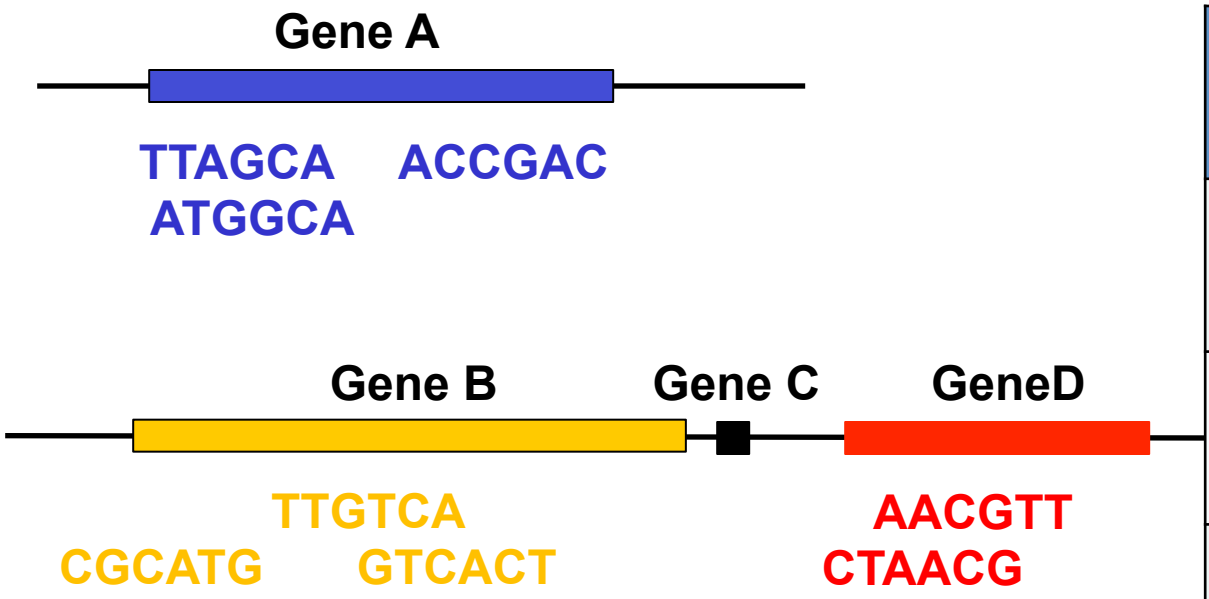
Next-generation Sequencing

Lecture 10

Align reads to Genome and count



Gene ID	Sample1
A	3
B	3
C	0
D	2

For a given gene, the number of reads aligned to the gene measures its expression level.

Differentially expressed gene Analysis Tools

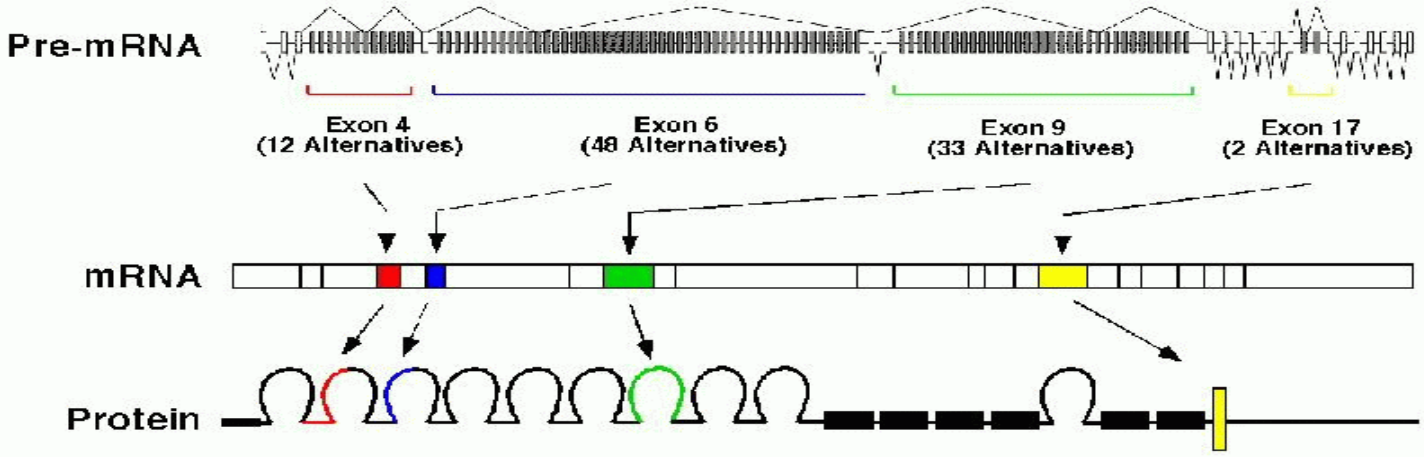
Tools	Statistics			speed
edgeR	Empirical Bayes estimation and exact tests based on the negative binomial distribution	Robinson et al., 2010	High TPR	media
DEseq	Negative binomial distribution.	Anders and Huber, 2010	Low TPR	media
NOISeq	Compares replicates within the same condition to estimate noise distribution of M (log-ratio) and D (absolute value of the difference).	Tarazona et al., 2011	High TPR	Data size
baySeq	Empirical Bayesian methods using the negative binomial distribution.	Hardcastle and Kelly, 2010		slow
TSPM		Auer and Doerge, 2011	Data size	media
BitSeq	a hierarchical log-normal model and determines the probability of differential expression by Bayesian model averaging	Glaus et al., 2012		
POME	Poisson mixed-effects model	Hu et al., 2012		

Applications of RNA-seq

- Gene expression
  - Expression of individual genes/loci
  - Quantitatively discriminate isoforms using junction reads and coverage of individual exons, introns, etc.
- Annotation
  - New features of the transcriptome: genes, exons, splicing, ncRNAs
- SNP
- Fusion gene detection

Alternative Splicing (AS)

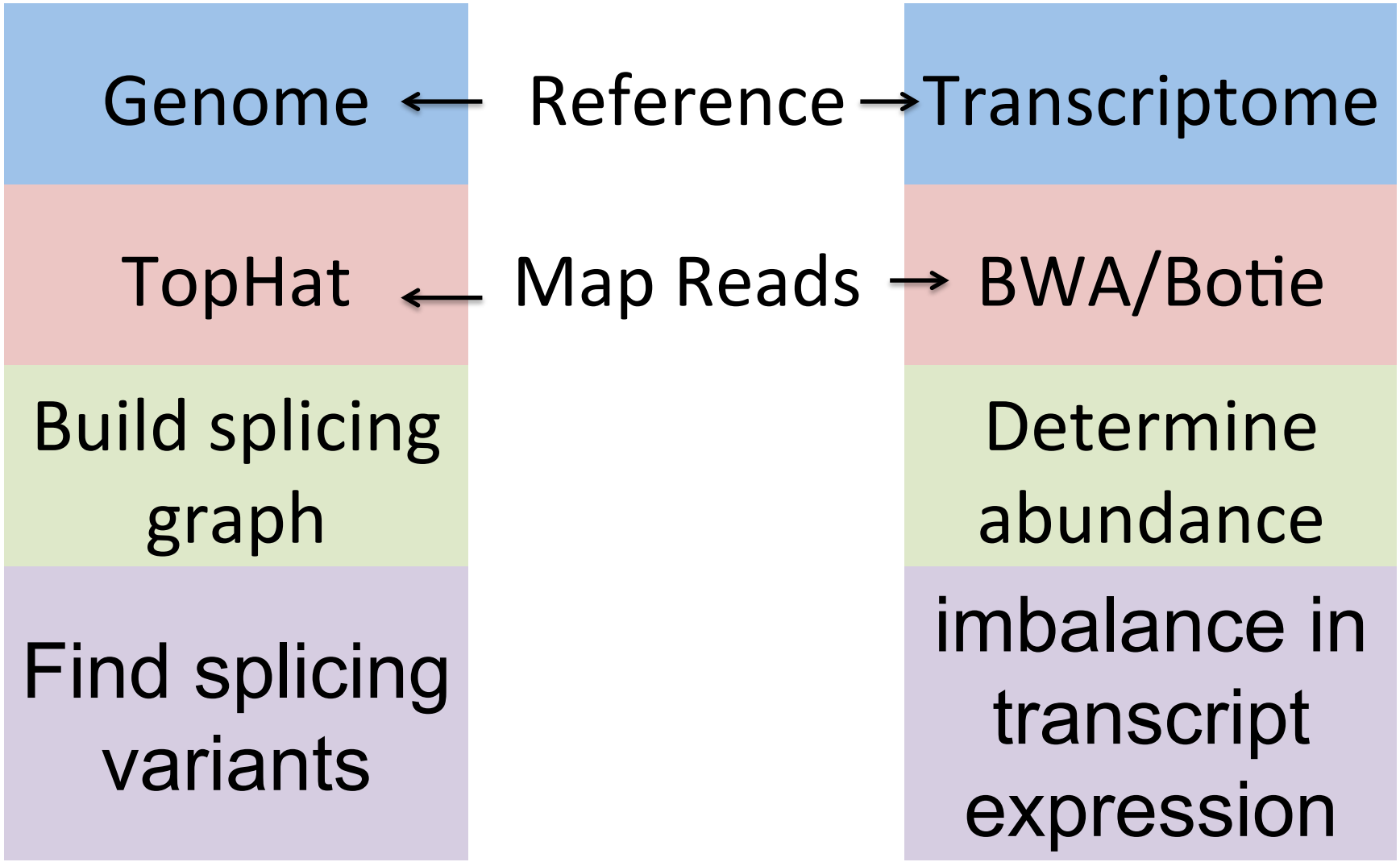
- 35% - 60% of human genes show AS
- process by which the gene's exons are pieced together in multiple ways forming mRNA during the RNA splicing.
- some genes have a huge number of isoforms (*slo* >500, *neurexin* >1000, *DSCAM* > 38000)



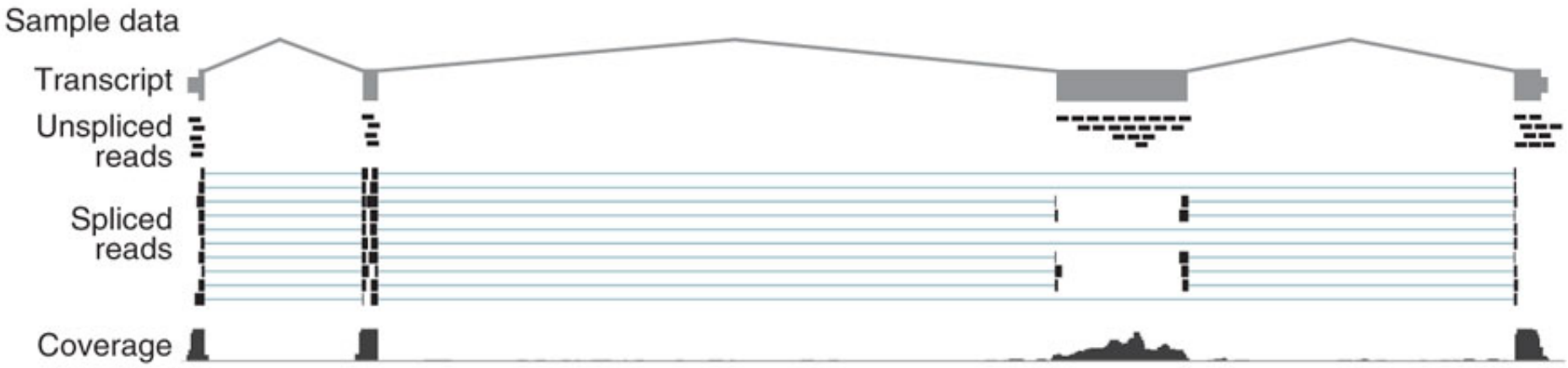
DSCAM axon guidance receptor

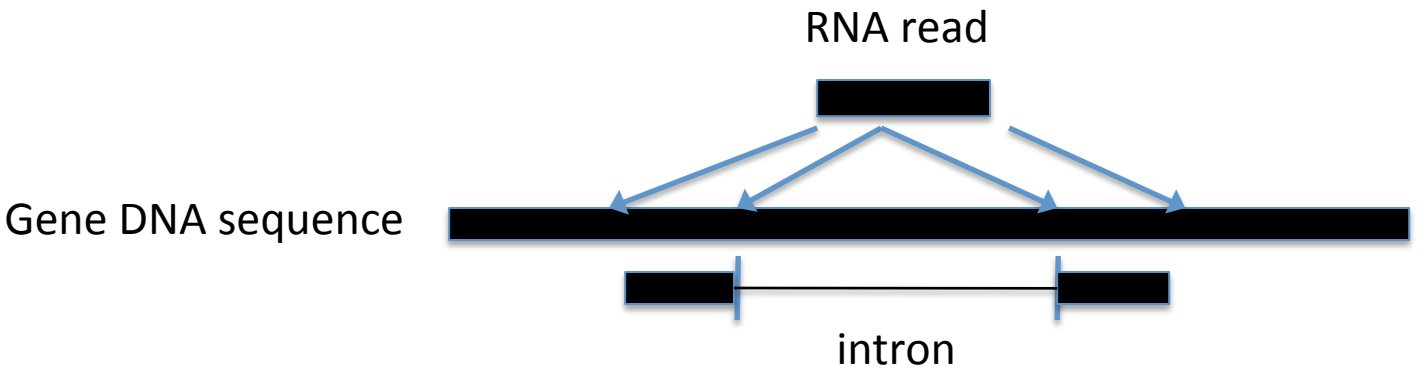
Taken from [**Graveley, 2001**]

Splicing site discovery pipelines

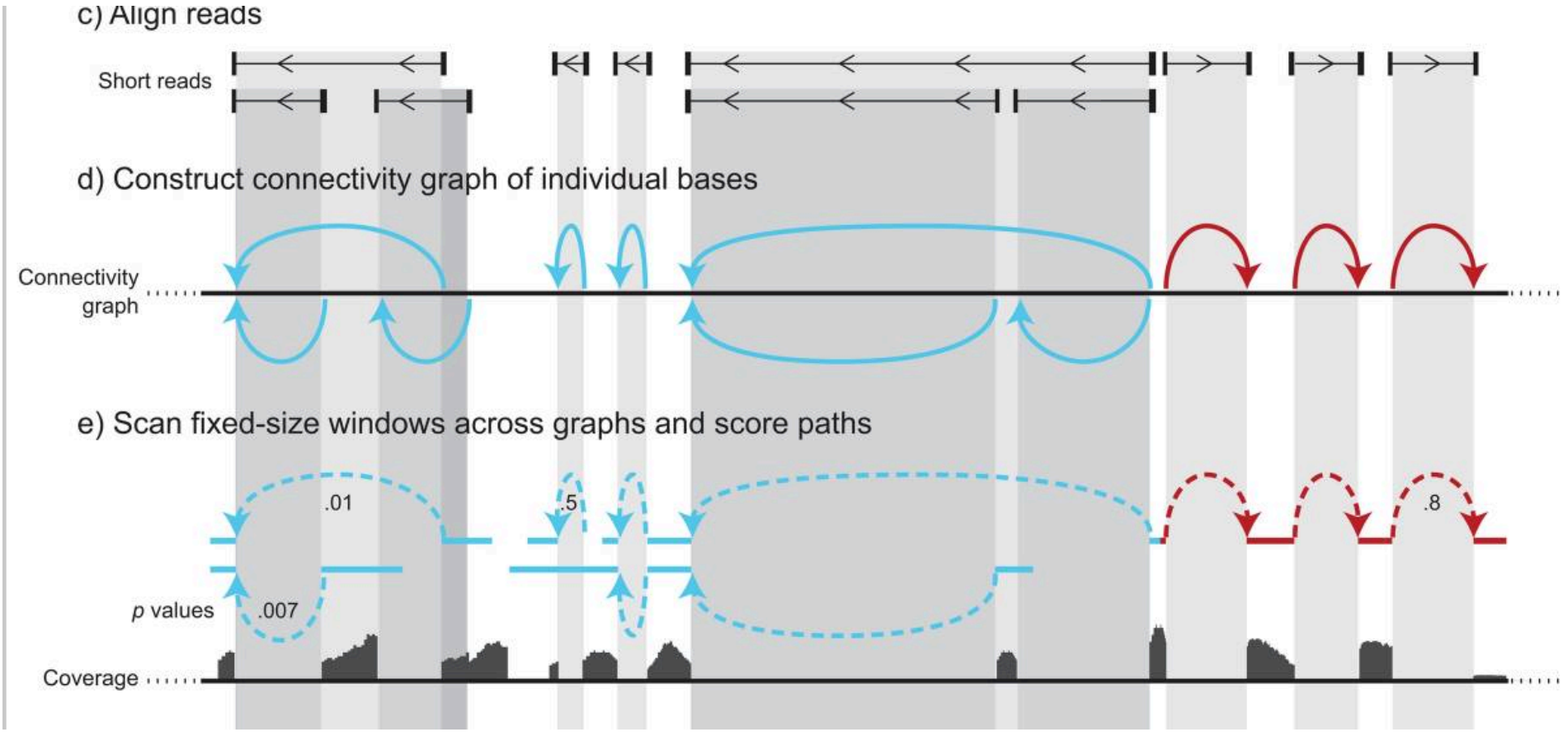


Splicing site discovery gapped mapping



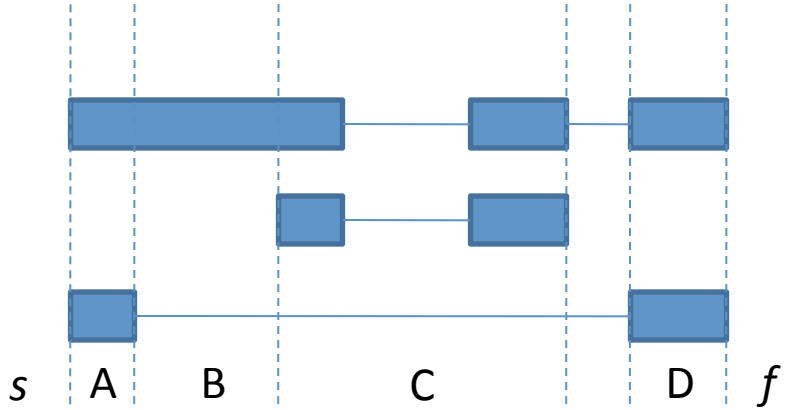


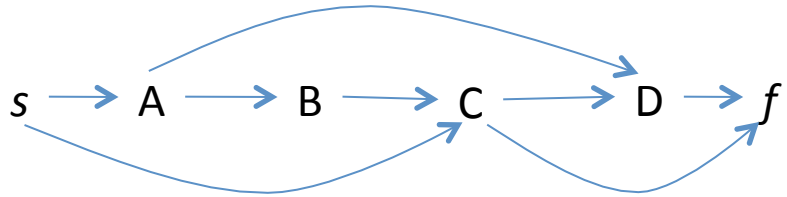
Splicing site and isoform discovery Using exon-graph



Guttman et al. Nature Biotechnology, 2010

Splicing graph and splicing variants





An edge in the splicing graph, called a *block*, represents a maximal sequence of adjacent exons or exon fragments that always appear together in a given set of splicing variants. Therefore, variants can be represented by sequence of blocks, e.g. {ABCD, C, AD}.

