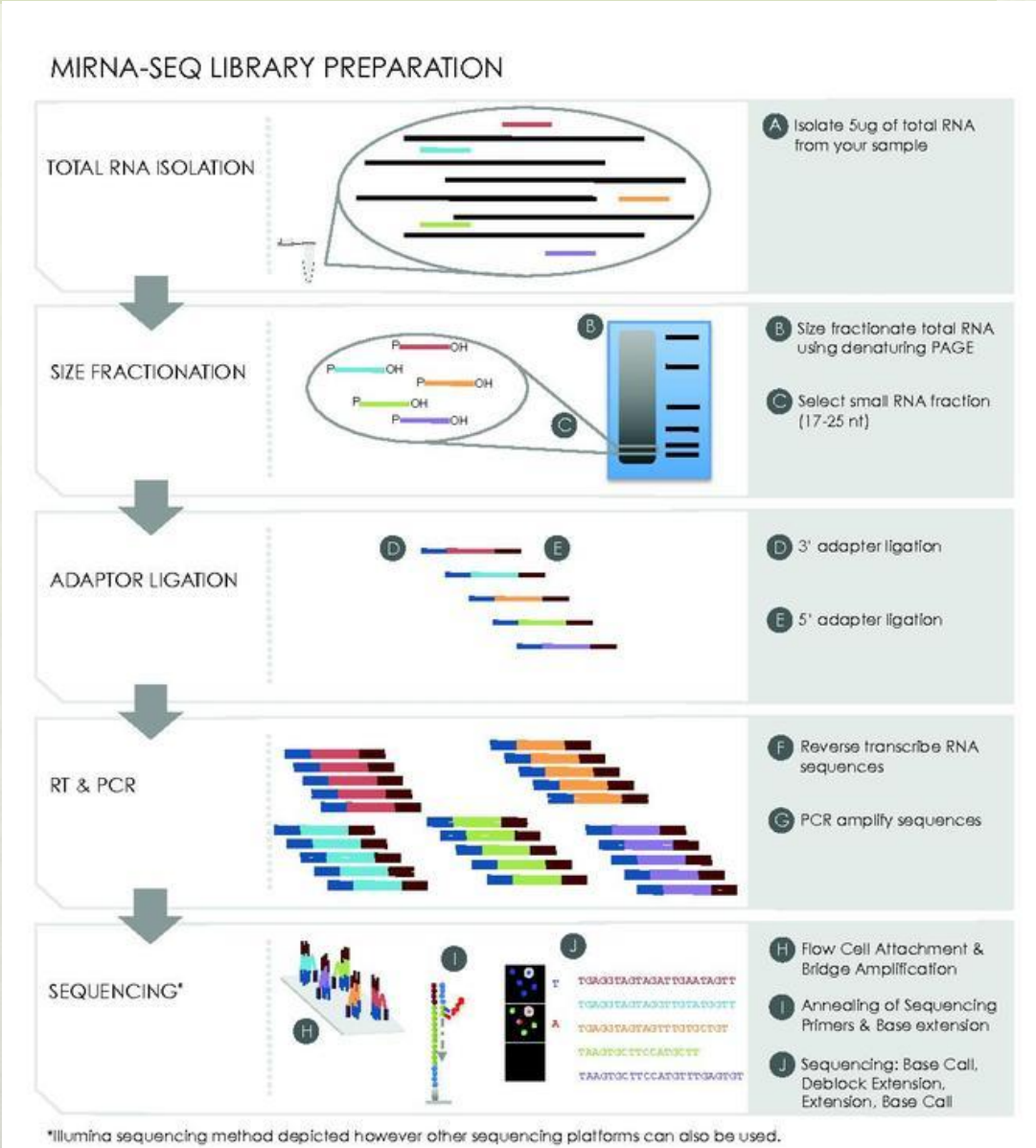
# microRNAs y secuenciación masiva (PII)

Genómica Funcional
Máster en Genética y Evolución

# Secuenciación masiva & genes de microRNAs

- Extracción del RNA total
- Ligar adaptadores
- Seleccionar por longitud
- Generar una librería de cDNA
- Secuenciar
- Análisis bioinformático




