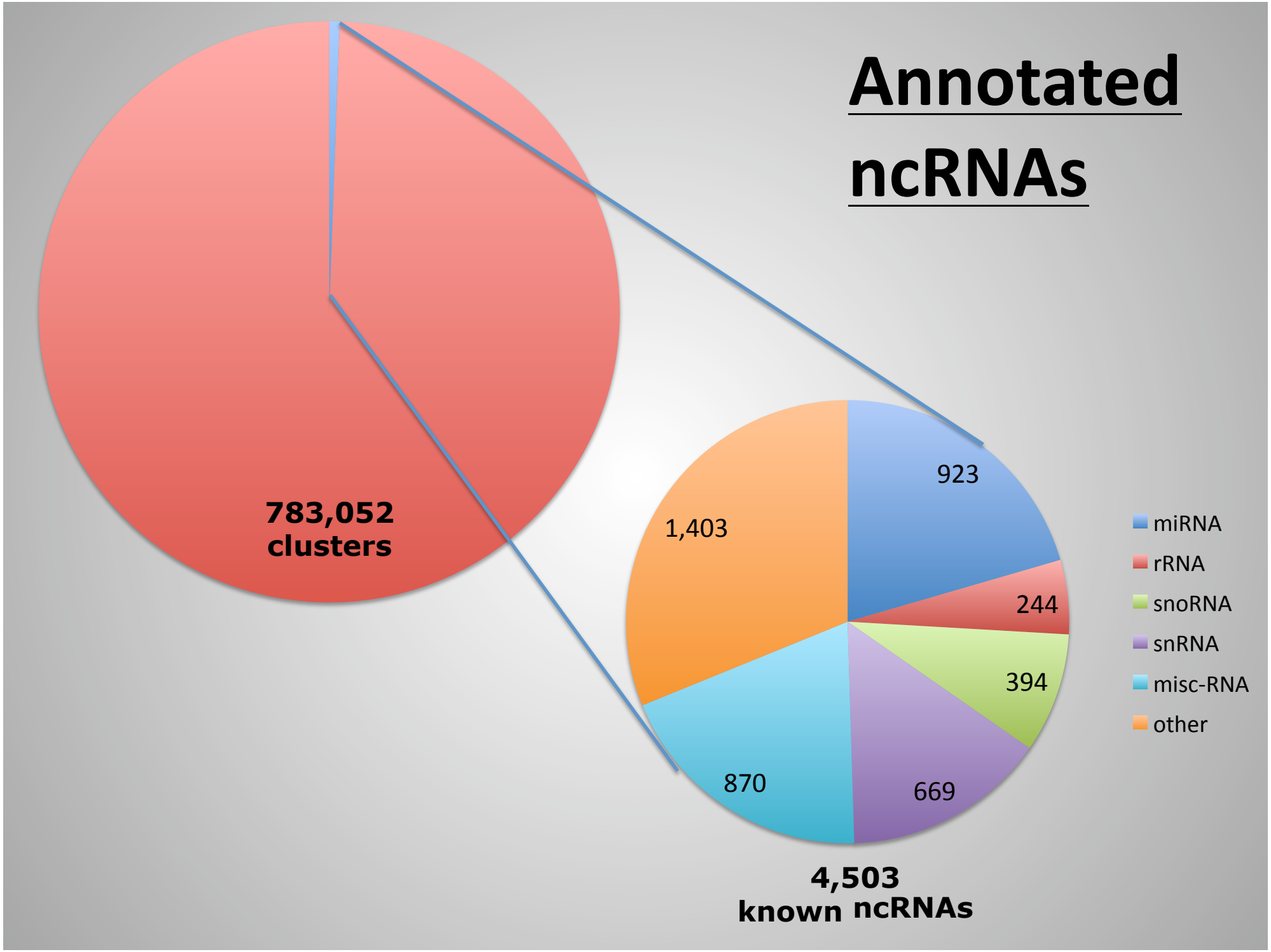
map reads to genome (segemehl)