Vertices *s* and *f* are included into graph, and are linked to the 5' and 3' of each variant, respectively. Each splicing variant corresponds to a directed path that goes from *s* to *t*. But note that some paths in the splicing graph do not correspond to real variants, e.g. {ABC, CD}.

Larcroix, et. al. WABI, 2009

non-coding RNA

- Messenger RNA (mRNA) is the RNA that carries information from DNA to the ribosome. The coding sequence of the mRNA determines the amino acid sequence in the protein that is produced.
- Sequencing of the human genome showed that there are only ~20,000 protein-coding genes, representing <2% of the total genomic sequence .
- Many RNAs do not code for protein and these so-called non-coding RNAs ("ncRNA") can be encoded by their own genes (RNA genes), but can also derive from mRNA introns.
- Non coding RNA -- Highly abundant and functionally important

Types of non-coding RNA

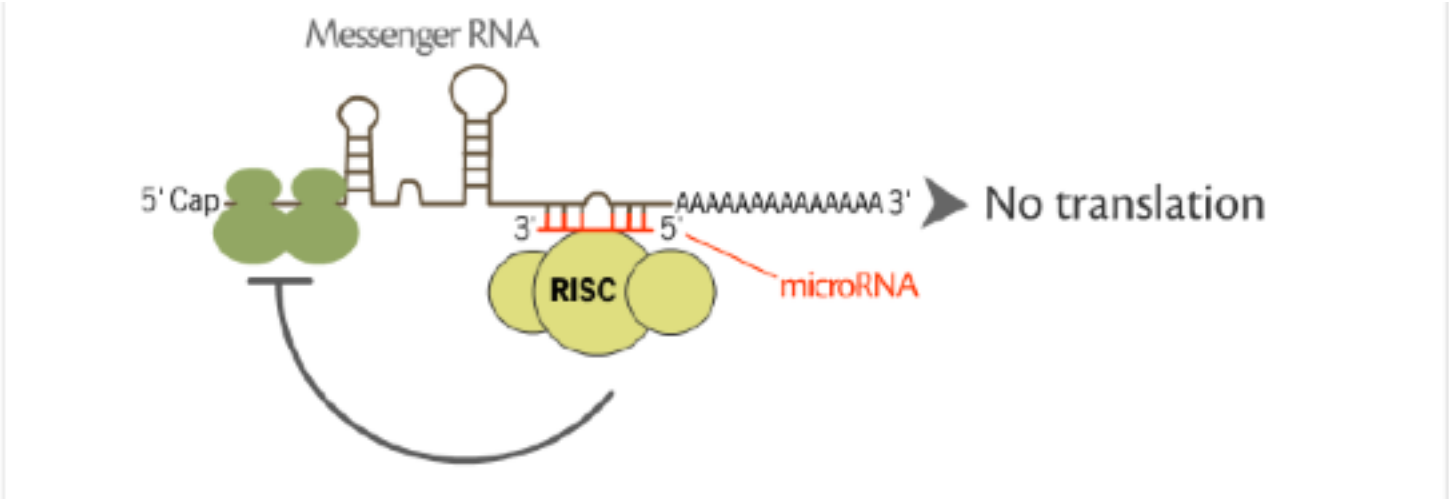
- transfer RNA (tRNA) and ribosomal RNA (rRNA),
- snRNAs- Small nuclear ribonucleic acid
- snoRNAs-Small nucleolar RNA
- Small RNAs (21-26nt): microRNAs, siRNAs, stRNAs, tony noncoding RNAs etc.
- long ncRNAs-Long non coding RNAs
- exRNAs-Extracellular RNA

Small RNAs

- Small regulatory RNAs can control mRNA stability or translation, or target epigenetic modifications to specific regions of the genome.
- MicroRNAs (miRNAs) are another class of endogenous small noncoding RNA molecules that regulate eukaryotic gene expression at posttranscriptional level.

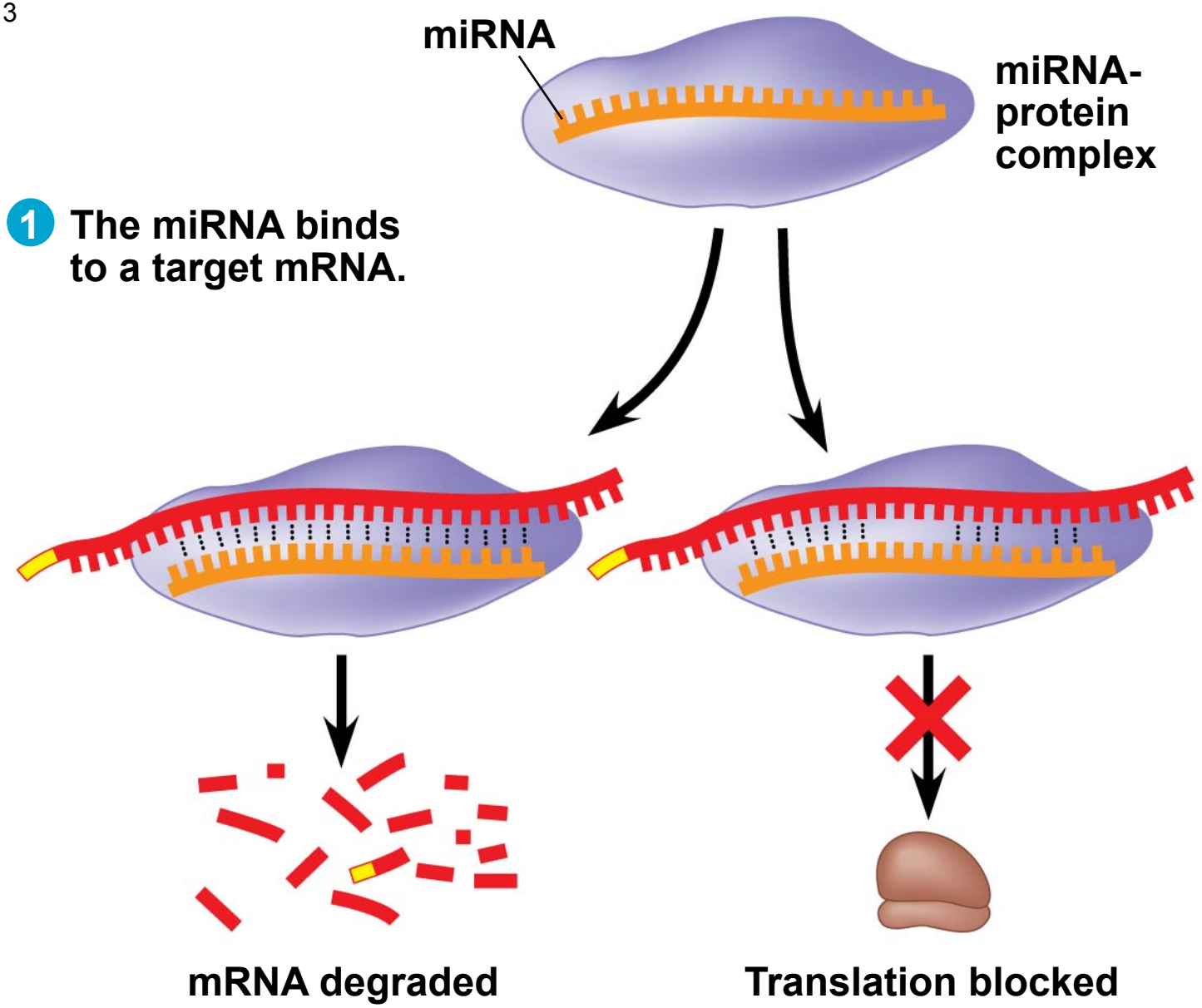
MicroRNA

- **miRNAs** are small single-stranded RNA (22 nt) molecules that can bind to complementary mRNA sequences



RISC = RNA-induced Silencing Complex

Figure 15.13

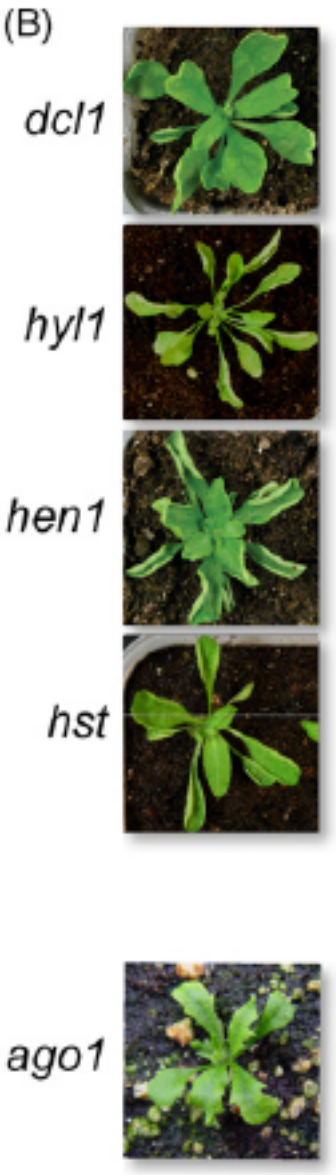

miRNA
miRNA-protein complex
1 The miRNA binds to a target mRNA.
mRNA degraded
Translation blocked
2 If bases are completely complementary, mRNA is degraded. If match is less than complete, translation is blocked.

miRNAs have diverse functions in animals

- Development:
  - Brain development (miR-430)
  - Muscle (miR-1)
  - Heart development (miR-1)
  - neuronal development (miR-124)
- Metabolism:
  - misregulation of miRNA causes metabolic disorders
- Immuno responses:
  - miRNA function as positive and negative regulator.
- Cancer:
  - miRNAs function as oncogenes or tumour suppressors
- Viral infection:
  - suppressor or enhancer

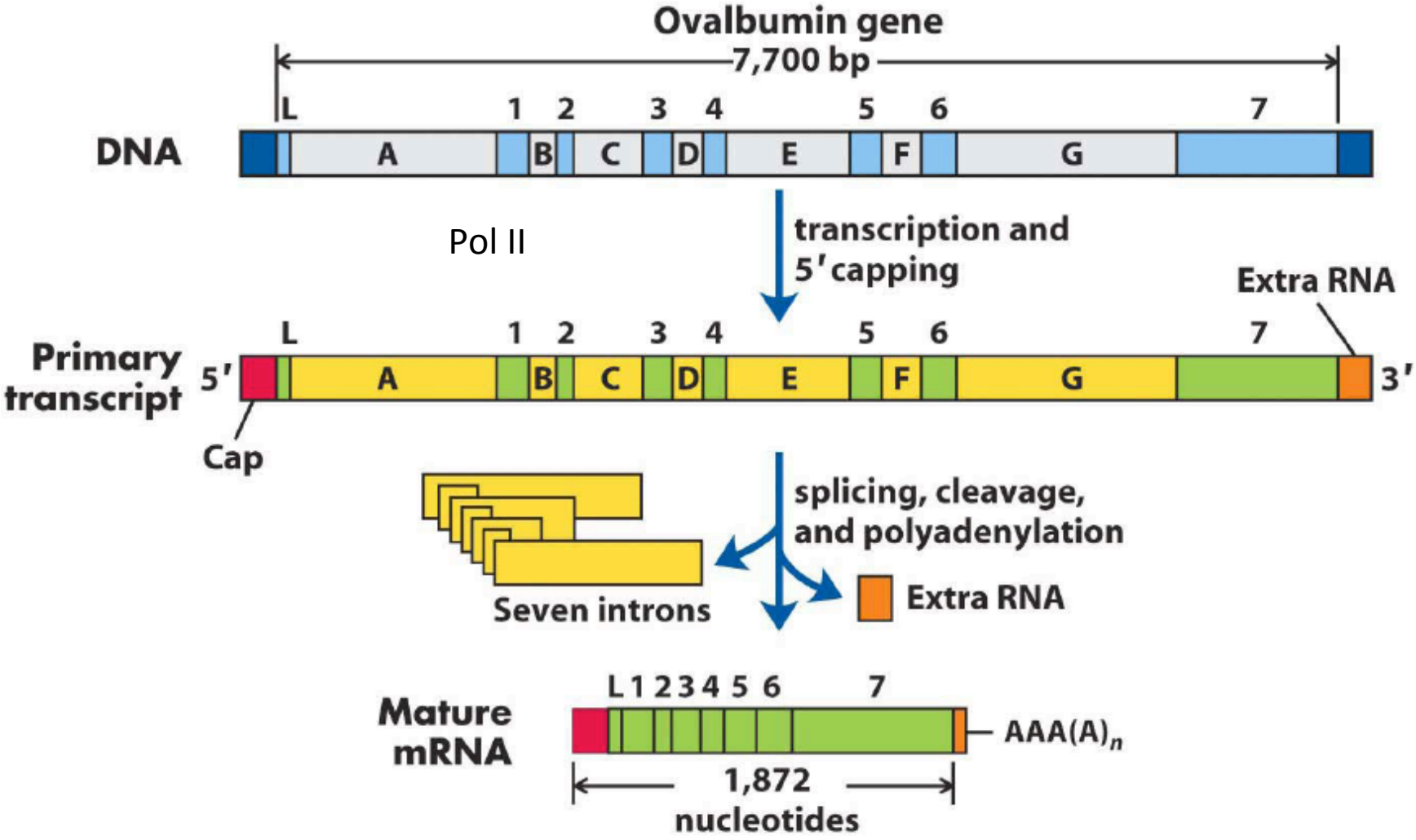
miRNAs have diverse functions in plants

- Development:
  Organ identity
- Stress responses (abiotic and biotic stress)
- Hormone biogenesis and signaling
- Metabolism