**Source:**
http://en.wikipedia.org/wiki/MicroRNA_Sequencing

# Análisis bioinformático

## El punto de partida

### Las secuencias de las lecturas en formato fastq

Cada lectura está representada por 4 líneas

```
@SRR037876 GSM522374_1:1:148:931:861
TAGTTCTACAGTCCGACGATCTCGTATGCCGTCTTC
+
BB@+?0:4@B@-@/A<3A7@-=@<1=@87=?<==9#
```

Secuencia de la lectura

Calidad de la lectura

**La longitud de las lecturas depende del número de ciclos en las secuenciación, frecuentemente entre 36 y 50 nucleótidos**

# Flujo del trabajo

## 1) Control de calidad

- Eliminar lecturas con baja calidad
  → **Muy importante** para la detección de variación de secuencia
  → **Potencialmente importante en la detección de isomiRs**

## 2) Detectar y eliminar el adaptador

- Se secuencia parte del adaptador si la molécula es mas corta que la lecturas (número de ciclos)
- El adaptador no alinea frente al genoma
- Hay que eliminarlo para poder mapear la lectura y detectar isomiRs

## 3) Colapsar las lecturas

- Unir lecturas únicas en una entrada única que consiste en
  - Secuencia & conteo (read count) (las veces una secuencia fue observada en un experimento)

## 4) Alinear las lecturas únicas frente a una referencia

# Quality control

```
@SRR037876 GSM522374_1:1:148:931:861
TAGTTCTACAGTCCGACGATCTCGTATGCCGTCTTC
+
BB@+?0:4@B@-@/A<3A7@-=@<1=@87=?<==9#
```

The base call quality is ASCII encoded
B=66; @=64, ?=63, etc1

**Convert to numbers**

66|66|64|43|63 …

**Convert to Phred Score (Q)**

Subtract X from this number
**Q(B) = 66-33 = 33**
X frequently is 33 but it might depend on the vendors protocol

