

CURSO DE CURTA DURAÇÃO - 2017

BIOINFORMÁTICA

BioME – CENTRO MULTIUSUÁRIO DE BIOINFORMÁTICA - UFRN

NEXT GENERATION SEQUENCING

Análise de Dados de Sequenciadores de Segunda Geração

Prof. Dr. JORGE ESTEFANO SANTANA DE SOUZA

E-mail: jorge@imd.ufrn.br



RNA-seq II



Bioinformatics
Multidisciplinary
Environment

Centro
Multiusuário
de Bioinformática



METRÓPOLE
DIGITAL



BIOINFORMÁTICA
UFRN



UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

²Bio
Instituto de
Bioinformática e
Biotecnologia

Objetivo:

Utilizar as ferramentas básicas de transcriptoma para obter o padrão de expressões dos mirRNAs de uma amostras.

Comandos Básicos de Linux:

Para trabalhar com nossos dados, vamos precisar saber alguns comandos básicos do Linux. Podem procurar mais informação no site:

<http://wiki.ubuntu-br.org/ComandosBasicos>

Ferramentas:

- 1- Linux.
- 2- WebServer.
- 3- cutadapt
- 4- mapper
- 5- miRDeep2

Inicial:

Login maquina local:

Login:

Senha:

Login no server:

```
ssh -p 4422 bif@10.7.5.38
```

Senha: bif0003

Inicial:

Pasta com dados iniciais:

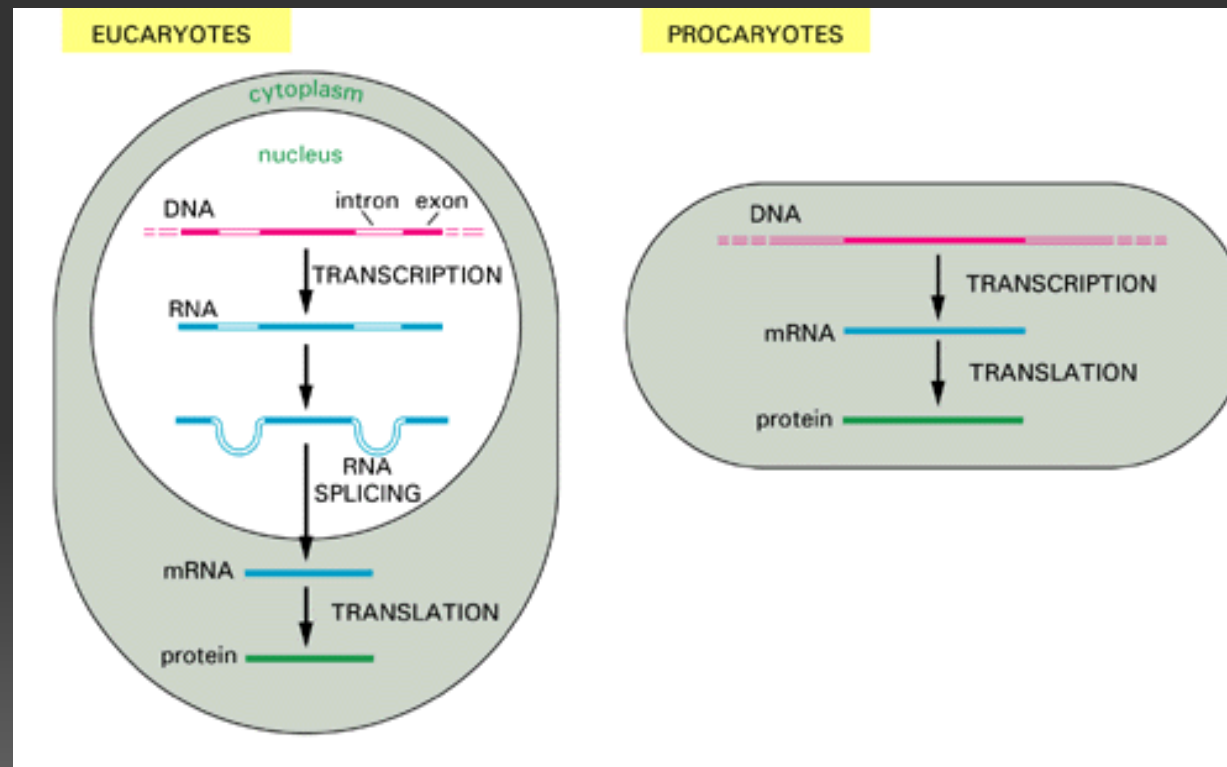
`/home/treinamento/NGS/RNAseq/`

Pasta servidor WEB:

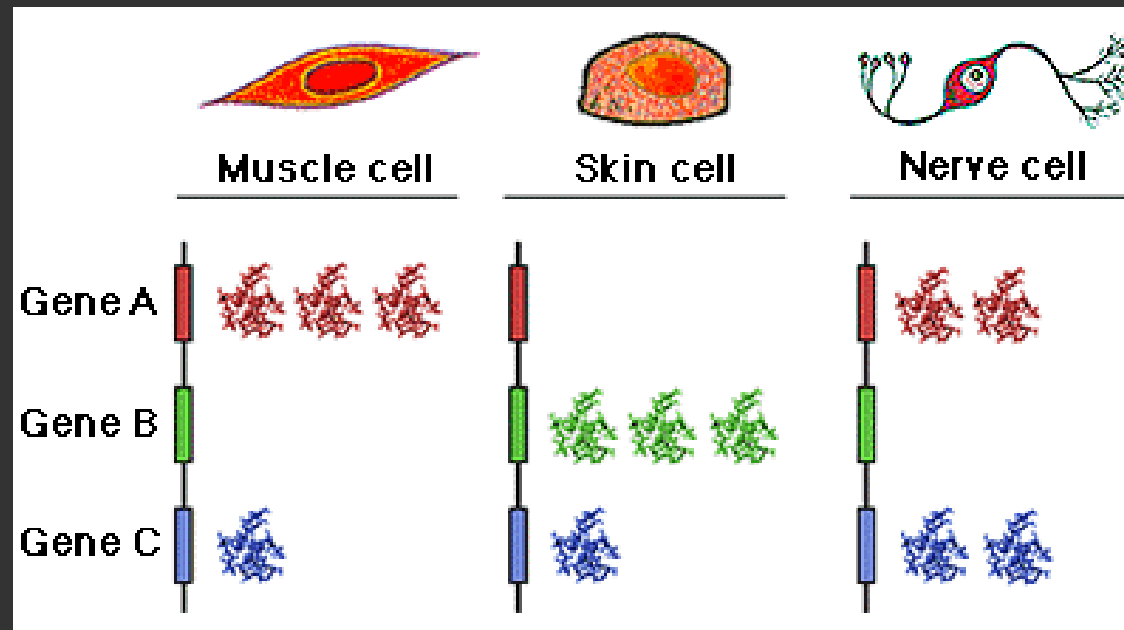
`/home/bif/public_html/`

Transcriptoma

Transcriptoma: conjunto total de RNAs em uma dada célula ou tecido.