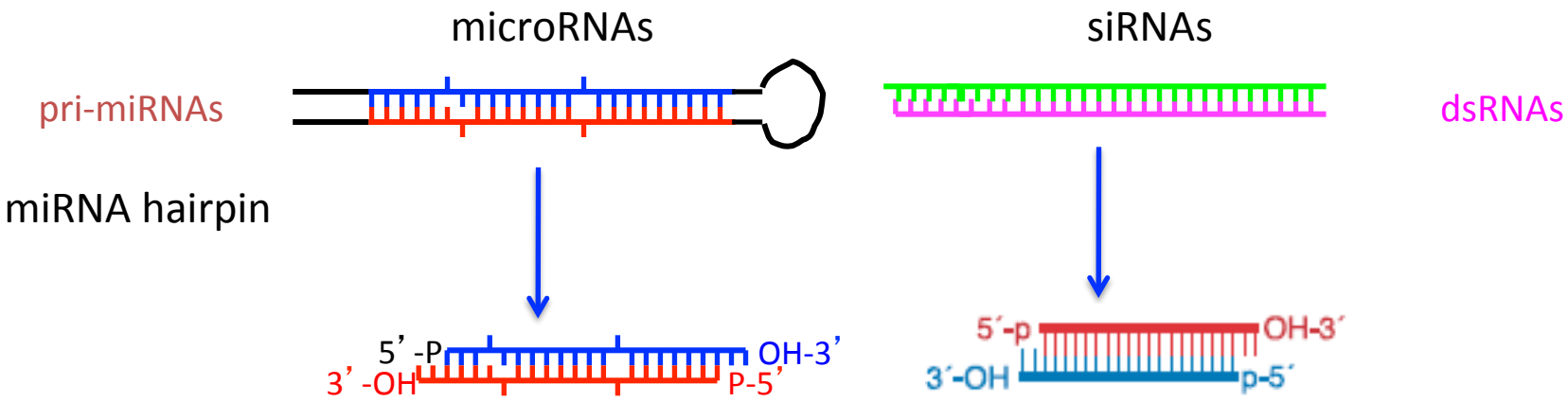


miRNA synthesis

Like mRNA, miRNA precursor is predominantly transcribed by Pol II



Structure of miRNA/miRNA* and siRNA duplex



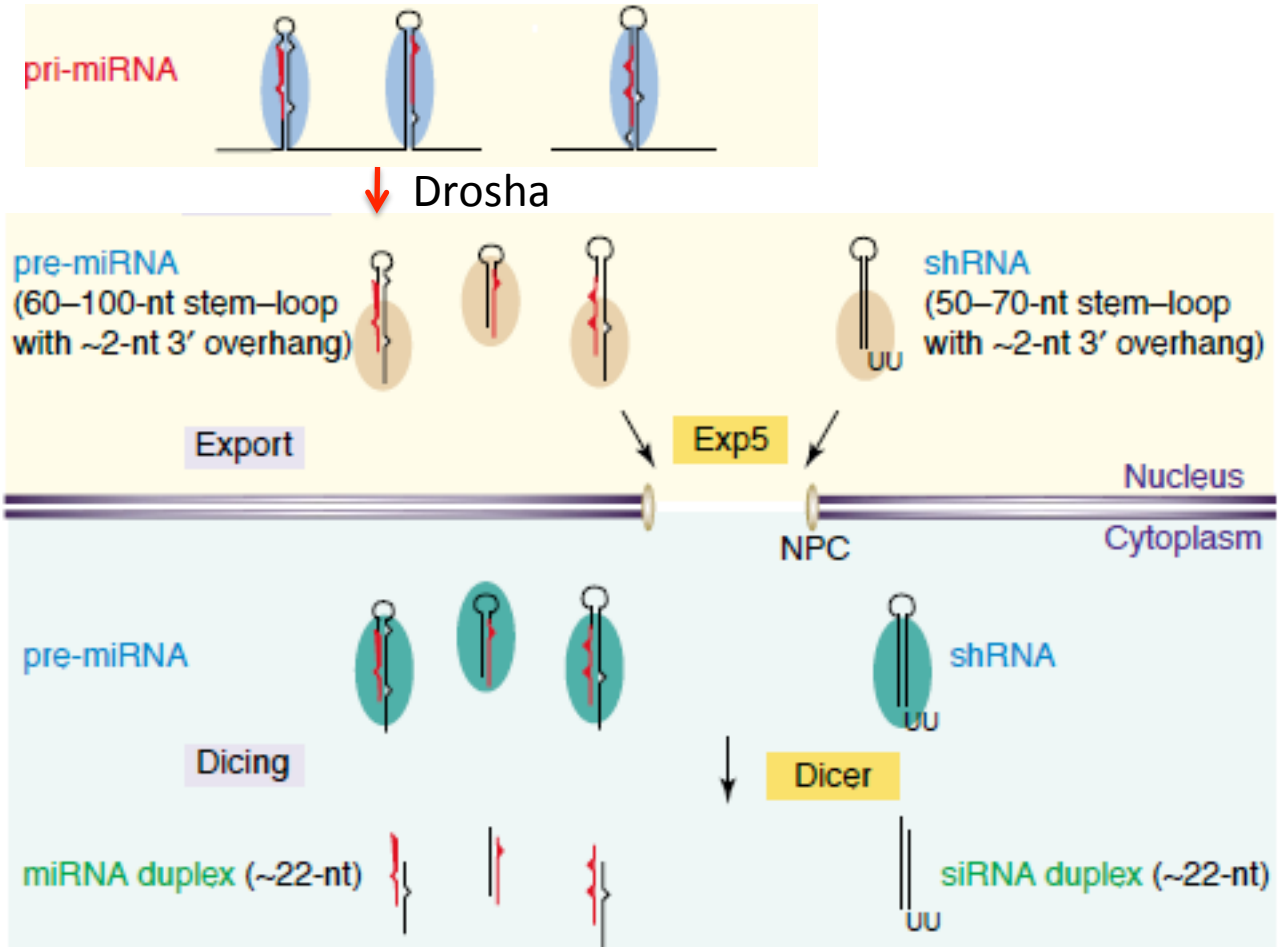
Cloning and sequencing small RNAs established that they are generated in a duplex form

The duplex has 2 nt overhang at 3’ end, 5’ phosphate at each strand. 3’ OH for miRNAs and most of siRNAs in many organisms but not in plants,

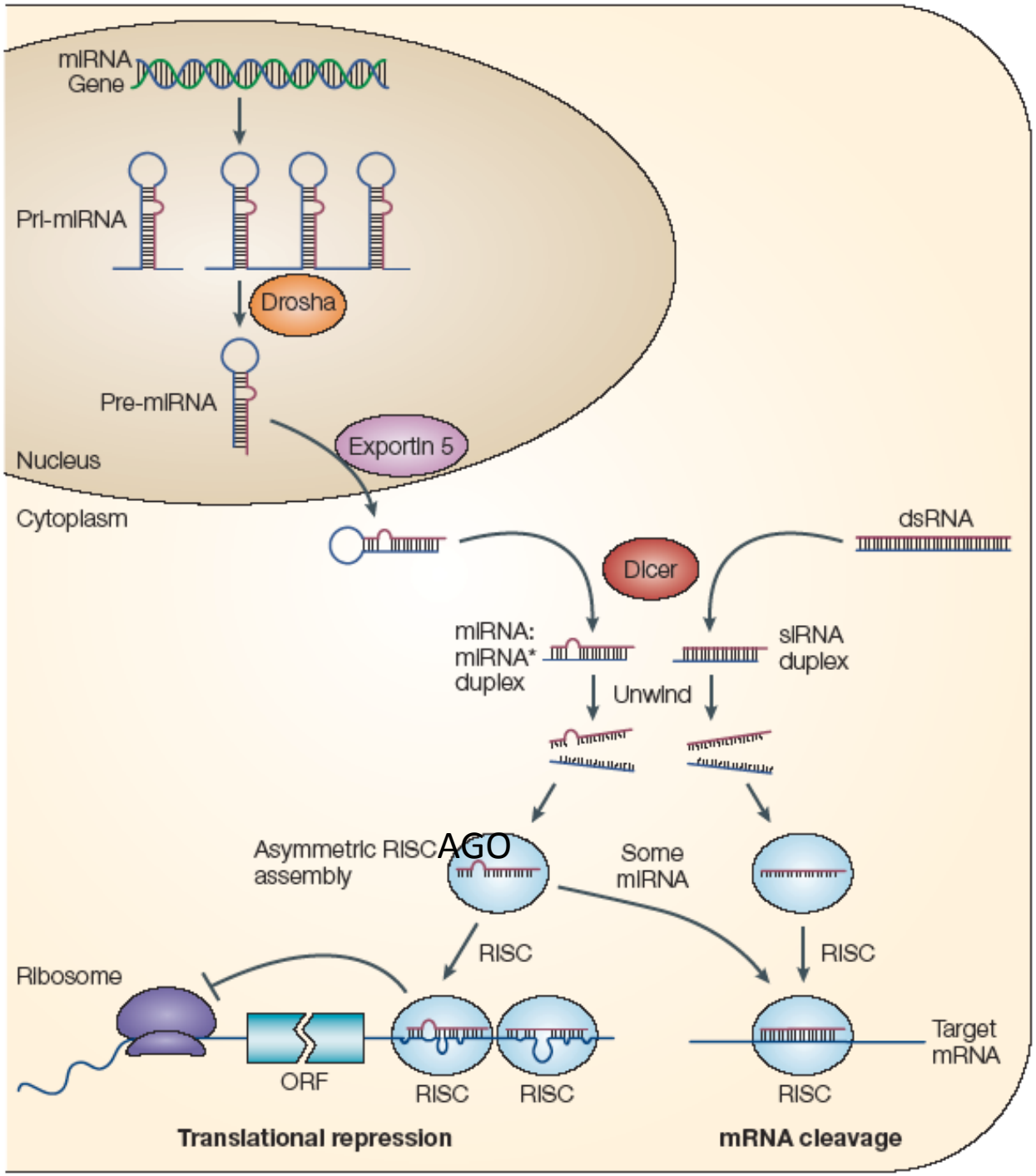
miRNA dupelx: generated from imperfect match stem loop transcripts by pol II or pol III

siRNA duplex: near perfect match dsRNAs generated from transgene, transposon, repeated-DNA or exogenous dsRNA

Model for miRNA biogenesis



Drosha processes pri-miRNA in nucleus to pre-miRNAs of ,70-nt, which are exported by Exp5. Upon export, Dicer participates in the second step (dicing) to produce miRNA duplexes.



The functional mechanism of miRNAs

Nature Reviews Genetics 5:522-531 (2004).

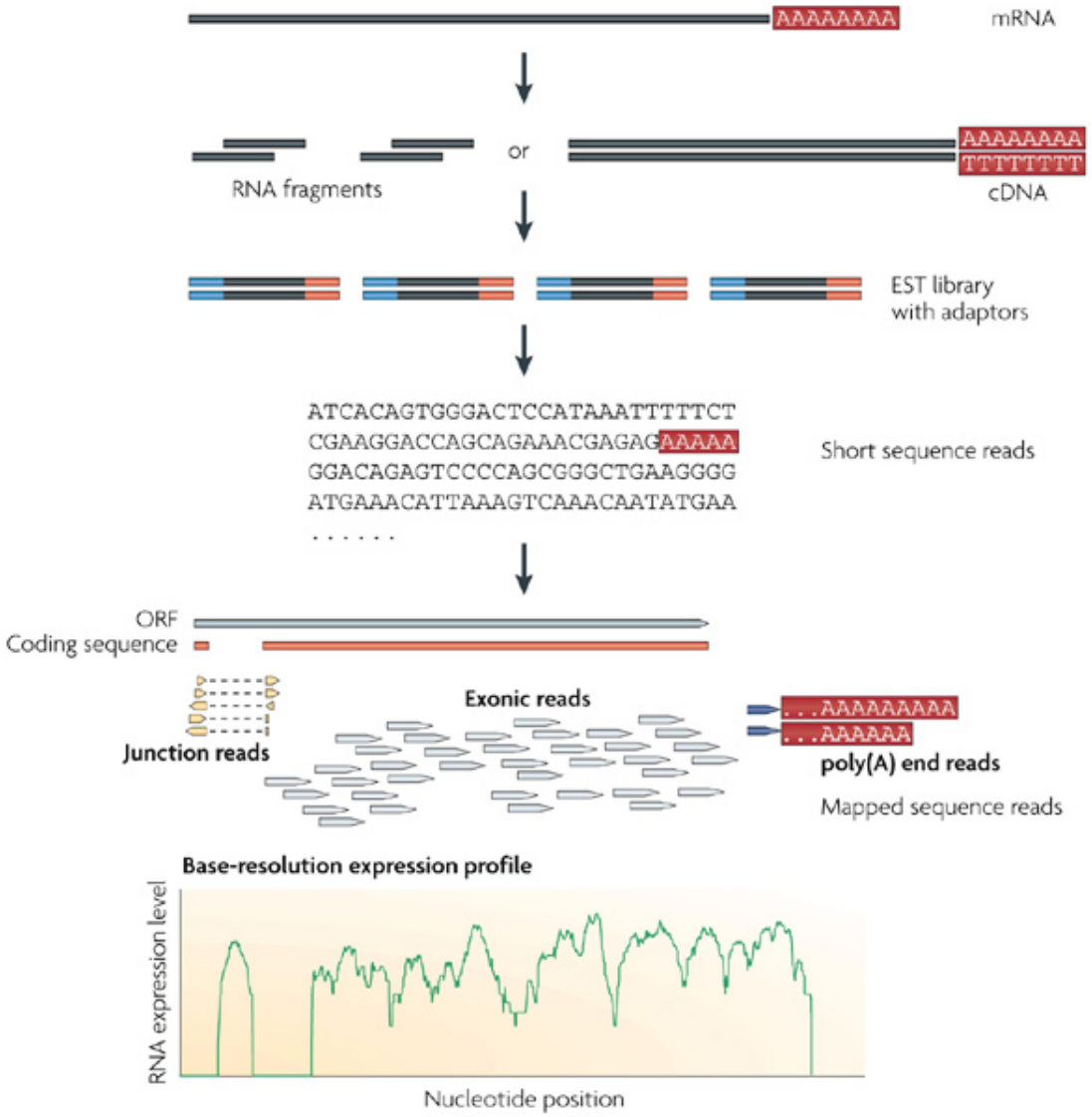
After generation, the miRNA strand of miRNA duplex is loaded into the RNA-induced silencing complex (RISC) to repression gene expression by translational inhibition or target cleavage. The catalytic component of RISC is the Argonaute protein.

RNA-seq Exp.

Extract sufficient mRNA from total using either poly-A selection or depletion of rRNA (RiboMinus).

