e-ISSN 2965-8047

Documentos

255

Concórdia, SC / Novembro, 2024

OBJETIVOS DE DESENVOLVIMENTO SUSTENTÁVEL

4 EDUCAÇÃO DE QUALIDADE

OBJETIVOS DE DESENVOLVIMENTO SUSTENTÁVEL

9 INDÚSTRIA, INOVAÇÃO E INFRAESTRUTURA

Explorando o universo dos microRNAs: manual para a predição de novos microRNAs

Embrapa
Suínos e Aves

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Suínos e Aves
Ministério da Agricultura e Pecuária

e-ISSN 2965-8047

Documentos 255

Novembro, 2024

Explorando o universo dos microRNAs: manual para a predição de novos microRNAs

Francelly Geralda Campos
Adriana Mércia Guaratini Ibelli
Haniel Cedraz de Oliveira
Maurício Egídio Cantão
Jane de Oliveira Peixoto
Mônica Corrêa Ledur
Simone Eliza Facioni Guimarães

Embrapa Suínos e Aves
Concórdia, SC
2024

Embrapa Suínos e Aves
Rodovia BR 153 - KM 110
89.715-899, Concórdia, SC
www.embrapa.br/fale-conosco/sac

Revisão de texto
Jean Carlos Porto Vilas Boas Souza

Projeto gráfico
Leandro Sousa Fazio

Diagramação
Vivian Fracasso

Foto da capa
stablediffusionweb.com

Publicação digital: PDF

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Suínos e Aves

Explorando o universo dos microRNAs: manual para a predição de novos microRNAs / Francelly Geralda Campos [*et al.*] – Concórdia : Embrapa Suínos e Aves, 2024.
PDF (20 p.) : il. color. – (Documentos / Embrapa Suínos e Aves, e-ISSN 2965-8047; 255)

1. Genética. 2. Bioinformática. I. Campos, Francelly Geralda. II. Ibelli, Adriana Mércia Guaratini. III. Oliveira, Haniel Cedraz de. IV. Cantão, Maurício Egídio. V. Peixoto, Jane de Oliveira. VI. Ledur, Mônica Corrêa. VII. Guimarães, Simone Eliza Facioni. VIII. Título. IX. Série.

CDD (21. ed.) 636.082 1

Claudia Antunez Arrieche (CRB-14/880)

Autores

Adriana Mércia Guaratini Ibelli
Bióloga, doutora em Genética Evolutiva e Biologia Molecular, analista da Embrapa Pecuária Sudeste, São Carlos, SP

Francelly Geralda Campos
Zootecnista, doutora em Zootecnia pela Universidade Federal de Viçosa, Viçosa, MG

Haniel Cedraz de Oliveira
Médico Veterinário, doutor em Zootecnia, Universidade Federal de Viçosa, Viçosa, MG

Jane de Oliveira Peixoto
Zootecnista, doutora em Zootecnia, pesquisadora da Embrapa Suínos e Aves, Concórdia, SC

Maurício Egídio Cantão
Bacharel em Processamento de Dados, doutor em Bioinformática, pesquisador da Embrapa Suínos e Aves, Concórdia, SC

Mônica Corrêa Ledur
Zootecnista, doutora em Genética e Melhoramento Animal, pesquisadora da Embrapa Suínos e Aves, Concórdia, SC

Simone Eliza Facioni Guimarães
Médica Veterinária, doutora em Ciência Animal, professora titular do Departamento de Zootecnia da Universidade Federal de Viçosa, Viçosa, MG

Apresentação

Os microRNAs (miRNAs) são uma classe importante de RNAs não codificantes. Estes são pequenos RNAs com tamanho entre 18 a 25 nucleotídeos de comprimento, localizados em regiões intrônicas ou exônicas. Os miRNAs, além de regular a expressão gênica, estão envolvidos na regulação de diversos processos biológicos, incluindo o ciclo celular, a diferenciação e o metabolismo. Em 2024, os cientistas Gary Ruvkun e Victor Ambros receberam o prêmio Nobel de medicina em reconhecimento pela descoberta dos microRNAs na década de 1990 que tem permitido elucidar parte de uma complexa rede de regulação da expressão dos genes e manutenção dos processos biológicos. Os avanços no sequenciamento de nova geração revolucionaram a capacidade de estudar os miRNAs, proporcionando maior detecção de miRNAs e permitindo a identificação em diferentes tecidos e estágios de desenvolvimento. No entanto, a tarefa computacional de identificar com precisão os miRNAs ainda permanece desafiadora devido à complexidade dos dados gerados e às sutis diferenças entre sequências. A análise bioinformática de dados de sequenciamento de miRNA é essencial para desvendar os complexos mecanismos de regulação genética e entender seu papel em uma variedade de processos biológicos. Assim, o tutorial elaborado fornece um passo a passo para a identificação de miRNAs, a partir de dados de sequenciamento de RNA em grande escala (RNA-Seq), utilizando como exemplo os suínos, mas que pode ser usado para diferentes espécies, fornecendo à comunidade científica um protocolo para conduzir suas análises de forma prática e eficiente. Ao seguir o fluxo de trabalho mencionado neste tutorial, os estudantes e pesquisadores poderão identificar microRNAs e posteriormente elucidar as vias biológicas e processos associados à sua regulação. Essa análise pode fornecer insights valiosos, permitindo uma compreensão mais abrangente dos sistemas biológicos, das interações entre diferentes níveis de regulação genética e melhor direcionar pesquisas futuras. A identificação de novos miRNAs por meio de análises bioinformáticas é de grande importância para entendermos diversos processos regulatórios, melhorando, assim, o conhecimento das regiões funcionais do genoma.

Mônica Corrêa Ledur
Pesquisadora da Embrapa Suínos e Aves

Sumário

Introdução 9

Ferramentas para identificação de microRNAs 10

Controle de qualidade 10

Mapeamento com Bowtie 11

Mapeamento e quantificação - mirDeep2 11

Passo-a-passo para a identificação de microRNAs 12

Controle de qualidade 12

Mapeamento 14

Identificação e predição dos microRNAs – mirDeep2 17

Considerações finais 19

Agradecimentos 19

Referências 19

Introdução

Os RNAs não codificantes (ncRNAs) foram considerados como “ruído genômico” ou “lixo” das regiões genômicas durante parte dos anos 2000 (Zhang *et al.*, 2019). No entanto, com os avanços nas técnicas de sequenciamento, foi possível comprovar que os ncRNAs desempenham papel fundamental em diversos processos biológicos (Ye *et al.*, 2021; López-Jiménez, Andrés-León, 2021). Entre os diversos tipos de ncRNAs, os microRNAs têm recebido considerável atenção devido à sua capacidade de regular a expressão gênica pós-transcricional (Jonas, Izaurralde, 2015).