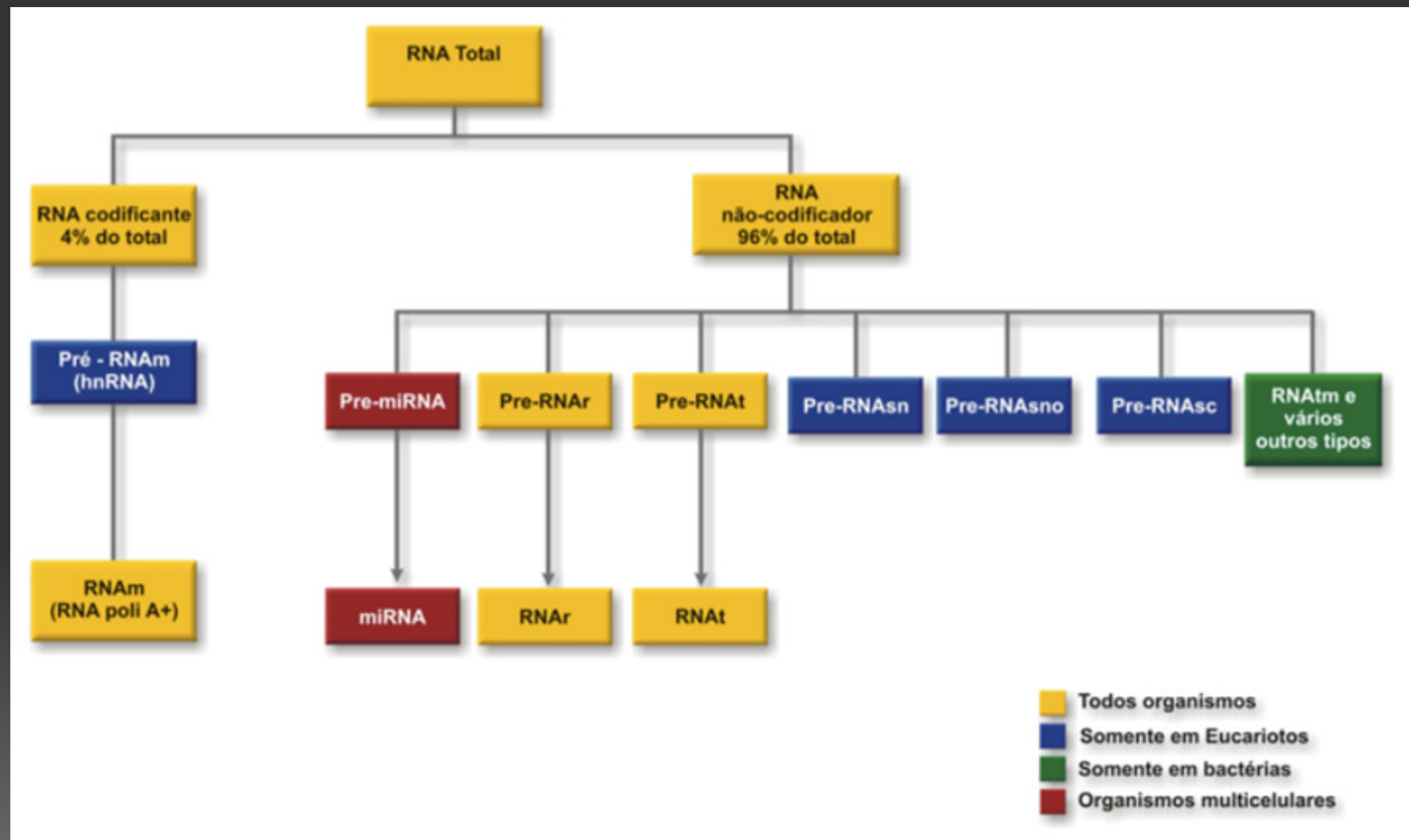


Transcriptoma



A expressão gênica diferencial é responsável pela diferenciação fenotípica entre células do mesmo organismo.

Transcriptoma: composição



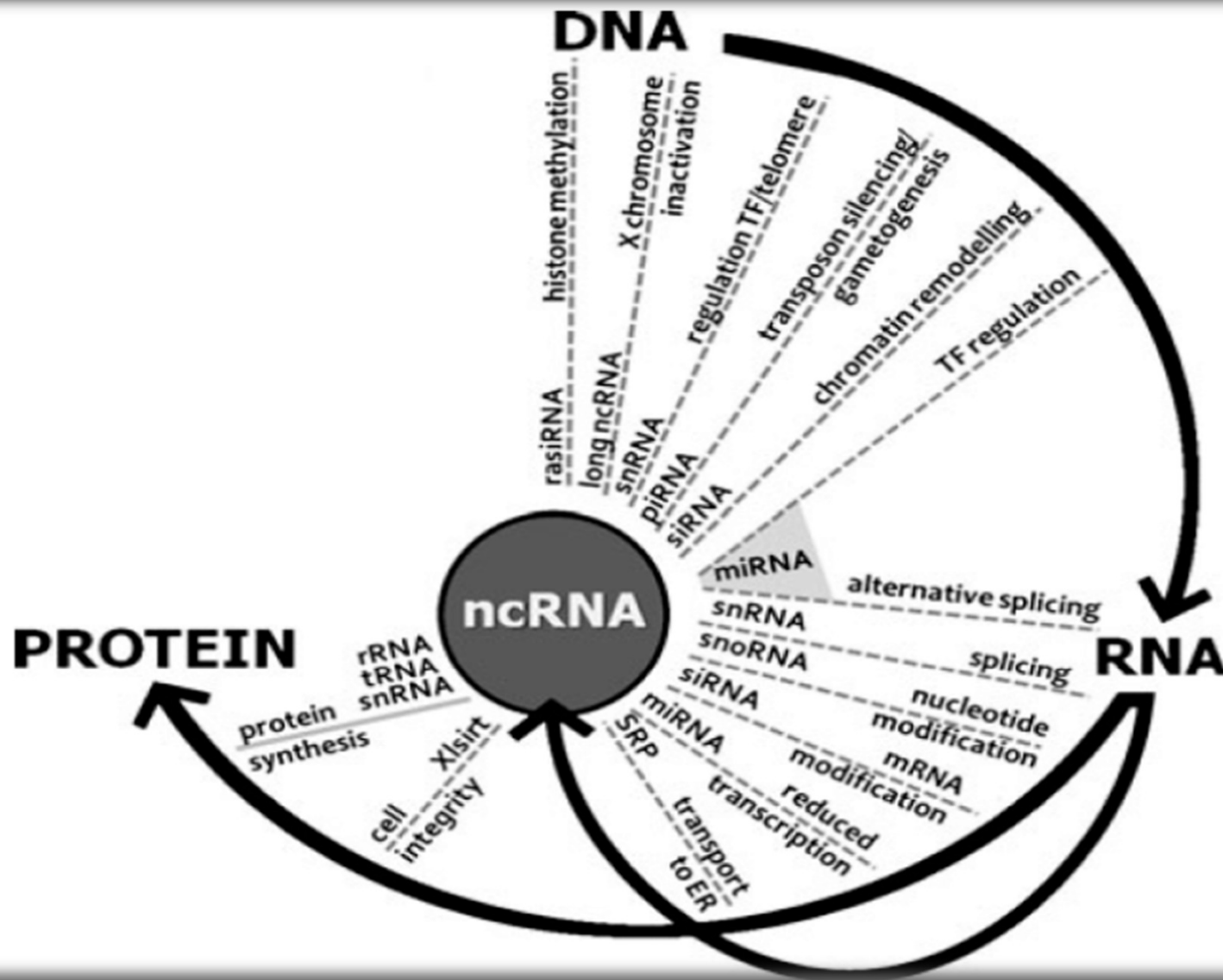
RNA não-codificador (ncRNA)

- Produtos de genes presentes no genoma.
- ncRNA não é traduzido para uma proteína.
- Participa nos mais diversos processos biológicos: regulação de ciclo celular, diferenciação de células e tecidos, desenvolvimento, apoptose, metabolismo.
- Presente nos diversos reinos dos seres vivos.
- A diversidade funcional de cada classe de ncRNAs ainda está longe de ser totalmente conhecida.

ncRNAs : principais classes

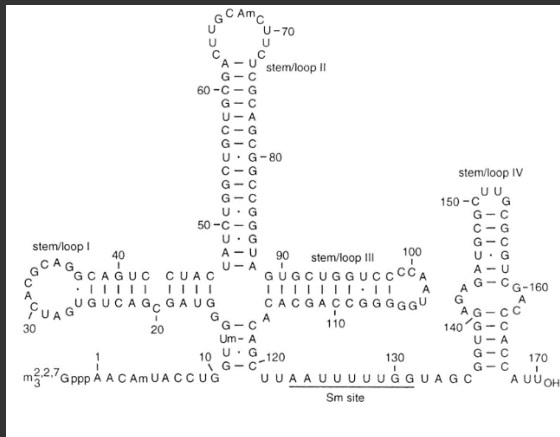
Nome	Descrição
microRNA	pequeno RNA (maduro 18-25 nt) que, em geral, contém um papel de regulação pós-transcricional.
piRNA	<i>piwi-interacting</i> RNAs – relacionado às células germinativas.
siRNA/RNAi	<i>small interfering</i> RNAs ou RNA de interferência – pequeno RNA (20-25 nt), que em geral, interfere na expressão de genes por silenciamento.
snoRNA	<i>small nucleolar</i> RNAs, encontra no núcleo das células eucarióticas, sendo relacionado a vários processos como: modificação de RNA etc
sRNA	Pequenos RNAs ou <i>small</i> RNA (< 200 nucleotídeo [nt]) em que se incluem os miRNA, siRNA e outros
TERC	Componente de Telomerase RNA
Longos ncRNAs	RNAs com mais de 200 nucleotídeos. Ex. Xist RNA
NAT	Transcrito antisense ou <i>Natural antisense transcripts</i>
SRP RNA	Partícula de reconhecimento de sinal ou <i>Signal Recognition Particle RNA</i>
Ribozymes	Rnase P, <i>Hammerhead</i> e RNAs intrônicos grupo I, II e III
Promoter-associated RNAs	RNAs recentemente descritos localizados em região promotora/TSS. Ex.: pasRNA, palRNA, tiRNA, CUT, PROMPT, TSSa-RNA etc.
Termini-associated RNAs	RNAs descritos localizados na da região terminadora de genes.