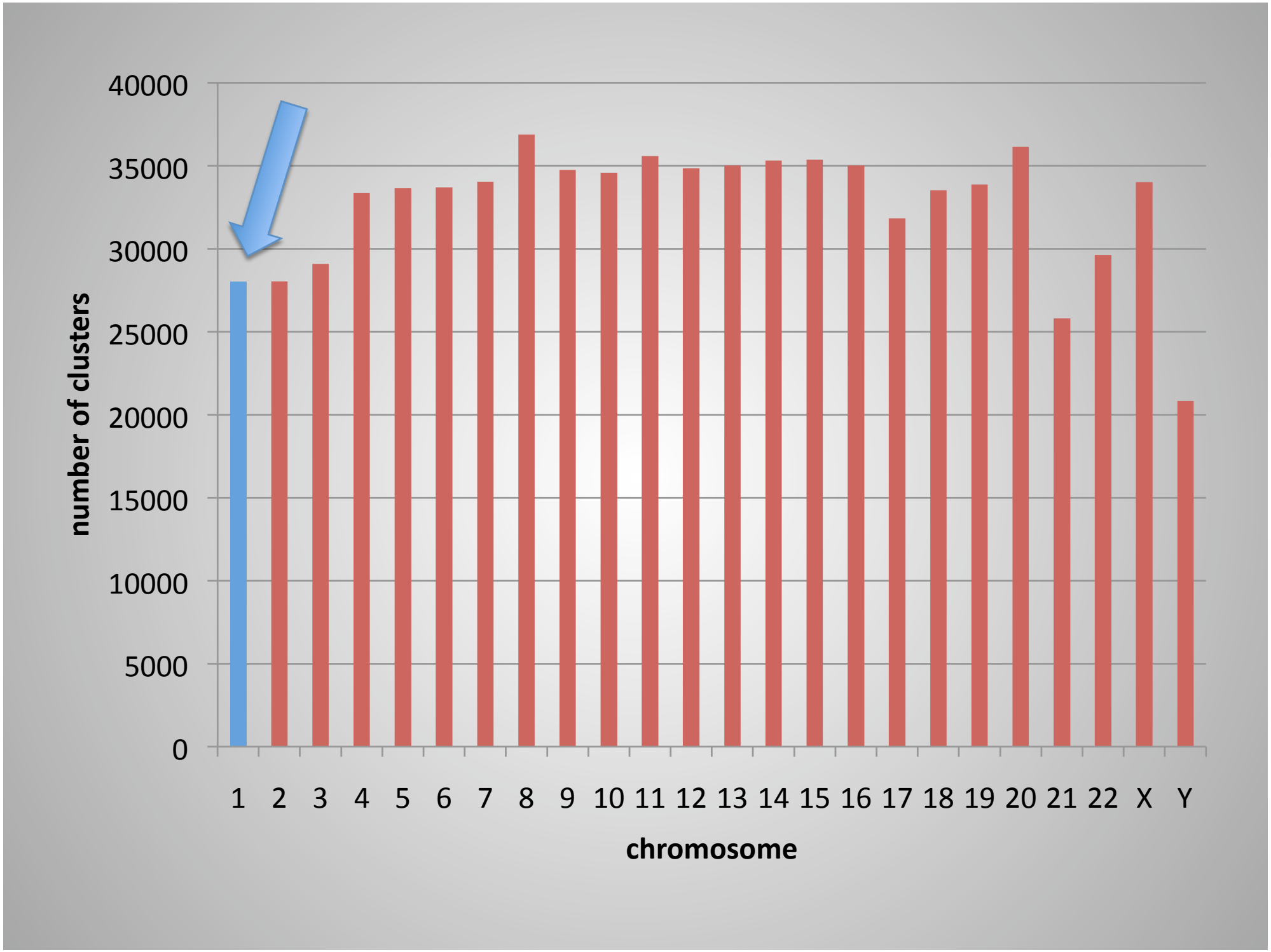
Cluster reads

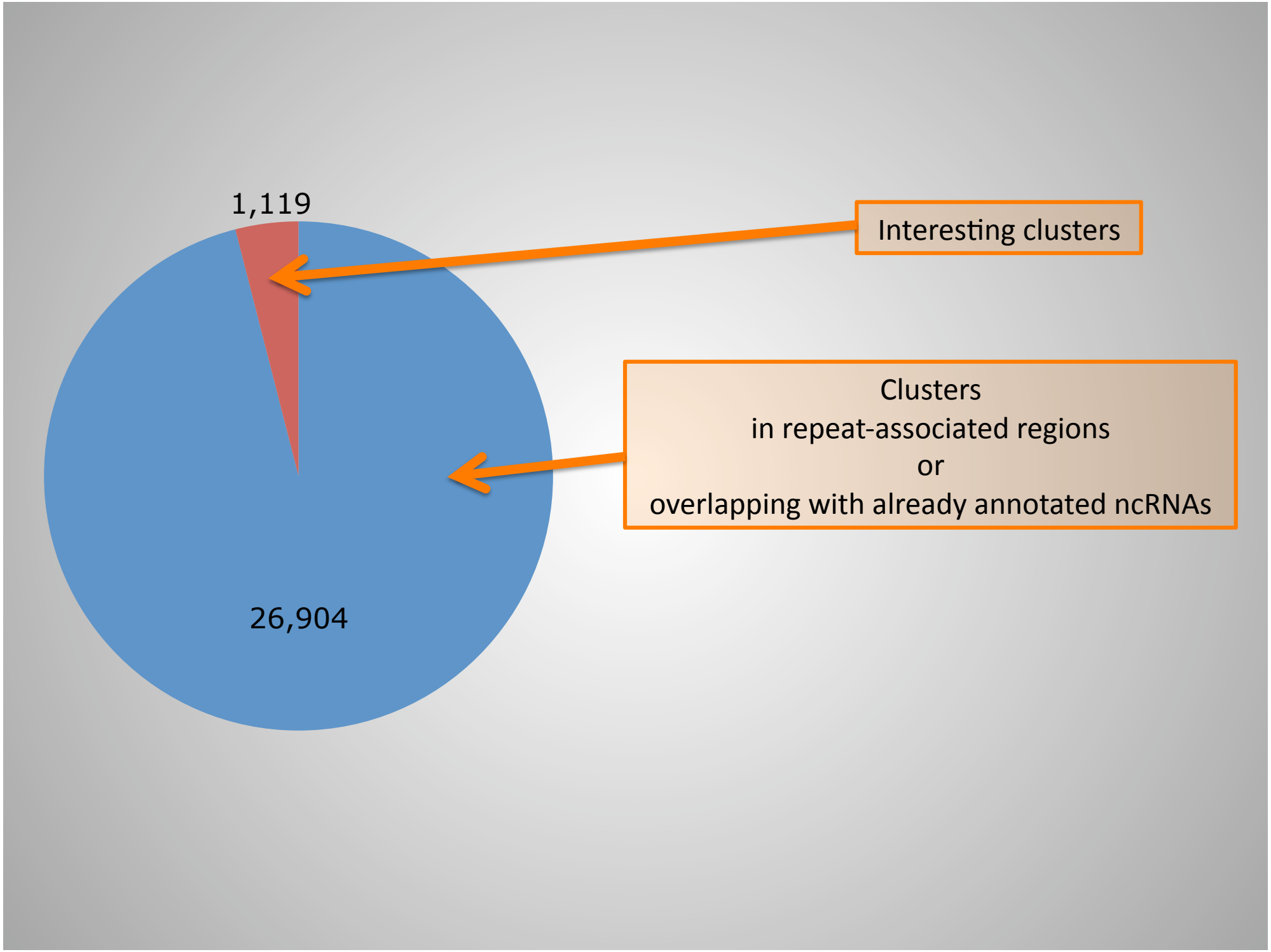
783,052 clusters

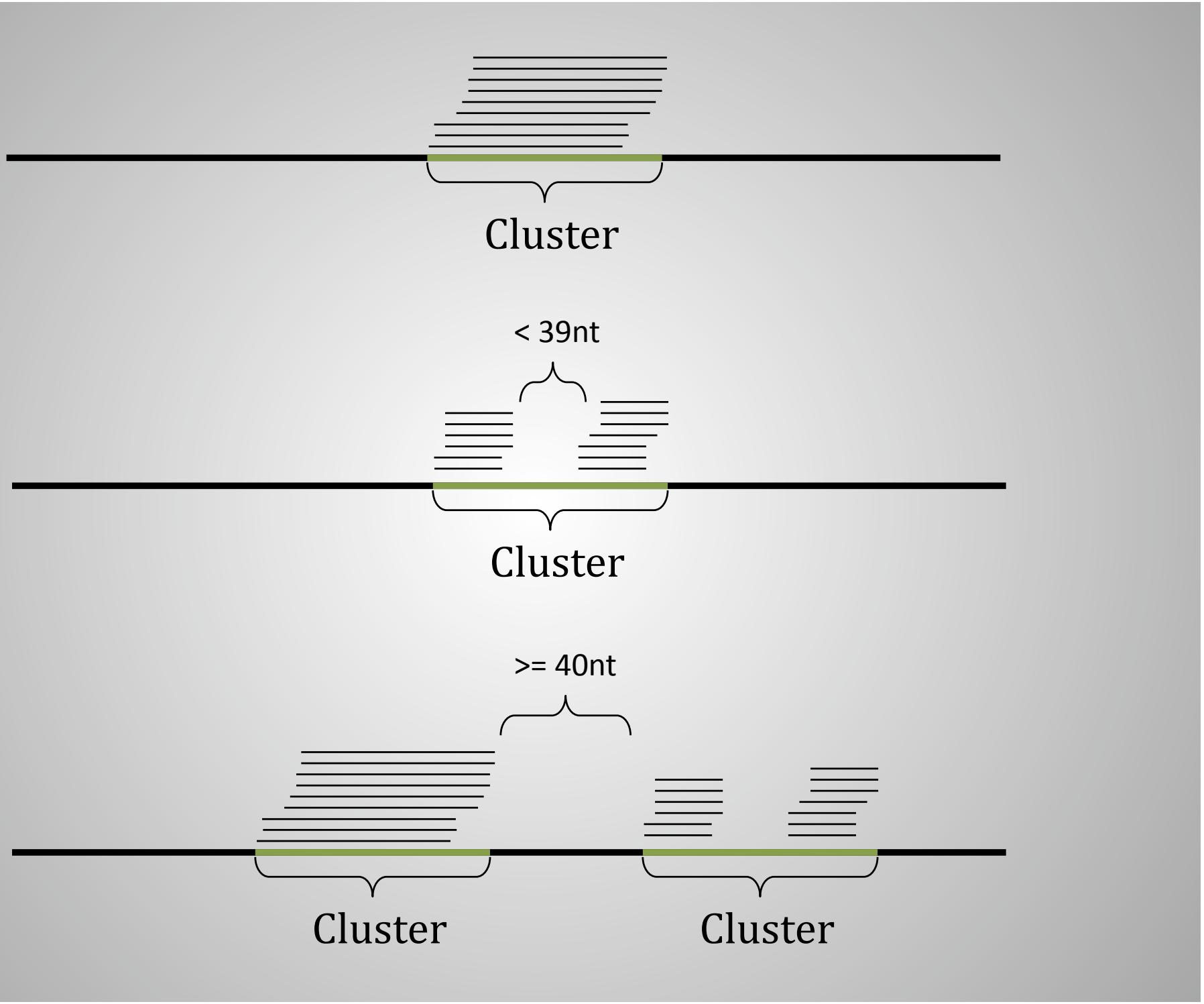

40000
35000
30000
25000
20000
15000
10000
5000
0
number of clusters
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
chromosome

Annotated ncRNAs