Non-poly(A) RNA can yield important noncoding RNA gene discovery

reads are aligned with the reference genome

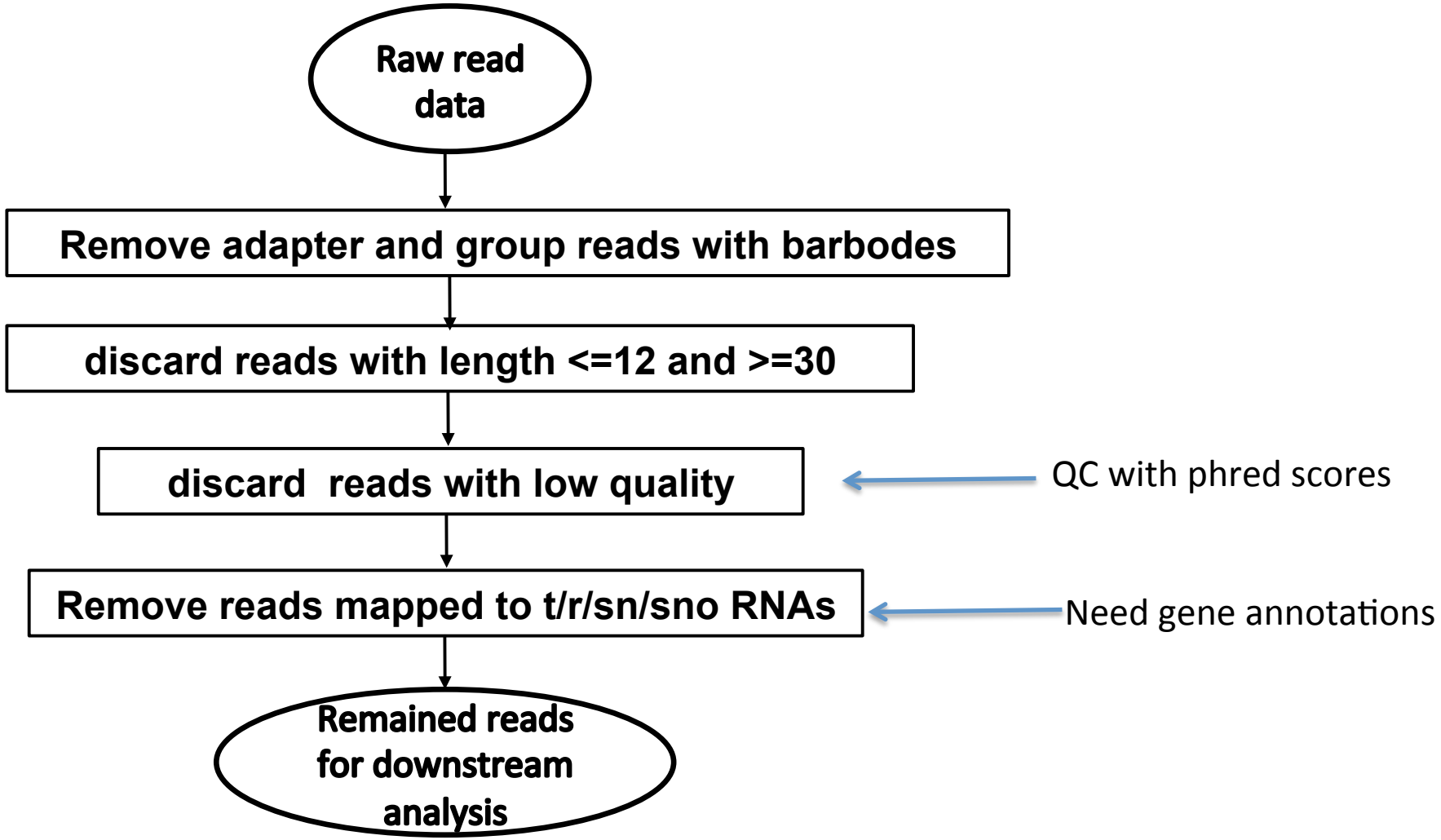


Wang, Gerstein, Snyser, 2009

miRNA-seq Data analysis

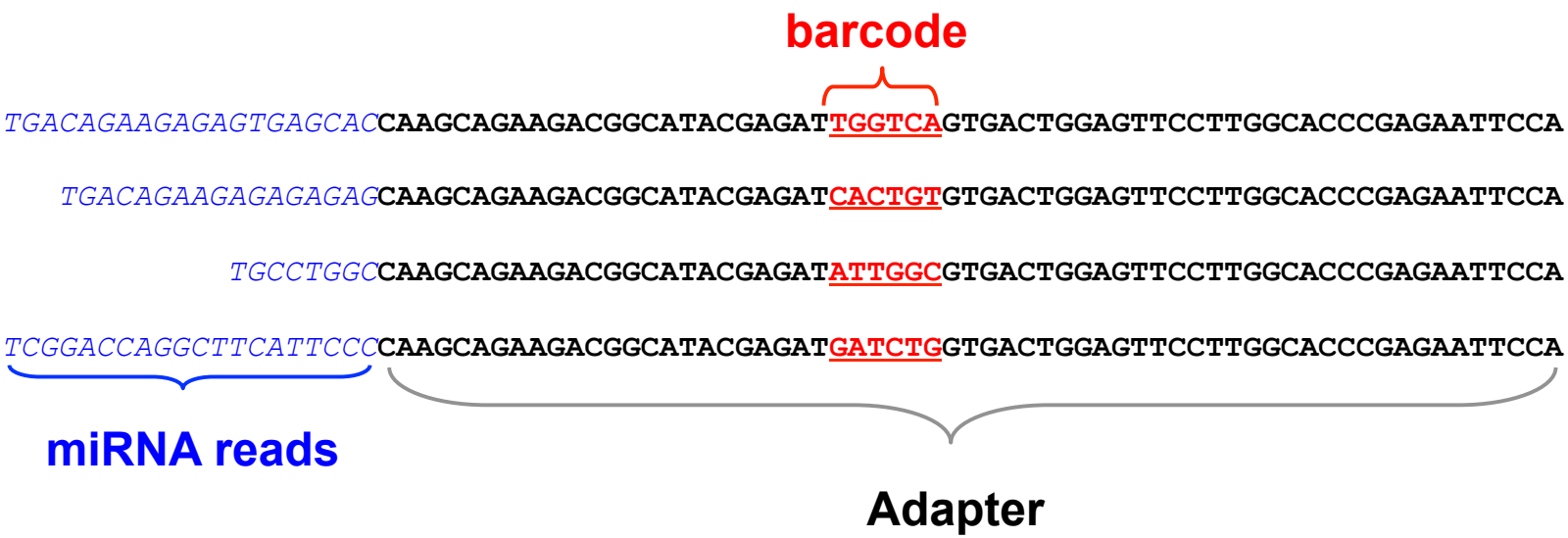
- Preprocessing
- Abundance analysis
- Imprecision analysis
- miRNA trimming and tailing analysis

Preprocessing



Preprocessing

An example of raw miRNA-seq data filtering

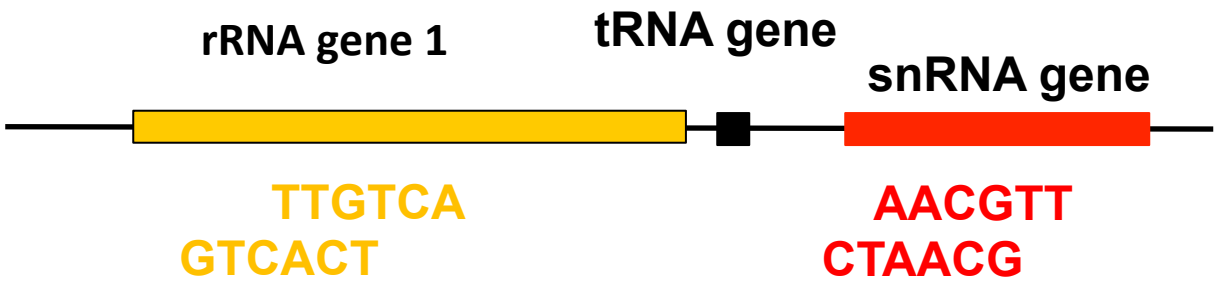


Useful tools for preprocessing

- Fastx Toolki
  http://hannonlab.cshl.edu/fastx_toolkit
  - Split by barcodes: fastx_barcode_splitter.pl
  - Remove adapters: fastx_clipper
  - removal of poor-quality reads: fastq_quality_filter
- FastQC:

http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

Remove reads mapped to t/r/sn/sno RNAs



Need to map reads to reference genome, and identify t/r/sn/sno RNA genes with gene annotation information (GFF files).

A typical results after preprocessing

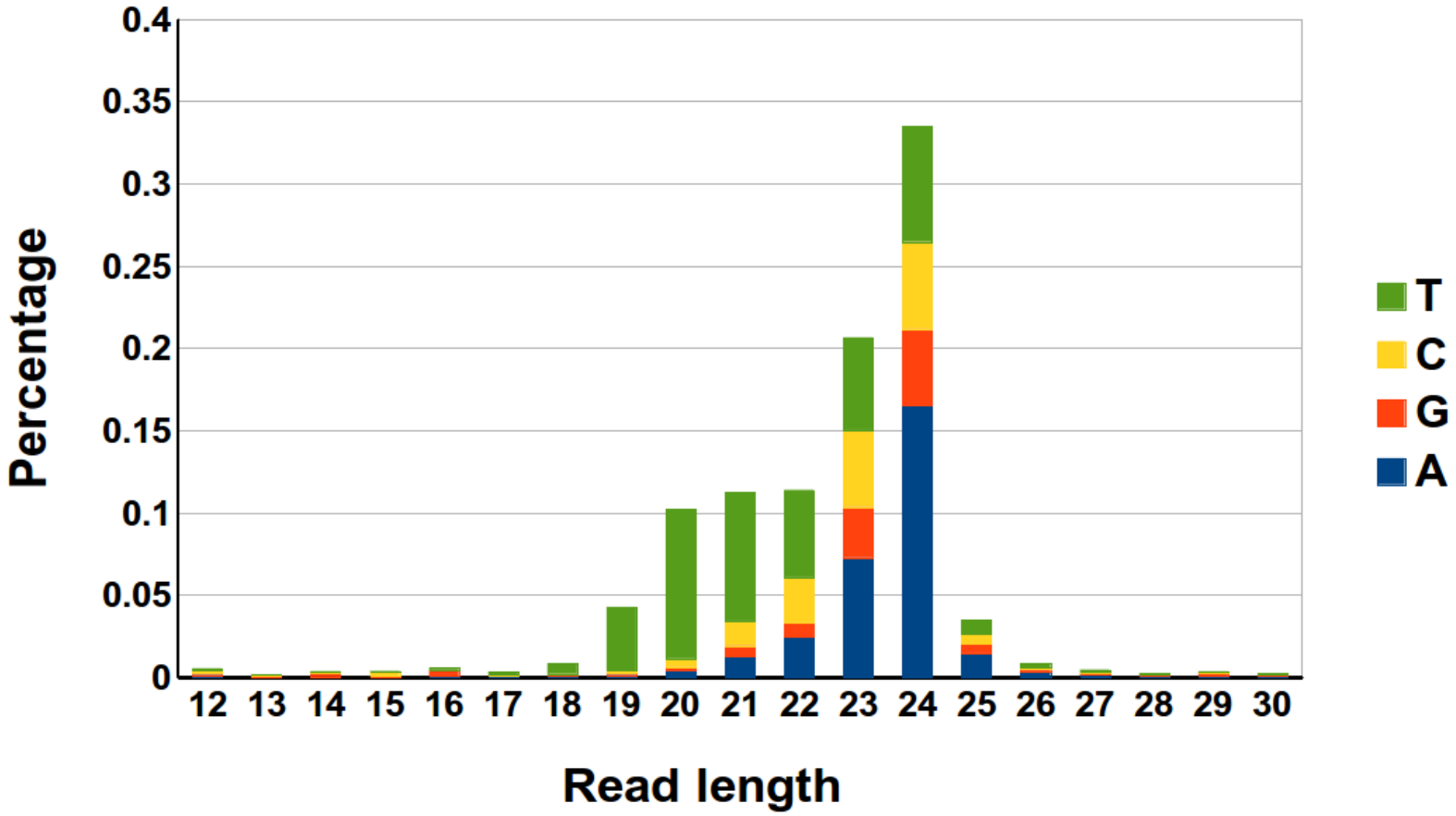
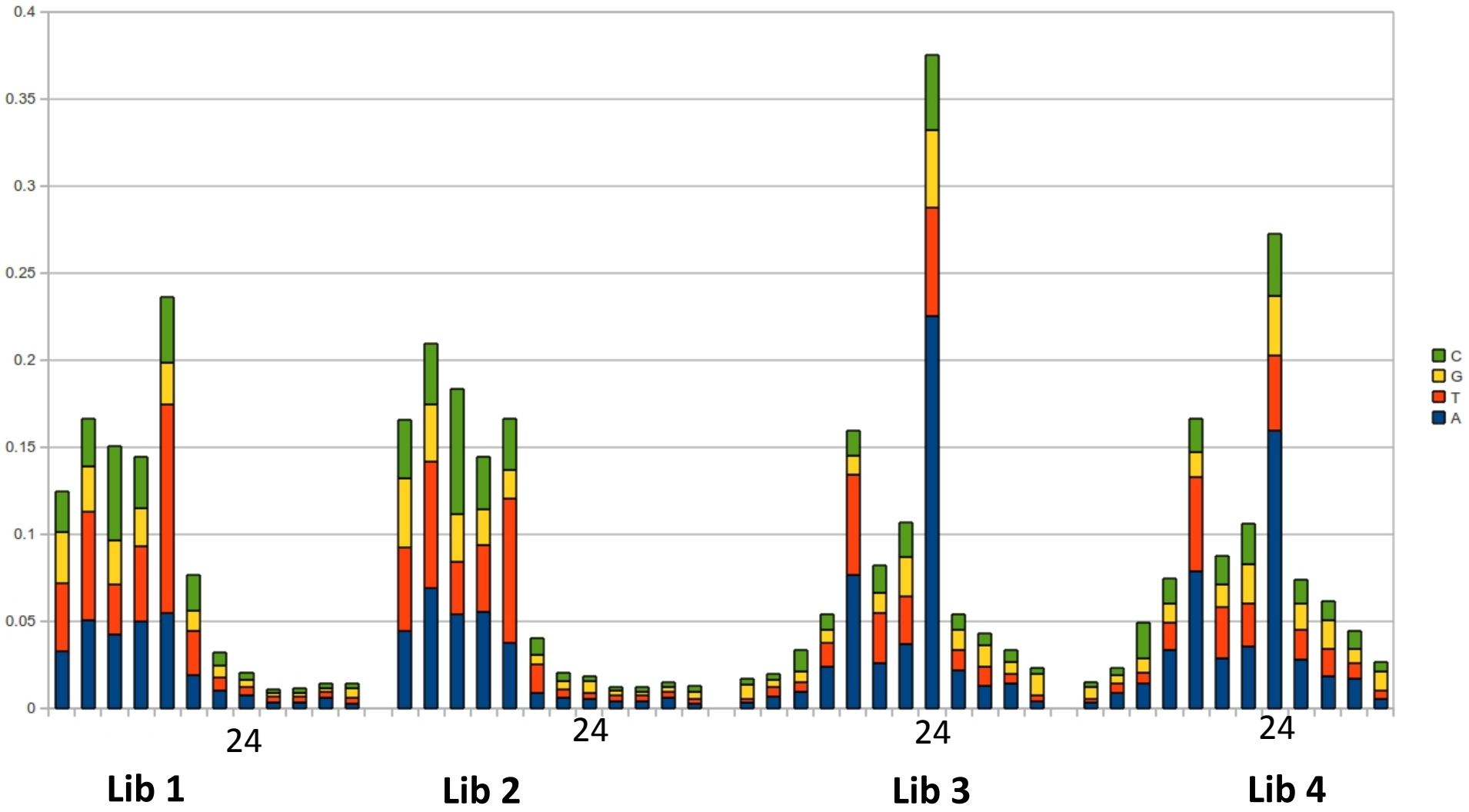
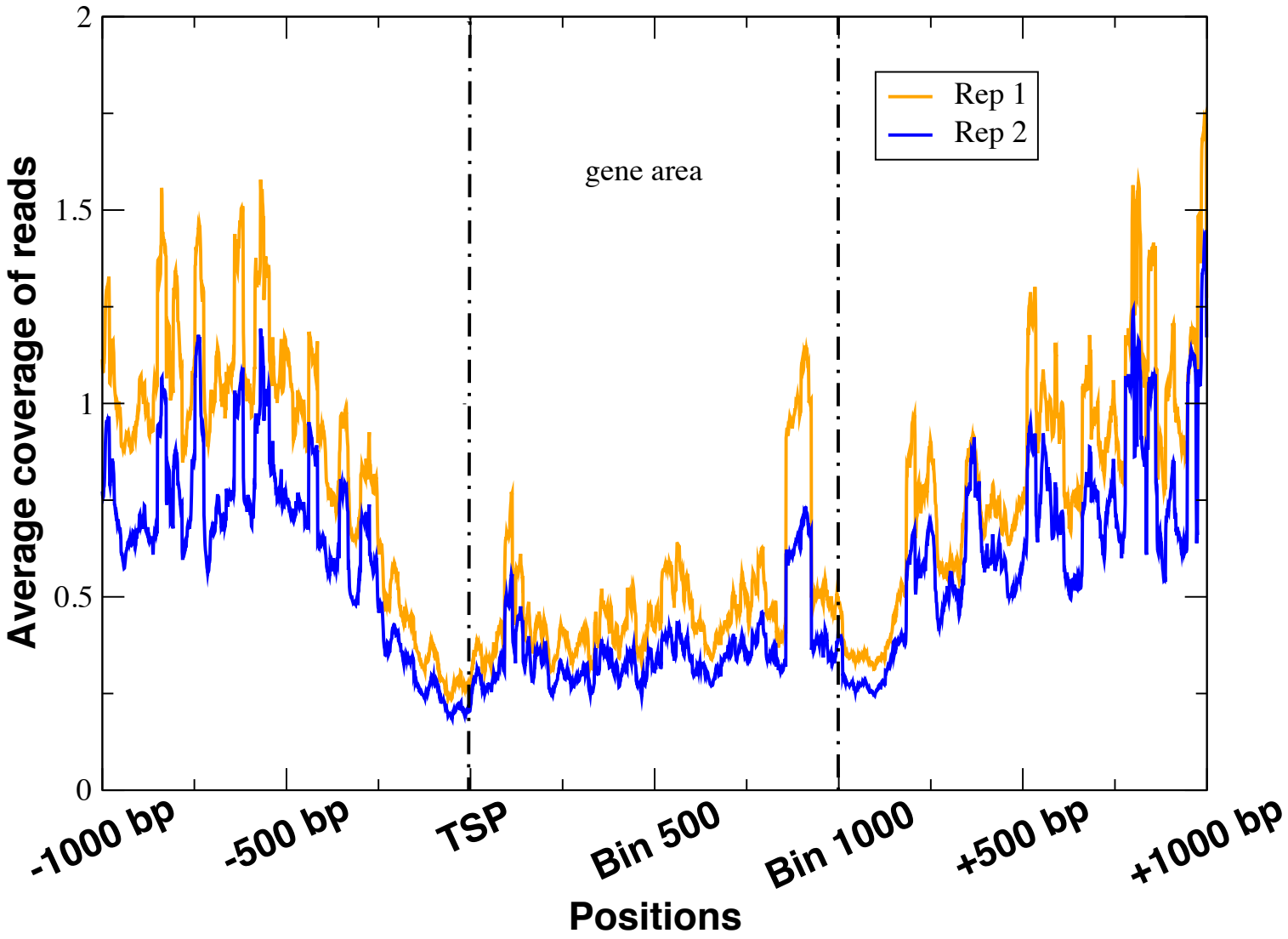


Figure 1: Typical length distribution of remained miRNA-seq reads

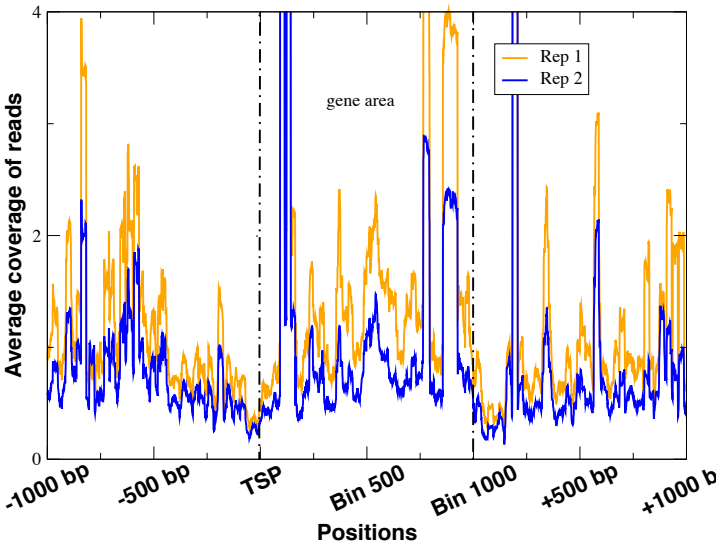
Which libraries are good?