ncRNAs: principais classes

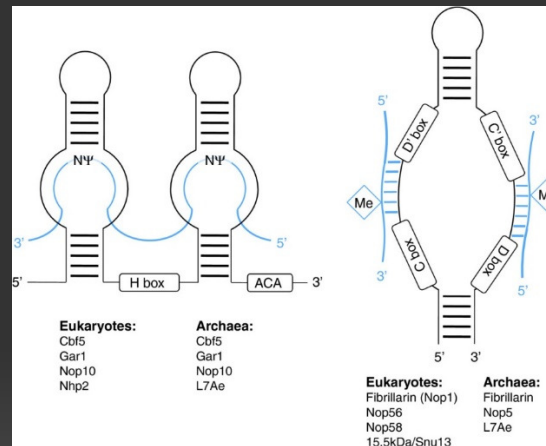


ncRNAs : estrutura secundária

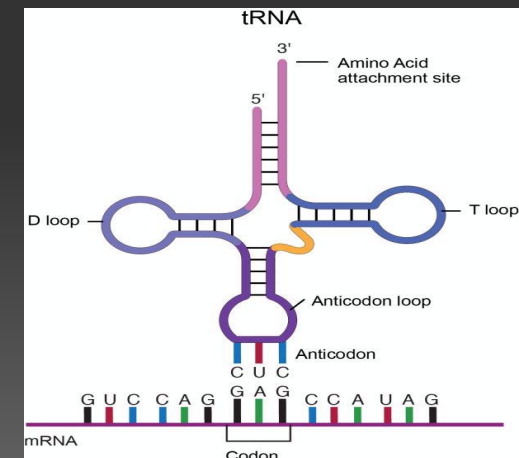
Pequeno RNA nuclear (snRNA)



Pequeno RNA nucleolar (snoRNA)



RNA transportador (tRNA)

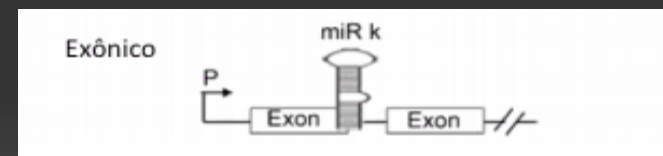
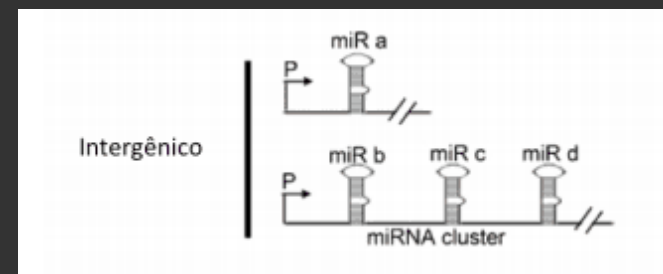
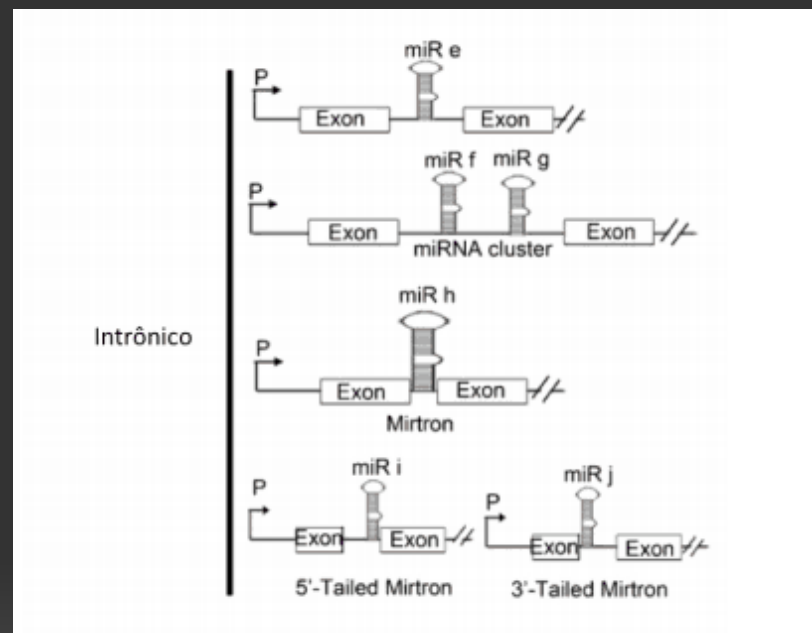


ncRNAs : tamanho (nt)

Classe de sRNAs	Tamanho (nt)	Organismos
miRNA	17-25	Plantas, algas, vírus animais, protistas
endosRNA	21-24	Plantas, fungos, animais
exosRNA	~24	Plantas, fungos, animais
natsRNA	21-24	Plantas
casiRNA	24	Plantas
tasiRNA	21	Plantas
piRNA	26-31	Células germinativas
piRNA-like	24-30	<i>Drosophila</i> , <i>C. elegans</i>
rasiRNA	26-31	Plantas, animais

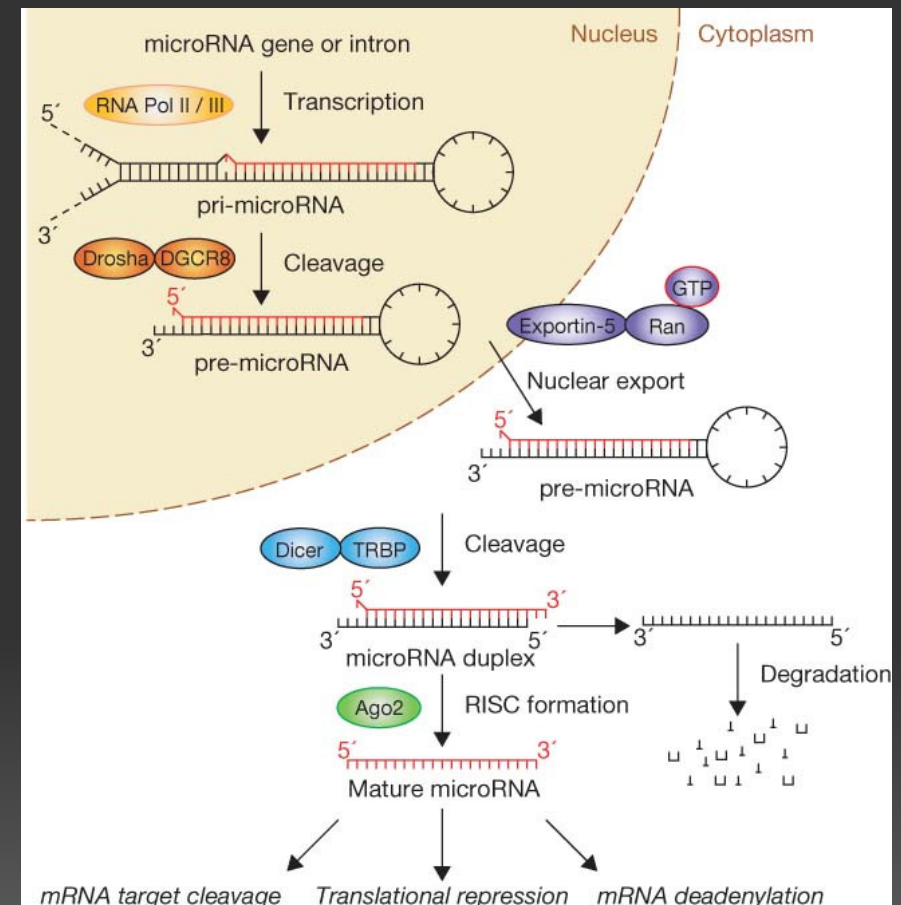
Classe de sRNAs	Tamanho (nt)	Organismos
tasiRNA	21	Plantas
piRNA	26-31	Células germinativas
piRNA-like	24-30	<i>Drosophila</i> , <i>C. elegans</i>
rasiRNA	26-31	Plantas, animais
tiRNAs	30-40	Leveduras, animais
tRFs	17-26	animais
snoRNA	60-300	Eucariotos, procariotos
snRNA	~150	Eucariotos

ncRNAs : organização no genoma



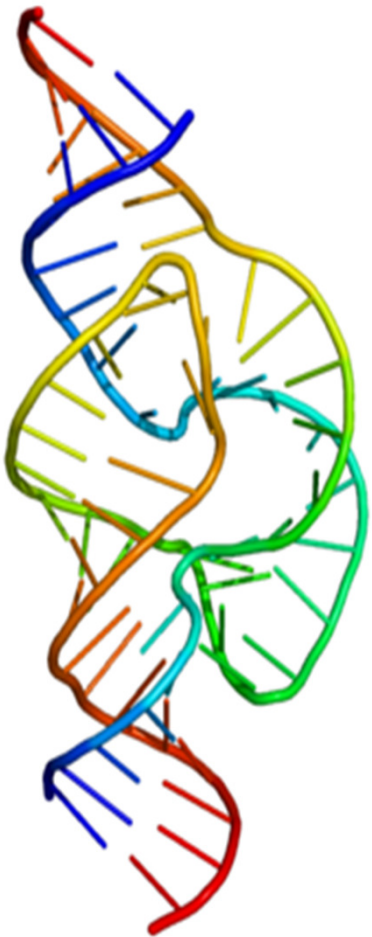
MicroRNA (miRNA)

- Micro RNA atua na regulação da expressão gênica de eucariotos superiores, diminuindo a expressão de seus alvos.
- Um único microRNA pode ter centenas de alvos diferentes.
- Mecanismo de complementaridade parcial com mRNAs, geralmente na região 3' UTR.



miRNAs: diferenças entre espécies

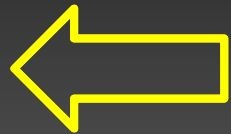
	Human	Fly
mature miRNAs	1100	186
known targets	111 (10%)	32 (17%)
pre-miRNAs	940	171
in introns	442 (47%)	60 (35%)
overlapping exons	47 (5%)	2 (1%)
overlapping UTRs	27 (3%)	7 (4%)
antisense	137 (15%)	18 (11%)
<10kb from another miRNA	297 (32%)	80 (47%)



- Quantos ncRNAs existem no genoma?
- Qual é o repertório total de estruturas e funções?
- Como as sequências de ncRNAs evoluem?
- Como podemos anotar essas sequências de ncRNAs?