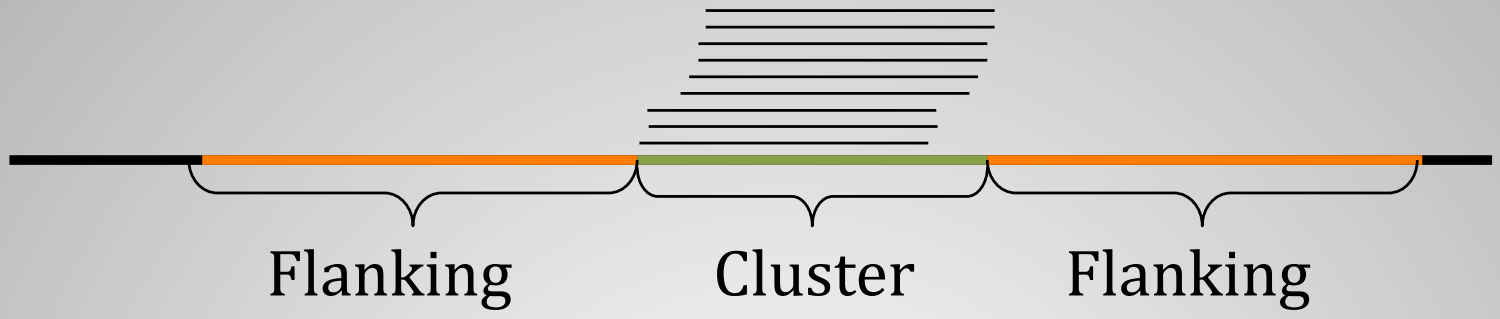

40000
35000
30000
25000
20000
15000
10000
5000
0
number of clusters
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
chromosome


1,119
26,904
Interesting clusters
Clusters
in repeat-associated regions
or
overlapping with already annotated ncRNAs