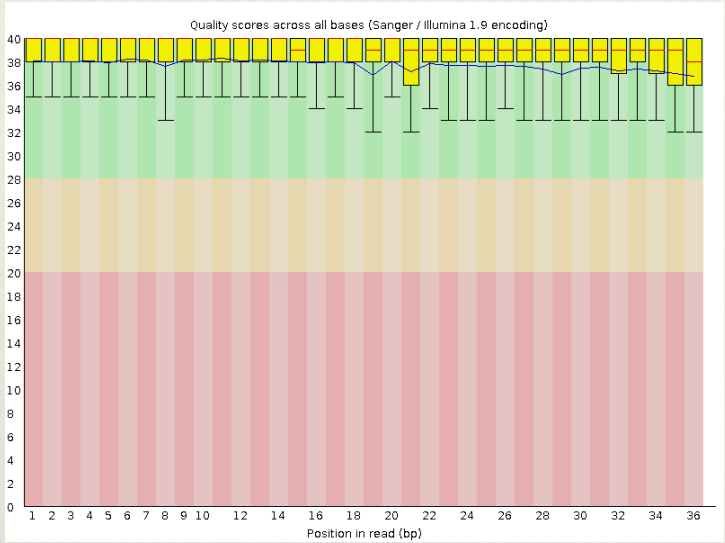
33|33|31|10|30 …

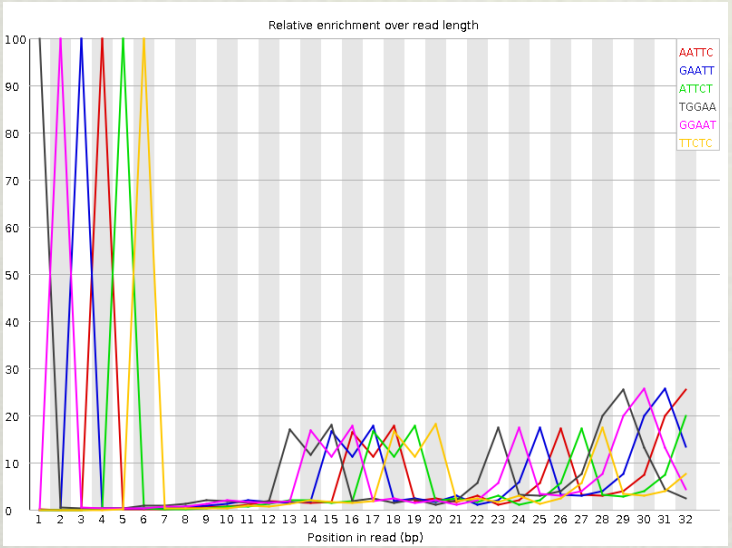
**Several possibilities**

- Trim the reads at the first base with Q < q
- Remove reads with mean Q < q
- Remove reads with min Q < q
- Remove reads if the number of bases with Q < q is over given threshold

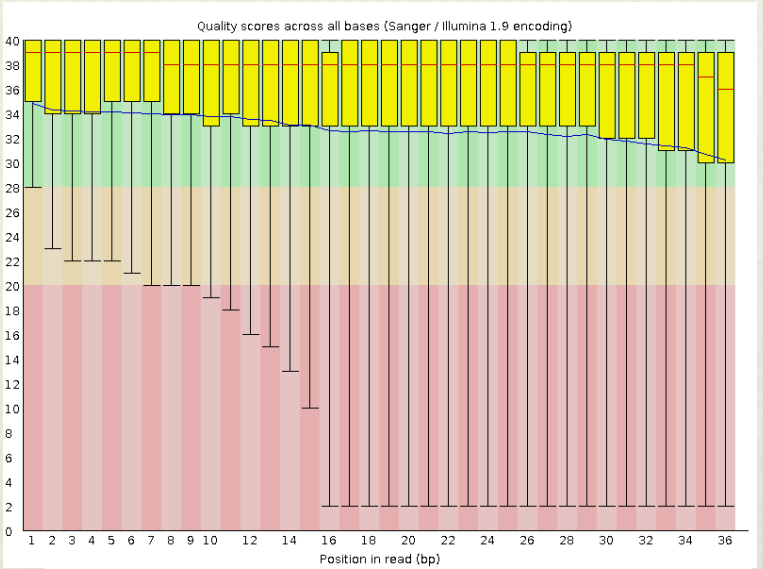
| Q | P (incorrect BC) | BC accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1000 | 99.9% |
| 40 | 1 in 10000 | 99.99% |

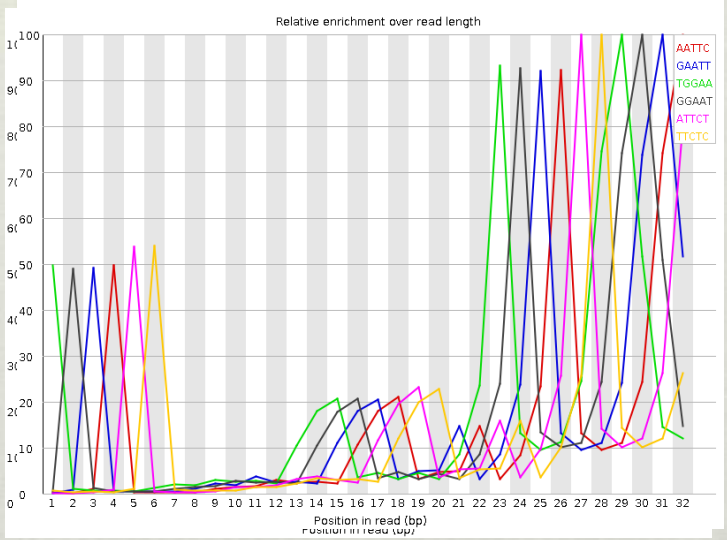
# Have a first glimpse on the data: fastqc