Principais abordagens

- Micro-arranjos
- Bioinformática *ab initio*: análise composicional, predição de estrutura secundária, homologia estrutural ou de sequências, sinais de promotores e terminadores.
- RNA-Seq



RNA-Seq: metodologia

- A metodologia central do RNA-Seq é simples:
 - Isolar o RNA alvo (mRNA ou ncRNA).
 - Fragmentar o RNA
 - Síntese do cDNA.
 - Sequenciamento.
 - Mapear as sequências no genoma.
- O maior número de vezes que um determinado transcrito for detectado, maior a sua abundância.
- Se um número grande de sequências for gerada, é obtido uma visão abrangente e quantitativa do transcriptoma.

RNA-Seq: estratégias

RNA-Seq



Enriquecido com
mRNAs

TruSeq Stranded Total
RNA with Ribo-Zero
Human/Mouse/Rat

TruSeq Stranded Total
RNA with Ribo-Zero
Gold

TruSeq Stranded Total
RNA with Ribo-Zero
Globin




Enriquecido com
ncRNAs

TruSeq Small RNA

Bancos de dados biológicos

- rRNA 5S ribosomal db, RDP11, European rRNA db
- tRNA GtRDB, Sprinzl
- RNase P RNase P DB
- SRP SRPDB
- tmRNA tmRNA website, tmRNA database
- uRNAs uRNADB
- snoRNAs snoRNABase, snoRNAdb, snoopy
- miRNAs miRBase, miRNAmmap, microRNAdb



miRBase

MANCHESTER 1804

[Home](#)
[Search](#)
[Browse](#)
[Genomics](#)
[Help](#)
[Download](#)
[Submit](#)

Stem-loop sequence MI0005813

Accession	MI0005813
ID	dme-mir-375
Description	Drosophila melanogaster miR-375 stem-loop
Stem-loop	<pre> --c gc --uu g au goga a cgggca gaa acuuugggccaag ga gcaaacu uc u gocugu oaa ugaauuugggou oo uguuuga ag o uoa -a uuuo g -- ---a o </pre> <input type="button" value="Get sequence"/>
Genome context	Coordinates (BDGP5.0) 2L: 857542-857632 [+] Overlapping transcripts: intergenic View flanking features
Gene family	MIPF0000114; mir-375



Mature sequence MIMAT0005472

Accession	MIMAT0005472
ID	dme-mir-375
Sequence	55 - uuugucguuugcuuaguu - 76 <input type="button" value="Get sequence"/>
Evidence	experimental; 454 [1-2], Solexa [2]
Predicted targets	MICROCOSM: dme-mir-375 TARGETSCAN: dme-mir-375

Bancos de datos biológicos



Rfam 11.0 (August 2012, 2208 families)

The Rfam database is a collection of RNA families, each represented by **multiple sequence alignments**, **consensus secondary structures** and **covariance models (CMs)**. [More...](#)

QUICK LINKS

SEQUENCE SEARCH

VIEW AN RFAM FAMILY

VIEW AN RFAM CLAN

KEYWORD SEARCH

TAXONOMY SEARCH

JUMP TO

YOU CAN FIND DATA IN RFAM IN VARIOUS WAYS...

Analyze your RNA sequence for Rfam matches

View Rfam family annotation and alignments

View Rfam clan details

Query Rfam by keywords

Fetch families or sequences by NCBI taxonomy

Go

Example

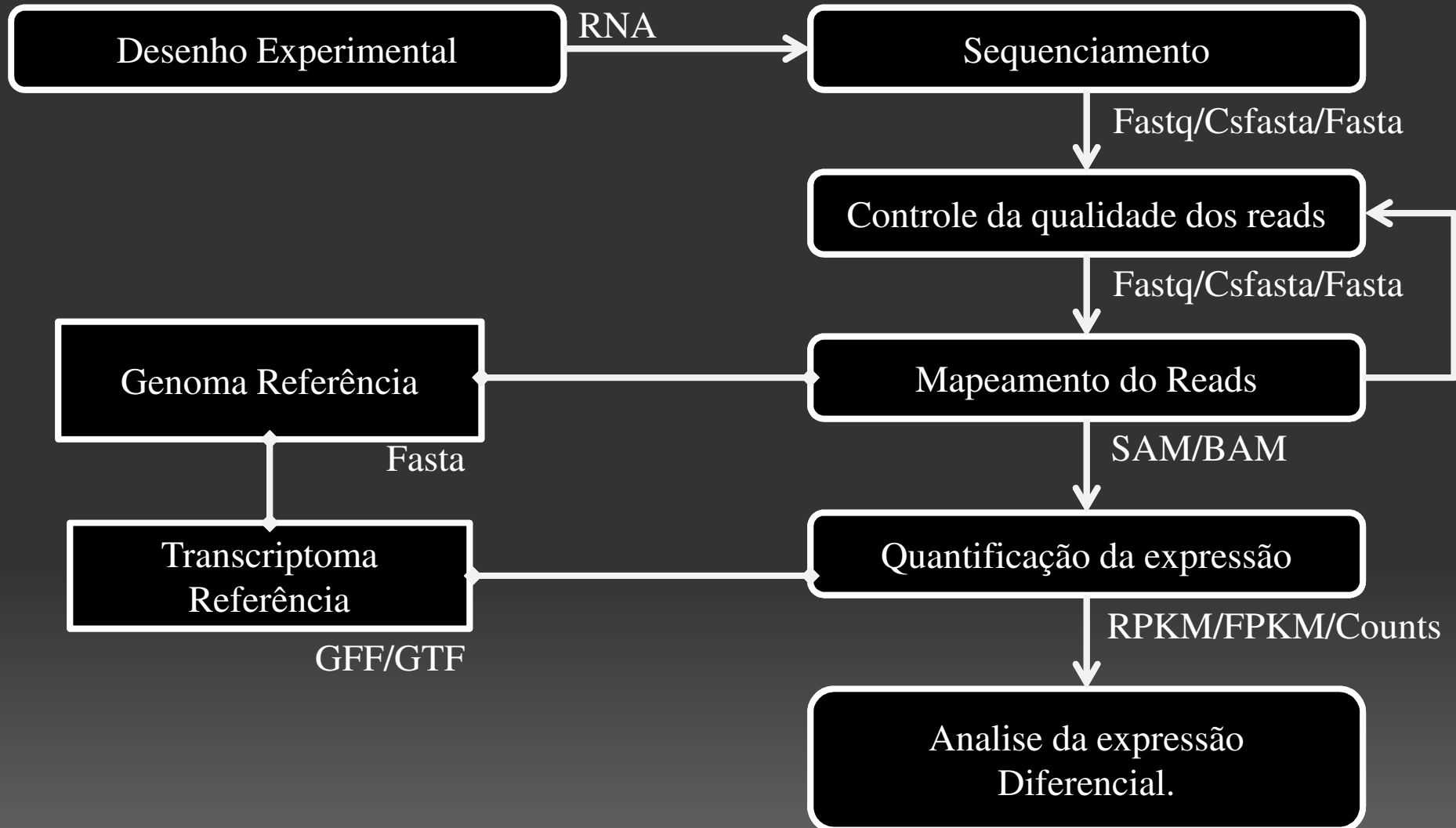
Enter any type of accession or ID to jump to the page for a Rfam family, sequence or genome