Os microRNAs são pequenos RNAs de aproximadamente 18-25 nucleotídeos de comprimento, que atuam como reguladores negativos da expressão gênica por meio do pareamento de bases com sequências-alvo complementares nos RNAs mensageiros (mRNAs). Essa interação pode levar à degradação do mRNA-alvo ou à inibição da sua tradução, influenciando assim diversos processos celulares (Lin *et al.*, 2013; O’Brien *et al.*, 2018). Os microRNAs estão localizados em regiões genômicas intrônicas ou nos éxons de genes codificadores de proteínas (Kim; Kim 2007).

O microRNA é tipicamente transcrito pela RNA polimerase II como um RNA primário, que é posteriormente processado em uma forma de precursor de microRNA em *hairpin* (pré-microRNA) pela enzima RNase III Drosha no núcleo e, então, exportado para o citosol pela exportina-5 (Bartel, 2004). No citosol, o pré-microRNA é clivado pela enzima Dicer em um duplex de microRNA, onde um dos braços é carregado na proteína Argonauta (AGO) dentro do complexo de silenciamento induzido por RNA (RISC) e usado como sequência guia para direcionar a ligação ao mRNA-alvo (Krol *et al.*, 2010). Um microRNA pode influenciar simultaneamente vários genes localizados na mesma via de sinalização celular (Li *et al.*, 2015). Eles interagem com seus alvos, que incluem não apenas mRNAs, mas também RNAs não codificantes longos (lncRNAs), pseudogenes e RNAs circulares (Tay *et al.*, 2014).

As abordagens biológicas e bioinformáticas permitem a descoberta de milhares de microRNAs em plantas e animais, os quais são depositados no miRBase, o principal repositório online de sequências e anotações de microRNAs. Sua versão mais recente (v22) abrange sequências de microRNA de 271 organismos, totalizando 38.589 precursores e 48.860 microRNAs maduros (Kozomara; Birgaoanu; Griffiths-Jones, 2019). Os avanços no sequenciamento de nova geração revolucionaram a capacidade de estudar os microRNAs, proporcionando maior detecção de microRNAs e permitindo a identificação em diferentes tecidos e estágios de desenvolvimento. No entanto, a tarefa computacional de identificar com precisão os microRNAs ainda permanece desafiadora devido à complexidade dos dados gerados e às sutis diferenças entre sequências.

A identificação do microRNA não é uma tarefa simples e direta, mas sim um processo complexo que envolve a integração de diferentes disciplinas e abordagens. A complexidade surge devido à natureza dos microRNAs, que são moléculas pequenas e com função altamente influente na regulação genética, e também devido à diversidade de métodos e técnicas necessárias para sua identificação e caracterização. Portanto, uma estratégia interdisciplinar, que combina conhecimentos e técnicas de áreas como biologia molecular, bioinformática e genômica, é necessária para a identificação eficaz e completa dos microRNAs. Assim, este manual tem como propósito de fornecer informações detalhadas e um guia completo para a identificação de microRNAs a partir de dados de sequenciamento em larga escala.

O manual está estruturado da seguinte maneira: (i) A seção 2 aborda as etapas e as principais ferramentas empregadas para a identificação de microRNAs, proporcionando uma visão geral da análise proposta. (ii) Na seção 3, essas ferramentas são aplicadas, seja por meio de scripts ou diretamente através de comandos do sistema operacional Linux, distribuição CentOS release 7.6 (usada no manual), em um guia passo a passo que demonstra como realizar a análise utilizando um conjunto de dados de sequenciamento de RNA de fetos suínos.

O presente trabalho está alinhado aos Objetivos de Desenvolvimento Sustentável (ODS): 4 – Educação e Qualidade, e contribuem para o atingimento da meta 4.3 – Técnica e superior; e também o ODS 9 – Indústria, Inovação e Infraestrutura, e contribuem

para o atingimento da meta 9.5 -Pesquisa científica e capacidade tecnológica.

Ferramentas para identificação de microRNAs

Para conduzir as análises mencionadas neste guia, são indispensáveis dados de sequenciamento de microRNAs, obtidos por meio de bibliotecas de RNA preparadas com o kit Illumina TruSeq Small RNA (Illumina). Aqui, utilizamos arquivos brutos de sequências no formato FASTQ e as análises são executadas em um ambiente Linux, utilizando a linguagem de shell script.

As etapas para identificar microRNAs incluem controle de qualidade, mapeamento e quantificação das sequências. Os programas utilizados são: Trimmomatic (Bolger *et al.*, 2014), Cutadapt (Martin, 2011), Bowtie (Langmead, 2010) e miRDeep2 (Friedländer *et al.*, 2012). Abaixo, será fornecido um resumo conciso dos softwares empregados em cada fase da análise em dados de sequenciamento de RNA provenientes de amostras de fetos de suínos.

Controle de qualidade

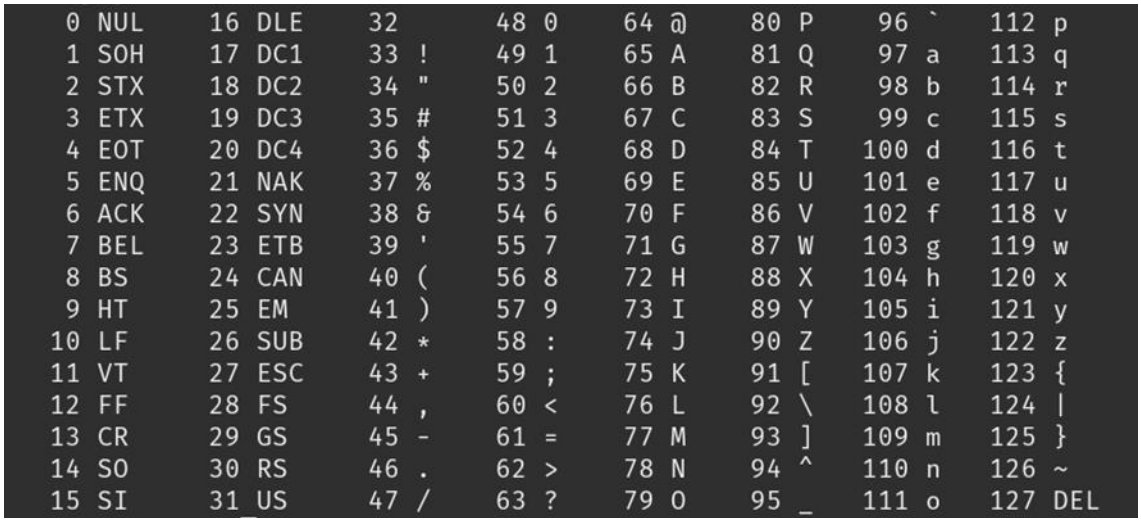
O objetivo do controle de qualidade consiste em remover bases e sequências de baixa qualidade, bem como sequências de adaptadores, para reduzir erros de sequenciamento e aumentar a qualidade dos dados analisados. Neste manual, utilizamos os programas Cutadapt para eliminar as sequências de adaptadores e Trimmomatic para remover sequências de baixa qualidade. O programa Cutadapt foi projetado para remover adaptadores de sequências curtas geradas a partir das novas plataformas de sequenciamento de DNA. No sequenciamento de pequenos RNAs, as leituras frequentemente são mais longas do que o próprio RNA, contendo partes do adaptador 3'. A remoção desses adaptadores é crucial para garantir a qualidade e a precisão das análises subsequentes (Martin, 2011). O programa Trimmomatic é uma ferramenta versátil projetada para pré-processar leituras de sequenciamento de alto rendimento, removendo bases e sequências de baixa qualidade. Ele emprega uma abordagem de janela deslizante para remover tais regiões. A remoção dessas sequências e regiões contribui para o aumento de precisão e a confiabilidade das análises, tornando-o uma ferramenta essencial para o processamento de dados de sequenciamento em diversas aplicações (Bolger *et al.* 2014).