- Very good sequencing quality
- Suspicious overrepresentation of k-mers




- “Acceptable” sequencing quality
- Less k-mer bias at initial positions

# Detectar el adaptador

**Mature microRNAs are around 21 nt long --- in the example below, the reads are 50 nt long**

**→(part of) the adapter is sequenced as well**

```
@SRR518946.15 DA19881:1:30F8JAAXX:8:1:8:919 length=50
TAGCTTATCAGACTGATGTTGACTCGTATGCCGTCTTCTGCTTGTGTGTT
+SRR518946.15 DA19881:1:30F8JAAXX:8:1:8:919 length=50
BBBB<BBBABCABBABB@@B>@?=B@7:@6A?=8>B>7?##########
```

**Align the adapter sequence to the read**

3' RNA Adapter
5' P-UCGUAUGCCGUCUUCUGCUUGU

**Parameters:**

- **Minimum length of detected adapter sequence (10 nt)**
- **Number of allowed mismatches**

```
@SRR518946.15 DA19881:1:30F8JAAXX:8:1:8:919 length=50
TAGCTTATCAGACTGATGTTGACTCGTATGCCGTCTTCTGCTTGTGTGTT
+SRR518946.15 DA19881:1:30F8JAAXX:8:1:8:919 length=50
BBBB<BBBABCABBABB@@B>@?=B@7:@6A?=8>B>7?##########
```

# Detectar el adaptador

```
@
TAAGTGGGAGGCCCTCGTATGCCGTCTTCTGCTTGTAAAAAAAAAAAATA
+
BCA6<>>BBAB?AC@AABACBAB?'>A:A>?@A@################
@
TGAGGTAGTAGATTGTATAGTTTCGTATGCCGTCTTCTGCTTGATTATGT
+
BC@2ABCC?BBA?=BABB??ABB@B@=A@?@B?AB;##############
@
TCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAATAATTTTTTTTTTTTTT
+
B@;>?BBB?3?BBBB@9@A?0<AA##########################
@
TTCAAGTAATCCAGGATAGGCTTCGTATGCCGTCTTCTGCTTTAATTTTT
+
>CBCBBA?@@CBB@?BA@7:?A7=>8/:7<>=29################
@
TAATACTGCCTGGTAATGATGACTCGTATGCCGTCTTCTGCTTGTTGTGG
+
BCCCCCBCCCCCC?>ACCCBCCBAB>>@@BBAB=;;?=B@##########
@
TGAGGTAGTAGATTGTATAGTTTCGTATGCCGTCTTCTGCTTGATTTTTT
+
BCAAA?BC:6<AABA>@:98=B:AA@A9>>@;??################
@
TAGCTTATCAGACTGATGTTGACTCGTATGCCGTCTTCTGCTTGTGTGTT
+
BBBB<BBBABCABBABB@@B>@?=B@7:@6A?=8>B>7?##########
```

# Detectar el adaptador

```
@
TAAGTGGGAGGCCC
+
BCA6<>>BBAB?AC
@
TGAGGTAGTAGATTGTATAGTT
+
BC@2ABCC?BBA?=BABB??AB
@

+

@
TTCAAGTAATCCAGGATAGGCTTCGTATGCCGTCTTCTGCTTTAATTTTT
+
>CBCBBA?@@CBB@?BA@7:?A7=>8/:7<>=29################
@
TAATACTGCCTGGTAATGATGAC
+
BCCCCCBCCCCCC?>ACCCBCCB
@
TGAGGTAGTAGATTGTATAGTT
+
BCAAA?BC:6<AABA>@:98=B
@
TAGCTTATCAGACTGATGTTGAC
+
BBBB<BBBABCABBABB@@B>@?
```

14 nt

22 nt

0 nt

Not trimmed (50 nt)