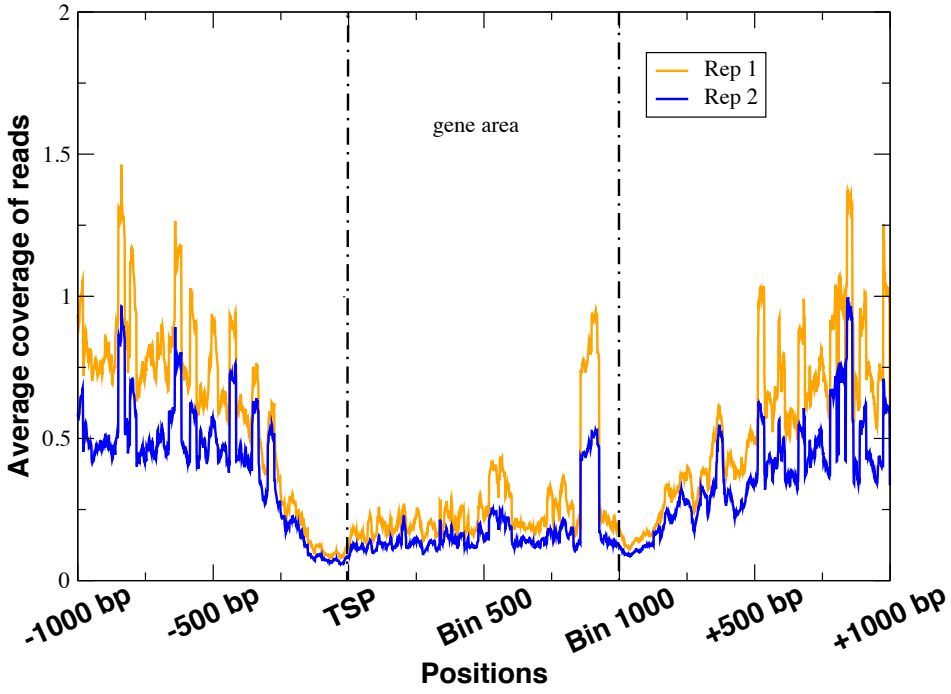
Locations of short reads on genes



Locations of short reads on genes



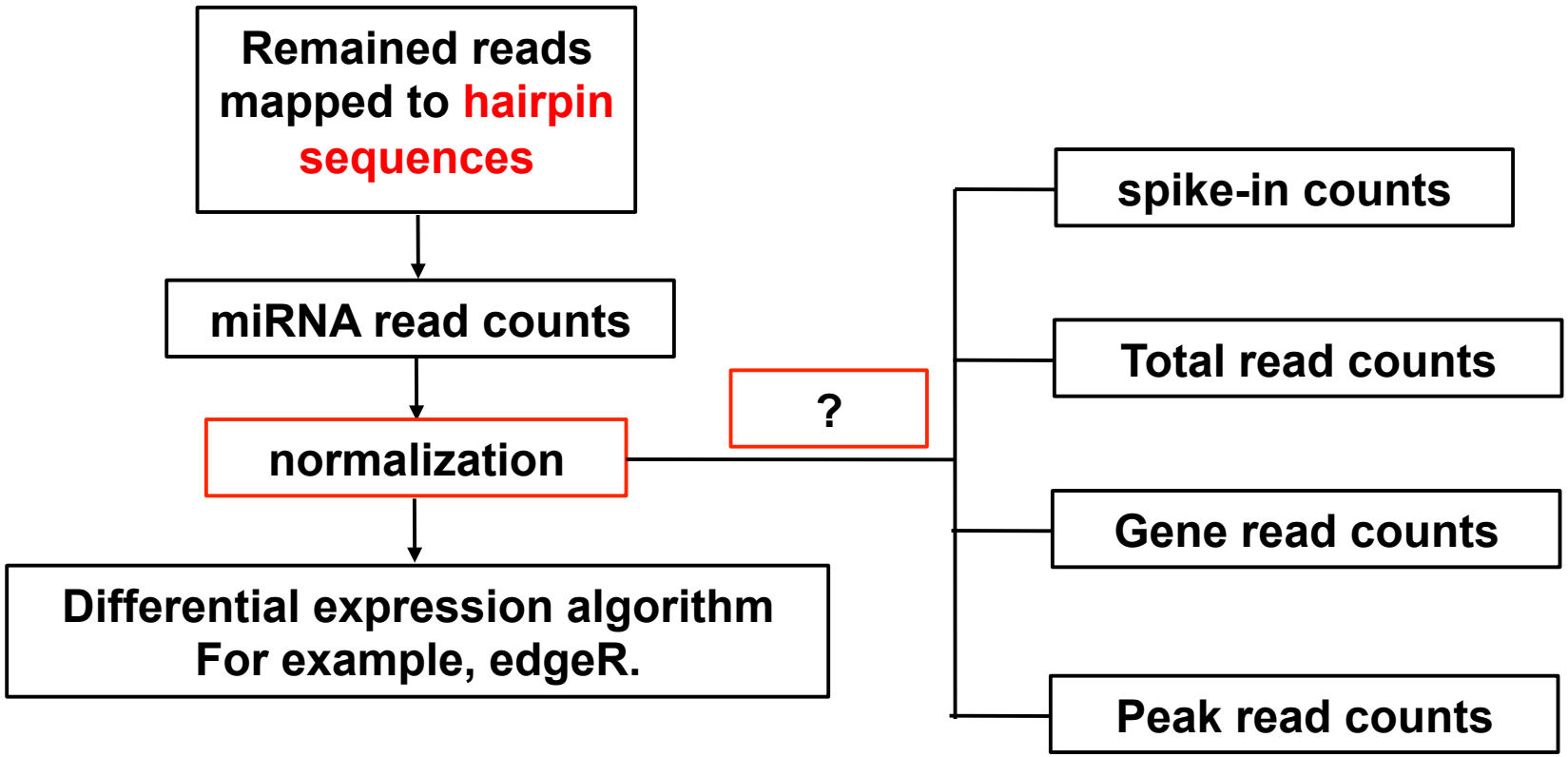
Pseudo genes



Coding genes

Abundance analysis

Pipeline of miRNA differential expression analysis



miRNA hairpins as references

- Use miRNA hairpin sequences as the reference sequences.
- Allow at most 1 mismatch.
- Get hairpin sequences from miRBase: the microRNA database. http://www.mirbase.org/

Differential expression analysis

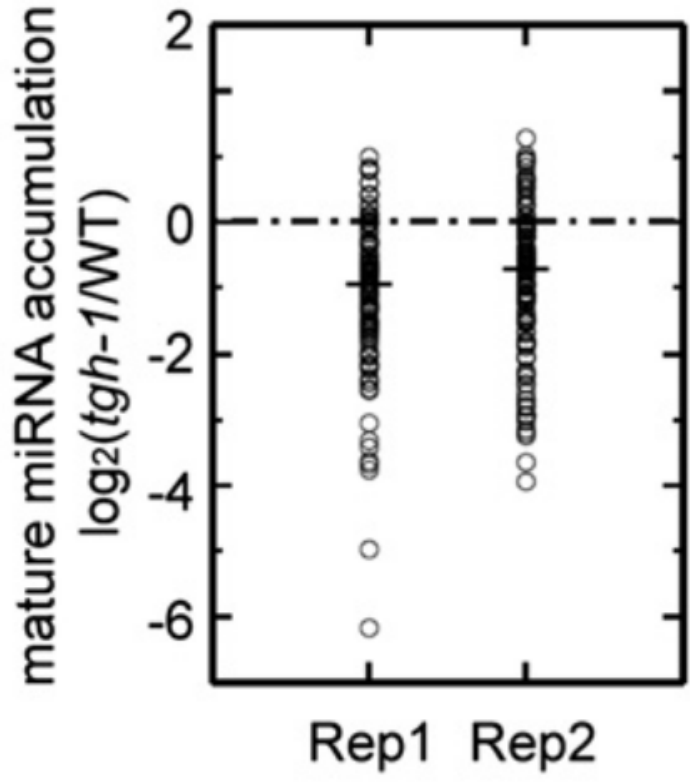
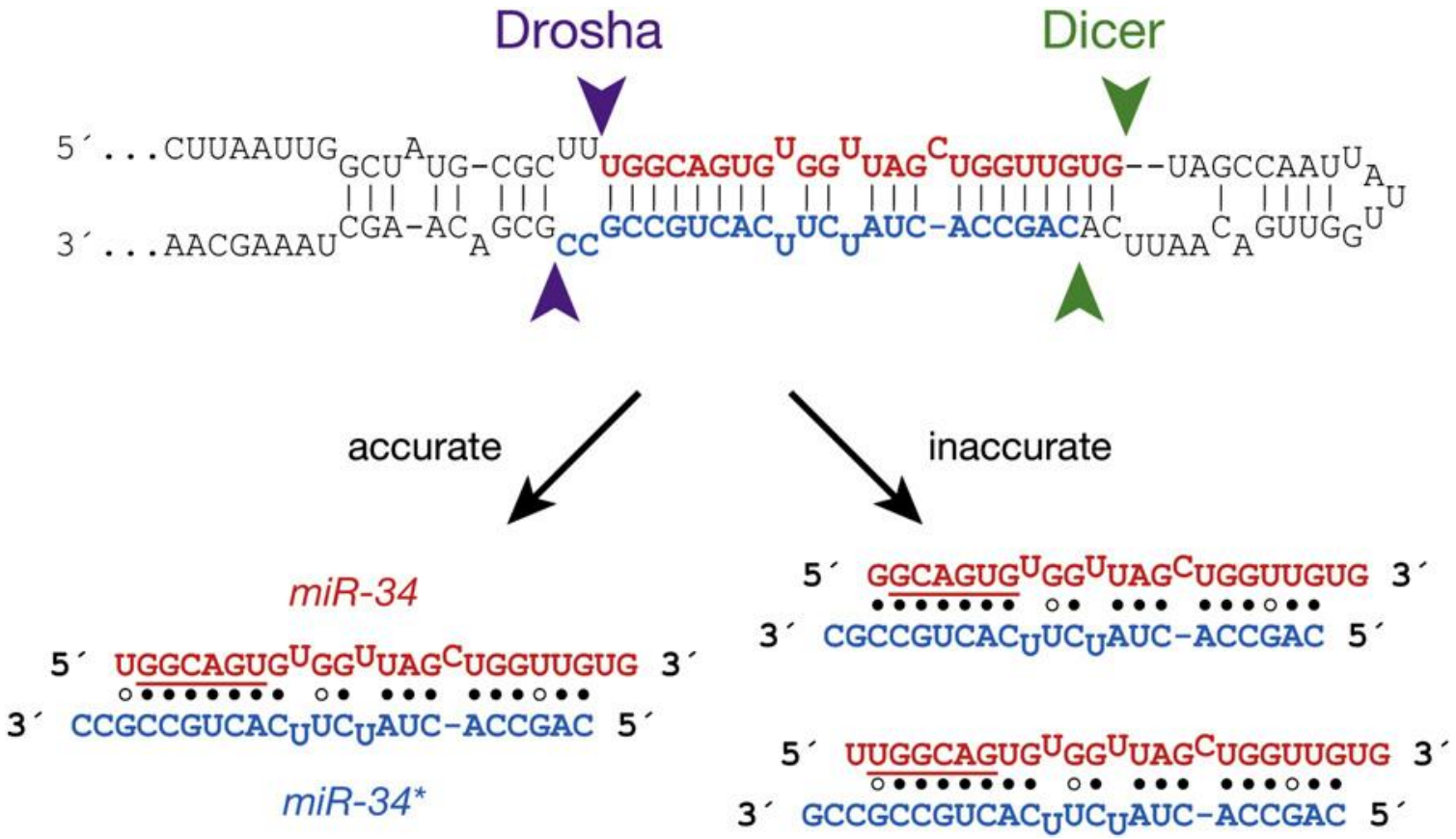


Figure 2: miRNA abundance was down regulated by TOUGH in Arabidopsis

Gene read counts was used as the normalization method.

Ren, Guodong, et al. "Regulation of miRNA abundance by RNA binding protein TOUGH in Arabidopsis." PNAS 109.31 (2012): 12817-12821.