Para usar os programas Cutadapt e Trimmomatic é necessário que os arquivos de sequências estejam no formato FASTQ (Figura 1). Esse tipo de arquivo foi desenvolvido para lidar com a saída das métricas básicas de qualidade de sequenciamento. As pontuações de qualidade e sequência são codificadas como caracteres ASCII (*American Standard Code for Information Interchange*). Um código ASCII é a representação numérica de um caractere (Figura 2). O formato utiliza quatro linhas para cada sequência. A primeira linha contém o cabeçalho iniciado pelo caractere '@', seguido pelo identificador de sequência e uma descrição opcional. A segunda linha corresponde à bases sequenciadas. A terceira linha inicia com o caracter "+" e pode ser utilizado para adicionar qualquer tipo de informação. Por fim, a quarta linha do arquivo FASTQ apresenta os valores de qualidade para cada base sequenciada, onde cada caractere, depois de convertido, representa o valor da qualidade Phred.

As quatro linhas de um arquivo FASTQ representam uma sequência, também chamada de *read* (leitura, em português). Uma *read* consiste em uma única sequência de nucleotídeos gerada pelo sequenciador de DNA. O tamanho da *read* pode variar de acordo com o equipamento e a biblioteca de sequenciamento utilizados. Os sequenciadores podem gerar dois tipos de *reads*: *single-end* (sequenciamento de uma extremidade) e paired-end (sequenciamento de ambas as extremidades). A limpeza e filtragem das *reads* é uma etapa importante para o pré-processamento dos dados brutos gerados.

```
@VH00451:1:AAANG3CM5:1:1101:39893:1000 1:N:0:CACCGG
GTCCCGTAGAACCGACCTTGCGTGGAATTCTCGGGT
+
C;CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
```

Figura 1. Representação de um arquivo FASTQ.



0	NUL	16	DLE	32		48	0	64	@	80	P	96	`	112	p
1	SOH	17	DC1	33	!	49	1	65	A	81	Q	97	a	113	q
2	STX	18	DC2	34	"	50	2	66	B	82	R	98	b	114	r
3	ETX	19	DC3	35	#	51	3	67	C	83	S	99	c	115	s
4	EOT	20	DC4	36	$	52	4	68	D	84	T	100	d	116	t
5	ENQ	21	NAK	37	%	53	5	69	E	85	U	101	e	117	u
6	ACK	22	SYN	38	&	54	6	70	F	86	V	102	f	118	v
7	BEL	23	ETB	39	'	55	7	71	G	87	W	103	g	119	w
8	BS	24	CAN	40	(	56	8	72	H	88	X	104	h	120	x
9	HT	25	EM	41	)	57	9	73	I	89	Y	105	i	121	y
10	LF	26	SUB	42	*	58	:	74	J	90	Z	106	j	122	z
11	VT	27	ESC	43	+	59	;	75	K	91	[	107	k	123	{
12	FF	28	FS	44	,	60	<	76	L	92	\	108	l	124	|
13	CR	29	GS	45	-	61	=	77	M	93	]	109	m	125	}
14	SO	30	RS	46	.	62	>	78	N	94	^	110	n	126	~
15	SI	31	US	47	/	63	?	79	O	95	_	111	o	127	DEL

Figura 2. Caracteres da tabela ASCII.

Mapeamento com Bowtie

O Bowtie é um alinhador rápido de sequência curta que utiliza a memória de forma eficiente e é voltado para alinhar grandes conjuntos de sequências curtas com genomas de referência. Antes de alinhar as *reads*, é necessário construir ou obter um índice apropriado do genoma de referência. Uma vez criado, um índice pode ser reutilizado para múltiplos alinhamentos. O programa Bowtie é compatível com os sistemas operacionais Linux, Mac OS X e Windows, possui código aberto e pode ser usado gratuitamente, facilitando a integração com outros softwares de bioinformática (Langmead, 2010).

Mapeamento e quantificação - mirDeep2

O miRDeep2 é um programa que identifica microRNAs e informa sobre candidatos de alta confiança que são detectados em múltiplas amostras independentes. O funcionamento é iniciado com uma verificação do formato dos arquivos de entrada. Em seguida, ocorre uma rápida quantificação dos microRNAs conhecidos se forem fornecidos arquivos de precursores do miRBase. Posteriormente, são removidos potenciais precursores de microRNA do genoma com base nos mapeamentos de leitura como guia. As leituras são analisadas para reter apenas mapeamentos de pelo menos 18 nucleotídeos (nt). As duas cadeias do genoma são escaneadas separadamente e a remoção é iniciada quando um agrupamento de leituras é encontrado. As sequências cobertas pelo agrupamento de leituras são removidas e a varredura continua até que não haja mais agrupamentos de leituras dentro de 70 nt. Após, é preparado um arquivo de assinatura utilizando-se a ferramenta bowtie-build para construir um índice de transformação de Burrows-Wheeler dos precursores removidos (Manzini, 2001; Li; Durbin, 2009), seguido pelo mapeamento do conjunto de leituras de sequenciamento para o índice usando bowtie. Na quarta etapa, as estruturas secundárias de RNA dos precursores potenciais são previstas usando o RNAfold e os valores de P randfold são calculados para um subconjunto dos precursores. O RNAfold é um programa que calcula a estrutura secundária de uma sequência de RNA. Isso é crucial para identificar estruturas de *hairpin* características dos precursores de microRNAs. O P randfold é uma medida estatística que ajuda a avaliar a probabilidade de que a estrutura de RNA prevista ocorra por acaso. Valores baixos de P randfold indicam que a estrutura observada é estatisticamente significativa e não é apenas um artefato aleatório. Na quinta etapa, os precursores potenciais são pontuados individualmente ou descartados pelo algoritmo central miRDeep2. Na sexta etapa, são examinadas as distribuições de pontuação e calculadas as estatísticas de desempenho (Friedländer *et al.*, 2012).