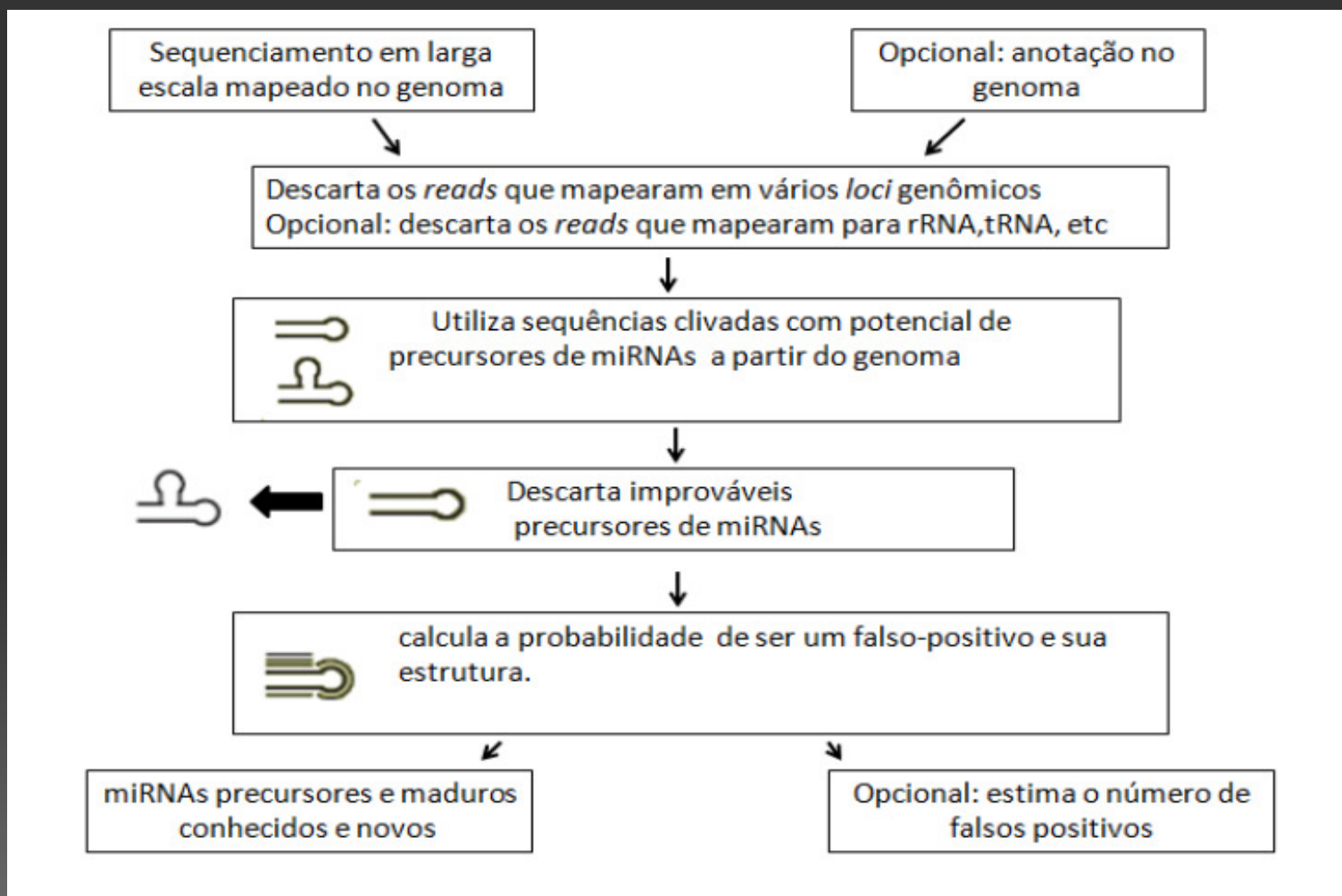
- Quantificar a expressão dos ncRNAs conhecidos.
- Identificar novos ncRNAs e sua localização no genoma.
- Quantificar expressão diferencial.
- Identificar mutações em ncRNAs.
- Anotar as sequências de ncRNA em bases biológicas.
- No caso de microRNAs, identificar os genes alvo.