Cluster
< 39nt
Cluster
>= 40nt
Cluster
Cluster

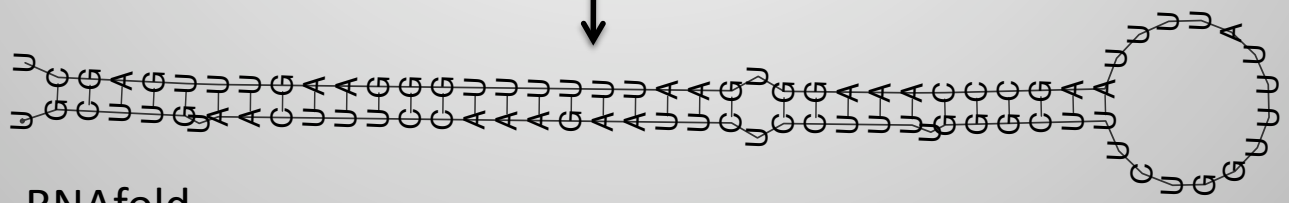
Bioinformatics pipeline

hairpin structure

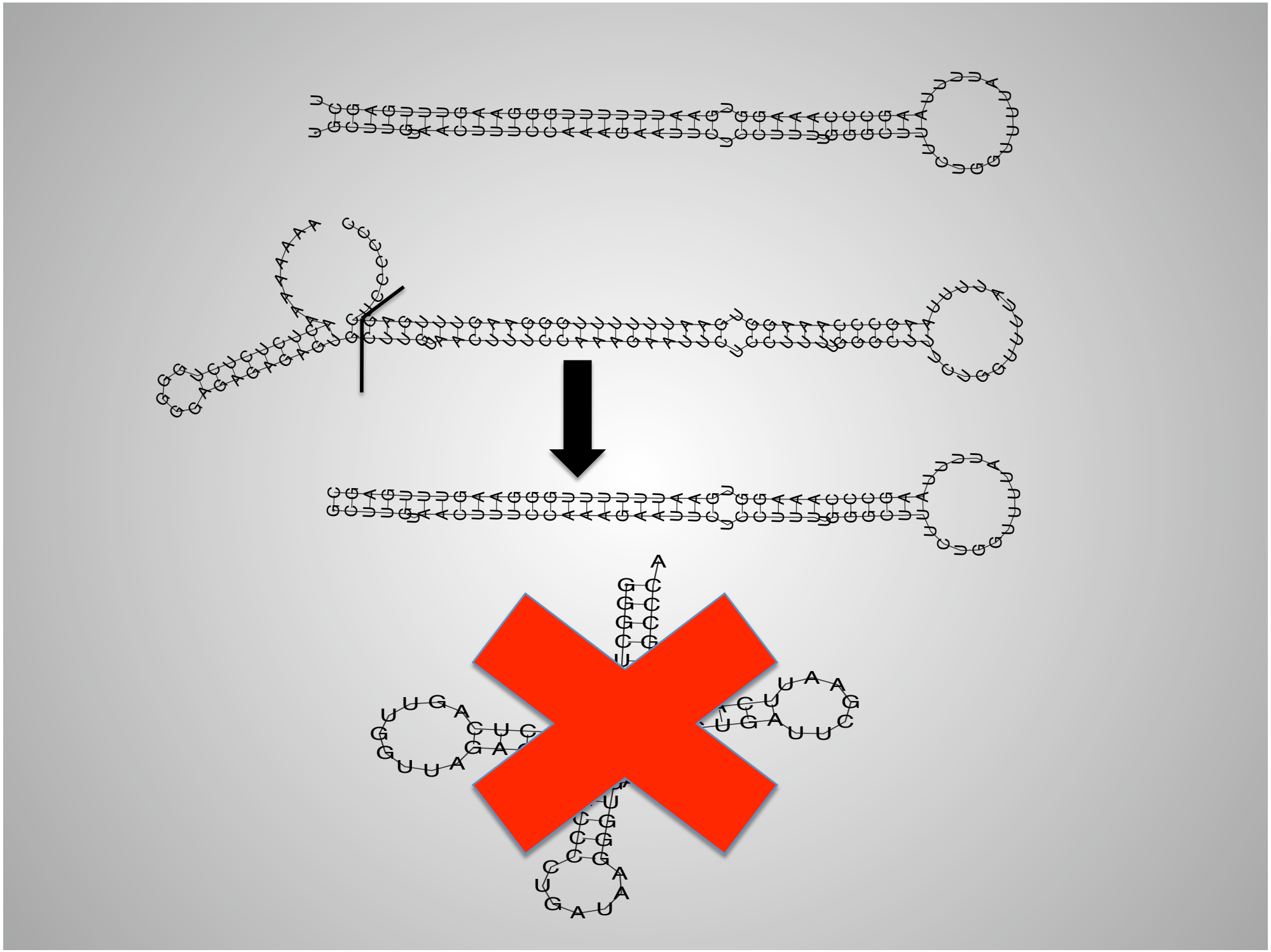


……(((((((((((((((((..((((((((((….)))))))))))))))))))))))))).)))))……..