Passo-a-passo para a identificação de microRNAs

Esta seção apresenta um procedimento passo a-passo que ilustra como realizar a análise para identificação de microRNAs, utilizando um conjunto de dados de RNA-Seq de um experimento com fetos de suínos (*Sus scrofa*). As duas amostras são provenientes do sequenciamento *single-end* Illumina, utilizando o kit TruSeq Small RNA. O sequenciamento foi de 36pb com a plataforma Illumina HiSeq2500. As amostras usadas neste manual foram coletadas de fetos de suínos de 35 dias, sendo uma amostra do sexo feminino (pig_mir_01) e outra do sexo masculino (pig_mir_02). As amostras utilizadas neste manual estão no formato FASTQ e compactadas pelo programa gzip. Os comandos correspondentes à cada etapa da análise devem ser executados no terminal do Linux, a partir do prompt de comandos. Em todos os casos, será apresentada uma descrição dos parâmetros usados e da saída (arquivo) resultante.

Primeiramente, é necessário criar um diretório para armazenar todos os programas e amostras usados neste documento. É importante que o nome do arquivo/diretório não possua acentos nem espaços.

```
# Todos os comandos são executados no terminal bash (Bourne Again SHell) do Linux

# Crie (mkdir) a pasta Analise_microRNAs

$ mkdir Analise_microRNAs

# Entre (cd) na pasta Analise_microRNAs

$ cd Analise_microRNAs
```

Os arquivos necessários para execução dessas análises estão disponíveis no link (https://drive.google.com/drive/folders/18MJPLeagn5t_6h1q3HEnhvtYIjmsm1lg). Abra o link em um navegador e faça o download dos arquivos pig_mir_01.fastq.gz e pig_mir_02.fastq.gz.

Crie o diretório 00-Fastq e mova (mv) os dois arquivos baixados para dentro desta pasta.

```
$ mkdir 00-Fastq

$ mv *.gz 00-Fastq
```

Controle de qualidade

Para maior confiabilidade, o controle de qualidade é realizado pelos programas Cutadapt e Trimmomatic. Certifique-se que você esteja na pasta que tenha os programas e os arquivos necessários para iniciar o controle de qualidade.

Controle de qualidade – Cutadapt

No terminal do sistema operacioal Linux, instale o programa Cutadapt com o comando abaixo:

```
#Versão do programa Cutadapt usada no manual foi 4.9

$ pip install cutadapt

#Para maiores informações sobre a instalação e uso visite a página do programa em:

https://cutadapt.readthedocs.io/en/stable/
```

O comando nohup do Linux permite que o processo continue executando mesmo se o terminal for fechado.

Será utilizado o caractere “>” para redirecionar a saída padrão do comando (tela) para um arquivo texto.

Será utilizado o comando “for” do linux para fazer um loop e executar todas as amostras simultaneamente.

O programa Cutadapt remove um adaptador específico das reads utilizando os seguintes parâmetros/arquivos abaixo:

- **a** → adaptador usado no kit;
- **m** → comprimento mínimo das reads;
- **j** → número de threads;
- ***max-n 0*** → remover reads com bases “N”;
- **o** → nome do arquivo de saída;
- **pig_mir_01_fastq.gz** → nome do arquivo de entrada.
- **pig-mir-01.report.cutadapt.txt** → arquivo de relatório. Todas as mensagens de saída e estatísticas do processo de corte serão salvas neste arquivo.

Para rodar o programa cutadapt execute os seguinte passos:

#Crie a pasta onde será salvo os resultados da análise do programa Cutadapt

$ mkdir 01-Cleaned

#Criando uma variável de nome “lib” contendo o nome das duas amostras

$ lib=(pig_mir_01 pig_mir_02)

 # Rodando as duas amostras simultaneamente usando o comando for

$ for i in ${lib[@]}; do nohup cutadapt -a TGGAT TCTCGGGTGCCAAGG -m 18 -j 5 --max-n 0 -o 01-Cleaned/”$i”_cut.fastq.gz 00-Fastq/”$i”.fastq.gz > 01-Cleaned/”$i”.report.cutadapt.txt & done

Arquivos após a execução do Cutadapt (Figura 3):

```
pig_mir_01_cut.fastq.gz
pig_mir_01.report.cutadapt.txt
pig_mir_02_cut.fastq.gz
pig_mir_02.report.cutadapt.txt
```

Figura 3. Arquivos gerados pela saída do programa Cutadapt.

Controle de qualidade – Trimmomatic

Para baixar o programa Trimmomatic versão 1.39 (usado no manual) execute os comandos abaixo:

- **wget** → comando usado para baixar arquivos da internet.
- **unzip** → descompacta os arquivos “.zip”.

$ wget http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.39.zip

$ unzip Trimmomatic-0.39.zip

 # Utilize o comando java -jar para executar programas feitos em linguagem Java.

$ java -jar Trimmomatic-0.39/trimmomatic-0.39.jar

#Para maiores informações sobre a instalação e uso visite a página do programa em:

http://www.usadellab.org/cms/?page=trimmomatic

O Trimmomatic é uma ferramenta amplamente utilizada para o pré-processamento de dados de sequenciamento. Uma das suas funções é realizar o controle de qualidade dos arquivos de sequenciamento, o que inclui a detecção e remoção de leituras curtas e de baixa qualidade. Segue abaixo uma explicação para os parâmetros utilizados para execução do Trimmomatic:

- **SE** → usada para indicar que a filtragem de qualidade é em single-end;
- **threads** → número de threads;
- **pig_mir_01_cut_trimmo.fastq.gz** → nome do arquivo de saída;
- **SLIDINGWINDOW:5:20** → esta opção é usada para realizar uma filtragem de qualidade de deslizamento de janela por base. O número 5 especifica o tamanho da janela, enquanto 20 é o valor mínimo da qualidade média dentro da janela;
- **LEADING:3** → remove bases com qualidade menor que 3 no início das sequências.
- **TRAILING:3** → remove bases com qualidade menor que 3 no final das sequências;
- **pig_mir_01.report.trimmo.txt** → arquivo de texto que contém um relatório sobre o processo de filtragem realizado.