22 nt

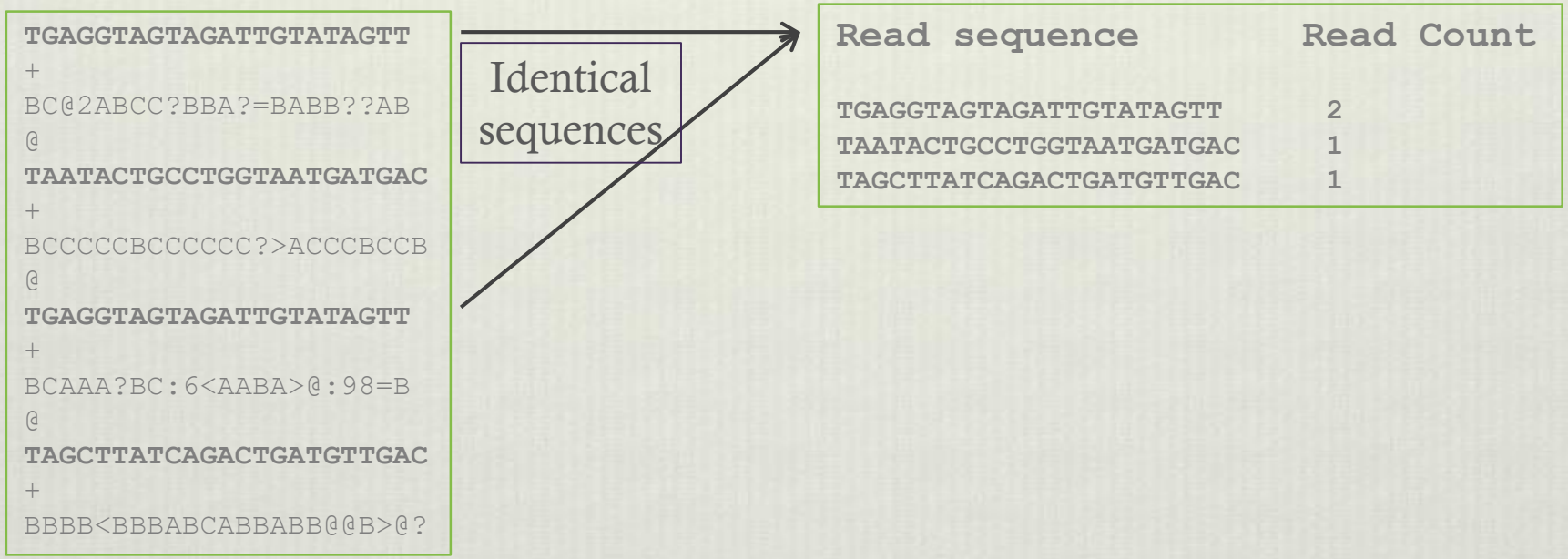
22 nt

23 nt

# Colapsar las lecturas

Convert adapter trimmed fastq file into read/count format

→ **Collapse "redundant reads" into unique reads & read count**



# Ficheros de entrada

## Fasta:

```
>1999420#1462
TGAGATGAAGCACTGTAGAAAA
>443945#1281
GGGAGCATCTCTCGGTCTATGCTGT
>633088#562
TAGATGAAGCACTGTAGCTCTT
>255762#230
GCATTGGTGGTAGAATTCTCGCC
>516042#97
TCAGATGAAGCACTGTAGCTCTT
>1582566#86
TAGATGAATCACTGTAGCTC
>1462753#79
TGGAATTATGGAAAATGACAGATGGC
>625879#40
GTTAAGATATCCCGGACGAGCCC
>517214#8
TCCTTTGGTATAGTGGTGAGTATCCC
>626077#2
TGACTTTGACCTGAGAGAAGAAGGC
```

## Read/count

```
TGAGATGAAGCACTGTAGAAAA   1462
GGGAGCATCTCTCGGTCTATGCTGT      1281
TAGATGAAGCACTGTAGCTCTT    562
GCATTGGTGGTAGAATTCTCGCC   230
TCAGATGAAGCACTGTAGCTCTT    97
TAGATGAATCACTGTAGCTC       86
TGGAATTATGGAAAATGACAGATGGC       79
GTTAAGATATCCCGGACGAGCCC    40
TCCTTTGGTATAGTGGTGAGTATCCC        8
TGACTTTGACCTGAGAGAAGAAGGC         2
```