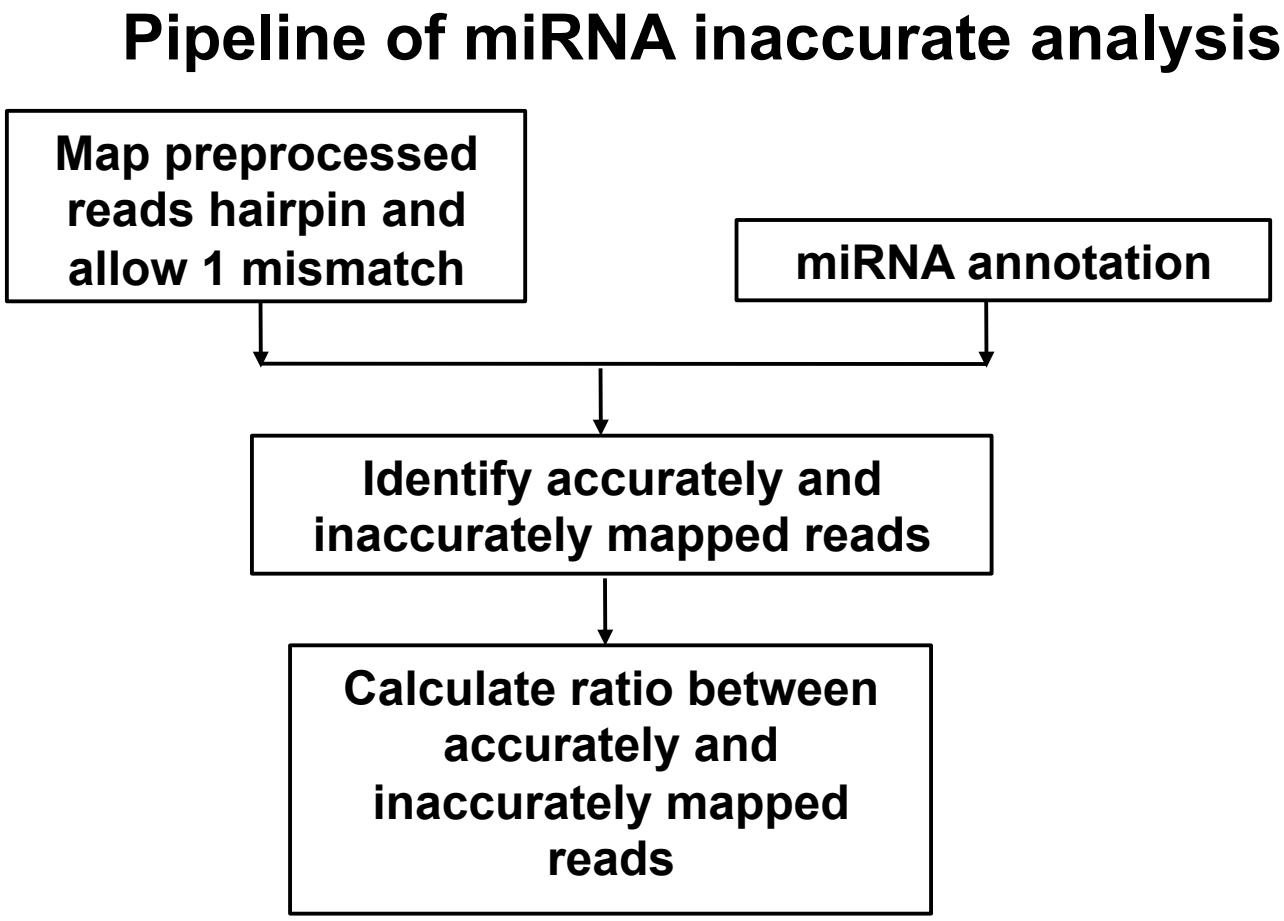
miRNA imprecision analysis



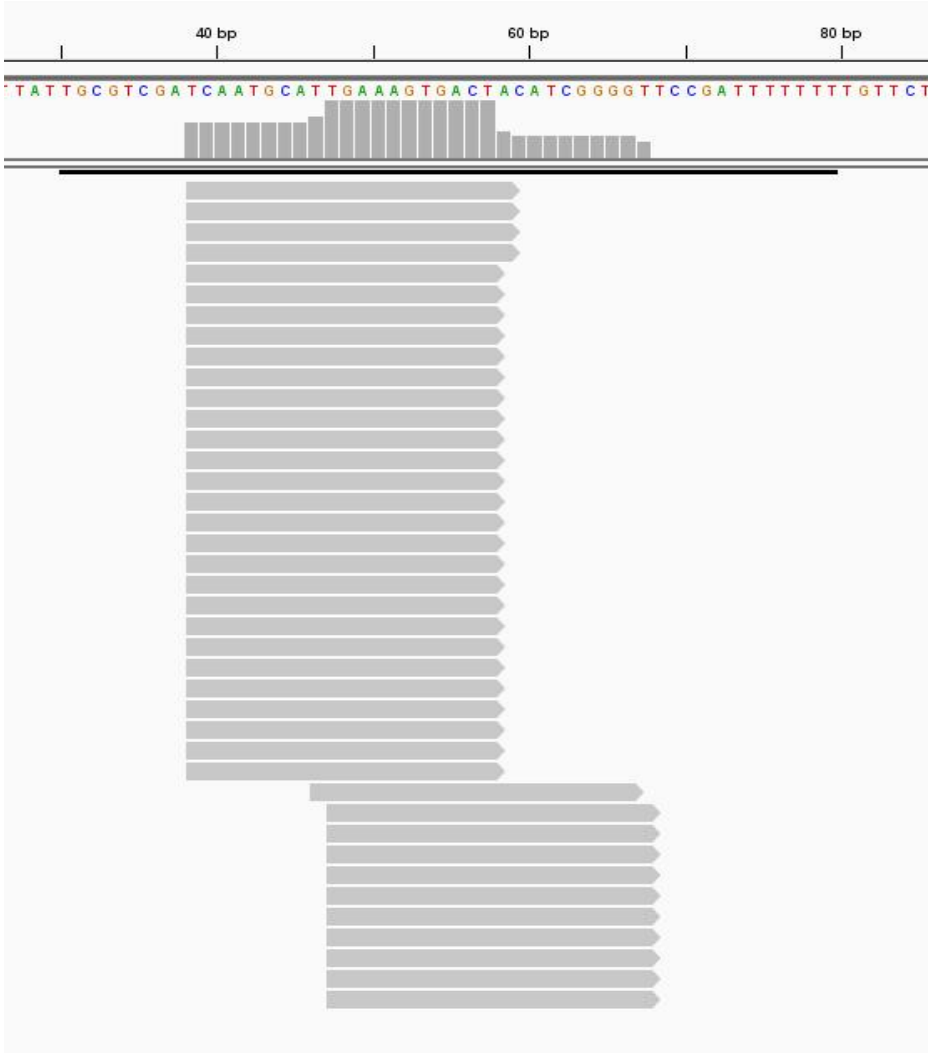
Inaccurate Processing of the 5' end of a miRNA are cleaved by Drosha and Dicer. In this duplex, the mature miRNA (red) is paired to a partially complementary miRNA* (blue).

Seitz et al. Current Biology (2008): 147-151.

miRNA imprecision analysis



miRNA imprecision analysis



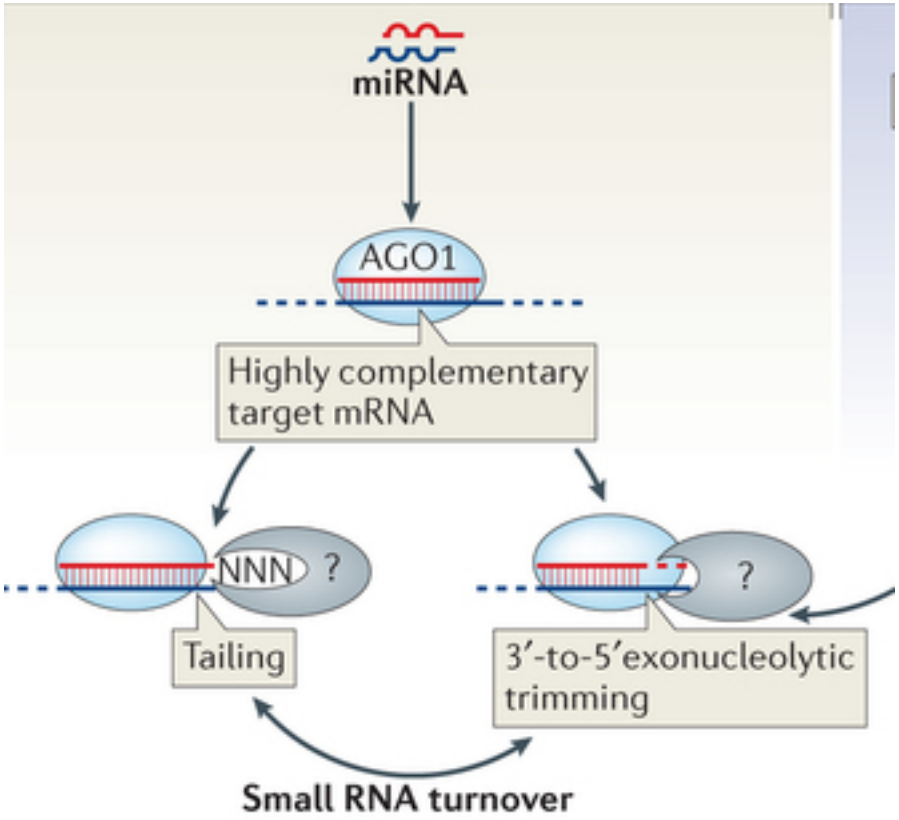
miRNA sequence data shows inaccurate of ath-miR161

Typical numbers of imprecision-reads in a mutant of Arabidopsis

ID	Precision-reads	imprecision-reads	Total-reads	ratio
ath-MIR166a	244333	821	245154	0.003348915
ath-MIR166b	229539	710	230249	0.003083618
ath-MIR166c	229495	647	230142	0.002811308
ath-MIR166d	229498	643	230141	0.002793939
ath-MIR166g	229417	646	230063	0.002807927
ath-MIR166f	229223	645	229868	0.002805958
ath-MIR166e	229216	648	229864	0.002819058
ath-MIR165a	95220	1176	96396	0.012199676
ath-MIR165b	94125	524	94649	0.005536244
ath-MIR158a	64164	268	64432	0.004159424
ath-MIR319a	48540	412	48952	0.008416408
ath-MIR319b	48372	446	48818	0.009135974
ath-MIR159a	18056	409	18465	0.022150014
ath-MIR396a	1805	10737	12542	0.856083559
ath-MIR159b	7576	373	7949	0.046924141
ath-MIR161	6139	257	6396	0.040181363
ath-MIR319c	4824	329	5153	0.063846303
ath-MIR162b	4501	178	4679	0.038042317
ath-MIR162a	4509	76	4585	0.016575791
ath-MIR403	3100	46	3146	0.014621742
ath-MIR858	2638	54	2692	0.020059435
ath-MIR168a	2528	103	2631	0.039148613
ath-MIR396b	2228	55	2283	0.024091108

miRNA trimming and tailing

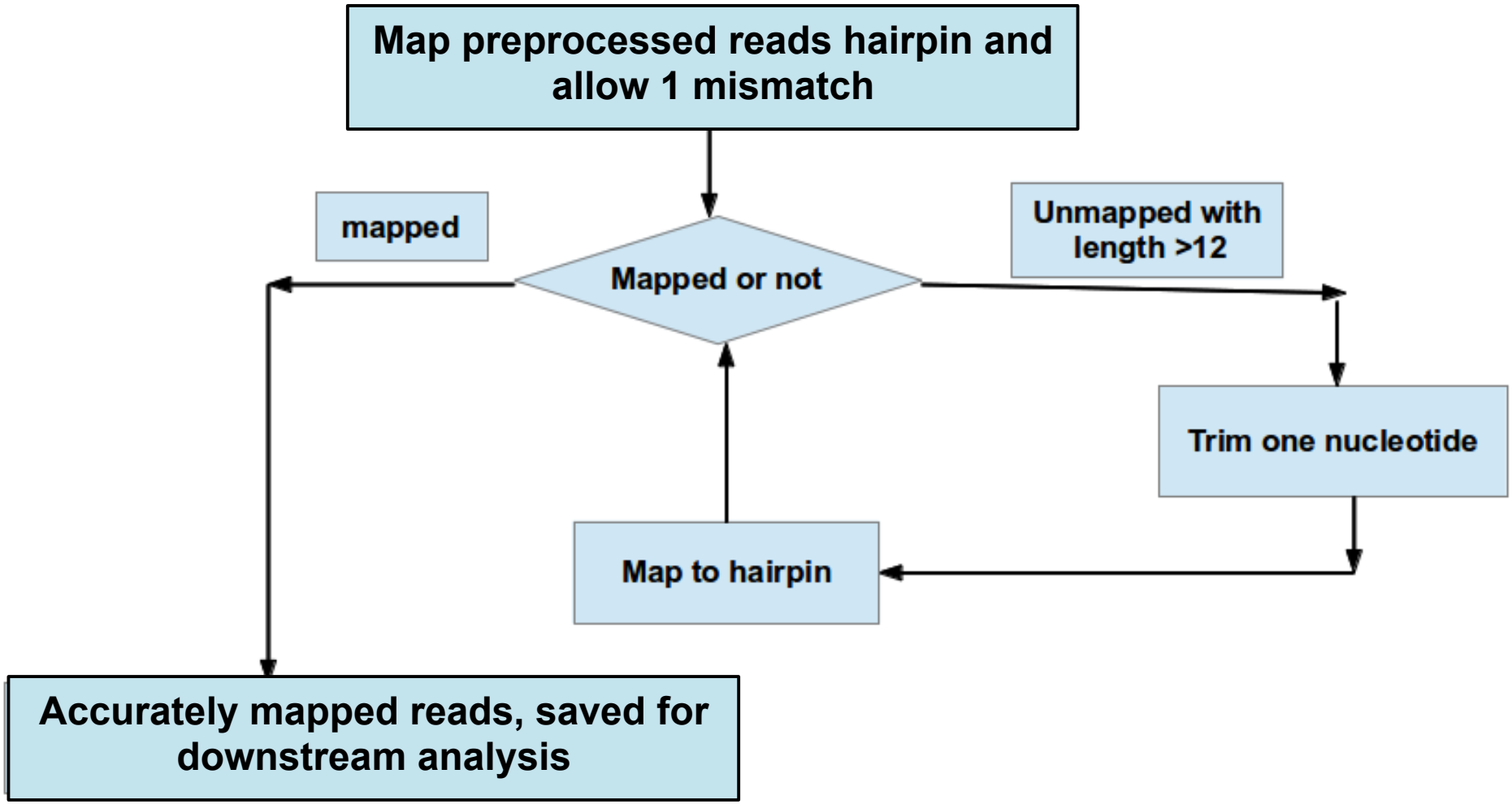
Various mechanisms have now been identified that regulate miRNA stability and that diversify miRNA sequences to create distinct isoforms. The production of different isoforms of individual miRNAs in specific cells and tissues may have broader implications for miRNA-mediated gene expression control.



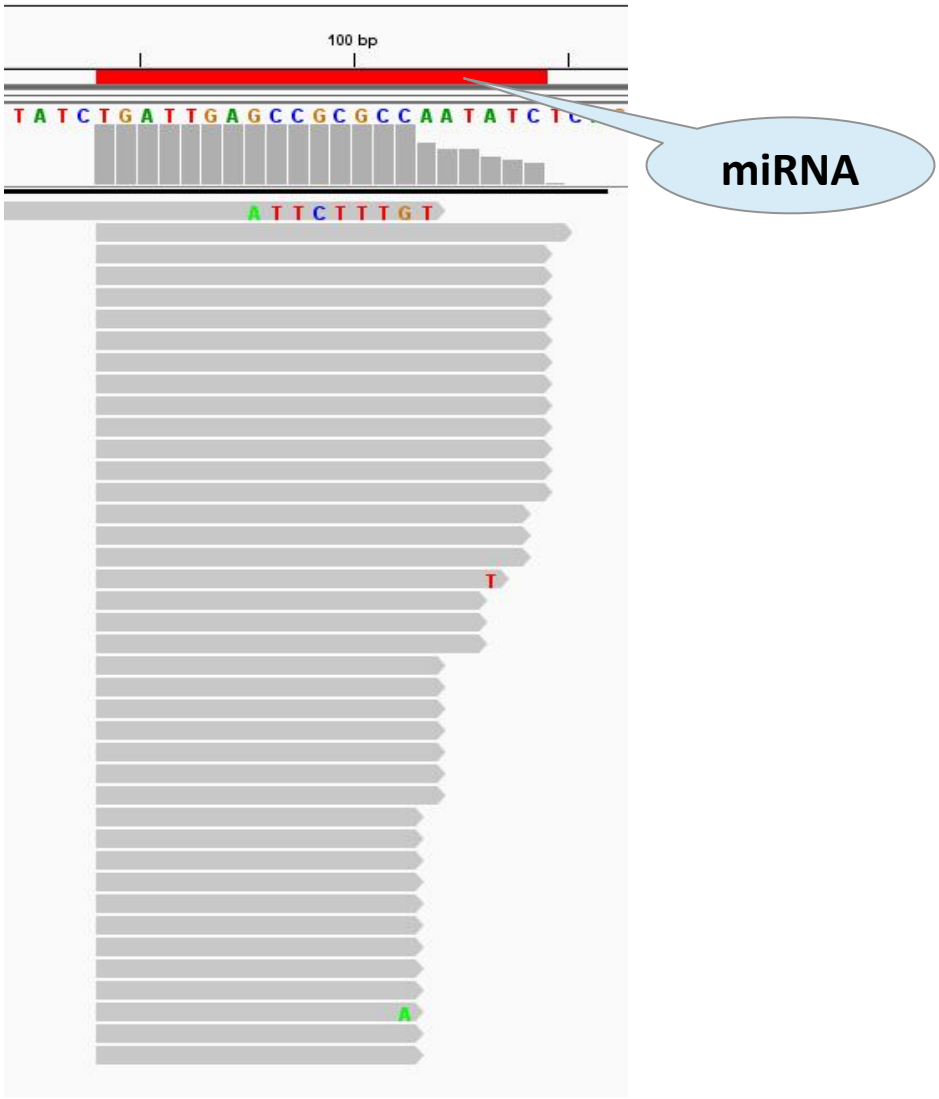
the addition of adenosine or uracil to the miRNA ('tailing')
the 3′-to-5′ exonucleolytic resection of the miRNA 3′ end ('trimming')

Ameres, and. Zamore. Nature Reviews Molecular Cell Biology 14.8 (2013): 475-488.