 # Para rodar as duas amostras simultaneamente

$ for i in ${lib[@]}; do java -jar Trimmomatic-0.39/trimmomatic-0.39.jar SE -threads 10 01-Cleaned/”$i”_cut.fastq.gz 01-Cleaned/”$i”_cut_trim.fastq.gz SLIDINGWINDOW:5:20 LEADING:3 TRAILING:3 MINLEN:18 2> 01-Cleaned/”$i”.report.trimmo.txt & done

Arquivos de saída após a execução do Trimmomatic (Figura 4):

```
pig_mir_01_cut_trim.fastq.gz
pig_mir_01.report.trimmo.txt
pig_mir_02_cut_trim.fastq.gz
pig_mir_02.report.trimmo.txt
```

Figura 4. Arquivos gerados pela saída do programa Trimmomatic.

Mapeamento

O mapeamento é realizado com o alinhador bowtie e a ferramenta mapper do mirDeep2. Antes de começar o mapeamento, é fundamental realizar o download de todos os arquivos necessários para a análise. Isso inclui as sequências de RNA transportador (tRNA) e RNA ribossômico (rRNA) do Rfam. Rfam é um banco de dados de famílias de RNA não codificantes, onde cada família é representada por um alinhamento de sequências múltiplas, uma estrutura secundária de consenso e um modelo de covariância para a anotação de RNAs não codificantes em conjuntos de dados de nucleotídeos, fornecendo um sistema uniforme para a anotação de RNA, que abrange informações sobre famílias de microRNA (Kalvari *et al.*, 2021).

Para baixar apenas as famílias de tRNA e rRNA de interesse, entre na página do Rfam em https://rfam.org e clique no link “browser”, depois “*entry type*”, selecione as famílias e clique em Submit (Figura 5). Neste manual foi utilizada a versão 14.10 do Rfam (Acessado em 07/2024).

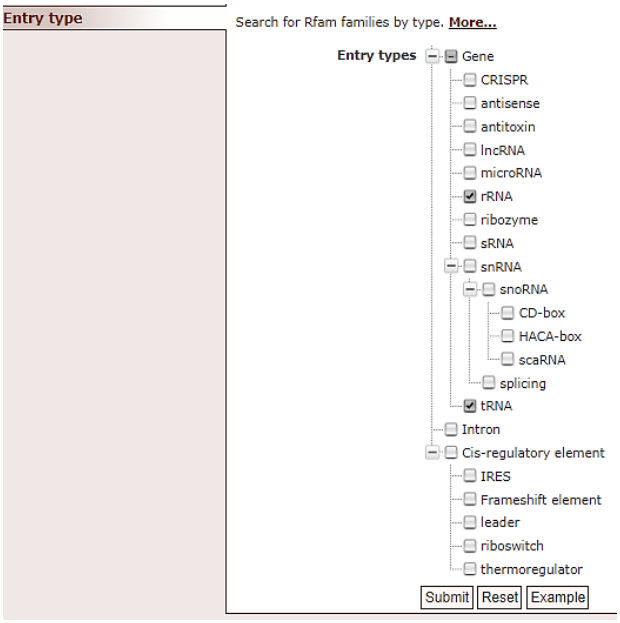


Figura 5. Tipos de famílias de RNAs na base de dados Rfam.

A base de dados exibe as colunas “Accession”, “ID”, “Type” e “Description” das famílias correspondentes aos tRNA e rRNA (Figura 6).

Accession	ID	Type	Description
RF00001	5S_rRNA	Gene; rRNA	5S ribosomal RNA
RF00002	5_8S_rRNA	Gene; rRNA	5.8S ribosomal RNA
RF00005	tRNA	Gene; tRNA	tRNA
RF00177	SSU_rRNA_bacteria	Gene; rRNA	Bacterial small subunit ribosomal RNA
RF01852	tRNA-Sec	Gene; tRNA	Selenocysteine transfer RNA
RF01959	SSU_rRNA_archaea	Gene; rRNA	Archaeal small subunit ribosomal RNA
RF01960	SSU_rRNA_eukarya	Gene; rRNA	Eukaryotic small subunit ribosomal RNA
RF02540	LSU_rRNA_archaea	Gene; rRNA	Archaeal large subunit ribosomal RNA
RF02541	LSU_rRNA_bacteria	Gene; rRNA	Bacterial large subunit ribosomal RNA
RF02542	SSU_rRNA_microsporidia	Gene; rRNA	Microsporidia small subunit ribosomal RNA
RF02543	LSU_rRNA_eukarya	Gene; rRNA	Eukaryotic large subunit ribosomal RNA
RF02545	SSU_trypano_mito	Gene; rRNA	Trypanosomatid mitochondrial small subunit ribosomal RNA
RF02546	LSU_trypano_mito	Gene; rRNA	Trypanosomatid mitochondrial large subunit ribosomal RNA
RF02547	mtPerm-5S	Gene; rRNA	Permuted mitochondrial genome encoded 5S rRNA
RF02554	ppoRNA	Gene; rRNA	mt-5S-like P. polycephalum
RF02555	hveRNA	Gene; rRNA	mt-5S-like H. vermiformis

Figura 6. Dados filtrados e mostrando apenas as famílias de tRNA e rRNA da base de dados do Rfam.

Copiar os dados da primeira coluna (Accession) com os nomes das famílias de tRNA e rRNA para um arquivo chamado nomes_tRNAs_rRNAs.txt, como mostra a Figura 7.

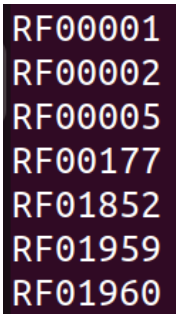


Figura 7. Conteúdo parcial do arquivo nomes_tRNAs_rRNAs.txt.

Para baixar todos os arquivos de famílias de interesse do banco RFAM de forma automática, execute o comando abaixo:

```
#Criar a pasta rfam para organizar os arquivos de rRNA e tRNA

$ mkdir rfam

$ for i in $(cat nomes_tRNAs_rRNAs.txt); do wget https://ftp.ebi.ac.uk/pub/databases/Rfam/CURRENT/fasta_files/$i.fa.gz -O rfam/$i.fa.gz; done
```

Depois de baixado, agrupe todos os arquivos em um único arquivo "*.fa*", usando o comando abaixo:

```
# zcat  Extrai o conteúdo de um arquivo compactado pelo programa gzip

$ zcat rfam/*.fa.gz > rfam/Rfam_tRNA_rRNA.fa
```

Mapeamento com o Bowtie

Para instalar o Bowtie, baixe o pacote na versão 1.3.1 (usado no manual) em https://sourceforge.net/projects/bowtie-bio/files/bowtie, descompacte o arquivo.tar.gz e acesse o diretório descompactado.

```
$ wget https://sourceforge.net/projects/bowtie-bio/files/bowtie/1.3.1/bowtie-1.3.1-linux-x86_64.zip

$ unzip bowtie-1.3.1-linux-x86_64.zip

#Adicionar a pasta do programa bowtie no PATH do linux

$ export PATH=$PATH:$PWD/bowtie-1.3.1-linux-x86_64:
```