# Asignar lecturas

```
>1999420#1462
TGAGATGAAGCACTGTAGAAAA
>443945#1281
GGGAGCATCTCTCGGTCTATGCTGT
>633088#562
TAGATGAAGCACTGTAGCTCTT
>255762#230
GCATTGGTGGTAGAATTCTCGCC
>516042#97
TCAGATGAAGCACTGTAGCTCTT
>1582566#86
TAGATGAATCACTGTAGCTC
```

**From which RNAs are these reads derived?**

↓

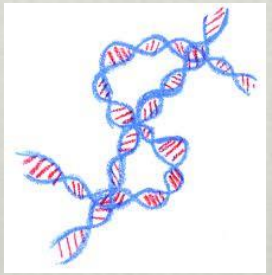
**Assign reads to a reference library**

1

Map reads to a set of known small RNA sequences

2

Map to the genome & genome annotations



# Asignar lecturas

1

**Input: read/count**

```
>1999420#12682
TGAGGTAGTAGGTTGTGTGGTT
>633088#5692
TGAGATGAAGCACTGTAGCTC
>255762#2630
TGAGGTAGTAGGTTGTATAGTT
>516042#1297
TCAGATGAAGCACTGTAGCTCTT
>443945#181
TGAGGTAGTAGGTTGTGTGGT
>1582566#86
TGAGGTAGTAGGTTGTATAGT
>1462753#79
TGGAATTATGGAAAATGACAGATGGC
>625879#40
GTTAAGATATCCCGGACGAGCCC
>517214#8
TACCCTGTAGAACCGAATTTGTG
>626077#2
TGACTTTGACCTGAGAGAAGAAGGC
```

**Alignment**
**Bowtie, Blast, etc**

**Sequence library (miRBase)**

```
>hsa-let-7b-5p
TGAGGTAGTAGGTTGTGTGGTT

> hsa-miR-143-3p
TGAGATGAAGCACTGTAGCTC

> hsa-let-7a-5p
TGAGGTAGTAGGTTGTATAGTT

> hsa-miR-509-3p
TGATTGGTACGTCTGTGGGTAG

> hsa-miR-10b-5p
TACCCTGTAGAACCGAATTTGTG
```

# Asignar lecturas

1

**Input: read/count**

>1999420#**12682 → read count**
TGAGGTAGTAGGTTGTGTGGTT
>633088#5692
TGAGATGAAGCACTGTAGCTC
>255762#2630
TGAGGTAGTAGGTTGTATAGTT
>516042#1297
TCAGATGAAGCACTGTAGCTCTT
>443945#181
TGAGGTAGTAGGTTGTGTGGT
>1582566#86
TGAGGTAGTAGGTTGTATAGT
>1462753#79
TGGAATTATGGAAAATGACAGATGGC
>625879#40
GTTAAGATATCCCGGACGAGCCC
>517214#8
TACCCTGTAGAACCGAATTTGTG
>626077#2
TGACTTTGACCTGAGAGAAGAAGGC

**Alignment**
**Bowtie, Blast, etc**

**Sequence library (miRBase)**

>hsa-let-7b-5p
TGAGGTAGTAGGTTGTGTGGTT

> hsa-miR-143-3p
TGAGATGAAGCACTGTAGCTC

> hsa-let-7a-5p
TGAGGTAGTAGGTTGTATAGTT

> hsa-miR-509-3p
TGATTGGTACGTCTGTGGGTAG

> hsa-miR-10b-5p
TACCCTGTAGAACCGAATTTGTG

**Sum the read count of all mapped reads**

| Name | Read Count |
|---|---|
| hsa-let-7b-5p | 12863 |
| hsa-miR-143-3p | 5692 |
| hsa-let-7a-5p | 2716 |
| hsa-miR-10b-5p | 8 |