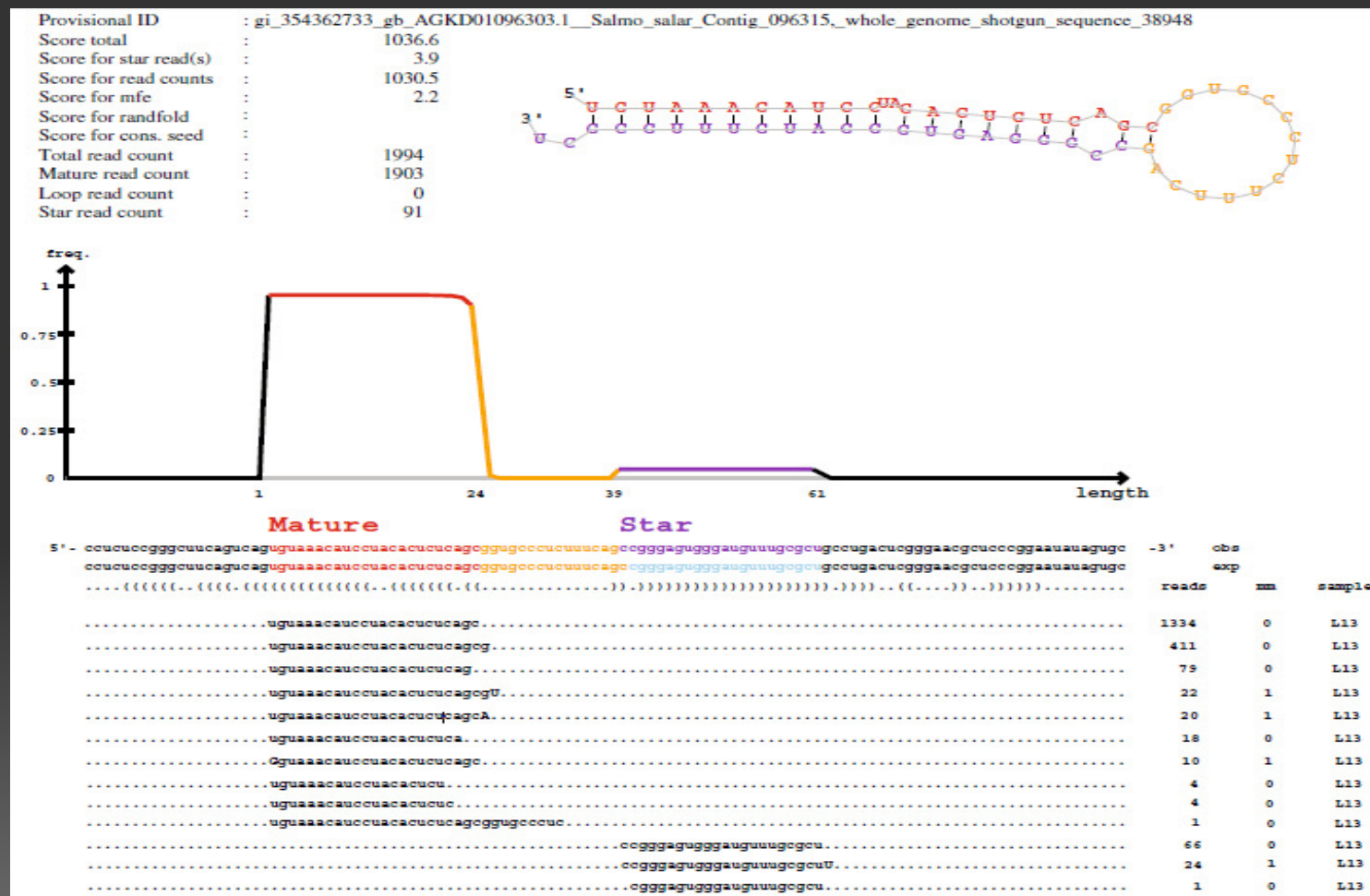


Fluxo de trabalho





Para microRNAs: MirDeep2

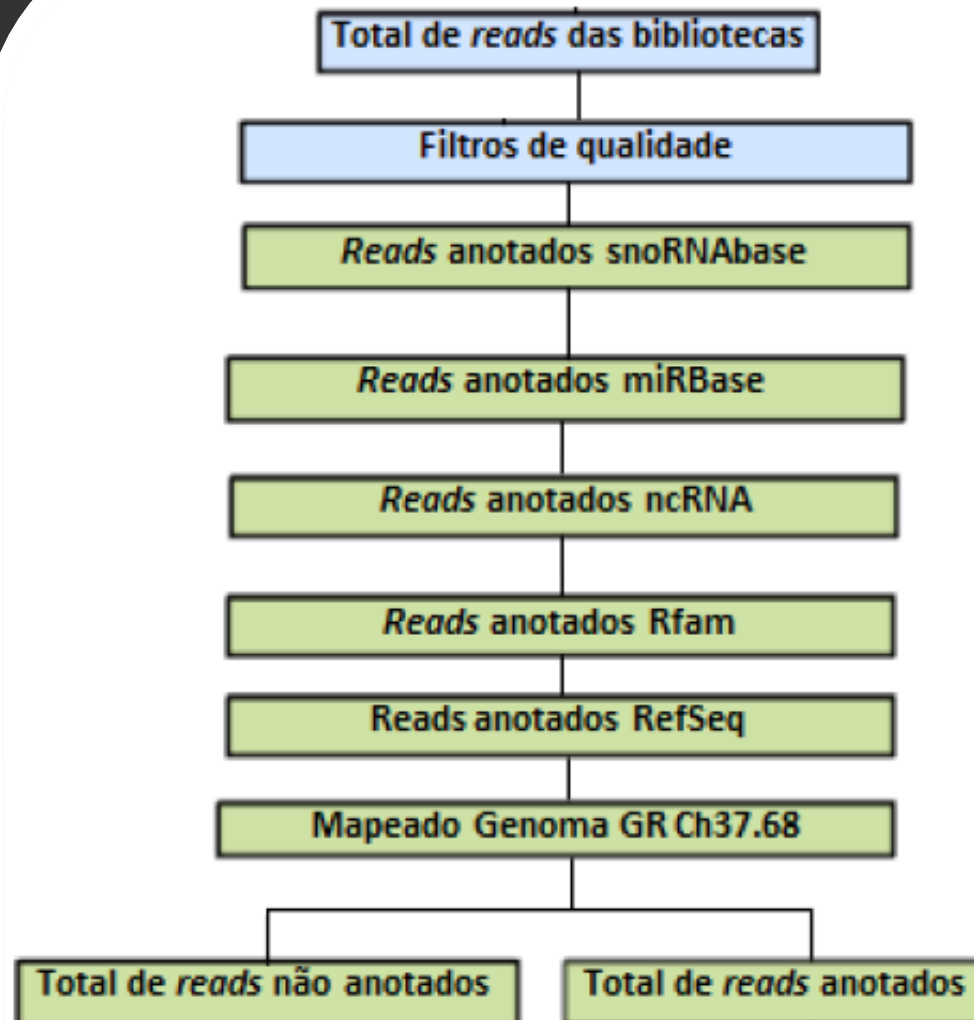




Para as demais famílias de ncRNA:

<p>Bowtie</p> <p>Extremely fast, general purpose short read aligner</p>	
<p>TopHat</p> <p>Aligns RNA-Seq reads to the genome using Bowtie</p> <p>Discovers splice sites</p>	

Busca de novos ncRNAs



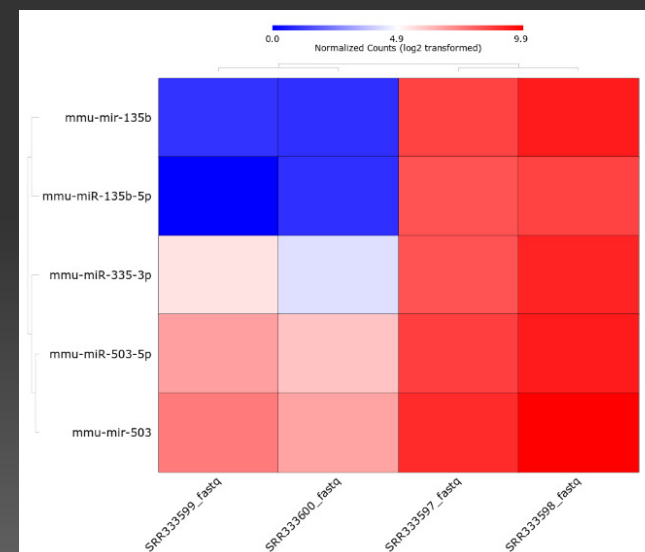
Expressão Diferencial

Resultado do MirDeep

hsa-let-7c-3p	17	hsa-let-7c
hsa-let-7d-5p	7719	hsa-let-7d
hsa-let-7d-3p	217	hsa-let-7d
hsa-let-7e-5p	2203	hsa-let-7e
hsa-let-7e-3p	7	hsa-let-7e
hsa-let-7f-5p	28898	hsa-let-7f-1
hsa-let-7f-1-3p	265	hsa-let-7f-1
hsa-let-7f-5p	28898	hsa-let-7f-1
hsa-let-7f-5p	28993	hsa-let-7f-2
hsa-let-7f-5p	28993	hsa-let-7f-2
hsa-let-7f-2-3p	45	hsa-let-7f-2
hsa-let-7g-5p	23244	hsa-let-7g
hsa-let-7g-3p	70	hsa-let-7g
hsa-let-7i-5p	659	hsa-let-7i
hsa-let-7i-3p	33	hsa-let-7i
hsa-miR-1	79	hsa-mir-1-1
hsa-miR-1	79	hsa-mir-1-1
hsa-miR-1	79	hsa-mir-1-2
hsa-miR-1	79	hsa-mir-1-2
hsa-miR-100-5p	6136	hsa-mir-100
hsa-miR-100-3p	1	hsa-mir-100
hsa-miR-101-5p	75	hsa-mir-101-1
hsa-miR-101-3p	7045	hsa-mir-101-1
hsa-miR-101-3p	7045	hsa-mir-101-1
hsa-miR-101-3p	7123	hsa-mir-101-2
hsa-miR-101-3p	7123	hsa-mir-101-2
hsa-miR-103a-3p	52629	hsa-mir-103a-1



(DESeq, EdgeR, SAMSeq)



Pipelines

sRNAMapper	Usa: FastQC, Bowtie or BWA
ncRNAAnnotator	Anota: tRNA, lincRNA, mt_tRNA, miRNA, rRNA, snRNA, snoRNA, and piRNA
CompaRNA	Faz Expressão Diferencial usando: DESeq, EdgeR, SAMSeq
ncRNAPredictor	Usa o Mirdeep2 para encontrar novos ncRNAs

Preparando os dados iniciais:

```
mkdir rna2
```

```
cd rna2
```

```
ln -s /home/treinamento/NGS/RNAseq/sample_data/SRR326279_R1.fastq .
```

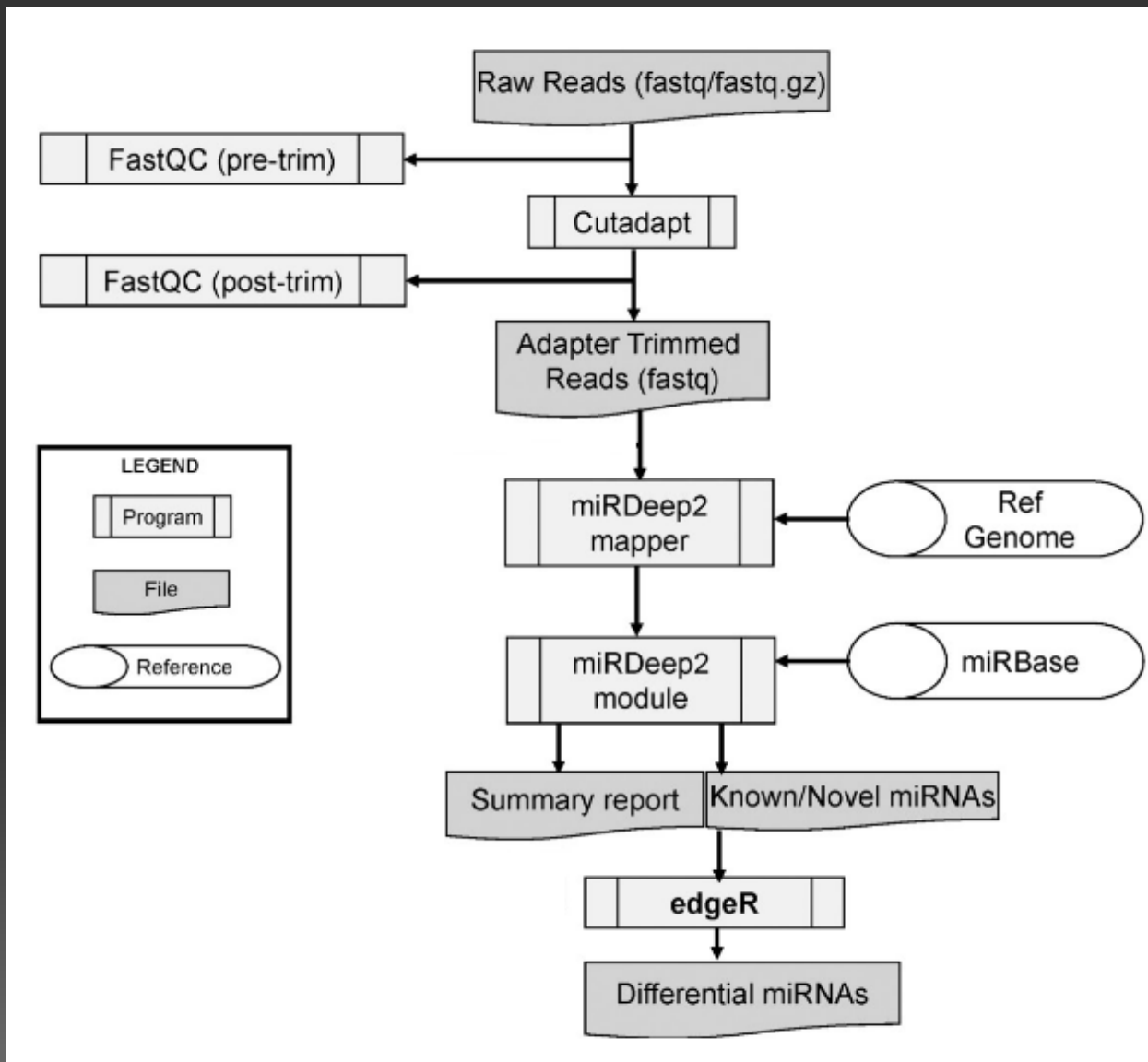
```
ln -s /home/treinamento/NGS/RNAseq/sample_data/SRR326280_R1.fastq .
```

```
ln -s /home/treinamento/NGS/RNAseq/sample_data/SRR326281_R1.fastq .
```