Criar o index usando Bowtie:

- **bowtie-build** → para criar um índice de sequência;
- **Rfam_tRNA_rRNA.fa** → arquivo (gerado acima) de entrada contendo as sequências de RNA que serão indexadas;
- **Rfam_tRNA_rRNA** → prefixo que será usado para nomear os arquivos de índice gerados.

```
$ bowtie-build --threads 16 rfam/Rfam_tRNA_rRNA.fa rfam/Rfam_tRNA_rRNA
```

Mapeamento usando o bowtie - Rfam:

- **v** → especifica o número máximo de mismatches permitidos durante o alinhamento das sequências;
- **a** → reporta todos os alinhamentos válidos, em vez de apenas o melhor alinhamento para cada sequência de leitura;
- **best** → instrui o programa a classificar os alinhamentos encontrados e relatar apenas o melhor alinhamento para cada sequência;
- **strata** → controla a exibição de resultados. Quando usada, agrupa os alinhamentos em "estratos", o que significa que alinhamentos idênticos ou equivalentes são agrupados juntos. Isso é útil para reduzir a redundância nos resultados;
- **threads** → número de threads;
- **un** → especifica o nome do arquivo de saída para as sequências que não foram alinhadas durante o processo de alinhamento;
- **1> /dev/null** → redireciona a saída padrão para /dev/null, ou seja, descarta todas as mensagens de saída padrão. Isso é útil para suprimir a saída padrão e manter apenas o arquivo de log para análise posterior;
- **2>** → redireciona a saída de erro padrão de um comando para um arquivo específico.

```
$ mkdir 02-Bowtie

$ for i in {01..02}; do nohup bowtie -v 0 -a --best
--strata --threads 16 -x rfam/Rfam_tRNA_rRNA
01-Cleaned/pig_mir_"$i"_cut_trim.fastq.gz --un
02-Bowtie/pig_mir_"$i"_cut_trim_unaligned_RFAM.
fastq 1> /dev/null 2> 02-Bowtie/pig_mir_"$i"_RFAM.
log & done
```

Os arquivos gerados na saída do comando acima são:

```
pig_mir_01_cut_unaligned_RFAM.fastq
pig_mir_01_RFAM.log
pig_mir_02_cut_unaligned_RFAM.fastq
pig_mir_02_RFAM.log
```

Figura 8. Arquivos gerados pelo mapeamento pelo programa Bowtie .

Usando o mirDeep2

Para usar o mirDeep2, você precisará baixar e instalar o programa. Siga os comandos abaixo para realizar a instalação:

```
#Versão do mirdeep2 utilizada no manual v.
0.1.3

$ git clone https://github.com/rajewsky-lab/mirdeep2

$ cd mirdeep2

$ perl install.pl

$ source ~/.bashrc && perl install.pl

#Adicionar a pasta bin no PATH do Linux

$ export PATH=$PATH:$PWD/bin:

$ cd ..

Para maiores informações sobre a instalação e uso
visite a página do programa em: https://github.com/
rajewsky-lab/mirdeep2
```

Iniciando o mapeamento com mapper

Para iniciar o mapeamento com o mapperl.pl é necessário juntar os dois arquivos gerados pela saída do programa bowtie em um único arquivo fastq, usando o comando a seguir:

```
$ cat 02-Bowtie/*_unaligned_RFAM.fastq >
02-Bowtie/all_cut_trim_unaligned_RFAM.fastq
```

Crie a pasta 02-Mapped para salvar os resultados do mapeamento pelo mirDeep2.

```
$ mkdir 02-Mapped
```

Antes de mapear as sequências no genoma do suíno, primeiro será necessário baixar o genoma e criar o index.

```
# Baixar o genoma pelo site do UCSC

$ mkdir reference

$ wget https://hgdownload.soe.ucsc.edu/golden
Path/susScr11/bigZips/susScr11.fa.gz -O reference/
susScr11.fa.gz

$ gunzip reference/susScr11.fa.gz
```

Criar o index do genoma do suíno, usando o comando bowtie-build:

- **susScr11.fa** → o arquivo de entrada contendo o genoma de referência.
- **susScr11** → prefixo que será usado para nomear os arquivos de índice gerados.

```
$ bowtie-build --threads 16 reference/susScr11.fa
reference/susScr11
```

Após ter criado os índices, os arquivos gerados são mostrados na Figura 9.

```
susScr11.rev.1.ebwt
susScr11.rev.2.ebwt
susScr11.1.ebwt
susScr11.2.ebwt
susScr11.3.ebwt
susScr11.4.ebwt
```

Figura 9. Arquivos de índices gerados pelo bowtie-build.

O mapeamento das sequências é realizado com o programa mapper.pl com os seguintes parâmetros:

- **h** → transforma o fastq em fasta.
- **j** → remove sequências que contenham caracteres diferentes de ACGTN.

- **l** → descarta leituras mais curtas do que um comprimento específico em nucleotídeos (18).
- **m** → colapsa as reads, removendo as sequências duplicadas - para deixar o mapeamento mais rápido.
- **p** → as leituras serão mapeadas para o genoma de referência fornecido, cujo arquivo de índice foi especificado.
- **s** → salva o fasta das reads colapsadas.
- **t** → nome do arquivo dos mapeamentos de leitura para o banco de dados de referência no formato arf do miRDeep2.
- **v** → exibe um relatório de progresso.
- **o** → número de threads.

```
$ mapper.pl 02-Bowtie/all_cut_trim_unaligned_RFAM.fastq -e -h -j -l 18 -m -p reference/susScr11 -s 02-Mapped/all_reads_collapsed.fa -t 02-Mapped/all_reads_collapsed_vs_genome.arf -v -o 64 2> 02-Mapped/report.log

#mover o arquivo de log do bowtie para a pasta 02-Mapped

$ mv bowtie.log 02-Mapped
```

A saída do mapper.pl são dois arquivos de log com os detalhes do mapeamento, o arquivo arf para a entrada no miRdeep2 e o arquivo com as reads colapsadas no formato fasta.

```
all_reads_collapsed.fa
all_reads_collapsed_vs_genome.arf
bowtie.log
report.log
```

Figura 10. Arquivos gerados pela execução do programa mapper.pl.

Identificação e predição dos microRNAs – mirDeep2

Para iniciar a etapa de identificação e predição dos microRNAs é necessário baixar alguns arquivos do banco de dados mirBase. O site miRBase fornece ampla gama de informações sobre microRNAs publicados, incluindo suas sequências, precursores de biogênese, coordenadas do genoma, referências bibliográficas, dados de expressão, de sequenciamento e anotações. O banco de dados do miRBase está disponível publicamente em http://mirbase.org e para este manual foi utilizada a versão 22.1.