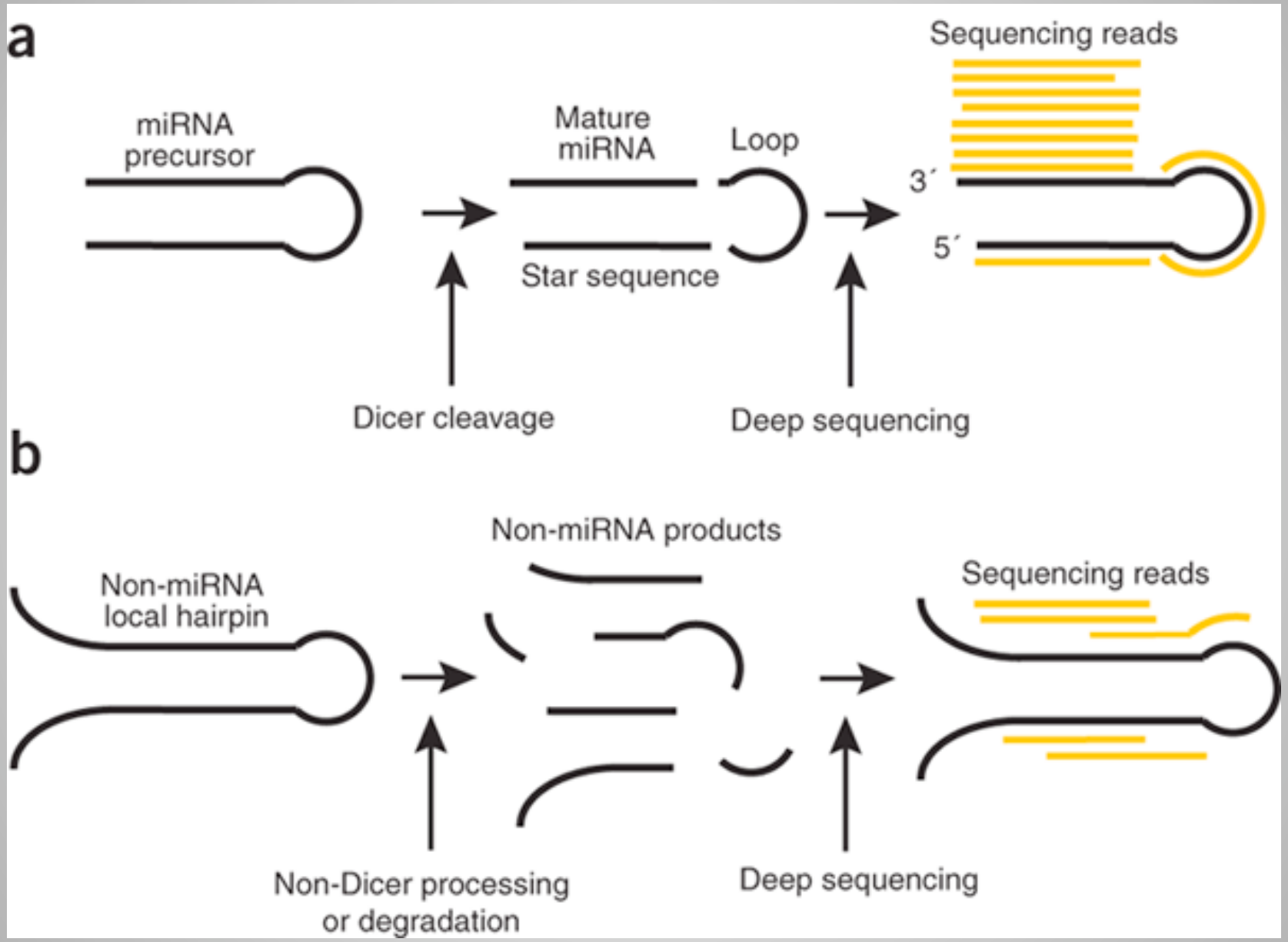
RNAplfold (L=120; W=120) & Nussinov (longest unbranched hairpin)