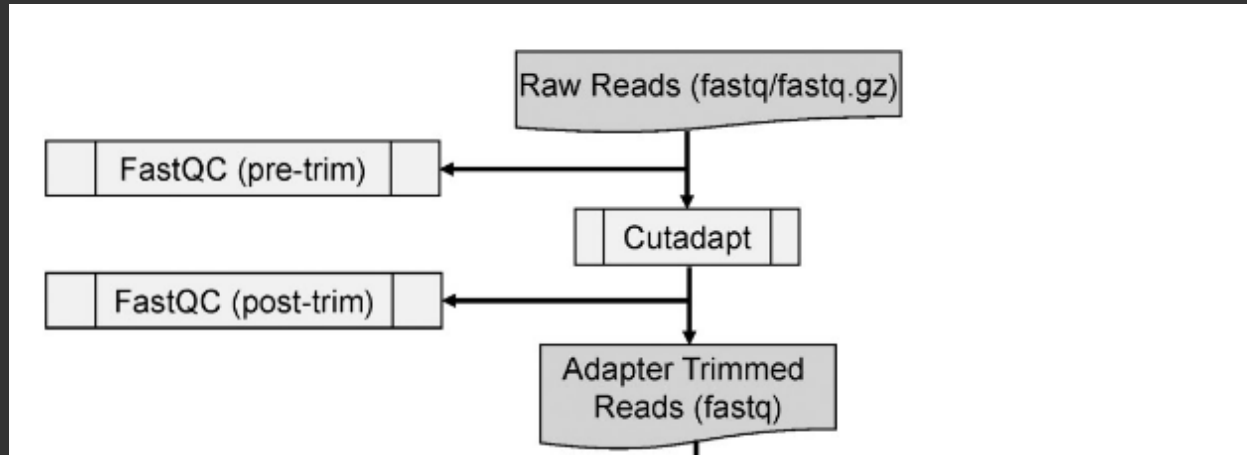
```
ln -s /home/treinamento/NGS/RNAseq/sample_data/SRR326282_R1.fastq .
```

```
ln -s /home/treinamento/NGS/RNAseq/small_ref/ .
```

Pipelines

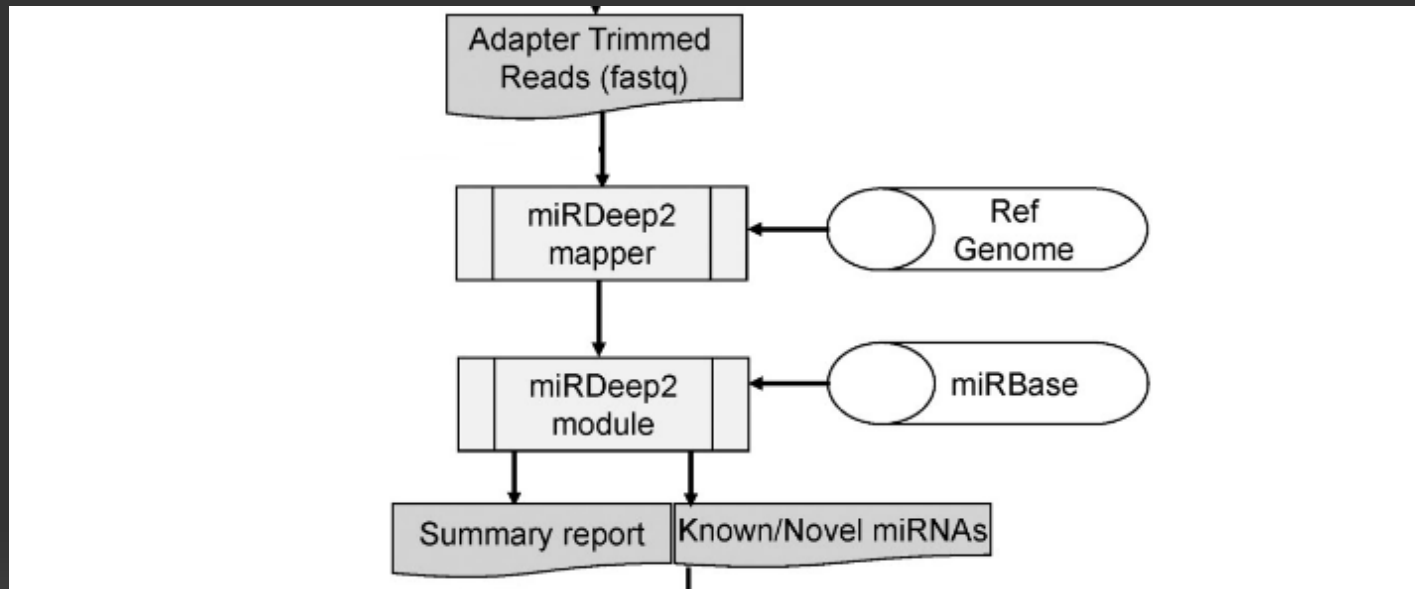


Pipelines

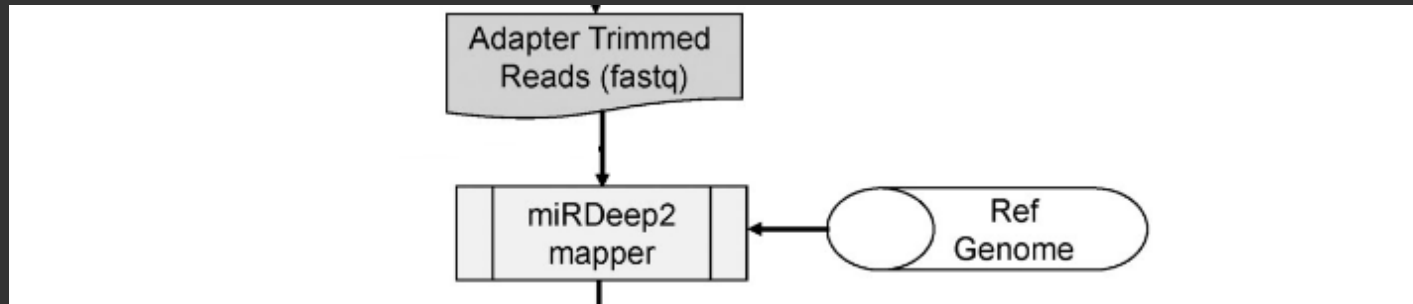


```
cutadapt -b AATCTCGTATGCCGTCTTCTGCTTGC -O 3 -m 17  
-f fastq SRR326279_R1.fastq > SRR326279_R1.ct.fastq
```

Pipelines



Pipelines



For instance, a typical `mapper.pl` command might look something like the following:

```
mapper.pl trimmed_cutadapt.fastq -e -p reference-genome -s processed_reads.fa -t mapped_reads.arf -h -m -i -j
```

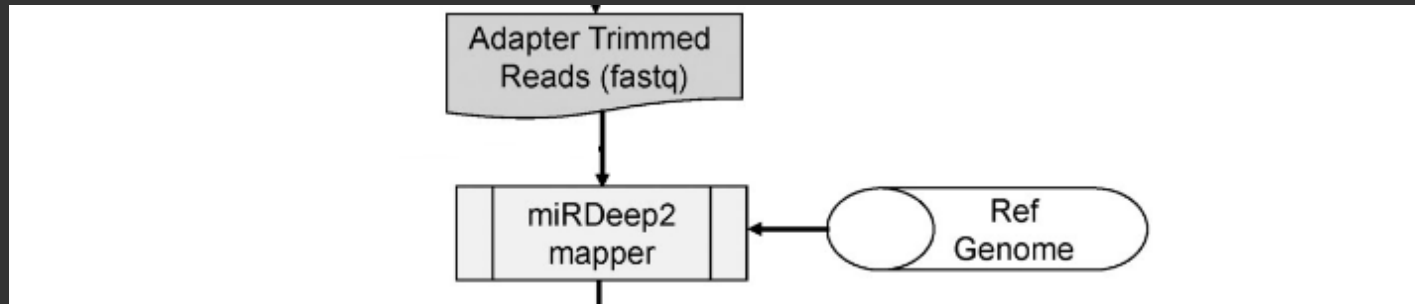
Where:

- "-e" input file is FASTQ format
- "-p" map to reference genome
- "-s" print processed reads to this file
- "-t" print reads mapping to this file
- "-h" parse to fasta format
- "-m" collapse reads
- "-i" convert RNA to DNA alphabet (to map against the genome)
- "-j" remove all entries that have a sequence that contains letters other than a,c,g,t,u,n,A,C,G,T,U,N

For more command options, use:

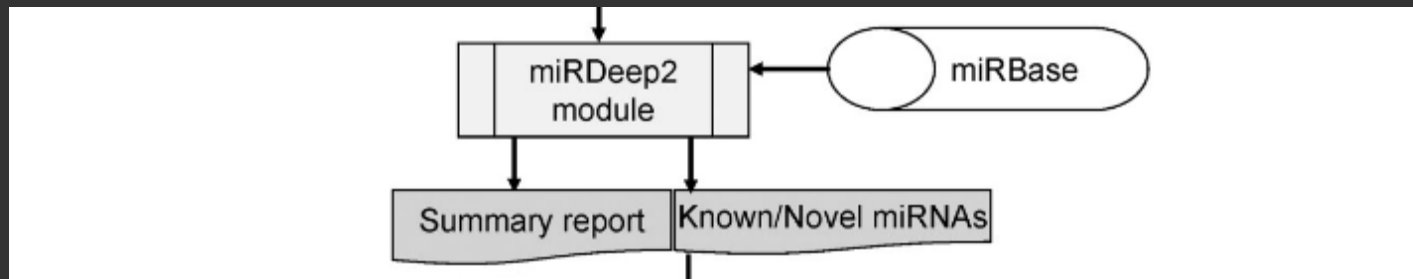
```
mapper.pl --help
```

Pipelines



```
mapper.pl SRR326279_R1.ct.fastq  
-e  
-p small_ref/hg19_chr1  
-s SRR326279.pr.fa  
-t SRR326279.mr.arf  
-h -m -i -j
```

Pipelines



For instance, a typical `miRDeep2.pl` command might look something like the following:

```
miRDeep2.pl processed_reads.fa genome.fa mapped_reads.arf mature.fa none  
hairpin.fa -t Mouse 2 > report.log
```

Where:

- "-t" species being analyzed

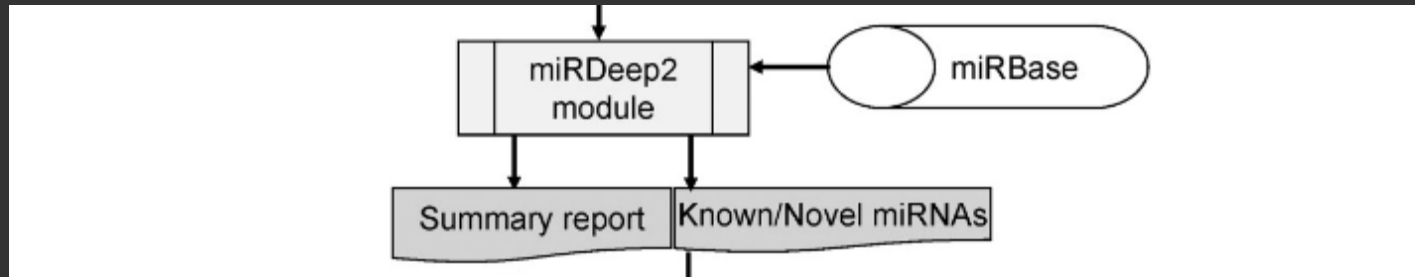
output:

This command will generate a directory with pdfs showing the structures, read signatures and score breakdowns of novel and known miRNAs in the data, an html webpage that links to all results generated (`result.html`), a copy of the novel and known miRNAs contained in the webpage but in text format which allows easy parsing (`result.csv`), a copy of the performance survey contained in the webpage but in text format (`survey.csv`) and a copy of the miRNA read signatures contained in the pdfs but in text format (`output.mrd`).

For more command options, use:

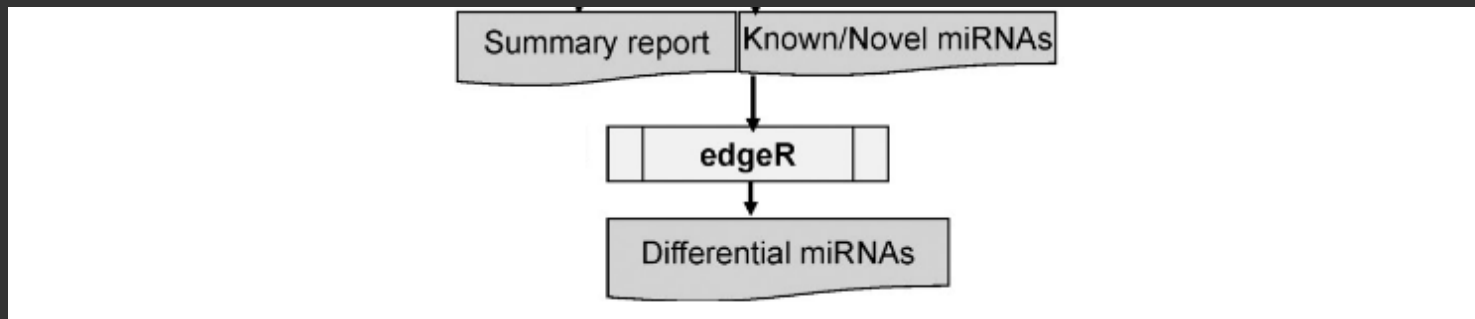
```
miRDeep2.pl --help
```

Pipelines



```
miRDeep2.pl SRR326279.pr.fa  
            small_ref/hg19_chr1.fa  
            SRR326279.mr.arf  
            small_ref/mature.hsa.dna.fa  
            none  
            small_ref/hairpin.hsa.dna.fa  
            -t Human 2> report.log
```

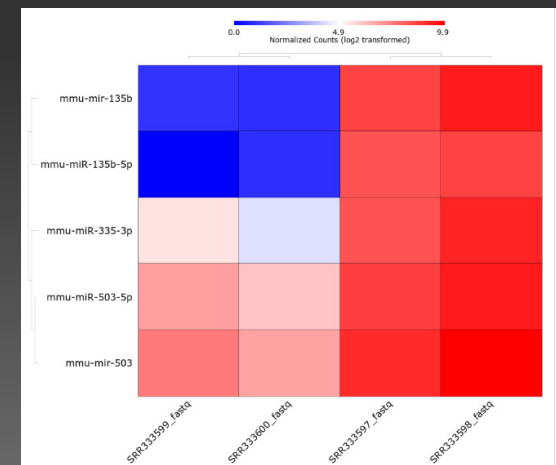

Pipelines



hsa-let-7c-3p	17	hsa-let-7c
hsa-let-7d-5p	7719	hsa-let-7d
hsa-let-7d-3p	217	hsa-let-7d
hsa-let-7e-5p	2203	hsa-let-7e
hsa-let-7e-3p	7	hsa-let-7e
hsa-let-7f-5p	28898	hsa-let-7f-1
hsa-let-7f-1-3p	265	hsa-let-7f-1
hsa-let-7f-5p	28898	hsa-let-7f-1
hsa-let-7f-5p	28993	hsa-let-7f-2
hsa-let-7f-5p	28993	hsa-let-7f-2
hsa-let-7f-2-3p	45	hsa-let-7f-2
hsa-let-7g-5p	23244	hsa-let-7g
hsa-let-7g-3p	70	hsa-let-7g
hsa-let-7i-5p	659	hsa-let-7i
hsa-let-7i-3p	33	hsa-let-7i
hsa-miR-1	79	hsa-mir-1-1
hsa-miR-1	79	hsa-mir-1-1
hsa-miR-1	79	hsa-mir-1-2
hsa-miR-1	79	hsa-mir-1-2
hsa-miR-100-5p	6136	hsa-mir-100
hsa-miR-100-3p	1	hsa-mir-100
hsa-miR-101-5p	75	hsa-mir-101-1
hsa-miR-101-3p	7045	hsa-mir-101-1
hsa-miR-101-3p	7045	hsa-mir-101-1
hsa-miR-101-3p	7123	hsa-mir-101-2
hsa-miR-101-3p	7123	hsa-mir-101-2
hsa-miR-103a-3p	52629	hsa-mir-103a-1



(DESeq, EdgeR, SAMSeq)



CURSO DE CURTA DURAÇÃO - 2017

BIOINFORMÁTICA

BioME – CENTRO MULTIUSUÁRIO DE BIOINFORMÁTICA - UFRN

Obrigado.

E-mail: jorge@imd.ufrn.br



Bioinformatics
Multidisciplinary
Environment

Centro
Multiusuário
de Bioinformática



INSTITUTO
METRÓPOLE
DIGITAL



²Bio
Instituto de
Bioinformática e
Biotecnologia