Trimming and tailing analysis pipeline

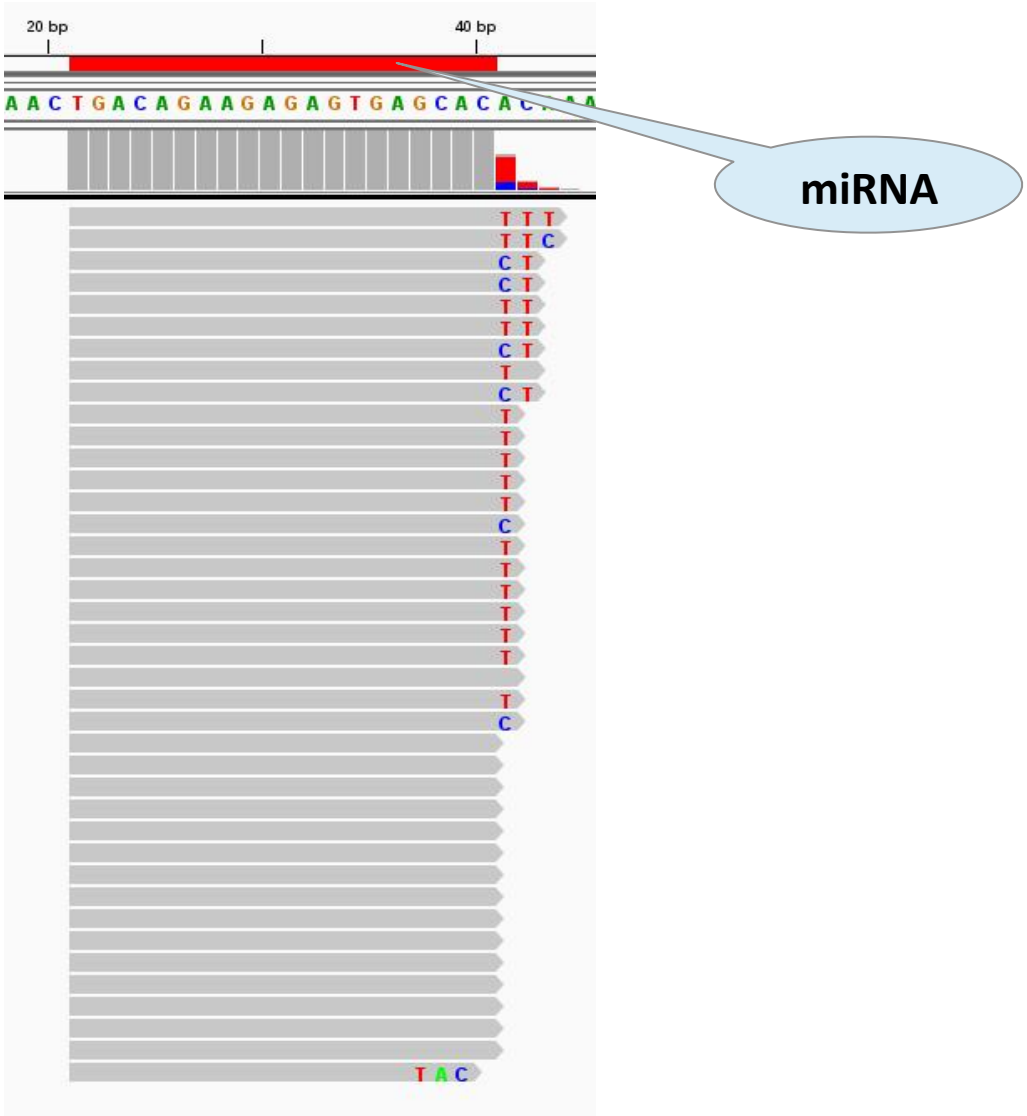


miRNA trimming and tailing



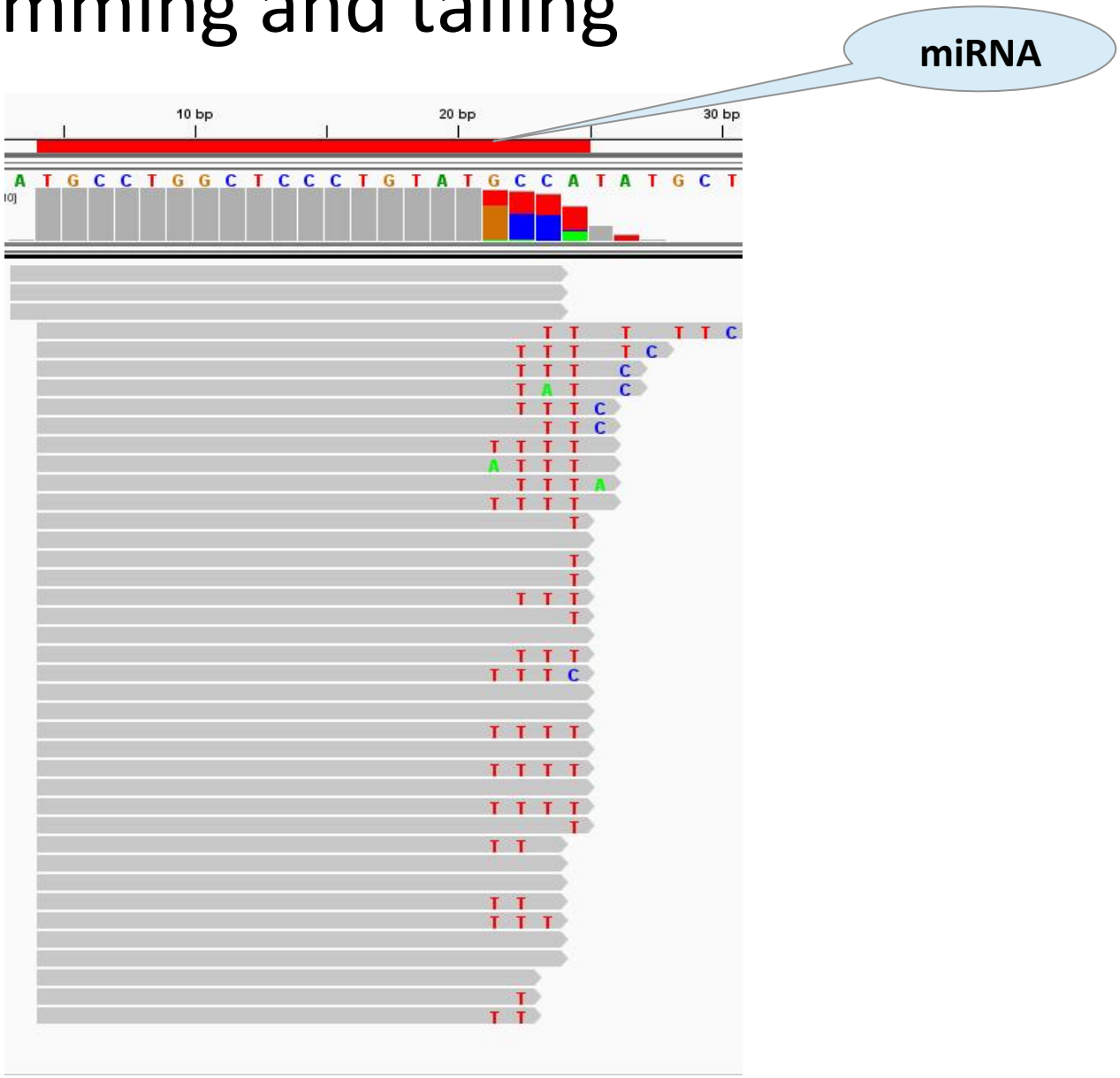
An example for trimming miRNA reads

miRNA trimming and tailing



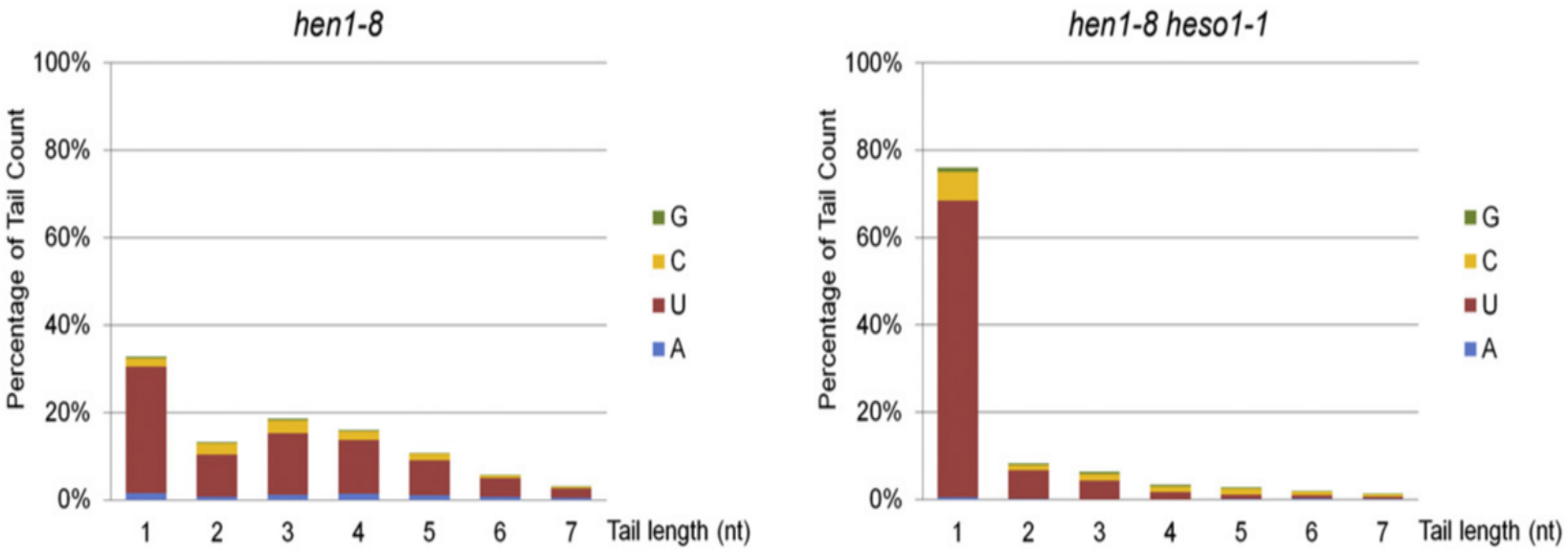
An example for miRNA tailing (some reads have tails)

miRNA trimming and tailing



An example of reads that have both trimming and tailing

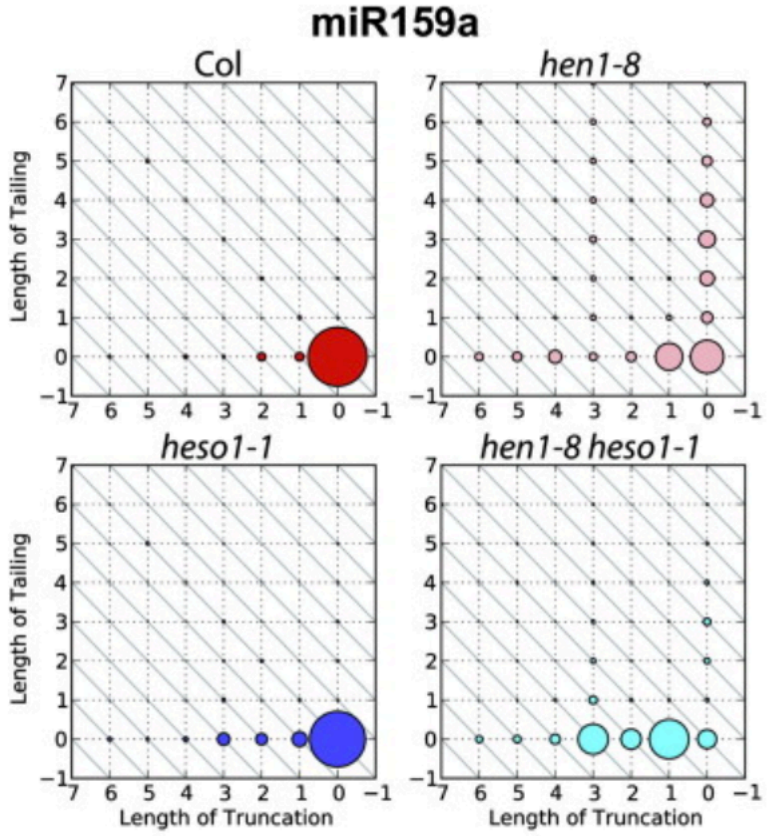
miRNA trimming and tailing



Tail length distribution and nucleotide frequencies in the tails of miR166a. The figure shows there was a shift toward shorter tails in the hen1-8 heso1-1 mutant as compared to the hen1-8 mutant.

Zhao, et al Current Biology (2012): 689-694.

miRNA trimming and tailing analysis



The distribution of trimmed and tailed reads of miR159a in different lines

Zhao, et al Current Biology (2012): 689-694.

Types of non-coding RNA

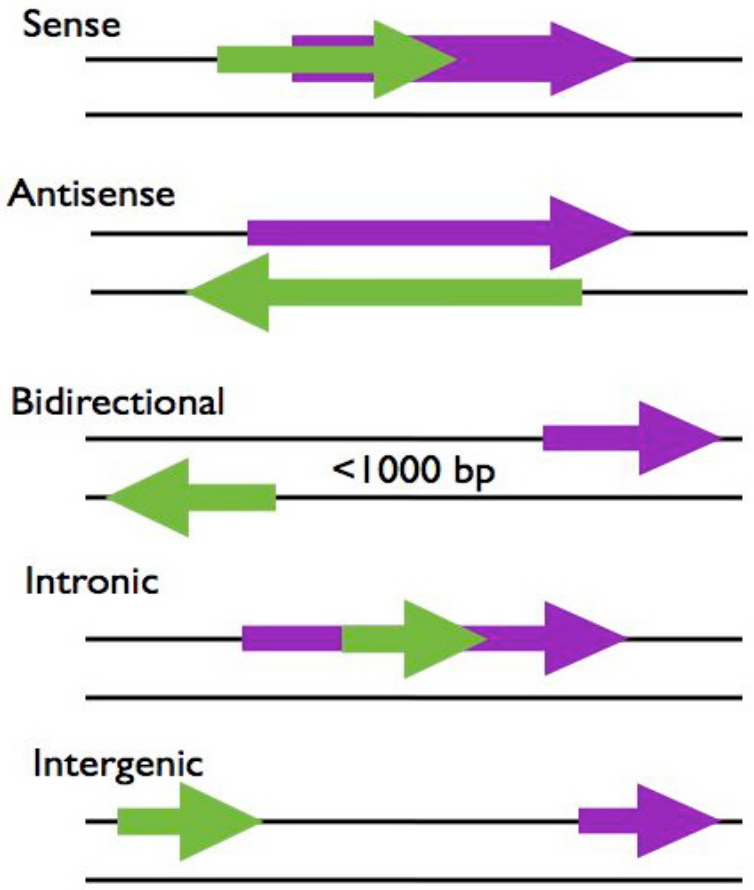
- transfer RNA (tRNA) and ribosomal RNA (rRNA),
- snoRNAs-Small nucleolar RNA
- microRNAs
- siRNAs
- snRNAs- Small nuclear ribonucleic acid
- exRNAs-Extracellular RNA
- long ncRNAs-Long non coding RNAs

Long non coding RNAs

- Long non-coding RNAs (long ncRNAs, lncRNA) are non-protein coding transcripts longer than 200 nucleotides.
- Non-coding RNAs play very important roles in regulation
- Recently, Long non-coding RNAs (lncRNA), longer than 200 nucleotides, have been discovered.
- LncRNAs have gained widespread attention as a potentially new and crucial layer of biological regulation
- Many lncRNA act by activating or repression the transcriptional activity of other genes.

Location of lncRNA in the genome.

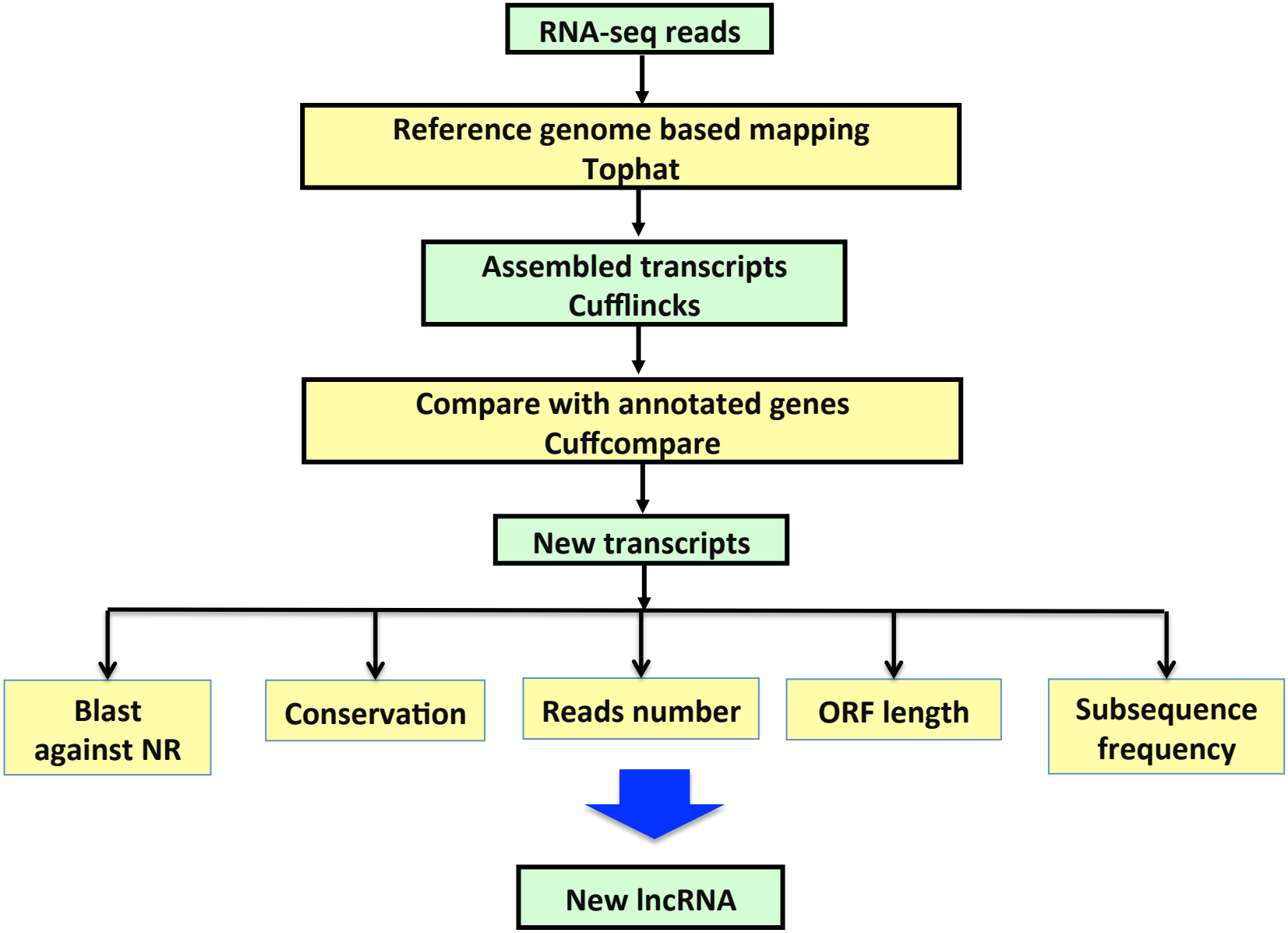
- LncRNAs can be categorized according to their proximity to protein coding genes in the genome, using this criteria lncRNAs are generally placed into five categories:



long non-coding RNA

- RNA-seq is a useful tool for discovery of new lncRNAs
- Many new lncRNAs are discovered by RNA-seq, but most them are in animal species.