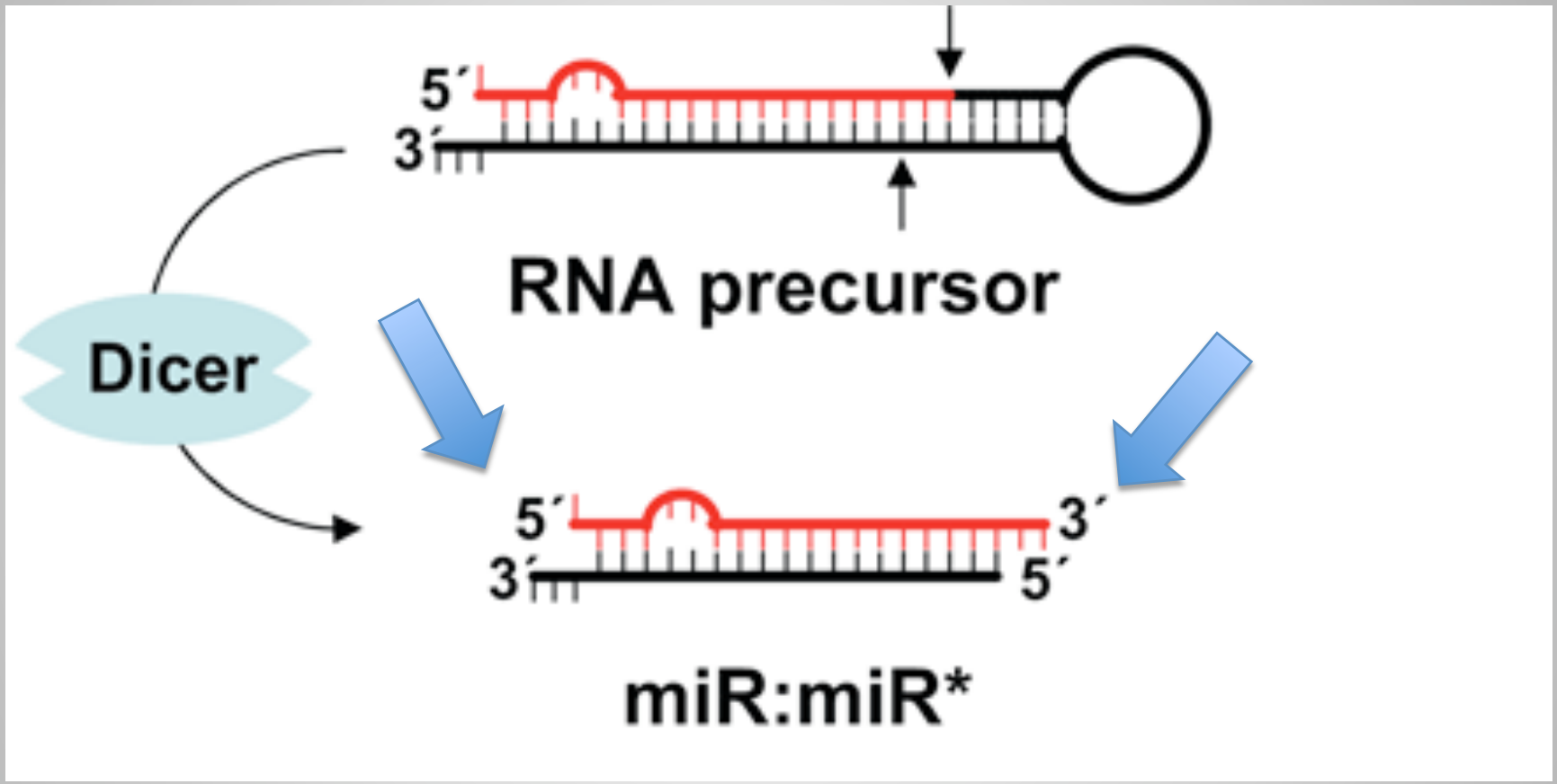
RNAfold



dicer cleavage



Friedländer 2008



Schwarz 2006

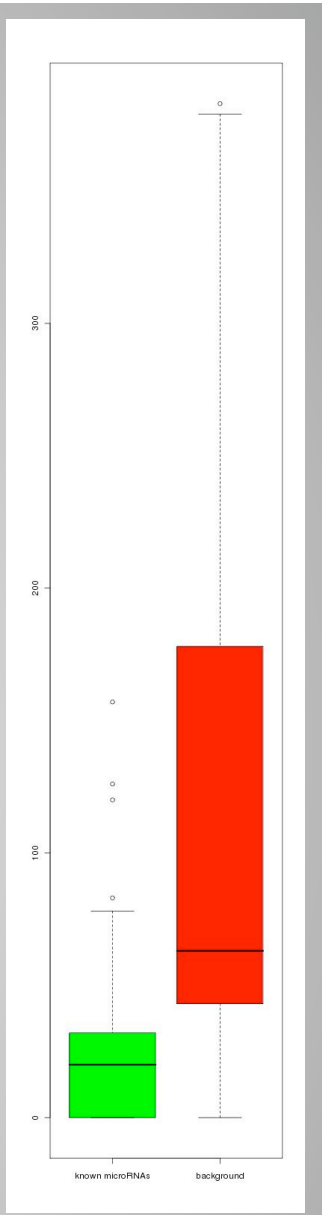
hairpin stability

((((.((((((((((((((..((((((((((....))))))))))))))))))))))).))))
..((((.((((.((((((((((((((..((((((((((....))))))))))))))))))))))).)))).)))).)
.((...((((.((((.((((((((((((((..((((((((((....))))))))))))))))))))))).)))).)))).)....)
....((((...((((.((((.((((((((((((((..((((((((((....))))))))))))))))))))))).)))).)))).)....))))))
(((.....(((...((((.((((((((((((((..((((((((((....))))))))))))))))))))))).))))))).....))))))

Stable hairpin structure

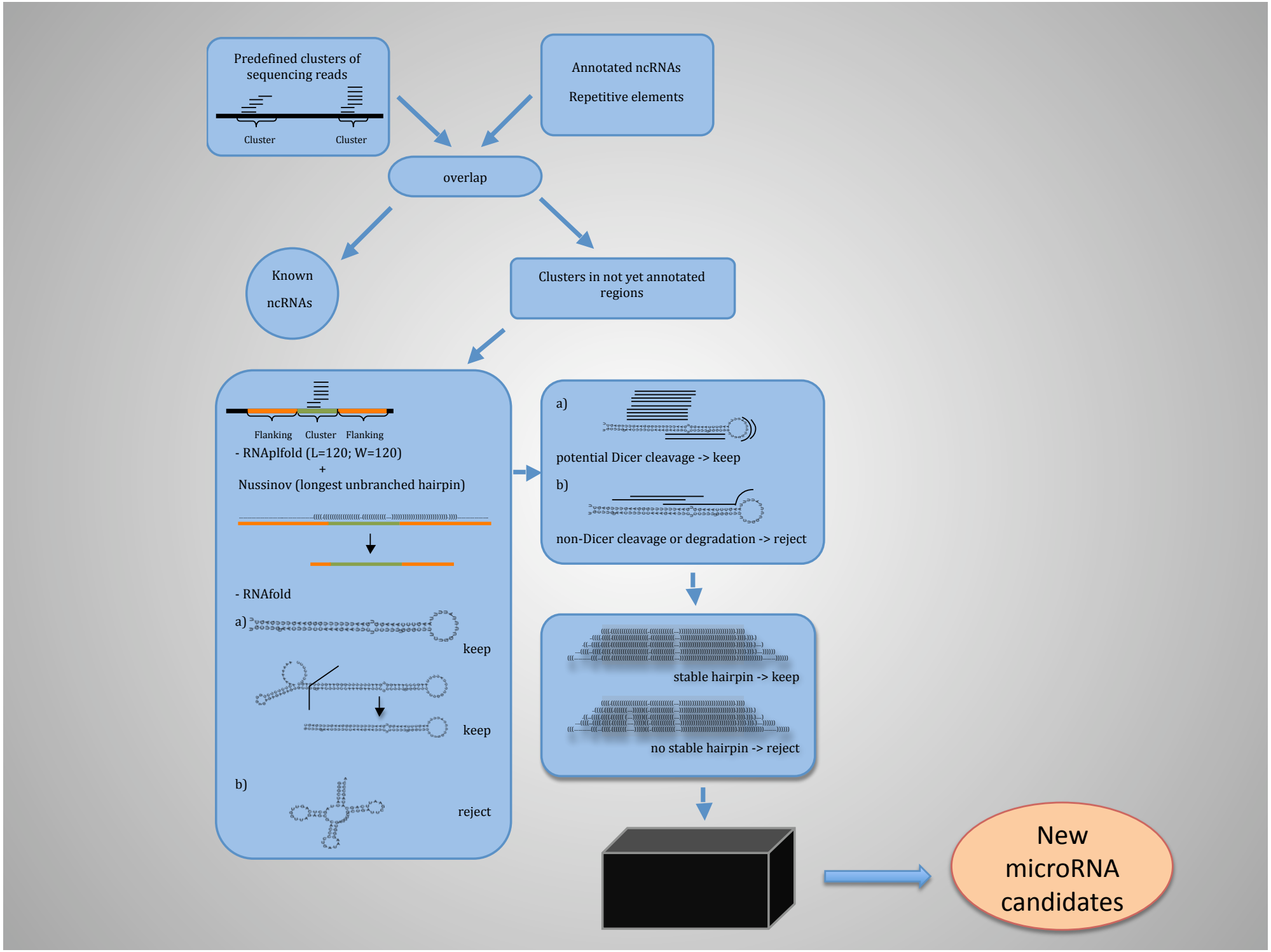
((((.((((((((((((((..((((((((((....))))))))))))))))))))))).))))
..((((.((((.((((((((((((((..((((((((((....))))))))))))))))))))))).)))).)))).)
.((...((((.((((.((((((((((((((..((((((((((....))))))))))))))))))))))).)))).)))).)....)
....((((((((((.....)))))))))).((((..((((((((((....))))))))))))))))))(((....))))))) ....))))))
.....((((((((((.....)))))))))). ((((..((((((((((....)))))))))))))))))) ((((....)))))) ....)))))(((..)))