# Differential Expression

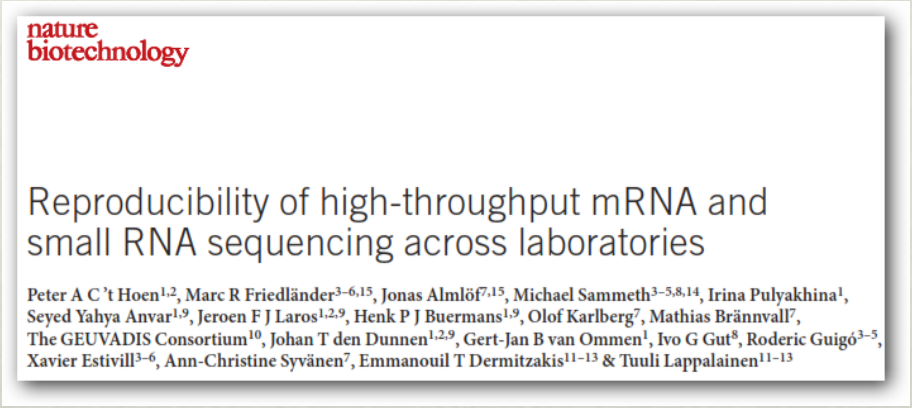
**Compare the expression profiles between two conditions**

The total yield of reads can (will) differ between two different samples → read counts cannot be used for comparison

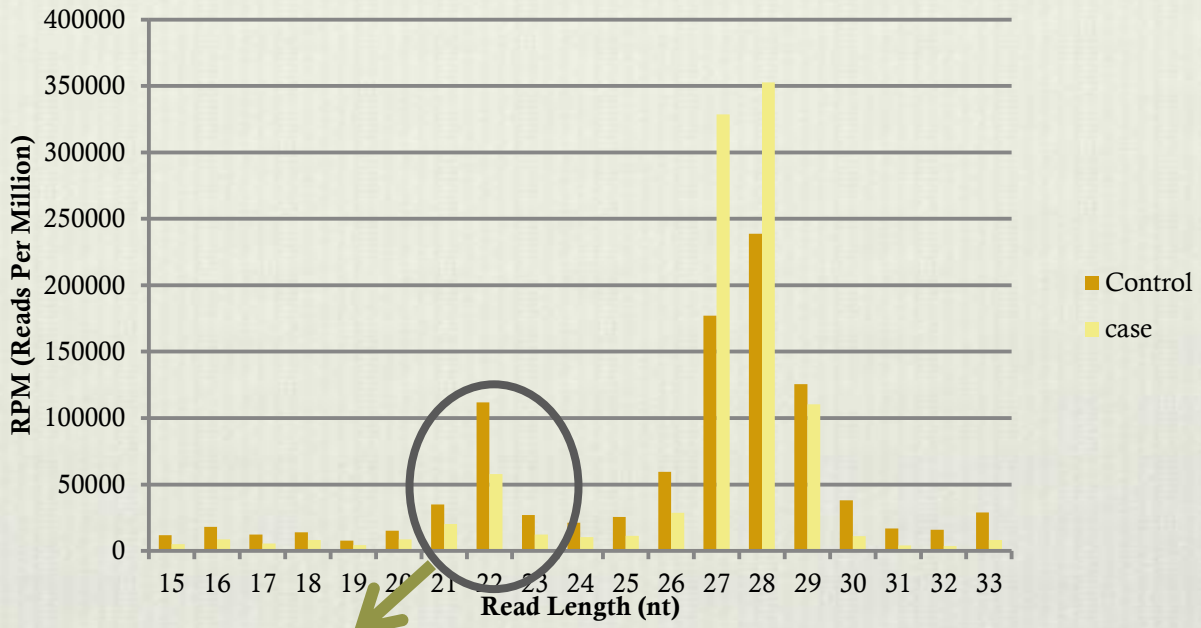
**Two possibilities:**

1. **The expression measure needs to be independent of the total number of reads (Reads Per Million)**

2. **The expression values need to be scaled, i.e. make the total number of reads the same in all conditions/samples**