Pipeline


RNA-seq reads
Reference genome based mapping
Tophat
Assembled transcripts
Cufflincks
Compare with annotated genes
Cuffcompare
New transcripts
Blast against NR
Conservation
Reads number
ORF length
Subsequence frequency
New lncRNA

An example for maize

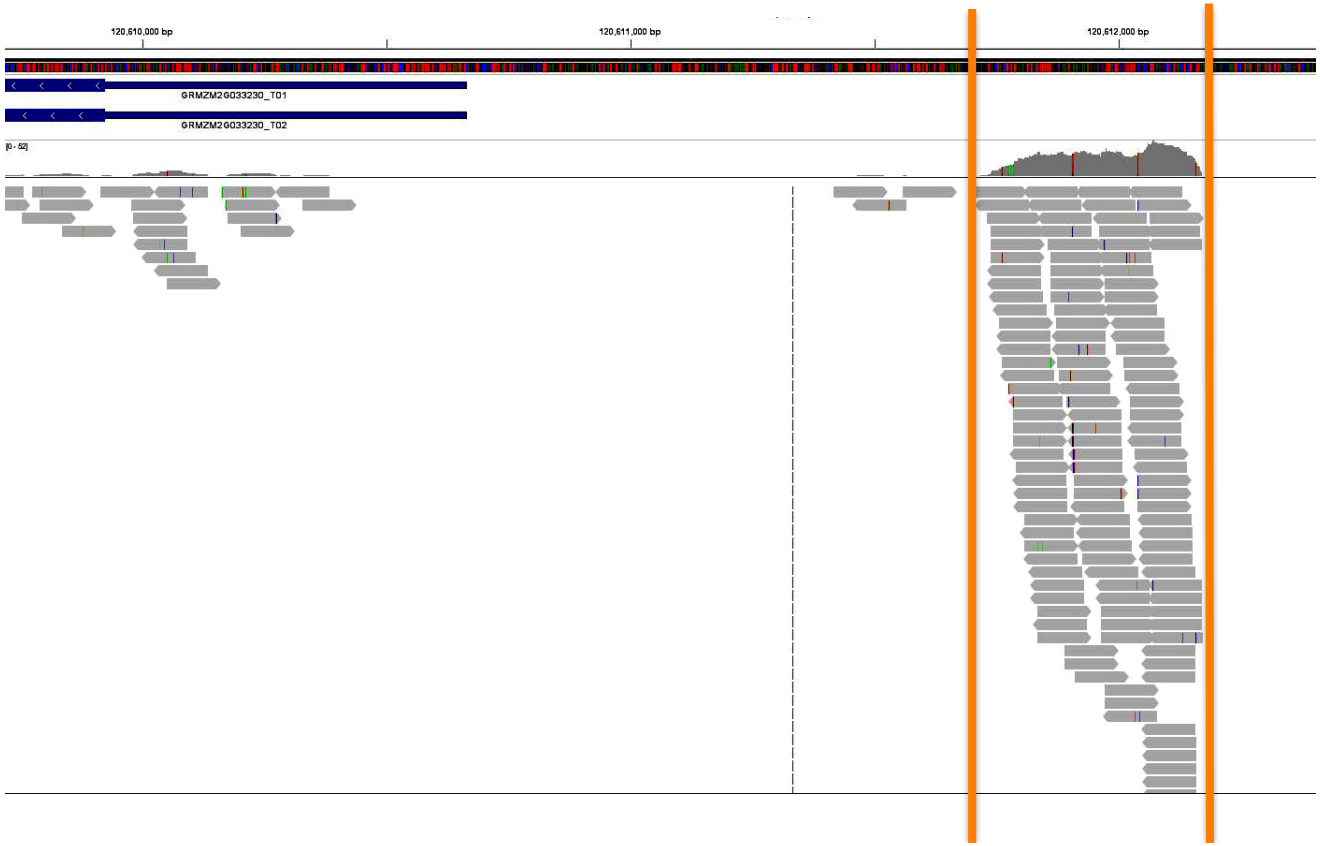
- RNA-seq reads from B73 maize line after submerging
- Total 2775 new transcripts are found
- The number of high quality new transcripts is 228. They appear in all three replicates and have more than 100 reads in each replicate
- There are 29 candidates of intergenic long non-coding RNAs

An example for maize

- These 29 candidates of intergenic lncRNAs
  - No homology to known protein coding genes
  - No discernible protein motif
  - No long open reading frame
  - Are less abundant (number of average reads = 200, which is smaller than other coding genes)
  - Contain fewer exons (most just have one exon).

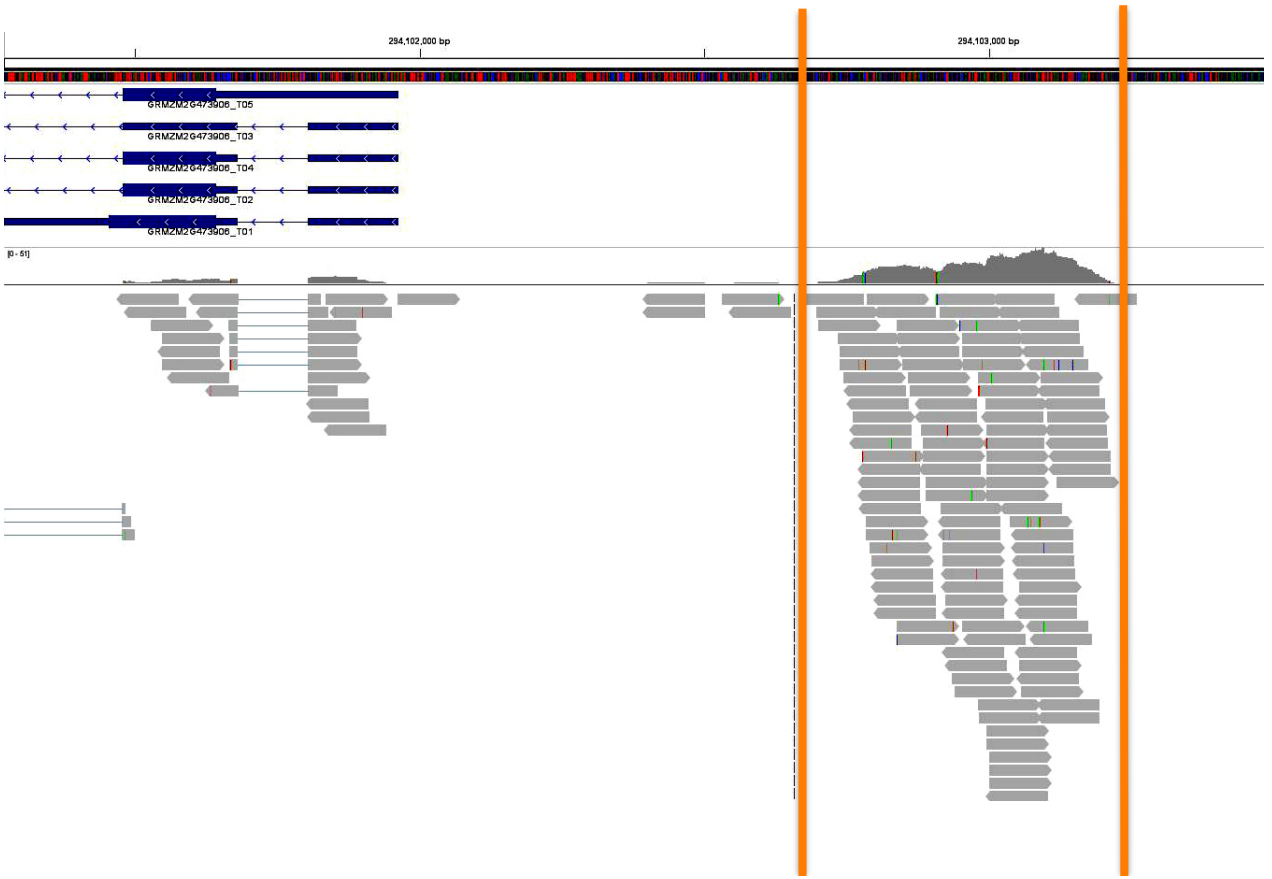
lncRNA Candidate 1

- Chr 8
- Near GRMZM2G033230
- Length: 473
- Reads number: 114



Candidate 2

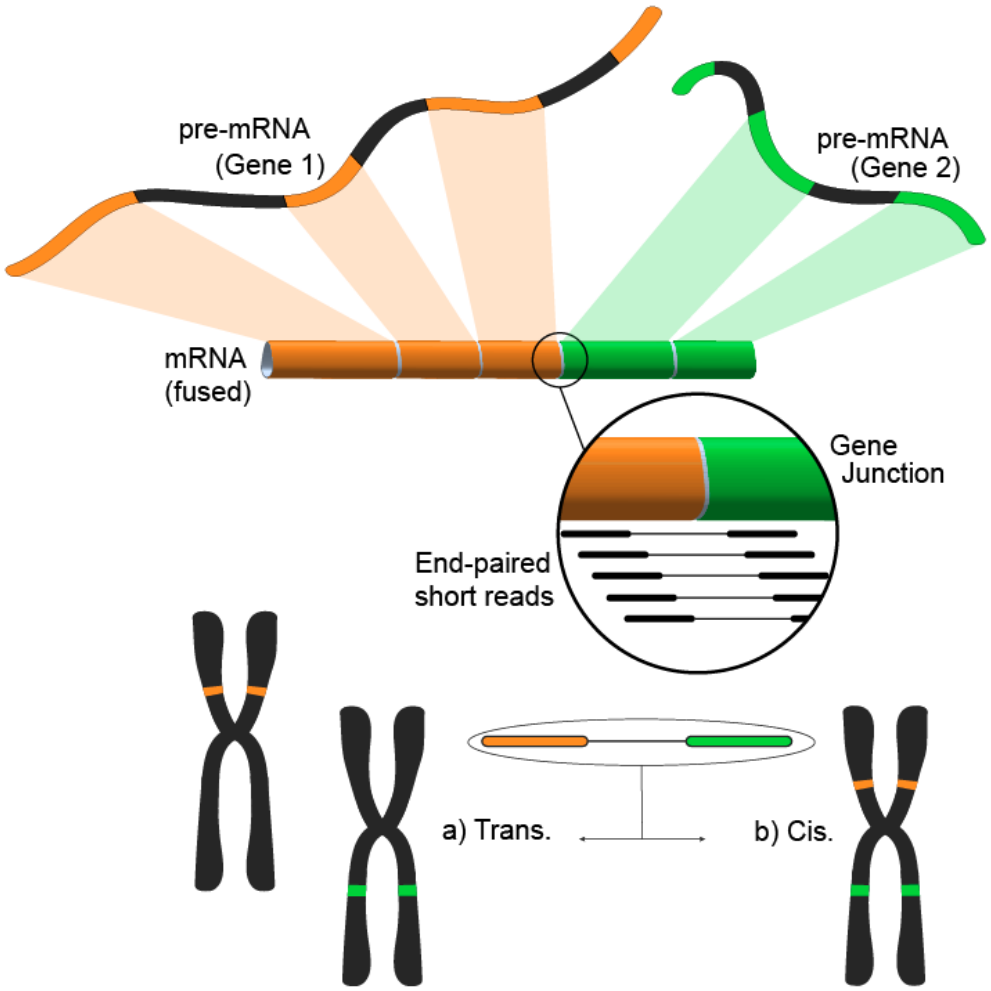
- Chr 1
- Near GRMZM2G473906
- Length: 560
- Reads number: 127



Applications of RNA-seq

- Gene expression
  - Expression of individual genes/loci
  - Quantitatively discriminate isoforms using junction reads and coverage of individual exons, introns, etc.
- Annotation
  - New features of the transcriptome: genes, exons, splicing, ncRNAs (next class)
- SNP
- Fusion gene detection

Other applications of mRNA-seq: gene fusion



- The unmapped short reads can then be further analyzed to determine whether they match an exon-exon junction where the exons come from different genes.
- An alternative approach is using pair-end reads, when potentially a large number of paired reads would map each end to a different exon, giving better coverage of these events.
- Novel combinations genes can be identified.

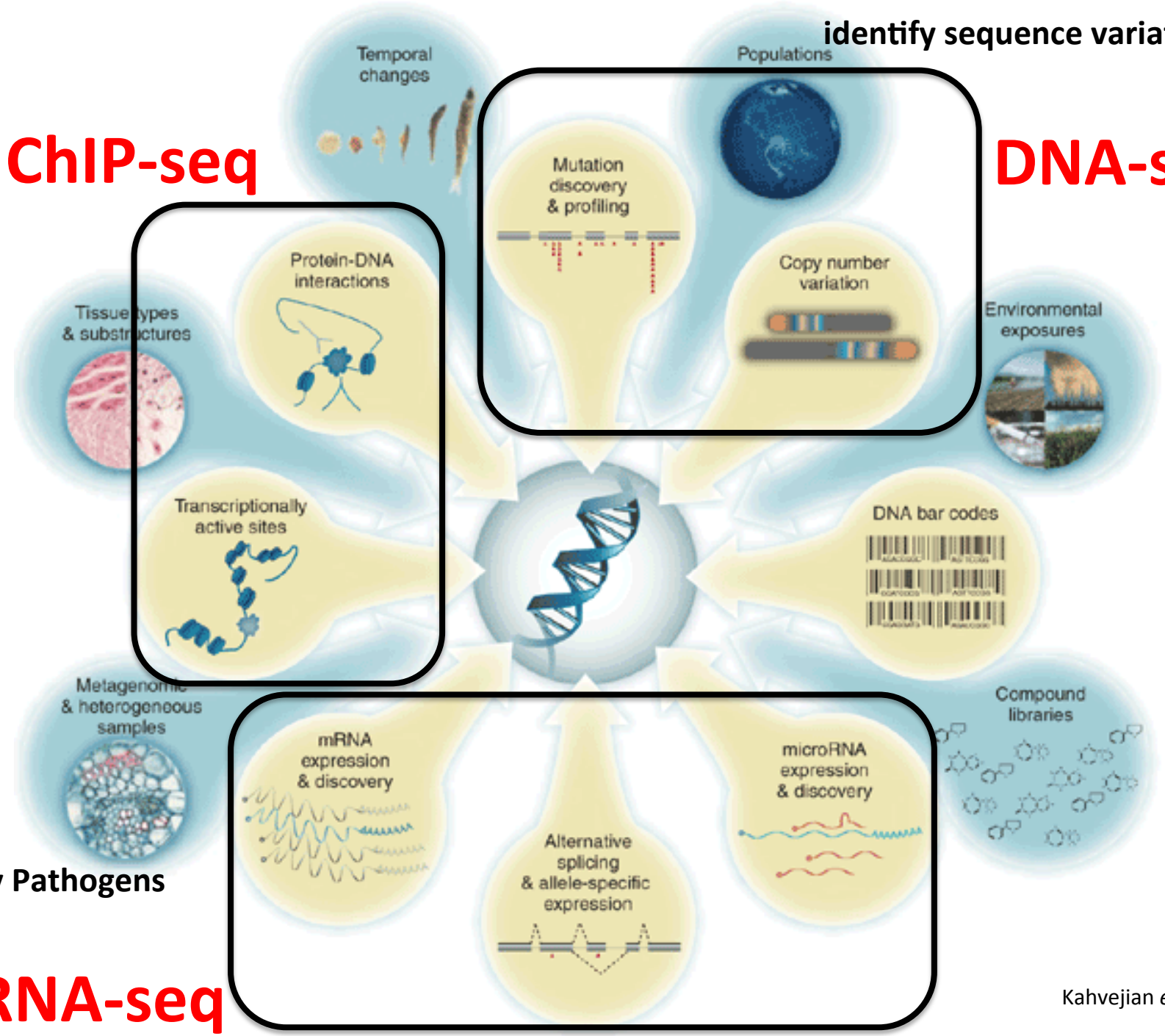
Acknowledgement: Wiki – mRNA-seq

Finding fusion genes

- A case: RNA-seq data for the leukemia K562 cell line
  - ~15 000 candidate fusion-genes found
  - ~85% candidate fusion-genes are known paralogs or have no protein product!!!
  - 15 candidate fusion-genes are found after additional filtering of candidate fusion-genes where the known BCR-ABL is number one candidate

- Filtering of candidate fusion-genes is highly necessary in order to reduce the large number of candidate fusion-genes (from ten of thousands to tens)!

identify sequence variations



Identify Pathogens

Kahvejian *et al*, 2008