No stable hairpin



P-value: 4e-7

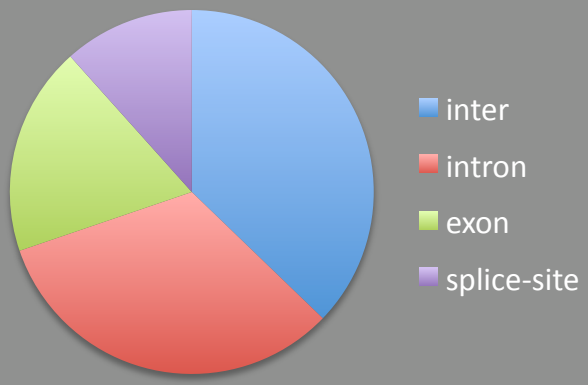
Workflow


Predefined clusters of sequencing reads
Cluster
Cluster
Annotated ncRNAs
Repetitive elements
overlap
Known ncRNAs
Clusters in not yet annotated regions
Flanking
Cluster
Flanking
- RNAplfold (L=120; W=120)
+
Nussinov (longest unbranched hairpin)
- RNAfold
a)
keep
keep
b)
reject
a)
potential Dicer cleavage -> keep
b)
non-Dicer cleavage or degradation -> reject
stable hairpin -> keep
no stable hairpin -> reject
New microRNA candidates

RNAmicro

Recall of known microRNAs (chrom1):
~50%

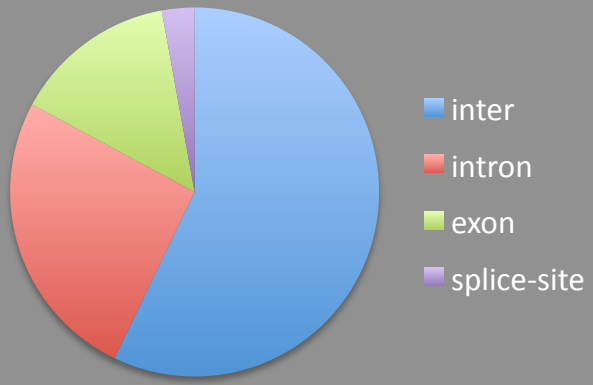
New candidates: **46**



tripletSVM

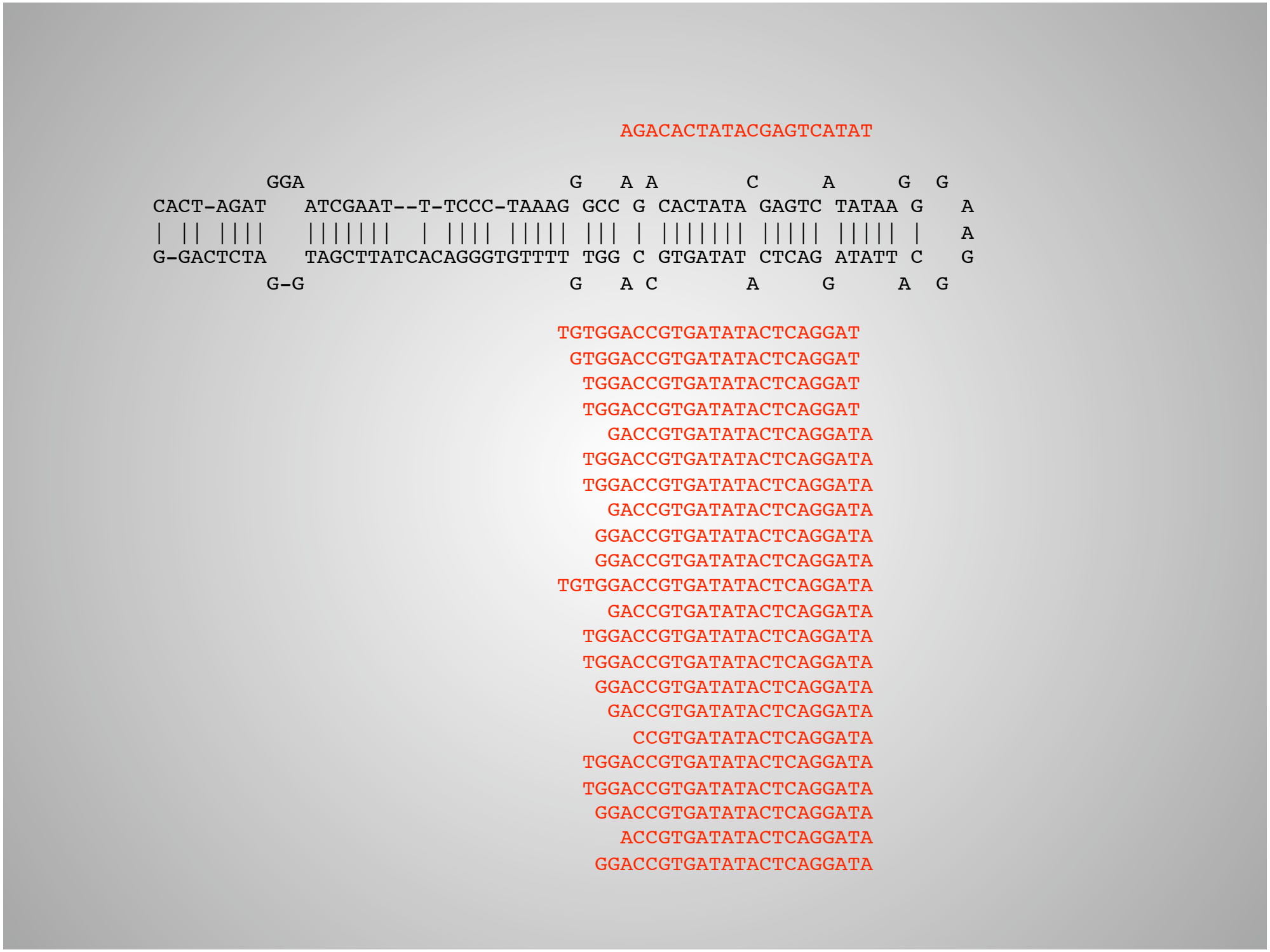
Recall of known microRNAs (chrom1):
~66%