Para baixar os arquivos “mature.fa” e “hairpin.fa” dos microRNAs do site miRBase siga os comandos abaixo:

```
$ mkdir mirbase

$ wget https://www.mirbase.org/download/mature.fa -O mirbase/mature.fa

$ wget https://www.mirbase.org/download/hairpin.fa -O mirbase/hairpin.fa
```

As sequências baixadas do banco de dados mirBase contêm a base “U” (uracila) ao invés de “T” (timina). Use o programa “rna2dna.pl” para realizar a conversão das bases conforme o comando abaixo:

```
$ rna2dna.pl mirbase/mature.fa > mirbase/mature.dna.fa

$ rna2dna.pl mirbase/hairpin.fa > mirbase/hairpin.dna.fa
```

Selecionar apenas as sequências de interesse (*sus scrofa* - ssc).

O comando “grep” do Linux vai filtrar todas as linhas nos arquivos “mature.dna.fa” e “ehairpin.dna.fa” que contenham o padrão “ssc” e salvar a sequência fasta de cada *read* nos arquivos respectivos, “exclusivos_ssc.mature.dna.fa” e “exclusivos_ssc.hairpin.dna.fa”. Para isso, use os seguintes parâmetros:

- **grep “ssc”** → filtra as linhas contendo o padrão “ssc”.
- **A1** → inclui a linha que corresponde ao padrão e a linha seguinte (para capturar a sequência completa no formato fasta).
- **no-group-separator** → não adicione caracteres como separador de grupos.
- **exclusivos_ssc.mature.dna.fa** → arquivo de saída com os microRNAs maduros.
- **exclusivos_ssc.hairpin.dna.fa** → arquivo de saída com os precursores dos microRNAs.

```
$ grep -A1 --no-group-separator “ssc” mirbase/mature.dna.fa > mirbase/exclusivos_ssc.mature.dna.fa

$ grep -A1 --no-group-separator “ssc” mirbase/hairpin.dna.fa > mirbase/exclusivos_ssc.hairpin.dna.fa
```

Crie uma pasta 03-microRNA.

```
$ mkdir 03-microRNA
```

Para executar o miRDeep2 e identificar microRNAs novos e conhecidos, execute os comandos abaixo:

- **all_reads_collapsed.fa** → arquivo com sequências das reads mapeadas.
- **susScr11.fa** → arquivo fasta do genoma de referência.
- **all_reads_collapsed_vs_genome.arf** → arquivo de mapeamento das reads contra o genoma.
- **exclusivos_ssc.mature.dna.fa** → microRNAs maduros exclusivos da espécie de interesse identificados.
- **none** → como não temos um arquivo de outros miRNAs conhecidos relacionados, deve-se colocar “none” nessa posição.
- **exclusivos_ssc.hairpin.dna.fa** → precursores de microRNAs exclusivos da espécie de interesse identificados.
- **b** → usa o valor mínimo de cut-off para o mirdeep2 score.
- **v** → remove pastas e arquivos temporários.

```
$ cd  03-microRNA

$ nohup miRDeep2.pl ../02-Mapped/all_reads_collapsed.fa ../reference/susScr11.fa ../02-Mapped/all_reads_collapsed_vs_genome.arf ../mirbase/exclusivos_ssc.mature.dna.fa none  ../mirbase/exclusivos_ssc.hairpin.dna.fa -v -b 4 2> report.log &
```

No nosso caso, como não há uma opção específica disponível para o genoma do suíno no miRDeep2, o parâmetro -t não foi utilizado.

Serão gerados os arquivos:

- miRNAs_expressed_data_da_analise.csv.
- error_data_da_analise.log.
- result_data_da_analise.html.
- Pasta pdfs_data_da_analise.
- Pasta chamada mirna_results_data_da_analise/ com o arquivo dos novos.

O mirDeep2 gera pdfs informativos sobre os microRNAs.

A Figura 11 mostra um novo microRNA em suínos relatado pelo miRDeep2. No canto superior esquerdo, o programa fornece o detalhamento da pontuação miRDeep2 para o microRNA relatado, juntamente com contagens de leitura para a sequência madura, loop e star. No canto superior direito, a figura mostra a estrutura secundária predita do hairpin do microRNA. A estrutura é codificada por cores para representar as diferentes regiões do hairpin associadas à biogênese do microRNA, como a região madura (vermelha), loop (amarela) e star (azul). As linhas pontilhadas abaixo da sequência do microRNA mostram as leituras alinhadas. Nucleotídeos incompatíveis são apresentados em letras maiúsculas e o parâmetro “mm” indica o número de incompatibilidades. Essas informações combinadas fornecem uma compreensão abrangente do novo microRNA.

O arquivo gerado com a extensão .html contém as informações dos miRNAs preditos na análise e pode ser aberto utilizando um navegador. Para manter somente os miRNAs preditos confiáveis nos resultados, a coluna que contém “no” para a informação de “randfold p-value” será removida. Para fazer isso, execute o comando abaixo:

```
$ awk -F’\t’ ‘/novel miRNAs predicted by miRDeep2/ {flag=1; next} flag && $0 == “”{exit} flag && $9 == “yes”’ result_*.csv | cut -f1 > novel_mirnas.txt

$ for i in $(cat novel_mirnas.txt); do grep  -A1 --no-group-separator “^>$i$” mirna_results_*/novel_mature_*_na.fa; done > novel_mirnas.fa
```

Com a execução do comando acima, nosso exemplo gerou um conjunto de 43 novos microRNAs preditos.