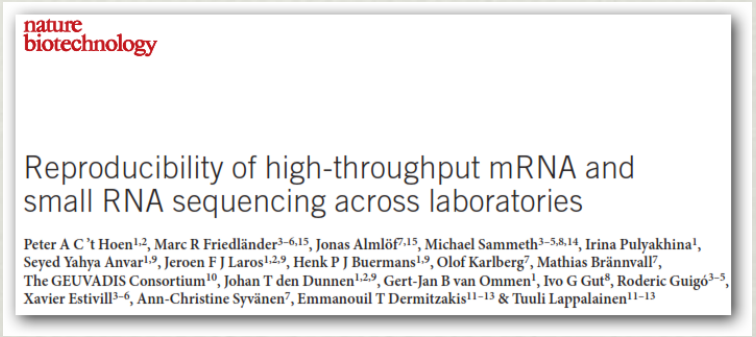


# Differential Expression




- Mature microRNAs are '**relatively**' less frequent in cases than in controls
- Could be due to overexpression of 27-29 nt RNAs in cases, i.e. the absolute abundance needs not to be different between cases and controls!

→ **Normalize for "each used library" separately!**


nature biotechnology

Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories

Peter A C 't Hoen[1,2], Marc R Friedländer[3–6,15], Jonas Almlöf[7,15], Michael Sammeth[3–5,8,14], Irina Pulyakhina[1], Seyed Yahya Anvar[1,9], Jeroen F J Laros[1,2,9], Henk P J Buermans[1,9], Olof Karlberg[7], Mathias Brännvall[7], The GEUVADIS Consortium[10], Johan T den Dunnen[1,2,9], Gert-Jan B van Ommen[1], Ivo G Gut[8], Roderic Guigó[3–5], Xavier Estivill[3–6], Ann-Christine Syvänen[7], Emmanouil T Dermitzakis[11–13] & Tuuli Lappalainen[11–13]

RPM of i'th microRNA: $$RPM_i = 10^6 \frac{RC_i}{RC_{microRNAs}}$$

# isomiRs

hsa-mir-99a

```
CCATTGGCATAAACCCGTAGATCCGATCTTGTGGTGAAGTGGACCGCACAAGCTCGCTTCTATGGGTCTGTGTCAGTGTG  7,156
.((((((((((.((((((((((..(((.((((((.((........)))))))).)))..))))))))).)))))))))..
           AACCCGTAGATCCGATCTTGT                                    ← 3′ length variant          3,815
           AACCCGTAGATCCGATCTTGTG                                   ← hsa-miR-99a-5p               961
           AACCCGTAGATCCGATCTTGTGA                                  ← Non-templated addition (A)   922
           AACCCGTAGATCCGATCTTG                                     ← 3′ length variant            266
           AACCCGTAGATCCGATCTTGTGT                                  ← Non-templated addition (U/T) 227
          AAACCCGTAGATCCGATCTTGT                                    ← 5′ length variant             74
           AACCTGTAGATCCGATCTTGT                                    ← 3′ length variant             65
           AACCCGTAGATCCGATCTT                                      ← 3′ length variant             54
           AACTCGTAGATCCGATCTTGT                                    ← 3′ length variant             30
           AACCCGTAGATCCGATCTTGTA                                   ← Non-templated addition (A)    27
           AACCCGTAGATCCGATCTTGA                                    ← Non-templated addition (A)    27
           AACCCGTAGATCCGATCTTGC                                    ← Non-templated addition (C)    25
           AACCCGTAGATCCGATCT                                       ← 3′ length variant             23
           AACCCTTAGATCCGATCTTGT                                    ← 3′ length variant             23
           AACCCGTAGATCCTATCTTGT                                    ← 3′ length variant             21

                                      CAAGCTCGCTTCTATGGGTCTGT                                        21
                                      CAAGCTCGCTTCTATGGGTCTGA                                        21

           AACCCGTAGATCCGATCTTGTGAA                                 ← Non-templated addition (2 A)  17
               CGTAGATCCGATCTTGT                                    ← Multiple length variant       12
           AACCCGTAGATCCGATCTTGTAA                                  ← Non-templated addition (2 A)  11

                                      CAAGCTCGCTTCTATGGGTCTG                                         11
```

# Detectar nuevos microRNAs


- Drosha/Dicer (DCL in plants) processing patterns can be detected
- Both mature microRNAs (both arms) are represented in the sample?
- 5' end of the mature microRNA shows less fluctuation
- Virtually all reads are organized in one or two clusters