New candidates: **41**



Example of novel microRNA candidate


>chr1:66866690-66866810 Cluster4847
TAGGACTCATATAGTGCCAGGTGT
TAGGACTCATATAGTGCCAGGT
TAGGACTCATATAGTGCCAGGT
TAGGACTCATATAGTGCCAGGTG
ATAGGACTCATATAGTGCCAG
ATAGGACTCATATAGTGCCAGGT
ATAGGACTCATATAGTGCCAGGT
ATAGGACTCATATAGTGCCAG
ATAGGACTCATATAGTGCCAGG
ATAGGACTCATATAGTGCCAGG
ATAGGACTCATATAGTGCCAGGTGT
ATAGGACTCATATAGTGCCAG
ATAGGACTCATATAGTGCCAGGT
ATAGGACTCATATAGTGCCAGGT
ATAGGACTCATATAGTGCCAGG
ATAGGACTCATATAGTGCCAG
ATAGGACTCATATAGTGCC
ATAGGACTCATATAGTGCCAGGT
ATAGGACTCATATAGTGCCAGGT
ATAGGACTCATATAGTGCCAGG
ATAGGACTCATATAGTGCCA
AGACACTATACGAGTCATAT
ATAGGACTCATATAGTGCCAGG
CACTAGATGGAATCGAATTTCCCTAAAGGGCCAGACACTATACGAGTCATATAAGGGAAGGCATTATAGGACTCATATAGTGCCAGGTGTTTTGTGGGACACTATTCGATGGATCTCAGG
(.(((((((...((((((((((((((((((.(((.(.(((((((.(((((.((((((.(.....).)))))).)))))).)))))))).).)))).)))))).))))).)..)))))))..)))))))


AGACACTATACGAGTCATAT
GGA G A A C A G G
CACT-AGAT ATCGAAT--T-TCCC-TAAAG GCC G CACTATA GAGTC TATAA G A
G-GACTCTA TAGCTTATCACAGGGTGTTTT TGG C GTGATAT CTCAG ATATT C G
G-G G A C A G A G
TGTGGACCGTGATATACTCAGGAT
GTGGACCGTGATATACTCAGGAT
TGGACCGTGATATACTCAGGAT
TGGACCGTGATATACTCAGGAT
GACCGTGATATACTCAGGATA
TGGACCGTGATATACTCAGGATA
TGGACCGTGATATACTCAGGATA
GACCGTGATATACTCAGGATA
GGACCGTGATATACTCAGGATA
GGACCGTGATATACTCAGGATA
TGTGGACCGTGATATACTCAGGATA
GACCGTGATATACTCAGGATA
TGGACCGTGATATACTCAGGATA
TGGACCGTGATATACTCAGGATA
GGACCGTGATATACTCAGGATA
GACCGTGATATACTCAGGATA
CCGTGATATACTCAGGATA
TGGACCGTGATATACTCAGGATA
TGGACCGTGATATACTCAGGATA
GGACCGTGATATACTCAGGATA
ACCGTGATATACTCAGGATA
GGACCGTGATATACTCAGGATA

Outlook

microRNA detection:

- new machine learning approach
- new features

ncRNA Classification:

- using locaRNA to search for similar structures as tRNA, snRNA, etc.
- include preprocessed genome-wide predictions for snoRNAs and snRNAs, etc.

New machine learning approach for Deep-Sequencing data

Features:

-Structure
- -Energy
- -Loop
- -5‘ and 3‘ Stems
- -Bulges

-Position of reads
- -Dicer cleavage (2nt overhang, valid products)
- -Distance to loop
- -Cluster size (shifted reads)

-Hairpin stability

- .....

Outlook

microRNA detection:

- new machine learning approach
- new features

ncRNA Classification:

- using locaRNA to search for similar structures as tRNA, snRNA, etc.
- include preprocessed genome-wide predictions for snoRNAs and snRNAs, etc.

Thanks To

Uni Leipzig

Peter F. Stadler
Clara I. Bermudez Santana (clustering, known ncRNAs)
Jana Hertel (microRNAs, RNAmicro)
Steve Hoffmann (segemehl)

TU Munich

Martin Sturm (machine learning)