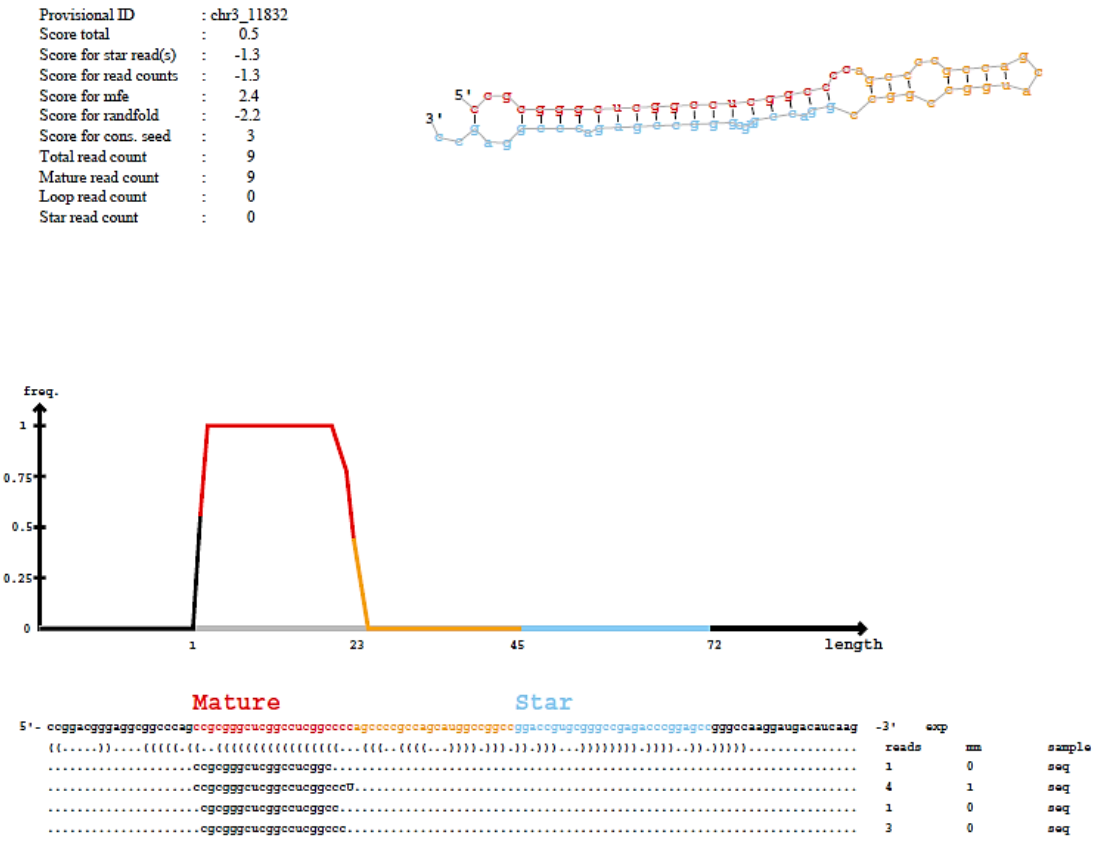


Figura 11. Pdf gerado pelo mirDeep2 para predição de um novo microRNA em suínos.

Considerações finais

A análise bioinformática de dados de sequenciamento de microRNA é essencial para desvendar os complexos mecanismos de regulação genética e entender seu papel em uma variedade de processos biológicos. Ao seguir o fluxo de trabalho mencionado neste manual, os pesquisadores poderão identificar novos microRNAs que podem estar envolvidos em vias biológicas e processos associados à sua regulação. Essa análise pode fornecer insights valiosos, permitindo uma compreensão mais abrangente dos sistemas biológicos, um melhor entendimento das interações entre diferentes níveis de regulação genética e um direcionamento mais aprimorado de pesquisas futuras.

Agradecimentos

Ao Instituto Nacional de Ciência e Tecnologia/ Ciência Animal (INCT/CA, Processo CNPq Processo 465377/2014-9) da Universidade Federal de Viçosa (UFV). A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES)/PROEX Processo 88887.844747/2023-00 pelo financiamento do projeto, pelo apoio ao projeto (Código de Financiamento 001) e pela bolsa de FGC. Ao CNPq pelo financiamento do projeto (Processo 402935/2021), pela bolsa de produtividade de MCL, AMGI e SEFG e pela bolsa de pós-doutorado de HCO.

Referências

BARTEL, D. P. Metazoan microRNAs. **Cell**, v. 173, n. 1, p. 20-51, 2018.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114-2120, 2014.

FRIEDLÄNDER, M. R.; MACKOWIAK, S. D.; LI, N.; CHEN, W.; RAJEWSKY, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. **Nucleic Acids Research**, v. 40, n. 1, p. 37-52, 2012.

JONAS, S., IZAURRALDE, E. Towards a molecular understanding of microRNA-mediated gene silencing. **Nature Reviews Genetics**, v.16, n. 7, p. 421-433, 2015.

KALVARI, I.; NAWROCKI, E. P.; ONTIVEROS-PALACIOS, N.; ARGASINSKA, J.; LAMKIEWICZ, K.; MARZ, M.; PETROV, A. I. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. **Nucleic Acids Research**, v. 49, n. D1, p. D192-D200, 2021.

KIM, Y. K.; KIM, V. N. Processing of intronic microRNAs. **The EMBO Journal**, v. 26, n. 3, p. 775-783, 2007.

KOZOMARA, A.; BIRGAOANU, M; GRIFFITHS-JONES, S. miRBase: from microRNA sequences to function. **Nucleic Acids Research**, v. 47, n. D1, p. D155-D162, 2019.

KROL, J.; LOEDIGE, I.; FILIPOWICZ, W. The widespread regulation of microRNA biogenesis, function and decay. **Nature Reviews Genetics**, v. 11, n. 9, p. 597-610, 2010.

LANGMEAD, B. Aligning short sequencing reads with Bowtie. **Current protocols in bioinformatics**, v. 32, n. 1, p. 11-7, 2010.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows–Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754-1760, 2009.

LI, Y., HUANG, J., GUO, M., ZUO, X. MicroRNAs regulating signaling pathways: potential biomarkers in systemic sclerosis. **Genomics, proteomics & bioinformatics**, v. 13, n. 4, p. 234-241, 2015.

LIN, X. Z.; LUO, J.; ZHANG, L. P.; WANG, W.; SHI, H. B.; ZHU, J. J. MiR-27a suppresses triglyceride accumulation and affects gene mRNA expression associated with fat metabolism in dairy goat mammary gland epithelial cells. **Gene**, v. 521, p. 15-23, 2013.

LÓPEZ-JIMÉNEZ, E.; ANDRÉS-LEÓN, E. The implications of ncRNAs in the development of human diseases. **Non-coding RNA**, v. 7, n. 1, p. 17, 2021.

MANZINI, G. An analysis of the Burrows—Wheeler transform. **Journal of the ACM**, v. 48, n. 3, p. 407-430, 2001.

MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet. Journal**, v. 17, n. 1, p. 10-12, 2011.

O'BRIEN, J.; HAYDER, H.; ZAYED, Y.; PENG, C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. **Frontiers in endocrinology**, v. 9, 2018.

TAY, Y., RINN, J., PANDOLFI, P. P. The multilayered complexity of ceRNA crosstalk and competition. **Nature,** v. 505, n. 7483, p. 344-352, 2014.

YE, J.; LI, J.; ZHAO, P. Roles of ncRNAs as ceRNAs in gastric cancer. **Genes**, v. 12, n. 7, p. 1036, 2021.

ZHANG, P.; WU, W.; CHEN, Q.; CHEN, M. Non-coding RNAs and their integrated networks. **Journal of Integrative Bioinformatics**, v. 16, n. 3, 2019.